FLIP ROBO

House Price Prediction

Submitted by:

Prateek AP

INTRODUCTION

Business Problem Framing

A house is one of the most basic needs of a human being. Hence Real Estate industry as a whole holds a very high rank in-term's contribution to the World Economy as a sector. The overall Real Estate business was valued at **US$ 6,883 Billion** in 2021. Many reports indicate Real Estate to contribute about 7% of the World Economy. Real Estate is among the top 5 sectors that contribute to world economy and is expected to see a CAGR of 10.5%.Now 10.5% CAGR provides a lot of opportunities for the businesses and with many established players in this highly populated sector and new companies looking to make their cut in to this Market, Data science comes as a very important tool to be competitive or have an edge over the others in the market with predictive analysis, which would help them formulate strategies and taking informed decisions and hence increase their bottom-line. Predictive modelling, Market Mix Modelling, Recommendation Systems are some of the machine learning techniques used for achieving the business goals for the housing companies. Our problem is related to one such housing company.

A US-based housing company named **Surprise Housing** has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below.

The company is looking at prospective properties to buy houses to enter the market. We are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not.

For this company wants to know:

- Which variables are important to predict the price of variable?

- How do these variables describe the price of the house?

Conceptual Background of the Domain Problem

Housing is one of the important and significant concepts in the real estate market which eventually increase the participation of either investing companies or individuals who are interested in buying the prospective properties across the globe. So, we are interested in this domain to solve the problem of predicting the appropriate prices for the properties across the Australian market so the client can maximize their profit by buying the properties at lower than actual price and selling them for higher price thus increasing the company's revenue.

We will be interested in knowing the factors that affect the price so we could formulate a buying strategy for a property.

Review of Literature

Our client has collected the housing details of the Australian real estate market, to analyze the data trends across the domain so as to get an insight on the various factors that are responsible that decide a price of a property. On Analysis, we have got dataset having records of approximately about individual 1400 entries of house details across Australian market. Reviewing the dataset, we have the raw dataset containing null, missing values in certain features which are most important to predict our target variable.

We also analyzed the data and scope according to many statistical parameters which describe the data loopholes which leads to wrong prediction. We now have to check & treat the data to get maximum accuracy for the prediction of target variable i.e., SalePrice. We see the house prices are between the price range of 35000 units to 755000 units.

Motivation for the Problem Undertaken

The main motivation behind picking up this problem is to get an in-depth insight on the Australian real estate trends to know the reason behind ups and downs of the market.

In order to maximize the profit of our client our Machine Learning model plays an important role to actually get the parameters which affect the sales price to make them earn profit out of the each and every individual house they deal with.

Analytical Problem Framing

Mathematical/ Analytical Modelling of the Problem

On analysis of the dataset, we got a dataset has target variable as continuous values, hence this is a Regression problem. We shall first interpret the data for our analysis by getting some statistical parameters like mean, median & quantiles and decide what are the factors that help predict the target variable. We have features having right or left skewness in most of the columns, which are treated using power_transfom. We will then identify & delete the outliers using quantiles method for our better model building. Other than that, we have some columns having null values which will be treated to replace the NAN values with the mean values if it's a Numerical dataset and replaced with mode if the data is of Object type.

We classify the data into Numerical, Discrete, Categorical values to analyze it using various plot.

Data Sources and their formats

Data set is for Australian Real Estate Market, the dataset file we are using is of the csv format. The dataset captures various features which is of temporal, continuous, discrete and categorical which will be used to analyze the behavior of the target variable.

Data Preprocessing Done

1. Loading Dataset
2. Looking at statistical parameters.
3. Feature Engineering.
4. Replacing the Null values by mean, mode or according the various parameters provided by the client in the literature.
5. Categorical values are encoded using One-Hot Encoding.
6. Number of Unique value records
7. Removing Outliers
8. Removing skewness
9. Checking VIF
10. PCA

Data Inputs- Logic- Output Relationships

The input data provided, helps to understand the house specifications, independent features which is responsible to find the sale price of house which is the target dependent variables. Any change in the individual features may change the target variable.

State the set of assumptions (if any) related to the problem under consideration

No such assumptions are made while analyzing or building the model.

Hardware and Software Requirements and Tools Used

Minimum Hardware Requirements:

- 8Gb+ of RAM
- 128 GB or more (256 GB recommended)
- 10Gb+ of free hard drive space
- Working keyboard
- Trackpad/Mouse
- Display
- A power Adapter.

Software's and Tool's Used:

- Jupyter Notebook
- NumPy
- Pandas
- Matplotlib
- Seaborn
- Plotly
- Scikit Learn

Model/s Development and Evaluation

Identification of possible problem-solving approaches (methods)

The outliers were corrected by IQR Method and the threshold was taken as 3.5, instead of 1.5 to limit the data loss, the regression models and KNN would not perform good as they are very sensitive to the outliers and the tree-based models or Ensemble techniques would perform better. Let’s run the various algorithms to check the same.

Testing of Identified Approaches (Algorithms)

- LinearRegression

  -
    ```
    Model Report:
    MSE 816256678.8115723
    Mean Absolute Error 20681.4116021397
    r2_score 84.26206208724409
    CV 76.80361724946184
    Difference between r2_score and cv is 7.4584448377822525
    ```
- Ridge

  -
    ```
    Model Report:
    MSE 816385236.1903766
    Mean Absolute Error 20681.4116021397
    r2_score 84.26206208724409
    CV 77.07961714635309
    Difference between r2_score and cv is 7.182444940891003
    ```
- Lasso

  -
    ```
    Model Report:
    MSE 816054513.5916703
    Mean Absolute Error 20681.4116021397
    r2_score 84.26206208724409
    CV 76.84376496196742
    Difference between r2_score and cv is 7.418297125276666
    ```
- ElasticNet

  -
    ```
    Model Report:
    MSE 1549183911.8860266
    Mean Absolute Error 20681.4116021397
    r2_score 84.26206208724409
    CV 72.23397294570952
    Difference between r2_score and cv is 12.028089141534565
    ```
- KNeighborsRegressor

  -
    ```
    Model Report:
    MSE 2093512560.302857
    Mean Absolute Error 20681.4116021397
    r2_score 84.26206208724409
    CV 64.56439587923313
    Difference between r2_score and cv is 19.697666208010958
    ```
- DecisionTreeRegressor

  -
    ```
    Model                                                 Report:
    MSE                                        2513132078.2095237
    Mean          Absolute          Error        20681.4116021397
    r2_score                                    84.26206208724409
    CV                                         54.413889016532565
    Difference between r2_score and cv is 29.848173070711525
    ```
- ExtraTreesRegressor

  -
    ```
    Model                                                  Report
    MSE                                        1016430748.5230589
    Mean          Absolute          Error        20681.4116021397
    r2_score                                    84.26206208724409
    CV                                          82.51635948164918
    Difference between r2_score and cv is 1.7457026055949143
    ```
- RandomForestRegressor

  -
    ```
    Model                                                 Report:
    MSE                                        1117211661.8226366
    Mean          Absolute          Error        20681.4116021397
    r2_score                                    84.26206208724409
    CV                                          80.24362322661895
    Difference between r2_score and cv is 4.018438860625139
    ```
- BaggingRegressor

  -
    ```
    Model                                                 Report:
    MSE                                          1248466933.45473
    Mean          Absolute          Error        20681.4116021397
    r2_score                                    84.26206208724409
    CV                                          77.06437349168492
    Difference between r2_score and cv is 7.197688595559171
    ```
- GradientBoostingRegressor

  -
    ```
    Model                                                 Report:
    MSE                                         960324585.3935404
    Mean          Absolute          Error        20681.4116021397
    r2_score                                    84.26206208724409
    CV                                          81.52593182521369
    Difference between r2_score and cv is 2.7361302620304
    ```
- AdaBoostRegressor

  -
    ```
    Model                                                 Report:
    MSE                                        1312407394.2840106
    Mean          Absolute          Error        20681.4116021397
    r2_score                                    84.26206208724409
    CV                                          73.79809822924052
    Difference between r2_score and cv is 10.463963858003567
    ```

Run and Evaluate selected models

The algorithms used for hyper parameter tuning are:

- RandomForestRegressor

  -
    ```
    Model                                                           Report
    MSE                                                  1218929673.860992
    Mean            Absolute             Error            20681.4116021397
    r2_score                                             84.26206208724409
    CV                                                    80.3848846892926
    Difference between r2_score and cv is 3.8771773979514847
    ```
- ExtraTreesRegressor

  -
    ```
    Model                                                           Report
    MSE                                                  983991981.1932831
    Mean            Absolute             Error            20681.4116021397
    r2_score                                             84.26206208724409
    CV                                                   82.61506684749486
    Difference between r2_score and cv is 1.6469952397492307
    ```
- GradientBoostingRegressor

  -
    ```
    Model                                                           Report
    MSE                                                  999511168.7553064
    Mean            Absolute             Error            20681.4116021397
    r2_score                                             84.26206208724409
    CV                                                   81.93426033017059
    Difference between r2_score and cv is 2.3278017570735017
    ```

Final Model:

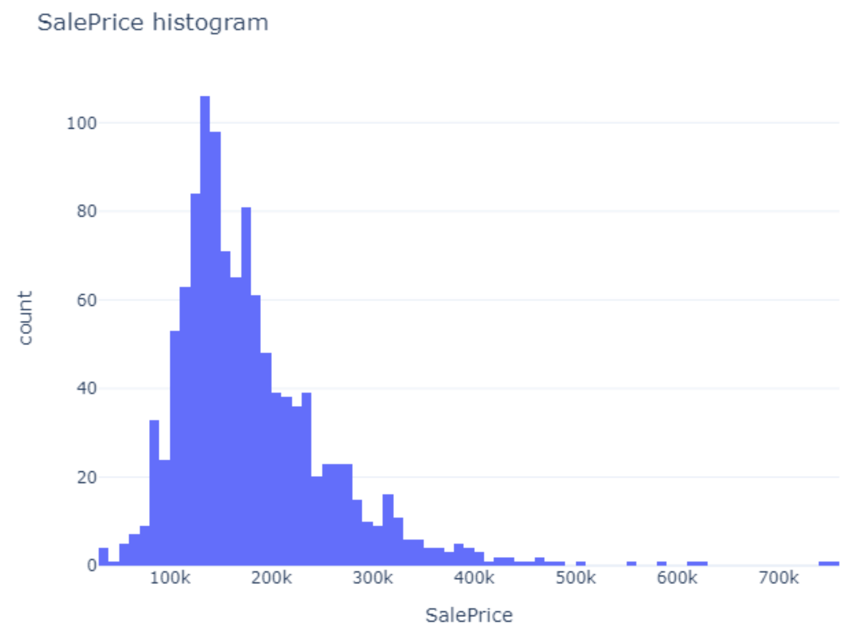
ExtraTreesRegressor with HyperParameter Tuning.

Key Metrics for success in solving problem under consideration

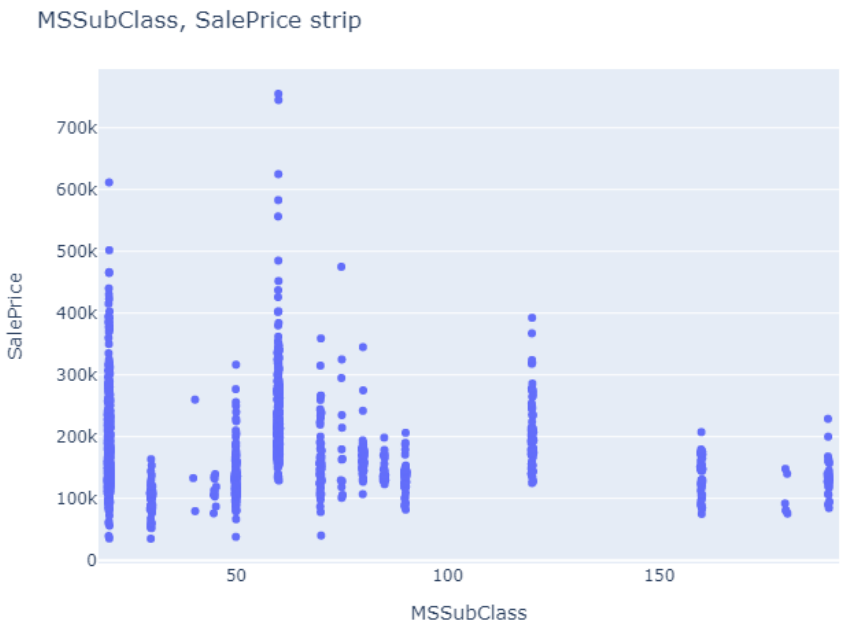
The Key Metrics used in solving the problem are:

- R2 Score
- Cross-Validation Score
- MSE
- RMSE

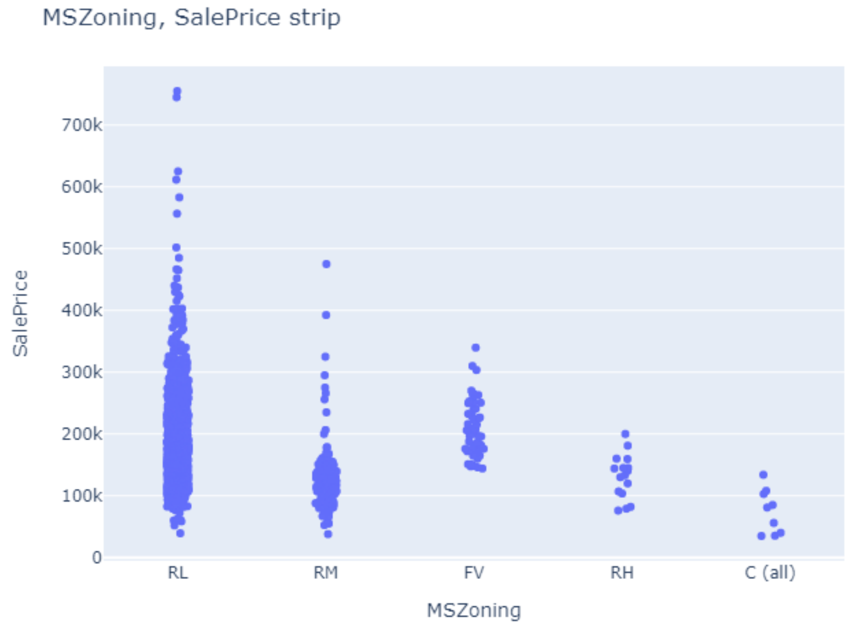
Visualizations:



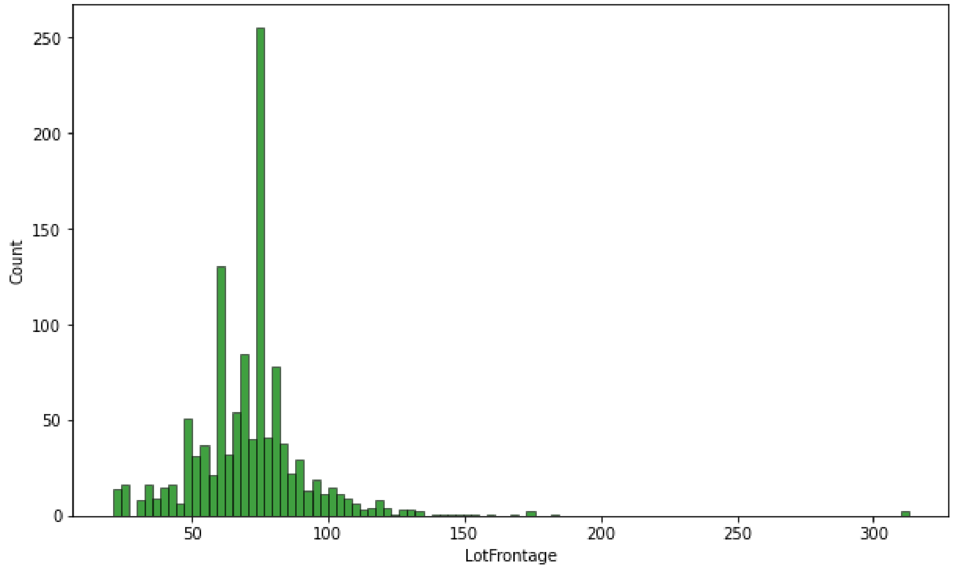
- We see that the majority of the houses sold are in the range 100-200k.



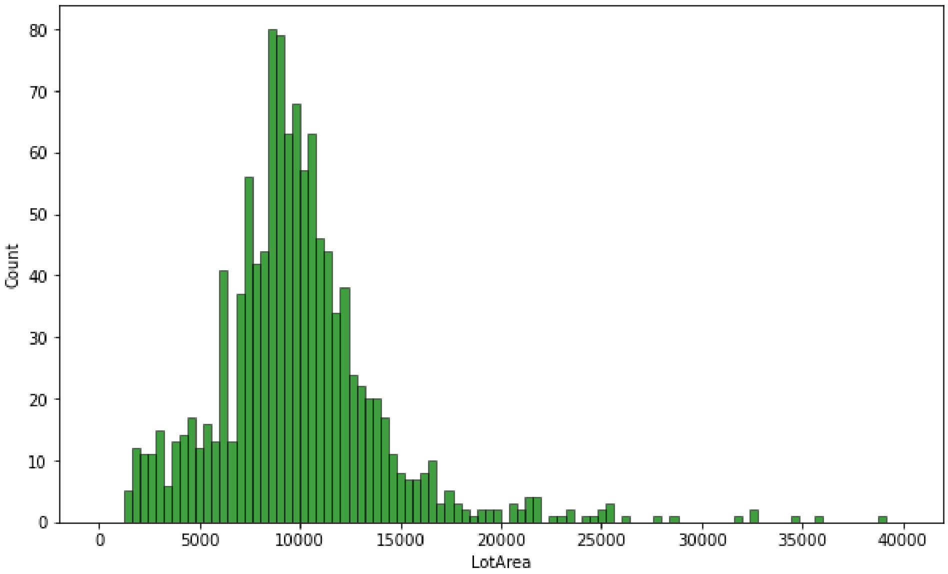
- We see that majority of the houses sold are class 20 and class 60(1-STORY 1946 & NEWER ALL STYLES and 2-STORY 1946 & NEWER ALL STYLES)
- We also see class 20 and class 60 fetch the highest prices.
- class 30 1-STORY 1945 & OLDER are sold the cheapest



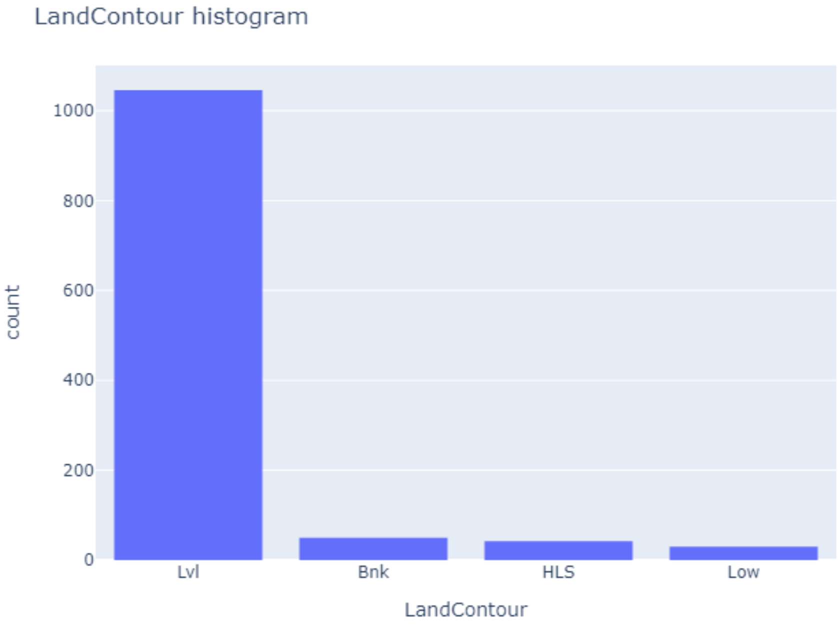
- ➢ We see that the majority of the houses sold were in Residential Low-density area and the least in commercial.
- ➢ We also see that the prices for Residential Low Density and Floating Village Residential were on higher end.
- ➢ Residential Medium Density and Commercial properties look to be on cheaper end.



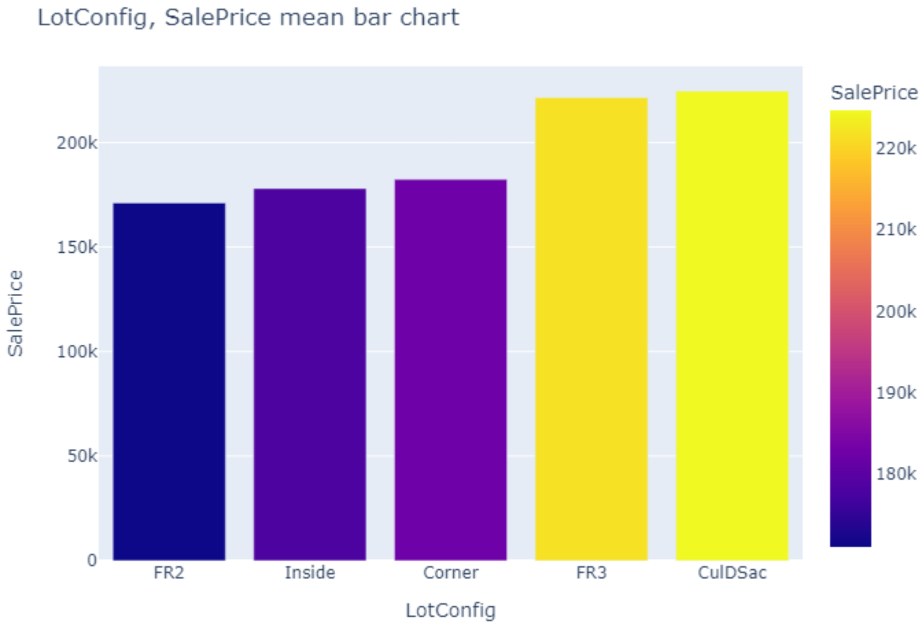
- ➢ Most of the properties seem to have a LotFrontage between 50 to 80.



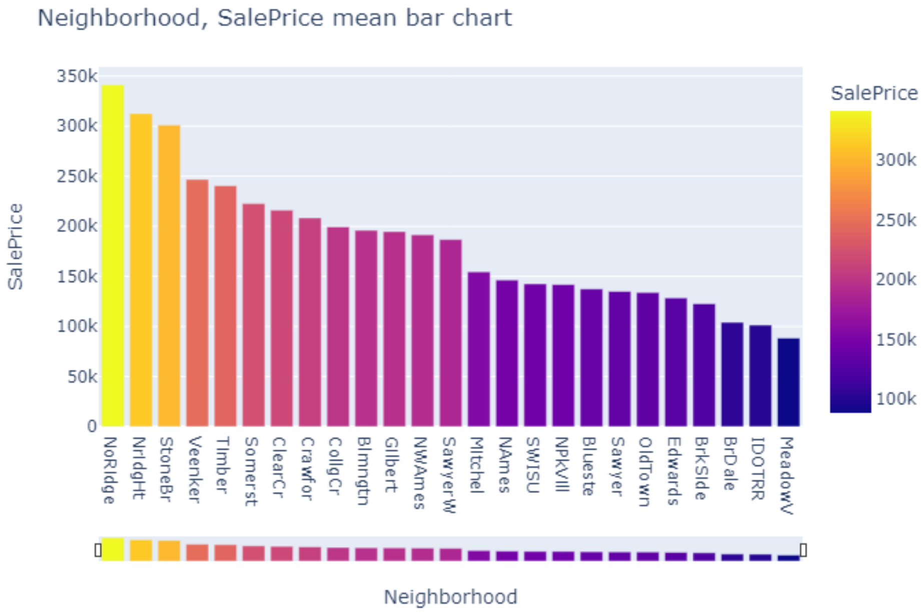
- ➢ We see majority of the properties tend to have a Lot Area between 7k to 13k.



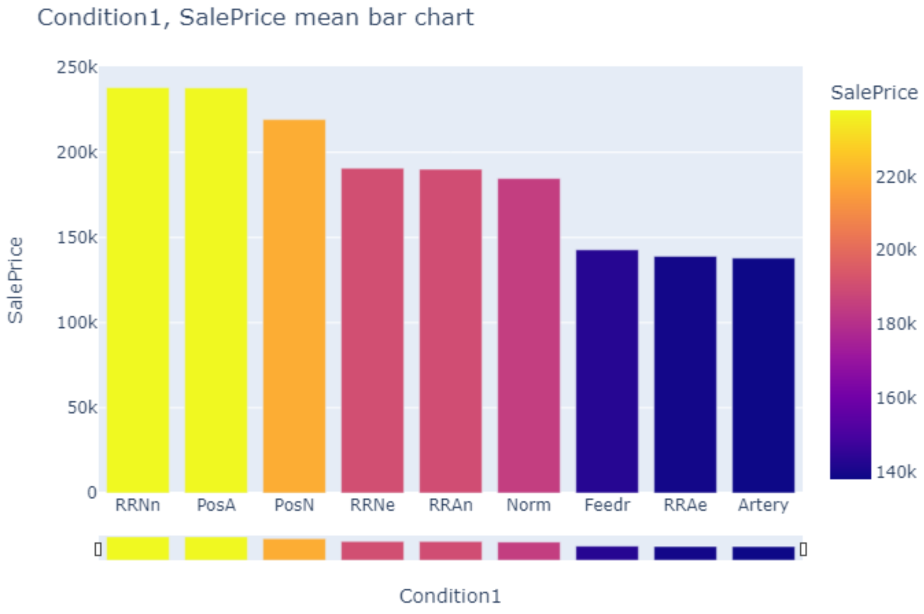
- ➢ We see the highest number of houses sold were Near Flat/Level contour.



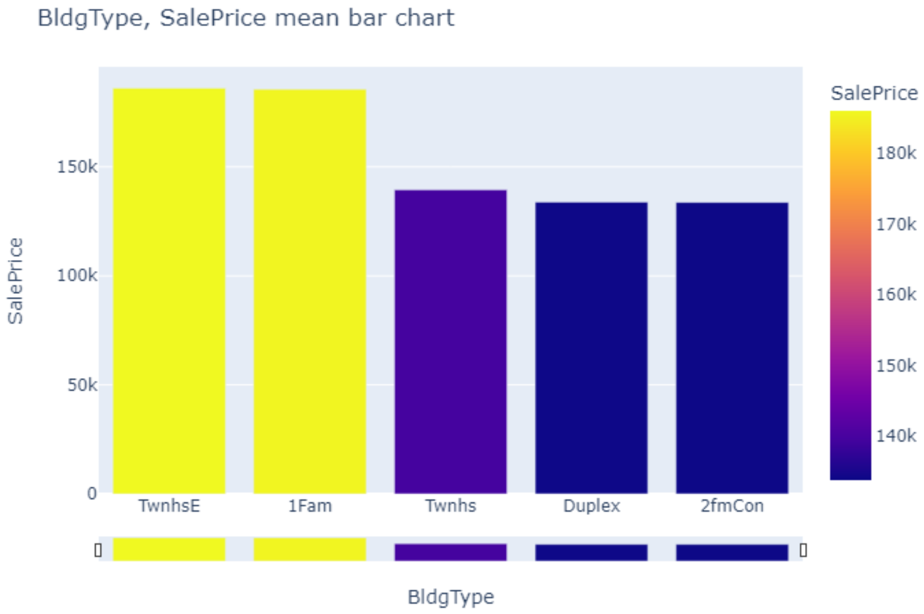
- We see the prices for Cul d sac, the property at the end of the road has a highest average price, followed by FR3(Frontage on 3 sides of property).
- FR2(Frontage on 2 sides of property) has the least average price.



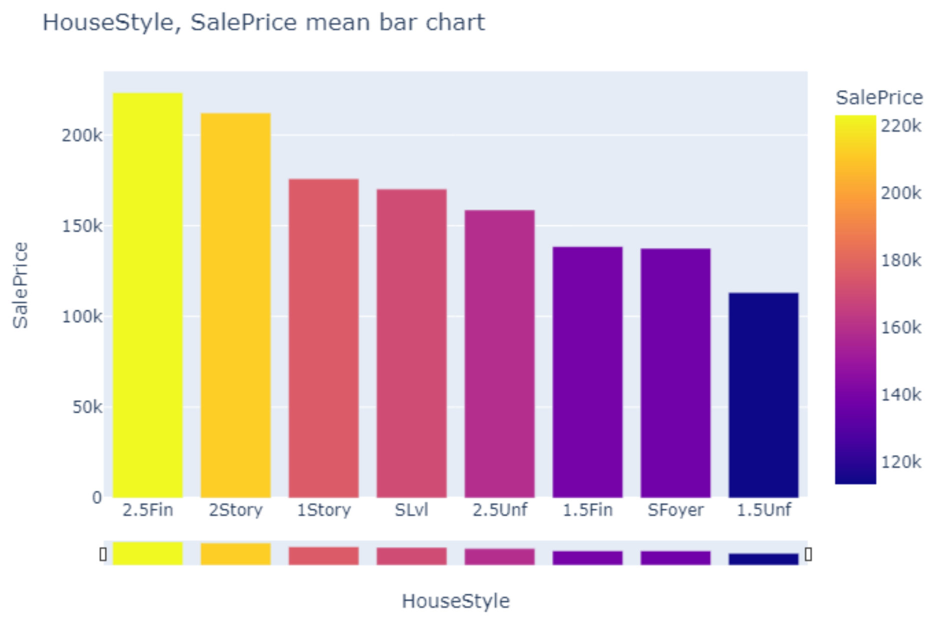
- Northridge followed by Northridge Heights and Stone Brook, have the highest prices.
- Meadow Village and Iowa DOT and Rail Road seem to have the least prices.



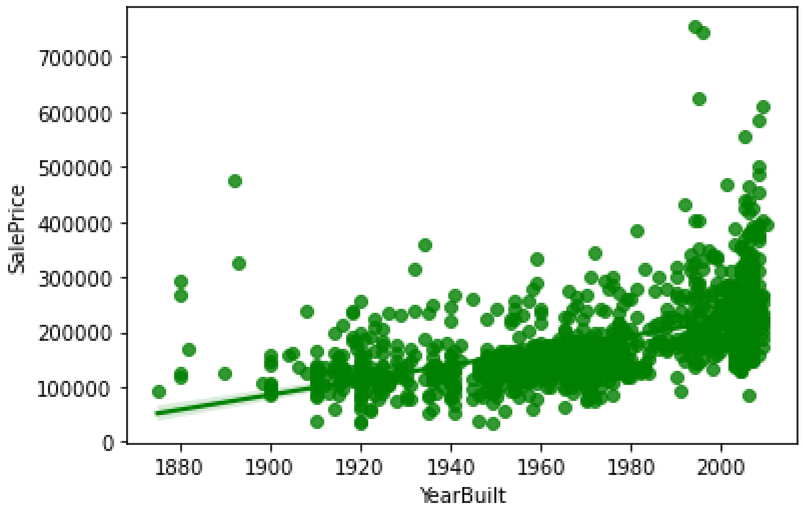
- ➢ Properties Within 200' of North-South Railroad and properties Adjacent to positive off-site feature have the highest price.
- ➢ Properties Adjacent to arterial street are cheaper.



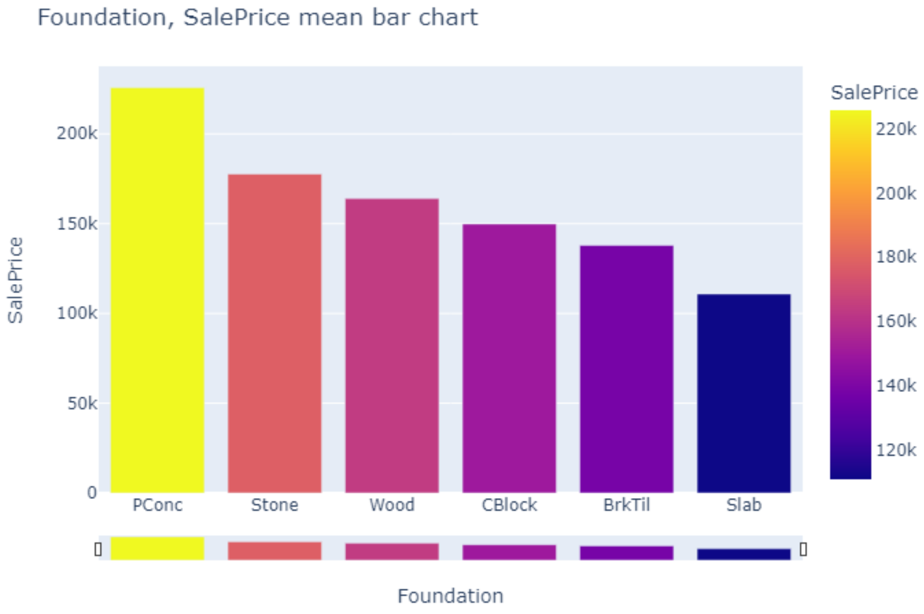
- ➢ Prices for Townhouse End Unit and Single-family Detached are Higher.
- ➢ Two-family Conversion; originally built as one-family dwelling prices are the cheapest.



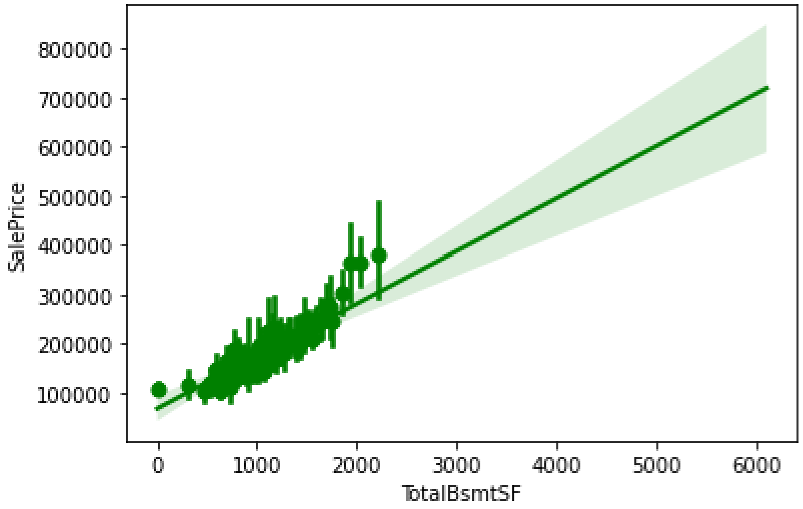
- ➢ Two and one-half story: 2nd level finished and two story have the highest price.
- ➢ One and one-half story: 2nd level unfinished has the lowest.



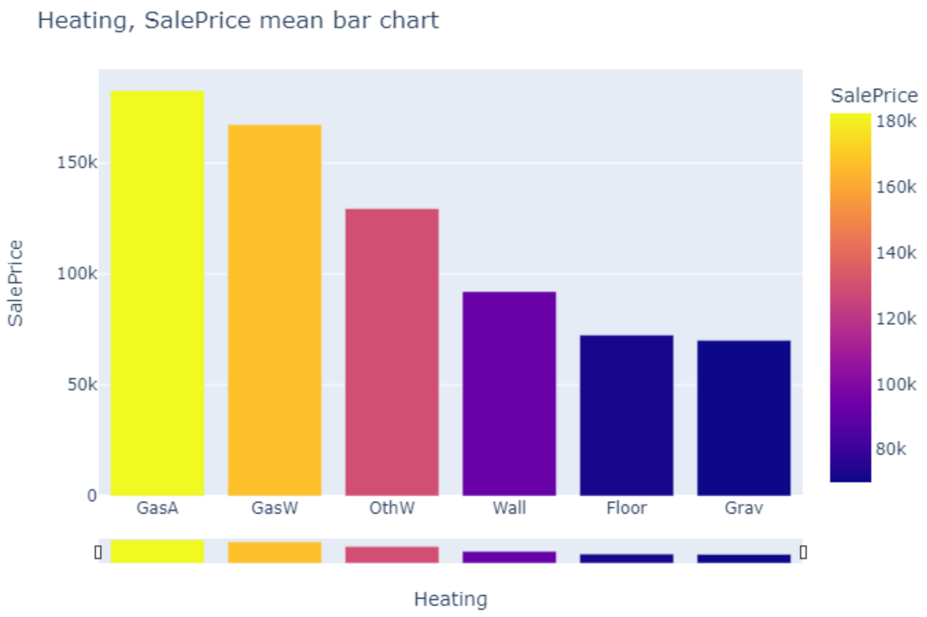
- ➢ From above, we see the new the house the higher the price.



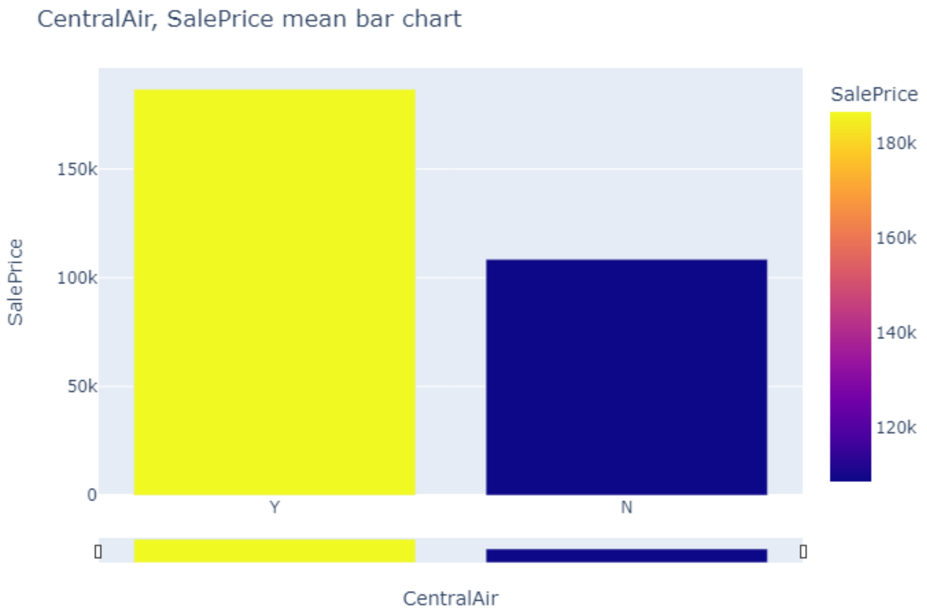
- ➢ Poured Concrete have the highest price and Slab the lowest.



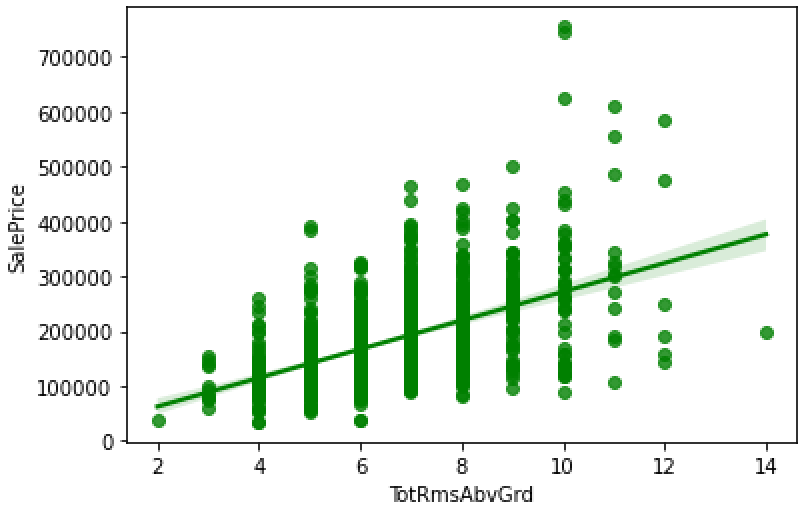
- ➢ We see that as the Basement area increases, so does the price.



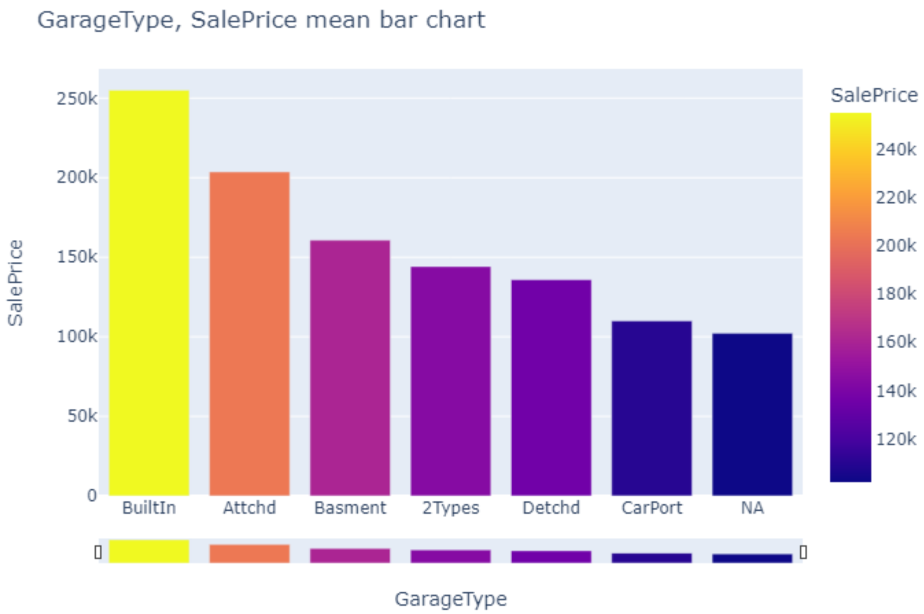
- Houses powered by Gas forced warm air furnace tend to have highest price.
- Homes with Gravity furnace the least.



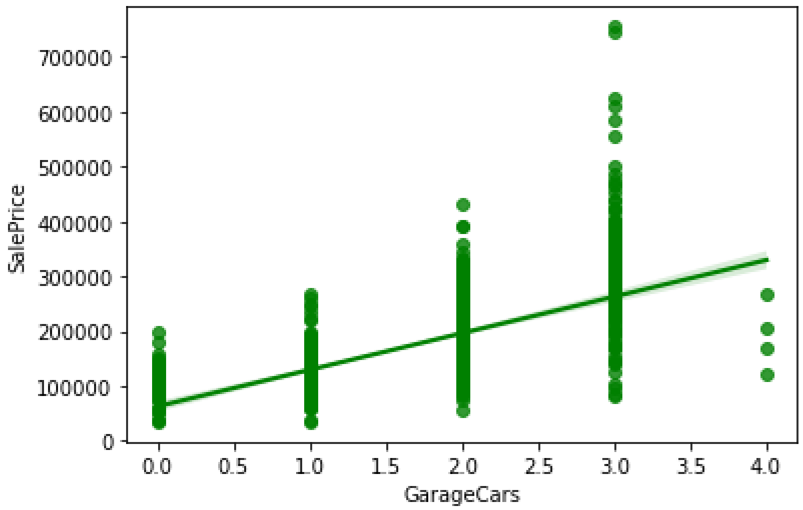
- We see that the houses with Central Air Conditioning have a significantly higher price compare to the once without.



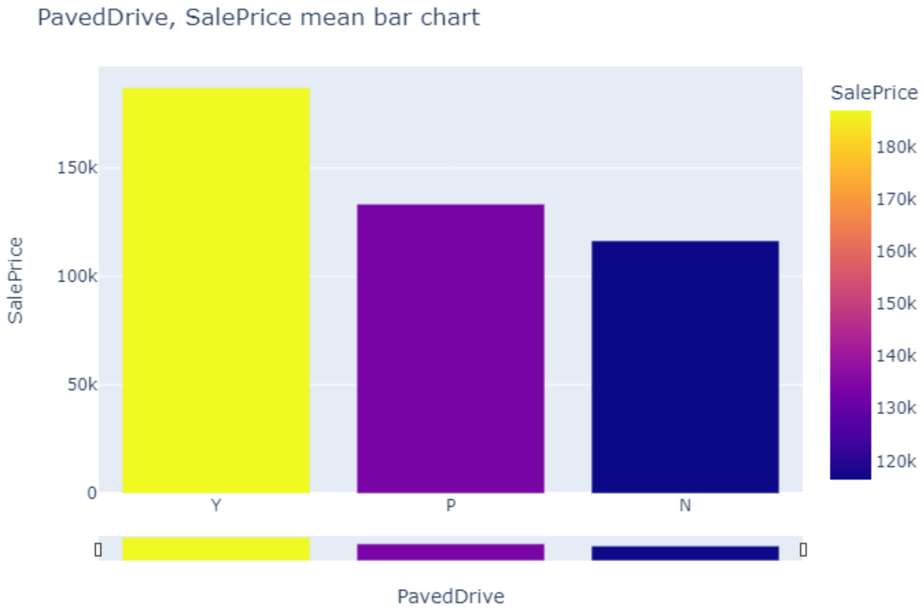
- We see TotalRmsAbvGrd and SalesPrice are linearly related, as the TotalRmsAbvGrd increase, so does the SalesPrice.



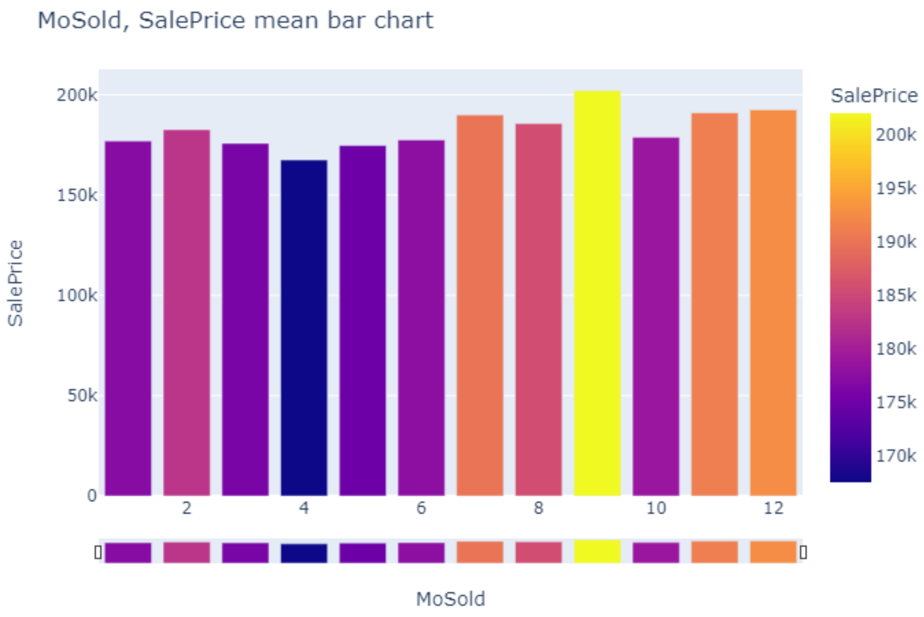
- Built-In (Garage part of house - typically has room above garage) Garage tend to fetch the highest price and Car Port the lowest.



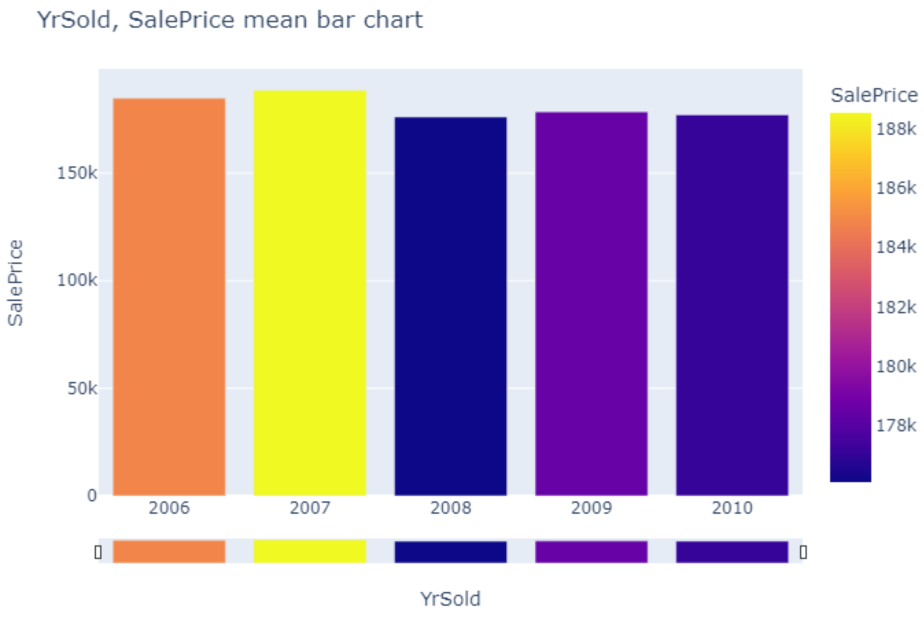
- We see as the number of car parking's increase, so does the price.



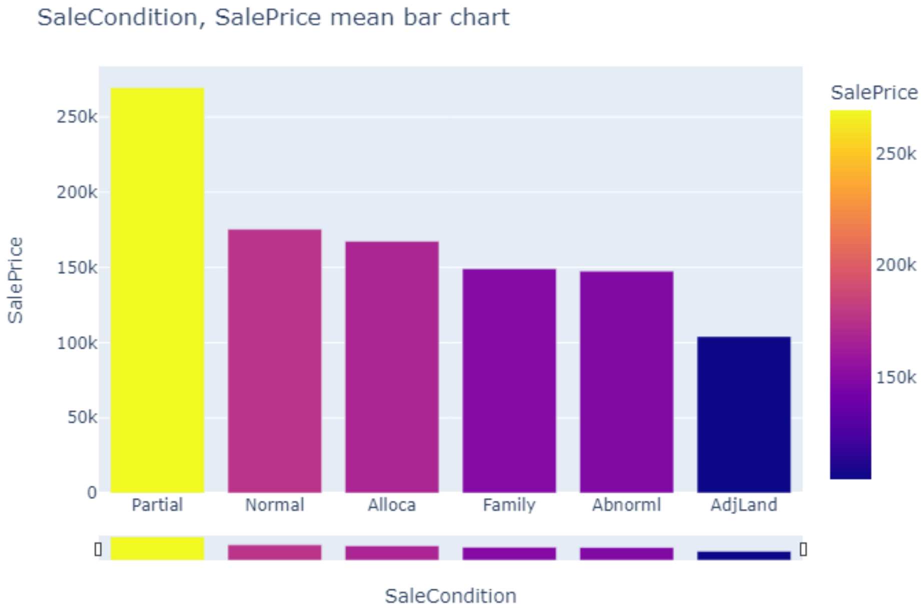
- We see the houses with Paved drive have a higher price than those without it.



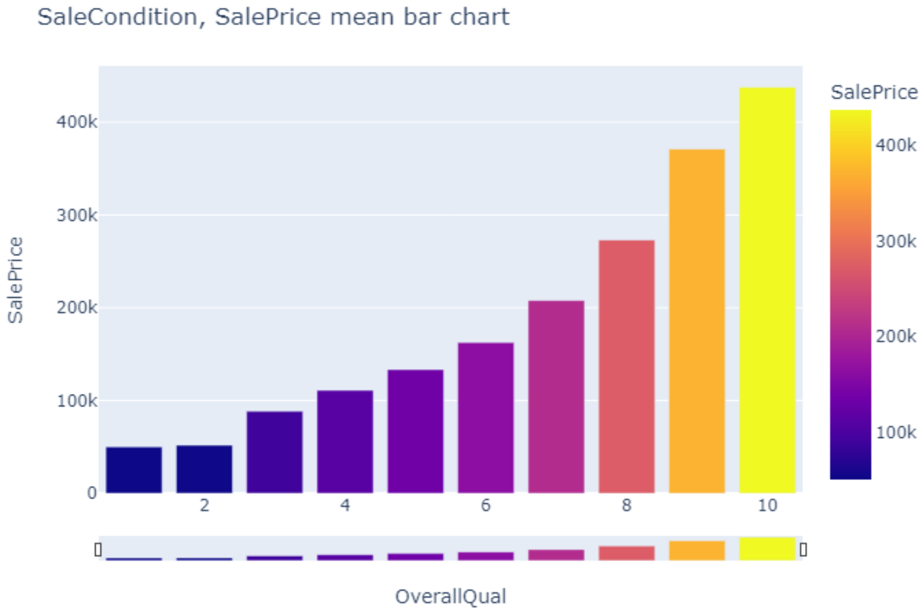
- ➢ We see the houses sold in the month of September tend to have Higher Price and the houses sold in April lower prices.
- ➢ We also see the houses sold in first 6 months of the year have in general lower price than the once sold in last 6.



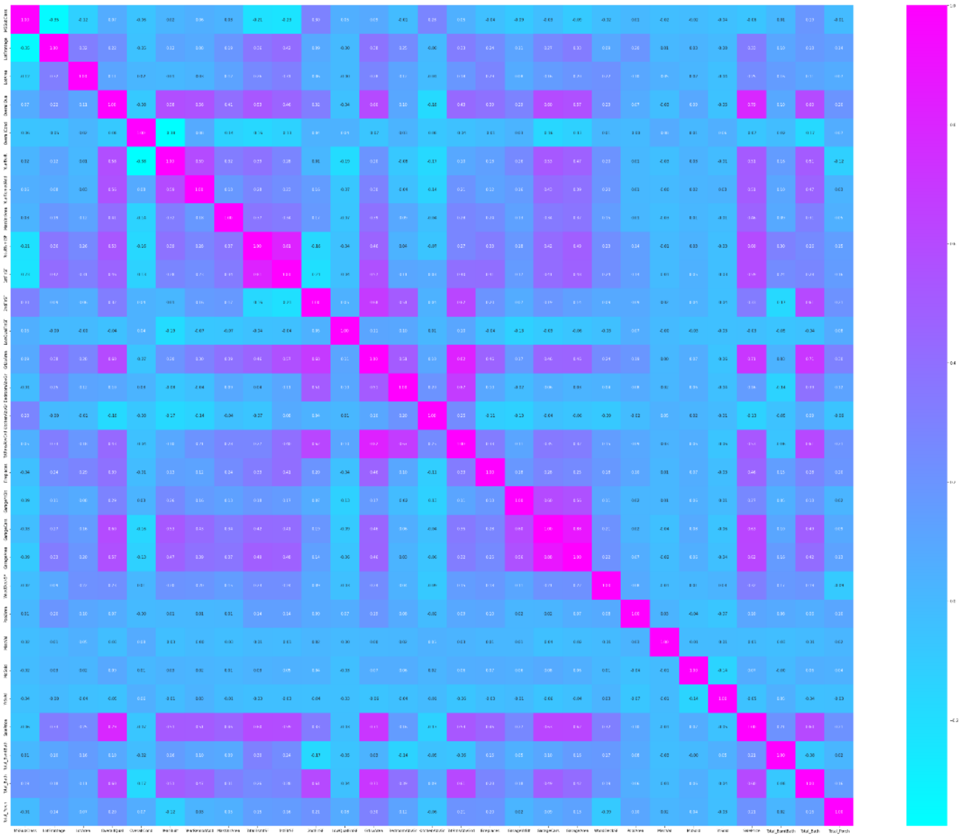
- ➢ We see that the prices for the houses were the lowest in 2008 the major reason being global financial crisis (GFC)



➢ We see that the Partial (Home was not completed when last assessed (associated with New Homes)) tend to cost the highest and Adjoining Land Purchase the least.



➢ We see that as the overall quality increases, so does the Price



- ➢ We see that the features have both +ve and -ve correlation with the Target variable.
- ➢ Weakest Correlation
  - OverallCond -0.065642
  - YrSold -0.045508
  - LowQualFinSF -0.032381
  - MiscVal -0.013071
  - MoSold 0.072764
- ➢ Strongest Correlation
  - GrLivArea 0.707300
  - OverallQual 0.789185

Interpretation of the Results

We have achieved an accuracy of 82.4% which is pretty good with a difference in cross-validation of 1.64%.

CONCLUSION

Key Findings and Conclusions of the Study

We have been able to build a model with 82% accuracy. If we had a dataset that was larger than the one provided, it would have helped us build a better Model.

Learning Outcomes of the Study in respect of Data Science

We needed more data to analyse and train the machine even better. Price was indicative based on the specifications. But having limited number of records to analyse which is indicative of under learning of the datapoints by the machine. So better the data records better the learning and training of the model. Model performs well with the most trained records. The data pre-processing was very important to get proper data for model building and prediction.

Limitations of this work and Scope for Future Work

The main limitation of our work is less data to analyse and train the machine. The Actual price and predicted price are having some difference which may lead to underfitting. This may be due to the training of model with less records and more features. So, in future if we need better performance from the model, we would need a dataset with more records so that the model can give better outcomes.