PROJECT NAME:

HOUSING PRICE PREDICTION

Submitted by:
LALBIAK ZAUVA

INTRODUCTION

- **Problem framing:**

  Houses are one of the necessary needs of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world's economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases.

  Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for 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. The company is looking at prospective properties to buy houses to enter the market.

  My task is 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:

  1. Which variables are important to predict the price of variable?
  2. How do these variables describe the price of the house?

- **Conceptual background of the domain problem:**

  1. **Overall Quality:** This is the most important feature which affects the price of the house. More the ratings (out of 10), more value is added to the sale price of the house.

2. **Gross Living Area**: This contributes the 2nd most and has positive correlation with the price. It is an important feature and contributes the most to the value of the house.
3. **Garage Cars/Area:** Most people give importance on the number of cars the garage can accommodate and this feature adds a huge value to the overall price.
4. **Total square feet of basement area:** The larger the basement area, more will be the price and this feature is among the top features which adds huge value to the house.
5. **Number of full bathrooms:** The larger the number of full bathrooms, more value will be added to the price.
6. **Open porch square feet:** Larger the area of the open porch, more will be the price of the house.

- **Analytical problem framing:**

1. **Mathematical/Analytical Modelling of the Problem:**
   To model the problem, we need to apply some mathematical and statistical tools to the data. So, in order to do this, we need to:
   a) Fill the missing values accordingly. If they are categorical, we can fill them with mode of the data. But in this project, applying our domain knowledge, I filled them with 'None' since filling null values to some rows with mode just does not make sense.
   b) Transforming the variables is the next step. Since the values are diverse which makes it difficult for our model to work on, I used Power Transform so that our data looks normally distributed. We need to have a normal distribution so that our model works best.
   c) Encoding the categorical data is the next step since the machine does not understand object type data and only understands numbers.
   d) We then need to scale the data and we are ready to build our model.

2. **Data sources and their formats:**

   The dataset is collected by a US-based company named Surprise Housing in a csv file format. The data looks something like:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Id	MSSubCla	MSZoning	LotFronta	LotArea	Street	Alley	LotShape	LandCont	Utilities	LotConfig	LandSlope	Neighborh	Condition	Condition	BldgType	HouseStyl	OverallQu	OverallCo
127	120	RL		4928	Pave		IR1	Lvl	AllPub	Inside	Gtl	NPkVill	Norm	Norm	TwnhsE	1Story	6	5
889	20	RL	95	15865	Pave		IR1	Lvl	AllPub	Inside	Mod	NAmes	Norm	Norm	1Fam	1Story	8	6
793	60	RL	92	9920	Pave		IR1	Lvl	AllPub	CulDSac	Gtl	NoRidge	Norm	Norm	1Fam	2Story	7	5
110	20	RL	105	11751	Pave		IR1	Lvl	AllPub	Inside	Gtl	NWAmes	Norm	Norm	1Fam	1Story	6	6
422	20	RL		16635	Pave		IR1	Lvl	AllPub	FR2	Gtl	NWAmes	Norm	Norm	1Fam	1Story	6	7
1197	60	RL	58	14054	Pave		IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story	7	5
561	20	RL		11341	Pave		IR1	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	6
1041	20	RL	88	13125	Pave		Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	4
503	20	RL	70	9170	Pave		Reg	Lvl	AllPub	Corner	Gtl	Edwards	Feedr	Norm	1Fam	1Story	5	7
576	50	RL	80	8480	Pave		Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1.5Fin	5	5
449	50	RM	50	8600	Pave		Reg	Bnk	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Fin	6	6
833	60	RL	44	9548	Pave		IR1	Lvl	AllPub	CulDSac	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	6
277	20	RL	129	9196	Pave		IR1	Lvl	AllPub	Inside	Gtl	Mitchel	Norm	Norm	1Fam	1Story	7	5
84	20	RL	80	8892	Pave		IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	5
888	50	RL	59	16466	Pave		IR1	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1.5Fin	5	7
1013	70	RL	55	10592	Pave		Reg	Lvl	AllPub	Inside	Gtl	Crawfor	Norm	Norm	1Fam	2Story	6	7
1154	30	RM		5890	Pave		Reg	Lvl	AllPub	Corner	Gtl	IDOTRR	Norm	Norm	1Fam	1Story	6	8
728	20	RL	64	7314	Pave		Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1Story	7	5
270	20	RL		7917	Pave		IR1	Lvl	AllPub	Corner	Gtl	Edwards	Norm	Norm	1Fam	1Story	6	7
1105	160	RM	24	2016	Pave		Reg	Lvl	AllPub	Inside	Gtl	BrDale	Norm	Norm	TwnhsE	2Story	5	5

3. **Data Pre-processing:**

   a) Missing values are filled accordingly. If they are categorical, they are filled with ‘None’; and if they are continuous, they are filled with ‘0’.
   b) Irrelevant columns are dropped.
   c) A distribution plot is plotted to see how the data is spread and to see the skewness of each feature.
   d) Then the skewness is handled using power transform.
   e) The categorical columns are encoded using Ordinal Encoder.
   f) A correlation plot is made to see how much features are related to the target variable. But no columns are dropped in this case.

4. **Tools and libraries:**
   a) NumPy is imported to deal with numbers
   b) Pandas is imported so that we can work on tables and dataframes
   c) Seaborn is also used to do some visualizations
   d) PowerTransformer is used to do transformations on the data
   e) OrdinalEncoder is used to encode the categorical data
   f) StandardScaler is used to scale the data
   g) Sklearn and metrics tools are imported to build the model and make predictions, and also to do check overfitting of the model.
   h) Pickle is imported to save the model for later purpose

- **Model development & Evaluation:**

1. The dataset is imported in its raw format which is a csv file. It needs to be cleaned by filling the null values accordingly and then the skewness needs to be dealt with using appropriate transformations on the data. Once the data is cleaned and skewness is minimized, some visualizations also help us understand which features are important and which are not.

2. Then, data needs splitting into train and test dataset. Out of different algorithms like XGB Regressor, Linear Regression, **XGB Regressor** is preferred to build our model and make predictions since it gives the

best accuracy and is the most consistent with the train and test dataset.

3. XGBoost:
   a) The main advantage of XGBoost is its lightning speed compared to other alogorithms.
   b) It uses parallel processing to increase its speed.
   c) XGBoost also handles missing values in the dataset.
   d) Using this algorithm, the accuracy score is 89.61%.

```
x_train,x_test,y_train,y_test = train_test_split(x_scaled,y,test_size = 0.2, random_state = 14)
xgb.fit(x_train, y_train)
```

```
XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
             colsample_bynode=1, colsample_bytree=1, enable_categorical=False,
             gamma=0, gpu_id=-1, importance_type=None,
             interaction_constraints='', learning_rate=0.300000012,
             max_delta_step=0, max_depth=6, min_child_weight=1, missing=nan,
             monotone_constraints='()', n_estimators=100, n_jobs=12,
             num_parallel_tree=1, predictor='auto', random_state=0, reg_alpha=0,
             reg_lambda=1, scale_pos_weight=1, subsample=1, tree_method='exact',
             validate_parameters=1, verbosity=None)
```

```
y_test_pred = xgb.predict(x_test)
print(f"The accuracy score is {r2_score(y_test,y_test_pred)*100:.2f} %")
```

```
The accuracy score is 89.61 %
```

Linear Regression:

a) It is a supervised ML algorithm to predict continuous data/label. Linear Regression is one of the most fundamental algorithms in the Machine Learning world.

b) Building blocks of a Linear Regression Model are:
   - Discreet/continuous independent variables
   - A best-fit regression line
   - Continuous dependent variable, i.e., a LRM predicts the dependent variable using a regression line based on the independent variables. The equation of the Linear Regression is:

$$Y = a + bX + e$$

```
x_train,x_test,y_train,y_test = train_test_split(x_scaled,y,test_size = 0.2, random_state = 3)
lr.fit(x_train, y_train)
```

```
▾ LinearRegression
LinearRegression()
```

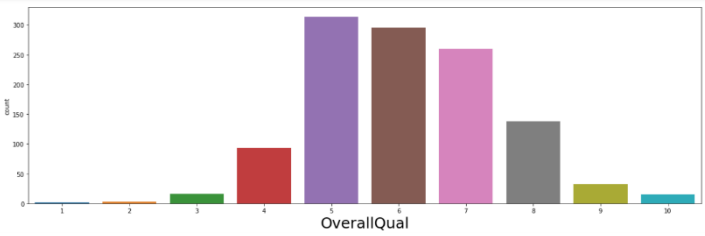
```
y_test_pred = lr.predict(x_test)
print(f"The accuracy score is {r2_score(y_test,y_test_pred)*100:.2f} %")
```

```
The accuracy score is 86.27 %
```

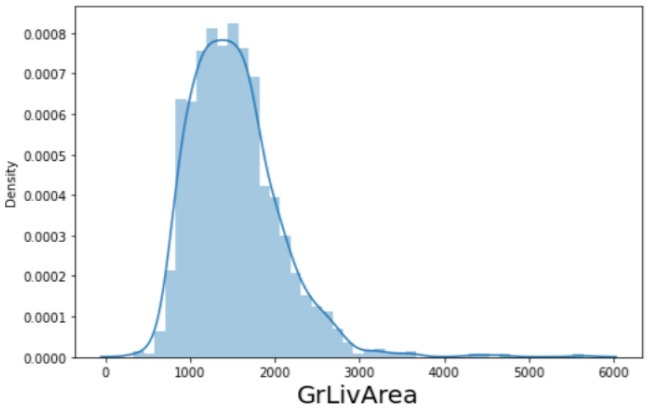
4. R2_score is used to find the accuracy of the prediction. It explains the proportion of the variance, i.e., the proportion of the variance in the observed data which the model explains, or the reduction in error over the null model. Its values lies between 0 and 1 and values closer to 1 means more variance is explained by the model, which then means better accuracy.

   Since our problem is a regression problem, we can only compute R2_score.

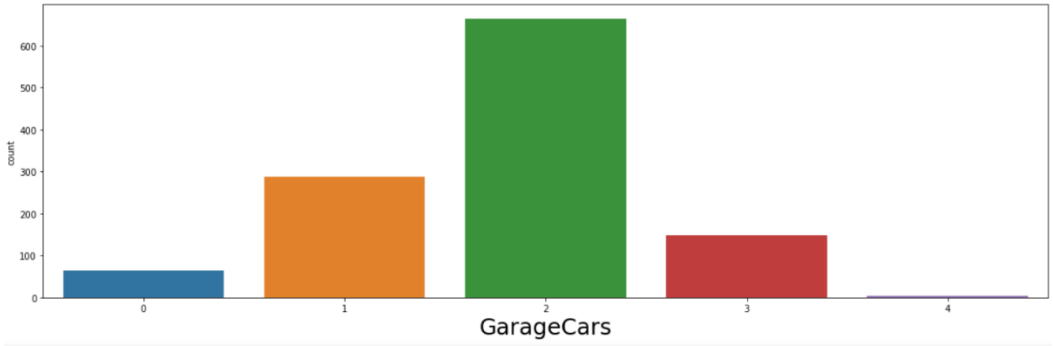
5. **Visualizations:**
   a) Overall quality has the highest correlation with the Price, and most of the houses are rated 5 out of 10.



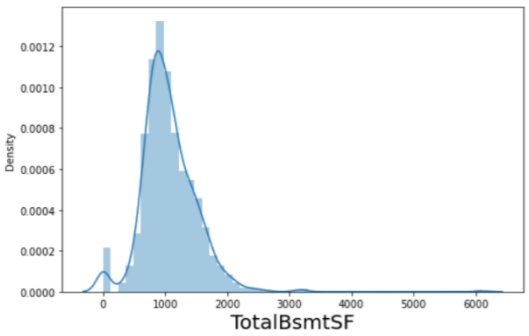
   b) Gross Living Area is the 2nd most important features having a positive correlation with the price.

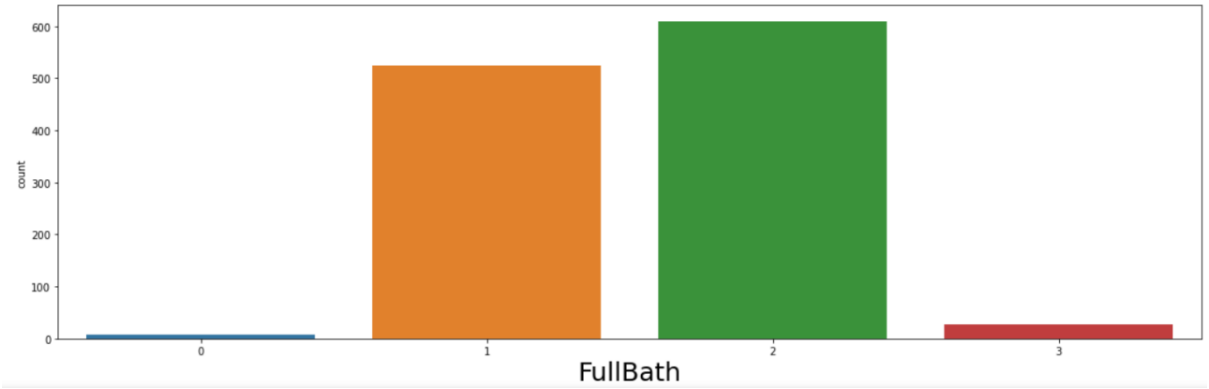


   c) Number of cars that the garage can accommodate is at the 3rd most important feature. And most of the houses can accommodate 2 cars.



d) Total square feet of basement is also among the top features. Larger the value, more will be the price.
e) More is the number of full bathrooms, more value will be added to the price. And most of the houses has 2 full bathrooms.





6. **Interpretation of some important results:**

a) Although overall quality has a high correlation, most of the houses are not rated 10 out of 10, but 5, 6 and 7 out of 10.
b) Most houses do not have a very large gross living area. This may be because people do not want to spend extra money when they can live with a moderately large gross living area.
c) Most houses accommodate only 2 cars followed by 1, 3 and 0 cars. Only very few accommodate 4 cars. It would be best to invest in a house which can accommodate at least 1 car.
d) Many full bathrooms are actually not necessary for most households since most families can live with 1 or 2 full bathrooms.

Conclusion

From the analysis of the given dataset, we can conclude that the sale price of houses depends on various factors like – how big the house is, how many bathrooms it has, how big is the garage, does it has a fireplace, how large is the porch, etc. Among different algorithms, XGBoost Regressor works best for this dataset.

While working on the project, there are so many null values, so domain knowledge really helped me to fill the null values. We cannot just fill them with mode or mean since it will not make sense.

The accuracy could be improved and some parameter tuning could definitely improve the results by using appropriate parameters.