

Predictive modeling PROJECT REPORT

PM



May 16, 2022

Brijal Panchal

Contents

[Problem 1: Linear Regression 2](#_Toc103632822)

[1.1. Exploratory Data Analysis for Problem 1 2](#_Toc103632823)

[1.2. Build various iterations of the Linear Regression model using appropriate variable selection techniques for the full data. 9](#_Toc103632824)

[1.3. Split the data into training (70%) and test (30%). Build the various iterations of the Linear Regression models on the training data and use those models to predict on the test data using appropriate model evaluation metrics 15](#_Toc103632825)

[Problem 2: Logistic Regression 15](#_Toc103632826)

[2.1. Exploratory Data Analysis for Problem 2 16](#_Toc103632827)

[2.2. Build various iterations of the Logistic Regression model using appropriate variable selection techniques for the full data. Compare values of model selection criteria for proposed models. Compare, as many criteria as you feel are suitable. 21](#_Toc103632828)

[2.3. Split the data into training (70%) and test (30%). Build the various iterations of the Logistic Regression models on the training data and use those models to predict on the test data using appropriate model evaluation metrics. 24](#_Toc103632829)

# Problem 1: Linear Regression

You are hired by a company named Gem Stones Co Ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of approximately 27,000 pieces of cubic zirconia (which is an inexpensive synthesized diamond alternative with similar qualities of a diamond).

Your objective is to accurately predict prices of the zircon pieces. Since the company profits at a different rate at different price levels, for revenue management, it is important that prices are predicted as accurately as possible. At the same time, it is important to understand which of the predictors are more important in determining the price.

Data Discerption

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Carat | Carat weight of the cubic zirconia. |
| Cut | Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal. |
| Colour | Colour of the cubic zirconia. With D being the best and J the worst. |
| Clarity | Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst in terms of avg. price) IF, VVS1, VVS2, VS1, VS2, SI1, Sl2, l1 |
| Depth | The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter. |
| Table | The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter. |
| Price | The Price of the cubic zirconia. |
| X | Length of the cubic zirconia in mm. |
| Y | Width of the cubic zirconia in mm. |
| Z | Height of the cubic zirconia in mm. |

Sample of the dataset:



Dataset has 10 variables and 26967 samples in which cut, color and clarity are ordinal categorical variables rest 7 are continuous variable.

## Exploratory Data Analysis

* Rows present 34 duplicate data in data frame. After removing, the duplicated rows there are 26933 samples are present for prediction.
* 697 missing values are there in the depth variable data column, Lets replace missing values of depth with mean of it, which is 61.745 as its mean and median are almost same for it.

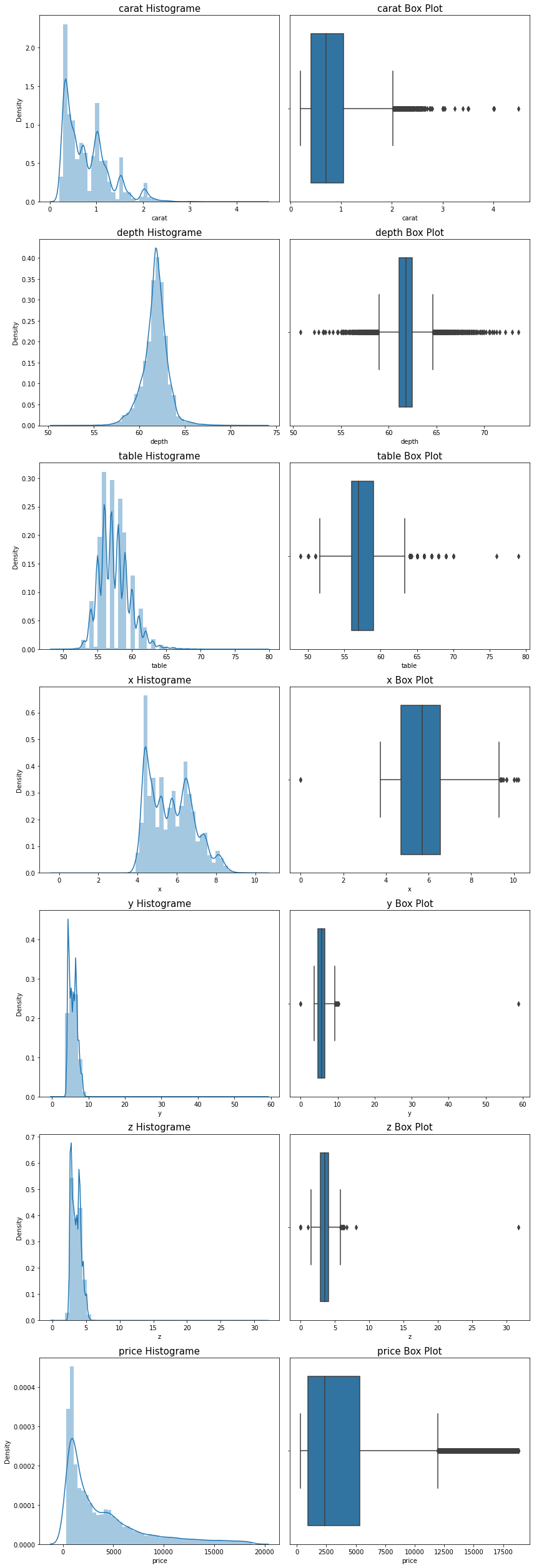
Summary of description

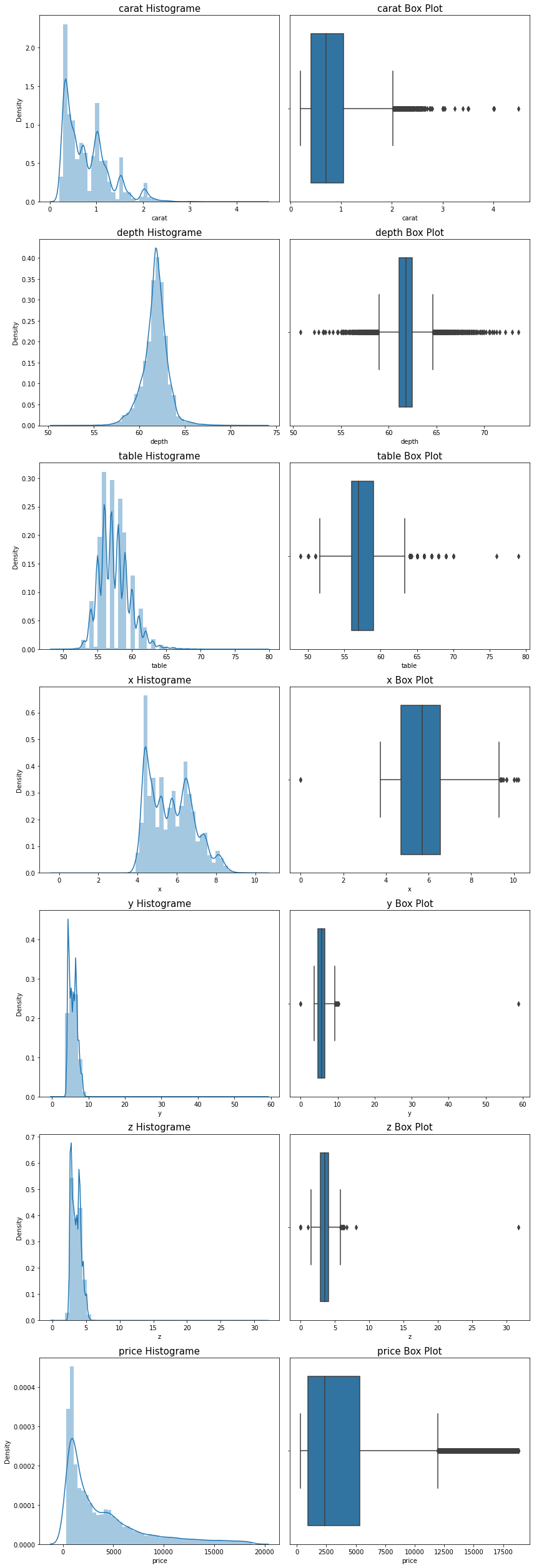


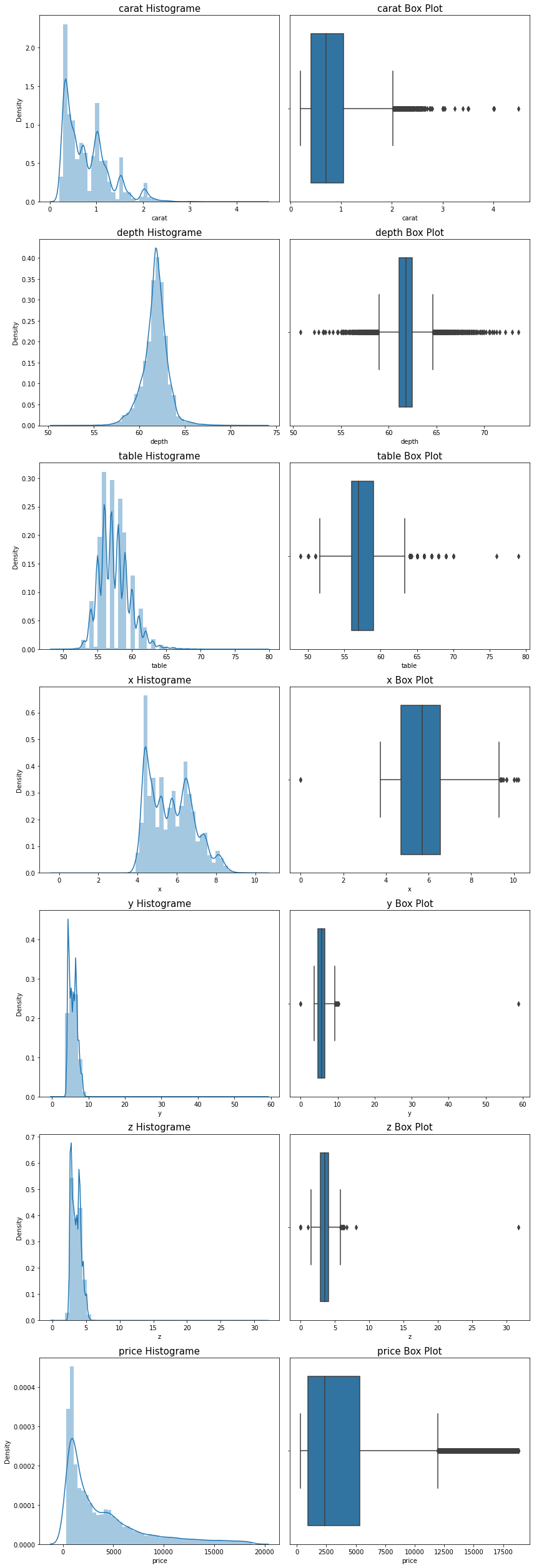
* From table we can see that carat has minimum value of 0.2, maximum value is 4.5 and mean and median are 0.7980 and 0.7 respectively while standard deviation is around 0.477
* depth has minimum value of 61.1, maximum value is 73.6 and mean and median are 61.74 and 62.5 respectively while standard deviation is around 1.39
* table has minimum value of 49, maximum value is 79 and mean and median are 57.45 and 57 respectively while standard deviation is around 2.23
* x has minimum value of 0, maximum value is 10.23 and mean and median are 5.73 and 5.69 respectively while standard deviation is around 1.127
* y has minimum value of 0, maximum value is 58.9 and mean and median are 5.73 and 5.7 respectively while standard deviation is around 1.165
* z has minimum value of 0, maximum value is 31.8 and mean and median are 3.53 and 3.52 respectively while standard deviation is around 0.719
* price vary from 362 to 18818 with average of 3937 and median of 2375 with standard deviation of around 4022.52

Univariate Analysis:

* Histogram and Box-plot of continuous variables

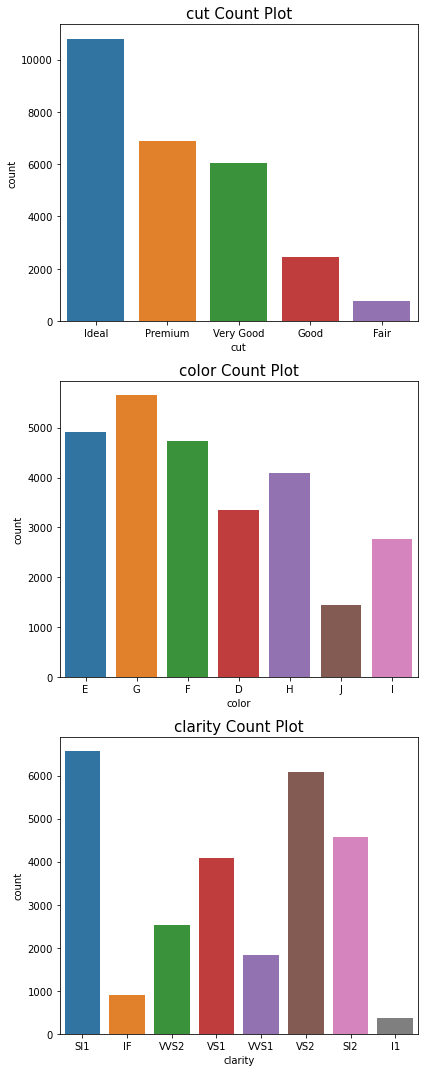


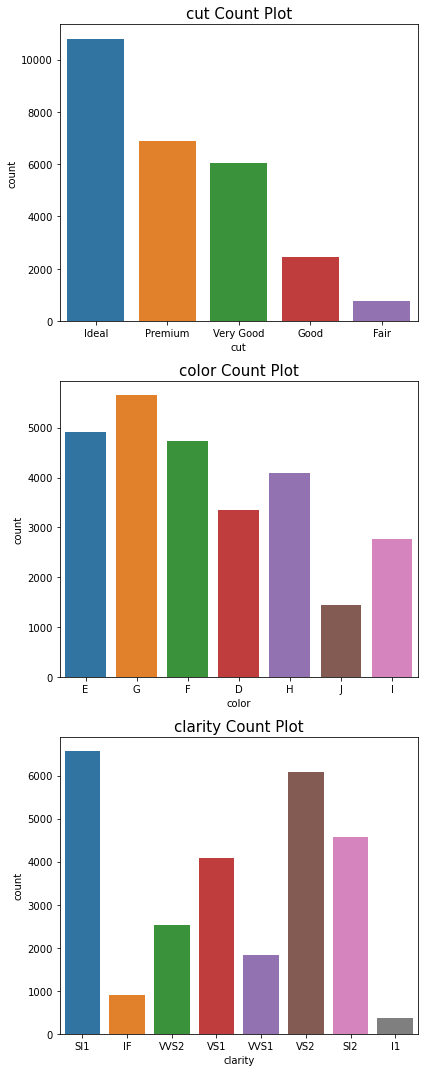




From above graph, we see that all continuous variables have outliers in their data in which price, cart and depth are heavily outliers in the data. depth follows normal distribution. While cart, table, x, y, z and price are right skewed data.

* Count plot for categorical variables

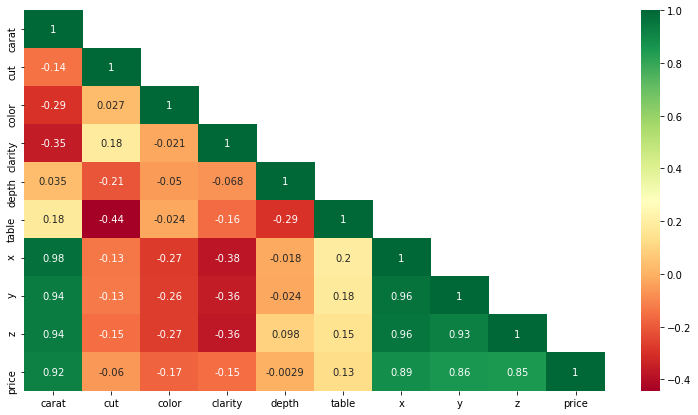




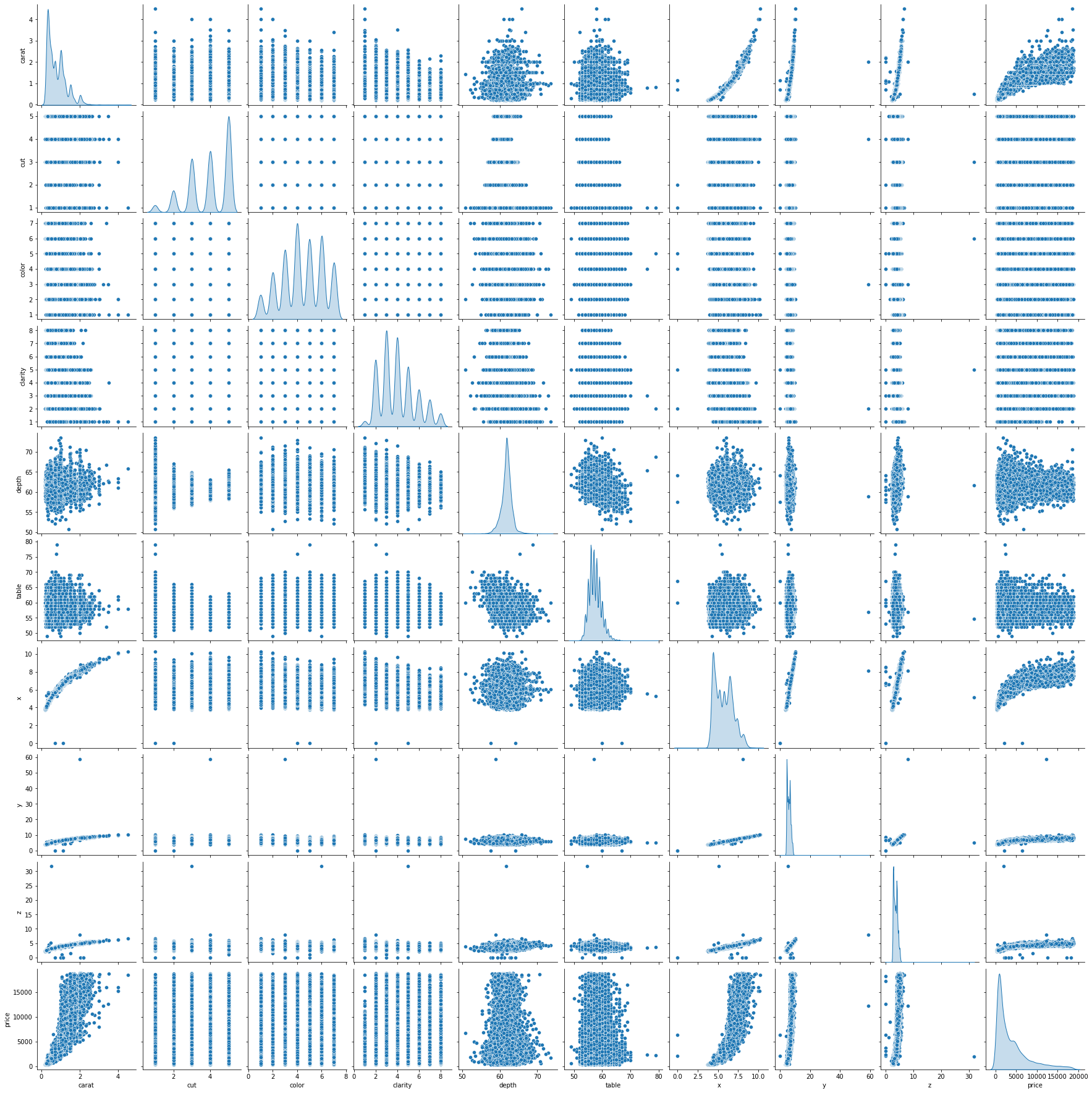
* There are 5 types of category available in cut variable which are as follow, Lets apply label encoding with order of 1 to 5 for Fair, Good, Very Good, Premium, Ideal.
  + Ideal 10816
  + Premium 6899
  + Very Good 6030
  + Good 2441
  + Fair 781
* There are 7 types of category available in color variable which are as follow, Lets apply label encoding with order of 1 to 7 for J to D.
  + G 5661
  + E 4917
  + F 4729
  + H 4102
  + D 3344
  + I 2771
  + J 1443
* There are 8 types of category available in clarity variable which are as follow, Lets apply label encoding with order of 8 to 1 for IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1.
  + SI1 6571
  + VS2 6099
  + SI2 4575
  + VS1 4093
  + VVS2 2531
  + VVS1 1839
  + IF 894
  + I1 365

Multivariate analysis:

* Heat map



* From heat map we can see that x,y,z are highly correlated as well inter co-related with each other.
* There are many variables, which are possible to create multicollinearity.
* Depth and cut variable has very low correlation with dependent variable price.
* Pair plot



From pair plot, we can see that there is positive relation present between x, y, z and carat predictors.

## Build various iterations of the Linear Regression model using appropriate variable selection techniques for the full data.

* Model 1 - Build the Liner Regression Model using all the variables on the full data using statsmodels library.
  + Summary of statistics:

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | price | **R-squared:** | 0.908 |
| **Model:** | OLS | **Adj. R-squared:** | 0.908 |
| **Method:** | Least Squares | **F-statistic:** | 2.966e+04 |
| **Date:** | Tue, 17 May 2022 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 11:44:44 | **Log-Likelihood:** | -2.2957e+05 |
| **No. Observations:** | 26933 | **AIC:** | 4.592e+05 |
| **Df Residuals:** | 26923 | **BIC:** | 4.592e+05 |
| **Df Model:** | 9 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Intercept** | 3373.8815 | 607.364 | 5.555 | 0.000 | 2183.417 | 4564.346 |
| **carat** | 1.091e+04 | 75.775 | 143.999 | 0.000 | 1.08e+04 | 1.11e+04 |
| **cut** | 117.5691 | 8.149 | 14.428 | 0.000 | 101.597 | 133.541 |
| **color** | 327.9433 | 4.605 | 71.209 | 0.000 | 318.917 | 336.970 |
| **depth** | -82.7053 | 6.667 | -12.406 | 0.000 | -95.772 | -69.638 |
| **clarity** | 500.4353 | 4.996 | 100.174 | 0.000 | 490.644 | 510.227 |
| **table** | -30.2031 | 4.198 | -7.194 | 0.000 | -38.432 | -21.974 |
| **x** | -916.2283 | 44.485 | -20.597 | 0.000 | -1003.420 | -829.036 |
| **y** | 29.6264 | 23.718 | 1.249 | 0.212 | -16.862 | 76.115 |
| **z** | -42.9107 | 38.744 | -1.108 | 0.268 | -118.852 | 33.030 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 5930.667 | **Durbin-Watson:** | 2.015 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 274511.173 |
| **Skew:** | -0.101 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 18.639 | **Cond. No.** | 6.98e+03 |

* + Check for multicollinearity in the predictor variables using Variance Inflation Factor (VIF): VIF of all predictors are 13.62 which is > 5, So all independent variables are correlated with each other
  + Adjusted R squared value is 0.908.
  + P-value of y and z is more than 0.05 as well as the condition number is large, 6.98e+03. This might indicate that there are strong multicollinearity or other numerical problems
* Model 2 - Drop the variable which has highest the P-value (i.e. z) > 0.05. Build a second iteration of the model on the full data. Check the VIF values of the predictor variables
  + Summary of statistics:

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | price | **R-squared:** | 0.908 |
| **Model:** | OLS | **Adj. R-squared:** | 0.908 |
| **Method:** | Least Squares | **F-statistic:** | 3.336e+04 |
| **Date:** | Tue, 17 May 2022 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 11:44:55 | **Log-Likelihood:** | -2.2957e+05 |
| **No. Observations:** | 26933 | **AIC:** | 4.592e+05 |
| **Df Residuals:** | 26924 | **BIC:** | 4.592e+05 |
| **Df Model:** | 8 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Intercept** | 3517.9473 | 593.273 | 5.930 | 0.000 | 2355.101 | 4680.794 |
| **carat** | 1.091e+04 | 75.773 | 143.994 | 0.000 | 1.08e+04 | 1.11e+04 |
| **cut** | 117.7632 | 8.147 | 14.455 | 0.000 | 101.795 | 133.732 |
| **color** | 327.9046 | 4.605 | 71.202 | 0.000 | 318.878 | 336.931 |
| **depth** | -85.1820 | 6.280 | -13.563 | 0.000 | -97.492 | -72.872 |
| **clarity** | 500.3665 | 4.995 | 100.168 | 0.000 | 490.575 | 510.157 |
| **table** | -30.0866 | 4.197 | -7.168 | 0.000 | -38.313 | -21.860 |
| **x** | -939.1701 | 39.368 | -23.856 | 0.000 | -1016.333 | -862.008 |
| **y** | 26.4919 | 23.549 | 1.125 | 0.261 | -19.665 | 72.649 |
| **Omnibus:** | 5931.279 | **Durbin-Watson:** | 2.015 |  |  |  |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 274645.211 |  |  |  |
| **Skew:** | -0.101 | **Prob(JB):** | 0.00 |  |  |  |
| **Kurtosis:** | 18.643 | **Cond. No.** | 6.81e+03 |  |  |  |

* + Adjusted R squared value is 0.908.
  + P-value of y is more than 0.05 as well as the condition number is large, 6.81e+03. This might indicate that there are strong multicollinearity or other numerical problems
  + Check for multicollinearity in the predictor variables using Variance Inflation Factor (VIF): VIF of all predictors are 13.62 which is > 5, So all independent variables are correlated with each other
* Model 3 - Drop the variable which has highest the P-value (i.e. y) > 0.05. Build a third iteration of the model on the full data. Check the VIF values of the predictor variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | price | **R-squared:** | 0.908 |
| **Model:** | OLS | **Adj. R-squared:** | 0.908 |
| **Method:** | Least Squares | **F-statistic:** | 3.813e+04 |
| **Date:** | Tue, 17 May 2022 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 11:45:01 | **Log-Likelihood:** | -2.2957e+05 |
| **No. Observations:** | 26933 | **AIC:** | 4.592e+05 |
| **Df Residuals:** | 26925 | **BIC:** | 4.592e+05 |
| **Df Model:** | 7 |  |  |
| **Covariance Type:** | nonrobust |  |  |

* + Summary of statistics:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Intercept** | 3553.6690 | 592.426 | 5.999 | 0.000 | 2392.483 | 4714.855 |
| **carat** | 1.091e+04 | 75.747 | 144.074 | 0.000 | 1.08e+04 | 1.11e+04 |
| **cut** | 117.5225 | 8.144 | 14.430 | 0.000 | 101.559 | 133.486 |
| **color** | 327.9120 | 4.605 | 71.204 | 0.000 | 318.885 | 336.939 |
| **depth** | -85.4990 | 6.274 | -13.627 | 0.000 | -97.797 | -73.201 |
| **clarity** | 500.4164 | 4.995 | 100.181 | 0.000 | 490.626 | 510.207 |
| **table** | -30.2846 | 4.193 | -7.222 | 0.000 | -38.504 | -22.065 |
| **x** | -913.6864 | 32.197 | -28.378 | 0.000 | -976.795 | -850.578 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 5934.455 | **Durbin-Watson:** | 2.015 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 275066.329 |
| **Skew:** | -0.102 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 18.655 | **Cond. No.** | 6.78e+03 |

* + Adjusted R squared value is 0.908.
  + The condition number is large, 6.81e+03. This might indicate that there are strong multicollinearity or other numerical problems
  + Check for multicollinearity in the predictor variables using Variance Inflation Factor (VIF): VIF of x predictors is infinite which is > 5, So x predictor has multicollinearity with other predictors
* Model 4 - Drop the variable which has the highest VIF (x variable) in the third iteration of the model. Build a forth iteration of the model. Check the VIF values of the predictor variables

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Intercept** | -2680.1408 | 558.350 | -4.800 | 0.000 | -3774.535 | -1585.747 |
| **carat** | 8823.4707 | 18.009 | 489.940 | 0.000 | 8788.171 | 8858.770 |
| **cut** | 120.2370 | 8.264 | 14.549 | 0.000 | 104.038 | 136.436 |
| **color** | 323.7082 | 4.671 | 69.300 | 0.000 | 314.553 | 332.864 |
| **depth** | -44.7045 | 6.198 | -7.213 | 0.000 | -56.852 | -32.557 |
| **clarity** | 522.7729 | 5.006 | 104.435 | 0.000 | 512.961 | 532.584 |
| **table** | -29.1542 | 4.255 | -6.851 | 0.000 | -37.495 | -20.813 |

* + Summary of statistics:

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | price | **R-squared:** | 0.906 |
| **Model:** | OLS | **Adj. R-squared:** | 0.906 |
| **Method:** | Least Squares | **F-statistic:** | 4.306e+04 |
| **Date:** | Tue, 17 May 2022 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 11:45:07 | **Log-Likelihood:** | -2.2996e+05 |
| **No. Observations:** | 26933 | **AIC:** | 4.599e+05 |
| **Df Residuals:** | 26926 | **BIC:** | 4.600e+05 |
| **Df Model:** | 6 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 5223.591 | **Durbin-Watson:** | 2.011 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 98243.939 |
| **Skew:** | 0.419 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 12.319 | **Cond. No.** | 6.28e+03 |

* + Adjusted R squared value is 0.906.
  + The condition number is large, 6.28e+03. This might indicate that there are strong multicollinearity or other numerical problems
  + Check for multicollinearity in the predictor variables using Variance Inflation Factor (VIF): VIF of all predictors are < 5, So there is no multicollinearity in predictors
* Hence best model is model 4 in which there is no multi collinearity present in predictors and having adjusted R squared value is 0.906

## Split the data into training (70%) and test (30%). Build the various iterations of the Linear Regression models on the training data and use those models to predict on the test data using appropriate model evaluation metrics

* Split the data into training (70%) and test (30%): with random\_state as 1, there are 188530 samples spliced as train data samples and 80800 samples spliced as test data samples
* Note: sklearn.linear\_model library is used for linear regression model building
* Model 1 - Building the model on the training data on whole data:
  + Train Accuracy Score : 0.9085604983839957
  + Test Accuracy Score : 0.9078450452130498
* Model 2 - Building the model on the training data on data except z predictor:
  + Train Accuracy Score : 0.908555196218629
  + Test Accuracy Score : 0.9078435412487519
* Model 3 - Building the model on the training data on data except y,z predictor:
  + Train Accuracy Score : 0.9085546719764409
  + Test Accuracy Score : 0.9078376151287759
* Model 4 - Building the model on the training data on data except x,y,z predictor:
  + Train Accuracy Score 0.9056128506663244
  + Test Accuracy Score : 0.9055879513287518
* RMSE score on training and testing data samples on each models:

|  |  |  |
| --- | --- | --- |
|  | **RMSE Training Data** | **RMSE Test Data** |
| **Model 1** | 1212.41848 | 1230.27385 |
| **Model 2** | 1212.45363 | 1230.28389 |
| **Model 3** | 1212.4571 | 1230.32344 |
| **Model 4** | 1231.80526 | 1245.24887 |

* From accuracy score and RMSE score model 3 would be the great fit as compare to model 4 as its accuracy is dropped in model 4

Conclusion: the final model of Part (II) and the proposed one in Part (III) are different. After calculating the accuracy and RMSE score of model 4 in Part (III), we can see that the accuracy is dropping in model 4 as compared to model 3. So we should go with model 3 as compared to model 4 having predictor carat, cut, color, clarity, depth, table and x.

# Problem 2: Logistic Regression

You are hired by a tour and travel agency which deals in selling holiday packages. You are provided details of 872 employees of a company. Among these employees, some opted for the package and some didn't. You have to help the company in predicting whether an employee will opt for the package or not on the basis of the information given in the data set. Also, find out the important factors on the basis of which the company will focus on particular employees to sell their packages.

Data Discerption

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Holiday\_Package | Opted for Holiday Package yes/no? |
| Salary | Employee salary |
| age | Age in years |
| edu | Years of formal education |
| no\_young\_children | The number of young children (younger than 7 years) |
| no\_older\_children | Number of older children |
| foreign | foreigner Yes/No |

Sample of the dataset:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Holliday\_Package** | **Salary** | **age** | **educ** | **no\_young\_children** | **no\_older\_children** | **foreign** |
| **1** | no | 48412 | 30 | 8 | 1 | 1 | no |
| **2** | yes | 37207 | 45 | 8 | 0 | 1 | no |
| **3** | no | 58022 | 46 | 9 | 0 | 0 | no |
| **4** | no | 66503 | 31 | 11 | 2 | 0 | no |
| **5** | no | 66734 | 44 | 12 | 0 | 2 | no |

Dataset has 7 variables and 872 samples. Dataset has six predictors with Holiday\_Package as categorical variable and rest are continuous variable. Foreign is dependent and categorical variable.

## Exploratory Data Analysis for Problem 2

* Rows present no duplicate data in data frame
* No missing value present in data frame

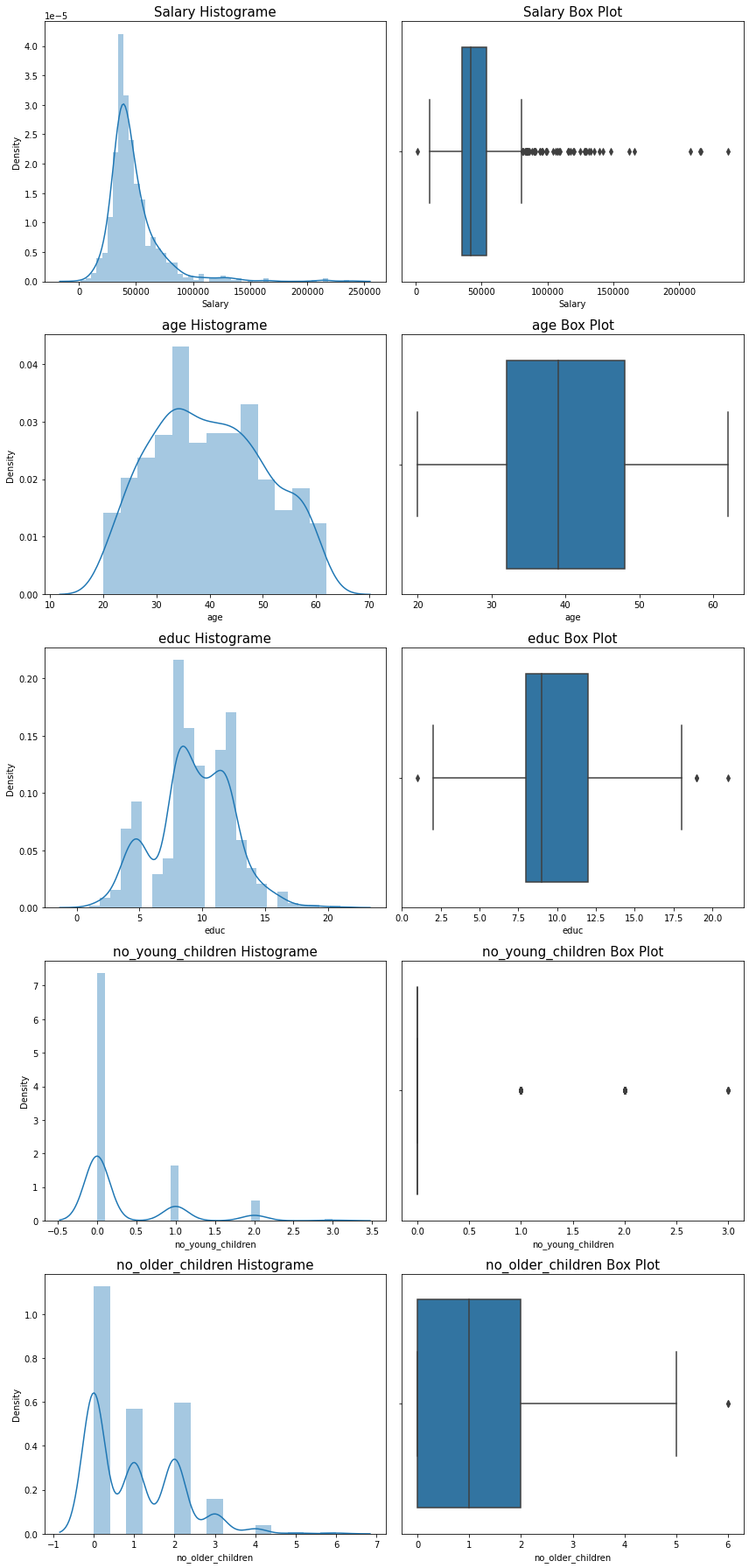
Summary of description

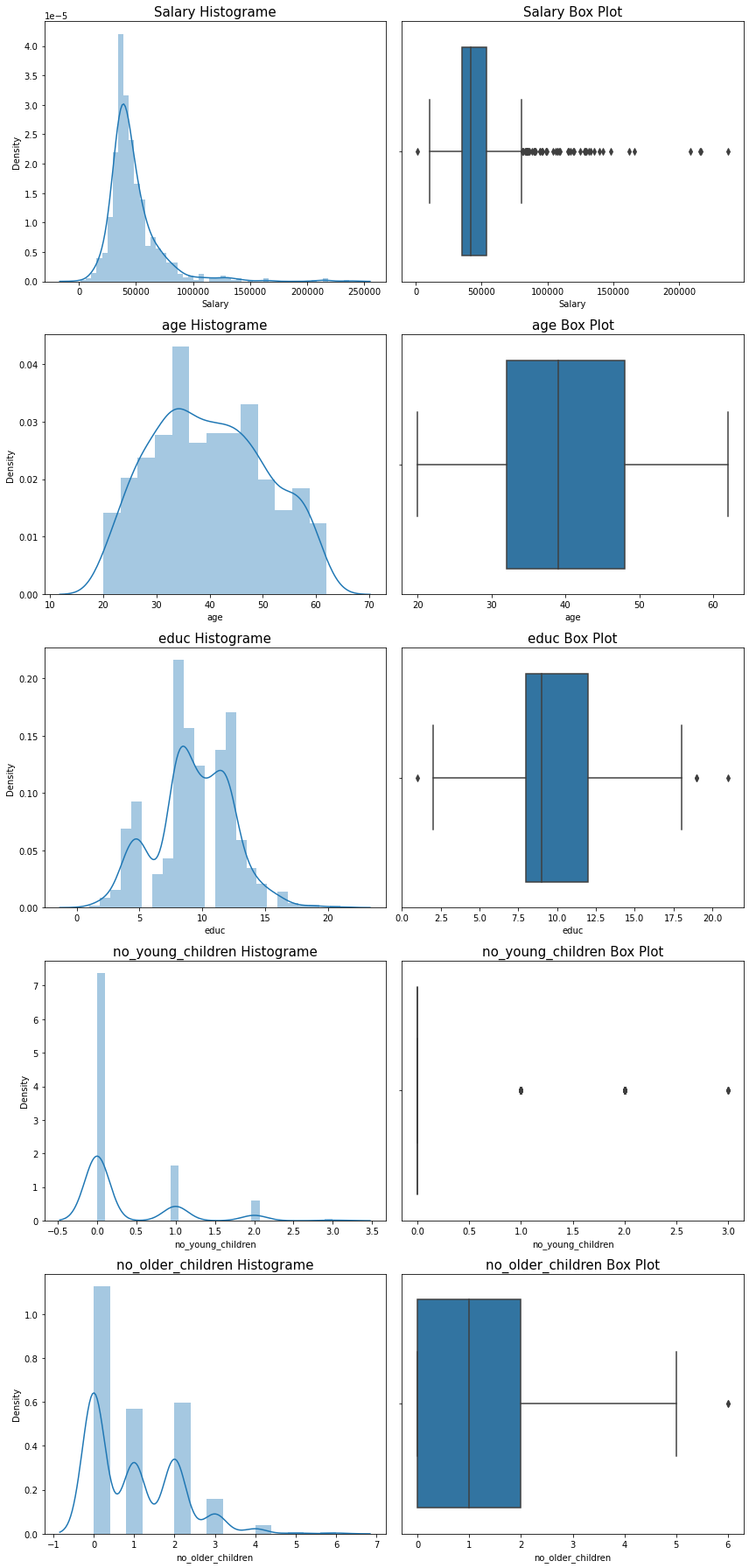
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **Salary** | 872 | 47729.172 | 23418.6685 | 1322 | 35324 | 41903.5 | 53469.5 | 236961 |
| **age** | 872 | 39.955275 | 10.551675 | 20 | 32 | 39 | 48 | 62 |
| **educ** | 872 | 9.307339 | 3.036259 | 1 | 8 | 9 | 12 | 21 |
| **no\_young\_children** | 872 | 0.311927 | 0.61287 | 0 | 0 | 0 | 0 | 3 |
| **no\_older\_children** | 872 | 0.982798 | 1.086786 | 0 | 0 | 1 | 2 | 6 |

* From table we can see that price vary from 1322 to 236961 with average of 47729.172 and median of 41903.5 with standard deviation of around 23418.6685
* age vary from 20 to 62 with average of 39.95 and median of 39 with standard deviation of around 10.55
* educ has minimum value of 1, maximum value is 21 and mean and median are 9.30 and 9 respectively while standard deviation is around 3.03
* no\_young\_children has minimum value of 0, maximum value is 3 and mean and median are 0.311 and 0 respectively while standard deviation is around 0.6128
* no\_older\_children has minimum value of 0, maximum value is 6 and mean and median are 0.98 and 1 respectively while standard deviation is around 1.086

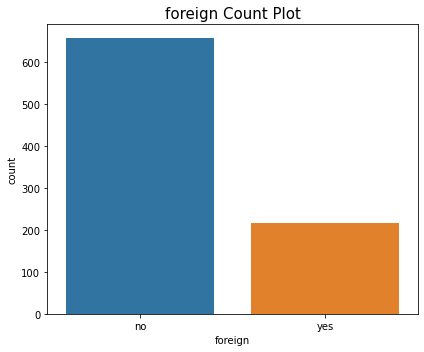
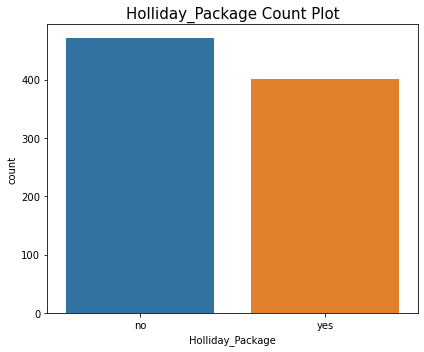
Univariate Analysis:

* Histogram and Box-plot of continuous variables





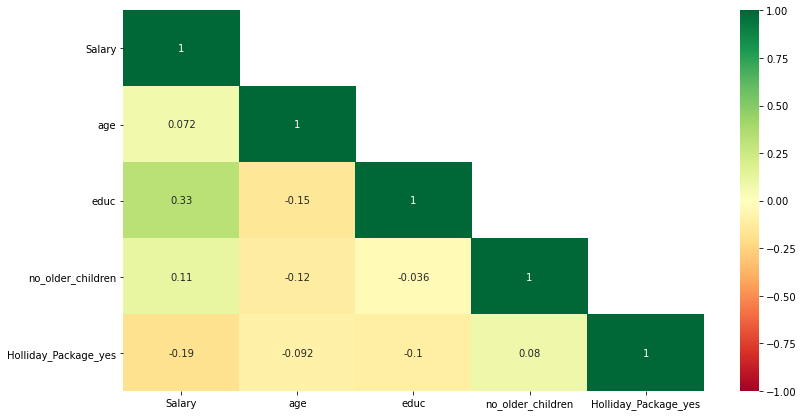
* Age and education is normally distributed while rest predictor are left skewed data samples
* All predictors are having outliers present in data.
* Dropping the variables of no\_young\_children as the majority of the values are zero.
* Count plot for categorical variables



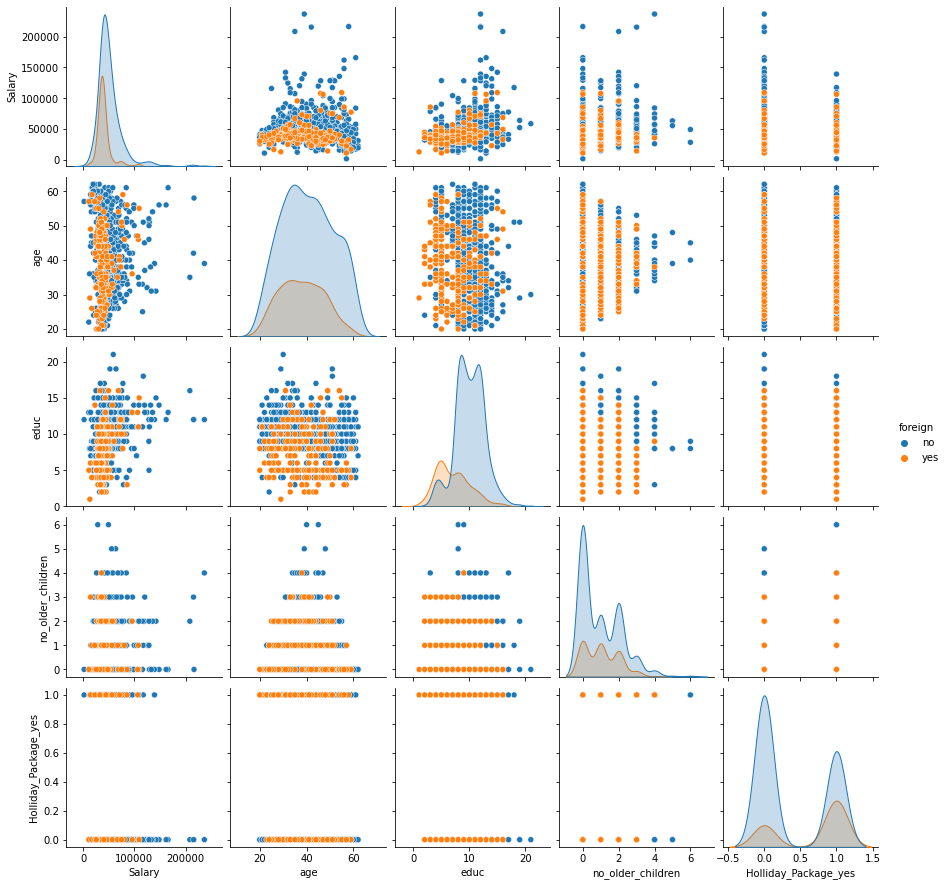
* There are 471 samples present as not opted for Holiday Package, while 401 samples are present with opted for Holiday Package
* There are 656 samples present as not foreigner, while 216 samples are present as foreigner.
* Lets apply label encoding method called ‘One-Hot encoding’ on Holliday\_Package predictor and drop Holliday\_Package\_no column

Multivariate analysis:

* Heat map



* + From heat map we can see there is weak co-relation between education and salary predictor
* Pair plot



* + Frome pair plot we can see there is no any strong co-relation present in predictor data set. So there is very strong possibilities that mullticolinearity present in dataset for predictors.
  + Converting the 'foreign' Variable into numeric by replacing no as 0 and yes as 1

## Build various iterations of the Logistic Regression model using appropriate variable selection techniques for the full data. Compare values of model selection criteria for proposed models. Compare, as many criteria as you feel are suitable.

* Model 1 - Build the Logistic Regression Model using all the variables on the full data using stats models library
  + Summary of statistics:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dep. Variable:** | foreign | **No. Observations:** |  | 872 |
| **Model:** | Logit | **Df Residuals:** |  | 866 |
| **Method:** | MLE | **Df Model:** |  | 5 |
| **Date:** | Tue, 17 May 2022 | **Pseudo R-squ.:** |  | 0.2469 |
| **Time:** | 15:50:25 | **Log-Likelihood:** |  | -367.62 |
| **converged:** | True | **LL-Null:** |  | -488.15 |
| **Covariance Type:** | nonrobust | **LLR p-value:** |  | 4.558e-50 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **Intercept** | 4.0980 | 0.591 | 6.931 | 0.000 | 2.939 | 5.257 |
| **Salary** | -1.162e-05 | 6.75e-06 | -1.722 | 0.085 | -2.49e-05 | 1.61e-06 |
| **age** | -0.0464 | 0.009 | -4.899 | 0.000 | -0.065 | -0.028 |
| **educ** | -0.3966 | 0.038 | -10.379 | 0.000 | -0.472 | -0.322 |
| **no\_older\_children** | -0.0502 | 0.087 | -0.578 | 0.563 | -0.221 | 0.120 |
| **Holliday\_Package\_yes** | 1.0770 | 0.190 | 5.674 | 0.000 | 0.705 | 1.449 |
|  |  |  |  |  |  |  |

* + Check for multicollinearity in the predictor variables using Variance Inflation Factor (VIF):
    - Salary VIF = 1.2
    - age VIF = 1.07
    - educ VIF = 1.17
    - no\_older\_children VIF = 1.05
    - Holliday\_Package\_yes VIF = 1.05
  + Probability of Salary and no\_older\_children is > 0.05 as well as VIF of Salary is highest, dropping Salary variable to improve the model and resolve multicollinearity in model
* Model 2 - Drop the variable which has the highest VIF value(Salary variable). Build a second iteration of the model on the full data( Note : Threshold value considered is VIF < 1.15)
  + Summary of statistics:

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | foreign | **No. Observations:** | 872 |
| **Model:** | Logit | **Df Residuals:** | 867 |
| **Method:** | MLE | **Df Model:** | 4 |
| **Date:** | Tue, 17 May 2022 | **Pseudo R-squ.:** | 0.2436 |
| **Time:** | 16:15:27 | **Log-Likelihood:** | -369.23 |
| **converged:** | True | **LL-Null:** | -488.15 |
| **Covariance Type:** | nonrobust | **LLR p-value:** | 2.707e-50 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **Intercept** | 3.8450 | 0.569 | 6.760 | 0.000 | 2.730 | 4.960 |
| **age** | -0.0485 | 0.009 | -5.115 | 0.000 | -0.067 | -0.030 |
| **educ** | -0.4167 | 0.037 | -11.328 | 0.000 | -0.489 | -0.345 |
| **no\_older\_children** | -0.0727 | 0.086 | -0.845 | 0.398 | -0.241 | 0.096 |
| **Holliday\_Package\_yes** | 1.1282 | 0.188 | 6.003 | 0.000 | 0.760 | 1.497 |

* + Check for multicollinearity in the predictor variables using Variance Inflation Factor (VIF):
    - age VIF = 1.05
    - educ VIF = 1.04
    - no\_older\_children VIF = 1.02
    - Holliday\_Package\_yes VIF = 1.03
  + Probability of no\_older\_children is > 0.05 , dropping no\_older\_children variable to improve the model and resolve multicollinearity in model
* Model 3 - Drop the variable which has the highest p-value>0.05 (no\_older\_children variable) in the second iteration of the model. Build a third iteration of the model
  + Summary of statistics:

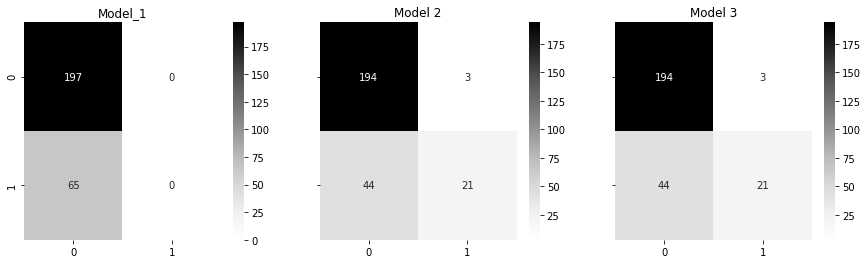
|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | foreign | **No. Observations:** | 872 |
| **Model:** | Logit | **Df Residuals:** | 868 |
| **Method:** | MLE | **Df Model:** | 3 |
| **Date:** | Tue, 17 May 2022 | **Pseudo R-squ.:** | 0.2429 |
| **Time:** | 16:20:50 | **Log-Likelihood:** | -369.59 |
| **converged:** | True | **LL-Null:** | -488.15 |
| **Covariance Type:** | nonrobust | **LLR p-value:** | 3.998e-51 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **Intercept** | 3.7486 | 0.558 | 6.722 | 0.000 | 2.656 | 4.842 |
| **age** | -0.0480 | 0.010 | -5.041 | 0.000 | -0.067 | -0.029 |
| **educ** | -0.4155 | 0.037 | -11.304 | 0.000 | -0.488 | -0.343 |
| **Holliday\_Package\_yes** | 1.1146 | 0.187 | 5.961 | 0.000 | 0.748 | 1.481 |

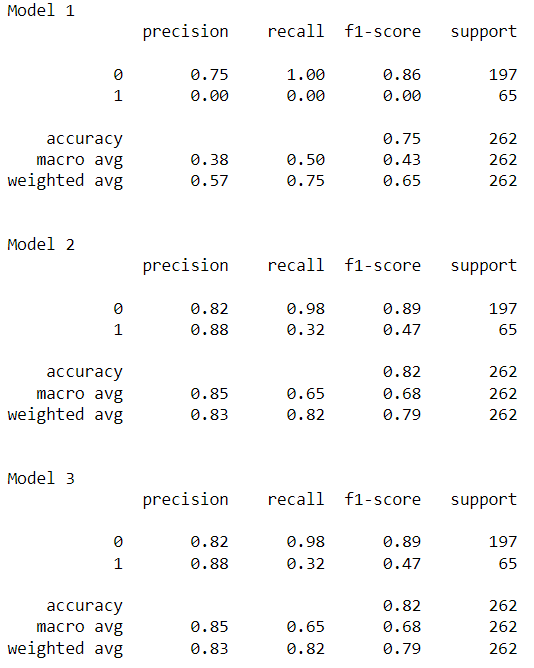
* + Check for multicollinearity in the predictor variables using Variance Inflation Factor (VIF):
    - age VIF = 1.04
    - educ VIF = 1.04
    - Holliday\_Package\_yes VIF = 1.02
  + As all predictors have p-Value < 0.05 there is no multicollinearity preset in model, hence Model 3 is best model as per the VIF and summary of model

## Split the data into training (70%) and test (30%). Build the various iterations of the Logistic Regression models on the training data and use those models to predict on the test data using appropriate model evaluation metrics.

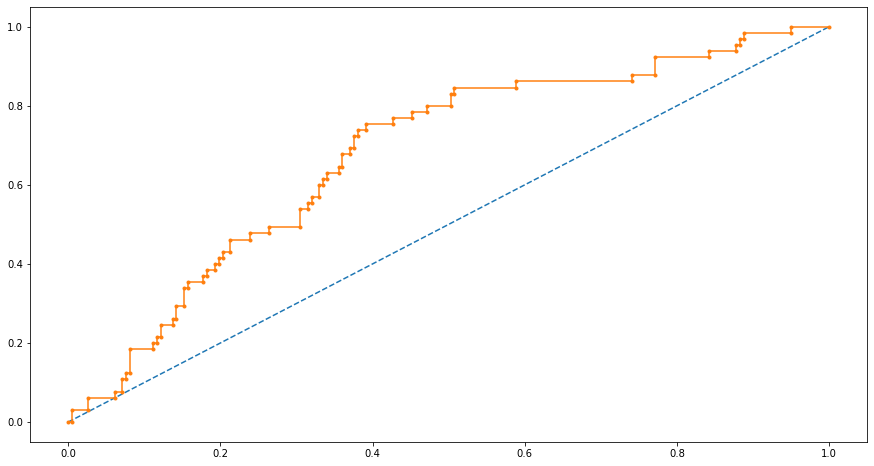
* Split the data into training (70%) and test (30%): with random\_state as 1, there are 610 samples spliced as train data samples and 262samples spliced as test data samples
* Note: sklearn.linear\_model library is used for logistic regression model building
* Model 1 - Building the model on the training data on whole data:
  + Train Accuracy Score : 0.7524590163934426
* Model 2 - Building the model on the training data on data except Salary predictor:
  + Train Accuracy Score : 0.8311475409836065
* Model 3 - Building the model on the training data on data except Salary , no\_older\_children predictor:
  + Train Accuracy Score : 0.8377049180327869
* Confusion Matrix summary statistics Evaluation on the Test Data:



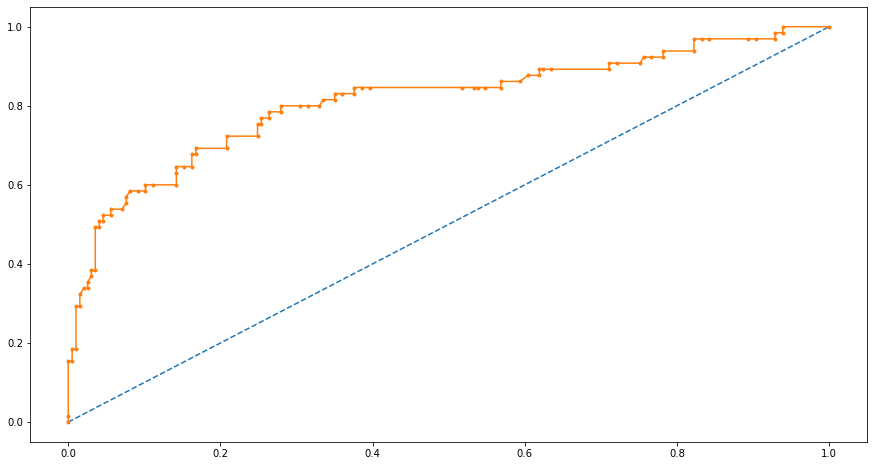
* Classification report on the Test Data:



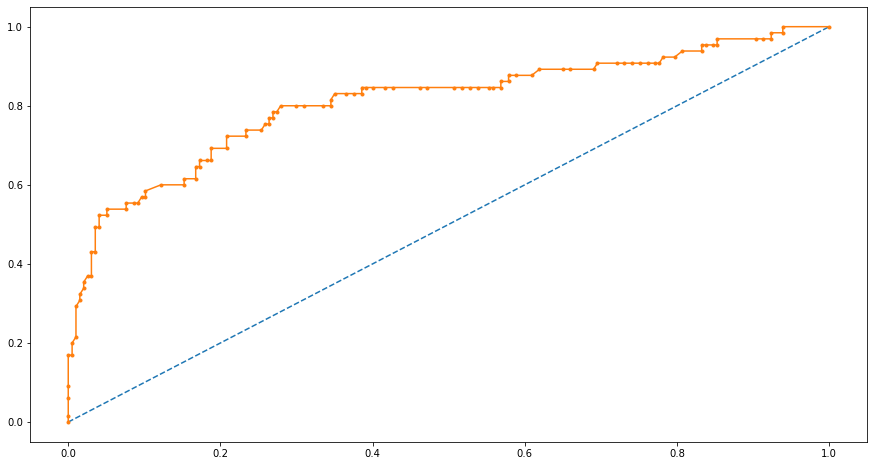
* Test AUC and ROC for the model 1:
  + AUC: 0.67911
  + ROC



* Test AUC and ROC for the model 2:
  + AUC: 0.81281
  + ROC



* Test AUC and ROC for the model 3:
  + AUC: 0.81000
  + ROC



* From Train and test confusion matrix summary, classification report, AUC and ROC curve of there are mild changes in model 2 and model 3. As AUC of model 2 is higher, it is preferable model for logistic regression.

Conclusion: the final model of Part (II) and the proposed one in Part (III) are different. After calculating the accuracy and confusion matrix summary, classification report, AUC and ROC curve of model 3 in Part (III), we can see that the accuracy is dropping in model 3 as compared to model 2. So we should go with model 2 as compared to model 3 having predictor age, educ, no\_older\_children, Holliday\_Package\_yes.