HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 3 (2024)

Making Housing Predictions Using ML Without Live Market Data

Justin Chen[1] and Soyoun Choi[#]

[1]Monta Vista High School, USA
[#]Advisor

ABSTRACT

This paper explores the importance of machine learning and the benefits machine learning can bring to the housing industry. The housing industry makes up an average of more than 15% of the US's yearly GDP and plays a significant role in the US economy. Additionally, the housing industry, specifically house prices, determines the environment each person grows up in by filtering out those who are unable to afford the costs of living in any given area. Having accurate house price information on hand can save many hours spent looking for affordable housing. After a brief introduction of Zillow, a house listing website that also contains a feature that provides highly accurate predictions of its users' house prices, the central question, "How high a degree of accuracy can house prices be predicted without information regarding the state of the current economy or house market?" is explored. While Zillow uses data such as market trends and past selling prices, only information regarding the physical traits of the house, such as its geographical location and its room count, are considered within this paper (Note that the scaling factor for many variables is messed up, meaning that despite numbers that don't intuitively seem correct, the relationship between each pair of data points remains unaffected). After visually identifying correlations between house value and the other variables with simple scatter plots, the relationships between different variables were further explored with different regression models, and a neural network was made based off of the previous exploration.

Introduction

Why Do House Prices Matter?

The housing market is one of the largest markets in the world, making up more than 15% of the US's GDP (gross domestic product, or the net value of goods and services provided within a single year) over the past half a decade (Han). Because it makes up such a large portion of the world's net spending each year, fluctuations within the housing market often directly correlate to the amount of economic growth experienced by a given country. For example, rising house prices can lead to increased spending on construction projects to take advantage of the heightened house prices and boost economic growth, while decreasing house prices would cause the opposite effect, decreasing spending on construction projects and slowing down economic growth. With a wide range of prices dependent on many factors, such as the neighborhood, the number of rooms, the geographical location of the house, and much more, house prices are incredibly difficult to predict with reliable accuracy. Small differences in any of the factors that contribute toward determining the worth of a house can lead to drastic differences between the actual selling price. This variation could lead to a nightmare for lower or middle-class families with limited funds that may be looking to move and could result in lots of wasted energy and time spent looking for an affordable house. This paper will explore how machine learning and artificial intelligence can benefit the housing industry as a whole, as well as to how high a degree of accuracy house pricing can be predicted given only the physical traits of the house.

Why Choose a Neural Network, Rather Than Having a Human Appraisal?

Appraisals are processes after which a human appraiser provides an evaluation of a property's estimated value based on its features and the housing market at a given time, and are often performed during the house purchasing process (Levinson). However, they are meant more for the buyer than the seller, as appraisers are often brought in by the party loaning to the buyer to ensure that the property is worth equal to or more than the amount being loaned. This way, after the buyer takes the loan using the house as collateral (meaning that in the scenario the buyer cannot pay the loan, the loaner has the right to confiscate the house), even if the buyer fails to pay off the loan, the loaning party will be able to seize and sell the house for the same amount or more than the amount they lost. Additionally, since these appraisals are generally intended for the latter stages of purchasing a house when a lender needs to determine whether the asking price of the house is reasonable, they don't help the owner with determining the house's listing price which needs to be identified before the house is actually put on the market. This means that people looking to sell their homes must also hire an appraiser of their own to determine a reasonable listing price, which could cost over $300, an amount that would be outside of many low-income families' budgets. The process of appraising a house also takes anywhere between 1 to 4 weeks, which could be detrimental to families in desperate financial situations. For example, if a serious car accident occurred to someone and their family did not have enough money on hand to pay for hospital bills, the multiple-week-long delay caused by the manual home appraisal process could prevent the person involved in the accident from being able to receive the care they need and could lead to their death.

A neural network, on the other hand, is a machine learning model that processes information in a way similar to the brain. It uses complex algorithms and large quantities of training data to determine "node weights" for each factor that it takes as input — essentially determining how strong the correlation is between a given variable and the end result. In the context of this paper, an example would be between a house's price and the number of rooms it has. Afterward, it would be able to produce, given information about the house, the same (or similar enough to make a reasonable estimate) information as an appraiser, except for free and within merely a few seconds. Although accessing the neural network and obtaining information from it may require access to a computer, more than 99.5% of libraries in the US offer access to computers. Additionally, even if there are no public computers available nearby, the cost of purchasing a low to medium-end laptop capable of running the neural network is much lower than that of hiring a professional appraiser.

What Is Zillow, and How Does This Project Relate to It?

Zillow is a popular website that provides house listing services for free to the general public for both rent and purchase. It's able to provide its services without charging users by instead charging property management companies, who advertise their listings on Zillow's website. Additionally, Zillow has its own neural network, the Zestimate, which can predict a fairly accurate house value for any users looking to sell/rent out their homes. Their neural network makes its predictions with lots of information such as market trends, on-market data such as selling values for other similar homes in the area, off-market data such as prior sales, house-specific details such as square footage, and much more. The neural network that will be discussed in this paper, on the other hand, will try to produce similar results to the neural network from Zillow without taking market trends and other economic factors into consideration. The only data that is loaded into the neural network explored in this paper is information regarding physical house traits such as its location and the number of rooms it contains.

Exploratory Data Analysis

Before actually building a neural network, it is always important to first identify potentially useful relationships between different variables. As this dataset had multiple thousands of data points, manually finding their relationships would be near impossible, or at least highly illogical. Instead, graphs of the data points were created to represent these values. Since the focus of this paper is the prediction of house prices, the main graphs used during this period were scatterplots of house prices, plotted against the other variables. The purpose of these graphs was to aid in visually identifying obvious correlations between pairs of variables, rather than to give exact information regarding what kind of correlation each pair had or the strength of its correlation. Additionally, the graphs were useful in identifying which relationships should be further explored or had potential.

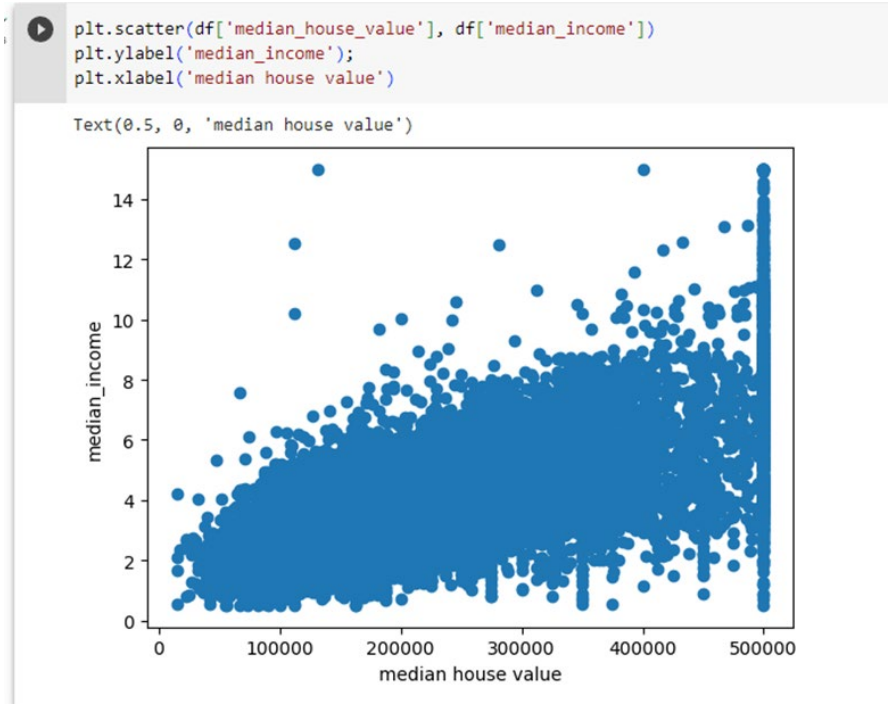

Figure 1. A graph of the correlation between house value and income

The graph above is an example of one of the scatterplots. The variable plotted on the x-axis is the house value, while the variable plotted on the y-axis is the homeowner's income. In this case, there is a very noticeable seemingly linear relationship between the pair of variables.

```
plt.scatter(df['median_house_value'], df['population'])
plt.ylabel('population');
plt.xlabel('median house value')
```

```
Text(0.5, 0, 'median house value')
```

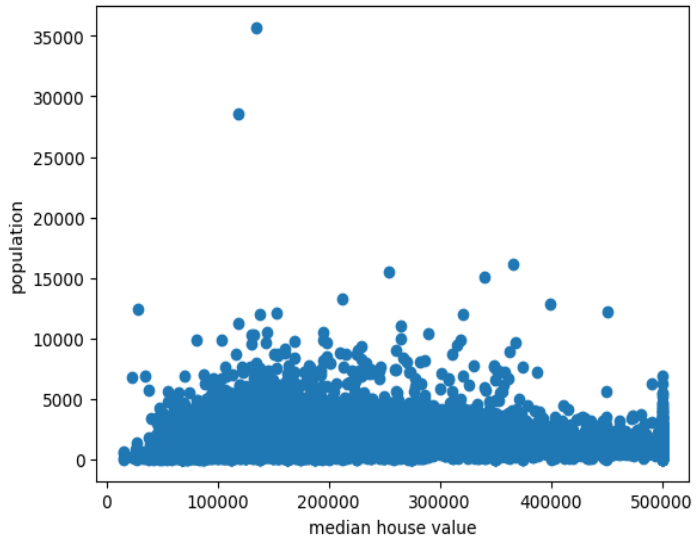

Figure 2. A graph of the correlation between house value and population

This is a different scatterplot that was created during the exploratory data analysis phase, where house value was plotted on the x-axis against the area's population on the y-axis. However, unlike the previous graph, which showed an obvious relationship between the two axes, this graph shows a much less obvious correlation, with only hints at a potentially cubic relationship described by the small bump shown between the values of 100000 and 200000. Due to the large amounts of data clumped together near the bottom of the graph, it is impossible to visually predict the relationship between the two variables. However, as the pair appears to potentially have some correlation, their relationship will later be explored using regression models.

Methodology

After cleaning up some unnecessary parts of the dataset and identifying areas of potential interest within the dataset, the process of interpreting the exact kinds of relationships between the different variables began. First, the original dataset of over 15,000 data points was split into two separate sets with a ratio of 7:3, where the 7 represented the ratio of data points that went into the training set and the 3 represented the ratio of data points that went into the set that would later be used for testing purposes. After this, the exact correlations between identified variables of interest were mapped out by comparing the different pairs to identify their relationships. Regression graphs were utilized for the comparisons, with the house value always on the x-axis, and a graph was created with each of the other remaining variables on the y-axis. Each graph contained regressions of the following kinds: linear, quadratic, cubic, quartic, and quintic. Additionally, the mean squared error, or the strength of the correlation between the x and the y values (with 0 being that the prediction is perfect), was calculated for each of the types of regressions to determine which form of regression produced optimal predictions within each graph. The optimal prediction model within each graph was always the one with the smallest amount of error or the one that was closest to a perfect prediction.

Example code for one of the regressions:

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error as mse
population = X_train['population'].values.reshape(-1, 1)
house_value = X_train['median_house_value'].values.reshape(-1, 1)
```

Purpose of code above: to create variables population and house_value that are shaped in a way such that different regressions can be performed.

```
X_train['house_value^2'] = (X_train['median_house_value']**2).values.reshape(-1, 1)
X_train['house_value^3'] = (X_train['median_house_value']**3).values.reshape(-1, 1)
X_train['house_value^4'] = (X_train['median_house_value']**4).values.reshape(-1, 1)
X_train['house_value^5'] = (X_train['median_house_value']**5).values.reshape(-1, 1)
```

Purpose of the code above: to create new columns in the dataset with already existing data raised to all necessary powers to simplify and factor out the code for the different regressions.

```
#population linear
linear = LinearRegression()
linear.fit(house_value, population)
lineMSE = mse(linear.predict(house_value), population)

#population quadratic
quad = LinearRegression()
quad.fit(houseval_quadratic, population)
quadMSE = mse(quad.predict(houseval_quadratic), population)

#population cubic
cube = LinearRegression()
cube.fit(houseval_cubic, population)
cubeMSE = mse(cube.predict(houseval_cubic), population)

#population quartic
quart = LinearRegression()
quart.fit(houseval_quartic, population)
quartMSE = mse(quart.predict(houseval_quartic), population)

#population quintic
quint = LinearRegression()
quint.fit(houseval_quintic, population)
quintMSE = mse(quint.predict(houseval_quintic), population)

plt.scatter(house_value, population)
plt.plot(linear.predict(house_value), c = 'r', marker = '.', label = 'linear')
plt.scatter(house_value, quad.predict(houseval_quadratic), c = 'black', marker = '.', label = 'quadratic')
plt.scatter(house_value, cube.predict(houseval_cubic), c = 'g', marker = '.', label = 'cubic')
plt.scatter(house_value, quart.predict(houseval_quartic), c = 'w',marker = '.', label = 'quart')
plt.scatter(house_value, quint.predict(houseval_quintic), c = 'purple', marker = '.', label = 'quint')
print('lineMSE = ', end = '')
print(lineMSE)
print('quadMSE = ', end = '')
print(quadMSE)
print('cubicMSE = ', end = '')
print(cubeMSE)
print('quartMSE = ', end = '')
print(quartMSE)
print('quintMSE = ', end = '')
print(quintMSE)
```

Purpose of code above: to create 5 different kinds of regressions for a given graph, house value vs. population in this case

Regression Graphs

Y-axis: Households

```
lineMSE = 148248.56061723025
quadMSE = 146092.85807662248
cubicMSE = 145935.49625669234
quartMSE = 146124.55472296965
quintMSE = 146507.44885785697
```

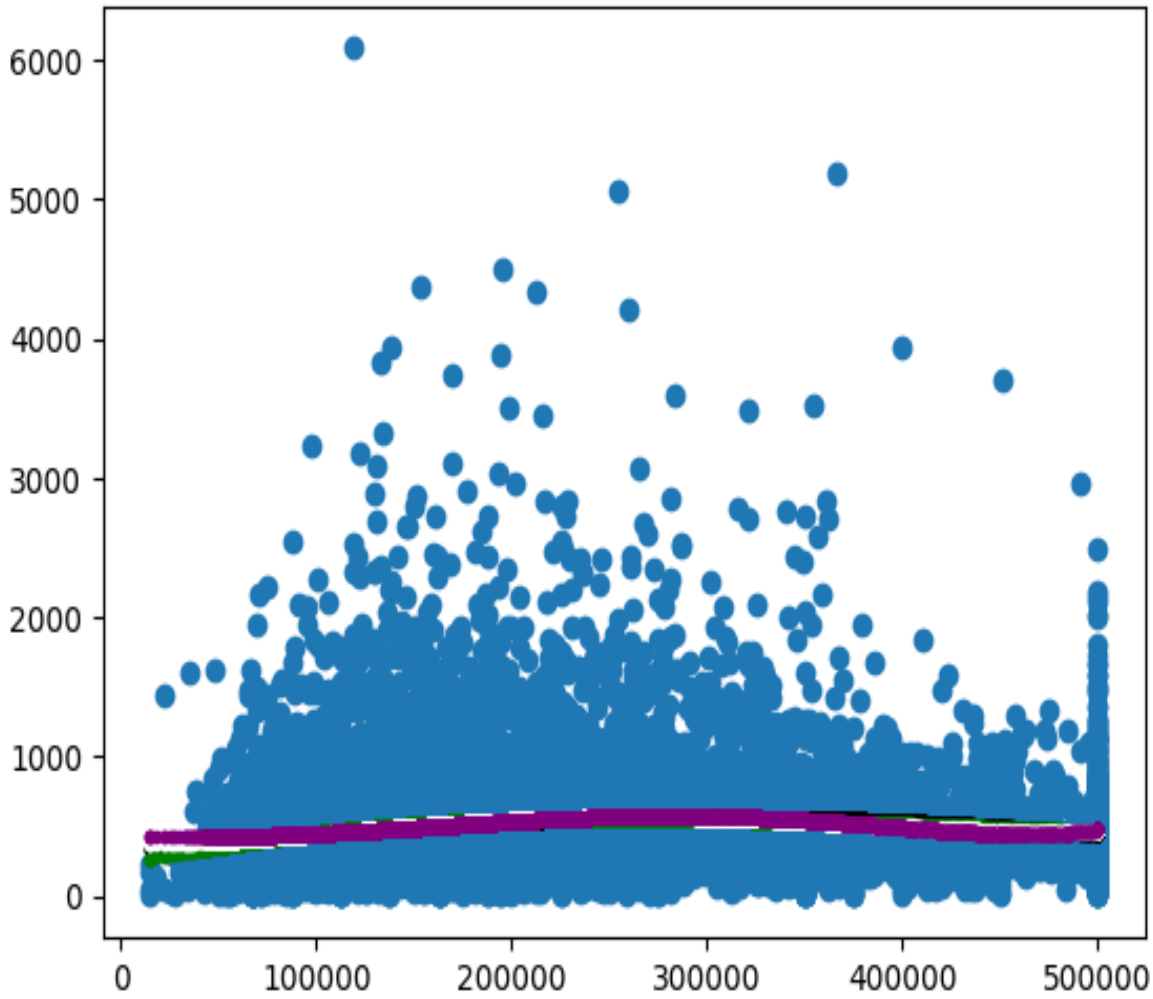

Figure 3. A graph containing the linear, quadratic, cubic, quartic, and quintic regressions of house value(x-axis) vs households(y-axis).

Above is the graph of house value against households. The mean squared error for cubic regression is the lowest in this particular graph, meaning that cubic regression fits the best for this particular graph.

Y-axis: Population

```
lineMSE = 1334253.4744410706
quadMSE = 1319133.3648495814
cubicMSE = 1316193.8678076607
quartMSE = 1318402.9026604511
quintMSE = 1322015.5588173107
```

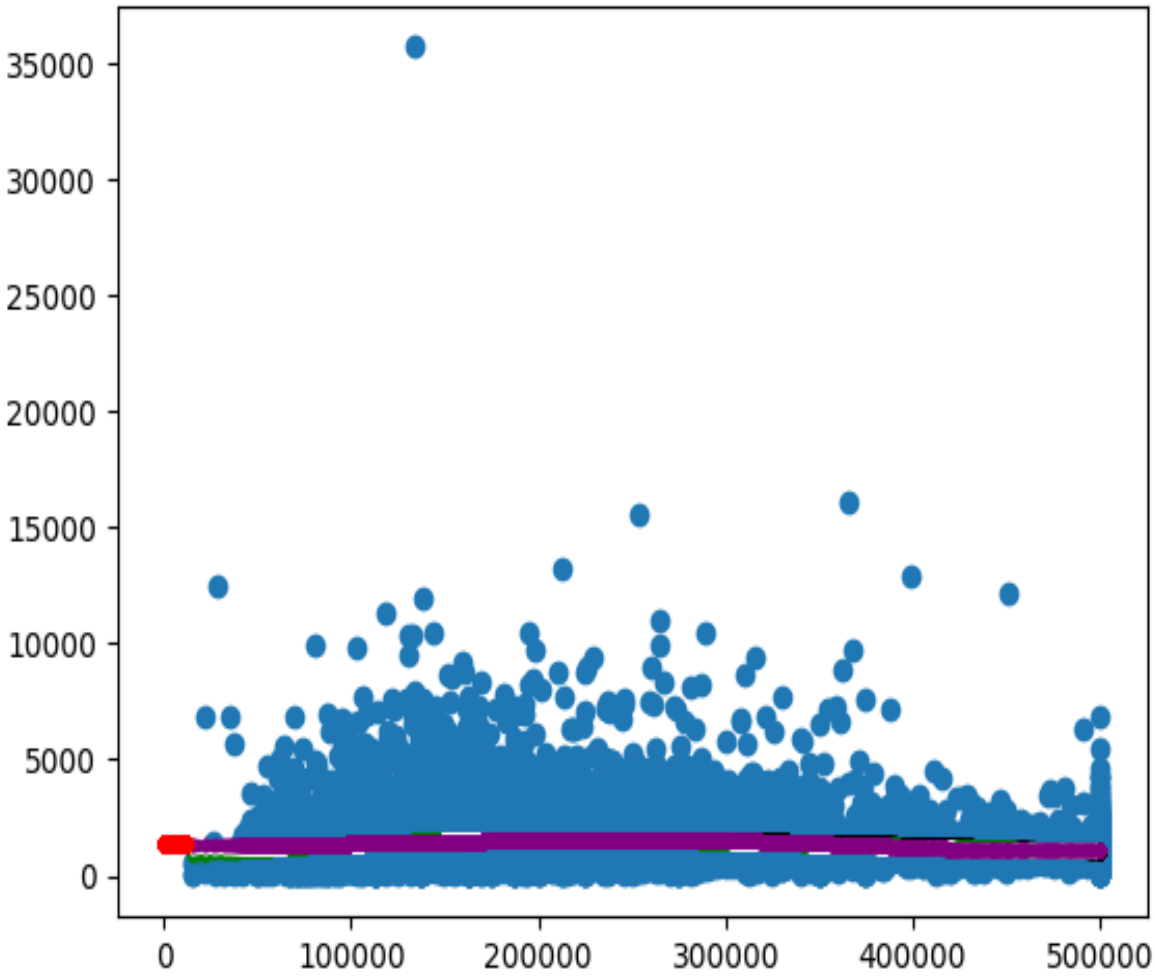

Figure 4. A graph containing the linear, quadratic, cubic, quartic, and quintic regressions of house value(x-axis) vs. population(y-axis).

Above is the graph of house value plotted against population. As with the graph comparing households and house values, the relationship between these two variables is also shown to be described with the lowest amount of error by a cubic regression.

Y-axis: Rooms

```
lineMSE = 4700825.44708672
quadMSE = 4664722.616149254
cubicMSE = 4664465.519384846
quartMSE = 4667951.779274081
quintMSE = 4673961.550753303
```

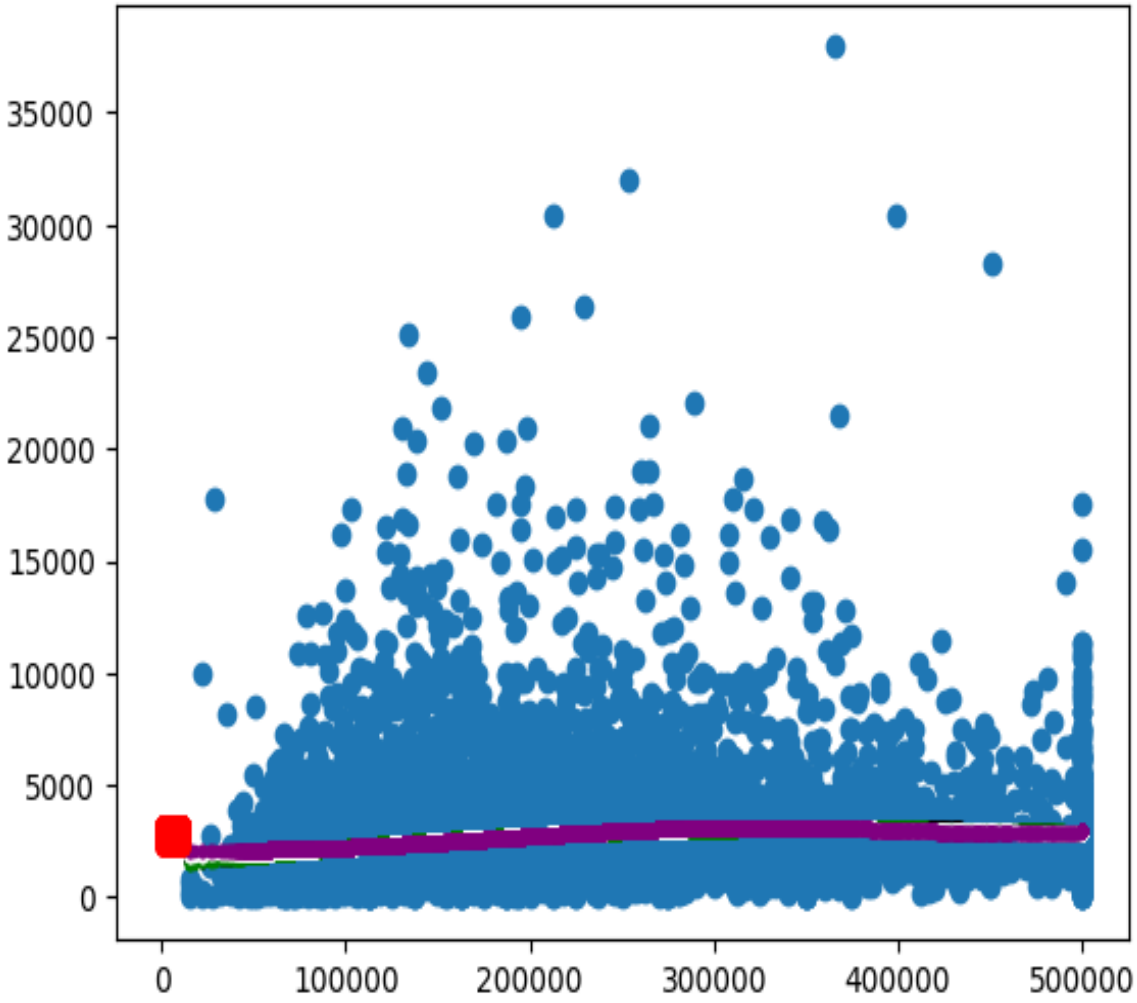

Figure 5. A graph containing the linear, quadratic, cubic, quartic, and quintic regressions of house value(x-axis) vs. rooms(y-axis)

The graph above describes the variables “house value” and “rooms” plotted against each other. Again, though not by much, a cubic regression is the best regression model for representing the relationship between these two variables.

Y-axis: Housing Age

```
lineMSE = 155.39186473184284
quadMSE = 154.58209809910278
cubicMSE = 154.49340137671444
quartMSE = 154.58097821785304
quintMSE = 154.63222914409985
```

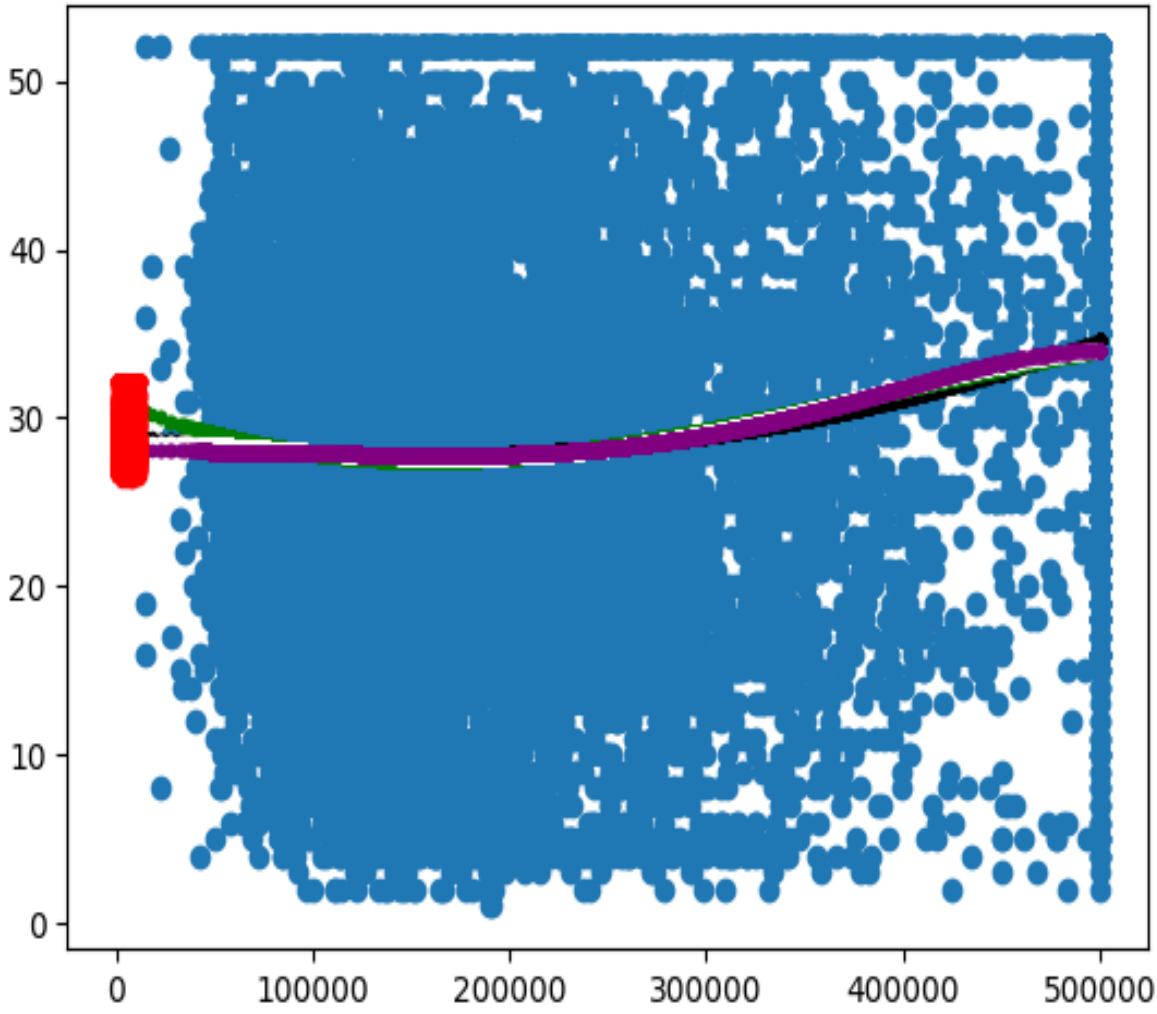

Figure 6. A graph containing the linear, quadratic, cubic, quartic, and quintic regressions of house value(x-axis) vs. housing age(y-axis).

The value shown on the y-axis of the graph above is the house's age, while the values on the x-axis are the house's value. Once again, a cubic regression most accurately describes the relationship between the two variables.

Y-axis: Income

```
lineMSE = 1.8508851365903802
quadMSE = 1.849931069575846
cubicMSE = 1.8321928560620975
quartMSE = 1.8265613266305916
quintMSE = 1.8379077782977753
```

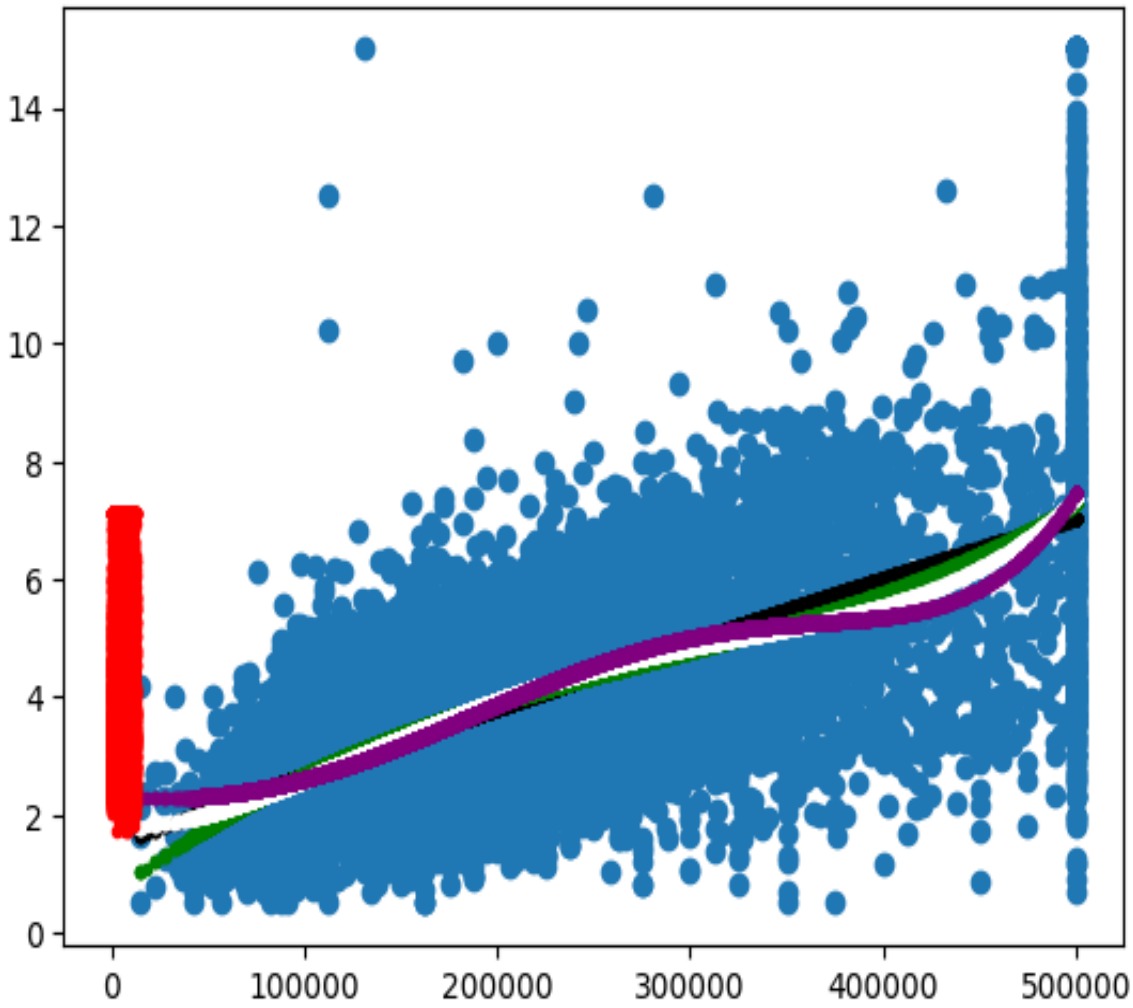

Figure 7. A graph containing the linear, quadratic, cubic, quartic, and quintic regressions of house value(x-axis) vs income(y-axis).

The graph above displays the variable “house value” plotted against the variable “income”. Although the graph had previously been visually identified as having a somewhat linear correlation, the regression models state that the most accurate model in describing the two variables’ relationships is the quartic model. Although the difference in mean squared error is trivial (only a difference of 0.024), buildups of large quantities of these non-optimal regression choices could lead to predictions that are of much lower precision in the latter stages of the neural network creation.

Y-axis: Latitude

```
lineMSE = 4.430355536289826
quadMSE = 4.334723376098095
cubicMSE = 4.261329710782332
quartMSE = 4.281382566965212
quintMSE = 4.344138744281738
```

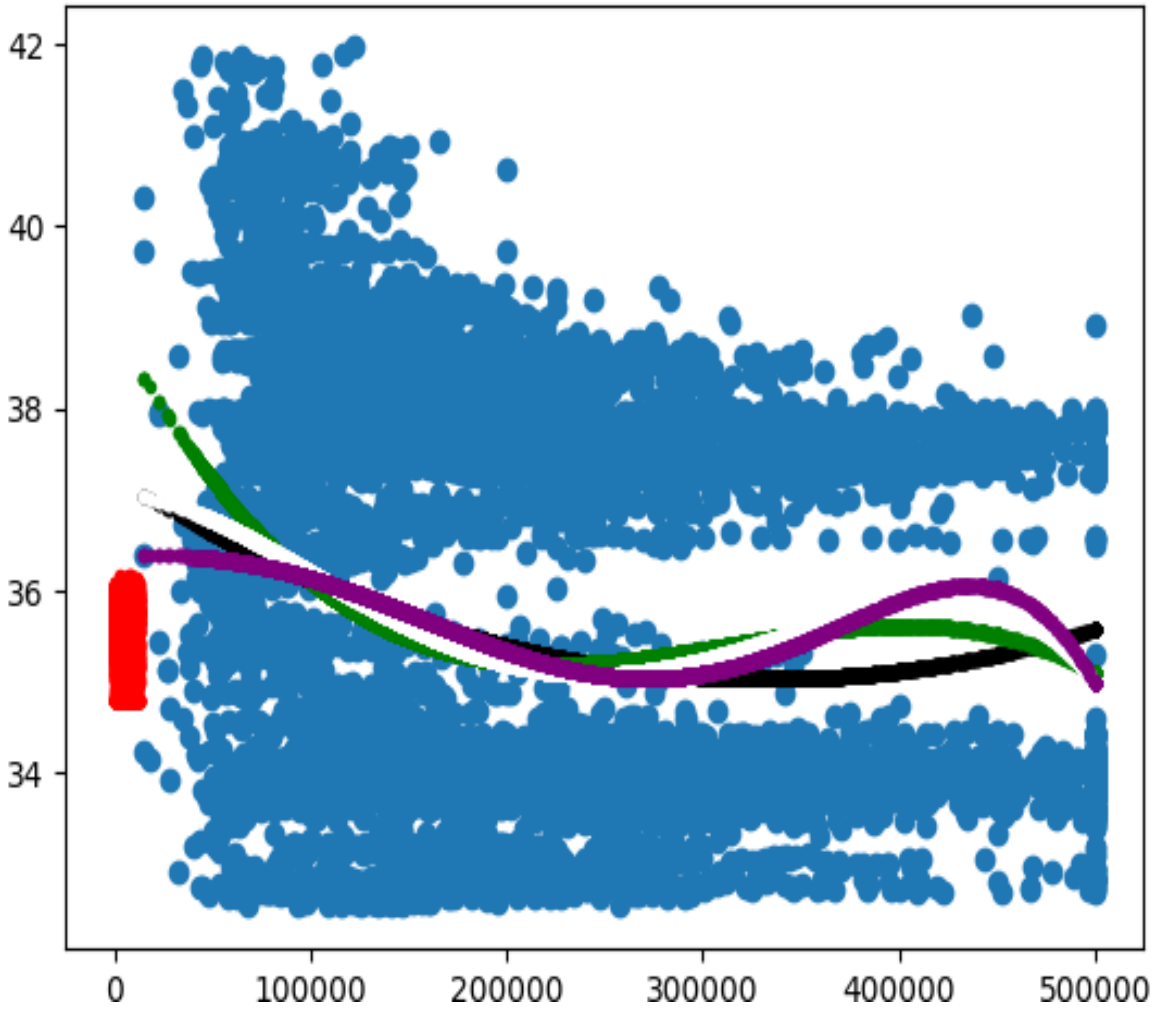

Figure 8. A graph containing the linear, quadratic, cubic, quartic, and quintic regressions of house value(x-axis) vs latitude(y-axis).

Above is the graph comparing latitude to house value. Although the graph looks somewhat erratic, and intuitively latitude may seem not to have any correlation with house values, the mean squared error values of the regressions indicate that the pair of variables does have a strong relationship. Specifically, the best description of their relationship is with a cubic regression.

Y-axis: Longitude

```
lineMSE = 3.967220965372894
quadMSE = 3.9622778023976557
cubicMSE = 3.9219115621675122
quartMSE = 3.929591158019126
quintMSE = 3.943884673096805
```

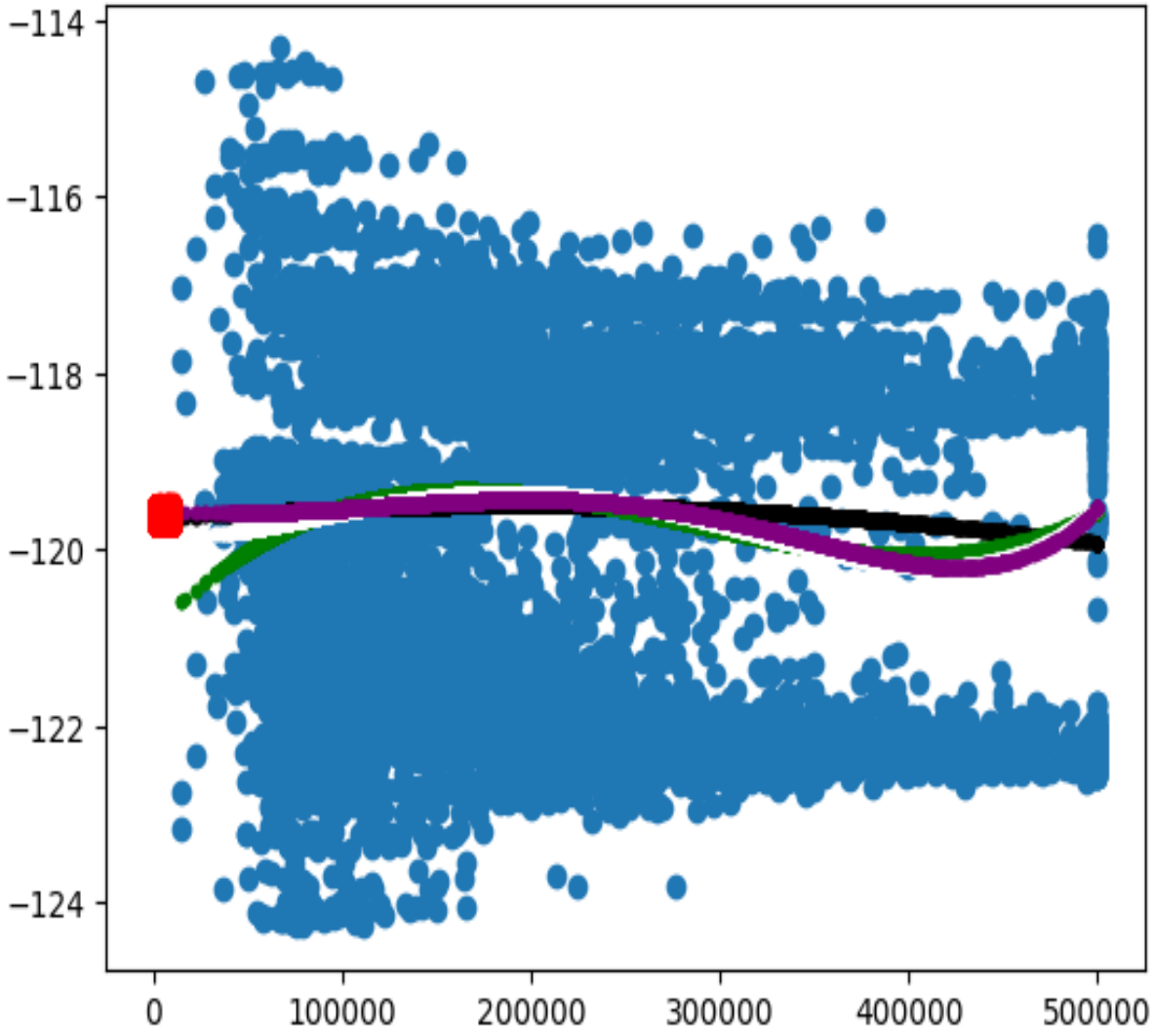

Figure 9. A graph containing the linear, quadratic, cubic, quartic, and quintic regressions of house value(x-axis) vs. longitude(y-axis).

This graph is quite similar to the previous graph describing latitude plotted against house value. The only differences between the two are that this graph replaces latitude with longitude and that the mean squared error values of these regressions are notably lower than those of the regressions of the previous graph.

Conclusion

After finishing up with the identification of relationships between variables, the process concludes with the creation of the neural network, the accumulation of all previously explored information. After normalizing all the data points (scaling them all down by the same factor) using the StandardScaler utility from the sklearn library, the neural network was created with the new scaled data using the Keras API from the TensorFlow library. Using a sequential model from the library, the neural network is trained on a set of data containing 70% of the 15,000+ total data points and features multiple layers of nodes with varying numbers of nodes in each layer. Each layer of nodes feeds into the following layer until eventually a single, final output node is reached: the predicted value. After inputting data points from the testing data set (the 30% of data points that were not used to train the neural network), the predicted house values are then compared with the actual values from the data points, and a mean squared error is calculated with the errors resulting from each of the testing data points. After many trials of models with different numbers of layers and different amounts of nodes on each layer, a model containing 6 layers of nodes with 64 nodes on the first layer, 32 nodes on the second layer, 16 nodes on the third layer, 8 nodes on the fourth layer, 4 nodes on the fifth layer, and a single node on the final layer proved to have the best outcome, with a mean squared error of 0.4728 (reference Appendix A for more details about each iteration). The model kept the majority of the data columns but dropped the columns 'longitude', 'latitude', and 'median housing age'', each of which caused far larger mean squared errors when included in the neural network. By dropping the mentioned data columns, the model was able to decrease the error by a total of almost 45% of the original error.

Future Direction

The results of the exploration conducted within this paper are limited by both time constraints and constraints within the dataset, such as a limited quantity of data points and a very limited amount of data within each said data point, having only 9 different variables total. With more time and a more extensive dataset, a higher degree of accuracy than that of the neural network built alongside the paper could be reached. Additionally, since all of the data points were retrieved from California, the results of the neural network can only be applied to houses within California, as conditions may vary from state to state. With data retrieved from wider geographical ranges, the coverage area of the neural network, or the area over which the neural network would be able to accurately predict house prices, would also increase. However, despite some shortcomings, the degree of accuracy achieved by this dataset is still enough to provide a good approximation of what the price of a house in California, without consideration of the economy, would be. A possible future step for the project could be to create a website connected to the neural network that could be used as an alternative to Zillow or as a backup if Zillow's website ever goes down so that regardless of where someone lives, they will always able to determine the rough price of their house for no cost and with little time spent. Another alternative step the project could take would be to increase the number of data points, as mentioned earlier, and to widen the coverage of the neural network. With many options on steps that could be taken in the future, the project, given enough time and opportunity, has plenty of room to grow.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

Opendoor | Sell your home the minute you're ready. (n.d.). Opendoor.com. Retrieved September 23, 2023, from https://www.opendoor.com/articles/how-long-does-an-appraisal-take

Runkle, L. (2022, August 11). Should Home Sellers Get an Appraisal Before Listing? Here's Why It'll Cost You. Real Estate News & Insights | Realtor.com®. https://www.realtor.com/advice/sell/should-home-sellers-get-an-appraisal-before-listing/#:~:text=%E2%80%9CAppraisals%20aren

Han, B. (n.d.). Real estate industry accounted for 16.9% of GDP in 2021. RealTrends. https://www.realtrends.com/articles/real-estate-industry-accounted-for-16-9-of-gdp-in-2021/#:~:text=Real%20estate%20industry%20accounted%20for%2016.9%25%20of%20GDP%20in%202021%20%2D%20RealTrends

Admin. (2015a, October 15). Internet access and digital holdings in libraries. Tools, Publications & Resources. https://www.ala.org/tools/libfactsheets/alalibraryfactsheet26#:~:text=Almost%20all%20public%20libraries%20(99.5,do%20not%20offer%20this%20service.

Corporate - about. Zillow. (n.d.). https://www.zillow.com/z/corp/about/

Arthur, M. (2022, April 13). How accurate is my zestimate, and can I influence it?. Zillow. https://www.zillow.com/learn/influencing-your-zestimate/

Fontinelle, A. (2021, December 22). How zillow makes money. Investopedia. https://www.investopedia.com/articles/personal-finance/110615/why-zillow-free-and-how-it-makes-money.asp