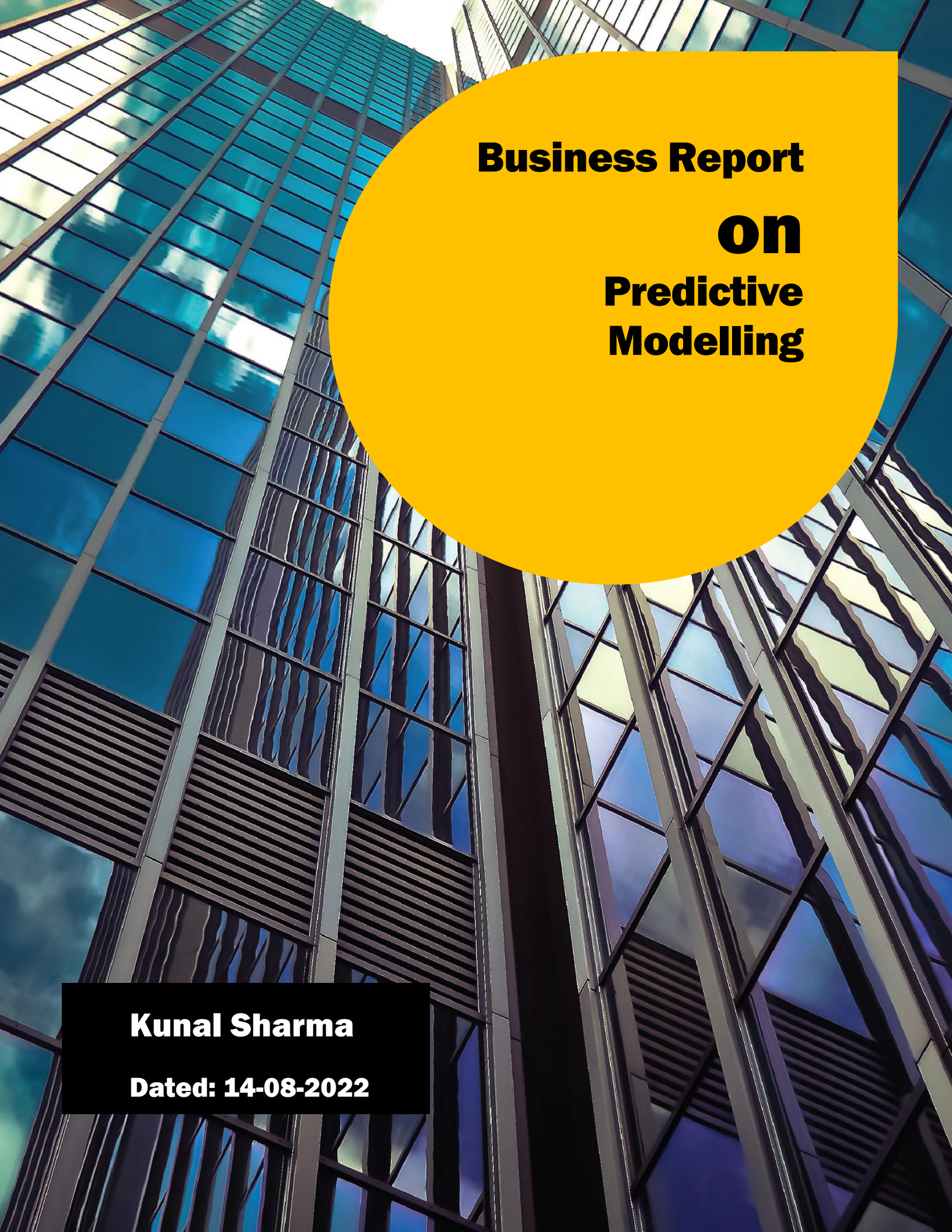


Business Report on Predictive Modelling

Kunal Sharma

Dated: 14-08-2022

Table of Contents

List of Figures

Problem 1: Linear Regression

Executive Summary

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:

- 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 worst and J the best.
- Clarity: Clarity refers to the absence of the Inclusions and Blemishes. (In order from Worst to Best in terms of avg price) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1
- Depth: The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.
- Table: The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
- Price: the Price of the cubic zirconia.
- X: Length of the cubic zirconia in mm.
- Y: Width of the cubic zirconia in mm.
- Z: Height of the cubic zirconia in mm.

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

Snapshot of DataFrame

Table 1: Snapshot of Dataframe

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
0	1	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	2	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
2	3	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	4	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
4	5	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779

In the given dataset, column 'Unnamed: 0' seems to be of no us hence for the further analysis, we are dropping the particular column from the sample.

After dropping the feature, following dataset has been obtained:

	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

Understanding Data and Missing Values

#	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

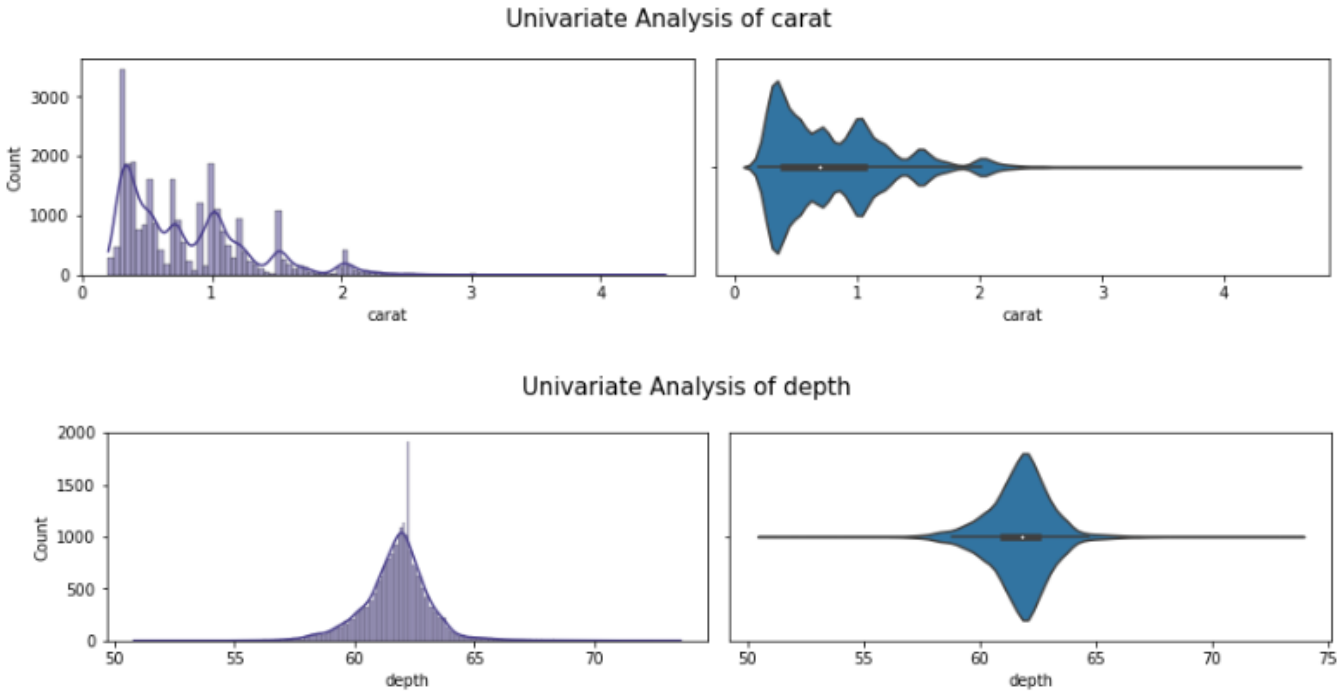
Table 2: Data Describe

	count	mean	std	min	25%	50%	75%	max
carat	26967.0	0.80	0.48	0.2	0.40	0.70	1.05	4.50
depth	26270.0	61.75	1.41	50.8	61.00	61.80	62.50	73.60
table	26967.0	57.46	2.23	49.0	56.00	57.00	59.00	79.00
x	26967.0	5.73	1.13	0.0	4.71	5.69	6.55	10.23
y	26967.0	5.73	1.17	0.0	4.71	5.71	6.54	58.90
z	26967.0	3.54	0.72	0.0	2.90	3.52	4.04	31.80
price	26967.0	3939.52	4024.86	326.0	945.00	2375.00	5360.00	18818.00

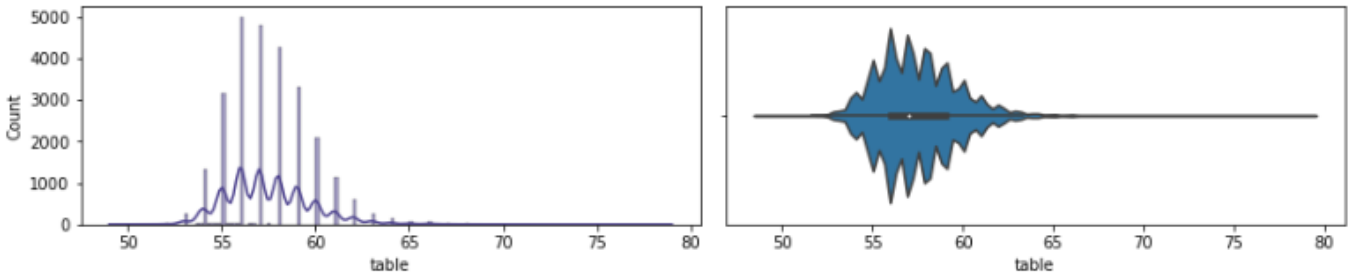
	cut	color	clarity
count	26967	26967	26967
unique	5	7	8
top	Ideal	G	SI1
freq	10816	5661	6571

Univariate Analysis

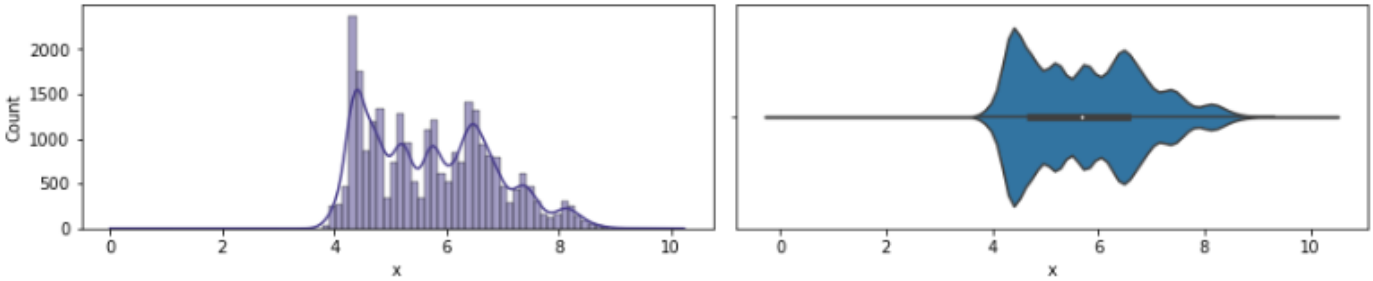
Figure 1: Univariate Analysis of Continuous Variables



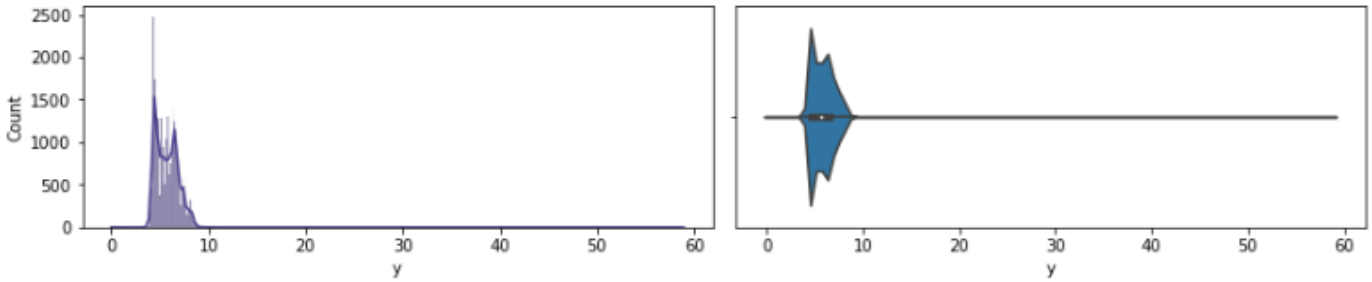
Univariate Analysis of table



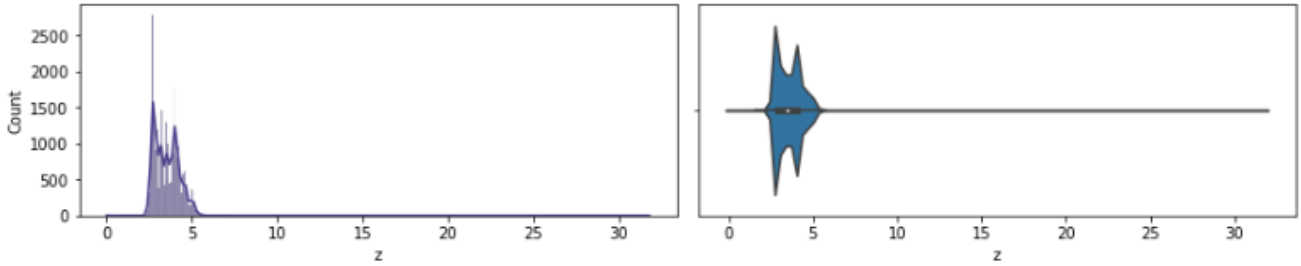
Univariate Analysis of x



Univariate Analysis of y



Univariate Analysis of z



Univariate Analysis of price

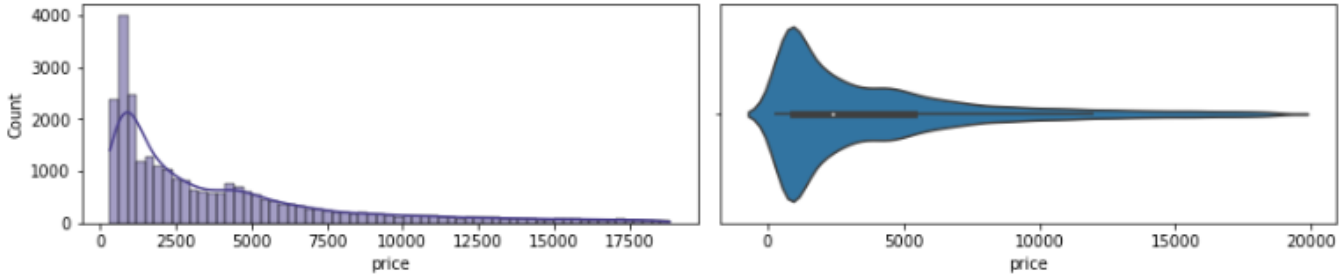
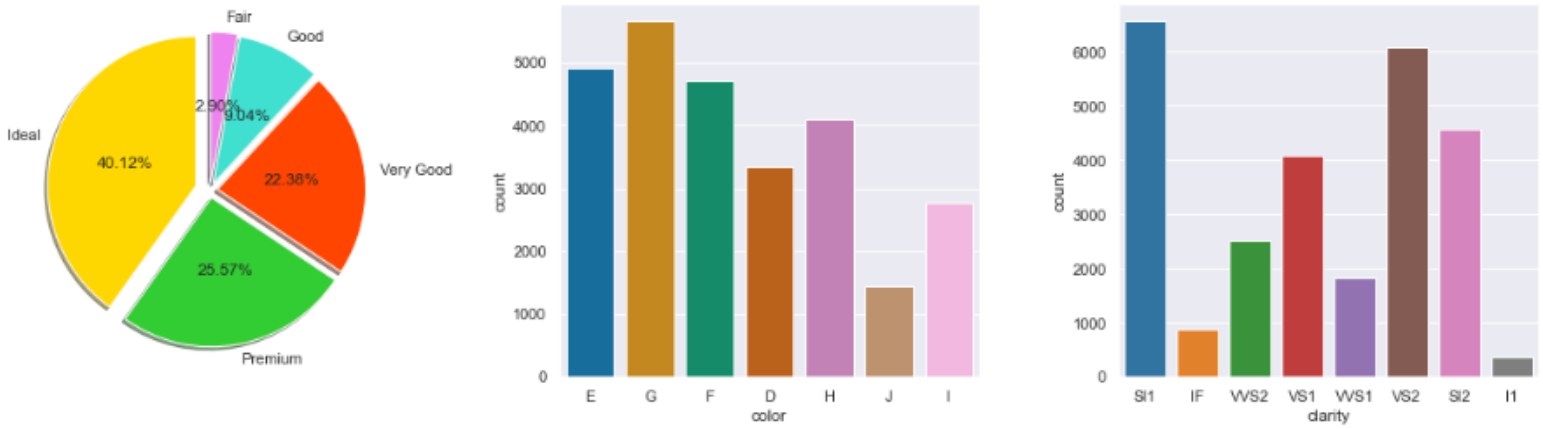


Table 2: Skewness

	carat	depth	table	x	y	z	price
Skewness	1.11	NaN	0.77	0.39	3.87	2.58	1.62

From table 2, we can observe that for depth column, NaN value is showing for skewness which can be due to the missing values present.

Figure 2 : Univariate Analysis of Categorical variables



Bi-Variate Analysis

Figure 3: Correlation matrix

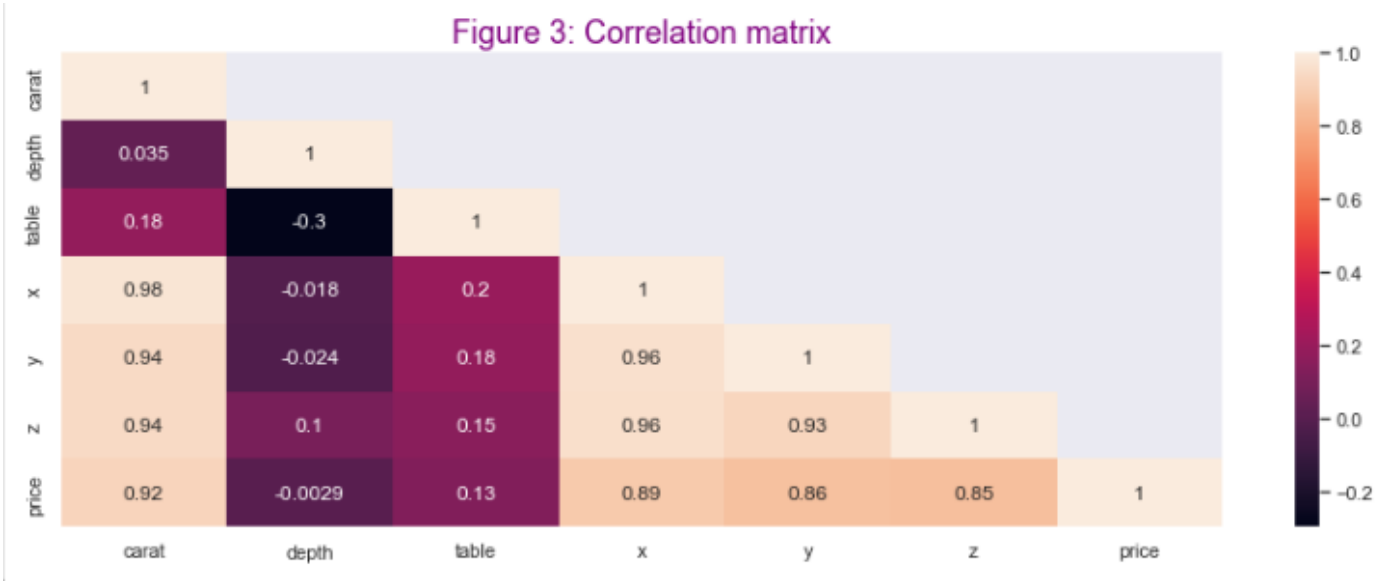
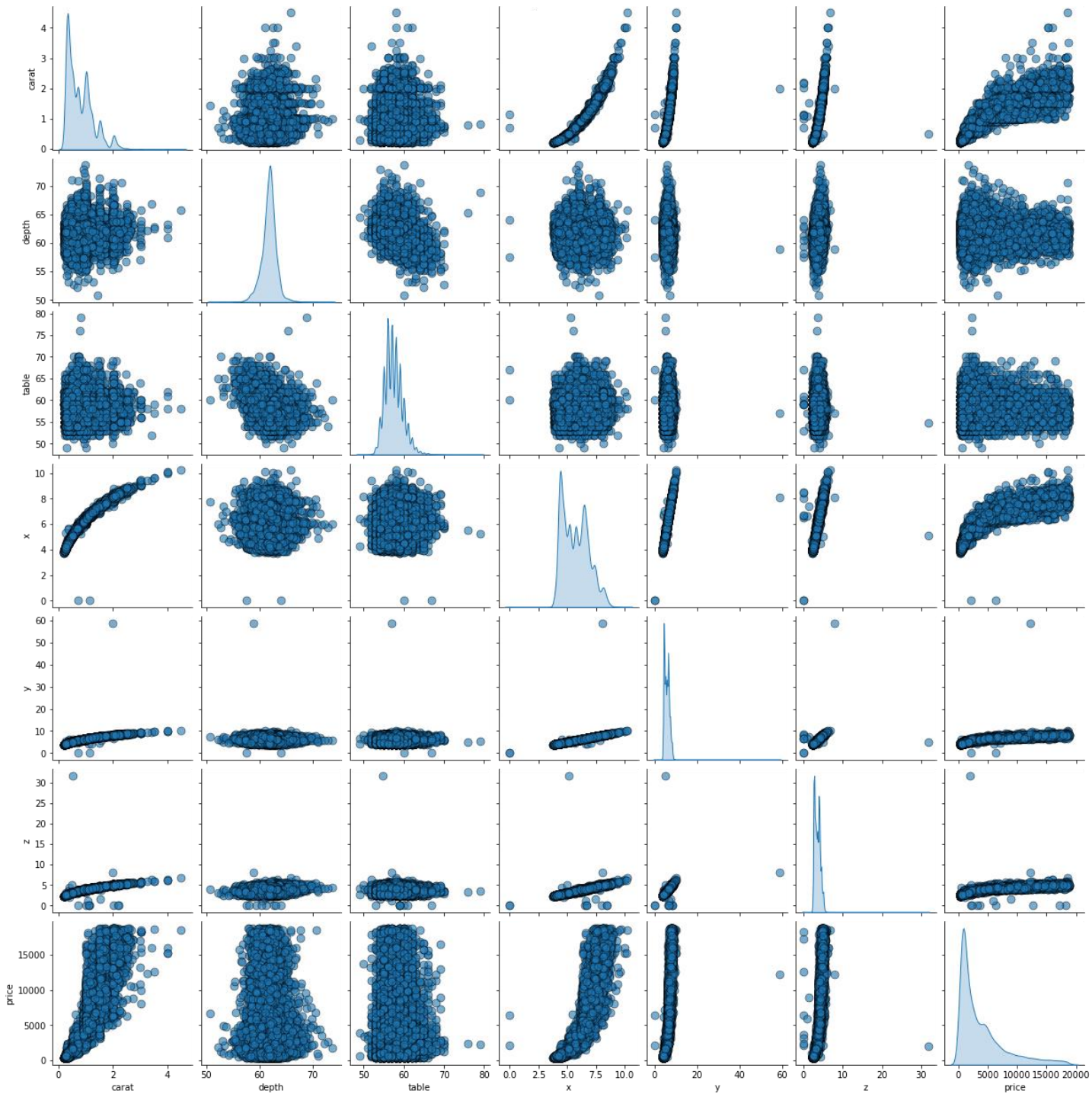


Figure-4: Pairplot Analysis of Continuous Variables



- A strong positive correlation can be observed between price w.r.t. carat, width, length and height.

Figure 5: Bi-Variate Analysis of Cut

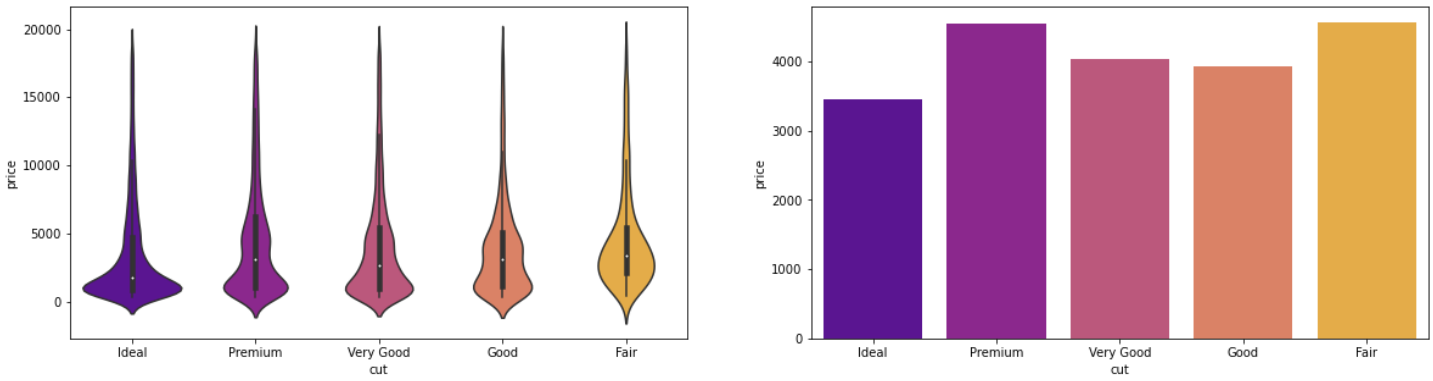


Figure 6: Bi-Variate Analysis of Color

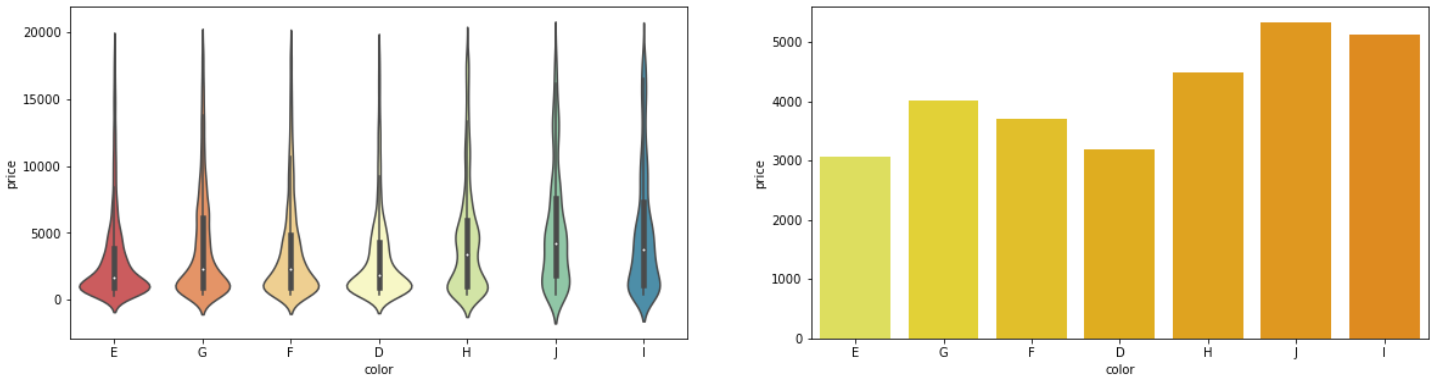
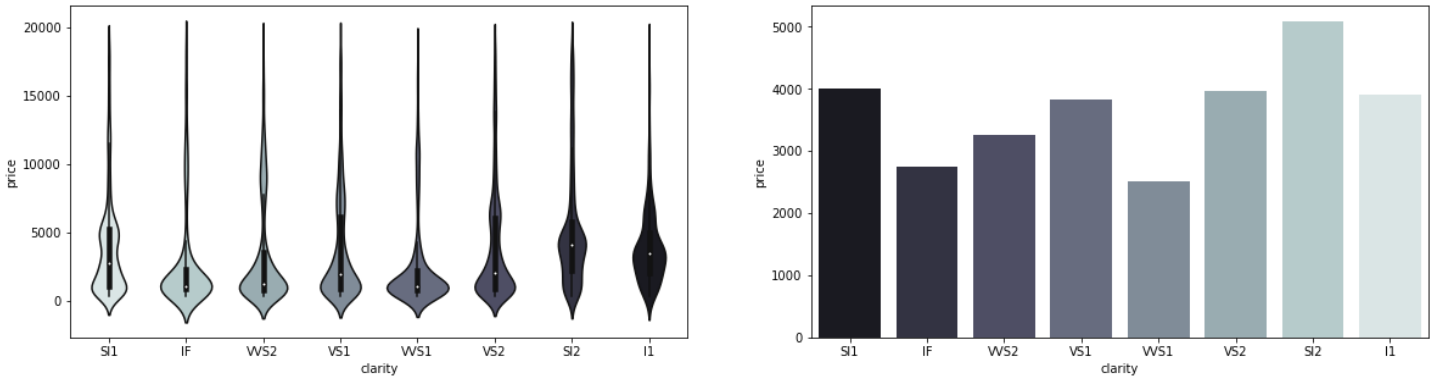


Figure 7: Bi-Variate Analysis of Clarity



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? Check for the possibility of combining the sub levels of a ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning.

From the given data, we have observed a total of 697 null values present in 'depth' column which has to be treated however since there are outliers present, we would have to impute values with the median values. Fillna() comes handy to impute null values at once.

After applying function, we can perceive from the below data that the null values has been treated successfully:

Table 3: Imputation of Null values

	carat	cut	color	clarity	depth	table	x	y	z	price
0	0	0	0	0	0	0	0	0	0	0

Also, from descriptive summary, we can observe a several '0' values present in x, y and z columns which is a clear indication of false entries since diamonds cannot be 2-dimensional. However, with only 8 values, data won't be affected that much hence we can chose to drop these invalid values.

After dropping the bad values, we can observe from the below table that the overall observations has been reduced to 26925 keeping the columns intact as 10. Furthermore, bad values has been treated successfully.

Table 4: Descriptive Summary

	carat	depth	table	x	y	z	price
count	26925.000000	26925.000000	26925.000000	26925.000000	26925.000000	26925.000000	26925.000000
mean	0.797821	61.746982	57.455305	5.729385	5.733152	3.538820	3936.249991
std	0.477085	1.393457	2.231327	1.126081	1.163820	0.717483	4020.983187
min	0.200000	50.800000	49.000000	3.730000	3.710000	1.070000	326.000000
25%	0.400000	61.100000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	0.700000	61.800000	57.000000	5.690000	5.700000	3.520000	2373.000000
75%	1.050000	62.500000	59.000000	6.550000	6.540000	4.040000	5353.000000
max	4.500000	73.600000	79.000000	10.230000	58.900000	31.800000	18818.000000

Moreover, we have observed from the data that various sub-categories has been provided for clarity which itself has sub-levels which indicates certain quality standards however this can be further truncated with a threshold of quality This shall help us to determine standard quality parameters for clarity of diamond.

Table 5.1: Ordinal summary

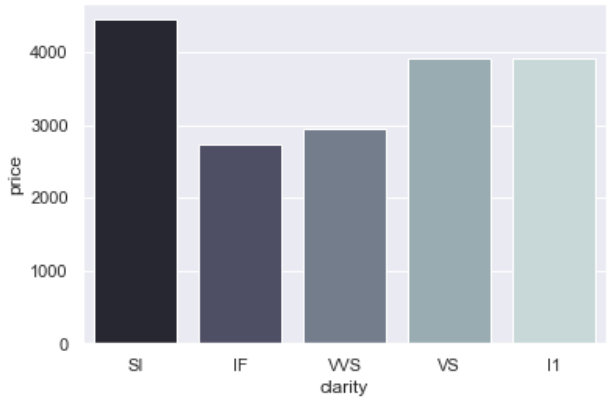
	SI1	VS2	SI2	VS1	VVS2	VVS1	IF	I1
clarity	6565	6093	4564	4087	2530	1839	891	364

As we can observe that ordinal variables SI has two sub-values SI1 and SI2 which can be merged into one and similar exercise can be done for VS as well as VVS.

Table 5.2: Ordinal summary post treatment

	I1	IF	VVS	VS	SI
clarity	364	891	4369	10180	11129

Figure 8: Clarity Analysis



After treating sub-ordinal variables, we can observe from Table 5.2 a much subtle and defined view for certain clarity parameters for diamond. Also, Figure 8 illustrates that no difference in terms of effect on price can be seen after applying feature engineering. SI still happens to be top contributing factor for price of diamond followed by VS and I1.

1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj. Rsquare. Compare these models and select the best one with appropriate reasoning.

After applying label encoding to the dataset, following output can be observed:

Table 6.1: Label Encoding Sample

	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.30	4	5	1	62.1	58.0	4.27	4.29	2.66	499
1	0.33	3	3	4	60.8	58.0	4.42	4.46	2.70	984
2	0.90	2	5	3	62.2	60.0	6.04	6.12	3.78	6289
3	0.42	4	4	2	61.6	56.0	4.82	4.80	2.96	1082
4	0.31	4	4	3	60.4	59.0	4.35	4.43	2.65	779

Table 6.2: Label Encoding Sample

	carat	cut	color	clarity	depth	table	x	y	z	price
0	float64	object	object	object	float64	float64	float64	float64	float64	int64

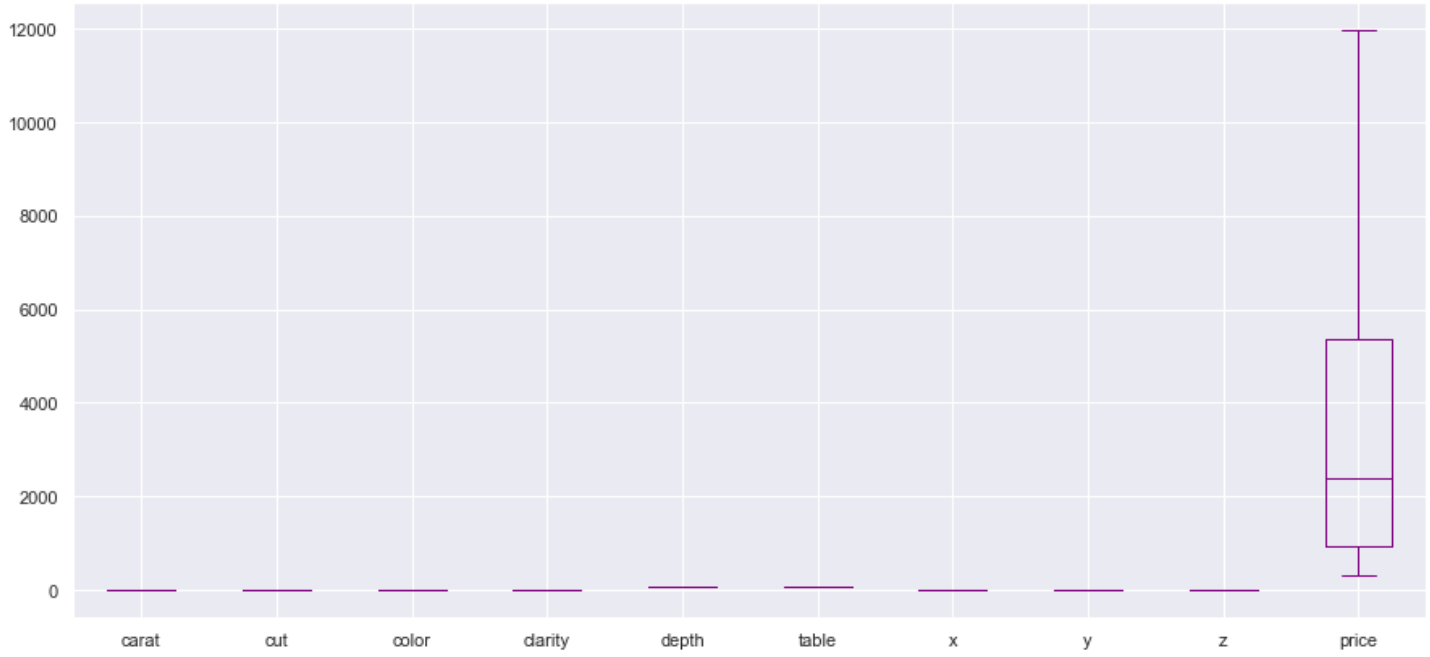
However, it can be seen from Table 6.2 that though the encoding has been done, yet, data type can be still seen as type object which has to be converted to type integer before modelling can be done.

Outlier Treatment

Before moving to further analysis and splitting data, outliers must be treated as it can impact model. Certain techniques can be applied such as IQR or Z-score method to treat outliers however for this analysis, we shall proceed with IQR method and treat extreme values with lower & upper whisker values.

After treating the outliers, we can observe from Figure 9 that our dataset does not have any extreme values present and suitable for further segmentation.

Figure 9: Outlier Treatment



Train-Test-Split

Train_test_split concepts is a preliminary step taken before any model is built. It is effective in building an optimum model which neither underfits nor overfits the data in such a fashion that it partitions the data into trained set and testing set.

Training Set or labelled data is an example given to the model to analyses and learn the patterns in the data which is about **70% of our total data**.

Testing set can be treated as 'unseen data' and is used for building the model. In other words, it is used to test the hypothesis generated by the model. To know the performance of any model, it should be tested on unseen data which is about **30% of our total data**.

Parameters for Train_test_split():

- Test_size – Helps to determine the size of test set.
- Train_size – Represents the proportion of data set in training set.
- Stratify – Split dataset in such a fashion that the ratio of class labels is constant. Specially used when the data is imbalanced.
- Shuffle – Shuffles the data before splitting.
- Random_state – Controls the shuffling of the dataset before splitting and makes the outcome reproducible with same result.

However, before applying splitting(), we need to ensure that no object type data is present and encoding needs to be applied otherwise as model only accepts integer data type.

Additionally, target variable needs to be dropped from original dataset and stored in a separate variable to use later for testing.

Further, splitting() has been applied keeping **30% as testing dataset** and following variables has been generated:

1. Xtrain – Training dataset without target variable
2. Xtest – Testing dataset without target variable
3. Ytrain – Training dataset with target variable which needs to be predicted
4. Ytest - Training dataset with target variable which needs to be predicted

Model 1: Linear Regression

Linear Regression is a procedure of predicting a real number methods of which model data with linear combination of the explanatory variables which can be expressed as:

*Dependent variable = (weight * independent variable) + constant*

Independent Variables modelled should follow a correlation with Target or dependent variable which thus indicates how closely their relationship follows a straight line.

The main function of Linear Regression is to obtain a best fit line representing linear combination in such a way that mean sum of squared error has been curtailed. '**Gradient Descent'** method comes in handy to determine Best-fit line in a model.

Co-efficient of Determinant

Denoted as R^2 (R-square), coefficient of determinant determines the fitness of linear model. Closer the value tends to 1 i.e. closer the data points to the line, the better the model is.

Accuracy of a R^2 is generally affected by **Regression error** i.e. variance captured between actual target variable and predicted variable.

Step 1: Calling Linear Regression model and fitting the model

Linear Regression fits a linear model with coefficients m = (m1, ..., m_n) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.

Step 2: Checking Co-efficients and Intercept

Coefficients or weights are used as representation of the data and is a combination of all standard deviations, covariance and correlation. It is allotted as one scale factor with every independent variable which depicts the amount of difference occur in Target variable for certain change in 'co-efficient' value in Independent variable. In other words, it determines the strength of variation in Target variable with change in Independent variable.

Intercepts or constant represents the mean value of the response or target variable when all the predictors (independent variables) are equal to zero.

After fitting the model, following coefficients can be observed:

- The coefficient for carat is: 8849
- The coefficient for cut is: 112
- The coefficient for color is: 262
- The coefficient for clarity is: 848
- The coefficient for depth is: 34
- The coefficient for table is: -17
- The coefficient for x is: -1461
- The coefficient for y is: 1653
- The coefficient for z is: -954

The intercept for our model is **-4827** which derives that with clarity, cut, color, etc. all being zero (0), then the mean price would be -4827 which is meaningless. Nevertheless, normalised technique i.e. z-score can be applied to make it nearly zero.

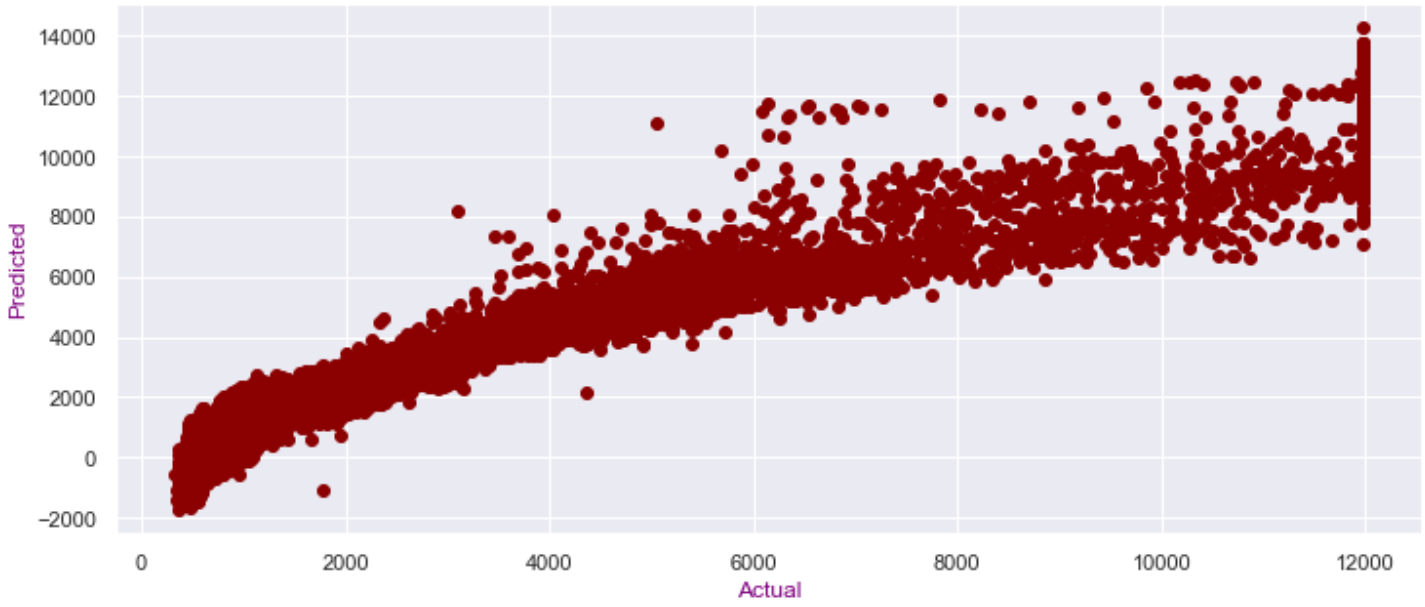
Step 3: Checking R^2 value i.e. Accuracy

The accuracy for trained model is:- **0.93**
The accuracy for test model is:- **0.93**

R-square i.e. coefficient of determinant determines the accuracy or efficiency of training & test set and for our regression model, we have observed score of 93% approx. for both training & testing set which indicates that model can be a right fit model.

Step 4: Predicting the variables and calculating errors

Figure 10: Test vs. Predict comparison



The above figure illustrates strength of dependency between target and predicted variable thus a linear plot, moderate correlation. However, there are lots of spread which indicates noise present in the data.

Root mean square of training data is: - **924.14**

Root mean square of testing data is: - **927.99**

RMSE of training data determines that predicted best line has a variance error of 924.14 w.r.t actual point whereas predicted best line for testing data is 927.99 varied from actual point.

Model 2: Stats models

OLS i.e. Ordinary Least method is another useful method used in Linear Regression to estimate the parameters by appropriately choosing the coefficients in such a fashion that it would reduce the overall Total Sum of square error.

Step 1: Concatenating train and test sets

Table 7: Concatenation of Training sets

	carat	cut	color	clarity	depth	table	x	y	z	price
16997	1.26	3.0	1.0	1.0	60.5	62.0	6.97	6.92	4.20	5292.0
24457	0.90	2.0	3.0	2.0	59.0	59.0	6.35	6.36	3.72	4484.0
16612	1.37	4.0	4.0	2.0	59.6	57.0	7.28	7.22	4.32	11649.0
308	0.84	1.0	3.0	2.0	63.6	57.0	5.98	5.93	3.79	3316.0
26652	2.00	2.0	0.0	2.0	60.8	62.0	8.09	8.12	4.93	11972.5

Table 8: Concatenation of Tests sets

	carat	cut	color	clarity	depth	table	x	y	z	price
22114	0.34	2.0	2.0	2.0	62.4	60.0	4.41	4.44	2.76	537.0
2275	0.30	4.0	5.0	2.0	61.2	55.0	4.35	4.31	2.65	844.0
19183	0.50	4.0	3.0	1.0	62.5	57.0	5.09	5.05	3.17	1240.0
5030	1.10	1.0	5.0	1.0	63.3	56.0	6.53	6.58	4.15	4065.0
25414	1.02	3.0	4.0	1.0	61.1	62.0	6.54	6.49	3.98	4057.0

Step 2: Hypothesis Testing

Hypothesis testing is an effective measure taken in OLS to determine the reliability of co-efficients or if there is a chance that target variable has a relationship with predictor in the universe.

Null Hypothesis i.e. H_0 : There is no relationship between Target & Predictor in the universe.

Alternate Hypothesis i.e. H_1 : There is relationship between Target & Predictor in the universe.

P-value suggests the probability of finding the identical coefficient value in the sample data w.r.t. Null Hypothesis.

Step 2: Computing Variance Inflation Factor (VIF) and building OLS model:

VIF identifies correlation between independent variables and strength of correlation. Mathematically, it can be computed as:

$$VIF = \frac{1}{1-r^2}$$

VIF = 1 indicates no correlation

1 < VIF < 5 indicated moderate collinearity

VIF > 5 indicates severe/critical Multicollinearity i.e. co-efficients are poorly estimated

Upon calculating VIF, following values can be obtained:

- VIF for carat is: 122.97
- VIF for cut is: 10.32
- VIF for color is: 5.51
- VIF for clarity is: 7.0
- VIF for depth is: 1220.72
- VIF for table is: 874.13
- VIF for x is: 10669.97
- VIF for y is: 9417.87
- VIF for z is: 3320.47

As we can observe, all the predictors seems to have high Multicollinearity. However, before doing revised analysis, we shall proceed to build OLS model and understand the p-val, R^2 and standard error.

After applying OLS formula, we can observe:

Summary as follow:

Table 9: OLS Summary

Dep. Variable:	price	**R-squared:**	0.929
Model:	OLS	**Adj. R-squared:**	0.929
Method:	Least Squares	**F-statistic:**	2.722e+04
Date:	Tue, 09 Aug 2022	**Prob (F-statistic):**	0.00
Time:	02:55:06	**Log-Likelihood:**	-1.5545e+05
No. Observations:	18847	**AIC:**	3.109e+05
Df Residuals:	18837	**BIC:**	3.110e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4826.7285	815.423	-5.919	0.000	-6425.030	-3228.427
carat	8848.5715	84.199	105.091	0.000	8683.534	9013.609
cut	112.0991	7.455	15.036	0.000	97.486	126.712
color	262.1516	4.171	62.851	0.000	253.976	270.327
clarity	847.9543	9.197	92.200	0.000	829.928	865.981
depth	33.6083	11.302	2.974	0.003	11.456	55.761
table	-17.0220	3.987	-4.269	0.000	-24.837	-9.207
x	-1461.1109	137.885	-10.597	0.000	-1731.378	-1190.843
y	1652.5259	135.540	12.192	0.000	1386.855	1918.197
z	-953.5113	141.708	-6.729	0.000	-1231.271	-675.752

From table 9, we can observe that the intercept is nearly identical to what we have built in Model 1 with coefficients values and R^2 also being identical. Overall p-value seems to be less than 0.05 thus we fail to accept the H_0 and statistically, can conclude that chances of having a relationship with price (Target variable) are not high.

Model 3: Linear Regression (Scaled)

Intercept of -4827 was meaningless in predicting and interpreting the predictors w.r.t target variables. Normalized scaling technique can be utilized to scale the data and understand the effects on the overall scores.

Note – Outliers treatment and Encoding is still a must before modelling.

After applying Z-Score scaling technique on data and fitting Linear Regression model on the scaled trained and test data, following scores has been observed:

- The coefficient for carat is: 1
- The coefficient for cut is: 0
- The coefficient for color is: 0
- The coefficient for clarity is: 0
- The coefficient for depth is: 0
- The coefficient for table is: 0
- The coefficient for x is: 0
- The coefficient for y is: 1

- The coefficient for z is: 0
- The Intercept of scaled data is: **0**

- The accuracy for Scaled trained model is:- 0.93
- The accuracy for Scaled test model is:- 0.93

- Root mean square of Scaled training data is 0.27
- Root mean square of Scaled testing data is 0.27

Model 4: Statsmodel (Scaled)

After applying z-score, following score can be observed:

VIF for scaled carat is: 32.94
VIF for scaled cut is: 1.51
VIF for scaled color is: 1.11
VIF for scaled clarity is: 1.21
VIF for scaled depth is: 4.45
VIF for scaled table is: 1.62
VIF for scaled x is: 417.0
VIF for scaled y is: 398.36
VIF for scaled z is: 234.85

Summary as follow:

Table 10: OLS Summary(Scaled)

Dep. Variable:	price	**R-squared:**	0.929
Model:	OLS	**Adj. R-squared:**	0.929
Method:	Least Squares	**F-statistic:**	2.755e+04
Date:	Tue, 09 Aug 2022	**Prob (F-statistic):**	0.00
Time:	02:50:13	**Log-Likelihood:**	-1876.8
No. Observations:	18847	**AIC:**	3774.
Df Residuals:	18837	**BIC:**	3852.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0030	0.002	1.515	0.130	-0.001	0.007
carat	1.1662	0.011	104.393	0.000	1.144	1.188
cut	0.0352	0.002	14.653	0.000	0.031	0.040
color	0.1303	0.002	63.480	0.000	0.126	0.134
clarity	0.1974	0.002	92.032	0.000	0.193	0.202
depth	0.0150	0.004	3.528	0.000	0.007	0.023
table	-0.0085	0.002	-3.405	0.001	-0.013	-0.004
x	-0.4103	0.042	-9.867	0.000	-0.492	-0.329
y	0.5106	0.041	12.557	0.000	0.431	0.590
z	-0.2197	0.031	-7.007	0.000	-0.281	-0.158

Root Mean Square of Error for Scaled Trained data is:- 0.27

Mean Square of Error for Scaled Test data is:- 0.27

After Scaling, we can observe VIF score has significantly dropped hence we can try dropping feature with highest variance to understand the impact on the overall VIF score along with R-square and p-val.

Following VIF can be observed:

Carat: 31.9
Cut: 1.5
Color: 1.11
Clarity: 1.2
Depth: 4.12
Table: 1.59
Y: 206.92
Z: 204.28

Summary as follow:

Dep. Variable:	price	**R-squared:**	0.889
Model:	OLS	**Adj. R-squared:**	0.889
Method:	Least Squares	**F-statistic:**	1.878e+04
Date:	Wed, 10 Aug 2022	**Prob (F-statistic):**	0.00
Time:	22:22:30	**Log-Likelihood:**	-6178.6
No. Observations:	18847	**AIC:**	1.238e+04
Df Residuals:	18838	**BIC:**	1.245e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0033	0.002	1.332	0.183	-0.002	0.008
cut	0.0390	0.003	12.913	0.000	0.033	0.045
color	0.1088	0.003	42.390	0.000	0.104	0.114
clarity	0.2190	0.003	81.645	0.000	0.214	0.224
depth	0.0451	0.005	8.459	0.000	0.035	0.056
table	0.0053	0.003	1.706	0.088	-0.001	0.011
x	0.2801	0.052	5.430	0.000	0.179	0.381
y	0.6817	0.051	13.355	0.000	0.582	0.782
z	0.0703	0.039	1.791	0.073	-0.007	0.147

Omnibus:	878.468	**Durbin-Watson:**	2.011
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1004.768
Skew:	0.564	**Prob(JB):**	6.57e-219
Kurtosis:	2.910	**Cond. No.**	50.2

Model Comparison

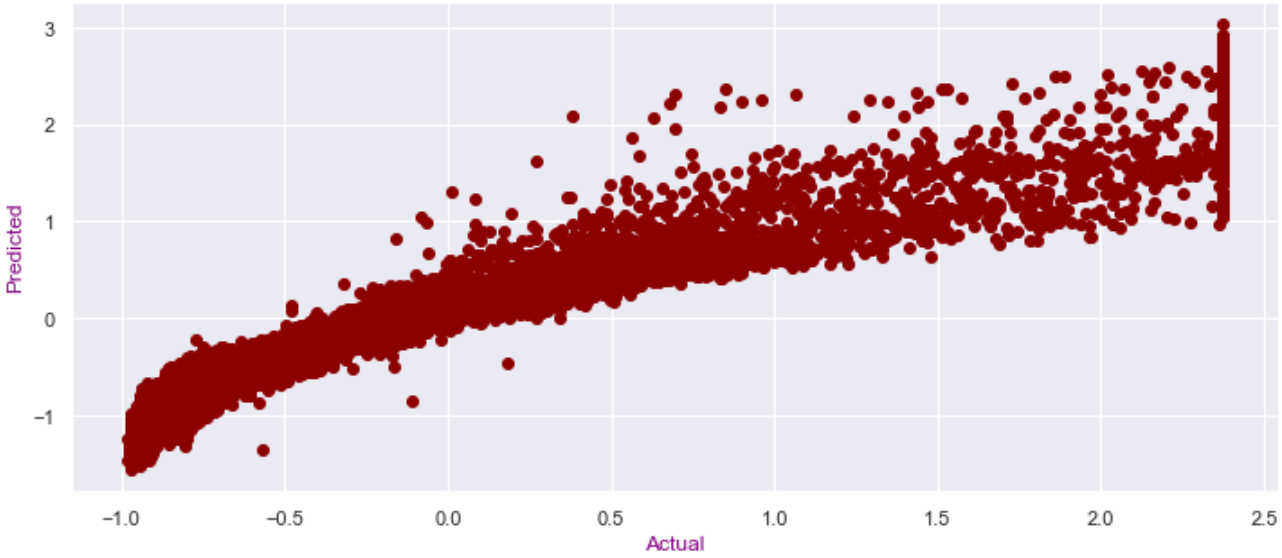
	Scaled		Unscaled	
	Linear Regression	Statsmodels	Linear Regression	Statsmodels
Co-efficient for carat	1	0	8848.571	8848.571
Co-efficient for cut	0	1	112.1	112.1
Co-efficient for color	0	0	262.15	262.15
Co-efficient for clarity	0	0	847.95	847.95
Co-efficient for depth	0	0	33.61	33.61
Co-efficient for table	0	0	-17.02	-17.02
Co-efficient for x	0	0	-1461.11	-1461.11
Co-efficient for y	1	0	1652.53	1652.53
Co-efficient for z	0	0	-953.51	-953.51
RMSE for Trained	0.27	0.27	924.14	924.14
RMSE for Test	0.27	0.27	927.99	927.99
Accuracy for Trained	0.93	-	0.93	0.929
Accuracy for Test	0.93	-	0.93	0.929

	Scaled	Unscaled
VIF for carat	32.94	122.97
VIF for cut	1.51	10.32
VIF for color	1.11	5.51
VIF for clarity	1.21	7
VIF for depth	4.45	1220.72
VIF for table	1.62	874.13
VIF for x	417	10669.97
VIF for y	398.36	9417.87
VIF for z	234.85	3320.47

	Statsmodels	
	Scaled	Unscaled
Adjusted R-square	0.929	0.929

	Linear Regression	
	Scaled	Unscaled
Intercept	0	-4827

Figure 11: Test vs. Predict comparison for Scaled



Observations:

- Overall, scaling haven't made any impact on the accuracy of the model for Linear Regression followed by Adjusted R-square for Statsmodels.
- Co-efficients which depicts deviation or variance in the data seems to have significantly dropped to 0 and 1 due to scalability in the data along with Intercept which earlier observed was -4827, however can be seen as 0 now, which does make more sense now to interpret.
- Regression error for both the models also seems to be more justifiable after normalization.
- Collinearity also seems to have reduced with scaled data. Except carat, x, y and z, variables seems to be have moderate Multicollinearity.
- Overall, it is safe to presume that **Linear Regression model** can be chosen as appropriate model with high Accuracy Score. On the other hand, R^2 seems to be impacted for Statsmodels and significantly seems dripping after variables been tweaked with highest variance.

1.4 Inference: Basis on these predictions, what are the business insights and recommendations. Please explain and summarize the various steps performed in this project. There should be proper business interpretation and actionable insights present.

Step 1: EDA

1. Price is the target variable while all other are the predictor's variables. Dataset consists of 26967 rows and 11 columns overall with 6 floats, 1 Integer and 3 object data types. Furthermore, a total of 697 null values has been

observed in depth column and 34 duplicate values which is 0.12% of the total data hence can be dropped. Also, Unnamed seems to be a column of no relevance, hence can be removed from further analysis.

2. Price ranges from 326.0 to 18818.0 whereas length, width and height seems to have '0' as minimum value which is not ideal for our case study needs to be further observed.
3. Heavy skewness has been observed in continuous attributes with all the features are positive or right skewed except 'Depth' which may not be reliable due to presence of null values and 'x' which happens to be slightly left skewed.
4. For most of the features in the sample, outliers are present which indicates that there are extreme values in the dataset.
5. 'Ideal' cut seems to dominate the data with 40% variables followed by color 'G' and clarity 'SI1' with 5653 & 6565 data points respectively.
6. A very strong positive correlation has been observed between Price and carat in such a way that if carat increase, price would also increase for diamond however it is quite evident from Figure that majority of the high-end pricing belongs to carat ranging from 1-2. Similarly, x, y and z also shares a strong positive correlation with pricing but not a linear one. For depth and table, data seems pretty scattered forming a cloud which could be due to lot of noise and/or presence of Multicollinearity.
7. Pricing has been expensive for Fair cut with an average spend of 4579.24 as well as for 'J' color with spend of 5329.70, 'I' happens to be best though. Moreover, SI happens to be the most expensive clarity category for Gems with an average spend of 5088.86 followed by VS and I1 at 3965.69 approx. Please observe that neither of the clarity techniques are best in categorization.

Step 2: Linear Regression and Stats modelling:

Final Linear Regression equation lies as follow:

*(0.0) * Intercept + (1.17) * carat + (0.04) * cut + (0.13) * color + (0.2) * clarity + (0.02) * depth + (-0.01) * table + (-0.41) * x + (0.51) * y + (-0.22) * z*

- When carat increases by one unit, diamond price increases by 1.17 units, keeping all other predictors constant.
- When cut increases by one unit, diamond price increases by 0.13 units, keeping all other predictors constant.
- When color increases by one unit, diamond price increases by 0.2 units, keeping all other predictors constant.
- When clarity increases by one unit, diamond price increases by 0.02 units, keeping all other predictors constant.
- When depth increases by one unit, diamond price increases by 0.02 units, keeping all other predictors constant.
- When table increases by one unit, diamond price decreases by 0.01 units, keeping all other predictors constant.
- When length(x) increases by one unit, diamond price decreases by 0.41 units, keeping all other predictors constant.
- When width (y) increases by one unit, diamond price increases by 0.51 units, keeping all other predictors constant.
- When height (z) increases by one unit, diamond price decreases by 0.22 units, keeping all other predictors constant.
- **Carat, cut, color, clarity and width (y)** happens to be most pivotal attributes for predicting the price.
- Linear relation can be seen between actual Target variable 'y' and predicted variable 'y' however with plethora of data spread which is an indication of noise i.e. unexplained variance in the data.
- As the training and testing set are almost inline with efficiency of 93% approx., we can derive to the fact that model can be treated as right-fit.

Statsmodels:

- VIF (Variance Inflation Factor) helped us to determine the strength of collinearity among predictors which helped us to build several stats models and compare efficiency and p-val for the best fit.
- We can see that R-squared and Adjusted R-square are the same i.e. **0.929** with overall P-value is less than alpha.
- Presented dataset is having length 'x' and height 'z' in cubic mm with negative coefficiency and since the p-val is also less than 0.05 for the attributes, we can conclude that higher the length and height of the diamond, lesser the price will be.
- Nevertheless, width 'y' happens to have a positive co-efficiency and p-val less than 0.05, hence, more the width of gem stone, higher the price will be.

Recommendations:

- The 'Fair Cut' or 'Less Ideal Cut' on diamond has been the most expensive followed by 'Premium' cut hence these should be considered for high profitable stones.
- Price has an increasing trend with width 'y' hence business should focus on stones with higher width.
- Business are recommended to produce gems with less height to increase profitability as great height might lead to dark appearance that could cause less profits.
- Focus should be on producing a flat diamond comparatively with less 'Table' for appropriate direction of light as this will lead to more profits.
- Claimed clarity technique 'IF' as best seems to be generating comparatively less price hence it is recommended for Business to relook the prices which might elevate the profits.

Problem 2: Package Prediction

Executive Summary

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

2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.

Snapshot of DataFrame

Table 11: Holiday Package Dataset

	Unnamed: 0	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	1	no	48412	30	8	1	1	no
1	2	yes	37207	45	8	0	1	no
2	3	no	58022	46	9	0	0	no
3	4	no	66503	31	11	2	0	no
4	5	no	66734	44	12	0	2	no

In the given dataset, column 'Unnamed: 0' seems to be of no us hence for the further analysis, we are dropping the particular column from the sample.

After dropping the feature, following dataset has been obtained:

	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

Understanding Data and Missing Values

Column	Non-Null Count	Dtype
Holliday_Package	872 non-null	object
Salary	872 non-null	int64
age	872 non-null	int64
educ	872 non-null	int64
no_young_children	872 non-null	int64
no_older_children	872 non-null	int64
foreign	872 non-null	object

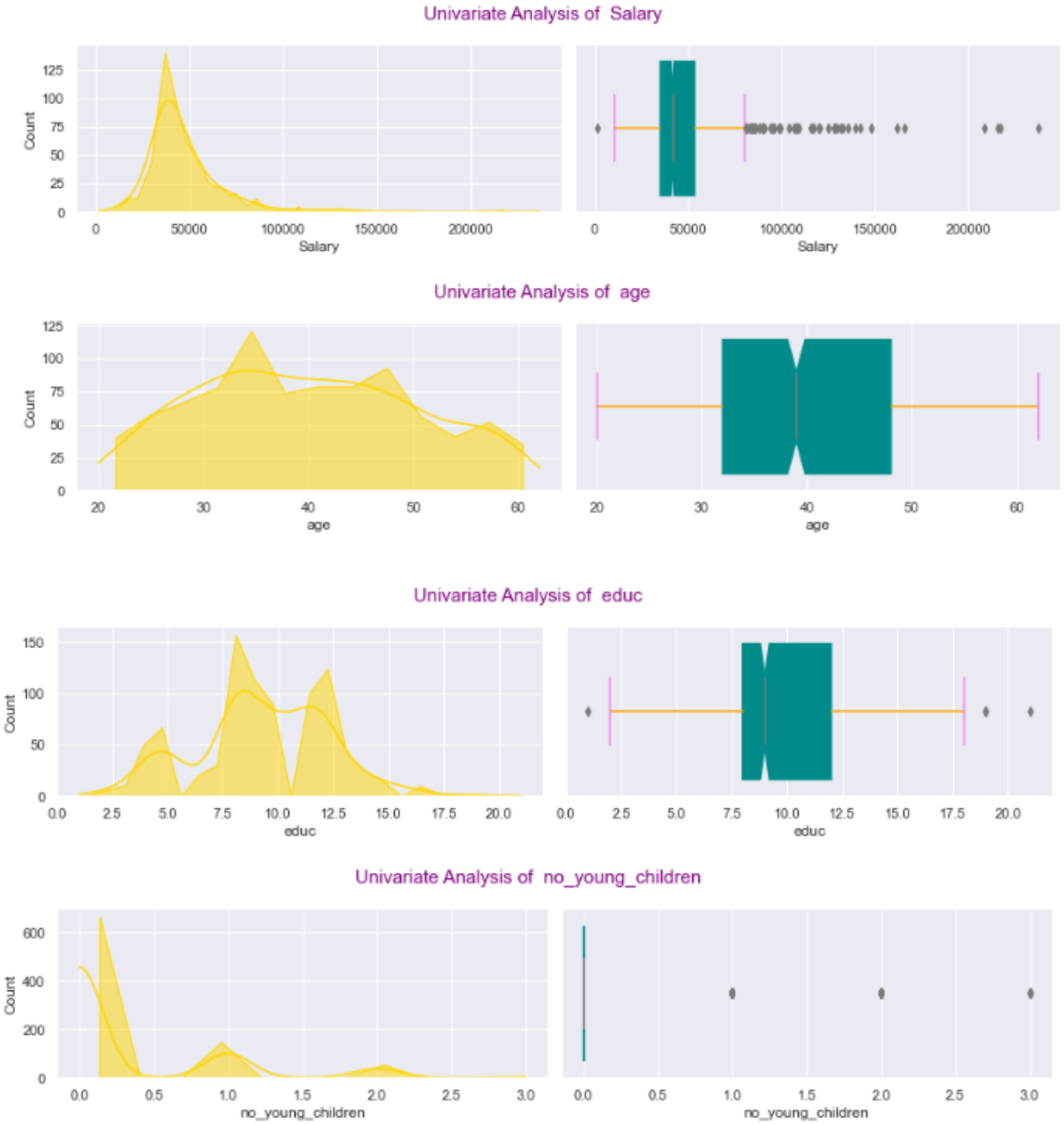
Table 12: Descriptive Summary of Continuous and Categorical variables

	Salary	age	educ	no_young_children	no_older_children
count	872.00	872.00	872.00	872.00	872.00
mean	47729.17	39.96	9.31	0.31	0.98
std	23418.67	10.55	3.04	0.61	1.09
min	1322.00	20.00	1.00	0.00	0.00
25%	35324.00	32.00	8.00	0.00	0.00
50%	41903.50	39.00	9.00	0.00	1.00
75%	53469.50	48.00	12.00	0.00	2.00
max	236961.00	62.00	21.00	3.00	6.00

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

Univariate Analysis

Figure 12: Univariate Analysis of Continuous variables



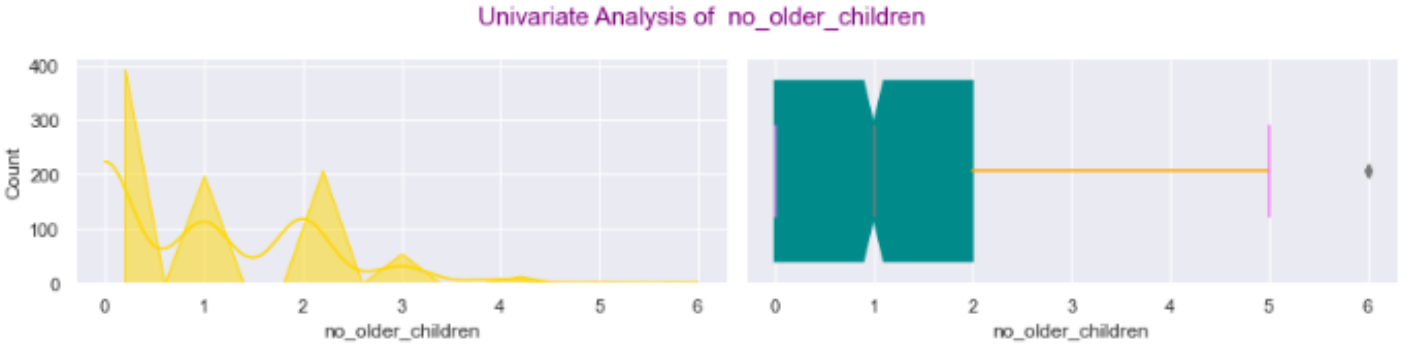
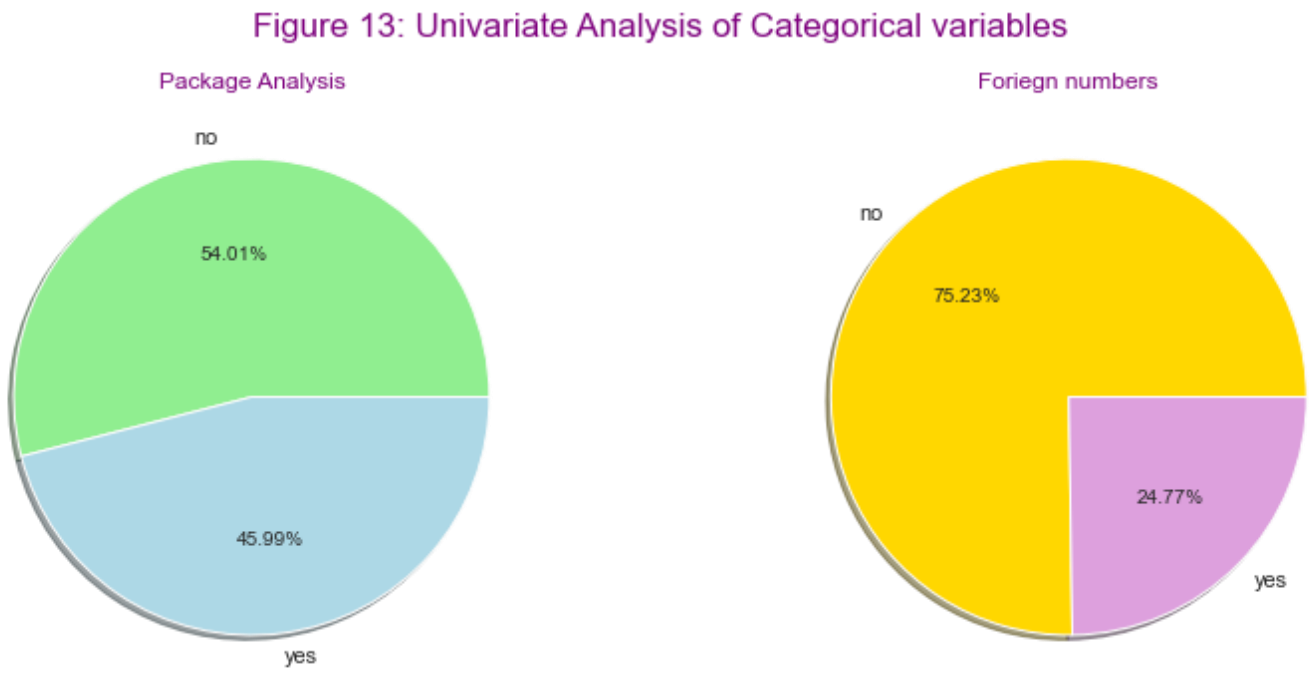
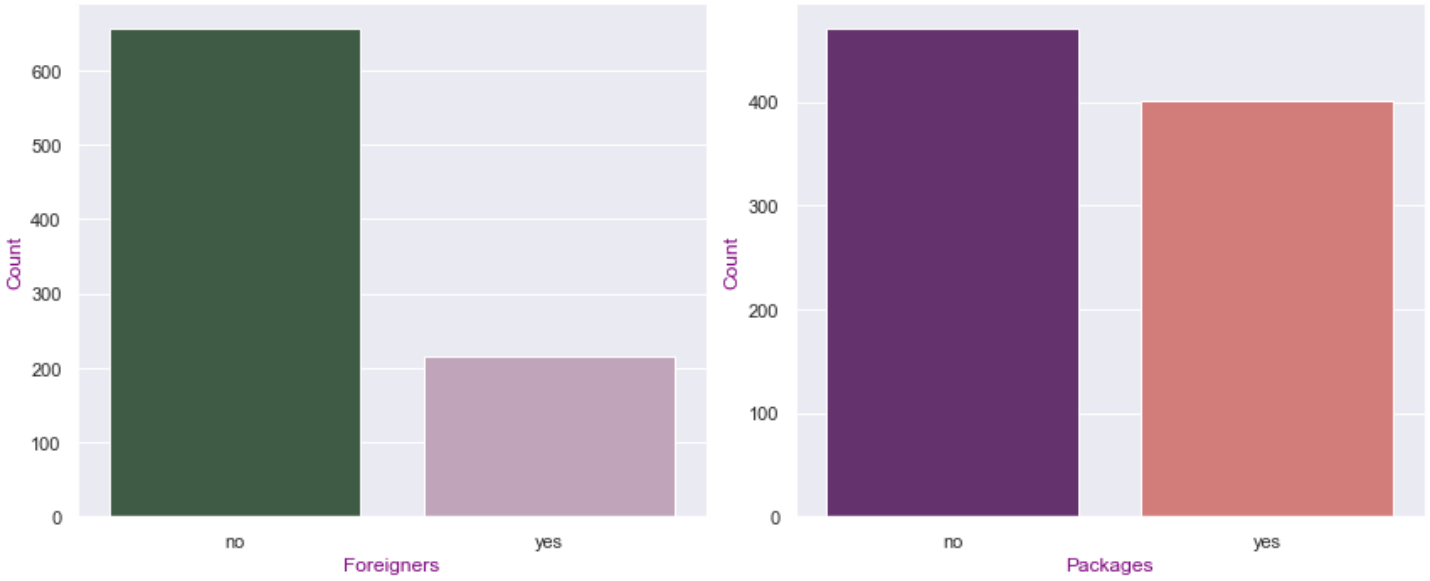


Table 13: Skewness

	Salary	Age	Education	Older Children	Younger Children
Skewness	3.1	0.15	-0.05	0.95	1.94





- Data seems to be quite imbalanced for both categorical variables. Employees, majority of which are non-foreigners, tend to be more biased towards not opting packages.

Bi-Variate Analysis

Figure 14: Correlation Matrix

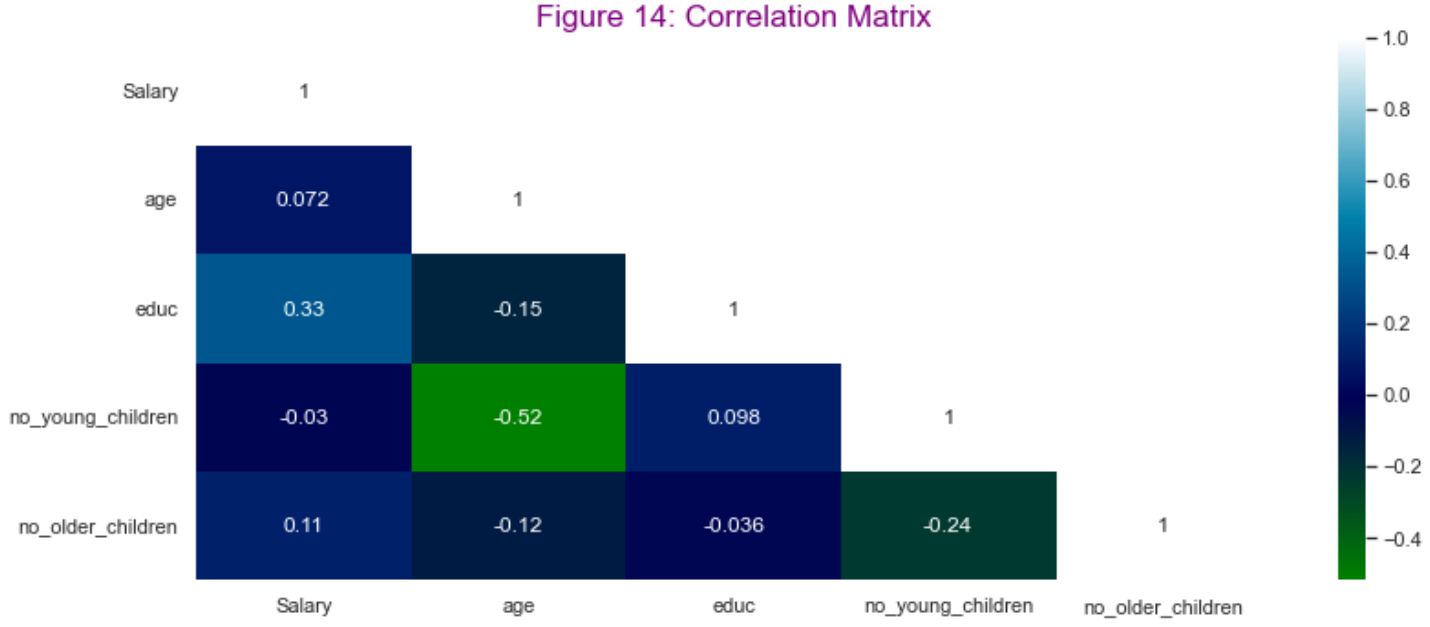


Figure 15: Pairplot Analysis of variables w.r.t Package

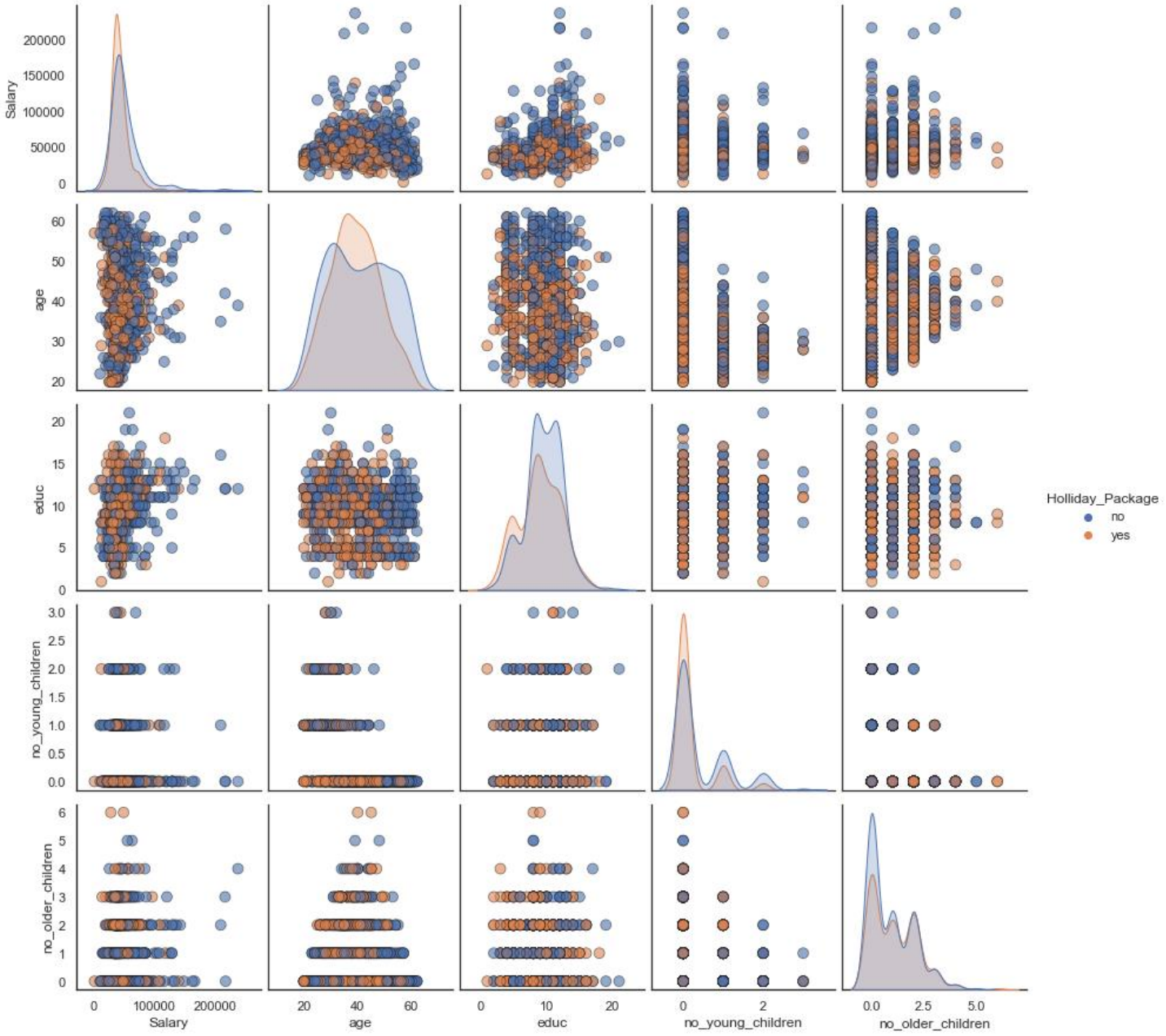


Figure 16: Pairplot Analysis of variables w.r.t. Foreigners

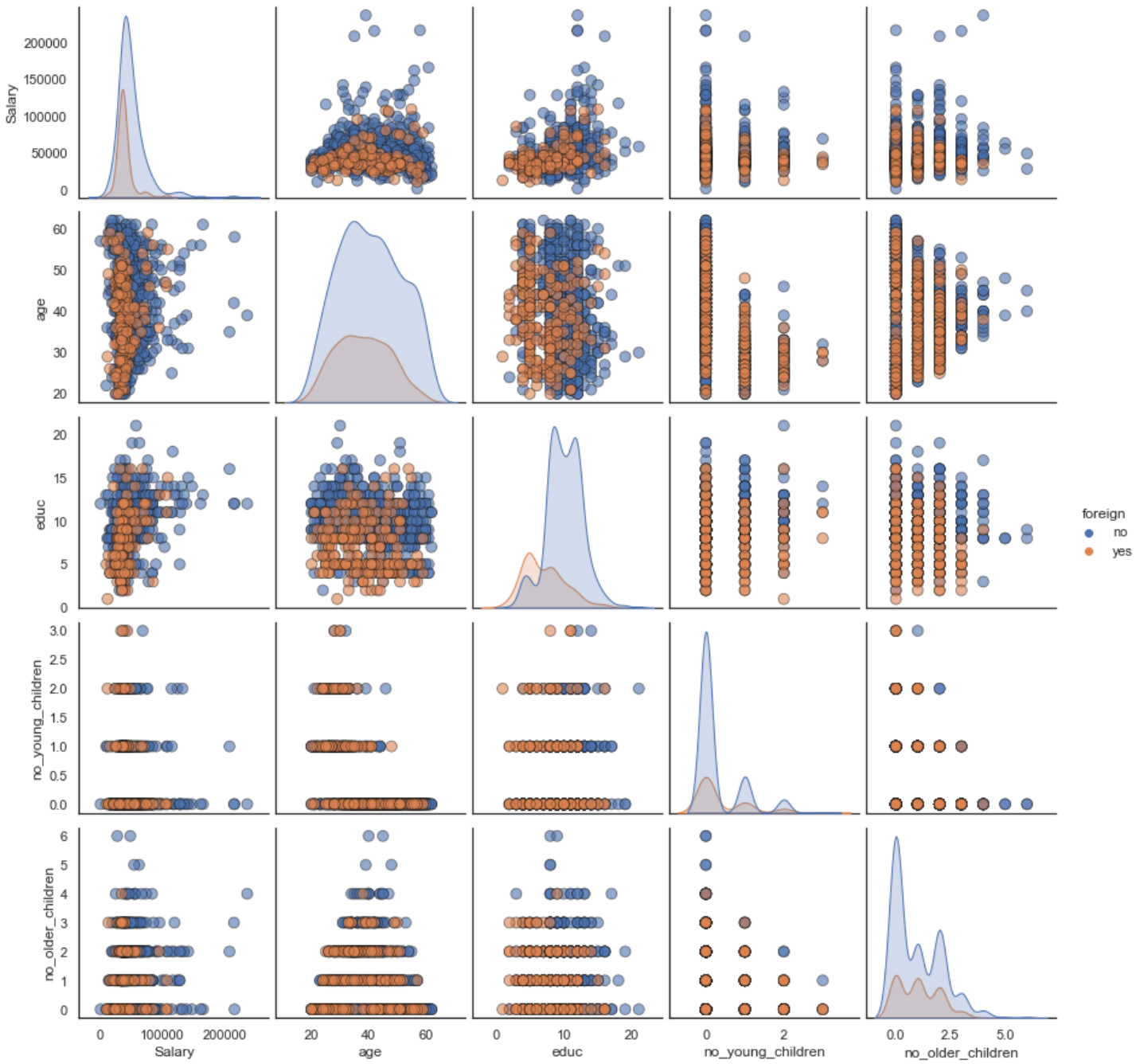


Figure 17: Bi-Variate Analysis of Holiday Package

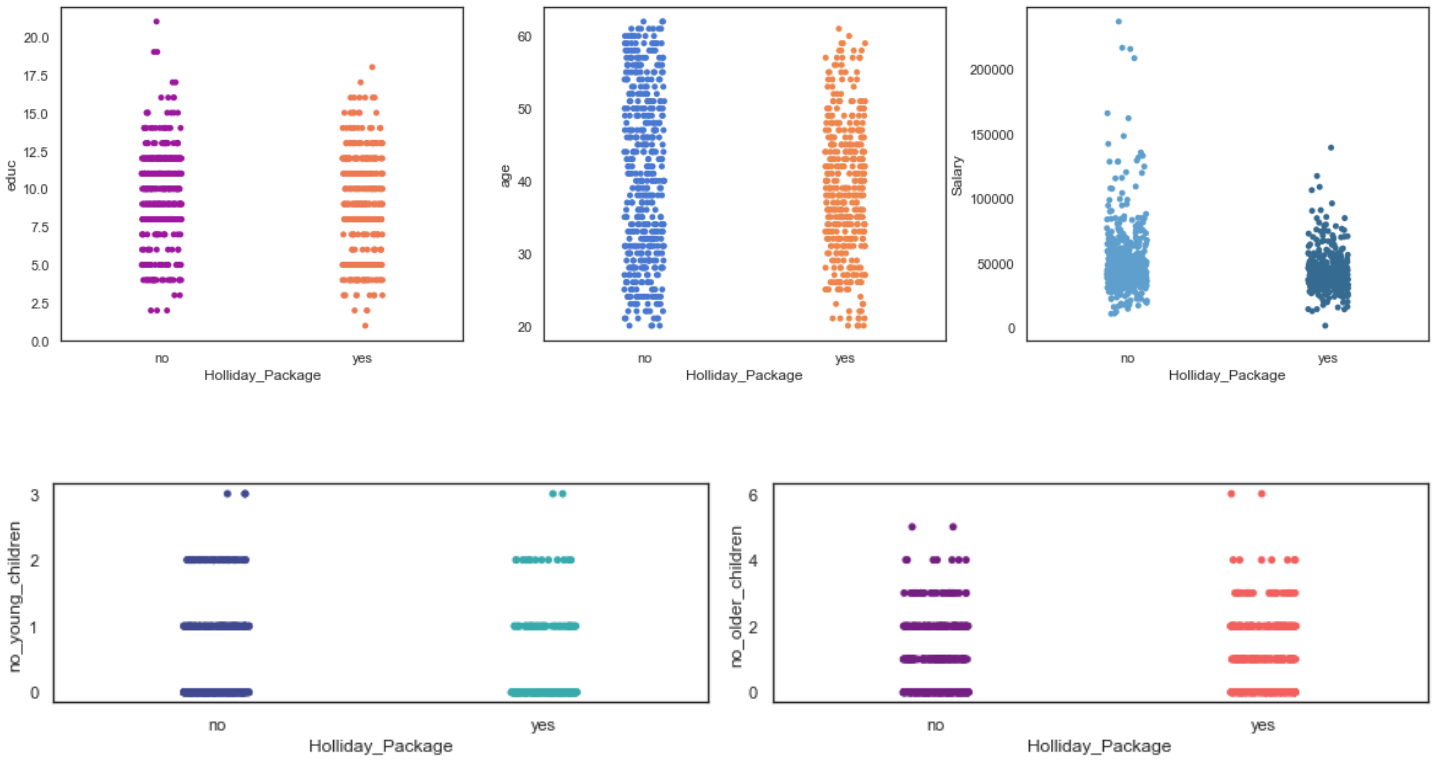


Figure 18: Bi-Variate Analysis of Foreign Class

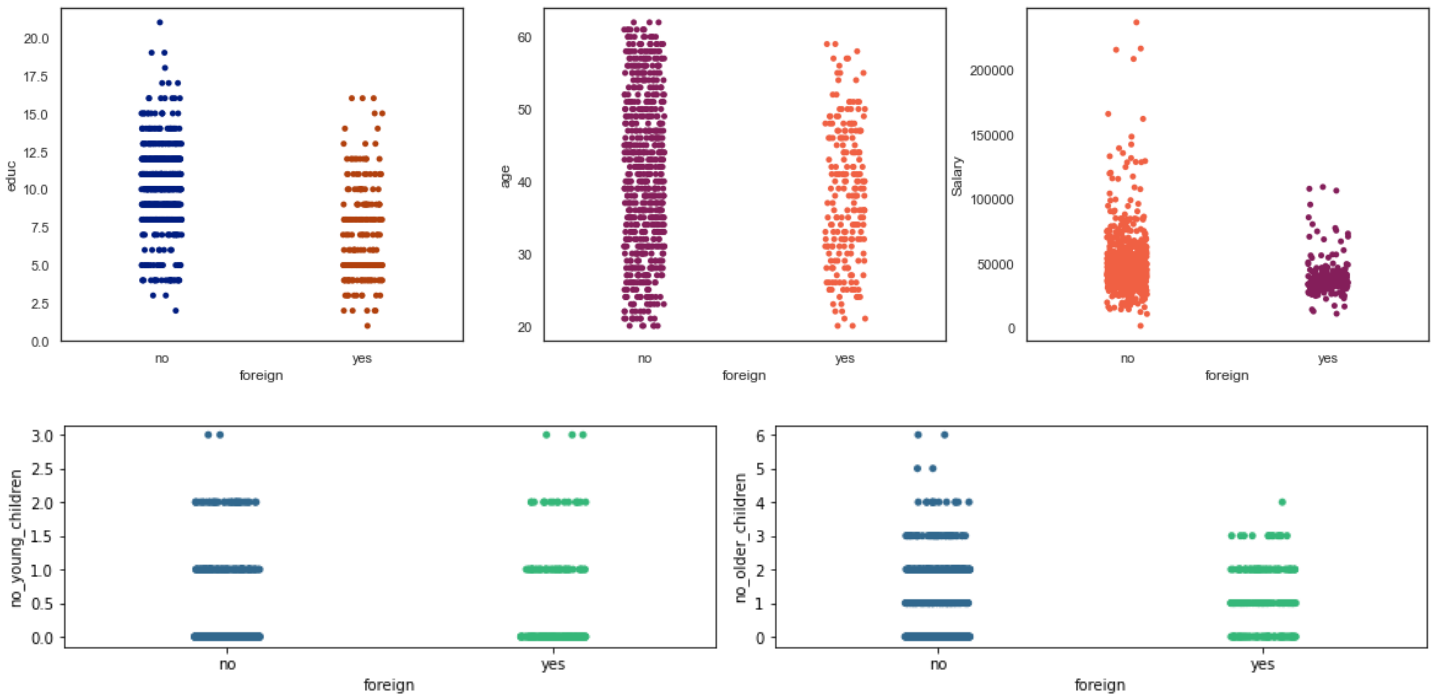


Figure 19: Class Analysis

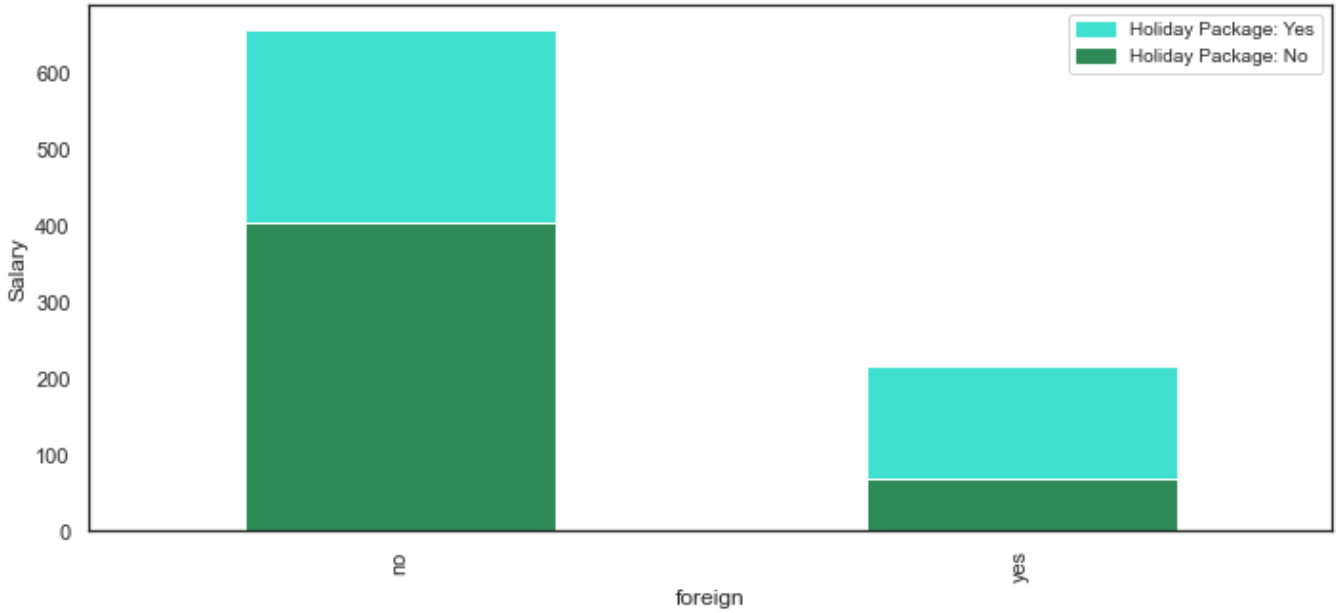


Figure 20: Class Analysis of Salary

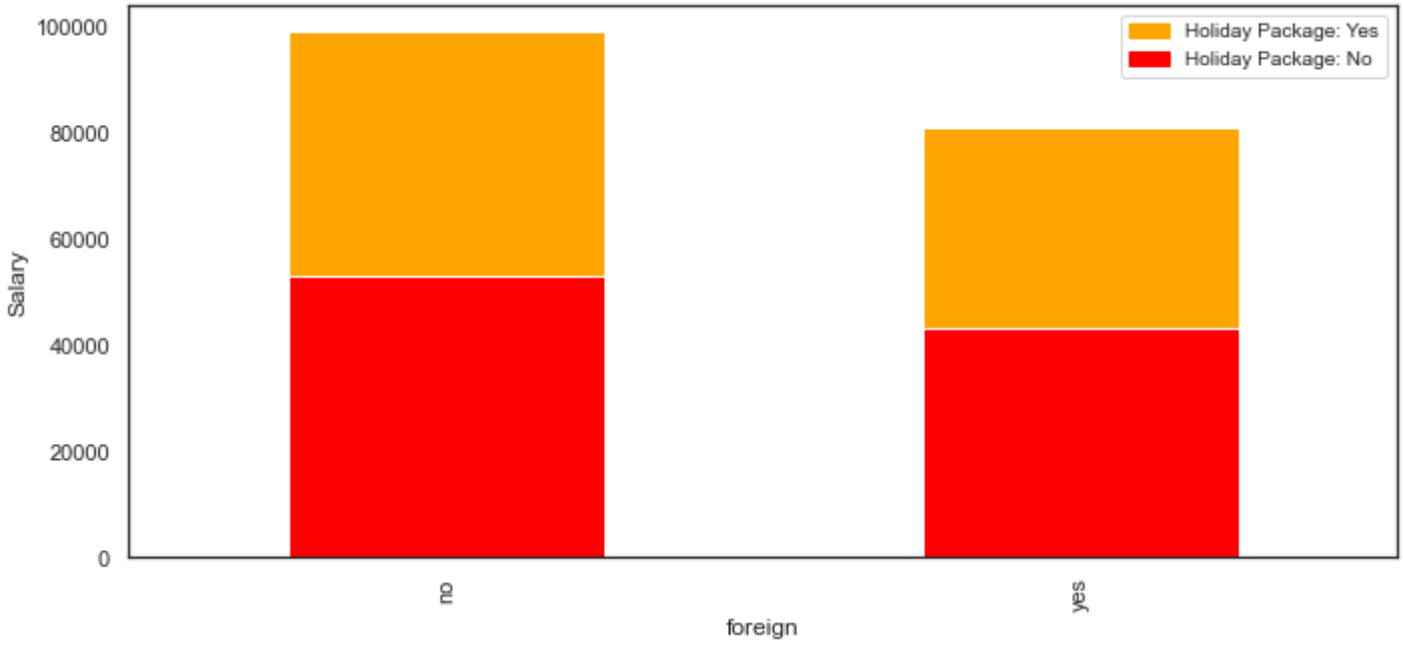


Figure 21: Class Analysis of Education

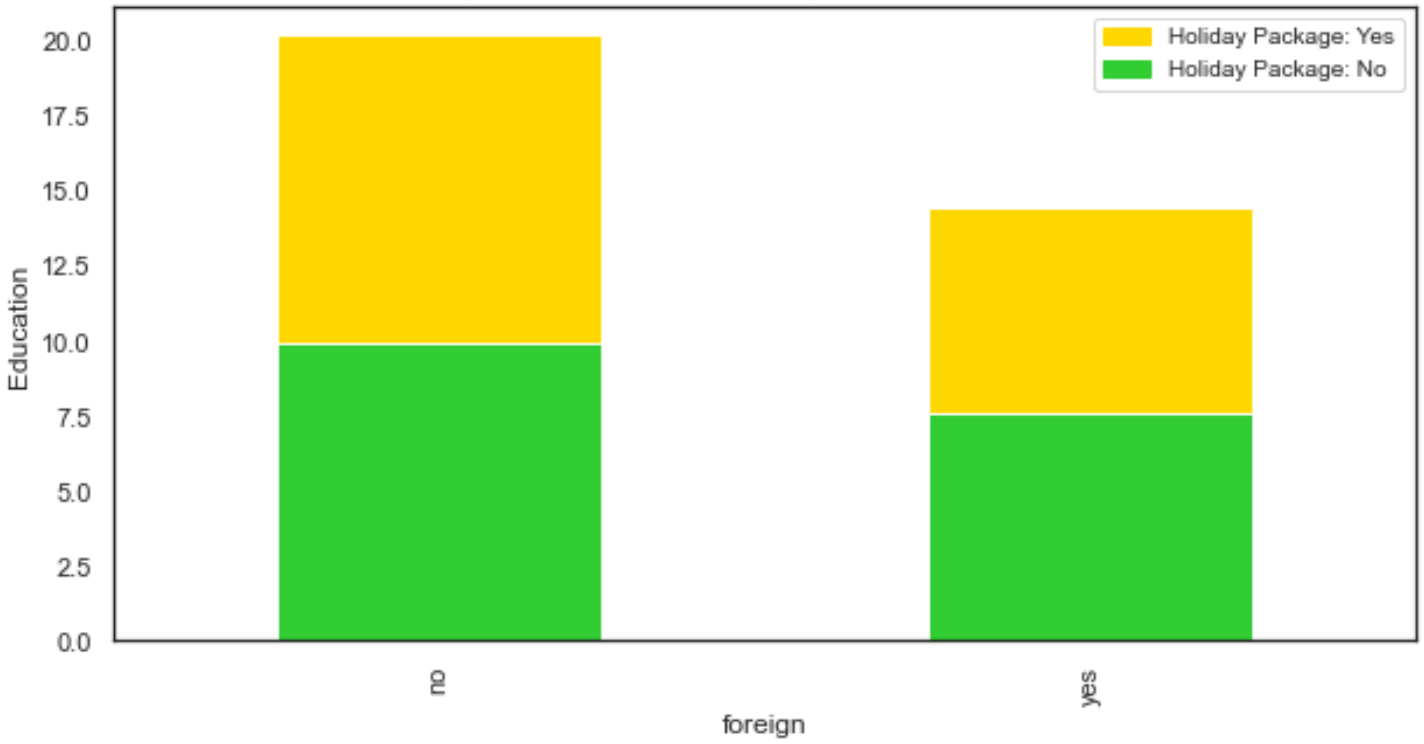
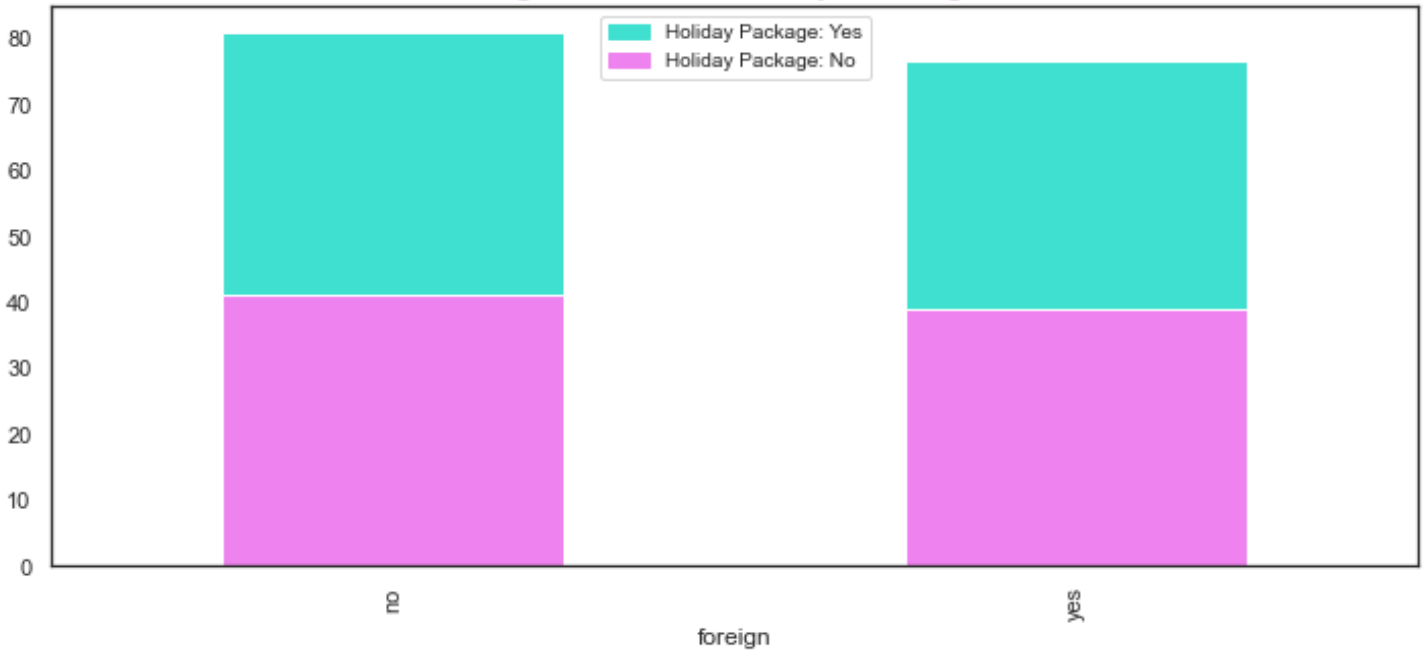


Figure 22: Class Analysis of Age



Inference:

- The above dataset consists of 872 observations and 8 features, six of them are of integer data type and 2 are of type object. There are no null values in the data with zero duplicates are present.
- Descriptive summary suggests that Age of employees ranges from 20 till 62 with an average age of 40, employees may or may not have enrolled in packages with most of employees (75%) considered are of age 48. Similarly, Salary of employees ranges from 1322 till 236961 units with average Salary being 47729.17 however mostly belongs in the slab of 53469.5 units. Furthermore, maximum number of younger children belonging are 3 are older children being 6.
- Holiday Packages seems to be quite imbalanced with 471 employees did not opt for packages and 401 did opt. In other words, 54% of our sample population choose to refuse the package and 45% are enrolled. Likewise, Probability of being a foreigner is 75% with 656 employees and not being a foreigner is 24% with 216 employees.
- Salary and number of Younger children seems to be heavily right skewed with majority of outliers above upper whisker whereas Age and Education seems to be normally distributed with minimal outliers i.e., there is one unique employee who had 21 years of formal education and is not a foreigner as well as not opted for any package. On the other hand, outlier below lower whisker suggest that there is one employee which stands out with only 1 year of formal education and is a foreigner as well as opted package. A lot of asymmetry has been observed with number of Older Children though it seems light skewed.
- Negative correlation has been noticed between independent variables except Education and Salary but with less collinearity. Predictors seems to be overlapping w.r.t. packages and foreign class which refers to the fact that data points are not able to discriminate well between classes and thus the probability of finding target variable will be close to 50% i.e. data points can be considered as weak predictors. For most of the relations, data seems to quite scattered with almost a cloud formation which could be due to unexplained variance.
- With more Age, probability seems to be quite less for opting package as well as being a foreigner. Average employee with Age 41 has chosen to opt and is not a foreigner whereas 37 seems to be not interested who's also a foreigner. Uniformly, with high Salary, probability of opting package is low. Employees who are non-foreigners with average salary of 53227 has chosen to refuse packages and employees who are foreigners with mean Salary of 37865 has opted for package. For Education, foreign employees with lowest tenure of formal education has opted and those with highest, tends to decline and/or non-foreign employees with average tenure of 10 years has refused for packages and with average tenure of 7 years seems to be interested.
- Holiday packages and Foreign class doesn't seems to be much impacted with number of younger children however probability of opting is more with more number of older children present.

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).

Table 14: Data Types

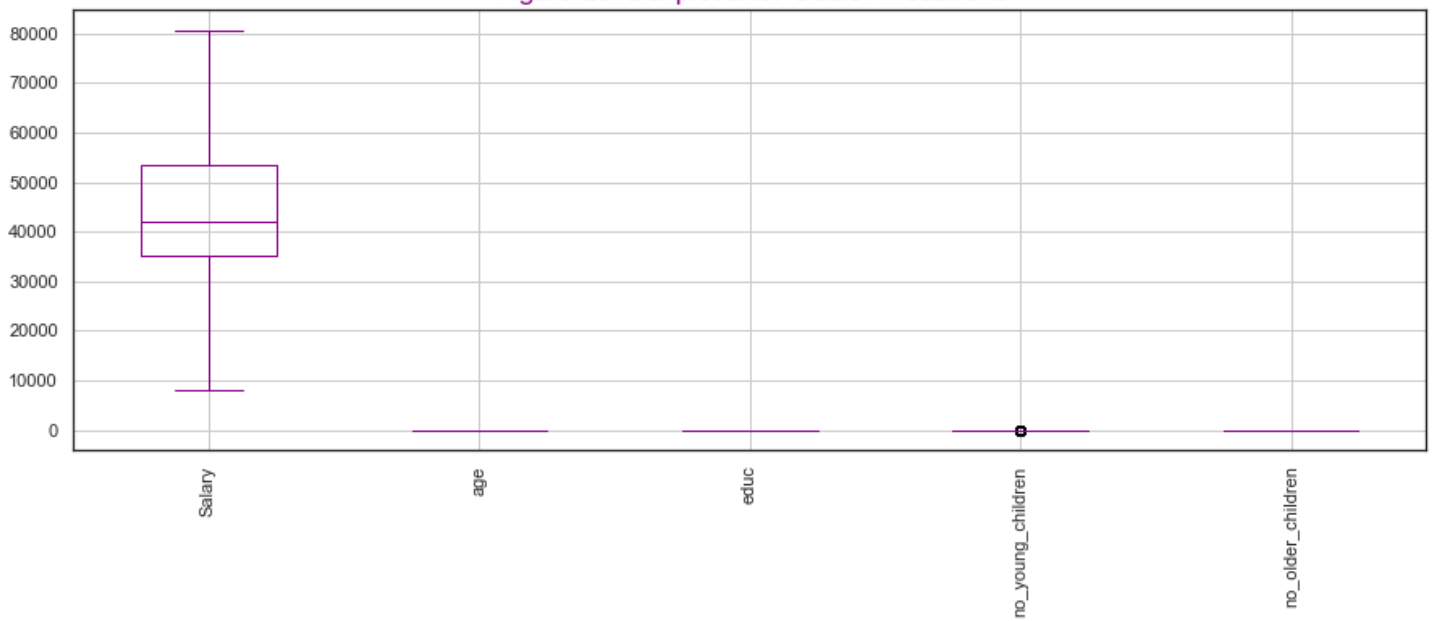
	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	object	int64	int64	int64	int64	int64	object

- From Table 14, we can behold a total of 2 objects variable which can be encoded or converted to type integer.

Outlier Treatment

Several outliers can be seen in sample features which can be treated before modelling using IQR technique. However, for this analysis, we will refrain from removing any outlier from feature **'no_young_children'** as majority of the data points defining the column are outliers hence removing them might leave us with no data to evaluate.

Figure 23: Boxplot after Outlier Treatment



- As we can notice, there are no outliers present except no_young_children.

Train-Test-Split

Further, splitting () has been applied keeping **30% as testing dataset** and following variables has been generated:

1. Xtrain – Training dataset without target variable
2. Xtest – Testing dataset without target variable
3. Ytrain – Training dataset with target variable which needs to be predicted
4. Ytest - Training dataset with target variable which needs to be predicted

Model 1 – Logistic Regression

Logistic Regression, also known as 'Logit' and/or Maximum-Entropy classifier, is another method of supervised learning for classification. Unlike Linear Regression, it accepts dependent variable as binary categories, however as integer type.

Logit assigns probabilities to different classes to which a particular data point is likely to belong.

It establishes relation between dependent and independent variables using regression with linear combination as follows:

$$Z = w.x + b$$

Wherein Z is our independent variable, w_i is the optimal weight assigned to an input feature which is determined using **'cross-entropy loss function'** followed by the approach of Gradient Descent, representing how pivotal the feature is for classification.

W is positive - Direct correlation with the class of interest

W is negative – Inverse relation with the class of interest

X being independent variable with b is bias (error).

Sigmoid Function()

Since weights and bias tends to be running numbers hence Z ranges from -∞ to + ∞ and needs to be converted into probability. Sigmoid function forms a s-curve assigning a threshold value, defining probability of either 0 or 1 and can be defined as:

$$P(Y = 1) = \text{Sigmoid}(Z) = \frac{1}{1+ e^{-z}}$$

$$P(Y = 0) = 1 – P(Y = 1)$$

Step 1: Calling Logit model and fitting the data

Logit model has been called using the below tuning parameters fitting where loss function and sigmoid has been utilized to form best s-curve:

Random_state: 1

Max_Iter: 10000

N_jobs = 2

Step 2: Checking Accuracy Scores

After applying sigmoid and cross-entropy to fit the training data, following scores can be observed:

Training Score is : 0.53

Testing Score is : 0.55

Step 3: Regularizing Model

With accuracy of 53 and 55%, models seems to be under-fit and seems of no usage in the production. However, GridSearchCV can be applied to perform an exhaustive search over specified parameter values for an estimator.

Note – Values provided inside lists can be relooked in case testing models generates less effective results.

Tuning Parameters used in GridSearchCV for Logit Model:

1. **Penalty:** It shrinks the coefficients of those dependent variables which are less contributive toward zero. This helps in optimizing our model for efficiency.
2. **Solver:** This helps in providing a weight variable to each record and fitting the Logit regression intercept. Variable with most weight record means to influence the model.
3. **Tolerance:** Threshold for the optimization.
4. **Cross-validation (CV)** - Helps in avoiding over-fitting and greedy nature of model by following a k-fold cross validation wherein K = number of times we want to run a model via different iterations.

Step 4: Fitting model on GridSearchCV

Training set needs to be fit again on GridSearchCV for generating best parameters that can be used for further model building.

Best_estimator_ is an effective function used inside GridSearchCV which helps in determining the best parameters out of list provided and can be used further.

After applying best_estimator() with cv = 3, following values has been obtained:

- **max_iter** = 10000
- **n_jobs** = 2
- **random_state** = 1
- **solver** = 'liblinear'
- **tol** = 0.00001

Step 5: Predicting on Train and Test Dataset

Table 15: Logit Probabilities of Test Data

	0	1	2	3	4
0	0.73	0.51	0.81	0.92	0.38
1	0.27	0.49	0.19	0.08	0.62

Table 15 is the sample of 5 observations depicting the Logit capability of defining probabilities and assigning to each record to classify predicting class. For eg: Record 0 has a probability of 0.73 of belonging to class 0 and 0.27 of belonging to class 1.

Model 2 – Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is another classification method used in Supervised Learning technique to predict observations where classes are **known and fixed.** It also utilizes approach of forming a linear combination and observing data from low-dimensional space by maximizing the between class scatter and minimize class scatter.

LDA constructs linear equation which minimizes the possibility of misclassification of cases into their respective classes i.e. building a best fit line which separates the 0 and 1 with least chances of classes being wrongly classified to their respective class.

It can be also be used as a dimensionality reduction technique by focusing on maximizing the separability of known classes in the target variable.

Step 1: Calling LDA model and fitting into trained data

LDA model has been called using the below tuning parameters to form a best line for separating probabilities:

- Solver = 'svd'
- Tol = 0.00001
- Shrinkage = None
- N_components = None

Step 2: Applying GridSearchCV and fitting grid_search with estimator

LDA uses a pre-defined approach of using parameters that are already defined in a way to predict model with best efficiency however we have decided to tweak our threshold value with grid search using **cv = 5**, which in result, provided us a value of '0.00001' which we will be using for further analysis.

Step 3: Predicting Trained and Test dataset

Similar to Logit, LDA also separates records with probabilities of belonging to a class (target) variable.

Table 16: LDA probabilities of Test Data

	0	1	2	3	4
0	0.73	0.47	0.76	0.93	0.4
1	0.27	0.53	0.24	0.07	0.6

Table 16 depicts a sample of 5 observations with probabilities of belonging to a specific class. As one can notice, record 0, 2 and 3 are completely biased towards class 0.

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.

Accuracy

For Logistic Regression model -

- Accuracy for Trained Data is: 0.66
- Accuracy for Test Data is: 0.67

For LDA Regression model -

- Accuracy for Trained Data is: 0.62
- Accuracy for Test Data is: 0.66

Confusion Matrix and Classification Report for Logistic Regression Model

Table 17: Classification report for Logit Trained Data

	precision	recall	f1-score	support
no	0.66	0.75	0.70	326
yes	0.66	0.55	0.60	284
accuracy			0.66	610
macro avg	0.66	0.65	0.65	610
weighted avg	0.66	0.66	0.65	610

Figure 24: Confusion Matrix of Trained Logistic model

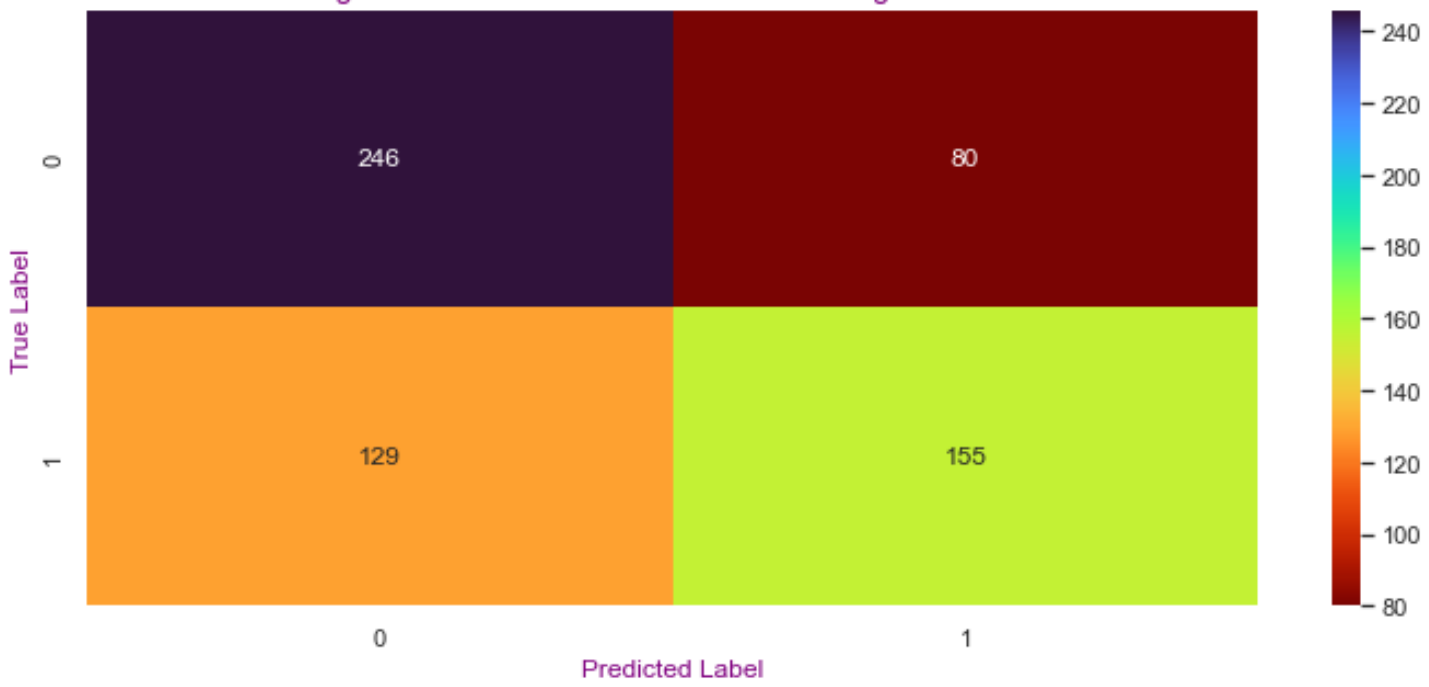
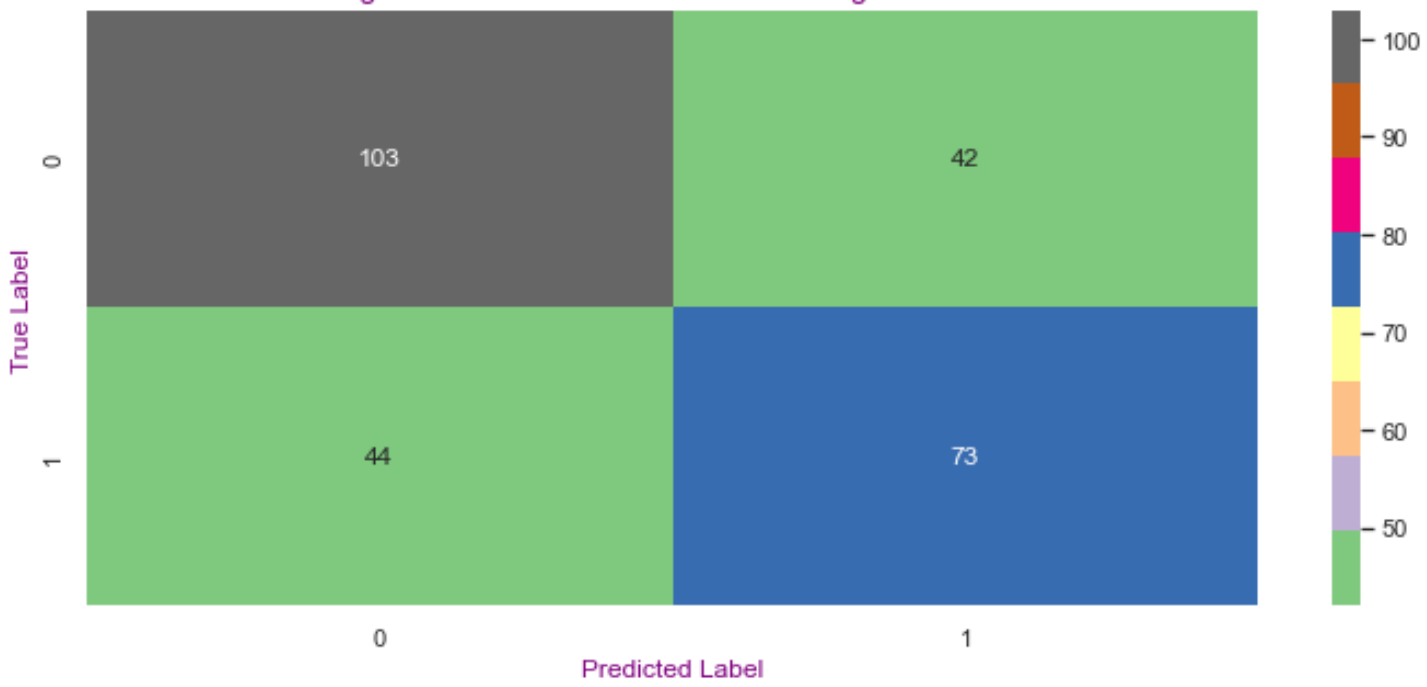


Table 18: Classification report for Logit Test Data

	precision	recall	f1-score	support
no	0.70	0.71	0.71	145
yes	0.63	0.62	0.63	117
accuracy			0.67	262
macro avg	0.67	0.67	0.67	262
weighted avg	0.67	0.67	0.67	262

Figure 25: Confusion Matrix of Test Logistic model



Confusion Matrix and Classification Report for Linear Discriminant Analysis

Table 19: Classification report for LDA Trained Data

	precision	recall	f1-score	support
no	0.64	0.68	0.66	326
yes	0.60	0.55	0.58	284
accuracy			0.62	610
macro avg	0.62	0.62	0.62	610
weighted avg	0.62	0.62	0.62	610

Figure 26: Confusion Matrix of Trained LDA model

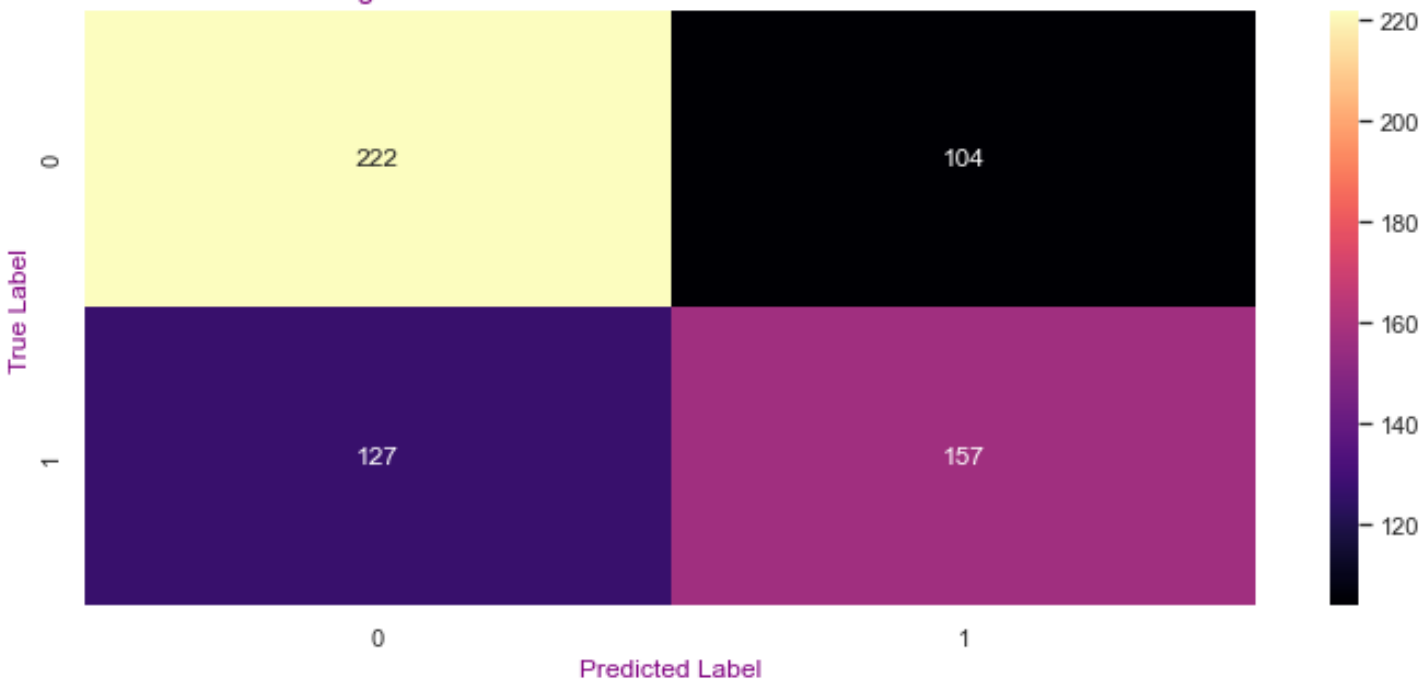
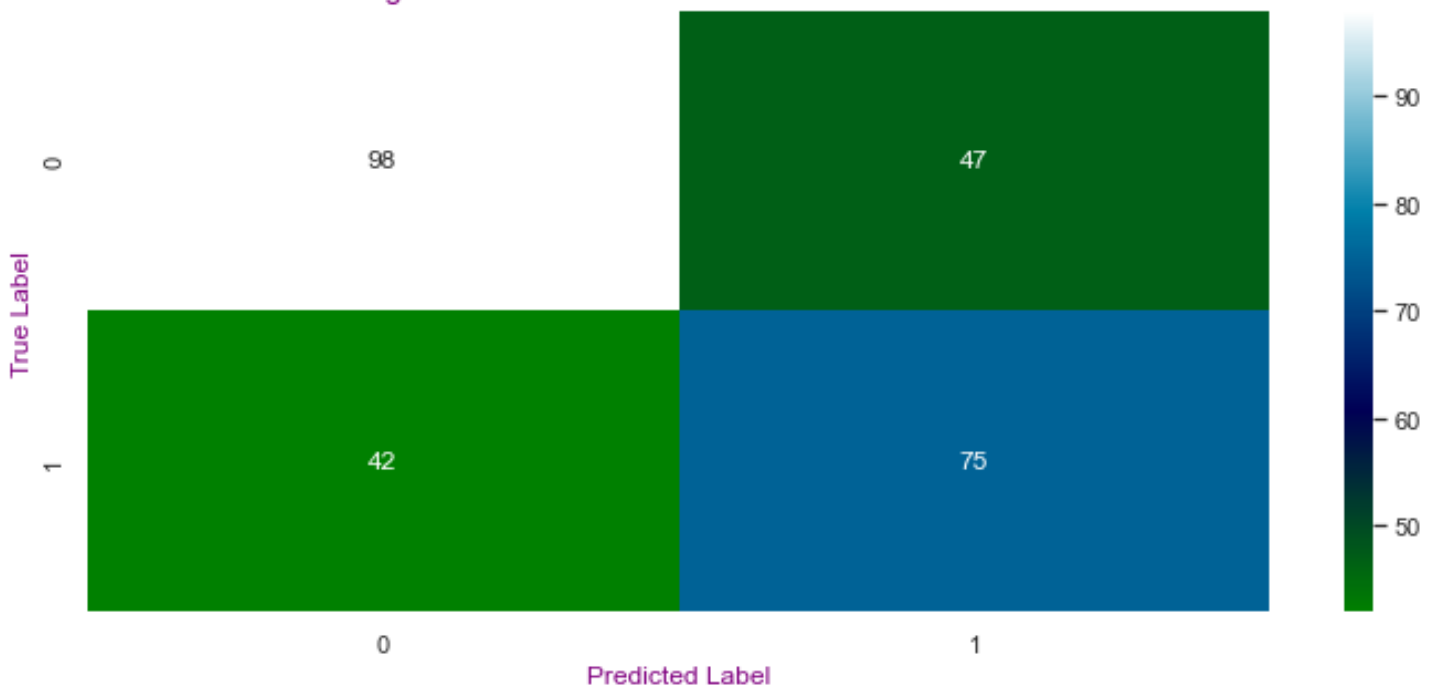


Table 20: Classification Report for LDA Test Data

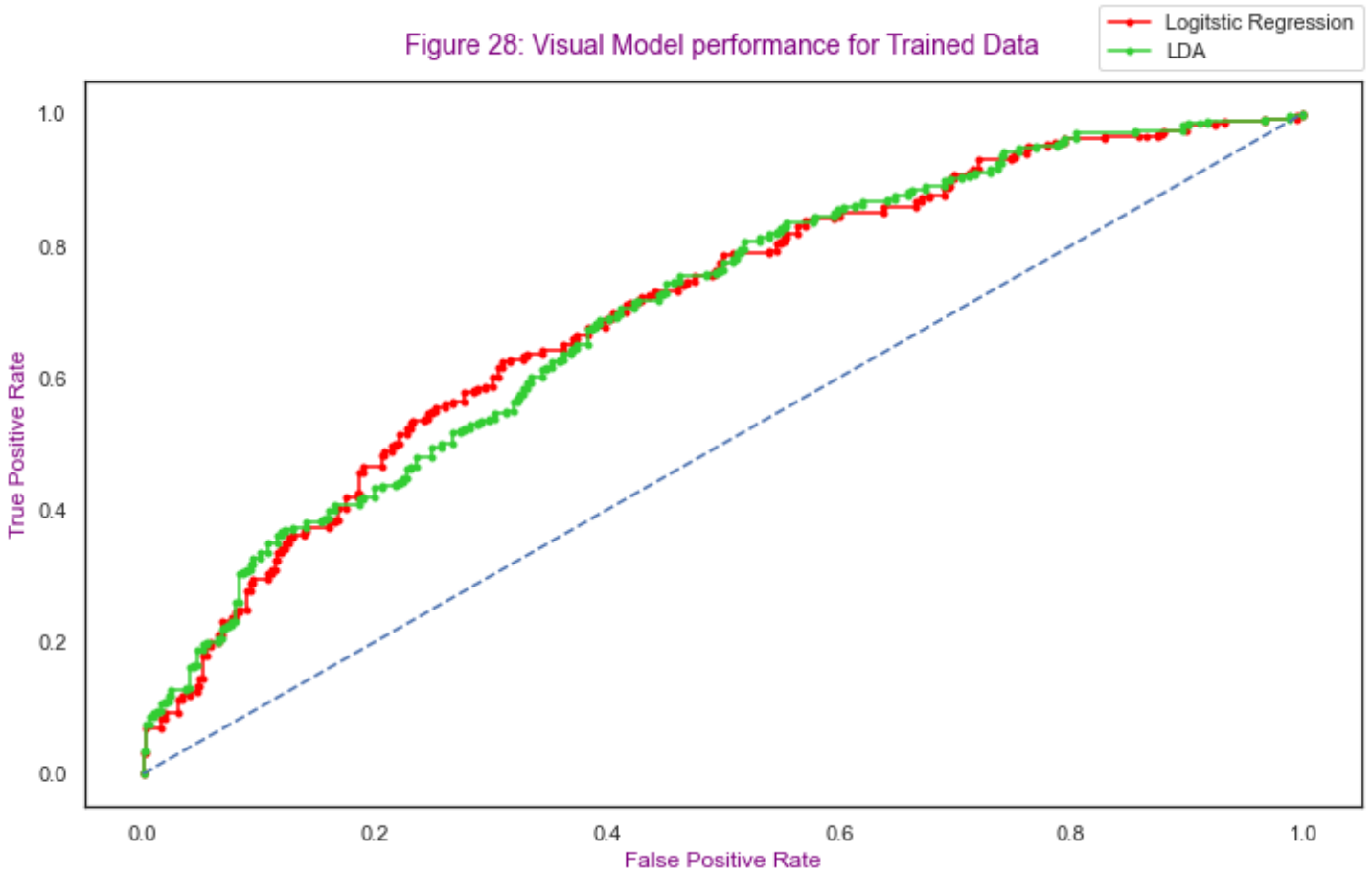
	precision	recall	f1-score	support
no	0.70	0.68	0.69	145
yes	0.61	0.64	0.63	117
accuracy			0.66	262
macro avg	0.66	0.66	0.66	262
weighted avg	0.66	0.66	0.66	262

Figure 27: Confusion Matrix of Test LDA model



ROC/AUC for Trained Data

Figure 28: Visual Model performance for Trained Data



- Area under the curve for Trained Logistic Regression Model is **70.22 %**
- Area under the curve for Trained LDA Model is **69.91 %**

ROC/AUC for Test Data

- Area under the curve for Test Logistic Regression Model is **68.64 %**
- Area under the curve for Test LDA Model is **68.75 %**

Model Comparison

Table 21: Model comparison

	Accuracy	AUC	Recall	Precision	F1 Score
LOGIT Train	0.66	70.22	0.55	0.66	0.60
LOGIT Test	0.67	68.64	0.62	0.63	0.63
LDA Train	0.62	69.91	0.55	0.60	0.58
LDA Test	0.66	68.75	0.64	0.61	0.63

Observations:

- Area under the curve of training set for Logistic Regression and LDA models stands at 70.22% and 69.91% respectively which indicates that to some extent, classes has been correctly classified. However, for test set, Logistic regression model has relatively minimal yet low AUC score of 68.64% and similarly, LDA has a score of 68.75% which indicates that model is under-fit. Also, for test and trained both, curve seems to be far from y-axis which shows the liberal nature of the model due to higher True positive rates.
- Overall, Testing set has performed better for both the models with Logistic Regression a score of 67%. Accuracy score for LDA also stands at 62% for Training and 66% Testing set.
- F1 score seems to be lower for both models however precision seems to be higher for trained and lower for LDA Test data. This could be due to imbalanced class and less True positives has been predicted.
- Recall seems to be higher for Testing data for both the models. LDA seems to have more Recall than Logistic Regression which illustrates the low False positives that has been predicted i.e. 42.
- Overall, model found successful in predicting class 0 more efficiently than class 1 since precision as well as recall seems to be higher for the same.
- From the analysis, it is safe to presume that LDA model seems to be more efficient as the model seems to be predicting less False positives hence high recall score for class 1 with almost identical f1 score and slightly higher AUC score i.e. less under-fit compared to Logistics model.

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

Insights:

For Predicting Class 0:

Precision: For both the models, 70% of the employees have been predicted for not opting Holiday Package out of all employees predicted to have opted.

Recall: Out of all the employees who have not opted for any package, 71% has been predicted correctly for Logistic model and 68% for LDA model.

For Predicting Class 1

Precision: Out of all the employees predicted to have opted for packages, 63% has been predicted for actually opting for Holiday Package under Logistic model and 61% under LDA.

Recall: Under Logistic Regression, 62% of the employee predicted to have actually opted for Holiday Package out of all the employees who actually opted for Package. Similarly, 64% has been predicted correctly under LDA.

- Overall, Logistic model happens to be 67% accurate and LDA seems to be 66% accurate.

Recommendations:

- Given dataset seems to be asymmetric with extreme values present. Package as well as foreigner frequency seems to be imbalanced as data speaks more about the employees who has not opted for package and/or are not foreigners. Collecting more data may help in balancing the overall model by reducing the overlapping of data and generating more efficient results.
- Salary with high tenure seems to be increasing however with higher occupancy, probability of opting package is low. Similarly, majority of the employees with salary more than 50000, happens to be more inclined towards not opting the package. Lucrative discounts can be leveraged by Business and/or custom packages can be made as per customer's preference/needs may drum up more sales.
- Business might want to dig deeper with employees who are not foreigners which happens to be 75% of our analysis and have high probability of not opting for Holiday package including those with high tenure, age and salary. Analysing their travelling trends, Special Plans can be offered to such customers for the holiday season.
- Separate Packages or Individualized package can be offered on per-project basis i.e. monthly packages can be set with fixed costing for those who are frequent travellers however not opting for packages. Additionally, add-ons might assist in attracting new employees with special monitoring, especially with employees who are tenured and high-salaried.

----------THE END ----------