Week 5 Assignment

Matthew Roland

2023-09-27

Loading & Cleaning the Data

```
##This code will load the csv from my github repo
flight_data <- read.csv("https://raw.githubusercontent.com/Mattr5541/DATA-607/main/Week%205/Flight_Data

##This will change any empty strings to NA
flight_data[flight_data == ""] <- NA

##Delete any rows that are completely populated by NA values
flight_data <- flight_data %>% filter(rowSums(is.na(.)) != ncol(flight_data))

##I want this dataset to exist in a long format, so I am populating any empty rows in the X column with
flight_data <- flight_data %>% fill(everything(), .direction = "down")

##Finally, I want to rename the variables for the sake of readability and consistency
flight_data <- flight_data %>% rename(Airport = X, Arrival = X.1, Los_Angeles = Los.Angeles, San_Diego =

kable(flight_data)
```

Airport	Arrival	Los_Angeles	Phoenix	San_Diego	San_Francisco	Seattle
ALASKA	on time	497	221	212	503	1841
ALASKA	delayed	62	12	20	102	305
AM WEST	on time	694	4840	383	320	201
AM WEST	delayed	117	415	65	129	61

Converting from wide to long

```
##This will pivot the data from wide to long
flight_long <- flight_data %>% pivot_longer(
  cols = -c(Airport, Arrival),
  names_to = "City",
  values_to = "Value"
)

kable(flight_long)
```

Airport	Arrival	City	Value
ALASKA	on time	Los_Angeles	497
ALASKA	on time	Phoenix	221
ALASKA	on time	San_Diego	212
ALASKA	on time	San_Francisco	503
ALASKA	on time	Seattle	1841
ALASKA	delayed	Los_Angeles	62
ALASKA	delayed	Phoenix	12
ALASKA	delayed	San_Diego	20
ALASKA	delayed	San_Francisco	102
ALASKA	delayed	Seattle	305
AM WEST	on time	Los_Angeles	694
AM WEST	on time	Phoenix	4840
AM WEST	on time	San_Diego	383
AM WEST	on time	San_Francisco	320
AM WEST	on time	Seattle	201
AM WEST	delayed	Los_Angeles	117
AM WEST	delayed	Phoenix	415
AM WEST	delayed	San_Diego	65
AM WEST	delayed	San_Francisco	129
AM WEST	delayed	Seattle	61

Now let's finish off with an analysis

First, I admittedly thought about finding mean and median values for the data. . . but then I realized that proportion/percentages would likely be more appropriate for this type of frequency database

```
##This will create group sums for arrivals by airport, without respect to city
arrive_sum <- flight_long %>% group_by(Airport, Arrival) %>%
  summarize(Sum = sum(Value))
```

```
## `summarise()` has grouped output by 'Airport'. You can override using the
## `.groups` argument.
```

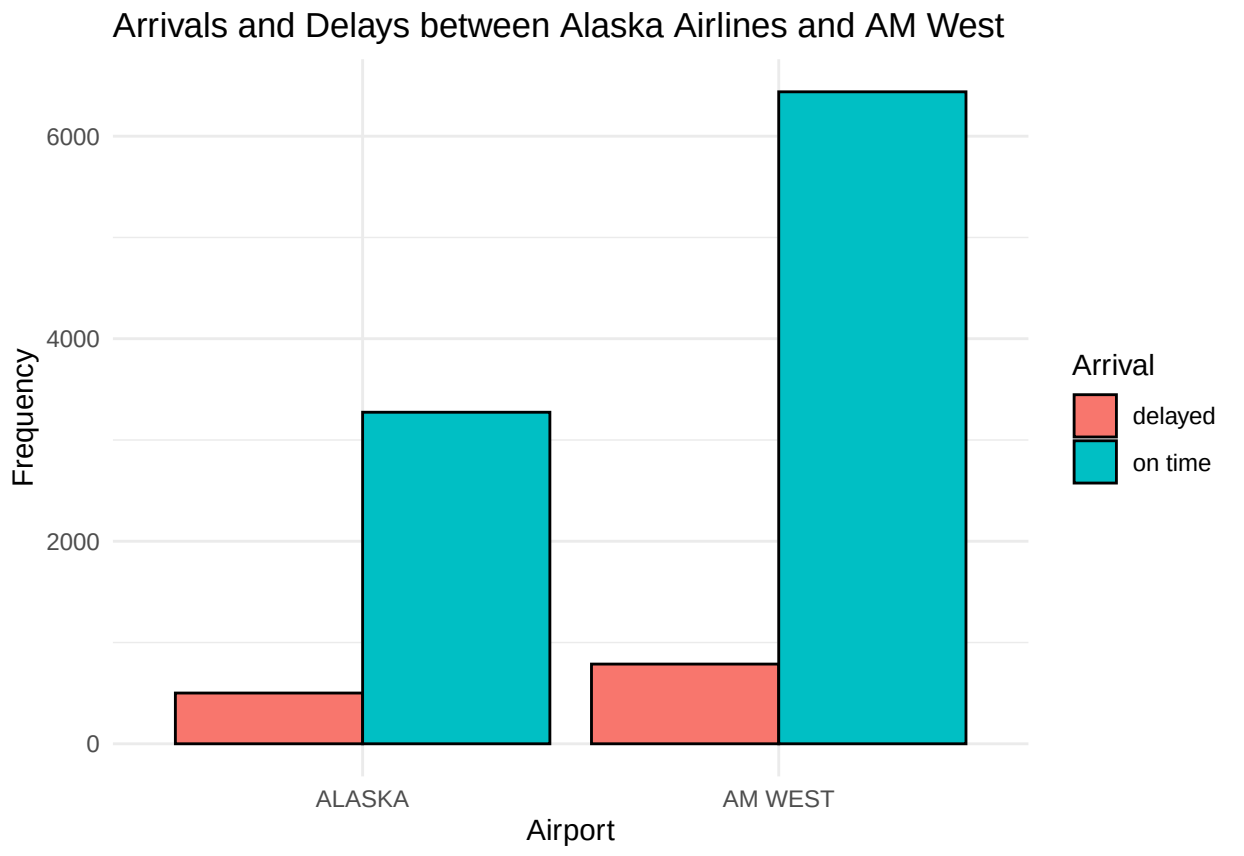
```
##This will calculate the sums for Alaska and AM West separately and put them into a new data frame
arrive_totals <- arrive_sum %>% group_by(Airport) %>%
  summarize(Totals = sum(Sum))

##Now, I will merge that dataframe into the arrive_sum dataframe so I can perform some calculations
arrive_sum <- arrive_sum %>% merge(arrive_totals, by = "Airport")

##And here are the calculations: Specifically, I want to divide the sums for on-time vs. delays by the
arrive_sum <- arrive_sum %>% mutate(Proportion = round((Sum / Totals) * 100, 2))

##And now, I want to drop the Totals column, since it was only for the sake of computation
arrive_sum <- arrive_sum %>% select(-c("Totals"))
##And I felt like adding a % string to the end of every value
arrive_sum$Proportion <- paste0(arrive_sum$Proportion, "%")

kable(arrive_sum)
```

Airport	Arrival	Sum	Proportion
ALASKA	delayed	501	13.27%
ALASKA	on time	3274	86.73%
AM WEST	delayed	787	10.89%
AM WEST	on time	6438	89.11%

As we can see here, Alaska airlines had a slightly higher percentage of delays when compared to AM West. Overall, however, both AM West and Alaska airlines had a relatively consistent proportion of on-time arrivals compared to delays

##Visual Representation

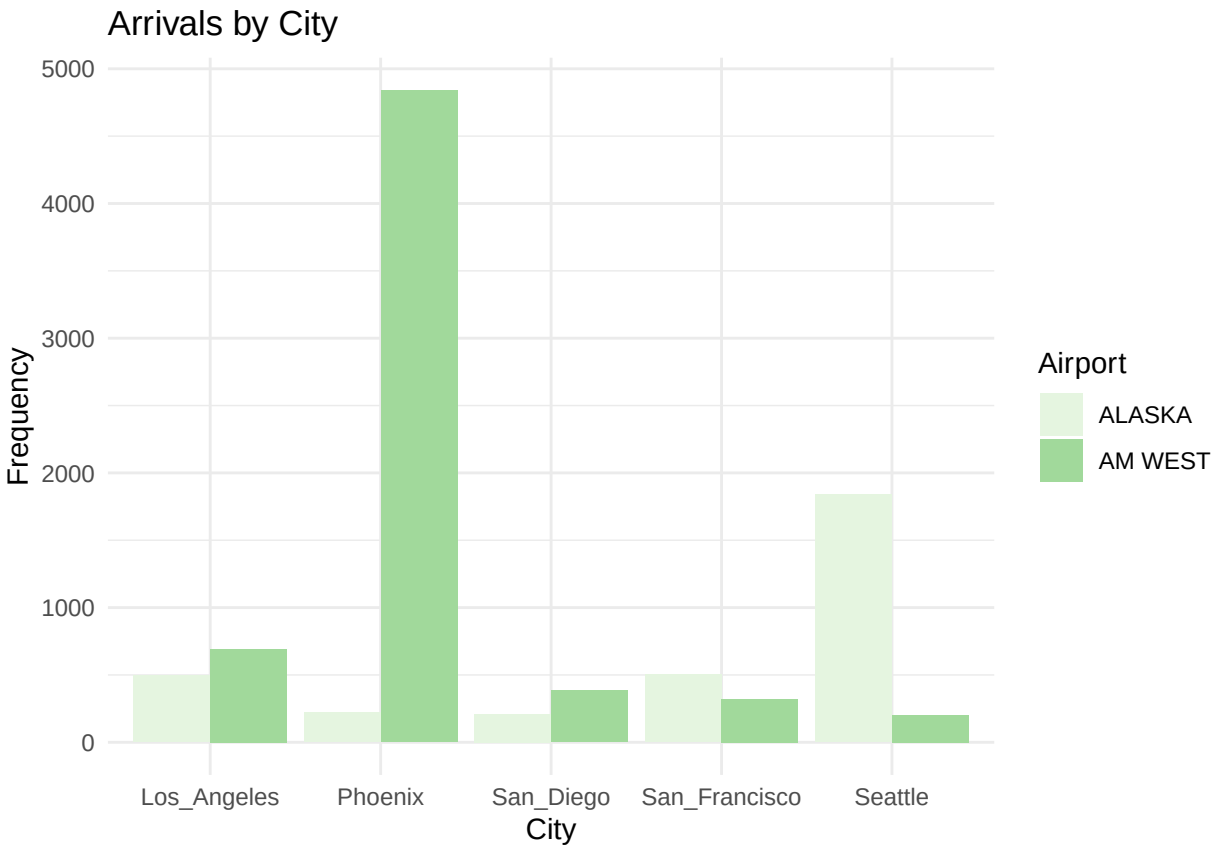
```
ggplot(arrive_sum, aes(x = Airport, y = Sum, fill = Arrival)) +
  geom_bar(stat = "identity", color = "black", position = position_dodge()) + theme_minimal() + labs(ti
```



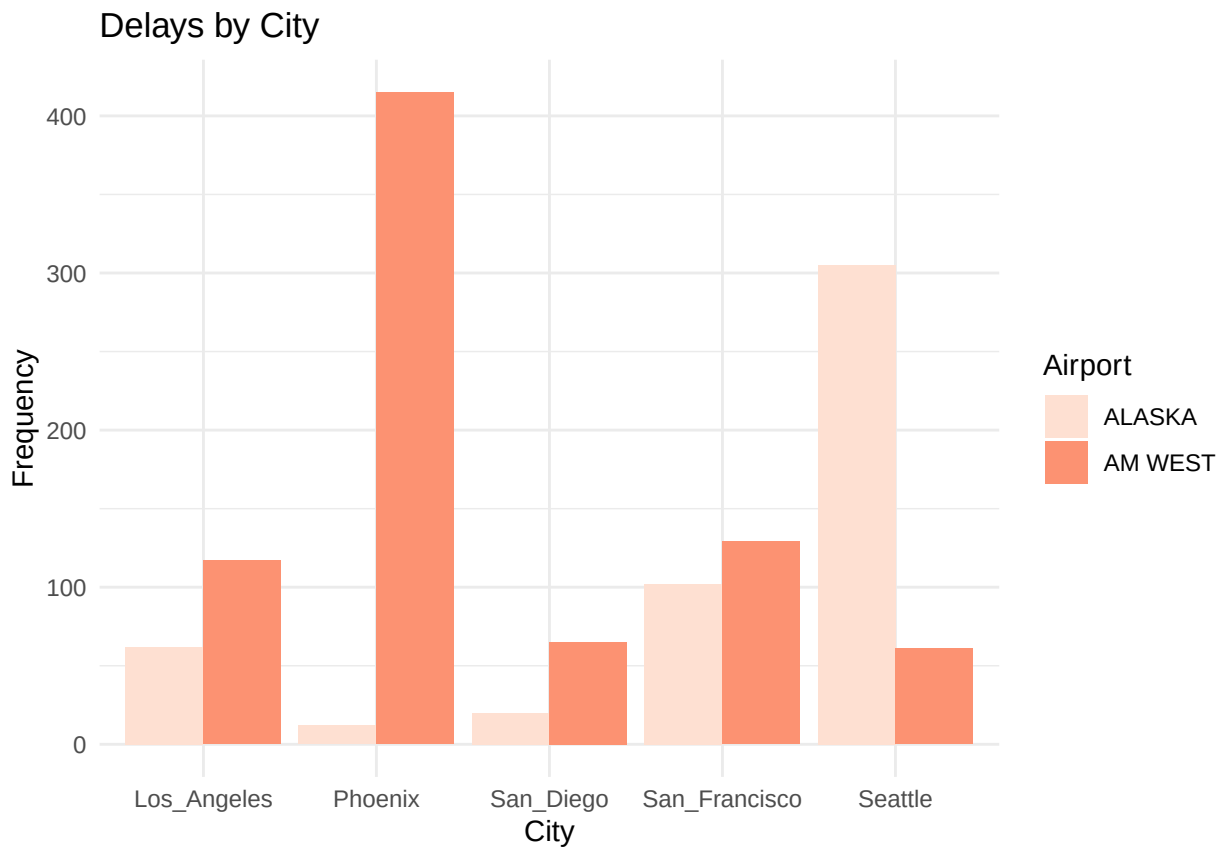

```
ggplot(arrive_sum, aes(x = Airport, y = Proportion, fill = Arrival)) +
  geom_bar(stat = "identity", color = "black", position = position_dodge()) + theme_minimal() + labs(ti
```

And finally, graphic representation of the frequency counts for arrivals and delays by city

```
flight_arrive <- flight_long %>% subset(Arrival != "delayed")

ggplot(flight_arrive, aes(x = City, y = Value, fill = Airport)) + geom_bar(stat = "identity", position =
```




```
flight_delays <- flight_long %>% subset(Arrival != "on time")

ggplot(flight_delays, aes(x = City, y = Value, fill = Airport)) + geom_bar(stat = "identity", position =
```



And with this, we can see that AM West flights from Phoenix have the largest amount of on-time arrivals, but also the largest amount of delays