Project on Predictive Modeling

Ravindra Ramesh Tabde
PGP -DSBA –LVC
August 2021
Date: 04-Feb-2022

Table of Contents

CONTENTS

LIST OF FIGURE

LIST OF TABLES

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.

Dataset for Problem 1: cubic_zirconia.csv

Approach of solution:

1. Exploratory Data Analysis for Problem 1
2. Build various iterations of the Linear Regression model using appropriate variable selection techniques for the full data.
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.

Data description for zirconia dataset and data type check:

Sr. No.	Feature Name	Description	Additional info	Data type
1	Carat	weight of the cubic zirconia	Physical weight of diamond	Numeric
2	Cut	Describe the cut quality of the cubic zirconia	Quality is increasing order Fair, Good, Very Good, Premium, Ideal.	Categorical (Ordinal)
3	Colour	Color of the cubic zirconia	D being the best and J the worst.	Categorical (Ordinal)
4	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, Sl1, Sl2, l1	Categorical (Ordinal)
5	Depth	The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter		Numeric
6	Table	The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter		Numeric
7	Price	Price of the cubic zirconia		Numeric
8	X	Length of the cubic zirconia	In mm	Numeric
9	Y	Width of the cubic zirconia	In mm	Numeric
10	Z	Height of the cubic zirconia	In mm	Numeric

Table 1-Description of Zirconia dataset

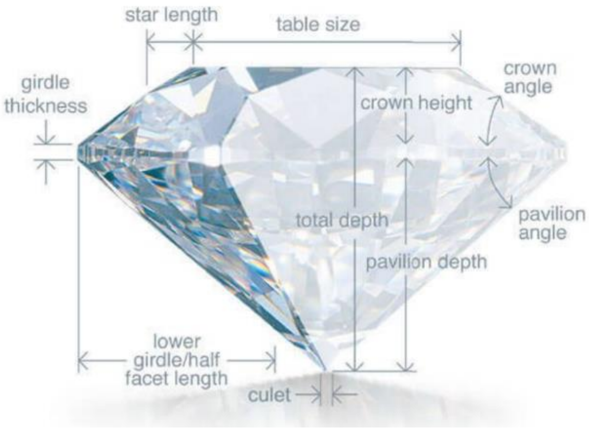


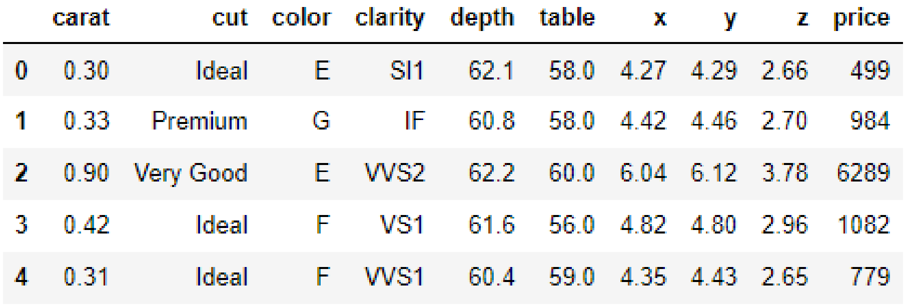
Fig. 1 –Dimensional proportions of Zirconia

Fig. 1- Shows the dimensional proportions of diamond to get an idea of the various dimensions of cubic zirconia.

1.1 Exploratory Data Analysis for Problem 1:

To start with the analysis let’s look at the sample data

Sample of dataset:



	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
2	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
4	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779

Table 2-sample of Zirconia dataset

Check for the types of variables and missing values in the dataset.

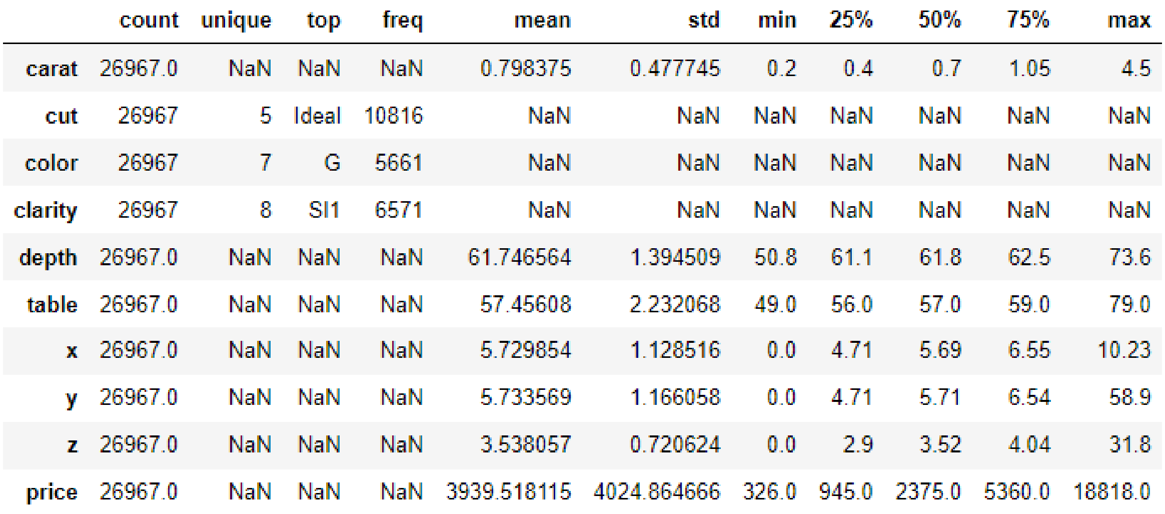
```
Range Index: 26967 entries, 0 to 26966
Data columns (total 10 columns):

No.  Column              Non-Null Count        DType
--------------------     --------------------  ----------
0  Carat                 26967 non-null        float64
1  Cut                   26967 non-null        object
2  Color                 26967 non-null        object
3  Clarity               26967 non-null        object
4  Depth                 26270 non-null        float64
5  Table                 26967 non-null        float64
6  x                     26967 non-null        float64
7  y                     26967 non-null        float64
8  z                     26967 non-null        float64
9  Price                 26967 non-null        int64
```

Observations:

- There are total 26967 rows and 10 columns in the dataset.
- Dependent variable price is of integer data type. Cut, Color and clarity is of object type. All other variable is of float64 datatype.
- We can see 9 independents variable and one target variable – price.
- From above data we can say that there are **missing values (697)** in the dataset for depth feature values in dataset.
- Also checked for duplicate values **there are 33 duplicate values** found in dataset.

Let's check for summary statistics of the dataset for all variable.

Summary Statistics:



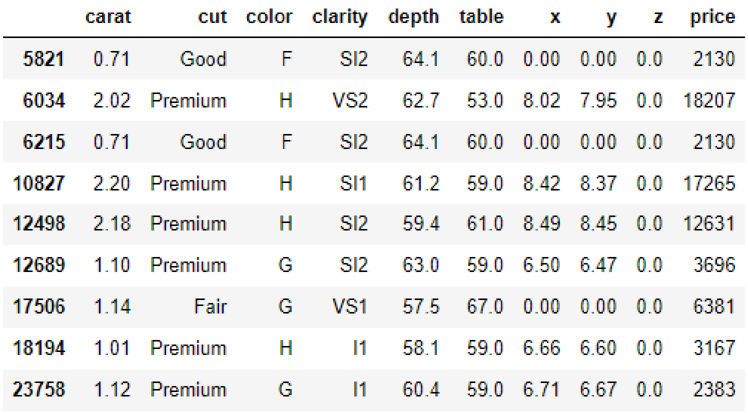
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
carat	26967.0	NaN	NaN	NaN	0.798375	0.477745	0.2	0.4	0.7	1.05	4.5
cut	26967	5	Ideal	10816	NaN	NaN	NaN	NaN	NaN	NaN	NaN
color	26967	7	G	5661	NaN	NaN	NaN	NaN	NaN	NaN	NaN
clarity	26967	8	SI1	6571	NaN	NaN	NaN	NaN	NaN	NaN	NaN
depth	26967.0	NaN	NaN	NaN	61.746564	1.394509	50.8	61.1	61.8	62.5	73.6
table	26967.0	NaN	NaN	NaN	57.45608	2.232068	49.0	56.0	57.0	59.0	79.0
x	26967.0	NaN	NaN	NaN	5.729854	1.128516	0.0	4.71	5.69	6.55	10.23
y	26967.0	NaN	NaN	NaN	5.733569	1.166058	0.0	4.71	5.71	6.54	58.9
z	26967.0	NaN	NaN	NaN	3.538057	0.720624	0.0	2.9	3.52	4.04	31.8
price	26967.0	NaN	NaN	NaN	3939.518115	4024.864666	326.0	945.0	2375.0	5360.0	18818.0

Table3 – Summary Statistics of Zirconia dataset

Observations:

- From the summary statistics we can see infer that, in the categorical data ideal, G and SI1 is the popular features for cut, color and clarity respectively as it has top frequency.
- There are 5 ,7 and 8 unique categories for cut, color and clarity feature respectively.
- Mean and median for depth feature is almost equal beside the fact that it has missing entries.
- Mean and median for table feature is almost equal.
- For price variable mean and median are varies more. It shows data must be skewed.
- Also there are 0 values shows for minimum length, width and height(x,y,z features). It is a invalid entries. In practical sense there will be no 2 dimensional or 1 dimensional diamonds.

Let’s check for data points which has either x or y or z as zero values.



	carat	cut	color	clarity	depth	table	x	y	z	price
5821	0.71	Good	F	SI2	64.1	60.0	0.00	0.00	0.0	2130
6034	2.02	Premium	H	VS2	62.7	53.0	8.02	7.95	0.0	18207
6215	0.71	Good	F	SI2	64.1	60.0	0.00	0.00	0.0	2130
10827	2.20	Premium	H	SI1	61.2	59.0	8.42	8.37	0.0	17265
12498	2.18	Premium	H	SI2	59.4	61.0	8.49	8.45	0.0	12631
12689	1.10	Premium	G	SI2	63.0	59.0	6.50	6.47	0.0	3696
17506	1.14	Fair	G	VS1	57.5	67.0	0.00	0.00	0.0	6381
18194	1.01	Premium	H	I1	58.1	59.0	6.66	6.60	0.0	3167
23758	1.12	Premium	G	I1	60.4	59.0	6.71	6.67	0.0	2383

Table 4 – Record with Zero Values

There are 9 record (ref Table 3) which has zero values. So for further analysis we can removed this records.

Checking for Null Values:

No.	Column	Null Count
0	Carat	0
1	Cut	0
2	Color	0
3	Clarity	0
4	Depth	697
5	Table	0
6	x	0
7	y	0
8	z	0
9	Price	0

From the descriptive statistics summary Table 2 .we have seen that mean and median of the depth feature are almost equal. (Mean=61.7, and median = 61.8)
We can impute null values with the median of depth.

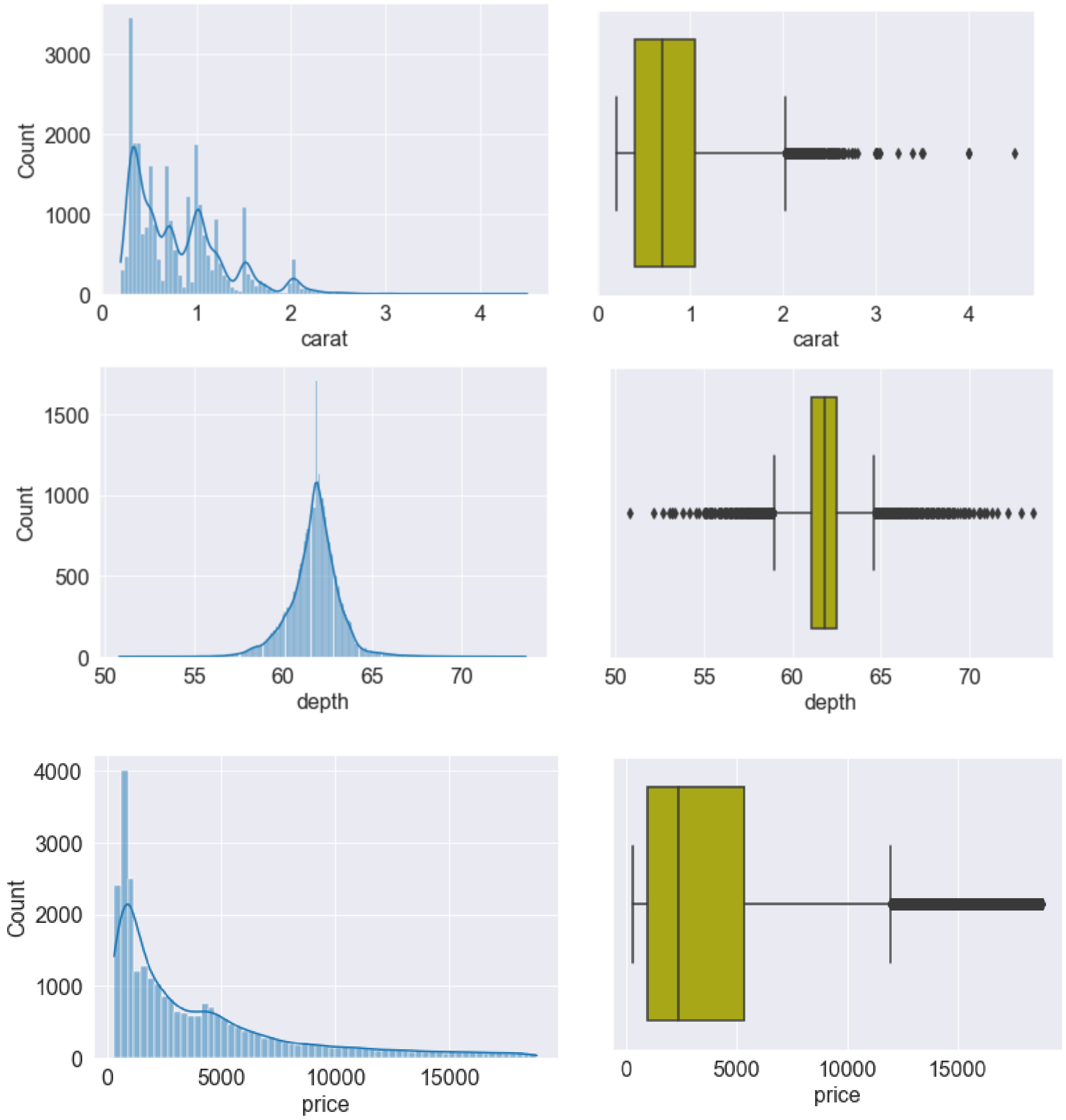
Checking for duplicate records:
There are some duplicate records found in dataset. (33 duplicate rows out of 26958).
Which is nearly 0.12 % of the total data. If the duplicate record are marginal compared with all data in that case we need to consider the duplicate record. As they will have impact on output of the model.
So in this case we have dropped the duplicated row as the percentage of duplicate record is low.

After removal of duplicate records and invalid data entries we have refined dataset having 26925 rows and 10 columns.

1.1.1 UNIVARIATE ANALYSIS

- We can carried out univariate analysis by using histogram and box plot for numerical continuous data and strip plot, count plot , pie plot for categorical data to find information or patterns in the data

Univariate Analysis: Continuous variable



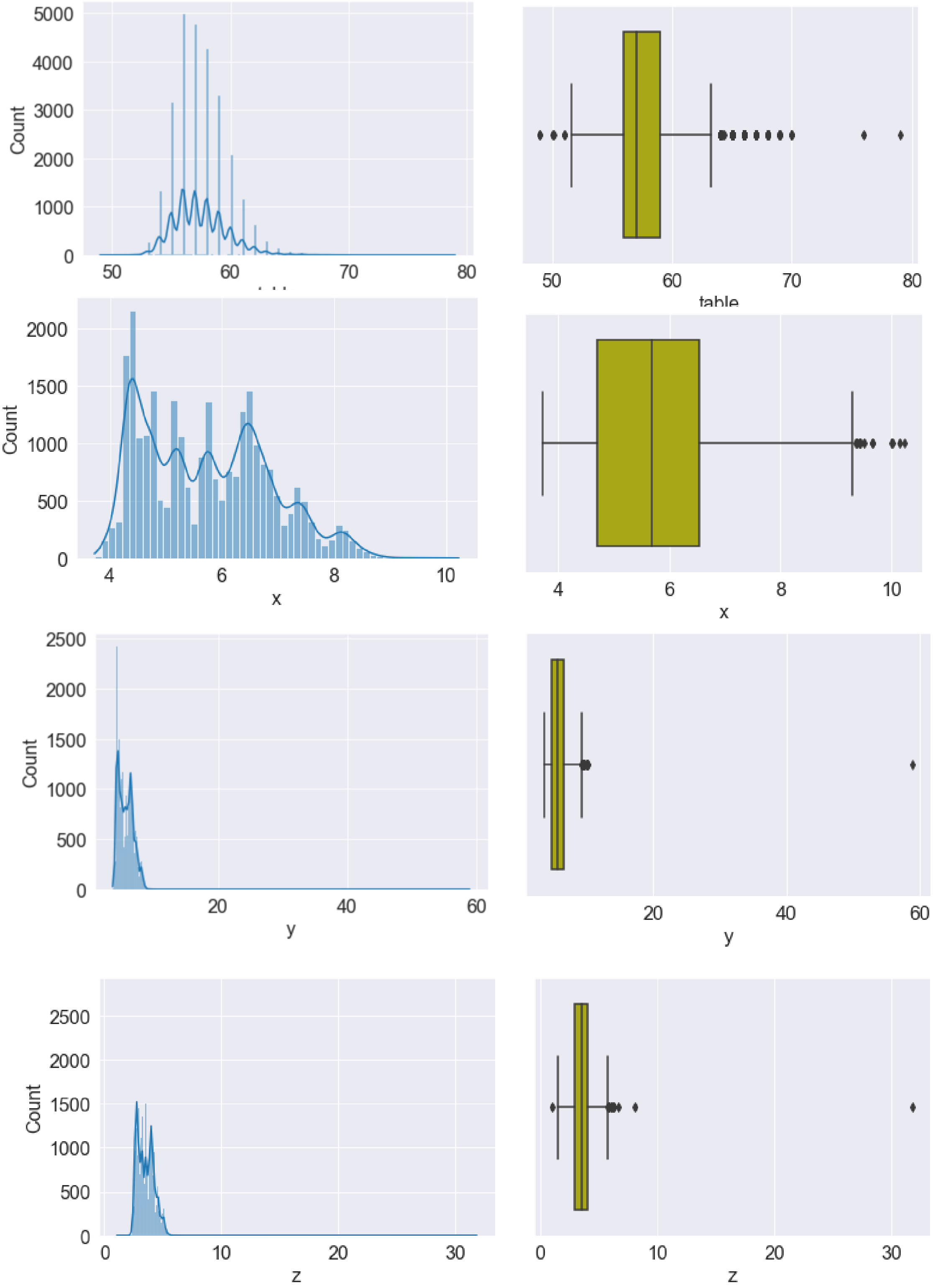


Fig. 2 – Histogram and Box plot for continuous variable Zirconia dataset

Observations:

- There are significant amount of outlier present in some variable, the features with data point that are far from the rest of dataset which will affect the outcome of our regression model. So we have to treat the outlier.

- We can see that the distribution of some quantitative features like 'carat' and the target feature 'price' are heavily right-skewed'.

- Though there are outlier in the dataset for depth variable but it shows more symmetric distribution.

Kurtosis & Skewness in Dataset:

Kurtosis is a measure of whether or not a distribution is heavy-tailed or light-tailed relative to a normal distribution.
Skewness is a measure of the asymmetry of a distribution. This value can be positive or negative.

Features	Skewness	kurtosis
carat	1.114871	1.212235
depth	-0.028403	3.863895
table	0.764890	1.579418
x	0.402010	-0.720965
y	3.888607	160.727513
z	2.639529	88.516471
price	1.619055	2.152993

Table 5 – Kurtosis and Skewness values for zirconia data variables

Observations:

- Data distribution in the variable y and z is right skewed as skewness values are relatively large 3.89 and 2.64 respectively followed by the price variable.
- Depth variable has more stable distribution with nearly 0 skewness.
- Kurtosis for y and z variable is way more than the other variable it shows more outlier's are present.

Univariate Analysis: Categorical variable:

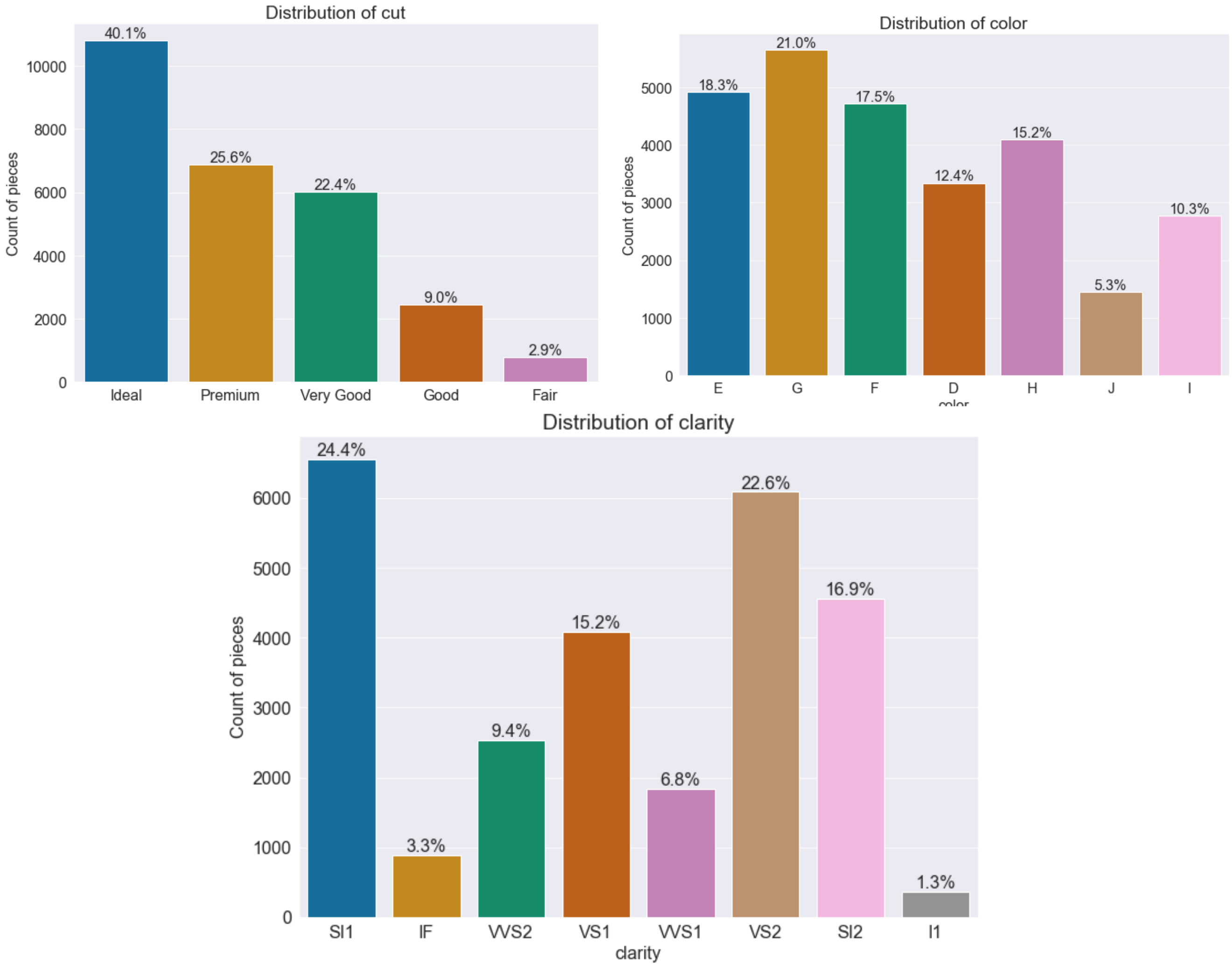


Fig. 3– Count Plot for categorical variables zirconia dataset

Observations:

- There are three categorical variables in data cut, color and clarity.
- Cut variable have 5 subtype among the all, ideal type cut have highest piece count followed by premium, very good, good and Fair.
- Color variable have 7 subtype. ‘G’ type color has highest piece count followed by
  E, F, H, D, I and J.
- Clarity variable have 8 subtype. ‘SI1’ type has highest piece count followed by VS2, SI2, VS1, VVS2, VVS1, IF and I1.

1.1.2 BIVARIATE ANALYSIS

We can carried out bivariate analysis by using box plot, violin plot, swarm plot and strip plot. Here we have used box plot and violin plot for continuous variable and categorical variable analysis.

Also for target variable vs continuous independent variable we have used regression plot.

Target variable – Independent continuous variable

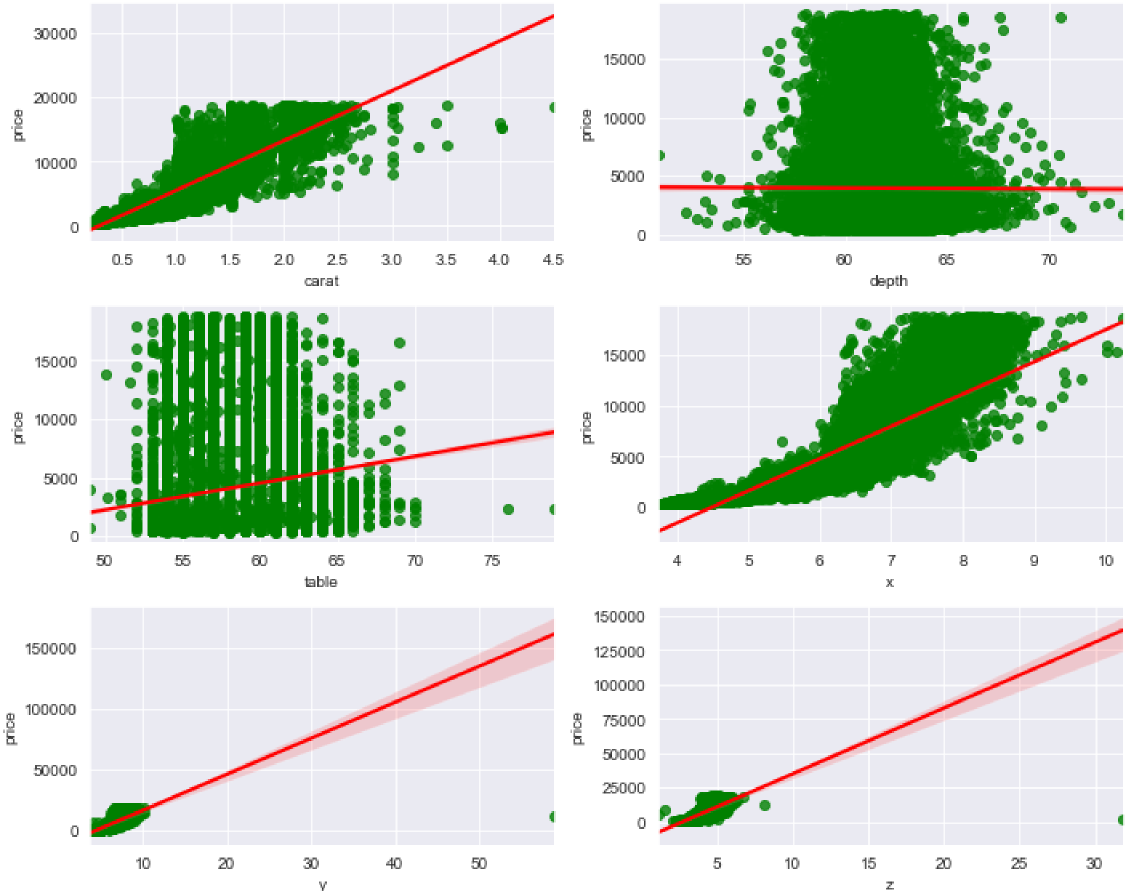


Fig. 4 – Bivariate analysis for target vs. independent continuous variables

Checking of correlation coefficient (r) and P value with the target variable

Features	r	p
carat	0.922400	0
depth	-0.002683	0.659802
table	0.126967	3.7026e-97
x	0.887467	0
y	0.857255	0
z	0.855775	0

Table 6 – r and p value for independent continuous variables

Coefficient of correlation r:

- r is always a number between -1 and 1.
- r > 0 indicates a positive association.
- r < 0 indicates a negative association.
- Values of r near 0 indicate a very weak linear relationship.
- The strength of the linear relationship increases as r moves away from 0 toward -1 or 1
- The extreme values r = -1 and r = 1 occur only in the case of a perfect linear relationship.

Significance level p value:

For the $t-statistic$ for every co-efficient of the Linear Regression the null and alternate Hypothesis is as follows:

- H0 : The variable is significant.¶
- H1: The variable is not significant.

Lower the p-value for the t-statistic more significant are the variables.

Observations:

We can infer from fig.5 and table.5

- Carat vs. Price: Pearson's coefficient of correlation is 0.92 it shows there is high positive correlation also p value is 0 it show significance of variable is much higher almost equal to 100%.
- Depth vs. Price: Coefficient of correlation is -0.0026 it show there is almost no correlation between variables.
- Table Vs. Price: Coefficient of correlation is 0.1269 it show there very weak correlation between variables with p value almost equal to 0.
- X,Y, Z vs Price: Coefficient of correlation is higher in positive side with r > 0.85 it shows X,Y and Z variable are positively correlated with the price.
- From the regression plot we can see there are few outlier values in the the Y and Z variable.

Target variable – Independent categorical variable

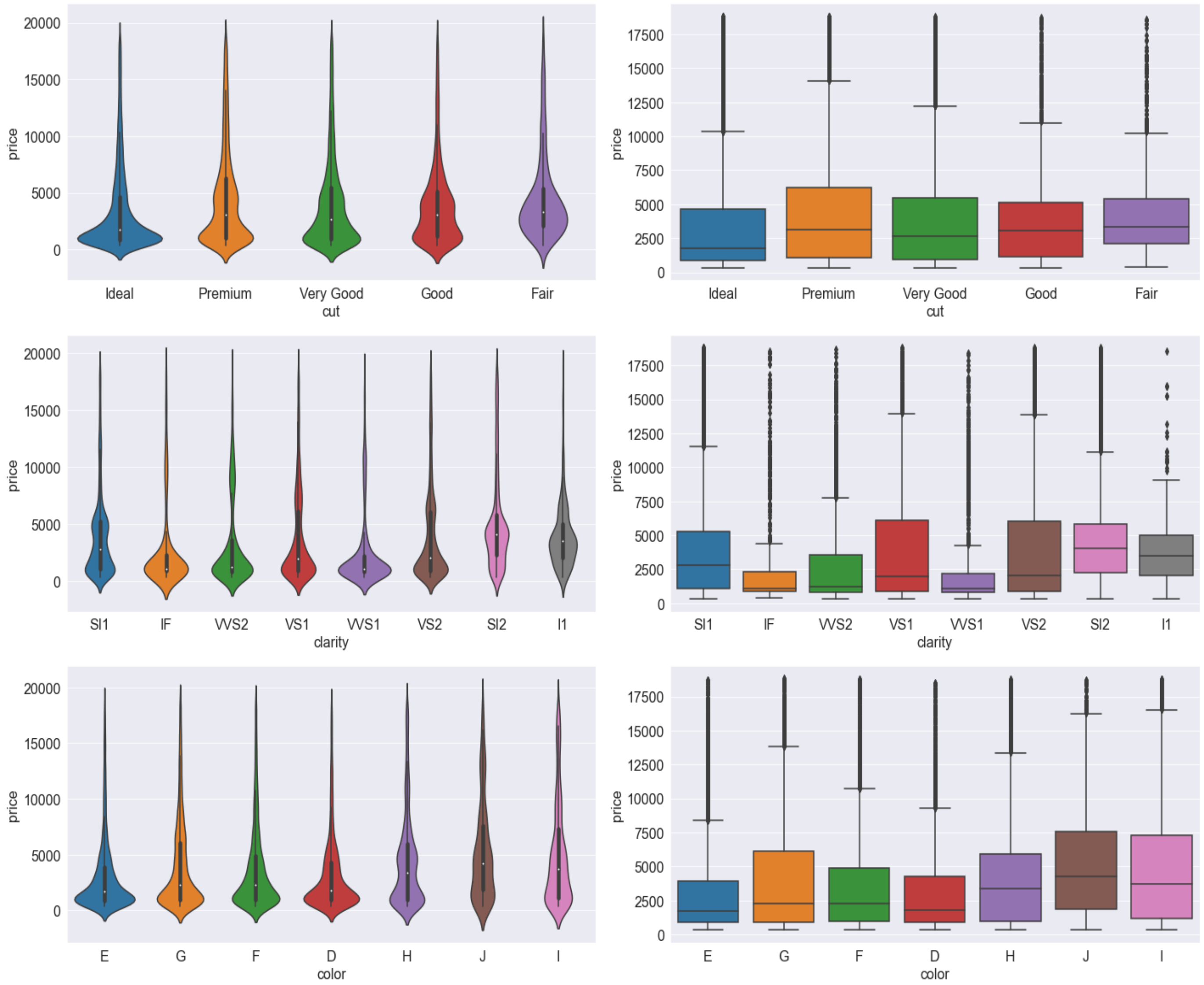


Fig. 5 – Bivariate analysis for target vs. independent categorical variables

Observations:

- For Cut Vs Price: The both box plot and violin plot shows the price distribution among the different categories of cut. Median for different categories within the same range. Ideal and fair has difference in the price distribution compare to other category we can see from violin plot.
- For Clarity Vs. Price: VS1 and VS2 has same price distribution and median price. As like SI1 and SI2. Clarity will be the significant predictor for price.
- For Color vs Price: The box plot and Violin plot shows ranges of median prices among the colour category but there is small differences in the median prices.

MULTIVARIATE ANALYSIS:

The pair plot helps us to understand the relationship between all the numerical values in the dataset. On comparing all the variables with each other we could understand the patterns or trends in the dataset.

Checking for correlation of numerical variable by taking cut as hue variable.

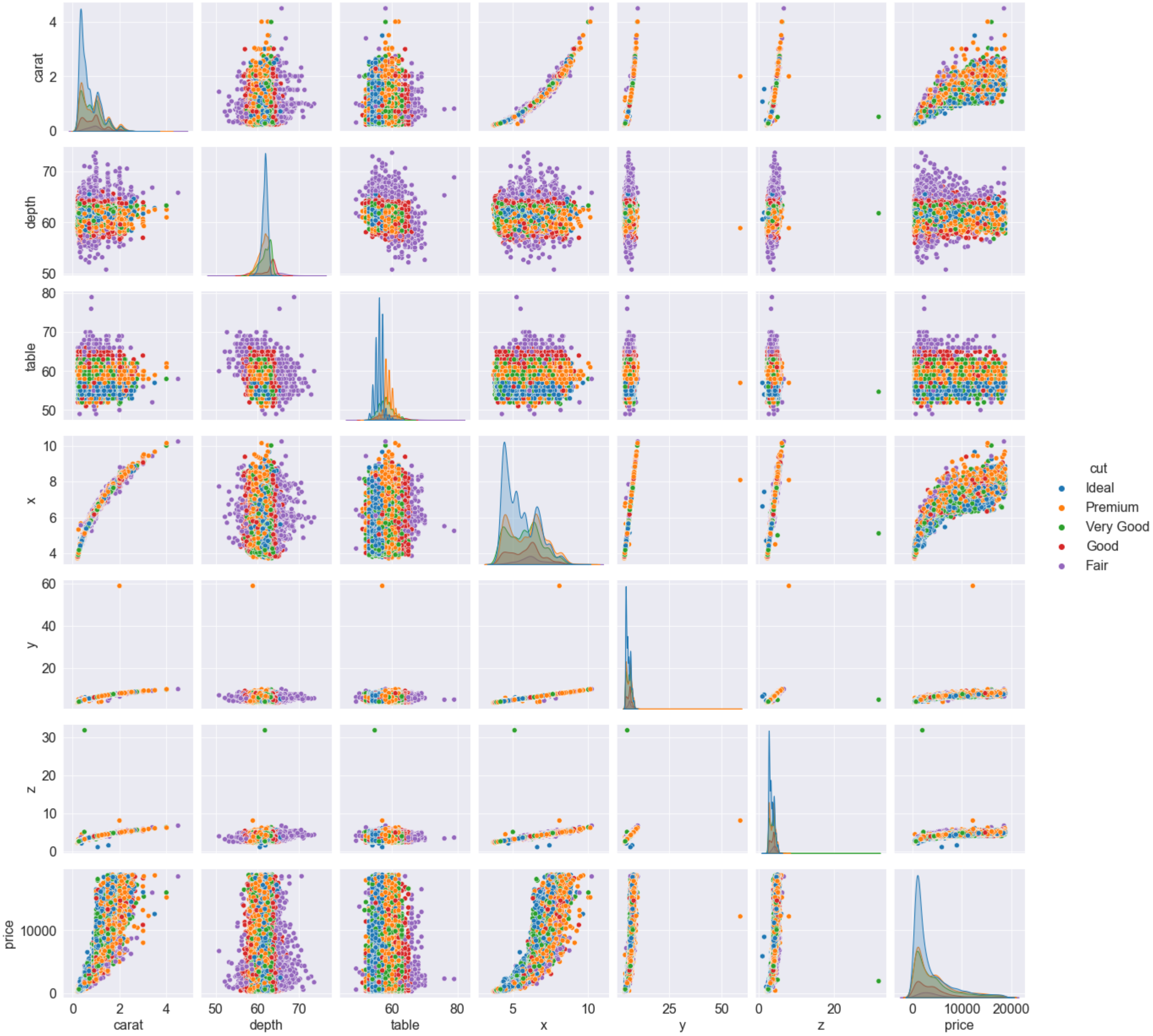


Fig.6 – Pair plot of numeric variable of Zirconia dataset

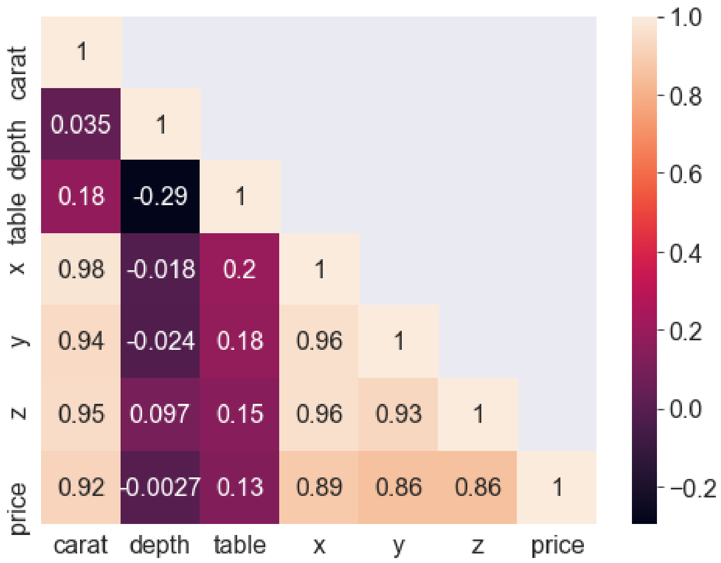


Fig. 7 – Heat map of correlation Matrix

Observation:

- From the pair plot and the heat map of correlation matrix we can see that variable X, Y, Z and carat has high correlation with target variable.
- There is no correlation between depth and price also very weak correlation between table and price variable.
- From pair plot we can see there are outlier in the y and z variable.
- We can see from pair plot there is multicolinearity found between the carat and X, Y, Z variable.

Action before building model:

- ➢ There are significant amount of outlier present in some variable, the features with data point that are far from the rest of dataset which will affect the outcome of our regression model. So we have to treat the outlier.

- ➢ There are some subcategories in the categorical variable having similar price distribution. From fig.6 inferences we can club those categories.

- ➢ For Cut variable we can club ideal and premium under to ideal premium. Also for good and very good to good_very_good by putting fair as separate category.

- ➢ Same we can club for clarity VS1 and VS2 to VS1_2 also for SI1 and SI2 to SI1_2.

- ➢ For color category D and E to D_E. also for J and I to J_I.

- ➢ For X,Y and Z variable we can combine this variable to Volume(x*y*z) as there is multicolinearity between these variable.

By taking the above action on data we can proceed for the model building.

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

As discuss in the EDA session there are 3 categorical variable in the dataset cut, color and clarity which are ordinal in nature. So we have to encode it before building the model sue to ordinal nature of variable we have to go for labelled encoding so that we can maintain the order importance of the variable.

Here **price is the dependent variable** and other are predicting variable or independent variable.
Here stats model library used with OLS (ordinary least squares) method for model building.
We have used here descriptive approach for model building.

Approach to Build Models

- Build model with all variable.
- Observe the R square and Adjusted R square value.
- Check for VIF for multicolinearity
- Remove column with multicolinearity
- If no multicolinearity found check for P values
- Remove feature with high p values if any

Model 1: Using all variables

The coefficient and intercept for model 1 as follow:

```
Intercept for model is - 5416.299789
Coefficient for carat is 8801.427231
Coefficient for cut is  230.355585
Coefficient for color is 302.237938
Coefficient for clarity is 586.088414
Coefficient for depth is 44.223963
Coefficient for table is -26.461954
Coefficient for x is -1518.395973
Coefficient for y is 1858.484905
Coefficient for z is -1214.563675
```

You can put this values in the linear regression equation for making the value base Prediction.

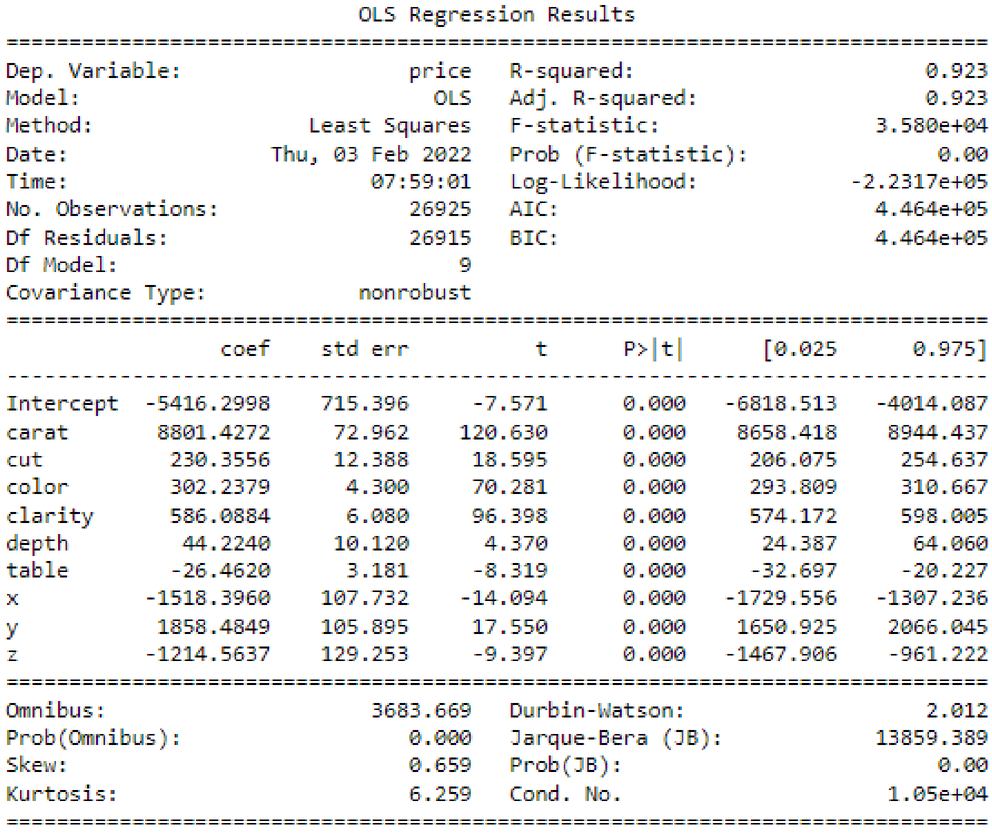
Check for VIF values for Model 1:

```
carat  VIF =  33.01
cut  VIF =  1.3
color  VIF =  1.11
clarity  VIF =  1.2
depth  VIF =  4.42
table  VIF =  1.37
x  VIF =  427.15
y  VIF =  406.86
z  VIF =  234.91
```

We can see VIF score for the variable used in this model found to be way higher than the cut off range taken 5 (4 variables). It shows there is a problem of Multicolinearity among the predictor variables.

To overcome the problem of multicolinearity we can further analyse for PCA technique how it will help us in the dimension reduction.

Let’s check for summary of model



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.923
Model:                            OLS   Adj. R-squared:                  0.923
Method:                 Least Squares   F-statistic:                 3.580e+04
Date:                Thu, 03 Feb 2022   Prob (F-statistic):               0.00
Time:                        07:59:01   Log-Likelihood:            -2.2317e+05
No. Observations:               26925   AIC:                         4.464e+05
Df Residuals:                   26915   BIC:                         4.464e+05
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -5416.2998    715.396     -7.571      0.000   -6818.513   -4014.087
carat       8801.4272     72.962    120.630      0.000    8658.418    8944.437
cut          230.3556     12.388     18.595      0.000     206.075     254.637
color        302.2379      4.300     70.281      0.000     293.809     310.667
clarity      586.0884      6.080     96.398      0.000     574.172     598.005
depth         44.2240     10.120      4.370      0.000      24.387      64.060
table        -26.4620      3.181     -8.319      0.000     -32.697     -20.227
x          -1518.3960    107.732    -14.094      0.000   -1729.556   -1307.236
y           1858.4849    105.895     17.550      0.000    1650.925    2066.045
z          -1214.5637    129.253     -9.397      0.000   -1467.906    -961.222
==============================================================================
Omnibus:                     3683.669   Durbin-Watson:                   2.012
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            13859.389
Skew:                           0.659   Prob(JB):                         0.00
Kurtosis:                       6.259   Cond. No.                     1.05e+04
==============================================================================
```

Observations:

- From the above summary of model we can infer that all variables have statistical significant as no p value is greater than the 0.05.
- Adjusted R square value seems to be good
- For variables in model_1 VIF score for 4 variables that is X, Y, Z and Carat is higher than the threshold 5. It shows problem of multicolinearity.
- This model explains 92% of response variable variation.

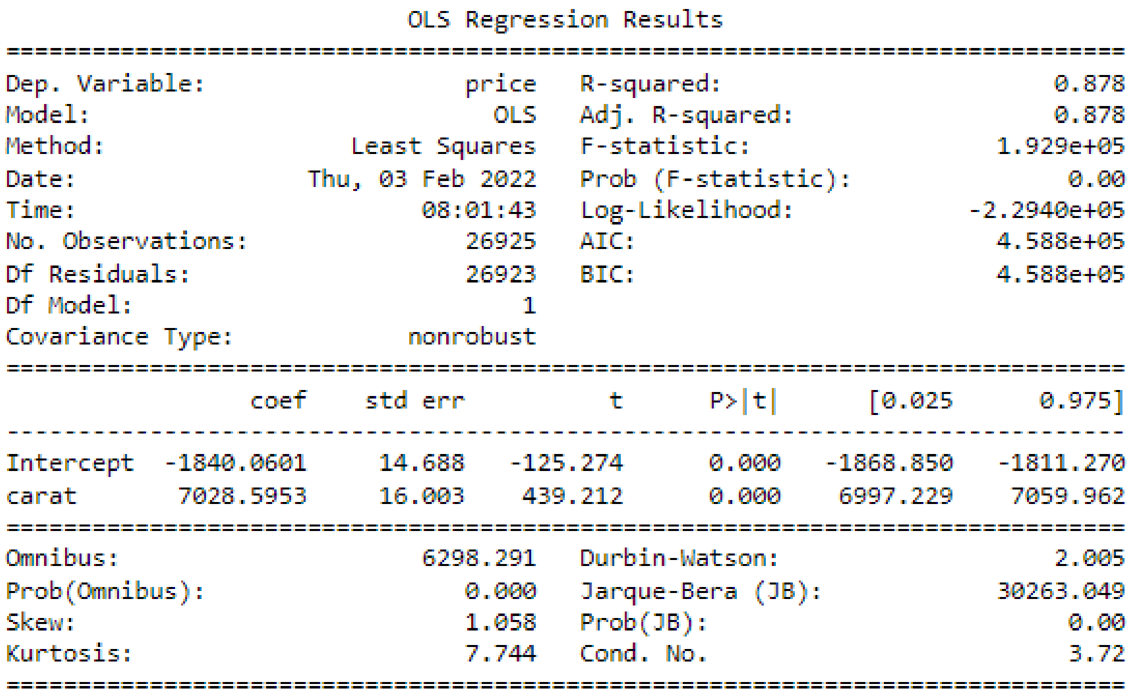
Model 2: Using carat as predictor for price

As we have checked in EDA there is a highest correlation for carat with price.
Using the single predictor variable we have built the model

The coefficient and intercept for model 2 as follow:

Intercept for model is – 1840.060070
Coefficient for carat is 7028.595315

Let's check for summary of model:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.878
Model:                            OLS   Adj. R-squared:                  0.878
Method:                 Least Squares   F-statistic:                 1.929e+05
Date:                Thu, 03 Feb 2022   Prob (F-statistic):               0.00
Time:                        08:01:43   Log-Likelihood:            -2.2940e+05
No. Observations:               26925   AIC:                         4.588e+05
Df Residuals:                   26923   BIC:                         4.588e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -1840.0601     14.688   -125.274      0.000   -1868.850   -1811.270
carat       7028.5953     16.003    439.212      0.000    6997.229    7059.962
==============================================================================
Omnibus:                     6298.291   Durbin-Watson:                   2.005
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            30263.049
Skew:                           1.058   Prob(JB):                         0.00
Kurtosis:                       7.744   Cond. No.                         3.72
==============================================================================
```

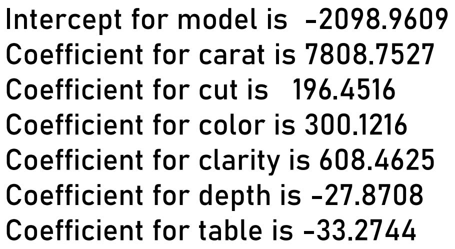
Observations:

- From the above summary of model we can infer that carat variables have more statistical significant as p value is less than 0.05.
- Adjusted R square value marginally decreases compare to model 1.
- As we see in the EDA carat are positively correlated with the price and coefficient value is also positive
- This model explains almost 88% of response variable variation.

Model 3: Using carat, cut, color, clarity, depth and table as predictor for price

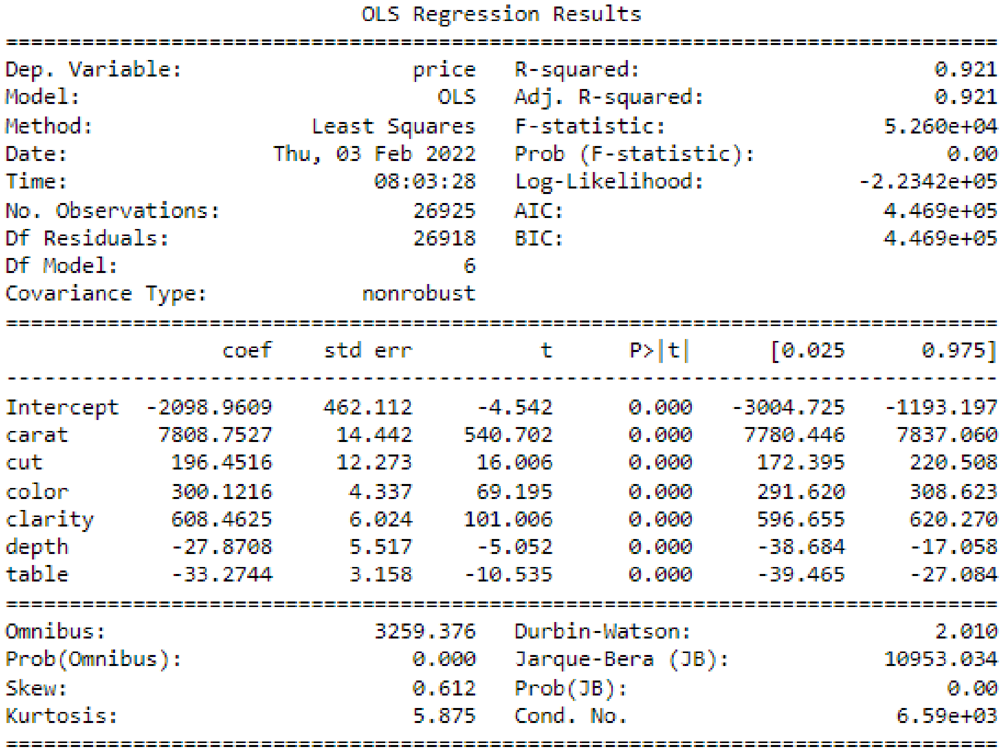
As the X, Y and Z variable have high collinearity with the carat variable we can build the model excluding the X, Y and Z variable.

The coefficient and intercept for model 3 as follow:



```
Intercept for model is  -2098.9609
Coefficient for carat is 7808.7527
Coefficient for cut is   196.4516
Coefficient for color is 300.1216
Coefficient for clarity is 608.4625
Coefficient for depth is -27.8708
Coefficient for table is -33.2744
```

Let's check for summary of model:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.921
Model:                            OLS   Adj. R-squared:                  0.921
Method:                 Least Squares   F-statistic:                 5.260e+04
Date:                Thu, 03 Feb 2022   Prob (F-statistic):               0.00
Time:                        08:03:28   Log-Likelihood:            -2.2342e+05
No. Observations:               26925   AIC:                         4.469e+05
Df Residuals:                   26918   BIC:                         4.469e+05
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -2098.9609    462.112     -4.542      0.000   -3004.725   -1193.197
carat       7808.7527     14.442    540.702      0.000    7780.446    7837.060
cut          196.4516     12.273     16.006      0.000     172.395     220.508
color        300.1216      4.337     69.195      0.000     291.620     308.623
clarity      608.4625      6.024    101.006      0.000     596.655     620.270
depth        -27.8708      5.517     -5.052      0.000     -38.684     -17.058
table        -33.2744      3.158    -10.535      0.000     -39.465     -27.084
==============================================================================
Omnibus:                     3259.376   Durbin-Watson:                   2.010
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            10953.034
Skew:                           0.612   Prob(JB):                         0.00
Kurtosis:                       5.875   Cond. No.                     6.59e+03
==============================================================================
```

Observations:

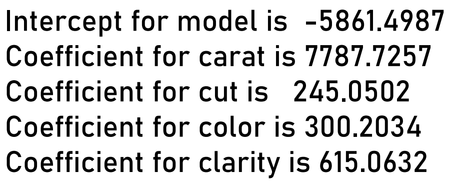
- From the above summary of model3 we can infer that all variables have statistical significant as p value is less than 0.05.
- Adjusted R square value almost similar to the model1
- Coefficient value are positive for carat, cut, color and clarity variable.
- Coefficient values are negative for depth and table variable.
- This model explains almost 92% of response variable variation.
- Also VIF values of variable for this model are less than 5 (VIF cut off value assumed 5). Which shows model is free of multicolinearity.

Model 4: Using carat, cut, color, clarity predictor for price

As the X, Y and Z variable have high collinearity with the carat variable we can build the model excluding the X, Y and Z variable.

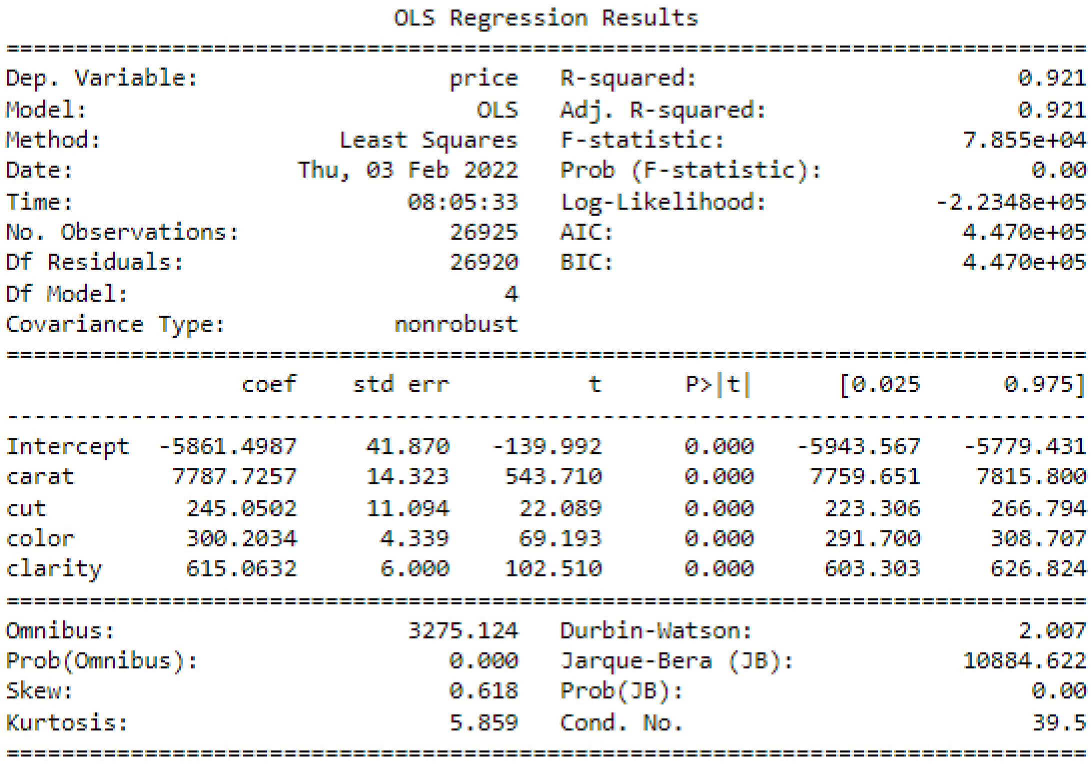
Also excluding depth and table variable as it shows weak relationship with price.

The coefficient and intercept for model 4 as follow:



```
Intercept for model is  -5861.4987
Coefficient for carat is 7787.7257
Coefficient for cut is  245.0502
Coefficient for color is 300.2034
Coefficient for clarity is 615.0632
```

Let's check for summary of model:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.921
Model:                            OLS   Adj. R-squared:                  0.921
Method:                 Least Squares   F-statistic:                 7.855e+04
Date:                Thu, 03 Feb 2022   Prob (F-statistic):               0.00
Time:                        08:05:33   Log-Likelihood:            -2.2348e+05
No. Observations:               26925   AIC:                         4.470e+05
Df Residuals:                   26920   BIC:                         4.470e+05
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -5861.4987     41.870   -139.992      0.000   -5943.567   -5779.431
carat       7787.7257     14.323    543.710      0.000    7759.651    7815.800
cut          245.0502     11.094     22.089      0.000     223.306     266.794
color        300.2034      4.339     69.193      0.000     291.700     308.707
clarity      615.0632      6.000    102.510      0.000     603.303     626.824
==============================================================================
Omnibus:                     3275.124   Durbin-Watson:                   2.007
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            10884.622
Skew:                           0.618   Prob(JB):                         0.00
Kurtosis:                       5.859   Cond. No.                         39.5
==============================================================================
```

Observations:

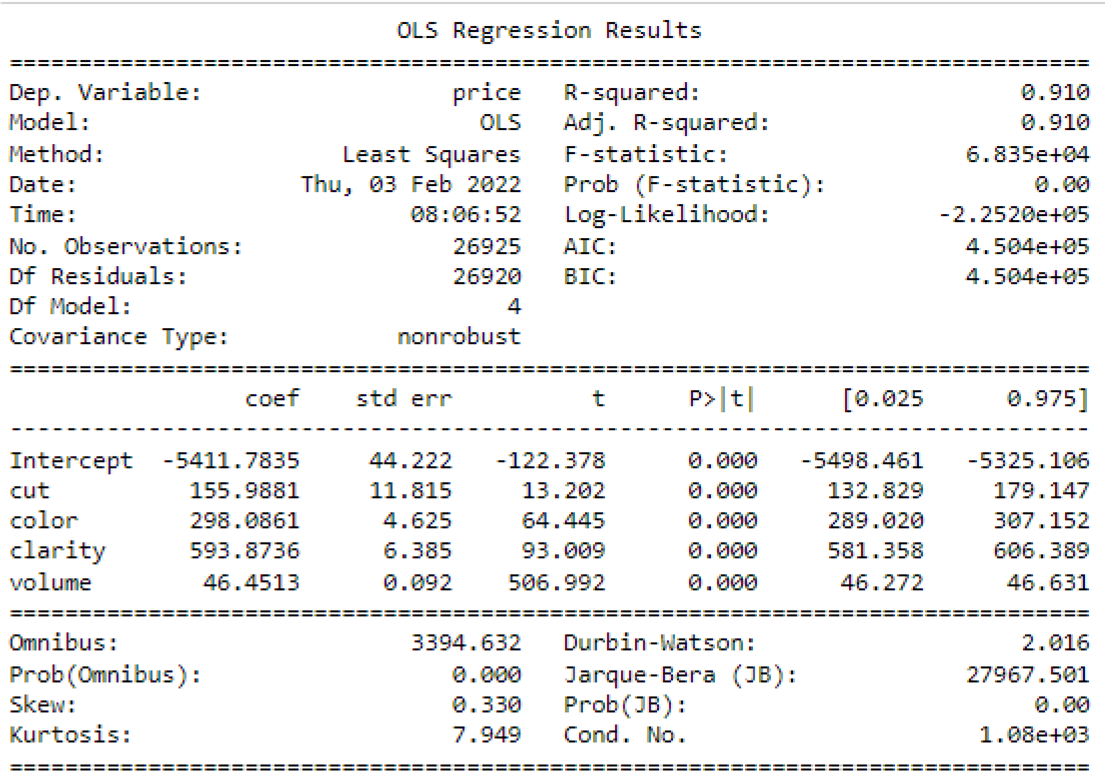
- From the above summary of model_4 we can infer that all variables have statistical significant as p value is less than 0.05.
- No change in the Adjusted R square value almost similar to the model1 and model3
- Coefficient value are positive for carat, cut, color and clarity variable.
- This model explains almost 92% of response variable variation.
- Also VIF values of variable for this model are less than 5(VIF cut off value assumed 5). Which shows model is free of multicolinearity.

<u>Model 5: Using cut, color, clarity and volume variable as predictor for price</u>
As the X, Y and Z variable have high collinearity with the carat variable we can replace carat with the derive feature volume (x*y*z)
Also excluding depth and table variable as it shows weak relationship with price.

The coefficient and intercept for model 5 as follow:

```
Intercept   -5411.783535
cut           155.988095
color         298.086087
clarity       593.873630
volume         46.451298
```

Let's check for summary of model:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.910
Model:                            OLS   Adj. R-squared:                  0.910
Method:                 Least Squares   F-statistic:                 6.835e+04
Date:                Thu, 03 Feb 2022   Prob (F-statistic):               0.00
Time:                        08:06:52   Log-Likelihood:            -2.2520e+05
No. Observations:               26925   AIC:                         4.504e+05
Df Residuals:                   26920   BIC:                         4.504e+05
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -5411.7835     44.222   -122.378      0.000   -5498.461   -5325.106
cut          155.9881     11.815     13.202      0.000     132.829     179.147
color        298.0861      4.625     64.445      0.000     289.020     307.152
clarity      593.8736      6.385     93.009      0.000     581.358     606.389
volume        46.4513      0.092    506.992      0.000      46.272      46.631
==============================================================================
Omnibus:                     3394.632   Durbin-Watson:                   2.016
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            27967.501
Skew:                           0.330   Prob(JB):                         0.00
Kurtosis:                       7.949   Cond. No.                     1.08e+03
==============================================================================
```

Observations:

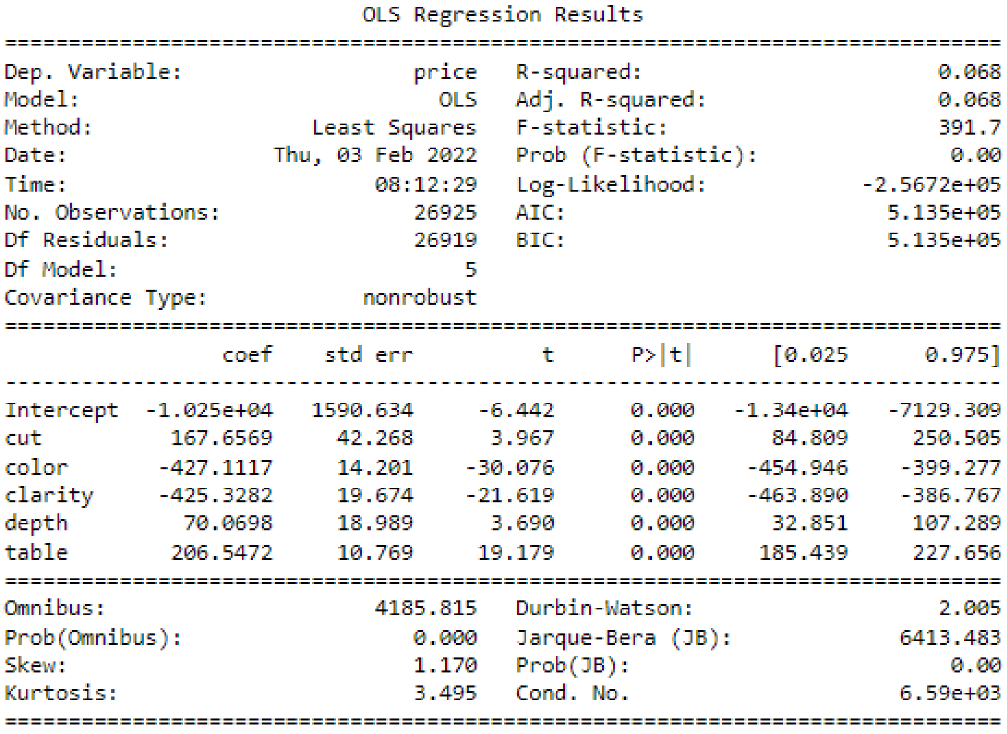
- From the above summary of model_5 we can infer that all variables have statistical significant as p value is less than 0.05.
- Adjusted R square value decreases as we drop carat variable
- Coefficient value are positive volume as like carat, cut, color and clarity variable.
- This model explains almost 91% of response variable variation.
- Also VIF values of variable for this model are less than 5(VIF cut off value assumed 5). Which shows model is free of multicolinearity.

Model 6: Using low VIF values (cut, color, clarity, depth and table variable)

We have checked for VIF values of all variable used volume as derived feature from (x*y*z)
VIF values for all variable are as follows:

carat VIF = 107.99
cut VIF = 1.25
color VIF = 1.11
clarity VIF = 1.16
depth VIF = 1.33
table VIF = 1.35
volume VIF = 107.13

From this above VIF values we have drop volume and carat variable as it has hig her VIF values.
Let's build the model and check for summary of model:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.068
Model:                            OLS   Adj. R-squared:                  0.068
Method:                 Least Squares   F-statistic:                     391.7
Date:                Thu, 03 Feb 2022   Prob (F-statistic):               0.00
Time:                        08:12:29   Log-Likelihood:            -2.5672e+05
No. Observations:               26925   AIC:                         5.135e+05
Df Residuals:                   26919   BIC:                         5.135e+05
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -1.025e+04   1590.634     -6.442      0.000   -1.34e+04   -7129.309
cut          167.6569     42.268      3.967      0.000      84.809     250.505
color       -427.1117     14.201    -30.076      0.000    -454.946    -399.277
clarity     -425.3282     19.674    -21.619      0.000    -463.890    -386.767
depth         70.0698     18.989      3.690      0.000      32.851     107.289
table        206.5472     10.769     19.179      0.000     185.439     227.656
==============================================================================
Omnibus:                     4185.815   Durbin-Watson:                   2.005
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6413.483
Skew:                           1.170   Prob(JB):                         0.00
Kurtosis:                       3.495   Cond. No.                     6.59e+03
==============================================================================
```

Observations:

- From the above summary of model_5 we can infer that all variables have statistical significant as p value is less than 0.05.
- Adjusted R square value decreases drastically as we drop carat and volume variable
- Coefficient value are positive for cut, depth and table variable.
- Coefficient value for color and clarity becomes negative
- This model explains only 7% of response variable variation.

1.2.1 Model Performance Comparison:

Out of the 6 model which we have built to predict price we can compare it by using the Adjusted R square values in the feature selection descriptive method.

Model	Predictors	Adjusted R^2
Model 1	All variable	0.9228771
Model 2	only carat	0.8775242
Model 3	carat,cut,color,clarity,depth,table	0.9213913
Model 4	carat,cut,color,clarity	0.9210704
Model 5	cut,color,clarity,volume	0.9103475
Model 6	cut,color,clarity,depth,table	0.067648

Table 7 – Model Performance comparison in descriptive approach

1.2.2 Model Selection for prediction:

- Model_4 and Model_5 shows highest Adj. R-squared values than the other models with less number of predictors variables
- Also Model_4 and Model_5 are free of multicolinearity as we have seen in model building.
- For prescriptive analysis model 4 is suitable with only 4 independent variable.
- For selecting one model out of these two need more data.

```
RMSE of Model_4: 973.7741948477665
RMSE of Model_5: 1037.8180349238996
```

After calculating the RMSE we have found RMSE for model 4 is lower than the model 5

Model 4 has lower RMSE value so model 4 would be the better fit model for prediction. Keeping the fact in mind that the best descriptive model might not be the best predictive model.

We can explore further with the predictive approach.

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.

- If we only wanted to predict using Linear Regression and were not looking for the model building aspect of it, we can do that as well.
- For this exercise, we will use the same variables as if we are used in the previous models.

Key Differences in Predictive Modelling.

- We will split the data into train and test and get an idea about the expected quality of predictions in future
- We will need to choose a metric of interest. Let's choose RMSE.
- Build the model on the training data and check the RMSE on the test data.

Note: We are going to build all the models, get their predictions and then go on to evaluate those models
We have use Sklearn library for linear regression

We will be doing a **70:30** split. 70% of the whole data will be used to train the data and then 30% of the data will be used for testing the model.

Dimension of train & test data

	rows	columns
X_train	18847	10
X_test	8078	10
Y_train	18847	
Y_test	8078	

Table 8- Dimension of train and test data

We can build the model as given below like in previous approach. This time we can make model on training data and predict it against the test data.

Model 1: Using all variables
Model 2: Using carat as predictor for price
Model 3: Using carat, cut, color, clarity, depth and table as predictor for price
Model 4: Using carat, cut, color, clarity predictor for price
Model 5: Using cut, color, clarity and volume variable as predictor for price
Model 6: Using low VIF values (cut, color, clarity, depth and table variable)

We have build the above six model on the training data and checked against the test data. We have used here the RMSE as a matrix of interest.

Let’s check out the RMSE on Train and Test data for price prediction for different model:

1.3.1 Model Performance Comparison:

Model	Predictors	R^2 Train	R^2 Test	RMSE Train	RMSE Test	Adjusted R^2
Model 1	carat, cut, color, clarity, depth, table, x, y, z	0.9228	0.9229	960.78	967.38	0.9228771
Model 2	carat	0.8764	0.8799	1215.57	1207.29	0.8775242
Model 3	carat,cut,color,clarity,depth,table	0.9213	0.9217	970.38	975.20	0.9213913
Model 4	carat, cut, color, clarity	0.9209	0.9215	972.80	976.18	0.9210704
Model 5	Cut , color, clarity, Volume(x*y*z)	0.9100	0.9111	1037.23	1039.25	0.9103475
Model 6	cut, color, clarity, depth, table	0.0675	0.6823	3339.48	3364.09	0.0676480

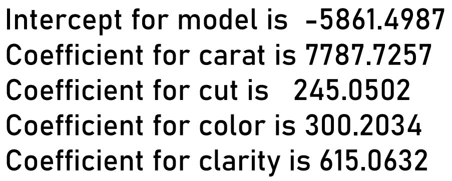
Table 9- Model performance with train test split

1.3.2 Model Selection for prediction:

From the above table model 4 and model 5 look better for prediction. As it has less number of variable and good RMSE on train as well as test data.
But for selecting one out of model 4 and model 5 it would be better to have more data for training and testing.
With this data model 4 looks more stable and balanced however model 5 also not bad choice.

Model 4: Using carat, cut, color, clarity predictor for price

The coefficient and intercept for model 4 as follow:



```
Intercept for model is  -5861.4987
Coefficient for carat is 7787.7257
Coefficient for cut is   245.0502
Coefficient for color is 300.2034
Coefficient for clarity is 615.0632
```

Conclusion:

- We have found the high dependency of price on the carat need to be analyse in detail with the domain expertise.
- Model 4 can be used in descriptive and predictive analysis.
- Using the descriptive approach we have found the model 4 is useful also the predictive approach come with the same result.
- Would suggest to work on more variable and more data to build the more stable model.
- Also from the model 5 we can found that coefficient of volume variable is positive it means price increases as volume of stone increases.
- As we can see from model building carat and volume are correlated among themselves. So volume also play key role in price of zirconia.
- In the EDA of price vs cut we have seen there is not much dependency of price on specific quality of cut. This has to be investigated by the company. Is there same trend in the market or need to take another action on this?
- Last but not lease the declaimer for this model full fledge testing needs to be done before deployment.

The price of a zirconia is often decided by multiple factors. The price determination, therefore seems difficult. An accurate predictive model can be valuable to businesses and consumers to determine the fair price of a diamond.
The goal of the project is to build a model that can accurately predict the price of a diamond potentially based on its weight, quality and dimension measurements.

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.

Dataset for Problem 2**:** Holiday_Package.csv

Approach of solution:

1. Exploratory Data Analysis for Problem 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.
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.

Data description for holiday package dataset and data type check:

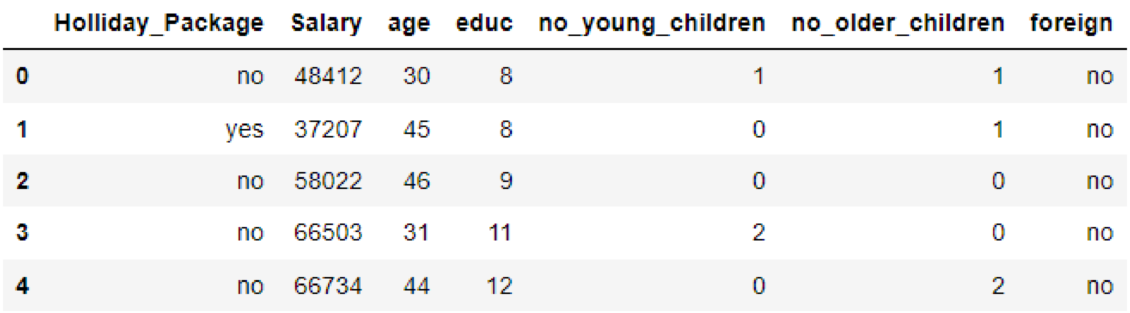
Sr. No.	Feature Name	Description	Data type
1	Holiday Package	Opted for Holiday package yes/no	Categorical
2	salary	Employee Salary	Numeric
3	age	Age in years	Numeric
4	edu	Years of formal education	Numeric
5	no_young_children	The number of young children (younger than 7 years)	Numeric
6	no_older_children	The number of older children	Numeric
7	foreign	Foreigner yes/no	Categorical

Table 10-Description of Holiday package dataset

2.1 Exploratory Data Analysis for Problem 2:

To start with the analysis let's look at the sample data

Sample of dataset:



	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	no	48412	30	8	1	1	no
1	yes	37207	45	8	0	1	no
2	no	58022	46	9	0	0	no
3	no	66503	31	11	2	0	no
4	no	66734	44	12	0	2	no

Table 11-sample of Holiday package dataset

Check for the types of variables and missing values in the dataset.

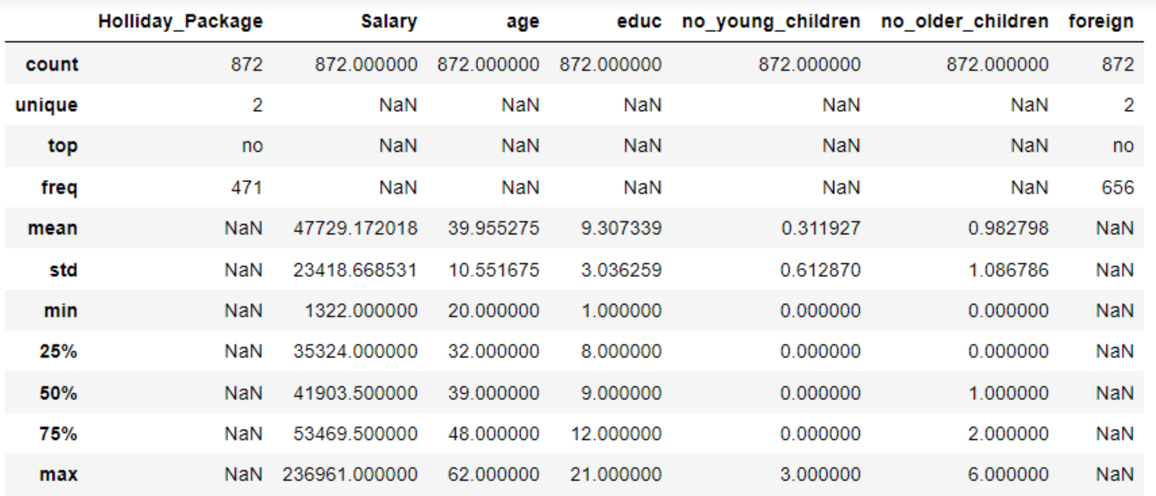
```
Range Index: 872 entries, 0 to 871
Data columns (total 7 columns):

No.  Column                Non-Null Count      DType
--------------------------  --------------------  ----------
0   Holiday_package         872 non-null         object
1   Salary                  872 non-null         int64
2   age                     872 non-null         int64
3   educ                    872 non-null         int64
4   no_young_children       872 non-null         int64
5   no_older_children       872 non-null         int64
6   foreign                 872 non-null         object
```

Observations:

- There are total 872 rows and 7 columns in the dataset.
- Dependent variable Holiday is of object data type. All other independent variable is of integer data type.
- We can see 6 independents variable and one target variable – Holliday package as opted yes/no.
- From above data we can say that there are **no missing values** in the dataset for depth feature values in dataset.
- Also checked for duplicate values **no duplicate values** found in dataset.

Let’s check for summary statistics of the dataset for all variable.

Summary Statistics:



	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
count	872	872.000000	872.000000	872.000000	872.000000	872.000000	872
unique	2	NaN	NaN	NaN	NaN	NaN	2
top	no	NaN	NaN	NaN	NaN	NaN	no
freq	471	NaN	NaN	NaN	NaN	NaN	656
mean	NaN	47729.172018	39.955275	9.307339	0.311927	0.982798	NaN
std	NaN	23418.668531	10.551675	3.036259	0.612870	1.086786	NaN
min	NaN	1322.000000	20.000000	1.000000	0.000000	0.000000	NaN
25%	NaN	35324.000000	32.000000	8.000000	0.000000	0.000000	NaN
50%	NaN	41903.500000	39.000000	9.000000	0.000000	1.000000	NaN
75%	NaN	53469.500000	48.000000	12.000000	0.000000	2.000000	NaN
max	NaN	236961.000000	62.000000	21.000000	3.000000	6.000000	NaN

Table 12 – Summary Statistics of holiday package dataset

Observations:

- From the summary statistics we can see infer that, employee salary ranges from 1322 to 236961 also age of the employee varies from 20 years to 62 years
- Education of employee varies from 1 to 21 years, average and median for education looks almost equal to 9 years.
- In dataset no of young children and no of old children variable is discrete info. As we see average no of young children and older children is in decimal value. So we have to convert it into object data type to taken as category for dependent variable interaction.
- There are two type of employee in data set foreigner and not foreigner.

2.1.1 UNIVARIATE ANALYSIS

- We can carried out univariate analysis by using histogram and box plot for numerical continuous data and strip plot, count plot , pie plot for categorical data to find information or patterns in the data

Univariate Analysis: Continuous variable

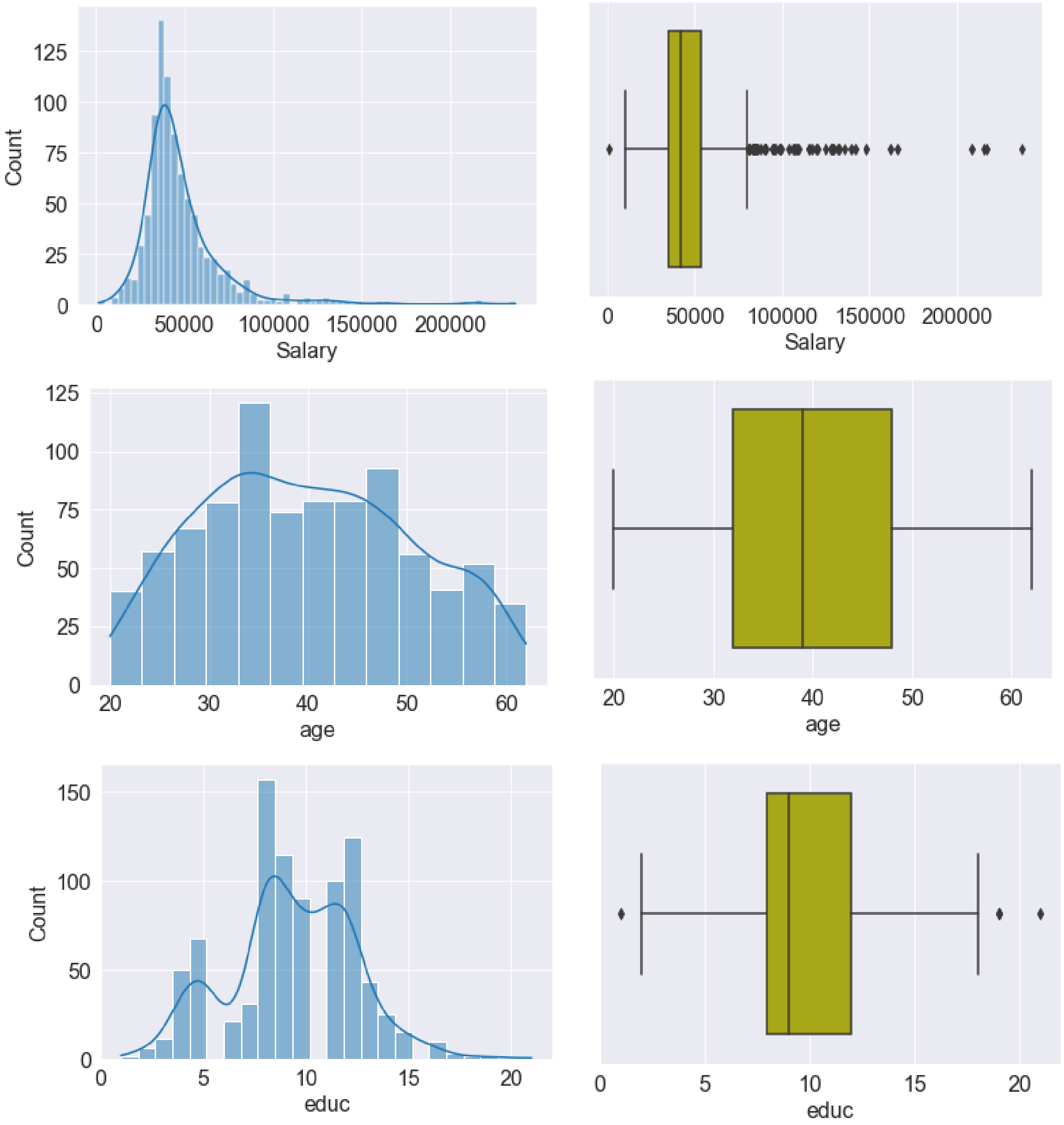


Fig. 8 – Histogram and Box plot for continuous variable holiday package dataset

Kurtosis & Skewness in Dataset:

Features	Skewness	kurtosis
Salary	3.103216	15.852557
age	0.146412	-0.909962
edu	-0.045501	0.005558
No_young children	1.946515	3.109892
No_older children	0.953951	0.676017

Table 13– Kurtosis and Skewness values for holiday package data variables

Observations:

- There are outlier present in salary variable which is valid for salary variables.
- From the histogram and the skewness value we can say the distribution of salary variable is right skewed.
- Education variable is look like normally distributed in spite the fact that it has few outlier present at both end (Few outlier in acceptable range).
- There is multimodal distribution observed in the education variable.
- The age feature almost follow normal distribution without any outlier in the data.

Univariate Analysis: Categorical variable

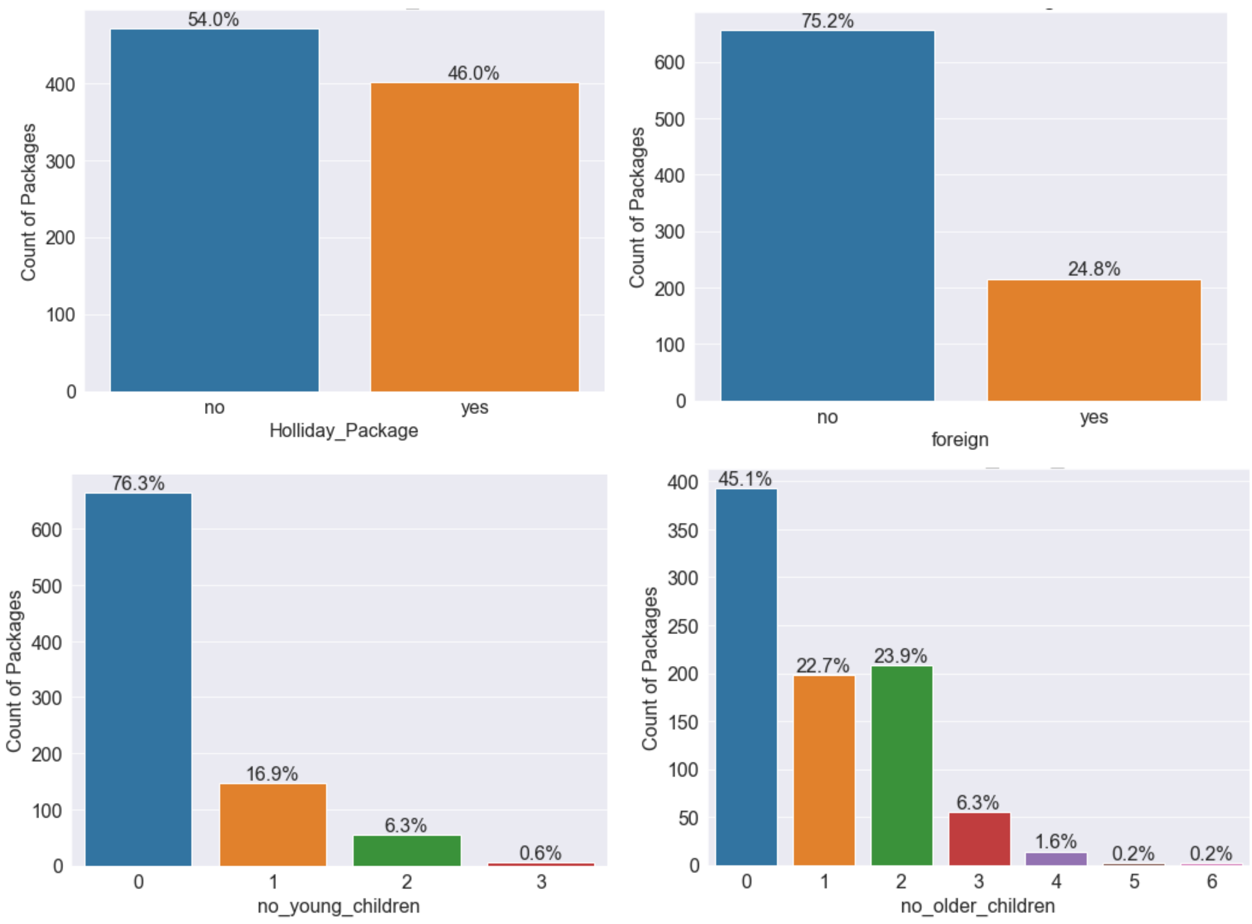


Fig. 9 – count plot for categorical variable holiday package dataset

Observations:

- 46 % of employee opted for holiday package, it looks target variable is balance in number of opted package or not.
- 25% of the employee are foreigner.
- Majority of employee do not have children as in young and older category.

2.1.2 BIVARIATE ANALYSIS

Target variable – Independent continuous variable

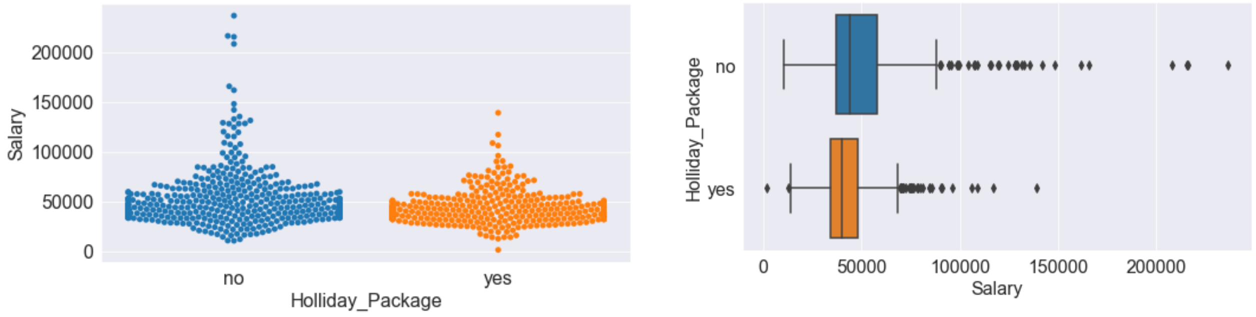


Fig. 10 – Swarm and box plot holiday package - Salary

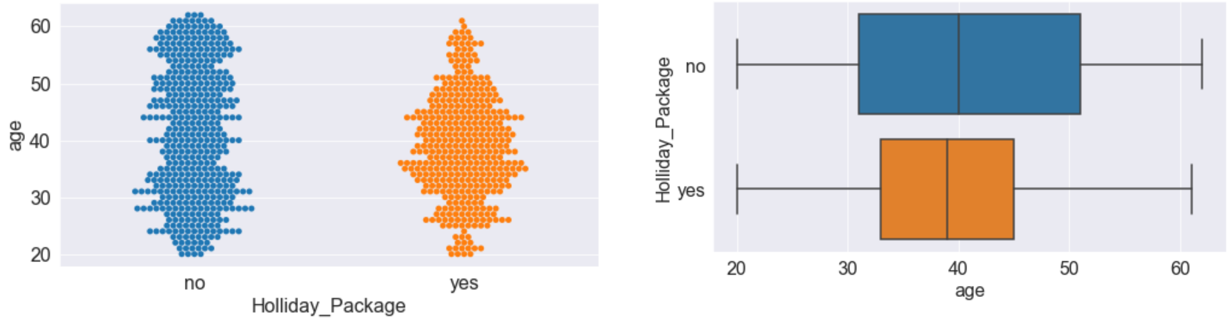


Fig. 11 – Swarm and box plot holiday package - age

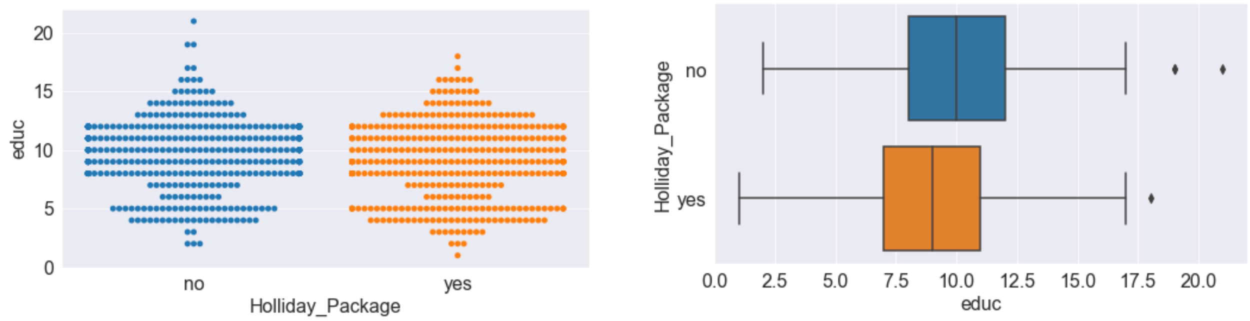


Fig. 12 – Swarm and box plot holiday package - educ

Observations:

- From the fig 10 There is no conclusive evidence for employee salary impacting of whether the employee opted for holiday package or not.
- Median age for holiday package opted for employee is less than the employee not opted but this will be more clearly with more data for evaluation.
- From the swarm plot the distribution of employee education opted for package is looks very similar to employee not opted.

Target variable – Independent categorical variable

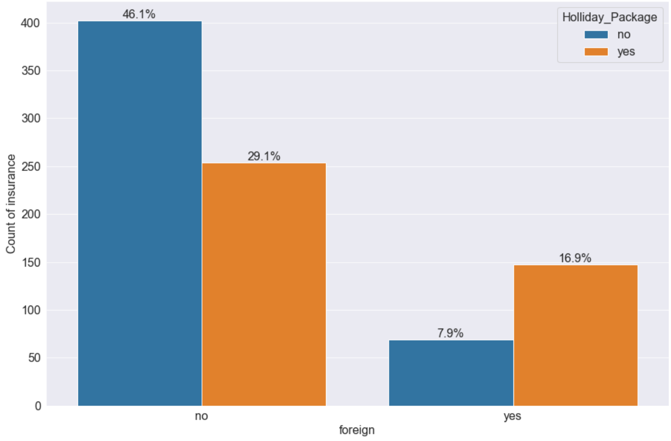


Fig. 13 – Bivariate analysis holiday package – Foreigner

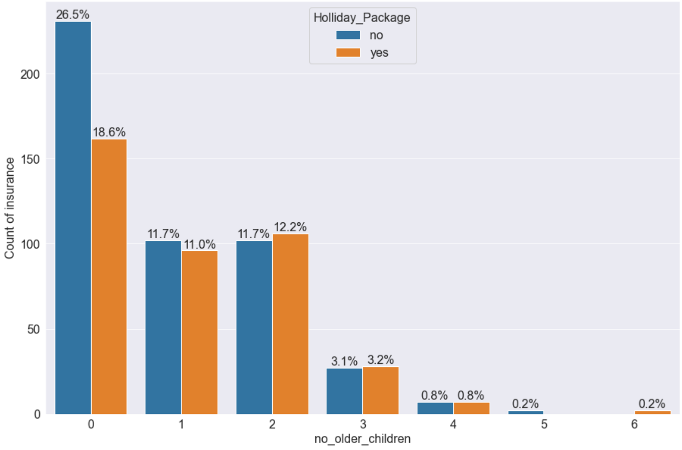


Fig. 14– Bivariate analysis holiday package – no_older children

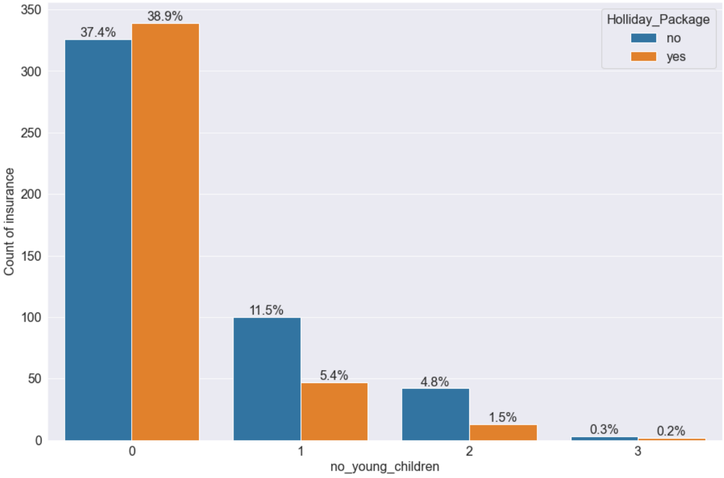


Fig. 15– Bivariate analysis holiday package – no_young children

Observations:

- From the fig.13 if the employee is foreigner employee then the average percentage package opted is more than other employee. This would be the good variable for package predication.
- Employees having no children have high chance for taking holiday package same will be true for both category as younger and older children.
- Employees with 1 or 2 young children have less probability for opting a holiday package compare to the other category.
- There is no sufficient data available for employee having more than 3 children as young or older. So we cannot conclude whether they opt for package or not basis of less data.

MULTIVARIATE ANALYSIS:

Checking for correlation of numerical variable by taking target variable holiday package as hue.

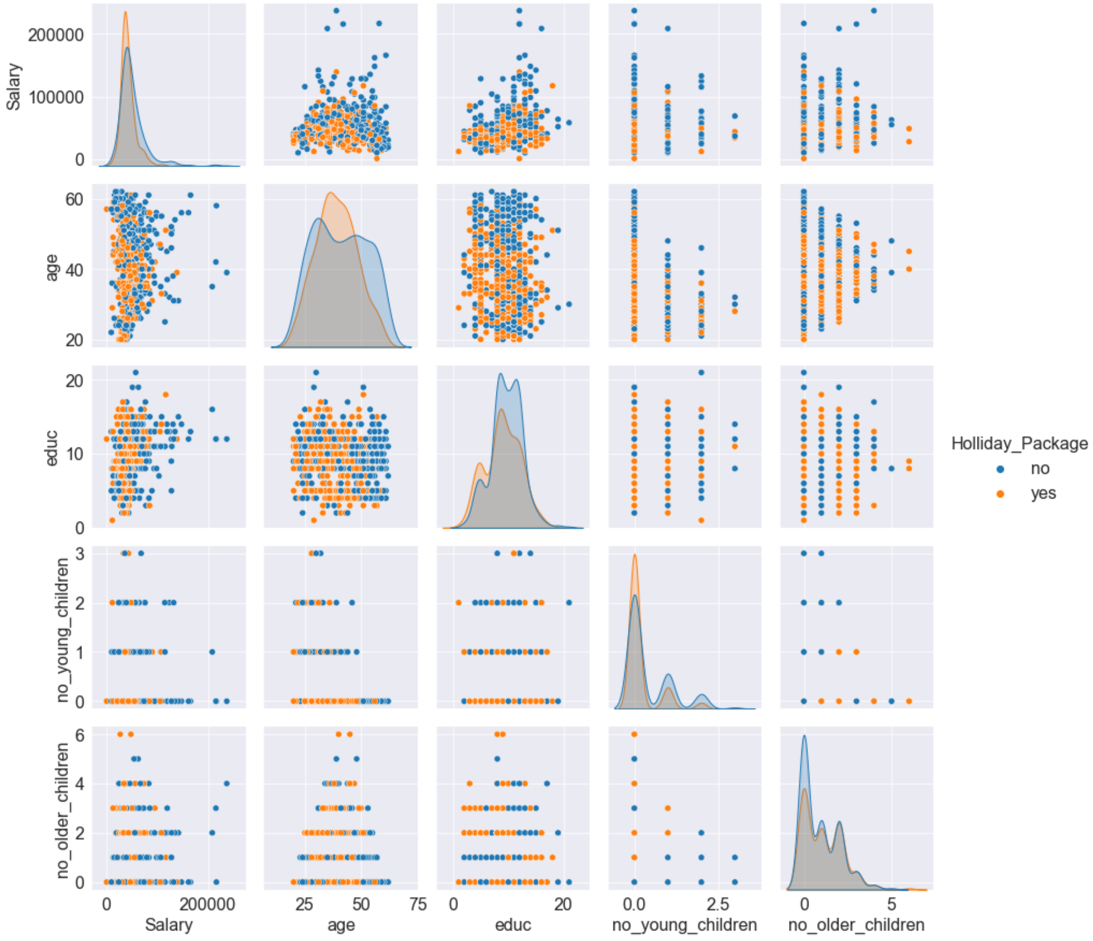


Fig. 16– Pair plot of numeric variable of holiday package dataset

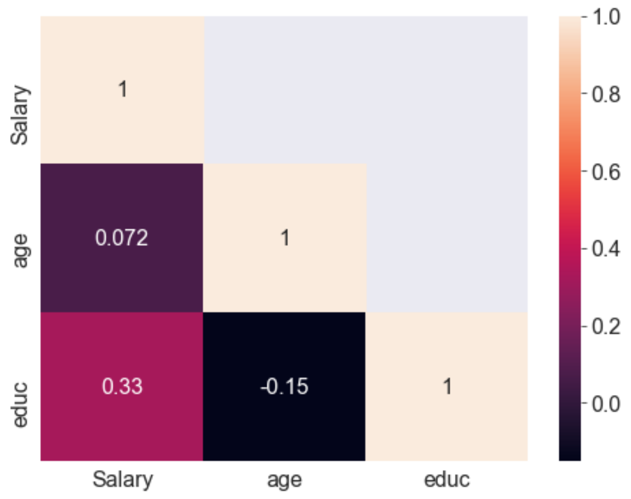


Fig. 17– Heat map matrix for Numerical variable of holiday package dataset

Observations:

- From the heat map of correlation matrix we can say there is no correlation between the salary, age and educ variable.
- From the pair plot educ variable shows bimodal distribution whereas number of old children and young children shows multimodal distribution
- Age follows the nearly normal distribution for employee who opted for holiday package.
- As of now from pair plot we cannot find the variable which is having strong impact on target variable as diagonal distribution overlapping each other for employee opted for holiday package or not.

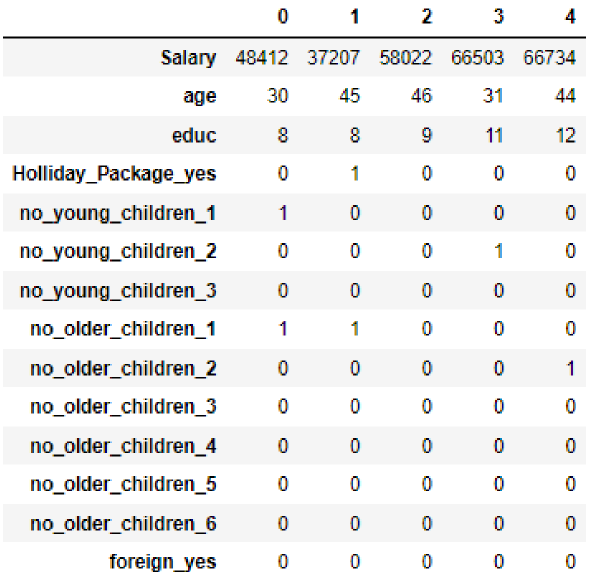
With this EDA analysis we take further for predicting the target variable i.e. Holiday package with using the other independent variable.
Also target variable is categorical (Yes/No) in nature we can able to do prediction by using the logistic regression, CART or RF techniques.

For this case we used the logistic regression technique with descriptive and Predictive approaches.

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.

As discuss in the EDA there are 3 numeric continuous variable in the dataset salary, age and educ. Other variable are categorical with including target variable Holiday package. We have to encode it before building the model as order of variable is not important here we opted for one hot dummy encoding.

After the encoding we have total 14 variable or columns in the data set (dropping first) As shown in sample of dataset below.



	0	1	2	3	4
Salary	48412	37207	58022	66503	66734
age	30	45	46	31	44
educ	8	8	9	11	12
Holliday_Package_yes	0	1	0	0	0
no_young_children_1	1	0	0	0	0
no_young_children_2	0	0	0	1	0
no_young_children_3	0	0	0	0	0
no_older_children_1	1	1	0	0	0
no_older_children_2	0	0	0	0	1
no_older_children_3	0	0	0	0	0
no_older_children_4	0	0	0	0	0
no_older_children_5	0	0	0	0	0
no_older_children_6	0	0	0	0	0
foreign_yes	0	0	0	0	0

Table 14 – Sample of dataset after encoding

We are building the model with descriptive approach as follows

Approach to Build Models

- Forward Selection
- Add columns and Check Adj Pseudo RSqaure
- Look at VIF values
- Remove column with high VIF value
- If no, multicollinearity is observed, remove columns based on p value

For model performance comparison we used **Pseudo R-square as criteria** of model selection.

Let’s build the models using this technique.

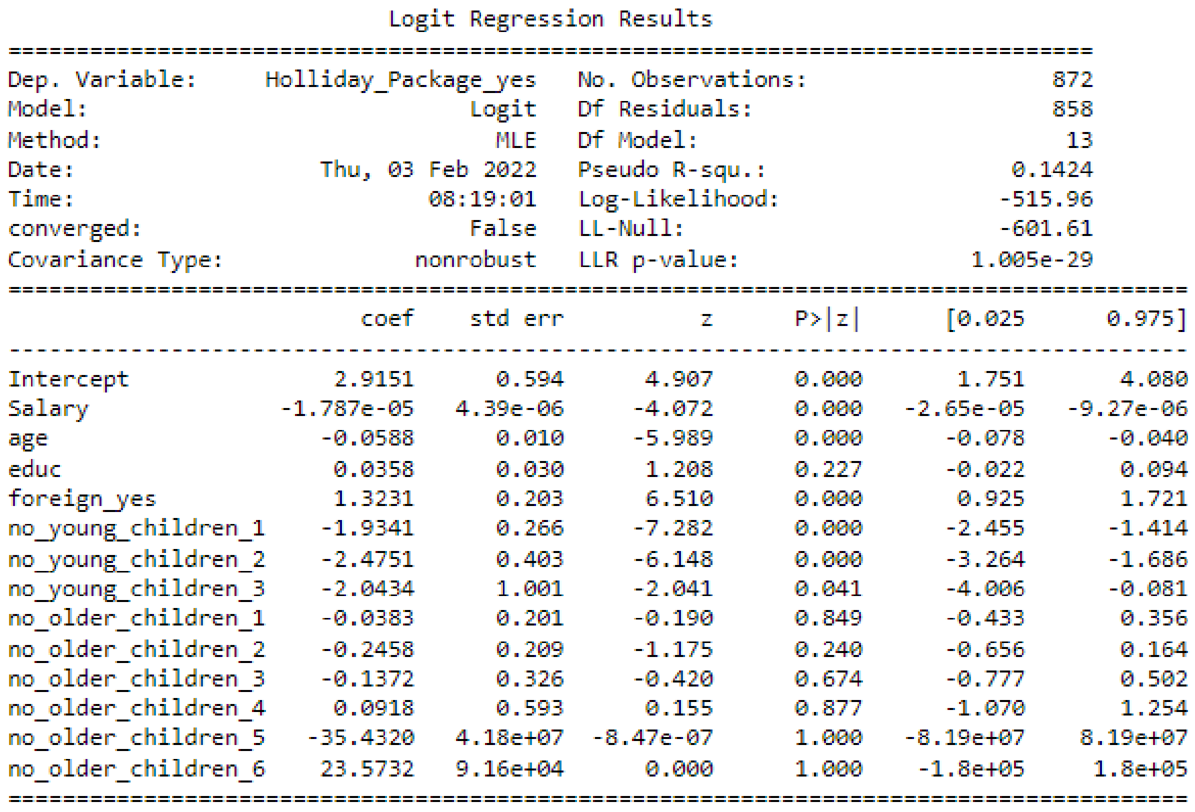
Model 1 : Using all variable for holiday package yes prediction

We used statsmodel library for descriptive approach to find the individual feature importance in the model.

The coefficient and intercept for model 1 as follow:

```
Intercept                2.915119
Salary                  -0.000018
age                     -0.058796
educ                     0.035850
foreign_yes              1.323125
no_young_children_1     -1.934109
no_young_children_2     -2.475101
no_young_children_3     -2.043436
no_older_children_1     -0.038296
no_older_children_2     -0.245837
no_older_children_3     -0.137160
no_older_children_4      0.091762
no_older_children_5    -35.432039
no_older_children_6     23.573207
```

Let's check for summary of model_1:



```
                           Logit Regression Results
================================================================================
Dep. Variable:     Holliday_Package_yes   No. Observations:                  872
Model:                            Logit   Df Residuals:                      858
Method:                             MLE   Df Model:                           13
Date:                  Thu, 03 Feb 2022   Pseudo R-squ.:                  0.1424
Time:                          08:19:01   Log-Likelihood:                -515.96
converged:                        False   LL-Null:                       -601.61
Covariance Type:              nonrobust   LLR p-value:                 1.005e-29
=======================================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept               2.9151      0.594      4.907      0.000       1.751       4.080
Salary              -1.787e-05   4.39e-06     -4.072      0.000   -2.65e-05   -9.27e-06
age                    -0.0588      0.010     -5.989      0.000      -0.078      -0.040
educ                    0.0358      0.030      1.208      0.227      -0.022       0.094
foreign_yes             1.3231      0.203      6.510      0.000       0.925       1.721
no_young_children_1    -1.9341      0.266     -7.282      0.000      -2.455      -1.414
no_young_children_2    -2.4751      0.403     -6.148      0.000      -3.264      -1.686
no_young_children_3    -2.0434      1.001     -2.041      0.041      -4.006      -0.081
no_older_children_1    -0.0383      0.201     -0.190      0.849      -0.433       0.356
no_older_children_2    -0.2458      0.209     -1.175      0.240      -0.656       0.164
no_older_children_3    -0.1372      0.326     -0.420      0.674      -0.777       0.502
no_older_children_4     0.0918      0.593      0.155      0.877      -1.070       1.254
no_older_children_5   -35.4320   4.18e+07  -8.47e-07      1.000   -8.19e+07    8.19e+07
no_older_children_6    23.5732   9.16e+04      0.000      1.000    -1.8e+05     1.8e+05
=======================================================================================
```

Observations:

- Intercept is positive for this model but for continuous variable all coefficient are low.it means there is no strong predictors for target variable. Last two variable shows significant values but data related to this point not enough to conclude.
- For model 1 Pseudo R-squ.: is 0.1424.
- There are insignificant variables in the model as p value higher than 0.05.
- Also we have to check for multicolinearity problem using the VIF score
- We can overcome this multicolinearity problem by implementing the PCA on dataset for dimensions reduction.

Checking for VIF score of variable.

```
Salary  VIF =  1.18
age  VIF =  1.75
educ  VIF =  1.42
no_young_children_1  VIF =  1.54
no_young_children_2  VIF =  1.32
no_young_children_3  VIF =  1.04
no_older_children_1  VIF =  1.25
no_older_children_2  VIF =  1.4
no_older_children_3  VIF =  1.14
no_older_children_4  VIF =  1.05
no_older_children_5  VIF =  1.01
no_older_children_6  VIF =  1.01
foreign_yes  VIF =  1.29
```

- Let's take VIF 2 as cut off value
- VIF of all variables found as less than 2 so we cannot drop variable base on VIF score for further model building.
- we can drop variables on basis of p values

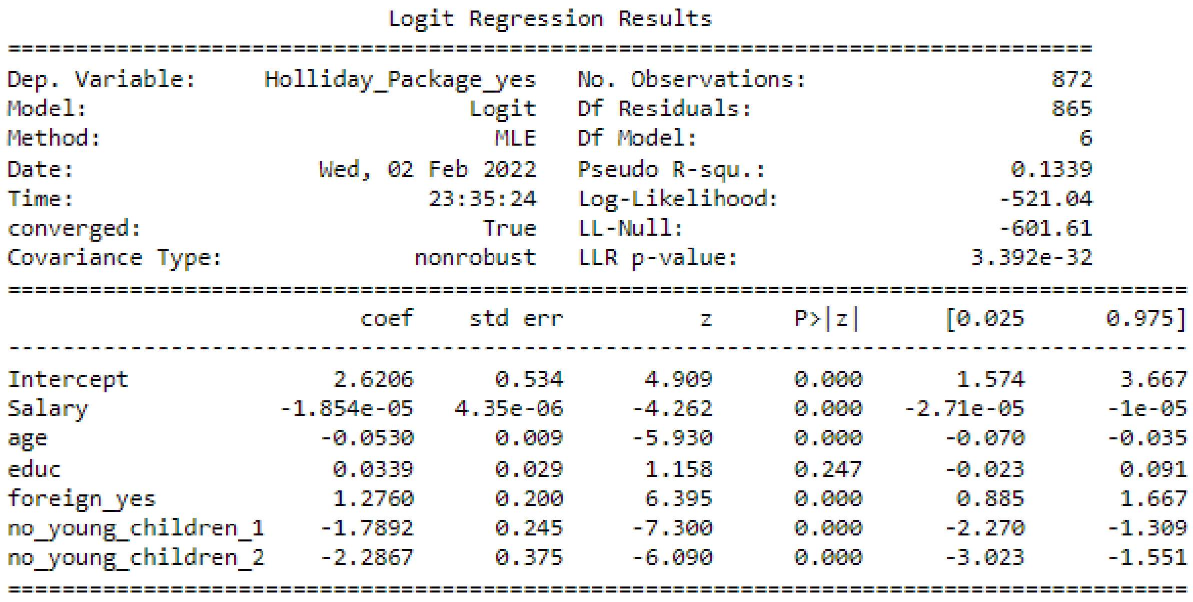
Model :2 Drop the variable which has the highest p-value>0.05 except education

As the P value of educ variable is also greater than 0.05 but it is independent variable and importance in a practical sense we can keep it as is in this model

The coefficient and intercept for model 2 as follow:

```
Intercept               2.620604
Salary                 -0.000019
age                    -0.052959
educ                    0.033891
foreign_yes             1.275983
no_young_children_1    -1.789203
no_young_children_2    -2.286750
```

Let's check for summary of model_2:



```
                           Logit Regression Results
==============================================================================
Dep. Variable:     Holliday_Package_yes   No. Observations:                  872
Model:                            Logit   Df Residuals:                      865
Method:                             MLE   Df Model:                            6
Date:                  Wed, 02 Feb 2022   Pseudo R-squ.:                  0.1339
Time:                          23:35:24   Log-Likelihood:                -521.04
converged:                         True   LL-Null:                       -601.61
Covariance Type:              nonrobust   LLR p-value:                 3.392e-32
=====================================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept               2.6206      0.534      4.909      0.000       1.574       3.667
Salary              -1.854e-05   4.35e-06     -4.262      0.000   -2.71e-05      -1e-05
age                    -0.0530      0.009     -5.930      0.000      -0.070      -0.035
educ                    0.0339      0.029      1.158      0.247      -0.023       0.091
foreign_yes             1.2760      0.200      6.395      0.000       0.885       1.667
no_young_children_1    -1.7892      0.245     -7.300      0.000      -2.270      -1.309
no_young_children_2    -2.2867      0.375     -6.090      0.000      -3.023      -1.551
=====================================================================================
```

Observations:

- For model 2 Pseudo R-squ.: is 0.1339 decreases as compare to Model1, but not that much significant decrease as this model uses only 6 variable as predictor.
- Education variable have high p value in this model as well. We have to build the model by dropping the educ variable as well.
- We have check for VIF score already so this model is free of multicolinearity.
- As we have check in EDA section Employee as a foreigner have significant potential to strongly predict target variable. Here coefficient for foreigner yes is positive with significant p value.
- Also number of young children 1 and 2 have negative coefficient values. It shows it contribute in predicting the target variable.
- Salary variable do not have significant contribution in prediction as its coefficient value is very less nearly zero.

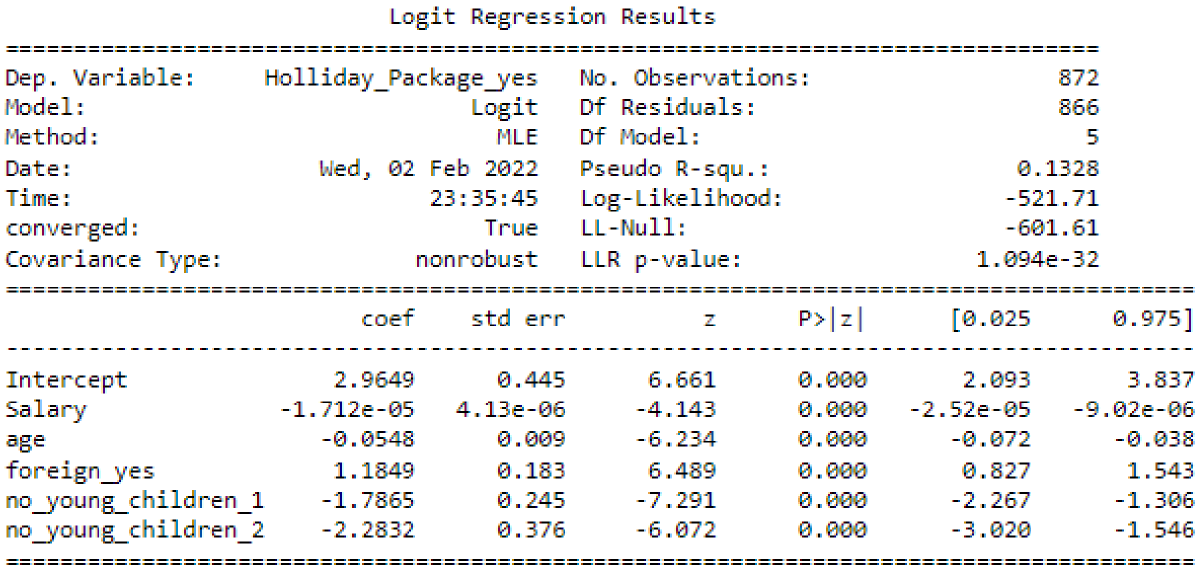
<u>Model :3 Drop the variable which has the highest p-value>0.05 (Drop educ variable)</u>

Let's build the new model dropping the education variable.(note this model is for cross examination in practice we never drop this kind of variable as it has practical importance)

The coefficient and intercept for model 3 as follow:

```
Intercept               2.964903
Salary                 -0.000017
age                    -0.054805
foreign_yes             1.184912
no_young_children_1    -1.786485
no_young_children_2    -2.28321
```

Let’s check for summary of model_3:



```
                           Logit Regression Results                           
==============================================================================
Dep. Variable:     Holliday_Package_yes   No. Observations:                  872
Model:                            Logit   Df Residuals:                      866
Method:                             MLE   Df Model:                            5
Date:                  Wed, 02 Feb 2022   Pseudo R-squ.:                  0.1328
Time:                          23:35:45   Log-Likelihood:                -521.71
converged:                         True   LL-Null:                       -601.61
Covariance Type:              nonrobust   LLR p-value:                 1.094e-32
=======================================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept               2.9649      0.445      6.661      0.000       2.093       3.837
Salary              -1.712e-05   4.13e-06     -4.143      0.000   -2.52e-05   -9.02e-06
age                    -0.0548      0.009     -6.234      0.000      -0.072      -0.038
foreign_yes             1.1849      0.183      6.489      0.000       0.827       1.543
no_young_children_1    -1.7865      0.245     -7.291      0.000      -2.267      -1.306
no_young_children_2    -2.2832      0.376     -6.072      0.000      -3.020      -1.546
=======================================================================================
```

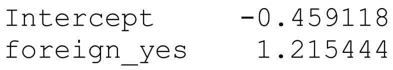
Observations:

- For model 3 Pseudo R-squ.: is 0.1328 not marginal decrease.
- All variables shows significance as p values are less than 0.05
- Still salary variable do not show significant coefficient value.

Model :4 Using foreigner_yes variable

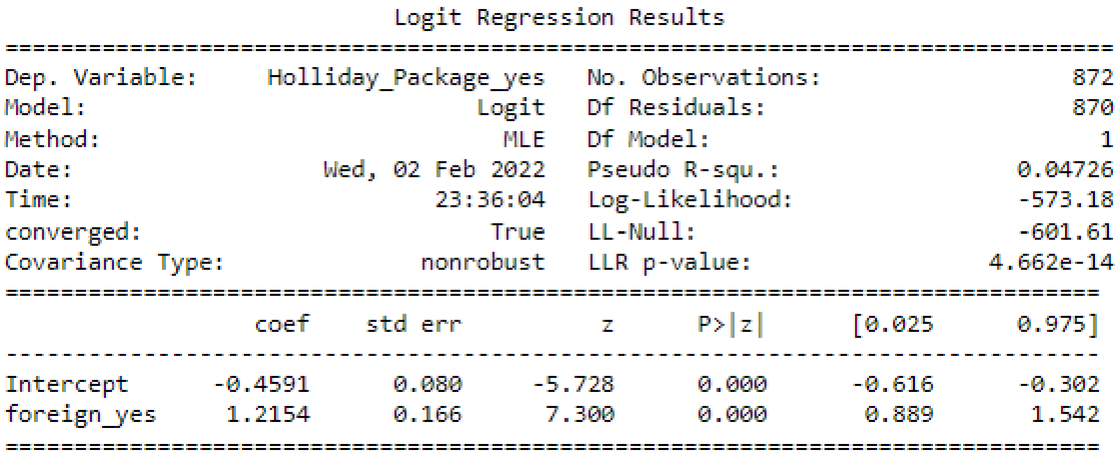
Let’s build model by using only foreigner as predictor. We have seen in a EDA this variable shows high significance in predicting the holiday package.

The coefficient and intercept for model 4 as follow:



```
Intercept     -0.459118
foreign_yes    1.215444
```

Let’s check for summary of model_4:



```
                           Logit Regression Results                           
==============================================================================
Dep. Variable:     Holliday_Package_yes   No. Observations:                  872
Model:                            Logit   Df Residuals:                      870
Method:                             MLE   Df Model:                            1
Date:                  Wed, 02 Feb 2022   Pseudo R-squ.:                 0.04726
Time:                          23:36:04   Log-Likelihood:                -573.18
converged:                         True   LL-Null:                       -601.61
Covariance Type:              nonrobust   LLR p-value:                 4.662e-14
===============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept      -0.4591      0.080     -5.728      0.000      -0.616      -0.302
foreign_yes     1.2154      0.166      7.300      0.000       0.889       1.542
===============================================================================
```

Observations:

- For model 4 Pseudo R-squ.: is 0.04726 decreases drastically. As it uses only one predictor for predicting.
- Coefficient value increases as compare to previous models.

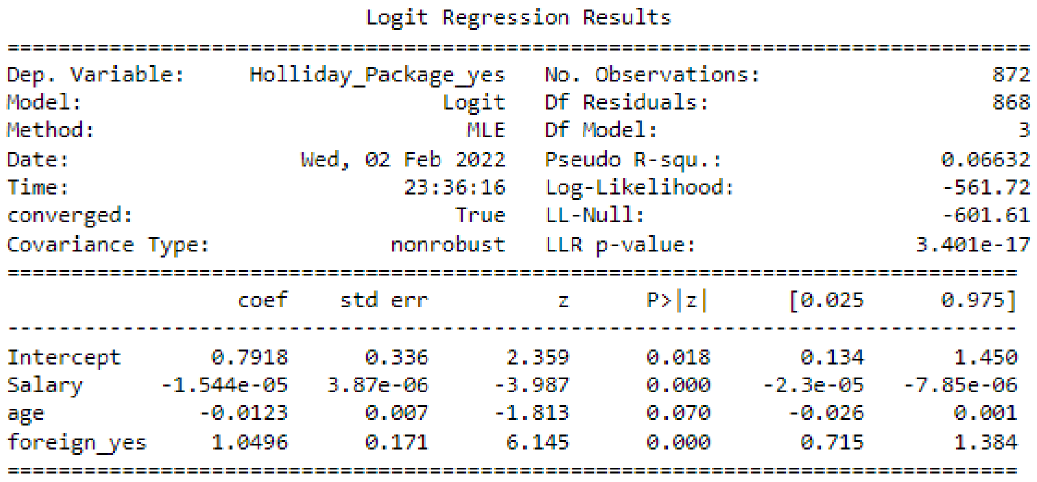
Model :5 Using Foreign_yes, age, salary

Let's build the new by using only foreigner and other continuous variable like age and salary.

The coefficient and intercept for model 5 as follow:

```
Intercept      0.791827
Salary        -0.000015
age           -0.012254
foreign_yes    1.049592
```

Let's check for summary of model_5:



```
                           Logit Regression Results
===============================================================================
Dep. Variable:     Holliday_Package_yes   No. Observations:                 872
Model:                            Logit   Df Residuals:                     868
Method:                             MLE   Df Model:                           3
Date:                  Wed, 02 Feb 2022   Pseudo R-squ.:                0.06632
Time:                          23:36:16   Log-Likelihood:               -561.72
converged:                         True   LL-Null:                      -601.61
Covariance Type:              nonrobust   LLR p-value:                3.401e-17
===============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept       0.7918      0.336      2.359      0.018       0.134       1.450
Salary      -1.544e-05   3.87e-06     -3.987      0.000    -2.3e-05   -7.85e-06
age            -0.0123      0.007     -1.813      0.070      -0.026       0.001
foreign_yes     1.0496      0.171      6.145      0.000       0.715       1.384
===============================================================================
```

Observations:

- For model 5 Pseudo R-squ.: is 0.06632 increases compare model4 .
- P value of age variable shows insignificance of the variable as p value>0.05.
- We have to investigate further by collecting more data why this shows insignificance in the prediction.

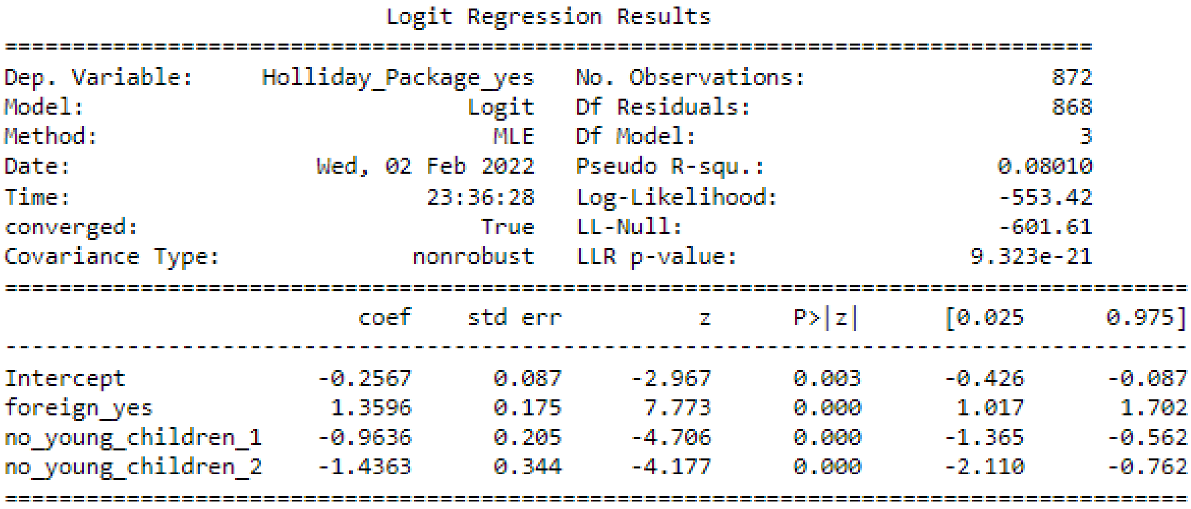
Model :6 Using Foreign_yes, no_young_children_1 and no_young_children_2

As we seen in the model building and EDA part no young children and foreigner is the good predictor and strongly impacting on target variable. Let's build a separate model by using this variables.

The coefficient and intercept for model 6 as follow:

```
Intercept              -0.256738
foreign_yes             1.359645
no_young_children_1    -0.963642
no_young_children_2    -1.436341
```

Let's check for summary of model_6:



```
                           Logit Regression Results
==============================================================================
Dep. Variable:     Holliday_Package_yes   No. Observations:                  872
Model:                            Logit   Df Residuals:                      868
Method:                             MLE   Df Model:                            3
Date:                  Wed, 02 Feb 2022   Pseudo R-squ.:                 0.08010
Time:                          23:36:28   Log-Likelihood:                -553.42
converged:                         True   LL-Null:                       -601.61
Covariance Type:              nonrobust   LLR p-value:                 9.323e-21
=======================================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept              -0.2567      0.087     -2.967      0.003      -0.426      -0.087
foreign_yes             1.3596      0.175      7.773      0.000       1.017       1.702
no_young_children_1    -0.9636      0.205     -4.706      0.000      -1.365      -0.562
no_young_children_2    -1.4363      0.344     -4.177      0.000      -2.110      -0.762
=======================================================================================
```

Observations:

- For model 6 Pseudo R-squ.: is 0.08010 increases compare model_5 .
- All variable has good significance as p value>0.05.
- The coefficient shows significance values. Number of young children increases the chance of opting a holiday package gets reduces.
- As employee is foreigner there will be more probably that he or she will opted for package.
- This are more suitable predictors for predicting the target variable

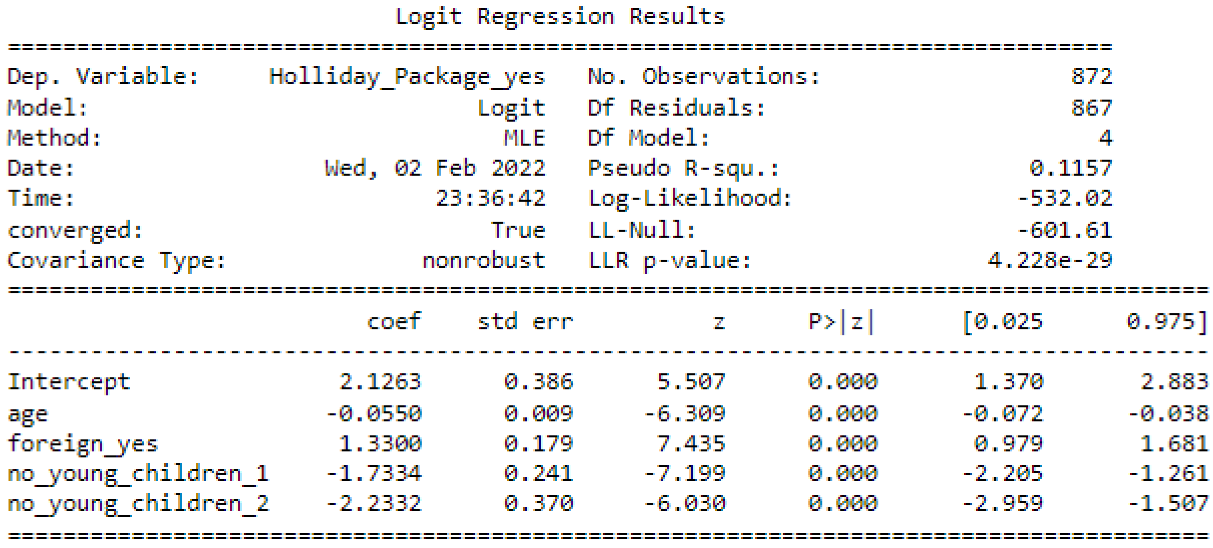
Model: 7 Using age, foreigners, no_young_children_1 and no_young_children_2

As we have seen in previous models salary variable not contributing significantly, so we can remove it from model 3.

The coefficient and intercept for model 7 as follow:

```
Intercept              2.126334
age                   -0.054978
foreign_yes            1.329957
no_young_children_1   -1.733389
no_young_children_2   -2.233161
```

Let’s check for summary of model_7:



```
                           Logit Regression Results
================================================================================
Dep. Variable:     Holliday_Package_yes   No. Observations:                  872
Model:                            Logit   Df Residuals:                      867
Method:                             MLE   Df Model:                            4
Date:                  Wed, 02 Feb 2022   Pseudo R-squ.:                  0.1157
Time:                          23:36:42   Log-Likelihood:                -532.02
converged:                         True   LL-Null:                       -601.61
Covariance Type:              nonrobust   LLR p-value:                 4.228e-29
=======================================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept               2.1263      0.386      5.507      0.000       1.370       2.883
age                    -0.0550      0.009     -6.309      0.000      -0.072      -0.038
foreign_yes             1.3300      0.179      7.435      0.000       0.979       1.681
no_young_children_1    -1.7334      0.241     -7.199      0.000      -2.205      -1.261
no_young_children_2    -2.2332      0.370     -6.030      0.000      -2.959      -1.507
=======================================================================================
```

Observations:

- For model 7 Pseudo R-squ.: is 0.1157 comparable to the best model till now .
- All variable has good significance as p value>0.05.
- This are model also suitable for predicting as significant Pseudo R square value with less number of feature.

2.2.1 Model Performance Comparison:

There is no direct measure of goodness of fit for a logistic regression. For a linear regression the total sum of squares and the residual sum of squares are two well defined quantities.

In case of logistic regression, these are not available. Hence it is difficult to quantify for a proposed logistic model, how much of the total variability in the data, it is able to explain.

Since the range of the pseudo-R2 is 0 to a number less than or equal to 1, the interpretation of the above values obtained in models is not easy. Instead of taking them as an absolute number, it is better to look at their relative values among the models under consideration.

Thus, it is clear that the model proposed as the Final Model has considerable higher Pseudo R2 values.

Let's compare model performance on the basis of Pseudo R2 values

Model	Predictors	Pseudo R^2
Model 1	All variable	0.142367
Model 2	Salary, age, educ, foreigner_ yes, no_young children_1, no_young children_2	0.133926
Model 3	Salary, age, foreigner_ yes, no_young children_1, no_young children_2	0.132806
Model 4	foreigner_ yes	0.047262
Model 5	Salary, age, foreigner_ yes	0.066315
Model 6	foreigner_ yes, no_young children_1, no_young children_2	0.080102
Model 7	age, foreigner_ yes, no_young children_1, no_young children_2	0.115672

Table 15 – Logistic Regression Model Performance comparison in descriptive approach

2.2.2 Model Selection:

- Model_2 , Model_3 and Model_7 shows high Pseudo R-squared values than the other models with less number of predictors variables
- These 3 models are free of multicolinearity as we have seen in model building.
- Model 1 shows highest Pseudo R square value but it has problem of multicolinearity with more number of variables
- For prescriptive analysis **model 2 is suitable** with only 6 independent variable.
- Also we can think of **model 7** as it shows significant value of Pseudo R square with only 4 independent variable.

For Model 2 Coefficient and intercept is:

```
Intercept               2.620604
Salary                 -0.000019
age                    -0.052959
educ                    0.033891
foreign_yes             1.275983
no_young_children_1    -1.789203
no_young_children_2    -2.286750
```

We can explore further with the predictive approach.

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.

- If we only wanted to predict using Logistic Regression and were not looking for the model building aspect of it, we can do that as well.
- For this exercise, we will use the same variables as we have used in the previous models.

Key Differences in Predictive Modelling.

- We will split the data into train and test and get an idea about the expected quality of predictions in future
- We will need to choose a metric of interest. Let's choose classification report with precision, recall, f1 score and accuracy of train and test.
- Build the model on the training data and check the classification report on the train and test data.

Note: We are going to build all the models, get their predictions and then go on to evaluate those models

We have use Sklearn library for linear regression

We will be doing a **70:30** split. 70% of the whole data will be used to train the data and then 30% of the data will be used for testing the model.

Dimension of train & test data

	rows	columns
X_train	610	14
X_test	262	14
Y_train	610	
Y_test	262	

Table 16- Dimension of train and test data for Holiday Package dataset

We used statisy function here to split the same proportion of target variable in train and test.

Model using parameter as follows:

Solver='newton-cg',
Penalty='none',
max_iter=1000,
Verbose=True,
n_jobs=2,
Random state=2

Note: We can also make the model on the basis of hyper parameter selection by using Grid search CV for this instance we used common parameter as listed above.

Now we can make model on training data and predict it against the training as well as test data and evaluate on the basis of classification report.
We can check previously build model for prediction purpose

Model 1: Using all variables
Model 2: Salary, age, educ, foreigner_ yes, no_young children_1, no_young children_2
Model 3: Salary, age, foreigner_ yes, no_young children_1, no_young children_2
Model 4: foreigner_ yes
Model 5: Salary, age, foreigner_ yes
Model 6: foreigner_ yes, no_young children_1, no_young children_2
Model 7:, age, foreigner_ yes, no_young children_1, no_young children_2

We have built the above six model on the training data and checked against the test data. We have used here the classification report as a matrix of interest.

Let's check out the models performance on Train and Test data for price package prediction for different model:

2.3.1 Model Performance Comparison:

In this case data is balance as holiday package opted and not opted in 46:54 ratio. i.e. 54 % employee not going to take package and 46 % employee opted for holiday package.

- If employee opted for package or not that is the area of interest then we will go for accuracy as a measure of performance.
- If company interest is to be focus on one category we will go for precision, recall and F1 score matrix of measure

Let's compare the performance of above 7 models on the prediction of test data.

Model	AUC	Precision		Recall		F1 Score		Accuracy	
		Train	Test	Train	Test	Train	Test	Train	Test
Model 1	0.750	0.66	0.70	0.56	0.57	0.60	0.63	0.67	0.69
Model 2	0.756	064	0.67	0.54	0.57	0.59	0.61	0.65	0.67
Model 3	0.754	0.65	0.68	0.57	0.56	0.60	0.61	0.66	0.68
Model 4	0.605	0.68	0.68	0.37	0.35	0.48	0.46	0.63	0.63
Model 5	0.670	0.68	0.66	0.38	0.34	0.49	0.45	0.63	0.62
Model 6	0.660	0.72	0.67	0.36	0.33	0.48	0.44	0.64	0.62
Model 7	0.732	0.64	0.65	0.53	0.53	0.58	0.59	0.65	0.66

Table 17 – Logistic Regression Model Performance comparison in Predictive approach

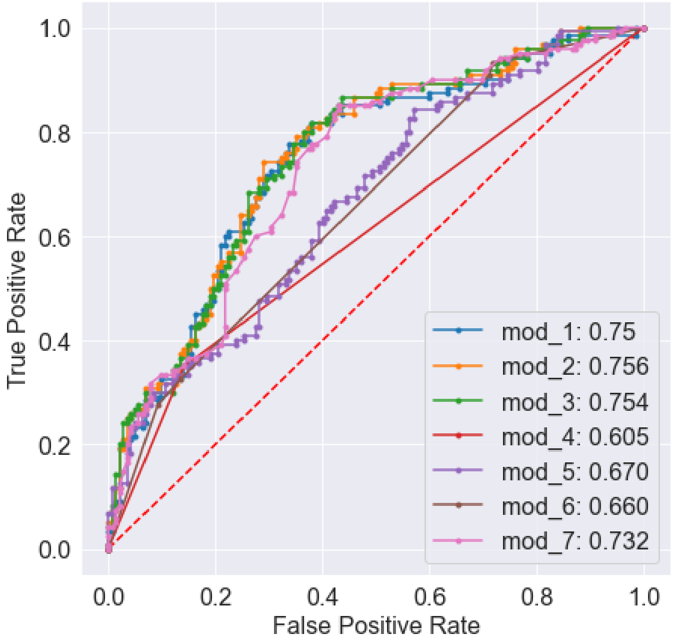


Fig. 18– AUC_ROC curve for prediction on Test data for all Models

Observations:

- From above we can see the accuracy score for all models on train and test almost all models shows more accuracy on test data than train except model5. It shows models are under fitted and needs more data for training.
- If company want to know only the yes or no kind of information just to predict employee will opt package or not the accuracy will be the measure of performance of model. In that case model 1, model 3 and Model 6 will be the better fit.
- Also we can see from the AUC ROC curve the model2, model 3 and Model 6 having good AUC score.
- If we looking for model with comparatively higher accuracy with less number of independent variables then model 6 will be the good choice having only 3 independent variable.
- Also if we see the accuracy score for model 1 then it is higher compare to all models with more number of variables.

Comparison with descriptive Approach:

From the descriptive approach we can see the model 2, model 3 and Model 7 was better fit models.
Here in predictive approach model 1, model 3 and Model 7 is the better fit models.

Let's discuss about the model 7 with which shows comparatively higher performance and consistence in train and test predictive approach.
We can see recall for train and test for model7 is equal also train and test accuracy score are almost equal. So model 7 will performed better in train and test beside the performance can be improved by interrogating new feature with more significance with respect to the prediction.

2.3.2 Model Selection for prediction:

Model 7 : age, foreigner_ yes, no young children_1, no, young children_2

- Classification report on Train data Model_7:

	precision	recall	F1-Score	support
0	0.65	0.74	0.70	329
1	0.64	0.53	0.58	281
accuracy			0.65	610
Macro avg	0.65	0.64	0.64	610
Weighted avg	0.65	0.65	0.64	610

Table.18 – Classification report on Train data

- Confusion Matrix on Train data:

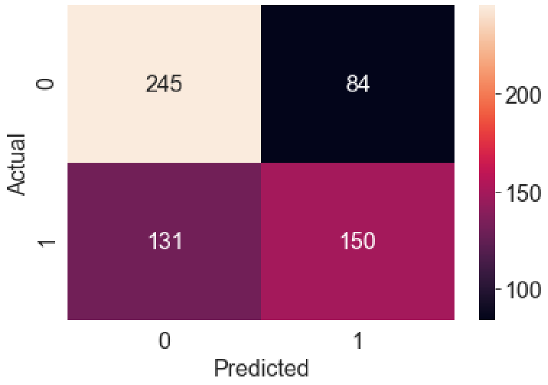


Fig.19 Heat map Confusion Matrix for Train data

Observations:

- Predicting employee who will opt for package
  - Precision (64%) – 64 % of prediction of Employee who will opt for package are correct.
  - Recall (53%) – 53 % of Employee who will opt for package correctly predicted.

- Predicting employee who will not opt for package
  - Precision (65%) – 65 % of prediction of Employee who will not opt for package are correct.
  - Recall (74%) – 74 % of Employee who will not opt for package correctly predicted.

- Model overall accuracy – 65 % of total prediction are correct.

- Classification report on Test data Model_7:

	precision	recall	F1-Score	support
0	0.66	0.76	0.71	142
1	0.65	0.53	0.59	120
accuracy			0.66	262
Macro avg	0.66	0.65	0.65	262
Weighted avg	0.66	0.66	0.65	262

Table.19 – Classification report on Test data

- Confusion Matrix on test:

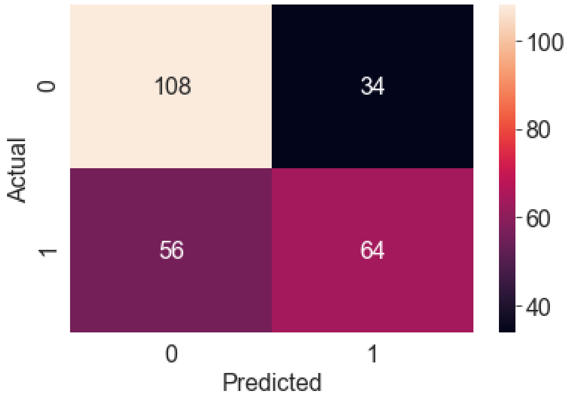


Fig. 20 Heat map Confusion Matrix for Test data

Observations:

➢ Predicting employee who will opt for package
  - Precision (65%) – 65 % of prediction of Employee who will opt for package are correct.
  - Recall (53%) – 53 % of Employee who will opt for package correctly predicted.

➢ Predicting employee who will not opt for package
  - Precision (66%) – 66 % of prediction of Employee who will not opt for package are correct.
  - Recall (76%) – 76 % of Employee who will not opt for package correctly predicted.

➢ Model overall accuracy – 66 % of total prediction are correct.

Conclusion:

For Descriptive logistic regression (Feature Analysis)

- From the EDA and the models we infer that as employee being a foreigner have more probability to opt for holiday packages
- Number of young children increases the probability for opting the holiday package decreases.
- Number of older children increases the probability of taking holiday package increases.
- The surprisingly salary variable not impacting the probability of holiday package opt.
- At some extent it is true that employee with higher age shown more interest to opt holiday package.

For predictive logistic regression (Prediction)

- As discuss in the comparison section model performing almost similar in both approaches. Model 7 with consistency in train and test performance can be used for prediction of employee opt or not opt holiday package.
- We can also do the analysis using CART and Random Forest technique to compare the accuracy and other performance measure to choose best model.

Recommendations:

- As we seen in EDA Foreigner employee are more interested to opt holiday packages. So company should come up with more attractive and required packages for this category of employee.
- Employee with no young children also have greater chance to opt for packages to company has to target these category as well.
- Employee with higher age have higher chance to opt packages so company can come up with the new packages for this category employee.

As all Model performance not up to the mark so company have to check for other feature for better model performance and good predictions.