

Introduction

The American Community Survey (ACS) provides vital information on a yearly basis about the nation and people. Through the ACS, we know more about jobs and occupations, educational attainment, veterans, whether people own or rent their home and other topics. Public officials, planners and entrepreneurs use this information to assess the past and plan the future. This data is used to plan hospital and schools, to support school lunch programs, to improve emergency services, to build bridges and to inform businesses looking to add jobs and expand to new markets and more. The ACS consists of 72 questions pertaining to the topics split into population and household characteristics.

- The population section is about employment, education, veterans, income and housing costs, commuting, disability and health insurance.
- The Housing data is about:
 - housing characteristics: these variables ask about plumbing, kitchen facilities, and other housing features to help identify areas with substandard housing. There are variables about size and age of housing also flag local problems like overcrowding, health hazards, and congestion. Through this data we can find out communities eligible for housing assistance, rehabilitation loans, and other program that help people afford decent, safe, and sanitary housing.
 - Owners and Renters: the questioners were asked whether they own or rent their home, and the amount of monthly rent or how much the home and property are worth. These statistics are used to analyze whether housing is affordable, protect owners and renters and allocate and find assistance programs. These statistics are used to understand changes in local housing markets, monitor affordability, qualify for assistance and reduce the tax revenue losses from vacant or abandoned properties. These data can be used to design and market homes, and home goods by Businesses.
 - People and Relationships: these parts of variables are about respondents' age, sex, race, Hispanic origin, and their relationship to others in the household. This information is used to monitor well-being, discrimination, and economic hardship. With the help of these information we can find groups such as single parents, low-income families, older people living alone, etc. and provide funds and services for these groups. Also, businesses use these estimates to evaluate local market demand for products and services.

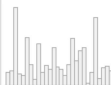
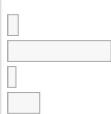
In this project I worked on Housing dataset and try to find some information about vulnerable groups and mostly focused on Renters.

Data Frame Summary

Housing

Dimensions: 7487361 x 14

Duplicates: 1327802

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing								
1	State [numeric]	Mean (sd) : 27.8 (15.9) min < med < max: 1 < 27 < 56 IQR (CV) : 30 (0.6)	51 distinct values		7487361 (100%)	0 (0%)								
2	Units.structure [factor]	1. others 2. SingleFamily.Detached 3. SingleFamily.Attached 4. Manufactured.Housing	<table><tr><td>442600</td><td>(6.6%)</td></tr><tr><td>4511282</td><td>(66.9%)</td></tr><tr><td>369354</td><td>(5.5%)</td></tr><tr><td>1423410</td><td>(21.1%)</td></tr></table>	442600	(6.6%)	4511282	(66.9%)	369354	(5.5%)	1423410	(21.1%)		6746646 (90.11%)	740715 (9.89%)
442600	(6.6%)													
4511282	(66.9%)													
369354	(5.5%)													
1423410	(21.1%)													

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
3	Tenure [factor]	1. Owned.mtg 2. Owned.free 3. Rented 4. Others	<div>2576950 (42.0%)</div> <div>1707678 (27.9%)</div> <div>1725632 (28.1%)</div> <div>121964 (2.0%)</div>		6132224 (81.9%)	1355137 (18.1%)
4	Food.Stamp [numeric]	Min : 1 Mean : 1.9 Max : 2	<div>1 : 775443 (11.3%)</div> <div>2 : 6097496 (88.7%)</div>		6872939 (91.79%)	614422 (8.21%)
5	HH.income [numeric]	Mean (sd) : 80345.2 (88145.5) min < med < max: -21500 < 57000 < 3209000 IQR (CV) : 71600 (1.1)	56465 distinct values		6132224 (81.9%)	1355137 (18.1%)
6	Grs.rent [numeric]	Mean (sd) : 1071 (612) min < med < max: 4 < 940 < 5022 IQR (CV) : 666 (0.6)	4303 distinct values		1725632 (23.05%)	5761729 (76.95%)
7	Property.Value [numeric]	Mean (sd) : 276504.7 (383572.9) min < med < max: 100 < 180000 < 6308000 IQR (CV) : 220000 (1.4)	2462 distinct values		4347580 (58.07%)	3139781 (41.93%)
8	Owner.costs [numeric]	Mean (sd) : 1257.3 (1050.8) min < med < max: 0 < 981 < 13392 IQR (CV) : 1151 (0.8)	9834 distinct values		4284628 (57.22%)	3202733 (42.78%)
9	No.rooms [numeric]	Mean (sd) : 6 (2.4) min < med < max: 1 < 6 < 30 IQR (CV) : 3 (0.4)	27 distinct values		6746646 (90.11%)	740715 (9.89%)
10	when.moved [factor]	1. 12 months or less 2. 13 to 23 months . 3. 2 to 4 years . 4. 5 to 9 years . 5. 10 to 19 years 6. 20 years or more	<div>738449 (12.0%)</div> <div>397311 (6.5%)</div> <div>1008962 (16.4%)</div> <div>1001457 (16.3%)</div> <div>1406809 (22.9%)</div> <div>1579202 (25.8%)</div>		6132190 (81.9%)	1355171 (18.1%)
11	Householder [factor]	1. others 2. single.m 3. single.f	<div>5792717 (94.5%)</div> <div>168265 (2.7%)</div> <div>171242 (2.8%)</div>		6132224 (81.9%)	1355137 (18.1%)
12	No.children [numeric]	Mean (sd) : 0.5 (1) min < med < max: 0 < 0 < 18 IQR (CV) : 1 (1.9)	19 distinct values		6132224 (81.9%)	1355137 (18.1%)
13	F.Kitchen [numeric]	Min : 0 Mean : 0 Max : 1	<div>0 : 6525705 (96.7%)</div> <div>1 : 220941 (3.3%)</div>		6746646 (90.11%)	740715 (9.89%)
14	F.Plumbing [numeric]	Min : 0 Mean : 0 Max : 1	<div>0 : 6570331 (97.4%)</div> <div>1 : 176315 (2.6%)</div>		6746646 (90.11%)	740715 (9.89%)

Summary Table

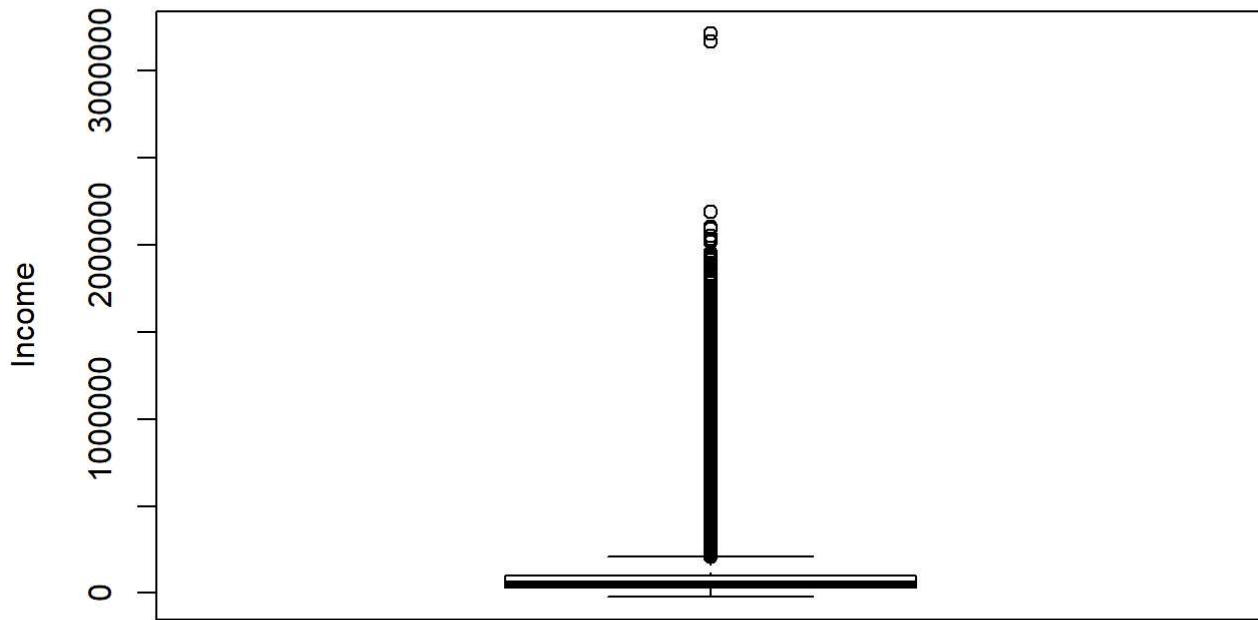
The summary table - as is shown - contains certain variables and related major data that are used in this analysis.

- **State** The first variable in the table refers to States of America. This survey analyzes the given housing data of united stated.
- **Units.Structure**
The second variable refers to Units in Structure. A structure is a separate building that either has open spaces on all sides or is separated from other structures by dividing walls. The units in structure are subdivided to four categories: **singlefamily.detached**, **singlefamily.attached**, **manufactured.Housing** and **other** groups that provide information on the housing inventory. According to the data the Unit Detached (a unit structure with open space on all four sides) is the most popular type of unit structure, 67% and manufactured housing (apartments containing 2 or more housing units) is second, 21%. More than 90% of given data is valid in this category. The data used to analyze the American community interest in marketing area.
- **Tenure** The third variable refers to tenure. The question was asked at occupied housing units. Occupied housing units are classified as **owned.mtg**, **owned.free**, **rented** and **other**. Tenure provides a measurement of home ownership, which has served as an indicator of the nation's economy for decades. This data shows that the most unites are under the mortgage, 42% and also around 28% of them are rented unites. Just near 28% of American has own unites without related loans. The data used to analyze the mortgage and real estate marketing in United States. The data also serve in understanding the characteristics of owner- occupied and renter- occupied units to aid builders, mortgage lenders, planning officials, government agencies, etc. More than 80% of given data is valid in this category.
- **Food.Stamp**
The forth variable refers to Food Stamp benefits. The Food Stamp is a funded program as one intended to permit low-income households to obtain a more nutritious diet. The data shows that about 89% of Americans don't receive national income support program. This variable also shows the tau of poverty status, to measure economic well- being, and to assess the need for assistance in each state and also is used for comparing the amount of income to determine the welfare and the comfort ratio. The results are reliable according to 92% validity.
- **HH.income** Householder income is the sum of the amounts reported separately for wage or salary income; net self-employment income; interest, dividends, or net rental or royalty income or income from estates and trusts, Social Security or Railroad retirement income, Supplemental Security Income, public assistance or welfare payments, retirement, survivor, or disability pensions, and all other income. Minimum amount of income is 21500 dollar and Maximum is 3209000 dollar per year. The analyse shows the average of incomes stands on 57000 dollar per year. Big difference in minimum and maximum amount of income refers to large class distinctions in American society. Also IQR is 71600 that shows the large difference amount of income between leisure and middle classes. We can see that income distribution is skwed. Income is a vital measure of general economic circumstances. Also income is used to understanding the relationship between the welfare and type of house holding.
- **Grs.rent** Gross rent is the contract rent plus the estimated average monthly cost of utilities and fuels. Gross rent is intended to eliminate differentials that result from varying practices with respect to the inclusion of utilities and fuels as part of the rental payment. The average rent is about 1000 dollar per month and maximum is 5000 dollar. The missing data category is more than 75% that shows the data is not stable completely. Also gross rent provides information on the monthly housing cost expenses for renters. When the data is used in conjunction with income data, the information offers an excellent measure of housing affordability and excessive shelter costs. Results show the standard of living in each estate and broadness of living abilities. Distribution of rent is alsı unnormal. The number of Max, min, and IQR says that the data suffer from outliers.
- **Property.value** The next variable refers on property value. The question was asked at housing units that were owned, being bought, vacant for sale, or sold not occupied at the time of the survey. Value is the respondent's estimate for selling price. The maximum value is 6,308,000 dollar and the average value shows 180,000 dollar. The difference

between most expensive property and average value, about six million dollar, shows the existence of outliers. 42% of data are missing value that should consider in analysis.

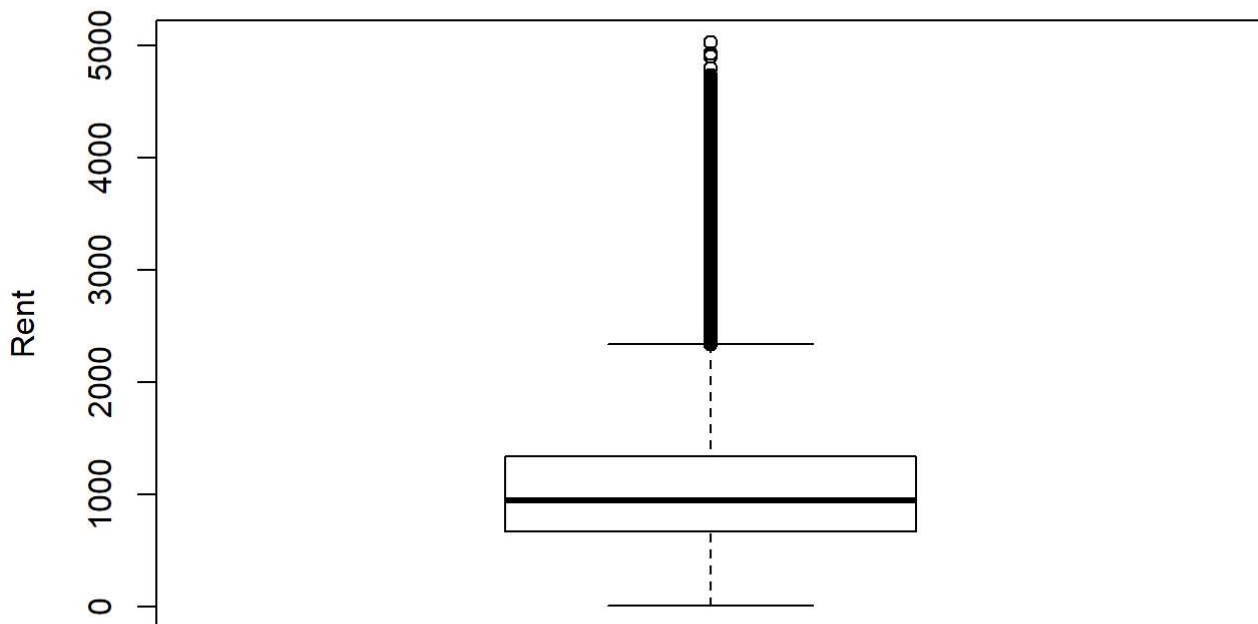
- **Owner.Costs** Selected monthly owner costs provide information on the monthly housing cost expenses for owners. Selected monthly owner costs are the sum of payments for mortgages, deeds of trust, contracts to purchase, or similar debts on the property; real estate taxes; fire, hazard, and flood insurance on the property; utilities; and fuels. This data used to determine living costs in each estate separately and finding the lowest and highest range of costs. The table shows the average cost is 1000 dollar per month. Also rate of IQR is 1151 and maximum figure stands on 13400. When the data is used in conjunction with income and gross rent data, the information offers an excellent measure of housing affordability and excessive shelter costs. The missing data is less than 50% in this survey.
- **No.rooms** Rooms provide the basis for estimating the amount of living and sleeping spaces within a housing unit. These data allow officials to plan and allocate funding for additional housing to relieve crowded housing conditions. The statistics on rooms are in terms of the number of housing units with a specified number of rooms. Survey found that the house in United States possess up 30 rooms in some houses. The intent of this data is to count the number of whole rooms used for living purposes. For each unit, rooms include living rooms, dining rooms, kitchens, bedrooms, finished recreation rooms, enclosed porches suitable for year- round use, and lodger's rooms. Also results determined that in average 6 rooms are in house. According to the 90% validity, it serves to aid in planning for future services and infrastructure, such as home energy assistance programs and the development of waste treatment facilities.
- **when.Moved** The next data refers to year householder moved into unit. These data help to measure neighbourhood stability and to identify transient communities. These variables show the year of the latest move by the householder. The intent is to establish the year the present occupancy by the householder began. The half number of householders, about 50%, stayed at the same place for a period of more than five years, 23% between 5 and 10 years and 26% more than 10 years. The results shows that a few groups of people tend to move each year, 12%. The data also is used to assess the amount of displacement caused by floods and other natural disasters through the states. Also it can be considered as an aid to evaluate the changes in service requirements.
- **Householder and No.children(Unmarried and Children)** The data on relationship to householder were derived from answers to Question 2 American Community Survey (ACS), 2017. From responses to this question numbers of related children, own children, unmarried partner households, and multigenerational households. An unmarried-partner household is a householder other than a married- couple household that includes a householder and an unmarried partner. An unmarried partner can be of the same sex or of the opposite sex as the householder. According to the data 94.5% of householder is in relationship. About 0.2% of male householder lives with male partner and 2.6% with female partner. The same range is seen for female householder, 0.2% lives with female partner and 2.6% with male partner 0.2%. Also Children include a son or daughter by birth, a stepchild, or adopted child of the householder, regardless of the child's age or marital status. The statistics used to determine the plus cost related to number of children that the householder should be pay in compare with non children householder. the maximum number is 18 which is much more than IQR and this shows there is outlier in this data. This survey tried to find that by increasing the number of children the gross rent would be increased or not.

Household Income Data



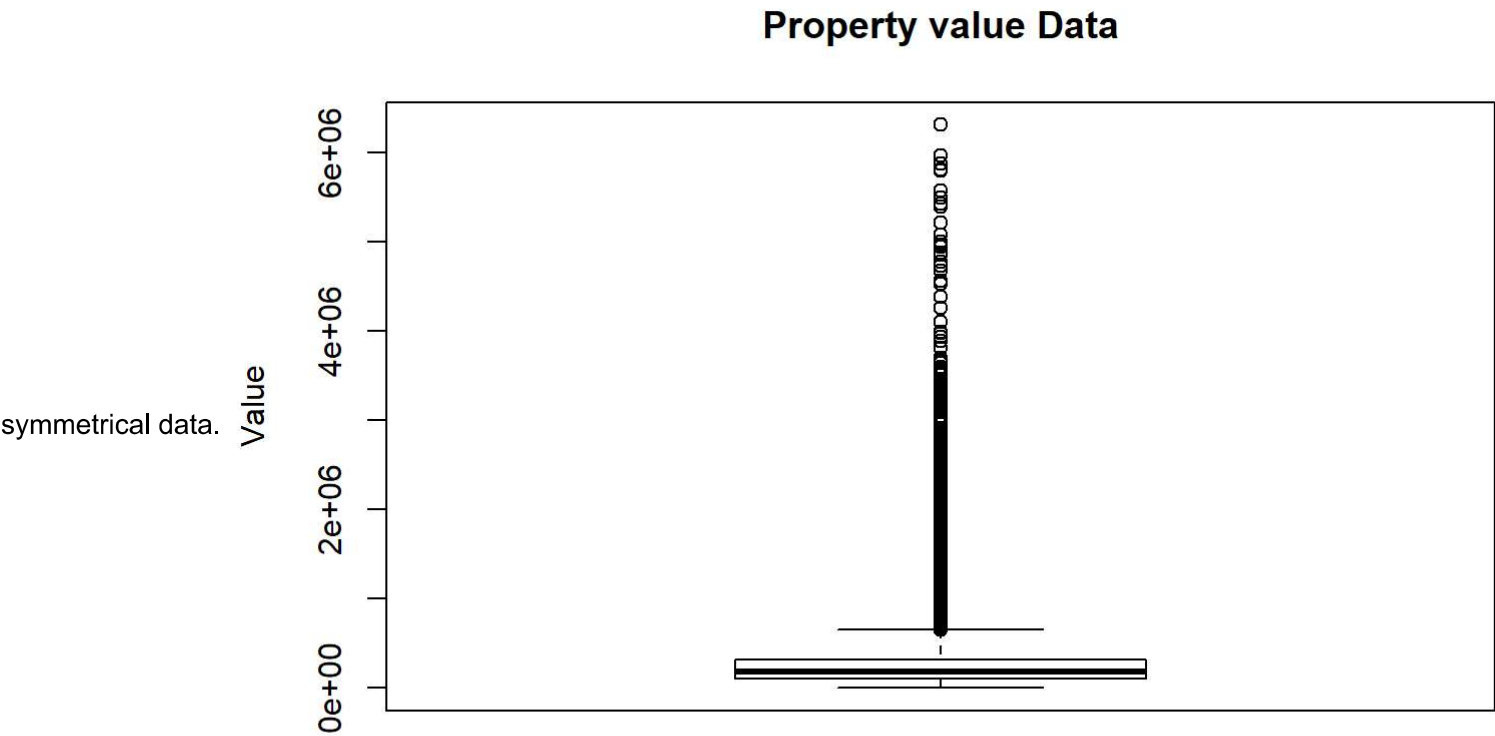
Boxplots are an excellent way to identify outliers and other data anomalies. The box plot of Household Income Data shows that there are also lots of outliers.

Gross Rent Data



In this box plot

we can see lots of dots beyond the extreme line that shows potntial outliers in Gross Rent Data. Also here we do not have a



In Property value Data , there are outliers too. we can see that it is a skewed data.

Methodology

Missing data can reduce the statistical power of a study and can produce biased estimates, leading to invalid conclusions. In this analysis in order to handle missing values Listwise deletion is used because of its simplicity and comparability across analyses. However, this method has some disadvantages. As the summary table illustrates about 70% of gross rent are missing. this percentage for household income and property value is about 40%. It seems removing missing value leads to reducing statistical power. Also, in this method we don't use all information and estimates may be biased.

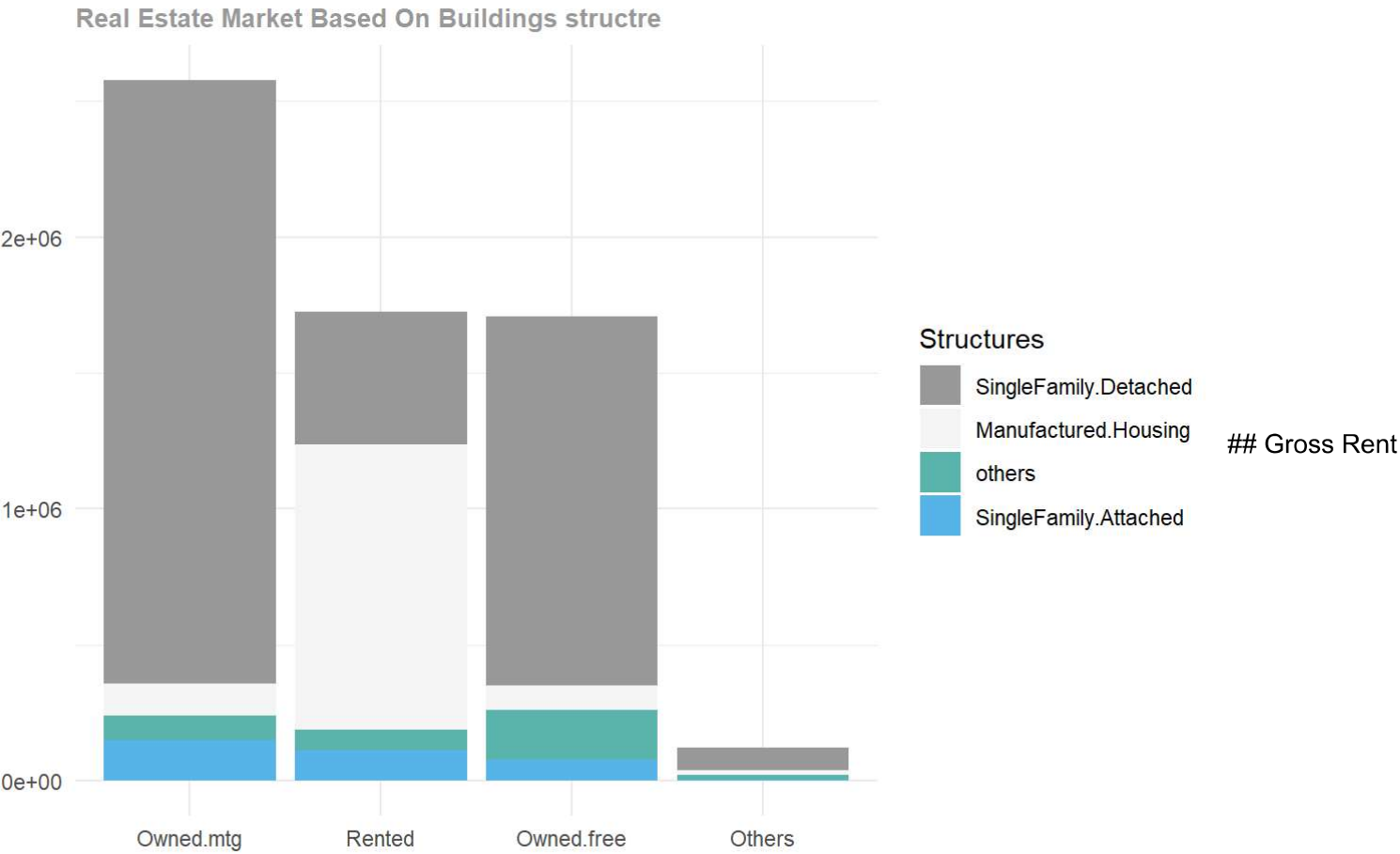
Outliers Most parametric statistics, like means, standard deviations, and correlations, and every statistic based on these, are highly sensitive to outliers. And since the assumptions of common statistical procedures, like linear regression and ANOVA, are also based on these statistics, outliers can really mess up our analysis. Regarding the plot, boxplot is the best for presenting the outliers. As the previous box plots shows, in household income, gross rent and property value datasets, we have many outliers. In this document, at first outliers are detected and then the rows containing the outliers were removed. Removing outliers from data can be good because they are not always practical in certain sets of data. However, the removal of them can be impractical as the data might not show the true results. On the other hand, removing them can be a good thing as it can provide us with the ability to perform statistical tests on data, which in turn can give us a better understanding of the data.as we will see below the impact of outliers on the result of regression.

Association between Buidings structures and Tenure

Here in order to check if there is an association between these two categorical variables, chi squared test is used. As the result shows There is sufficient evidence ($P\text{-value} < 0.05$) that Tenure and type pf structures are associated.

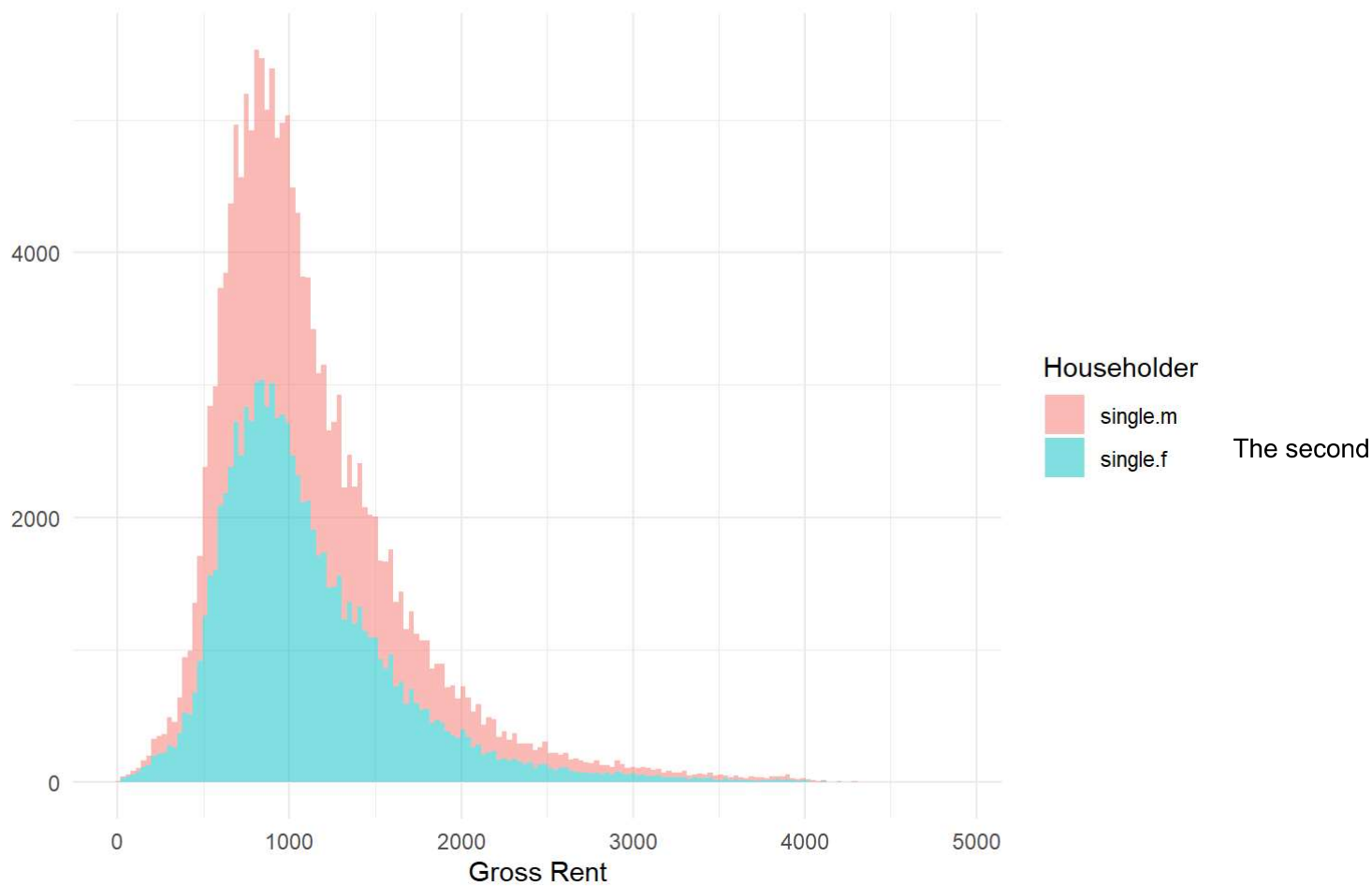
statistic	p.value	parameter	method
2546721	0	9	Pearson's Chi-squared test

The plot shows that most of Americans prefer to buy a single-family house on the other hand apartments and manufactural building are popular among renters. We can also see that most of householders live in their owned home and the number of owners who have mortgages is much more than who doesn't have. Finally we can add that the detached units are the most popular one among Americans.

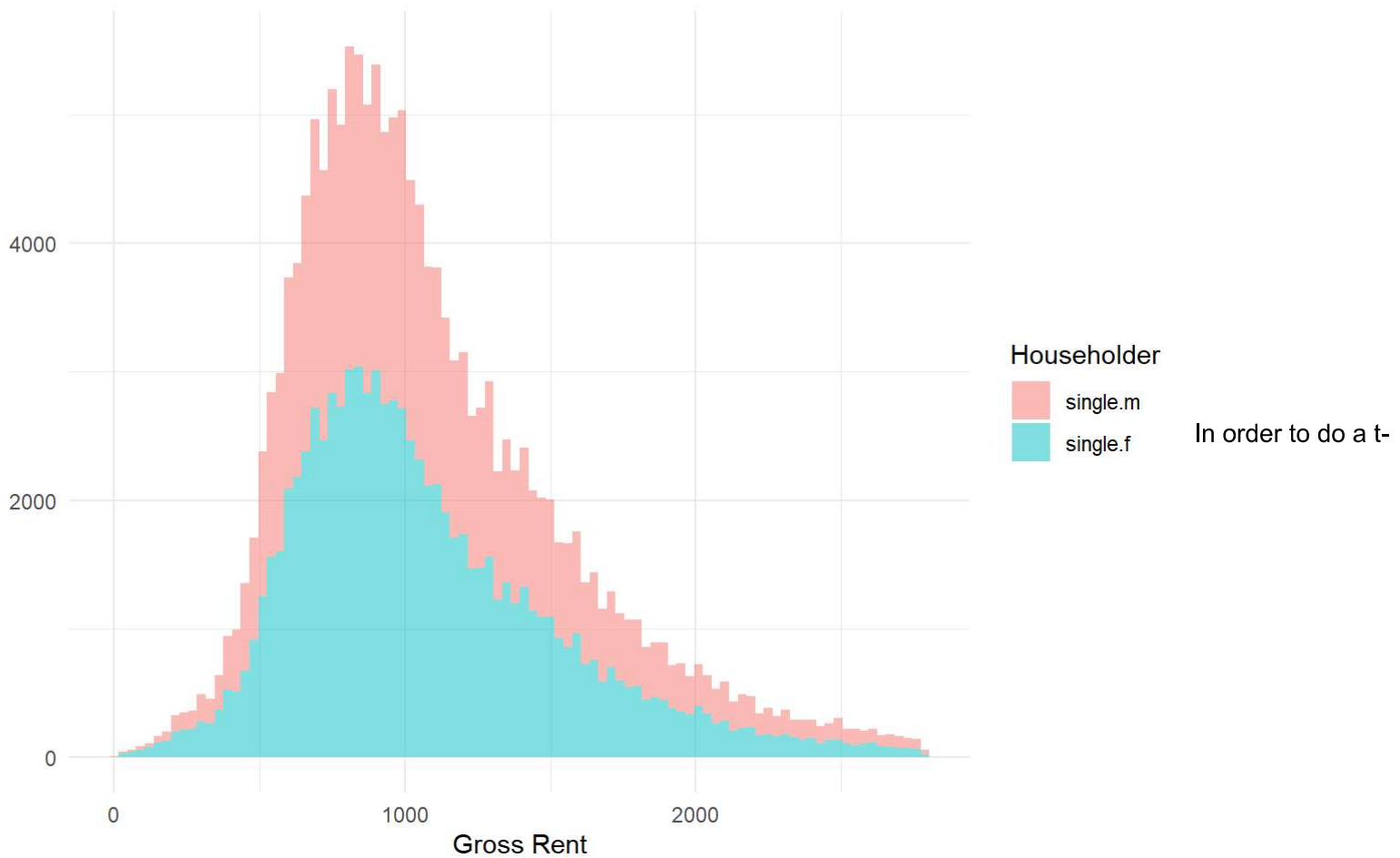


for single males and single females.

Here we want to know if the mean Gross Rent for single females is significantly different from that of single males. So we can do a t-test to find the answer. At first, we check the normality of data. The plot shows that we do not have a normal dataset.As we saw in the begining we have lots of outliers in Gross Rent Data. So the normality was checked after removing outliers.



histogram has less skewness and kurtosis but it does not seem as a normal distribution yet.



test we need to calculate the variance of each group. Because the data is not normal and it is a big data the levene test is used to check if the variance of each group is equal or not . here the p-value is very small (less than 0.05) so we can say that the variances are not equal.

term	df	statistic	p.value
group	1	102.4981	0
	161366	NA	NA

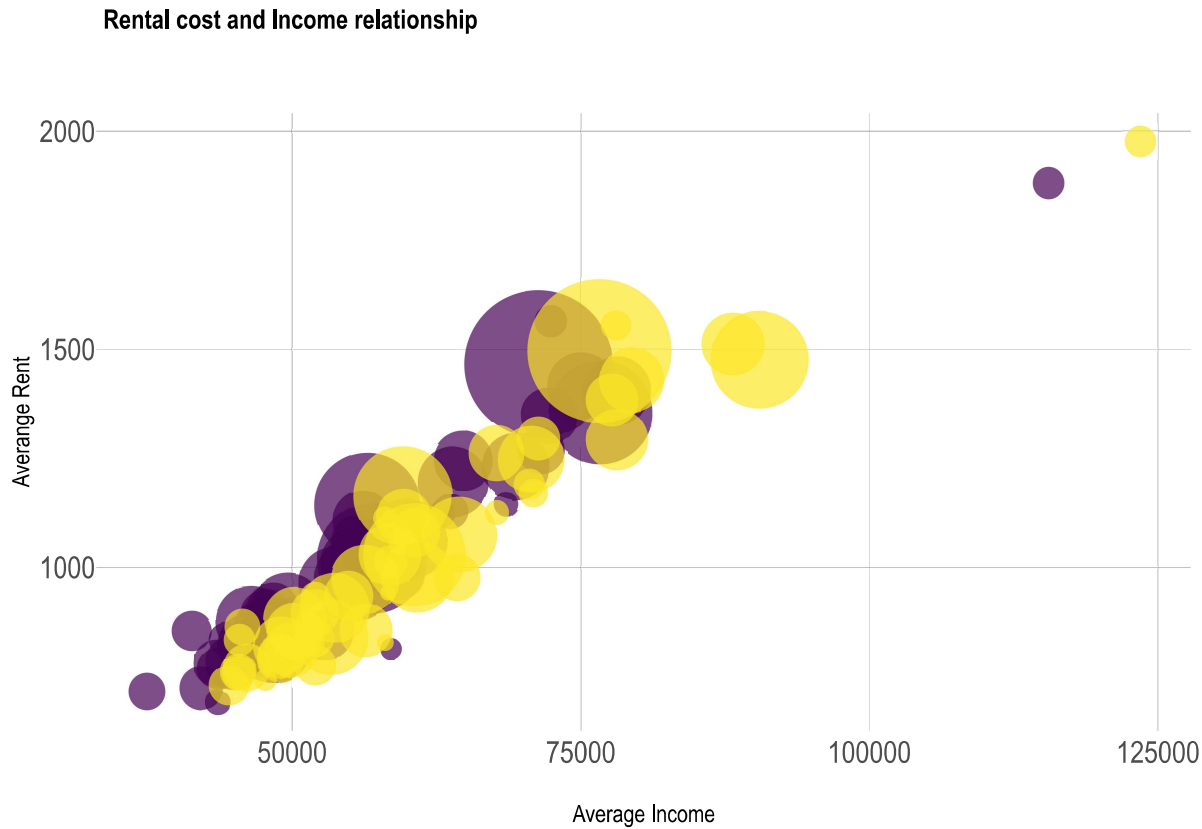
Finally we can do a t-test. We obtained p-value less than 0.05, then we can conclude that the mean of gross rent for two groups are not similar.

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
25.46736	1096.294	1070.827	10.77862	0	153827.9	20.83639	30.09833	Welch Two Sample t-test	two.sided

Single Family Renters.

This bubble plot shows the Gross Rent to Income ratio for single males and single females for each State. At first glance, we can see a positive relationship between income and rent cost so we can say people with higher incomes prefer to live in a place with higher rent maybe with better conditions and the ratio is the same for all states and for both groups the trend is same. this plot also helps us to know what is the States in the extreme part of the graphic, or what is the one out of the general trend. For instance, Average Rent and average Income for the district of Columbia are much higher than in other states. Also,

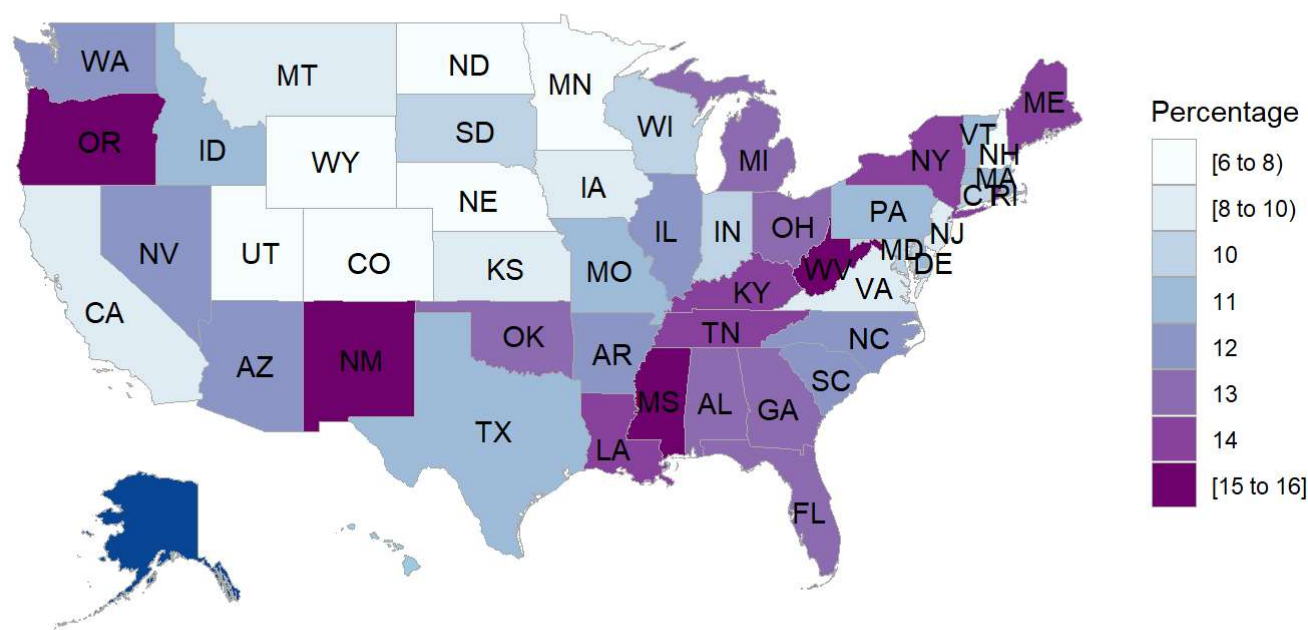
the biggest bubbles are related to California, New York, Texas so we can say in these States there are many more single families than other states. another interesting thing is that the differences between average income for single males and females in California are high we can see the same thing in Texas.



Supplemental Nutrition Assistance Program Benefits

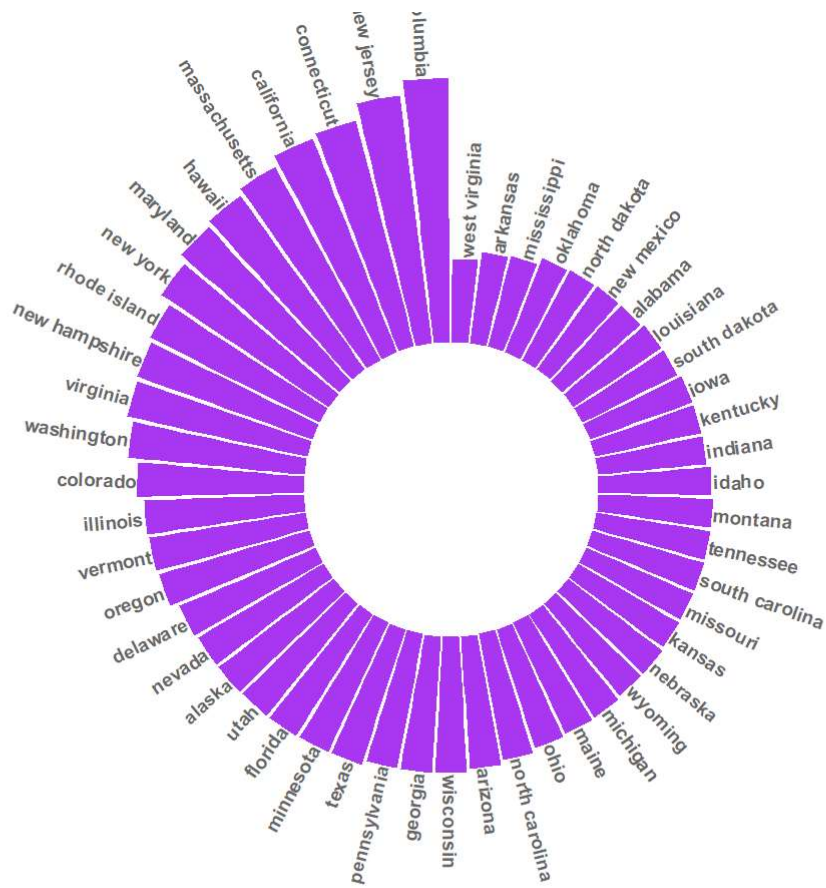
SNAP is a program for low-income households to obtain a more nutritious diet. Here the percentage of households received this program for each state are calculated. We can say states like Wyoming, Utah, Colorado, Nebraska a smaller percentage of households benefit from this program. At first glance 4 white states in the center of map attract attentions . white color says that just 6 to 8 percentage of household in these states receiving SNAP which means there are fewer low-income household but It seems better to compare the interpretation of this plot with poverty rate for each state because if these states are poor it means that the SNAP does not work properly there.

Percentage of Households recieving SNAP Per State



Owner Cost

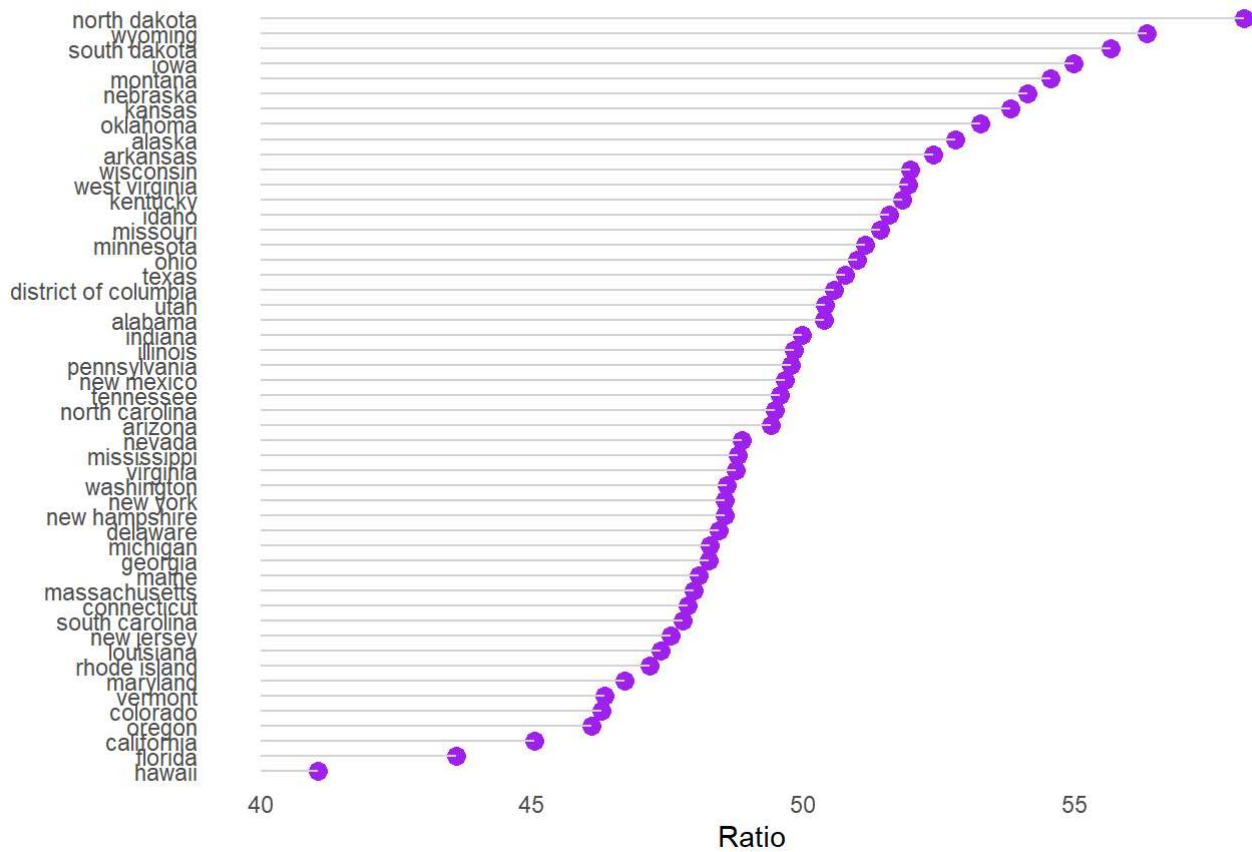
This plot compares Owner Cost for each state. We can say that Owner Cost in District of Colombia, New Jersey, Connecticut, California, Massachusetts is much higher than others. On the other hand, states like West Virginia, Arkansas, Mississippi and Oklahoma have the lowest Owner Cost.



Rent-to-Income Ratio

In the plot we can see In North Dakota, Wyoming, South Dakota, and Iowa more than 55 % of household income is spent on renting a house which is a big number. Hawaii, Florida and California are the best states to live for tenants.

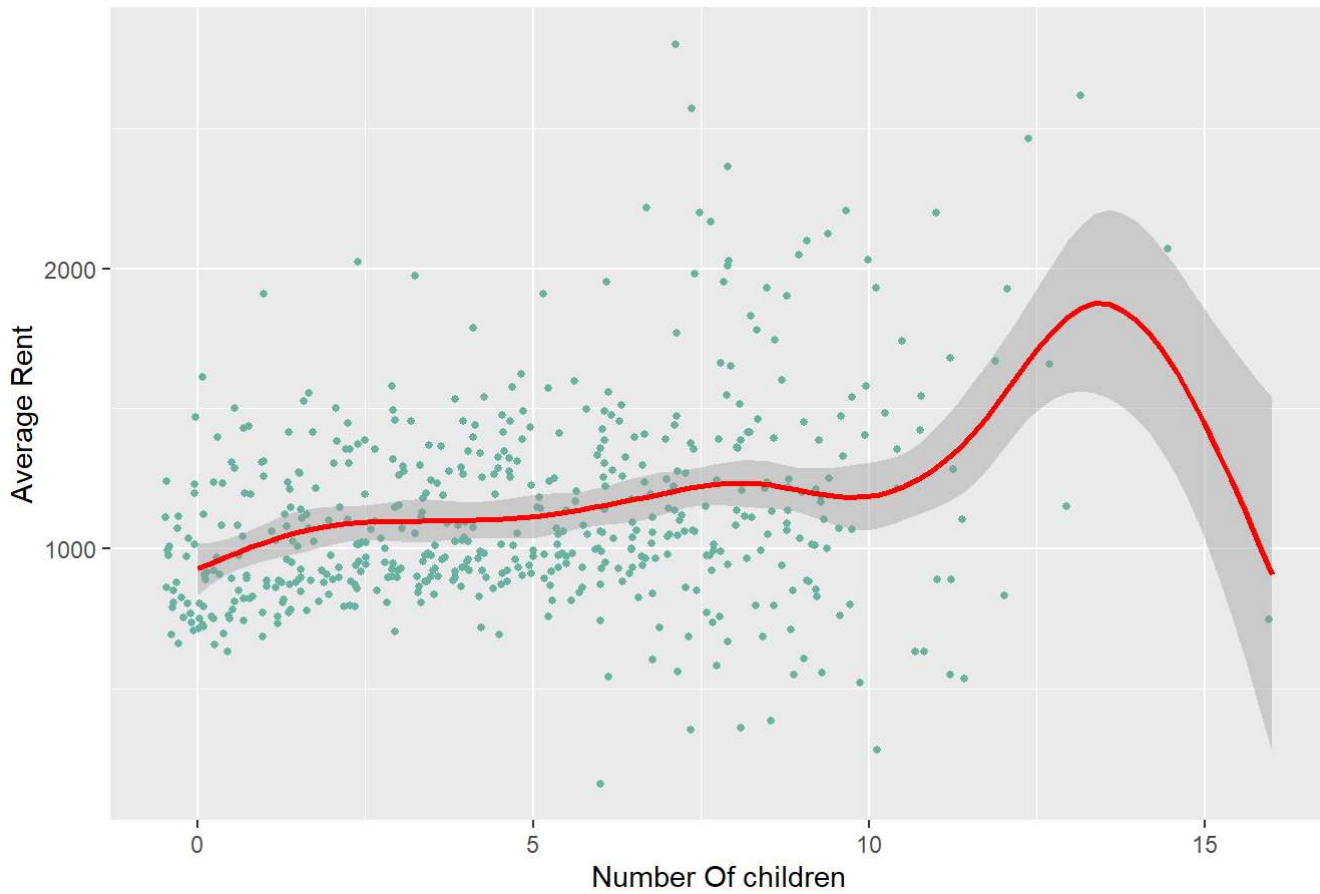
Rent-to-Income Ratio



Relationship between Number of children and Rent Cost.

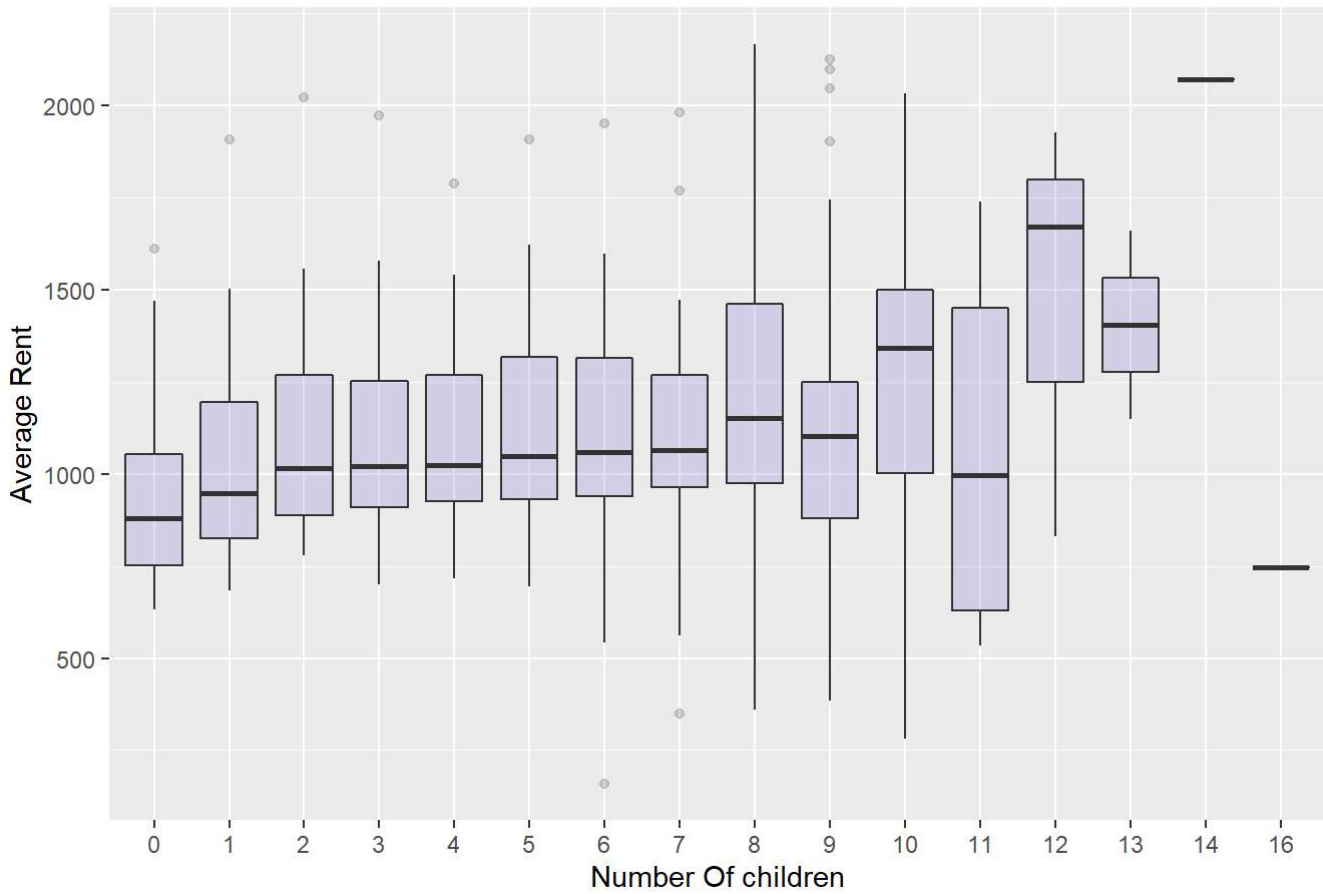
The plot shows the relation between these two variables.

Relationship between Number of children and Rent Cost in the USA



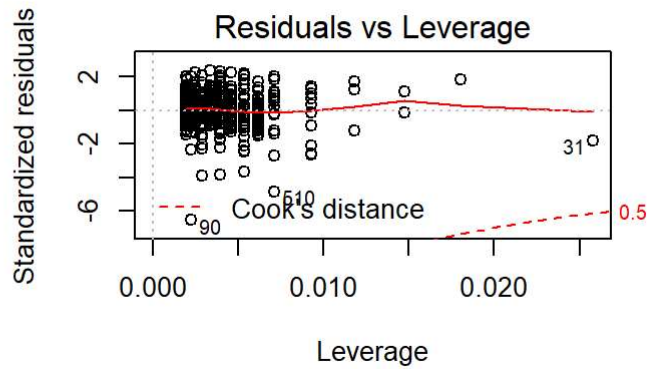
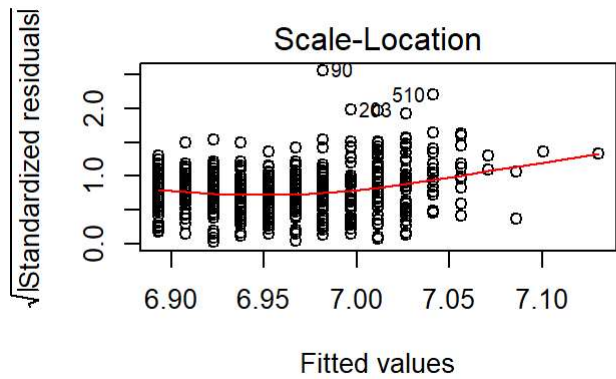
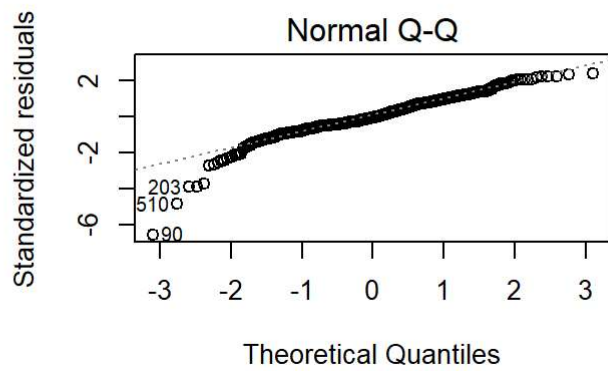
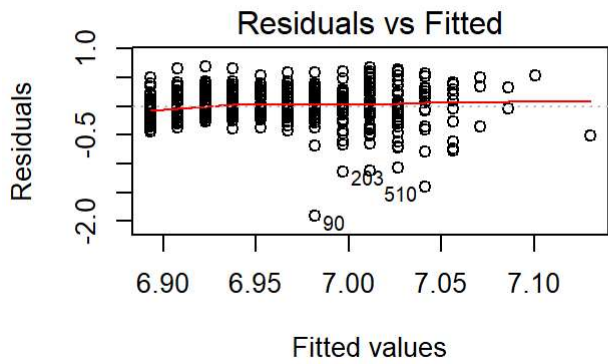
As it was mentioned there are lots of outliers in this dataset so before conducting the regression the outliers was removed.

Distribution of Gros Rent Cost for number of children (without outliers)



Before

conducting a regression model we should check the regression assumptions.



Looking at the **Residuals vs Fitted plot**, we see that the red line is perfectly flat. This tells us that there is no discernible non-linear trend to the residuals. Furthermore, the residuals appear to be equally variable across the entire range of fitted values. There is no indication of non-constant variance.

The second plot **Normal Q-Q** shows residuals are normally distributed because residuals follow a straight line but the residuals deviate from the diagonal line in the lower tail because of outlier.

The Scale-Location plot shows whether our residuals are spread equally along the predictor range. In this case, the red smooth line is not horizontal also the spread around the red line varies with the fitted values which suggest heteroskedasticity.

The last plot is **the Residuals vs Leverage plot** tells us which points have the greatest influence on the regression. Here there are no points outside the dotted line and we can say our plot doesn't show any influential cases.

- The Residuals section : the median is which is close to 0 . However, the 3Q and 1Q are not close to each other in magnitude. The max is and min is that are far from each other in magnitude so here we can see that the distribution of the residuals is not symmetrical.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.8928174	0.0229300	300.602791	0.0000000
No.children	0.0148514	0.0040068	3.706529	0.0002328

- Coefficients section
 - Estimate: Intercept means that rent cost for family without children would be 689.2817378 and No.children that consider slope of our regression line means that for every 1 children the rent cost increases by 0.0148514.
 - std. Error: is saying that rent Cost when there is no children can vary from 6.8928174 by 0.02293. Besides, the increase of rent cost for a children can vary from 0.0148514 by 0.0040068.
 - t value and Pr(>|t|): based on these two columns we can say there is a relation between the number of children and rental cost as p-value for t-test is $2.32798110 \times 10^{-4}$ and we can reject null hypothesis as it is less than .05 .here p-value is a very small value that means number of children is probably an excellent addition to our model.
- Residual standard error: the actual rent cost for all of children can deviate from the true regression line by approximately 0.2919337, on average.
- Multiple R-squared : here R^2 is 0.0258854 which means the model can explain 2.5885362% of the total variability.
- F-statistic: Because the number of our data is big in order to say there is a relation between number of children and rent cost we do not need a large F-statics. Here it is 13.7383543 also p-value is less than .05 and we can reject null hypothesis that means there is a relation between number of children and rent cost.

Discussion

The survey shows that tenants in some states have financial problems as, in some states the rent cost is half of householders income and short-term contracts are trend in these states .besides it seems that some programs such as SNAP is not executed there. which these two factor cause some economic problems for people who live in these states. However , I think for do a better analysis more factore should be considered.

Incorporating the weight (and inflation adjustment) is essential for getting correct and meaningful results, but because of complexity I ignor these factors.

I checked the relation between number of childrens and rent cost . before starting the anlysis there is no relation between two variables but it seems there are a strong relationship. On the other hand becuse removing outliers and missing values I think may be the model is not as correct as it shows.