Week5_Assignment

Heleine Fouda

2023-10-07

This assignment will consist in three main tasks. First, we will create a .CSV file from the chart provided. Then, we will read the information from our .CSV file into R, and perform some data cleaning and data transformation as needed. Finally, we will use the cleaned data to analyze and compare the arrival delays of the two airline carriers listed in the data set.

Setting up our Environment

Let's first set up our work environment.

The Data Set

Let's create a .CSV file from the provided pdf document.

```
#  Creating a .CSV file
flights_data<- data.frame(
Cities= c("Los Angeles", "Phoenix", "San Diego", "San Francisco", "Seattle"),
  Alaska_on_time = c(497, 221,  212, 503, 1841),
  Alaska_delayed = c(62, 12, 20, 102, 305),
  Am_west_on_time = c(694, 4840, 383, 320, 201),
  Am_west_delayed = c(117,415, 65, 129, 61)
)

file_path <- "flights_data.csv"

write.csv(flights_data, file = file_path, row.names = FALSE)

cat("Data has been saved to", file_path, "\n")
```

```
## Data has been saved to flights_data.csv
```

Let's read the .CSV file information into R

```
# Reading the .CSV file info into R
flights_data <- read.csv("flights_data.csv", sep = ",")

flights_data
```

```
##          Cities Alaska_on_time Alaska_delayed Am_west_on_time Am_west_delayed
## 1   Los Angeles            497             62            694             117
## 2       Phoenix            221             12           4840             415
## 3     San Diego            212             20            383              65
## 4 San Francisco            503            102            320             129
## 5       Seattle           1841            305            201              61
```

Data exploration

Let's take a peek at the data set

```
# Glimpse of the data set
glimpse(flights_data)
```

```
## Rows: 5
## Columns: 5
## $ Cities          <chr> "Los Angeles", "Phoenix", "San Diego", "San Francisco"~
## $ Alaska_on_time  <int> 497, 221, 212, 503, 1841
## $ Alaska_delayed  <int> 62, 12, 20, 102, 305
## $ Am_west_on_time <int> 694, 4840, 383, 320, 201
## $ Am_west_delayed <int> 117, 415, 65, 129, 61
```

```
#  Data Head
head(flights_data)
```

```
##          Cities Alaska_on_time Alaska_delayed Am_west_on_time Am_west_delayed
## 1   Los Angeles            497             62             694             117
## 2       Phoenix            221             12            4840             415
## 3     San Diego            212             20             383              65
## 4 San Francisco            503            102             320             129
## 5       Seattle           1841            305             201              61
```

```
# | Data Structure
str(flights_data)
```

```
## 'data.frame':    5 obs. of  5 variables:
##  $ Cities         : chr  "Los Angeles" "Phoenix" "San Diego" "San Francisco" ...
##  $ Alaska_on_time : int  497 221 212 503 1841
##  $ Alaska_delayed : int  62 12 20 102 305
##  $ Am_west_on_time: int  694 4840 383 320 201
##  $ Am_west_delayed: int  117 415 65 129 61
```

Below is the summary statistics

```
# Data summary
summary(flights_data)
```

```
##     Cities          Alaska_on_time   Alaska_delayed  Am_west_on_time
##  Length:5           Min.   : 212.0   Min.   : 12.0   Min.   : 201
##  Class :character   1st Qu.: 221.0   1st Qu.: 20.0   1st Qu.: 320
##  Mode  :character   Median : 497.0   Median : 62.0   Median : 383
##                     Mean   : 654.8   Mean   :100.2   Mean   :1288
##                     3rd Qu.: 503.0   3rd Qu.:102.0   3rd Qu.: 694
##                     Max.   :1841.0   Max.   :305.0   Max.   :4840
##  Am_west_delayed
##  Min.   : 61.0
##  1st Qu.: 65.0
##  Median :117.0
##  Mean   :157.4
##  3rd Qu.:129.0
##  Max.   :415.0
```

Data Transformation and Data Cleaning

Let's new columns or variables to our data frame

```
# Adding On-time column
New_flights<-flights_data|>
  select(Cities, Alaska_on_time, Am_west_on_time,  Alaska_delayed,Am_west_delayed)|>
  mutate(On_time = (Am_west_on_time) + (Alaska_on_time))
New_flights
```

```
##          Cities Alaska_on_time Am_west_on_time Alaska_delayed Am_west_delayed
## 1   Los Angeles            497             694             62             117
## 2       Phoenix            221            4840             12             415
## 3     San Diego            212             383             20              65
## 4 San Francisco            503             320            102             129
## 5       Seattle           1841             201            305              61
##   On_time
## 1    1191
## 2    5061
## 3     595
## 4     823
## 5    2042
```

```
# Adding the Delayed column
new_flights <-New_flights|>
mutate(Delayed = ( Alaska_delayed) + (Am_west_delayed))
new_flights
```

```
##          Cities Alaska_on_time Am_west_on_time Alaska_delayed Am_west_delayed
## 1   Los Angeles            497             694             62             117
## 2       Phoenix            221            4840             12             415
## 3     San Diego            212             383             20              65
## 4 San Francisco            503             320            102             129
## 5       Seattle           1841             201            305              61
##   On_time Delayed
## 1    1191     179
## 2    5061     427
## 3     595      85
## 4     823     231
## 5    2042     366
```

```
# Adding the Airlines column
new_flights <-New_flights|>
mutate(Airlines = ( Alaska_delayed) + (Am_west_delayed) + (Alaska_on_time) + (Am_west_on_time))
new_flights
```

```
##          Cities Alaska_on_time Am_west_on_time Alaska_delayed Am_west_delayed
## 1   Los Angeles            497             694             62             117
## 2       Phoenix            221            4840             12             415
## 3     San Diego            212             383             20              65
## 4 San Francisco            503             320            102             129
## 5       Seattle           1841             201            305              61
##   On_time Airlines
## 1    1191     1370
## 2    5061     5488
## 3     595      680
## 4     823     1054
## 5    2042     2408
```

```r
#  Delayed Flights

flights_delay <- new_flights|>
  group_by(Cities) |>
  reframe(Total_delayed = flights_data$Alaska_delayed + flights_data$Am_west_delayed)

flights_delay
```

```
## # A tibble: 25 x 2
##    Cities      Total_delayed
##    <chr>               <int>
##  1 Los Angeles           179
##  2 Los Angeles           427
##  3 Los Angeles            85
##  4 Los Angeles           231
##  5 Los Angeles           366
##  6 Phoenix               179
##  7 Phoenix               427
##  8 Phoenix                85
##  9 Phoenix               231
## 10 Phoenix               366
## # i 15 more rows
```

```r
# Delayed flights arranged in descending order
flights_delay <- new_flights|>
  group_by(Cities) |>
  reframe(Total_delayed = flights_data$Alaska_delayed + flights_data$Am_west_delayed)|>
  arrange(desc(Total_delayed))

flights_delay
```

```
## # A tibble: 25 x 2
##    Cities        Total_delayed
##    <chr>                 <int>
##  1 Los Angeles             427
##  2 Phoenix                 427
##  3 San Diego               427
##  4 San Francisco           427
##  5 Seattle                 427
##  6 Los Angeles             366
##  7 Phoenix                 366
##  8 San Diego               366
##  9 San Francisco           366
## 10 Seattle                 366
## # i 15 more rows
```

```r
# On-time flights
flights_ont <- new_flights|>
  group_by(Cities) |>
  reframe(Total_ontime = flights_data$Alaska_on_time+ flights_data$Am_west_on_time)

flights_ont
```

```
## # A tibble: 25 x 2
##    Cities      Total_ontime
```

```
##    <chr>              <int>
##  1 Los Angeles         1191
##  2 Los Angeles         5061
##  3 Los Angeles          595
##  4 Los Angeles          823
##  5 Los Angeles         2042
##  6 Phoenix             1191
##  7 Phoenix             5061
##  8 Phoenix              595
##  9 Phoenix              823
## 10 Phoenix             2042
## # i 15 more rows
```

Let's find the number of on-time flights and delayed flight per carrier or airline

```
# Delayed and on- time flights - per airline

flights_data <-flights_data|>
  mutate(Delayed =  Am_west_delayed + Alaska_delayed
)
flights_data
```

```
##          Cities Alaska_on_time Alaska_delayed Am_west_on_time Am_west_delayed
## 1   Los Angeles            497             62             694             117
## 2       Phoenix            221             12            4840             415
## 3     San Diego            212             20             383              65
## 4 San Francisco            503            102             320             129
## 5       Seattle           1841            305             201              61
##   Delayed
## 1     179
## 2     427
## 3      85
## 4     231
## 5     366
```

Let's now put the data into a format that makes the analysis easier

```
# Pivot-longer
pivot_longer(flights_data, cols = 2:5, names_to = "Airline", values_to = "Count")
```

```
## # A tibble: 20 x 4
##    Cities        Delayed Airline         Count
##    <chr>           <int> <chr>           <int>
##  1 Los Angeles       179 Alaska_on_time    497
##  2 Los Angeles       179 Alaska_delayed     62
##  3 Los Angeles       179 Am_west_on_time   694
##  4 Los Angeles       179 Am_west_delayed   117
##  5 Phoenix           427 Alaska_on_time    221
##  6 Phoenix           427 Alaska_delayed     12
##  7 Phoenix           427 Am_west_on_time  4840
##  8 Phoenix           427 Am_west_delayed   415
##  9 San Diego          85 Alaska_on_time    212
## 10 San Diego          85 Alaska_delayed     20
## 11 San Diego          85 Am_west_on_time   383
## 12 San Diego          85 Am_west_delayed    65
## 13 San Francisco     231 Alaska_on_time    503
## 14 San Francisco     231 Alaska_delayed    102
```

```
## 15 San Francisco     231 Am_west_on_time   320
## 16 San Francisco     231 Am_west_delayed   129
## 17 Seattle           366 Alaska_on_time   1841
## 18 Seattle           366 Alaska_delayed    305
## 19 Seattle           366 Am_west_on_time   201
## 20 Seattle           366 Am_west_delayed    61
```

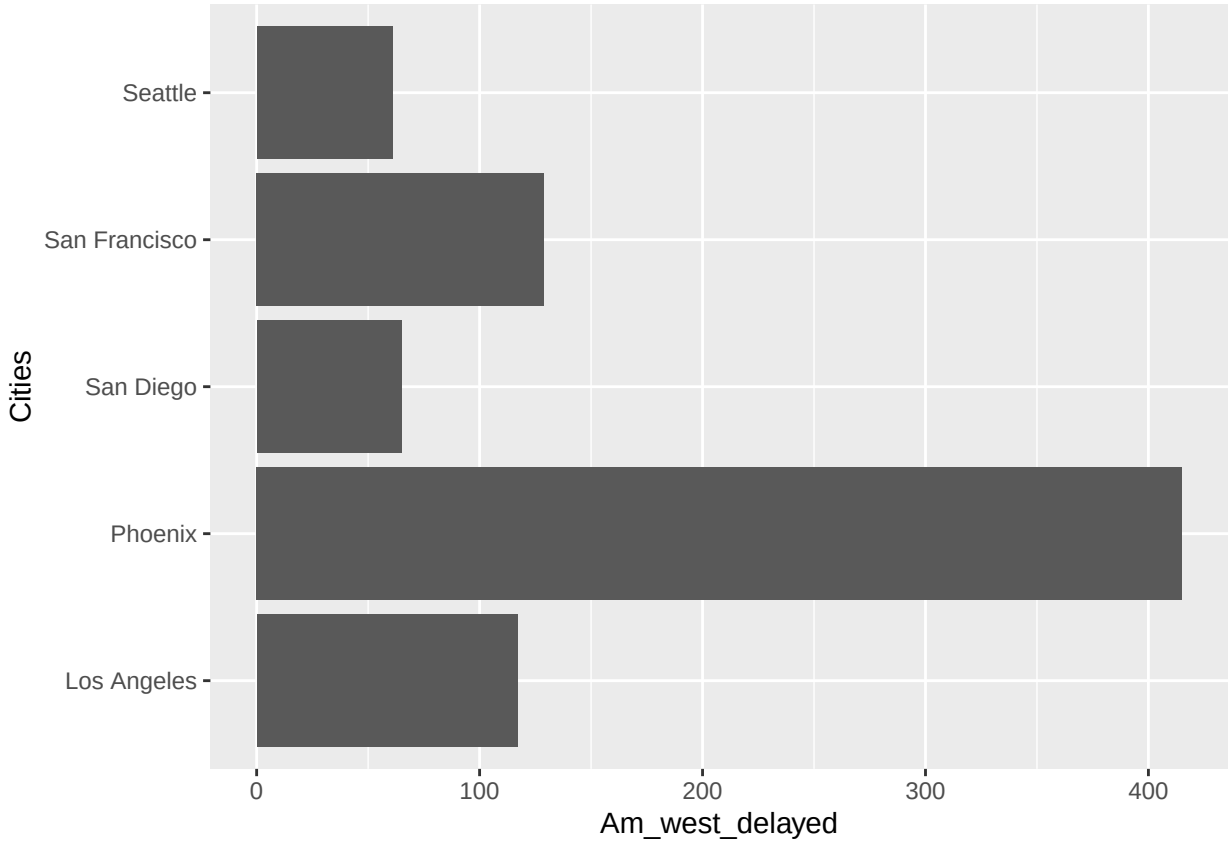
```
flights_data <- c("Cities", "Airline", "Delayed", "Count")

flights_data
```

```
## [1] "Cities"  "Airline" "Delayed" "Count"
```

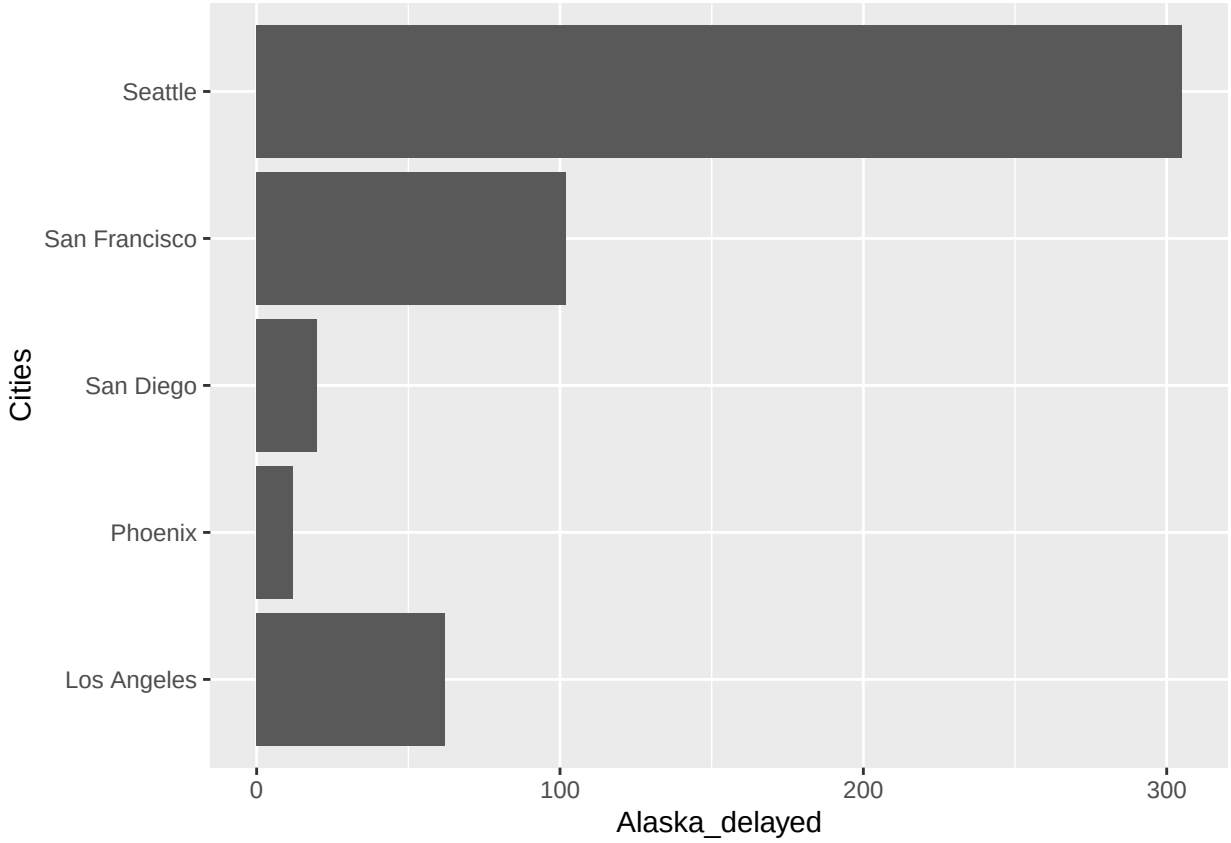
Data Visualization

```
# AM_West delays- by -Cities
df <- data.frame(new_flights)
  ggplot(data = df, aes(x = Am_west_delayed, y = Cities))+
  geom_bar(stat="identity")
```




```
# Alaska delays by cities

ggplot(data = new_flights, aes(x = Alaska_delayed, y = Cities))+
  geom_bar(stat="identity")
```



Findings & Conclusion

From the charts above, one sees that San Francisco is the Destination where both Alaska and AM West seems to experience an almost similar level of delays.