Learn Git and GitHub without any code!

Using the Hello World guide, you'll start a branch, write comments, and open a pull request.

Read the guide

amandabgaeta / **King-County-Housing-Modeling**

forked from learn-co-curriculum/dsc-phase-2-project

<> Code ⇄ Pull requests ▷ Actions ⊞ Projects 📖 Wiki ⊘ Security ⤒ Insights ⚙ Settings

main ▾

Go to file Add file ▾ Code ▾

This branch is 9 commits ahead of learn-co-curriculum:main.

⇄ Pull request ± Compare

amandabgaeta README update ⋯ 5 minutes ago 🕐 18

data	Initial commit	4 months ago
images	remove student notebook	15 minutes ago
scratchWork	updates with final notebook and images	18 minutes ago
.canvas	Create .canvas	4 months ago

About

No description, website, or topics provided.

📖 Readme

⚖ View license

Releases

No releases published
Create a new release

📄	.gitignore	Initial commit	4 months ago
📄	CONTRIBUTING.md	Initial commit	4 months ago
📄	LICENSE.md	Initial commit	4 months ago
📄	Notebook.ipynb	updates with final notebook and images	18 minutes ago
📄	README.md	Update README.md	9 minutes ago
📄	presentation.pdf	presentation added	6 minutes ago

README.md ✏️

Introduction: Business Problem

COVID has made working from home a trend that will continue. With a job less dependent on living in a certain area, people are reassessing where they live whether that is within the same area or a completely different state or country. Another side effect of COVID has been the extreme reduction in interest rates, which have driven home purchases to skyrocketed.

This model has been constructed with both of these in mind, for those interested in buying a home in King County. Buyers should be able to progress decisioning on what zipcodes they can target in their home search along with what they can expect in relation to home features and quality within their budget. This should aid in their understanding on what they want to prioritize and/or sacrifice to get a home that fits their needs and is in their price range. Information derived can also point home buyers to research other things about the areas they are interesting like schooling for future family.

Model Coverage Map

The Data

This project uses the King County House Sales dataset, which can be found in `kc_house_data.csv` in the data folder in this repo. The description of the column names can be found in `column_names.md` in the same folder.

Data Processing

Notebook.pdf includes cleansing, editing, functions, and feature engineering that was utilized in mode iterations.

The /scratchWork folder has variations of data analysis, model approaches, and notebook iterations that were used to get to the final Notebook.pdf

Model Iteration

Notebook.pdf includes an outline of the nine model iterations that were completed. From first to the last iteration of the model, the R2 was increased from 0.68 to 0.83, and RMSE was reduced from 204,432 USD to 90,878 USD -- meaning the final model can predict within ~91,000 USD of the pricing.

Data Analysis

Zipcodes were a key driver of price in this model. Thus, analysis was approached by analyzed the top 5 positive price driving zipcodes and the top 5 negative price driving zipcodes. In addition, top home features that drove price beyond zipcode location were plotted and analyzed.

The /images folder holds the images of these plots.

Recommendations on Model Utilization In Guiding Buyer Search

If new to the area or starting with no information

- 98072: Woodinville (Northeast of Seattle, outside of city
- 98027: Issaquah (far East of Seattle, outside of the city)
- 98118: Seattle (Seward Park, Ranier Valley)

If area is priority and/or buyers have more budget (Top Positive Zipcodes):

- 98039: Medina
- 98004: Bellevue
- 98112: (Seattle: Mann, Central Area)
- 98102: (Seattle: Eastlake, Cascade)

- 98109: (Seattle: Westlake, Cascade)

Negative zipcodes

- Address as buyer drives the conversation

Future Work

Buyer Homework When Utilizing This Model

Based on importance of area and features highly correlated with price, they should assess what is most important to them in a home. This will help us take more specific steps in their search. Since much of the price drivers are dependent on zipcode, it would be useful for buyers to do more contextual and qualitative research about the area as well.

Additional Data Analysis and Modeling

Due to the weight on zipcodes in driving price, there should be further research on the areas these zipcodes cover. On the surface the homes have seemingly similar stats beyond price, so there are clearly things not represented in the dataset that are driving these prices. In general, King County has plenty of open source data to explore and add to this model. Some thats stood out include schools and test scores, public transit/walk/bike scores, as well as crime in specific areas (https://www.kingcounty.gov/services/data.aspx).