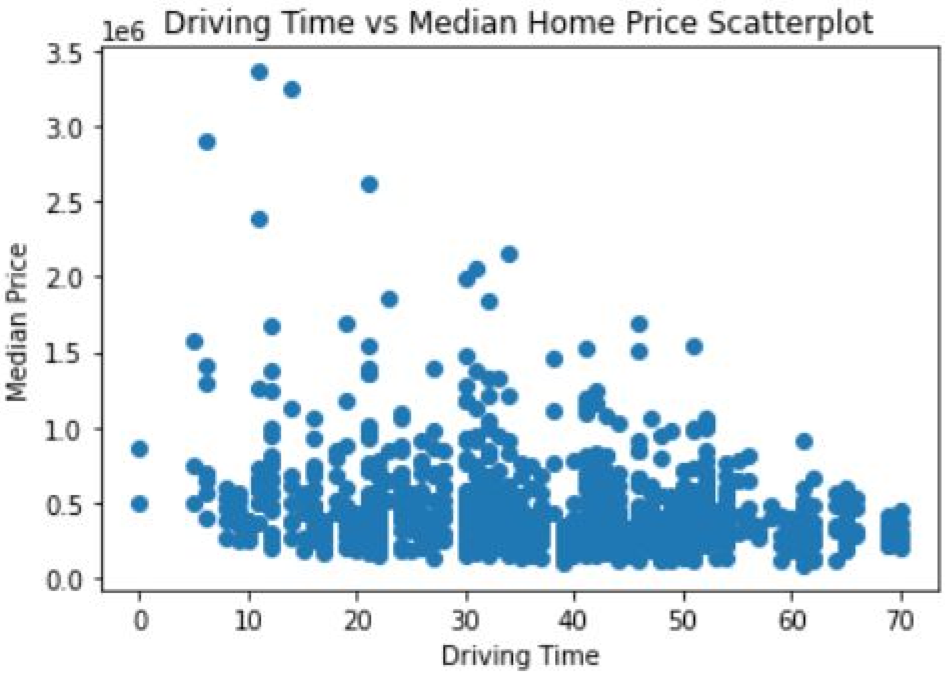
Gurteg Singh
ECN 145

Problem Set 1

1.) Logfile etc.
2.) How many observations are in the dataset? Summarize the data and report the average values for the commuting time variables? What is the longest driving commute to Downtown Boston?

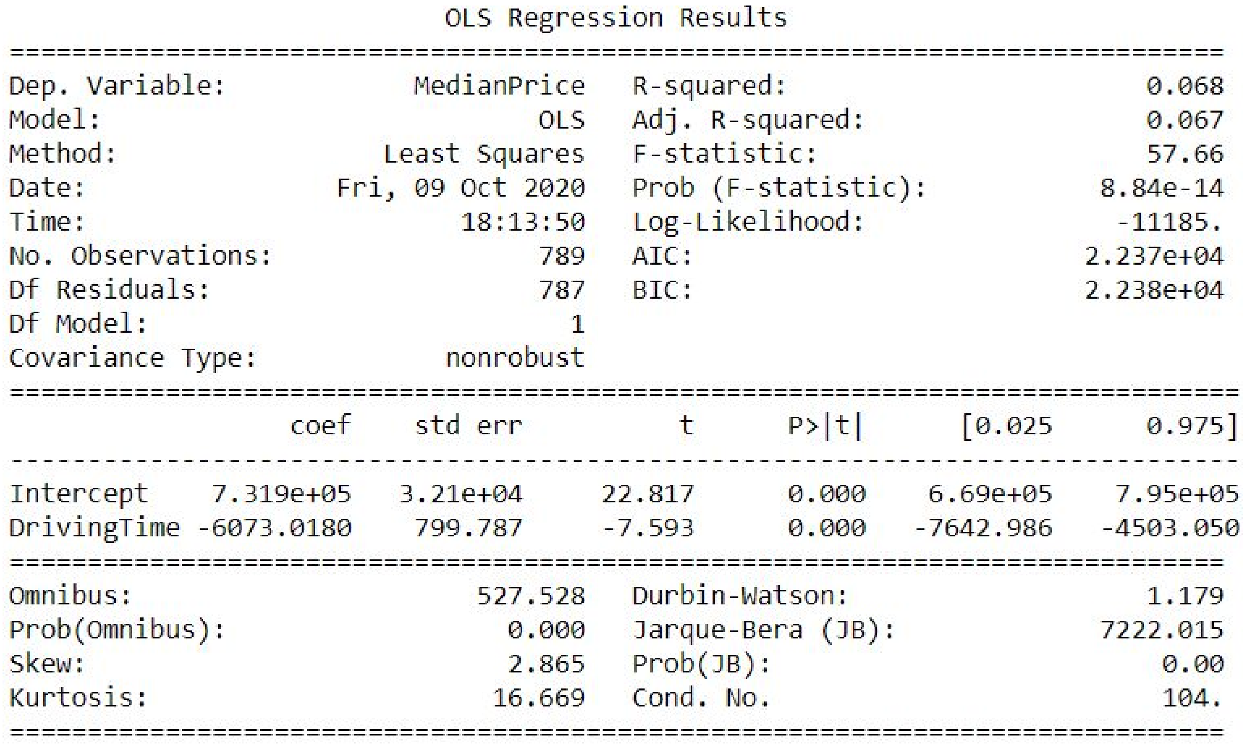
There are 1,140 observations (however only 881 MedianPrice counts)
Average DrivingTime = 39.991228 minutes
Average TransitTime = 92.416667 minutes
Longest Driving Time to Downtown Boston = 62.440000 minutes

3.)



Here, we can see that the overall correlation between MedianPrice and DrivingTime is negative. Hence, on average, as driving time increases, MedianPrice tends to decrease. The units of MedianPrice are in $100,000 intervals, and the units of DriveTime are in one minute intervals.

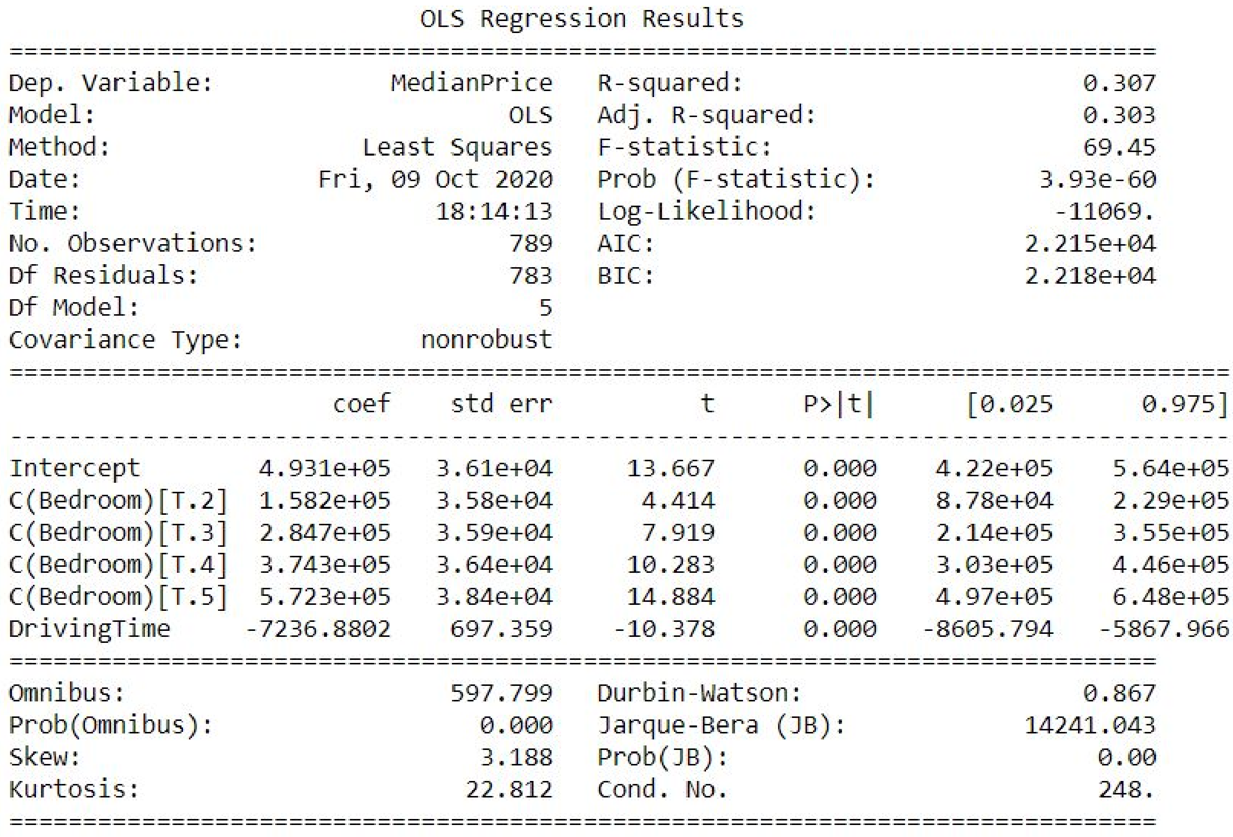
4.)



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            MedianPrice   R-squared:                       0.068
Model:                            OLS   Adj. R-squared:                  0.067
Method:                 Least Squares   F-statistic:                     57.66
Date:                Fri, 09 Oct 2020   Prob (F-statistic):           8.84e-14
Time:                        18:13:50   Log-Likelihood:                -11185.
No. Observations:                 789   AIC:                         2.237e+04
Df Residuals:                     787   BIC:                         2.238e+04
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept   7.319e+05   3.21e+04     22.817      0.000    6.69e+05    7.95e+05
DrivingTime -6073.0180   799.787     -7.593      0.000   -7642.986   -4503.050
==============================================================================
Omnibus:                      527.528   Durbin-Watson:                   1.179
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7222.015
Skew:                           2.865   Prob(JB):                         0.00
Kurtosis:                      16.669   Cond. No.                         104.
==============================================================================
```

Here, we can see that this model has some correlation, and now covers 6.8% of the data’s variation. We can also see that on average, each additional minute of DriveTime reduces housing prices (MedianPrice) by about $6073.02. This goes along with our postulate that housing prices tend to drop farther from the city. Our coefficient of correlation is significant with a T value of 22.817 letting us reject the null hypothesis that DrivingTime is not correlated with MedianPrice, and our intercept starts around $73,190.

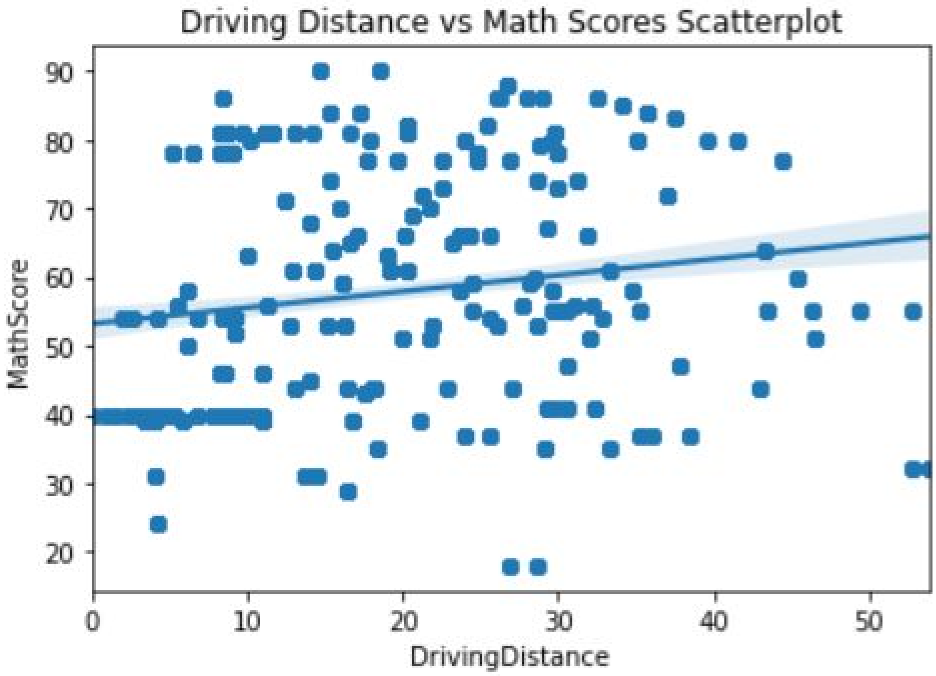
5.)



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            MedianPrice   R-squared:                       0.307
Model:                            OLS   Adj. R-squared:                  0.303
Method:                 Least Squares   F-statistic:                     69.45
Date:                Fri, 09 Oct 2020   Prob (F-statistic):           3.93e-60
Time:                        18:14:13   Log-Likelihood:                -11069.
No. Observations:                 789   AIC:                         2.215e+04
Df Residuals:                     783   BIC:                         2.218e+04
Df Model:                           5
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept       4.931e+05   3.61e+04     13.667      0.000    4.22e+05    5.64e+05
C(Bedroom)[T.2] 1.582e+05   3.58e+04      4.414      0.000    8.78e+04    2.29e+05
C(Bedroom)[T.3] 2.847e+05   3.59e+04      7.919      0.000    2.14e+05    3.55e+05
C(Bedroom)[T.4] 3.743e+05   3.64e+04     10.283      0.000    3.03e+05    4.46e+05
C(Bedroom)[T.5] 5.723e+05   3.84e+04     14.884      0.000    4.97e+05    6.48e+05
DrivingTime    -7236.8802    697.359    -10.378      0.000   -8605.794   -5867.966
==============================================================================
Omnibus:                      597.799   Durbin-Watson:                   0.867
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            14241.043
Skew:                           3.188   Prob(JB):                         0.00
Kurtosis:                      22.812   Cond. No.                         248.
==============================================================================
```

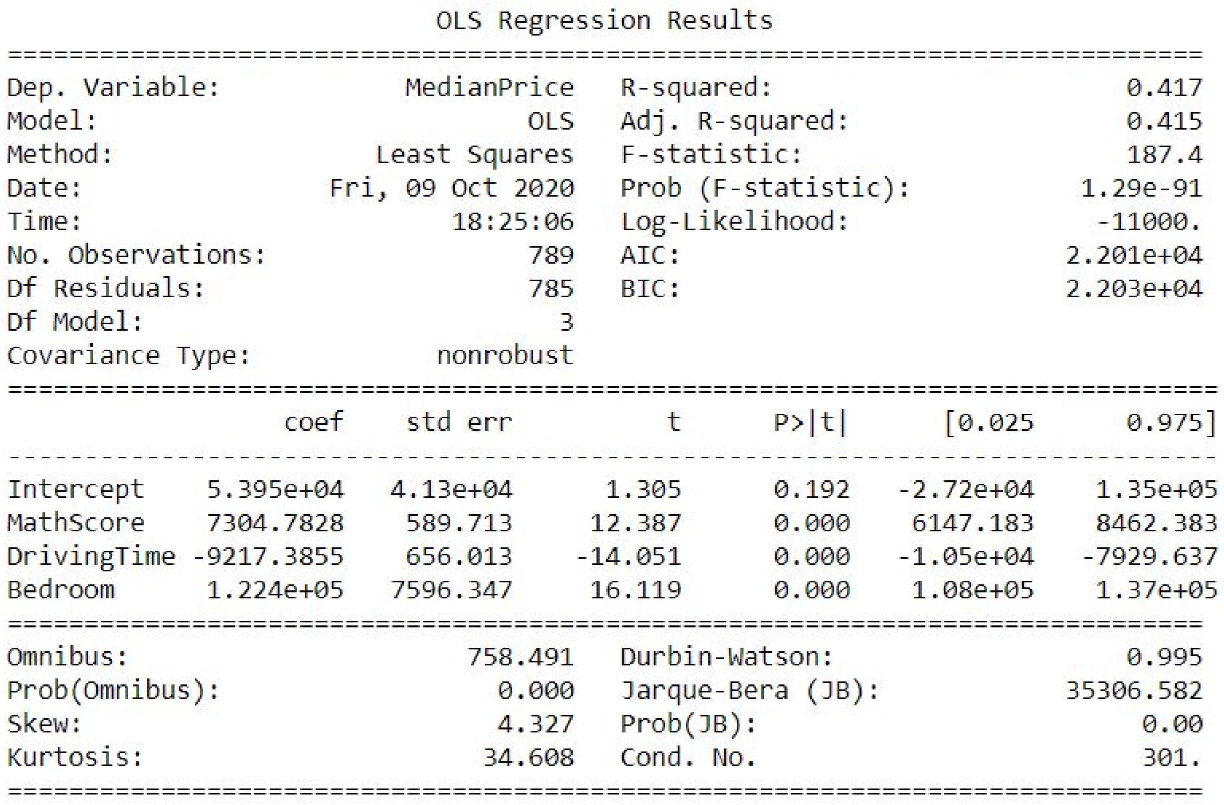
Here we can see that each coefficient of Bedrooms = 2 through 5 are statistically significant, and each tends to increase in chronological order based on our base value, Bedrooms = 1 which is now incorporated into our intercept. For example, *holding all other variables constant,* a house with five Bedrooms costs about $57,230 more than a house with only one bedroom, similarly a house with 4, 3, or 2 bedrooms, would on average cost $37,430, $28,470, & $15,820 more than a house with a single bedroom respectively. Our R-Squared has increased and our model now covers around 30.7% of the data variation, and our Adjusted R-Squared falls slightly more due to the increase in variables. In addition, our intercept decreases as more variables are accounted for, decreasing our preliminary omitted variable bias.

6.)



Here we can see a weaker trend between Mathscores and DrivingDistance. It seems to be that there is little correlation between the variables, although the line of best fit does suggest a slightly positive correlation, that is; each additional mile an individual must drive to reach the city, they're math scores tend to rise slightly. In addition, the terms of homoscedasticity may be getting violated as the data variation tends to increase farther and farther down our plot.

7.)



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            MedianPrice   R-squared:                       0.417
Model:                            OLS   Adj. R-squared:                  0.415
Method:                 Least Squares   F-statistic:                     187.4
Date:                Fri, 09 Oct 2020   Prob (F-statistic):           1.29e-91
Time:                        18:25:06   Log-Likelihood:                -11000.
No. Observations:                 789   AIC:                         2.201e+04
Df Residuals:                     785   BIC:                         2.203e+04
Df Model:                           3
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept    5.395e+04   4.13e+04      1.305      0.192   -2.72e+04    1.35e+05
MathScore    7304.7828    589.713     12.387      0.000    6147.183    8462.383
DrivingTime -9217.3855    656.013    -14.051      0.000   -1.05e+04   -7929.637
Bedroom      1.224e+05   7596.347     16.119      0.000    1.08e+05    1.37e+05
==============================================================================
Omnibus:                      758.491   Durbin-Watson:                   0.995
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            35306.582
Skew:                           4.327   Prob(JB):                         0.00
Kurtosis:                      34.608   Cond. No.                         301.
==============================================================================
```

The coefficient of math is “correct” in this model, as it is statistically significant. What this model tells us is that - *holding all other variables constant* - for each additional percentage point increase in the math score, MedianPrice tends to rise on average, $7304.78.

8.) Based on our previous scatterplot, we saw a positive correlation between DriveDistance and MathScore. Hence, as math scores tend to rise, DriveTime will also increase, thereby reducing the value of MedianPrice housing, as it is located further away from the city. Also we can take away from our #6 scatterplot that our slope is less than 1, alluding to the fact that the delta in drive time is greater than that of MathScore for each additional increase in percentage point. Hence we have to travel much further to get an increase is mathscore. Now that we have even more variables in our model, DrivingTime ecompasses less omitted variable bias, and shows that the community is even more sensitive to commute time.

9.) Yes, the coefficient both makes sense as we saw previously that the people in this city prefer a shorter commute, along with the fact that it is statistically significant. Therefore, we can reject the null, that the Beta value of DrivingTime is not correlated.

10.) One reason this may be the case is that a large home may be inhabited by a family. A smaller home may be owned/rented by a student who may prefer a cheaper home slightly further from the city and be less sensitive to DriveTime. **Part B:** One way to test this correlation is to add an interaction term between Bedroom and Mathscore and see if it is statistically significant, and whether it’s negatively or positively correlated.