Data607 - Assignment5

Amit Kapoor

2/29/2020

Introduction

Data manipulation is one of the most important part of Data Science. The purpose of this assignment is to perform data manipulation using R packages tidyr and dplyr. Data manipulation involves data rearrangement, manipulation and its analysis to make it ready for applicable model.

Problem Statement

We have been provided the data for 2 airlines that describes arrival delays for both of them across five destinations.The task is to create a csv file with the given data and use R libraries tidyr and dplyr as needed to tidy and transform data and then perform analysis to compare the arrival delays for the two airlines.

Solution

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

Using read.csv function we populated flights_df from my github repository https://raw.githubusercontent.com/amit-kapoor/data607/master/week5/flighdetails.csv. We first dropped the blank row from data and then used gather function from tidyr package to gather data in City and Flight Count and then used arrange function from dplyr package by Airline.

Next using spread function from tidyr package, spread the data along arrival to make it wide and then rename columns. Then We used mutate function from dplyr package created new columns Delayed_Perc and OnTime_Perc columns.

We used all these functions to have a 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.

```
# read the data from csv
flights_df <- read.csv("https://raw.githubusercontent.com/amit-kapoor/data607/master/week5/flighdetails
flights_df
```

```
##         X     X.1 Los.Angeles Phoenix San.Diego San.Francisco Seattle
## 1  Alaska on time         497     221       212           503    1841
## 2         delayed          62      12        20           102     305
## 3                          NA      NA        NA            NA      NA
## 4 AM West on time         694    4840       383           320     201
## 5         delayed         117     415        65           129      61
```

We do have NAs in the 3rd row. Since its a small dataset, we can simply delete the NAs from 3rd row.

```
# drop row having NAs and rename columns
flights_df <- flights_df %>%
  drop_na() %>%
  rename(Airline=X, Arrival=X.1, "Los Angeles"=Los.Angeles, "San Diego"=San.Diego, "San Francisco"=San.

flights_df
```

```
##   Airline Arrival Los Angeles Phoenix San Diego San Francisco Seattle
## 1  Alaska on time         497     221       212           503    1841
## 2         delayed          62      12        20           102     305
```

```
## 4 AM West on time         694    4840       383           320      201
## 5         delayed         117     415        65           129       61
```

From above its visible that the rows with delayed arrival do not have airline populated. Here we know the 2nd row is for Alaska and 4th for AM West.

```
# Replace the blank value in the 2nd row Airline column with Alaska
flights_df$Airline[2] <- "Alaska"
# Replace the blank value in the 4th row Airline column with AM West
flights_df$Airline[4] <- "AM West"

flights_df
```

```
##   Airline Arrival Los Angeles Phoenix San Diego San Francisco Seattle
## 1  Alaska on time         497     221       212           503    1841
## 2  Alaska delayed          62      12        20           102     305
## 4 AM West on time         694    4840       383           320     201
## 5 AM West delayed         117     415        65           129      61
```

In the next few steps we are going to make data available for analysis using gather, arrange, spread, rename and mutate functions. We will add 2 new columns for %delay and %ontime analysis.

```
# Gather data in City and Flight Count and then arrange by Airline
flights_df <- flights_df %>%
  gather("City", "Flight_Count", 3:7) %>%
  arrange(Airline)

flights_df
```

```
##    Airline Arrival          City Flight_Count
## 1   Alaska on time   Los Angeles          497
## 2   Alaska delayed   Los Angeles           62
## 3   Alaska on time       Phoenix          221
## 4   Alaska delayed       Phoenix           12
## 5   Alaska on time     San Diego          212
## 6   Alaska delayed     San Diego           20
## 7   Alaska on time San Francisco          503
## 8   Alaska delayed San Francisco          102
## 9   Alaska on time       Seattle         1841
## 10  Alaska delayed       Seattle          305
## 11 AM West on time   Los Angeles          694
## 12 AM West delayed   Los Angeles          117
## 13 AM West on time       Phoenix         4840
## 14 AM West delayed       Phoenix          415
## 15 AM West on time     San Diego          383
## 16 AM West delayed     San Diego           65
## 17 AM West on time San Francisco          320
## 18 AM West delayed San Francisco          129
## 19 AM West on time       Seattle          201
## 20 AM West delayed       Seattle           61
```

```
# spread along arrival to make it wide and then rename columns

flights_df <- flights_df %>%
  spread("Arrival", "Flight_Count") %>%
  rename(Delayed="delayed", OnTime="on time")
```

```
flights_df
```

```
##    Airline          City Delayed OnTime
## 1   Alaska   Los Angeles      62    497
## 2   Alaska       Phoenix      12    221
## 3   Alaska     San Diego      20    212
## 4   Alaska San Francisco     102    503
## 5   Alaska       Seattle     305   1841
## 6  AM West   Los Angeles     117    694
## 7  AM West       Phoenix     415   4840
## 8  AM West     San Diego      65    383
## 9  AM West San Francisco     129    320
## 10 AM West       Seattle      61    201
```

```
# mutating new columns Delayed_Perc and OnTime_Perc
flights_df <- flights_df %>%
  mutate(total_count = Delayed+OnTime) %>%
  mutate(Delayed_Perc = (Delayed/total_count)*100) %>%
  mutate(OnTime_Perc = (OnTime/total_count)*100) %>%
  select(1:7, -5)       # dropped total_count column

flights_df
```

```
##    Airline          City Delayed OnTime Delayed_Perc OnTime_Perc
## 1   Alaska   Los Angeles      62    497    11.091234    88.90877
## 2   Alaska       Phoenix      12    221     5.150215    94.84979
## 3   Alaska     San Diego      20    212     8.620690    91.37931
## 4   Alaska San Francisco     102    503    16.859504    83.14050
## 5   Alaska       Seattle     305   1841    14.212488    85.78751
## 6  AM West   Los Angeles     117    694    14.426634    85.57337
## 7  AM West       Phoenix     415   4840     7.897241    92.10276
## 8  AM West     San Diego      65    383    14.508929    85.49107
## 9  AM West San Francisco     129    320    28.730512    71.26949
## 10 AM West       Seattle      61    201    23.282443    76.71756
```

Data Analysis

For Data analysis, subsetting the data for both airlines Alaska and AM West for further analysis.

```
# create dataframe for Alaska
alaska_df <- subset(flights_df, flights_df$Airline == "Alaska")

# create dataframe for AM West
amwest_df <- subset(flights_df, flights_df$Airline == "AM West")
```

```
# summary for alaska airline data
summary(alaska_df)
```

```
##     Airline      City              Delayed          OnTime
##         :0   Length:5           Min.   : 12.0   Min.   : 212.0
##  Alaska :5   Class :character   1st Qu.: 20.0   1st Qu.: 221.0
##  AM West:0   Mode  :character   Median : 62.0   Median : 497.0
##                                 Mean   :100.2   Mean   : 654.8
##                                 3rd Qu.:102.0   3rd Qu.: 503.0
##                                 Max.   :305.0   Max.   :1841.0
##   Delayed_Perc     OnTime_Perc
```

```
##  Min.   : 5.150   Min.   :83.14
##  1st Qu.: 8.621   1st Qu.:85.79
##  Median :11.091   Median :88.91
##  Mean   :11.187   Mean   :88.81
##  3rd Qu.:14.212   3rd Qu.:91.38
##  Max.   :16.860   Max.   :94.85
```

```
# summary for am west airline data
summary(amwest_df)
```

```
##     Airline      City              Delayed          OnTime
##         :0   Length:5           Min.   : 61.0   Min.   : 201
##  Alaska :0   Class :character   1st Qu.: 65.0   1st Qu.: 320
##  AM West:5   Mode  :character   Median :117.0   Median : 383
##                                 Mean   :157.4   Mean   :1288
##                                 3rd Qu.:129.0   3rd Qu.: 694
##                                 Max.   :415.0   Max.   :4840
##   Delayed_Perc     OnTime_Perc
##  Min.   : 7.897   Min.   :71.27
##  1st Qu.:14.427   1st Qu.:76.72
##  Median :14.509   Median :85.49
##  Mean   :17.769   Mean   :82.23
##  3rd Qu.:23.282   3rd Qu.:85.57
##  Max.   :28.731   Max.   :92.10
```

Analysis for Delayed Arrival

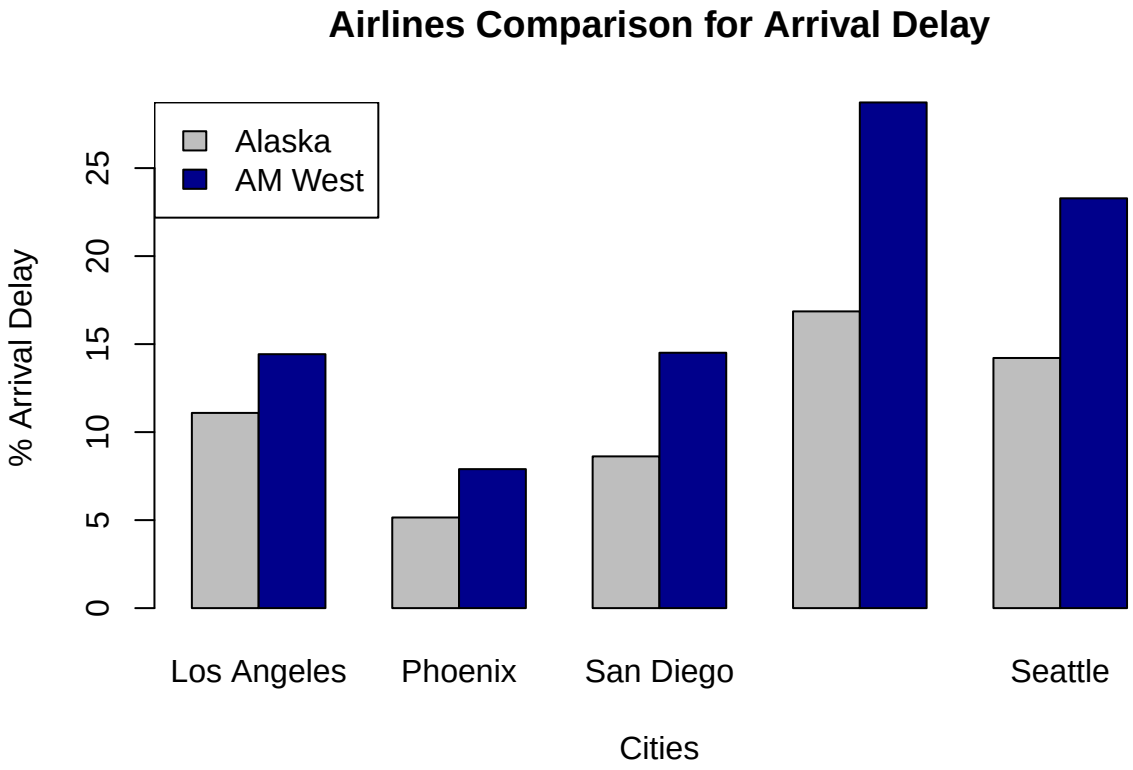
Will all the data munging performed above, we are ready to draw data analysis now for Delayed arrival by seeing side by side comparison for both the airlines. First we created matrix for delayed percentage data for both the airlines and then draw graphs to further analyze.

```
# create matrix with Delayed_Perc by Airline
arrival_delay <- matrix(c(alaska_df$Delayed_Perc,
                      amwest_df$Delayed_Perc), nrow = 2, ncol = 5, byrow = T)

# rename columns as City Names
colnames(arrival_delay) <- amwest_df$City

arrival_delay
```

```
##      Los Angeles  Phoenix San Diego San Francisco  Seattle
## [1,]    11.09123 5.150215   8.62069      16.85950 14.21249
## [2,]    14.42663 7.897241  14.50893      28.73051 23.28244
```

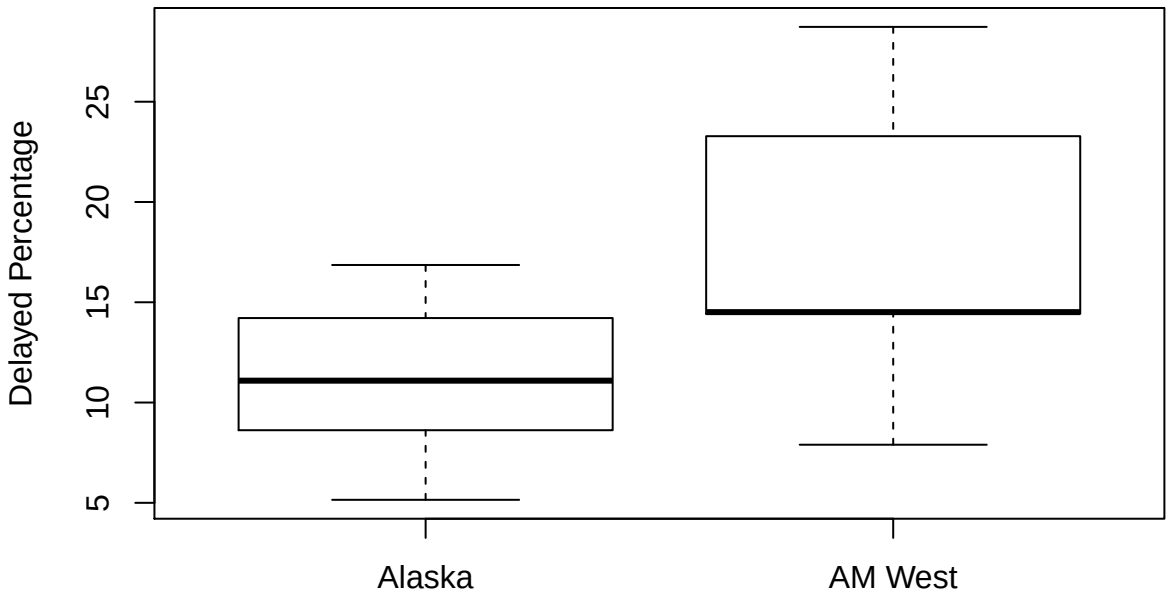
Using barplot function bar graph is plotted for side by side comparison for Arrival Delay of both airlines.

```
# Draw barplot
barplot(arrival_delay,
        main= "Airlines Comparison for Arrival Delay",
        beside=TRUE,
        legend.text=TRUE,
        col = c("grey", "darkblue"),
        xlab = "Cities",
        ylab = "% Arrival Delay")
legend("topleft", legend=c("Alaska", "AM West"), fill = c("grey", "darkblue"))
```



Using boxplot function bar graph for Arrival Delay of both airlines, we can see the delayed percentage distribution of airlines.

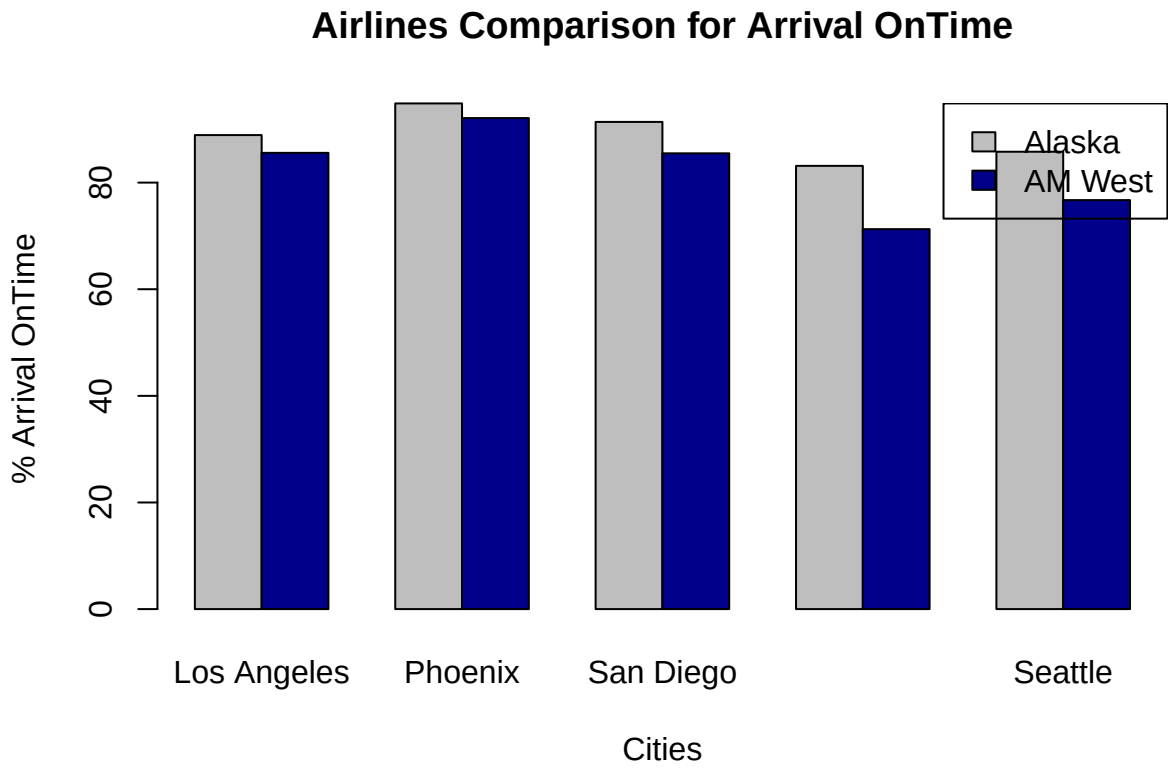
```
boxplot(alaska_df$Delayed_Perc,
        amwest_df$Delayed_Perc,
        names = c("Alaska", "AM West"),
        ylab="Delayed Percentage")
```



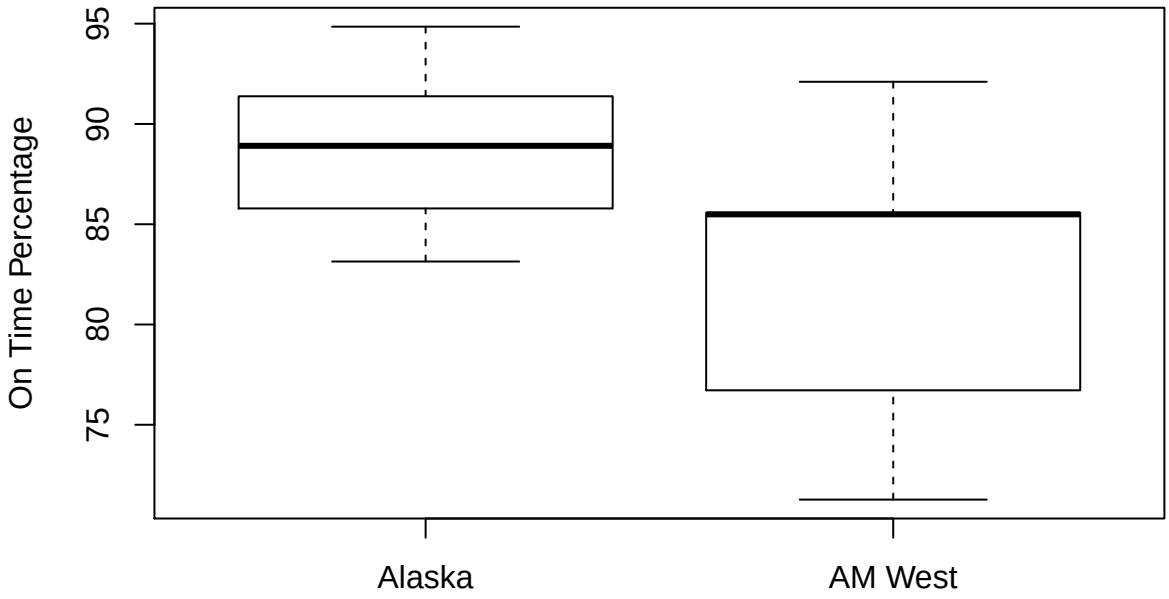
Analysis for On Time Arrival

Here is an additional analysis for OnTime arrical for both the airlines. Similar seps has been performed to create matrix for on time arrival data and then drawn the barplot and boxplot.

```
# create matrix with OnTime_Perc by Airline
arrival_ontime <- matrix(c(alaska_df$OnTime_Perc,
                      amwest_df$OnTime_Perc), nrow = 2, ncol = 5, byrow = T)

# rename columns as City Names
colnames(arrival_ontime) <- amwest_df$City

# Draw barplot
barplot(arrival_ontime,
        main= "Airlines Comparison for Arrival OnTime",
        beside=TRUE,
        legend.text=TRUE,
        col = c("grey", "darkblue"),
        xlab = "Cities",
        ylab = "% Arrival OnTime")
legend("topright", legend=c("Alaska", "AM West"), fill = c("grey", "darkblue"))
```




```
# boxplot for ontime arrival data
boxplot(alaska_df$OnTime_Perc,
        amwest_df$OnTime_Perc,
        names = c("Alaska", "AM West"),
        ylab="On Time Percentage")
```



Summary/Conclusion

By analyzing the data for both the Alaska & AM West airlines, it is clear that the mean % arrival delay for Alsaka airline (62) is less that of AM West airline (117). Also from the Bar chart drawn above we can conclude that AM West airline has more % delays as compared to Alaska airline for all the cities. San Francisco has highest % delay for both the airlines.

For Ontime arrival, we can see Alaska airlines has more % arrival ontime in all the cities.

References

https://rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf