Hamed Sayed
Danny Garcia
5/09/2024

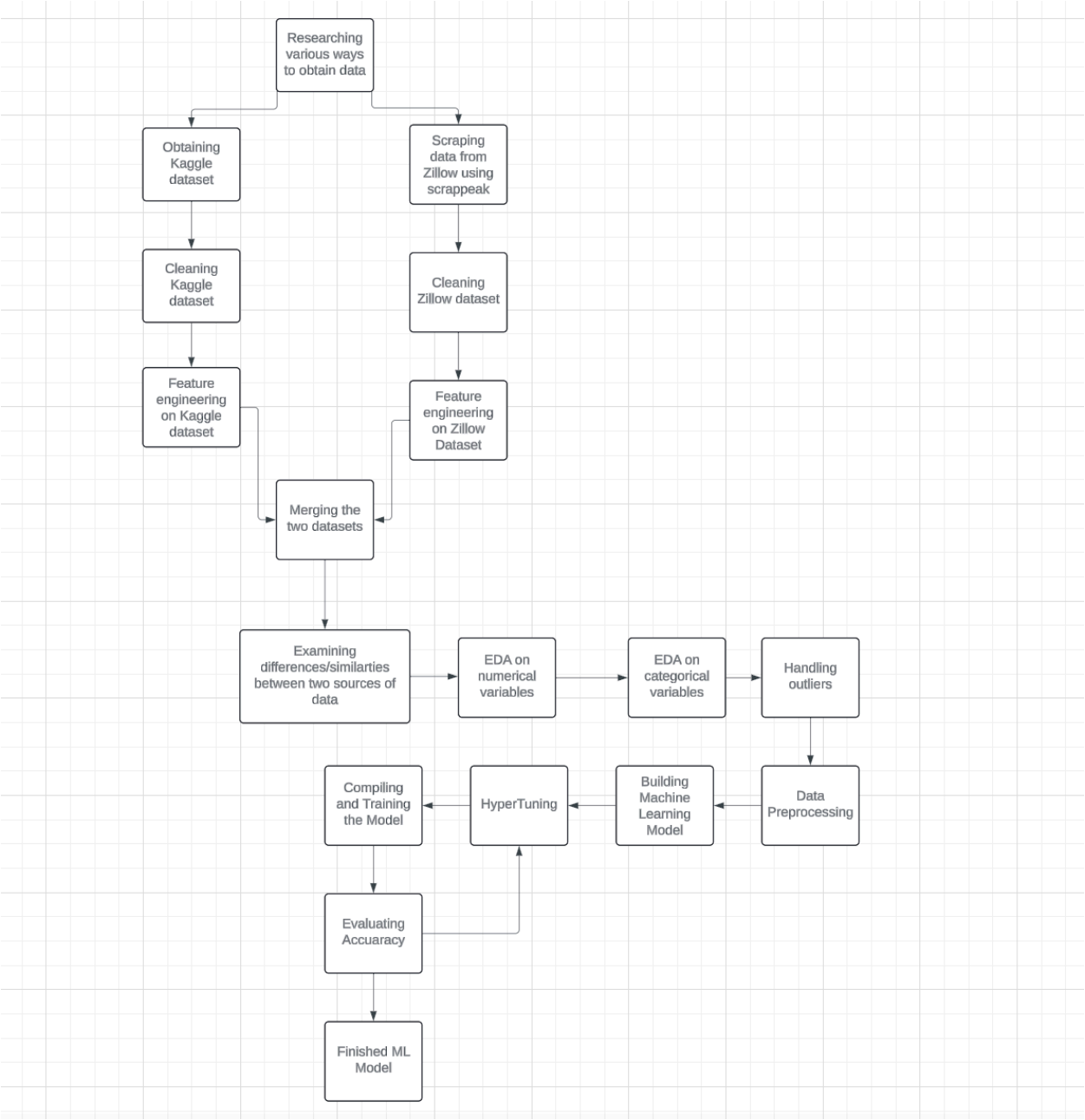
Analysis on New York Housing Market

Introduction:

For this project we decided to examine data regarding the New York Housing Market. We got our data from two sources. One of them was from the Kaggle New York Housing Market dataset. This dataset contained many valuable variables with regards to houses for sale in New York. For our other source of data, we used the Zillow scraper API from Scrapeak.com. This API is free for 1000 credits per account so it sufficed us during the project. The API allowed us to perform any search on a city/county on zillow then scrape all the listings from the search results. We used this API to perform search results on the different New York counties. Since our Kaggle dataset originally only contained five counties, scraping was useful so we could get a more overall view of housing across New york. It also allowed us to compare the two datasets to see how they differ.

We first examined our data and performed any cleaning/preprocessing that was needed; we also performed some feature engineering and pulled some information from our other columns. After that we used seaborn and examined trends in our data. After our exploration we created some machine learning models that would predict the housing price based on certain variables. This project was important because it can serve as a guide for anyone looking to buy a home in New York. By filling in their desired house features, square footage, number of bathrooms, city, etc. they will be able to get an estimate on how much it will cost them.

Methods:



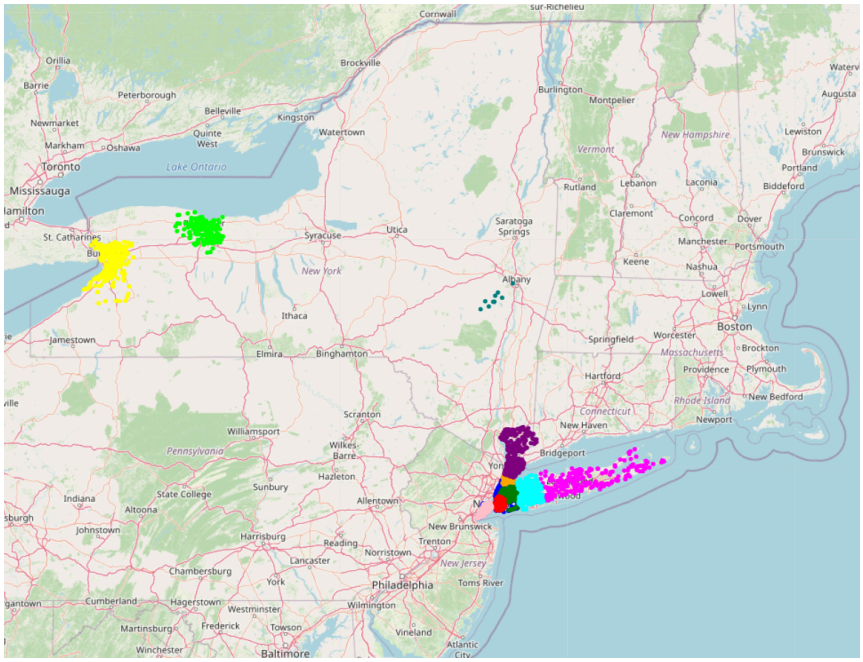
After our two sources of data were obtained we had to individually clean our two sources of data before merging. For both of our datasets we dropped duplicate rows and null columns. In our Kaggle dataset we dropped six columns and in our Zillow dataset we dropped many columns since it came with 95 columns. In both of our datasets we were able to perform some feature

engineering and extract a city column from the address column. We used an apply function that would retrieve the city from the address column and stored the city in a separate city column. In our Kaggle dataset there were some listings that were not valid property types so we dropped those rows. There were also some columns that did not have a county associated with them so we used our city column to match a county to them. After our datasets were cleaned we merged them but made sure to create a column that would specify the source of our data. We conducted exploratory data analysis to examine trends and find information about our data. After that we then used a correlation matrix to perform feature selection for our machine learning models. Once our features were decided, we built the pipeline for our machine learning models, using one hot encoding for our categorical variables. After we ran our data through our pipeline we then fed the data to our different machine learning models. We aimed to increase our model performance through hypertuning and tested our models accuracy with our test sets.

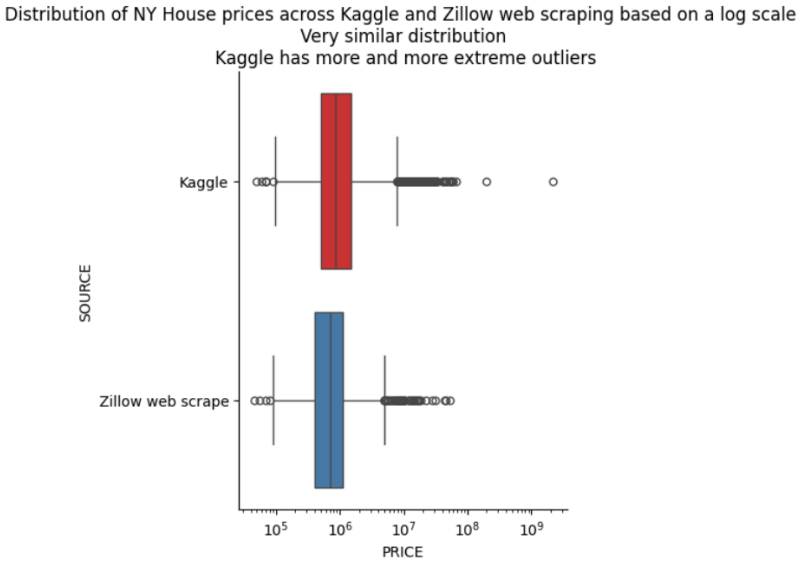
Materials:

For this project, we used the Kaggle New York Housing Market dataset as well as the Zillow Scraper API for our dataset. We wrote our code for data processing, data visualization, and machine learning model on Google Colab Notebook. A CSV file was also used for the input file.

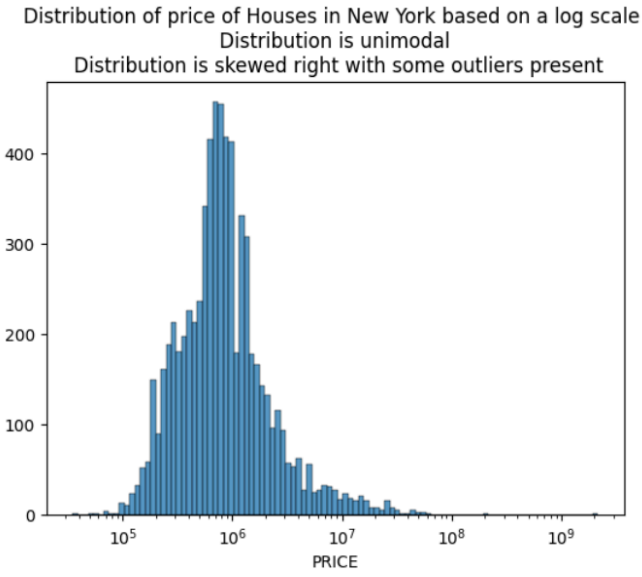
Results:



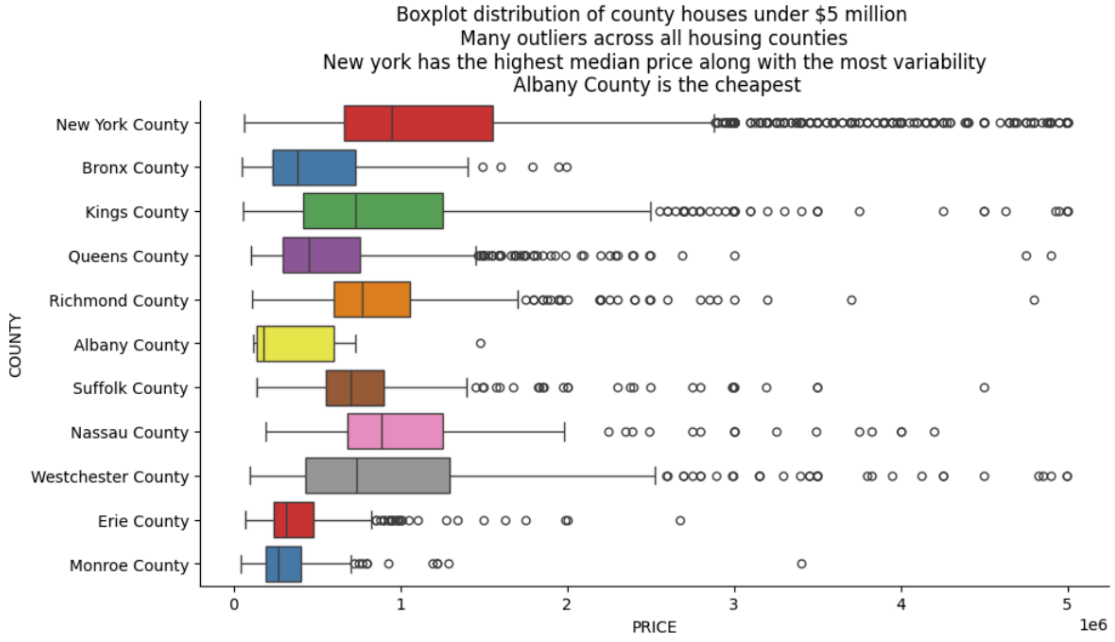
The above plot is an interactive plot we created using folium. Each color represents one of our eleven New York counties in our dataset and each dot represents a specific listing. When working with this map, users are able to select individual listings to receive more information about that specific property. Our dataset contained listings from New York county (blue), Queens county (green), Kings county (red), Bronx county (orange), Westchester county (purple), Erie county (yellow), Nassau county (cyan), Suffolk county (magenta), Monroe county (lime), Richmond county (pink), and Albany county (Teal). From a glance we can see that there are not that many property listings when it comes to Albany county.



When comparing our two sources of data, they contained nearly identical distributions across numerical attributes. The attribute with the most difference between the two datasets appeared to be the price. As seen in the box plot distribution of our two sources of data, they have similar IQR's as well as medians; however our Kaggle dataset seems to have more extreme outliers. Our more extreme outlier could potentially be a user error by whoever uploaded the dataset to Kaggle since after doing more research we were not able to find any property that was listed for that price online.



The above histogram displays our property price distribution based on a log scale. We can see that our distribution is clearly unimodal with the mode right below 10^6. Our price distribution is skewed right and there are outliers on both sides however there appears to be more outliers towards the right side of our graphs. From our graph we can infer that there are more properties which are priced above the median than there are below it. Since our price distribution did not follow a normal distribution which would be evident through the bell curve, it made us consider how to deal with outliers before constructing our machine learning models.



The above plot shows a box plot distribution of our county property prices of listings under 5 million dollars. We can clearly see that there are many outliers across all housing counties. New York has the highest median price along with the most variability when it comes to price. Albany county is the cheapest, however it is important to note that we had the least amount of samples for Albany county so it may not be a true indicator of its price.

When looking at all of our plots created we were able to make several conclusions about our dataset. When comparing our Kaggle and Zillow web scraped data, they were almost

identical however the Kaggle dataset had more extreme outliers when it came to price. We found that larger houses with more bedrooms and bathrooms were typically more expensive which is what we expected. We found that multi-family homes are the most expensive, followed by townhouses, houses, condos, and co-ops. Along with that, townhouses have the highest variability when it comes to price with co-ops having the least. When looking at the prices of counties, New York county was on average the most expensive however it also had the most variability when it came to real estate prices. Nassau County was the second most expensive while Albany county was the cheapest. Monroe County had the least variability when it came to real estate prices.

Regarding the machine learning model, the machine learning model that had the highest accuracy was only able to reach an accuracy of up to 75%. We built various models with different model architectures and different hyperparameters, which gave us mixed results. The first Machine Learning model that we created was a Convolutional Neural Network model which consisted of 3 neural networks and an Adam optimizer. We also used the mean squared error loss function during training because we felt like that was the best fit for the purpose of the model. After specifying certain hyperparameters, training, and evaluating the model, the CNN model only reached a test accuracy of about 45%.

The second model was a Support Vector Regression model which had the lowest accuracy out of all the models with a test accuracy of about 4.5%. It wasn't until we trained our third model that we reached our highest test accuracy out of all the models. Our third Machine Learning model was a Gradient Boosting Regression model. We actually ended up building two different GBR models because the first one we trained with the necessary hyperparameters and after training and evaluating the model, it had a test accuracy of 75%. Therefore, we decided to

train a second GBR model, this time tuning it using a grid search and feature selection. However, the second GBR model actually ended up having a lower test accuracy than the first one, with a test accuracy of about 67%. We believe that the second Gradient Boosting Regression model had more fine-grained control over the model's performance which allowed for overfitting, especially with the limited dataset that we had, therefore it had a lower accuracy at predicting the housing market price than the simpler GBR model.

Discussion:

There were a couple of technical challenges that we encountered during this project. The first challenge we encountered was finding another source of data other than our Kaggle data. We wanted to find a way to pull live data just to ensure we are obtaining up to date values on the current New York housing market. We first tried to see if Zillow had their own API to do this. We found out that although Zillow does have their own API, it does not allow you to obtain live listings of houses, rather it allows you to obtain different information regarding current housing metrics. Although it was interesting, it was not what we were looking for in this task. We then looked at Redfin. Redfin has a site where you can download CSV data however it does not have an API. Redfin's data also does not allow you to obtain data regarding live housing. Through more research we came across Scrappeak's Zillow web scraper. This API allowed us to perform any search on a neighborhood/county and scrape all the listings associated with that search. This was very helpful since it gave us a secondary data source that was from live listings. The one downside to this API is that since we were using the free version, we were limited to the amount of searches we had. An area we can improve upon with this project is to upgrade our accounts so we can perform more searches. This way we can perform a search on every county in New York and not just the top eleven most populated ones.

Another technical challenge that we had was that our machine learning model wasn't as accurate as we wanted it to be. We first started by creating a Convolutional Neural Network model with 3 neural layers. After training the model, it only reached an accuracy of about 45%. Our second ML model, which used a Support Vector Regression model, was even less accurate than the first model. It wasn't until we created our third model which was a Gradient Boosting Regression model that we reached an accuracy that we were somewhat satisfied with. However, since our most accurate model only reached an accuracy of 75%, we would still like our model to be more accurate. We tried tuning the GBR model with a grid search and feature selection, trying to make the model more accurate, but this led to an even less accurate model. We theorized that this happened because of overfitting due to a limited dataset. For continuing work, we believe trying different model architectures is our best approach.

Instructions:

Open the .ipynb file on Google Colab

Note: You can use one of our API keys in the code cells or create your own by following the instructions below

1. Setting up Scrappeak Account:

- If you don't have a Scrappeak account, visit scrappeak and create one.
- After logging in, navigate to the "Plans" section and select "Try for free" to activate your free trial.
- Go to "Scrapers" and choose the Zillow Scraper API, then select "try for free."
- Click on the "API and Documentation" tab, and copy your API key provided there.

- Paste your API key into the second code cell

2. Obtaining Kaggle Data:

- Visit the New York Housing Market Kaggle dataset or download from Google Drive (skip to the last bullet point if downloaded from drive)
- Download the dataset by clicking on the "Download" button.
- Extract the contents of the downloaded zip file
- Run the first code cell and choose NY-House-Dataset.csv to upload

3. Running The Program:

- Ensure that one API key is provided in the second code cell. One API key should be left uncommented. If API keys have exceeded their credit limit, please provide your own API key by following the instructions in Step 1.
- Run the entire program
- Program will process and clean the dataset, providing visual aides to better understand the dataset.
- It will then train a Machine Learning Model to predict the future housing market prices