ANA 515 Assignment 2

Yatin Pawar

11/12/2021

Description of dataset

This dataset of “Airline safety” from Five Thirty Eight’s GitHub data repository. It is trying to answer the question, “should travelers avoid flying airlines that have had crashes in the past?” Accidents happen either due to human errors or due to natural events. Is it true that past data of accidents make Airline vulnerable? An effort made to study data from Aviation Safety Network’s database of 30 years for 56 airlines. The 30-year period down into two halves: first from 1985 to 1999, and then from 2000 to 2014. The comparison to see if data shows any improvements or degradation of safety measures. If we identify a correlation, that will imply that crash risk is persistent — predictable to some extent based on the airline.

The data was collected in CSV file. The column names are Name of airline, Number of available seat kilometers 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, and Fatalities from 2000 to 2014. Except Airline name, rest of the data is numeric.

```
#Next line url is to get the data from GitHub to R Studio wd.
#using read.csv to read data from csv file
url <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/airline-safety/airline-safety.csv"

airline_safety <- read.csv(url)

#Need to install below libraries to run the functions within them.
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.1
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## Warning: package 'tidyr' was built under R version 4.1.1
```

```
## Warning: package 'readr' was built under R version 4.1.1
```

```
## Warning: package 'purrr' was built under R version 4.1.1
```

```
## Warning: package 'dplyr' was built under R version 4.1.1
```

```
## Warning: package 'stringr' was built under R version 4.1.1
```

```
## Warning: package 'forcats' was built under R version 4.1.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.1.1
```

```
library(bslib)
```

```
## Warning: package 'bslib' was built under R version 4.1.1
```

```
##
## Attaching package: 'bslib'
```

```
## The following object is masked from 'package:utils':
##
##     page
```

```
#Renaming better column names
names(airline_safety)[3]<-"Incidents_1985-99"
names(airline_safety)[4]<-"Fatal_Accidents_1985-99"
names(airline_safety)[5]<-"Fatalities_1985-99"
names(airline_safety)[6]<-"Incidents_2000-14"
names(airline_safety)[7]<-"Fatal_Accidents_2000-14"
names(airline_safety)[8]<-"Fatalities_2000-14"
```

```
#This next chunk is inline code. Inline code puts the text with the output of the function in the document.
```

This dataframe has 56 rows and 8 columns. The names of the columns and a brief description of each are in the table below:

Coulmn Names and Description

```
library(knitr)
col_desc_airline_safety<-data.frame(
Names = c("Airline","avail_seat_km_per_week","Incidents_1985-99","Fatal_Accidents_1985-99","Fatalities_1985-99","Incidents_2
000-14","Fatal_Accidents_2000-14","Fatalities_2000-14"),
  Description = c(
  "Airline (asterisk indicates that regional subsidiaries are included)",
  "Available seat kilometers flown every week",
  "Total number of incidents, 1985–1999",
  "Total number of fatal accidents, 1985–1999","Total number of fatalities, 1985–1999", "Total number of incidents, 2000–201
4", "Total number of fatal accidents, 2000–2014", "Total number of fatalities, 2000–2014")
)

knitr::kable(head(col_desc_airline_safety[, 1:2]), "simple")
```

Names	Description
Airline	Airline (asterisk indicates that regional subsidiaries are included)
avail_seat_km_per_week	Available seat kilometers flown every week
Incidents_1985-99	Total number of incidents, 1985–1999
Fatal_Accidents_1985-99	Total number of fatal accidents, 1985–1999
Fatalities_1985-99	Total number of fatalities, 1985–1999
Incidents_2000-14	Total number of incidents, 2000–2014

```
#Create a new dataset with the name 'safetydata2' from the dataset ‘airline_safety’ that shows columns   "Incidents_2000-1
4","Fatal_Accidents_2000-14","Fatalities_2000-14"
library(dplyr)
safetydata2 <- airline_safety %>%
  select("Incidents_2000-14", "Fatal_Accidents_2000-14","Fatalities_2000-14")
```

Statistical summary of data from year 2000 to 2014 as below:

```
#summary of the dataset safetydata2
sumdata<-summary(safetydata2)

print (sumdata)
```

```
##  Incidents_2000-14 Fatal_Accidents_2000-14 Fatalities_2000-14
##  Min.   : 0.000    Min.   :0.0000          Min.   :  0.00
##  1st Qu.: 1.000    1st Qu.:0.0000          1st Qu.:  0.00
##  Median : 3.000    Median :0.0000          Median :  0.00
##  Mean   : 4.125    Mean   :0.6607          Mean   : 55.52
##  3rd Qu.: 5.250    3rd Qu.:1.0000          3rd Qu.: 83.25
##  Max.   :24.000    Max.   :3.0000          Max.   :537.00
```