Twitter Airline Sentiment Data Analysis

by Sreya Vadlamudi

Data Quality

For this project, the Twitter US Airline Sentiment dataset was analyzed. In this dataset, the variables including were the following: tweet_id, airline_sentiment, airline_sentiment_confidence, negativereason, negativereason_confidence, airline, name, retweet_count, text, tweet_coord, tweet_created, tweet_location, and user_timezone.

Overall Data Quality: The data was very unorganized filled with data that was not necessary or useful for analysis such as tweet_coord. Furthermore, there was a lot of missing data making it more difficult to find a pattern through figures without dropping the data for certain analysis such as for the negativereason column which explained the negative reason as to why someone did not like an airline experience. This quality of data made it more difficult to analyze, especially since there was a significant amount of null data which could changed the outcome of the results if it was listed.

Written information: In regards to the written information such as name, airline name, and text, name was not very significant for any analyzing since knowing the name of those who tweeted will not improve the airline. The most significant written column was the airline name and the text because sentiment score of how positively or negatively a twitter user felt could be found with text. Text was a good column as well as airline because there was not a lot of missing data. Unfortunately, the inconsistency in capitalization in the text column proved to be an issue because it made it harder to code for.

Numerical information: In regards to the numerical information provided in this dataset, they proved to be beneficial since not a lot of data was missing from them such as airline_sentiment_confidence and negativereason_confidence. These were easy to understand as well and interpret. These columns helped understand how the twitter users felt about their tweets and how confident they were in them.

•

Statistics for Figures

Statistics:

- I took the averages of airline_sentiment_confidence and negativereason_confidence just to get an understanding of how confidence people were in what they said on average.

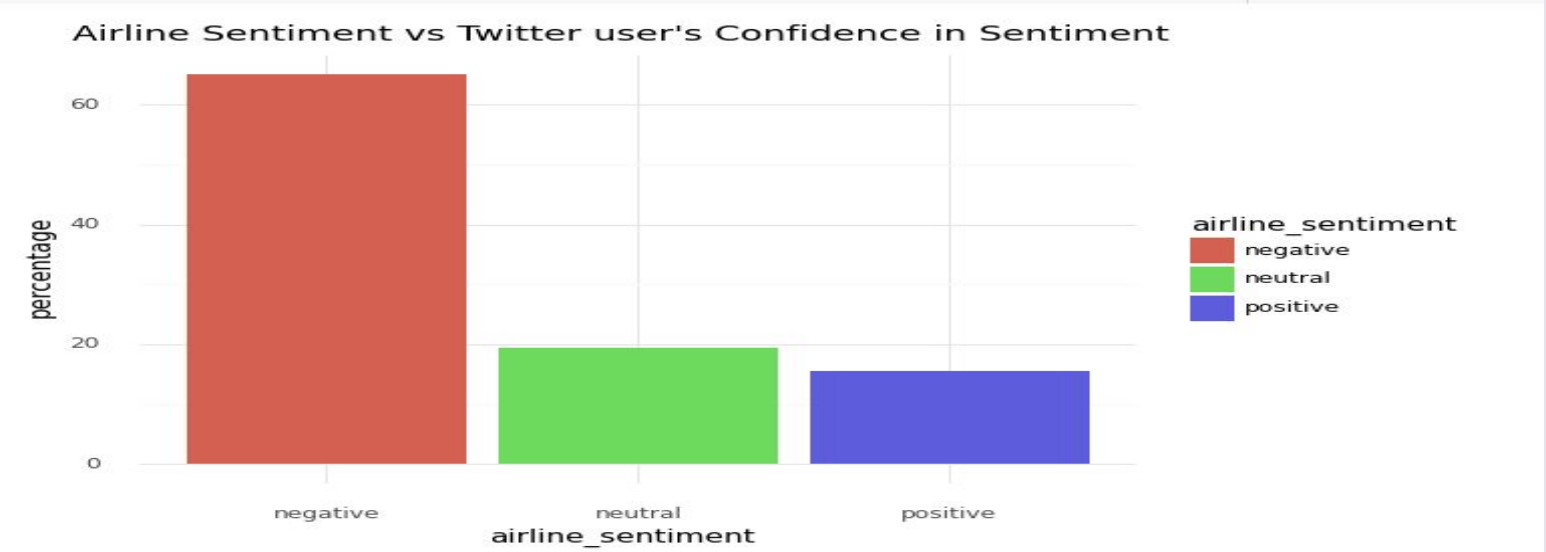
Average airline_sentiment_confidence: .899

The average airline_sentiment_confidence average shows that people were very confidence in stating whether there experience was positive, negative, or neutral which means the figures will be fairly accurate since they were confident in their opinions.

Average negativereason_confidence: .637

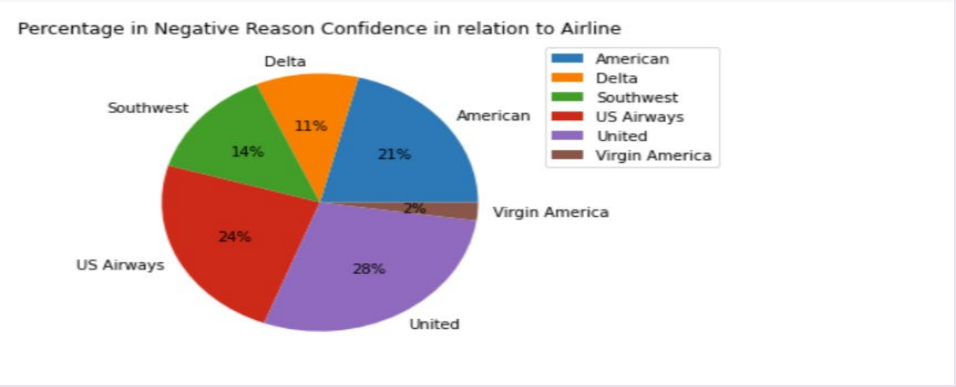
The confidence in the reasoning for the negative reason is a bit lower, but still above 50% so they were fairly confident in why they had a negative experience which once again is good to know for analysis purposes. Their confidence shows that the results I get can be fairly accurate in showing users concerns.

Figure 1



For this figure, based on the statistics from the previous slide, I wanted to look more into positive vs negative vs neutral in regards to people's I took in all the data in order to see how many people chose positive vs neutral vs negative, and how confident they were in what airline sentiment they chose. For the figure, I created a bar graph with airline_sentiment of positive, neutral, and negative on the x-axis and the percentage of confidence in the y-axis. Through this figure, it was obvious just like with the averages that people were much more confident in stating that they had a negative experience with an airline with about 80% or higher. After negative, the second highest was neutral at a little less than 20% confidence, and the last was positive around 15%. With this it was evident most people felt negatively, so the next thing I wanted to do was figure out why in detail.

```
import matplotlib.pyplot as plt
import numpy as np
twitterData.groupby(['airline']).sum().plot(
    kind='pie', y='negativereason_confidence', autopct='%1.0f%%')
plt.title("Percentage in Negative Reason Confidence in relation to Airline")
plt.legend(bbox_to_anchor=(1.05, 1.0), loc='upper left')
plt.tight_layout()
plt.ylabel("")
plt.show()
```




```
[ ]  import matplotlib.pyplot as plt
     import numpy as np

     twitterData.groupby(['negativereason']).sum().plot(
         kind='pie', y='negativereason_confidence', autopct='%1.0f%%')

     plt.legend(bbox_to_anchor=(1.05, 1.0), loc='upper left')
     plt.tight_layout()
     plt.ylabel("")
     plt.title("Percentage in Negative Reason Confidence in relation to Negative Reason")
     plt.show()
```

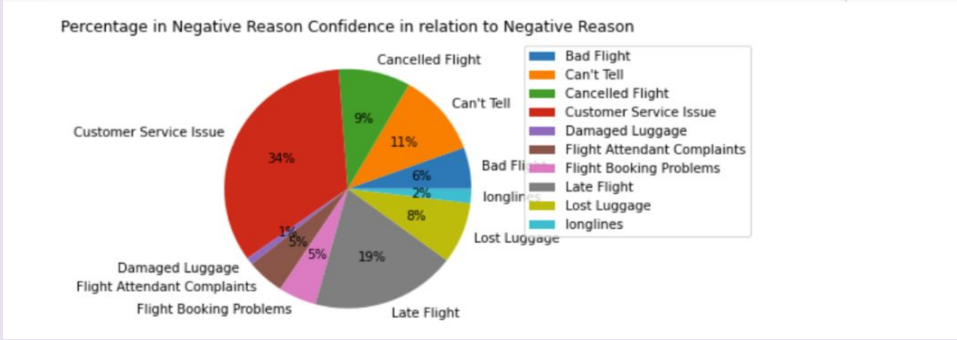


Figure 2 (left)

Figure 3 (right)

Figure 4

```
import matplotlib.pyplot as plt
import numpy as np
ggplot(twitterData, aes(x="airline", y="retweet_count", fill = "user_timezone"))
+ labs(x="Airline", y= "Amount of Retweets")
+ theme_minimal()+ggtitle("Airline vs Retweet Count based on Timezone")
+geom_bar(stat='identity')
```

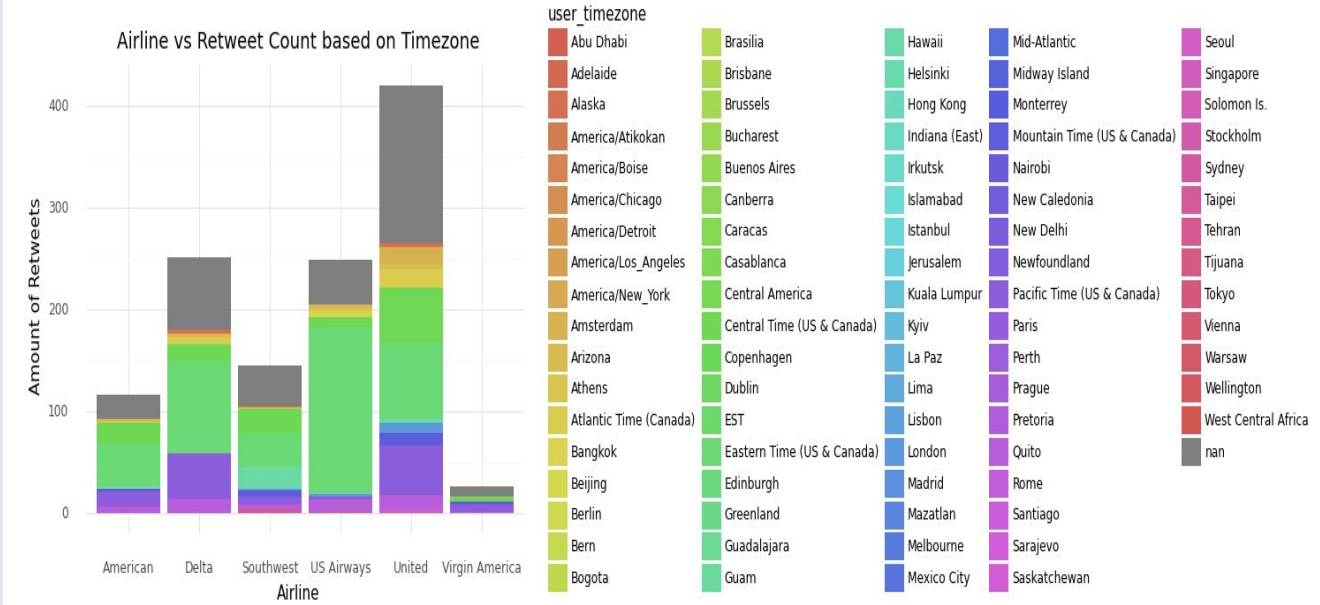
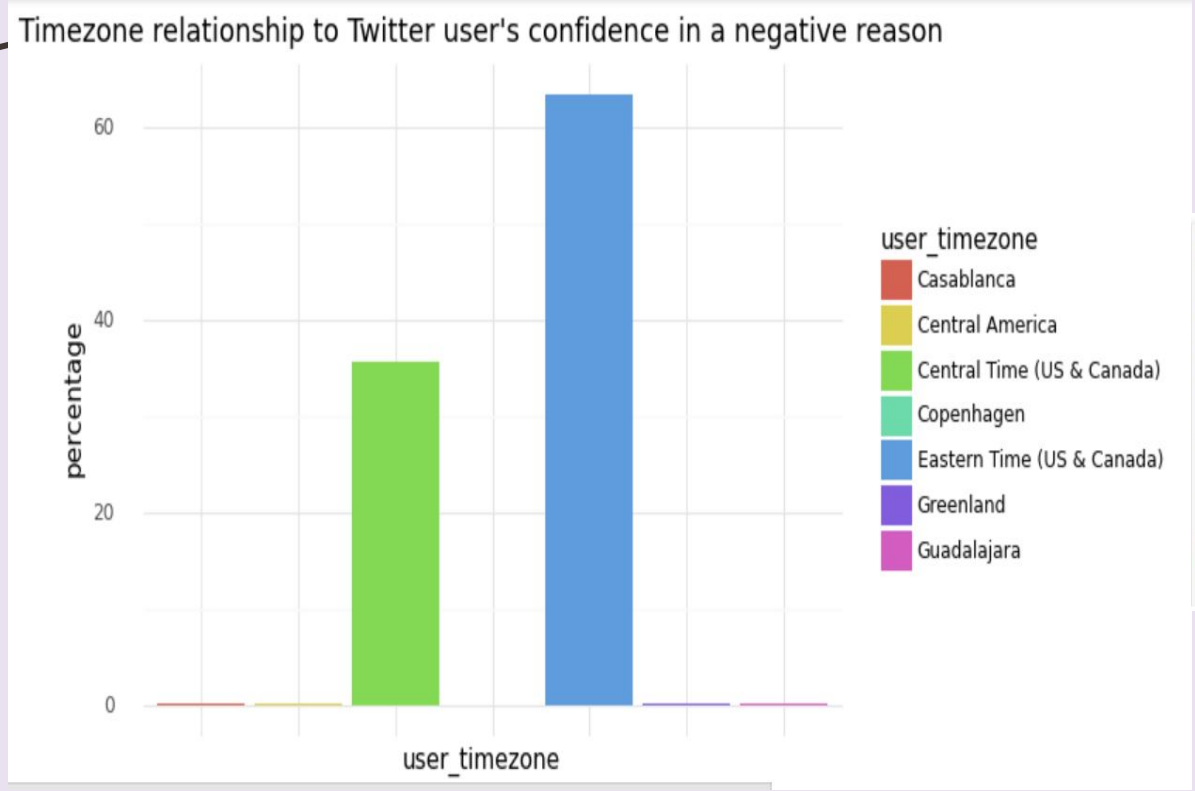


Figure 5



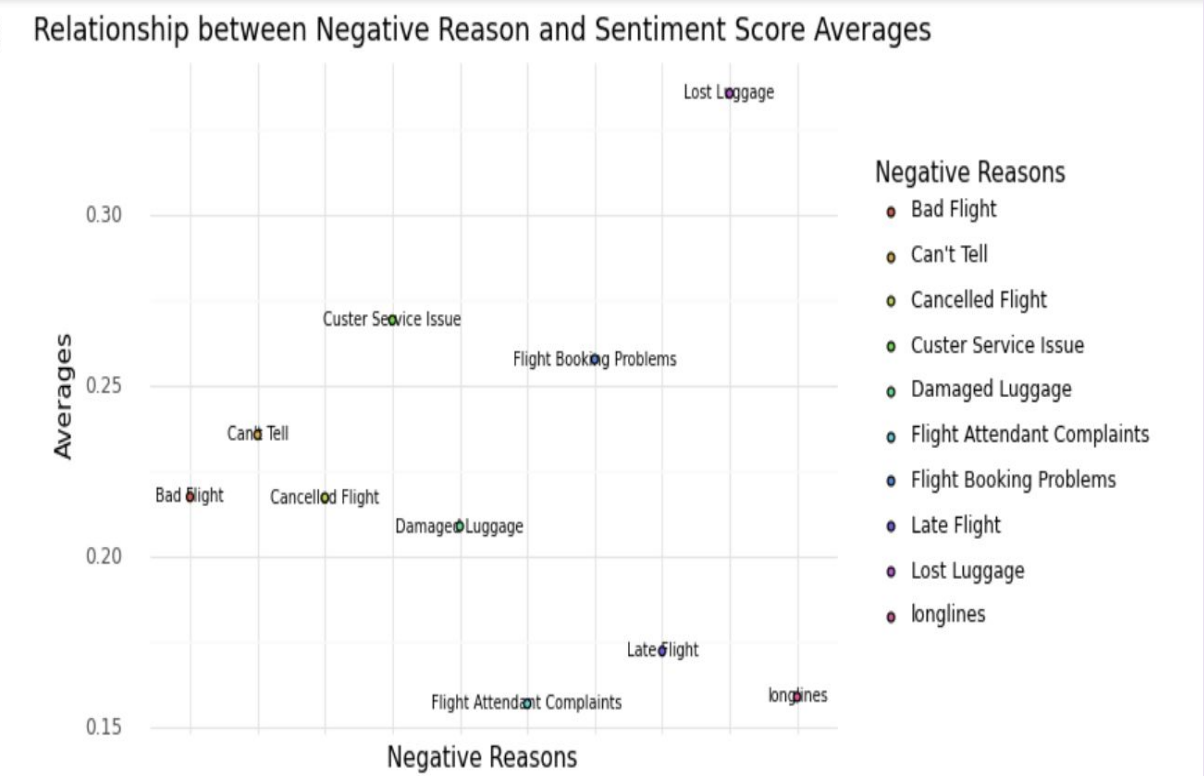

```
#Figure 2
twitterDataMix3 = (twitterData["user_timezone"] == "Casablanca") & (twitterData["airline"]=="United")
 | (twitterData["user_timezone"] == "Central America") & (twitterData["airline"]=="United")
| (twitterData["user_timezone"] == "Central Time (US & Canada)")& (twitterData["airline"]=="United")
| (twitterData["user_timezone"] == "Copenhagen") & (twitterData["airline"]=="United")
| (twitterData["user_timezone"] == "Dublin") & (twitterData["airline"]=="United")
| (twitterData["user_timezone"] == "Eastern Time (US & Canada)")& (twitterData["airline"]=="United")
| (twitterData["user_timezone"] == "Greenland")& (twitterData["airline"]=="United")
| (twitterData["user_timezone"] == "Guadalajara")& (twitterData["airline"]=="United")
twitterDataMix4 = twitterData.loc[twitterDataMix3]

twitterDataMix4["percentage"] = (twitterDataMix4["negativereason_confidence"] /
                twitterDataMix4["negativereason_confidence"].sum()) * 100
print(ggplot(twitterDataMix4, aes(x = "user_timezone", y = "percentage", fill = "user_timezone"))
+ ggtitle("Timezone relationship to Twitter user's confidence in a negative reason")
+ theme_minimal()+theme(axis_text_x=element_blank())+ geom_bar(stat='identity'))
```

Figures 1-5 Explanations

Figures 2 and 3: For these figures I wanted to see which airlines people felt the most negatively about as well as the specific reasons why people felt so negatively about the airline experience overall. To do this, I only had to look at the negative reason confidence and negative reason data. Since some of the negative reason data was missing values, I decided to drop the null values since they would not add any significance to the figures I made. For these I decided to make two pie charts. The first pie chart/Figure 2 measured in percentages how confidently people felt about their negative sentiment towards a certain airline. Through this figure, it was evident that people felt the most negatively towards United with 28% and second worst was US Airways with 24%. In addition, the second pie chart/Figure 3 was similar since it also measured the confidence level in a person's negative sentiment, but this time it measured in percentage how they felt about the negative reasons that affected their experiences. Through this figure, it could be seen that the biggest thing people had issues with were customer service issues and the second biggest problem was flights being late with 34% and 19% consecutively. Overall, people were most confident in the fact that United was the worst airline and customer service problems were the biggest issues.

Figures 4 and 5: For these figures, I wanted to see in more detail the relationship between confidence and the retweet_count, and I kept all the data just to see which airline had the most missing data. I created a bar graph in relation to airlines and retweet_count to see who felt strong enough to retweet something that was stated about certain airlines, and I wanted to see which timezone had the most retweets. Overall, although there was a significant amount of null data where we could not see what timezone a retweet was from, especially for United, United still had the most amount of retweets meaning people felt strongly enough to retweet(over 400 retweets). Even with the missing data, there were a few places that had the most retweets which were Casablanca, Central America, Central Time(US & Canada), Copenhagen, Eastern Time (US & Canada), Greenland, Guadalajara. Based on this information, I created the next figure to look more in detail at which timezone out of the seven above had the most negative connotation in regards to airline experience, specifically United airlines in order to see which time zones United airlines should improve its airline experience for. Through this figure, Twitter users were most confident in their negative sentiment for United in the timezones of Central Time(US & Canada)(35%) and Eastern Time(US & Canada)(Above 60%).

Sentiment Score Figure And Explanation



After seeing that Eastern Time(US & Canada) and Central Time(US & Canada) were the worst for people in regards to using United airlines, I wanted to see what the sentiment scores on the text/tweet specific to United in these time zones would be in comparison to the negative reasons. I decided to take the averages of sentiment scores specific to text that were relevant to United airlines that applied to these two time zones. Once I took the averages, I wanted to see which sentiment scores were more negative for what negative reasons and compare the two. Therefore, I wanted to see the relationship between these sentiment scores and the negative reasons people hated in order to improve United airlines in these time zones. For this I used a scatter plot, and overall, the sentiment scores for the tweets/text were most negative for the negative reasons of Flight Attendant Complaints, Late Flight, and longlines in the specific areas of Central Time and Eastern Time(US & Canada) for United airlines.

Insights(What I've Learned)

Overall from all of these figures, I learned quite a bit about the airlines, the timezones, and the negative reasons people felt were the reason they had a negative airline experience. Primarily, most people felt negatively and were fairly confident in the way the felt in comparison to feeling positively or neutral to an airline. I was surprised that it was found that customer service issues was one of the biggest issues when it comes to the negative reasons in overall airline quality. I did not know this was such a struggle for people at the airport. Furthermore, the worst airline, which was also seen in figures such as the pie charts and had the most retweets showing people had a lot to say about this airline was United airlines. Along with this, in regards to timezones, I learned that specifically to United airlines, which was the one I focused on, Central Time(US & Canada) and Eastern Time(US & Canada) stand out meaning more focus needs to put in there timezones for airline quality. I also saw that the negative reasons for specifically United airlines in these timezones was different from the overall negative reasons for every airline. United emphasized issues in longlines and Late Flights while the overall issue for all airlines was customer service issues.

Conclusions based on Analysis

From these insights and analysis, it would be best for all airlines to work on improving their quality in customer service and working to actually benefit the customers who travel with their airline so they can have a more positive experience. All airlines should also fix the issue of late flights. Since late flights are not controllable sometimes, I recommend keeping people in the loop so they are not blindsided and feel important. This would automatically improve airline quality for them.

Furthermore, through sentiment score averages, other prominent negative reasons that could also be fixed, especially for United airlines, would be Flight Attendant Complaints, longlines, and Late Flights. United airlines should be work on fixing these issues in the Central and Eastern timezones of US and Canada so the tweets are less negative in those areas. In order to do this, I personally would recommend that in the Central and Eastern timezones, those working on flights and in the airports need to spend more time taking care of people's concerns and making sure the people feel understood. I once again would recommend that United airlines review the feedback that they get from the people to change their methods. For example, with late flights, give people other solutions or key ifnromation so they once again do not feel overwhelmed or angry enough to tweet ot retweet. With longlines, people should prepare in advance so airlines should allow them to check in online so that the lines go by more quickly. TSA precheck would also be a solution for longlines.

Overall, fixing these main issues would definitely limit the negativereason_confidence for people and hopefully more would feel confidently about their airline experiences.

Cleaning up Data(Appendix)

Steps taken to clean up data in Excel:

- Get rid of duplicate values that there would be no skewed figures(there were 36 duplicate values in this dataset that were unnecessary)
- Deleted rows that were completely useless in the sense that most of the column was null data with one in every ten cells having data. I also knew I would not use these columns for analysis. The columns that I deleted were airline_sentiment_gold, negativereason_gold, and tweet_cord. Airline_sentiment_gold and negativereason_gold did not have any valuable information that I could use to analyze so I stuck to using airline_sentiment, airline_sentiment_confidence, negativereason, and negativereason_confidence.
- In order to find a sentiment score for specific airlines, I needed to check through code if the text contained the airline name. The inconsistency in capitalization of the airline name in different tweets made checking for it difficult. I used the excel function PROPER to fix the capitalization in the text column.

Overall Summary/Action List of Steps(Appendix Continued)

1. Clean data by deleting unnecessary data, delete duplicates, and fix capitalization
2. Take averages of airline_sentiment_score and negativereason_score in order to see how confidence people are on their opinions
3. Create 6 different figures in order to understand the reasons for poor airline quality and more details on which airlines in what timezones:
    a. Figures on negativereason in relation to negativereason_confidence and airline in relation to negativereason_confidence
    b. Figures on retweet relation to airlines in specific timezones and negativereason_confidence in relation to the timezone specific to United airlines
    c. Figure in order to measure the average sentiment scores in relationship to negative reasons when using United airlines.
4. Take insights and conclusions over what was learned from the data and figures

Sources used in order to help with specific parts of the code:

1. Making a pie chart: https://www.geeksforgeeks.org/how-to-create-pie-chart-from-pandas-dataframe/
2. Sentiment score analysis: https://melaniewalsh.github.io/Intro-Cultural-Analytics/05-Text-Analysis/04-Sentiment-Analysis.html#:~:text=Calculate%20Sentiment%20Scores&text=polarity_scores()%20and%20input%20a,score%20between%20%2D1%2D1.

Sentiment Analysis Figure Code(Appendix)

```
from numpy.lib.function_base import average

from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

# Initialize VADER so we can use it later
sentimentAnalyser = SentimentIntensityAnalyzer()

import pandas as pd
pd.options.display.max_colwidth = 14641

def calculate_sentiment(text):
    # Run VADER on the text
    scores = sentimentAnalyser.polarity_scores(text)
    # Extract the compound score
    compound_score = scores['compound']
    # Return compound score
    return compound_score

United_DataFrame = twitterData[twitterData["textClean"].str.contains("@United", na=False)]

twitterDataMix = (United_DataFrame["negativereason"] == "Bad Flight") & (United_DataFrame["user_timezone"] == "Central Time (US & Canada)") | (United_DataFrame["negativereason"] == "Bad Flight") & (United_DataFrame["user_timezone"] == "Eastern Time (US & Canada)")
twitterDataMix2 = United_DataFrame.loc[twitterDataMix]
twitterDataMix2.head()

twitterDataMix3 = (United_DataFrame["negativereason"] == "Can't Tell") & (United_DataFrame["user_timezone"] == "Central Time (US & Canada)") | (United_DataFrame["negativereason"] == "Can't Tell") & (United_DataFrame["user_timezone"] == "Eastern Time (US & Canada)")
twitterDataMix4 = United_DataFrame.loc[twitterDataMix3]
twitterDataMix4.head()

twitterDataMix5 = (United_DataFrame["negativereason"] == "Cancelled Flight") & (United_DataFrame["user_timezone"] == "Central Time (US & Canada)") | (United_DataFrame["negativereason"] == "Cancelled Flight") & (United_DataFrame["user_timezone"] == "Eastern Time (US & Canada)")
twitterDataMix6 = United_DataFrame.loc[twitterDataMix5]
twitterDataMix6.head()

twitterDataMix7 = (United_DataFrame["negativereason"] == "Customer Service Issue") & (United_DataFrame["user_timezone"] == "Central Time (US & Canada)") | (United_DataFrame["negativereason"] == "Customer Service Issue") & (United_DataFrame["user_timezone"] == "Eastern Time (US & Canada)")
twitterDataMix8 = United_DataFrame.loc[twitterDataMix7]
twitterDataMix8.head()

twitterDataMix9 = (United_DataFrame["negativereason"] == "Damaged Luggage") & (United_DataFrame["user_timezone"] == "Central Time (US & Canada)") | (United_DataFrame["negativereason"] == "Damaged Luggage") & (United_DataFrame["user_timezone"] == "Eastern Time (US & Canada)")
twitterDataMix10 = United_DataFrame.loc[twitterDataMix9]
twitterDataMix10.head()

twitterDataMix11 = (United_DataFrame["negativereason"] == "Flight Attendant Complaints") & (United_DataFrame["user_timezone"] == "Central Time (US & Canada)") | (United_DataFrame["negativereason"] == "Flight Attendant Complaints") & (United_DataFrame["user_timezone"] == "Eastern Time (US & Canada)")
twitterDataMix12 = United_DataFrame.loc[twitterDataMix11]
twitterDataMix12.head()

twitterDataMix13 = (United_DataFrame["negativereason"] == "Flight Booking Problems") & (United_DataFrame["user_timezone"] == "Central Time (US & Canada)") | (United_DataFrame["negativereason"] == "Flight Booking Problems") & (United_DataFrame["user_timezone"] == "Eastern Time (US & Canada)")
twitterDataMix14 = United_DataFrame.loc[twitterDataMix13]
twitterDataMix14.head()

twitterDataMix15 = (United_DataFrame["negativereason"] == "Late Flight") & (United_DataFrame["user_timezone"] == "Central Time (US & Canada)") | (United_DataFrame["negativereason"] == "Late Flight") & (United_DataFrame["user_timezone"] == "Eastern Time (US & Canada)")
twitterDataMix16 = United_DataFrame.loc[twitterDataMix15]
twitterDataMix16.head()
```

```
twitterDataMix17 = (United_DataFrame["negativereason"] == "Lost Luggage") & (United_DataFrame["user_timezone"] == "Central Time (US & Canada)") | (United_DataFrame["negativereason"] == "Lost Luggage") & (United_DataFrame["user_timezone"] == "Eastern Time (US & Canada)")
twitterDataMix18 = United_DataFrame.loc[twitterDataMix17]
twitterDataMix18.head()

twitterDataMix19 = (United_DataFrame["negativereason"] == "longlines") & (United_DataFrame["user_timezone"] == "Central Time (US & Canada)") | (United_DataFrame["negativereason"] == "longlines") & (United_DataFrame["user_timezone"] == "Eastern Time (US & Canada)")
twitterDataMix20 = United_DataFrame.loc[twitterDataMix19]
twitterDataMix20.head()

score1 = 0
for i in twitterDataMix2["textClean"]:
    score1 = calculate_sentiment(i) + score1

score2 = 0
for i in twitterDataMix4["textClean"]:
    score2 = calculate_sentiment(i) + score2

score3 = 0
for i in twitterDataMix6["textClean"]:
    score3 = calculate_sentiment(i) + score3

score4 = 0
for i in twitterDataMix8["textClean"]:
    score4 = calculate_sentiment(i) + score4

score5 = 0
for i in twitterDataMix10["textClean"]:
    score5 = calculate_sentiment(i) + score5

score6 = 0
for i in twitterDataMix12["textClean"]:
    score6 = calculate_sentiment(i) + score6

score7 = 0
for i in twitterDataMix14["textClean"]:
    score7 = calculate_sentiment(i) + score7

score8 = 0
for i in twitterDataMix16["textClean"]:
    score8 = calculate_sentiment(i) + score8

score9 = 0
for i in twitterDataMix18["textClean"]:
    score9 = calculate_sentiment(i) + score9

score10 = 0
for i in twitterDataMix20["textClean"]:
    score10 = calculate_sentiment(i) + score10
```

```
average1 = score1/len(twitterDataMix2["textClean"])
average2 = score2/len(twitterDataMix4["textClean"])
average3 = score3/len(twitterDataMix6["textClean"])
average4 = score4/len(twitterDataMix8["textClean"])
average5 = score5/len(twitterDataMix10["textClean"])
average6 = score6/len(twitterDataMix12["textClean"])
average7 = score7/len(twitterDataMix14["textClean"])
average8 = score8/len(twitterDataMix16["textClean"])
average9 = score9/len(twitterDataMix18["textClean"])
average10 = score10/len(twitterDataMix20["textClean"])

data1 = [["Bad Flight", average1],["Can't Tell", average2],["Cancelled Flight", average3],["Custer Service Issue", average4],["Damaged Luggage", average5],["Flight Attendant Complaints", average6],["Flight Booking Problems", average7],["Late Flight", average8],["Lost Luggage", average9],["longlines",
dataAverage = pd.DataFrame(data1, columns = ["Negative Reasons", "Averages"])

dataAverage.head()


ggplot(dataAverage, aes(x="Negative Reasons", y="Averages", fill = "Negative Reasons"))+ geom_text(aes(label=dataAverage["Negative Reasons"]), size = 7)+ggtitle("Relationship between Negative Reason and Sentiment Score Averages")+ theme_minimal() + geom_point() + theme(axis_text_x=element_blank())
```