STATS 15 Final Project - Analyzing Airfares

Agastya Rao, Bruno Cardenas-Bourges, Jacob Bianchi, and Margo Novikov

Background Information

Guiding Questions and Scope of Analysis

The airplane was one of the most significant and influential inventions of the 20th century, and the airline industry is an exciting and ever-evolving field that will continue to shape the way the world travels for years to come. As a result of the creation of the airplane, countless new connections have been formed between continents, countries, and people who live thousands of miles apart. Aviation has fostered economic growth across the globe, and it has facilitated long-distance travel for millions. The continental United States has a strong network of airways, connecting nearly every state and major city throughout the country.

For travelers, the price of airline tickets is a critical factor when making decisions. With customers trying to optimize their spending and trying to find the cheapest flights possible, we decided to explore **what factors affect the price of an airline ticket**. However, since we are all based in LA and have a special connection to the city, along with the fact that our original data consisted of over 9 million rows, we decided to focus our analysis on flights departing only from LAX. To this end, the motivating question of our project is as follows: **what factors affect the price of a ticket for a flight departing from LAX?**

General Information and Background

What is the ariline industry?

The airline industry is a collection of businesses and companies known as airlines. These airlines offer transportation services to both travelers and cargo. For our purposes, we will be focusing on the travel portion of the airline industry.

How does a customer book a ticket?

Customers can book tickets through a desired airline, select their origin and destination airports, and choose a convenient departure/arrival date and time. Then, the user is typically shown a list of flights and their corresponding prices. Some of the offered flights may be direct, meaning that they connect the two airports in one contiguous flight. Other flights may have connections, where the passengers need to disembark at a different airport and transfer onto another airplane. Although the customer would pay one fee for the entire trip, our dataset treats each connection (also known as a layover flight) as its own observation, and the total price is split up accordingly.

Many passengers typically look to book a flight round-trip. They stay at the destination for a given period of time, and return back to the origin location. Although the customer only pays once, our dataset treats these as two separate flights, so a round-trip ticket would count for two separate observations, and the price per ticket would be divided among those two flights. On the other hand, a one-way ticket is a flight from one location to another, with no return flight. These are represented as a single observation.

What is the significance of a hub?

Oftentimes, these connection flights, also known as layovers, are routed to a hub airport. Hub airports serve as transfer points between flights, and they often have direct flights to many airports. The hub system allows air travel between two smaller airports to be economically feasible for both the airline and customer. Offering flights from a small airport to a hub airport (rather than between two small airports), helps to reduce empty seats on the flight, thus decreasing costs.

However, each airline typically operates through its own "primary" hub. Many of the flights through the airline will be routed to its hub airport, which allows the airline to focus its operations at that location. This relationship means that a hub airport/city is served primarily by a single airline. This drives up costs for passengers since there is little competition from other airlines at a hub airport.

Background on LAX

Los Angeles International Airport (LAX) is one the most famous and one of the busiest airports in the world. It serves the metropolitan area of Los Angeles, California. In fact, in terms of the number of passengers traveling through the airport, it ranks second in the United States and third in the world (according to Airports Council International). Serving as the busiest airport on the west coast, LAX offers a large number of flights to a variety of airports throughout the world. Our analysis will focus on flights where LAX is the origin airport.

Information About the Dataset

Reliability and trustworthiness

We found the dataset on Kaggle, a website that allows users to upload data and share it with others. (https://www.kaggle.com/datasets/zernach/2018-airplane-flights) The "2018 Airplane Flights" dataset was uploaded by a data scientist who created it by compiling data from the Bureau of Transportation Statistics - Office of Airline Information. Specifically, he used the Airline Origin and Destination Survey (DB1B). This dataset is a 10% random sample of airline tickets throughout the United States, collected from reporting carriers. We determined that the data was trustworthy and reliable, given that it was collected and compiled by a government agency.

Before uploading to Kaggle, the data scientist condensed the dataset to flights that took place during 2018 and did some minor data tidying. He made his approach (as well as a step-by-step guide) to cleaning the dataset available on GitHub.

Basic data structure

The dataset (when filtered for LAX as the origin airport) consists of 415,297 observations of 14 variables, where each observation represents a single ticket order. A ticket order is a purchase made by a user for 1 or more tickets on a flight.

We will only be using 7 of the 14 variables included in the dataset. The other variables were arbitrary numbers that would not affect the price of an airplane ticket, for example the ID number for the order, the origin area code, the destination area code, etc.

us-airports.csv

This dataset contains some basic information about airports located within the United States. We will only be using it to determine the state/city in which each airport is located. This dataset was found on `ourairports.com`, and, after confirming some of the observations, we determined this data to be trustworthy.

It contains 29,658 observations of 23 variables. We will only be using three of these variables: `iata_code`, `state` (which we renamed from `region_name`), and `municipality`.

Explanation of variables

1. ***Quarter (int):*** The quarter refers to the time of the year during which the flight was flown. Quarter 1 refers to flights flown between January and March, quarter 2 refers to flights flown between April and June, quarter 3 refers to flights flown between July and September, and quarter 4 refers to flights flown between October and December.
2. ***Origin (chr):*** This variable refers to the three letter code of the airport out of which the flights depart from. The original data set consisted of 263 distinct origin cities, but we are focusing our analysis on flights only departing from LAX.
3. ***Dest (chr):*** This variable refers to the three letter code of the airport at which the flight lands. In our dataset, there are 119 distinct destinations of flights departing from LAX. As you will see in our exploratory data analysis later on, however, we focused most of our graphs on the top 10 destinations in terms of number of flights landing at these destinations in order to simplify our analysis and make our insights more directed and digestible.
4. ***Miles (int):*** This variable refers to the number of miles traveled during each flight. Each origin-destination combo corresponds with one value for the "miles" variable.
5. ***AirlineCompany (chr):*** This variable refers to the two-letter airline company code that the user used from start to finish. There are 12 distinct airlines departing from LAX in our dataset.
6. ***NumTicketsOrdered (num):*** This variable refers to the number of tickets purchased during the order. Since each row represents a single order, some orders may contain multiple tickets, resulting in a need for this variable.
7. ***PricePerTicket (num):*** This variable represents the price per ticket in the order. It is not the total cost of the order. This price is determined by a multitude of factors, many of which we will investigate in this project. **This is our response variable.**
8. ***iata_code (chr):*** Comes from `us-airports.csv`. It refers to the 3 letter abbreviation of an airport, as specified by the International Air Transport Association.
9. ***state (chr):*** Comes from `us-airports.csv`. It refers to the full name of the state where an airport is located.
10. ***municipality (chr):*** Comes from `us-airports.csv`. It refers to the name of the city/region where an airport is located.

Exploratory Data Analysis

Cleaning the data

Loading the data

First, we must begin by loading in our dataset. As mentioned earlier, it is over 9 million rows long. Therefore, we are going to read it in using `fread()`, a function of the `data.table` package to speed up this process.

```r
tickets <- fread("flight_data.csv")
```

Additionally, let's load in the `us-airports.csv`. This will allow us to see where each airport is located.

```
airports <- read_csv("us-airports.csv")
```

Joining the data

Now, we will join our two datasets together. We will perform a left join to make sure that we preserve all rows in `tickets`. We will join on the 3-letter airport abbreviation (for the destination), since this variable is present in both datasets.

```
tickets_joined <- tickets %>%
  left_join(airports, by = c("Dest" = "iata_code"))
```

Next, let's select all of the variables we will be using. We are going to rename `region_name` to `state` for clarity.

```
tickets_joined <- tickets_joined %>%
  select(Quarter, Origin, Dest, Miles, AirlineCompany, NumTicketsOrdered, PricePerTicket,
         region_name, municipality) %>%
  rename(state = region_name)
```

Filtering the data

Since we are focusing our analysis on LAX as the origin airport, let's filter our data to fit that restriction.

```
tickets_LAX <- tickets_joined %>%
  filter(Origin == "LAX")
```

Now, let's do a quick analysis for any NA values.

```
tickets_LAX %>%
  summarize_all(funs(sum(is.na(.))))
```

```
##   Quarter Origin Dest Miles AirlineCompany NumTicketsOrdered PricePerTicket
## 1       0      0    0     0              0                 0              0
##   state municipality
## 1   439          439
```

Because all of the NAs appear in the `state` and `municipality` columns, there is likely an issue caused by our left join. Let's investigate.

```
tickets_LAX %>%
  filter(is.na(state) | is.na(municipality)) %>%
  group_by(Dest) %>%
  summarize(n())
```

```
## # A tibble: 2 x 2
##   Dest  ‘n()‘
##   <chr> <int>
## 1 SJU     249
## 2 STT     190
```

We can see that we are missing states and regions that correspond to SJU (Luis Munoz Marin International Airport) and STT (Cyril E. King Airport). After some further research, we found that SJU is located in Puerto Rico, and STT is in the US Virgin Islands. Both of these areas are territories of the United States, not states. Therefore, it makes sense that `us-airports.csv` excludes these areas. Since the NA values will not interfere with our analysis for the most part, we will keep these observations in the data.

None of the other variables have NA values.

Structure of the data

Finally, let’s look at the structure and head of our data to make sure that everything is as expected.

```
tickets_LAX %>%
  str()
```

```
## Classes ’data.table’ and ’data.frame’:   415297 obs. of  9 variables:
##  $ Quarter          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Origin           : chr  "LAX" "LAX" "LAX" "LAX" ...
##  $ Dest             : chr  "ORD" "PHL" "PHL" "PHL" ...
##  $ Miles            : num  1744 2402 2402 2402 2402 ...
##  $ AirlineCompany   : chr  "AA" "AA" "AA" "AA" ...
##  $ NumTicketsOrdered: num  1 2 1 3 2 2 1 2 2 2 ...
##  $ PricePerTicket   : num  200.7 67 67.5 104 109 ...
##  $ state            : chr  "Illinois" "Pennsylvania" "Pennsylvania" "Pennsylvania" ...
##  $ municipality     : chr  "Chicago" "Philadelphia" "Philadelphia" "Philadelphia" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
tickets_LAX %>%
  head()
```

```
##    Quarter Origin Dest Miles AirlineCompany NumTicketsOrdered PricePerTicket
## 1:       1    LAX  ORD  1744             AA                 1         200.65
## 2:       1    LAX  PHL  2402             AA                 2          67.00
## 3:       1    LAX  PHL  2402             AA                 1          67.50
## 4:       1    LAX  PHL  2402             AA                 3         104.00
## 5:       1    LAX  PHL  2402             AA                 2         109.00
## 6:       1    LAX  PHL  2402             AA                 2         111.00
##           state municipality
## 1:     Illinois      Chicago
## 2: Pennsylvania Philadelphia
## 3: Pennsylvania Philadelphia
## 4: Pennsylvania Philadelphia
## 5: Pennsylvania Philadelphia
## 6: Pennsylvania Philadelphia
```

All of our variables appear to be in order. Now, we can move on to data visualizations!

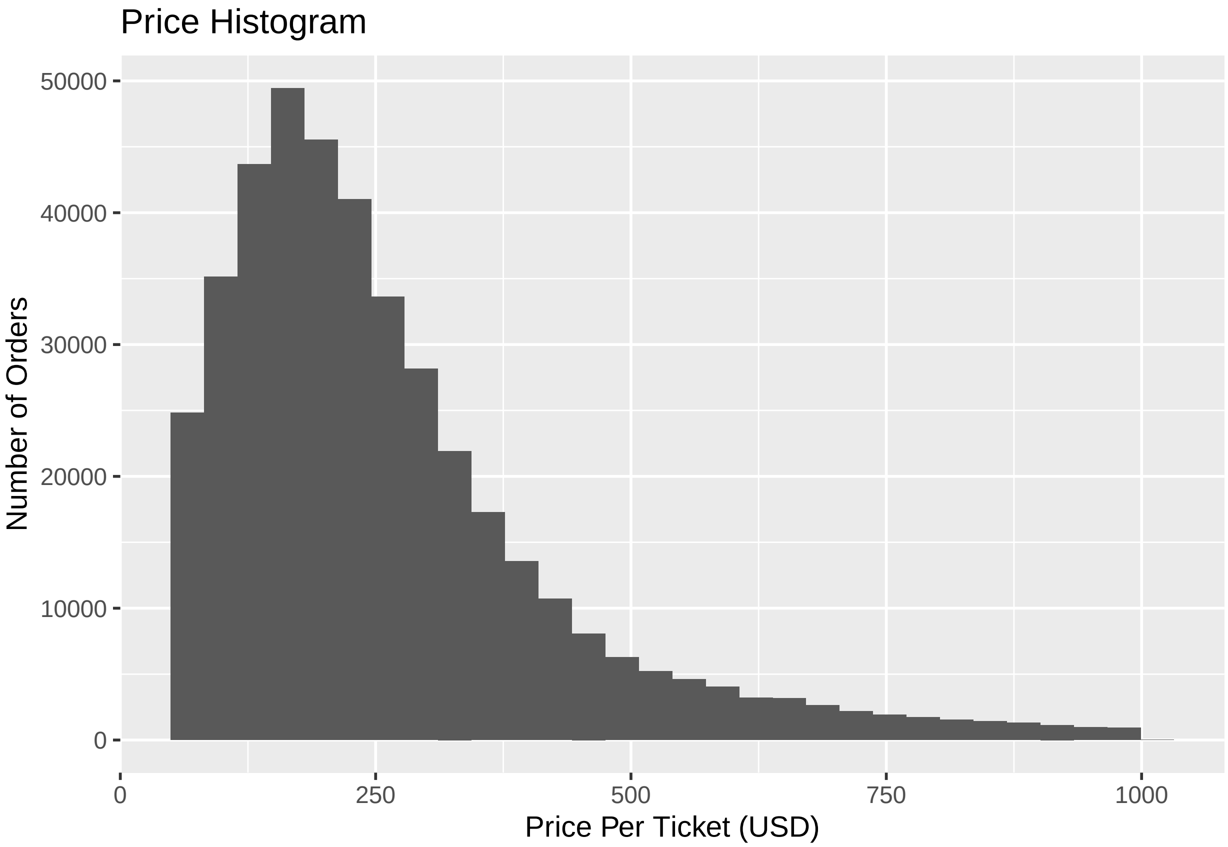
`PricePerTicket` distribution

Price Histogram

Let’s plot the distribution of the ticket prices, and examine it.

```
tickets_LAX %>%
  ggplot(aes(x = PricePerTicket)) +
  geom_histogram() +
  labs(title = "Price Histogram",
       x = "Price Per Ticket (USD)",
       y = "Number of Orders")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Additionally, let's examine some summary statistics.

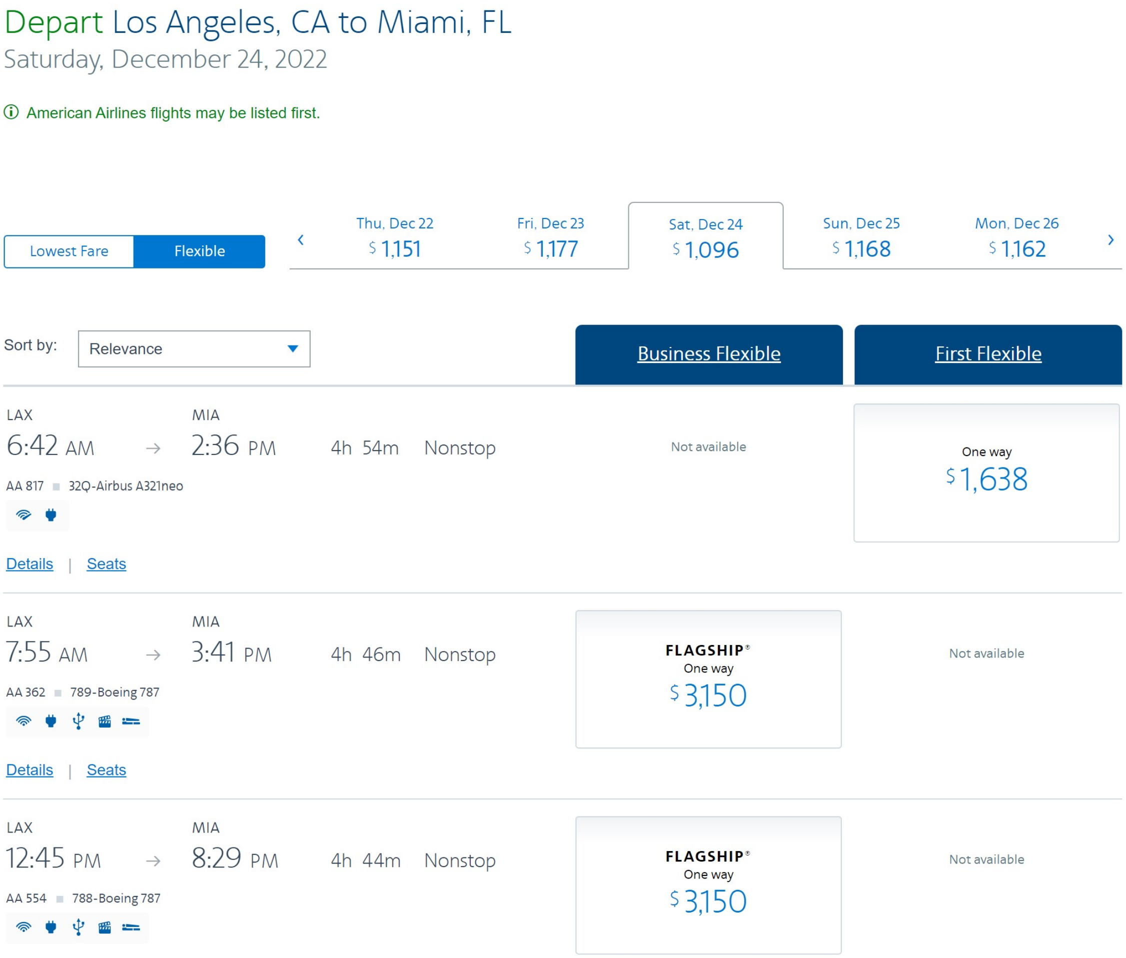
```
tickets_LAX %>%
  summarize(mean = mean(PricePerTicket),
            median = median(PricePerTicket),
            min = min(PricePerTicket),
            max = max(PricePerTicket),
            sd = sd(PricePerTicket))
```

```
##       mean median min  max       sd
## 1 262.1202  219.5  50 1000 167.9422
```

The histogram of price per ticket is unimodal and right-skewed, with a median of about 219.5 USD and a mean of about 262.1 USD. The minimum price was 50 USD and the maximum price was 1,000 USD. The standard deviation is about 167.9 USD, which is relatively large compared to the range of the data. This means that there is a large amount of variance within the `PricePerTicket` variable.

Examining the hard cap at 1,000 USD

It is also important to note that the data has a hard maximum value for price. We believe that some airplane tickets can easily be more expensive than 1,000 USD. We decided to make sure that this was the case. Below is an example ticket pricing from American Airlines. We selected a flight from Los Angeles to Miami, and picked business/first class. Also, this flight is scheduled for Christmas Eve, a time when travelling is at an extreme high.



Depart Los Angeles, CA to Miami, FL

Saturday, December 24, 2022

ⓘ American Airlines flights may be listed first.

Thu, Dec 22	Fri, Dec 23	Sat, Dec 24	Sun, Dec 25	Mon, Dec 26
$ 1,151	$ 1,177	$ 1,096	$ 1,168	$ 1,162

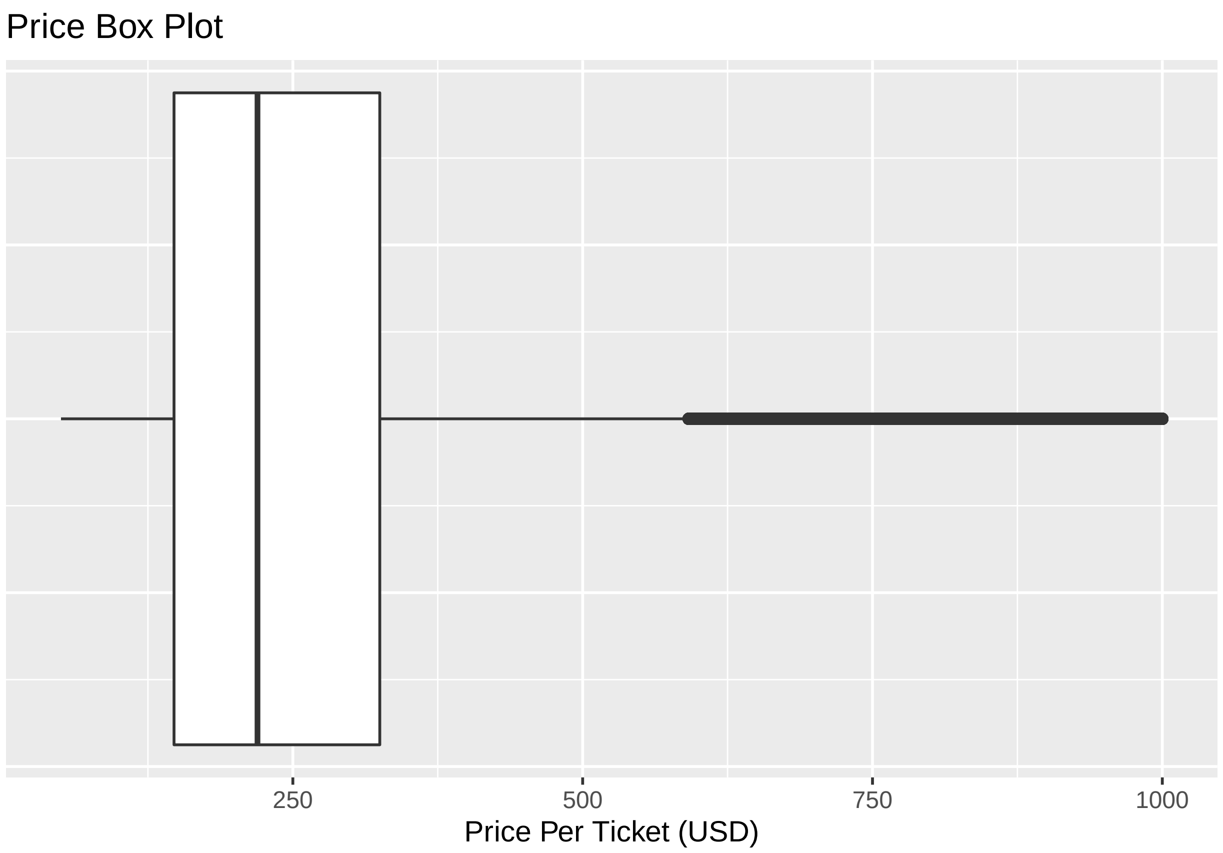
LAX	MIA			Business Flexible	First Flexible
6:42 AM AA 817 ■ 32Q-Airbus A321neo	2:36 PM	4h 54m	Nonstop	Not available	One way $ 1,638
7:55 AM AA 362 ■ 789-Boeing 787	3:41 PM	4h 46m	Nonstop	FLAGSHIP® One way $ 3,150	Not available
12:45 PM AA 554 ■ 788-Boeing 787	8:29 PM	4h 44m	Nonstop	FLAGSHIP® One way $ 3,150	Not available

We can see that these prices are well over 1,000 USD. Now, it is almost certain that the dataset was limited to prices under that value.

Examining outliers

Although the histogram is skewed right, there does not appear to be any outliers immediately noticeable. Let's examine this distribution with a box plot to take a further look.

```r
tickets_LAX %>%
  ggplot(aes(x = PricePerTicket)) +
  geom_boxplot() +
  labs(title = "Price Box Plot",
       x = "Price Per Ticket (USD)") +
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank())
```



Again, we see that the distribution is skewed right. The box plot shows that while the data does range from 50 USD to 1,000 USD, it has a much smaller interquartile range (IQR).

```
quantile(tickets_LAX$PricePerTicket, c(.25, .75))
```

```
##   25%   75%
## 147.5 325.0
```

By calculating the 25th (Q1) and 75th (Q3) percentile, we can see that 50% of the data is located between 147.5 USD and 325 USD. Additionally. the darker line located on the right side of the plot represents values that are identified as outliers. These outliers are calculated with the formula `1.5 * IQR + Q3`

```
IQR <- quantile(tickets_LAX$PricePerTicket, .75) - quantile(tickets_LAX$PricePerTicket, .25)
print(unname((1.5 * IQR) + quantile(tickets_LAX$PricePerTicket, .75,)))
```

```
## [1] 591.25
```

By this formula, any price above about 591.3 USD is considered an outlier. Although these values do skew the distribution of price, we have decided against removing them. Because airplane prices can become much more expensive than this, these outliers are not unreasonable values. Therefore, it is important to consider them in our analysis.

Most "Popular" Routes from LAX

Let's take a look at what the most "popular" destinations are when leaving from LAX. We will do this by seeing which orders occurred the most throughout 2018.

```
tickets_LAX_area <- tickets_LAX %>%
  group_by(Dest, state, municipality) %>%
  summarize(count = n()) %>%
  arrange(desc(count))
tickets_LAX_area %>%
  head(10)
```

```
## # A tibble: 10 x 4
## # Groups:   Dest, state [10]
##    Dest  state         municipality  count
##    <chr> <chr>         <chr>         <int>
##  1 JFK   New York      New York      26232
##  2 BOS   Massachusetts Boston        17888
##  3 ORD   Illinois      Chicago       17614
##  4 HNL   Hawaii        Honolulu      15948
##  5 ATL   Georgia       Atlanta       14442
##  6 SFO   California    San Francisco 13903
##  7 EWR   New Jersey    Newark        13873
##  8 MCO   Florida       Orlando       13528
##  9 DEN   Colorado      Denver        13337
## 10 SEA   Washington    Seattle       12439
```

From this, we can see the destinations where the most orders were purchased. JFK, which is located in New York City (a large urban area), has the highest count of orders with over 25,000. We can see that all of these airports are located in large cities and popular destinations

Although we can discover the most "popular" routes from LAX purely from numbers, it is important to gain a visual understanding of these numbers and regions. Let's create a map to examine the states in which the most flights arrived from LAX.

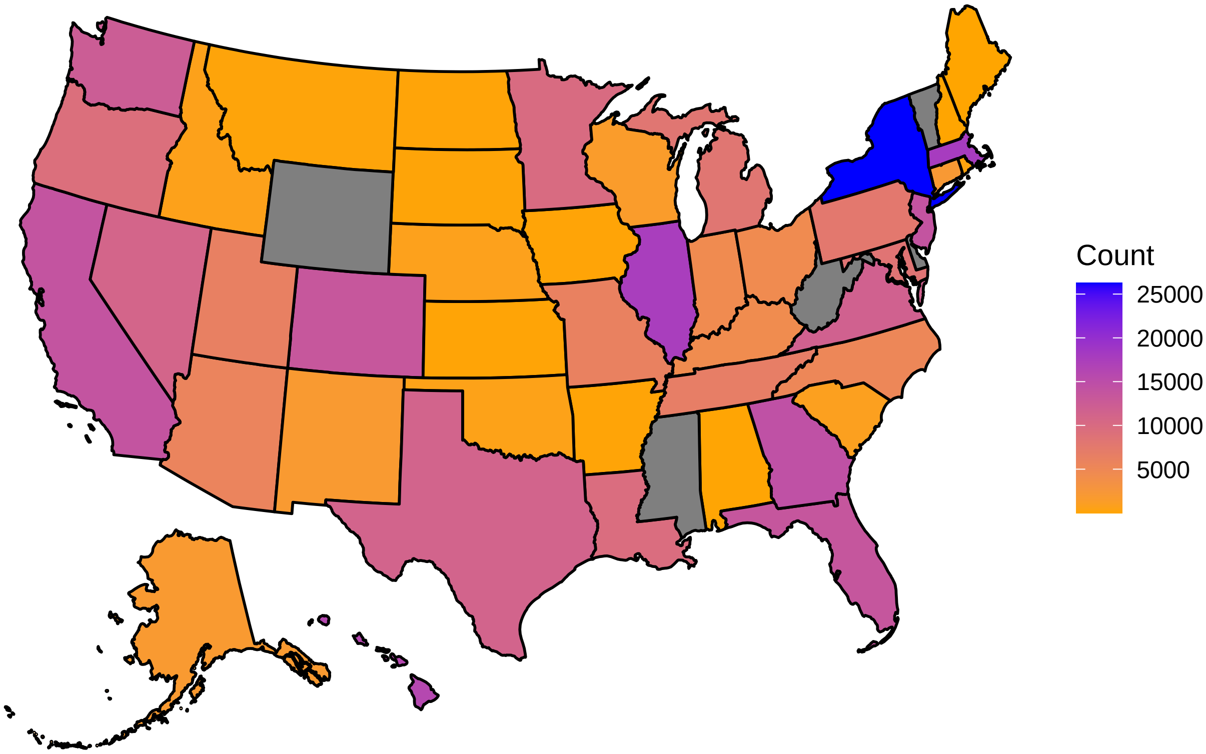
First, we'll need to load in and create some necessary tables

```
states_map <- us_map("state") %>%
  mutate(long = x, lat = y, region = tolower(full))
not_in_LAX <- setdiff(state.name, tickets_LAX_area$state)
tickets_LAX_area <- tickets_LAX_area %>%
  ungroup() %>%
  add_row(state = not_in_LAX) %>%
  mutate(state_lower = tolower(state))
```

Now, we can create the map!

```
tickets_LAX_area %>%
  ggplot(aes(map_id = state_lower)) +
  geom_map(aes(fill = count), map = states_map, color = "black") +
  expand_limits(x = states_map$long, y = states_map$lat) +
  scale_fill_gradient(low = "orange", high = "blue") +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank()) +
  labs(title = "Map: Flights to Each State",
       x = "", y = "", fill = "Count") +
  theme_void()
```

Map: Flights to Each State



In order to better understand the data in the map, we searched for the most popular routes including LAX. In 2018 Airfare watchdog, a site that finds the best available trips, examined more than 20,000 routes and recorded the data. They compiled a list of the most popular flyring routes within the United States. Within their top 10 list, these rankings included LAX:

- 8th: Los Angeles (LAX) – Denver (DEN) Colorado
- 7th. Los Angeles (LAX) – Seattle (SEA) Washington
- 6th: Los Angeles (LAX) – Las Vegas (LAS) Nevada
- 4th: Los Angeles (LAX) – Chicago (ORD) Illinois
- 2nd: Los Angeles (LAX) – San Francisco (SFO) California
- 1st: Los Angeles (LAX) – New York (JFK) New York

Our data and map don't fully correspond to the top 10 list because we, unlike the site, we examined only flights out of LAX instead of both ways. Taking this into consideration, there were a few popular states in our map that weren't mentioned in the site: Florida, Georgia, Hawaii, Massachusetts, Texas, Virginia, and New Jersey.

- New Jersey and Massachusetts (much like New York) provide transportation to NE America whilst Florida and Georgia provide transportation to SE America. These states allow individuals to most conveniently access different parts of this country.
- Hawaii is an island, so it can be accessed only by plane/boat and California is the closest state to Hawaii. Vacations makes flights to Hawaii extremely popular.
- Texas contains many large cities and its airports provide access to the south and Virginia's airports provide access to the middle of the east states.

Additionally, let's take a look at the states where we have no flights leaving from LAX.

```
print(not_in_LAX)
```

```
## [1] "Delaware"      "Mississippi"   "Vermont"       "West Virginia"
## [5] "Wyoming"
```

After a bit of research, we discovered that Delaware has no major, national airport. This would explain why there are not flights to that state. Additionally, after researching flights online, we were unable to find direct routes from LAX to any airports in Mississippi, Vermont, West Virginia, or Wyoming. Although it is still possible to fly to these states from LAX, a connection flight would be neccessary.

Examining Miles Traveled

After examining this map, we wondered if flights that have shorter distance to travel (i.e., destinations closer to LAX) will cost less for the traveler.

Destination and mileage

First of all, while examining the data, we noticed that each destination only had one unique mileage associated with it. Let's confirm that this is the case with a short bit of code.

```
tickets_LAX %>%
  group_by(Dest, Miles) %>%
  summarize() %>%
  head(10)
```

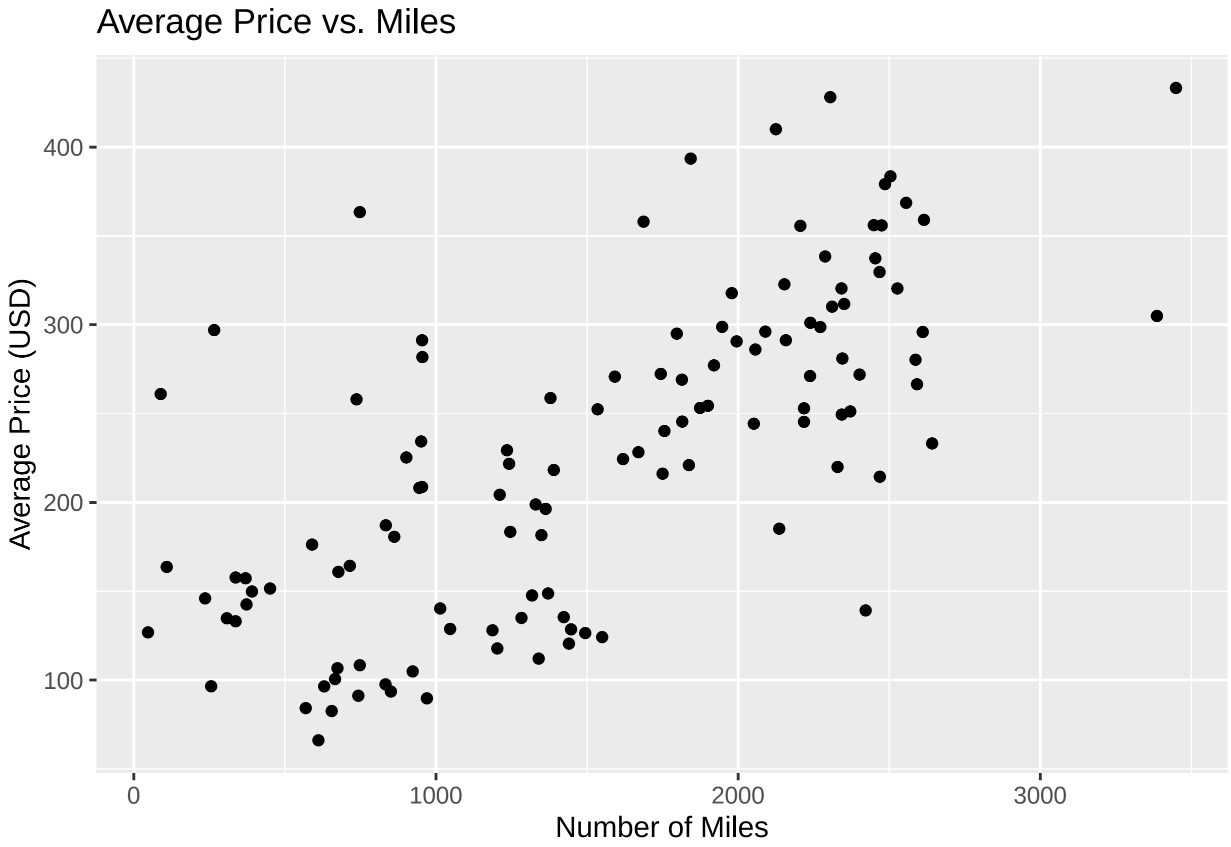
```
## # A tibble: 10 x 2
## # Groups:   Dest [10]
##    Dest  Miles
##    <chr> <dbl>
##  1 ABQ     677
##  2 AGS    2090
##  3 ALB    2468
##  4 AMA     955
##  5 ANC    2345
##  6 ASE     737
##  7 ATL    1947
##  8 AUS    1242
##  9 BDL    2527
## 10 BHM    1815
```

After examining this table, we confirmed that no destination appeared more than once. Therefore, each destination has a singular value for mileage associated with it.

Miles vs. PricePerTicket

Now that we have established that each destination corresponds with one mileage, we know that a single value of `Miles` will corresponded with multiple values for `PricePerTicket`, since prices differ even though the destination may be the same. This is because of various factors like the time of the year, the airline company, and more. Thus, in order to arrive at a singular value for price for each distinct value of miles, we found the average price for each distinct value of miles.

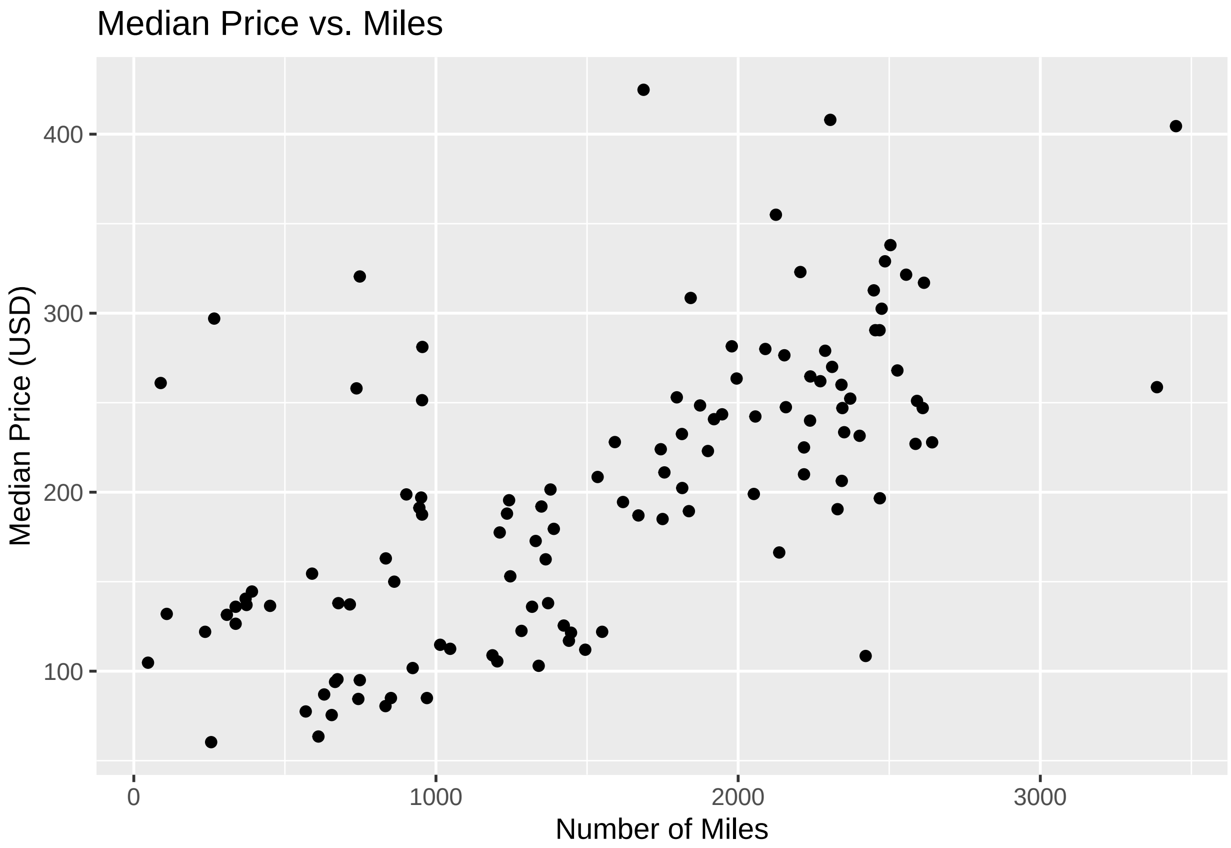
```
tickets_LAX %>%
  group_by(Dest) %>%
  mutate(mean = mean(PricePerTicket)) %>%
  select(Miles, mean, Dest) %>%
  distinct() %>%
  ggplot(aes(x = Miles, y = mean)) +
    geom_point() +
  labs(title = "Average Price vs. Miles",
       x = "Number of Miles",
       y = "Average Price (USD)")
```



The resulting scatter plot of average price vs number of miles shows a fairly strong, positive linear relationship between price and miles. This makes sense because of several reasons, as noted by Via Travelers (https://viatravelers.com/what-affects-flight-prices/): longer flights means more fuel needed, paying the crew more, increased overflight fees, extra food and drinks, among others.

Sometimes, it's important to look at the difference between the mean and median of a variable. The mean is skewed easily, so it might misrepresent the data. Let's examine this same graph, except this time we will look at the Median Price.

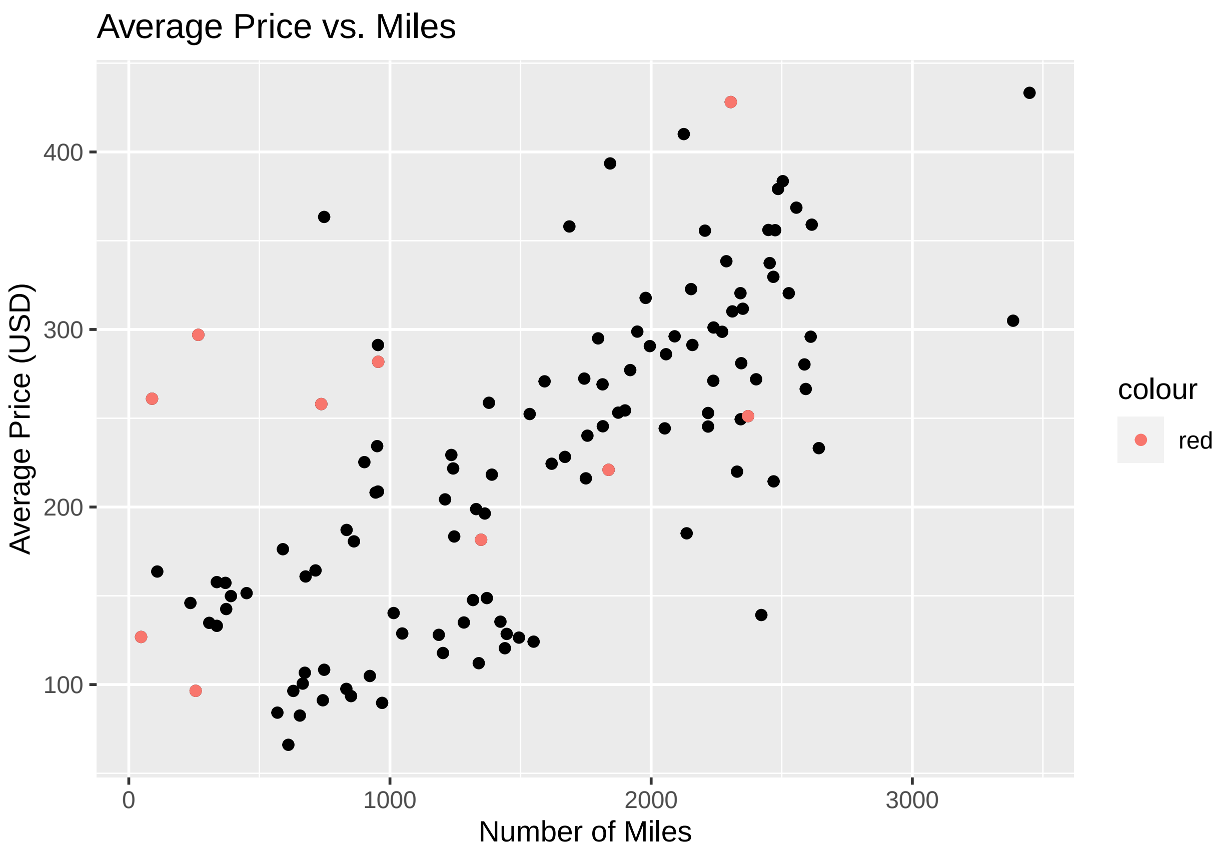
```
tickets_LAX %>%
  group_by(Dest) %>%
  mutate(median = median(PricePerTicket)) %>%
  select(Miles, median, Dest) %>%
  distinct() %>%
  ggplot(aes(x = Miles, y = median)) +
  geom_point() +
  labs(title = "Median Price vs. Miles",
       x = "Number of Miles",
       y = "Median Price (USD)")
```



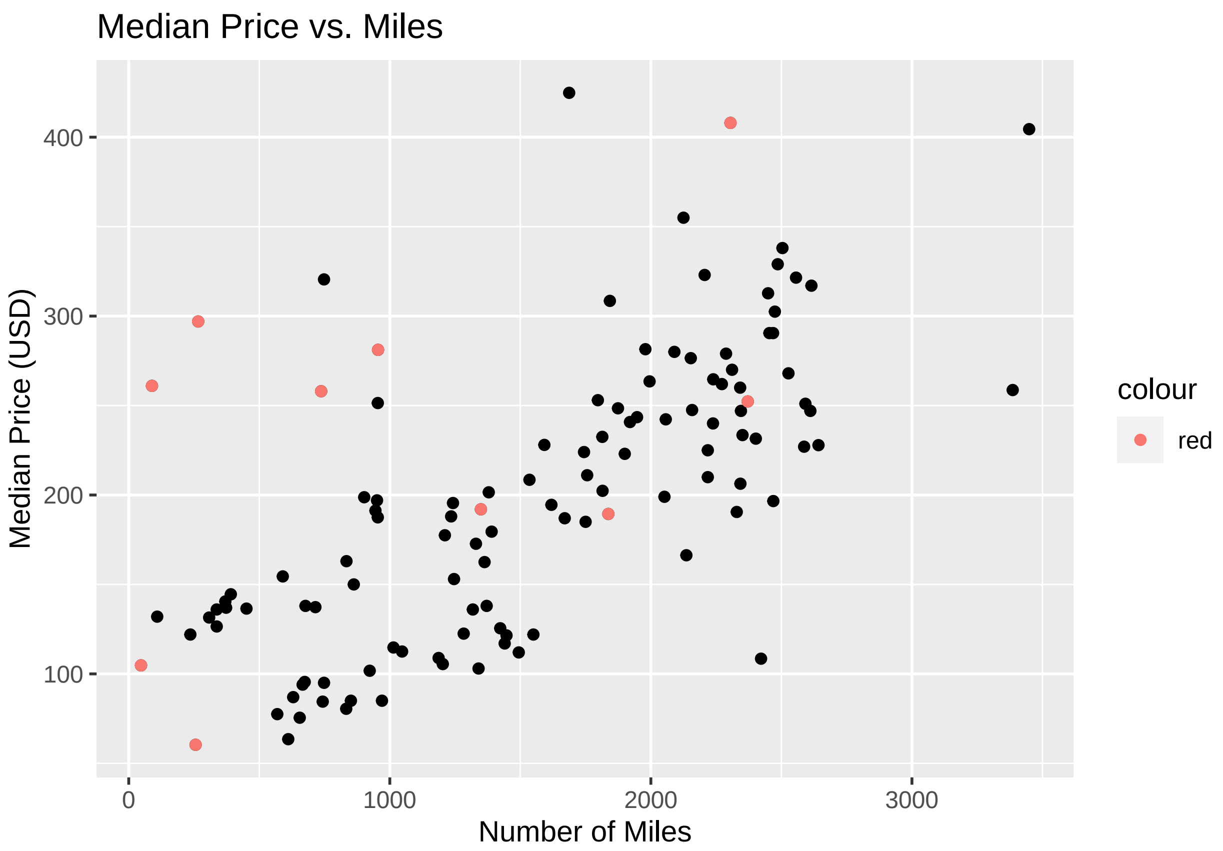
After looking at this graph, in addition to the graph of average price vs. miles, we noticed that there are some points that do not fit the trend of this data. After a bit of digging, we discovered that those destinations (each data point represents one destination) have limited observations in the dataset. Let's find out which airports have less than 10 observations and label them on our graphs.

```
less_10_flights <- tickets_LAX %>%
  group_by(Dest) %>%
  mutate(mean = mean(PricePerTicket), median = median(PricePerTicket), sum = n()) %>%
  select(Dest, Miles, mean, median, sum) %>%
  distinct() %>%
  filter(sum < 10)

tickets_LAX %>% group_by(Dest) %>%
  mutate(mean = mean(PricePerTicket)) %>%
  select(Miles, mean, Dest) %>%
  distinct() %>%
  ggplot(aes(x = Miles, y = mean)) +
  geom_point() +
  geom_point(data = less_10_flights, aes(x = Miles, y = mean, color = "red")) +
  labs(title = "Average Price vs. Miles",
       x = "Number of Miles",
       y = "Average Price (USD)")
```



```
tickets_LAX %>% group_by(Dest) %>%
  mutate(median = median(PricePerTicket)) %>%
  select(Miles, median, Dest) %>%
  distinct() %>%
  ggplot(aes(x = Miles, y = median)) +
  geom_point() +
  geom_point(data = less_10_flights, aes(x = Miles, y = median, color = "red")) +
  labs(title = "Median Price vs. Miles",
       x = "Number of Miles",
       y = "Median Price (USD)")
```



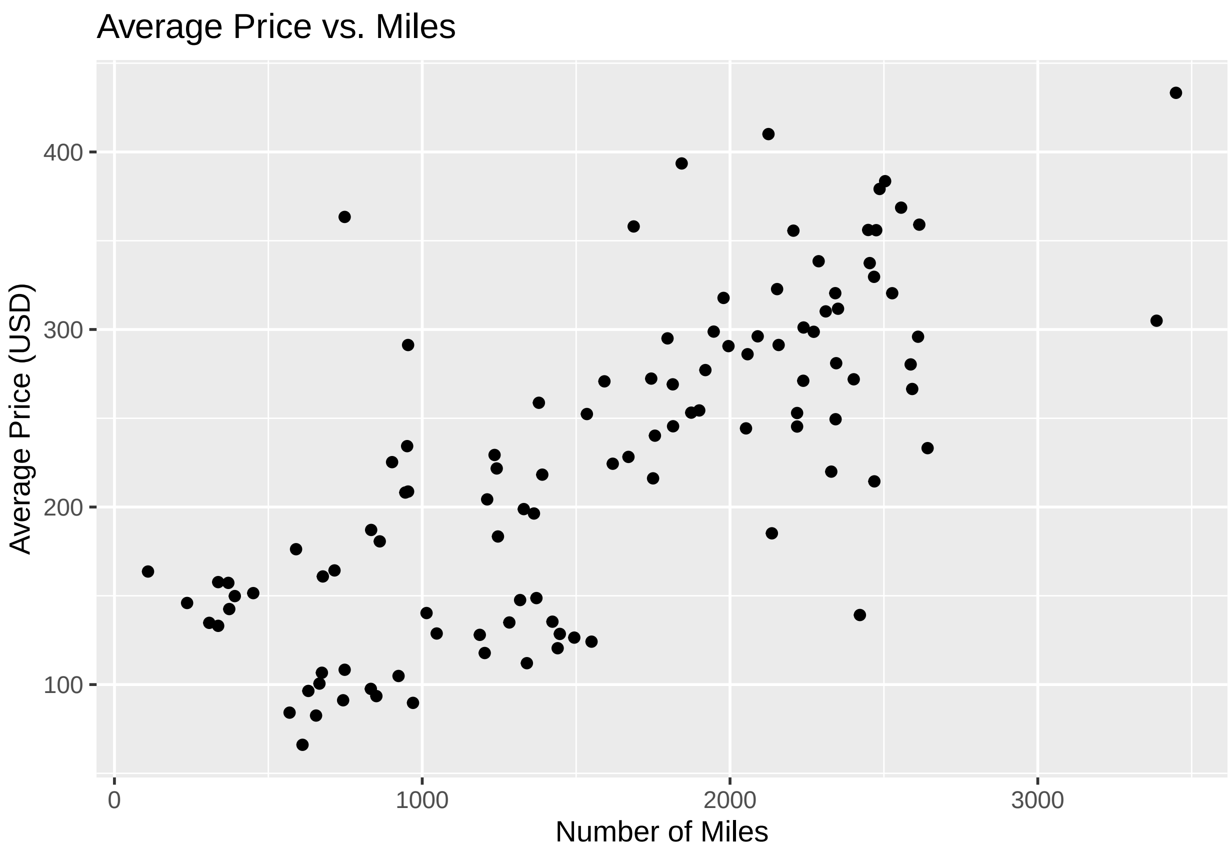
We can see that, by visualizing these data points, many of them do not fit the trend of this data. Because they have limited observations, we will remove them from our data.

```
tickets_LAX <- tickets_LAX %>%
  group_by(Dest) %>%
  mutate(num_obs = n()) %>%
  filter(num_obs >= 10)
```

Let's examine our graphs once more to make sure everything looks correct!

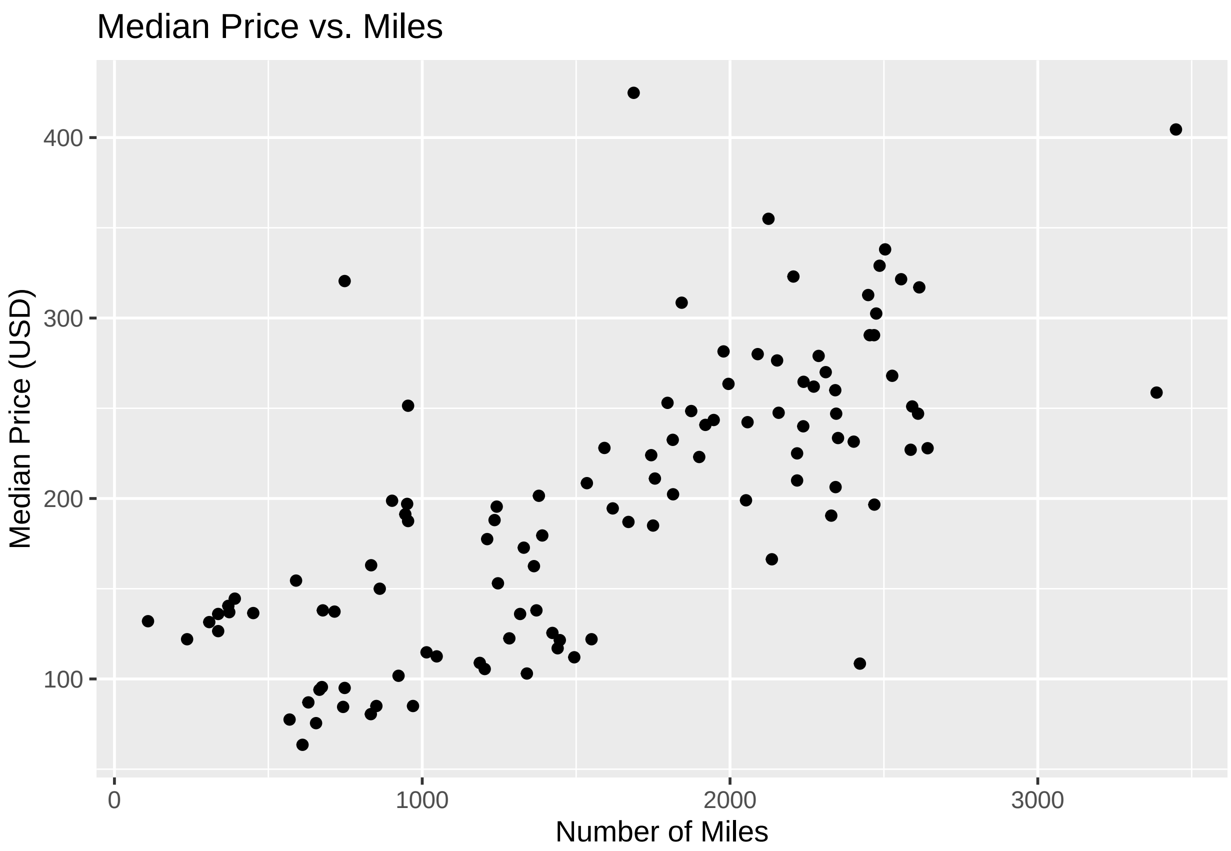
```
LAX_miles_price_mean <- tickets_LAX %>%
  group_by(Dest) %>%
  mutate(mean = mean(PricePerTicket)) %>%
  select(Miles, mean, Dest) %>%
  distinct()

ggplot(LAX_miles_price_mean, aes(x = Miles, y = mean)) +
  geom_point() +
  labs(title = "Average Price vs. Miles",
       x = "Number of Miles",
       y = "Average Price (USD)")
```



```r
LAX_miles_price_med <- tickets_LAX %>%
  group_by(Dest, Miles) %>%
  mutate(median = median(PricePerTicket)) %>%
  select(Miles, median, Dest) %>%
  distinct()

ggplot(LAX_miles_price_med, aes(x = Miles, y = median)) +
  geom_point() +
  labs(title = "Median Price vs. Miles",
       x = "Number of Miles",
       y = "Median Price (USD)")
```



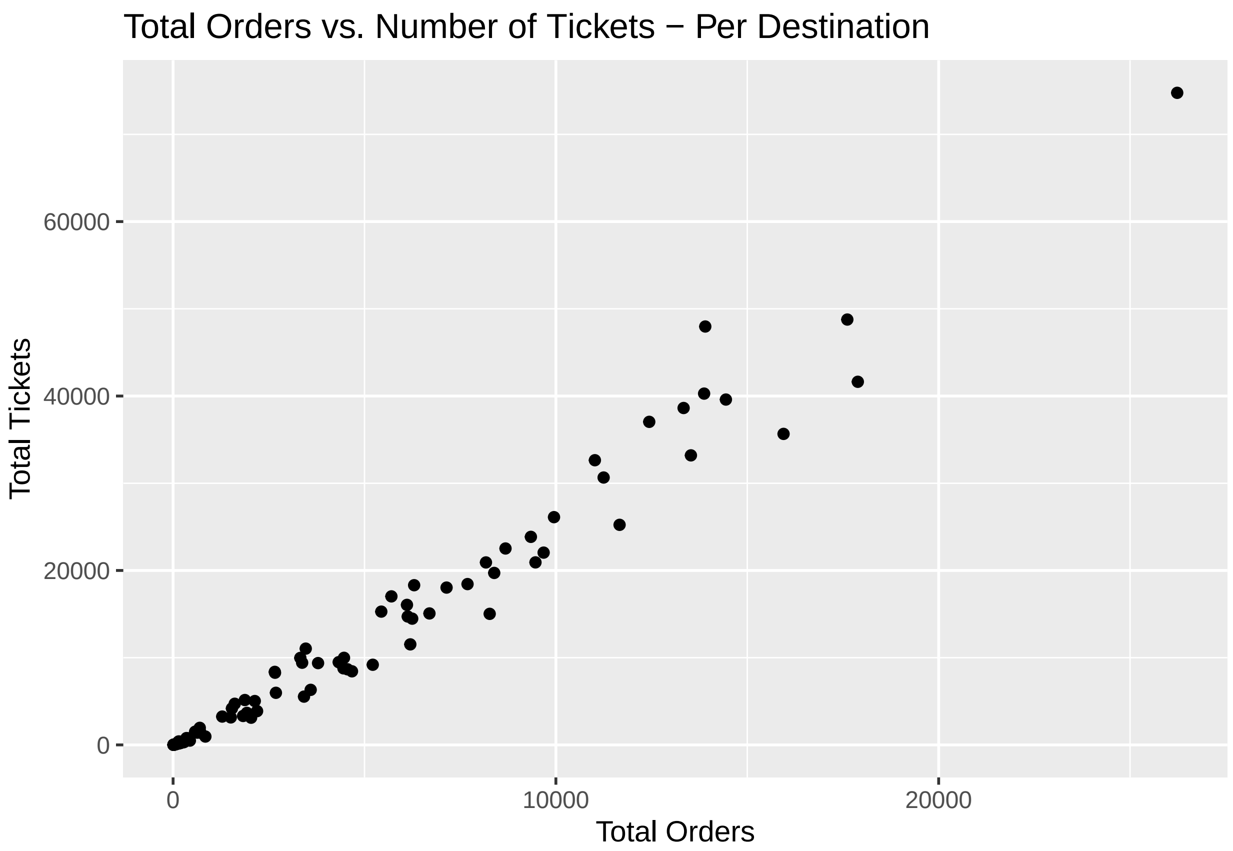
Selecting the Top 10

For the remainder of our exploratory data analysis, we'd like to focus on the "top 10" most popular destinations from LAX. As explored previously, we can find this using the pure number of observations for each destination. However, we wondered if using the `NumTicketsOrdered` variable would be a better way to determine this.

Let's examine a scatter plot comparing the two to see the difference between them.

```r
LAX_orders_tickets <- tickets_LAX %>%
  group_by(Dest) %>%
  summarize(total_orders = n(), total_tickets = sum(NumTicketsOrdered))

ggplot(LAX_orders_tickets, aes(x = total_orders, y = total_tickets)) +
  geom_point() +
  labs(title = "Total Orders vs. Number of Tickets - Per Destination",
       x = "Total Orders",
       y = "Total Tickets")
```



From this, we see that there is a very strong, positive correlation between the number of tickets ordered and the number of orders. Therefore, we have decided that using either method to determine the top 10 is appropriate. We will use the same method utilized previously.

```
top_10 <- tickets_LAX %>%
  group_by(Dest) %>%
  summarize(count = n()) %>%
  arrange(desc(count)) %>%
  select(Dest) %>%
  head(10) %>%
  pull()
top_10
```

```
##  [1] "JFK" "BOS" "ORD" "HNL" "ATL" "SFO" "EWR" "MCO" "DEN" "SEA"
```

Here, we can see the destination airports that will be included in the `top_10_LAX` dataset.

```
top_10_LAX <- tickets_LAX %>%
  filter(Dest %in% top_10)
```

Now, lets look at our new dataset!

```
top_10_LAX %>%
  head()
```

```
## # A tibble: 6 x 10
## # Groups:   Dest [5]
##   Quarter Origin Dest  Miles AirlineComp~1 NumTi~2 Price~3 state munic~4 num_obs
##     <int> <chr>  <chr> <dbl> <chr>           <dbl>   <dbl> <chr> <chr>     <int>
## 1       1 LAX    ORD    1744 AA                  1    201. Illi~ Chicago   17614
## 2       1 LAX    DEN     862 AA                  1    257. Colo~ Denver    13337
## 3       1 LAX    ORD    1744 AA                  1    193. Illi~ Chicago   17614
## 4       1 LAX    ATL    1947 AA                  1    445. Geor~ Atlanta   14442
## 5       1 LAX    BOS    2611 AA                  1    767. Mass~ Boston    17888
## 6       1 LAX    JFK    2475 AA                  1    757. New ~ New Yo~   26232
## # ... with abbreviated variable names 1: AirlineCompany, 2: NumTicketsOrdered,
## #   3: PricePerTicket, 4: municipality
```

```
top_10_LAX %>%
  str
```

```
## grouped_df [159,204 x 10] (S3: grouped_df/tbl_df/tbl/data.frame)
##  $ Quarter          : int [1:159204] 1 1 1 1 1 1 1 1 1 1 ...
##  $ Origin           : chr [1:159204] "LAX" "LAX" "LAX" "LAX" ...
##  $ Dest             : chr [1:159204] "ORD" "DEN" "ORD" "ATL" ...
##  $ Miles            : num [1:159204] 1744 862 1744 1947 2611 ...
##  $ AirlineCompany   : chr [1:159204] "AA" "AA" "AA" "AA" ...
##  $ NumTicketsOrdered: num [1:159204] 1 1 1 1 1 1 1 1 1 1 ...
##  $ PricePerTicket   : num [1:159204] 201 257 193 445 767 ...
##  $ state            : chr [1:159204] "Illinois" "Colorado" "Illinois" "Georgia" ...
##  $ municipality     : chr [1:159204] "Chicago" "Denver" "Chicago" "Atlanta" ...
##  $ num_obs          : int [1:159204] 17614 13337 17614 14442 17888 26232 14442 17614 17888 15948 ...
##  - attr(*, "groups")= tibble [10 x 2] (S3: tbl_df/tbl/data.frame)
##   ..$ Dest : chr [1:10] "ATL" "BOS" "DEN" "EWR" ...
##   ..$ .rows: list<int> [1:10]
##   .. ..$ : int [1:14442] 4 7 111 112 113 114 115 989 990 991 ...
##   .. ..$ : int [1:17888] 5 9 57 58 59 61 116 117 118 119 ...
##   .. ..$ : int [1:13337] 2 2151 2152 2153 2154 2155 2156 2157 2158 2159 ...
##   .. ..$ : int [1:13873] 6087 6088 6089 6090 7093 7094 7095 7096 7097 7098 ...
##   .. ..$ : int [1:15948] 10 1777 2528 2529 2530 2531 2532 2533 2534 2535 ...
##   .. ..$ : int [1:26232] 6 62 63 64 65 66 67 68 69 70 ...
##   .. ..$ : int [1:13528] 60 879 4159 4160 4161 4162 4163 4164 4165 4166 ...
##   .. ..$ : int [1:17614] 1 3 8 6091 6092 6093 6094 6096 6097 9377 ...
##   .. ..$ : int [1:12439] 11 12 13 14 15 16 17 18 19 20 ...
##   .. ..$ : int [1:13903] 5577 5578 5579 5580 5581 5582 5583 5584 6095 10195 ...
##   .. ..@ ptype: int(0)
##   ..- attr(*, ".drop")= logi TRUE
```

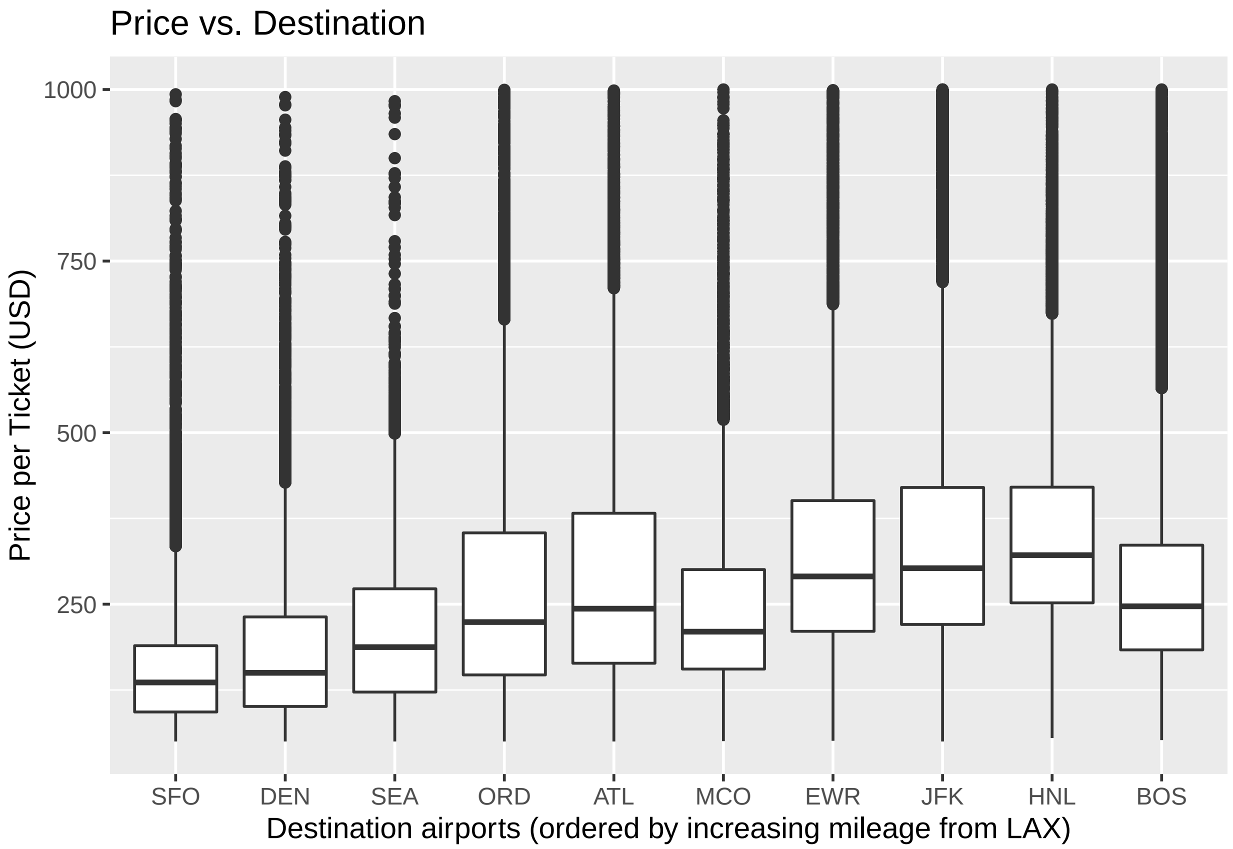
We see that our data has been reduced to a little less than 160,000 observations. This is much more manageable and will lend itself to much better analysis. **Unless otherwise stated, we will use this dataset from this point forward in the exploratory data analysis.**

`Dest` affecting `PricePerTicket`

Using our new `top_10_LAX` dataset, let's examine how these destinations differ in the price per ticket. We will use box plots to look at the variance within the `PricePerTicket` variable for each destination

Because these destinations have no inherent order to them, we will arrange them by their distance from LAX, in terms of mileage.

```r
top_10_LAX %>%
  ggplot(aes(x = reorder(Dest, +Miles), y = PricePerTicket)) +
  geom_boxplot() +
  labs(title = "Price vs. Destination",
       x = "Destination airports (ordered by increasing mileage from LAX)",
       y = "Price per Ticket (USD)")
```



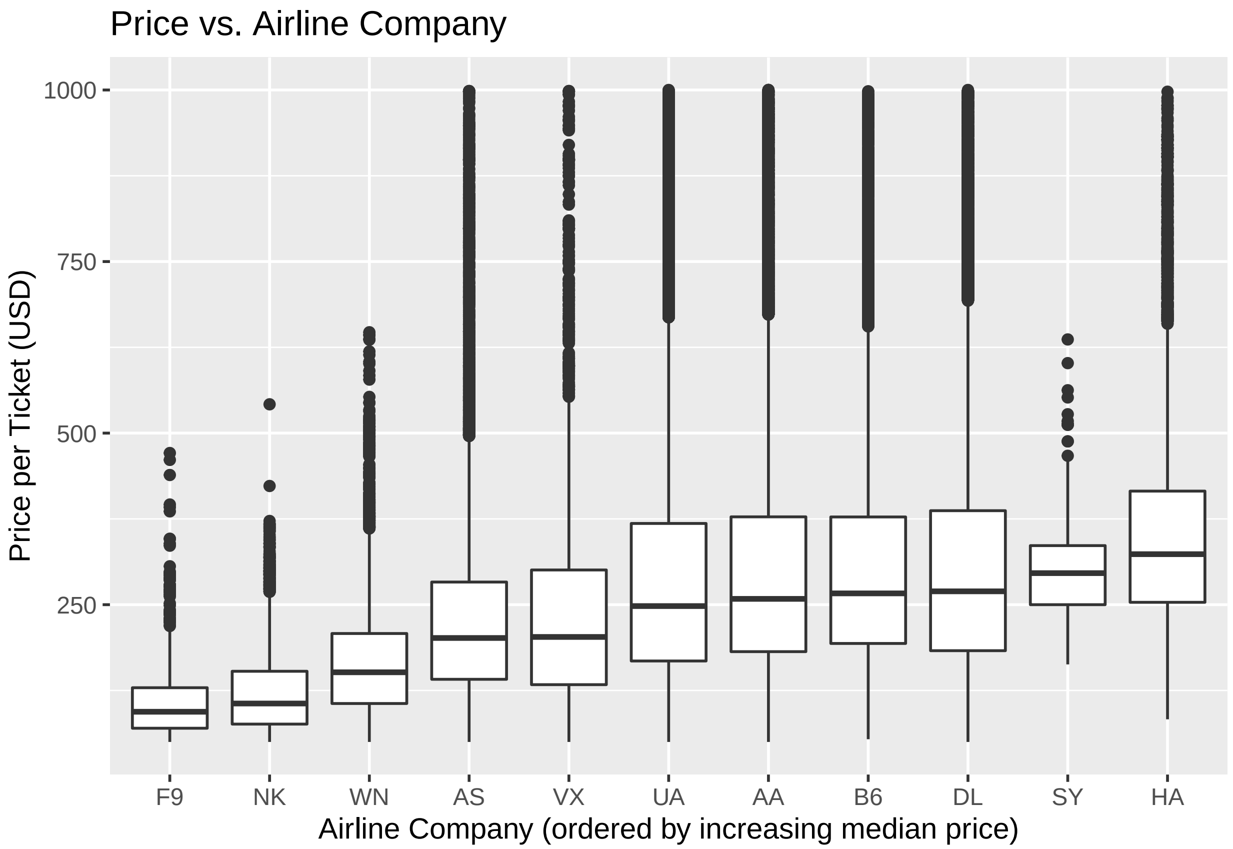
These box plots show the distribution of for each destination in the top 10 destinations. Just based on the order, we can see that SFO is the closest to LA and Boston is the furthest from LA.

There does seem to be a slight positive correlation as the number of miles increases; however, MCO (Orlando International Airport) and BOS (Boston Logan International Airport) do not necessarily follow this pattern. We will investigate and explain this relationship later.

`AirlineCompany` affecting `PricePerTicket`

Let's look for a similar relationship between the `PricePerTicket` and the airlines that serve our top 10 destinations! We will order these by increasing median price.

```r
top_10_LAX %>%
  group_by(AirlineCompany) %>%
  mutate(med_airline_price = median(PricePerTicket)) %>%
  ggplot(aes(x = reorder(AirlineCompany, +med_airline_price), y = PricePerTicket)) +
  geom_boxplot() +
  labs(title = "Price vs. Airline Company",
       x = "Airline Company (ordered by increasing median price)",
       y = "Price per Ticket (USD)")
```



We see that F9 (Frontier Airlines) and NK (Spirit Airlines) are clearly the cheapest airlines, with median prices of well below 125 USD. WN (Southwest Airlines) is another airline that stands out as being cheap with a median price of 151.5 USD. Upon some research, these statistical results make sense. Frontier Airlines' Wikipedia article states that it is a "major ultra low-cost U.S. carrier headquartered in Colorado", Spirit Airlines' Wikipedia article states that it is a "major ultra low-cost U.S. carrier headquartered in Florida", and Southwest Airlines' Wikipedia article states that it is the "world's largest low-cost carrier."

On the other end of the spectrum, the airline with the highest median price is HA (Hawaiian Airlines). This makes sense because of several reasons, as cited by The Cold Wire (https://www.thecoldwire.com/why-are-flights-to-hawaii-so-expensive/): Hawaii is comparatively far from the rest of the country, Hawaii is a high-in-demand vacation destination, fuel costs in Hawaii are high because of its limited supply, among other reasons.

An interesting airline to consider is that of SY (Sun Country Airlines). Its median price of $296 USD is very high but its maximum price is only 636.5 USD, which is well below the maximum prices of most other airlines. With some research, we find that Sun Country's Wikipedia article states that it is an "American ultra-low-cost passenger and cargo airline," which is interesting because it has the second highest median price. Upon further analysis, we find that the mean price for Sun Country Airlines is 302.2 USD, which is still fairly high but relatively lower compared to its positioning when comparing the median prices of the airlines. With more research, we discovered that in early 2018, Sun Country Airlines redesigned their airplane interior, removing first class cabins. This is one possible explanation as to why the maximum price is lower.

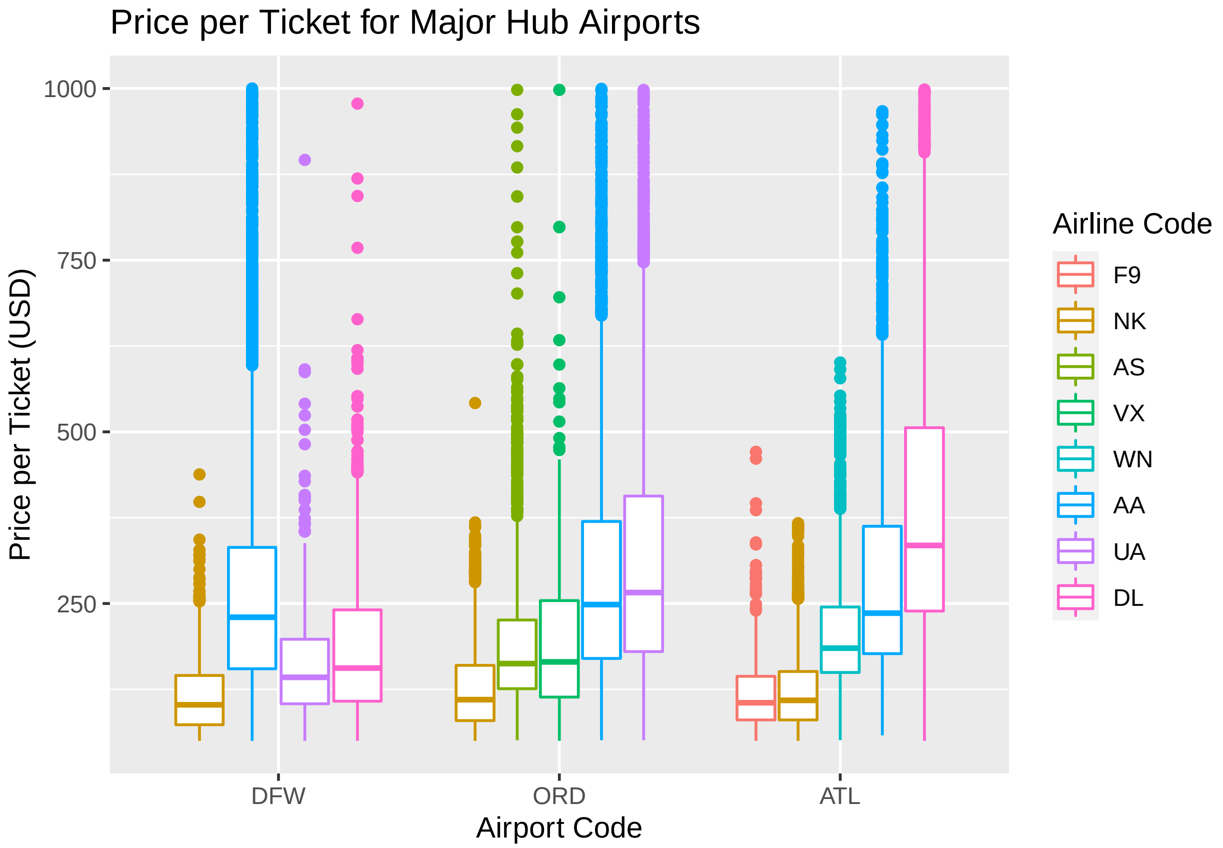
Examining Major Hub Airports

An extremely important aspect of our exploratory data analysis is examining how one variable could affect the `PricePerTicket` *in conjunction with* another variable. We will accomplish by looking at the major hub airports. We decided to look at the respective hubs for three airlines that we have flown the most with:

- Chicago O'Hare International Airport (ORD) - Primary hub for United Airlines
- Hartsfield-Jackson Atlanta International Airport (ATL) - Primary hub for Delta Airlines
- Dallas/Fort Worth International Airport (DFW) - Primary hub for American Airlines

Because DFW is not contained in the `top_10_LAX` dataset, we will be using `tickets_LAX` to create this graph. Additionally, we will order the destinations by increasing mileage, and order the airports by increasing *overall* median, similar to what was done in the previous two graph.

```r
tickets_LAX %>%
  filter(Dest %in% c("ORD", "ATL", "DFW")) %>%
  group_by(AirlineCompany) %>%
  mutate(med_airline_price = median(PricePerTicket)) %>%
  ggplot(aes(x = reorder(Dest, +Miles), y = PricePerTicket,
             color = reorder(AirlineCompany, +med_airline_price))) +
  geom_boxplot() +
  labs(title = "Price per Ticket for Major Hub Airports",
       x = "Airport Code",
       y = "Price per Ticket (USD)",
       color = "Airline Code")
```

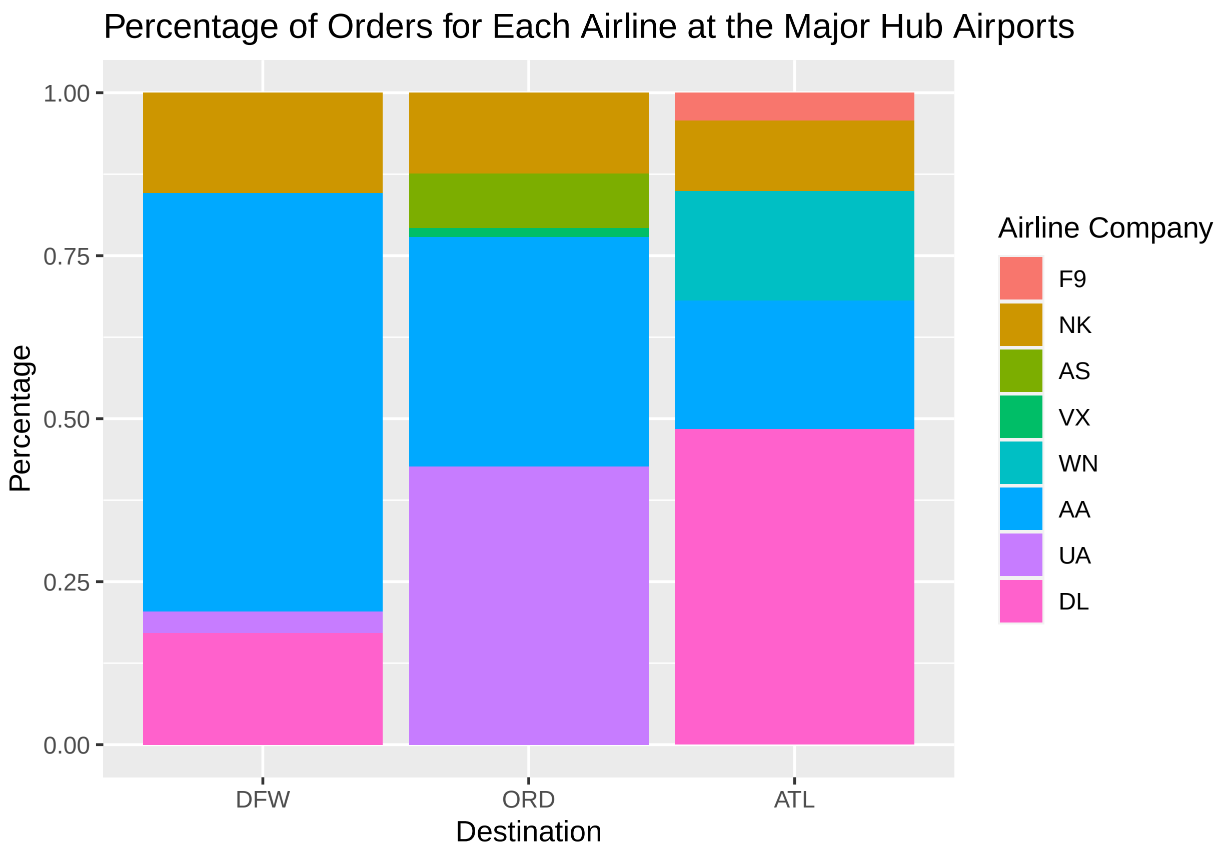


From the graph, we noticed that for the destination of Dallas, American Airlines is the most expensive airline, for the destination of Chicago, United Airlines is the most expensive airline, and for the destination of Atlanta, Delta Airlines is the most expensive airline. The most expensive airline for each airport is the airline for which that airport is a primary hub.

This discovery that airlines with airport hubs have the highest prices for these destinations compared to other airlines makes sense. For instance, The Cold Wire (https://www.thecoldwire.com/why-is-delta-so-expensive/#:~:text=For%20example%2C%20Delta%20Airlines%20has,airline%20to%20compete%20with%20it.) notes that Delta Airlines is one of few major airline in Atlanta and, "because of this, Delta's ticket prices to and from Atlanta, specifically, tend to be more expensive than ticket prices elsewhere." Competition is less tight because Delta owns a much larger percent of the market share of ATL, and can therefore raise their prices at that location.

In fact, let's examine this further by looking at the number of orders for flights arriving at each of these airports, separated by airline.

```r
tickets_LAX %>%
  filter(Dest %in% c("ORD", "ATL", "DFW")) %>%
  group_by(AirlineCompany) %>%
  mutate(med_airline_price = median(PricePerTicket)) %>%
  ggplot(aes(x = reorder(Dest, +Miles),
             fill = reorder(AirlineCompany, +med_airline_price))) +
  geom_bar(position = "fill") +
  labs(title = "Percentage of Orders for Each Airline at the Major Hub Airports",
      x = "Destination", y = "Percentage", fill = "Airline Company")
```

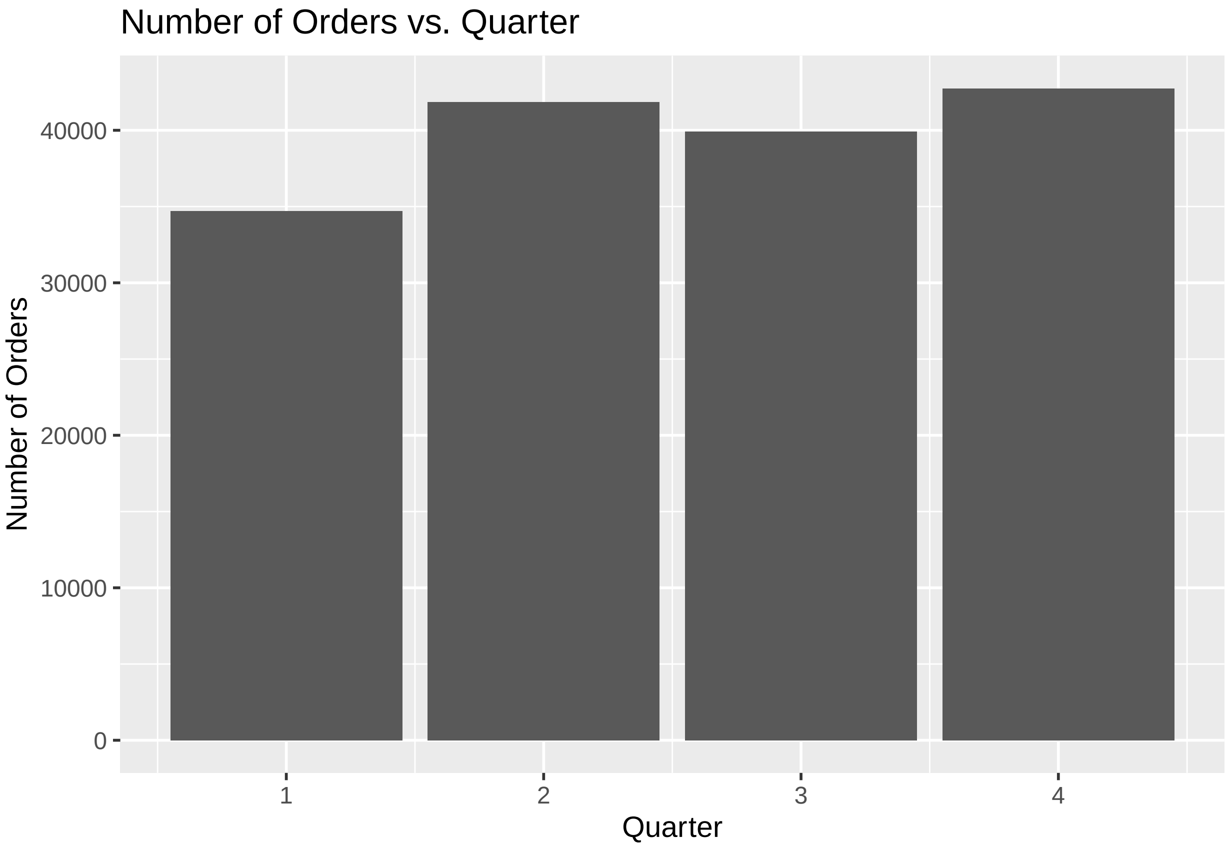


We can make some very interesting observations from this graph, as well! We see that for DFW, American Airlines controls the majority of flights entering from LAX. The same applies for ATL, where the majority of entering flights are with Delta Airlines. However, we see that both American and United Airlines each run nearly 40% of the flights arriving at ORD (United Airline's percentage is marginally larger). After a bit of research, we found that ORD is a secondary hub airport for American Airlines Therefore, it follows that American Airlines would control a large portion of flights entering ORD.

`Quarter` affecting `PricePerTicket`

Let's first examine how the `Quarter` variable is distributed with a bar chart.

```
top_10_LAX %>%
  ggplot(aes(x = Quarter)) +
  geom_bar() +
  labs(title = "Number of Orders vs. Quarter",
       y = "Number of Orders")
```

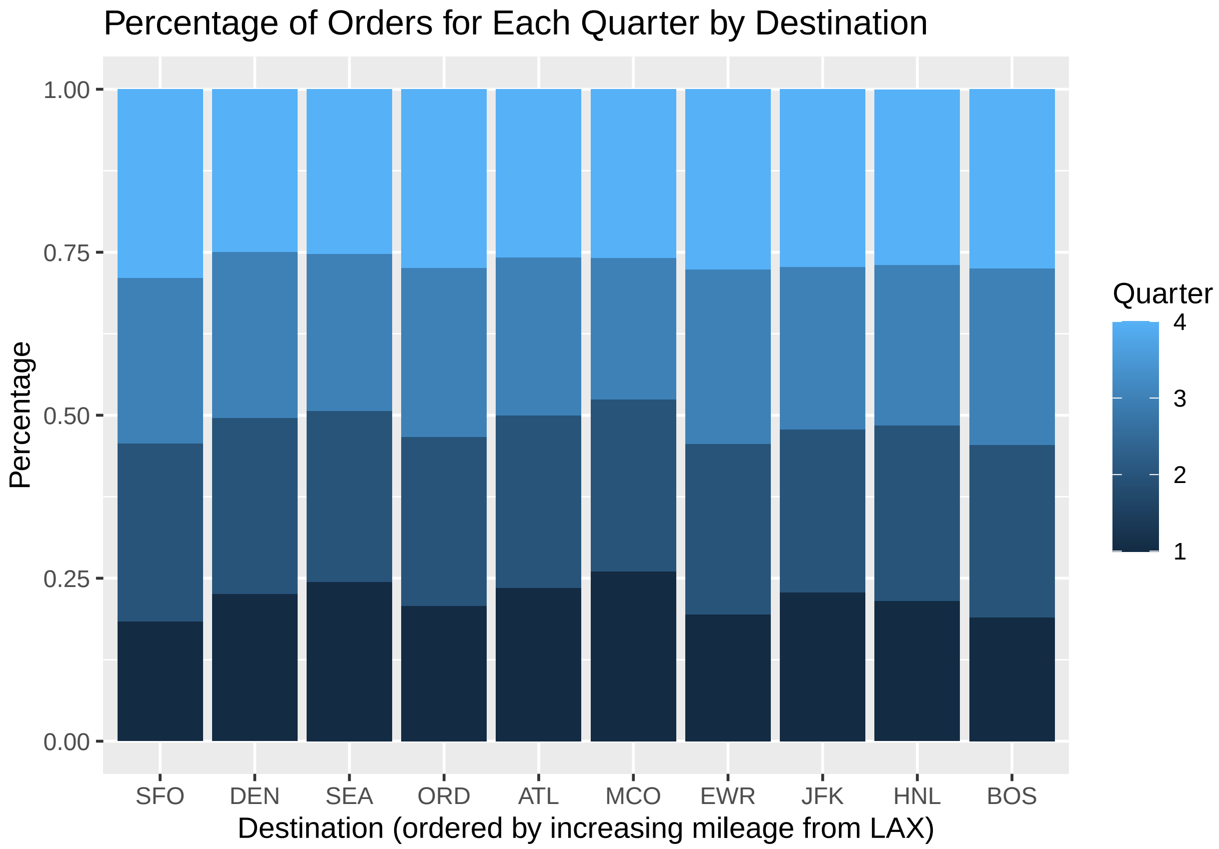


The number of orders per quarter seems relatively similar, at around the 4,000 orders per quarter range. However, quarter 1 does seem to have a notably lower number of orders with slightly less than 3,500 that quarter. We hypothesize that this is because there is less reason for most people to fly during the months of January to March. For Christmas holidays, people tend to fly towards the end of December, which is not accounted for in quarter 1. The only holiday during quarter 1 would be spring break/Easter break, but this does not affect many working individuals because they usually continue with their work during these months. The times of the year during which working individuals take flights for holidays is usually during the summer and winter, which is why we see quarter 2 and quarter 4 having the most number of orders.

Let's see if any destination has more orders during a particular quarter.

```
top_10_LAX %>%
  group_by(Dest, Miles, Quarter) %>%
  summarize(count = n()) %>%
  ggplot(aes(x = reorder(Dest, +Miles), y = count, fill = Quarter)) +
  geom_bar(position = "fill", stat = "identity") +
  labs(title = "Percentage of Orders for Each Quarter by Destination",
       x = "Destination (ordered by increasing mileage from LAX)",
       y = "Percentage")
```

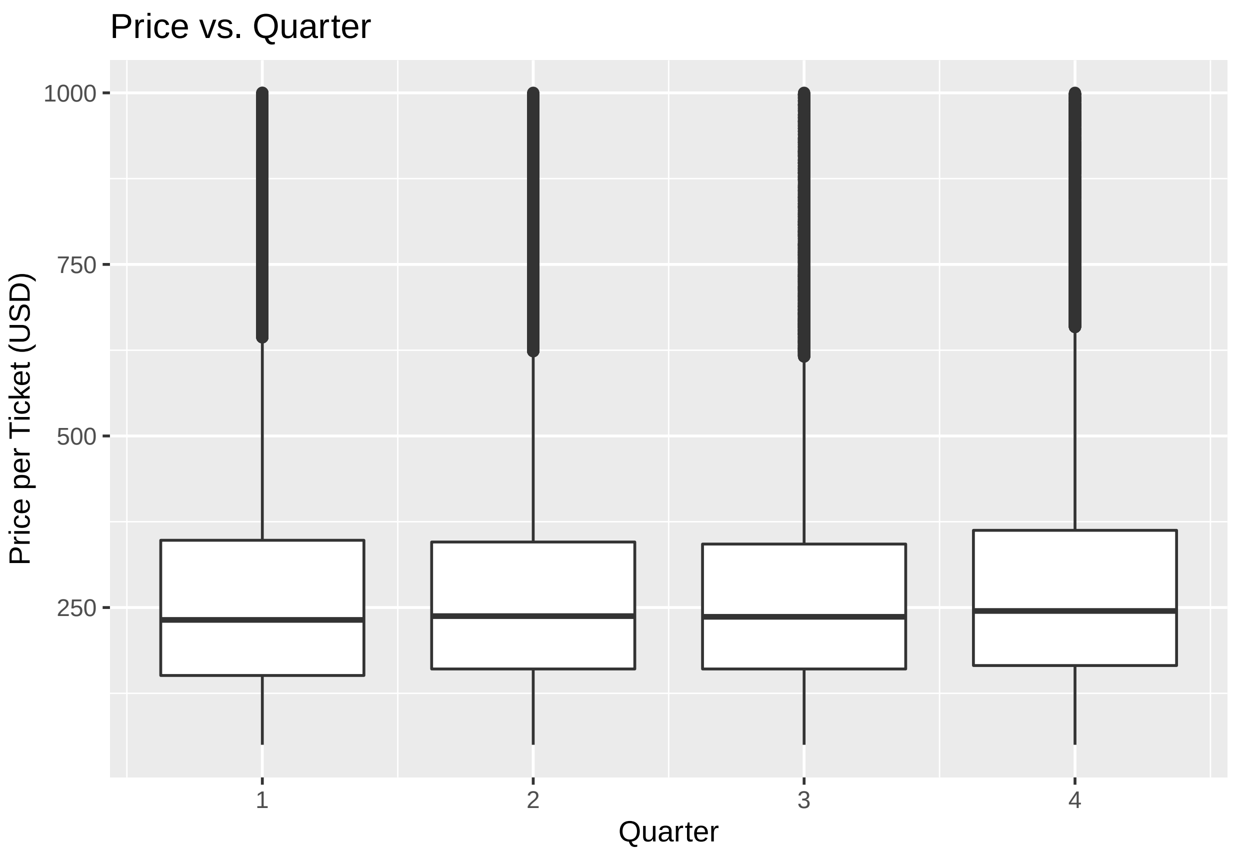
```
## 'summarise()' has grouped output by 'Dest', 'Miles'. You can override using the
## '.groups' argument.
```



Again, it is difficult to make any strong conclusions from this graph. Each percentage is relatively close to the rest, and there are no airports that stick out as outliers during a particular quarter.

With that, let's examine how the quarter might relate to the price per ticket.

```
top_10_LAX %>%
  ggplot(aes(group = Quarter, x = Quarter, y = PricePerTicket)) +
  geom_boxplot() +
  labs(title = "Price vs. Quarter",
       x = "Quarter",
       y = "Price per Ticket (USD)")
```



The box plots of price to quarter look very similar, suggesting that the quarter during which flights from LAX flew to the top 10 destinations in 2018 might have no impact on the price of airline tickets. It may be worth noting that the IQRs of the plots from Q1 and Q4 are slightly larger than those of the Q2 and Q3 box plots. Although the quarter may have no noticeable impact on the price per ticket, it is certainly possible that the month, week, day, or time will all impact the price. Those variables are not present in our dataset, so we are unable to analyze them, but further research into those attributes would be interesting.

The number of Airlines serving a destination

As noted in our explanation of hub airports, the amount of competition of airlines at an airport could definitely impact the price of a ticket to that airport. We'd like to examine this fact further.

First, let's find out how many airlines serve each of our top 10 destinations.

```
top_num_airlines <- top_10_LAX %>%
  group_by(Dest, AirlineCompany) %>%
  summarize() %>%
  group_by(Dest) %>%
  summarize(num_airlines = n()) %>%
  arrange(desc(num_airlines))
top_num_airlines
```

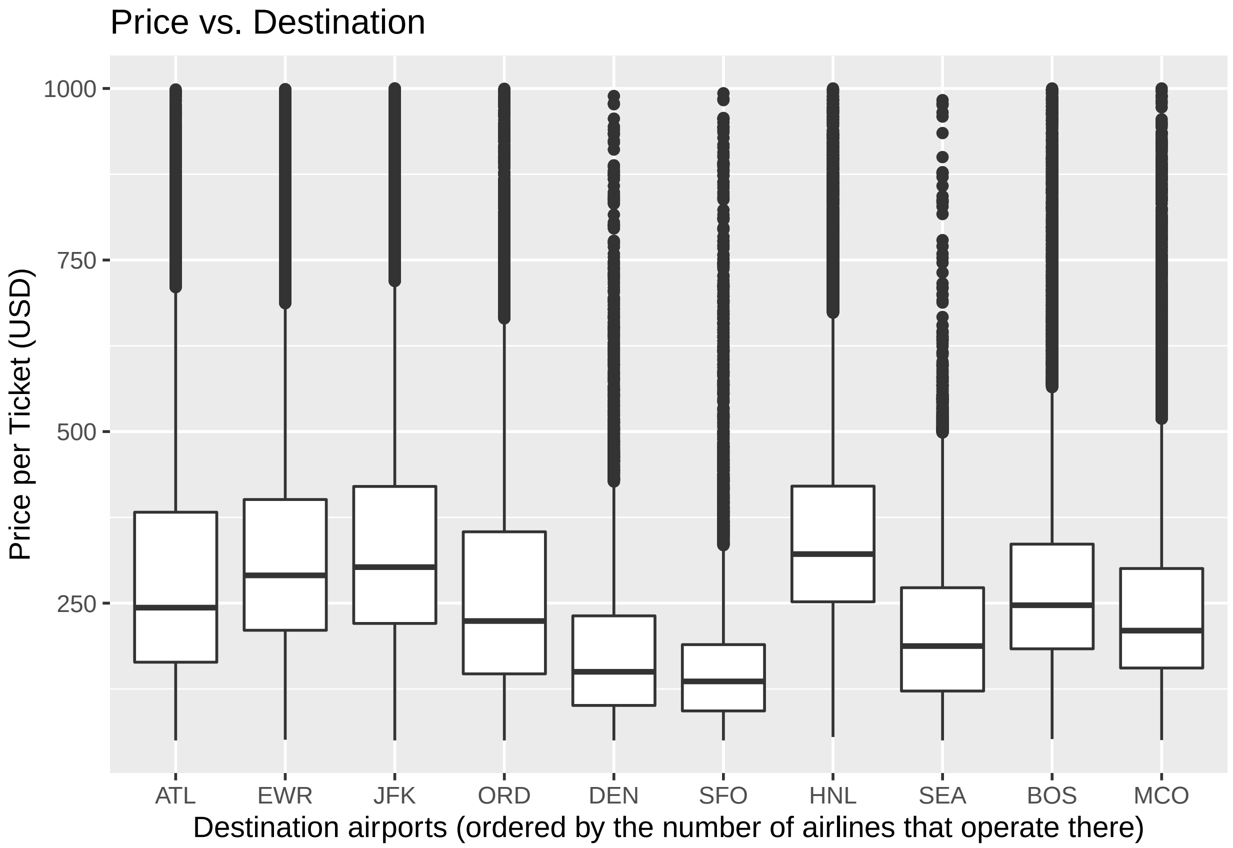
```
## # A tibble: 10 x 2
##    Dest  num_airlines
##    <chr>        <int>
##  1 MCO              9
##   2 BOS              8
##   3 HNL              7
##   4 SEA              7
##   5 DEN              6
##   6 SFO              6
##   7 ATL              5
##   8 EWR              5
##   9 JFK              5
##  10 ORD              5
```

Now, let's join these tables together to add `num_airlines` as a variable to `top_10_LAX`.

```
top_10_LAX <- top_10_LAX %>%
  left_join(top_num_airlines, by = "Dest")
```

With that done, we can examine how this variable might affect the price! Let's create a box plot to examine the `Dest` and `PricePerTicket` variables. We have created this plot earlier, but now we will reorder the destinations by the number of airlines that operate at that location.

```
top_10_LAX %>%
  ggplot(aes(x = reorder(Dest, +num_airlines), y = PricePerTicket)) +
  geom_boxplot() +
  labs(title = "Price vs. Destination",
       x = "Destination airports (ordered by the number of airlines that operate there)",
       y = "Price per Ticket (USD)")
```



Before constructing the box plots, we hypothesized that there would be a downward trend in the median price as we move toward looking at destinations with more number of airlines flying to those destinations due to the law of supply: the greater the number of suppliers, the lower the price of the goods/service. However, looking at our data, there is only a small, if any, apparent relationship.

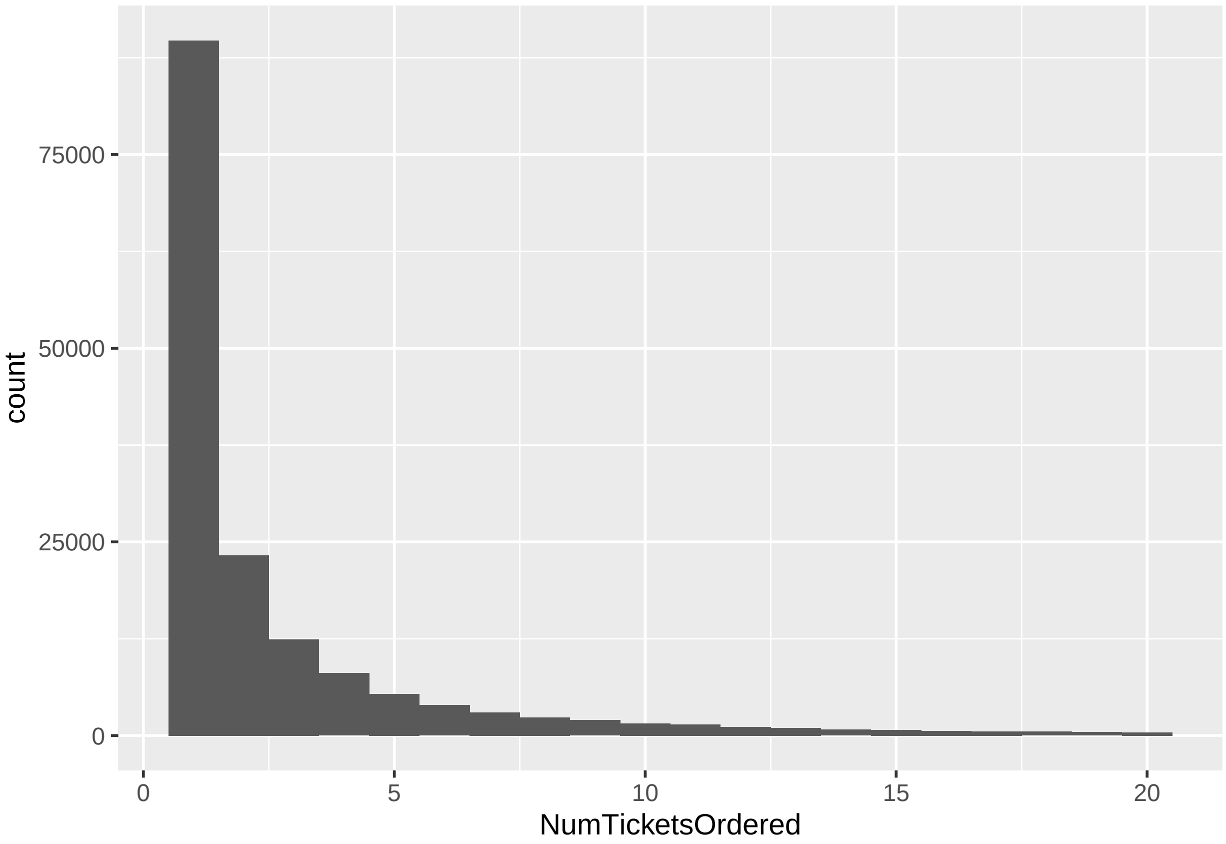
This could be perplexing at first, but it's important to note that there are so many other factors at play here that are not shown in the graph. The biggest one would be the number of miles from LAX to each of the destinations; in an earlier graph, we saw that there was a positive correlation between miles and price. So, we can safely conclude that the number of miles to each destination has a significantly larger effect on price than the number of airlines flying from LAX to each destination. This is not to say, however, that the number of airlines to each destination does not impact the price to that destination; we just don't have sufficient statistical evidence from this limited sample of 10 destinations to make a conclusive claim.

One extremely important observation, however, is that MCO (Orlando International Airport) and BOS (Boston Logan International Airport) are the top 2 airports in terms of number of airlines operating there. Earlier, when analyzing how `PricePerTicket` was affected by `Dest`, we ordered the destinations by miles away from LAX to search for a trend. We found a trend, but noted that MCO and BOS didn't exactly fit within it. The reason for that disparity may be the number of airlines. BOS and MCO both have a high amount of airline competition, which may be the reason that their prices were lower than expected when evaluating the trend.

`NumTicketsOrdered` affecting `PricePerTicket`

Finally, let's examine the number of tickets placed during an order. We'll begin by looking at the distribution of this variable.

```
top_10_LAX %>%
  ggplot(aes(x = NumTicketsOrdered)) +
  geom_histogram(bins = 20)
```

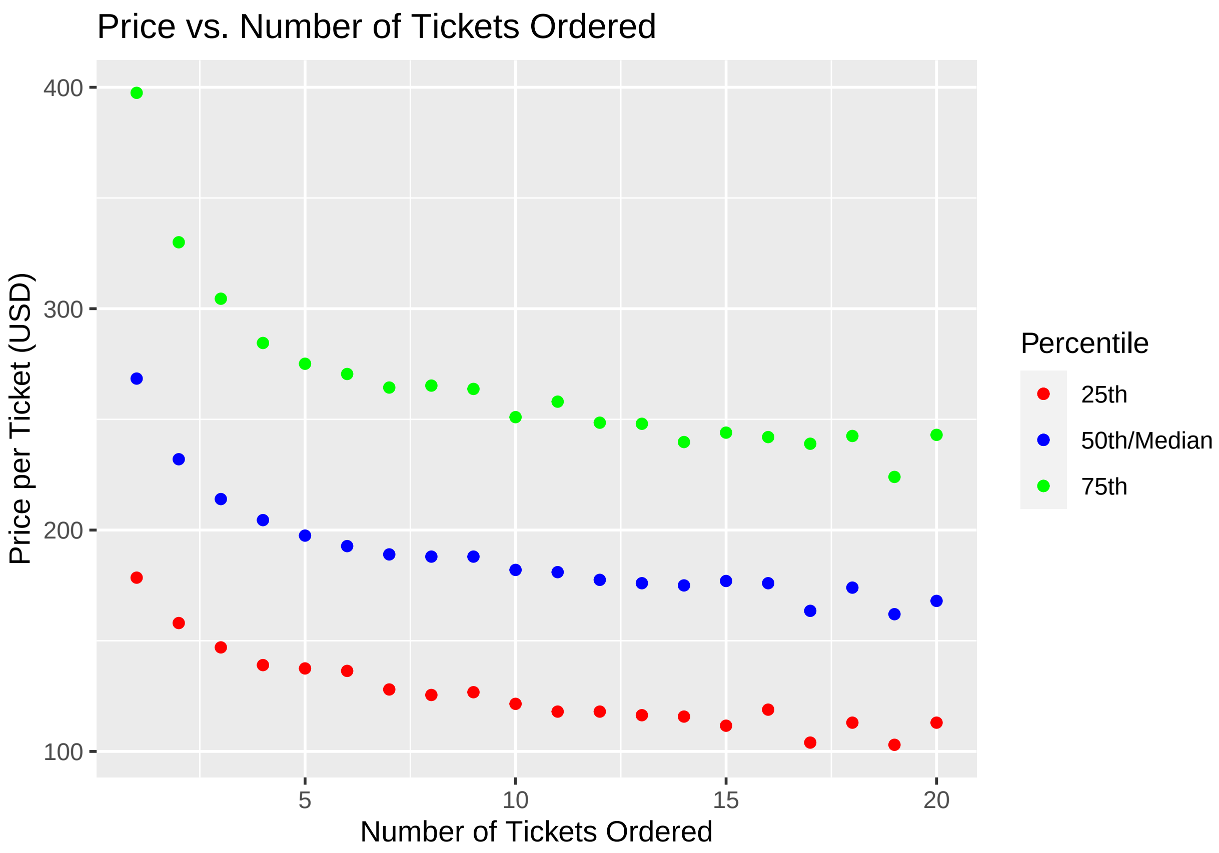


From this, we can see that the distribution is skewed left, with the mode number of tickets ordered being 1 and 2 tickets. It was not very often that an order of over 10 tickets was placed.

Let's further examine this variable by seeing how it correlates to `PricePerTicket`. We will do this by calculating the 25th, 50th, and 75th percentiles of `PricePerTicket` for each value of `NumTicketsOrdered`. We plotted it this way because `NumTicketsOrdered` is not a discrete variable. Therefore, if we created a normal scatterplot, all we would see is vertical lines of points. We wouldn't be able to make any conclusions about trends within the data.

```
top_10_LAX_percentiles <- top_10_LAX %>%
  group_by(NumTicketsOrdered) %>%
  mutate(p25 = quantile(PricePerTicket, .25),
         med = quantile(PricePerTicket, .5),
         p75 = quantile(PricePerTicket, .75)) %>%
  select(NumTicketsOrdered, p25, med, p75) %>%
  distinct()

ggplot(top_10_LAX_percentiles, aes(x = NumTicketsOrdered)) +
  geom_point(aes(y = p25, color = "25th")) +
  geom_point(aes(y = med, color = "50th/Median")) +
  geom_point(aes(y = p75, color = "75th")) +
  scale_color_manual(name = "Percentile",
                     values = c("25th" = "red", "50th/Median" = "blue", "75th" = "green")) +
  labs(title = "Price vs. Number of Tickets Ordered",
       x = "Number of Tickets Ordered",
       y = "Price per Ticket (USD)")
```



We see that there is a negative relationship between the two variables, but this relationship does not look linear; instead, the relationship is inversely proportional, meaning that the decreasing relationship will taper off as the number of tickets ordered increases.

After doing some research to understand the reasoning behind this relationship, we found that, according to Investopedia, (https://www.investopedia.com/articles/personal-finance/032416/cheapest-way-buy-two-or-more-airline-tickets.asp#:~:text=Key%20Takeaways,the%20lowest%2Dcost%20seats%20available.), "if you purchase multiple tickets in a single transaction, the price will be the same for each of the tickets." This would initially imply that our graph should be a horizontal line; however, there are other factors to consider that explain the relationship shown in our graph. For example, passengers who buy only one ticket are more likely to buy first class/business class flights with premium features than passengers who buy 10 tickets at once, who are likely big families or bigger groups who are already spending a lot of money and would therefore pick the more basic/economy class tickets. This theory explains why the two variables are inversely proportional with one another.

Data Modeling and Inference

ANOVA test

In the Exploratory Data Analysis section, we saw that the prices of flights to different destinations are significantly different. We wanted to confirm this using an ANOVA test.
We decided to explore the question of whether there is a statistically significant difference in the mean prices of flights going from LAX among the top 10 destinations. These was our Null Hypothesis: the true mean prices of flights to all destinations are equal. This was our Alternative Hypothesis: the true mean prices of flights to all destinations are not equal.

```
  my_data <- top_10_LAX %>%
  select(Dest, PricePerTicket)
  model <- aov(PricePerTicket ~ Dest, data = my_data)
  summary(model)
```

```
##                 Df    Sum Sq  Mean Sq F value Pr(>F)
## Dest             9 7.448e+08 82753248    3080 <2e-16 ***
## Residuals   159194 4.277e+09    26864
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA test resulted in an extremely low p-value of less than 2 x 10^(-16). We have sufficient statistical evidence to suggest that the mean prices of flights going from LAX to the top 10 destinations are different and that these differences in prices are not occurring due to randomness.

Chi-Square

Another hypothesis test we wanted to conduct was a chi-squared test to explore the question: is there a statistically significant difference between the observed distribution of the number of flights per quarter and a uniform distribution of flights per quarter across all 4 quarters? This was our Null Hypothesis: the number of flights taken in each quarter are uniformly distributed. This was our Alternative Hypothesis: the number of flights taken in each quarter are not uniformly distributed.

```
chisq.test(table(top_10_LAX$Quarter))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(top_10_LAX$Quarter)
## X-squared = 978.9, df = 3, p-value < 2.2e-16
```

The resulting p-value of less than 2.2 x (10^-16) gives us sufficient evidence to reject the null hypothesis. We have enough evidence to suggest that the number of flights taken in each quarter is not uniformly distributed.

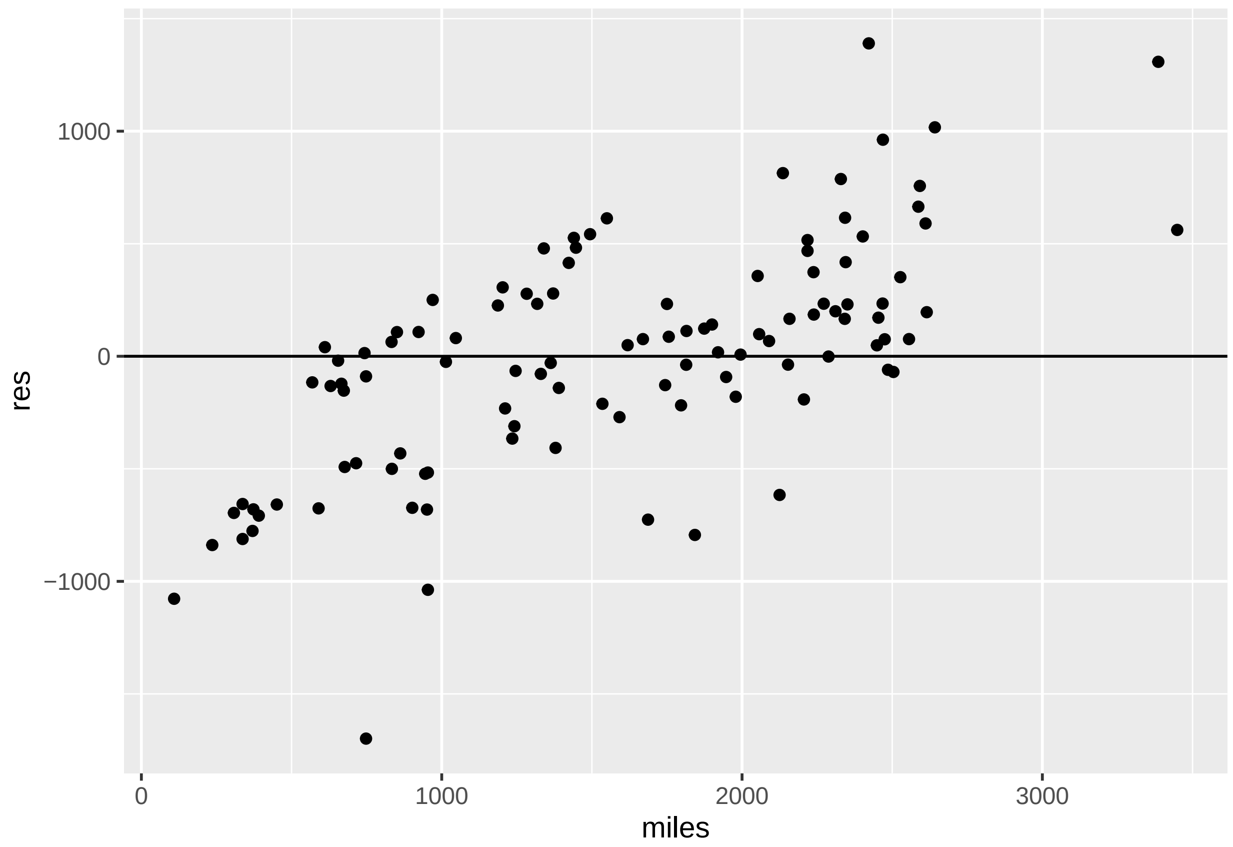
Linear Regression

In addition to comparing the prices of flights to different destinations, we wanted to model the paired data in various scatterplots from the Exploratory Data Analysis Section. Specifically, we used the `lm()` and `mgvc::gam()` linear modeling functions to test for association and strength. The first three examples also include the relevant residual plots to confirm whether a linear model is appropriate. (All graphs are have had outliers removed.)

Our first two paired examples are of number of miles against mean and median price, respectively.

```
miles_price_mean_model <- lm(Miles ~ mean, data = LAX_miles_price_mean)

mpm_residuals <- tibble(miles = LAX_miles_price_mean$Miles, res = resid(miles_price_mean_model))
ggplot(mpm_residuals, aes(x = miles, y = res)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



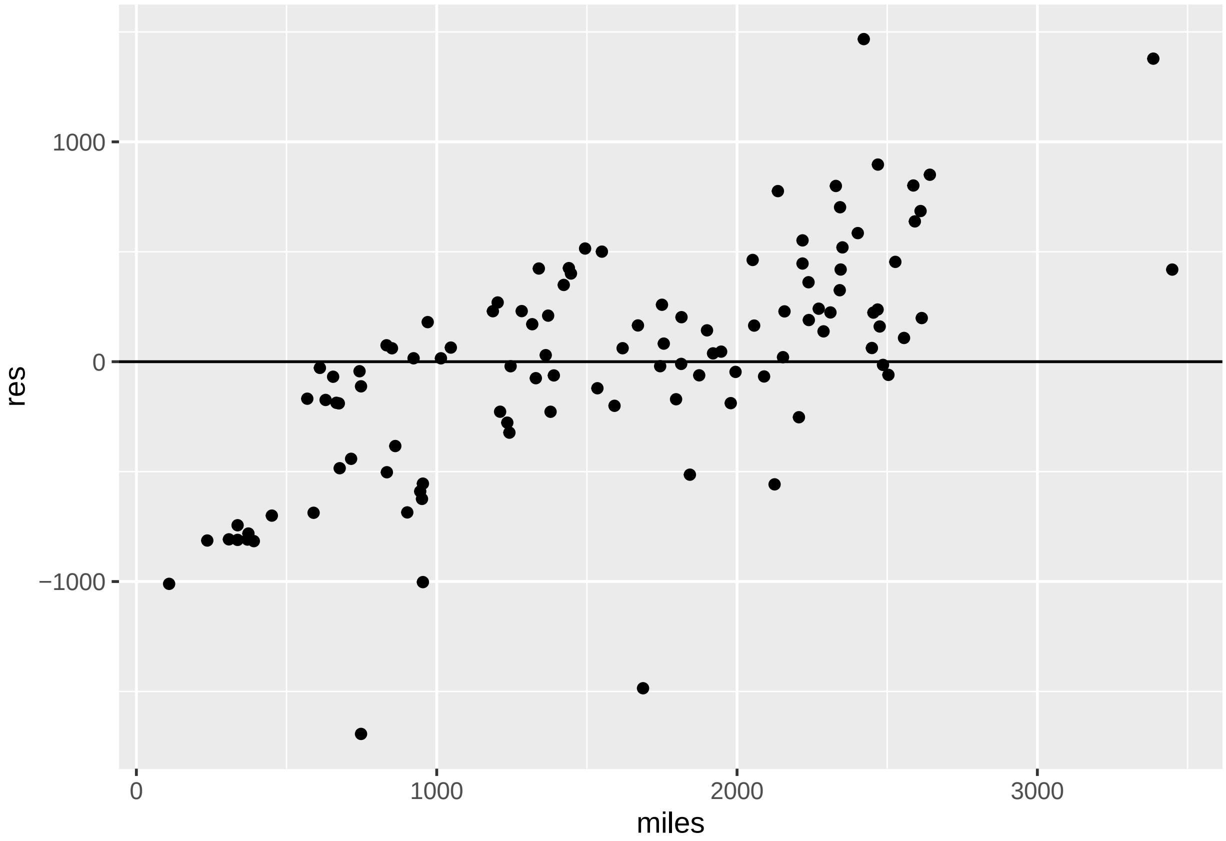
```
summary(miles_price_mean_model)
```

```
##
## Call:
## lm(formula = Miles ~ mean, data = LAX_miles_price_mean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1698.61  -231.77    48.68   250.42  1390.17
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 153.6709   132.5478   1.159    0.249
## mean          6.3098     0.5457  11.563   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 506 on 107 degrees of freedom
## Multiple R-squared:  0.5554, Adjusted R-squared:  0.5513
## F-statistic: 133.7 on 1 and 107 DF,  p-value: < 2.2e-16
```

Calling the `summary()` function on the relevant linear model, we see that the adjusted R-squared value is 0.7553 and p < 2e-16. This indicates statistically significant evidence of a correlation (our null hypothesis being that there is no association between the variables) which is moderately strong.

```
miles_price_med_model <- lm(Miles ~ median, data = LAX_miles_price_med)

mpmd_residuals <- tibble(miles = LAX_miles_price_med$Miles, res = resid(miles_price_med_model))
ggplot(mpmd_residuals, aes(x = miles, y = res)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



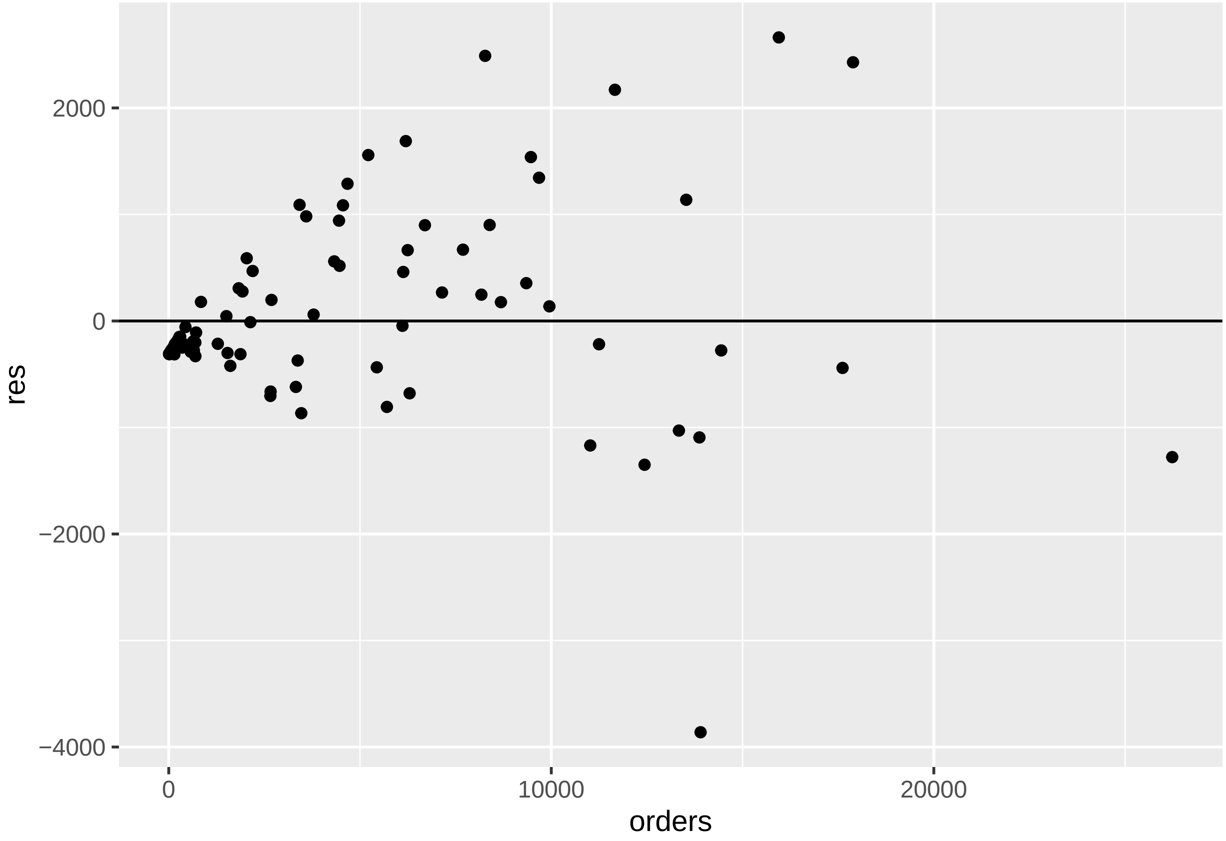
```
summary(miles_price_med_model)
```

```
##
## Call:
## lm(formula = Miles ~ median, data = LAX_miles_price_med)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1693.04  -227.46    37.85   258.96  1467.31
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 193.9923   136.1974   1.424    0.157
## median        7.0111     0.6409  10.939   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 521.5 on 107 degrees of freedom
## Multiple R-squared:  0.5279, Adjusted R-squared:  0.5235
## F-statistic: 119.7 on 1 and 107 DF,  p-value: < 2.2e-16
```

Again, in this example p < 2e-16 and our adjusted R-squared value is 0.7317, indicating a moderate to moderately strong correlation. We can verify that a linear model is appropriate in both of the past examples by observing the respective residual plots. Since there is no pattern in either plot, we have no issues.

```
orders_to_tickets_model <- lm(total_orders ~ total_tickets, data = LAX_orders_tickets)

ot_residuals <- tibble(orders = LAX_orders_tickets$total_orders, res = resid(orders_to_tickets_model))
ggplot(ot_residuals, aes(x = orders, y = res)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



```
summary(orders_to_tickets_model)
```

```
##
## Call:
## lm(formula = total_orders ~ total_tickets, data = LAX_orders_tickets)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3861.2  -301.4  -244.2   247.0  2662.0
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.160e+02  9.760e+01   3.238   0.0016 **
## total_tickets 3.637e-01  5.815e-03  62.545   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 835.5 on 107 degrees of freedom
## Multiple R-squared:  0.9734, Adjusted R-squared:  0.9731
## F-statistic:  3912 on 1 and 107 DF,  p-value: < 2.2e-16
```

The next example is of total orders against total tickets per destination. Similarly to the previous graphs, the `summary()` function returns a p-value of <2e-16, but this time adjusted R-squared is 0.9734, meaning that the correlation is very strong. A quick review of the residual plot confirms that a standard linear model is appropraite.

```
fit_p25 <- gam(NumTicketsOrdered ~ p25, data = top_10_LAX_percentiles)
fit_med <- gam(NumTicketsOrdered ~ med, data = top_10_LAX_percentiles)
fit_p75 <- gam(NumTicketsOrdered ~ p75, data = top_10_LAX_percentiles)

summary(fit_p25)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## NumTicketsOrdered ~ p25
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.35733    4.32598  10.716 3.05e-09 ***
## p25         -0.28349    0.03385  -8.374 1.27e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =  0.784   Deviance explained = 79.6%
## GCV = 8.3852  Scale est. = 7.5466    n = 20
```

```
summary(fit_med)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## NumTicketsOrdered ~ med
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 48.11914    5.61545   8.569 9.09e-08 ***
## med         -0.19872    0.02942  -6.755 2.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =  0.701   Deviance explained = 71.7%
## GCV = 11.611  Scale est. = 10.45     n = 20
```

```
summary(fit_p75)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## NumTicketsOrdered ~ p75
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.80907    5.62164   7.615 4.91e-07 ***
## p75         -0.12112    0.02086  -5.807 1.68e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =  0.633   Deviance explained = 65.2%
## GCV = 14.287  Scale est. = 12.858    n = 20
```

Lastly, we wanted to model our graph displaying price in quartiles vs. number of tickets ordered. However, a clear a consistent curve in the points across the 25th, 50th, and 75th percentile (all included on the original graph) suggested that a linear model may not fit the data right. We instead used the `mgcv::gam()` model, which `ggplot::geom_smooth()` defaulted to. We also treated the dots representing the 25th, 50th, and 75th percentiles as independent when running the `summary()` function, creating identical individual models for each one.

All p-values were again less than 2e-16, meaning that there is statistically significant evidence of an association. Our adjusted R-squared values were 0.777, 0.707, and 0.667 for p25, med, and p75, respectively, indicating moderate to moderately strong associations. Looking at the graph, it makes sense that the R-squared values consistently decrease from p25 onwards since the pattern that all the models follow is more extreme at p75 and less extreme at p25.

Conclusion

After researching, executing our exploratory data analysis, and performing data modeling and tests, we found that the price of an airline ticket, when leaving from LAX, is influenced by several variables. For example, we found that the price per ticket is directly correlated with the number of miles traveled, and we found that this relationship can be well assessed with a linear model. Additionally, we discovered an inverse relationship between the number of tickets ordered and the price per ticket, which was intriguing. As the number of tickets ordered increased, the price per ticket decreased, but it decreased at a decreasing rate. We found that, although the `Quarter` variable did not impact the price much, there was a statistically significant deviation from a uniform distribution in the number of orders made per quarter. We found relationships between the price per ticket and the destination airport. Specifically, we discovered interesting results when looking at 3 hub airports. The airline which had a hub at that airport had significantly higher prices. Additionally, we found that for MCO (Orlando International Airport) and BOS (Boston Logan International Airport) the price per ticket may be influenced by the number of airlines operating at that destination. Namely, airports with more competition generally had lower prices.

However, we are aware that there are *many* more variables impacting the price of an airline ticket. For example, the time of day of the flight, the week/month of the flight, the oil and gasoline market at the time of the flight, employment rates, how far in advance the order was booked, and many more. These variables were not present in our dataset, so we could not examine them. It would be interesting to look further into this topic and discover how these attributes might affect the price of an airplane ticket.