* Project - Housing Prices Prediction
* Objective – To learn and apply the techniques of Linear Regression to the given data of House Prices of Washington area of the United States. Eventually fit a multilinear model to the given dataset and see how predictions are done on unseen dataset.
* Dataset description - The dataset consists of house prices from King County an area in the US State of Washington, this data also covers Seattle. The dataset was obtained from Kaggle. The dataset consisted of historic data of houses sold between May 2014 to May 2015.

The dataset consisted of 21 variables and 21597 observations.

* Variable description –

id - Unique ID for each home sold

date - Date of the home sale

price - Price of each home sold

bedrooms - Number of bedrooms

bathrooms - Number of bathrooms, where .5 accounts for a room with a toilet but no shower

sqft\_living - Square footage of the apartments interior living space

sqft\_lot - Square footage of the land space

floors - Number of floors

waterfront - A dummy variable for whether the apartment was overlooking the waterfront or not

view - An index from 0 to 4 of how good the view of the property was

condition - An index from 1 to 5 on the condition of the apartment,

grade - An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.

sqft\_above - The square footage of the interior housing space that is above ground level

sqft\_basement - The square footage of the interior housing space that is below ground level

yr\_built - The year the house was initially built

yr\_renovated - The year of the house’s last renovation

zipcode - What zipcode area the house is in

lat - Lattitude

long - Longitude

sqft\_living15 – Updated square footage area of living space based on year 2015

sqft\_lot15 - Updated square footage of land space based on year 2015

* Importing libraries –

1. Numpy – library to work with arrays.
2. Pandas – Contains functions for cleaning, manipulating, and analysing the data.
3. Matplotlib – To make plots.
4. Seaborn – To make plots.
5. Sklearn – To use train\_test\_split, to scale the variables.

* Data Cleaning –

1. Checked for null values in the dataset – found none.
2. Checked and dropped duplication in IDs – found 177.

* Feature Engineering –

1. Calculated age of the house by changing the date format and keeping the renovation date in consideration, removed other columns like id, year built, sold year, renovated date.
2. Houses having condition 1,2 on a scale of 1-5 are very less hence replaced with 1.5.
3. Clubbed the 70 unique zip codes into four main categories namely – ‘Average, Good, very good, Posh’ grouping them based on the mean prices for each zip code.

* EDA –

1. Outliers were identified for each 21 variables by using means of scatter plot and then comparing other featured from same category ids to be considered as outliers, then removed.
2. Checked the variables for multicollinearity using bivariate analysis.
3. Variables having very less correlation with target variables are dropped.
4. Dummy variables for ‘Zipcode variable’ were made.

* Model building –

1. Data was split into train and test with ratio of 70:30.
2. Different method of scaling like StandardScaler, MinMaxScaler were used and maximum accuracy came with MinMaxScaler.
3. Build our model cautiously keeping an eye on p-value and used cross-validation technique to select the important/driver variables in our model. But the relationship between predicted and actual values of price weren’t showing linear relationship hence we decided to do log transformation.

Before log transformation :

Chart, scatter chart

Description automatically generated

After log transformation:

Chart, scatter chart

Description automatically generated

1. At last, 7 feature that came to be affecting the house prices were –

Bathroom

Sqft\_living

Waterfront

View

Condition

Grade

Zipcode

And predictions were made.

* Residual Analysis –

1. We checked the assumption of normality of residuals using Q-Q Plot and histogram as shown below:

Chart, line chart

Description automatically generated

Chart, histogram

Description automatically generated

1. Correlation between actual points and fitted vales was checked to be very high.
2. To check the goodness of fit of model:

Scatter plot for fitted vs actual values with a line having slope 1 was made, the points lie close to the reference line.

1. To check the presence of heteroscedasticity:

Scatter plot for residuals vs fitted was made. If the points seem to scatter randomly, it is safe to assume that the errors are homoscedastic. It was random.