DATA607 HW5

Vincent Bianco

3/1/2020

Introduction: The purpose of this assignment is to use tidy, transform, and analyze a given table of flights from two different airlines to 5 different cities. This data shows the amount of flights which arrive on time or delayed to a given destination.We will use tidyr to reshape the data into an easier format for analysis, then we will use ggplot to create an informative visualization to see which airline is performing better.

Loading up necessary libraries.

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

```
##
## 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)
```

Reading in the CSV file with destination and flight times data.

```
destinationfile <- read.csv("https://raw.githubusercontent.com/biancov/DATA607HW5/master/destinations.cs
destinationfile
```

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

Filling in missing row and column names.

```
names(destinationfile)[1] = "Airline"
names(destinationfile)[2] = "Arrival Status"
destinationfile$Airline[2] <- destinationfile$Airline[1]
destinationfile$Airline[5] <- destinationfile$Airline[4]
destinationfile
```

```
##   Airline Arrival Status 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
## 3                                 NA      NA        NA            NA      NA
## 4  AMWEST        on time         694    4840       383           320     201
## 5  AMWEST        delayed         117     415        65           129      61
```

Lets make the Arrival Statuses (On time and delayed) as columns, and the Airline + Destination as part of each row entry.

```
destinationfile <- drop_na(destinationfile)

destinationfilenew <- destinationfile %>%
  gather("Destinations","Flights",3:7) %>%
  spread("Arrival Status","Flights")
destinationfilenew
```

```
##    Airline  Destinations delayed on time
## 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   AMWEST   Los.Angeles     117     694
## 7   AMWEST       Phoenix     415    4840
## 8   AMWEST     San.Diego      65     383
## 9   AMWEST San.Francisco     129     320
## 10  AMWEST       Seattle      61     201
```

Lets create a "Total" and Delay Rate" column, showing the total number of flights and the proportion of those flights which are delayed for each airline at each destination.

```
destinationfilefinal <- mutate(destinationfilenew,
      Total = delayed + as.numeric(destinationfilenew$'on time'),
      Delay_Rate = 100*delayed/Total)
destinationfilefinal
```

```
##    Airline  Destinations delayed on time Total Delay_Rate
## 1   ALASKA   Los.Angeles      62     497   559  11.091234
## 2   ALASKA       Phoenix      12     221   233   5.150215
## 3   ALASKA     San.Diego      20     212   232   8.620690
## 4   ALASKA San.Francisco     102     503   605  16.859504
## 5   ALASKA       Seattle     305    1841  2146  14.212488
## 6   AMWEST   Los.Angeles     117     694   811  14.426634
## 7   AMWEST       Phoenix     415    4840  5255   7.897241
## 8   AMWEST     San.Diego      65     383   448  14.508929
## 9   AMWEST San.Francisco     129     320   449  28.730512
## 10  AMWEST       Seattle      61     201   262  23.282443
```

Now we can summarize this data to see which airline is having the most delays and at which cities are most of the delays occuring across all airlines.

```
destinationfilefinal %>%
  group_by(Airline) %>%
  summarize(mean(Delay_Rate))
```

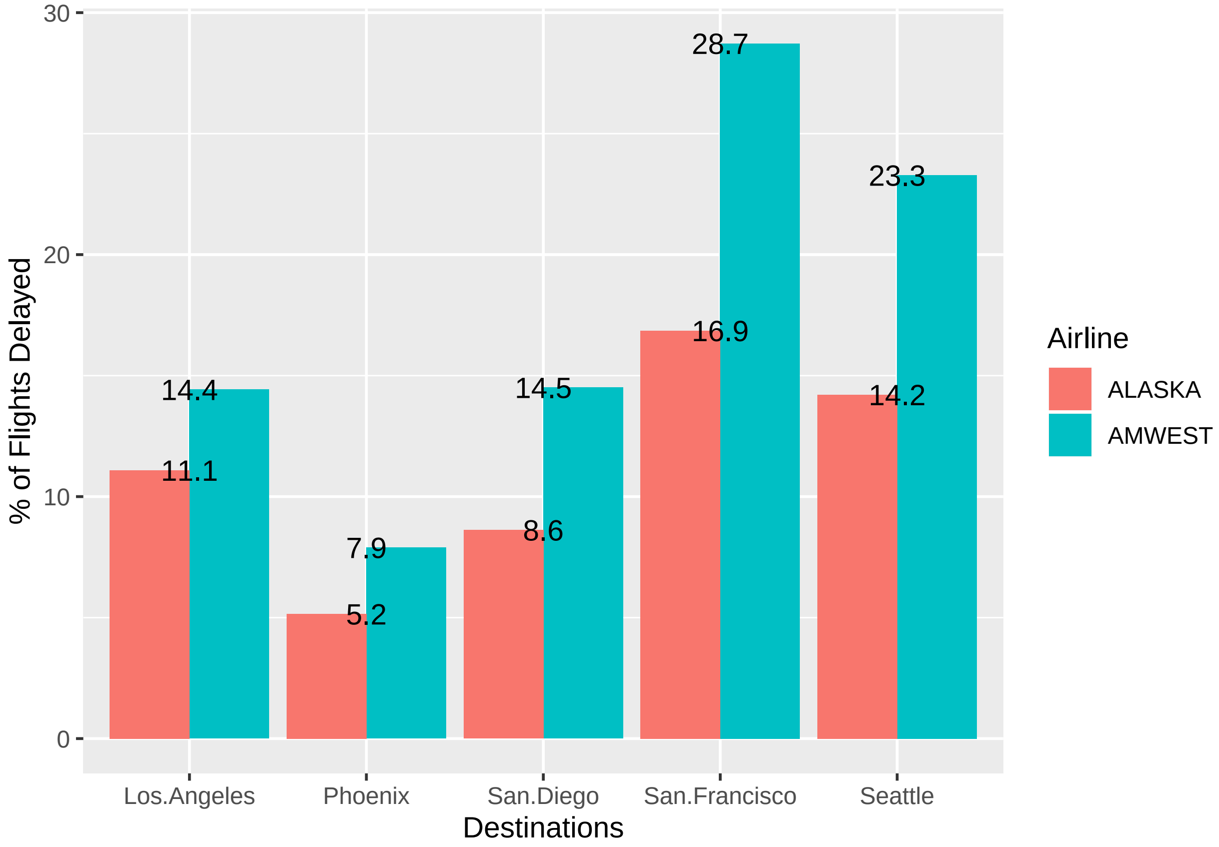
```
## # A tibble: 2 x 2
##   Airline `mean(Delay_Rate)`
##   <chr>                <dbl>
## 1 ALASKA                11.2
## 2 AMWEST                17.8
```

```
destinationfilefinal %>% group_by(Destinations) %>% summarize(mean(Delay_Rate))
```

```
## # A tibble: 5 x 2
##   Destinations  `mean(Delay_Rate)`
##   <chr>                      <dbl>
## 1 Los.Angeles                12.8
## 2 Phoenix                     6.52
## 3 San.Diego                  11.6
## 4 San.Francisco              22.8
## 5 Seattle                    18.7
```

For both airlines, we can use a bar chart to compare the difference in delay rates for each destination.

```
ggplot(destinationfilefinal, aes(x=Destinations,y=Delay_Rate)) +
geom_bar(aes(fill=Airline), stat = "identity",position=position_dodge()) +
ylab("% of Flights Delayed") +
geom_text(aes(label=round(Delay_Rate,1)), color = "black",position = position_dodge(0.9))
```



Here we can see that across all the destinations, AMWEST is experiencing a significantly higher percentage of delays.