Housing Project

pinal

12/17/2019

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

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

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
##   ST  ADJINC ACR AGS ELEP RNTM RNTP TEN  VALP VEH YBL  FINCP HHL  HINCP HUPAC
## 1  1 1061971  NA  NA   NA   NA   NA  NA    NA  NA  NA     NA  NA     NA    NA
## 2  1 1061971   1  NA  350   NA   NA   2 25000   3   2 151000   1 151000     4
## 3  1 1061971   1  NA  300   NA   NA   1 80000   1   5     NA   1  39930     4
## 4  1 1061971   3   1  220    2  100   3    NA   2   6  11400   1  11400     2
## 5  1 1061971  NA  NA   60    2   80   3    NA   0   5     NA   1   3900     4
##   TAXP
## 1   NA
## 2    3
## 3    6
## 4   NA
## 5   NA
```

```
##   ST  ADJINC ACR AGS ELEP RNTM RNTP TEN   VALP VEH YBL FINCP HHL HINCP HUPAC
## 1 16 1061971   2   1   50   NA   NA   4     NA   1   9    NA   1 10800     4
## 2 16 1061971  NA  NA    1    2  600   3     NA   2   2  3500   1  3500     1
## 3 16 1061971   2   1  140   NA   NA   2 250000   2   4 92900   1 92900     4
## 4 16 1061971  NA  NA   80    2  420   3     NA   0   6    NA   1 30020     4
## 5 16 1061971   1  NA  160   NA   NA   1 140000   3   3 45000   1 45000     1
##   TAXP
## 1   NA
## 2   NA
## 3   17
## 4   NA
## 5   18
```

```
##   ST  ADJINC ACR AGS ELEP RNTM RNTP TEN   VALP VEH YBL  FINCP HHL  HINCP HUPAC
## 1 29 1061971  NA  NA   80    2  720   3     NA   2   5     NA   1  37100     4
## 2 29 1061971  NA  NA   NA   NA   NA  NA     NA  NA  NA     NA  NA     NA    NA
## 3 29 1061971   2   1  240   NA   NA   1 175000   2   5  76000   1  76000     4
## 4 29 1061971   1  NA   90   NA   NA   1 200000   2   7 119850   1 119850     2
## 5 29 1061971  NA  NA    1    2  340   3     NA   1   5     NA   1   7850     4
##   TAXP
## 1   NA
## 2   NA
## 3   29
## 4   48
## 5   NA
```

```
##   ST  ADJINC ACR AGS ELEP RNTM RNTP TEN   VALP VEH YBL  FINCP HHL  HINCP HUPAC
## 1 42 1061971   2  NA   NA   NA   NA  NA     NA  NA   1     NA  NA     NA    NA
## 2 42 1061971   1  NA   70   NA   NA   4     NA   0   1     NA   1   7300     4
## 3 42 1061971   3   1  140   NA   NA   1  95000   1   1     NA   1  11000     4
## 4 42 1061971   2   1   90   NA   NA   2 900000   2   3 111600   1 111600     4
## 5 42 1061971   1  NA   NA   NA   NA  NA 240000  NA   3     NA  NA     NA    NA
##   TAXP
## 1   NA
## 2   NA
## 3   40
## 4   64
## 5   NA
```

```
##   ST  ADJINC ACR AGS ELEP RNTM RNTP TEN  VALP VEH YBL  FINCP HHL  HINCP HUPAC
## 1  1 1061971  NA  NA   NA   NA   NA  NA    NA  NA  NA     NA  NA     NA    NA
## 2  1 1061971   1  NA  350   NA   NA   2 25000   3   2 151000   1 151000     4
## 3  1 1061971   1  NA  300   NA   NA   1 80000   1   5     NA   1  39930     4
## 4  1 1061971   3   1  220    2  100   3    NA   2   6  11400   1  11400     2
## 5  1 1061971  NA  NA   60    2   80   3    NA   0   5     NA   1   3900     4
##   TAXP
## 1   NA
## 2    3
## 3    6
## 4   NA
## 5   NA
```

Modifying dataset

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.

```
##    State Adjustment.Factor Lot.Size Agriculture.Sales Electricity.Cost
## 1      1           1061971       NA                NA               NA
## 2      1           1061971        1                NA              350
## 3      1           1061971        1                NA              300
## 4      1           1061971        3                 1              220
## 5      1           1061971       NA                NA               60
## 6      1           1061971        1                NA              100
## 7      1           1061971        1                NA              240
## 8      1           1061971        2                 1              130
## 9      1           1061971       NA                NA              130
## 10     1           1061971        2                 1               80
##    Meal.Included Monthly.Rent Tenure Property.Value No.Of.Vehicles
## 1             NA           NA     NA             NA             NA
## 2             NA           NA      2          25000              3
## 3             NA           NA      1          80000              1
## 4              2          100      3             NA              2
## 5              2           80      3             NA              0
## 6             NA           NA      1          18000              1
## 7             NA           NA      1         390000              3
## 8             NA           NA      2         120000              0
## 9              2          340      3             NA              1
## 10            NA           NA      1         160000              6
##    Year.Property.Built Family.Income Household.Language Household.Income
## 1                   NA            NA                 NA              NA
## 2                    2        151000                  1          151000
## 3                    5            NA                  1           39930
## 4                    6         11400                  1           11400
## 5                    5            NA                  1            3900
## 6                    2            NA                  1            5400
## 7                    8        136000                  1          136000
## 8                    2         52600                  1           52600
## 9                   10            NA                  1          103000
## 10                   4         81600                  1           81600
##    Children.Present Taxes
## 1                NA    NA
## 2                 4     3
## 3                 4     6
## 4                 2    NA
## 5                 4    NA
## 6                 4     3
## 7                 4    26
## 8                 4     5
## 9                 4    NA
## 10                4    10
```

```
##      State       Adjustment.Factor    Lot.Size       Agriculture.Sales
##  Min.   : 1.00   Min.   :1011189   Min.   :1.0       Min.   :1
##  1st Qu.:13.50   1st Qu.:1029257   1st Qu.:1.0       1st Qu.:1
##  Median :27.00   Median :1035988   Median :1.0       Median :1
##  Mean   :26.00   Mean   :1038179   Mean   :1.3       Mean   :1
##  3rd Qu.:38.25   3rd Qu.:1045195   3rd Qu.:1.0       3rd Qu.:1
##  Max.   :48.00   Max.   :1061971   Max.   :3.0       Max.   :6
##                                    NA's   :1249777   NA's   :3567841
##  Electricity.Cost Meal.Included      Monthly.Rent         Tenure
##  Min.   :  1.0    Min.   :1         Min.   :   4.0    Min.   :1.0
##  1st Qu.: 70.0    1st Qu.:2         1st Qu.: 520.0    1st Qu.:1.0
##  Median :120.0    Median :2         Median : 800.0    Median :2.0
##  Mean   :138.4    Mean   :2         Mean   : 939.7    Mean   :1.9
##  3rd Qu.:190.0    3rd Qu.:2         3rd Qu.:1200.0    3rd Qu.:3.0
##  Max.   :650.0    Max.   :2         Max.   :3900.0    Max.   :4.0
##  NA's   :745527   NA's   :3136435   NA's   :3136435   NA's   :745527
##  Property.Value    No.Of.Vehicles   Year.Property.Built Family.Income
##  Min.   :    100   Min.   :0.0      Min.   : 1.0        Min.   : -21500
##  1st Qu.:  96000   1st Qu.:1.0      1st Qu.: 3.0        1st Qu.:  38500
##  Median : 180000   Median :2.0      Median : 5.0        Median :  70000
##  Mean   : 285679   Mean   :1.8      Mean   : 5.1        Mean   :  94895
##  3rd Qu.: 334000   3rd Qu.:2.0      3rd Qu.: 7.0        3rd Qu.: 117000
##  Max.   :6308000   Max.   :6.0      Max.   :21.0        Max.   :3164000
##  NA's   :1782533   NA's   :745527   NA's   :415249      NA's   :1885717
##  Household.Language Household.Income  Children.Present     Taxes
##  Min.   :1.0        Min.   : -21500   Min.   :1.0      Min.   : 1.0
##  1st Qu.:1.0        1st Qu.:  28300   1st Qu.:3.0      1st Qu.:18.0
##  Median :1.0        Median :  57000   Median :4.0      Median :32.0
##  Mean   :1.4        Mean   :  80918   Mean   :3.4      Mean   :34.7
##  3rd Qu.:1.0        3rd Qu.: 101000   3rd Qu.:4.0      3rd Qu.:52.0
##  Max.   :5.0        Max.   :3164000   Max.   :4.0      Max.   :68.0
##  NA's   :745527     NA's   :745527    NA's   :745527   NA's   :1817340
```

A. 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)

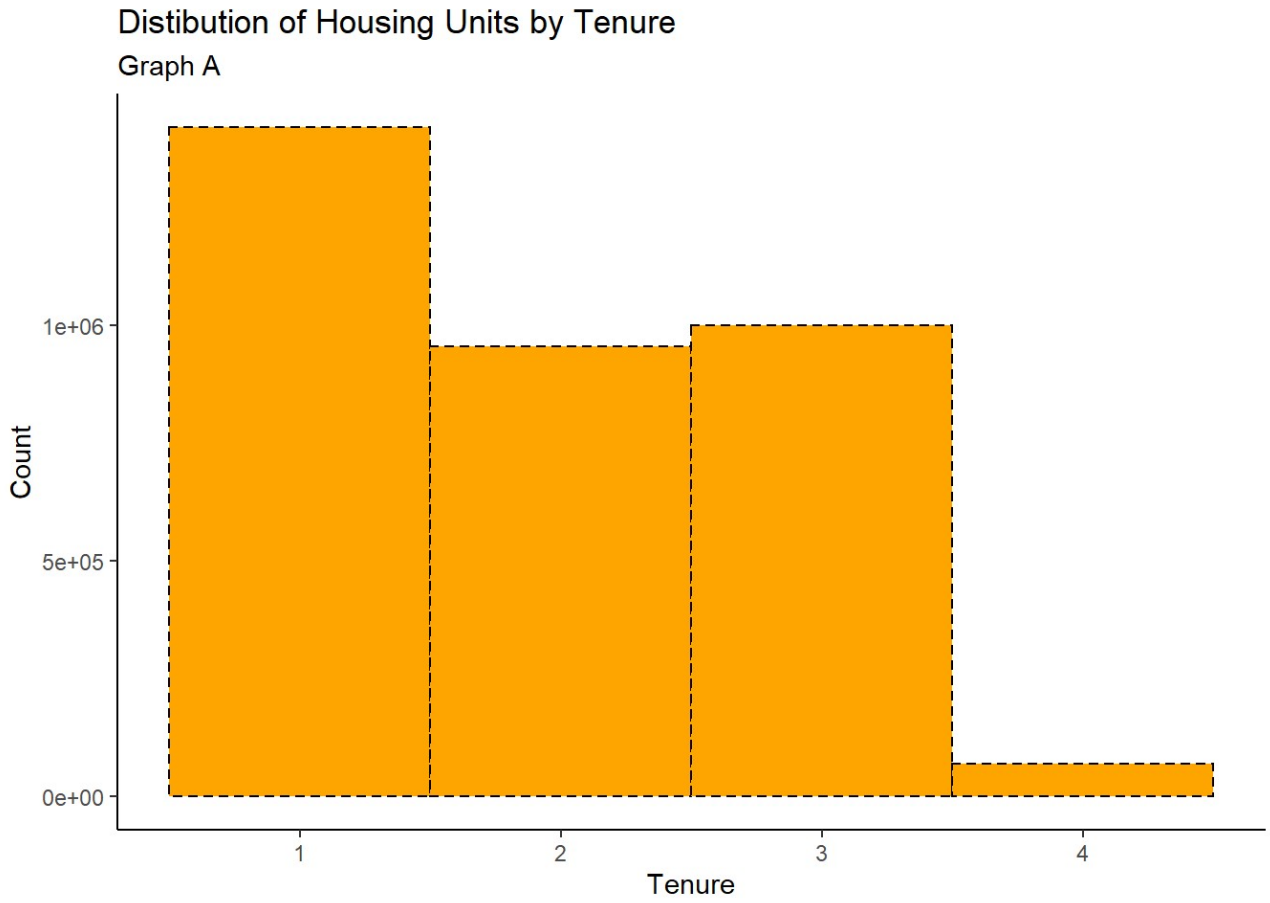
1 .Owned with mortgage or loan (include home equity loans)

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.



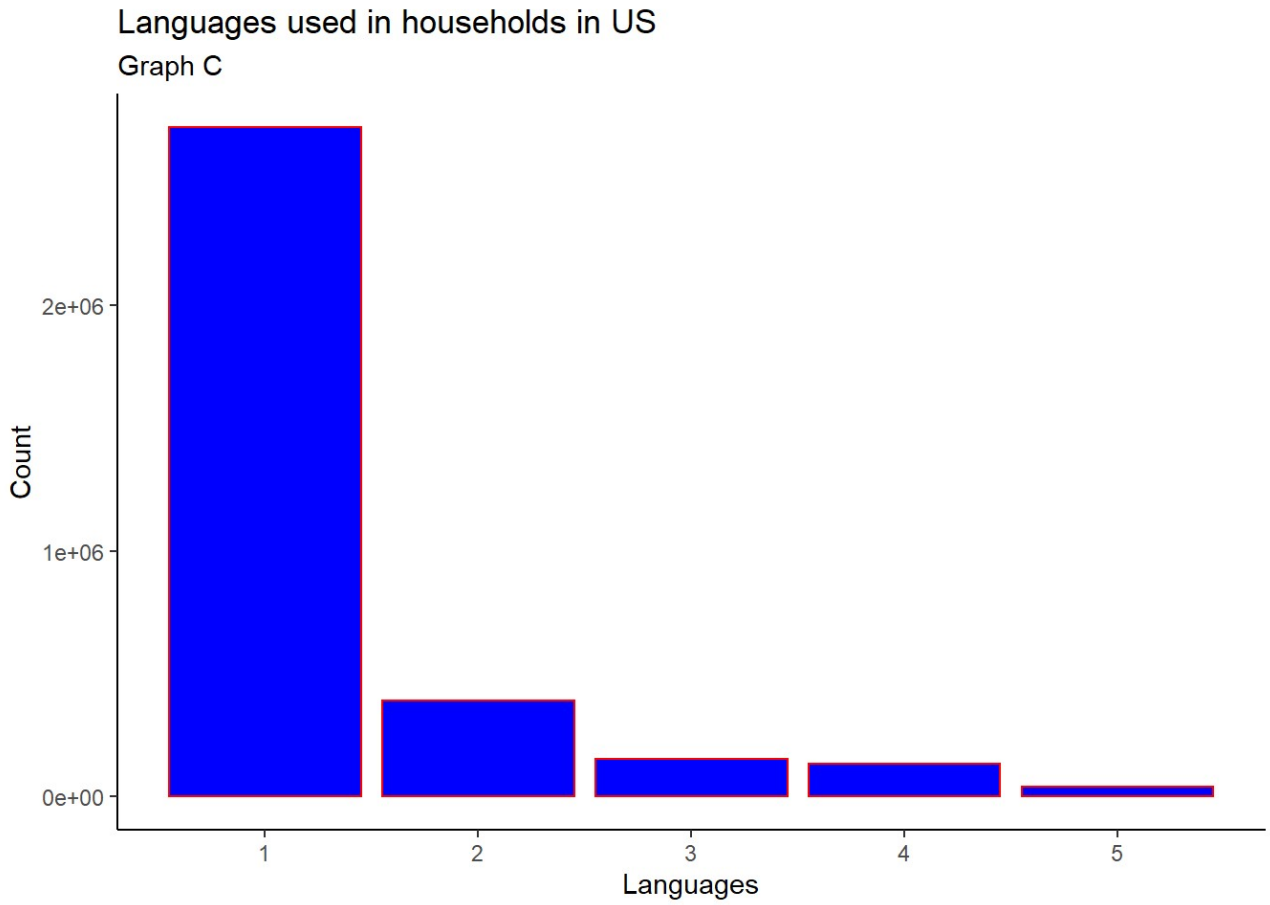
From Graph A , it is evident that Housing Units are mostly “Owned with mortgage or loan” and there are smaller number of housing units which are “Occupied without payment of rent” . However, there is no much difference between housing units which are “Owned Free and Clear” and “Rented” .

For Graph B , look at the second last graph .

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” .

B. 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

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 is shown in the last graph “Housing Units by Household languages”(which is the last graph in the report)



C. 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

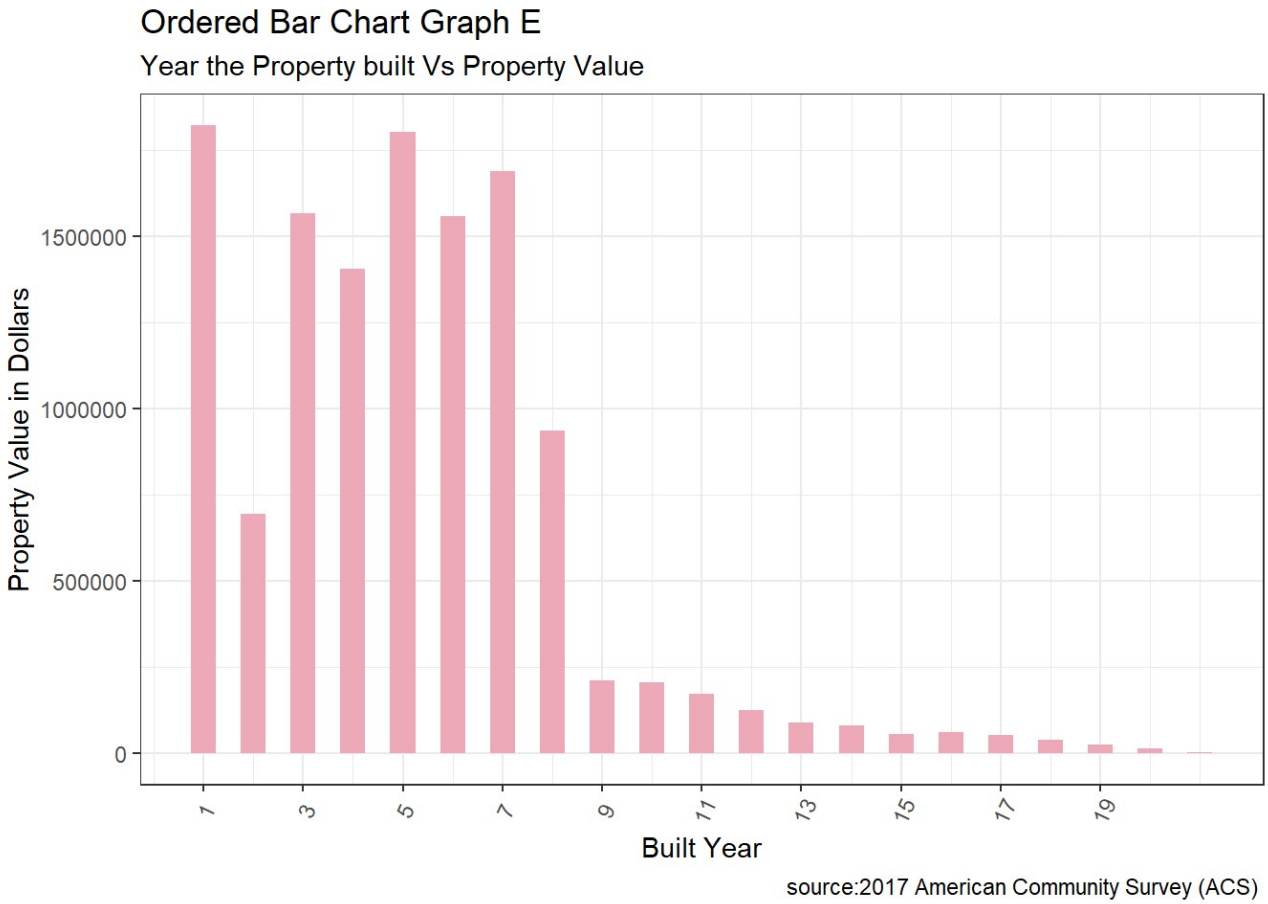
15 .2011

16 .2012

17 .2013

18 .2014

19 .2015
20 .2016
21 .2017



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 .

D. 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

```
## Loading required package: acs
```

```
## Loading required package: stringr
```

```
## Loading required package: XML
```

```
##
## Attaching package: 'acs'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## The following object is masked from 'package:base':
##
##     apply
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
## Loading required package: maps
```

```
## Loading required package: sp
```

```
## Checking rgeos availability: FALSE
##      Note: when rgeos is not available, polygon geometry     computations in maptoo
ls depend on gpclib,
##      which has a restricted licence. It is disabled by default;
##      to enable gpclib, type gpclibPermit()
```

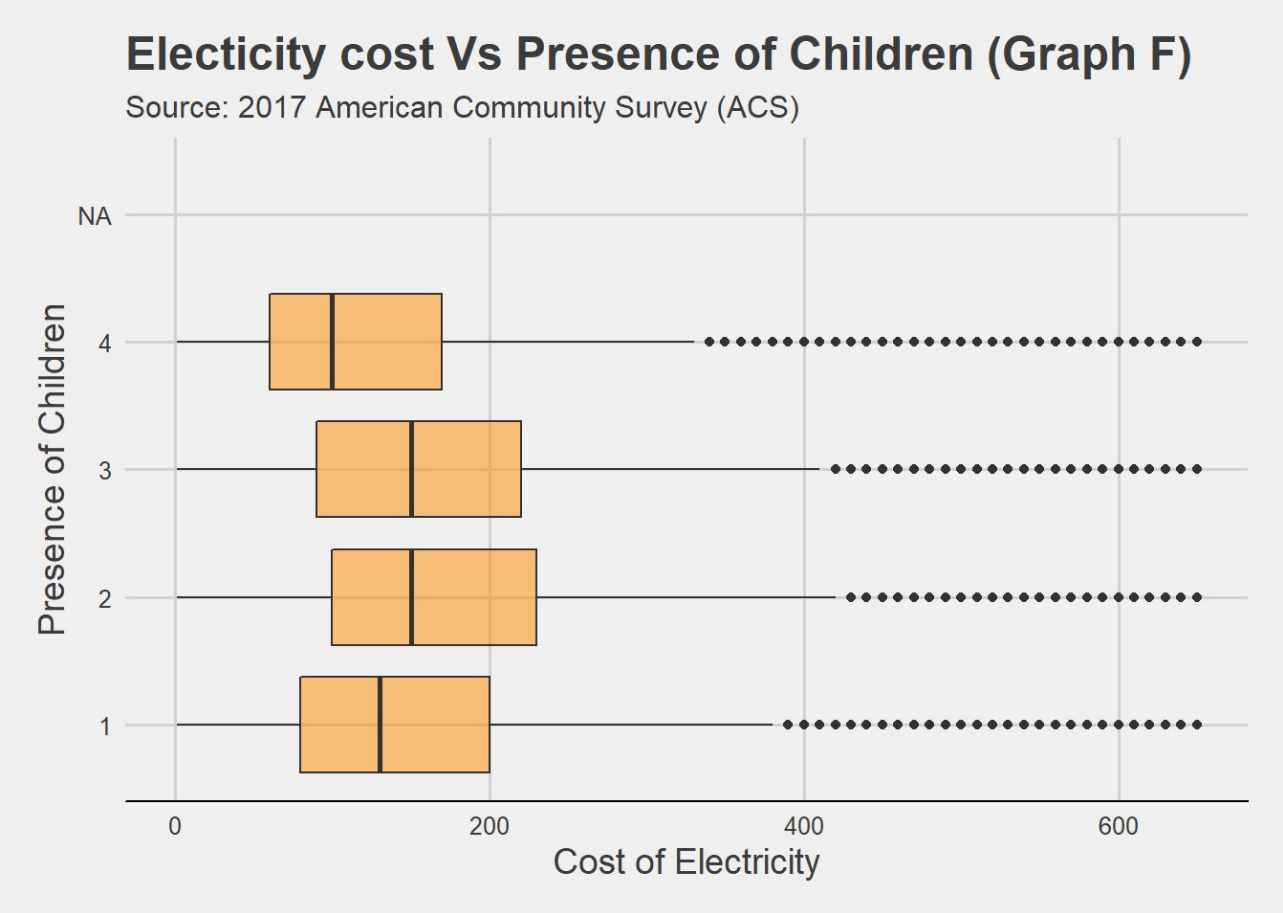
```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:acs':
##
##     combine
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## rgdal: version: 1.4-8, (SVN revision 845)
##  Geospatial Data Abstraction Library extensions to R successfully loaded
##  Loaded GDAL runtime: GDAL 2.2.3, released 2017/11/20
##  Path to GDAL shared files: C:/Users/Owner/Documents/R/win-library/3.6/sf/gdal
##  GDAL binary built with GEOS: TRUE
##  Loaded PROJ.4 runtime: Rel. 4.9.3, 15 August 2016, [PJ_VERSION: 493]
##  Path to PROJ.4 shared files: C:/Users/Owner/Documents/R/win-library/3.6/sf/proj
##  Linking to sp version: 1.3-2
```

```
## # A tibble: 4,194,300 x 16
## # Groups:   Children.Present [5]
##    State Adjustment.Fact… Lot.Size Agriculture.Sal… Electricity.Cost
##    <int>            <int>    <int>            <int>            <int>
##  1     1          1061971       NA               NA               NA
##  2     1          1061971        1               NA              350
##  3     1          1061971        1               NA              300
##  4     1          1061971        3                1              220
##  5     1          1061971       NA               NA               60
##  6     1          1061971        1               NA              100
##  7     1          1061971        1               NA              240
##  8     1          1061971        2                1              130
##  9     1          1061971       NA               NA              130
## 10     1          1061971        2                1               80
## # … with 4,194,290 more rows, and 11 more variables: Meal.Included <int>,
## #   Monthly.Rent <int>, Tenure <int>, Property.Value <int>,
## #   No.Of.Vehicles <int>, Year.Property.Built <int>, Family.Income <int>,
## #   Household.Language <int>, Household.Income <int>, Children.Present <int>,
## #   Taxes <int>
```

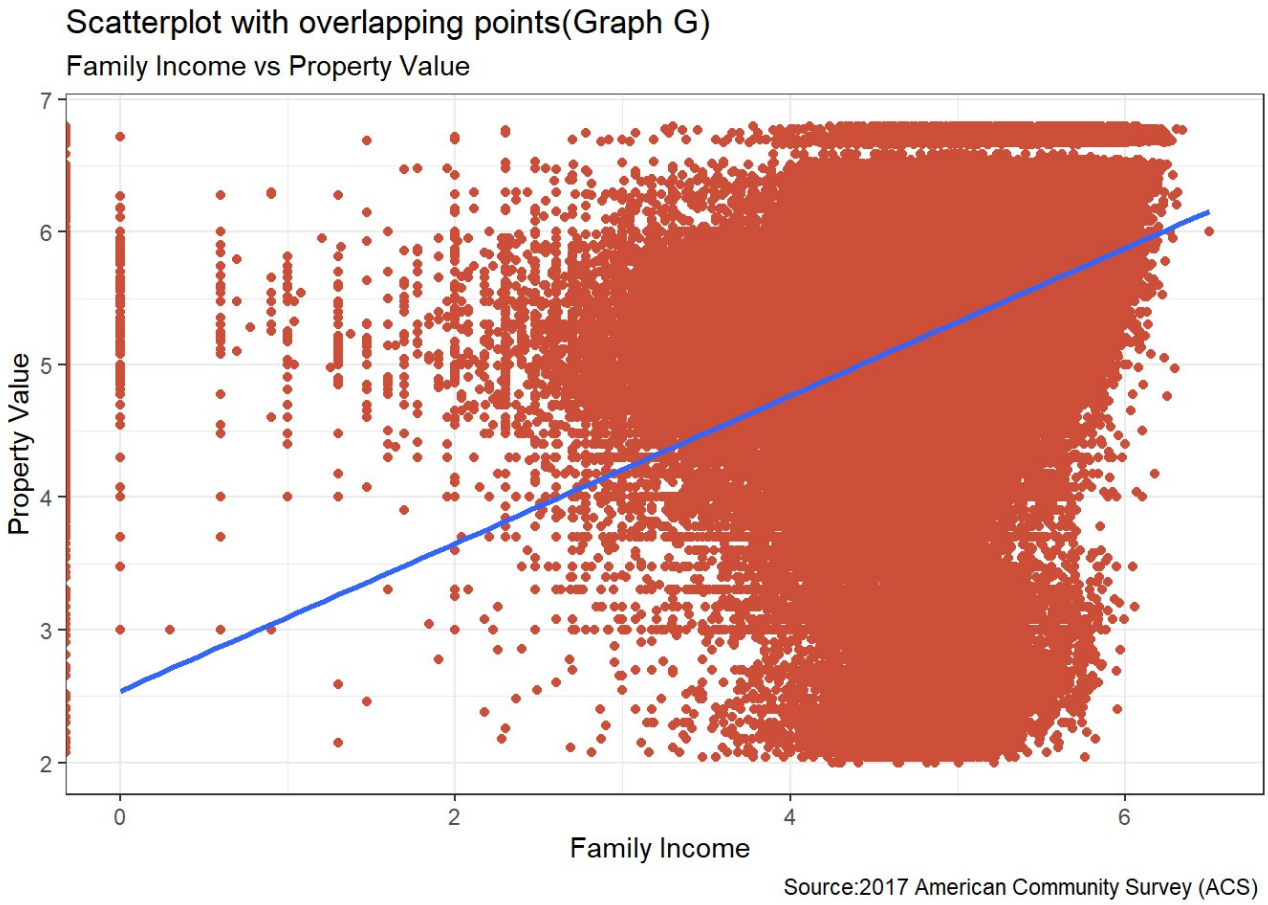


Above Graph F shows the affect of presence of children on Electricity cost of housing units. It is evident that presence of children in the housing unit can be attributed to an increase in electricity consumption . However, Age group of the children does not influence the electricity consumption charges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

E. 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.

```
## Warning in data(Housing.Unit.Survey, package = "ggplot2"): data set
## 'Housing.Unit.Survey' not found
```

```
## Warning in FUN(X[[i]], ...): NaNs produced

## Warning in FUN(X[[i]], ...): NaNs produced

## Warning in FUN(X[[i]], ...): NaNs produced
```



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 .

F.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 :

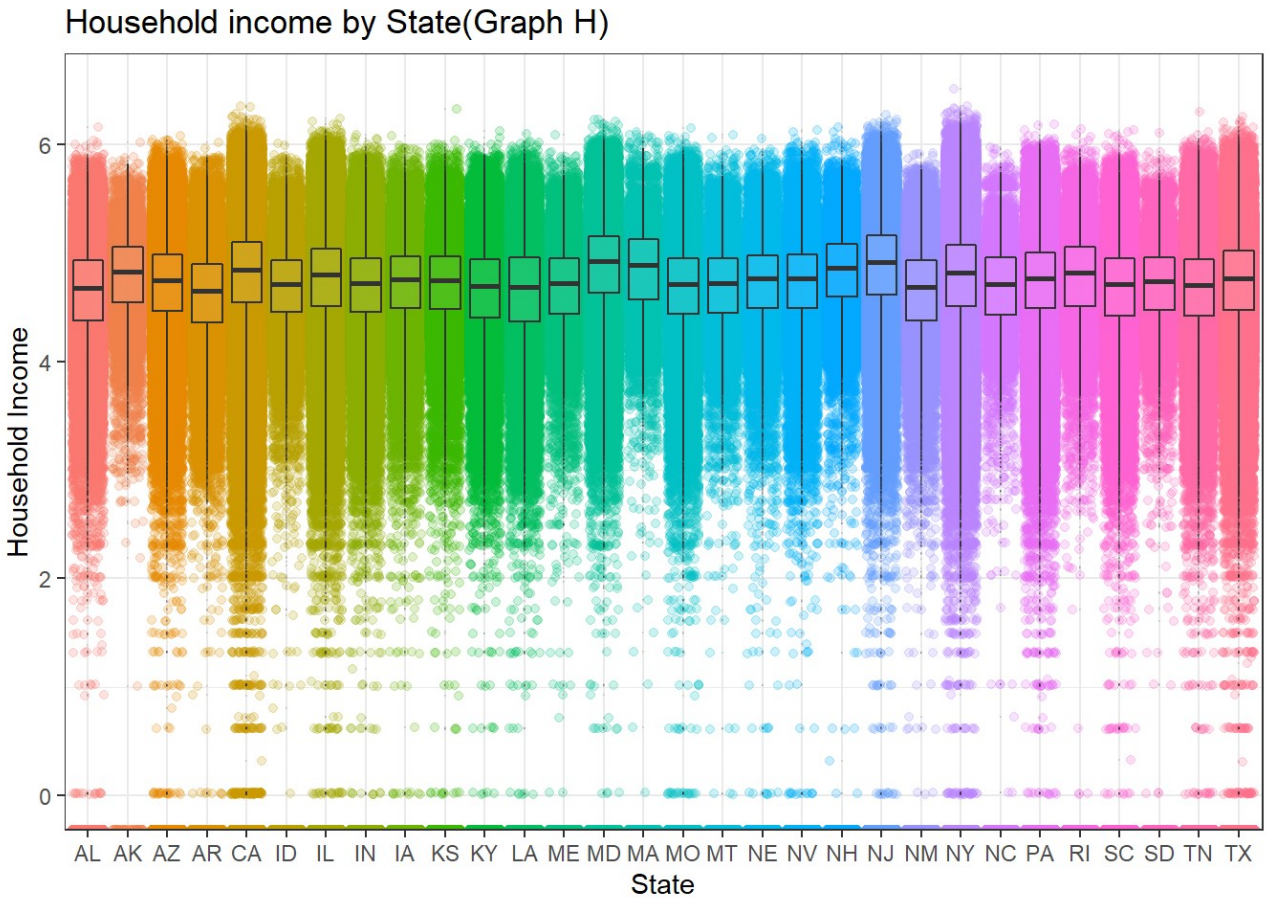
From Graph H and I , it is evident that the Family Income of Housing Units State wise and Household Income of Housing Unit State wise doesn't have much differences.

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.

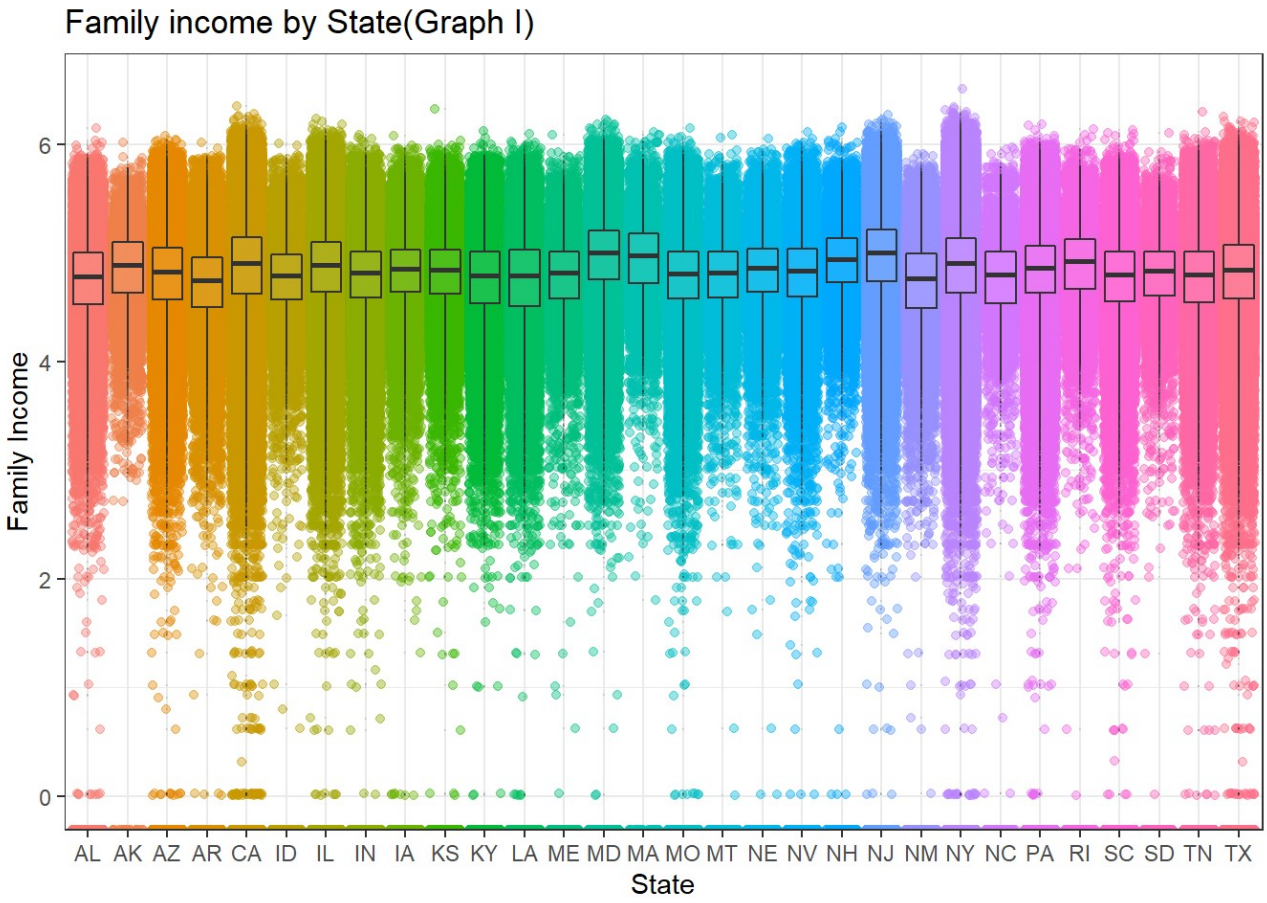
Note : We have used ADJINC variable which is an adjustment factor for housing dollar amounts. By dividing ADJINC by 1,000,000 , we can obtain the inflation adjustment factor and multiplying it to the PUMS variable value , we can adjust it to 2017 dollars. Variables requiring ADJINC on the Housing Unit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

```
##   State Adjustment.Factor Lot.Size Agriculture.Sales Electricity.Cost
## 1    AL           1061971       NA                NA               NA
## 2    AL           1061971        1                NA              350
## 3    AL           1061971        1                NA              300
## 4    AL           1061971        3                 1              220
## 5    AL           1061971       NA                NA               60
##   Meal.Included Monthly.Rent Tenure Property.Value No.Of.Vehicles
## 1            NA           NA     NA             NA             NA
## 2            NA           NA      2          25000              3
## 3            NA           NA      1          80000              1
## 4             2          100      3             NA              2
## 5             2           80      3             NA              0
##   Year.Property.Built Family.Income Household.Language Household.Income
## 1                  NA            NA                 NA               NA
## 2                   2        151000                  1           151000
## 3                   5            NA                  1            39930
## 4                   6         11400                  1            11400
## 5                   5            NA                  1             3900
##   Children.Present Taxes Inflation.Factor Household.Income.New
## 1               NA    NA         1.061971                   NA
## 2                4     3         1.061971            160357.621
## 3                4     6         1.061971             42404.502
## 4                2    NA         1.061971             12106.469
## 5                4    NA         1.061971              4141.687
##   Family.Income.New
## 1                NA
## 2         160357.62
## 3                NA
## 4          12106.47
## 5                NA
```

```
## Warning in FUN(X[[i]], ...): NaNs produced

## Warning in FUN(X[[i]], ...): NaNs produced

## Warning in FUN(X[[i]], ...): NaNs produced
```




```
## Warning in FUN(X[[i]], ...): NaNs produced

## Warning in FUN(X[[i]], ...): NaNs produced

## Warning in FUN(X[[i]], ...): NaNs produced
```



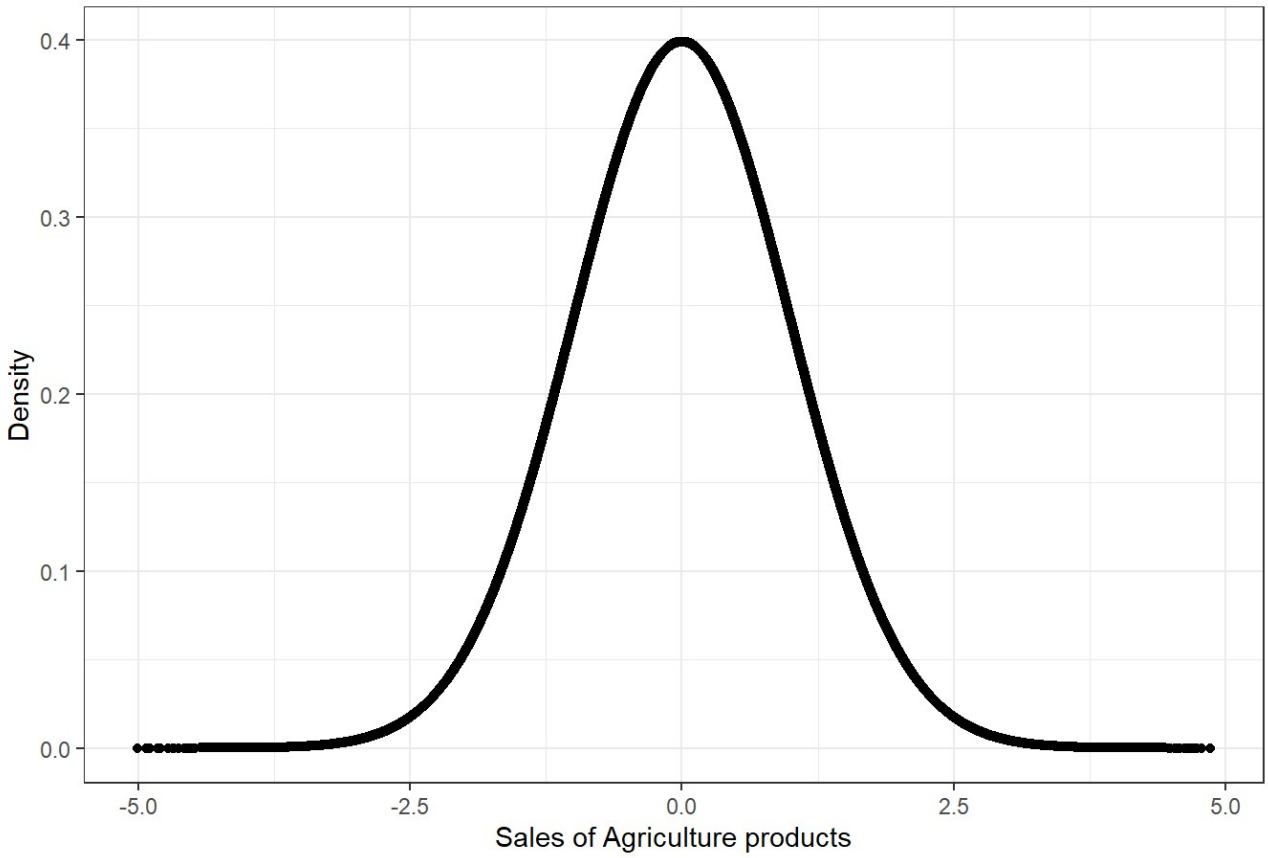
G. Density Plot for Sales of agriculture products
The density (the probability of a particular value) for the normal distribution is calculated using dnorm. While it is technically mathematically impossible to find the exact probability of a number from a continuous distribution, this is an estimate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

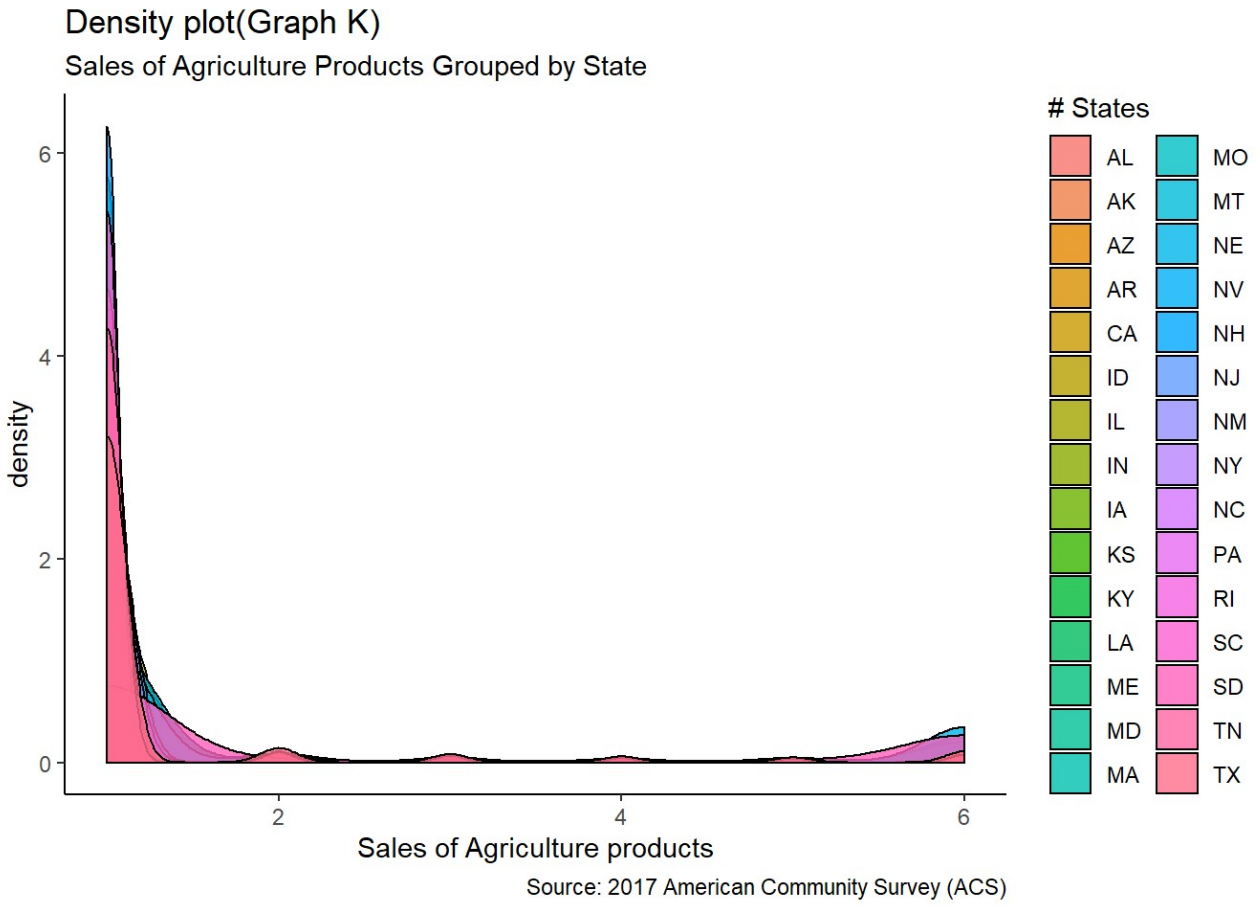
We have used these functions and prepared the below density plot for sale of agriculture products.

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

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 .

Graph J


0.4
0.3
0.2
0.1
0.0
Density
-5.0
-2.5
0.0
2.5
5.0
Sales of Agriculture products



A chi-square test tests a null hypothesis about the relationship between two variables. It requires categorical variables, usually only two, but each may have any number of levels.
Null hypothesis : There is no connection between lot size and number of vehicles.

```
##
##                                    No Vehicles 1 Vehicle 2 Vehicles
##   House on less than one acre                91467    606299     884603
##   House on one to less than ten acres        15422     99397     201516
##   House on ten or more acres                  4767     23910      51618
##
##                                    3 Vehicles 4 Vehicles 5 Vehicles
##   House on less than one acre             342826     107258      27586
##   House on one to less than ten acres     112473      42142      12865
##   House on ten or more acres               33595      14036       4867
##
##                                    6 or more Vehicles
##   House on less than one acre                   10427
##   House on one to less than ten acres            6625
##   House on ten or more acres                     3226
```

```
##
##  Pearson's Chi-squared test
##
## data:  ChisqTest
## X-squared = 59216, df = 12, p-value < 0.00000000000000022
```

Conclusion: At a 5% significance level, the data provides sufficient evidence (P-value < 0.005) that we reject the null hypothesis and it is evident that lot size and number of vehicles in the housing units are associated.

Checking if there is a relation between Property Value and Taxes .

```
##
##  Pearson's product-moment correlation
##
## data:  Housing.Unit.Survey$Property.Value and Housing.Unit.Survey$Taxes
## t = 740.79, df = 2376958, p-value < 0.00000000000000022
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4320548 0.4341204
## sample estimates:
##       cor
## 0.4330882
```

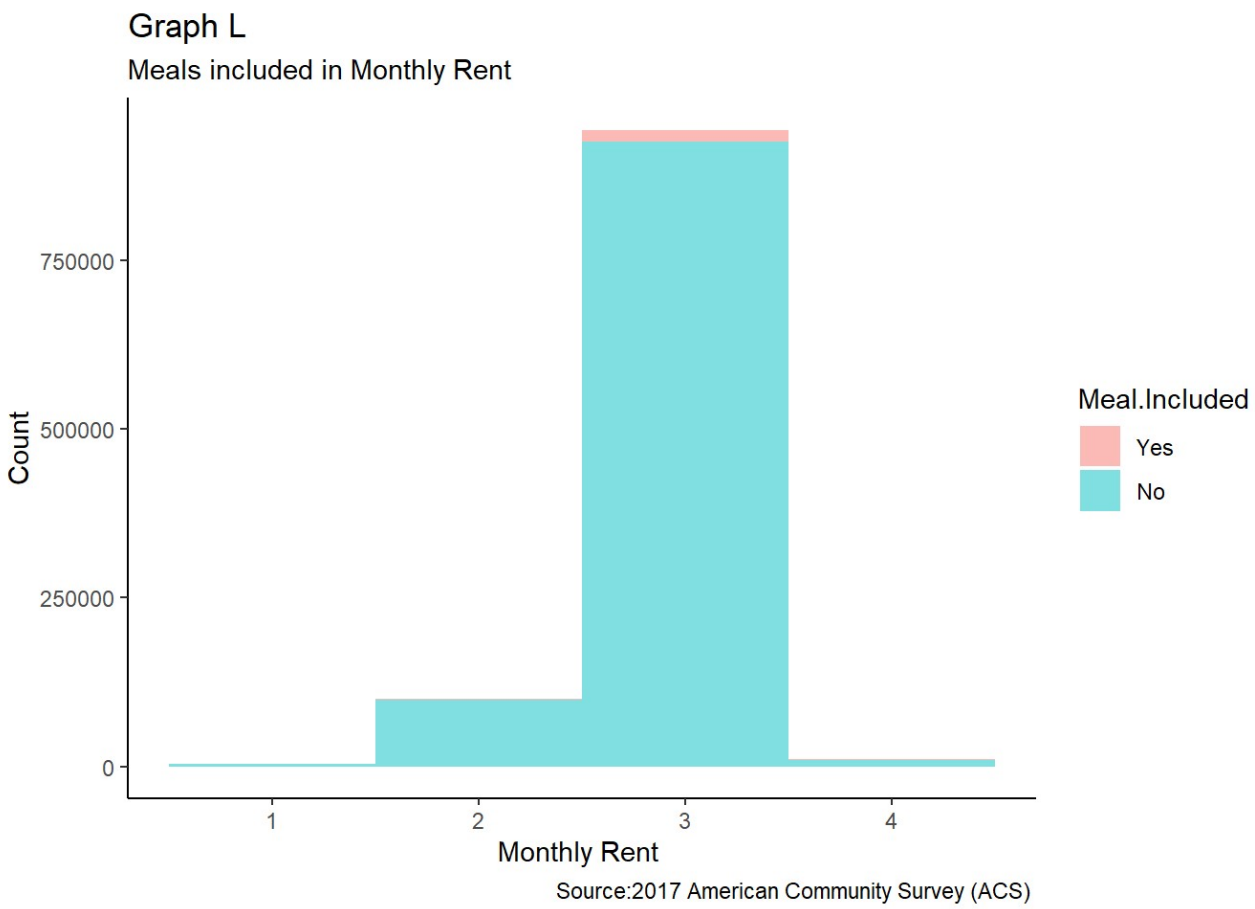
Conclusion : This strong positive correlation makes sense because higher the property value higher the taxes. They are co-related

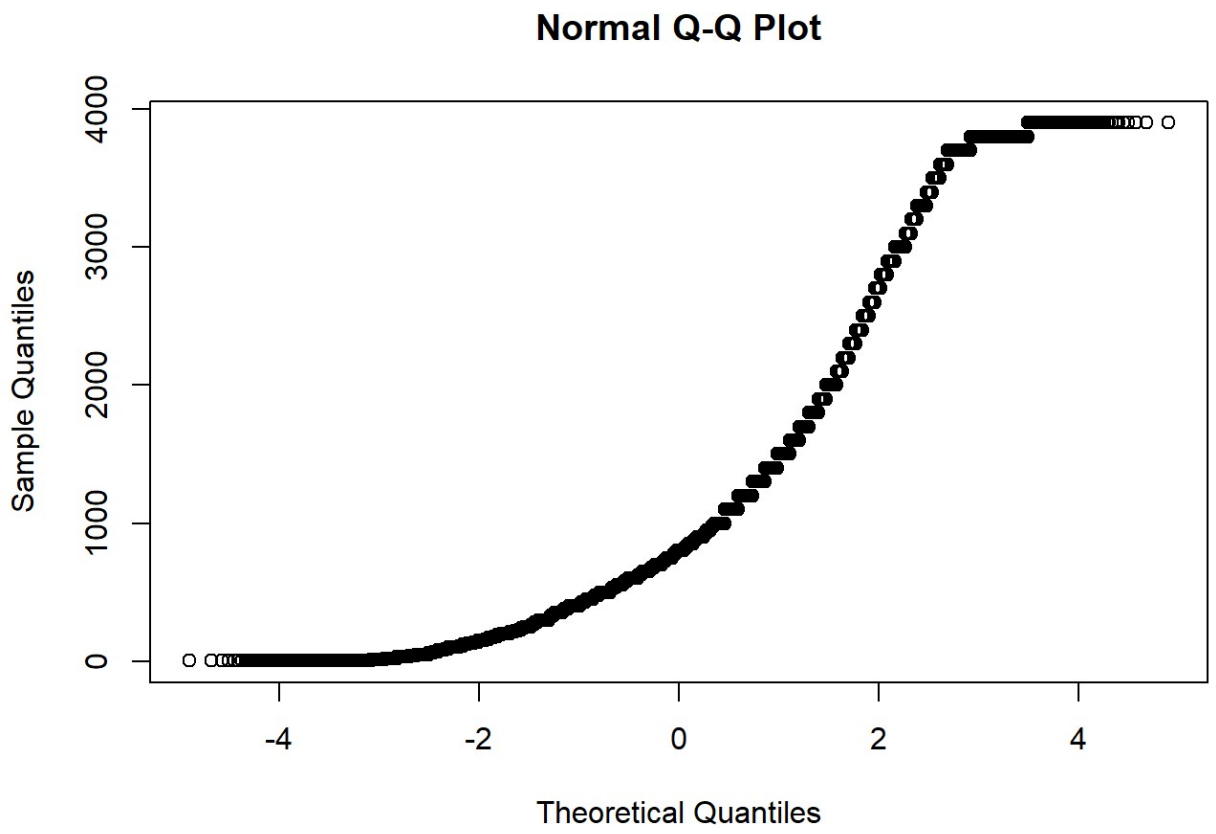
H. Meals Included in Monthly rent affects Rent

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 :

```
##   Meal.Included Monthly.Rent
## 1           Yes     892868.0
## 2            No     364959.1
```


Graph L
Meals included in Monthly Rent
750000
500000
250000
0
Count
1
2
3
4
Monthly Rent
Source:2017 American Community Survey (ACS)
Meal.Included
Yes
No

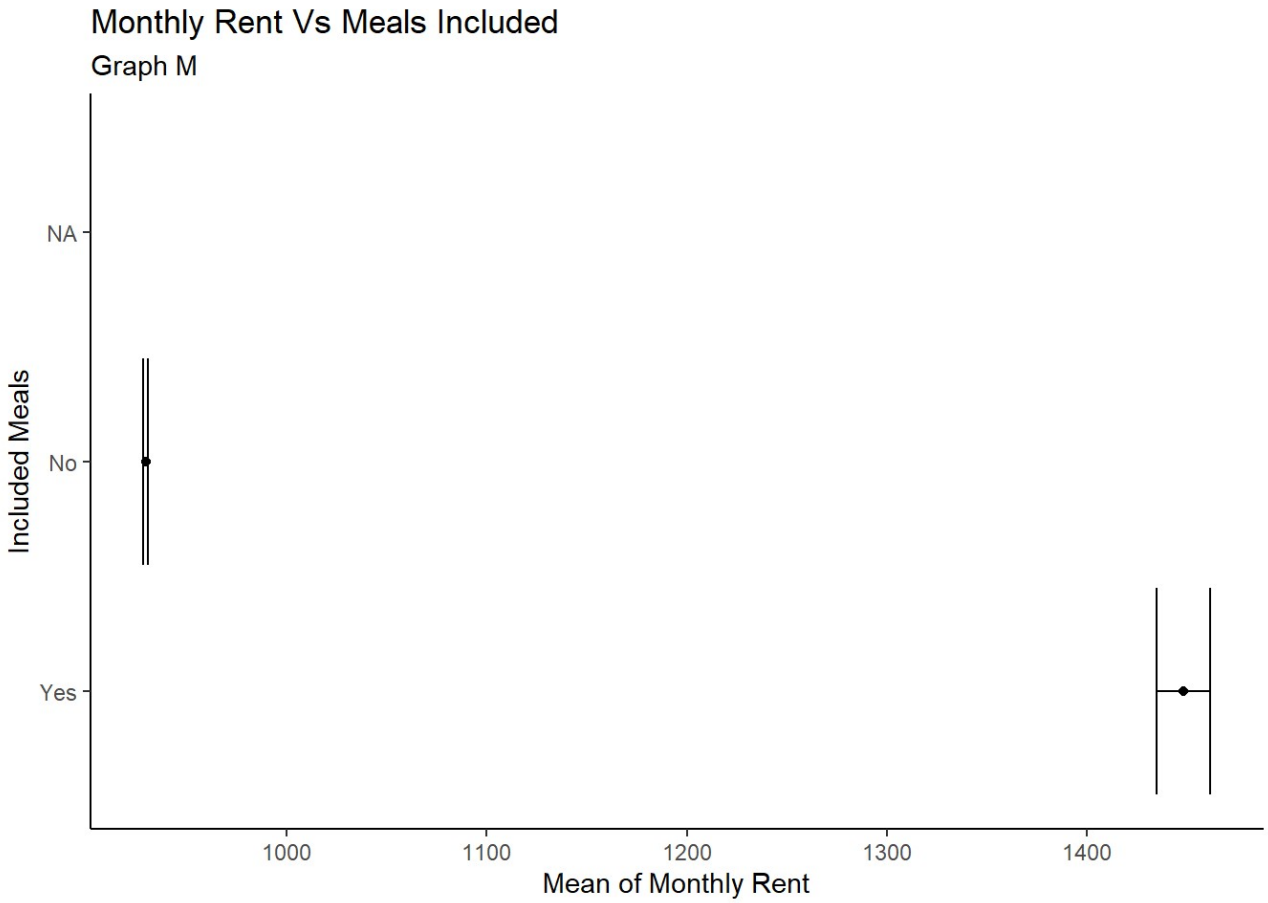



```
##
##  Two Sample t-test
##
## data:  Monthly.Rent by Meal.Included
## t = 119.06, df = 1057863, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  509.7730 526.8371
## sample estimates:
## mean in group Yes  mean in group No
##         1448.1456          929.8406
```

```
## Warning: Factor `Meal.Included` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

```
## # A tibble: 3 x 5
##   Meal.Included mean.monthlyrent sd.monthlyrent lower upper
##   <fct>                    <dbl>          <dbl> <dbl> <dbl>
## 1 Yes                      1448.           945. 1435. 1461.
## 2 No                        930.           604.  929.  931.
## 3 <NA>                       NA             NaN    NA    NA
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

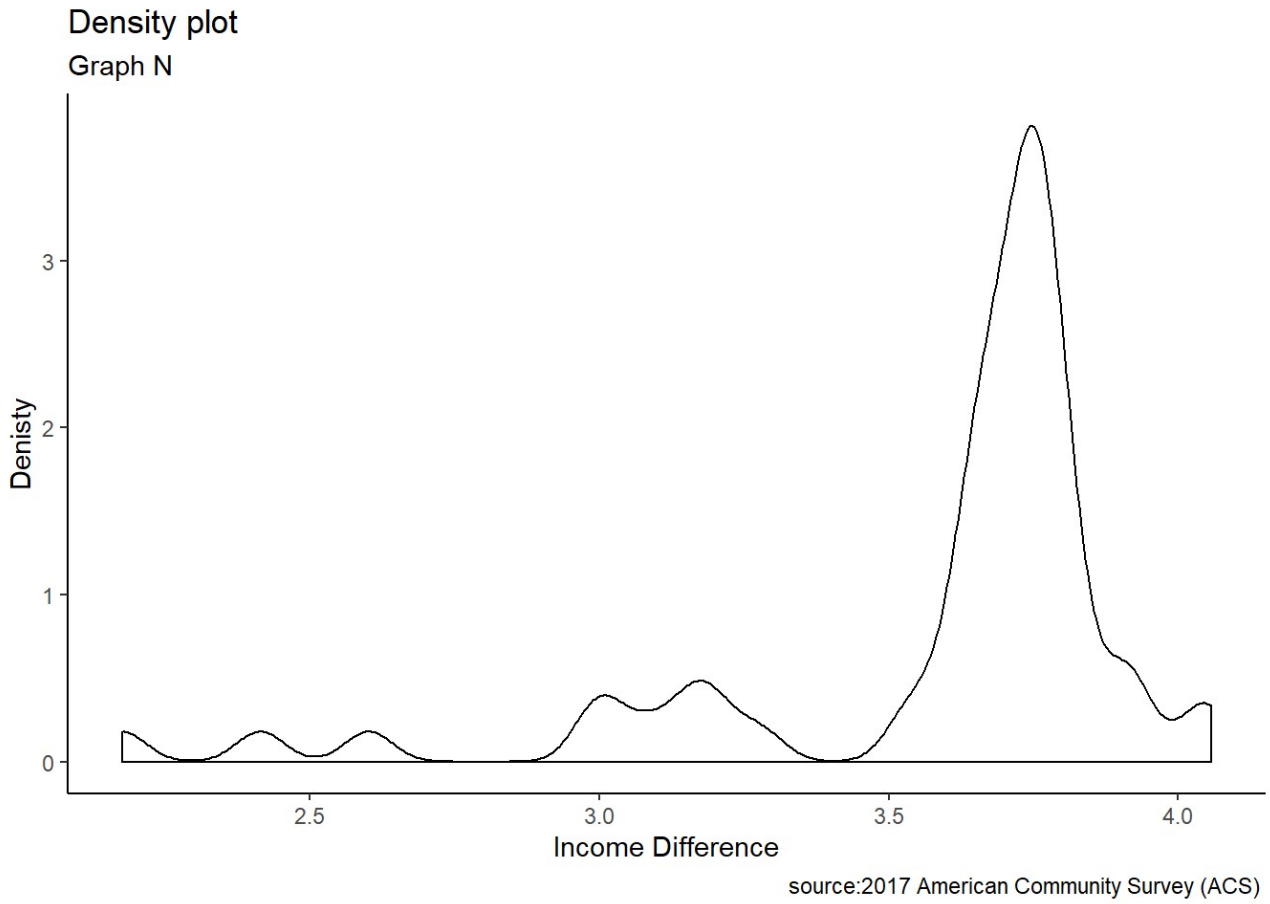


We have performed two sample paired T-tests for family income and household income.
Null hypothesis : Both are not same
Form the t-test results, it is evident that p-value is less than 0.05 means we can reject the null hypothesis and conclude that both income are same.

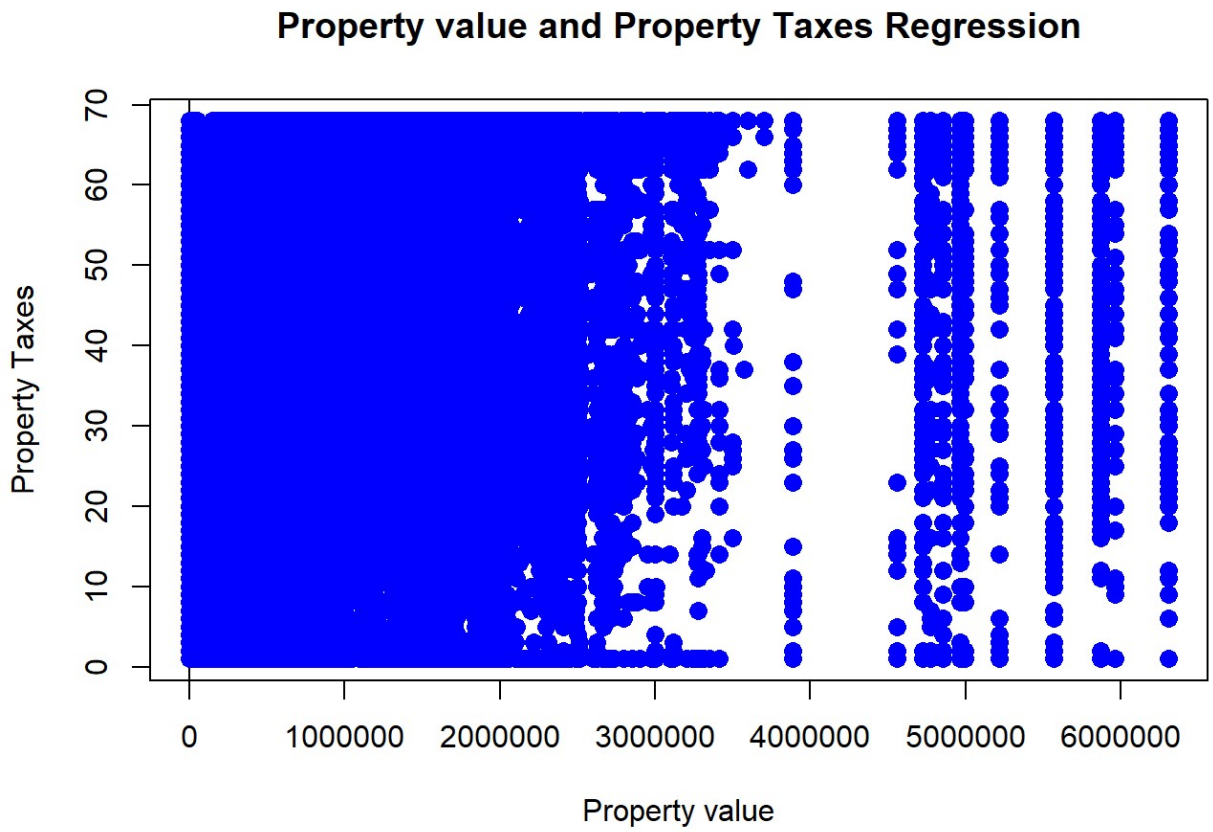
```
##
##  Paired t-test
##
## data:  Housing.Unit.Survey$Family.Income and Housing.Unit.Survey$Household.Income
## t = -224.51, df = 2308582, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1625.201 -1597.071
## sample estimates:
## mean of the differences
##               -1611.136
```

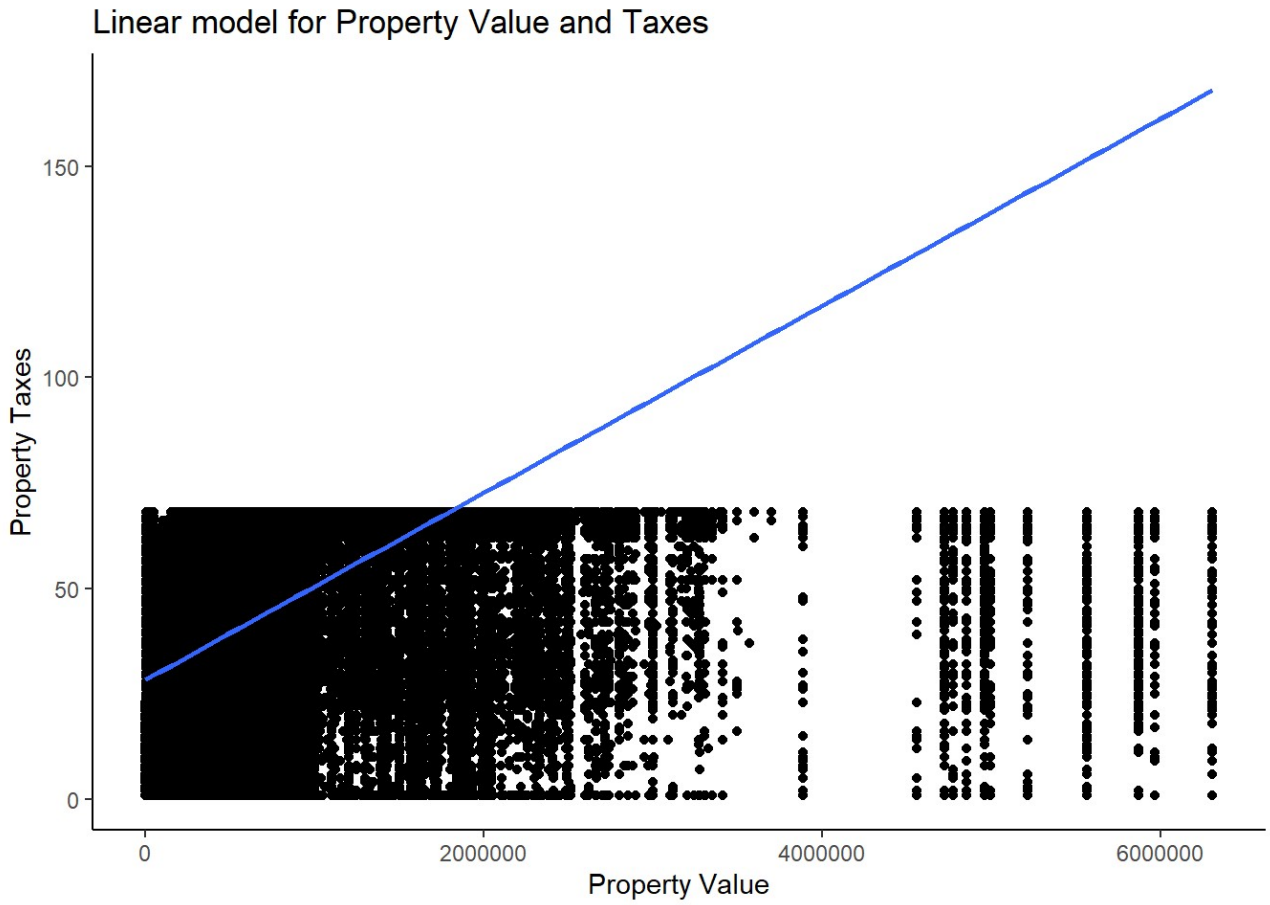
```
## Warning in FUN(X[[i]], ...): NaNs produced

## Warning in FUN(X[[i]], ...): NaNs produced
```

```
## Warning: Removed 1 rows containing missing values (geom_vline).
```



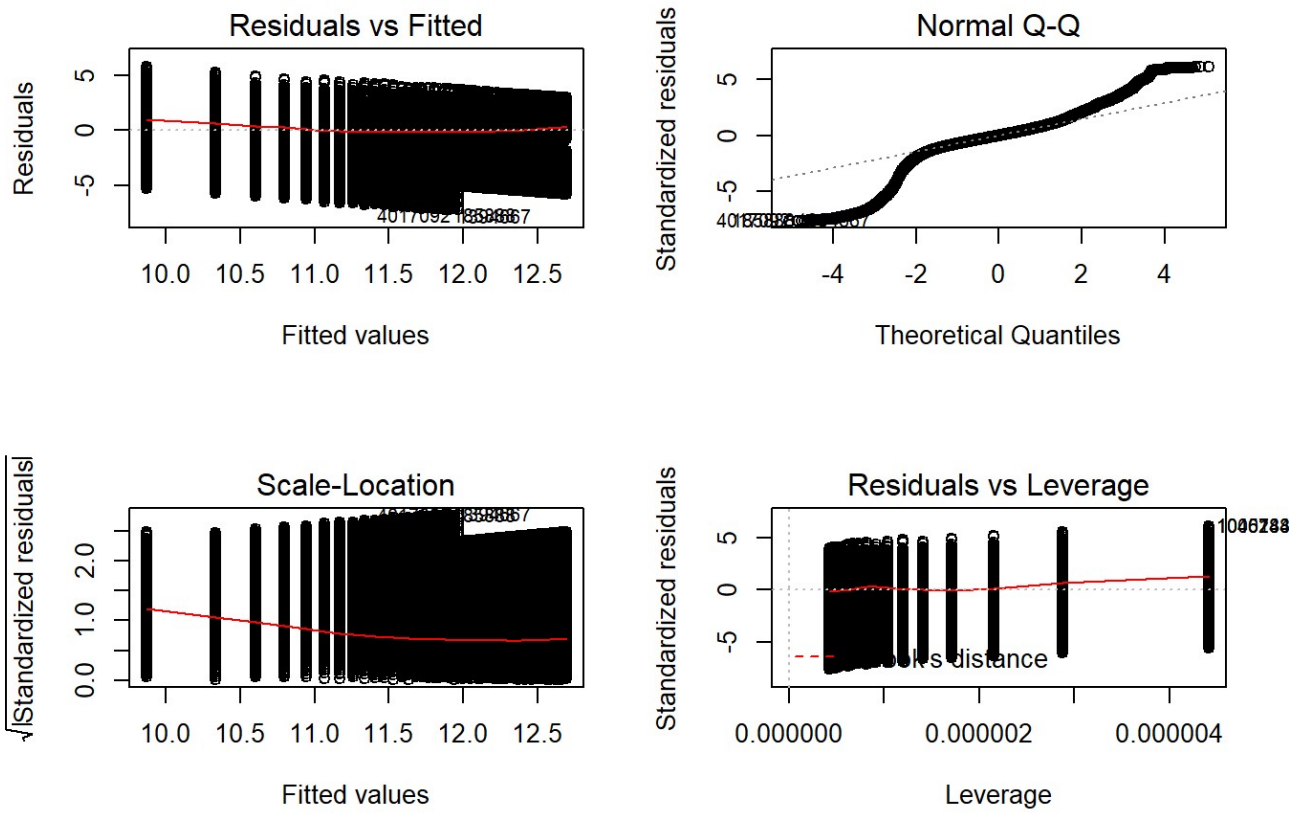
I. Linear Regression model for Property Value and Taxes


Property value and Property Taxes Regression
Property Taxes
0
10
20
30
40
50
60
70
0
1000000
2000000
3000000
4000000
5000000
6000000
Property value




```
##
## Call:
## lm(formula = V ~ TA, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -567017 -144820  -45741   40023 6306857
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -7317.85     462.86  -15.81 <0.0000000000000002 ***
## TA           8460.81      11.42  740.79 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 369300 on 2376958 degrees of freedom
##   (1817340 observations deleted due to missingness)
## Multiple R-squared:  0.1876, Adjusted R-squared:  0.1876
## F-statistic: 5.488e+05 on 1 and 2376958 DF,  p-value: < 0.00000000000000022
```

Conclusion : Residuals are essentially the difference between the actual observed response values (distance to stop) and the response values that the model predicted based on the best fit line. Residuals difference here is strong negative . Because of the negative and missing values in the dataset estimate coefficient value is negative here for our model , which suggests there will be no taxes on property . However , if property values increases then Taxes value will be 8460.81 (Approximately)



Graph 1: There is a clear indication of non-linearity present in this plot. Furthermore, we see that the variance appears to be increasing in fitted value.

Graph 2 :The residuals appear highly non-normal. Both the lower tail and upper tail are heavier than we would expect under normality.This may be due to the non-constant variance issue we observed in the Residuals vs. Fitted plot.

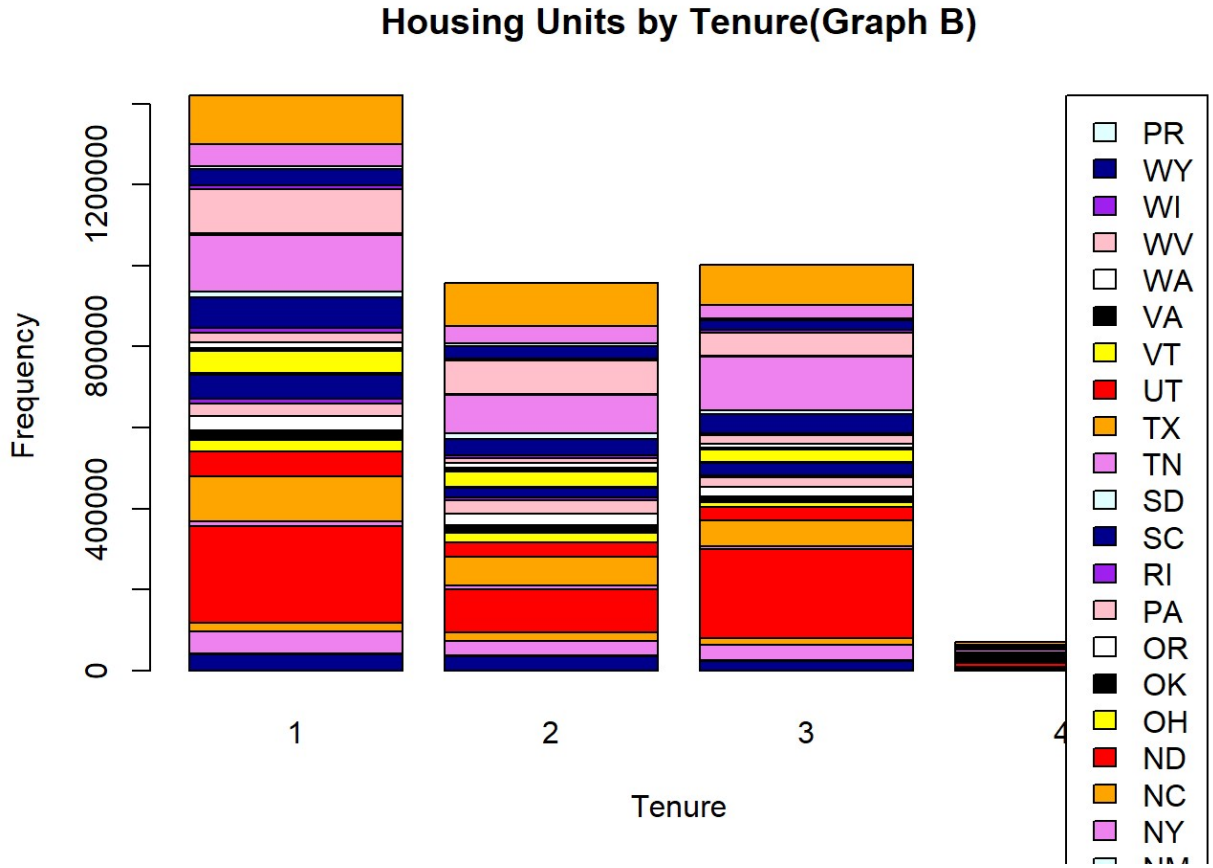
Graph 3 :We see a clear increasing trend in residual variance that runs through most of the plot. This is indicated by the upward slope of the red line, which we can interpret as the standard deviation of the residuals at the given level of fitted value.

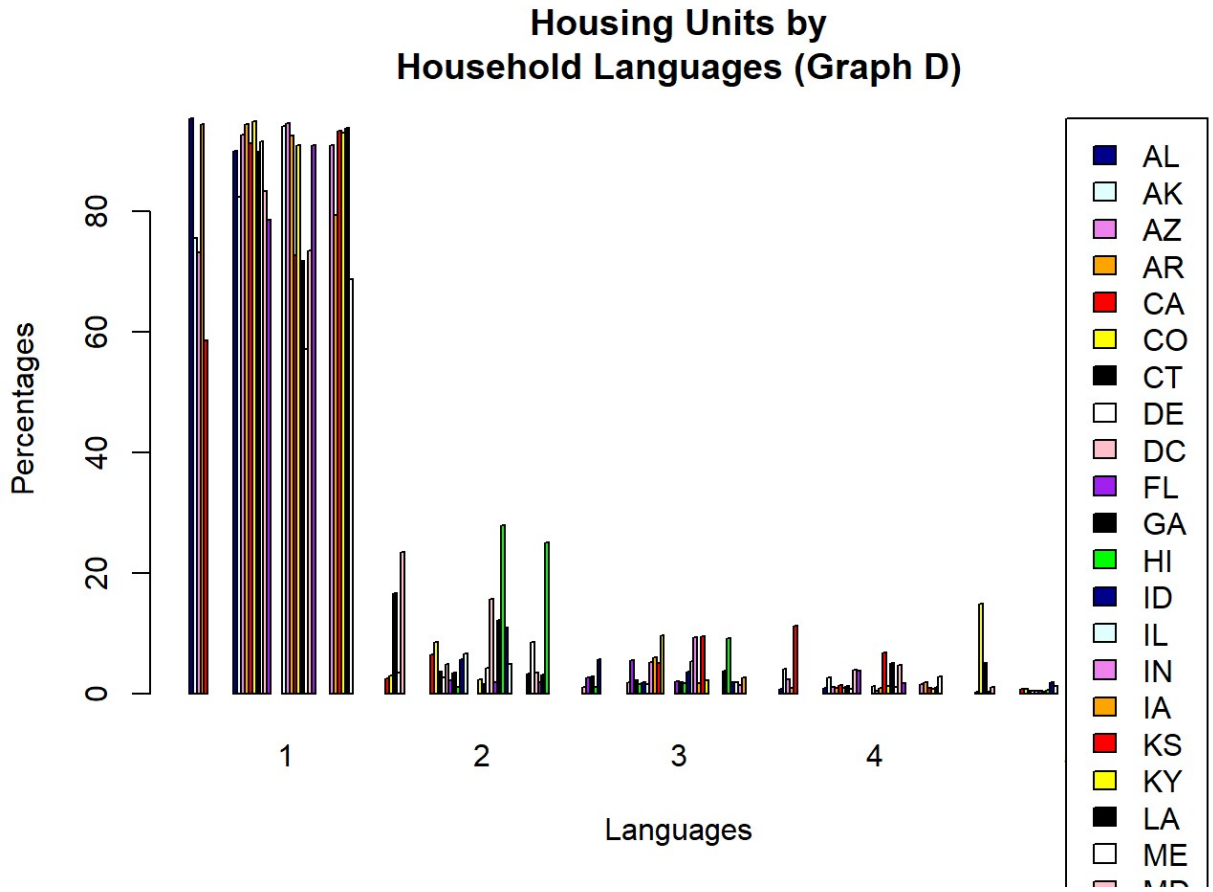
Graph 4 : None of the points appear to be outliers.

Explanation for the below Two graphs has been given above.

```
##
##          1      2      3      4
##   AL  38671  33568  22740   3430
##   AK   3955   3510   3068    647
##   AZ  53915  35114  37746   2685
##   AR  21570  21285  14928   2179
##   CA 237428 107020 221584   8895
##   CO      0      0      0      0
##   CT      0      0      0      0
##   DE      0      0      0      0
##   DC      0      0      0      0
##   FL      0      0      0      0
##   GA      0      0      0      0
##   HI      0      0      0      0
##   ID  13737   8899   7196    768
##   IL 109870  70966  64267   3961
##   IN  61757  35845  32184   2359
##   IA  28016  22538  12914   1339
##   KS  22822  19688  14113   1326
##   KY  36899  28633  21980   2645
##   LA  30680  32158  24087   3309
##   ME  12124   9189   5032    599
##   MD  59053  23435  29701   1417
##   MA   5458   2468   3527    130
##   MI      0      0      0      0
##   MN      0      0      0      0
##   MS      0      0      0      0
##   MO  52582  37715  31146   2889
##   MT   7597   7617   4622    900
##   NE  15306  13000   8976    925
##   NV  21952  11575  20367    772
##   NH  13152   7576   5801    372
##   NJ  75936  39708  47455   1962
##   NM  13567  14411   9160   1175
##   NY 139291  95094 131843   6090
##   NC   4491   2576   2762    284
##   ND      0      0      0      0
##   OH      0      0      0      0
##   OK      0      0      0      0
##   OR      0      0      0      0
##   PA 109839  81851  57810   5086
##   RI   9328   4665   7039    278
##   SC  40336  31992  23163   2798
##   SD   6158   6574   3909    508
##   TN  53961  41869  33778   3598
##   TX 121451 105519  98258   7331
##   UT      0      0      0      0
##   VT      0      0      0      0
```

```
##   VA      0      0      0      0
##   WA      0      0      0      0
##   WV      0      0      0      0
##   WI      0      0      0      0
##   WY      0      0      0      0
##   PR      0      0      0      0
```


Housing Units by Tenure(Graph B)
Frequency
0
400000
800000
1200000
1
2
3
4
Tenure
PR
WY
WI
WV
WA
VA
VT
UT
TX
TN
SD
SC
RI
PA
OR
OK
OH
ND
NC
NY



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 langauges 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 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.