melfriedman / KingHousing

<> Code | Issues | Pull requests | Actions | Projects | Wiki | Security | Insights | Settings

main | Go to file | Add file | Code

melfriedman Update README.md ... 2 minutes ago 38

data	Add files via upload	9 days ago
images	more pictures added	5 minutes ago
practice_notebooks	organized	2 hours ago
presentation	organized	2 hours ago
.gitignore	Initial commit	13 days ago
Final_notebook.ipynb	Add files via upload	7 hours ago
README.md	Update README.md	2 minutes ago
final_notebook.pdf	resave	19 minutes ago

README.md

KingHousing

Flatiron Phase 2 project with Multiple Regression Analysis

M3 Consulting

Our team is composed of Mark Patterson, Matthew Zhang, and Mel Friedman. Our goal is to find affordable housing in Seattle, Washington and the surrounding King County area using predictive data modeling.

Overview

About

Multiple Regression Analysis on King County Houses

#machine-learning
#multiple-linear-regression
#flatiron-school-project

Readme

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 3

melfriedman Mel Friedman
markp-rankin
mzcode98

Languages

Jupyter Notebook 100.0%

Housing data for King County was provided by Flatiron School which included a little over 20,000 houses sold in 2014 and 2015. This included information such as prices, how many bedrooms and bathrooms, square feet(lot, basement, above ground, living area), the condition of the house and possible views, when it was built and if it was renovated, the zipcode and coordinates, as well as some basic data on the neighborhood.

Our intention is to use our model to help the Salazar family, a family of four, purchase their first home. Their household income is $75,000 and with a down payment of $10,000 they have been approved for a mortgage of $316,000. With the tech boom of the new millennium, and King County being home to tech giants Amazon and Microsoft, the housing market has never been more competitive. We aim to use our model to find the Salazar family a home that meets all their needs and is in their price frame, along with many other hardworking families across king county.

Goals

1. Find which features help predict home prices.

- What features can be minimized to bring down a home price?

2. Look into housing locations to see if there is any relation to price.

- Where in King County have the most affordable houses for new buyers

3. Create an accurate model with low error that includes important features that homeowners want in the price range that they can afford.

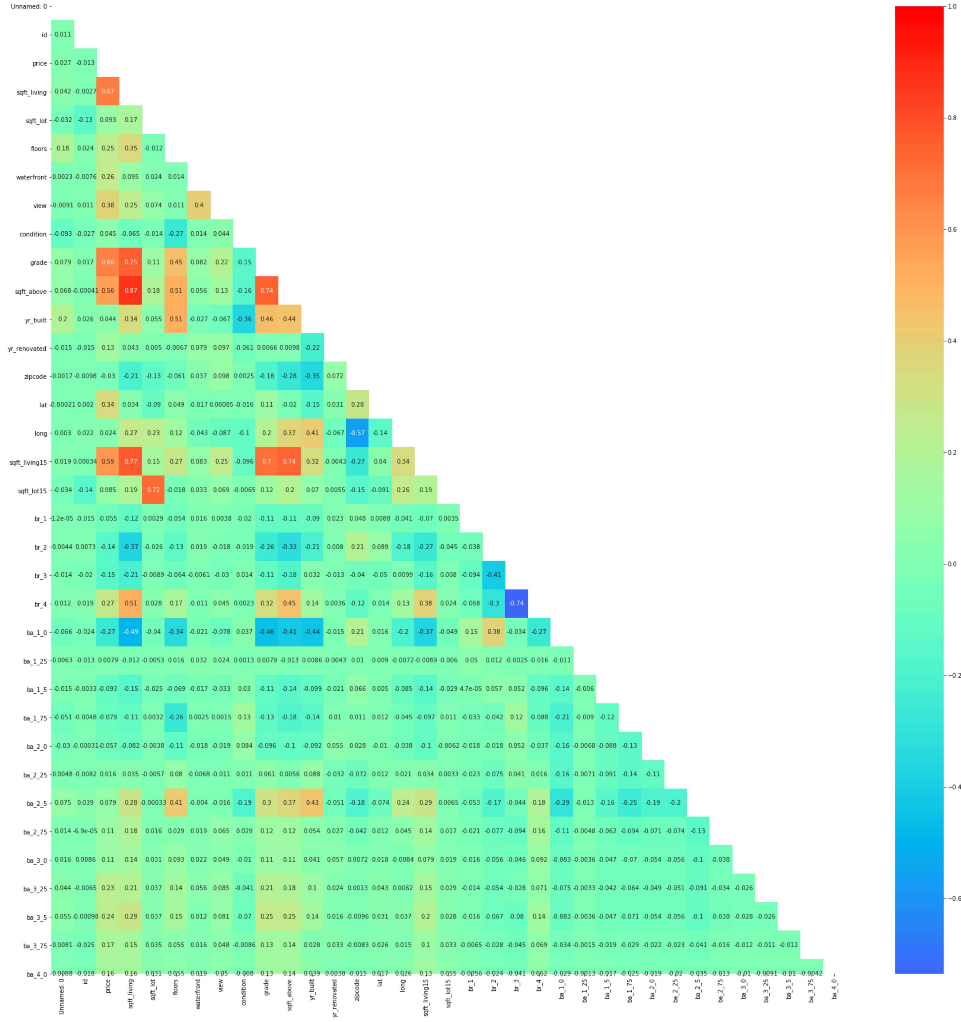
Milestones

Exploratory Data Analysis

Our EDA was started with a brief search for null values and then splitting values into their own dataframes based on the type of house feature they were to make early visualizations easier to read.

```
# Seperate groups of features into seperate dataframes: counts, size, condition
df_counts = df[['price','floors', 'bedrooms', 'bathrooms', 'waterfront', 'view']]
df_condition = df[['price','condition', 'grade', 'yr_built', 'yr_renovated', 'lat', 'long', 'zipcode']]
df_size = df[['price','sqft_lot', 'sqft_living', 'sqft_above', 'sqft_basement', 'sqft_lot15', 'sqft_living15']]
```

After this was done, we checked linearity and correlation (linearity was checked for each individual dataframe, while correlation was plotted using the original dataframe).



Outliers were deleted, such as a house with 33 bedrooms and houses where a sale price two standard deviations away from the mean. At this point we felt ready to start our first basic model.

Model Approach A

Created a preliminary model for inference and focused on low and medium priced houses in the range of \$154,000 to \$605,000. Eventually this was split to just include medium priced homes which further limited the data to \$315,000 to \$605,000. All seven of the models we created had poor R^2 levels (approximately 0.10). After further EDA we learned that this grouping had poor linearity patterns.

Model Approach B

Adjusted data to include houses only in our low priced range (\$154,000 to \$315,000) as well as one-hot encode condition, grade, bedroom, and bathrooms columns. We found that our R^2 score went up, but not by much and only ever got as high as 0.188.

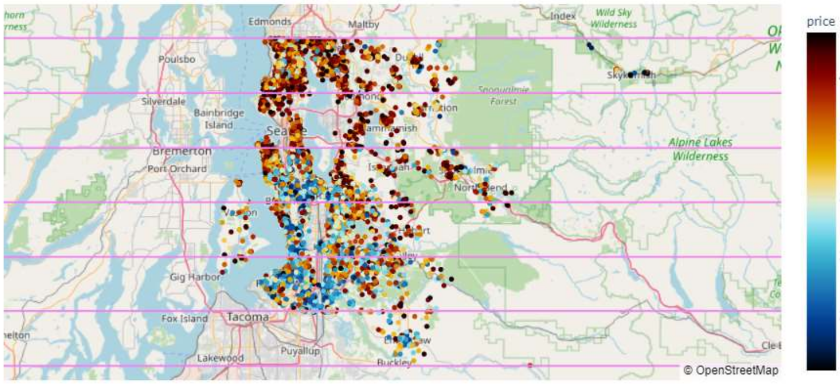
Model Approach C

Continued modeling with the dataset used in approach B, but did some transformations on our data. These transformations included getting the log of grade (np.log()) and min-max scaling on the other thirteen predictor variables with the exception of price, yr_built, and yr_renovated. R^2 did not have much significant change and only increased to 0.2.

Model Approach D

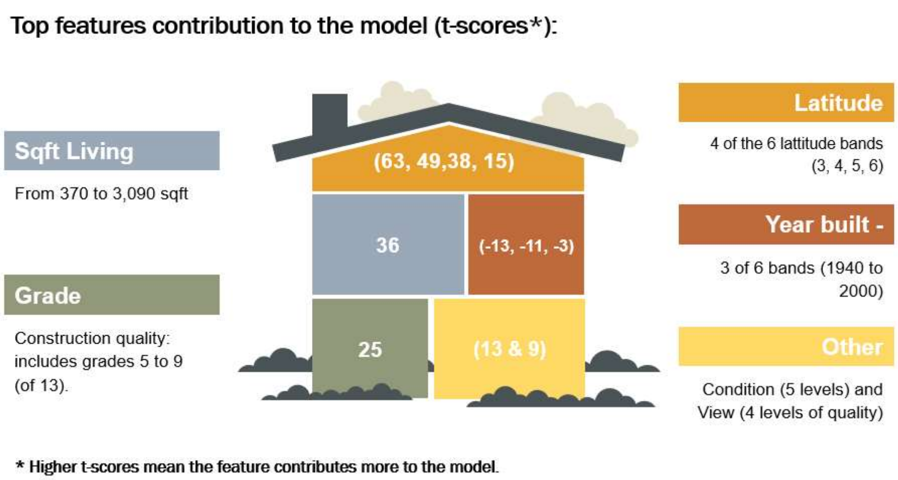
After not seeing significant changes from data transformations we decided to add more predictor variables related to locational data (latitude, longitude, and zipcode columns). Based on mapping where houses were located and how much they sold for, we could tell that houses sold in southern King County were generally less expensive than houses in the north. Once these predictor variables were added the R^2 went up all the way to 0.492.

We then created 6 bands in equal length for the latitude as well as another 6 bands on 20 year intervals for how old the homes are. After these inclusions we made further adjustments and deleted predictors that had a p-value higher than 0.05 as well as check for mutli-collinearity and delete and conflicting predictors. After these final adjustments, we ended up with an R^2 of 0.518.



Predictive Modeling

From our final model we can enter in a predictor such as how many bedrooms is wanted or how old of a house someone would want to predict what the price may be. For the case of the Salazar family, a 1,200sqft house in our 5th latitude grouping would cost approximately $365,000 which is out of their budget, however a house of the same square foot in the 3rd latitude grouping would cost approximately $256,000 which is comfortably in their price range.



Next Steps

Going forward we'd like to make more adjustments to our model so we can continue to increase our R^2. Looking back, our model's predictive power may have been lessened by limiting the house prices to the lower range. Linear regression may also not be the best modeling approach for our data and what we aim to get out of it, and so as we go forward we will be exploring other machine learning methods.

In addition, we'd like to do more analysis on different home types and how they affect price (duplex, townhouse, condos, etc), as well as taking a deeper dive into location data such as walkability and proximity to transit lines, grocery stores, medical centers, and other essential businesses. Additional features and predictors that were not included in our data set that we would like to gather and add would be information on parking (such as if there is an included parking spot, a garage, or street parking) and if the home has a private yard (especially important when including data for duplexes, townhouses, and condos).

Because our data only included houses sold in 2014 and 2015, we could potentially be missing major data due to the impact of COVID-19. In this ever changing world and going into the future it will be important to capture this data and to see how this could potentially change our model to best help families find affordable housing in this post-COVID world.