IIM
RANCHI
बहुमुखविकासो गन्तव्यः

Business Statistics Assignment (MBA-BA)

Application of Regression in Machine Learning

GROUP MEMBERS-

ADITYA GURBAXANI	BA0004-21
JIMIT GANDHI	BA0011-21
RITESH RAI	BA0027-21
VYOM AGARWAL	BA0037-21

Table of Contents

Table of Contents

List of Figures

1. Introduction

Currently, Real Estate Industry is one of the very few industries that has not utilized the true potential of machine learning and modern mathematical tools to the fullest. Predicting the price of houses is highly complicated as many factors play a significant role. Therefore, it becomes essential to consider all the factors critically and understand their impact on house pricing. Regression is one of the crucial methods because it gives us more insight into this type of data.

Several factors come into the picture when we consider house prices. These factors or independent variables can have a significant impact on the house price (dependent variable). In the data, the major impacting factors considered for the house prices are the number of bedrooms, floors, total land size, house, bathrooms, and condition. These factors majorly decide the cost of the homes.

The factors mentioned are all independent of each other and significantly impact housing prices. The analysis is carried out by implementing the Linear Regression Model to fit the data and determine the relationship between the factors and house prices.

The dataset we used to formulate the equation for predicting the price is of Washington, USA. After cleansing of the dataset, we have more than 21000 observations to carry forward our regression analysis.

2. Linear Regression

2.1. Simple Linear Regression

Linear Regression formulated the relationship between two variables by fitting a linear equation to the observed data. One variable is considered an explanatory variable, the other a dependent variable. In our business problem, we attempt to model the relationship of the predicted value of the house with six different independent variables. These six factors/variables will be explained further in detail with their importance to this problem.

Now let us consider the relationship between house price and the living area (in sqft.). The Simple Linear Regression Model is defined as shown below.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Where,

$Y_i: Dependent\ Variable$

$\beta_0: Population\ Y\ Intercept$

$\beta_1: Population\ Slope\ Coefficent$

$X_i: Independent\ Variable$

$\varepsilon_i: Random\ Error\ term$

Applying simple linear Regression yields a regression line, as shown in Figure-1.

Python Code

```
sns.jointplot(x='sqft_living',y='price', kind='reg', data=df, height=8.27,
                  joint_kws = { 'color': 'green', 'line_kws': {'color':'red'}})
```

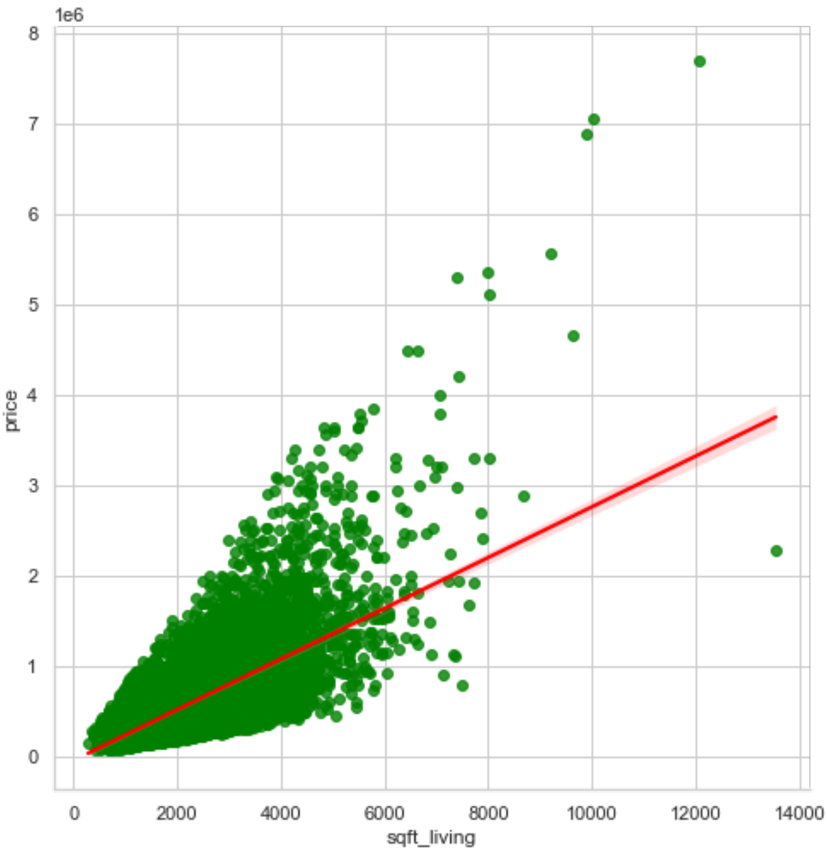

FIGURE-1: SCATTERPLOT OF HOUSE PRICES AND LIVING AREA WITH REGRESSION LINE

The scatterplot shown in the figure has house prices on the Y-Axis and the Living Area (in sqft.) on the X-Axis. The red line is the regression line obtained by fitting the data points into the Simple Linear Regression Model.

The Simple Linear Regression Equation or Prediction Line estimates the population regression line and is defined as below.

$$\hat{Y}_i = b_0 + b_1 X_i$$

Where,

$\hat{Y}_i: Estimated\ or\ Predicted\ Y\ value\ for\ i^{th}\ observation$

$b_0: Estimate\ of\ the\ regression\ intercept$

$b_1: Estimate\ of\ the\ regression\ slope$

$X_i: Value\ of\ X\ for\ i^{th}\ observation$

2.2.Multiple Linear Regression

Multiple linear Regression, also known simply as multiple Regression, is a statistical technique that uses various independent variables to predict the outcome of a dependent variable. The goal is to formulate the linear relationship between the independent variables and the dependent (dependent) variables. In our business problem, we used the six above-mentioned independent variables to calculate and formulate the equation.

The Multiple Regression Model examines the relationship between the dependent variable (Y) and two or more independent variables (X_k)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

Where,

β_0: Y-Intercept

$\beta_1, \beta_2, \ldots \beta_k$: Population slopes

$\varepsilon_i: Random\ Error$

The coefficients in this multiple regression model are therefore estimated using sample data. These estimated coefficients give us the Multiple Regression Equation with k independent variables.

$$\widehat{Y_i} = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki}$$

Where,

$\widehat{Y_i}: Estimated\ (or\ predicted)\ value\ of\ Y$

$b_0: Estimated\ Intercept$

$b_1, b_2, \ldots\ b_k: Estimated\ slope\ coefficients$

3. Machine Learning Model

The machine learning model used for the business problem is the Linear Regression. Linear Regression is included in the '*Linear Models*' package with the python '*scikit learn*' library. The Model is based on supervised learning and utilizes labeled data for training.

The Model performs the regression task, i.e., it considers the input variables and corresponding output variables. The Model then fits the data points into the multiple regression equation. This task yields multiple regression lines. To select the best fit line, the Model computes the residual sum of squares, denoted as r^2 (r squared value). It selects the line having minimum r^2 value as the final regression line.

In Figure-1, we can see the solid red line and some areas near the red line shaded with red colour. The red colour line is the best fit line, i.e., the line having minimum r^2 value. The shaded red area is all the other possible regression lines computed by the Model. While Figure-1 shows the visualization for just one factor, we will be considering multiple factors in the Model for predicting the house prices.

4. Business Problem

4.1. Defining the problem

Predicting the price of a house is complicated in Real Estate Industry. There are many factors due to which the price of a house varies. Therefore, it becomes difficult for both retailers and customers to determine the price. This statistical analysis aims to help us understand the relationship between the functions of the house and how these variables are used to predict the price of the house.

4.2. Analysis of factors

We start by identifying the significant factors that need to be considered while deciding the house prices. Out of the various factors that may impact house prices, we have identified six key factors that shall be considered for the analysis. We have carried out an analysis for every factor to deeply understand their role in determining the house prices.

4.2.1. Number of Bedrooms

One of the factors for predicting the price of a house is the number of rooms in that house.

Python Code

```
sns.boxplot(x='bedrooms', y='price', data=df)
```

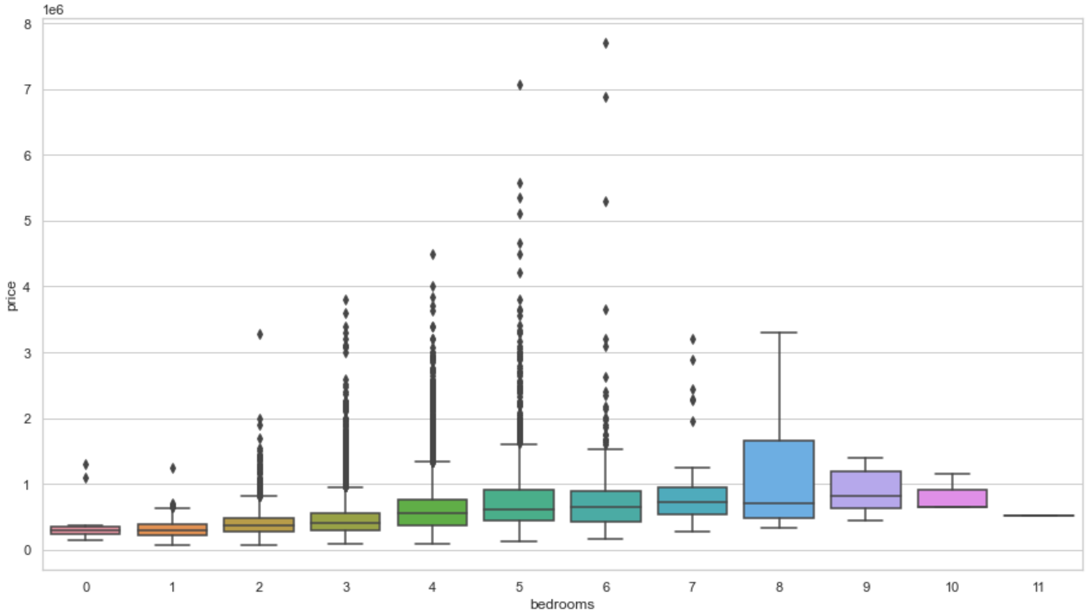

FIGURE-2: BOXPLOT SHOWING THE RANGE OF HOUSE PRICES FOR DIFFERENT NUMBERS OF BEDROOMS

From the plotted boxplot showing number of bedrooms vs. housing price, the following conclusion can be derived:

- A gradual rise in prices with an increase in the number of bedrooms
- The range of bedrooms in our dataset varies from 0 to a maximum value of 11
- The price range also varies a lot when we have more than five bedrooms in the house and similarly very less variation in case of a smaller number of bedrooms.

4.2.2. Number of Bathrooms

Another factor that plays a role in predicting the price of the house is the number of bathrooms.

Python Code

```
sns.boxplot(x='bathrooms', y='price', data=df)
```

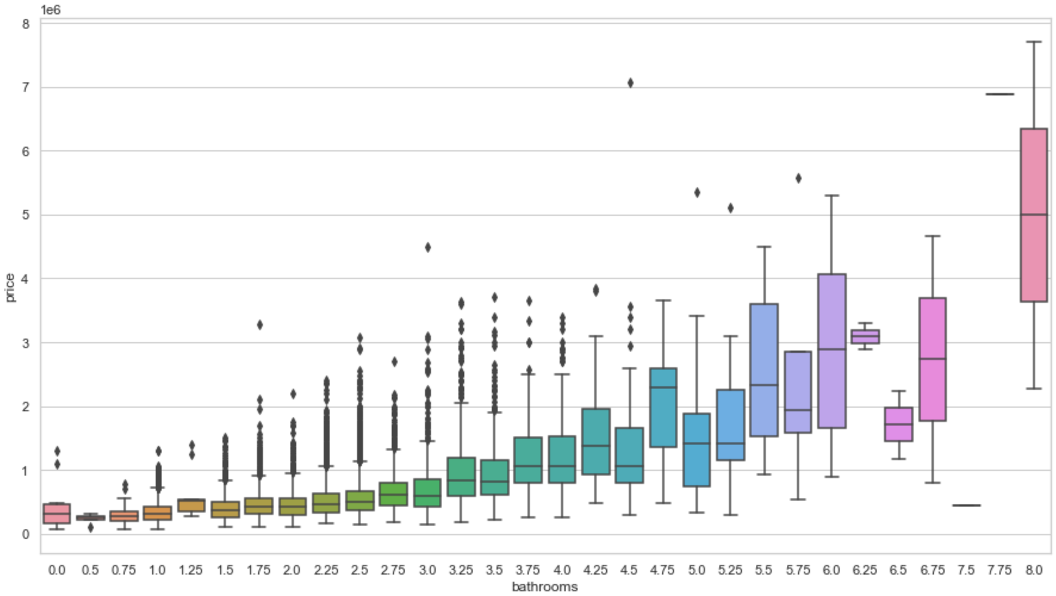

FIGURE-3: BOXPLOT SHOWING THE RANGE OF HOUSE PRICES FOR DIFFERENT NUMBERS OF BATHROOMS

From Figure-3, we can analyze and understand how the price of a house differs with a different number of bathrooms in a house. In our dataset, we have houses with 0 bathrooms to houses with eight bathrooms. With the increase in the number of bathrooms, we can notice a gradual increase in the price of the house. Also, we can see that the house's price is less when the number of bathrooms is less in a house, but deviation tends to increase with an increase in the number of bathrooms.

4.2.3. Living Area of the House

This factor includes the total area at which the house is built in the square feet area.

Python Code

```
sns.distplot(df['sqft_living'], color="green")
```

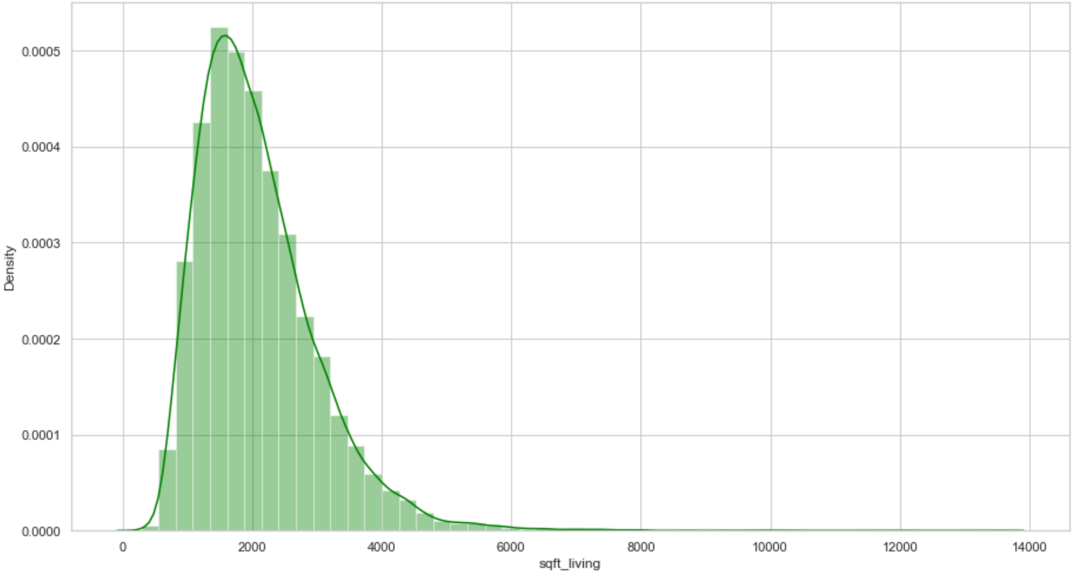

FIGURE-4: HISTOGRAM SHOWING THE DISTRIBUTION OF HOUSE LIVING AREA (IN SQFT)

From the Figure-4, we can analyze and understand the distribution of houses in terms of living area. Distribution is right-skewed with the maximum number of houses available are of 2000 sqft living area. Houses with more than 5000 sqft living are very few as they fall under the luxury category.

4.2.4. Total Area of the House

This factor includes the total area of the property, including the house, lawn, etc., in the square feet area. The range of total square feet is available in the dataset is exceptional. Therefore, it is impossible to represent the whole data set in one single graph and get conclusive evidence from it.

4.2.5. Number of floors in the House

This factor comprises a number of floors present in the house. From Figure-5, we can analyze and understand how the price of a house differs with a different number of floors in a house. The range in our dataset for number of floors in the house varies 1 to 3.5 (where 3.5 signifies 3 complete floors with half terrace and half floor).

Python Code

```
sns.boxplot(x='floors', y='price', data=df)
```

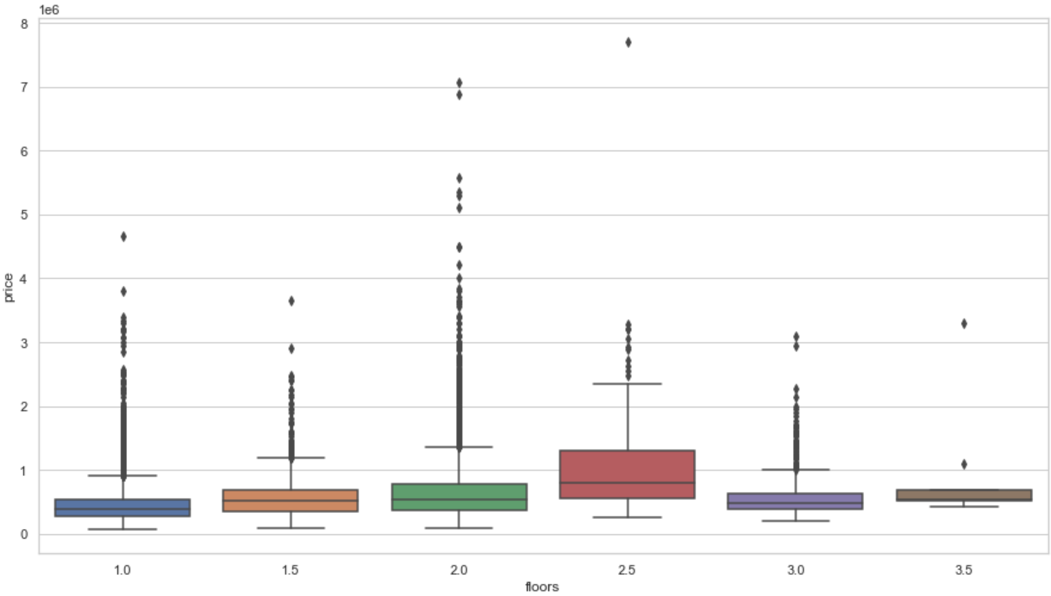

FIGURE-5: BOXPLOT SHOWING THE RANGE OF HOUSE PRICES FOR THE DIFFERENT NUMBER OF FLOORS

Although the maximum number of houses lie in 1 floor category followed by 2 floor houses but deviation in housing prices is most in case of 2.5 floor houses. Number of floors in a major factor when deciding house prices.

4.2.6. Condition of house

This is the rating of current condition of the house. Each house is rated on a scale from 1 to 5 according to its condition. From the Figure-6, we can also analyze and understand how the price of a house differs with the condition. The condition of a house plays a significant role in predicting the price of a house irrespective of other variables.

We can notice the increase in the price of the house with better house condition. The condition of houses varies from 1 to 5, with 1 being the lowest rating for house condition and 5 being the highest. The majority of houses lie in 3rd condition followed by 4th.

Python Code

```
sns.boxplot(x='condition', y='price', data=df)
```

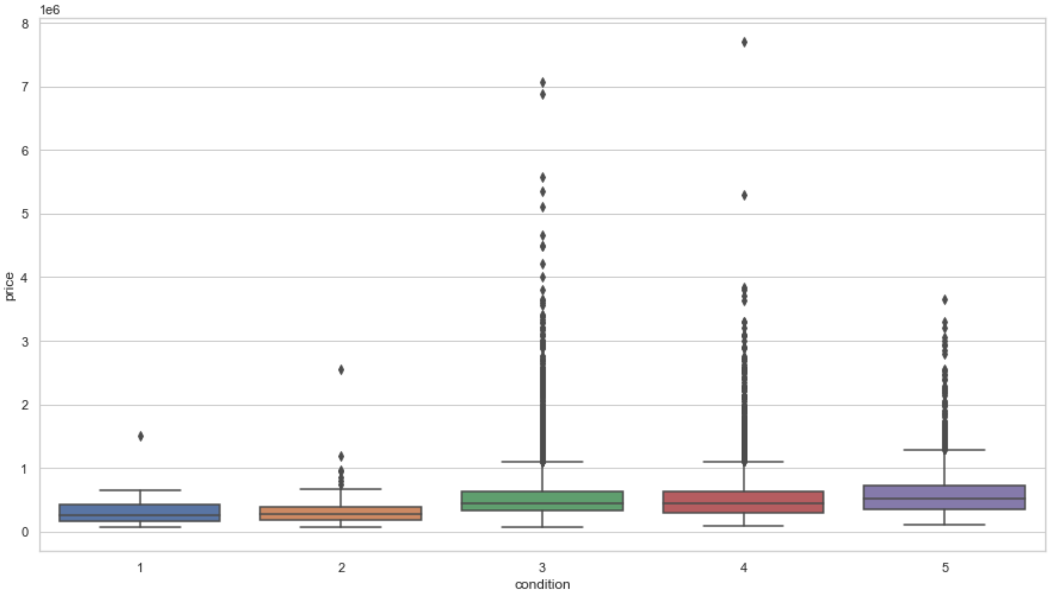

FIGURE-6: BOXPLOT SHOWING THE RANGE OF HOUSE PRICES FOR THE DIFFERENT CONDITION RATINGS

5. Applying Regression Model

The regression model has been applied using python libraries- *pandas, NumPy, matplotlib, seaborn & sklearn*. The *pandas'* library provides us the interface to read and manipulate the data. The *NumPy* library has been used for some matrix manipulations. The *matplotlib* and *seaborn* libraries have been used to generate visualization in the form of charts. The *sklearn or sci-kit learn* library contains the Linear Regression model, which has been used for fitting the data.

Python Code

```
import pandas as pd
import NumPy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

We start by reading the data from CSV (Comma Separated Values) file and load the data into a Pandas Data Frame. We list down the different columns and find out the size of the data.

Python Code

```
df = pd.read_csv('./final_data.csv')
print(df.columns)
print(df.shape)
```

Next, we have analyzed the data. We create some visualizations to understand the range and distribution of the values for the independent variables.

Python Code

```
df.drop(['price'], axis=1).hist(figsize=(15,20))
```

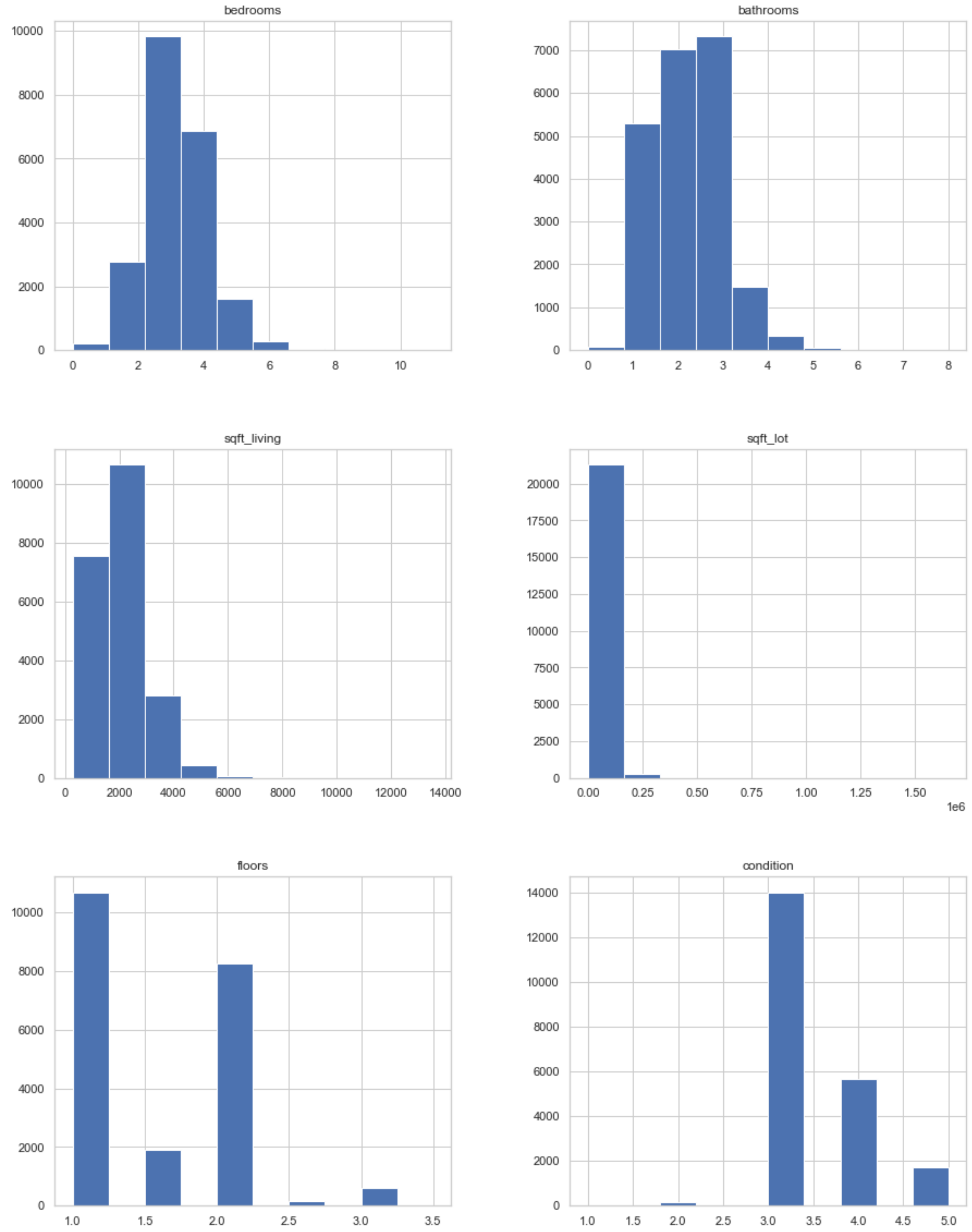

FIGURE-7: RANGE AND DISTRIBUTION OF VALUES FOR ALL THE INDEPENDENT VARIABLES

Next, we examine the correlation between the different columns. We plot the correlation coefficient using the heat map.

Python Code

```
# Generate a mask for the upper triangle
mask = np.triu(np.ones_like(df.corr(), dtype=bool))
# Draw the heatmap for correlation matrix with the mask
sns.heatmap(df.corr(), mask=mask, cmap='Spectral', vmax=.3, center=0,
            square=True, linewidths=.5, cbar_kws={"shrink": .9})
```

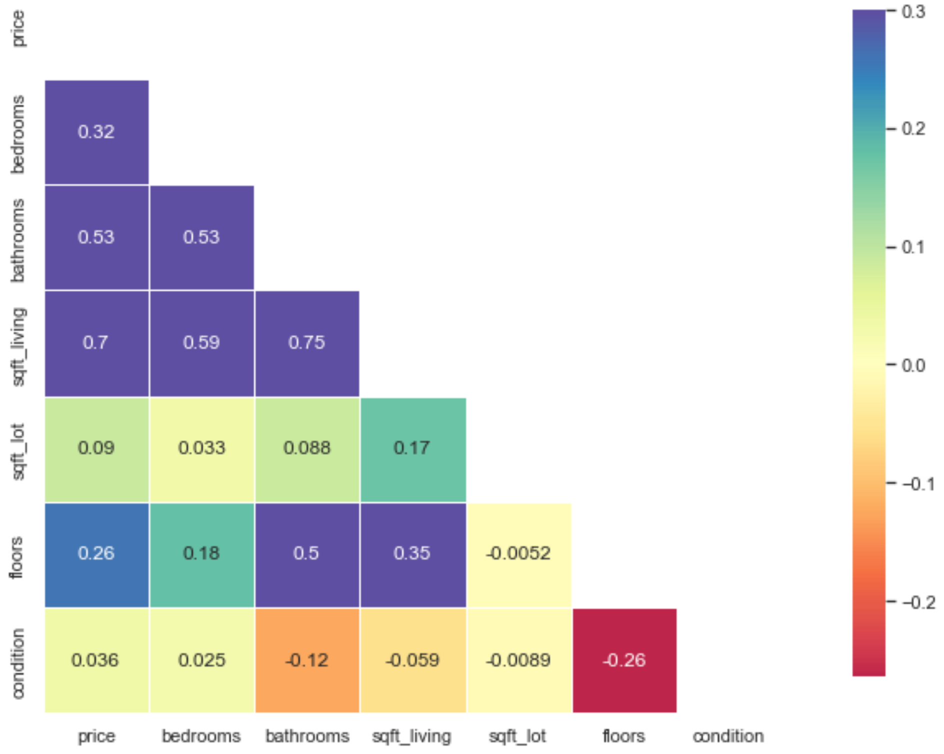

FIGURE-8: CORRELATION MATRIX

Next, we place all the independent variable columns in the X data frame and the dependent variable column in the Y data frame. An instance of the Linear Regression model is created, and the data is fitted into this Model.

Python Code

```
from sklearn.linear_model import LinearRegression
x = df.drop(['price'], axis=1)
y = df['price']
model = LinearRegression()
model.fit(x, y)
```

5.1. Coefficients and Intercept

Upon successfully fitting the data into the linear regression model, the intercept and coefficients were computed. These were then viewed on the python console.

Python Code

```
#Print the Intercept and Coeffecents
#Estimating the values upto 3 decimal places
print("b_%1d: %11.3f" % (0,model.intercept_))

for i in range(len(model.coef_)):
    print("b_%1d: %11.3f" % (i+1,model.coef_[i]))
```

Console Output

```
b_0: -108676.020
b_1:  -68318.358
b_2:   11424.827
b_3:     315.553
b_4:      -0.376
b_5:   14184.491
b_6:   53659.969
```

Regression Statistics

Multiple R	0.720075852
R Square	0.518509233
Adjusted R Square	0.518375516
Standard Error	254788.6864
Observations	21612

5.2. Testing the fitness of the Regression Model (F-test)

ANOVA Table

	df	**SS**	**MS**	**F**
Regression	6	1.51037E+15	2.51728E+14	3877.676283
Residual	21605	1.40254E+15	64917274707	
Total	21611	2.91291E+15		

Step-1: State Null Hypothesis and Alternative Hypothesis

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$
There is no relationship between house price and the 6 selected factors.

$H_a: Atleast\ one\ coeffficent \neq 0$
There is some relationship between the house prices and at least one of the 6 selected factors.

Step-2: Level of significance
$\alpha = 0.05$ (assumed)

Step-3: Find Critical value F_α
Numerator Degree of Freedom = 6, Denominator Degree of Freedom = 21605
$F_{0.05} = 3.092$

Step-4: Find the test statistic F_{STAT}
From ANOVA table, $F_{STAT} = 3877.676$

Step-5: Rejection Rule
$F_{STAT}(3877.676) > F_{0.05}(3.092)$, reject H_0

There is significant evidence to support that there is a relation between the house prices and the 6 selected factors, i.e., number of bedrooms, number of bathrooms, living area, total area, number of floors, and the condition rating of the house.

6. Results

As per the Regression Model output, the intercept and coefficient considered for predicting the house prices are shown below.

	Coefficients
Intercept	-108676.0202
bedrooms	-68318.35795
bathrooms	11424.8266
sqft_living	315.5532858
sqft_lot	-0.375502888
floors	14184.49105
condition	53659.96937

The multiple linear regression equation for this business problem is shown below.

$$
\begin{aligned}
\widehat{House\ Price} = {} & -108676.02 - 68318.358 * (\#\text{Bedrooms}) + 11425 * (\#\text{Bathrooms}) \\
& + 316 * (\text{Living Area}) - 0.37 * (\text{Total Area}) + 14185 * (\#\text{Floors}) \\
& + 53660 * (\text{Condition Rating})
\end{aligned}
$$

The model score is computed by the python console.

Python Code

```
#Getting the score for the Model created
model.score(x, y)
```

Console Output

```
0.5185092327321064
```

7. Conclusion

According to our regression analysis, the r^2 value obtained is 0.52. Hence it can be concluded that the price of the house is moderately dependent on the 6 selected variables i.e., number of bedrooms, number of bathrooms, living area, total area, number of floors, and the condition rating of the house. From the F test we carried out, we got significant evidence to support a relationship between the house prices and the 6 selected factors.

We have used machine learning (python programming) to analyze these 6 factors further and plot their relationship with housing prices. The machine learning model Linear Regression (from Scikit Learn Python Library) has been used.

Different graph plots are used, including box plot, histogram, Correlation matrix, etc for analysing the data. Box plot is used for understanding the variation of house prices with the number of bedrooms, number of bathrooms, number of floors, and condition of the house. The histogram is used in plotting the living area of the house concerning housing prices. We have also plotted all the independent variables to understand the range and distribution of the values, followed by the correlation between the different variables using a heat map.

8. References

(1) Predicting House Prices with Machine Learning - https://towardsdatascience.com/predicting-house-prices-with-machine-learning-62d5bcd0d68f

(2) Predicting House Prices with Machine Learning - https://medium.com/@manilwagle/predicting-house-prices-using-machine-learning-cab0b82cd3f

(3) Difference Between Algorithm and Model in Machine Learning - https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning/

(4) Predicting House Prices - https://www.kaggle.com/burhanykiyakoglu/predicting-house-prices/notebook

(5) Multiple Linear Regression- http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm

(6) Multiple Linear Regression- https://towardsdatascience.com/multiple-linear-regression-8cf3bee21d8b

(7) ML | Linear Regression- https://www.geeksforgeeks.org/ml-linear-regression/

(8) Linear Regression in Machine Learning Definition, Advantage & uses - https://www.mygreatlearning.com/blog/linear-regression-in-machine-learning/

(9) A Simple Introduction to ANOVA (with applications in Excel) - https://www.analyticsvidhya.com/blog/2018/01/anova-analysis-of-variance/

(10) Python Seaborn Library- https://seaborn.pydata.org/index.html

(11) Python Numpy Library- https://numpy.org/doc/stable/user/basics.rec.html

(12) Python Pandas Library- https://pandas.pydata.org/pandas-docs/stable/index.html

(13) Python MatPlotLib Library- https://matplotlib.org/stable/contents.html

(14) Python Scikit Learn Library, Linear Regression Model- https://scikit-learn.org/stable/modules/linear_model.html

9. Appendix

(1) Jupyter Notebook contain python codes and outputs
https://github.com/adityagurbaxani/py-house-prices/blob/main/BS_Assignment_MLR.ipynb

(2) Data File (CSV format) – 21612 records
https://github.com/adityagurbaxani/py-house-prices/blob/main/final_data.csv