Data Analytics Case Study: King County Housing Prices

Holly Ringsak

Introduction

Overview: Discover variables that influence the housing market in King County, Washington, and create a predictive model to accurately predict housing prices.

Purpose & Context: This project was created as part of my data analytics course through CareerFoundry to demonstrate my abilities in exploratory visual analysis in Python, as well developing a hypothesis and testing them through various analytical approaches, and presenting my findings through Tableau.

Objective: To build an interactive dashboard visually showcasing well-curated results of an advanced exploratory analysis conducted in Python.

Role: Data Analyst

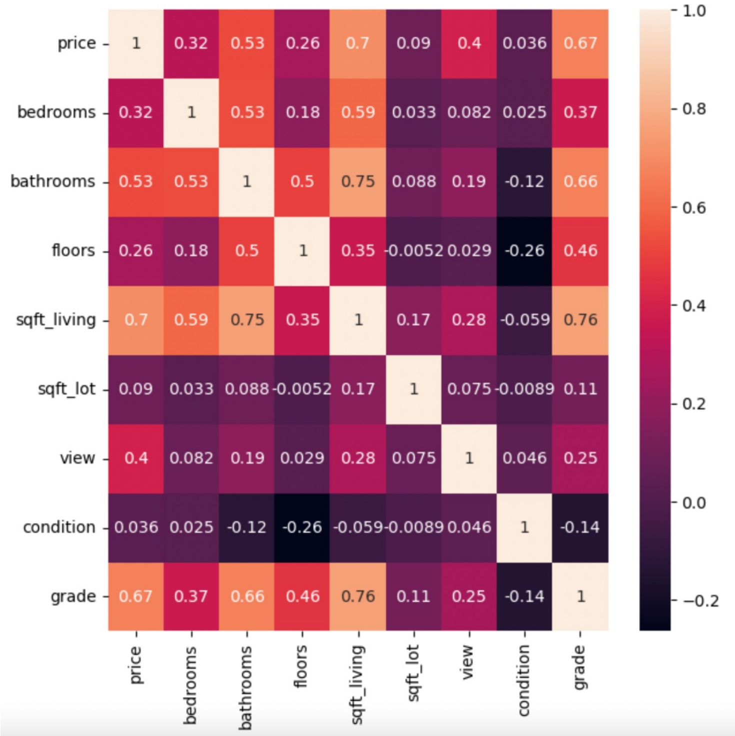
Project Duration: 2 weeks

Tools Used:

- Excel
- Python
- Tableau

Analysis Process:

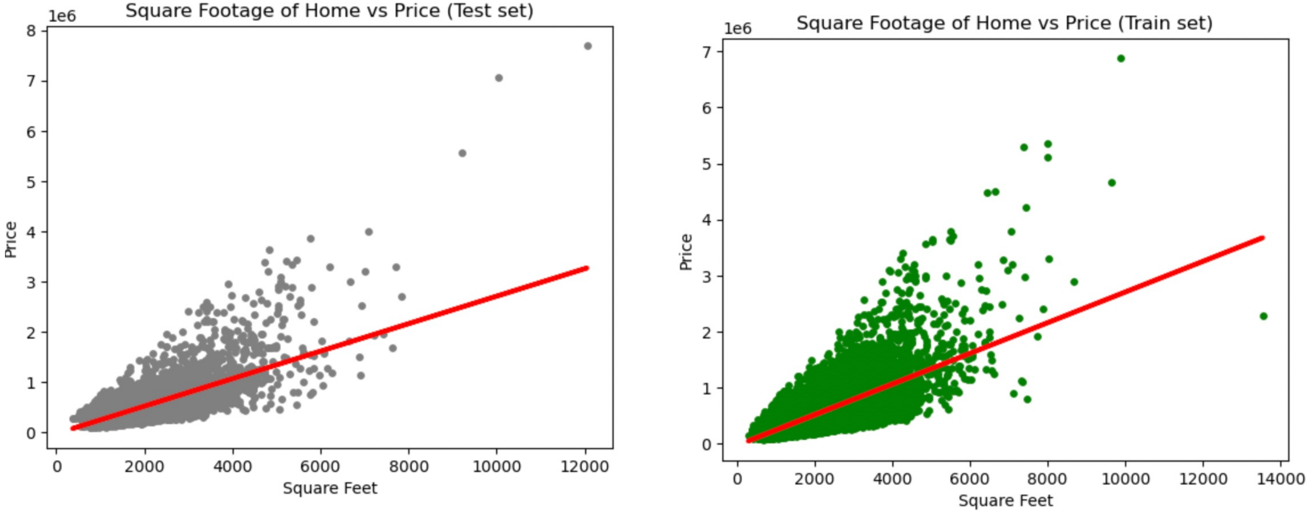
- Sourcing Open Data : I selected the House Sales in King County, USA dataset for my analysis
- Exploring Relationships: Conducted exploratory analysis in Python to test the correlation of different variables with the pricing of homes.
- Geographical Visualizations with Python: Sourced a shapefile with the same location as my data set to perform geospatial analysis.
- Supervised Machine Learning, Regression: Prepare data for regression analysis, split data into training and test sets, run linear regression and analyze the model performance.



	price	bedrooms	bathrooms	floors	sqft_living	sqft_lot	view	condition	grade
price	1	0.32	0.53	0.26	0.7	0.09	0.4	0.036	0.67
bedrooms	0.32	1	0.53	0.18	0.59	0.033	0.082	0.025	0.37
bathrooms	0.53	0.53	1	0.5	0.75	0.088	0.19	-0.12	0.66
floors	0.26	0.18	0.5	1	0.35	-0.0052	0.029	-0.26	0.46
sqft_living	0.7	0.59	0.75	0.35	1	0.17	0.28	-0.059	0.76
sqft_lot	0.09	0.033	0.088	-0.0052	0.17	1	0.075	-0.0089	0.11
view	0.4	0.082	0.19	0.029	0.28	0.075	1	0.046	0.25
condition	0.036	0.025	-0.12	-0.26	-0.059	-0.0089	0.046	1	-0.14
grade	0.67	0.37	0.66	0.46	0.76	0.11	0.25	-0.14	1

Correlation heatmap showing the correlation between different variables in the dataset

Results of linear regression test on the price and square-footage of a home. Model did not perform well.



```
Slope:  [[273.70843969]]
Mean squared error:   72796296390.32614
R2 score:   0.5057892303219628
```

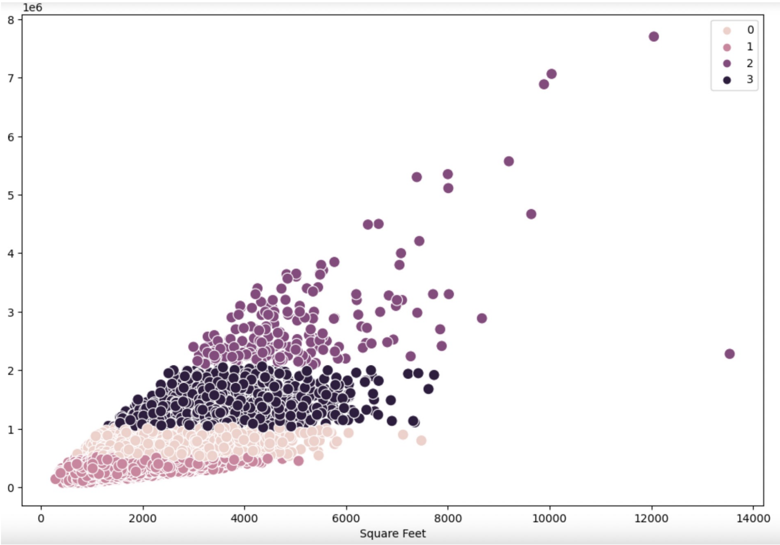
```
Slope:  [[273.70843969]]
Mean squared error:  66507995357.54318
R2 score:  0.4860368981218999
```

Analysis Process:

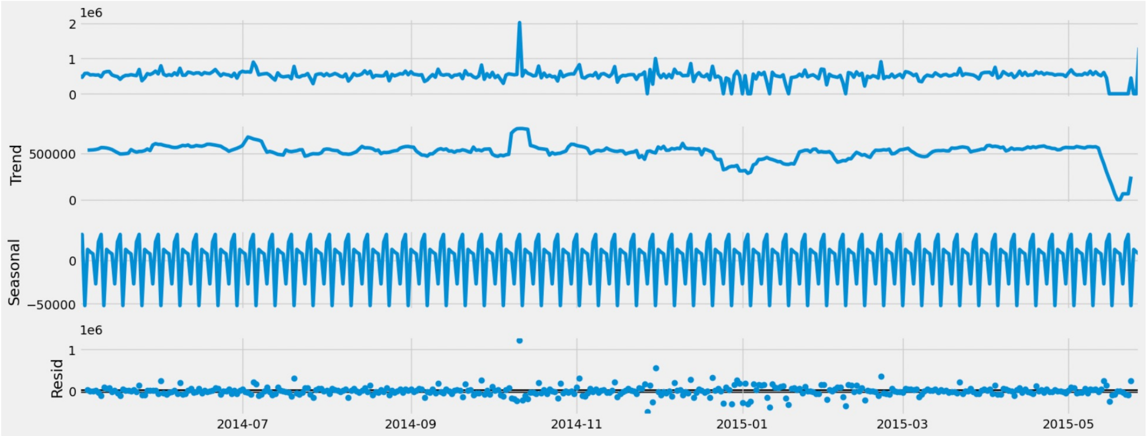
- Unsupervised Machine Learning, Clustering: Used the elbow technique to determine optimal number of clusters, ran k-means algorithm, created a variety of visualizations using clustered data, calculated descriptive statistics.

- Sourcing & Analyzing Time Series Data: Prepared data for time series analysis, split the time series into 4 components (as seen on the right), tested for stationarity using the Dickey-Fuller test and an autocorrelation test, stationarized the data through differencing.

- Creating Data Dashboards: Created a storyboard in Tableau to present findings made in Python. *Link to Tableau Presentation*



Plotted 4 clusters based on results of the elbow technique

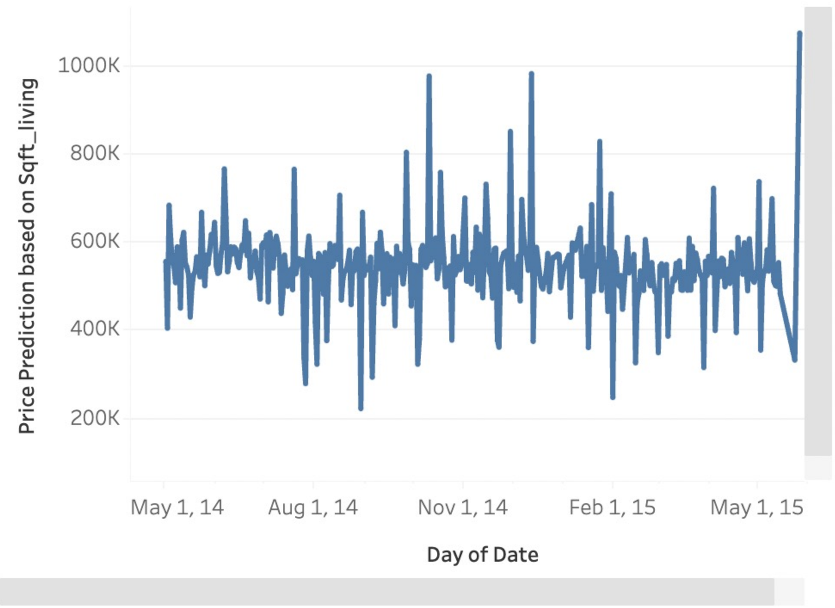


Applied an additive decomposition model to split the time series to show the underlying trend, seasonality, and noise.

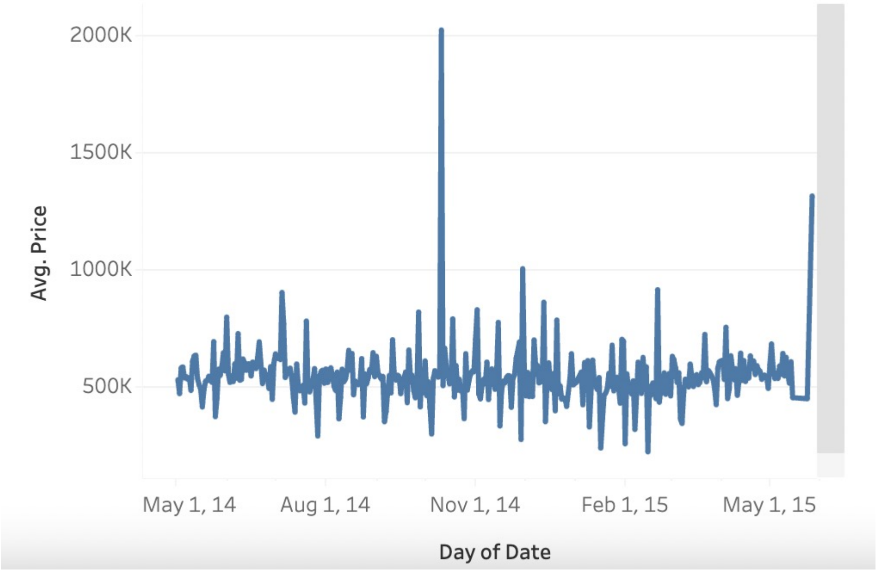
Reflection:

- Although the grade and square footage of the homes were the variables with the greatest correlation to the price, the correlation is not strong enough to build an accurate predictive model.
- Next steps: In order to make an accurate predictive model there need to be more variables with a higher correlation to price or to each other. The dataset I analyzed was sourced from King County's government website which has many other public datasets that could potentially be merged with this one to provide more insights.

Prediction of Price based on Sqft_living variable



Actual Average Price


Avg. Price
2000K
1500K
1000K
500K
May 1, 14
Aug 1, 14
Nov 1, 14
Feb 1, 15
May 1, 15
Day of Date