Iron Regression Quest

David Arques
May 23rd

Project Overview 📋

In this project, we are given a dataset of **House sale prices in Seattle**, US and our **Goal** is to:

1. **Analyze and clean** the data to apply in machine learning
2. **Apply** different **supervised regression machine learning** models.
3. **Assess the results** and choose the best model to deploy.

Dataset 💾

Description: House sale prices in Seattle (King County), US.

Timeframe: May 2014 to May 2015

Size: (21613 rows, 21columns)

Columns:

- **id:** Unique identifier for each house.
- **date:** Date of house sale.
- **bedrooms:** Number of bedrooms.
- **bathrooms:** Number of bathrooms per bedroom.
- **sqft_living:** Interior living space area.
- **sqft_lot:** Land space area.
- **floors:** Number of house floors.
- **waterfront:** Presence of waterfront view.
- **view:** Number of house viewings.
- **condition:** Overall house condition.
- **grade:** Overall grade based on King County grading system.
- **sqft_above:** Area excluding the basement.
- **sqft_basement:** Basement area.
- **yr_built:** Year of house construction.
- **yr_renovated:** Year of house renovation.
- **zipcode:** ZIP code area.
- **lat:** Latitude coordinate.
- **long:** Longitude coordinate.
- **sqft_living15:** Interior living space of nearest 15 neighbors in 2015.
- **sqft_lot15:** Land space of nearest 15 neighbors in 2015.
- **TARGET > price:** Sale price of the house (prediction target).

Data Cleaning ✨

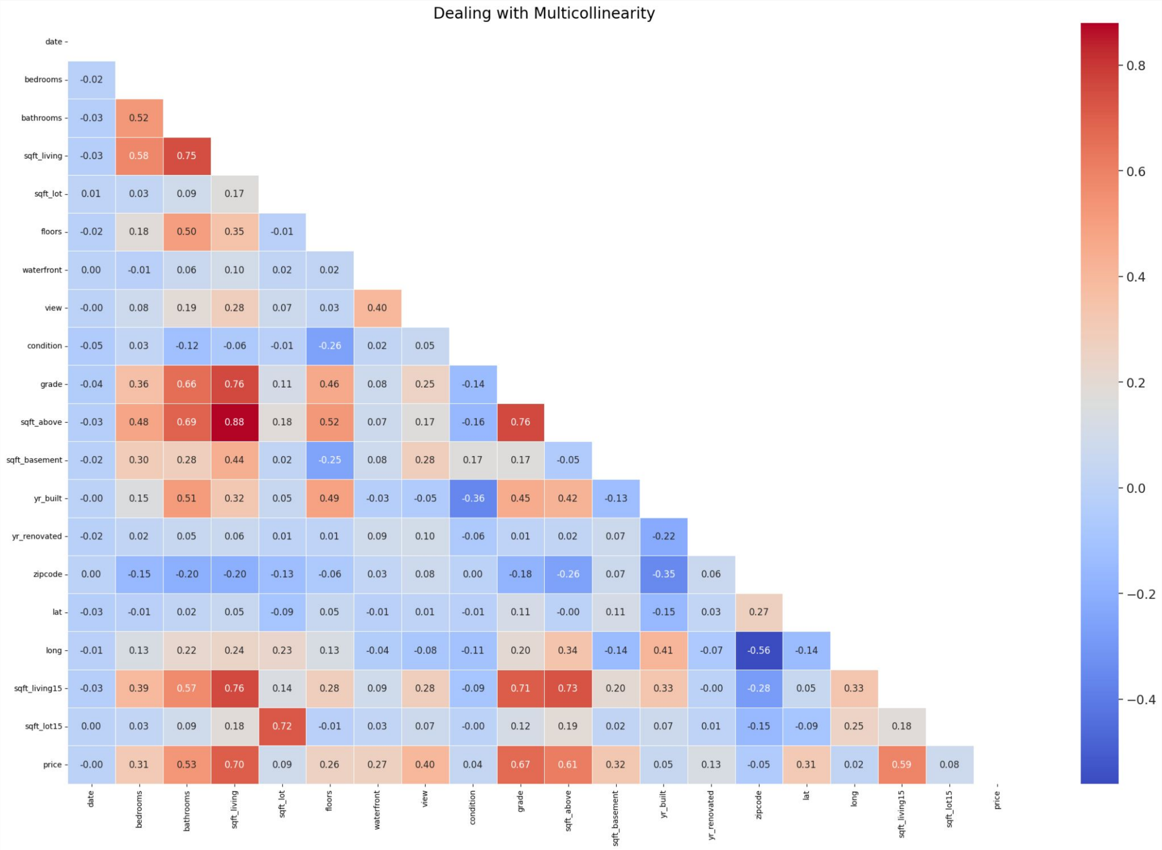
Actions:

- Using house **id**'s as **index**
- Changing dates to **datetime**
- **All columns** are already **numerical**
- **No null** values
- Moving **[price]** to the right as **target**

Heatmap Conclusions:

The **highest correlations** are with **sqft_living** and **sqft_above**, but both are **below 0.8 with** the target **price**.

Not removing any column due to low correlation over target [Price].

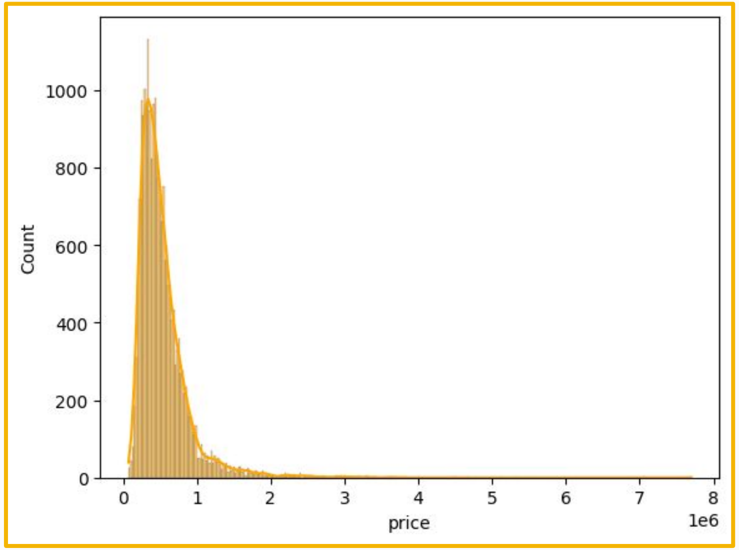


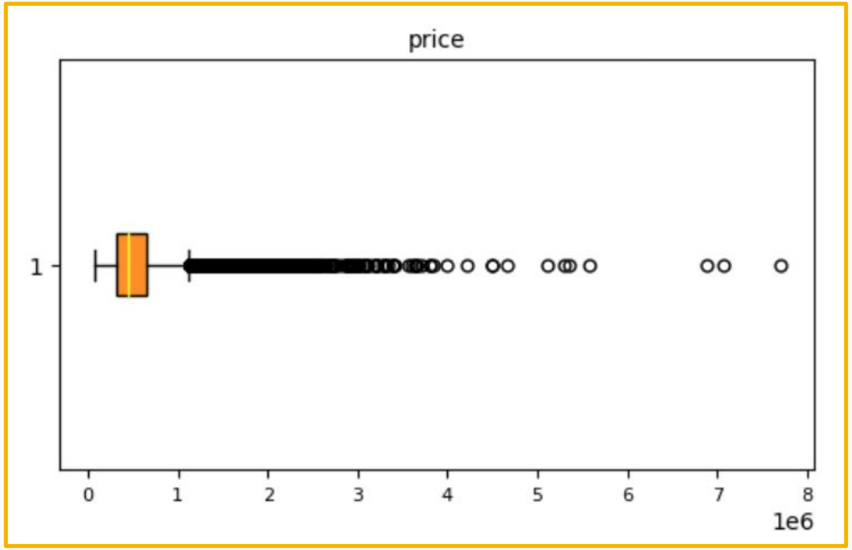
Data Exploration 📊

Insights:

No normal distribution of price data, with skewness towards lower prices

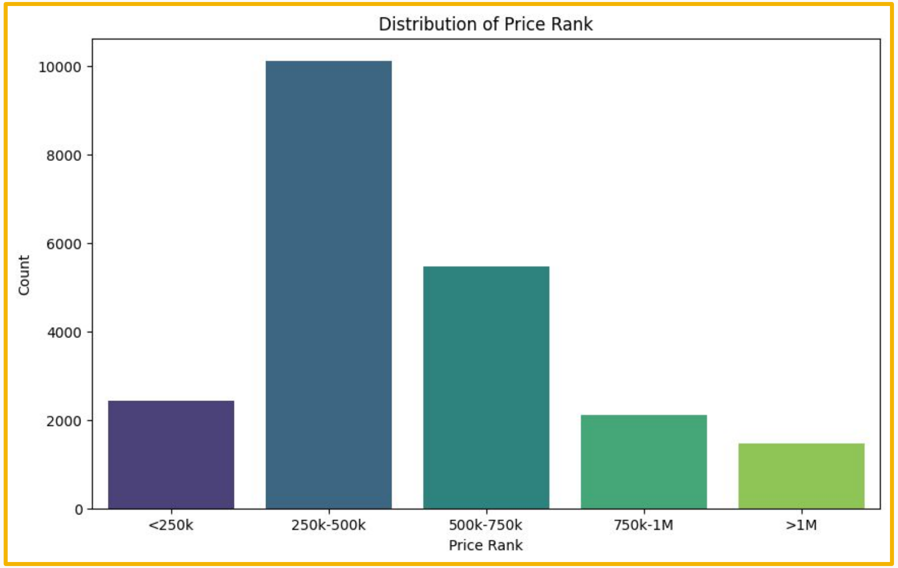
High number of **outliers** with high house prices

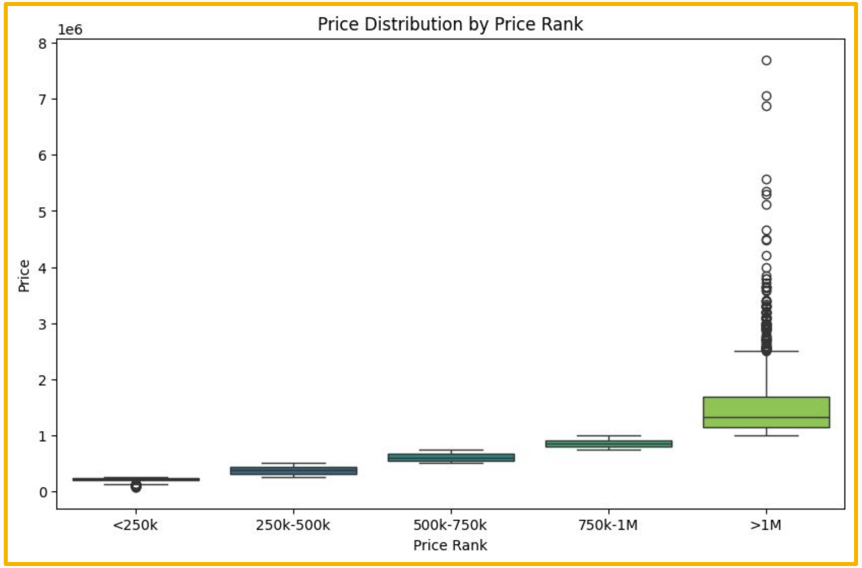




Target Exploration 🏠

- **Price_ranks**: (<250k), (250k-500k), (500k-750k), (750k-1M), (>1M)
- **Most** of the house prices are **within two rank prices (250k-500k) + (500k-750k)**
- **Most** of teh **outliers** are within teh **price rank** of houses **above 1M**

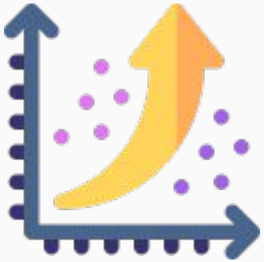




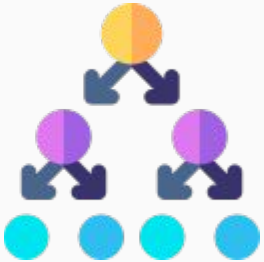
Regression ML Models tested

Train-test Split = 70% Train / 30% Test

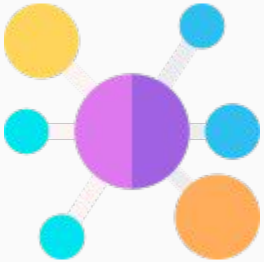
Linear Regression



Decision Tree Regressor



K-Nearest Neighbor



Results 🎯

```
Linear Regression R2 = 0.6995
Linear Regression RMSE = 208296.7277
Linear Regression MSE = 43387526779.3553
Linear Regression MAE = 127486.8026
```

```
Decision Tree R2 = 0.7366
Decision Tree RMSE = 194996.9105
Decision Tree MSE = 38023795092.7977
Decision Tree MAE = 100868.8284
```

```
Key N Neighbour R2 = 0.4932
Key N Neighbour RMSE = 270495.0558
Key N Neighbour MSE = 73167575226.5868
Key N Neighbour MAE = 164982.8095
```

Normalization

Applying Normalization with **MinMaxScaler** to the Decision Tree Regressor

Before

```
Decision Tree R2 = 0.7366
Decision Tree RMSE = 194996.9105
Decision Tree MSE = 38023795092.7977
Decision Tree MAE = 100868.8284
```

After Normalization

```
Tree Model 2 R2 = 0.7386
Tree Model 2 RMSE = 194262.7442
Tree Model 2 MSE = 37738013765.2635
Tree Model 2 MAE = 101625.9621
```

Conclusions

- Decision tree is the best model to predict accurate house sale price in the Seatle (US) market:
    - Has the highest R2 score and Lowest RMSE, MSE and MAE metrics

```
Decision Tree R2 = 0.7366
Decision Tree RMSE = 194996.9105
Decision Tree MSE = 38023795092.7977
Decision Tree MAE = 100868.8284
```

- Applying normalization with ‘MinMaxScaler’ to decision tree model it slightly improves the results

```
Tree Model 2 R2 = 0.7386
Tree Model 2 RMSE = 194262.7442
Tree Model 2 MSE = 37738013765.2635
Tree Model 2 MAE = 101625.9621
```