 Krispy3 / KingCountySales

Public

Flatiron School Phase 2 Project

☆ 0 stars

🍴 0 forks

☆ Star

▼

👁 Unwatch ▼

<> Code

🔍 Issues

🔗 Pull requests

🎬 Actions

📁 Projects


📖 Wiki

🛡 Security

📈 Insights

🔗 main ▼

⋮

 Krispy3 finalized ...

1 minute ago ⌚ 13

View code

☰ README.md

✎

# King County Sales

---

Flatiron School Phase 2 Project

Harrison Carter

Flatiron School NYC Data Science cohort 05162022



## Project Description

---

The goal of this project is to use the dataset provided by King County, WA to make predictions and inferences about housing prices according to potential predictors therein. The focal point is to use multiple linear regression to explore the parameters that correlate with my target variable (sale price) in order to build an acceptable model for analysis. All of the computational work here is done in Python and associated libraries (primarily sklearn). The primary predictor here for price is the Zip code for the associated residence, secondary predictors being the square footage of the associated living space, its waterfront juxtaposition or lack thereof, and ZIP code.

## Methodology

---

First read the data. Initially, we display a heatmap to help inform choices moving forward. here we see that the sqft\_living has a relatively high correlation, and we note that for later incorporation into the final model. We then prepare our first model and displays its regression results for inspection of the initial model between ZIP code and log price. By comparing ZIP code, we already account for around half of the variance in the data with the r squared value of 0.533. Rule of thumb is that the skew should be less than +/-0.5 for proper regression and the kurtosis should not exceed 6 to ensure normality. Here, skew is too high at 0.594 and kurtosis is borderline, with a value of 5.132. We can fix this going forward.

I selected the living area to incorporate into the multiple regression to boost the r squared value, indicating a much better fit and capturing much more variance. Using the log of the square footage for the living area does not change the r squared, but does eliminate some kurtosis and much of the skew. Both of these are further reduced by incorporating waterfront availability into the regression. This leaves us with a p-value of 0.833.

Performing cross validation on this model with our dataset yields a similar p-value to the test set, within 0.004. This means that our model is not overfit, provided that there is no data leakage (a problem that did occur and was fixed).

## Conclusion

---

This model takes two major factors that alone do not determine the price of a house and combines them to better infer their additive relationship to price, and tempers the skewness and kurtosis to acceptable levels by incorporating waterfront availability. In short, we can take this model and use it to analyze real world housing data with high confidence, accounting for prospectively 83.3 % of the variance in the data.

## Releases

No releases published

[Create a new release](#)

## Packages

No packages published

[Publish your first package](#)

## Languages

● Jupyter Notebook 100.0%