Predicting House Prices in King County, USA

Akinbanji Akinyera
06-05-2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of all methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis results
  - Predictive Analysis results

Introduction

- The objective of the project is to determine the market price of a house in King county, USA given a set of features. This is achieved by analyzing and predicting housing prices using attributes or features such square footage, number of bedrooms, number of floors etc.
- Audience is Real Estate Investment Trust board

Methodology

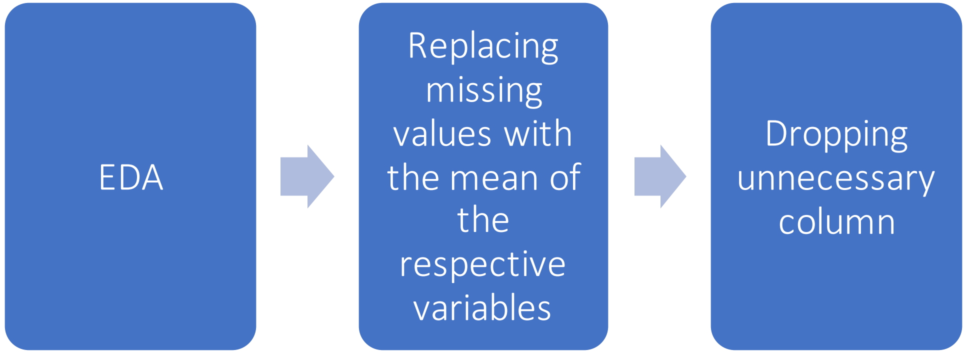
- Data Collection:
- Perform data wrangling
- Perform Exploratory Data Analysis
    - Coefficients of the variables in relation to Price are determined
- Perform predictive analysis using regression models

Data Collection

- Data was extracted using Pandas
- Source: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DA0101EN-SkillsNetwork/labs/FinalModule_Coursera/data/kc_house_data_NaN.csv

Data Wrangling

- Exploratory Data Analysis was performed on the dataset initially
- Missing values were dealt with
- Redundant variable was dropped.



EDA

Replacing missing values with the mean of the respective variables

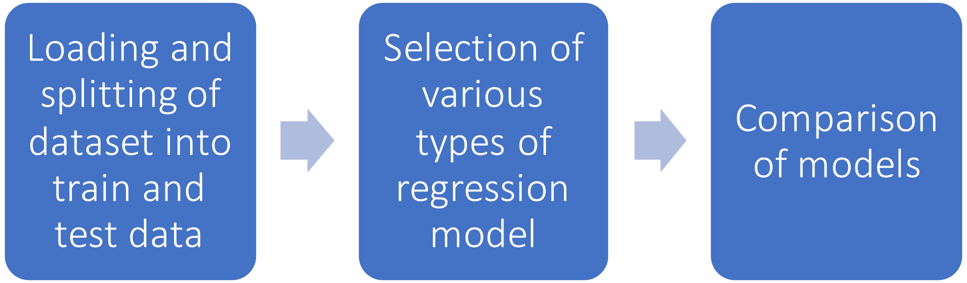
Dropping unnecessary column

Exploratory Data Analysis

- Exploratory data analysis was performed by determining the coefficients of the variables in relation to price

Predictive Analysis (Regression)

- The dataset was loaded and split into train and test data
- Simple Linear, Multivariate Linear and Multivariate Polynomial regression models were selected. The models were trained using the train dataset
- Models were compared based on R square and Mean Squared Error (MSE) values and the best model was eventually chosen



Loading and splitting of dataset into train and test data

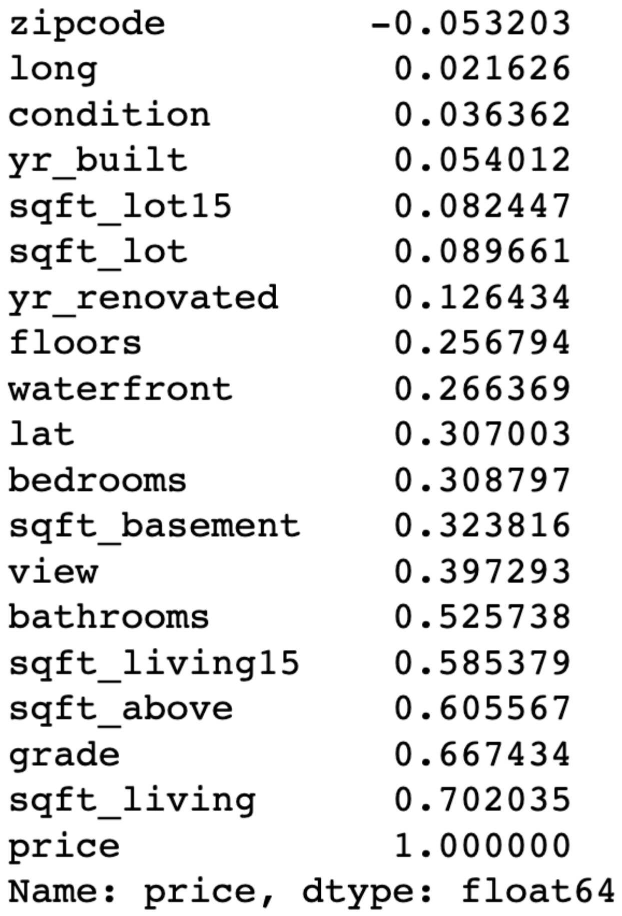
Selection of various types of regression model

Comparison of models

Results

- Exploratory data analysis result
- Predictive analysis results

Exploratory Data Analysis Result

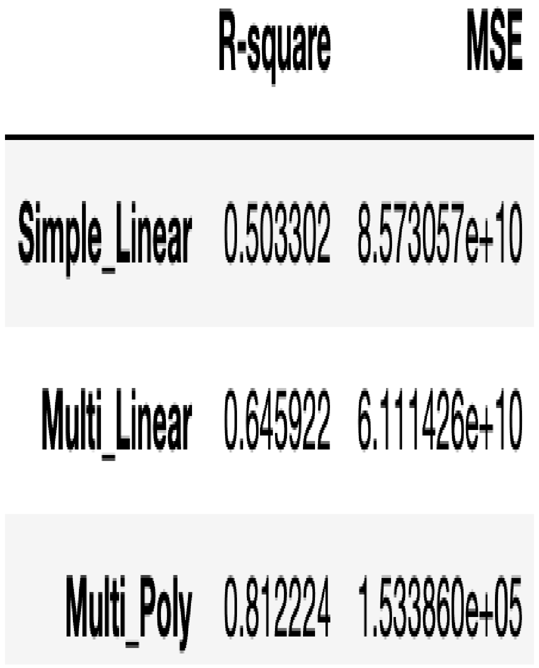


```
zipcode          -0.053203
long              0.021626
condition         0.036362
yr_built          0.054012
sqft_lot15        0.082447
sqft_lot          0.089661
yr_renovated      0.126434
floors            0.256794
waterfront        0.266369
lat               0.307003
bedrooms          0.308797
sqft_basement     0.323816
view              0.397293
bathrooms         0.525738
sqft_living15     0.585379
sqft_above        0.605567
grade             0.667434
sqft_living       0.702035
price             1.000000
Name: price, dtype: float64
```

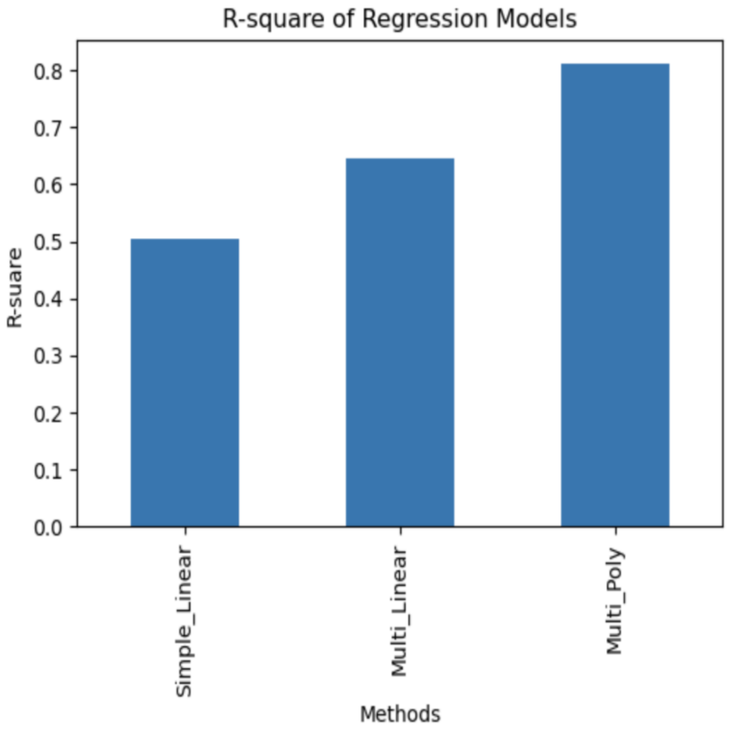
From the result, it can be seen than sqft_living is mostly correlated with price while floors, waterfront, lat, bedrooms, sqft_basement, view, bathrooms, sqft_living15, sqft_above and grade are somewhat correlated

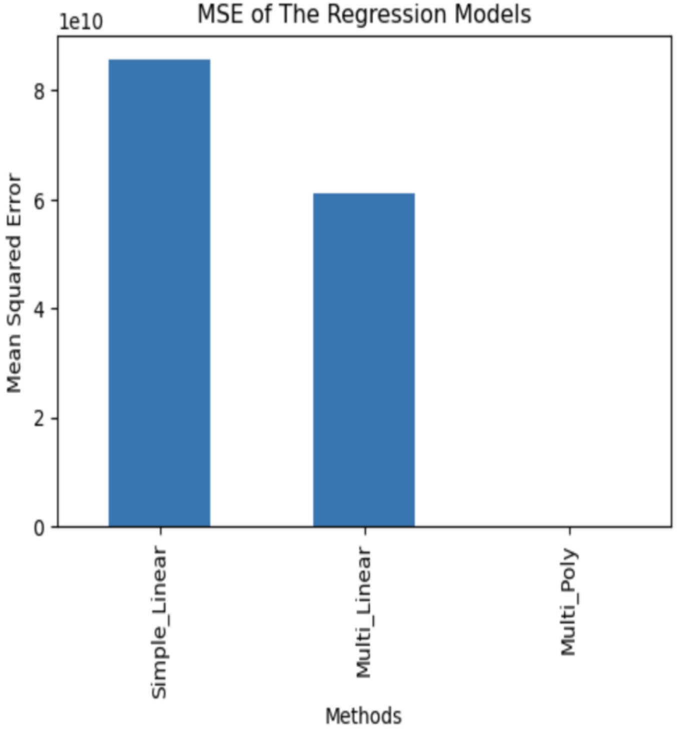
Predictive Analysis Result (Regression)

Regression Accuracy



	R-square	MSE
Simple_Linear	0.503302	8.573057e+10
Multi_Linear	0.645922	6.111426e+10
Multi_Poly	0.812224	1.533860e+05

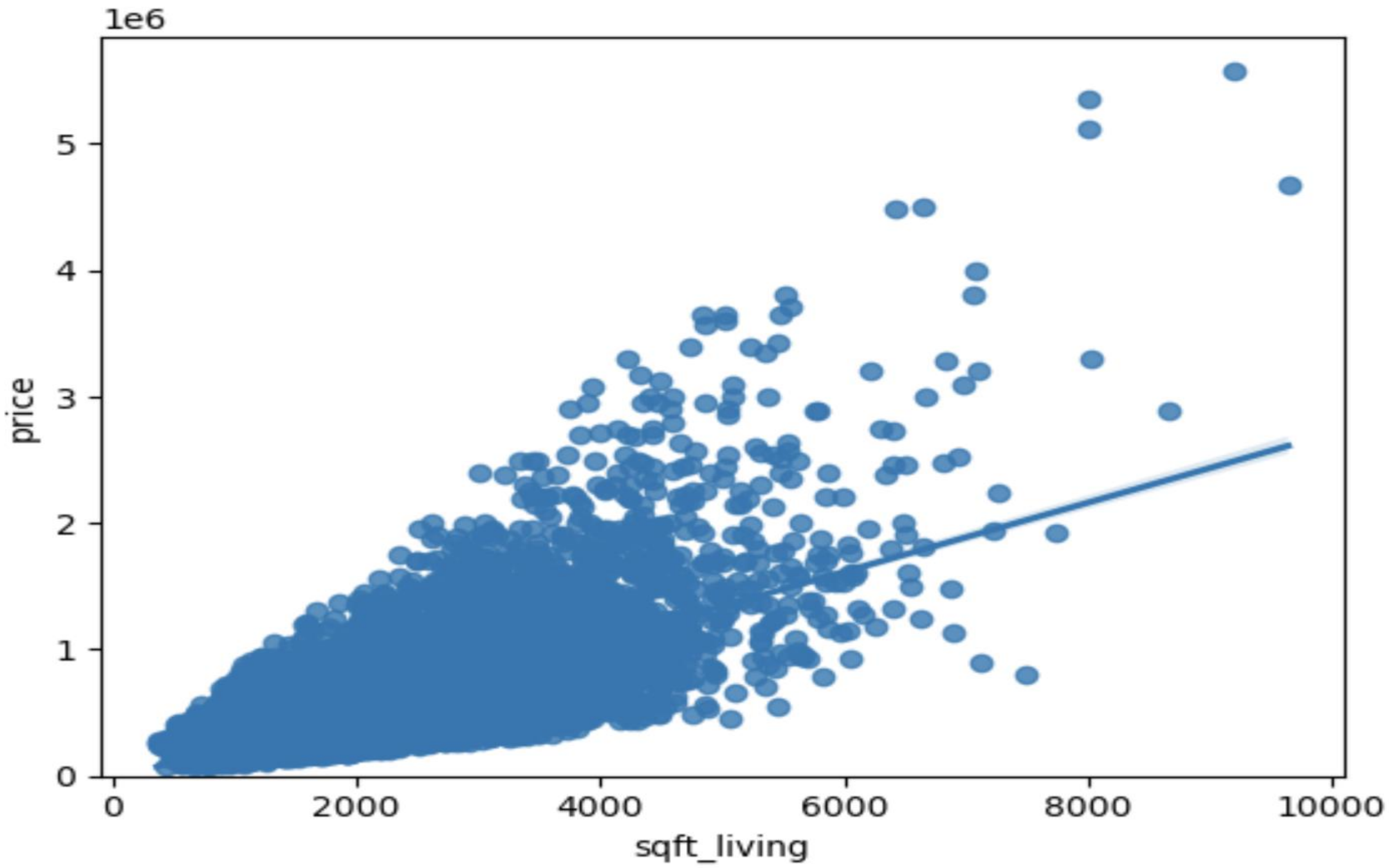


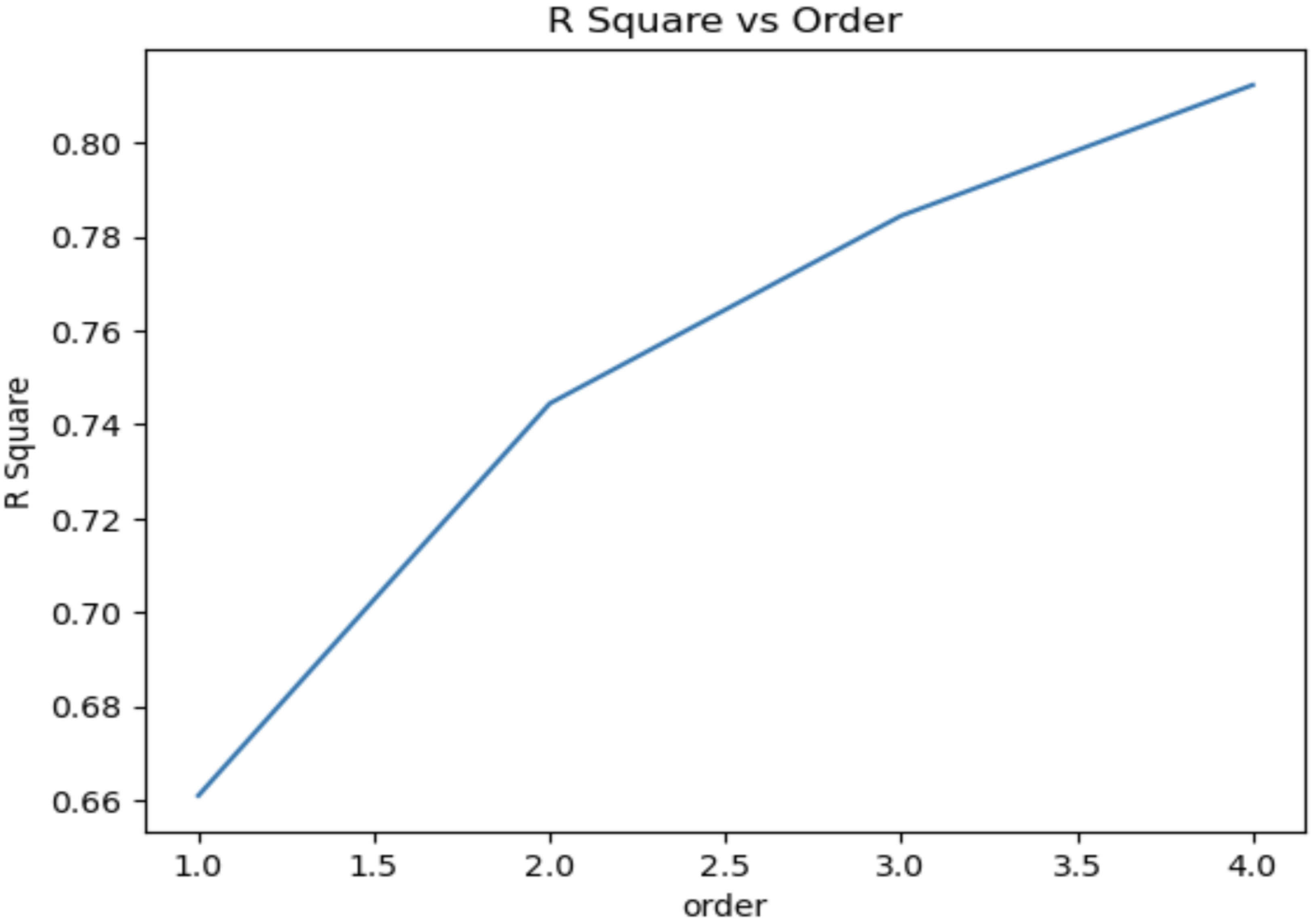


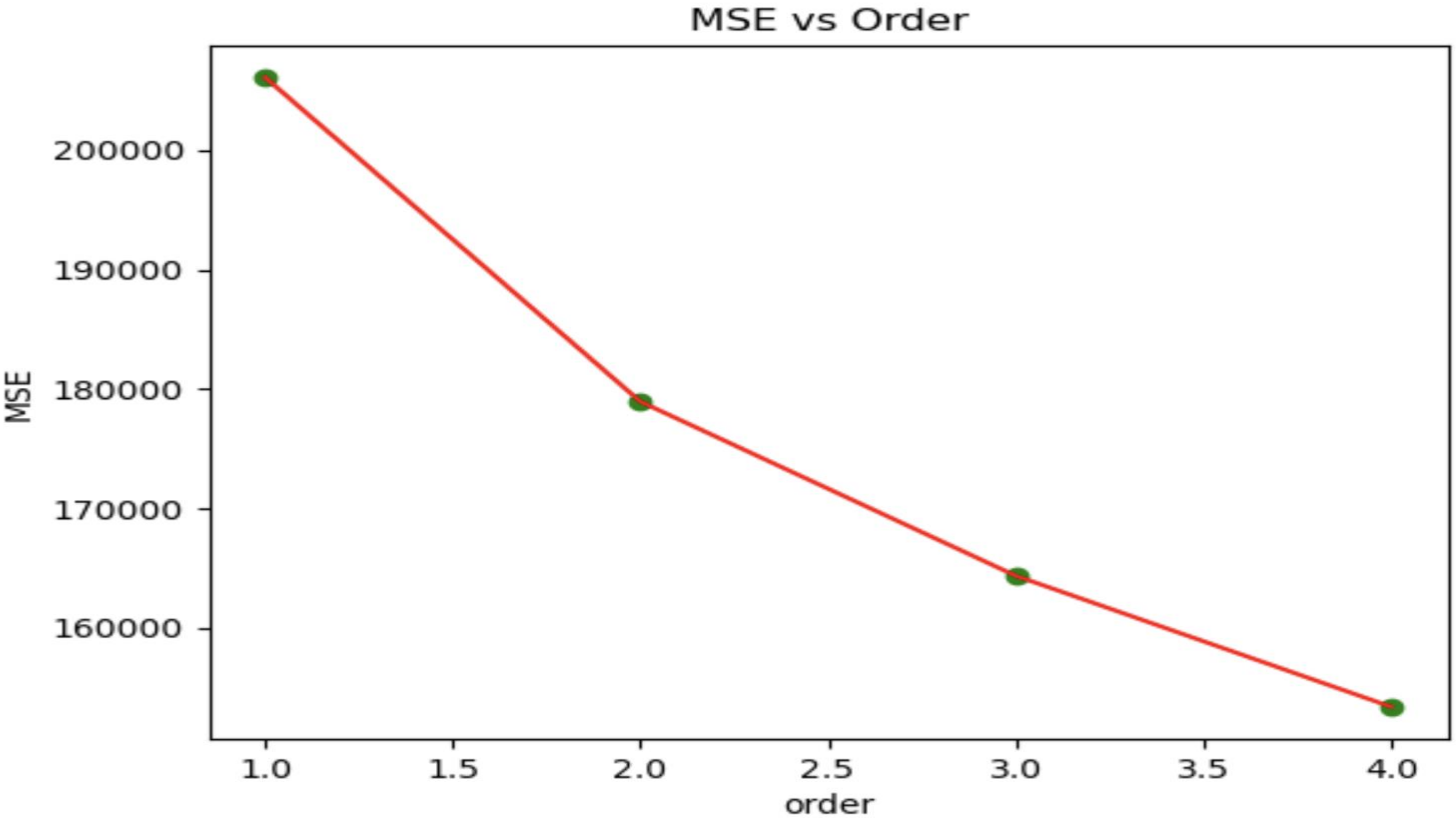
Conclusion

- Multivariate Polynomial Regression has the highest r square value of 0.812224 and the lowest mean squared error of 153386. Therefore, it is the preferred model
- Sqft_living has the most effect in determining the market price of a house in King county, USA
- Floors, waterfront, bedrooms, lat, sqft_basement, view, bathrooms, sqft_living15, sqft_above and grade are somewhat determinants of price as well.

APPENDIX


1e6
price
0
1
2
3
4
5
0
2000
4000
6000
8000
10000
sqft_living


R Square vs Order
R Square
0.66
0.68
0.70
0.72
0.74
0.76
0.78
0.80
1.0
1.5
2.0
2.5
3.0
3.5
4.0
order

MSE vs Order
MSE
200000
190000
180000
170000
160000
order
1.0
1.5
2.0
2.5
3.0
3.5
4.0