FREDRICKKYEKI /
dsc-phase-2-project

Code | Pull requests | Actions | Projects | Wiki | Security | Insights

Phase II project of the Flat-Iron curriculum for Data Science

View license

0 stars | 452 forks | 0 watching | Activity

Public repository · Forked from learn-co-curriculum/dsc-phase-2-project-v2-3

main

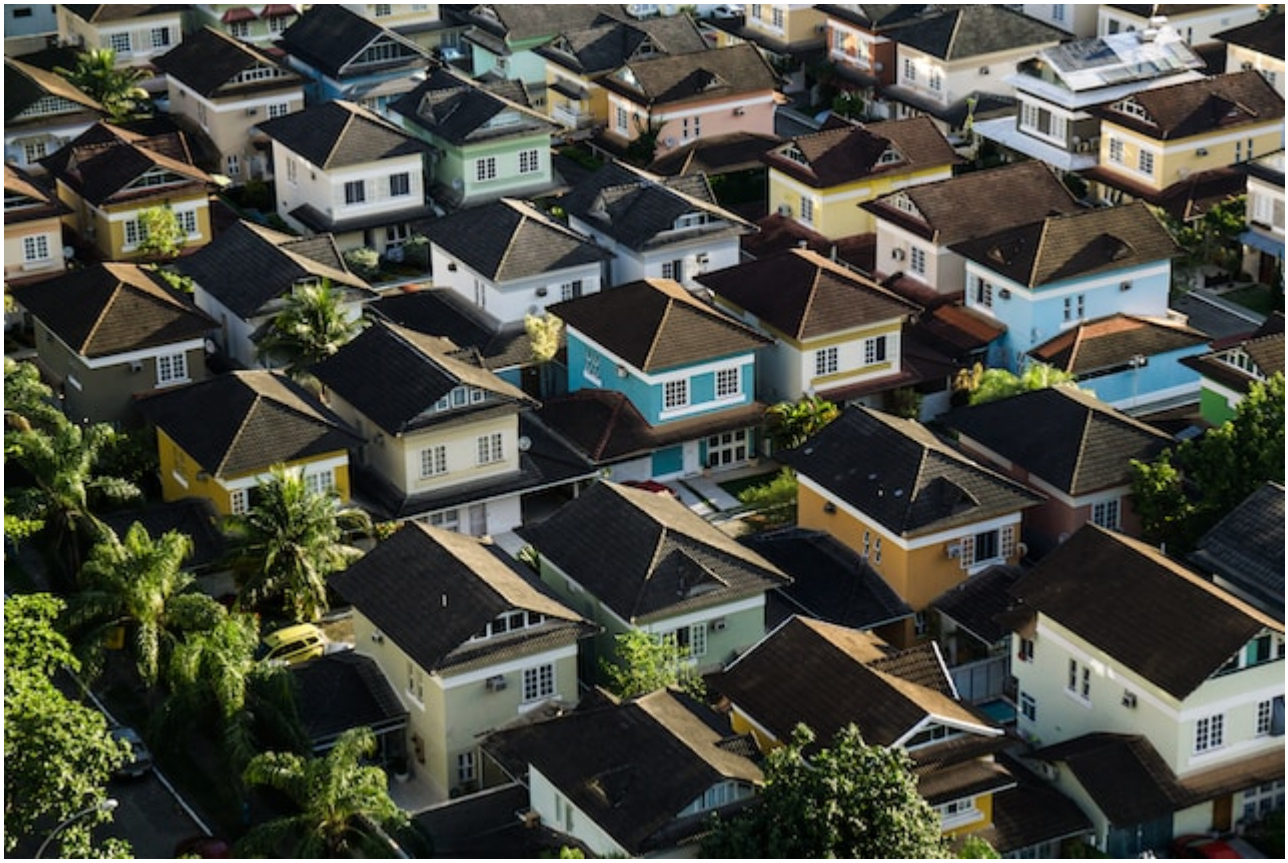
Branches | Tags

This branch is 9 commits ahead of learn-co-curriculum:main. Contribute | Sync fork

FREDRICKKYEKI [update]: add presentation pdf from dennis ... 26 minutes ago 25

View code

README.md



Phase 2 Project:

King County Housing Data: Regression Modeling.

Overview

This project focuses on analyzing housing sales data in King County, Northwest USA, using multiple linear regression modeling. The goal of this project is to provide valuable insights to homeowners, real estate agencies, and other stakeholders regarding factors that influence house prices and to make data-driven recommendations.

Contents

Project Goal

The primary goal of this project is to perform a comprehensive analysis of house sales data in King County, Northwest USA, using multiple linear regression modeling techniques. This analysis aims to provide valuable insights into the factors that influence house prices in the region and make data-driven recommendations for homeowners, real estate agencies, and other stakeholders.

Stake Holders

The intended audience for this project includes:

- **Homeowners**: Individuals looking to enhance the value of their properties through informed decision-making regarding renovations and improvements.
- **Real Estate Agencies**: Companies and agents seeking data-driven insights to assist their clients in buying and selling homes effectively.
- **Data Science Professionals**: Professionals in the field of data science and analytics interested in understanding how regression modeling can be applied to real-world business problems.

Business Problem

The business problem addressed in this project revolves around the need to empower homeowners with insights that can help them make decisions regarding home renovations that may positively impact the estimated value of their properties. Additionally, the project aims to assist real estate agencies in providing valuable guidance to their clients based on data-driven analysis.

Dataset

The dataset utilized for this analysis is the King County House Sales dataset, accessible via the `kc_house_data.csv` file within the project's data folder. This dataset contains various attributes related to houses sold in King County, including details on size, location, condition, and sale prices. A comprehensive description of the column names and data attributes can be found in the `column_names.md` file in the same folder.

Analysis Steps

The project encompasses several key analysis steps:

1. **Data Exploration**: Exploring the dataset to gain a deep understanding of its characteristics, including data distribution, the presence of missing values, and the identification of potential outliers.
2. **Data Preprocessing**: Cleaning and preprocessing the dataset, which includes addressing missing values, encoding categorical variables, and ensuring data quality.
3. **Feature Selection**: Identifying the most relevant features for building regression models, focusing on those that significantly influence house prices.
4. **Model Building**: Constructing multiple linear regression models using various sets of selected features to predict house prices effectively.
5. **Model Evaluation**: Assessing the performance of the regression models using appropriate evaluation metrics and statistical techniques.
6. **Interpretation**: Interpreting the model results, including coefficients, their significance, and their real-world implications.
7. **Recommendations**: Providing clear, data-driven recommendations to homeowners and stakeholders based on the analysis and model findings.

Model Results

The final regression model achieved an R-squared value of approximately 59.5%, indicating that 59.5% of the variance in house prices is explained by the selected predictor variables. Key factors influencing house prices include:

- square footage of living space
- the number of bathrooms

- bedrooms
- floors
- lot size
- year built
- view quality

Requirements

To run the analysis, you'll need the following:

- python3 (3.8.5)
- conda 4.10.3 (Jupyter Notebook)
- Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, statsmodels

Authors: fountain_pen:

- Fredrick Kyeki
- Ivy Mwanza
- Dennis Kobia
- Ben Ochoro

Feel free to reach out with any questions or feedback!

Repository Structure

```
├── README.md <- The top-level README for reviewers of this project
├── student.ipynb <- Narrative documentation of analysis in Jupyter notebook
├── presentation.pdf <- PDF version of project presentation
├── data <- Both sourced externally and generated from code
└── images <- Both sourced externally and generated from code
```

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%