Data-607 Week-5 Assignment

Student Name : Sachid Deshmukh

Date : 09/29/2018

- GitHub Location for rmd file
- GitHub Location for pdf file
- RPubs location of published file
- Output file location file

1] Library Initialization

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.4.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

2] Read Data

```
Airline.Schedule = read.csv("C:/MSDS/MSDSQ1/Data607/Week5/AirlineSchedule.csv", stringsAsFactors = F)
```

2] Preview Data

```
head(Airline.Schedule)
```

```
##   Airline  Status  LA   PH  SD  SF  SEA
## 1  Alaska On Time 497  221 212 503 1841
## 2  Alaska Delayed  62   12  20 102  305
```

```
## 3 AM West On Time 694 4840 383 320  201
## 4 AM West Delayed 117  415  65 129   61
```

```
str(Airline.Schedule)
```

```
## 'data.frame':    4 obs. of  7 variables:
##  $ Airline: chr  "Alaska" "Alaska" "AM West" "AM West"
##  $ Status : chr  "On Time" "Delayed" "On Time" "Delayed"
##  $ LA     : int  497 62 694 117
##  $ PH     : int  221 12 4840 415
##  $ SD     : int  212 20 383 65
##  $ SF     : int  503 102 320 129
##  $ SEA    : int  1841 305 201 61
```

3] Convert Dataframe into tall format. Convert city columns into row

```
Airline.Schedule = Airline.Schedule%>%gather(City, FlightCount, LA:SEA)
head(Airline.Schedule)
```

```
##   Airline  Status City FlightCount
## 1  Alaska On Time   LA         497
## 2  Alaska Delayed   LA          62
## 3 AM West On Time   LA         694
## 4 AM West Delayed   LA         117
## 5  Alaska On Time   PH         221
## 6  Alaska Delayed   PH          12
```

4] We are interested in only delayed Flight Count. Let's filter for Status = Delayed

```
Airline.Delayed = Airline.Schedule%>%dplyr::filter(Status=="Delayed")
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.3
```

5] Analyze Summary Statistics

```
summary((Airline.Delayed%>%dplyr::filter(Airline == "Alaska")%>%data.frame())$FlightCount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    12.0    20.0    62.0   100.2   102.0   305.0
```

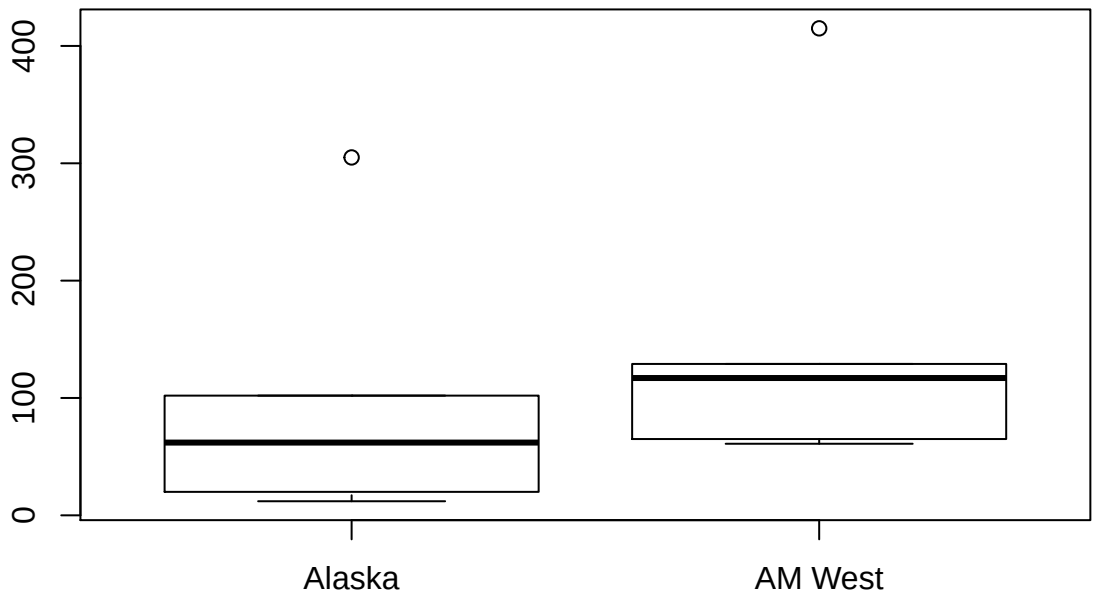
```
summary((Airline.Delayed%>%dplyr::filter(Airline == "AM West")%>%data.frame())$FlightCount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    61.0    65.0   117.0   157.4   129.0   415.0
```

We can see that delayed flight count number are on different scale for Alaska and AM West airlines. Let's confirm this with boxplot

6] Box plot to confirm difference in delayed flight counts

```
boxplot(Airline.Delayed$FlightCount ~ Airline.Delayed$Airline)
```



Boxplot also confirms that scale for flight delay counts for Alaska and AM West airlines are different. We can see two extreme observations which suggest some cities are more prone for flight delayes compared to others

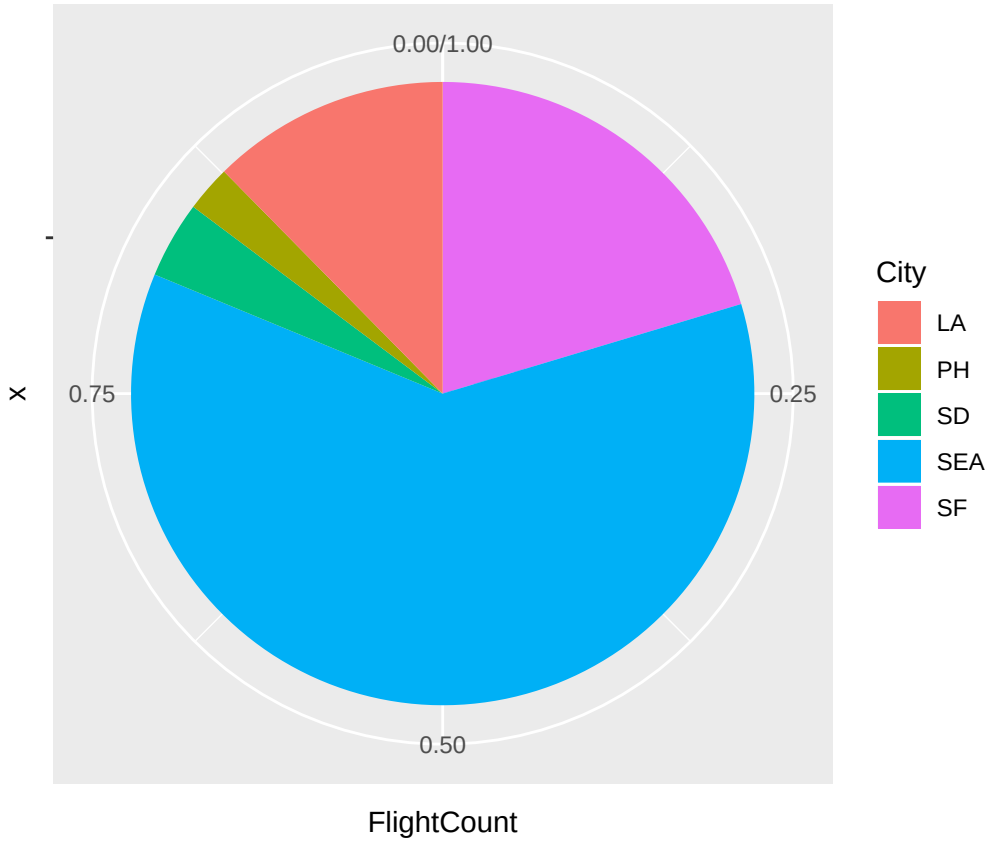
Since the scale is different we can't compare these stats directly. It's better to convert them to proportions for comparison purpose

```
Alaska.Sum = sum((Airline.Delayed%>%dplyr::filter(Airline == "Alaska")%>%data.frame())$FlightCount)
AMWest.Sum = sum((Airline.Delayed%>%dplyr::filter(Airline == "AM West")%>%data.frame())$FlightCount)

Airline.Alaska.Delayed = Airline.Delayed%>%dplyr::filter(Airline == "Alaska")%>%dplyr::mutate(FlightCou

Airline.AMWest.Delayed = Airline.Delayed%>%dplyr::filter(Airline == "AM West")%>%dplyr::mutate(FlightCo

Airline.Delayed = rbind(Airline.Alaska.Delayed, Airline.AMWest.Delayed)
```

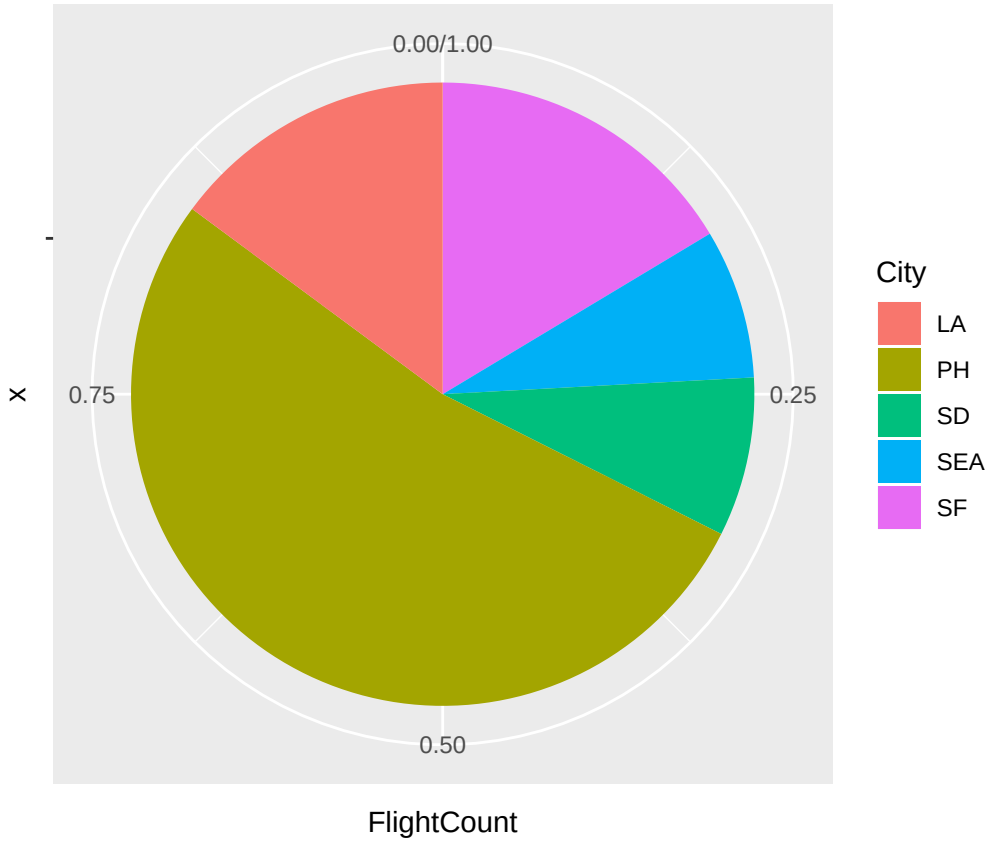
7] Let's visualize proportions with pie chart

```
DrawPieChart = function(data)
{
  bp<- ggplot(data, aes(x="", y=FlightCount, fill=City))+
  geom_bar(width = 1, stat = "identity")
  pie <- bp + coord_polar("y", start=0)
  pie
```

```
}
DrawPieChart(Airline.Delayed%>%dplyr::filter(Airline == "Alaska")%>%data.frame())
```



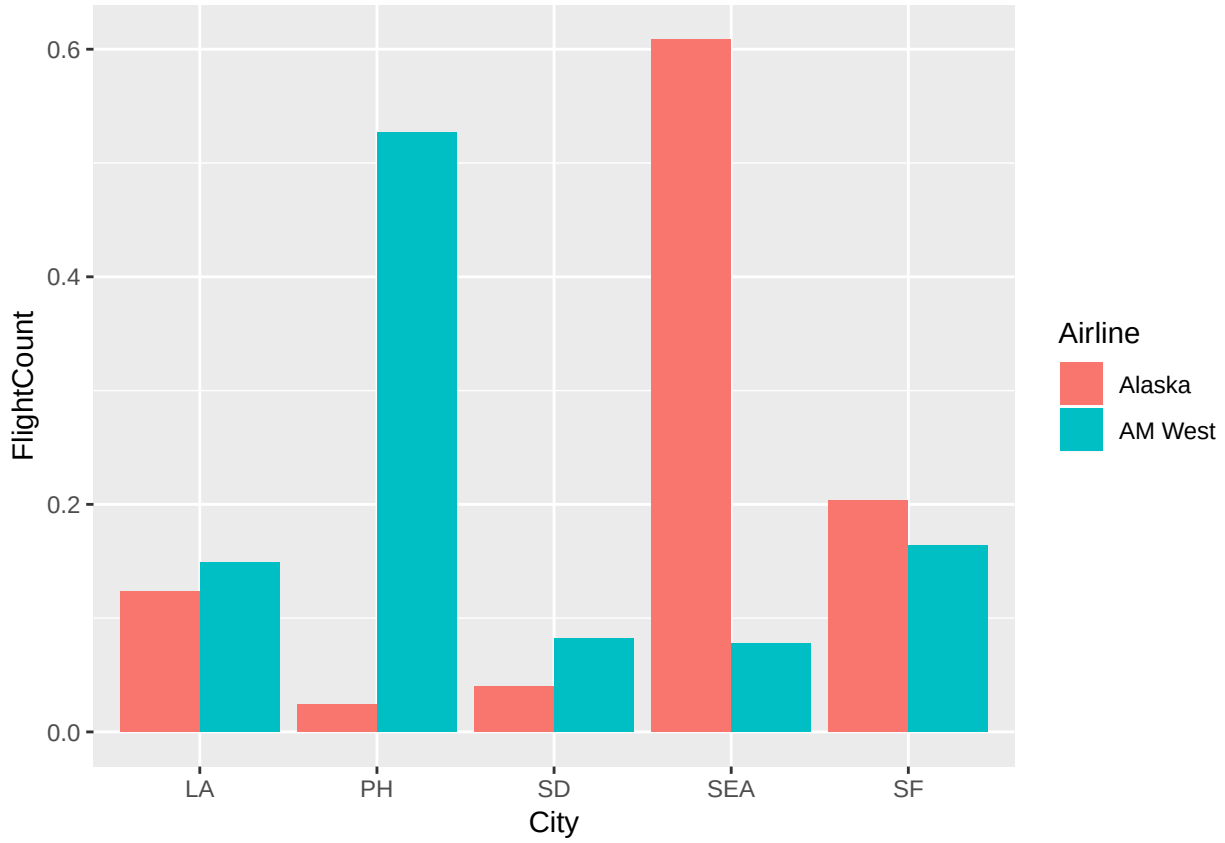

```
DrawPieChart(Airline.Delayed%>%dplyr::filter(Airline == "AM West")%>%data.frame())
```



Pie chart suggests that Alaska airline is more prone for delays in Seattle however AM West airline is more prone for delayes in Phoenix

8] Let's visualize proportions with Bar Chart

```
DrawBarChart = function(data)
{
  ggplot(data=data, aes(x=City, y=FlightCount, fill=Airline)) +
  geom_bar(stat="identity", position=position_dodge())
}

DrawBarChart(Airline.Delayed)
```



Bar chart also reveals that AM West is more prone for delayes in Phoenix and Alaska is more prone in delayes in Seattle

9] Is Alaska airline more prone for delayes compared to AM West Airline?

Let's see summary stats and boxplot

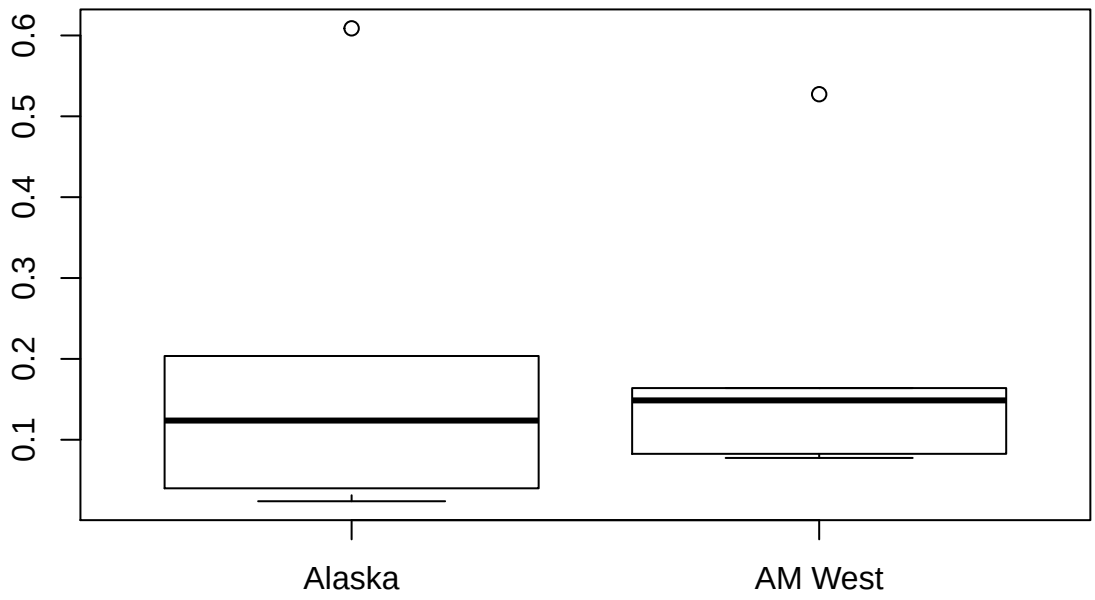
```
summary((Airline.Delayed%>%dplyr::filter(Airline == "Alaska")%>%data.frame())$FlightCount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.02395 0.03992 0.12375 0.20000 0.20359 0.60878
```

```
summary((Airline.Delayed%>%dplyr::filter(Airline == "AM West")%>%data.frame())$FlightCount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.07751 0.08259 0.14867 0.20000 0.16391 0.52732
```

```
boxplot(Airline.Delayed$FlightCount ~ Airline.Delayed$Airline)
```



Above sumary stats suggests that both Alaska and AM West airline have same avg percentage of filght delay whihc is 20%. From boxplot we can observe that Alaska airline have higher variablity 2% to 20% (18% IQR) in general compared to AM West airline whihc ranges from 7% to 16% (9% IQR)

From the above statistics it is difficult to conclude that one airline is more prone for delayes compared to other airline. Let's perform t test and chisquare test of independence to conclude our findings

```
Chisq.Delayed = Airline.Schedule%>%dplyr::filter(Status=="Delayed")
Alaska.D.Cnt = Airline.Schedule%>%dplyr::filter(Status=="Delayed" & Airline=="Alaska")%>%dplyr::select(C
AMWest.D.Cnt = Airline.Schedule%>%dplyr::filter(Status=="Delayed" & Airline=="AM West")%>%dplyr::select
chisqdf = data.frame(Alaska.D.Cnt$City, Alaska.D.Cnt$FlightCount, AMWest.D.Cnt$FlightCount)
names(chisqdf) = c("City", "AlaskaCount", "AMWestCount")
chisq.test(chisqdf[,-1])
```

```
##
##  Pearson's Chi-squared test
##
## data:  chisqdf[, -1]
## X-squared = 550.53, df = 4, p-value < 2.2e-16
```

```
t.test(Airline.Alaska.Delayed$FlightCount, Airline.AMWest.Delayed$FlightCount)
```

```
##
##  Welch Two Sample t-test
##
## data:  Airline.Alaska.Delayed$FlightCount and Airline.AMWest.Delayed$FlightCount
## t = -2.0423e-16, df = 7.5547, p-value = 1
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.3166456  0.3166456
## sample estimates:
## mean of x mean of y
##       0.2       0.2
```

1] ChiSquare test above have p value of 2.2e-16. We can reject null hypothesis that Delay in airline is independent of cities. We can conclude that particular city has significant impact on delay in airlines

2] Student t test above have p value of 1. We can't reject null hypotheisis that difference in mean for flight delayes between two airlines is 0. We can conclude that both alirlines have equal likelihood of flight delays