Pinalben Kamleshbhai Patel

Student ID : 0675587

AMOD 5250 - Final Project

Data Analytics with R

**Data Summary :**

The **Public Use Microdata Sample (PUMS)** contains a sample of actual responses to the **American Community Survey (ACS)**. The PUMS dataset includes variables for nearly every question on the survey, as well as many new variables that were derived after the fact from multiple survey responses (such as poverty status). Each record in the file represents a single person, or, in the household-level dataset, a single housing unit. In the person-level file, individuals are organized into households, making possible the study of people within the contexts of their families and other household members. PUMS files for an individual year, such as 2017, contain data on approximately one percent of the United States population. PUMS files covering a five-year period, such as 2013-2017, contain data on approximately five percent of the United States population.

For this analysis , we have limited our scope to only "**Housing Record**" data for United States. There are four files A, B, C and D which have various variables, but we selected 16 variables to perform our analysis.

Variable names and their description which are used for analysis :

* ST - State
* ADJINC - Adjustment Factor
* ACR - Lot Size
* AGS - Agriculture sales
* ELEP - Electricity Cost
* RNTM - Meal Included in Rent
* RNTP - Monthly Rent
* TEN - Tenure
* VALP - Property Value
* VEH - Number of Vehicles
* YBL - Year in which structure was built
* FINCP - Family Income
* HHL - Household Language
* HINCP - Household Income
* HUPAC - Presence of Children
* TAXP - Property Taxes

There are certain relationships I can find in these variables such as we can see how housing units are divided by Tenure and household languages used in the different housing units. How presence of children affects the electricity cost of the housing units ? How Monthly rent of the unit and if Meals are included in rent is related ?

What is the relation between Year the structure was built and the Property value ? Family Income and Household Income have any similarities ? and How Property value and Family income are related ?

From the subsets of datasets, we have prepared four dataframes which are pums\_fileA, pums\_fileB, pums\_fileC and pums\_fileD . By combining all four dataframes into one we get out final dataset which is **Housing.Unit.Survey** ,which we will use for the analysis .

Using “colnames” function , we have changed the variable names into more descriptive column names. By using “factor” , we have labeled some integer values such as for State 1 is AL for Alabama . For Meal Included in Rent Yes for 1 and No for 2 and so on.

**Table 1 : Summary of the data and its distribution**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Min** | **1st Qu** | **Median** | **Mean** | **3rd Qu** | **Max** |
| **State** | 1 | 13.5 | 27 | 26 | 38.25 | 48 |
| **Lot.Size** | 1 | 1 | 1 | 1.3 | 1 | 3 |
| **Agriculture.Sales** | 1 | 1 | 1 | 1 | 1 | 6 |
| **Electricity.Cost** | 1 | 70 | 120 | 138.4 | 190 | 650 |
| **Monthly.Rent** | 4 | 520 | 800 | 939.7 | 1200 | 3900 |
| **Tenure** | 1 | 1 | 2 | 1.9 | 3 | 4 |
| **Property.Value** | 100 | 96000 | 180000 | 285679 | 334000 | 6308000 |
| **No.Of.Vehicles** | 0 | 1 | 2 | 1.8 | 2 | 6 |
| **Family.Income** | -21500 | 38500 | 70000 | 94895 | 117000 | 3164000 |
| **Household.Language** | 1 | 1 | 1 | 1.4 | 1 | 5 |
| **Household.Income** | -21500 | 28300 | 57000 | 80918 | 101000 | 3164000 |
| **Children.Present** | 1 | 3 | 4 | 3.4 | 4 | 4 |
| **Taxes** | 1 | 18 | 32 | 34.7 | 52 | 68 |
| **Year.Property.Built** | 1 | 3 | 5 | 5.1 | 7 | 21 |

We have used ADJINC variable which is an adjustment factor for housing dollar amounts. By dividing ADJINC by 1,000,000 , we can obtain the inflation adjustment factor and multiplying it to the PUMS variable value , we can adjust it to 2017 dollars. Variables requiring ADJINC on the Housing Unit file are FINCP and HINCP.

ADJINC – Adjustment.Factor, FINCP – Family.Income, HINCP – Household.Income

**Findings :**

In the first section, I talked about some relationships which I intended to analyze here . Here, in this section , I have provided my analysis and findings for those relationships .

* **Distribution of housing units by Tenure**

If we look at the PUMS Data Dictionary we will find the description of numbers here :

**Tenure**

b .N/A (GQ/vacant)

1 .Owned with mortgage or loan (include home equity loans)

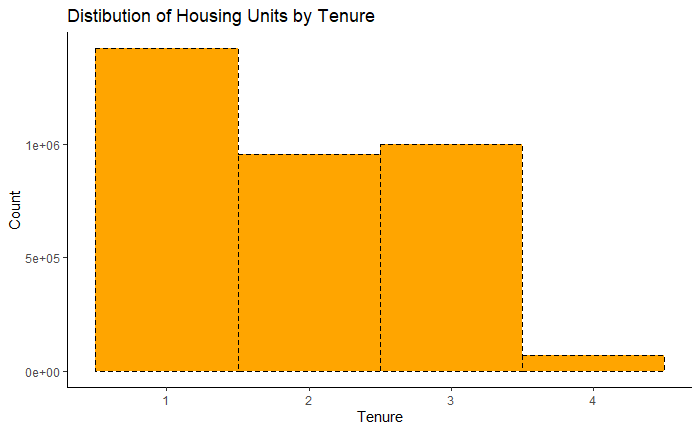
2 .Owned free and clear

3 .Rented

4 .Occupied without payment of rent

Here, we have used two kinds of graphs to explain how Housing Units are distributed by Tenure.

**Graph A :**

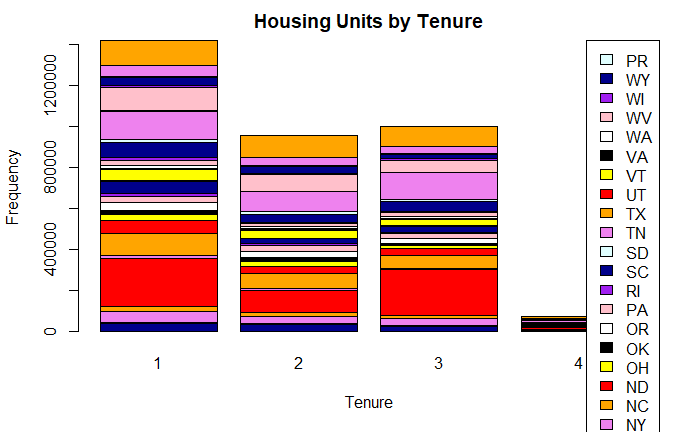


From Graph A , it is evident that Housing Units are mostly “Owned with mortgage or loan” and there are smaller number of housing units which are “Occupied without payment of rent” . However, there is no much difference between housing units which are “Owned Free and Clear” and “Rented” .

We did the same analysis by using Tenure and State and using Bar chart and result is as below :

**Graph B :**

We can see that the results are same as Graph B , however we can also derive that UT(Utah) , ND(North Dakota) ,NY(New York) and TN(Tennessee) are the states where people are having equal status for “Owned housing units “ and “ Rented housing Units” .



* **Distribution of housing units by household languages**

If we look at the PUMS Data Dictionary we will find the description of numbers here :

**Household language**

b .N/A (GQ/vacant)

1 .English only

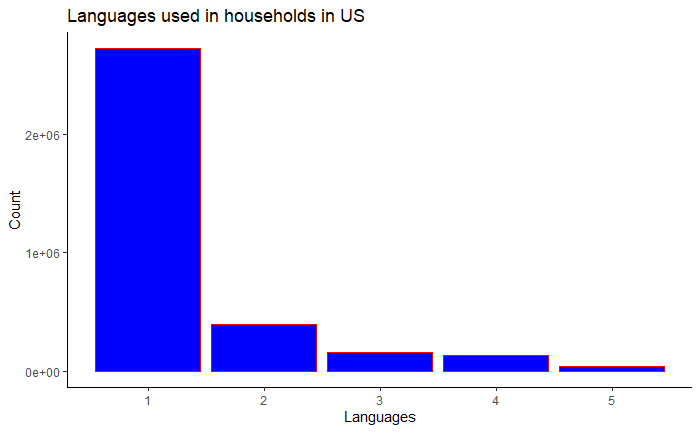
2 .Spanish

3 .Other Indo-European languages

4 .Asian and Pacific Island languages

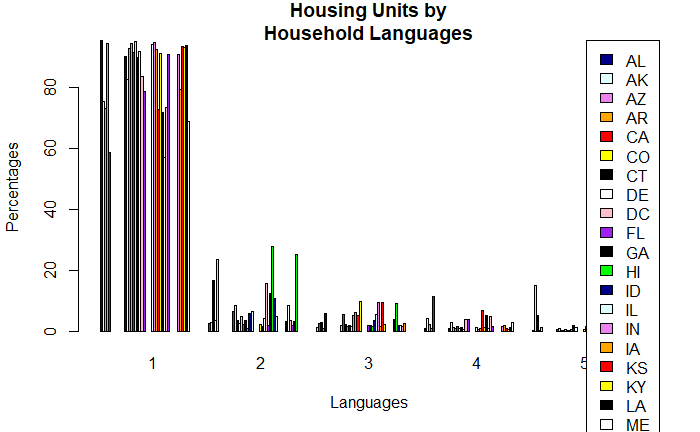
5 .Other language

**Graph C :**



From the above Graph C, it is evident that if we look at overall United states , English is the most used language in the households and Spanish takes the second place . However, state wise distribution of household languages in United States can be represented as below:

**Graph D :**



* **Value of Property is affected by in which era they have built**

If we look at the PUMS Data Dictionary we will find the description here :

When structure first built 01 .1939 or earlier

02 .1940 to 1949

03 .1950 to 1959

04 .1960 to 1969

05 .1970 to 1979

06 .1980 to 1989

07 .1990 to 1999

08 .2000 to 2004

09 .2005

10 .2006

11 .2007

12 .2008

13 .2009

14 .2010

15 .2011

16 .2012

17 .2013

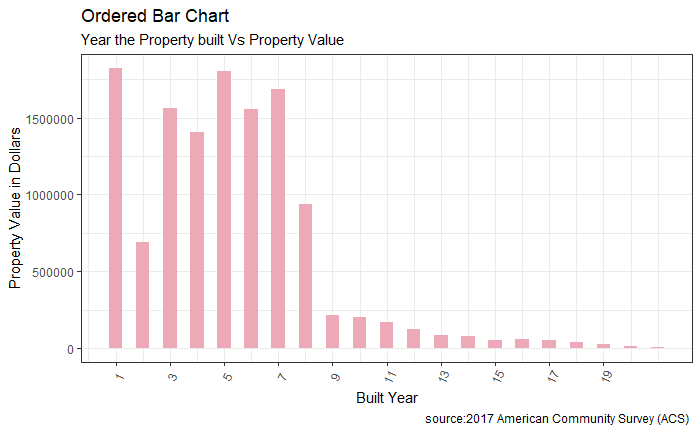
18 .2014

19 .2015

20 .2016

21 .2017

**Graph E :**



From the below graph E, we can identify that Property values are getting decreased over time , when the structure was first built it has the highest values and then values are decreasing over time . It means we can conclude that old properties now have least values .

* **Electricity cost affected by Presence of Children**

If we look at the PUMS Data Dictionary we will find the description here :

HH presence and age of children

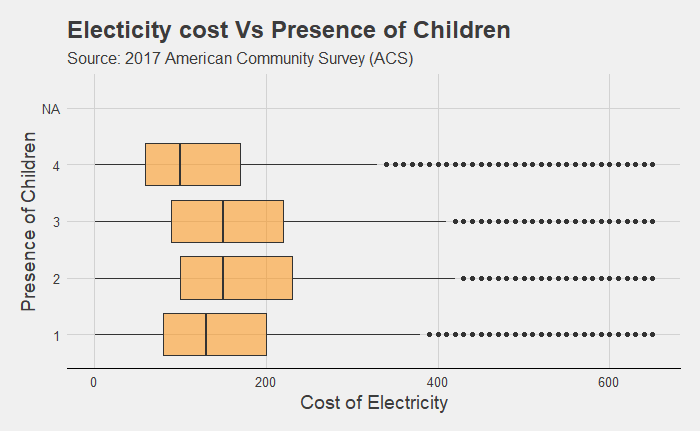
1 .With children under 6 years only

2 .With children 6 to 17 years only

3 .With children under 6 years and 6 to 17 years

4 .No children

**Graph F :**

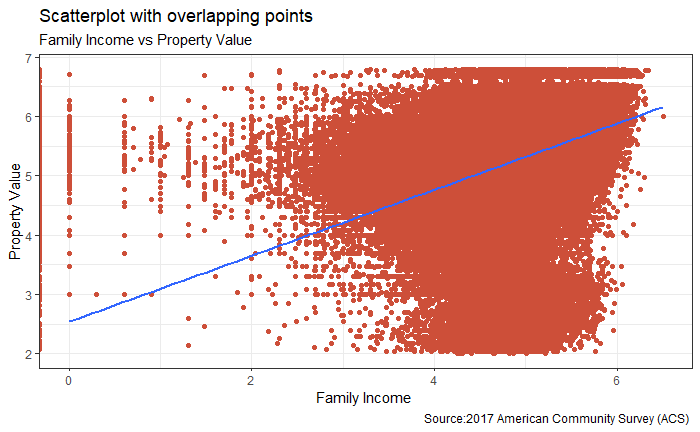


Above Graph F shows the affect of presence of children on Electricity cost of housing units. It is evident that presence of children in the housing unit can be attributed to an increase in electricity consumption . However, Age group of the children does not influence the electricity consumption charges .

* **Relationship between Family Income and Property Value**

Here, both variables are quantitative variables property value and Family income. I have used Scatterplot with smooth line plot to describe how these two are related.

**Graph G :**



We can observe from the above plot that the highest family income group is prevalent to high property value . We can see the straight smooth vertical line which is going upward showing linear regression between both the values.

In this plot, there are outliers also , which suggests that there can be some family groups where their family income is not that high, but property value is .

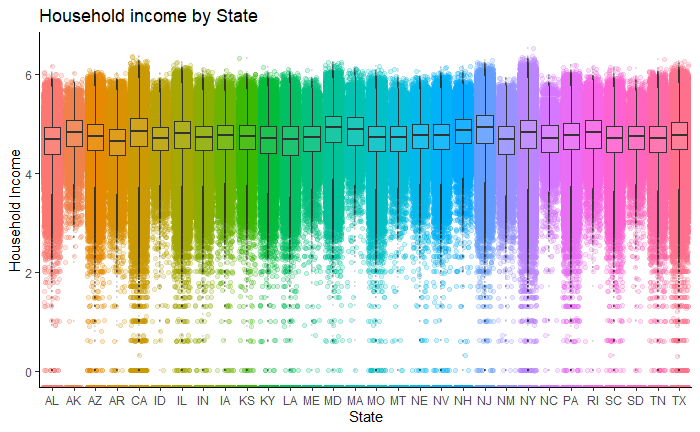
* **Household Income by State and Family Income by State**

Here , we are analyzing the difference between Household Income and Family Income by providing two different graphs for state wise incomes . Look at the results :

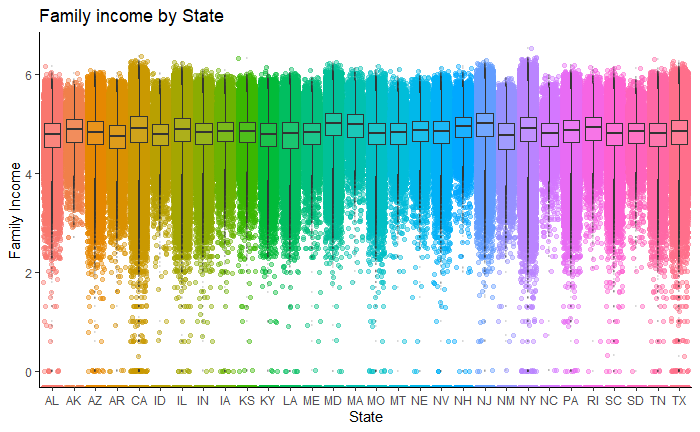
From Graph H and I , it is evident that the Family Income of Housing Units State wise and Household Income of Housing Unit State wise doesn’t have much differences.

The same can be seen from the summary of these two variables from the table 1. Their Mean and Median values don’t have much difference as well . So, we can conclude in United states , people are earning in families as much as they are earning while they are living with different people. Their average income is same.

**Graph H :**



**Graph I :**



* **Density Plot for Sales of agriculture products**

The density (the probability of a particular value) for the normal distribution is calculated using dnorm. While it is technically mathematically impossible to find the exact probability of a number from a continuous distribution, this is an estimate of the probability. Like with rnorm, a mean and standard deviation can be specified for dnorm.

We have used these functions and prepared the below density plot for sale of agriculture products.

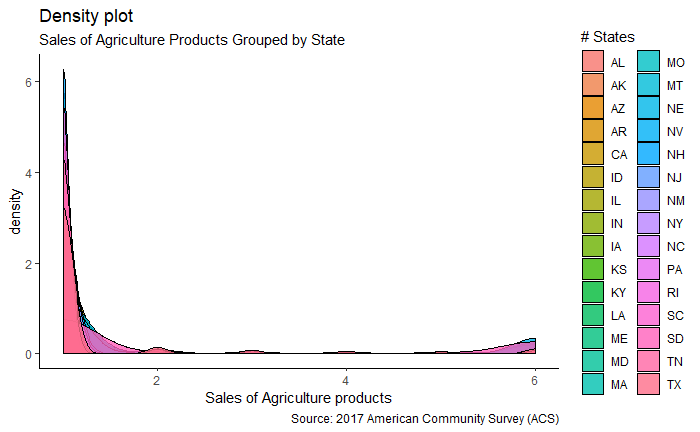
From the Graph J , we can say that around 4 % of values have Zero sales .

**Graph J :**



**Graph K :**

If we look at the Graph K then it is evident that sales of agriculture products grouped by state has better results. We can observe that States such as TX,TN,PA,SD,SC,NY have better number of sales of agriculture products .



* **Meals Included in Monthly rent affects Rent**

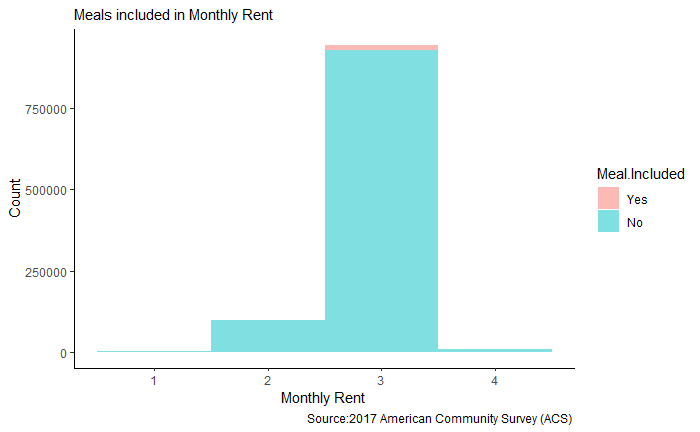
We have performed T-test and checked whether if meals are included in the monthly rent then rent is high or not .

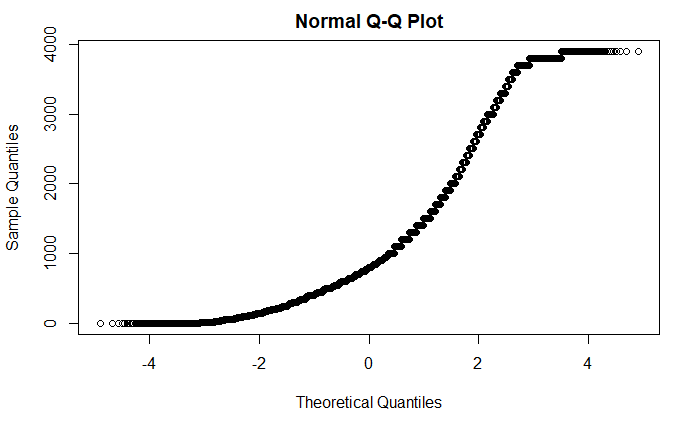
Null Hypothesis : Rent is not higher if meals are included in the rent

From the t-test results , we can see that p-value is less than 0.05 so, we will reject the null hypothesis and conclude that rent is higher if meals are included in it.

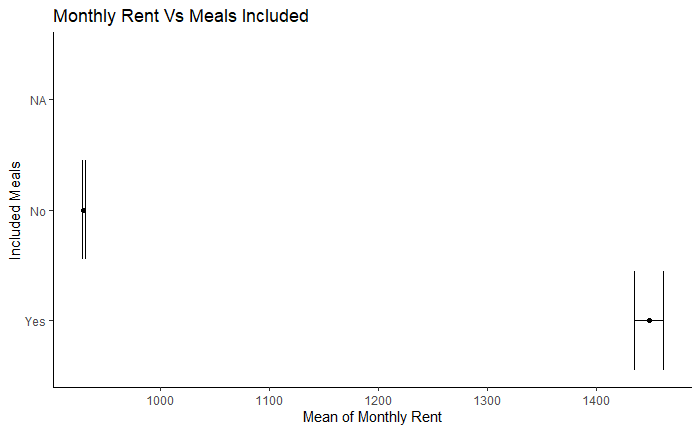
We can check the same thing with the graphs provided below :

**Graph L :**





**Graph M :**



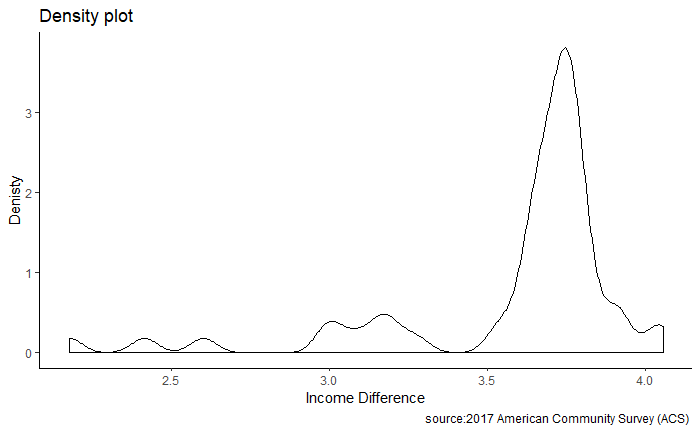
* **Income Difference between Household Income and Family Income**

**Graph N :**

We have performed two sample paired T-tests for family income and household income.

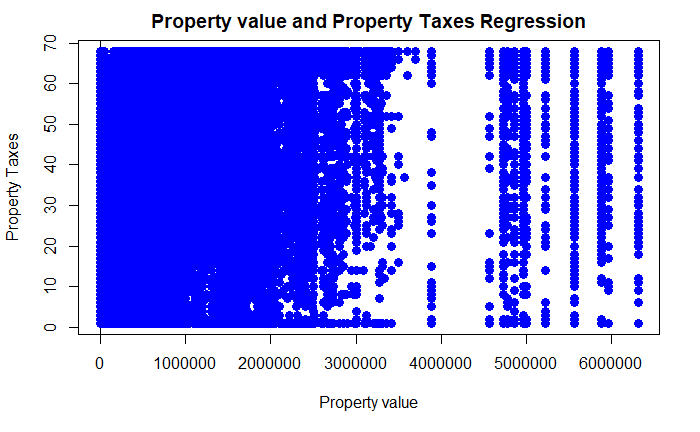
Null hypothesis : Both are not same

Form the t-test results, it is evident that p-value is less than 0.05 means we can reject the null hypothesis and conclude that both income are same.

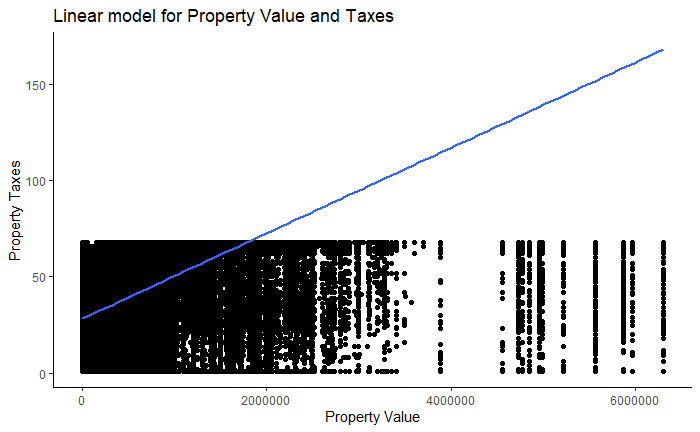


* **Linear Regression model for Property Value and Taxes**

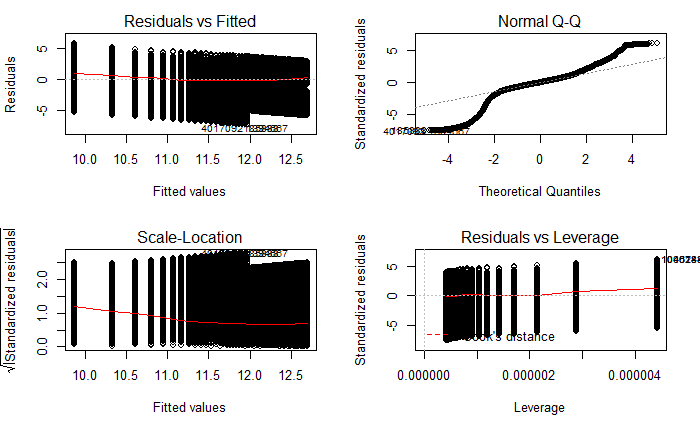
**Graph O :**



**Graph P:**



**Graph Q:**



**Discussion :**

After completing various statistical analysis on the data, I come to the conclusion that for different relationships which I have identified earlier I found some results. Such as I realized that Family Income and Household Income are not related but they are almost same for Housing units in United states.

For distribution of housing units by Tenure and household languages , we found out that Most housing units are owned with loan or Rented in United states and English and Spanish are the most used languages in States.

Results have also shown that Value of the property is decreasing by time . If some house is built in the 1900s then its value was highest that time but now by time it is decreasing.

We also realized that the number of children affects the electricity cost, but age group of children doesn't matter.

Family Income and Property values were also related positively . Such as If income is higher then property value is also high.

We are also able to predict the sale of agriculture products for a year for housing units.

If meals are included in the rent then monthly rent of that Housing unit is also increasing that also we have seen .

We also found linear regression between Property Value and Property Taxes. However, there are some limitations to my analysis as I have avoided all the missing values in my analysis . I think that could differ the actual results. I am not 100 % sure of my analysis there could be some loopholes.

There is room for a lot of future research, with this data. I only analyze a little bit of it . There is so much more we can do to study this data set.