Vishal Maru

Assignment 2

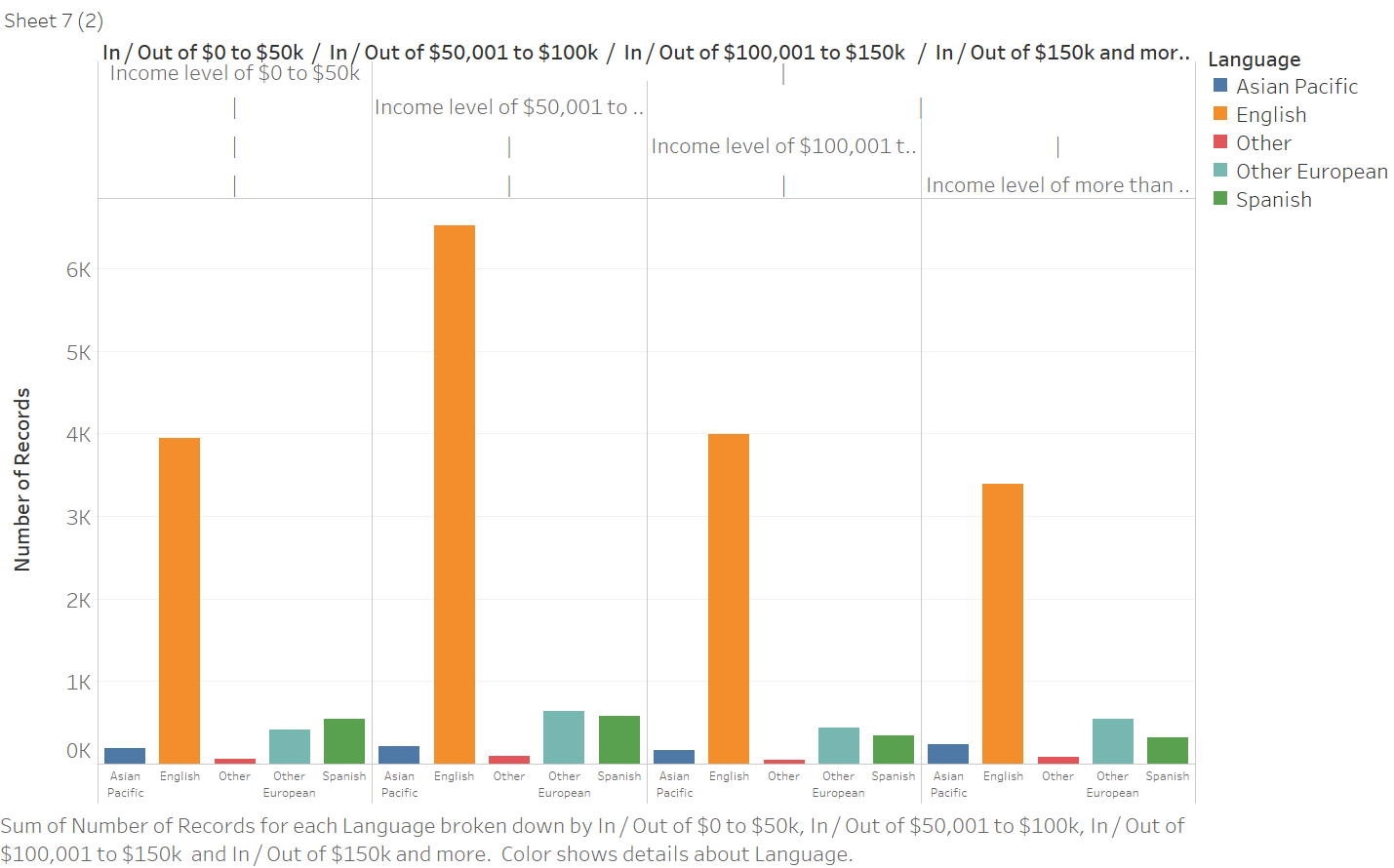
IT 418 – Foundations of Data Science

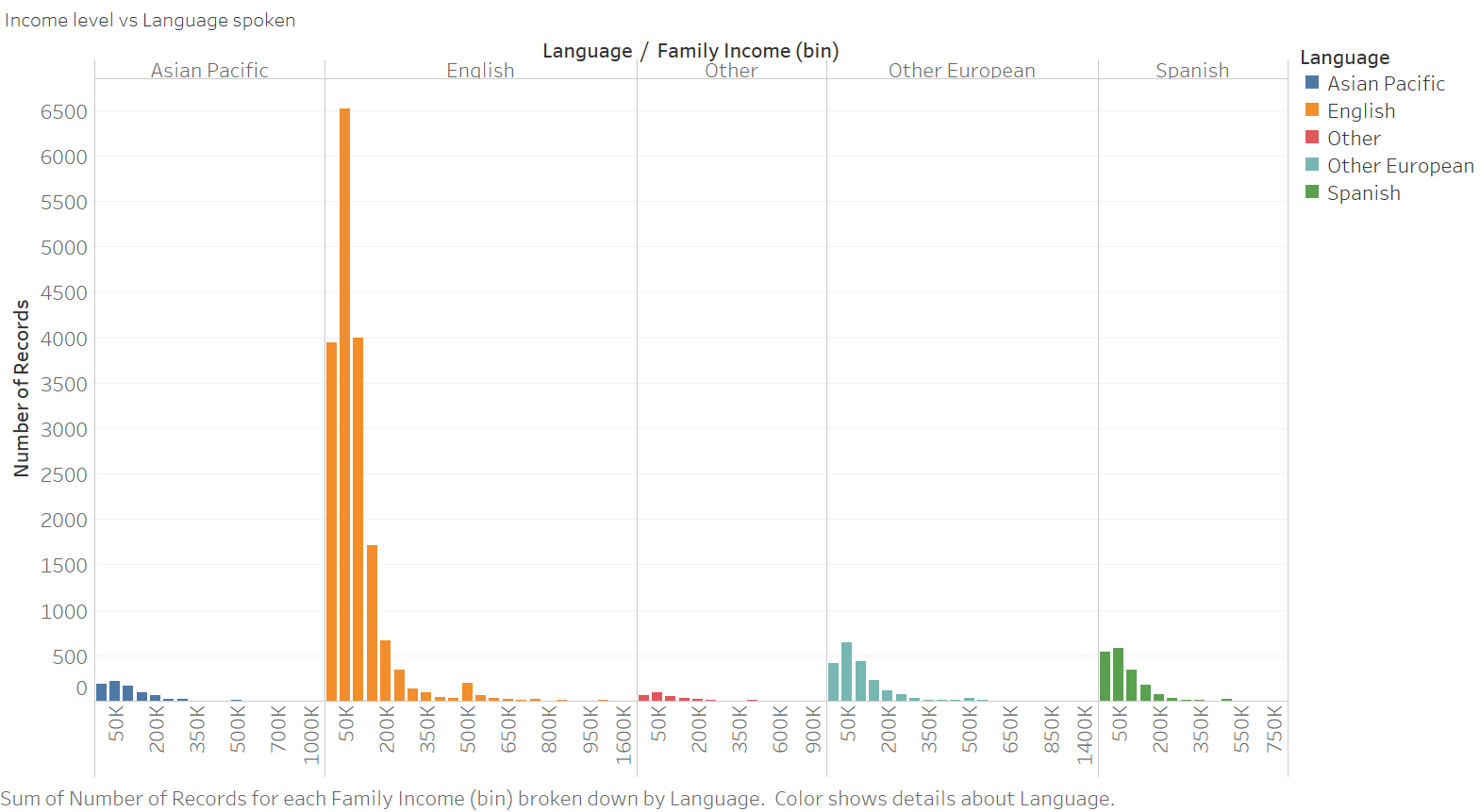
Dr. Rajeev Bukralia

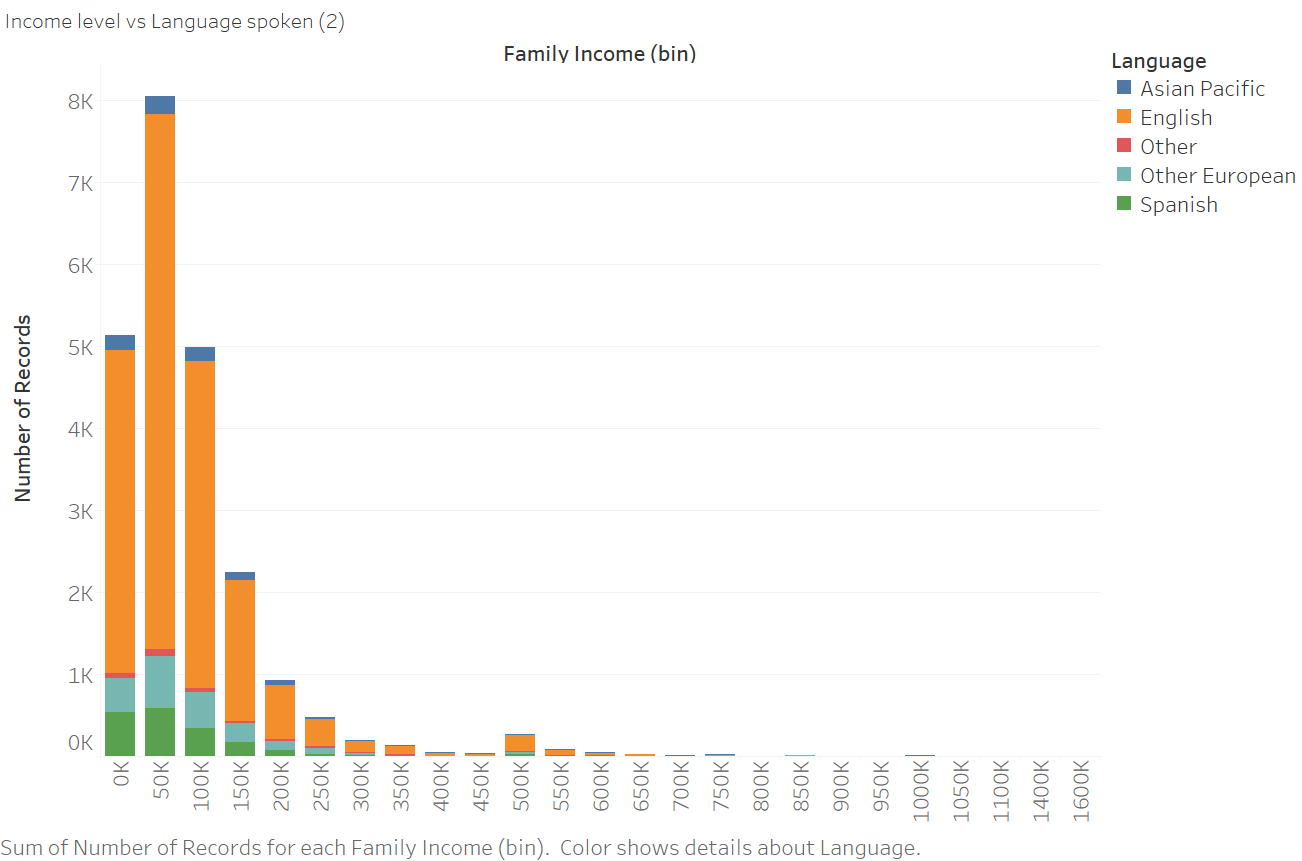
May 3, 2020

1. **TABLEAU VISUALIZATIONS**

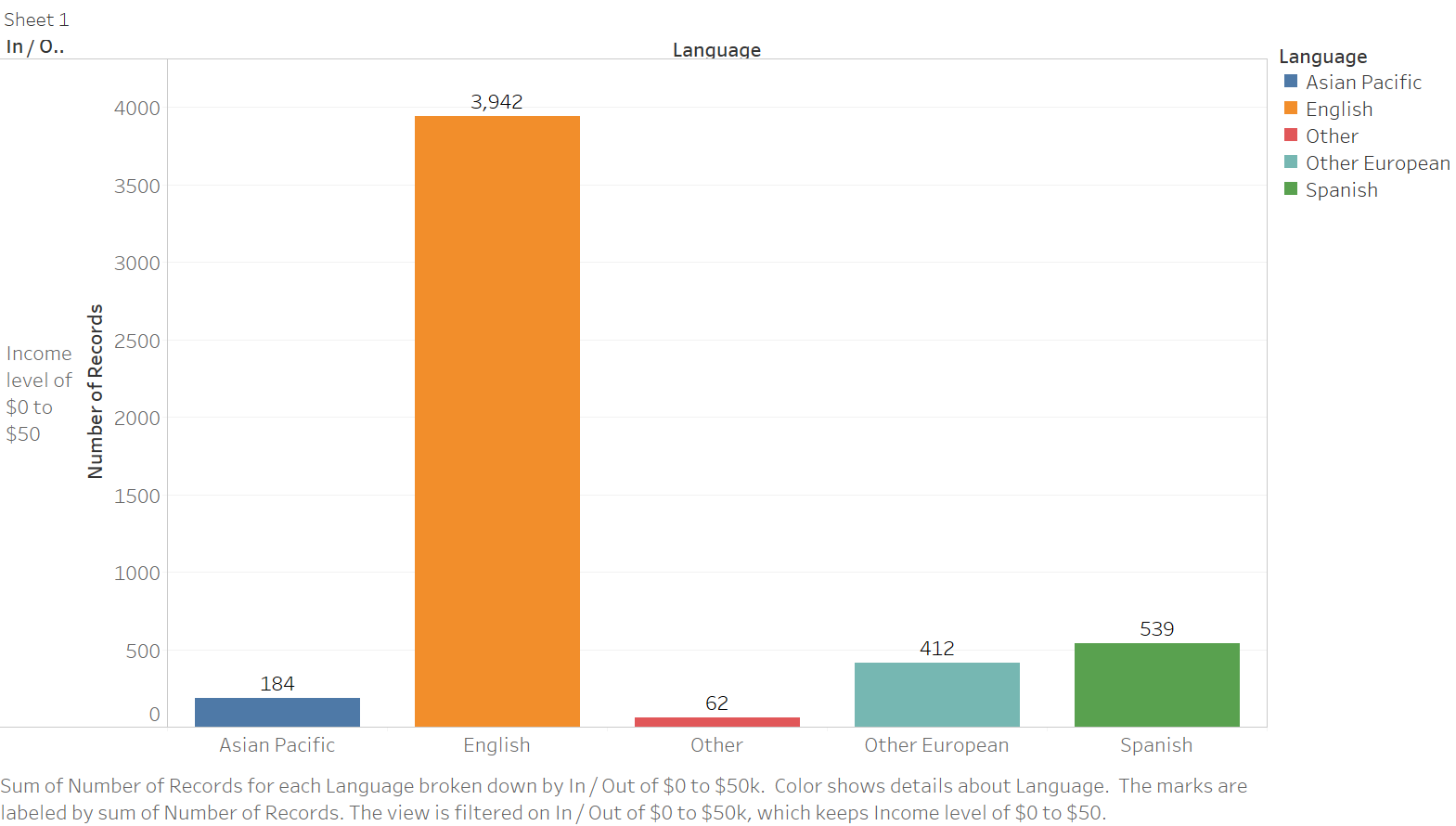
A) Bar chart for Income level groups and Language spoken at home



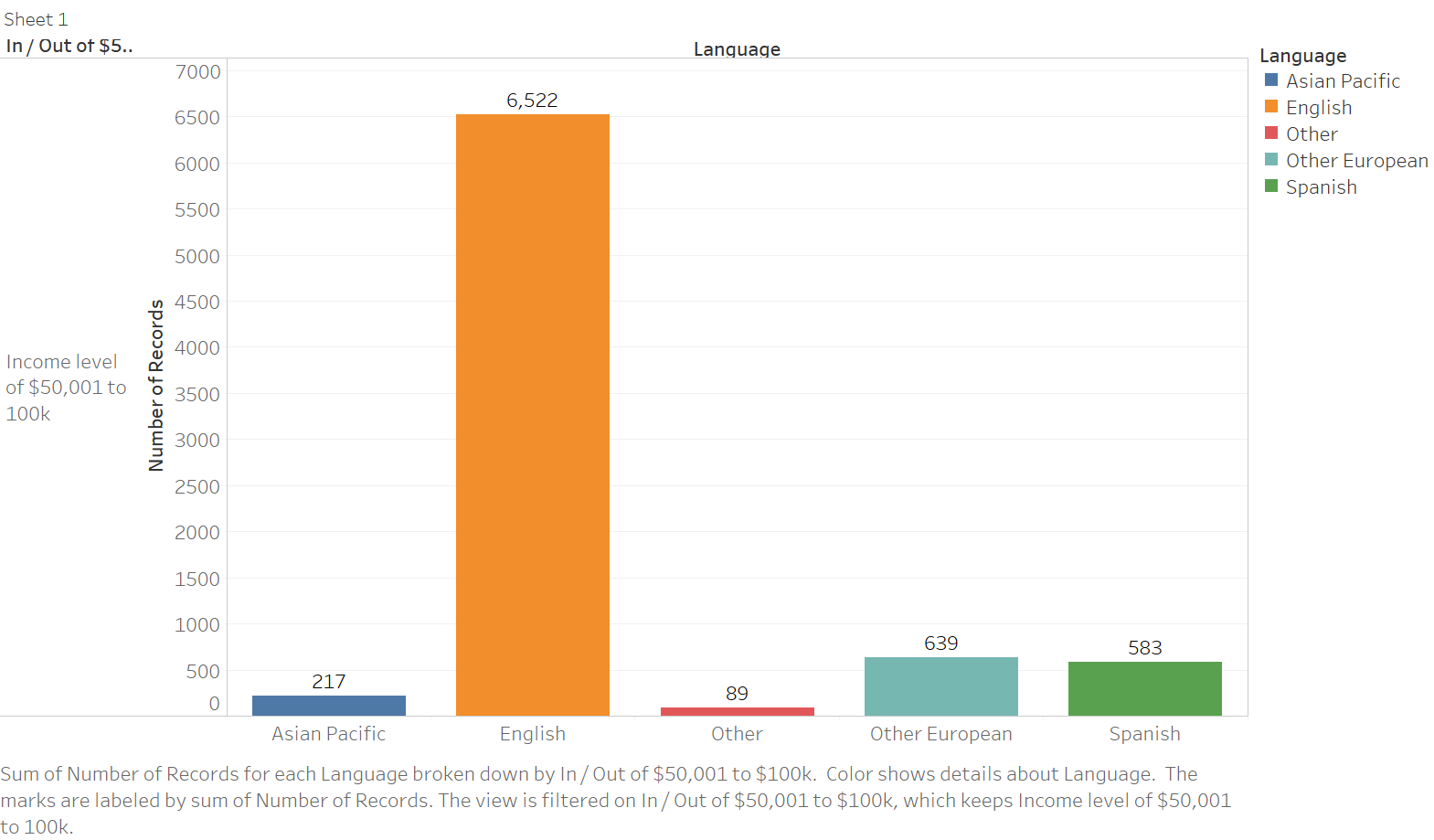




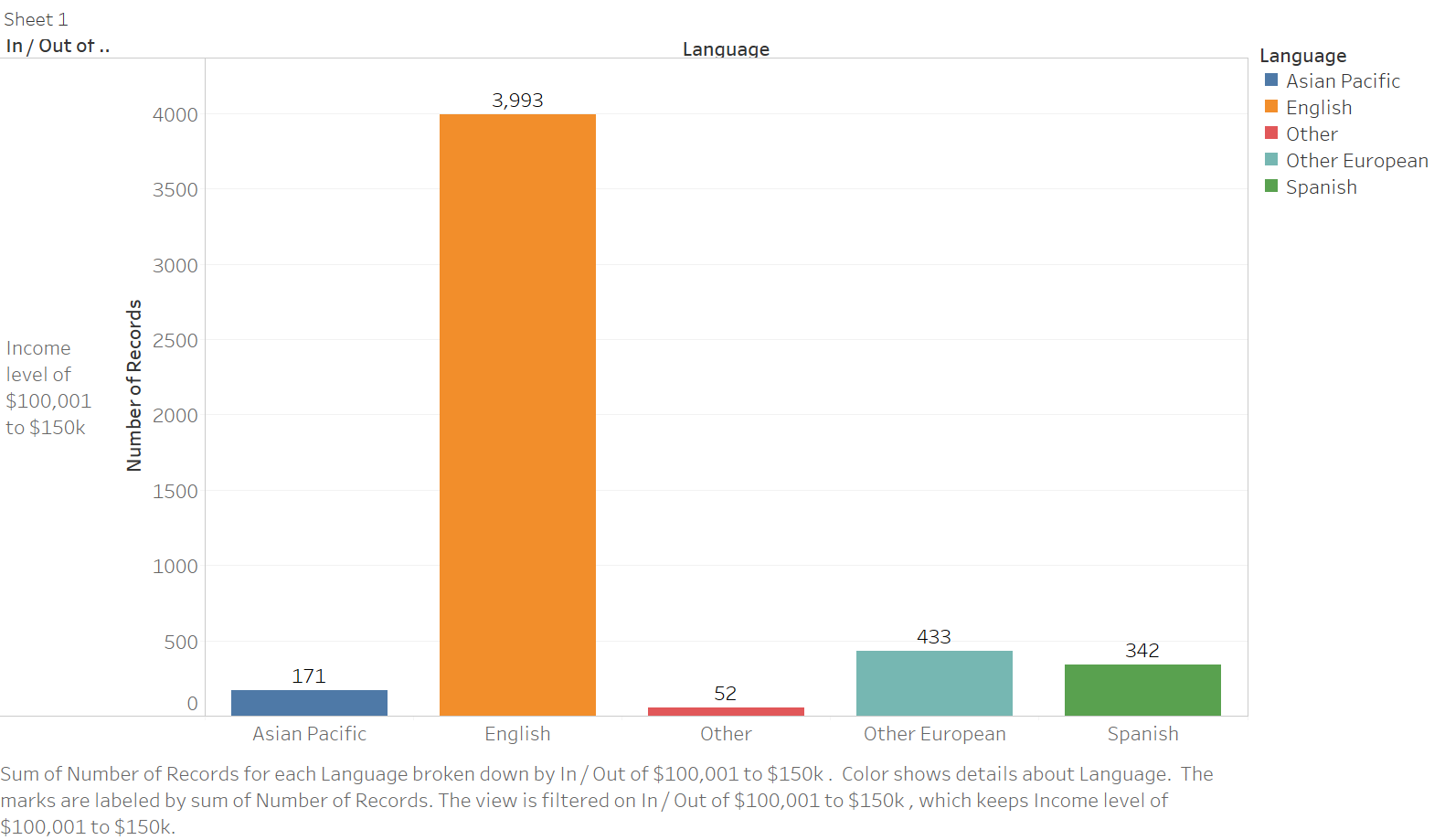
Number of households according to their language with Income level of $0 to $50k



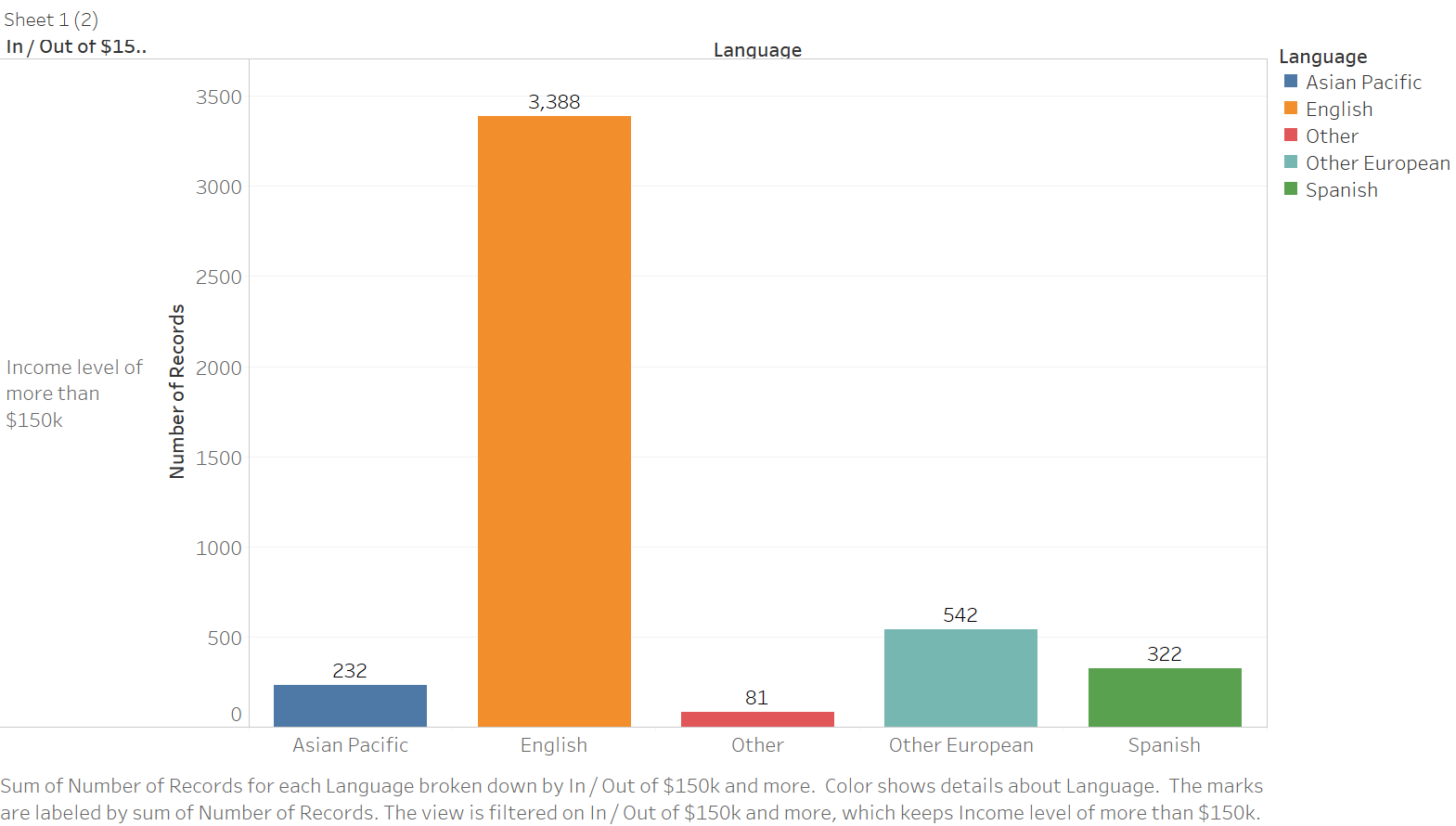
Number of households according to their language with Income level of $50,001 to $100k



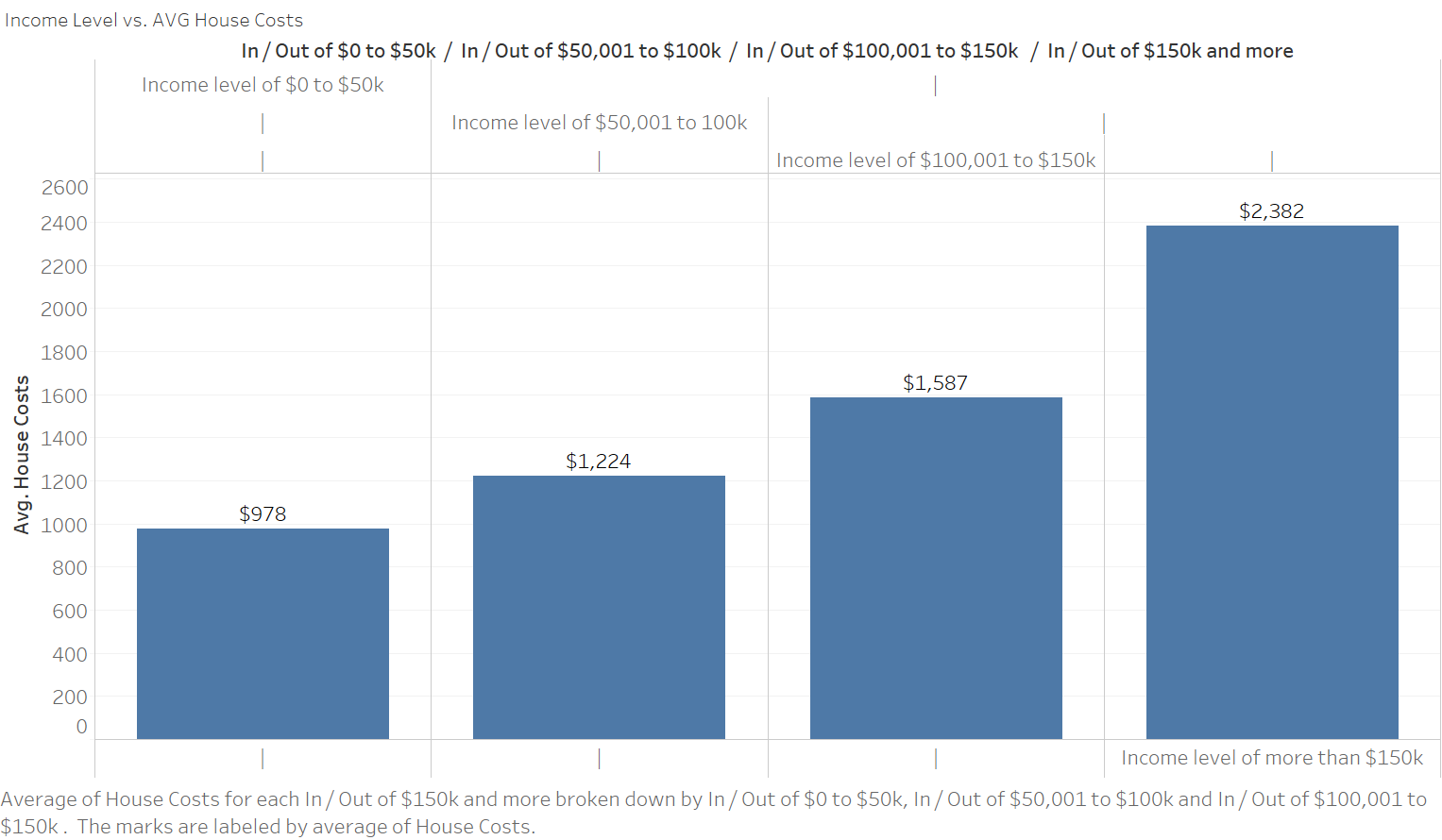
Number of households according to their language with Income level of $100,001 to $150k



Number of households according to their language with Income level of more than $150k

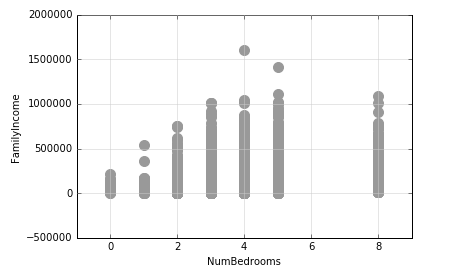


1. B) Relationship between Family Income levels and Average House Cost

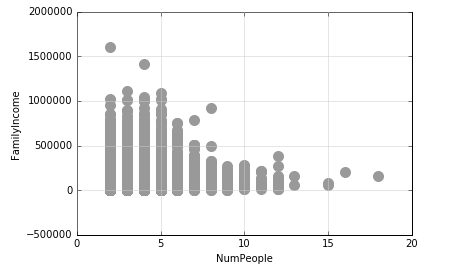


1. **SCATTERPLOTS FOR FAMILY INCOME AND EACH DEPENDENT VARIABLE**

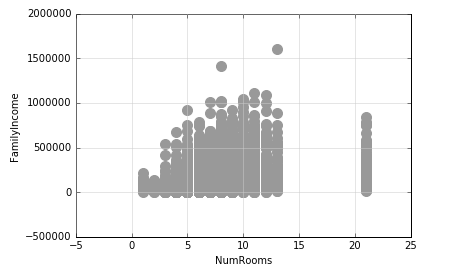
Family Income and Number of Bedrooms



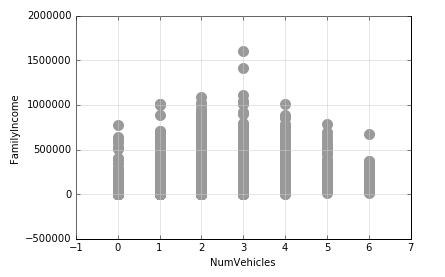
Family Income and Number of People



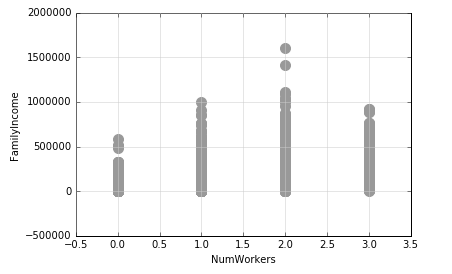
Family Income and Number of Rooms



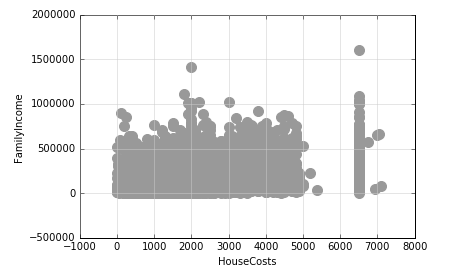
Family Income and Number of Vehicles



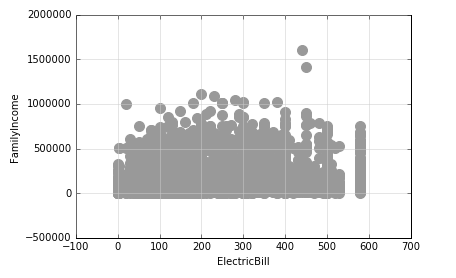
Family Income and Number of Workers



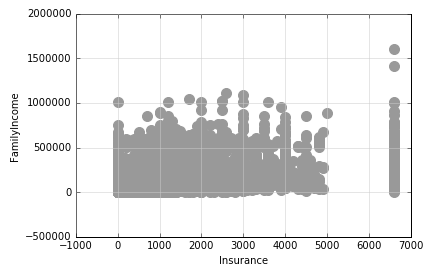
Family Income and House Costs



Family Income and Electric Bill



Family Income and Insurance



**3 MULTIPLE REGRESSION MODEL**

Four multiple regression models are presented here each with two variations of datasets and four independent variables. Second dataset had changes depending on these conditions:

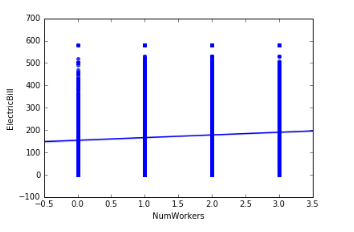
1. Family Income cannot be less than sum of expenses mentioned in the dataset (House costs, Insurance, and Electric Bill).
2. If Number of workers is zero than the Family Income cannot be more than zero.
   1. Savings cannot be considered income.
   2. External income like personal loans, bank loans, government and other benefits cannot be accounted as Income.
   3. Based on these assumptions, these rows were considered invalid.

These conditions were used to see if they bring any significant change to the analysis but they did not.

The Set of variables that gave the highest r-squared value at 0.295 were NumWorkers + HouseCosts + Insurance + ElectricBill. Theses were with the original untouched dataset. With the edited dataset (a and b == True), it dropped to 0.294. Same for, when 1600 rows were deleted (a or b == True) but that is not shown here.

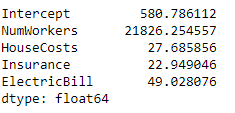
This set of variables was specifically chosen because even when they have a very small correlation with Family Income, theirs is the highest among all the other independent variables. When you compare this model with linear regression model of family income and number of workers (r square = 0.05), this multiple regression model (r square = 0.295) proves to be a far better option for the analysis. This model can explain ~30% variability of our dependent variable (family income) around the mean. Which is better than any other model in the dataset. Yes, we have a slightly higher number of variables, but the abs r square value is not penalized which also means adding variables is not bringing anything significant. P value is 0.000 for these variables so we can reject the null hypothesis that these variables are not associated with level of family income. This does not mean that hypothesis has almost 100% probability of being true. This extremely high p value can be explained by high values on side of the dataset, huge outliers, and vastly distributed data points.

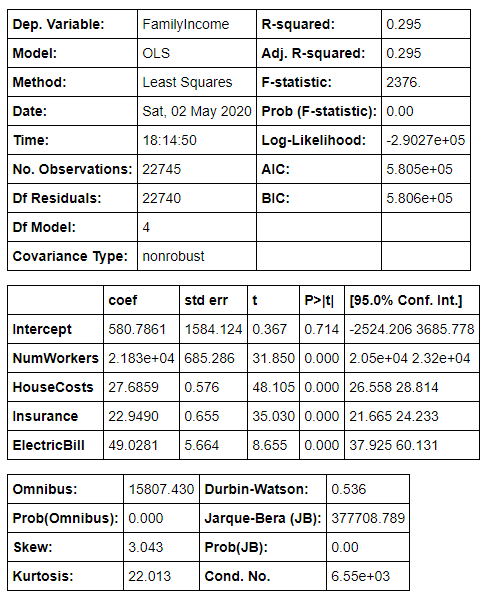
Multicollinearity is slightly above 0 for the independent variables as shown from one of the examples below.



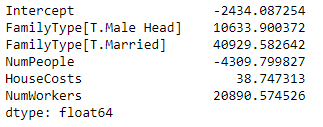
1. Without changes to the data

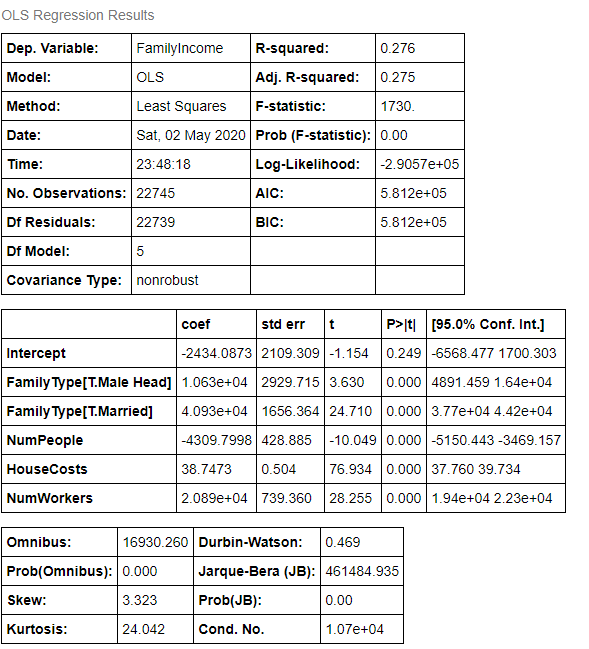
NumWorkers + HouseCosts + Insurance + ElectricBill



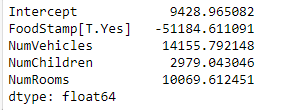


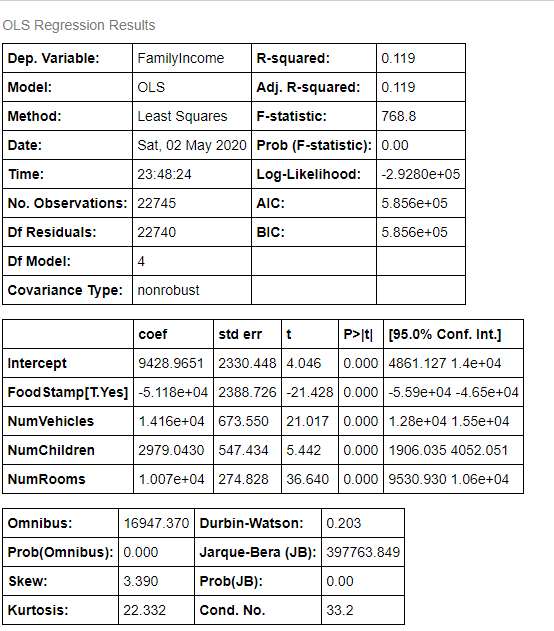
FamilyType + NumPeople + HouseCosts + NumWorkers



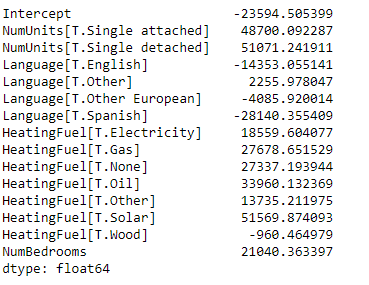


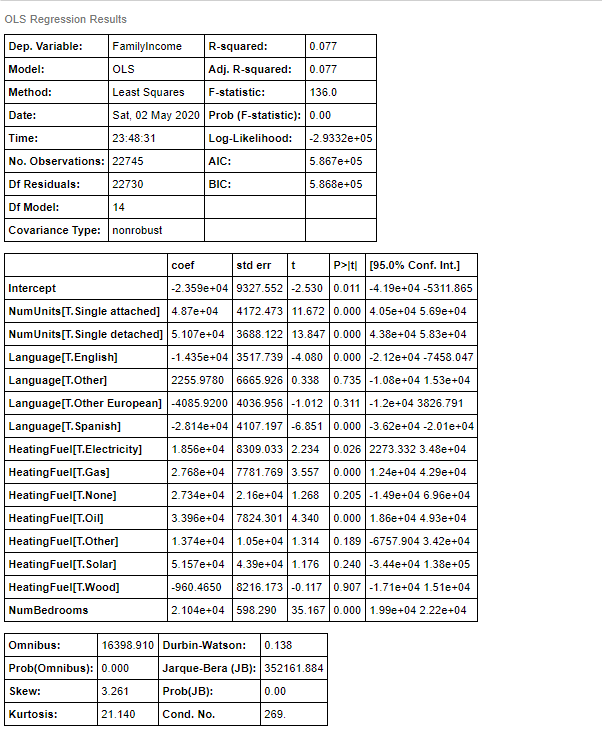
NumVehicles + NumChildren + NumRooms + FoodStamp





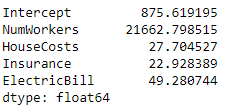
NumBedRooms + NumUnits + Language + HeatingFuel

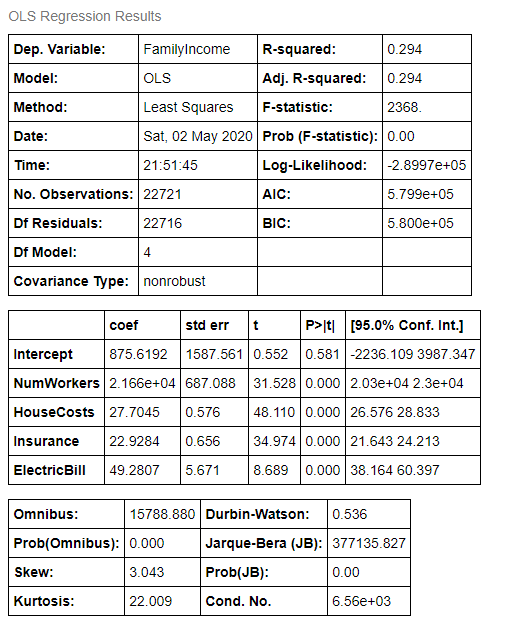




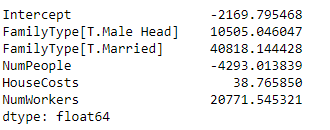
With changes to the data

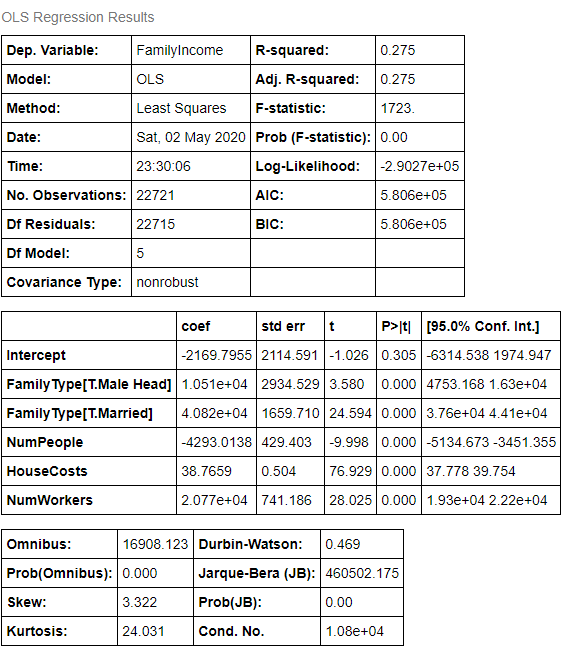
NumWorkers + HouseCosts + Insurance + ElectricBill



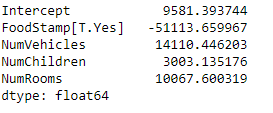


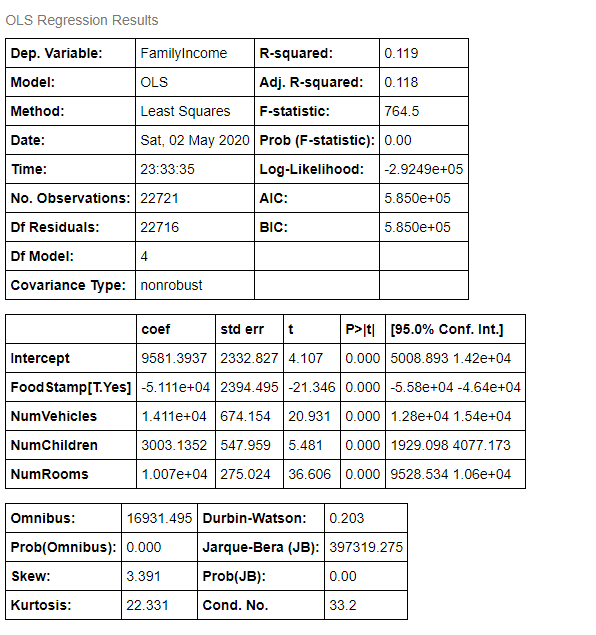
FamilyType + NumPeople + HouseCosts + NumWorkers



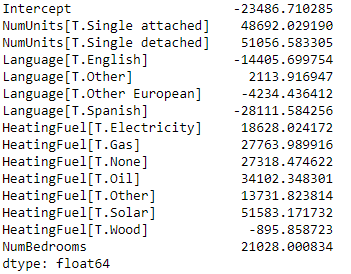


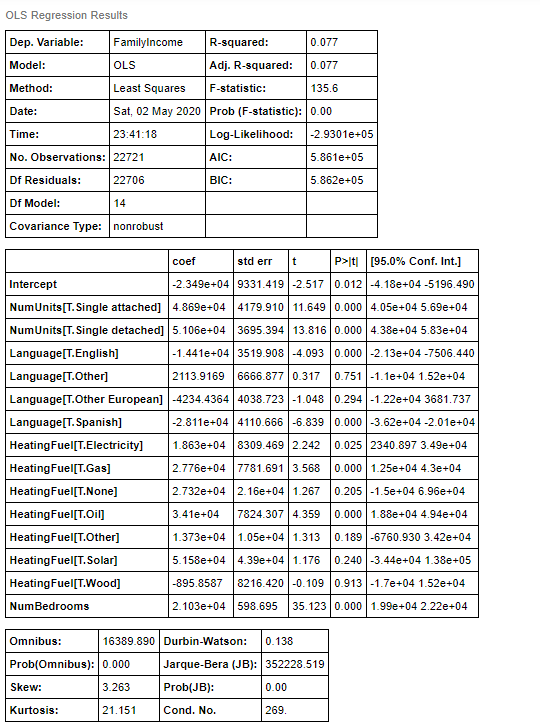
NumVehicles + NumChildren + NumRooms + FoodStamp





NumBedRooms + NumUnits + Language + HeatingFuel



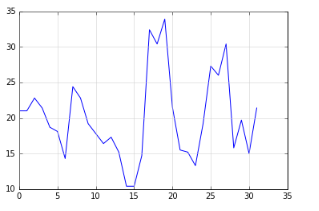


**4 CARS.CSV VISUALIZATIONS**

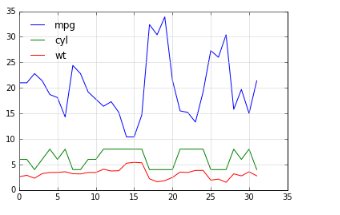
**Examine cars Dataset**

First column with car names has no title. There are no null values. There are three data types: float64, int64, and object with 32 rows in total.

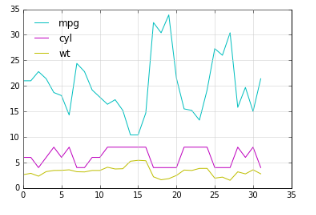
**Line Chart using mpg**



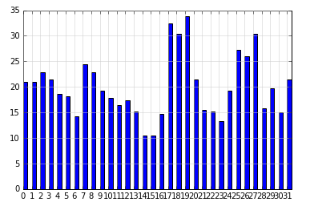
**Create a plot using cyl, wt, and mpg**



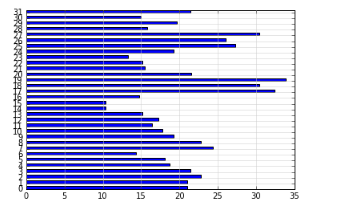
**Change the colors of the line plot**



**Create a bar chart for mpg**



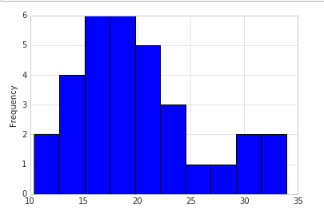
**Make the above chart horizontal**



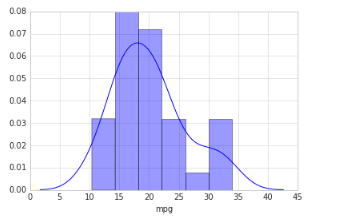
**What is the maximum mpg in your cars dataset**

33.8999999

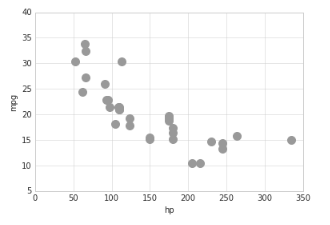
**Create a histogram for mpg**



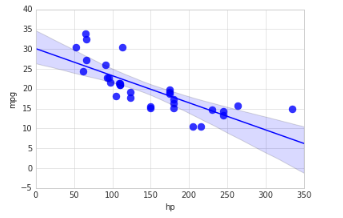
**Create a distribution plot for mpg using Seaborn**



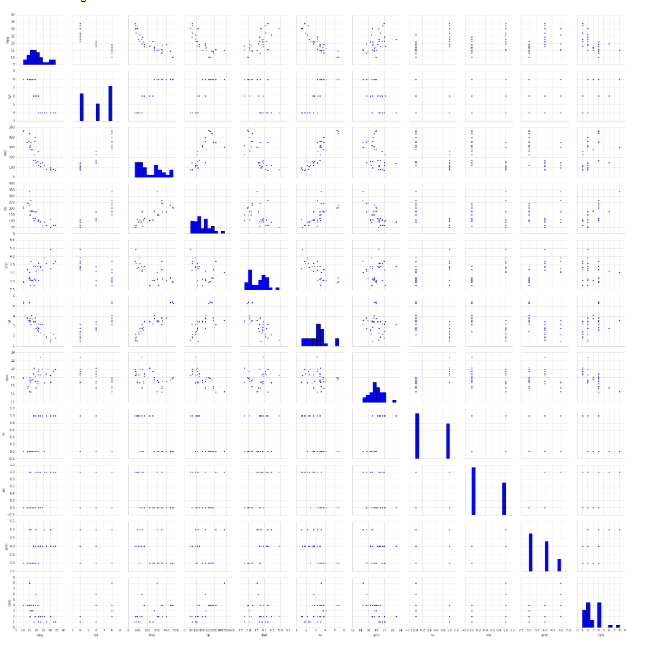
**Create a Scatterplot (x=hp, y=mpg)**



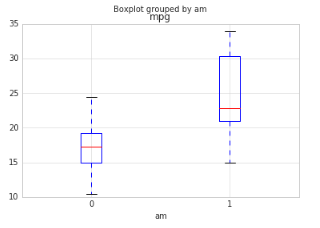
**Use Seaborn for scatterplot and linear regression model fit**



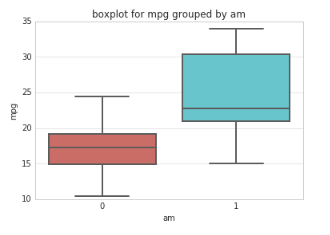
**Use Seaborn to create pair plot for all attributes**



**Create a boxplot for mpg grouped by am**



**Using Seaborn create boxplot for mpg grouped by am**



**Insights for above visualizations**

1. Mpg(miles per gallon) decreases as hp (car’s horsepower) increases. Negative relationship
2. Cyl (Number of cylinders) is lowest with the highest numbers of mpg’s.
3. Wt(weight) is lowest with the highest numbers of mpg’s.
4. (Considering 0 = manual and 1 = automatic) automatic cars have a higher median than Q3 of manual cars for mpg.
5. Cyl (cylinder) and wt (weight) have a positive relationship.