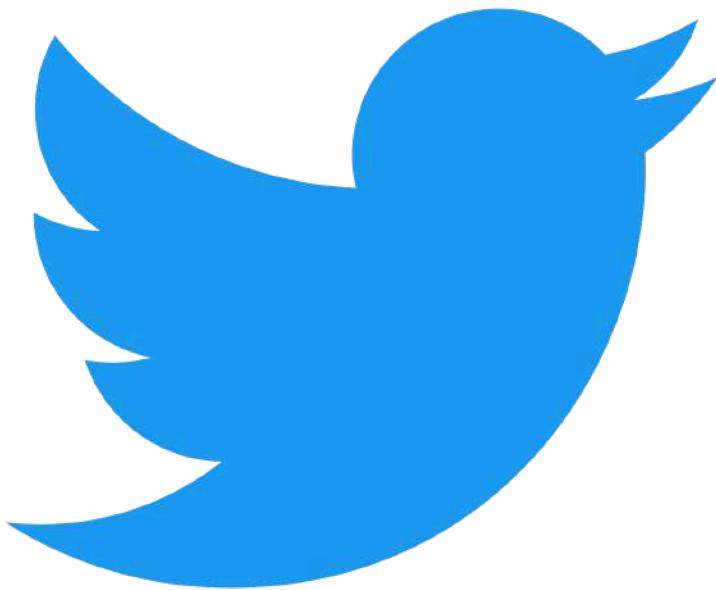


Group 12, Team 3 “Twitter”



D4. Final Report

D4. Final Report

PMAAD - Preprocessat i Models Avançats d'Anàlisi de Dades

GROUP COMPONENTS

Aubach Altes, Artur

Fernández Benavente, Julia

Minguella Torra, Maria

Park Gonzalez, Nicolas Gil

Planell Bosch, Noel Nathan

INDEX

GROUP COMPONENTS	1
INDEX	1
MOTIVATION OF THE PROJECT	8
DESCRIPTION OF DATA	9
Data Source and Process	9
Data Structure	9
Exclusion and Inclusion Criteria	9
Metadata	9
Descriptive Analysis of Raw Data	14
WORKING PLAN	15
DIVISION OF TASKS	15
PREPROCESSING	19
Data Acquisition and Data Homologation	19
Data Cleaning	19
Feature Extraction	20
Feature Selection	22
Correlation matrix	22
Decision Tree	23
Random Forest	24
PCA	25

Selecting the optimal number of principal components	25
Analysis of our PCA	28
Projection of categorical variables	30
Conclusion	31
Missing Data Treatment	32
Deterministic Imputation	32
Types of Missing Values	33
Looking for patterns	33
Little's Test	34
Choosing the best imputation method	35
MICE - Multiple Imputation by Chained Equations	36
MiMMi - Mixed Intelligent-Multivariate Missing Imputation	38
Random Forest	39
Comparison of MICE, MiMMi and RF	40
Results	44
OUTLIERS	46
Preliminary Steps	46
Strategy	47
Univariate Approach	47
Bivariate Approach	48
Mahalanobis Distance	49
Multivariate Approach (PCA)	50
LOF (Local Outlier Factor)	51
Isolation Forest	51
Outlier Treatment	54
Execution	55
Descriptive analysis	56
Univariate analysis	56
User's profile	57
Link basic color	57
Sidebar basic color	58
Created	59
User	61
Gender	61
Gender.confidence	62
Unit state	64
Trusted judgement	65
Continent	66

User's behavior	67
Retweet count	67
Tweet count	68
Favorite number	69
Tweet average word length	70
Tweet word count	71
Description word count	72
Description average word length	73
Bivariate analysis	74
retweet_count x fav_number	76
tweet_length x fav_number	78
link_basic_color x sidebar_basic_color	80
gender x sidebar_basic_color	82
gender x link_basic_color	84
gender x fav_number	86
gender x tweet_length	88
MCA - Multiple Correspondence Analysis	90
Main Objective	90
Preparation	90
Analysis of Results	94
Variability Explanation	97
Associations between variables	97
Position in plot	101
Relationships between variables	101
Individual Plots	103
Summary	105
TIME-SERIES-CLUSTERING	106
Preliminary Steps	107
Analysis of Hypothesis 1	107
Visualization of Collaborator's Confidence in Determining User's Gender Throughout the Year	107
Visualization of Collaborator's Confidence in Determining User's Gender Grouped by Months	108
Analysis of Hypothesis 2	111
Clustering	111
Visualization of Collaborator's Confidence in Determining User's Gender Grouped by Months and Clustering	114
Summary	116
CLUSTERING	117

CURE	117
Our implementation: GOWER's distance	118
CURE results	119
DBSCAN	122
Ensuring that its use is appropriate	122
Tuning of hyperparameters	123
Analyzing the results	124
OPTICS	128
Tuning of hyperparameters	128
Analysing the results	130
PROFILING	132
CURE	132
Filtering out irrelevant variables	132
Random Forest	132
Inference tests	134
Link basic color	135
Sidebar basic color	136
Gender	138
Gender confidence	139
Continent	141
Privacy	142
Tweet count	143
Favorite number	145
Tweet average word length	146
Tweet word count	148
Description word count	149
Description average word length	151
CPG	153
TLP	155
Traffic Lights Panel	155
User	156
User's Behaviour	157
Annotated-Traffic Lights Panel	160
Bivariate analysis profiling	162
User's profile	162
User	162
User's behaviour	163
Cluster template	165

DBSCAN	167
Filtering out irrelevant variables	167
Bivariate plots and inference tests	167
Link basic color	168
Sidebar basic color	170
Created	172
Gender	174
Gender.confidence	176
Continent	178
Privacy	180
Tweet count	182
Favorite number	184
Tweet average word length	186
Tweet word count	188
Description word count	190
Description average word length	192
CPG	195
TLP	196
Cluster template	198
TEXT MINING	200
PREPROCESSING	200
Spam Cleaning	201
Language Detection And Translation	201
Character And Number Cleaning	202
Tokenization	203
Stemming	203
Lemmatizer	204
Stopwords	205
Results	205
Final analysis	207
SENTIMENT ANALYSIS	208
Word Frequency by Gender	209
Unique Words by Gender	212
Male Unique Words	212
Female Unique Words	213
Brand Unique Words	215
Unknown Unique Words	216
Polarity Scores	218

Sentiment Scores	223
Summary	232
LSA	233
Distance between documents	233
Relationship between words	234
Politicians	234
Country	235
Coding	236
Ideology	238
Singers	239
Others	239
Conclusions	240
LDA	241
Choosing Number of Topics	241
Topic Visualization with PCA	247
Topic Description	248
FACTORIAL ANALYSIS	249
Correspondence analysis	249
Results	250
Text	250
Plots	251
Description	252
Plots	253
Correspondence analysis on generalised lexical table	254
Text	254
Description	257
CLUSTERING DOCUMENTS	261
Clustering the Tweets	261
Word Cloud	261
Term Document Matrix	262
K-Means Clustering	264
Determining the Number of Clusters	264
Clustering Results	266
Visualizing with PCA	266
Cluster Similarity Matrix	270
Cluster Word Clouds	273
Conclusions	276
Hierarchical Clustering	277

Clustering the User Description	278
Word Cloud	278
Term Document Matrix	280
Determining the Number of Clusters	281
Clustering Results	284
Visualizing with PCA	284
Cluster Similarity Matrix	289
Cluster Word Clouds	291
Conclusions	295
Hierarchical Clustering	296
GEOSPATIAL	298
Data	298
Geographical Representation	299
Heatmap of User Gender	300
Point Map Displaying User Locations	301
North America Distribution	301
North Europe Distribution	302
Analysis of Spatial-Temporal Data	303
User Growth in North America (2006-2015)	304
User Growth in North Europe (2006-2015)	310
Geographic Predominance of Gender by Zone	315
Gender distribution	316
Gender Predominance Heatmap	318
North America Gender Predominance	319
North Europe Gender Predominance	320
GANTT DIAGRAM	322
CONCLUSIONS	323
ANNEX	332

MOTIVATION OF THE PROJECT

In recent years, social media has become an essential part of our daily lives. Twitter is one of the most popular social media platforms, and millions of people use it to share their opinions, ideas, and experiences. However, predicting the gender of Twitter users based on their tweets, profile colors, and other features is a challenging task. The motivation behind this project is to explore people's biases when it comes to gender and how it affects the prediction of gender in social media. We will analyze a database of Twitter users and their predicted gender labels to identify any patterns or biases in the data.

To achieve this goal, we will first perform a thorough preprocessing step on the database. This step includes feature engineering, outlier detection, and missing value treatment to ensure the quality and integrity of the data.

Additionally, we will perform a descriptive analysis to get a better understanding of our database. This will be done first through a univariate approach and afterwards using a bivariate analysis, which will allow us to fully grasp each variable as well as the relationship between each other.

Next, we will use Multiple Correspondence Analysis (MCA) to explore the relationships between different features and the predicted gender labels of the users. MCA is a powerful statistical technique that allows us to visualize and analyze high-dimensional categorical data.

In addition to MCA, we will also perform clustering analysis using Cure, DBSCAN, and Optics algorithms. Clustering will help us identify any groups or clusters of users with similar features and gender labels.

Furthermore, we will also apply time series clustering techniques to explore any patterns or trends in the data over time.

Overall, this project aims to provide insights into people's biases when it comes to gender and how it affects the prediction of gender in social media. The results of this study may have implications for a wide range of fields, including marketing, social media analytics, and data science, and may help us better understand the limitations and biases of gender prediction processes in social media.

DESCRIPTION OF DATA

Data Source and Process

The source of the data can be found at [Kaggle](#), where information about the variables and the process from the owners of the data set are explained in detail. To get the data the only step that was necessary was downloading its CSV file. As this dataset has all the necessary variables that are required for this project (textual, geographical, temporal...) with enough values for each one, adding additional datasets wasn't needed. Therefore, the only phase required to get the data was a simple download.

Data Structure

The database for this project consists of 20,000 user profiles from Twitter. Each row in the database represents a single user profile and includes various variables such as one of their tweets, their description, the colors of their profile, their location, and other features. In addition, each row includes a predicted gender label, which is the gender that people believe the user is, as well as how many people have judged this user's gender.

Exclusion and Inclusion Criteria

Given the size of the database (20,000), which is too large for efficient analysis and studies, we decided to take a random sample of 7000 observations. This sample is a representative subset of the larger database and allows us to draw valid conclusions about the relationships between the variables of interest. By establishing clear inclusion and exclusion criteria and taking a random sample, we can ensure the validity and reliability of our analysis.

Metadata

In the next page, we will include a metadata of the initial variables.

After the preprocessing stage, we will include a final metadata which will include the variables we have extracted from the data, as well as the removal of variables that we have decided to exclude from the study. Therefore, we recommend that metadata be checked out as well, as it will be more useful to comprehend our project.

Variable	Modalities	Meaning	Type	Measuring Unit	Missing Code	Measuring procedure	Range	Role	Question
Unit_id		Id of the user	Numerical		-		[8157192 26, 8157579 85]	Explanatory	What is the id of the unit or observation?
Golden	TRUE / FALSE	If the observation is included in the golden standard	Boolean		-			Explanatory	Is this unit included in the golden standard?
	TRUE								
	FALSE								
Unit_state		If the unit's judgement is finished.	Categorical		-			Explanatory	What is the state of the unit currently? Has the judgment been finished?
	Finalized	The judgment is finished							
	Golden	In process of judgment							
	Not Golden	No judged							
Trusted_judgments		Number of trusted judgments of the user profile	Numerical	Judgements	-		[3, 274]	Explanatory	How many contributors have judged this unit?
Last_judgment_at		Date of last contributor judgment	Date	month/day/year hour: minute	NA		[2015-10-26 21:36, 2015-10-27 2:48]	Explanatory	What is the last time a contributor has judged this unit?
Gender		The predicted gender by the contributors	Categorical		NA			Response	What is the gender that the contributors have classified the user of this unit by?

	Male	The contributors believe the user is male							
	Female	The contributors believe the user is female							
	Brand	The contributors believe the user is a brand							
Gender:confidence		Contributor's confidence in the predicted gender	Numerical	Confidence	NA		[0, 1]	Response	How confident were the contributors in their gender judgment?
Profile_y_n	yes, no	If the contributors believe in the existence of the profile	Boolean		-			Explanatory	Do the contributors believe in the existence or availability of the profile?
	Yes	The contributors believe in the existence of the profile							
	No	The contributors don't believe in the existence of the profile							
Profile_y_n:confidence		Confidence in the existence/non-existence of the profile	Numerical	Confidence	-		[0, 1]	Explanatory	How confident are the contributors of their classification of the profile's availability?
Created		Date when the profile was created	Date	month/day/year hour: minute	-	[2006-8-5 16:04, 2015-10-26 13:19]		Explanatory	When was the profile of the unit created?

Description		The user's profile description	Textual		NA			Explanatory	What is the user's profile description?
Fav_number		Tweets the user has favorited	Numerical	Tweets	-			Explanatory	How many tweets has the user liked?
Gender_gold	Male, Female, or Brand	The true gender for gold observations	Categorical		NA			Response	What is the true gender of the user for gold observations?
	Male	The profile is accurately classified as Male							
	Female	The profile is accurately classified as Female							
	Brand	The profile is accurately classified as a Brand							
Link_color		The link color on the profile	Categorical	Hex Value	-			Explanatory	What is the link color of the profile?
Name		The user's name	Textual		-			Explanatory	What is the user's name?
Profile_y_n_gold	YES (for gold units) NA (others)	The existence/no n-existence of the profile for gold standard units	Categorical		NA			Explanatory	If the profile is gold, what is the true value for the existence or availability of the profile?
	YES	The user is in the golden standard and exists.							
	NA	Not included in the golden standard							

Profileimage		A link to the profile image	Textual		-	Web scraping Twitter		Explanatory	What is the profile image of the user?
Retweet_count		Number of times the user has retweeted	Numerical	Retweets	-	Web scraping Twitter	[0,330]	Explanatory	How many times has the user retweeted?
Sidebar_color		Color of the profile sidebar	Categorical(Color)	Hex value	-	Web scraping Twitter		Explanatory	What is the color of the sidebar of the user's profile?
Text		Text of a random one of the user's tweets	Textual		-	Web scraping Twitter		Explanatory	What is the text of the tweet of this user?
Tweet_coord		The coordinates of the user		(latitude, longitude)	NA	Web scraping Twitter		Explanatory	What are the coordinates of the location in which the user uploaded the tweet?
Tweet_count		Number of tweets that the user has posted	Numerical	Tweets	-	Web scraping Twitter	[1, 2680199]	Explanatory	How many tweets has the user posted?
Tweet_created		When the random tweet was created	Date	month/day/year hour: minute	-	Web scraping Twitter	[10/26/15 12:39, 10/26/15 13:20]	Explanatory	When was the tweet created?
Tweet_id		The tweet id of the random tweet	Numerical		-	Web scraping Twitter	[6,5873*10^17, 65874*10^17]	Explanatory	What is the id of the tweet?
Tweet_location	Cities around the world	Location of the tweet	Categorical	month/day/year hour: minute	NA	Web scraping Twitter		Explanatory	What is the user's location (city)?
User_timezone	Time zones	The timezone of the user	Categorical		NA	Web scraping Twitter		Explanatory	What is the user's time zone?

Descriptive Analysis of Raw Data

An analysis of raw data is important because it allows us to identify any issues or irregularities in the data that need to be addressed during the preprocessing stage. By analyzing the raw data, we can identify missing values, outliers, inconsistent data, and other data quality issues that need to be addressed to ensure that our results are accurate and reliable. Additionally, by studying the raw data, we may be able to identify new variables that could be important for our analysis. The more we can clean and prepare our data before we begin the analysis, the more accurate and useful our results will be.

Our raw data consists of user profiles from Twitter, which includes a variety of variables such as their tweet, description, profile colors, and location.

However, upon inspection, we discovered that some locations provided by users were not valid, and therefore we need to eliminate these invalid locations. Additionally, we observed that the text variables "tweet" and "description" have weird symbols that are not suitable for analysis and need to be removed. This tells us that we will have to go through a data cleaning process.

Furthermore, the profile colors are written in hex code which makes it difficult to study these variables. We will need to transform them into a more usable format. We also believe that we can extract interesting variables from the "tweet" and "description" variables such as their length or size of words. Therefore, we will also perform a feature extraction step.

Furthermore, we have identified missing values in the dataset that we will need to address. We also need to study the possible outliers in the dataset to determine their impact on our analysis. Consequently, it is necessary to perform a treatment of missing data, as well as an analysis of outliers.

Lastly, it is important to recognize which variables are useful when it comes to our study. For this reason, apart from centering on the variables we are interested in, we will perform a feature selection stage in order to confirm that these variables are useful for our analysis.

WORKING PLAN

DIVISION OF TASKS

	Tasks	Artur	Julia	Maria	Nico	Noel
D3 (1st Part)						
Preprocessing	Data Acquisition and Homogeneity	X				X
	Feature Engineering		X		X	X
	Feature Selection		X			
	Missing Data			X	X	
	Outliers				X	
Univariate descriptive analysis				X		
Bivariate descriptive analysis					X	
Additional descriptive statistics		X				
Data Set description according to results		X				
Clustering	Mixed Metrics				X	
	CURE Strategy				X	
	Construction of ontology based distances					X
	Clustering with mixed Data (Generalized Gibert Distance)		X			X
	DBSCAN			X		

	OPTICS			X		
	Comparison of Cluster Results		X			X
Time Series			X			
Automatic Interpretation of Classes	Profile Validation	X				
	CPG - Consumer Packaged Goods		X		X	
	TLP - Traffic Light Protocol			X		X
	Comparison and Conclusions	X				
MCA - Multiple Correspondence Analysis		X				X
FMA - Multiple Factor Analysis				X		
D3 (2nd part)						
Space-Time Modeling			X			X
Text Mining	Sentiment Analysis Models		X	X		
	LSA - Latent Semantic Analysis			X		
	Topic Modeling				X	X
Latent Variable Modeling	Clustering with Significant Values	X			X	
	ANCOVA					X
D4						
Motivation of the work			X			
Data Source presentation						X

Formal description of data structure and metadata	X				
Detailed description of Preprocessing and data preparation			X		
Basic statistical descriptive analysis				X	
Working Plan	X				
Folder to be delivered					
Presentation slides		X	X		
R Scripts	X				
Bibliography	X				
README.txt					X
Subfolder with intermediate data files				X	

PREPROCESSING

Preprocessing is a crucial step in any statistical project as it involves cleaning and transforming raw data to make it suitable for analysis. In this segment, we will explain the various techniques involved in preprocessing that we have applied to our database, which include data cleaning, data imputation, feature engineering, missing imputation as well as outlier analysis. All these techniques are important in ensuring that the data is reliable, accurate, and suitable for statistical analysis.

Data Acquisition and Data Homologation

In order to carry out a successful statistical analysis, it is crucial to ensure that the data is properly homogenized and acquired. Data homogenization involves the process of standardizing and cleaning the data to ensure that it is consistent and free from errors. We achieved this by removing any duplicates or inconsistencies that may exist in the data. It is important to ensure that the data is collected in a systematic and organized manner to avoid bias or errors that may affect the analysis. By properly homogenizing and acquiring the data, it will be possible to conduct an accurate statistical analysis that will provide valuable insights into the gender distribution of Twitter users.

Data Cleaning

In the first part of the preprocessing stage of our project, we focused on cleaning the text variables to ensure that the data was suitable for analysis. This involved removing any weird symbols or characters that could not be analyzed, as these could potentially impact the accuracy of our results.

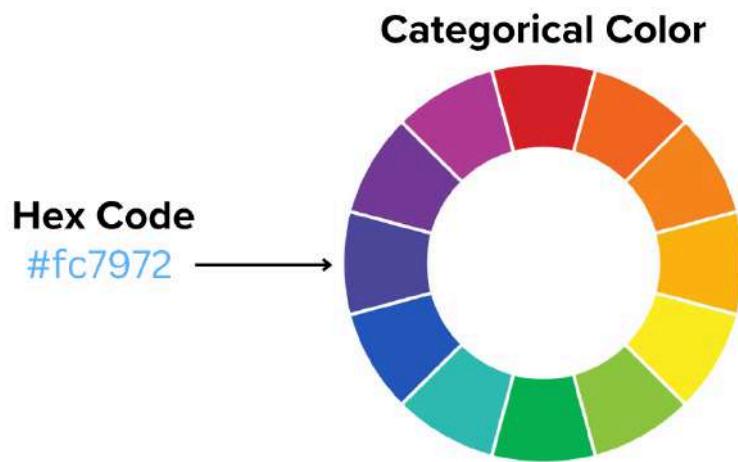
We also looked at the values in the variables to check for any inconsistencies or errors. For example, in the *sidebar color* and *link color* variables, which had color in hex, we found incorrect color codes that were not valid, and we replaced them with missing values (NA) to indicate that the data was missing or unreliable. Similarly, in the "location" variable, we found user-imputed locations that were nonsensical, and we replaced these with missing values as well.

By cleaning the variables in this way, we were able to ensure that the data we used for our analysis was reliable and accurate, which is essential in any statistical project.

Feature Extraction

Feature extraction is a critical step in statistical analysis as it involves identifying and extracting meaningful information from raw data that can be used for further analysis. Feature extraction involves transforming data from its original form into a set of features that can be used to build models or make predictions. It can include techniques such as dimensionality reduction, data transformation, and categorical encoding.

In our project, we focused on extracting different features from our dataset in order to perform further analysis. One feature we extracted was the basic color of each observation's link and sidebar (variables `sidebar_color` and `link_color`), which we transformed into a categorical variable. We did this by calculating the distance of each observation's color to the basic colors and then labeling its color to the basic color with the smallest distance. This allowed us to analyze the data in a more meaningful way, as we could group observations by basic color and investigate any patterns or trends within those groups, for example, if we noticed any trend with the predicted gender and the color that observation used.



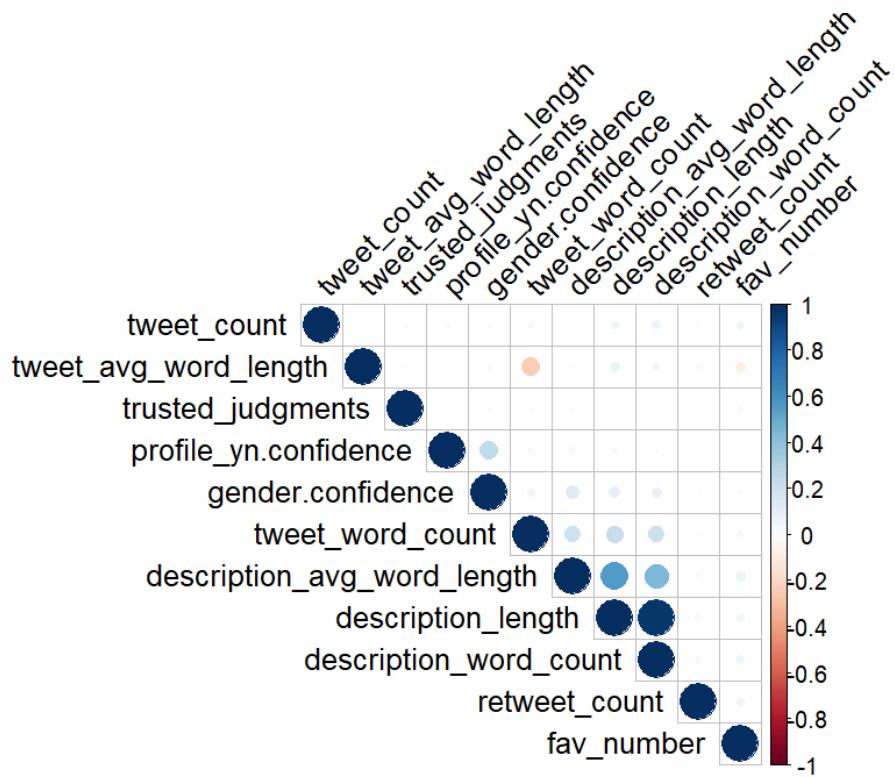
Another feature we extracted was information about the tweet and description of each observation, such as the average word length or word count. This can be useful for identifying patterns or trends in language use or for predicting certain outcomes based on the language used in the tweet or description. We created these variables to observe if people connected, for example, longer tweets with women? or maybe lengthier words with businesses?

We also extracted location variables, which we converted into a categorical variable named continent and a geographical one named location with longitude and latitude. This allowed us to analyze the data spatially and identify any geographical patterns or trends. We also included a variable named privacy for those who had chosen not to show their location, which we will explain later in this project the reason for this addition in more detail.

Overall, feature extraction is an important step in statistical analysis as it allows for the identification of meaningful patterns and trends in data that can be used for further analysis or prediction. By extracting relevant features from our dataset, we were able to gain a deeper understanding of our data and make more informed conclusions based on our findings.

Feature Selection

Correlation matrix

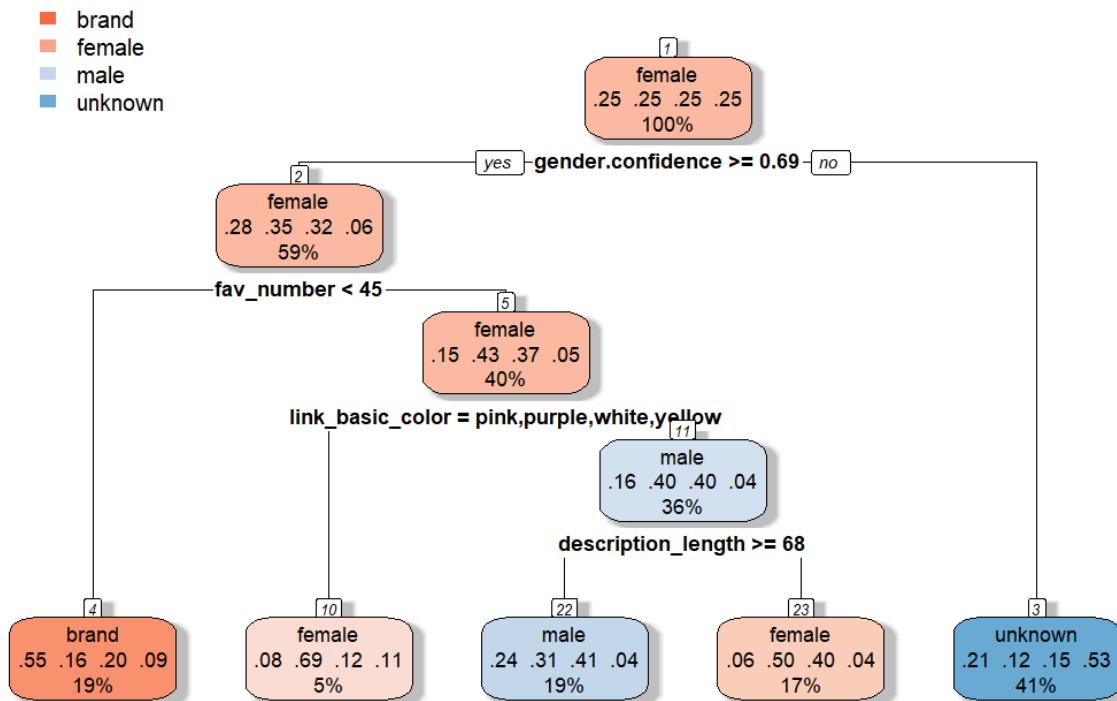


To begin our feature selection analysis, we first examined the correlation matrix using Pearson correlation as the metric. Our goal was to identify which variables were highly correlated with each other, so that we could remove any unnecessary variables and simplify our model.

The variables extracted from the description text (length, word count, and average word length) were found to be moderately to strongly positively correlated with each other. The variable *tweet_word_count* had a weak positive correlation with them and a negative correlation with *tweet_avg_word_length*. The remaining variables showed little to no correlation with each other.

Based on our findings, we decided to remove the *description_length* variable from our analysis, as it was highly correlated with other description-related variables, and retaining it could lead to multicollinearity issues.

Decision Tree

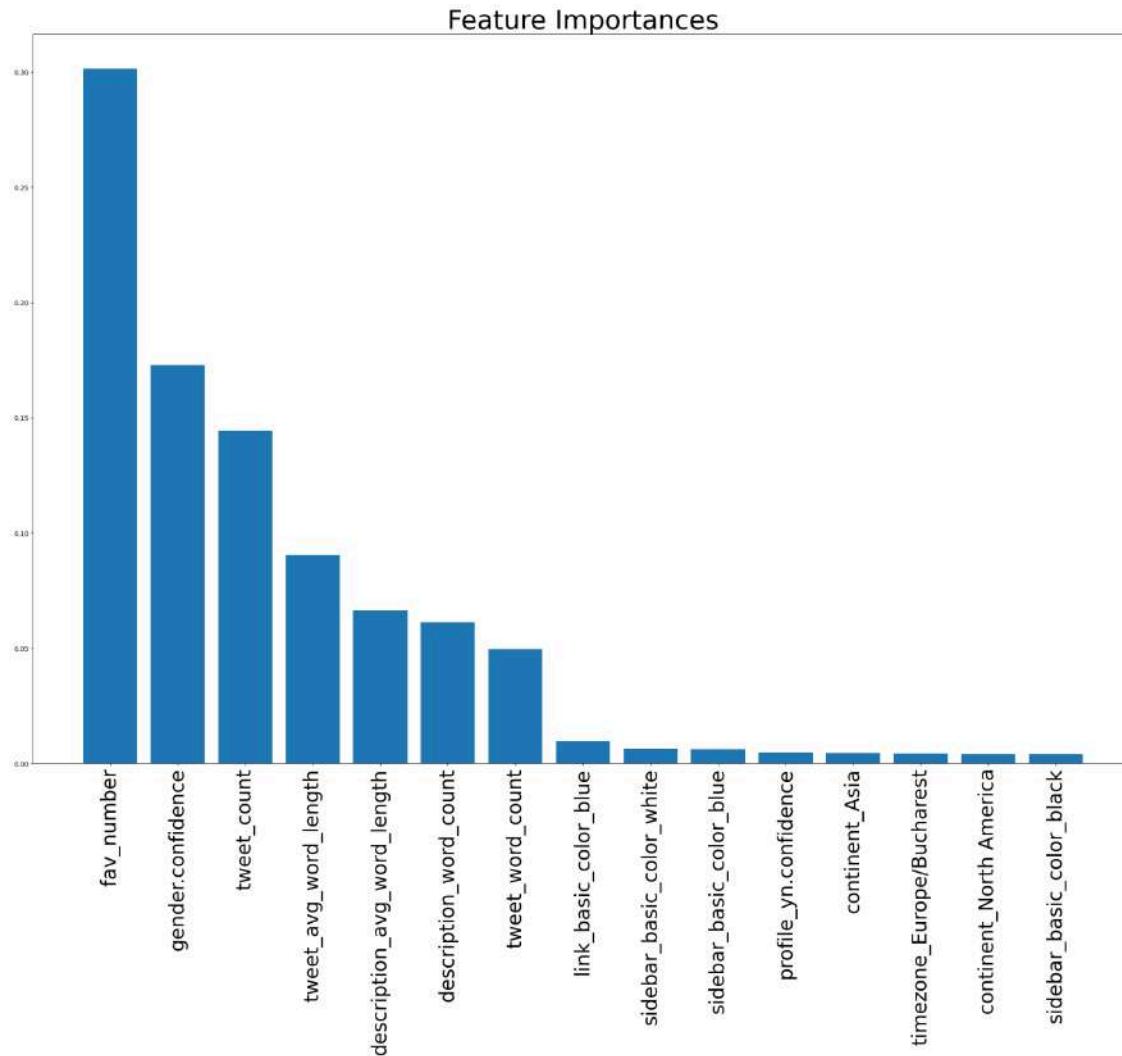


A decision tree is a popular machine learning algorithm that recursively splits a dataset based on the most informative features, aiming to maximize the information gain at each step. This makes decision trees a useful tool for identifying the most important variables in a dataset, as they can quickly reveal which features are most predictive of the outcome of interest.

To train our decision tree, we balanced the gender variable by using upsampling. Specifically, we added random instances of the less represented categories. Our data didn't require additional preprocessing steps, and we trained the decision tree with a depth of four. Our findings revealed that the variable 'gender.confidence' with a value greater than or equal to 0.69 was the most important feature for distinguishing the 'unknown' gender category from the other categories.

When it comes to the 'brand' category, the most relevant variable was 'fav_number' with a value less than 45. Furthermore, the colors pink, purple, white, and yellow in the 'link_basic_color' variable were found to be more indicative of a female category than a male one. Similarly, a 'description_length' with a value less than 68 was more predictive of a female category than a male one. It must be taken into account that the precision of our decision tree has been 50%, therefore, the conclusions drawn from it are not conclusive.

Random Forest



We have decided to use a random forest for our analysis because it is an ensemble of multiple decision trees, which makes it more robust against overfitting compared to a single decision tree. However, the trade-off for this increased robustness is a reduction in interpretability. Nevertheless, a random forest still allows us to assess the importance of each variable in the model.

As a preprocessing step, we applied one-hot encoding to the categorical variables and balanced the target variable, gender. The resulting feature importance plot, shown above, reveals that *fav_number* is the most important variable, followed by *gender.confidence*, *tweet_count*,

tweet_avg_word_length, *description_avg_word_length*, *description_word_count*, and *tweet_word_count*. The other variables have less than 3% importance, but notable ones include *link_basic_color_blue*, *sidebar_basic_color_white*, and *sidebar_basic_color_blue*.

It is important to note that the accuracy of this model is only 30% on the testing dataset, which is only slightly better than random guessing (which would yield 25%). Therefore, we must be cautious in interpreting the results. While these variables seem to be the most important, the overall performance of the model is not very good.

PCA

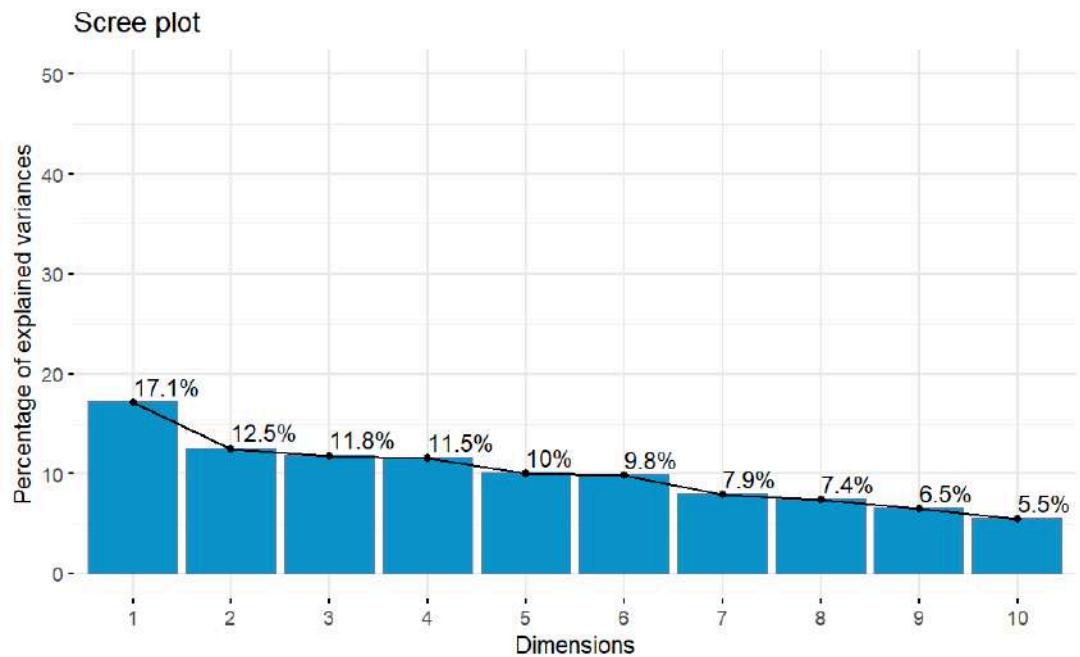
Principal component analysis (PCA) is a popular unsupervised technique used to reduce the dimensionality of data by representing it in a lower-dimensional space while retaining as much of the original data variation as possible. This can improve interpretability and enable visualization of data with many variables.

The PCA finds a set of orthogonal components that are ordered by the amount of variance they explain in the data. To perform PCA, the data is centered by subtracting the mean of each attribute, and the eigenvectors and eigenvalues of the covariance matrix are computed. The components are a linear combination of the original variables and are used to represent the data in a new space. It is important to address outliers and missing data, which can negatively impact the accuracy of the PCA results.

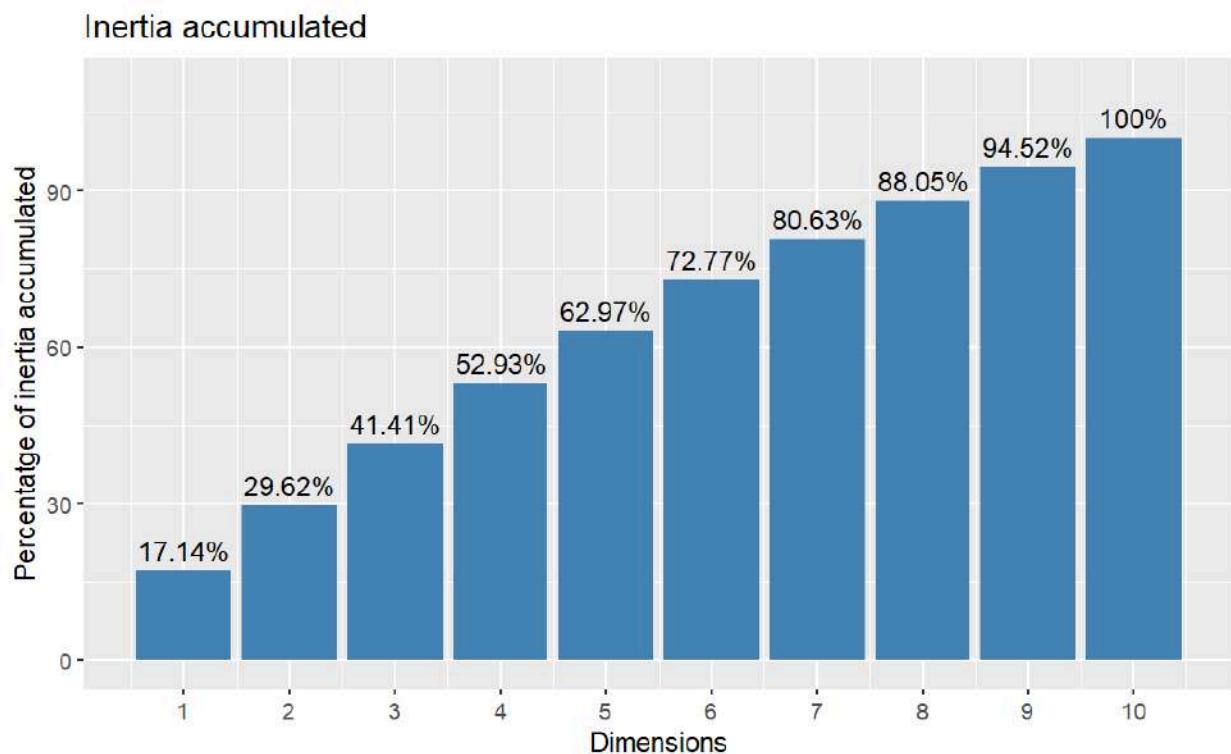
In this section, we perform PCA on our dataset using the FactoMineR library to compute the components, factoextra to visualize the results, and corrplot to create correlation matrices.

Selecting the optimal number of principal components

To begin, we are going to calculate the PCs of our dataset. In this next graph we can observe the percentage of variances that are explained by each principal component:



Next up, we have included a graph that shows the accumulated inertia of each dimension:



Finally, we collected the eigenvalue, the variance and cumulative variance in a table:

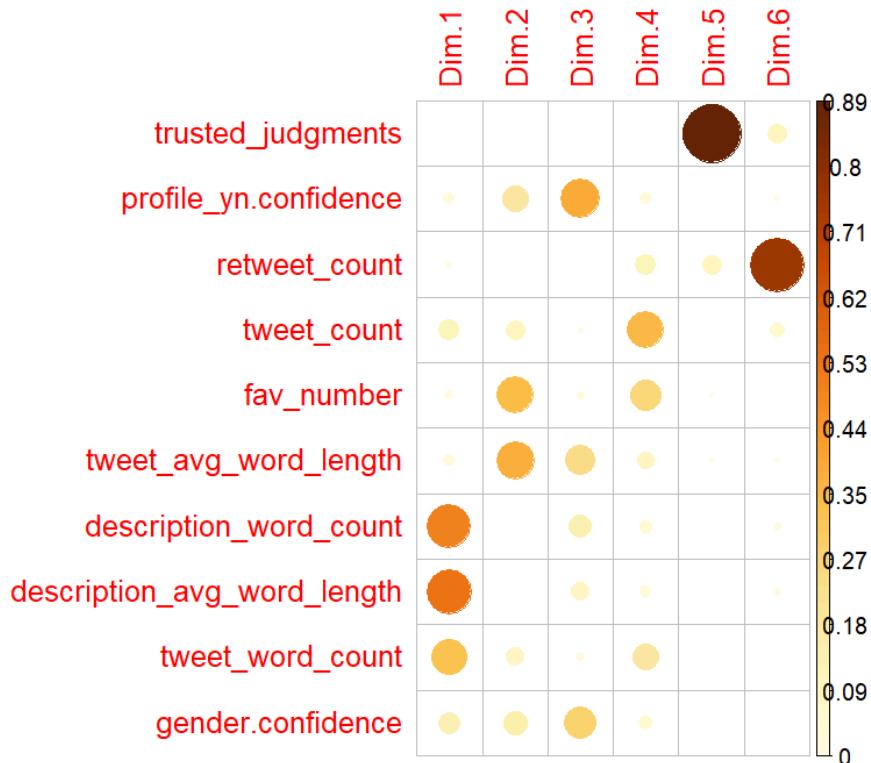
	Eigenvalue	Percentage of variance	Cumulative percentage of variance
comp 1	1.714	17.138	17.138
comp 2	1.248	12.483	29.620
comp 3	1.179	11.789	41.409
comp 4	1.152	11.522	52.932
comp 5	1.004	10.037	62.969
comp 6	0.980	9.801	72.770
comp 7	0.786	7.864	80.634
comp 8	0.742	7.415	88.049
comp 9	0.647	6.471	94.519
comp 10	0.548	5.481	100.000

Our method for determining the number of principal components to retain is based on the eigenvalues of the components and the cumulative variance explained by them. Specifically, we select all principal components with an eigenvalue greater than 1 and aim for a cumulative variance close to 70%.

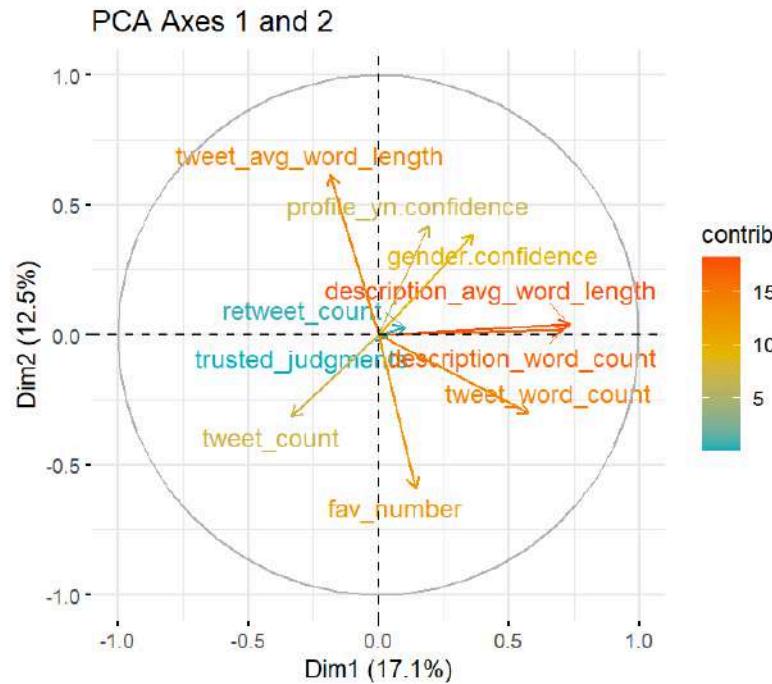
In our particular case, this approach has led us to select 6 principal components. Although the eigenvalue of the sixth component is slightly below 1, it is still close enough that we have decided to retain it.

Analysis of our PCA

The following graph shows how well represented each variable is in each PC, i.e., which variables are mostly represented by each PC.

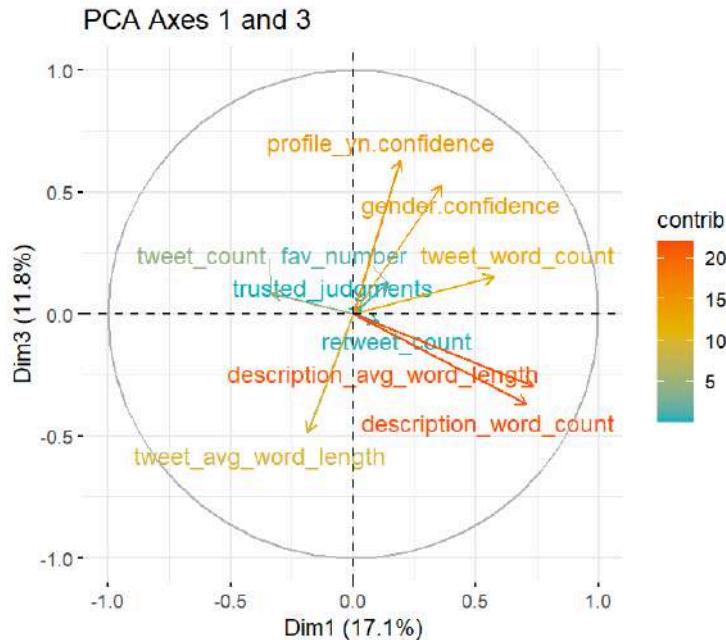


The variables that have the highest contribution to the first principal component are *description_word_count* and *description_avg_word_length*, while *tweet_word_count* have a moderate contribution. *Gender.confidence* has a small contribution to this dimension. In the second principal component, *Fav_number* and *tweet_avg_word_length* have the highest contribution. *Profile_yn.confidence* has a high contribution, while *gender.confidence* and *tweet_avg_word_length* have a moderate contribution. In the fourth principal component, *tweet_count* and *fav_num* have a moderate contribution, while *tweet_word_count* has a small contribution. *Trusted_judgements* has the highest contribution to the fifth principal component. Finally, *retweet_count* has the highest contribution to the sixth principal component.



Taking a look into this first graphic where we are representing the first and second PC, we can see *the description_avg_word_length* and *description_word_count* made a high positive contribution to the first dimension. Both variables showed a similar angle, close to the X axis, indicating that they have a strong correlation with each other and are important in explaining the overall variability of the data in the first component.

On the other hand, *Tweet_avg_word_length* had a high positive contribution to the second dimension, but *fav_number* had a negative contribution to the second dimension.

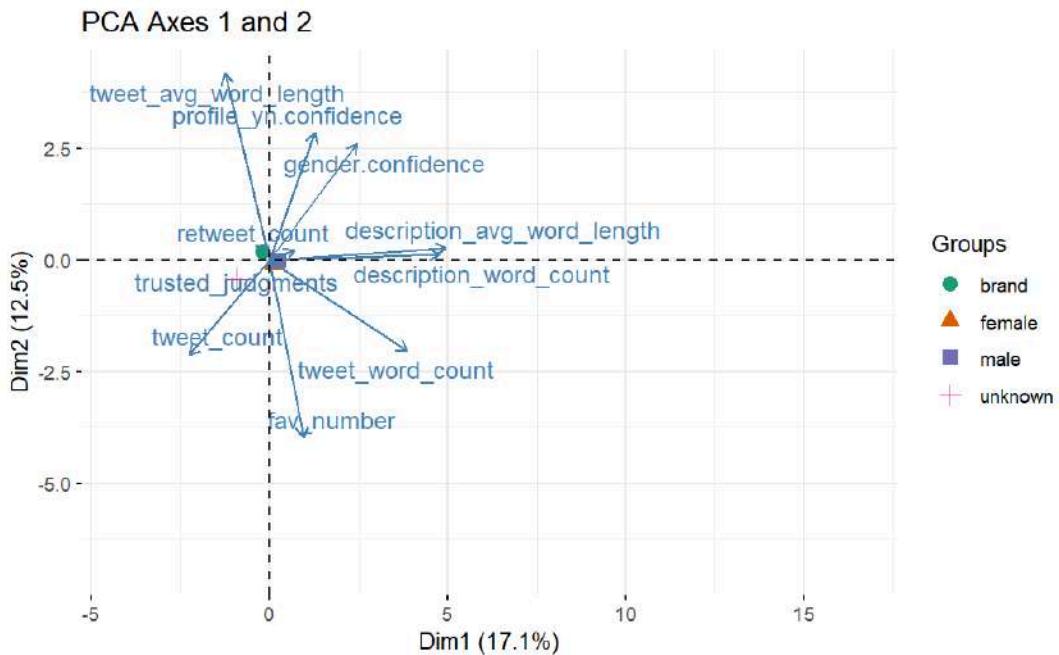


In the graph, we can see that the variables *description_avg_word_length* and *description_word_count* form an angle of minus 30 degrees with respect to the positive X-axis, indicating that they have a positive contribution to the first dimension and a negative contribution to the third position. On the other hand, *profile_yn.confidence* has a positive contribution to the third dimension, while *tweet_avg_word_length* has a negative contribution.

Furthermore, *gender.confidence* and *tweet_word_count* have a positive contribution in both dimensions, but their contribution is relatively smaller compared to the other variables.

Projection of categorical variables

We attempted to project our categorical variables onto the dimensions of the PCA as centroids. However, we were unable to derive any meaningful insights. The following graph depicts the gender variable projection onto the plane of dimensions 1 and 2. All the centroids, except the missing category, are clustered in the middle.



Conclusion

During the PCA analysis, we observed that `description_avg_word_length` and `description_word_count` provide similar information, as they had similar magnitude and angle in different dimensions. This suggests that we should take into account their joint importance in subsequent analyses.

Missing Data Treatment

The treatment of missing values is a critical step in statistical data analysis. It is essential to carefully handle missing values to ensure that the resulting analysis and conclusions are accurate and reliable. Missing values can arise due to various reasons such as data entry errors, participant non-response, or study design. In our database, we find two different types: entry errors (i.e. color hex-codes imputed incorrectly) as well as participant non-response (i.e. users not disclosing information about their location, users with no description in their account...).

Deterministic Imputation

We have encountered some missing values in several variables, including location, city, and timezone. Some of these missing values could potentially be imputed based on the availability of information in other variables. Specifically, if a user's city is known, it is possible to deduce their location (latitude and longitude) and timezone.

Therefore, we propose to use a conditional imputation approach for these missing values. That is, we will first impute the missing values for the city variable using an appropriate imputation method. Then, we will use this imputed city value to deduce the missing values for the location and timezone variables. This will be done using a deterministic approach, as the relationship between the city, location, and timezone variables is known and can be directly inferred.

It is important to note that any remaining missing values that cannot be deduced through this approach will require a separate imputation strategy, which will be outlined in the next section of this report.

Types of Missing Values

There are different types of missing values, and the appropriate treatment depends on the nature of the missing data. The three main types of missing values are Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR).

In MCAR, the missing data are unrelated to any other observed or unobserved variable in the data set. In this case, the missing data can be ignored, and any imputation method can be used without biasing the results.

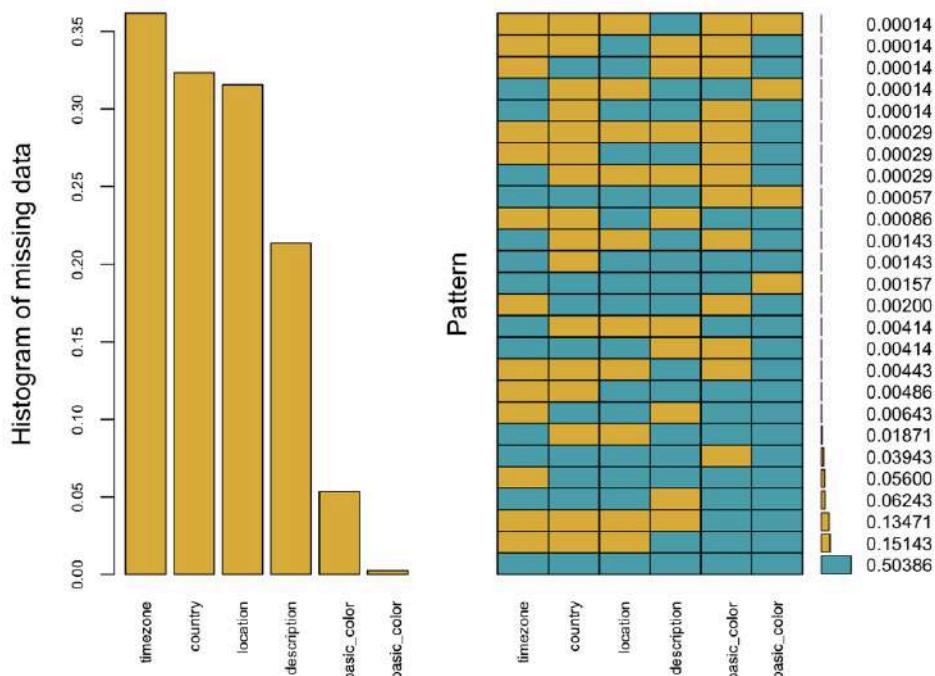
In MAR, the missing data are related to other observed variables in the data set, but not to the missing values themselves. In this case, the missing data can be imputed using a variety of methods, such as regression imputation or multiple imputation, as long as the imputation model accounts for the relationship between the missing values and the observed variables.

In MNAR, the missing data are related to the missing values themselves, and cannot be explained by the observed data. In this case, imputation methods may introduce bias, and the best approach is to replace NA's by a new category that explains their missing status.

Looking for patterns

A thorough understanding of the pattern of missing data is essential for drawing valid conclusions from statistical analyses and for ensuring the reliability and reproducibility of research findings. That's why we have plotted this histogram of missing data, as well as a plot that displays missing data as yellow and non-missing data as blue, in order to understand the patterns in our database more comprehensively:

Missing Data Pattern



Before analyzing the plot, it is important to note that we haven't plotted every variable, specially not the derived variables, only those important to the study of missings.

First off, if we look at the histogram, we can observe how we have 36% of missing data in the variable *timezone*, followed by 32% in country and 31% in location. Afterwards, we have 21% of missing values in *description*. Lastly, there is 5% of missing values in *link_basic_color* and 0.2% in *sidebar_basic_color*

From this plot, we can gather that half of our dataset has no missing data. Additionally, 15% of our database only has missings in the variables related to location. Another big portion of our data, 13% to be precise, has missings in the variables related to location as well as description. Afterwards, the patterns get less important as they only describe less than 1-5% of our data. Some of these are observations with no timezone, no description, missing on sidebar color or missing on both country and location. Eventually the patterns describe about less than 1% of data which does not tell us very important or useful information.

Little's Test

To check the type of missing values we have, we performed a Little's Test. Little's test is a statistical test used to determine whether the missing values in a data set are Missing Completely at Random (MCAR) or not. To perform Little's test, we examined the correlation matrix of the variables in our database, including missing values, and calculated the test statistic. The test statistic follows a chi-squared distribution, and the p-value indicates the level of significance of the test. These were the results obtained from the test:

```
> mcar_test(db)
p-value: 0
```

We obtained a p-value of 0, indicating strong evidence against the null hypothesis that the missing values are MCAR. This result suggests that the missing values in our data set are not missing randomly, and we need to use an appropriate approach taking this into consideration.

First off, we have missing values in a variable that pertains to the location of our users. The reason some observations have missing values is that these users have chosen not to display this information. It is important to recognize that, as the test proved, these missing values are not Missing Completely at Random (MCAR), but are instead Missing Not at Random (MNAR). This is because the missingness is related to the value of the variable itself and is not randomly distributed across the observations.

When imputing missing values in a MNAR scenario, it is crucial to recognize that imputing these values using a model can introduce bias into the analysis. This is because imputation models typically assume that the missing data are MCAR or MAR, and this assumption may not hold in the MNAR scenario. Specifically, by imputing these values, we would be presupposing that these observations behave the same way as the other observations, which is not the case since they have fundamentally different characteristics.

Therefore, it is essential to recognize the nature of the missingness in the data and carefully consider the potential biases introduced by imputing the missing values. For this reason, we will include a new category in our location variables named *Missing*, in order to avoid introducing new biases to our database. We also thought it would be interesting to study the relationship between users who chose not to display location information on Twitter, by creating a new variable named *Privacy*, set to True if one of the location variables was missing.

A similar thing happens when it comes to the variable *description*, which has NAs for users with no description. We will use the same approach as with the location variables, adding a *No description* category to this variable.

Additionally, we also have another variable with missing values, which pertains to the color of the profile of the Twitter users (sidebar color and link color). The reason for missingness in this variable is due to an input error, where the hex color codes were entered as an incorrect value. It is important to recognize that in this case, the missing values are likely to be Missing Completely at Random (MCAR) since there is no underlying relationship between the missing data and the observed or unobserved variables in the data set. Therefore, we will be using imputation methods to replace these values.

Choosing the best imputation method

Since we have a relatively small amount of missing data to complete in our dataset, it is important to take advantage of this by selecting an appropriate imputation method that can accurately fill in the missing values. In order to choose the best imputation method among MIMMI, MICE, and Random Forest, we created false missing values in three variables (*fav_number*, *gender* and *gender_confidence*) and used these three methods to impute these values.

By comparing the performance of the three imputation methods on the false missing values, we can determine which method provides the most accurate imputations for our data. The method that performs the best on the false missing values will be the most appropriate method to use for imputing the actual missing values in the dataset. This approach helps to ensure that the

imputation method is tailored to the specific characteristics of our data and can provide the most accurate imputations possible. By selecting the most appropriate imputation method, we can ensure the validity and reliability of our statistical analysis and produce meaningful insights from our data.

Hereunder we have represented our experiment in a table in order to visualize it:

	MIMMI	MICE	Random Forest
Parameters	- dist = "gower" - clusters = 30	- m = 10 - max iter = 50 - method = pmm	-
Performance Measures	<ul style="list-style-type: none"> - Comparison of distribution plot of predicted values and real values - Root mean squared error (RMSE) measure for numerical variables - Accuracy for categorical variables 		

	Numerical var.		Categorical var.		
Variables with false missings	Fav_number (num.) 250 imputed missings	Gender.confidence (num.) 250 imputed missings	Gender (cat.) 250 imputed missings	Sidebar_basic_color (cat.) 250 imputed missings	Link_basic_color (cat.) 250 imputed missings

MICE - Multiple Imputation by Chained Equations

The Multivariate Imputation by Chained Equations (MICE) method is a popular imputation technique used to handle missing data. MICE imputes missing data by creating multiple imputed datasets and uses regression models to estimate the missing values. The MICE algorithm imputes missing values one variable at a time, while taking into account the observed values in the other variables in the dataset.

One advantage of the MICE method is that it can handle missing data in both continuous and categorical variables. It is also able to capture the uncertainty in the imputations by generating multiple imputed datasets, which allows for more accurate statistical analyses. Additionally, MICE can handle missing data with missingness at random (MAR) assumptions, which is a common type of missing data in practice. One disadvantage of the MICE method is that it can be

computationally intensive, particularly when dealing with a large number of variables or when the dataset is very large.

We chose to set the hyperparameter m to 10, which means that for each missing value, we imputed 10 different values based on the observed data. This helped to increase the accuracy and reduce the uncertainty of our imputations. We also selected the "pmm" method, which stands for predictive mean matching, to impute the missing data. This method involves matching each missing value with a predicted value based on the observed data, and then randomly selecting one of the matched values as the imputed value. It is a well-established imputation method that has been used in many research studies and statistical software packages, such as R and SAS. Finally, we set the maximum number of iterations to 40, which means that the MICE algorithm will repeat the imputation process 40 times to obtain a stable and reliable imputation result. These choices were based on previous research and best practices in imputation methods, and were deemed appropriate for our specific dataset.

Hereunder is the code used to execute this method, as well as the results obtained when compared to the actual values:

```
> mice(df_missings, m=10, maxit=40, meth=c("pmm"), seed=200)
```

Here are the performance results:

Performance Measures	Numerical Variables		Categorical Variable		
	Fav_likes	Gender Confidence	Gender	Sidebar_basic_color	Link_basic_color
RMSE (num.)	5483.5	0.1586092	-		
Accuracy (cat)			33.6 %	40.4 %	50.8 %

MIMMI - Mixed Intelligent-Multivariate Missing Imputation

The MIMMI method stands for Mixed Intelligent-Multivariate Missing Imputation. This method takes the assumption that observations close to each other are more similar, and builds on that by creating clusters (of similar observations) and imputing the local mean of the cluster for the missing observations.

To do that, we have to select a small number of relevant variables with no missing values. In our case, we chose these variables:

- *created*, turned numerical for this imputation, since it could affect the amount of tweets the person has liked, for example.
- *retweet_count*
- *tweet_count*
- *tweet_length*, since the larger the tweet is, the more information it is more likely to give and therefore the more confidence people will have in assessing their gender.
- *tweet_word_count*, for the same reason as above.

Afterwards, clusters will be made from this method with hierarchical clustering using the Gower metric, which works well for both categorical and numerical variables. The number of clusters tends to be a high number (20-30). For our case, we chose 30 clusters because it seemed the more fitting option.

An advantage to this method, as opposed to more simple ones, is the increase in accuracy resulted. Even so, it is more time consuming and might require expert knowledge to choose the relevant variables.

Using the program given to us in this class, we executed missing in our data:

```
> MiMMi(df_missings)
```

Hereunder we have inserted the variables resulted from this imputation:

Performance Measures	Numerical Variables		Categorical Variable		
	Fav_likes	Gender Confidence	Gender	Sidebar_basic_color	Link_basic_color
RMSE (num.)	4470.06	0.116718	-		
Accuracy (cat.)	-		39.6 %	63.6 %	66.4 %

As can be seen, performance measures for MiMMi seem to be better than for MICE. Even so, this does not mean it is the best imputation method for this database, since it is also very important to compare the distributions of each imputation to the real distribution.

Random Forest

Random Forest imputation is a machine learning-based approach for imputing missing values in datasets. The method is based on the Random Forest algorithm, which is a decision tree-based algorithm used for classification and regression tasks.

The RF imputation method has several advantages. Firstly, it can handle both numeric and categorical variables, which makes it a versatile imputation method. Secondly, it can capture non-linear relationships between variables, which may be missed by other imputation methods, such as mean imputation or regression imputation.

However, like any imputation method, the RF imputation method has its limitations. It assumes that the missing data mechanism is MAR, and the imputed values are dependent on the observed data. Therefore, it is important to assess the validity of these assumptions before using the RF imputation method to impute missing values. Also, the RF imputation method may not perform well in datasets with a small sample size or a high proportion of missing data. In such cases, alternative imputation methods may be more appropriate.

Since in our case we are not dealing with a small sample size or a high proportion of missing data, we believe this method aligns well with our data. Even so, it has to be said that the way we are imputing these “false” missings for this experiment is at random, which might not work well for this method since it assumes missing data is MAR.

We used an R function from the package missForest, which already applies the Random Forest method to imputation of missings:

```
> missForest(df_missings)
```

Hereunder we have inserted the results of this last method:

Performance Measures	Numerical Variables		Categorical Variable		
	Fav_likes	Gender Confidence	Gender	Sidebar_basic_color	Link_basic_color
RMSE (num.)	4597.721	0.1422671	-		

Accuracy (cat.)	-	51.6 %	62.4 %	54 %
---------------------------	---	--------	--------	------

Again, this method seems to exceed the previous ones, but we still have to analyze the distributions.

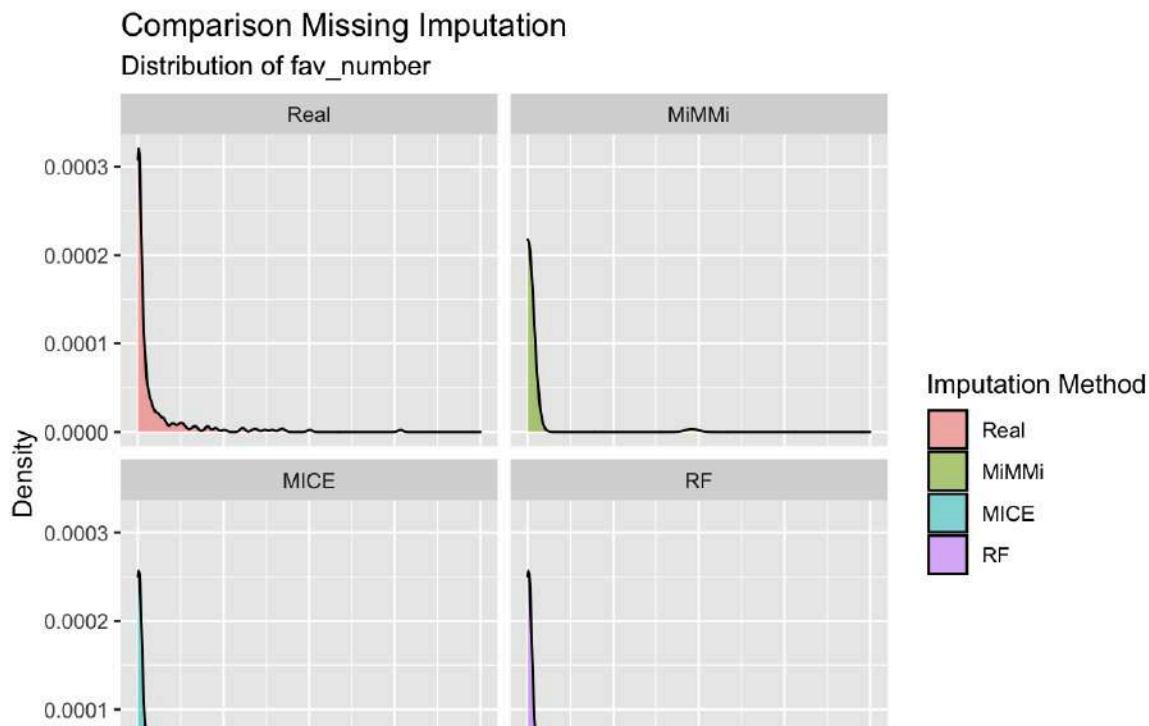
Comparison of MICE, MiMMi and RF

After comparing the performance of different imputation methods using metrics such as MSE or accuracy, the next step would be to compare the distributions of the imputed values from each method with the real values. This step is important because imputing missing values can introduce bias and distort the distribution of the data.

If the distribution of the imputed values differs significantly from the real distribution, this could lead to incorrect statistical inferences and negatively impact the validity of the results. For example, if the imputed values are biased towards higher or lower values, this could affect the mean and standard deviation of the variable, which could impact the results of subsequent analyses.

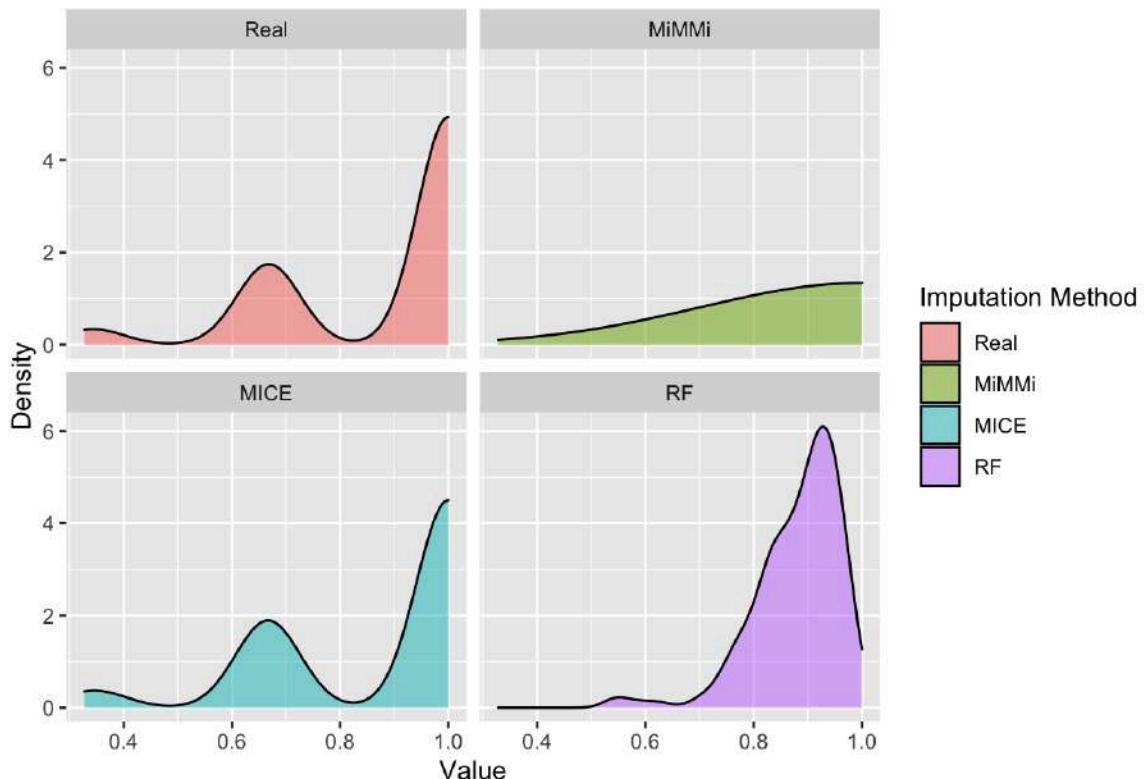
Therefore, it is important to select an imputation method that produces imputed values with distributions that are as similar as possible to the real distribution of the variable. By comparing the distributions of the imputed values from different imputation methods with the real values, we can identify the method that best preserves the distribution of the data and therefore produces the most accurate imputations.

Hereunder is the comparison of distributions of our numerical variables: *fav_likes* and *gender_confidence*:



Comparison Missing Imputation

Distribution of gender confidence



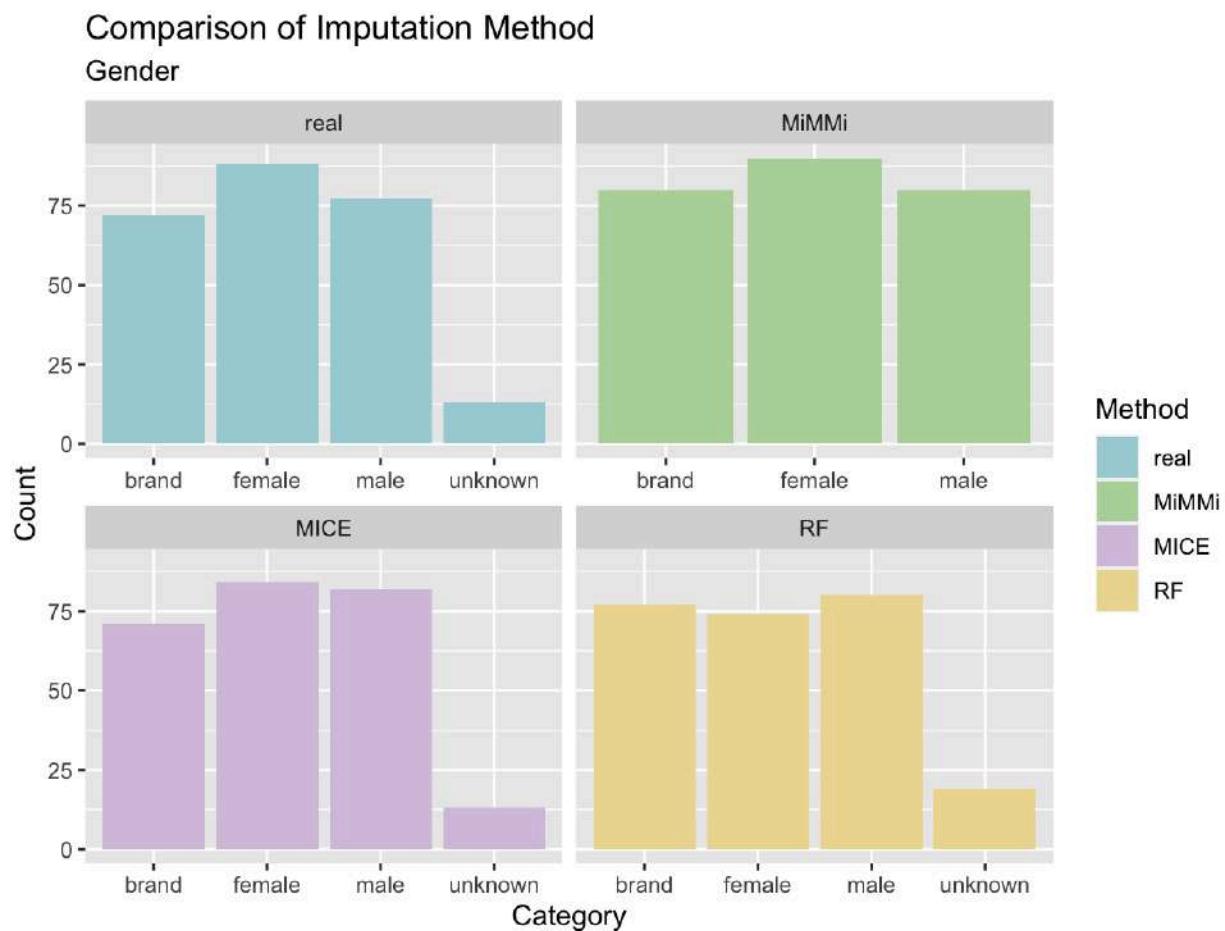
As for the *fav_likes* variable, we find that most distributions are similar to the real values. Specially MICE and RF, which seem to capture the pattern a bit better than MiMMi.

With respect to the variable *gender confidence*, we find a big difference for each method. This is because the distribution of real values makes up a really specific pattern, with three spikes, one at 0.35, 0.65 and 1 approximately. As for MiMMi, it seems that it has not really found the pattern as it does not take into account the increases, and it simply has increased in density as the confidence goes up. Looking at MICE, we can observe that it's the method that has best adapted to this pattern, looking really similar to the real distribution. Lastly, when it comes to Random Forest, it has done a better job than MiMMi but has not really accomplished the distribution

needed. It does recognize the two last peaks, although it gives much more frequency to the second one opposed to the first.

Therefore, although the performance measures of the numerical variables related to the MiMMi and RF method are better than for MICE, we can observe that MICE has adapted better to our data and, at least for the numerical variables, seems like the safer choice in order to avoid changing the distribution of our variables.

Let's look at the boxplot of *Gender* now:



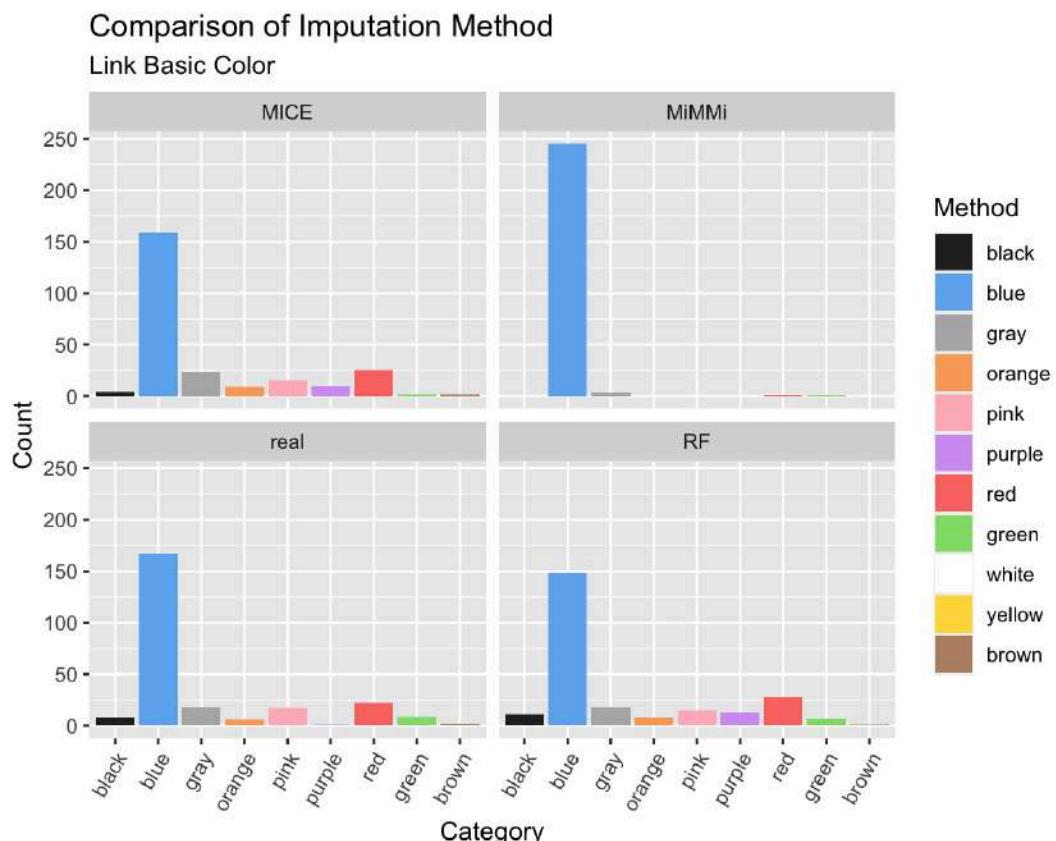
Again, it seems that MICE has adapted better to the real proportions of our variables, having female be the most frequent category, followed by male, brand and then unknown.

As for RF, which is the second best method for this variable, we find that although it has kept the proportion of unknown lows, the female category is the third lowest proportion, which makes it differ from the real proportions.

Lastly, MiMMi keeps the proportion of the most frequent categories but completely dismisses the category *unknown*.

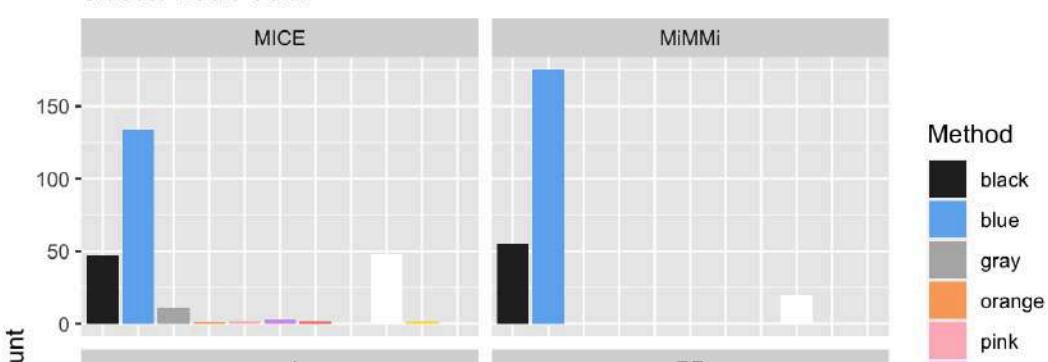
This again, proves that it is important to look ahead at more than the accuracy measure, which had a better performance score for MiMMi than for MICE.

Lastly, we will look at the two color variables:



Comparison of Imputation Method

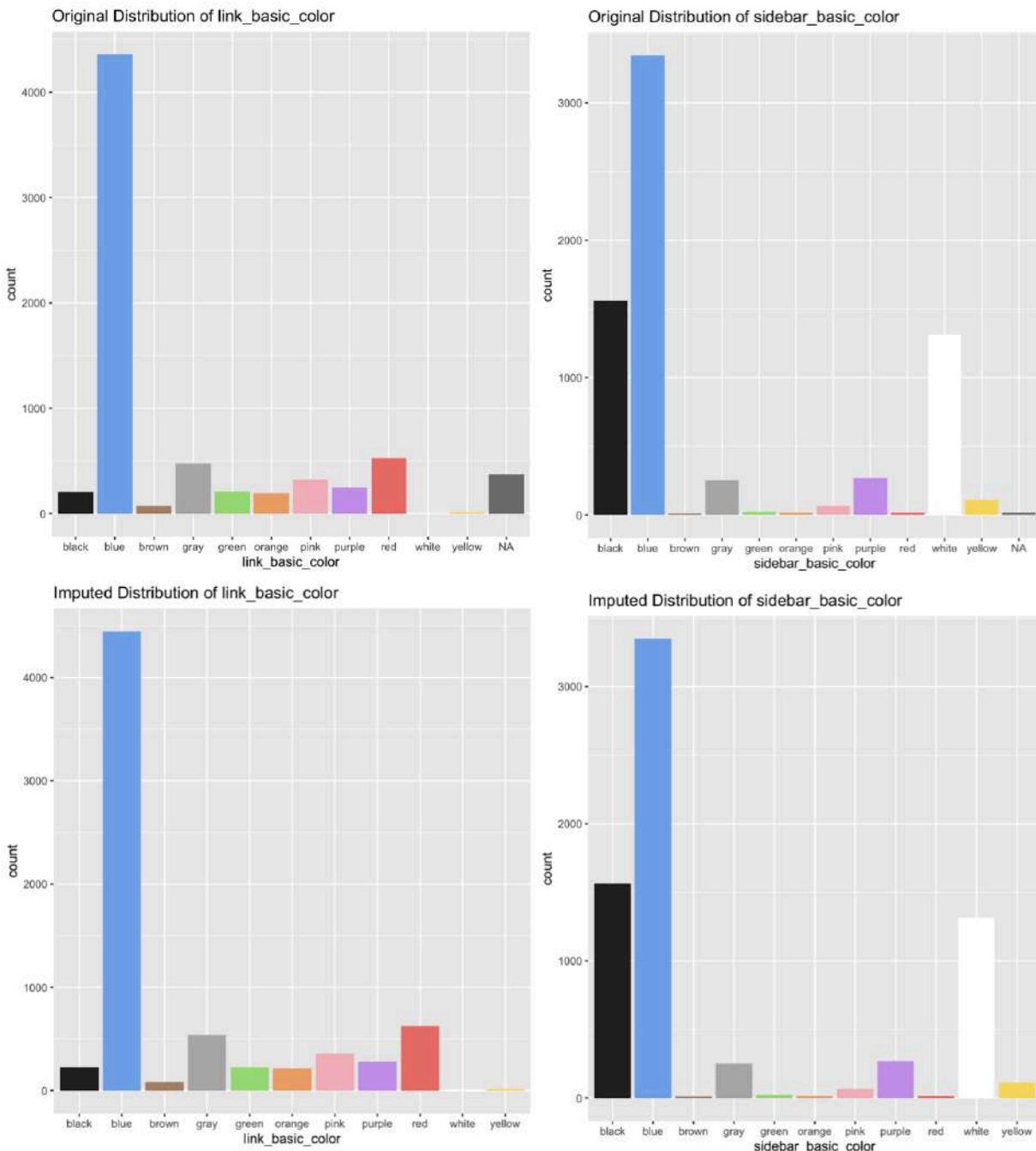
Sidebar Basic Color



Results

After performing missing imputation on the "sidebar_basic_color" and "link_basic_color" variables using MICE, we compared the distribution of the imputed values to the original distribution. Here are the resulting plots:

Comparison of Original and Imputed Distributions



As we can see in the plots, the imputed distribution remains similar to the original distribution, indicating that the imputation did not significantly alter the underlying data distribution.

This is a desirable outcome, as imputation methods aim to fill in the missing values in a way that preserves the underlying structure and patterns of the data. A good imputation method should not significantly alter the distribution or relationships within the data, as doing so could introduce bias and affect the validity of the subsequent analyses.

Therefore, the fact that the imputed distribution remains similar to the original distribution provides confidence that the imputation method used was appropriate and effective in filling in the missing values.

OUTLIERS

Data analysis includes an essential process known as outlier detection, which aims to identify atypical or anomalous values in a dataset. These outliers are values that are outside the normal or expected range and may be due to measurement errors, inaccurate records, or simply extreme values that do not follow the normal distribution of the data.

Outlier detection is crucial, as they can significantly impact the results of statistical analyses and models. Additionally, outliers can negatively affect the accuracy and robustness of the models, as they can distort the relationship between variables and responses. Therefore, it is essential to carry out proper detection and treatment of outliers to ensure the quality and validity of the results obtained in data analysis. Numerous techniques can detect outliers, such as univariate, bivariate, multivariate (PCA) approaches, LOF, Mahalanobis distance, Isolation Forest, among others. However, we will focus solely on these mentioned methods, as they are the ones covered in class.

Preliminary Steps

We begin with a database consisting of 26 variables, many of which are unnecessary for outlier detection. Therefore, the first step we will take is to remove such irrelevant variables using logical reasoning:

X.1: This is an identifier (ID), so it does not influence outlier detection.

X: This is another identifier (ID) and, like the previous case, does not affect the process.

Text: This is a character-type (chr) variable.

Description: This is also a character-type (chr) variable.

Profile_yn: Contains a single value, making it irrelevant for analysis.

Created: This is a date (date) variable that has not yet been processed adequately.

Location: This corresponds to geographical coordinates.

X_unit_state: Includes a single value, making it useless for outlier analysis.

description_avg_word_length: 2nd generation variable.

description_word_count: 2nd generation variable.

description_length: 2nd generation variable.

tweet_avg_word_length: 2nd generation variable.

gender.confidence: Target variable.

gender: Target variable.

By eliminating these unnecessary variables, we can focus on the relevant variables and improve the efficiency and accuracy of outlier detection in our database.

Strategy

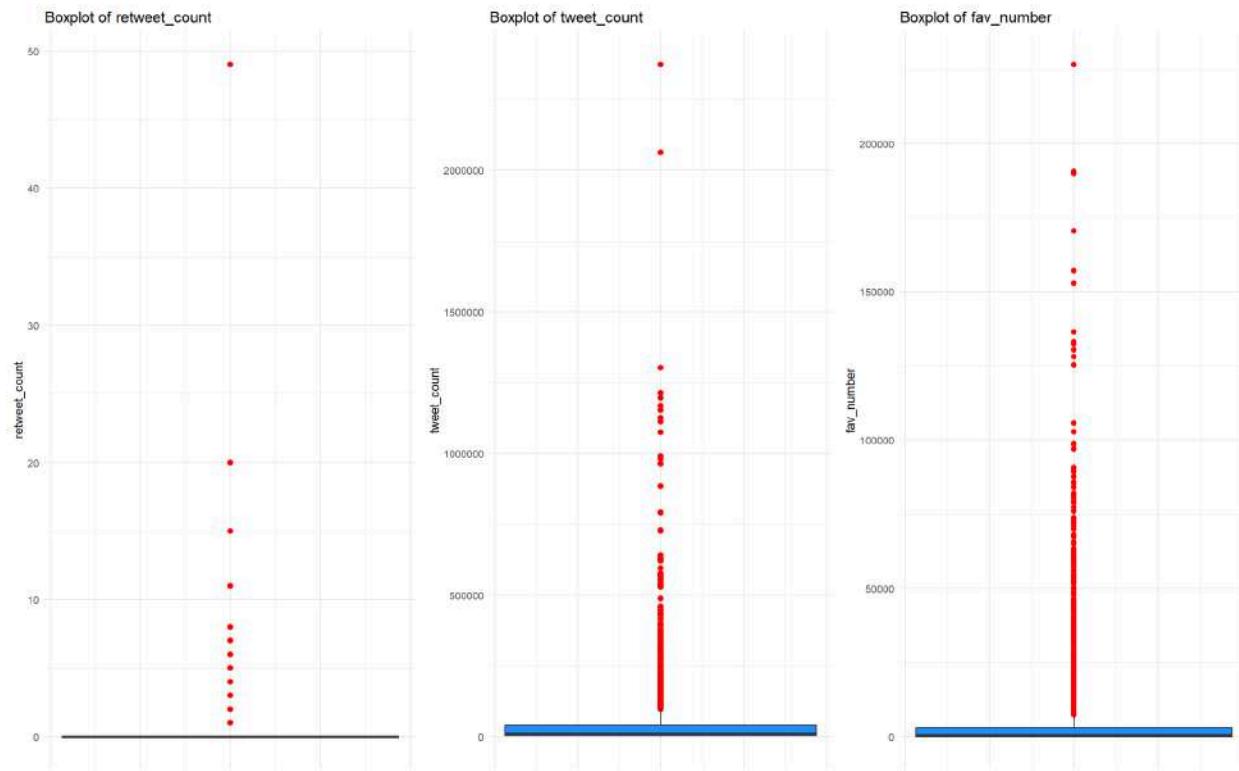
As we mentioned previously, we will address outlier detection using 7 different methods. We will analyze each of them to determine which is most suitable to apply in our case.

Univariate Approach

We have selected this method due to its simplicity and effectiveness in identifying atypical values in each numeric variable individually. It is useful when working with datasets containing variables with normal and non-normal distributions. Moreover, this approach provides a solid foundation for understanding the individual behavior of each variable and how they are affected by extreme values.

The IQR (Interquartile Range) method works by calculating the range between the first quartile (25th percentile) and the third quartile (75th percentile) of a dataset. Outliers are then determined by examining any data points that fall below the first quartile minus 1.5 times the IQR or above the third quartile plus 1.5 times the IQR. This approach is particularly useful for detecting outliers in datasets with skewed distributions or those containing extreme values.

We will apply this method to numeric variables such as '*retweet_count*', '*tweet_count*', '*fav_number*' and '*gender.confidence*'. Specifically, we will use the IQR (Interquartile Range) method to identify outliers.



As we can see in the graphs, at first glance, 3 of the 3 variables might contain outliers. However, in the case of retweet_count, these do not appear to be actual outliers but rather extreme values that should be taken into account and are not irrational.

On the other hand, in the case of tweet_count and fav_number, we can observe illogical values, such as a user who has published 2,000,000 tweets or has given 200,000 likes. These values are clearly identified as outliers.

Taking into account the analysis carried out with the univariate approach and the generated graphs, we can conclude that this method is suitable for identifying outliers in some of the numeric variables present in our dataset.

Bivariate Approach

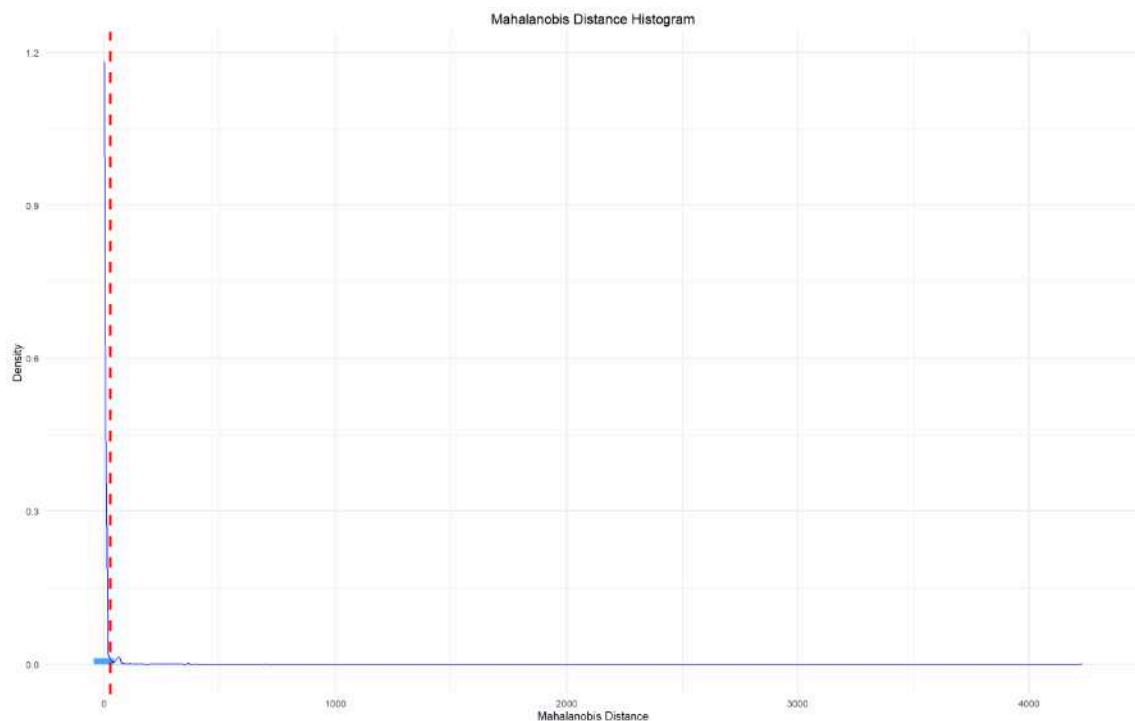
We have not selected this approach because, although it can be useful for analyzing the relationship between two variables, its ability to detect atypical values is limited compared to multivariate methods, such as Mahalanobis distance and isolation forest, which can analyze the interaction among multiple variables simultaneously.

Mahalanobis Distance

The selection of Mahalanobis distance as a method for outlier detection is based on its ability to take into account the correlation between variables in the process. This aspect is relevant when working with non-normally distributed variables, as other approaches, such as univariate or bivariate, might not be sufficient to identify outliers in the interaction among multiple variables. The Mahalanobis distance is a robust method that can address correlation and non-normality in the dataset.

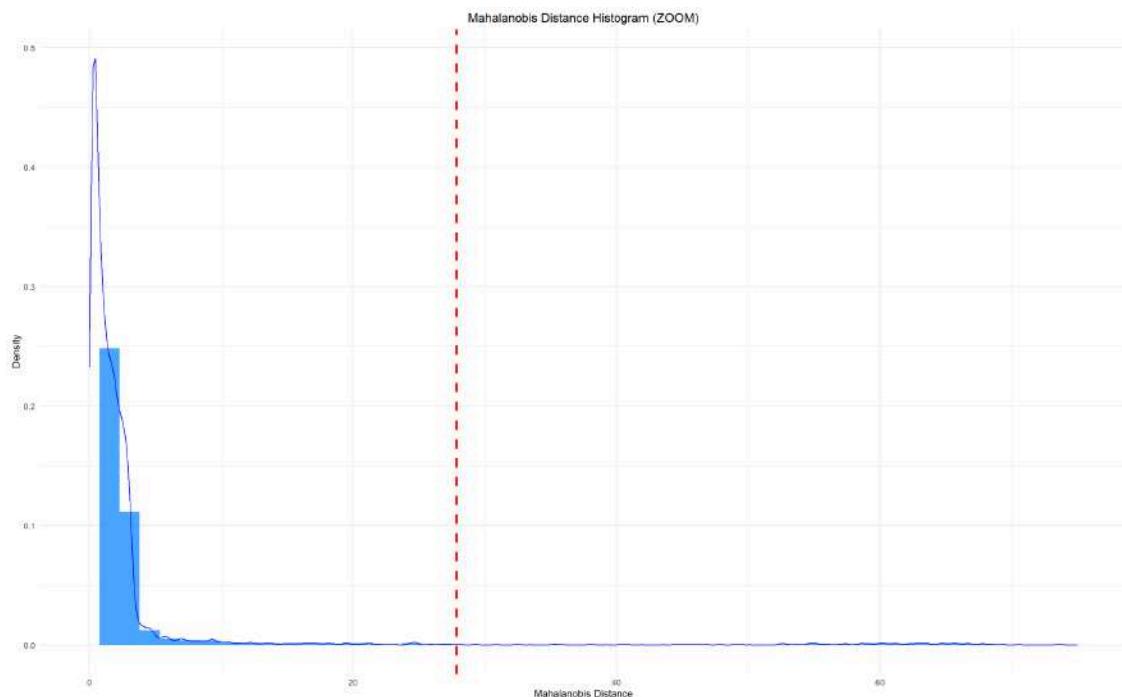
The Mahalanobis distance works by measuring the distance of a data point from the mean of the data, considering the covariance between variables. In other words, it takes into account the shape and orientation of the data distribution. This method is particularly useful for multivariate outlier detection, as it can reveal the relationships between variables that other methods might overlook.

To apply this method, we again select only numeric variables. After performing various tests, we have determined that the optimal threshold is 0.001. This value is small enough to ensure greater rigor and avoid the elimination of truly important values.



The graph presented is a histogram generated using GGplot2, which illustrates the distribution of Mahalanobis distances while emphasizing the critical value with a dashed red vertical line. This critical value, or threshold, is used to identify outliers based on Mahalanobis distance. Points that have Mahalanobis distances greater than the threshold and fall to the right of the red line are considered potential outliers in the multivariate space. Visualizing the threshold with the red line allows for a more straightforward determination of which points warrant further investigation or possible removal from the dataset due to their outlier status.

In the first graph, we can see that the majority of the values are close to 0, yet the graph displays distance values of up to 4,000. This could be due to extreme outliers in the dataset, possibly caused by the values of fav_number = 305,390 and tweet_count = 2,372,591. As a result, we decided to create a second graph with a zoomed-in view to better observe the results.



By presenting the second graph, we can more effectively analyze the distribution of Mahalanobis distances and gain a clearer understanding of potential outliers. This additional visualization offers valuable insights into the data and supports more informed decisions when addressing these outliers.

Multivariate Approach (PCA)

Although Principal Component Analysis (PCA) is useful for reducing dimensionality and detecting outliers, I did not select it because Mahalanobis Distance and Isolation Forest already effectively address multivariate analysis and can handle non-normally distributed variables, which are present in the dataset.

LOF (Local Outlier Factor)

I did not choose this method because, although it is useful for detecting outliers in datasets with clusters, it is more sensitive to parameter selection and may be less robust compared to Isolation Forest. Additionally, Isolation Forest is more computationally efficient and can handle high-dimensional datasets more easily.

Isolation Forest

We have chosen the Isolation Forest method due to its effectiveness and efficiency as a machine learning algorithm in detecting outliers in high-dimensional datasets. By building random decision trees and isolating atypical observations, Isolation Forest is capable of identifying outliers without assuming a specific distribution for the variables, making it a solid choice in combination with the other selected methods.

The Isolation Forest algorithm works by recursively partitioning the dataset into smaller subsets using randomly selected features and split values. In this process, each observation is isolated, and the number of steps required to isolate an observation represents its isolation score. Outliers tend to have shorter isolation paths because they are typically further from the majority of the data points. By averaging the isolation scores across multiple decision trees, the algorithm assigns an anomaly score to each observation, which can be used to determine if a data point is an outlier.

It is worth mentioning that part of the code comes from the shared class folder (PMAAD-LAB). When applying the code, we encountered an error related to the 'ranger' package, a dependency of 'solitude', which has a limit on the number of levels it can handle in an unordered categorical variable. In this case, the variables "country" and "timezone" have more than 53 levels. To solve this, we grouped the less frequent levels in both variables into a category called 'Other'.

Next, we tackled the selection of the best parameters for Isolation Forest. Since we do not have prior information about the outliers, we applied a grid search and compared the mean depth (mean_depth) of the trained models.

Grid search: This technique consists of specifying a set of possible values for each parameter that needs to be optimized. Then, a model is trained for each possible combination of these parameter values, and the performance of each model is evaluated.

Comparison of mean depth (mean_depth): The mean depth of a tree in the Isolation Forest model is a heuristic measure of the model's performance in detecting outliers. In general, a higher mean depth value indicates that the model is more effective at separating outliers from normal observations.

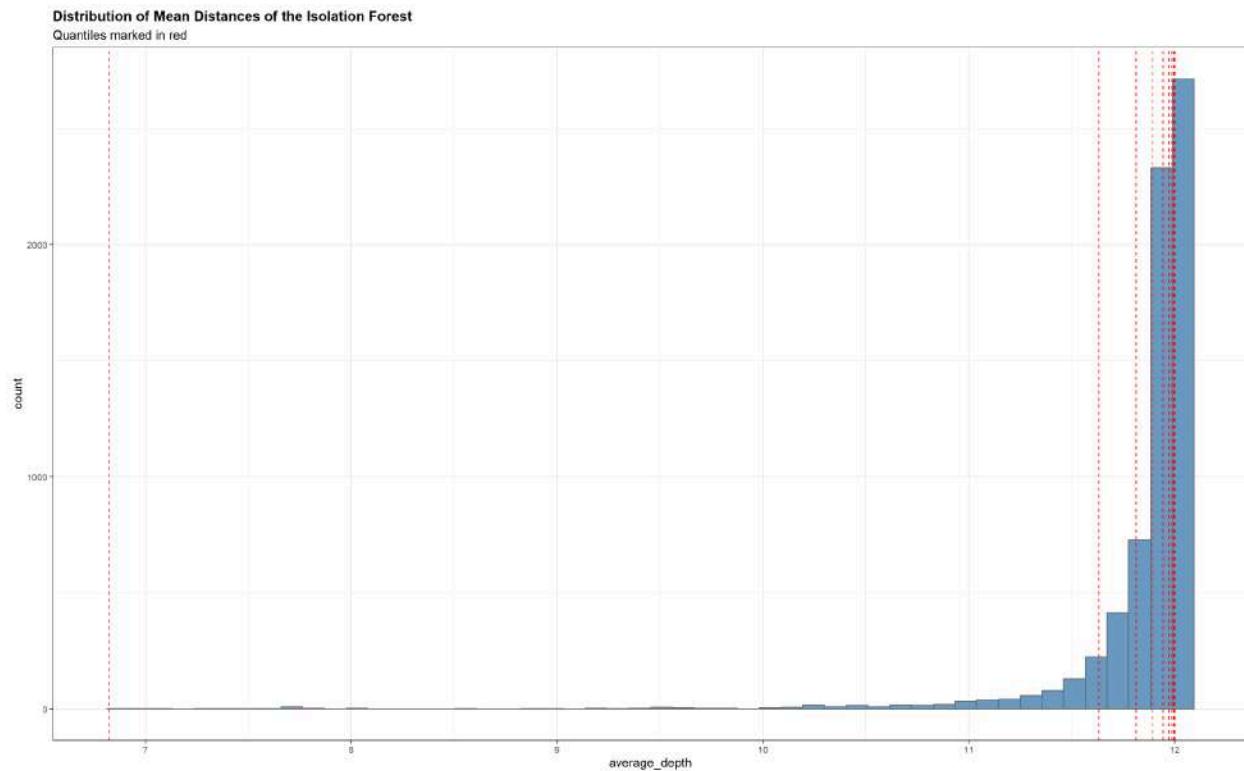
In this case, the performance of each trained model with different parameter combinations was evaluated using mean depth as the criterion. The model with the highest mean depth was selected, and in case of a tie, the model with fewer trees was chosen, as this generally implies lower computational complexity and shorter training and prediction times.

The results obtained were as follows:

sample_size_num_trees	mean_depth
3475_250	11.83752
3475_750	11.83535
3475_500	11.83280
3400_500	11.83133
3400_250	11.83038
3400_750	11.82890

According to the results, the model with the parameters sample_size = 3475 and num_trees = 250 has the highest mean depth (mean_depth = 11.83752). Although the difference in mean depth between the models is small, fewer trees generally mean lower computational complexity and shorter training and prediction times.

Once the optimal parameters were selected, we evaluated the model and represented the results in a graph. The graph shows the distribution of the mean distances of the Isolation Forest, highlighting the quantiles in red.



In the histogram, the horizontal axis represents the average depth values calculated by the Isolation Forest algorithm, and the vertical axis represents the frequency of these values in the dataset. The bars in the histogram, colored in steel blue, show the distribution of the data points according to their mean distances. The red dashed vertical lines in the graph represent quantiles of the average depth values.

These quantiles divide the data into equal intervals, with each interval containing the same proportion of data points. The red lines help to visualize the distribution of the data and make it easier to identify potential outliers, as data points with higher average depths are more likely to be considered outliers.

After thorough testing, we determined that the optimal threshold for detecting outliers using the Isolation Forest method is 0.01. This threshold value ensures a balance between accurately identifying genuine outliers and minimizing the risk of mistakenly classifying normal observations as outliers.

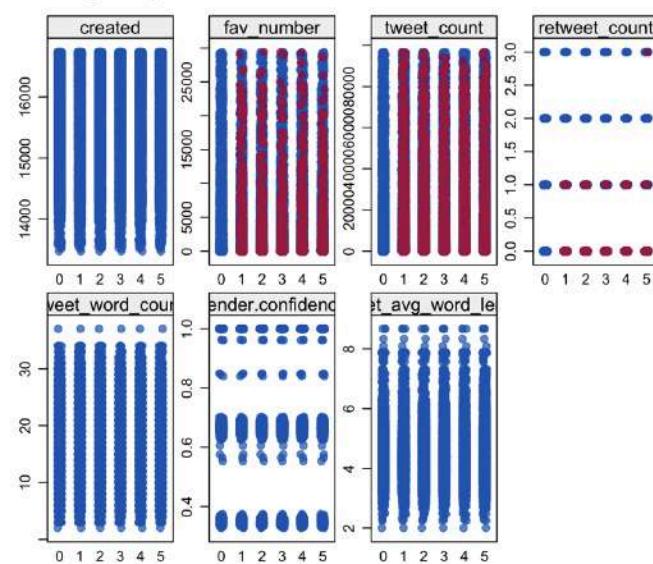
Outlier Treatment

After identifying outliers, the next step is the treatment to handle them. Since we have chosen a very small threshold that ensures these outliers come from errors (e.g. a user having an abnormal number of liked tweets), the best step forward is replacing them with missing values and imputing new values. We chose this treatment because outliers can affect the performance of clustering algorithms like CURE, DBSCAN, and OPTICS as well as MCA, which we will be performing in later sections and the best option in our case is to try to reduce the effects of outliers by replacing them as if they were missing values.

These missing values will then be imputed using the Multiple Imputation by Chained Equations (MICE) method, which we have previously determined to be the best method for our data. After imputing the missing values using MICE, we have checked the distribution of the data by creating a stripplot of the imputations. This allows us to ensure that the distribution of the imputed values remains similar to the original distribution of the data, and that our imputations are accurate and reliable. By following these steps, we can improve the overall quality and reliability of our statistical analysis.

Hereunder is the stripplot of the imputations:

Missings Imputation Outliers MICE



As can be seen from the graph, the MICE method has correctly imputed these values. When it comes to *fav_number* and *tweet_count*, we imputed the complete 5 imputations with this function:

```
> data <- Complete(mice_imp)
```

This function returns a mix of the 5 imputations, which for *fav_number* and *tweet_count* are all correct. On the other hand, for *retweet_count*, we chose to last imputed values only, since we felt they were the ones that adapted better to our data.

Execution

In conclusion, we have applied multiple outlier detection methods to our dataset, including the univariate approach using the IQR method, Mahalanobis distance, and Isolation Forest. Each of these methods has demonstrated effectiveness in identifying atypical values within our data. By combining these techniques and thoroughly analyzing the results, we were able to detect a total of 1,387 outliers in the dataset.

It is important to note that we used very small threshold values for each method; specifically, the Mahalanobis distance threshold is 0.001 and the Isolation Forest threshold is 0.01. These low thresholds ensure that the identified data points are indeed outliers, which makes us confident in treating them similarly to missing values.

These outliers can have a significant impact on the results of our statistical analyses and models. Therefore, identifying and addressing these outliers is crucial to ensure the quality and validity of our data analysis. By utilizing a combination of methods, we have maximized our ability to detect and analyze potential outliers effectively.

Overall, our outlier detection strategy has proven to be both effective and efficient, allowing us to identify and address atypical values in our dataset. This thorough analysis has laid a solid foundation for any subsequent statistical analyses and modeling, ensuring the reliability and accuracy of our results.

Descriptive analysis

Univariate analysis

This report presents an individual analysis of all variables within the preprocessed database, which has been divided into three sections. It should be noted that not all variables have plots included in this analysis, as the number of modalities exceeds the capacity for accurate visualization on a plot. The variables that don't have plots included in this analysis are country, timezone, and created.

Firstly, we analyze the users' profiles to gather information about what we know of the users that are part of the data. This category includes variables such as the sidebar or the link color. This will give us a general overview of the different types of profiles we will be analyzing.

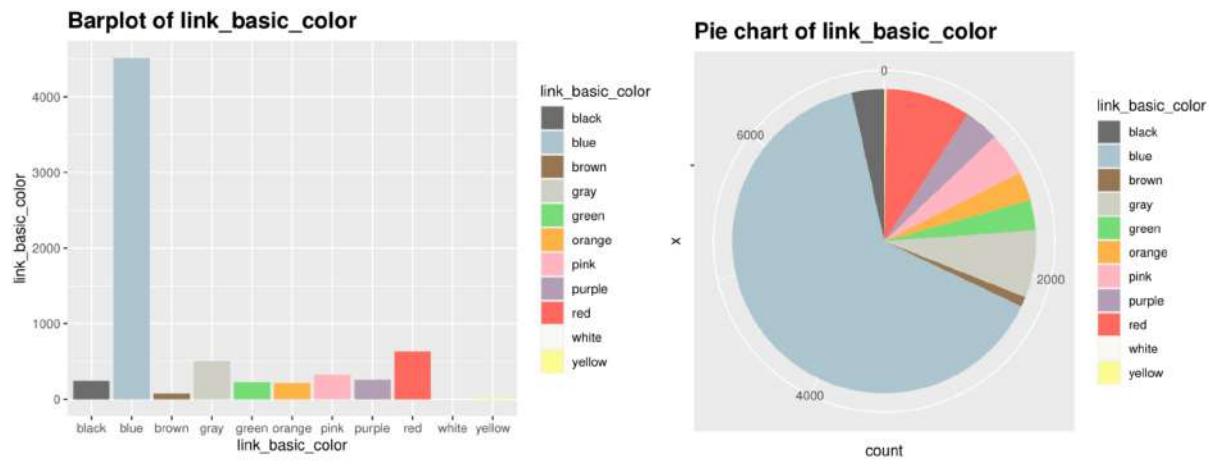
Secondly, we will study the profile user itself, such as which is their gender (which will be our response variable later on in the study) or where they are from.

And finally, we will inspect the user's behavior. We will dig into their number of tweets or how many tweets they've liked for instance.

With the univariate analysis, we aim to take the data, summarize it, and then find some pattern in the data.

User's profile

Link basic color

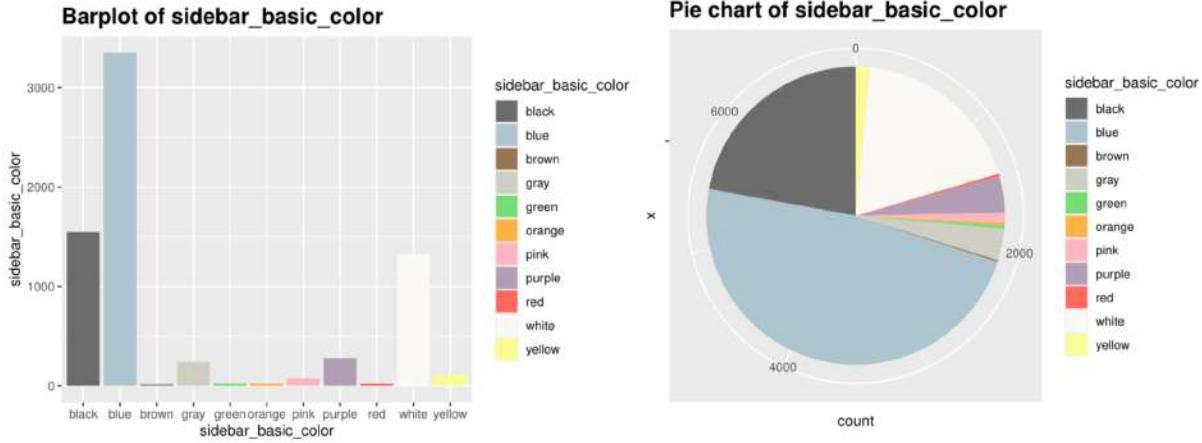


The pie chart and barplot show the distribution of colors in the link. The distribution is clearly not uniform, with blue being the mode and more than half of the respondents being part of it.

So, we can see that most links are blue, followed by red and grey in the distance. There's quite a lot of dispersion between the first and the second most voted modalities so that shows that most other colors are underrepresented. The least represented modality is white, being this the minimum.

This color dominance may be attributed to the fact that blue is the default setting.

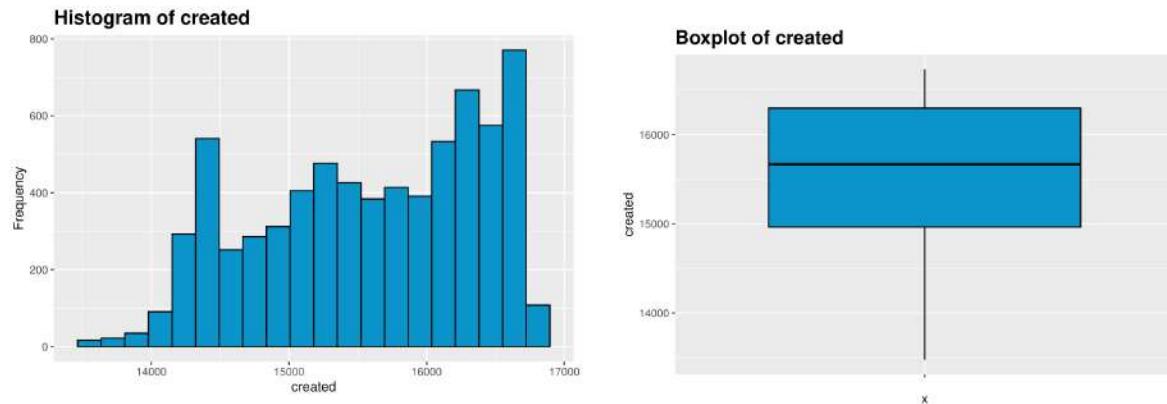
Sidebar basic color



The pie chart and barplot show the distribution of colors on the sidebar. The distribution is clearly not uniform, with blue being the mode and almost half of the respondents being part of it.

So, we can see that most links are blue, followed by grey and white in the distance. There's some dispersion between the first and the second most voted modalities so that shows that most other colours are not as represented. The least represented modality is brown, being this the minimum.

Created



This variable has been created from the transformation of the date type format that indicates when has the profile been created to numeric values.

The two graphical representations of created are skewed, particularly, we can notice that they present a normal distribution slightly moved to the right. That means that it's left-skewed because the peak veers to the right.

As such, the median is encountered a little to the right of the center of the plot. This kind of distribution, however, has also led to the fact that none of the values is considered an outlier.

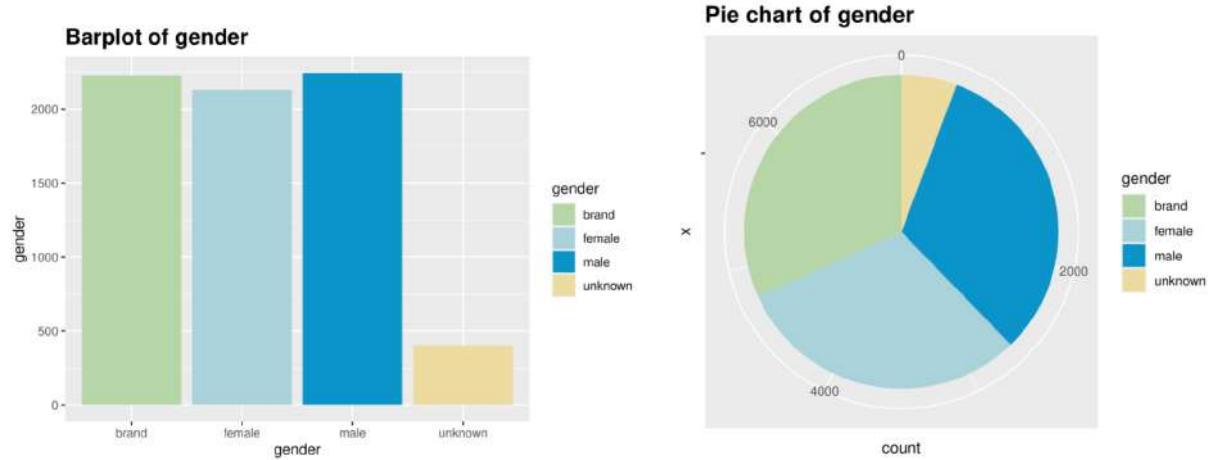
This pattern exemplifies the rise in popularity of the social network since as time passes the more people have created their accounts.

Going on with the summary statistics, and starting with the robust ones, for the variable gender confidence the median is 15667, as observed in the box plot. The first quartile (Q1) is 14962 and the third quartile (Q3) is 16298. For this variable, the summary in five numbers (number of days that have passes since January 1, 1970) is the following:

Min	1st Quartile	Median	Mean	3rd Quartile	Max
13475	14962	15667	15590	16298	16734

User

Gender

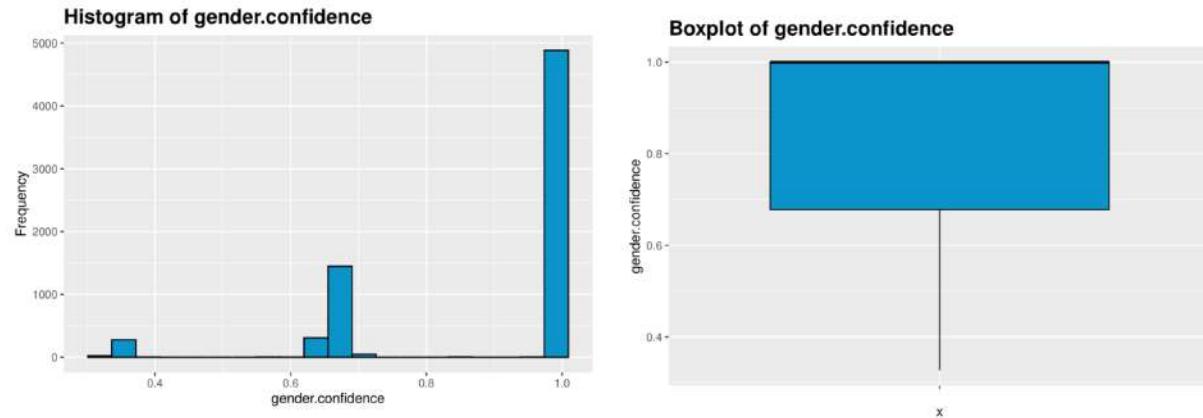


The pie chart and barplot show the distribution of the gender variable. The distribution is almost uniform, with the exception of the 'unknown' modality, which is underrepresented in comparison.

In this case, the mode is male, followed by brand and female in that order. The gender 'brand' refers that the profile that corresponds to that of a company instead of an individual.

So, we can see that most links are male, followed by brand and female. As stated previously, the least represented modality is 'unknown', being this the minimum.

Gender.confidence



The two graphical representations of gender confidence show a distribution which could be compared to three batches. The frequency of each class is organized in an ascending manner. For in, the first batch has the lowest frequency, the last the highest and so on.

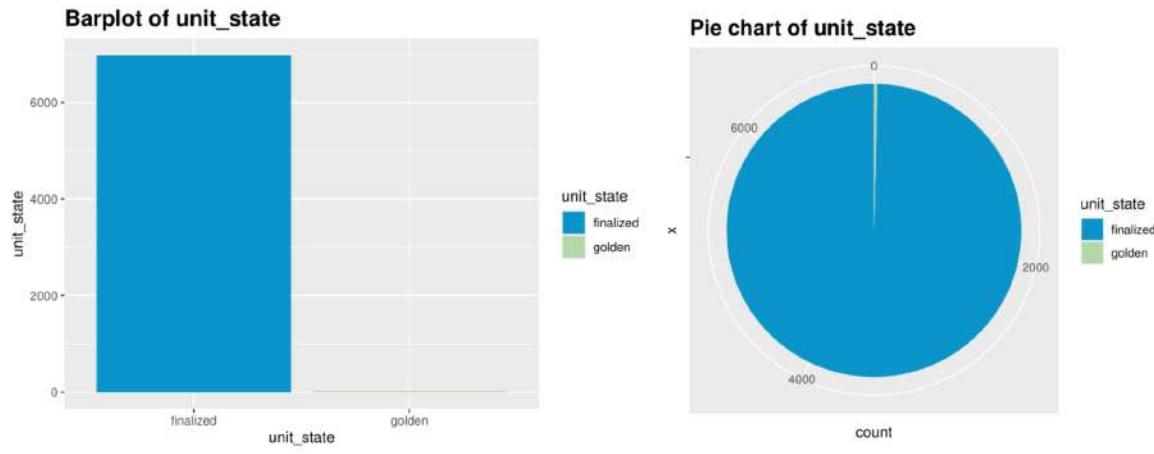
As such, the median is encountered at the rightmost point of the possible range which is 1. This kind of distribution, however, has also led to the fact that none of the values is considered an outlier.

This pattern exemplifies that most of the people that answered the survey to predict gender have the utmost confidence in their prediction.

Going on with the summary statistics, and starting with the robust ones, for the variable gender confidence the median is 1, as observed in the box plot. The first quartile (Q1) is 0.6781 and the third quartile (Q3) is 1. For this variable, the summary in five numbers (in % between 0 and 1) is the following:

Min	1st Quartile	Median	Mean	3rd Quartile	Max
0.3272	0.6781	1	0.8858	1	1

Unit state

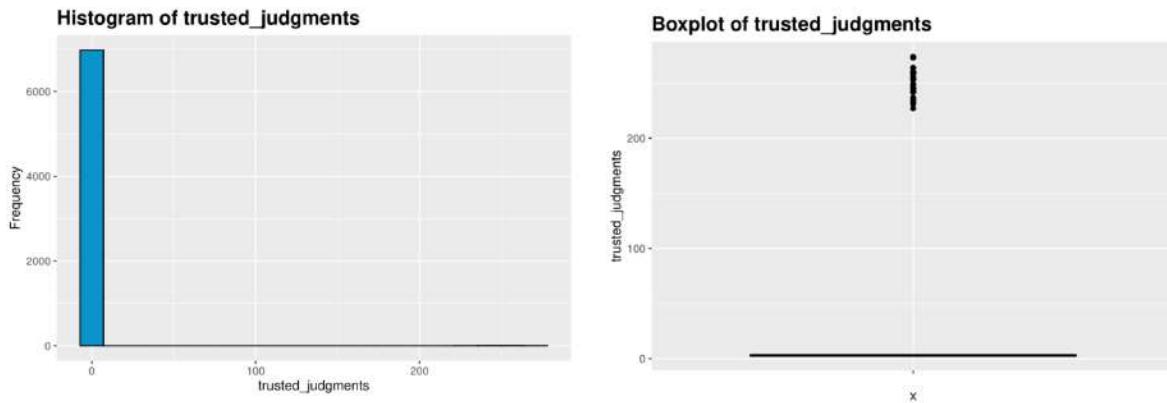


The pie chart and barplot show the distribution of the unit_state variable. The distribution only has two modalities, which are finalized and golden, with golden being grossly underrepresented.

In this case, the mode is the other modality, finalized as it is almost the entirety of the dataset.

As stated before, the least represented modality is golden, being this the minimum.

Trusted judgement



The two graphical representations of the retweet count variable are skewed, particularly, we can notice that they're asymmetrical to the left. That means that it's right-skewed because the peak veers to the left. Which in this case means that the vast majority of tweets have a retweet count of three.

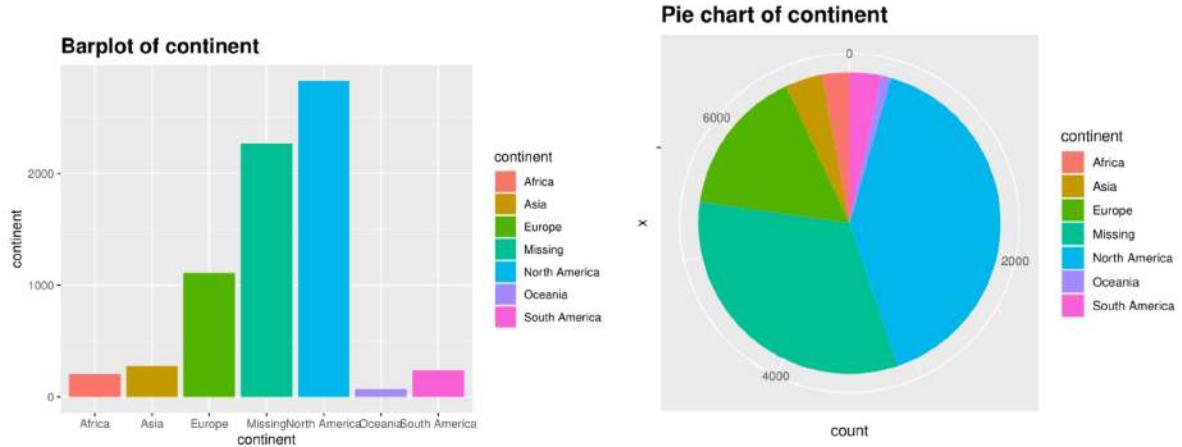
This means that anything outside of 3 is considered an outlier.

This pattern is likely because if the tweeter profile is not golden the dataset ensured that it had 3 observations. And, as the majority of profile are not golden, this leads to the same distribution in trusted judgements as well.

Going on with the summary statistics, and starting with the robust ones, for this variable the median is 3, as observed in the box plot. The first quartile (Q1) is 3 and the third quartile (Q3) is 3. For this variable, the summary is the following:

Min	1st Quartile	Median	Mean	3rd Quartile	Max
3	3	3	3.883	3	274

Continent

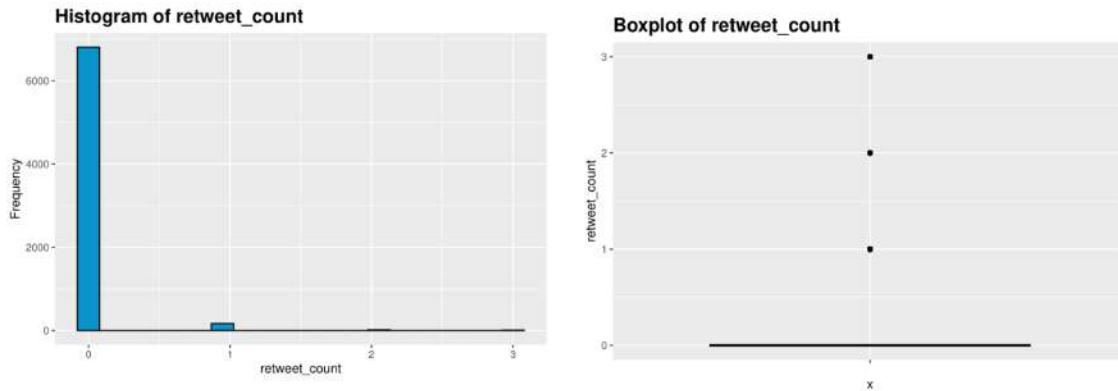


The pie chart and barplot show the distribution of the continent variable. The distribution is clearly not uniform, with North America being the mode and almost half of the respondents being part of it.

So, we can see that most users are from North America, followed by 'Missing' (which is that they didn't want to disclose a location) and Europe. There's quite a lot of dispersion between the first and the second most voted modalities if we don't take into account the 'Missing' modality. This shows that most other continents are underrepresented. The least represented modality is Oceania, being this the minimum.

User's behavior

Retweet count



The two graphical representations of the retweet count variable are skewed, particularly, we can notice that they're asymmetrical to the left. That means that it's right-skewed because the peak veers to the left. Which in this case means that the vast majority of tweets have a retweet count of zero.

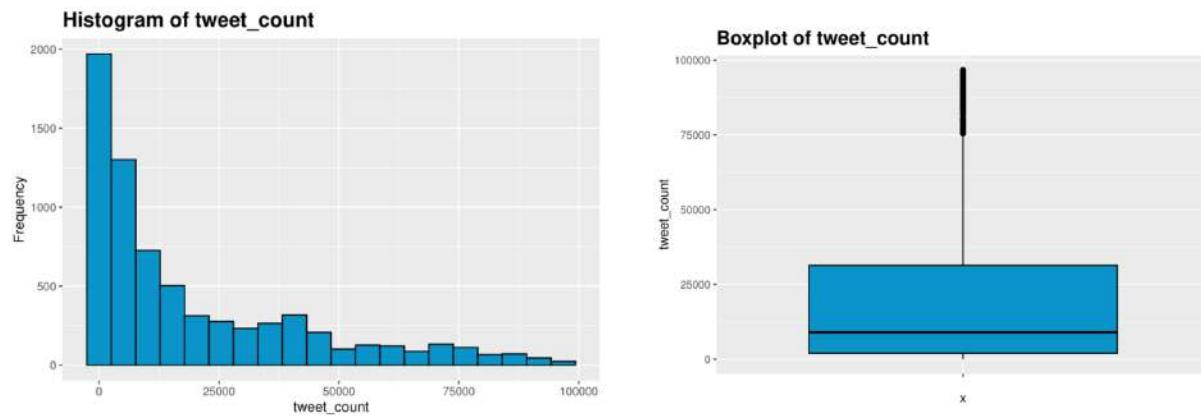
This means that anything that is retweeted, even if once, is considered an outlier. Yet, when retweets do occur, they are found in the range of 1 to 10 retweets, with the majority falling between 1 and 5.

This pattern is likely because most tweets are not particularly noteworthy or interesting enough to be shared by other users. As a result, tweets that receive a large number of retweets are rare and are likely to be seen as particularly notable or impactful by their audience.

Going on with the summary statistics, and starting with the robust ones, for this variable the median is 0, as observed in the box plot. The first quartile (Q1) is 0 and the third quartile (Q3) is 0. For this variable, the summary in five numbers (in number of retweets) is the following:

Min	1st Quartile	Median	Mean	3rd Quartile	Max
0	0	0	0.03214	0	3

Tweet count



The graphical representation of the tweet count variable shows almost the same pattern as the retweet count variable.

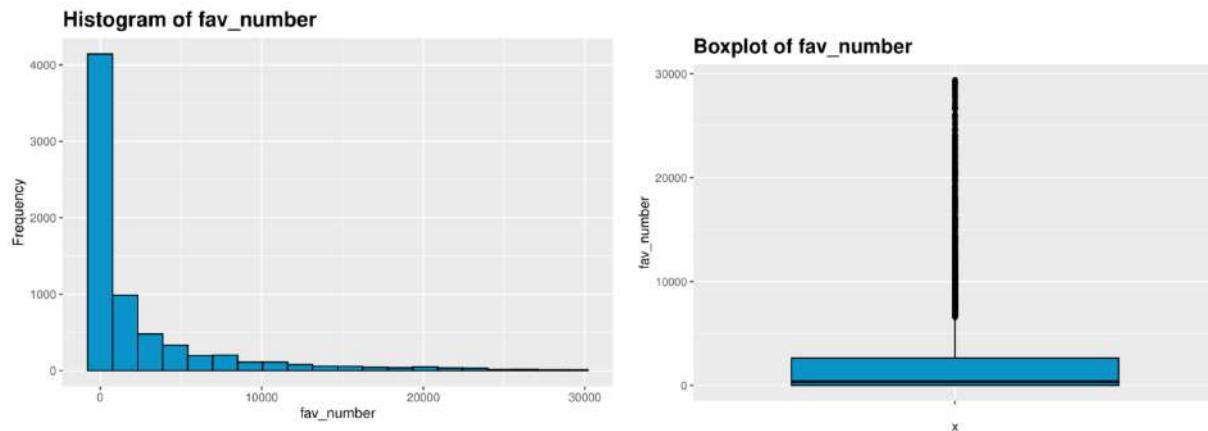
The distribution is less concentrated at zero, as we can see that the third quartile now reaches up to 31329 tweet count. But, we can still observe a skewness in the data, as the peak of the distribution veers towards the left.

In this case, this means that only the users that are prolific tweeter writers are considered outliers. This is expected, as a prolific tweet writer has at least a 75,000 tweet count. Which is remarkably high when taking into consideration that the majority of users are at 0.

Going on with the summary statistics, and starting with the robust ones, for the variable tweet count the median is 8936, as observed in the box plot. The first quartile (Q1) is 1960 and the third quartile (Q3) is 31329. For this variable, the summary in five numbers (in number of tweets) is the following:

Min	1st Quartile	Median	Mean	3rd Quartile	Max
1	1960	8936	19478	31329	96748

Favorite number



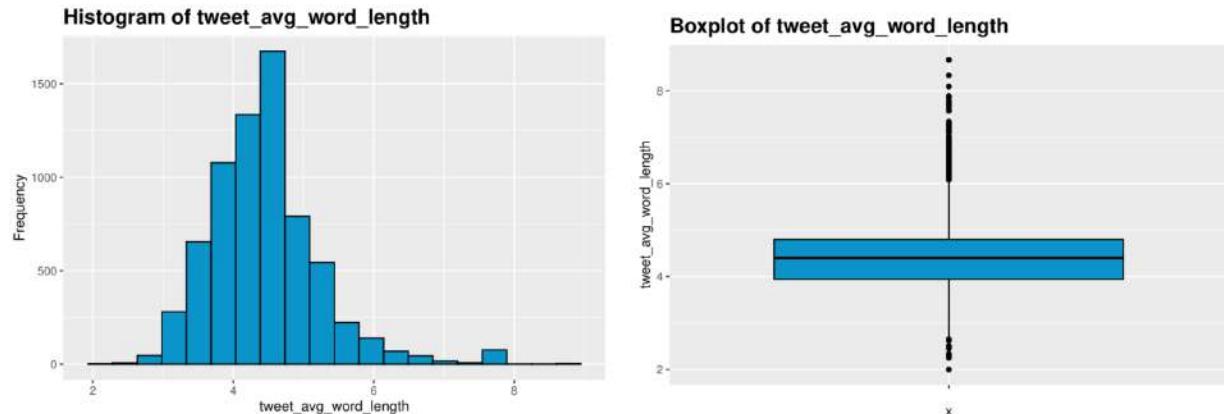
The graphical representation of the favorite number variable shows almost the same pattern as the tweet count variable but is even more dispersed.

The distribution is less concentrated at zero. Yet, we can see that the quartiles have the same proportions as the previous variable. We can still observe a skewness in the data, as the peak of the distribution veers towards the left. In this case, this means that only the users that have a huge number of favourited tweets are considered outliers. This is expected, as this type of user needs to have at least more than 5,000 favourited tweets. Which is remarkably high when taking into consideration that the majority of users are at 0.

Going on with the summary statistics, and starting with the robust ones, for this variable the median is 359, as observed in the box plot. The first quartile (Q1) is 6 and the third quartile (Q3) is 2634. For this variable, the summary is the following:

Min	1st Quartile	Median	Mean	3rd Quartile	Max
0	6	359	2536	2634	29410

Tweet average word length



The two graphical representations of the average tweet length variable are skewed, particularly, we can notice that they present a normal distribution slightly moved to the left. That means that it's right-skewed because the peak veers to the left. Which in this case means that the average word length of a tweet resides between 4 and 5 characters.

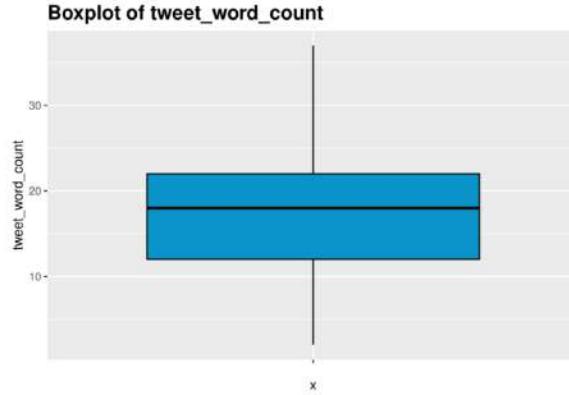
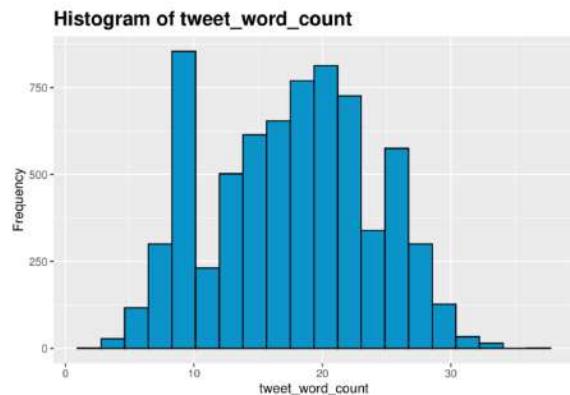
This means that any word below and above this range is considered an outlier. This is to be expected, as there exist shorter and longer words than the average.

This pattern exemplifies what is the common length in words of the English lexicon when speaking of casual language.

Going on with the summary statistics, and starting with the robust ones, for this variable the median is 4.4, as observed in the box plot. The first quartile (Q1) is 3.941 and the third quartile (Q3) is 4.8. For this variable, the summary is the following:

Min	1st Quartile	Median	Mean	3rd Quartile	Max
2	3.941	4.4	4.44	4.8	8.667

Tweet word count



The two graphical representations of the tweet word count variable aren't skewed, although we can notice the distribution moved to the left a little. This shift in distribution is due to the high number of tweets with a word count of 9, which skews the distribution.

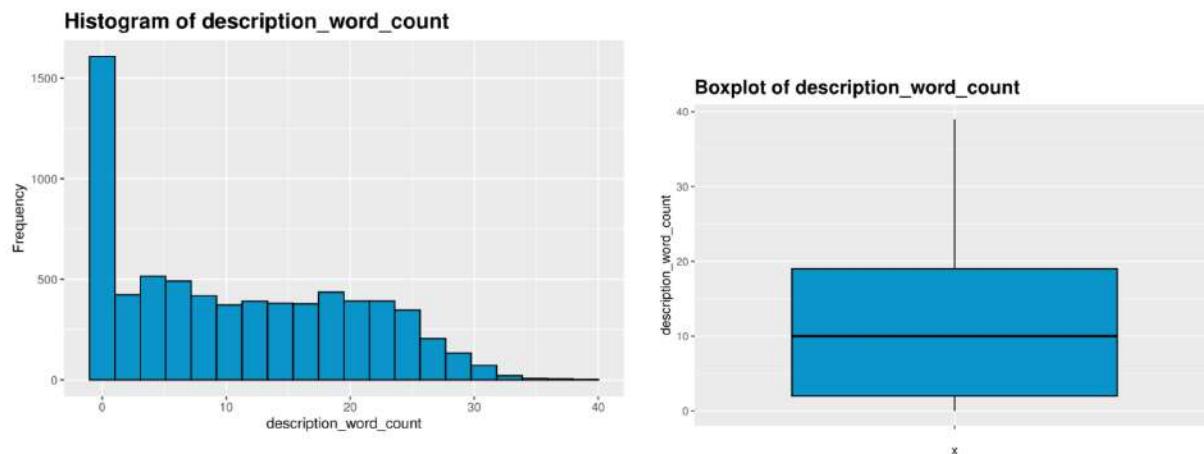
Having the median at 17 and a range between 0 and a little past 30 means there's no possibility for outliers, as can be seen in the boxplot.

This pattern exemplifies how most tweets follow a common distribution in terms of word count. This tendency could be explained by the existence of a character number limit.

Going on with the summary statistics, and starting with the robust ones, for this variable the median is 18, as observed in the box plot. The first quartile (Q1) is 12 and the third quartile (Q3) is 22. For this variable, the summary is the following:

Min	1st Quartile	Median	Mean	3rd Quartile	Max
2	12	18	17.64	22	37

Description word count



The graphical representation of the description word count variable shows almost the same pattern as the tweet count variable. As the tail shown in the histogram is much larger this time.

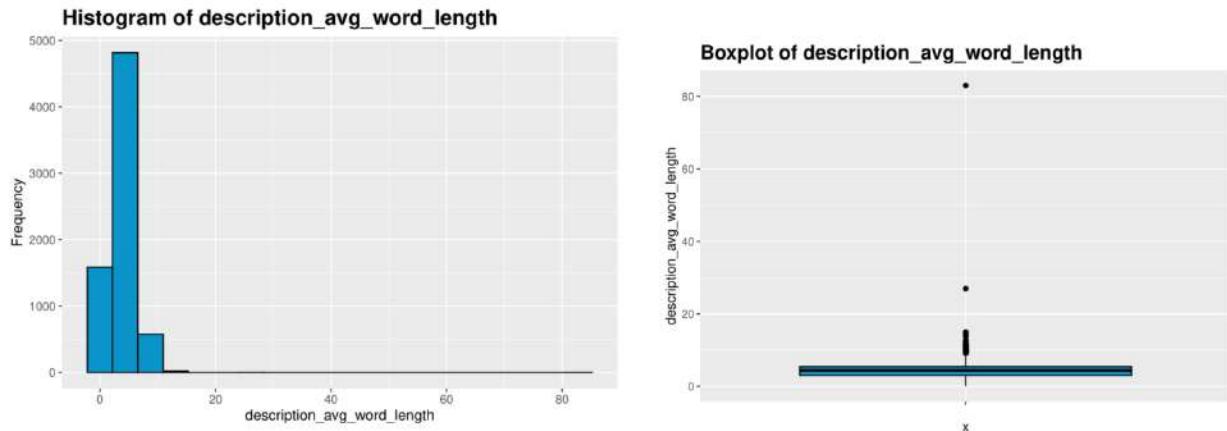
The distribution is less concentrated at zero, as we can see that the third quartile now reaches almost 20 words. But, we can still observe a skewness in the data, as the peak of the distribution veers towards the left.

In this case, this means that there are no possible outliers. This is expected, as most people don't have a description but those that do have a character limit that doesn't steer too far off from 0.

Going on with the summary statistics, and starting with the robust ones, for this variable the median is 10, as observed in the box plot. The first quartile (Q1) is 2 and the third quartile (Q3) is 19. For this variable, the summary is the following:

Min	1st Quartile	Median	Mean	3rd Quartile	Max
0	2	10	10.99	19	39

Description average word length



The distribution is less concentrated at zero, as we can see that now the value with peak frequency is around 5 characters, followed up by 0 and 10. But, we can still observe a skewness in the data, as the peak of the distribution veers towards the left.

This means that any word above 10 characters is considered an outlier. This is to be expected, as there exist longer words than the average.

In this case, this means that only the users that are prolific description writers are considered outliers.

Going on with the summary statistics, and starting with the robust ones, for this variable the median is 4.429, as observed in the box plot. The first quartile (Q1) is 3 and the third quartile (Q3) is almost 5.435. For this variable, the summary is the following:

Min	1st Quartile	Median	Mean	3rd Quartile	Max
0	3	4.429	3.913	5.435	83

Bivariate analysis

In this report, we have studied and analyzed the relations between pairs of variables in order to discover if they are related in any way. We have picked different pairs of variables, represented as the *Explanatory variable* \times *Response variable*, where the response variable is the one dependent on the explanatory variable, according to our criteria. Of course, a relationship of dependence may not exist between the variables but finding out is the purpose of this analysis.

There are three types of variable pairs we can study:

- **Quantitative x Quantitative**

- *retweet_count* \times *fav_number*
 - *tweet_length* \times *fav_number*

We want to see which variables play a factor in the amount of likes a user has given. We will study whether the amount of retweets or the tweet length of the user is correlated to *fav_number*.

- **Qualitative x Qualitative**

- *link_basic_color* \times *sidebar_basic_color*
 - *gender* \times *sidebar_basic_color*
 - *gender* \times *link_basic_color*

We are also interested in seeing if their gender might affect the colors they chose to personalize their profile. Does a specific gender have more tendency to choose a color? We have also decided to analyze how people decide to combine their link and sidebar color.

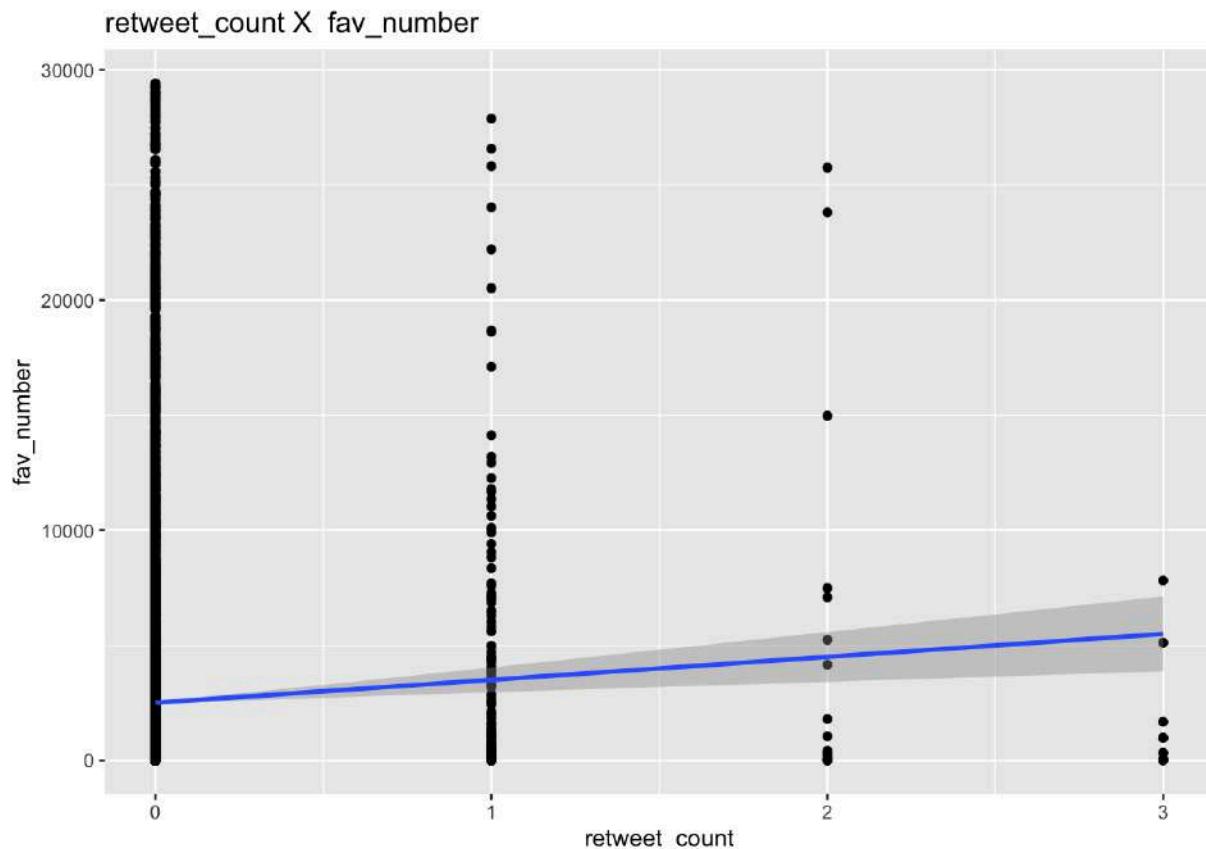
- **Quantitative x Qualitative**

- *gender* \times *fav_number*
 - *gender* \times *tweet_length*

Finally, we are also going to analyze if gender is a discriminating factor when it comes to *fav_number* or the tweet's length. Does a specific gender have a tendency to like more tweets? Do they also tend to write longer tweets?

retweet_count x fav_number

First up, let's analyze the relationship between *retweet_count* (the number of retweets the user's tweet has) and *fav_number* (how many tweets the person has liked)



Outliers

Before studying the actual relationship between our two variables, we must examine the number of outliers found in them. Looking at the scatter plot we see no apparent outliers, this could be because of our preprocessing where we eliminated the outliers according to a threshold.

Scatter plot

The plot shows us the number of retweets a user has on the X axis, this variable doesn't allow decimal numbers which is why we see the points scattered in three columns, which is the range of this variable. In this

plot, we are visualizing the relationship of this variable with the number of tweets the user has liked.

We can see that the majority of our individuals lie in the 0 retweets, however, the amount of likes they have varies between a range of 0 and 30 000 likes. However, although it is hard to see, the majority of these users lie at the bottom portion of like number. In any case, this means that if a user has 0 retweets it is going to be hard to estimate the number of likes only based on this variable. However, we can see that as the number of retweets progresses there are less individuals with a higher amount of likes.

Linear relationship

It is going to be hard to determine whether these two variables have a linear relationship, since our database consists of a short range of retweet number per user and the majority of these users have 0 retweets.

Covariance & correlation

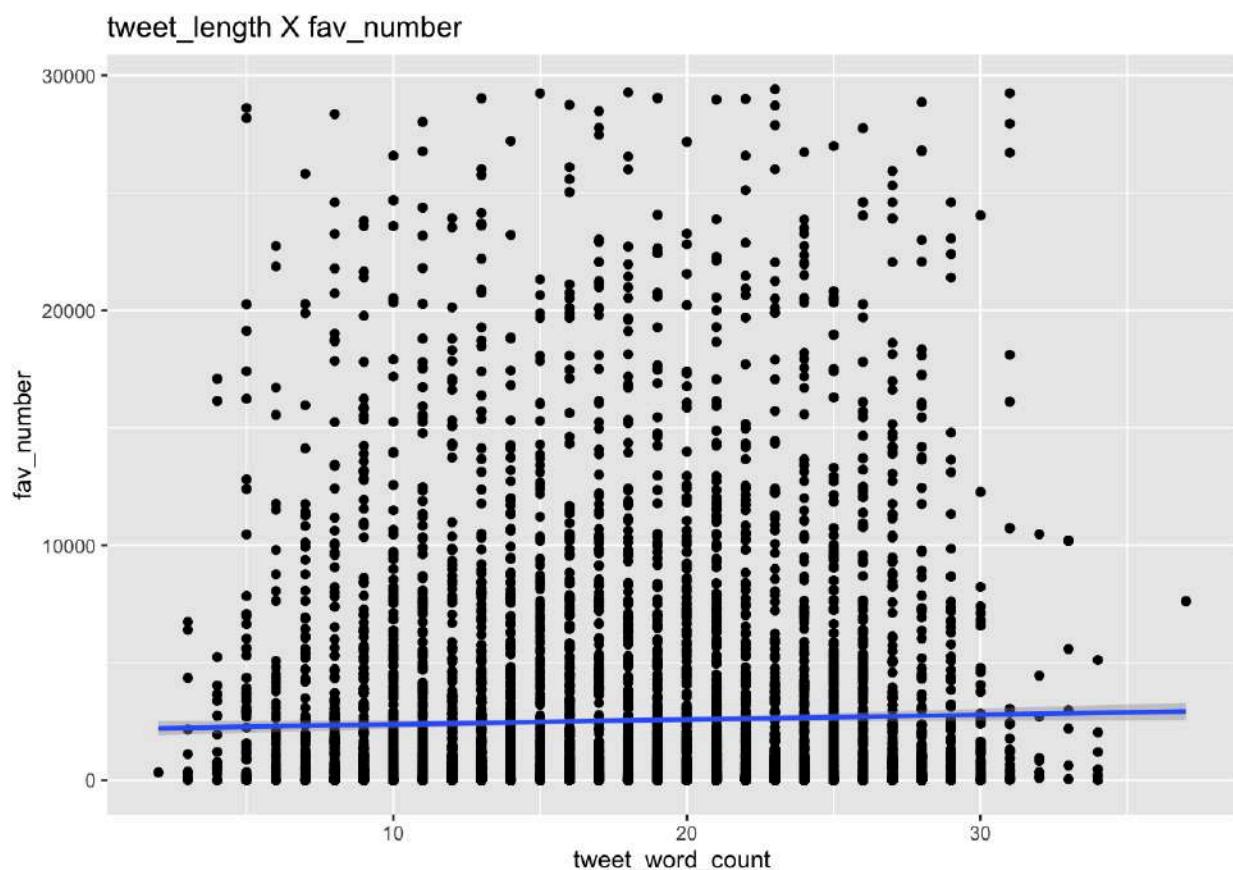
Now we will evaluate the covariance and correlation measures between the two values. Using the R functions to calculate them, as seen below, we obtain a covariance score of 42.67, since this value is not standardized, and strongly depends on the scale of the variables. On the other hand, correlation not only allows us to measure the direction of the relationship, but also its strengths. Correlation ranges from -1 to 1, the closer to these values, the stronger the relationship, however as we can see, the correlation is near 0 in this case, indicating that there is no relationship between the both.

Hereunder the covariance and correlations results have been inserted:

```
> cov(retweet_count, fav_number)
[1] 42.67007
> cor(retweet_count, fav_number)
[1] 0.04269973
```

tweet_length x fav_number

Next up, we will analyze if there is any relationship between *tweet_length* (the length of the tweet), with *fav_number* (the number of tweets the user has liked).



Outliers

It is important to ensure that there are no visible outliers, in order to confirm our outlier treatment's success. As can be seen from the plot, it seems that there is no clear outlier.

Scatter plot

We can observe how most of the observations fall on the 0 to 30

`tweet_word_count`, while the `fav_number` variable has most observations between 0 to 10000. If we look at the scatter part with the most density, we can observe a normal distribution, where values around 20 (`tweet_word_count`) are in the middle of the distribution. This tells us that the majority of the population has around 20 words in their tweet, while they like from 0 to 10000 tweets. The rest of the population that have more or less words in their tweet, also have a smaller `fav_number`. It is interesting to see that, apart from this normal distribution, we find another part of the population above this distribution. This part represents the people who are above the threshold of 10.000 likes. We can see how, for this part of the population, there is no clear number of word count where they are centered.

Linear relationship

As is represented with the blue, we can observe the relationship between the two variables. Considering that the line practically has no slope, this indicates that there is no strong linear relationship between these two variables. This means that, if the word count increases, the user's number of liked tweets does not necessarily increase or decrease proportionally.

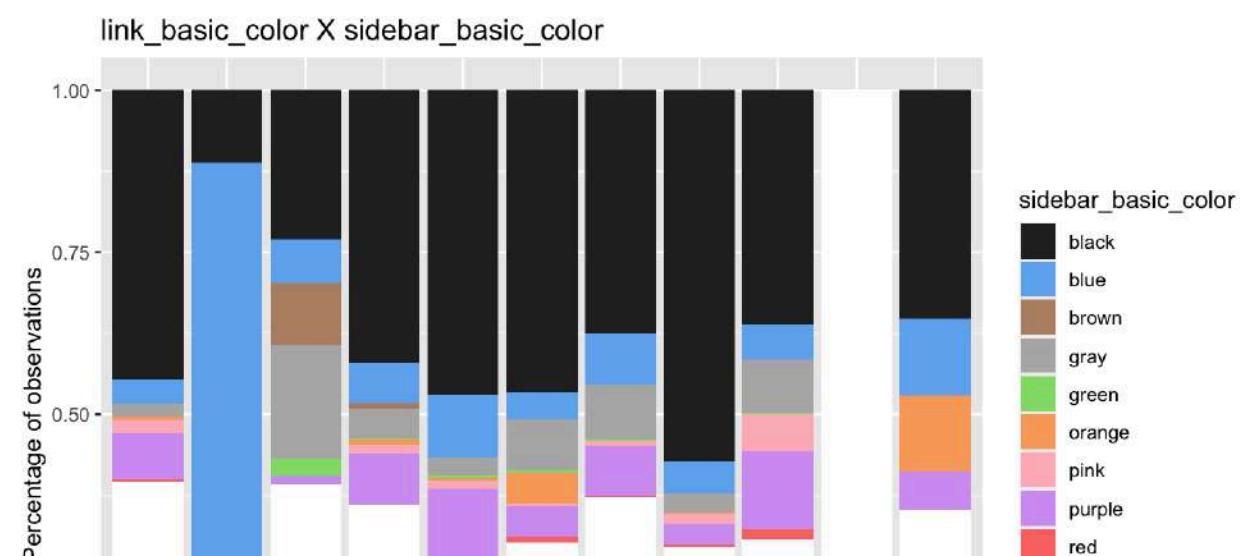
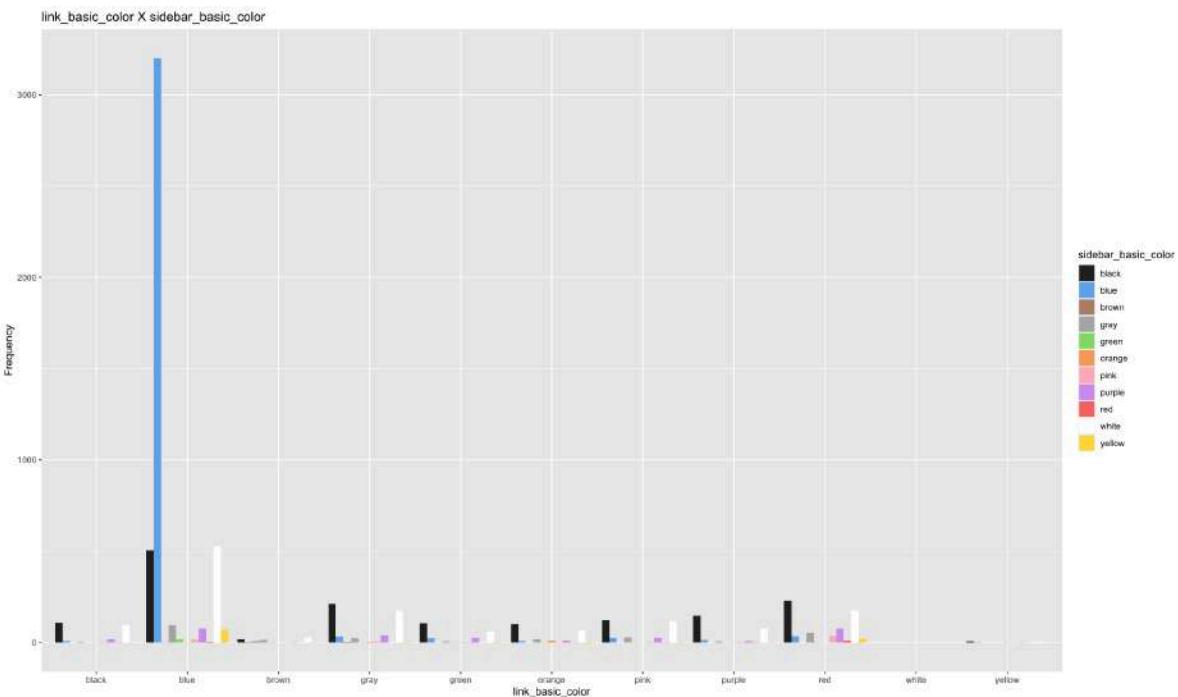
Covariance & correlation

Lastly, the covariance and correlation results for these two variables (which we have inserted below) indicate what we have seen on the graph. While there is technically a positive correlation between the two variables, it is not big enough to be significant.

```
> cov(tweet_word_count, fav_number)
[1] 781.3878
> cor(tweet_word_count, fav_number)
[1] 0.02622478
```

link_basic_color x sidebar_basic_color

Let's look at the relationship between two categorical variables now: link basic color and sidebar basic color:



As we can see, the first graph is not particularly useful since, as we have seen in the univariate analysis, blue is the predominant color, and it does not allow us to correctly interpret the relative proportions of the categories. For this reason, the second graph, which consists of the percentage of observations, is much more useful. As we can see, the proportion of users who have the same *sidebar_basic_color* as *link_basic_color* increases when it is one of these colors: black, blue, brown, orange, red and white. Additionally, we can observe an increase in proportion between different colors. For example, we find that, when the link color is brown, the proportion of sidebar color being gray increases significantly. Same thing happens with sidebar purple, which is increased when the link is red. It is also interesting to see that people with white on the sidebar also have white on the link color. This could be due to many reasons, such as that white is the default color and therefore people who have it mean that they haven't changed their profile.

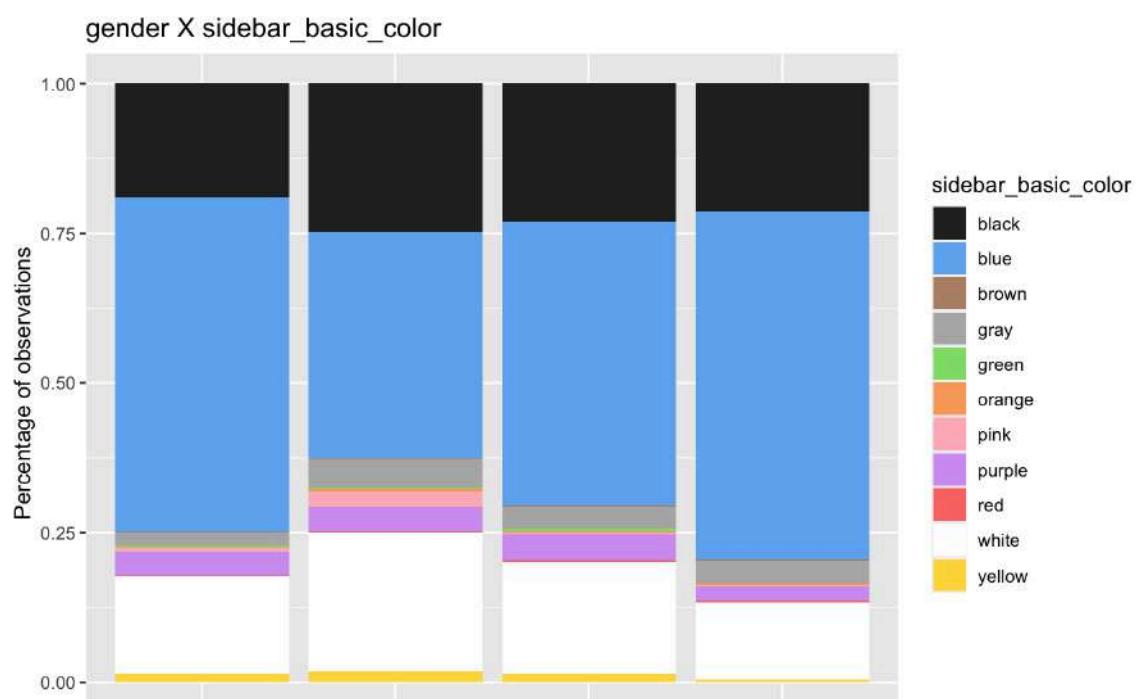
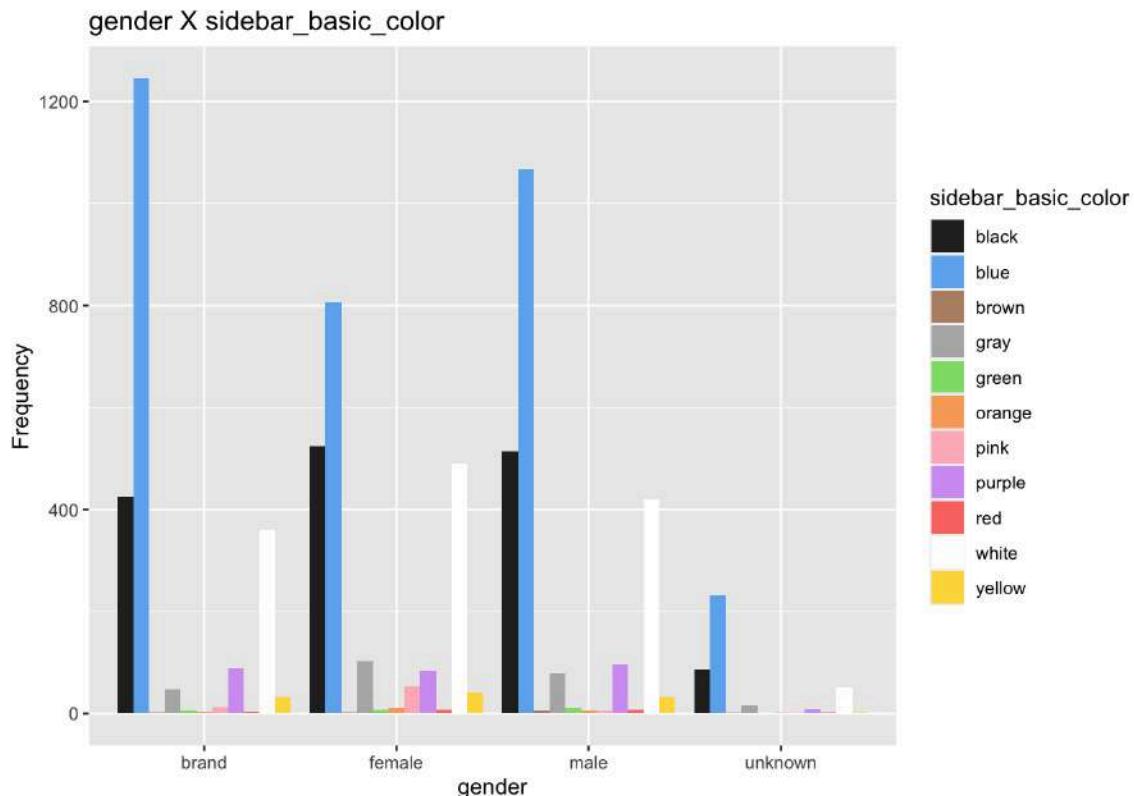
All of what we have seen tells us that the *link_basic_color* and *sidebar_basic_color* do have a relationship between each other. Nevertheless, it is useful to prove our interpretation with a chi-square test, which is used to determine if there is a significant association between the variables:

```
> chisq.test(link_basic_color, sidebar_basic_color, simulate.p.value =  
TRUE)  
  
Pearson's Chi-squared test with simulated p-value (based on 2000  
replicates)  
  
data: link_basic_color and sidebar_basic_color  
X-squared = 3746.2, df = NA, p-value = 0.0004998
```

As we can see, the p-value is smaller than 0.05, and therefore, the null hypothesis that there is no relationship can be rejected. This proves that there is an association between the link color of the profile and the sidebar color.

gender x sidebar_basic_color

Next up, we will be analyzing if there is a relationship between *gender* and *sidebar_color*.



As we can see, it seems that blue, black and white are all the most used colors. There is an interesting thing with these common colors which is that, for the *female* category, the *blue* color is significantly smaller compared to the other genders (while still being the most common). This could come from society's association of *blue* with more "masculine" roles or with brands. Additionally, this association of gender with color can also be observed with the color pink, which has an increased proportion for *female*. Again, this might be due to the association between pink and *femininity*. Additionally, we see that the colors gray, purple and yellow are the next most common colors. We can see how, for gender *unknown*, the proportion of users with yellow decreases.

Just as before, we can perform a chi-square test on these two variables. Here are the results:

```
> chisq.test(gender, sidebar_basic_color, simulate.p.value = TRUE)

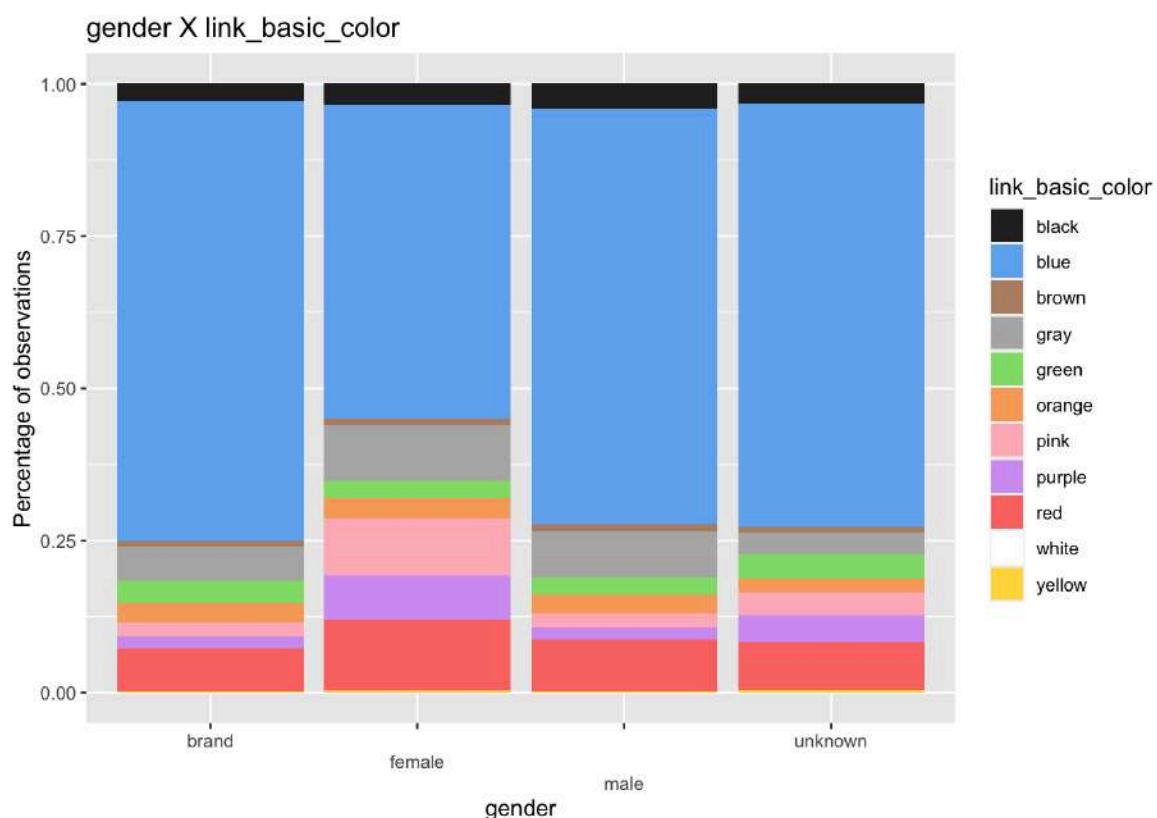
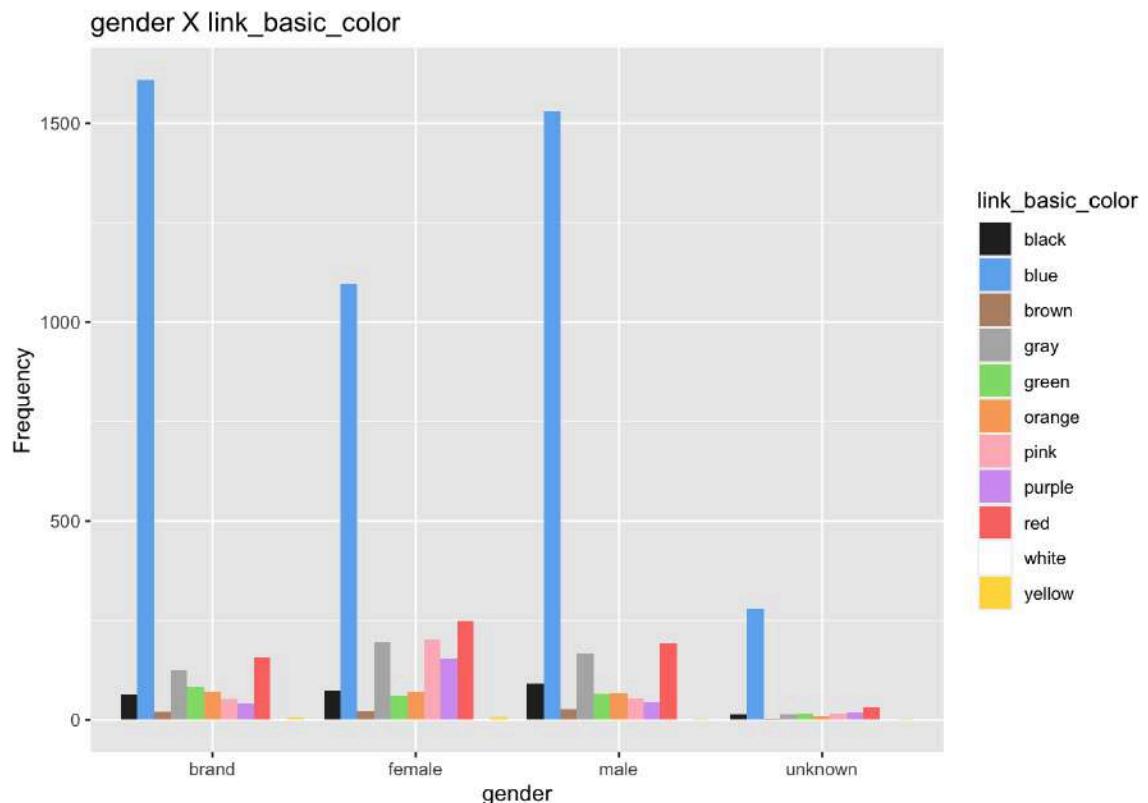
Pearson's Chi-squared test with simulated p-value (based on 2000
replicates)

data: gender and sidebar_basic_color
X-squared = 241.94, df = NA, p-value = 0.0004998
```

As the p-value is smaller than 0.05 (our significance level), we can reject the null hypothesis and confirm that there is an association between the predicted gender of a user and the color of the user's sidebar.

gender x link_basic_color

We can also analyze the relationship between *gender* and the other color variable, *link*:



As we can see, for this variable, the most common colors are blue, red, gray and purple. It is interesting to see how we can observe the same dynamics we found between the relationship with *gender* and *sidebar_basic_color*. For example, the proportion of the color blue also diminishes for users who have been classified as *females*, while the proportion of *pink* increases. It also seems that the proportion of purple and red increases too for the category *female*. As for the other gender categories, there are not many more interesting patterns, other than an increase of the proportion of *gray* for the *male* gender and of *green* for the *unknown* color.

Next up, let's perform the chi-square test:

```
> chisq.test(gender, link_basic_color, simulate.p.value = TRUE)
```

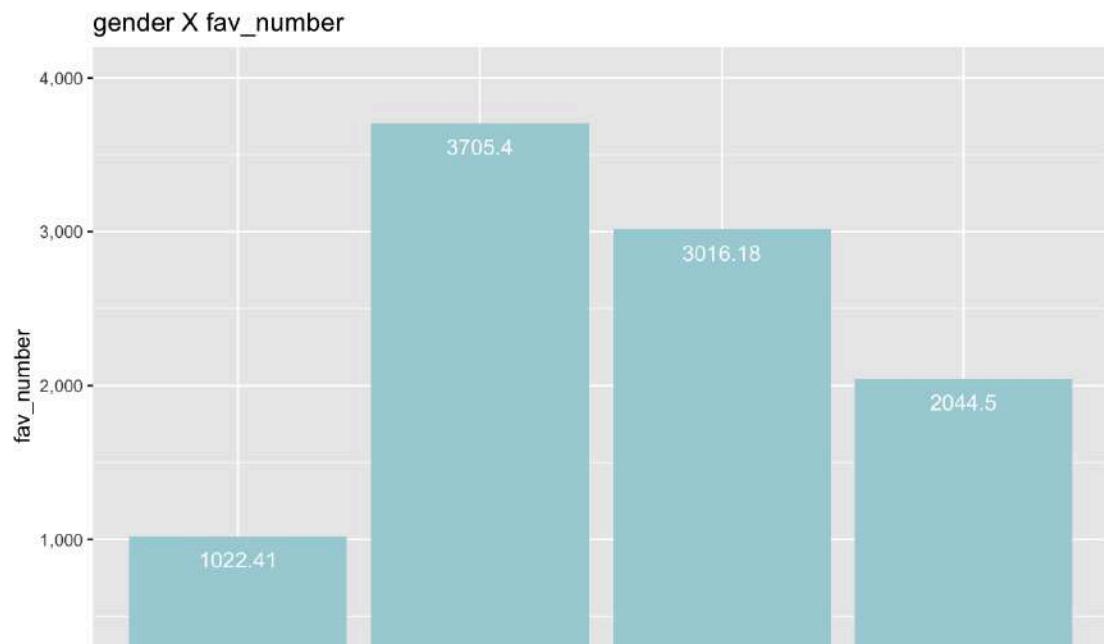
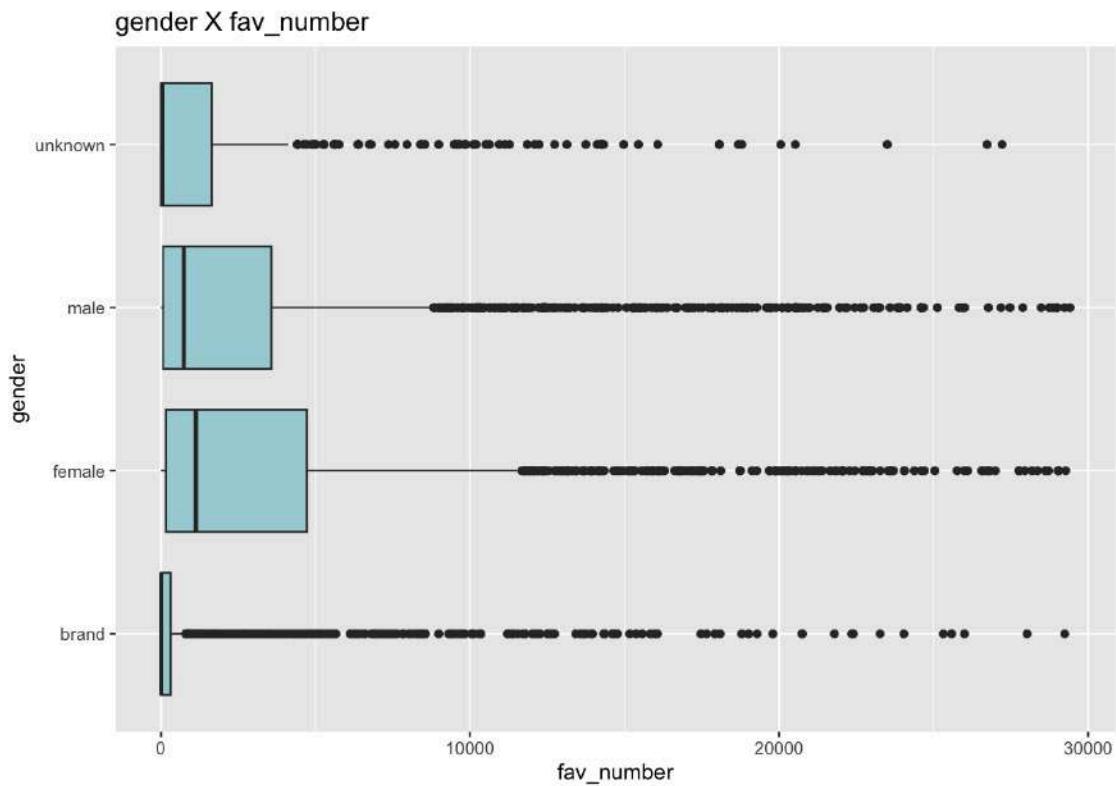
```
Pearson's Chi-squared test with simulated p-value (based on 2000  
replicates)
```

```
data: gender and link_basic_color  
X-squared = 420.61, df = NA, p-value = 0.0004998
```

As we can observe, the p-value is also below the significance threshold of 0.05. Therefore, we can reject the null hypothesis and conclude that there is a significant association between the predicted *gender* and *link color*.

gender x fav_number

Next up, we will be analyzing numerical variables with categoricals. First, it is the *gender* variable with *fav_number*. Hereunder is a boxplot of our variables:



As we can see, the gender *brand*'s interquartile range is the one with the lowest values. This makes sense, as a brand or company is not going to use Twitter as a form of entertainment where they will be liking tweets, but merely as a form of advertising. Next up, the one with the lowest interquartile range is *unknown*, followed by *male* and, lastly, *female*. This tells us that users who have more likes are more likely to be classified as *females*. Additionally, we can see that having fewer likes makes the account seem harder to classify when it comes to gender.

This can also be seen in the next table, which shows the *fav_number* average of each gender:

```
> centroide
   gender fav_number
1   brand  1022.411
2 female  3705.403
3   male   3016.180
4 unknown 2044.502
```

As we can see, the pattern is confirmed. It is interesting to see that the highest value (for female) and lowest (for brand), have a difference of 2700 likes.

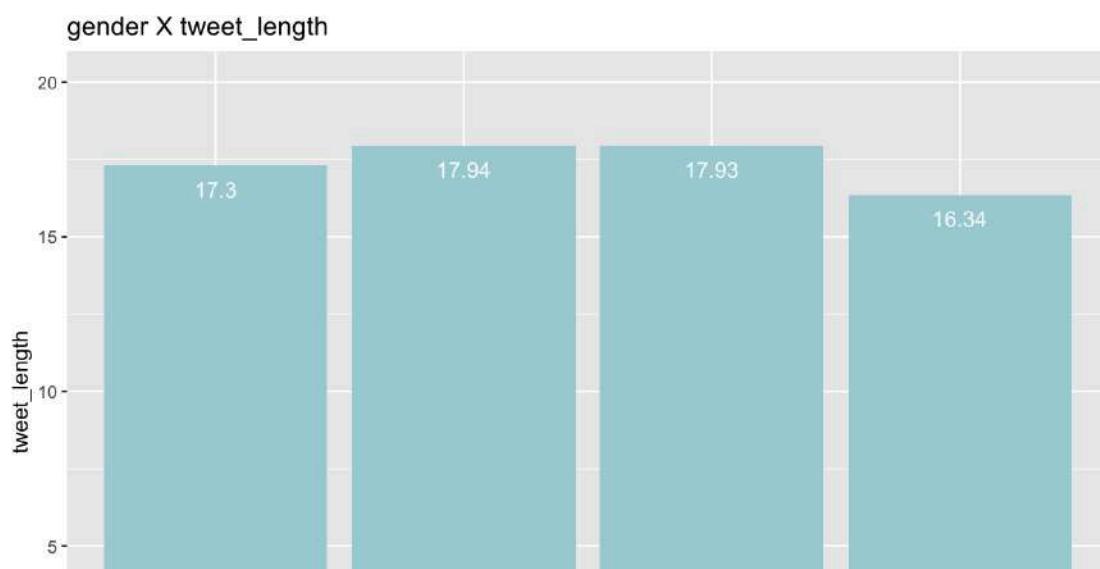
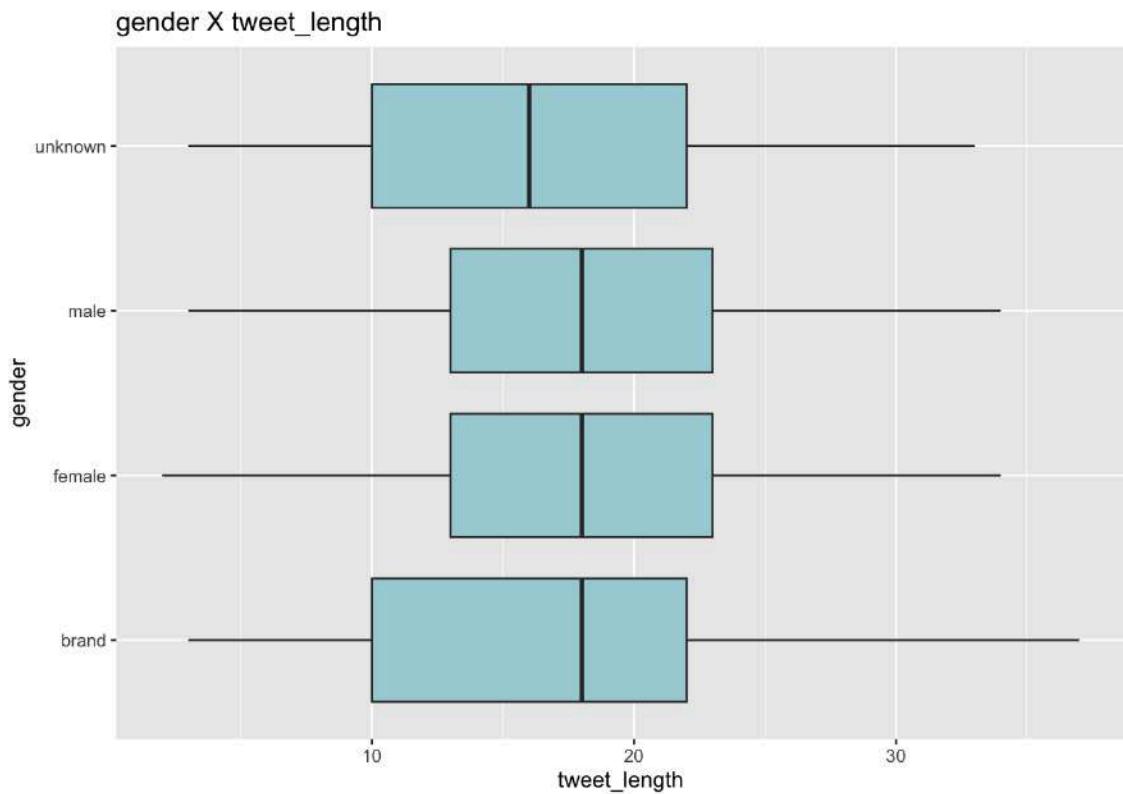
Next up are the values for the standard deviation:

```
> aggregate(fav_number, by=list(gender), FUN=sd)
  Group.1      x
1   brand 3071.352
2 female 5587.451
3   male  5145.537
4 unknown 4408.136
```

As we can see, the lowest value is for the brand. This tells us that, when it comes to the association of brands, we find the smallest range of values in which *fav_number* can be found. On the other hand, female and male are the highest.

gender x tweet_length

Next up, here are the boxplots and barplots of the variable *gender* and *tweet_length*:



For this next pair of variables, the difference does not seem as significant as *fav_number*. Nevertheless, we can find some valuable information: First off, *brand* and *unknown* have similar interquartile range, while the median is a bit higher for *brand*. This tells us that, while the 25 to 75% of users that have been classified as brands or unknown have the same tweet length (approximately 10 to 22), the 50% mark (or median) is higher for those who are classified as brands. When it comes to *male* and *female*, both have very similar median values (17.93). This tells us that the variable *tweet_length* does not significantly affect the prediction of a user being a man or a woman, but it does affect this distinction of a user being a person or an entity or unknown. This makes sense because, the longer the tweet, the more information you have about the user, and therefore the less likely you are to classify it as *unknown*.

We can also observe the average of each gender:

```
> centroide
  gender tweet_length
1   brand    17.30341
2 female    17.93941
3   male     17.93090
4 unknown   16.34500
```

As we can observe the values range from 16.3 to 17.9, with *unknown* being the lowest and *female* the highest (very close to *male*).

We can also look at the standard deviation to derive some conclusions:

```
> aggregate(tweet_word_count, by=list(gender), FUN=sd)
  Group.1      x
1   brand 5.903311
2 female 6.304357
3   male 6.201288
4 unknown 6.487387
```

Just as with *fav_number*, the category *brand* has the smallest standard deviation. It seems that when people think of a *brand* user, there is a clearer image or idea of what the profile looks like and behaves, compared to the other gender categories, which have higher standard deviations.

MCA - Multiple Correspondence Analysis

In this next section we will be performing Multiple Correspondence Analysis (MCA). MCA is a powerful statistical technique that allows us to analyze the relationship between multiple categorical variables. By mapping the categories of different variables onto a common set of axes, MCA helps us to identify patterns and relationships in complex datasets that may not be immediately apparent.

In our project, we will be using MCA to study the relationship between gender and several other categorical variables, including continent, privacy, and color. By examining the relationships between these variables, we hope to gain insights into how gender is related to other important dimensions of identity and experience.

The categorical variable of gender has been the subject of much study and debate, as it is a fundamental aspect of identity that shapes our experiences in a variety of ways. By using MCA to explore the relationship between gender and other categorical variables, we can gain a deeper understanding of the complex ways in which gender intersects with other dimensions of identity and experience.

Main Objective

Our main objective with this analysis is to see how our main variable, *gender*, is related to our other categorical variables, such as *privacy*, *continent*, *link color* and *sidebar color*. This information will tell us how people relate gender to other features of a person. For this reason, we are using these 5 variables as active variables to execute MCA, while using additional quantitative variables such as *fav_number* (how many tweets the user has liked), *tweet_count* (how many tweets this person has written), *tweet_word_count* (how many words the tweet from the user has), *tweet_avg_word_length* (the average length of the words of the user's tweet) and *gender.confidence* (the confidence of the people judging the user's gender on their prediction) as well as *created* (the time the user created the account, converted to a numerical value).

We did not use other quantitative variables such as *retweet_count* (how many tweets the user has retweeted), as it had a near zero variance.

Preparation

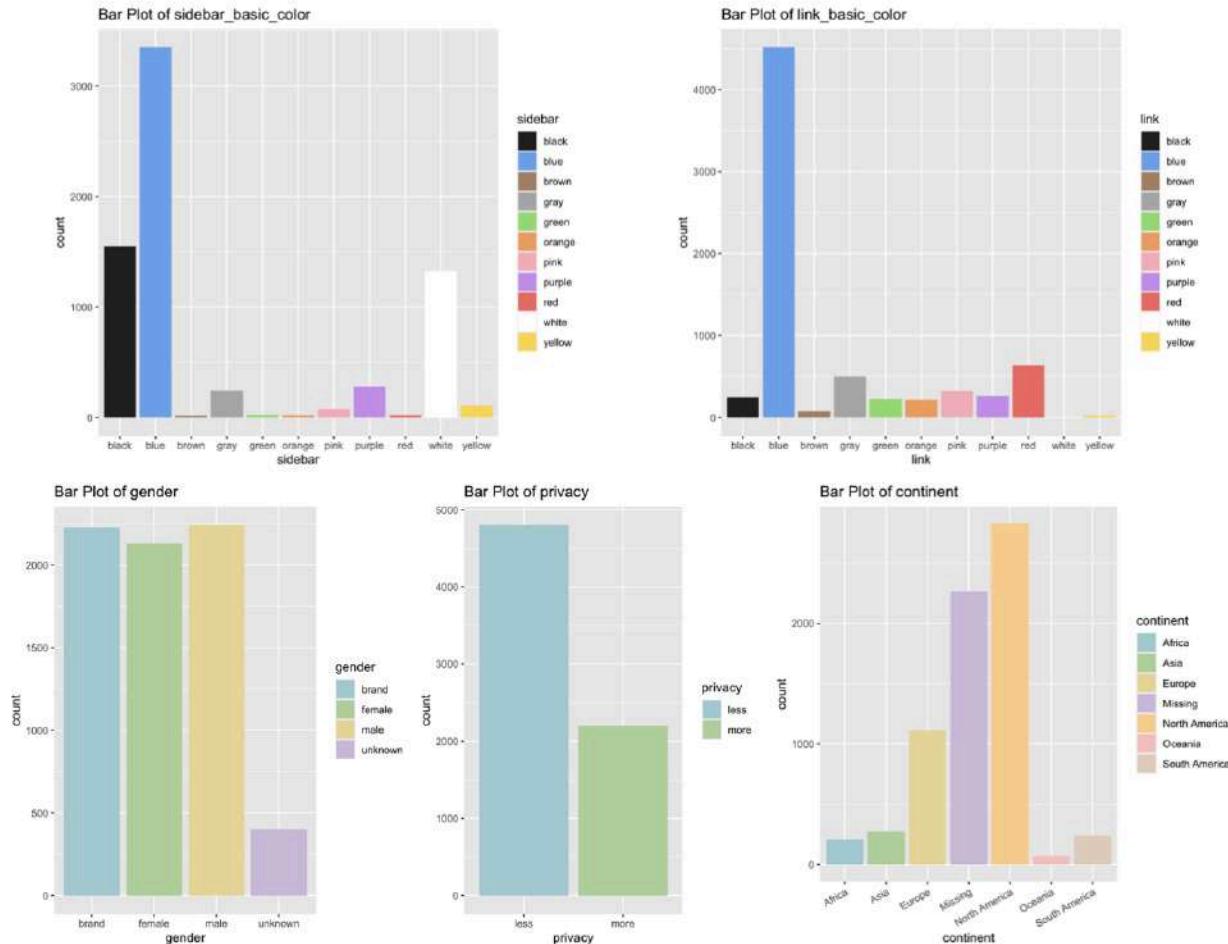
When performing MCA (Multiple Correspondence Analysis), it is important to check for infrequent modalities in the categorical variables included in the analysis. These infrequent modalities could potentially affect the MCA results, leading to incorrect interpretations of the data.

Infrequent modalities may result in a lack of variability within the data, which can cause the MCA to over-emphasize the importance of more frequent modalities, leading to a potential bias in the analysis. Moreover, if infrequent modalities are overlooked, important patterns or relationships within the data may be missed, which could lead to inaccurate interpretations of the MCA results.

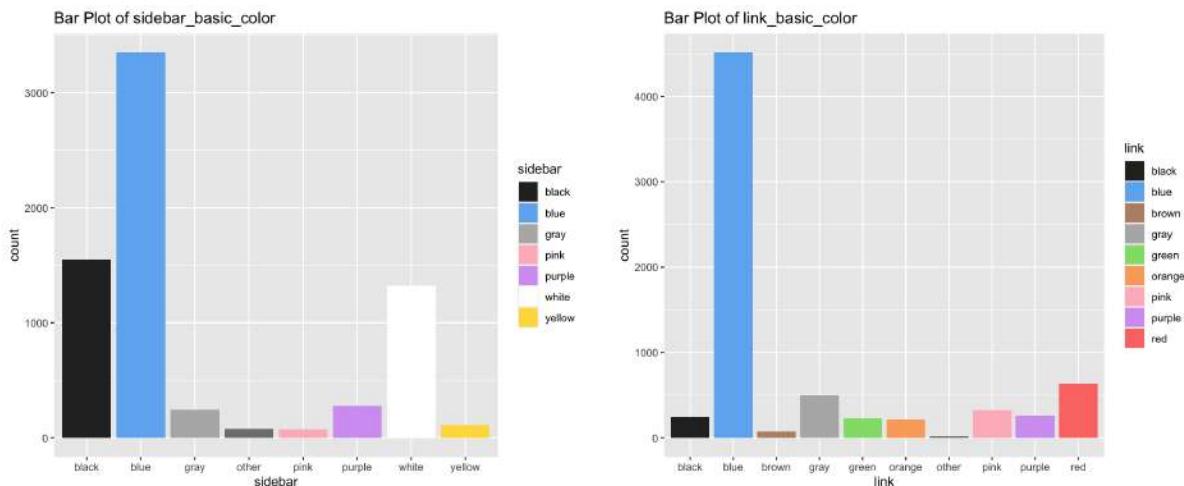
Therefore, it is important to identify infrequent modalities in the categorical variables before conducting the MCA, and to consider how they may impact the analysis. This can involve re-grouping or combining infrequent modalities with similar modalities to increase the overall sample size and variability within the data. Additionally, it is important to keep in mind the relative frequency of modalities and their contribution to the overall analysis, so that they can be appropriately interpreted in the context of the MCA results.

Hereunder is the barplot of each of the 5 qualitative variables we are using for MCA:

Bar Plot of Categorical Variables used for MCA



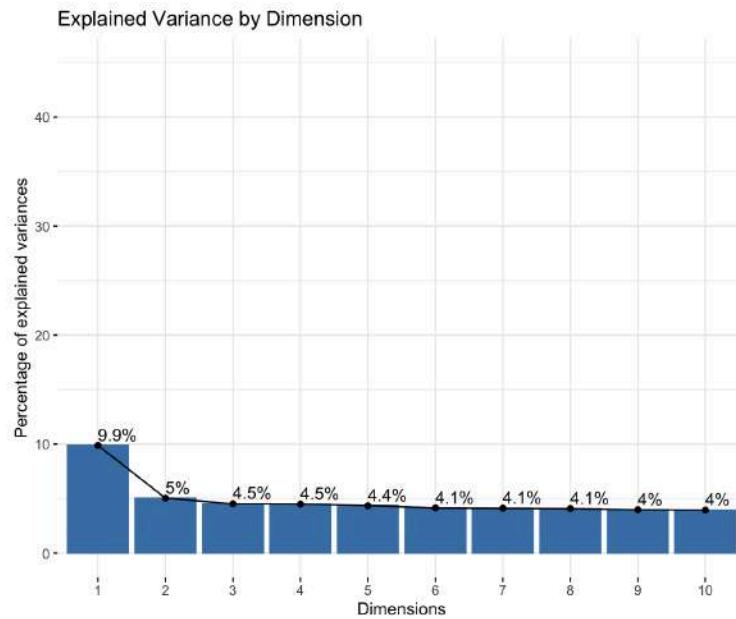
As can be seen from these plots, there are a few colors in both the *sidebar_basic_color* as well as *link_basic_color* that are very infrequent. For this reason, we will be grouping these colors (brown, green, orange and red for *sidebar*, and white and yellow for *link*) into a category named *Others*. Hereunder we have inserted the resulting bar plot after this transformation:



With this in mind, we executed MCA:

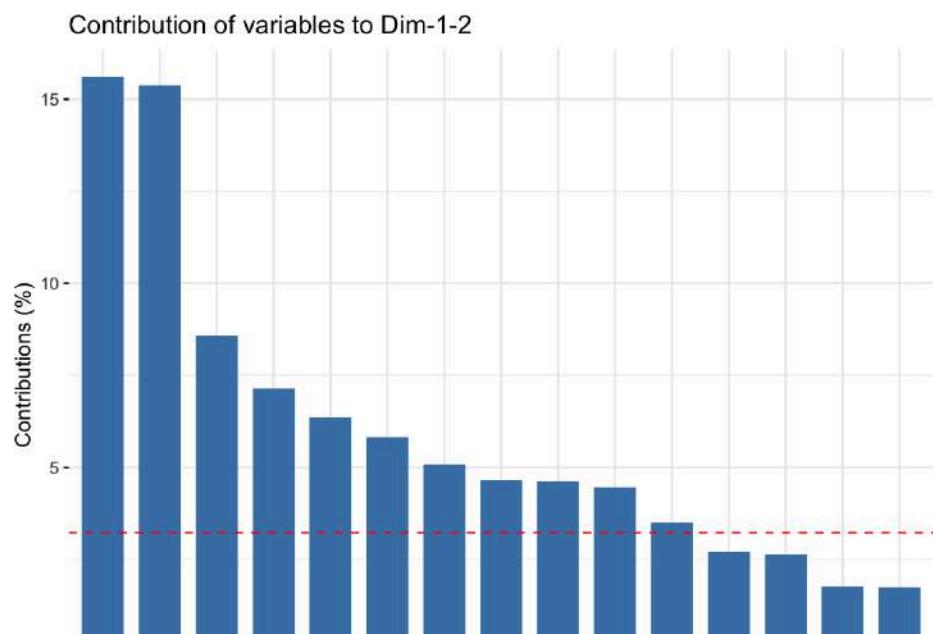
```
> mca <- MCA(data, ind.sup = active_index, quanti.sup = quanti_index,  
graph = FALSE, axes = c(1,2))
```

First off, we will look at the amount of total variance explained by each dimension:



The usual approach is to keep the dimension up until the “elbow point”, in which the slope becomes less pronounced. In our case, that means keeping the first two dimensions.

Additionally, we can see how much each modality contributes to the dimensions we have kept:

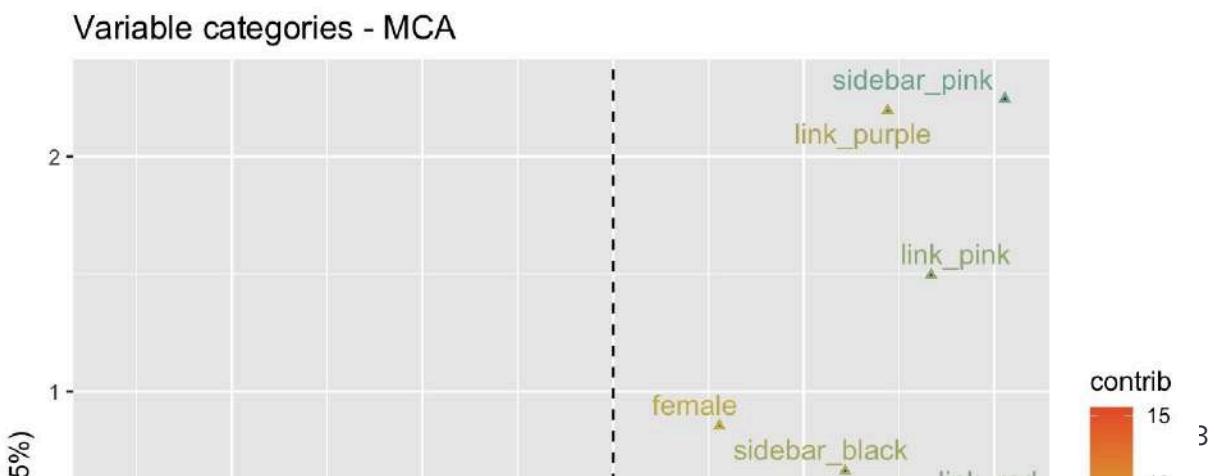


As we can see, both *privacy* modalities contribute a lot to the MCA dimensions. Additionally, both *Male* and *Female* contribute too. Additionally, colors such as blue, purple, pink and black contribute. Lastly, when it comes to continents, we can see how it is the variable that contributes the least, with the categories North America and Europe being the ones that contribute the most of this variable.

Analysis of Results

When analyzing the plot of the first two MCA dimensions there are several things that we will be looking for. First, we will examine the distribution of the categories of each variable and assess if any variable or category is strongly associated with each other. This can be observed by looking at the proximity of the categories to each other on the plot. Additionally, we are going to assess the variability in the data and identify the variables that contribute most to the structure of the plot, as they may be the most important in explaining the variability in the data. Finally, we will also examine any patterns or trends that may emerge from the plot, as they may provide valuable insights into the underlying relationships between the variables.

Let's start by including the resulting plot of the first two dimensions of the MCA (the plot is included in the annex of this report as a single landscape page):

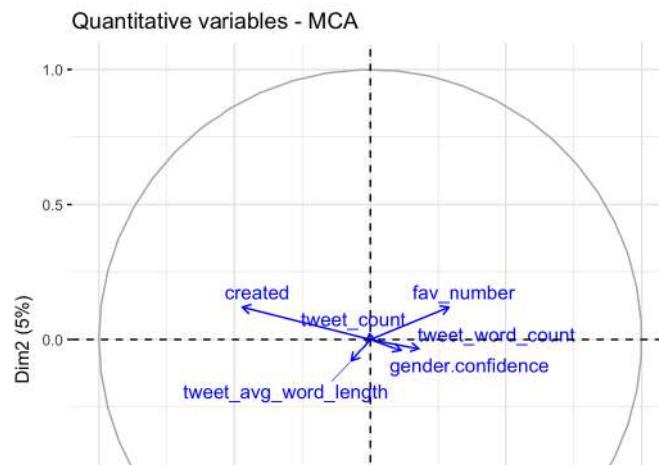


Here we have our variable's categories plotted on the first two dimensions of the MCA.

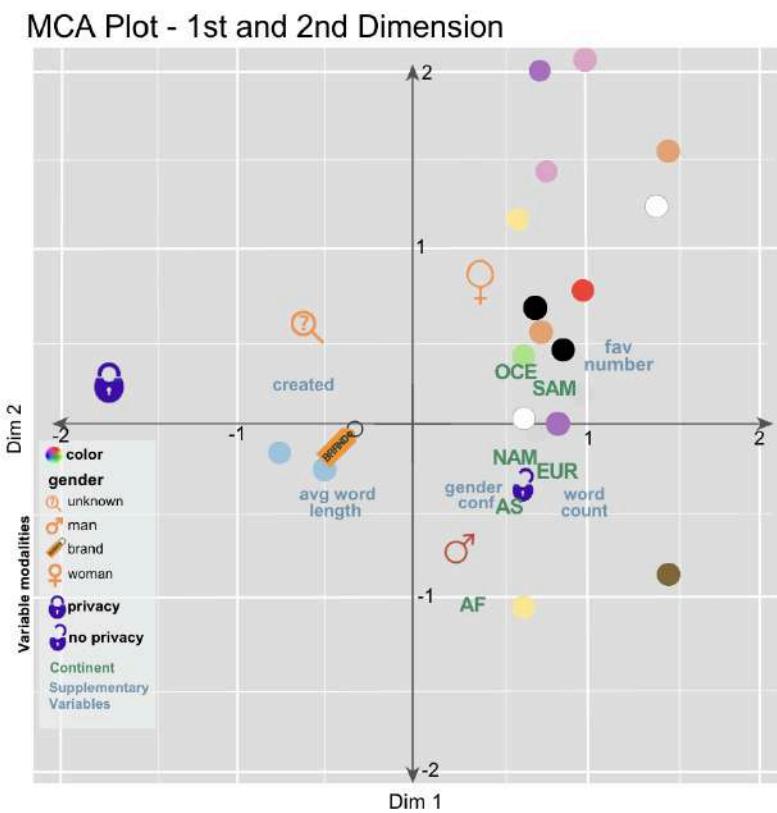
Additionally, they are coloured with the value of their contribution to the dimensions. We can see modalities such as *More* (from the variable privacy, meaning they are more private since they have chosen not to disclose their location information) as well as *Missing* have the highest contribution. The next modalities that contribute most to the MCA are purple, female, less (from the *Privacy* variable), and blue. Others are male, North America, black and pink.

Unfortunately, this plot is very hard to read as most modalities are found on the right of the axis and have long names. For this reason, we decided to make a visual representation of this plot by changing the modalities names by a symbol that represents it best. This will allow us to make a more thorough analysis of the MCA, as well as making it easier to draw conclusions by simply looking at the plot.

But before looking at that, we can also check the additional quantitative variables, which can be seen independently on this graph:



While this graph can be better interpreted, it still does not tell us much since it does not relate it to the active quantitative variables. For this reason, we also added these variables into our visualization of the MCA. Here under we can find the results:



As can be seen in the graph, we can observe the *color* modalities represented as color spheres, while the *gender* modalities are represented with orange symbols (explained on the lower-left legend) as well as the *privacy* variable represented with blue symbols. Additionally, the continent modalities are written with their abbreviations in green (NAM: North America, SAM: South America, OCE: Oceania, AS: Asia, EUR: Europe and AF: Africa). Lastly, supplementary variables are represented in purple.

Next up, we will be looking at this graph to draw conclusions about the relationship between these variables as well as the database as a whole.

Variability Explanation

First up, we can identify variables that contribute most to the structure. This can be done by looking at the distance between the variables and the origin. Variables that are farthest from the origin contribute most to the structure of the plot, indicating that they may be the most important in explaining the variability in the data. Additionally, it suggests that these variables have a stronger association with the dimensions captured by the plot, and may be the most important variables in explaining the patterns and relationships between the categorical variables in the data. In other words, these variables may be the most informative for understanding the underlying structure of the data and the relationships between the categories.

In our case, we find that the ones that are farthest from the origin are the colors pink and purple for the 1st dimension as well as the “more” level from the variable *privacy* for the 2nd dimension. The fact that the color pink is farthest from the origin in the 1st dimension suggests that users with a pink profile color are significantly different from users with other colors. This may indicate that users with pink profiles have unique characteristics or behaviors on Twitter. Furthermore, the fact that the *privacy* variable is farthest from the origin in the 2nd dimension suggests that users who are private about their location are significantly different from users who are not. This may indicate that the *privacy* setting has a strong impact on user behavior or characteristics.

Associations between variables

Next up, we are looking for associations between variables. For example, if there is a positive association between profile color and *privacy*, it would mean that users with certain profile colors are more likely to have certain privacy settings. Similarly, if there is a negative association between gender and continent, it would mean that certain values of these two variables tend to occur less often together than would be expected by chance.

This is because, when two modalities are close to each other, it means that they come hand in hand (if one is true, the other usually is true too). On the other hand, if a modality is in the opposite direction of the other and very distant, it means that they are usually not found together in observations.

We wrote in these tables a summary of the associations between gender and other variables, as well as color and the other variables:

Gender	POSITIVE ASSOCIATIONS	NEGATIVE ASSOCIATIONS
Male	Gender Confidence Word Count Africa Asia Yellow Blue Brown No Privacy	Created Unknown Pink Purple
Female	Pink Black Fav_number Oceania South America	Blue Avg_word_length Privacy
Brand	Blue Avg_word_length Privacy	Most colors Fav_number
Unknown	Created Privacy	Gender Confidence Word Count

Colors (only most relevant)	POSITIVE ASSOCIATIONS	NEGATIVE ASSOCIATIONS
Pink and Purple	Female	Brand
Blue	Brand	Female
Yellow + Brown	Male	Unknown
Most colors	Female	Brand Unknown

Let's analyze these results. First off, we can see that there is a positive association with male and the color green, yellow and brown. This tells us that when people see these colors on a user profile, they generally expect the profile to be a man's. Additionally, it is positively associated with gender confidence and word count. This means that people expect longer tweets from men, as well as meaning that when they guess that the user is a man, they are more confident about it (compared to when they guess it is a woman, brand...). Additionally, it is correlated with the continents Africa and Asia. It is also close to the *No Privacy* modality, meaning that predicted *male* users tend to be more open about their information.

It is also interesting to see a negative association with *Created* (the date the user's profile was created). This means that the older the account is, the more likely it is to be predicted as a man. Lastly, it is negatively correlated with Pink and Purple. This was expected, considering that pink and purple is generally viewed as a more "feminine" color by society.

Let's move on to the associations that the modality *Female* has. The most interesting part about this is the association with colors such as *Pink* and *Purple*, which as we talked about, have always been viewed as colors for women. This proves that this idea is still held nowadays. It is also positively associated with the continents Oceania and South America. Additionally, it is positively related to *fav_liked*. This means that people believe that users that have liked more tweets are generally women.

On the other hand, the modality *Female* is negatively associated with the color Blue. This tells us that if the color of the sidebar and/or link is Blue, people will most likely not guess that the user is a woman. Again, this brings us back to the stereotype of colors being gendered. Apart from pink and purple being viewed as more feminine, blue has always been the "masculine" color. This proves that this conception is still present in our society today. It is also interesting to see that it is negatively correlated with the average word length of a tweet. That means that tweets that have shorter words are generally expected to have come from women. This can be an explanation of the misogyny found in our society, which expects women to be the inferior gender. We believe this because lengthier words are generally related to higher intelligence since it proves the use of more complex (and therefore longer) words. Therefore, this might mean that there is still an association between women being less intellectually capable.

Next up, we have the modality *Brand*. We can see that it is positively associated with the color *Blue*, as well as with *avg_word_length*. This tells us that people generally associate brands with users with a blue profile and lengthier words. This might be interesting for brands or companies that want to appear more relatable and closer to people, because it can tell them that using colors such as blue and tweeting words that are longer will distance them from the general

population. Additionally, brand users are closer to being more private about their location than being open about it.

It is also important to add that it is negatively associated with *fav_number*, meaning that accounts that like fewer tweets are generally classified as brands.

Lastly, let's look at the *Unknown* gender. This can tell us what are the things that make a user harder to classify. In this case, it is the variables *created* and *Privacy* that are positively associated with this variable. This tells us that people who have not been identified as a gender are generally accounts that have been created recently and are more private (haven't disclosed information about their location).

Position in plot

This positioning of the continent variables on the right side of the y-axis could indicate that there is less variability in color and privacy preferences between users from different continents, as well as gender. Alternatively, it could suggest that the variation in color and privacy preferences is not as strongly related to the continent variable as it is to other variables.

This positioning of gender categories on different quadrants indicates that there are clear differences between the gender categories in terms of their preferences or characteristics related to the other variables in the analysis. This could suggest that the gender variable is an important factor in explaining the variability in the data and should be considered when examining the relationships between the other variables. Moreover, the fact that each gender category is plotted on a different quadrant may suggest that there are different patterns of association between the gender variable and the other variables in the analysis. Additionally, we can see how the modalities on the right refer to a *person*, while the variables on the left refer to an entity (brand or unknown). This tells us that the 2nd dimension explains if the user is an entity, while the 1st dimension explains whether it is a female or male, or a brand or unknown.

It is also interesting to see the gender that is plotted on the opposite quadrant of each category. For example, female and brand are opposite to each other, meaning that they generally have opposite profiles. On the other hand, male and brand are opposite to each other too.

Relationships between variables

Next up, we will look at the position that certain variables take in relation to each other.

Gender and Color: Most colors are plotted close to *Female*, while *Male* has fewer colors closer. Additionally, *Brand* is only close to blue and *Unknown* is very far from most colors. This pattern may suggest that there are differences in color preferences between genders. Users predicted as females may be more likely to choose a wide range of colors, which could explain why they are closely plotted with most colors in the MCA plot. On the other hand, males may be more likely to choose a narrower range of colors, which would explain why they are plotted with fewer

colors. Additionally, people with an unknown gender don't have a clear pattern when it comes to color.

In order to confirm this, we decided to perform a chi-square test, which is a statistical test used to determine whether there is a significant association between two categorical variables. The test compares the observed frequencies of the data with the expected frequencies under the assumption of independence between the two variables. Hereunder are the results obtained:

> <code>chi_squared_test(gender,link_color)</code>	> <code>chi_squared_test(gender,sidebar_color)</code>
<code>p-value = 6.789265e-72 < 0.05</code>	<code>p-value = 3.231741e-38 < 0.05</code>

Since the p-value for both the sidebar color and the link color is under our chosen significance level (0.05), we can reject the null hypothesis that these variables are independent from each other. This proves that Gender (or more precisely the prediction of *gender*) is related to both the sidebar and link profile.

Gender and Privacy: The brand or unknown gender category is closely plotted with the variable privacy, while the male and female are plotted with less privacy. This suggests that there may be a strong association between the privacy variable and the brand or unknown gender category. This pattern could mean that the brand or unknown gender category is more likely to be a user that values privacy more, compared to the male or female categories. Let's perform the chi-square test for these two variables next:

> <code>chi_squared_test(gender,privacy)</code>
<code>p-value = 1.36527e-47 > 0.05</code>

The results indicate that we can not reject the null hypothesis, meaning that we can not reject the idea that they are independent from each other. This tells us that there is no significant association between these two variables. We were not expecting these results, as the modalities brand and unknown were relatively closer to the "more privacy" setting, while Male and Female were closer to "less privacy". We believe that this means that although there might be an association, it is not strong enough to be statistically significant.

Gender and Continent: As we have seen, the *Continent* modalities do not seem to have a strong association with *Gender*. We have seen some continents being plotted closer with *Female* and some with *Male*, but it does not seem like a big association. We can perform the chi-square test to confirm this:

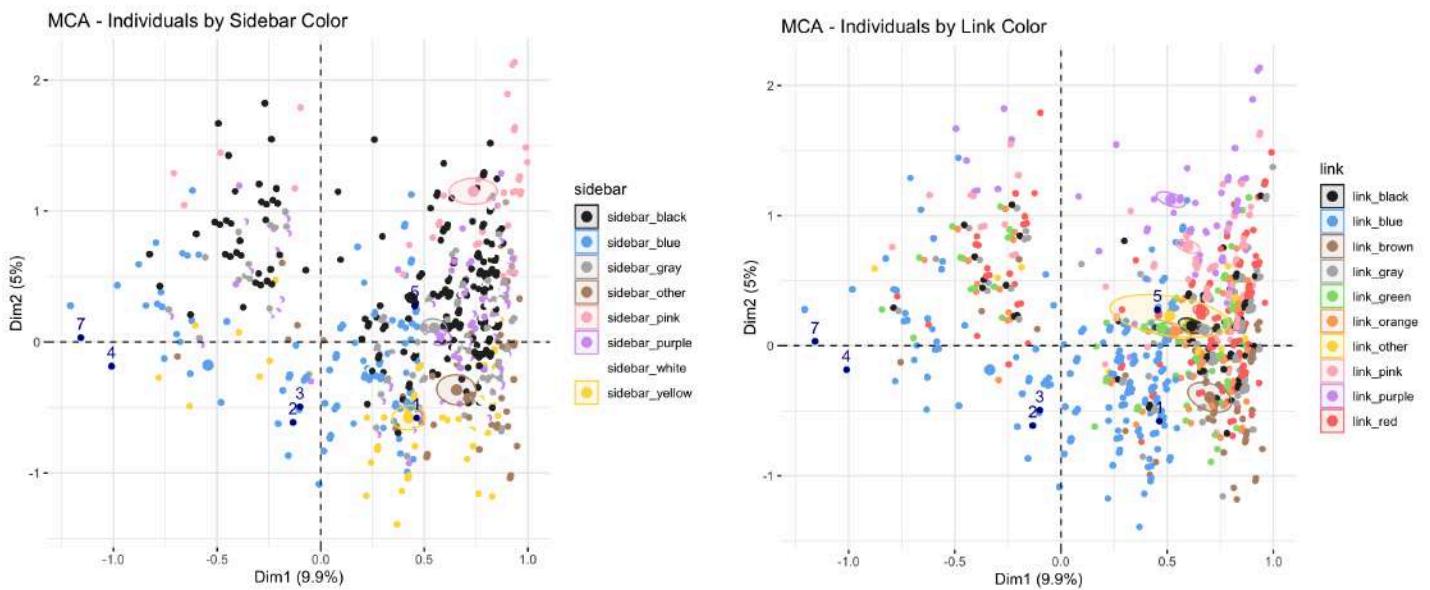
```
> chi_squared_test(gender,privacy)
```

```
p-value = 1.001209e-37 > 0.05
```

As can be seen, this test can not reject the null hypothesis of no association between variables.

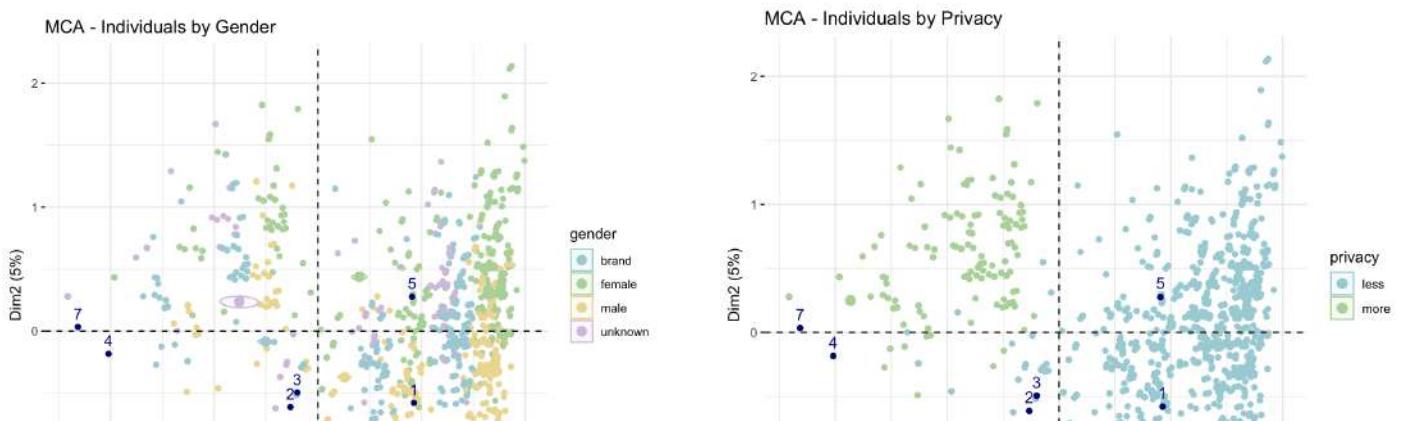
Individual Plots

Everything we have extracted from the MCA can also be looked at with these plots of the individuals (as opposed to the variables). For example, here are the individuals plotted on the dimensions, based on their link and sidebar color:



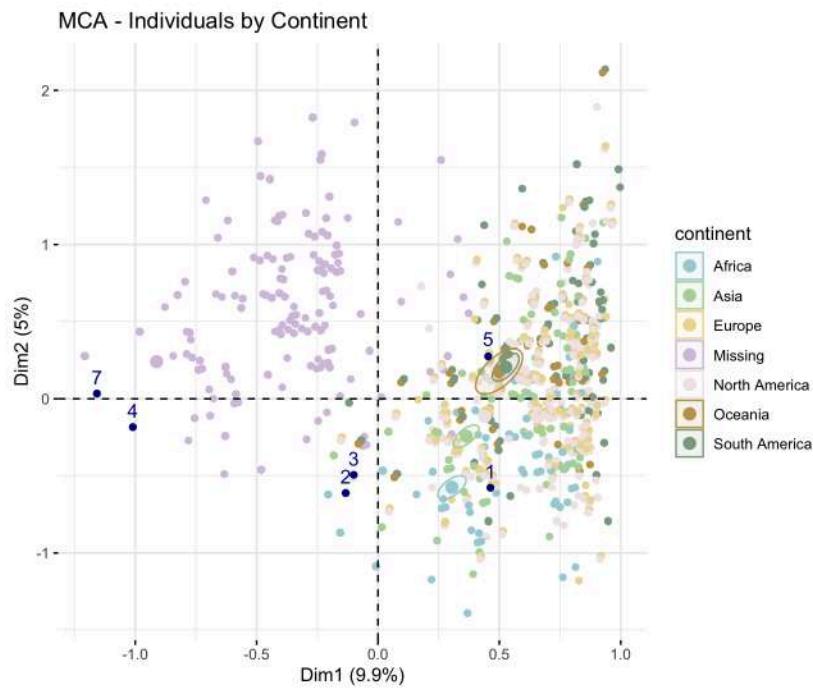
As we can see, most of the blue observations are found on the lower left quadrant and lower right, which is where most of the male and brand observations are found. Additionally, for the female quadrant (top-right) we find most of the purple and pink observations. Furthermore, the yellow is found on the lower part of the graph, specially on the right. Lastly, we can see how the quadrant where we find the unknown gender, there is no clear color pattern. This confirms the relationships we have established earlier between gender and color modalities.

Now we will check the plot of the individuals based on their gender and privacy settings:



As for these two plots, we find that the gender variable is not so clearly divided, having different smaller clusters for each modality instead of just one for each. As for the privacy variable, we can observe that it clearly divides the population in two big clusters: those that value privacy more and those that value it less.

Lastly, this is the plot with the observations coloured with their continent:



We can see how it divides the data into two big clusters: those that have missing and those that don't. Additionally, we can see that most Africa observations are found on the lower right, while Oceania and South America are on the upper right.

Summary

Hereunder we have summarized the whole analysis in this table, comparing the conclusions we have extracted from the MCA:

	Male	Female	Brand	Unknown
Color				No strong association
Tweet	Long tweets	Short words	Long words	Short tweets
Profile	Old profile	High number of liked tweets	Low number of liked tweets	Recently new profile
Privacy	Low	Medium	Medium-High	High
Gender Confidence	High	Medium	Medium-Low	Low
Continent	Asia Africa Europe North America	Oceania South America Europe North America	No strong association	No strong association
Opposite Profile	Unknown	Brand	Female	Male

TIME-SERIES-CLUSTERING

In this section of the study, we focus on the analysis of the "Twitter User Gender Classification" database, which contains information about Twitter users, such as account creation dates and confidence of the collaborators in determining the user's gender. From this database, we have formulated two primary hypotheses:

- Hypothesis 1: It is possible that individuals who created their Twitter accounts during holiday periods are more difficult to determine their gender by collaborators compared to those who registered at other times of the year.
- Hypothesis 2: It may be easier for collaborators to determine the gender of older Twitter users compared to newer users.

To investigate these hypotheses, we employed the Time-Series-Clustering method. This technique is well-suited for analyzing temporal data and uncovering hidden patterns in time series by effectively grouping data based on temporal similarities. Time-Series-Clustering is an approach that analyzes and clusters time series according to their similarities. The objective of this method is to partition the dataset into groups (clusters) such that the time series within each group are as similar as possible to each other and as distinct as possible from the time series in other groups. This method is valuable for identifying hidden patterns and trends in data over time, which facilitates a more comprehensive understanding and analysis of the phenomena under study.

In our case, Time-Series-Clustering is a suitable method for addressing our hypotheses since it enables us to identify and group Twitter users based on the evolution of the collaborator's confidence in determining the user's gender over time. By implementing this method, we can determine if there are substantial differences in collaborator's confidence among various user groups, allowing us to evaluate the validity of our hypotheses.

Preliminary Steps

Before delving into the analysis of the proposed hypotheses, several preliminary steps were required to prepare and clean the data. Initially, the "df\$created" variable was converted into a date data type to facilitate the temporal analysis of collaborator's confidence in determining user's gender based on the account creation dates of Twitter users.

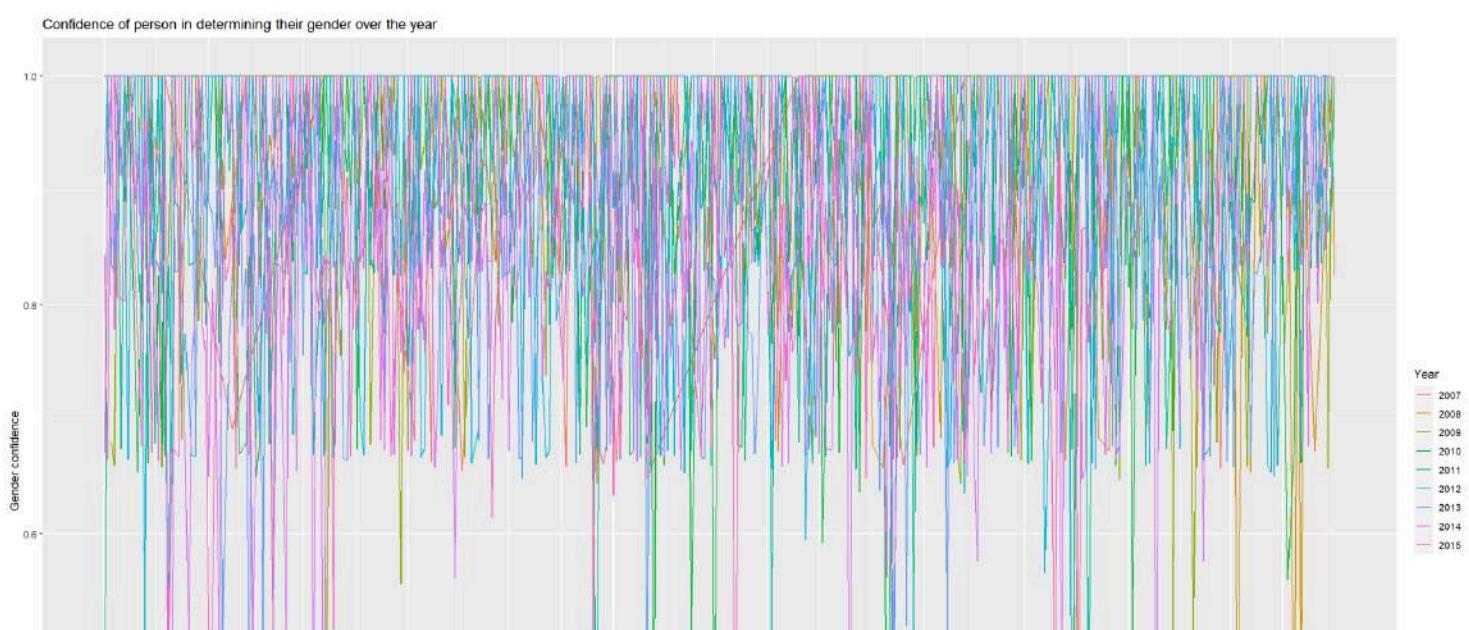
Upon executing this conversion, it was discovered that only one value in the database corresponded to the year 2006. Given that this single data point is insufficient to yield meaningful information and could potentially cause issues when applying Time-Series-Clustering, the decision was made to remove this record from the year 2006. By doing so, we ensured that the analysis relies on more consistent and representative data reflecting the temporal trends in collaborator's confidence in determining user's gender.

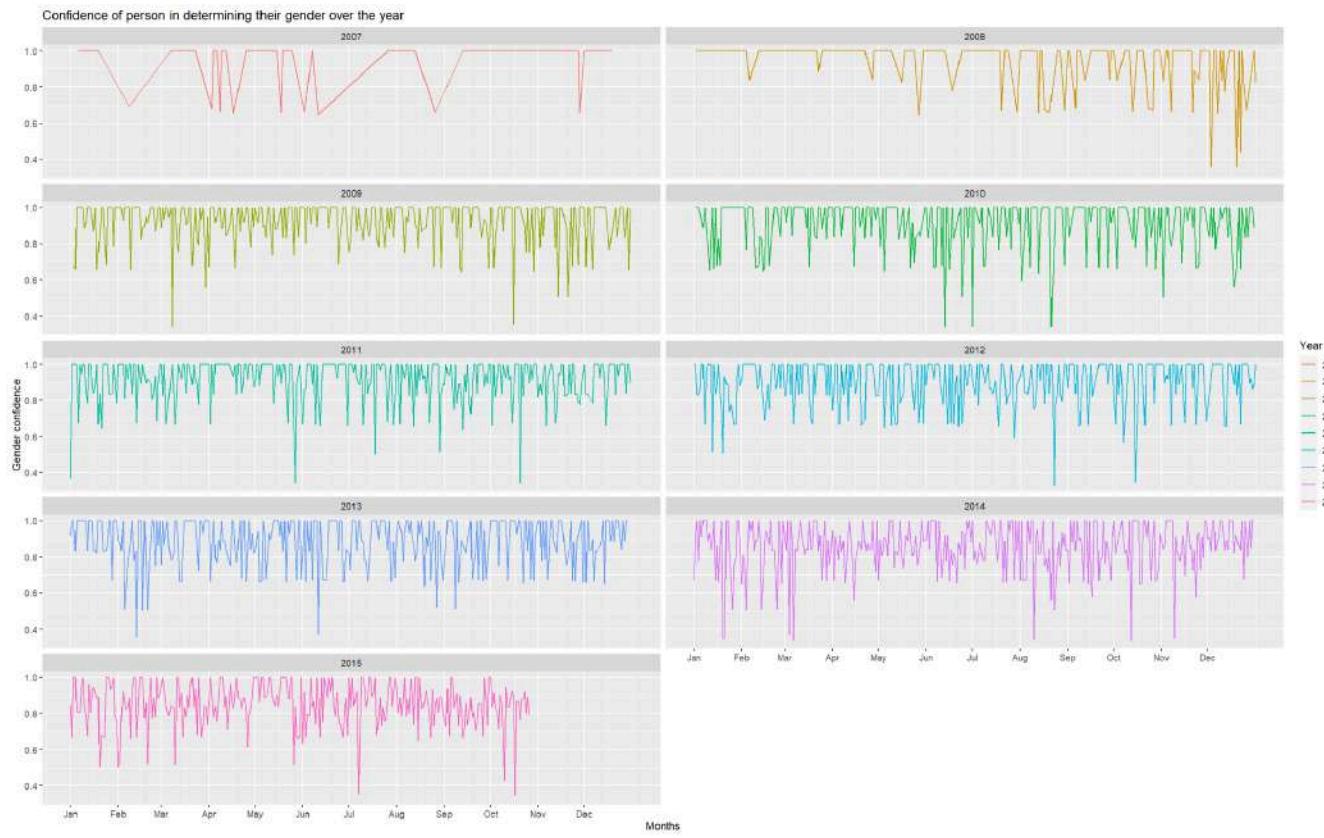
Analysis of Hypothesis 1

To evaluate Hypothesis 1, which posits that it might be more difficult to determine the gender of individuals who establish a Twitter account during holiday seasons compared to those who registered at other times of the year, we opted to generate a visual representation by creating a graph to illustrate the progression of collaborator's confidence in determining user's gender over time.

Visualization of Collaborator's Confidence in Determining User's Gender Throughout the Year

Initially, we created a graph that exhibits all the years collectively, with distinct color-coded lines representing each year. This allows for the observation of the general progression of collaborator's confidence in determining user's gender over time.





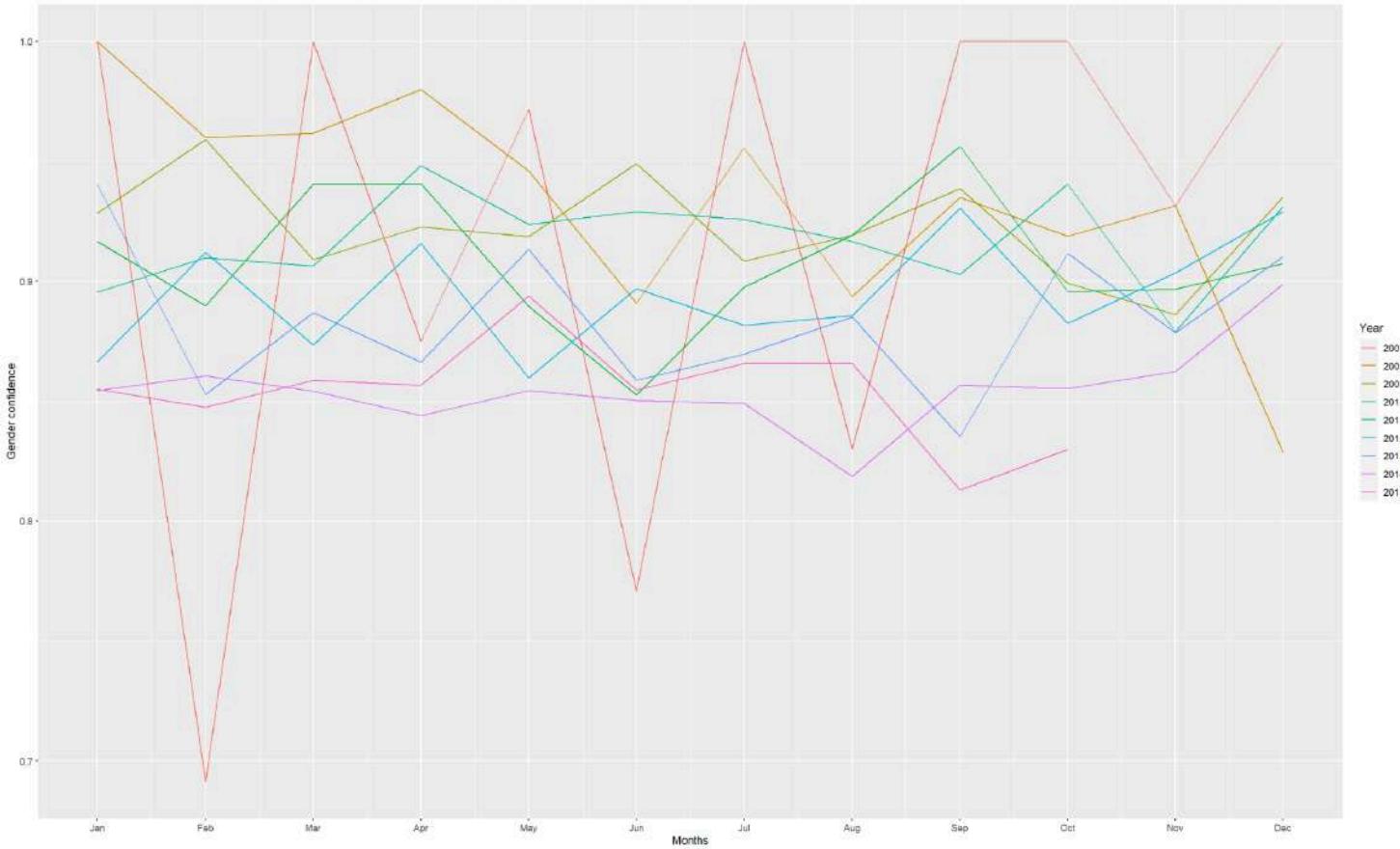
Next, to enhance data visualization and enable a more straightforward comparison between the years, we devised a graph using the "facet_wrap" function, which separates each year into individual panels. This clearly displays the behavior of collaborator's confidence in determining user's gender throughout the year for each of the years present in the dataset.

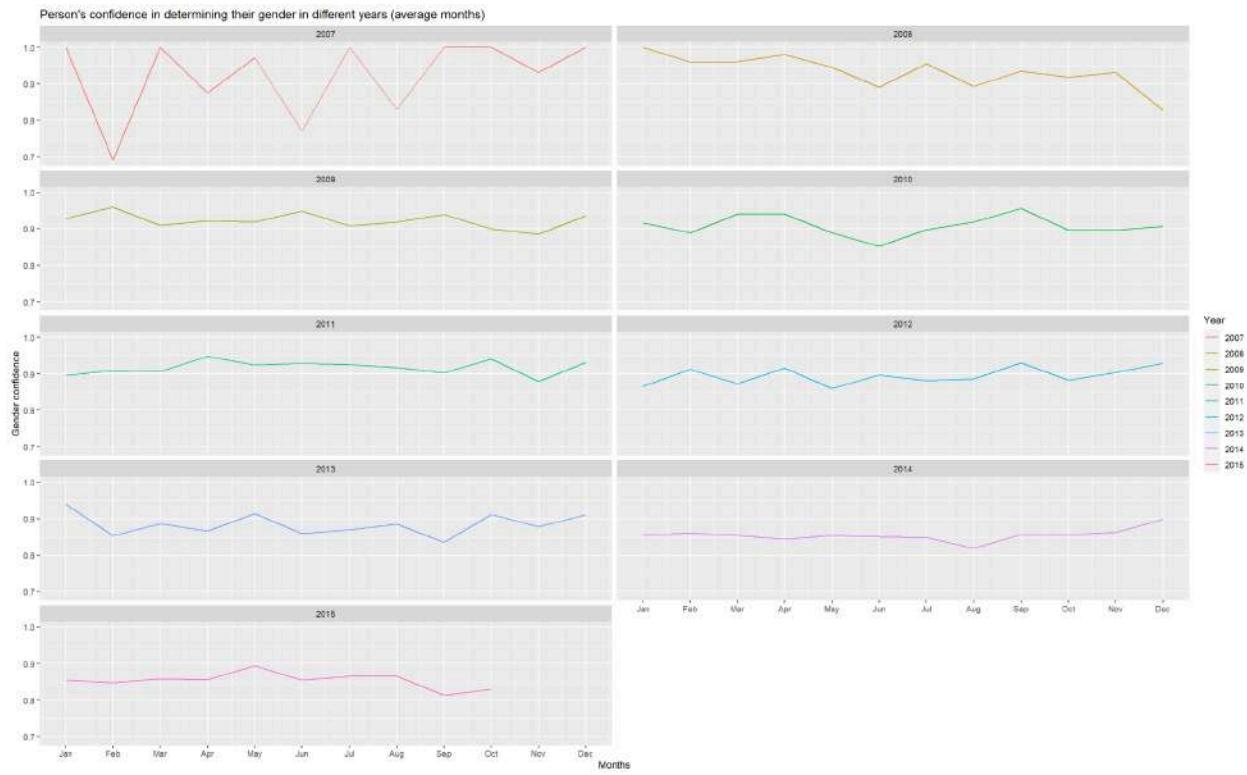
Next, to enhance data visualization and enable a more straightforward comparison between the years, we devised a graph using the "facet_wrap" function, which separates each year into individual panels. This clearly displays the behavior of collaborator's confidence in determining user's gender throughout the year for each of the years present in the dataset.

Visualization of Collaborator's Confidence in Determining User's Gender Grouped by Months

Since we could not reach a clear conclusion in the previous section when analyzing collaborator's confidence in determining user's gender throughout the year, we have decided to modify our approach and analyze collaborator's confidence grouped by months instead of by days. This will allow us to observe if there is any clearer pattern or trend based on the months of the year.

Person's confidence in determining their gender in different years (average months)





Just like in the previous analysis, we generated two types of graphs: one that shows all the years together and another that separates each year into a different panel using the "facet_wrap" function. Both graphs show the average collaborator's confidence in determining user's gender for each month instead of each day, which allows us to simplify the visualization and facilitate the identification of trends.

After analyzing the graphs grouped by months, we can observe that, although there are some variations in collaborator's confidence in determining user's gender throughout the months, no clear pattern is identified that would allow us to confirm or refute Hypothesis 1. There is no consistent trend based on vacation periods that suggests that it is more difficult to determine the gender of individuals who join Twitter during those periods.

Based on the results presented, we can draw some additional conclusions about the relationship between collaborator's confidence in determining user's gender and the date of creation of their Twitter accounts:

- During the first year of Twitter's existence (2007), a greater variability in the average collaborator's confidence in determining user's gender is observed, oscillating between 0.69 and 1.0. Starting from 2008, the average collaborator's confidence stabilizes in a

narrower range between approximately 0.8 and 0.95.

- Throughout the years, there does not seem to be a clear trend in the average collaborator's confidence in determining user's gender based on the month of account creation. In general, the average collaborator's confidence varies throughout the months, but no specific pattern is identified.
- Between 2007 and 2010, there are months with a perfect average collaborator's confidence in determining user's gender (1.0), which indicates that some individuals had absolute certainty about their gender during those periods. However, starting from 2011, the average collaborator's confidence does not reach the value of 1.0 again, which suggests a change in the perception or expression of gender identity on the platform.

In summary, the results suggest that collaborator's confidence in determining user's gender based on the date of creation of the Twitter account has undergone changes over time, with greater variability in the early years. No clear pattern is identified based on the month of account creation.

Analysis of Hypothesis 2

Clustering

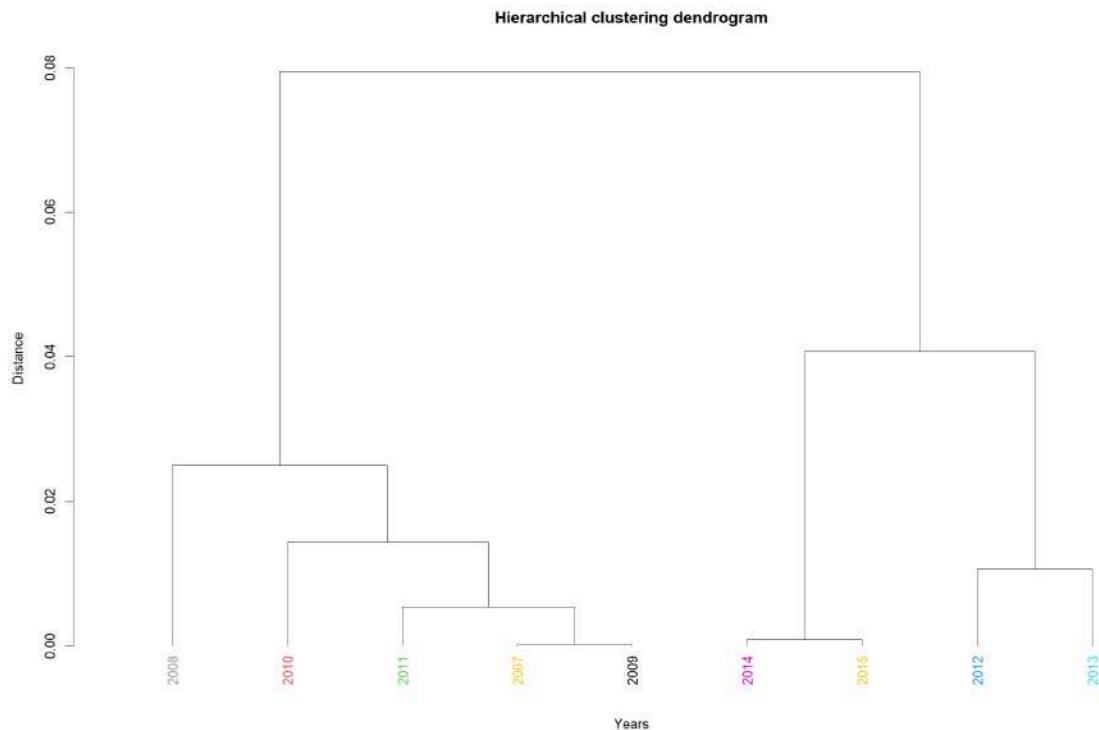
For the clustering analysis, we have chosen to use the average collaborator's confidence values per month rather than the average values per day. This decision is informed by several factors that make the monthly analysis more suitable for testing Hypothesis 2: "Perhaps it is easier to determine the gender of older Twitter users compared to newer users."

By aggregating the data monthly, we reduce noise in the data, making it easier to identify trends and patterns. Moreover, working with monthly values enhances the efficiency of the clustering analysis.

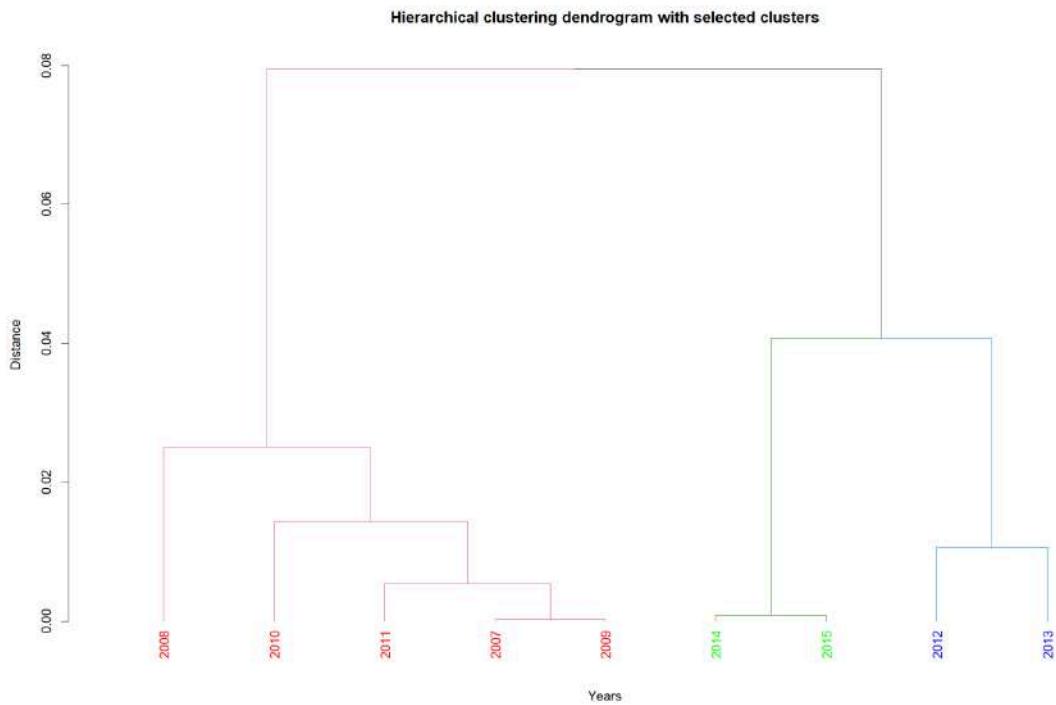
As part of our analysis, we have grouped clusters using years based on average collaborator's confidence in determining user's gender. This enables us to assess whether a correlation exists between users' age on the platform and collaborator's confidence in determining their gender.

In our clustering analysis, we have employed Euclidean distance as a measure of dissimilarity between data points. Euclidean distance is one of the most prevalent metrics for calculating the distance between two points in a multidimensional space, defined as the length of the shortest segment connecting two points.

Additionally, we have utilized hierarchical clustering with the complete linkage method to group years based on average collaborator's confidence in determining user's gender. The complete linkage method is an agglomerative clustering approach that merges clusters according to the maximum distance between all pairs of observations in two distinct groups. This method tends to generate more compact and balanced clusters compared to alternative linkage approaches.



To determine the optimal number of clusters, we created a dendrogram using the results of the hierarchical clustering. The dendrogram is a visual representation of the hierarchical structure of the clusters, with the height of the branches indicating the distance between the groups.



Upon examining the generated dendrogram, we can observe that the optimal number of clusters for our data is 3. The groups are divided as follows:

Cluster 1: 2007, 2008, 2009, 2010, and 2011

Cluster 2: 2012 and 2013

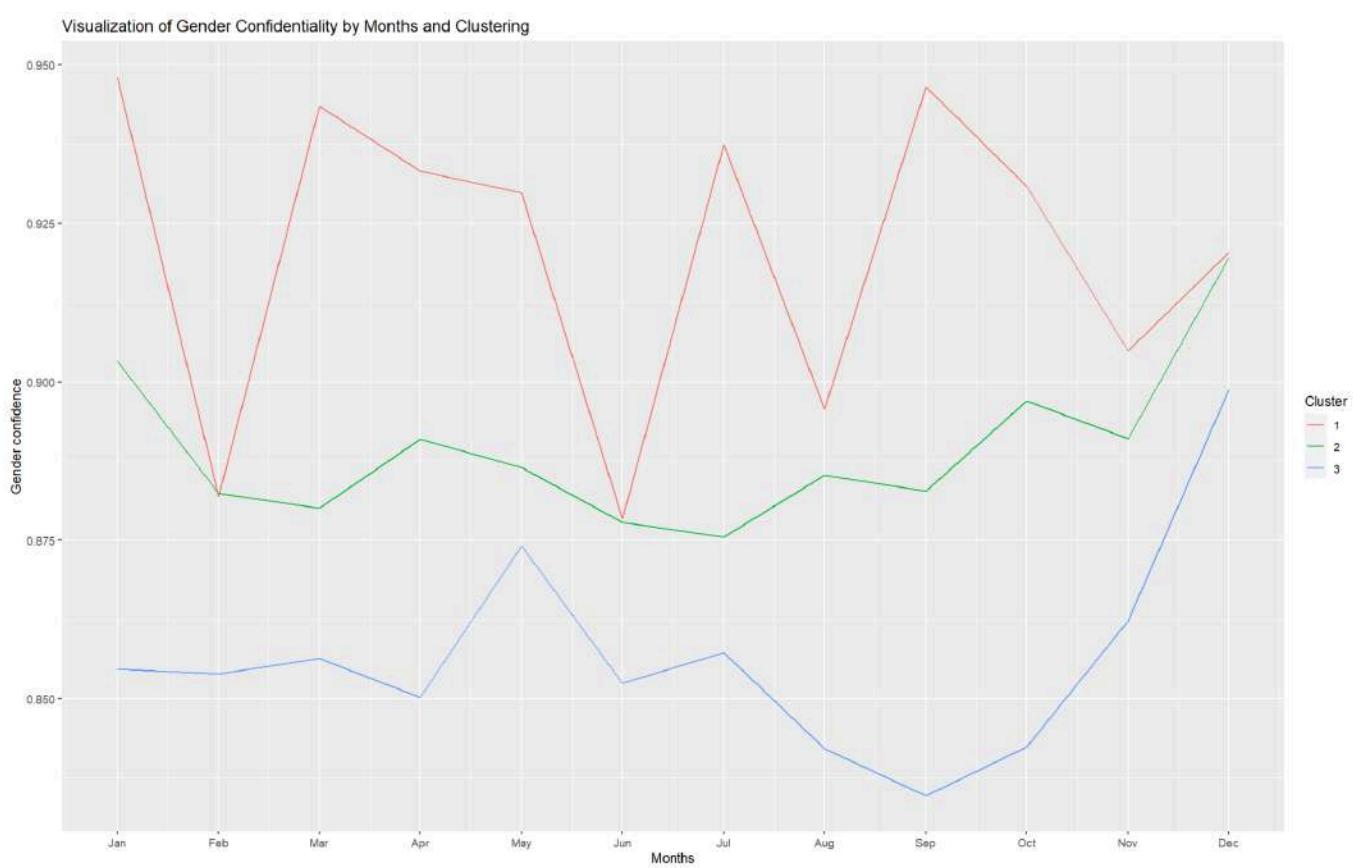
Cluster 3: 2014 and 2015

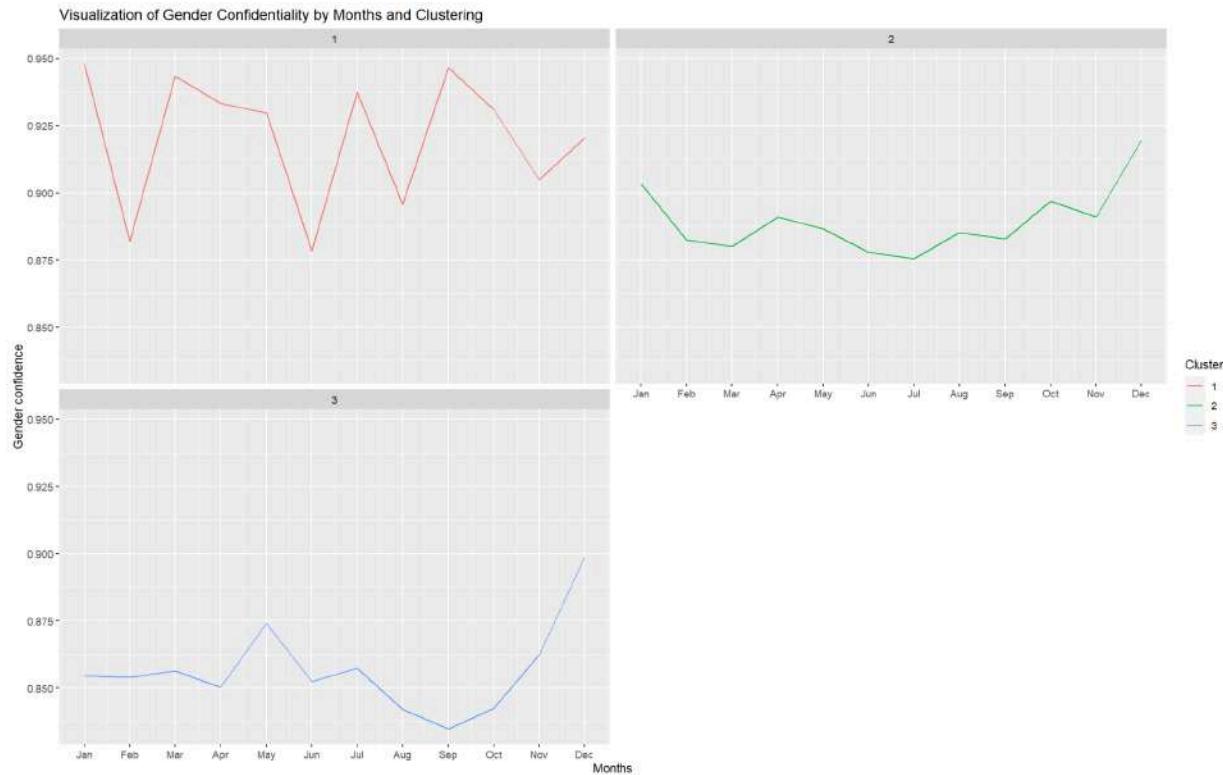
Interestingly, the clusters have formed with consecutive years, suggesting a trend in the ease of determining user gender over time. We can infer that the ability to determine user gender has experienced changes over the years on Twitter, and these groups represent different levels of certainty in gender identification among the collaborators.

The clustering analysis offers a clearer perspective on how the confidence in determining users' gender has evolved on Twitter over the years. The grouping of years into consecutive clusters implies that there have been changes in the certainty with which collaborators can identify users' gender, which could be related to the age of users' Twitter accounts and their experience on the platform. This finding partially supports Hypothesis 2, proposing that it may be easier to determine the gender of users with older Twitter accounts compared to those with newer accounts.

Visualization of Collaborator's Confidence in Determining User's Gender Grouped by Months and Clustering

After classifying the data into clusters, we proceed to calculate the average confidence in determining users' gender for each month and cluster, allowing us to represent a single line per cluster in the graph.





We have used the same graph format used in the section "Analysis of Hypothesis 1" to analyze the results. Two graphs are generated: one with a single line per cluster that shows the evolution of collaborator's confidence in determining user's gender over the months, and another that divides the data into separate panels for each cluster using facet_wrap. Both visualizations allow for comparing the trends in collaborator's confidence in each cluster and could support our Hypothesis 2 about the relationship between the age of users' Twitter accounts and the ease of determining their gender.

After analyzing the results obtained in the visualization of collaborator's confidence by month and clustering, we can confirm our initial hypothesis that it is easier to determine the gender of users with older Twitter accounts compared to those with newer accounts. This can be clearly observed in the trend of the mean values of collaborator's confidence of the clusters, where cluster 1 (2007-2011) has the highest values, followed by cluster 2 (2012-2013), and finally cluster 3 (2014-2015). We observe that collaborator's confidence tends to decrease with new users.

From the presented results, we can draw some additional conclusions, as we have previously done:

In Cluster 1, the average collaborator's confidence fluctuates within a range of approximately 0.88 to 0.95. There is no specific pattern observed based on the month of account creation. Overall, it can be inferred that collaborator's confidence in Cluster 1 was more stable and higher compared to the other two clusters.

In Cluster 2, the average collaborator's confidence ranges between approximately 0.88 and 0.92, indicating a decrease compared to Cluster 1. This could be the result of increasing diversity in the perception or expression of gender identity on the platform during that period.

Cluster 3 exhibits an even greater decrease in average collaborator's confidence, with values primarily ranging between 0.83 and 0.90. This could indicate that in more recent years, the confidence of collaborators in determining users' gender on Twitter has declined or become more diverse.

The results suggest that collaborator's confidence in determining users' gender based on the date of Twitter account creation and clusters has undergone changes over time. The average collaborator's confidence was higher in the earlier years (Cluster 1) and decreased in the subsequent clusters (2012-2015). No clear pattern is identified based on the month of account creation within each cluster.

Summary

In this study, we employed Time-Series-Clustering to analyze the "Twitter User Gender Classification" database in order to address two hypotheses related to collaborator's confidence in determining users' gender and the timing of account creation on the platform. The main findings of our analysis are as follows:

- Hypothesis 1: The analysis does not support the hypothesis that it is more difficult to determine the gender of individuals who created their Twitter accounts during holiday periods compared to those who registered at other times of the year. No discernible pattern or trend based on vacation periods was identified in the data.

-
- Hypothesis 2: Our analysis partially supports the hypothesis that it is easier to determine the gender of users with older Twitter accounts compared to those with newer accounts. The clustering analysis revealed that the confidence of collaborators in determining users' gender has evolved over the years on Twitter, with users who created their accounts from 2007 to 2011 showing higher confidence on average compared to users who created their accounts from 2012 to 2015.

The results of this study contribute to a better understanding of how the ease of determining users' gender has changed over time on the Twitter platform. These findings can be utilized to inform further research on the factors that influence the confidence of collaborators in determining users' gender and the role of social media platforms in shaping and reflecting these identities.

CLUSTERING

In the following section we will be exploring several clustering methods. We will take a look at CURE, DBSCAN and OPTICS clustering. For each section we will later perform advanced profiling on the clusters generated and try to determine the quality of the clustering by this.

CURE

CURE or Clustering Using Representatives is an algorithm designed to solve some of the challenges of traditional clustering algorithms which can be sensitive to the initial selection of

cluster centers, like k-means, and do not tend to work well with large datasets, since the computational cost may be high, such as hierarchical clustering.

The CURE clustering technique works by first selecting a smaller sample from the dataset, which will be representative points from the dataset, which are called *medoids*. The medoids are chosen by randomly sampling a small fragment of the original dataset.

The following step is to cluster this sample into k clusters, in our code we use hierarchical clustering. Once the clustering is done, we must find the center of each cluster. After this, for each cluster we will select a certain amount of medoids belonging to said cluster that are closest to the initial center calculated, we will use this new subset to update the center of each cluster. This is a way to assure that the center is more robust. The number of medoids used to update the center is up to the user to decide, but it is usually determined by a percentage r of the total number of medoids in the cluster.

The final step is to assign the remaining data into the clusters by finding the closest center for each point, in this way we are clustering all the remaining points just by calculating k distances per point.

The main advantage of CURE clustering is that it can be used with large datasets to save computing time, since the number of medoids tends to be much smaller than the number of data points and the most complex part, the traditional clustering algorithm, is done only to this subset. However, it may not work as well with datasets that contain a high number of outliers and noise.

Our implementation: GOWER's distance

For this exercise, we were given a code that allowed us to run CURE only on numeric data. Our goal was to be able to implement Gower's distance, since our dataset contains categorical data as well. Our hope is that we will obtain a better clustering since we will be able to give more information with the categorical data.

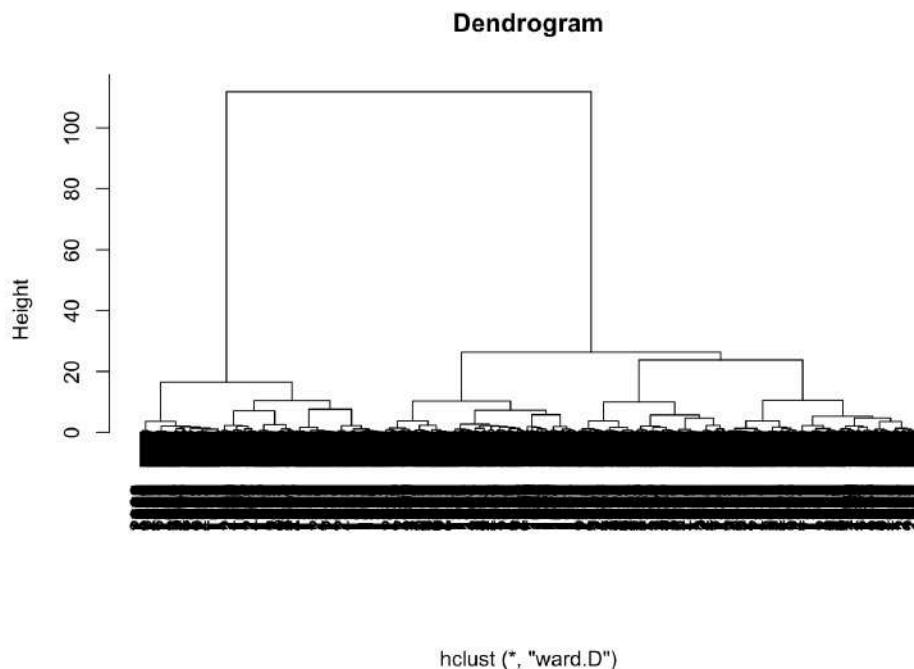
For the clustering of the medoids, we used daisy, provided by a library from R called "cluster". This library allows us to hierarchically cluster the smaller sample using Gower's distance by previously calculating the distance/dissimilarity matrix.

The main challenge of the programming was to figure out how to calculate the center. Originally, our code used k-means to find it, but k-means does not work with categorical data. Finally we decided to separate the numerical variables to find the center using k-means, and then find the center for each categorical variable by selecting the mode.

To assign each of the remaining points we used the function `gower_dist` provided by the library “`gower`” to find the distances of each point to each center.

CURE results

Firstly we will comment on the initial parameters fed into the system, the value for the previously mentioned r will be 20% of the points per cluster. The distance we will be using to calculate the centers will be Gower’s, evidently. For the clustering of the medoids we will use Ward distance on the dissimilarity matrix. And finally, the number of clusters k will be 4, as we have determined by looking at the dendrogram we plotted below, generated from the hierarchical clustering of the medoids.



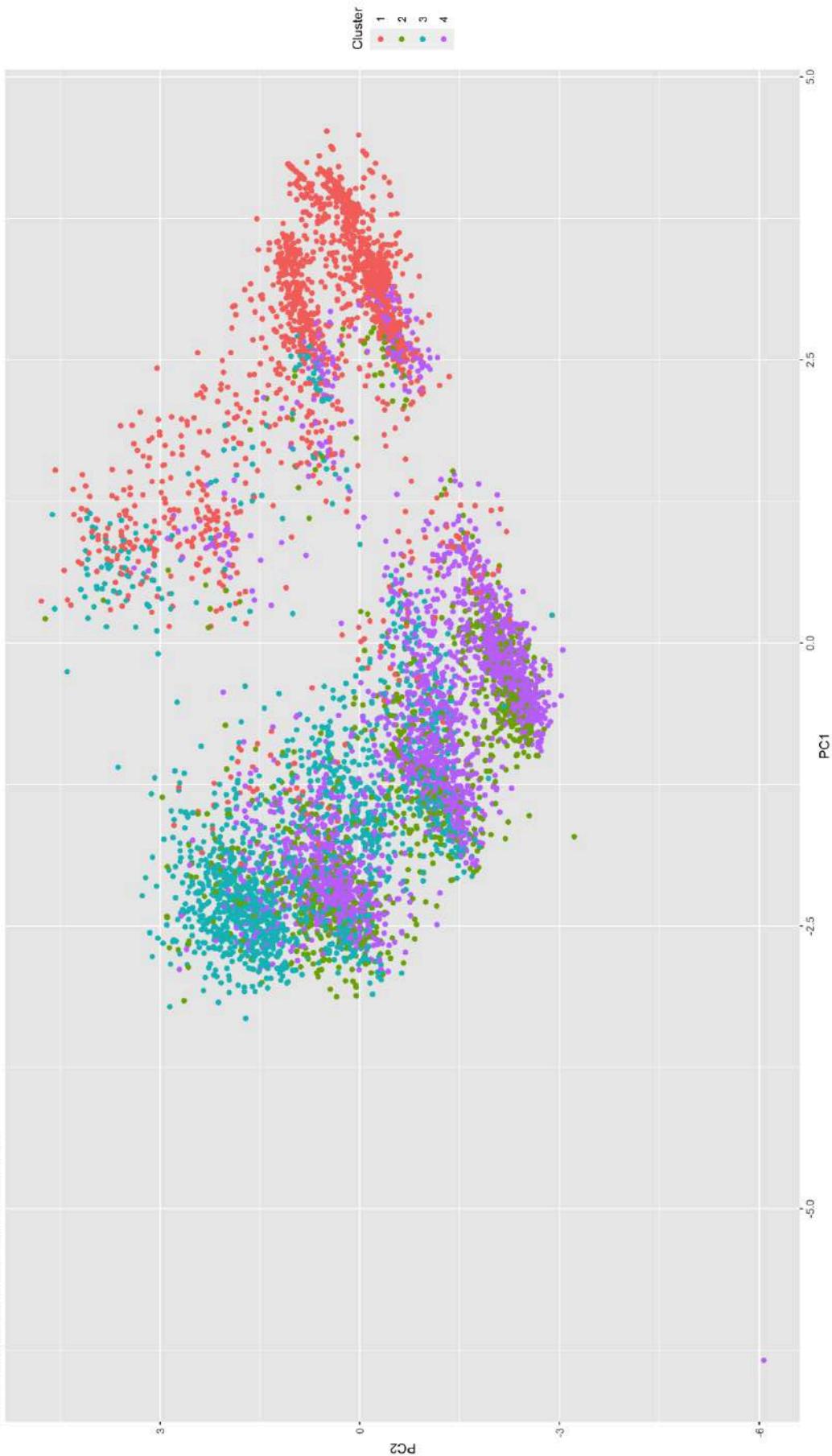
Once clustered, we checked the number of points assigned to each cluster. Keeping in mind that our dataset has 7000 rows, we can see that the individuals are pretty evenly distributed among clusters.

```
> table(CUREclusteredData$cluster)
```

1	2	3	4
1954	1082	1632	2332

Next, we used PCA to visualize the clustering in a two dimensional space. The graph is on the next page. From what we can see, the PCA is not clearly allowing us to appreciate each cluster, since they are situated in a much higher dimension. However, in the graph it looks like we can see several layers of each cluster. We can distinguish three purple regions, which we can assume are joined together in the higher dimensions, the same goes for the green cluster's points. On top of the purple and green points, we see a region of blue points indicating a third cluster, we can also see some blue points in between the layers of purple, which we can also hypothesize to be layers of the blue cluster. We can also observe that to the right there are some points which are separated from the rest, mostly composed of red points, also separated into what looks to be layers.

CURE clustering visualization using PCA



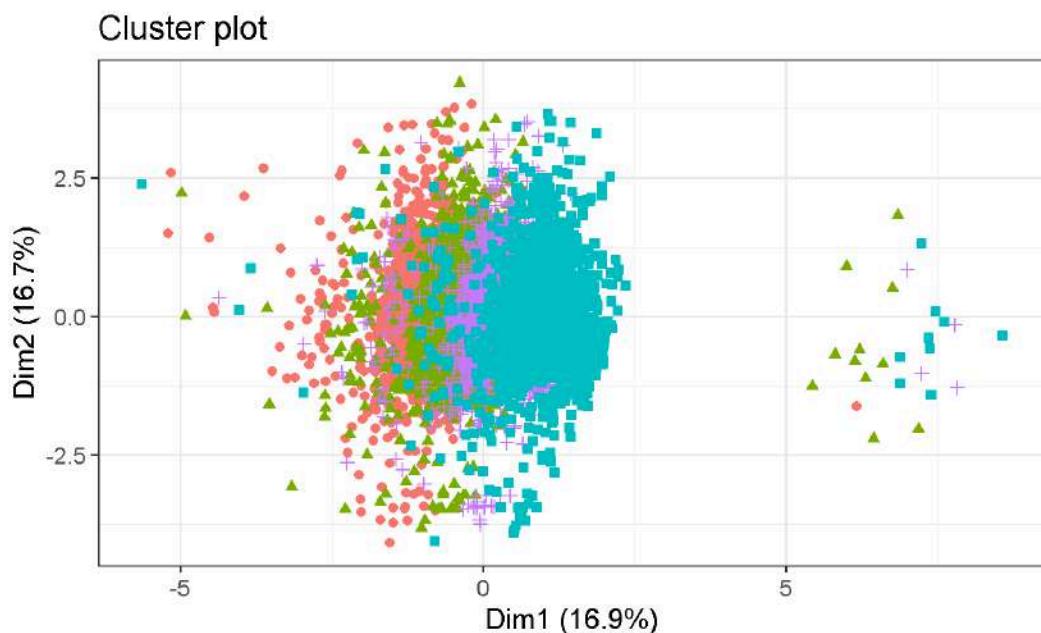
DBSCAN

In the upcoming section, we will implement DBSCAN, which is a density-based clustering algorithm used to identify clusters of data points in a dataset. This algorithm is especially effective for datasets that have complex and irregular shapes. It is important to mention that all data used in this section has been normalised, as DBSCAN has a high distance sensitivity.

In our project, we will apply DBSCAN to analyze the resulting clusters obtained from numerical values, examine their unique characteristics, and determine if they share any similarities with the gender-based cluster. By doing so, we can investigate whether certain patterns inherent in the numerical data of our dataset relate to the gender of the users in our data.

Ensuring that its use is appropriate

However, before proceeding with DBSCAN, we need to ensure that it's the appropriate clustering algorithm to use. We can determine this by first analyzing the results of a PCA on the data. A scatter plot generated from the PCA revealed distinct forms in the plot, as shown in Figure X, indicating the need for DBSCAN for clustering.



In particular, the scatter plot shows two primary clusters of points with different densities. The less-dense cluster may simply be outliers. Furthermore, the main cluster also includes a trail of points that could also be outliers. This suggests that using other clustering algorithms, such as k-means, would likely combine both an outlier and non-outlier data in the clusters, as demonstrated in the provided example of a k-means clustering result.

Given the need for DBSCAN, we can proceed with tuning the hyperparameters.

Tuning of hyperparameters

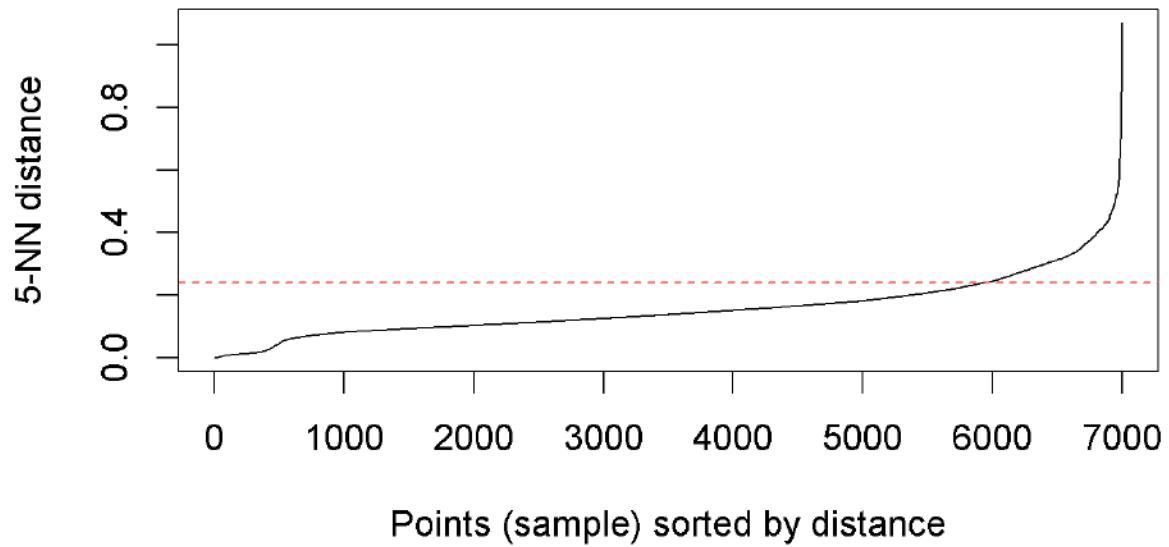
In the case of DBSCAN, this algorithm only possesses 2 hyper tuning parameters that are modifiable, min-points and epsilon.

Starting with min-points, its main function is that of specifying the smallest number of data points required to form a dense region. In other words, a cluster is only formed if it contains at least the "min-points" number of data points. If a data point has fewer than "min points" neighboring points within a distance of "epsilon," it's considered a noise point or an outlier.

Following with epsilon, it specifies the maximum distance between two points for them to be considered part of the same cluster. A larger value of "epsilon" means that points farther away from each other can still be part of the same cluster.

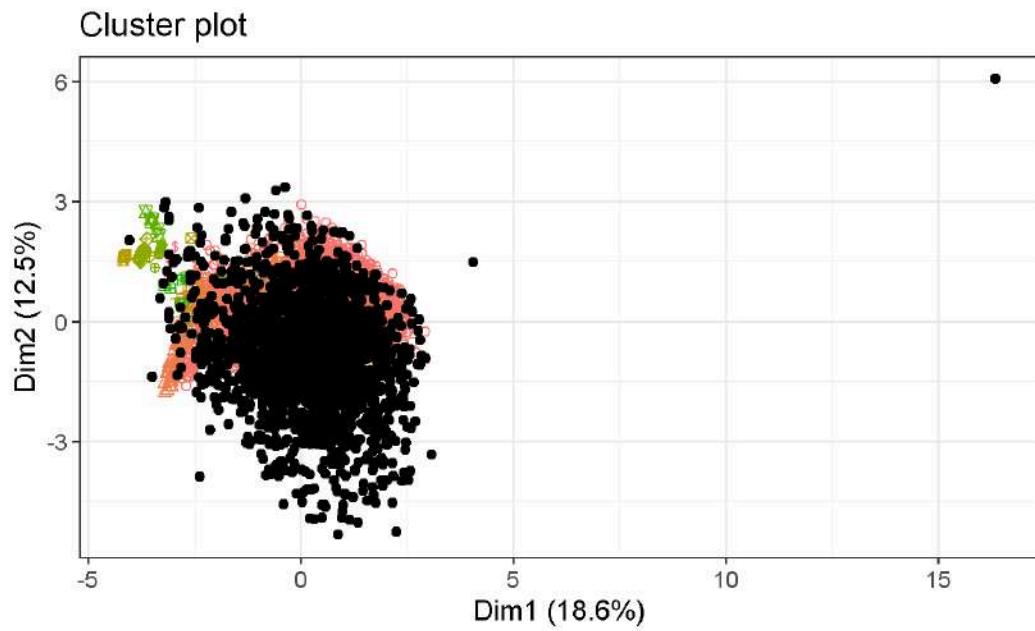
Having set the definitions for each parameter we calculate the min-points by first multiplying the number of rows of numerical values in our dataset by a percentage. In this case, the latter has been chosen to be 0.25% as it is the standard. This results in a min-points value of 18.

Afterward, we calculate epsilon. To do so, we make use of an elbow plot. In the case of DBSCAN, the elbow plot is obtained by using a KNN distance plot. This means that the desired value we wish to get is that it crosses the curve at the point with the most change in the slope. Consequently, this resulted in an epsilon value of 0.24 as seen in the figure down below.



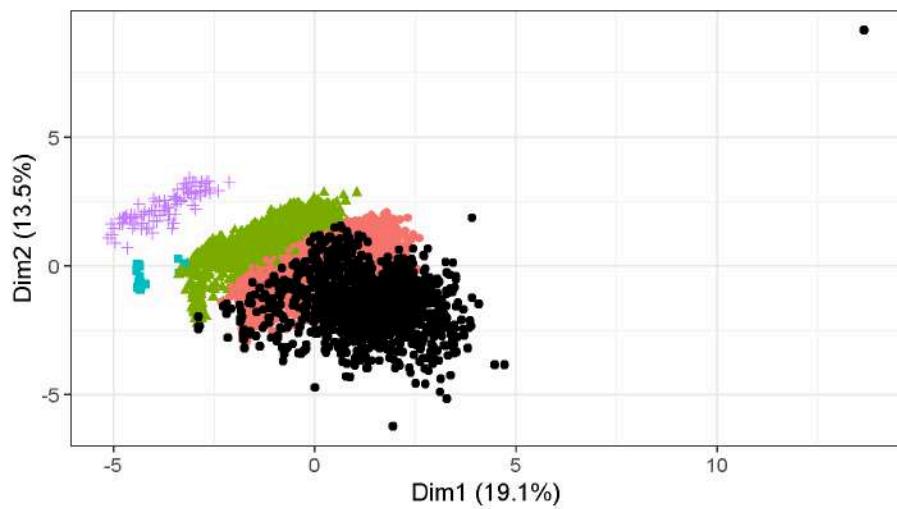
Analyzing the results

With the parameters tuned, the only step left is to execute the algorithm. In the following figures, there's the comparison of the resulting plots after executing the program with the default parameters and the parameters chosen prior, the latter from two different angles.

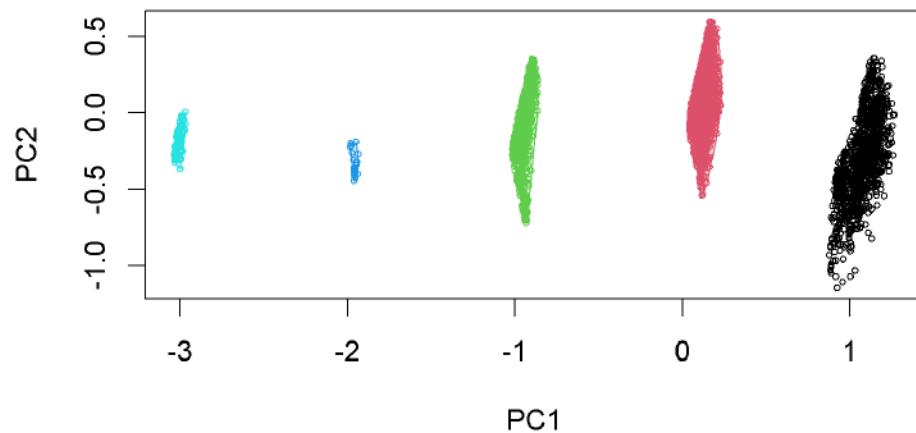


It can be seen with default parameters in the figure prior that the angle of the plot doesn't allow us to perceive much. However, we can see how the 'layer' at the front is all considered an outlier, which does permit us some insight into our data. It seems as if there's the possibility of the data being structured by stacked layers. And if that is true, if those layers are defined, we can encounter very distinct groups of individuals to study and to observe if they correspond to their gender counterparts.

Cluster plot



Convex cluster Hulls, $\text{eps} = 0.24$



This first hypothesis appears to be true, as with the new plots we gain a new view of the data, which exemplifies what was theorized earlier. The resulting plot is divided into four distinct and defined layers. The final clusters (by number order) have respectively 4365, 1407, 36 and 92 members each.

Now, what's left to study is their characteristics and if those characteristics correspond to those of our target variable.

OPTICS

OPTICS is another density-based clustering algorithm that shares many similarities with DBSCAN. However, the key difference between the two is that OPTICS extends the density-based approach of DBSCAN to handle clusters of varying densities and scales.

In our study, we aimed to compare the effectiveness of OPTICS and DBSCAN for clustering our dataset. To accomplish this, we evaluated the quality and definition of the clusters produced by each algorithm.

Tuning of hyperparameters

As with DBSCAN, OPTICS requires tuning of its hyperparameters, including 'min-points' and epsilon. To do this efficiently, we utilized a grid search to automatically find the optimal values.

This grid search employed the silhouette method. The silhouette method assigns a score to each data point based on how well it belongs to its assigned cluster. A high score indicates that the data point is well-suited to its cluster. So, by calculating the mean score across all data points, we can identify the epsilon and min-point values that produce the highest score. Our analysis revealed that an epsilon value of 0.9 and a value of 5 for min-point were optimal.

Additionally, there's also a final hyperparameter to tune which is the reachability plot, to be more precise, its cutting point.

But, before being able to explain what a reachability plot entails, the concept of reachability distance must be explained.

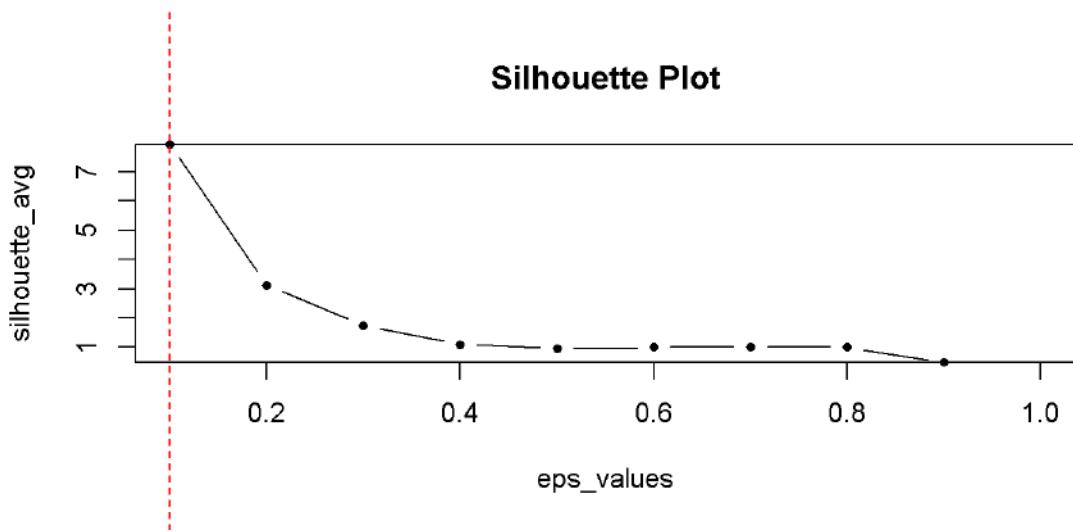
In a broader definition, reachability distance is a measure of the proximity between two points in a dataset, taking into account the density of the points surrounding them. In OPTICS, the reachability distance between two points is defined as the maximum distance between them and the nearest point with a higher density. In other words, the reachability distance measures the minimum density required to reach a point from another point within a specified distance.

With that definition covered, we can proceed the explanation of what is a reachability plot in this context.

A reachability plot is a visualization tool that allows for variable density extraction of clusters in a dataset. By combining the reachability distances and the data set ordering (arranging the data points in a dataset based on their reachability distances), we can create a plot where the Y-axis represents the point density and the X-axis represents the ordering of the points. This allows us to identify clusters as peaks in the plot, where the height of the peak corresponds to the density of the cluster.

Cutting the reachability plot at a single value is similar to using a threshold in the DBSCAN algorithm. Points above the threshold are considered noise, and points below the threshold are assigned to clusters. Each break in the plot from left to right signifies the start of a new cluster. By adjusting the threshold, we can control the sensitivity of the clustering algorithm and identify clusters of different sizes and densities.

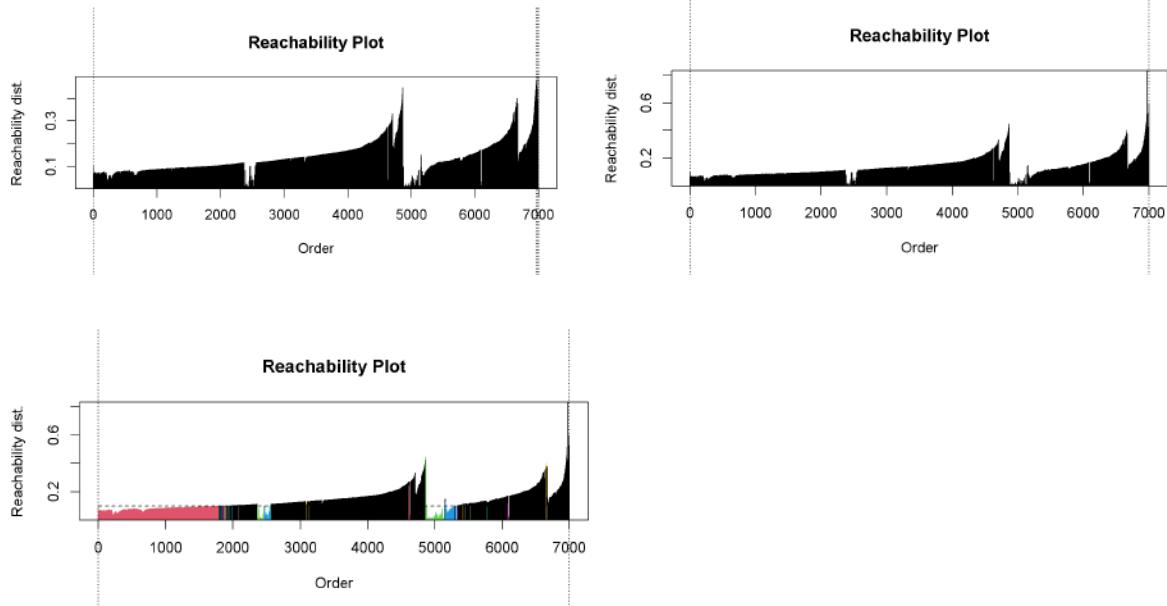
To decide the cutting point, another silhouette method was employed. In this case, we applied a sequence of 10 values ranging from 0.1 to 1 with a 0.1 interval. Then, these values will also be applied as the epsilon value while executing the silhouette method on the OPTICS for all the values. The final silhouette plot is the following:



As mentioned earlier, in the silhouette method you need to select the highest value point in the plot, which in this case will be 0.1.

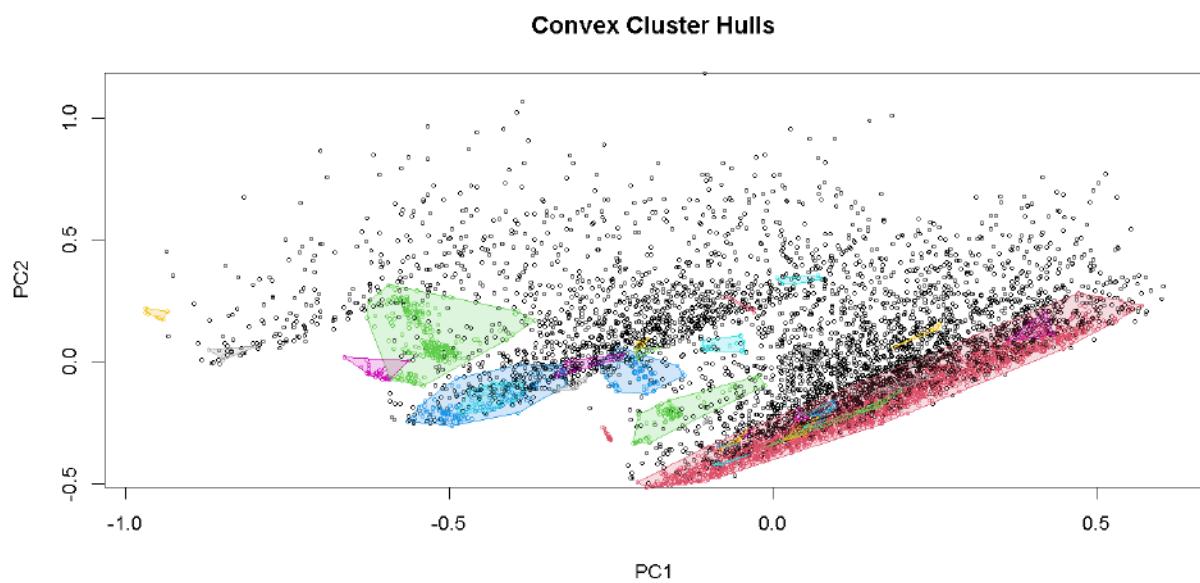
Having the cutting point selected, we can now cut the reachability plot, as seen in the following figure. From left to right is the evolution of the reachability plot, first the with default parameters,

then after finding the optimal values for the OPTICS adn finally after applying the cutting point.



Analysing the results

Applying the same visualization method used on DBSCAN, we can see the resulting clusters.



As we can observe, the use of OPTICS for this dataset appears to be a worst option compared to other clustering algorithms such as DBSCAN. This is because the resulting don't appear to be neither defined nor encapsalute all the data points that they are supposed to, leading to the creation of multiple clusters of less than 10 individuals, as we can see in hte image down below with all the resulting clusters.

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	
4351	1794	11	6	6	8	5	15	2	4	4	3	4	3	4	5	6	3	107	79	15	3	5	
23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39							
5	9	15	284	148	21	9	4	7	2	5	3	1	5	27	7	5							

PROFILING

CURE

Filtering out irrelevant variables

Random Forest

We can use Random Forests to assess the importance of a feature in separating the clusters by measuring the decrease in accuracy or Gini index that results when that feature is randomly permuted while all other features remain unchanged. These measures are used to assess the relative importance of features in predicting the outcome variable.

We can use this to determine which features are most relevant when separating our clusters. The way to do this is to train a RF with our clustered dataset to predict in which cluster a row belongs. Then we can check the Mean Decrease Accuracy and GINI to see the most relevant features.

The training of the model was done with 80% of the clustered dataset. We reserved 20% to test the accuracy of the model. It is important to also test the accuracy of the RF to know if we can trust the importance of the features that the RF tells us and other conclusions we may extract by looking at the Mean Decrease measures.

The accuracy of the model we trained is quite good at 83%:

```
> print(paste("Accuracy:", round(accuracy, 2)))
[1] "Accuracy: 0.83"
```

We will now explain how to interpret the importance measures:

- **Mean Decrease Accuracy**

Mean Decrease Accuracy, or MDA, is a measure of feature importance. The MDA of a given feature evaluates how the accuracy decreases when that feature is removed from the model. That is, it quantifies the decrease in the model's accuracy when a specific feature is removed from the training process. A higher MDA score indicates that the feature is more important in raising the model's accuracy.

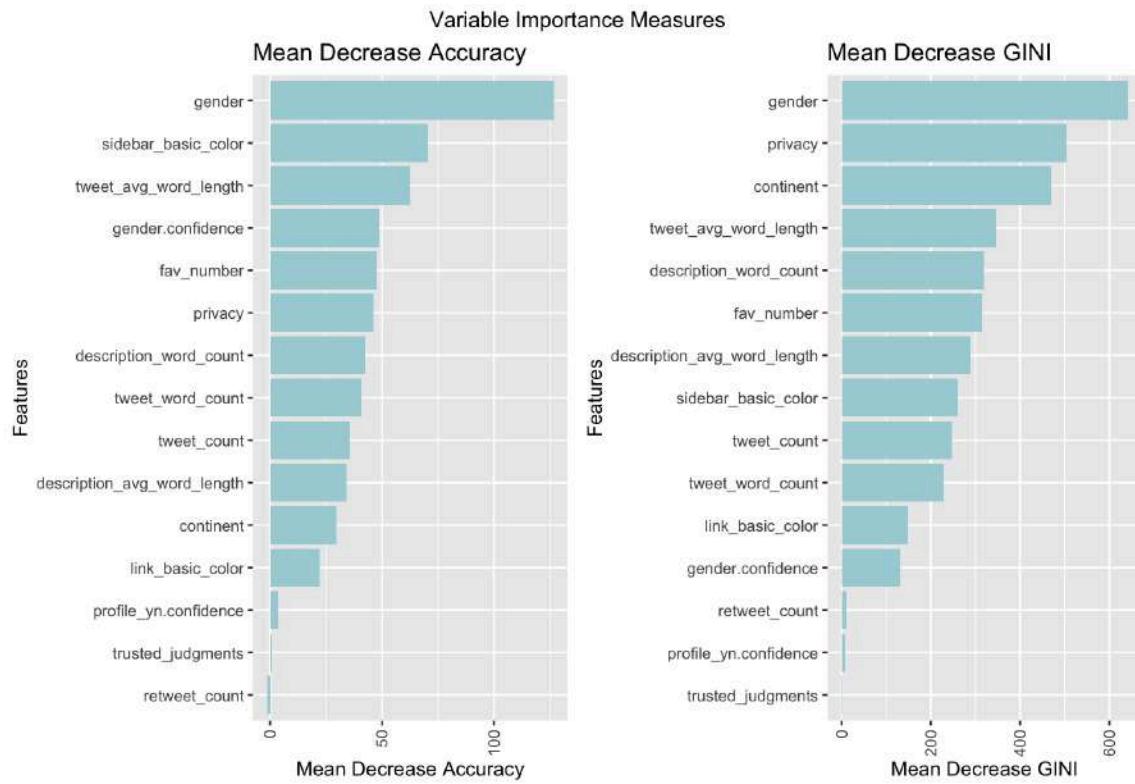
- **Mean Decrease Gini**

Mean Decrease Gini, or MDG, is a measure of feature importance. The MDG of a feature evaluates the reduction in impurity caused when a feature is randomly permuted, this is measured with the gini index. This means, it measures how much the impurity of the node decreases when a particular feature is used for splitting in the tree. A higher MDG score indicates that the feature is more important for reducing impurity in the decision tree. A high impurity means that the node will not be good at separating data into classes accurately.

In the following graph we plotted the MDA and MDG importance measures for each variable we used in the CURE clustering. The importance measures were obtained from the RF we trained, as previously mentioned. Each feature is ranked by importance, from highest to lowest. The higher the score the more important the feature.

As we can see the most important features in separating the clusters are Gender, Privacy, Continent, Sidebar color and Average word length. Gender being the most distinctive feature of them all as it has both the highest MDA and MDG.

It is important to note that these measures only provide a relative ranking of feature importance. However, looking at the accuracy, we believe that we can trust, more or less, the conclusions we obtained. The most irrelevant features are profile_yn.confidence, trusted_judgments and retweet_count, which we will dismiss from our inference test analysis.

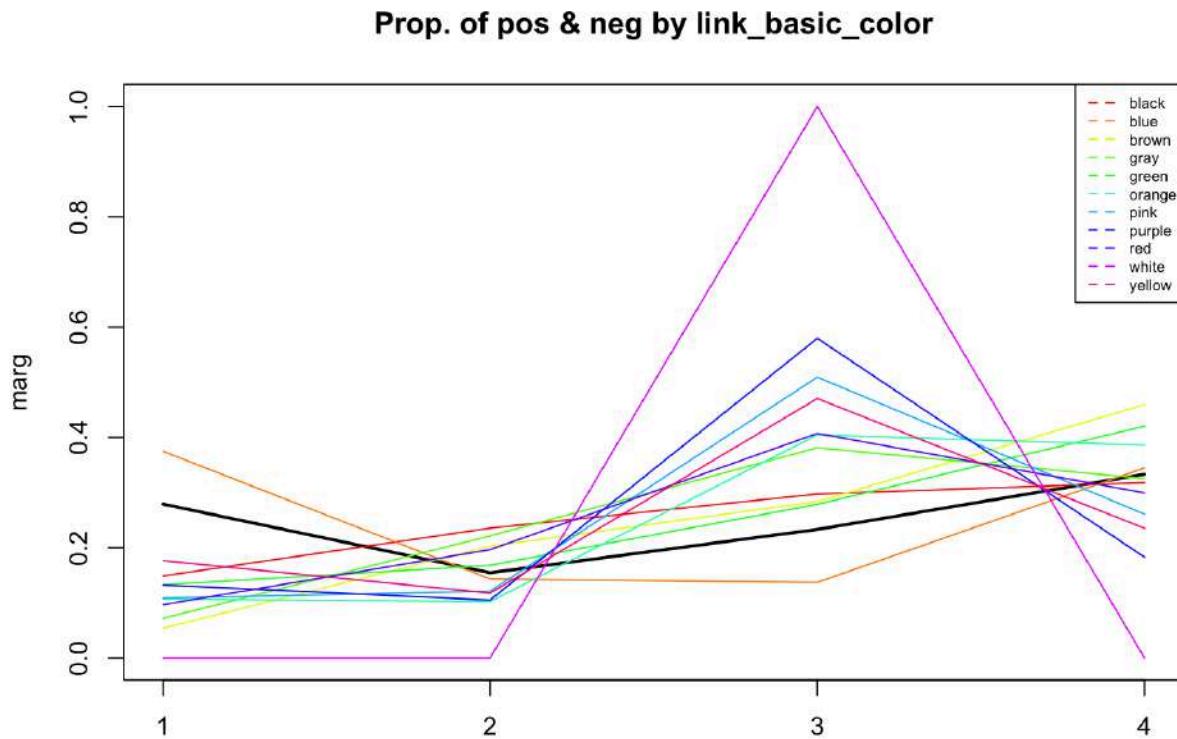


Inference tests

After filtering out the non-relevant variables. We proceed to do several plots to represent the distributions of numerical variables and to characterize each cluster, and conduct inference tests.

We conducted a series of inference tests to ensure that the results for each cluster are actually different from each other or if any observed differences are simply due to chance. The tests done are for quantitative variables the p-values of an ANOVA, Kruskal-Wallis and Valors tests. For the qualitative variables we will make use of the p-value of a ValorsTest too.

Link basic color



The chart above shows the probability that each category of the link basic color variable belongs to a cluster. In other words, it plots the percentage of individuals with a certain category that are in each cluster, for example, in this case, nearly a 100% of the individuals with a white link color are in cluster 3, or nearly 40% of the individuals that have a blue link color are in cluster 1. In the chart, we can see that the third cluster contains a higher percentage of the color white. On the other hand, in the other clusters, we see the opposite as the probability now is around 0. It can also be seen how most colors with the exception of black, blue, and brown are at a lesser probability percentage in the other clusters that are not the third and the complimentary happens in the third.

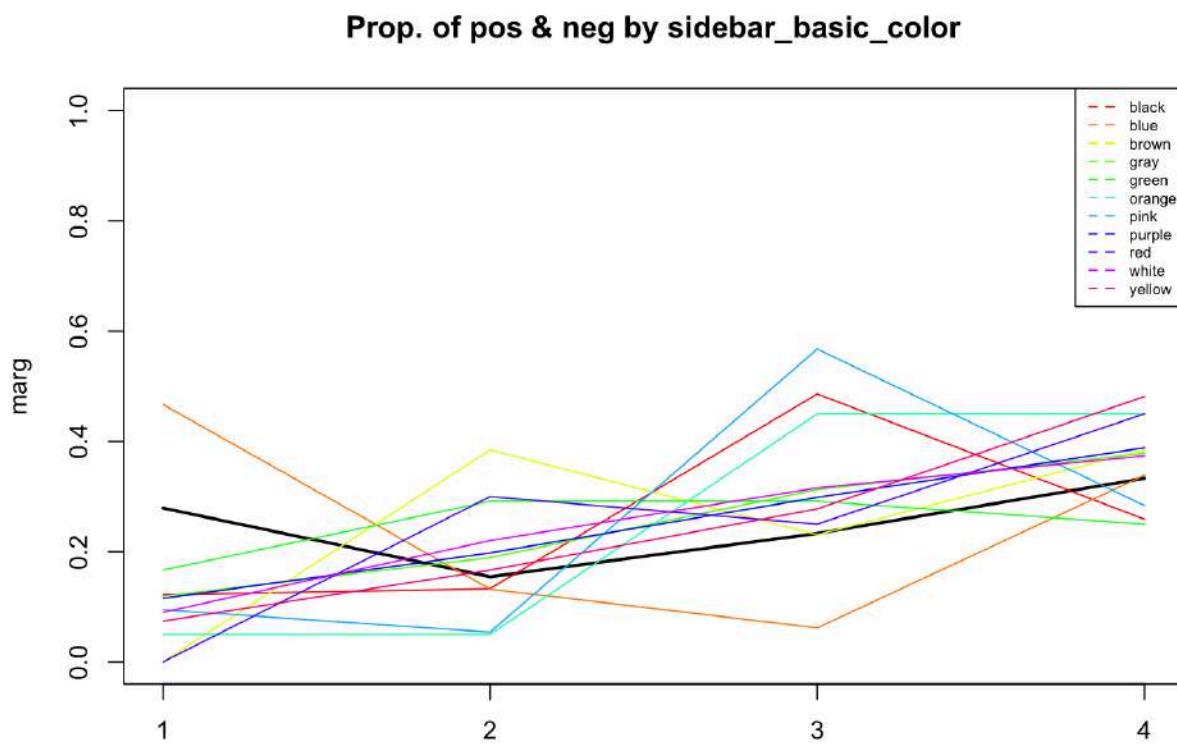
We have also conducted a Chi-square test between cluster and link basic color to determine whether there is a significant association between these two categorical variables. The p-value

we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the categorical variables.
- H_1 : significant association between the categorical variables.

	Chi-squared
P-value	p-value < 2.2e-16

Sidebar basic color



The chart above shows the probability that each category of the sidebar basic color variable belongs to a cluster. In the chart, we can see that the third cluster contains a higher tendency to

the colors orange, pink, and black and a lower probability of blue. It can also be seen in the second cluster that it has the complementary color probabilities to the first.

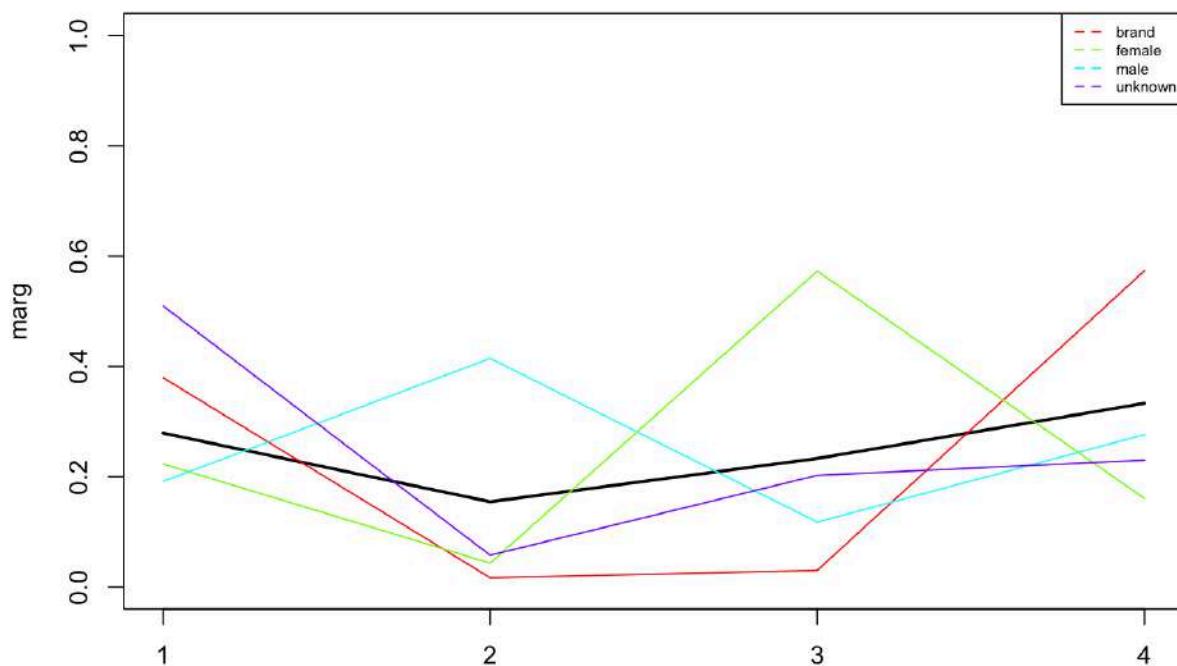
We have also conducted a Chi-square test between cluster and sidebar basic color to determine whether there is a significant association between these two categorical variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the categorical variables.
- H_1 : significant association between the categorical variables.

Chi-squared	
P-value	p-value < 2.2e-16

Gender

Prop. of pos & neg by gender



The chart above shows the probability that each category of the gender variable belongs to a cluster. In the chart, we can see that the second, third and fourth cluster contain a lower percentage of users of unknown gender. We can also see how the most predominant genders are male and female with a percentage over 80%. We see that cluster 2 is the cluster with highest proportion of the males and cluster 3 is the one with the highest proportion of females.

We have also conducted a Chi-square test between cluster and gender to determine whether there is a significant association between these two categorical variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

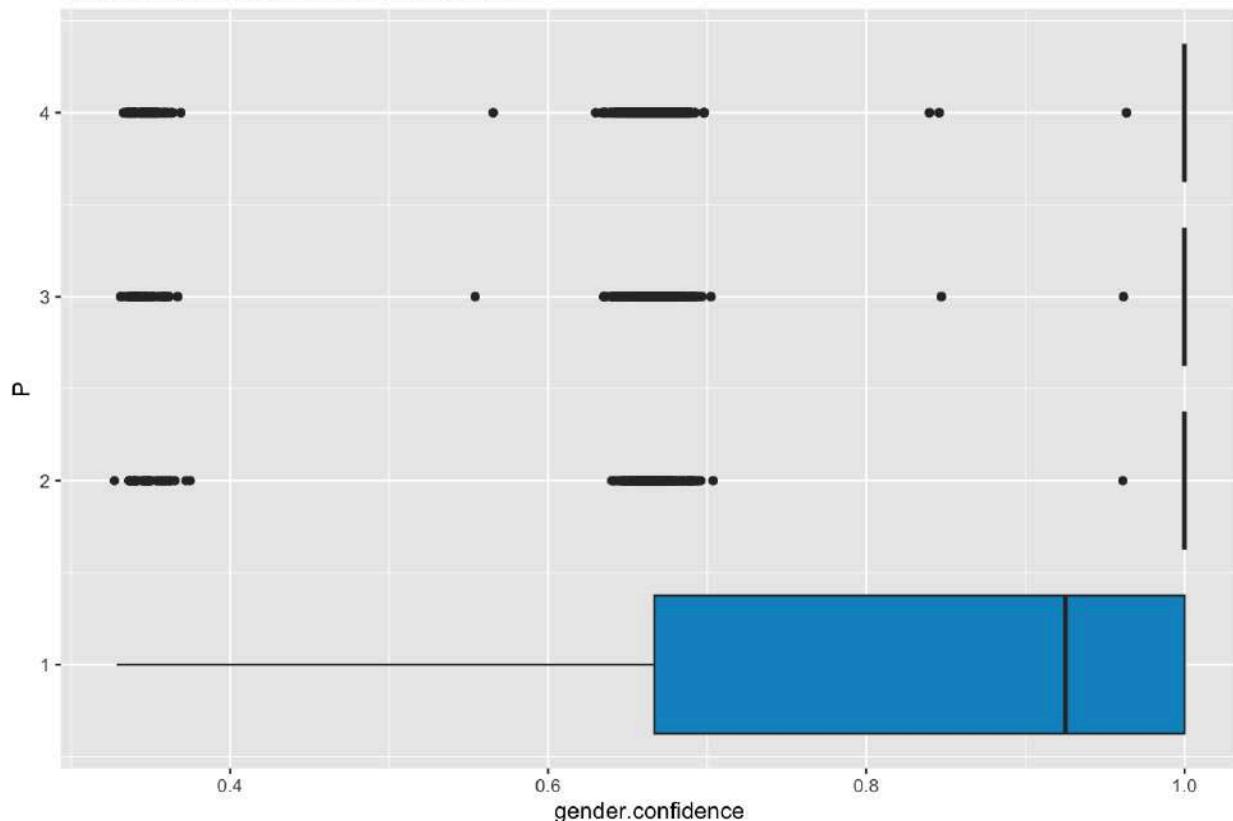
- H_0 : no significant association between the categorical variables.

- H_1 : significant association between the categorical variables.

	Chi-squared
P-value	p-value < 2.2e-16

Gender confidence

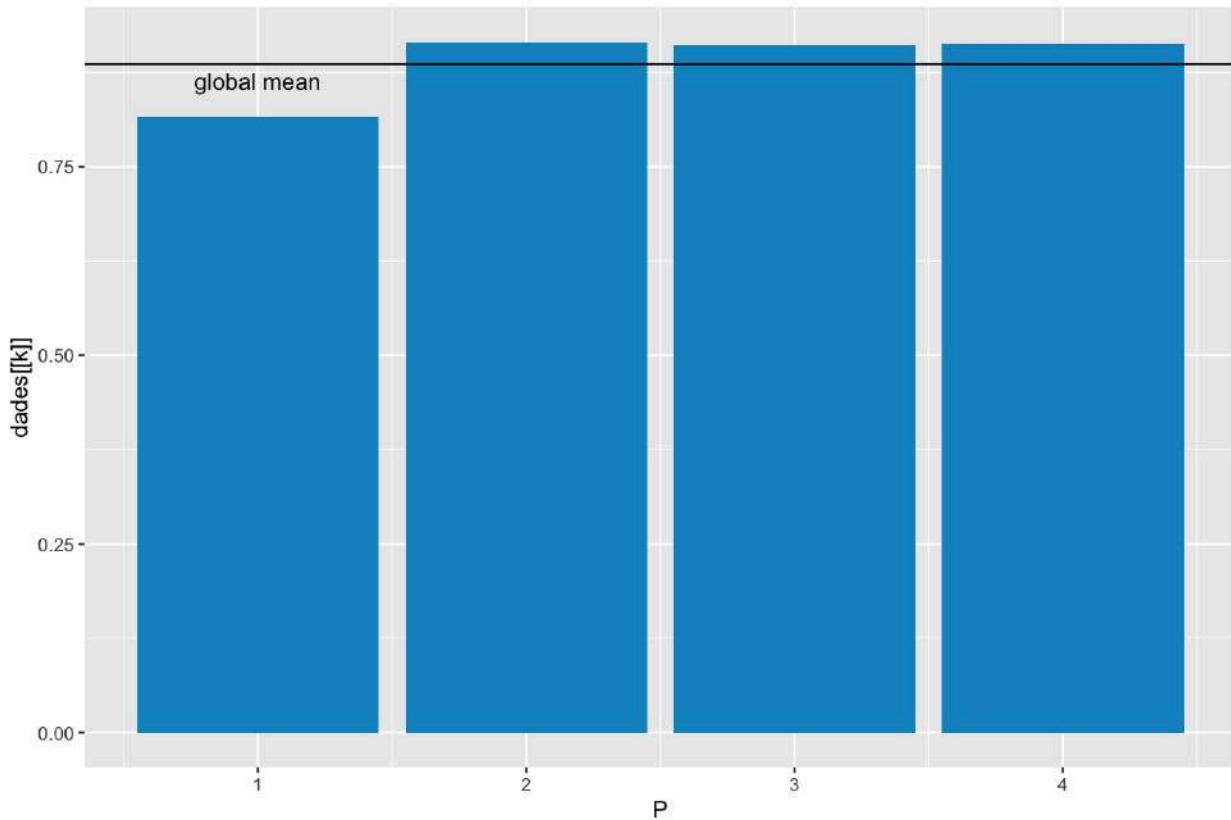
Boxplot of gender.confidence by Class



Upon analyzing its distribution, it becomes clear that there are notable differences among the clusters. The three batch pattern previously mentioned in the univariate analysis is also noticeable.

The second, third and fourth clusters display a similar pattern at the end of the possible range of values, with little to no difference and barely any interquartile range. The first cluster, however, has a notable difference of lower values compared to the rest and a much larger quartile range.

Means of gender.confidence by Class



Furthermore, upon examining the mean of each cluster, it is evident that the first cluster is composed of individuals with considerably lower gender confidence than all others. What's more, the second, third and fourth clusters present a mean surpassing the global mean.

We have also conducted a ValorsTest, ANOVA and Kruskal-Wallis tests between cluster and gender confidence to determine whether there is a significant association between these two variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

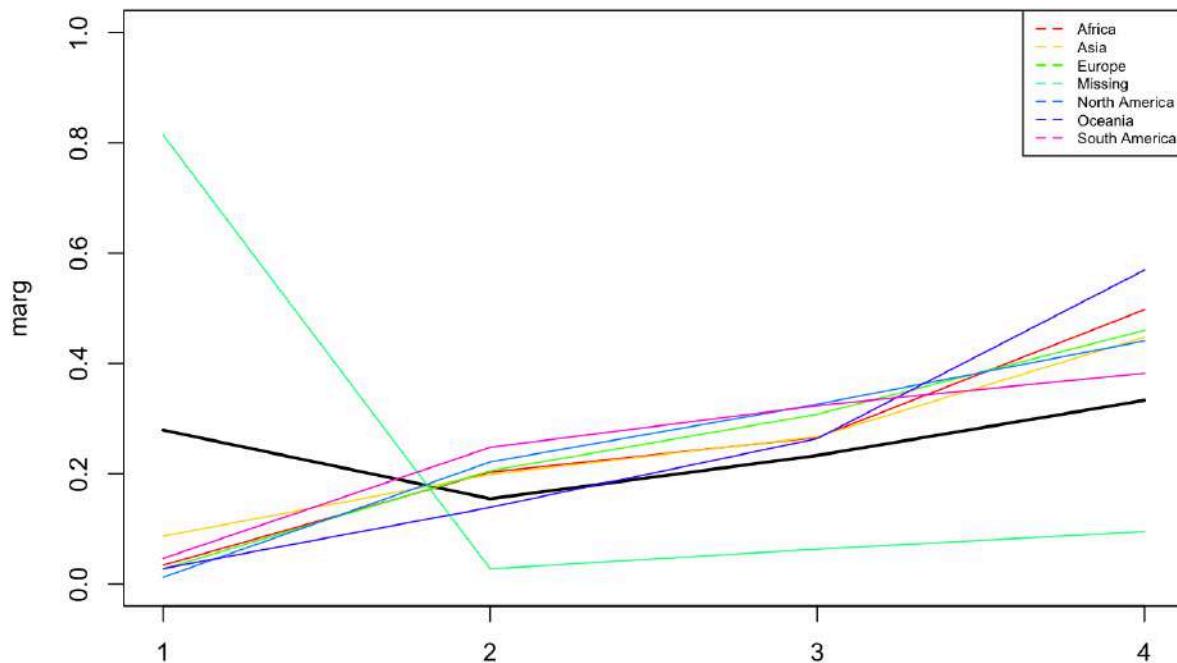
- H_0 : no significant association between the variables.
- H_1 : significant association between the variables.

	ValorsTest	ANOVA	Kruskal-Wallis
P-value	0.000000e+00 3.651832e-09	9.71064476814508e-7 5	2.04541937885151e-9 5

	7.509929e-11
	4.664572e-18

Continent

Prop. of pos & neg by continent



The chart above shows the probability that each category of the continent variable belongs to a cluster. In the chart, we can see that the first cluster contains a very high percentage of users not disclosing their location. We can also see in cluster 4 how the most predominant continents are Africa and North America with a percentage over 50%, most of the Europeans are also found here. In clusters 2, 3, 4, North America is one of the most predominant along with South America, these clusters also have very little presence of Missing. It can also be seen how the first and fourth clusters are more or less flipped versions of each other.

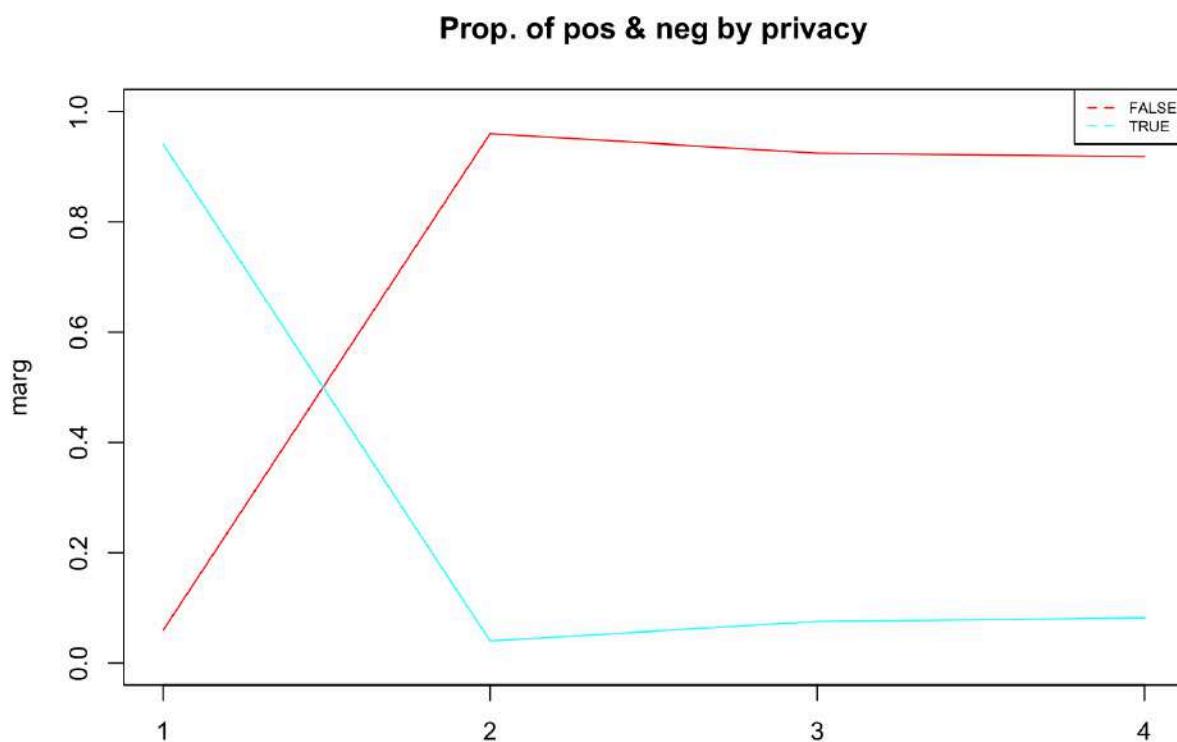
We have also conducted a Chi-square test between cluster and continent to determine whether there is a significant association between these two categorical variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and

we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the categorical variables.
- H_1 : significant association between the categorical variables.

Chi-squared	
P-value	p-value < 2.2e-16

Privacy



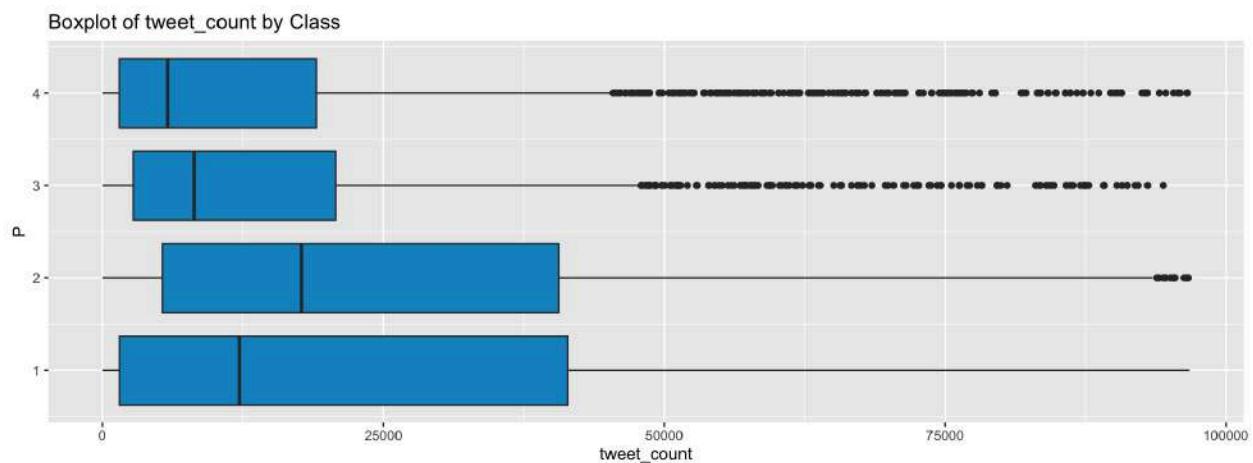
The chart above shows the probability that each category of the privacy variable belongs to a cluster. In the chart, we can see that the clusters 2,3 and 4 contain a lower percentage of not disclosing personal information. On the other hand, in the first cluster we see that the opposite is true.

We have also conducted a Chi-square test between cluster and privacy to determine whether there is a significant association between these two categorical variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the categorical variables.
- H_1 : significant association between the categorical variables.

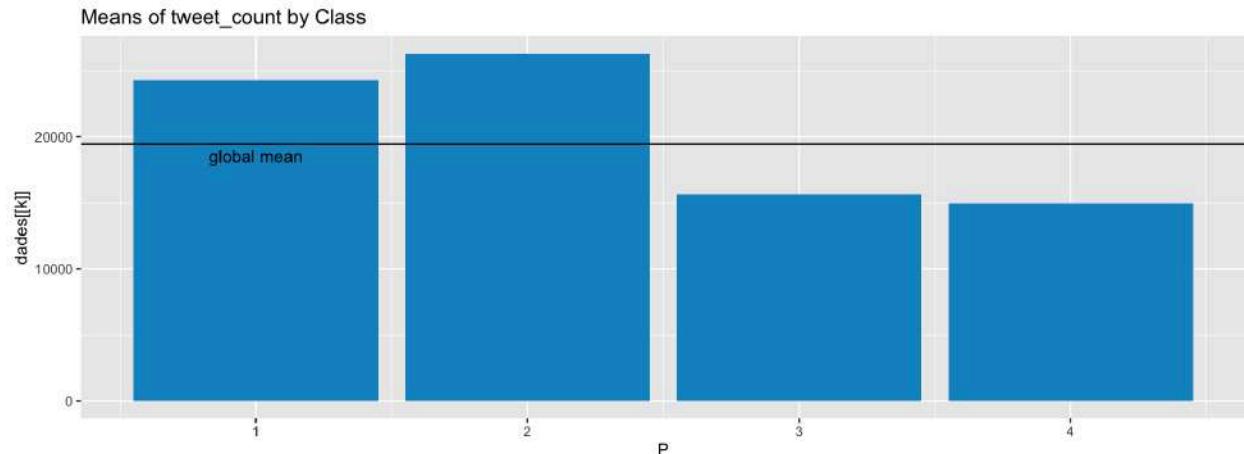
	Chi-squared
P-value	p-value < 2.2e-16

Tweet count



Upon analyzing its distribution, it becomes clear that there are notable differences among the clusters.

The third and fourth clusters display a similar pattern, with the only differences being the length of the right whisker and the median, they both have the smallest interquartile range. The first cluster, however, has a notable difference of higher values compared to the rest and a much larger quartile range, with no outliers.



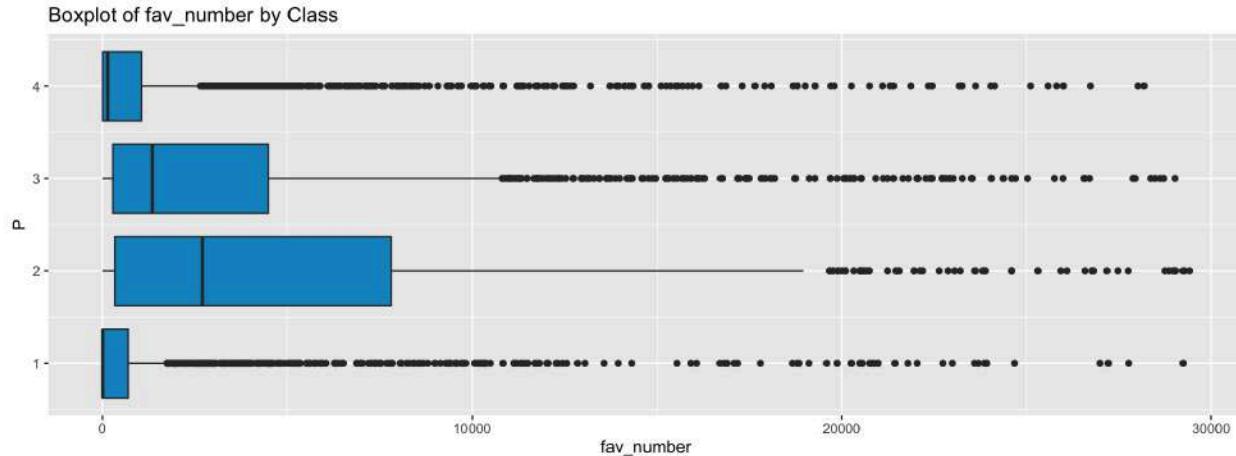
Furthermore, upon examining the mean of each cluster, it is evident that the third and fourth clusters are individuals with a considerably lower tweet count than all others, being under the global mean. The second is the opposite with the highest mean. Both 1 and 2 are considerably over the global mean level.

We have also conducted a ValorsTest, ANOVA and Kruskal-Wallis tests between cluster and tweet count to determine whether there is a significant association between these two variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the variables.
- H_1 : significant association between the variables.

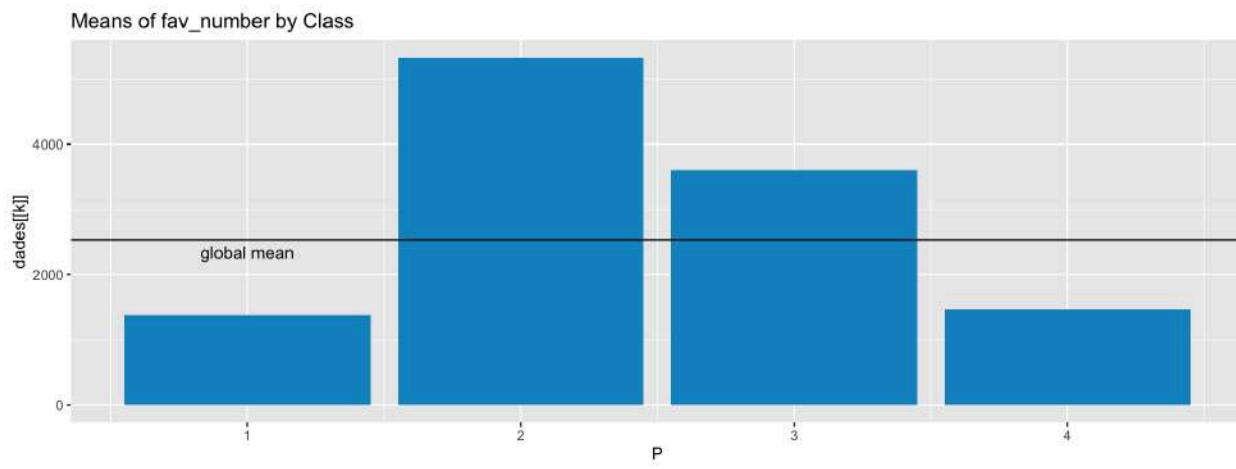
	ValorsTest	ANOVA	Kruskal-Wallis
P-value	3.690310e-27 5.154397e-26 1.798561e-14 0.000000e+00	1.22252812448677e-62	4.36395336943845e-50

Favorite number



Upon analyzing its distribution, it becomes clear that there are notable differences among the clusters.

The first and fourth clusters display a similar pattern, with the only differences being the length of the right whisker, a slightly shorter interquartile range and outliers. The second cluster, however, has a notable difference of higher values compared to the rest and a much larger quartile range. The third cluster is like a middle ground in between the 2nd and 4th clusters.



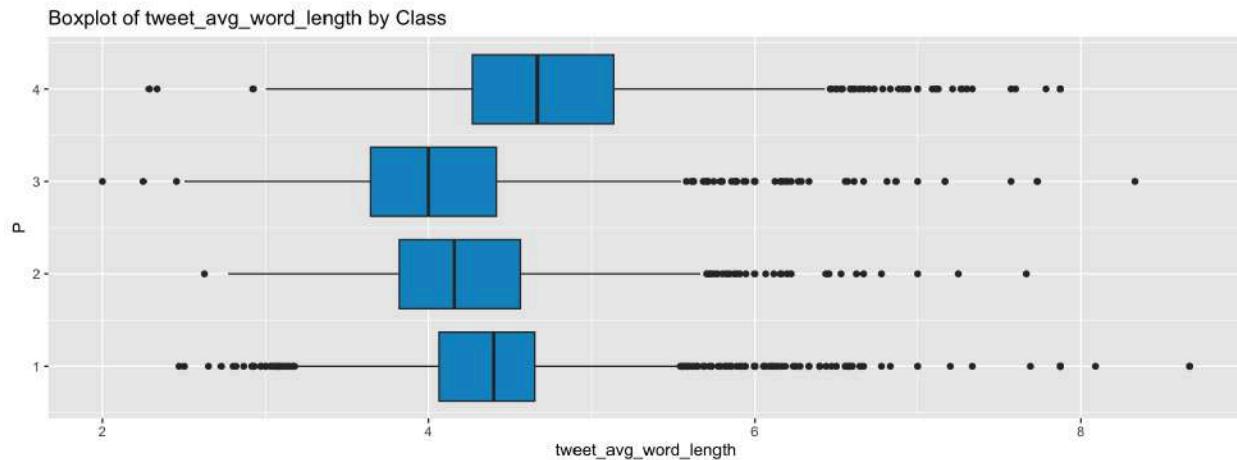
Furthermore, upon examining the mean of each cluster, it is evident that the second cluster is composed of individuals with a considerably higher favorite count than all others. What's more, the second and third clusters present a global surpassing mean, while the other cluster left are noticeably below average.

We have also conducted a ValorsTest, ANOVA and Kruskal-Wallis tests between cluster and favorite number to determine whether there is a significant association between these two variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the variables.
- H_1 : significant association between the variables.

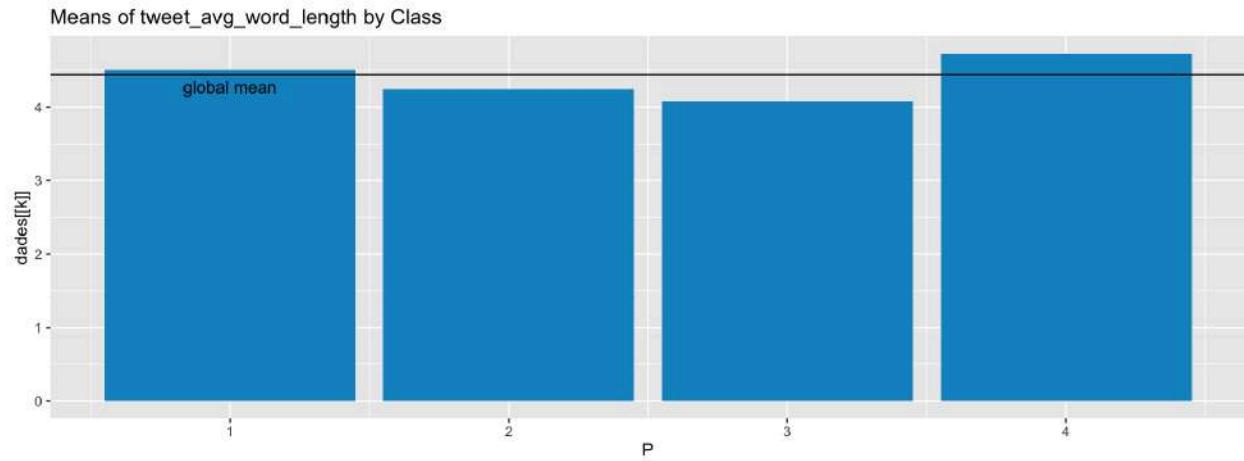
	ValorsTest	ANOVA	Kruskal-Wallis
P-value	0.000000e+00 1.630657e-92 1.720509e-24 0.000000e+00	4.43958618574133e-07	2.84061636293101e-298

Tweet average word length



Upon analyzing its distribution, it becomes clear that these clusters are quite similar in terms of this variable.

The first, second and third clusters display the most similar patterns. The fourth cluster, however, has a slight difference of higher values compared to the rest and a the largest quartile range.



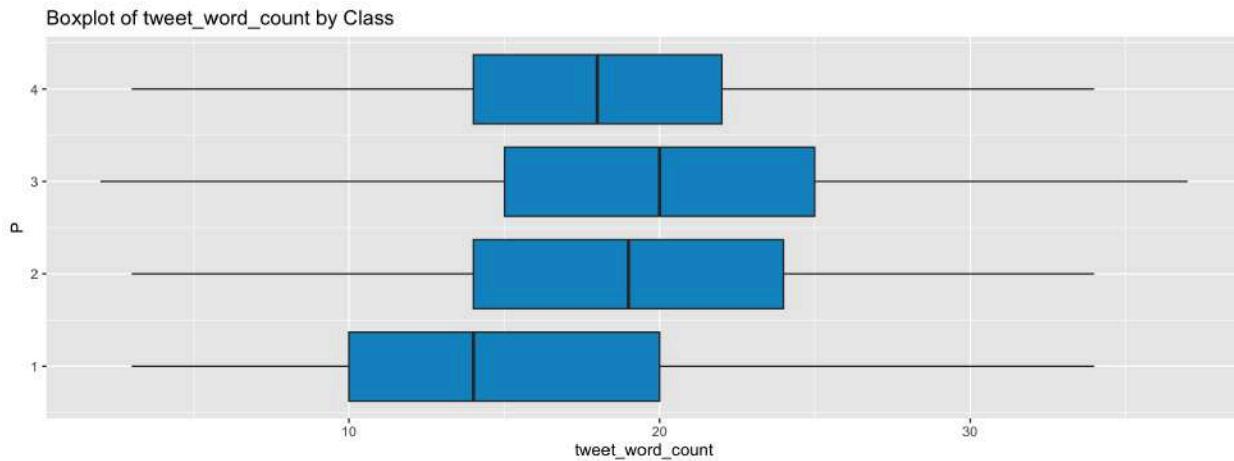
Furthermore, upon examining the mean of each cluster, it is evident that all clusters present a mean value that is very similar to the global mean.

We have also conducted a ValorsTest, ANOVA and Kruskal-Wallis tests between cluster and tweet average word length to determine whether there is a significant association between these two variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the variables.
- H_1 : significant association between the variables.

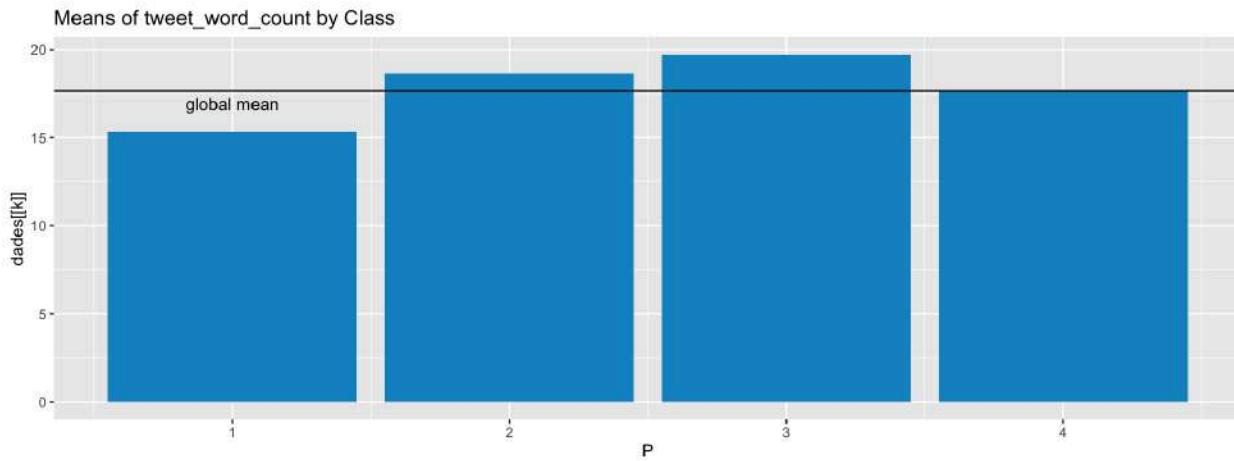
	ValorsTest	ANOVA	Kruskal-Wallis
P-value	5.942864e-06 0.000000e+00 0.000000e+00 4.622458e-99	1.7417296311717e-182	3.06128362885691e-215

Tweet word count



Upon analyzing its distribution, it becomes clear that these clusters are quite similar in terms of this variable.

The second, third and fourth clusters display the most similar patterns. The first cluster, however, has a notable difference of lower values compared to the rest, but a similar quartile range.



Again, upon examining the mean of each cluster, we see that all clusters present a mean value that is very similar to the global mean, except for the first one, which is significantly lower.

We have also conducted a ValorsTest, ANOVA and Kruskal-Wallis tests between cluster and tweet word count to determine whether there is a significant association between these two variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the

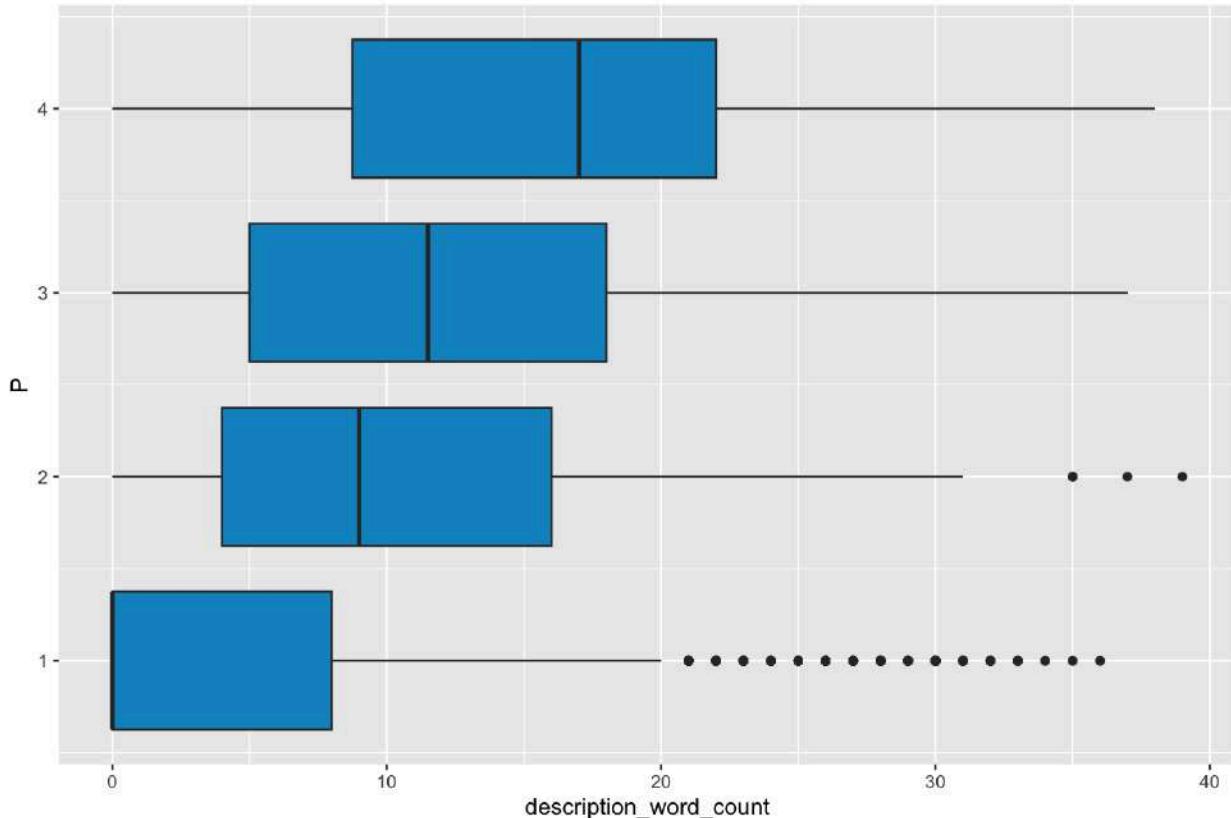
null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the variables.
- H_1 : significant association between the variables.

	ValorsTest	ANOVA	Kruskal-Wallis
P-value	0.000000e+00 2.383678e-09 4.326244e-53 3.121404e-01	1.05910325853188e-95	6.74005657756277e-108

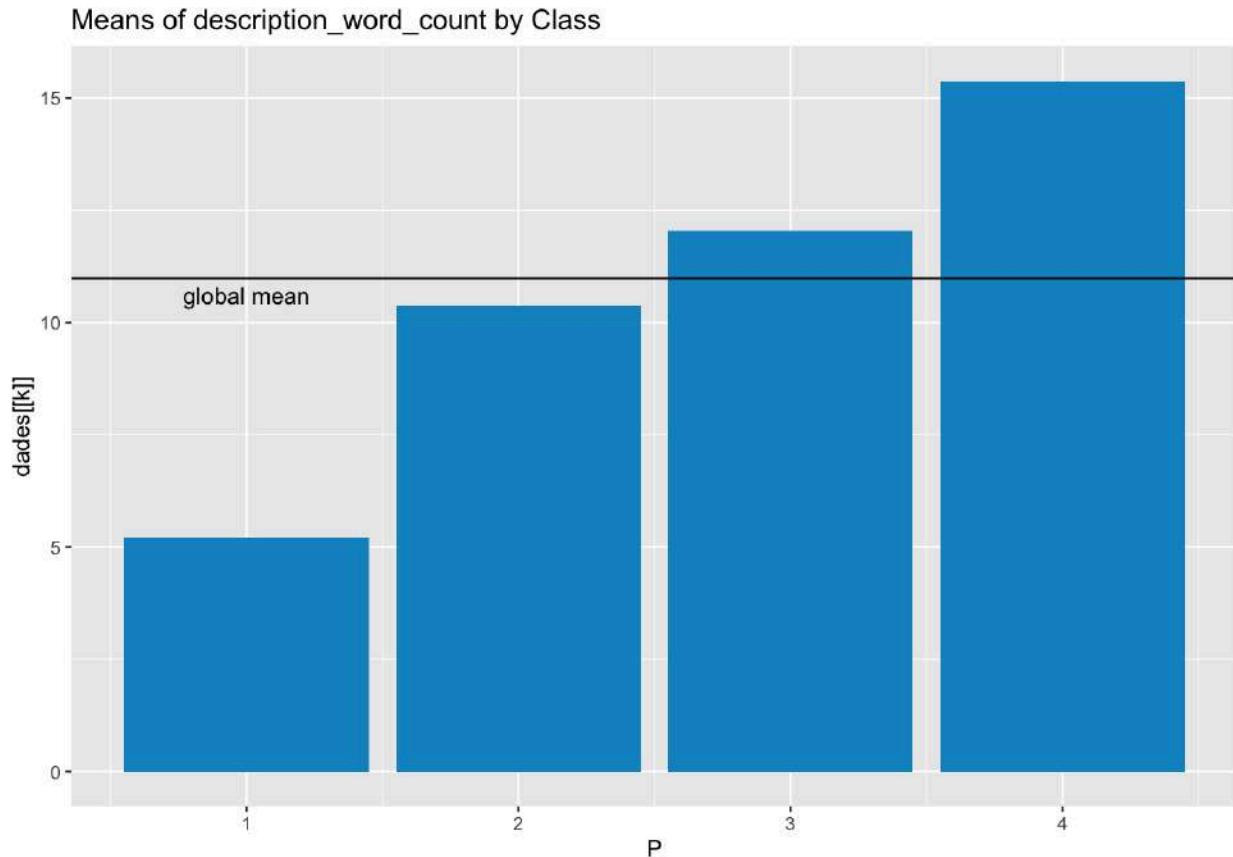
Description word count

Boxplot of description_word_count by Class



Upon analyzing its distribution, it becomes clear that there are notable differences among the clusters.

All of the clusters sit at different medians, although their interquartile range is similar, except the first one. The first cluster is the one with the lowest amount of words in the description, while 4 is the one with the most words.



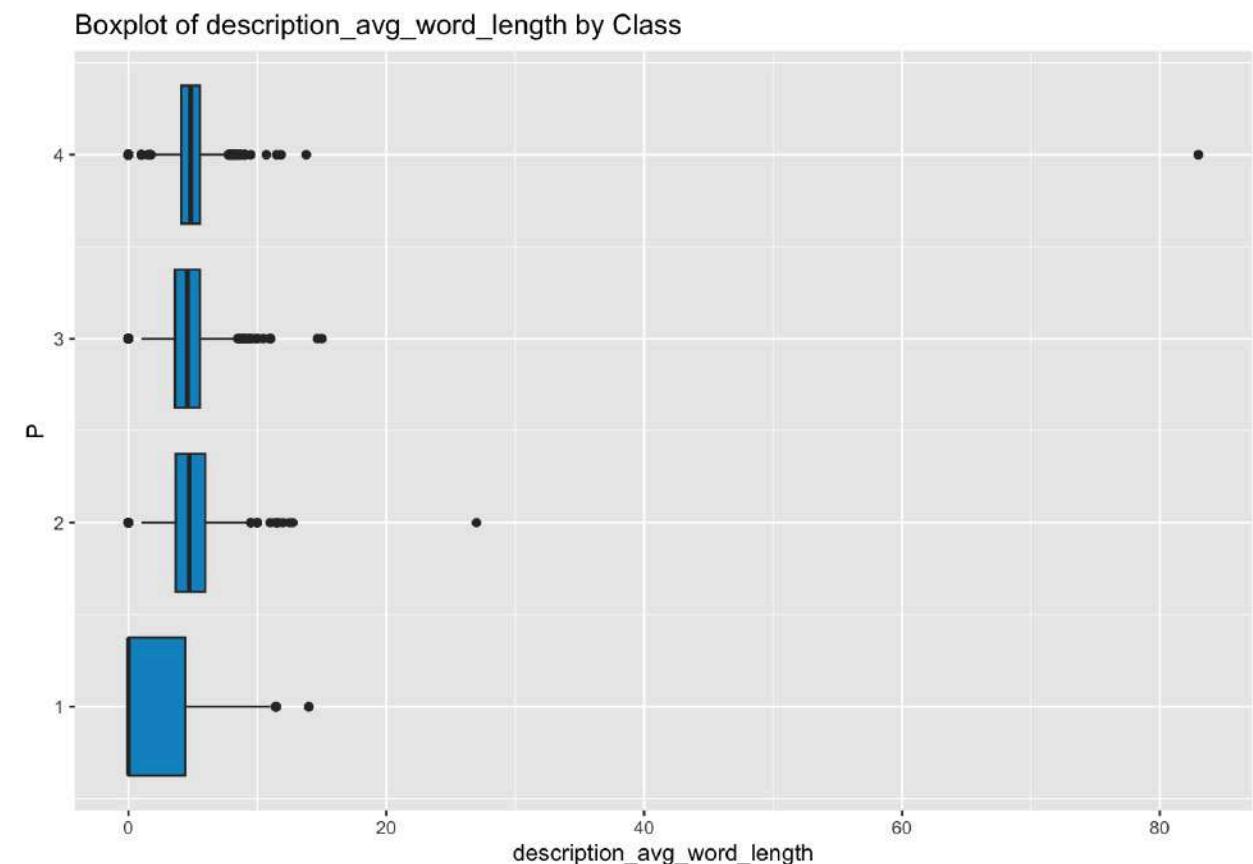
Furthermore, upon examining the mean of each cluster, it is evident that the first cluster is composed of individuals with a considerably lower description word count than all others. What's more, the fourth cluster presents a global surpassing mean by a lot, while the other cluster left are around the average.

We have also conducted a ValorsTest, ANOVA and Kruskal-Wallis tests between cluster and description word count to determine whether there is a significant association between these two variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the variables.
- H_1 : significant association between the variables.

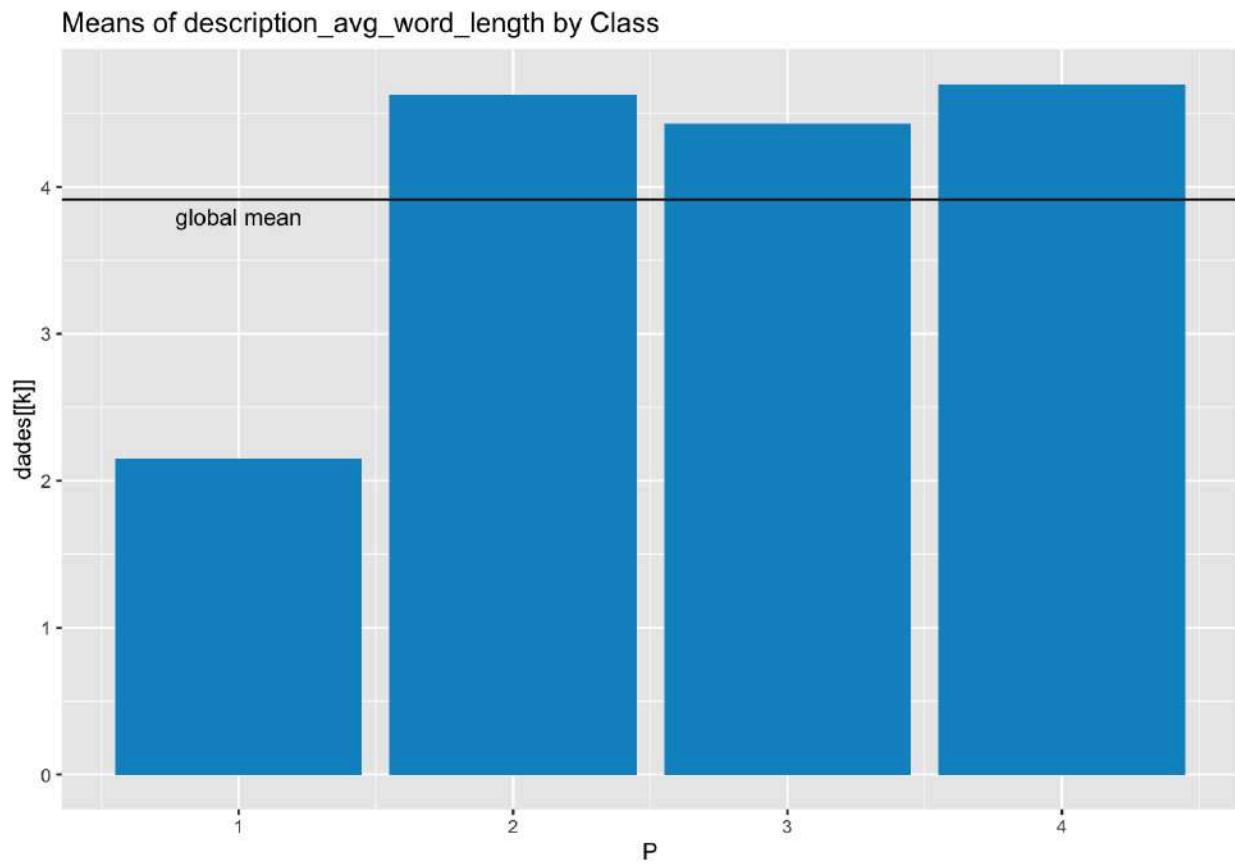
	ValorsTest	ANOVA	Kruskal-Wallis
P-value	0.000000e+00 7.798111e-03 6.143817e-08 1.679746e-168	1.29774463796134e-2 94"	0

Description average word length



Upon analyzing its distribution, it becomes clear that there are notable differences among one of the clusters.

The second, third and fourth clusters display a similar pattern, with the only difference being the slightly varying interquartile range. The first cluster, however, has a notable difference of lower values compared to the rest and a much larger quartile range.



Furthermore, upon examining the mean of each cluster, it is evident that the first cluster is comprised of individuals with a considerably lower favorite count than all others. What's more, the first cluster presents a mean well below the global, while the other clusters have similar means and are above the global.

We have also conducted a ValorsTest, ANOVA and Kruskal-Wallis tests between cluster and description average word length to determine whether there is a significant association between these two variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the variables.
- H_1 : significant association between the variables.

ValorsTest

ANOVA

Kruskal-Wallis

P-value	0.000000e+00 4.315622e-23 2.654705e-20 1.373851e-71	4.90350365867409e-248	6.94217211150219e-2 48
---------	--	-----------------------	---------------------------

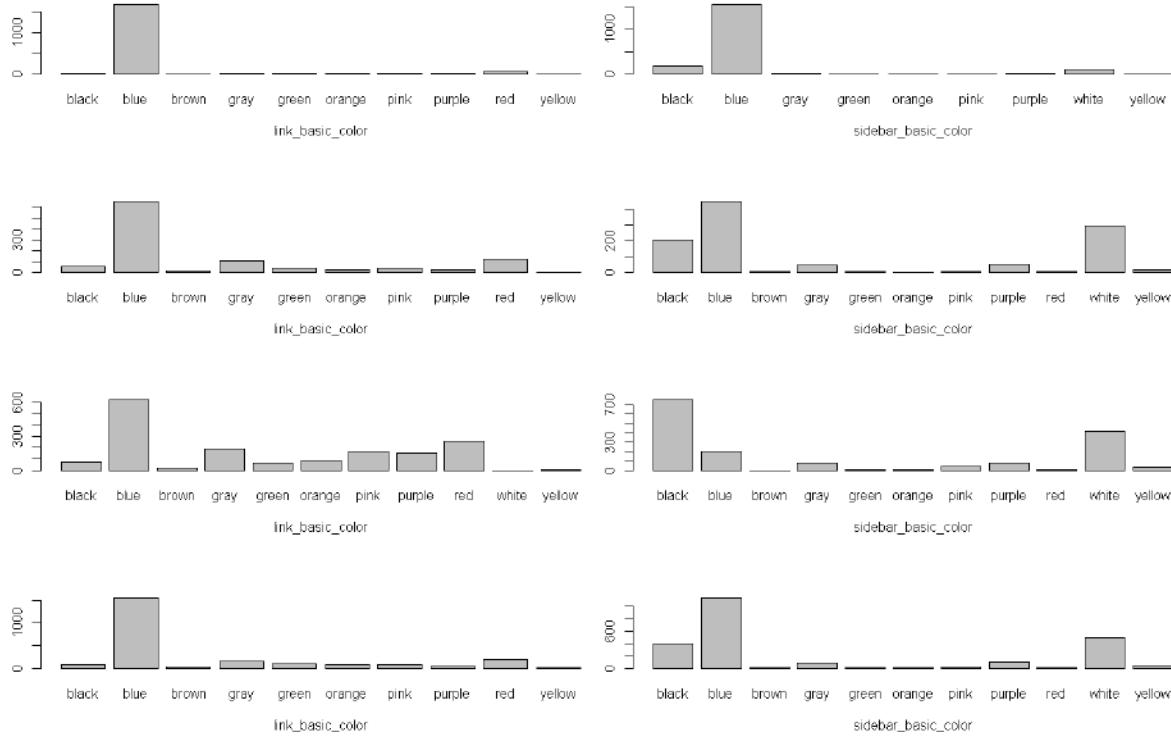
CPG

Furthermore, we utilized a class panel graph to provide a comprehensive overview of each cluster's characteristics. A class panel graph is a type of visualization that presents multiple graphs or plots in a grid-like layout, where each panel represents a subset of the data, such as a histogram for numerical variables or barplots for categorical variables. By utilizing a CPG, we could examine the unique attributes of each cluster and their relationship to one another.

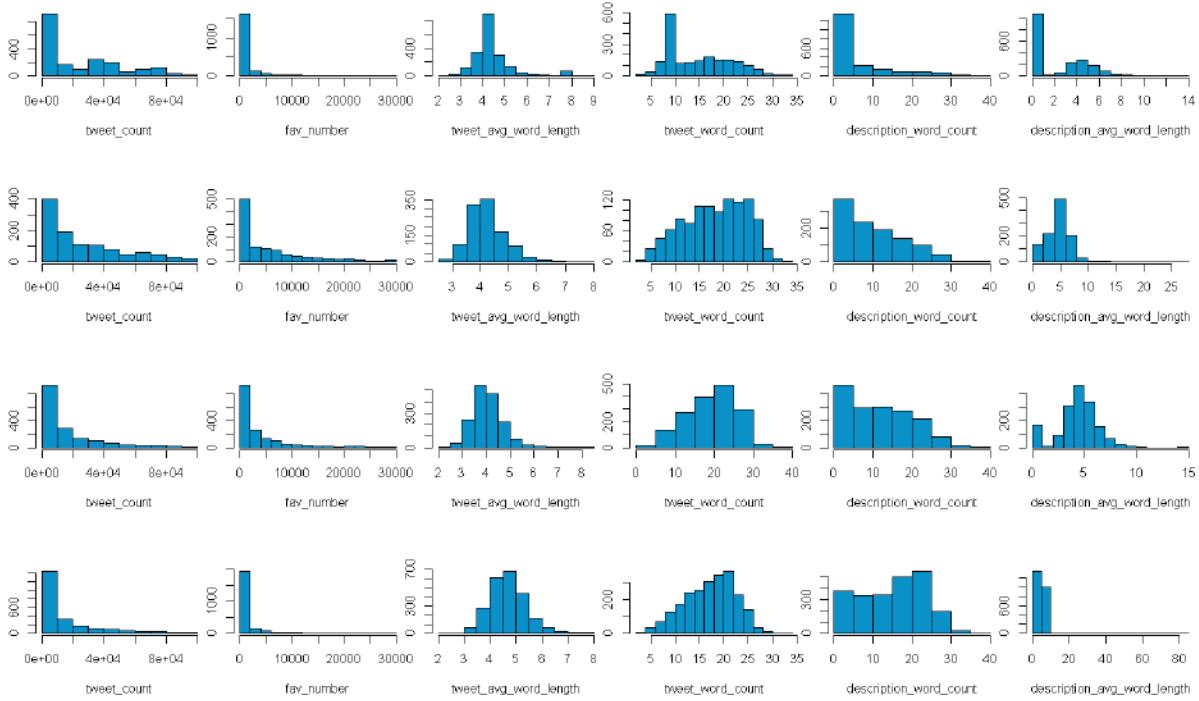
It's important to note that the variable arrangement in a CPG has a significant impact on the cluster's overall interpretation. Thus, we grouped our variables into three sections, mirroring the univariate analysis conducted earlier in this report, all of which focus on user-centered factors. These sections include the user's profile characteristics, such as their personalization preferences; the user's personal information, such as their continent of residence; and the user's behavior, such as the characteristics of their tweets.

As a result, we created the CPG by views presented in the same first as the figure below. The other view are on Annex 4 for better visibility.

User's profile



User's behavior



TLP

Traffic Lights Panel

A traffic light panel is a graphical display that uses the colors of traffic lights (red, yellow, and green) to indicate the status of a set of metrics or indicators.

In a traffic light panel, each metric or indicator is represented by a single cell or box, and the color of the cell indicates its status. Typically, green indicates that the metric is performing well, yellow indicates that there may be some issues or concerns, and red indicates that there is a problem that needs to be addressed.

In clustering, the TLP can be used to easily summarize the range of numeric variables for each cluster into a plot, this will help us determine the quality of the clustering if we see that the ranges are different or not.

In our case, we decided that the status, or the color of each cell, would be determined by where the mean of a variable for a particular cluster falls in the quantiles of that variable. In other words,

first we calculate the quantiles for each variable with the full dataset, then we compute the mean for that variable in each cluster, and then we see in which quantile it lands.

This will help us identify if our clustering algorithm has done a correct job at grouping the variables into separate ranges with each cluster. Then we will also be able to compare each cluster based on this.

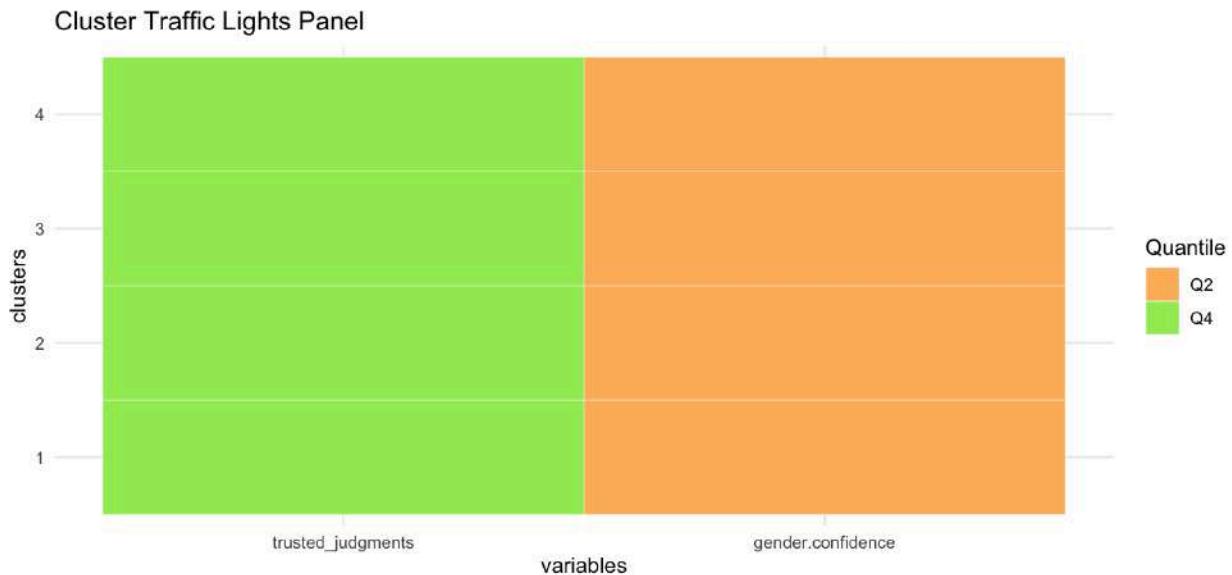
The color code we used initially was red for the 1st quartile, indicating a low value of said numerical variables, yellow for the 2nd and 3rd quartile, and green for the 4th quartile, indicating a higher range of values for that numerical variable in the cluster. However, we noticed that most of the cells in our TLP were yellow, so we decided to add another color, orange to represent the 2nd quartile, which means that only the 3rd quartile will be colored yellow.

Next, we will comment on the results of our TLP looking at the views we constructed. These views include the user's profile characteristics, such as their personalization preferences; the user's personal information, such as their continent of residence; and the user's behavior, such as the characteristics of their tweets. In the case of TLP, we can only use numerical variables as we defined the color criterion as being the belonging of a mean to a quantile, which can only be calculated with numerical values. Therefore the remaining views we can assess with the TLP are User and User's Behaviour.

User

For this view, as we can see in the following plot, the CURE didn't really distinguish the range of these variables in the different clusters, meaning they were probably less relevant in the clustering. As we can see the means for these variables all fall in the same cluster, which means that the points in these clusters are probably spread out. Another possible reason for this is that there is very little variance in those variables, meaning that almost all the points have the same value.

A way to check whether it's one or the other is by using the annotated Traffic Lights Panel, which we will see in the next section. However, for now, we can conclude that the clusters weren't generated based on this view (the users' characteristics).



User's Behaviour

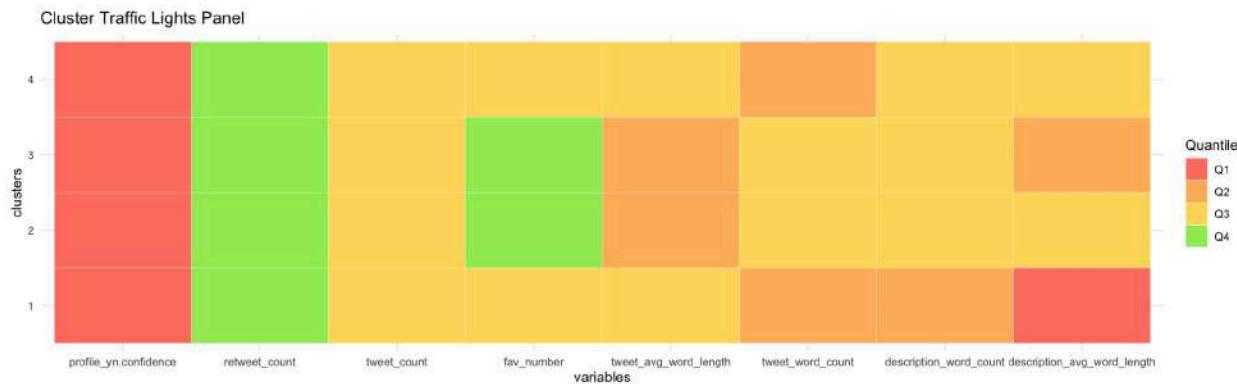
For this view, we will look at if the clustering was based on the variables that explain the user's behaviour. Immediately we can see a little more variation of colors in the plot. The first three variables; profile_yn.confidence, retweet_count and tweet_count, seem to not have been determining factors in this clustering, since all of their means fall within the same quantile. As mentioned before we will also check the annotated-TLP.

On the other hand the rest of the variables seem to be separated into more distinguishable ranges in the clusters. The mean of fav_number, or the number of likes a user has given is in Q3 for clusters 1 and 4 and it's in Q4 for clusters 2 and 3, this means that clusters 2 and 3 consist of individuals that are more active on the platform in terms of liking tweets.

The mean for avg_word_length, or the average word length of the user's tweet is lower for users in clusters 2 and 3, since it falls in the 2nd quartile of the variable. However avg_word_length is higher for the clusters 1 and 4, with their means falling within the 3rd quartile.

Now looking at the description_word_count, or nº of words in the user's description, it seems that all the clusters are pretty similar, with all their means falling within the 3rd quartile, except for cluster 1, that has a lower word count, with the average falling within the 2nd quartile.

Finally, the description_avg_word length, or the average word length of the user's description, is higher for clusters 2 and 4, and lower for clusters 1 and 3, especially for cluster 1.



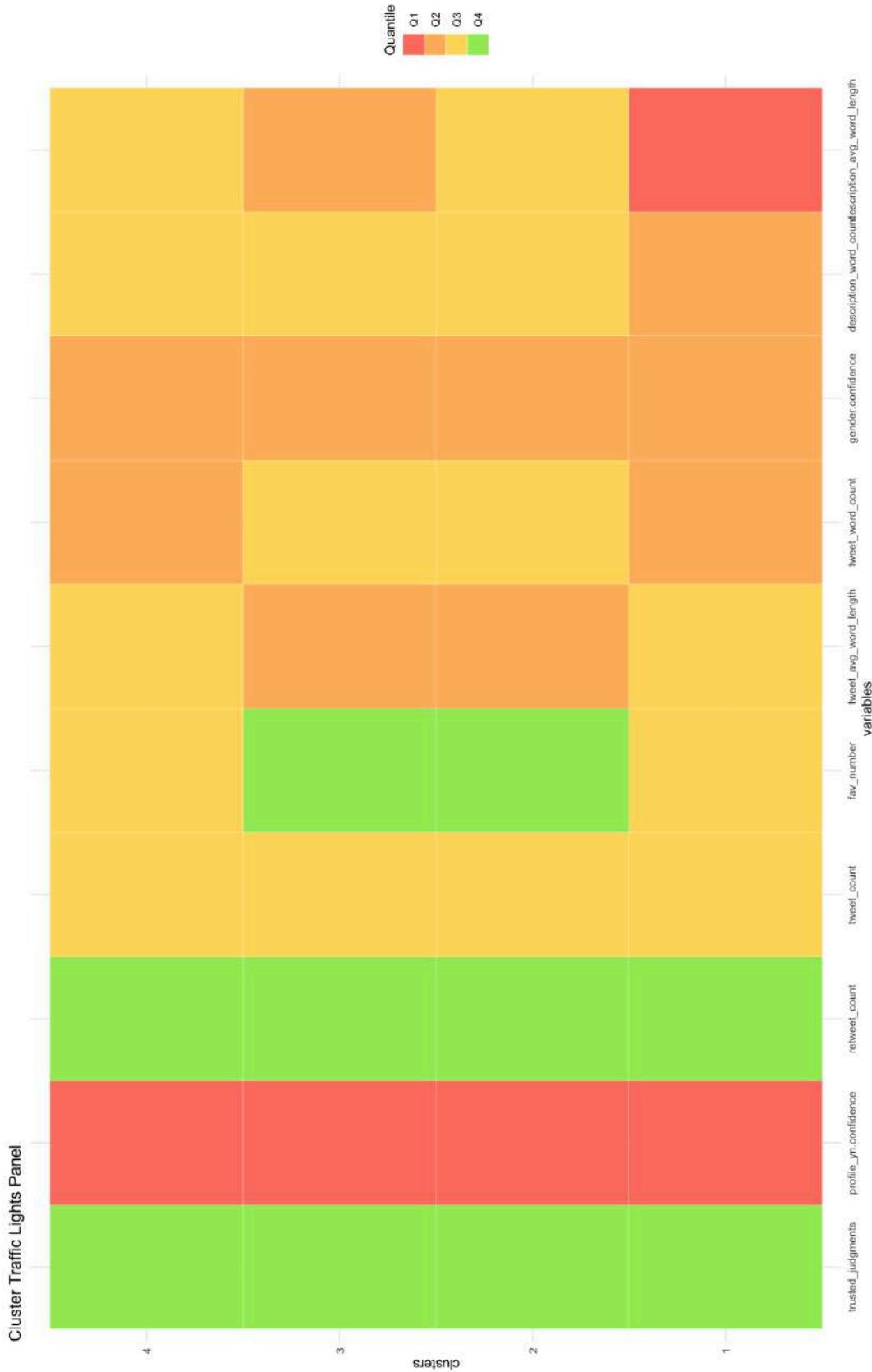
From this analysis, we can conclude that cluster 1 tends to have users that are less active in terms of giving likes, said users use longer words in their tweet, but use less words. And they have the shortest descriptions out of all the clusters, using also the shortest words in them.

Moving on to cluster 2, these users are more active in terms of giving likes, but use shorter words in their tweets, on the other hand, they also use more words in them. They also have some of the longest words in their description.

For cluster 3, these users are also more active in terms of giving likes, use shorter words in their tweets, and use more words in them. They also have smaller words in their descriptions compared to cluster 2.

Finally, for cluster 4, the users are quite similar to cluster 1: they are less active giving likes, have longer words in their tweet, but use less words. But they do use more words and longer in their description as compared to cluster 1.

In the following page we plotted the full TLP and in the following section we will comment the annotated-TLP.



Annotated-Traffic Lights Panel

The Annotated-Traffic Lights Panel, or a-TLP, is similar to the regular TLP except its cells are shaded based on the coefficient of variation (CV) for each variable per cluster. The CV is a statistical measure for relative variability in datasets. The formula for the coefficient of variation is:

$$CV = (\text{standard deviation} / \text{mean}) * 100\%$$

It tends to be a percentage, or ranged in between 0 and 1, but if the variables are not normalized it may need to be standardized dividing by the range of each variable. Even so, the value may surpass 1 if the standard deviation is really high compared to the mean.

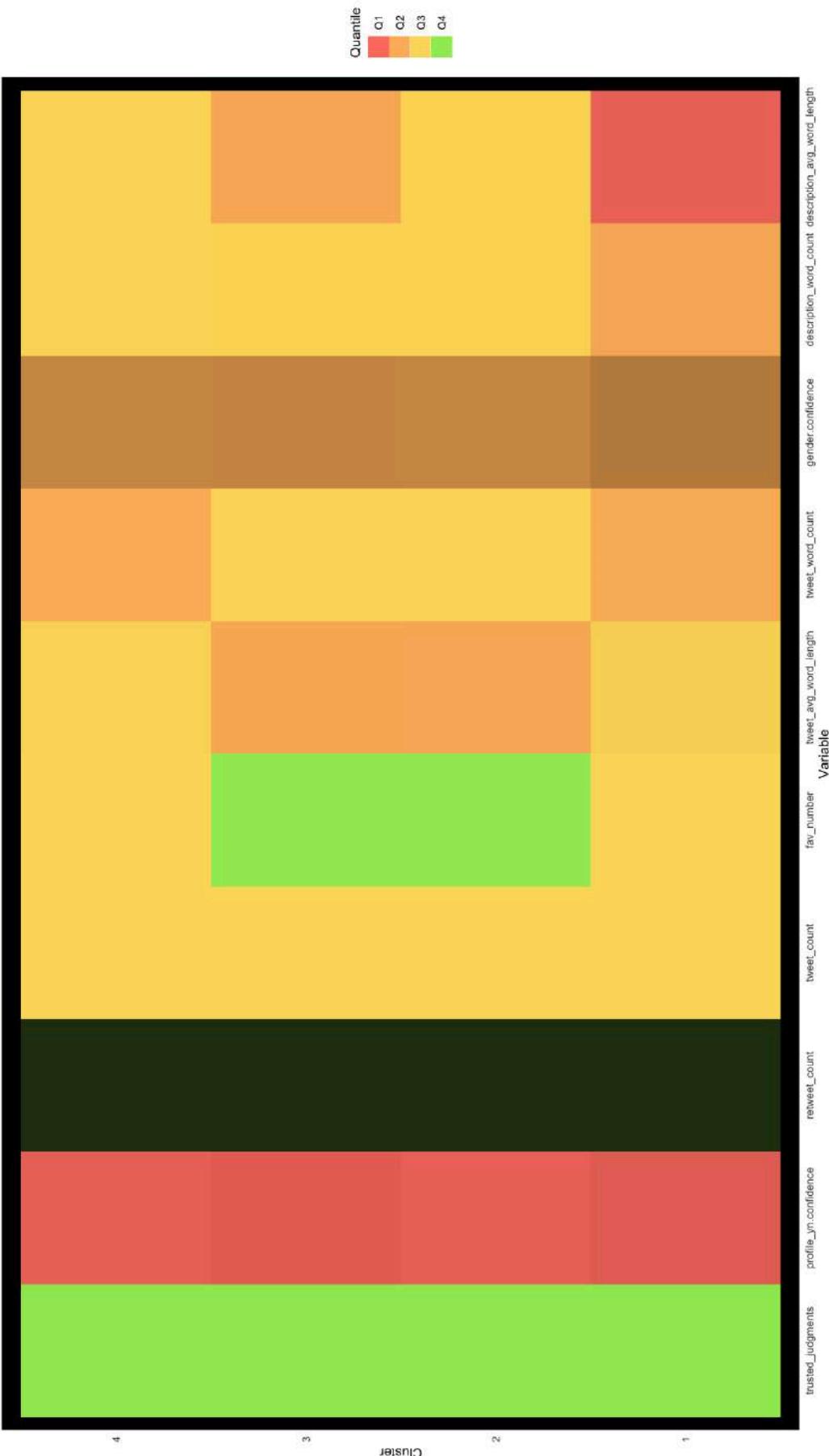
A high coefficient of variation indicates that the standard deviation of a data set is relatively high compared to the mean of the data set. In other words, it means that the data points are more spread out and less clustered around the mean.

Applying this to the TLP means that we will be able to see if the points are all close to the mean or more spread out for each variable and cluster. This will give us useful information, since we will be able to see whether the clustering grouped points of a specific small range of the variable which are in a given quantile (meaning that the clustering separated the data based on the ranges of this variable), or if the clustering grouped very spread out points and the mean coincidentally falls into said quantile (meaning that the clustering wasn't based on this variable, or it had little importance).

In the a-TLP, the darker the cell is the higher the CV is. Meaning that the points for that variable are more spread out in the cluster, which, in its turn, means that the clustering in that cluster likely wasn't based on that variable.

In the following page, we have plotted our a-TLP. We will now discuss the results. First we will comment on the User view variables: trusted_judgments and gender.confidence; as we can see the colors of the trusted_judgments column are very close to the original colors, this means that the points are close together in each cluster, so this likely means that this variable has low variance, since all the clusters are in the same quartile and have low CV. On the other hand, gender.confidence has become pretty dark with the CV values, meaning that there is a wide range of values in each cluster, so the clustering was most likely not based on these variables.

annotated-Traffic Lights Panel



Now talking about the User's Behaviour View, we can see that most of its variables are pretty close to the original color, meaning a low CV for each and that the points are close together, so the clustering was based on them. Nevertheless, there are some exceptions, the main one is retweet_count, which is the darkest column in all of the a-TLP, in other words, each cluster has a very wide range of values of retweet count, so this variable didn't intervene in the creation of the clusters, since the points of each cluster aren't grouped close to the mean and the means are all in the same quartile, meaning there is likely no distinction in the ranges of the variable when clustered.

Some other variables that appear darker are profile_yn.confidence, tweet_avg_word_length and tweet_avg_word_count. This means that we can't completely trust the TLP classification since the points are more spread out than would be ideal and that these variables were less relevant in the clustering.

Bivariate analysis profiling

For the categorical variables we couldn't perform the TLP, so instead we decided to conduct a bivariate analysis, where we looked at the proportion of the categories of each variable per cluster. We decided to look at it by percentage of presence of a category in the cluster, since this will allow a more fair comparison since the clusters are different in size.

Looking at the plot beneath we will analyse the different views for the categorical variables.

User's profile

- **link_basic_color:** for the color of the link, we see that cluster 1 is mostly comprised of blue individuals, and cluster 3 has notably fewer blue individuals in proportion than the rest, with more proportion of red, purple, pink and grey individuals. For clusters 2 and 4 they mostly follow the proportion of the whole population/dataset, although cluster 2 has a higher proportion of red and grey individuals than cluster 4. We can conclude that for cluster 2 and 4 this variable didn't really participate in the separation.
- **sidebar_basic_color:** for the sidebar color we see the same pattern as for the link color. This variable mostly participated in the distinction of clusters 1 and 3.

User

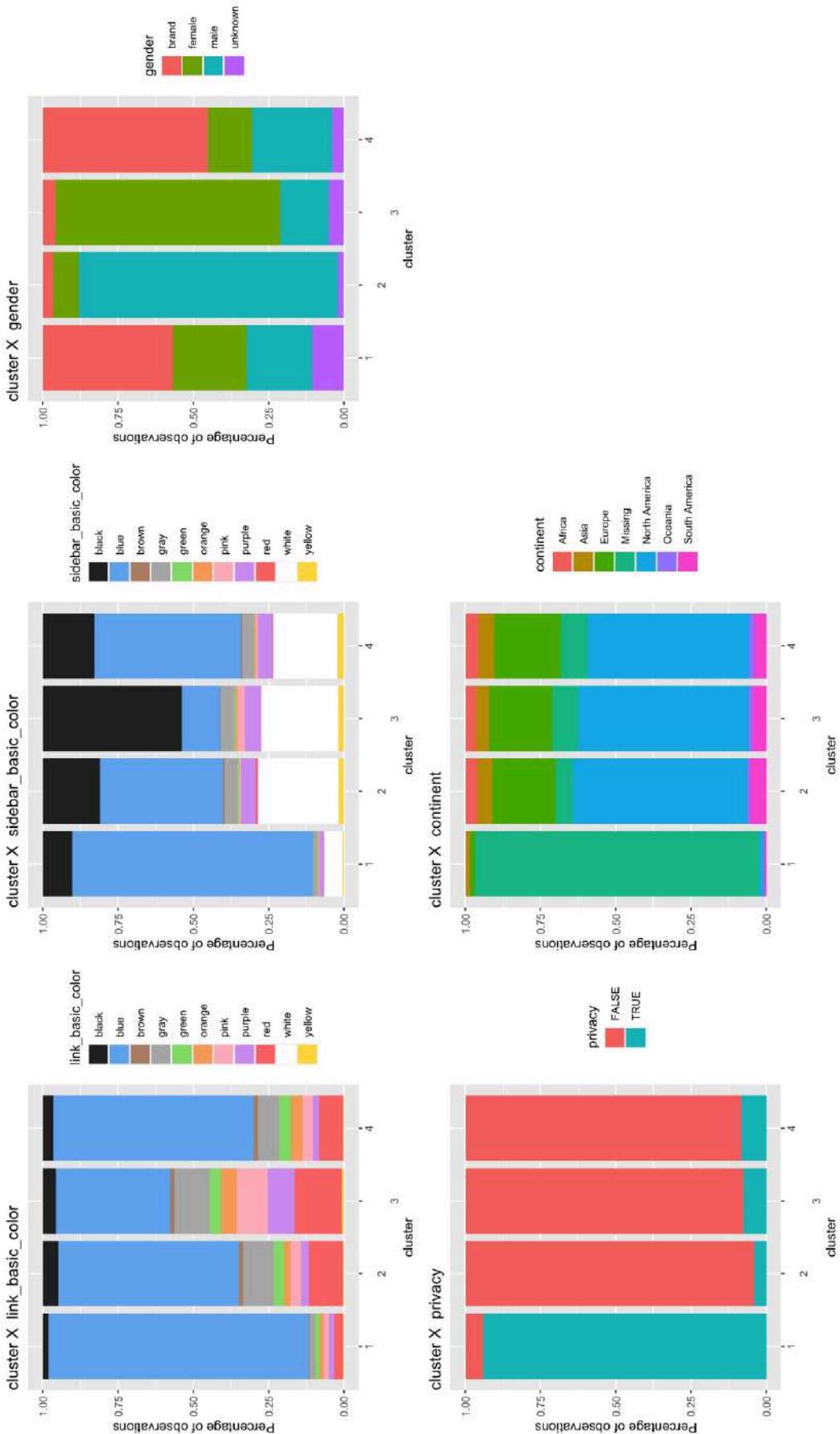
- **gender:** this variable is interesting, in cluster 2 we can see that the vast majority of individuals are male, and for cluster 3 we see the opposite, the majority is females. For cluster 4 the majority is brand, but it is not as overwhelming and same goes for cluster 1.

To conclude, cluster 2 is mostly males, cluster 3 is mostly females, and clusters 1 and 4 are mostly brands but there is a presence of the other genders.

- **continent:** For *continent*, we see that clusters 2, 3, 4 follow similar proportions with very few missings, meaning that these clusters didn't discriminate in terms of this variable. However, cluster 1 is almost completely composed of missing individuals.

User's behaviour

- **privacy:** for behaviour we see a similar pattern as in *continent*, in cluster 1 most of the individuals want privacy, and for the other clusters the vast majority, doesn't mind about privacy and disclosed their location. This similarity between privacy and continent is likely due to the fact that privacy is a feature extracted from location, where if the location was missing privacy was set to TRUE, and the same goes from continent, which was also extracted from the same variable, so naturally if the location was missing, the continent would also be "Missing".



Cluster template

After all the tools have been implemented in order to aid with the profiling of the clusters obtained from CURE, we can now proceed to make a template for the characterization of the main features that distinguish each cluster.

This template will be characterized by a grid with all the variables comprising the observations made prior about each cluster so that we can get a general idea on each one of them. This template will also be grouped by views.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
User's profile				
Link basic color	Blue	Black	White Purple	Brown Gray Orange
Sidebar basic color	Blue	Brown	Pink Black Grey	Black Orange Red
User's personal information				
Gender	Brand or Unknown	Male	Female	Brand or male
Gender.confidence	Lowest	Mid-high	Mid-high	Mid-high
Continent	Missing	South America North America	North America South America	Oceania Africa Europe
Privacy	High	Low	Low	Low
User's behavior				
Tweet count	High	High	Low	Low
Favorite number	Low	Highest	High	Low

Tweet average word length	No great distinction	No great distinction	No great distinction	Highest
Tweet word count	Lowest	No great distinction	No great distinction	No great distinction
Description word count	Lowest	Mid-low	Mid-high	Highest
Description average word length	Lowest	High	High	High

So we could describe each cluster the following way:

- **Cluster 1:** Predominantly composed of Missing continent profiles made by either a brand or unknown, with a high privacy who write average twitter posts in terms of word length but lower amount of words and a low gender confidence. It also likes to keep its description at a minimum in all aspects.
- **Cluster 2:** A male profile with average twitter posts in terms of word length and word count with a slightly higher gender confidence who doesn't value privacy. High in favorite number.
- **Cluster 3:** A female profile who chose, and has low privacy. Its tweets are average in word length and word count. Usually from North and South America.
- **Cluster 4:** A brand or male profile with high word length, but average traits in terms of word count. It also possesses a long and wordy description and low privacy setting.

DBSCAN

Filtering out irrelevant variables

In DBSCAN we also tried to implement a random forest to filter out irrelevant variables but encountered that the resulting accuracy from it was subpar, meaning that the results of the random forest weren't reliable enough to be taken into consideration.

For that reason, we resorted to the univariate analysis. We had previously seen that some variables present almost no variance if not any, meaning that they can't be an essential factor to the clustering. Thus, we decided to eliminate these kind variables. This resulted in the elimination of the variables unit_state, trusted_judgments, profile_yn.confidence and retweet_count.

Bivariate plots and inference tests

After filtering out the non-relevant variables, we proceed to do a series of bivariate plots to represent the distributions of both numerical variables and the different modalities for categorical variables.

While doing so, we also conducted a series of inference tests to ensure that the results for each cluster are actually different from each other or if any observed differences are simply due to chance. The tests done are for quantitative variables the p-values of an ANOVA, Kruskal-Wallis and Valors tests. For the qualitative variables we will make use of the p-value of a chi-squared test.

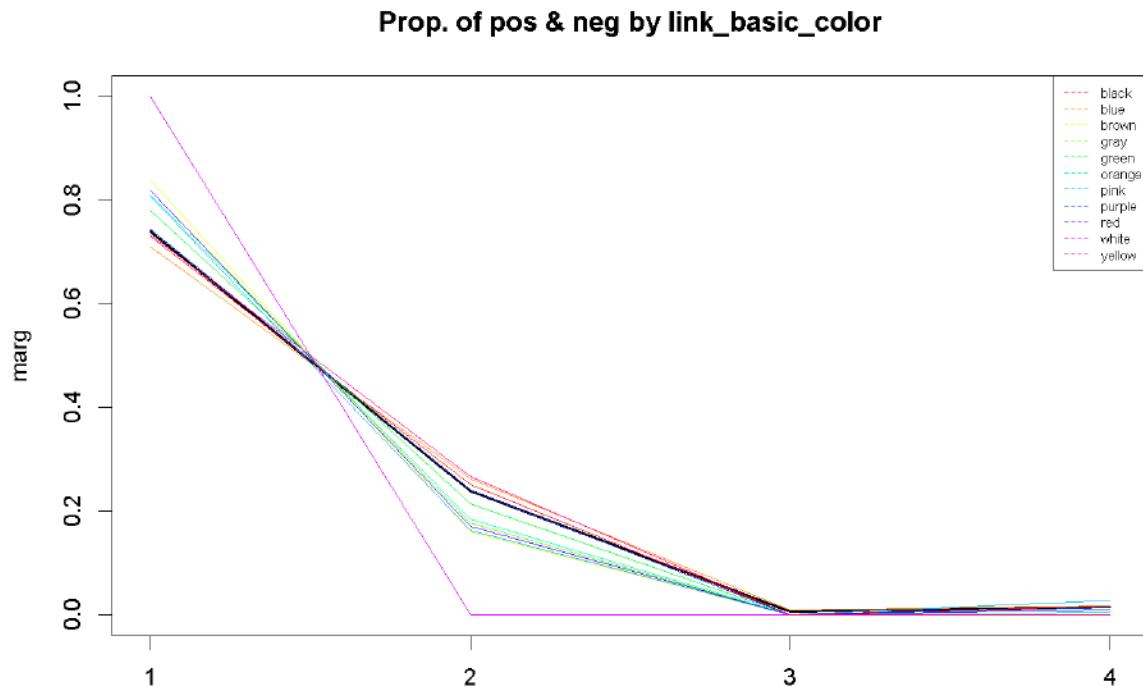
In all those cases, a p-value that is lower than the 0.05 threshold will help reject the null hypothesis that there is no significant difference between the clusters and conclude that the observed differences are statistically significant.

This procedure will depend on the type of variable:

- For quantitative variables, we will use a boxplot for each cluster and a bar chart showing the mean of each cluster.
- For qualitative variables, we will use a graph showing the probability of each category being in a cluster. In addition, we will also use two bar charts to see the distribution of each variable in the clusters, both will be bar charts, and one will be a stacked chart, in other words, it will be a single bar per cluster, where each bar will be subdivided by

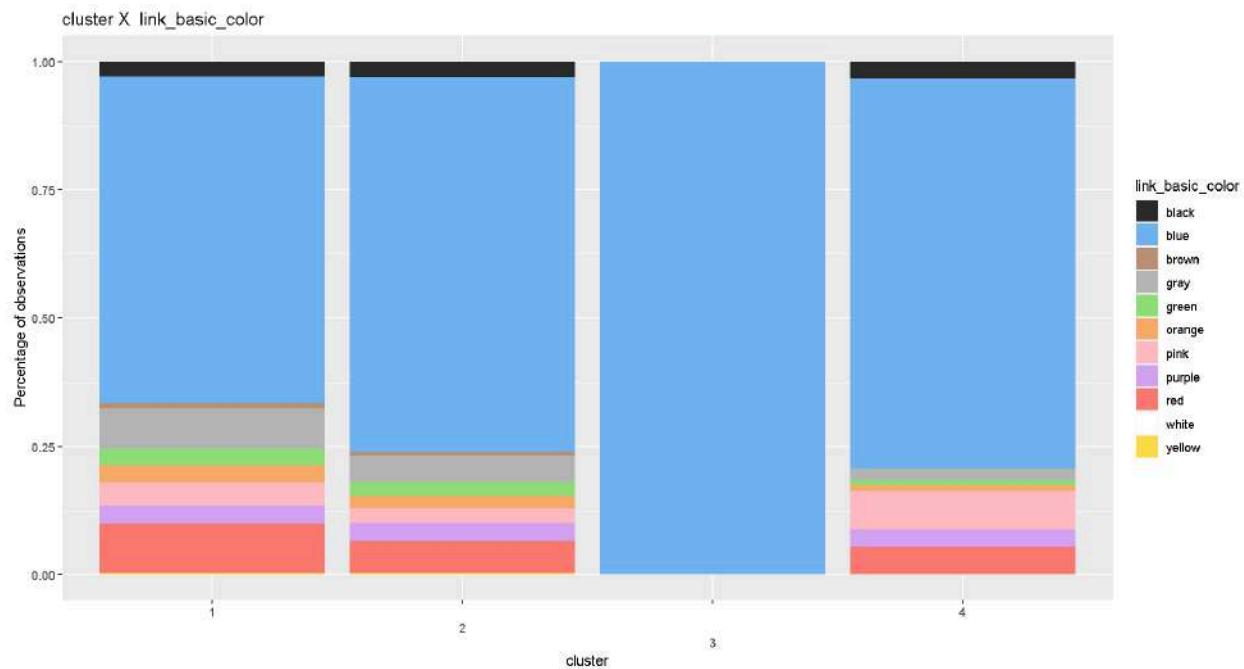
categories. The other graph will be a dodge bar chart, i.e. in each cluster there will be one bar representing each category.

Link basic color



The chart above shows the probability that each category of the link basic color variable belongs to a cluster. In the chart, we can see that the first cluster contains a higher percentage of the color white. On the other hand, in the other clusters, we see the opposite as the probability now is around 0. It can also be seen how most colors with the exception of black, blue, and yellow are at a lesser probability percentage in the second cluster and the complimentary happens in the first.

As we have observed considerable differences, we have added an additional chart to allow us to better observe these.



In the bar charts, we can see that the second cluster has more percentage of pink color and a slightly smaller percentage of orange. In contrast, the first cluster has more percentage of brown color.

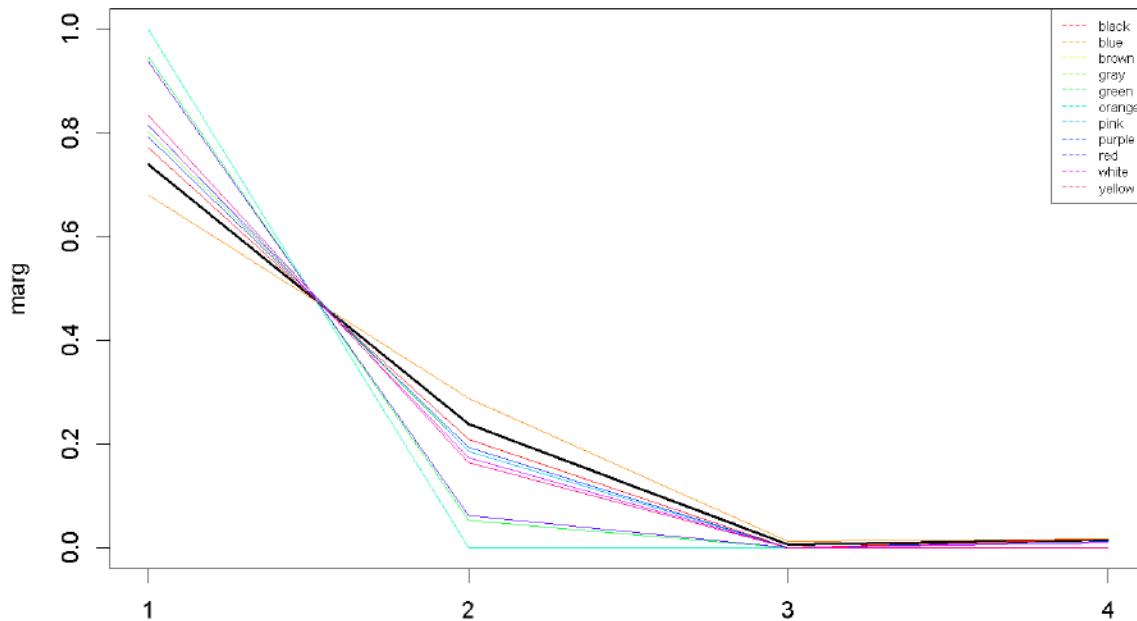
We have also conducted a Chi-square test between cluster and link basic color to determine whether there is a significant association between these two categorical variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the variables.
- H_1 : significant association between the variables.

Chi-squared	
P-value	8.98e-07

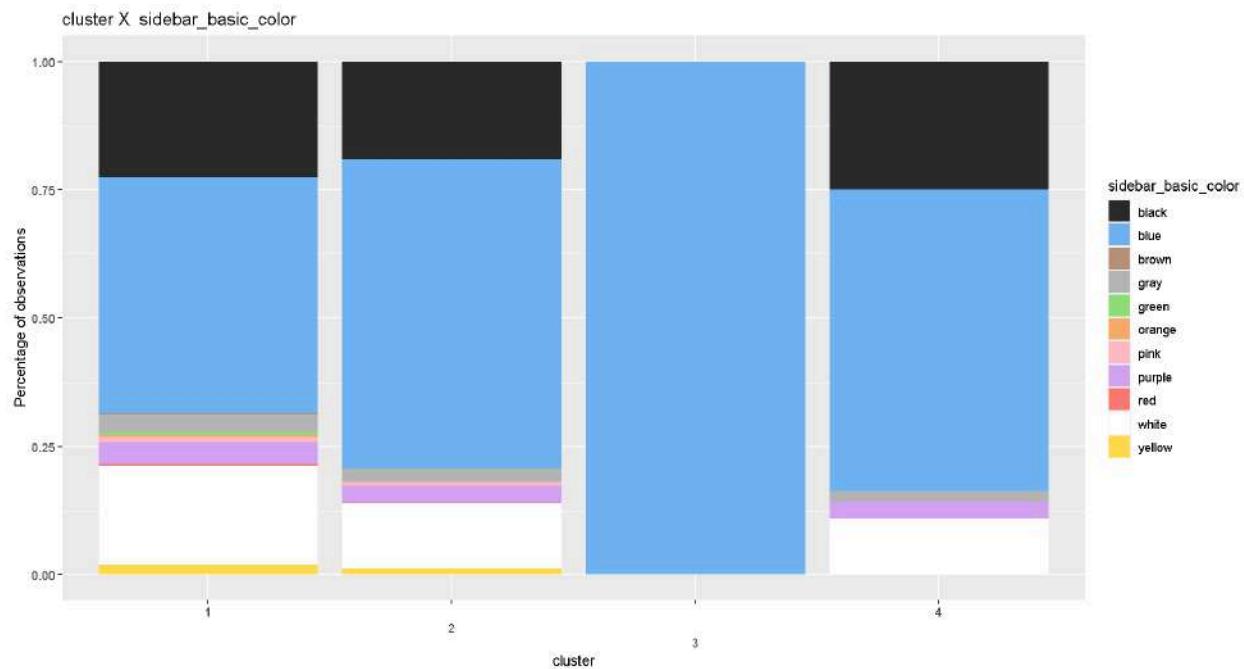
Sidebar basic color

Prop. of pos & neg by sidebar_basic_color



The chart above shows the probability that each category of the sidebar basic color variable belongs to a cluster. In the chart, we can see that the first cluster contains a higher percentage of the colors orange, grey, and red and a lower percentage of blue. It can also be seen in the second cluster it has the complementary color probabilities to the first.

As we have observed considerable differences, we have added an additional chart to allow us to better observe these.



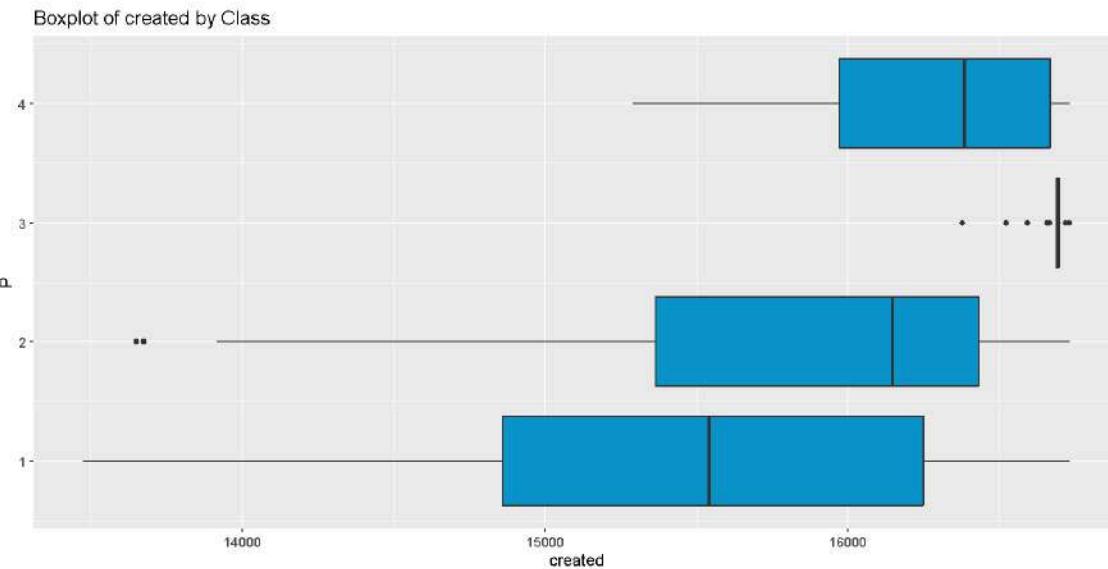
In the bar charts, we can see that the first cluster has more percentage of pink, black, and white color. In contrast, the third cluster has no percentage of black color.

We have also conducted a Chi-square test between cluster and sidebar basic color to determine whether there is a significant association between these two categorical variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the variables.
- H_1 : significant association between the variables.

	Chi-squared
P-value	2.2e-16

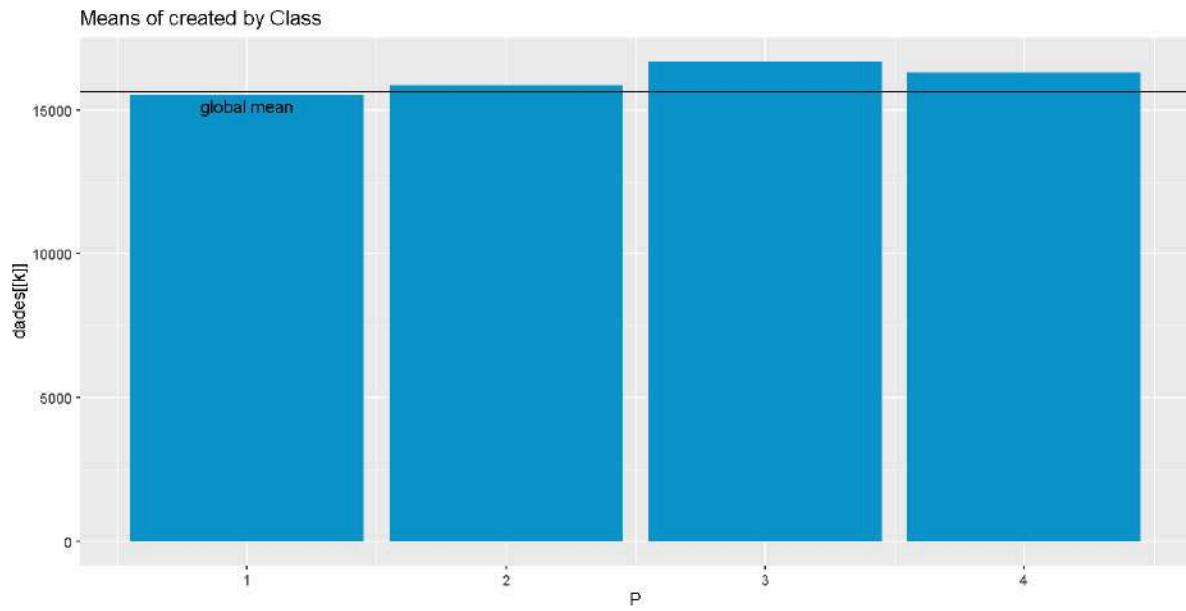
Created



Upon analyzing its distribution, it becomes clear that there are notable differences among the clusters.

The first and second clusters display a similar pattern, with the only differences being the length of the left whisker and median. The third cluster, however, has a notable difference of higher values compared to the rest and a much smaller quartile range.

The fourth cluster also presents the same negative skewness as the first two clusters but with a smaller quartile range that is further pushed to the right, and a smaller left whisker.



Furthermore, upon examining the mean of each cluster, it is evident that all clusters present a mean value close to the global mean. The only notable exception present from that rule is the third cluster has a slightly higher mean value.

To be more precise, the mean of all clusters in order are 15540, 16147, 16695 and 16297 respectively.

We have also conducted a ValorsTest, ANOVA and Kruskal-Wallis tests between cluster and created to determine whether there is a significant association between these two categorical variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

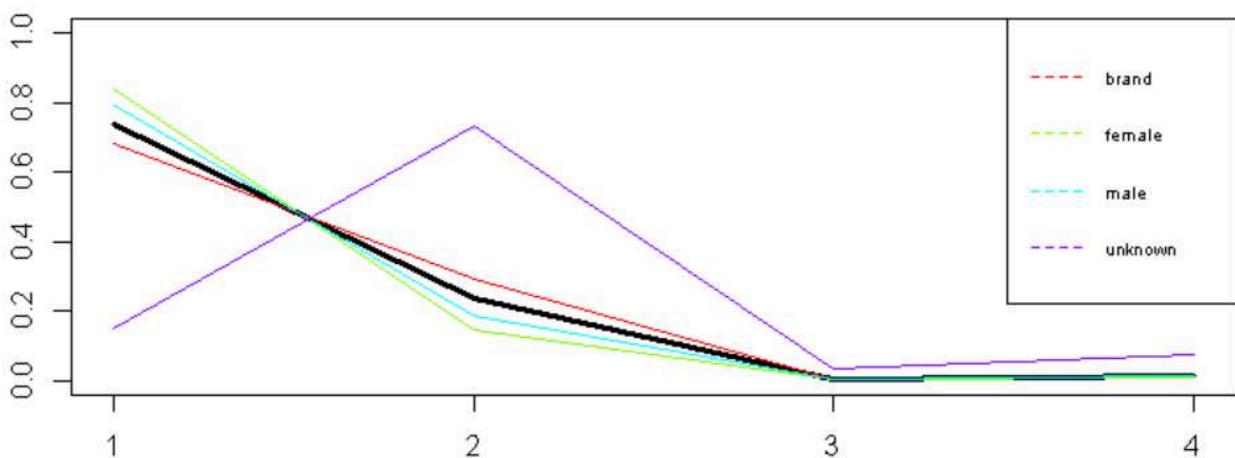
- H_0 : no significant association between the variables.
- H_1 : significant association between the variables.

	ValorsTest	ANOVA	Kruskal-Wallis
P-value	0.000000e+00 6.299774e-39	7.6705080081545e-1 78	2.44600745985682e-76

	1.024954e-15		
	1.595732e-16		

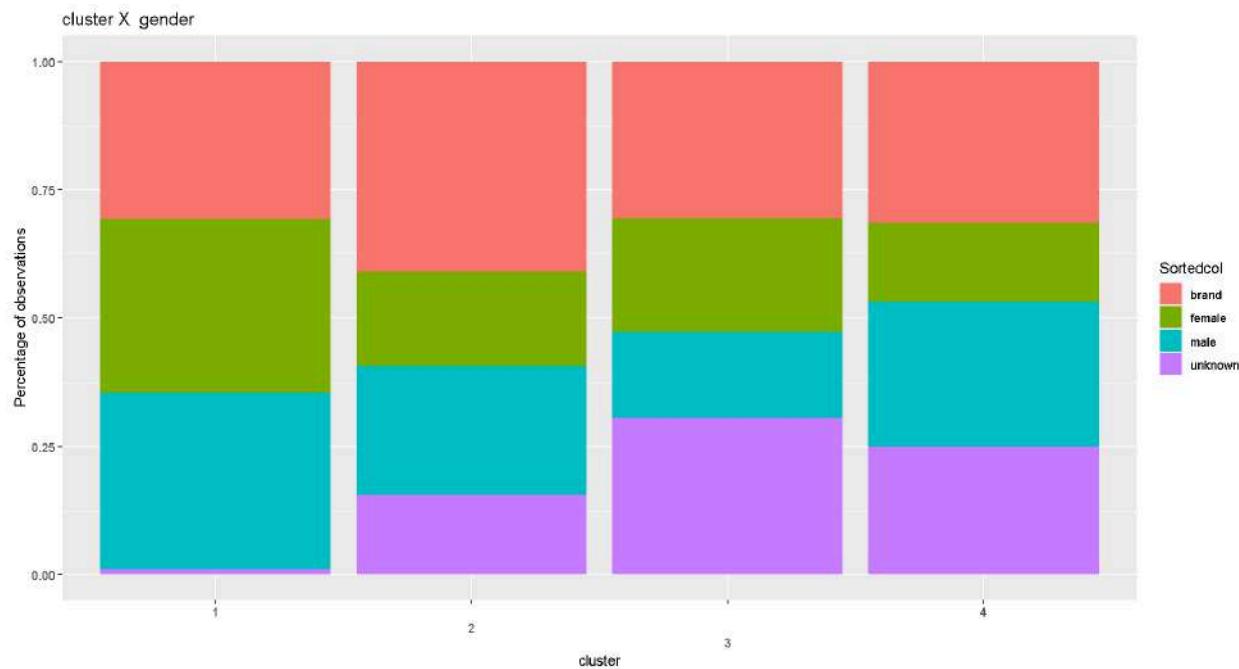
Gender

Prop. of pos & neg by gender



The chart above shows the probability that each category of the gender variable belongs to a cluster. In the chart, we can see that the first cluster contains a lower percentage of users of unknown gender. We can also see how the most predominant genders are male and female with a percentage over 80%. On the other hand, in the second cluster, we see that the opposite is true. It can also be seen how the first and second clusters are flipped versions of each other.

As we have observed considerable differences, we have added an additional chart to allow us to better observe these.



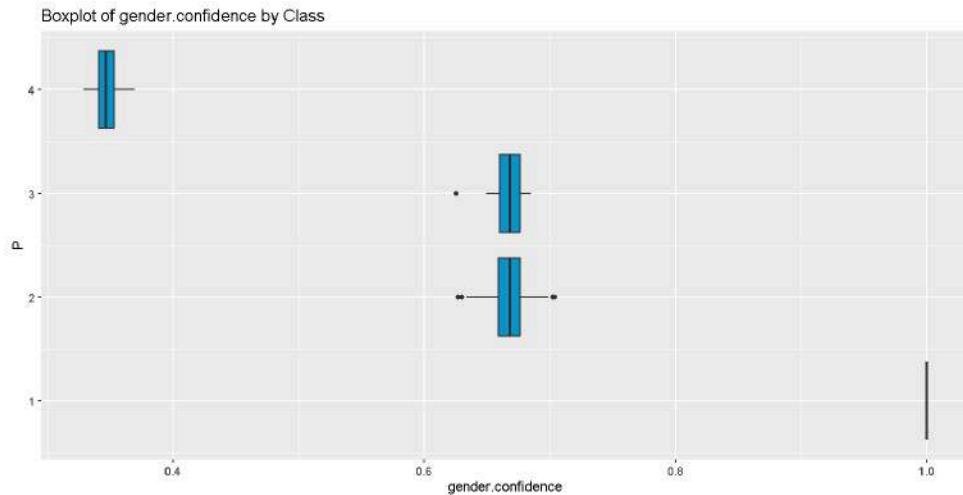
In the bar charts, we can see that the first cluster has a higher proportion of both male and female as mentioned prior. The highest percentage of brand is found on the second cluster. It is also noticeable that the unknown modality is mostly predominant on cluster 3.

We have also conducted a Chi-square test between cluster and gender to determine whether there is a significant association between these two categorical variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the variables.
- H_1 : significant association between the variables.

	Chi-squared
P-value	2.2e-16

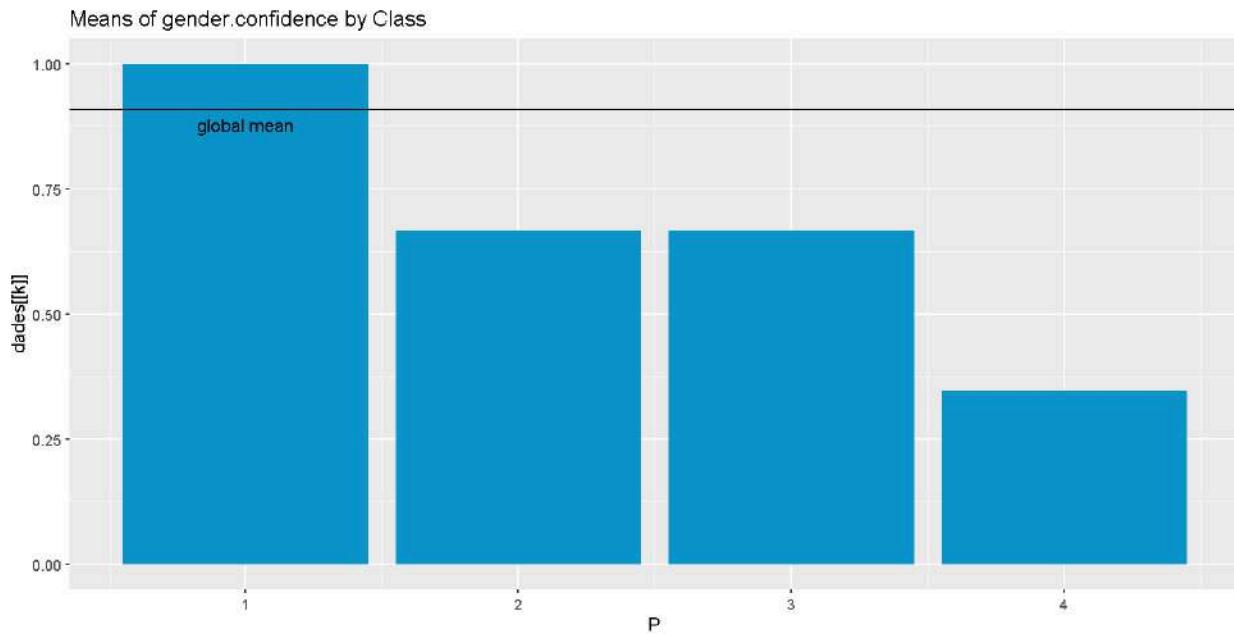
Gender.confidence



Upon analyzing its distribution, it becomes clear that there are notable differences among the clusters. It is also noticeable the three batch pattern previously mentioned in the univariate analysis.

The second and third clusters display a similar pattern at the center of the possible range of values, with the only difference being the length of the whiskers. The first cluster, however, has a notable difference of higher values compared to the rest and a much smaller quartile range.

The fourth cluster presents a slightly smaller interquartile range that is pushed to the lowest values.



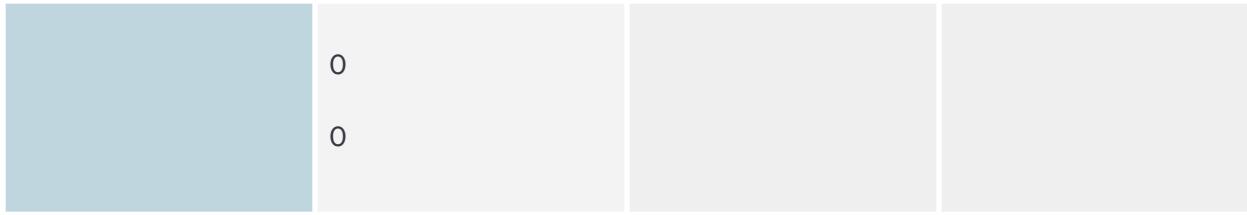
Furthermore, upon examining the mean of each cluster, it is evident that the fourth cluster is comprised of individuals with considerably lower gender confidence than all others. What's more, the first cluster presents a global surpassing mean, while the second and third clusters' means are slightly below it.

To be more precise, the mean of all clusters in order are 1, 0.6673, 0.667 and 0.3473 respectively.

We have also conducted a ValorsTest, ANOVA and Kruskal-Wallis tests between cluster and gender confidence to determine whether there is a significant association between these two variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

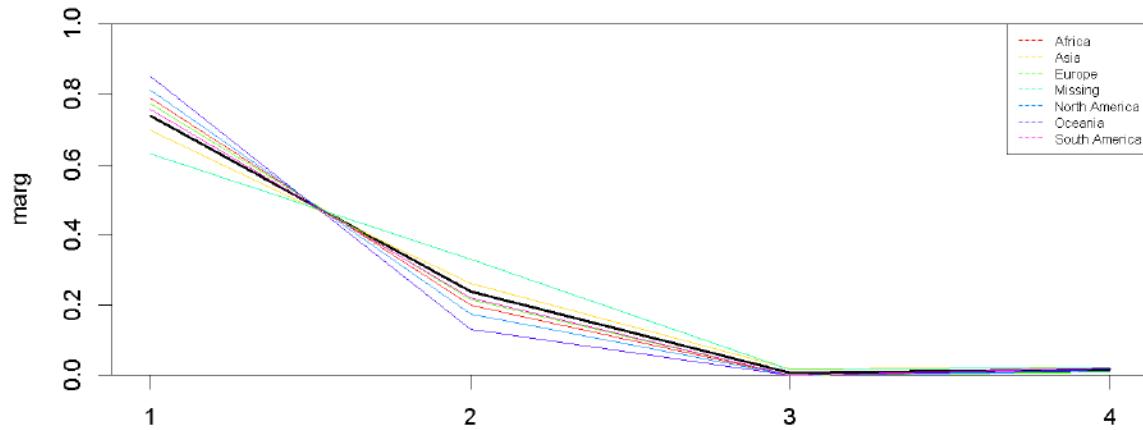
- H_0 : no significant association between the variables.
- H_1 : significant association between the variables.

	ValorsTest	ANOVA	Kruskal-Wallis
P-value	0 0	0	0



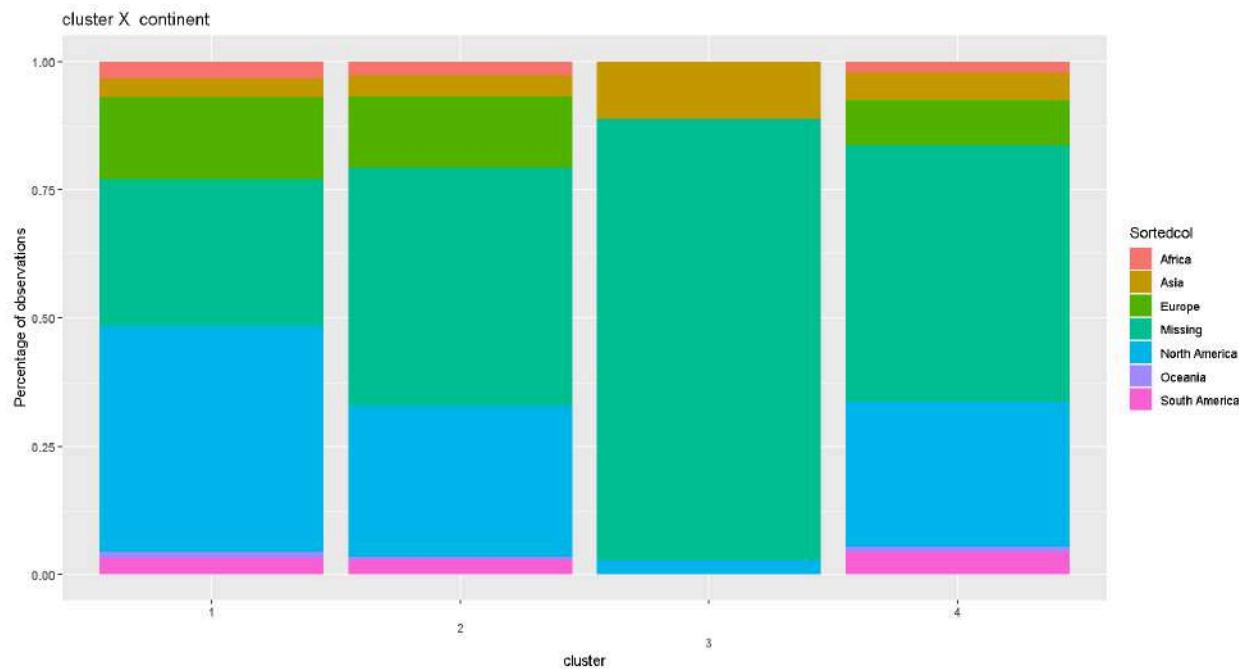
Continent

Prop. of pos & neg by continent



The chart above shows the probability that each category of the continent variable belongs to a cluster. In the chart, we can see that the first cluster contains a lower percentage of users not disclosing their location and Asia. We can also see how the most predominant continents are Oceania and North America with a percentage over 80%. On the other hand, in the second cluster, we see that the opposite is true. It can also be seen how the first and second clusters are flipped versions of each other.

As we have observed considerable differences, we have added an additional chart to allow us to better observe these.



In the bar charts, we can see that the first cluster has a higher proportion of North America, not disclosing their location and Africa. In contrast, on the rest of the clusters, it appears that they have a higher proportion of not disclosing their location distribution. Only in the third cluster, there's more presence of Asia compared to other modalities.

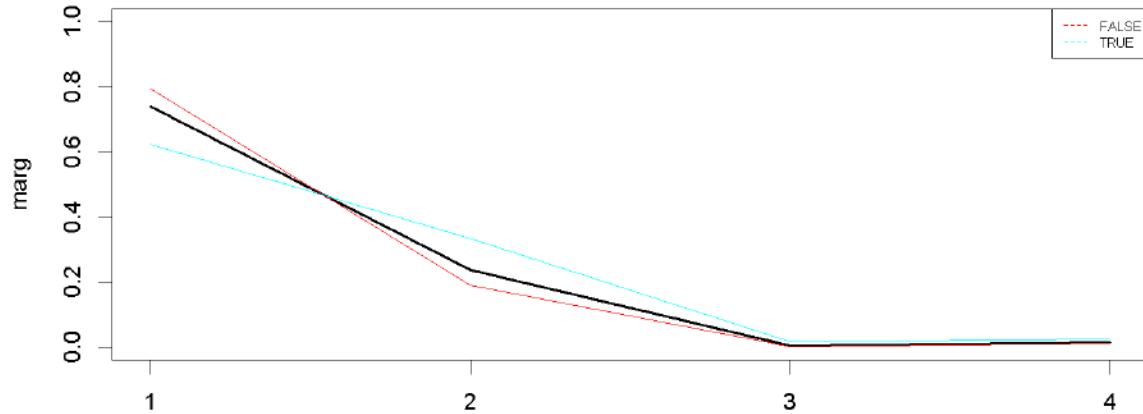
We have also conducted a Chi-square test between cluster and continent to determine whether there is a significant association between these two categorical variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the variables.
- H_1 : significant association between the variables.

	Chi-squared
P-value	2.2e-16

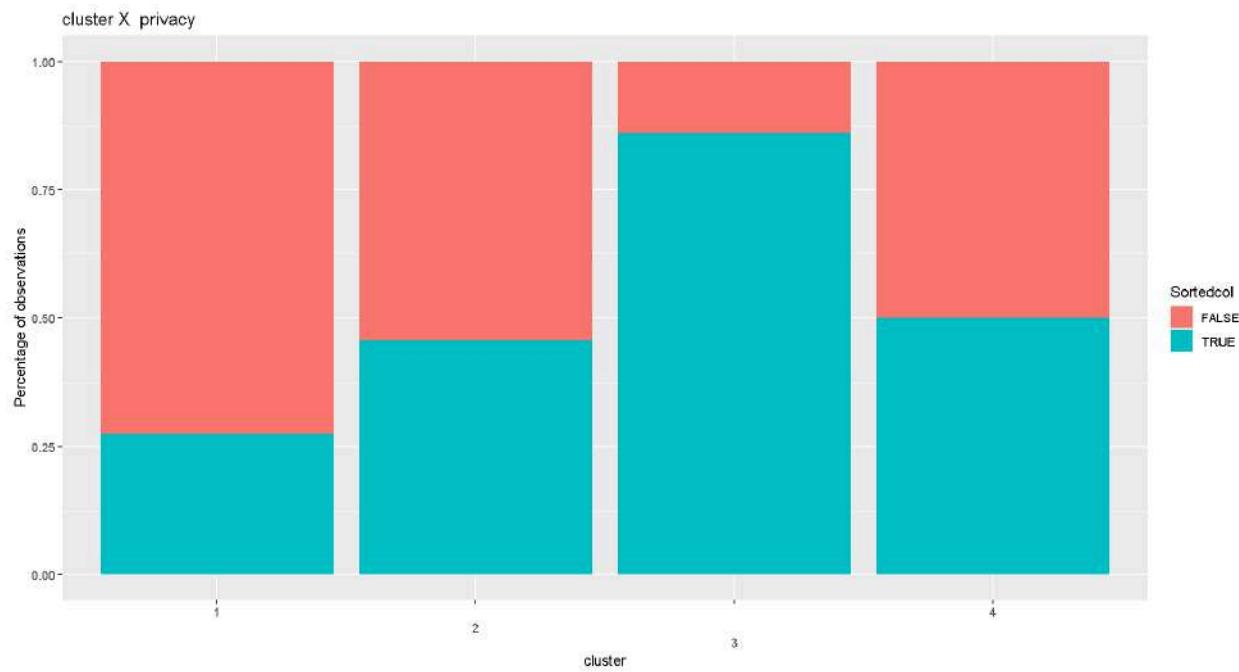
Privacy

Prop. of pos & neg by privacy



The chart above shows the probability that each category of the privacy variable belongs to a cluster. In the chart, we can see that the first cluster contains a lower percentage of not disclosing personal information. On the other hand, in the other clusters, we see that the opposite is true. It can also be seen how the first and second clusters are flipped versions of each other.

As we have observed considerable differences, we have added an additional chart to allow us to better observe these.



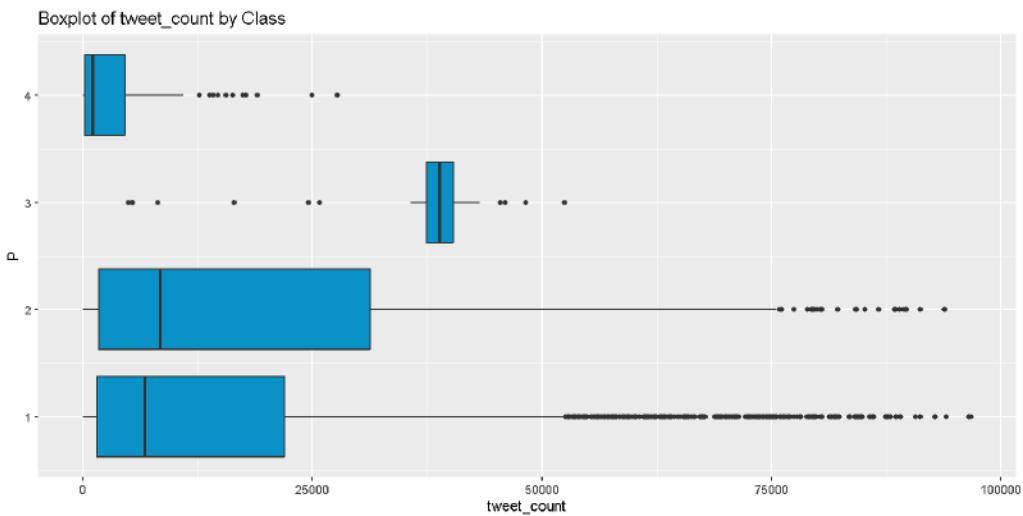
In the bar charts, we can see that the first cluster has a lower percentage of privacy, as non-privacy occupies two-thirds of the overall barplot. In contrast, the second and fourth clusters appear to have a 50/50 distribution. Only in the third cluster there's more presence of privacy compared to the other modality.

We have also conducted a Chi-square test between cluster and privacy to determine whether there is a significant association between these two variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the variables.
- H_1 : significant association between the variables.

	Chi-squared
P-value	2.2e-16

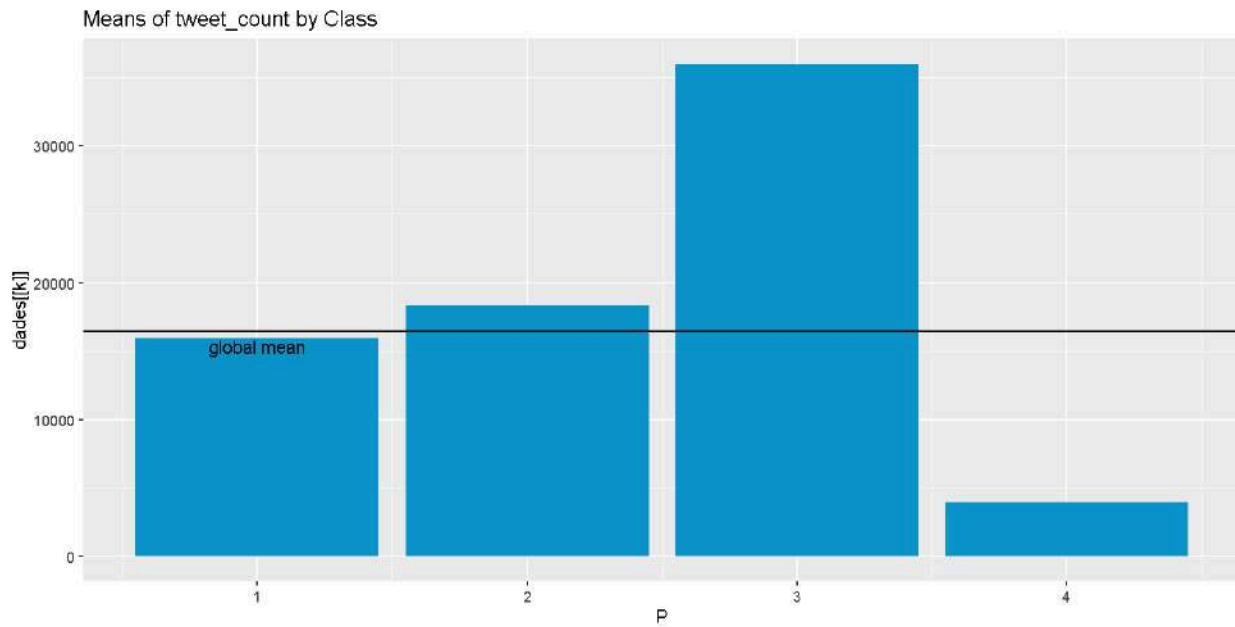
Tweet count



Upon analyzing its distribution, it becomes clear that there are notable differences among the clusters.

The first and second clusters display a similar pattern, with the only differences being the length of the right whisker and the median. The third cluster, however, has a notable difference of higher values compared to the rest and a much smaller quartile range.

The fourth cluster also presents the same positive skewness as the first two clusters but with a smaller quartile range.



Furthermore, upon examining the mean of each cluster, it is evident that the third cluster is comprised of individuals with a considerably higher tweet count than all others and the fourth the opposite. The other clusters left are just a global mean level.

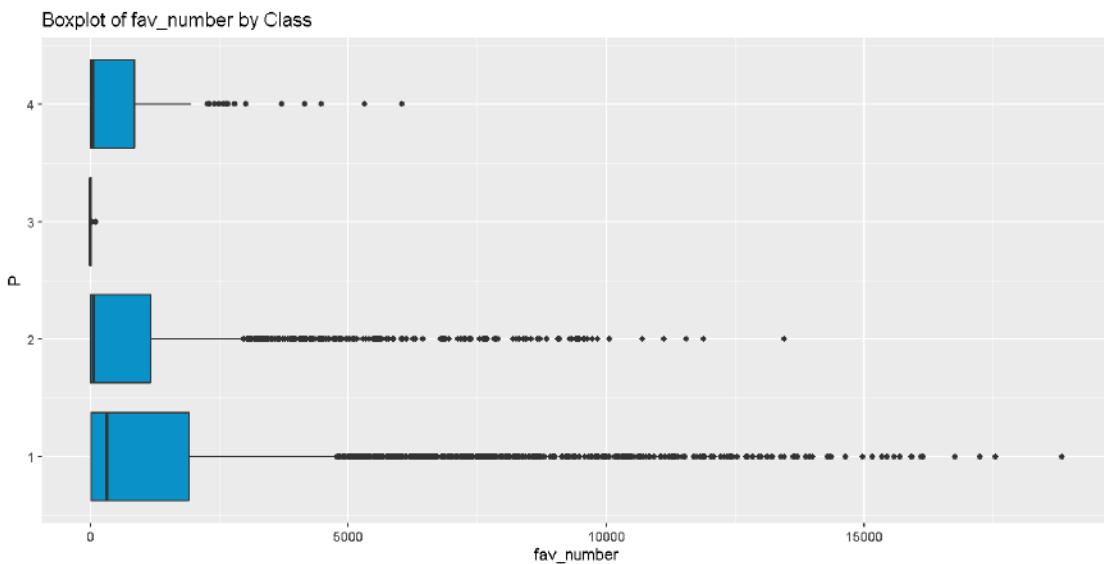
To be more precise, the mean of all clusters in order are 6742, 8449, 38858 and 1130 respectively.

We have also conducted a ValorsTest, ANOVA and Kruskal-Wallis tests between cluster and tweet count to determine whether there is a significant association between these two variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the variables.
- H_1 : significant association between the variables.

	ValorsTest	ANOVA	Kruskal-Wallis
P-value	8.347790e-04 5.118399e-05 8.948527e-09 3.936587e-09	6.00890757215565e-46	1.54463566759605e-23

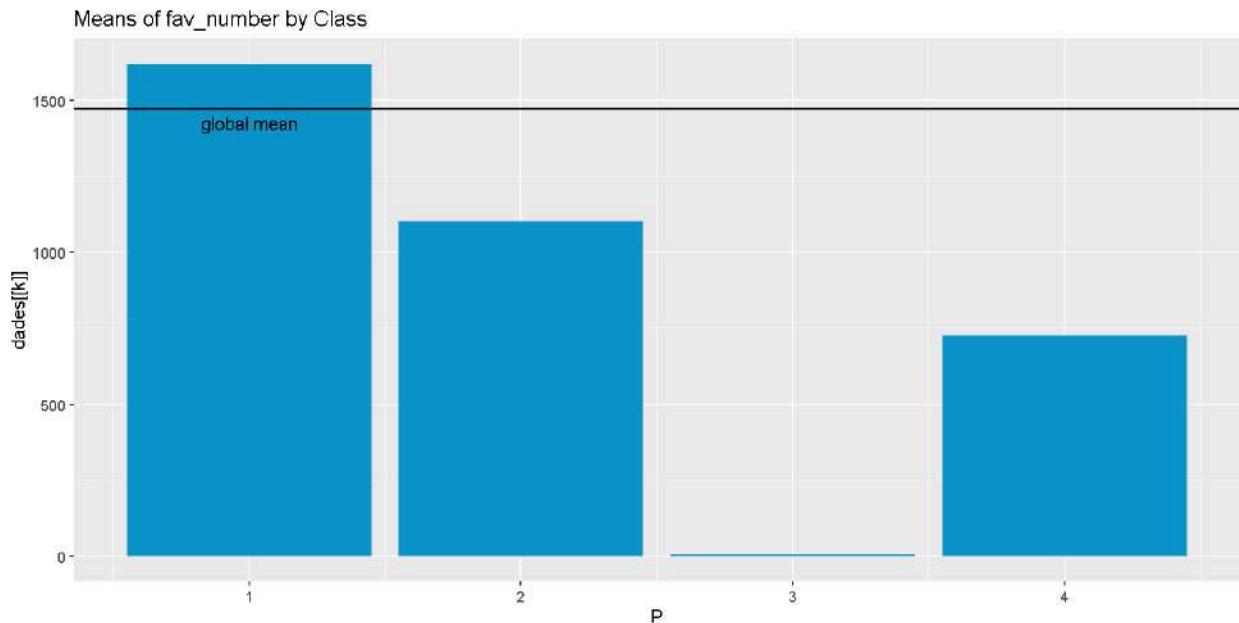
Favorite number



Upon analyzing its distribution, it becomes clear that there are notable differences among the clusters.

The second and fourth clusters display a similar pattern, with the only differences being the length of the right whisker and outliers. The third cluster, however, has a notable difference of lower values compared to the rest and a much smaller quartile range.

The first cluster also presents the same positive skewness as the first two clusters but with a bigger quartile range, right whiskers, and outlier trail.



Furthermore, upon examining the mean of each cluster, it is evident that the third cluster is comprised of individuals with a considerably lower favorite count than all others. What's more, the first cluster presents a global surpassing mean, while the other cluster left are slightly below average.

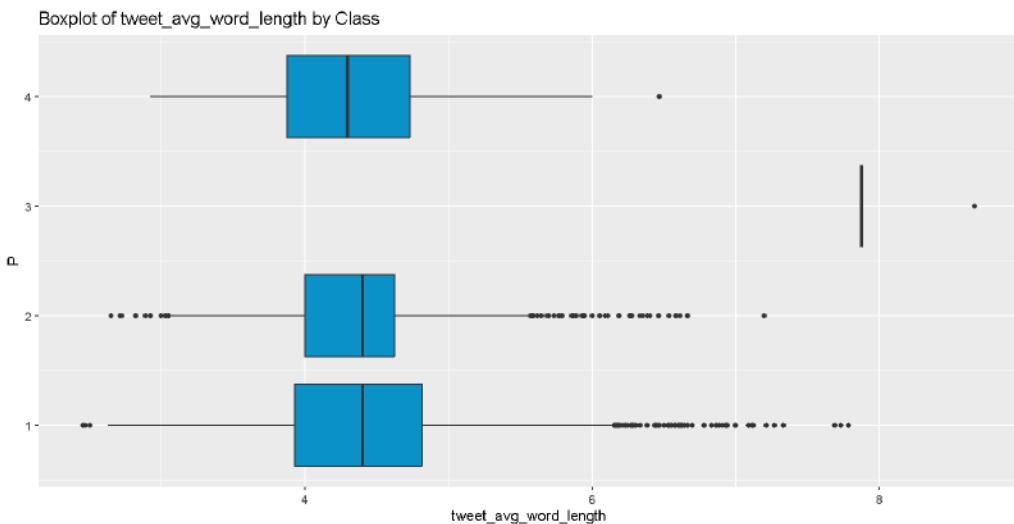
To be more precise, the mean of all clusters in order are 1619, 1101, 6.639 and 725.4 respectively.

We have also conducted a ValorsTest, ANOVA and Kruskal-Wallis tests between cluster and favorite number to determine whether there is a significant association between these two variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the variables.
- H_1 : significant association between the variables.

	ValorsTest	ANOVA	Kruskal-Wallis
P-value	1.296330e-13 4.864684e-10 3.595991e-04 2.816377e-03	1.02900360227989e-154	1.65514837597303e-32

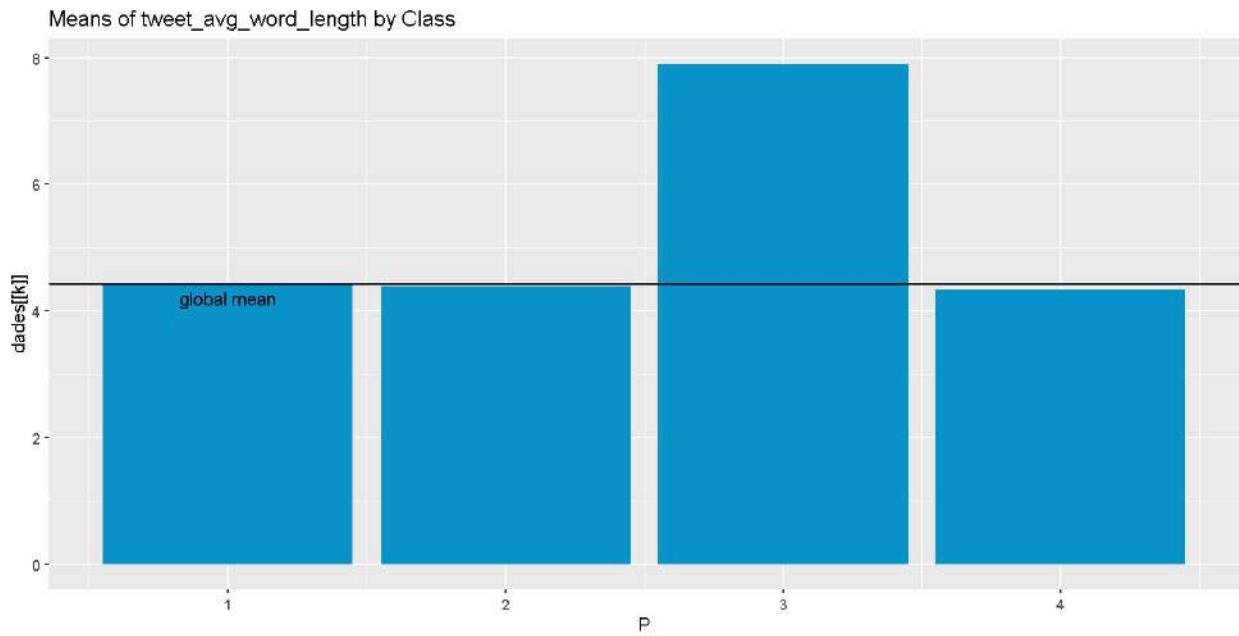
Tweet average word length



Upon analyzing its distribution, it becomes clear that there are notable differences among the clusters.

The first and second clusters display a similar pattern, with the only difference being the length of the whiskers. The third cluster, however, has a notable difference of higher values compared to the rest and a much smaller quartile range.

The fourth cluster also presents the same negative skewness as the first two clusters but with a median that is of a slightly lower value.



Furthermore, upon examining the mean of each cluster, it is evident that all clusters except the third one present a mean value that is identical to the global mean. The only notable exception present from that rule, as mentioned prior, is the third cluster which has a much higher mean value.

To be more precise, the mean of all clusters in order are 4.415, 4.384, 7.897 and 4.331 respectively.

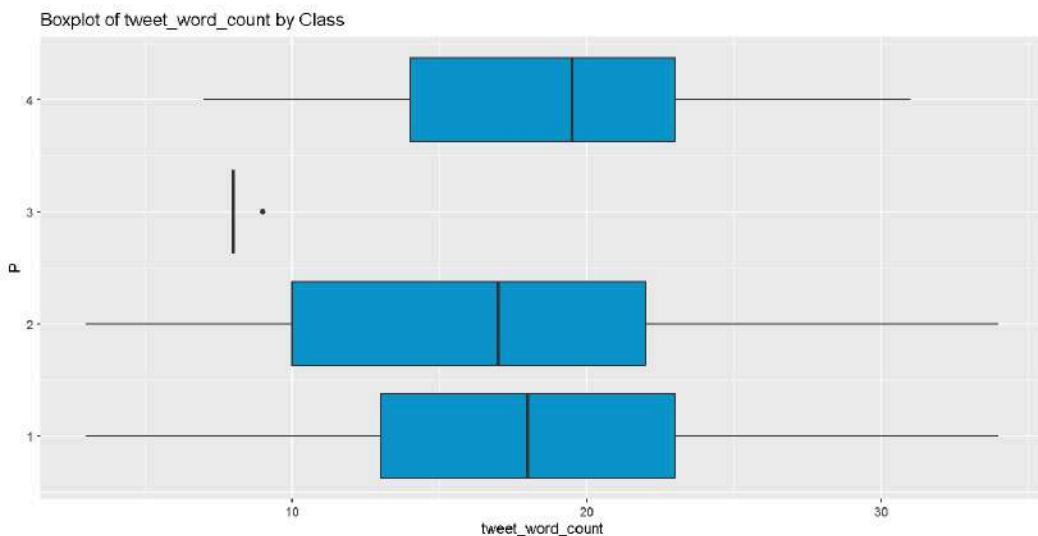
We have also conducted a ValorsTest, ANOVA and Kruskal-Wallis tests between cluster and tweet average word length to determine whether there is a significant association between these two variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the variables.
- H_1 : significant association between the variables.

	ValorsTest	ANOVA	Kruskal-Wallis
P-value	1.294098e-02 4.991165e-03	1.24846741763883e-1 65	1.83155958584308e-23

5.374436e-170
9.961665e-02

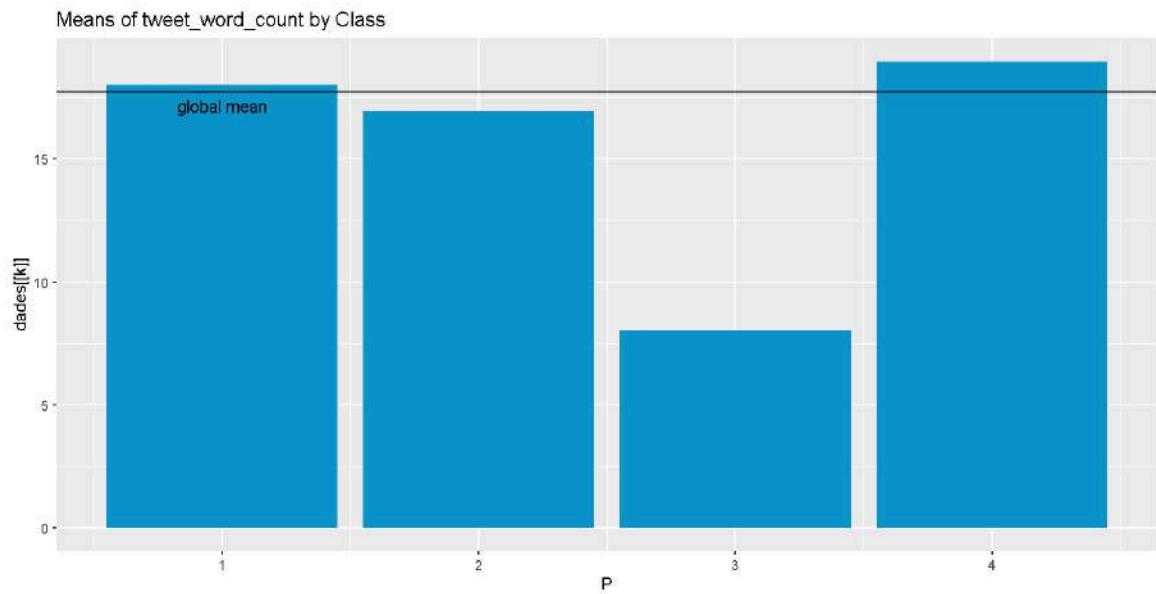
Tweet word count



Upon analyzing its distribution, it becomes clear that there are notable differences among the clusters.

The first and fourth clusters display a similar pattern, with the only difference being the length of the interquartile range and whiskers and the median, with the first cluster being the smallest of the two in the first category and a higher median. The third cluster, however, has a notable difference of lower values compared to the rest and a much smaller quartile range.

The second cluster presents a slightly bigger interquartile range pushed to the left but the median is just slightly lower when compared to the first and fourth clusters.



Furthermore, upon examining the mean of each cluster, it is evident that the third cluster is comprised of individuals with a considerably lower tweet word count than all others. What's more, the fourth cluster presents a global surpassing mean, while the first cluster is at the average value and the second slightly below it.

To be more precise, the mean of all clusters in order are 18.03, 16.96, 8.028 and 18.96 respectively.

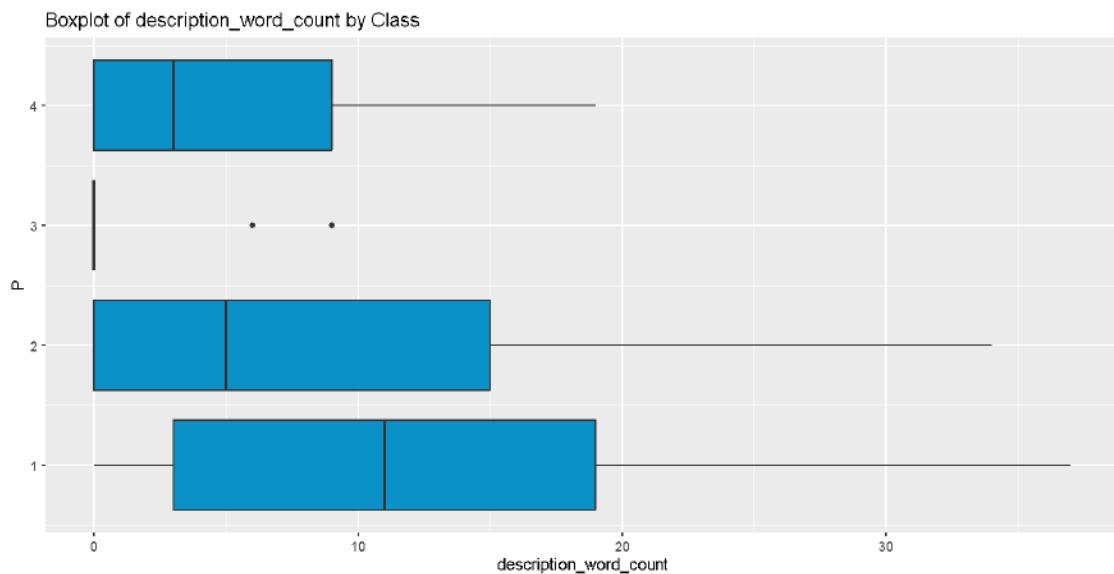
We have also conducted a ValorsTest, ANOVA and Kruskal-Wallis tests between cluster and tweet word count to determine whether there is a significant association between these two variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the variables.
- H_1 : significant association between the variables.

	ValorsTest	ANOVA	Kruskal-Wallis
P-value	5.553902e-11 2.421358e-08 0.000000e+00 2.531438e-02	2.6086457407089e-315	3.80101622134831e-28



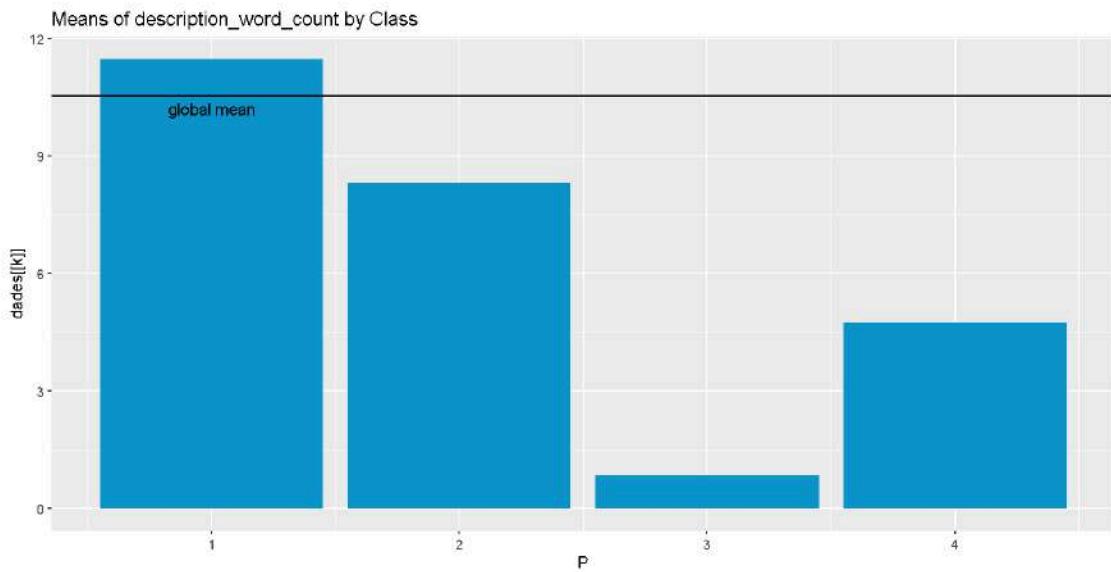
Description word count



Upon analyzing its distribution, it becomes clear that there are notable differences among the clusters.

The second and fourth clusters display a similar pattern in the boxplot, with the only differences being the length of the left whisker and interquartile range. The third cluster, however, has a notable difference of lower values compared to the rest and a much smaller quartile range.

The first cluster also presents the same positive skewness as the first two clusters but with an interquartile range that is further pushed to the right.



Furthermore, upon examining the mean of each cluster, it is evident that the third cluster is comprised of individuals with a considerably lower description word count than all others. What's more, the first cluster presents a global surpassing mean, while the other cluster left are slightly below average.

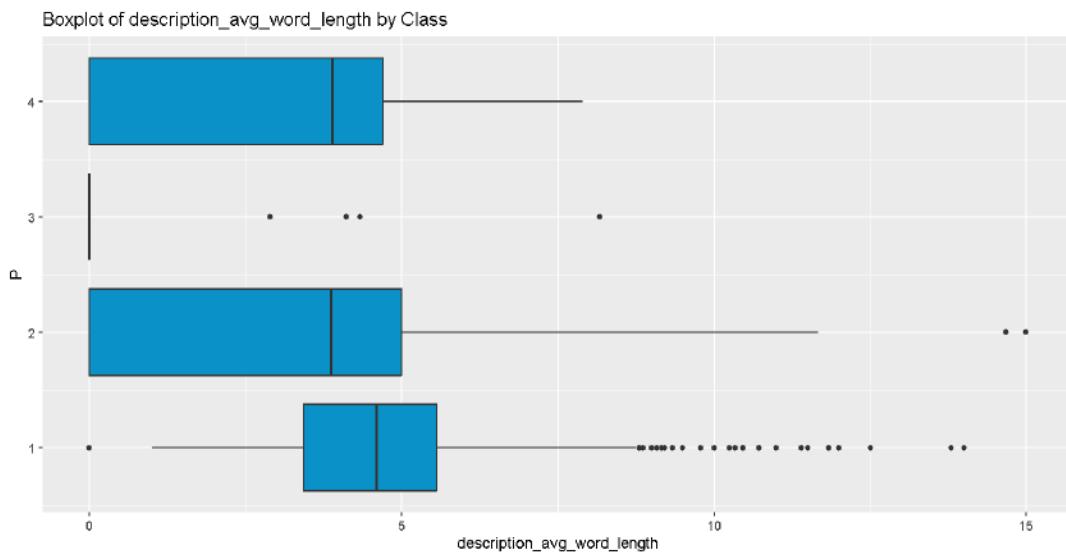
To be more precise, the mean of all clusters in order are 11.46, 8.304, 0.8333 and 4.728 respectively.

We have also conducted a ValorsTest, ANOVA and Kruskal-Wallis tests between cluster and description word count to determine whether there is a significant association between these two variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the variables.
- H_1 : significant association between the variables.

	ValorsTest	ANOVA	Kruskal-Wallis
P-value	1.389738e-39 0.000000e+00 5.529099e-11 2.764415e-10	1.72488152145487e-5 6	3.02426156860567e-55

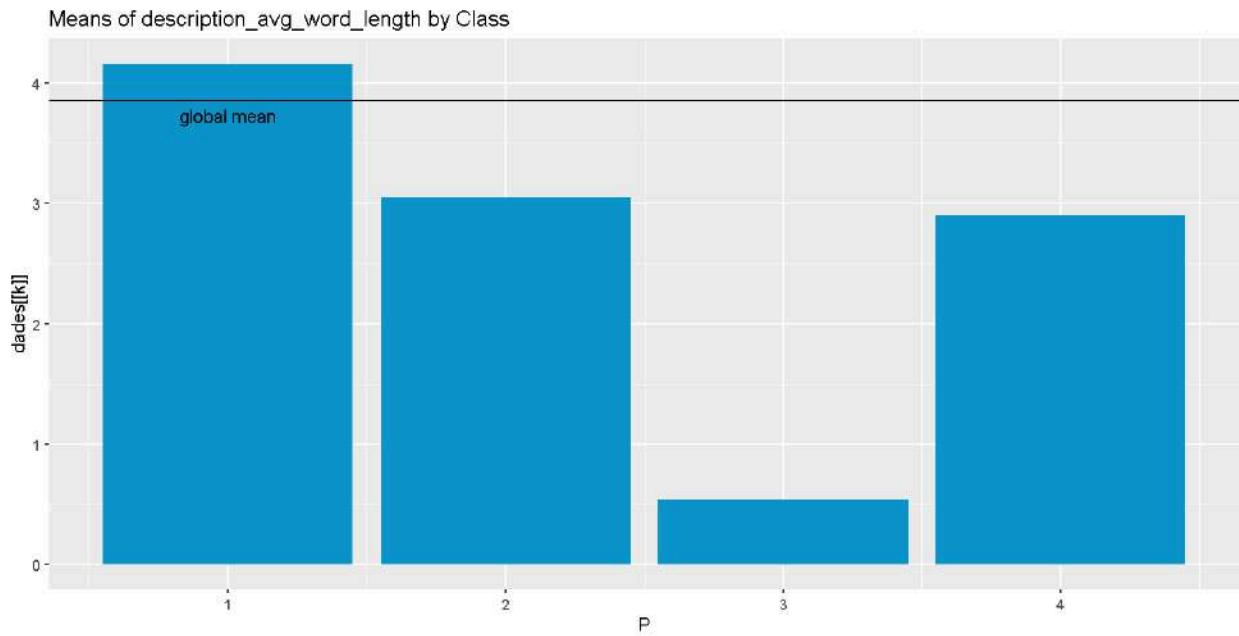
Description average word length



Upon analyzing its distribution, it becomes clear that there are notable differences among the clusters.

The second and fourth clusters display a similar pattern, with the only difference being the length of the right whisker. The third cluster, however, has a notable difference of lower values compared to the rest and a much smaller quartile range.

The first cluster presents a much smaller interquartile range further pushed to the right but the median is just slightly higher when compared to the second and fourth clusters.



Furthermore, upon examining the mean of each cluster, it is evident that the third cluster is comprised of individuals with a considerably lower favorite count than all others. What's more, the first cluster presents a global surpassing mean, while the other cluster left are slightly below average.

To be more precise, the mean of all clusters in order are 4.159, 3.052, 0.5417 and 2.896 respectively.

We have also conducted a ValorsTest, ANOVA and Kruskal-Wallis tests between cluster and description average word length to determine whether there is a significant association between these two variables. The p-value we obtained being below 0.05 indicates that there is strong evidence against the null hypothesis and we can conclude that there is significant association between the two variables. Therefore, this variable is relevant in the clustering.

- H_0 : no significant association between the variables.
- H_1 : significant association between the variables.

	ValorsTest	ANOVA	Kruskal-Wallis
P-value	9.384019e-58 0.000000e+00	1.23109722510975e-3 6	8.11229226393761e-5 2

	2.220446e-16		
	8.123410e-05		

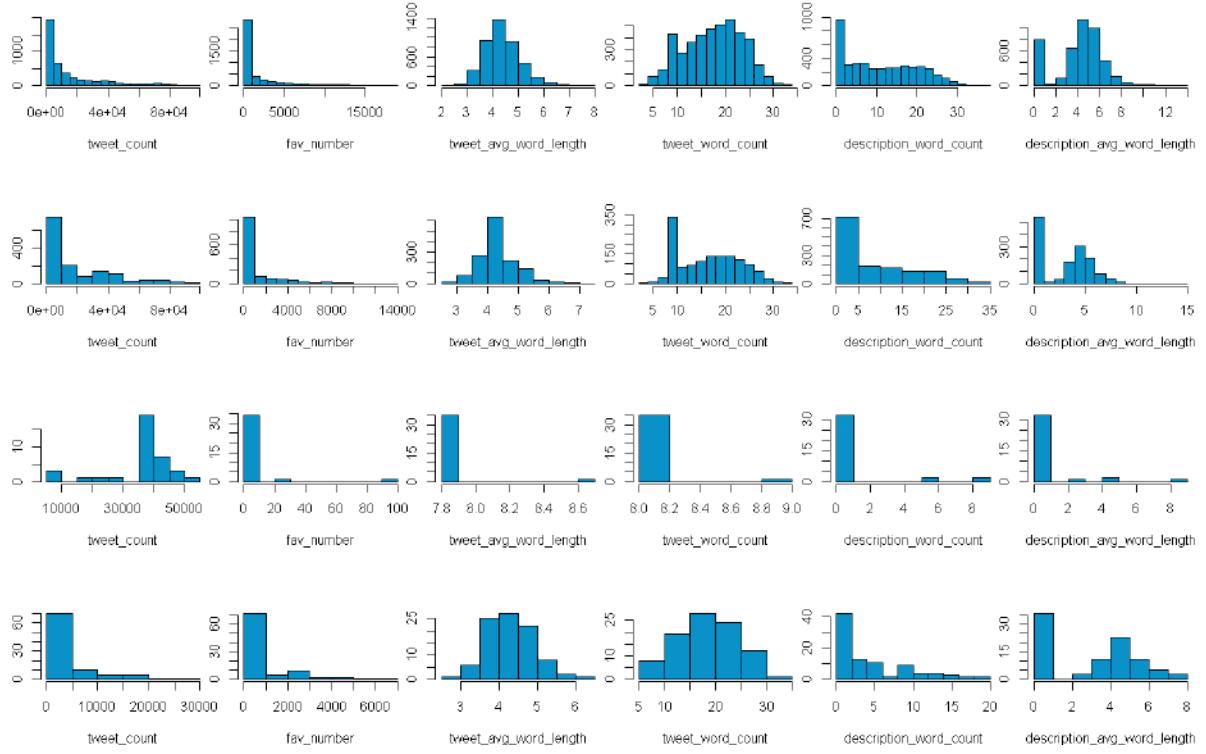
CPG

Furthermore, we utilized a class panel graph to provide a comprehensive overview of each cluster's characteristics. A class panel graph is a type of visualization that presents multiple graphs or plots in a grid-like layout, where each panel represents a subset of the data, such as a histogram for numerical variables or barplots for categorical variables. By utilizing a CPG, we could examine the unique attributes of each cluster and their relationship to one another.

It's important to note that the variable arrangement in a CPG has a significant impact on the cluster's overall interpretation. Thus, we grouped our variables into three sections, mirroring the univariate analysis conducted earlier in this report, all of which focus on user-centered factors. These sections include the user's profile characteristics, such as their personalization preferences; the user's personal information, such as their continent of residence; and the user's behavior, such as the characteristics of their tweets.

As a result, we created the CPG by views (in numerical order of clusters) presented in the same first as the figure below. The other 2 views are on Annex 5 for better visibility.

User's behavior



TLP

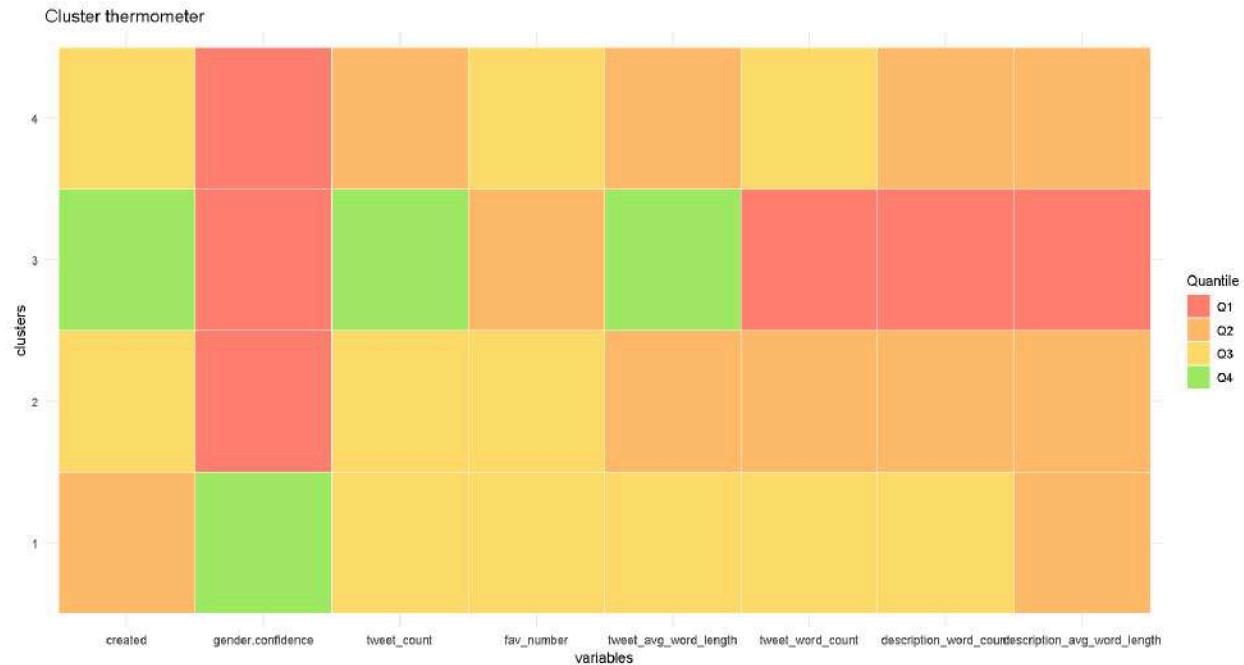
Looking at how this resulting plot grid doesn't intuitively display the relation between each cluster, we decided that the next step to proceed is the elaboration of a traffic lights panel, as it can help showcase more easily the disparity in the values of some variables between clusters.

A traffic lights panel is a type of visual display that provides an easy-to-understand summary of a particular dataset, using the colors of traffic lights (red, yellow, and green) to indicate the status of variables.

Each color represents, in the case of quantitative variables, the absence or abundance of the variable in question compared to the other clusters. As such, red is tied to a bad or negative value, yellow to a medium or neutral value, and green to a good or positive value. What is

considered a bad value are the higher or lower values of the variable depending on the variable's semantics. In our study, red values will significantly lower values.

Hereunder is the resulting graph:



We can see how cluster 2 and 4 share some commonalities even if there are some distinctions in the variables tweet_count, created and tweet_word_word_count. Apart from that, we can see how cluster 3 is the most distinguishable cluster form all the others by their almost absence of medium values.

Cluster template

After all the tools have been implemented in order to aid with the profiling of the clusters obtained, in this case, from DBSCAN, we can now proceed to make a template for the characterization of the main features that distinguish each cluster.

This template will be characterized by a grid with all the variables comprising the observations made prior about each cluster so that we can get a general idea on each one of them. This template will also be grouped by views.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
User's profile				
Link basic color	Blue Red Gray	Blue	Blue	Blue Pink
Sidebar basic color	Blue Black White	Blue Black White	Blue	Blue Black
Created	Old profile	Medium	Young	Medium
User's personal information				
Gender	Male or Female (lowest % of unknown)	Brand or unknown	Brand or unknown	brand or male
Gender.confidence	Highest	Low	Low	Low
Continent	North American Missing Europe (Highest % N.America and Europe)	Missing North America Europe	Missing (Highest % Asian)	Missing North America (Highest % S. America)
Privacy	Least privacy	Most privacy	Medium privacy	Medium privacy
User's behavior				

Tweet count	Medium	Medium	High	Medium-low
Favorite number	Medium	Medium	Medium-Low	Medium
Tweet average word length	Medium	Medium-low	High	Medium-low
Tweet word count	Medium	Medium-low	Low	Medium
Description word count	Medium-high	Medium	Low	Medium-low
Description average word length	Medium-high	Medium	Low	Medium

So we could describe each cluster the following way:

- **Cluster 1:** Predominantly comprised by old North-American profiles made by either a male or female, with a low privacy who write average tweeter posts in both terms of count and word length and a high gender confidence.
- **Cluster 2:** A brand profile or made from an unknown gender with also average tweeter posts which an average gender confidence who also highly values privacy.
- **Cluster 3:** Either a brand profile or made from an unknown gender who chose all blue in profile personalization, is young and has a non-disclosed location but privacy is not the most predominant choice. Its tweets are high in numbers, low in content but the words are longer. It also likes to keep its description at a minimum in all aspects.
- **Cluster 4:** A brand or male profile who tweets rarely but when it does the tweet possess average traits. It also possesses a average description and privacy setting.

Curiously enough, some of the clusters present the same characteristics as the conclusions extracted from the MCA by gender. We can see how cluster 1 shares the majority of its traits with observations encountered with a male profile, with the exception of tweet length as in this cluster it is only average. Cluster 3 also shares a sizeable number of similarities with the observations made with the brand profile. Furthermore, it appears as if Cluster 2 shares a lot of commonalities with the unknown study.

TEXT MINING

PREPROCESSING

Textual data preprocessing plays a crucial role in any analysis that involves unstructured information, such as tweets or user profile descriptions on Twitter. Unlike numerical data, which often already comes in a structured form and ready for analysis, textual data comprises a variety of elements, from words and phrases to emoticons and URL links, which could complicate their interpretation.

In this phase of our analysis, we will focus on the 'text' and 'description' attributes of our database titled "Twitter User Gender Classification". The 'text' attribute encompasses the content of a random tweet from the user, while 'description' incorporates the user's profile description. Both attributes are a rich source of information about the user's preferences, interests, and possibly gender, thus making it vitally important to prepare them properly for subsequent analysis.

Textual data preprocessing is carried out differently from numerical data, as it requires unique techniques and tools. While numerical data can be cleaned and normalized using mathematical and statistical techniques, textual data needs a distinct set of operations such as spam detection and removal, language detection and translation, cleaning of characters and numbers, tokenization, stemming, and stopword removal.

In this section, we will describe in detail how we will approach each of these preprocessing stages and what visualization tools we will use to examine and better understand our data. This approach will enable us to maximize the value extracted from our textual data, making it easier for us to generate more accurate and meaningful conclusions in the later phases of our analysis.

In summary, in this preprocessing section, our goal is to clean and prepare our textual data for analysis and modeling. This process will include a series of specific tasks such as cleaning spam, language detection and translation, cleaning of characters and numbers, tokenization, stemming, and stopword removal. By the end of this process, our textual data will be ready to be used in our gender classification analysis.

Spam Cleaning

As part of the preprocessing process, we performed an initial cleanup of words that do not provide relevant value to our analysis. This step allowed us to streamline the process and reduce computational costs of subsequent procedures, such as translation.

To carry out this initial cleaning, we used regular expressions, a very effective tool for identifying and removing unwanted text patterns. In our case, we focused on three types of elements that are often common in tweets but do not provide significant value for our gender classification task.

- The first element we removed were user mentions, which on Twitter are denoted with an "@" followed by the username. Although these mentions are an integral part of communication on Twitter, for our analysis, they do not provide useful information about the gender of the user we are classifying.
- The second element we removed were links. Twitter allows users to include links to websites in their tweets, but these links, like user mentions, do not provide us with valuable information about the user's gender.
- Lastly, we noticed that there was a recurring phrase, "No description", in the description field of some users. This phrase indicates that the user has not provided any profile description, and therefore, we also removed it from our data.

With this initial cleanup, we have managed to reduce the amount of irrelevant information in our text data, which will allow us to focus on the words and phrases that really matter in the next steps of our analysis.

Language Detection And Translation

Due to the international diversity of Twitter, tweets and profile descriptions can be written in multiple languages. For our study, it's essential to interpret the content of these texts regardless of the language they are written in. Therefore, we opted to translate the 'text' and 'description' attributes to a unified language: English.

The first step in this process involved language identification. We recognized the language of each tweet and description with the aim of subsequently translating it. This language identification is crucial to ensure that the translation is accurate and consistent.

Once the language was identified, we proceeded with the translation of the texts. For this purpose, we established a connection with the Google Translate API. Google Translate is an online translation tool that supports a vast range of languages, known for its accuracy and efficiency.

Translating the texts into English provides multiple benefits to our analysis. First, it allows us to interpret the meaning of the tweets and descriptions without taking into account their original language. This is essential, as phrases like "Hola, ¿qué tal el día?" and "Hi, what's up today?" convey the same idea but in different languages.

Second, the translation facilitates the subsequent step of our preprocessing: the removal of stopwords, lemmatization, among others. By having all texts in a single language, we can consistently and effectively apply these preprocessing techniques, thus ensuring that our data are optimally prepared for subsequent analysis.

Character And Number Cleaning

Having carried out the translation and initial spam cleaning of our textual data, the next step involves the cleaning of characters and numbers, as well as the normalization of uppercase and lowercase letters. In this section, we focus on removing elements that are not letters of the English alphabet and converting all letters to lowercase, as these do not provide relevant information for our gender classification analysis.

To carry out this task, we turned again to regular expressions. These allow us to effectively identify and remove unwanted character patterns. We have decided to keep only the characters of the English alphabet, removing all numbers, punctuation marks, and other symbols. In addition, we have converted all characters to lowercase to ensure consistency in our analysis.

Given that we have already translated our texts into English, we can carry out this cleaning and normalization without the risk of removing characters that may have their own meaning in other languages. However, we have established an exception for words beginning with "#". On Twitter, these words are used as hashtags and often convey additional meaning that may be relevant to

our analysis. For example, "#love" may convey an emotion or preference more intensely than simply "love".

This process of cleaning characters, numbers, and normalization of uppercase/lowercase letters helps us further minimize noise in our textual data and focus on the words and phrases that are truly important for our gender classification analysis. With these cleaner and more focused data, we are better prepared to extract meaningful conclusions in the subsequent stages of our analysis.

Tokenization

After cleaning the texts of spam, translating them into English, removing superfluous characters and numbers, and normalizing to lowercase, the next step is tokenization. Tokenization is the process of breaking down text into more compact units, called tokens, which in this case will be individual words.

For this purpose, we turned to the NLTK (Natural Language Toolkit) library, a resource widely used for natural language processing in Python. Specifically, we applied the 'punkt' function of NLTK for tokenization of our words. This tool allows us to effectively and accurately split texts into individual words, facilitating their subsequent analysis.

However, we established an exception for words that contain "#". On Twitter, as we mentioned earlier, words preceded by "#" are hashtags and can contain additional meaning. If we applied standard tokenization, the "#" would be separated from the word, which could result in the loss of this additional meaning. Therefore, we incorporated an exception in our tokenization process to keep these words with "#" intact.

Tokenization is an essential step in our preprocessing of textual data, as it allows us to analyze our texts at the word level, greatly facilitating subsequent stages of our analysis, such as the removal of stopwords, lemmatization, and more.

Stemming

Once tokenization is complete, the next step in our preprocessing of textual data is stemming. Stemming is a procedure that simplifies words down to their root or basic form, removing affixes such as suffixes or prefixes.

The purpose of stemming is to consolidate the meaning of words by minimizing them to their most essential form. For example, the words "running", "runner", "ran", and "runs" would be simplified to their common root "run". This allows us to treat all these words as one, facilitating a simpler interpretation and analysis of our texts.

Stemming is particularly useful in our case, as we are working with textual data from Twitter, which can contain a wide diversity of words and grammatical structures. By minimizing these words to their root, we manage to simplify our data and focus on the most relevant semantic content of our texts.

It's important to note that stemming can have limitations and does not always provide the most correct or "lemmatized" basic form of a word.

Therefore, stemming is a valuable tool in our text data preprocessing process, as it allows us to consolidate the meaning of words and simplify our data for subsequent analysis.

Lemmatizer

After the tokenization of texts, the next step in our preprocessing is stemming. This process involves reducing words to their root or "stem", which allows us to focus on the primary meaning of words and decrease the dimensionality of our data.

For example, the words "running", "runner", and "run" originate from the same root "run", and they all encapsulate a similar meaning related to the action of running. By applying stemming, all these words are simplified to "run", which facilitates treating them as the same word in our analysis.

To make this process more efficient and precise, we have used part of speech tags (POS tags). These tags provide us with information about the grammatical function of each word (whether it is a noun, verb, adjective, etc.), which allows us to perform more appropriate stemming. For instance, the word "running" could be a present participle verb or an adjective, and its root may vary depending on its function in the sentence. By using POS tags, we can ensure that we apply stemming pertinently.

Stemming is a powerful technique that allows us to simplify our texts and focus on the main meaning of the words. This process will help us gain more precise and meaningful insights in the subsequent stages of our gender classification analysis.

Stopwords

The final stage of our preprocessing process is the removal of filler words, also known as "stopwords". These are common words in a language (for example, "the", "a", "is", "in" in English) that often do not significantly contribute to the understanding of a text's content. By removing these words, we manage to reduce the volume of data to be analyzed and focus on words that truly provide meaning, which is our main goal.

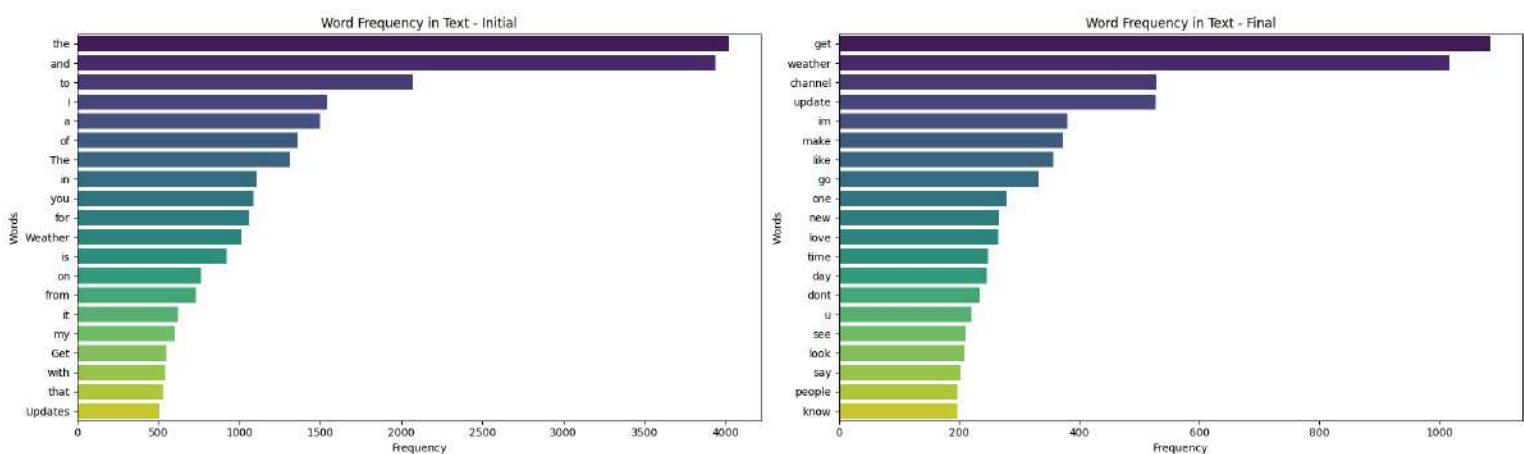
To identify and remove these filler words, we rely on the list of stopwords provided by the natural language processing library NLTK for English: `stopwords.words('english')`. This list covers most of the commonly used words in English that generally have reduced semantic value.

Thus, the removal of stopwords is a powerful tool that allows us to simplify our texts and focus on words with more informative content.

Results

In this section, we will examine how our preprocessing has impacted the data. It should be noted that while developing the code, we informally checked the results to fine-tune the code. However, here we will compare the results from the initial and final data of the two text variables:

This plot represents the 20 most frequent words in the tweets before and after preprocessing:

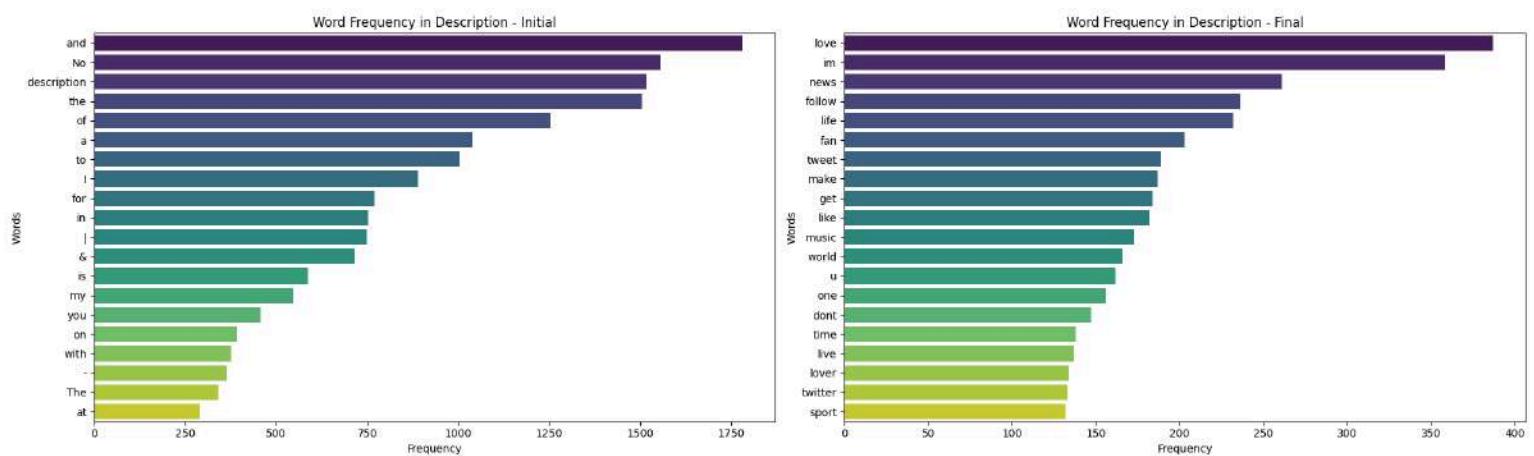


Initially, the dataset was dominated by stopwords such as 'the', 'and', 'to', 'I', 'a', 'of', 'The', 'in', 'you', 'for'. These words, while frequently used in the English language, offer little insight when analyzing the content of tweets. The prominence of the word 'Weather' hints that the initial dataset might have contained numerous tweets pertaining to weather conditions.

Upon preprocessing, a dramatic shift in the list of frequently used words is evident. With the stopwords eliminated, the remaining words convey more specific and useful information. Words like 'get', 'weather', 'channel', 'update' imply discussions centered around weather news or updates. Interestingly, 'im' is seen frequently, likely a representation of "I'm", suggesting that contractions may not have been adequately handled during preprocessing. Other words such as 'make', 'like', 'go', 'one', 'new', 'love', 'time', 'day', 'dont', 'u', 'see', 'look', 'say', 'people' suggest a wide array of topics and interactions.

Interestingly, despite these changes, the frequency distribution of words seems to have remained relatively consistent before and after preprocessing.

This plot represents the 20 most frequent words in the descriptions before and after preprocessing:



Initially, the most common words include a plethora of stopwords such as 'and', 'No', 'the', 'of', 'a', 'to', 'I', 'for', 'in'. Symbols like 'I' and '&' also make an appearance, along with the word 'description', which is likely part of default or incomplete profile descriptions in user profiles.

After preprocessing, the list of most common words appears to have significantly changed. The stopwords have been purged, and the remaining words seem to convey more specific information. Words like 'love', 'im', 'news', 'follow', 'life', 'fan', 'tweet', 'make', 'like', 'get', 'music', 'world', 'u', 'one', 'dont', 'time', 'lover', 'twitter', 'live' suggest a variety of themes and attitudes. Interestingly, contractions such as 'im' and 'dont' appear frequently, suggesting that the preprocessing might not have correctly handled these.

The frequencies of the words have shifted after preprocessing, indicating that many stopwords have been eliminated and the data dimensionality has been reduced.

Final analysis

Preprocessing textual data, in this case, tweets and Twitter user descriptions, has proven to be a crucial and effective step for extracting meaningful information and reducing data dimensionality. The removal of stopwords and other non-informative elements, such as certain symbols, has allowed more significant words to stand out in the analyses.

In the case of tweets, removing stopwords and other non-informative elements has unveiled more concrete and useful themes, centered around news or weather updates, among others. This suggests that the dataset likely contained a large number of tweets related to weather conditions.

Regarding user descriptions, preprocessing has allowed us to discern a wider variety of themes and attitudes, such as love, music, and life, among others. The resulting words offer a more diverse and useful insight into how users choose to describe themselves. Again, the issue with contractions was observed.

Moreover, despite substantial transformations in the composition of the most common words, the frequency distribution of words seems to have remained relatively consistent before and after preprocessing. This indicates that, although preprocessing alters the data representation, the underlying structure may stay consistent.

In general, these analyses demonstrate the importance and utility of preprocessing in extracting valuable information from textual data in the context of social media analysis. Although preprocessing may present challenges, such as handling contractions, the benefits in terms of data dimensionality reduction and enhancing the relevance of extracted information are clear.

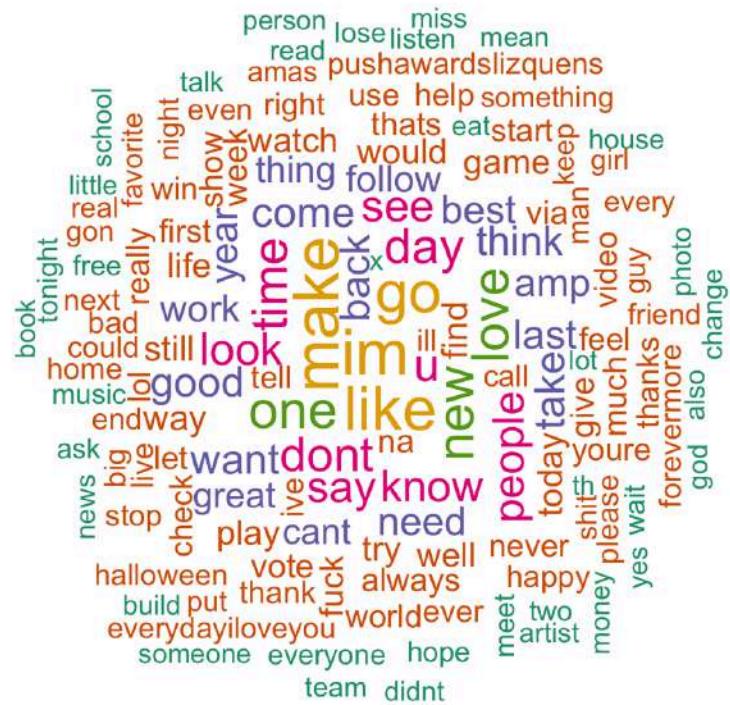
SENTIMENT ANALYSIS

Sentiment analysis, also known as opinion mining, is a technique that enables us to extract subjective information from text and classify it as positive, negative, or neutral. By combining sentiment analysis with gender identification in tweets, we can gain a comprehensive understanding of gender-specific patterns in sentiment expression and uncover nuances that may influence the ways in which individuals interact, perceive, and communicate on social media platforms. Understanding these patterns can aid in tailoring marketing campaigns, identifying target demographics and promoting gender-inclusive discussions.

In our case, we are found with two textual variables: *text* (the tweet of the user), as well as *description* (the description of the user's profile). We will be analyzing both and trying to extract conclusions on how their predicted gender, continent or privacy settings might interact with the sentiment used in their tweet.

For example, these are the most used words in our database:

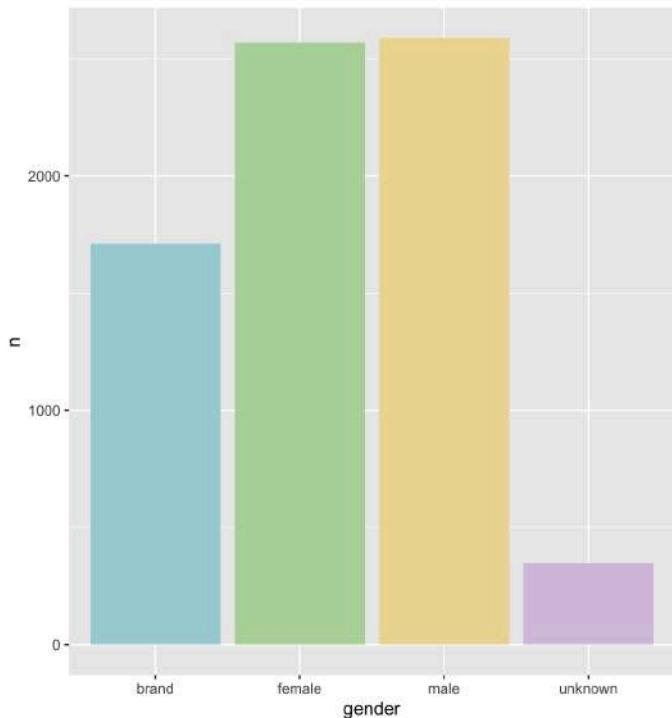
Most Used Words in Tweets



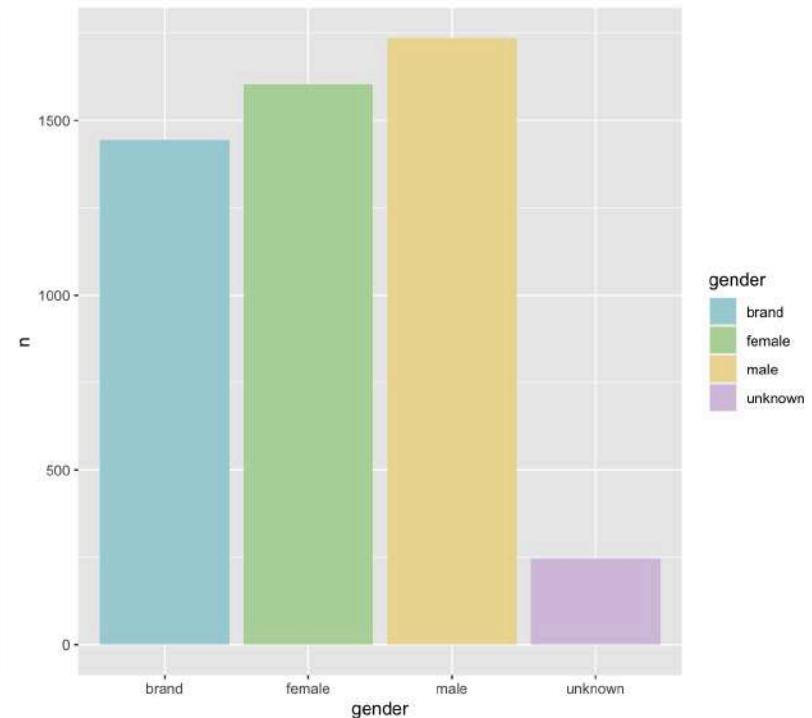
Word Frequency by Gender

First off, we have taken the preprocessed database with the clean text and we are starting off by looking at the different amount of words used by each predicted gender, for both the description and the tweet:

Total Tweet Word Count by Gender

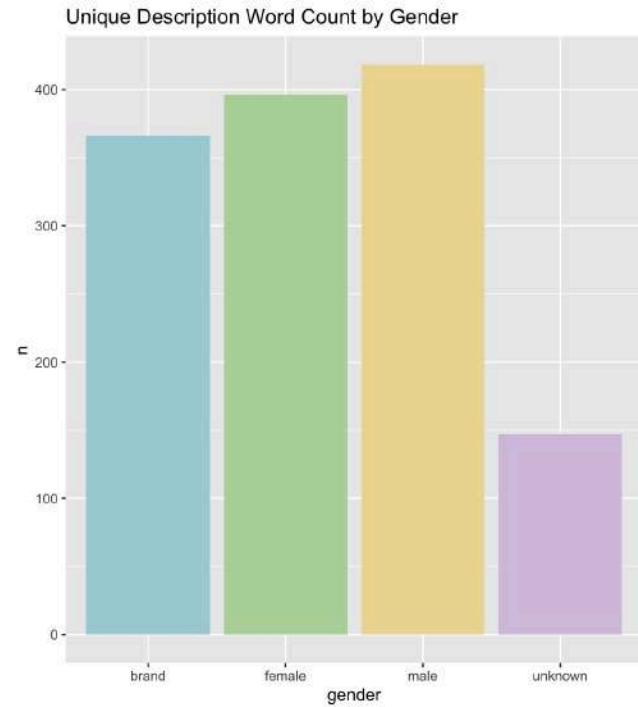
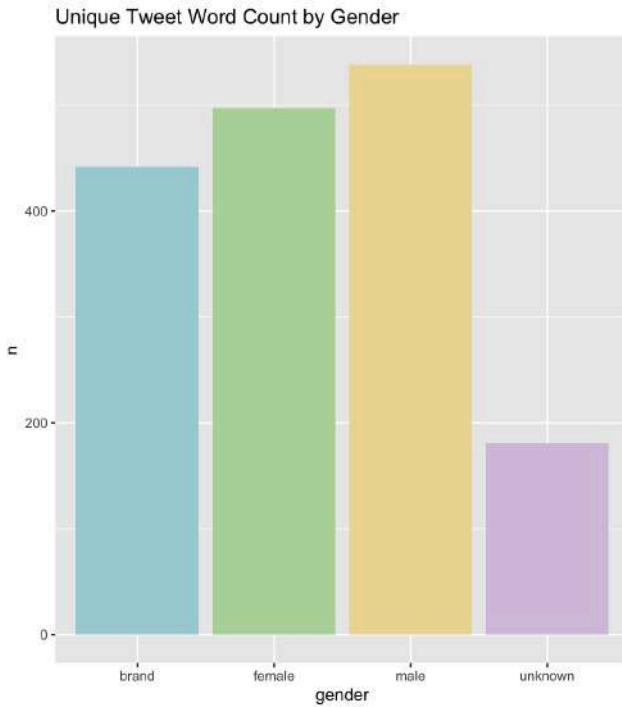


Total Description Word Count by Gender



As we can see, the male and female tweets seem to have the same amount of words in total, while the brand has less words and the unknown has less than $\frac{1}{4}$ of the amount the *male* and *female* categories have. This makes sense, since it tells us that human users are the ones that have longer tweets, followed by brands. And those accounts that have not been able to be classified have a low amount of total words, making it harder to predict their gender. As for the description, we find that the category *male* has the highest total word count, followed closely by *female* and then *brand*. This tells us that, when it comes to description, *brand* users are more similar to the *male* and *female* users. It is also interesting to see that predicted *male* users have a higher description count than *female*.

Although this gives us some information, it is more useful to directly check the unique word counts used by each gender, as it will not represent the length of the tweets but more so the different words used:



gender

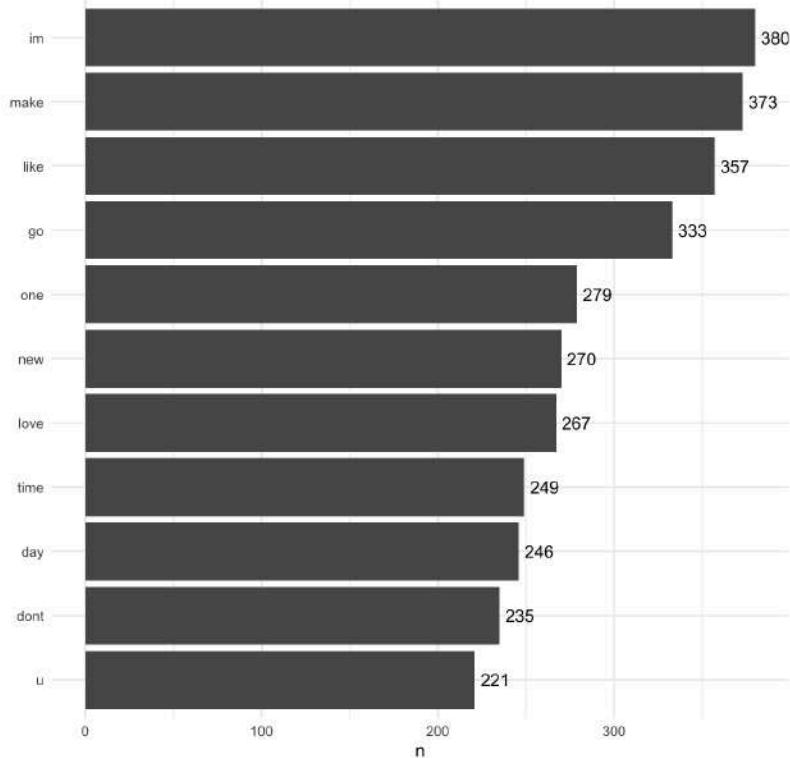
- brand
- female
- male
- unknown

Here, we can observe the same pattern for both the description and the tweet's text. We can see that those users predicted as males have the highest word counts, followed by *female*, *brand* and lastly *unknown*. It will be interesting to see which are the words that are found in the *male* category that are not used in the other genders, which we will do soon.

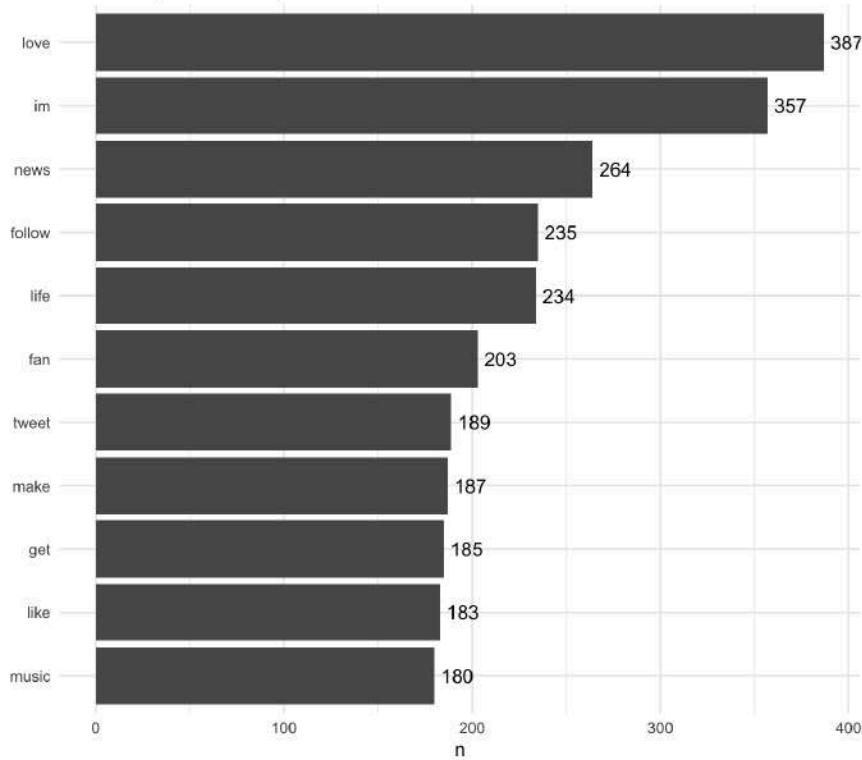
Additionally, we can see that the tweets have (overall) more unique words used than the description, which tells us that descriptions have a tendency to use the same words.

But before that, let's take a look at what are the most used words overall in the tweet and description of our users.

Most Frequent Tweet Words



Most Frequent Description Words



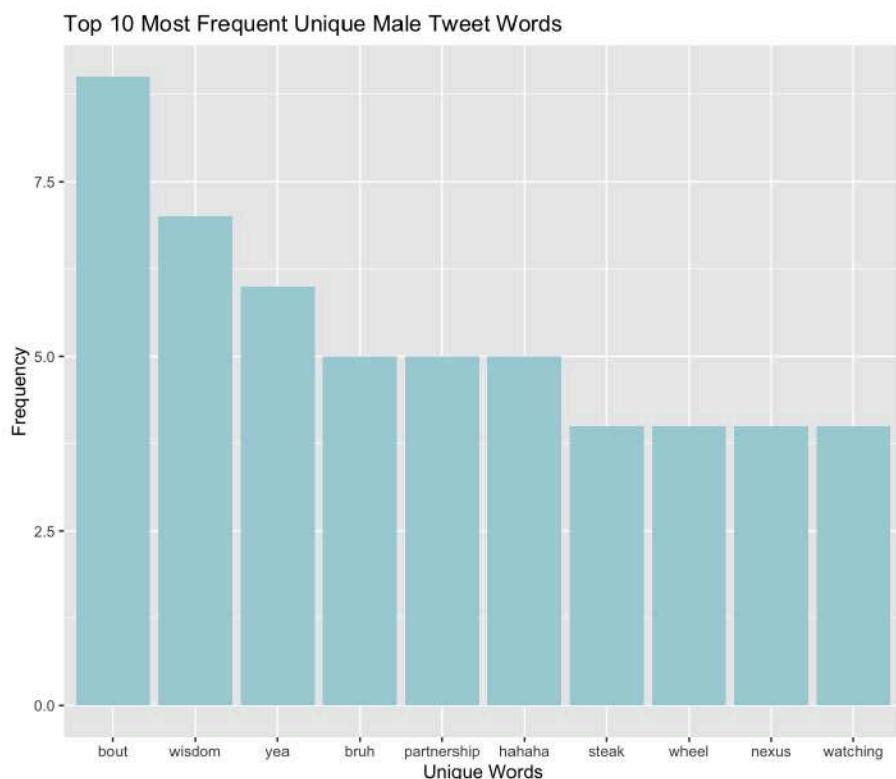
From these two plots, we can conclude that the most used words for the tweets are mainly verbs (update, get, i'm, make, like, go...). This makes sense, since tweets usually express an action or opinion happening in the moment. On the other hand, description, although having some verbs too (such as tweet, make, get, like, follow), seems more centered around catching the eye of whoever is checking out their account, instead of expressing something as with the tweets. For example, there is a lot of use of the word *follow*, which invites the person visiting their account to follow them. The description also has words which seem to describe the person, such as fan (who they are a fan of, which could be a sports team, a singer...), or the music they listen to. Lastly, it seems that both have love on their most used words, while for description it is the most used and for tweets the least one.

Unique Words by Gender

As we explained before, it would be really interesting to see what words are unique to only one gender. This can tell us what words are very specific to each gender and serve as great factors that influence the prediction of someone's gender. Take into account that these unique words have a very small frequency, which means that we can not necessarily confidently trust the conclusions obtained.

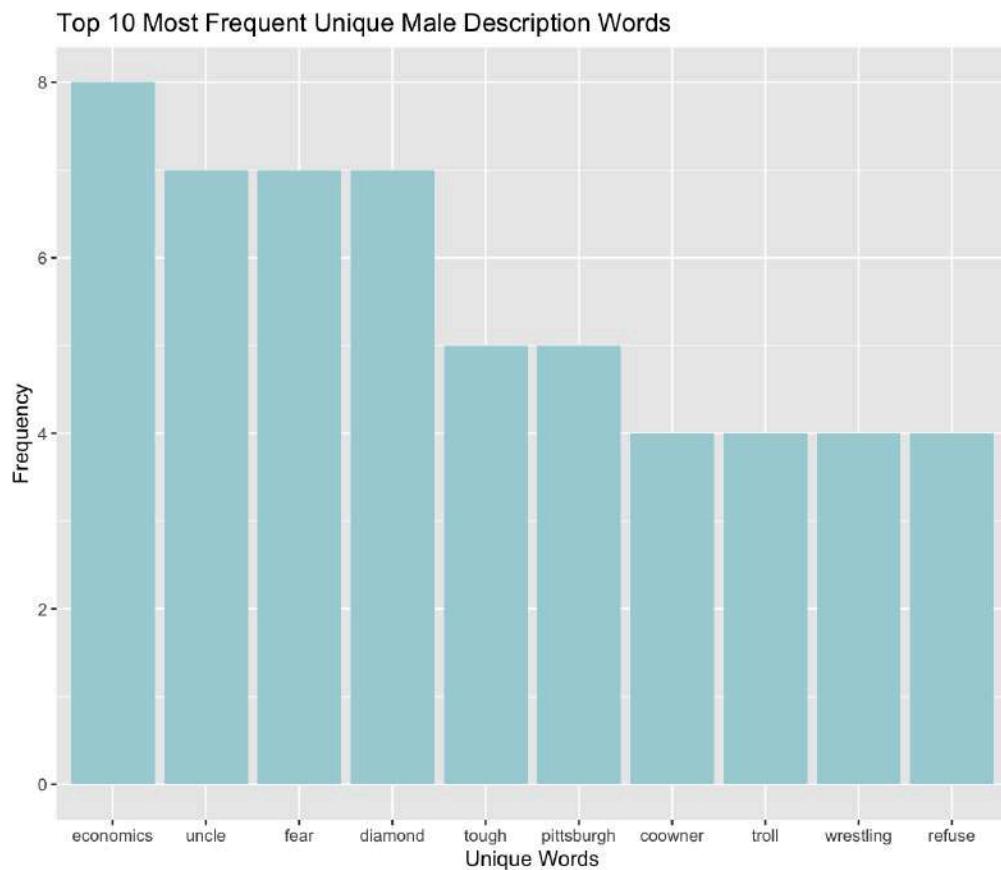
Male Unique Words

Here is the plot for the words unique to the male users:



We can see that the most used words are *bout*, *wisdom* and *yea*, with a frequency of 9, 7 and 6 respectively. Additionally, we also find words such as *steak*, *wheel*, and *bruh*, which are typically associated with the *male* gender.

Hereunder is the same plot, but for the description words:

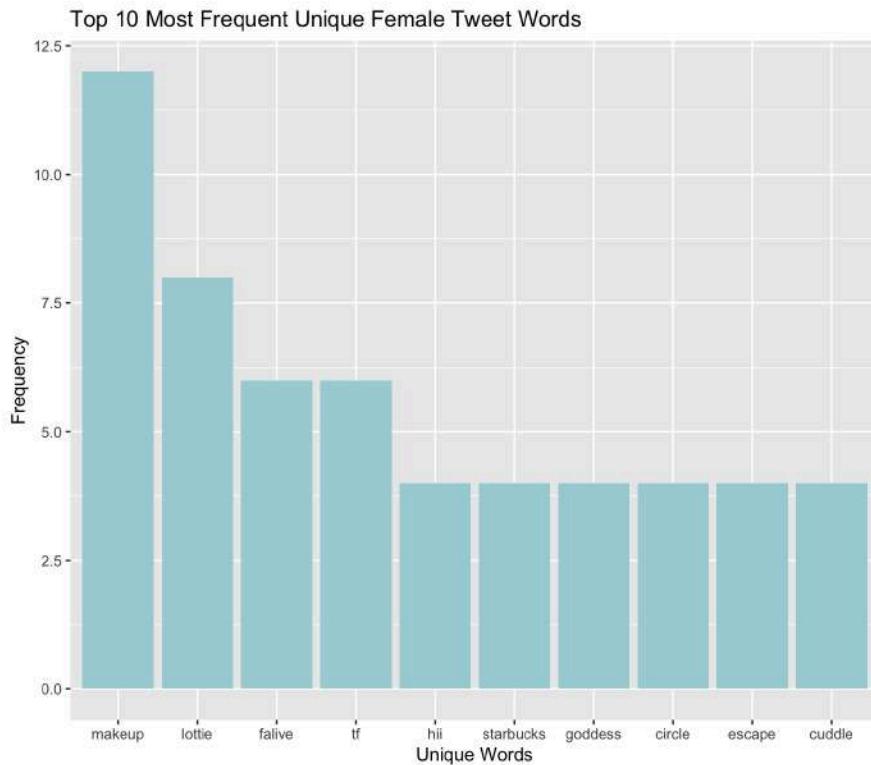


We can see the words that only predicted *males* have used to define themselves. We observe things such as *economics*, *uncle*, *fear*, *diamond*, *tough*, *co-owner*, *troll*, *wrestling*... These are mostly things stereotypically related to masculinity.

Female Unique Words

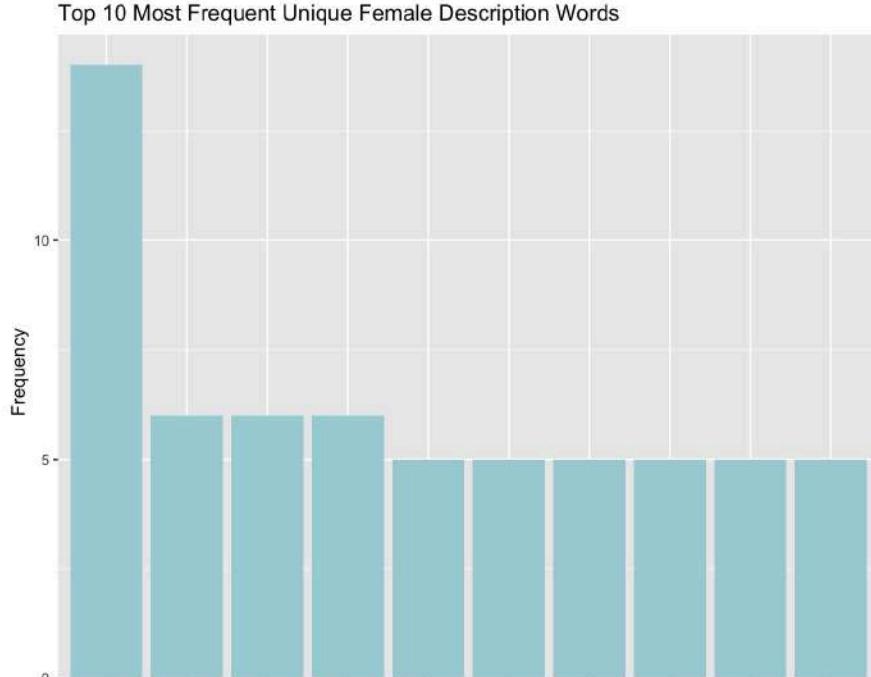
Next up, we will be looking at the unique words found only in the users that have been classified as *women*. This will allow us to gain some insight into what words are typically associated with women.

First off, we will start off with the most frequent unique female tweet words:



We can see words typically associated with the *female* gender such as makeup, starbucks, goddess and cuddle. This shows that there is still a connection between our perception of women to associations with makeup, affection (for the word cuddle), interest in shopping and coffee (starbucks)...

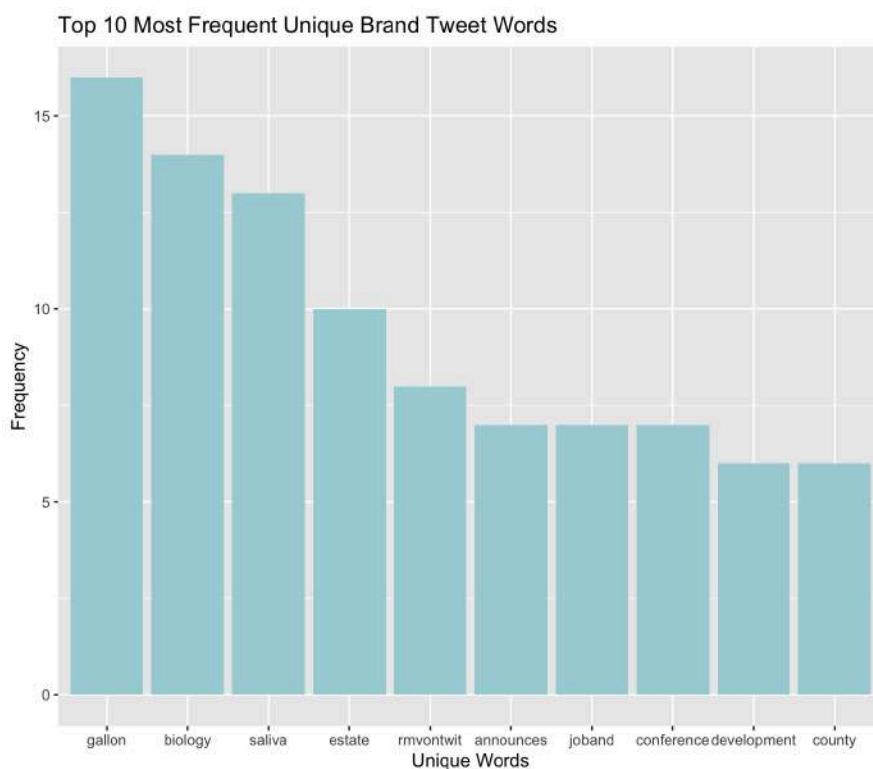
Hereunder, we can also observe the same plot for the description words of predicted *female* users:



We can also see the word *makeup*, as well as *mixer*, *ambassador* and *nurse*. Afterwards, we find a list of names, such as Ariana, Tom, Shawn and Sandra. This could be mentions to celebrities or simply their own names.

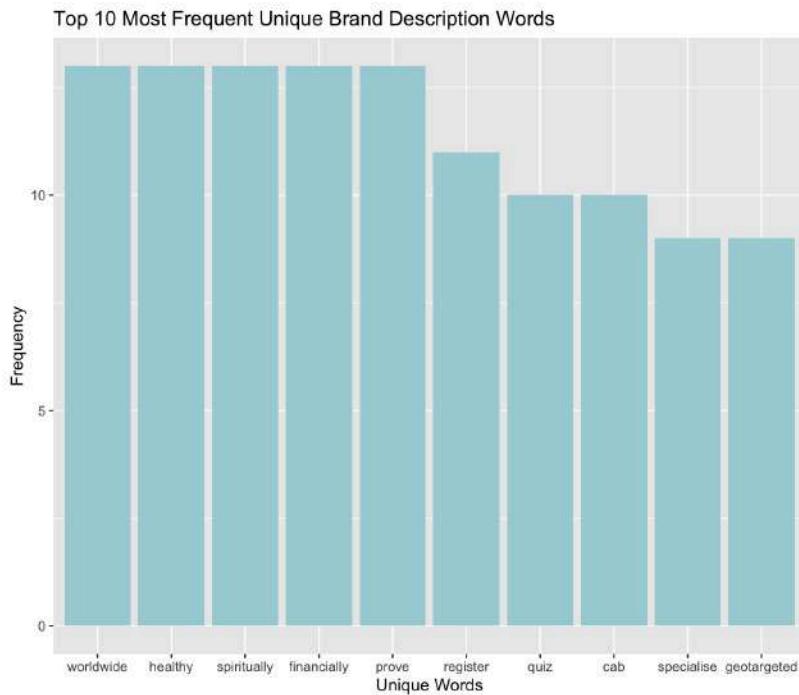
Brand Unique Words

Next up, we will look at the unique words that are only used by brand users:



As we can see, the words unique to the brand users are remarkably formal and sophisticated compared to others. We see some that talk about science such as *biology*, *saliva*, and *gallon*... On the other hand, other words are *estate*, *conference*, *county*, *development*, *announces*... All of these words point to talk about their company and possible announcements or news.

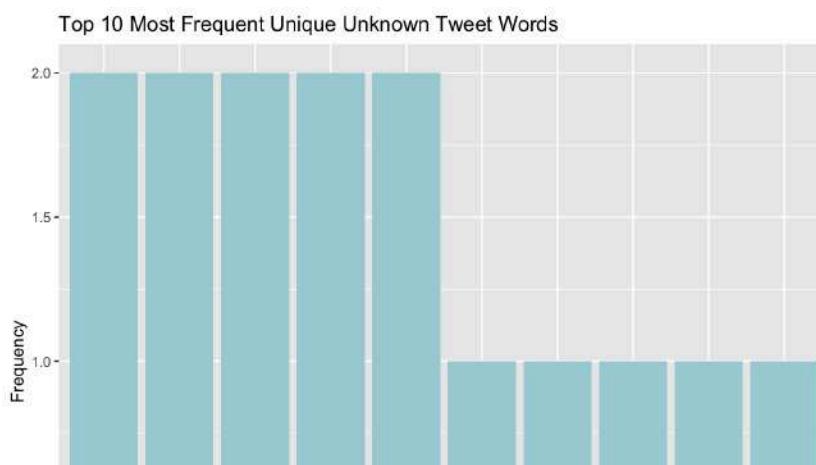
Let's look at the description words now:



We have words that could very well be used in advertising, such as *worldwide*, *healthy*, *spiritually*, *financially*, *register*, *specialise*, *geotargeted*... It seems that the users classified as brands use adjectives that are typically used to describe companies.

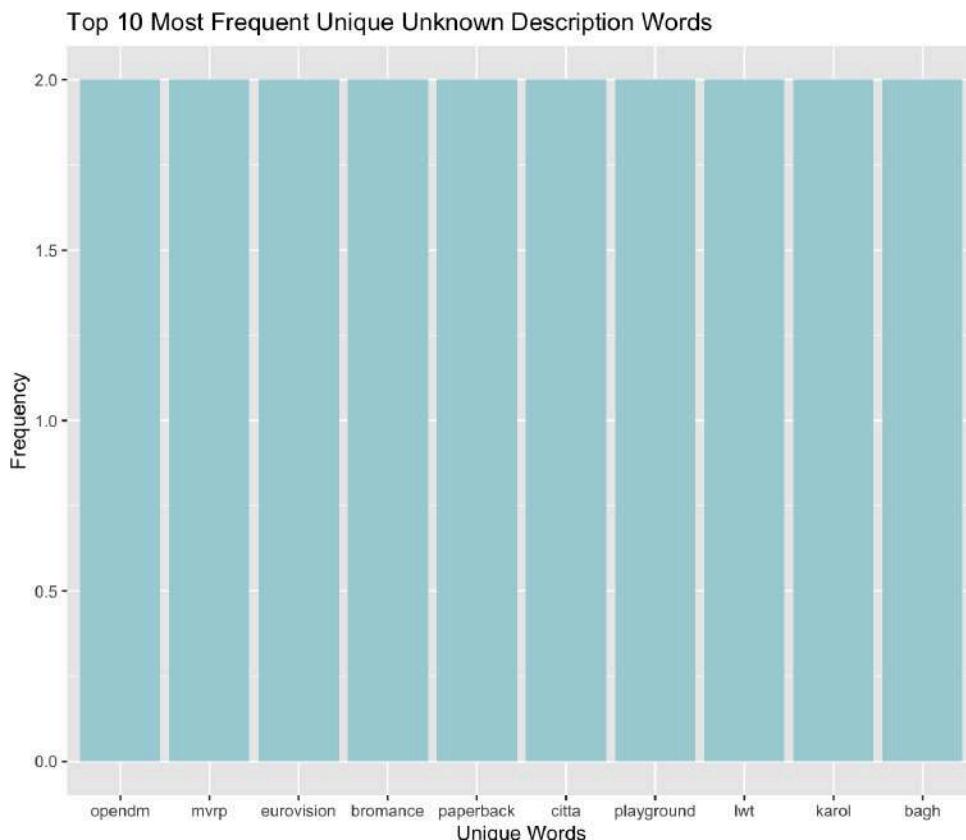
Unknown Unique Words

And lastly, these are the words unique to the *unknown* gender category:



We can see that the frequency of these words goes from 1 to 2, which is not very conclusive. Additionally, these words are mainly typing errors which explains why they are infrequent.

As for the description words, we observe the same pattern:



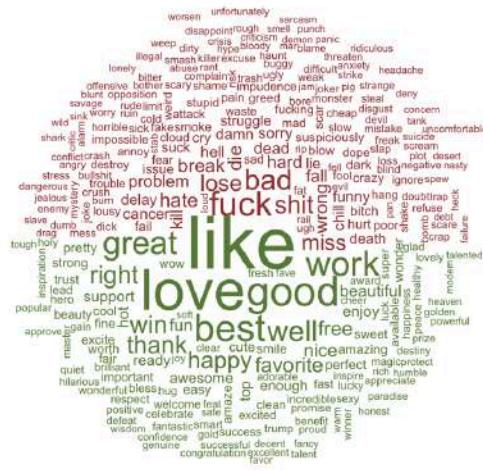
Very low frequency, with words such as *opendm*, *mvrp*, *eurovision*, *bromance*, *paperback*, *citta*, *playground*... Sources such as [Urban Dictionary](#) say that MVRP is an acronym for Multiverse Roleplay, and is used when users are roleplaying a character from another universe. This explains the difficulty of identifying their gender, as all the information displayed about their account only shows information about the gender they are taking on as a role. Urban Dictionary also states that LWT is an acronym for “Laughing With Tears”.

Polarity Scores

The first thing we can do to extract some information from our database, apart from directly analyzing the words used, is to look at the polarity scores of each tweet. We can do this by using a polarity dictionary, which has assigned words and how positive or negative they are.

Hereunder is a plot of the most used positive and negative words for both the tweet and description:

Most used Negative and Positive Tweet Words



Most used Negative and Positive Description Words



As we can see, both of them have similarities. For example, they both have *love* and *like* as most used positive words. We find general adjectives on the tweets such as great, right, good, happy, favorite. On the other hand, we find adjectives related to a person such as lover or enthusiast.

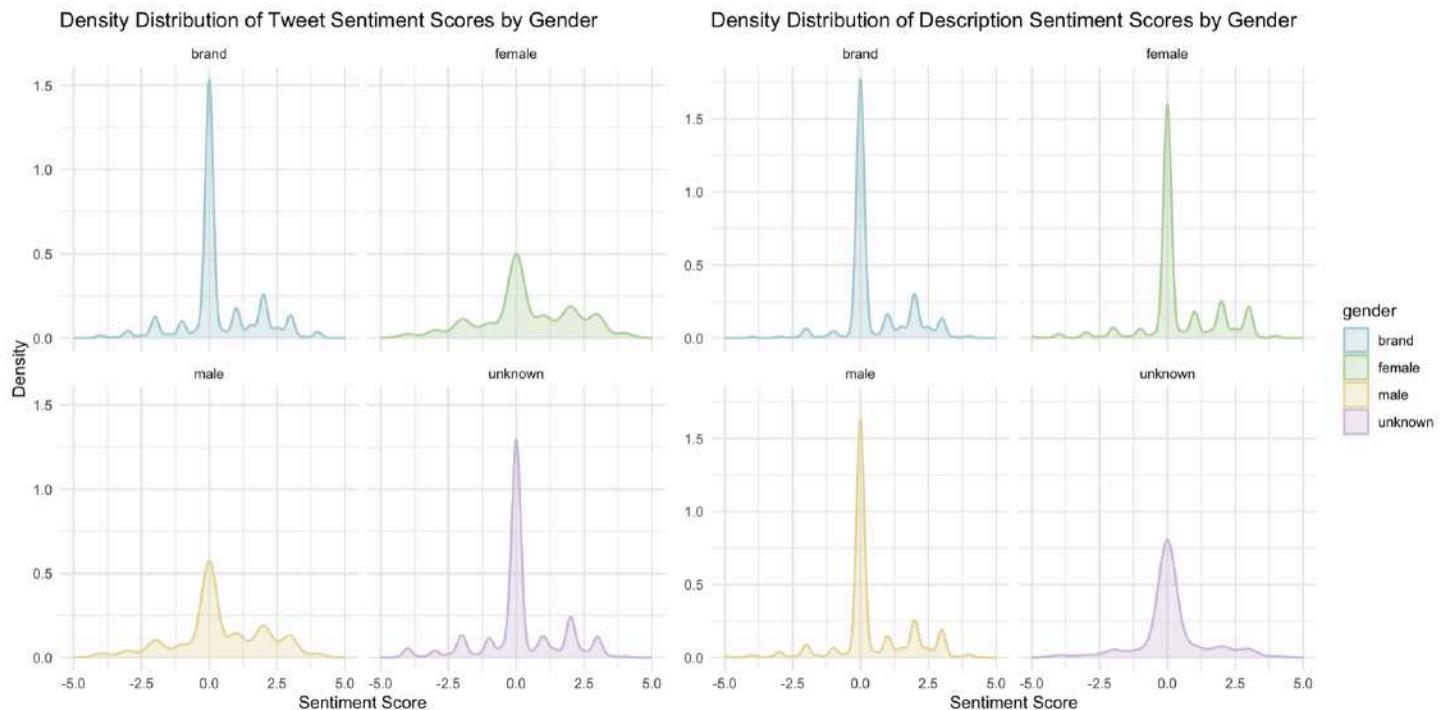
As for the negative words, we find that the tweets are filled with swear words as well as verbs such as break, die, hate, miss and kill. For descriptions, we see words such as conservative, joke, break, addict, RIP to be the most used ones.

We can also analyze the proportion of positive and negative tweets on our database:

Tweets		Description	
Positive	Negative	Positive	Negative
0.5694289	0.4305711	0.6687037	0.3312963

As we can see, most of our database is positive, but people seem to be much more positive in their description than their tweets.

First off, we found it was convenient to see the distribution of sentiment scores by gender. We did this by getting the average score of each tweet, and then plotting the density distribution of those scores for each gender. Hereunder is the plot:



As for the tweet sentiment score, we can see how, overall, the *brand* seems to be the category that is most centered around 0, followed by the *unknown* category. This tells us that these two are usually very neutral when they are tweeting, while for *male* and *female* categories, they seem to be much more opinionated and either *positive* or *negative*. This was to be expected as, usually, brands maintain a neutral and professional image.

When it comes to the description, the differences between *brand* and the *male* and *female* categories decrease. It seems that classified *male* and *female* users are much more neutral in their profile descriptions than their tweets. As for the *unknown*, they are much less neutral than before. We can also observe the average and standard deviation of these scores:

Sentiment Score		Female	Male	Brand	Unknown
Average	Tweet	0.399	0.388	0.349	0.161
	Description	0.423	0.423	0.495	0.135
Standard Dev.	Tweet	1.73	1.65	1.35	1.43
	Description	1.37	1.34	1.14	1.22

We can observe how the accounts predicted as *female* have the most positive tweets on average, as well as the biggest standard deviation. They are closely followed by *male*, then *brand* and then close to 0, we find the *unknown* category.

When it comes to description, we find that the *brand* category has the highest average, followed by both the *male* and *female* categories.

In order to obtain a conclusion about these sentiment scores, we performed an ANOVA model, whose summary provides information about the statistical significance of *gender* in explaining the variation in the sentiment score.

```
> anova_model <- aov(Points ~ gender, data = tweets_points)
> summary(anova_model)

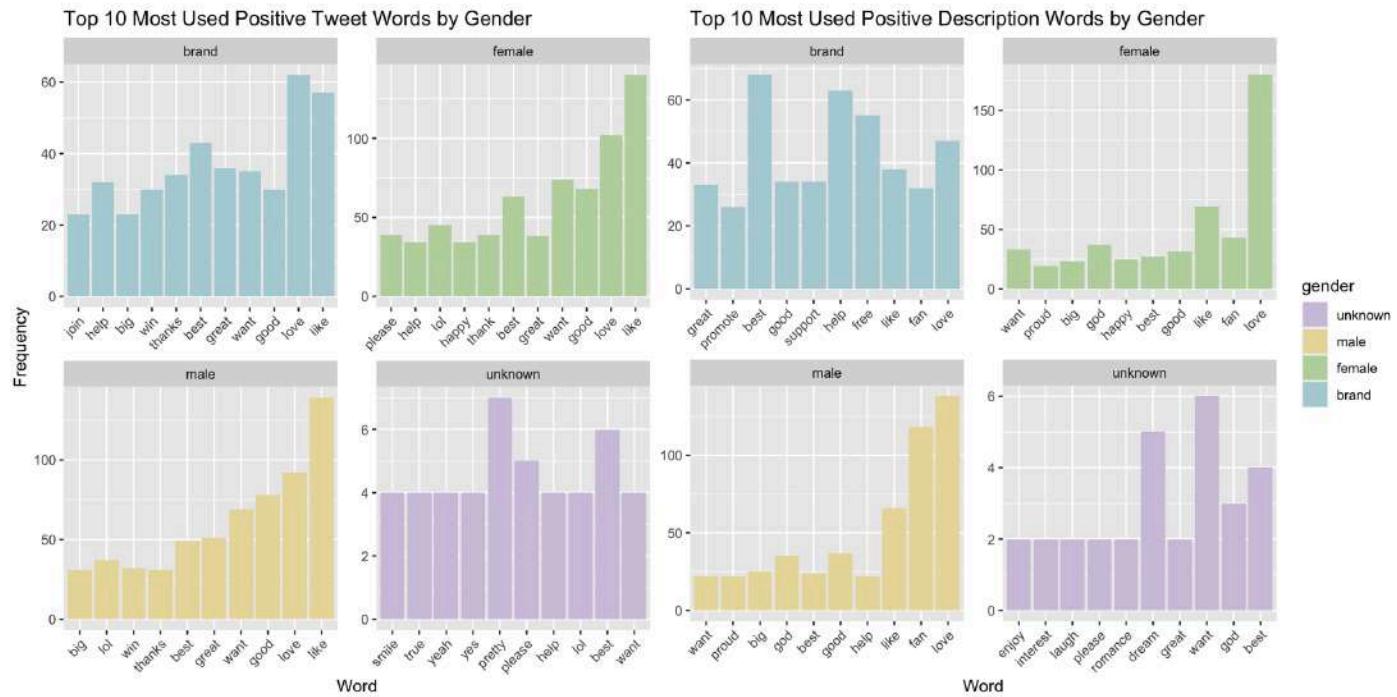
Df Sum Sq Mean Sq F value Pr(>F)
gender      3     21   6.951   2.802 0.0384 *
Residuals  6996 17352   2.480
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To interpret these results, it is important to know that the mean sum of squares is 6.951, which represents the average variability explained by the 'gender' variable. The F value is 2.802, which is the ratio of mean squares between groups of different gender to mean squares within groups (residuals). It tests the overall significance of the *gender* variable in explaining the variation in the sentiment score.

Based on the ANOVA summary, the p-value of 0.0384 is less than the conventional significance level of 0.05. This suggests that the *gender* variable is statistically significant in explaining the variation in the sentiment score of a user's tweet.

Now that we are more familiar with the sentiment score of our database and how it relates to gender, we wanted to analyze the most used positive and negative words for each gender, to see if there is a tendency for a gender to use a certain word when being positive or negative.

First off, here are the 10 most used positive tweet and description words:

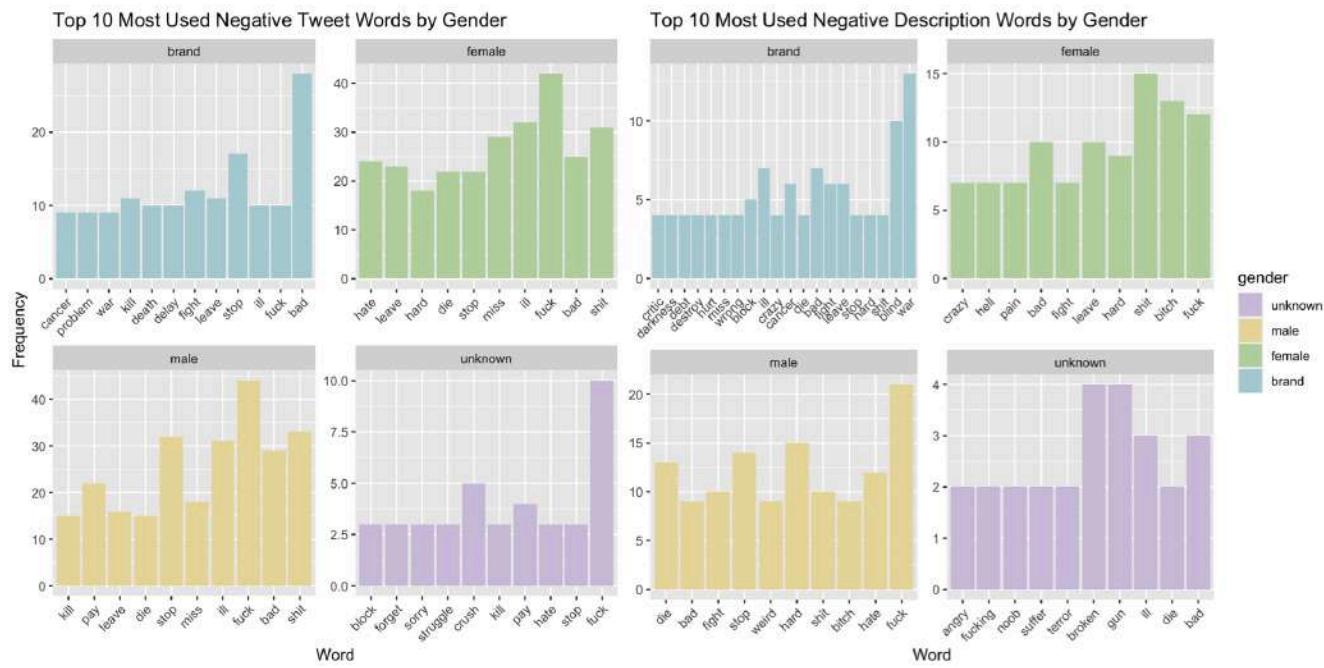


As we can observe, although the relative frequencies vary from one gender to the other, the words that appear are pretty similar. The only differences we can observe are words such as *please* are only found in the *female* and *unknown* category. Additionally, the positive words for the description are the same for both *male* and *female*.

As for the *brand* category we find words such as *promote*, *support*, *free*, *help* which are not found in the other category.

We also find that *god* is one of the most positive words found in the *female*, *male* and *unknown* categories. This tells us that one of the ways people define themselves most positively is with the word *god*.

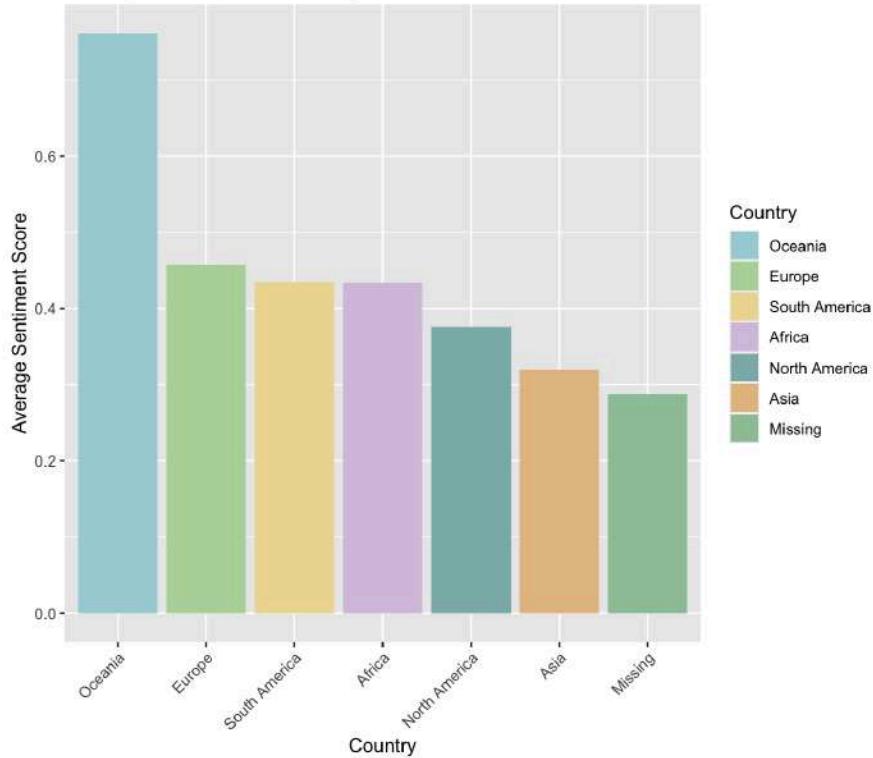
Now, we will be looking at the most used negative words:



We can observe how, when it comes to the brand category, there are almost no curse words. It specifically has words that are not negative as in violent or aggressive, but simply about difficult topics of society, such as war, death, cancer, debt... On the other hand, both male and female had more personal negative words, which are usually used in a more aggressive way and often when criticizing something, such as bad, shit, die, leave, stop... As for the description, female had more frequencies of words such as *crazy* and *b*tch* compared to male classified users. Lastly, for the unknown category, it seems that their vocabulary includes internet-based expressions like *fan*, *crush*, *noob* and *block*, which are frequently used in online contexts.

Lastly, we wanted to add another perspective looking at the average scores for each continent:

Average Sentiment Score by Continent



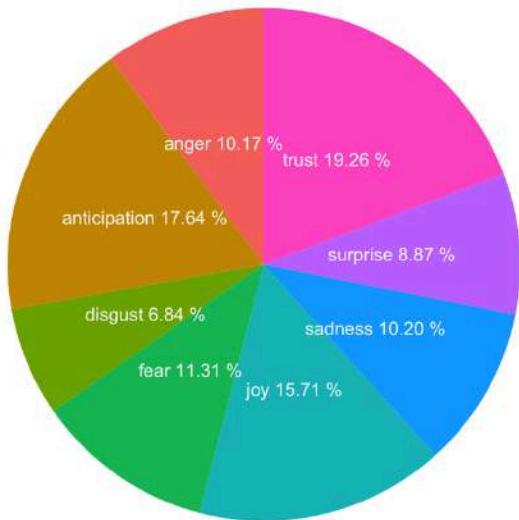
This could be representative of many things. For example, the continents with the highest average (*Oceania, Europe, South America* and *Africa*) might have a more positive culture or a better socio-political situation. As for the most negative, *North America, Asia* and those with no location disclosed, the opposite might be true.

Sentiment Scores

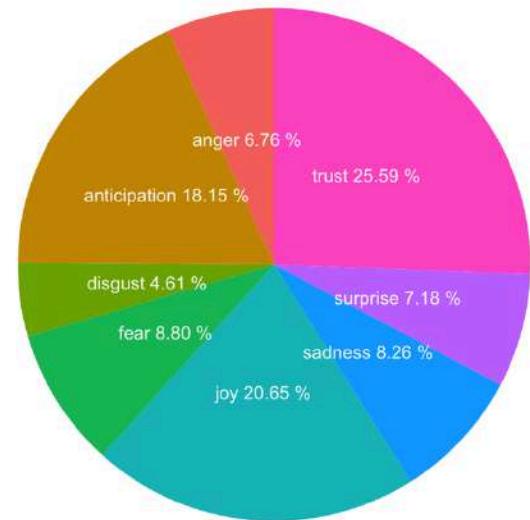
Until now, we have been analyzing using the polarity score to get our results. But it is also possible to use other sentiment scores such as a measure of different feelings: *anger, anticipation, disgust, fear, joy, sadness, surprise* and *trust*. To do so, we will use the *nrc* database, which classifies words into one of these 8 categories.

First off, we will analyze the overall proportion of these sentiment categories in our dataframe, to gain insight on the overall trends in sentiment.

Sentiment Proportion in Tweets



Sentiment Proportion in Description



Sentiment anger disgust joy surprise
 anticipation fear sadness trust

Sentiment anger disgust joy surprise
 anticipation fear sadness trust

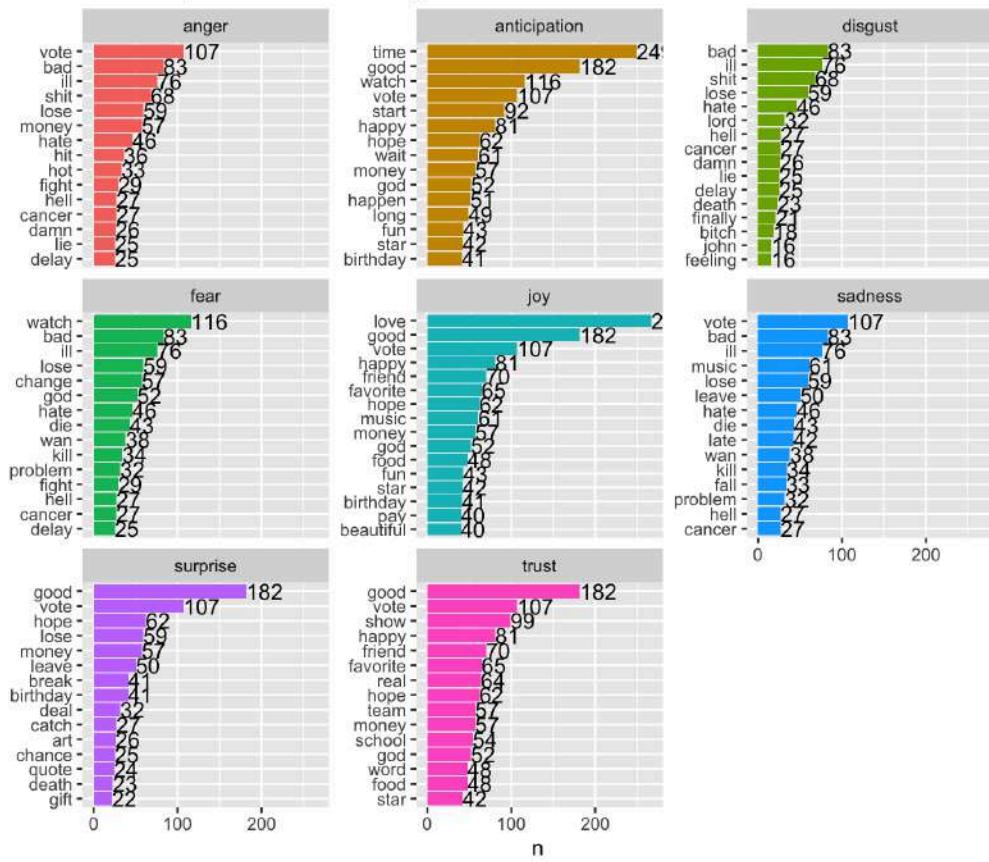
Overall, we can see how joy and anticipation are the main sentiments found in Twitter, for both tweets and description. Additionally, disgust and anger have the lowest proportions.

If we compare both, it is interesting to see how there is much more joy and trust in user's descriptions, while there is much more anger, disgust, fear and sadness in their tweets.

This tells us that people are more positive overall, as they use more positive sentiments such as joy and anticipation, and don't use negative sentiments such as disgust or anger. Still, it seems that people are even more positive when it comes to defining their descriptions compared to their tweets. This makes sense, as people tend to use the user's description to define themselves, and they are more likely to do so in a good fashion. On the other hand, Twitter is used to spread negativity a lot of times, and that is mainly done through tweets, as it is a tool for expression.

Moving on, we can also obtain a graph like this one, showing the most frequent words used of each sentiment:

Most frequent tweet words by sentiment

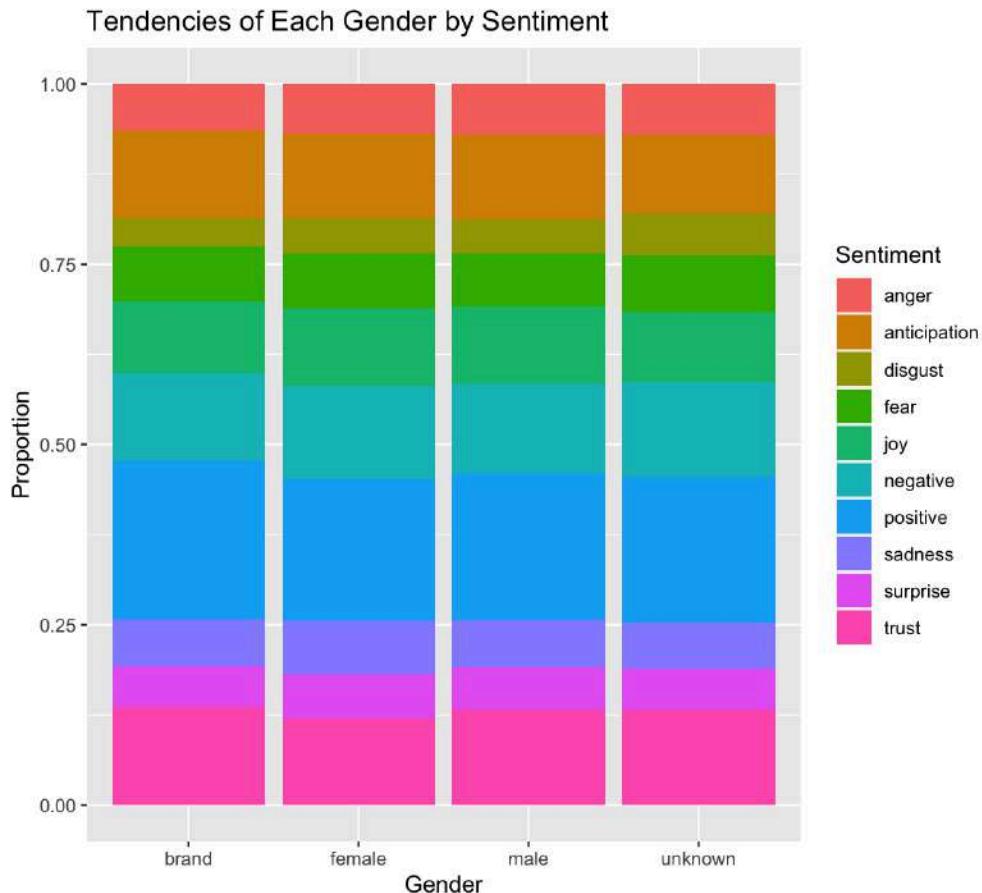


This can be very helpful in order to understand what type of conversations people are having about each emotion.

For example, what are people scared of? We know that, when tweeting about fear, they are talking about *cancer, hell, ill, god, hate, die, kill, fight...* From these results, It seems that people are worried about disease, death and war. On the other hand, what are things that people are happy about? We can see what are the most frequent words used when talking about joy, such as *love, happy, friend, hope, music, money, food, birthday...* It seems that people are mostly happy about their relationships, important events and food.

When it comes to their descriptions, we can see that they use *trustful* words such as *lover, god, official, friend, real, professional, proud, team, school...* They are typically adjectives that describe themselves positively and prove they are a real person. We can also observe a lower frequency of negative sentiments on the description, such as *anger, fear or distrust*, which are far less frequent than on their tweets.

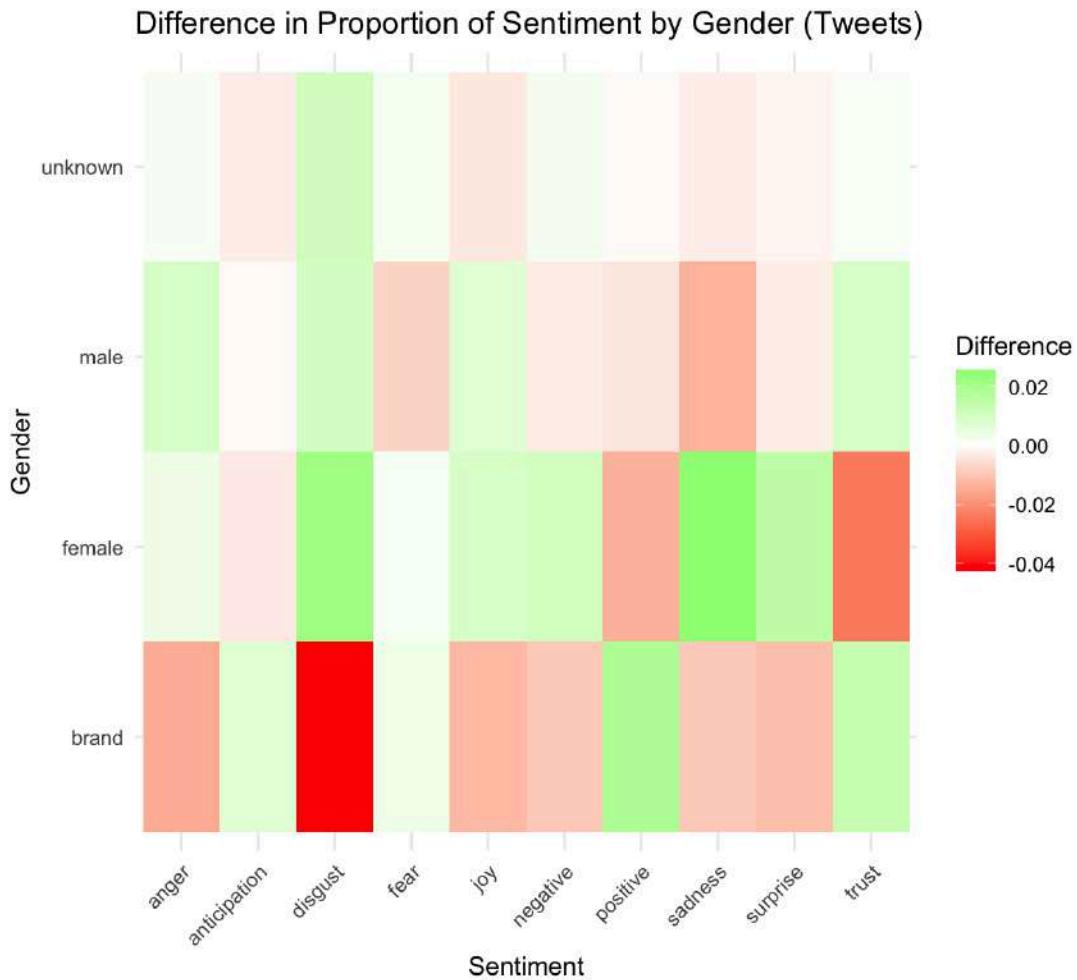
Additionally, we can see what demographic is happiest, saddest, when it comes to the words they are tweeting about. Here is a plot of what feeling each predicted gender's tweets and description convey:



As we can see, it is hard to draw conclusions from this graph due to the unbalanced frequency of tweets by gender.

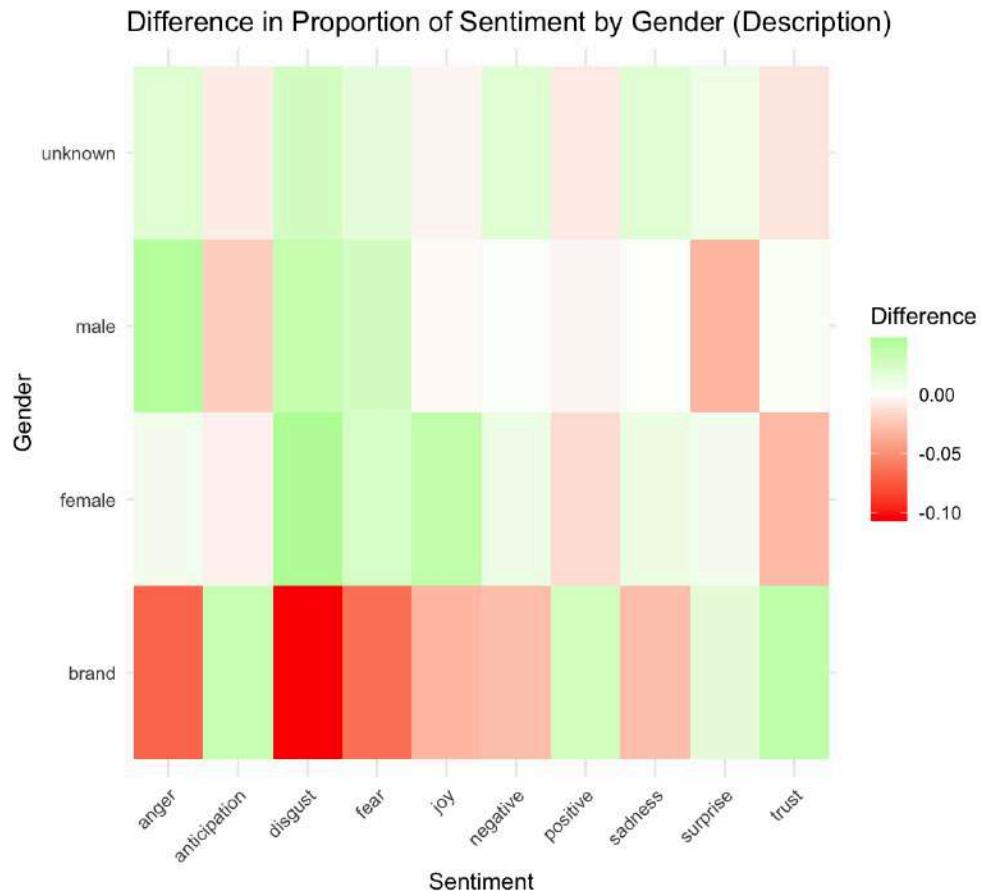
For that reason, we calculated the difference between the proportion of each gender with the proportion of the gender in each sentiment. If the difference is 0, then it means that the gender does not have a positive or negative tendency towards that sentiment. On the other hand, if it is positive it will have a positive relationship with that sentiment, and the same with the negative differences.

Here is the resulting plot for the tweets:



As we can see, for the *unknown* gender, it is more likely to express feelings of disgust, while it less likely uses words of anticipation, joy, sadness or surprise. On the other hand, the predicted gender *male* is more likely to convey feelings of anger, joy and trust, while it is less likely to use words of sadness or fear. When it comes to the *female* category, they are more likely to express disgust, sadness, surprise and joy. But they are less likely to be positive or display words of trust. Lastly, when it comes to brands, we observe a very big difference in using words of anger and disgust, as well as joy, negativity, sadness or surprise. On the other hand, they are much more likely to be positive and trustful.

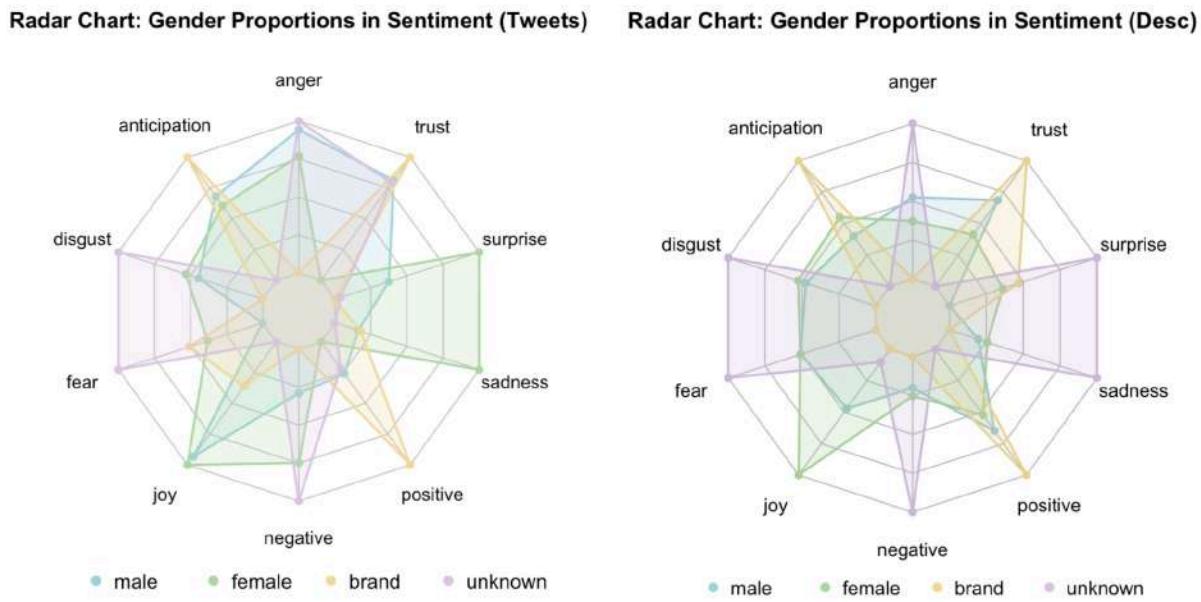
Here is the same plot for the description text:



As we can see, it follows a similar pattern than the one for tweets. An interesting difference is that they are more likely to express feelings or words of sadness and fear in their description, compared to their tweets. Additionally, they are more likely to express feelings of anger, even more so than in their tweets. As for women, they have the same differences, although they are less pronounced when it comes to their description. Lastly, for brands, we see that they have bigger negative differences for anger, disgust and fear. This tells us that, while they do express fear in their tweets, they are much less likely to do so in their descriptions. On the other hand, they are much more likely to use words of surprise in their description.

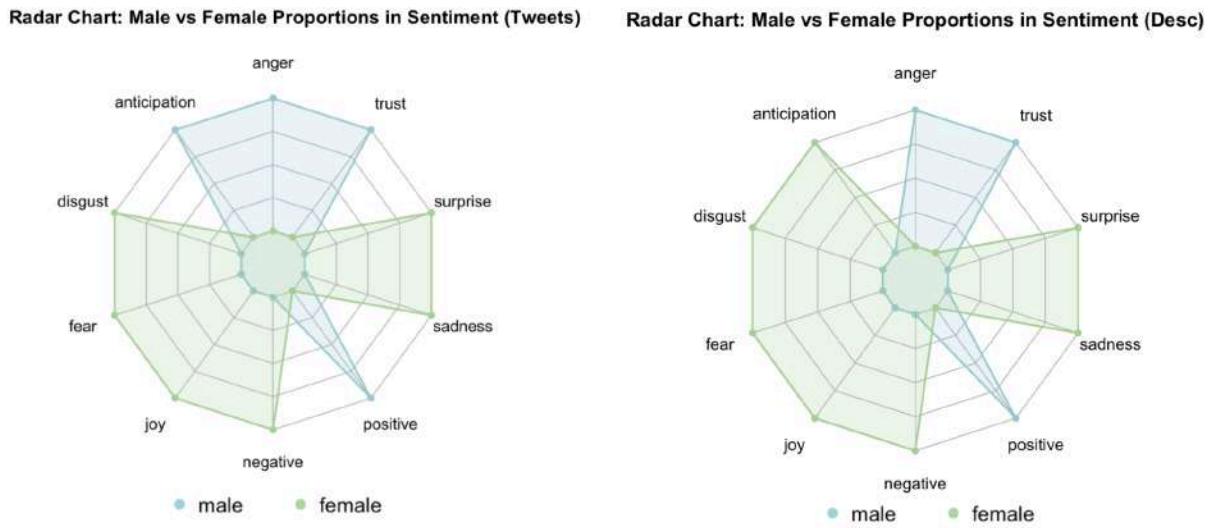
We can also observe these differences using a different visualization: a radar chart. To do so, we are plotting the relative proportions of each gender in the different sentiments in order to see the tendencies.

We will start off with the general spider chart:



As we can see, the *unknown* category seems to be in a lot of extremes, especially those that are negative. For example, they have the highest proportion when it comes to *anger*, *disgust*, *fear* and *negative* for tweets, and also *surprise* and *sadness* for description. This might indicate that these accounts are mainly *trolls* that are found frequently on the internet, especially Twitter. On the other hand, we see the *female* category dominate the sentiments of *joy*, *surprise* and *sadness* for tweets. When it comes to *brands*, they dominate *trust*, *anticipation* and *positive* emotions. As for *males*, they don't appear in any extreme.

Unfortunately, this plot has too many groups and it is hard to draw more conclusions from it. For that reason, we have plotted only the *male* and *female* categories of the *predicted gender* on a spider chart:

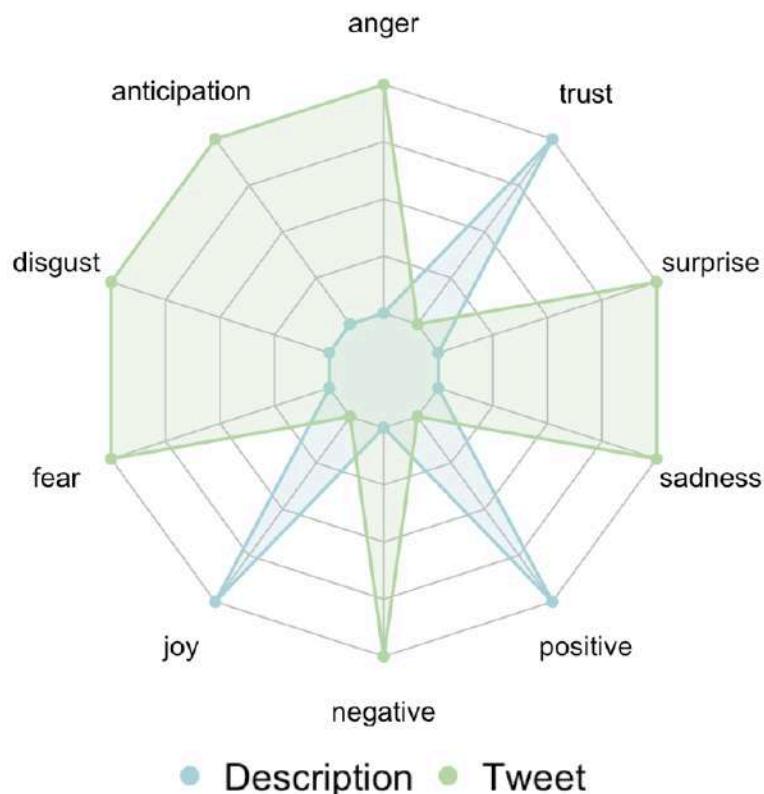


As we can observe, the *male* category uses anger, trust and positivity more. On the other hand, the *female* category uses more feelings of fear, surprise, sadness, disgust, joy and negativity. This supports what we saw earlier in the difference in proportion.

The only change we see between the description and tweets is that men have a bigger proportion when it comes to *anticipation* in their tweets and women in their descriptions.

We also wanted to check whether each sentiment predominated in people's tweets or descriptions. For that reason, we plotted this graph:

Radar Chart: Tweet vs Description Proportions in Sentiment



As we can see, negative sentiments such as *anger*, *disgust*, *fear*, *sadness* and *negative* are found on more proportion on tweets, while positive feelings such as *joy*, *positive* and *trust* are found more frequently on user's descriptions.

This tells us that, while people are more negative when tweeting, they tend to be more positive when defining themselves in their profile.

Now that we have analyzed these plots, it seems that there is a relationship between the predicted gender of the user and the sentiment expressed in their tweet. Let's try to back this impression with a chi-squared test, in order to see if there is a statistically significant dependency with the two:

Here are the results of performing the test on our tweets:

```
contingency_table <- table(tweets$gender, tweets$sentiment)
chi_sq_test <- chisq.test(contingency_table)

Pearson's Chi-squared test

data: contingency_table
X-squared = 54.719, df = 27, p-value = 0.001241
```

Additionally, we also performed the test on the description's text:

```
contingency_table <- table(description$gender, description$sentiment)
chi_sq_test <- chisq.test(contingency_table)

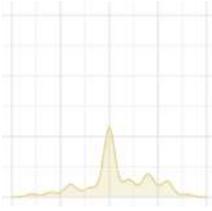
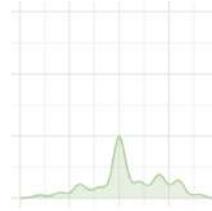
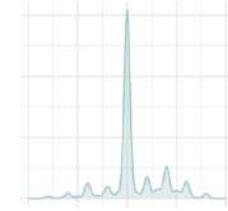
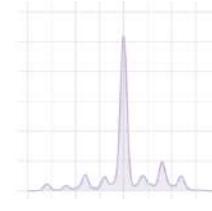
Pearson's Chi-squared test

data: contingency_table
X-squared = 268.78, df = 27, p-value < 2.2e-16
```

As we can see, both tests have a p-value that is smaller than 0.05, confirming that there is a significant association between the sentiment of the tweet/description and the gender that the user has been classified as. This tells us that, when people are classifying a user's tweet, the sentiment that the tweet and description are significant in their prediction.

Summary

We wanted to summarize the conclusions found in this chapter in this table:

	Male	Female	Brand	Unknown
Length	Long tweets and description	Long tweets, medium description	Medium tweets, long description	Short tweets and description
Unique Words	Typically masculine words (<i>bruh</i>)	Typically feminine words (makeup, starbucks...)	Professional and scientific words	Online jargon
Polarity Scores				
Sentiment Scores	+ Anger, Disgust, Joy, Trust - Sad, Fear	+ Sadness, Disgust - Positive, Trust	+ Positive, Trust, Anticipation - Disgust, Fear, Anger	+ Disgust, Anger - Trust, Anticipation
Description vs Tweets	More positive in their descriptions			

As a conclusion, we can say that there is still an association between gender and sentiment.

LSA

LSA (Latent Semantic Analysis), also known as Latent Semantic Analysis, is a powerful technique widely used in natural language processing and information retrieval. Its purpose is to uncover meaningful patterns and relationships between words by analyzing the semantic structure of a collection of documents or texts.

The core idea behind LSA lies in representing the texts mathematically using a matrix that connects the documents with the terms they contain. Through a process called singular value decomposition (SVD), this matrix is transformed to reduce dimensionality and extract the most significant features. By doing so, LSA captures the underlying semantic information and enables comparisons between documents based on the meaning of words, rather than relying solely on exact word matches.

In this section, we only present the work done with the "description" variable since the "text" variable (individuals' tweets) either yielded similar exploration results to "description" or had limited interpretability. This is because they were more challenging to preprocess properly.

Distance between documents

	Brand	Female	Male	Unknown
Brand	0.003885232	0.00193392	0.002016844	0.001284431
Female	0.00193392	0.004105585	0.003273889	0.00179327
Male	0.002016844	0.003273889	0.003365631	0.001406511
Unknown	0.001284431	0.00179327	0.001406511	-0.001004335

We have divided the dataset by gender. In each partition, we have computed the average distance between documents in relation to other partitions. Specifically, we have taken all the descriptions of men and compared them with those of women. The similarity between descriptions is relatively low. It is worth noting that the descriptions among women show the highest resemblance, while the descriptions of men and women are closer to each other than to

brand and unknown. Unknown exhibits the least similarity in its descriptions compared to the others.

Relationship between words

In this section, we have created a list of words that we consider interesting and have searched for words that are associated with them, meaning they are closer in the term matrix. To select this list of words, we have relied on the results of sentiment analysis and, at the same time, have searched for topics related to user descriptions on Twitter, such as politics, ideology, celebrities they follow, and so on. Based on these topics, we have chosen which words to explore.

Politicians

obama		trump		clinton	
foxnews	0.699	makeamericagreatagain	0.763	hillary	1.000
hick	0.699	bluelivesmatter	0.695	proguns	1.000
likud	0.699	godhelpamerica	0.695	prostaterights	1.000
nationalism	0.699	obamasucks	0.695	prolife	0.877
partysaudiisis	0.699	donald	0.669	american	0.385
pres	0.699	lettrumpbetrum p	0.644	believe	0.336
truthin	0.677	trumpian	0.644	mother	0.309
misinformation	0.677	voting	0.644	humanrights	0.280
dislike	0.409	immigration	0.426	change	0.276
president	0.232	bencarson	0.407	genuinely	0.239
evil	0.200	sencruz	0.407	proud	0.231

When examining the words related to politicians such as Obama, Trump, and Clinton, we can observe certain patterns and associations. However, it's important to note that these associations

are based on a limited set of words and may not fully capture the complexities of their political careers or public perceptions.

Regarding Obama, some of the words associated with him include media outlets like Fox News, his role as a president, and political parties like Likud (although this association may not be accurate as Likud is an Israeli political party). Additionally, there are negative terms like "evil," "dislike," and "hick" that express animosity towards him.

In the case of Trump, the identified words include slogans such as "Make America Great Again" and "Blue Lives Matter" (which refers to support for the police). There's also a mention of Trump's book titled "Let Trump Be Trump." It's worth noting that these slogans represent some of the key messages associated with Trump's campaign and presidency. Additionally, derogatory terms towards Obama like "Obama sucks" are mentioned, along with references to Republican politicians like Ben Carson and Senator Cruz.

Concerning Clinton, the words associated with her primarily focus on ideological positions such as being "pro-guns," "pro-life," and supporting "states rights" and "human rights." The terms "mother" and "believe" are also mentioned, although their relevance may not be as certain.

Country

spain		usa		germany		british	
ram	0.695	fangooodsdisclamer	0.538	wan	0.617	mattjcom	0.568
latin	0.620	han	0.538	oitnb	0.609	consul	0.507
total	0.552	varley	0.491	ireland	0.572	saskatchewan	0.507
ale	0.524	handcraft	0.467	ouat	0.544	amdg	0.454
visual	0.517	cardinalnation	0.449	republic	0.513	jesuit	0.454

theatre	0.496	collar	0.449	czech	0.490	obedience	0.454
cricket	0.468	longhornnation	0.449	denyque	0.418	province	0.438
cycle	0.467	adrift	0.439	nullisecunda	0.418	whostar	0.438
horse	0.451	classroom	0.430	wetswim	0.418	arabic	0.434
atheist	0.441	fouryear	0.430	kardashians	0.409	accent	0.434
affair	0.384	packabroad	0.430	bear	0.281	hindi	0.434
communism	0.292	labor	0.408	casualty	0.265	egyptian	0.434

When examining countries, we have not discovered significant findings pertaining to the terms explored. Nonetheless, for Germany, we have identified connections with other nations such as the Czech Republic and Ireland. Regarding the term "British," it is associated with Saskatchewan, a province in Canada, as well as with Arabic, Hindi, and Egyptian. Furthermore, we have come across words like "province" and "consul" that bear some relevance.

Coding

python		developer	
androidios	1.000	tattooer	0.506
cynic	1.000	dreamersucceeder	0.506
django	1.000	cusenydc	0.466
picky	1.000	originally	0.466
ruby	1.000	plugin	0.465
intj	0.740	trainee	0.465
javascript	0.729	xenoncraft	0.465
java	0.718	lost	0.422
fanboy	0.698	tec	0.422
hacker	0.602	beatlemaniac	0.416
swift	0.583	engineeer	0.416
software	0.314	ambidextrous	0.412
caralho	0.281	immorally	0.412
engineer	0.255	ragdoll	0.412
obessed	0.219	css	0.372

The words related to Python are primarily programming languages such as Java, Swift, and JavaScript. There are also Python libraries like Django, Picky, and Cynic. Additionally, there are more words related to the world of computer science such as software, Android, and iOS. Lastly, there are words related to attributes a person may have when programming in Python, such as engineer, obsessed, fanboy, and hacker.

In the case of "developer," some words are not directly related. The clearest ones are CSS, a styling language, and plugin.

Ideology

conservative		tax		freedom	
jewish	0.555	alternatefinance	0.683	democracy	0.706
hippie	0.518	cashflow	0.683	iranfreedomi	0.653
amendment	0.489	crowdfunding	0.683	mrsmaryam	0.603
scare	0.474	finanz	0.683	rajavi	0.603
hypocrisy	0.465	goto	0.635	maryamrajavi	0.587
granddaughter	0.429	funding	0.526	enforcer	0.577
granny	0.429	citn	0.525	rightfulness	0.577
fiscal	0.356	upschool	0.525	systemic	0.577
separation	0.356	ekiti	0.525	mouthy	0.561
brat	0.341	competence	0.525	spiritualist	0.561
liberal	0.337	calculator	0.511	theosophist	0.561
capitalism	0.280	investment	0.413	humanrights	0.209

For the ideology, we find that the word "conservative" is associated with Jews, amendment, fiscal, separation, liberal, capitalism, which would be ideological elements where conservatives are either in favor or against. We also find that there are words associated with antiquity associated with the word "conservative" (granddaughter, granny).

For the word "tax," we find that it is mainly related to words associated with the economy and finance. For example, we find words like crowdfunding, funding, competence, investment, alternate finance, cash flow. We also find a connection with Nigeria, where "CITN" stands for Chartered Institute of Taxation of Nigeria, and "Ekiti" is a state in Nigeria.

Lastly, we see that for the word "freedom," we mainly find words related to Iran. These words have likely emerged from criticisms against the regime. Among these words, we find the names of important Iranian politicians who fought against the system there, such as Mrs. Maryam and Rajavi. The Iranian regime has a strong theological component, which we can observe with words like theosophist and spiritualist. Other words that are or may be related to opposition to the Iranian regime are rightfulness, systemic, human rights, and democracy.

Singers

bieber		shawn		taylor	
beif	0.760	idk	0.693	tortilla	0.706
justin	0.646	maejor	0.693	thoctober	0.642
idk	0.575	mahogany	0.693	thseptemeber	0.642
maejor	0.575	johnes	0.673	lucy	0.642
loooooove	0.569	marizer	0.673	stseptember	0.642
comee	0.534	fcp	0.673	sytycdc	0.635
justinnn	0.534	aaron	0.650	alison	0.591
youuuu	0.534	mendeserts	0.595	grammy	0.591

In the case of Justin Bieber, we haven't found any additional terms that contribute relevant meaning to "Bieber". However, in the case of Shawn, more connections can be established as he is related to other singers like Maejor and Mahogany. Additionally, in the case of Taylor, we find her songs such as "The October" and "The September". We can also mention other singers like Alison and one of the most important music awards, the Grammy.

Others

makeup		bitcoin	
vloggerblogger	0.829	revolution	0.477
gyaru	0.604	earn	0.455

ulzzang	0.604	opportunity	0.337
barbies	0.532	simple	0.334
beautyblogger	0.532	enthusiast	0.313
beautiful	0.532	hardworking	0.299
beneath	0.491	money	0.276
marilyn	0.491	developer	0.263
artiststylist	0.483	interested	0.255
womenwife	0.483	debt	0.243
roughneck	0.483	learn	0.243
muameredith	0.463	use	0.230

When we have explored words related to makeup, we have come across terms associated with influencers such as vloggerblogger, beatublogger, artiststylist, Marilyn (Monroe). We have also found gyaru and ulzzang, which are terms related to female aesthetics associated with makeup. They originate from South Korea and Japan, respectively.

Conclusions

After conducting this analysis, we have observed that the relationships found in our LSA strongly depend on the temporal context in which the data was collected. This was evident with the word "freedom," which was closely associated with Iran. Furthermore, the descriptions of individuals are heavily polarized, as we were able to compare when exploring the words "Obama" and "Trump." Lastly, it is worth noting that our text sample is relatively small given the complexity of the problem we are facing, analyzing Twitter text.

LDA

Next up, we will be performing LDA.

Before we start, it needs to be said that Twitter is notorious for its brevity, with users expressing their thoughts in 280 characters or less. Such limited space can pose challenges to topic modeling algorithms like LDA, as short texts may lack sufficient context and contain less representative vocabulary. Furthermore, the informal nature of tweets, including abbreviations, acronyms, and emoticons, adds noise to the data, potentially impacting the quality of topic extraction.

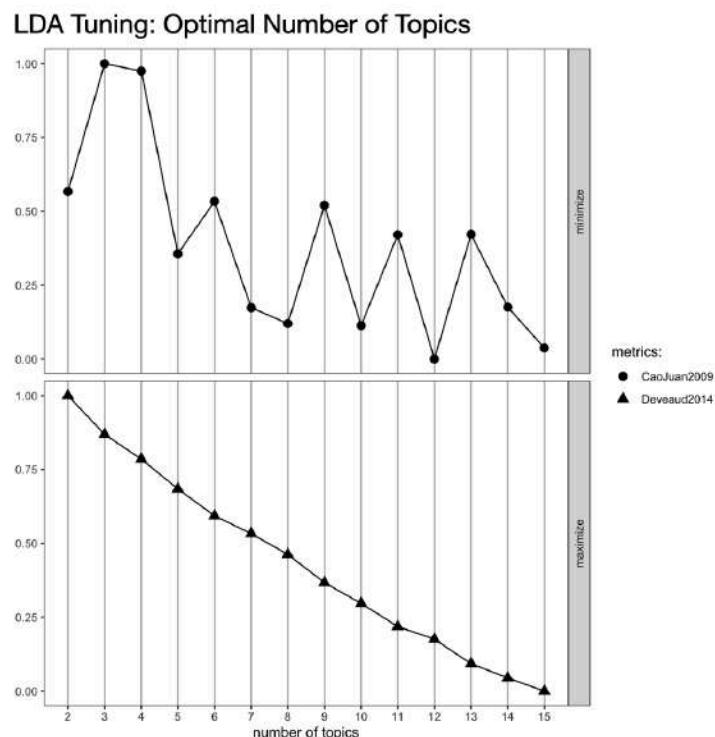
Still, we hope to gain some insightful knowledge, as this dynamic nature of Twitter provides an opportunity to capture and explore trending topics and emerging discussions.

The first thing we did was transform our database into a corpus, in order to process it into a *Document Text Matrix*. Afterwards, we applied term frequency weighting to capture relative importance of words as well as mitigating the impact of common words.

Choosing Number of Topics

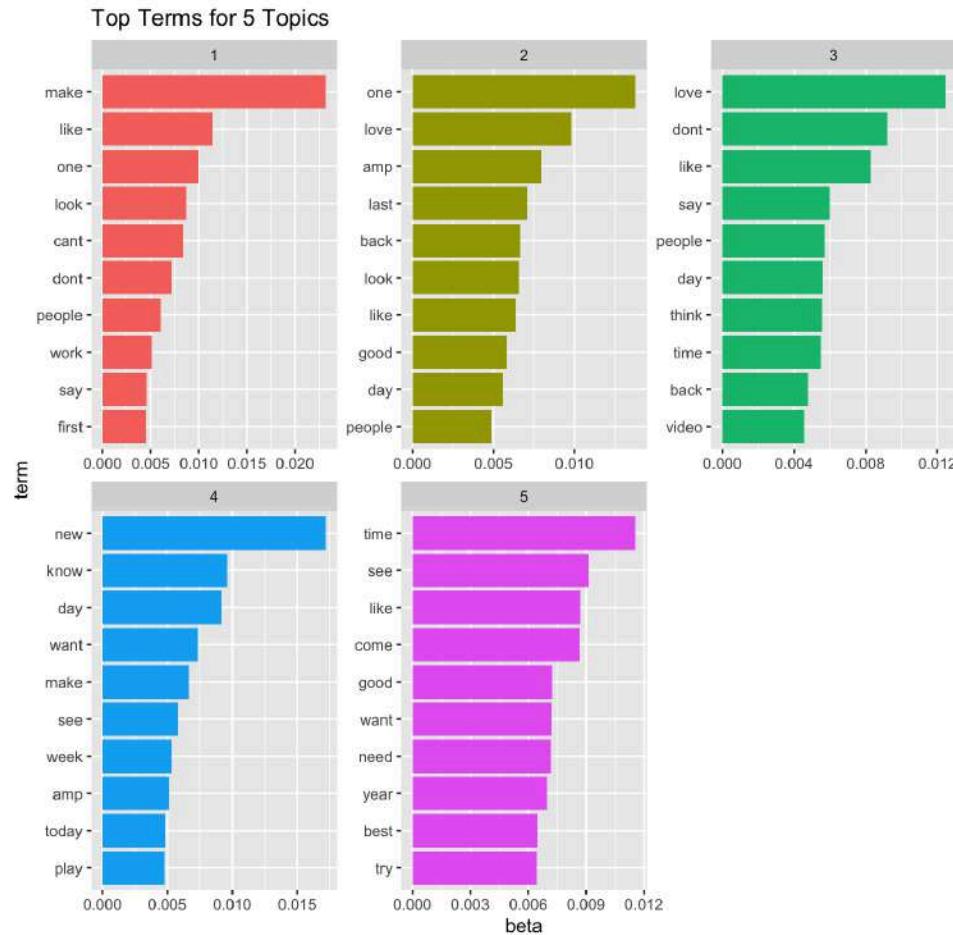
Now it is time to choose the hyperparameters, such as the number of topics we want to find. To do this, we have to check the results obtained from LSA and Document Clustering. As for LSA, .When it comes to Document Clustering, we found that 8 topics would be optimal

We can also look at this graph, which plots two metrics that can be taken into account when deciding the number of topics:



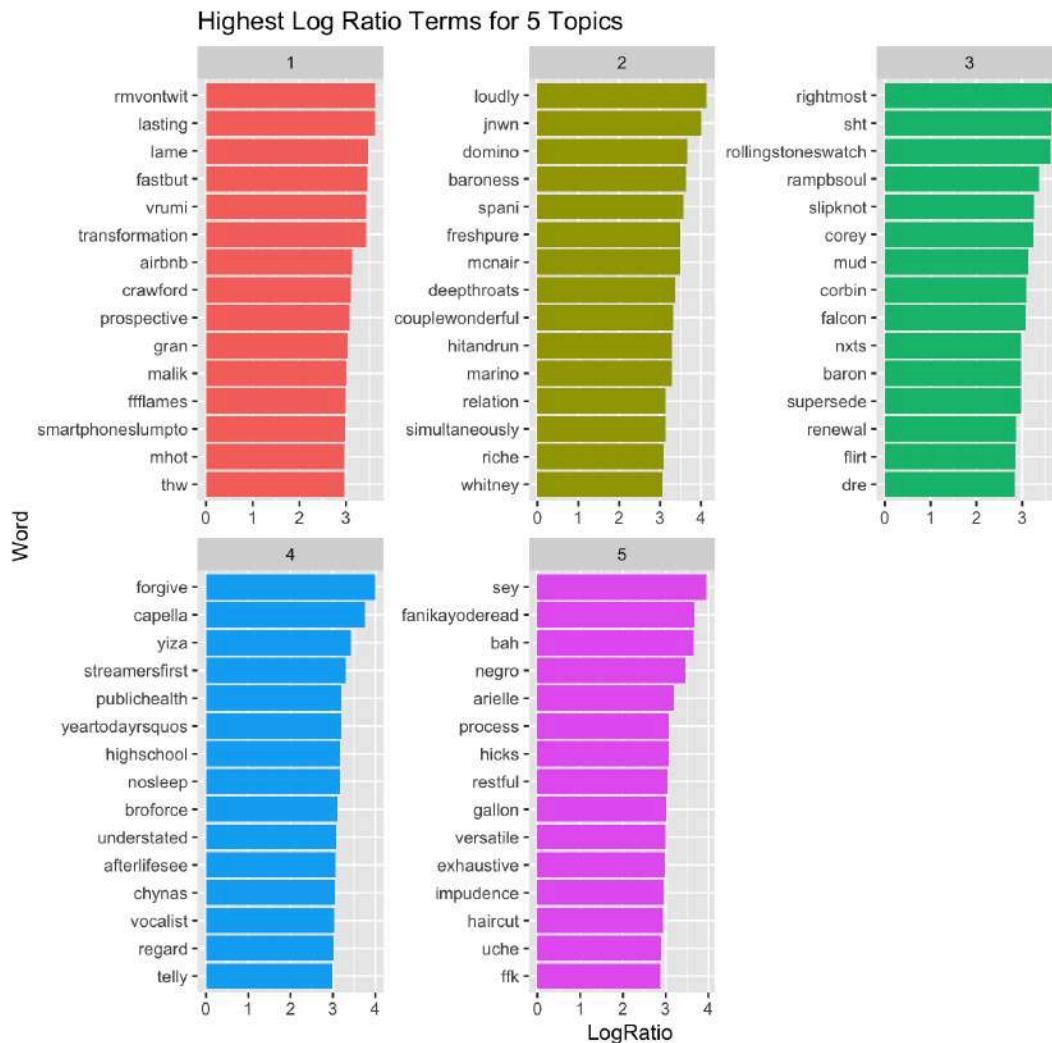
If we follow the results obtained from the graph, the most optimal number of topics is found on 5 or 8, since it points to the best combination of measures. For that reason, we will be trying these two measures.

If we perform LDA with 5 topics, we obtain these results as the top terms for each topic:



As we can see, it seems that most words found are typical words that don't help in the definition of the topics. Some words are found in multiple topics. There are some unique words, such as *work* and *first* for Topic 1, *amp* and *last* for Topic 2, *day* and *video* for Topic 3, *know*, *see* and *play* for Topic 4, and *try*, *best* and *year* for Topic 5. Since this does not give us much information, we can also plot the words with the highest log ratio for each topic, calculated as the log of the beta of that word in that topic, divided by the average of the beta of that word in all other topics. These will help us see words that are not only found a lot overall, but found on each topic specifically and not in other topics.

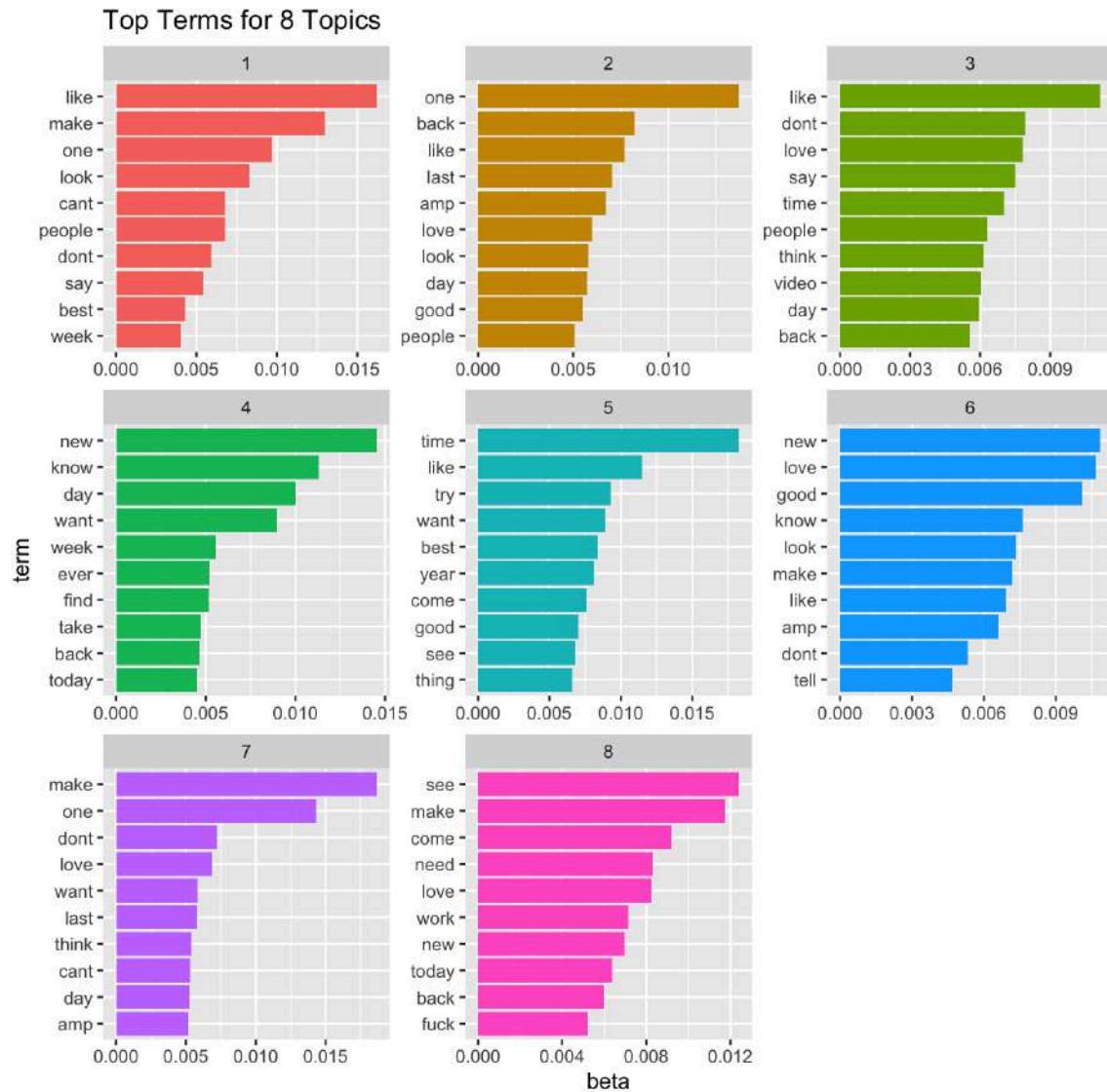
These are the results obtained:



We see words such as *lame*, *transformation*, *airbnb*, *prospective*, *crawford* on the first topic. As for the second, we find words such as *loudly*, *simultaneously*, *couple*, *relation*... As for the third one, we see words such as *rolling stones*, *corey*, *nxts*, *flirt*, *dre*... On the other hand, for topic 4 we find words such as *vocalist*, *regard*, *afterlife*, *no sleep*, *high school*, *forgive*, *public health*... On the last topic, number 5, we find words such as *process*, *restful*, *versatile*, *exhaustive* and *haircut*. Unfortunately, these words don't help paint a picture of our topics either.

Since we haven't found very interesting results using 2 topics, we will perform LDA with 8 topics now.

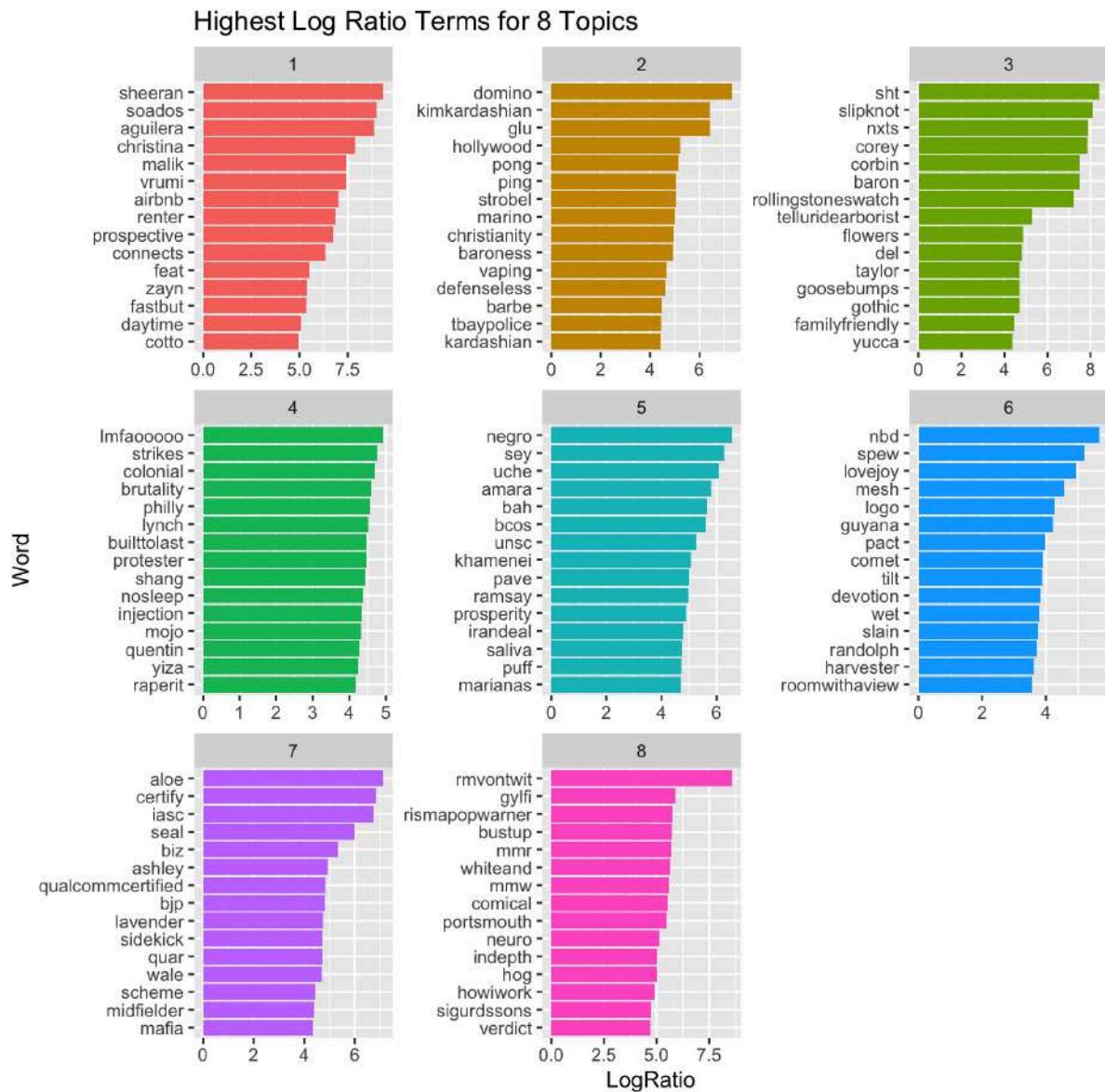
Hereunder we have included a plot of the top terms for each topic:



As we can see, we encounter the same problem here. The top words are repeated and don't give a lot of information. We have some unique words such as *f*ck* for Topic 8, *video* for Topic 3, and *week* for Topic 1. Still, we lack more knowledge to correctly identify the theme of each topic.

We will now analyze the words with the highest log ratio, in order to see if we can obtain more information.

Hereunder is the plot for the highest log ratio terms:



We can see how the first topic has a lot of singers, such as Christina Aguilera, Zayn Malik, Ed Sheeran... Therefore, we can conclude it encapsulates pop singers.

As for the second one, we find words such as Kim Kardashian, vaping and Hollywood. We also find entertainment games such as ping pong and domino. If we generalized this topic, it could be classified as *entertainment*.

The third topic also has references to music such as Taylor Swift, *goosebumps* (a Travis Scott song), Gothic, Rolling Stones, nxts, slipknot... Apart from Taylor Swift, the others could be

classified as metal or alternative music. Therefore, we could say this is the Alternative Music topic.

As for the fourth one, we find political words such as strikes, colonial, brutality, protestor, lynch, shang (the dynasty ruling in China from about the 18th to the 12th centuries)... We will classify this as a social and political topic.

Looking at the words for the 5th topics, we find the supreme Iran leader Khamenei, as well as the words *Iran Deal*, *UNSC* (United Nations Security Council, or ONU). It seems this topic is related to the Iran deal of 2015 with the US which prevented Iran from acquiring a nuclear bomb.

Moving on to the 6th topic, we find *LoveJoy*, an indie rock band, *spew*, a style of experimental music. We also see *Randolph*, which might be related to the basketball player Anthony Randolph. We also find the country of Guyana and *Harvester*, which might relate to a violent shooting video game. We could classify this as a mixture of music, sports and videogames.

As for the 7th topic, we find words related to a certification by QualComm (a big semiconductor and mobile telephone company), with words such as certify and qualcommcertified. We also find words such as mafia, scheme and BJP, which stands for Bharatiya Janata Party, a political party in India. In 2015, a gangster from a mafia killed the leader of the BJP. Additionally, after doing some research, we found that many Indian people believe that the BJP is turning India into a “mafia” republic. We also find the IASc, the Indian Academy of Science. Overall, it seems that this topic pertains to Indian politics.

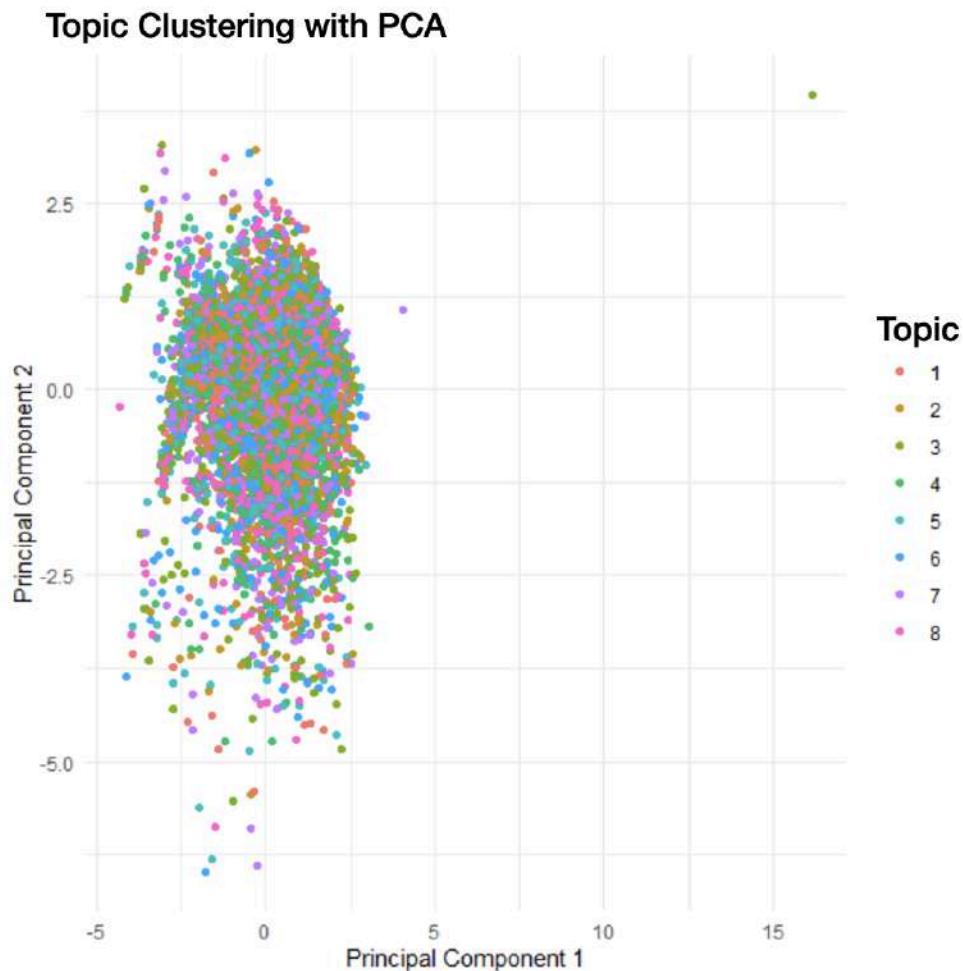
Lastly, we find a bizarre combination of two topics: On the one hand, it seems to talk about the history of the first Scandinavian king, *Gylfi*, as well as the word *sigurdssons*, which we believe refers to Sigurd Ragnarsson, a scandinavian viking warrior. On the other hand, it talks about gaming words such as MMR, which stands for match making rating and MMW, a commonly used gaming mouse.

As we can see, it seems that each topic is either a reference to entertainment, music, video games, politics and sports. Unfortunately, some combine themes that seem unrelated with each other. But that is to be expected when dealing with text from Twitter, as it will not be able to encapsulate the entirety of topics displayed.

Topic Visualization with PCA

To check if there are some unseen patterns in our topics, we performed a PCA.

Hereunder is the plot of the PCA on the first two dimensions:



Unfortunately, we find no pattern in this plot. We also performed a bivariate analysis in which we plotted some variables against the topic each observation pertained to. As we expected from the PCA results, they also did not add any information.

Topic Description

Lastly, we wanted to summarize the topics we have found in this table:

Number	Name	Description
1	Pop Singers + Apartment Hunting	Names of Pop Singers, as well as AirBnb themed.
2	Entertainment: TV and Games	Themes of Hollywood, Kardashians and games
3	Alternative and Rock Music	Music terms, bands and artists mostly related to alternative and rock music
4	Social Revolution and Politics	Themes of social revolution and politics
5	Iran Politics	Iran deal from the US, to prevent Iran from acquiring bombs.
6	Entertainment: Alternative Music, Video Games and Sports	References to alternative music, video games and sports
7	Indian Politics	References to BJP, mafia, IASc
8	Scandinavian History + Video Games Lingo	References to Scandinavian kings and warriors, as well as acronyms used in the video games world.

FACTORIAL ANALYSIS

Factor analysis is a statistical technique widely employed in social sciences and other fields to uncover the underlying structure and relationships among a set of observed variables. It is a powerful tool for reducing the complexity of data and identifying latent factors that influence the observed variables. By extracting common factors, factor analysis enables researchers to gain valuable insights into the underlying dimensions or constructs that drive the observed patterns.

With the intent of finding more in-depth about the relationship between the words found inside our tweets and corroborating the findings mentioned prior in our semantic analysis, in this section, we did a factorial analysis. As with the semantic analysis, two kinds of text will be treated as input, the text of each tweet and the description of each profile.

To be more precise, 2 analyses have been done. One only analyzed the words inside of each tweet and one added more depth to this previous analysis with the addition of some categorical variables into the mix.

Correspondence analysis

Starting with the base analysis, to realize a correspondence analysis, the input text must be in a very concrete format, which is a matrix form. To do so, we applied the use of the CountVectorizer function found on the superml library. This function allows us to convert the entire 7000 rows of tweet text into a count vectorizer matrix, which we will later use in this analysis.

However, due to space limitations, we can't do a count vectorizer that takes into account every word found in every tweet. As such, to do this transformation we must first determine what will be the number of common words that will cut into the final matrix.

As such, taking into account the large number of possible candidates we decided to settle with 20 words out of the available thousands, as we believe that it is a big enough number to give some real insight into the relationships of each word and explain with further depth our findings in sentiment analysis but low enough that it doesn't take a toll on the computer's resources and doesn't overcloud our interpretation, as more than a hundred will surely hinder our ability to interpret them all.

In the case of the profile description, we chose 20 for the exact same reason.

Results

In this subsection, we will analyze the resulting values and plots about the base analysis previously mentioned. We will first take into consideration all the descriptive analysis results we have obtained and then move on to the actual plots showing the relationship pattern.

Text

First of all, we need to ensure that there is a strong association between the word as we pretend to analyze their relationship with each. For that reason, we performed a chi-square which yielded these results:

Chi-squared	61092.39 (p-value = 1.961716e-10)
-------------	------------------------------------

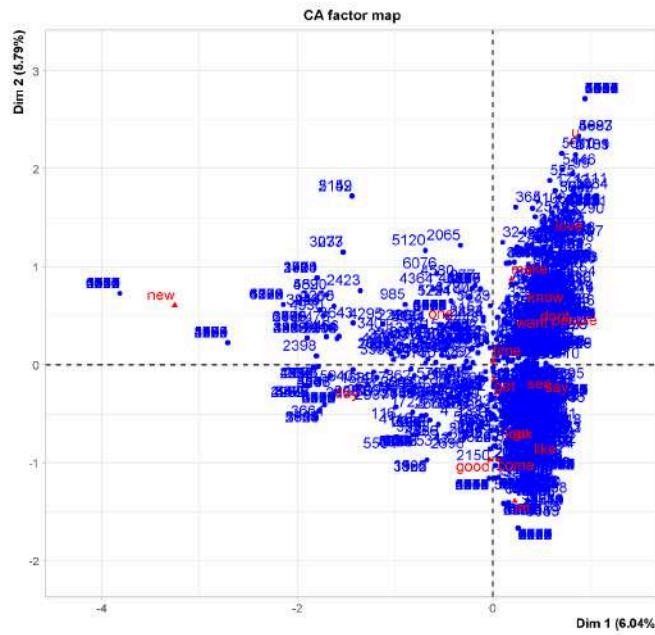
As we can see the results suggest a strong association between the words we intend on studying. This is because it implies that there is likely a significant relationship or dependence between the variables, and they are not independent of each other.

As with the eigenvalues¹, just observing a fraction of the results we can see how the most relevant dimension is the first that explains an astonishing only 6.04% of the total variance. From then on, the percentage gradually decreases, as can be seen in the second dimension with 5.79%. Further than that is lower than 4%, and as such, we decided to not consider it.

Looking further into it, we can observe which words are the ones the most well represented in each chosen dimension. For example, we can see how the word is well represented in the dimensions in the following bar plot.

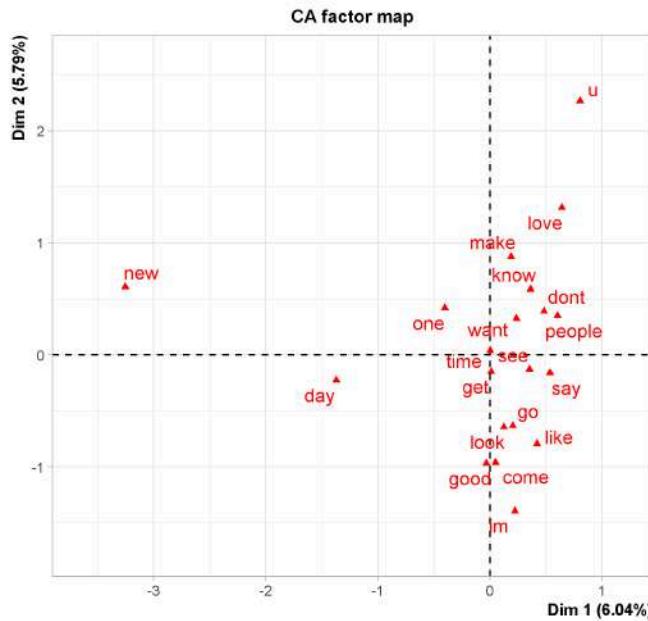
¹ All results obtained from eigenvalues can be found on the Annex 6 section.

Plots



For starters, we can see how the plot depicting with both each row and the most frequent words shows some clear patterns already. We can see a distinct pattern present between the left half and the right half of the plot, with the left being more disperse and wide and the right the exact opposite. In addition, there's also a main difference between the top and bottom halves as we can see how in there's a clear division in the border between the quadrants.

But this picture only shows us the combination of both terms we intend to study, in which you can see the row distribution but the word distribution not so much, so, let's see the latter separately to gather further insight into this situation.



We can see how a cluster with some outliers have formed. The main dispersion is to the far left, that appears to be associated with new and the like, another far north with u. The main cluster are grouped alongside the center compared to the other two.

From this information we can hypothesise that maybe the main cluster refers to words with a bigger probability that a more human candidate has written compared to the left half which contains more neutral and formal words less found in informal speech. We will examine this in more detail when we look into the results of the correspondence analysis on generalised lexical table.

Description

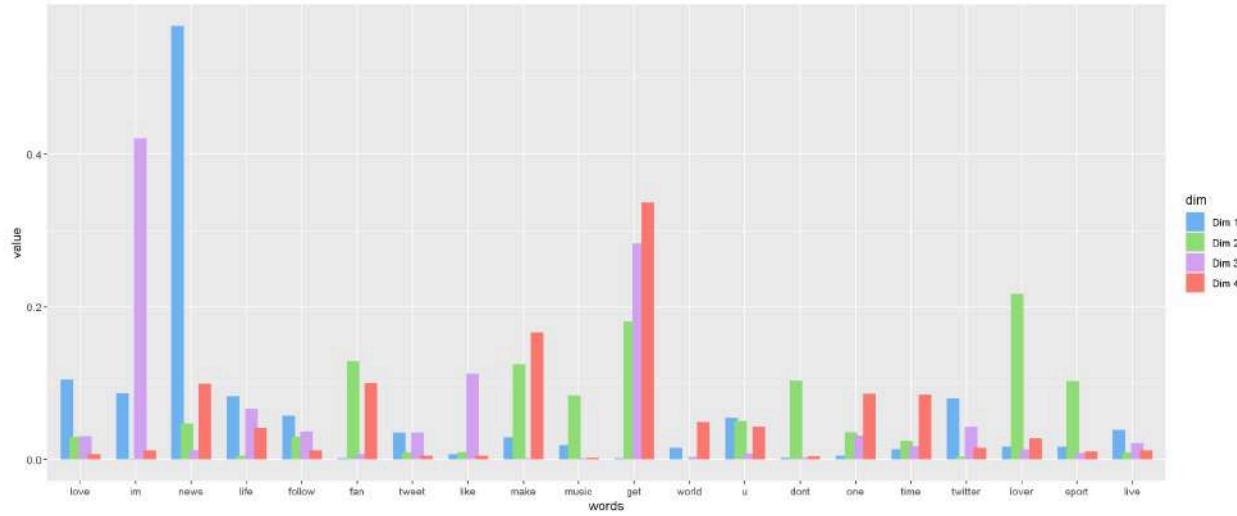
Moving on to the description results we can see that in terms of chi-square values and eigenvalues, we can draw pretty much the same conclusions found in the text portion. And here's the proof:

The chi-square of independence between the two variables is equal to:

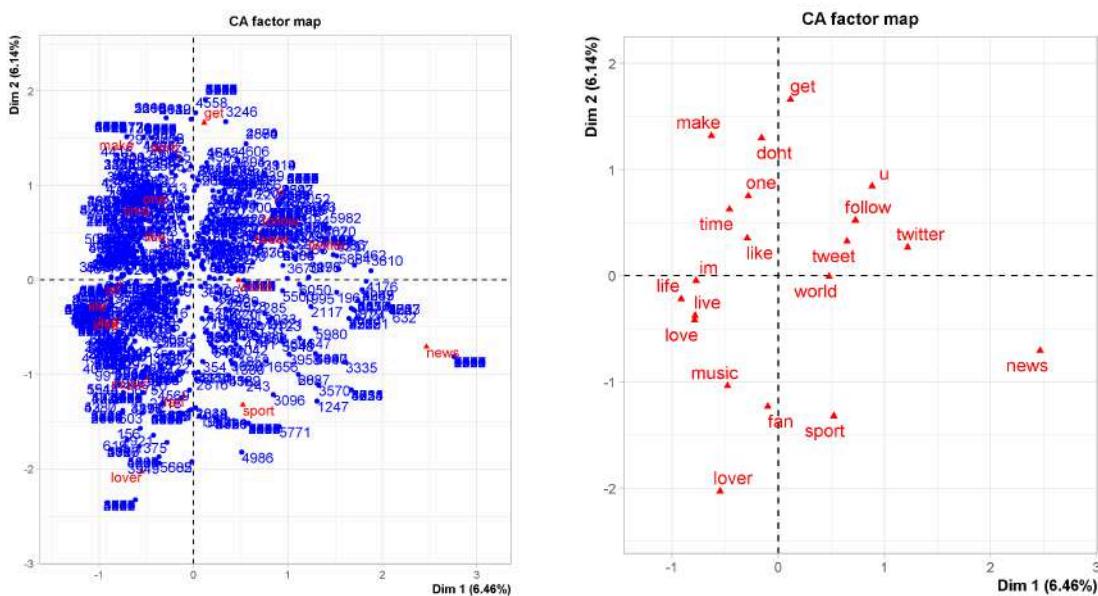
Chi-squared	46118.81 (p-value = 2.716162e-16)
-------------	------------------------------------

And in this case the eigenvalues for the two chosen dimensions are 6.46% and 6.14% respectively.

Here's the most representative words for each dimension too.



Plots



Moving on to the results, we will begin analyzing the ones obtained from the categorical variables, we can see that results compared to the text differ greatly. In this case, each quadrant appears to present its own cluster as each quadrant border is almost devoid of any point whereas the clusters are in the middle of quadrant densely packed except for the fourth which is slightly less dense and populated.

Examining closer the places of each of the most frequent words, their placement seems to also go according to the cluster density, having in the fourth quadrant the least amount of words. There seems to be a slight pattern where in the third quadrant are all words related to love, live and music, and how into the fourth there's the word world and sports and further into it news which could signify accounts who talk about current events.

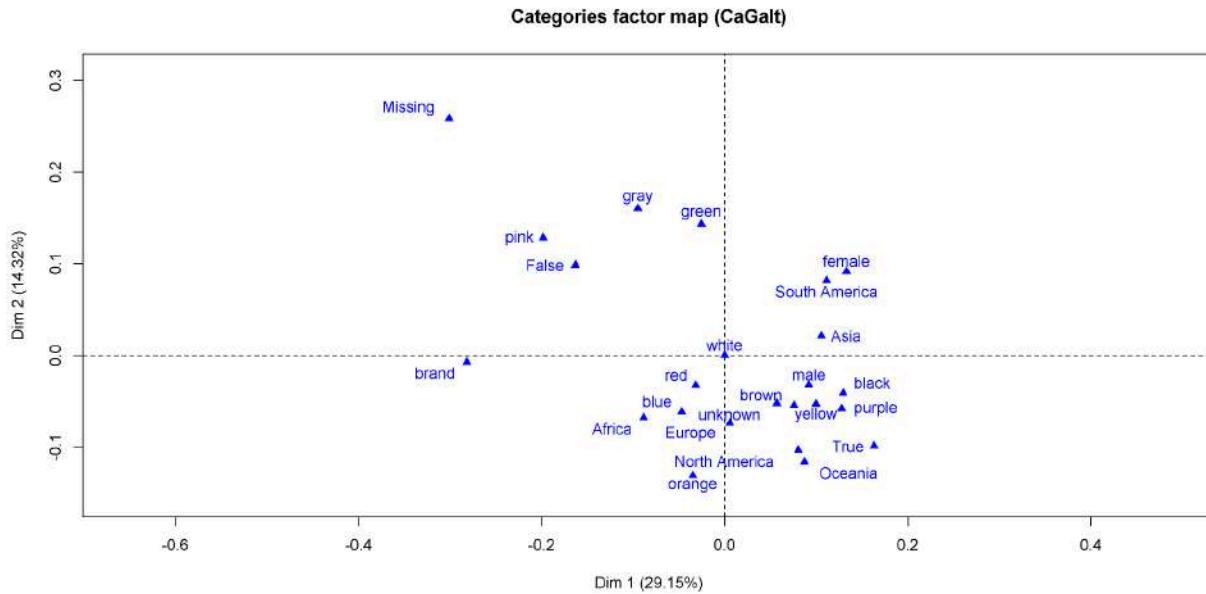
Correspondence analysis on generalised lexical table

In this subsection we will now implement an extension of the correspondence analysis using a generalised lexical table. This extension is characterized by the addition of categorical modalities with the help of an MCA or a PCA depending on their type. This way, we can now observe in the plot directly the associations between different them and the different kinds of words.

To ensure that it is reasonable to move on with this extension, we also did a quick peek into eigenvalues just as before.

Text

Beginning with the qualitative variables plots, we obtained from the text how almost all variance just comes from the first dimension with an 29.15% while from the second dimension onwards it steadily decreases from 14.32% in the categorical analysis and in the numerical.

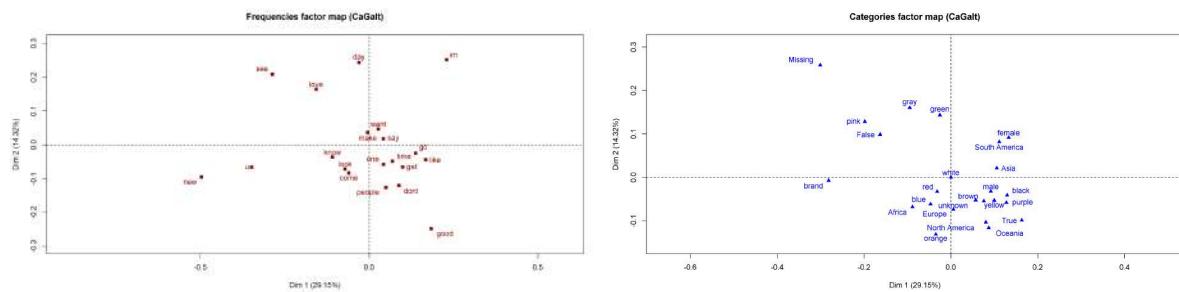


As for the results, we obtained four plots (two for each pair of dimensions) to show how all the modalities, rows, and words are scattered.

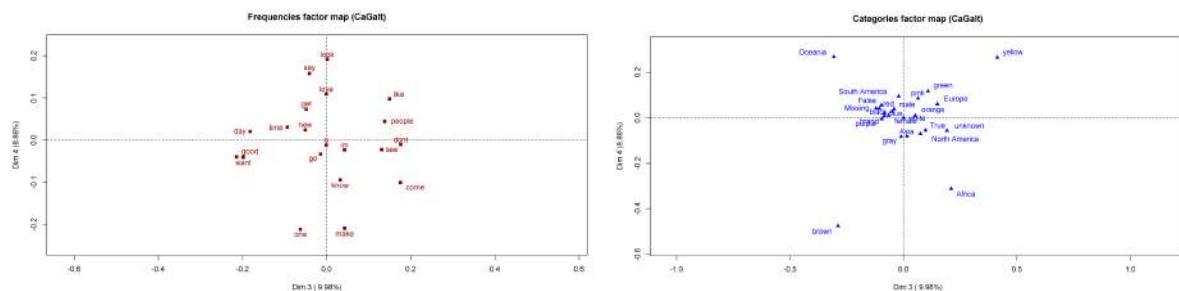
With the first pair, we can observe a clear division between the associations of the gender modalities male and female with the gender modality brand, with unknown being located in the middle. Now looking vertically we can also see this kind of disparity between colors, especially grey, pink and green at the top and orange at the bottom for the link color. And diagonally, in opposite ends there's the modality of having privacy to the top left and not having it at the bottom right.

In addition, there's also the Missing modality in the continent variable which is by far the most far from the center at the top-left.

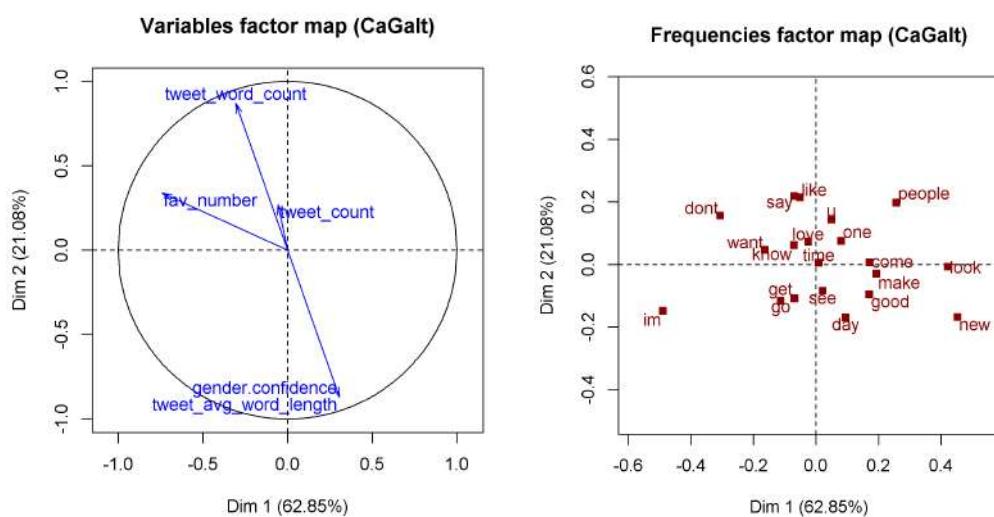
From all these observations we can gather that being a certain gender or another can also greatly impact how much privacy you decide to have. For instance, the gender's on the left (brand) have no privacy, which makes absolute sense as it is in the companies best interest to make themselves known on a wide scale. And on the opposite side, gender's male and female are from individuals who are more likely to protect their identity. And as for unknown, that it is right in the fence of both sides just proves further this assumption.



Onto the words, the main thing is the clusters observed earlier, so remembering that bit of information mentioned earlier, we can prove true our hypothesis where people think that brand profiles have more tendency to use more advertising friendly wording such as new, u, look etc. and how they think a human profile tends to a more personal lexicon such as using im.

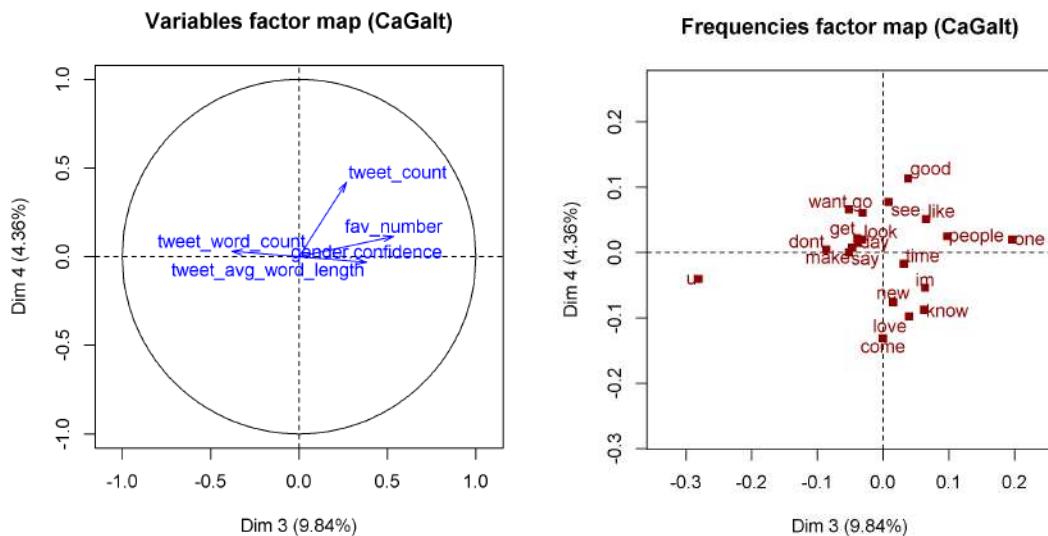


The second pair of dimensions unfortunately doesn't seem to bear a lot of information that hasn't already been told in contrast, as it only shows the diagonal oppositions between the modalities of Oceania and Africa and yellow and brown link colors.



As for the quantitative variables, their eigenvalues for the selected dimensions are 81.82%, 15.21%, 1.53% and 1.16% in that order.

Just looking into the first dimension pair, we can clearly see that the average tweet word length and the gender confidence go in the opposite direction as the rest of the numeric variables by a lot of length. We can also appreciate some words that seem to not share any association with the numerical variables such as im or people.

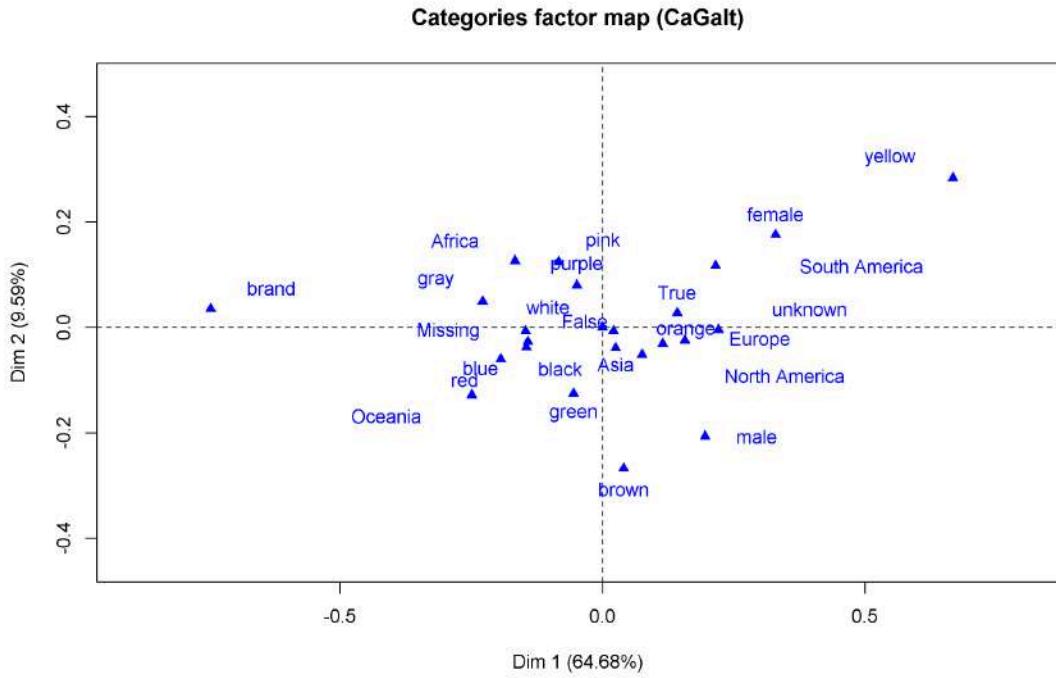


Unfortunately, looking at the second pair, just like with the qualitative variables, it doesn't seem to be any kind of useful association, as for instance, one would be the word 'good' and the number of tweets.

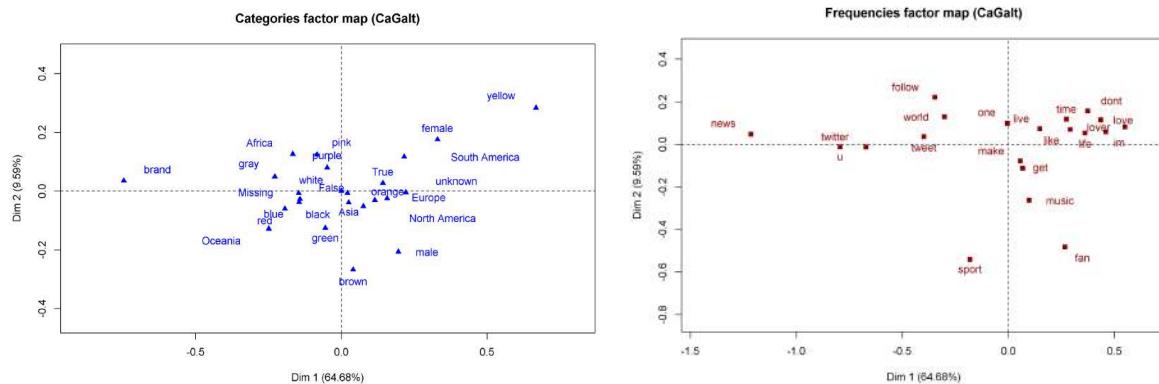
Description

As with the description here are the eigenvalues we obtained:

We can see how the first dimension now holds less dominance in terms of variance percentage with now a 64.68% but the next dimensions compensate for it with the percentage inside the 4 and 9.59 intervals.



With the first pair of dimensions, most of the modalities are distributed across the center although there are a few notable exceptions. These would be the modalities of the gender brand, male and female and the brown, orange and yellow for the link colors. We can see the same gender distribution already mentioned in the text results, with the main difference that now the unknown gender is located in the middle between the modalities male and female.

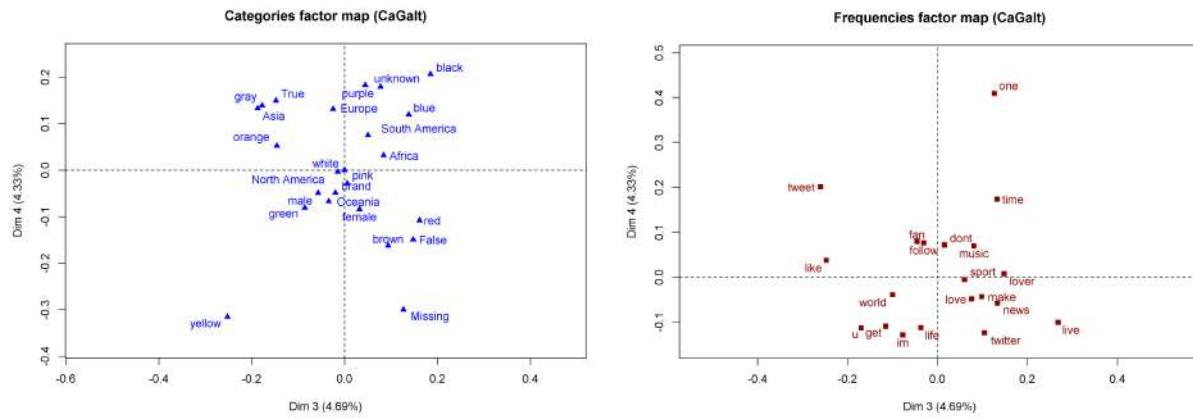


Regarding the words, we can see some obvious associations between their placement and the results seen in the previous plot. For instance, words such as twitter, news, or world have almost

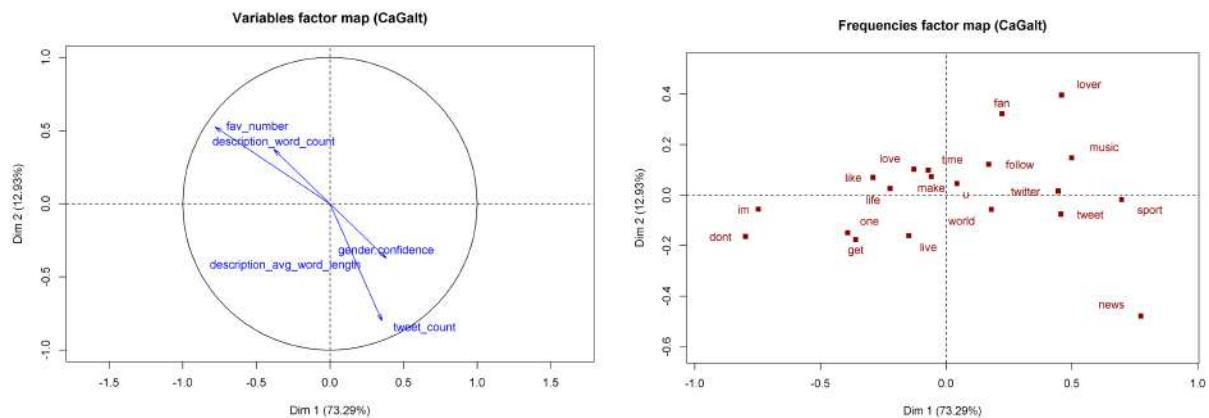
the same placement as the brand gender, meaning that this kind of topic must be a popular association words for these kinds of profiles.

Another clear relationship is between the female gender and the word lover, love and like which have almost the same placement. Which contributes to the findings mentioned before on the sentiment analysis about how influenced people are with societal stereotypes on gender.

On the contrary, the male gender is situated on the fourth quadrant with words such as like, music, or fan and the brown and yellow link color. There's also the word sport near it, which again reinforces the idea about stereotypes mentioned in the paragraph above.

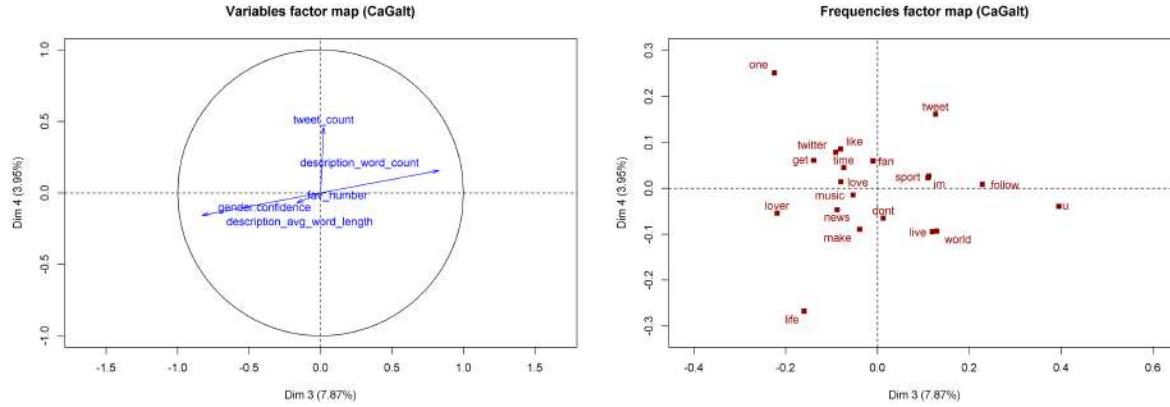


As far as the other dimension pair goes, the only information extracted from it is that words like live or news are associated with low privacy.



Looking into the numerical variables, we have the following variance percentages: 71,79%, 13,89%,

Curiously enough, we can see how the variables of fav_number and description_word_count are diagonally opposed to the rest, especially gender, confidence and tweet_count. And the only apparent information we can extract is that tweets with the word news may correlate with having a higher tweet count. There's also the fact that the majority of words aren't associated with any of the numerical variables.



As far as this dimension pair goes, we can also see how a major part of the words aren't associated with these quantitative variables. Additionally, we can see how the relationship between the numeric variables present in the first pair of dimensions still lingers in this pair too.

CLUSTERING DOCUMENTS

Clustering the Tweets

For the first clustering, we will use the tweets of our database, they can be found in the variable *text*.

Word Cloud

First we will look at the word cloud generated by our tweets after preprocessing them (removing stopwords, punctuation, duplicated tweets, numbers and other symbols). The word cloud is a visual representation of the most frequently used words in a corpus. The plot below shows 80 words maximum.



The most popular words, as we can see, are *like*, *one*, *time* and *love*, among others. But we detect a lot of words that don't bring much meaning to the analysis, some examples being *get*, *let* and *gon*, which are words that don't bring us a lot of information about the text.

To solve this we can remove these words manually by taking a look at the initial plot. After depurating a little more this is the result:



The results of this are primarily composed by:

- Words which seem to be related to the whole **social media field**, such as *follow*, *video* or *like*.
 - **Slang vocabulary** like *lol* and other swear words, this is to be expected, since twitter is an informal social media platform.
 - We see a mix of positive and negative words too, like *good* and *bad*, *never* and *always*, which we observed are antonyms.

Term Document Matrix

The next step is to create a Term Document Matrix, this will transform our documents into numerical data that our algorithms can understand. The document generated by our data has this shape:

```
<<TermDocumentMatrix (terms: 12218, documents: 6107)>>
Non-/sparse entries: 43691/74571635
Sparsity           : 100%
Maximal term length: 89
Weighting          : term frequency (tf)
```

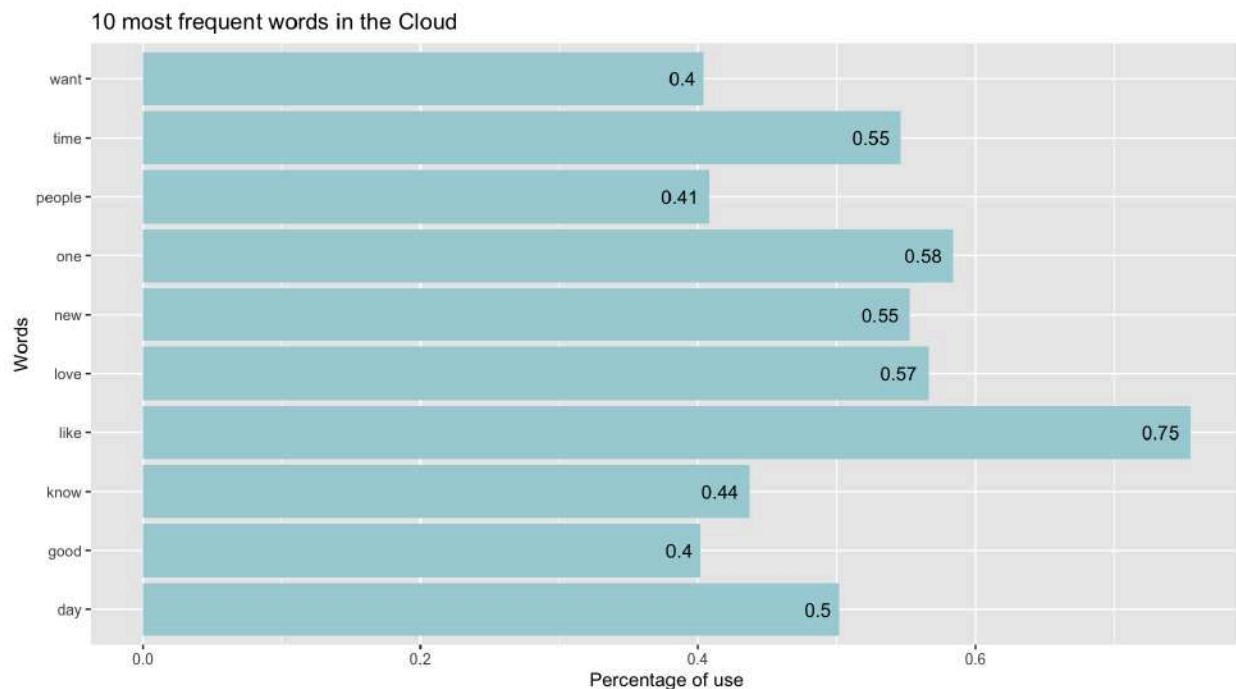
The dimension of the matrix is 12239×6107 , indicating that our corpus has 12239 different words, which are represented by the columns, and 6107 documents, which are represented by the rows of the matrix. Each cell located at ij will indicate whether the word i is present in the document j .

document represented by the column j , and how many times it is present in it, basically, the cell contains the number of times the word appears in the tweet.

With the Term Document Matrix we can easily calculate the frequency of each word and find the top words. Here's our database's top 20:

	palabra	frec		palabra	frec
like	like	340	come	come	177
one	one	263	think	think	173
love	love	255	take	take	172
new	new	249	back	back	170
time	time	246	best	best	151
day	day	226	work	work	145
know	know	197	year	year	138
people	people	184	still	still	124
want	want	182	great	great	119
good	good	181	way	way	119

We will visualize the frequency of the top 10 words. Each word will have its percentage of use associated with it in relation to the total amount of words in the tweets of the whole dataset. We can see that the most popular word by difference is *like*, followed by the words *one* and *love*.



K-Means Clustering

Determining the Number of Clusters

For the Document Clustering we will be using the K-Means clustering which is a popular method used in this area. Before we proceed, if we recall the shape of the Term Document Matrix, we have too many terms, which will slow down the performance of the algorithm, aside from that the sparsity of the matrix is 100%, this means most words rarely appear in tweets, so the matrix has a lot of zeros.

To solve this problem we can remove sparse terms, by setting a *sparsity limit* which will not allow terms of higher sparsity to be in the matrix. We have set this limit to 98%, usually one would set a lower Sparsity, but we couldn't set the limit any lower or we would end up with little to no terms. We are left with 17 terms and the sparsity of this matrix is now reduced to 97%.

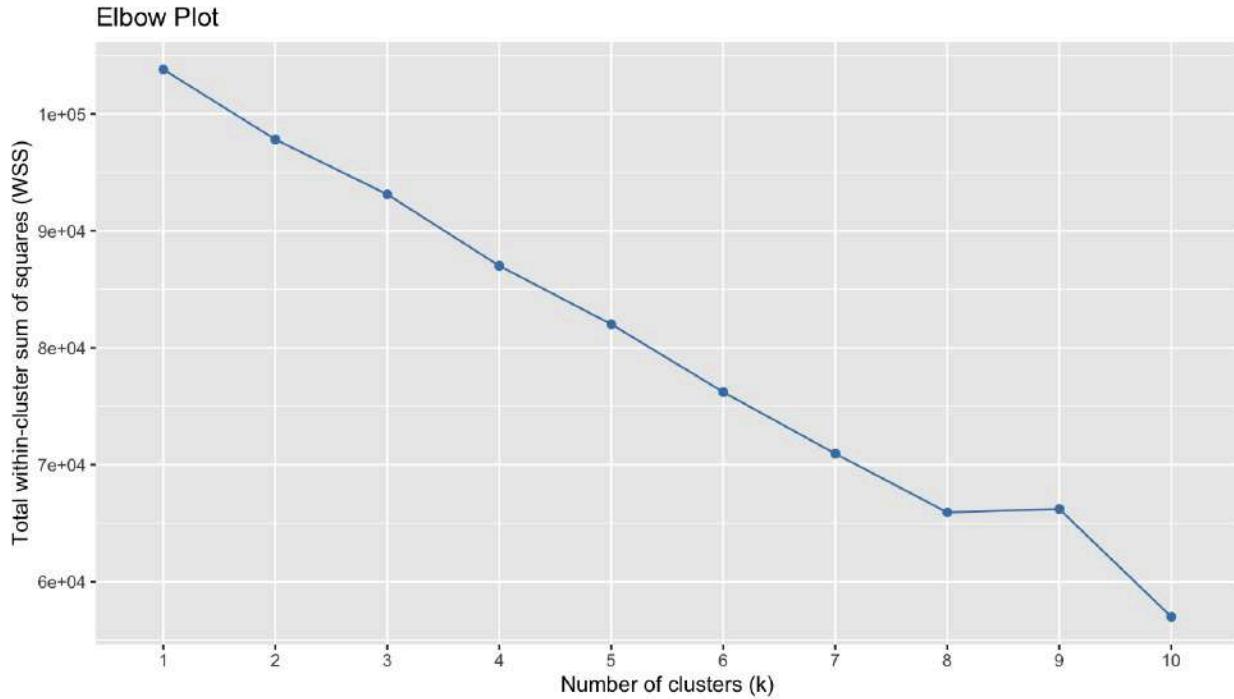
```
<<TermDocumentMatrix (terms: 17, documents: 6107)>>
Non-/sparse entries: 3263/100556
Sparsity           : 97%
Maximal term length: 6
Weighting          : term frequency (tf)
```

Next, we will transpose the matrix since we will be clustering the documents (now situated in the rows) by the amount of each term they contain (now in the columns, and can be interpreted as variables). Following this step, we will normalize the matrix to improve the clustering's performance. We used standardization or Z-score normalization.

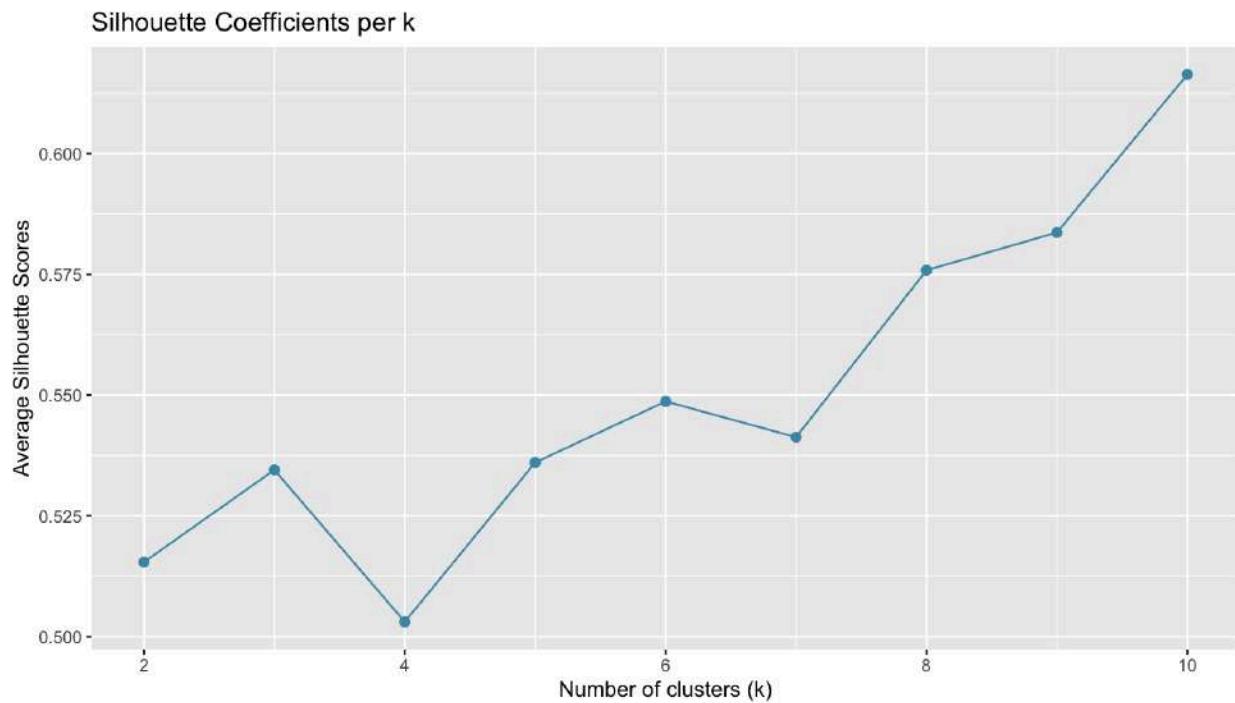
Before we create the clusters, we need to determine the number of clusters we need to create. To do this, first we decided to conduct an Elbow Plot. To create an elbow plot we need to set a range of values k we want to try, k being the number of clusters the K-Means will create. Then we will run K-Means for each k in the range we defined. For each one of these clusters we will calculate the Within-Cluster Sum of Squares (WCSS), which is a measure of the quality of the clustering that indicates the compactness of the clusters, so it is a measure we want to reduce.

The objective of this plot is to select the smallest WCSS while trying to maintain a reduced number of clusters. This plot usually follows the shape of an elbow and one would select the k that creates the point where the decrease of WCSS starts to slow down.

The plot our data generated doesn't follow this shape, it looks more like a straight line, however we can spot an elbow at $k=8$.



To further explore the optimal k , we tried to plot the Average Silhouette Scores for each cluster, the Silhouette Score is a measure of the quality of the cluster as well that measures how well each point fits within its cluster. This measure ranges from -1 to 1 and needs to be maximized. The conclusion we reached with the plot below and the plot above is that our data seems to want to be grouped into many clusters. We hypothesize that these plots didn't come out right due to the sparseness of the matrix. However we chose $k=8$, due to the small elbow point we observed.



Clustering Results

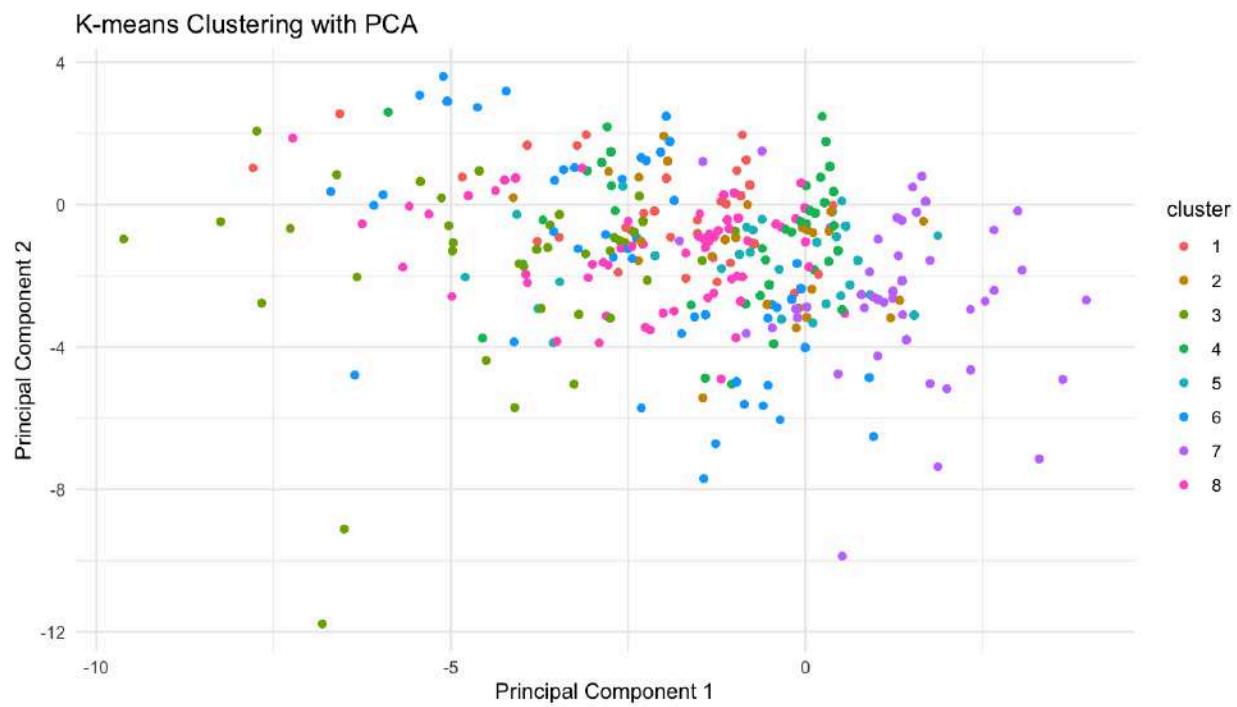
Once we have applied the K-Means clustering to our document term matrix, these are the results. The number of data points per cluster is the one below, the row above indicates the number of the clusters, and the value below indicates the number of documents in each cluster.

```
> summary(pca_scores$cluster)
  1   2   3   4   5   6   7   8
156 134 156 4441 135 295 276 514
```

As we can observe, the clusters aren't equally distributed, most of them being within a similar range, except for the 4th cluster, which contains most of the points of our matrix. This may, again, be due to the fact that we have very few terms that are sparse.

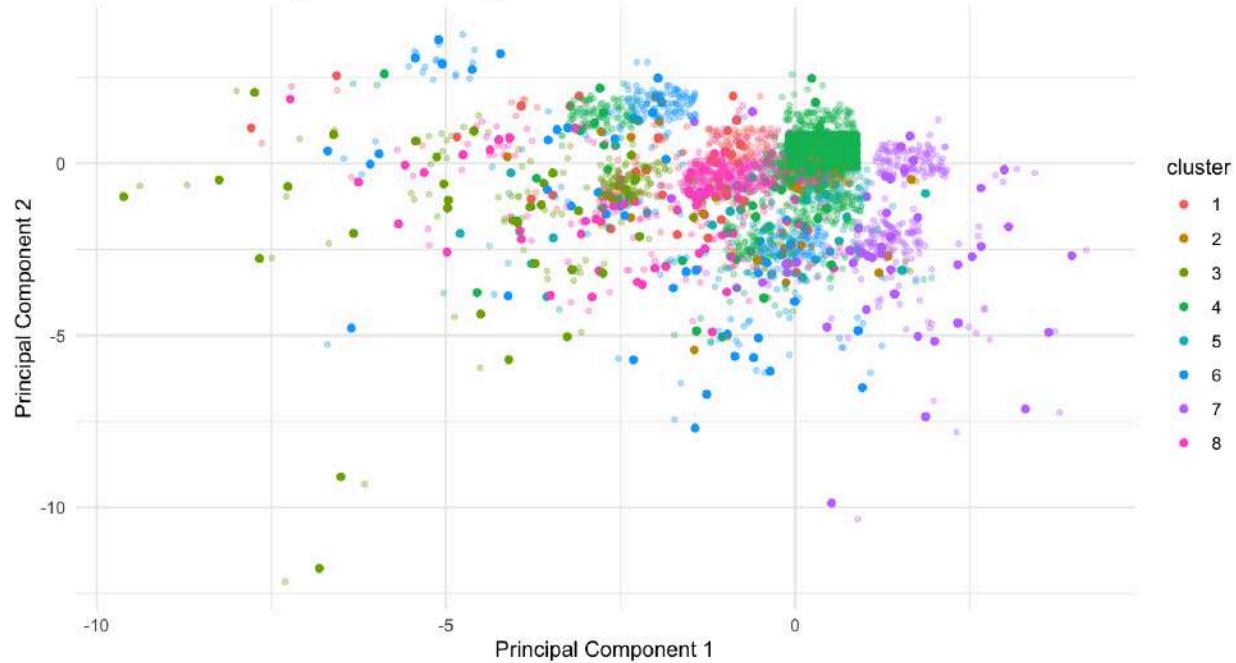
Visualizing with PCA

Nevertheless, we visualized the clustering with PCA, we created the dimensions using the Document Term Matrix. The result is below. At first glance, it doesn't look like there are over 6000 points in the plot, this is because the points are stacked on top of each other, due to the sparsity and scarcity of terms.



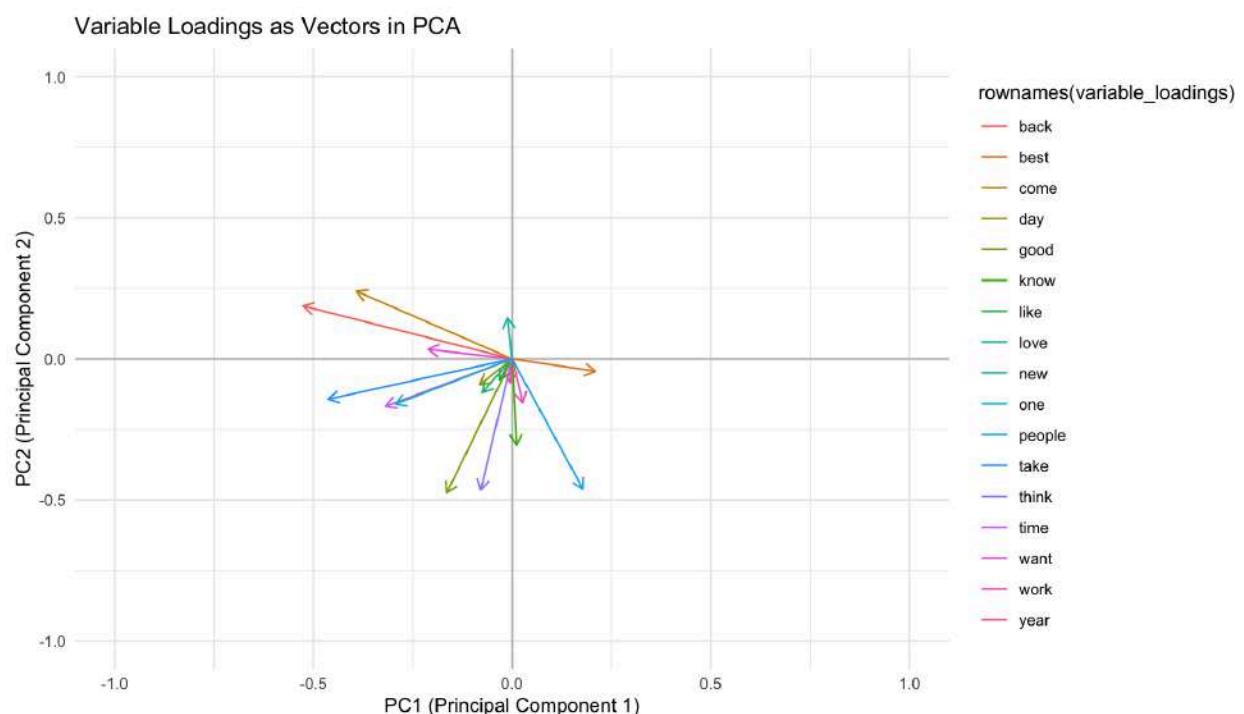
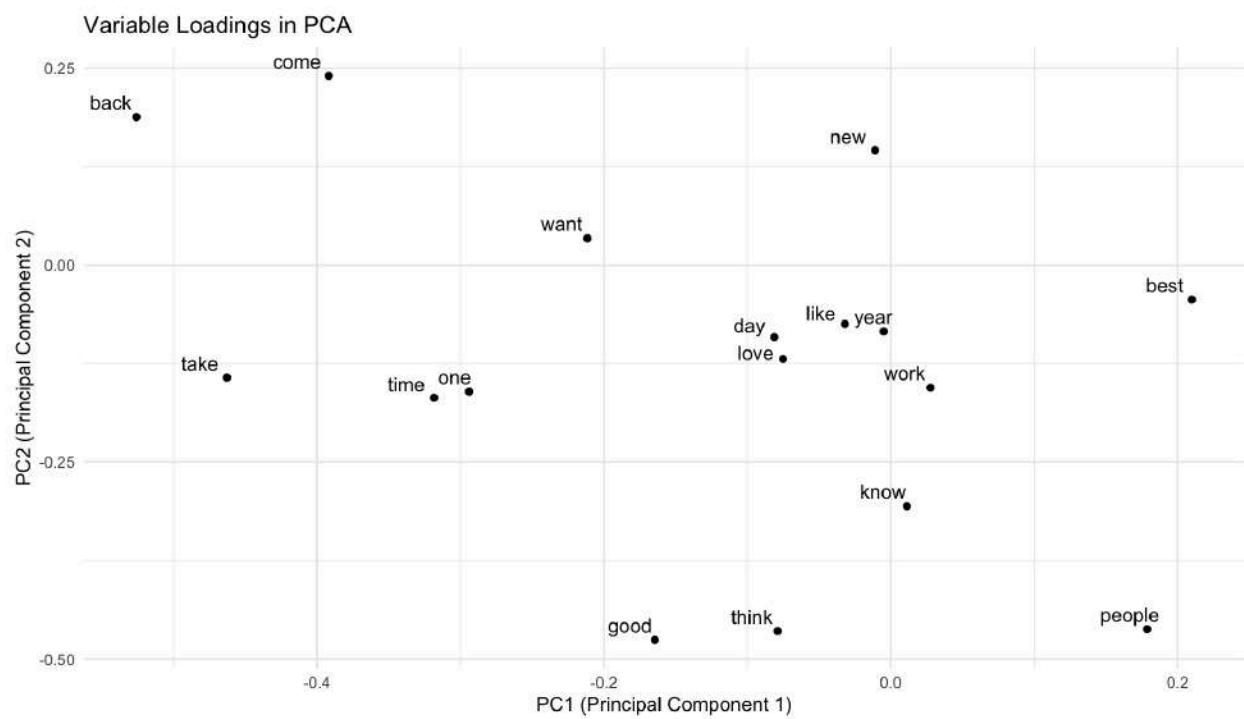
A better way to visualize the clusters in this case, will be to use “jitter”, which is a mode that randomly scatters the overlapped points to a nearby area within a radius that can be adjusted to allow the visualization of the number of points that are stacked. This is the result of adding jitter:

K-means Clustering with PCA adding Jitter

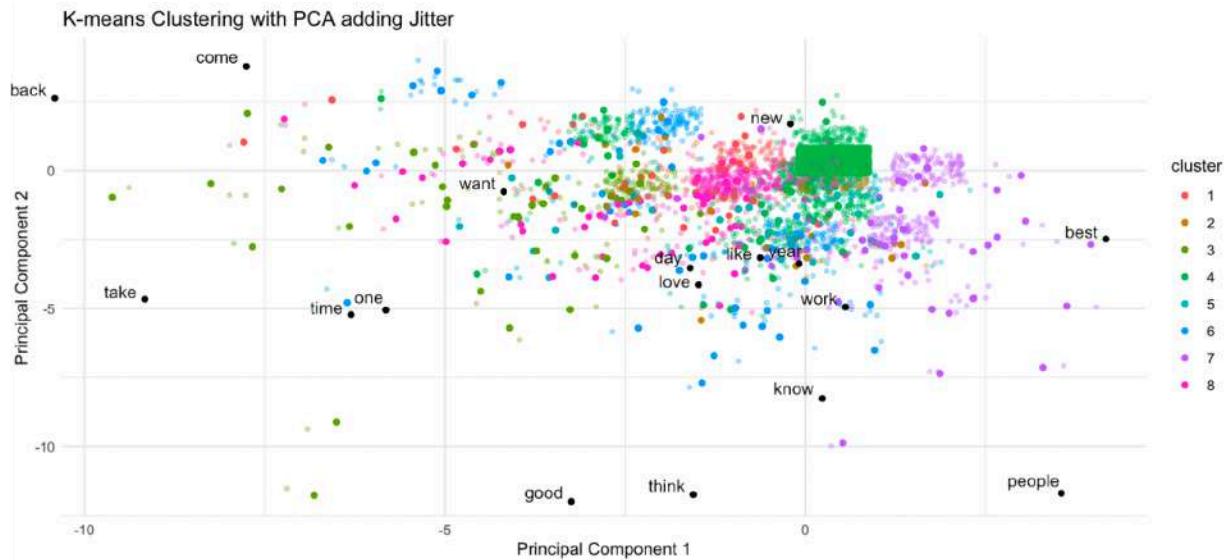


Now we can see many points and how they're condensed in the plot. We observe that the densest area is the green cluster near (0,0). These points belong to the 4th cluster that had the most points as we've discussed before.

To better understand each cluster, we also plotted the variable loadings, which allow us to see how each variable or term, in this case, interacts and affects each PCA dimension.



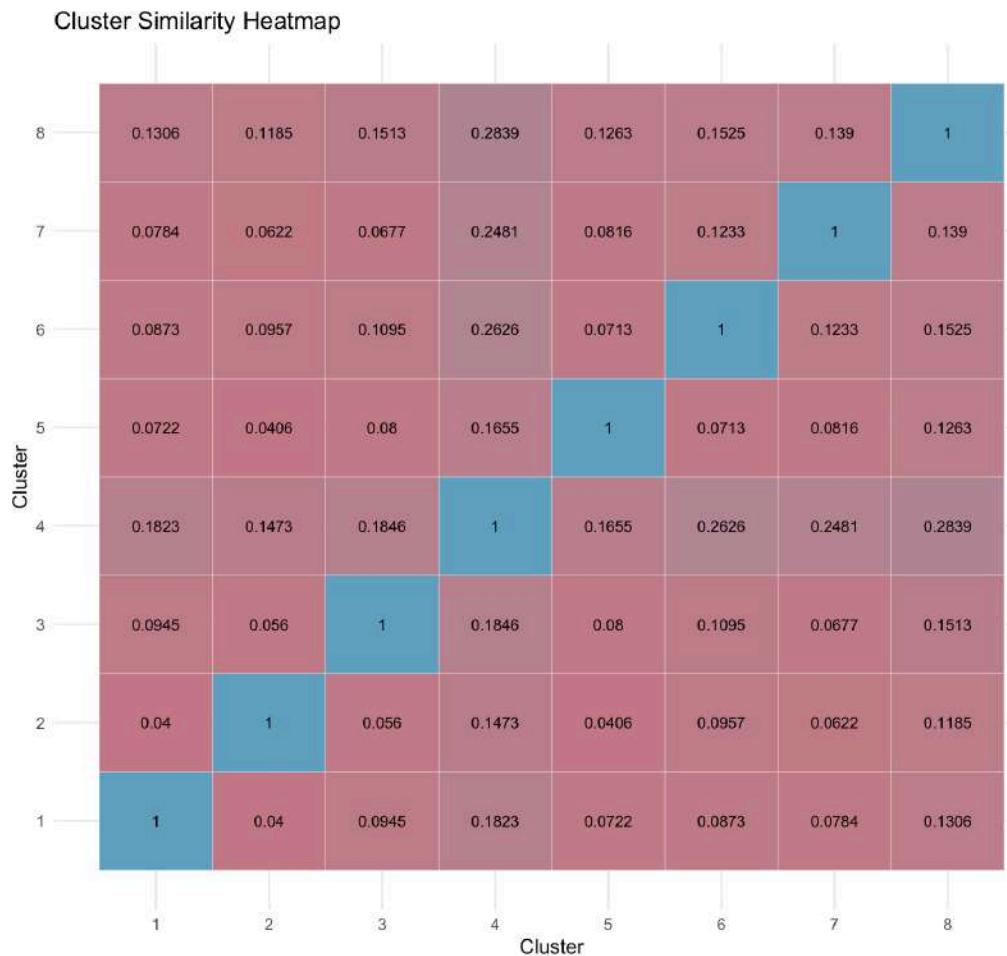
The plot below shows how the clusters are aligned with the variable loadings. The term *new* seems to be closest to cluster 4, *best* and *people* seems to be closest to cluster 7, but the other terms are harder to see the relation since the clusters are more dispersed around these areas.



Cluster Similarity Matrix

To further deepen our analysis, we decided we wanted to create something similar to a confusion matrix but to compare the clusters pairwise. The objective is to, in each cell provided by the cluster pairs, compute the Cosine Similarity between the union of the documents that belong to each cluster. This way, we can apply what we previously learned in LSA to compare the clusters.

As we said before, we joined all the documents by their assigned clusters into a vector of length $k=8$, these will be the documents we test their similarity on. The result is the following.



The result of this similarity matrix is as expected. The Cosine Similarity of one cluster compared with itself is 1, which is the maximum value this measure can take, however the value of the documents compared with the documents that are not itself are very low, ranging from 0.05 to 0.24 approximately. This makes sense, since we would hope that the clustering would separate the documents into clusters in a way that they would be different from each other across the other clusters. The aim is to obtain clusters that are not similar to each other, so we can group the documents appropriately.

One thing to observe is that in the line of Cluster 4, the lines that go across it are slightly more blue, this means that this cluster is more similar to the other clusters. This makes the most sense, since Cluster 4 is the largest cluster and it occupies a more centric area among the rest of the clusters.

To further rank the similarity among clusters we ordered them in descending order (excluding the reflexive similarity). As we can observe, our analysis is correct, since the top 5 Cosine Similarity measures are amongst pairs that contain the Cluster 4.

The most dissimilar pairs usually include the Cluster 2, which we don't see very well represented in our PCA plot.

Cluster1	Cluster2	Similarity
10	4	0.2839
12	4	0.2626
14	4	0.2481
16	3	0.1846
18	1	0.1823
20	4	0.1655
22	6	0.1525
24	3	0.1513
26	2	0.1473
28	7	0.1390
30	1	0.1306
32	5	0.1263
34	6	0.1233
36	2	0.1185
38	3	0.1095
40	2	0.0957
42	1	0.0945
44	1	0.0873
46	5	0.0816
48	3	0.0800
50	1	0.0784
52	1	0.0722
54	5	0.0713
56	3	0.0677
58	2	0.0622
60	2	0.0560
62	2	0.0406
64	1	0.0400

Cluster Word Clouds

To finalize this clusterings analysis we will present the Word Clouds associated with each cluster. We included the 10 most frequent words, since we considered the frequency of the rest to be too low to give a proper conclusion.

Cluster 1

Word Cloud - Cluster 1



For Cluster 1, these are the 10 most popular words, we can't distinguish any topics amongst them.

Cluster 2

Word Cloud - Cluster 2



The word cloud for Cluster 2 doesn't seem to have any major topics, but it seems to reminisce on the year the user has lived, possibly reminiscing about how good it was, how old they feel, how the next year will be.

Cluster 3

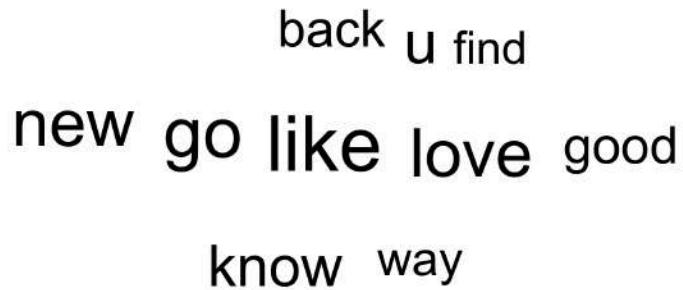
Word Cloud - Cluster 3



For Cluster 3, these are the 10 most popular words, we can't distinguish any topics amongst them. If we stretch it a little we may be able to think that maybe the user wants to turn back time and change things, but we can't say for sure.

Cluster 4

Word Cloud - Cluster 4



For Cluster 4, we see similar words as in the previous clusters. As we commented before, Cluster 4 is the most similar to the others, so this makes sense. The theme of this cluster seems to be about finding something the user loves or likes, but again it's hard to say since the words don't give a lot of information.

Cluster 5

Word Cloud - Cluster 5



For Cluster 5, the most mentioned term is *work*, and we see terms that are usually associated with it, such as *start*, *day*, *go* and *home*; “start work”, “work day”, “go to work”, “work from home” being possible examples.

Cluster 6

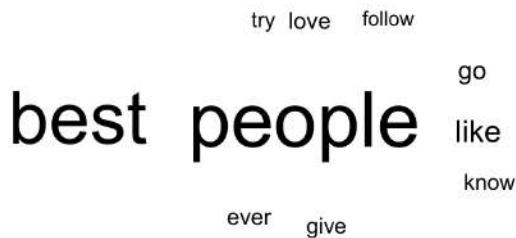
Word Cloud - Cluster 6



For Cluster 6, the main terms seem to be *think* and *come*. We can't think of any specific themes when we see this cluster.

Cluster 7

Word Cloud - Cluster 7



The main words for this cluster are *best* and *people*. The theme for this cluster seems to revolve around people that the user is fond of, loves or follows.

Cluster 8

Word Cloud - Cluster 8



Finally, for Cluster 8, the most frequent terms are *time*, *one* and *day*. We don't know exactly what this is referring to, but possibly it refers to new horizons.

Conclusions

The conclusions of this clustering aren't very solid, the word clouds for each cluster don't give us much information so it's not possible to be certain about the theme of these tweets solely by looking at this. The cause for this could be many reasons, maybe K-means was not the best choice, or maybe it's just our corpus which is not appropriate for this type of analysis.

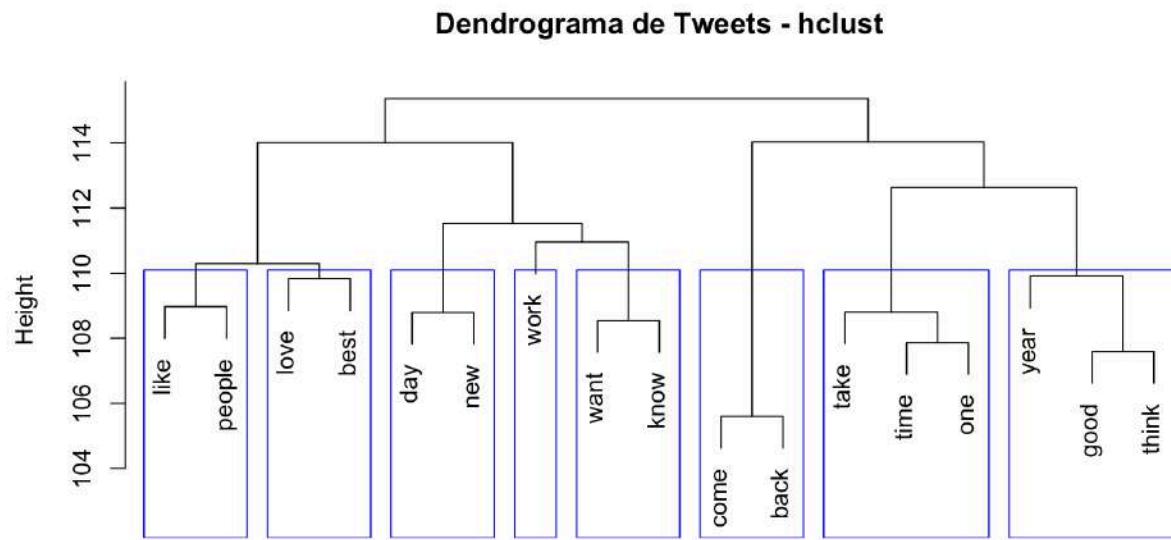
Hierarchical Clustering

There is another type of clustering commonly used for Documents Clustering. In this case, we tried to cluster the words based on the documents. This consists of transposing again our Term Document Matrix. The results of this differ from the ones we obtained with K-Means.

We see the following word associations:

- *like + people*: a likely theme of the tweets being about meeting people.
- *love + best*: a likely theme of the tweets being about loving the best of some category.
- *new + day*: a likely theme of the tweets being about starting a new day.
- *work*: this is self explanatory.
- *want + know*: a theme of curiosity towards a subject.
- *come + back*: a likely theme of the tweets being about someone or something coming back into trend or just into the users' life.
- *take + time + one*: for the theme of these tweets we're not quite sure what it could be.
- *year + good + think*: a likely theme of the tweets being about how the users think they had a good year.

As we have mentioned before, we're not sure about how reliable this conclusion is, based solely on this information. Especially in this case, since we aren't even taking a look into the tweets' contents.

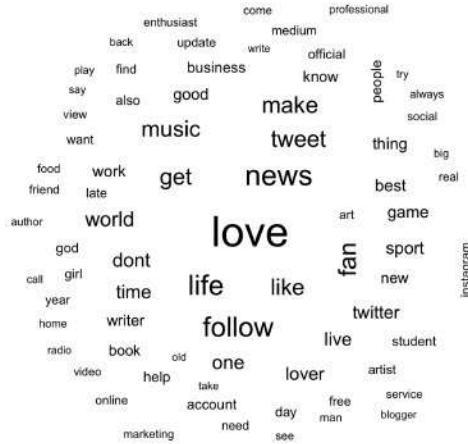


Clustering the User Description

This time we will be clustering the tweet's user description. The documents can be found under the variable *description* in our dataset.

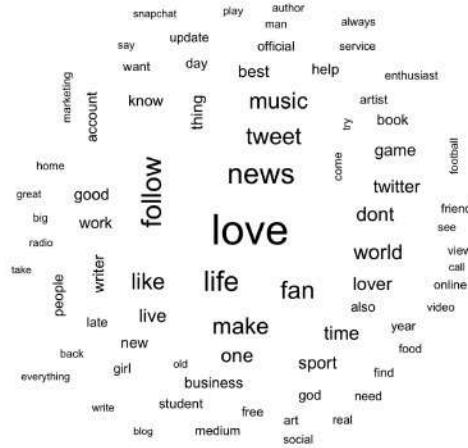
Word Cloud

Same as previously we will take an initial look into the Corpus' Word Cloud generated by our user descriptions after preprocessing them (removing stopwords, punctuation, duplicated entries, numbers and other symbols).



The most popular words this time are *love*, *life*, *news* and *music*, among others. But we detect some words that don't bring much meaning to the analysis, some examples being *get* and *instagram*, which are words that don't bring us a lot of information about the text.

To solve this we can remove these words manually by taking a look at the initial plot. After depurating a little more this is the result:



The results of this are primarily composed by:

- Words which seem to be related to **twitter**, such as *follow*, *tweet* or *twitter*.

- **Business vocabulary** like *business, service and marketing*.
- We see a mix of hobbies too, like *art, football, game...*

Term Document Matrix

The next step is to create a Term Document Matrix, this will transform our documents into numerical data that our algorithms can understand. The document generated by our data has this shape:

```
<<TermDocumentMatrix (terms: 12855, documents: 5161)>>
Non-/sparse entries: 40572/66304083
Sparsity           : 100%
Maximal term length: 130
Weighting          : term frequency (tf)
```

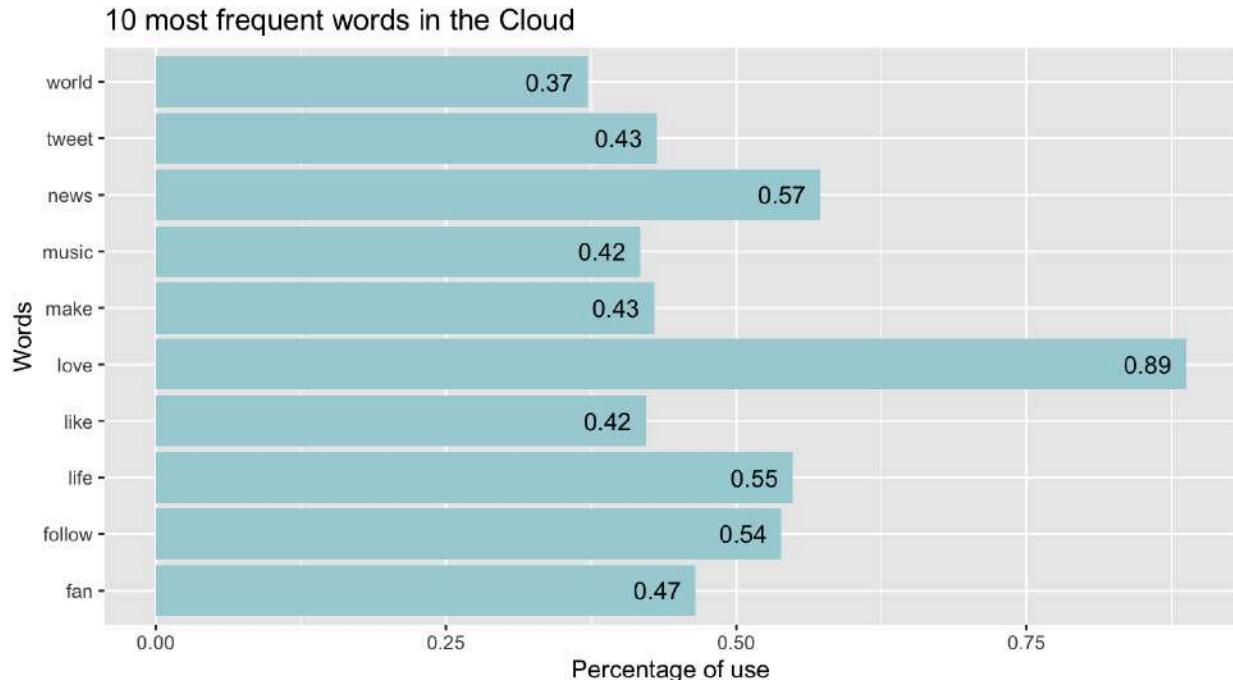
The dimension of the matrix is 12859×5161 , indicating that our corpus has 12859 different words, which are represented by the columns, and 5161 documents, which are represented by the columns of the matrix. Each cell located at ij will indicate whether the word i is present in the document represented by the column j , and how many times it is present in it, basically, the cell contains the number of times the word appears in the tweet.

With the Term Document Matrix we can easily calculate the frequency of each word and find the top words. Here's our database's top 20:

palabra freq			palabra freq		
love	love	372	dont	dont	144
news	news	240	one	one	143
life	life	230	time	time	136
follow	follow	226	live	live	133
fan	fan	195	lover	lover	129
tweet	tweet	181	twitter	twitter	125
make	make	180	sport	sport	122
like	like	177	game	game	115
music	music	175	best	best	114
world	world	156	thing	thing	110

We will visualize the frequency of the top 10 words. Each word will have its percentage of use associated with it in relation to the total amount of words in the tweets of the whole dataset. We

can see that the most popular word in the descriptions by difference is *love*, followed by the words *news* and *life*.



Determining the Number of Clusters

For the Document Clustering we will be using the K-Means clustering which is a popular method used in this area. Before we proceed, if we recall the shape of the Term Document Matrix, we have too many terms, which will slow down the performance of the algorithm, aside from that the sparsity of the matrix is 100%, this means most words rarely appear in tweets, so the matrix has a lot of zeros.

To solve this problem we can remove sparse terms, by setting a *sparsity limit* which will not allow terms of higher sparsity to be in the matrix. We have set this limit to 98%, usually one would set a lower Sparsity, but we couldn't set the limit any lower or we would end up with little to no terms. We are left with 21 terms and the sparsity of this matrix is now reduced to 97%.

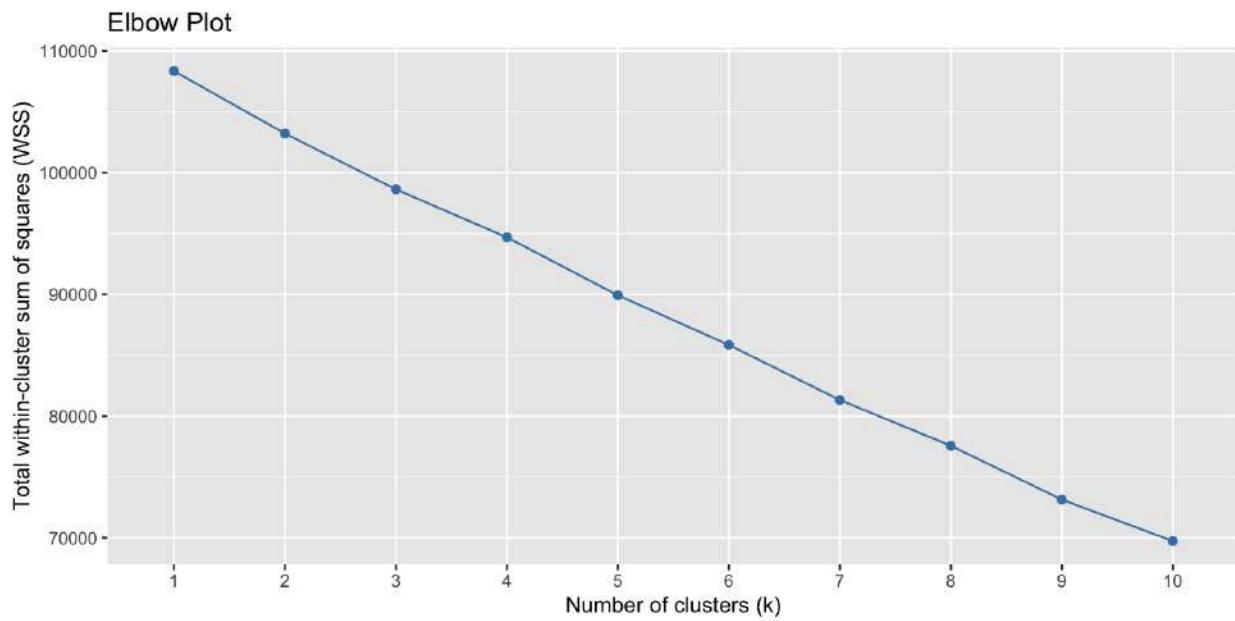
```
<<TermDocumentMatrix (terms: 21, documents: 5161)>>
Non-/sparse entries: 3266/105115
Sparsity : 97%
Maximal term length: 7
```

Weighting : term frequency (tf)

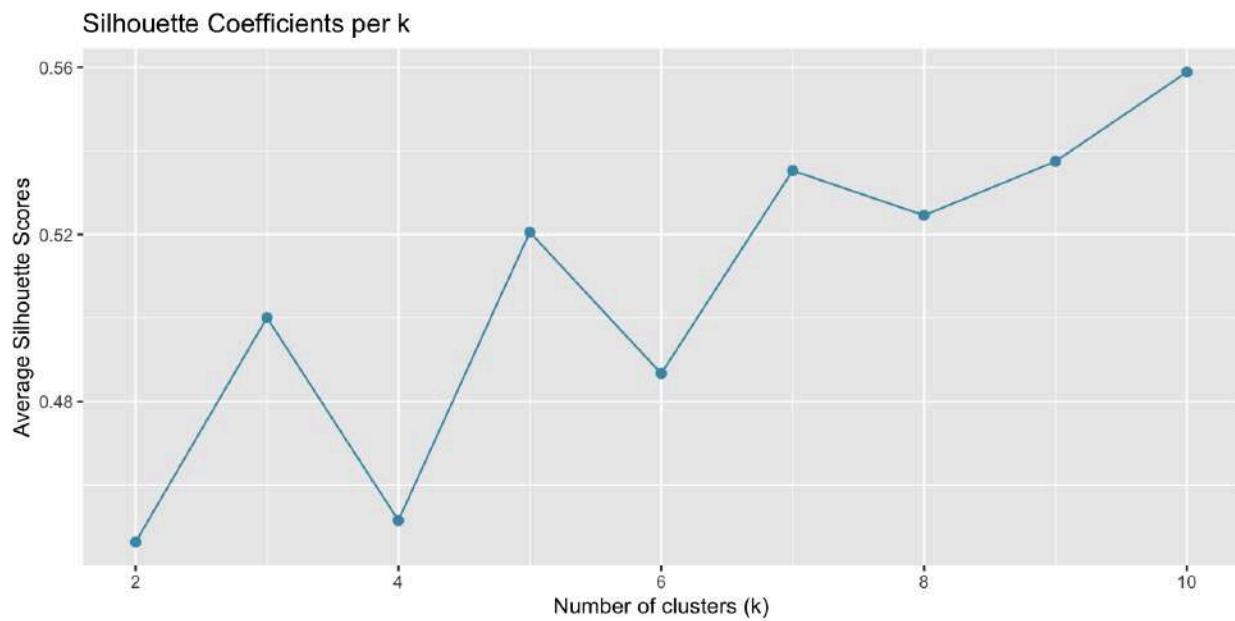
Same as before, we will transpose the matrix since we will be clustering the documents (now situated in the rows) by the amount of each term they contain (now in the columns, and can be interpreted as variables). Following this step, we will normalize the matrix to improve the clustering's performance. We used standardization or Z-score normalization.

Again, before we create the clusters, we need to determine the number of clusters we need to create. To do this, first we decided to conduct an Elbow Plot. To create an elbow plot we need to set a range of values k we want to try, k being the number of clusters the K-Means will create. Then we will run K-Means for each k in the range we defined. For each one of these clusters we will calculate the Within-Cluster Sum of Squares (WCSS), which is a measure we want to reduce as explained in the previous clustering.

The plot our data generated doesn't follow this shape, it looks more like a straight line, and we can't spot an elbow in it.



To further explore the optimal k , we tried to plot the Average Silhouette Scores for each cluster, once again. The conclusion we reached with the plot below and the plot above is that our data seems to want to be grouped into many clusters. We hypothesize that these plots didn't come out right due to the sparseness of the matrix. However we chose $k=7$, due to the high increase in the Silhouette Score at that point.



Once we have applied the K-Means clustering to our document term matrix, these are the results. The number of data points per cluster is the one below, the row above indicates the number of the clusters, and the value below indicates the number of documents in each cluster.

Clustering Results

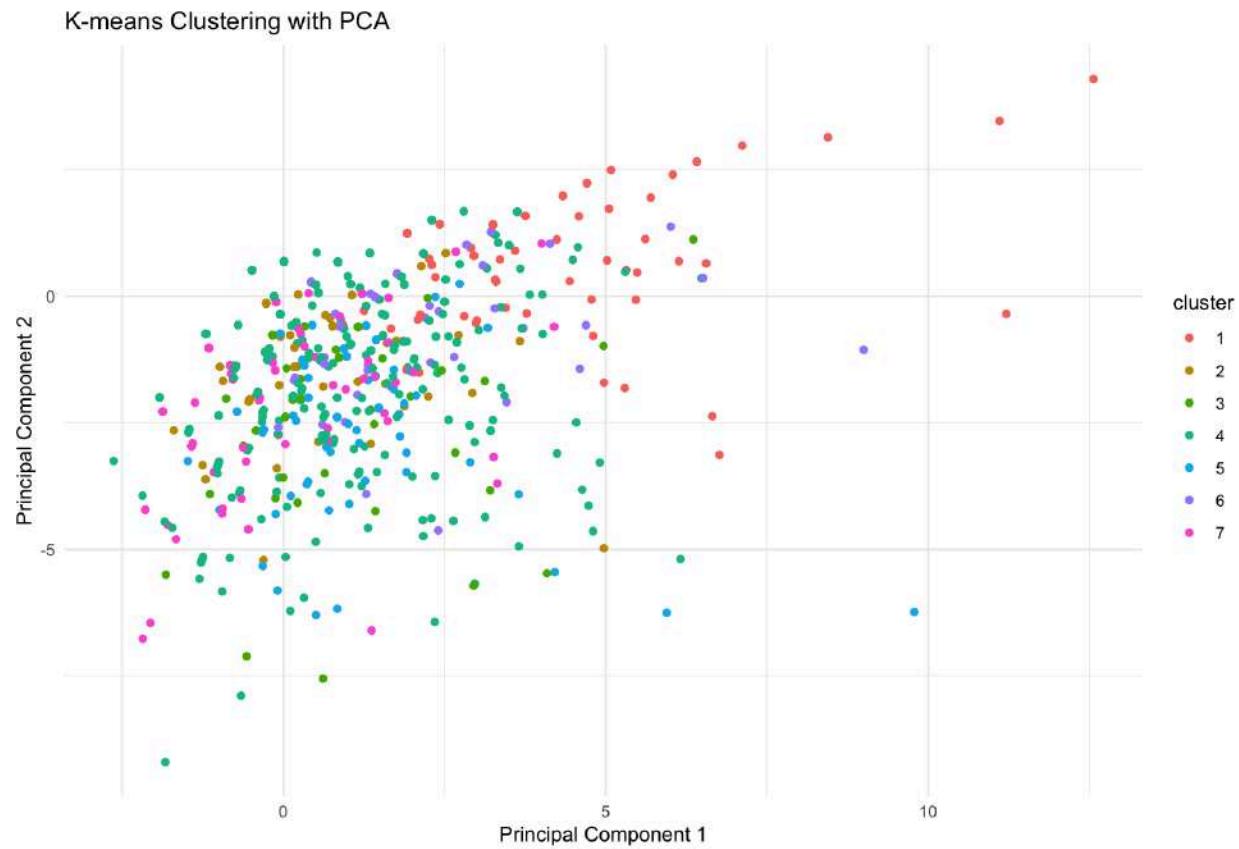
Once we have applied the K-Means clustering to our document term matrix, these are the results. The number of data points per cluster is the one below, the row above indicates the number of the clusters, and the value below indicates the number of documents in each cluster.

```
> summary(pca_scores$cluster)
  1   2   3   4   5   6   7
204 102 108 4242 153 117 235
```

As we can observe, the clusters aren't equally distributed, most of them being within a similar range, except for the 4th cluster, which contains most of the points of our matrix. This may, again, be due to the fact that we have very few terms that are sparse.

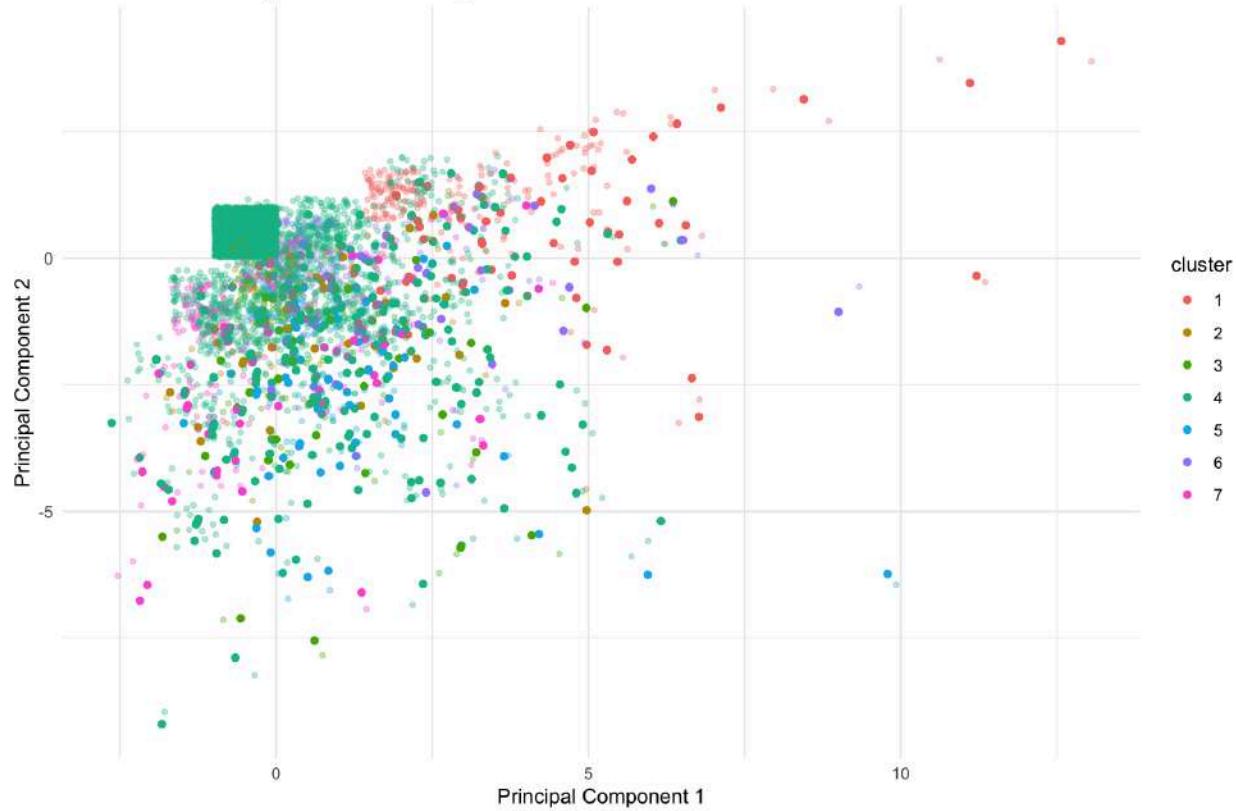
Visualizing with PCA

But same as before, we visualized the clustering with PCA, we created the dimentions using the Document Term Matrix. The result is below. We observe the same pattern as in the previous clustering, it doesn't look like there are over 6000 points in the plot, this is because the points are stacked on top of each other, due to the sparsity and scarcity of terms.



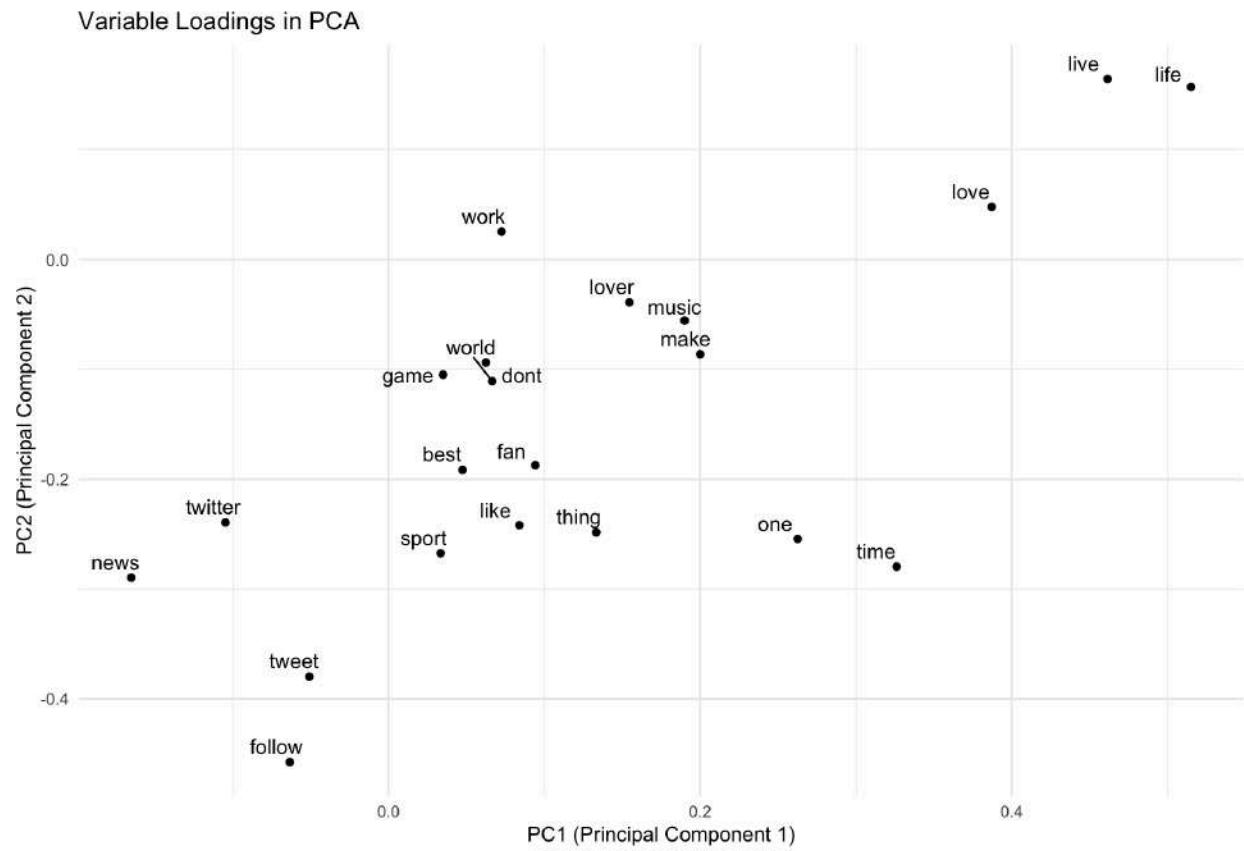
When we add “jitter” we allow the visualization of the number of points that are stacked by scattering them. This is the result of adding jitter:

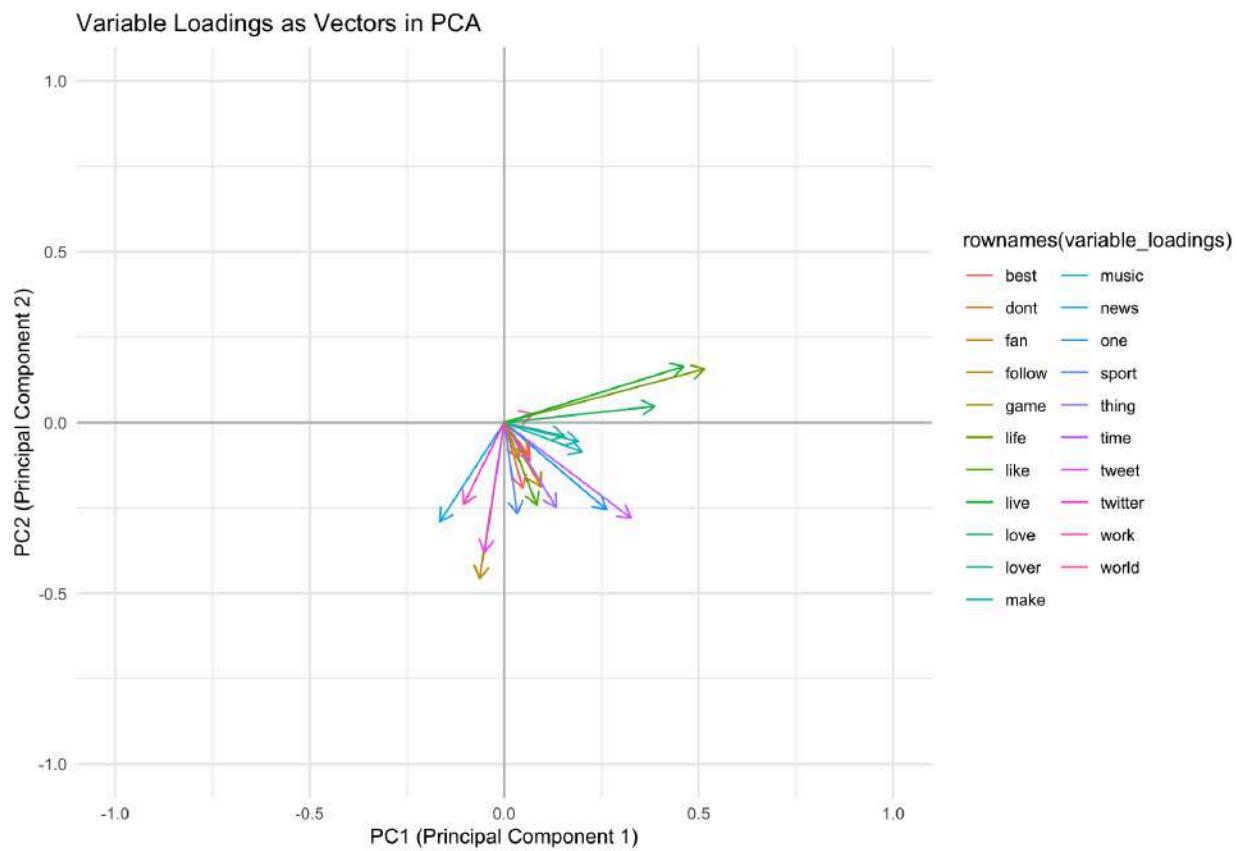
K-means Clustering with PCA adding Jitter



Now we can see many points and how they're condensed in the plot. We observe that the densest area is the dark green cluster near (0,0). These points belong to the 4th cluster that had the most points as we've discussed before.

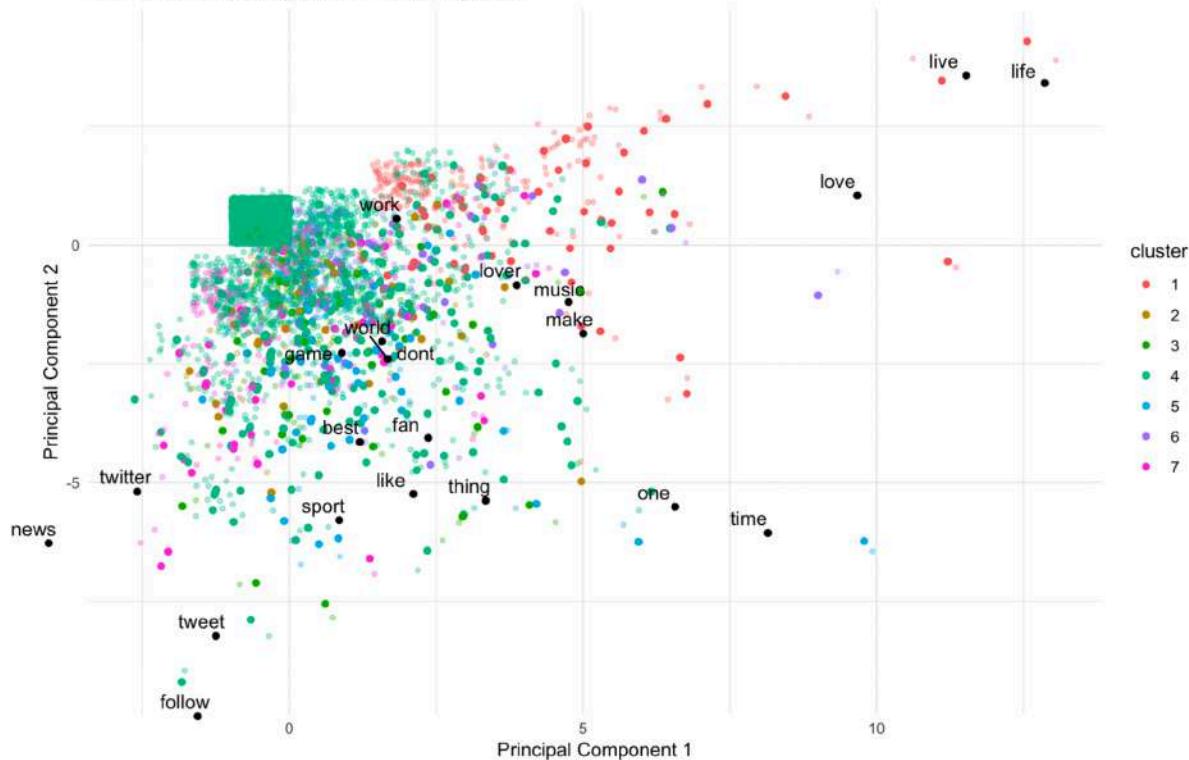
To better understand each cluster, we also plotted the variable loadings, which allow us to see how each variable or term, in this case, interacts and affects each PCA dimension.





The plot below shows how the clusters are aligned with the variable loadings. The term *work* seems to be closest to cluster 4, *live* and *life* seem to be closest to cluster 1, but the other terms are harder to see the relation since the clusters are more dispersed around these areas.

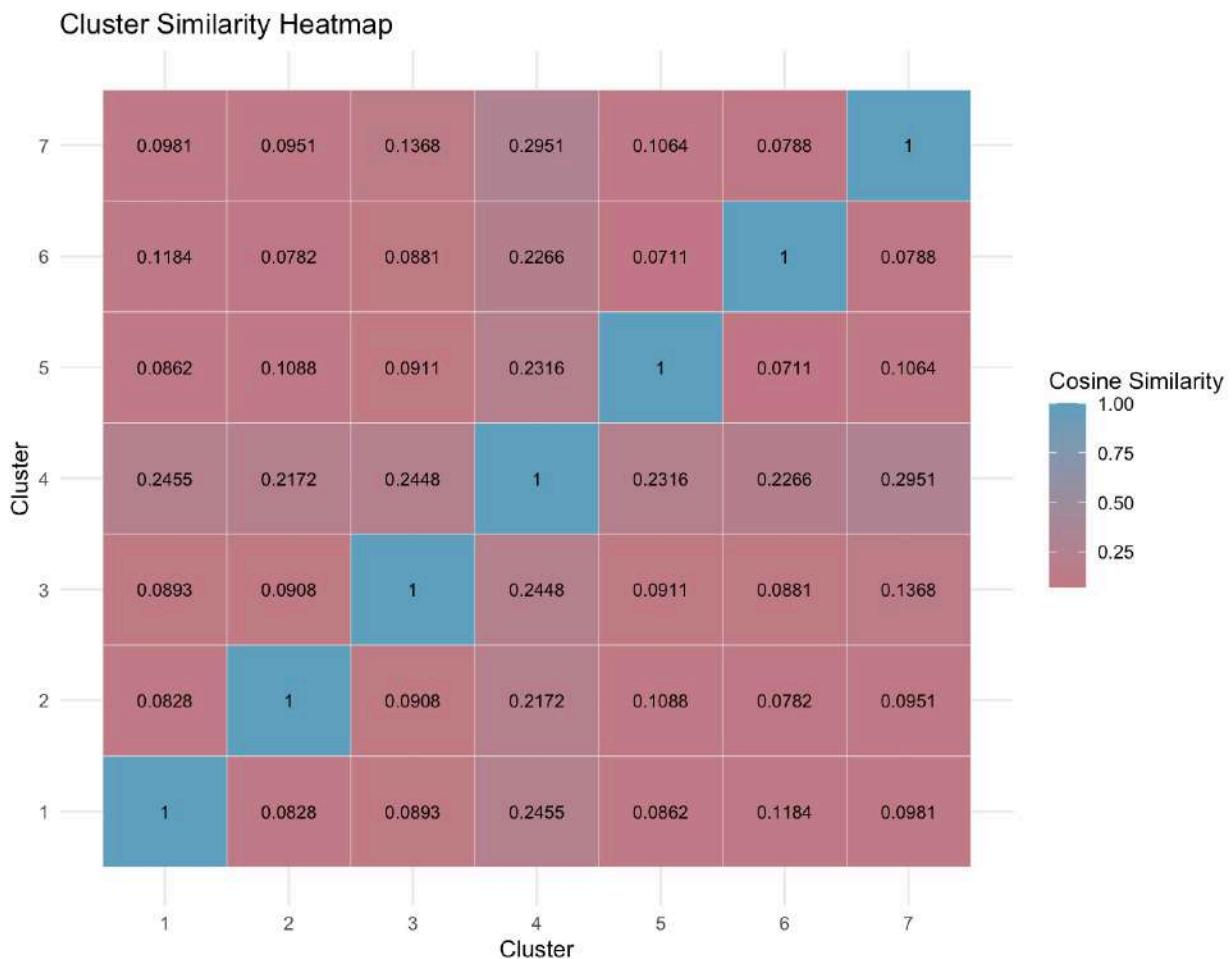
K-means Clustering with PCA adding Jitter



Cluster Similarity Matrix

Like in the previous section, we decided we wanted to create something similar to a confusion matrix but to compare the clusters pairwise. The objective is to, in each cell provided by the cluster pairs, compute the Cosine Similarity between the union of the documents that belong to each cluster. This way, we can apply what we previously learned in LSA to compare the clusters.

As we said before, we joined all the documents by their assigned clusters into a vector of length $k=7$, these will be the documents we test their similarity on. The result is the following.



The result of this similarity matrix is as expected. The Cosine Similarity of one cluster compared with itself is 1, which is the maximum value this measure can take, however the value of the documents compared with the documents that are not itself are very low, ranging from 0.07 to 0.29 approximately. This makes sense, since we would hope that the clustering would separate the documents into clusters in a way that they would be different from each other across the other clusters. The aim is to obtain clusters that are not similar to each other, so we can group the documents appropriately.

One thing to observe is that in the line of Cluster 4, the lines that go across it are slightly more blue, this means that this cluster is more similar to the other clusters. This makes the most sense, since Cluster 4 is the largest cluster and it occupies a more centric area among the rest of the clusters.

To further rank the similarity among clusters we ordered them in descending order (excluding the reflexive similarity). As we can observe, our analysis is correct, since the top 5 Cosine Similarity measures are amongst pairs that contain the Cluster 4.

The most dissimilar pairs usually include the Cluster 6, which we don't see very well represented in our PCA plot.

	Cluster1	Cluster2	Similarity
9	4	7	0.2951
11	1	4	0.2455
13	3	4	0.2448
15	4	5	0.2316
17	4	6	0.2266
19	2	4	0.2172
21	3	7	0.1368
23	1	6	0.1184
25	2	5	0.1088
27	5	7	0.1064
29	1	7	0.0981
31	2	7	0.0951
33	3	5	0.0911
35	2	3	0.0908
37	1	3	0.0893
39	3	6	0.0881
41	1	5	0.0862
43	1	2	0.0828
45	6	7	0.0788
47	2	6	0.0782
49	5	6	0.0711

Cluster Word Clouds

To finalize this clusterings analysis we will present the Word Clouds associated with each cluster. We included the 10 most frequent words, since we considered the frequency of the rest to be too low to give a proper conclusion.

Cluster 1

Word Cloud - Cluster 1



For Cluster 1, we were excited to see words that made sense together. The topic we can extract is that the users in this cluster are probably very optimistic, with a *live*, *laugh*, *love* kind of view of life. Life is about enjoyment, loving but also working.

Cluster 2

Word Cloud - Cluster 2



For Cluster 2, we also were able to quickly find the relation between the top words. the topic we can extract is that the users in this cluster are homebodies: they like to stay home, playing video games, movies or listening to music.

Cluster 3

Word Cloud - Cluster 3



In this cluster, we had a hard time trying to deduce a topic, we decided it's most likely related to users that are showing love to their best friends on their description, aside from being people who ask for followers and like being online.

Cluster 4

Word Cloud - Cluster 4



For this cluster, we think that the topic of these user's descriptions are related to being a fan and loving music, so likely they're displaying their affection as fans towards certain artists or bands. Another curious topic in here is *news*, maybe the description is mentioning news tweets.

Cluster 5

Word Cloud - Cluster 5



Cluster 5 has quite similar terms to Cluster 4, excluding *news* (and *tweet* and *u*) all the words from Cluster 4 are here. So likely this cluster represents a similar topic to Cluster 5, differing in the news part.

Cluster 6

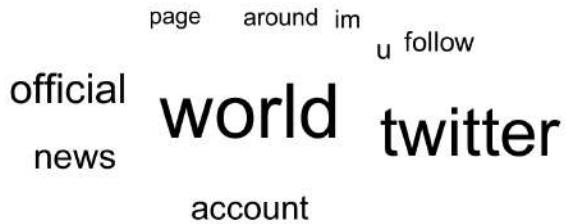
Word Cloud - Cluster 6



Cluster 6 is a fun topic, it likely represents users who are cat or animal lovers, who like coffee, food and music.

Cluster 7

Word Cloud - Cluster 7



Cluster 7 is similar to Cluster 4 again, but in a different way than Cluster 5. This cluster has more mentions of news, probably world news. And it seems to be composed of users claiming to be official pages. In contrast to Cluster 4 and 5 there is no mention of music.

Conclusions

The conclusions of this clustering aren't 100% certain. However, in this section we were able to extract what seem to be more believable topics from each cluster. We know these tropes of users: the cat lover, the news agency, the person that loves video games... This made the analysis all the more interesting to us.

One thing we observed is the similarity seen between the word clouds of Cluster 4, 5 and 7. It makes sense that they would be the most similar since Cluster 4 is the largest cluster that is also in a quite centric position in the PCA plot we saw. We also see that the Cosine similarities between clusters 4 and 7 and 4 and 5 are among the highest ones in our similarity ranking. However the cosine similarity of 5 and 7 is not as high as the previously mentioned pairs, this checks out, as we hypothesized that each one of these clusters was similar to a section of the Cluster 4.

Hierarchical Clustering

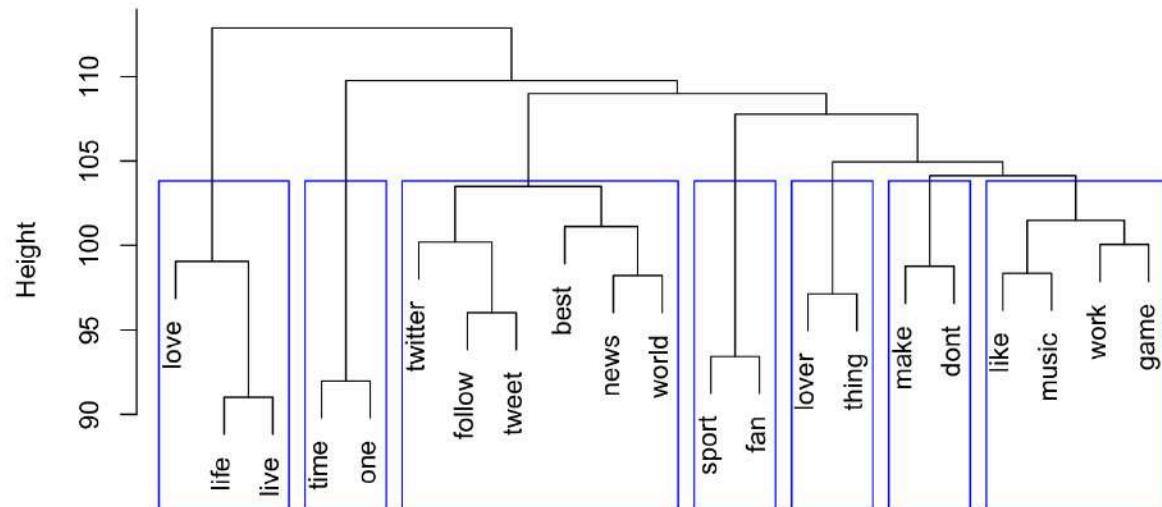
Finally we did the same Hierarchical Clustering as before. In this case, we tried to cluster the words based on the documents. This consists of transposing again our Term Document Matrix. The results of this differ from the ones we obtained with K-Means.

We see the following word associations:

- *love + live + life*: a likely theme of the descriptions being about living life to the fullest.
- *one + time*: a likely theme of the descriptions being about something being a one time thing.
- *twitter + follow + tweet*: a likely theme of the descriptions being about twitter users advertising their accounts. Additionally, it is joined with *best + news + word*, a theme probably indicating a news platform.
- *sports + fan*: we didn't observe this theme in our K-Means clustering so it is interesting to see how the different methods may affect the outcome.
- *lover + thing*: for the theme of these descriptions we're not quite sure what it could be. But it is related to people loving an array of things.
- *make + dont*: for the theme of these descriptions we're not quite sure what it could be.
- *like + music + work + game*: for the theme of these descriptions we think they're about people liking music, explaining what work they do or what games they enjoy.

As we have mentioned before, we're not sure about how reliable this conclusion is, based solely on this information. Especially in this case, since we aren't even taking a look into the tweets' contents.

Dendrograma de Description - hclust



GEOSPATIAL

In this "Geospatial" section, we will embark on an intensive and detailed study of geographical trends within the Twitter database, using the Type II - Point Processes approach for this purpose. This analysis will allow us to uncover, interpret, and visualize the spatial relationships and interactions that exist among Twitter users. By doing this, we will be in a privileged position to identify significant and emerging patterns that may not be immediately evident at first glance.

Within this framework, we will focus on key aspects of geospatial analysis such as geographical representation of data, spatial-temporal data analysis, and geographic predominance of gender by zone.

Geographical representation involves a detailed visualization of the location and distribution of Twitter users. Here, we will use visualization techniques like heatmaps and point maps to clearly display the geographical distribution of users based on their gender.

The spatial-temporal data analysis will focus on assessing how the distribution of Twitter users has changed over time. This analysis will enable us to identify any emerging trends and better understand the dynamics at play.

Lastly, we will tackle geographic predominance of gender by zone. This part of the analysis will allow for a better understanding of differences among locations in terms of the gender distribution of Twitter users.

In summary, this geospatial section is intended to shed light on the complex and multifaceted geographical patterns that emerge from Twitter data, providing a deeper and more nuanced understanding of user interactions and behaviors across different geographical locations.

Data

To carry out our geospatial analysis, we will use the dataset "db_outliers_missings.csv". This dataset is the result of a previous cleaning and preprocessing process in which outliers and missing values from the original dataset were addressed.

Within this dataset, we will focus on three specific attributes: 'lon' (longitude), 'lat' (latitude), and 'created' (profile creation date). The longitude and latitude represent the geographical coordinates where the user's profile was created, which will enable us to perform the geospatial

analysis. As for the 'created' attribute, we will only consider the year of profile creation, as our interest lies in analyzing the temporal trends in the data.

In addition, to facilitate the analysis of gender distribution, we will divide our dataset into two new sets: "Male" and "Female". These datasets will represent users who have the "male" and "female" label respectively in the "gender" attribute. This division will allow us to more effectively analyze and compare geographical and temporal trends among Twitter users of both genders.

In summary, the correct selection and preparation of data is a critical step in our analysis. By ensuring that our data is clean and relevant to our research questions, we are in the best possible position to derive valid and meaningful conclusions.

Geographical Representation

In this section, we will focus on the geographical representation of our data, with special emphasis on the geospatial distribution of our Twitter users without distinguishing gender. This study will allow us to analyze patterns of clustering and geographical dispersions. Although the analysis is performed with the data divided into two sets, "Male" and "Female", we do not intend in this section to perform an analysis of the difference in gender concentration by region. That aspect will be addressed later in the "Geographic Predominance of Gender by Zone" section.

A visually powerful method to carry out this task is through a "Heatmap of User Gender". Heatmaps are useful tools that allow us to represent multivariate data in a way that is easy to interpret and visually appealing. Each point on the map represents the geographical location of a user, and the color of the point indicates the density of users in that region, providing a visual representation of the "heat" or activity in that area.

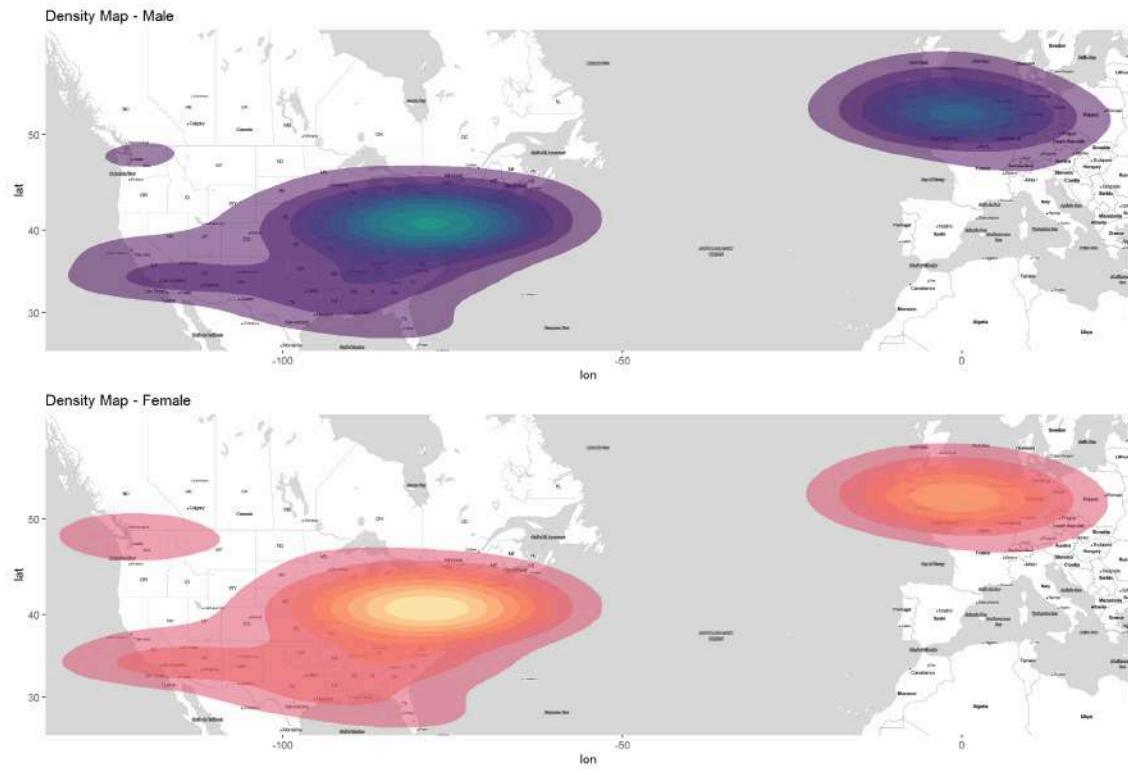
In our case, the "Heatmap of User Gender" will offer a global vision of the behavior of our data, allowing us to quickly and efficiently identify regions with a higher concentration of Twitter users.

In addition, as part of our geographical representation, we will implement additional analysis using a "Point Map Displaying User Locations". This type of representation will allow us to visualize the specific location of each user, providing an additional level of detail that will complement the general view provided by the heatmap.

By combining these two visualization techniques, we will obtain a detailed and complete view of the geographical distribution of our users, which will give us a solid foundation for subsequent analyses. This dual representation will better understand the density and location of users based on the location where they created their profiles.

Heatmap of User Gender

After carrying out the visualization using the "Heatmap of User Gender" for the "Male" and "Female" datasets, we have observed some notable geographical trends in the distribution of our data. Specifically, there is a predominant concentration of users in two particular regions: North America and Northern Europe.



In North America, the highest concentration of users is located in the states of Virginia, Maryland, and New Jersey, in the United States. On the other hand, in Northern Europe, the highest concentration of users is located in the United Kingdom, particularly in London.

This finding has significant implications for our analysis. Due to the geographical concentration of our data in these two regions, we have decided to split our analysis, treating North America and Northern Europe as distinct regions in the following subsections.

This decision to break down our regions of interest is based on the premise that conducting more specific and detailed analyses of each region will allow us to extract more significant and granular patterns and trends, which could be lost in a high-level analysis that groups all locations.

In addition, this detailed analysis allows us to infer that our data primarily represents Twitter users from these two regions. It's important to note that any analysis and conclusions derived from our data will be primarily centered on these two geographical groups. Thus, the characteristics, trends, and patterns we identify may be more representative of these users in North America and Northern Europe.

[Point Map Displaying User Locations](#)

Continuing with our analysis within the Geographical Representation section, we will now focus on a more granular and detailed approach with the use of a "Point Map". In this section, we strive to provide a more accurate and detailed representation of user locations in each region, to better understand the geographical trends and distributions of our users.

The "Point Map" will display the locations where Twitter users created their profiles, providing a more detailed view than a simple count of users in each region. Each point on the map will represent the approximate geographical location of a particular user.

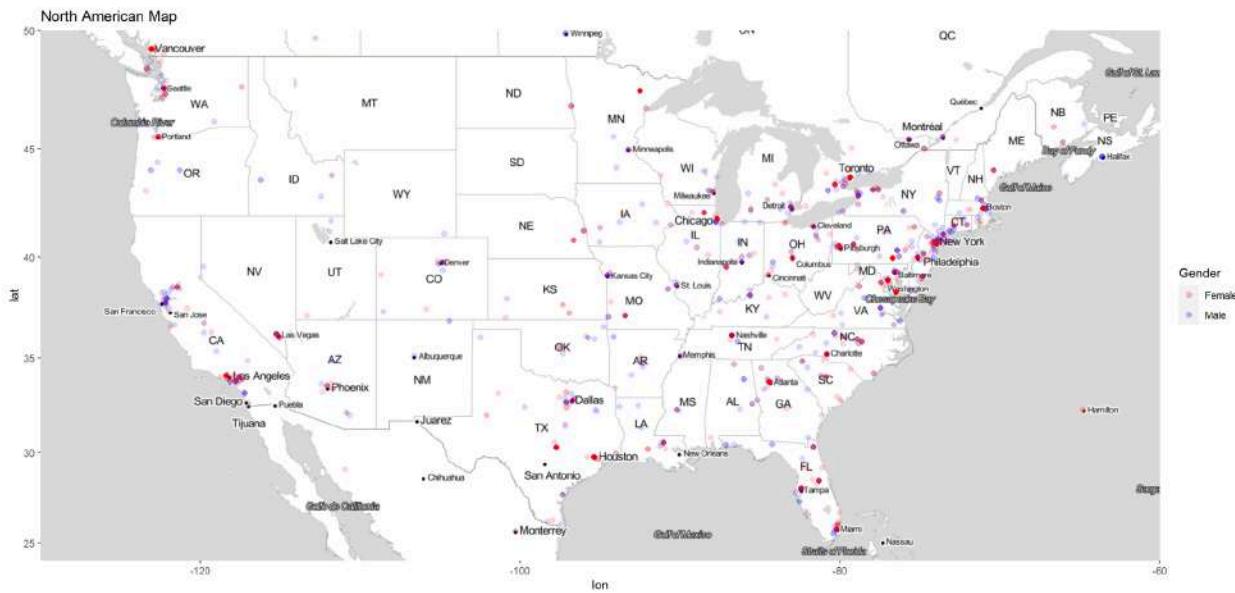
It's important to clarify that, due to privacy considerations and in an effort to protect user identities, each point will be represented larger on the map. This decision ensures that the exact location of any individual user cannot be identified, yet still provides us with a useful and effective representation of the overall distribution of our users.

Moreover, we want to emphasize that, despite this more detailed representation of user locations, we have not collected or present any additional information that could be used to identify a specific user. Our analysis is strictly limited to the geographical and temporal attributes of the data, and we maintain a strong commitment to protecting the privacy and identity of all users represented in our dataset.

With this more specific and detailed view of user locations, we can progress in our analysis and delve deeper into the trends and characteristics of Twitter users in our regions of interest.

North America Distribution

After implementing the "Point Map Displaying User Locations", we were able to visualize the distribution of users in North America in a much clearer and specific way.



In the analysis of the distribution in North America, a clear concentration of users in state capitals and major cities was observed. The presence of users in cities such as San Francisco, New York, and Philadelphia stood out especially.

This pattern suggests that users from these metropolitan areas are particularly active on Twitter, whether due to cultural, demographic, economic, or other factors. However, it is important to note that this distribution does not necessarily represent the entirety of Twitter users in North America, but rather the areas of highest activity within our database.

The distribution in North America presents an interesting pattern and provides us with valuable perspective on the geographical trends of our users.

North Europe Distribution

Continuing with our "Point Map Displaying User Locations" analysis, we turned our attention to the distribution of users in Northern Europe. As anticipated from the Heatmap analysis, the United Kingdom stands out as the main point of activity in the region.



We noticed a significant density of users in the capital cities of the United Kingdom, such as London, Manchester, and Liverpool. However, the concentration in these cities is not as dramatic as observed in the United States. In fact, a somewhat more dispersed distribution is observed compared to North America.

In addition, it is relevant to highlight the notable presence of users in Paris. Despite not being within the United Kingdom, this city demonstrated considerable user activity.

In general terms, it is evident that there are more users in North America than in Northern Europe, which we could also infer from the Heatmap analysis. This pattern of distribution offers a fascinating insight into the geography of our user base and allows us to explore regional trends and correlations in future analyses.

Analysis of Spatial-Temporal Data

In this section, we turn our attention to conducting a spatio-temporal analysis of our data. The purpose of this analysis is to observe the growth of the social network in the regions of North America and Northern Europe over time, providing additional context and a deeper layer of understanding to our data.

Spatial and temporal data are often combined to show how geographic features and phenomena change over time. In our case, this combination will be useful to understand the expansion and growth of Twitter as a social network in these specific regions.

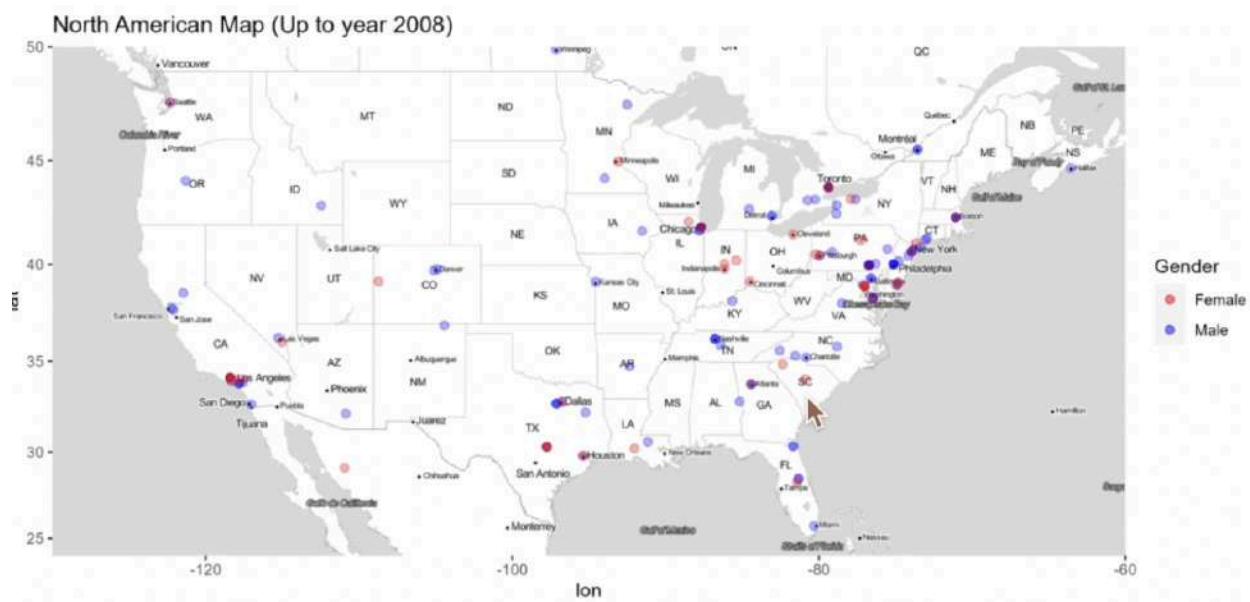
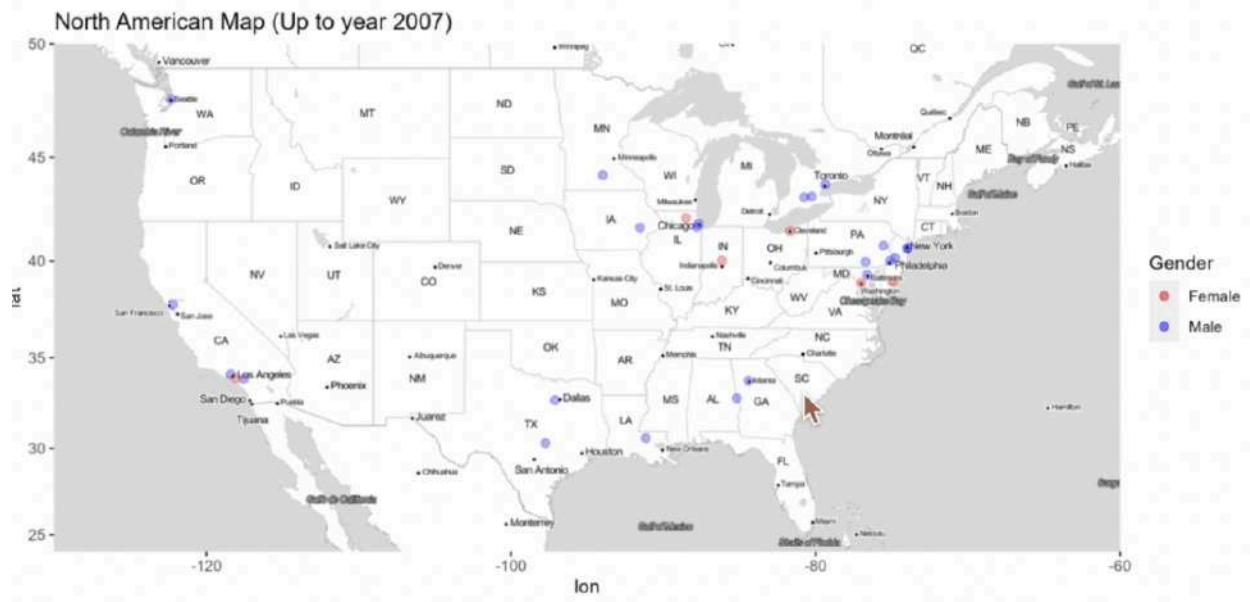
We will use the "Male" and "Female" dataset, which contains information from Twitter user profiles created between the years 2006 and 2015. Our goal is to observe the increase in the number of users each year during this period, which will allow us to obtain a visual representation of how the social network has grown and developed in these two specific regions.

In the following subsections, we will perform this user growth analysis for North America and Northern Europe separately, which will allow us to compare and contrast growth trends and patterns in these two regions.

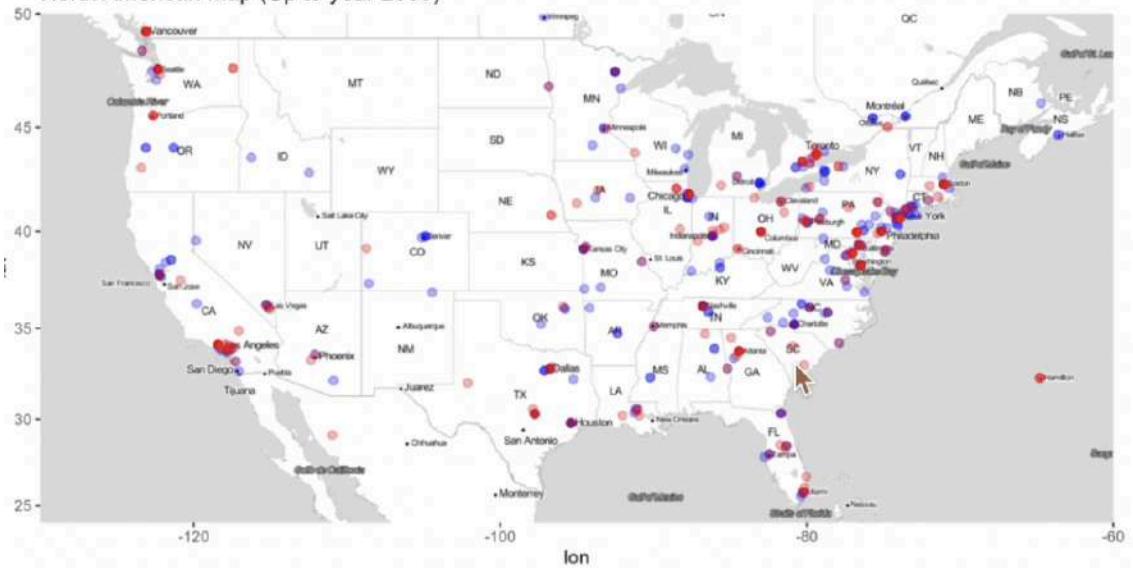
This spatio-temporal analysis will allow us to understand not only where our users are located but also when they joined the platform, adding an additional dimension to our understanding of user behavior in these regions.

User Growth in North America (2006-2015)





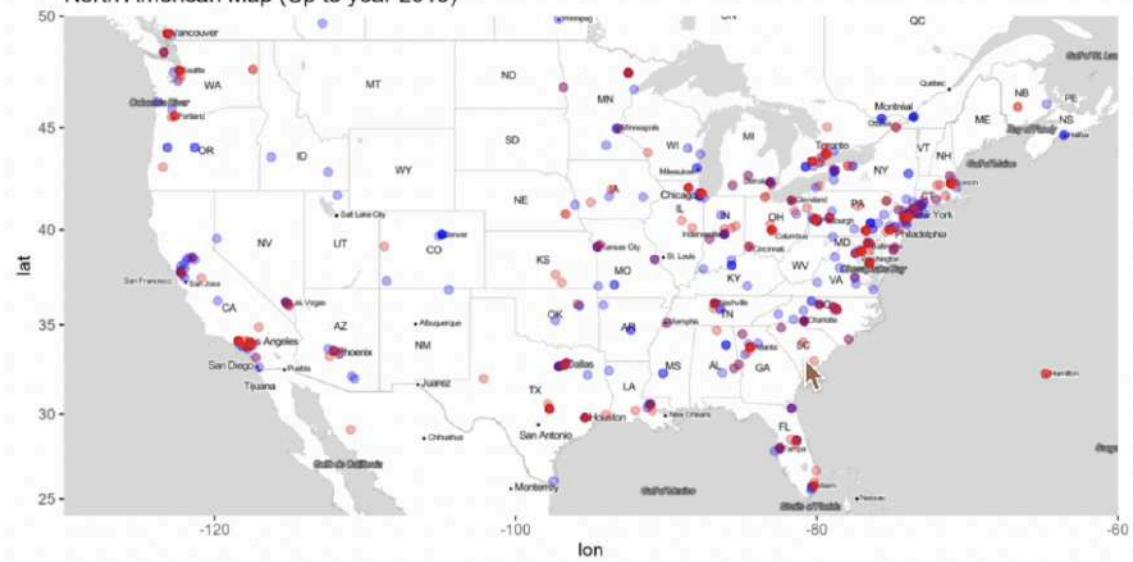
North American Map (Up to year 2009)



Gender

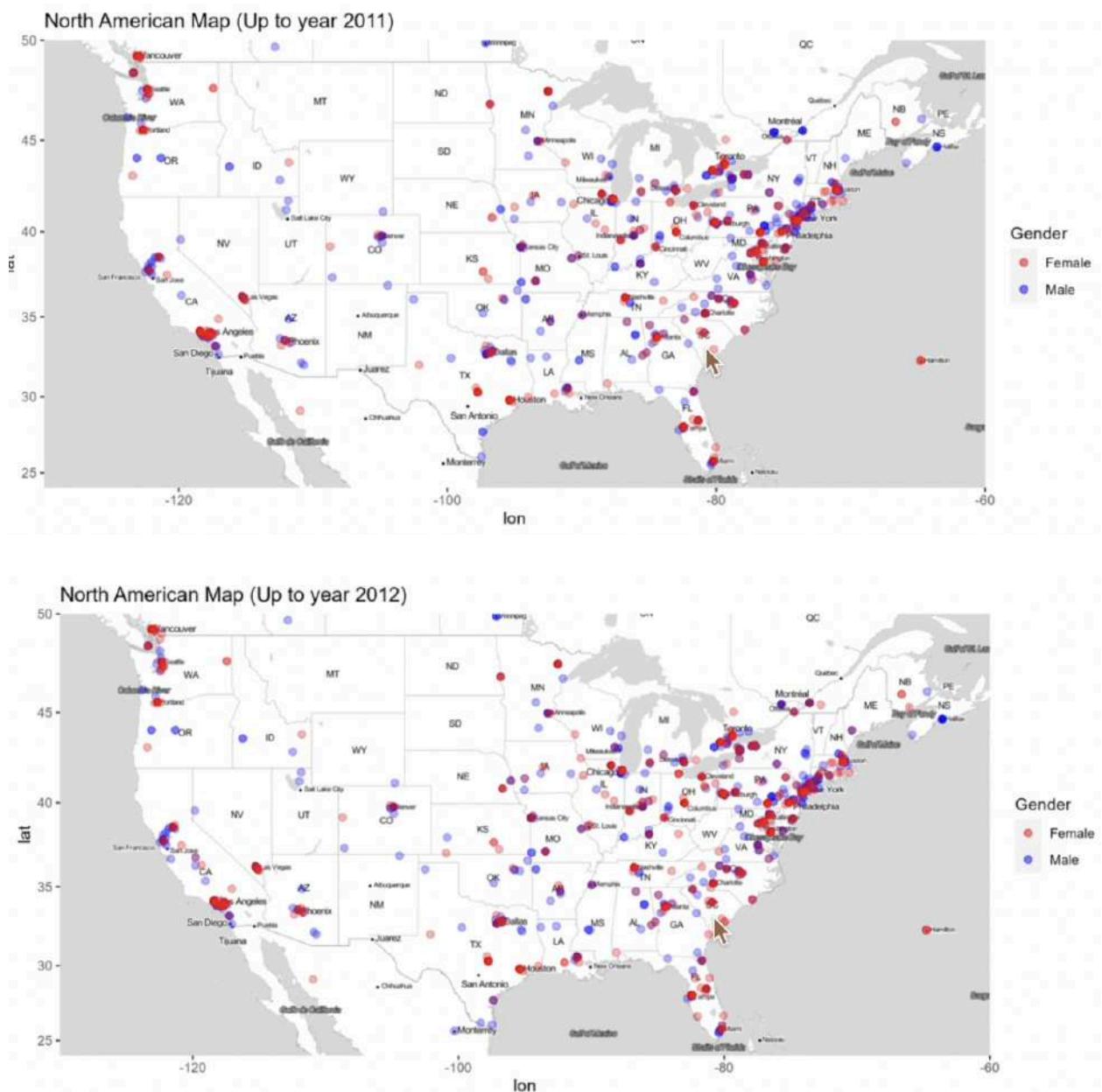
- Female
- Male

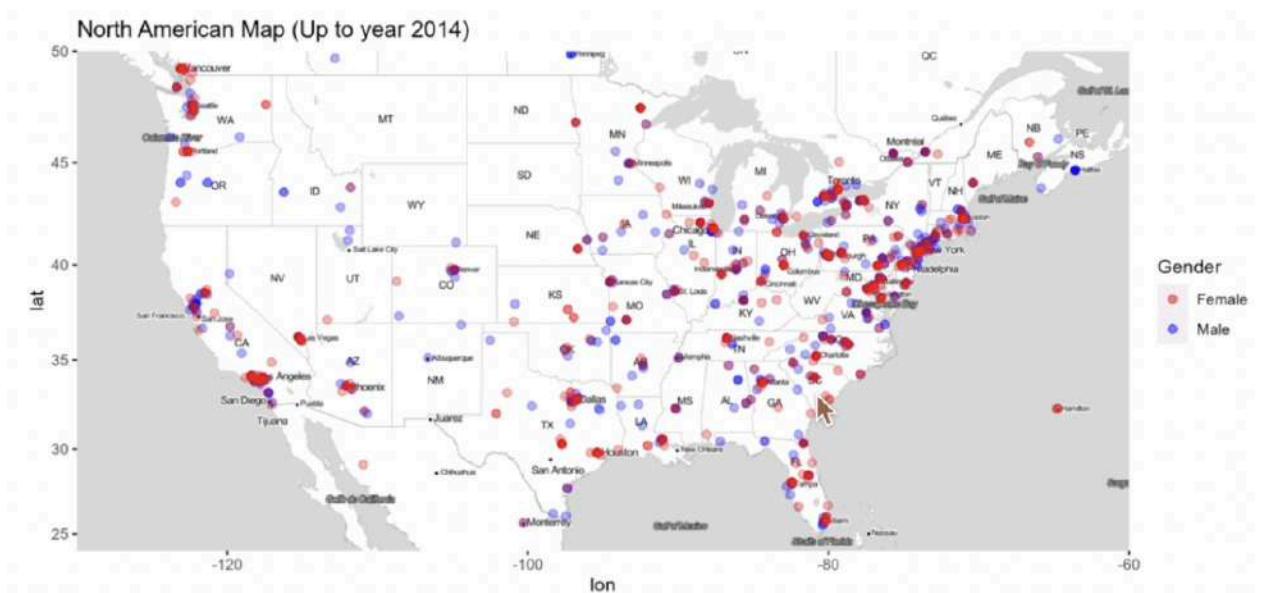
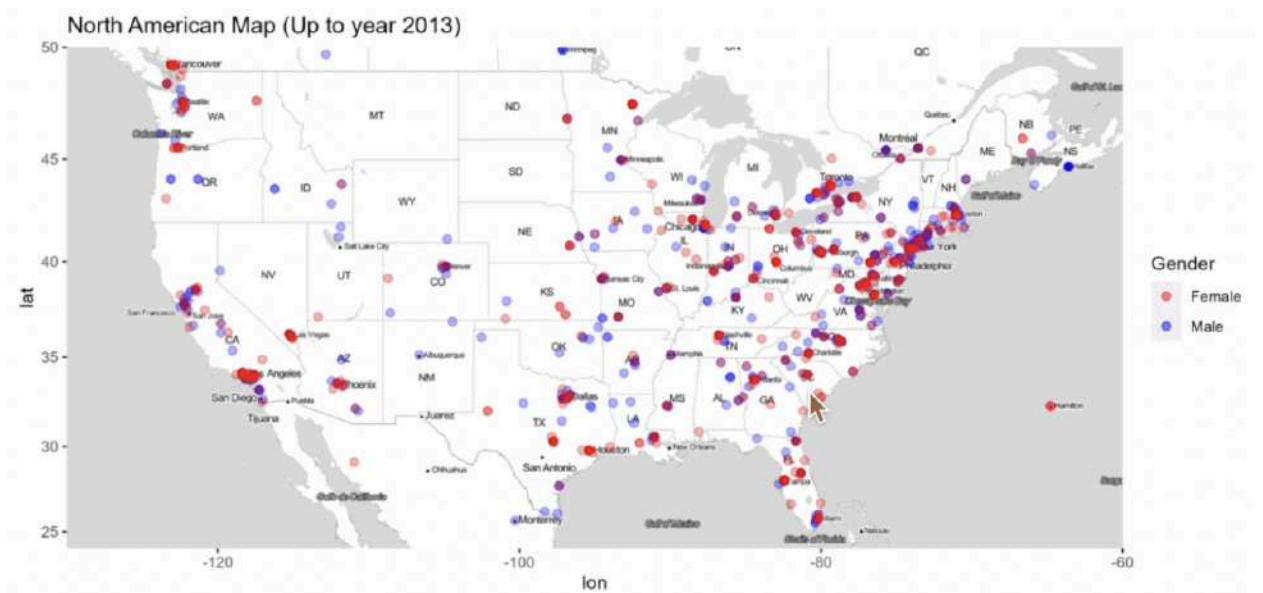
North American Map (Up to year 2010)

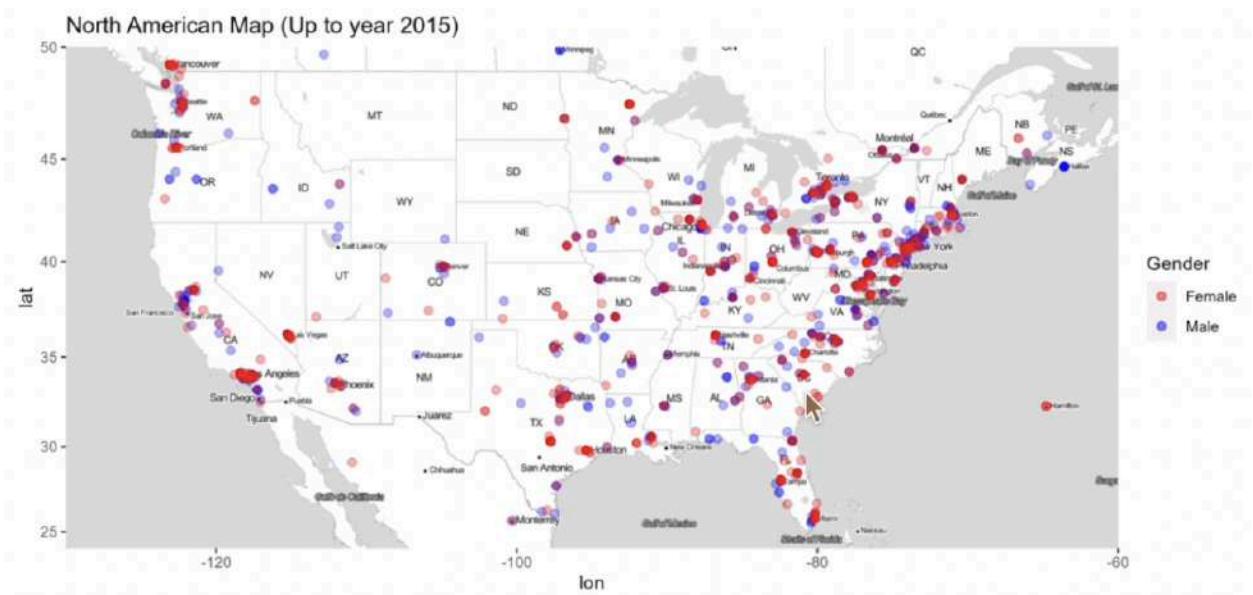


Gender

- Female
- Male







Focusing on the growth of Twitter users in North America from 2006 to 2015, intriguing and valuable patterns are revealed. Looking at the data, one of the most notable findings is that initially, there were no female users in our database. This fact suggests that male users were the pioneers in adopting Twitter in this region, according to our database.

Moreover, in exploring the geographic distribution of user growth over time, we notice a fascinating trend. In the early years, the concentration of users was mainly centered around capitals and larger metropolitan areas. Cities like San Francisco, New York, and Philadelphia stand out as the first hubs of Twitter adoption in North America.

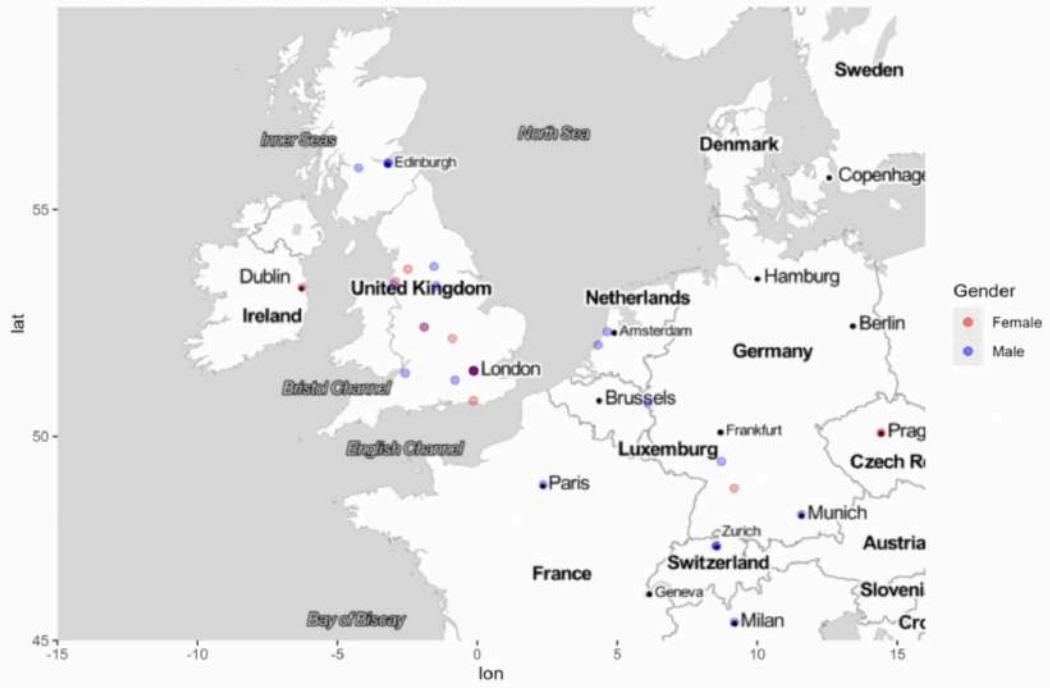
As we move forward in time, we can observe a diffusion of Twitter adoption from these capitals to non-metropolitan regions. This observation reflects how the social networking platform gained popularity and expanded beyond urban centers to more suburban and rural areas.

These observations highlight the evolution of Twitter adoption in North America.

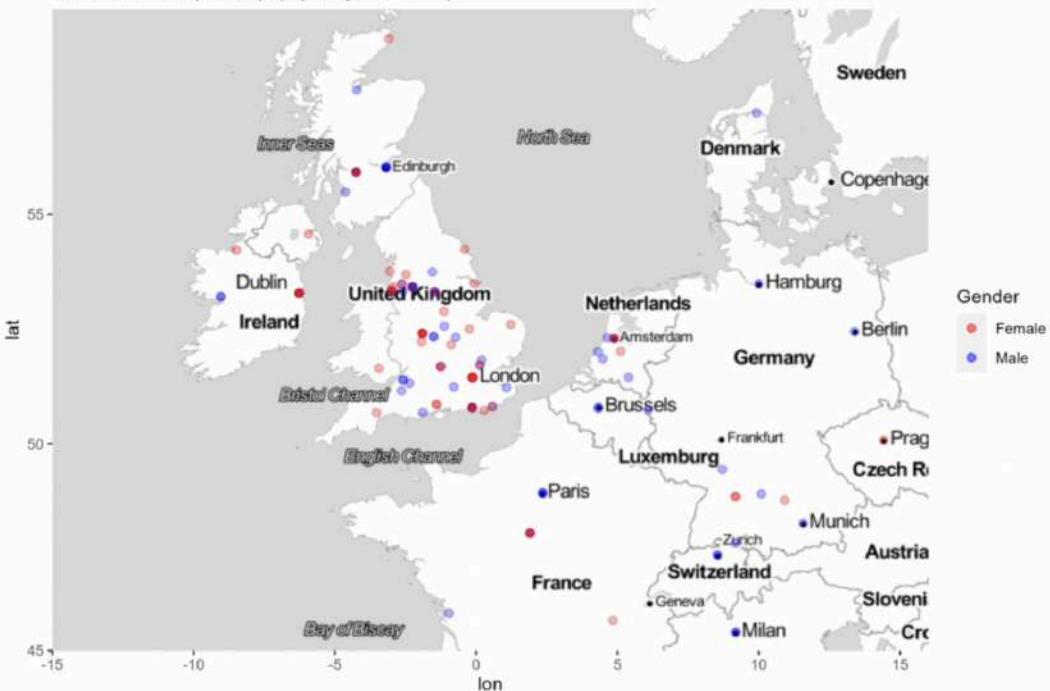
User Growth in North Europe (2006-2015)



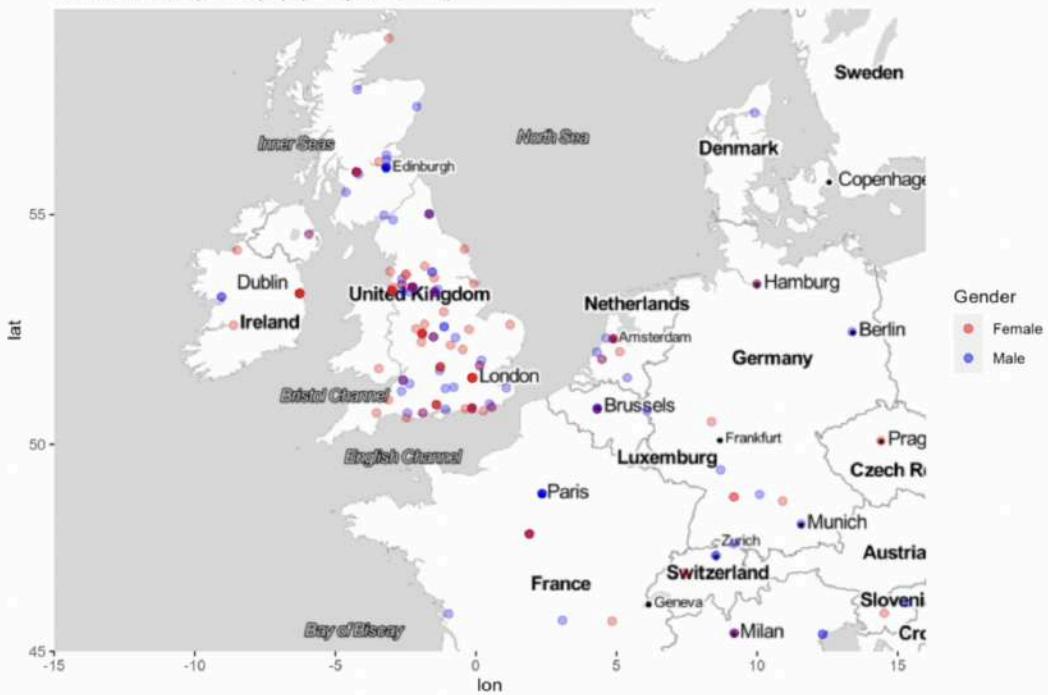
Northern Europe Map (Up to year 2008)



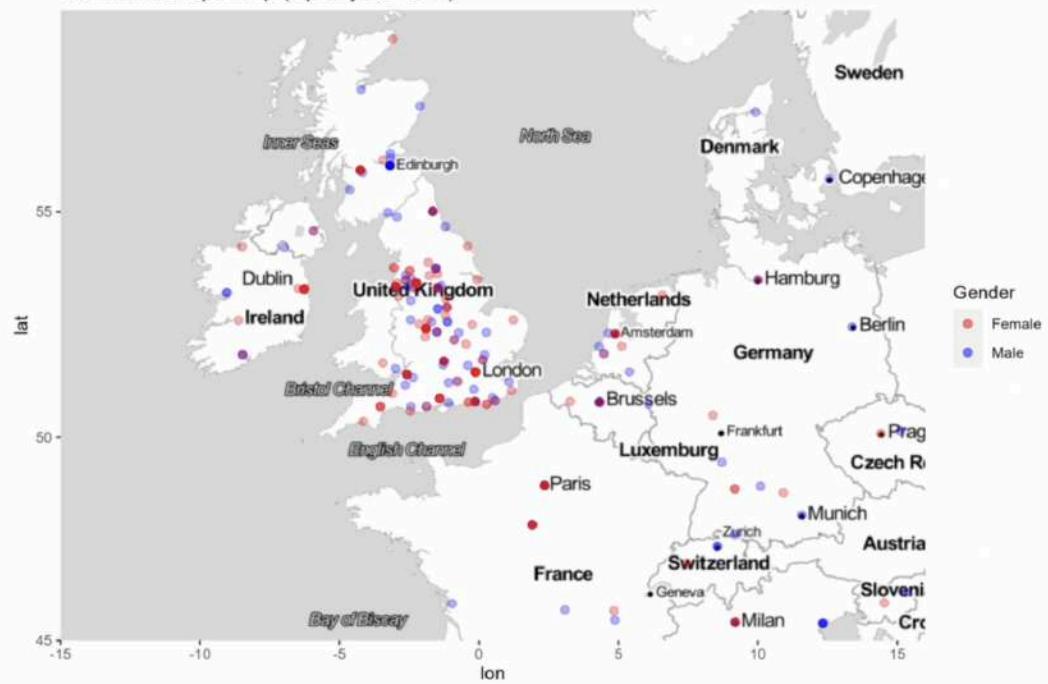
Northern Europe Map (Up to year 2009)



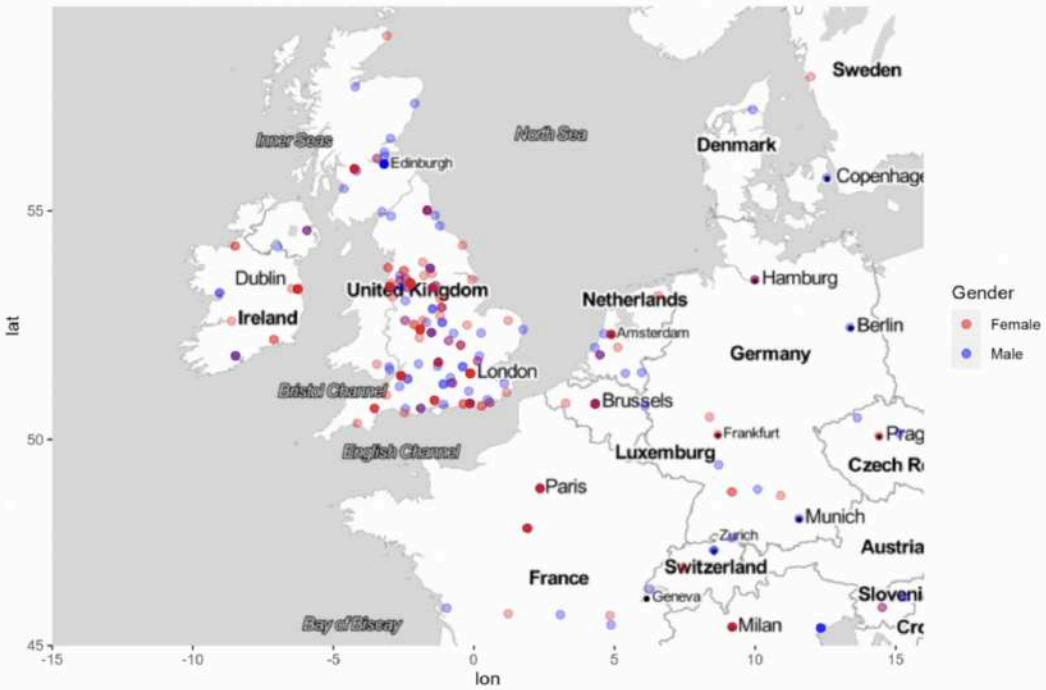
Northern Europe Map (Up to year 2010)



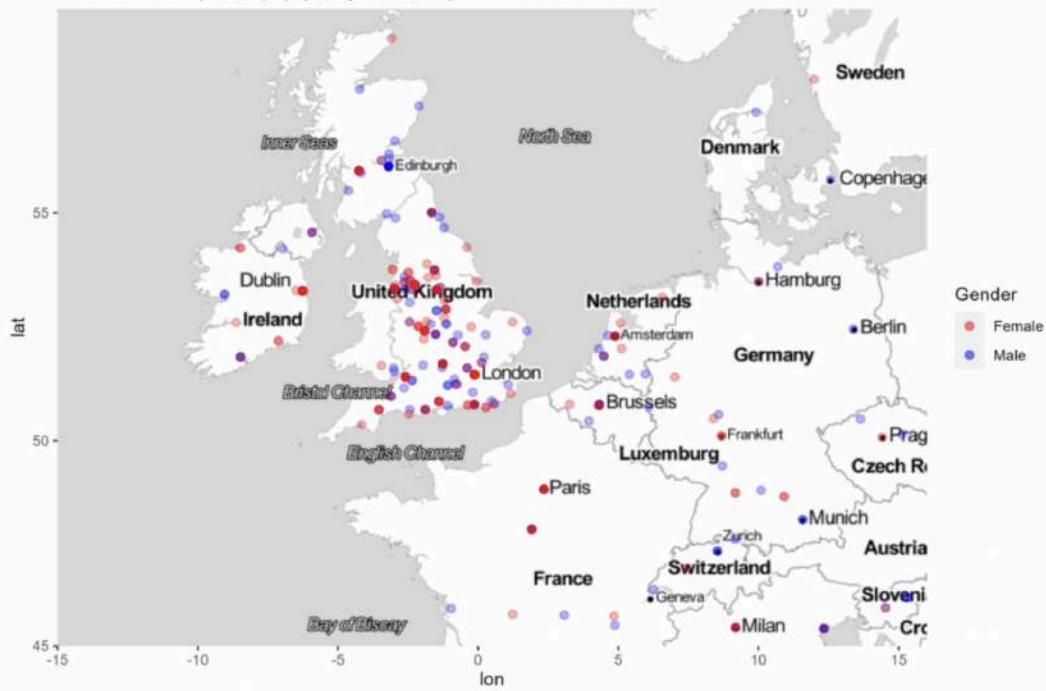
Northern Europe Map (Up to year 2011)

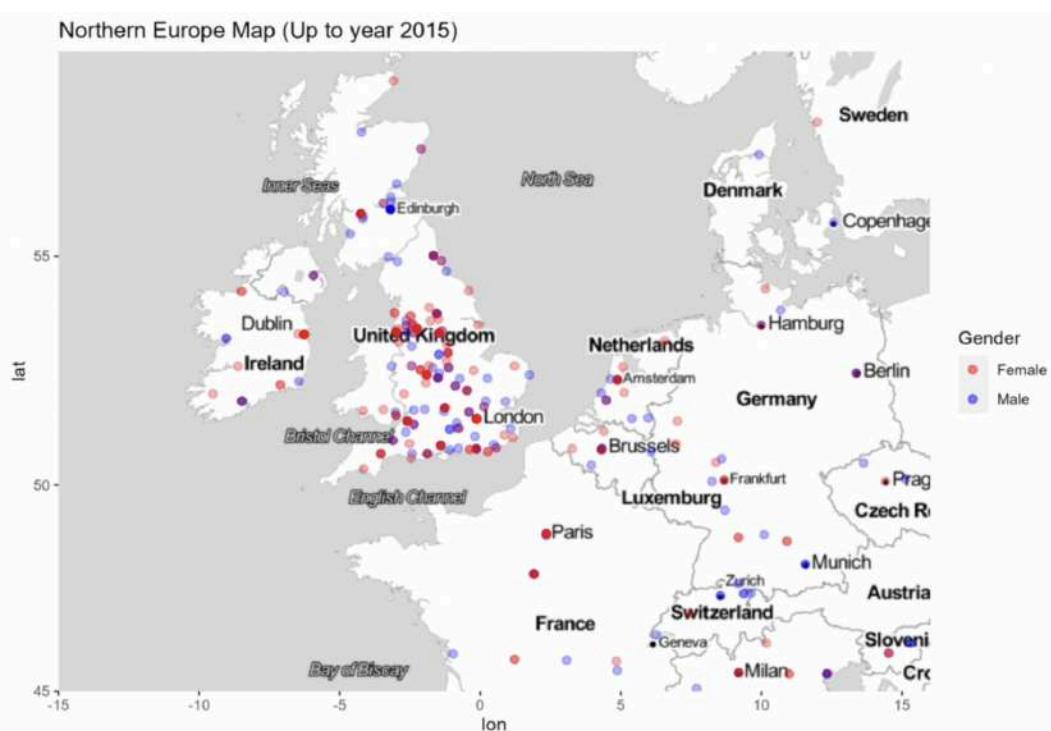
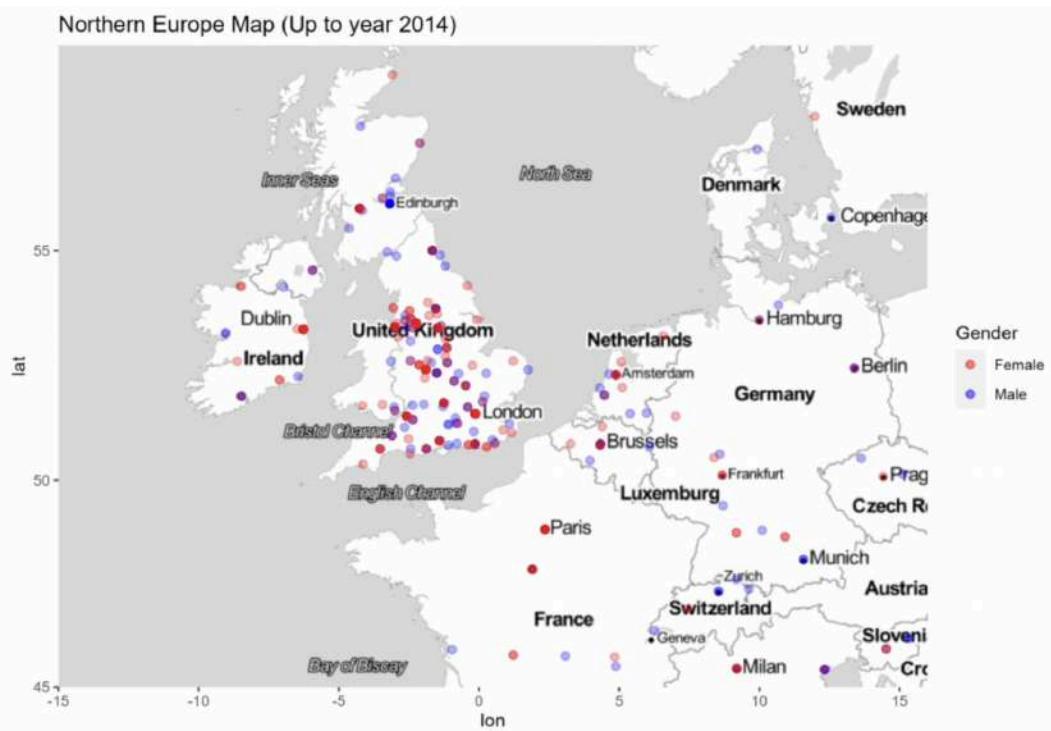


Northern Europe Map (Up to year 2012)



Northern Europe Map (Up to year 2013)





Continuing with the analysis, we focus on the growth of Twitter users in Northern Europe during the same time period, from 2006 to 2015. In line with the trend observed in North America, the

data reveals a similar pattern in Northern Europe. Male users were again the first to adopt Twitter, according to our database.

In terms of geographic distribution, the early Twitter users in Northern Europe were mainly concentrated in the capitals. We can see notable concentrations of users in cities such as London, Manchester, and Liverpool. This pattern reaffirms the trend observed in North America that major urban centers were the first to adopt Twitter.

Moreover, it's interesting to note that although user concentration gradually expands to non-metropolitan areas over time, this phenomenon appears to be less pronounced in Northern Europe than in North America. This could be indicative of different social network adoption patterns between these two regions.

These observations demonstrate the utility of our spatiotemporal analysis approach, as it allows us to identify and compare user growth trends in different geographic regions. Through this analysis, we can gain a more detailed insight into the evolution of Twitter adoption in Northern Europe and how it compares to other parts of the world.

Geographic Predominance of Gender by Zone

In this section, we will focus our analysis on the gender predominance in each of the regions under study: North America and Northern Europe. Our aim is to determine which gender is most prevalent in each region, providing a deeper insight into the geographic distribution of our users based on their gender.

Initially, we will carry out a general calculation of the number of male and female users in each region, and present these data in terms of percentages. This analysis will provide a panoramic view of the gender predominance in each region, allowing us to identify, on a large scale, if there is a notable disparity in the distribution of users of a specific gender.

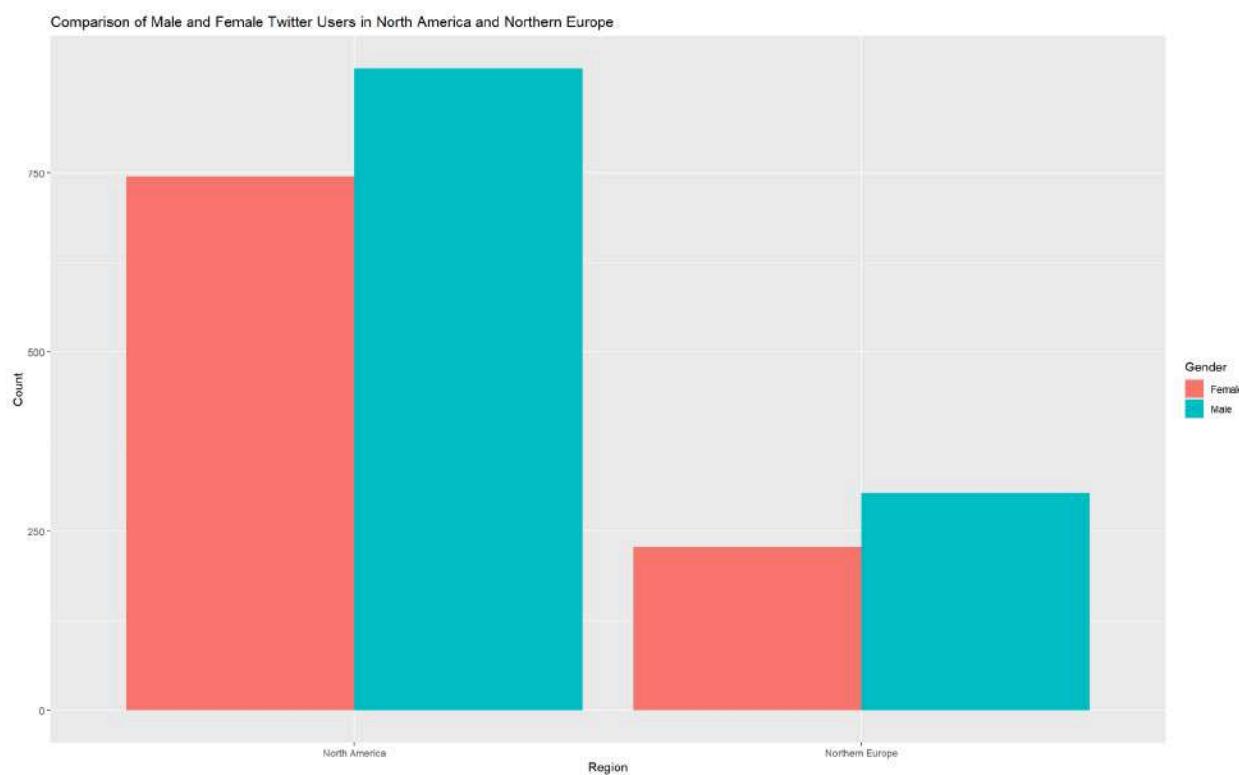
Next, we will broaden our analysis to provide a more granular view of gender predominance in each region. To do this, we will once again resort to the visualization technique of heat maps, which have proven to be an effective tool to represent our observations in a clear and visually intuitive way.

In this "Gender Predominance Heatmap", the colors will represent the relative density of users of a specific gender in a region. This will allow us to more accurately identify the areas where one gender is more predominant and understand how this distribution compares to that of the other gender.

By combining these approaches, we will be able to provide a detailed and nuanced assessment of gender predominance in the regions of North America and Northern Europe, adding an additional layer of complexity to our understanding of the geography of Twitter users.

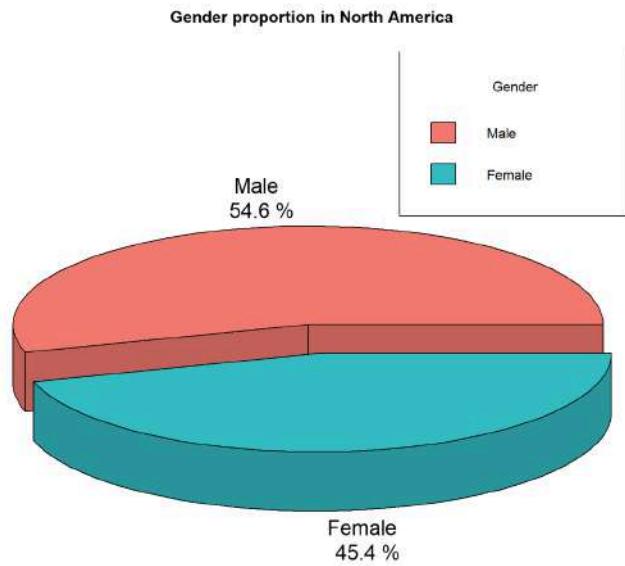
Gender distribution

To provide a more detailed view of the gender distribution in the regions of North America and Northern Europe, we have carried out two types of graphical analyses: bar charts and 3D pie charts.

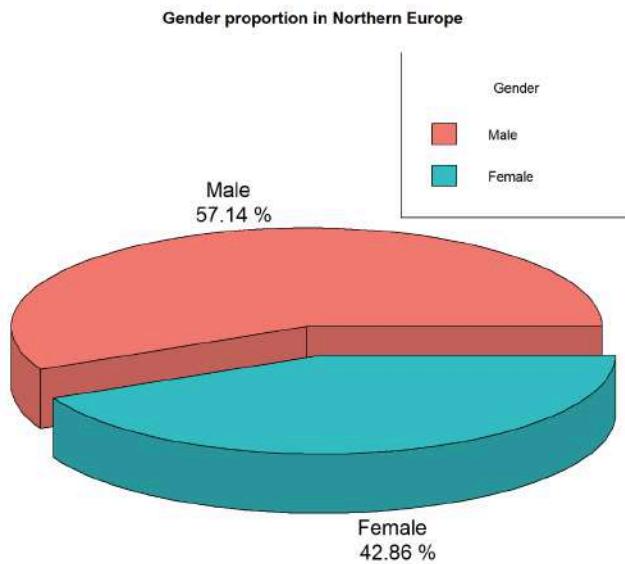


The bar chart provides a clear and straightforward view of the number of users of each gender in each region. In the North America region, we found approximately 750 female users and 900 male users. On the other hand, in Northern Europe, the data reveal a smaller total number of users, with around 240 female users and 300 male users.

Although these numbers provide us with a quantitative view of the gender distribution, they do not give us an accurate picture of the proportion of each gender in each region. For this, we turn to 3D pie charts, which allow us to visualize gender proportions in a more intuitive way.



The pie chart for North America shows that male users account for 54.6% of the total, while female users represent 45.4%.



Meanwhile, the chart for Northern Europe reveals a distribution slightly more skewed towards male users, with 57.14% male users and 42.86% female users.

These analyses provide us with both a quantitative and proportional view of the gender distribution in our study regions, allowing us to understand more accurately the predominance of each gender in each region.

Gender Predominance Heatmap

While bar charts and pie charts provide useful information about the overall gender distribution in our study regions, we wanted to delve deeper and analyze gender predominance at a more localized level. To do this, we have created gender predominance heat maps.

In these maps, the predominance of the `male` gender is represented in a more blue tone (#34bdc3), the predominance of the `female` gender in a more pinkish tone (#f27a70), while areas of gender equality (or those for which we do not have data) are indicated in white.

To create these maps, we have used the `kde2d` function from the MASS package in R, which estimates a bivariate density using a kernel. Although the technical details of this function are beyond the scope of this report, the basic idea is that it allows us to estimate the density of occurrences (in this case, gender predominance) in a two-dimensional matrix (our geographical coordinates).

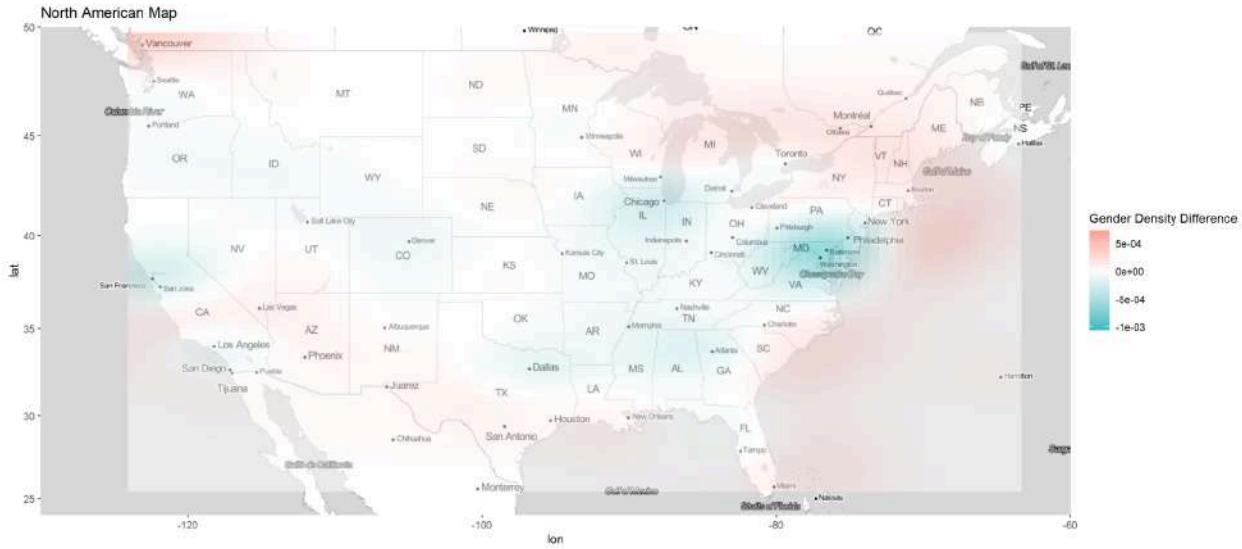
In simple terms, we have created a grid over our study regions, and for each point on the grid, the `kde2d` function estimates the density of users of each gender based on their proximity to that point.

Esta es la función que hemos definido para calcular estas densidades:

```
calc_density <- function(data, n = 100) {  
  lims <- c(range(data$lon), range(data$lat))  
  grid <- expand.grid(lon = seq(lims[1], lims[2], length.out = n),  
                      lat = seq(lims[3], lims[4], length.out = n))  
  dens <- kde2d(data$lon, data$lat, n = n, lims = lims)  
  grid$z <- as.vector(dens$z) # Convert dens$z to a vector  
  return(grid)  
}
```

Using this technique, we have managed to generate heat maps that reveal gender predominance in each region at a much more detailed and localized level than before. In the following subsections, we will analyze these maps for each of our study regions.

North America Gender Predominance



Upon observing the gender predominance heat map for North America, several notable features stand out.

Firstly, we can see that male predominance is manifested in areas such as Maryland, Chicago, and San Francisco. In these places, the blue hue of our heat map is more pronounced, indicating a higher concentration of male Twitter users.

However, this does not imply that female presence is nonexistent or even low in these areas. Rather, it indicates that, comparatively, there are more male users than female users in these specific locations.

In contrast, Vancouver displays a predominance of female users, which is denoted by a more pinkish hue on our heat map. This suggests that, at least in this city, female users are more active or numerous on Twitter compared to male users.

It's important to note that these findings are the result of the information that is available in our data set and should not be interpreted as a definitive representation of the gender distribution of all Twitter users in North America.

In the following subsection, we will repeat this analysis for the Northern Europe region.

North Europe Gender Predominance

When exploring gender predominance in Northern Europe, we observe some interesting trends.

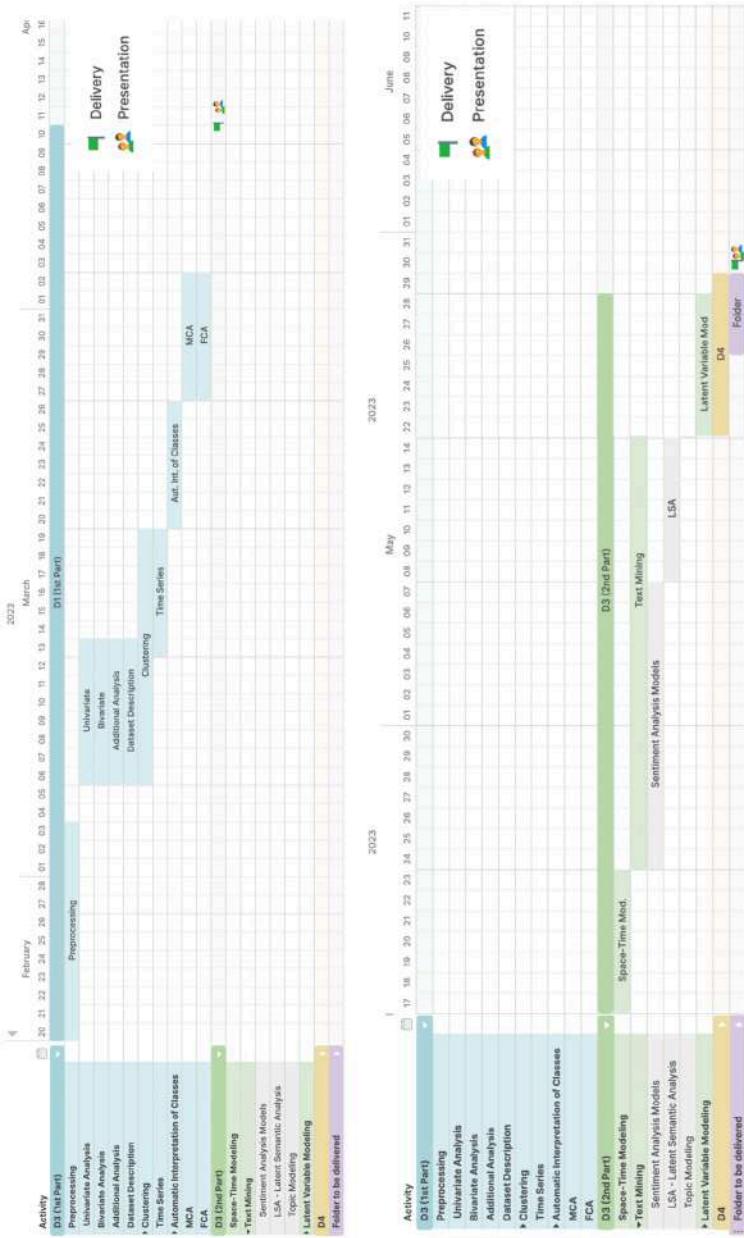
In the case of London and Manchester, male predominance is apparent. The more pronounced blue tone on the heat map in these areas indicates a higher concentration of male Twitter users. Although female presence is also notable, the data suggests that men constitute a larger proportion of Twitter users in these cities.

However, in the outskirts of London, we observe a slightly pinker tone. This suggests a slight predominance of women, which could indicate that female users are slightly more active or numerous in these areas than male users. However, it is important to note that this predominance is rather limited and less pronounced than what is observed in the cities.

As mentioned earlier, these findings are specific to our dataset and should not be interpreted as a definitive representation of the gender distribution of all Twitter users in Northern Europe. However, they provide an interesting insight into gender dynamics in Twitter usage in these areas.



GANTT DIAGRAM



In this section, we will comment on our progress, ranging from our previous expectations to the current reality. It appears that, despite just having vague knowledge on what tasks we had to do in the time we did the initial Gantt that our expectations were met. The main reason why being that the statements we put in it were vague enough as to give us a good range of liberty if there were more work in a task than what was previously thought.

CONCLUSIONS

FEATURE ENGINEERING

During the feature extraction stage of our statistical project, we focused on transforming our data into a more manageable format that would enable us to draw meaningful insights from it. One of the key transformations we made was converting the color variable from a hex code to a categorical variable. This allowed us to more easily analyze the relationship between color and other variables in our dataset. Another important feature we extracted was the continent variable, which provided a broader, more general overview of the location of users in our dataset. By grouping users by continent, we were able to gain a better understanding of geographical trends and patterns in our data. Furthermore, we also extracted the word count and average word length of each observation's tweet and description. This allowed us to identify trends in language usage, such as whether users tend to write longer or shorter tweets, and whether there are any variations in language usage across different demographics. Overall, these feature extraction techniques enabled us to gain deeper insights into our data and draw more accurate conclusions from our statistical analysis.

During feature selection, we noticed a strong correlation between the added variables that represented the length of the description. Consequently, we had to discard one of them, namely, the variable "description_length". Moreover, we discovered that numeric variables generally had more predictive power in determining gender, but the categorical variable "link_basic_color" also played a significant role in gender prediction.

MISSING DATA TREATMENT

In the missing data treatment stage of our statistical project, we aimed to explore the effectiveness of different imputation methods for handling missing data. We applied three different techniques: MICE, MiMMi, and Random Forest, on real values in our dataset. We then compared the method distributions to the real distributions and evaluated the performance measures (RMSE and Accuracy) to determine the best imputation method for our dataset. After careful consideration, we selected MICE as the best method, based on its ability to preserve the original distribution of the variables. We used MICE to impute both the real missing values and the outliers that we wanted to replace. Finally, we ensured that the distribution of the variables

remained consistent after the imputation process. Overall, our findings suggest that MICE is a reliable method for handling missing data in our dataset.

OUTLIERS

In the data analysis process, outlier detection plays a critical role in identifying anomalous or atypical values that may skew results or distort relationships between variables. By employing multiple methods, including univariate IQR, Mahalanobis distance, and Isolation Forest, we have successfully detected 1,387 outliers in our dataset. By using low thresholds for each method (0.001 for Mahalanobis distance and 0.01 for Isolation Forest), we ensured that the identified data points are genuine outliers.

Outlier treatment involved replacing these values with missing values and imputing new ones using the Multiple Imputation by Chained Equations (MICE) method. This process minimized the impact of outliers on our dataset and improved the overall quality and reliability of our statistical analysis.

In conclusion, our strategy for detecting and addressing outliers has been both effective and efficient. It has laid a solid foundation for subsequent statistical analyses and modeling, ensuring the accuracy and validity of our results.

MCA - MULTIPLE CORRESPONDENCE ANALYSIS

In our MCA (Multiple Correspondence Analysis) analysis, we aimed to identify relationships and patterns in our dataset by analyzing the associations between categorical variables. Our analysis revealed that the variables "predicted gender", color, and privacy were the most significant contributors to the variance in our dataset.

We found that there is a strong relationship between gender and color (supported by a chi-squared test). The biggest associations between the two were found to be blue and brand users, as well as between pink and purple and the "female" gender category. However, we did not find any significant relationship between continents.

Furthermore, we added additional quantitative variables to expand our analysis. This revealed that people tend to be more confident in their prediction when it is for male users. Additionally, we have found an association between those classified as male users and tweets with more words. In contrast, predicted female users tended to "like" more tweets and use shorter words in

their tweets. Additionally, we observed that unknown gender was more commonly associated with newer accounts, as opposed to male users.

Overall, our MCA analysis allowed us to gain a deeper understanding of the relationships between categorical variables in our dataset, and provided valuable insights into the behavior and preferences of different user demographics.

TIME SERIES CLUSTERING

In this section, we explore the relationship between the confidence of collaborators in determining users' gender and Twitter account creation dates using Time-Series-Clustering. To test Hypothesis 1, we created visualizations that represent the evolution of collaborators' confidence in determining users' gender over the year and grouped by months, but no clear pattern was identified to support the hypothesis.

For Hypothesis 2, we used Euclidean distance clustering and hierarchical clustering with the complete linkage method. We found that the optimal number of conglomerates was three, which were formed with consecutive years. This indicated a trend in the confidence of collaborators in determining users' gender over time. After visualizing the confidence of collaborators in determining users' gender by months and groupings, we partially confirmed Hypothesis 2, which suggests that it is easier to determine the gender of users with older Twitter accounts compared to those with newer accounts.

Overall, our analysis revealed that the confidence of collaborators in determining users' gender based on Twitter account creation dates has evolved over time, and it is easier to determine the gender of users with older accounts compared to those with newer accounts. No clear pattern was identified based on the account creation month or holiday periods.

CLUSTERING

During the clustering stage of our study, our objective was to observe how the clustering of our data was affected by implementing the CURE, DBSCAN, and OPTICS clustering methods. To do so, we conducted a k-means with PCA analysis to evaluate whether the latter two methods were necessary. We found that due to the complex shape of the resulting plot, simpler clustering methods would not suffice.

For CURE, we implemented and modified the code to be able to use CURE with categorical variables by using Gower's distance. We determined the amount of clusters by looking at the dendrogram. Then we were able to plot the results using PCA to visualize the clustering.

For DBSCAN, we implemented an elbow plot to calculate epsilon and used a percentage of the numerical values in our dataset to determine min-points. This resulted in a plot with several layers, one of which was heavily influenced by outliers while the other four were well-defined clusters.

In the case of OPTICS, we employed a grid search with a built-in silhouette method to find the optimal hyperparameters. Then, we used another silhouette method to determine the cutting point of the reachability plot. However, the clusters obtained with OPTICS were subpar, as they allowed for the creation of clusters with fewer than 10 individuals. As a result, we did not use the clusters obtained from OPTICS in the profiling step.

PROFILING

In the profiling step, we used the resulting clusters from both CURE and DBSCAN.

In both cases, we first implemented a series of tests to filter out any redundant variables. An example of a test we have done is the random forest that assesses the importance of a feature by measuring the decrease in accuracy or Gini index that results when that feature is randomly permuted while all other features remain unchanged.

Afterward, we did a bivariate plot for all the relevant variables to analyze the distinct characteristics that each cluster possesses and inference test to make sure that the results for each cluster are actually different from each other and not by chance. Consequently, a CPG and its simplified version (TLP) were made to group each variable into a view to simplify cluster interpretation and shade light on their traits and how they mingle together. These views were inspired by the sections made in the univariate analysis, which are the

For CURE, the clusters can be summarised as the following:

- **Cluster 1:** Predominantly composed of Missing continent profiles made by either a brand or unknown, with a high privacy who write average twitter posts in terms of word length but lower amount of words and a low gender confidence. It also likes to keep its description at a minimum in all aspects.

- **Cluster 2:** A male profile with average twitter posts in terms of word length and word count with a slightly higher gender confidence who doesn't value privacy. High in favorite number.
- **Cluster 3:** A female profile who chose, and has low privacy. Its tweets are average in word length and word count. Usually from North and South America.
- **Cluster 4:** A brand or male profile with high word length, but average traits in terms of word count. It also possesses a long and wordy description and low privacy setting.

In DBSCAN, this resulted in a group of clusters where they shared a significant amount of similarities to the study made in the MCA section by gender. These clusters can be summarized in the following way:

- **Cluster 1:** Predominantly comprised of old North-American profiles made by either a male or female, with low privacy who write average tweeter posts in both terms of count and word length and high gender confidence.
- **Cluster 2:** A brand profile made from an unknown gender with also average tweeter posts which average gender confidence and also highly values privacy.
- **Cluster 3:** Either a brand profile or made from an unknown gender who chose all blue in profile personalization, is young, and has a non-disclosed location but privacy is not the most predominant choice. Its tweets are high in numbers and low in content but the words are longer. It also likes to keep its description at a minimum in all aspects.
- **Cluster 4:** A brand or male profile who tweets rarely but when it does the tweet possesses average traits. It also possesses an average description and privacy setting.

TEXT MINING PREPROCESSING

In this study, we underwent an extensive preprocessing of textual data from Twitter, focusing on tweets and user descriptions for gender classification. Our preprocessing involved several key stages: initial cleanup to remove spam, language detection and translation to English, character and number cleaning along with letter case normalization, tokenization, stemming, and stopwords removal. In each stage, we used specialized techniques such as regular expressions, Google Translate API, and NLTK library functions.

In the tweets, once superfluous content was eliminated, salient themes emerged around news and weather updates, indicating the dataset likely contained numerous weather-related tweets. For user descriptions, the removal of less significant elements unveiled a wider array of themes and attitudes such as love, music, and life, providing more nuanced insights into user self-representation. However, the handling of contractions presented a challenge in

preprocessing. The frequency distribution of words stayed relatively consistent throughout, implying that while preprocessing modifies the data representation, the foundational structure remains unchanged.

SENTIMENT ANALYSIS

Our main objective when performing sentiment analysis was to see if the sentiment of the description and tweet of the user affected people's classification of the user's gender.

One of the things we did was a simple analysis of the words used. For example, we checked which words were unique to each gender and not used in the rest. For those users classified as males or females, we found typically masculine words for male users and typically feminine words for female users. As for brands, they used more professional words, and the unknown gender used internet slang.

We also used polarity scores and classified each text with a score. One of the things we did was plot the density distribution in which we saw that those classified as brands or unknown remained mainly neutral, while female and male were a bit more distributed to the extremes. Using an ANOVA model, we confirmed that gender was statistically significant in explaining polarity scores.

Afterwards, we applied sentiment classification, in which each text was classified in one of 8 sentiments. By plotting the overall proportion we saw that the main sentiment used were joy and anticipation, and the least were anger and disgust.

We also plotted the difference in proportion of each gender, and we could see that: males had a higher proportion in anger, disgust and joy, and a lower proportion in fear and sadness. For users classified as females, there was a higher proportion in fear and sadness and a lower proportion in positivity and trust. Lastly, we found that brands had very low proportions in all negative sentiments, while they had higher proportions in positivity and trust. We also used a chi-squared test to confirm the statistical relationship between these variables.

Overall, these were the conclusions, in which we saw that the stereotypes of what words each gender uses as well as the feelings they are expected to express are present in our data.

LSA

Using Latent Semantic Analysis (LSA), we have compared the similarity of documents based on their genres and found that, in general, the documents exhibit significant differences from one another.

Subsequently, we delved into the analysis of words that we have identified as particularly interesting, seeking their contextual associations. It became evident that the relationships uncovered through our LSA approach are highly influenced by the temporal context in which the data was collected. A striking example of this is the word "freedom," which exhibited a strong association with Iran. Moreover, when exploring terms such as "Obama" and "Trump," we observed a marked polarization in the descriptions of individuals.

Finally, it is important to acknowledge that our text sample, comprising Twitter data, is relatively small in size considering the complexity of the problem we are addressing.

LDA

We performed LDA on the tweet's text variable in order to see what different topics our users are tweeting about.

The first thing we had to do was pick the number of topics. Document clustering concluded that 8 was the best number, while the LDA tuning metrics said that 2, 5 or 8 were the best. For this reason, we experimented with 5 and 8 and saw that 8 got more interesting results.

As for the top terms, we weren't able to find useful information as the words were not very revealing. For that reason, we used the highest log ratio of words in each topic to define the topics better.

Although the themes found were a bit random, we were able to generalize each topic. We found topics that were political, some about music and others about entertaining.

Although we have found some interesting results, it has been challenging to do so since Twitter has so many different topics.

FACTORIAL ANALYSIS

In this section, we implemented a correspondence analysis with and without a generalized lexicon table, and for that, we selected the 20 most frequent words to ensure meaningful interpretation without excessive noise.

Then, we did a chi-squared test to ensure that the relationships we later find aren't by chance and we calculated the eigenvalues to make certain we took the appropriate number of dimensions to consider for the interpretation.

And while examining our results, we found out which words were the best represented by these chosen dimensions. In the text, the most important ones of such are the word new for dimension 1 and come for dimension 2. As for the description, news for dimension1 and im for dimension3.

Finally, using the findings we've observed with the results, we found that for the base correspondence analysis that in the description there were certain groups of words with related topics such as love, twitter, or events.

And for the extension, the main conclusion we found is that there's a clear influence of societal stereotypes and gender norms in the predictions. For instance, in the description, we can see how the female gender is associated with the words love or lover.

Moreover, we saw how there's an association between privacy and gender the user thought the profile was from an individual or a brand as we can see in the text results, the brand gender is near the no privacy modality whereas both female and male were linked to the opposite.

GEOSPATIAL

Through the geospatial analysis of the gender distribution of Twitter users in North America and North Europe, we have identified significant patterns and valuable insights.

The analysis of the geographical distribution of users suggests a notable concentration of users in certain regions of both zones. The spatial-temporal data has allowed us to understand the growth of the social network in these regions. It's notable that in both regions, men were the first to adopt Twitter. In the early years, users were mainly concentrated in the capitals, but over time, there was an increasing adoption of Twitter in non-capital areas.

Gender predominance has emerged as a relevant factor in this study. While men predominate in absolute terms in both regions, the female presence is significant in the user base.

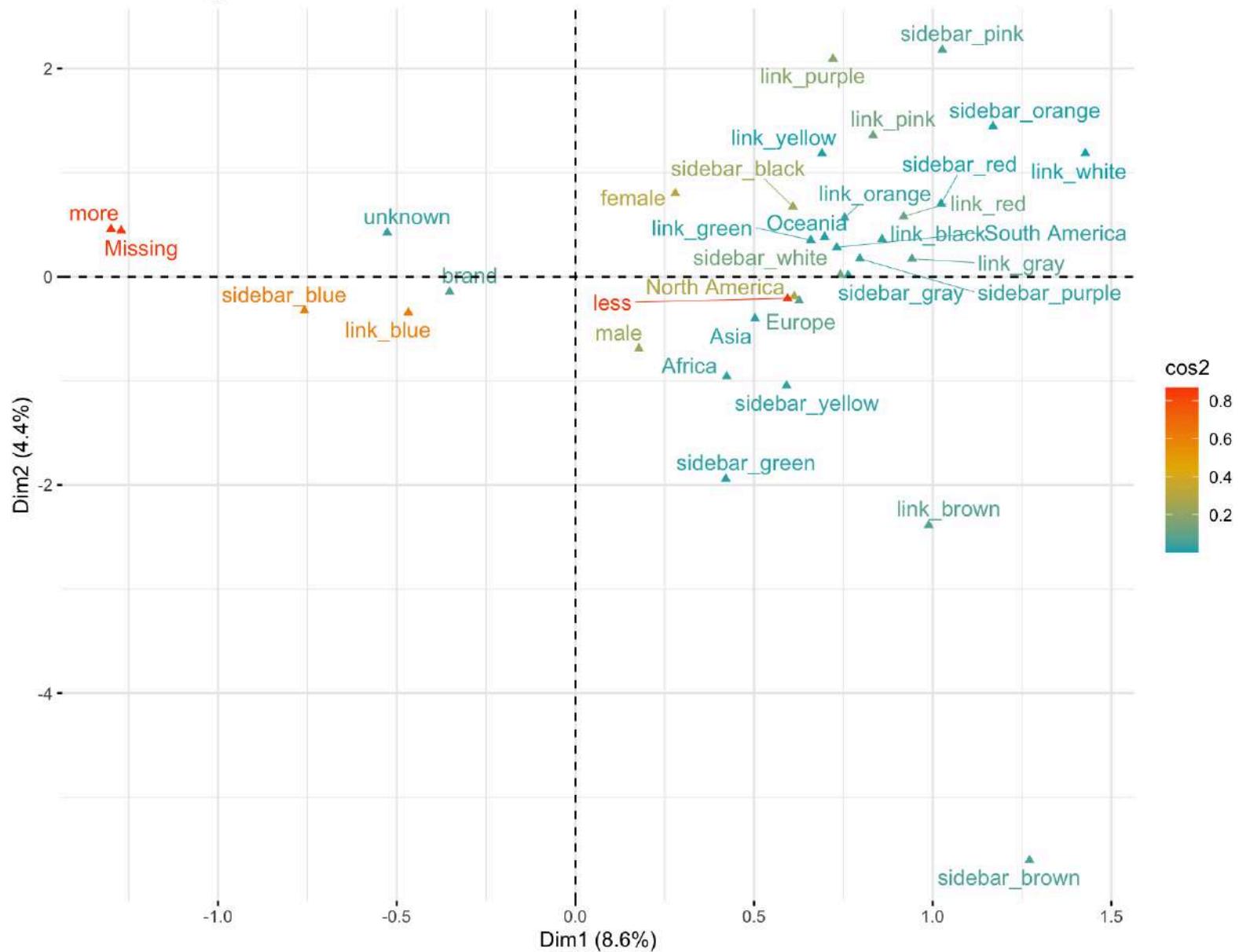
Even more detailed is the analysis of gender predominance by specific location. In North America, men predominate in Maryland, Chicago, and San Francisco, while women are more prevalent in Vancouver. In North Europe, London and Manchester show a male predominance, while on the outskirts of London, though by a very narrow margin, women predominate.

These findings illustrate the complexity of Twitter adoption and usage in terms of gender and geography. However, it should be kept in mind that these results are specific to our dataset and might not be fully representative of the global landscape. Future analyses could benefit from including more regions and exploring other demographic variables.

ANNEX

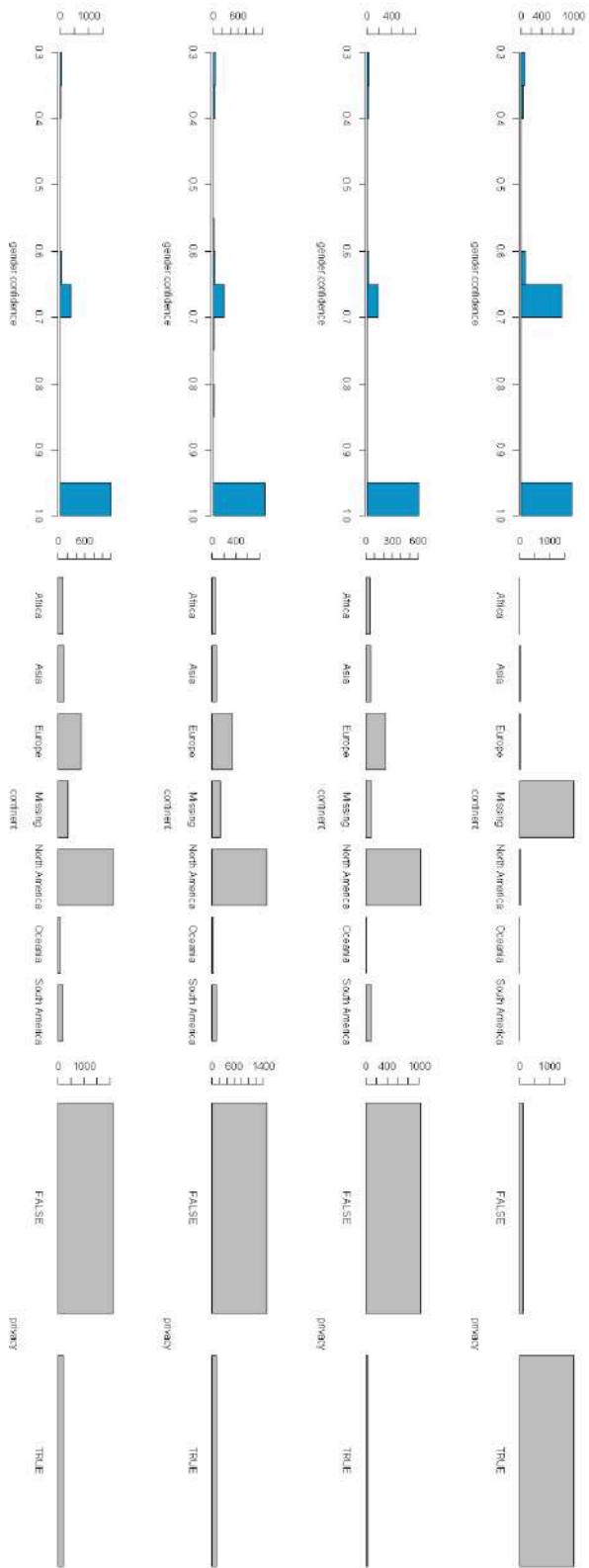
Annex 3. MCA

Variable categories - MCA



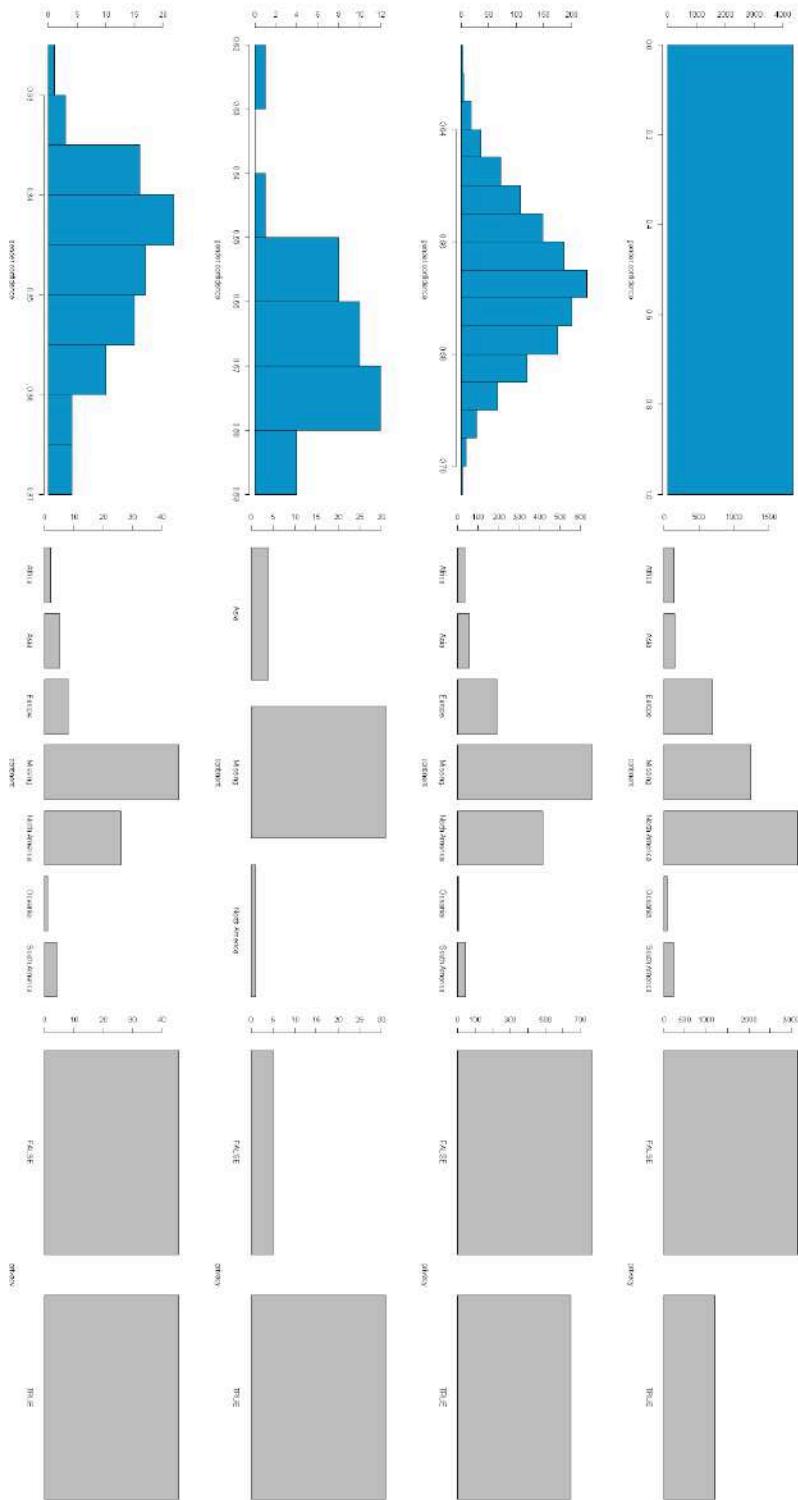
Annex 4. CPG CURE

User's personal information

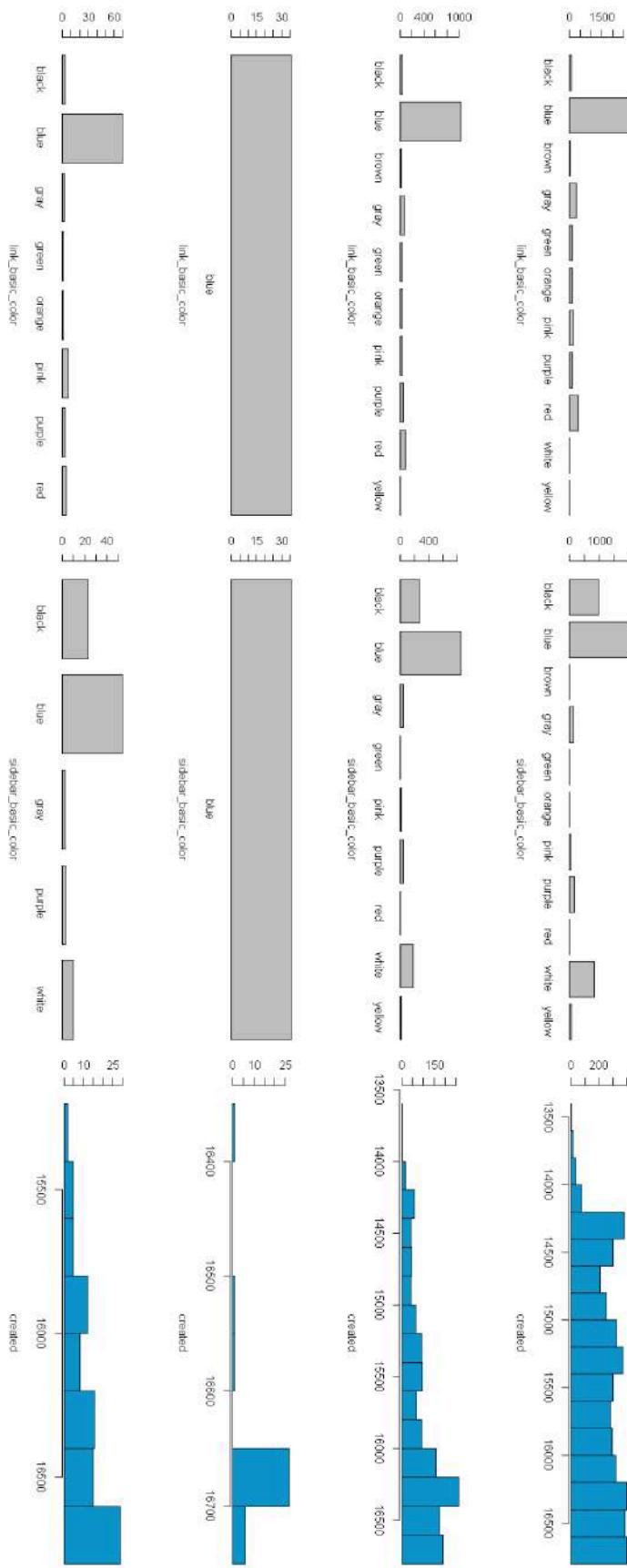


Annex 5. CPG DBSCAN

User's personal information



User's profile



Annex 6. Factorial analysis eigenvalues

CA text:

Eigenvalues		Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8
Variance		0.727	0.697	0.685	0.680	0.671	0.664	0.656	0.648
% of var.		6.035	5.786	5.687	5.640	5.572	5.510	5.447	5.380
Cumulative % of var.		6.035	11.821	17.508	23.148	28.720	34.230	39.677	45.057
		Dim.9	Dim.10	Dim.11	Dim.12	Dim.13	Dim.14	Dim.15	Dim.16
Variance		0.642	0.640	0.632	0.622	0.615	0.606	0.602	0.595
% of var.		5.324	5.308	5.244	5.165	5.103	5.029	4.998	4.942
Cumulative % of var.		50.381	55.689	60.933	66.098	71.201	76.230	81.229	86.170
		Dim.17	Dim.18	Dim.19					
Variance		0.578	0.561	0.528					
% of var.		4.796	4.653	4.381					
Cumulative % of var.		90.966	95.619	100.000					

CA description:

Eigenvalues		Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8
Variance		0.798	0.759	0.741	0.724	0.704	0.700	0.687	0.673
% of var.		6.456	6.136	5.996	5.853	5.693	5.658	5.560	5.446
Cumulative % of var.		6.456	12.592	18.588	24.442	30.135	35.793	41.353	46.798
		Dim.9	Dim.10	Dim.11	Dim.12	Dim.13	Dim.14	Dim.15	Dim.16
Variance		0.669	0.657	0.651	0.631	0.627	0.603	0.596	0.565
% of var.		5.407	5.315	5.265	5.107	5.072	4.880	4.822	4.568
Cumulative % of var.		52.205	57.521	62.786	67.893	72.964	77.844	82.666	87.234
		Dim.17	Dim.18	Dim.19					
Variance		0.549	0.520	0.509					
% of var.		4.444	4.208	4.114					
Cumulative % of var.		91.678	95.886	100.000					

CAGALT text:

Eigenvalues								
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8
Variance	0.033	0.016	0.011	0.010	0.008	0.007	0.006	0.006
% of var.	29.146	14.319	9.985	8.881	6.909	6.228	5.395	5.251
Cumulative % of var.	29.146	43.466	53.451	62.332	69.241	75.469	80.864	86.115
	Dim.9	Dim.10	Dim.11	Dim.12	Dim.13	Dim.14	Dim.15	Dim.16
Variance	0.004	0.003	0.003	0.002	0.001	0.001	0.001	0.000
% of var.	3.559	2.772	2.577	1.886	1.162	0.751	0.555	0.399
Cumulative % of var.	89.673	92.445	95.023	96.909	98.070	98.821	99.376	99.775
	Dim.17	Dim.18	Dim.19					
Variance	0.000	0.000	0.000					
% of var.	0.133	0.089	0.003					
Cumulative % of var.	99.908	99.997	100.000					

CAGALT description:

Eigenvalues								
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8
Variance	0.232	0.034	0.017	0.016	0.012	0.010	0.009	0.008
% of var.	64.678	9.592	4.690	4.328	3.346	2.767	2.450	2.147
Cumulative % of var.	64.678	74.269	78.960	83.288	86.634	89.401	91.851	93.998
	Dim.9	Dim.10	Dim.11	Dim.12	Dim.13	Dim.14	Dim.15	Dim.16
Variance	0.005	0.005	0.004	0.003	0.002	0.001	0.001	0.001
% of var.	1.415	1.313	1.017	0.727	0.515	0.347	0.329	0.218
Cumulative % of var.	95.413	96.726	97.743	98.469	98.984	99.331	99.661	99.879
	Dim.17	Dim.18	Dim.19					
Variance	0.000	0.000	0.000					
% of var.	0.067	0.054	0.000					
Cumulative % of var.	99.946	100.000	100.000					