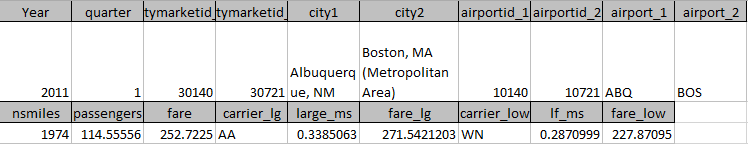
**Technical Summary**

**Purpose**

The purpose of our project is to analyze the factors that affect air fare ticket prices and examine how seasonality and region play a role in pricing.

**Data Set**

The fare price data set from 1996 to 2014 was downloaded from [www.transportation.gov](http://www.transportation.gov). There are 19 columns and 160,000 rows in the original data set.



**Regression Model**

**Model 1:**

Dependent variable: the lowest fare

Independent variables: miles, passengers, carrier(with the lowest fare price) market share.

R^2=0.2114

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.271e+02 4.601e-01 276.30 <2e-16 \*\*\*

nsmiles 4.394e-02 2.390e-04 183.84 <2e-16 \*\*\*

passengers -2.341e-02 3.382e-04 -69.22 <2e-16 \*\*\*

lf\_ms 2.793e+01 5.148e-01 54.25 <2e-16 \*\*\*

**Model 2:**

We want to introduce fuel price to this model. we download plane fuel price from <http://www.indexmundi.com/commodities/?commodity=jet-fuel>, and use average price for each quarter.

Dependent variable: the lowest fare

Independent variables: miles, passengers, carrier(with the lowest fare price) market share, fuel price

R^2=0.2446

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.105e+02 4.924e-01 224.32 <2e-16 \*\*\*

nsmiles 4.336e-02 2.340e-04 185.28 <2e-16 \*\*\*

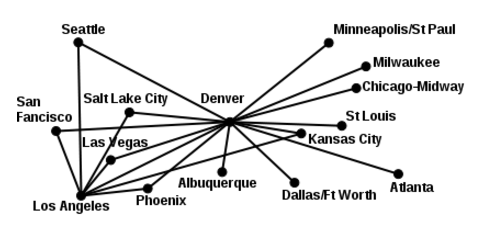
passengers -2.482e-02 3.314e-04 -74.90 <2e-16 \*\*\*

lf\_ms 2.666e+01 5.040e-01 52.90 <2e-16 \*\*\*

fuel\_price 1.311e+01 1.570e-01 83.56 <2e-16 \*\*\*

**Model 3:**

We want to introduce region to this model. In our original data set, we have airport, city and market could be used as region. After some research, we decided to use “hub” as our dummy variable. Hubs are those cities which are used by those airlines as their transit city. For example, as the graph shows, Denver and Los Angeles are used as hubs. We downloaded a hub list with 33 airport hub, then combined it with our original data set. For depart or destination airports, if one of both belongs to hub, we identify it as 1. If not, we identify it as 0.



Dependent variable: the lowest fare

Independent variables: miles, passengers, carrier (with the lowest fare price) market share, fuel price, if\_hub

R^2=0.245

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.129e+02 5.688e-01 198.436 <2e-16 \*\*\*

nsmiles 4.331e-02 2.341e-04 185.026 <2e-16 \*\*\*

passengers -2.409e-02 3.423e-04 -70.395 <2e-16 \*\*\*

lf\_ms 2.600e+01 5.100e-01 50.985 <2e-16 \*\*\*

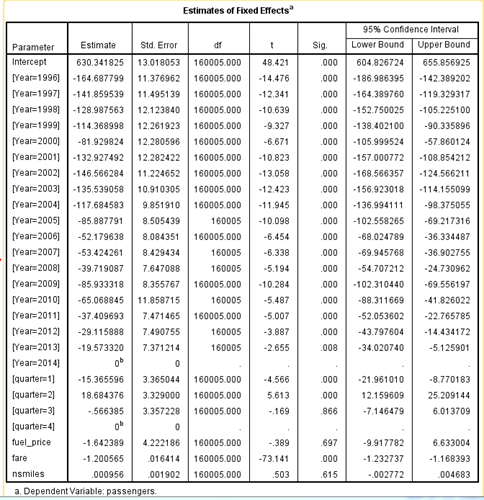
fuel\_price 1.311e+01 1.569e-01 83.575 <2e-16 \*\*\*

if\_hub -3.147e+00 3.714e-01 -8.472 <2e-16 \*\*\*

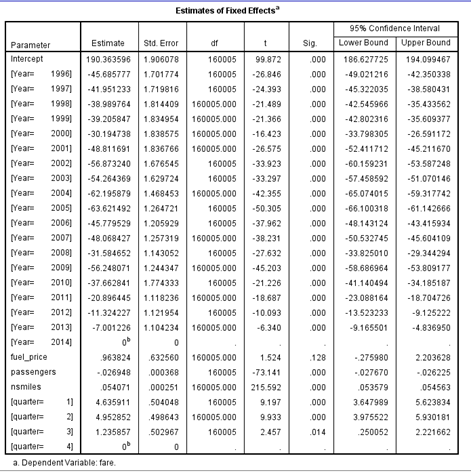
**Determining Busy Season:**

Panel data analysis conducted in SPSS shows that Quarter 2 (April, May, and June) is the busiest time for airlines with the highest number of passengers flying. We attribute this high number of passengers to summer break travel. Quarter 4 (Oct-Dec) had the second highest number of passengers flying which we attributed to winter holiday travel (Output 1). The average number of passengers flying per route (two cities), per day in our model was 630. This shows that the average number of passengers per route has increased by about 164 people since 1996. Output 2 shows the results of a panel data analysis of fare which showed that Quarter 4 had the lowest airfare and Quarter 2 had the highest airfare (increase of only $4.95 from Quarter 4’s price).

Output 1:



Output 2:



**Final Model**

After using SPSS to help us select quarter 2 and 4 as the busiest seasons for air travel, we were able to identify busy as 1 and not busy as 0 in our final model.

Dependent variable: the lowest fare

Independent variables: miles, passengers, carrier (with the lowest fare price) market share, fuel price, if\_hub, busy\_season

R^2=0.2454

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.117e+02 5.935e-01 188.215 < 2e-16 \*\*\*

nsmiles 4.331e-02 2.340e-04 185.041 < 2e-16 \*\*\*

passengers -2.415e-02 3.423e-04 -70.547 < 2e-16 \*\*\*

lf\_ms 2.600e+01 5.099e-01 50.997 < 2e-16 \*\*\*

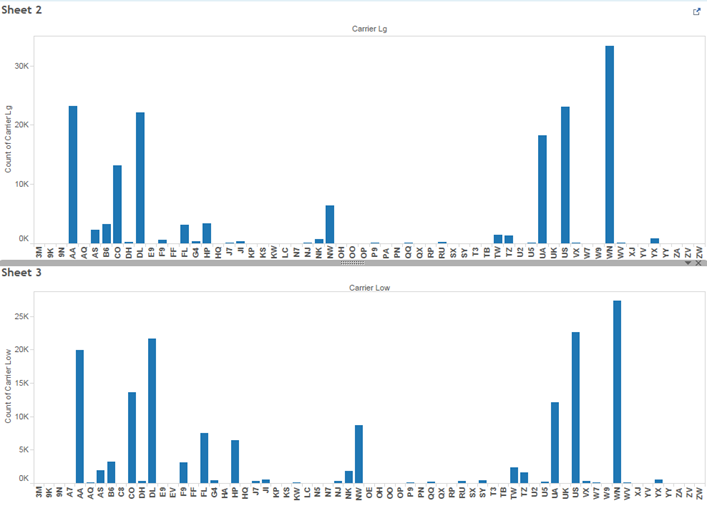
fuel\_price 1.317e+01 1.571e-01 83.839 < 2e-16 \*\*\*

if\_hub -3.130e+00 3.713e-01 -8.429 < 2e-16 \*\*\*

busy\_season 2.168e+00 3.168e-01 6.844 7.71e-12 \*\*\*

This model shows that longer distance will make price higher. More passengers on a flight results in a lower price. We think the drop in price is due to the fact that the plane will fly with empty seats, so that the carrier can give a lower price per customer. When fuel price is high, the price will be high. If the airport is a hub, the price will be low, because customers have a lot of choices. During a busy season, the price will be higher.

For market share we found something very interesting. This model shows more market share will lead to lower price. When we aggregate the frequency of carriers in lowest price and highest price, we found the distribution is similar. Our findings suggest that carriers have different price strategies. For example, the highest one is Southwest and that carrier likes posting out the highest price ticket to compensate for also posting the lowest price tickets.



**Caveat**

The R^2 getting improved as we bring fuel price, hub and time series. All variables are significant, which means they are all playing very important role in fare price. But the R^2 is low. Because fare pricing is very complicated and we cannot just use a linear regression with limited variables to do the prediction. But we can use our conclusion for further analysis about time series and region.

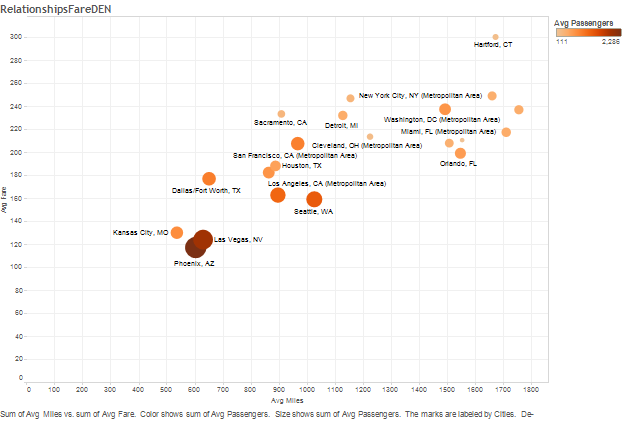
**Visualizations**

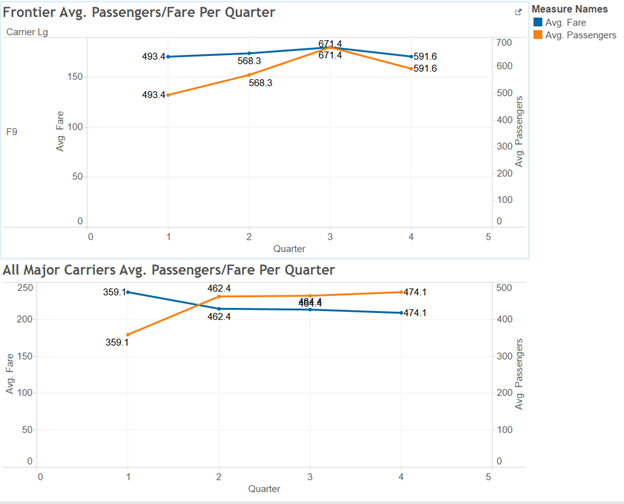
As a final step we used Tableau to more closely examine the airfare associated with routes to and from Colorado. Table1a with data from 1996-2014 was used for the analysis. We subset the data into all records that relate to Colorado.

Our first objective was to see how the number of passengers and miles could affect air fare.

We can see that the further the destination city it is, the higher the airfare it gets and the fewer passengers they have.

Further analysis could focus on region/specific city.





Average fare and passengers comparison between frontier and all major carriers by quarter

1. F9 has lower average fare compare to all major carriers’ average fare
2. However, F9 has higher passengers than all major carriers’ (A. when calculating average passengers for all major carriers, there might some outliers B. lower fare attracts more passengers)
3. Overall average number of passengers traveling from/to Denver is the lowest in Quarter 1. We believe that because Quarter 1 is Jan-Mar, it is immediately after the winter break when people go back to work after the holidays and travel less.
4. For Frontier, average passengers and average fare goes up and down together
5. In Quarter 3 (July-Sep) there is the highest fare and passengers flow which may be due to summer break, especially in July. It would be helpful to have had the data segmented by month in our data, but we can only speculate based on quarters.
6. Our visualizations indicated that, in general, airfare and the average number of passengers move in opposite directions over time. As the number of passengers increases, the average airfare decreases.