1. First we download the dataset and choose the tool that we want to use for the rest of project.
    a. Dataset can be found here: https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment?resource=download
    b. The tool I chose to use is Rstudio.
2. The next step was to clean up the data and find what variables I want to use. This involves first importing the dataset into Rstudio to take a look at the data.
    a. I chose to use the columns "airline_sentiment", "airline_sentiment_confidence", “negativereason”, "airline", "text", “tweet_created”, and "user_timezone".
    b. Next, I checked for any missing values
        i. I fixed “negativereason” by replacing null values with “No reason”
        ii. I fixed “user_timezone” by replacing null values with “No timezone”
    c. Important observations of the dataset are logged in Dataset_Assessment.md
    d. Now the dataset is ready to go into the models and analysis.

3. I first worked on my distribution of sentiment by airline plot
    a. This was a bar chart using airline_sentiment as the y value, and face wrapping by airline.
        i. Also changed the colors and added labels to alter the aesthetics of the chart.
4. I then worked on the distribution of sentiment based on airline and day of the week
    a. First I separated the tweets into the days of the week based on date
    b. Then plot the data in a bar chart separating by airline and using different colors to represent positive, negative, and neutral.
5. To further investigate this I did another box and whisker plot into the distribution of Confidence score by sentiment
    a. This was done by creating a box and whisker plot of all of the sentiment data.
6. Next was the Top 10 Negative Reasons
    a. This was another fairly simple bar chart that showcases why each negative review was left.
    b. I sorted all of the values based on frequency then created a new dataframe with the frequencies of the top 10 reasons.
    c. This data frame went into a bar chart that got sorted based on descending order.
        i. Titles and colors then added for aesthetics
7. This lead me to my next question of whether or not the sentiment can be predicted using a model
    a. I chose a logistic regression model that uses a 80/20 split for training and test data
    b. Next we get predictions from the model and create a confusion matrix using that predicted data
    c. The last step is to graph the model using ggplot2 again.
8. The report concludes with the executive summary and analysis after the models and visualizations have been created

a. All of the analysis information is compiled in the Insight Report and Executive Summary

The Github for all of the files can be found here: https://github.com/dapak2002/MGSC-410-Homework-1