PG-DSBA Dec_A 2020

Predictive Modelling Project

Submitted by: F Maria Jasmine

Date of Submission: 06/06/2021

Table of Contents

II. Problem Statement: Logistic Regression and Linear Discriminant Analysis

Problem 1: Linear Regression

You are hired by a company Gem Stones co ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of almost 27,000 cubic zirconia (which is an inexpensive diamond alternative with many of the same qualities as a diamond). The company is earning different profits on different prize slots. You have to help the company in predicting the price for the stone on the bases of the details given in the dataset so it can distinguish between higher profitable stones and lower profitable stones so as to have better profit share. Also, provide them with the best 5 attributes that are most important.

Data Dictionary:

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.
Color	Colour of the cubic zirconia.With D being the best and J the worst.
Clarity	cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, FL = flawless, I3= level 3 inclusions) FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3
Depth	The Height of a 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.

Problem statement:

Using the Linear regression model, the Prices for the CZ diamonds are predicted and the 5 main important attributes which have high impact on the Price is selected.

1.1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA). Perform Univariate and Bivariate Analysis.

Data Description:

1. The Cz diamond data set has 26967 observations and 10 features excluding the Un-named column.
2. There are 33 duplicate data in the data. They are not removed. Sale of two CZ diamonds can have similar specifications and the duplicates may not exactly indicate repetitions. No duplicates removed.
3. There were 697 null values in the depth feature. Other all features had 0 null values. The null values in depth are imputed using (Height/Width *100) of that particular row. The data dictionary has mentioned depth as "The Height of a cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.".
4. The given data has 7 Numeric features and 3 Discrete features.
5. Numerical Data summary:

Table: 1.1: Numerical Description Summary

	count	mean	std	min	25%	50%	75%	max
carat	26967.0	0.798375	0.477745	0.2	0.40	0.70	1.05	4.50
depth	26270.0	61.745147	1.412860	50.8	61.00	61.80	62.50	73.60
table	26967.0	57.456080	2.232068	49.0	56.00	57.00	59.00	79.00
x	26967.0	5.729854	1.128516	0.0	4.71	5.69	6.55	10.23
y	26967.0	5.733569	1.166058	0.0	4.71	5.71	6.54	58.90
z	26967.0	3.538057	0.720624	0.0	2.90	3.52	4.04	31.80
price	26967.0	3939.518115	4024.864666	326.0	945.00	2375.00	5360.00	18818.00

- 75% of the observations has a carat size of 1.05. Only 25% of the observations has greater 1.05 carat size.
- The 75% of the observations have price less than 5500 because the carat size is less than 1.05. When the weight is low, price is also low.
- The ideal depth percentage is 59 – 62.6 %. Too Low depth % - dark appearance as diamond does not reflect light as well. dull appearance as diamond loses out the light. Increase in depth % will not increase the price. No linear relationship.
- The ideal table percentage is 54 – 57% it is calculated by dividing the Table width by diameter. Too low Table % - light gets trapped and leaks out the sides of the diamonds. Too high Table % - light does not reflect off of the crown and it makes it dull.

- Anomalies: Dimensions x, y and z has zeros. When one of the dimensions is zero the other two dimensions cannot exist. The rows with zero in dimensions are removed.
- Anomalies: y and z has very high measurements for 2 observations which are irrelevant with the other two dimensions. Hence, the extreme outliers are also removed.

6. Discrete Data summary.

Table 1.2: Discrete Description summary

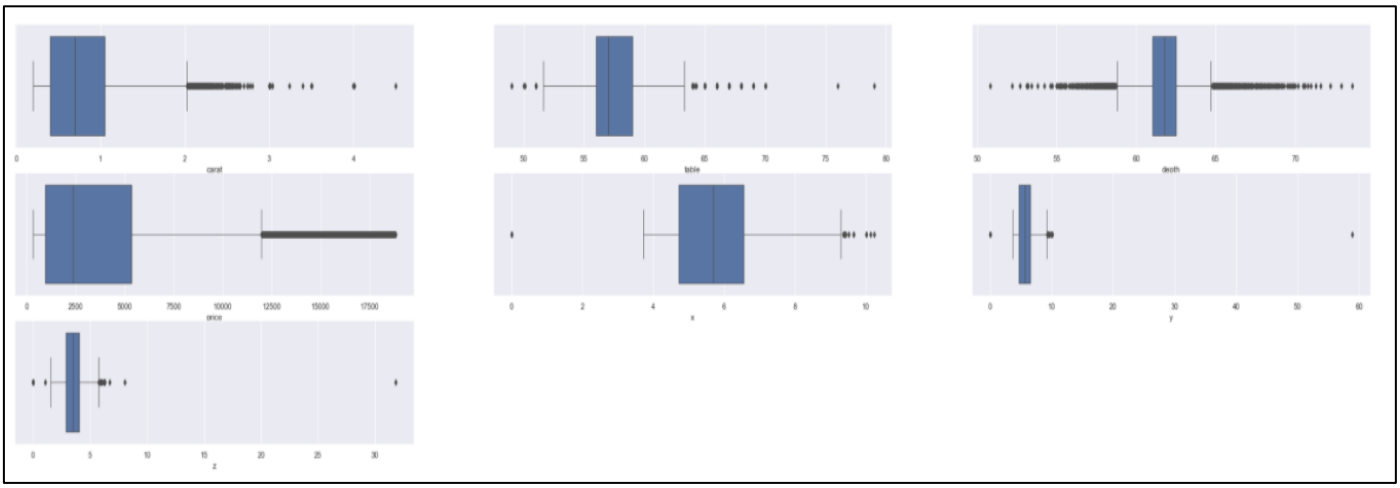
	count	unique	top	freq
cut	26967	5	Ideal	10816
color	26967	7	G	5661
clarity	26967	8	SI1	6571

- Cut category as 5 classes, Colour has 7 classes and Clarity has 8 classes.
- The Ideal Cut, G Colour and SI1 Clarity have high frequencies.

Outliers Treatment:

1. Outliers in Depth, Table, x, y and z are removed using Inter quartile range (Winsorizarion).
2. As the size of the carat increases, price also increase. Treating the outliers in Carat feature using IQR will assign high price to the less weight carats. Only 0.26% of the data has carat size more than 2.5. These special sizes can be analysed separately. For the model, Carat size (< = 2.5) is considered.
3. The extreme outliers in z and y are also removed.
4. Treating outliers in Target variable – Price will affect the predictions. No outlier treatment for Price variable.

Fig 1.1: Boxplot representation before treating outliers and Anomalies:



Exploratory Data Analysis:

Table 1.3: Coefficient of Variance (CV%):

The coefficient of variation is a statistical measure of the dispersion of data points around the mean.			
Feature	**Mean**	**Standard Deviation**	**CV%**
Carat	0.798	0.478	59.840
Depth	61.745	1.413	2.288
Table	57.456	2.232	3.885
x - Length	5.730	1.129	19.695
y - Width	5.734	1.166	20.337
z - Height	3.538	0.721	20.368
Price	3939.518	4024.865	102.166

The data points in Carat and Price are highly dispersed. The other features' CV% are comparatively less.

Table 1.4: Skewness and Kurtosis:

Feature	Skewness	Skew Type	Kurtosis	Kurtosis Type
Carat	0.9997	> 0.5 < 1 |Slightly right skewed	0.5103	< 3 | Platykurtic - Low Peak
Depth	-0.2287	(-0.5 to +0.5) | Approximately symmetrical	0.1759	< 3 | Platykurtic - Low Peak
Table	0.4835	(-0.5 to +0.5) | Approximately symmetrical	-0.0136	< 3 | Platykurtic - Low Peak
x - Length	0.3724	(-0.5 to +0.5) | Approximately symmetrical	-0.8104	< 3 | Platykurtic - Low Peak
y - Width	0.3694	(-0.5 to +0.5) | Approximately symmetrical	-0.8151	< 3 | Platykurtic - Low Peak
z - Height	0.366	(-0.5 to +0.5) | Approximately symmetrical	-0.8139	< 3 | Platykurtic - Low Peak
Price	1.6224	> 1 | Extremely right skewed	2.1917	< 3 | Platykurtic - Low Peak

Fig 1.2: Univariate Analysis: Numerical columns:

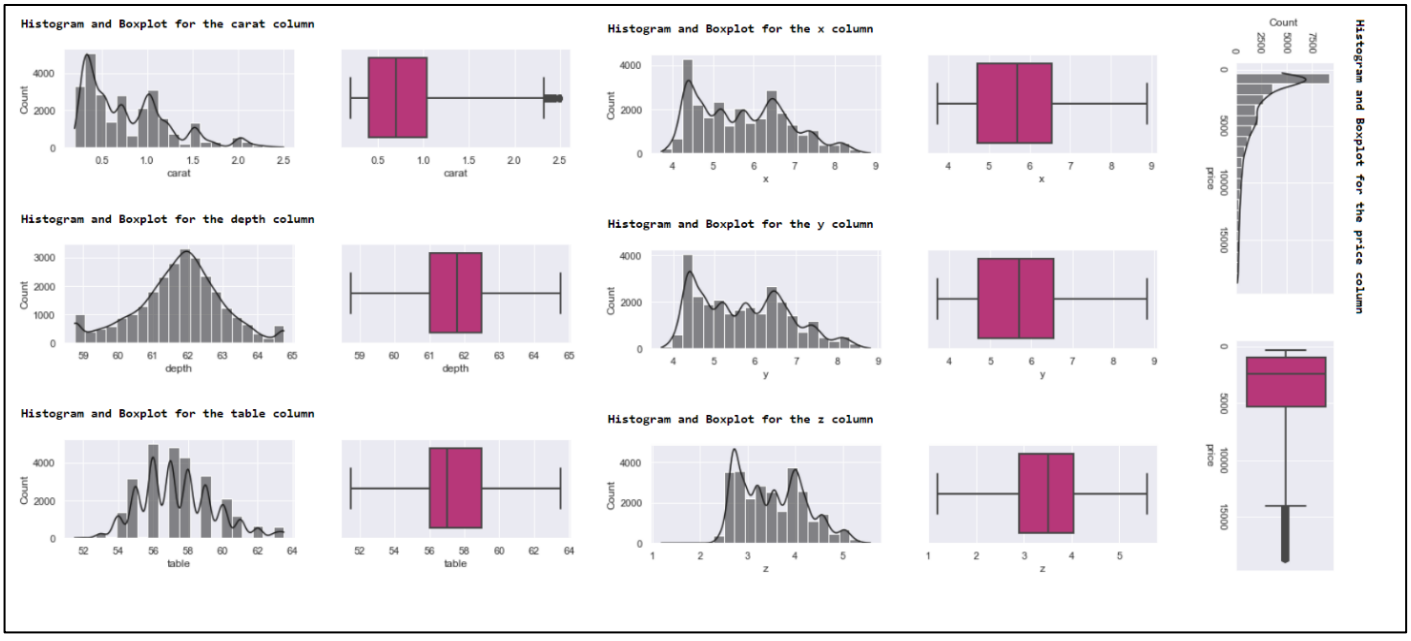
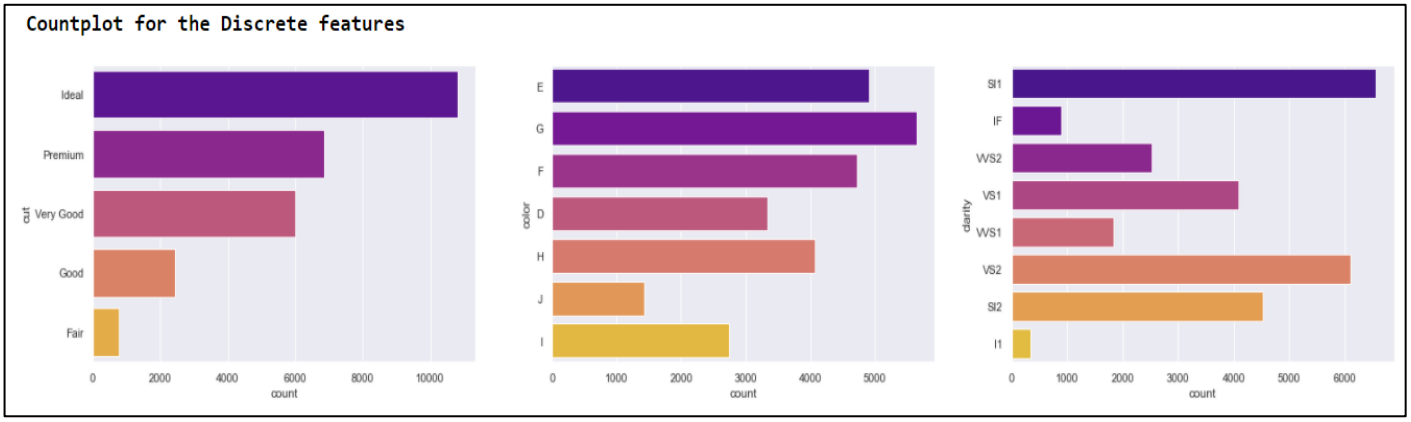
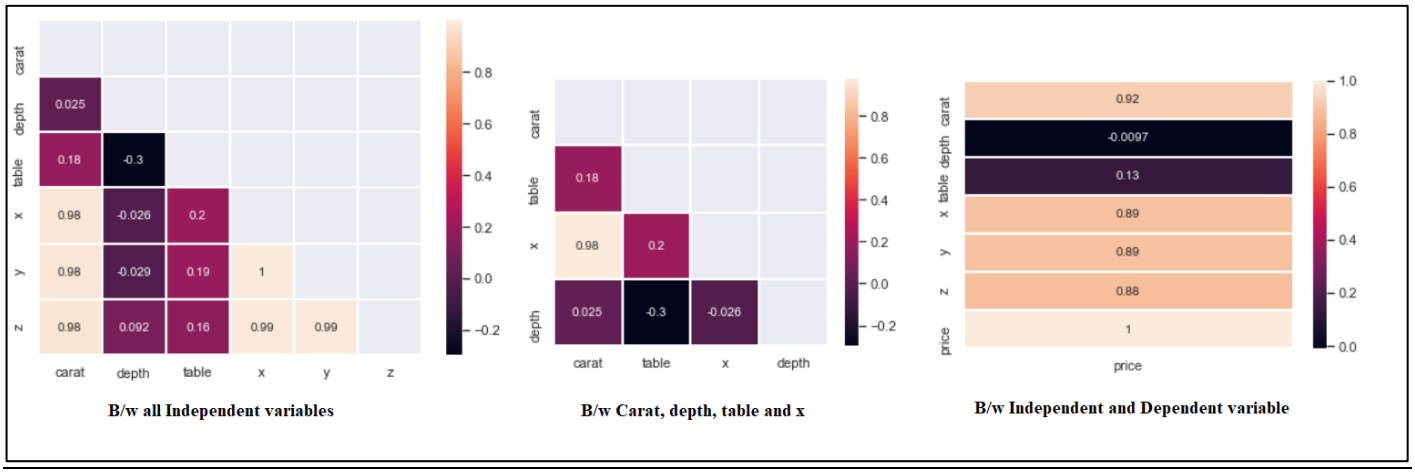


Fig 1.3: Univariate Analysis: Discrete columns:



Bivariate Analysis:

Fig 1.4: Bivariate Analysis: Correlation plots:

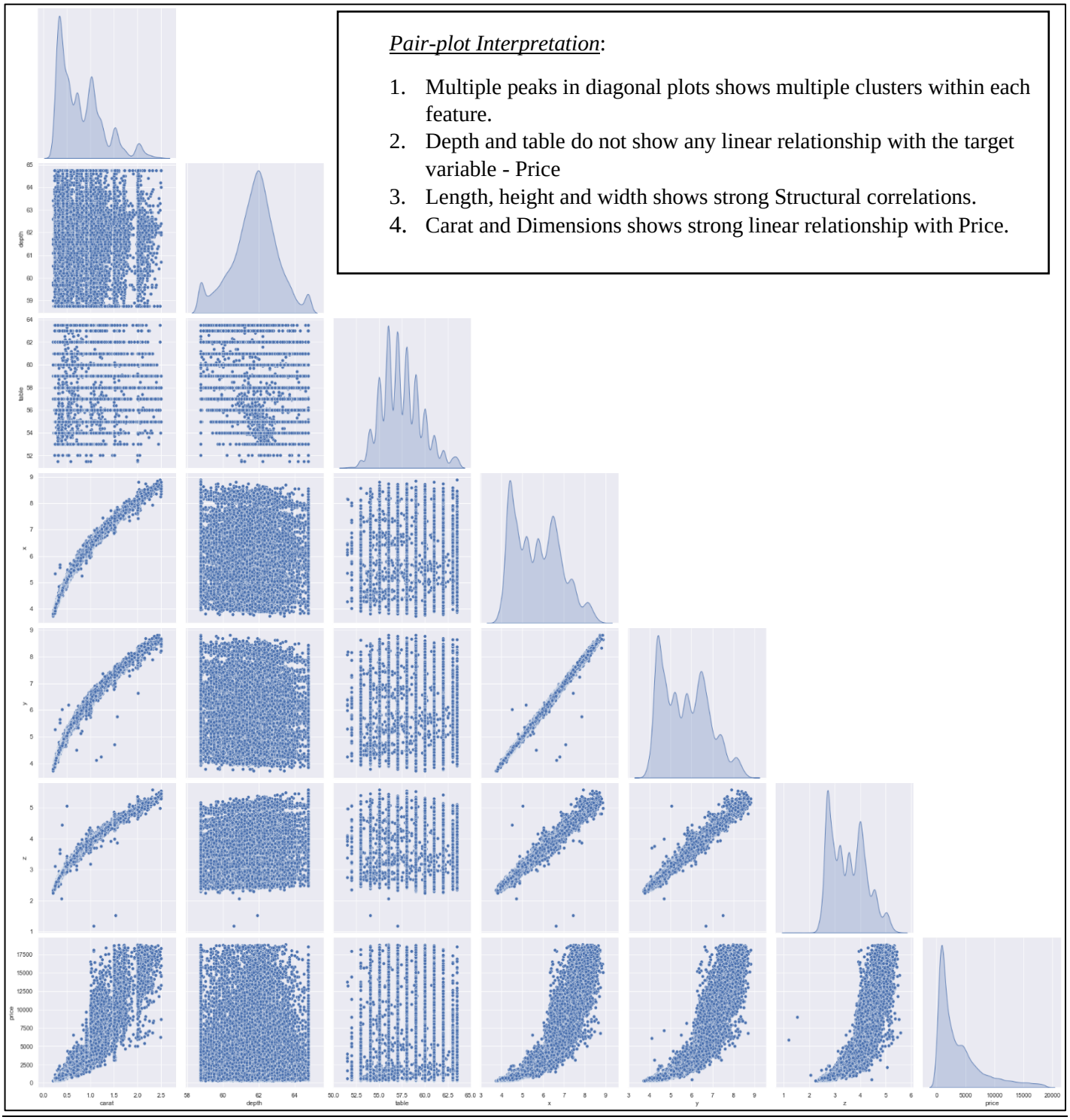


1. In Fig 1.4. The first plot shows strong multicollinearity between the independent variables due to Structural correlation between x, y and z dimensions.

2. The second plot is created after removing y and z (height and width). The information from y and z is available with depth feature. This plot shows collinearity between carat and x. Increase in carat size increases length of the diamonds.

3. Third plot shows the linear relationship between the Predictors and response variables. Very strong linear relationship between Carat, x, y and z and Price can be seen. We can say that these features are important for predicting the Price.

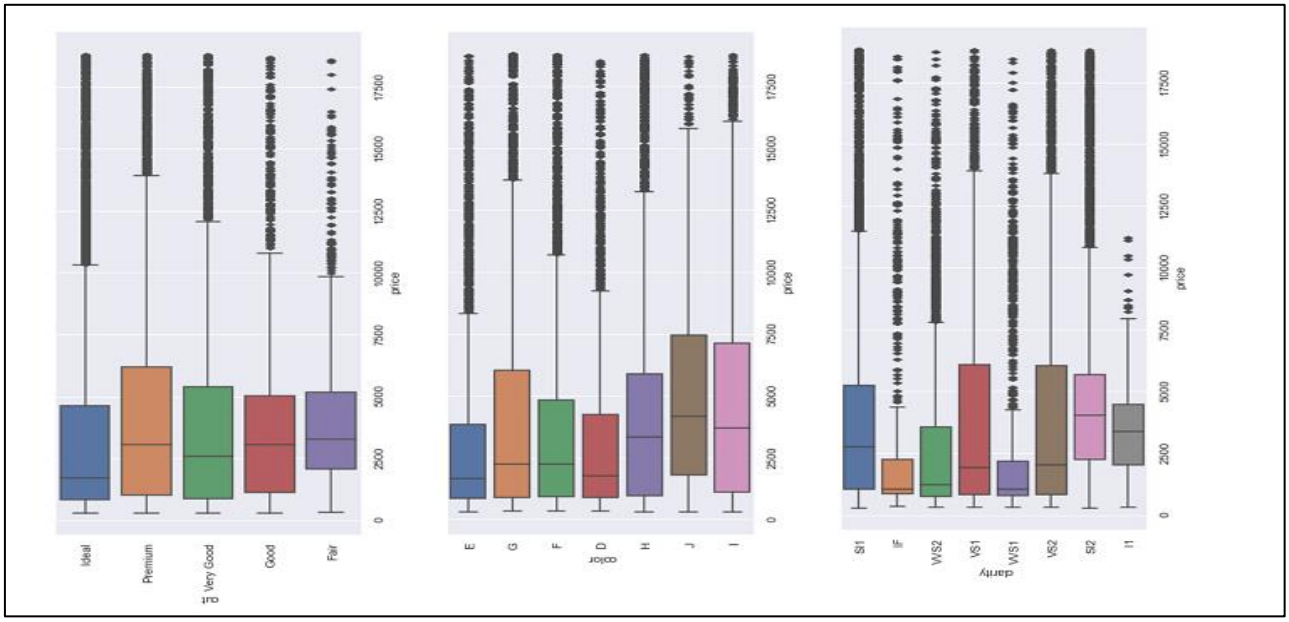
Fig 1.5: Bivariate Analysis: Pair Plots



Pair-plot Interpretation:

1. Multiple peaks in diagonal plots shows multiple clusters within each feature.
2. Depth and table do not show any linear relationship with the target variable - Price
3. Length, height and width shows strong Structural correlations.
4. Carat and Dimensions shows strong linear relationship with Price.

Fig 1.6: Bivariate Analysis: Boxplot for Discrete columns and Target variable:



The Median price for Cut- Ideal, Color- E, D, F, Clarity – IF, VVS1, VVS2 are low. But these attributes have lot of outliers. The Cz diamonds with these features though their average price are low they are in demand and can be sold at a higher price. In other words, diamonds with these features bring more profit.

Fig 1.7: Bivariate Analysis: Depth and Table:

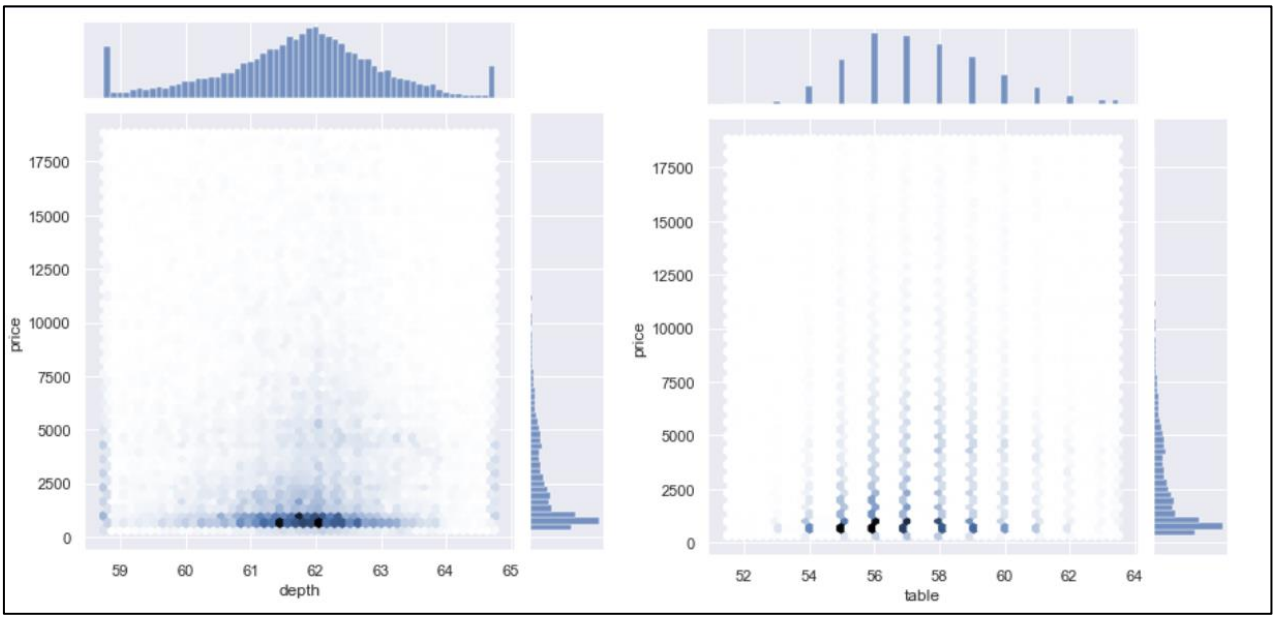
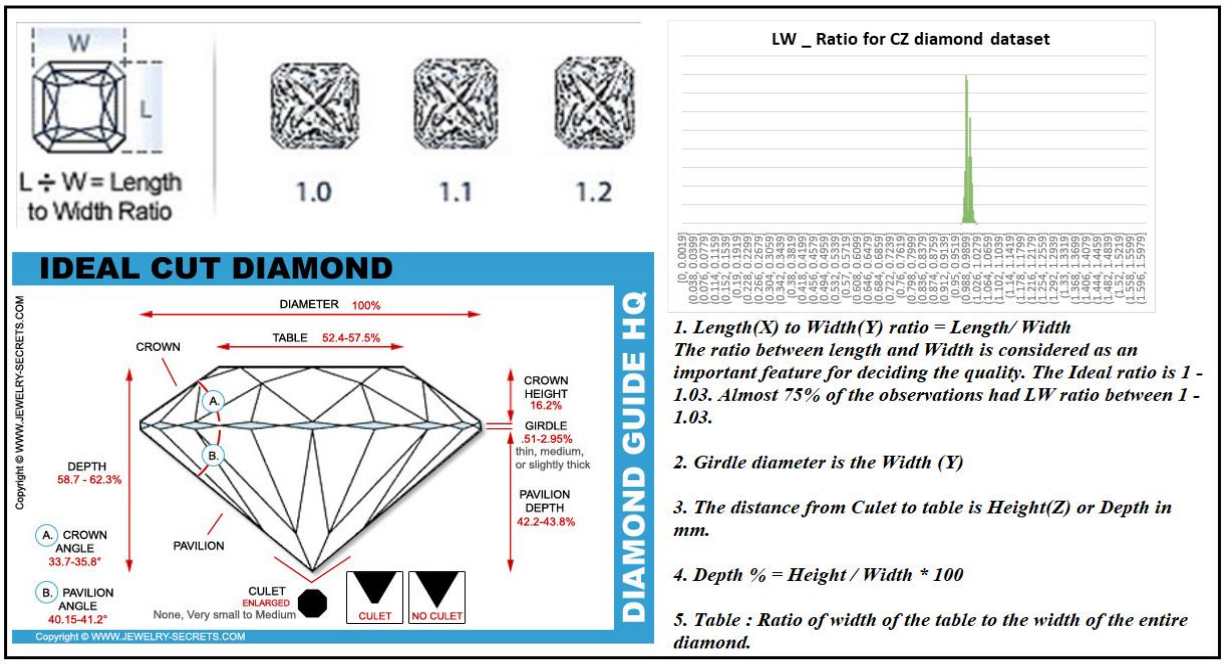


Fig 1.8.1 and 1.8.2: Understanding the Measurements of the CZ diamonds



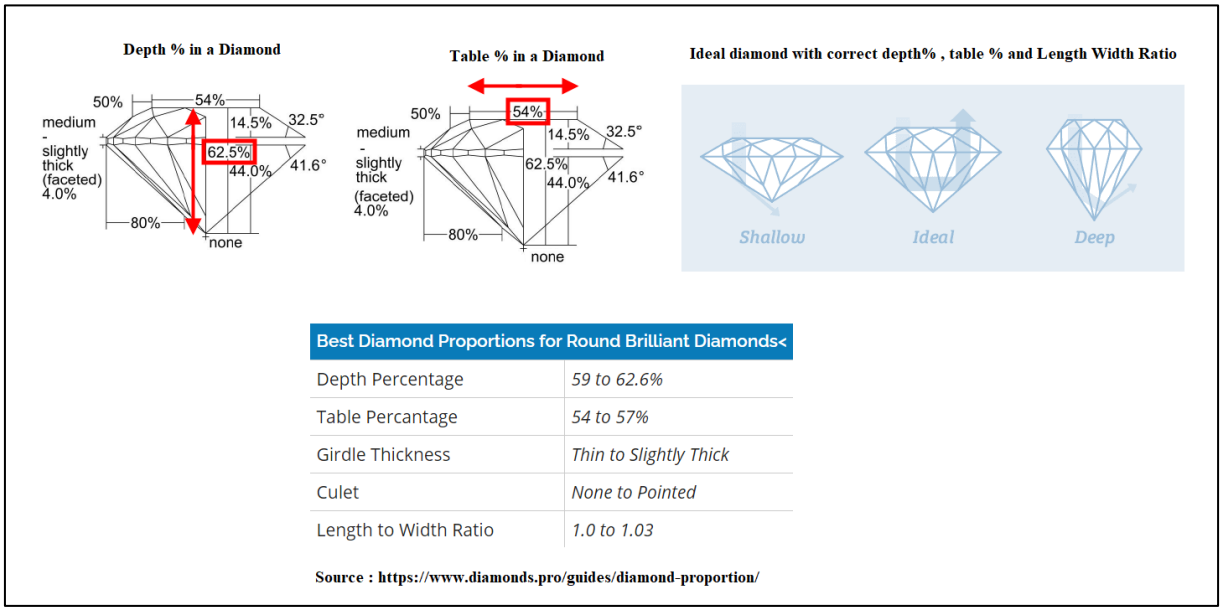
1. Length(X) to Width(Y) ratio = Length/ Width
The ratio between length and Width is considered as an important feature for deciding the quality. The Ideal ratio is 1 - 1.03. Almost 75% of the observations had LW ratio between 1 - 1.03.

2. Girdle diameter is the Width (Y)

3. The distance from Culet to table is Height(Z) or Depth in mm.

4. Depth % = Height / Width * 100

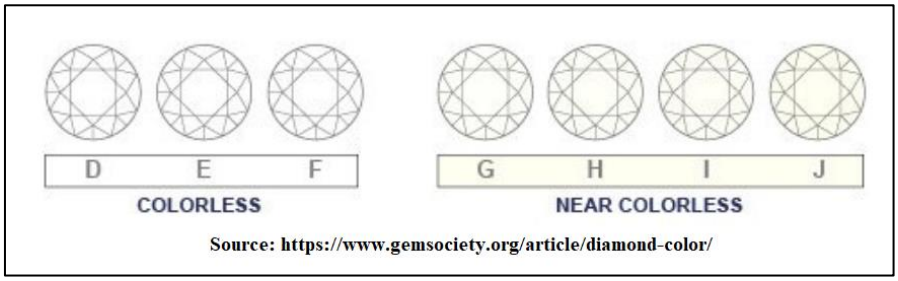
5. Table : Ratio of width of the table to the width of the entire diamond.



Best Diamond Proportions for Round Brilliant Diamonds<	
Depth Percentage	*59 to 62.6%*
Table Percantage	*54 to 57%*
Girdle Thickness	*Thin to Slightly Thick*
Culet	*None to Pointed*
Length to Width Ratio	*1.0 to 1.03*

Source : https://www.diamonds.pro/guides/diamond-proportion/

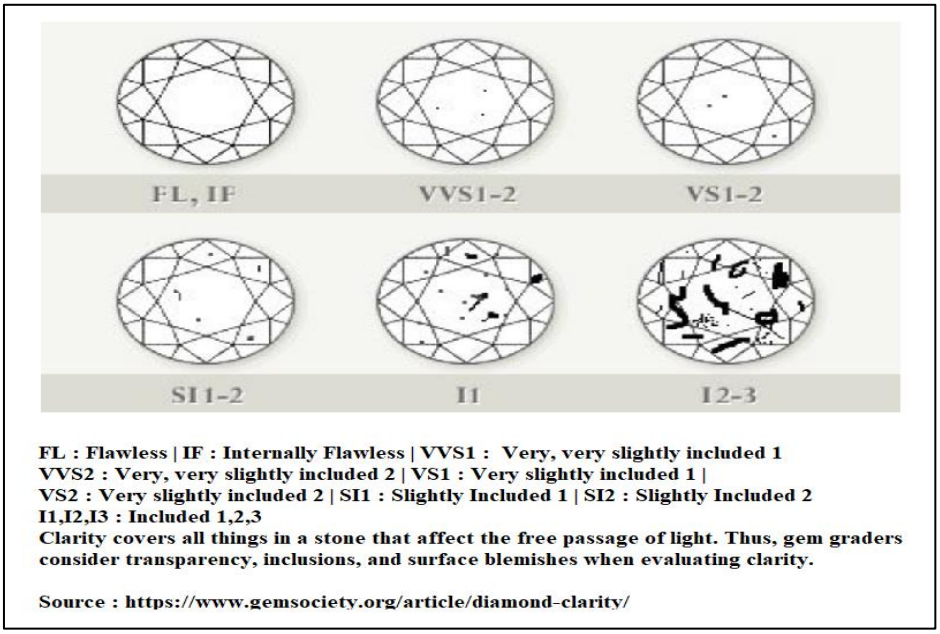
Fig 1.9: Diamond colours:



Source: https://www.gemsociety.org/article/diamond-color/

- The most highly valued diamonds have no color.

Fig 1.10: Diamond Clarity:



FL : Flawless | IF : Internally Flawless | VVS1 : Very, very slightly included 1
VVS2 : Very, very slightly included 2 | VS1 : Very slightly included 1 |
VS2 : Very slightly included 2 | SI1 : Slightly Included 1 | SI2 : Slightly Included 2
I1,I2,I3 : Included 1,2,3
Clarity covers all things in a stone that affect the free passage of light. Thus, gem graders consider transparency, inclusions, and surface blemishes when evaluating clarity.

Source : https://www.gemsociety.org/article/diamond-clarity/

1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Do you think scaling is necessary in this case?

There were 697 null values in the depth feature. Other all features had 0 null values. The null values in depth are imputed using (Height/Width *100) of that particular row. The data dictionary has mentioned depth as "The Height of a cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.".

```
df_cz['depth'].fillna((df_cz['z']/df_cz['y'])*100,inplace=True)
```

Why scaling?

Variance Inflation Factor quantifies the severity of multicollinearity in an ordinary least squares regression analysis.

Table 1.5: Variance Inflation Factors

	VIF Factor	features
3	10373.233401	length
4	9300.779064	width
5	2892.012582	height
1	942.782640	depth
2	717.617959	table
0	101.218699	carat

	VIF Factor	features
2	607.691497	length
1	520.343424	table
3	507.939945	depth
0	85.217709	carat

	VIF Factor	features
0	26.845034	carat
3	26.839207	length
1	1.170601	depth
2	1.143972	table

	VIF Factor	features
0	1.0	(CaratDimensions,)
2	1.0	(Depth,)
3	1.0	(Color,)
1	1.0	(Cut&Table,)
4	1.0	(Clarity,)

1. First data frame shows the VIF values on Unscaled data with all features. The VIF values are very high.
2. Second data frame shows the VIF values for the data without Height and width variables. The VIF values are still very high.
3. Third data frame shows the VIF values on scaled data. Scaling reduced the Multicollinearity. But the VIF values for carat and length are still high.
4. The last data frame shows the VIF values on PCA components. The Multi collinearity is totally removed.
5. The VIF values on scaled data and on PCA data are very low when compared to other values.
6. Also, the features in data are found in different scales. Carat size is in ‘mg’, dimensions are in ‘mm’, Depth and table are in ‘%s’ and the Price is in monetary value. In order to equal weightage to all the features, scaling is needed.
7. However, for this particular data, Models are performed on unscaled data despite its high multi collinearity issue in order to understand the Coefficients effectively.

Model Pre- processing:

Label Encoding:

Cut: Ideal – 1, Premium – 2, Very Good – 3, Good – 4, Fair – 5.

Clarity: From FL to I3 ➔ 1 to 11

Color: From D to J ➔ 1 to 5

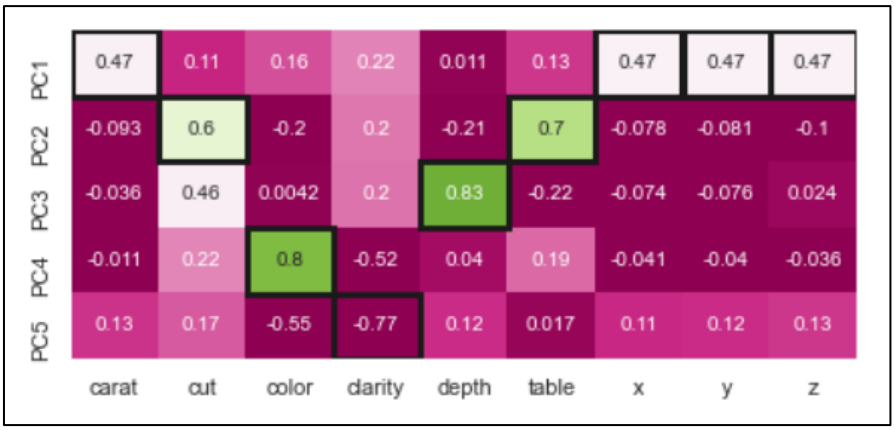
One hot Encoding:

One hot encoding is performed using get_dummies () function.

PCA Components:

Principle components analysis is performed using Sklearn decomposition library. N components = 5. The 5 components have explained 95% of the variances in the data. Below is the heatmap of the components explaining the variance across different features.

Fig 1.11: Heatmap showing the components and their captured variances



A new data frame using the components are created with new column names. CaratDimensions: Carat, x, y and z, CutTable: Cut and Table feature, Depth, Color and Clarity.

4 main Assumptions of Linear Regression model:

1. **Linear relationship**: b/w the independent variable, x, and the dependent variable, y.
2. **Independence**: The independent variables are actually independent and not collinear.
3. **Homoscedasticity and Autocorrelation**: The residuals have constant variance at every level of x.
4. **Normality**: The residuals of the model are normally distributed.

1.3 Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using Rsquare, RMSE.

- The independent and dependent variables are separated. (X and y)
- Using Train Test Split the 26886 observations are separated as Train – 18820 and Test – 8066.
- Random state is given. This ensures that the splits that you generate are reproducible

Linear Regression Models:

In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

4 different models are executed for the given data. As the data has multi-collinearity issue and outliers, PCA is used as a pre-processing step and model is executed on the PCA data with Label encoding. The other type is on the Unscaled – One hot encoded data. The types of the model that are executed are given below:

Model 1: Single Linear regression model on Unscaled data with Label encoding - Carat

Model 2: Multiple Linear regression model on data with PCA components with Label encoding.

Model 3: Multiple Linear regression model on Unscaled data with One hot encoding

Model 4: Multiple Linear regression model using Sklearn library. (Unscaled data with One hot Encoding)

Model 5: Multiple Linear regression model using Sklearn library. (On PCA Components)

Linear Regression Model 1: Single Linear Regression model on Carat feature.

The carat size has very strong linear relationship with Price. So, a single linear regression model with only Carat feature is executed.

Data: Unscaled data with Label Encoding

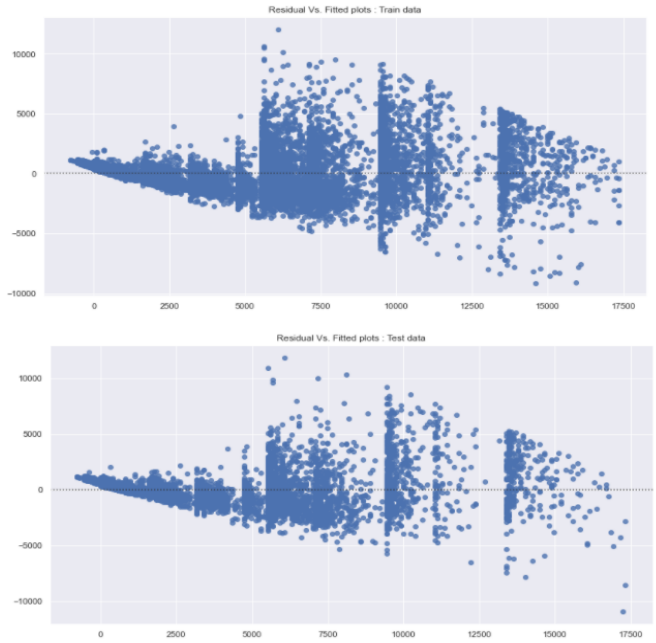
Fig 1.12: Model 1 Output and Residual plot:

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.853
Model:	OLS	**Adj. R-squared:**	0.853
Method:	Least Squares	**F-statistic:**	4.666e+04
Date:	Thu, 03 Jun 2021	**Prob (F-statistic):**	0.00
Time:	00:17:14	**Log-Likelihood:**	-70491.
No. Observations:	8066	**AIC:**	1.410e+05
Df Residuals:	8064	**BIC:**	1.410e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2343.6557	33.178	-70.638	0.000	-2408.694	-2278.618
carat	7866.4723	36.419	216.001	0.000	7795.082	7937.862

Omnibus:	2262.654	**Durbin-Watson:**	2.001
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	15847.712
Skew:	1.160	**Prob(JB):**	0.00
Kurtosis:	9.463	**Cond. No.**	3.69



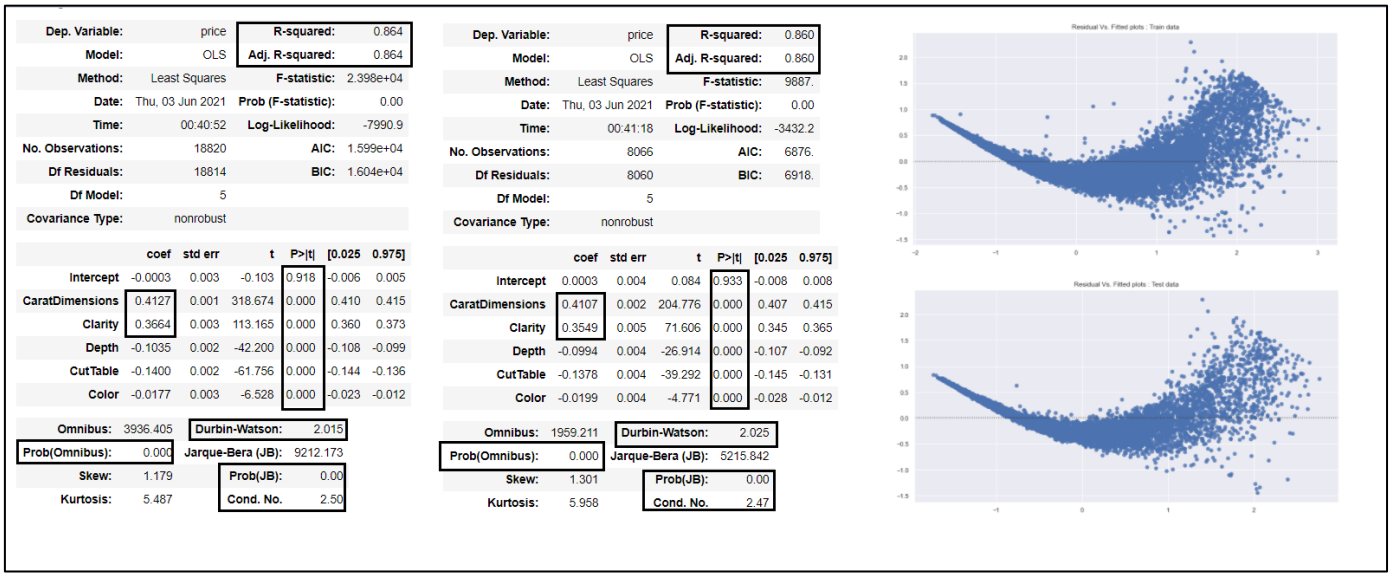
1. **Intercept p-value P>|t|:** p-value < 0.05. The negative intercept is not acceptable. This may be due to the outliers, multicollinearity and violation in one of the assumptions. When X is 0, Y cannot be in negative.
2. **coef/P>|t|:** p-value < 0.05. Carat has impact on Price. Increasing Carat by 1 unit will be associated with an average increase of β1(7847.38) in Y **(Assumption #1)**
3. **Prob (F-statistic):** The coefficient is not equal to zero. There is a linear relationship between carat and price variable. **(Assumption #1)**
4. **Prob (Omnibus)/Prob (JB):** p-value is < 0.05. The residuals form a normal distribution. **(Assumption #4)**. Skew is slightly high and Kurtosis shows high peak due to outliers.
5. **Durbin-Watson:** Value less than 2 is expected. Model value is very close to 2. The residuals have constant variance at every level of x. **(Assumption #3)**
6. **Cond. No.** Value less than 30. No multi collinearity. **(Assumption #2)**
7. **R-squared/ Adjusted R-squared:** The feature carat is able to explain 85% variance in the Price variable.

Linear Regression Model 2: Multiple Linear Regression model on 5 PCA Components.

Due to high multi collinearity the two out of three Dimension features take a negative coefficient. These dimensions are strongly and positively correlated to the Price variable. The Correlation coefficient and the regression slope sign should be same. Due to high multi collinearity the co-efficients are affected. Hence, the data is pre-processed using Principle components analysis. PCA reduces the reduces the Column dimensions and multi collinearity. The VIF factors also showed good results on the data with principal components.

Data: Scaled data | PCA Components | Label Encoding

Fig 1.13: Model 2 Output and Residual plots on Train and Test data



Dep. Variable:	price	R-squared:	0.864
Model:	OLS	Adj. R-squared:	0.864
Method:	Least Squares	F-statistic:	2.398e+04
Date:	Thu, 03 Jun 2021	Prob (F-statistic):	0.00
Time:	00:40:52	Log-Likelihood:	-7990.9
No. Observations:	18820	AIC:	1.599e+04
Df Residuals:	18814	BIC:	1.604e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0003	0.003	-0.103	0.918	-0.006	0.005
CaratDimensions	0.4127	0.001	318.674	0.000	0.410	0.415
Clarity	0.3664	0.003	113.165	0.000	0.360	0.373
Depth	-0.1035	0.002	-42.200	0.000	-0.108	-0.099
CutTable	-0.1400	0.002	-61.756	0.000	-0.144	-0.136
Color	-0.0177	0.003	-6.528	0.000	-0.023	-0.012

Omnibus:	3936.405	Durbin-Watson:	2.015
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9212.173
Skew:	1.179	Prob(JB):	0.00
Kurtosis:	5.487	Cond. No.	2.50

Dep. Variable:	price	R-squared:	0.860
Model:	OLS	Adj. R-squared:	0.860
Method:	Least Squares	F-statistic:	9887.
Date:	Thu, 03 Jun 2021	Prob (F-statistic):	0.00
Time:	00:41:18	Log-Likelihood:	-3432.2
No. Observations:	8066	AIC:	6876.
Df Residuals:	8060	BIC:	6918.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0003	0.004	0.084	0.933	-0.008	0.008
CaratDimensions	0.4107	0.002	204.776	0.000	0.407	0.415
Clarity	0.3549	0.005	71.606	0.000	0.345	0.365
Depth	-0.0994	0.004	-26.914	0.000	-0.107	-0.092
CutTable	-0.1378	0.004	-39.292	0.000	-0.145	-0.131
Color	-0.0199	0.004	-4.771	0.000	-0.028	-0.012

Omnibus:	1959.211	Durbin-Watson:	2.025
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5215.842
Skew:	1.301	Prob(JB):	0.00
Kurtosis:	5.958	Cond. No.	2.47

1. **Intercept p-value P>|t|:** p-value > 0.05. The intercept has no impact on dependent variable as the independent variables are scaled and the mean is zero. The value of Y @ X = 0". The intercept becomes zero when X is zero.
2. **coef/P>|t|:** p-value < 0.05. All features do have an impact on the Price **(Assumption #1)**
3. **Prob (F-statistic):** The coefficient is not equal to zero. There is a linear relationship between independent and price variable. **(Assumption #1)**
4. **Prob (Omnibus)/Prob (JB):** p-value is < 0.05. The residuals form a normal distribution. **(Assumption #4)**. Skew is slightly high and Kurtosis shows high peak due to outliers.
5. **Durbin-Watson:** Value less than 2. The residuals have constant variance at every level of x. **(Assumption #3)**
6. **Cond. No.** Value less than 30. No multi collinearity. **(Assumption #2)**
7. **R-squared/ Adjusted R-squared:** The given set of features were able to explain 86% variance in the Price variable.

Important Variables from Model 2: Carat, Dimensions (x, y and z) and Clarity

Linear Regression Model 3: Multiple Linear Regression Model on Unscaled data with One hot Encoding:

Train data size: 17866 and Test data size: 7658

Data: Un-scaled data | One hot Encoding for Discrete variables:

From the above models it is clear that Carat, Clarity and Dimensions have important role to play in predicting the price.

1. For easy interpretation Linear regression model on Unscaled data is performed though this model has high multi collinearity.
2. Model3 on full data had very high multi collinearity and the values were not up to the mark.
3. Hence, this model is further tuned by:

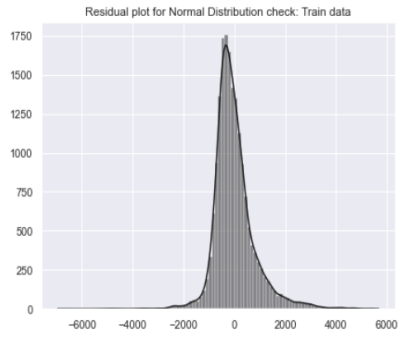
- Data has only observations where Price <= 13000. Why 13000? The 95% quantile is approximately 13000.
- The Attributes like Depth, VS1, E, F, Good and Very Good had p-values more than 0.05. Which shows that they do not have an impact on Price on Population data. Hence, they were removed one by one.

Fig 1.14: Model 3 Output and Residual plots on Train Data

Dep. Variable:	price	R-squared:	0.896
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	9642
Date:	Fri, 04 Jun 2021	Prob (F-statistic):	0.00
Time:	15:23:20	Log-Likelihood:	-1.4796e+05
No. Observations:	17866	AIC:	2.960e+05
Df Residuals:	17849	BIC:	2.961e+05
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	250.1510	155.856	1.605	0.109	-55.341	555.643
carat	9488.2854	103.570	91.612	0.000	9285.277	9691.293
x	-1768.1750	141.025	-12.538	0.000	-2044.598	-1491.752
y	1420.0584	136.098	10.434	0.000	1153.294	1686.823
z	-684.1005	85.681	-7.984	0.000	-852.043	-516.158
IF	1156.7117	41.444	27.910	0.000	1075.478	1237.946
VVS1	1022.8967	30.093	33.991	0.000	963.911	1081.883
VVS2	1083.1883	25.715	42.123	0.000	1032.785	1133.592
I1	-2619.2196	63.479	-41.261	0.000	-2743.644	-2494.795
D	351.6263	22.898	15.356	0.000	306.743	396.509
E	201.1335	19.731	10.194	0.000	162.459	239.808
Ideal	237.4449	17.424	13.627	0.000	203.292	271.598
Premium	113.9405	20.938	5.442	0.000	72.901	154.980
Fair	-597.8598	47.222	-12.661	0.000	-690.419	-505.300
J	-1266.6066	33.893	-37.370	0.000	-1333.041	-1200.172
VS2	471.3091	17.983	26.209	0.000	436.062	506.557
H	-279.4424	21.302	-13.118	0.000	-321.197	-237.688

Omnibus:	4625.866	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	22049.662
Skew:	1.180	Prob(JB):	0.00
Kurtosis:	7.904	Cond. No.	245.



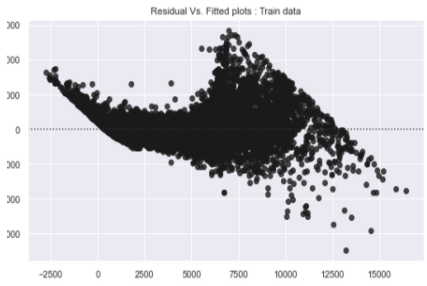
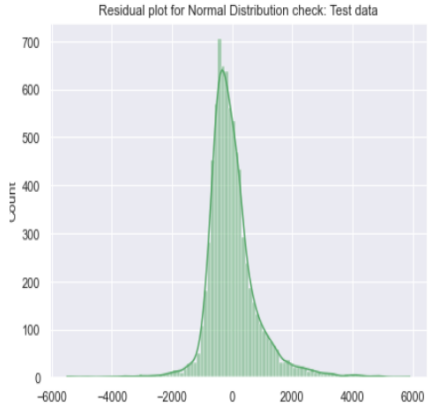


Fig 1.15: Model 3 Output and Residual plots on Test Data

Dep. Variable:	price	R-squared:	0.896
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	4109.
Date:	Fri, 04 Jun 2021	Prob (F-statistic):	0.00
Time:	15:25:49	Log-Likelihood:	-63455.
No. Observations:	7658	AIC:	1.269e+05
Df Residuals:	7641	BIC:	1.271e+05
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	20.2665	241.217	0.084	0.933	-452.586	493.119
carat	9322.0270	159.506	58.443	0.000	9009.352	9634.702
x	-1234.4887	195.418	-6.317	0.000	-1617.562	-851.415
y	1209.6956	186.774	6.477	0.000	843.566	1575.825
z	-1106.0407	145.927	-7.579	0.000	-1392.098	-819.983
IF	1245.3851	63.200	19.705	0.000	1121.496	1369.275
VVS1	1014.0052	44.926	22.571	0.000	925.939	1102.072
VVS2	1139.6994	39.236	29.047	0.000	1062.786	1216.613
I1	-2870.8759	96.517	-29.745	0.000	-3060.075	-2681.677
D	337.7813	34.965	9.660	0.000	269.239	406.323
E	219.2306	30.314	7.232	0.000	159.807	278.654
Ideal	210.0020	26.779	7.842	0.000	157.508	262.496
Premium	104.4329	32.040	3.259	0.001	41.626	167.240
Fair	-419.5385	71.097	-5.901	0.000	-558.909	-280.168
J	-1251.8450	52.295	-23.938	0.000	-1354.357	-1149.333
VS2	479.3093	28.017	17.108	0.000	424.388	534.230
H	-284.2763	32.336	-8.791	0.000	-347.665	-220.888

Omnibus:	1968.194	Durbin-Watson:	2.006
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9172.189
Skew:	1.173	Prob(JB):	0.00
Kurtosis:	7.821	Cond. No.	240.



1. **Intercept p-value P>|t|:** p-value < 0.05. When X = 0, the intercept for test data is 20 Which is means when X= 0, the Price (y) = 20.
2. **coef/P>|t|:** p-value < 0.05. All features in the model do have an impact on the Price **(Assumption #1)**
3. **Prob (F-statistic):** The coefficient is not equal to zero. There is a linear relationship between independent and price variable. **(Assumption #1)**
4. **Prob (Omnibus)/Prob (JB):** p-value is < 0.05. The residuals form a normal distribution. **(Assumption #4)**. Skew is slightly high and Kurtosis shows high peak due to outliers.
5. **Durbin-Watson:** Value very near to 2. The residuals have constant variance at every level of x. **(Assumption #3)**
6. **Cond. No.** Multi collinearity is reduced by removing Depth and other variables where p-value was more than 0.05 **(Assumption #2)**
7. **R-squared/ Adjusted R-squared:** The given set of features were able to explain 90% variance in the Price variable.

Important Variables from Model 3: Carat, Dimensions – y (width), Clarity - IF, VVS1, VVS2, VS2, Color - D, E, Cut - Ideal and Premium

Note: The co-efficients of dimensions Length and height is shown negative because of multi collinearity.

Linear Regression Model 4: Linear Regression model using Sklearn library:

In order to compare the coefficients from Stats Model 3 (Unscaled One hot encoded data) with Sklearn coefficients, same data is used for Sklearn library model.

Data: Un-scaled data | One hot Encoding for Discrete variables:

Table 1.6: Coefficients from Sklearn Linear Model 4:

	carat	y	IF	VVS1	VVS2	D	VS1	E	F	VS2	G	Ideal	Premium	Very Good	depth
0	9984.0	1112.0	1108.0	980.0	977.0	665.0	624.0	478.0	442.0	333.0	284.0	205.0	177.0	99.0	6.0

Important Variables from Model 4: Carat, Dimensions – y (width), Clarity - IF, VVS1, VVS2, Color - D, E, Cut - Ideal and Premium

Table 1.7: Model performance metrics (Unscaled data):

	Train	Test
R-Squared	92.38	92.19
Intercept	993.18	
Mean Squared Error	692012.21	671576.48
Root Mean Squared Error	831.87	819.50
Mean Absolute Error	566.64	566.13

The given set of independent variables were able to explain 92% variance in the dependent variable.

The RMSE and MAE is low for unscaled data.

Mean Absolute Error (MAE), like the RMSE, the MAE measures the prediction error. It is the average absolute difference between observed and predicted outcomes, MAE = mean (abs (observed - predicted)). MAE is less sensitive to outliers compared to RMSE.

Linear Regression Model 5: Multiple Linear regression model using Sklearn library. (On PCA Components)

Data: Scaled data | PCA components:

Table 1.8: Coefficients from Sklearn Linear Model 5

	CaratDimensions	Clarity	Color	Depth	Cut&Table
0	0.4127	0.3664	-0.0177	-0.1035	-0.14

Table 1.9: Model performance metrics (Unscaled data):

	Train	Test
R-Squared %	86.4366	85.9667
Intercept	-0.0003	
Mean Squared Error	0.1373	0.1369
Root Mean Squared Error	0.3705	0.3700
Mean Absolute Error	0.2827	0.2830

The given set of independent variables were able to explain 85 % variance in the dependent variable.

The RMSE and MAE is low for scaled PCA components data.

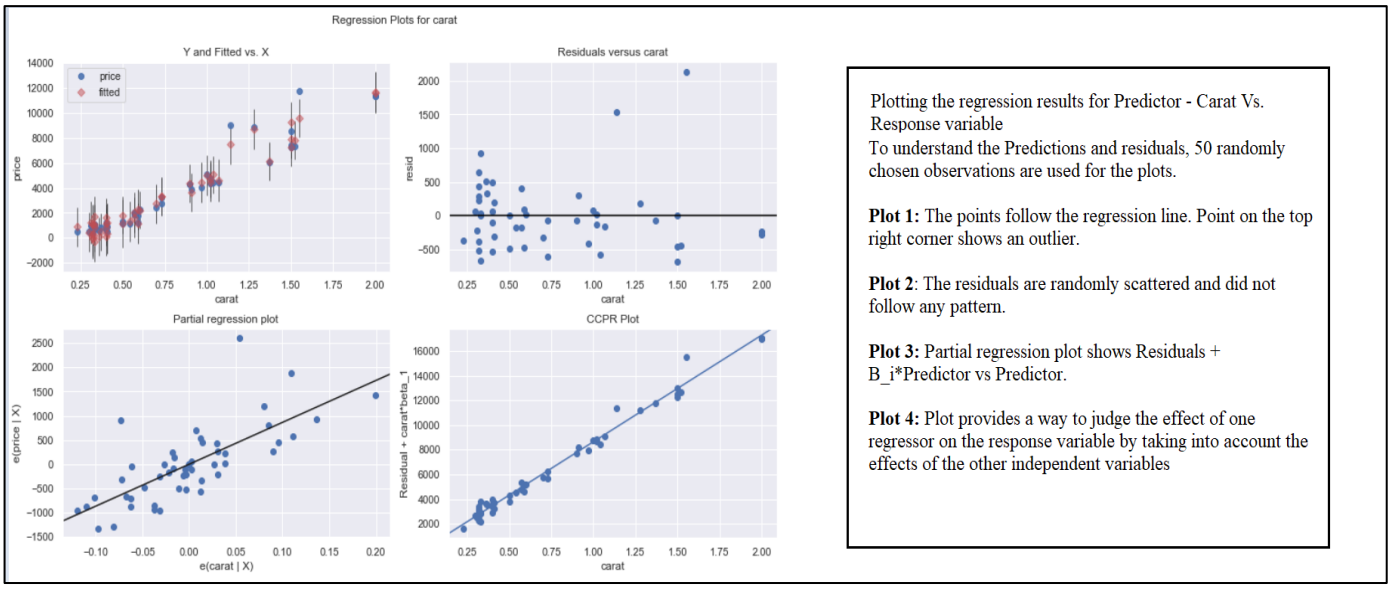
Model Comparison:

Linear Model on PCA: As the data is scaled and multi collinearity is removed, the model is able to explain the variances in length, width and height together. The mentioned three dimensions play a vital role in price determination along with Carat size. PCA model has given high priority to Carat, Dimensions (x, y and z) and Clarity.

Linear Model on Unscaled One hot encoded data: For easy interpretation this model is performed. And yes, it has given a similar insight (check the Fig 1.14 and 1.15 for attributes importance). But it did not perform well on explaining the Dimensions (x, y and z). Due to high multi collinearity the co efficients are impacted among the three variables and they are assigned negative regression slope to any two variables among them. The correlation coefficient and regression slope signs should be same.

Linear Model using Sklearn library: The coefficients and R-squared from this model is similar to the stats model (Model 3). The Errors metrics are given in Table: 1.7 and 1.9

Fig 1.16: Regression result plot: (Predictor – Carat Vs. Price)



Plotting the regression results for Predictor - Carat Vs. Response variable
To understand the Predictions and residuals, 50 randomly chosen observations are used for the plots.

Plot 1: The points follow the regression line. Point on the top right corner shows an outlier.

Plot 2: The residuals are randomly scattered and did not follow any pattern.

Plot 3: Partial regression plot shows Residuals + B_i*Predictor vs Predictor.

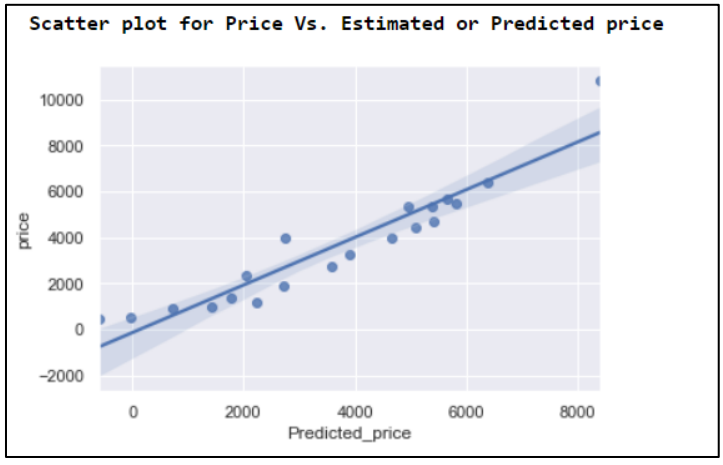
Plot 4: Plot provides a way to judge the effect of one regressor on the response variable by taking into account the effects of the other independent variables

Table 1.10: Overview of a Data frame with Predictions and residuals.

The predictions and the residuals are concatenated to the test data, which will help for further analysis.

	carat	depth	table	x	y	z	I1	IF	SI1	SI2	...	I	J	Fair	Good	Ideal	Premium	Very Good	price	Predicted_price	Residuals
22361	1.10	62.7	58.0	6.58	6.54	4.11	0	0	1	0	...	1	0	0	0	0	1	0	4435	5079.103311	-644.103311
7526	1.03	61.6	61.0	6.51	6.44	3.99	0	0	0	0	...	0	0	0	0	0	1	0	5364	5378.075094	-14.075094
8069	0.40	61.4	57.0	4.72	4.79	2.92	0	0	0	0	...	0	0	0	0	1	0	0	926	718.364511	207.635489
2629	1.08	59.4	59.0	6.68	6.75	3.99	0	0	0	0	...	0	0	0	0	0	1	0	6415	6380.818641	34.181359
9426	0.71	62.5	60.0	5.61	5.65	3.52	0	0	0	0	...	0	0	0	0	0	0	1	2745	3588.842835	-843.842835

Fig 1.17: Scatterplot for Price vs. Estimated price.



1.4 Inference: Basis on these predictions, what are the business insights and recommendations.

Best 5 attributes of for Price Prediction:

- *Carat ➔ Increasing Carat by 1 gm will be associated with an average increase of β1(9936.01) in y.
- The carat size for the given data is < 4.00 this falls under the Commercial size category.
- Measurements (Length, *Width and Height) ➔ Model 2
- Clarity → *IF, *VVS2, *VVS1, VS2 and SI1 shows Positive impact on price ➔Model 3 & 4
- Color → *D and E shows Positive impact on price ➔Model 3& 4
- Cut → Ideal and Premium shows Positive impact on price ➔Model 3 & 4

Note: The attributes with '*' are the top 5 attributes.

Features like Color -I, J, Clarity – I1, SI2, Cut – Fair have negative impact on predicting the Price.
The Gem Stones co ltd., can earn more profit on CZ diamonds of carat size less than 1.0 (Average carat size of the data without outliers)
The Gem Stones co ltd., can quote the prices of CZ diamonds according to the mentioned attributes. In the given data the diamond features like Ideal, Premium, D, E, IF had very low Median price but huge number of outliers were found within each feature, which means the prices can be quoted high for the diamonds with the given attributes (Check Fig 1.6). The model result also expresses same attributes. Fig 1.14 and 1.15

Fig 1.18: Model Linear Equation from Model 2(scaled data):

$$Y(Price) = Intercept + \beta1*CaratDimensions(X1) + \beta2*Clarity(X2) + \beta3*Depth(X3) + \beta4*CutTable(X4) + \beta5*Color(X5)$$

$$Y(Price) = 0.0003 + 0.4107 * X1 + 0.3549 * X2 - 0.0994 * X3 - 0.1378 * X4 - 0.0199 * X5$$

Conclusion:

Cubic zirconia is a man-made gemstone. It's a crystalline form of zirconium dioxide that's hard and colorless. The given data set is analysed and pre-processed for the Linear regression model. Four models executed and compared co-efficients from the model. The 4C's of CZ diamond quality i.e., Carat, measurements of the diamond, Clarity-IF, VVS1, VVS2, Cut – Ideal, Premium and Color-D, E played a major role in predicting the price.

Problem 2: Logistic Regression and LDA

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 Dictionary:

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

Problem statement:

Using Logistic regression and Linear Discriminant analysis with given set of data we have to find whether an employee will prefer a holiday package that is being sold by a tour and travel agency company.

2.1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA). Perform Univariate and Bivariate Analysis.

Data Description:

1. The given data set has 872 employees' details and 8 different features.
2. No duplicates in the data set with and without the Unique ID.
3. No null values in the given data set.
4. There are two continuous features, 3 Numeric categorical features and 2 discrete features.
5. Numerical data summary:

Table 2.1: Numerical features Summary

	count	mean	std	min	25%	50%	75%	max
Salary	872.0	47729.172018	23418.668531	1322.0	35324.0	41903.5	53469.5	236961.0
age	872.0	39.955275	10.551675	20.0	32.0	39.0	48.0	62.0
educ	872.0	9.307339	3.036259	1.0	8.0	9.0	12.0	21.0
no_young_children	872.0	0.311927	0.612870	0.0	0.0	0.0	0.0	3.0
no_older_children	872.0	0.982798	1.086786	0.0	0.0	1.0	2.0	6.0

- The difference between Mean salary and the Maximum salary is very huge. It clearly shows the presence of outliers.
- The mean and median age is very close and the difference between mean age and maximum age is less. The age feature forms normal distribution.
- The company has very vast diversity of employees. The number of years of education ranges from 1 to 21.
- Employees with zero kids and employees with more than 5 kids are also observed.

6. Discrete Data summary:

Table 2.2: Discrete features Summary

	count	unique	top	freq
Holliday_Package	872	2	no	471
foreign	872	2	no	656

- The highest frequency in Target variable is the Class 'No'. 54% of the employees have not preferred the holiday package.
- 75% of the employees are not Foreigners.

- The Salary of one the employee is very low, it is 1322. This seems to be a wrong entry.
- The feature names are renamed using the rename function.

```
df.rename(columns={'age' :'Age', 'educ': 'No_years_Education',
                   'no_young_children' :'No_Young_kids','no_older_children' :'No_Elder_kids',
                   'Holliday_Package':'Holiday_package','foreign' : 'Foreigner'},inplace=True)
```

Outliers Treatment:

- Age do not have any outliers.
- Model performed well with outliers than without outliers. Hence outliers are not treated.
- Anomaly in the Salary feature is treated while treating outliers in Salary (Refer to Data description last point)

Fig 2.1: Crosstabs for Number of years of Education and Number of Older kids:

No_years_Education	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	21
Holiday_package																				
no	0	3	2	21	27	10	15	90	59	55	62	79	18	14	7	4	2	0	2	1
yes	1	3	9	29	40	11	16	67	55	35	38	45	25	11	8	6	1	1	0	0

No_Elder_kids	0	1	2	3	4	5	6
Holiday_package							
no	231	102	102	27	7	2	0
yes	162	96	106	28	7	0	2

- The frequencies for Number of years Education less than 4 and more than 17 are very low. Hence the data points below and above the years 4 and 17 are clubbed together.
- The frequencies for Number of Elder kids above 5 are very low. Hence, observations with more than 5 children are clubbed together.

Fig 2.2: Crosstabs for Number of years of Education and Number of Older after binning:

No_years_Education	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Holiday_package														
0	26	27	10	15	90	59	55	62	79	18	14	7	4	5
1	42	40	11	16	67	55	35	38	45	25	11	8	6	2

No_Elder_kids	0	1	2	3	4	5
Holiday_package						
0	231	102	102	27	7	2
1	162	96	106	28	7	2

Exploratory Data Analysis and Data Visualization:

Table 2.3: Coefficient of Variance (CV%):

The coefficient of variation is a statistical measure of the dispersion of data points around the mean.			
Feature	**Mean**	**Standard Deviation**	**CV%**
Salary	47729.17	23418.67	49.00
Age	39.96	10.55	26.00

- The CV% for the Salary is high. This expresses that the data points are highly dispersed from the mean. Feature salary is unstable.

Table 2.4: Skewness and Kurtosis:

Feature	Skewness	Skew Type	Kurtosis	Kurtosis Type
Salary	3.1235	> 1 | Extremely right skewed	15.9501	> 3 | Leptokurtic - High Peak
Age	0.1464	(-0.5 to +0.5) | Approximately symmetrical	-0.91	< 3 | Platykurtic - Low Peak

Data Visualization:

Fig 2.3: Univariate Analysis: Numerical Column:

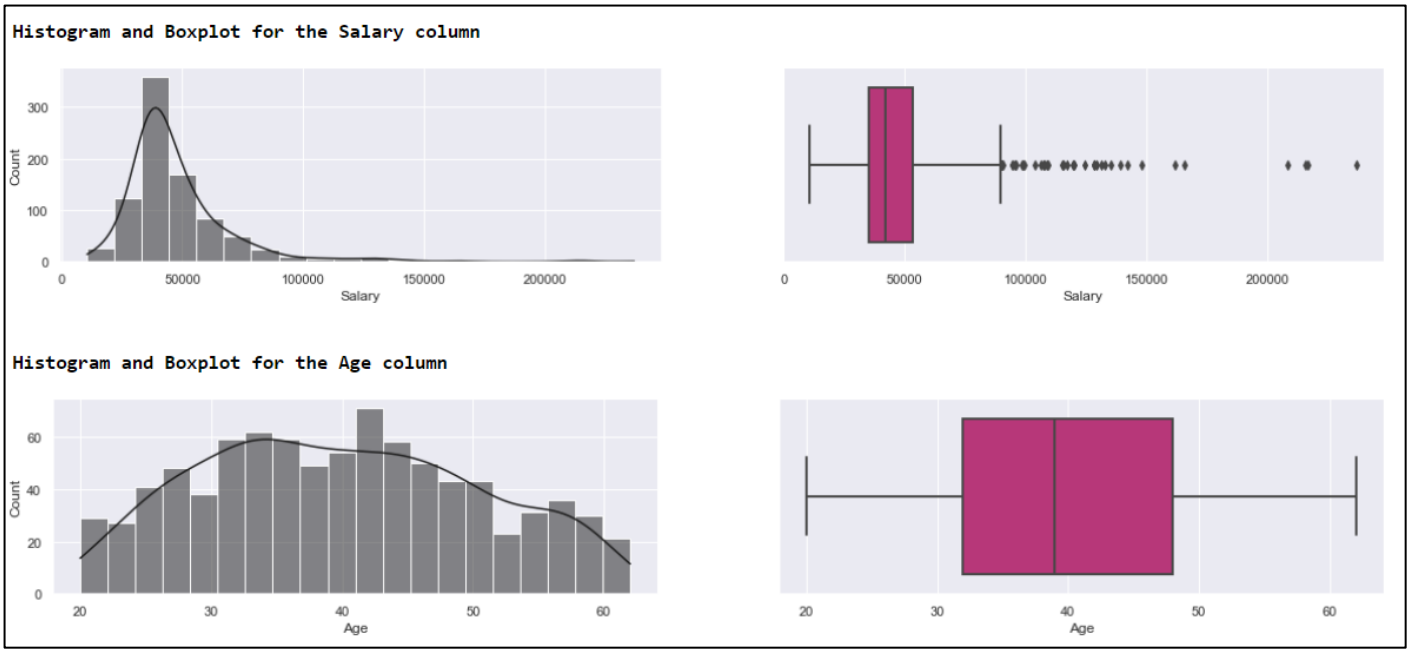
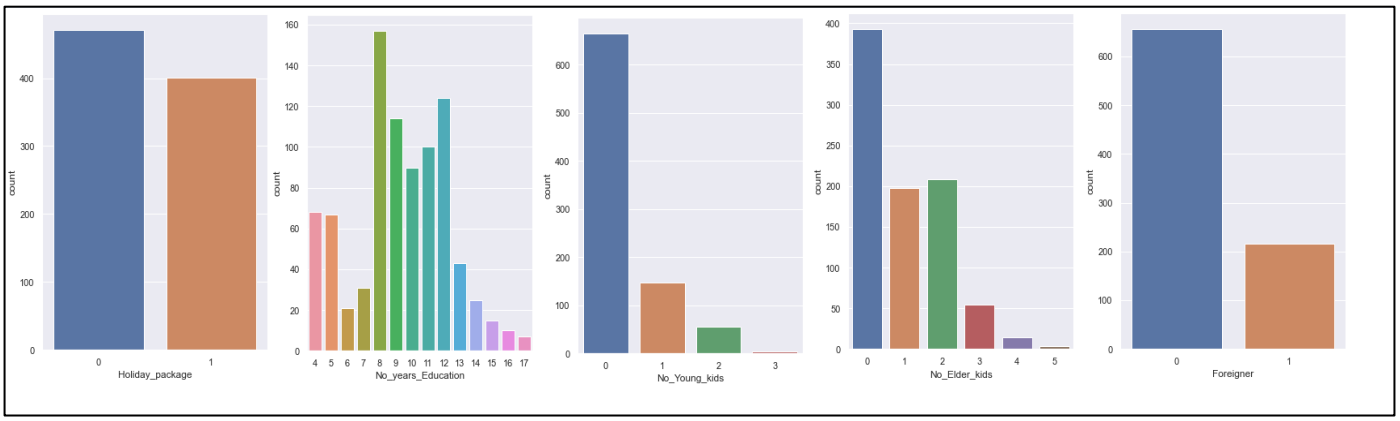


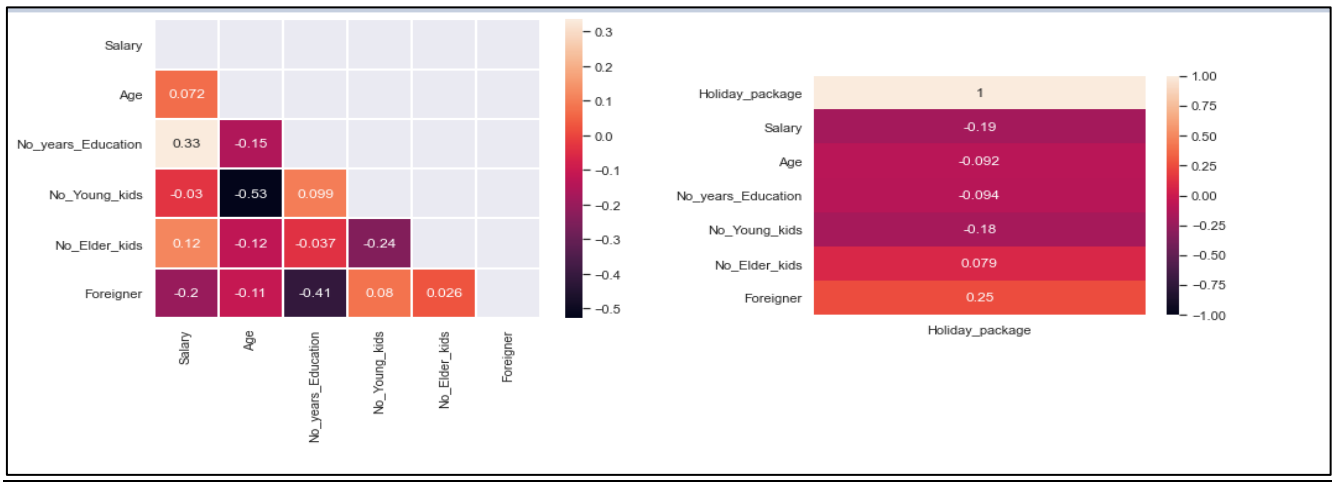
Fig 2.4: Univariate Analysis: Non continuous and Discrete column:



Univariate Analysis summary:

1. The highest peak in the histogram of Salary clearly shows that most of the samples are clustered around 47000 to 50000 salary range. Extremely Right skewed distribution. The boxplot shows no outlier in the lower boundary - Anomaly treated.
2. The kurtosis and skew are very low for Age, denoting that there are no outliers and the samples are almost evenly distributed.
3. The preference for Holiday package which is out target variable has 54:45 ratio. i.e., No = 54%, Yes = 45%.
4. More employees have 8 years of education and the next highest groups are 5,9 and 12 years of education.
5. Almost 31% of employees do not have Children.
6. Only 24% of the Employees are foreigners.

Fig 2.5: Bivariate Analysis: Correlation plot:



Heatmap interpretation:

1. The first plot shows the relationship between the independent variables. No high multi collinearity found in the data. For Logistic regression data with high multi collinearity is not preferred. Hence, the data is suitable for the model.
2. When r is between -0.4 to +0.4 there does not exist a linear relationship between the Predictor and response variable. The second heatmap plot shows that none of the features are highly linearly related to the target.
3. The Foreign feature has weak positive correlation and No. of young kids and Salary has very weak negative correlation.
4. Therefore, with the given data set and their correlations let's build the Logistic and LDA model and predict the target.

Fig 2.6: Bivariate Analysis: Pair plot:

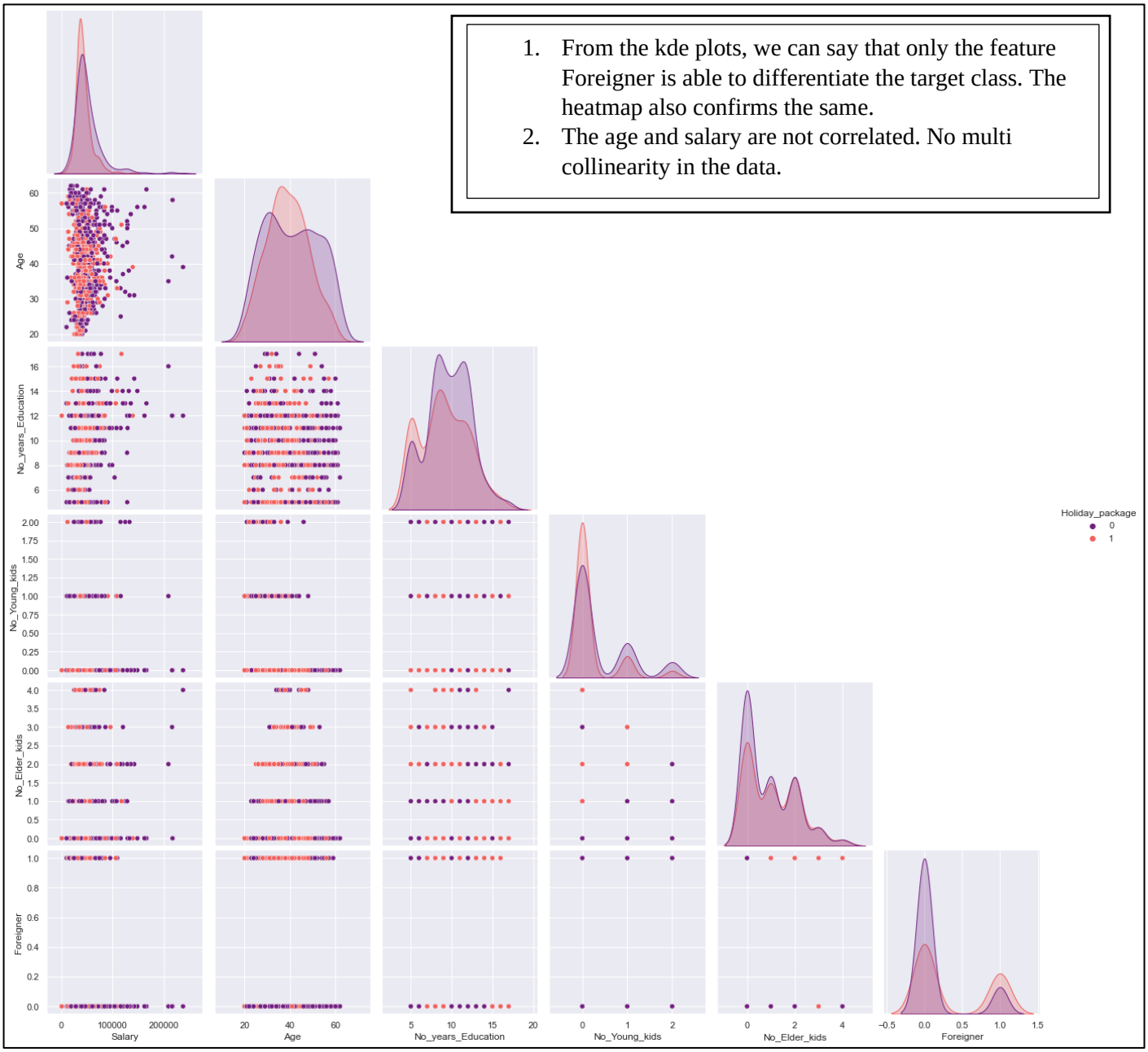
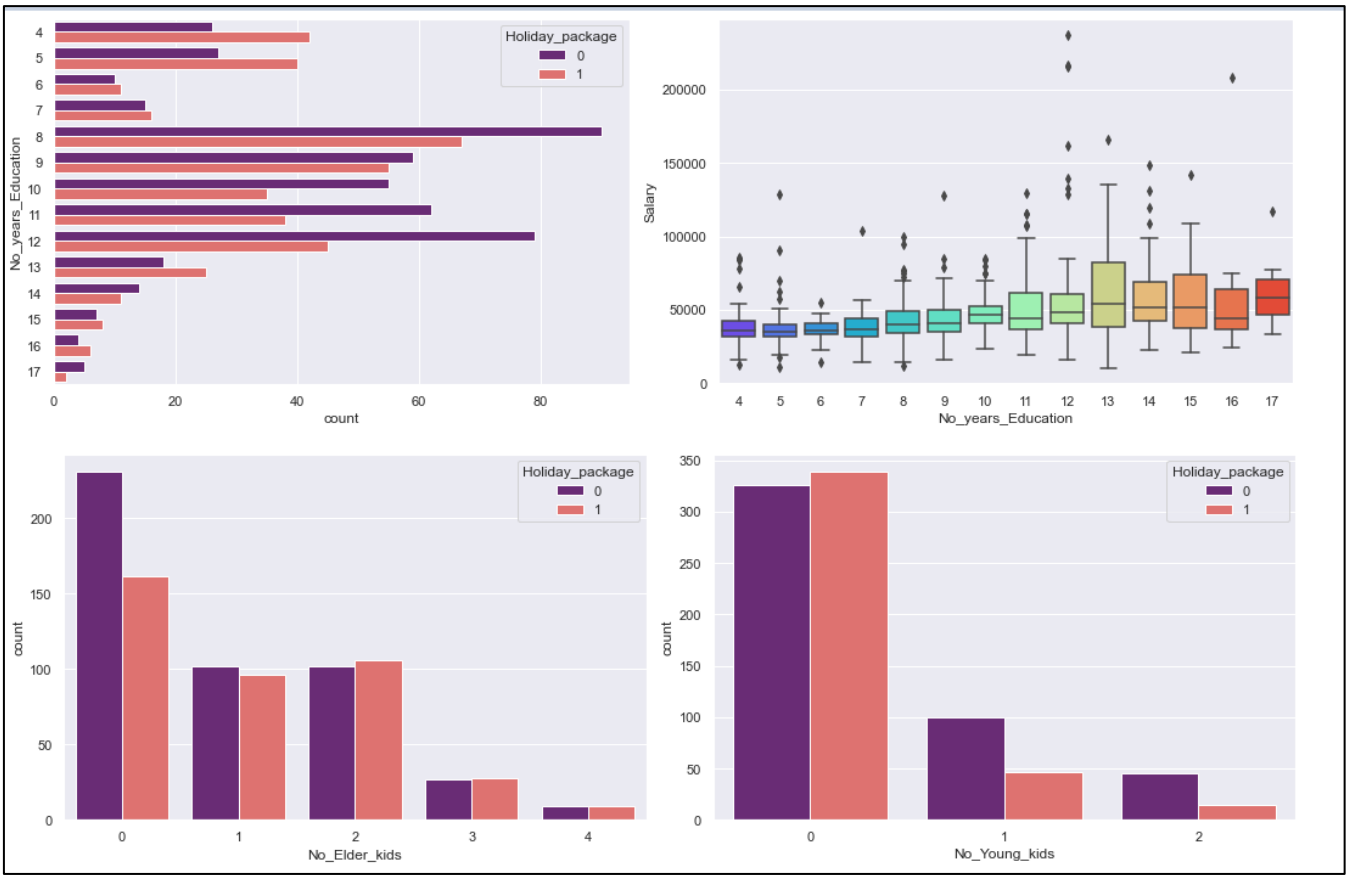
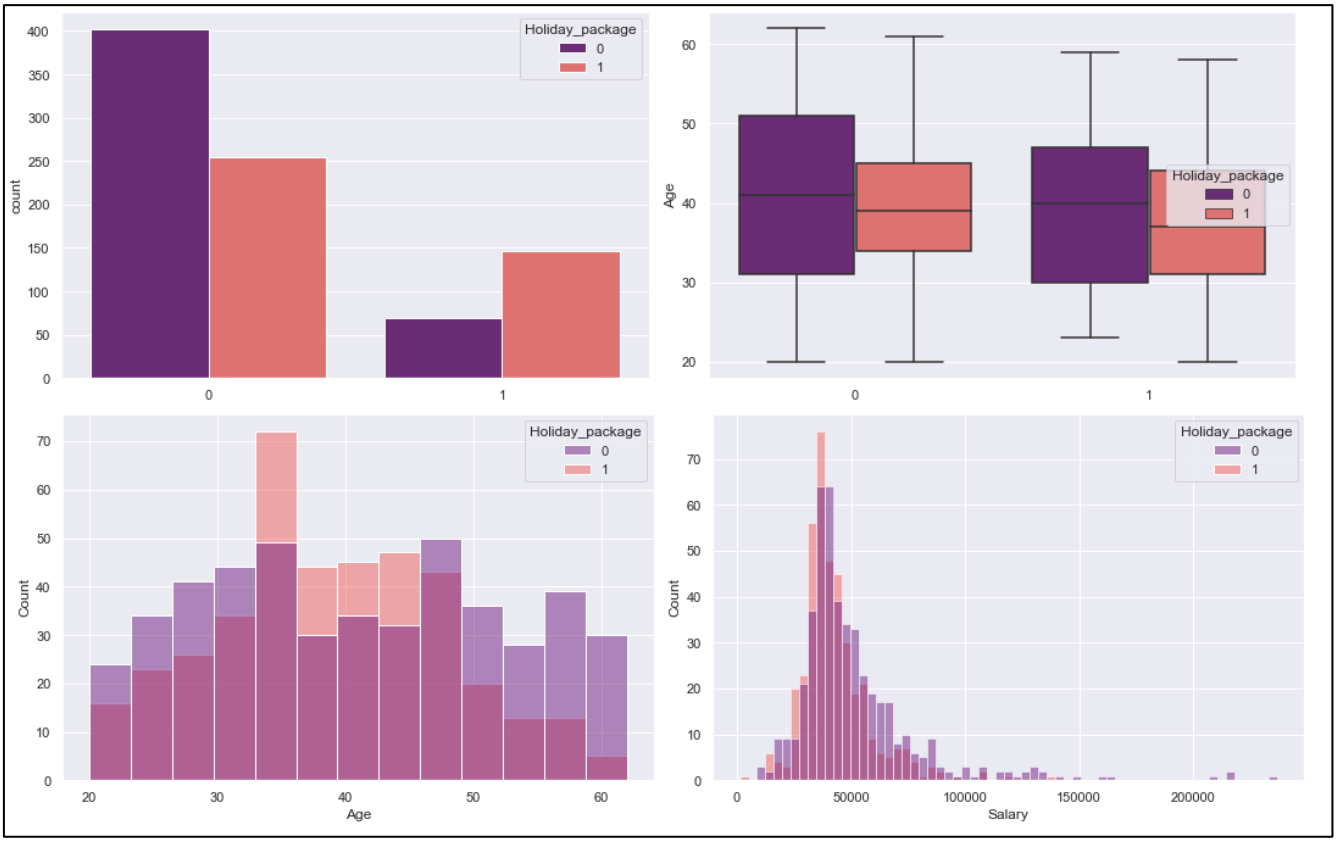


Fig 2.7: Bivariate Analysis: Independent variables vs Target variable as hue





Bivariate analysis summary:

1. The Salary of employees is not strongly correlated with the number of years of education. As the number of years of education increases the outliers in the Salary also increases. This clearly shows that there is mixed level of employees in the data.
2. Employees with up to *4 and 5 years* of education shows more interest for holiday package than other groups. Almost half of the employees from the group *9,13,15 and 16 years of education* choose the holiday package.
3. No significant relationship between the target and employees with elder kids. But a very weak negative relationship with No of Young kids is noticed. *Employees with 0 young kids prefer holiday package.*
4. The only positively correlated variable is Foreigner. *Almost 68% of the employees who are foreigners prefers the holiday tour package* and their mean age is 37 years.
5. *The Median age for 'Yes' category is less than the other group.*
6. *The Median salary for the employees who prefer package is 39,809* and the other category has a median salary of 43,940.
7. Most of the outliers i.e., *employees earning more than 1,00,000 do not prefer for a Holiday package.*

2.2 Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis).

Logistic Regression Assumptions:

1. The target variable should be categorical. The Holiday Package variable which is the target is categorical.
2. No collinearity is expected. But there is Multi collinearity in the data but not very high. The VIF is the measures the multi collinearity in the independent variables.

Table 2.5: Variance Inflation Factor:

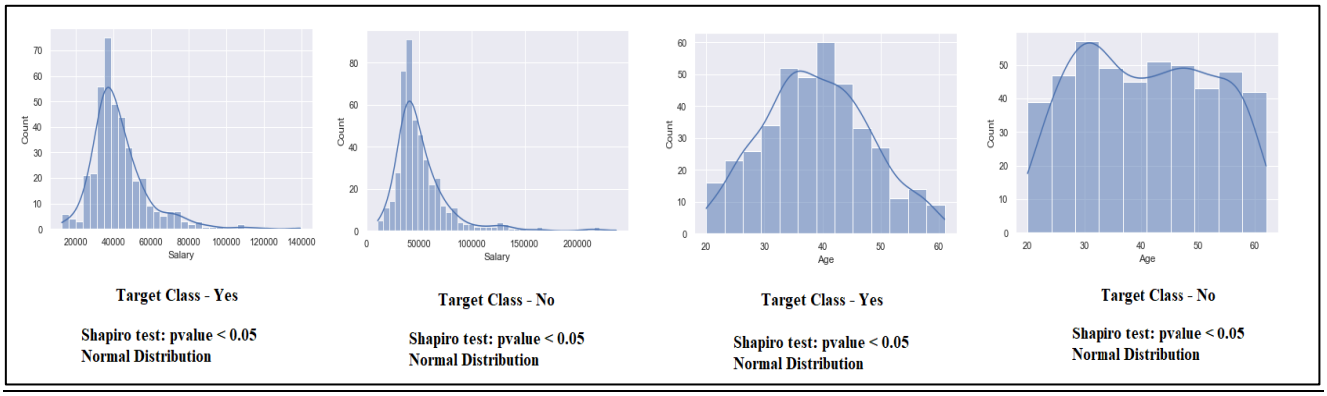
	VIF Factor	features
0	9.830872	No_years_Education
5	8.067121	Age
4	6.098640	Salary
2	1.893281	No_Elder_kids
1	1.543474	No_Young_kids
3	1.426691	Foreigner

3. The class of interest is 1 and the other class is coded as 0.

Linear Discriminant Analysis (LDA) Assumptions:

1. The independent variables are normally distributed for each class. Below figure shows that continuous independent variables follow a normal distribution for both classes.

Fig 2.8: Normality Assumption check:



2. There is no equal variance between the independent variables.

Fig 2.9: Equal variance Assumption check:

Salary	1.872645e+08	Salary	2.825694e+08
Age	8.193002e+01	Age	1.348445e+02
No_years_Education	9.742344e+00	No_years_Education	7.607544e+00
No_Young_kids	2.535910e-01	No_Young_kids	4.593938e-01
No_Elder_kids	1.182257e+00	No_Elder_kids	1.130867e+00
Foreigner	2.327805e-01	Foreigner	1.253015e-01
Target Class - Yes		**Target Class - No**	

3. LDA works well even when the assumptions are violated. (Duda, et al., 2001)'

Model Pre-Processing:

1. The Predictors and response variables are separated as X and y.
2. The Test split ratio is 70:30 | Train – 70% and Test 30%
3. 872 observations are split into 610 and 262 for Train and Test.
4. Random state: This ensures that the splits that you generate are reproducible

Models executed:

1. Model 1: Logistic Regression Model using Sklearn library
2. Model 2: Linear Discriminant Analysis using Sklearn library
3. Model 3: Logistic Regression using Stats model library

Model 1: Logistic Regression Model using Sklearn library

Logistic regression is probably the most important supervised learning classification method. It estimates relationship between a dependent variable (target) and one or more independent variable (predictors) where dependent variable is categorical/nominal. Using Grid search cv Logit model is executed.

Fig 2.10: Parameters given in Grid search CV:

```
Params = {'penalty':['l2','none'],
      'solver':['newton-cg','liblinear'],
         'tol':[0.0001,0.00001]}
```

- Newton-cg and Liblinear – for small data sets with less dimensions.
- Tolerance – smaller the value better the accuracy is.
- Penalty: should be given according to the solver type. Both the given solver type supports l2 penalty.

Fig 2.11: Best Parameters:

```
Grid_search.best_params_
```

```
{'penalty': 'none', 'solver': 'newton-cg', 'tol': 0.0001}
```

Model 2: LDA using Sklearn library (Fig:2.12)

```
#Build LDA Model and fit the data
clf = LinearDiscriminantAnalysis(solver='eigen',tol=0.0001)
LDA_Model =clf.fit(Xtrain,ytrain)
```

Model 3: Logistic Regression using Stats model library (Fig 2.13)

Logit Regression Results

Dep. Variable:	Holiday_package	**No. Observations:**	610
Model:	Logit	**Df Residuals:**	604
Method:	MLE	**Df Model:**	5
Date:	Sun, 06 Jun 2021	**Pseudo R-squ.:**	0.1128
Time:	16:11:50	**Log-Likelihood:**	-373.15
converged:	True	**LL-Null:**	-420.60
Covariance Type:	nonrobust	**LLR p-value:**	6.251e-19

	coef	std err	z	P>|z|	[0.025	0.975]
Salary	-1.527e-05	5.03e-06	-3.033	0.002	-2.51e-05	-5.4e-06
Age	-0.0190	0.006	-3.026	0.002	-0.031	-0.007
No_years_Education	0.1198	0.029	4.093	0.000	0.062	0.177
No_Young_kids	-1.1439	0.196	-5.846	0.000	-1.527	-0.760
Foreigner	1.4127	0.221	6.400	0.000	0.980	1.845
No_Elder_kids	0.1050	0.080	1.311	0.190	-0.052	0.262

1. Features like Foreigner and Number of young children shows high impact on predicting the target than other features. The bivariate analysis plots also confirmed the same.
2. Salary and Age shows minimalistic negative impact on the target.
3. The p-value for Number of elder kids is > 0.05. This is not a good predictor for the response variable.

2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.

Fig 2.14: Model comparison:

	Logit_Train	Logit_Test	LDA_Train	LDA_Test
Accuracy	67.38	66.41	65.90	69.85
AUC	73.41	72.55	73.47	72.44
Recall	58.06	51.64	57.35	57.38
Precision	66.39	68.48	64.26	72.16
F1 Score	61.95	58.88	60.61	63.93

Test Metrics:

	A : Yes | P : Yes	A : No| P : No	A : No | P : Yes	A : Yes | P : No	TP + TN	FP + FN	Specificity	Sensitivity	G-Mean	MCC
Confusion Matrix	*TP*	*TN*	*FP*	*FN*	*Accuracy*	*Misclassification*	*TN/TN+FP*	*TP/TP+FN*		
Logit	63.00	111.00	29.00	59.00	174.00	86.00	0.79	0.52	0.64	0.32
LDA	**70.00**	**113.00**	**27.00**	**52.00**	**183.00**	**82.00**	**0.81**	**0.57**	**0.68**	**0.39**
	A : Actual | P : Predicted									

1. Accuracy: This measure evaluates overall efficiency of the model. LDA model have a better accuracy score (Accuracy Score = TP+TN / TP+TN+FP+FN)
2. Area Under Curve: AUC score for both the models are similar. Let us look at the other metrics to finalize the model.
3. Classification metrics:
    - ***True Positive:*** Employee interested in holiday package predicted as interested**
    - True Negative: Employee not interested in holiday package predicted as not interested.
    - False Positive: Employee not interested in holiday package predicted as interested.
    - ***False Negative:*** Employee interested in holiday package predicted as not interested**
4. ***False negative has more impact on the business. The tour agency may miss prospective customers for holiday package who are under False negative category.***
5. Precision: It is measure for model's exactness. The count of False positives is less in LDA model. (Precision =TP / TP+FP) A higher precision value is a indication of good classifier.
6. Recall: It is the percent of Positive cases identified as model. The recall score for Logistic regression model is lesser than LDA model. LDA model is able to catch a greater number of Positive cases thereby reducing the False Negatives.
7. F1 Score: F1 is the harmonic mean of Precision and Recall. LDA has better Precision and Recall scores, hence the F1 score for LDA is better than Logistic regression.
8. G-mean score: It is the square root of Sensitivity and Specificity. This measure is important in the avoidance of overfitting the negative class and under fitting the positive class. LDA has a slight high G-mean score.

Fig 2.15: Logistic Regression and Linear Discriminant Analysis: Classification Report, Confusion Matrix plot and ROC curve:

Logit :Classification Report : Train Data

	precision	recall	f1-score	support
0	0.68	0.75	0.71	331
1	0.66	0.58	0.62	279
accuracy			0.67	610
macro avg	0.67	0.67	0.67	610
weighted avg	0.67	0.67	0.67	610

Logit :Classification Report : Test Data

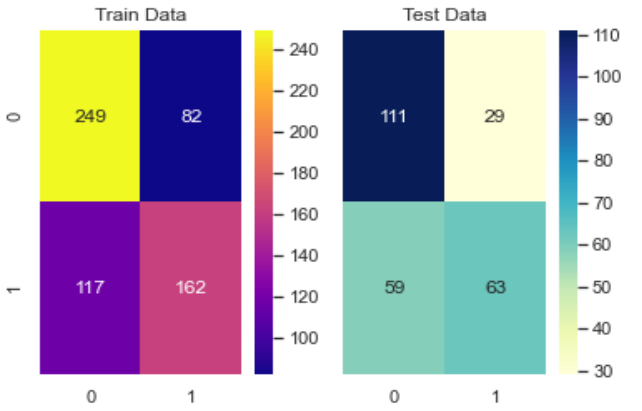
	precision	recall	f1-score	support
0	0.65	0.79	0.72	140
1	0.68	0.52	0.59	122
accuracy			0.66	262
macro avg	0.67	0.65	0.65	262
weighted avg	0.67	0.66	0.66	262

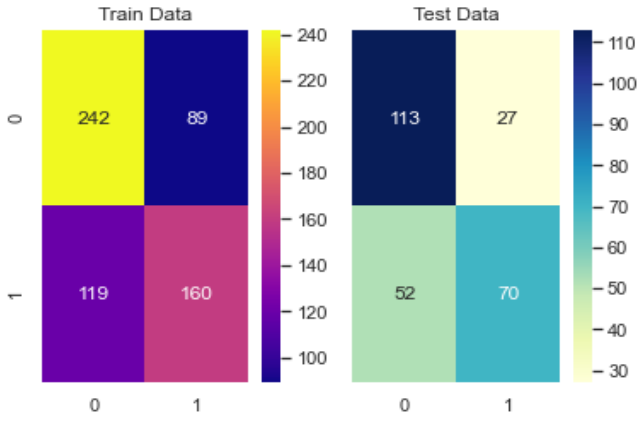
LDA :Classification Report : Train Data

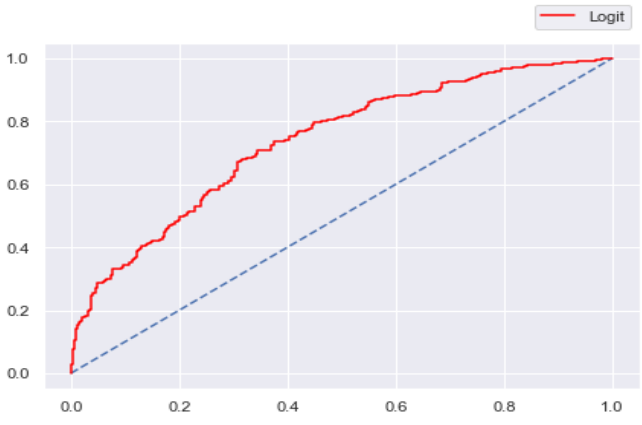
	precision	recall	f1-score	support
0	0.67	0.73	0.70	331
1	0.64	0.57	0.61	279
accuracy			0.66	610
macro avg	0.66	0.65	0.65	610
weighted avg	0.66	0.66	0.66	610

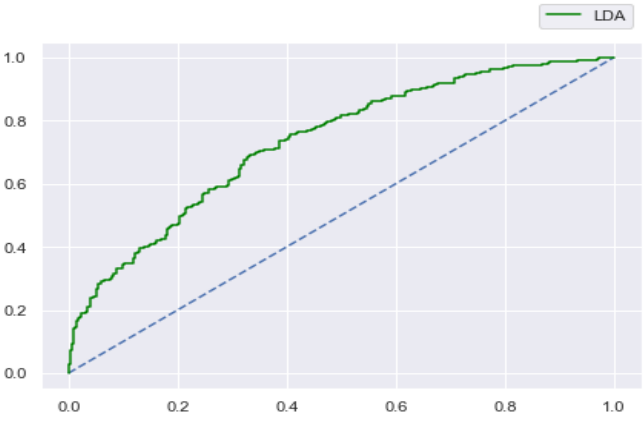
LDA :Classification Report : Test Data

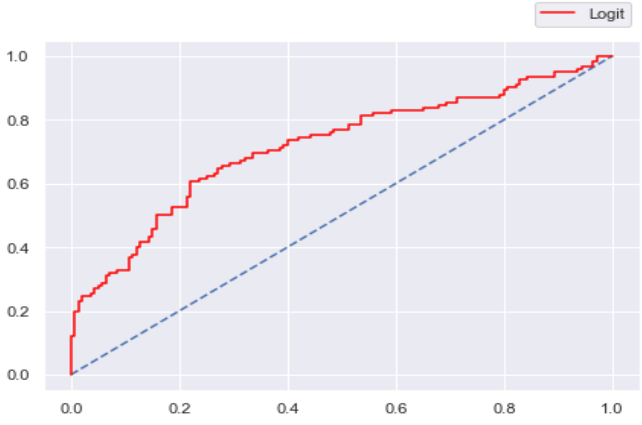
	precision	recall	f1-score	support
0	0.68	0.81	0.74	140
1	0.72	0.57	0.64	122
accuracy			0.70	262
macro avg	0.70	0.69	0.69	262
weighted avg	0.70	0.70	0.69	262

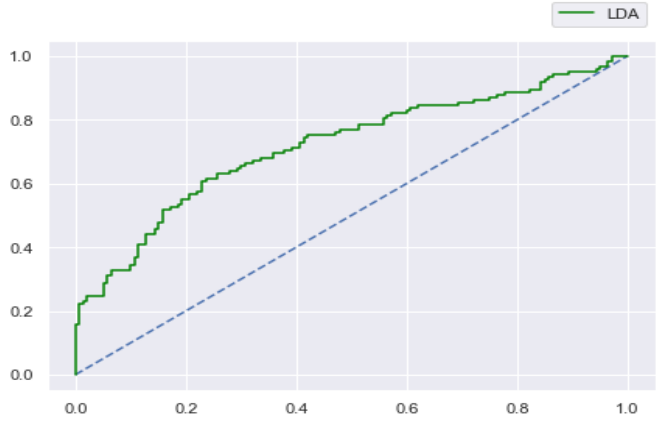








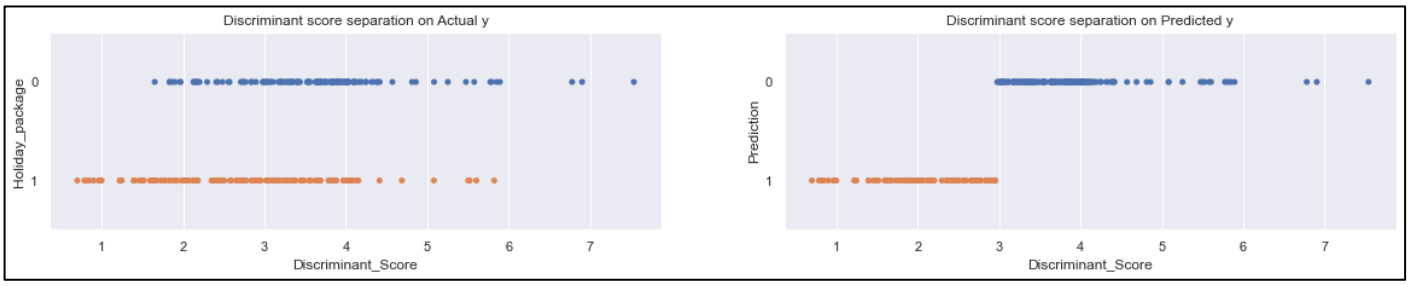




9. Matthew's Correlation Coefficient: It is a correlation coefficient between the observed and predicted classifications. The value ranges from -1 to +1. LDA have MCC as 0.39. LDA model did not either have perfect or worst prediction but predicted values better than Logistic regression model.

From the Classification metrics we can conclude that 'Linear Discriminant model is better than Logistic regression.

Fig 2.16: Linear Discriminant Analysis Model: Discriminant score analysis B/W Actual and Predictions:



If Discriminant score less than 3 it is predicted as 'YES' and more than 3 it is predicted as 'NO',

Table 2.6 Overview of 10 observations from a data frame with Actual y, predicted y and Predicted probabilities:

	Salary	Age	No_years_Education	No_Young_kids	No_Elder_kids	Foreigner	Holiday_package	Predicted_Probabilities	Predicted_Holiday_Package
287	44476.0	36	12	0	2	0	1	0.565939	1
313	50066.0	55	9	0	0	0	1	0.289119	0
740	32175.0	58	5	0	0	1	1	0.553591	1
147	142183.0	31	15	0	2	0	0	0.313394	0
491	36308.0	25	16	2	0	0	0	0.156674	0

Fig 2.17: Unstandardized Coefficients from Logistic regression and Linear Discriminant analysis:

	Salary	Age	No_years_Education	No_Young_kids	No_Elder_kids	Foreigner
0	-0.00002	-0.03461	0.07128	-1.33848	0.0369	1.02155

Logistic Regression

	Salary	Age	No_years_Education	No_Young_kids	No_Elder_kids	Foreigner
0	-0.00001	-0.05566	0.03149	-1.437	-0.0342	1.14735

Linear Discriminant Analysis

1. Foreigner and Number of young children's coefficients have high weightage.
2. Other all coefficients are very low.
3. Intercept from LDA model is 2.60697. Then the probability of having the outcome will be > 0.5.

4. Discriminant score = Intercept (β_0) + (β_1) * Foreigner - (β_2) * No of Young kids + (β_3) * No of years of Experience - (β_4) * Age - (β_5) * Salary
5. Discriminant score = 2.60697+ 1.14735 * Foreigner – 1.437 * No of Young kids + 0.03149 * No of years of Experience – 0.05566 * Age – 0.00001 * Salary

2.4 Inference: Basis on these predictions, what are the insights and recommendations.

Fig 2.18: Histogram representation of TP, TN, FP and FN in the Test data (262 observations):

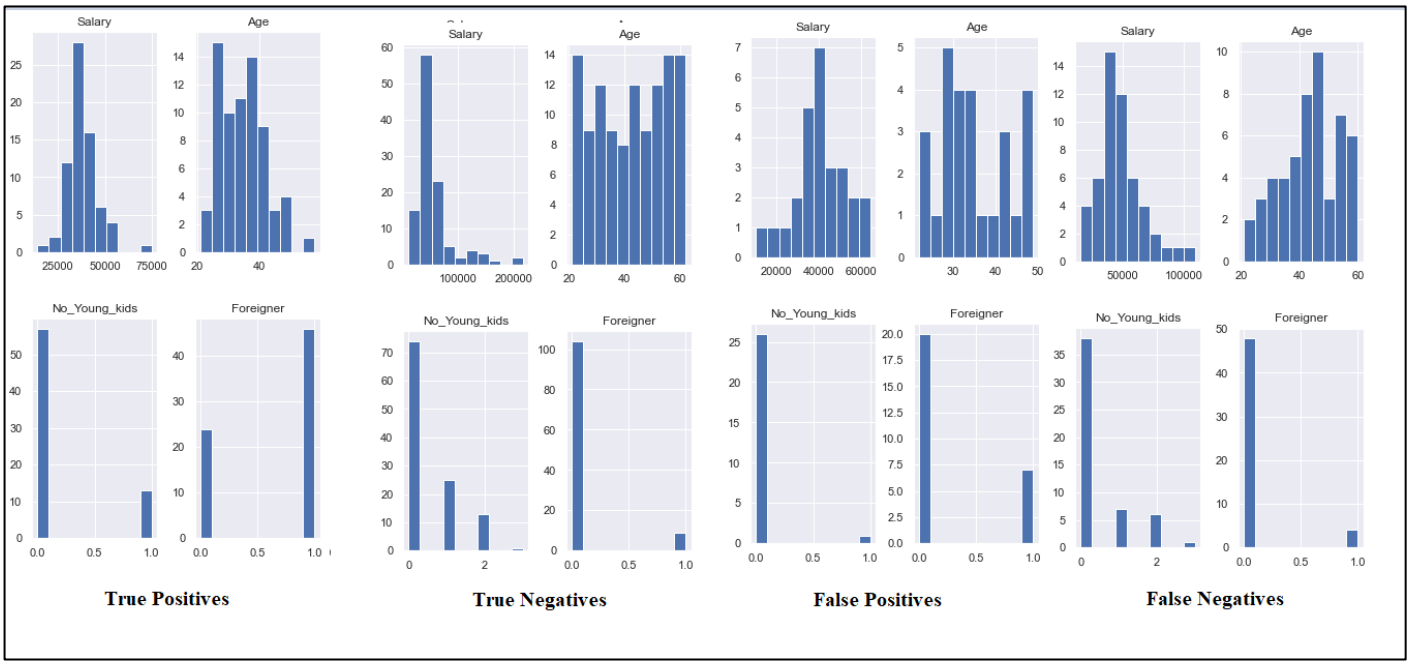


Fig 2.19: Classification metric scores on different Probability Cut-offs:

```
for j in np.arange(0.3,0.7,0.1):
    custom_prob = j
    custom_cutoff_data=[]
    for i in range(0,len(ytrain)):
        if np.array(LDA_Model.predict_proba(Xtrain)[:,1])[i] > custom_prob:

            a=1
        else:
            a=0
        custom_cutoff_data.append(a)
    print('Cutoff:',round(j,3),'\t','Accuracy Score',round(metrics.accuracy_score(ytrain,custom_cutoff_data),4),
                                    'F1 Score',round(metrics.f1_score(ytrain,custom_cutoff_data),4),
                                      'Precision',round(metrics.precision_score(ytrain,custom_cutoff_data),4),
                                        'recall',round(metrics.recall_score(ytrain,custom_cutoff_data),4)
         )
```

```
Cutoff: 0.3      Accuracy Score 0.5934 F1 Score 0.6737 Precision 0.5322 recall 0.9176
Cutoff: 0.4      Accuracy Score 0.6689 F1 Score 0.6824 Precision 0.6078 recall 0.7778
Cutoff: 0.5      Accuracy Score 0.659 F1 Score 0.6061 Precision 0.6426 recall 0.5735
Cutoff: 0.6      Accuracy Score 0.6443 F1 Score 0.5057 Precision 0.6938 recall 0.3978
```

1. Foreign employees and employees with 0 to 1 young kid are the prospective customers.
2. By looking at the histogram plot we can notice that few foreign employees and more number of employees with 0 young kids are found in False positives. Displaying the reviews and experience of the employees who have already travelled and letting them know the attractive features of the travel and tour package could draw their interest.
3. Creating a good Budget-friendly package will attract employees who earn less than 100000.
4. Exclusive packages for employees who are earning more than 100000.
5. Honeymoon and adventurous tour packages for employees who have no kids and young employees.
6. Reducing the cut-off probability i.e., when an observation has 0.3 as its predicted probability then the observations fall under 'Yes' class. Similarly for 0.4 for other cut-offs.
7. Cut-off like 0.3 and 0.4 shows good recall and F1 scores.
8. The business problem is about predicting the employee preference in choosing the holiday package. Therefore, the company can use 0.3, 0.4 and 0.5 cut-off probabilities in order to get high positive rate and low false negatives.
9. Marketing strategies can be implemented for employees under False positives.
10. The above strategies can be implemented to increase the number of customers for tour package.

Conclusion:

With the given set of observations and features we are asked to predict whether an employee of a particular company will choose a holiday package. The data is analysed using Logistic regression and Linear discriminant analysis to classify and predict classes YES and NO. A foreign employee or an employee with 0 or 1 young kid opts for package. Insights from Exploratory data analysis matches with the model's output. The weightage of coefficients to these features are higher than other coefficients. Employees from the above-mentioned categories can be considered as their prospective targets. As the data is not scaled, feature-salary impacted the performance of the data due to its high scale. By scaling the data and reducing the cut-off probability we can get better classification metrics.

List of libraries imported in the Jupyter Codebook:

Common Libraries:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
sns.set()
%matplotlib inline
from sklearn.model_selection import train_test_split
from statsmodels.stats.outliers_influence import variance_inflation_factor
```

Problem Statement:1

```
from sklearn.linear_model import LinearRegression
import statsmodels.formula.api as smf
from sklearn.preprocessing import StandardScaler
from matplotlib.patches import Rectangle
from sklearn.decomposition import PCA
```

Problem Statement: 2

```
from sklearn.model_selection import train_test_split,GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn import metrics
from sklearn.metrics import
roc_auc_score,roc_curve,classification_report,confusion_matrix,plot_confusion_matrix,accuracy_score
```