Variables I Kept:

1. **Airline:** Name of the Airline
2. **Airline\_sentiment:** Whether the Airline Sentiment is Positive, Negative or Neutral
3. **Airline\_sentiment\_confidence:** How confident the sentiment analysis is that the airline sentiment is correct
4. **Negativereason:** Negative reason why the tweet was posted (Ie Late Flight)
5. **Negativereason\_confidence:** How confident the sentiment analysis is that the negative reason is correct
6. **Retweet\_count:** Number of times the Tweet was retweeted
7. **Text:** The actual Tweet text

I kept the variable, airline, to keep track of which airline was connected to each tweet. To get a better understanding of each tweet's sentiment, I kept airline\_sentiment and airline\_sentiment\_confidence. To get a better understanding of each negavetime tweet, I kept negativereason and negativereason\_confidence. To understand if there were users who had the same experience, I kept rewteeet\_count. And finally, to understand of there was a correlation between text length and airline sentiment, I kept the variable text. All the variables I kept were needed to answer a business question.

Variables I Removed:

1. **Tweet\_id:** The Twitter user’s ID
2. **Airline\_sentiment\_gold:** No Data Provided
3. **Negativereason\_gold:** No Data Provided
4. **Tweet\_coord:** Coordinates from where the Tweet was sent
5. **User\_timezone:** Timezone from where the Tweet was sent
6. **Tweet\_location:** The location from where the Tweet was sent
7. **Name:** Name of Twitter User
8. **Tweet\_created:** Date and time for when each Tweet was posted

Some variables like, airline\_sentiment\_gold and negativereason\_gold, were left blank. Since there were no values in these two columns, I decided not to use them. I did not really see the point in having these two variables in the dataset, since the entire columns were null. I wonder why they were added to the dataset.Other variables, like tweet\_location, either had blank cells or had strange locations such as “Let’s Play” or “1/1 loner squad”. I did not find these variables very helpful and felt like they wouldn’t present any insights. Not having each exact location, ie “Los Angeles”, made it hard to understand each tweet location. If they had given the precise location it would have been easier to use the column and see where the most common locations for the tweets for each airline. Everytime I would plot a graph using Tweet\_location, User\_timezone or tweet\_created, there would be so many different time zones, coordinates or dates and time. I decided it was not going to answer any of the business questions, so I removed them. Finally, I tried to use Name and tweet\_id, however I could not find mch use for them and they did not answer any business questions, so I removed them.