**Homework Assignment 2**

**STA 141A A02**

**Submitted By:**

**Question1:** Unzip and load the airfare dataset. Convert the columns to appropriate data types, then separate table

1a and 6 into different variables (to help you avoid double counting). You don’t need to write an answer

for this question, but please mark the code for this question in the appendix.

**Solution 1:**

**a)** From the column name "table" in the dataset airfare provided I extracted the entries corresponding to "1a" and "6"

and stored them in the table1a and table6 respectively. This can be accomplished by using split command on grouping

column "table" in dataset.(Code in hw\_code\_shivang\_soni doc.)

**Question2:**  What time span does the data cover? Do any quarters or years in that span have no data? Check

separately for table 1a and table 6. In addition, check both tables for patterns in the missing values.

**Solution 2:**

**a)**The time span which data covers is 21 yrs from 1996-2017 .Which can be found out using the range command on

years in the table. Furthermore time span which data covers in terms of quarter is from 1-4 which can also be found

using range command on quarters in table.

**b)** Yes there is no data for **Year 2016: Quarter 2**  and **Year 2017: Quarter 2,3,4** for table1a. This can be found

by using table command.

**c)** There is no data for **Year 2017: Quarter 2,3,4** for table6. This can be found by using table command.

**d)** Intable6 there is presence of "NA" values in **Column:airport1,airport2,airport\_id1,airport\_id2.** Furthermore

there is a relation i.e. for **city2=" Los Angeles, CA (Metropolitan Area)"** and also **city1= "Aspen"** there is "NA"

values present for the "lg\_fare", lg\_carrier" fields in both the tables.

**Question 3:** In 2017, which cities have the most connections to other cities? Which have the least? How do these

results compare to 10 years earlier? 20 years earlier? Which cities have increased connectivity the most?

**Solution3**:

**a)**The table below presents the list of cities having highest connections.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| CITY | Atlanta, GA (Metropolitan Area) | Dallas/Fort Worth, TX | Chicago, IL | Boston, MA (Metropolitan Area) | Washington, DC (Metropolitan Area) | Denver, CO |
| Connections | 71 | 70 | 69 | 68 | 67 | 67 |

Table 1: Cities having most connections with other cities

b)

Table 2: Cities having least connections

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| CITY | Allentown/Bethlehem/Easton, PA | Amarillo, TX | Atlantic City, NJ | Asheville, NC | Augusta, GA | Bakersfield, CA |
| Connections | 1 | 1 | 1 | 1 | 1 | 1 |

**c)** Almost the same trend in airport having high and low connections. Below is plot for connectionsOn analysis of this graph it tells that there are more number of cities For city 2 from 1997 to 2007 there are decrease in connection and similarly there is again a decline in connections from 2007 to 2017. Whereas for city 1 to show same trend in connections.

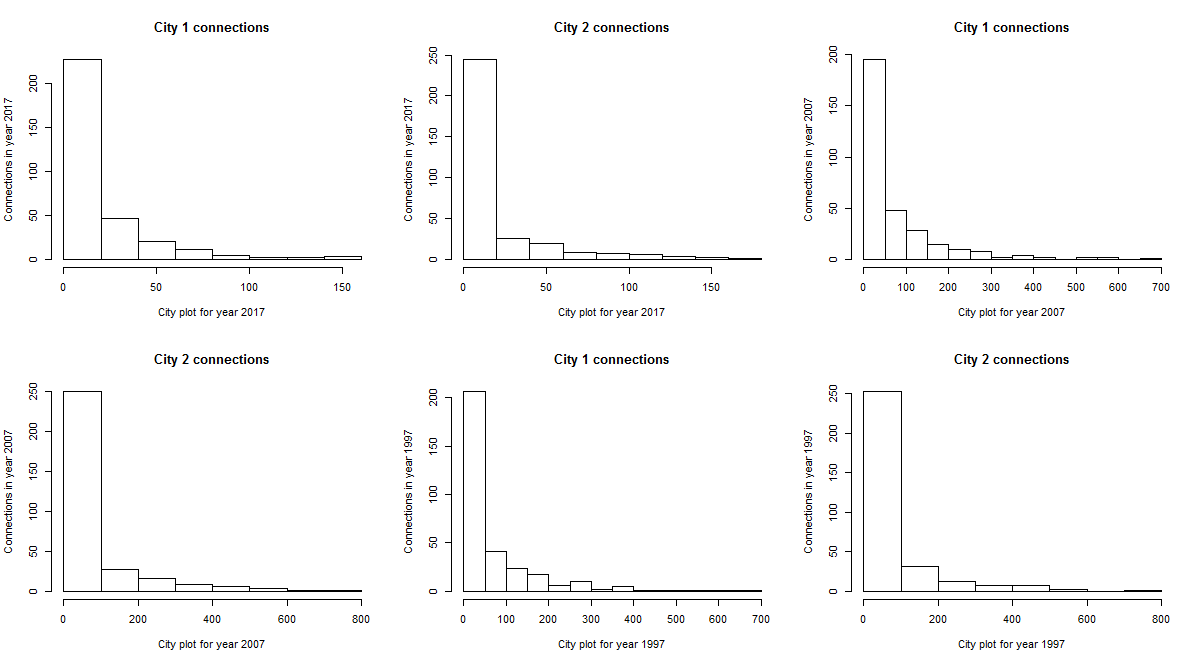
****

Figure 1: Cities list on x axis and connections on y axis

**d)** Cities like Washington has increased its connection Whereas the cities like Detroit has decreased its connections. (Code in hw\_code\_shivang\_soni doc.)

**Question 4**: How has the approximate number of **total** passengers per quarter changed over the years? Create a

graphic to show this and comment on patterns you see. Some quarters have a sharp decline in number

of total passengers. What might explain these?

**Solution 4:**

**a)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Date in Quarter** | 1996-02-15 | 1996-05-15 | 1996-08-15 | 1996-11-15 | 1997-02-15 | 1997-05-15 |
| **Total Passenger** | 165486.2 | 191638.6 | 185492.6 | 187782.1 | 184771.0 | 212122.5 |
| **Date in Quarter** | 1997-08-15 | 1997-11-15 | 1998-02-15 | 1998-05-15 | 1998-08-15 | 1998-11-15 |
| **Total Passenger** | 209034.0 | 206867.8 | 193263.9 | 219692.5 | 212179.9 | 215447.5 |

Table 3:Total Passengers Per Quarter

and so on till 2017...

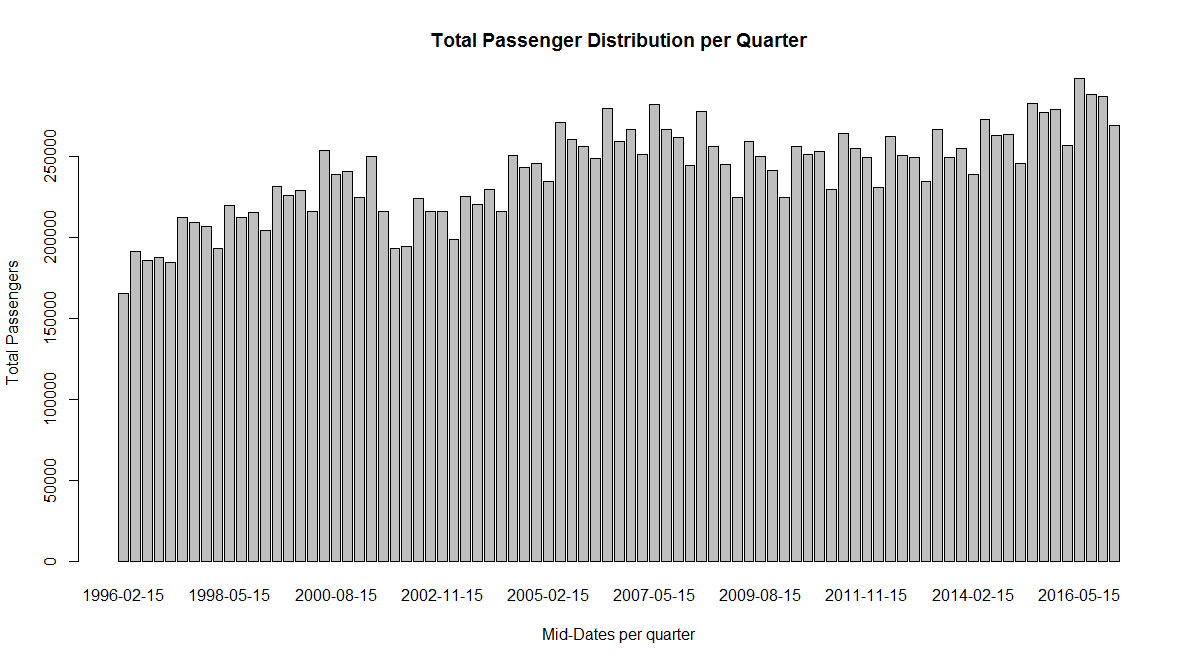
**b)** There is first increasing trend in the airfare then steep decrease and then again increase pattern as seen from the table containing data about total passengers per quarter.****

Figure 2:Barplot for Total passenger per Quarter

**c)** There is a decline in total passenger data from year 2000-2002. Main causes might be increase in airport fare as mean airport fare in 2000 is more as compare to 1999. Other cause for the same might be some sort of disaster etc. like 9/11.

**Question5:** The average fares in the dataset are in nominal dollars (the actual price in dollars at the time). Inflation can confound conclusions based on nominal dollars over time. To deal with this, statisticians convert nominal dollars to real dollars. The conversion formula is explained at the end of this document.

Load the CPI dataset. Create a new column real17\_fare in the table 6 airfare dataset that has the average fare converted to real Q1 2017 dollars. You don’t need to write an answer for this question, but please mark the code for this question in the appendix.

**Solution 5:** After merging the two tables i.e. table6 with cpi table. Calculation for real17\_fare is done based on real Q1 2017 dollar value. Keeping CPI base 2017 as scalar constant a new column is created in newly merged table which hold this value i.e. " real17\_fare ". (Code in hw\_code\_shivang\_soni doc.)

**Question 6:** How have airfares changed over time? Use fares in real Q1 2017 dollars to investigate this graphically.

Comment on patterns you see.

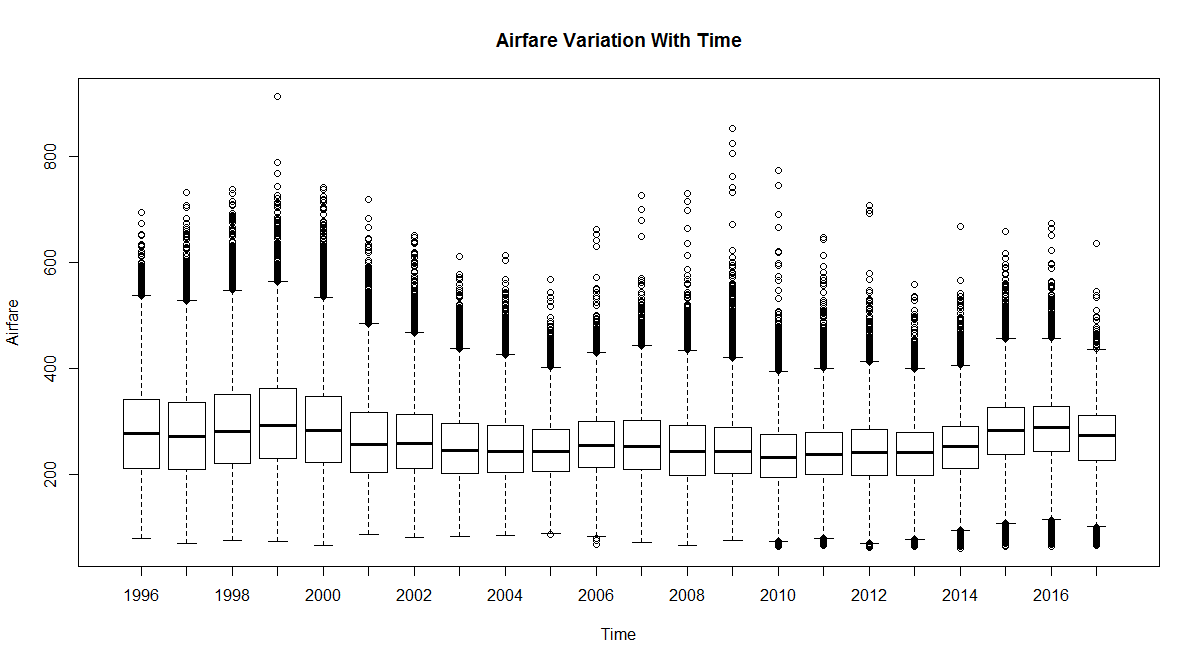
**Solution 6: a) **

Figure 3: Changes in Airfare with time

**b)** The airfares changed over time as seen from the graph is as follows: first increases with time then decreases and again follow same pattern.

The reason for decrease in airfare in 2005 can be accounted as a cause of 9/11. Another reason which accounts for is change in number of passenger travelling.

**Question 7.** For 2015, what is the relationship between fare and distance? Use table 1a to investigate this visually and by using an appropriate statistical model or test. Comment on what you can infer from each and whether there is any disagreement. State the assumptions, use diagnostics to check whether they hold, and comment on how this affects your conclusions. Repeat your analysis with table 6. Comment on differences between your two results and why these occur.

**Solution 7: a)** Initially there is a linear relationship in the graph b/w miles travelled and fare and then the deviation in the linear relationship is due to presence of outlier in table which makes the fitted line in graph deviates from what is normal.

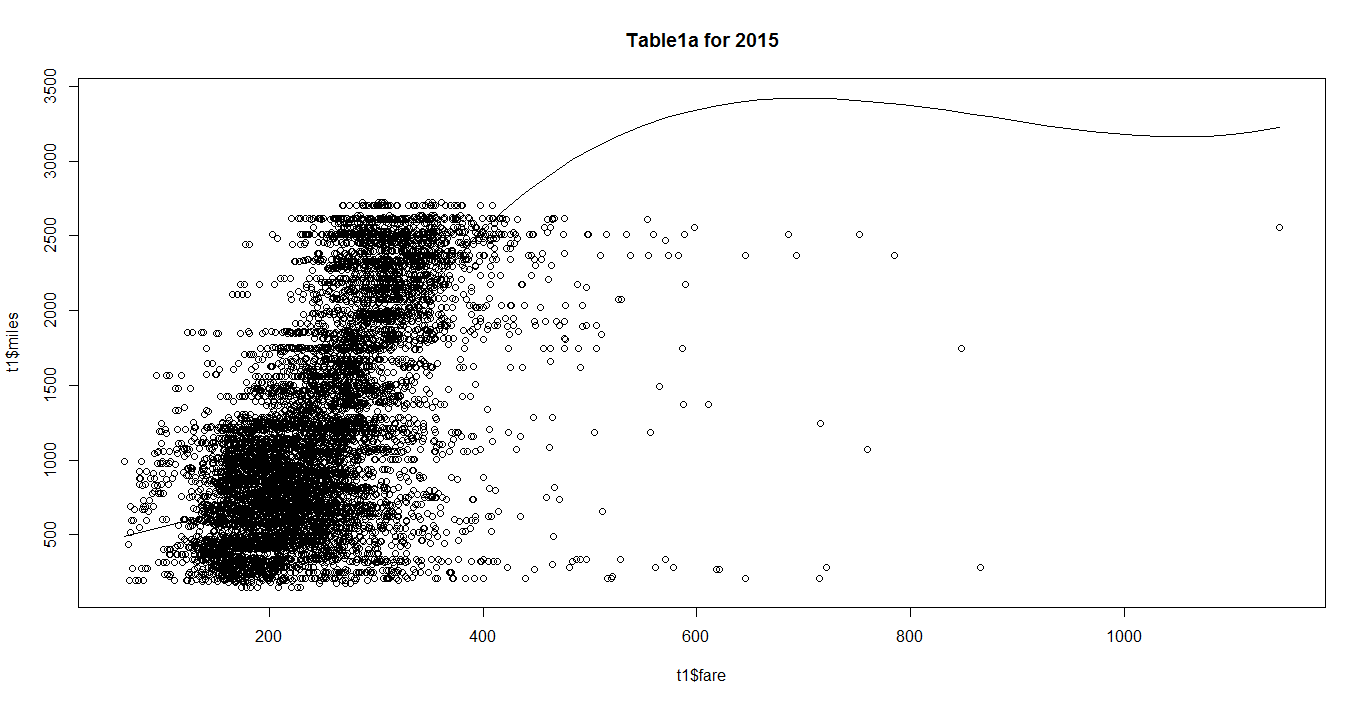
****

Figure 4:Scatter plot b/w distance and fare for table1a in year 2015

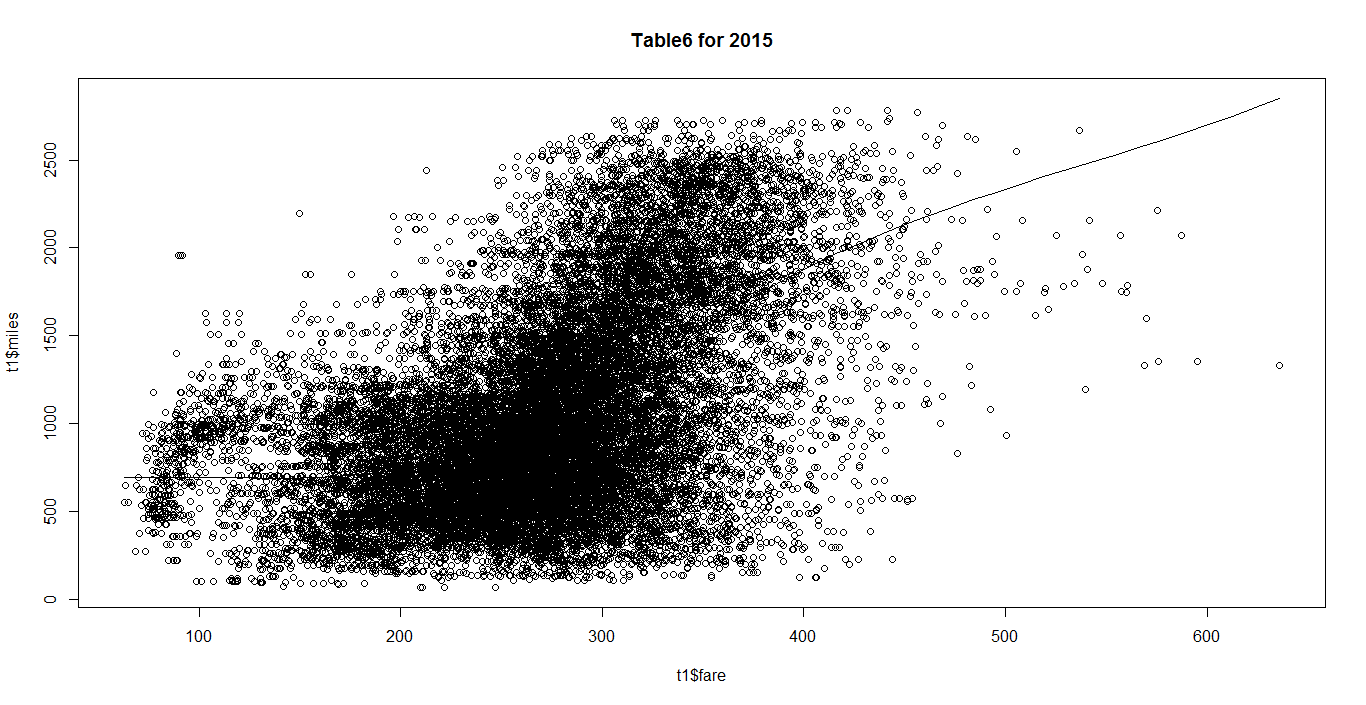
**b)** This shows a linear relationship between miles and fare and no such deviation due to presence of outliers are seen.****

Figure 5:Scatter plot b/w distance and fare for table6 in year 2015

**c)**

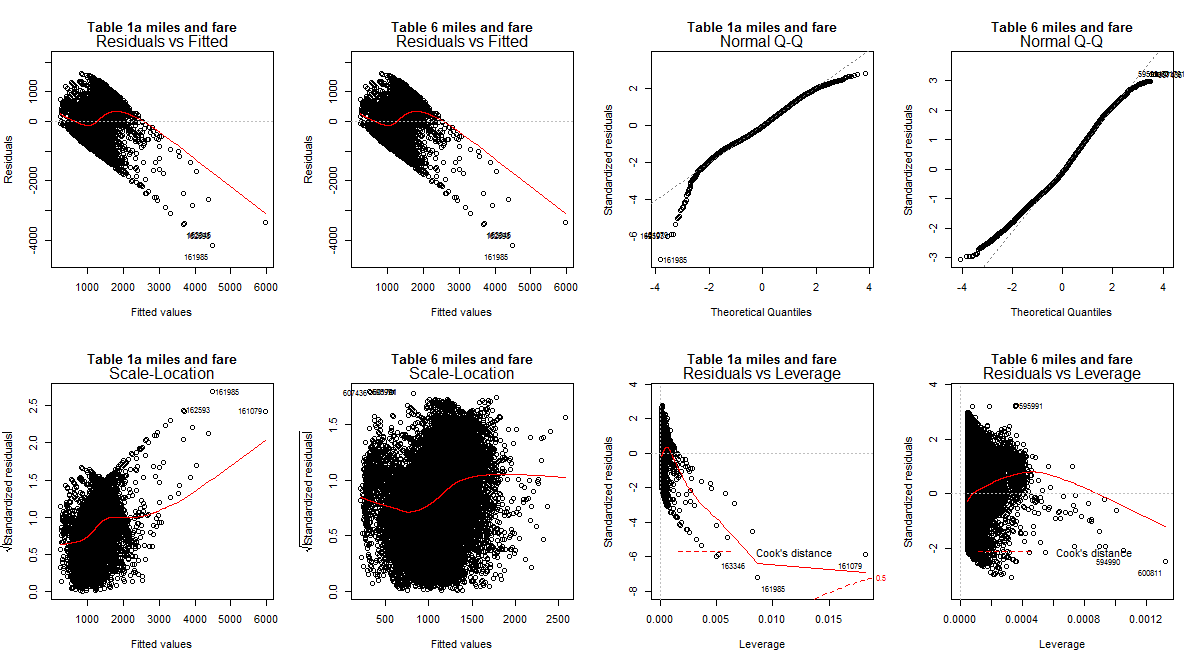


Figure 6:Diagnostics analysis for table1a and table6 data

For the above mentioned graph

**Table1a:** 1)Residuals v/s Fitted values: There is a non linear relationship between the two which indicates that there is presence of non linearity in the dataset.

2.)QQ Plot: The QQ plot for this distribution data shows the normalized trend.

3.)Scale Local: Almost constant variance with some deviation but not exactly linear

4.)Cook's distance: There is an outlier in graph but that can be neglected as having less cook's distance.

On the basis of above conditions we can conclude the relationship between fares and mile is nonlinear.

**Table6:** 1)Residuals v/s Fitted values: There is a non linear relationship between the two which indicates that there is presence of non linearity in the dataset.

2.)QQ Plot: The QQ plot for this distribution is left skewed but most of it data shows the normalized trend.

3.)Scale Local: This is to check constant variance as there is not a linear trend so this graph does not show any constant variance.

.4.) Cook's distance: There is an outlier in graph but that can be neglected as having less cook's distance.

On the basis of above conditions we can conclude the relationship between fares and mile is nonlinear.

**Question 8:** Modify the model or test you used in the previous question to consider average daily passengers in

addition to distance. Recheck the assumptions and then comment on what you can infer about the relationship between fare and average daily passengers. Is there any difference between the results for table 1a and table 6?

**Solution 8:**

**a)**

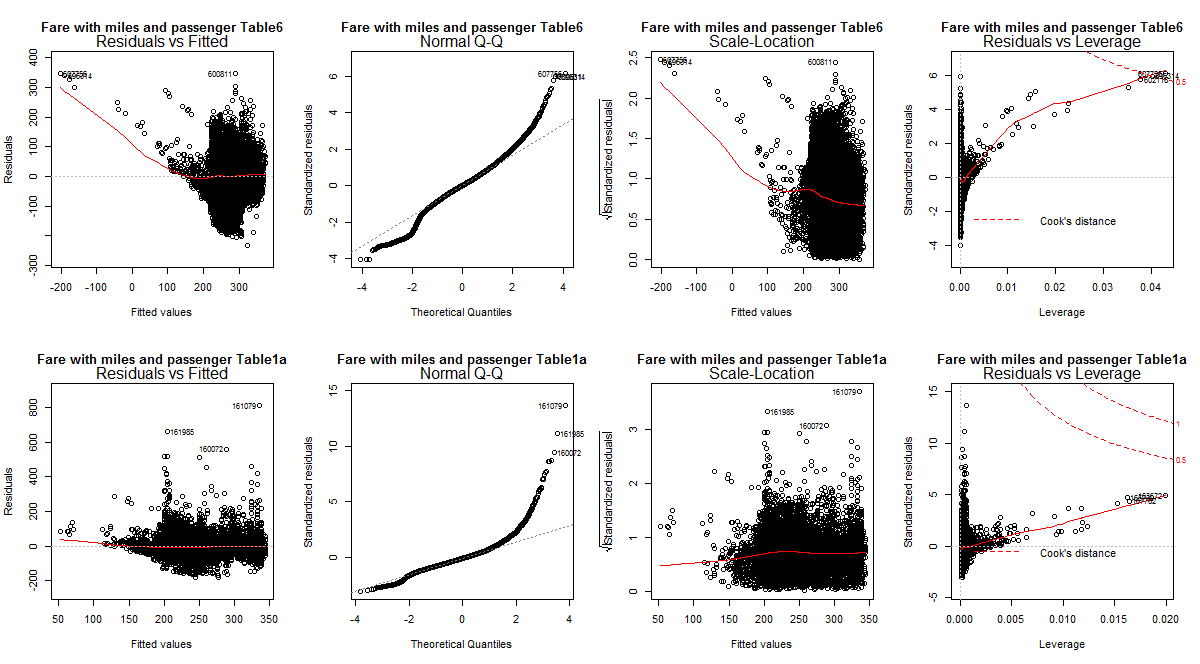
****

Figure 7:Table1a and Table6 Diagnostic analysis

**b)**

For the above mentioned graph

**Table1a:** 1)Residuals v/s Fitted values: There is a non linear relationship between the two which indicates that there is presence of non linearity in the dataset.

2.)QQ Plot: The QQ plot for this distribution data is high tailed but most of it data shows the normalized trend.

3.)Scale Local: Almost constant variance with some deviation but not exactly linear

4.)Cook's distance: Measure of high leverage and error .So outliers in this graph moderate cook's distance.

On the basis of above conditions we can conclude the relationship between fares with mile and also that to passenger is nonlinear..

**Table6:** 1)Residuals v/s Fitted values: There is a non linear relationship between the two which indicates that there is presence of non linearity in the dataset.

2.)QQ Plot: The QQ plot for this distribution is high tailed but most of it data shows the normalized trend.

3.)Scale Local: This is to check constant variance as there is not a linear trend so this graph does not show any constant variance.

4.)Cook's distance: High cook's distance for the outliers.

On the basis of above conditions we can conclude the relationship between fares with mile and also that to passenger is nonlinear.

**c)** There is a difference in distribution for both table1a and table6 as can be inferred from the graph. Based on the graph we can tell both data data does not show same trend in normality etc. Whereas two table have different cook's distance for the outliers.

**Question 9:** For 2015, identify city pairs where the carrier with the largest market share has fares below the average

for that city pair. Investigate these using graphics, statistics, or models (as you see fit). Comment on patterns you find.

**Solution 9:**

**a)**

|  |  |  |
| --- | --- | --- |
| **SNO** | **CITY1** | **CITY2** |
| **1** | Allentown/Bethlehem/Easton, PA | Dallas/Fort Worth, TX |
| **2** | Allentown/Bethlehem/Easton, PA | Denver, CO |
| **3** | Allentown/Bethlehem/Easton, PA | Nashville, TN |
| **4** | Allentown/Bethlehem/Easton, PA | St. Louis, MO |
| **5** | Allentown/Bethlehem/Easton, PA | Jacksonville, FL |

Table 4: City pair with carrier with largest market share below avg fare.

**b)** This has a skewed QQ plot with non constant variance and not linearity in the data.with p value less than 0.05. so that means nonlinear trend between the data.

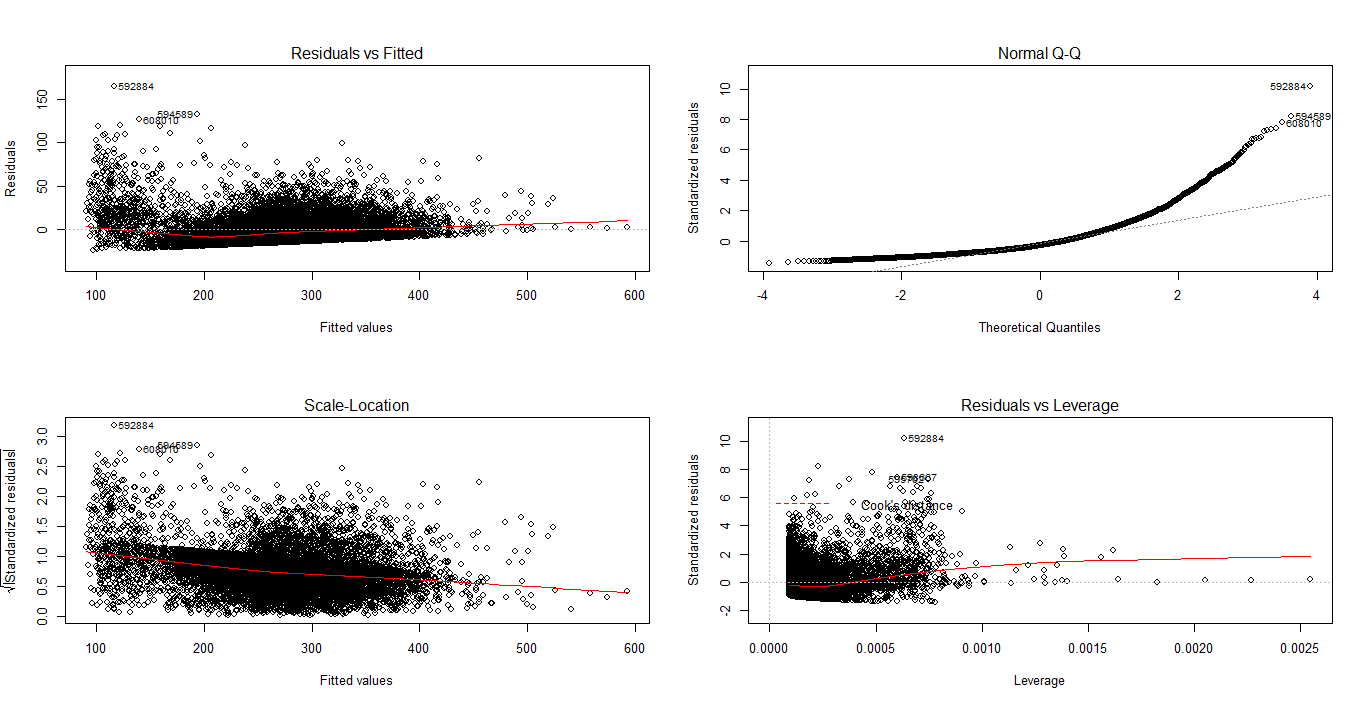
****

Figure 8:Diagnostic analysis for city pair with largest market share below avg fare

**Question 10.** Use table 1a to compare Sacramento (SMF), Oakland (OAK), San Francisco (SFO), and San Jose (SJC). How do fares differ between these airports? Which airport has the most long-distance connections and how does this compare to the others? Do these results differ by year?

**Solution 10.**

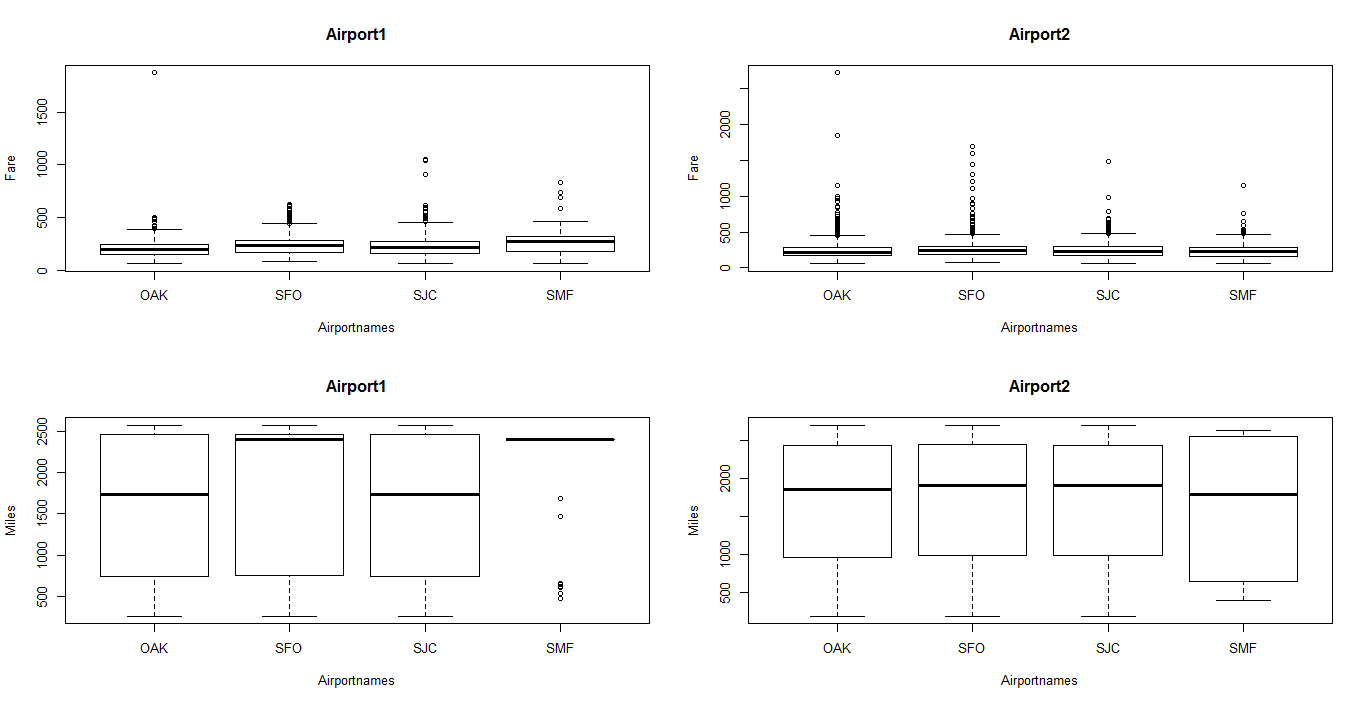
****

Figure 9:Boxplot for how fare and distance b/w the cities differ.

**a)** The graphics shown above indicates how the fares differ b/w the airport so as per the graph **Airport1: SJC** has high fare as compare to other airport. Whereas for **Airport2:SFO** has high fare.

**b)** As shown in the graphics above the most long distance in miles connections are for **Airport1:SMF** and for **Airport2:SJC.**

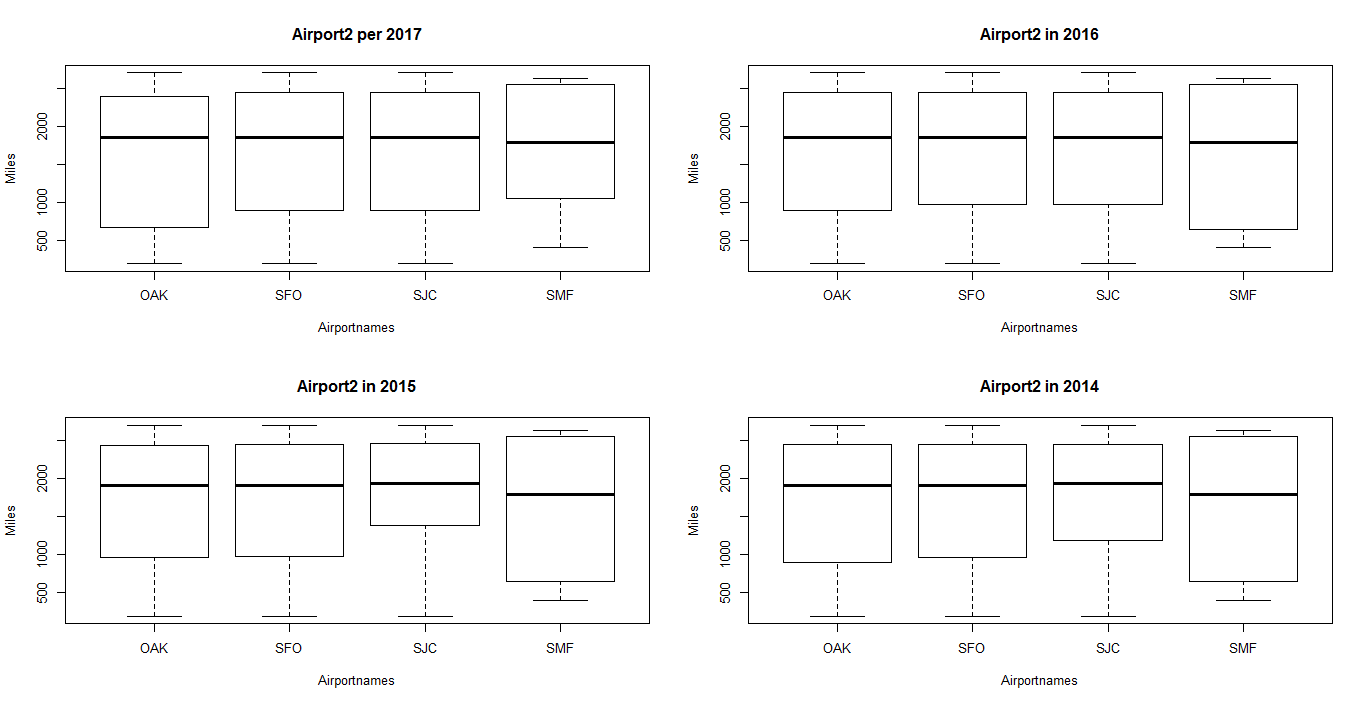
****

Figure 10:How distance vary with time for the city under consideration

**c)** Graphics above indicates how the miles for long distance connections changed over last 4 years. as there is a difference in 1st quartiles for OAK and SMF as seen for year 2017 and 2016.Whereas difference in 1st quartiles of SJC and SMF for year 2015-2014

**References:**

1.) **From Wikipedia**: https://www.bls.gov/cpi/ (CPI related information)

2.) **From Stack overflow**: https://stackoverflow.com/questions/18304672/sapply-in-r-how-to-use (Usage of sapply)

: https://stat.ethz.ch/R-manual/R-devel/library/base/html/merge.html (Merging in R)

3.)**Piazza:** Instructor and T.A's help.