# Report

*Bryana Gutierrez, Lily Li, Erica Wong*

*December 8, 2015*

## Cleaning Data

```r
av_data_1 <- read.delim('../rawdata/aviation_data.txt', header = TRUE,
                        sep = '|' , stringsAsFactors = FALSE)
air_data_1 <- read.csv('../rawdata/airplane_crashes.csv',
                       stringsAsFactors = FALSE)

# This takes only the airplane data fromt the data frame as opposed to  all
# aircrafts.
av_data_2 <- subset(av_data_1, Aircraft.Category == ' Airplane '|
                      Aircraft.Category == '  ')
air_data_2 <- air_data_1[!grepl('airship', air_data_1$Type), ]

# We will merge these two date tables so that air_data provides information
# from 1920-1981 and av_data information from 1982-2015. So now we extract
# only the years we want from each data frame.
acc_date <- av_data_2$Event.Date
av_data_2$year <- as.numeric(substr(acc_date, start = nchar(acc_date)-4,
                                    stop = nchar(acc_date)))
av_data_3 <- subset(av_data_2, year >= 1982)

ac_date <- air_data_2$Date
air_data_2$year <- as.numeric(substr(ac_date, start = nchar(ac_date)-3,
                                     stop = nchar(ac_date)))
air_data_3 <- subset(air_data_2, year <= 1981 & year >= 1920)

# In av_data the columns 'Location' and 'Country' should be combined as well
# as 'Make' and 'Model'
av_data_3$new_location <- paste0(av_data_3$Location, av_data_3$Country)
av_data_3$Type <- paste0(av_data_3$Make, av_data_3$Model)

# In air_data the order of data ia currently in ascending order, but to match
# av_data we want it to be in descending order.
air_data_4 <- air_data_3[nrow(air_data_3):1,]

# Now to extract the information (columns) that we want
av_data_4 <- av_data_3[c('Event.Date', 'new_location', 'Type')]
air_data_5 <- air_data_4[c('Date', 'Location', 'Type')]

# We don't want the crash information for crashes missing 'Type' information.
av_data_5 <- subset(av_data_4, Type != '    ')
air_data_6 <- subset(air_data_5, Type != '')

# Rename column names so that they are the same for both data frames.
names(av_data_5) <- c('date', 'location', 'type')
names(air_data_6) <- c('date', 'location', 'type')
```

```r
# Now we combine the two data frames.
data <- rbind(av_data_5, air_data_6)

# Changing the 'type' column to all lowercase
data$type <- tolower(data$type)

# Changing the 'location' column to all lowercase
data$location <- tolower(data$location)

# Showing the head of the cleaned data
head(data)
```

```
##           date                          location
## 1  11/06/2015       gonzales, la  united states
## 2  11/04/2015   mount pocono, pa  united states
## 3  11/04/2015        wakeman, oh  united states
## 4  11/04/2015         haines, ak  united states
## 6  11/03/2015       tecumseh, mi  united states
## 7  11/03/2015   fayetteville, ar  united states
##                                type
## 1                     cessna  120
## 2               schweizer  269c 1
## 3  american autogyro  sparrowhawk
## 4                     cessna  180
## 6                piper  pa 18-150
## 7       cirrus design corp  sr22t
```

```r
# Exporting the 'data' file into the data folder.
write.csv(data, file = '../data/data.csv')
```

## Part 1: Monthly Data

In this part of the project, we were looking to see which of the months had the most plane crashes. In order to do this, we will need to first load the necessary code and packages.

```r
# packages needed
library(stringr)

# function needed
source('../code/num_name.R')
```

Our cleaned data set does not have a column that just contains the months when each accident happens, so we will had to make our own by doing the following.

```r
# importing data
data <- read.csv('../data/data.csv', stringsAsFactors = FALSE)

# removing first column due to repeated index
data$X <- NULL

# extracting months
```

```
data$month <- as.numeric(str_extract(data$date, '[0-9]+'))

# labeling months with words using the num_name function.
dates <- data$month
month <- num_name(dates)

# seeing the new dataframe
head(data)
```

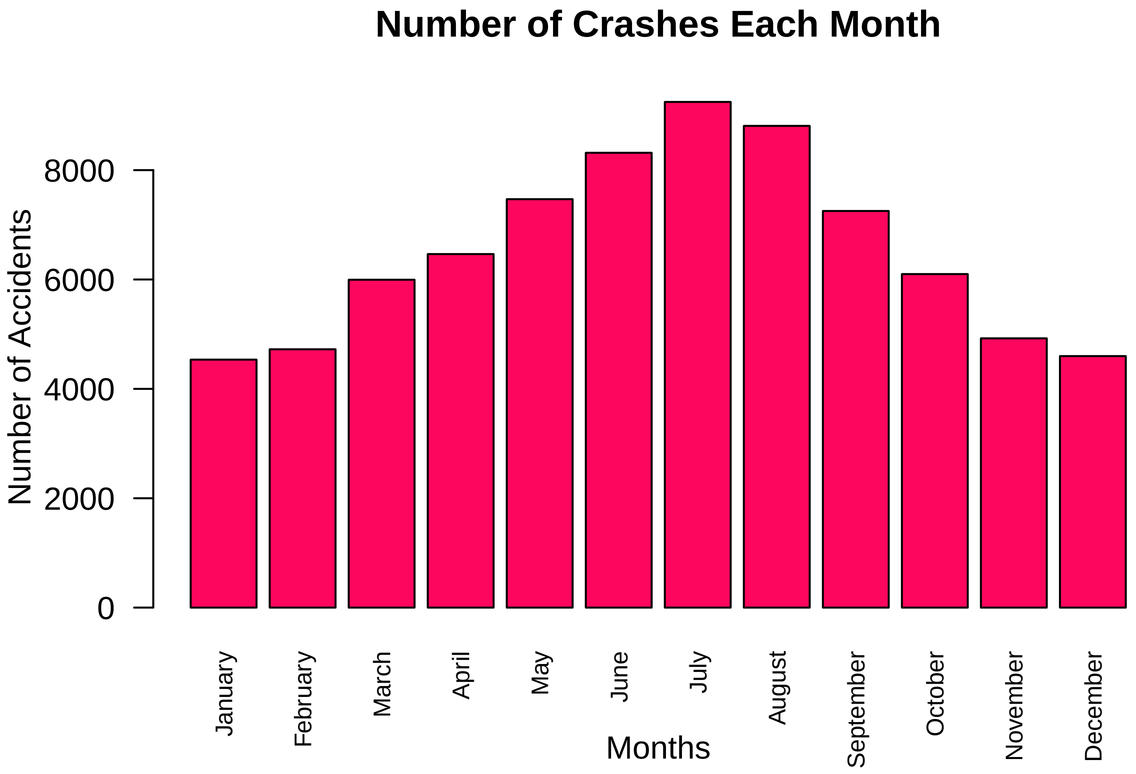
```
##           date                          location
## 1  11/06/2015       gonzales, la  united states
## 2  11/04/2015   mount pocono, pa  united states
## 3  11/04/2015        wakeman, oh  united states
## 4  11/04/2015         haines, ak  united states
## 5  11/03/2015       tecumseh, mi  united states
## 6  11/03/2015   fayetteville, ar  united states
##                               type month
## 1                     cessna  120     11
## 2               schweizer  269c 1     11
## 3  american autogyro  sparrowhawk     11
## 4                     cessna  180     11
## 5                piper  pa 18-150     11
## 6       cirrus design corp  sr22t     11
```

With this column, we are now ready to graph to see our findings.

```
# Number of Entries Each Month
month_names <- c('January', 'February', 'March', 'April', 'May', 'June',
                 'July', 'August', 'September', 'October', 'November',
                 'December')

monthly_data = NULL
for (name in month_names){
  total_per_month <- sum(month == name)
  monthly_data <- c(monthly_data, total_per_month)
}

# Graphing the months
barplot(monthly_data, xlab = 'Months',
        ylab = 'Number of Accidents',
        main = 'Number of Crashes Each Month',
        names.arg = month_names,
        las = 2,
        cex.names = .75,
        col = '#FD075E')
```



Our original belief was that the number of accidents every month would be relatively uniform, but instead, we found that there are more accidents that occur during the middle of the year, with the most occurring in July. We found this idea to be interesting and surprising. The number of accidents could have increased due to there being more traveling being done during the summer due to kids being on summer vacation. However, during the summer seasons, there are also a lot of tropical storms , thunderstorms, and changes in the weather that could affect how safe it is to travel and increases the chance of getting into a plane accident. According to a New York Times article, the number of airplane accidents could be greater in the summer months because that is typically when people fly their private airplanes. A smaller, private aircraft is more likely to get into an accident than a commercial plane. The article states that, "Statistics from the N.T.S.B. show that general aviation aircraft average nearly seven accidents per 100,000 flight hours, compared with an average of 0.16 accidents per 100,000 hours for commercial airlines" (Fowler).

Finally, we will export our graphics.

```
# Exporting the Graphics
# PDF
pdf('../plots_and_graphics/number_of_crashes_each_month.pdf')
barplot(monthly_data, xlab = 'Months',
        ylab = 'Number of Accidents',
        main = 'Number of Crashes Each Month',
        names.arg = month_names,
        las = 2,
        cex.names = .75,
        col = '#FD075E')
dev.off()
```

```
## pdf
##   2
```

```
# PNG
png('../plots_and_graphics/number_of_crashes_each_month.png', res = 96)
barplot(monthly_data, xlab = 'Months',
        ylab = 'Number of Accidents',
        main = 'Number of Crashes Each Month',
        names.arg = month_names,
        las = 2,
        cex.names = .75,
        col = '#FD075E')
dev.off()
```

```
## pdf
##   2
```