eDreams ODIGEO

Federico Cirigliano

CASE STUDY 3 - Airline Popularity on Twitter

Questions and Answers

1. Drive a descriptive analysis of the Tweets dataset that would allow us to:

● Understand and compare the popularity of the different airlines. Could you extract any conclusion?

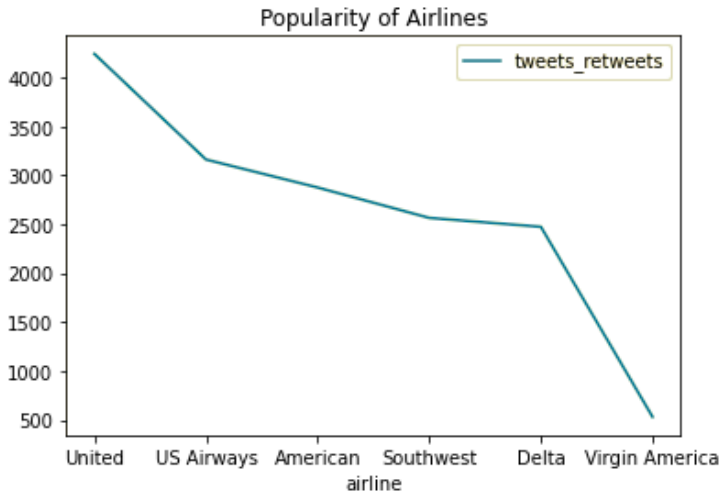

Figure 1. Tweets+ReTweets by Aerolínea.

In order to compare the popularity of the airlines, I created “tweets_retweets” measure. This is the additions of quantity of tweets and quantity of retweets for each airline. As can be seen in the graph in *Figure 1*, three groups of airlines could be identified according popularity. “United" could be placed in the most popular group, one level below there are "US Airways", "America", "Southwest" and "Delta" and lastly "Virgin America". It is also observed that there is a significant difference between the most popular and least popular airline vs. the rest of the airlines.

● Analyze the tweets distribution based on the type of sentiment and the day of the week. Could you extract any conclusion?

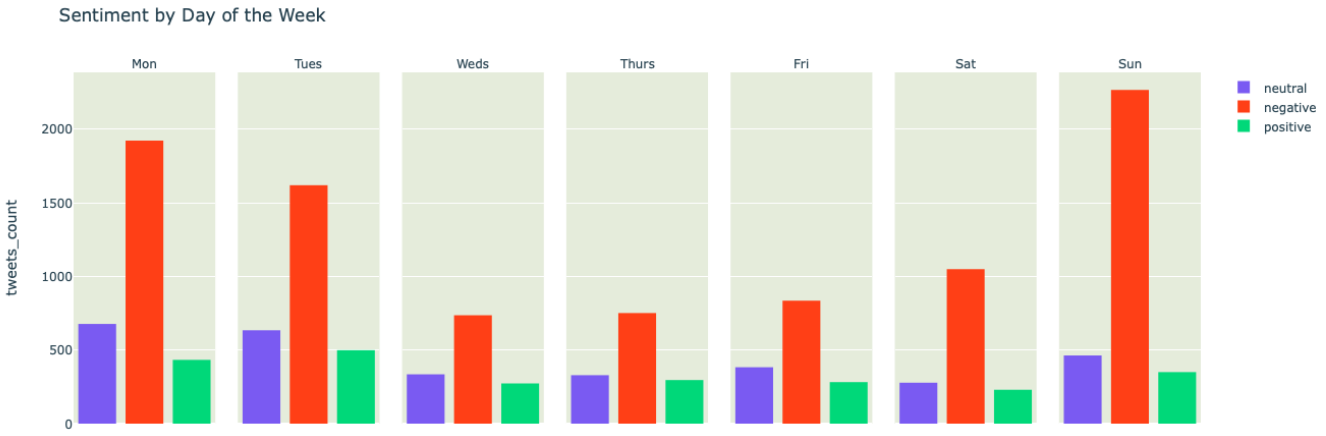

Figure 2. Type of sentiment of tweets by day of the week.

In order to achieve this point, I created “day_of_week” and “day_of_week_name” columns. The first is the number of the day of the week (used for results order) and the second is the name. Both of them were calculated from “tweet_created” column. Then I generated dataframes with the amount and percentage of each type of sentiment for each day.
Analyzing the number and percentage of tweets according to type of sentiment and days of the week, it can be observed that on Sunday, Monday and Tuesday approximately twice as many tweets are written as on the rest of the days of the week.
On the other hand, it was observed that negative tweets predominated every day. Then on a smaller scale neutral tweets and finally positive tweets. Sunday is when the highest percentage of negative tweets is found (73% vs. an average of 60%).

● Analyze the tweets distribution based on the type of sentiment and the hour of the day. Could you extract any conclusion?

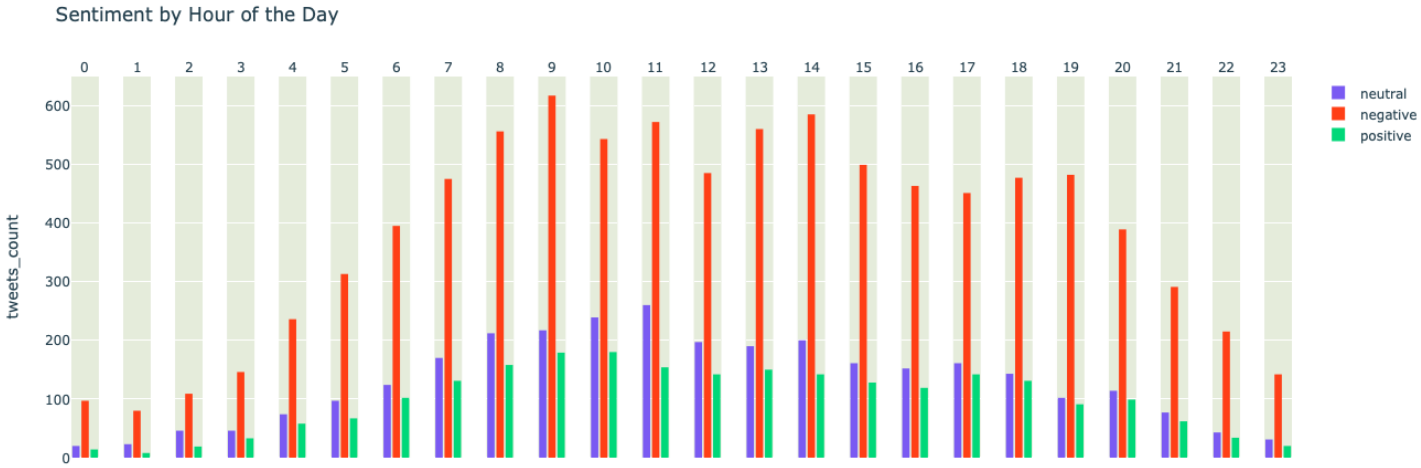

Figure 3. Type of sentiment of tweets by hour of the day.

In order to achieve this point, I created “hour_of_the_day” column. This was calculated from “tweet_created” column. Then I generated dataframes with the amount and percentage of each type of sentiment for hour of the day.
Analyzing the number and percentage of tweets based on sentiment type and time of day, it can be observed that from 8:00 am to 7:00 pm (working hours) 70% of tweets are generated.
On the other hand, the ratio between positive, neutral and negative tweets is similar in each hour of the day.

2. Merge the Tweets dataset with the Airline_Delay one by the date and time of the tweets creation, and analyze if there is any correlation between tweets and cancelled flights.

Clarification: there are 20.517 rows with "CANCELLED" = 1 and 20.059 rows of these has “DEP_TIME” NULL values. So when I merge the Tweets dataset with the Airline_Delay by the date and time of the tweets creation, most of these rows are lost.

In order to merge the datasets, I created “date_time” column in both of them. In *Airline_Delay* dataset “date_time” column is the addition of “YEAR”*100000000+“MONTH”*1000000+DAY_OF_MONTH*10000+“DEP_TIME”. In Tweets dataset “date_time” column is the addition of year*100000000+month*1000000+ day*10000+

hour*100+minute. All of those fields were calculated from “tweet_created” column. Finally I merge the datasets by these columns.
In order to answer this point, I generated a heat map of correletions and then a frequency table only between “airline_sentiment” and “cancelled” columns. In both I could not see a correlation between these variables. To test whether this insight is statistically significant or not, we use the chi-square test of independence.
The *chi-square test* of independence is used to determine whether there is an association between two or more categorical variables. In this case, we would like to test whether the type of sentiment of tweets (“airline_sentiment” column) has any association with the cancelled flights (“cancelled” column).
H^0: The variables are not correlated with each other. This is the H^0 used in the *chi-square test.* P-value is the probability of H^0 being true. If p-value>0.05 then only we accept the assumption (H^0). This means, if two variables are correlated, then the p-value will come very close to zero. In our case, the p-value of *chi-square test* is: 2.716792428089885e-91. As evident, the p-value is less than 0.05, hence we reject the null hypothesis that the type of sentiment of tweets is not associated with cancelled flights.
In the other hand, a *correlation test* is another method to determine the presence and extent of a linear relationship between two quantitative variables. In order to execute it, I created a new numerical variable from the type of sentiment of tweets. If the value is “neutral” then “o”, if “positive” then “1” and if “negative” the “2”.
In this case, we would like to statistically test again whether exists a correlation between the type of sentiment of tweets (now numerical) and cancelled flights. Finally the p-value=7.147340032950288e-49 is smaller than 0.05, so we reject the null hypothesis too that the relationship between the type of sentiment of tweets and cancelled flights is not significant.

Conclusion: in both tests *chi-square* and *linear regression correlation test* the null hypothesis was rejected. So the answer is the variables type of sentiment of tweets and cancelled flights **are correlated**.