CzarProCoder /
Real-Estate-Data-Analysis-Project

Code · Issues · Pull requests · Actions · Projects · Wiki · Security · Insights · Settings

View license

0 stars · 1 fork · 1 watching · 10 Branches · 0 Tags · Activity

Public repository

main · 10 Branches · 0 Tags · Go to file t · Go to file · + · Add file · Code · ···

CzarProCoder Merge pull request #25 from CzarProCoder/Evaclaire_branch ··· 17 minutes ago

data	Include .geojson file for map boundaries	5 days ago
images	Add .pdf files and fig 10-12	40 minutes ago
pdfs	Update notebook .pdf file	32 minutes ago
.canvas	Add initial data files and jupyter notebook	last week
.gitignore	Add initial data files and jupyter notebook	last week
CONTRIBUTING.md	Add initial data files and jupyter notebook	last week
LICENSE.md	Add initial data files and jupyter notebook	last week
README.md	Finalize on the README	45 minutes ago
presentation.pptx	Add .pdf files and fig 10-12	40 minutes ago
student.ipynb	Add .pdf files and fig 10-12	40 minutes ago

README · License

REAL ESTATE DATA ANALYSIS PROJECT

*** GROUP MEMBERS ***

- Prossy Nansubuga Kamau -prossykamau@gmail.com
- Evaclaire Wamitu - evamunyika@gmail.com
- Julius Kinyua - juliusczar36@gmail.com
- Joan Nyamache - kerubonyamache@gmail.com
- Elizabeth Masai -elizabethchemtaim@gmail.com
- Kelvin Mwaura - kelvin.mwaura1@student.moringaschool
- Mourine kitili- mourinekitilimourine@gmail.com
- Allan Kiplagat - allankiplgat@gmail.com
- Mitch Mathiu - mmuriithi92@gmail.com

PROJECT OVERVIEW

The project aims to assit the real estate agents in King County to adrress the need for precise house price estimation by developing a robust predictive model .To achieve this,we will undertake indepth analysis of the real estate data provided ,which includes ,historical sales, current listings, property attributes, and other relevant features The goal of the analysis i to find the most infuential factors driving house prices and ascertain their correlation with each other.

BUSINESS PROBLEM

In this real estate market,estimation of house prices accurately is vital for home owners that is both buyers and sellers.Achieving this goal depends heavily on identifying and understanding the key factors influencing house prices. If these factors are not well taken into consideration,stakeholders may have a hard time in making sound decisions leading to potential loses to both parties.

DATA UNDERSTANDING

This project will make use of data from King County Housing Dataset.The Dataset has 21597 entries and 21 columns ,on of them being the price column which is the target variable while the rest will be used to make predictions.The dataset contains categorical and numerical columns, with data types of integers, objects, and floats.

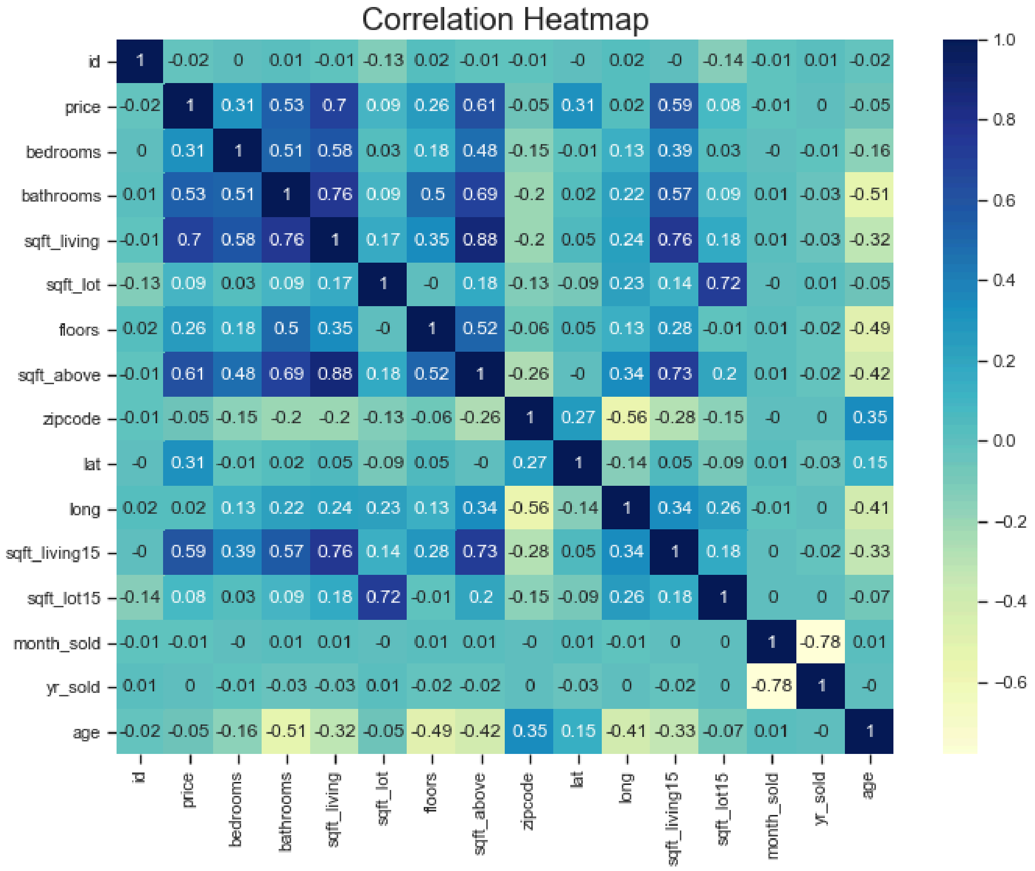
DATA PREPARATION AND CLEANING

This is marked by importing relevant libraries such as pandas,seaborn,matplotlib,statsmodel and scipy to be used in cleaning,analysis and modelling.Dataset is then loaded using pd as we observe our columns to understand the independent variables to analyse with the price.Most of the columns are numerical making it suitable for regression analysis.
Data cleaning involves checking for validity,accuracy,completeness consistency and uniformity of data.We will drop id and date column since there is no use for it,check for null values and replace them and also fill the missing values.

DATA VISUALIZATION

Before performing modelling,visualization is done to analyse some of the trends in the data. A Correlation Heat Map wil be created to identify the variables that most correlates with the target variable-Price.This is also used to check for multicollinearity of features.

HeatMap

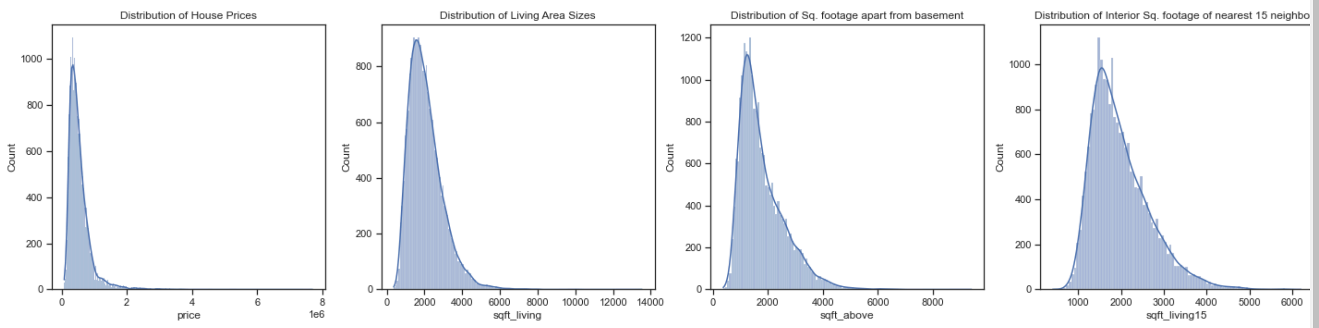


	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	sqft_above	zipcode	lat	long	sqft_living15	sqft_lot15	month_sold	yr_sold	age
id	1	-0.02	0	0.01	-0.01	-0.13	0.02	-0.01	-0.01	-0	0.02	-0	-0.14	-0.01	0.01	-0.02
price	-0.02	1	0.31	0.53	0.7	0.09	0.26	0.61	-0.05	0.31	0.02	0.59	0.08	-0.01	0	-0.05
bedrooms	0	0.31	1	0.51	0.58	0.03	0.18	0.48	-0.15	-0.01	0.13	0.39	0.03	-0	-0.01	-0.16
bathrooms	0.01	0.53	0.51	1	0.76	0.09	0.5	0.69	-0.2	0.02	0.22	0.57	0.09	0.01	-0.03	-0.51
sqft_living	-0.01	0.7	0.58	0.76	1	0.17	0.35	0.88	-0.2	0.05	0.24	0.76	0.18	0.01	-0.03	-0.32
sqft_lot	-0.13	0.09	0.03	0.09	0.17	1	-0	0.18	-0.13	-0.09	0.23	0.14	0.72	-0	0.01	-0.05
floors	0.02	0.26	0.18	0.5	0.35	-0	1	0.52	-0.06	0.05	0.13	0.28	-0.01	0.01	-0.02	-0.49
sqft_above	-0.01	0.61	0.48	0.69	0.88	0.18	0.52	1	-0.26	-0	0.34	0.73	0.2	0.01	-0.02	-0.42
zipcode	-0.01	-0.05	-0.15	-0.2	-0.2	-0.13	-0.06	-0.26	1	0.27	-0.56	-0.28	-0.15	-0	0	0.35
lat	-0	0.31	-0.01	0.02	0.05	-0.09	0.05	-0	0.27	1	-0.14	0.05	-0.09	0.01	-0.03	0.15
long	0.02	0.02	0.13	0.22	0.24	0.23	0.13	0.34	-0.56	-0.14	1	0.34	0.26	-0.01	0	-0.41
sqft_living15	-0	0.59	0.39	0.57	0.76	0.14	0.28	0.73	-0.28	0.05	0.34	1	0.18	0	-0.02	-0.33
sqft_lot15	-0.14	0.08	0.03	0.09	0.18	0.72	-0.01	0.2	-0.15	-0.09	0.26	0.18	1	0	0	-0.07
month_sold	-0.01	-0.01	-0	0.01	0.01	-0	0.01	0.01	-0	0.01	-0.01	0	0	1	-0.78	0.01
yr_sold	0.01	0	-0.01	-0.03	-0.03	0.01	-0.02	-0.02	0	-0.03	0	-0.02	0	-0.78	1	-0
age	-0.02	-0.05	-0.16	-0.51	-0.32	-0.05	-0.49	-0.42	0.35	0.15	-0.41	-0.33	-0.07	0.01	-0	1

From the above HeatMap,there is relatively strong positive correlations between price and sqft_living at 0.7, sqft_above at 0.61, sqft_living15 at 0.59 and number of bathrooms at 0.53. The weakest inverse correlations were between price and zipcode and age at -0.05, and month sold at -0.01.

Histograms

This will be used to explore the distribution of variables with the strongest positive relationships with price.



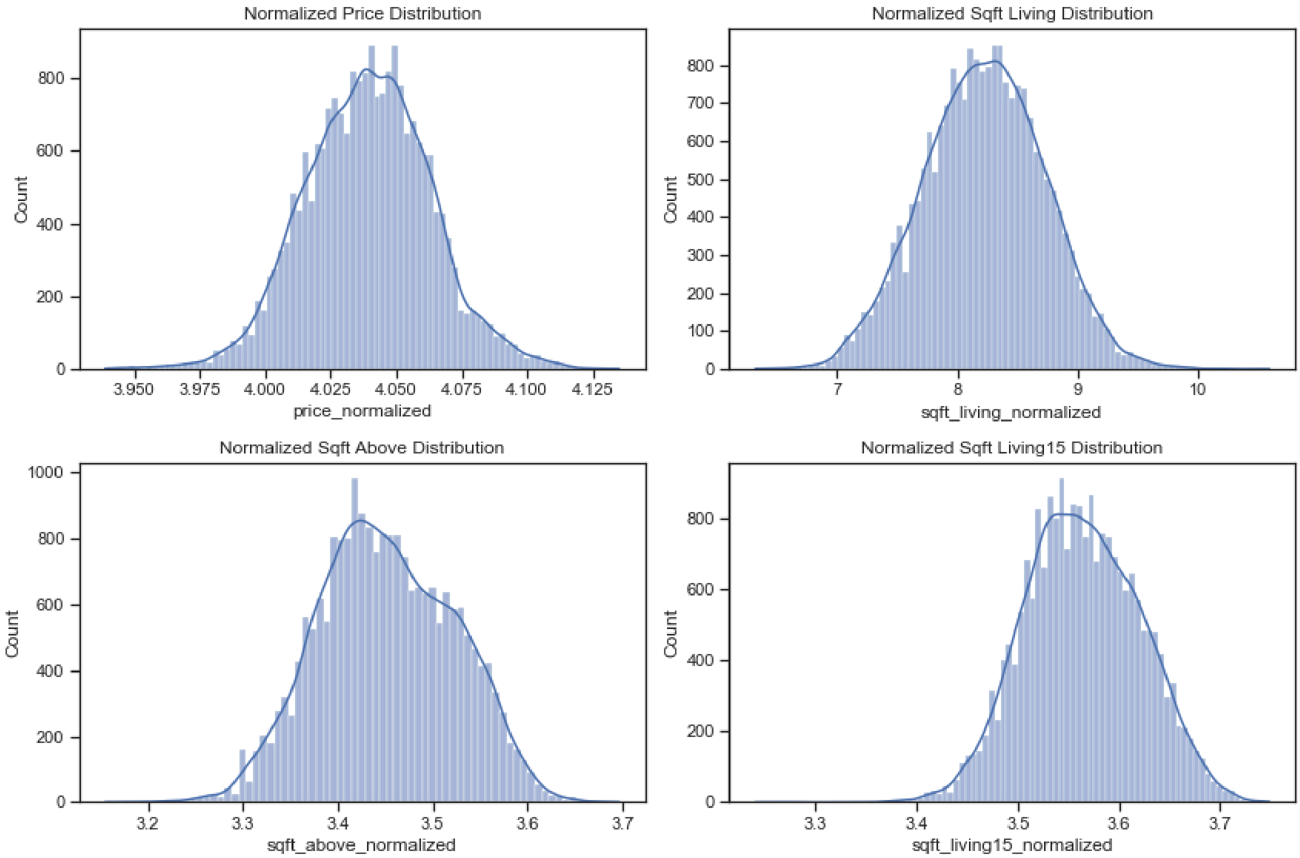
From the histograms above, we can conclude that;

1. Distribution of House Prices- the distribution appears slightly right-skewed with a skewness of 4.02 .This indicates that there are more houses at higher price renges than lower ones. The kurtosis of the price plot is 34.54 which is much greater than 3,indicating leptokurtic distribution that is, the distribution has a sharp peak and heavy tails.
2. Distribution of Living Area Sizes(sqft_living)-The skewness of the sqft_living plot is 1.47 ,implicating a moderately positively skewed distribution.The peak of the distribution suggests that most houses have living areas clustered around a specific size range, with fewer houses having very small or very large living areas.
3. Distribution of Sq. Footage apart from the basement-The skewness of the sqft_above plot is 1.45,which indicates a moderately positively skewed distribution distribution.

4. Distribution of Interior Sq. footage of nearest 15 neighbors-The skewness of the sqft_living15 plot is 1.11, indicating a moderately positively skewed distribution.This suggests that houses in a given neighborhood tend to have similar interior square footage sizes.

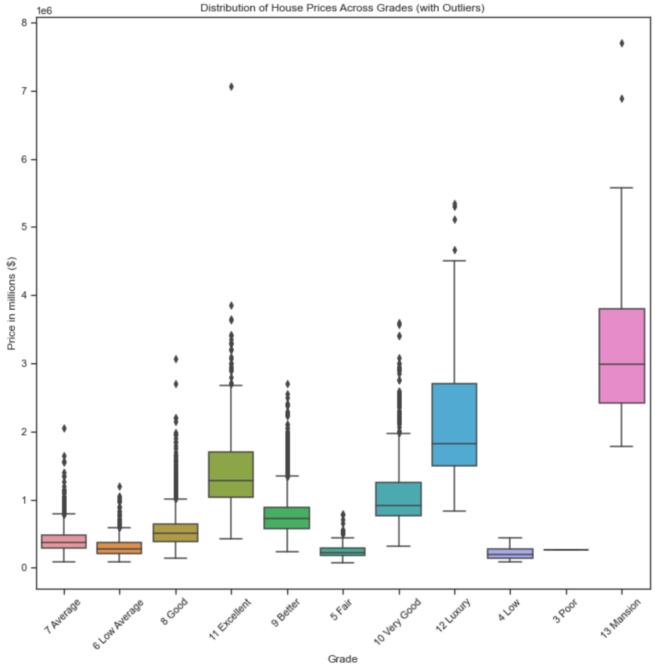
Bar Plots

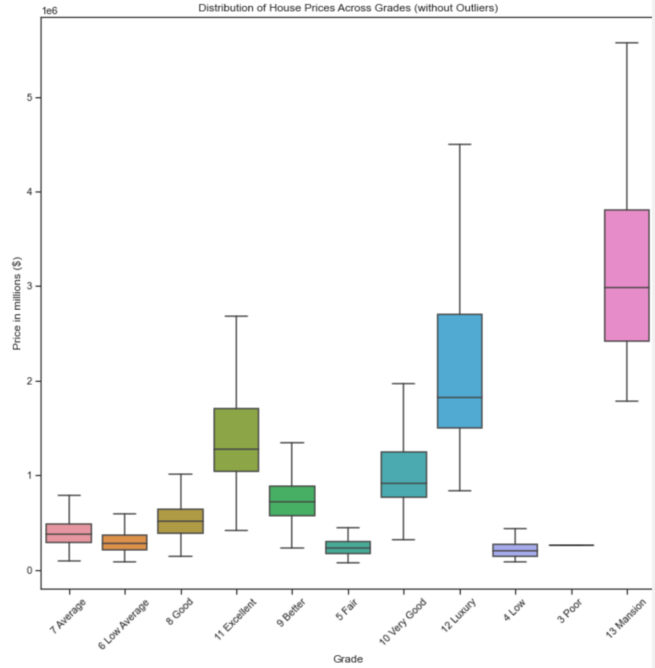
These visuals will help analyze the variation of average house prices as per the overall condition of the house ,that is whether the house is located in a water front or not,the quality of views from the house and the number of levels in the house.

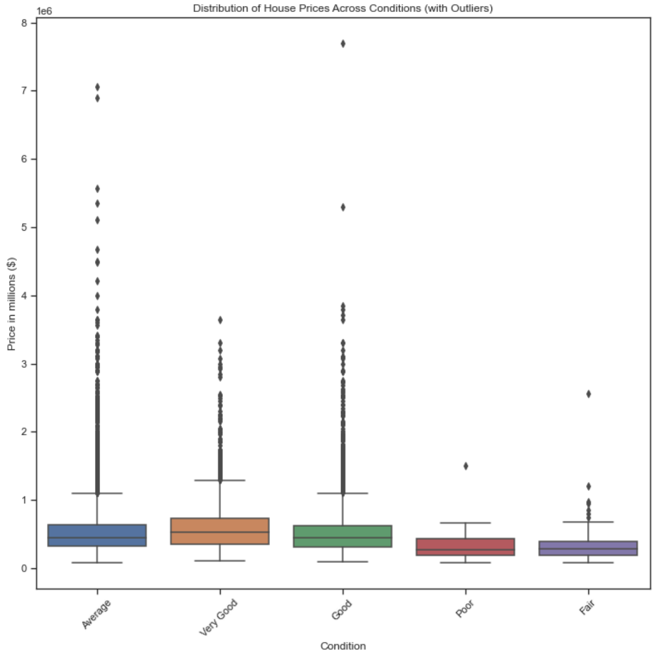


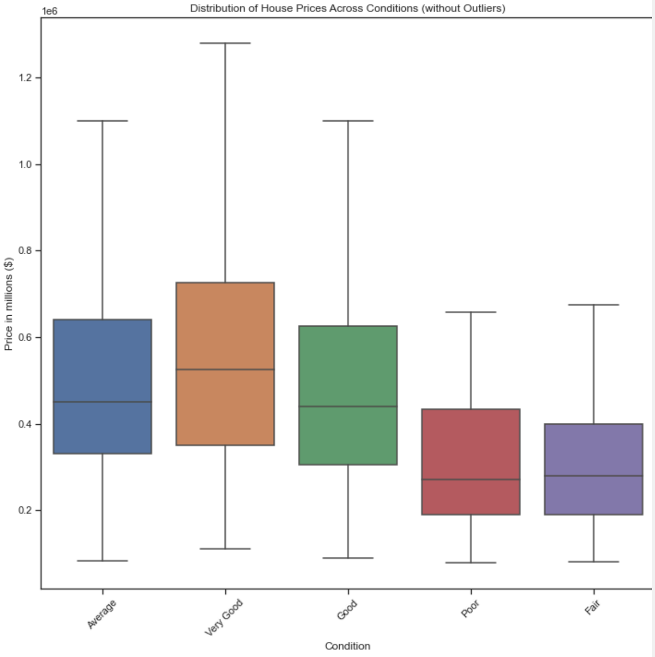
Scatter Plots

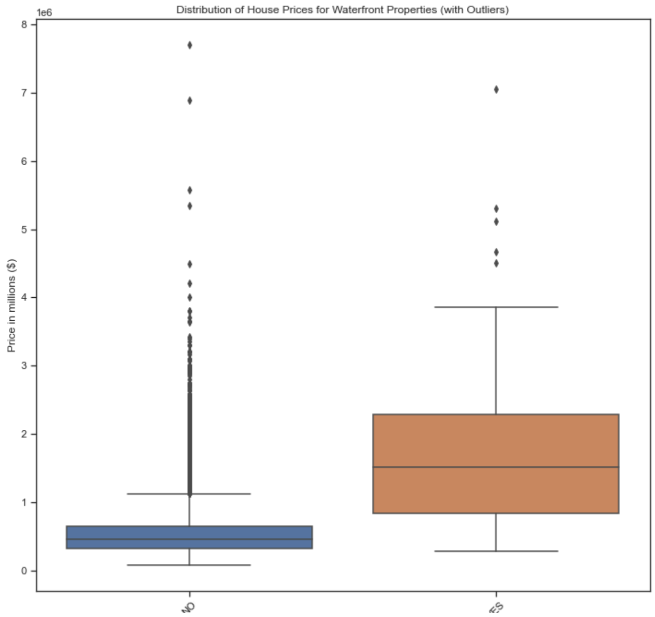
We will create a grid of scatter plots to visualize and better explore the relationships between variables and check if they pass the assumptions of linear regression.That is Linearity,Homoscedasticity and normality assumptions.

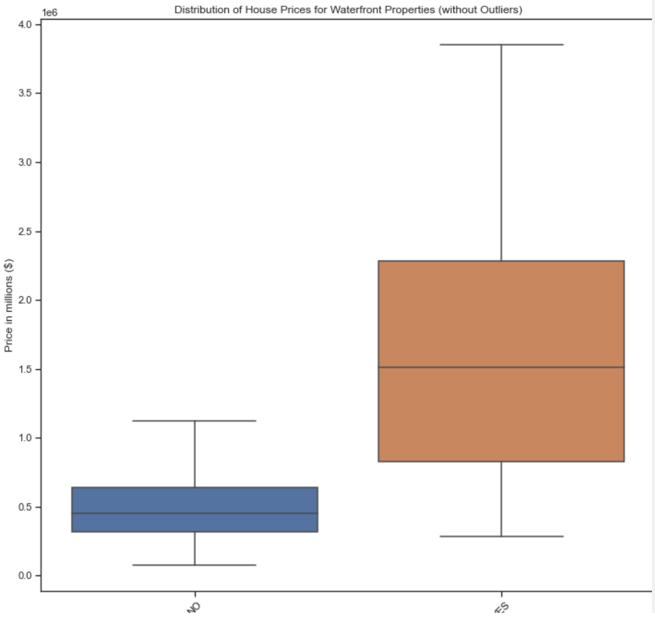

Distribution of House Prices Across Grades (with Outliers)
1e6
Price in millions ($)
0
1
2
3
4
5
6
7
8
7 Average
6 Low Average
8 Good
11 Excellent
9 Better
5 Fair
10 Very Good
12 Luxury
4 Low
3 Poor
13 Mansion
Grade


Distribution of House Prices Across Grades (without Outliers)
1e6
Price in millions ($)
0
1
2
3
4
5
7 Average
6 Low Average
8 Good
11 Excellent
9 Better
5 Fair
10 Very Good
12 Luxury
4 Low
3 Poor
13 Mansion
Grade


Distribution of House Prices Across Conditions (with Outliers)
1e6
Price in millions ($)
0
1
2
3
4
5
6
7
8
Average
Very Good
Good
Poor
Fair
Condition


Distribution of House Prices Across Conditions (without Outliers)
1e6
Price in millions ($)
0.2
0.4
0.6
0.8
1.0
1.2
Average
Very Good
Good
Poor
Fair
Condition


Distribution of House Prices for Waterfront Properties (with Outliers)
1e6
Price in millions ($)
0
1
2
3
4
5
6
7
8
NO
ES


Distribution of House Prices for Waterfront Properties (without Outliers)
1e6
Price in millions ($)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
NO
ES

MODELING

In this section,Ordinary Least Squares (OLS) will be used and the values that will be looked at in the summary will be R-squared and P-values. R-squared communicates the level of variance around our target variable(Price) that can be explained by the model.P-value on the other hand is used to check the null hypothesis ,that is if there is a relationship between the target(price) and the chosen variables.If the p-value is not less than 0.05,we fail to reject the null hypothesis

General Overview

We will take a general overview on the variables that have a strong correlation with the price. Based on the visualization done above,it is evident that the square footage of living space (sqft_living) shows the strongest positive correlation with the price, marked at 0.7,this indicates a significant impact of the size of the living area to price.On the other hand the year the house was built has a weaker positive correlation of 0.5 .

Baseline Model

In this model ,we will use price as the dependent variable and sqft_living as the independent variable to determine the coefficient and the y-intercept. The model has an R-squared value of 0.49 expolaining 49% variation in price ,making it statistically significant.The intercept and coefficient for sqft_living are approximately -$43,990 and 281, respectively, both of which are statistically significant. We will also come up with a qqplot which gives an indication that using a polynomia regression for price and sqft_living would have been more effective .

First Model

This model incorporates all the other independent variables to understand their impact on price.We will introduce dummies for the categorical data. This model has an approximate R-squared of 0.6 indicating a 60% variance in price.It i however off by $156,659 as given by our mean absolute error.

Second Model

Since our dataset suggests that some columns are stored in object form but are supposed to be numeric suc as yr_renovated and sqft_basement,we will convert them into numeric.The transformation will help to train our machine learning model which requires numeric data.We will create a new multiple linear regrerssion model which will include th new numeric columns created.We will perform one-hot encoding on the view column. The new model has an adjusted R squared of approximately 52% ,this explains 52% of the variance in price .The model is statistically significant.However the predictions are off by $169,937.This makes the previous model better than this model.

Third Model

We will build another model using the Waterfront column after which we will split the dataset into training and testing sets .The model is statistically significant with an adjusted R-squared value of approximately 55%,this explains 55 % of variation in price.However,the model's predictions are off by $167,435,amking the previous model better than this.

Fourth Model

We will build another model using the condition column . The model is statistically significant with an adjusted R-squared value of approximately 53%, indicating a 53% of variation in price.

after including all the categorical variables independently into our model, we have come to the conclusion that, we will go with the first multiple linear regression model ,that is, 'First Model' as it takes into account all the independent variables that are highly correlated with price and it also is the model with the highest r sqaured value of .592. It explains about 60% percent of the variance in price.

Linear Regression

After testing and training our model,regression metrics like r2_score, mean_absolute_error and mean_squared_erro will be used to evaluate the performance of our machine learning model.

A summarised interpretation of the the metrics is as below;

1. r-squared value of approximately 0.551 indicates that the linear regression model explains about 55% of the variance in the target variable price.

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 8

Languages

● HTML 85.3% ● Jupyter Notebook 14.7%

Suggested workflows

Based on your tech stack

SLSA Generic generator

Configure

Generate SLSA3 provenance for your existing release workflows

Jekyll using Docker image

Configure

Package a Jekyll site using the jekyll/builder Docker image.

More workflows

Dismiss suggestions