# DATA 607: WEEK 5 Assignment Tidying and Transforming Data

James Williams

#STEP 1: READ IN DATA Load libraries and import CSV file

```
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)
flights <- as_tibble(read.csv("https://raw.githubusercontent.com/jambawilliams/DATA607WEEK5/master/Flig
flights
```

```
## # A tibble: 5 x 7
##   X       X.1     Los.Angeles Phoenix San.Diego San.Francisco Seattle
##   <fct>   <fct>         <int>   <int>     <int>         <int>   <int>
## 1 ALASKA  on time         497     221       212           503    1841
## 2 <NA>    delayed          62      12        20           102     305
## 3 <NA>    <NA>             NA      NA        NA            NA      NA
## 4 AM WEST on time         694    4840       383           320     201
## 5 <NA>    delayed         117     415        65           129      61
```

#STEP 2: TIDY DATA Rename columns, remove rows with null values, fill empty cells with preceding values

```
flights <- flights %>% rename(Airline = X)
flights <- flights %>% rename(Status = X.1)
flights <- flights %>% rename(Los_Angeles = Los.Angeles)
flights <- flights %>% rename(San_Diego = San.Diego)
flights <- flights %>% rename(San_Francisco = San.Francisco)
flights <- flights %>% filter(!is.na(flights$Los_Angeles))
flights <- flights %>% fill(Airline)
flights
```

```
## # A tibble: 4 x 7
##   Airline Status  Los_Angeles Phoenix San_Diego San_Francisco Seattle
##   <fct>   <fct>         <int>   <int>     <int>         <int>   <int>
## 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
```

#STEP 3: COMPARE ARRIVAL DELAYS Transform tibble to organize data by destination and delay count to compare airline performance

```
delay <- gather(flights, "Destination", "n", 3:7)
delay <- spread(delay, "Status", "n")
delay <- arrange(delay, desc(Destination))
delay <- delay %>% rename(Delayed = delayed)
delay <- delay %>% rename(On_Time = "on time")
delay
```

```
## # A tibble: 10 x 4
##    Airline Destination   Delayed On_Time
##    <fct>   <chr>           <int>   <int>
##  1 ALASKA  Seattle           305    1841
##  2 AM WEST Seattle            61     201
##  3 ALASKA  San_Francisco     102     503
##  4 AM WEST San_Francisco     129     320
##  5 ALASKA  San_Diego          20     212
##  6 AM WEST San_Diego          65     383
##  7 ALASKA  Phoenix            12     221
##  8 AM WEST Phoenix           415    4840
##  9 ALASKA  Los_Angeles        62     497
## 10 AM WEST Los_Angeles       117     694
```

Compare total number of delays between airlines

```
delay %>% group_by(Airline) %>% summarise(Delayed=sum(Delayed), On_time=sum(On_Time))
```

```
## # A tibble: 2 x 3
##   Airline Delayed On_time
##   <fct>     <int>   <int>
## 1 ALASKA      501    3274
## 2 AM WEST     787    6438
```

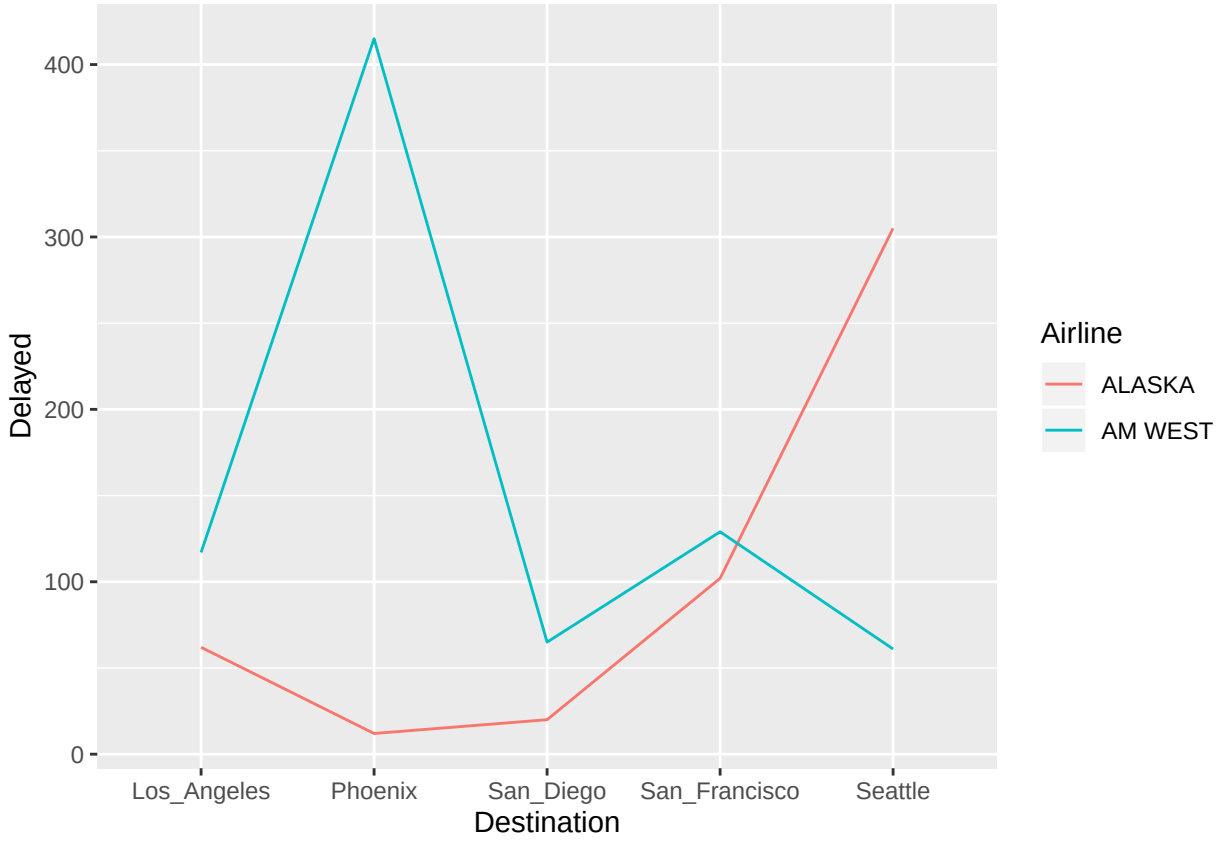
Compare number of delays between destinations

```
delay %>% group_by(Destination) %>% summarise(Delayed=sum(Delayed), On_time=sum(On_Time))
```

```
## # A tibble: 5 x 3
##   Destination   Delayed On_time
##   <chr>           <int>   <int>
## 1 Los_Angeles       179    1191
## 2 Phoenix           427    5061
## 3 San_Diego          85     595
## 4 San_Francisco     231     823
## 5 Seattle           366    2042
```

Visualize airline performance

```
graph <- ggplot(delay, aes(x = Destination, y = Delayed))+
  geom_line(aes(color=Airline, , group = Airline))
graph
```




#STEP 4: CONCLUSIONS In absolute terms, American West had more delays than Alaska. American West though also had twice as many flights as Alaska. American West actually had a lower rate of delay (12%) compared to Alaska (15%). Phoenix had the most number of delays of any destination while San Diego had the least. San Francisco had the highest rate of delays of any destination at 28%, whereas San Diego had the lowest at 8%.