**COMPARITIVE STUDY ON REGRESSION TECHNIQUES**

**Dissertation Submitted in partial fulfillment of the requirements for the**

**award of the degree of Master of Science [Biostatistics]**

**By**

**PRIYADHARSHINI M**

**Register No: 1813422046009**

**Under the Guidance of**

**Dr. JANANI B, M.Sc., M.Phil.,**

**Assistant Professor**

****

**RESEARCH DEPARTMENT OF STATISTICS**

**S.D.N.B. VAISHNAV COLLEGE FOR WOMEN (AUTONOMOUS)**

**CHROMEPET, CHENNAI-600044**

**(Affiliated to University of Madras, Re-Accredited with ‘A+’ Grade by NAAC)**

**APRIL 2020**

**SHRIMATHI DEVKUNVAR NANALAL BHATT VAISHNAV**

**COLLEGE FOR WOMEN (AUTONOMOUS) CHROMEPET,**

**CHENNAI-44.**



**M.Sc. (BIOSTATISTICS)**

**2018-2020**

This is to certify that the dissertation entitled “**COMPARITIVE STUDY ON REGRESSION TECHNIQUES**” is a record of research work done by PRIYADHARSHINI M in the Department of Statistics during the period 2019-2020.This dissertation is an independent work on the part of the candidate under the supervision Dr. JANANI B, M.Sc., M.Phil., Assistant Professor, Department of Statistics, S.D.N.B. Vaishnav College for Women (Autonomous) Chromepet, Chennai-44.

Dissertation Supervisor Department of Statistics

(INCHARGE)

**SHRIMATHI DEVKUNVAR NANALAL BHATT VAISHNAV**

**COLLEGE FOR WOMEN (AUTONOMOUS) CHROMEPET,**

**CHENNAI-44.**



**M.Sc. (BIOSTATISTICS)**

**2018-2020**

NAME : PRIYADHARSHINI M

REG.NO : 1813422046009

This is to certify that this project work entitled “**COMPARITIVE STUDY ON REGRESSION TECHNIQUES**” is a bonafide record of the work done by PRIYADHARSHINI M in partial fulfillment of the requirements for the award of the degree of Master of Science in Biostatistics in SHRIMATHI DEVKUNVAR NANALAL BHATT VAISHNAV COLLEGE FOR WOMEN (AUTONOMOUS), CHROMEPET,CHENNAI-44, affiliated to the University of Madras.

Dissertation Supervisor Department of Statistics

(INCHARGE)

Submitted to the University of Madras for the M.Sc., Project Viva-Voce held on 28-03-2020 at SHRIMATHI DEVKUNVAR NANALAL BHATT VAISHNAV COLLEGE FOR WOMEN (AUTONOMOUS) CHROMEPET, CHENNAI-44.

INTERNAL EXAMINER EXTERNAL EXAMINER

**ACKNOWLEDGEMENT**

Before anything praise be to the god for all things.

I would also like to dedicate my thanks to Dr.R.Geetha, M.Phil., Ph.D., Principal for providing all facilities and resource for the completion of project.

I would like to express my sincere gratitude to Dr.D.Annapoorni, M.Phil., Ph.D., Associate Professor and Head of the Department of Statistics (S.D.N.B Vaishnav College for Women) for guidance.

I place on record, my sincere gratitude to Mrs Suriya, M.Phil., Assistant Professor and Head of the Department of BioStatistics (S.D.N.B vaishnav College for Women) for her great efforts, constant guidance and encouragement.

I would like to express my special thanks and gratitude to my supervisor Dr.JANANI B, Assistant Professor for her great efforts and guidance throughout the preparation of this research.

The support of my family and friends throughout this work is very much appreciated.

I would also like to dedicate my solitary thanks to all the staffs in the Department of Biostatistics (SDNB Vaishnav College for Women) for their assistance to me throughout my years of study.

**ABSTRACT**

Usually, House price index represents the summarized price changes of residential housing. While for a single family house price prediction, it needs more accurate method based on house type, size, build year, local amenities, and some other factors which could affect house demand and supply. With limited dataset and data features, a practical and composite data pre-processing, creative feature is examined in this paper. The paper proposes a comparative study on Linear and advanced Regression techniques such as LASSO and Ridge Regression model to predict individual house price. In this paper we discuss about the best regression that is used for the prediction of future housing prices. For the selection of prediction methods we compare and explore various prediction methods. We utilize Ridge and LASSO regression as our model because of its adaptable and probabilistic methodology on model selection. The research aim is to check how well the Ridge regression and Lasso regression works for the prediction of house pricing. The result from this research proved combination of advanced regression technique is suitable and the minimum prediction error is obtained for prediction of house pricing.

**Index Terms**: - House Prediction; Regression Analysis, LASSO Regression, Ridge Regression.

**CHAPTERS**

**INDEX**

**1. INTRODUCTION ------------------------------------------------------------------- 8**

**2. REVIEW OF LITERATURE ----------------------------------------------------- 11**

**3. DATA DESCRIPTION AND METHODOLOGY ---------------------------- 14**

**4. ANALYSIS AND INTERPRETATION ---------------------------------------- 21**

**5. FINDINGS AND CONCLUSION ----------------------------------------------- 34**

**BIBLIOGRAPHY --------------------------------------------------------------------- 35**

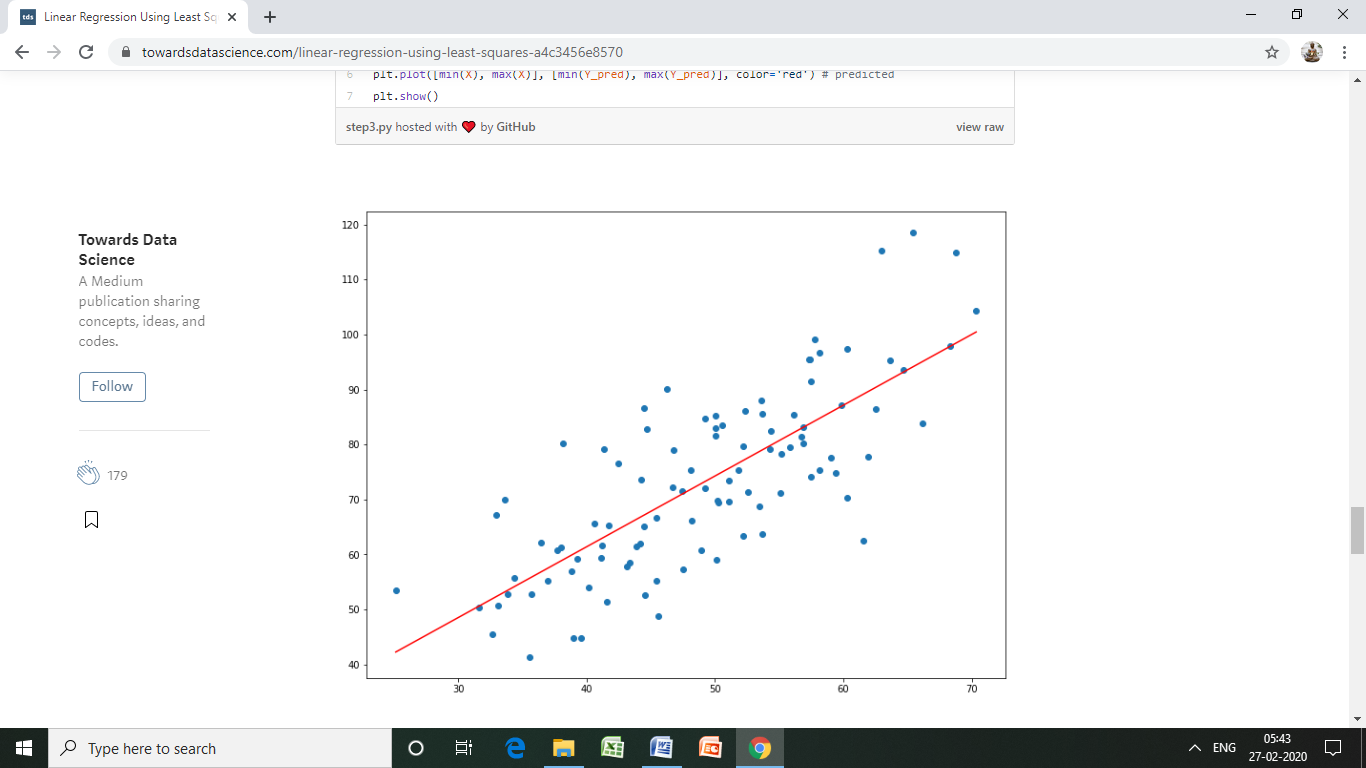
**APPENDIX ------------------------------------------------------------------------------ 36**

**CHAPTER-1**

**INTRODUCTION**

**Linear regression**

Regression analysis is the statistical method for predictive modeling, and it is one of the most commonly used methods in many scientific fields such as engineering, the physical and chemical sciences, and the social sciences, sociology, geology, etc. Satisfying the assumptions such as collinearity between variables ought to be significant issue in data science.



Let’s see an easy example for simple linear regression. As per our example our friend has a dog and she is interested to know about her pet a bit deeper. She specifically wants to know how many times her pet wags its tail when she pats him on his head for a certain number of times. To analyse this she collected some data on her own by observing. For a single pat he wags twice, for three pats he wagged five times, for four pats he wagged seven times and so on. By using these values she plotted a scatter plot by making number of pats on x axis and number of wags on y axis. Now the simple linear regression equation can be applied for the observation collected by the following eqation

Wags = (Weight \* Pats) + Bias

By substituting the observed values on the equation we can obtain the predicted wags for certain number of pats.

Advanced level tools such as Lasso and Ridge regression methods are designed to overcome such problem.

**Ridge regression**

Ridge regression is a way to create a parsimonious model when the number of predictor variables in a set exceeds the number of observations, or when a data set has multicollinearity (correlations between predictor variables).

Tikhivov’s method is basically the same as ridge regression, except that Tikhonov’s has a larger set. It can produce solutions even when your data set contains a lot of statistical noise (unexplained variation in a sample).

**Ridge Regression vs. Least Squares**

Least squares regression isn’t defined at all when the number of predictors exceeds the number of observations; It doesn’t differentiate “important” from “less-important” predictors in a model, so it includes all of them. This leads to overfitting a model and failure to find unique solutions. Least squares also has issues dealing with multicollinearity in data. Ridge regression avoids all of these problems. It works in part because it doesn’t require unbiased estimators; While least squares produces unbiased estimates, variances can be so large that they may be wholly inaccurate. Ridge regression adds just enough bias to make the estimates reasonably reliable approximations to true population values.

**Shrinkage**

Ridge regression uses a type of shrinkage estimator called a ridge estimator. Shrinkage estimators theoretically produce new estimators that are shrunk closer to the “true” population parameters. The ridge estimator is especially good at improving the least-squares estimate when multicollinearity is present.

**LASSO: (Least Absolute Shrinkage and Selection Operator)**

It is an advanced regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistic model it produces.

Originally, it was discovered in geophysics literature in 1986 and later independently rediscovered and popularized by Rober Tibshirani.

Lasso was originally formulated for least square models and this simple case reveals a substantial amount about the behavior of the estimator.

**Lasso Vs Ridge**

Ridge and Lasso work by penalizing the magnitude of coefficients of features along with minimizing the error between predict and actual observations.The difference is how they assign penalty to the coefficients, Ridge regression assign penalty for square of the magnitude of the coefficients

Optimization problem is that, we wish to find either a maximum or a minimum of a specific function. The function that you want to optimize is usually called the loss function (or cost function). The loss function is defined for each machine learning algorithm you use, and this is the main metric for evaluating the accuracy of your trained model.

**Statistical models used in prediction:**

The common statistical models that can be used for the prediction are

* Support Vector Machine
* Random Forest
* Regression
* Decision Tree
* Neural Network

**Motivation of the study:**

I compare linear regression with advanced regression techniques like Ridge and LASSO regression so that it can be used for better prediction for any kind of dataset.

**Objective:**

To find out better regression technique that gives almost accurate prediction.

**Study Area:**

Comparing Regression techniques for the purpose of predicting the Zillow’s Home Value Prediction data from Kaggle was used.

**Analysis:**

The secondary data from Kaggle is used to find what better Regression analysis can be used to perform prediction.

**Chapterization:**

**Introduction** gives a brief content on what my study is about and the importance of it.

**Literature’s Review** consists a summary of the knowledge about a specific area of my research.

**Data Description and Methodology** chapter explains the procedure I did, to evaluate the reliability and validity of the research. That is, this chapter contains the process of data collection, data preprocessing, data analysis, system analysis, system implementation, system testing.

**Analysis and Interpretation** chapter gives the interpretation of the model.

**Finding and Conclusion** chapter gives the judgments based on the data.

**CHAPTER-2**

**REVIEWS OF LITERATURE**

This chapter provides a review of the recent research articles based on the difference and the prediction of ridge and lasso regression techniques for various fields.

**Literature review:**

**Mayooran Thevaraja(University of Jaffna), Azizur Rahman(Charles Sturt University), Mathew Gabirial(Minnesots State University) has done a work on “COMPARING LINEAR, RIDGE AND LASSO REGRESSION TECHNIQUES USING WINE DATA”.**

This research paper mainly deals with comparing the linear regression with Ridge and Lasso regressions on the Vinho Verde white wine test data from Minho (northwest) region of Portugal. Using the Vinho Verde white wine test data, a statistical analysis is made for testing the advantages of each of the three regression methods using R software with ‘glmnet’ package. The Wine Quality dataset describes red and white variants of the “Vinho Verde” wine. In this paper only white wine data has been chosen for the analysis. The data contains 11 physiological predictors which encompass the physical and chemical characteristic of “Vinho Verde”. The RSS value calculated on train set for OLS is less than that for Ridge and Lasso. Calculating the RSS values on the test data set provides a good way to assess the regression model in contrast to using a single dataset. It has been given that Ridge and Lasso gives less Rss values, which means that Ridge and Lasso are the best fitting models.

**Jose Manuel Pereiraª\*, Mario Bastoa, Amelia Ferreira da Silvab - IPCA - Polytechnic Institute of Cavado and Ave, Campus do IPCA, 4750-810 Barcelos, Portugal, IPP - Polytechnic Institute of Oporto, Rua Dr. Roberto Frias, 4200-465 Oporto, Portugal.**

“The Logistic Lasso and Ridge Regression in Predicting Corporate Failure”. The prediction of corporate bankruptcy is a phenomenon of interest to investors, creditors, borrowing firms, and governments alike. Many quantitative methods and distinct variable selection techniques have been employed to develop empirical models for predicting corporate bankruptcy. For the present study the lasso and ridge approaches were undertaken, since they deal well with multicolinearity and display the ideal properties to minimize the numerical instability that may occur due to overfitting. The models were employed to a dataset of 2032 non-bankrupt firms and 401 bankrupt firms belonging to the hospitality industry, over the period 2010-2012. The results showed that the lasso and ridge models tend to favor the category of the dependent variable that appears with heavier weight in the training set, when compared to the stepwise methods implemented in SPSS.

**A SURVEY ON REGRESSION ESTIMATE WITH LASSO METHOD**

**Ms. Anjali, Mr. Rakesh Shivhare, Ms. Komal Pandey3 Mr. Mukesh Dixit, Research Scholar (CSE Department), REC Bhopal.**

Several quantitative methods along with distinct variable choice techniques have been working to develop observed models for predicting commercial bankruptcy. For the current study the lasso as well as ridge approach is undertaken, since they agreement well through multicolinearity along with show the ideal property to minimize the arithmetical instability that might occur due

to over fitting. They fit Lasso but they compare both ridge and lasso to analyse the data in a more detailed format. SPSS tool has be used in this dataset to fit the model for ridge and lasso regression.

**“Feature Selection using LASSO” by Valeria Fonti.**

This research paper explain and discuss the use of the LASSO method to address the feature selection task. Feature selection is a crucial and challenging task in the statistical modeling field, there are many studies that try to optimize and standardize this process for any kind of data, but this is not an easy thing to do. The paper includes the an introduction of feature selection task, the LASSO method, and apply the LASSO feature selection property to a Linear Regression problem, and the results of the analysis on a real dataset are shown. And fnally, the same analysis is repeated on a Generalized Linear Model in particular with a Logistic Regression Model for

a high-dimensional dataset and the findings of the scientific study of J.Chen and Z.Chen [4] are presented in the research paper.

**“Ridge and Lasso Regression Models for Cross-Version Defect Prediction” - Xiaoxing Yang ,School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China.**

**Wushao Wen ,School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China.**

Sorting software modules in order of defect count can help testers to focus on software modules with more defects. One of the most popular methods for sorting modules is generalized linear regression. Both Ridge and LASSO regression models are used for cross-version defect prediction. Cross-version defect prediction is an approximated to real applications. It constructs prediction models from a previous version of projects and predicts defects in the next version. Experimental results based on 11 projects from the PROMISE repository consisting of 41 different versions show that: 1) there exist severe multicollinearity problems in the experimental datasets; 2) both Ridge and LASSO regression models perform better than linear regression and negative binomial regression for cross-version defect prediction; and 3) compared with two best methods(Ridge and LASSO regression) in sorting software modules according to the predicted number of defects. It has been shown that Ridge regression has comparable performance and less model construction time compared to LASSO regression.

**CHAPTER-3**

**DATA DESCRIPTION AND METHODOLOGY**

**Data description**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **VARIABLE** | **DESCRIPTION** |
| 1 | MSSubClass | The building class |
| 2 | LotArea | Lot size in square feet |
| 3 | OverallQual | Overall material and finish quality |
| 4 | OverallCond | Overall condition rating |
| 5 | YearBuilt | Original construction date |
| 6 | 1stFlrSF | First Floor square feet |
| 7 | 2ndFlrSF | Second floor square feet |
| 8 | BsmtFullBath | Basement full bathrooms |
| 9 | BedroomAbvGr | Number of bedrooms above basement level |
| 10 | KitchenAbvGr | Number of kitchens above grade |
| 11 | TotRmsAbvGrd | Total rooms above grade (does not include bathrooms) |
| 12 | GarageCars | Size of garage in car capacity |
| 13 | WoodDeckSF | Wood deck area in square feet |
| 14 | ScreenPorch | Screen porch area in square feet |
| 15 | SalePrice | The property's sale price in dollars. |

Here, SalePrice is the target variable that we're trying to predict.

**Methodology**

Data

Descriptive Measures:

* Frequency table
* Bar graph

Regression techniques:

* Linear regression
* Ridge regression
* Lasso regression

Statistical Tools Used:

* SPSS
* Python.

**Data:**

Any bit of information that is expressed in a value or numerical number is data. Data is basically a collection of information, measurements or observations.

Raw data is an initial collection of information. This information has not yet been organized. After the very first step of data collection, you will get raw data.

Discrete data is that which is recorded in whole numbers, like the number of children in a school or number of tigers in a zoo. It cannot be in decimals or fractions.

Continuous data need not be in whole numbers, it can be in decimals. Examples are the temperature in a city for a week, your percentage of marks for the last exam etc.

**Descriptive Measures**

**Frequency table:**

Frequency distribution in statistics provides the information of the number of occurrences (frequency) of distinct values distributed within a given period of time or interval, in a list or table.

**Mean:**

The mean (or average) is the most popular and well-known measure of central tendency. It can be used with both discrete and continuous data. The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have n values in a data set and they have values x1, x2, ..., xn, the population mean, usually denoted by µ is,

**Median:**

The median is the middle number in a sorted, ascending or descending, list of numbers and can be more descriptive of that data set than the average.

Median = {(n + 1) ÷ 2}th value

The median is sometimes used as opposed to the mean when there are outliers in the sequence that might skew the average of the values.

If there is an odd amount of numbers, the median value is the number that is in the middle, with the same amount of numbers below and above.

If there is an even amount of numbers in the list, the middle pair must be determined, added together, and divided by two to find the median value.

**Mode:**

The mode is the most commonly observed value in a set of data.In many cases, the modal value will differ from the average value in the data.

**Bar** **graphs**:

A bar graph is a way to represent data graphically by using bars or columns of different lengths. They are graphs we use to compare various items or choices or to show how something changes over a period of time. They can also be utilized to show the frequency of certain data. Other names for bar chart are bar charts or column graphs.

**Regression techniques**

**Linear regression:**

The simple linear regression model, in which you aim at predicting n observations of the response variable, Y, with a linear combination of m predictor variables, X, and a normally distributed Ridge Regression or term with variance σ2:

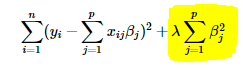
,

.

As we don't know the true parameters , we have to estimate them from the sample. In the Ordinary Least Squares (OLS) approach, we estimate them as in such a way, that the sum of squares of residuals that is deviated from the best fit line is made as small as possible. In other words, we minimize the loss function which is the mean square error. The most common method for fitting a regression line is the method of least-squares. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0).

**Ridge regression:**

Recently, high-dimensional shrinkage and parameter selection techniques have been of great importance. Tikhonov regularization is another name for Ridge Regression that deals with many predictors and an ill-conditioned model matrix (i.e. is not invertible or near singular). Fitting a model with many predictors and no regularization results in a non-unique LSE’s solution. Furthermore, LSE depends on where if the , then doesn’t have an inverse. However, Ridge Regression has the ability to overcome these hurdles by constraining the coefﬁcient estimates; hence, it can reduce the estimator’s variance and introduce some bias.

Similar to LSE, Ridge Regression coefﬁcients are estimated by minimizing the loss function, 

If lambda is zero then you can imagine we get back OLS. However, if lambda is very large then it will add too much weight and it will lead to under-fitting. Having said that it’s important how lambda is chosen. This technique works very well to avoid over-fitting issue. Tikhonov regularization tries to ﬁnd estimates that ﬁt the data reasonably well.

**LASSO regression:**

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of muticollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

Lasso Regression adds “absolute value of magnitude” of coefficient as penalty term to the loss function.



If lambda is zero then we will get back OLS whereas very large value will make coefficients zero hence it will under-fit.

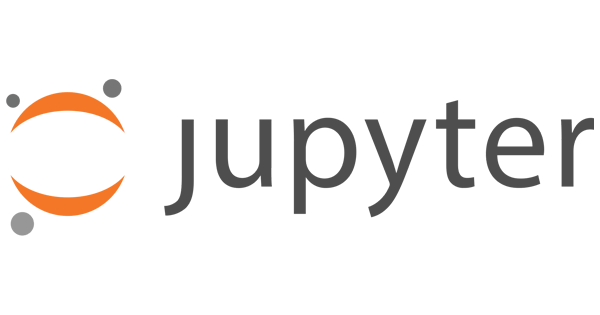
**Statistical tools**

The two major statistical tool used for computing the house price dataset are

* SPSS
* Python

**SPSS:ibm2.png (Statistical Programming for Social Sciences)**

IBM SPSS Statistics, the world’s leading statistical software, is designed to solve business and research problems by means of ad hoc analysis, hypothesis testing, geospatial analysis and predictive analytics. Organizations use SPSS Statistics to understand data, analyze trends, forecast and plan to validate assumptions, and drive accurate conclusions. In the house pricing dataset we use SPSS for checking out the frequency and a graphical representation of the dataset.

**Python: **

Project Jupyter exists to develop open-source software, open-standards, and services for interactive computing across dozens of programming languages. The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

**CHAPTER-4**

**ANALYSIS AND INTERPRETATION**

**Data description:**

**Frequency table:**

To perform prediction using advanced regression analysis, I took Zillow’s Home Value Prediction data from Kaggle which has 15 variables. The following information provides the information about the variables that are used to perform prediction with advanced regression analysis.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | MSSubClass | LotArea | OverallQual | OverallCond | YearBuilt |
| N | Valid | 1460 | 1460 | 1460 | 1460 | 1460 |
| Missing | 0 | 0 | 0 | 0 | 0 |
| Mean | | 56.90 | 10516.83 | 6.10 | 5.58 | 1971.27 |
| Median | | 50.00 | 9478.50 | 6.00 | 5.00 | 1973.00 |
| Mode | | 20 | 7200 | 5 | 5 | 2006 |
| Variance | | 1789.338 | 99625649.650 | 1.913 | 1.238 | 912.215 |
| Minimum | | 20 | 1300 | 1 | 1 | 1872 |
| Maximum | | 190 | 215245 | 10 | 9 | 2010 |
| Sum | | 83070 | 15354569 | 8905 | 8140 | 2878051 |

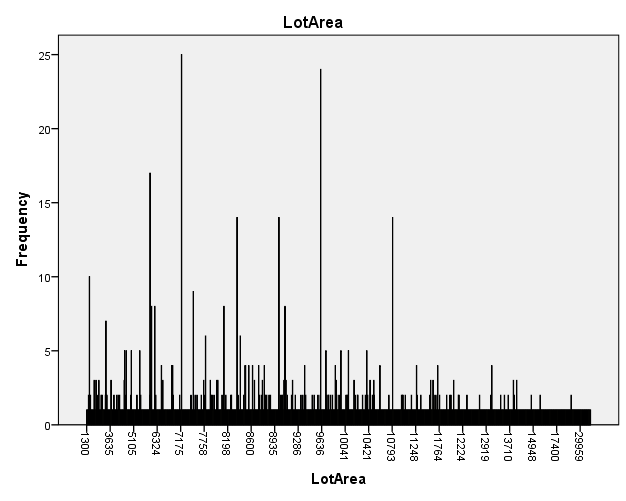
|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | @1stFlrSF | @2ndFlrSF | | BsmtFullBath | | BedroomAbvGr | | KitchenAbvGr |
| N | | Valid | | 1460 | 1460 | | 1460 | | 1460 | | 1460 |
| Missing | | 0 | 0 | | 0 | | 0 | | 0 |
| Mean | | | | 1162.63 | 346.99 | | .43 | | 2.87 | | 1.05 |
| Median | | | | 1087.00 | .00 | | .00 | | 3.00 | | 1.00 |
| Mode | | | | 864 | 0 | | 0 | | 3 | | 1 |
| Variance | | | | 149450.079 | 190557.075 | | .269 | | .665 | | .049 |
| Minimum | | | | 334 | 0 | | 0 | | 0 | | 0 |
| Maximum | | | | 4692 | 2065 | | 3 | | 8 | | 3 |
| Sum | | | | 1697435 | 506609 | | 621 | | 4185 | | 1528 |
|  | | | | | | | | | | | | |
|  | | | TotRmsAbvGrd | | | GarageCars | | WoodDeckSF | | ScreenPorch | SalePrice | |
| N | Valid | | 1460 | | | 1460 | | 1460 | | 1460 | 1460 | |
| Missing | | 0 | | | 0 | | 0 | | 0 | 0 | |
| Mean | | | 6.52 | | | 1.77 | | 94.24 | | 15.06 | 180921.20 | |
| Median | | | 6.00 | | | 2.00 | | .00 | | .00 | 163000.00 | |
| Mode | | | 6 | | | 2 | | 0 | | 0 | 140000 | |
| Variance | | | 2.642 | | | .558 | | 15709.813 | | 3108.889 | 6311111264.297 | |
| Minimum | | | 2 | | | 0 | | 0 | | 0 | 34900 | |
| Maximum | | | 14 | | | 4 | | 857 | | 480 | 755000 | |
| Sum | | | 9516 | | | 2580 | | 137597 | | 21989 | 264144946 | |

**Interpretation:**

The dataset contains total of 1460 values with no missing values. The above table gives mean, median, mode, variance, minimum, maximum of all the variables present in the dataset. This table gives detailed structure of how all the variables are present in the dataset.

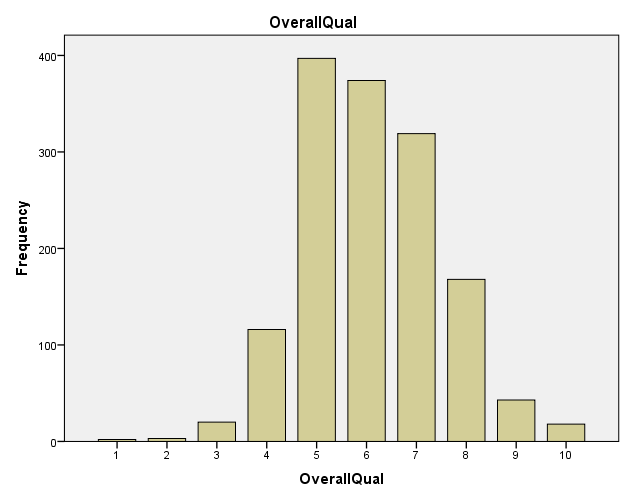
**Bar graph:**

A simple and easy way of representing the variables present in Zillow’s Home Value Prediction data is the bar graph that gives the frequency for each variables present in the dataset.

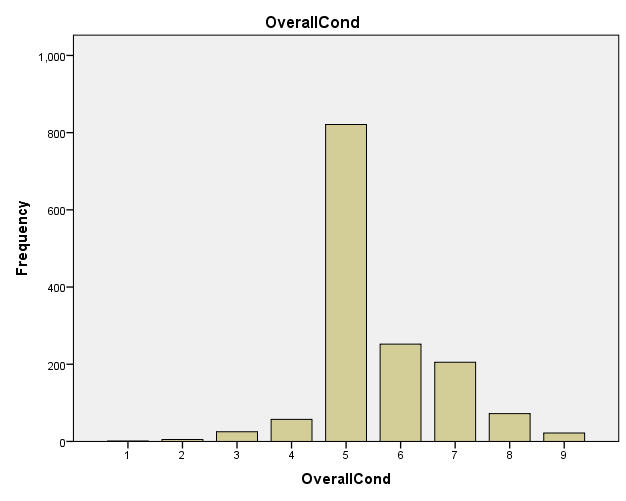


LotArea is the variable which measures the Lot size in square feet. From the graph, we could see nearly 25 houses are built within 7175 area square feet and few more with 9636 square feet and

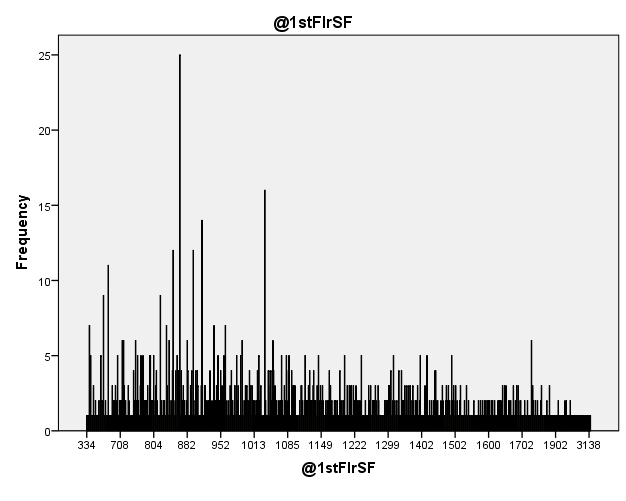
many houses are built within the size that ranges from 5105 to 10793.



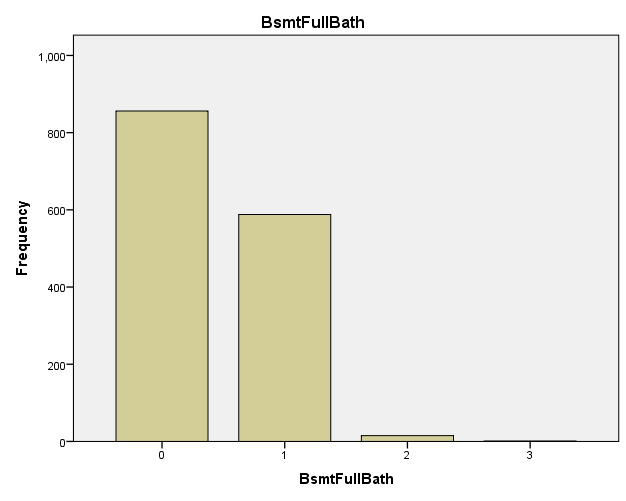
OverallQual is the variable that gives overall material and finish quality of the house. From the above graph we could see that, the nearly 300-400 houses are built with average and good quality.



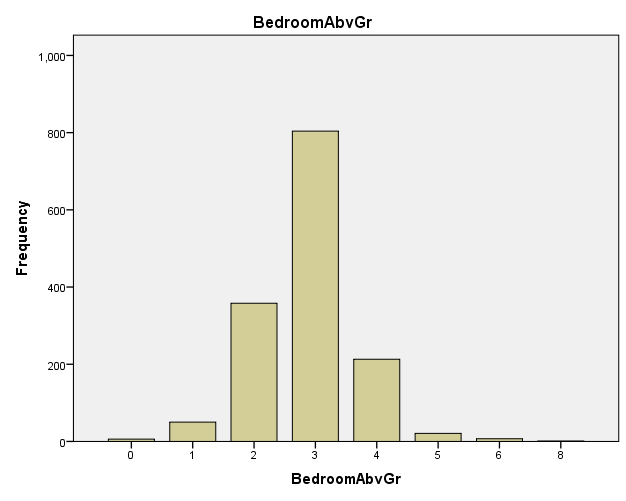
“OverallCond” is the variable which gives the rating for Overall condition for a house from 1-9. From the above bar graph we could see that most of houses are in good condition from which most of the ratings for the condition of house lies within 4-8.



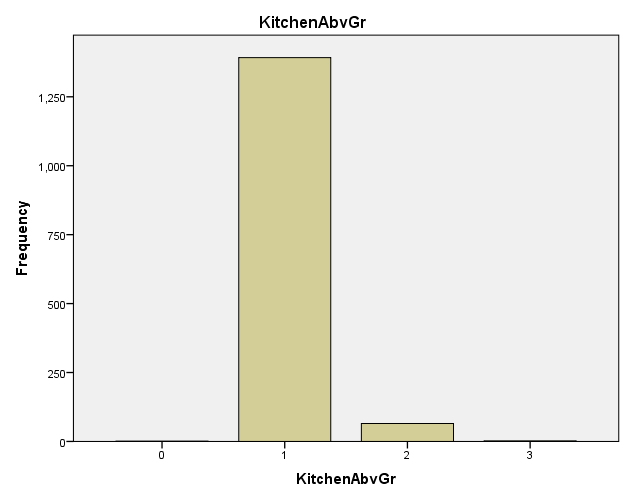
1stFlrSF represents the square feet of first floor. It has been shown that most of the houses have square feet that range from 804-882 and some more houses contain square feet range from 1013-1085.



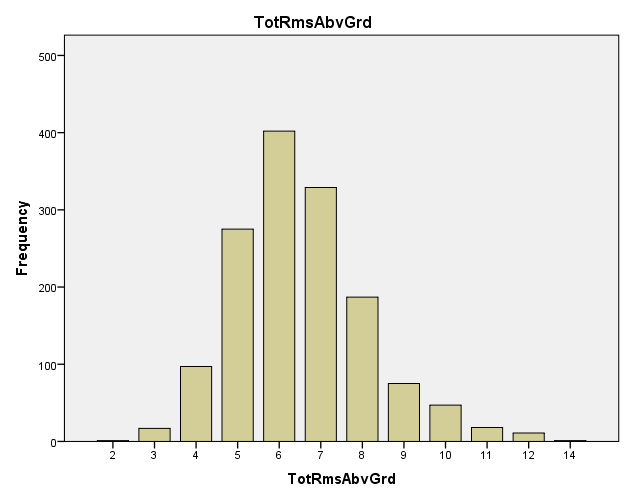
BsmtFullBath is a variable which represent Basement full bathrooms from which we can see that in some houses there is only 1bathroom in the basement and 2 bathrooms available as a rare case.



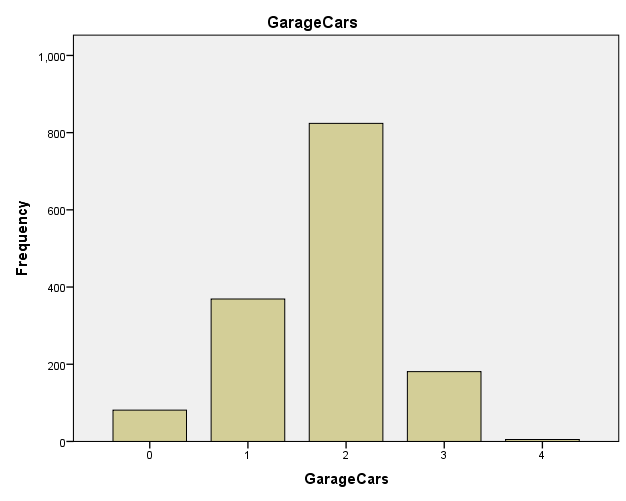
BedroomAbvGr is the Number of bedrooms above basement level. Most houses contain almost 2-4 bedrooms above basement level.



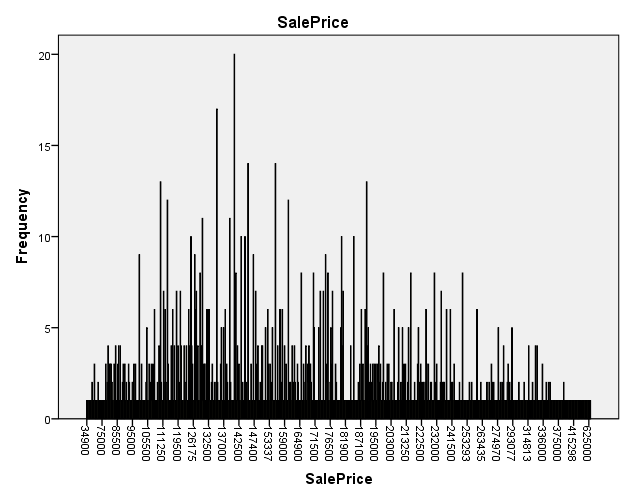
KitchenAbvGr is the number of kitchens contained above ground. Almost all houses contain 1 kitchen and only few houses have 2 kitchens.



TotRmsAbvGrd represents total rooms above grade (which does not include bathrooms). Most houses are filled with 4-9 rooms available.



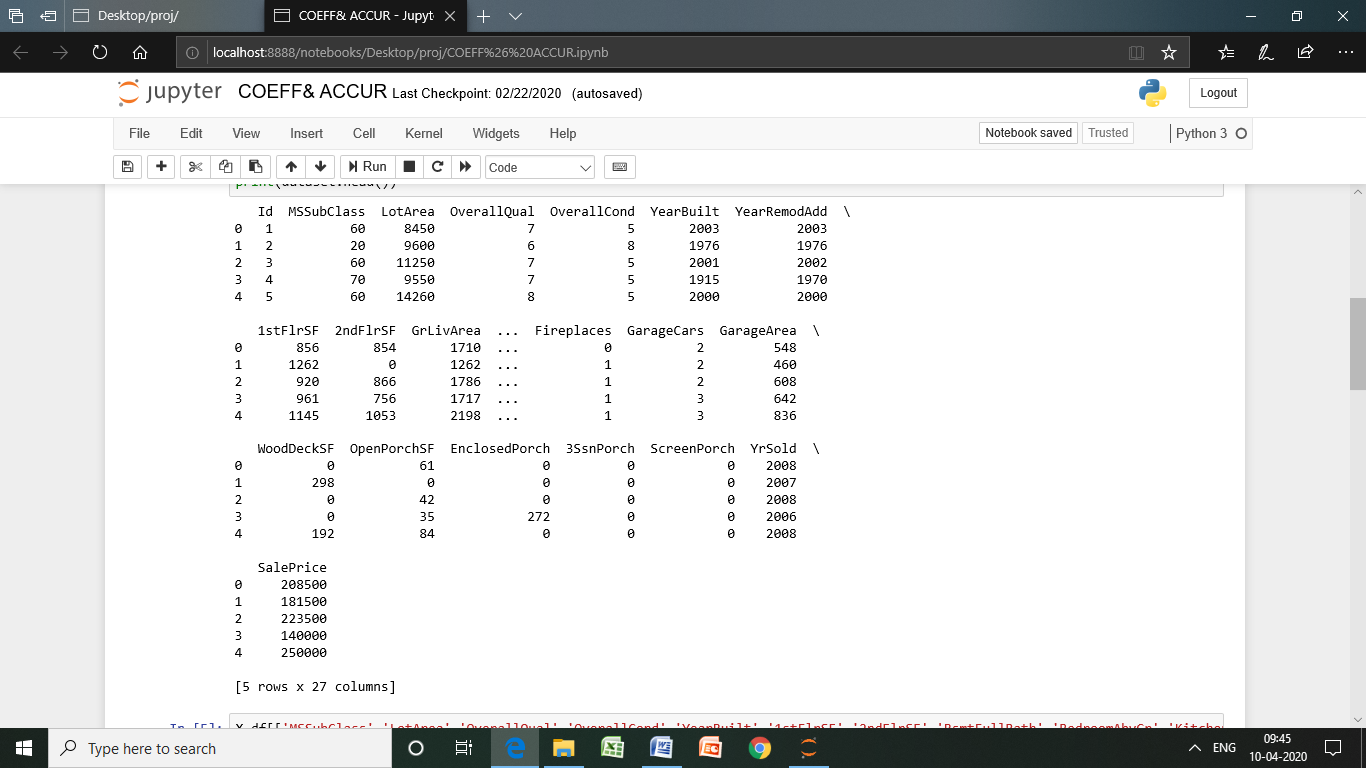
GarageCars is the size of garage in car capacity that is to be filled. Some houses don’t have garage car area and many houses have capacity of 2 cars and few houses have capacity of 1 car and very few have capacity of size to be 3.



SalePrice is the price of the property which is in dollars. This is the target variable that we're trying to predict.

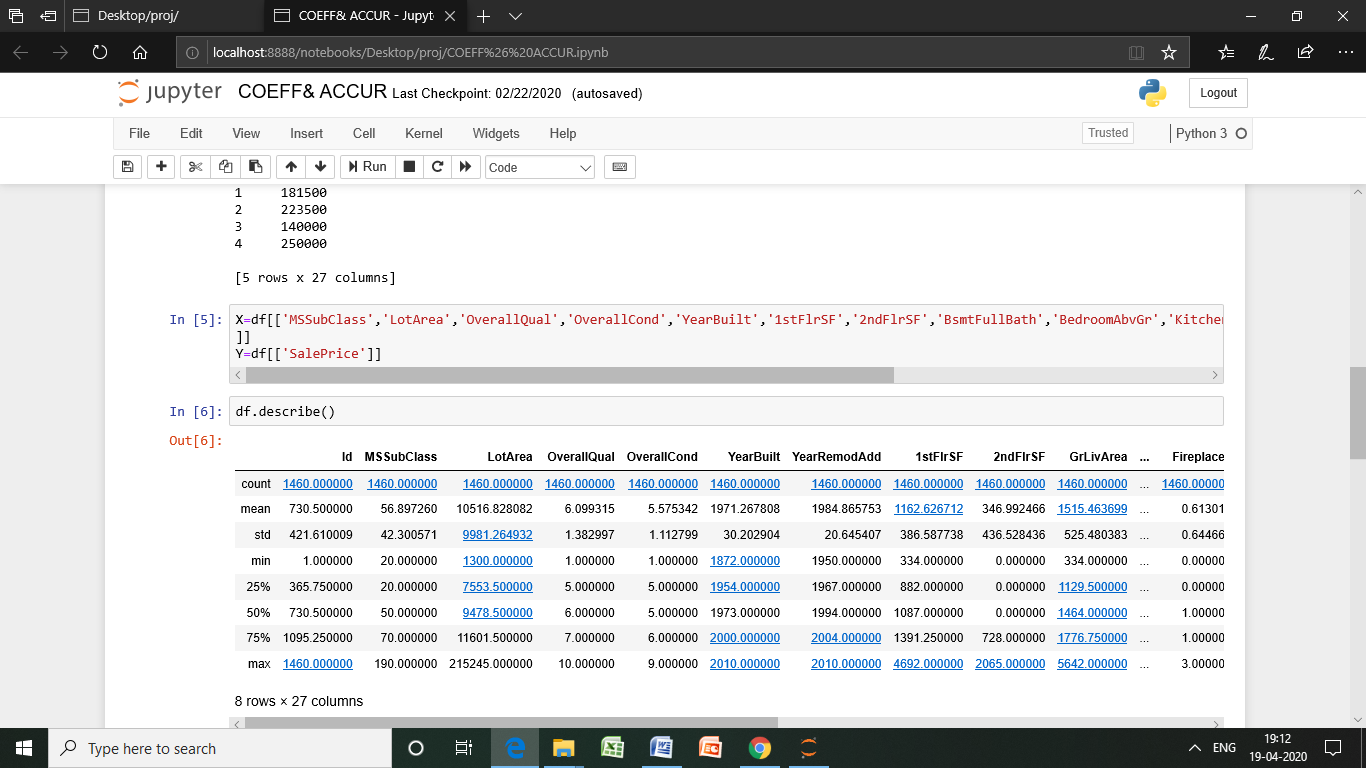
**REGRESSION TECHNIQUES**

By using Jupyter Notebook, we move on to further analysis for this dataset. While viewing the description of the dataset some of the variable names and only the first 5 observations of those variables are being shown just to see the overall view of imported data.



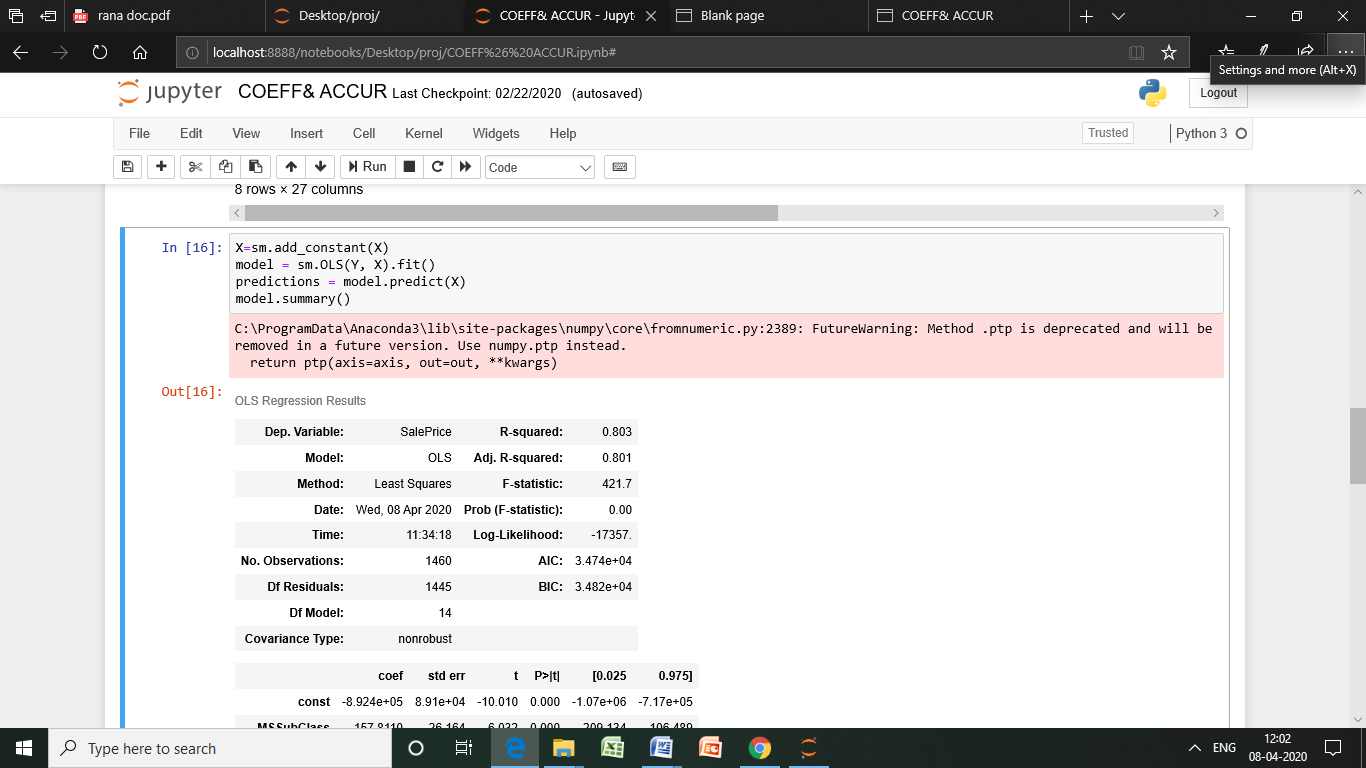
**Linear regression**

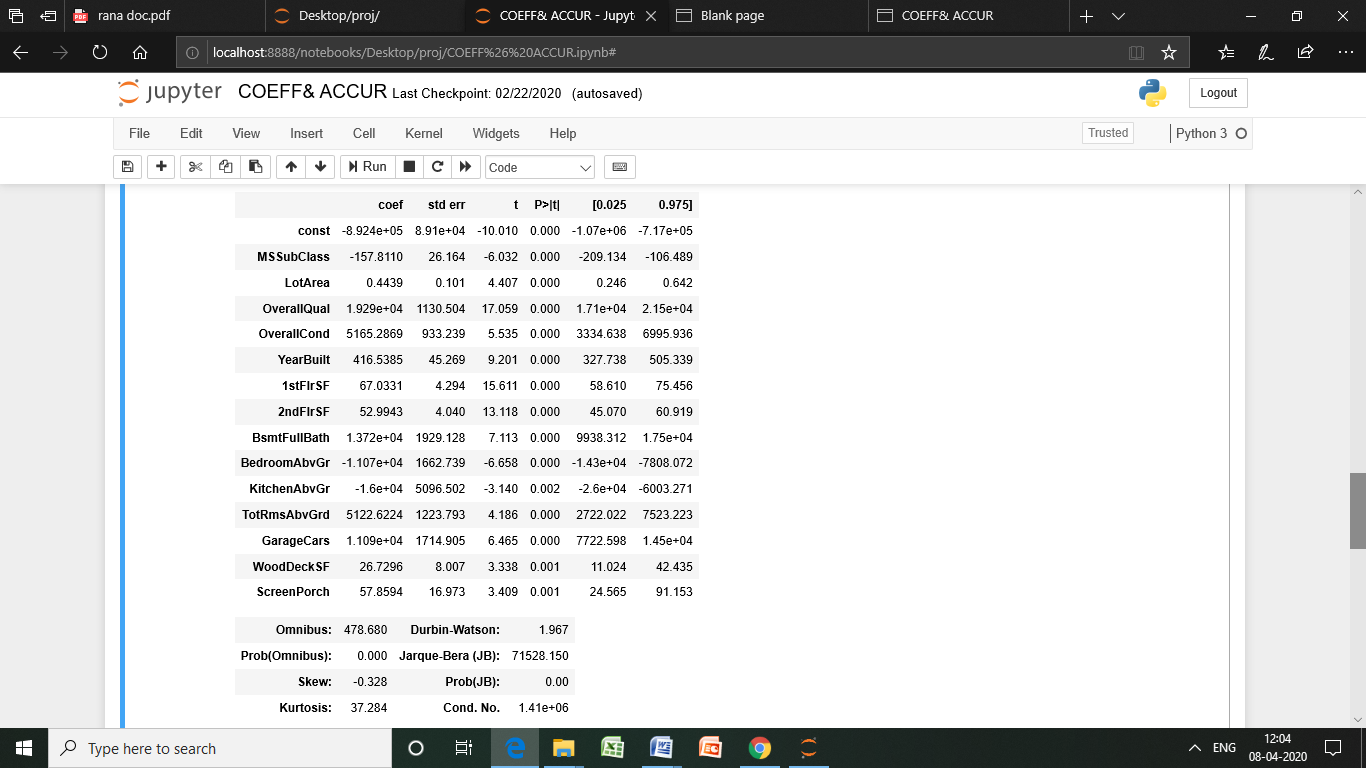
Assuming X to be the independent variable which contains all of the variables in the dataset and Y as the depenent variable (i.e., a variable that needs to be predicted using regression technique).



After assumption of variables, we predict the model using OLS (Optimum Least Square) estimator which gives the following results.

**OLS (Optimum Least Square)**



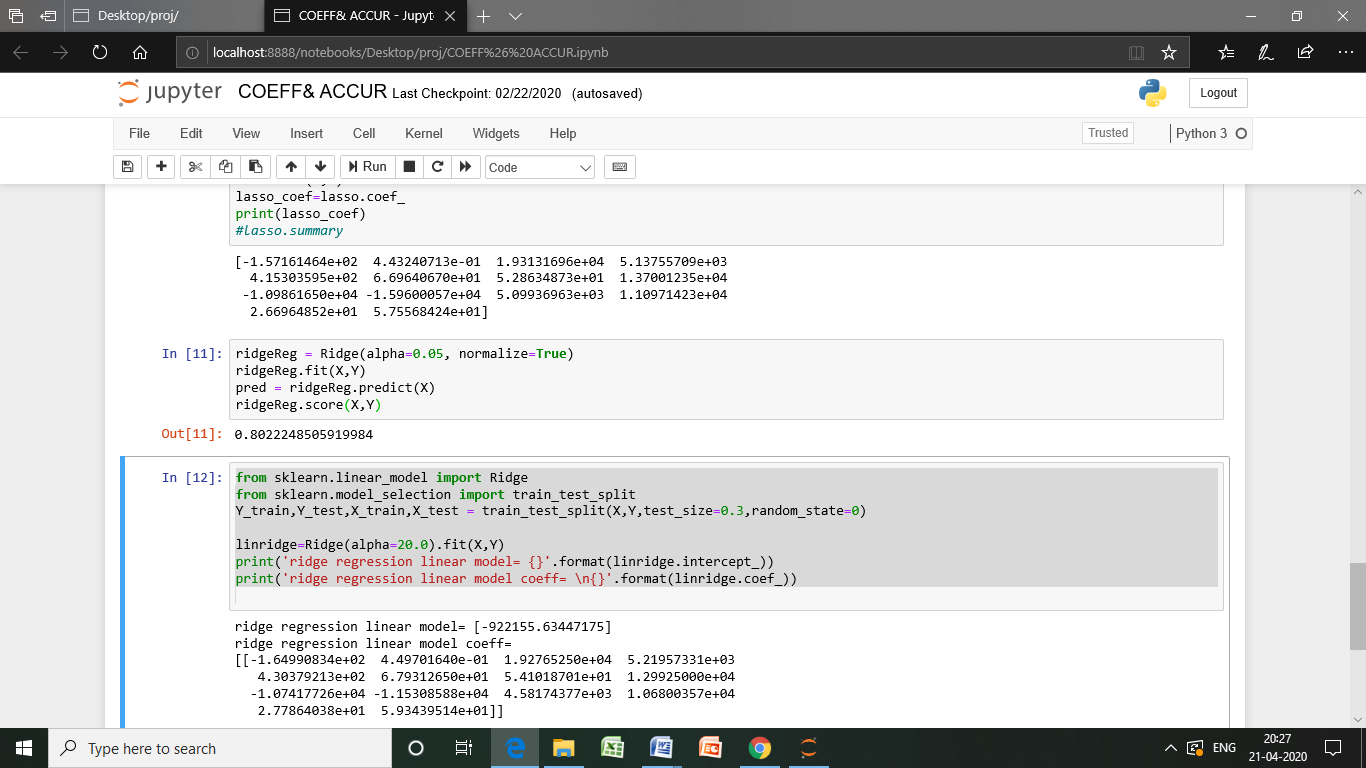


With this OLS regression results, the R square is 0.803 which defines an accuracy of 80.3% and the Adjusted R squared value ranges to be 0.801 that gives the accuracy of 80.1% for the house pricing data with 1460 observations and it also shows that the model used is the Optimum Least Square model with Least Square method. The P-values and the coefficients in this analysis display whether the relationship in the model is statistically significant or not and the nature of those relationships. Here, we have reduced the P-value all the variables that are not significant. The variables that are significant are more important for house price prediction. MSSubClass, LotArea, OverallQual, OverallCond, YearBuilt, 1stFlrSF, 2ndFlrSF, BsmtFullBath, BedroomAbvGr, KitchenAbvGr, TotRmsAbvGrd, GarageCars, WoodDeckSF, ScreenPorch are the variables which gives more impact for prediction of house. The coefficients describe the mathematical relationship between each independent variable and the dependent variable. The variables which have P value more than or near to 0.05 are being reduced to increase the accuracy level of the house pricing data.

After this model prediction with OLS regression for linear regression, we further on move to check the accuracy and see the difference between the coefficients of Linear regression, Ridge and LASSO regression.

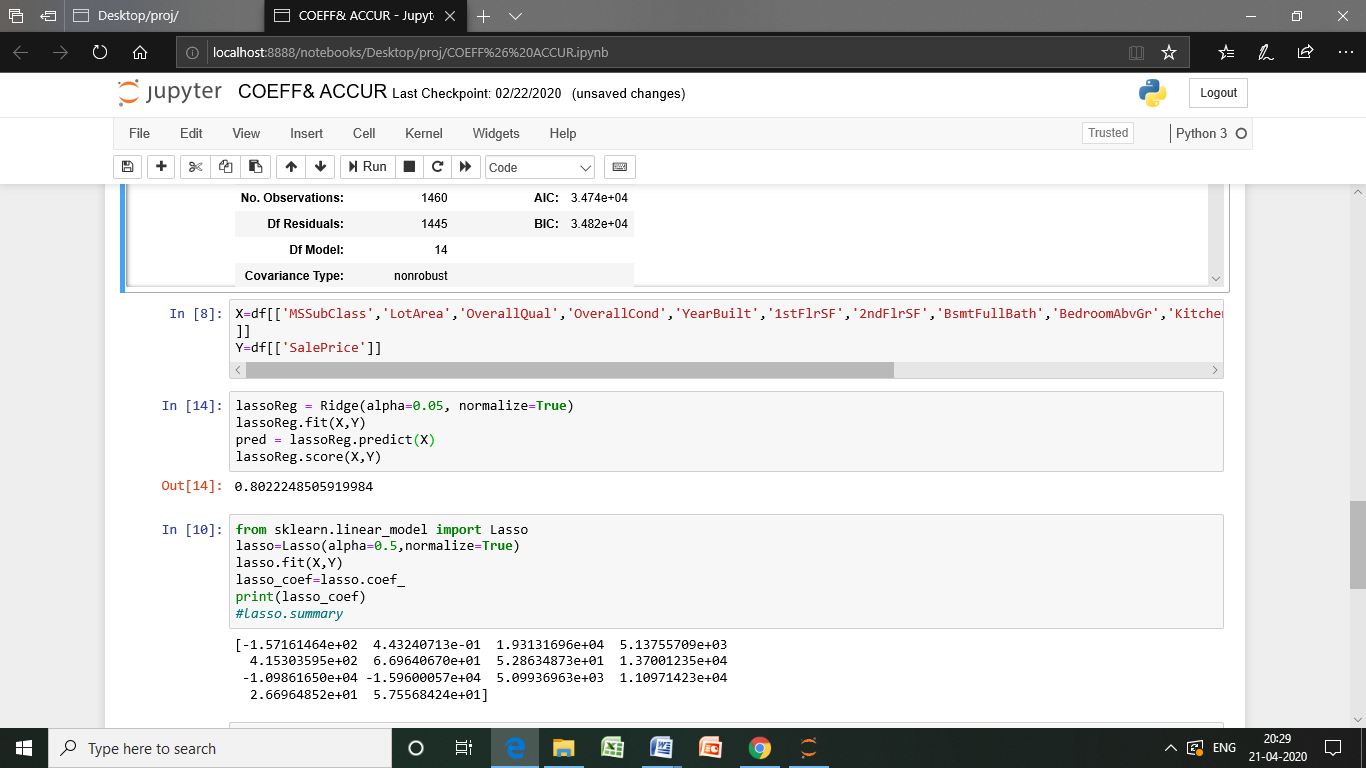
**Comparison of regression techniques**

Ridge regression



Here we acquired the accuracy of Ridge regression to be 80.22%.

LASSO regression



The accuracy of LASSO regression seems to be same as of ridge regression which is counted to be 80.22%. compared to linear regression we get better results with ridge and Lasso regression.

But, there seems no difference between Ridge and Lasso regression techniques since multicollinearity do not exist in this dataset. Difference for Ridge and Lasso regression exists only in big data analysis or when there is mulicollinearity between variables.

The following table gives us the differences between the coefficients that are estimated for linear regression, and advanced regression (i.e., Ridge and Lasso regression).

|  |  |  |  |
| --- | --- | --- | --- |
| **VARIABLES** | **COEFFICIENTS** | | |
| Linear regression | Ridge | Lasso |
| Const | -892400.00 | 0.00 | 0.00 |
| MSSubClass | -157.81 | -164.99 | -157.16 |
| LotArea | 0.44 | 0.45 | 0.44 |
| OverallQual | 19290.00 | 19276.53 | 19313.17 |
| OverallCond | 5165.29 | 5219.57 | 5137.56 |
| YearBuilt | 416.54 | 430.38 | 415.30 |
| 1stFlrSF | 67.03 | 67.93 | 66.96 |
| 2ndFlrSF | 52.99 | 54.10 | 52.86 |
| BsmtFullBath | 13720.00 | 12992.50 | 13700.12 |
| BedroomAbvGr | -11070.00 | -10741.77 | -10986.17 |
| KitchenAbvGr | -16000.00 | -11530.86 | -15960.01 |
| TotRmsAbvGrd | 5122.62 | 4581.74 | 5099.37 |
| GarageCars | 11090.00 | 10680.04 | 11097.14 |
| WoodDeckSF | 26.73 | 27.79 | 26.70 |
| ScreenPorch | 57.86 | 59.34 | 57.56 |

While comparing the coefficient of variables of Ridge and Lasso regression, we could clearly observe that the coefficient of Lasso regression has almost been minimised compared to as the Ridge regression. This shows that Lasso regression uses shrinkage and the shrinkage is where data values are shrunk towards a central point. Thus, we could clearly observe that Lasso and Ridge regression gives better results compared to linear regression and while comparing Lasso and Ridge regression, Lasso gives better and accurate results compared to Ridge regression. So, the home value prediction gives better results while it is being predicted with advanced regression techniques which are the Ridge and Lasso regression techniques.

**CHAPTER-5**

**FINDINGS AND CONCLUSION**

I took Zillow’s Home Value Prediction data from Kaggle which has 1460 observations with 15 variables to compare linear regression with advanced regression techniques (i.e., Ridge and Lasso regression). Before entering into the comparison for the regression techniques, I perform visualization of variables using bar chat in SPSS software. While working on comparison of variables, I initially checked the accuracy and coefficients of linear regression with OLS model I obtained an accuracy of 80.1% after reducing the insignificant variables from the dataset. Then moving on to check the accuracy of Ridge and LASSO regression I acquired accuracy of 80.2% from which I obviously see difference between linear and advanced regression techniques. Moving on to the further analysis on comparison, I collected coefficients of all there regression techniques from which I could see the shrinkage of coefficients which shrunk towards zero and it is very transparent that Ridge and LASSO regression perform a way better than Linear regression. When I compare Ridge and Lasso regression technique I prefer LASSO regression to use for prediction of Zillow’s home value data since variable shrinkage can been seen better compared to Ridge regression.

**BIBLIOGRAPHY**

* 1. Beginning Programming with Python For Dummies (For Dummies Series) by John Paul Mueller.

# Introduction to the analysis of learning algorithms: ridge regression and lasso by Ryota Tomioka.

# Python Data Analytics: Data Analysis and Science Using Pandas, matplotlib, and

# the Python Programming Language by Fabio Nelli.

* 1. “Linear, Ridge and Lasso Regression comprehensive guide for beginners”- <https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/>
  2. Python Programming. Python Programming for Beginners, Python Programming for Intermediates by Adam Stewart.
  3. “Visual Explanation of Ridge Regression and LASSO”- <https://www.slideshare.net/kaz_yos/visual-explanation-of-ridge-regression-and-lasso>

# Ridge/Lasso Regression, Model selection by Xuezhi Wang.

**APPENDIX**

import pandas as pd

import numpy as np

import statsmodels.api as sm

from sklearn import metrics

from sklearn.linear\_model import Ridge

from statsmodels.stats.outliers\_influence import variance\_inflation\_factor

from sklearn.metrics import mean\_squared\_error

df=pd.read\_csv('c:/Users/HP/Desktop/proj/house.csv')

df.shape

df.head()

dataset = pd.DataFrame(df)

print(dataset.head())

X=df[['MSSubClass','LotArea','OverallQual','OverallCond','YearBuilt','1stFlrSF','2ndFlrSF','BsmtFullBath','BedroomAbvGr','KitchenAbvGr','TotRmsAbvGrd','GarageCars','WoodDeckSF','ScreenPorch']]

Y=df[['SalePrice']]

df.describe()

X=sm.add\_constant(X)

model = sm.OLS(Y, X).fit()

predictions = model.predict(X)

model.summary()

lassoReg = Ridge(alpha=0.05, normalize=True)

lassoReg.fit(X,Y)

pred = lassoReg.predict(X)

#pred = ridgeReg.predict(np.array([9550,7,5,756,1,0,3,7,1,642]).reshape(1,10))

​

from sklearn.linear\_model import Lasso

lasso=Lasso(alpha=0.5,normalize=True)

lasso.fit(X,Y)

lasso\_coef=lasso.coef\_

print(lasso\_coef)

#lasso.summary

ridgeReg = Ridge(alpha=0.05, normalize=True)

ridgeReg.fit(X,Y)

pred = ridgeReg.predict(X)

#pred = ridgeReg.predict(np.array([9550,7,5,756,1,0,3,7,1,642]).reshape(1,10))

ridgeReg.score(X,Y)

#print(pred)

#pred.head()

from sklearn.linear\_model import Ridge

from sklearn.model\_selection import train\_test\_split

Y\_train,Y\_test,X\_train,X\_test = train\_test\_split(X,Y,test\_size=0.3,random\_state=0)

​

linridge=Ridge(alpha=20.0).fit(X,Y)

print('ridge regression linear model= {}'.format(linridge.intercept\_))

print('ridge regression linear model coeff= \n{}'.format(linridge.coef\_))

​