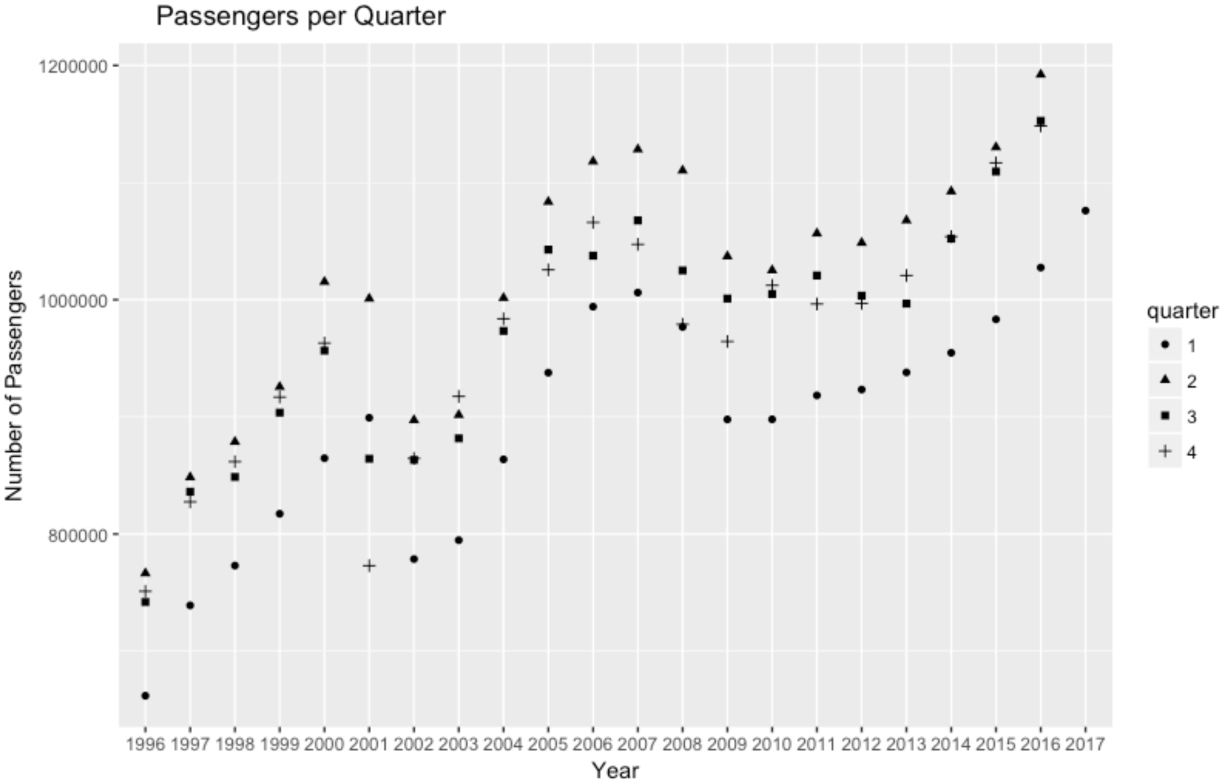
Matthew Heath

Homework 2

2. The data covers 1996 through the first quarter of 2017. There is missing data for every quarter and year, but there are no quarters or years that have no data. In table 1a, 1996 and 1997 have the most missing data, with hundreds of flights missing either 3 or 6 values. 1996 and 1997 also have the most missing values in table 6, with hundreds of flights missing 7 or 10 values. In table 1a one can see that as time goes on there are less missing values, although this association is not very strong. Also, every flight in table 6 has at least 4 missing values in each year. This is because table 6 omits specific airport details, which eliminates the variables airport1, airport2, airport_id1, and airport_id2, thus causing "missing" data. There is no apparent pattern across quarters regarding missing data.

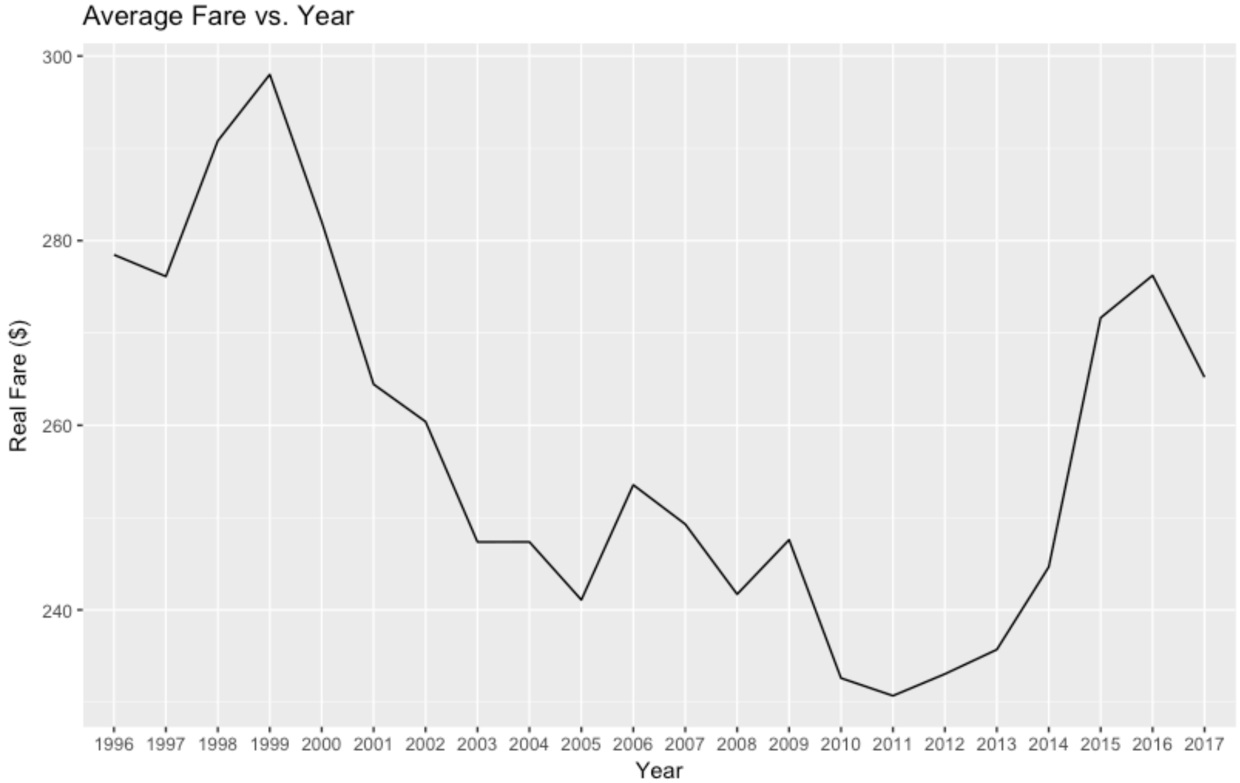
3. In 2017, Los Angeles (202), Dallas/ Fort Worth (192), Washington D.C. (184), and Phoenix (183), have the most connections, with the number of connections represented within the parentheses. I calculated this by combining two tables, one which represented the city from which the flight departed and one which represented the city from which the flight left. Besides the many cities with zero connections in 2017, there are 30 cities with only 1 connection. Among these cities are Waco, TX; Santa Maria, CA; Joplin, MO; and Ogden, UT. Since we only have data from the first quarter of 2017, I will only compare this to the first quarter data from 2007 and 1997 instead of the total data from each year. In 2007, Los Angeles (200), Washington D.C. (191), Las Vegas (183), and New York City (177) had the most connections. Apart from the cities with zero connections, there were 29 cities with only one connection in 2007, many of which are different from the cities with one connection in 2017. For example, in 2007 there was only one connection for cities such as Oxnard, CA; Topeka, KS; and Stockton, CA. In 1997, Los Angeles (203), Washington D.C. (187), New York City (186), and Chicago (180) had the most connections. Once again, apart from the sizable amount of cities with zero connections, there were 40 cities with only one connection, such as Santa Fe, NM; Modesto, CA; Carlsbad, CA; and Athens, GA. In order to see which cities have increased connectivity the most, I separately subtracted the 1997 and 2007 tables from the 2017 table. I found that between 2007 and 2017, Sanford, FL (43), Punta Gorda, FL (30), and Austin, TX (27) gained the most connections, with the number of connections gained in the parentheses. It seems that the southeastern part of the United States, and Florida in particular, gained the most connections. This pattern is also mostly true when looking at the differences between 2017 and 1997, as the previous three cities were among the top four regarding the most gained connections, although Salt Lake City was now among the top 3. (1)

4.



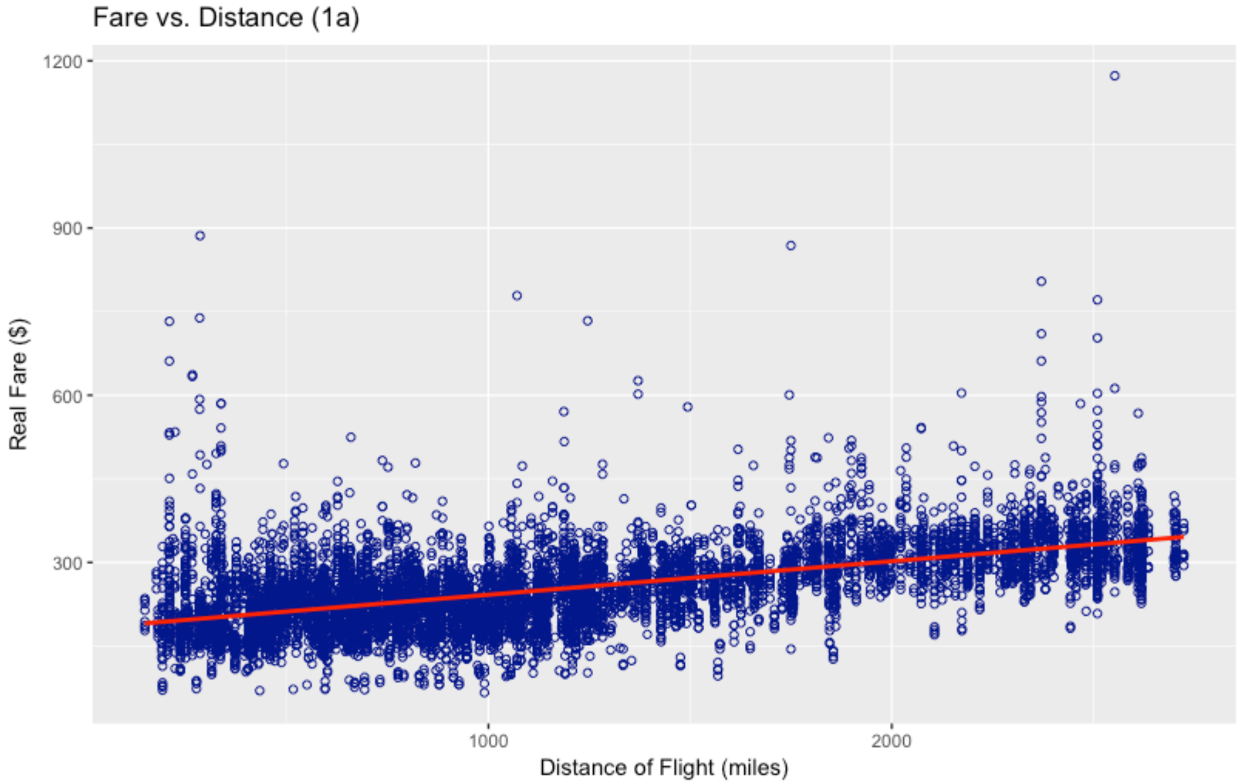
Generally speaking, the total number of passengers per quarter has increased through the years. I determined this by looking at the mean total number of passengers per day for each quarter and comparing between the years. One interesting trend is that the number of passengers greatly decreased in the fourth quarter of 2001 and took a few years to get back to the regular growth rate, which is likely due to fear after the September 11 terrorist attacks. Another dip can be seen around 2008, which can easily be explained by the beginning of the recession, which would prevent people from spending money on flights. I noticed that in nearly every year, the number of passengers in the first quarter is much smaller than the number of passengers for every other quarter. This is likely because it is immediately after the holiday season and most people are getting back to work. Since the first quarter is composed of January, February, and March, it happens to fall directly between Christmas breaks and Spring breaks for school, which are two of the likeliest times for families to make large vacations, many of which require flights. (2) (3)

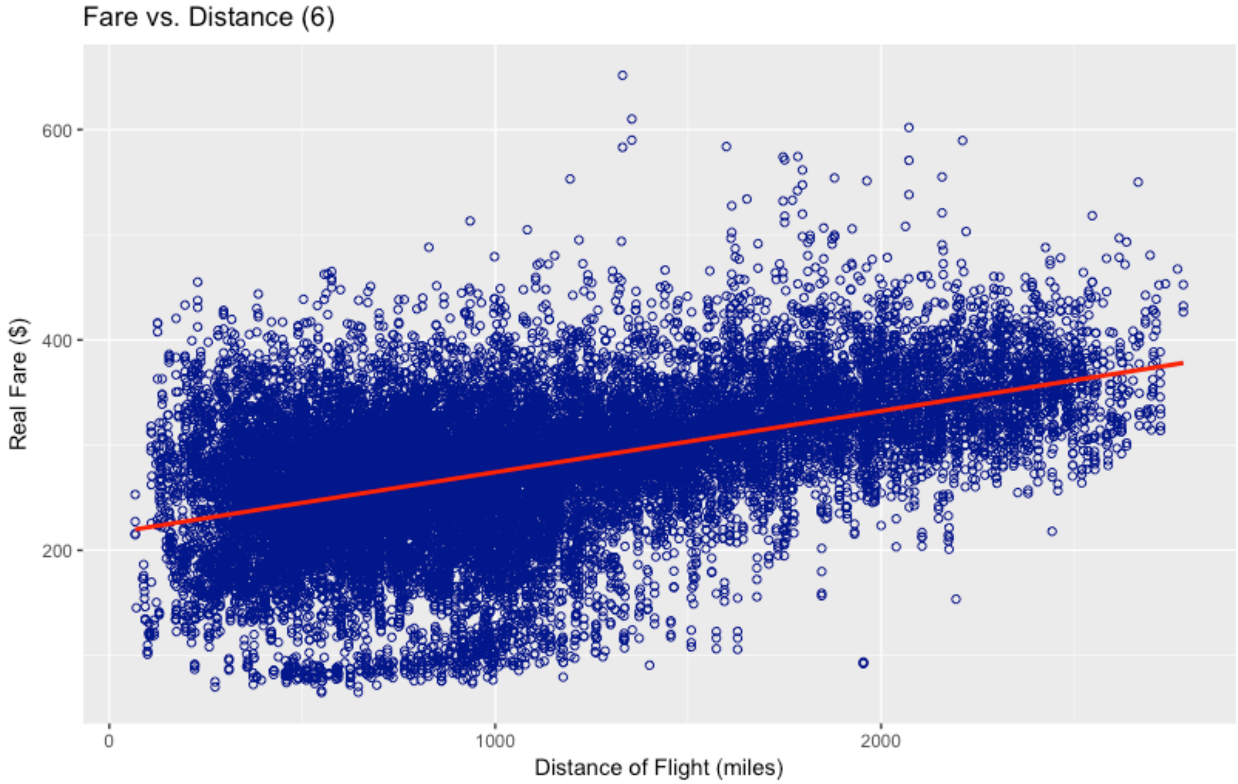
6.

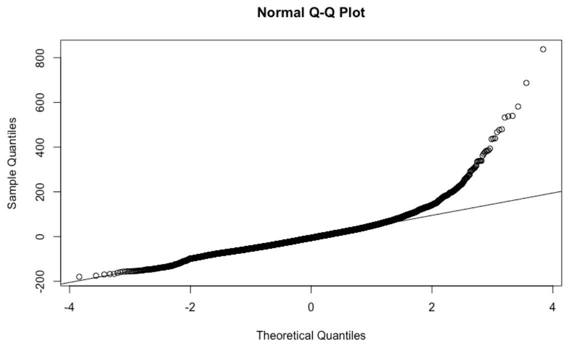


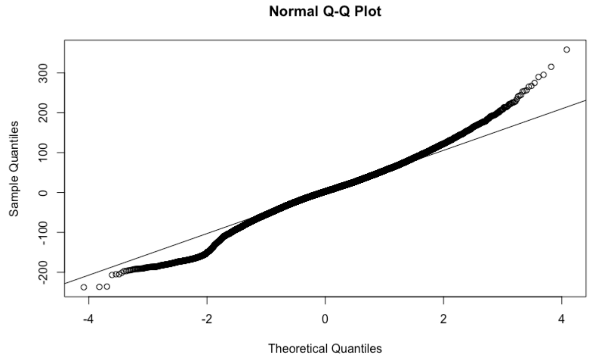
In order to see how airfares have changed over time, I first converted the nominal fares into real fares, which allows me to adjust for inflation. To calculate this I used the formula given, in which we divided the quarter 1 cpi of the base year by that of the year we are calculating, then multiplying the result by the average wages. For the cpi of the year we are calculating, I used the cpi for March for each year, as this is the final cpi at the end of the first quarter. A plot of these averages shows us that airfares have tended to decrease over time, with fares bottoming out in 2010 at around $235 per flight. However, since 2010 fares have been increasing again. Although there may be many factors causing this trend, I believe that the most major cause is likely the state of the economy. The lowest fares occur when the US is about to enter or in the middle of the recession, and as the recession ends the fares begin to rise again. This makes sense, as there will be less demand for flights during a poor economy, and therefore cheaper prices.

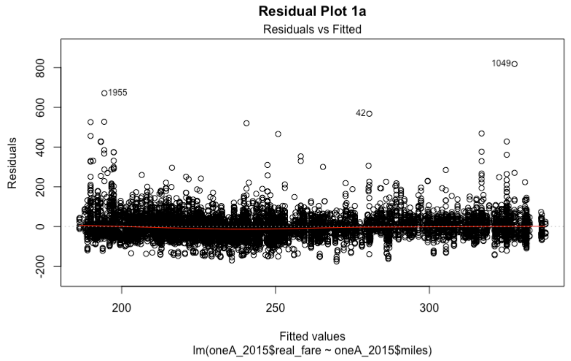
7.

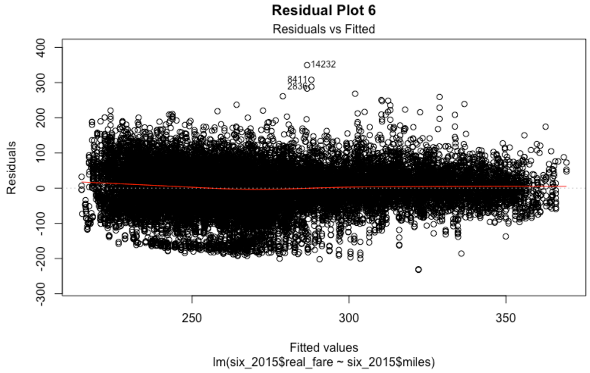

Fare vs. Distance (1a)
1200
900
600
300
Real Fare ($)
1000
2000
Distance of Flight (miles)


Fare vs. Distance (6)
600
400
200
Real Fare ($)
0
1000
2000
Distance of Flight (miles)



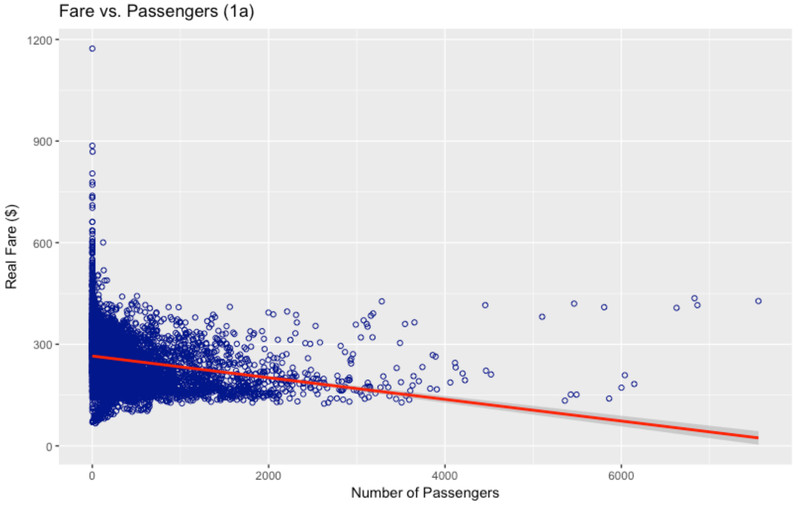


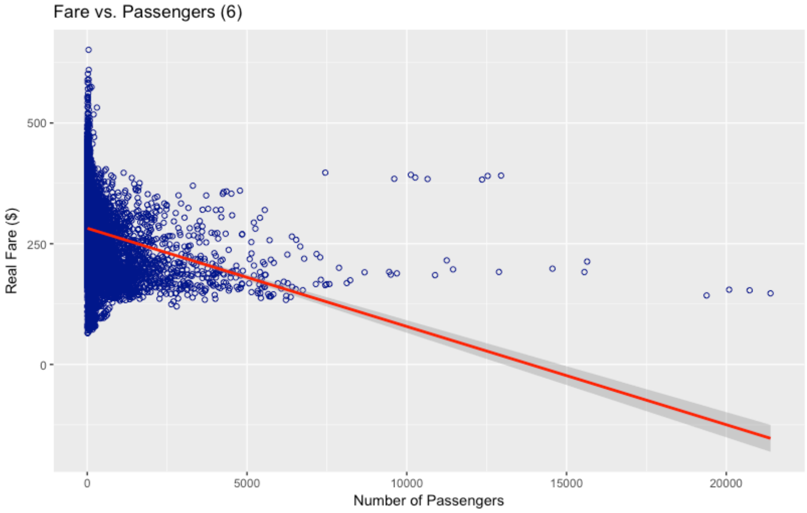


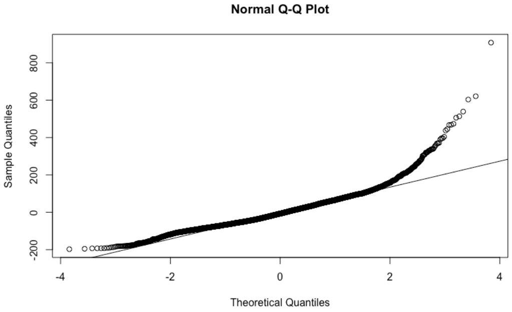


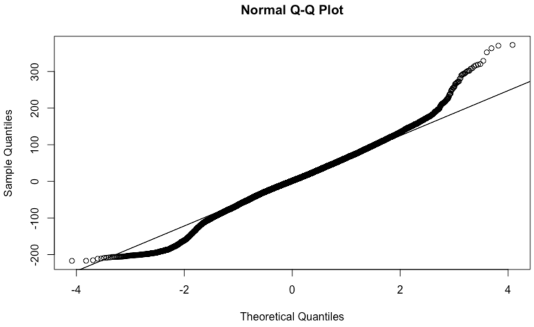
In order to determine the relationship between fare and distance, I ran a linear regression on the two variables in both tables 1a and 6. For table 1a, I got the model: y= 177.6043+0.059x, where y= fare and x=distance in miles. This infers that there is a positive linear relationship between the two variable, where the fare increases as the distance increases. This is supported by the scatter plot. For table 6, I got the model: y= 210.9846+0.057x, where y= fare and x= distance in miles. This is very similar to the model for table 1a, as this also shows that the fare increases as the distance increases. However, the amount that the fare increases per each additional mile is lower in table 6 then it is in table 1a, and the intercept for 1a is much smaller than the intercept for 6. As with table 1a, there is no disagreement between the model and the graph for table 6, as both show a positive linear relationship. As with all linear regressions, I had to make a few assumptions. These assumptions are that the relationship between the two variables is linear, the errors of the data are normally distributed and independent (no autocorrelation), and that the residuals have equal variances (homoscedasticity). One assumption that may not be met is the normality of residuals. By looking at the normal probability plot of the residuals, the data appears to be skewed right, with a particularly strong skew in table 1a. This prevents us from being able to reliably use the t-test and p-values. There is no pattern in the residual plots, so one can safely determine that the data is homoscedastic and linearly associated. I also noticed that there are some outliers that may be skewing our model. Finally, there is no reason to believe that the residuals are not independent, as there should be no correlation based on time or other associated variables.

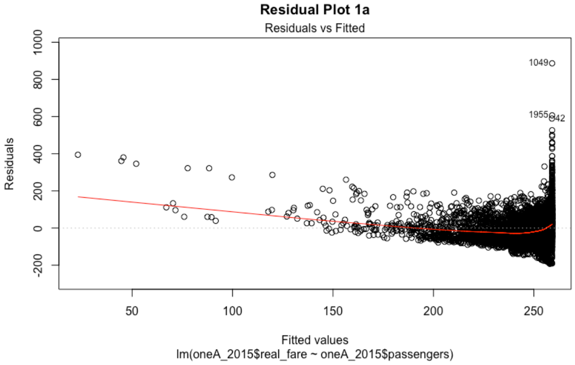
8.

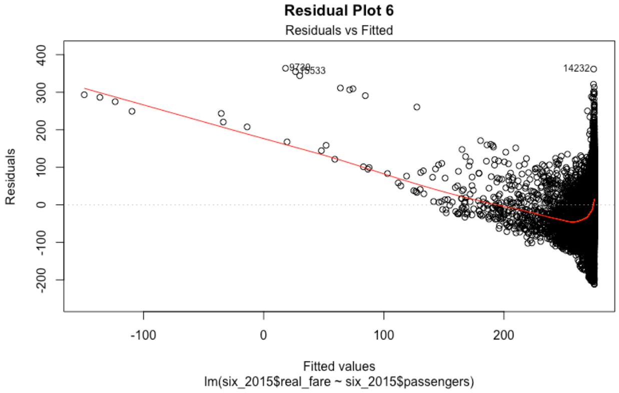

Fare vs. Passengers (1a)
Real Fare ($)
1200
900
600
300
0
0
2000
4000
6000
Number of Passengers


Fare vs. Passengers (6)
Real Fare ($)
500
250
0
0
5000
10000
15000
20000
Number of Passengers


Normal Q-Q Plot
Sample Quantiles
800
600
400
200
0
-200
-4
-2
0
2
4
Theoretical Quantiles

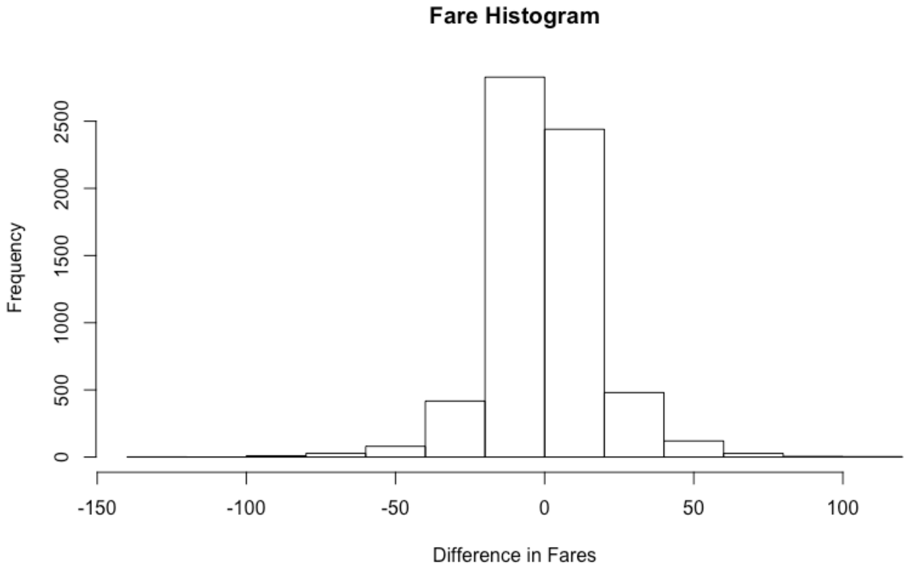

Normal Q-Q Plot
Sample Quantiles
300
200
100
0
-100
-200
-4
-2
0
2
4
Theoretical Quantiles

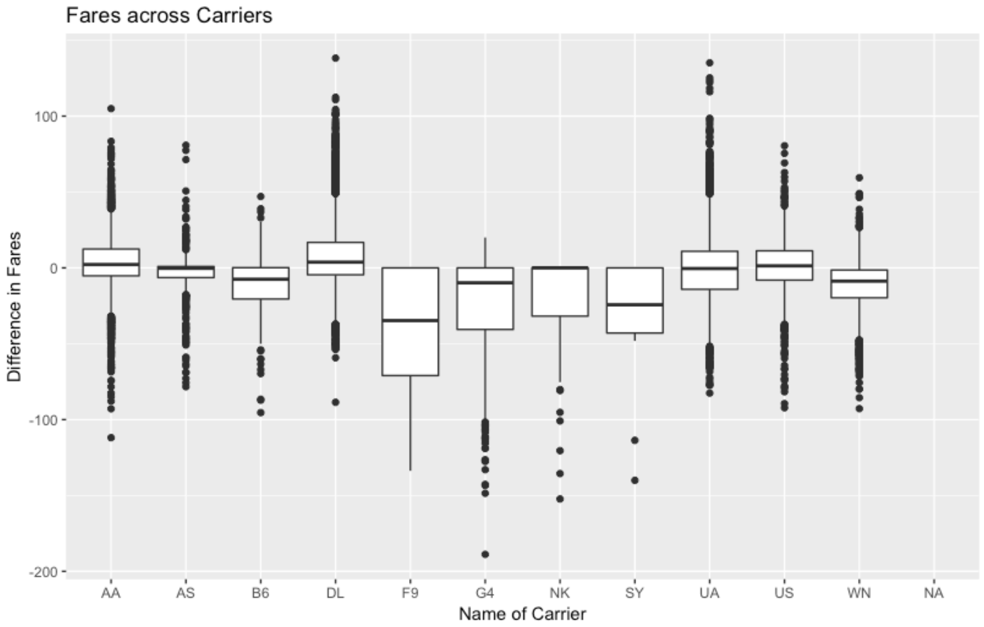




In order to determine the relationship between the mean number of passengers and airfare, I ran another linear regression. For table 1a I got the model: y= 259.09-0.031x, where y= airfare and x= # of passengers. For table 6 I got the model y= 275.484-0.019x. Although these models appear to indicate that there is a negative relationship, by looking at the graph one can see that there is no apparent relationship at all. This is also proven by looking at a summary of the regressions, which both have very low R- squared values, indicating weak or no linear association. Even though it is apparent that there is no linear association, I still checked the other assumptions and found the same violation of residual normality that appeared in the relationship between airfare and distance. When checking the residual plot one can find further evidence that the data is not linearly associated, as the residuals do not fall along the line y=0. Furthermore, there is a clear pattern for the residuals, as they fan out as x increases. This displays a lack of constant error variance, or heteroscedasticity. There is no major difference in the results between tables 1a and 6.

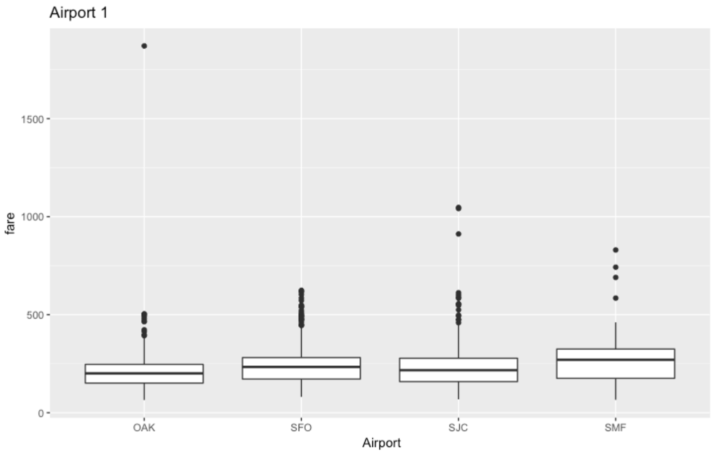
9.

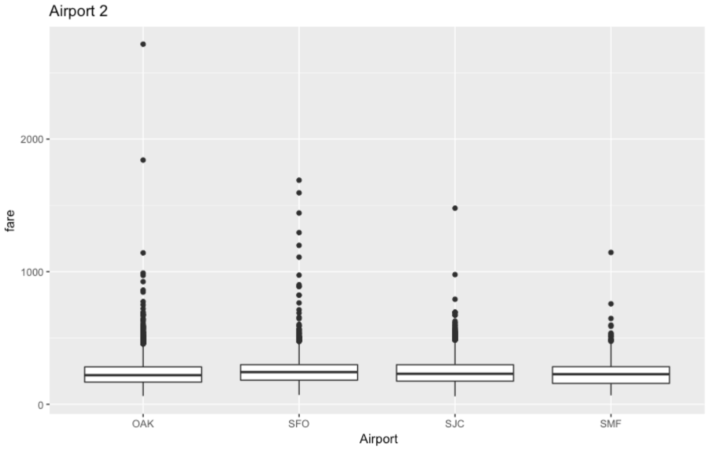


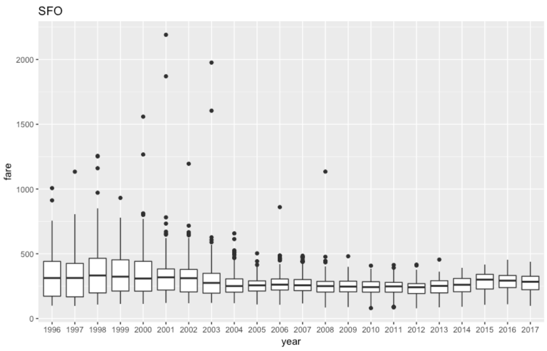


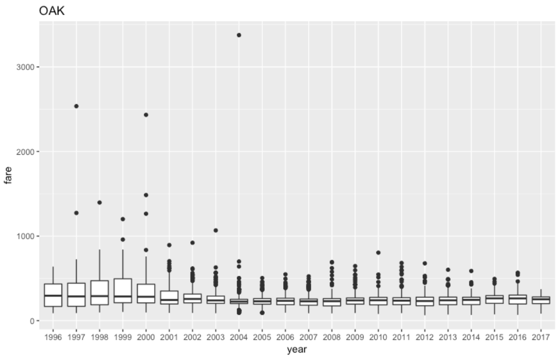
In order to find city pairs in which the carrier with the largest market share had fares below the average, I subtracted the average fares from the fares of the largest carrier, then calculated the average of this difference for each city. If the final answer was negative, then the largest carrier for that specific pair of cities had fare below the average. After performing these calculations I found many pairs whose largest carrier had fares below the average. In fact, a little more than half of all pairs of cities had a difference below zero, as shown in the histogram. I then investigated trends in these differences using box plots. I found little association between the difference in fares and most of the other variables, however I did find that the carrier itself played a role. By looking at the boxplot, one can see that most of the largest carriers have median fares that are pretty close to average. However, the carriers “F9”, “G4”, “NK”, and “SY” all have medians below zero and/or are heavily skewed. This indicates that these four carriers routinely have fares lower than average, and therefore cities in which these carriers hold the largest share will have the fare of their largest carrier lower than average. I tried regressing the differences against various other measures but did not find any strong linear associations, as the only strong associations involved measures that were included in the calculation to find the difference in fares.

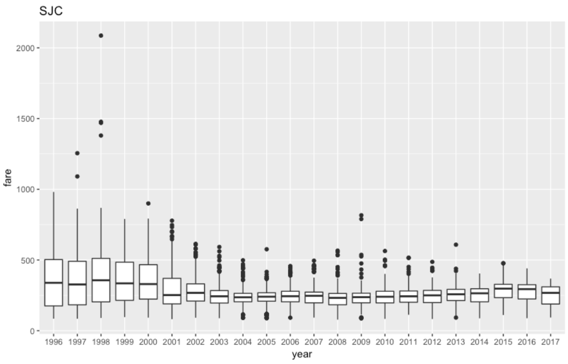
10.

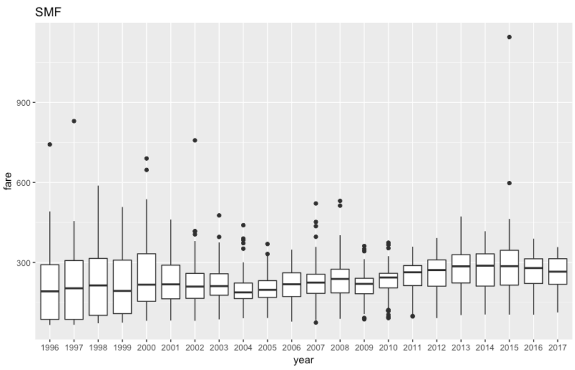

Airport 1


Airport 2


SFO


OAK


SJC


SMF

Overall, the fares between each of the four airports are very similar, although SFO tends to have slightly higher fares. This makes sense, as all four airports are very close to each other and therefore are competing with each other. In order to stay competitive, the fares must be close, as customers are likely to choose the cheapest option. The results by year tend to loosely follow the pattern that was observed for passengers by year. In years in which there were fewer passengers, the fares tend to be a bit cheaper. This makes sense, as the reduced demand leads to reduced prices, as suppliers want to entice the customers to buy more flights via cheaper prices. This trend is true across all four airports. In order to calculate the airport with the most long distance flights, I first had to figure out what made a flight "long distance." I determined that 2500 miles would be a good cutoff, then I sorted each flight for each airport into two categories: long flight and other. SFO had the most long flights, with 1335. This is pretty close to the 1317 long flights from SJC, but substantially more than the 1291 from OAK and 836 from SMF. However, there are many more total flights out of SFO than there are from SMF, which makes the comparison between the two airports somewhat unfair. These results do not significantly differ by year.

(1) https://stackoverflow.com/questions/12897220/how-to-merge-tables-in-r

(2) https://www.rdocumentation.org/packages/base/versions/3.4.1/topics/split

(3) http://www.sthda.com/english/wiki/ggplot2-axis-ticks-a-guide-to-customize-tick-marks-and-labels