rfb-vibe / **phase-2-project** Public

Using multiple linear regression modeling to analyze house sales in King County, WA

☆ **0** stars **0** forks

☆ Star ▾ Unwatch ▾

Code Issues Pull requests Actions Projects Wiki Security

main ▾ ···

rfb-vibe Merge branch 'main' of https://github.com/rfb-vibe/phase... ··· 2 minutes ago **13**

View code

README.md

Housing Sale Price Analysis in King County, WA

Authors: Rebecca Frost-Brewer

Business Understanding

Emerald City Realtors serves the King County community, providing prospective home sellers with guidance on how to improve the value of their home prior to listing.

- **Stakeholder**: Emerald City Realtors
- **Business Problem**: Emerald City Realtors need to provide prospective home sellers with guidance on how to improve the value of their home prior to listing, including the predicted increase in value expected based on improvements to particular features.
- **Business Question**: What features of their home can prospective home sellers change or improve to increase the value of their home, and by amount could this increase be specific to certain features?

These recommendations will be valuable to Emerald City Realtors because they will help prospective home sellers confidently ascertain how they can improve the value of their home, and if the investment is worth the cost.

Data Understanding

This project uses the King County House Sales dataset because Emerald City Realtors and its prospective homesellers are all based in King County. The dataset includes all data of single-family home sales from 2014-2015. The dataset itself can be found in `kc_house_data.csv` in the data folder of this GitHub repository along with the descriptions of the features, found in `column_names.md` Further information about the features can be found on the King County Assessor Website

The original dataset includes sales data for 21,597 homes with 20 different features, which include:

- `date` - Date house was sold
- `price` - Sale price (prediction target)
- `bedrooms` - Number of bedrooms
- `bathrooms` - Number of bathrooms
- `sqft_living` - Square footage of living space in the home
- `sqft_lot` - Square footage of the lot
- `floors` - Number of floors (levels) in house
- `waterfront` - Whether the house is on a waterfront
- `view` - Quality of view from house
- `condition` - How good the overall condition of the house is. Related to maintenance of house
- `grade` - Overall grade of the house. Related to the construction and design of the house
- `sqft_above` - Square footage of house apart from basement
- `sqft_basement` - Square footage of the basement
- `yr_built` - Year when house was built
- `yr_renovated` - Year when house was renovated
- `zipcode` - ZIP Code used by the United States Postal Service

Data Processing

To assist with creating sound models, we completed some data cleaning including:

- Dropping unrelated features to our business question (ID, sale date, zipcode, latitude, longitude, lot size, and the lot size and living space of a home's 15 closest neighbords)
- Dummy-encode categorical variables (`condition` and `grade`)
- Create binary variables for waterfront, view, and renovation status

Modeling

We are showing correlation and using regression coefficients in this analysis to be able to show the relationship between one or more features with sale price.

Using regression and interpreting correlation coefficients is effective for this business problem because it will allow for us to determine how sale price is impacted by different features and to what degree.

Buildng complex models with multiple features allows for us to be able to make more accurate, data-driven predictions.

Regression Results

In our final model comprising of all features except that of `cond_Poor` , `grade_12 Luxury` , and `reno_status` , our model's performance based on its adjusted R-squared improved from 38.98 percent to 57.5 percent.

Further, the Mean Absolute Error improved from our baseline score of 131878.02 to 106248.25, which is good.

In our final model, all features have a statistically significant linear relationship with sale price.

- While holding all other variables constant, the addition of a bathroom increases sale price by 29,020 dollars
- While holding all other variables constant, the addition of one floor level increases sale price by 41,040 dollars
- While holding all other variables constant, improving a home's condition from Average to Very Good increases sale price by 38,810 dollars
- While holding all other variables constant, improving a home's grade from Better to High Quality increases sale price by 82,180 dollars

Recommendations

1. Improve the grade of your home (construction quality) at a minimum to High Quality. An improvement from Better to High Quality is predicted to increase the sale price by 82,180 dollars
2. Adding an additional bathroom to your home is predicted to increase its sale price by 29,020 dollars
3. Each additional square foot of living space is predicted to add 81.12 dollars to the sale price; a 600-square foot addition would be predicted to increase the sale price by 48,672

Limitations and Next Steps

Our model only explains 57.5 percent of the variation in sale price, so we ought to be cautious with our predictions and conclusions. Further, our final model does have high levels of heteroscedasticity, which violates one of the assumptions of linear regression, such that our conclusions may be premature without additional manipulation of the data.

Next Steps:

- Collect more recent sales data for more accurate representation of the market
- Investigate influence of zipcode on sale price

For More Information

Please review our full analysis in our Jupyter Notebook or our presentation.

For any additional questions, please contact **Rebecca Frost-Brewer (frostbrewerr@gmail.com)**

Repository Structure

```
├── README.md                               <- The top-level README
for reviewers of this project
├── jnb-phase-2-project.ipynb           <- Narrative documentation of
analysis in Jupyter notebook
├── phase-2-project-presentation.pdf    <- PDF version of project
presentation
├── img                                 <- images
└── data                              <- Sourced externally
```

Releases

No releases published

Create a new release

Packages

No packages published

Publish your first package

Languages

- ● **Jupyter Notebook** 100.0%