Exploratory Data Analysis on Facebook Data

Mitchell Keomoungkhoune
Department of Systems Science and Industrial Engineering
State University of New York at Binghamton, Binghamton NY 13902

Abstract:

Data refers to the information that is accumulated through observation to collect facts and statistics for analysis and evidence. Data is considered the qualitative and quantitative characteristics of information that are often gathered in industry to better comprehend the large amounts of intelligence in their domain or within customer patterns and behaviors. Ultimately, this information is used to make informative and precise business decisions based on the analysis and evaluation of the data. For this project, EDA is used to evaluate data collected from Facebook that contains information specific to the user and the amount or type of activity the user is subject to. The objective of this project was to manipulate the pseudo-Facebook dataset to help identify patterns and characteristics of the data in order to make actionable conclusions. And with the resulting information identify significant patterns and behaviors from the data in order to make critical business decisions and predictive machine learning models. After conducting EDA on the Facebook dataset, the analysis suggested that users within the 20-29 age range were most inclined to use Facebook, there were more male users than female users from the dataset, female users have greater friend counts and initiated friendships, users are typically absent from using the Facebook longer as age increases, and the mobile Facebook app is more popular than the Facebook online website.

Keywords: Data, exploratory, Facebook, analysis

1. Introduction

This project used exploratory data analysis to conduct an initial assessment on the information from the Facebook dataset on user characteristics and behaviors on the social media platform. The collected data on Facebook users is a large high dimensional dataset containing different variables and attributes that gives no basic or immediate standalone value. However, with EDA chief leaders of Facebook can use aspects of the data that can develop precise business decisions as well as aiding in predictive modeling operations. Primary data attributes within the dataset includes qualitative variables such as the date of birth and age of the user, user ID, and gender while quantitative data included attributes such as friend count, likes received, mobile likes, and tenure. Using EDA, significant patterns, trends, implications, and inferences were able to be drawn from the raw data on Facebook users to develop actionable conclusions. This information can also be used to build machine learning models to predict the behavior of Facebook users based on the user specific information and the trends and patterns of the users' activity.

2. Background

2.1. Significance of Data and Analytics

As a result of the digital age transformation on industry there are vast amounts of information and raw data that hold significant hidden messages and implications that are waiting to be discovered and unlocked. In order to hold competitiveness in the market many companies and enterprises designate large amounts of resources to analyze and evaluate the data surrounding their business and the information on customers that could better their overall operational efficiency. There is a large significance on data quality and analytics that is typically utilized in every aspect of a businesses' operation. In an article by Forbes on the "The Age of Analytics and The Importance of Data Quality" explains that "49 percent of respondents said analytics helps them make better decisions, 16 percent says that it better enables key strategic initiatives, and 10 percent say it helps them improve relationships with both customers and business partners." [4]. Subsequently, the emphasis on collecting, analyzing, and understanding relevant data helps businesses make more informed and superior decisions, identify, and resolve problems, comprehend current or potential performances, and improve overall processes and efficiency [5].

Big data is a modern example of data emphasis or implementation into enterprises and businesses in a growing digital age. Big data refers to the systematic analysis of extracted information from datasets too large and complex for conventional data-processing methods. In an industry where digital transformation and automated technology is becoming more and more prevalent both enterprises and small-scale businesses rely on the collection and evaluation of data to run their business. Figure 1 shows the usage of big data in three major industries: telecommunications, financial services, and healthcare where the vast amounts of data are often managed by concepts of big data applications. The adoption of big data has grown exponentially where the number of companies investing in big data and/or some form of AI application grew from 17 % in 2015 has grown to 97.2 % in 2018 [7]. With the vast amount of information in the digital space of near all domains in industry, many businesses turn to EDA to assess the available data.

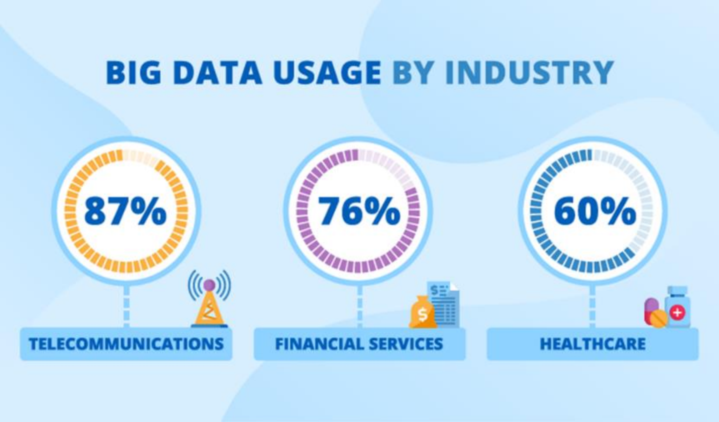


Figure 1. Big data usage in telecommunications, financial services, and healthcare [7]

2.2. Exploratory Data Analysis

EDA or Exploratory data analysis is the crucial process of performing initial analysis on data to discover patterns, identify anomalies, tests hypotheses, or check assumptions on the data. This analysis is often expressed through the use of summary statistics and graphical or data visualization methods such as boxplots, histograms, scatterplots, joint plots, multi-variate charts, parallel coordinates, etc. EDA is considered the first critical steps for data scientists, data analyst, and other data related specialist who need to evaluate and investigate the data and summarize its principal characteristics.

Exploratory data analysis was first developed and promoted by John Tukey, an American mathematician widely known for the development of the FFT (Fast Fourier Transform) box plot and algorithm as well as the Tukey's range test in statistical analysis [6]. Data analysis was defined by Tukey as the "Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data" which represented the basis of EDA [6]. Tukey used and promoted EDA to other statisticians to explore raw data and develop hypotheses or assumptions that could possibly lead to new information.

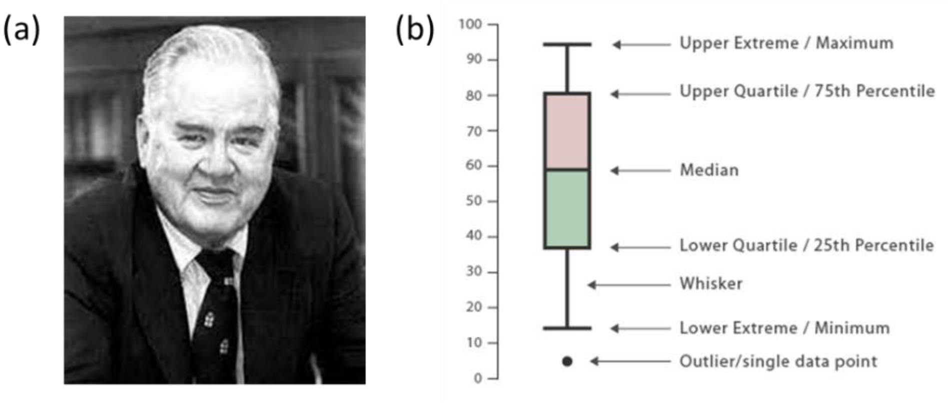

(a)
(b)
100
90
80
70
60
50
40
30
20
10
0
Upper Extreme / Maximum
Upper Quartile / 75th Percentile
Median
Lower Quartile / 25th Percentile
Whisker
Lower Extreme / Minimum
Outlier/single data point

Figure 2. (a) John Wilder Tukey (b) Tukey's box plot [9]

In 1977, Tukey wrote a book, "*Exploratory Data Analysis"*, that emphasized the over concentration of statistical hypothesis testing or confirmatory data analysis and suggested that there should be more emphasis on utilizing the data to create the hypotheses. This shifting ideology conveyed the importance of EDA by stating that the data collected from real-world industries should be the factor that dictates the nature of hypotheses based on the real-world. In this book he also designated the following objectives of EDA [6]:

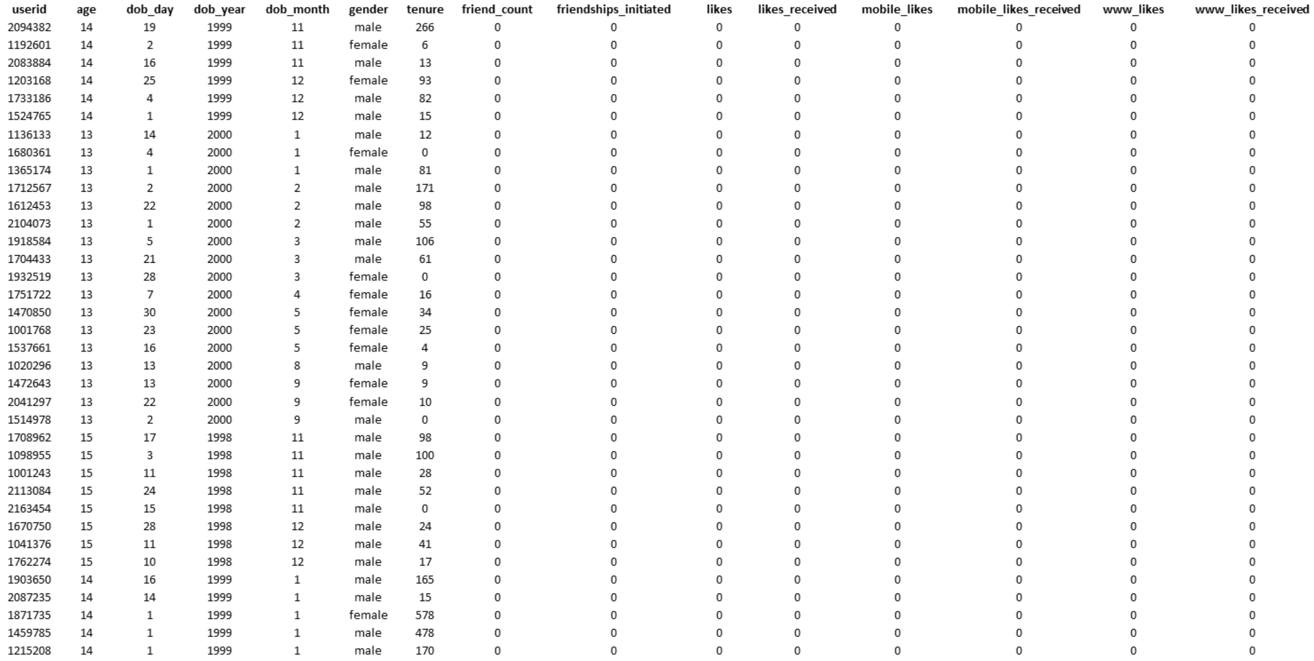
1. Suggest hypotheses about the causes of observed phenomena
2. Assess assumptions on which statistical inference will be based
3. Support the selection of appropriate statistical tools and techniques
4. Provide a basis for further data collection through surveys of experiments

Examples of EDA applications can be found among data scientist who use this method often to investigate, analyze, and summarize main characteristics of the dataset. An example would be EDA's utilization in clinical trials where EDA can help identify outliers in a patient population. EDA is also often used in retail as well where this analysis is used to identify customer spending patterns or units sold over time [3].

2.3. Facebook Dataset

Founded in 2004, Facebook is one of the most dominant social media platforms in the world. Currently, Facebook has 2.74 billion active users per month, dominates 59 % of the world's social networking populace, and is the third-most visited website and the second-most downloaded mobile app even compared to other social media giants such as Twitter, Snapchat, and Instagram [8]. Ultimately, Facebook is a media titan that constantly and continually floods the digital world with endless and endless amounts of data and knowledge. That being said it is critical for Facebook executives and data specialist to effectively and precisely manage the large amounts of data in order to maintain and take advantage of the available information to allocate resources and initiatives based on unveiling statistics.

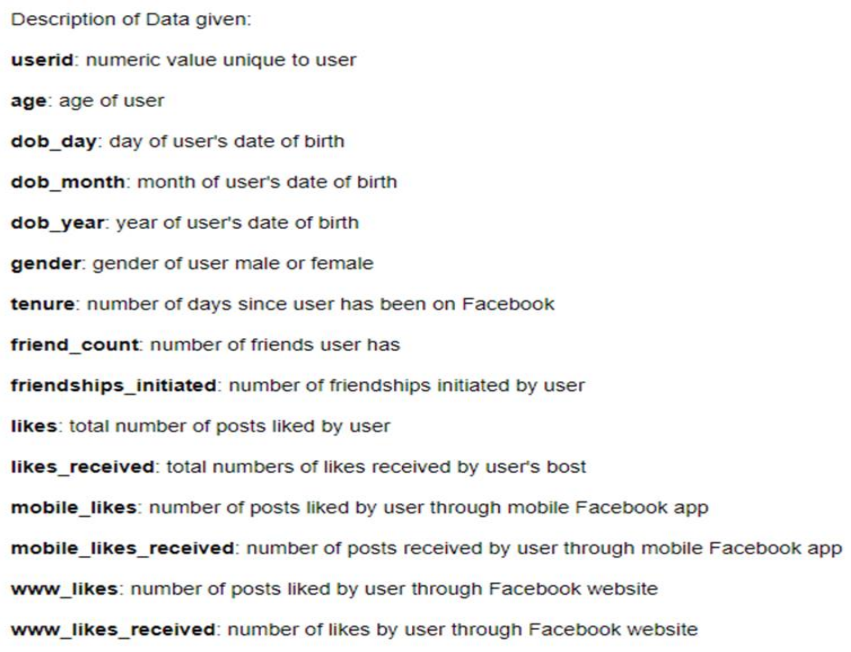
The Facebook dataset contains 99003 observations along with fifteen different independent variables. This dataset primarily consists of user information and the activity each user is subject to through Facebook "likes" using the mobile Facebook application or the online Facebook website. Figure 3 shows a sample of the Facebook dataset.



userid	age	dob_day	dob_year	dob_month	gender	tenure	friend_count	friendships_initiated	likes	likes_received	mobile_likes	mobile_likes_received	www_likes	www_likes_received
2094382	14	19	1999	11	male	266	0	0	0	0	0	0	0	0
1192601	14	2	1999	11	female	6	0	0	0	0	0	0	0	0
2083884	14	16	1999	11	male	13	0	0	0	0	0	0	0	0
1203168	14	25	1999	12	female	93	0	0	0	0	0	0	0	0
1733186	14	4	1999	12	male	82	0	0	0	0	0	0	0	0
1524765	14	1	1999	12	male	15	0	0	0	0	0	0	0	0
1136133	13	14	2000	1	male	12	0	0	0	0	0	0	0	0
1680361	13	4	2000	1	female	0	0	0	0	0	0	0	0	0
1365174	13	1	2000	1	male	81	0	0	0	0	0	0	0	0
1712567	13	2	2000	2	male	171	0	0	0	0	0	0	0	0
1612453	13	22	2000	2	male	98	0	0	0	0	0	0	0	0
2104073	13	1	2000	2	male	55	0	0	0	0	0	0	0	0
1918584	13	5	2000	3	male	106	0	0	0	0	0	0	0	0
1704433	13	21	2000	3	male	61	0	0	0	0	0	0	0	0
1932519	13	28	2000	3	female	0	0	0	0	0	0	0	0	0
1751722	13	7	2000	4	female	16	0	0	0	0	0	0	0	0
1470850	13	30	2000	5	female	34	0	0	0	0	0	0	0	0
1001768	13	23	2000	5	female	25	0	0	0	0	0	0	0	0
1537661	13	16	2000	5	female	4	0	0	0	0	0	0	0	0
1020296	13	13	2000	8	male	9	0	0	0	0	0	0	0	0
1472643	13	13	2000	9	female	9	0	0	0	0	0	0	0	0
2041297	13	22	2000	9	female	10	0	0	0	0	0	0	0	0
1514978	13	2	2000	9	male	0	0	0	0	0	0	0	0	0
1708962	15	17	1998	11	male	98	0	0	0	0	0	0	0	0
1098955	15	3	1998	11	male	100	0	0	0	0	0	0	0	0
1001243	15	11	1998	11	male	28	0	0	0	0	0	0	0	0
2113084	15	24	1998	11	male	52	0	0	0	0	0	0	0	0
2163454	15	15	1998	11	male	0	0	0	0	0	0	0	0	0
1670750	15	28	1998	12	male	24	0	0	0	0	0	0	0	0
1041376	15	11	1998	12	male	41	0	0	0	0	0	0	0	0
1762274	15	10	1998	12	male	17	0	0	0	0	0	0	0	0
1903650	14	16	1999	1	male	165	0	0	0	0	0	0	0	0
2087235	14	14	1999	1	male	15	0	0	0	0	0	0	0	0
1871735	14	1	1999	1	female	578	0	0	0	0	0	0	0	0
1459785	14	1	1999	1	male	478	0	0	0	0	0	0	0	0
1215208	14	1	1999	1	male	170	0	0	0	0	0	0	0	0

Figure 3. Facebook dataset sample

This dataset was provided by Kaggle, a data science community forum that also offers a large data repository. This particular dataset on the information of Facebook users was sourced from 2013 and primarily acts as a pseudo dataset to exercise data related training. Some variables in this dataset include dob_day, tenure, or likes which represent the date of birth of the user, the number of days the user has been on Facebook, and total number of posts likes by the user, respectively. Figure 4 shows each feature and the correlating definition.



Description of Data given:

userid: numeric value unique to user

age: age of user

dob_day: day of user's date of birth

dob_month: month of user's date of birth

dob_year: year of user's date of birth

gender: gender of user male or female

tenure: number of days since user has been on Facebook

friend_count: number of friends user has

friendships_initiated: number of friendships initiated by user

likes: total number of posts liked by user

likes_received: total numbers of likes received by user's bost

mobile_likes: number of posts liked by user through mobile Facebook app

mobile_likes_received: number of posts received by user through mobile Facebook app

www_likes: number of posts liked by user through Facebook website

www_likes_received: number of likes by user through Facebook website

Figure 4. Description of given data and its features

3. Methodology

3.1. Phases of EDA

The phases of EDA represent the necessary steps for each stage of the overall process. The first phase typically identifies the problem statement. For the project, the problem statement and Phase 1 was the Facebook data and potential implications that could be deciphered from it. Typically, in Phase 1 or the problem identification phase questions can be raised to better understand or clarify the problems. For this example, questions like "What age group are most inclined to use Facebook based on the dataset?" or "Are users more likely to use the mobile app or the Facebook website?" were asked in order to clarify the problem and establish directions for data manipulation. Phase 2 is the pre-profiling stage where the raw data undergoes a summarization and check for any missing values. Phase 3 is the data pre-processing steps which typically include data cleaning and manipulation methods. Phase 4 is the data visualization phase where graphical tools and data visualization methods are used to better represent the implications of the data. Lastly, Phase 5 is where results are summarized and interpreted based on the findings demonstrated by the visualization and graphical tools.

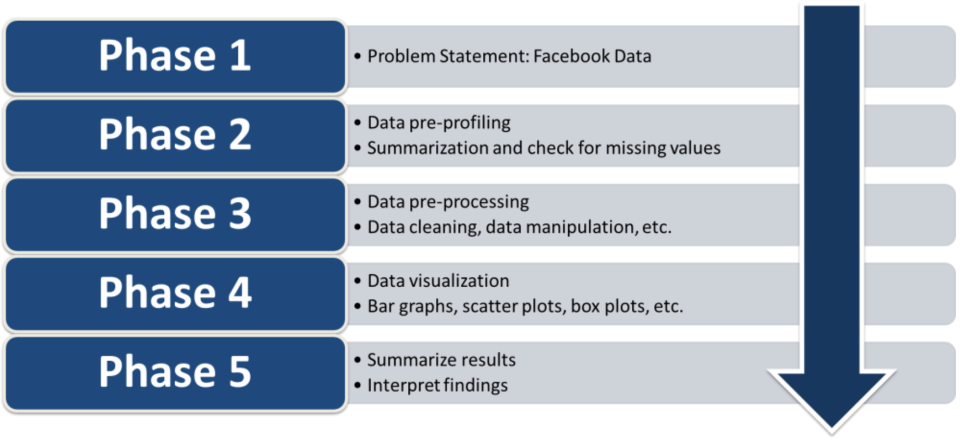
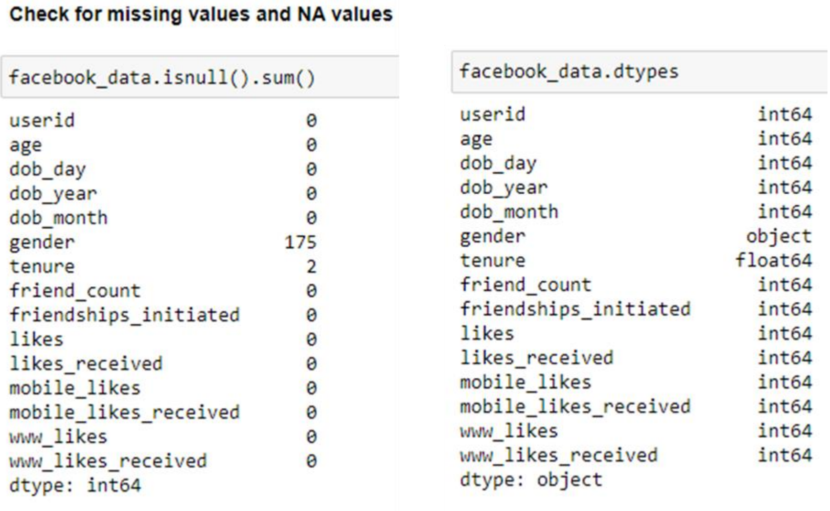

Phase 1
• Problem Statement: Facebook Data
Phase 2
• Data pre-profiling
• Summarization and check for missing values
Phase 3
• Data pre-processing
• Data cleaning, data manipulation, etc.
Phase 4
• Data visualization
• Bar graphs, scatter plots, box plots, etc.
Phase 5
• Summarize results
• Interpret findings

Figure 5. Phases of exploratory data analysis

3.2. Data Pre-Profiling

Before applying EDA to the Facebook dataset initial steps needed to be taken to properly manipulate the data to identify major characteristics of the dataset. Phase 1 or the data pre-profiling phase involves summarizing the raw data and monitoring for missing values and NA values that are within the dataset. Figure 6 shows that from the given Facebook data there are 175 NA values for gender and two missing values for tenure. The data type was also identified where gender was of type "object" and tenure was of type "float".



Check for missing values and NA values

```
facebook_data.isnull().sum()
```

```
userid                    0
age                       0
dob_day                   0
dob_year                  0
dob_month                 0
gender                  175
tenure                    2
friend_count              0
friendships_initiated     0
likes                     0
likes_received            0
mobile_likes              0
mobile_likes_received     0
www_likes                 0
www_likes_received        0
dtype: int64
```

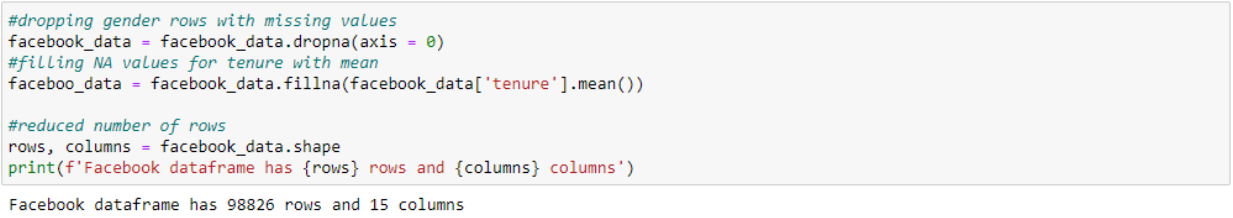
```
facebook_data.dtypes
```

```
userid                     int64
age                        int64
dob_day                    int64
dob_year                   int64
dob_month                  int64
gender                    object
tenure                   float64
friend_count               int64
friendships_initiated      int64
likes                      int64
likes_received             int64
mobile_likes               int64
mobile_likes_received      int64
www_likes                  int64
www_likes_received         int64
dtype: object
```

Figure 6. Data pre-profiling of Facebook dataset by checking for null values and identifying data type

3.3. Data Pre-processing

After data pre-profiling is complete data pre-processing is conducted to better format the data. Data pre-processing is the transformation of raw data into a more comprehensive configuration. Raw data is often incomplete, involves errors, and is inconsistent therefore data pre-processing is implemented to resolve these problems. The first step in data pre-processing for the Facebook dataset was to address the missing values in the dataset by dropping the rows with "NA" values for gender and filling the missing values for tenure with the mean of the column. Once this implemented the output of the rows shown in Figure 7 have decreased to 98826.



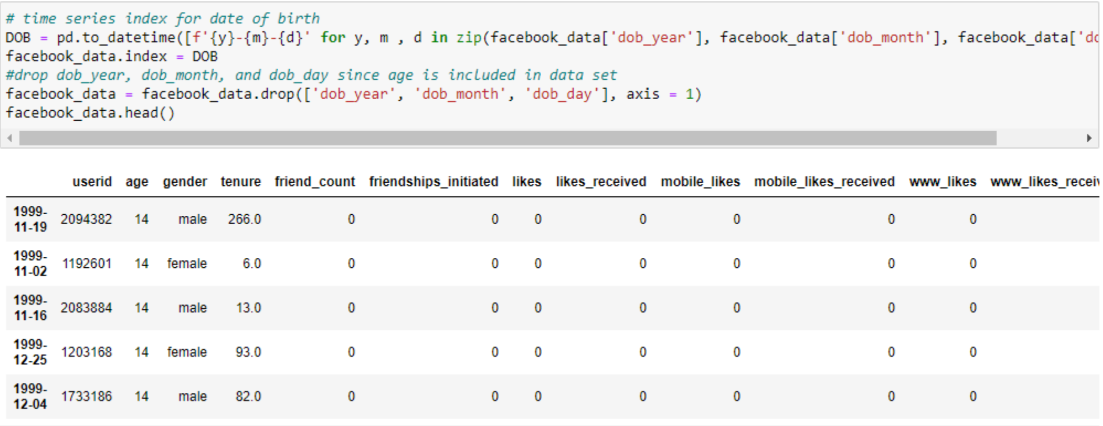
```
#dropping gender rows with missing values
facebook_data = facebook_data.dropna(axis = 0)
#filling NA values for tenure with mean
faceboo_data = facebook_data.fillna(facebook_data['tenure'].mean())

#reduced number of rows
rows, columns = facebook_data.shape
print(f'Facebook dataframe has {rows} rows and {columns} columns')
```

```
Facebook dataframe has 98826 rows and 15 columns
```

Figure 7. Python code displaying data pre-processing methods by dropping missing values in gender column and filling mean values for missing cells in tenure column

Columns for dob_year, dob_month, and dob_day were also dropped due to the age feature already containing relevant information in regard to these variables which is shown in Figure 8. Furthermore, the year, month, and day of the user's date of birth were reformatted into the first column in the "YYYY-MM-DD" configuration.



```
# time series index for date of birth
DOB = pd.to_datetime([f'{y}-{m}-{d}' for y, m , d in zip(facebook_data['dob_year'], facebook_data['dob_month'], facebook_data['do
facebook_data.index = DOB
#drop dob_year, dob_month, and dob_day since age is included in data set
facebook_data = facebook_data.drop(['dob_year', 'dob_month', 'dob_day'], axis = 1)
facebook_data.head()
```

	userid	age	gender	tenure	friend_count	friendships_initiated	likes	likes_received	mobile_likes	mobile_likes_received	www_likes	www_likes_recei
1999-11-19	2094382	14	male	266.0	0	0	0	0	0	0	0	
1999-11-02	1192601	14	female	6.0	0	0	0	0	0	0	0	
1999-11-16	2083884	14	male	13.0	0	0	0	0	0	0	0	
1999-12-25	1203168	14	female	93.0	0	0	0	0	0	0	0	
1999-12-04	1733186	14	male	82.0	0	0	0	0	0	0	0	

Figure 8. Dropped dob_year, dob_month, and dob_day columns due to age already incorporating this information

3.4. Data Visualization and Graphical Tools

The first EDA visualization that was explored was the age distribution for all users within the Facebook dataset using a bar graph. From this graph you can see that the majority of users are roughly in the range of 15-33 years of age. Meaning from this dataset Facebook users are generally teenagers and young to middle-aged adults.

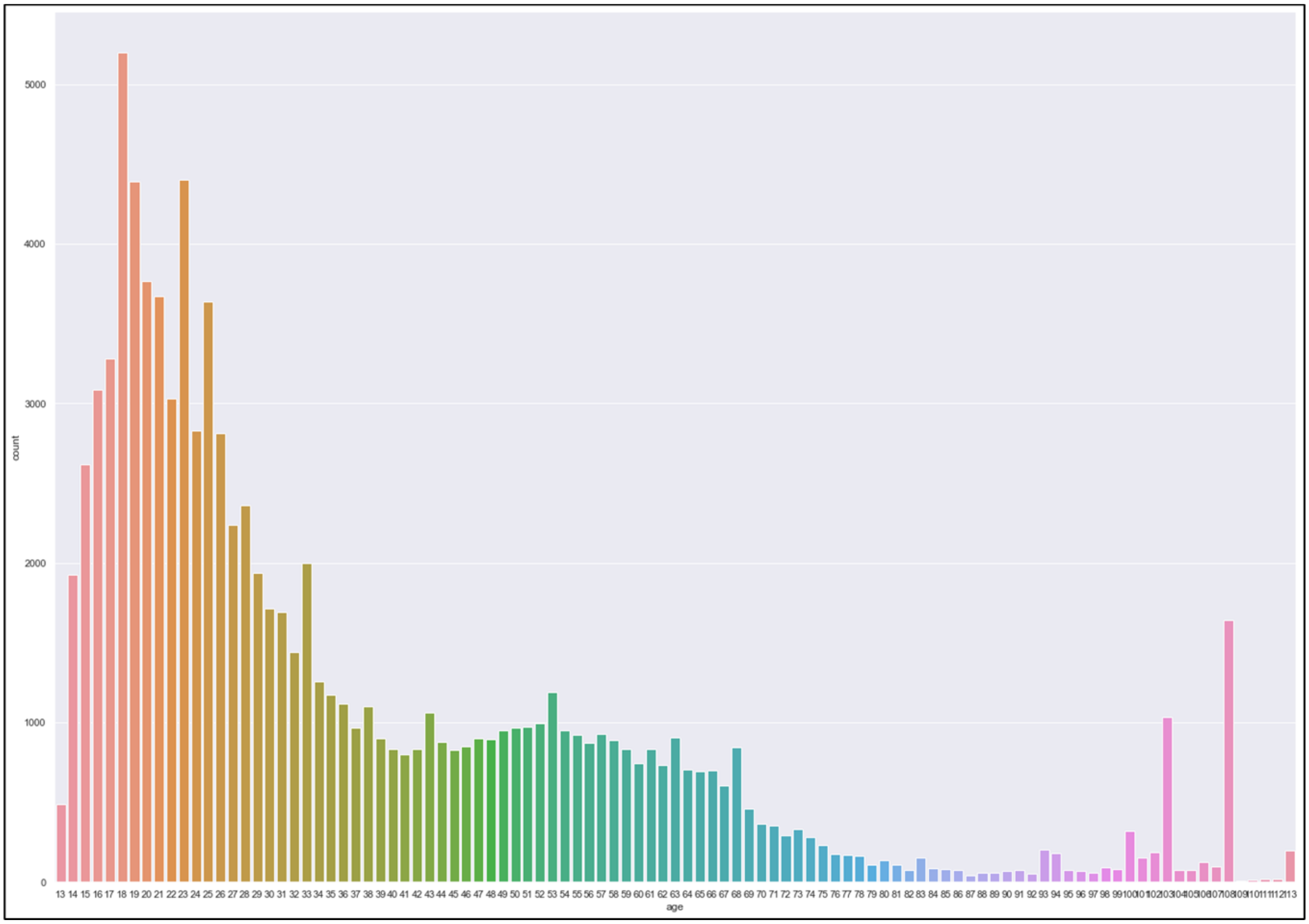


Figure 9. Distribution of age of Facebook users for dataset

Further initial investigation of the age of Facebook users from the dataset showed that the variable friend_count or the amount of friend the user has is higher for teenagers and young adults which is shown in the joint plot in Figure 10.

This trend corresponds with the bar graph of age distribution shown in Figure 9 demonstrating that the large number of users for this age range and friend count should be positively correlated.

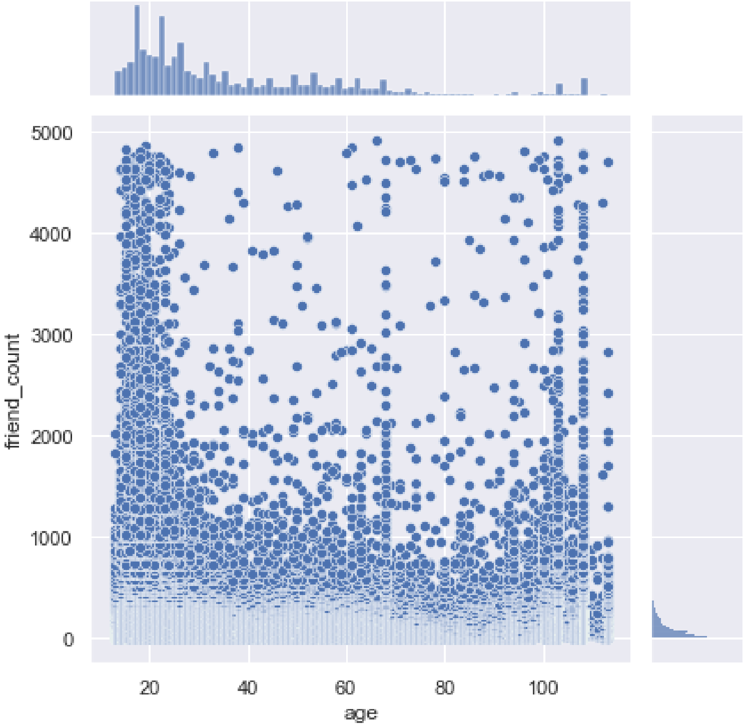


Figure 10. Joint plot of user age in relation to friend count

Gender among the users from the Facebook dataset was also investigated by using a pie chart shown in Figure 11. And from this data visualization you can see that there are more male users (59.3 %) than female users (40.7%) from the data.

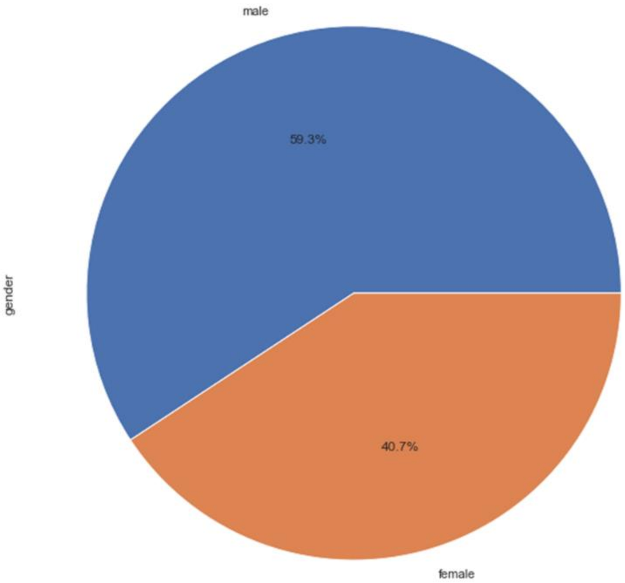
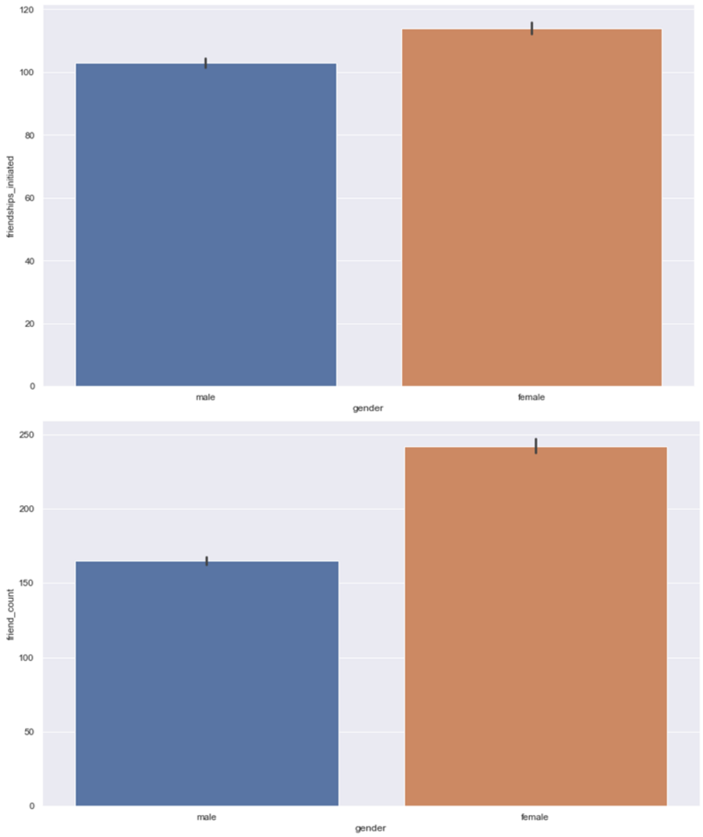


Figure 11. Percentage of gender among Facebook users in dataset

However, despite there being more male Facebook users when viewing the friend count and the friendship_initiated or number of initiated friendships by the user you can see that female users have greater numbers for each variable.

This instance is shown in Figure 12 by plotting bar graphs showing the relationship between gender with initiated friendships and friend count. This could possibly be an implication that female users are more active in their Facebook profiling.



Next, further exploration of gender was conducted by displaying the relationship between gender and designated age groups shown in Figure 13. For each gender, the majority of users fell within the 20-29 age range shown in the red box of Figure 13. This depiction further supports the initial evidence that most users fall within the teenager and young to middle-age adult range.

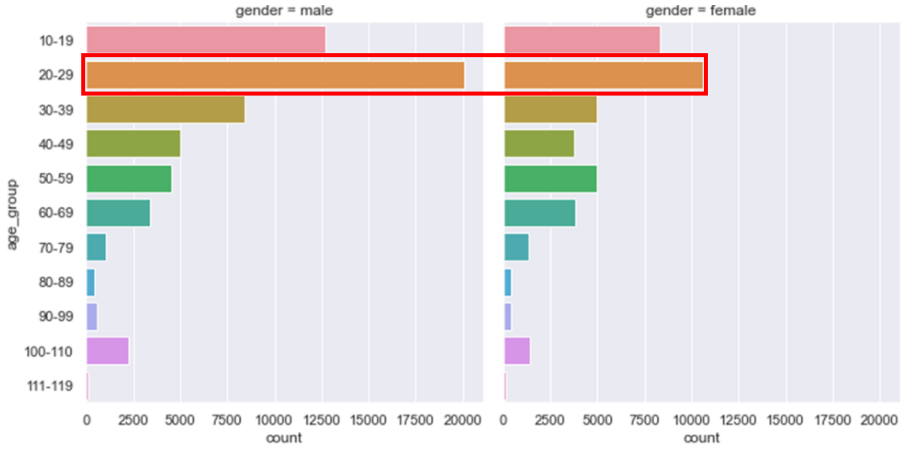


Figure 13. Bar graph for age group distribution for each gender

Investigation of tenure or the number of days since a user has been on Facebook in connection with both gender and age groups. Figure 14 shows box plots of how tenure is affected by gender or age group. From these plots it shows that generally male Facebook users have lower tenure than females or that male users tend to spend less time away from using Facebook than females. When viewing age groups and how this variable affects tenure you can see that tenure is lower for young age groups. A clear pattern also shown by this plot is the increasing tenure as age increases. This suggests that for age groups in the uppermost range of Facebook users tend to be disengaged from this particular social media platform for longer periods of time.

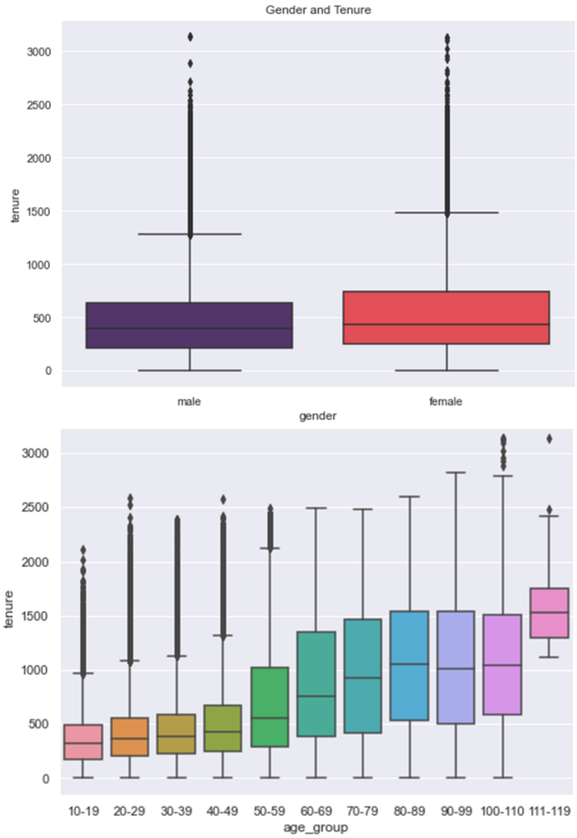


Figure 14. Box plot of tenure in connection with gender and age groups

Overall Facebook activity was also explored by plotting line graphs of mobile_likes, mobile_likes_received_, www_likes, and www_likes_received and its correlation to male and female Facebook users. The objective of this visualization was to determine the overall activity for each gender and the popularity difference between the Facebook mobile app and the Facebook website. This investigation is shown in Figure 15, where the blue and orange line represent overall activity from Facebook users using the mobile app while the red and green line represents Facebook activity from users utilizing the Facebook website. From this graph we can see that the majority of users from the dataset liked or received likes on posts on the mobile app.

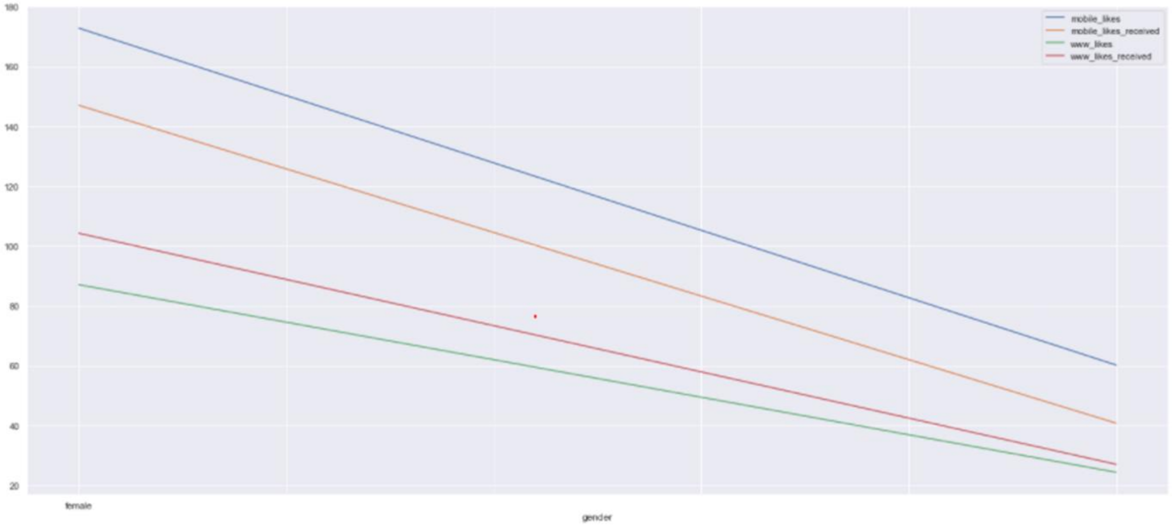


Figure 15. Line graph of gender and correlation between mobile Facebook activity and Facebook website activity

A heat map was also generated to show the pair-wise correlation for each independent variable shown in Figure 16 where lighter colored spaces represent variables with high positive correlation and darker spaces represented high negative correlation. This plot effectively demonstrates and identifies if any variables have some direct effecting relationship between another independent variable based on the Facebook dataset. From this heat map we can see that there is high correlation for variables such as the relationship between friend count and initiated friendships as well as the connection between total likes received and mobile likes received.

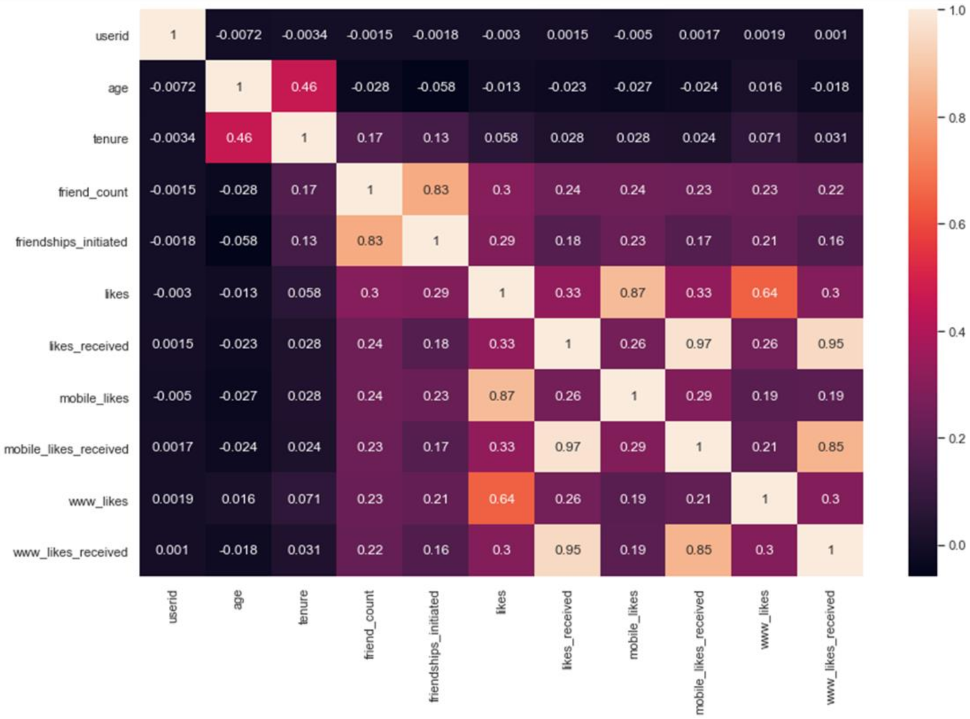


Figure 16. Heat map showing correlation between each variable of the Facebook data

4. Summary of Results

In summary based on the results from the EDA conducted on the Facebook data it was found that the age group of the majority of Facebook users were within the 20-29 age range and that among all user's males (59.3%) made up most of the Facebook population in the dataset as opposed to females (40.7 %). Although, despite there being more male Facebook users' females had higher rates for friend count and friendships initiated. Tenure was also investigated between age groups and gender and was found that generally tenure is lower for males and younger age groups. The usage of the Facebook mobile app and the Facebook website were also explored to determine which platform type was more prominent than the other. It was found that for both male and female users the majority of Facebook activity or likes posted and likes received were carried out on the mobile app at 63.4 % of total likes while activity on the Facebook website made up 36.6 % of total likes. A heat map was generated to visually display the pair-wise correlation for each variable and was found that there was a high positive correlation between friend count and initiated friendships as well as likes received, and mobile likes received.

5. Conclusions

In conclusion, exploratory data analysis was conducted on the Facebook dataset to identify patterns or trends to generate implications into actionable conclusions based on the generated data visualizations and graphical tools. From this analysis the following actionable conclusions were made on the Facebook dataset on user information and activity on Facebook:

1. Users within ages 20-29 are most prone to use Facebook
2. There are more male users than female users on Facebook
3. Females have more Facebook friends and initiated friendships
4. Users typically are absent from Facebook longer as age increases
5. Facebook mobile app is more popular than Facebook website

6. Future Work and Research

For future work and research other EDA visualizations can be made to further analyze the dataset for additional trends and patterns as well as hypothesis testing or assumption checking. Often is EDA implemented in tandem with machine learning or predictive modeling applications as well. With the resulting information and conclusions deriving from the EDA of the Facebook data machine learning models are able to be built to predict user behavior and patterns creating an extremely robust predictive model due to the large amount of available data. The conclusions from EDA on the dataset are also able to identify key areas or objectives where resources can be allocated towards. For example, from the dataset it was concluded that the Facebook mobile app was more popular than the Facebook website; further investigation should be taken in order to find the root cause for this trend and possibly focus directives towards increasing activity for users to use the website more.

References

[1]. Patil, Prasad. "What Is Exploratory Data Analysis?" *Medium*, Towards Data Science, 23 May 2018, towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15.

[2]. By: IBM Cloud Education. "What Is Exploratory Data Analysis?" *IBM*, www.ibm.com/cloud/learn/exploratory-data-analysis#:~:text=discovery%20process%20today.-,Why%20is%20exploratory%20data%20analysis%20important%20in%20data%20science%3F,interesting%20relations%20among%20the%20variables.

[3]. Simplilearn. "Exploratory Data Analysis: Techniques, Best Practices & Applications." *Simplilearn.com*, Simplilearn, 13 Oct. 2020, www.simplilearn.com/exploratory-data-analysis-article.

[4]. Panoho, Kale. "Council Post: The Age Of Analytics And The Importance Of Data Quality." *Forbes*, Forbes Magazine, 1 Oct. 2019, www.forbes.com/sites/forbesagencycouncil/2019/10/01/the-age-of-analytics-and-the-importance-of-data-quality/?sh=254b629b5c3c.

[5]. "Why Is Data Important for Your Business?" *RSS*, 9 Mar. 2020, www.grow.com/blog/data-important-business.

[6]. “Exploratory Data Analysis.” *Wikipedia*, Wikimedia Foundation, 7 May 2021, en.wikipedia.org/wiki/Exploratory_data_analysis.

[7]. Baturina, Olga. “40 Stats and Real-Life Examples of How Companies Use Big Data.” *ScienceSoft Footer Icon*, ScienceSoft, 21 Oct. 2020, www.scnsoft.com/blog/big-data-use-cases-stats-and-examples.

[8]. “47 Facebook Stats That Matter to Marketers in 2021.” *Social Media Marketing & Management Dashboard*, 11 Jan. 2021, blog.hootsuite.com/facebook-statistics/.

[9]. Ribecca, Severino. “A Look at Box Plots.” *Visualoop*, 6 Apr. 2015, visualoop.com/blog/32470/a-look-at-box-plots.