Internal Assessment

How is Machine Learning Used to Improve Systems over Time through Experience?

Number of Pages: 20

Appendix with data included after the reference page

**Introduction**:

Computers, coding, and technology are fundamental parts of our twenty-first century society. For those in Generation Z, technology has been integrated into our lives from the day we were born, through the toys we played with and the smartphones our family used. As someone who is engrossed in the world of networking and computer science, I carefully considered the ways that these topics will influence our future. Wanting to explore the specific things that will majorly affect our future lives, I decided to tackle the idea of machine learning. What is machine learning? According to IBM, it is “a branch of artificial intelligence (AI) focused on building applications that learn from data and improve their accuracy over time without being programmed to do so.” In essence, machine learning uses previously known and inputted data to accurately predict future data. Importantly, any machine learning program is constantly given new information and data points, which can then be used to make better predictions. Once I heard about machine learning, it immediately piqued my interest with its unique applications to our daily life. After researching it more, I became fascinated by how this process is used in complex deep learning techniques to make computers more intelligent and operate closer to how humans do, which is essential to developing autonomous robots and self-driving vehicles.

Through my investigation, I will be explaining the principles of machine learning, how it is used in our everyday lives, and how it works mathematically through a real-life example. Although complex examples of machine learning can only be accomplished through coding a computer algorithm, my example will involve the prediction of house prices, which is simple enough to be shown on paper.

**Investigation**

Machine learning is everywhere – it is used in our phones through their virtual phone assistants, in our emails through spam filtration, and in advanced self-driving cars, which we will use in our day-to-day lives within the next decade. Phone assistants like Siri use machine learning to improve their voice recognition and understand our sentences better, while email spam filtration is improved through machine learning by better recognizing suspicious emails. Most impressively, machine learning is applied in self-driving cars to improve their detection of road hazards such as people, potholes, signs, and more. High-tech companies like Apple and Tesla use machine learning so their products can learn from their mistakes, and the more data that their products collect, the better their algorithms become. Not only is machine learning already important to our lives right now, but as technology advances, it will become much more ubiquitous and essential to society as a whole in the future.

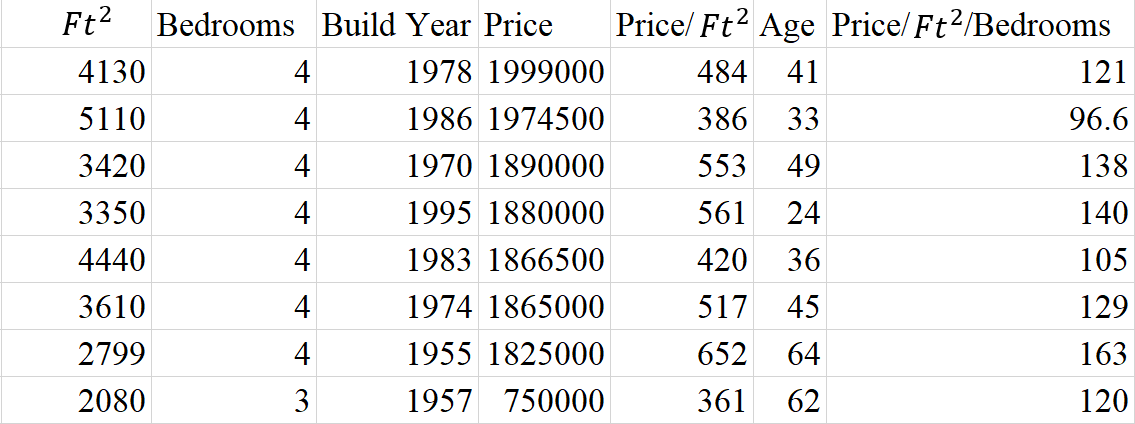
Personally, I only recently discovered the prevalence and importance of machine learning in our lives. Instead of machine learning, my interest laid in the idea of artificial intelligence (AI). While I had heard the term “machine learning” before, I did not know about its relevance to artificial intelligence until I began to work on this paper. As it turns out, machine learning is a subset of artificial intelligence and is essential in pushing AI applications forward. Once I understood this relationship, I became eager to learn more about how machine learning works, how it is used, and how to actually engage in it myself. In fact, I even attended an online machine learning camp at DigiPen Institute of Technology to discover more during the summer.

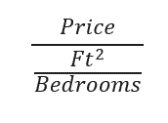
That brings us to an important question: how *does* machine learning actually work? This process is based on the mathematical principles of statistics, probability, calculus, and linear algebra. The core principle underlying machine learning is the development of a mathematical algorithm that keeps evolving as the amount of data to analyze increases. The process where the algorithm grows “smarter” over time is known as training.

To explain the core principles of machine learning, I decided to consider a case study of real estate prices. These prices rely on a multitude of different parameters, such as age, number of rooms, location, condition, square footage, and more. Thus, it is difficult to find accurate pricing predictions based on standard mathematical tools, because the distribution of prices against any given parameter looks very random. As there are many parameters, I have selected a few to demonstrate how to develop an algorithm that can be used in the machine learning process. Instead of using a whole range of parameters that would result in multi-dimensional complex algorithms, I have simplified the task to only use two. I took a few of the most important parameters – price, size, age, and number of bedrooms – and narrowed them down into two simpler variables. For the x (independent) variable, I used the age of the house, and for the y (dependent) variable, I combined price, size, and bedrooms into one parameter by dividing price per square feet per number of bedrooms.

The current objective is to find the line of regression, or best fit line; the best fit line is the line that goes through a set of data points which most closely resembles the pattern of the scatterplot. This line of best fit will provide a mathematical relationship between x and y, allowing us to hopefully predict the full price of a house just with the age provided. It is important to note that a linear regression might not be the appropriate relationship between these two variables; other approaches, like a nonlinear regression, may be more accurate in representing the relationship between the two variables. Nonetheless, we will continue with the assumption that linear regression provides sufficient approximation of the actual prediction in order to demonstrate how machine learning works in this case study.

To begin, here is a small data sample of 8 houses sold in my county from the Northwest Multiple Listing Service (Northwest MLS) database, which is a database that provides information on property sales and history:





Based on this set of data, we can graph it onto a scatter plot:

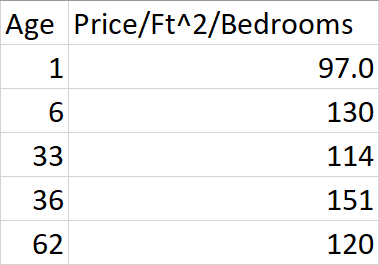
With only 8 data points, this graph does not provide a good representation of the housing market, because it includes only a small portion of houses sold in the local area. In order for the data set to be more representative, one should rely on a larger pool of data. Here is another scatter plot with 42 data points:

This newer graph looks different since it includes a larger data population. In machine learning, a program uses the data on the scatter plot to calculate the line of best fit; as the program is provided with more data, the line changes and becomes more representative of the predicted market prices. In order to find the best line of regression, we will be using the concept of the error function:

The error is the distance between the predicted value and the actual value; the smaller the error is, the more accurate the prediction is. Within the sigma equation, the point y1 (an inputted data point) is being subtracted from the line equation mxi + b (the calculated line of fit) in order to find the smallest distance between the point and the line. Only when each data point is equally spaced from the line will you get the best line of fit. Using the error equation and another equation:

We can find out the values of m and b through a system of equations. Firstly, we need to make sure that the partial derivative of the equation for b and m equals to 0 as to minimize the error.

Solving this by hand with a large sample would be very long and repetitive, so as a demonstration, I will be using this simple data table with only 5 elements to show how the math would be calculated:





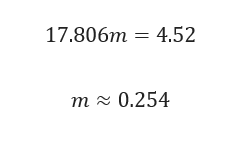


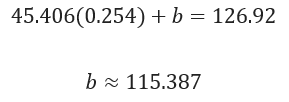
Now that we have the equations and our variables, I can plug the numbers into them.

Which simplifies to:

Now, this large equation can be simplified into something much easier to understand:

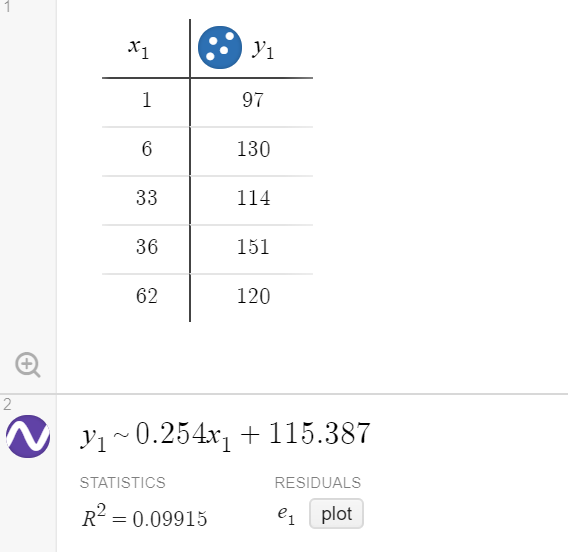
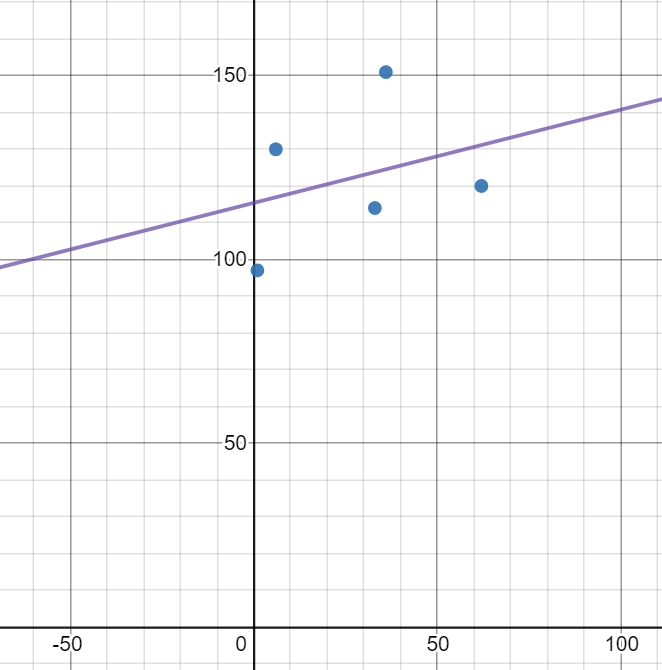
Now we can plug our values into the other equation, which is the summary of *mx+b*. Once we put in the same data, we can use a system of equations to solve for m and b, getting our line.

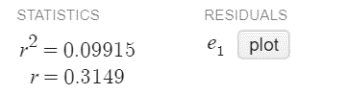






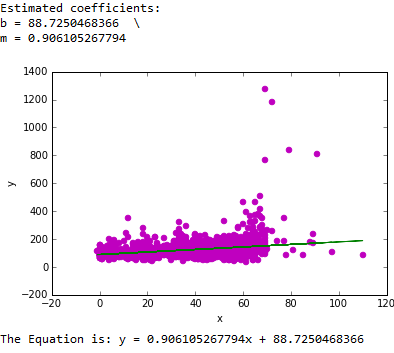
For this small set of data, we have found the equation to be about . The equation represents the line describing the relationship between age, or m, and the price per square feet per number of bedrooms, or y. We can check our work by inputting the data and graph into Desmos:





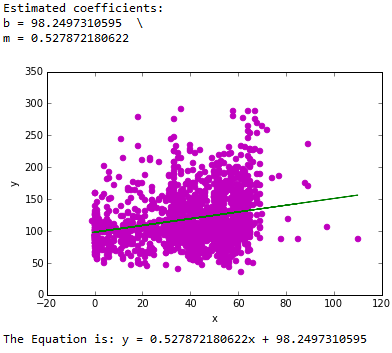
As we can see from this data, the work ended up bringing us the best fit line through the points. The R-score of the graph is shown as 0.315, which means there is some correlation with the line to the points, but it is not a strong correlation as we only have five data points to work with. An R-value closer to 1 would mean a near perfect correlation, so we can see that this line does not completely represent the relationship between the data points. The problem with the previously-described method is that as there is more data, the work becomes exponentially more tedious. In the real world, the difficulty of the calculations makes it near-impossible to accomplish by hand with much larger datasets, especially since actual uses of machine learning can use hundreds of thousands or millions of datasets. For this reason, we use computer programs to accomplish this, instead of calculating the millions of data points that would be used by hand. To show how this works, I used a Python-based, open-source machine learning algorithm that I incorporated into my own Python program. As someone who has been learning Python for many years, it was amazing to learn that not only were many machine learning algorithms made in Python, but also that I too could use one myself. Using online resources, I studied how these algorithms work in Python, so I could then incorporate it into my own program for this exploration.

In order to make the graph as accurate as possible, I obtained actual local MLS data of house sales in recent years. I narrowed down the data to sales that were between $500,000 and $2,000,000 and increased the sampling data used from the previous example of 42 data sets to more than 1500 different house sales data sets. This resulted in the following graph:



Line of Best Fit through Datapoints with Outliers



The main problem in this graph, though, is the outliers. For an accurate prediction, you need to remove the outliers, because they represent the exceptions and can make the line of best fit less accurate and less representative of the rest of the data. The outliers are found by taking the points that fall out of the confidence interval, which gives the amount of uncertainty within the sample data. In this case, the confidence interval was about 95%, so I removed the 5% of points that were outside the interval. After removing all the outliers, this gives the final graph with the x-axis being the age and the y-axis being the sell price per *Ft2* per number of bedrooms:

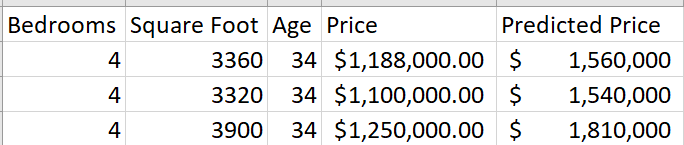
Line of Best Fit through Datapoints





With this data, we could use the age of a house, number of bedrooms, and size to predict the price of a house. For example, I’ll use a house that is 34 years old, has four bedrooms, and is 3500 square feet to find the predicted price of the house. Using the age of the house and the predicted equation:

We get a price per feet squared per bedroom of about $116. I can now multiply this by the number of bedrooms and total square footage to get the overall price, which is approximately $1,620,000. From the large dataset, there are exactly three houses that have the same age and number of bedrooms and have similar square footage. By comparing the actual prices of these similar houses with their predicted prices based on the calculated equation, the accuracy of the predictions can be evaluated. This table shows the results of this comparison:



The table shows that there is a significant difference between  the predicted prices versus the real prices, indicating a lack of accuracy in our line of fit. The biggest similarity between the real prices and predicted prices is that both the predicted and real prices increase with square footage.

Based on this, we can see that our line of best fit is not very accurate to the real, local data. In the real world, the price of the house is dependent on many other factors besides just size, bedrooms, and age, such as location, number of bathrooms and bathtubs, whether there’s a basement, and more. Once you begin adding these amounts of parameters, the graph becomes more and more complicated, and a linear regression would not be accurate enough; it would need a more complex line of best fit.

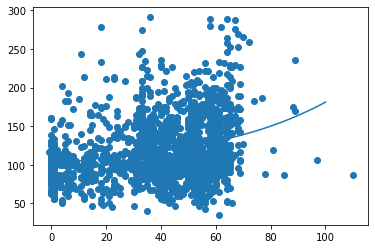
The next level of prediction would be to move from a linear regression to a polynomial regression, which would help make the line fit more accurately with the data set. Because this would make the math a much longer, more manual process, coding algorithms can be used to efficiently solve the problem, like I did with my Python program.

Polynomial regression is used when the correlation between multiple variables is not linear. Based on our sample data, a simple linear regression does not accurately reflect the relationship between age and price, so we can use a polynomial regression instead to decrease the error and get a better prediction. The equation of a polynomial regression includes exponents within it to make a better prediction by reducing the error. To solve this, we would need to find the coefficients to form an equation relating house age with house price.

Solving an equation like takes a lot more work than solving a linear equation. Firstly, it is necessary to set up a new error function for the polynomial.

The math for solving a polynomial regression equation is similar to the math applied for linear regression, but it additionally requires the need to solve for more coefficients. To solve this in a more efficient way, there is a matrix equation available which can be solved to find the values of each coefficient for the equation. This matrix equation for any degree of polynomial is the following:

The process of solving this equation manually is also very complex, much more than solving a linear regression, so we also will be using a Python program to find a polynomial of best fit for our data. Using a new program to solve for a third-degree polynomial regression, the following graph and equation is produced, with an R squared value of 0.0640:



Polynomial Line of Best Fit through Datapoints



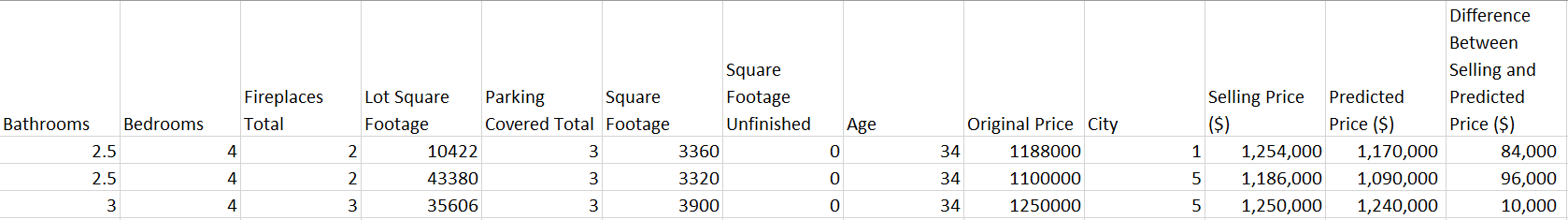
Going further, using a 36th degree polynomial regression instead of a 3rd degree polynomial, a higher R-squared score of about 0.124 is produced, but the resultant graph is too large to be displayed in this paper.

         The regression approaches discussed, linear and polynomial regression, have been based on using a single independent variable for predictions. An alternative approach would be extending this simple linear regression to a multiple linear regression that predicts a single value based on more than one independent variable. By using only two variables with the previous regressions, we limit the extent to which the analysis can be accurate. However, by adding even more input variables, it is possible to improve the prediction while reducing the error. An equation for a multiple regression would be as follows:

In this equation, theta represents the coefficient for each variable, and each xn would correspond to a different independent variable. By adding more than one independent variable, the resultant graph becomes multidimensional, which cannot be represented on paper. Therefore, the best way to solve such a complex equation would be through a Python program. For this program, the dependent variable is just the final selling price, while the independent variables included: bathrooms, bedrooms, total amount of fireplaces, the square footage of the lot, the total parking covered, the square footage of the house, the amount of unfinished square footage, the age, the original price, and the location. These variables correspond respectively in the regression as x1 to x10. Accounting for all of these variables, the final equation can be written as:

+ 69200

This equation, unlike the equations calculated through linear and polynomial regression, had a very high r-squared score of 0.957; this indicates it has higher accuracy in predicting housing prices, making it a more useful and applicable tool. If this equation is applied to the three houses previously considered to predict their selling price, a much more accurate prediction of the selling price (compared to the previous simple linear regression) emerges. Where our previous predictions based on linear regression were off by hundreds of thousands of dollars, these predictions based on multiple regression are only off by tens of thousands of dollars.



While multiple linear regression may appear to be the best approach to get the highest accuracy, it is not always necessary to use multiple regression to calculate equations with high correlation. Depending on the relationship between data points in any given dataset, any of the three approaches discussed could be used. For data sets which appear to have a linear relationship, linear regression can be used accurately enough, but it becomes too inaccurate for non-linear relationships to make high quality predictions. Polynomial regression, like linear regression, only uses a single independent variable to make a prediction, but it is the more accurate approach for nonlinear relationships between the independent and dependent variables. Finally, with multiple linear regression, the ability to have several independent variables enables more accurate predictions for complex, real world situations, but the multiple inputs make it unable to be easily visualized with a scatterplot, unlike linear and polynomial regressions.

While much of the previous discussion has centered on analyzing different regression approaches, regression analysis is not the same as machine learning. Instead, regression analysis is a tool used heavily in machine learning. What differentiates machine learning from traditional regression analysis is the constant addition of new data to the data set, which a machine learning algorithm uses to evolve the regression equation over time, thus making its predictions from that equation more accurate over time. Through my investigation, I specifically demonstrated how an increasing amount of data in simple linear regression allows for the development of better lines of fit. In addition, I demonstrated how when one increases the amount of input variables like in multiple regression,  the prediction will become more accurate, especially if the new input is closely related to the dependent variable. Interestingly, even if the new variable is not closely related to the dependent variable, the prediction made by the regression will not be negatively affected. In real machine learning algorithms, the multiple regression approach is the more common approach used, reflecting the reality that most real-world data sets are multidimensional and constantly have data added to it. The consistent addition of data to complex machine learning algorithms is how voice recognition in smartphones, like Siri, becomes better with more usage, and how self-driving cars learn from their mistakes to become better at driving. Prior to this investigation, I had always passively observed the application of machine learning in our daily lives, but only now do I fully understand the complicated mathematical mechanisms behind how it works and how it is used.

**Conclusion**

Overall, my goal with this assessment was to learn how machine learning works and to demonstrate its basics with a real-life example. Using a case study of real-life house prices, I attempted to find an equation based on linear regression. I first used scatter plots to show how increasing amounts of data changes the scatter plot and trend shown in the plot. Then, I took a very small sample of data and used a sigma equation to solve for the slope and y-intercept of a linear equation, showing how the line of best fit was calculated. After this, I used a much larger sample of data to create a best-fit line through a computer algorithm, which was graphed on the scatterplot. This best-fit line allowed for the prediction of house prices using certain parameters.  Subsequently, polynomial, and multiple linear regression were used to find equations that further improved these predictions using additional parameters. Lastly, these regression approaches were connected with the broader principles of machine learning to clearly demonstrate how it works.

At the beginning of this assessment, I knew very little about machine learning, but I’ve now become invested in understanding its real-world future implications. Through my research, I learned that the correlation of the prices and the input data through linear and polynomial regression did not provide conclusive results; even when I added more data, the resultant equations still were not accurate enough. In order to create a better predictive algorithm for real estate sales prices, I needed to make a more complex prediction through multiple linear regression. By using multiple linear regression through code I compiled, I found a much more accurate equation to predict housing prices in the real world. Having considered these different approaches, I am now more inclined to researching statistical regression further and applying other approaches, like binary and multivariate regression, to get even more accurate predictions in other interesting scenarios.

In coding for this investigation, I realized that the true advantage of machine learning is that people can easily evolve their algorithms by using computer tools to create optimal results. In machine learning, this is known as “training”; it is when you are continuously adding data and “training” your algorithm to provide correct results. After this training phase, the algorithm goes to an inference state, where the trained machine learning algorithm is used to make real predictions. The implications of this training mechanism reach widely in our society. Just imagine using facial recognition technology in your doctor’s office to diagnose rare diseases, or your doctors using algorithms to catch a rare cancer early. This isn’t something that you just have to imagine; this is our emerging reality, all made possible by machine learning algorithms. It’s incredible what this technology will mean for the future of human society. Hopefully, I showed how advanced mathematics through machine learning can create prediction algorithms that will improve and impact our everyday lives.

References

Grant, Peter. “Understanding Multiple Regression.” *Medium*, Towards Data Science, 21 Jan. 2021, towardsdatascience.com/understanding-multiple-regression-249b16bde83e#:~:text=Multiple%20regression%20is%20an%20extension,of%20a%20different%20physical%20parameter.

Hurwitz, Judith, and Daniel Kirsch. *Machine Learning for Dummies*. IBM, *Machine Learning for Dummies*, [www.ibm.com/downloads/cas/GB8ZMQZ3](about:blank).

IBM Cloud Education. “What Is Machine Learning?” *IBM*, 15 July 2020, www.ibm.com/cloud/learn/machine-learning.

“Linear Regression (Python Implementation).” *GeeksforGeeks*, 29 Nov. 2018, [www.geeksforgeeks.org/linear-regression-python-implementation/](about:blank).

“Machine Learning - Multiple Regression.” *Python Machine Learning Multiple Regression*, www.w3schools.com/python/python\_ml\_multiple\_regression.asp.

“Machine Learning.” *Wikipedia*, Wikimedia Foundation, en.wikipedia.org/wiki/Machine\_learning.

“ML: Multiple Linear Regression Using Python.” *GeeksforGeeks*, 23 Feb. 2020, www.geeksforgeeks.org/ml-multiple-linear-regression-using-python/.

“MLS Database.” *Northwest MLS*, [www.nwmls.com/](about:blank).

Muthu. “Maths behind Polynomial Regression.” *Muthukrishnan*, 17 June 2018, muthu.co/maths-behind-polynomial-regression/.

Pant, Ayush. “Introduction to Linear Regression and Polynomial Regression.” *Medium*, Towards Data Science, 16 Jan. 2019, towardsdatascience.com/introduction-to-linear-regression-and-polynomial-regression-f8adc96f31cb.

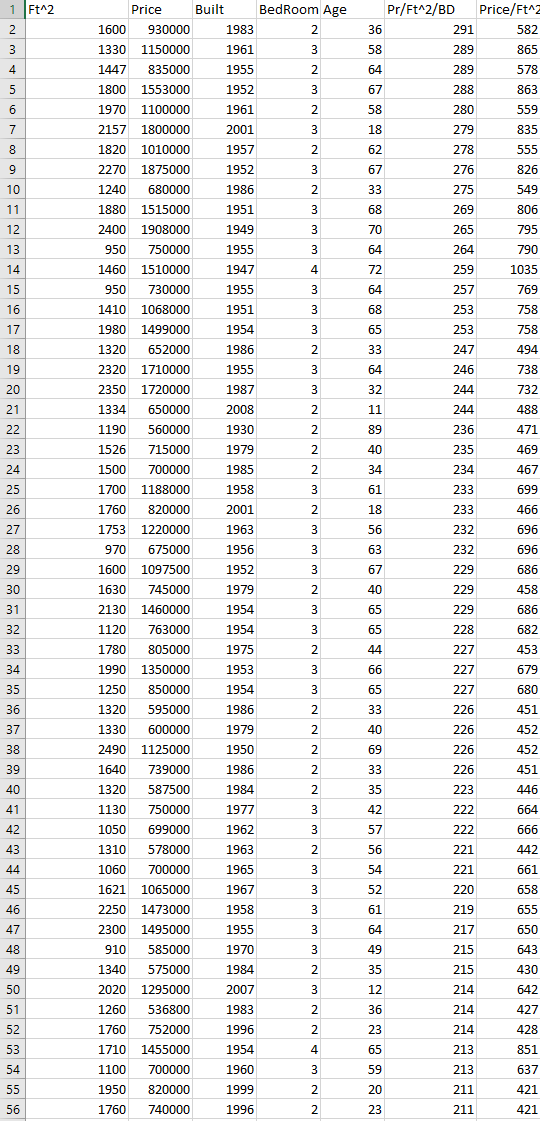
“Partial Derivative.” *Wikipedia*, Wikimedia Foundation, en.wikipedia.org/wiki/Partial\_derivative.

Que, Andrew. “Polynomial Regression.” *Mathematics of Polynomial Regression*, polynomialregression.drque.net/math.html.

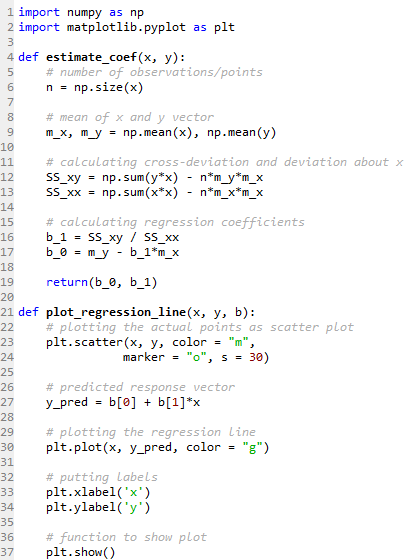
Shukla, Pratik. “Complete Guide On Linear Regression Vs. Polynomial Regression With Implementation In Python.” *Medium*, DataDrivenInvestor, 4 June 2020, medium.datadriveninvestor.com/complete-guide-on-linear-regression-vs-polynomial-regression-with-implementation-in-python-964c64c28aa8.

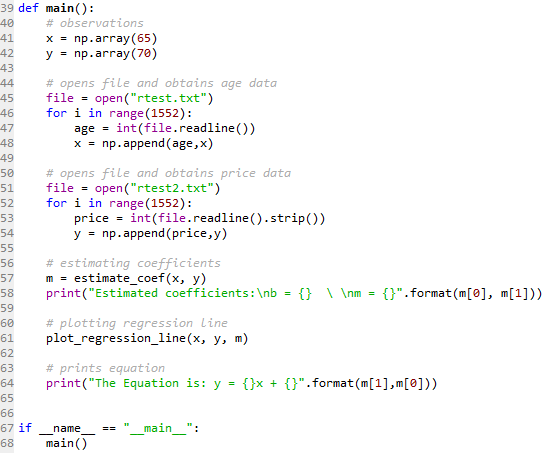
“Training versus Inference.” *Paul DeBeasi*, 14 Feb. 2019, blogs.gartner.com/paul-debeasi/2019/02/14/training-versus-inference/

Appendix



This was a portion of the data I collected on local real-estate from MLS for use with linear regression. Along with this list, I used an additional 1497 lines of data on houses from our area.

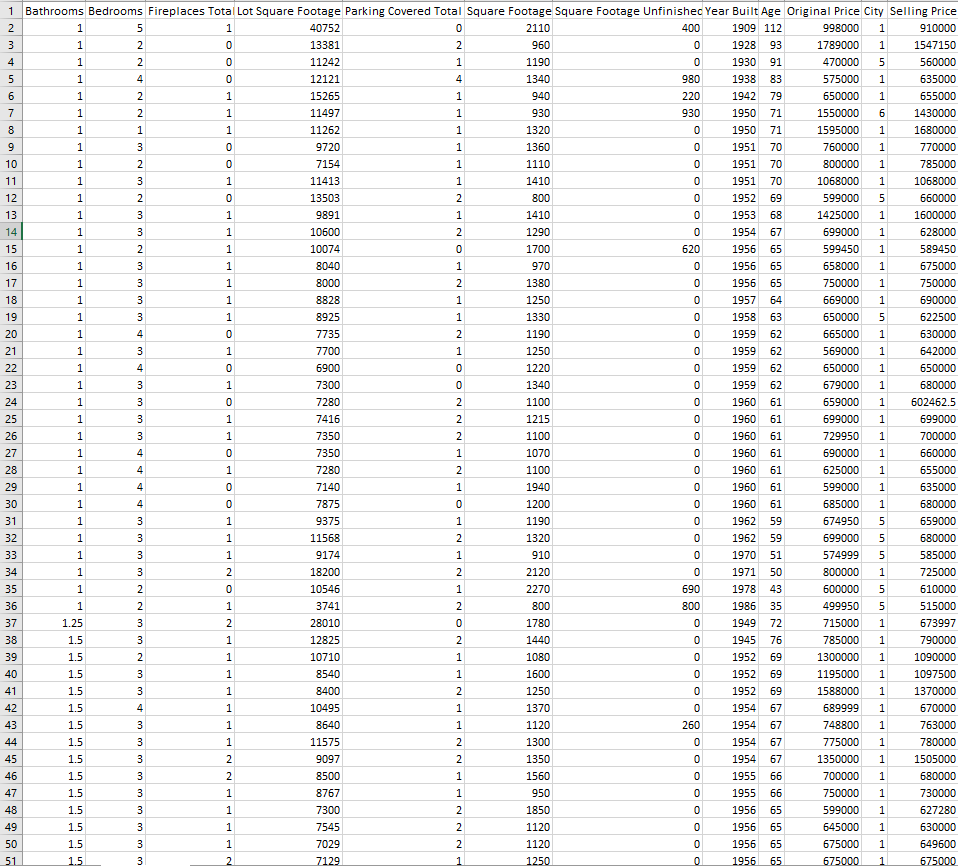




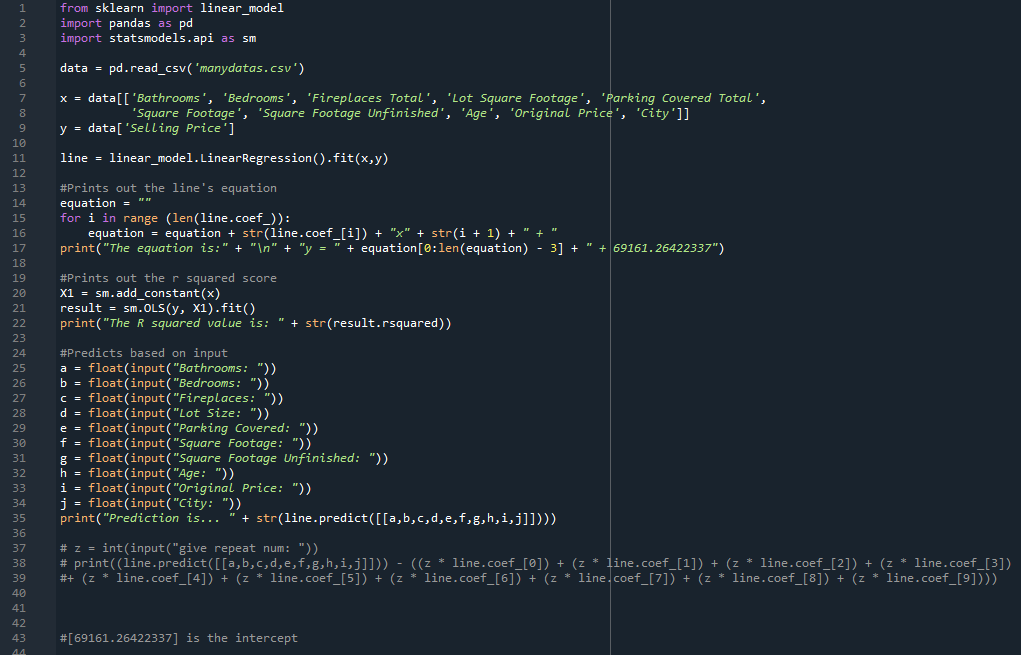
This was the code I used with my large list of data to create a graph and an equation for linear regression.



This was the code I used with my large list of data to create a graph and an equation for polynomial regression.



This was a portion of the data I collected on local real-estate from MLS for use with multiple linear regression. Along with this list, I used an additional 1478 lines of data on houses from our area



This was the code I used with my large list of data to create a prediction equation using multiple linear regression.