Week5 Assignment

Vishal Arora

March 2, 2019

Introduction

Data manipulation is one of the most important and critical part of Data Science.It is the 3rd step in the overall process of CRISP-DM.

> *CRISP-DM* : **Cross Industry Standard Process for Data Mining.**

Data preperation/manipulation is the process where data is rearranged, manipulated and prepared for the Analysis to be fed into Model.

Problem statment

We have been given data for 2 airlines which fly to certain cities with thier arrival times, whether OnTime or Delayed.Create a csv file with the data and manipulate the data to do analysis and infer about the delay arrivals for both airlines and summarize the same at end.

Solution

The R packages used for the solution are as below.
* **dplyr**
* **tidyr**
* **stringr**
* **graphics**
* **kableExtra**

Using read.csv function we populated df_airlines from airlines csv. *KableExtra* package was used for styling the table to display the loaded airline data in tabular form. Gather funtion from tidyr package was used for transforming wide table structure to long table structure. We gathered coulmns "City" and "NoofTimes".

Next Using *arrange* function from dplyr package, we sorted data on column *'Airlines'* As analysis needs to done on arrival delayed for both airlines for each city,using spread function from *tidyr* package on the arrival column to get two different columns for Ontime and delayed .Using *mutate* function we added three new columns to the table one for percentage delayed , percentage of flights on Time and one for total number of flights.

All this functions gave as final table structure to draw analysis graphs. As graphs needs to be plotted for two different airlines we *subset* table for two airlines. Alaska and AM West.

```
df_airline <- read.csv("airlines.csv",header = TRUE,sep=",")

kable(df_airline) %>%
  kable_styling(bootstrap_options = c("striped","hover","condensed","responsive"),full_width   = F,posi
  row_spec(0, background ="gray")
```

Airlines	Arrival	LosAngles	Phoenix	San_Dieago	Sanfrancisco	Seatle
Alaska	OnTime	497	221	212	503	1841
Alaska	Delayed	62	12	20	102	305
AM West	OnTime	694	4840	383	320	201
AM West	Delayed	117	415	65	129	61

```
longData <- gather(df_airline,"city","NoOfTimes", 3:7)
longData <- arrange(longData,Airlines)
finaltabDF <- spread(longData,Arrival,4)

finaltabDF <- select(finaltabDF , 1:4 ) %>%
              mutate(TotalPerRow = (Delayed + OnTime)) %>%
              mutate(PercDelayed = (Delayed/TotalPerRow)*100) %>%
              mutate(PercOnTime = (OnTime/TotalPerRow)*100)

finaltabDF <- select(finaltabDF,1:7,-5)
alaskaDF <- subset(finaltabDF, finaltabDF$Airlines == 'Alaska')
westDF <-  subset(finaltabDF, finaltabDF$Airlines == 'AM West')

kable(finaltabDF) %>%
  kable_styling(bootstrap_options = c("striped","hover","condensed","responsive"),full_width   = F,posi
  row_spec(0, background ="gray")
```

Airlines	city	Delayed	OnTime	PercDelayed	PercOnTime
Alaska	LosAngles	62	497	11.091234	88.90877
Alaska	Phoenix	12	221	5.150215	94.84979
Alaska	San_Dieago	20	212	8.620690	91.37931
Alaska	Sanfrancisco	102	503	16.859504	83.14050
Alaska	Seatle	305	1841	14.212488	85.78751
AM West	LosAngles	117	694	14.426634	85.57337
AM West	Phoenix	415	4840	7.897241	92.10276
AM West	San_Dieago	65	383	14.508929	85.49107
AM West	Sanfrancisco	129	320	28.730512	71.26949
AM West	Seatle	61	201	23.282443	76.71756

Using kable to have the sumamry for both airlines DF in table, but instead of displaying we used save_kable function to save the output in html format when the .Rmd file is run in R in the working directory.

```
kable(summary(alaskaDF)) %>%
  kable_styling(bootstrap_options = c("striped","condensed"),full_width = F,position =
  "left",font_size = 12) %>%
  row_spec(0, background ="gray") %>%
  save_kable(file = "SummaryForAlaskaAirlines.pdf", self_contained = T)

kable(summary(westDF)) %>%
  kable_styling(bootstrap_options = c("striped","condensed"),full_width = F,position =
  "left",font_size = 12) %>%
  row_spec(0, background ="gray") %>%
  save_kable(file = "SummaryForAMWestAirlines.pdf", self_contained = T)
```

Using Barplot function bar graphs is plotted for two airlines (Alaska , MD West) for five diferent cites where percentage of delayed in on Y axis and cities on X axis.

```
Val <-matrix(c(alaskaDF$PercDelayed,westDF$PercDelayed),nrow=2,ncol=5,byrow=TRUE)
colnames(Val) <- alaskaDF$city

barplot(Val,main="Alaska vs AM West Airlines", names.arg=c(alaskaDF$city),xlab="Cities", ylab="% Arrival

legend("topleft", c("Alaska","AM West"), fill = c("Grey","cyan"))
```

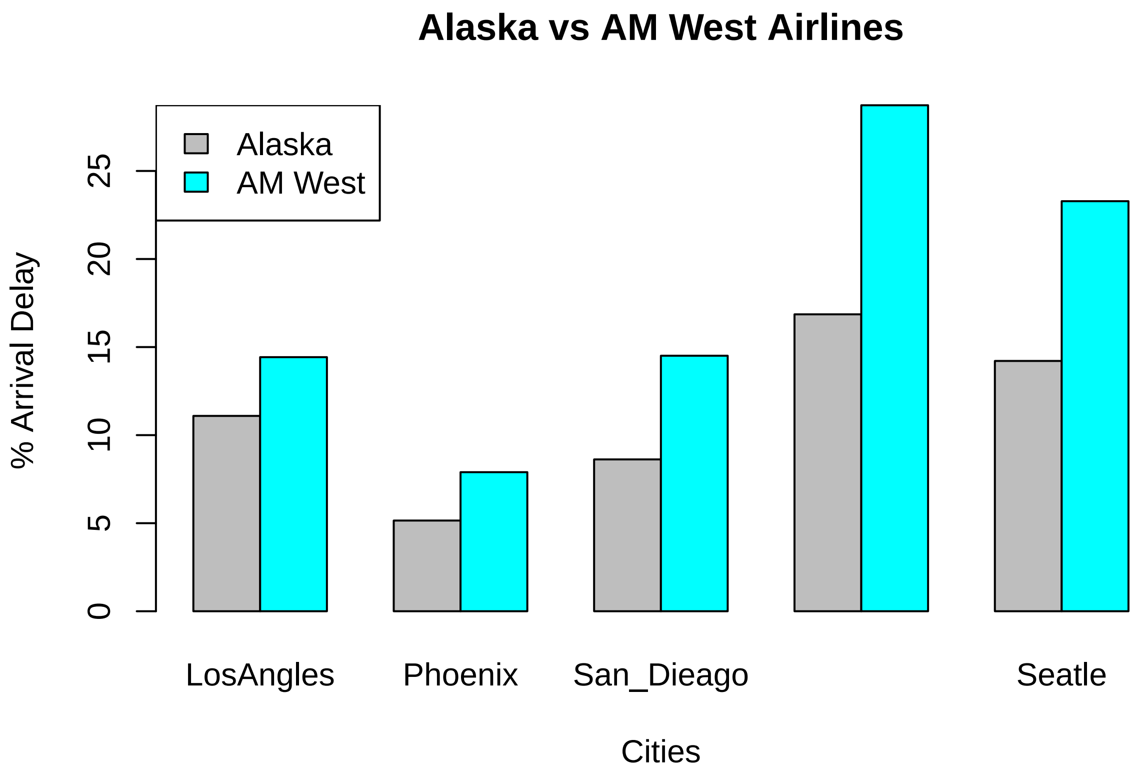

Figure 1: Fig1:Arrival delays in Percentage for both Airlines.

Summary

By Looking at the summary for both the Alaska & AM West airlines dataframe(s) it is clearly visible that the mean %arrival delay for Alsaka airline is less that AM West airline. This is also clearly visible from the Bar chart drawn above where we can visually compare that the % arrival delay for both airlines for each city and we can safely conclude that AM West airline has more delays compared Alaska airline..