Mitchell Layton

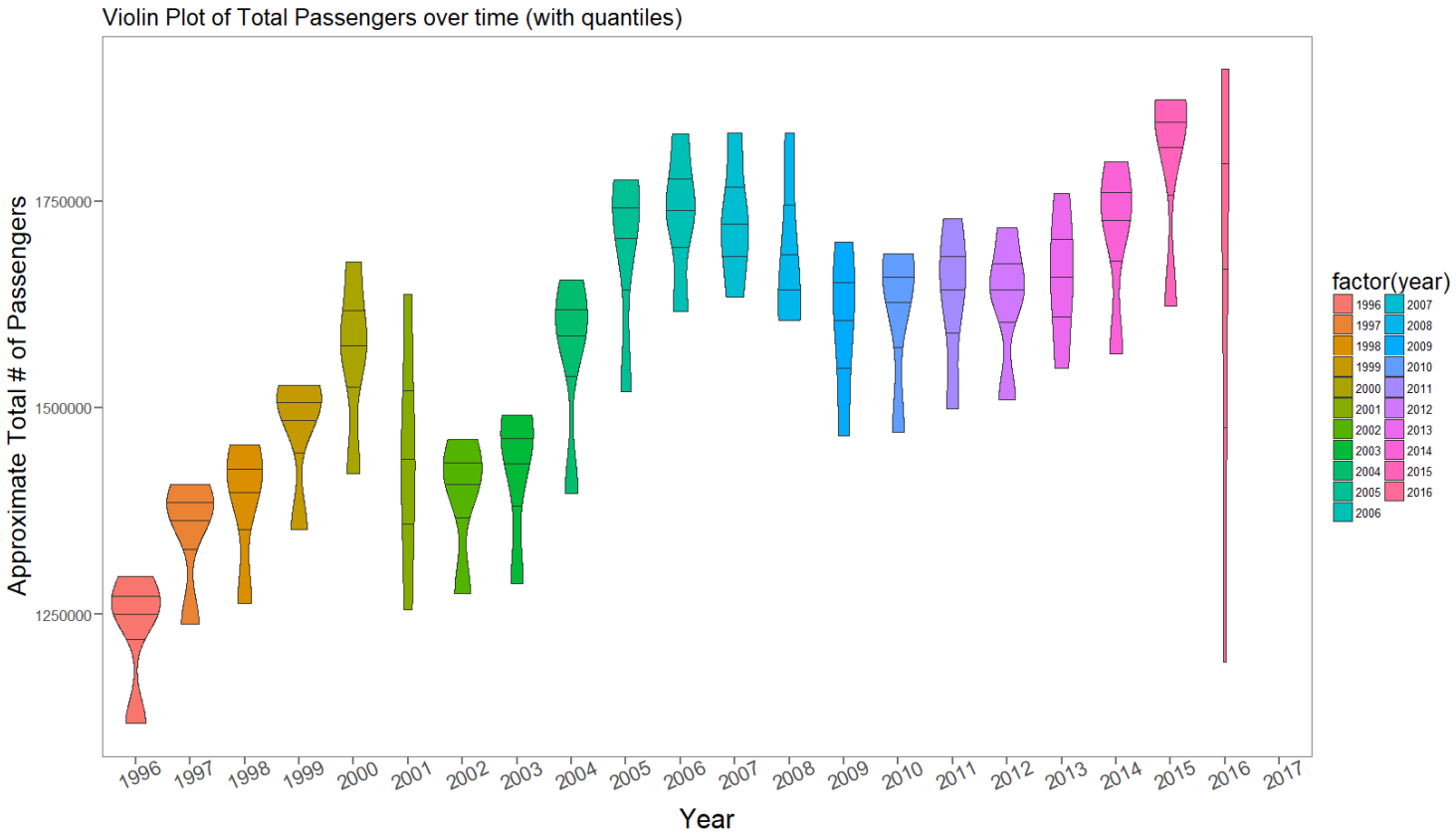
912307956

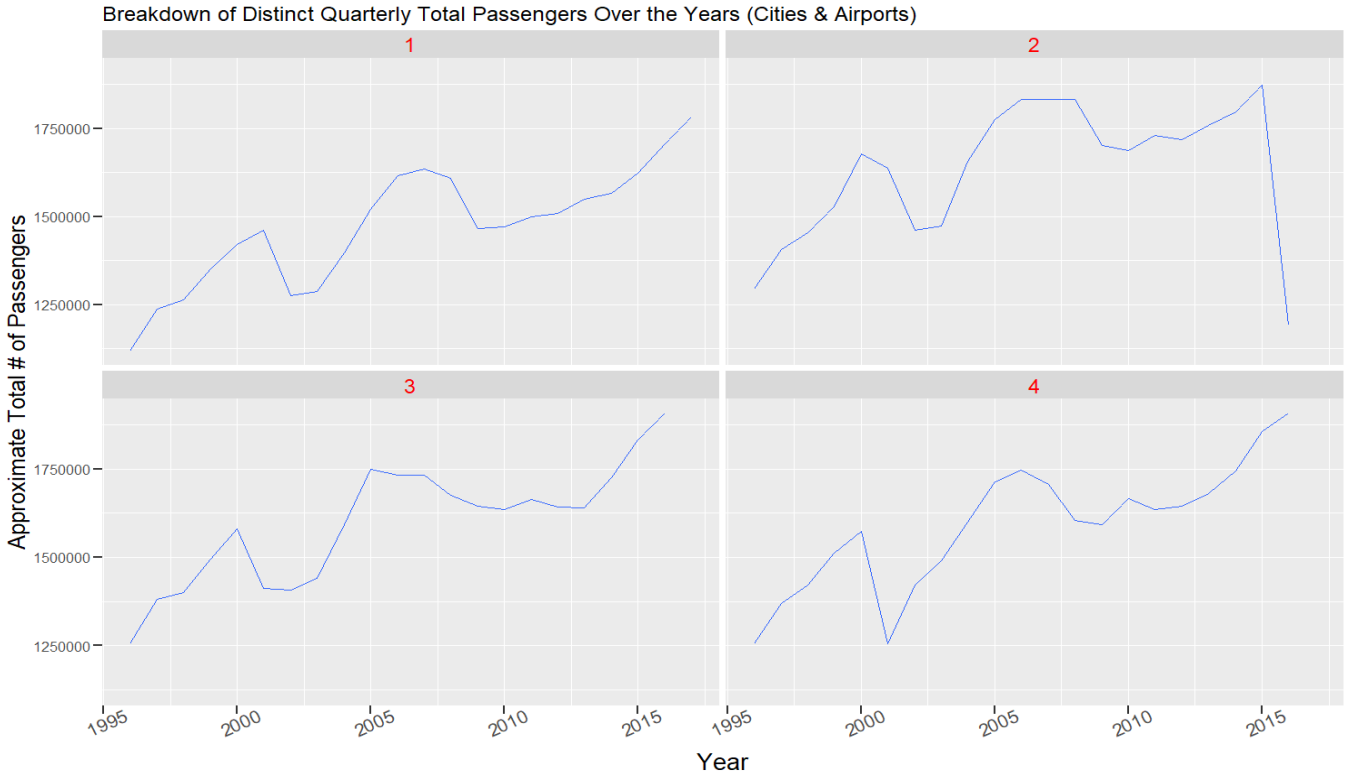
STA 141A Homework Assignment #2 – Due 10/31 @ 5:00 PM

1) (code in appendix)
2) The timespan that this data set covers is from quarter 1 in 1996, to quarter 4 in 2017. The data has certain characteristics such as when you check for NA's based on table 1a and 6 you find that there are 460,682 missing values in both the airport1 and airport2 variables which are all from table 6 and none from table 1a. There are other variables to check for NA patters as well such as the low and large market share variables both dependent on table 1a and 6. It shows that for the low market share variable, table 1a has much more NA's at 1399 compared to tables 6 at 568. Furthermore, for the large market share variable table 1a consists of 1331 compared to table 6 at 142. A trend that piggybacks off the previously stated is when you split the data on table 1a and 6, you find that the low_fare variable follows the low_marketshare variable exactly, and same goes for the lg_fare variable having the same exact missing values as the lg_marketshare. Since table 1a refers to the information regarding flights between pairs of airports
3) In 2017, the cities with the most connections to other cities are Atlanta, GA (Metropolitan Area) and Washington, DC (Metropolitan Area) with 13771 and 15945 connections respectively. The cities with the least connections to other cities in 2017 were Brainerd, MN and Greenville, MS with both 1 connections each. When comparing results from 10 years earlier in 2007, we find that the cities with the most connections were again Atlanta, GA and Washington, DC but this time with 651 and 762 connections respectively. The cities with the least connections in 2007 were Brunswick, GA and Allentown/Bethlehem/Easton, PA with 1 connection each. Lastly, when compared to results 20 years earlier in 1997, we find that the cities with the most connections were again Atlanta, GA and Washington, DC but with 667 and 748 connections respectively. Alternatively, the cities with the least connections in 1997 were Athens, GA and Brownsville, TX with 1 connection each.

   The population growth of the two cities Atlanta, GA and Washington DC rose over 15 times the amount in 10 years from 2007 to 2017 which means each year their connections on average grew 150%. Comparatively so, their 1997-2007 growth gap was almost nonexistent because it went from 651 to 667 and 762 to 748 which are pretty small differences over the course of 10 years.

   The cities which had the largest increase in connectivity for the city1 data using table 6 were: [Atlanta GA, Chicago IL, Boston MA, Dallas/Fort Worth TX, and Denver CO]. The cities with the largest increase in connectivity for the city2 data using table 6 were: [Washington DC, San Francisco CA, New York City NY, Seattle WA, and St. Louis MO]

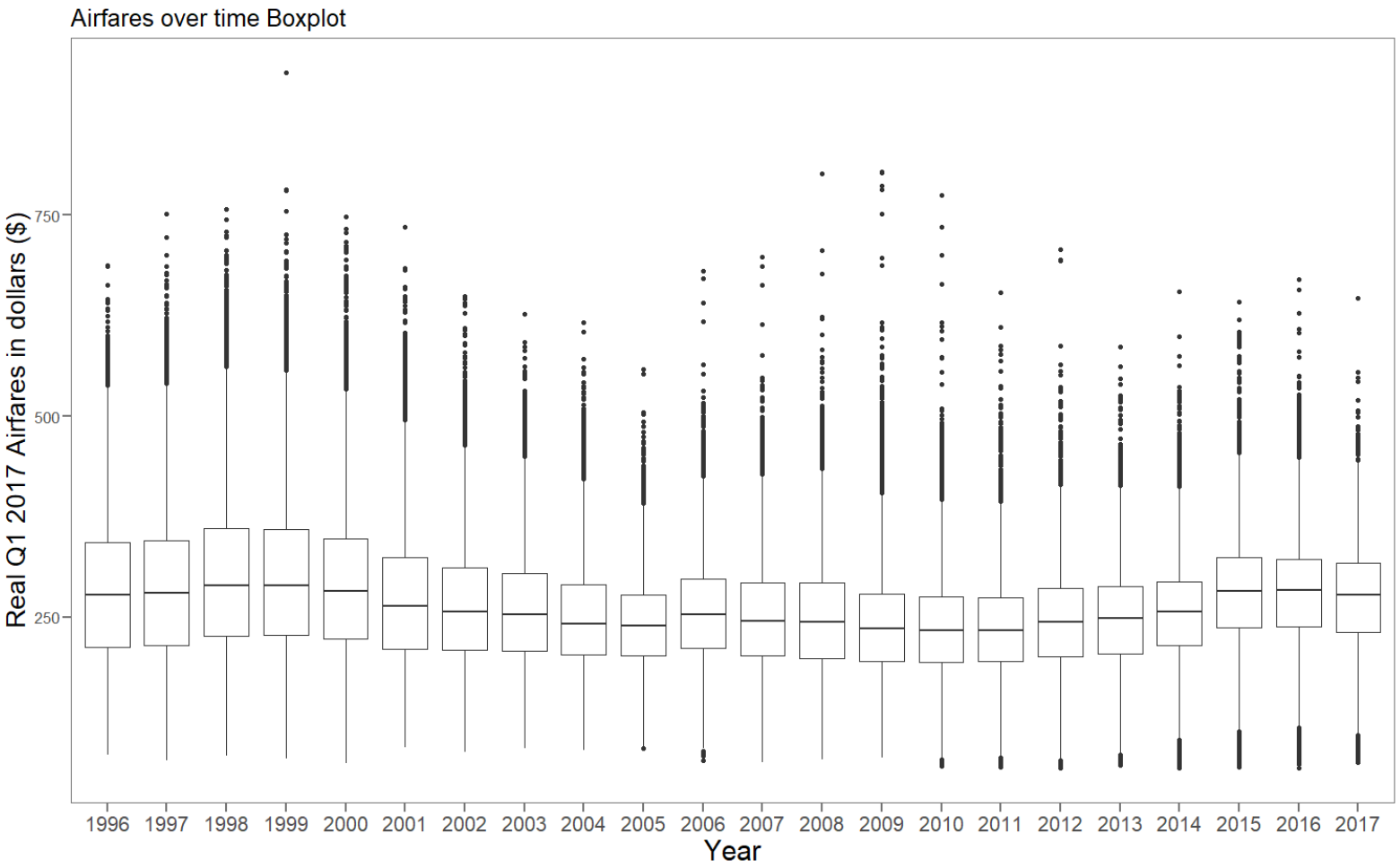




4) Above are a couple of graphical representations of the approximate total number of passengers over the years. The first plot is a violin plot which tracks the approximate total number of passengers over the span of the years of the data. For each year, it shows a graphical breakdown of the distributional breakdown and quantiles of the total passengers. This plot is

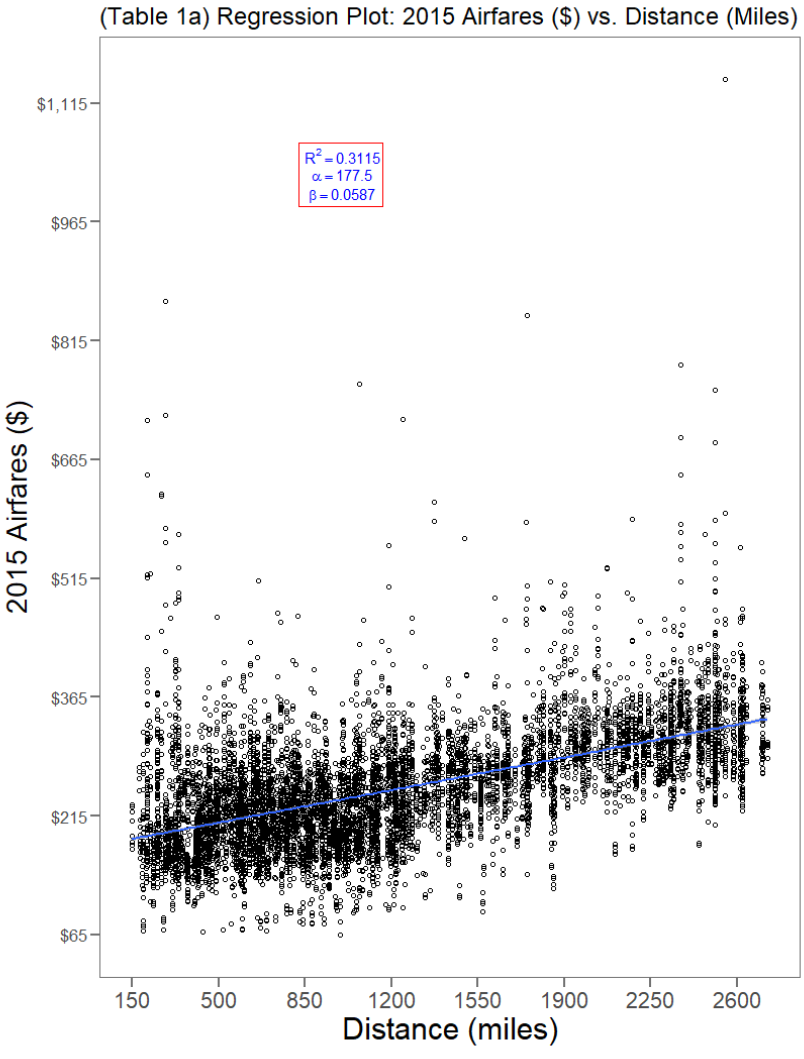
useful here in that it tells us us the long term trends of total passengers as well as the within variability of the total passengers. What we can interpret from this plot is that we notice a wave like function that increased over time. It's hard determine the exact cause of increased passengers over time. It may be from increased population growth over time, and while I think that is part of it, there are other factors that aid in the growth. The next plot is line density 2d plot which looks at the total number of passnegers over the years faceted on each of the 4 quarters of the year over time. In general the long term patterns shows an average increasing number in total passengers. One can notice some interesting observations from the second plot such as in 2001 where 9/11 occurred, we see a large dip in total passengers in Q3 near 2001 where the incident would have occurred. Consequently, looking at Q4 and the following years Q1 and Q2 we start by noticing the large dip in Q4 as a result of the airplane terrorist attack incident followed by a plateau for about a year in Q1 and Q1 respectively. Not only can we see the 2008 dip in Q1 and Q2 from the line plots, but we also notice that downward trend in 2008 on our violin plot as well, which might be correlated to the years 2008-2010 when the housing market crashed and created a recession with inceasing unemployment. Other than the previously mentioned sharp declines in the total number of passengers. Lastly, I think the huge drip in our graphics towards 2017 is from the lack of passengers data from that year. We only have Q1 data and since it the plot is basing it off a yearly basis, it makes sense it has a huge drop given we only have data for the first quarter.

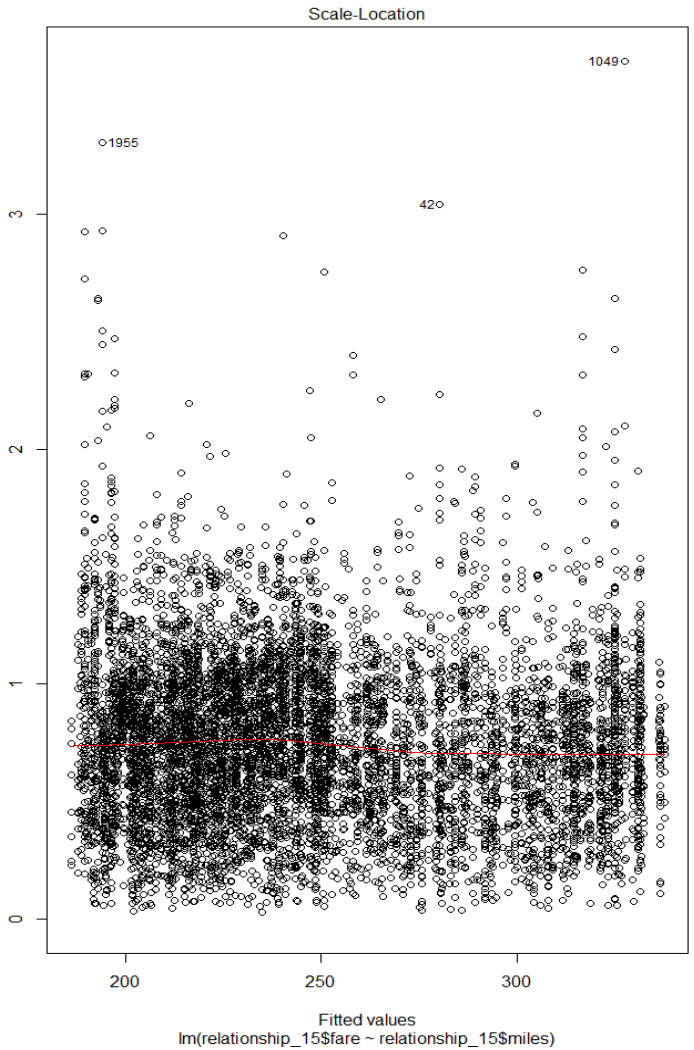
5) (Code for #5 in appendix below)
6) (Boxplot graphic below) Given the discrete x variable (years) and the continuous y-variable (Real Q1 Airfare in dollars) I saw it best to show how airfares have changed over time. It might have been better graphically if we were able to receive more accurate date time data so that we

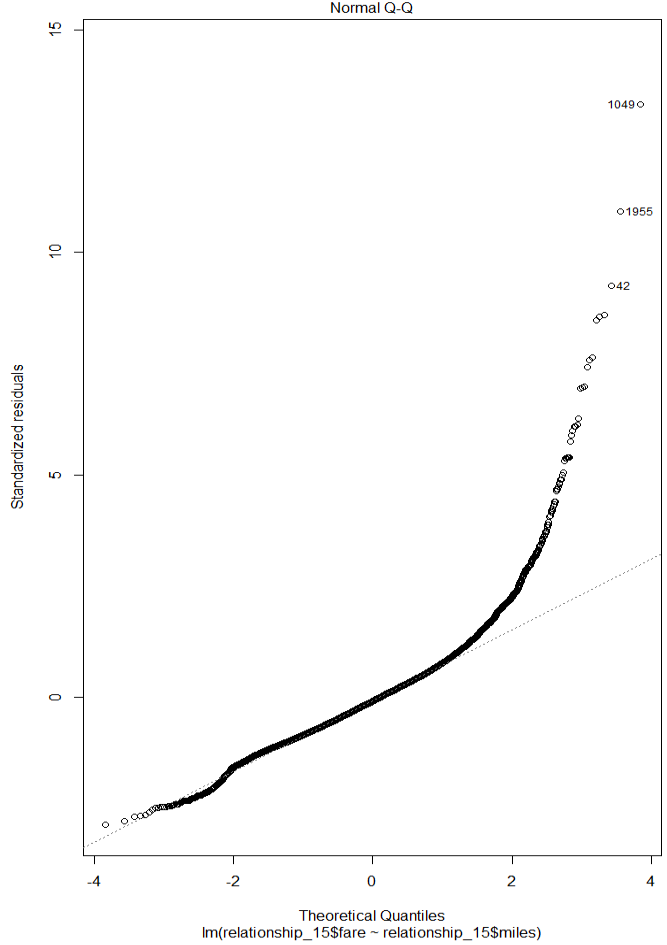


could have had a continuous x-axis and thus have an accompanying smooth line plot. Here we notice very slight fluctuations in the mean airfares over time resulting in rising and falling but ultimately coming back to where it was in 1997. Another interesting note is that there is a very large number of high airfare outliers in the data. It could mean that there will always be premium on traveling and people shelling out a bunch of money for first class flights. There aren't many lower airfare outliers in the data until it becomes a consistent feature of the data starting in 2010 and steadily rising. In general, the mean and the interquartile range does not seem to significantly change much over the years.

7) Starting with the table 1a data, I constructed a linear model plot based off the outcome variable (fare) and the independent variable (miles). While there are restrictions to my model which I will later discuss, the linear model works effectively for this data for most of observations. Looking at the normal Q-Q plot we see a very tight line following the normal line but it strays a little bit before and then after about the 1.75 theoretical quantile. Our data holds with regards to some of the assumptions. For one there is some linear relationship between the independent and dependent variables and can be seen in the linear densely packed observations in the scatterplot. Furthermore, the variables are multivariate normal for most of the data. I understand that the Q-Q plot does not hold the line for all observations but we can explain the problems, limitations, and outliers. For table 1a airfare data vs distance we notice a slight

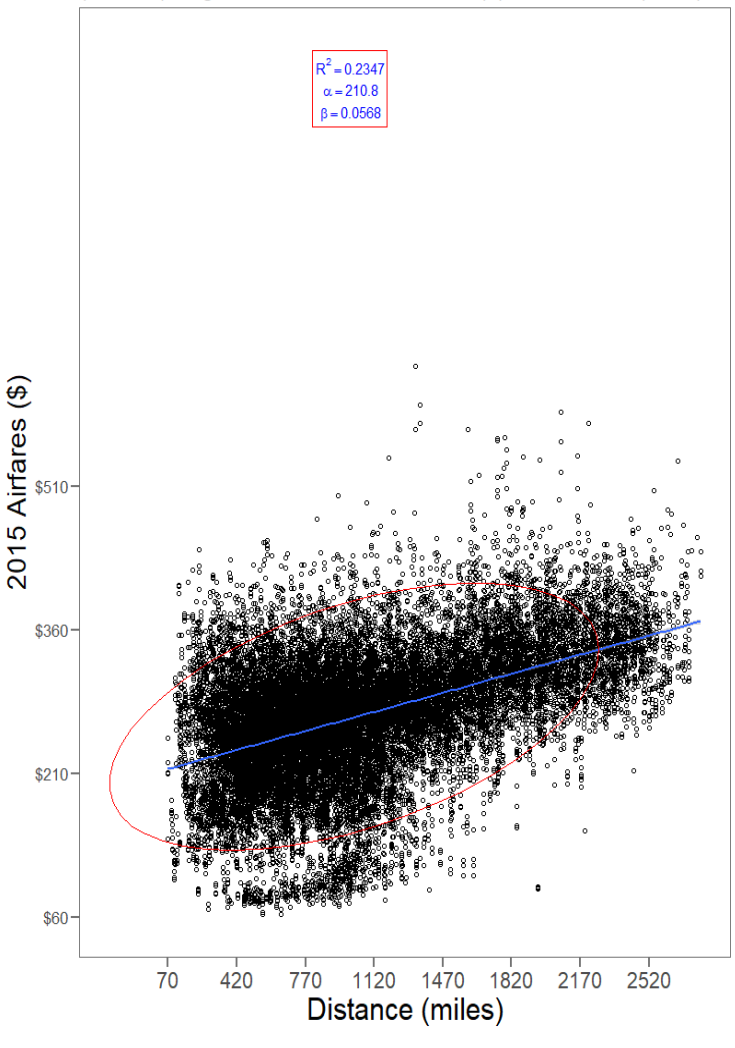


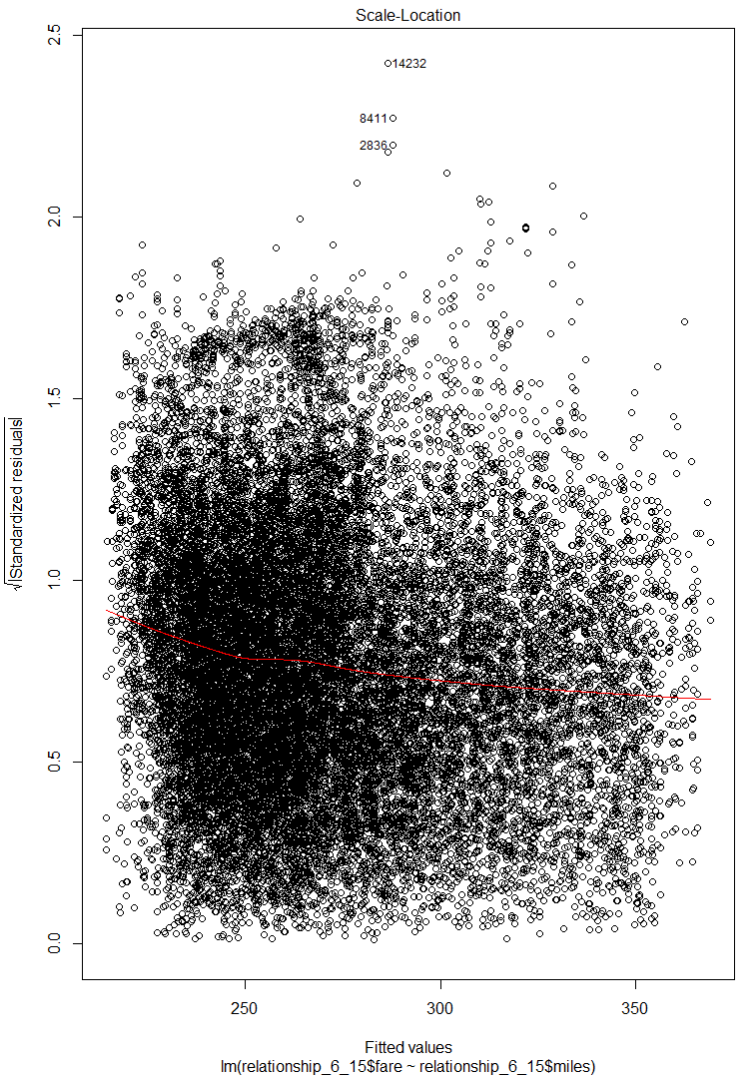


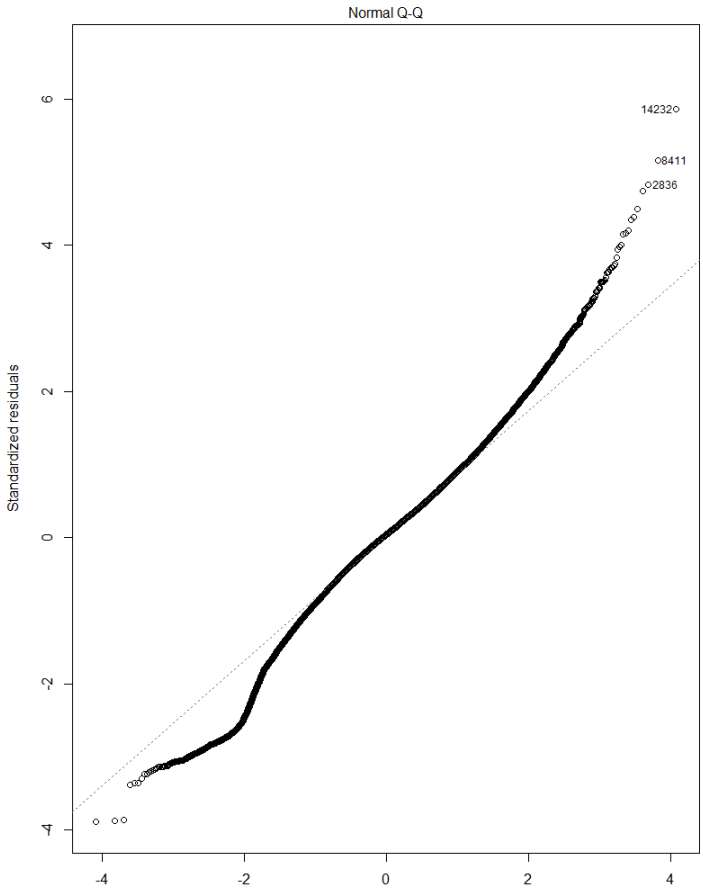


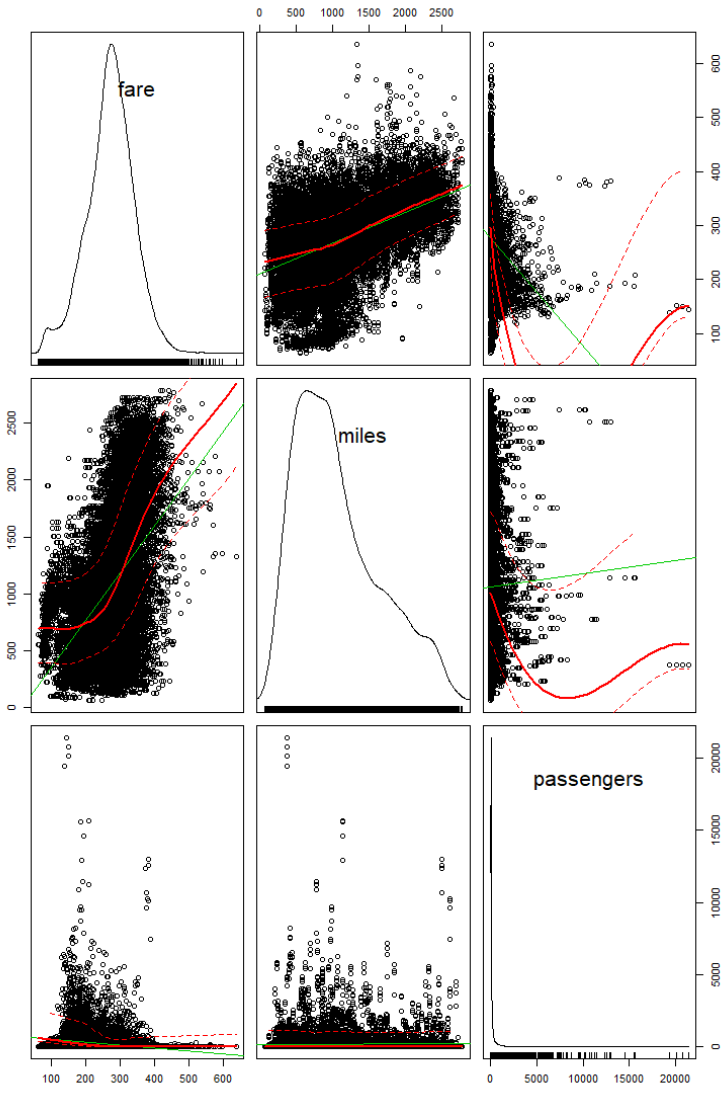
positive linear trend in the data with the $R^2 = 0.3115$ which tells of a weak linear trend. Statistically speaking this results in approximately 31% of variation in airfare costs can be explained by our model. Furthermore, the standardized residual plot shows that the relationship looks approximately normal with a constant looking variance. This is flawed in that data in the upper quartile cannot be explained by normality assumptions because of outliers or another external cause. Maybe the deviations can be interpreted as for people who are flying such long distances, they prefer maximum comfort and amenities and therefore result in the fare variations. Furthermore, the flights themselves need to accommodate more gas, time, and resources, for the long flights. Once the flight distances pass a certain mileage, the linearity breaks down and we will see a more rapid trend in prices increasing because of these combined factors. Another limitation of data is that it that more people are traveling shorter to moderate flight distances in general compared to those who are traveling further distances which could affect the linear model.

(Table 6) Regression Plot: 2015 Airfares ($) vs. Distance (Miles)





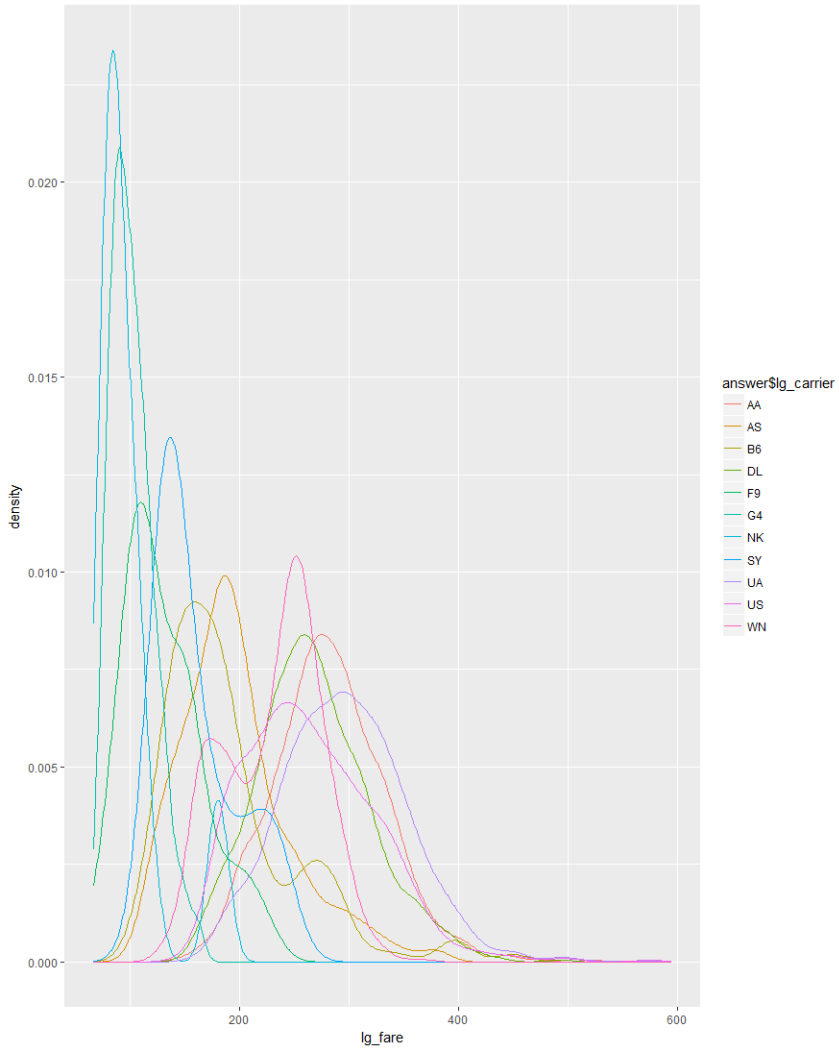




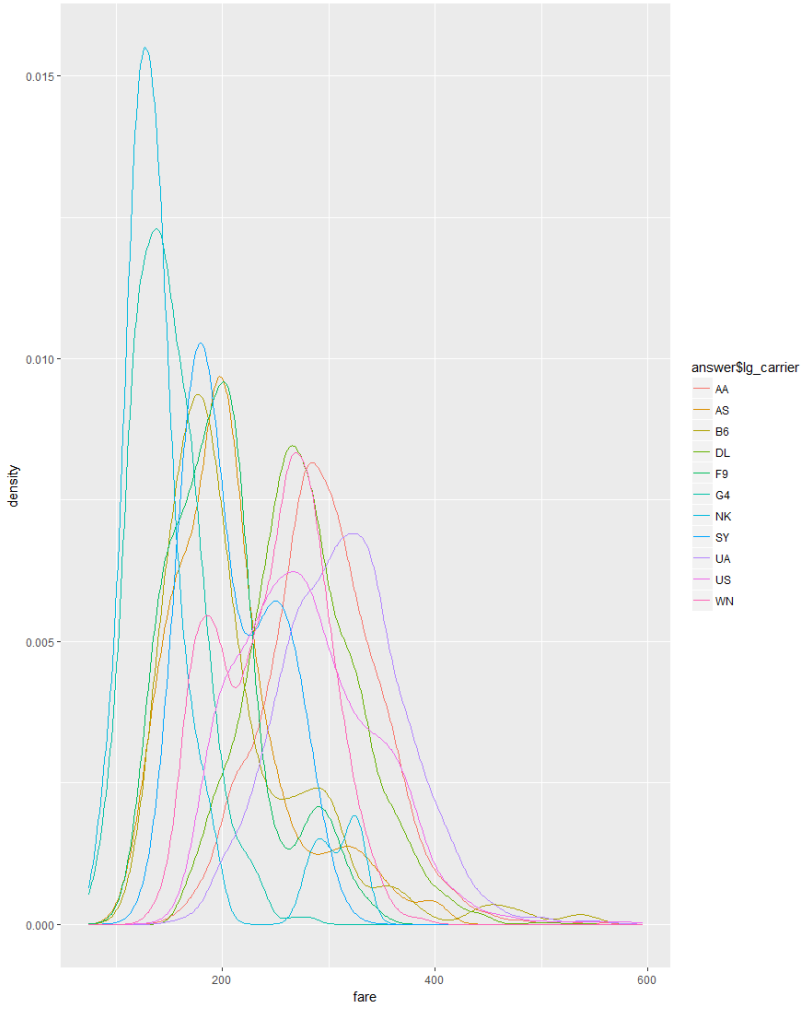
Undergoing the same process of linear modeling I got this visualization with table 6 data and added a confidence ellipse of 95% for more visual inferencing. Since table 6 contains information about flights between pairs of cities and not between airports like table 1a, we notice a different distribution of our data. Not only are there more observations in general, but we notice a very dense clustering of observations towards the center-left of the ellipse. Another important note is that the standardized residual plot of table 6 data strays a just a little further from being approximately normal because of the huge clustering of fitted airfare values. While it's hard to say whether the difference in the visual distribution of our data comes from just having more observations than table 1a, it's clear there is more noise in the data for table 6 observations and may lead to a harder interpretation of the inferential statistics. Like previously mentioned the normal Q-Q plot shows the assumption of the data coming from a normally distributed population for about a little over half of the quantiles. I believe there are more capable models of interpreting such a large and slightly convoluted data. In general, when looking at both table 1a and 6 plots, I would say one could notice a slightly positive linear trend with respect to airfares vs miles.

8) (The matrix scatterplot to the left is for this questions). For this question, I started off by trying to change my model by using the glm (general linear model) function to try and determine if a certain family of distributions would better fit my data. When I ended up finding that the model did not better suit my data I omitted it. While I'm not sure this was the best approach available, trying to accurately model the data deemed a bit challenging. After trying each family of distributions in glm() and failing to see better results in residuals vs. fitted or the normal Q-Q plot, I decided to just use the linear model function to try and model the data while accounting for any immediate deviations by explanations of the data and any potential violations of the assumptions. By adding another term to the linear model, the multivariate data visualization becomes a little harder because the extra variable needs to be accounted for in an appropriate way.

Recalling upon information from STA 108, I was trying to determine if it was necessary to just make it resemble the model, $y_{ij} = a + b_1X_1 + b_2X_2 + e_{ij}$ or rather try adding an interaction term or not. Regardless of how I played around with the interaction term by either multiplying both the miles and passengers or adding them then explicitly putting in their interaction, it did not drastically change the fit of the residuals and Q-Q plot. This can further be supported by the calculation of the Pearson coefficient. When calculated, between x and z (miles and passengers), the coefficient was -0.0622 which suggests a close to 0 negative correlation. Therefore, we may suggest the possibility of just interpreting the slope of our outcome variable. Furthermore, the summary of the models shows that our residual standard error for the models were right around 59 which gives us an idea of how far observed airfare costs are from the predicted or fitted airfare costs. This leads me to believe there is so much noise in our data that it is difficult to sort through the noise without cleaning or including more variables to clean the data more and just resorted to a linear model for the modeling of this huge dataset. In all the Q-Q plots for both table 1a and 6, and for both questions 7 and 8, we notice a very clear valid normality assumption in the middle of the plot until it reaches its outer theoretical quantiles where it becomes largely skewed. When comparing the summaries of the models used for table 1a in both question 7 and 8, we notice an almost negligible difference in the residual quartile values, intercept, mile slope, and residual standard error. The model in question 8 with an added predictor variable and account for interaction did not make a change to airfare costs. I thought it the graphic to the left, the matrix scatter plot where the diagonals are the histograms line distributions and then each paired diagonal is tested against each other to determine if there are any clear trends of linearity or signs of warning. In this case we can rule out passengers' data as being meaningless in our interpretation of these results and focus more on the fares and the miles



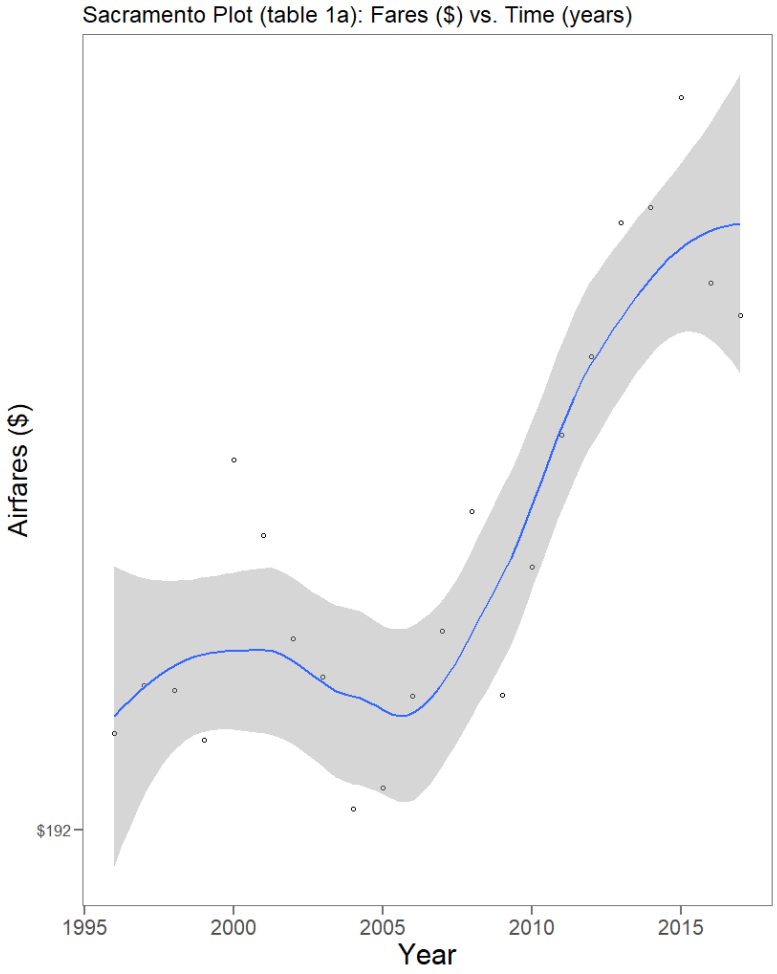
9) Both the histograms for the 'fare' and 'lg_fare' columns of our specific data for problem 9 show that they are approximately normally distributed and thus we can run a 2-sample t-test to determine whether the means of the two groups are equal or not. There are some assumptions to this test such as both groups are sampled from normal distributions with equal variances. Next, we shall state that the null hypothesis is that the two means are equal and the alternative is that they aren't. I calculated the variances of both the fares and the large fares and they came out with very close values. Thus, we can perform out 2 sample t-test on the fares and large fares (result table in appendix). We reject the null hypothesis that the true difference in means is equal to zero because our p-value < alpha. So, we accept the alternative hypothesis and are

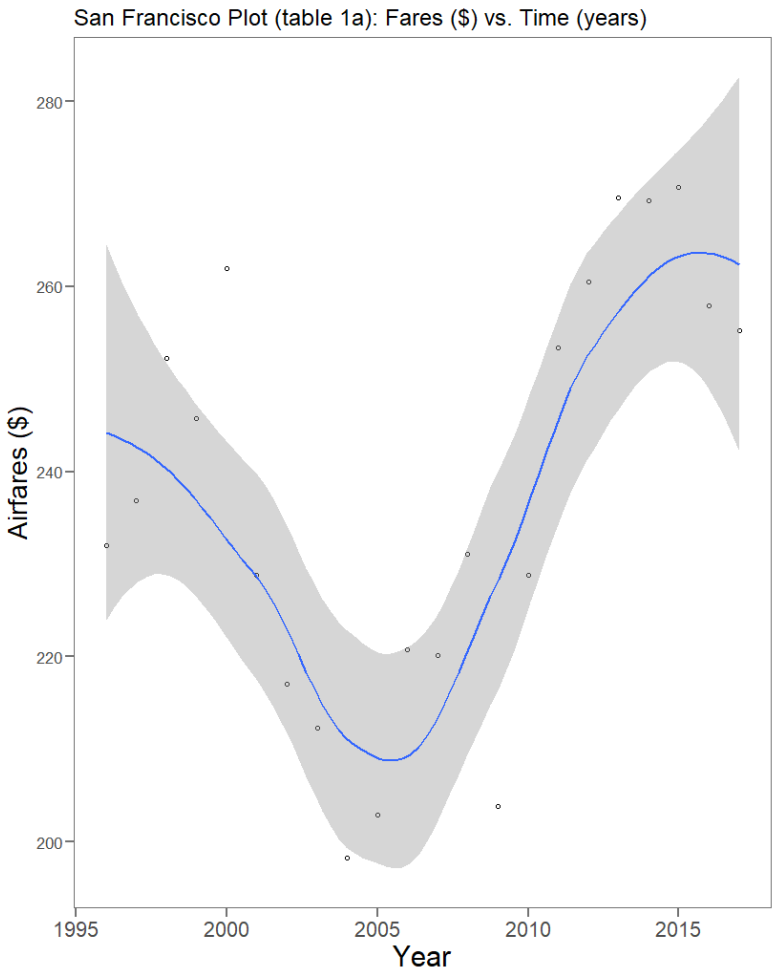


95% confident that the true difference in means is between [14.943, 15.572]. This means that for the carriers with the largest market share, they are on average between [14.943, 15.572] less than the mean fares of other carriers. The following colored histograms are the breakdown of both the large fair and regular fare average which shows higher peak densities for

10) Using table 1a to compare the given list of airports, we found some interesting points. For one, I was unable to find Oakland and San Jose in either the city1 or city2 variables in the data for table 1a. Thus, I concluded with analysis on just the cities I could get which were Sacramento and San Francisco. For the first question, the fares differ between these airports in that the mean difference in fares ($) is almost 10 dollars less through Sacramento than San Francisco at respective prices of $226.99 and $236.72. This could make sense when thinking about it because San Francisco is in the Bay Area which has typically higher prices of basically everything compared to living in Sacramento. Also, there are probably many places that you can go departing from SF than Sacramento which may be much further away, and as we have noted in the previous questions, that airfares slightly increase with distance and other factors as well which may play into why SF has a higher average fare price. Next, I could find that San Francisco had the most long-distance connections at 2704 miles and with 755 connections between itself and Boston, MA. There might be a strong case to advocate for Sacramento having the largest number of long-distance connections even though the distance in miles is short by almost a couple hundred. I found that Sacramento has 2341 long-distance connections with New York (Metropolitan Area) at 2553 miles. So, while SF may have the longest distance of long-distance connections, coming in a close second place Sacramento has over 3 times the amount of those long-distance connections. I found these results interesting because SF had the most long-distance connections even though San Francisco has a lower average distance (miles) to its connections that does Sacramento. It showed that Sacramento had 1754.49 average miles compared to San Francisco at 1718.42. Maybe there are possible geographical and socioeconomic factors that attribute the smaller mean distance flights from SF. Potential geographical or socioeconomic factors such as people who have the money are willing to pay for shorter distanced flights compared to those in Sacramento who will drive further places without consideration of air travel. Maybe people from SF are more likely to travel in general to places that are closer than people who travel from Sacramento. I'm just speculating without concrete data of the populations of people who depart from these airports but I'm just trying to add speculation. One could also calculate between the populations of Sac and SF, whether there is a statistically significant difference between the means of the distances to determine if the

differences need further interpretation or inferencing. When looking at the findings for how fares differ between Sacramento and San Francisco airports by year, we notice a decently linear trend by Sacramento's data with an adjusted $R^2 = 0.565$ which suggests a moderate linear correlation between fares over the years. The data seems to have reasonably equal variances across the variables from the plot. Comparatively, the plot for San Francisco does not look close to Sacramento's. It has much higher increased variability and does not resemble linearity much at all. It starts out by showing that fares are high to start in the late 1990's, then they go way down throughout the 2000's, and then they shoot back up after 2010. The data's density plot would seem to resemble that of a cosine graph. In general Sacramento's fares over time stay much more consistent and linear but both equal out to similar fare distribution towards the end of the years. (Below are the graphs accompanying the results dependent upon year)





Appendix

```
# references http://ggplot2.tidyverse.org/reference/index.html

# http://www.sthda.com/english/wiki/ggplot2-axis-scales-and-transformations

# for ggplot, and plot trasformations and edits

# http://erdavenport.github.io/R-ecology-lesson/04-dplyr.html for R tutorials and basics

# Other references are scattered throughout code right before each piece of codes that needs
referencing.

library(tidyverse)
library(ggmosaic)
library(plyr)
library(ggthemes)
library(grid)
library(sqldf)
library(reshape)
library(plotly)
library(scales)
library(car)
library(RColorBrewer)

# Note I commented out my variables p1, p2, p3, p4 which are the residual and Q-Q... etc
# plots of the linear models in 7 and 8 b/c it messed up R script when running whole thing.

set.seed(1234)

airfare <- read_csv("E:/Davis/STA 141A/DATA/airfare.zip")
cpi_1996_2017 <- read_csv("E:/Davis/STA 141A/DATA/cpi_1996_2017.csv")

# 1)

   table_1a = filter(airfare, table %in% c("1a"))
   table_6 = filter(airfare, table %in% c("6"))

#2)
   # To find the timespan that the data covers we use the "year" column vector
   min_year_1a = min(table_1a$year)
   max_year_1a = max(table_1a$year)
   min_quarter_1a = min(table_1a$quarter)
   max_quarter_1a = max(table_1a$quarter)

   min_year_6 = min(table_6$year)
   max_year_6 = max(table_6$year)
   min_quarter_6 = min(table_6$quarter)
   max_quarter_6 = max(table_6$quarter)
```

```
   min_year = min(cbind(min_year_1a,min_year_6))
   max_year = max(cbind(max_year_1a,max_year_6))
   min_quarter = min(cbind(min_quarter_1a,min_quarter_6))
   max_quarter = max(cbind(max_quarter_1a,max_quarter_6))
   cat("The timespan that this data covers is from quarter",min_quarter, min_year,"to quarter",
max_quarter, max_year)
   # (1996-2017)

   # To find any gaps in the "years" or "quarters" column vectors where there are NA's.

   NAs_Quarter_1a = table_1a$quarter[is.na(table_1a$quarter)]
   NAs_Quarter_6 = table_6$quarter[is.na(table_6$quarter)]
   NAs_Year_1a = table_1a$year[is.na(table_1a$year)]
   NAs_Year_6 = table_6$year[is.na(table_6$year)]

   # To look for any missing NA's or patterns

   table(is.na(table_1a$lg_carrier))
   table(is.na(table_6$lg_carrier))

   table(is.na(table_1a$low_carrier))
   table(is.na(table_6$low_carrier))
   # Notice that table 1a have the most missing values in both carrier variables.

   table(is.na(table_1a$low_fare))
   table(is.na(table_6$low_fare))

   table(is.na(table_1a$lg_fare))
   table(is.na(table_6$lg_fare))

   table(is.na(table_1a$low_marketshare))
   table(is.na(table_6$low_marketshare))

   table(is.na(table_1a$lg_marketshare))
   table(is.na(table_6$lg_marketshare))

#3)
   # Create a table from table 6 of both city 1 and city 2 and find the max value of each to determine the
most frequent cities
   # This tells us that those cities have the most connections with other cities
   by_year1 = split(table_6$city1, table_6$year, drop = T)
   by_year1 = table(by_year1$`2017`)
   max1 = max(by_year1)
   ATL = names(by_year1[which.max(by_year1)])
   by_year2 = split(table_6$city2, table_6$year, drop = T)
   by_year2 = table(by_year2$`2017`)
   max2 = max(by_year2)
   DC = names(by_year2[which.max(by_year2)])
   cat("\nIn 2017, the cities with the most connections were,",ATL,"and",DC,"with",
max1,"and",max2,"connections respectively.")

   new_table3 = table(table_6$city1)
   new_table4 = table(table_6$city2)
```

```
   MN = names(new_table3[which.min(new_table3)])
   MS = names(new_table4[which.min(new_table4)])
   max3 = min(new_table3)
   max4 = min(new_table4)
   cat("\nIn 2017, the cities with the least connections were,",MN,"and",MS,"with",
max3,"and",max4,"connections respectively.")

   # Now we split the data by the city and year using the split() functions based on year 2007.
   # Then we will store name and the count of each variable
   by_year1 = split(table_6$city1, table_6$year, drop = T)
   by_year1 = table(by_year1$`2007`)
   max1 = max(by_year1)
   ATL = names(by_year1[which.max(by_year1)])
   by_year2 = split(table_6$city2, table_6$year, drop = T)
   by_year2 = table(by_year2$`2007`)
   max2 = max(by_year2)
   DC = names(by_year2[which.max(by_year2)])
   cat("\nIn 2007, the cities with the most connections were,",ATL,"and",DC,"with",
max1,"and",max2,"connections respectively.")

   by_year3 = split(table_6$city1, table_6$year, drop = T)
   by_year3 = table(by_year3$`2007`)
   min1 = min(by_year3)
   GA = names(by_year3[which.min(by_year3)])
   by_year4 = split(table_6$city2, table_6$year, drop = T)
   by_year4 = table(by_year4$`2007`)
   min2 = min(by_year4)
   TX = names(by_year4[which.min(by_year4)])
   cat("\nIn 2007, the cities with the least connections were,",GA,"and",TX,"with",
min1,"and",min2,"connections respectively.")

   # Lastly we need to split the data according to the city based on year 1997
   by_year1 = split(table_6$city1, table_6$year, drop = T)
   by_year1 = table(by_year1$`1997`)
   max1 = max(by_year1)
   ATL = names(by_year1[which.max(by_year1)])
   by_year2 = split(table_6$city2, table_6$year, drop = T)
   by_year2 = table(by_year2$`1997`)
   max2 = max(by_year2)
   DC = names(by_year2[which.max(by_year2)])
   cat("\nIn 1997, the cities with the most connections were,",ATL,"and",DC,"with",
max1,"and",max2,"connections respectively.")

   by_year3 = split(table_6$city1, table_6$year, drop = T)
   by_year3 = table(by_year3$`1997`)
   min1 = min(by_year3)
   GA = names(by_year3[which.min(by_year3)])
   by_year4 = split(table_6$city2, table_6$year, drop = T)
   by_year4 = table(by_year4$`1997`)
   min2 = min(by_year4)
   TX = names(by_year4[which.min(by_year4)])
   cat("\nIn 1997, the cities with the least connections were,",GA,"and",TX,"with",
min1,"and",min2,"connections respectively.")
```

```
   # Lastly, we need to find the max value the largest increased connectivity

   by_city1 = split(table_6$city1, table_6$year, drop = T)
   by_1997_1 = table(by_city1$`1997`)
   by_2017_1 = table(by_city1$`2017`)

   by_city2 = split(table_6$city2, table_6$year, drop = T)
   by_1997_2 = table(by_city2$`1997`)
   by_2017_2 = table(by_city2$`2017`)

   x_1997 = data.frame(count(by_1997_1))
   x_2017 = data.frame(count(by_2017_1))

   xx_1997 = data.frame(count(by_1997_2))
   xx_2017 = data.frame(count(by_2017_2))

# For city 1
city1 = merge(x_1997,x_2017, by = "x.Var1")
# To delete the unneccesary columns
drops = c("freq.x", "freq.y")
city1 = city1[ , !(names(city1) %in% drops)]
colnames(city1) = c("City/State", "    (YEAR) 1997", "    (YEAR) 2017")

city1$increased_connections = (city1$`    (YEAR) 2017` - city1$`    (YEAR) 1997`)
city1$increased_connections = abs(city1$increased_connections)
city1 = city1 %>%
         arrange(desc(increased_connections))

biggest_city1 = head(city1, 5)
biggest_city1

# For city2
city2 = merge(xx_1997,xx_2017, by = "x.Var1")
# To delete the unneccesary columns
drops = c("freq.x", "freq.y")
city2 = city2[ , !(names(city2) %in% drops)]
colnames(city2) = c("City/State", "    (YEAR) 1997", "    (YEAR) 2017")

city2$increased_connections = (city2$`    (YEAR) 2017` - city2$`    (YEAR) 1997`)
city2$increased_connections = abs(city2$increased_connections)
city2 = city2 %>%
  arrange(desc(increased_connections))

biggest_city2 = head(city2, 5)
biggest_city2

#4)

   # Aggregate function is best here because it is like tapply but the output is a data frame where we
specify our numerical variable first followed by the categorical variables and the FUN function to apply

   agg_funct_cities = aggregate(passengers ~ quarter + year, table_6, sum)
```

```
   agg_funct_airports = aggregate(passengers ~ quarter + year, table_1a, sum)

   agg_funct = aggregate(passengers ~ quarter + year, airfare, sum)
   # plot function using ggplot and the 2d density geom function. We set the data equal to our
agg_function for citites and thenn pass the year on the x axis, and passengers on the y to the aes().
   # Then I faceted the data based on the quarters so we can see the yearly trends of each of the
quarters of passnegers over time.
     plot1 = ggplot(data = agg_funct, aes(year, passengers)) +
       geom_density_2d(stat="identity") +
       facet_wrap(~ quarter, nrow = 2) +
          labs(x = "Year", y = "Approximate Total # of Passengers", title = "Breakdown of Distinct Quarterly
Total Passengers Over the Years (Cities & Airports)") +
            theme(plot.title = element_text(size = rel(1.65))) +
            theme(axis.title.y = element_text(size = rel(1.77))) +
            theme(axis.text.x = element_text(size = rel(1.75))) +
            theme(axis.text.x=element_text(angle=25, hjust= .5)) +
            theme(axis.title.x = element_text(size = rel(1.85))) +
            theme(axis.text.y = element_text(size = rel(1.35))) +
            theme(axis.ticks.length=unit(0.25,"cm")) +
            theme(axis.ticks = element_line(size = rel(2))) +
            theme(plot.margin=unit(c(1,1,1.5,1.2),"cm")) +
            theme(strip.text.x = element_text(size = 17, colour = "red"))

     plot1

     plot2 = ggplot(data = agg_funct, aes(factor(year), passengers)) +
           geom_violin(scale = "count", adjust = 1.25, aes(fill = factor(year)), draw_quantiles = c(0.25, 0.5,
0.75)) +
                  labs(x = "Year", y = "Approximate Total # of Passengers", title = "Violin Plot of Total
Passengers over time (with quantiles)") +
                    theme_few() + theme(plot.title = element_text(size = rel(1.65))) +
                    theme(axis.title.y = element_text(size = rel(1.77))) +
                    theme(axis.text.x = element_text(size = rel(1.75))) +
                    theme(axis.text.x=element_text(angle=25, hjust= .5)) +
                    theme(axis.title.x = element_text(size = rel(1.85))) +
                    theme(axis.text.y = element_text(size = rel(1.35))) +
                    theme(axis.ticks.length=unit(0.25,"cm")) +
                    theme(axis.ticks = element_line(size = rel(2))) +
                    theme(plot.margin=unit(c(1,1,1.5,1.2),"cm")) +
                    theme(strip.text.x = element_text(size = 22, colour = "red", angle = 22.5)) +
                    theme(legend.title = element_text(size = rel(1.65)))

     plot2

#5)
     # Learned of the reshape() methodology of this problem from Piazza about question 5 from
anonymous
     # Prefer SQL statements sometimes because of my experience from last summer internship
     temp = sqldf("SELECT year, quarter, fare FROM airfare")
     # Remap factor levels for easier merge and join
     temp$quarter = mapvalues(temp$quarter, from = c("1", "2", "3", "4"), to = c("Q1", "Q2", "Q3", "Q4"))
```

```
    # Method: Found CPI for Q1 - Q4 by taking average of 3 month groupings
    cpi_1996_2017$Q1 = ((cpi_1996_2017$Jan) + (cpi_1996_2017$Feb) + (cpi_1996_2017$Mar)) / 3
    cpi_1996_2017$Q2 = ((cpi_1996_2017$Apr) + (cpi_1996_2017$May) + (cpi_1996_2017$Jun)) / 3
    cpi_1996_2017$Q3 = ((cpi_1996_2017$Jul) + (cpi_1996_2017$Aug) + (cpi_1996_2017$Sep)) / 3
    cpi_1996_2017$Q4 = ((cpi_1996_2017$Oct) + (cpi_1996_2017$Nov) + (cpi_1996_2017$Dec)) / 3

    # To delete certain columns
    drops = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Oct", "Nov", "Dec", "HALF1",
"HALF2")
    cpi_1996_2017 = cpi_1996_2017[ , !(names(cpi_1996_2017) %in% drops)]

    cpi_1996_2017 = reshape(cpi_1996_2017, varying = c("Q1","Q2","Q3","Q4"), v.names = "CPI",
direction = "long")

    colnames(cpi_1996_2017) = c("year","Sep","quarter","CPI","id")
    drops = c("id")
    cpi_1996_2017 = cpi_1996_2017[ , !(names(cpi_1996_2017) %in% drops)]

    CPI_b = sqldf("SELECT Sep FROM cpi_1996_2017 WHERE Year = 2017")
    CPI_b = as.double(head(CPI_b, 1))

    drops = c("Sep")
    cpi_1996_2017 = cpi_1996_2017[ , !(names(cpi_1996_2017) %in% drops)]

    # Creating new altered name for airfare so to not affect and get rid of the table 1a observations since
I am merging the cpi_1996_2017 data with table_6 data only.
    airfare_merged = merge(cpi_1996_2017, table_6, by = c("year", "quarter"), all.y = T)

    airfare_merged$real17_fare = (airfare_merged$fare)*(CPI_b/(airfare_merged$CPI))

#6)
    # We want real Q1 2017 airfare fares over time

    fares_over_time = sqldf("SELECT year, real17_fare FROM airfare_merged ORDER BY year")

    box_plot = ggplot(data = airfare_merged, aes(x = factor(airfare_merged$year), y =
airfare_merged$real17_fare)) +
      geom_boxplot() +
      labs(x = "Year", y = "Real Q1 2017 Airfares in dollars ($)", title = "Airfares over time Boxplot") +
      theme_few() + theme(plot.title = element_text(size = rel(1.65))) +
      theme(axis.title.y = element_text(size = rel(1.77))) +
      theme(axis.text.x = element_text(size = rel(1.75))) +
      theme(axis.title.x = element_text(size = rel(1.85))) +
      theme(axis.text.y = element_text(size = rel(1.35))) +
      theme(axis.ticks.length=unit(0.25,"cm")) +
      theme(axis.ticks = element_line(size = rel(2))) +
      theme(plot.margin=unit(c(1,1,1.5,1.2),"cm")) +
      theme(strip.text.x = element_text(size = 22, colour = "red", angle = 22.5)) +
      theme(legend.title = element_text(size = rel(1.65)))

    box_plot

#7)
```

```
    # references: http://t-redactyl.io/blog/2016/05/creating-plots-in-r-using-ggplot2-part-11-linear-regression-plots.html
    #           http://www.sthda.com/english/wiki/ggplot2-axis-scales-and-transformations
   # Part 1: Find a relationship between fare and distance for year 2015 using an appropriate statistical model or test
      # reload airfare to ensure all table values are there

      relationship_15 = sqldf("SELECT year, miles, fare, `table` FROM airfare WHERE year = 2015 AND`table` LIKE '%1a%'")

      model_q7 = lm(relationship_15$fare ~ relationship_15$miles)

      summary(model_q7)
      # p1 = plot(model_q7)

    reg_plot1 = ggplot(data = relationship_15, aes(x = relationship_15$miles, y = relationship_15$fare)) +
                geom_point(shape = 1) + geom_smooth(method = lm) +
                labs(title = "(Table 1a) Regression Plot: 2015 Airfares ($) vs. Distance (Miles)") +
                scale_x_continuous(name = "Distance (miles)", breaks = round(seq(min(relationship_15$miles) + 2, max(relationship_15$miles) + 2, by = 350),5)) +
                # changing axis scales and tick amount referece: https://stackoverflow.com/questions/11335836/increase-number-of-axis-ticks
                scale_y_continuous(name = "2015 Airfares ($)", breaks = round(seq(min(relationship_15$fare), max(relationship_15$fare), by = 150),0), labels = dollar) +
                  theme_few() + theme(plot.title = element_text(size = rel(1.49))) +
                  theme(axis.title.y = element_text(size = rel(1.77))) +
                  theme(axis.text.x = element_text(size = rel(1.75))) +
                  theme(axis.title.x = element_text(size = rel(1.85))) +
                  theme(axis.text.y = element_text(size = rel(1.35))) +
                  theme(axis.ticks.length=unit(0.25,"cm")) +
                  theme(axis.ticks = element_line(size = rel(2))) +
                  theme(plot.margin=unit(c(1,1,1.5,1.2),"cm")) +
                  theme(strip.text.x = element_text(size = 22, colour = "red", angle = 22.5))

      reg_plot1 = reg_plot1 + annotate("rect", xmin = 825, xmax = 1165, ymin = 985, ymax = 1065, fill="white", colour="red") +
                    annotate("text", x = 1000, y = 1050, label = "R^2 == 0.3115", colour = "blue", parse = T) +
                    annotate("text", x = 1000, y = 1025, label = "alpha == 177.5", colour = "blue", parse=T) +
                    annotate("text", x = 1000, y = 1000, label = "beta == 0.0587", colour = "blue", parse=T)
      reg_plot1

      relationship_6_15 = sqldf("SELECT year, miles, fare, `table` FROM airfare WHERE year = 2015 AND `table` LIKE '%6%'")

      model_q7_6 = lm(relationship_6_15$fare ~ relationship_6_15$miles)

      summary(model_q7_6)
      # p2 = plot(model_q7_6)

      reg_plot2 = ggplot(data = relationship_6_15, aes(x = relationship_6_15$miles, y = relationship_6_15$fare)) +
                  geom_point(shape = 1) + geom_smooth(method = lm) +
                  labs(title = "(Table 6) Regression Plot: 2015 Airfares ($) vs. Distance (Miles)") +
```

```
        scale_x_continuous(name = "Distance (miles)", breaks =
round(seq(min(relationship_6_15$miles) + 3, max(relationship_6_15$miles) + 3, by = 350),5)) +
        # changing axis scales and tick amount referece:
https://stackoverflow.com/questions/11335836/increase-number-of-axis-ticks
        scale_y_continuous(name = "2015 Airfares ($)", breaks =
round(seq(min(relationship_6_15$fare) - 3, max(relationship_6_15$fare) - 3, by = 150),0), labels =
dollar) +
         theme_few() + theme(plot.title = element_text(size = rel(1.49))) +
         theme(axis.title.y = element_text(size = rel(1.77))) +
         theme(axis.text.x = element_text(size = rel(1.75))) +
         theme(axis.title.x = element_text(size = rel(1.85))) +
         theme(axis.text.y = element_text(size = rel(1.35))) +
         theme(axis.ticks.length=unit(0.25,"cm")) +
         theme(axis.ticks = element_line(size = rel(2))) +
         theme(plot.margin=unit(c(1,1,1.5,1.2),"cm")) +
         theme(strip.text.x = element_text(size = 22, colour = "red", angle = 22.5))


    reg_plot2 = reg_plot2 + stat_ellipse(colour = "red") + annotate("rect", xmin = 805, xmax = 1185,
ymin = 885, ymax = 965, fill="white", colour="red") +
              annotate("text", x = 1000, y = 950, label = "R^2 == 0.2347", colour = "blue", parse = T) +
              annotate("text", x = 1000, y = 925, label = "alpha == 210.8", colour = "blue", parse=T) +
              annotate("text", x = 1000, y = 900, label = "beta == 0.0568", colour = "blue", parse=T)

    reg_plot2


#8)

    # Scatterplot matrices reference: https://www.statmethods.net/graphs/scatterplot.html
    scatter_data1 = sqldf("SELECT fare, miles, passengers FROM airfare WHERE year = 2015 AND `table`
LIKE '%1a%'")
    # From library car: gives us a matrix of plots which can compare variables and their histrograms

    y = scatter_data1$fare # Outcome variable
    x = scatter_data1$miles # Predictor 1
    z = scatter_data1$passengers # Predictor 2

    # Pearson coefficient will tell us how strong the collinearity is between our two variables to
determine if we should intrept the slope of y.
    cor(x, z, method = "pearson")

    # Model with 2 predictor variables
    multi_variate1 = lm(y ~ x + z)
    summary(multi_variate1)
    # p3 = plot(multi_variate1)

    # Model with 2 predictor variables and the interaction term.
    multi_variate2 = lm(y ~ x + z + x:z)
    summary(multi_variate2)
    # p4 = plot(multi_variate2)

    scatterplotMatrix(scatter_data1)
```

```
# Repeat above operations on table 6
      scatter_data2 = sqldf("SELECT fare, miles, passengers FROM airfare WHERE year = 2015 AND `table` LIKE '%6%'")
      y1 = scatter_data2$fare # Outcome variable
      x1 = scatter_data2$miles # Predictor 1
      z1 = scatter_data2$passengers # Predictor 2

      cor(x1, z1, method = "pearson")

      # Model with 2 predictor variables
      multi_variate3 = lm(y1 ~ x1 + z1)
      summary(multi_variate3)
      # p3 = plot(multi_variate1)

      # Model with 2 predictor variables and the interaction term.
      multi_variate4 = lm(y1 ~ x1 + z1 + x1:z1)
      summary(multi_variate4)
      # p4 = plot(multi_variate2)

      scatterplotMatrix(scatter_data2)

#9)  T-test refer     # We want 2015 pairs of cities where the carrier with largest market share has fares below the average for that city pair

    # Thought Process: Select piars of cities,

    # find the carrier(s) with largest mkt share that has fares below calculated avg of 2 citites mean fares

      select_pairs = sqldf("SELECT city1, city2, fare, lg_carrier, lg_fare FROM table_6 WHERE year = 2015 ORDER BY lg_carrier")

      answer = sqldf("SELECT * FROM select_pairs WHERE lg_fare < fare")

      hist(answer$lg_fare)

      hist(answer$fare)

      # Closely equal variances

      v1 = var(answer$lg_fare)

      v2 = var(answer$fare)

      # 2 sample T-test refernces: https://statistics.berkeley.edu/computing/r-t-tests

      t.test(answer$fare, answer$lg_fare, alternative = "two.sided", paired = T)

      ggplot(data = answer, aes(lg_fare))+

        stat_density(aes(group = answer$lg_carrier, color = answer$lg_carrier), position =  "identity", geom = "line")

      ggplot(data = answer, aes(fare))+

       stat_density(aes(group = answer$lg_carrier, color = answer$lg_carrier), position =  "identity", geom = "line")

      t-test reference: https://statistics.berkeley.edu/computing/r-t-tests

          Paired t-test

          data:  answer$fare and answer$lg_fare
```

```
        t = 95.046, df = 10666, p-value < 2.2e-16
        alternative hypothesis: true difference in means is not equal to 0
        95 percent confidence interval:
         14.94298 15.57231
        sample estimates:
        mean of the differences
                    15.25765



#10)

    q_10_1 =  sqldf("SELECT city1 as q10_CA_cities, city2 as Others, fare, miles, year FROM table_1a
WHERE city1 LIKE '%Sacramento%' OR city1 LIKE '%San Francisco%' OR city1 LIKE '%Oakland%' OR city1
LIKE '%San Jose%' GROUP BY fare, city1 ORDER BY city1")

    q_10_2 =  sqldf("SELECT city2 as q10_CA_cities, city1 as Others, fare, miles, year FROM table_1a
WHERE city2 LIKE '%Sacramento%' OR city2 LIKE '%San Francisco%' OR city2 LIKE '%Oakland%' OR city2
LIKE '%San Jose%' GROUP BY fare, city2 ORDER BY city2")

    q_10_main = rbind(q_10_1, q_10_2)

        # NOTE: Both San Jose and Oakland were not found in either city1 or city2 for table_1a

        sqldf("SELECT city1, city2 FROM table_1a WHERE city1 LIKE '%Oakland%' OR city2 LIKE
'%Oakland%'")

    sqldf("SELECT city1, city2 FROM table_1a WHERE city1 LIKE '%San Jose%' OR city2 LIKE '%San Jose%'")

        # Fares differeing between airports.

        fare_differences = aggregate(fare ~ q10_CA_cities, q_10_main, mean)

       # mean of fares shows that Sac is cheaper on average by about $10

        distance_differences = aggregate(miles ~ q10_CA_cities, q_10_main, mean)

       # mean distances in airports: even though Sac has slightly higher mean, below will show SF has most
long distance connections.

    long_dist_connect = sqldf("SELECT q10_CA_cities, Others, miles, count(miles) as Num_Connections
FROM q_10_main WHERE miles = 2704 ORDER BY miles DESC")

       # shows that SF has the most longest distance connections from SF to Boston MA at 2704 miles
occuring 755 times.

    long_dist_connect_1 = sqldf("SELECT q10_CA_cities, Others, miles, count(miles) as Num_Connections
FROM q_10_main WHERE q10_CA_cities LIKE '%Sacramento%' ORDER BY miles DESC")

       # shows Sac's distance connections from Sac to NY at 2553 miles with 2341 connections

        # Previous (first qustions) analysis but depending upon years now:

        fare_differences_1 = aggregate(fare ~ q10_CA_cities + year, q_10_main, mean)

    sac = sqldf("SELECT * FROM fare_differences_1 WHERE q10_CA_cities LIKE '%Sacramento%'")

    sf = sqldf("SELECT * FROM fare_differences_1 WHERE q10_CA_cities LIKE '%San Fran%'")

         # plot of sac data: fare vs year

    sac_plot = ggplot(data = sac, aes(x = year, y = fare)) +

      geom_point(shape = 1) + geom_smooth() +

      labs(title = "Sacramento Plot (table 1a): Fares ($) vs. Time (years)", x = "Year", y = "Airfares ($)") +
```

```
    theme_few() + theme(plot.title = element_text(size = rel(1.49))) +
    theme(axis.title.y = element_text(size = rel(1.77))) +
    theme(axis.text.x = element_text(size = rel(1.75))) +
    theme(axis.title.x = element_text(size = rel(1.85))) +
    theme(axis.text.y = element_text(size = rel(1.35))) +
    theme(axis.ticks.length=unit(0.25,"cm")) +
    theme(axis.ticks = element_line(size = rel(2))) +
    theme(plot.margin=unit(c(1,1,1.5,1.2),"cm")) +
    theme(strip.text.x = element_text(size = 22, colour = "red", angle = 22.5))
  sac_plot

   # plot of sf data: fare vs year
sf_plot = ggplot(data = sf, aes(x = year, y = fare)) +
 geom_point(shape = 1) + geom_smooth() +
 labs(title = "San Francisco Plot (table 1a): Fares ($) vs. Time (years)", x = "Year", y = "Airfares ($)") +
    theme_few() + theme(plot.title = element_text(size = rel(1.49))) +
    theme(axis.title.y = element_text(size = rel(1.77))) +
    theme(axis.text.x = element_text(size = rel(1.75))) +
    theme(axis.title.x = element_text(size = rel(1.85))) +
    theme(axis.text.y = element_text(size = rel(1.35))) +
    theme(axis.ticks.length=unit(0.25,"cm")) +
    theme(axis.ticks = element_line(size = rel(2))) +
    theme(plot.margin=unit(c(1,1,1.5,1.2),"cm")) +
    theme(strip.text.x = element_text(size = 22, colour = "red", angle = 22.5))
  sf_plot
```