JJ Goh

Data assessment

Notable data quality issues were encountered, and the methods used to mitigate the inconsistencies are as follows:

1) **Missing values for variables 'negativereason', 'tweetcord', 'negativereason_gold', 'airlinesentiment_gold'.**
   Mitigation: When analyzing variable 'negativereason', I filtered out the missing variables by using filter() function so missing values will not affect analysis. I removed the rest of the variables as they contained more than 80% of missing values.

2) **Variable 'tweet_created' included time of the tweet which created trouble for date grouping**
   Mitigation: Removed timestamp from the variable by using function as.Date

3) **Inconsistent values for the same attribute (e.g. New York represented as "NY", "this place called NY")**
   Mitigation: On excel, edited and used a regular expression to represent a location to ensure consistency

4) **Variable 'text' contains special characters such as hashtags and punctuation marks, and stop words**
   Mitigation: Removed these special characters by using gsub() function. When analyzing for words itself, tokenized the tweets and removed all stop words from tidytext