Assignment 1: Twitter Sentiment Data

Luis Otero

3/13/2020

Complete Data Assessment of Dataset and Outline Findings

This dataset taken from Kaggle recorded the most prevalent problems of each major US Airline and the different reasons for all the negative reception that each airline receives. The dataset contains 14640 rows and 15 columns with each row being a tweet that was posted between February 16 to 24 of the year 2015.

A breakdown of the attributes follows:

Attribute	Type of Variable
tweet_id	Numerical
airline_sentiment	Categorical
airline_sentiment_confidence	Numerical
negativereason	Categorical
negativereason_confidence	Numerical
airline	Categorical
airline_sentiment_gold	Categorical
name	Categorical
negativereason_gold	Categorical
retweet_count	Numerical

Attribute	Type of Variable
text	Categorical
tweet_coord	Categorical
tweet_created	Categorical
tweet_location	Categorical
user_timezone	Categorical

Right off the bat, the dataset appeared to be very inconsistent and untidy as there were numerous missing values in different columns. For instance, there are 5462 missing values in the negative reason column. Instead of dropping all the missing values from the column, I will just fill them in as "Other." The empty spaces that appeared in the negative reasons attribute are due to the corresponding sentiment recording being neutral or positive. Since, the negative reason column only records an instance if the airline sentiment column recorded "negative," it would make sense to have empty spaces if the sentiments were neutral or positive. There also appeared to be a number of columns that were unnecessary such as airline_sentiment_gold, negative_reason_gold and tweet_id. I concluded they were unnecessary since I would not require a user's social media ID to draw the conclusions I desire and the airline_sentiment_gold and negative_reason_gold contained missing values that spanned more than half of the dataset.

There were six different airlines presented in the dataset and all of them have headquarters that are based in the United States. The tweet_coord column can be used to pinpoint the origins of the tweets and observe if location has any impact on the kind of sentiment the user is experiencing. However, the coordinates column has about 80% of its data missing so pinpointing the locations would just be for curiosity's sake. United, US Airways, and American Airlines are the top three companies that have the largest number of negative sentiment tweets while Virgin America has the smallest number in all of the three sentiments. This must be due to the fact that Virgin America is the smallest out of the three airlines and that the airline only conducts flights between metropolitan cities on the West Coast.

The main questions that I am aiming to answer are: what are the top negative reasons people would post about airlines on Twitter, which airlines need to step and take advantage of this feedback, and is there a relationship between negative reasons and the confidence behind those reasons?

List of Steps Taken

1. Explore dataset
2. Look at basic statistics of data
3. Divide data into tables to draw more concrete descriptive statistics
4. Make data modifications beforehand to make visualizations more appealing

5. Make visualizations to generate early answers and conclusions
6. Find which words are most frequent in people's tweets
7. Establish the origin of the tweets with the given coordinates
8. Observe any connections between columns in the dataset (Linear Model)
9. Make conclusions

Analyze data set to produce insight report

Loading Dataset

```r
setwd('/Users/luiscarlosotero/Documents/2019-2020/Applied_Analytics')

Tweets <- read.csv(here::here("twitter-airline-sentiment", "Tweets.csv"),
                             stringsAsFactors = FALSE)
```

Loading Libraries

Names of Variables and Dimensions

```r
#Names of Attributes
names(Tweets)
```

```
##  [1] "tweet_id"                     "airline_sentiment"
##  [3] "airline_sentiment_confidence" "negativereason"
##  [5] "negativereason_confidence"    "airline"
##  [7] "airline_sentiment_gold"       "name"
##  [9] "negativereason_gold"          "retweet_count"
## [11] "text"                         "tweet_coord"
## [13] "tweet_created"                "tweet_location"
## [15] "user_timezone"
```

```r
#Dimensions of Dataset
dim(Tweets)
```

```
## [1] 14640    15
```

The dataset contains 14640 rows and 15 columns.

```r
#Summary Statistics
summary(Tweets)
```

```
##     tweet_id          airline_sentiment  airline_sentiment_confidence
##  Min.   :5.676e+17   Length:14640       Min.   :0.3350
##  1st Qu.:5.686e+17   Class :character   1st Qu.:0.6923
##  Median :5.695e+17   Mode  :character   Median :1.0000
##  Mean   :5.692e+17                      Mean   :0.9002
##  3rd Qu.:5.699e+17                      3rd Qu.:1.0000
##  Max.   :5.703e+17                      Max.   :1.0000
##
##  negativereason     negativereason_confidence   airline
##  Length:14640       Min.   :0.000             Length:14640
##  Class :character   1st Qu.:0.361             Class :character
##  Mode  :character   Median :0.671             Mode  :character
##                     Mean   :0.638
##                     3rd Qu.:1.000
##                     Max.   :1.000
##                     NA's   :4118
##  airline_sentiment_gold     name           negativereason_gold
##  Length:14640           Length:14640       Length:14640
##  Class :character       Class :character   Class :character
##  Mode  :character       Mode  :character   Mode  :character
##
##
##
##
##  retweet_count          text             tweet_coord
##  Min.   : 0.00000   Length:14640       Length:14640
##  1st Qu.: 0.00000   Class :character   Class :character
##  Median : 0.00000   Mode  :character   Mode  :character
##  Mean   : 0.08265
##  3rd Qu.: 0.00000
##  Max.   :44.00000
##
##  tweet_created      tweet_location     user_timezone
##  Length:14640       Length:14640       Length:14640
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

As shown above in the summary statistics of the sentiment analysis dataset, the majority of the attributes are categorical or text data.

```
#Count of Airline_Sentiment
Tweets %>% group_by(airline_sentiment) %>% summarize(Count = n())
```

```
## # A tibble: 3 x 2
##   airline_sentiment Count
##   <chr>             <int>
## 1 negative           9178
## 2 neutral            3099
## 3 positive           2363
```

There is a obviously a greater amount of tweets that are classifed as negative than those that are neutral and positive.

```
#Count of Airlines and Corresponding Sentiments
tab1 <- table(Tweets$airline, Tweets$airline_sentiment)

tab1
```

```
##
##                  negative neutral positive
##   American           1960     463      336
##   Delta               955     723      544
##   Southwest          1186     664      570
##   United             2633     697      492
##   US Airways         2263     381      269
##   Virgin America      181     171      152
```

United, US Airways, and American Airlines are the top three companies that have the largest number of negative sentiment tweets while Virgin America has the smallest number in all of the three sentiments. This must be due to the fact that Virgin America is the smallest out of the three airlines and that the airline only conducts flights between metropolitan cities on the West Coast.

Early Data Modifications

```
#Filling in all empties as NAs so that they can be counted
Tweets <- Tweets %>% mutate_all(na_if, "")
```

```r
# Store the length of each tweet into a new column.
Tweets <- Tweets %>% mutate(text_length = sapply(Tweets$text, function(x) nchar(x)))

# Set those extra-long tweets to NA
Tweets$text_length[Tweets$text_length > 170] <- NA
```

Visualizing Amount of Missing Variables in Dataset

```r
#Checking missing values
gg_miss_var(Tweets) + theme_stata() +
  labs(x = "Variables",
       y = "Missing Values",
       title = "Missing Values in Dataset") +
  theme(plot.title = element_text(size = rel(2))) +
  theme(panel.border = element_rect(linetype = "dashed", fill = NA)) +
  theme(axis.line = element_line(size = 3, colour = "grey80")) +
  theme(axis.text = element_text(colour = "black")) +
  theme(axis.ticks = element_line(size = 1))+
  theme(axis.title.x = element_text(size = rel(2))) +
  theme(axis.title.y = element_text(size = rel(2), angle = 90)) +
  theme(axis.text.y.left = element_text(angle = 0, size = 14)) +
  theme(axis.text.x.bottom = element_text(size = 16))
```

Negativereason_gold and airline_sentiment_gold have about 80% to 90% of their data missing, which would deem them unncessary in this exploratory data analysis.

```
#Checking for Any NAs
sum(is.na(Tweets$negativereason))
```

```
## [1] 5462
```

There are 5462 missing values in the negative reason column. Instead of dropping all the missing values from the column, I will just fill them in as "other."

```
#Filling in the NAs in negativereason
Tweets <- Tweets %>% mutate(negativereason = replace_na(negativereason, 'Other'))
```

```
#Averages and Variances in Negative Reason
Tweets %>% group_by(negativereason) %>%
  summarize(Avg_Confidence = mean(negativereason_confidence, na.rm = TRUE),
            Var_Confidence = var(negativereason_confidence, na.rm = TRUE),
            Count = n()) %>% arrange(desc(Avg_Confidence))
```

```
## # A tibble: 11 x 4
```

```
##    negativereason              Avg_Confidence Var_Confidence Count
##    <chr>                                <dbl>          <dbl> <int>
##  1 Lost Luggage                         0.813         0.0486   724
##  2 Cancelled Flight                     0.783         0.0536   847
##  3 Customer Service Issue               0.780         0.0500  2910
##  4 Late Flight                          0.769         0.0538  1665
##  5 Damaged Luggage                      0.733         0.0549    74
##  6 Flight Attendant Complaints          0.660         0.0568   481
##  7 Bad Flight                           0.632         0.0543   580
##  8 Can't Tell                           0.630         0.0509  1190
##  9 Flight Booking Problems              0.607         0.0477   529
## 10 longlines                            0.594         0.0493   178
## 11 Other                                0             0       5462
```

```
#Averages and Variances in Airline Sentiment
Tweets %>% group_by(airline_sentiment) %>%
  summarize(Avg_Confidence = mean(airline_sentiment_confidence, na.rm = TRUE),
            Var_Confidence = var(airline_sentiment_confidence, na.rm = TRUE),
            Count = n()) %>% arrange(desc(Avg_Confidence))
```

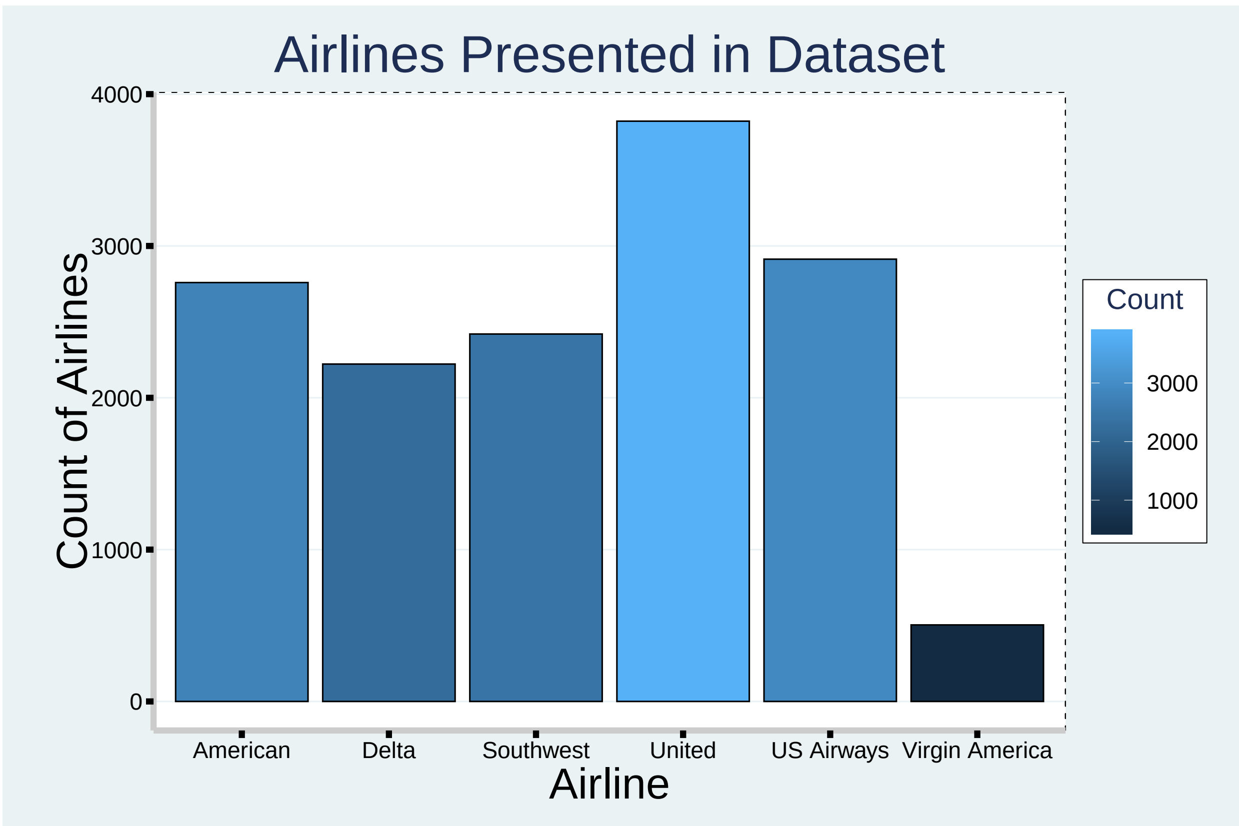
```
## # A tibble: 3 x 4
##   airline_sentiment Avg_Confidence Var_Confidence Count
##   <chr>                      <dbl>          <dbl> <int>
## 1 negative                   0.933         0.0191  9178
## 2 positive                   0.872         0.0322  2363
## 3 neutral                    0.823         0.0344  3099
```

Graphs

Airlines Presented in Dataset and Number of Times Presented

```
by_airline <- Tweets %>% group_by(airline) %>%
  summarise(Count = n())

ggplot(by_airline, aes(x = airline, y = Count, fill = Count)) +
  geom_bar(color = "black", stat = "identity") +
  labs(x = "Airline", y = "Count of Airlines",
       title = "Airlines Presented in Dataset") + theme_stata() +
  theme(plot.title = element_text(size = 36)) +
  theme(panel.border = element_rect(linetype = "dashed", fill = NA)) +
  theme(axis.line = element_line(size = 2, colour = "grey80")) +
  theme(axis.text = element_text(colour = "black")) +
```

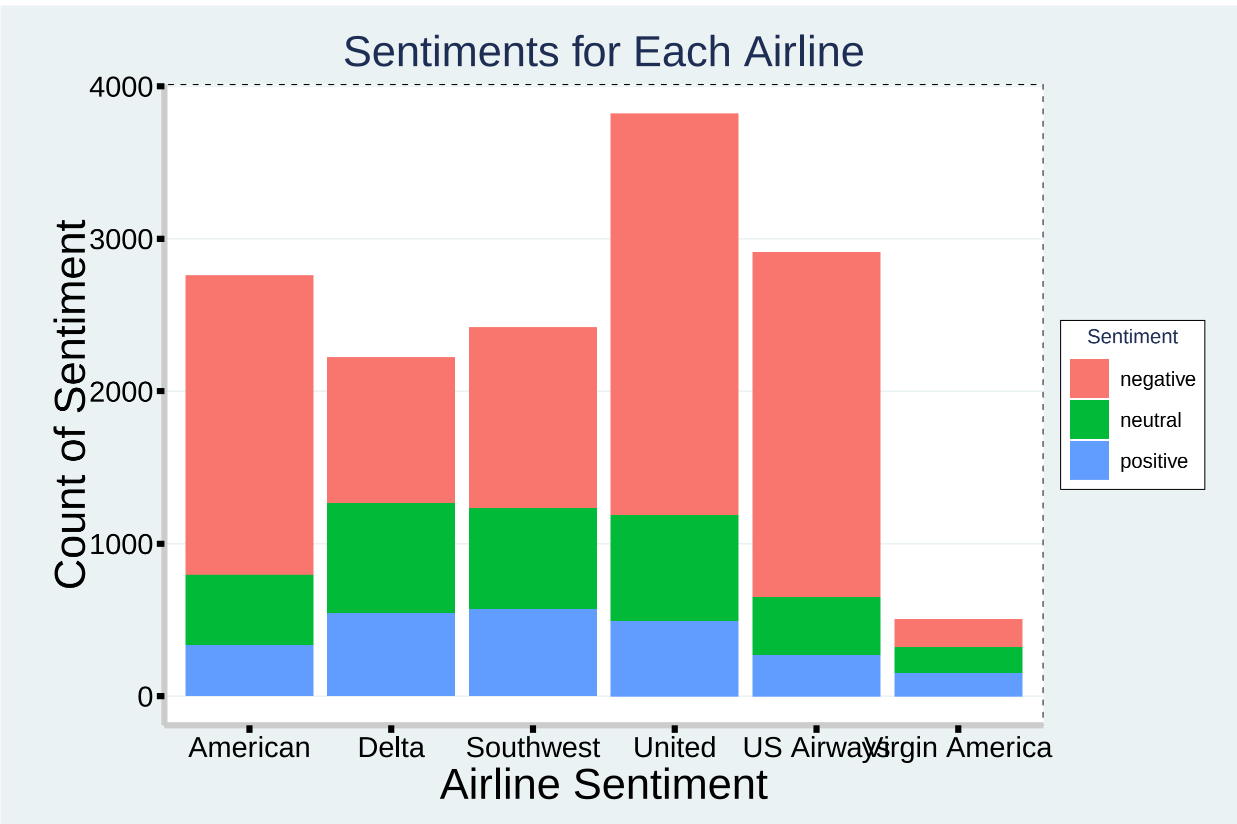
```
theme(axis.ticks = element_line(size = 2)) +
theme(axis.title.x = element_text(size = 30)) +
theme(axis.title.y = element_text(size = 30, angle = 90)) +
theme(axis.text.y.left = element_text(angle = 0, size = 16)) +
theme(axis.text.x.bottom = element_text(size = 16)) +
theme(legend.position = "right") +
theme(legend.title = element_text(size = 20)) +
theme(legend.text = element_text(size = 16)) +
theme(legend.key.size = unit(1, "cm"))
```



Visualizations of Sentiments for Each Airline

	airline	num_tweets	percent_negative	percent_neutral	percent_positive
1	US Airways	2913	77.69	13.08	9.23
2	American	2604	71.58	16.63	11.79
3	United	3822	68.89	18.24	12.87
4	Southwest	2420	49.01	27.44	23.55
5	Delta	2222	42.98	32.54	24.48
6	Virgin America	504	35.91	33.93	30.16

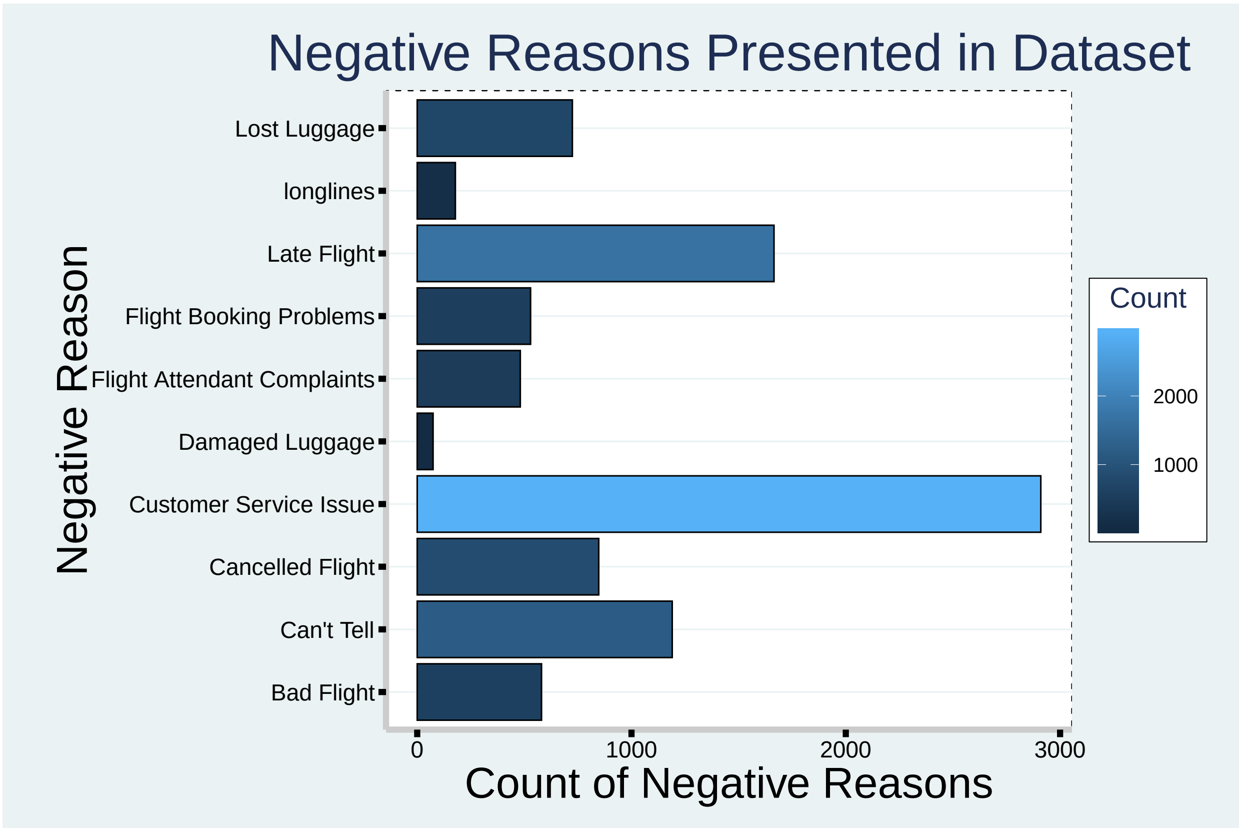
```
airline_sentiment <- as.data.frame(table(Tweets$airline, Tweets$airline_sentiment))
colnames(airline_sentiment) = c("Airline","Sentiment","Freq")

ggplot(airline_sentiment, aes(x = Airline, y = Freq, fill = Sentiment)) +
  geom_bar(stat = 'identity') +
  labs(x = "Airline Sentiment", y = "Count of Sentiment",
       title = "Sentiments for Each Airline") + theme_stata() +
  theme(plot.title = element_text(size = 30)) +
  theme(panel.border = element_rect(linetype = "dashed", fill = NA)) +
  theme(axis.line = element_line(size = 2, colour = "grey80")) +
  theme(axis.text = element_text(colour = "black")) +
  theme(axis.ticks = element_line(size = 2)) +
  theme(axis.title.x = element_text(size = 30)) +
  theme(axis.title.y = element_text(size = 30, angle = 90)) +
  theme(axis.text.y.left = element_text(angle = 0, size = 20)) +
  theme(axis.text.x.bottom = element_text(size = 20)) +
  theme(legend.position = "right") +
  theme(legend.title = element_text(size = 14)) +
  theme(legend.text = element_text(size = 14)) +
  theme(legend.key.size = unit(1, "cm")) +
  theme(strip.text = element_text(size = 14))
```



United, US Airways and American received the most negative reactions.

Visualization of Negative Reasons

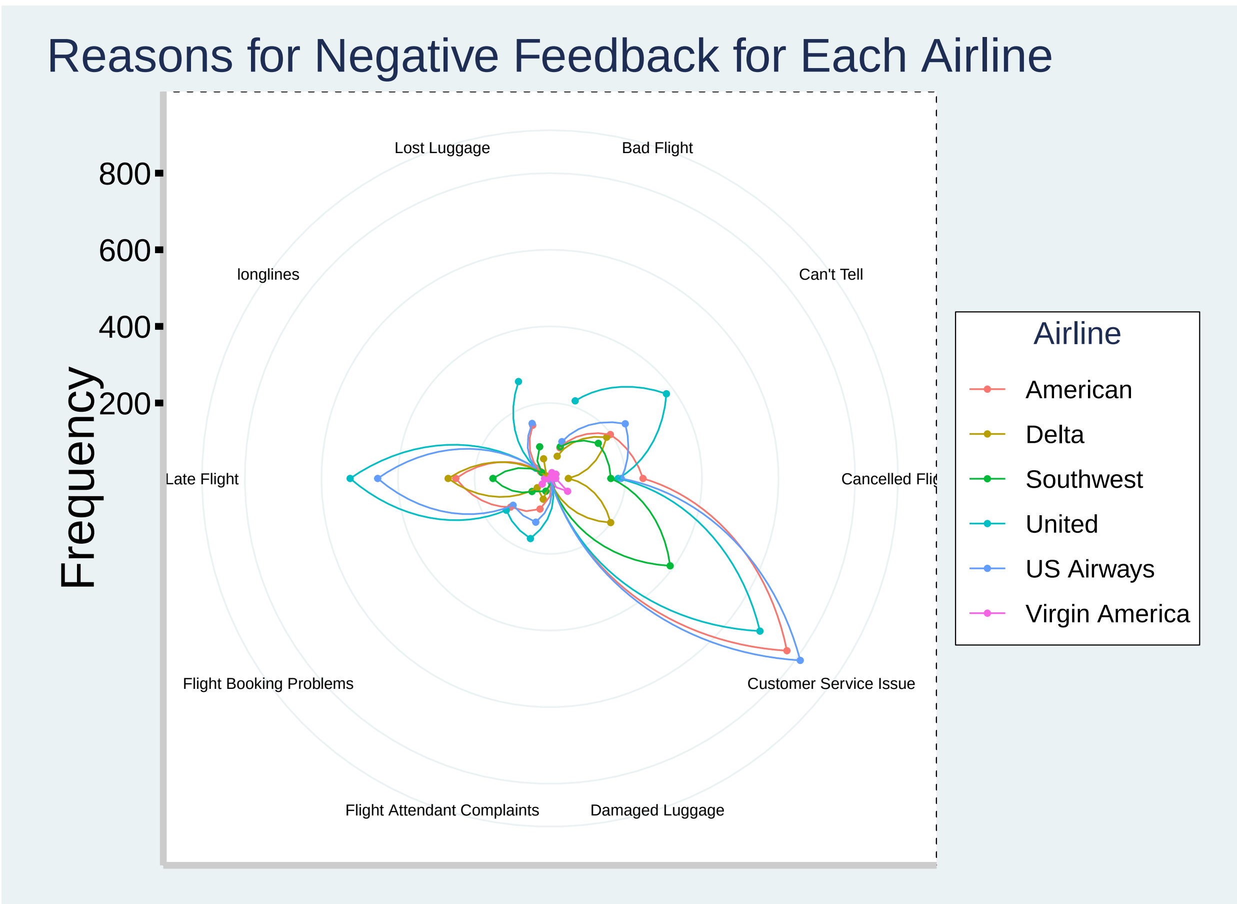
```
by_reason <- Tweets %>% filter(negativereason != "Other") %>%
  group_by(negativereason) %>%
  summarise(Count = n())

ggplot(by_reason, aes(x = negativereason, y = Count, fill = Count)) +
  geom_bar(color = "black", stat = "identity") +
  labs(x = "Negative Reason", y = "Count of Negative Reasons",
       title = "Negative Reasons Presented in Dataset") + theme_stata() +
  theme(plot.title = element_text(size = 36)) +
  theme(panel.border = element_rect(linetype = "dashed", fill = NA)) +
  theme(axis.line = element_line(size = 2, colour = "grey80")) +
  theme(axis.text = element_text(colour = "black")) +
  theme(axis.ticks = element_line(size = 2)) +
  theme(axis.title.x = element_text(size = 30)) +
  theme(axis.title.y = element_text(size = 30, angle = 90)) +
  theme(axis.text.y.left = element_text(angle = 0, size = 16)) +
  theme(axis.text.x.bottom = element_text(size = 16)) +
  theme(legend.position = "right") +
  theme(legend.title = element_text(size = 20)) +
  theme(legend.text = element_text(size = 14)) +
  theme(legend.key.size = unit(1, "cm")) +
  coord_flip()
```



Reasons Behind Each Negative Reason for Each Company

```
Tweets_negative = Tweets %>% filter(negativereason != "Other")

globalSentReasons = as.data.frame(table(Tweets_negative$negativereason,
                                        Tweets_negative$airline))
colnames(globalSentReasons) = c("Reason","Airline", "Freq")

ggplot(globalSentReasons,
       aes(y = Freq, x = Reason, group = Airline, colour = Airline)) + theme_stata() +
  coord_polar() + geom_point() + geom_path() +
  labs(y = "Frequency",
       title = "Reasons for Negative Feedback for Each Airline", x = NULL) +
  theme(plot.title = element_text(size = 30)) +
  theme(panel.border = element_rect(linetype = "dashed", fill = NA)) +
  theme(axis.line = element_line(size = 2, colour = "grey80")) +
  theme(axis.text = element_text(colour = "black")) +
  theme(axis.ticks = element_line(size = 2)) +
  theme(axis.title.x = element_text(size = 30)) +
  theme(axis.title.y = element_text(size = 30, angle = 90)) +
```

```
  theme(axis.text.y.left = element_text(angle = 0, size = 20)) +
  theme(axis.text.x.bottom = element_text(size = 20)) +
  theme(legend.position = "right") +
  theme(legend.title = element_text(size = 20)) +
  theme(legend.text = element_text(size = 16)) +
  theme(legend.key.size = unit(1, "cm")) +
  theme(plot.subtitle = element_text(size = 16))
```

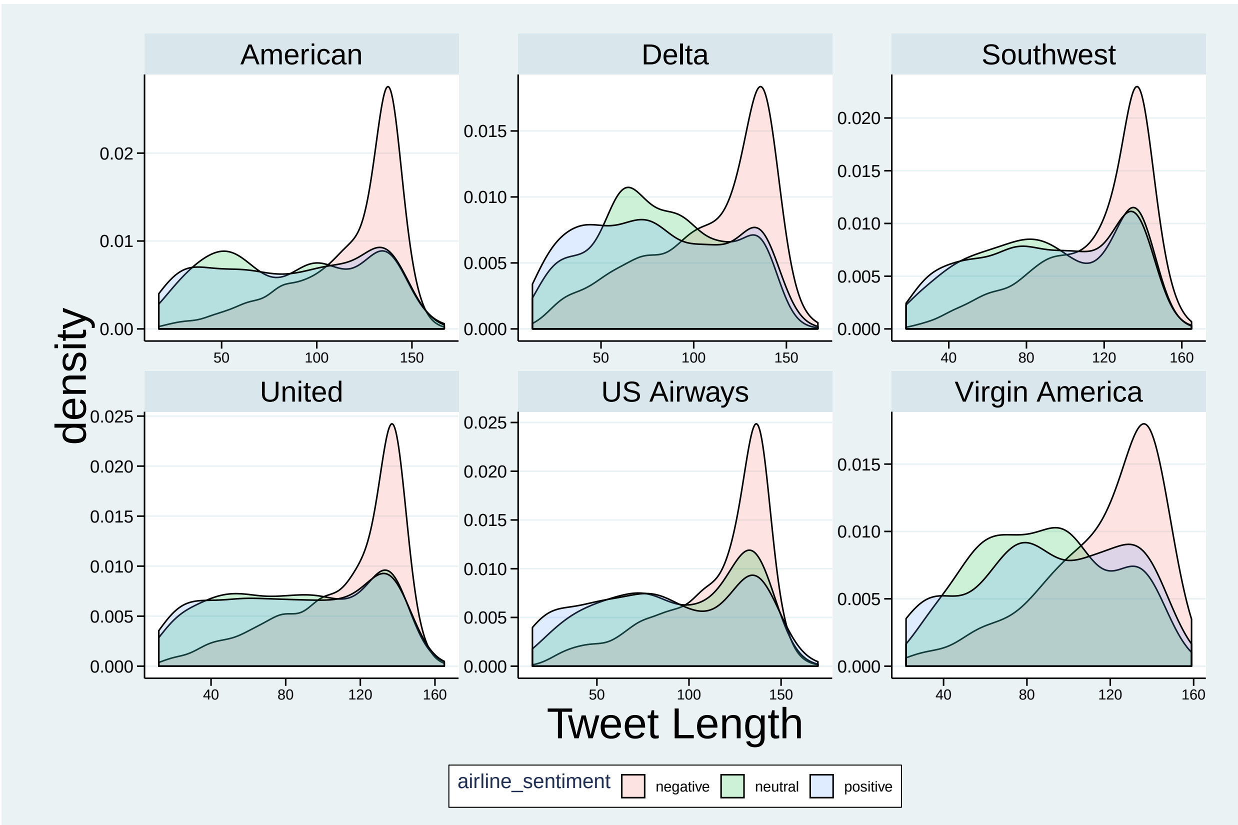


Visualization of Tweet Length by Sentiment

```
ggplot(Tweets, aes(x = text_length,
    fill = airline_sentiment)) +
  geom_density(alpha = 0.2) +
  facet_wrap(~airline, scale = 'free') +
  labs(x = 'Tweet Length') + theme_stata() +
  theme(axis.title.x = element_text(size = 30)) +
  theme(axis.title.y = element_text(size = 30, angle = 90)) +
```

```
  theme(axis.text.y.left = element_text(angle = 0, size = 12)) +
  theme(strip.text = element_text(size = 20))
```

```
## Warning: Removed 2 rows containing non-finite values (stat_density).
```



Most Frequent Words in Positive Sentiment

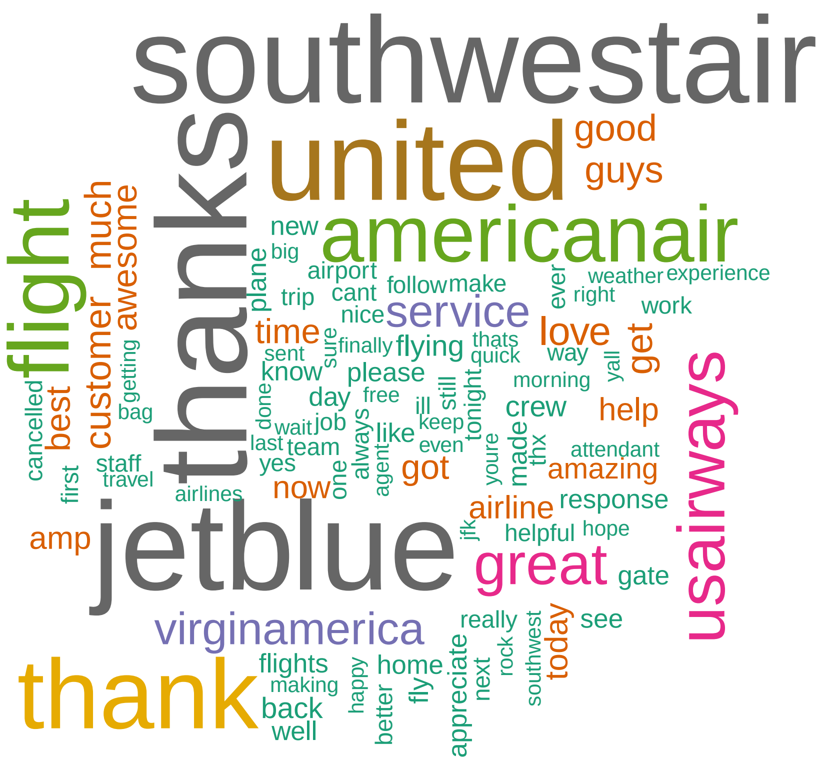
```
Tweets$text <- as.character(Tweets$text)
Tweets_tidy <- Tweets %>%
  unnest_tokens(word, text)
```

```
positive <- Tweets_tidy %>%
  filter(airline_sentiment == "positive")

# Taking out prepositional phrases
list <- c("to", "the","i", "a", "you", "for", "on", "and", "is", "are", "am",
          "my", "in", "it", "me", "of", "was", "your", "so","with", "at", "just", "this"
          "http", "t.co", "have", "that", "be", "from", "will", "we", "an", "can")
```

```r
positive <- positive %>%
  filter(!(word %in% list))

wordcloud(positive[,16], max.words = 100, rot.per = 0.30,
          colors = brewer.pal(8, "Dark2"))
```

```
## Warning in tm_map.SimpleCorpus(corpus, tm::removePunctuation):
## transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, function(x) tm::removeWords(x,
## tm::stopwords())): transformation drops documents
```

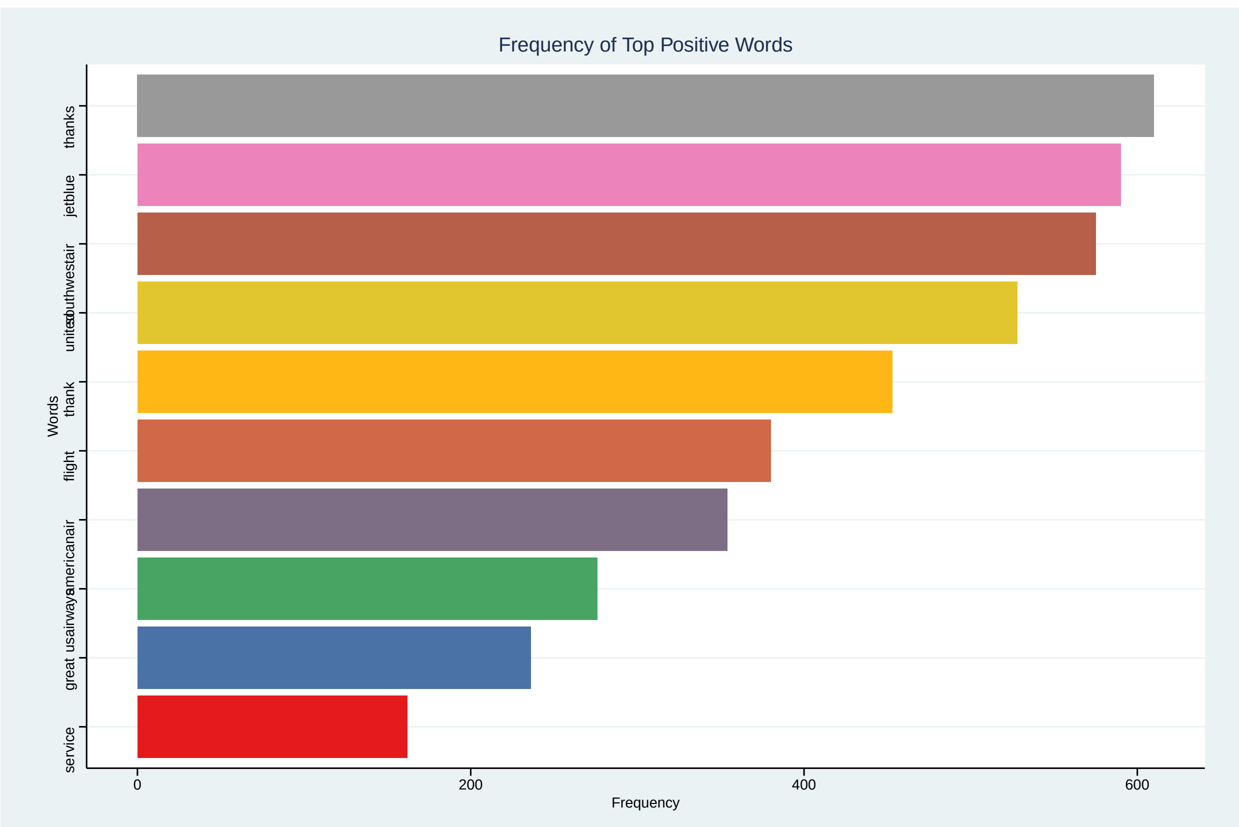



```r
positive <- positive %>%
  count(word, sort = TRUE) %>%
  rename(freq = n)
```

```r
positive <- positive %>%
  top_n(10)
```

```
## Selecting by freq
```

```
colourCount = length(unique(positive$word))
getPalette = colorRampPalette(brewer.pal(9, "Set1"))

# The Top 10 Most Frequent Words in Positive Tweets
positive %>%
  mutate(word = reorder(word, freq)) %>%
  ggplot(aes(x = word, y = freq)) + theme_stata() +
  labs(x = "Words", y = "Frequency",
       title = "Frequency of Top Positive Words") +
  geom_col(fill = getPalette(colourCount)) +
  theme(plot.title = element_text(size = 30)) +
  theme(panel.border = element_rect(linetype = "dashed", fill = NA)) +
  theme(axis.line = element_line(size = 2, colour = "grey80")) +
  theme(axis.text = element_text(colour = "black")) +
  theme(axis.ticks = element_line(size = 2)) +
  theme(axis.title.x = element_text(size = 30)) +
  theme(axis.title.y = element_text(size = 30, angle = 90)) +
  theme(axis.text.y.left = element_text(angle = 0, size = 24)) +
  theme(axis.text.x.bottom = element_text(size = 24)) +
  coord_flip() + theme_stata()
```

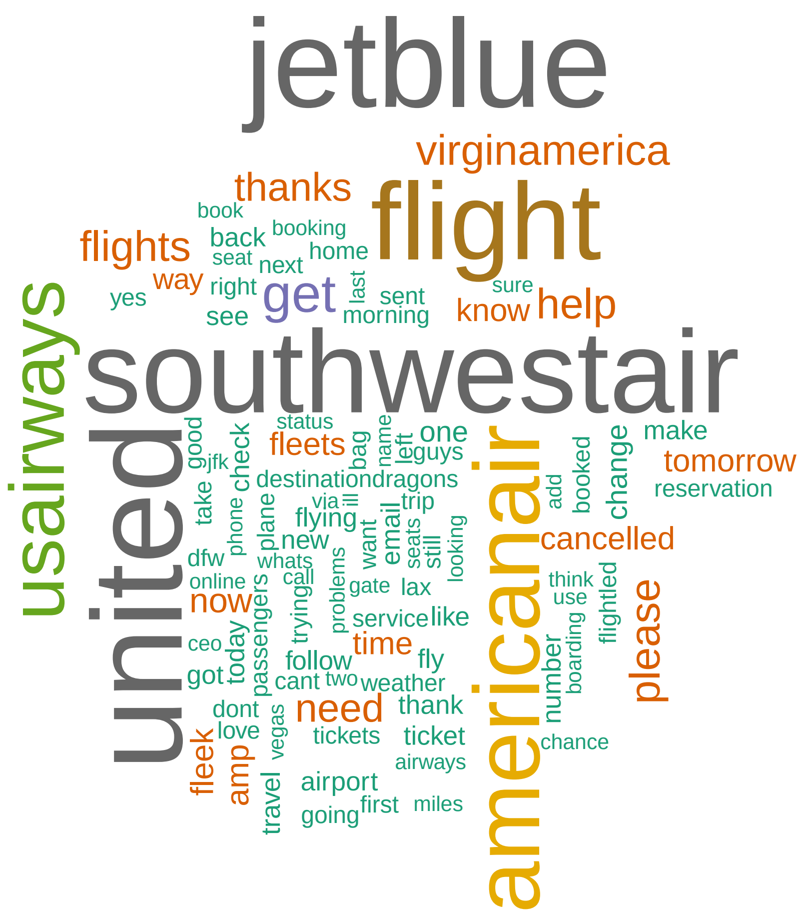


Most Frequent Words in Neutral Sentiment

```r
neutral <- Tweets_tidy %>%
  filter(airline_sentiment == "neutral")

neutral <- neutral %>%
  filter(!(word %in% list))

wordcloud(neutral[,16], max.words = 100, rot.per = 0.30,
          colors = brewer.pal(8, "Dark2"))
```

```
## Warning in tm_map.SimpleCorpus(corpus, tm::removePunctuation):
## transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, function(x) tm::removeWords(x,
## tm::stopwords())): transformation drops documents
```

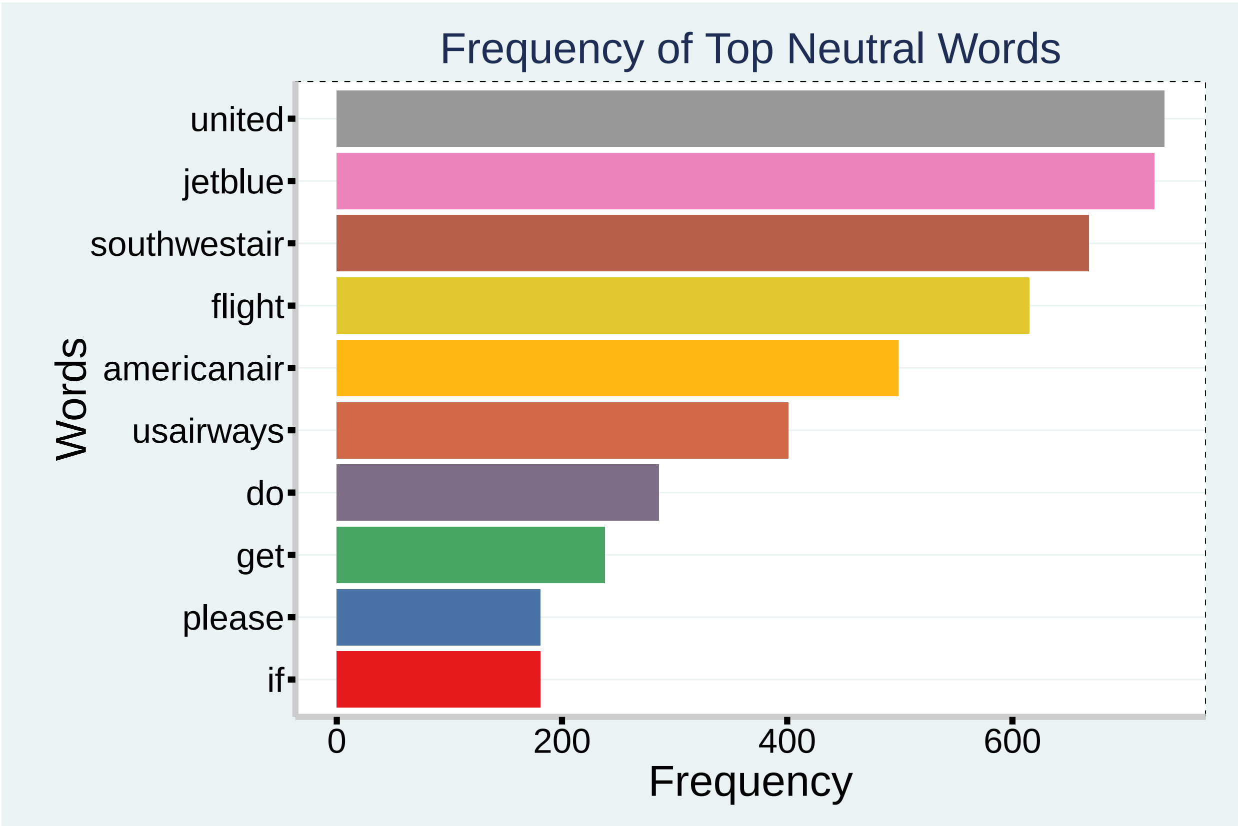


```r
neutral <- neutral %>%
  count(word, sort = TRUE) %>%
  rename(freq = n)
```

```r
neutral <- neutral %>%
  top_n(10)
```

```
## Selecting by freq
```

```r
colourCount = length(unique(neutral$word))
getPalette = colorRampPalette(brewer.pal(9, "Set1"))

# The Top 10 Most Frequent Words in Neutral Tweets
neutral %>%
  mutate(word = reorder(word, freq)) %>%
  ggplot(aes(x = word, y = freq)) + theme_stata() +
  labs(x = "Words", y = "Frequency",
       title = "Frequency of Top Neutral Words") +
  geom_col(fill = getPalette(colourCount)) +
  theme(plot.title = element_text(size = 30)) +
  theme(panel.border = element_rect(linetype = "dashed", fill = NA)) +
  theme(axis.line = element_line(size = 2, colour = "grey80")) +
  theme(axis.text = element_text(colour = "black")) +
  theme(axis.ticks = element_line(size = 2)) +
  theme(axis.title.x = element_text(size = 30)) +
  theme(axis.title.y = element_text(size = 30, angle = 90)) +
  theme(axis.text.y.left = element_text(angle = 0, size = 24)) +
  theme(axis.text.x.bottom = element_text(size = 24)) +
  coord_flip()
```

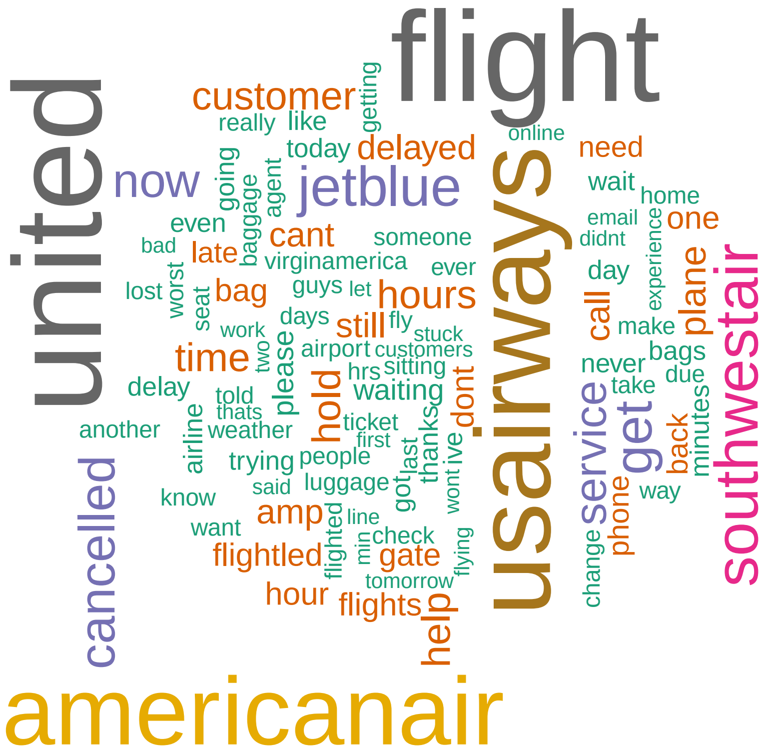


Most Frequent Words in Negative Sentiment

```
negative <- Tweets_tidy %>%
  filter(airline_sentiment == "negative")

negative <- negative %>%
  filter(!(word %in% list))

wordcloud(negative[,16], max.words = 100, rot.per = 0.30,
          colors = brewer.pal(8, "Dark2"))
```

```
## Warning in tm_map.SimpleCorpus(corpus, tm::removePunctuation):
## transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, function(x) tm::removeWords(x,
## tm::stopwords())): transformation drops documents
```



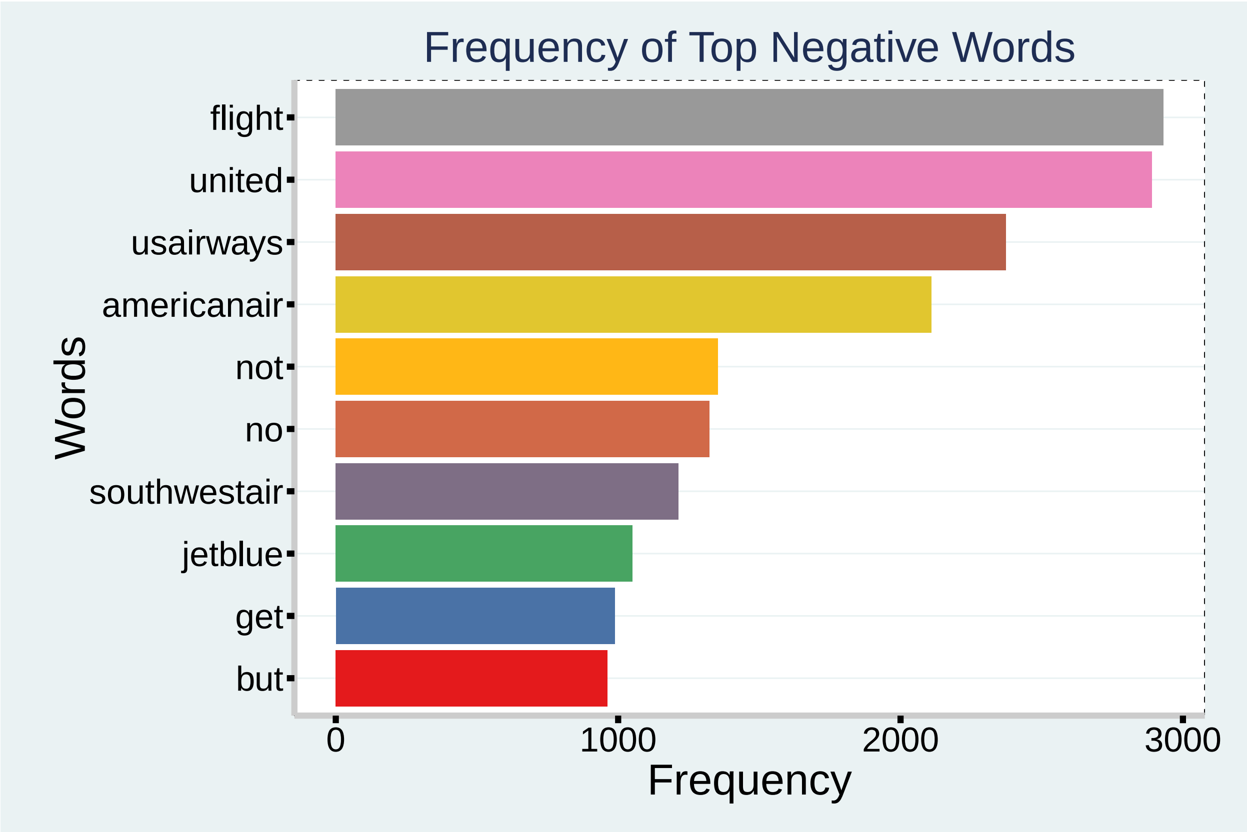

```r
negative <- negative %>%
  count(word, sort = TRUE) %>%
  rename(freq = n)
```

```r
negative <- negative %>%
  top_n(10)
```

```
## Selecting by freq
```

```r
colourCount = length(unique(negative$word))
getPalette = colorRampPalette(brewer.pal(9, "Set1"))

# The Top 10 Most Frequent Words in Negative Tweets
negative %>%
  mutate(word = reorder(word, freq)) %>%
  ggplot(aes(x = word, y = freq)) + theme_stata() +
  labs(x = "Words", y = "Frequency",
       title = "Frequency of Top Negative Words") +
  geom_col(fill = getPalette(colourCount)) +
  theme(plot.title = element_text(size = 30)) +
  theme(panel.border = element_rect(linetype = "dashed", fill = NA)) +
  theme(axis.line = element_line(size = 2, colour = "grey80")) +
  theme(axis.text = element_text(colour = "black")) +
```

```
  theme(axis.ticks = element_line(size = 2)) +
  theme(axis.title.x = element_text(size = 30)) +
  theme(axis.title.y = element_text(size = 30, angle = 90)) +
  theme(axis.text.y.left = element_text(angle = 0, size = 24)) +
  theme(axis.text.x.bottom = element_text(size = 24)) +
  coord_flip()
```



Tweet Locations

```
location = Tweets$tweet_coord
location = location[!is.na(location)]
location = as_tibble(location)
```

```
## Warning: Calling `as_tibble()` on a vector is discouraged, because the behavior is li
## This warning is displayed once per session.
```

```
location = select(location, location = value)
location$location = as.character(location$location)
```

```r
location_2 <- location %>%
  filter(location != "[0.0, 0.0]") %>%
  count(location)
```

```r
location_coords = strsplit(location_2$location, ',')

lat = NULL
long = NULL
for (i in 1:length(location_coords)) {
    lat = c(lat, substring(location_coords[[i]][1], 2)) # removes first character which
    long = c(long, location_coords[[i]][2])
}

location_2$lat <- lat
location_2$long <- long
```

```r
# remove "]" from coordinates
location_2$long = substr(location_2$long, 1, nchar(location_2$long)-1)

location_2$lat = as.numeric(location_2$lat)
location_2$long = as.numeric(location_2$long)
```

```r
options(repr.plot.width = 10, repr.plot.height = 7)
require(maps)
```

```
## Loading required package: maps
```

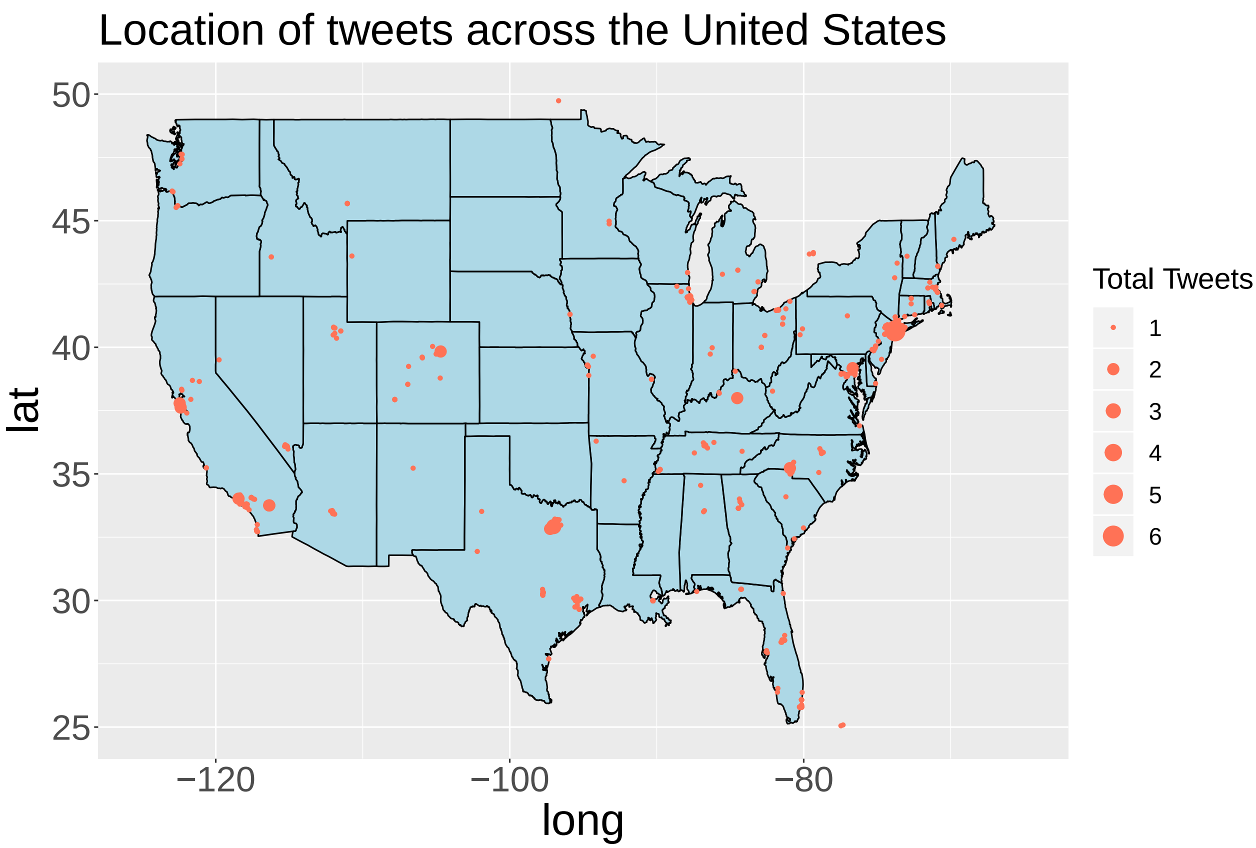
```
##
## Attaching package: 'maps'
```

```
## The following object is masked from 'package:purrr':
##
##     map
```

```r
states <- map_data("state")
ggplot() +
geom_polygon(data = states,
             aes(x = long, y = lat, group = group),
             colour="black", fill = 'lightblue')+
  ggtitle("Location of tweets across the United States") +
  geom_point(data = location_2,
             aes(x = long, y = lat, size = n),
```

```r
            color="coral1") + scale_size(name="Total Tweets") +
  xlim(-125, -65) + ylim(25, 50) +
  theme(plot.title = element_text(size = 30)) +
  theme(axis.title.x = element_text(size = 30)) +
  theme(axis.title.y = element_text(size = 30, angle = 90))  +
  theme(axis.text.y.left = element_text(angle = 0, size = 24)) +
  theme(axis.text.x.bottom = element_text(size = 24)) +
  theme(legend.title = element_text(size = 20)) +
  theme(legend.text = element_text(size = 16)) +
  theme(legend.key.size = unit(1, "cm"))
```

```
## Warning: Removed 56 rows containing missing values (geom_point).
```



Among the States, the tweets are spread out but are more centered around the East Coast in the NYC region.

Models

Data Cleaning

```r
# Checking NAs in negativereason_confidence
sum(is.na(Tweets$negativereason_confidence))
```

```
## [1] 4118
```

```r
# Dropping NAs in negativereason_confidence
Tweets <- Tweets %>% drop_na(negativereason_confidence)

# Dropping NAs in tweet_location and user_timezone
Tweets <- Tweets %>% drop_na(tweet_location) %>% drop_na(user_timezone)

# Eliminating Unnecessary Columns
Tweets <- Tweets %>% select(-c("tweet_coord", "tweet_id",
                               "airline_sentiment_gold",
                               "negativereason_gold",
                               "retweet_count",
                               "name"))

# Splitting Data into Train and Test Sets
num_rows <- nrow(Tweets)
train_idx <- sample(1:num_rows, floor(0.8*nrow(Tweets)))
Tweets_Train <- Tweets %>% slice(train_idx)
Tweets_Test <- Tweets %>% slice(-train_idx)
```

Linear Model

```r
mod1 <- lm(negativereason_confidence ~ negativereason + airline,
           data = Tweets_Train)

summary(mod1)
```

```
##
## Call:
## lm(formula = negativereason_confidence ~ negativereason + airline,
##     data = Tweets_Train)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.49507 -0.12081 -0.00136  0.20786  0.41852
##
## Coefficients:
##                                            Estimate Std. Error t value
## (Intercept)                               0.6185255  0.0153960  40.175
## negativereasonCan't Tell                  0.0090342  0.0166426   0.543
## negativereasonCancelled Flight            0.1588164  0.0180500   8.799
## negativereasonCustomer Service Issue      0.1544816  0.0150527  10.263
## negativereasonDamaged Luggage             0.0987879  0.0372430   2.653
## negativereasonFlight Attendant Complaints  0.0328177  0.0203620   1.612
## negativereasonFlight Booking Problems     -0.0237364  0.0198522  -1.196
## negativereasonLate Flight                 0.1511282  0.0159962   9.448
## negativereasonlonglines                  -0.0093663  0.0257866  -0.363
## negativereasonLost Luggage                0.1815063  0.0185859   9.766
## negativereasonOther                      -0.6171691  0.0162460 -37.989
## airlineDelta                              0.0007391  0.0119015   0.062
## airlineSouthwest                          0.0018141  0.0110292   0.164
## airlineUnited                            -0.0133138  0.0094112  -1.415
## airlineUS Airways                         0.0191342  0.0099030   1.932
## airlineVirgin America                    -0.0107052  0.0205278  -0.521
##                                          Pr(>|t|)
## (Intercept)                               < 2e-16 ***
## negativereasonCan't Tell                  0.58727
## negativereasonCancelled Flight            < 2e-16 ***
## negativereasonCustomer Service Issue      < 2e-16 ***
## negativereasonDamaged Luggage             0.00802 **
## negativereasonFlight Attendant Complaints  0.10710
## negativereasonFlight Booking Problems      0.23190
## negativereasonLate Flight                 < 2e-16 ***
## negativereasonlonglines                   0.71645
## negativereasonLost Luggage                < 2e-16 ***
## negativereasonOther                       < 2e-16 ***
## airlineDelta                              0.95048
## airlineSouthwest                          0.86936
## airlineUnited                             0.15724
## airlineUS Airways                         0.05341 .
## airlineVirgin America                     0.60205
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2106 on 4294 degrees of freedom
## Multiple R-squared:  0.6015, Adjusted R-squared:  0.6001
## F-statistic: 432.1 on 15 and 4294 DF,  p-value: < 2.2e-16
```

```r
#Training set predictions
preds_lm_train <- predict(mod1, Tweets_Train)

#Test set predictions
preds_lm_test <- predict(mod1, newdata = Tweets_Test)

#Train R2 and RMSE
R2(preds_lm_train, Tweets_Train$negativereason_confidence)
```

```
## [1] 0.6015247
```

```r
RMSE(preds_lm_train, Tweets_Train$negativereason_confidence)
```

```
## [1] 0.2102163
```

```r
#Train R2 and RMSE
R2(preds_lm_test, Tweets_Test$negativereason_confidence)
```

```
## [1] 0.5797474
```

```r
RMSE(preds_lm_test, Tweets_Test$negativereason_confidence)
```

```
## [1] 0.2139533
```

```r
#Dropping Certain Negative Reasons
Tweets_Train <- Tweets_Train %>% filter(negativereason != "Can't Tell",
                                        negativereason != "longlines")
```

```r
mod2 <- lm(negativereason_confidence ~ negativereason + airline,
           data = Tweets_Train)

summary(mod2)
```

```
##
## Call:
## lm(formula = negativereason_confidence ~ negativereason + airline,
##     data = Tweets_Train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.49123 -0.11732 -0.00462  0.21325  0.42112
##
## Coefficients:
##                                             Estimate Std. Error t value
## (Intercept)                                 0.621893   0.015433  40.296
## negativereasonCancelled Flight              0.156905   0.017866   8.782
## negativereasonCustomer Service Issue        0.154045   0.014899  10.339
## negativereasonDamaged Luggage               0.099180   0.036831   2.693
## negativereasonFlight Attendant Complaints  0.033446   0.020137   1.661
## negativereasonFlight Booking Problems      -0.023739   0.019637  -1.209
## negativereasonLate Flight                   0.152169   0.015827   9.615
## negativereasonLost Luggage                  0.181133   0.018380   9.855
## negativereasonOther                        -0.617273   0.016073 -38.404
## airlineDelta                               -0.008204   0.012741  -0.644
## airlineSouthwest                            0.008085   0.011664   0.693
## airlineUnited                              -0.018211   0.009958  -1.829
## airlineUS Airways                           0.012300   0.010435   1.179
## airlineVirgin America                      -0.019270   0.020899  -0.922
##                                            Pr(>|t|)
## (Intercept)                                 < 2e-16 ***
## negativereasonCancelled Flight              < 2e-16 ***
## negativereasonCustomer Service Issue        < 2e-16 ***
## negativereasonDamaged Luggage               0.00712 **
## negativereasonFlight Attendant Complaints  0.09681 .
## negativereasonFlight Booking Problems       0.22678
## negativereasonLate Flight                   < 2e-16 ***
## negativereasonLost Luggage                  < 2e-16 ***
## negativereasonOther                         < 2e-16 ***
## airlineDelta                                0.51967
## airlineSouthwest                            0.48822
## airlineUnited                               0.06751 .
## airlineUS Airways                           0.23859
## airlineVirgin America                       0.35655
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2083 on 3710 degrees of freedom
## Multiple R-squared:  0.6411, Adjusted R-squared:  0.6399
## F-statistic: 509.9 on 13 and 3710 DF,  p-value: < 2.2e-16
```

```r
#Dropping Certain Negative Reasons
Tweets_Train <- Tweets_Train %>% filter(negativereason != "Flight Attendant Complaints",
                                        negativereason != "Flight Booking Problems")
```

```r
mod3 <- lm(negativereason_confidence ~ negativereason +
             airline + text_length + user_timezone,
           data = Tweets_Train)

summary(mod3)
```

```
##
## Call:
## lm(formula = negativereason_confidence ~ negativereason + airline +
##     text_length + user_timezone, data = Tweets_Train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49976 -0.11620 -0.00852  0.21209  0.40057
##
## Coefficients:
##                                         Estimate Std. Error t value
## (Intercept)                            0.6187458  0.1481335   4.177
## negativereasonCancelled Flight         0.1542481  0.0178252   8.653
## negativereasonCustomer Service Issue   0.1519058  0.0148302  10.243
## negativereasonDamaged Luggage          0.0970618  0.0377883   2.569
## negativereasonLate Flight              0.1540326  0.0158021   9.748
## negativereasonLost Luggage             0.1828426  0.0184059   9.934
## negativereasonOther                   -0.6135734  0.0162815 -37.685
## airlineDelta                          -0.0056241  0.0139382  -0.404
## airlineSouthwest                       0.0028161  0.0124253   0.227
## airlineUnited                         -0.0214297  0.0108303  -1.979
## airlineUS Airways                      0.0065578  0.0112611   0.582
## airlineVirgin America                 -0.0212406  0.0229046  -0.927
## text_length                            0.0001358  0.0001191   1.140
## user_timezoneAdelaide                  0.2154562  0.2529022   0.852
## user_timezoneAlaska                   -0.0676388  0.1518703  -0.445
## user_timezoneAmerica/Atikokan         -0.0907976  0.2529069  -0.359
## user_timezoneAmerica/Chicago           0.0317247  0.1539049   0.206
## user_timezoneAmerica/Los_Angeles       0.2111689  0.1730458   1.220
## user_timezoneAmerica/New_York         -0.0360688  0.1602563  -0.225
## user_timezoneAmsterdam                 0.0377150  0.1539675   0.245
## user_timezoneArizona                  -0.0485281  0.1483825  -0.327
## user_timezoneAthens                   -0.0960745  0.1657246  -0.580
## user_timezoneAtlantic Time (Canada)   -0.0043033  0.1472516  -0.029
## user_timezoneBeijing                  -0.1454265  0.1884922  -0.772
## user_timezoneBerlin                    0.0619282  0.2072856   0.299
## user_timezoneBrasilia                 -0.1363275  0.1580628  -0.862
## user_timezoneBrisbane                  0.0158523  0.1885721   0.084
```

```
## user_timezoneBrussels                     0.0033542  0.2524203   0.013
## user_timezoneBuenos Aires                 0.2193439  0.2529842   0.867
## user_timezoneCaracas                     -0.0901103  0.1883609  -0.478
## user_timezoneCasablanca                  -0.0827688  0.1730418  -0.478
## user_timezoneCentral America              0.2366457  0.2528124   0.936
## user_timezoneCentral Time (US & Canada)  -0.0033022  0.1463990  -0.023
## user_timezoneCopenhagen                  -0.2100455  0.2063058  -1.018
## user_timezoneDublin                       0.0552283  0.1884176   0.293
## user_timezoneEastern Time (US & Canada)  -0.0141896  0.1462946  -0.097
## user_timezoneEdinburgh                    0.0019960  0.2524104   0.008
## user_timezoneGreenland                   -0.0257186  0.1734506  -0.148
## user_timezoneGuadalajara                 -0.1174011  0.2066602  -0.568
## user_timezoneGuam                        -0.2204629  0.2062539  -1.069
## user_timezoneHawaii                       0.0086608  0.1516361   0.057
## user_timezoneHelsinki                    -0.2285248  0.1688124  -1.354
## user_timezoneIndiana (East)              -0.0970480  0.1563957  -0.621
## user_timezoneIrkutsk                      0.2386367  0.2527422   0.944
## user_timezoneIslamabad                   -0.1108333  0.2066135  -0.536
## user_timezoneJerusalem                   -0.0862064  0.2072062  -0.416
## user_timezoneLa Paz                      -0.1205732  0.2066810  -0.583
## user_timezoneLondon                       0.0091855  0.1482765   0.062
## user_timezoneMadrid                       0.2044547  0.2526924   0.809
## user_timezoneMelbourne                   -0.0259120  0.1793312  -0.144
## user_timezoneMid-Atlantic                -0.0903680  0.1789432  -0.505
## user_timezoneMountain Time (US & Canada) -0.0134492  0.1472021  -0.091
## user_timezoneNairobi                     -0.1268094  0.2526927  -0.502
## user_timezoneNew Delhi                    0.1476681  0.1658811   0.890
## user_timezonePacific Time (US & Canada)   0.0029424  0.1464684   0.020
## user_timezoneParis                       -0.1409776  0.1686953  -0.836
## user_timezonePretoria                    -0.1199446  0.2528954  -0.474
## user_timezoneQuito                        0.0051722  0.1468233   0.035
## user_timezoneRome                        -0.2026249  0.2062763  -0.982
## user_timezoneSantiago                    -0.0101862  0.1685380  -0.060
## user_timezoneSeoul                        0.2123722  0.2065781   1.028
## user_timezoneSolomon Is.                 -0.0078334  0.2527235  -0.031
## user_timezoneStockholm                   -0.0779365  0.2071320  -0.376
## user_timezoneSydney                      -0.0583240  0.1615098  -0.361
## user_timezoneTehran                      -0.0074153  0.1684665  -0.044
## user_timezoneVienna                      -0.0154948  0.2525970  -0.061
## user_timezoneWellington                   0.0032184  0.2524191   0.013
##                                          Pr(>|t|)
## (Intercept)                              3.03e-05 ***
## negativereasonCancelled Flight            < 2e-16 ***
## negativereasonCustomer Service Issue      < 2e-16 ***
## negativereasonDamaged Luggage              0.0103 *
```

```
## negativereasonLate Flight                 < 2e-16 ***
## negativereasonLost Luggage                < 2e-16 ***
## negativereasonOther                       < 2e-16 ***
## airlineDelta                               0.6866
## airlineSouthwest                           0.8207
## airlineUnited                              0.0479 *
## airlineUS Airways                          0.5604
## airlineVirgin America                      0.3538
## text_length                                0.2542
## user_timezoneAdelaide                      0.3943
## user_timezoneAlaska                        0.6561
## user_timezoneAmerica/Atikokan              0.7196
## user_timezoneAmerica/Chicago               0.8367
## user_timezoneAmerica/Los_Angeles           0.2224
## user_timezoneAmerica/New_York              0.8219
## user_timezoneAmsterdam                     0.8065
## user_timezoneArizona                       0.7437
## user_timezoneAthens                        0.5621
## user_timezoneAtlantic Time (Canada)        0.9767
## user_timezoneBeijing                       0.4405
## user_timezoneBerlin                        0.7651
## user_timezoneBrasilia                      0.3885
## user_timezoneBrisbane                      0.9330
## user_timezoneBrussels                      0.9894
## user_timezoneBuenos Aires                  0.3860
## user_timezoneCaracas                       0.6324
## user_timezoneCasablanca                    0.6325
## user_timezoneCentral America               0.3493
## user_timezoneCentral Time (US & Canada)    0.9820
## user_timezoneCopenhagen                    0.3087
## user_timezoneDublin                        0.7695
## user_timezoneEastern Time (US & Canada)    0.9227
## user_timezoneEdinburgh                     0.9937
## user_timezoneGreenland                     0.8821
## user_timezoneGuadalajara                   0.5700
## user_timezoneGuam                          0.2852
## user_timezoneHawaii                        0.9545
## user_timezoneHelsinki                      0.1759
## user_timezoneIndiana (East)                0.5350
## user_timezoneIrkutsk                       0.3451
## user_timezoneIslamabad                     0.5917
## user_timezoneJerusalem                     0.6774
## user_timezoneLa Paz                        0.5597
## user_timezoneLondon                        0.9506
## user_timezoneMadrid                        0.4185
```

```
## user_timezoneMelbourne                     0.8851
## user_timezoneMid-Atlantic                  0.6136
## user_timezoneMountain Time (US & Canada)   0.9272
## user_timezoneNairobi                       0.6158
## user_timezoneNew Delhi                     0.3734
## user_timezonePacific Time (US & Canada)    0.9840
## user_timezoneParis                         0.4034
## user_timezonePretoria                      0.6353
## user_timezoneQuito                         0.9719
## user_timezoneRome                          0.3260
## user_timezoneSantiago                      0.9518
## user_timezoneSeoul                         0.3040
## user_timezoneSolomon Is.                   0.9753
## user_timezoneStockholm                     0.7067
## user_timezoneSydney                        0.7180
## user_timezoneTehran                        0.9649
## user_timezoneVienna                        0.9511
## user_timezoneWellington                    0.9898
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2059 on 3245 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6776, Adjusted R-squared:  0.671
## F-statistic: 103.3 on 66 and 3245 DF,  p-value: < 2.2e-16
```

References

- https://www.kaggle.com/crowdflower/twitter-airline-sentiment
- https://tidyr.tidyverse.org/reference/replace_na.html
- https://towardsdatascience.com/create-a-word-cloud-with-r-bde3e7422e8a