PROJECT

HOUSE PRICE PREDICTION USING DATA MINING

GROUP MEMBERS:

MAHSA RAHIMIAN

RUMANA SULTANA

COURSE NAME: DATA MINING

INSTRUCTOR: DR. FARNOUSH BANAEI-KASHANI

THE UNIVERSITY OF COLORADO DENVER

GITHUB LINK: HTTPS://GITHUB.COM/RUMANACU/HOUSE-PRICE-PREDICTION

OUTLINE

- Problem statement
- Methods
- Tools
- Results
- Lessons learned
- References

PROBLEM STATEMENT

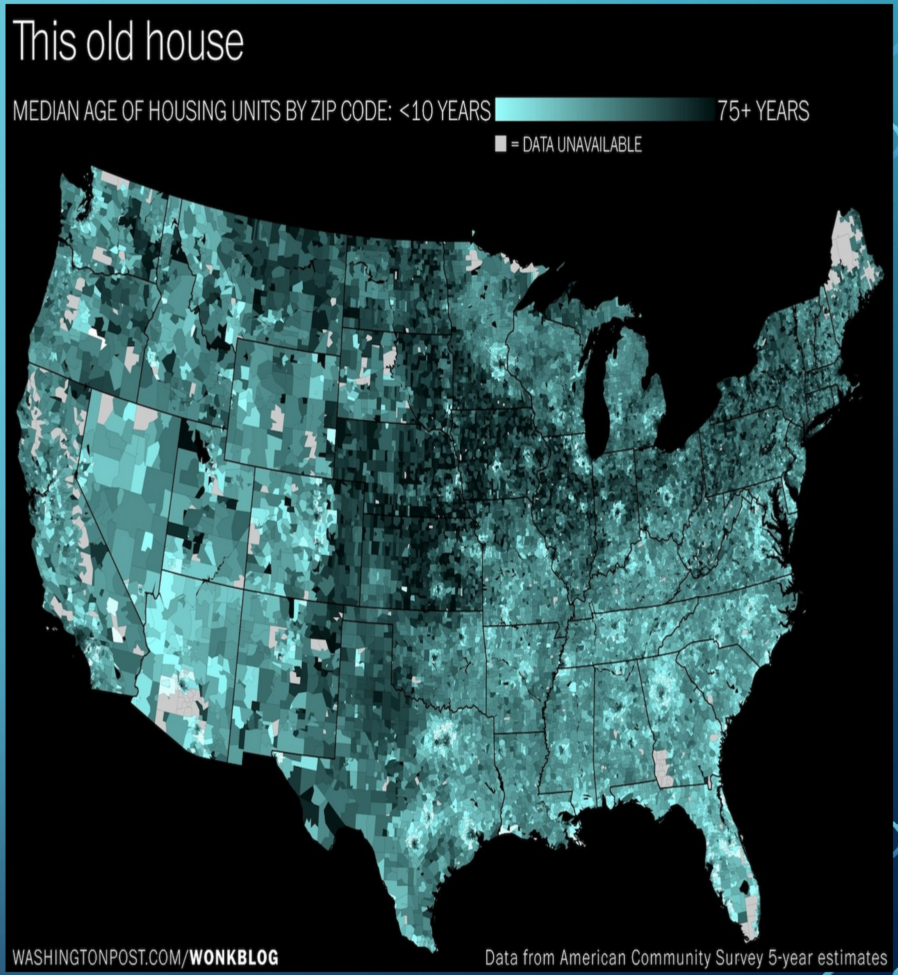
- Our goal in this project is to predict efficient house pricing for investors based on their priorities and budgets.
- In this paper, we will use different data mining strategies and linear regression algorithms to predict prices by analyzing current house prices, and then predict the future prices.

PROBLEM STATEMENT

- This project can impact anybody who is looking for house, real estate’s companies, and those who wants to invest on proper properties.

DATA SET AND DATA PROCESSING

In this project the first step that we are performing i data preprocessing. the dataset for this project has been obtained from Zillow.com. We used Zillow Home Value Index (ZHVI) [1] dataset that is a smoothed, seasonally adjusted measure of the typical home value and market changes across a given region and housing type. It reflects the typical value for homes in the 35th to 65th percentile range.The dataset has different type of time series housing price data per month from january 2000 to October 2021.


This old house
MEDIAN AGE OF HOUSING UNITS BY ZIP CODE: <10 YEARS
75+ YEARS
= DATA UNAVAILABLE
WASHINGTONPOST.COM/WONKBLOG
Data from American Community Survey 5-year estimates

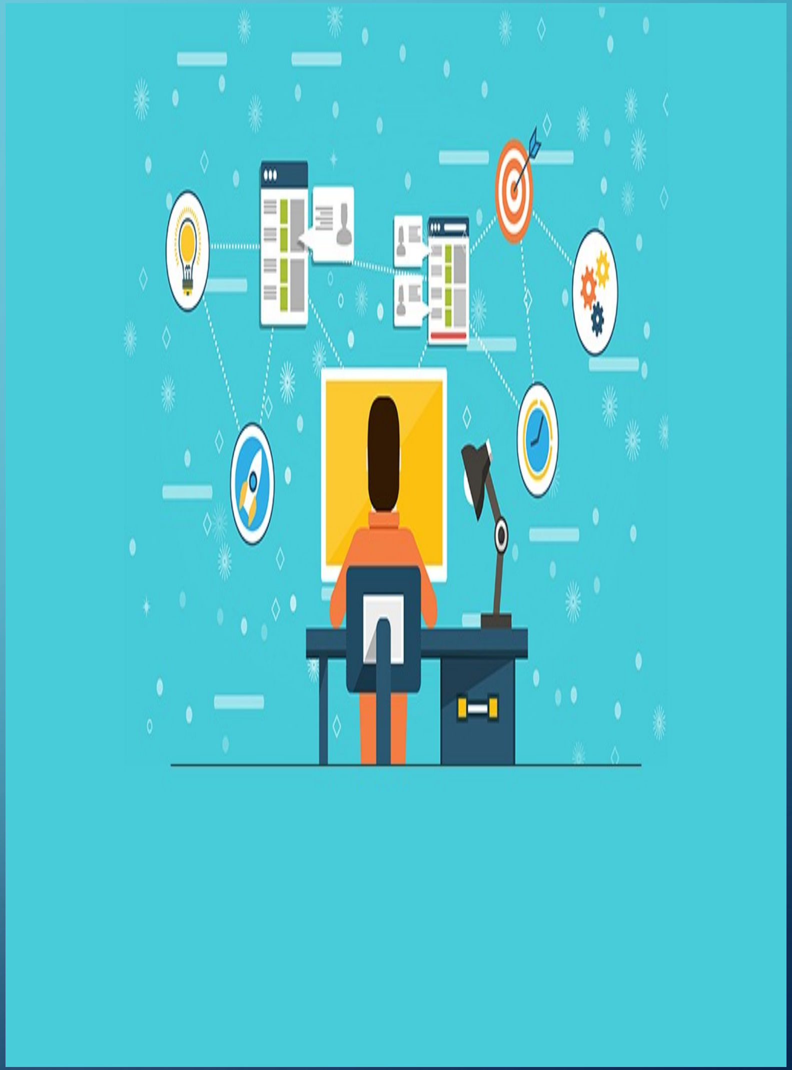
DATA ANALYSIS

In this step we can discover the implicit patterns of the data, which in turn helps choose appropriate machine learning approaches.

MODEL SELECTION

Before building models, the data should be processed accordingly so that the models could learn the patterns more

efficiently.



METHODS

SOME DATA VISUALIZATION

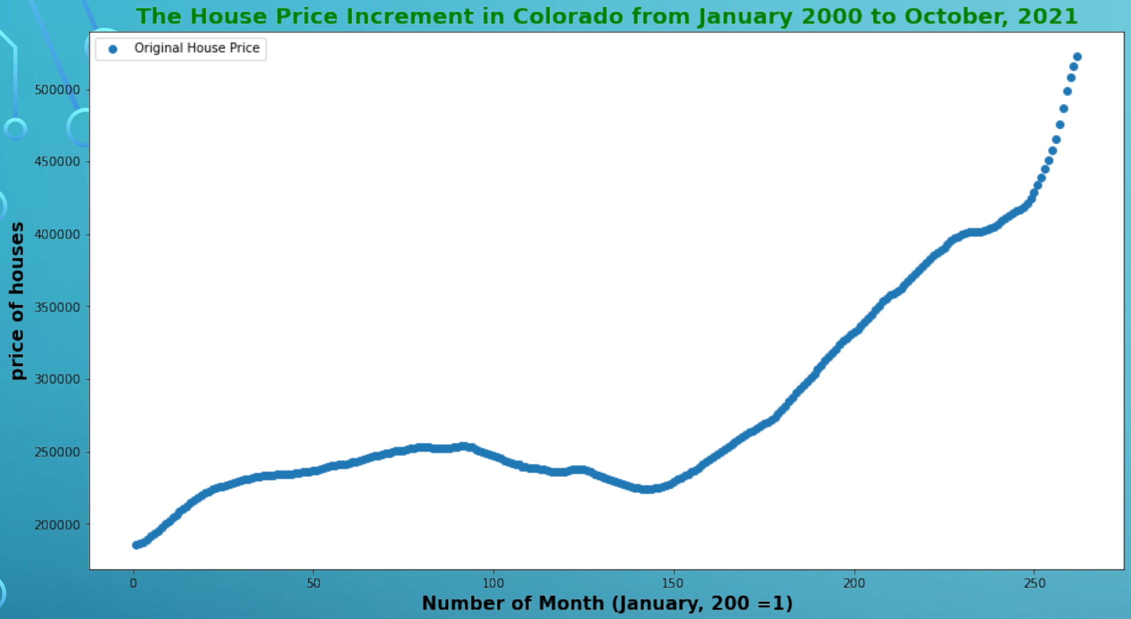


Figure 1: Increment in house price in Colorado from January 2000 to now

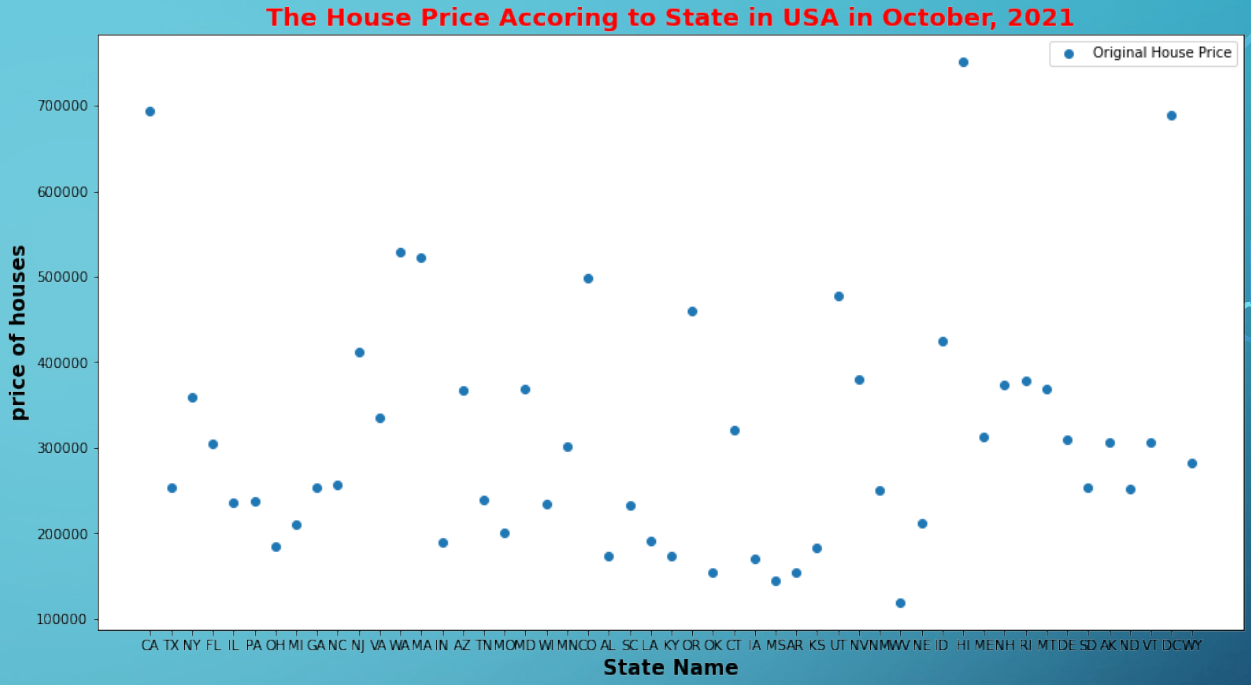


Figure 2: Statewide plot of house price on October, 2021

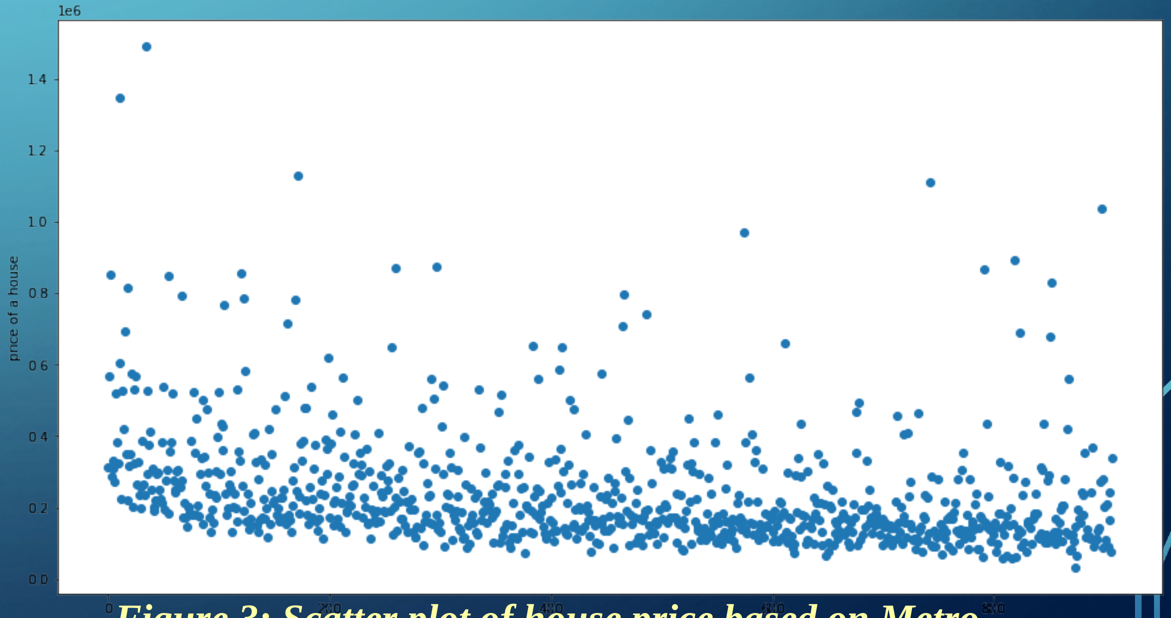


Figure 3: Scatter plot of house price based on Metro

- Metro-based Data set form Zillow
- Taking Average Price of all months

METHODS

SOME DATA VISUALIZATIONS

- Metro-based Data set form Zillow
- Taking Average Price of all months

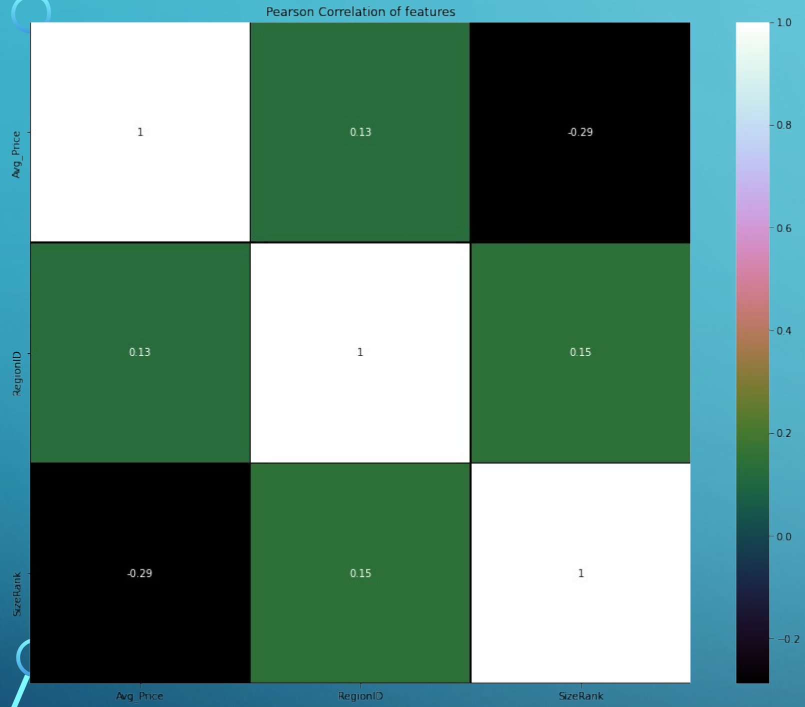
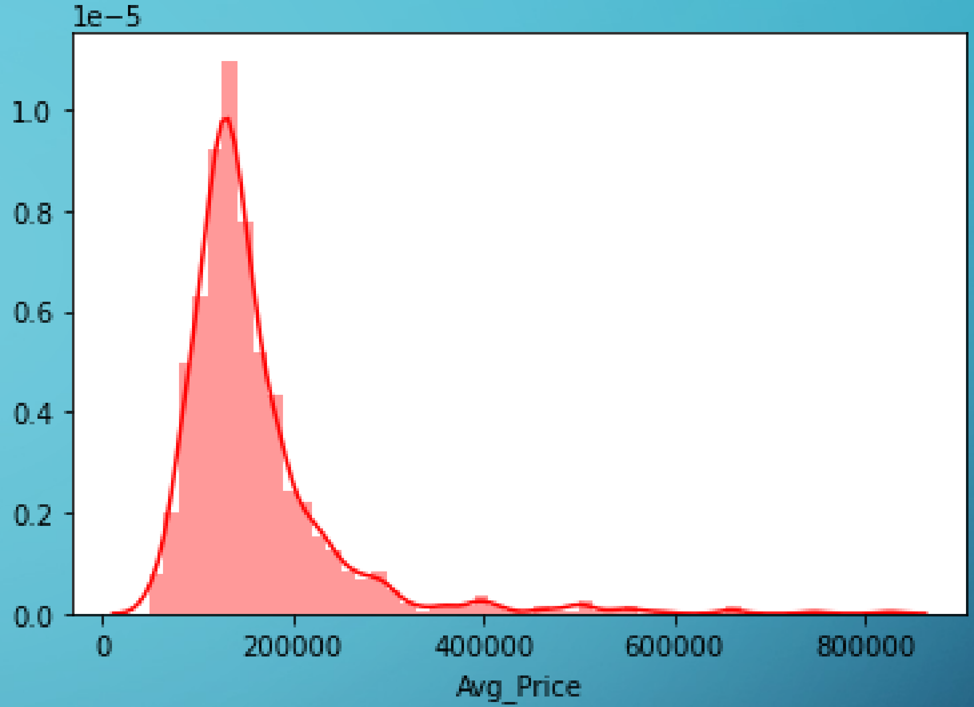


Figure 4: Pearson Correlation between features with average price (Metro)

Figure 5: Distribution plot on average price in Metro data



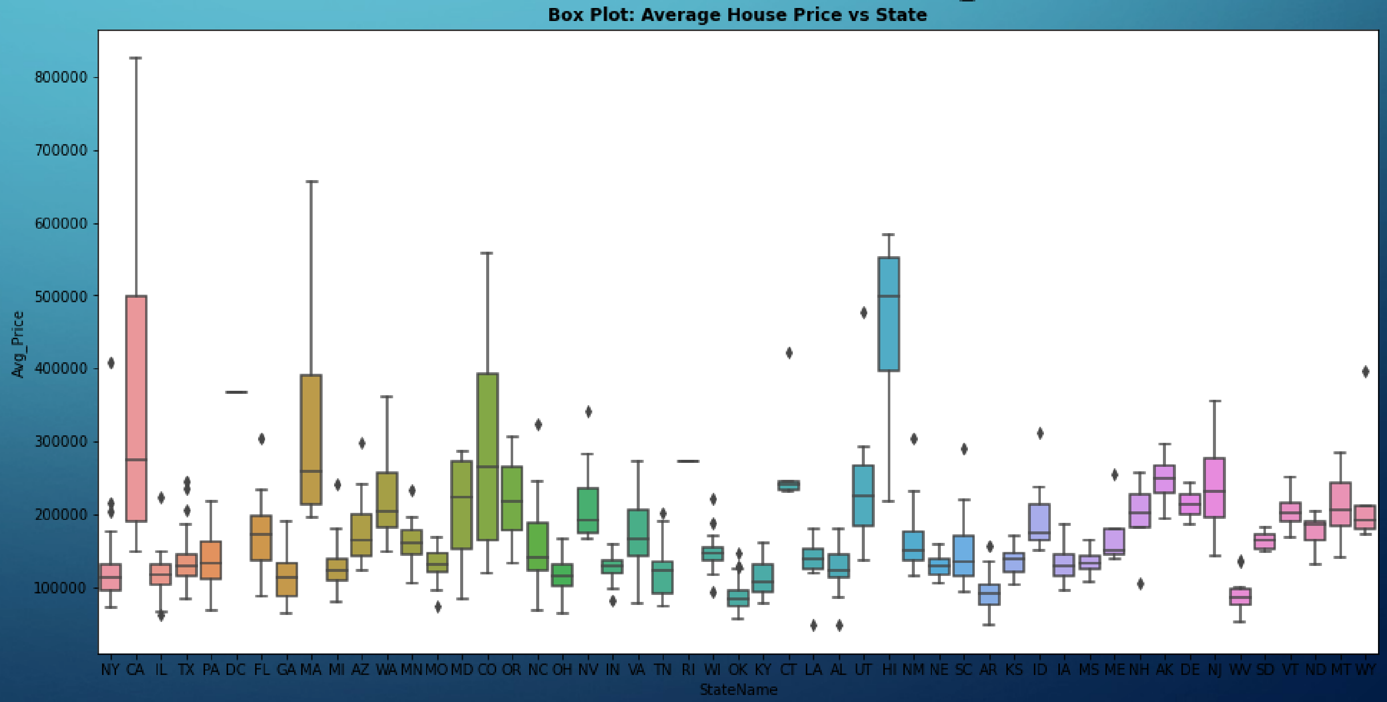


Figure 6: Box plot of average house price of metropolitan data (Statewise)

METHODS

DATA PREPARATION

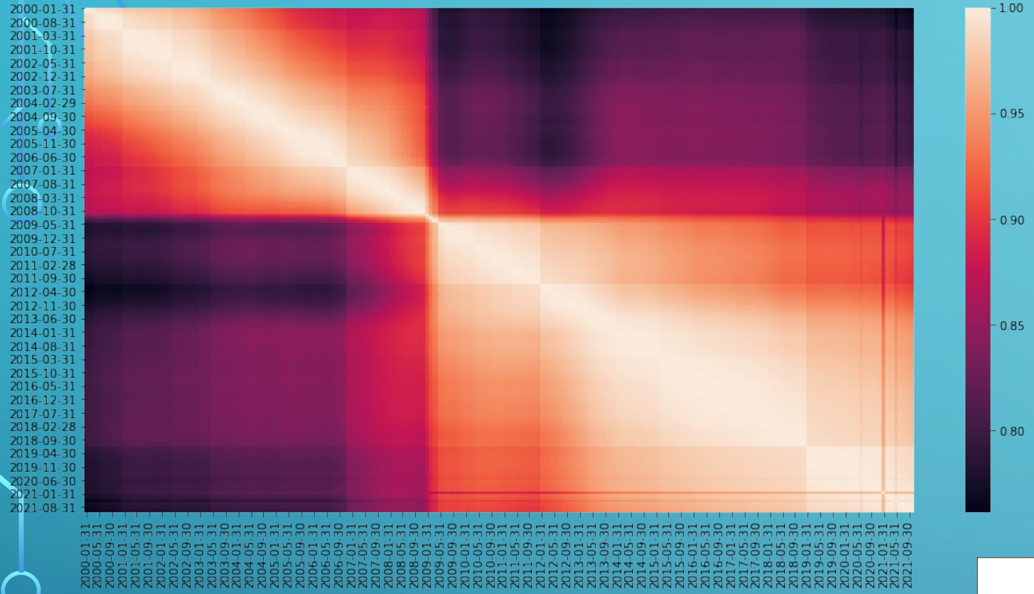
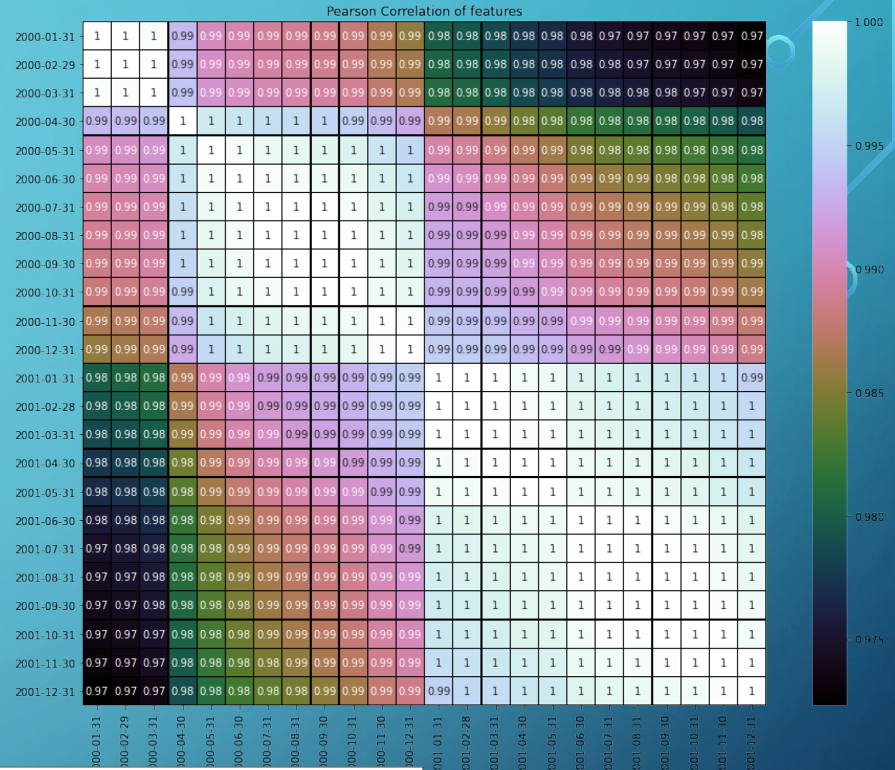


Figure 7: Correlation between all months from 2000 to 2021

Figure 8: The last 24 months house prices correlation



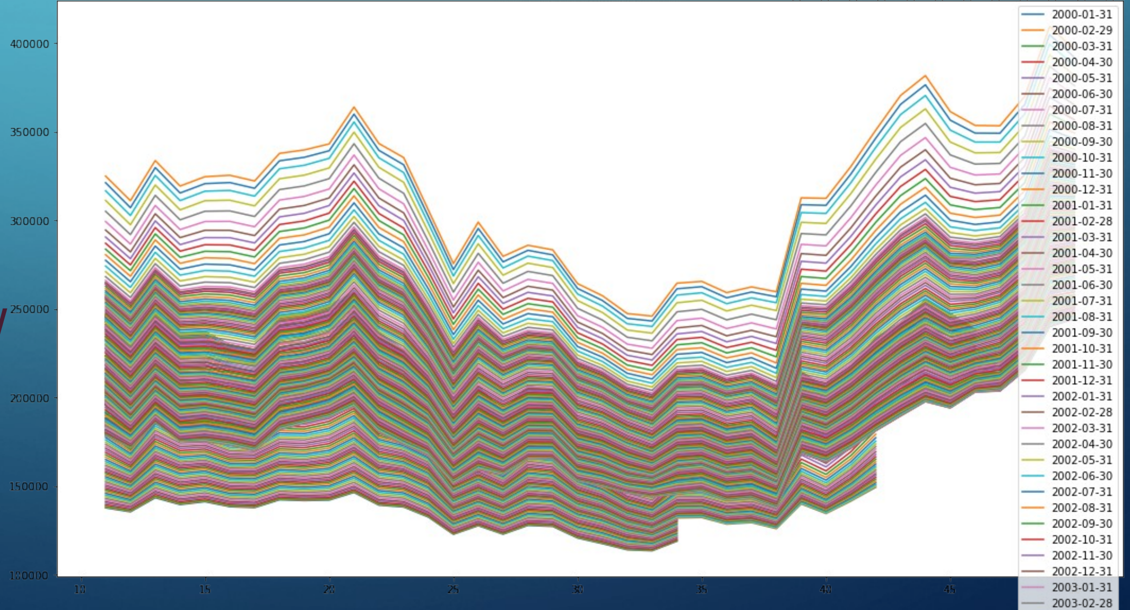


Figure 9: All months' price plot after data preparation (State based)

- Metro-based dataset form Zillow
- Taking prices of all months

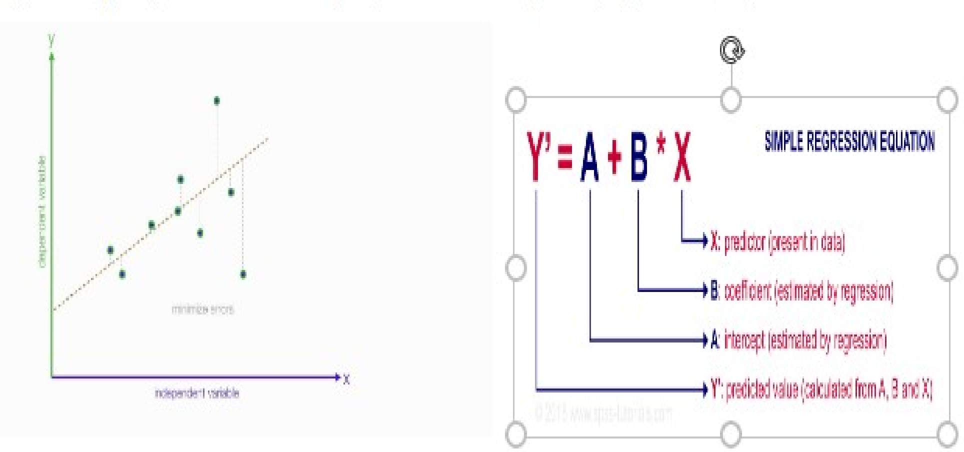
METHODS

MODEL DEVELOPMENT

Algorithm:

- Calculate x_mean and y_mean.
- Calculate the difference between (x[i], x_mean) and (y[i],y_mean)
- Calculate the square of the difference between (x[i], x_mean)
- and sum
- Calculate the products of the difference between (x[i], x_mean) and (y[i],y_mean) and sum
- Calculate the coefficient, B by the equation B =(diffx * diffy)/(diffx *diffx)
- Calculate the intercept A by the equation A=y_mean- (B* x_mean)
- Finally calculate the prediction by the equation: Y_pred=B*x+A and draw the regression line.

Figure 10: Linear regression model


y
dependent variable
minimize errors
x
independent variable
Y' = A + B * X
SIMPLE REGRESSION EQUATION
X: predictor (present in data)
B: coefficient (estimated by regression)
A: intercept (estimated by regression)
Y': predicted value (calculated from A, B and X)
© 2015 www.spss-tutorials.com

[4] Source: http://mybooksucks.com

[4] Source: https://spss-tutorials.com

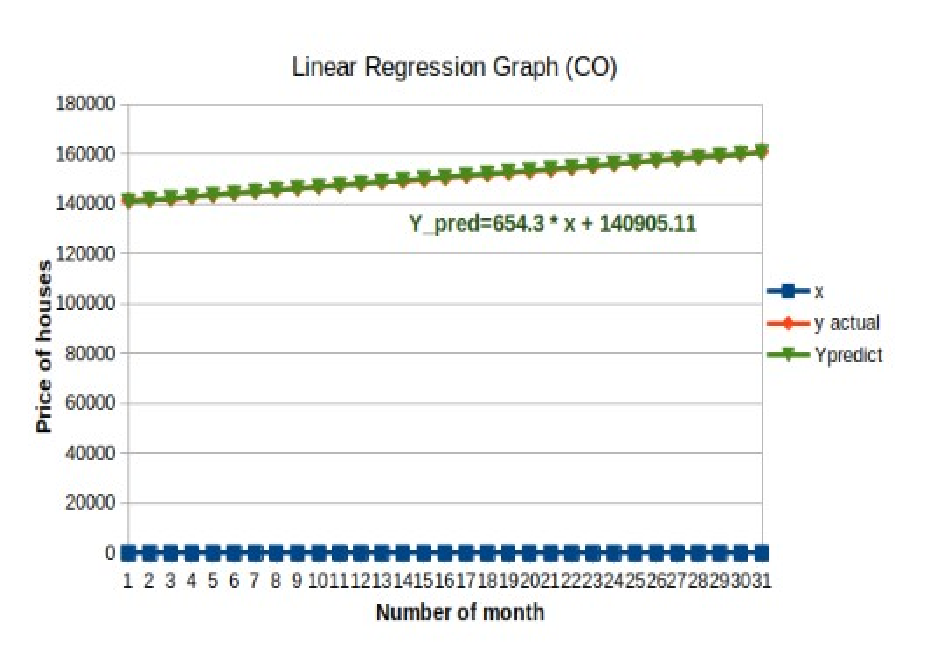


Figure 11: Graph of preliminary hand calculation analysis

TOOLS

- **MATplotlib** has user-friendly plotting toolbox for data visualizations
- **Python3 with Jupyter Notebook** is very enriched with machine and deep learning libraries.
- **NumPy** makes easy complex machine and deep learning numerical operations with large dataset.
- **SciPy** contains different modules for optimization, linear algebra, integration and statistics that is very helpful for data analysis.
- **Scikit-learn**, a machine learning library for Python has various algorithms like support vector machine, random forests, and k-neighbors, and it supports Python numerical and scientific libraries like NumPy and SciPy.

RESULTS (TRAINING)

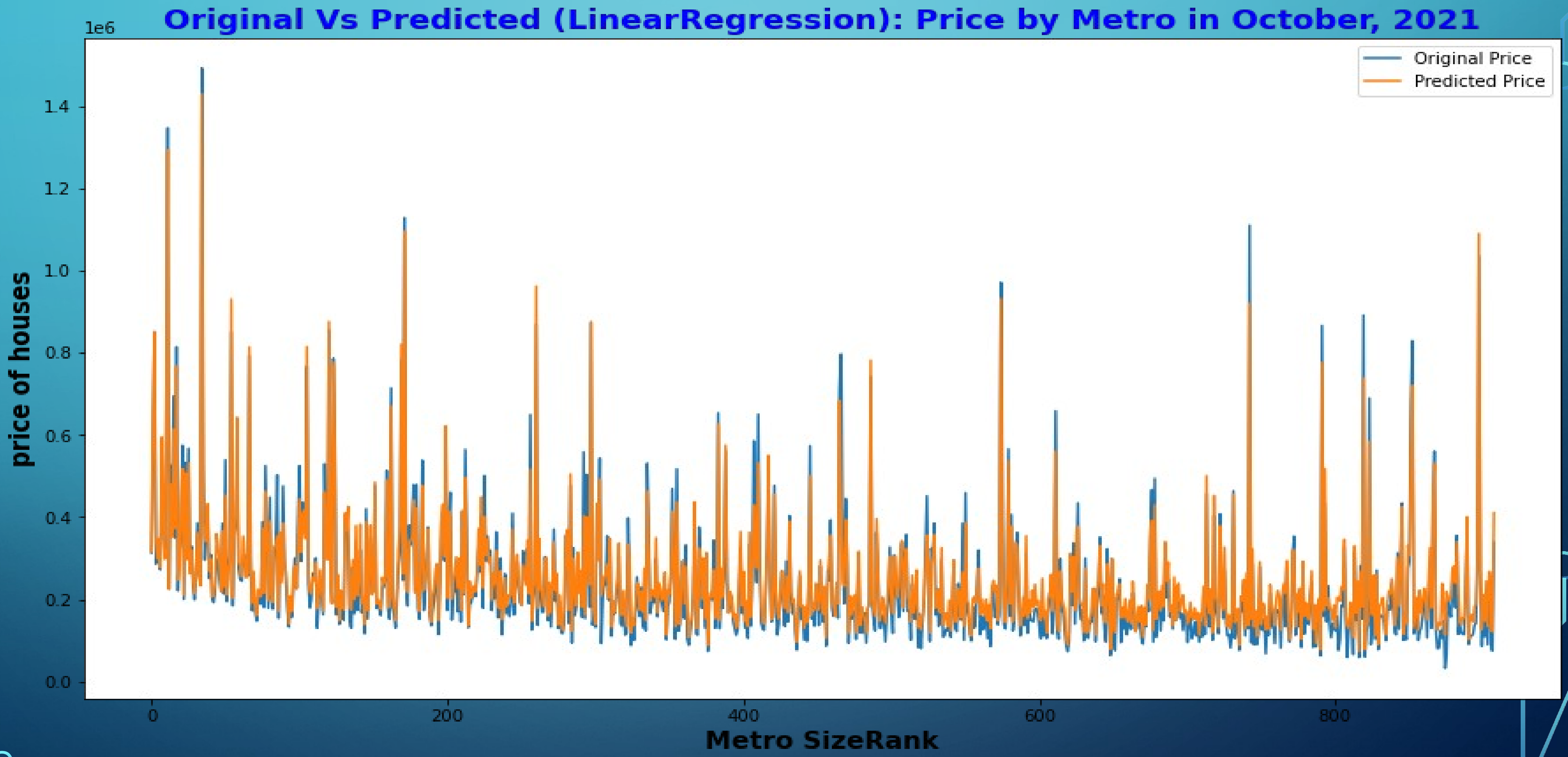


Figure 12: Original vs. Predicted result on Metro-based data Linear regression model

RESULTS (TRAINING)

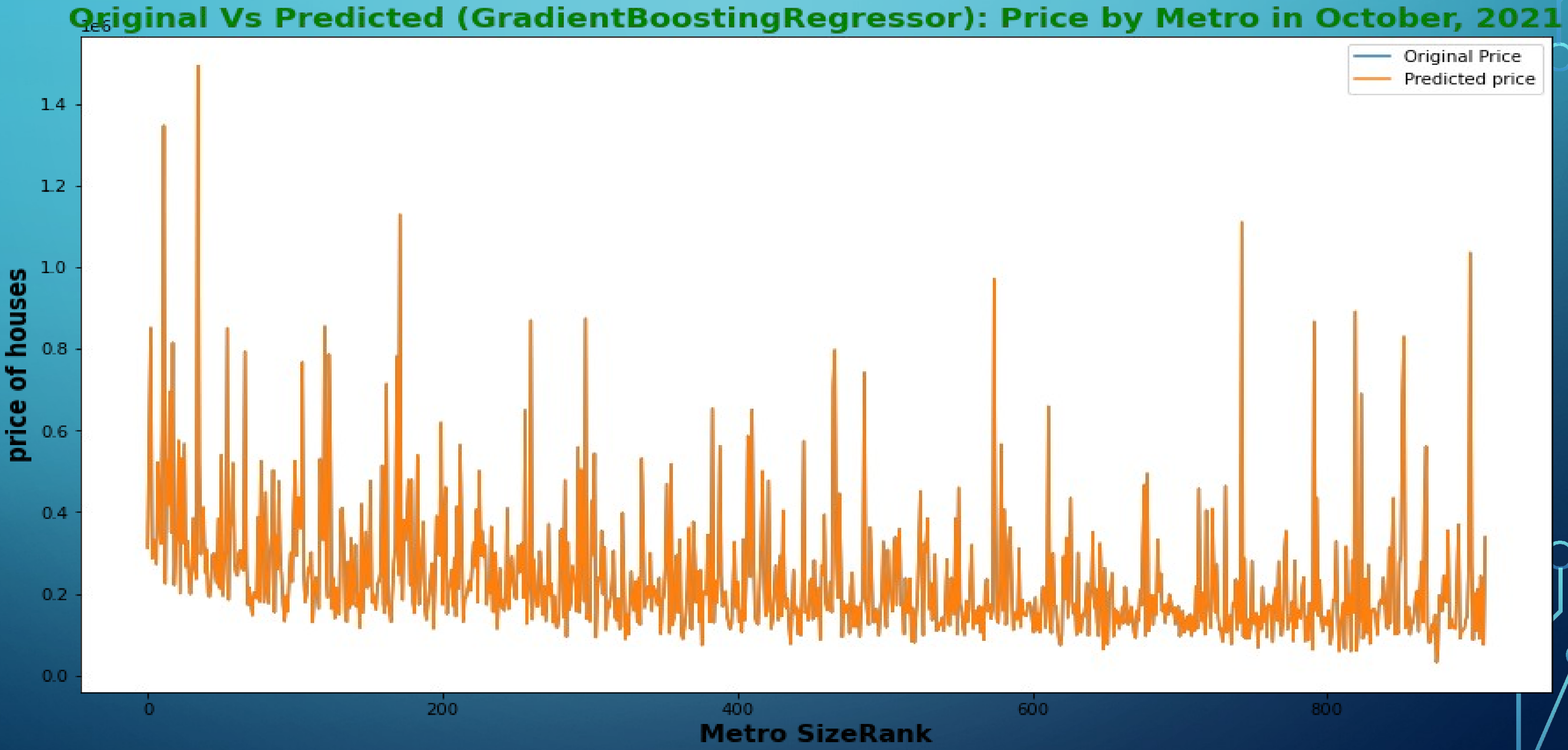


Figure 13: Original vs. Predicted result GradientBoostingregression on metro-based data

RESULTS (TRAINING)

Table : Most and Less expensive house price

Most Expensive		Less Expensive	
Predicted	**Original**	**Predicted**	**Original**
1493019.88	1493020.00	31775.00	31775.00
1346994.95	1346995.00	58553.00	58553.00
1128825.92	1128826.00	58558.00	58558.00
1110212.99	1110213.00	59964.00	59964.00
1035467.98	1035468.00	62956.99	62957.00

RESULTS (TRAINING)

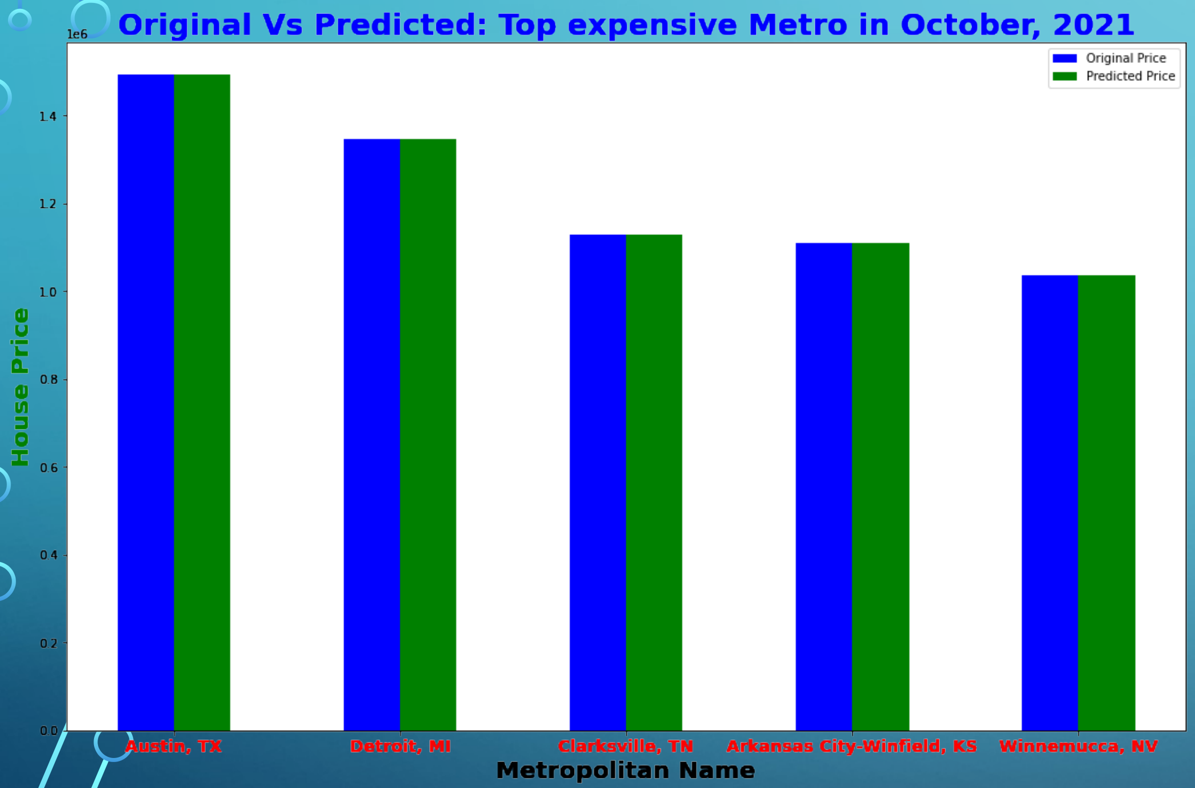


Figure 14: Most expensive metropolitan (predicted vs. original)

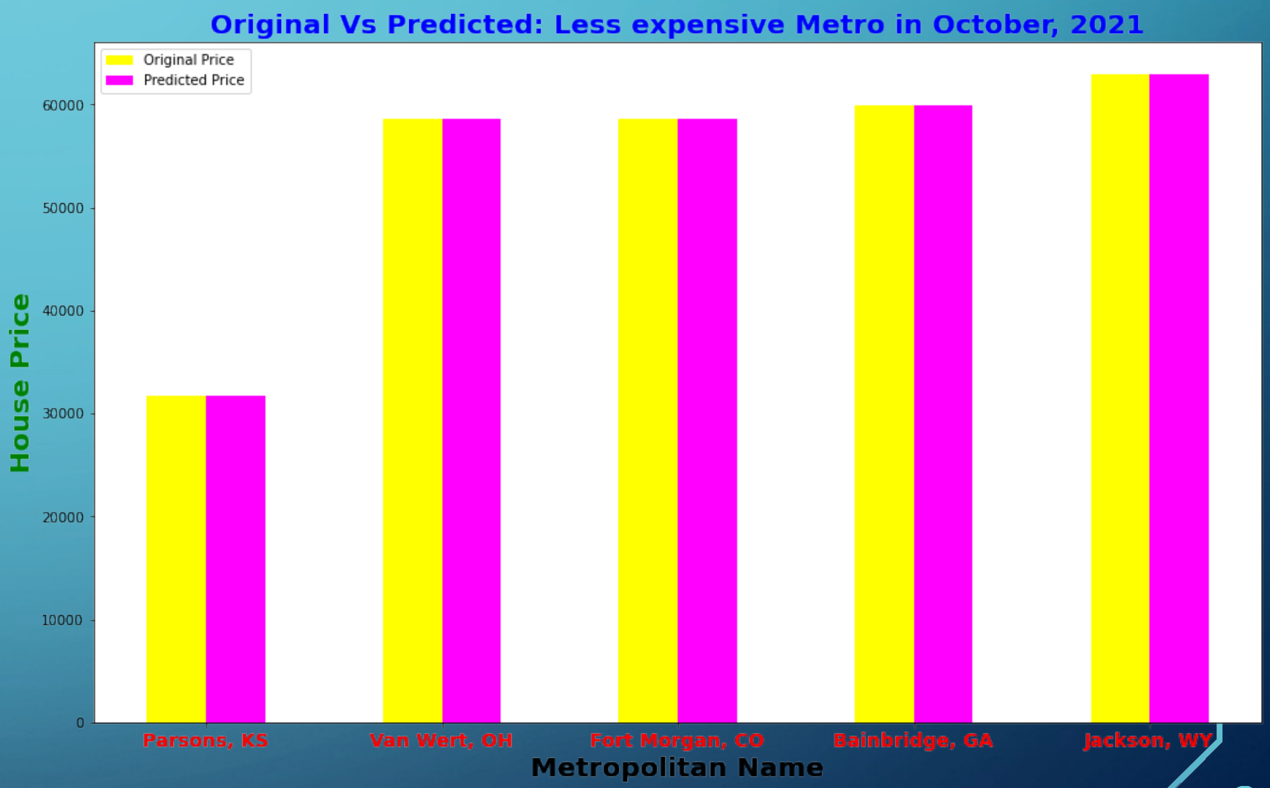


Figure 15: Less expensive metropolitan name (predicted vs. original)

RESULTS (TESTING)

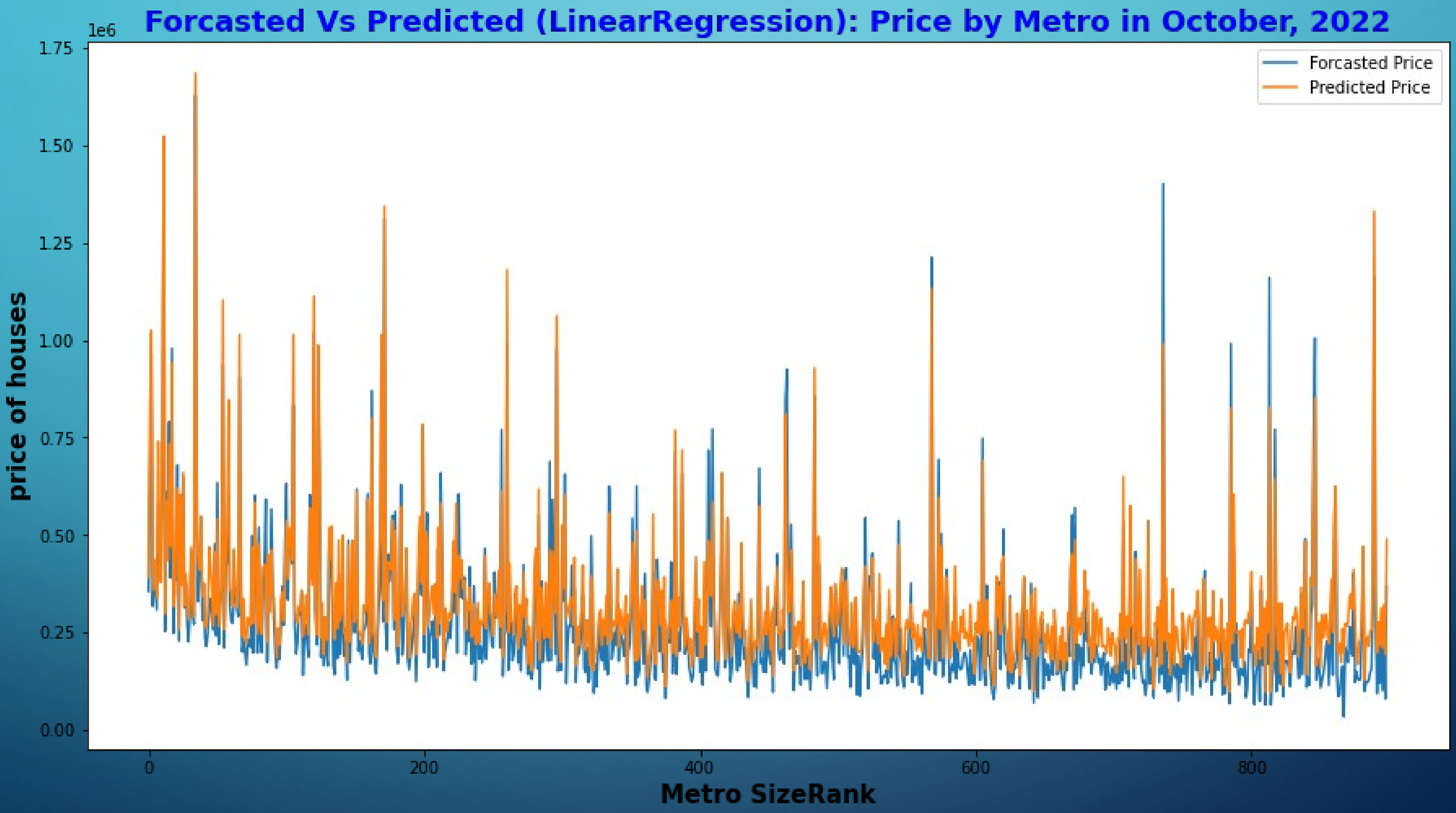


Figure 16: Predicted vs. Original house price on October 2022

RESULTS (TESTING)

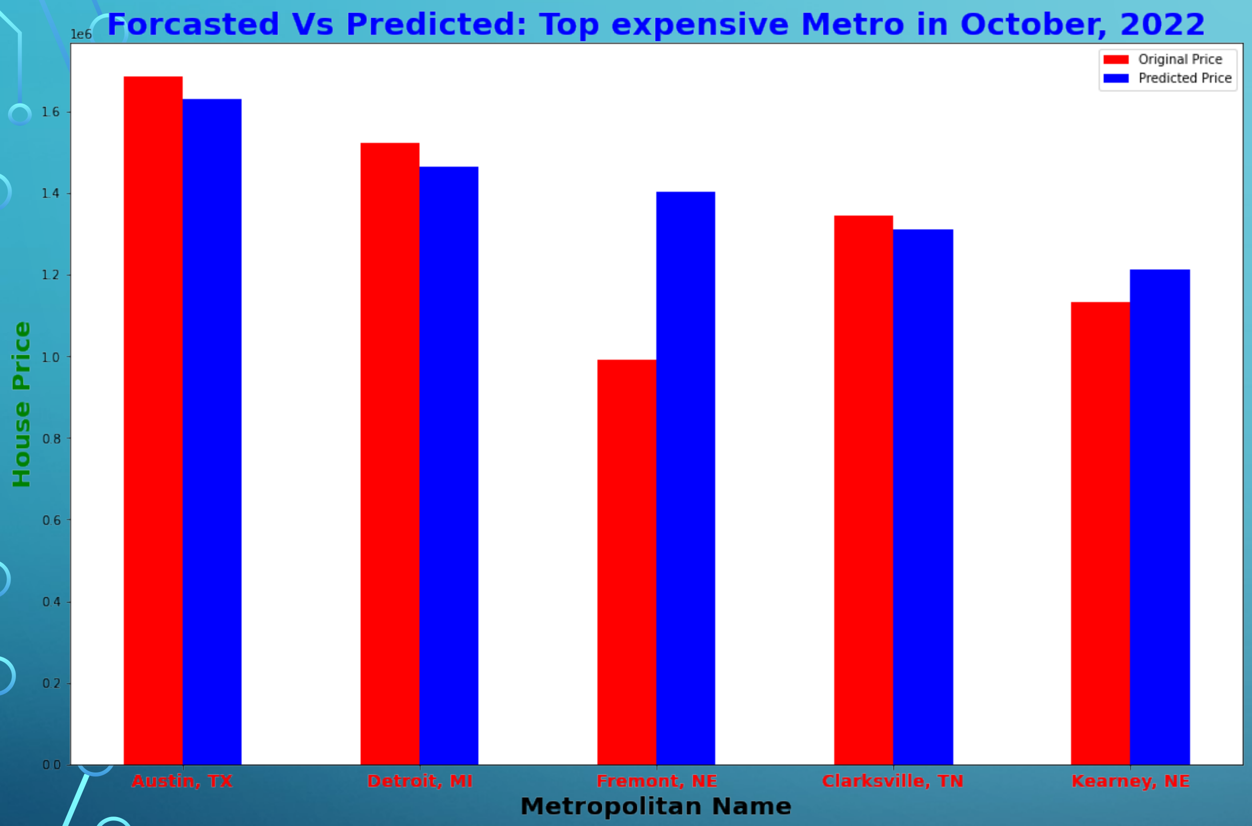
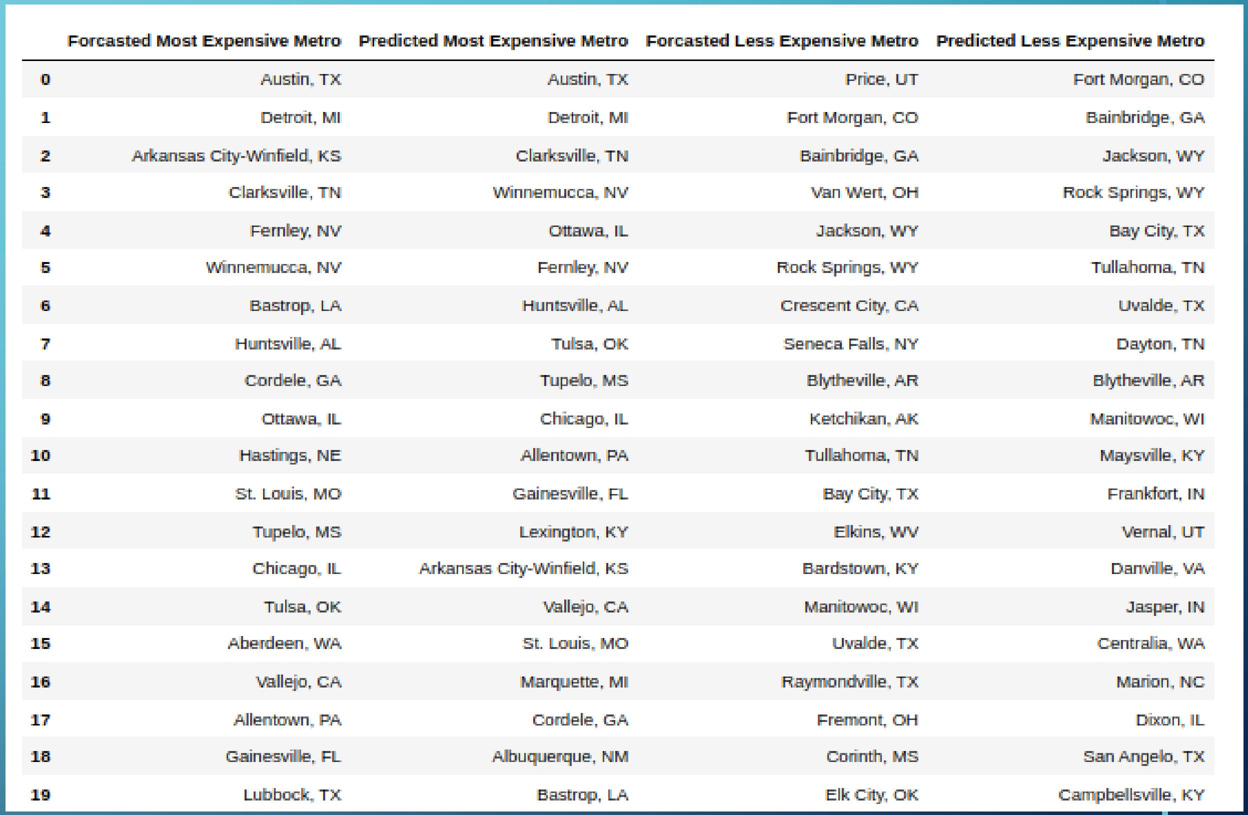


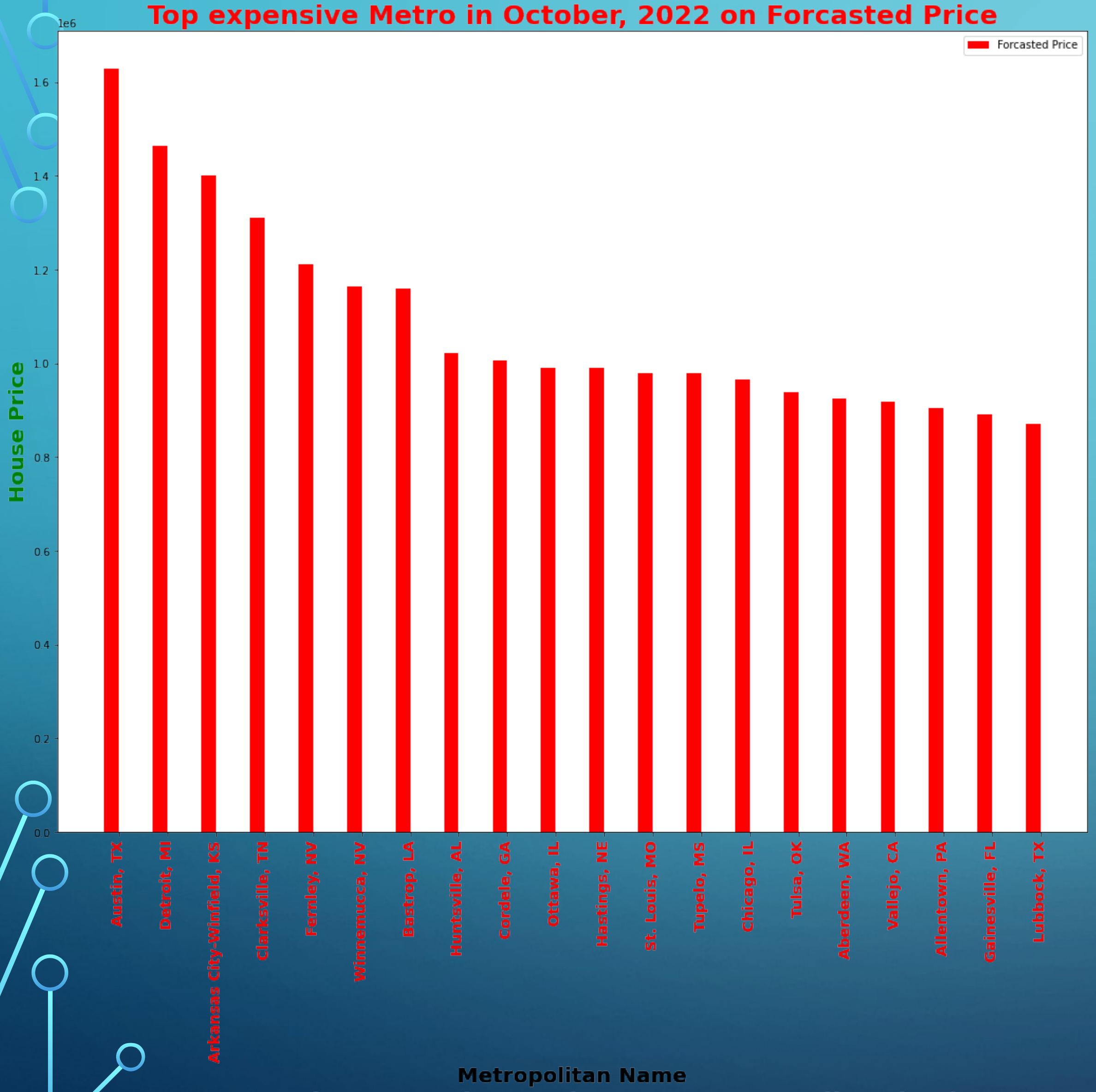
Figure 17: Top 5 metro house price on October, 2022



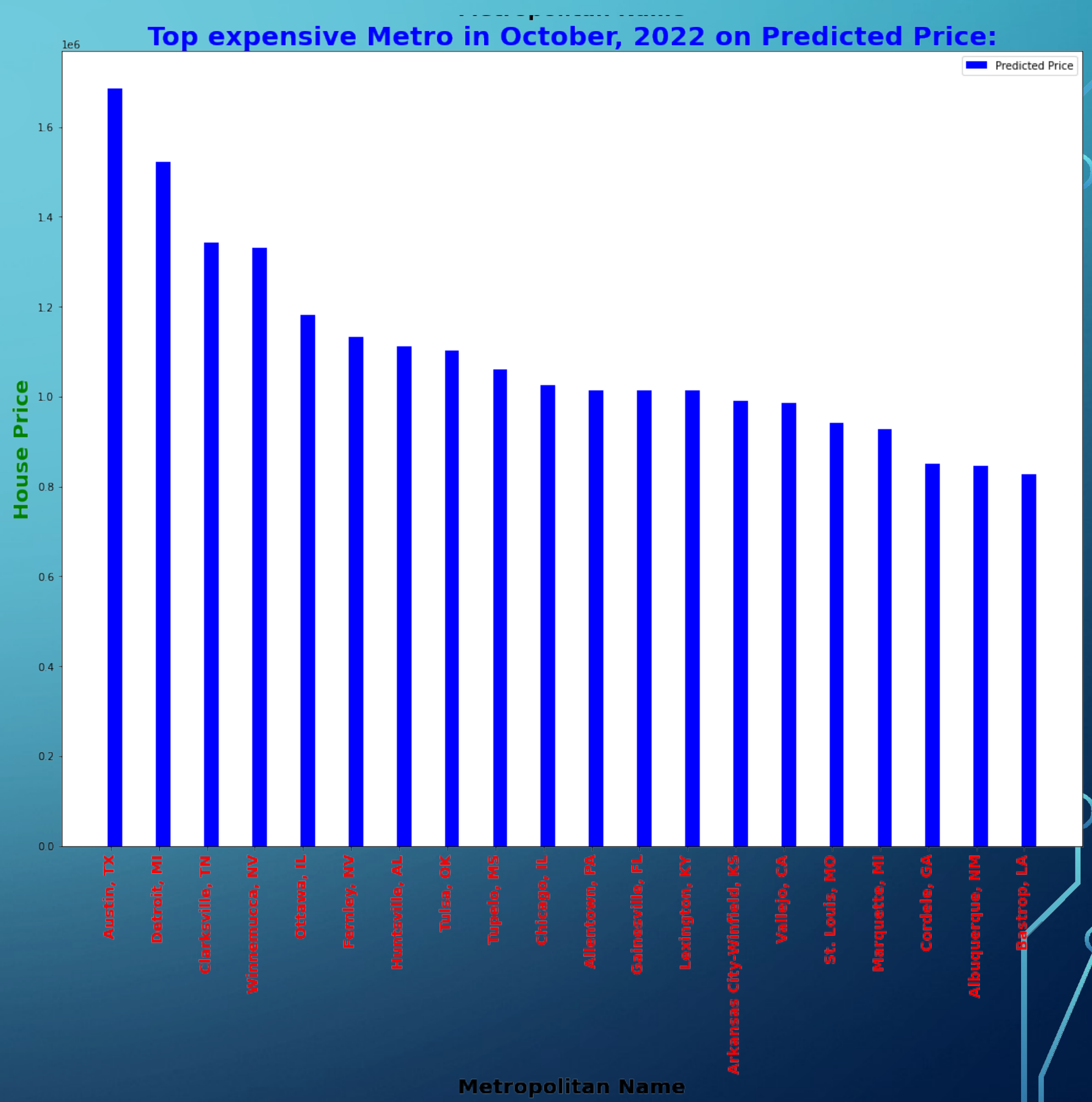
	Forcasted Most Expensive Metro	Predicted Most Expensive Metro	Forcasted Less Expensive Metro	Predicted Less Expensive Metro
0	Austin, TX	Austin, TX	Price, UT	Fort Morgan, CO
1	Detroit, MI	Detroit, MI	Fort Morgan, CO	Bainbridge, GA
2	Arkansas City-Winfield, KS	Clarksville, TN	Bainbridge, GA	Jackson, WY
3	Clarksville, TN	Winnemucca, NV	Van Wert, OH	Rock Springs, WY
4	Fernley, NV	Ottawa, IL	Jackson, WY	Bay City, TX
5	Winnemucca, NV	Fernley, NV	Rock Springs, WY	Tullahoma, TN
6	Bastrop, LA	Huntsville, AL	Crescent City, CA	Uvalde, TX
7	Huntsville, AL	Tulsa, OK	Seneca Falls, NY	Dayton, TN
8	Cordele, GA	Tupelo, MS	Blytheville, AR	Blytheville, AR
9	Ottawa, IL	Chicago, IL	Ketchikan, AK	Manitowoc, WI
10	Hastings, NE	Allentown, PA	Tullahoma, TN	Maysville, KY
11	St. Louis, MO	Gainesville, FL	Bay City, TX	Frankfort, IN
12	Tupelo, MS	Lexington, KY	Elkins, WV	Vernal, UT
13	Chicago, IL	Arkansas City-Winfield, KS	Bardstown, KY	Danville, VA
14	Tulsa, OK	Vallejo, CA	Manitowoc, WI	Jasper, IN
15	Aberdeen, WA	St. Louis, MO	Uvalde, TX	Centralia, WA
16	Vallejo, CA	Marquette, MI	Raymondville, TX	Marion, NC
17	Allentown, PA	Cordele, GA	Fremont, OH	Dixon, IL
18	Gainesville, FL	Albuquerque, NM	Corinth, MS	San Angelo, TX
19	Lubbock, TX	Bastrop, LA	Elk City, OK	Campbellsville, KY

Figure 18: List of Top20 metro name from prediction and forecasting

RESULTS (TESTING)



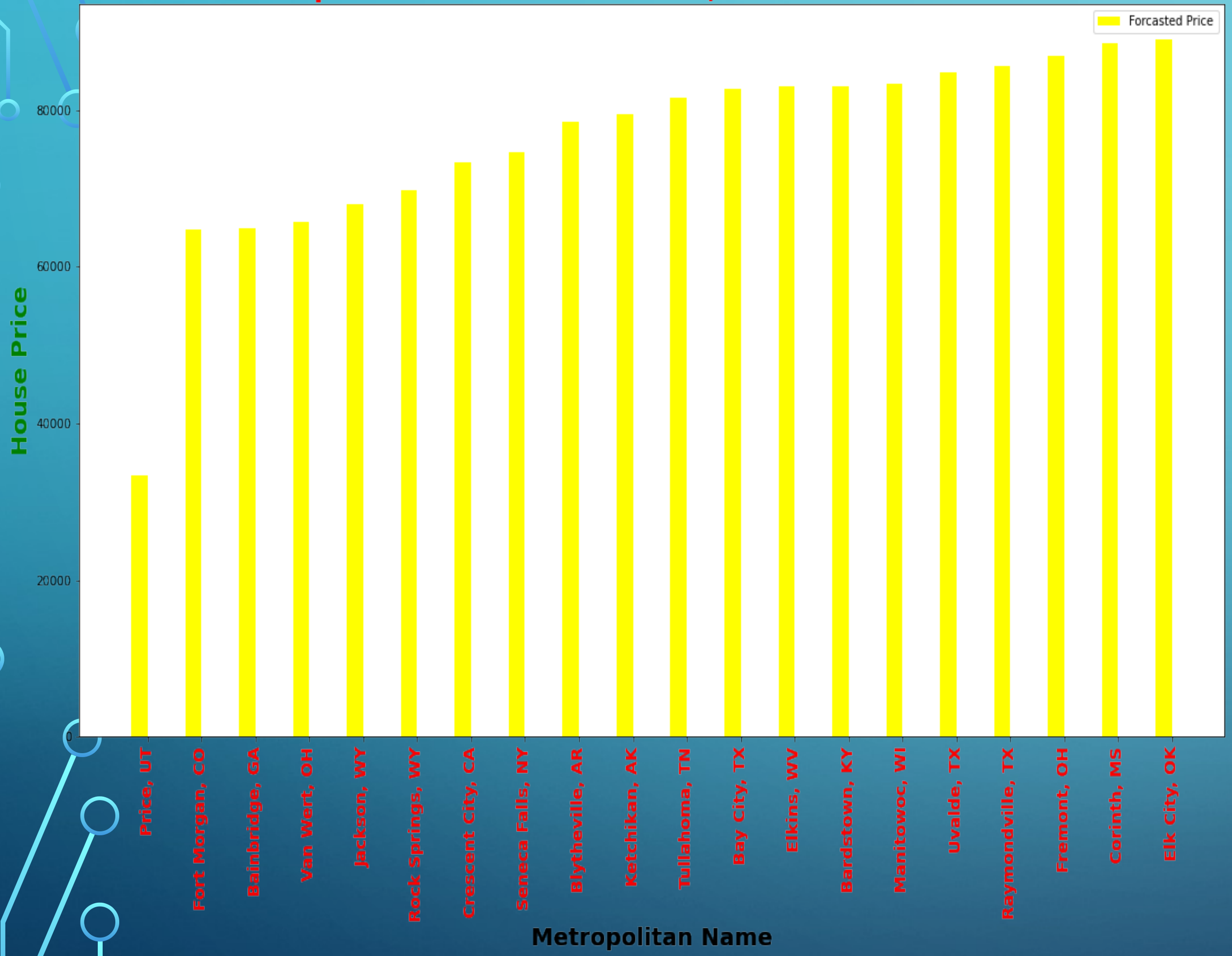
Top expensive Metro in October, 2022 on Forcasted Price



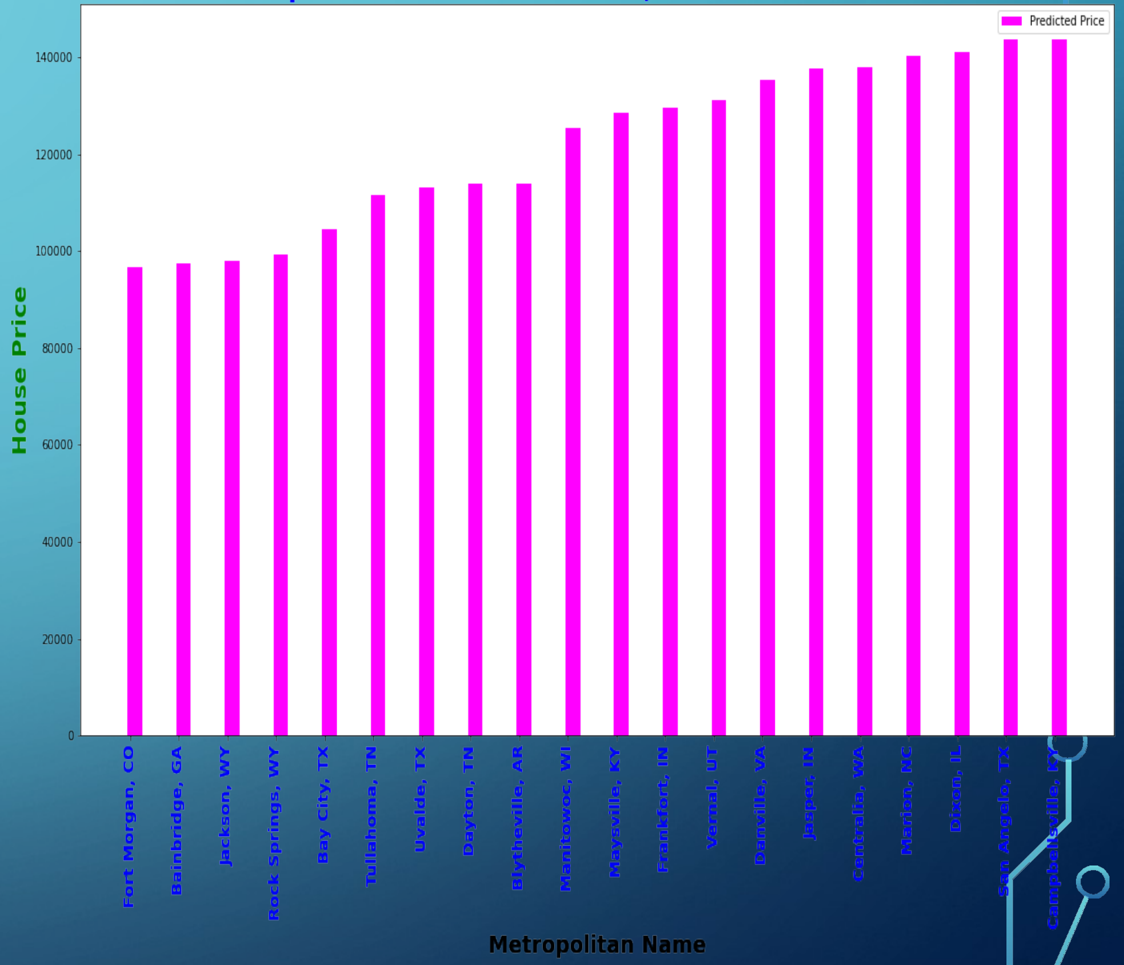
Top expensive Metro in October, 2022 on Predicted Price:

RESULTS (TESTING)

Less expensive Metro in October, 2022 on Forcasted Price



Less expensive Metro in October, 2022 on Predicted Price:



Lessons learned

- To develop a Linear model for any prediction
- Data preprocessing and preparing for model fitting
- Data visualization in different aspects
- Various data mining techniques

CHALLENGES AND FUTURE PLAN

- Accuracy is only 45% for 1st model. We want to increase it.

- Accuracy for 2nd model is above 99%. However, testing accuracy is not good. We want to develop a strategy to figure out it.

- We will work for best suggestion for people with reasonable price and excellency on other factors

THANK YOU!

References

[1] “Housing data,” Zillow Research, 25-Mar-2021. [Online]. Available: https://www.zillow.com/research/data/. [Accessed: 14-Sep-2021].

[2] N. Bhagat, A. Mohokar, and S. Mane, “House price forecasting using data mining,” International Journal of Computer Applications, vol. 152, no. 2, pp. 23–26, 2016.

[3] V. Valkov, “Predicting house prices with linear Regression: Machine learning from SCRATCH (PART II),” Medium, 05-Jul-2019. [Online]. Available: https://towardsdatascience.com/predicting-house-prices-with-linear-regression-machine-learning-from-scratch-part-ii-47a0238aeac1. [Accessed: 14-Sep-2021].

[5] T. Gupta, “Data preprocessing in Python,”Medium, 25-Dec-2020. [Online]. Available: https://towardsdatascience.com/data-preprocessing-in-python-b52b652e37d5. [Accessed: 14-Sep-2021].

[6] Online source. Link: https://github.com/jaskirat111/Housing-Price-Prediction-using-Advanced-ML-Algorithms

[7] Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, *174*, 433-442.