Abstract

Our field study concerns house prices in King Country, USA. between the time period May 2014 - May 2015. We will go through data exploration to identify most important features and empirically study how the various factors influence the house prices, our regression analysis revealed the best fit model to predict the price of the house.

Design

This project is one of the T5 Data Science BootCamp requirements. Data provided by Kaggle has been used in this project.

Data

The dataset is provided in .CSV format, its contains a single file which study how the various factors influence the house prices, this dataset contains house sale prices for King County, which includes Seattle, and includes 21,613 rows and 21 columns such as price, sqft_living year built among other things.

Algorithm

- Read the data
- Split data by cross into Train and Test.
- Cleaning the dataset by check if there are any missing values, dropping outliers, dropping the duplicates.
- Rename some columns such as view to property_quality, yr_built to year_built, yr_renovated to year_renovated and grade to design_of_build to be more readable.
- Visualization the correlation between features and target.
- Modeling and Test to improve R^2, and Rmse.

Tools

- Numpy and Pandas for data manipulation
- Matplotlib and Seaborn for plotting.
- Sklearn for machine learning and statistical modeling.

Communication

The slides are provided here.