**Keangcheang Ung**

keangcheang@gmail.com

979-250-2429

**Feature Selection Modeling for Predicting House Price in Nashville**

Springboard Capstone Project 1

### **Problem and Objective**

Real estate in the US has been an interesting topic for not only developers or investors but also to academic researchers or data enthusiast who wants to understand more about how to predict the house price. In this project, we will focus on the real estate in Nashville where the markets trends indicated an increase of $19,000 about 7% in median home sale over the past year. The average price per square foot for this same period rose to $195, up from $180.

The objective of this project was to determine the features that affects house price in Nashville using linear regression with regularization. In order to achieve this, we had to take several procedures like collecting and preparing the data set ready for modeling. Visualization for this project was essential to find the insight from different features against price. We trained the data set for modeling and test it with the tested data set. During this process, we also had to use regularization like Lasso for feature selections.

### **Data Wrangling and Visualization**

With regard to the capstone project which is about feature selections of important attributes that determine house price in Nashville, there were some data wrangling techniques used to prepare the data and ready for modeling.

First of all, checking the format of each feature regarding their datatypes and their observation sizes was my first step. This way I would know if there should be any datatype conversions I should take or any missing values I had to tackle. In this dataset, there were 56,000+ observations with 29 variables and 31 columns (List of features shown in Table 1). There were several features that contained missing values such as the exterior wall, tax district …etc. The house price attribute was numeric as well as land value, building value and total value which were essential to this project.

Table 1: List of Features

|  |  |
| --- | --- |
| **Features** | **Type** |
| Sale Price | Numeric |
| Legal Reference | String |
| Sold as Vacant | String |
| Multiple Parcels Involved in Sale | String |
| Owner Name | String |
| Address | String |
| City | String |
| State | String |
| Acreage | Numeric |
| Tax District | String |
| Neighborhood | Numeric |
| image | String |
| Land Value | Numeric |
| Building Value | Numeric |
| Total Value | Numeric |
| Finished Area | Numeric |
| Foundation Type | String |
| Year Built | Numeric |
| Exterior Wall | String |
| Grade | String |
| Bedrooms | Numeric |
| Full Bath | Numeric |
| Half Bath | Numeric |

Next, I had to check the distribution of price variable whether it was a normal distribution and make correction according to that. The result showed that the price is not normally distributed (Figure 1) and it was hard to interpret, thus I took the log transformation of the price variable so it can be interpretable and became a bell-shaped distribution with a little of skewness to the left. This indicated that some houses were sold higher than the average price which was logical and rational to the current housing market in Nashville. Moreover, total value, land value and building value were also converted to log value so that I could plot them against price in order to check their correlations (Figure 1). Like expected, they all had a positive relationship with price. In addition to this, I also made a heat-map with the correlation matrix to delve more into their relationship with price as shown in Figure 2. Land value, total value, finished value and full bath had moderate correlations with price.

Figure 1: Log Sale Price Distribution and Scatter plot of Log Sale Price and Log Total Value

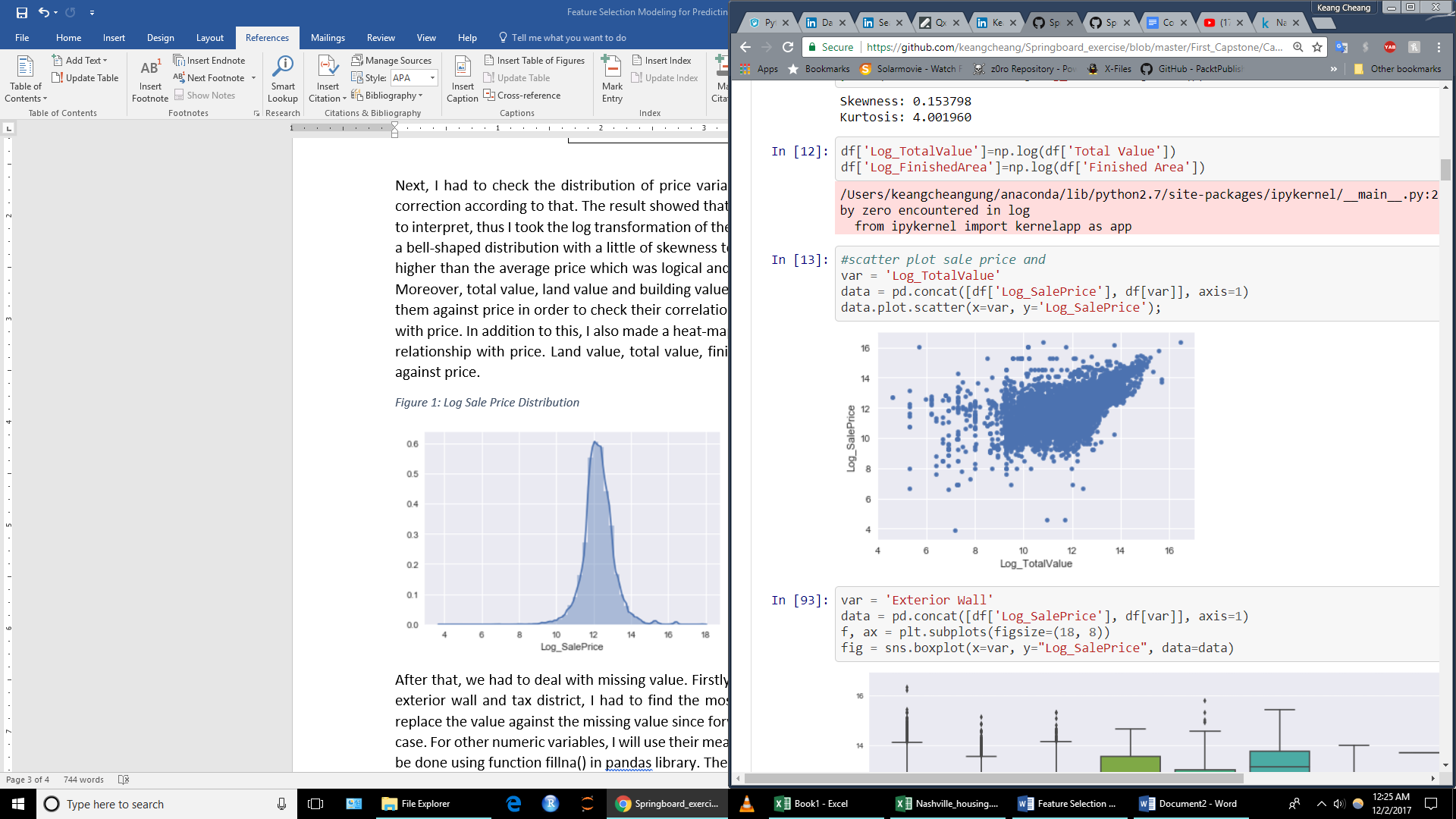
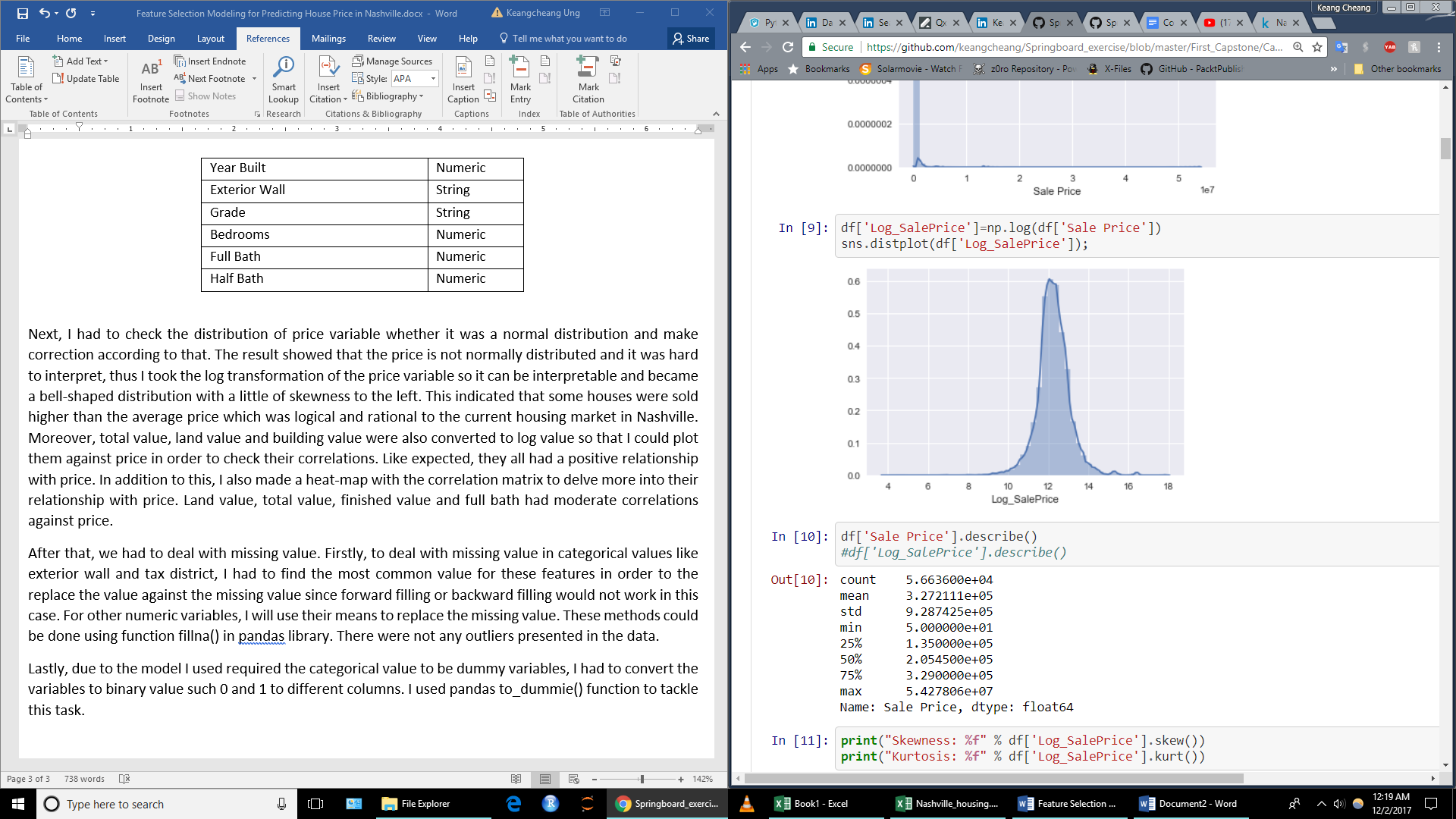
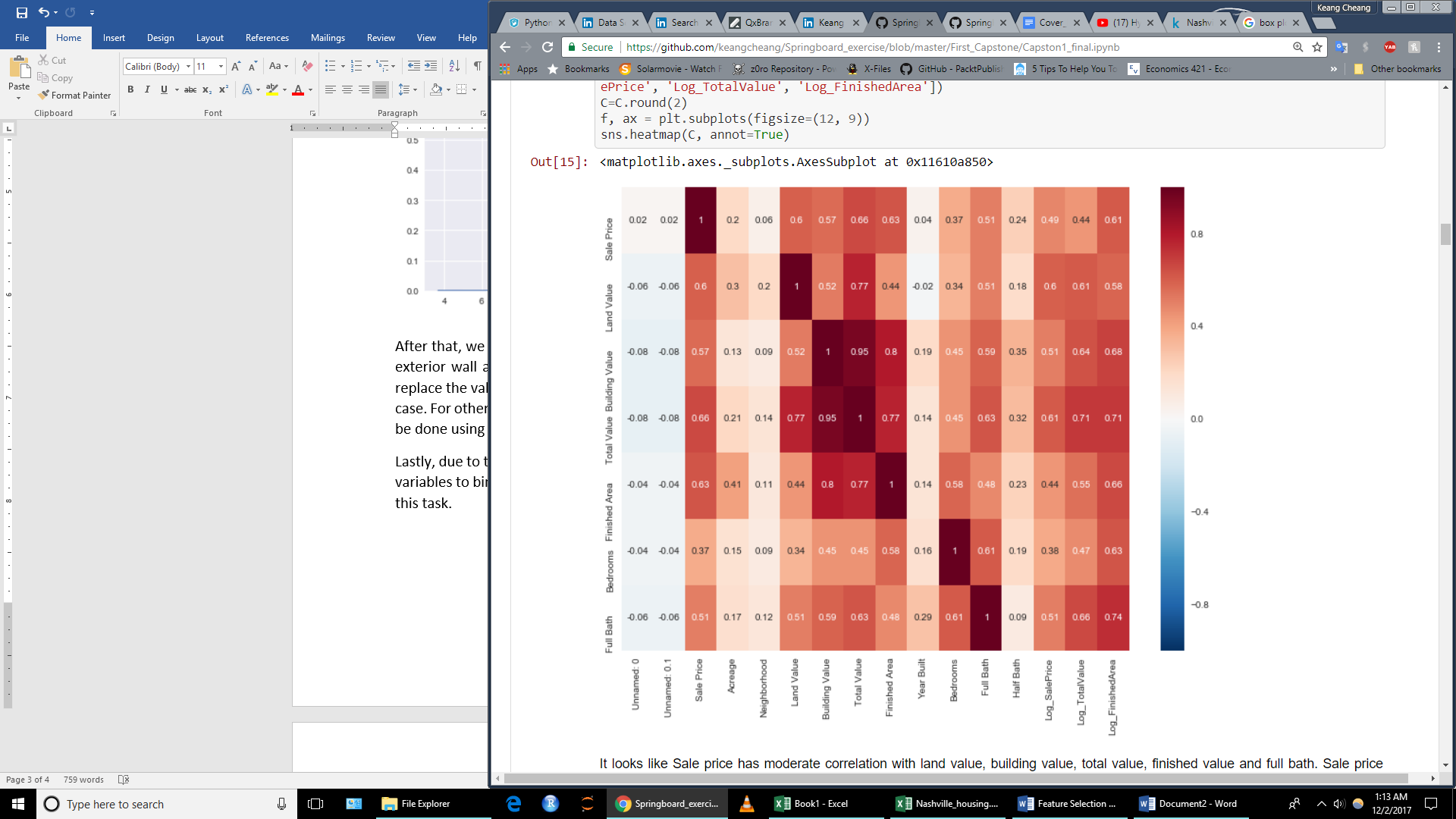
****

Figure 2: Heatmap with Correlation

After that, we had to deal with missing value. Firstly, to deal with missing value in categorical values like exterior wall and tax district, I had to find the most common value for these features in order to the replace the value against the missing value since forward filling or backward filling would not work in this case. For other numeric variables, I will use their means to replace the missing value. These methods could be done using function fillna() in pandas library. There were not any outliers presented in the data.

Lastly, due to the model I used required the categorical value to be dummy variables, I had to convert the variables to binary value such 0 and 1 to different columns. I used pandas to\_dummie() function to tackle this task.

1. **Modeling and Result**

Since our target value or dependent variable was continuous, we have to use machine learning that support this format and one that simple but powerful model was linear regression. Before we going any further, we had to discuss the concept of BLUE (Best Linear Unbiased Estimators) which was very important for the assumption we were going to make for this model to work. These are the important assumptions for this model:

* The residual term must be normally distributed which mean it has mean of zero and standard deviation of 1
* The residual term must be homoscedasticity and independent
* There should be no perfect collinearity
* The sample data set must be random

In the modeling phase, we had to split the data into training and test set so that we can avoid over-fitting. We used Skit-learn library to split the data using 70/30 splits and results showed: Training and testing set sizes (21131, 32) (9056, 32). With this training set, we had to train the data with linear regression by fitting the target variable with independent variables. Next, we used this fitting value to compare with the test data set to generate the accuracy score or a matric of MSE. The accuracy scores were as following:

**Testing Score: 0.411476963149**

**Training MSE: 0.276040455527**

**Testing MSE: 0.269870736571**

The score of 0.41 was not necessarily good but also not so bad either and that is why we believe using regularization might be able to improve the results of this score. The MSE of training and test were very similar which indicated there should be no over-fitting presented in the model. In addition to this result, we also had not forget the statistical result summarized below. The table described each coefficient with its parameter values and standard error as well as p-value. There were two main coefficients that are not statistically significant: the bedroom and G\_E. They were not significant because their p-value were higher than 5% of significance level we assumed for this model. The R2 known as the coefficient of determination for the model is 0.42 which was not bad either and it reflected how much information that the independent variables helped describe the data. The adjusted R2 was 0.419 which was slightly lower than R2.

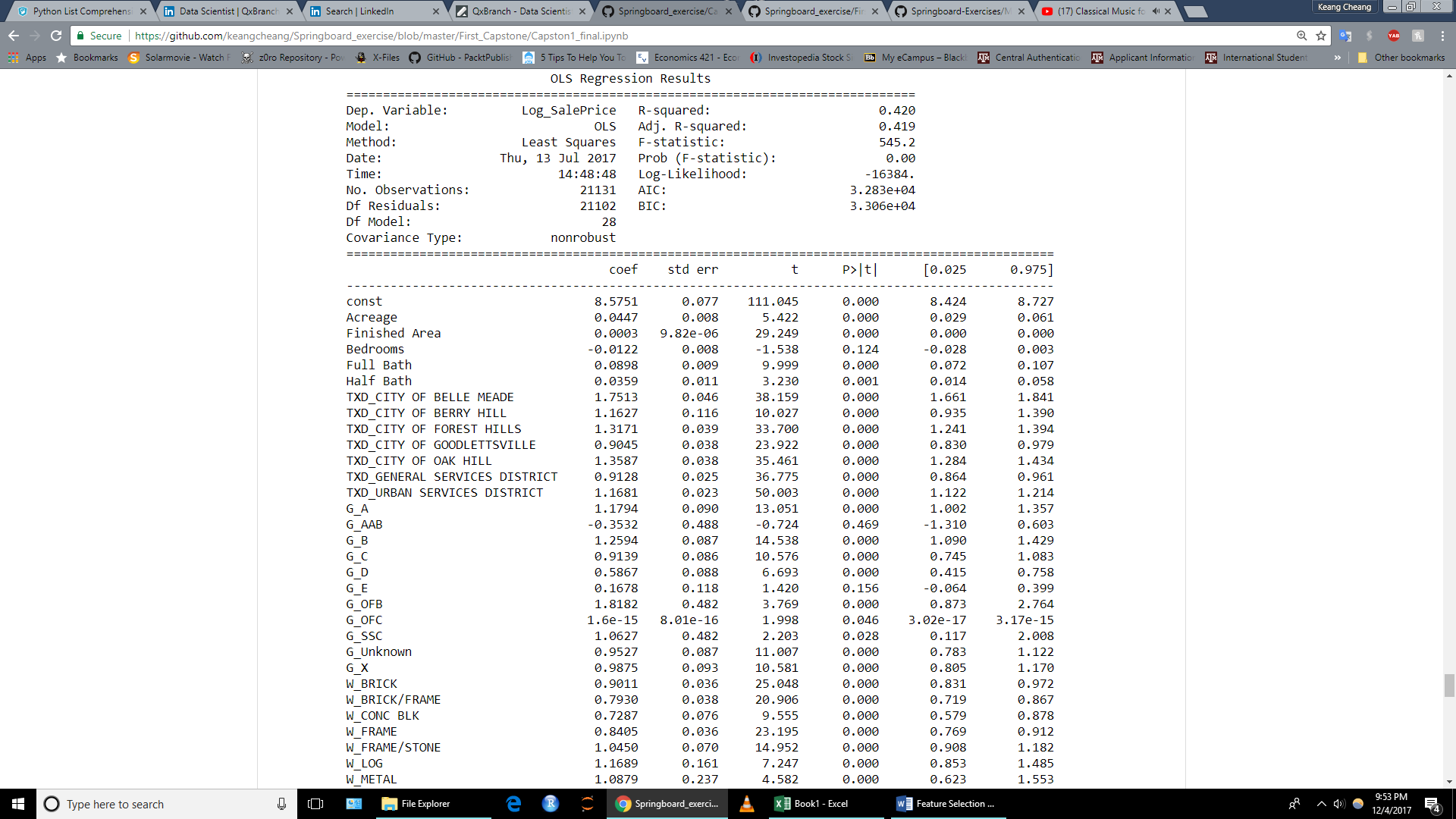
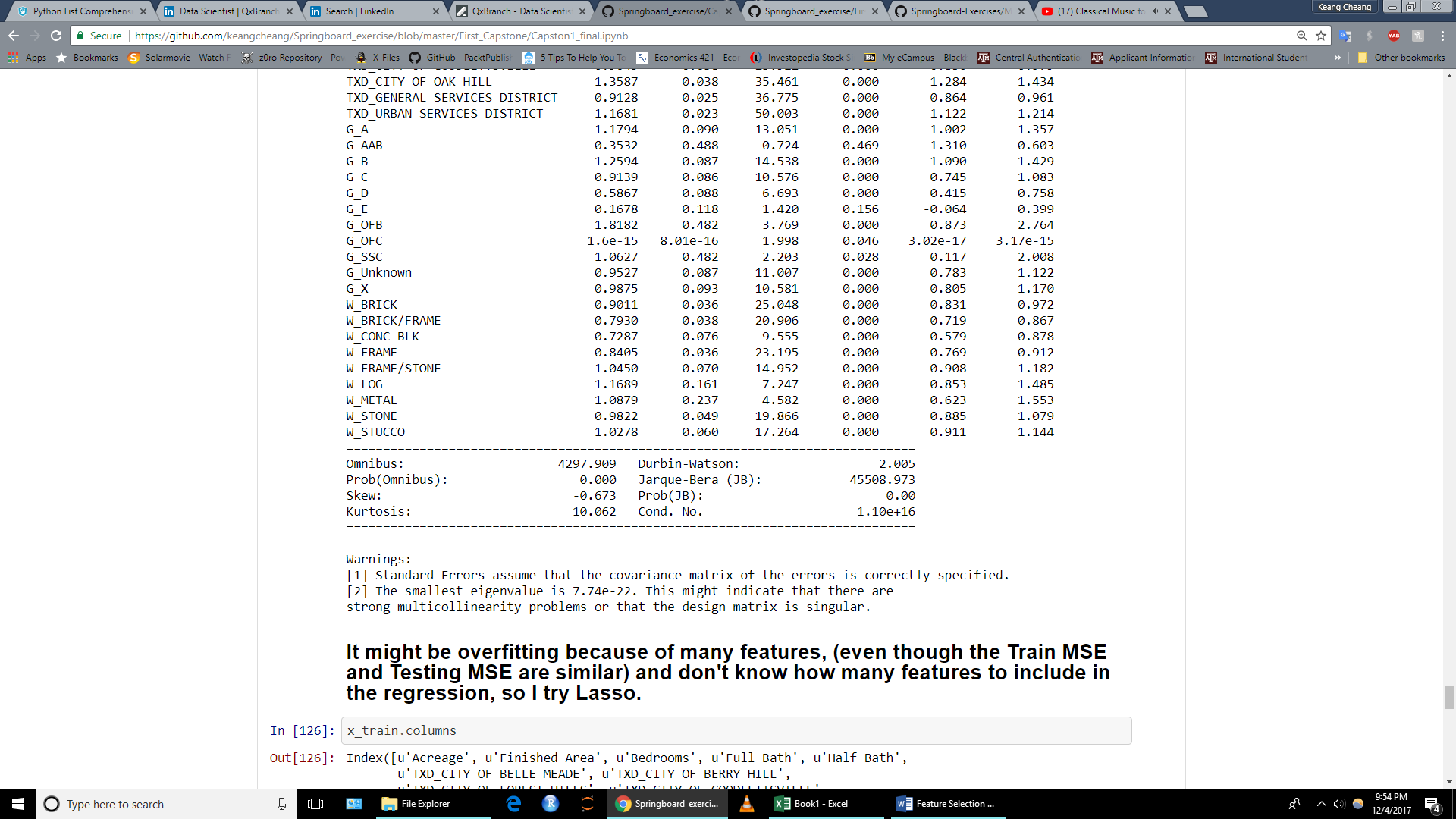
In order to improve our result, we would use regularization like L1 or Lasso to intervene. Lasso would perform both variable selection and regularization. After we ran Lasso, we got our result as following:

Testing Score: 0.34325587227

Training MSE: 0.306110557796

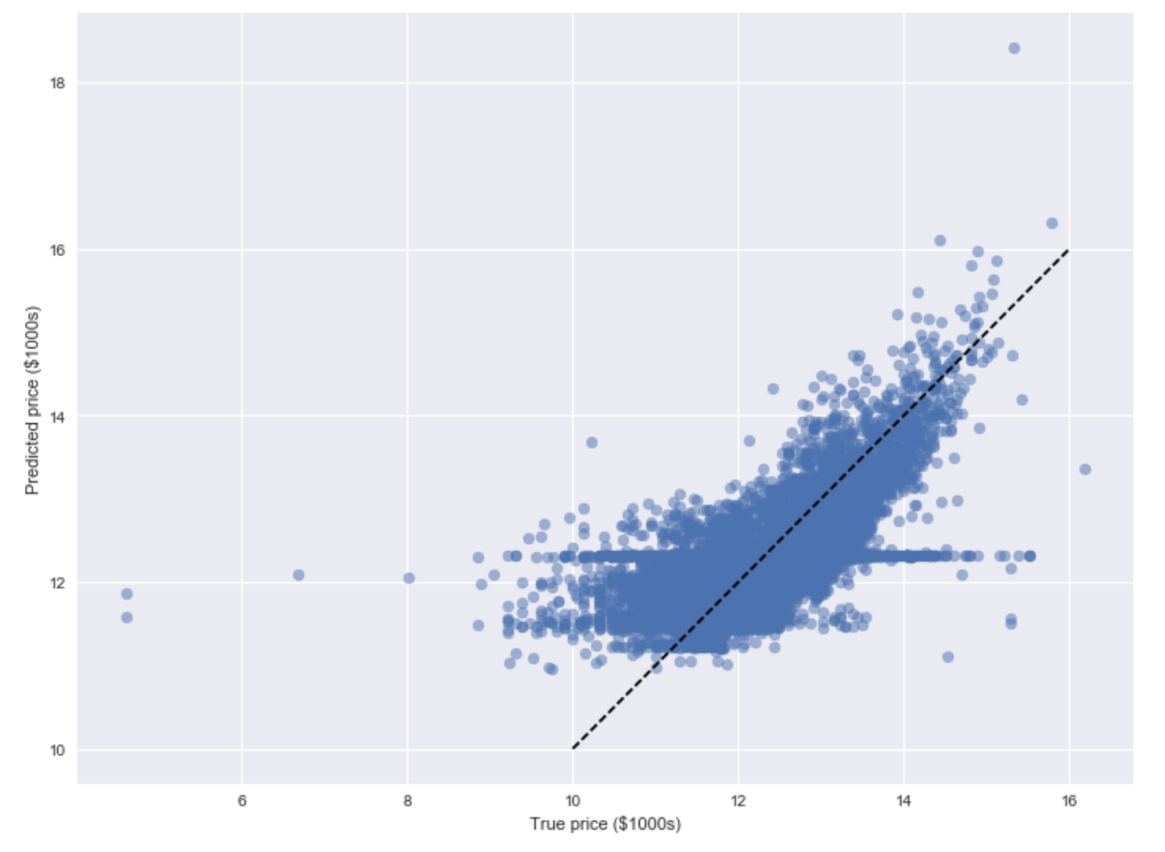
Testing MSE: 0.301153923281

These results were a little worse than the normal linear regression. The score was only 0.34 whereas previous score was 0.41, so the model was worse off than before. The MSE for both train and test were still similar to each other 0.306 and 0.301, respectively. However, even the score of Lasso did not improve, this model was still better in the sense that it reflected what coefficient were important and what not. The result of these terms were presented below:



**Selected features:**

(**False**, 'Acreage'), (**True**, 'Finished Area'), (**True**, 'Bedrooms'), (**True**, 'Full Bath'), (**False**, 'Half Bath'), (**False**, 'TXD\_CITY OF BELLE MEADE'), (**False**, 'TXD\_CITY OF BERRY HILL'), (**False**, 'TXD\_CITY OF FOREST HILLS'), (**False**, 'TXD\_CITY OF GOODLETTSVILLE'), (**False**, 'TXD\_CITY OF OAK HILL'), (False, 'TXD\_GENERAL SERVICES DISTRICT'), (**False**, 'TXD\_URBAN SERVICES DISTRICT'), (**False**, 'G\_A '), (**False**, 'G\_AAB '), (**False**, 'G\_B '), (**False**, 'G\_C '), (**False**, 'G\_D '), (**False**, 'G\_E '), (**False**, 'G\_OFB '), (**False**, 'G\_OFC '), (**False**, 'G\_SSC '), (**False**, 'G\_Unknown'), (True, 'G\_X '), (**False**, 'W\_BRICK'), (**False**, 'W\_BRICK/FRAME'), (False, 'W\_CONC BLK'), (**False**, 'W\_FRAME'), (**False**, 'W\_FRAME/STONE'), (**False**, 'W\_LOG'), (**False**, 'W\_METAL'), (**False**, 'W\_STONE'), (**False**, 'W\_STUCCO')

 The terms above showed us that the important features were bedroom, full bath, finished areas and G\_X (grade excellent). The scatter plot presented here displayed the relationship between the true price and the predicted price. The plot looked reasonable with the fitted line drew across the data points in the middle.

1. **Conclusion**

This project gave a new insight to Nashville real estate by offering what features important to the buyers and consequently drove up the price. By improving bedroom, having a full bath and bigger finished areas as well as in an excellent condition, the price of the house would be going up because the buyers willing to pay more. There were certainly ways to improve this model was to use cross validation and hyper parameter tuning on the dataset. This way would polish the results and accuracy score higher due to less over-fitting.

### Reference

* <https://www.trulia.com/real_estate/Nashville-Tennessee/market-trends>
* https://www.kaggle.com/tmthyjames/nashville-housing-date