

# Assignment 04

Ted Kim (Team Member: Seung Min Song)

2022-09-28

## 1. Data Read and Transform

Retrieve data from csv file into a data table *dt\_wider*.

```
dt_wider <- as.data.table(read.csv('https://raw.githubusercontent.com/blacksmilez/DATA607/main/Assignme
```

Define the names of empty columns 1 and 2 with column names *Airlines* and *Status*.

```
colnames(dt_wider)[1:2] = c('Airlines', 'Status')
dt_wider
```

##	Airlines	Status	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:	AM WEST	on time	694	4840	383	320	201
## 4:		delayed	117	415	65	129	61

Fill in the empty Airlines cells (probably two rows combined) with the cell value directly above.

RDocumentation. := Assignment by reference

RDocumentation. shift Fast lead/lag for vectors and lists

```
dt_wider[, Airlines := ifelse(Airlines != '', Airlines, shift(Airlines))]
dt_wider
```

##	Airlines	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:	AM WEST	on time	694	4840	383	320	201

```
## 4: AM WEST delayed      117      415      65      129      61
```

Use `pivot_longer()` function to create new column named *Air\_Port* and insert city name into *Air\_Port* column.

```
dt_long <- dt_wider %>%
  pivot_longer(
    cols = c('Los_Angeles', 'Phoenix', 'San_Diego', 'San_Francisco', 'Seattle'),
    names_to = 'Air_Port',
    values_to = 'Flights')
dt_long
```

```
## # A tibble: 20 x 4
##   Airlines Status Air_Port    Flights
##   <chr>    <chr>   <chr>      <int>
## 1 ALASKA  on time Los_Angeles    497
## 2 ALASKA  on time Phoenix      221
## 3 ALASKA  on time San_Diego    212
## 4 ALASKA  on time San_Francisco  503
## 5 ALASKA  on time Seattle    1841
## 6 ALASKA  delayed Los_Angeles    62
## 7 ALASKA  delayed Phoenix     12
## 8 ALASKA  delayed San_Diego     20
## 9 ALASKA  delayed San_Francisco  102
## 10 ALASKA  delayed Seattle     305
## 11 AM WEST  on time Los_Angeles    694
## 12 AM WEST  on time Phoenix    4840
## 13 AM WEST  on time San_Diego     383
## 14 AM WEST  on time San_Francisco  320
## 15 AM WEST  on time Seattle     201
## 16 AM WEST  delayed Los_Angeles    117
## 17 AM WEST  delayed Phoenix     415
## 18 AM WEST  delayed San_Diego     65
## 19 AM WEST  delayed San_Francisco  129
## 20 AM WEST  delayed Seattle      61
```

## 2. Analysis for Arrival Delays

To perform analysis to compare the arrival delays for the two airlines, we tried to see the delayed frequency for both airlines and see which airports have higher delayed rate than the average.

A. Calculate the **delayed\_rate** percentage of each carrier for each city.

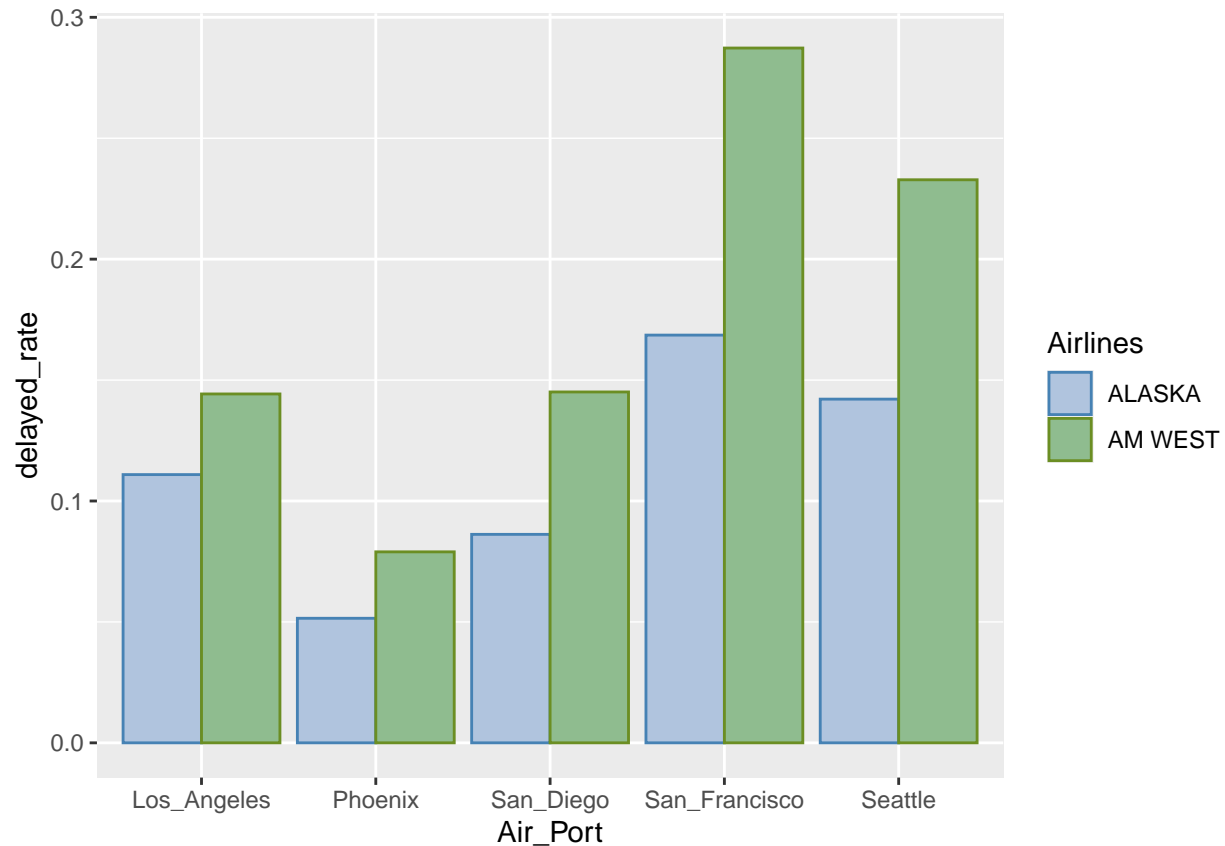
```
dt_summary <- dt_long %>%
  group_by(Airlines, Air_Port) %>%
  summarise(
    on_time = Flights[Status == 'on time'],
    delayed = Flights[Status == 'delayed'],
    total = Flights[Status == 'delayed'] + Flights[Status == 'on time'],
    delayed_rate = Flights[Status == 'delayed']
      / (Flights[Status == 'delayed'] + Flights[Status == 'on time']))
dt_summary
```

```
## # A tibble: 10 x 6
## # Groups:   Airlines [2]
##   Airlines Air_Port      on_time delayed total delayed_rate
##   <chr>      <chr>        <int>    <int> <int>         <dbl>
## 1 ALASKA    Los_Angeles      497      62   559         0.111
## 2 ALASKA    Phoenix         221     12   233         0.0515
## 3 ALASKA    San_Diego        212     20   232         0.0862
## 4 ALASKA    San_Francisco    503    102   605         0.169
## 5 ALASKA    Seattle        1841    305  2146         0.142
## 6 AM WEST   Los_Angeles      694    117   811         0.144
## 7 AM WEST   Phoenix        4840    415  5255         0.0790
## 8 AM WEST   San_Diego        383     65   448         0.145
## 9 AM WEST   San_Francisco    320    129   449         0.287
## 10 AM WEST  Seattle         201     61   262         0.233
```

B. ggplot to draw a geom\_bar graph to compare two carriers.

```
ggplot(data = dt_summary, aes(x = Air_Port, y = delayed_rate,
                              fill = Airlines, color = Airlines)) +
  geom_bar(stat='identity', position='dodge') +
  scale_color_manual(values = c('SteelBlue', 'OliveDrab')) +
```

```
scale_fill_manual(values = c('LightSteelBlue', 'DarkSeaGreen'))
```



C. Calculate the mean delayed for each carrier. And, check whether the airports' delayed rate is above or below average.

```
dt_summary <- dt_summary %>%
  group_by(Airlines) %>%
  mutate(
    mean_delay = mean(delayed_rate),
    above_below_avg = ifelse(delayed_rate > mean(delayed_rate), 'above', 'below')
  )
```

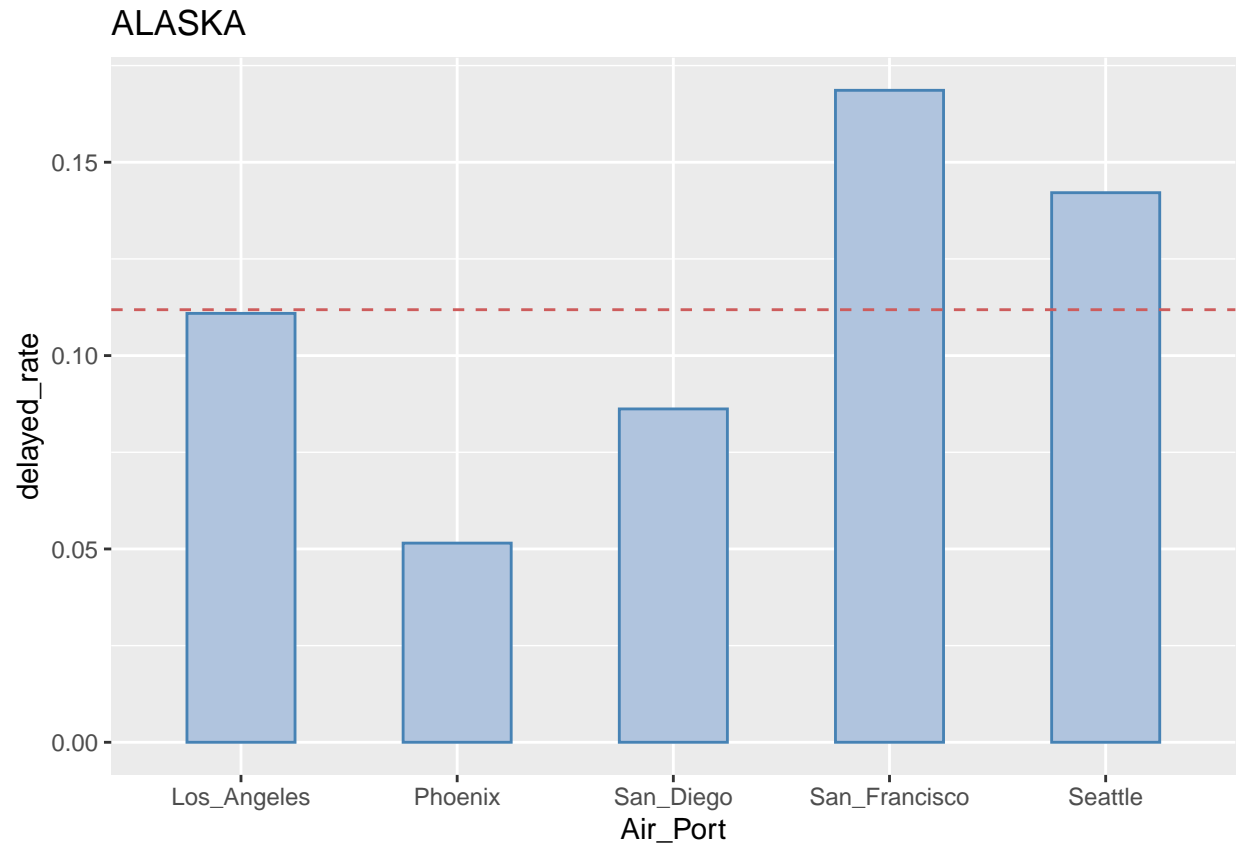
```
dt_summary
```

```
## # A tibble: 10 x 8
## # Groups:   Airlines [2]
##   Airlines Air_Port      on_time delayed total delayed_rate mean_delay above~1
##   <chr>      <chr>      <int>    <int> <int>      <dbl>    <dbl> <chr>
## 1 ALASKA    Los_Angeles    497     62  559      0.111    0.112 below
```

```
## 2 ALASKA Phoenix 221 12 233 0.0515 0.112 below
## 3 ALASKA San_Diego 212 20 232 0.0862 0.112 below
## 4 ALASKA San_Francisco 503 102 605 0.169 0.112 above
## 5 ALASKA Seattle 1841 305 2146 0.142 0.112 above
## 6 AM WEST Los_Angeles 694 117 811 0.144 0.178 below
## 7 AM WEST Phoenix 4840 415 5255 0.0790 0.178 below
## 8 AM WEST San_Diego 383 65 448 0.145 0.178 below
## 9 AM WEST San_Francisco 320 129 449 0.287 0.178 above
## 10 AM WEST Seattle 201 61 262 0.233 0.178 above
## # ... with abbreviated variable name 1: above_below_avg
```

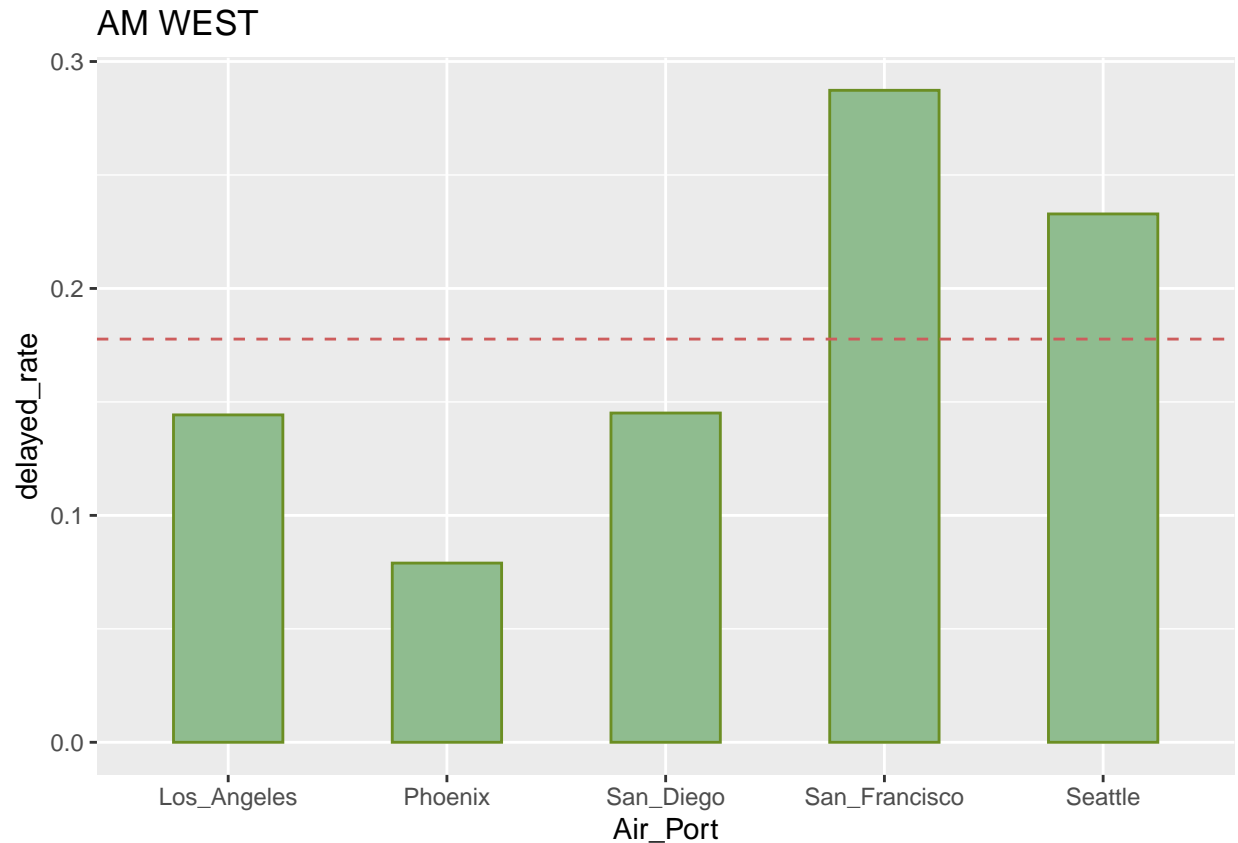
Draw `geom_bar` graph to compare Alaska airline's delay frequency in each airport with average delay rate.

```
mean_delay = unique(dt_summary[dt_summary$Airlines == 'ALASKA',]$mean_delay)
p <- dt_summary %>%
  filter(Airlines == 'ALASKA') %>%
  ggplot(aes(x = Air_Port, y = delayed_rate)) +
  ggtitle('ALASKA') +
  geom_bar(stat = 'identity', color = 'SteelBlue',
           fill = 'LightSteelBlue', width = 0.5)
p + geom_hline(yintercept = unique(mean_delay), linetype='dashed', color = 'IndianRed')
```



Draw `geom_bar` graph to compare AM West airline's delay frequency in each airport with average delay rate.

```
mean_delay = unique(dt_summary[dt_summary$Airlines == 'AM WEST',]$mean_delay)
p <- dt_summary %>%
  filter(Airlines == 'AM WEST') %>%
  ggplot(aes(x = Air_Port, y = delayed_rate)) +
  ggtitle('AM WEST') +
  geom_bar(stat = 'identity', color = 'OliveDrab',
           fill = 'DarkSeaGreen', width = 0.5)
p + geom_hline(yintercept = unique(mean_delay), linetype='dashed', color = 'IndianRed')
```



### 3. Conclusion

AM West has a higher delay frequency in every city than Alaska, and AM West has a higher average delay percentage than Alaska. Both airlines have two airports above average and three below average. San Francisco and Seattle are the most delayed cities based on this data set. Moreover, this data set is not sufficient to clearly identify which airline is better in general. This could be analyzed deeper if it contains the year, month, departure delay, and arrival delay data. Year and month data is useful in analyzing trends over time or comparing seasonal airline performance. Departure delay and arrival delay data are useful to spectate how severe the delay is. Because to some people, a delay of five to ten minutes may not be considered a delay. Therefore, if there is no big difference in price and service, Alaska with fewer delays looks better than AM West.

- GitHub - <https://github.com/blacksmilez/DATA607/tree/main/Assignment04>
- RPubS - <https://rpubs.com/blacksmilez/951635>