**METCS521 Final Project: Predicting Housing Prices with Linear Regression**

*Authors: Tomáš Horníček, Wangzhen Li*

**Introduction**

For our project, we took a large dataset, consisting of nearly 3000 house entries, and tried to build a linear regression model, that would help us predict the housing prices in the coming years. Along with creating a linear regression model, we have also explored different variables, which affect our main dependent variable: Sale Price. We have created graph visualizations for these variables, in order to get a good idea of how different houses, with different features, will cost in the future. For our project we have created three python files; *download\_housingdataset.ipynb*, *graphs.ipynnb* and *Linear Regression Model.ipynb*.

For the user to run our program and explore how we obtained our data; they need to take the following steps:

1. Run the *download\_housingdataset.ipynb* to download the dataset.
2. Run the *graphs.ipynnb* to explore how we manipulated the dataset, to obtain the results we have and to see how certain factors affect the housing prices.
3. Run the *Linear Regression Model.ipynb*, to explore the linear regression model we built, to predict the housing prices

**Downloading the Dataset**

To analyze the housing prices, the user must first download the dataset. In order to do that they have to run, line by line, the *download\_housingdataset.ipynb* file in *Jupyter Notebook*. The script fetches the *URL* of the dataset page and writes the file to the project directory. This file will be then read, by the *graphs.ipynnb* and the *Linear Regression Model.ipynb* files, which then work with the dataset to obtain the information we need.

**Manipulating Data and Exploring Variables Affecting Housing Prices**

In the file graphs.*ipynnb* file, we have explored different variables that affect the main dependent variable: Sale Price. First we had to manipulate the dataset, making it readable and useful for the project. The *graphs.ipynnb* file first reads the downloaded dataset and then *“dummifies”* the dataset. *“Dummifing”* the dataset is crucial, as we needed the dataset to contain values, which we could then compare against each other. Because the dataset consisted of numbers and strings, we first mapped the dataset to contain only numbers. We have achieved that through giving the dataset key and value pairs, replacing the text with numbers.

Our next step was to find out what data is useful for our project. We found that out by correlating each category with all the categories and then looked at how closely each category is correlated to the Sale Price. We took the categories that have a high correlation value to be categories that have the absolute value of correlation 0.7 or higher and the categories that have low correlation to have the absolute value of correlation 0.3 or lower. We then have created a dataset without the dropped low correlation values. In our newly created dataset, we still faced a problem of having too many null values. We could’ve approached it in two way, either drop the houses with null values or fill in the null value with the one closes to it. Our group felt that dropping the null values would keep our findings as accurate as possible, therefore we decided to go with this option.

Now, that we created a base dataset, we used the SKLearn library to generate a linear regression model. { Talk about the linear regression model, train data, house built 1998 }.

From the base dataset, that we have edited to become useful for us, we then took out variables, with high correlation to the Sale Price and that we found interesting. We then created sub datasets with these variables, and our dependent variable; Sale Price, and visualized them using graphs.

Chart, scatter chart

Description automatically generated

*Figure 1: Garage Area(sqft) vs. Sale Price(USD)*

The first graph that we visualized was *Figure1*, that showed the correlation between garage area and sale price. Based upon this we can roughly estimate the sale prices of houses based on their garage size. We can do that by obtaining the gradient of the line of best fit and plotting the gradient, with the garage area to the equation of the line, to give us the sale price of the house. There are many other factors that affect the sale price( as we can see with the anomalies that have a higher price than others and garage area ~ 800 ), however with this graph we can see that the bigger the garage area, the bigger the sale price.

Chart, bar chart, histogram

Description automatically generated

*Figure 2: Houses with over 1400 sqft above ground area: Year Built vs Sale Price(USD)*

Chart, bar chart, histogram

Description automatically generated

*Figure 3: Houses with under 1400 sqft above ground area: Year Built vs Sale Price(USD)*

Next we wanted to compare the square footage of the above ground level of the houses and how that compared to the Sale price over the years. For this graph visualization, we split the dataset into two datasets; one containing the houses with under 1400 sqft above level and one containing houses with over 1400 sqft. Next we took the average price of the houses for the given year for the two categories and compared it with the sale price. This has resulted in graphs *Figure 2*: showing the houses with over 1400 sqft above ground area and how their sale price progressed throughout the years, and it also shows us *Figure 3:* showing the houses with under 1400 sqft above ground area and how their sale price progressed throughout the years. We then compared the two graphs and come to the obvious conclusion that the bigger the square footage above ground, the higher the sale price. However, comparing the two graphs shows us much more. From the two graphs we can read information such as how did the square footage affect the price of the houses for the given year. We can achieve that by comparing th percentage increase for the houses with over 1400 sqft and houses under 1400 sqft. For example, comparing the houses in the early 2000’s we can see that the difference between houses under/over 1400 sqft is close to 100%. However, in the 1950’s - 1960 ‘s, the difference between houses under/over 1400 sqft is between 10-25%. Based upon these findings we can conclude that the square footage played a bigger role in the sale price, in the early 2000’s than it did in the 1950’s - 1960 ‘s.

Chart, scatter chart

Description automatically generated

*Figure 4: Year Remodel added vs. Sale Price*

Next, we found it interesting to explore how the year a remodel was added to the house, affected the house price. The product of our findings can be seen in *Figure 4.* We can observe an interesting trend in *Figure 4*. We can see that, if the remodeling of the house was done before the year 1995, it didn’t really influence the sale price. If the remodeling was done after the year 1995, we can see that the newer the remodeling, the higher the sale price. This could be for a number of reasons; we think that it is because of technology advancements in the 1990’s throughout to 2010. House technology, such as fingerprint entries, security cameras, etc. was advancing at that time and logically the newer and more advanced technology, the higher the sale price.

Chart, bar chart

Description automatically generated

*Figure 5: Masonry Veneer Type(“None”: 0, “Brick Face”: 1, “Stone”: 2, “Cinder Block”: 3 ) vs. Sale Price(USD)*

Finally in *Figure 5*, is a visualization of each masonry veneer type against the sale price. For this graph visualization, we have taken the average of all the houses having the same veneer type and compared it to the sale price. This gives us an idea of which masonry veneer type is most costly and will give us a very rough idea in predicting the price of a house based on the veneer type.

**Linear Regression Model**

{Linear Regression Model -> add info from Linear Regression Model}

**Conclusion**