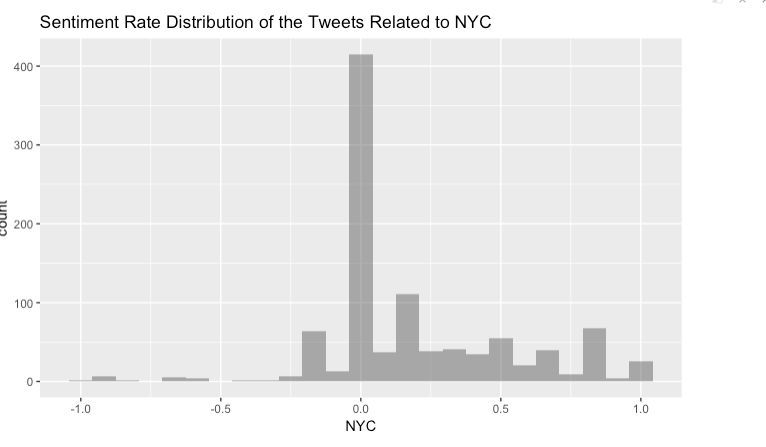
As the top two metropolitans in U.S., New York City (NYC) and Los Angeles (LA) attract a lot of people to visit or settle down. This project is going to study that the sentiment rates of tweets related to NYC and LA. Sentiment rates of tweets represent the emotions of the tweeter users and are defined from -1 to 1. If the sentiment rate is greater than 0, we could consider the tweet as a positive tweet, If the sentiment rate is less than 0, we could consider the tweet as a negative tweet, If the sentiment rate is equal to 0, we could consider the tweet as a neutral tweet.

**Part1 NYC VS LA**

Firstly, let's use t-test to test whether the sentiment rates of New York City tweets are different from that of Los Angeles tweets. The null hypothesis is that there is no difference between their sentiment rate. The p-value of this test is 0.05166. So, there is no enough evidence at the 0.05 significance level to claim that the sentiment rates of New York City tweets are different from that of Los Angeles tweets. What we should pay attention is that the P value is very closed to 0.05.

A screenshot of a cell phone

Description automatically generated

We want to know more about the difference between their sentiment rates. We can create histograms of the sentiment rates. The two histograms do look like similar. Let’s check the details numbers. The total amount of positive tweets related to NYC is 487, The total amount of neutral tweets related to NYC is 409, The total amount of negative tweets related to NYC is 104.The total amount of positive tweets related to LA is 423, The total amount of neutral tweets related to NYC is 498, The total amount of negative tweets related to NYC is 79. We can see LA have less positive and negative tweets than NYC, more neutral tweets than NYC.

Let’s compare the positive sentiment rates of tweets related to NYC and LA. We can see that the minimum rate of NYC is 0.01, which is lower than that of LA’s minimum rate 0.02778. But NYC has the higher mean rate than LA. In addition, they have same median rate. So, we could say the positive emotion of NYC tweets is a little bit higher than LA tweets generally, although the positive emotion of LA tweets starts from a higher point.

A black sign with white text

Description automatically generated

Now Let’s compare the negative sentiment rate of tweets related to NYC and LA. We can see that the number in below table are negative. Since the distance from 0 to the number shows how strong the emotion is, we need compare the absolute value of the below negative numbers. In this case, higher absolute value means stronger negative emotions. We could find that NYC tweets have higher median, higher mean and higher max number. So, we could say the negative emotion of NYC tweets is higher than LA tweets generally.

A close up of a white wall

Description automatically generated

**Part2 Residents VS Tourists**

We can create a binary classification based on the attribute “place” to distinguish residents and tourists of each city.

Below table shows the sentiment rate distribution of NYC residents’ tweets and NYC tourists’ tweets. We can see the rates from residents’ tweets have a less minimum, greater maximum and greater mean. So, we could say the emotion of residents’ tweets have a larger swing.

A close up of a white wall

Description automatically generated

Let’s use t-test to compare the sentiment rates of NYC residents’ tweets and all NYC-related tweets. The null hypothesis is that there is no difference between sentiment rates of NYC tweets and the sentiment rates of NYC residents' tweets. P-value of the test is 0.1057. So, there is no enough evidence at the 0.05 significance level to claim that the sentiment rates of NYC tweets are different the sentiment rates of NYC residents' tweets.

Another t-test compares the sentiment rates of NYC tourists’ tweets and all NYC-related tweets. The null hypothesis is that there is no difference between the sentiment rates of NYC tweets and the sentiment rates of NYC tourists' tweets. P-value of the test is 0.002577. So, this is enough evidence at the 0.05 significance level to claim that the sentiment rates of NYC tweets are different the sentiment rates of NYC tourists' tweets.

We have done the part of data set related to NYC now. We come to the other group. Below table shows the sentiment rate distribution of LA residents’ tweets and tourists’ tweets. We can see the biggest difference are the minimum rates. So we could say the most negative tweet from LA tourists is more negative than that from LA residents.

A close up of a street

Description automatically generated

Use the same tests to test the data set related to LA.

The first null hypothesis is there is no difference between the sentiment rates of all LA-related tweets and only LA residents' tweets. P-value of the test is 0.2736, which means there is no enough evidence at the 0.05 significance level to claim that the sentiment rates of LA tweets are different the sentiment rates of LA residents' tweets.

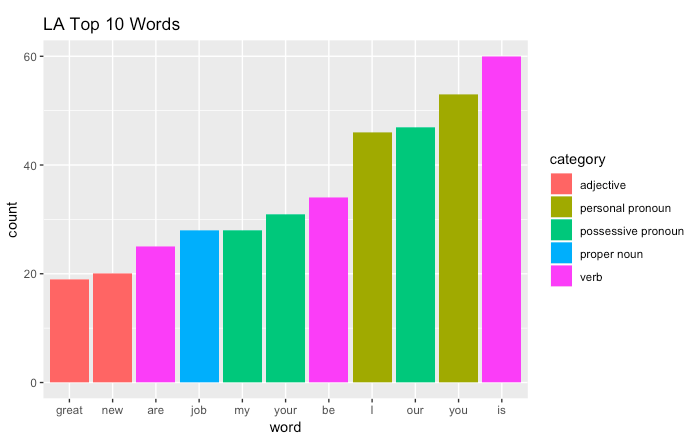
The second null hypothesis is that there is no difference between the sentiment rates of all LA-related tweets and only LA tourists' tweets. The P-value of the test is 0.8627, which means there is no enough evidence at the 0.05 significance level to claim that the sentiment rates of LA tweets are different the sentiment rates of LA tourists' tweets.

In conclusion, the sentiment rates of NYC tweets are closed to the sentiment rates of NYC residents’ tweets. The sentiment rates of LA tweets are closed to the sentiment rates of both subgroups.

**Part3 Most Frequency Words**

At last, let's create a graph for top 10 words of each city. Since we want to find the words more representative, preposition words (e.g. of, in, at), conjunction words (e.g. and) and article words (e.g. the), which could appear many times for building a sentence, were filtered. It should be noticed that different selection rules for key words should be applied depending on purpose.

A screenshot of a cell phone

Description automatically generated

We can find some interesting words in the two Top Ten Words graphs, such as “Mothers”, “Day”, “soul” in NYC tweets, “job”, “new” in LA tweets. Considering the analysis is carried out near 2019 mothers’ day, so we could say the top 10 words are time sensitive. Another interesting thing is that there is no same noun words or adjective words in the NYC tweets and LA tweets. So, we could conclude that there could be culture differences in the two big cities.