Homework #3

Sarah Derda

Due Date: October 16, 2020

## Question #1: Setting Global Options

knitr::opts\_chunk$set(echo = TRUE,   
 warning=FALSE,   
 message=FALSE,   
 error=FALSE)  
  
library(readr)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.2 ✓ dplyr 1.0.2  
## ✓ tibble 3.0.4 ✓ stringr 1.4.0  
## ✓ tidyr 1.1.2 ✓ forcats 0.5.0  
## ✓ purrr 0.3.4

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

library(ggplot2)  
library(ggthemes)  
library(titanic)  
library(stringr)

## Question #2: Editing HW#2 Plot

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

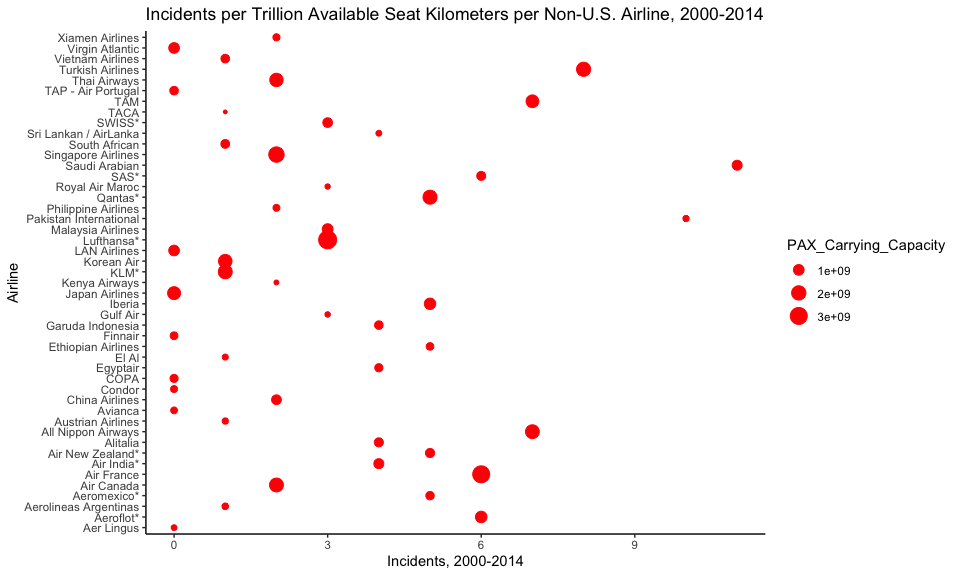
## # A tibble: 6 x 8  
## airline avail\_seat\_km\_p… incidents\_85\_99 fatal\_accidents… fatalities\_85\_99  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 Aer Li… 320906734 2 0 0  
## 2 Aerofl… 1197672318 76 14 128  
## 3 Aeroli… 385803648 6 0 0  
## 4 Aerome… 596871813 3 1 64  
## 5 Air Ca… 1865253802 2 0 0  
## 6 Air Fr… 3004002661 14 4 79  
## # … with 3 more variables: incidents\_00\_14 <dbl>, fatal\_accidents\_00\_14 <dbl>,  
## # fatalities\_00\_14 <dbl>

summary(airline)

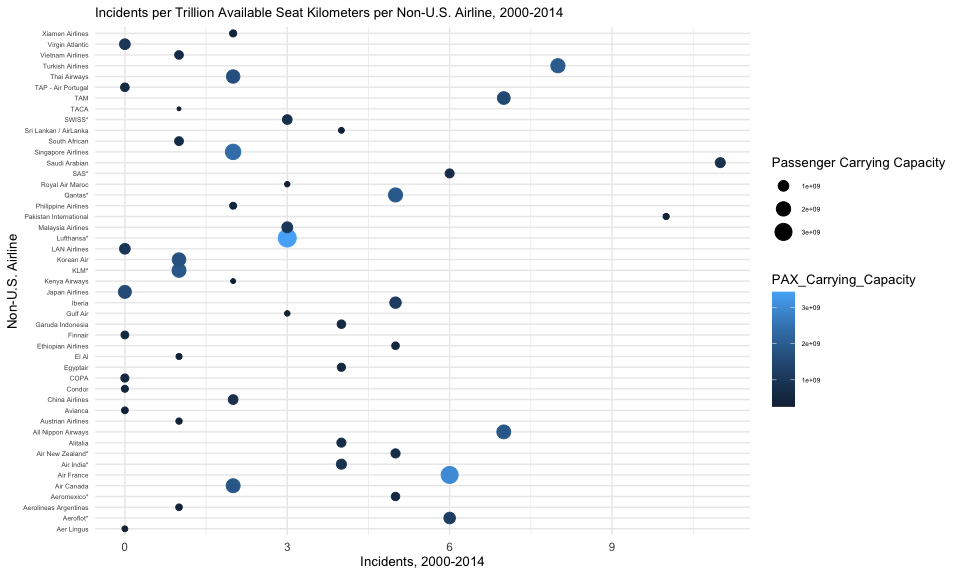
## airline avail\_seat\_km\_per\_week incidents\_85\_99   
## Length:56 Min. :2.594e+08 Min. : 0.000   
## Class :character 1st Qu.:4.740e+08 1st Qu.: 2.000   
## Mode :character Median :8.029e+08 Median : 4.000   
## Mean :1.385e+09 Mean : 7.179   
## 3rd Qu.:1.847e+09 3rd Qu.: 8.000   
## Max. :7.139e+09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

# Cleaning Data  
  
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"))

# Plotting Data  
  
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()



# Editing Plot  
  
ggplot(airline2, aes(x = Incidents\_00\_14, y = Airline,   
 color = PAX\_Carrying\_Capacity,  
 size = PAX\_Carrying\_Capacity)) +  
 geom\_point() +   
 labs(x = "Incidents, 2000-2014", y = "Non-U.S. Airline", title = "Incidents per Trillion Available Seat Kilometers per Non-U.S. Airline, 2000-2014", size = "Passenger Carrying Capacity") +   
 theme\_minimal() +  
 theme(axis.title = element\_text(size = 10),   
 plot.title = element\_text(size = 10),   
 axis.text.y = element\_text(size = 5),  
 legend.title = element\_text(size = 10),  
 legend.text = element\_text(size = 5))



Among the six guidelines for good graphics, the first graph above depicts high data density and using clear, meaningful labels. Aiming for high data density allows the viewer to easily see the message you're trying to convey as a coder. The higher the data to ink ratio is, the less "ink" there will be to process, which takes less time to read overall. Using clear, meaninful labels is also beneficial for readability. Providing descriptive labels for viewers allow them to interpret the graph better by clearly stating what each aspect of the graph conveys.

The second graph was edited. There was a background grid added in order to look at the points more easily between the airlines and how many incidents they had per passenger carrying capacity. The text was also enlarged to make it easier to read. Furthermore, the color of the points were changed in order to depict the size of passenger carrying capacity. Font sizes were also adjusted in order to distuinguish the titles and the general text.

## Question #3: Titanic Dataset

# Import Data & Observe  
  
tt <- titanic\_train  
head(tt)

## PassengerId Survived Pclass  
## 1 1 0 3  
## 2 2 1 1  
## 3 3 1 3  
## 4 4 1 1  
## 5 5 0 3  
## 6 6 0 3  
## Name Sex Age SibSp Parch  
## 1 Braund, Mr. Owen Harris male 22 1 0  
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female 38 1 0  
## 3 Heikkinen, Miss. Laina female 26 0 0  
## 4 Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35 1 0  
## 5 Allen, Mr. William Henry male 35 0 0  
## 6 Moran, Mr. James male NA 0 0  
## Ticket Fare Cabin Embarked  
## 1 A/5 21171 7.2500 S  
## 2 PC 17599 71.2833 C85 C  
## 3 STON/O2. 3101282 7.9250 S  
## 4 113803 53.1000 C123 S  
## 5 373450 8.0500 S  
## 6 330877 8.4583 Q

survdata <- tt %>%  
 as.data.frame() %>%  
 select("Survived", "Age")  
  
class(survdata$Survived)

## [1] "integer"

# Change Variable from Integer to Factor  
  
survdata$Survived <- as.factor(survdata$Survived)  
class(survdata$Survived)

## [1] "factor"

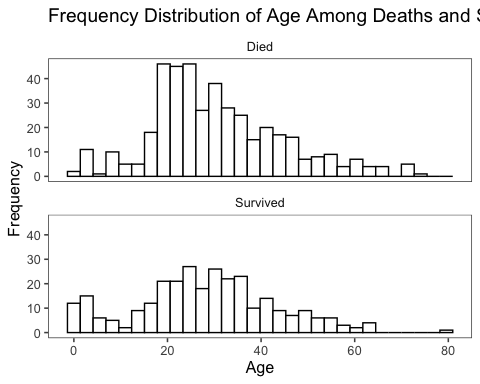
levels(survdata$Survived)

## [1] "0" "1"

# Relabel  
  
survdata <- mutate(.data = survdata,  
 Survived = fct\_recode(.f = Survived,  
 Died = "0",  
 Survived = "1"))  
  
# Top Ages of Survived and Died  
  
top\_surv <- survdata %>%  
 filter(Survived == "Survived") %>%  
 arrange(desc(Age)) %>%  
 slice(1:10)  
  
top\_died <- survdata %>%  
 filter(Survived == "Died") %>%  
 arrange(desc(Age)) %>%  
 slice(1:10)  
  
# Merge to Form 10 Rows  
  
join\_surv <- full\_join(top\_surv, top\_died)  
print(join\_surv)

## Survived Age  
## 1 Survived 80.0  
## 2 Survived 63.0  
## 3 Survived 63.0  
## 4 Survived 62.0  
## 5 Survived 62.0  
## 6 Survived 60.0  
## 7 Survived 60.0  
## 8 Survived 58.0  
## 9 Survived 58.0  
## 10 Survived 58.0  
## 11 Died 74.0  
## 12 Died 71.0  
## 13 Died 71.0  
## 14 Died 70.5  
## 15 Died 70.0  
## 16 Died 70.0  
## 17 Died 66.0  
## 18 Died 65.0  
## 19 Died 65.0  
## 20 Died 65.0

# Plot  
  
ggplot(survdata, aes(x = Age)) +   
 geom\_histogram(fill = "white", color = "black") +   
 facet\_wrap(~ Survived, ncol = 1) +   
 theme\_few() +  
 labs(y = "Frequency",   
 title = "Frequency Distribution of Age Among Deaths and Survivals")



# Two-Row Table Details  
  
survdata\_df <- survdata %>%  
 group\_by(Survived) %>%  
 summarize(N = n(), mean = mean(Age, na.rm = TRUE), na = sum(is.na(Age)))  
print(survdata\_df)

## # A tibble: 2 x 4  
## Survived N mean na  
## <fct> <int> <dbl> <int>  
## 1 Died 549 30.6 125  
## 2 Survived 342 28.3 52

# Stat Test  
  
t.test(survdata$Age[survdata$Survived=="Died"], survdata$Age[survdata$Survived=="Survived"])

##   
## Welch Two Sample t-test  
##   
## data: survdata$Age[survdata$Survived == "Died"] and survdata$Age[survdata$Survived == "Survived"]  
## t = 2.046, df = 598.84, p-value = 0.04119  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 0.09158472 4.47339446  
## sample estimates:  
## mean of x mean of y   
## 30.62618 28.34369

## Question #4: Word Puzzle

words\_alpha\_url <- read\_tsv("https://raw.githubusercontent.com/dwyl/english-words/master/words\_alpha.txt", col\_names = "word")  
  
words\_alpha <- words\_alpha\_url  
  
wordphrase <- words\_alpha %>%  
 filter(str\_detect(word, "^.{8}$")) %>%  
 filter(str\_detect(word, "^[Mm]")) %>%  
 mutate(first = str\_sub(word, 1, 4),  
 second = paste0(str\_sub(word, 6, 8), str\_sub(word, 5, 5))) %>%  
 filter(first %in% words\_alpha$word & second %in% words\_alpha$word)

I was not able to find the wordphrase. With the help of Nick, I was able to reduce the list to 215 words; however, all of them look like they have a different sound after moving the first letter of the second word to the end of the eight word phrase.