Housing Market Prediction in King County, USA

A Data Analysis for Real Estate Investment Trust

Outline

- Introduction
- Tools Used for Data Analysis
- Modules of Data Analysis
- Conclusion
- Recommendations

Introduction

- **Background**
  - The Trust intends to invest in the residential real estate.

- **Purpose**
  - Determine the market price of a house.
  - Analyze and predict housing prices using attributes or feature.

- **Tools used**
  - Jupiter Notebook
  - Python
  - Python libraries:
    - Pandas
    - Matplotlib
    - Numpy
    - Seaborn
    - Scikit-learn
  - PowerPoint

Introduction

- **Modules**
  1. Data preprocessing
  2. Data wrangling
  3. Exploratory Data Analysis
  4. Model Development
  5. Model Evaluation and Refinement
- **Conclusion**
  - Summarize the results and key patterns
- **Recommendation**
  - Suggestions for Real Estate investments

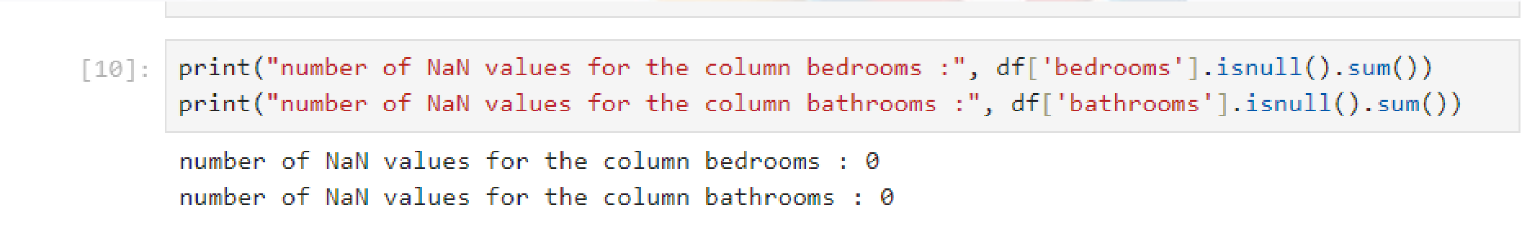
Module 1: Data Preprocessing

- Import the dataset
- Data includes information of homes sold in King County between May 2014 and May 2015.
- Attributes of dataset →
- Generate a statistical summary of the dataframe

Variable	**Description**
id	A notation for a house
date	Date house was sold
price	Price is prediction target
bedrooms	Number of bedrooms
bathrooms	Number of bathrooms
sqft_living	Square footage of the home
sqft_lot	Square footage of the lot
floors	Total floors (levels) in house
waterfront	House which has a view to a waterfront
view	Has been viewed
condition	How good the condition is overall
grade	overall grade given to the housing unit, based on King County grading system
sqft_above	Square footage of house apart from basement
sqft_basement	Square footage of the basement
yr_built	Built Year
yr_renovated	Year when house was renovated
zipcode	Zip code
lat	Latitude coordinate
long	Longitude coordinate
sqft_living15	Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area
sqft_lot15	LotSize area in 2015(implies-- some renovations)

Module 2: Data wrangling

- Remove redundant columns
- Find missing values
- Replace missing values
- Check if there is a missing value

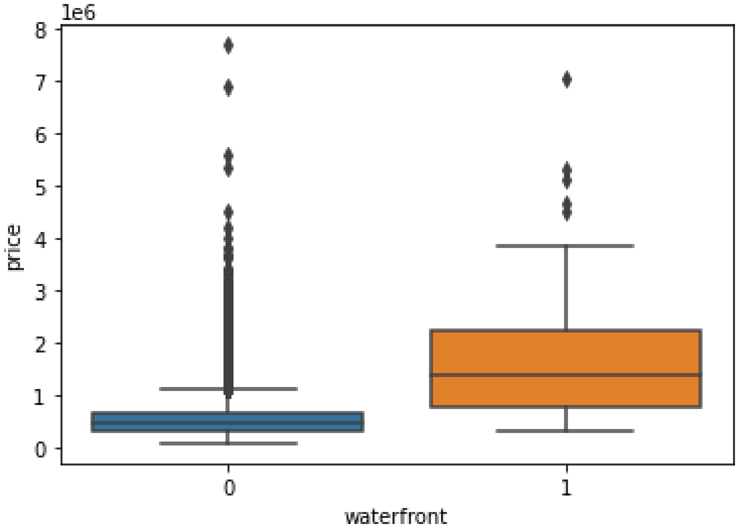


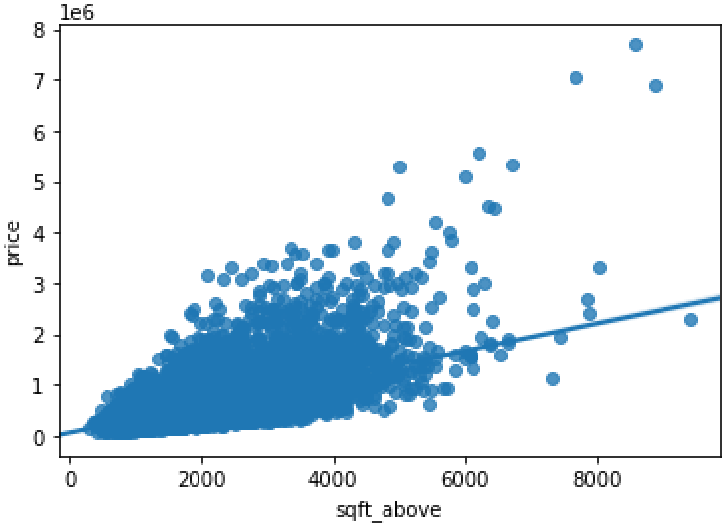
```
[10]: print("number of NaN values for the column bedrooms :", df['bedrooms'].isnull().sum())
      print("number of NaN values for the column bathrooms :", df['bathrooms'].isnull().sum())
```

```
number of NaN values for the column bedrooms : 0
number of NaN values for the column bathrooms : 0
```

Module 3: Exploratory Data Analysis

- Count the number of houses with unique floor values
- **Boxplot**: Determine whether houses with a waterfront view or without a waterfront view have more price outliers.
- **Regplot**: Determine if the area is negatively or positively correlated with price.





Module 4: Model development

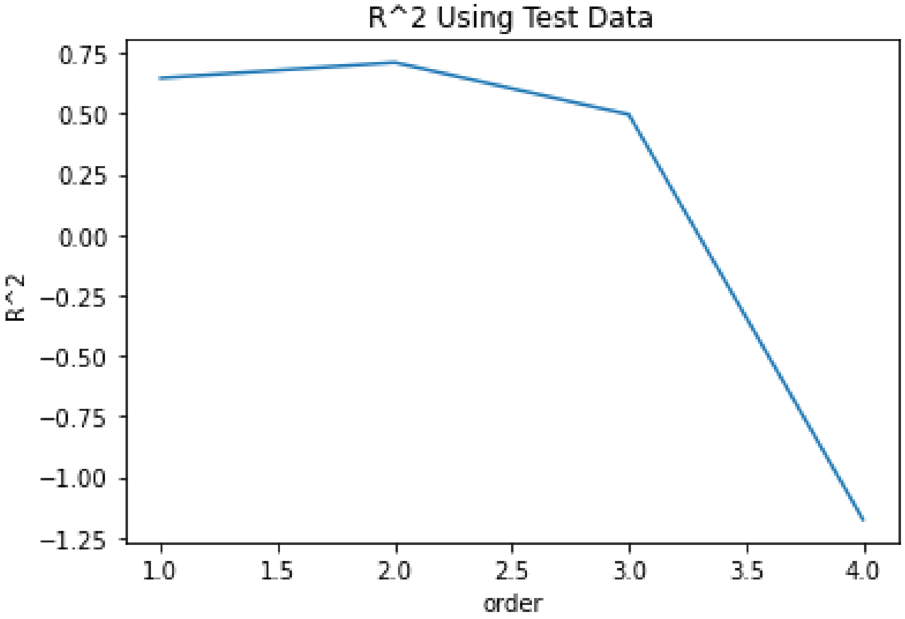
- Fit a linear regression model
- Predict the 'price' using the feature 'sqft_living’
- Predict the 'price' using multiple features:
  - Floors
  - Waterfront
  - Lat
  - Bedrooms
  - Sqft_basement
  - View
  - Bathrooms
  - ...

- Calculate R^2 for both predictions

R^2	‘price’
‘long’	0.00047
‘sqft_living’	0.49285
‘floors’, ‘waterfront’, ‘lat’, ‘bedrooms’, ‘sqft_basement’ ‘view’, ‘bathrooms’, …	0.65766

Module 3: Model Evaluation and Refinement

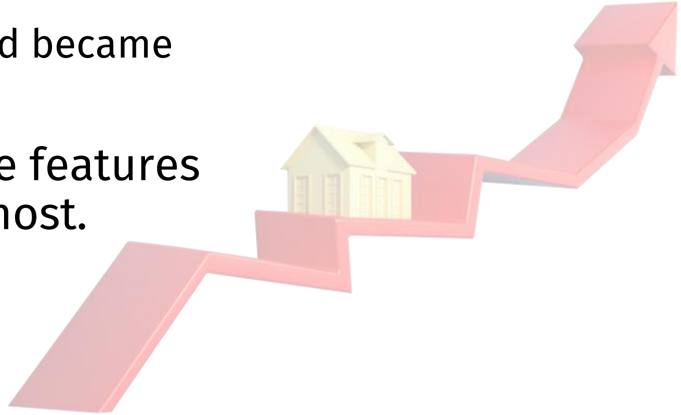
- Split the data into training and testing sets
- Fit a Ridge regression object using the training data
  - set the regularization parameter to 0.1
- Calculate the R^2 utilising the test data provide



The graph shows how the R^2 changes on the test data for different order polynomials.

Conclusion

- This project used linear regression and polynomial regression to
    - Analyzed the relationship between the independent variable (features of data) and dependent variable (price)
- According to the R-squared scores, the house price is influenced by multiple features including the size, the number of bathrooms, and the location.
    - When the model added more features, the R-squared became higher
- There is a need for developing models to test those features and find the ones affecting the housing price the most.



Recommendation

- Stakeholders should diversity the Real Estate due to the reason that the price is affected by many factors.
  - Invests in different types of property to cater for different customers
- The results have some limitations:
  - The dateset only contains one-year data from 2014-2015, may not be accurate to predict the distant future.
  - There are some factors that are not included in this prediction such as safety index of the district, proximity to amenities, school and hospital. These could be used to further improve the prediction model.