Data Analytics

&

Data Driven Decision

A.A. 2017-2018

Facebook Analysis

Data Analysis on “Facebook Comment Volume” Dataset

Davide Mariotti
Stefano Cortellessa
Luca Grillo

Indice

1. Introduction:

The aim of this project is to do a deep analysis on a dataset in order to achieve a scope. In our case we have chosen the "Facebook Comment Volume" dataset, a set containing a lot of useful information regarding a Facebook post (like number of comments, number of shares, etc.).
The aim of our analysis is to answer at the following question: *May the length of a post influence its reading and consequently also the number of comments and shares it will receive?*
For answering to this problem, we walked through two different preliminary phases:

- Data Cleaning of the dataset,
- Exploratory analysis.

After these we performed two Machine Learning technics named *Supervised Learning* and *Unsupervised Learning* to predict which will be the shares for a post in the future.

First of all, we considered a significant number of datasets, each of which very interesting as in terms of topics covered as for their attribute's configuration. We gradually discarded some of them, keeping all sets that at the same time had a better structure and aroused our interest.
In the end we opted for "Facebook Comment Volume".

The most important reason we have chosen the Facebook dataset was the argument factor; we were looking for a set in which we could apply all concepts learned during the course, but at the same time the dataset had to capture our attention and it had to be stimulant for subsequent studies and analysis. A second criterion was dimension; we were aiming for a sufficiently large dataset, both regarding entries and attributes. In fact, working with a large number of information, we are able to produce more consistent results for a large number of studies, like explorative analysis, supervised learning and unsupervised learning. A third parameter for the choice was the structure; the dataset on which we desired to apply our work had to contain significant values, namely attributes able to adequately describe the application reality of interest.

Some important domain specific concepts are discussed below:

- ***Public Group/Facebook Page***: It is a public profile specifically created for businesses, brands, celebrities etc.
- ***Post/Feed:*** These are basically the individual stories published on page by administrators of page.
- ***Comment:*** It is an important activity in social sites, that gives potential to become a discussion forum and it is only one measure of popularity/interest towards post is to which extent readers are inspired to leave comments on document/post.
- ***Share:*** It is another important activity in social sites, that allow people to share posts so that other people can read and comment on these posts.

2. Dataset Description:

In the following table the meaning of each column is explained:

Column Name	Column Description
Page Popularity/likes	It is a feature that defines users support for specific comments, pictures, wall posts, statuses, or pages.
Page Checkin's	Describes how many individuals so far visited this place. This feature is only associated with the places eg: some institution, place, theatre etc.
Page Talking About	This is the actual count of users who are 'engaged' and interacting with that Facebook Page. The users who actually come back to the page, after liking the page. This include activities such as comments, likes to a post, shares by visitors to the page.
Page Category	This defined the category of source of document eg: Local business or place, brand or product, company or institution, artist, band, entertainment, community etc. The category is defined by an integer number.
C1	Total comment count before selected base date/time.
C2	Comment count in last 24 hrs w.r.t to selected base date/time.
C3	Comment count is last 48 hrs to last 24 hrs w.r.t to base date/time.
C4	Comment count in first 24 hrs after publishing the document, but before the selected base date/time.
C5	The difference between C2 and C3.
min C1	Define the min of the variable C1 grouped by pages.
max C1	Define the max of the variable C1 grouped by pages.
mean C1	Define the Mean of the variable C1 grouped by pages.
median C1	Define the Median of the variable C1 grouped by pages.
standard_deviation C1	Define the Standard deviation of the variable C1 grouped by pages.
min C2	Define the min of the variable C2 grouped by pages.
max C2	Define the Max of the variable C2 grouped by pages.
mean C2	Define the Mean of the variable C2 grouped by pages.
median C2	Define the Median of the variable C2 grouped by pages.
standard_deviation C2	Define the Standard deviation of the variable C2 grouped by pages.
min C3	Define the min of the variable C3 grouped by pages.
max C3	Define the max of the variable C3 grouped by pages.
mean C3	Define the Mean of the variable C3 grouped by pages.

median C3	Define the median of the variable C3 grouped by pages.
standard_deviation C3	Define the Standard deviation of the variable C3 grouped by pages.
min C4	Define the min of the variable C4 grouped by pages.
max C4	Define the max of the variable C4 grouped by pages.
mean C4	Define the Mean of the variable C4 grouped by pages.
median C4	Define the Median of the variable C4 grouped by pages.
standard_deviation C4	Define the Standard deviation of the variable C4 grouped by pages.
min C5	Define the min of the variable C5 grouped by pages.
max C5	Define the Max of the variable C5 grouped by pages.
mean C5	Define the Mean of the variable C5 grouped by pages.
median C5	Define the median of the variable C5 grouped by pages.
standard_deviation C5	Define the standard deviation of the variable C5 grouped by pages.
Base Time	Selected time in order to simulate the scenario. Decimal (0-71) Encoding.
Post length	Character count in the post.
Post Share Count	These features count the number of shares of the post, that how many peoples had shared this post on to their timeline.
Post Promotion Status	To reach more people with posts in News Feed, individual promote their post and these features tells that whether the post is promoted (1) or not (0).
H Local	This describes the H hrs, for which we have the target variable/ comments received.
Monday	Indicates if the post was posted on Monday (0-1).
Tuesday	Indicates if the post was posted on Tuesday (0-1).
Wednesday	Indicates if the post was posted on Wednesday (0-1).
Thursday	Indicates if the post was posted on Thursday (0-1).
Friday	Indicates if the post was posted on Friday (0-1).
Saturday	Indicates if the post was posted on Saturday (0-1).
Sunday	Indicates if the post was posted on Sunday (0-1).
Monday_Base_Time	Indicates the day on which the post was published on selected base date/time.

Tuesday_Base_Time	Indicates the day on which the post was published on selected base date/time.
Wednesday_Base_Time	Indicates the day on which the post was published on selected base date/time.
Thursday_Base_Time	Indicates the day on which the post was published on selected base date/time.
Friday_Base_Time	Indicates the day on which the post was published on selected base date/time.
Saturday_Base_Time	Indicates the day on which the post was published on selected base date/time.
Sunday_Base_Time	Indicates the day on which the post was published on selected base date/time.
Target Variable	The number of comments in next H hrs (H is the variable 'H Local').

The dataset can be found at this link:
http://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset.

3. Data Cleaning

Data Cleaning is one of the most important part of the project. It's a process able of guaranteeing, with a certain level of reliability, the correctness of a large quantity of data. So, it's responsible of a correct analysis since its aim is to clean the dataset from the values that doesn't make sense, to check if they are correct with respect to the 'logic' of the dataset and to check if they are different from *null.* Furthermore, in this phase all duplicates and empty columns are deleted.

In this phase, the following operation have been performed:

- All the columns have been checked to be sure that there *are not null values* in the dataset,
- All the *values have been checked to be sure that they are all positive*, specifically Page Checkin's, Likes, Page Talking About, C1, C2, C3, C4, Base Time, Post Length, Post Share Count columns have been controlled,
- All the *values in the days columns have been checked to be 0 or 1*,
- All the *values in Base Data Time have been checked to be valid*,
- All *values inside C1 have been checked to be correct*. In particular that:
    - C1 >= C2 + C3
    - C4 <= C1
    - C5 = C2 - C3
- All *useless columns have been removed* (Post Promotion Status) and
- All *duplicated rows have been removed*.

4. Exploratory Analysis

Exploratory data analysis is an approach for analysing datasets to summarize their main characteristics, often with visual methods. It refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.
It is a good practice to understand the data first and try to gather as many insights from it.

Before starting to perform statistics on our principal goal described in section 1, we did some preliminary tests to better understand the dataset. Thus, the following operation have been performed:

- Statistics about Post length variable.

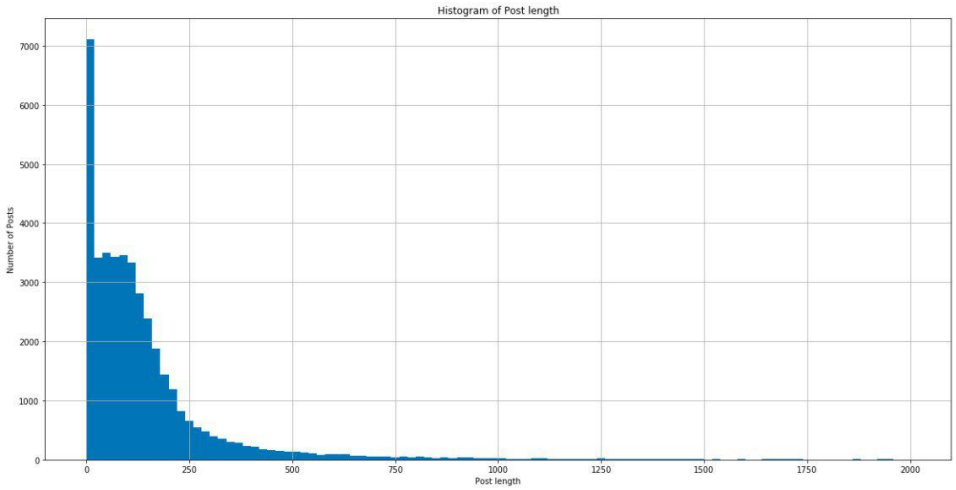


Figure 1: Histogram of Post Length

What can be seen from figure 1 is that the post length is concentrated between 0 and 200. It Is easily visible that 7000 posts have length 0. This kind of post are photos, links or videos without any description.

- Statistic about number of posts that published every day.

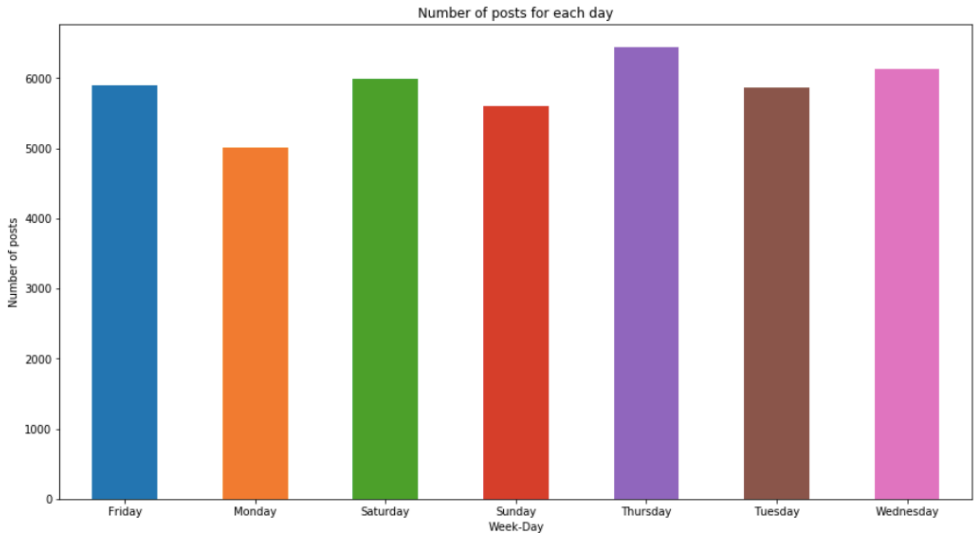


Figure 2: Number of posts for each day

Around 5000 are published every day as we can see from the histogram above.
It is also easy to see that Thursday is the day with the highest number of publication with around 6000 posts.

- Statistics about number of comments (C2, C3 and C4) of all the dataset.

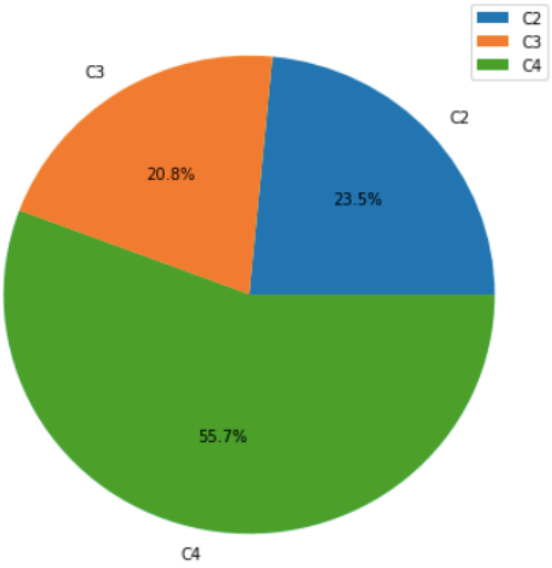


Figure 3: Pie chart about number of comments of all the dataset

What can be seen from figure 3 is that most of the comments are made during the period C4, so between 48 and 72 hours from the BaseDateTime.

- Statistics about the trend of comments in the 4 ranges, respectively C1, C2, C3 and C4.

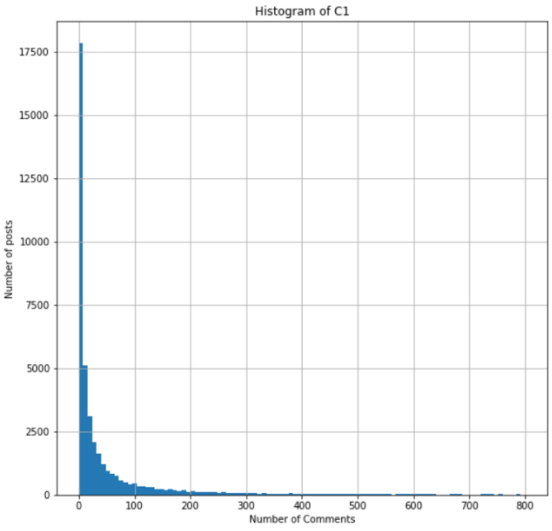


Figure 4: Histogram of C1

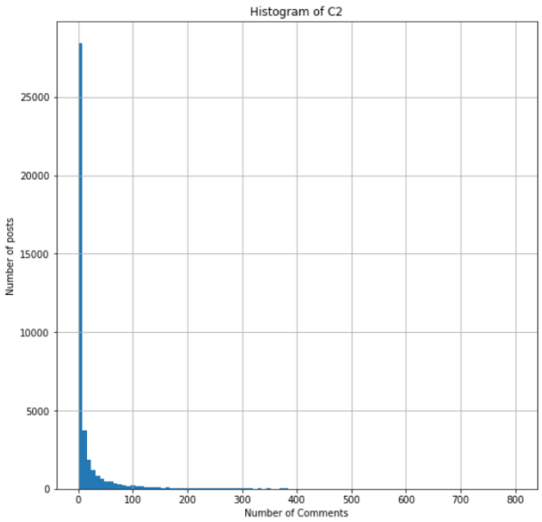


Figure 5: Histogram of C2

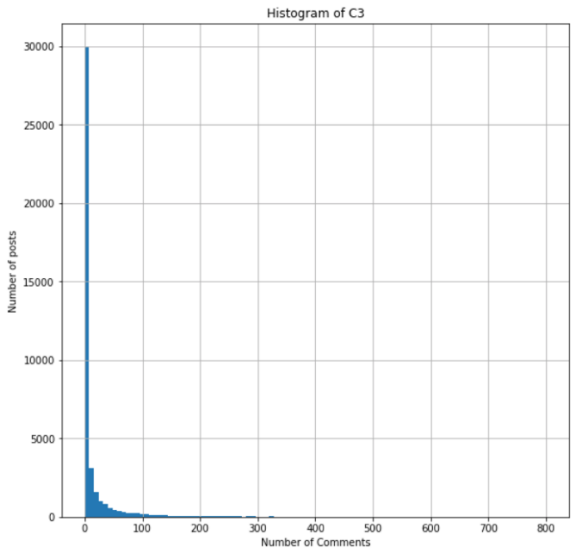


Figure 6: Histogram of C3

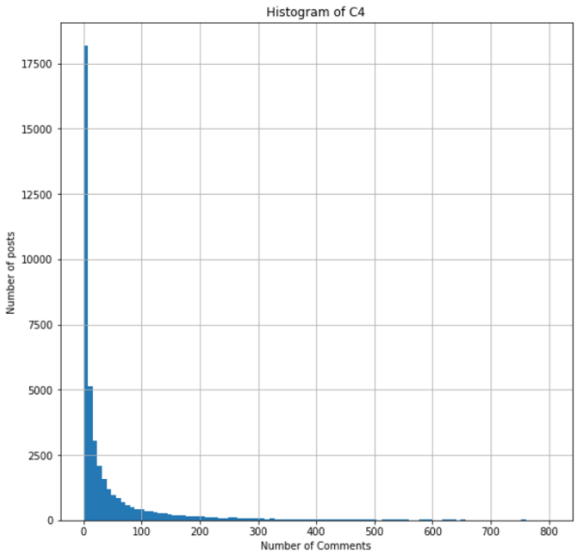


Figure 7: Histogram of C4

- Statistics about the variable 'Page Talking About '.

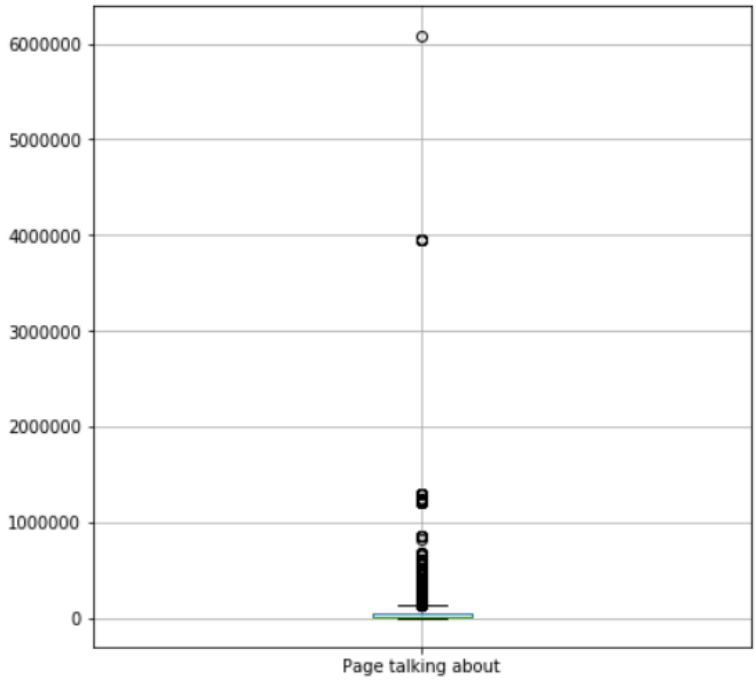


Figure 8: Boxplot of Page Talking About

- Statistics about the variable 'Post Share Count'.

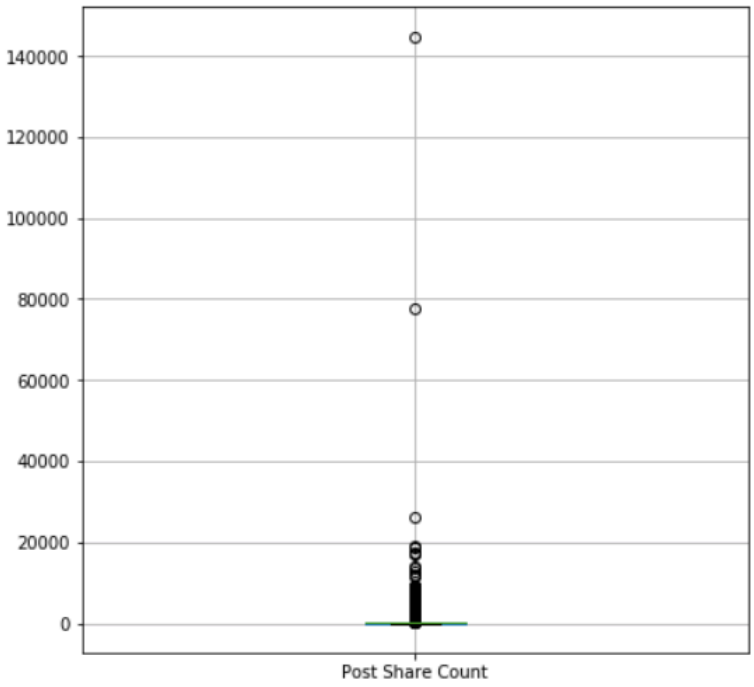


Figure 9: Boxplot of Post Share Count

- Statistics about likes for all pages.

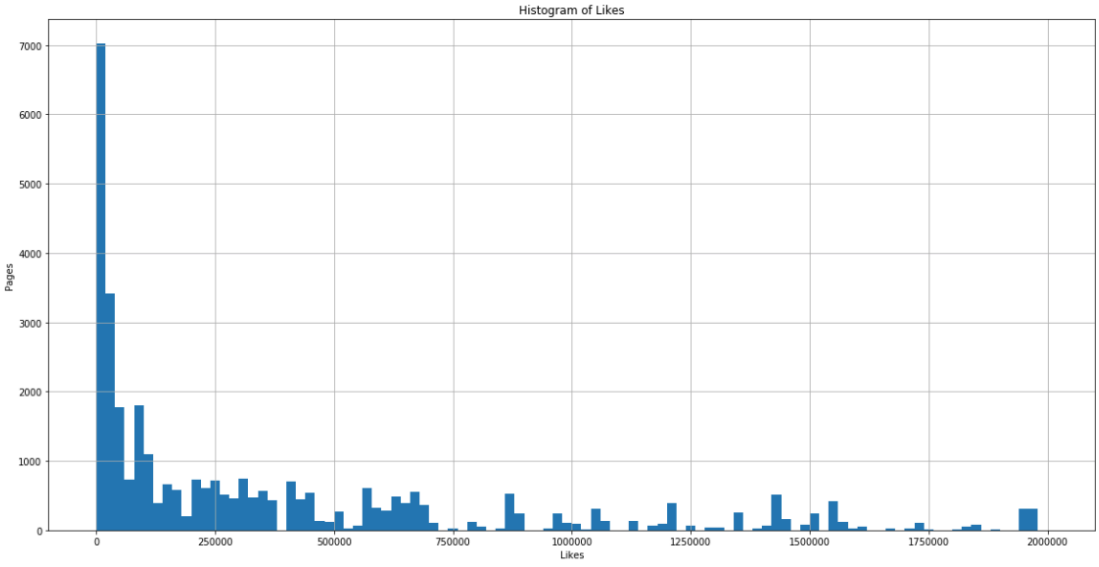


Figure 10: Histogram about Likes

The previous histogram shows that around 7000 posts have a very small number of likes. Thus, decreasing the range between 0 and 500, we figured out the exact number of likes. Moreover, we checked that there were no pages with 0 likes, since they would not be useful for our future research. A deeper analysis is showed below.

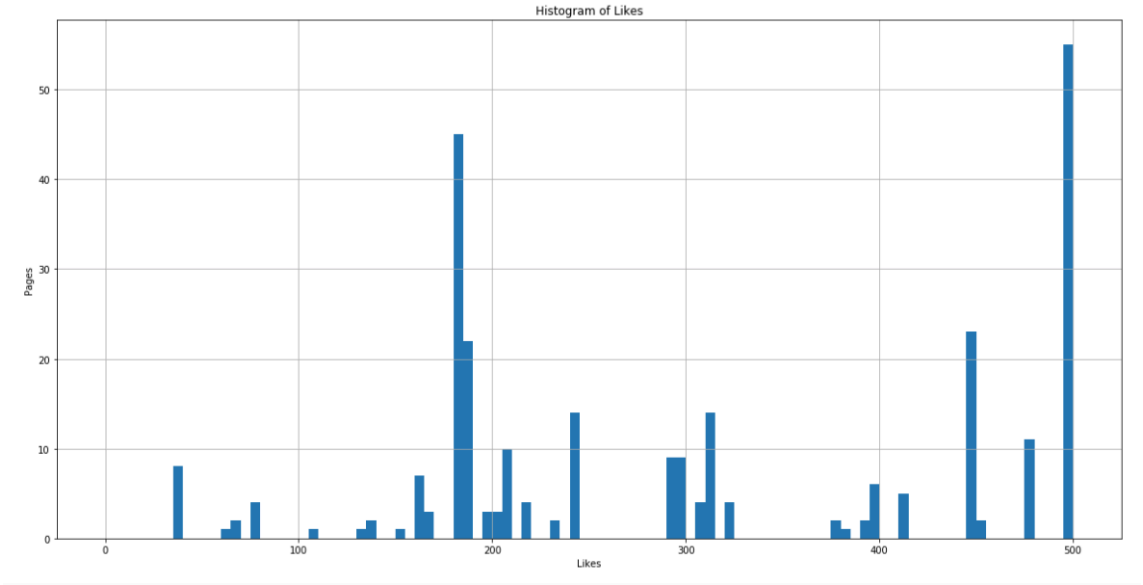


Figure 11: More detailed visualization about Likes

- Statistics about post category.

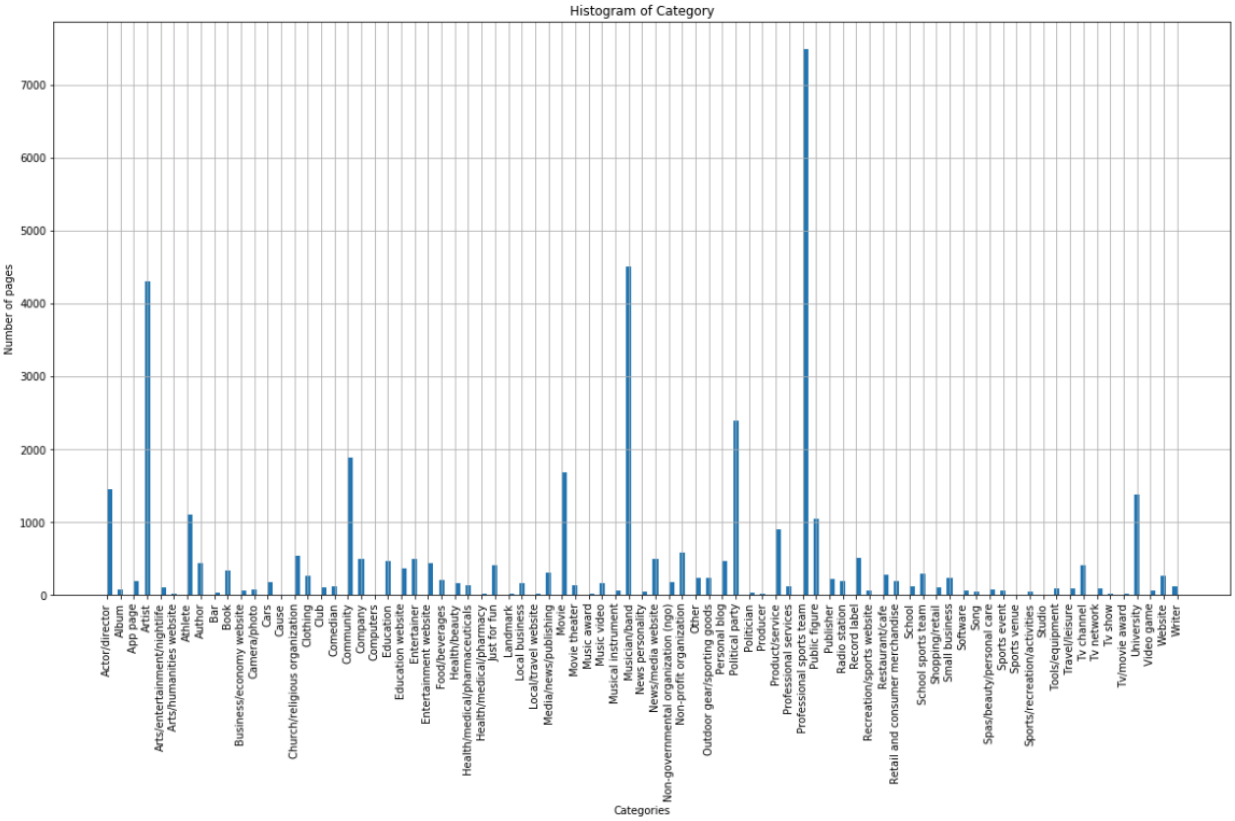


Figure 12: Histogram about categories

As can be seen from figure 12, ***Professional Sports Team*** is the most recurrent category.

Instead, from this point has been started the analysis regarding the principal goal we want to achieve. All statistics from now are made to answer to the following question (already explained in section 1).

"May the length of a post influence its reading and consequently also the number of comments and shares it will receive?"

To answer, initially the posts have been divided between short and long. Looking at several Facebook posts, short posts are those with a maximum of 400 characters. As a result, long posts will be those with at least 400 characters.

- Statistics about posts that have been posted in the first 24h, between 24h and 48h and between 48h and 72h (These hours are in reference to the base time).

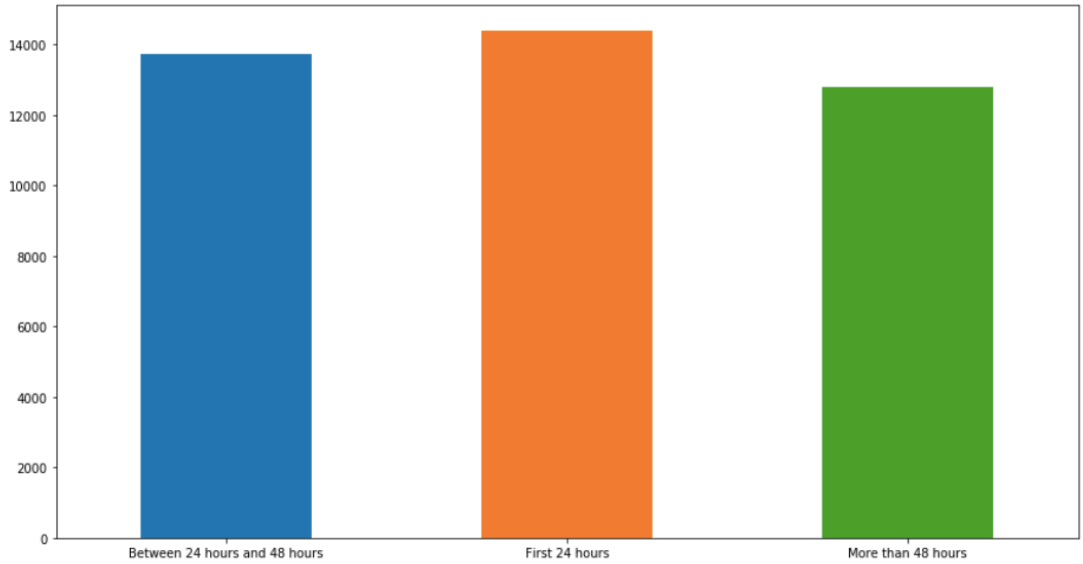


Figure 13: Histogram about posts divided by range

- Number of long and short posts.

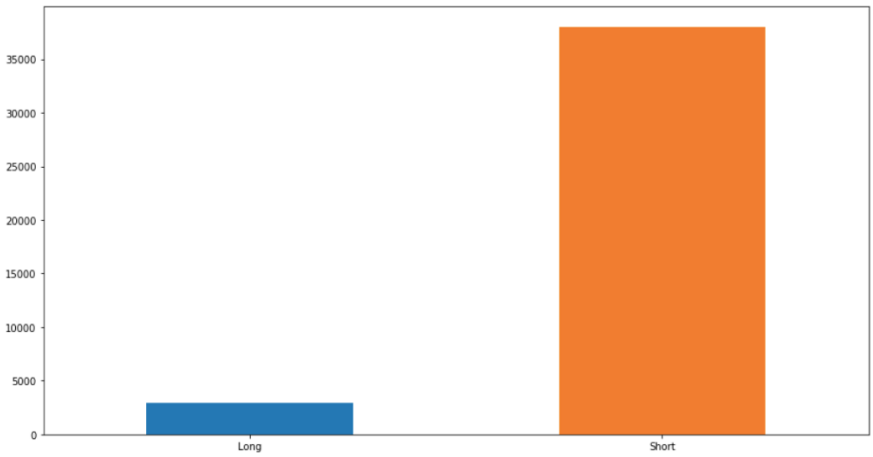


Figure 14: Histogram about number of long and short posts

- Number of shares about long and short posts.

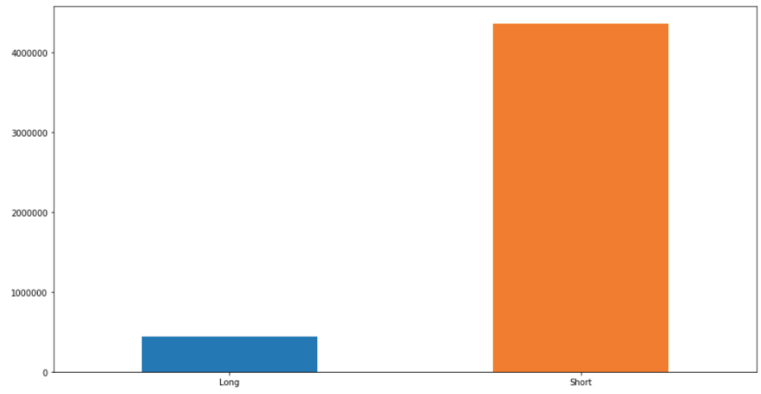


Figure 15: Histogram about number of shares divided for long and short posts

As it is easy to see, the ***short posts achieve a significantly higher number of shares***.

- Statistics about how many posts have been shared in the first 24h (divided between short and long).

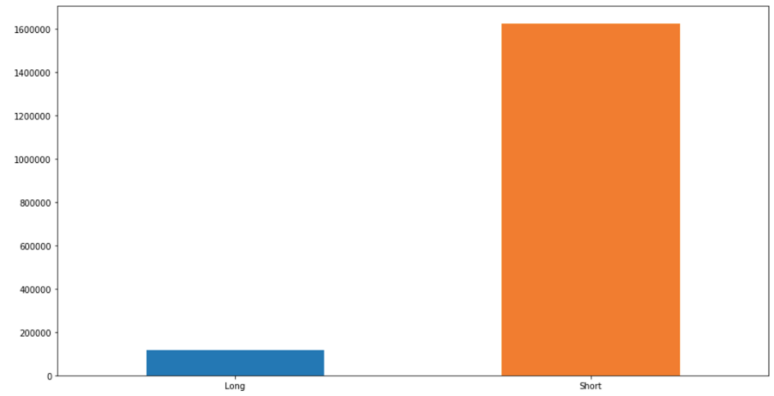


Figure 16: Histogram about number of shares divided for long and short posts in the first 24h

- Statistics about how many posts have been shared between 24h and 48h (divided between short and long).

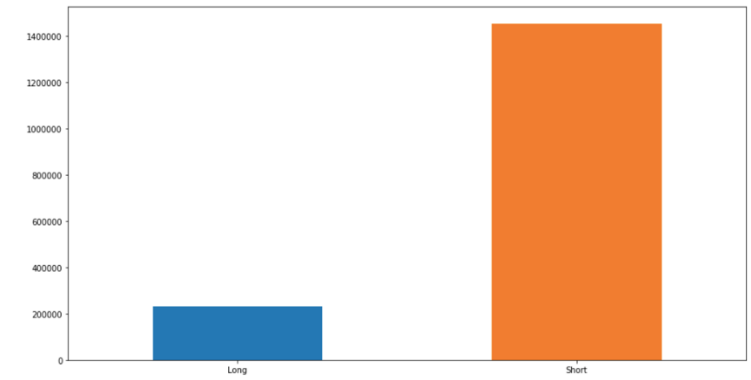


Figure 17: Histogram about number of shares divided for long and short posts between 24h & 48h

- How many posts have been shared between 48h and 72h (divided between short and long).

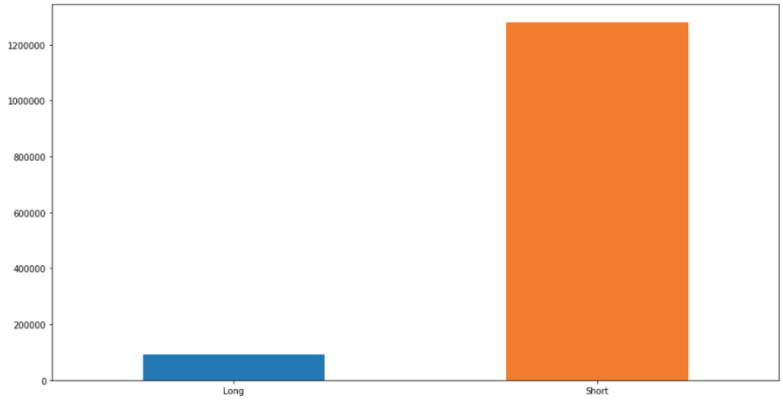


Figure 18: Histogram about number of shares divided for long and short posts between 48h & 72h

Even here, it is easy to see that ***short posts are those with more shares***.

- Statistics about the number of comments that the posts have been received in the first 24h (divided between short and long).

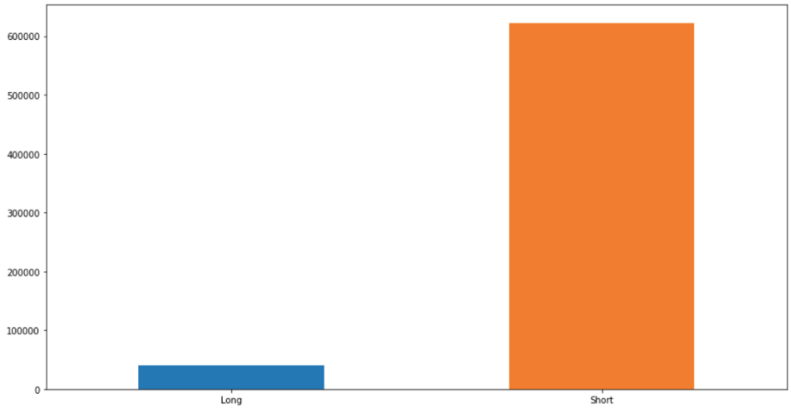
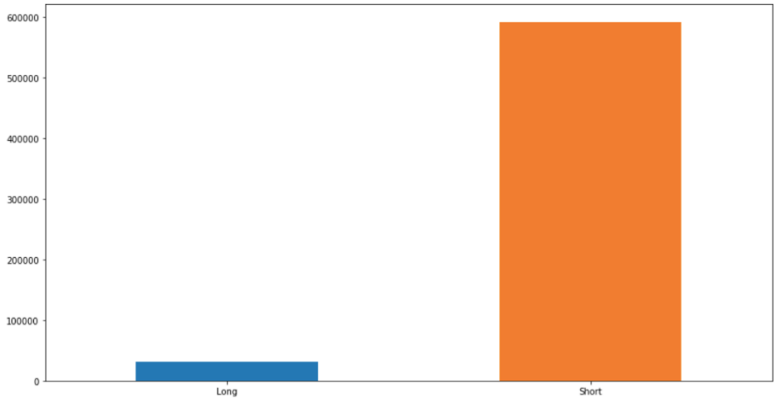


Figure 19: Histogram about number of comments divided for long and short posts in the first 24h

- Statistics about the number of comments that the posts have been received between 24h and 48h (divided between short and long).



- Statistics about the number of comments that the posts have been received between 48h and 72h (divided between short and long).

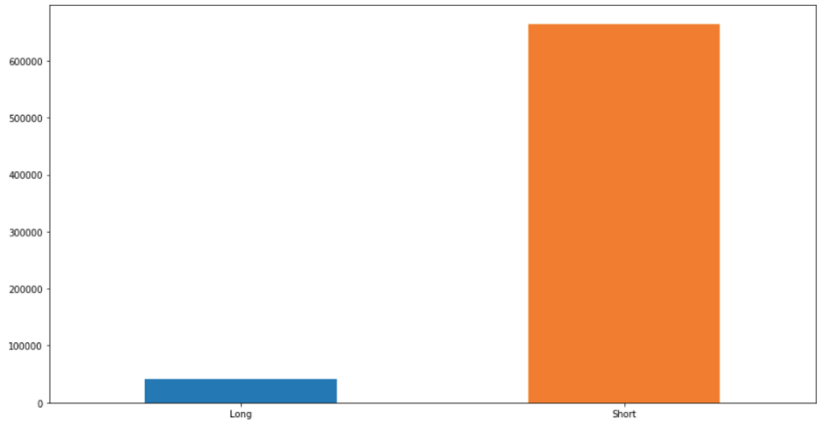


Figure 21: Histogram about number of comments divided for long and short posts between 48h & 72h

As can be seen from all the study performed above, the length of a post might influence its shares and the number of comments. So:

- It is figured out a possible 'good' post length.

We ordered in descending order the data frame based on the number of shares and we took the first 100 rows. Thereafter, we ordered again in descending order the resulting rows based on the number of all the comments and we took the first 50 rows. By averaging the length of the post between these first 50 rows, *we discovered that the 'good' post length is around **107***.

5. Unsupervised Learning

Unsupervised Learning is a machine learning technique to find patterns in data. The data given to unsupervised algorithm are not labelled, which means only the input variables (X) are given with no corresponding output variables. In this technic, the algorithms are left 'to themselves' to discover interesting structures in the data.

Since the values in the dataset are all numerical a ***K-means analysis*** can be performed. The only exception concerns the category column, being a categorical variable. Thus, the standard k-means algorithm is not directly applicable to categorical data for various reasons. The sample space for categorical data is discrete and doesn't have a natural origin. For instance, a Euclidean distance function on such a space is not really meaningful. K-means analysis does not support missing values in the data. As we checked in the beginning of the work, the dataset does not have any null value, so the K-Means analysis can be performed.

On cluster analysis algorithms all depend on the concept of measuring the distance (or some other measure of similarity) between the different observations that we are trying to cluster. If one of the variables is measured on a much larger scale than the other variables, then whatever

measure we use will be overly influenced by that variable. For this reason, to standardize variables has been decided.

First of all, the dataset has been split in training and test sets. The first consists of 70% of the observations and the second consists of the other 30% of the observations.

When using k-means clustering, users need some way to determine whether they are using the right number of clusters. One method to validate the number of clusters is the elbow method. The idea of the ***elbow method*** is to run k-means clustering on the dataset for a range of values of k (k from 1 to 10 in our case), and for each value of k calculate the ***Sum of Squared Errors*** (SSE). Then, plot a line chart of the SSE for each value of k. If the line chart looks like an arm, then the "elbow" on the arm is the value of k that is the best.

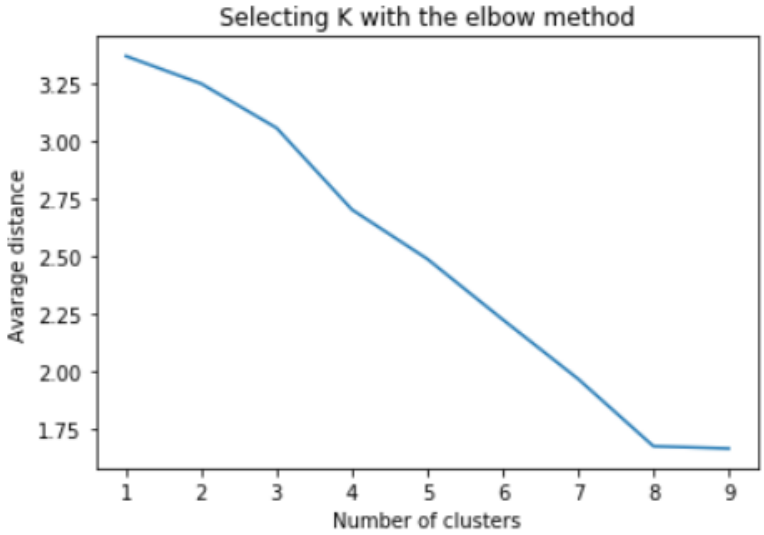


Figure 22: Elbow method

From the plot above can be seen that the average distance decreases when the number of cluster increases and that a good number of cluster is equal to 8. For this reason, the ***8-cluster solution*** has been interpreted.

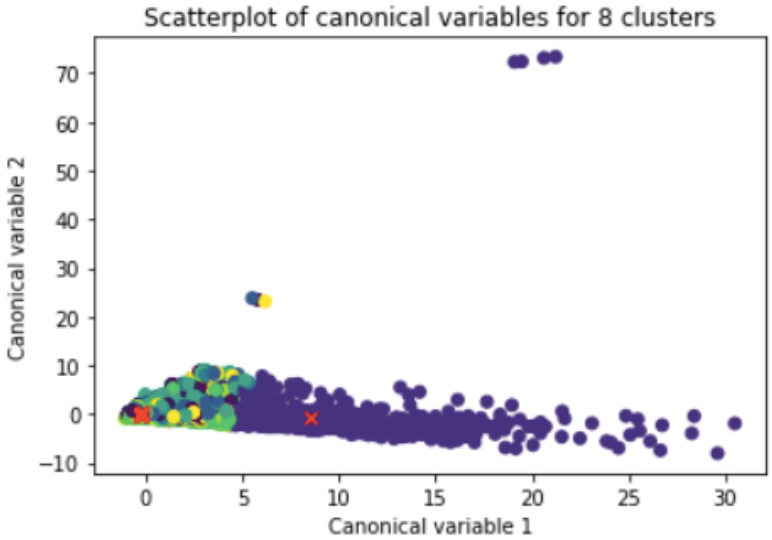


Figure 23: 8-cluster solution

The scatterplot above shows that the seven clusters excluding the one more spread out are dense pack meaning that the observation within the clusters are pretty highly correlated between each other and within cluster's variance is relative low. This overlap means that there is not good separation between these seven clusters. On the other hand, the cluster with center at position around (8, -0.7) shows a better separation and the observations are more spread out indicating less correlation among the observations and high within cluster's variance.

6. Supervised Learning

Supervised learning is a machine learning task, specifically it is a function that maps an input to an output based on example input-output pairs. It infers a function from labelled training data consisting of a set of training example. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value. A supervised learning algorithm analyses the training data and produces an inferred function, which can be used for mapping new examples. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

In order to carry out the Supervised Learning task on the Dataset, among all possible Machine Learning algorithms we have chosen to use the ***Decision Tree Learning*** and ***Neural Network*** Learning as predictive model.
The aim in this section is to predict the target variable 'Class' that represents a specific category of ***the number of post shares***.

First of all, the number of shares of a post has been divided in 4 different categories that indicate respectively:

- A low number of shares that indicates a number between 1 and 49 and it is identified by the number '1'
- A medium number of shares that indicates a number between 50 and 199 and it is identified by the number '2'
- A good number of shares that indicates a number between 200 and 699 and it is identified by the number '3'
- A high number of shares that indicates a number grater then 700 and it is identified by the number '4'

6.1 Decision Trees

Decision trees are a powerful prediction method and extremely popular. It works for both continuous as well as categorical output variables.

While implementing the decision tree the following two phases have been performed:

1. Building Phase
    - Pre-process the dataset.
    - Split the dataset from train and test using Python sklearn package.
    - Train the classifier.
2. Operational Phase
    - Make predictions.
    - Calculate the accuracy.

To *pre-process data*, the dataset has been divided into attributes and labels and will then divide the resultant data into both training and test sets. By doing this, the algorithm can be trained on one set of data and then tested out on a completely different set of data that hasn't seen yet. This provides you with a more accurate view of how your trained algorithm will actually perform.

Since when the data are comprised of attributes with varying scales, many machine learning algorithms can benefit from rescaling the attributes to all have the same scale.
Often this is referred to as normalization and attributes are often rescaled into the range between 0 and 1. It is useful for algorithms that weight inputs like regression and neural networks and algorithms that use distance measures like *K-Nearest Neighbours*.

Successively, *gini index* and *information gain* have been used. Both of these methods are used to select from n attributes of the dataset which attribute would be placed at the root node or the internal node. After this, the algorithm has been trained.

Finally, the number of shares has been predicted and the ***accuracy score*** has been calculated to be really next to ***78%.***

6.2 Neural Network

Neural networks are computing systems vaguely inspired by the biological neural networks that constitute animal brains. The neural network itself is not an algorithm, but rather a framework for many different machine learning algorithms to work together and process complex data inputs. Such systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules.
A neural network is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in the brain, can transmit a signal from one artificial neuron to another. An artificial neuron that receives a signal can process it and then signal additional artificial neurons connected to it.

The technic that has been used is **Multi-layer Perceptron (MLP)** that is a supervised learning algorithm that learns a function by training on a dataset. Given a set of features $X = x_1, x_2, ..., x_m$ and a target Y, it can learn a non-linear function approximator for either classification or regression.

First of all, the input set X and the output set Y of variables have been created encoding the latter with one-hot encoding technic to achieve a better result.
Train and test sets have been created splitting the dataset respectively in 90%-10%.
Finally, the network has been trained on the training set and the Y set has been predicted.
As a result, the number of shares (Y) predicted have an ***accuracy score*** calculated to be really next to ***76%.***

7. Conclusions

Working on the dataset, the goal we set ourselves, that is to see how much influence a short post (with a number of characters less than 400) compared to a long one (with a number of characters more than 400) has, has been achieved by our analysis thanks to the development of various phases. The first has been the one referred ***to clean the dataset*** (section 2), *eliminating all those columns that were repetitive or empty and checking that all numerical values within it were good and not corrupted*, in order to conduct a correct analysis. Successively has been possible to go through the ***exploratory analysis*** (section 3). This showed us that for every week-day, at most, *a number close to 6000 posts are published*. ***The day when there are more publications is Thursday***, so

exactly in the middle of the week, while ***the day when there are fewer publications is Monday***, with a number close to 5000 posts. Analysis of the comments was carried out, which showed that ***the period in which more comments are made and more likes are placed is between 48 and 72 hours after the publication of a post***. An analysis was also carried out on the number of likes that a post receives on average, from which it was found that more than 7000 posts, out of a total of 40948, receive less than 400. Another analysis that was carried out was about categories. In particular, has been analysed ***the category that shows the most interactions, which was found to be 'Professional Sports Team'***. After analysing all these 'secondary' quantities that are related to what is our main purpose, the final analysis was initiated and completed. To do this, the posts with more than 400 characters were initially separated from those with less than the same number, identifying them respectively as 'Long Posts' and 'Short Posts'. This showed that about 90% of posts have less than 400 characters. *This already mean that a post defined as short is certainly preferred over one defined as long.* Referring to this, it was also possible to note that the former is shared about 90% more. Finally, deepening the study, it was noticed that this is done more in the ***24 hours following the sharing of the post for short posts***, while ***between 24h and 48h for long ones***. In addition to the number of shares, the number of comments received from these two types of posts was also analysed. The results were almost identical, clearly showing that short length posts are:

- preferred by users,
- have more interactions.

For this phase we have also tried to give an ***'ideal' number of characters to obtain the largest possible number of iterations (represented by likes, shares and comments) which was found to be about 107***. Finally, ***Supervised and Unsupervised learning*** were performed. In the first case, two different Machine Learning prediction methodologies have been implemented:

- 'Decision Tree' and
- 'Neural Network',

both with the aim of predicting the number of shares that a post will have by identifying the 4 different classes explained in section 6. In both cases, an ***accuracy level close to 76% has been achieved***. Concerning the Unsupervised learning, we adopted a ***clustering algorithm called K-Means Clustering*** that clustered our data points into a number (8) of mutually exclusive clusters. After determining the right number of clusters through elbow method, we interpreted the 8-cluster solution and plotted them.
The resulting scatterplot shows that the seven clusters excluding the one more spread out are dense pack meaning that the observation within the clusters are pretty highly correlated between each other and within cluster's variance is relative low. This overlap means that there is not good separation between these seven clusters. On the other hand, the cluster with centre at position around (8, -0.7) shows a better separation and the observations are more spread out indicating less correlation among the observations and high within cluster's variance.