Assignment 5

Alejandro Osborne

March 3, 2018

```
library(RMySQL)
```

```
## Loading required package: DBI
```

```
library(tidyverse)
```

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

```
## -- Attaching packages ---------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.2     v dplyr   0.7.4
## v tidyr   0.7.2     v stringr 1.2.0
## v readr   1.1.1     v forcats 0.2.0
```

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

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

```
## -- Conflicts ------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidyr)
library(dplyr)
library(tidyselect)
```

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

```
##
## Attaching package: 'tidyselect'
```

```
## The following objects are masked from 'package:dplyr':
##
##     contains, ends_with, everything, matches, num_range, one_of,
##     starts_with
```

```
library(ggplot2)
library(stringr)
library(zoo, warn.conflicts = FALSE)
```

Establishing connection between R and Database

```
loadsql <- dbDriver("MySQL")
openlink = dbConnect(MySQL(), user='root', password ='celeborn', dbname='arrivals', host='localhost')
```

Query to get Data

```
AirportDataset<-"SELECT * FROM arrivals"
arrivals <-dbGetQuery(openlink, AirportDataset)
dbDisconnect(openlink)
```

```
## [1] TRUE
```

Loading Queried Data into data frame

```
FlightData <- data.frame(arrivals)
FlightData
```

```
##   Airline FlightStatus LosAngeles Phoenix SanDiego SanFrancisco Seattle
## 1  Alaska      on time        497     221      212          503    1841
## 2              delayed         62      12       20          102     305
## 3 AM West      on time        694    4840      383          320     201
## 4              delayed        117     415       65          129      61
```

Initial data is missing labels for 2 rows, will fix with "zoo" Library

```
FlightData$Airline[FlightData$Airline == ""] <- NA
FlightData$Airline <- na.locf(FlightData$Airline, option="locf")
FlightData <- FlightData
FlightData
```

```
##   Airline FlightStatus LosAngeles Phoenix SanDiego SanFrancisco Seattle
## 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
```

Finally Tidying the data (Stacking)

```
NewFlightData<-gather(FlightData, "City", "FlightCount", 3:7)
tidied_up <- arrange(NewFlightData, FlightStatus)
tidied_up
```

```
##    Airline FlightStatus         City FlightCount
## 1   Alaska      delayed   LosAngeles          62
## 2  AM West      delayed   LosAngeles         117
## 3   Alaska      delayed      Phoenix          12
## 4  AM West      delayed      Phoenix         415
## 5   Alaska      delayed     SanDiego          20
## 6  AM West      delayed     SanDiego          65
## 7   Alaska      delayed SanFrancisco         102
## 8  AM West      delayed SanFrancisco         129
## 9   Alaska      delayed      Seattle         305
## 10 AM West      delayed      Seattle          61
## 11  Alaska      on time   LosAngeles         497
## 12 AM West      on time   LosAngeles         694
## 13  Alaska      on time      Phoenix         221
```

```
## 14 AM West      on time      Phoenix        4840
## 15  Alaska      on time     SanDiego         212
## 16 AM West      on time     SanDiego         383
## 17  Alaska      on time SanFrancisco         503
## 18 AM West      on time SanFrancisco         320
## 19  Alaska      on time      Seattle        1841
## 20 AM West      on time      Seattle         201
```

Summary Statistics to give us a general overview of what the numbers can vaguely describe to us

```
tidied_up %>% group_by(Airline) %>% filter(FlightStatus == "delayed") %>% summarise(mean = mean(FlightC
```

```
## # A tibble: 2 x 7
##   Airline  mean   min   max median stdev total
##   <chr>   <dbl> <dbl> <dbl>  <int> <dbl> <int>
## 1 Alaska   100.   12.  305.     62  120.   501
## 2 AM West  157.   61.  415.    117  147.   787
```

We seek to find the last bit needed to reach a conclusion of any kind regarding this data. We create a field that describes the ratio of flights delayed/on time to the amount of flights scheduled.

```
NewTidied <- tidied_up %>% group_by(Airline, City) %>% arrange(Airline) %>% mutate(CityCounts = sum(Flig
NewTidied
```

```
## # A tibble: 20 x 6
## # Groups:   Airline, City [10]
##    Airline FlightStatus City         FlightCount CityCounts NewRatio
##    <chr>   <chr>        <chr>              <int>      <int>    <dbl>
##  1 Alaska  delayed      LosAngeles            62        559   0.111
##  2 Alaska  delayed      Phoenix               12        233   0.0515
##  3 Alaska  delayed      SanDiego              20        232   0.0862
##  4 Alaska  delayed      SanFrancisco         102        605   0.169
##  5 Alaska  delayed      Seattle              305       2146   0.142
##  6 Alaska  on time      LosAngeles           497        559   0.889
##  7 Alaska  on time      Phoenix              221        233   0.948
##  8 Alaska  on time      SanDiego             212        232   0.914
##  9 Alaska  on time      SanFrancisco         503        605   0.831
## 10 Alaska  on time      Seattle             1841       2146   0.858
## 11 AM West delayed      LosAngeles           117        811   0.144
## 12 AM West delayed      Phoenix              415       5255   0.0790
## 13 AM West delayed      SanDiego              65        448   0.145
## 14 AM West delayed      SanFrancisco         129        449   0.287
## 15 AM West delayed      Seattle               61        262   0.233
## 16 AM West on time      LosAngeles           694        811   0.856
## 17 AM West on time      Phoenix             4840       5255   0.921
## 18 AM West on time      SanDiego             383        448   0.855
## 19 AM West on time      SanFrancisco         320        449   0.713
## 20 AM West on time      Seattle              201        262   0.767
```

Finally we find The rations for the airlines in general (we only look at the Delayed flights) and we can see by a decent margin that AM West has more delayed flights than Alaska.

```
NewTidied %>% group_by(Airline) %>% filter(FlightStatus == "delayed") %>% summarise(mean = mean(NewRati
```

```
## # A tibble: 2 x 6
##   Airline  mean    min   max median standard_deviation
##   <chr>   <dbl>  <dbl> <dbl>  <dbl>              <dbl>
## 1 Alaska  0.112 0.0515 0.169  0.111             0.0459
## 2 AM West 0.178 0.0790 0.287  0.145             0.0821
```