IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

Prediction of House Price Using Linear Regression

[1]Siddharth Tomar, [2]Sunny Arora, [3]Yatharth Khansali, [4]Rahul Yadav, [5]Shubham Kumar

[1]Undergraduate Student, [2]Assistant Professor, [3]Undergraduate Student, [4]Undergraduate Student, [5]Undergraduate Student
Dept. Information Technology
Dr. Akhilesh Das Gupta Institute of Technology & Management (Affiliated to Guru Gobind Singh Indraprastha University),
Delhi, India.

Abstract: House price trend is important to predict housing prices without any biasness to help both the buyers and sellers for a better decision. Hence, house price trends are not only a concern for buyers but also for sellers. It also indicates the current economic situation regarding housing. The relationship between real estate and the economy is an important motivating factor. This dataset is obtained from www.kaggle.com. Comparison of different features like 'Number of Bedrooms', 'Area of construction', various amenities like 'garage', 'basement', etc. Here, 'Linear Regression' has been used to train the dataset. The objective of this project is to find out efficient pricing of a house and to calculate accuracy of the prediction. This application will help customers to invest in an estate without approaching an agent. It also decreases the risk involved in the transaction.

Index Term **- House price, prediction, linear regression, real estate, economy.**

1. Introduction

House price prediction is an innovative solution to understand the genuine house pricing. In this context, machine learning driven technologies are used to generate selling price of the houses and the father of machine learning is John Mc Carthy [1]. Machine learning can be defined as training a machine using different algorithms so it can behave intelligent in future by learning from previous examples rather than just storing and retrieving data items like a database system.

Machine learning is a blessing for today's technology, and it is growing at an expeditious rate. Machine learning is used in various major concepts such as Image and Speech Recognition, prediction, Product recommendations, Self-driving cars, Email Spam and Malware Filtering, Sentiment Analysis, Banking Domain etc.

Machine learning is requisite because of its extensive range of applications and its magnificent ability to learn from previous examples to provide solutions to complex problems efficiently, effectively and quickly [2]. The aim of this project is to calculate efficient house so that in future it will help buyers and sellers in knowing the trends of the market and getting the best available price of the property. It will also help the real estate investors to know the trend and get maximum possible profit.

2. Algorithm Used: Linear Regression

Linear regression is an approach to model the relationship between a dependent variable and independent variable. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data [3].

A linear regression line has an equation of the form $Y = a + bX$, where X is the independent variable and Y is the dependent variable.

2.1 Hypothesis of Linear Regression

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$$

Y is the predicted value
θ_0 is the bias term.
$\theta_1,\ldots,\theta_n$ are the model parameters
$x_1, x_2,\ldots,x_n$ are the feature values.

2.2 The cost Function

To define and measure the error of our model we define the cost function as the sum of the squares of the residuals

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h(x^i) - y^i)^2$$

2.3 Gradient descent

Gradient descent is a generic optimization algorithm used in many machine learning algorithms. It iteratively tweaks the parameters of the model in order to minimize the cost function.

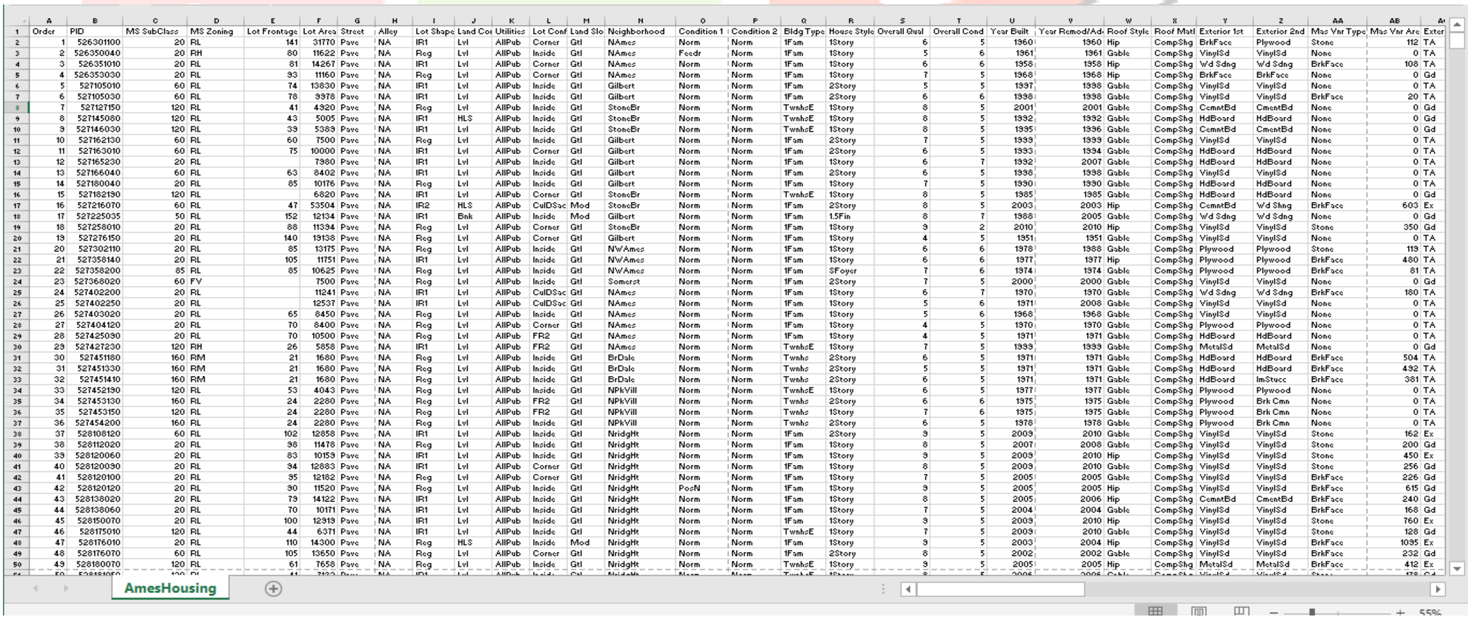
$$\theta_0 = \theta_0 - \frac{\alpha}{m}\sum_{i=1}^{m}(h(x^i) - y^i)$$

$$\theta_1 = \theta_1 - \frac{\alpha}{m}\sum_{i=1}^{m}(h(x^i) - y^i)x_1^i$$

3. Methodology

3.1 Selection and cleaning

At first data available in .csv format is taken from Kaggle in order to perform further steps. Data cleaning is the next major step performed after data selection. Different functions are created to fix the missing and null data values in the dataset and it is also confirmed that the data type is correct.



Order	PID	MS SubClass	MS Zoning	Lot Frontage	Lot Area	Street	Alley	Lot Shape	Land Coı	Utilities	Lot Conf	Land Slo	Neighborhood	Condition 1
1	526301100	20	RL	141	31770	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NAmes	Norm
2	526350040	20	RH	80	11622	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Feedr
3	526351010	20	RL	81	14267	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NAmes	Norm
4	526353030	20	RL	93	11160	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	NAmes	Norm
5	527105010	60	RL	74	13830	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm
6	527105030	60	RL	78	9978	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm
7	527127150	120	RL	41	4920	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	StoneBr	Norm
8	527145080	120	RL	43	5005	Pave	NA	IR1	HLS	AllPub	Inside	Gtl	StoneBr	Norm
9	527146030	120	RL	39	5389	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	StoneBr	Norm
10	527162130	60	RL	60	7500	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Gilbert	Norm
11	527163010	60	RL	75	10000	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	Gilbert	Norm
12	527165230	20	RL		7980	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm
13	527166040	60	RL	63	8402	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm
14	527180040	20	RL	85	10176	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Gilbert	Norm
15	527182190	120	RL		6820	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	StoneBr	Norm
16	527216070	60	RL	47	53504	Pave	NA	IR2	HLS	AllPub	CulDSac	Mod	StoneBr	Norm
17	527225035	50	RL	152	12134	Pave	NA	IR1	Bnk	AllPub	Inside	Mod	Gilbert	Norm
18	527258010	20	RL	88	11394	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	StoneBr	Norm
19	527276150	20	RL	140	19138	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	Gilbert	Norm
20	527302110	20	RL	85	13175	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NWAmes	Norm
21	527358140	20	RL	105	11751	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	NWAmes	Norm
22	527358200	85	RL	85	10625	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NWAmes	Norm
23	527368020	60	FV		7500	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm
24	527402200	20	RL		11241	Pave	NA	IR1	Lvl	AllPub	CulDSac	Gtl	NAmes	Norm
25	527402250	20	RL		12537	Pave	NA	IR1	Lvl	AllPub	CulDSac	Gtl	NAmes	Norm
26	527403020	20	RL	65	8450	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm
27	527404120	20	RL	70	8400	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	NAmes	Norm
28	527425090	20	RL	70	10500	Pave	NA	Reg	Lvl	AllPub	FR2	Gtl	NAmes	Norm
29	527427230	120	RH	26	5858	Pave	NA	IR1	Lvl	AllPub	FR2	Gtl	NAmes	Norm
30	527451180	160	RM	21	1680	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	BrDale	Norm
31	527451330	160	RM	21	1680	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	BrDale	Norm
32	527451410	160	RM	21	1680	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	BrDale	Norm
33	527452190	120	RL	53	4043	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NPkVill	Norm
34	527453130	160	RL	24	2280	Pave	NA	Reg	Lvl	AllPub	FR2	Gtl	NPkVill	Norm
35	527453150	120	RL	24	2280	Pave	NA	Reg	Lvl	AllPub	FR2	Gtl	NPkVill	Norm
36	527454200	160	RL	24	2280	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NPkVill	Norm
37	528108120	60	RL	102	12858	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	NridgHt	Norm
38	528112020	20	RL	98	11478	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm
39	528120060	20	RL	83	10159	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	NridgHt	Norm
40	528120090	20	RL	94	12883	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NridgHt	Norm
41	528120100	20	RL	95	12182	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	NridgHt	Norm
42	528120120	20	RL	90	11520	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	PosN
43	528138020	20	RL	79	14122	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	NridgHt	Norm
44	528138060	20	RL	70	10171	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	NridgHt	Norm
45	528150070	20	RL	100	12919	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	NridgHt	Norm
46	528175010	120	RL	44	6371	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	NridgHt	Norm
47	528176010	20	RL	110	14300	Pave	NA	Reg	HLS	AllPub	Inside	Mod	NridgHt	Norm
48	528176070	60	RL	105	13650	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	NridgHt	Norm
49	528180070	120	RL	61	7658	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm

Order	Condition 2	Bldg Type	House Style	Overall Qual	Overall Cond	Year Built	Year Remod/Ad	Roof Style	Roof Matl	Exterior 1st	Exterior 2nd	Mas Vnr Type	Mas Vnr Are	Exter
1	Norm	1Fam	1Story	6	5	1960	1960	Hip	CompShg	BrkFace	Plywood	Stone	112	TA
2	Norm	1Fam	1Story	5	6	1961	1961	Gable	CompShg	VinylSd	VinylSd	None	0	TA
3	Norm	1Fam	1Story	6	6	1958	1958	Hip	CompShg	Wd Sdng	Wd Sdng	BrkFace	108	TA
4	Norm	1Fam	1Story	7	5	1968	1968	Hip	CompShg	BrkFace	BrkFace	None	0	Gd
5	Norm	1Fam	2Story	5	5	1997	1998	Gable	CompShg	VinylSd	VinylSd	None	0	TA
6	Norm	1Fam	2Story	6	6	1998	1998	Gable	CompShg	VinylSd	VinylSd	BrkFace	20	TA
7	Norm	TwnhsE	1Story	8	5	2001	2001	Gable	CompShg	CemntBd	CmentBd	None	0	Gd
8	Norm	TwnhsE	1Story	8	5	1992	1992	Gable	CompShg	HdBoard	HdBoard	None	0	Gd
9	Norm	TwnhsE	1Story	8	5	1995	1996	Gable	CompShg	CemntBd	CmentBd	None	0	Gd
10	Norm	1Fam	2Story	7	5	1999	1999	Gable	CompShg	VinylSd	VinylSd	None	0	TA
11	Norm	1Fam	2Story	6	5	1993	1994	Gable	CompShg	HdBoard	HdBoard	None	0	TA
12	Norm	1Fam	1Story	6	7	1992	2007	Gable	CompShg	HdBoard	HdBoard	None	0	TA
13	Norm	1Fam	2Story	6	5	1998	1998	Gable	CompShg	VinylSd	VinylSd	None	0	TA
14	Norm	1Fam	1Story	7	5	1990	1990	Gable	CompShg	HdBoard	HdBoard	None	0	TA
15	Norm	TwnhsE	1Story	8	5	1985	1985	Gable	CompShg	HdBoard	HdBoard	None	0	Gd
16	Norm	1Fam	2Story	8	5	2003	2003	Hip	CompShg	CemntBd	Wd Shng	BrkFace	603	Ex
17	Norm	1Fam	1.5Fin	8	7	1988	2005	Gable	CompShg	Wd Sdng	Wd Sdng	None	0	Gd
18	Norm	1Fam	1Story	9	2	2010	2010	Hip	CompShg	VinylSd	VinylSd	Stone	350	Gd
19	Norm	1Fam	1Story	4	5	1951	1951	Gable	CompShg	VinylSd	VinylSd	None	0	TA
20	Norm	1Fam	1Story	6	6	1978	1988	Gable	CompShg	Plywood	Plywood	Stone	119	TA
21	Norm	1Fam	1Story	6	6	1977	1977	Hip	CompShg	Plywood	Plywood	BrkFace	480	TA
22	Norm	1Fam	SFoyer	7	6	1974	1974	Gable	CompShg	Plywood	Plywood	BrkFace	81	TA
23	Norm	1Fam	2Story	7	5	2000	2000	Gable	CompShg	VinylSd	VinylSd	None	0	Gd
24	Norm	1Fam	1Story	6	7	1970	1970	Gable	CompShg	Wd Sdng	Wd Sdng	BrkFace	180	TA
25	Norm	1Fam	1Story	5	6	1971	2008	Gable	CompShg	VinylSd	VinylSd	None	0	TA
26	Norm	1Fam	1Story	5	6	1968	1968	Gable	CompShg	VinylSd	VinylSd	None	0	TA
27	Norm	1Fam	1Story	4	5	1970	1970	Gable	CompShg	Plywood	Plywood	None	0	TA
28	Norm	1Fam	1Story	4	5	1971	1971	Gable	CompShg	HdBoard	HdBoard	None	0	TA
29	Norm	TwnhsE	1Story	7	5	1999	1999	Gable	CompShg	MetalSd	MetalSd	None	0	Gd
30	Norm	Twnhs	2Story	6	5	1971	1971	Gable	CompShg	HdBoard	HdBoard	BrkFace	504	TA
31	Norm	Twnhs	2Story	5	5	1971	1971	Gable	CompShg	HdBoard	HdBoard	BrkFace	492	TA
32	Norm	Twnhs	2Story	6	5	1971	1971	Gable	CompShg	HdBoard	ImStucc	BrkFace	381	TA
33	Norm	TwnhsE	1Story	6	5	1977	1977	Gable	CompShg	Plywood	Plywood	None	0	TA
34	Norm	Twnhs	2Story	6	6	1975	1975	Gable	CompShg	Plywood	Brk Cmn	None	0	TA
35	Norm	Twnhs	1Story	7	6	1975	1975	Gable	CompShg	Plywood	Brk Cmn	None	0	TA
36	Norm	Twnhs	2Story	6	5	1978	1978	Gable	CompShg	Plywood	Brk Cmn	None	0	TA
37	Norm	1Fam	2Story	9	5	2009	2010	Gable	CompShg	VinylSd	VinylSd	Stone	162	Ex
38	Norm	1Fam	1Story	8	5	2007	2008	Gable	CompShg	VinylSd	VinylSd	Stone	200	Gd
39	Norm	1Fam	1Story	9	5	2009	2010	Hip	CompShg	VinylSd	VinylSd	Stone	450	Ex
40	Norm	1Fam	1Story	8	5	2009	2010	Gable	CompShg	VinylSd	VinylSd	Stone	256	Gd
41	Norm	1Fam	1Story	7	5	2005	2005	Gable	CompShg	VinylSd	VinylSd	BrkFace	226	Gd
42	Norm	1Fam	1Story	9	5	2005	2005	Hip	CompShg	VinylSd	VinylSd	BrkFace	615	Gd
43	Norm	1Fam	1Story	8	5	2005	2006	Hip	CompShg	CemntBd	CmentBd	BrkFace	240	Gd
44	Norm	1Fam	1Story	7	5	2004	2004	Gable	CompShg	VinylSd	VinylSd	BrkFace	168	Gd
45	Norm	1Fam	1Story	9	5	2009	2010	Hip	CompShg	VinylSd	VinylSd	Stone	760	Ex
46	Norm	TwnhsE	1Story	7	5	2009	2010	Gable	CompShg	VinylSd	VinylSd	Stone	128	Gd
47	Norm	1Fam	1Story	9	5	2003	2004	Hip	CompShg	VinylSd	VinylSd	BrkFace	1095	Ex
48	Norm	1Fam	2Story	8	5	2002	2002	Gable	CompShg	VinylSd	VinylSd	BrkFace	232	Gd
49	Norm	TwnhsE	1Story	9	5	2005	2005	Hip	CompShg	MetalSd	MetalSd	BrkFace	412	Ex

Fig.1. DATA SET

3.1.1 Exploratory Data Analysis:

Exploratory Data Analysis was conducted to explore the relationship between Sale Price with every other features. For numeric features, the linear relationship was examined using a heatmap and correlation coefficients. For categorical data, bar plot was created to visualize the mean Sale Price across categories.

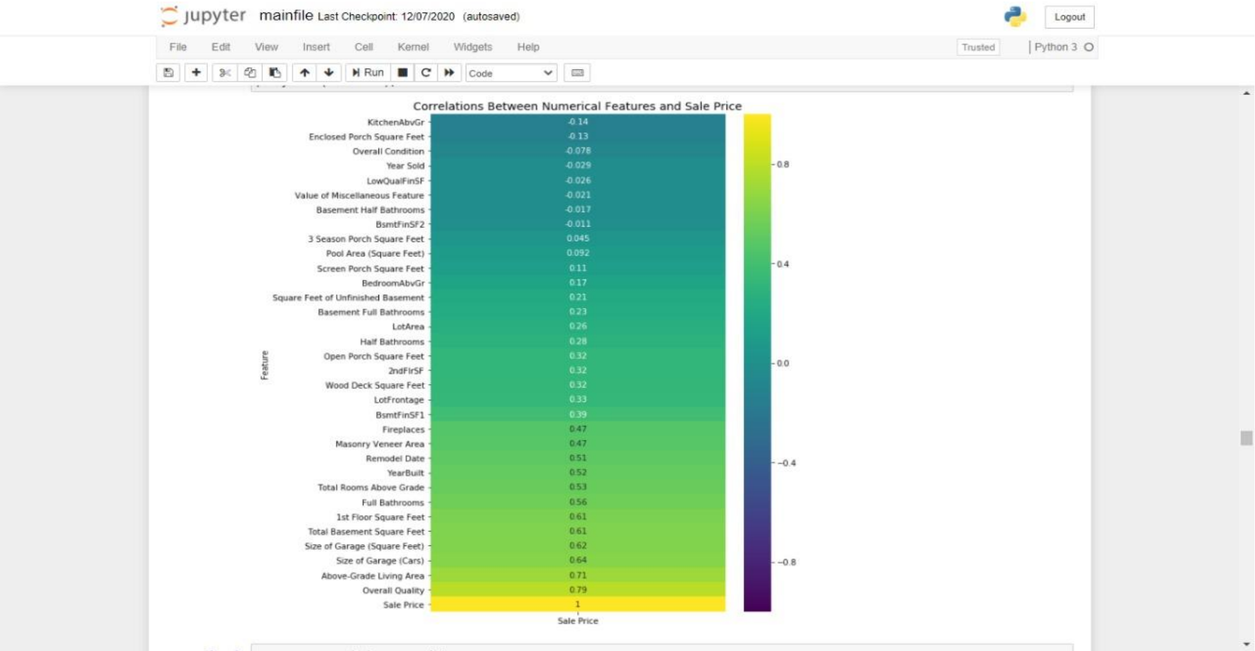


FIG.2 HEATMAP : Represents The Relationship Between Sale Prices And Numeric Features.

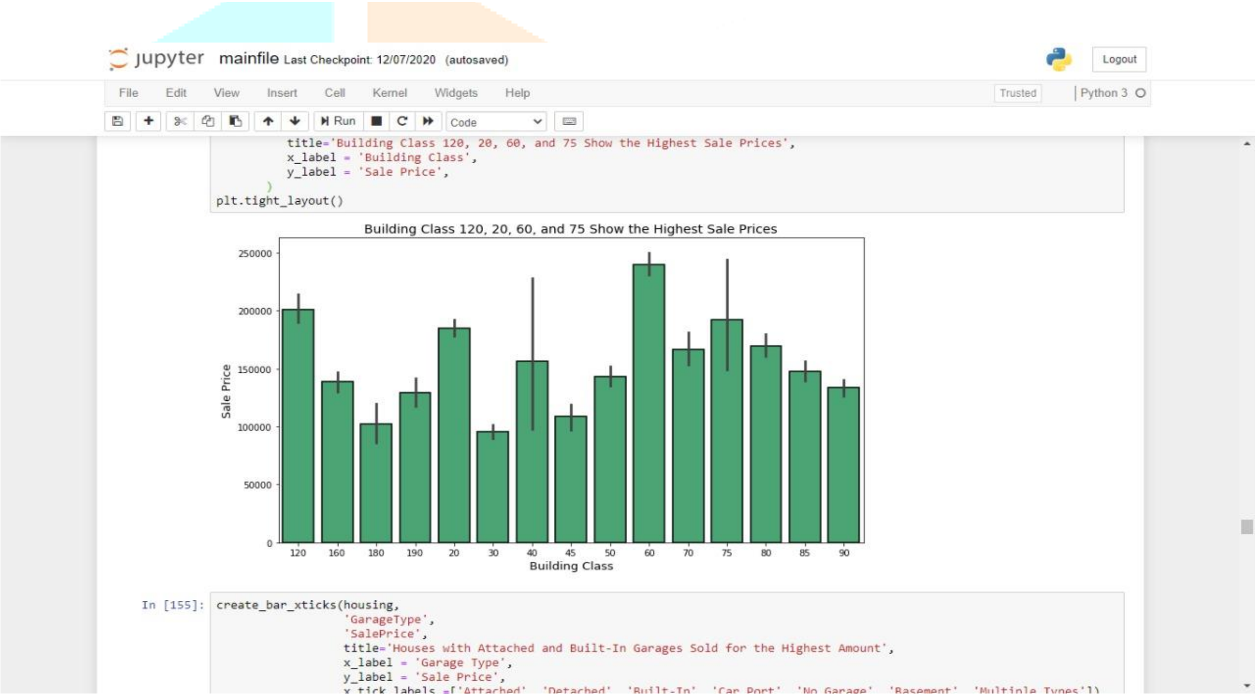


FIG.3 FIG.3 BAR GRAPH: Represents the relationship between sale prices and categorical data.

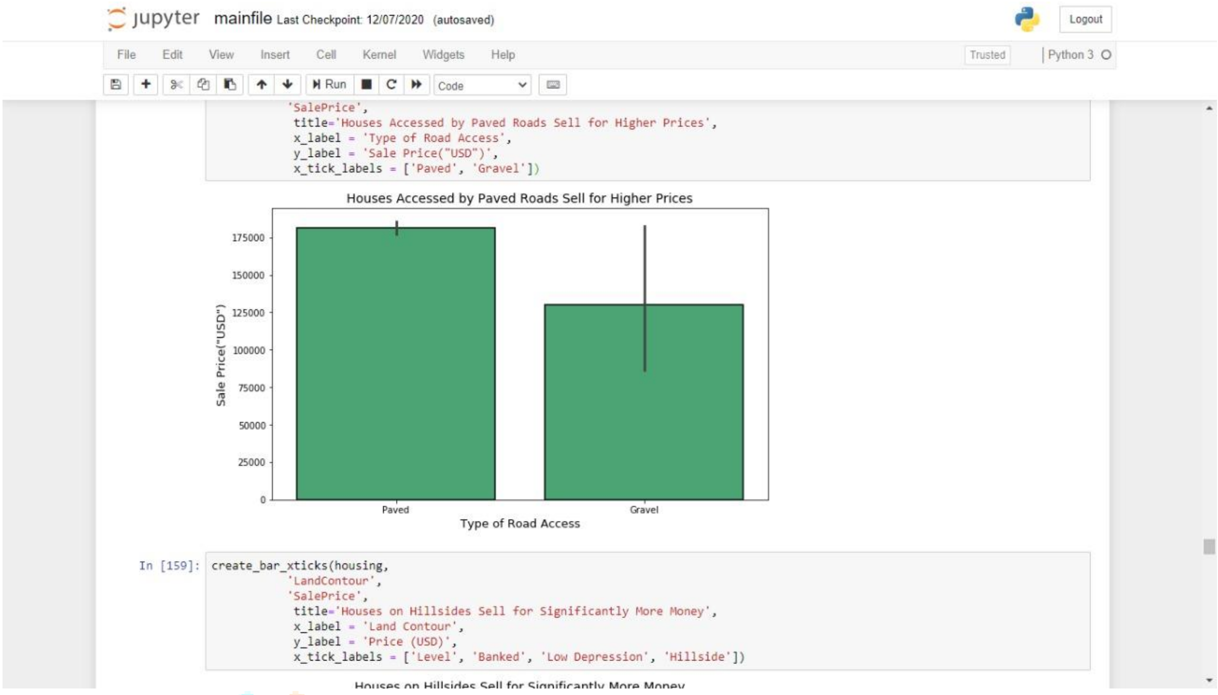


FIG.4 Houses Accessed by Paved Roads Sell for Higher Prices.

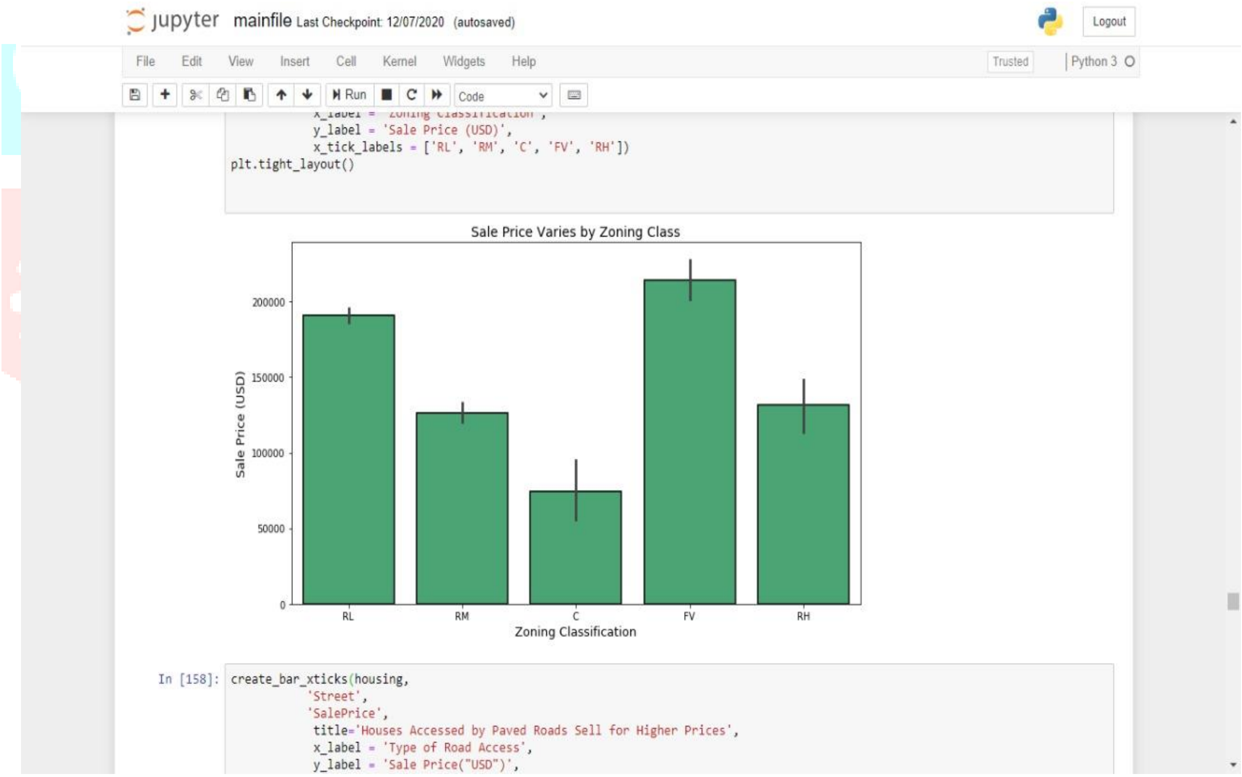


Fig.5 Sales Prices Varies by Zoning Class.

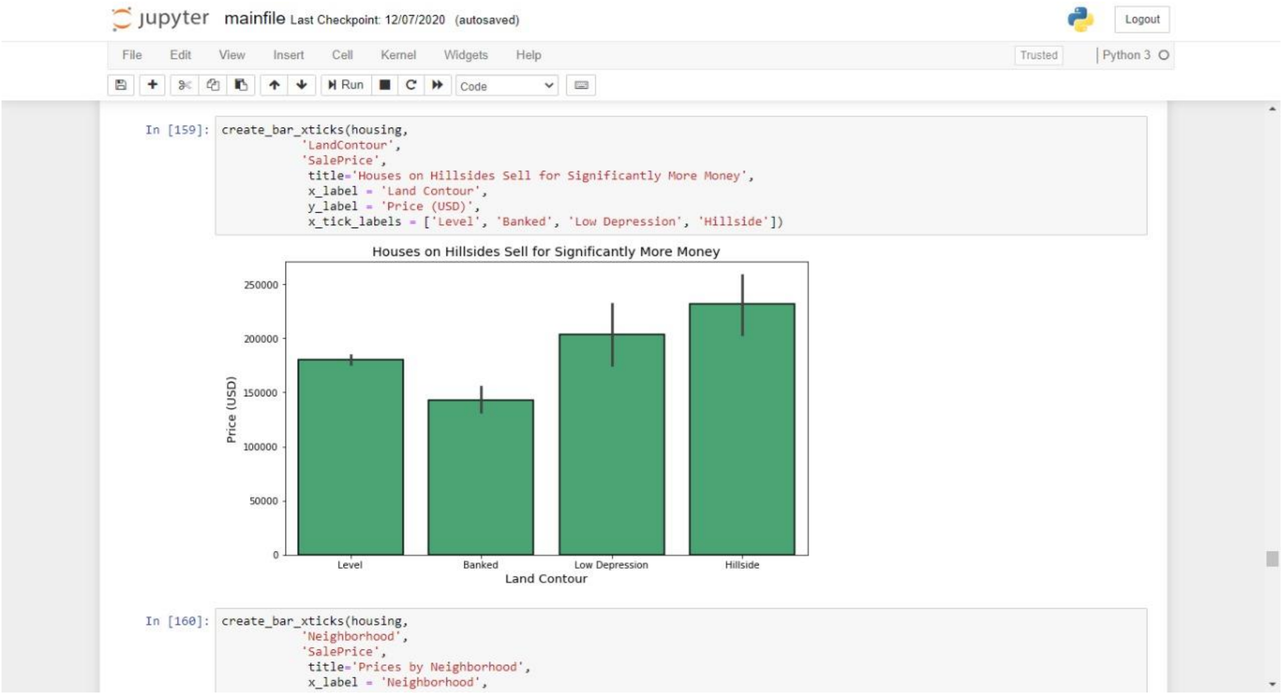


Fig.6 Houses on Hillside Sell for Significantly More Money.

3.2 Feature Engineering

In this phase, features are engineered to reduce dimensionality of the data and to account for the patterns and clusters that emerge during Exploratory Data Analysis. During feature engineering, a categorical variable was coded and removed from the data frame. All categorical variables of interest were dummified.

4. RESULT

```
lr_prelim = LinearRegression()
```

```
lr_prelim.fit(X_train_1, y_train_1);
```

```
display_R2_scores(lr_prelim, X_train_1, y_train_1, X_test_1, y_test_1)
```

```
The mean cross validation score for this model is 0.8195.
The training score for this model is 0.94.
The testing score for this model is 0.713.
```

FIG.7 Output Image of Code and Accuracy Of Linear Regression

5. CONCLUSION

We have learnt about the concepts of linear regression and gradient descent. The system makes optimal use of the Linear Regression Algorithm. The system makes use of data in the most efficient way. The linear regression algorithm helps to fulfil customers by increasing the accuracy of estate choice and reducing the risk of investing in an estate. A lot's of features that could be added to make the system more widely acceptable. The model's accuracy in predicting house price was measured by cross validation.

6. REFERENCES

[1] McCarthy, J.; Minsky, M.L.; Rochester, N.; Shannon, C.E. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. AI Mag. 2006, 27, 12.

[2] Nihar Bhagat, Ankit Mohokar, Shreyash Mane, 2016, House Price Forecasting using Data Mining.

[3] G. Naga Satish, Ch. V. Raghavendran, M.D. Sugnana Rao, Ch.Srinivasulu, 2019, House Price Prediction Using Machine Learning.

[4] Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, Wayan Firdaus Mahmudy, Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization. 2017.

[5] Ayon Dey, Machine Learning Algorithms: A Review 2016