Week 5 Assignment

Umais Siddiqui

September 30, 2017

Rpubs Link: http://rpubs.com/umais/data607_assignment5

Github Link: https://github.com/umais/DATA-607/tree/master/Week-5Assignment

Tidying and Transforming Data

Assignment Overview

In this assignment we will be using the TidyR and Dplyer packages to perform some analysis on the arrival delays for two Airlines.

Data Set

The data set that we will be using will be from MySQL data source. First step would be to use the RMySQL package to connect to the MySQL database and retrieve the data.

```
mydb = dbConnect(MySQL(), user='root', password='Welcome@1', dbname='flights', host='localhost')

rs = dbSendQuery(mydb, "SELECT * FROM AirlineArrival;")

df=fetch(rs, n=-1)

head(df)
```

```
##   Airline ArrivalStatus LosAngeles Phoenix San_Diego San_Francisco 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
```

Using tidyR to transform the data

As we can see from the results that the data set is in a wide format. What we would like to do in our first step is transform the data in to a long format so that we have the following columns Airline,ArrivalStatus,Cities, NumberOfFlights. We can achive this using the tidyR function called gather.

```
#Gather Function from tidyR
airlines2<- gather(df,Cities,NumberOfFlights,LosAngeles:Seattle)

head(airlines2)
```

```
##   Airline ArrivalStatus     Cities NumberOfFlights
## 1  Alaska       On Time LosAngeles             497
## 2  Alaska       Delayed LosAngeles              62
## 3 AM West       On Time LosAngeles             694
## 4 AM West       Delayed LosAngeles             117
## 5  Alaska       On Time    Phoenix             221
```

```
## 6  Alaska       Delayed    Phoenix              12
```

Using Dplyer to filter results

```
#Using the Filter Function from DPlyer
delayedFlights=filter(airlines2,ArrivalStatus=="Delayed")
delayedFlights
```

```
##    Airline ArrivalStatus        Cities NumberOfFlights
## 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     San_Diego              20
## 6  AM West       Delayed     San_Diego              65
## 7   Alaska       Delayed San_Francisco             102
## 8  AM West       Delayed San_Francisco             129
## 9   Alaska       Delayed       Seattle             305
## 10 AM West       Delayed       Seattle              61
```

Using DPlyer Pipeline and summarise function in DPlyer to look at the total number of flights delayed by each airline

We can see the comparison between how many flights are delayed by each ailrine

```
delayedFlights %>%
       group_by(Airline,ArrivalStatus)%>%
       summarise(total=sum(NumberOfFlights))
```

```
## Source: local data frame [2 x 3]
## Groups: Airline [?]
##
##   Airline ArrivalStatus total
##     <chr>         <chr> <int>
## 1  Alaska       Delayed   501
## 2 AM West       Delayed   787
```