**CapstoneProject -He Case Study**

**Project Report**

**Aparna Nair**

**aparna.nair.vijay@gmail.com**

**Business Scenario**

A banking institution requires actionable insights from the perspective of Mortgage-Backed Securities, Geographic Business Investment and Real Estate Analysis.

The objective is to identify white spaces/potential business in the mortgage loan. The mortgage bank would like to identify potential monthly mortgage expenses for each of region based on factors which are primarily monthly family income in a region and rented value of the real estate. Some of the regions are growing rapidly and Competitor banks are selling mortgage loans to subprime customers at a lower interest rate. The bank is strategizing for better market penetration and targeting new customers. A statistical model needs to be created to predict the potential demand in dollars amount of loan for each of the region in the USA. Also, there is a need to create a dashboard which would refresh periodically post data retrieval from the agencies. This would help to monitor the key metrics and trends.

The dashboard must demonstrate relationships and trends for the key metrics as follows: number of loans, average rental income, monthly mortgage and owner’s cost, family income vs mortgage cost comparison across different regions. The metrics are described not to limit the dashboard to these few only.

* **Domain:**Real Estate

**DATASET DESCRIPTION**

Following are the themes the fields fall under Home Owner Costs: Sum of utilities, property taxes.

* Second Mortgage: Households with a second mortgage statistics.
* Home Equity Loan: Households with a Home equity Loan statistics.
* Debt: Households with any type of debt statistics.
* Mortgage Costs: Statistics regarding mortgage payments, home equity loans, utilities and property taxes
* Home Owner Costs: Sum of utilities, property taxes statistics
* Gross Rent: Contract rent plus the estimated average monthly cost of utility features
* Gross Rent as Percent of Income Gross rent as the percent of income very interesting
* High school Graduation: High school graduation statistics.
* Population Demographics: Population demographic statistics.
* Age Demographics: Age demographic statistics.
* Household Income: Total income of people residing in the household.
* Family Income: Total income of people related to the householder.

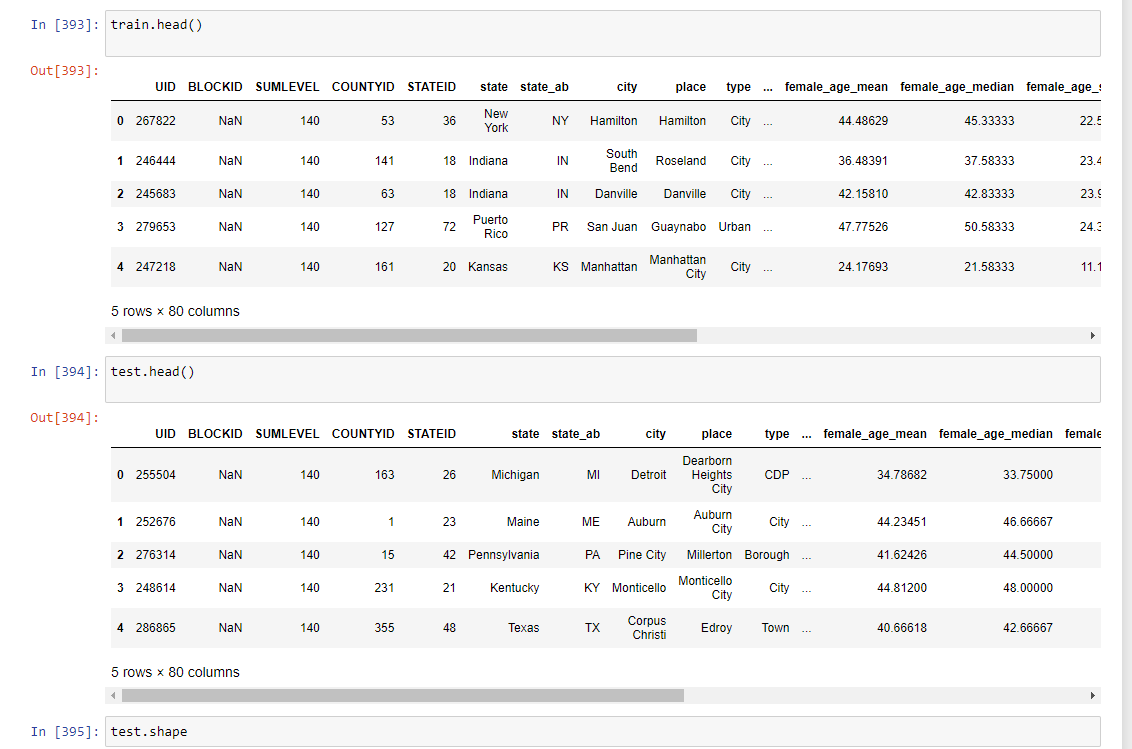
**Statistical algorithm execution – Python code and outputs**

1. Objective1) Import data

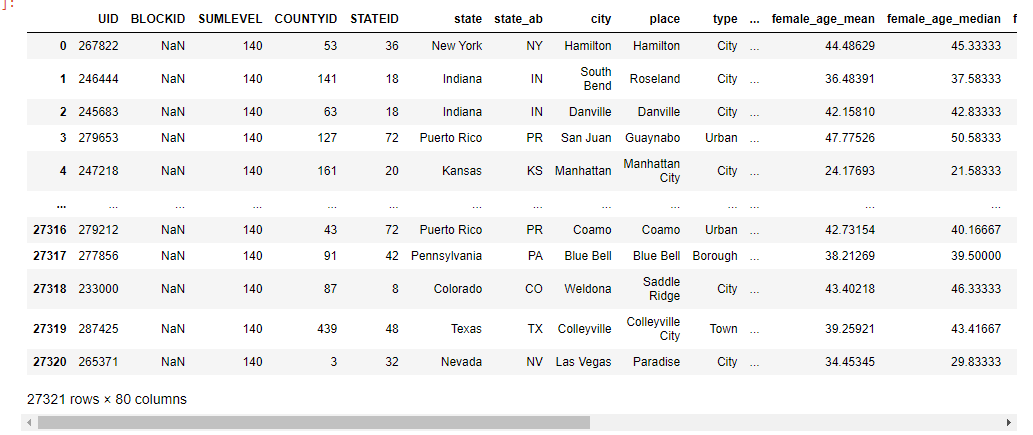


Imported the train and test data

1. 2.Objective 2) Figure out the primary key and look for the requirement of indexing



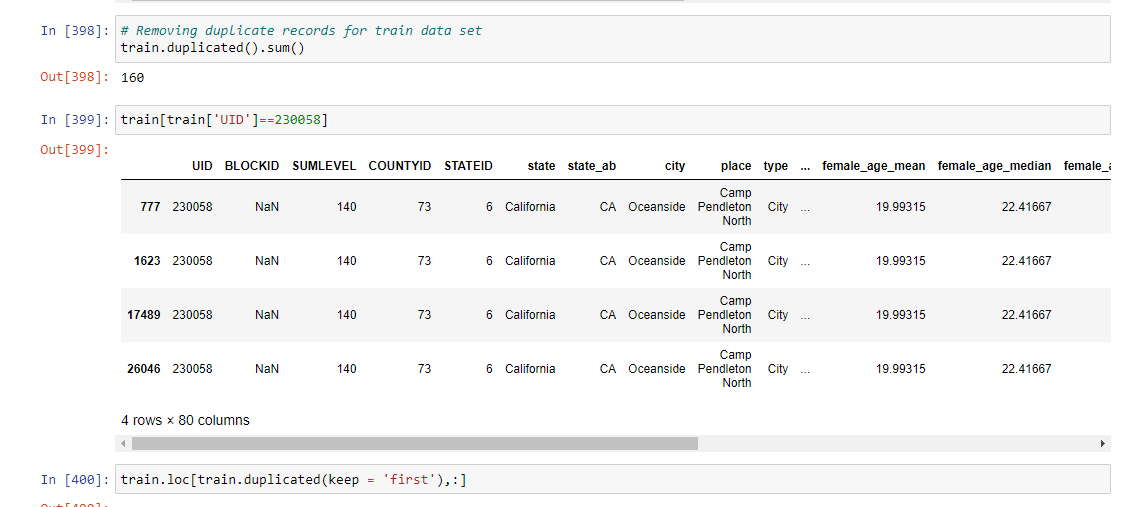


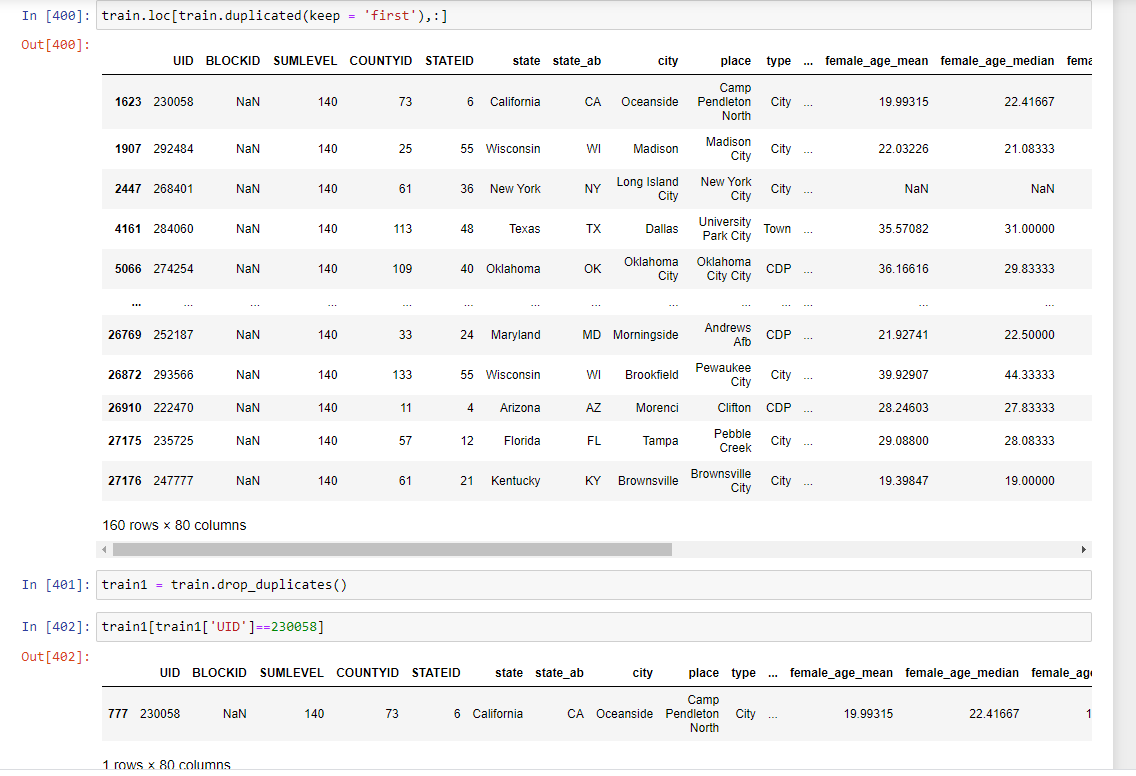


UID is identified as the primary key

3.Objective 3 ) Gauge the fill rate of the variables and devise plans for missing value treatment. Please explain explicitly the reason for the treatment chosen for each variable.

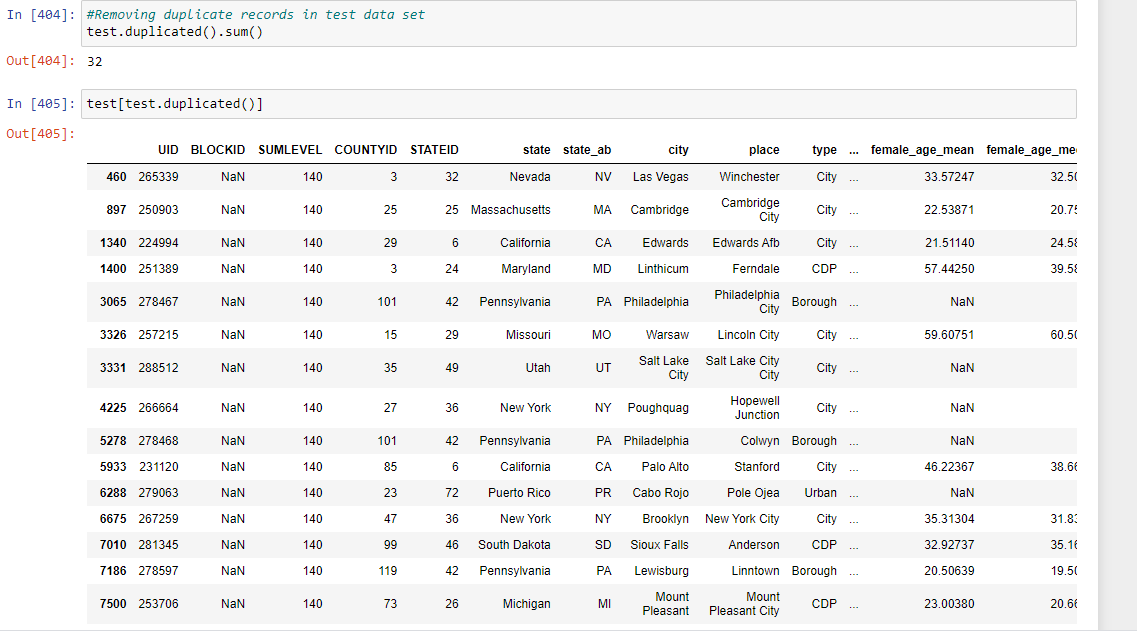
Removing the duplicate rows from train data set

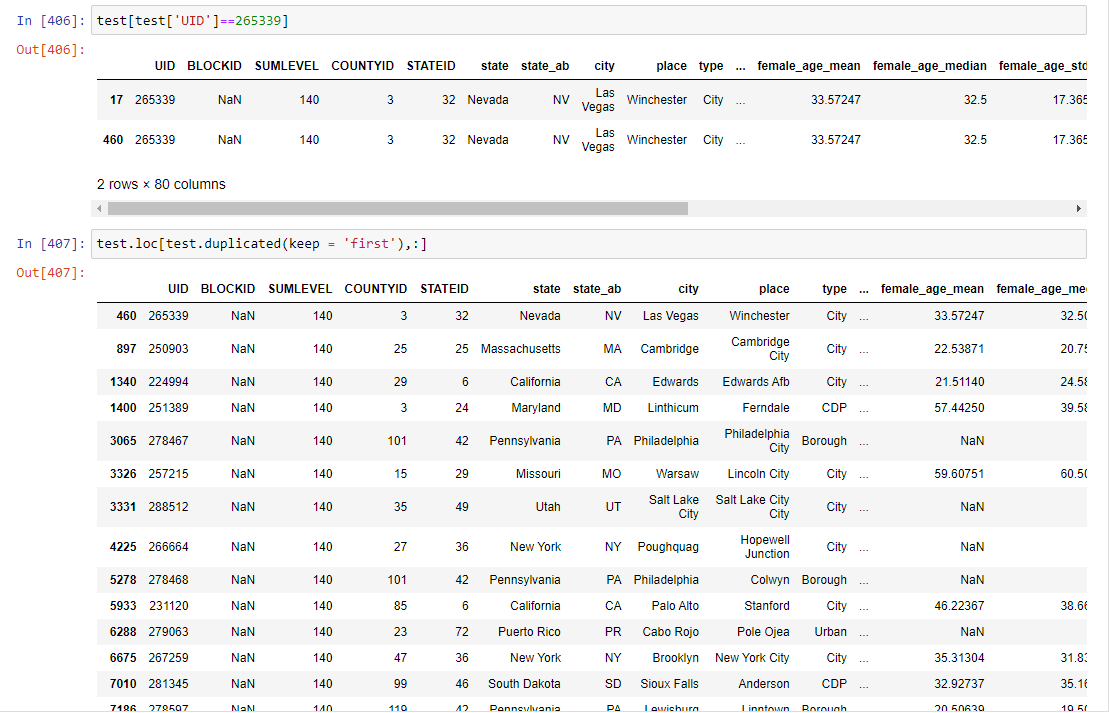


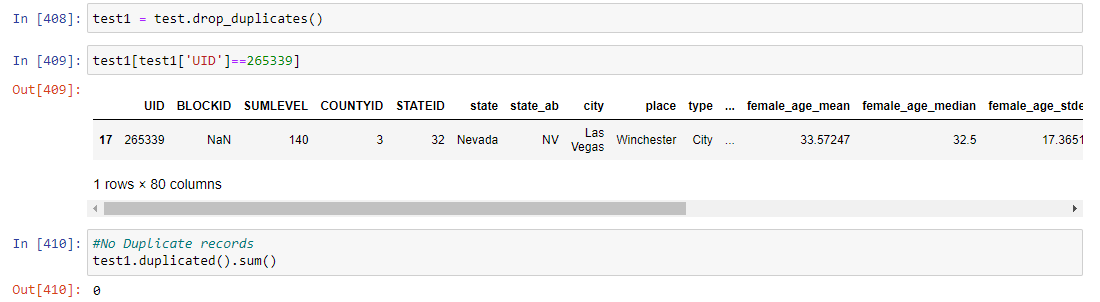


Some rows having same values were repeating . so keeping the first row deleted all the duplicate rows in train dataset

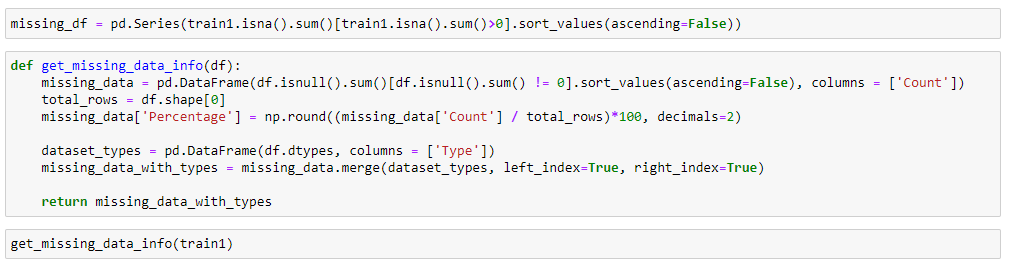
The same was done for test dataset also.

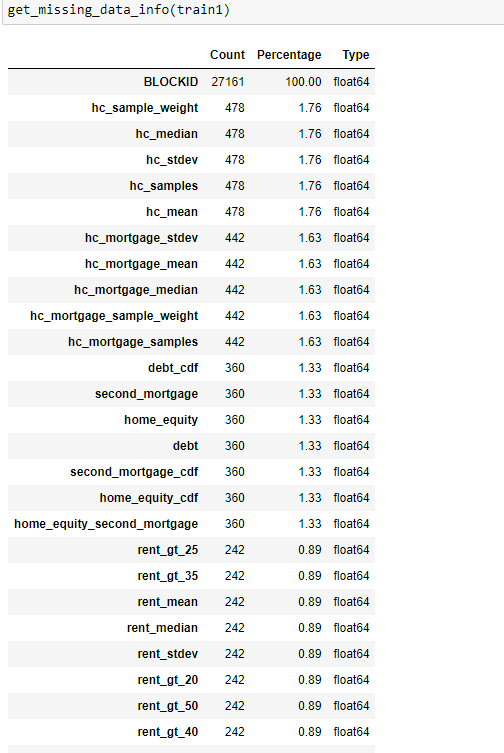




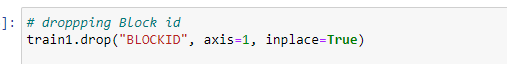


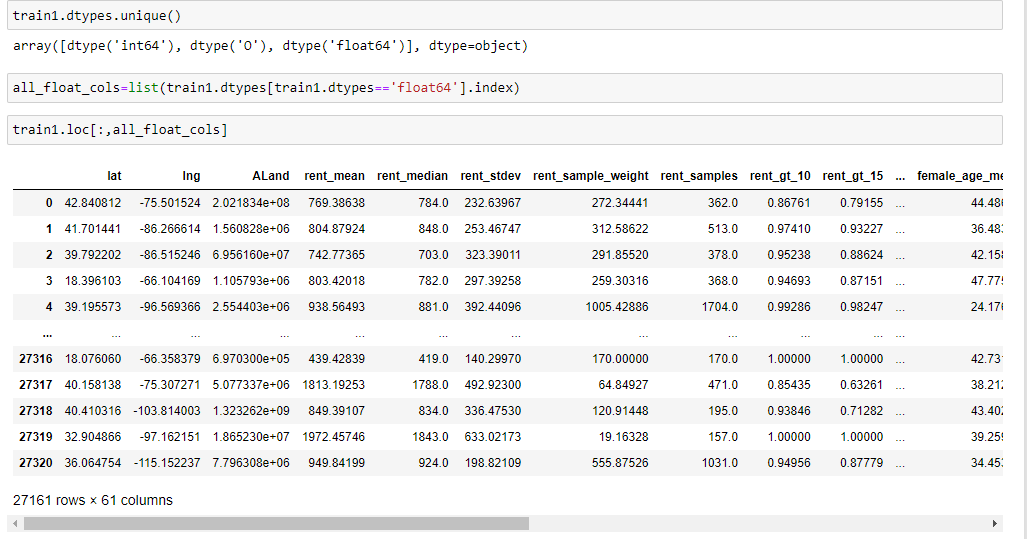
# Finding missing values



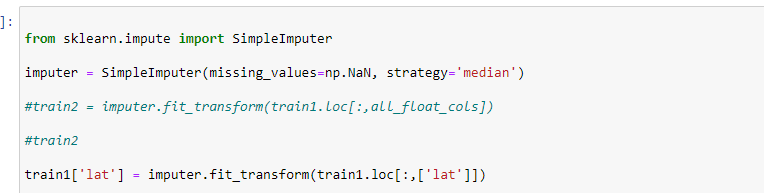


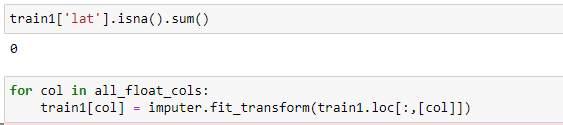
Dropping BLOCKID since it has 100 % missing values





Replacing the missing values with median by using simpleimputer

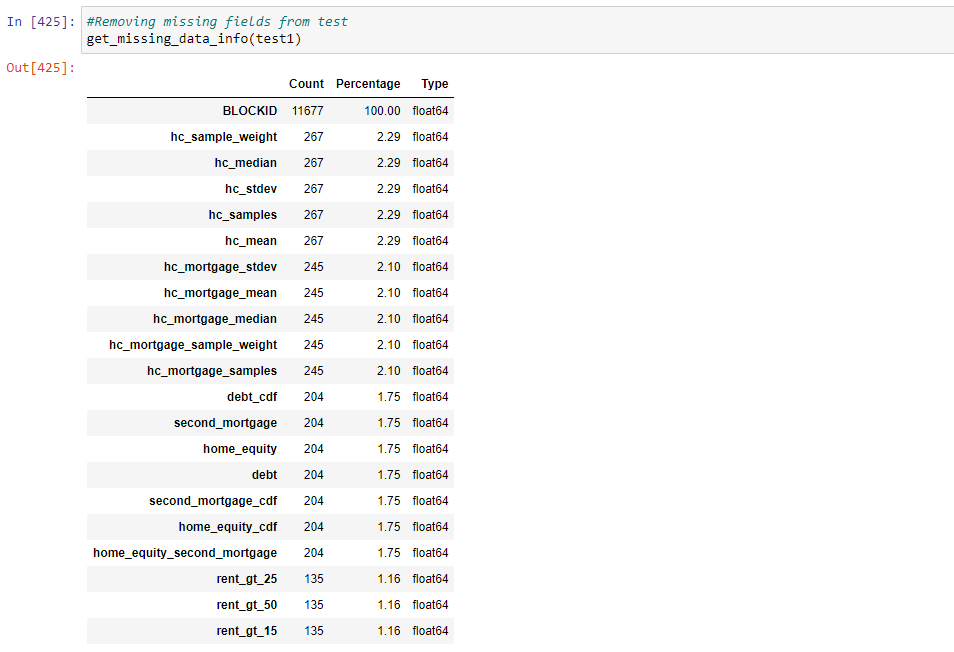






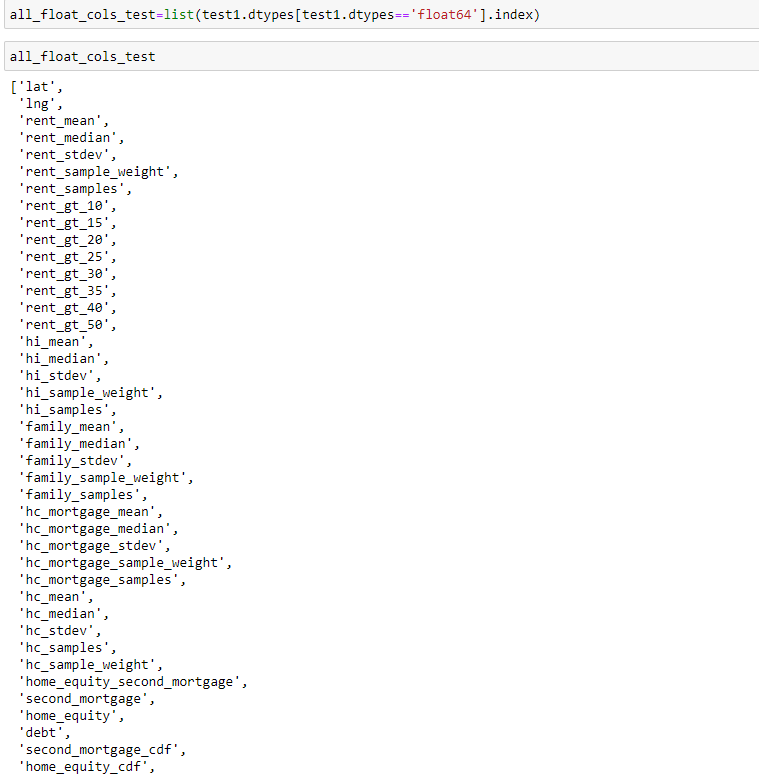
No missing values present in train dataset

Removing missing values in test dataset

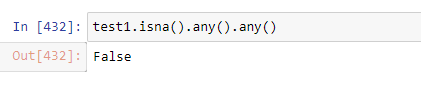


dropping BLOCKID from test dataset









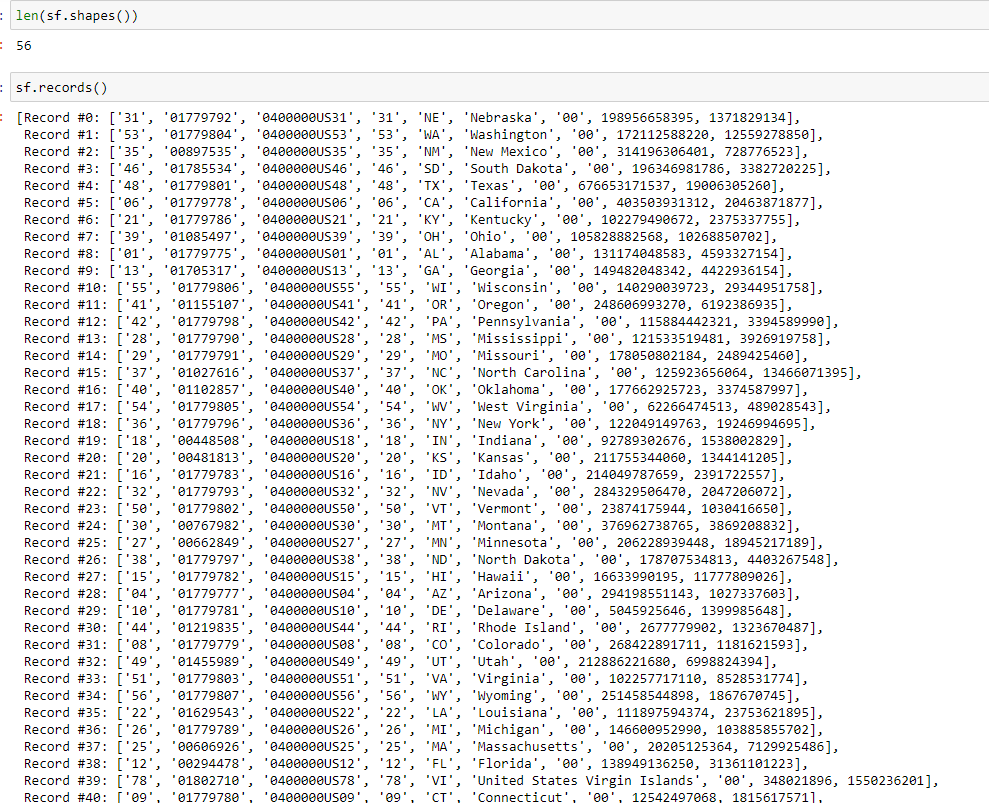
No missing values in test dataset

1. Objective 4) Understanding homeowner costs are incredibly valuable because it is positively correlated to consumer spending which drives the economy through disposable income. Perform debt analysis. You may want to follow the following steps:

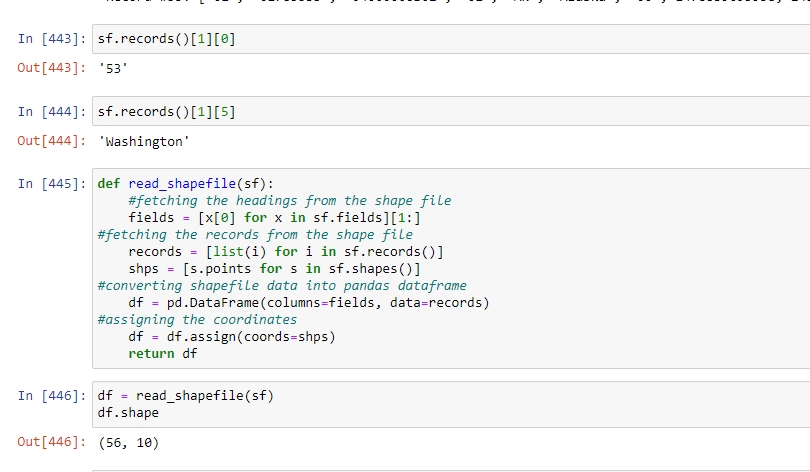
* Explore the top 2,500 locations where the percentage of households with a second mortgage is the highest and percent ownership is above 10%. Visualize using geo-map. You may keep the upper limit for the percent of households with a second mortgage to roughly 50%.

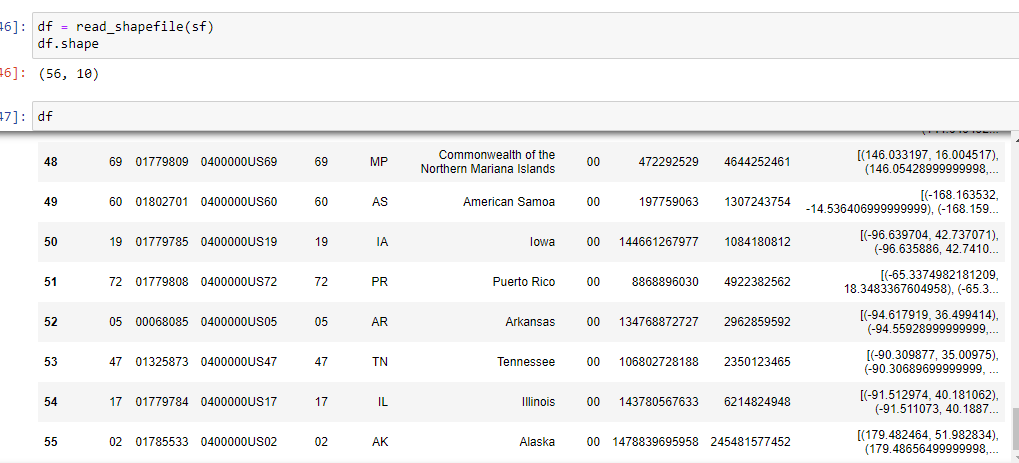
Getting the shape file



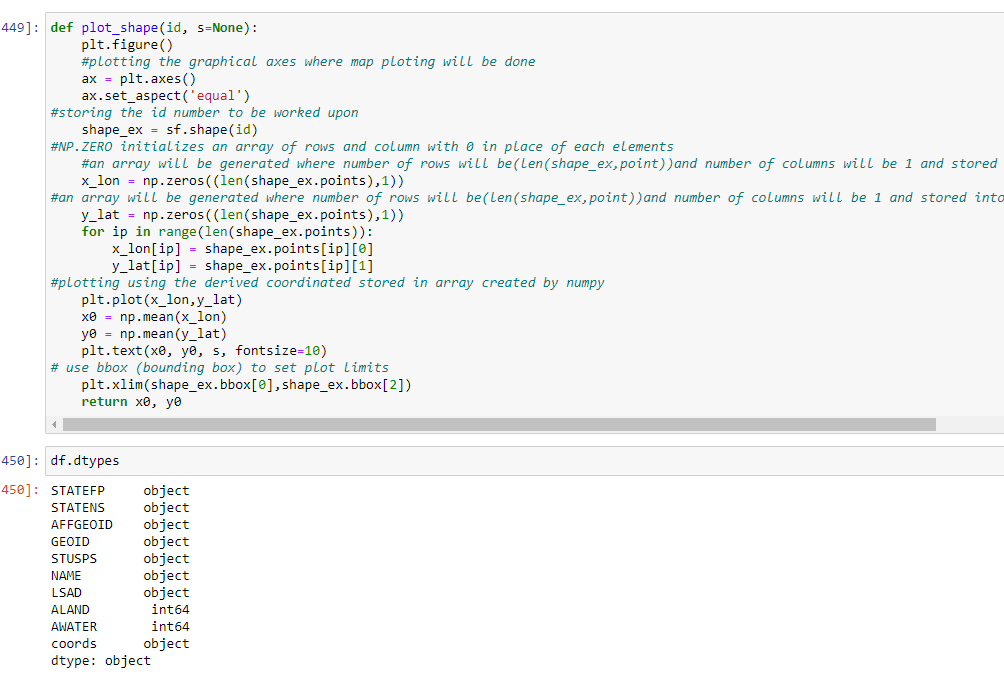


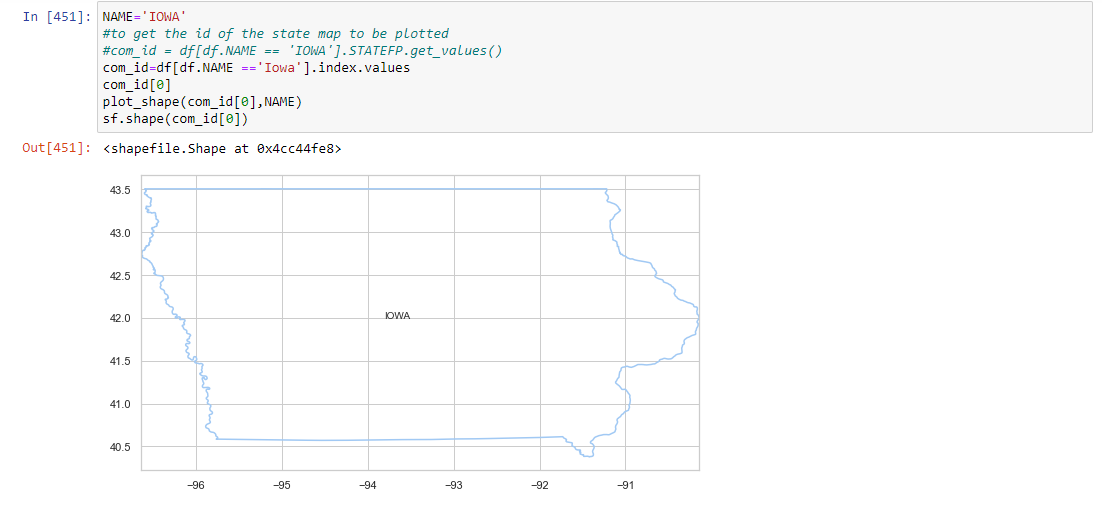
fetching the headings from shapefile and converting it in to a dataframe



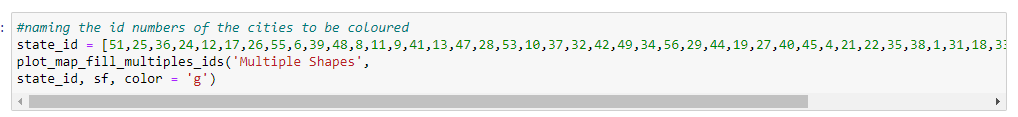


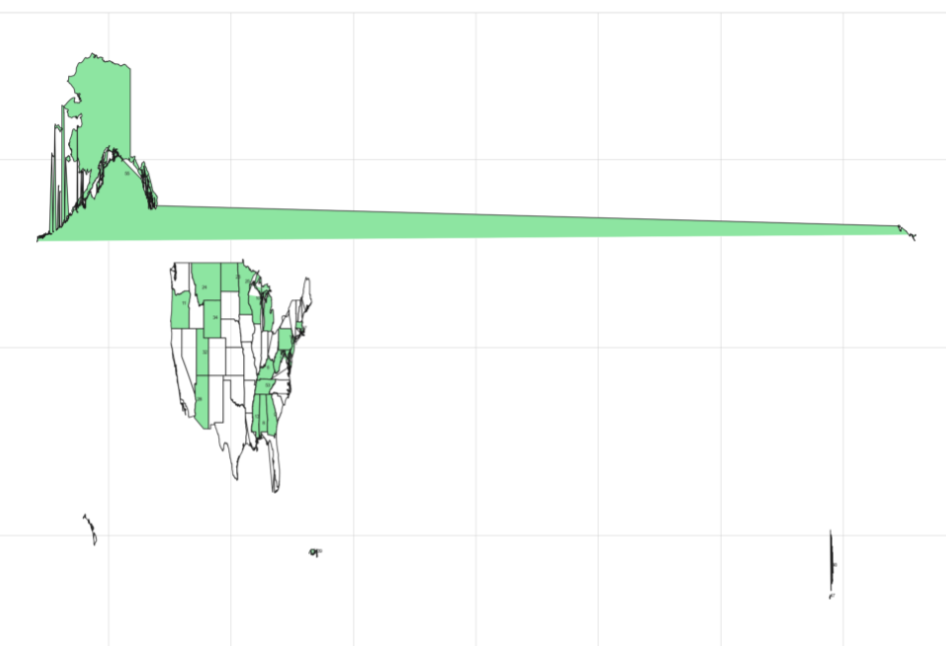
settings for plotting











* Bad debt is the debt you should avoid at all costs such as a second mortgage or home equity loan. Conversely, Good debt is all other debt not including second mortgage or home equity loan.

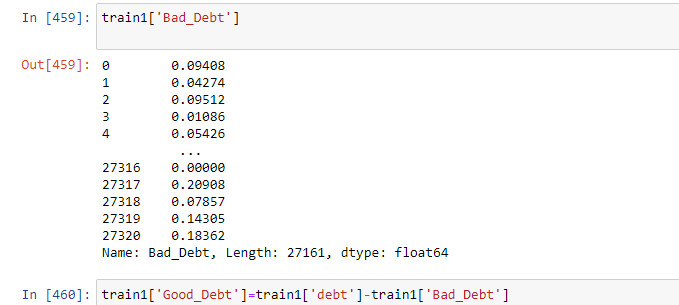
Bad Debt Equation:

Bad Debt = P (Second Mortgage ∩ Home Equity Loan)

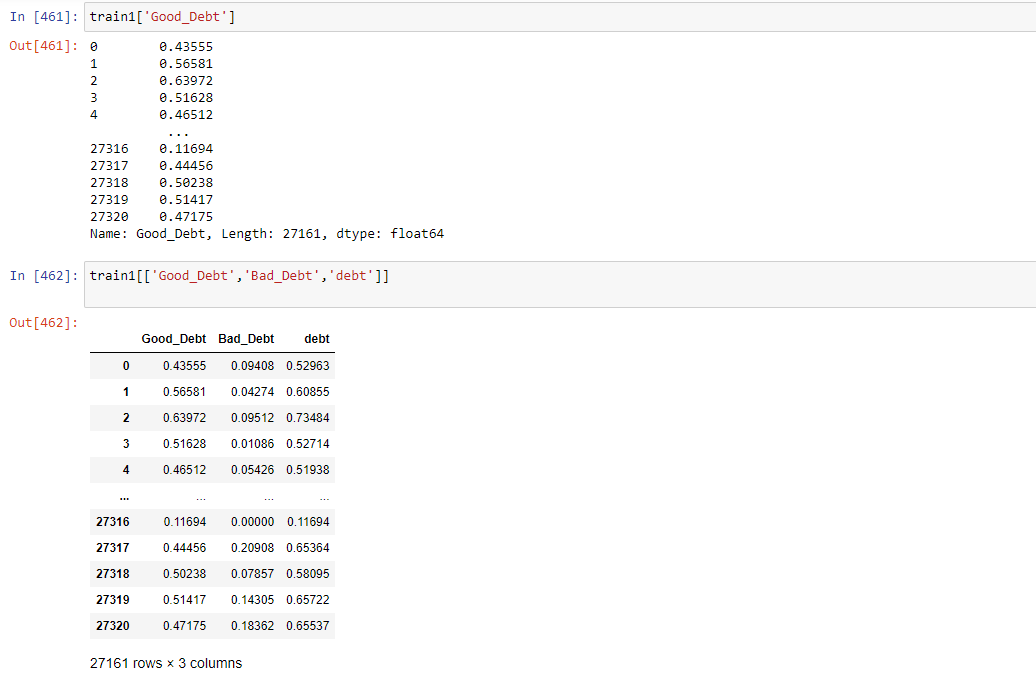
Bad Debt = second\_mortgage + home\_equity - home\_equity\_second\_mortgage

Creating Bad Debt column



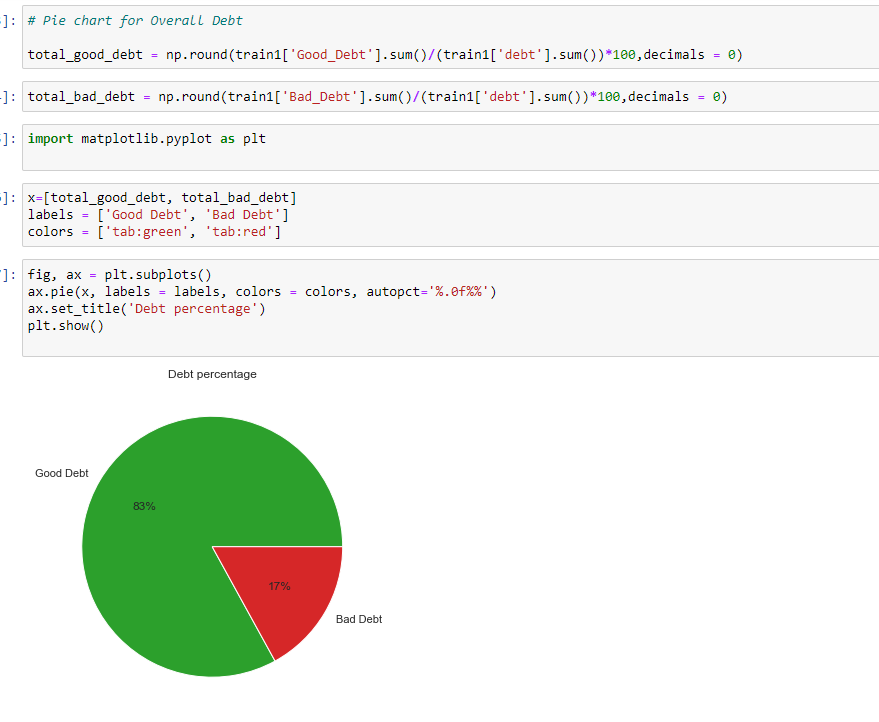


Creating Good debt column

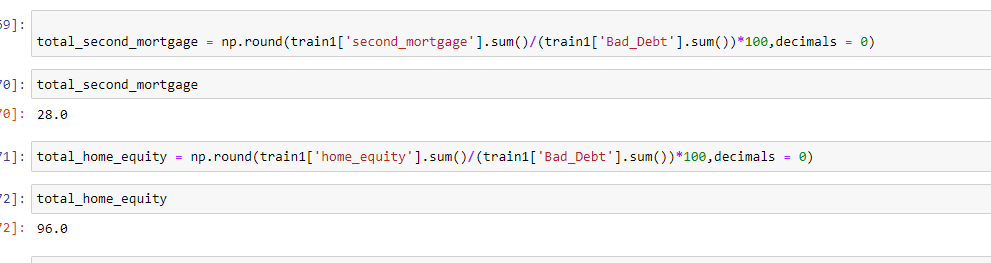


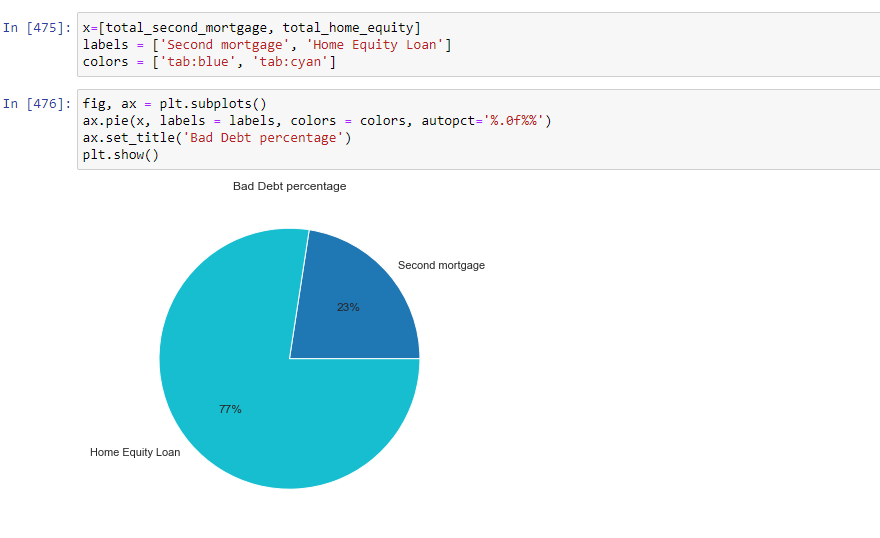
* Create pie charts (Venn diagram) to show overall debt (% bad and good debt) and bad debt (2 mortgage and home equity loan).

Pie chart for Overall debt



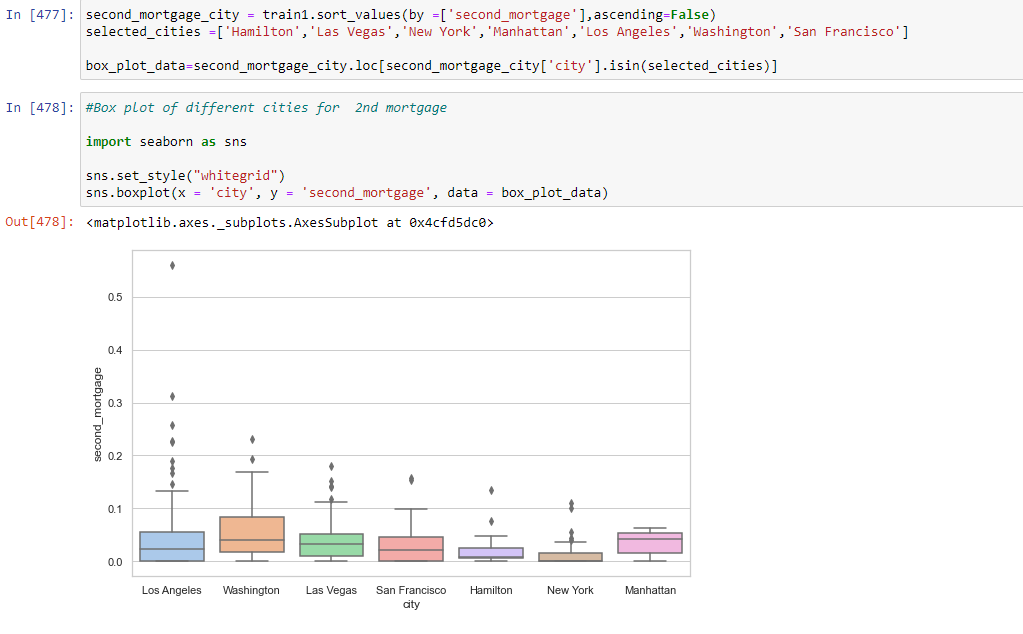
Pie chart for Bad Debt





* Create Box and whisker plot and analyze the distribution for 2nd mortgage, home equity, good debt and bad debt for different cities.

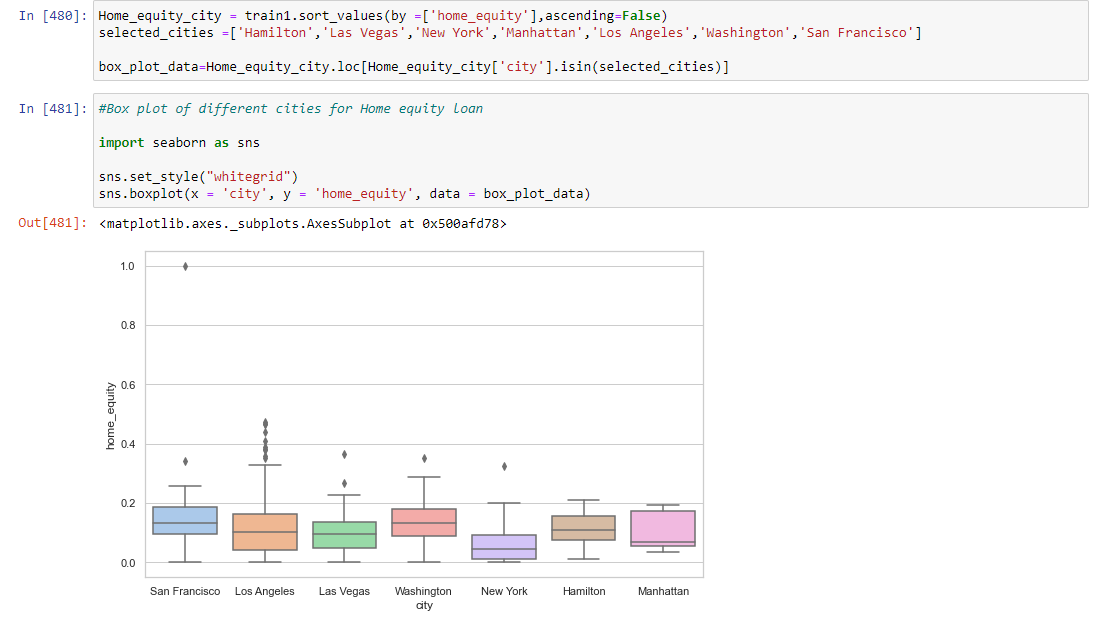
Box plot for Second mortgage



# Los Angeles is having lot of outliers

#Median second\_mortgage for Washington and Manhattan is highest among the observed cities

Box plot for Home equity

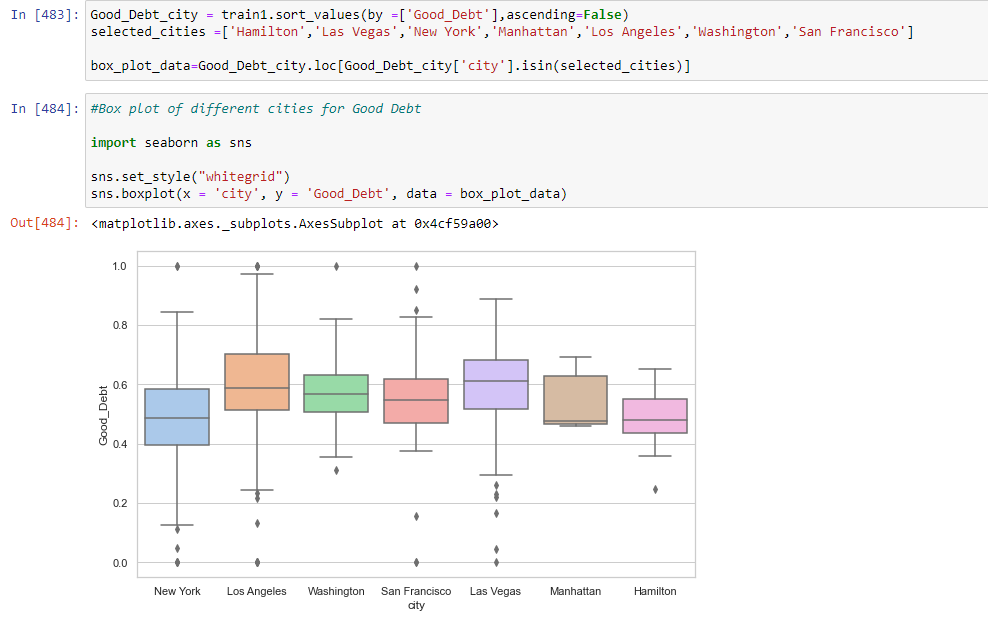


# Los Angeles is having lot of outliers forb home equity loan

# Washington,Hamilton & San Fransico is nearly having a normal distribution for home equity loan with few outliers

#Median home\_equity loan is highest for San Fransissco & Washington city

Box plot for Good Debt

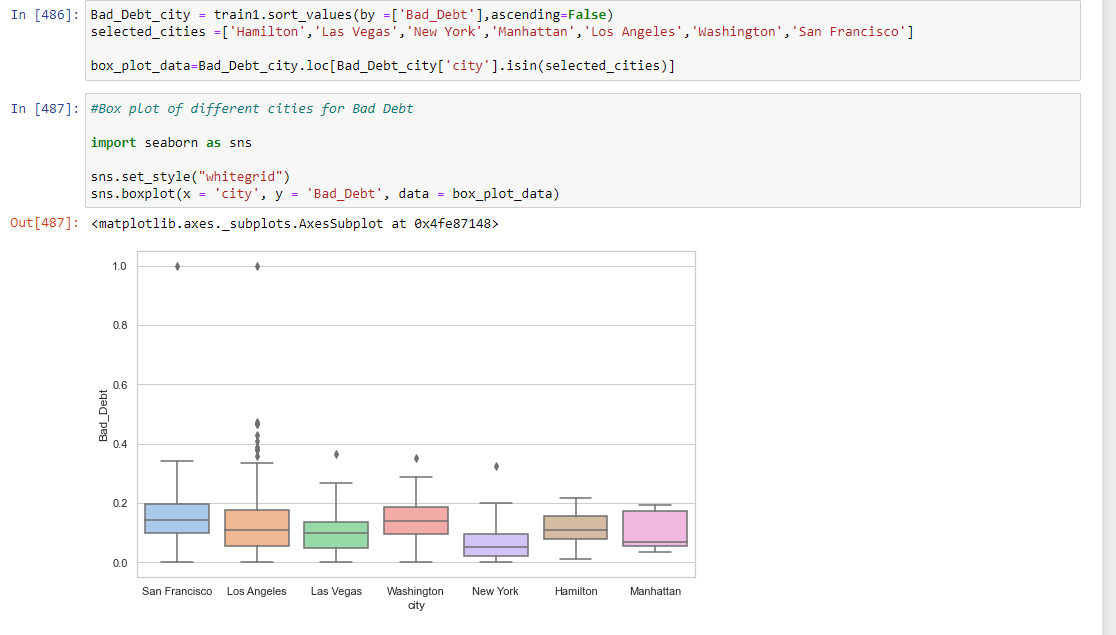


#New York is having a near normal distribution for Good Debt and Las Vegas i s having lot of outliers below the first Quartile

#Median distribution is different for all the cities

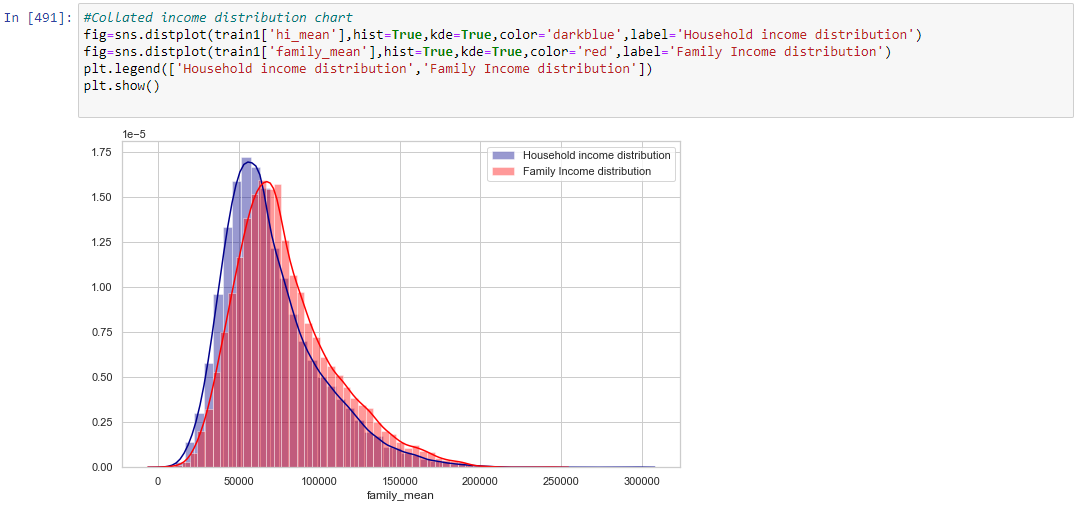
#Median Good debt distribution for New York is 0.5 which is less than Las Vegas and Los Angeles with 0.6

Box plot for Bad debt



#Median for Bad debt in New York is less than 0.1

* Create a collated income distribution chart for family income, house hold income and remaining income.



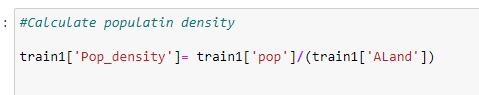
Household income and family income distribution have a large area of overlap and slightly right skewed.

Objective 5) Perform EDA and come out with insights into population density and age. You may require deriving new fields (Make sure to weight averages for accurate measurements):

* Population density (hint-use ‘pop’ and ‘Aland’ to calculate)
* median age (hint-use the variables ‘male\_age\_median’, ‘female\_age\_median’, ‘male\_pop’, ‘female\_pop’)

Visualize the findings using appropriate chart type.

Population density

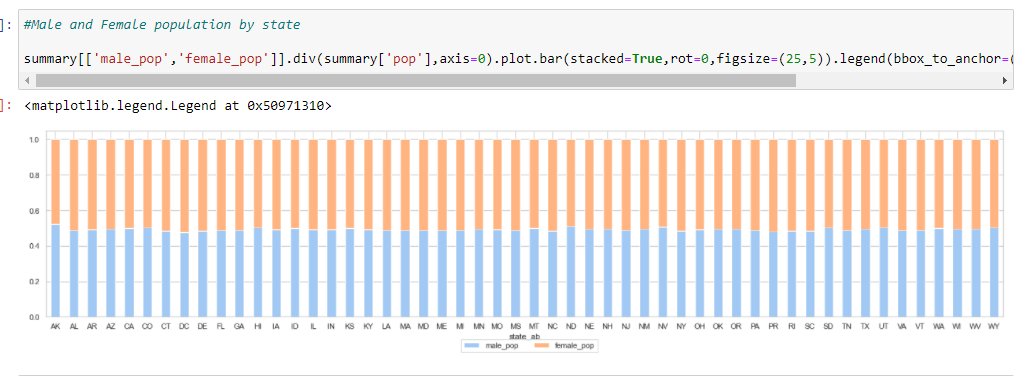




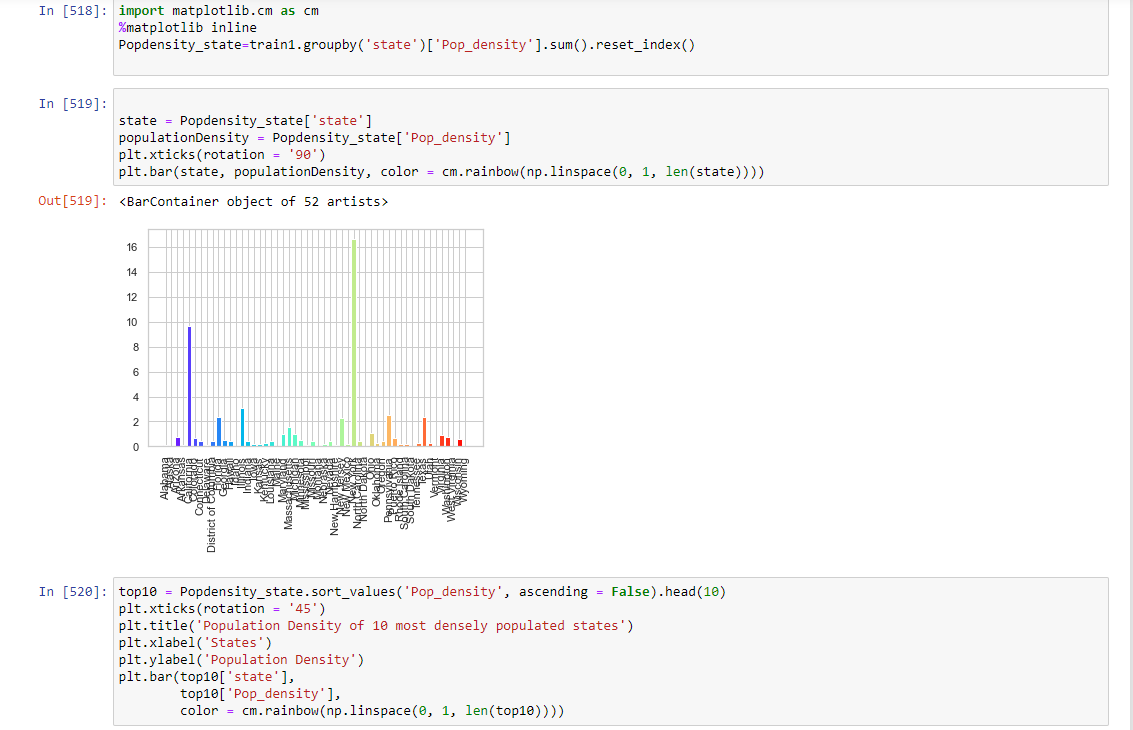
Median age



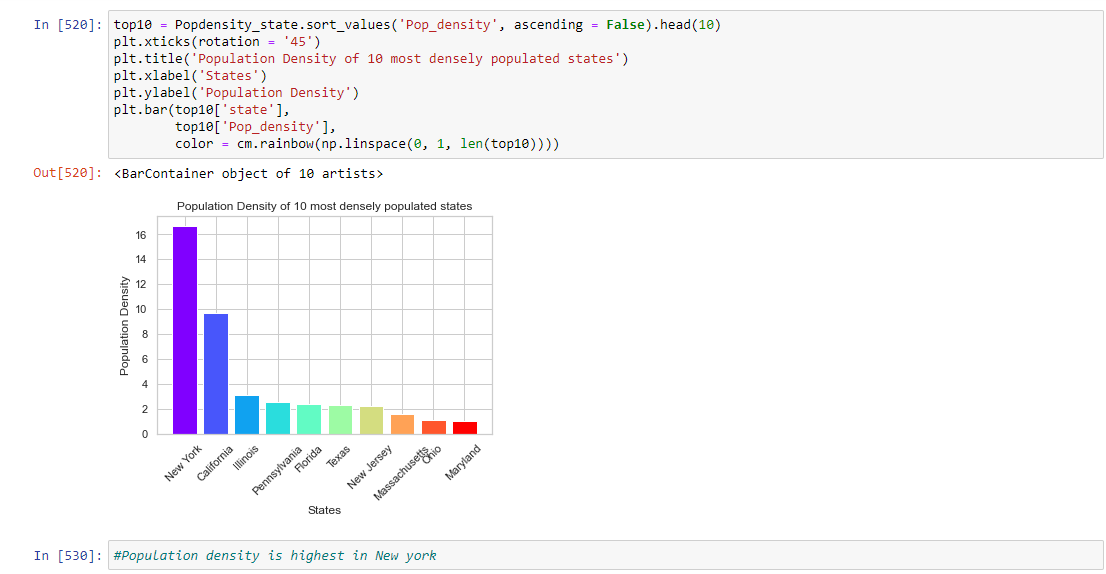




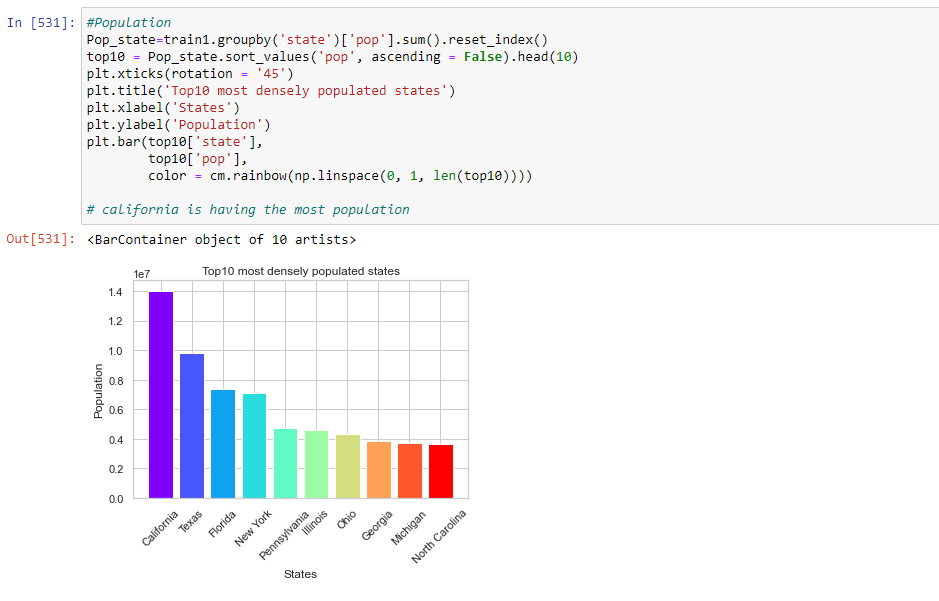
Male and Female population by state



Top 10 densely populated states



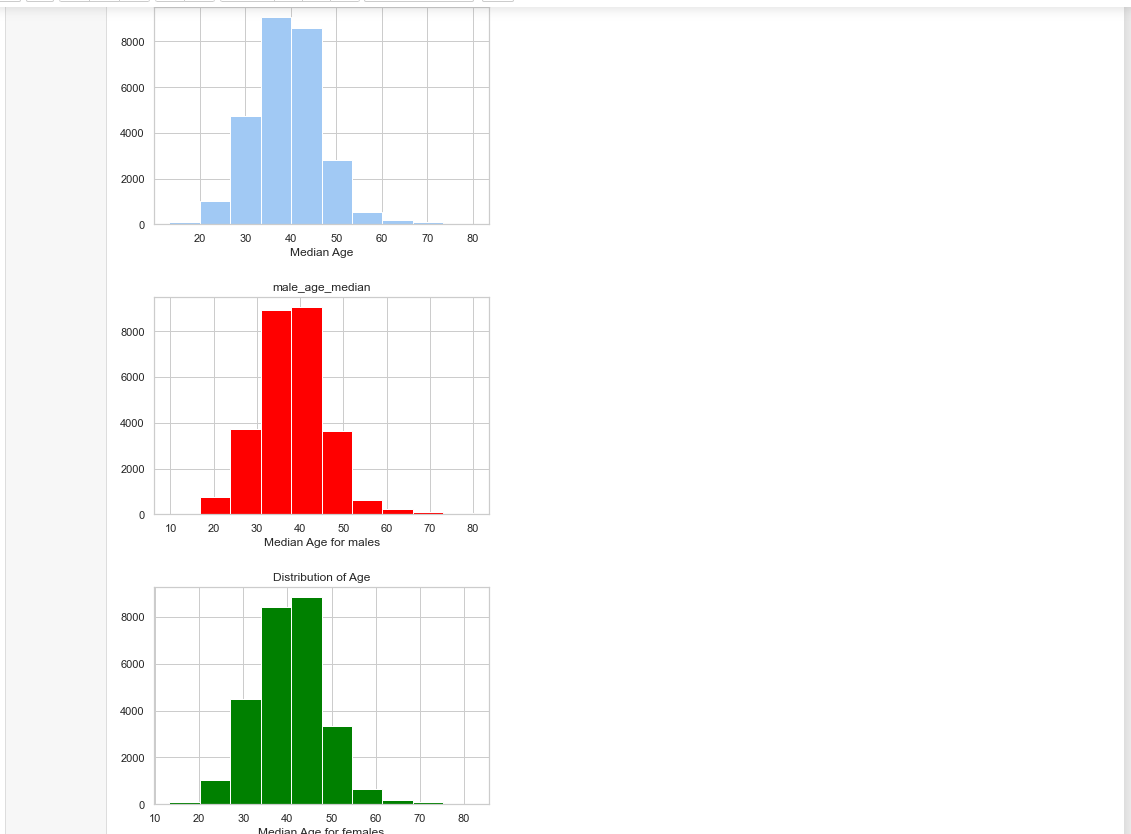
Population density is highest in New York



California is having the highest population

Histogram for age distribution

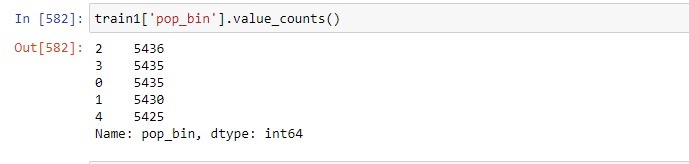


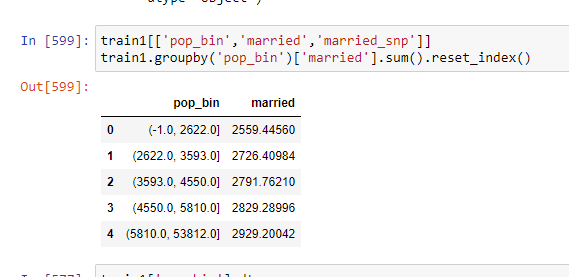


1. Objective 6 ) Create bins for population into a new variable by selecting appropriate class interval so that the no of categories(bins) don’t exceed 5 for the ease of analysis. Analyze the married, separated and divorced population for these population brackets. Visualize using appropriate chart type.

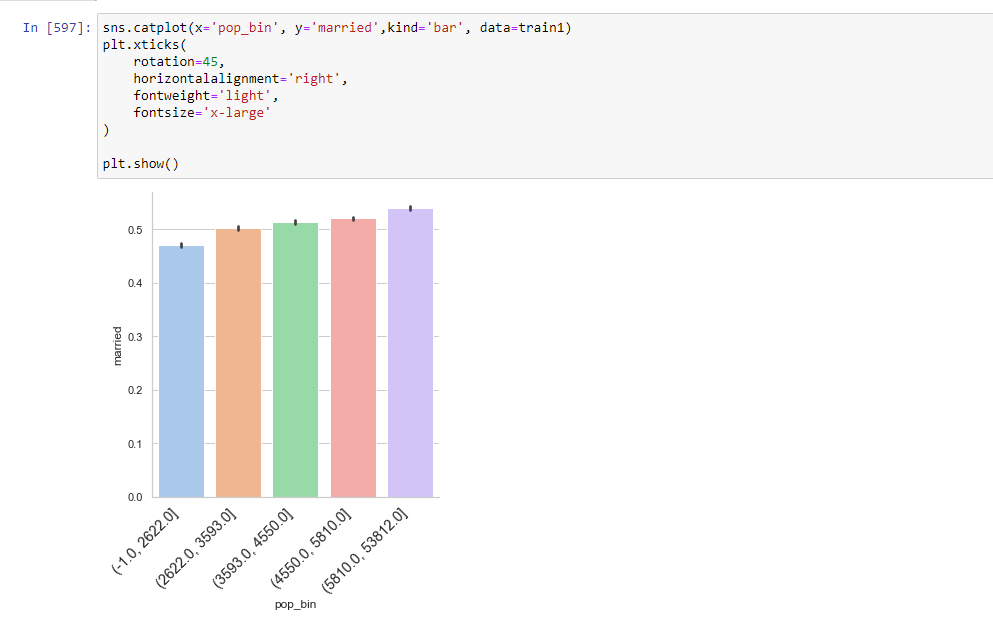
Created bins for population using qcut

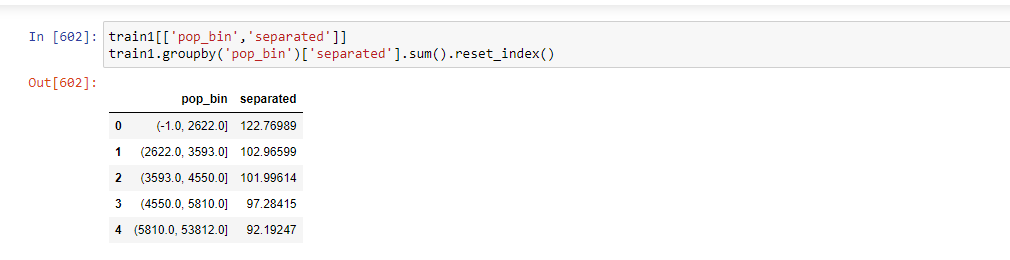




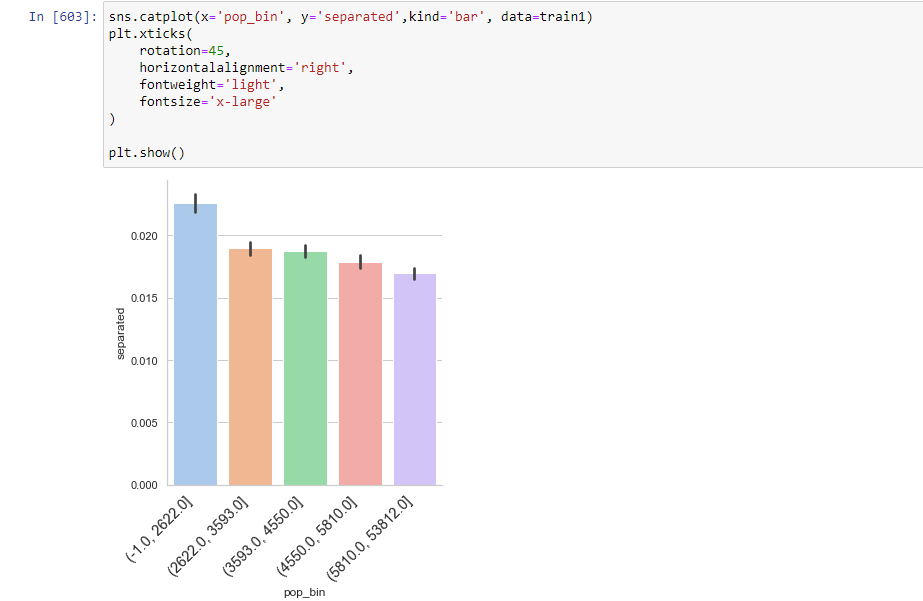


Analysing married population using catplot





Analysing separated population using catplot



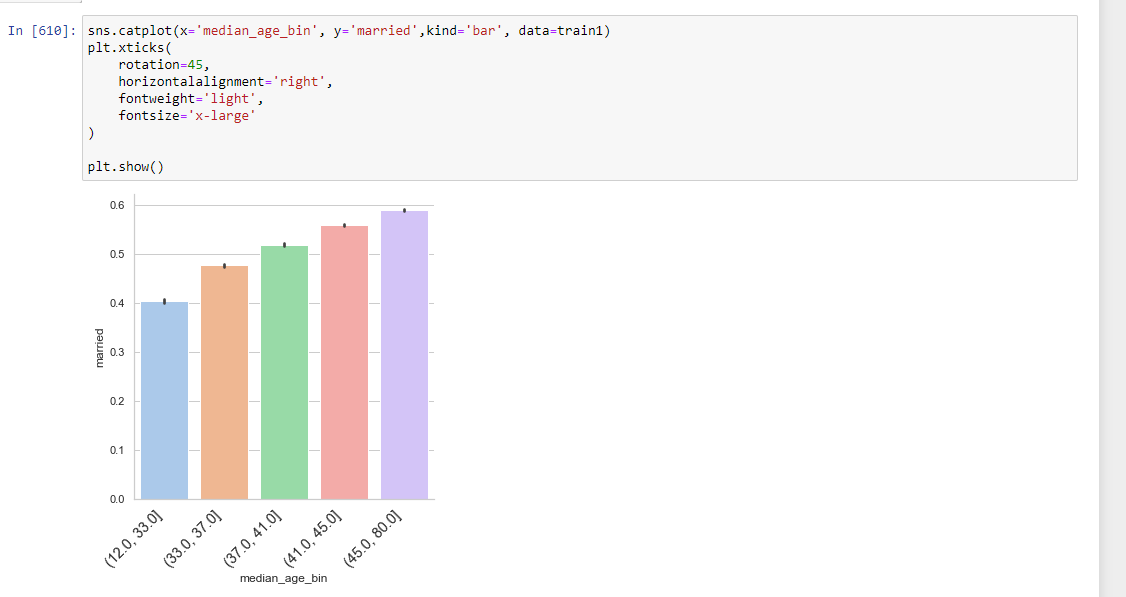
Analysing divorced population using catplot



Creating bin for median\_age



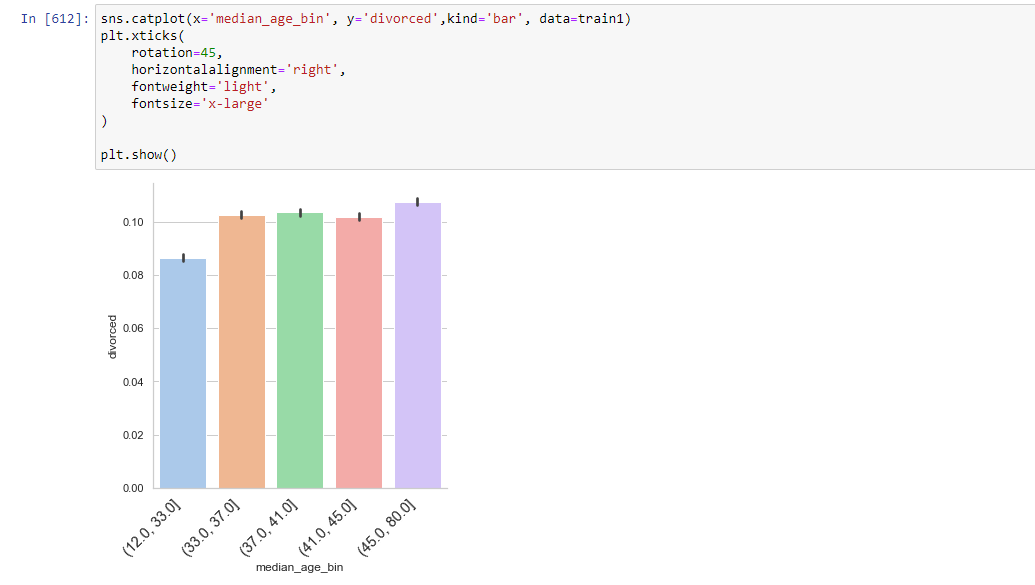
Catplot for married age distribution



Catplot for separated age distribution



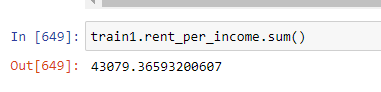
Catplot for divorced age distribution



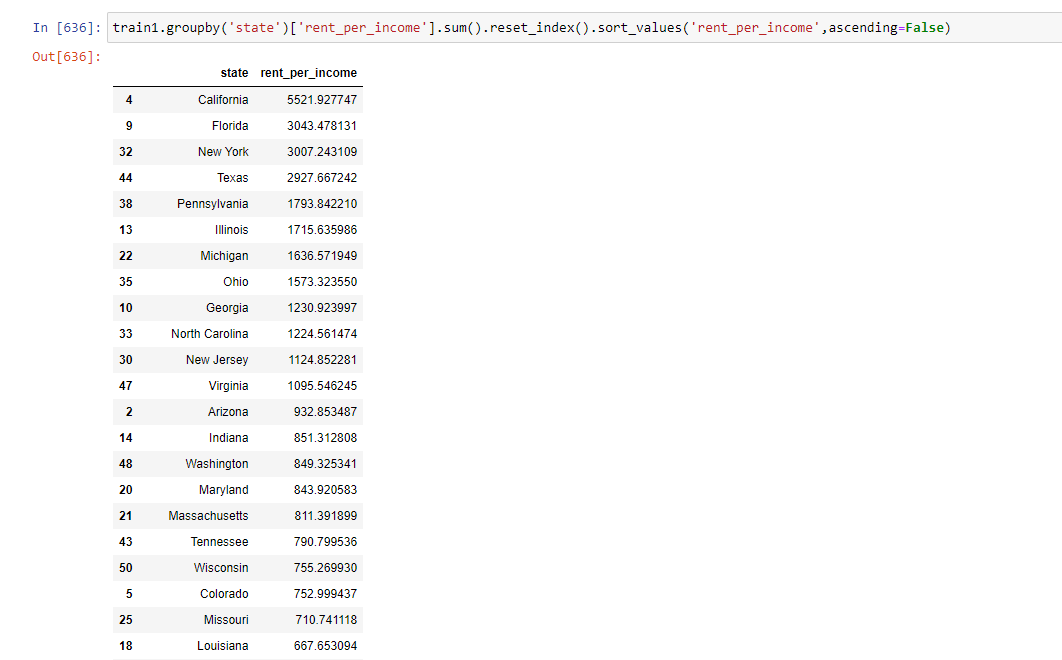
Objective 7) Please detail your observations for rent as a percentage of income at an overall level and for different states.

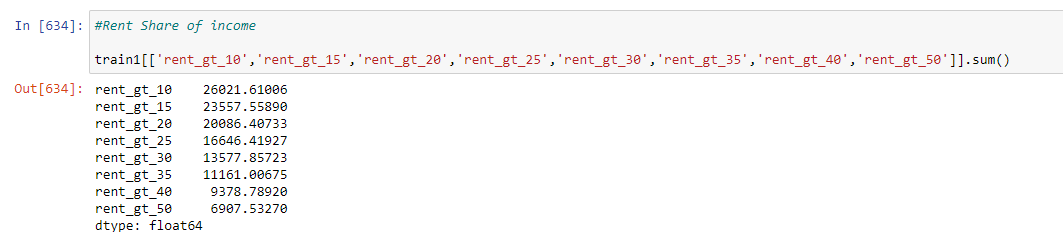
Rent as percentage of income at an overall level

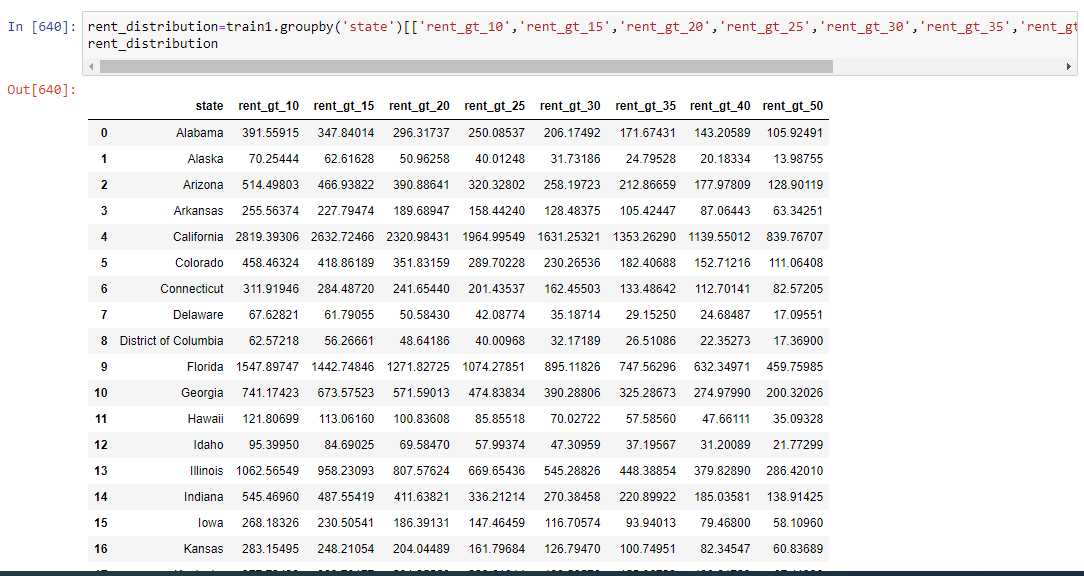


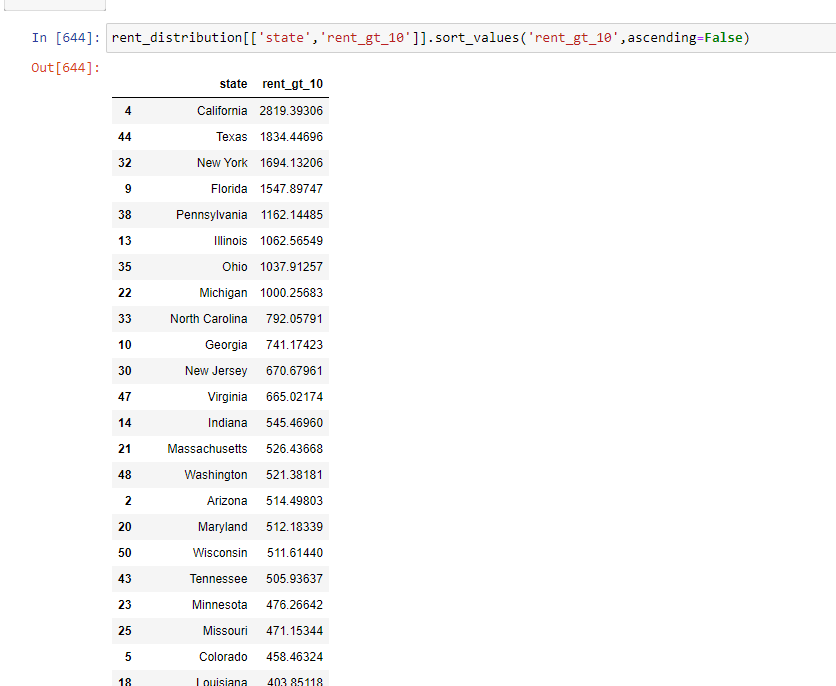


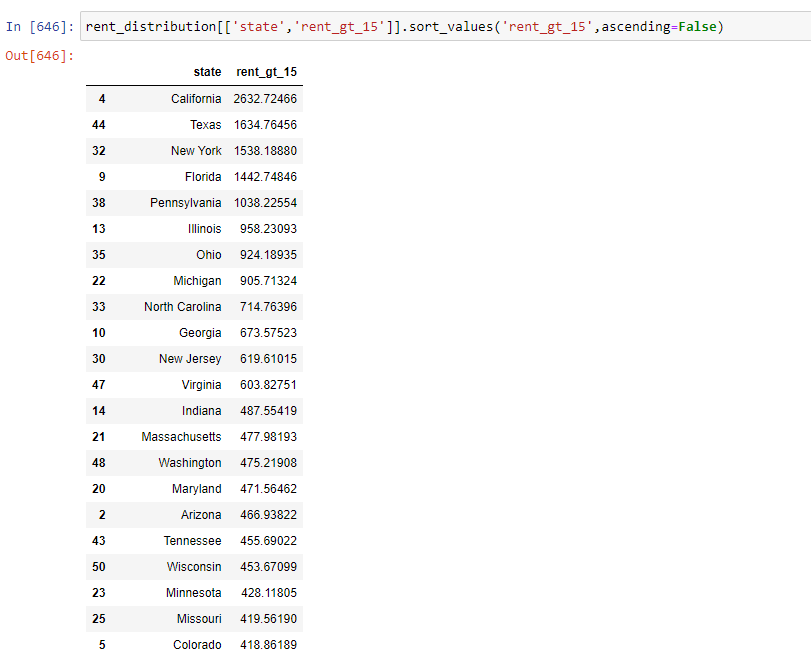
At State level



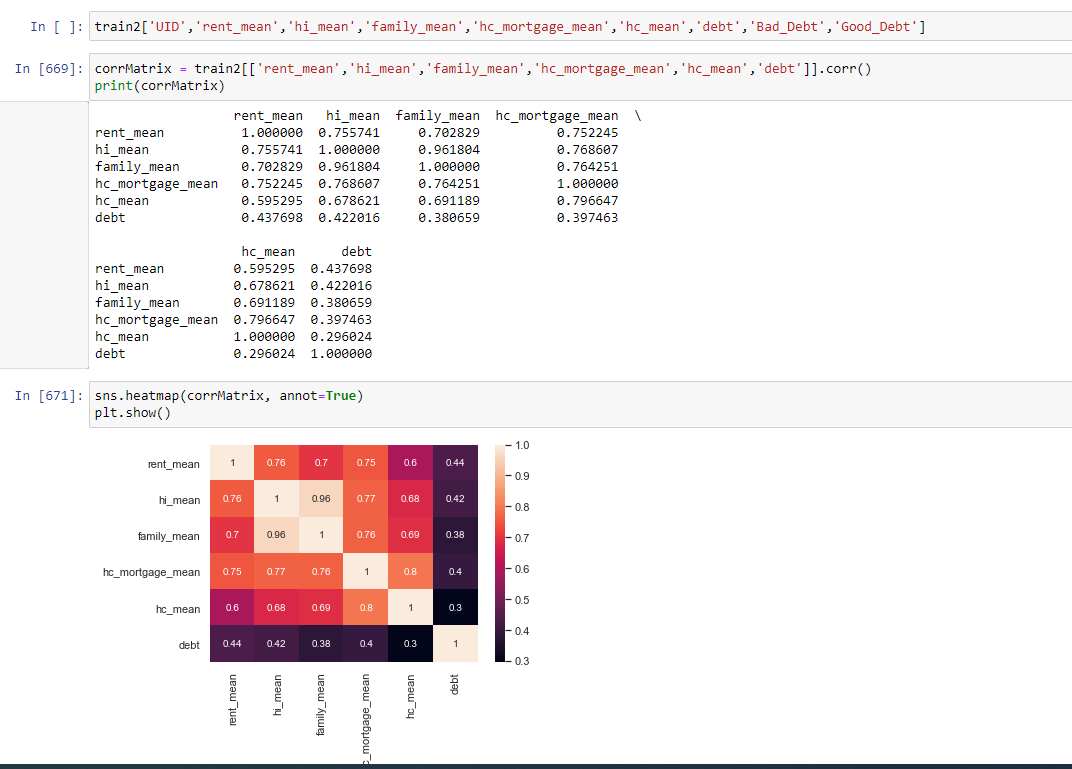




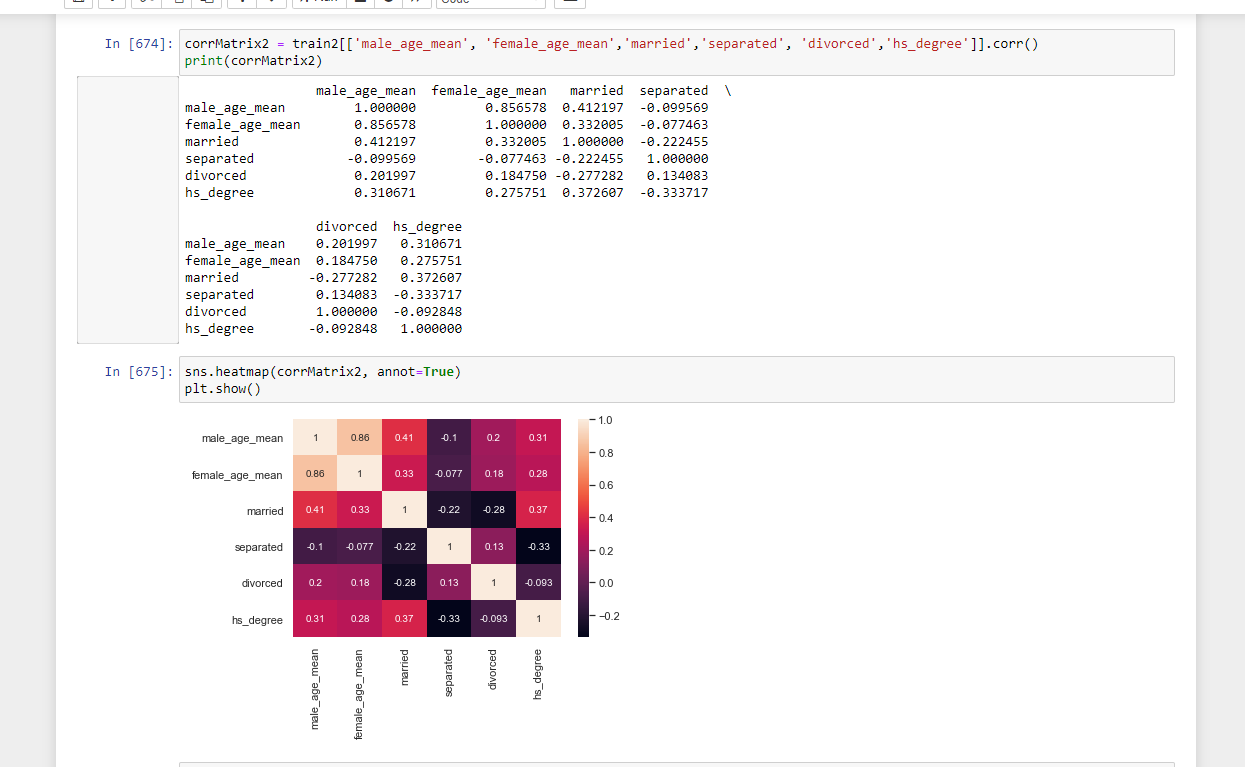


California is having the high share of rent as income.

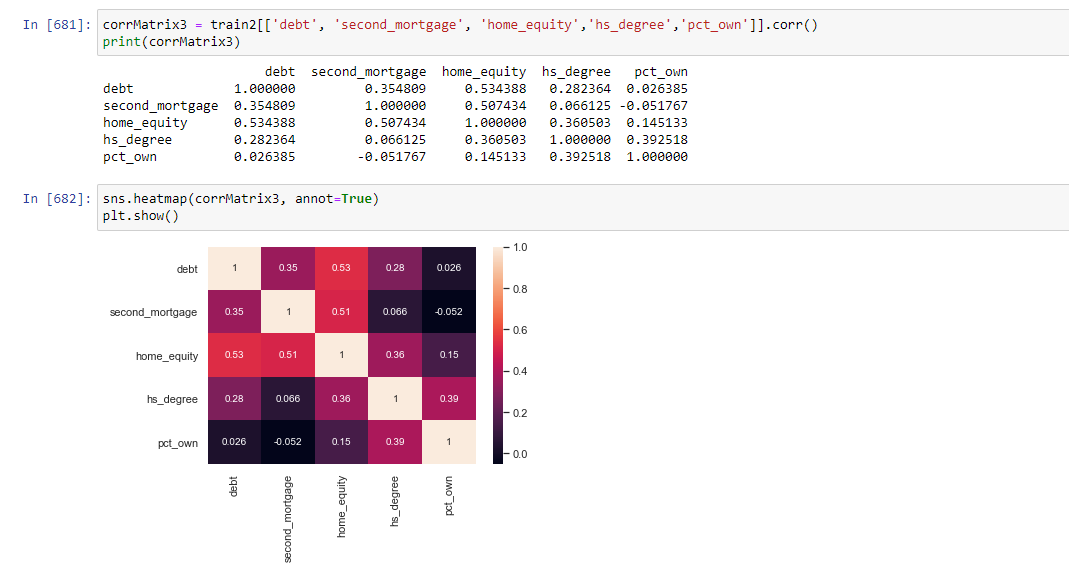
1. Objective8) Perform correlation analysis for all the relevant variables by creating a heatmap. Describe your findings.



hc\_mortage\_mean has a high positive correlation with rent\_mean,hi\_mean,hc\_mean and family\_mean.



Most of the fields are having low weak correlation here

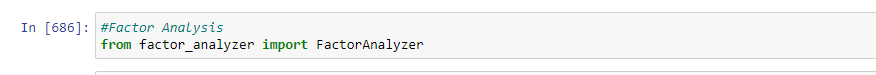


Debt,second\_mortagage and home\_equity is having high correlation.

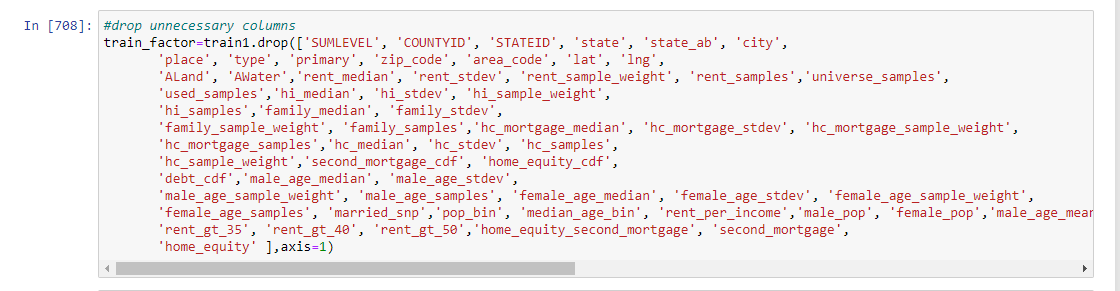
1. Objective 9) The economic multivariate data has a significant number of measured variables. The goal is to find where the measured variables depend on a number of smaller unobserved common factors or latent variables. Each variable is assumed to depend on a linear combination of the common factors, and the coefficients are known as loadings. Each measured variable also includes a component due to independent random variability, known as "specific variance" because it is specific to one variable. Obtain the common factors and then plot the loadings. Use factor analysis to find latent variables in our dataset and gain insight into the linear relationships in the data

* Highschool graduation rates
* Median population age
* Second Mortgage Statistics
* Percent Own
* Bad Debt Expense

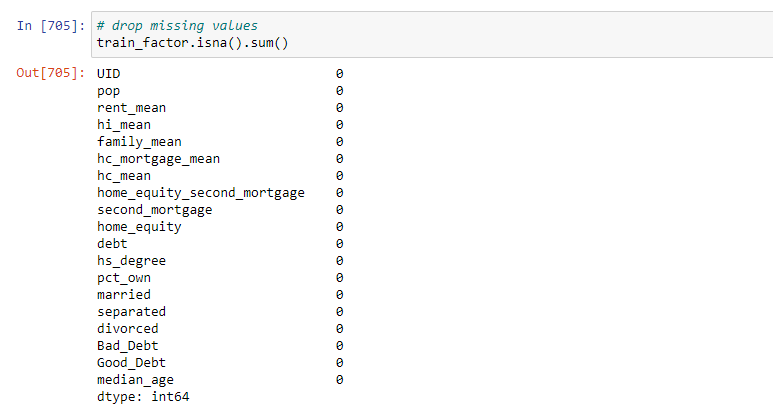
Factor Analysis



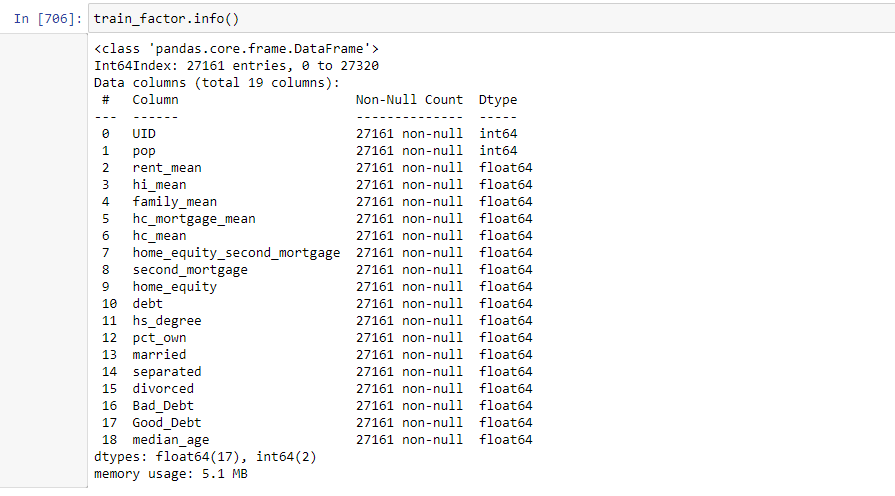
Dropping unnecessary columns



Dropping missing values



No missing values in the datset

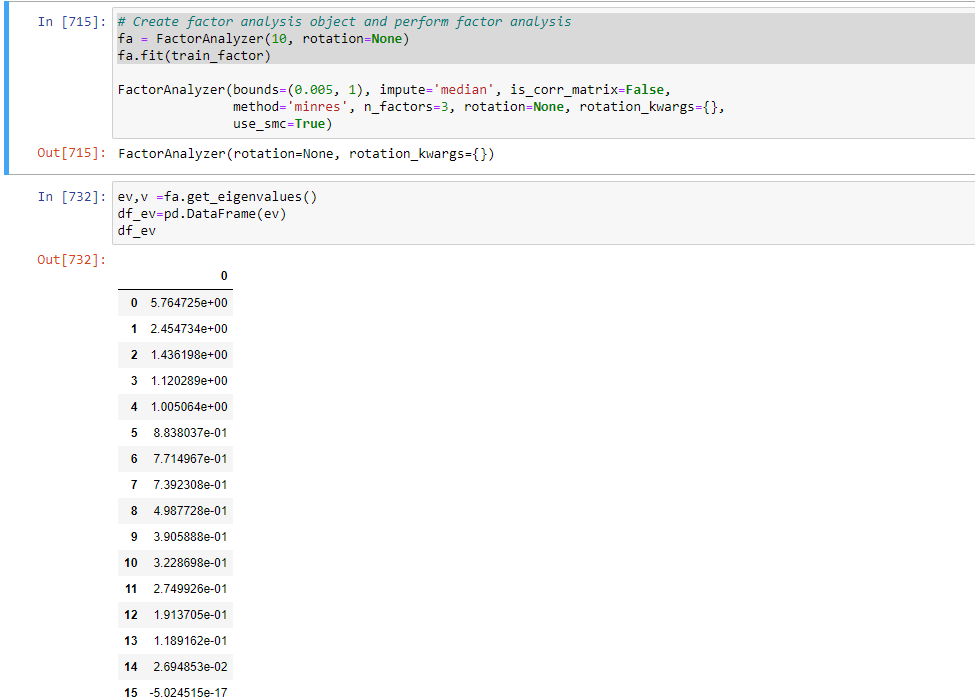


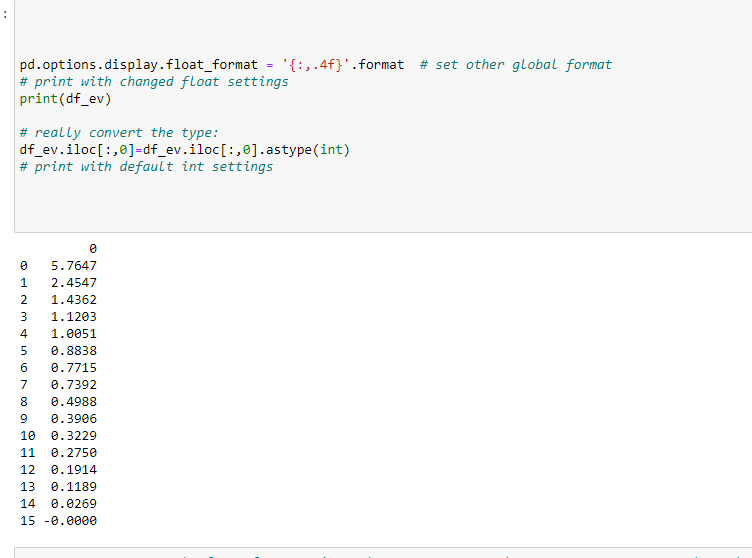


#The overall KMO for our data is 0.73, which is excellent. This value indicates that we can proceed with your planned factor analysis

#Test measures the suitability of data for factor analysis. It determines the adequacy for each observed variable and for the complete model. KMO estimates the proportion of variance among all the observed variable. Lower proportion id more suitable for factor analysis. KMO values range between 0 and 1. Value of KMO less than 0.6 is considered inadequate.

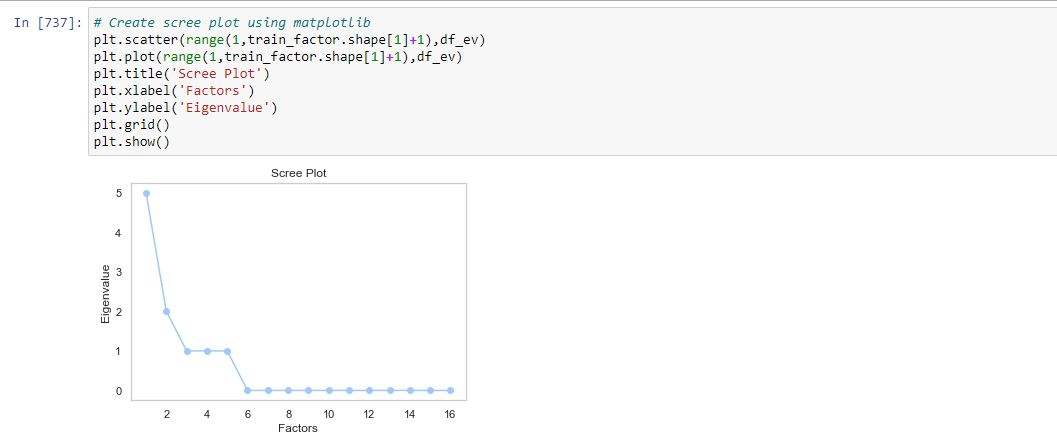
Creating factor analysis objects

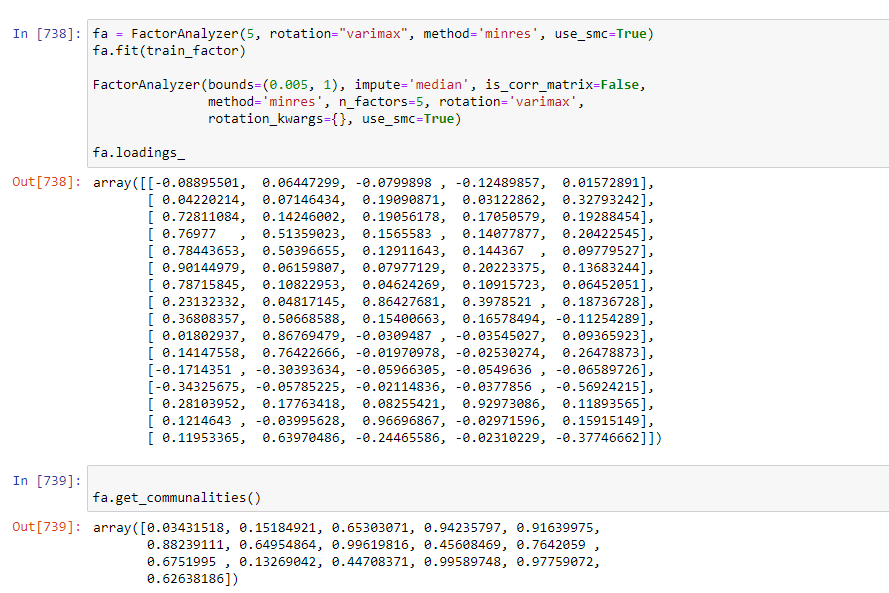




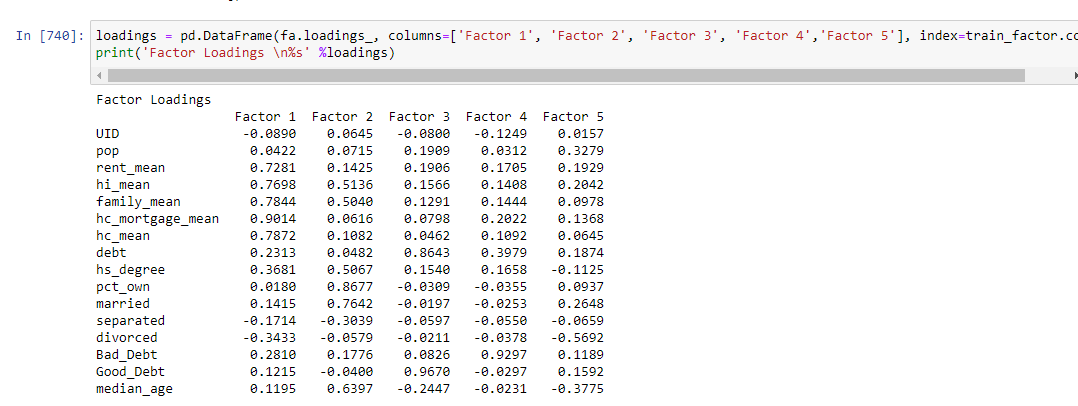
#Here, you can see only for 5-factors eigenvalues are greater than one. It means we need to choose only 5 factors (or unobserved variables)

Scree plot





Loadings



# Factor 1 - hc\_mortgage\_mean,hc\_mean,family\_mean,hi\_mean,rent\_mean

#Factor 2- pct\_own,married,hi\_mean,hs\_degree,median\_age

#Factor 3 - debt,Good\_Debt,

#Factor 4- Bad\_Debt,debt

#Factor 5 -Pop,married,divorced

Objective 10)Build a linear Regression model to predict the total monthly expenditure for home mortgages loan; please refer - ‘deplotment\_RE.xlsx’.

Column hc\_mortgage\_mean is predicted variable. This is mean monthly mortgage and owner costs of specified geographical location.

Note: *Exclude loans from prediction model which have NaN values for hc\_mortgage\_mean. NaN represents not a number/missing values.*

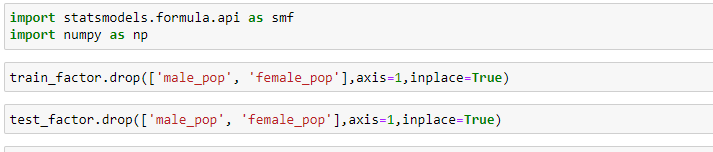
* Run a model at a Nation level. If the accuracy levels and R square are not satisfactory proceed to below step
* Run another model at State level. There are 52 states in USA.

Considerations: Keep below considerations while building a linear regression model

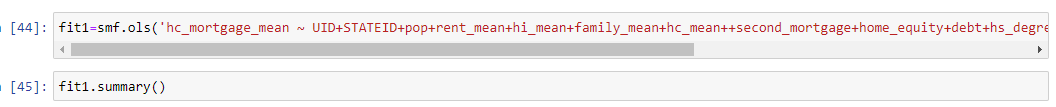
* Variables should have significant impact on predicting Monthly mortgage and owner costs
* Utilize all predictor variable to start with initial hypothesis
* R square of 60% and above should be achieved
* Ensure Multi-collinearity does not exist in dependent variables
* Test if predicted variable is normally distributed

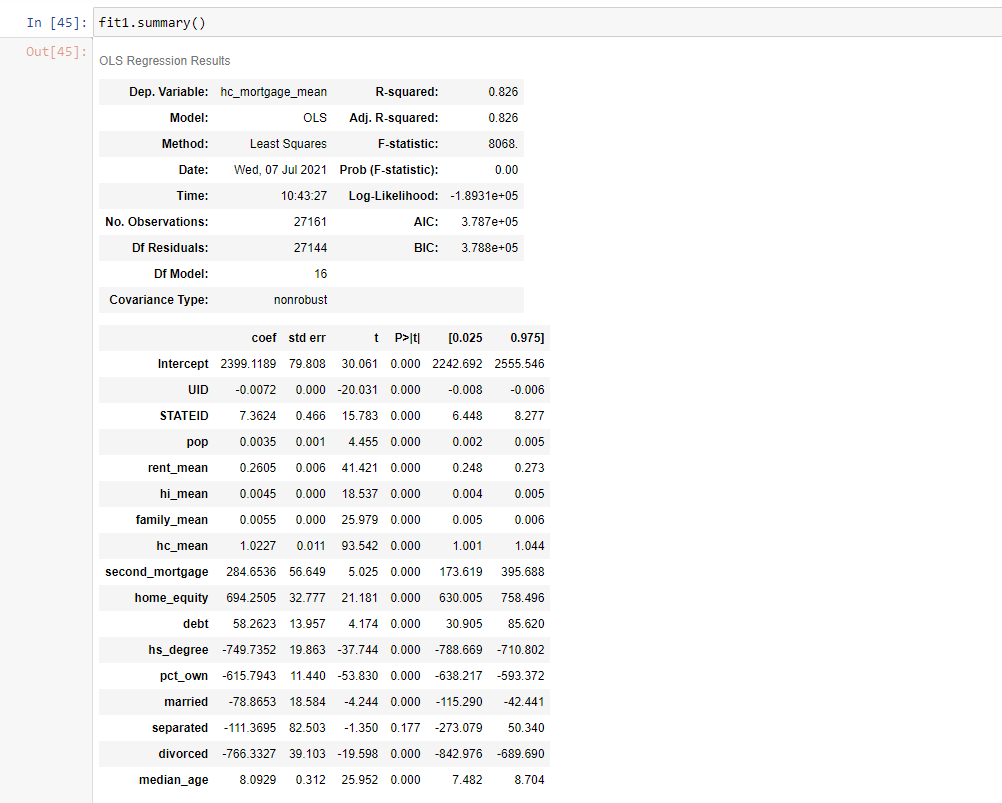
Dropping unnecessary variables





fit1

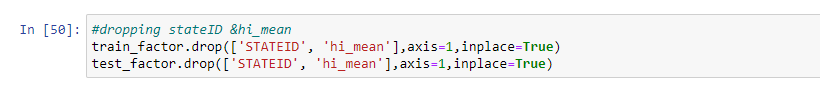




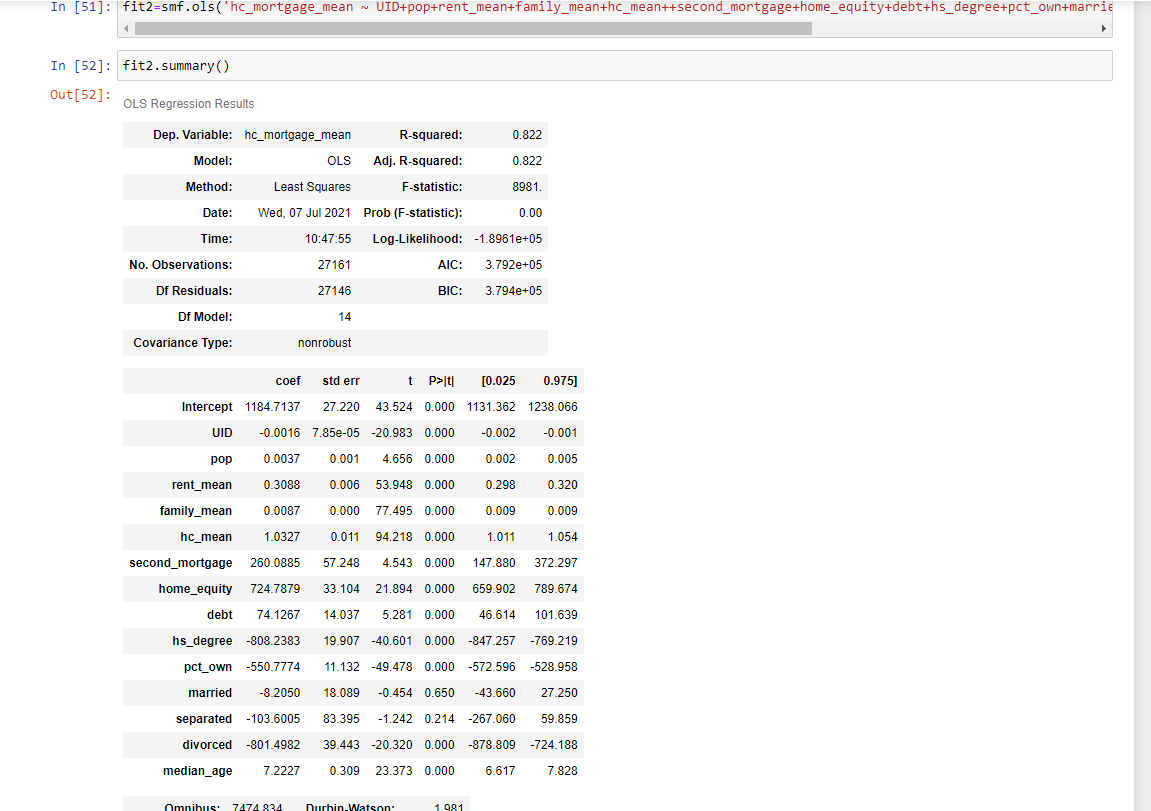
Checking multicollinearity for fit 1 variables



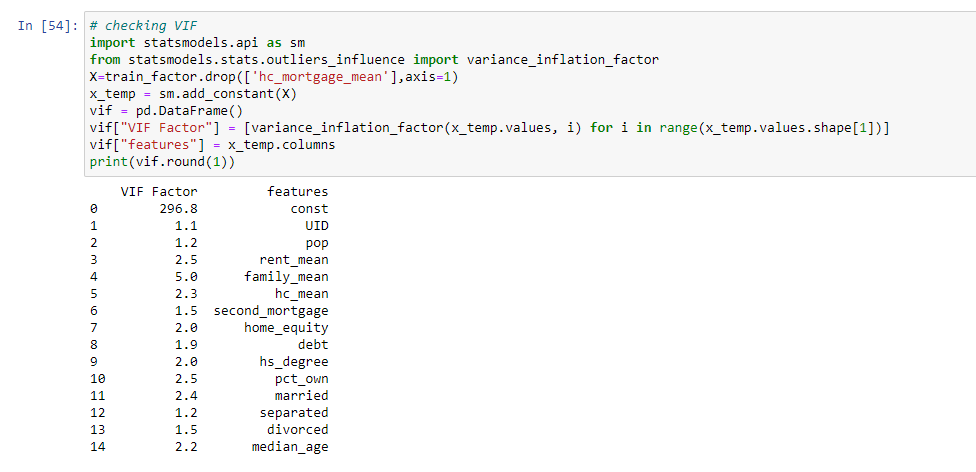
Dropping STATEID and hi\_mean because of high multicollinearity

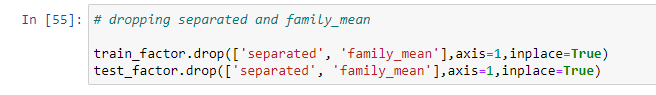


fit2



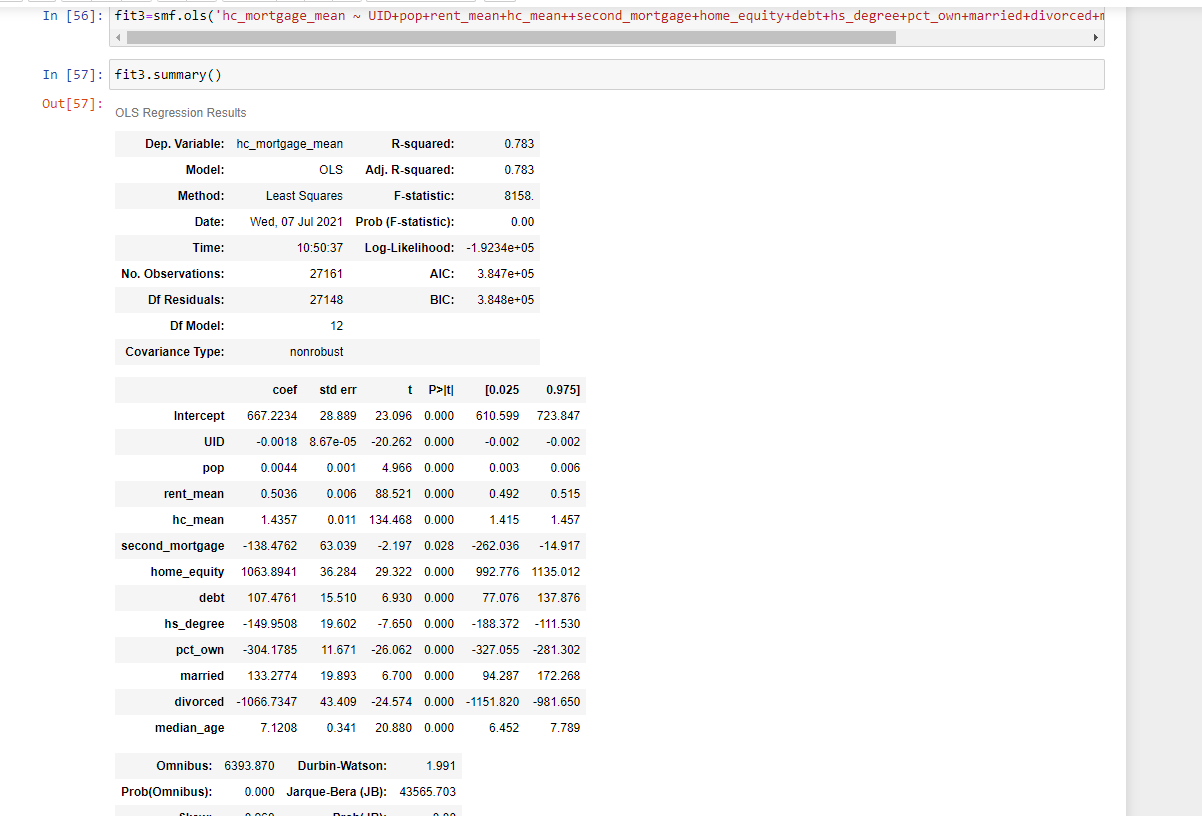
Checking VIF





Dropping separated(p>0.05) and family\_mean because of high multicollinearity

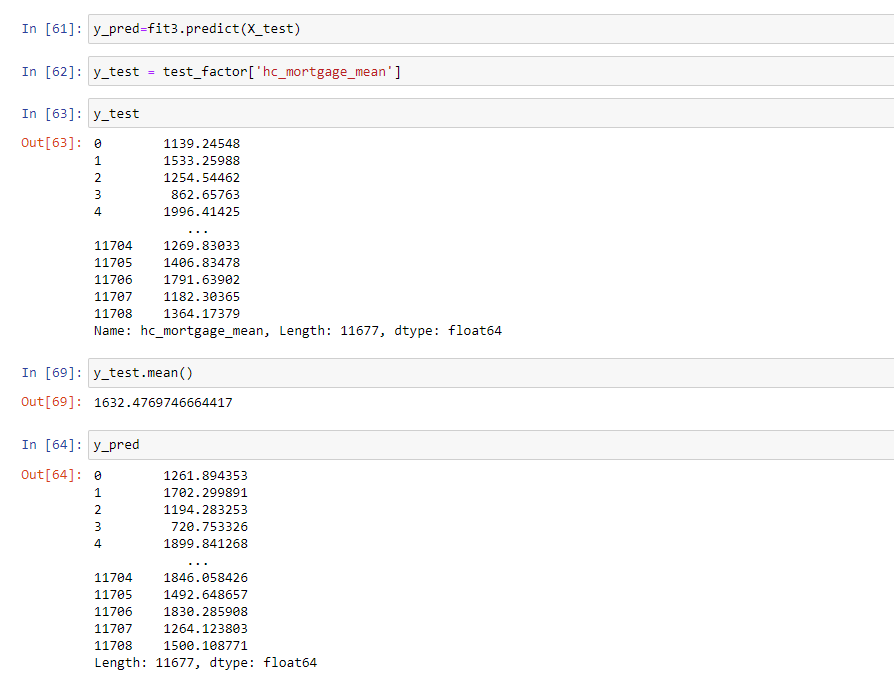
fit 3

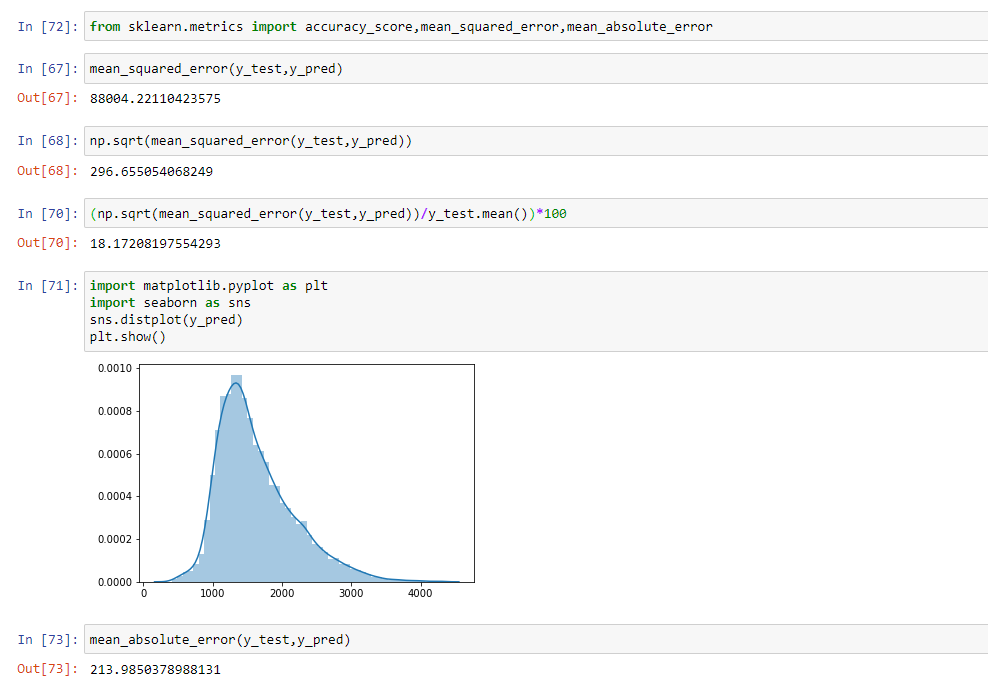


R2 and Adj R2 is .783

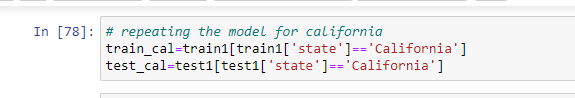


Predicting





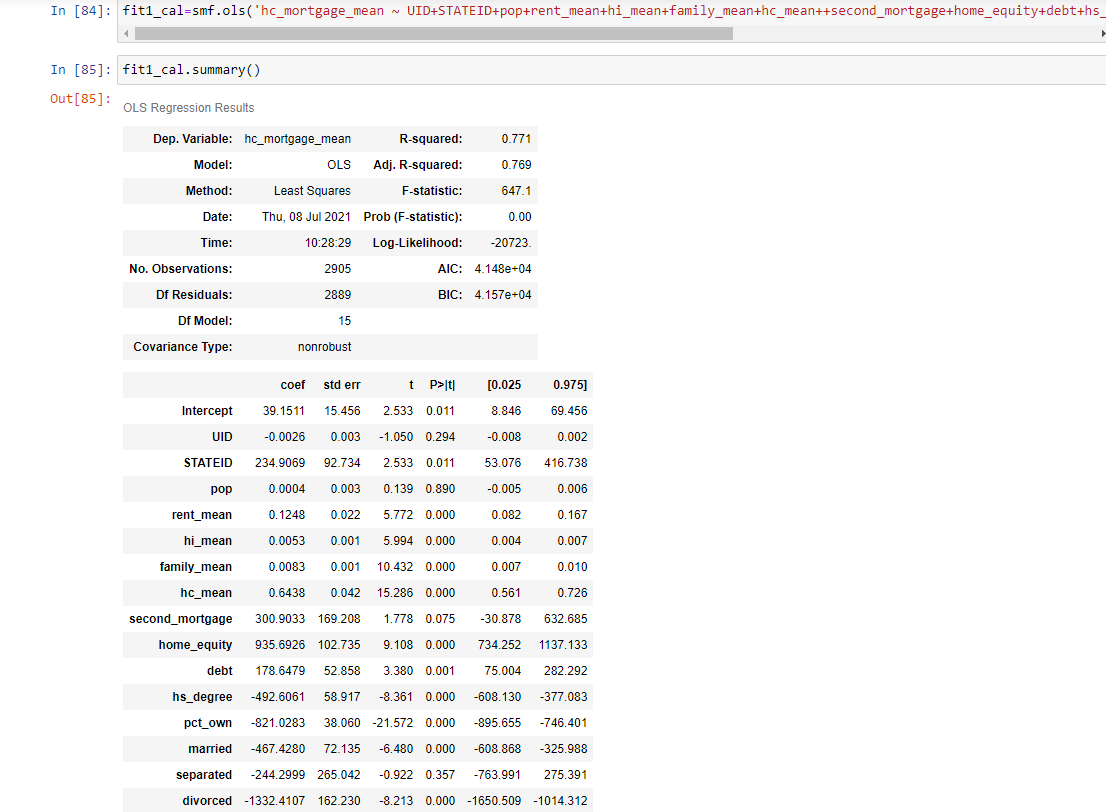
Repeating the model for California



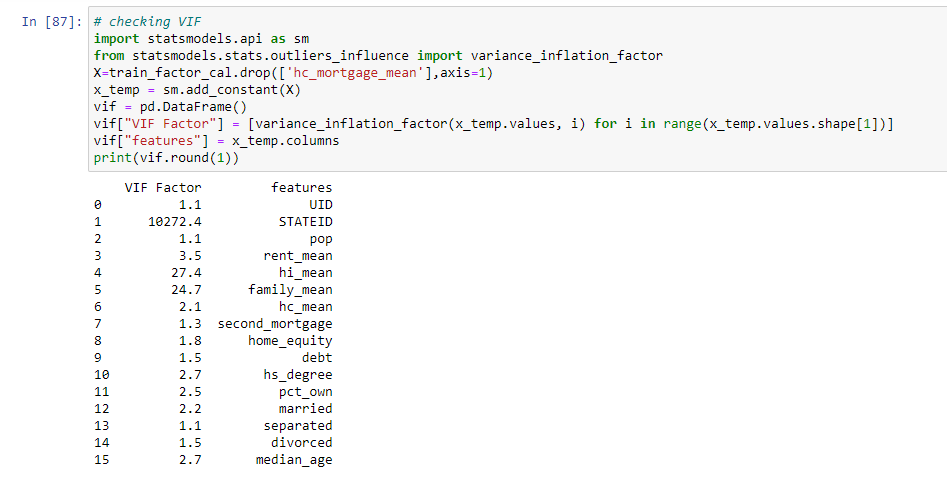
Dropping unnecessary variables



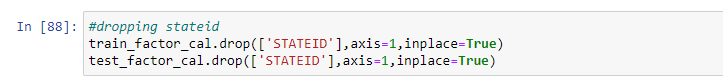
Fit1



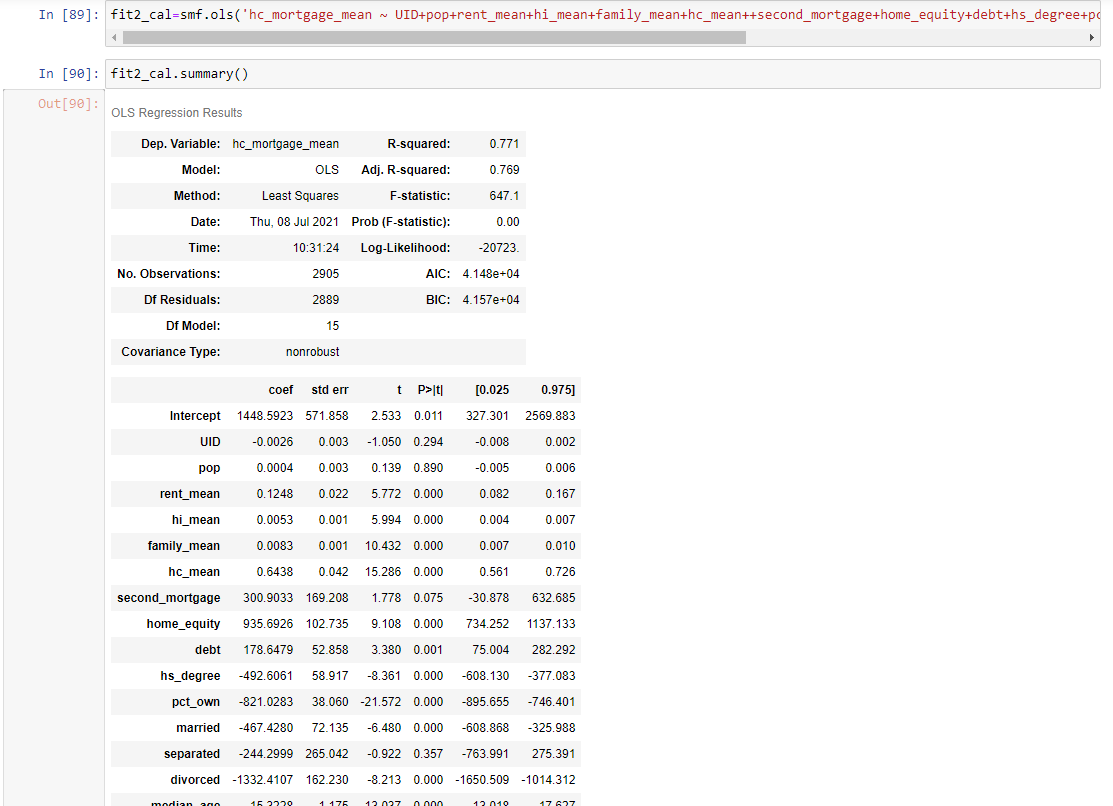
Checking VIF

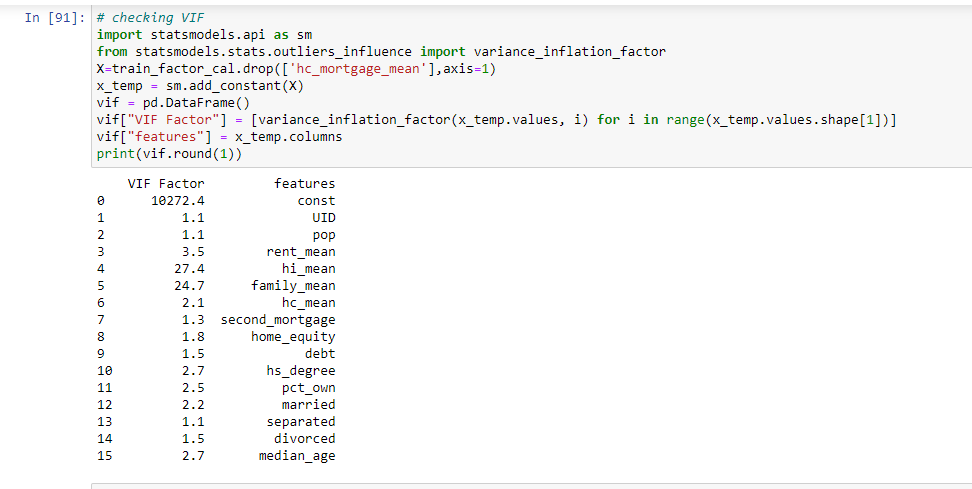


Dropping stated because of might multicollinearity



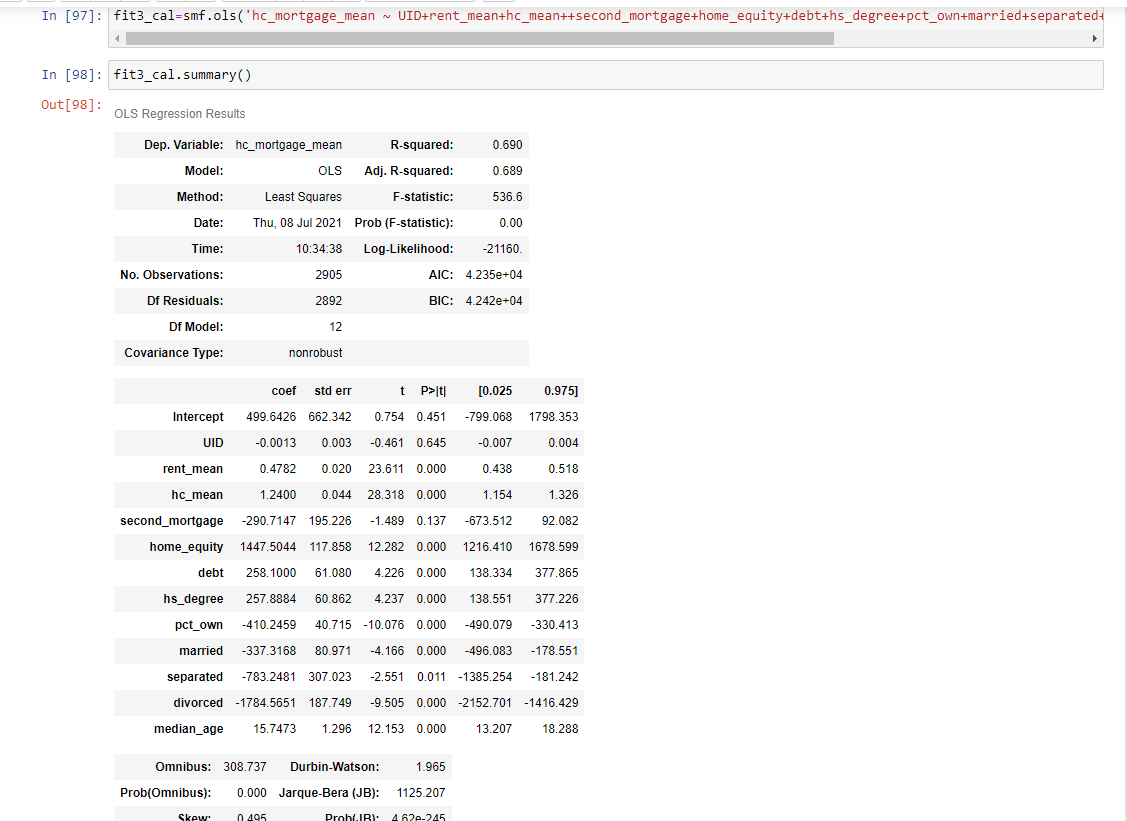
Fit2

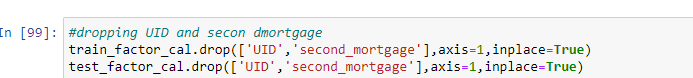




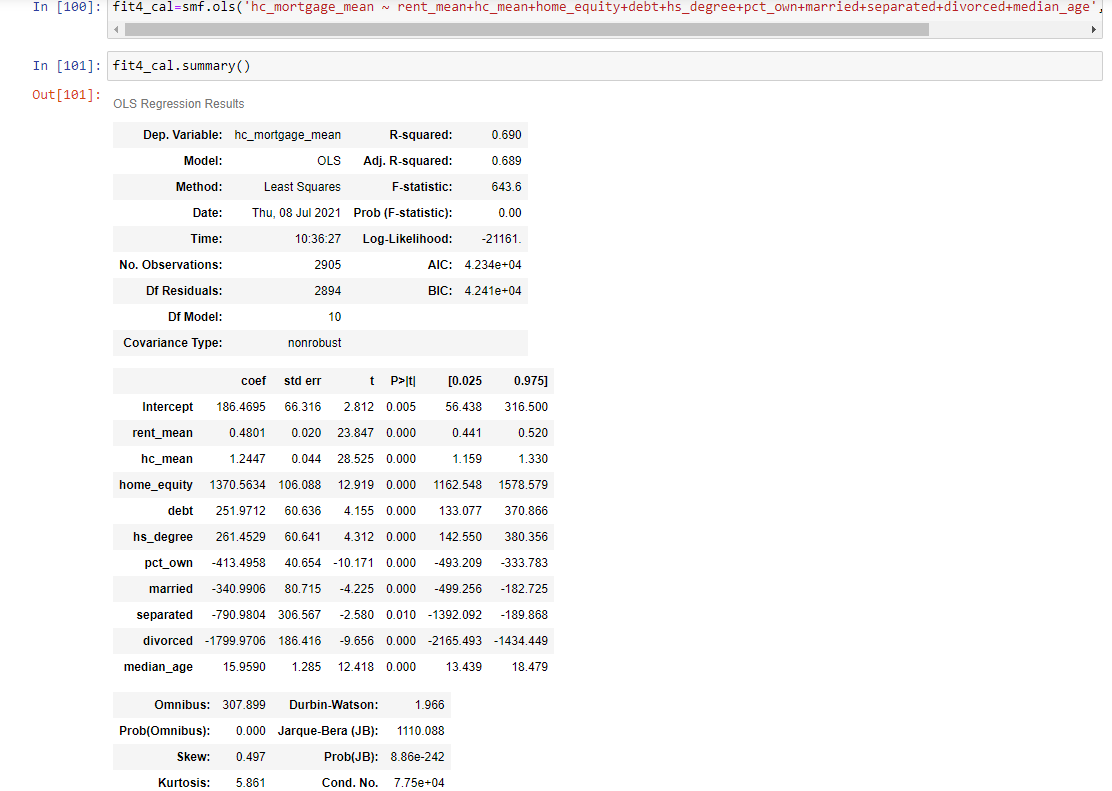
Dropping hi\_mean,family\_mean because of high VIF and pop( high p value)

fit 3



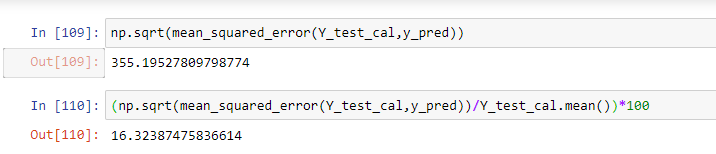


Fit 4



Predicting





1. Objective11) Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business. The dashboard must entail the following:
2. Box plot of distribution of average rent by type of place (Village, urban, town etc.)
3. Pie charts (Venn diagram) to show overall debt (% bad and good debt) and bad debt (2 mortgage and home equity loan)
4. Explore the top 2,500 locations where the percentage of households with a second mortgage is the highest and percent ownership is above 10%. Visualize using geo-map.
5. Heat map for correlation matrix
6. Pie chart to show the population distribution across different types of places (Village, urban, town etc.)

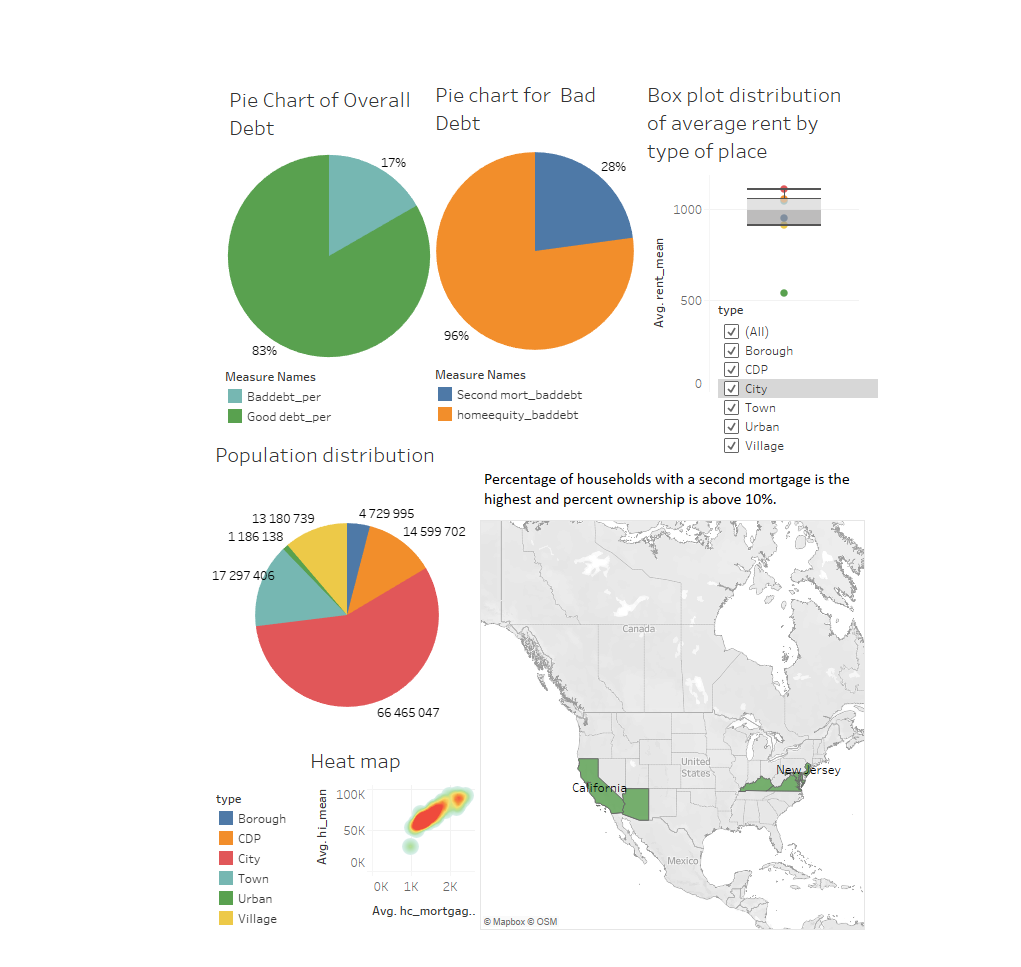


Tableau Public link to dashboard - https://public.tableau.com/app/profile/aparna6616/viz/CapstoneProject\_RealEstate/CapstoneProject1