RDA - US Airlines Twitter Sentiment Analysis

Christopher Brandenburg

26 April 2016

Exploratory Data Analysis Twitter Sentiment Analysis - American Airlines

Introduction: The data we are going to look at in this paper is a data dump of 14640 observations with 15 variables of Twitter data regarding openly voiced criticism in from of tweets of US airline customers. These tweets are tied to a tweet ID and user ID

The 15 variables include tweet_id <- unique ID per tweet airline_sentiment <- factor of 3 levels: negative, neutral and postive airline_sentiment_confidence <- confidence score of airline_sentiment classification negativereason <- reason for complaint extracted from tweet negativereason_confidence <- confidence score for negativereason airline <- airline mentioned in the tweet airline_sentiment_gold <- factor of 3 levels: negative, neutral and postive name <- twitter username negativereason_gold <- other reasons, identify what "gold" menas retweet_count <- how many times a tweet was retweeted text <- content of the tweet tweet_coord <- coordinates of the tweet (incase location services are activated) tweet_created <- creation date of the tweet tweet_location <- locationin format city, state (very messy) user_timezone <- timezone the user posted the tweet in

Variables that could be of particular interest to us would be airline_sentiment negativereason airline airline_sentiment_gold (figure out the difference between the two) negativereason_gold name retweet_count

Values we could potentially work with would be ch_id, program_duration, watching_time, timeslot, date, zipcode and coef. uid might be useful for aggregating data as well as ch_id.

Questions to answer: 1. How do different airlines stack up in feedback tweets they have received? - which airlines is doing particularly bad - whats the biggest issue they have 2. Are there users that are particularly loud? - are they heard 3. Are users more likely to voice criticism vs praise?

Loading the different libraries

```
library(ggplot2)
library(dplyr)
library(gmodels)
library(maps)
```

Loading the data

```
df = read.csv("tweets.csv")
```

Summary Statistics of the data

```
#summary statistics of all the variables in the dataset
summary(df)
```

```
##     tweet_id          airline_sentiment airline_sentiment_confidence
##  Min.   :5.676e+17   negative:9178     Min.   :0.3350
##  1st Qu.:5.686e+17   neutral :3099     1st Qu.:0.6923
##  Median :5.695e+17   positive:2363     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
##                        :5462   Min.   :0.000
##  Customer Service Issue:2910   1st Qu.:0.361
##  Late Flight           :1665   Median :0.671
##  Can't Tell            :1190   Mean   :0.638
##  Cancelled Flight      : 847   3rd Qu.:1.000
##  Lost Luggage          : 724   Max.   :1.000
##  (Other)               :1842   NA's   :4118
##            airline     airline_sentiment_gold          name
##  American      :2759           :14600         JetBlueNews:   63
##  Delta         :2222   negative:   32         kbosspotter:   32
##  Southwest     :2420   neutral :    3         _mhertz    :   29
##  US Airways    :2913   positive:    5         otisday    :   28
##  United        :3822                          throthra   :   27
##  Virgin America: 504                          rossj987   :   23
##                                               (Other)    :14438
##                               negativereason_gold retweet_count
##                                         :14608    Min.   : 0.00000
##  Customer Service Issue                 :   12    1st Qu.: 0.00000
##  Late Flight                            :    4    Median : 0.00000
##  Can't Tell                             :    3    Mean   : 0.08265
##  Cancelled Flight                       :    3    3rd Qu.: 0.00000
##  Cancelled Flight\nCustomer Service Issue:    2    Max.   :44.00000
##  (Other)                                :    8
##                        text                                 tweet_coord
##  @united thanks          :    6                                   :13621
##  @AmericanAir thanks     :    5   [0.0, 0.0]                      :  164
##  @JetBlue thanks!        :    5   [40.64656067, -73.78334045]:    6
##  @SouthwestAir sent      :    5   [32.91792297, -97.00367737]:    3
##  @AmericanAir thank you!:    4   [40.64646912, -73.79133606]:    3
##  @united thank you!      :    4   [18.22245647, -63.00369733]:    2
##  (Other)                 :14611   (Other)                         :  841
##                      tweet_created          tweet_location
##  2015-02-24 09:54:34 -0800:    5                 :4733
##  2015-02-24 11:43:05 -0800:    4   Boston, MA     : 157
##  2015-02-23 06:57:24 -0800:    3   New York, NY   : 156
##  2015-02-23 10:58:58 -0800:    3   Washington, DC: 150
##  2015-02-23 14:18:58 -0800:    3   New York       : 127
##  2015-02-23 15:25:46 -0800:    3   USA            : 126
##  (Other)                  :14619   (Other)        :9191
##                       user_timezone
##                              :4820
##  Eastern Time (US & Canada):3744
##  Central Time (US & Canada):1931
##  Pacific Time (US & Canada):1208
##  Quito                       : 738
##  Atlantic Time (Canada)      : 497
##  (Other)                     :1702
```

```
str(df)
```

```
## 'data.frame':    14640 obs. of  15 variables:
##  $ tweet_id                    : num  5.7e+17 5.7e+17 5.7e+17 5.7e+17 5.7e+17 ...
```

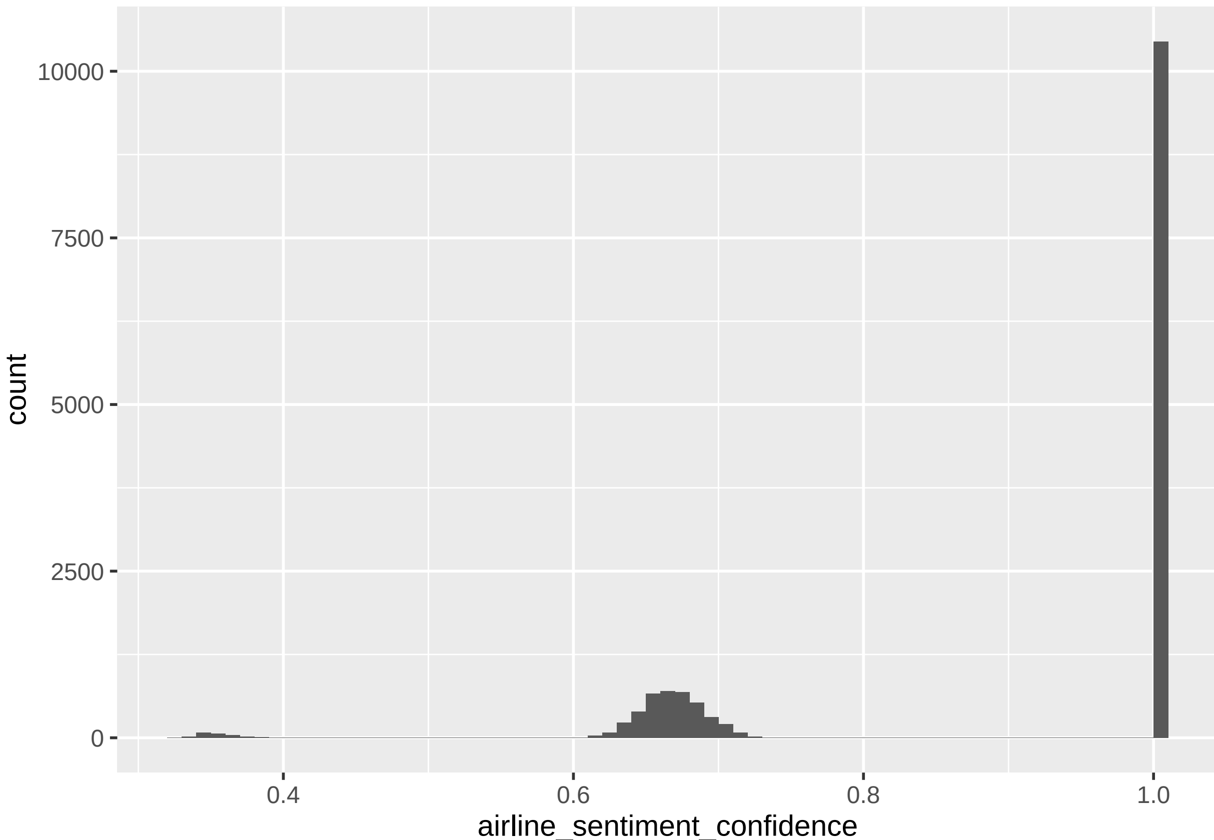
```
##  $ airline_sentiment           : Factor w/ 3 levels "negative","neutral",..: 2 3 2 1 1 1 3 2 3 3 ...
##  $ airline_sentiment_confidence: num  1 0.349 0.684 1 1 ...
##  $ negativereason              : Factor w/ 11 levels "","Bad Flight",..: 1 1 1 2 3 3 1 1 1 1 ...
##  $ negativereason_confidence   : num  NA 0 NA 0.703 1 ...
##  $ airline                     : Factor w/ 6 levels "American","Delta",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ airline_sentiment_gold      : Factor w/ 4 levels "","negative",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ name                        : Factor w/ 7701 levels "0504Traveller",..: 4050 5396 7679 5396 5396 5
##  $ negativereason_gold         : Factor w/ 14 levels "","Bad Flight",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ retweet_count               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ text                        : Factor w/ 14427 levels "\"LOL you guys are so on it\" - me, had this
##  $ tweet_coord                 : Factor w/ 833 levels "","[-33.87144962, 151.20821275]",..: 1 1 1 1 1
##  $ tweet_created               : Factor w/ 14247 levels "2015-02-16 23:36:05 -0800",..: 14212 14170 1
##  $ tweet_location              : Factor w/ 3082 levels "","  || san antonio, texas||",..: 1 1 1221 1
##  $ user_timezone               : Factor w/ 86 levels "","Abu Dhabi",..: 33 64 29 64 64 64 64 64 64 33
```

```
#per column the number of missing values
colSums(is.na(df))
```

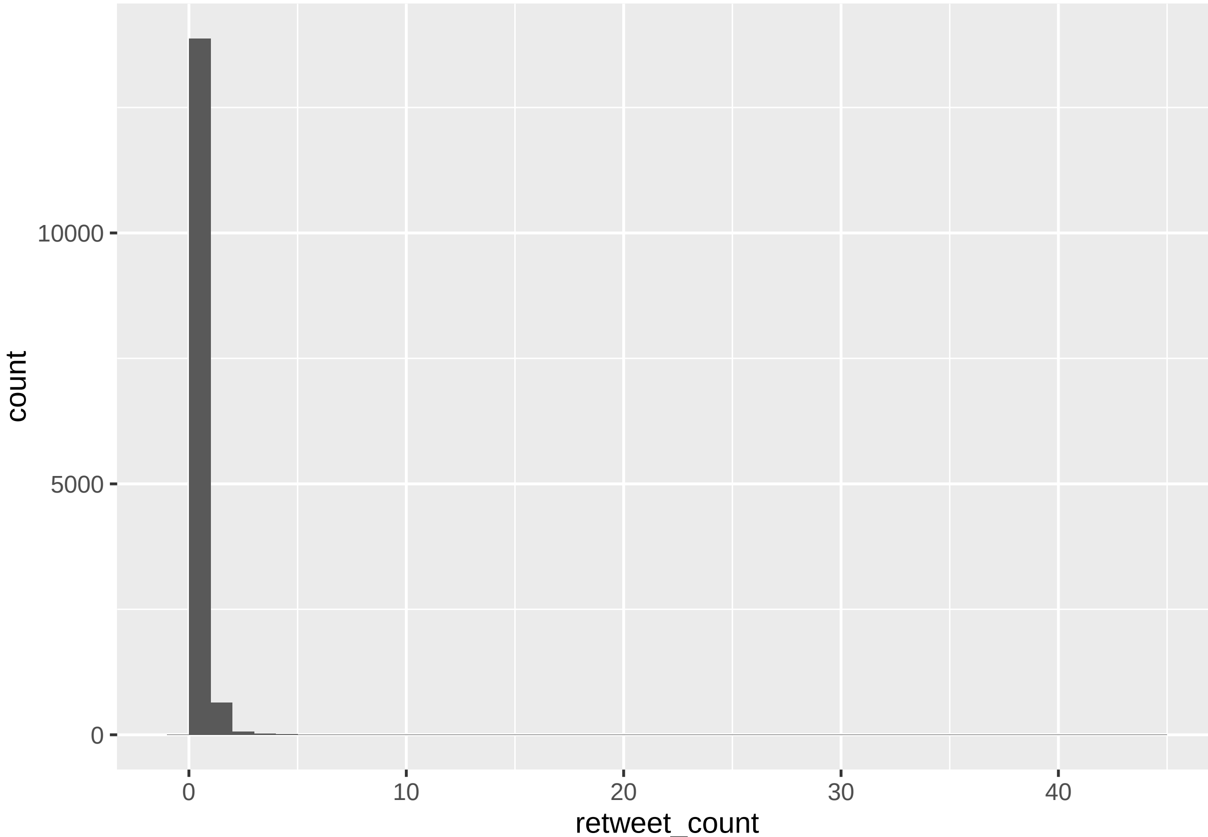
```
##                     tweet_id            airline_sentiment
##                            0                            0
## airline_sentiment_confidence               negativereason
##                            0                            0
##    negativereason_confidence                      airline
##                         4118                            0
##       airline_sentiment_gold                         name
##                            0                            0
##          negativereason_gold                retweet_count
##                            0                            0
##                         text                  tweet_coord
##                            0                            0
##                tweet_created               tweet_location
##                            0                            0
##                user_timezone
##                            0
```

Histograms to check distributions

```
#airline sentiment confidence
ggplot(df, aes(x=airline_sentiment_confidence)) + geom_histogram(binwidth=0.01)
```




```r
#retweet count
ggplot(df, aes(x=retweet_count)) +
geom_histogram(binwidth=1)
```



Many scores are between 0.5 and 0.7 confidence, most are 100 confident in the negativereason.

Looking at cities to check content of the field and counts

```
cities = df %>%
  group_by(tweet_location) %>%
  summarise(count= length(tweet_location)) %>%
  filter(count > 2)

head(cities,10)
```

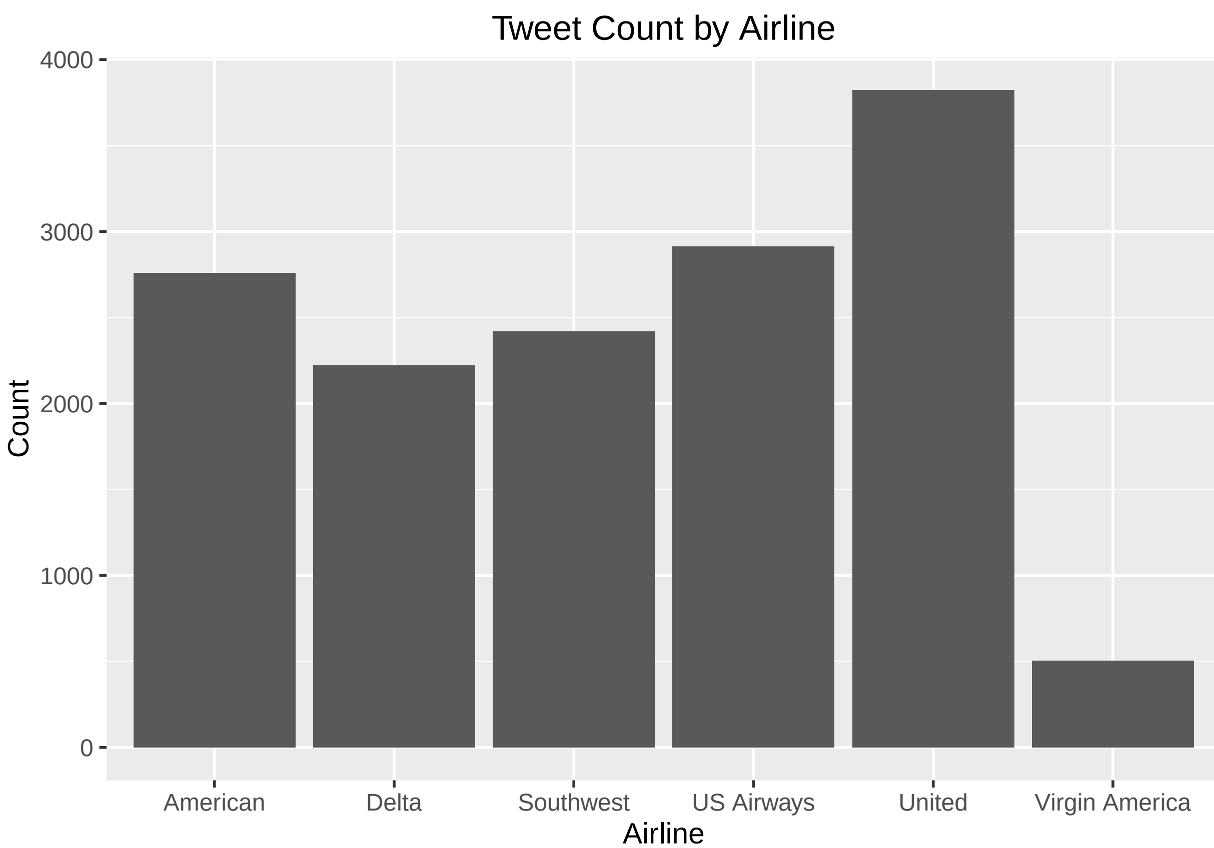
```
## Source: local data frame [10 x 2]
##
##           tweet_location count
##                   (fctr) (int)
## 1                         4733
## 2           Mexico, D.F.     3
## 3     #ManorvilleInExile     5
## 4                 #Omaha     5
## 5  #Westford #marketing      5
## 6               'Straya'     3
## 7                  'Zona     3
## 8        1/1 loner squad     6
## 9                10 ring     5
## 10                 20001     4
```

Nothing useful to find here at this point in time.

1. How do different airlines stack up in feedback tweets they have received?

- which airlines is doing particularly bad
- whats the biggest issue they have

```
ggplot(df, aes(x=airline)) + geom_bar() +
  ggtitle("Tweet Count by Airline") +
  xlab("Airline") +
  ylab("Count")
```



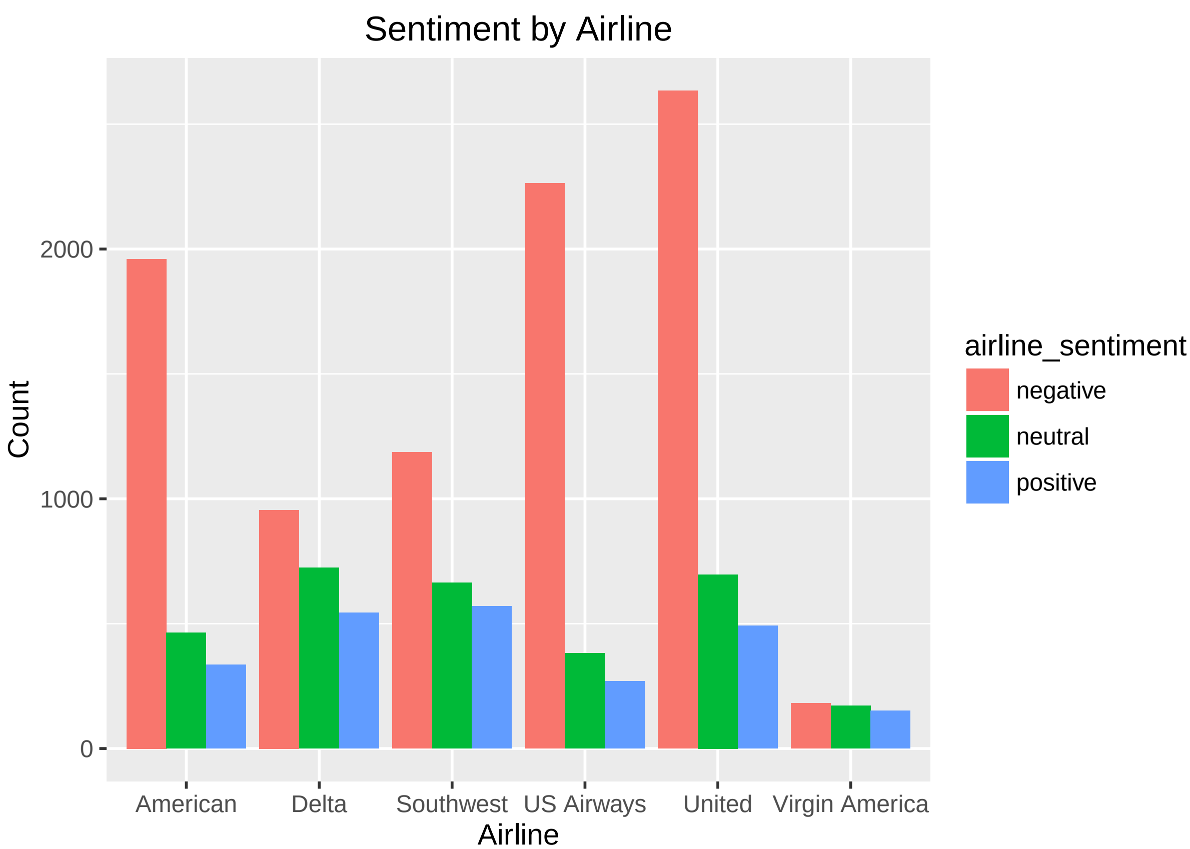

```
tweet_airline = df %>%
  group_by(airline) %>%
  summarise(count= length(airline))
print(tweet_airline)
```

```
## Source: local data frame [6 x 2]
##
##          airline count
##           (fctr) (int)
## 1       American  2759
## 2          Delta  2222
## 3      Southwest  2420
## 4     US Airways  2913
## 5         United  3822
## 6 Virgin America   504
```

Judging by the chart and the table output it becomes clear that United has the highest count of tweets aimed at them. American, Delta, Southwest and US Airways seem to have similar amounts of tweets aimed at them while Virgin America has the least amount of tweets directed at them.

We will now look into how the sentiment levels behave for each of the airlines.

```r
ggplot(df, aes(x=airline, fill =airline_sentiment )) +
  geom_bar(position="dodge") +
  ggtitle("Sentiment by Airline") +
  xlab("Airline") +
  ylab("Count")
```




```r
sentiment_airline = df %>%
  group_by(airline, airline_sentiment) %>%
  summarise(count= length(airline))
print(sentiment_airline)
```

```
## Source: local data frame [18 x 3]
## Groups: airline [?]
##
##          airline airline_sentiment count
##           (fctr)            (fctr) (int)
## 1       American          negative  1960
## 2       American           neutral   463
## 3       American          positive   336
## 4          Delta          negative   955
```

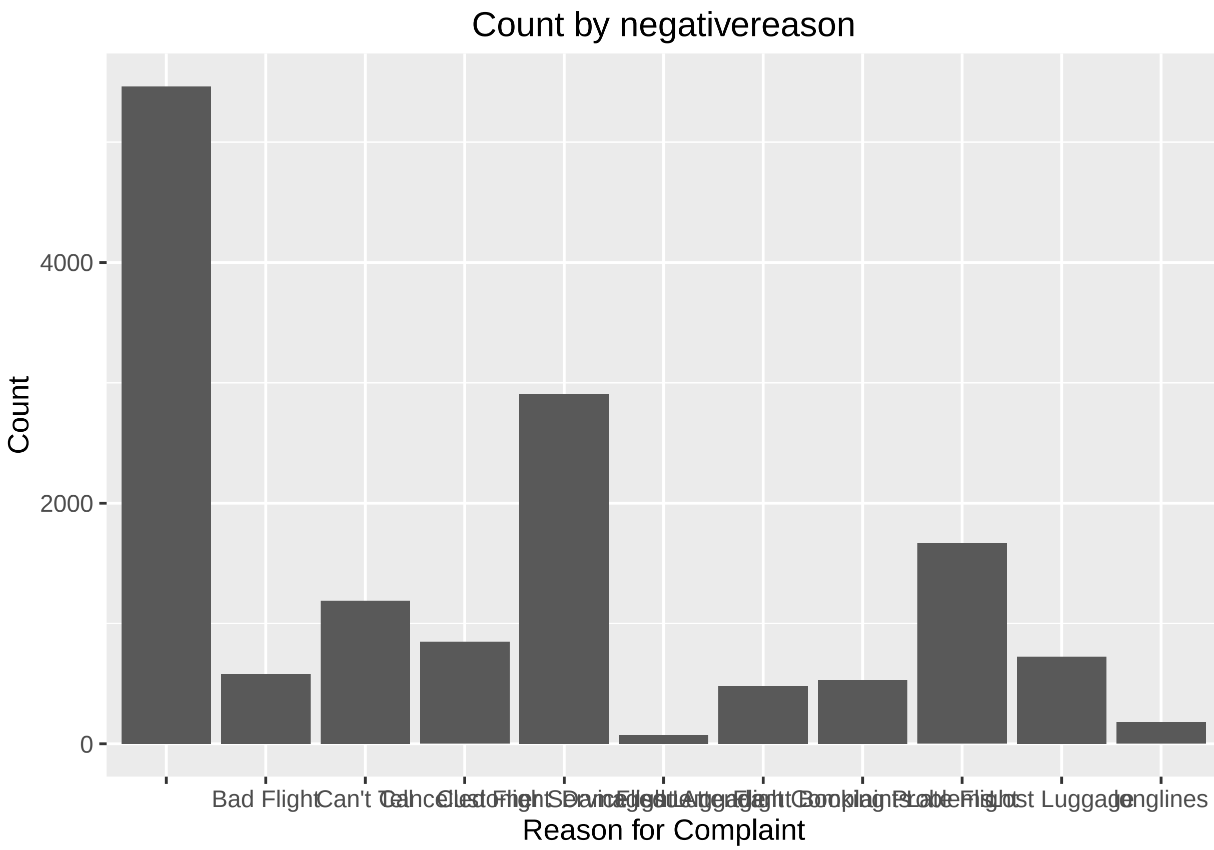
```
## 5           Delta           neutral   723
## 6           Delta          positive   544
## 7       Southwest          negative  1186
## 8       Southwest           neutral   664
## 9       Southwest          positive   570
## 10     US Airways          negative  2263
## 11     US Airways           neutral   381
## 12     US Airways          positive   269
## 13         United          negative  2633
## 14         United           neutral   697
## 15         United          positive   492
## 16 Virgin America          negative   181
## 17 Virgin America           neutral   171
## 18 Virgin America          positive   152
```

We can see that American, US Airways and United clearly have significantly higher amounts of negative tweets compared to neutral and positive tweets. People seem to complain a lot about these three companies. Delta and Southwest also exhibit a higher number of negative tweets but not as significant as the previous 3. Virgin America with the lowest number of tweets seems to have balance between negative, neutral and positive.

The question now arises what exactly the reasons are for people complaining to airlines.

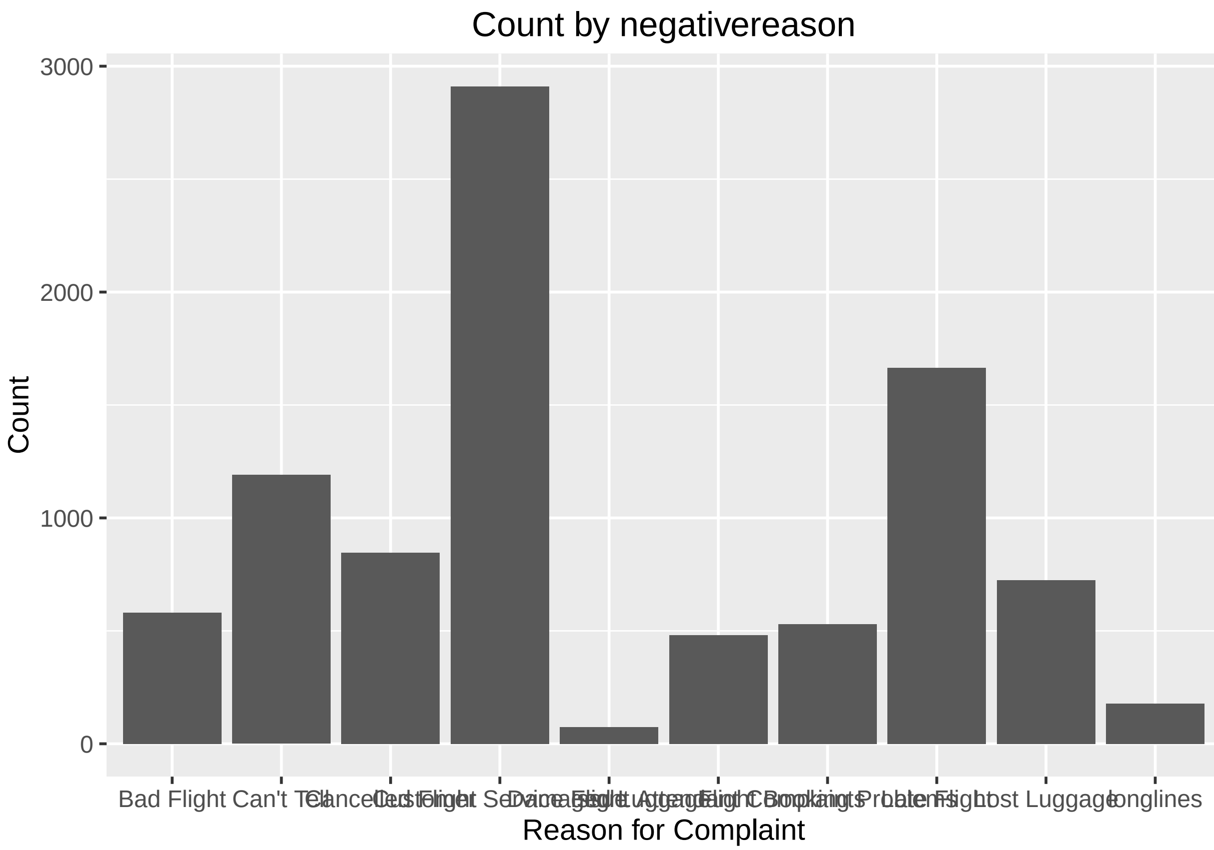
```r
complaints = df %>%
  group_by(negativereason) %>%
  summarise(count=length(negativereason))

ggplot(df, aes(x=negativereason)) + geom_bar() +
  ggtitle("Count by negativereason") +
  xlab("Reason for Complaint") +
  ylab("Count")
```




```
df1 = df %>%
  filter(negativereason !="")

ggplot(df1, aes(x=negativereason)) + geom_bar() +
  ggtitle("Count by negativereason") +
  xlab("Reason for Complaint") +
  ylab("Count")
```




```
print(complaints)
```

```
## Source: local data frame [11 x 2]
##
##                negativereason count
##                        (fctr) (int)
## 1                              5462
## 2                  Bad Flight   580
## 3                  Can't Tell  1190
## 4            Cancelled Flight   847
## 5      Customer Service Issue  2910
## 6             Damaged Luggage    74
## 7  Flight Attendant Complaints   481
## 8      Flight Booking Problems   529
## 9                 Late Flight  1665
## 10               Lost Luggage   724
## 11                  longlines   178
```

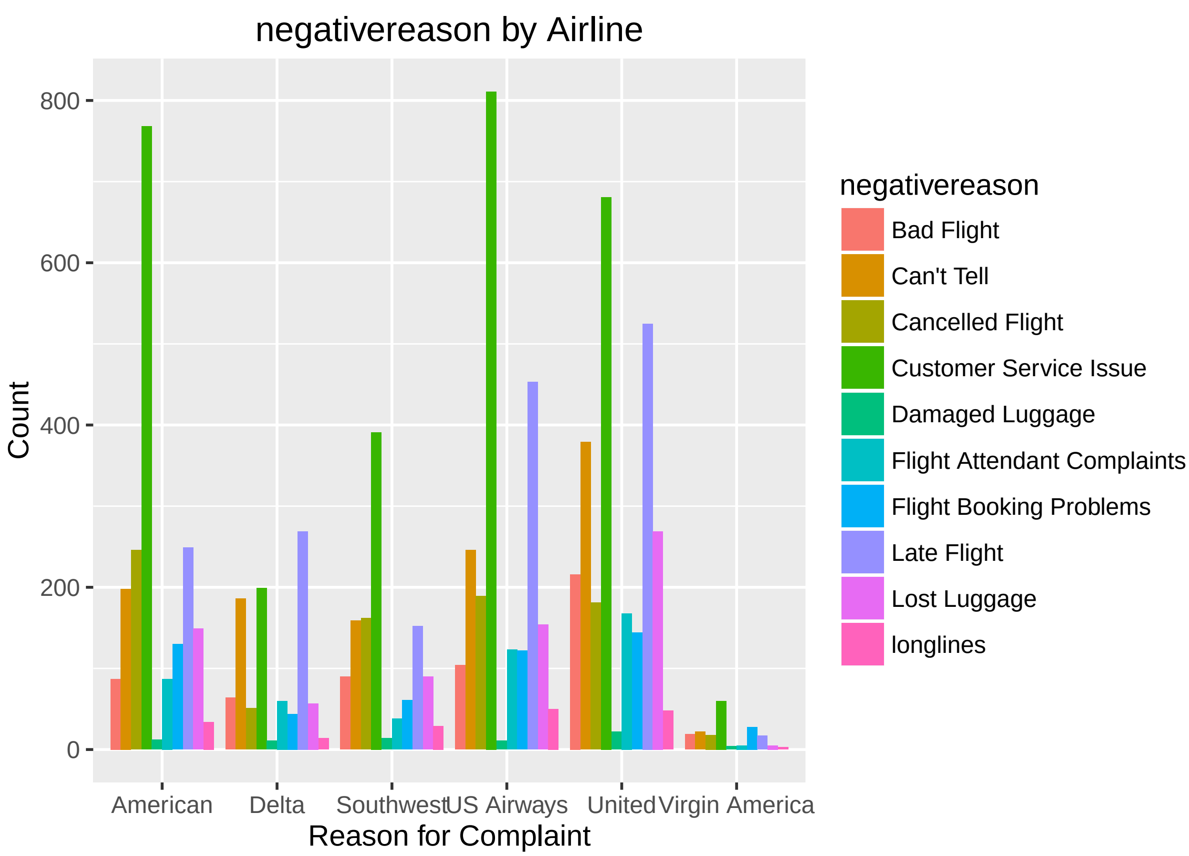
Looking at the table complaints we can see that the three key issues customers complain about are "Customer Service Issues" with 2910 cases (without further information we can't dive deeper into this), "Late flight" is the second most mentioned reason with 1665 cases. "Cant tell" is the third biggest with 1190 cases but there is not more information to be extracted from this.

Breaking this down by Airline might yield a better picture to provide us with an indicator of how badly different airlines are handling CS issues and complaints.

```
complaints_airline = df %>%
  group_by(airline, negativereason) %>%
  summarise(count=length(negativereason))

df1 = df %>%
  filter(negativereason !="")

ggplot(df1, aes(x=airline, fill=negativereason)) + geom_bar(position = "dodge") +
  ggtitle("negativereason by Airline") +
  xlab("Reason for Complaint") +
  ylab("Count")
```




```
print(complaints_airline)
```

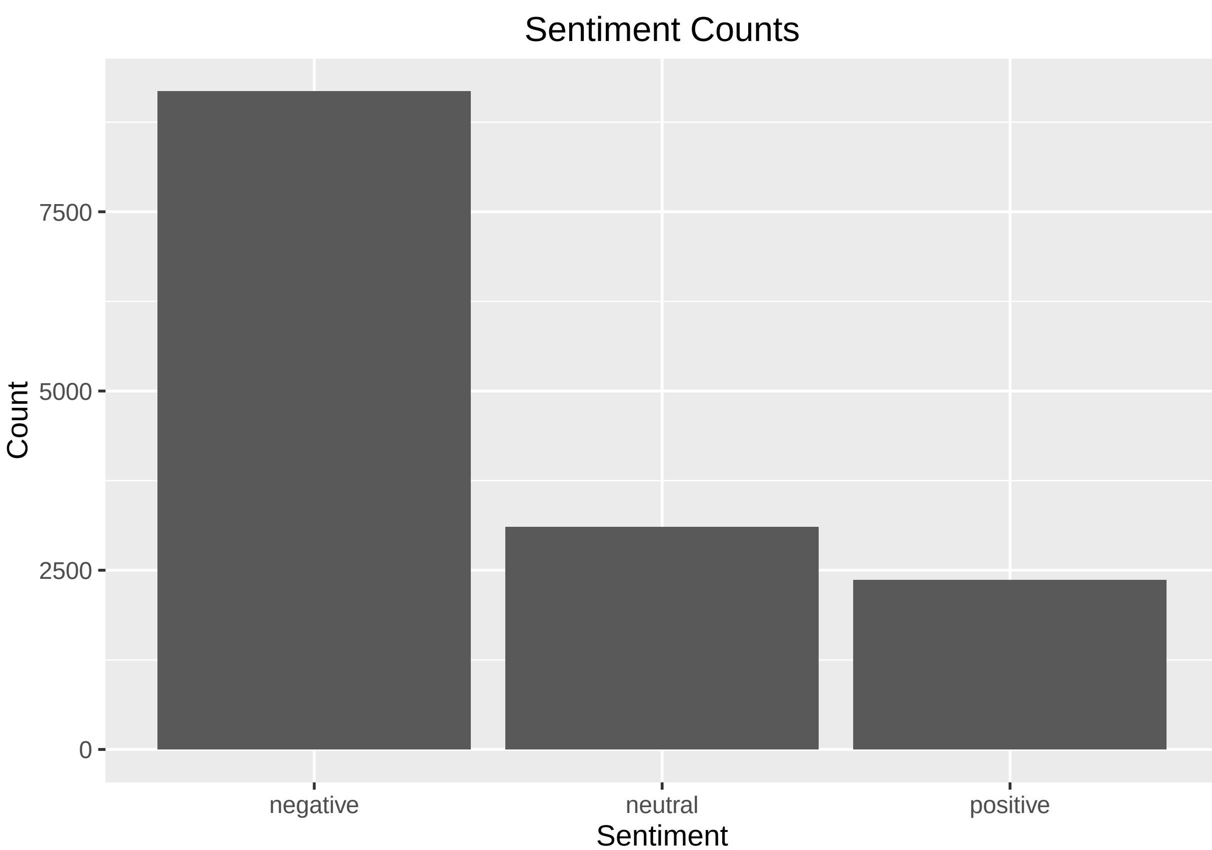
```
## Source: local data frame [66 x 3]
## Groups: airline [?]
##
##      airline             negativereason count
##       (fctr)                     (fctr) (int)
## 1   American                              799
## 2   American                 Bad Flight    87
## 3   American                 Can't Tell   198
## 4   American           Cancelled Flight   246
## 5   American     Customer Service Issue   768
## 6   American            Damaged Luggage    12
```

```
## 7  American Flight Attendant Complaints    87
## 8  American     Flight Booking Problems   130
## 9  American                 Late Flight   249
## 10 American                Lost Luggage   149
## ..      ...                         ...   ...
```

Customer Service Issues stand out for American Airlines, US Airways and United. Southwest seems to be dealing with this as well. Delta has its most complaints coming from Late Flights as well as United and US Airways who seem to have a similar issue. Breaking down complaint reasons by reason for Virgin America we can see that they are very rare,however customer service seems to be a small issue as well. We take the assumption that Virgin America carries out less flights compared to their competitors so following this the number of complaints will be lower as well.

3. Are users more likely to voice criticism vs praise? H0: Users are more likely to tweet if they have something to complain about.

```
ggplot(df, aes(x=airline_sentiment)) +
  geom_bar() + ggtitle("Sentiment Counts") + xlab("Sentiment") + ylab("Count")
```



Judging by the chart we can clearly see that customers are much more likely to voice negative criticism compared to neutral sentinment or positive sentiment.

Options: Following this inital exploration of the data and some small insights we can continue to propose a variety of ways in which data in this form could be used.

Problem: Airlines (all of them except Virgin America) receive a high number of complaints. Most are one of three reasons. Public complaints on Twitter can reach a very wide audience and have implications for how customers perceive the brand and how many mistakes the make.

Solution: Potentially build a stream-based early warning system that identify complaints that might reach critical mass (high number of @mentions plus pickung up retweet speed, both things that need to be investigated further). They could potentially react quickly to Customer Service Issues, directly deal with lost bag claims etc. Airlines would like to avoid bad publicity in this form and have a monetary incentive/budget for a tool like this. We might want to investigate a product/tool in this from in more detail.