ANA 515 Assignment 2

sunil salunke

6/19/2022

Section 1: Description of the data

This data set contains information about airlines and their safety incidents. This data is measuring information about safety incidents, fatal accidents, and fatalities over duration of 1985 to 2014 across 56 airlines. This safety information is divided in 3 ways. Incidents column has information about number of safety incidents with the airline. Fatal accidents column has information about number of fatal accidents. Fatalities column has information about total number of fatalities. This safety information is divided into two time periods i.e. 1985 to 1999 and 2000 to 2014. This data set also has information about available seat km flown every week.

The data source for this information is from aviation-safety.net website. This website is maintained by flight safety foundation. Some of the questions that I would like to answer are, should travelers avoid flying airlines that has crash in the past? Is there a correlation and consistency between number of fatalities from one period to another? What are the number of fatalities by each accident? Is there a correlation and consistency between number of incidents from one period to another?

This data is saved in CSV file format on github. This is a flat file. This is a delimited file and comma is the delimiter used in the file.

Section 2: Reading the data into R

```
#using read.csv to read data from csv file from a URL,
url <-
↪  "https://raw.githubusercontent.com/fivethirtyeight/data/master/airline-safety/airline-safety.csv"
airline_safety_df <- read.csv(url)

#head of output data frame airline safety
knitr::kable(head(airline_safety_df))
```

airline	avail_seat_km_per_week	incidents_85_99	fatal_accidents_85_99	fatalities_85_99	incidents_00_14	fatal_accidents_00_14	fatalities_00_14
Aer Lingus	320906734	2	0	0	0	0	0
Aeroflot*	1197672318	76	14	128	6	1	88
Aerolineas Argentinas	385803648	6	0	0	1	0	0
Aeromexico*	596871813	3	1	64	5	0	0
Air Canada	1865253802	2	0	0	2	0	0
Air France	3004002661	14	4	79	6	2	337

comment: I have used read.csv function to read the data from github url. This is a base r function with in-built functionality to read CSV file.

Section 3: Clean the data

```
#import packages
library(tidyverse)
library(summarytools)
```

```
#let's clean data by creating a separate data frame for 1985-1999
airline_safety_df_85_99 <- select(airline_safety_df,airline, avail_seat_km_per_week,
↪  incidents_85_99, fatal_accidents_85_99, fatalities_85_99)

#let's rename columns incidents_85_99 to incidents, fatal_accidents_85_99 to
↪  fatal_accidents, and fatalities_85_99 to fatalities in airline_safety_df_85_99 data
↪  frame
airline_safety_df_85_99 = airline_safety_df_85_99 %>% rename(incidents = incidents_85_99,
↪  fatal_accidents = fatal_accidents_85_99, fatalities = fatalities_85_99)

#let's check output data frame for 1985-1999 with columns renamed
knitr::kable(head(airline_safety_df_85_99))
```

airline	avail_seat_km_per_week	incidents	fatal_accidents	fatalities
Aer Lingus	320906734	2	0	0
Aeroflot*	1197672318	76	14	128
Aerolineas Argentinas	385803648	6	0	0
Aeromexico*	596871813	3	1	64
Air Canada	1865253802	2	0	0
Air France	3004002661	14	4	79

```
#let's arrange our data by number of incidents in descending manner
airline_safety_df_85_99 <- airline_safety_df_85_99 %>% arrange(desc(incidents))

#let's check output data frame for 1985-1999 with incidents arranged in descending manner
knitr::kable(head(airline_safety_df_85_99))
```

airline	avail_seat_km_per_week	incidents	fatal_accidents	fatalities
Aeroflot*	1197672318	76	14	128
Ethiopian Airlines	488560643	25	5	167
Delta / Northwest*	6525658894	24	12	407
American*	5228357340	21	5	101
United / Continental*	7139291291	19	8	319
US Airways / America West*	2455687887	16	7	224

```
#let's clean  data by creating a separate data frame for 2000-2014
airline_safety_df_00_14 <- select(airline_safety_df,airline, avail_seat_km_per_week,
↪  incidents_00_14, fatal_accidents_00_14, fatalities_00_14)
```

```
#let's rename columns incidents_00_14 to incidents, fatal_accidents_00_14 to
↪  fatal_accidents, and fatalities_00_14 to fatalities in airline_safety_df_00_14 data
↪  frame
airline_safety_df_00_14 = airline_safety_df_00_14 %>% rename(incidents = incidents_00_14,
↪  fatal_accidents = fatal_accidents_00_14, fatalities = fatalities_00_14)

#let's check output data frame for 1985-1999 with columns renamed
knitr::kable(head(airline_safety_df_00_14))
```

airline	avail_seat_km_per_week	incidents	fatal_accidents	fatalities
Aer Lingus	320906734	0	0	0
Aeroflot*	1197672318	6	1	88
Aerolineas Argentinas	385803648	1	0	0
Aeromexico*	596871813	5	0	0
Air Canada	1865253802	2	0	0
Air France	3004002661	6	2	337

Section 4: Characteristics of the data

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

A descriptive table:

```
library(knitr)
columns_summary <- data.frame(
Columns = c(colnames(airline_safety_df)),
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, 2000-2014", "Total number of fatal accidents, 2000-2014", "Total
↪  number of fatalities, 2000-2014")
)

kable(columns_summary, caption = "column_summary")
```

Table 5: column_summary

Columns	Description
airline	Airline (asterisk indicates that regional subsidiaries are included)
avail_seat_km_per_week	Available seat kilometers flown every week
incidents_85_99	Total number of incidents, 1985–1999
fatal_accidents_85_99	Total number of fatal accidents, 1985–1999
fatalities_85_99	Total number of fatalities, 1985–1999
incidents_00_14	Total number of incidents, 2000–2014

Columns	Description
fatal_accidents_00_14	Total number of fatal accidents, 2000–2014
fatalities_00_14	Total number of fatalities, 2000–2014

Section 5: Summary statistics

Subset the dataset:

Picking three columns to use summary function:

```
airline_safety_df_subset <- select(airline_safety_df, incidents_85_99,
↪  fatal_accidents_85_99, fatalities_85_99)
```

Produce a summary of the subset: checking for missing values

```
kable(airline_safety_df_subset %>%
  summarise_all(list(missing_count = ~ sum(is.na(.)))))
```

incidents_85_99_missing_count	fatal_accidents_85_99_missing_count	fatalities_85_99_missing_count
0	0	0

There are no missing values in above 3 columns.

Produce a summary of the subset:summary statistics

```
Summarytable<-summary(airline_safety_df_subset) #creates the summary
Summarytable #prints the summary in the output
```

```
##  incidents_85_99  fatal_accidents_85_99 fatalities_85_99
##  Min.   : 0.000   Min.   : 0.000        Min.   :  0.0
##  1st Qu.: 2.000   1st Qu.: 0.000        1st Qu.:  0.0
##  Median : 4.000   Median : 1.000        Median : 48.5
##  Mean   : 7.179   Mean   : 2.179        Mean   :112.4
##  3rd Qu.: 8.000   3rd Qu.: 3.000        3rd Qu.:184.2
##  Max.   :76.000   Max.   :14.000        Max.   :535.0
```

Above is the summary of these 3 columns.

There are no missing values in above 3 columns.

For 1985–1999 period, across different airlines, total number of minimum incidents are 0 and total number of maximum incidents are 76.

For 1985–1999 period, across different airlines, total number of minimum fatal accidents are 0 and total number of maximum fatal accidents are 14.

For 1985–1999 period, across different airlines, total number of minimum fatalities are 0 and total number of maximum fatalities are 535.