ERHS 535 Homework #2

Sarah Derda

Due Date: October 2, 2020

## Section 1: Description of Data

This data is describing various airline companies safety records and whether or not these airlines have experienced incidents, fatal incidents, and/or fatalities based on data from the years 1989 to 1999 and 2000 to 2014. The airline safety data was collected from the “Aviation Safety Network’s Database.” The coder named Nate Silver broke the data into the two previosuly mentioned time periods. Silver wanted to look into whether people were less inclined to fly a particular airline because of their crash history - whether it was their fault or not. He had also accounted for the number of available seat kilometers since one airline could fly more or less than another ultimately adjusting this value. He also looked into the correlation between the time periods to see if there was a huge difference. He found there was no correlation. Along with Silver’s findings, I think it would be useful to look into a portion of which airlines had the least and the most incidents, fatal incidents, and/or fatalities overall. It would be interesting to look into the combined occurrences as well. Some researchers might be interested in this data to see if there patterns across other time periods. If the data would be available, I think it would be interesting to look into the impact of a pandemic, such as the occurrence of COVID-19, and see if it affected airline safety based on fatalities, fatal incidents, and/or incidents. For this assignment, I’ll be focusing on only non-U.S. airlines. The asterisk for each airline means that regional subsidiaries were included. As for the format of the data, it is a comma-separated value file, which is a delimited text file that uses a comma delimiter to separate values or fields within the dataset.

## Section 2: Reading the Data into RStudio

In order to read in this dataset, you would need to install the “readr” package and call upon it’s library. There are multiple ways you could import the “csv” data into RStudio. One way is by saving the file onto your computer to a specific name and importing it using the “Import Dataset” button under “File.” A second way is calling upon the directory where it is located. Another way is by using the “paste0” function to create an object with the url since we’re importing it through GitHub and then reading it into RStudio with “read\_csv” by inputting the file name used to import the data from the internet. The last option is shown below.

I did not need to use any special options to upload this data. The dates from the original dataset were sectioned into two time periods mentioned previously, so no dates needed to be formatted as they were already included into the variable name. I did end up renaming some of the columns for easier readability. In order to look at non-U.S. airlines only, I filtered out all non-U.S. airlines. This was a fairly simple dataset, so I didn’t need to arrange anything using arrange() or add any other variables using mutate() for my interests.

# Upload libraries to use functions.  
  
library(knitr)  
library(readr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyverse)

## ── Attaching packages ──────────────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✓ ggplot2 3.3.0 ✓ purrr 0.3.4  
## ✓ tibble 3.0.3 ✓ stringr 1.3.1  
## ✓ tidyr 1.0.2 ✓ forcats 0.3.0

## ── Conflicts ─────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggplot2)

airline\_url <- paste0("https://raw.githubusercontent.com/fivethirtyeight/data/master/airline-safety/airline-safety.csv")  
  
airline <- read\_csv(file = airline\_url)

## Parsed with column specification:  
## cols(  
## airline = col\_character(),  
## avail\_seat\_km\_per\_week = col\_double(),  
## incidents\_85\_99 = col\_double(),  
## fatal\_accidents\_85\_99 = col\_double(),  
## fatalities\_85\_99 = col\_double(),  
## incidents\_00\_14 = col\_double(),  
## fatal\_accidents\_00\_14 = col\_double(),  
## fatalities\_00\_14 = col\_double()  
## )

head(airline)

## # A tibble: 6 x 8  
## airline avail\_seat\_km\_p… incidents\_85\_99 fatal\_accidents…  
## <chr> <dbl> <dbl> <dbl>  
## 1 Aer Li… 320906734 2 0  
## 2 Aerofl… 1197672318 76 14  
## 3 Aeroli… 385803648 6 0  
## 4 Aerome… 596871813 3 1  
## 5 Air Ca… 1865253802 2 0  
## 6 Air Fr… 3004002661 14 4  
## # … with 4 more variables: fatalities\_85\_99 <dbl>, incidents\_00\_14 <dbl>,  
## # fatal\_accidents\_00\_14 <dbl>, fatalities\_00\_14 <dbl>

summary(airline)

## airline avail\_seat\_km\_per\_week incidents\_85\_99   
## Length:56 Min. :2.594e+08 Min. : 0.000   
## Class :character 1st Qu.:4.740e+08 1st Qu.: 2.000   
## Mode :character Median :8.029e+08 Median : 4.000   
## Mean :1.385e+09 Mean : 7.179   
## 3rd Qu.:1.847e+09 3rd Qu.: 8.000   
## Max. :7.139e+09 Max. :76.000   
## fatal\_accidents\_85\_99 fatalities\_85\_99 incidents\_00\_14   
## Min. : 0.000 Min. : 0.0 Min. : 0.000   
## 1st Qu.: 0.000 1st Qu.: 0.0 1st Qu.: 1.000   
## Median : 1.000 Median : 48.5 Median : 3.000   
## Mean : 2.179 Mean :112.4 Mean : 4.125   
## 3rd Qu.: 3.000 3rd Qu.:184.2 3rd Qu.: 5.250   
## Max. :14.000 Max. :535.0 Max. :24.000   
## fatal\_accidents\_00\_14 fatalities\_00\_14  
## Min. :0.0000 Min. : 0.00   
## 1st Qu.:0.0000 1st Qu.: 0.00   
## Median :0.0000 Median : 0.00   
## Mean :0.6607 Mean : 55.52   
## 3rd Qu.:1.0000 3rd Qu.: 83.25   
## Max. :3.0000 Max. :537.00

airline2 <- airline %>%   
 rename(Airline = airline,  
 PAX\_Carrying\_Capacity = avail\_seat\_km\_per\_week,  
 Incidents\_85\_99 = incidents\_85\_99,  
 Fatal\_Accidents\_85\_99 = fatal\_accidents\_85\_99,  
 Fatalities\_85\_99 = fatalities\_85\_99,  
 Incidents\_00\_14 = incidents\_00\_14,  
 Fatal\_Accidents\_00\_14 = fatal\_accidents\_00\_14,  
 Fatalities\_00\_14 = fatalities\_00\_14) %>%  
 filter(Airline %in% c("Aer Lingus", "Aeroflot\*", "Aerolineas Argentinas", "Aeromexico\*",  
 "Air Canada", "Air France", "Air India\*", "Air New Zealand\*",  
 "Alitalia", "All Nippon Airways", "Austrian Airlines", "Avianca",  
 "British Airways", "Cathay Pacific", "China Airlines", "Condor",  
 "COPA", "Egyptair", "El Al", "Ethiopian Airlines", "Finnair",  
 "Garuda Indonesia", "Gulf Air", "Iberia", "Japan Airlines",  
 "Kenya Airways", "KLM\*", "Korean Air", "LAN Airlines", "Lufthansa\*",  
 "Malaysia Airlines", "Pakistan International", "Philippine Airlines",  
 "Qantas\*", "Royal Air Maroc", "SAS\*", "Saudi Arabian", "Singapore Airlines",  
 "South African", "Sri Lankan / AirLanka", "SWISS\*", "TACA", "TAM",  
 "TAP - Air Portugal", "Thai Airways", "Turkish Airlines", "Vietnam Airlines",  
 "Virgin Atlantic", "Xiamen Airlines"))

## Section 3: Characteristics of the Data

There are 47 rows and 8 columns in the “airline2” dataset.

airline\_table <- data.frame(letters = c("Airline", "PAX\_Carrying\_Capacity", "Incidents\_85\_99",  
 "Fatal\_Accidents\_85\_99", "Fatalities\_85\_99", "Incidents\_00\_14",  
 "Fatal\_Accidents\_00\_14", "Fatalities\_00\_14"),  
 letters = c("Airline name", "Available seats per kilometer per week", "Incidents from 1985 to 1999", "Fatal accidents from 1985 to 1999", "Fatalities from 1985 to 1999", "Incidents from 2000 to 2014", "Fatal accidents from 2000 to 2014", "Fatalities from 2000 to 2014"),  
 letters = c("None", "None", "None", "None", "None", "None", "None", "None"))  
  
kable(airline\_table, align = c("l", "l", "l"),  
 caption = "Airline Characteristics",   
 col.names = c("Column Name", "Description", "Units"))

Airline Characteristics

|  |  |  |
| --- | --- | --- |
| Column Name | Description | Units |
| Airline | Airline name | None |
| PAX\_Carrying\_Capacity | Available seats per kilometer per week | None |
| Incidents\_85\_99 | Incidents from 1985 to 1999 | None |
| Fatal\_Accidents\_85\_99 | Fatal accidents from 1985 to 1999 | None |
| Fatalities\_85\_99 | Fatalities from 1985 to 1999 | None |
| Incidents\_00\_14 | Incidents from 2000 to 2014 | None |
| Fatal\_Accidents\_00\_14 | Fatal accidents from 2000 to 2014 | None |
| Fatalities\_00\_14 | Fatalities from 2000 to 2014 | None |

## Section 4: Summary Statistics

There were no missing values from any of the airlines. The summaries are shown below.

sum(is.na(airline2))

## [1] 0

airline\_summaries <- airline2 %>%  
 summarize(min\_inc\_00\_14 = min(Incidents\_00\_14), max\_inc\_00\_14 = max(Incidents\_00\_14),  
 mean\_inc\_00\_14 = mean(Incidents\_00\_14), na\_inc\_00\_14 = sum(is.na(Incidents\_00\_14)),  
 min\_facc\_00\_14 = min(Fatal\_Accidents\_00\_14), max\_facc\_00\_14 = max(Fatal\_Accidents\_00\_14),  
 mean\_facc\_00\_14 = mean(Fatal\_Accidents\_00\_14), na\_facc\_00\_14 = sum(is.na(Fatal\_Accidents\_00\_14)),  
 min\_fat\_00\_14 = min(Fatalities\_00\_14), max\_fat\_00\_14 = max(Fatalities\_00\_14),  
 mean\_fat\_00\_14 = mean(Fatalities\_00\_14), na\_fat\_00\_14 = sum(is.na(Fatalities\_00\_14))) %>%  
 pivot\_longer(col = min\_inc\_00\_14:na\_fat\_00\_14,  
 names\_to = "Summary",  
 values\_to = "Value")  
  
print(airline\_summaries)

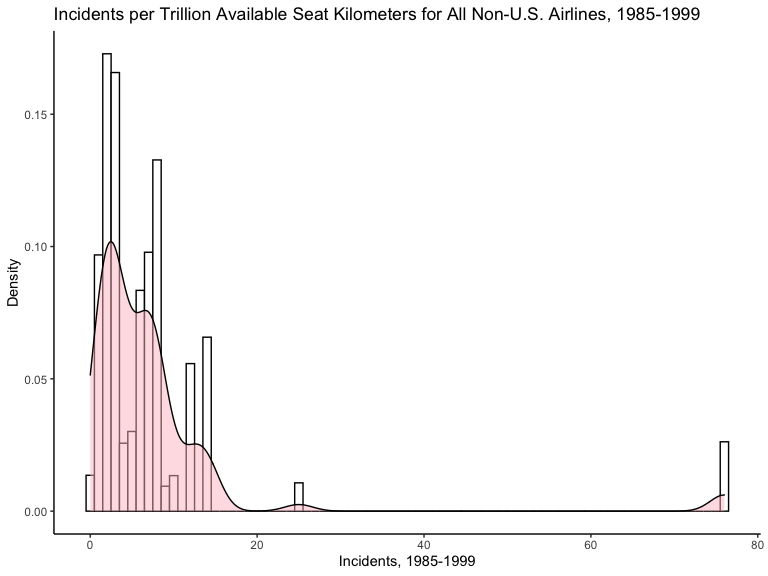
## # A tibble: 12 x 2  
## Summary Value  
## <chr> <dbl>  
## 1 min\_inc\_00\_14 0   
## 2 max\_inc\_00\_14 11   
## 3 mean\_inc\_00\_14 3.04   
## 4 na\_inc\_00\_14 0   
## 5 min\_facc\_00\_14 0   
## 6 max\_facc\_00\_14 2   
## 7 mean\_facc\_00\_14 0.574  
## 8 na\_facc\_00\_14 0   
## 9 min\_fat\_00\_14 0   
## 10 max\_fat\_00\_14 537   
## 11 mean\_fat\_00\_14 51.5   
## 12 na\_fat\_00\_14 0

## Section 5: Visualizing Data

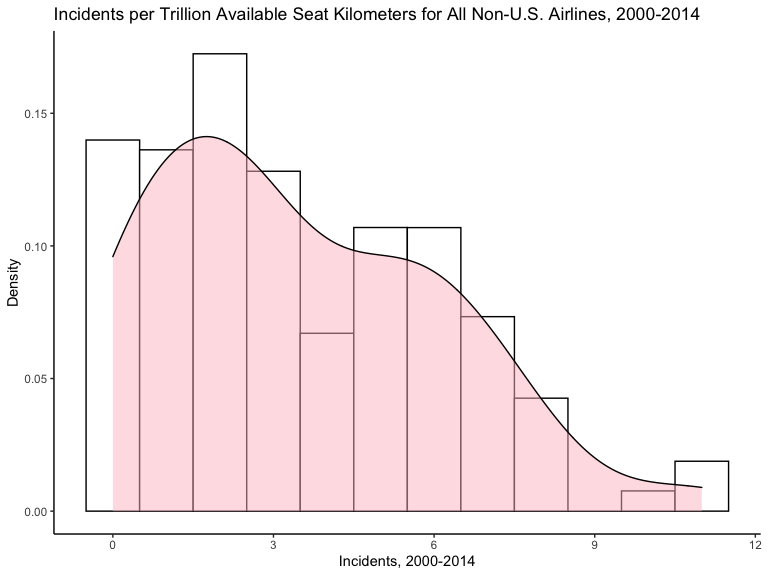
While first visualizing the data, I wanted to look into comparing incidents between the two time periods among non-U.S. airlines. This would allow us to get a better glimpse into the overall distrubution or frequency of incidents in non-U.S. airlines using histograms that are weighted by passenger carrying capacity (seats per kilometers per week). Secondly, by producing scatterplots showing which non-U.S. airlines had the most incidents, we have some insight into which *might* have more of a risk of an incident occurring. Again, this is only examining incidents as that has some indication of something going wrong with the flight rather than looking into fatalities, which indicates that the incident or accident was fatal. Furthermore, these graphs could aid in looking at histories of an individual airline - U.S. or non-U.S based. Also, I think it would be interesting comparing two countries’ airlines if a researcher wanted to examine countries that seem to portray more of a risk in safety. In producing these graphs, two good graphing techniques that were met were 1) high data density and 2) making the order meaningful since there were two time periods involved. High data density was met in the ink to data ratio within the graphs. The order was meaningful due to the fact that we’re comparing the time period from 1985-1999 and 2000-2014. However, it is important to note that when comparing these, the values along the x-axis are different between the two time periods.

As we can see from the graphs, it looks like there were more incidents among all airlines from 1985 to 1999 compared to the time period from 2000 to 2014. From 1985 to 1999, Aeroflot had the highest number of incidents. From 2000 to 2014, Saudi Arabian had the highest amount of incidents, but it was a far less number than Aeroflot.

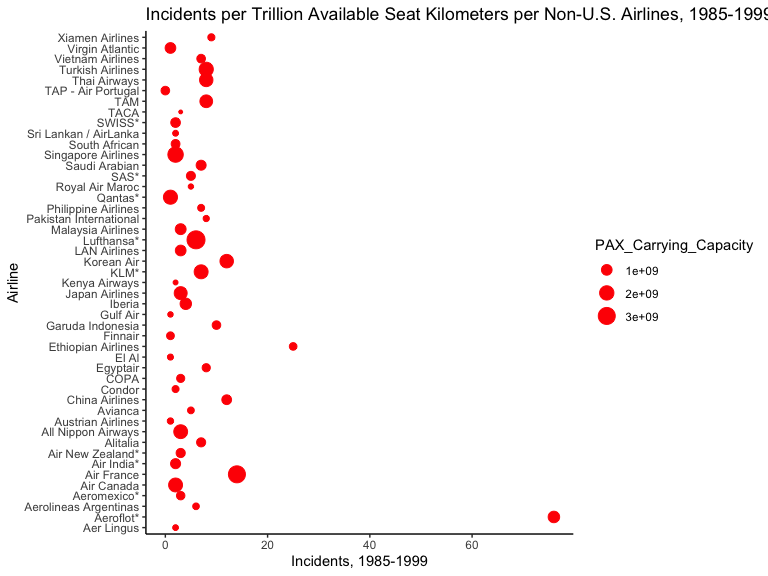
# Statistical Graphing - Histograms  
  
p1 <- ggplot(airline2, aes(x = Incidents\_85\_99, y = ..density.., weight = PAX\_Carrying\_Capacity)) +   
 geom\_histogram(binwidth = 1, color = "black", fill = "white") +   
 labs(x = "Incidents, 1985-1999", y = "Density", title = "Incidents per Trillion Available Seat Kilometers for All Non-U.S. Airlines, 1985-1999") +   
 theme\_classic() + geom\_density(alpha=.5, fill="pink")  
p1



p2 <- ggplot(airline2, aes(x = Incidents\_00\_14, y = ..density.., weight = PAX\_Carrying\_Capacity)) +   
 geom\_histogram(binwidth = 1, color = "black", fill = "white") +   
 labs(x = "Incidents, 2000-2014", y = "Density", title = "Incidents per Trillion Available Seat Kilometers for All Non-U.S. Airlines, 2000-2014") +  
 theme\_classic() + geom\_density(alpha=.5, fill="pink")  
p2



# Non-statistical Graphing - Scatterplots  
  
p3 <- ggplot(airline2, aes(x = Incidents\_85\_99, y = Airline, size = PAX\_Carrying\_Capacity)) +  
 geom\_point(color = "red") +   
 labs(x = "Incidents, 1985-1999", y = "Airline", title = "Incidents per Trillion Available Seat Kilometers per Non-U.S. Airlines, 1985-1999") +   
 theme\_classic()   
p3



p4 <- ggplot(airline2, aes(x = Incidents\_00\_14, y = Airline, size = PAX\_Carrying\_Capacity)) +  
 geom\_point(color = "red") +   
 labs(x = "Incidents, 2000-2014", y = "Airline", title = "Incidents per Trillion Available Seat Kilometers per Non-U.S. Airline, 2000-2014") +   
 theme\_classic()  
p4

