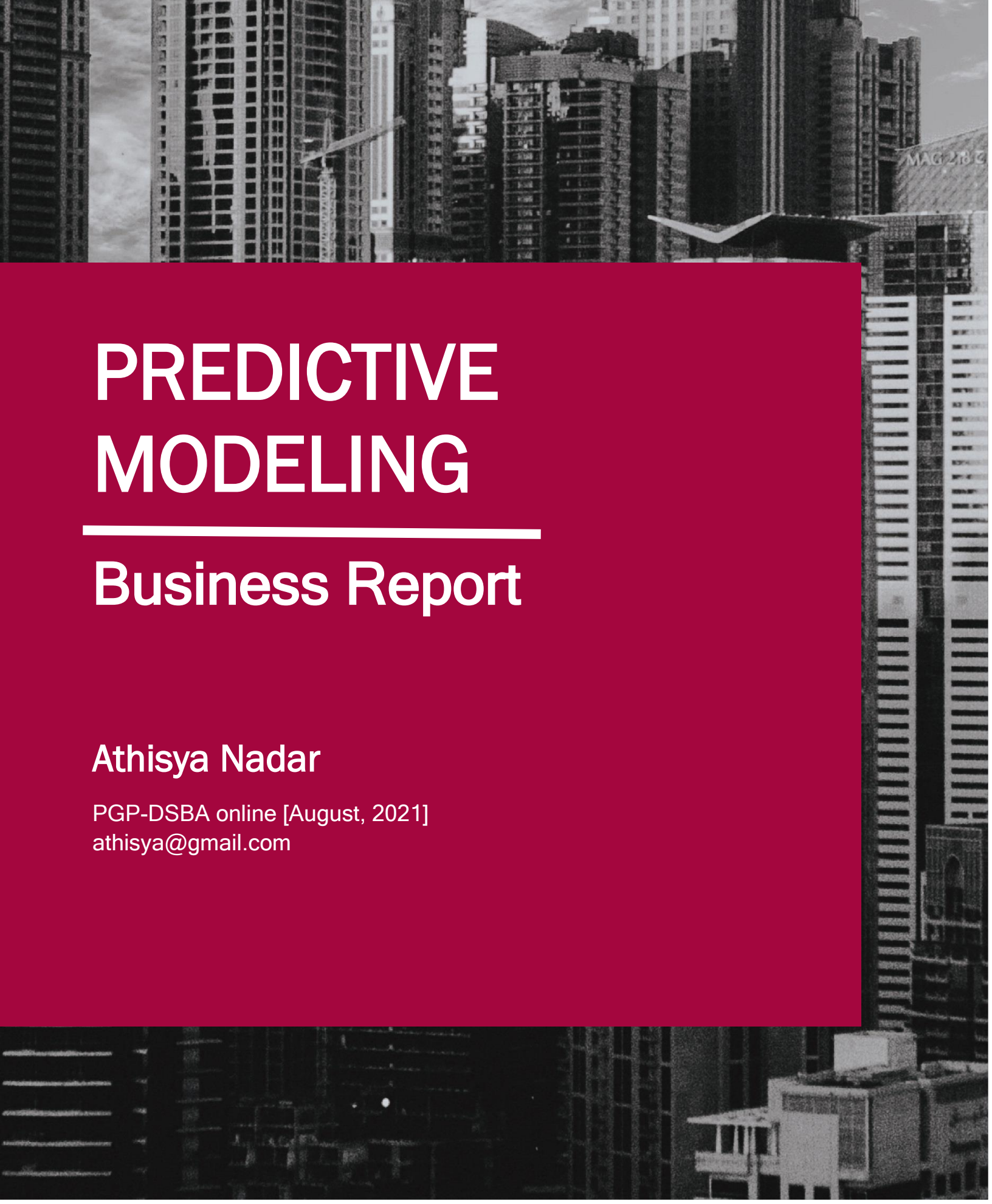


PREDICTIVE MODELING

Business Report

Athisya Nadar

PGP-DSBA online [August, 2021]
athisya@gmail.com

Table of Contents

Table of Figures: .....

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 worst and J the best.
Clarity	cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, IF = flawless, l1= level 1 inclusion) 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). Perform Univariate and Bivariate Analysis.

Regression analysis is one of the most commonly used tools to find a relationship (linear or non-linear) between a response and one or more predictors and exploit that relationship in predicting the expected value of the response for certain values of the predictor(s) with maximum accuracy possible.

Dependent variable or Response: It is the variable of interest that one wants to model or predict using one or more variables whose values are known.

Independent variable(s) or Predictor(s): It is assumed that the response depends on one or more predictors. These variables are independent and a model is formulated identifying the explicit relationship between the response and the predictor(s).

Types of Regression:

I. Simple Linear Regression: When the response is assumed to have a linear dependence on one single predictor.

II. Multiple Linear Regression: When the response is assumed to have a linear dependence on multiple predictors.

Let us now look at the head of the given data, which consists of 26967 rows and 11 columns.

	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

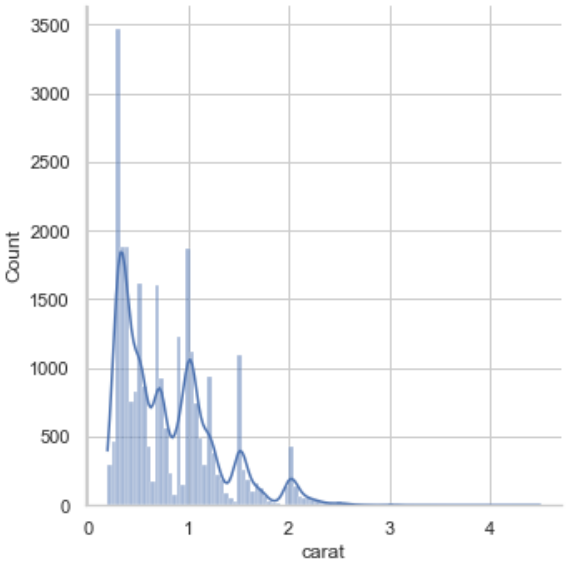
When we deep dive and look at the description, we can see

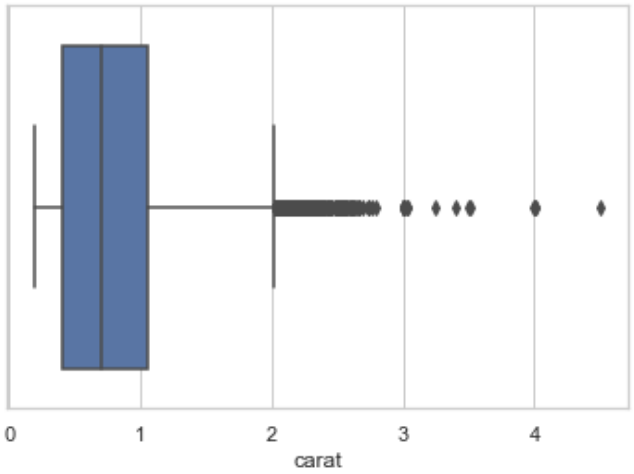
- price of the gem ranges from 326 to 3939.52
- there are 5 unique cuts with ideal being the most frequent 10816
- there are 7 unique colours with G being the most frequent 5661
- there are 8 unique clarity with SI1 being most frequent 6571
- Carat weight of the cubic zirconia ranges from 0.2 to 4.5

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Unnamed: 0	26967.0	NaN	NaN	NaN	13484.0	7784.846691	1.0	6742.5	13484.0	20225.5	26967.0
carat	26967.0	NaN	NaN	NaN	0.798375	0.477745	0.2	0.4	0.7	1.05	4.5
cut	26967	5	Ideal	10816	NaN	NaN	NaN	NaN	NaN	NaN	NaN
color	26967	7	G	5661	NaN	NaN	NaN	NaN	NaN	NaN	NaN
clarity	26967	8	SI1	6571	NaN	NaN	NaN	NaN	NaN	NaN	NaN
depth	26270.0	NaN	NaN	NaN	61.745147	1.41286	50.8	61.0	61.8	62.5	73.6
table	26967.0	NaN	NaN	NaN	57.45608	2.232068	49.0	56.0	57.0	59.0	79.0
x	26967.0	NaN	NaN	NaN	5.729854	1.128516	0.0	4.71	5.69	6.55	10.23
y	26967.0	NaN	NaN	NaN	5.733569	1.166058	0.0	4.71	5.71	6.54	58.9
z	26967.0	NaN	NaN	NaN	3.538057	0.720624	0.0	2.9	3.52	4.04	31.8
price	26967.0	NaN	NaN	NaN	3939.518115	4024.864666	326.0	945.0	2375.0	5360.0	18818.0

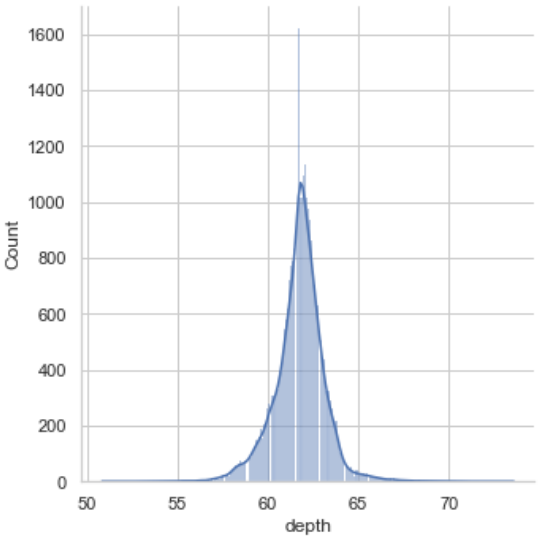
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 11 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Unnamed: 0  26967 non-null  int64  
 1   carat       26967 non-null  float64
 2   cut         26967 non-null  object 
 3   color       26967 non-null  object 
 4   clarity     26967 non-null  object 
 5   depth       26270 non-null  float64
 6   table       26967 non-null  float64
 7   x           26967 non-null  float64
 8   y           26967 non-null  float64
 9   z           26967 non-null  float64
 10  price       26967 non-null  int64  
dtypes: float64(6), int64(2), object(3)
memory usage: 2.3+ MB
```

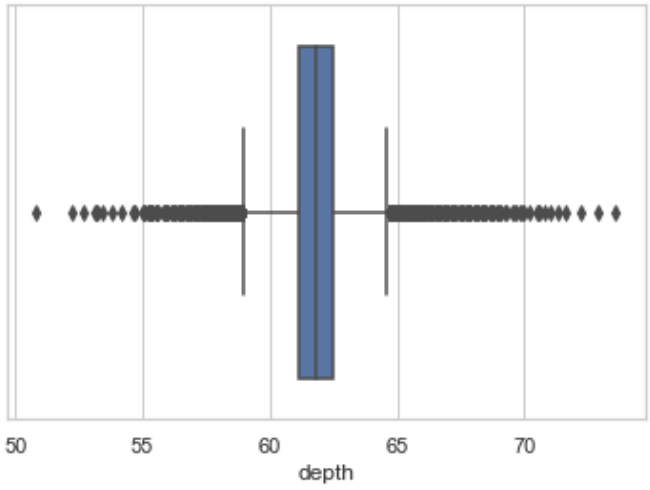
There are 697 null values present in depth column which we have imputed with mean values. Let us now perform univariate and bivariate analysis:



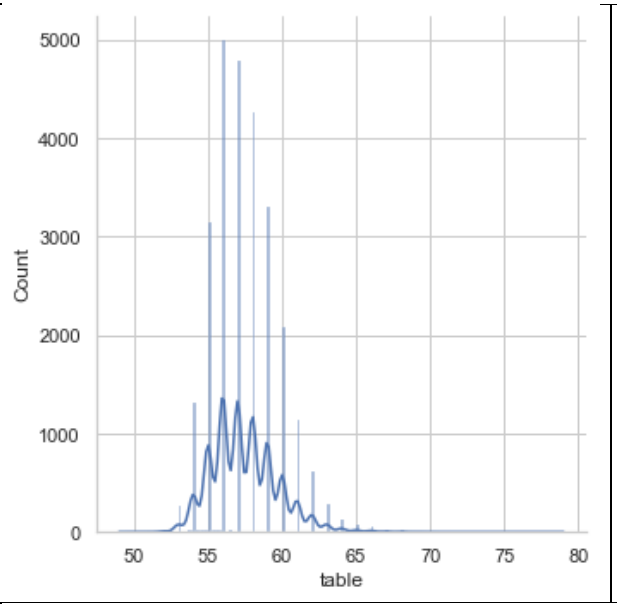


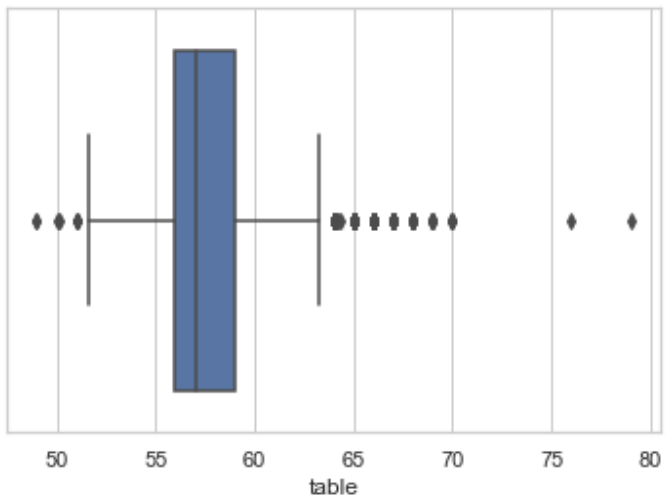

```
Description of carat
-----------------------------------------------------
count    26967.000000
mean         0.798375
std          0.477745
min          0.200000
25%          0.400000
50%          0.700000
75%          1.050000
max          4.500000
Name: carat, dtype: float64 Distribution of carat
-----------------------------------------------------
```



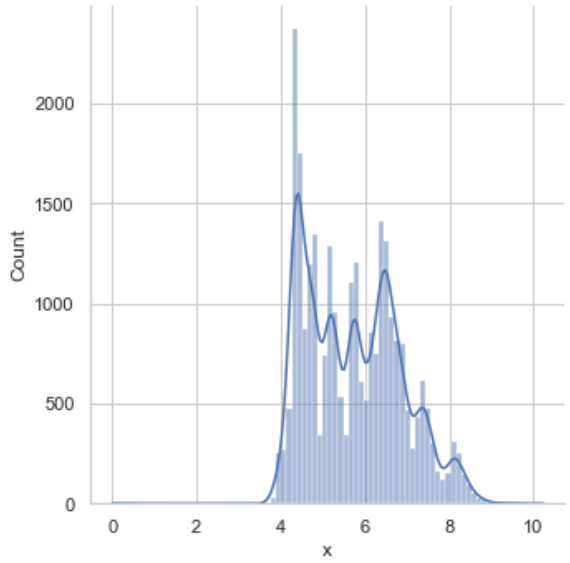


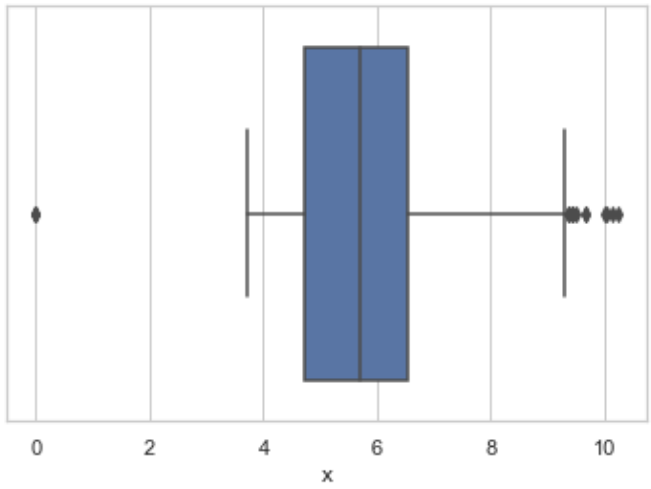
```
Description of depth
-----------------------------------------------------
count    26967.000000
mean        61.745147
std          1.394481
min         50.800000
25%         61.100000
50%         61.800000
75%         62.500000
max         73.600000
Name: depth, dtype: float64 Distribution of depth
-----------------------------------------------------
```



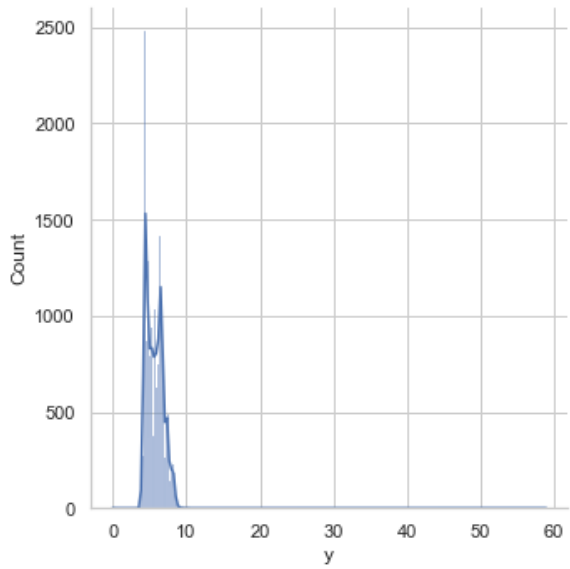


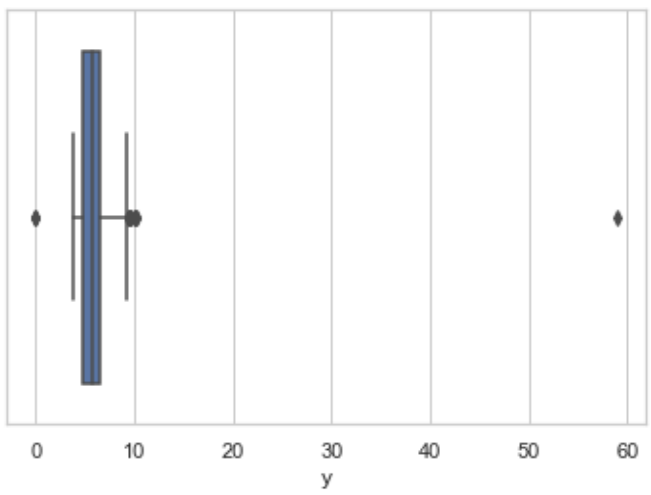

```
Description of table
-----------------------------------------------
count    26967.000000
mean        57.456080
std          2.232068
min         49.000000
25%         56.000000
50%         57.000000
75%         59.000000
max         79.000000
Name: table, dtype: float64 Distribution of table
-----------------------------------------------
```



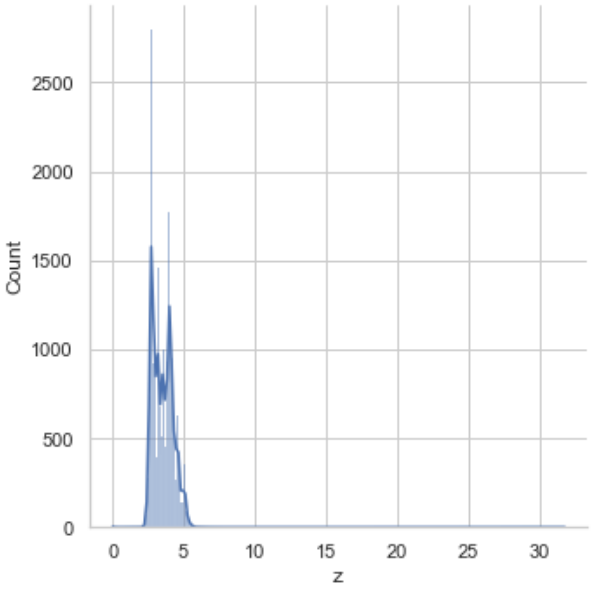


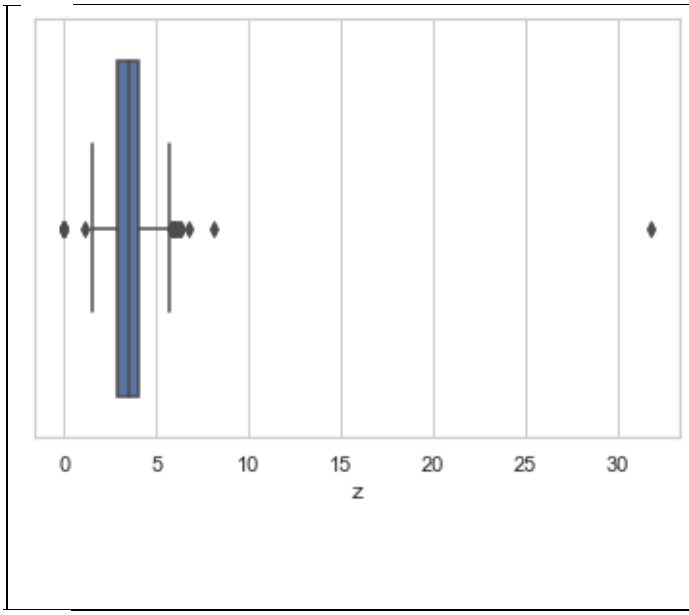

```
Description of x
-----------------------------------------
count    26967.000000
mean         5.729854
std          1.128516
min          0.000000
25%          4.710000
50%          5.690000
75%          6.550000
max         10.230000
Name: x, dtype: float64 Distribution of x
-----------------------------------------
```



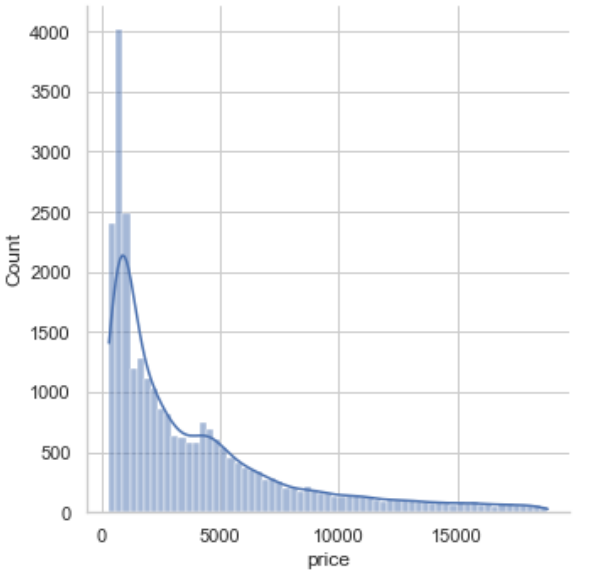


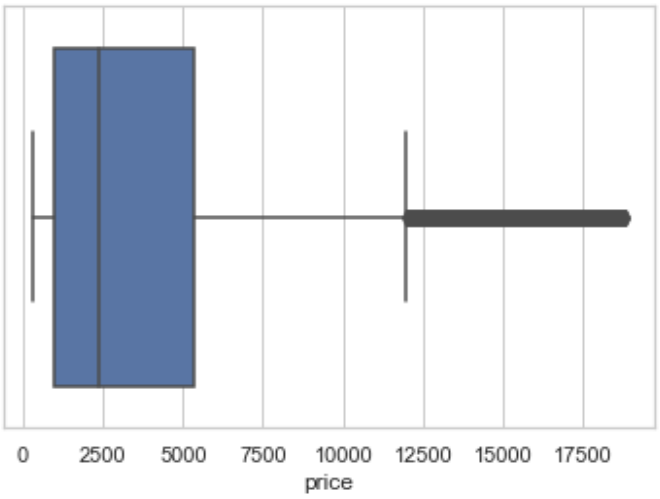
```
Description of y
----------------------------------------
count    26967.000000
mean         5.733569
std          1.166058
min          0.000000
25%          4.710000
50%          5.710000
75%          6.540000
max         58.900000
Name: y, dtype: float64 Distribution of y
----------------------------------------
```



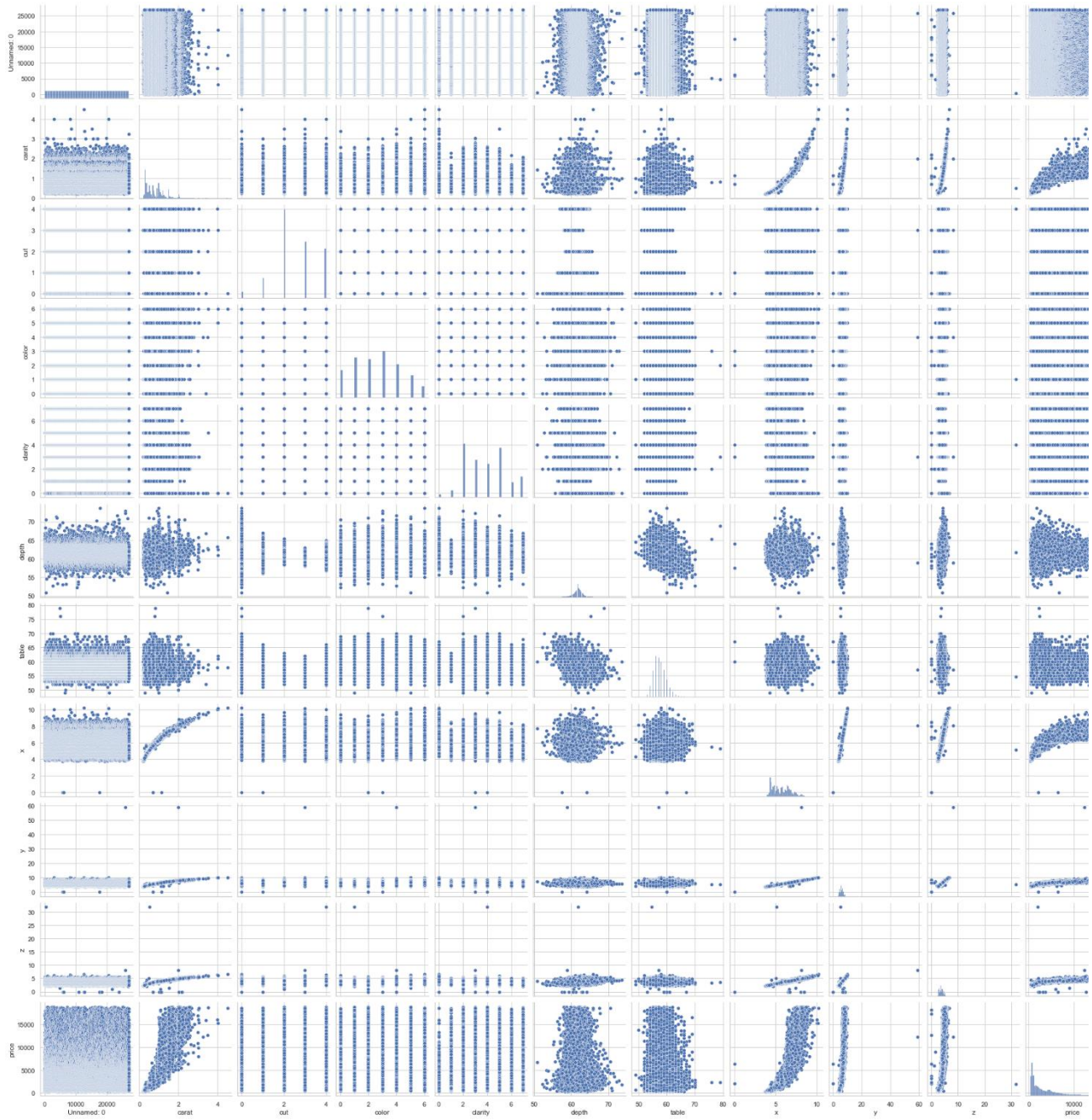


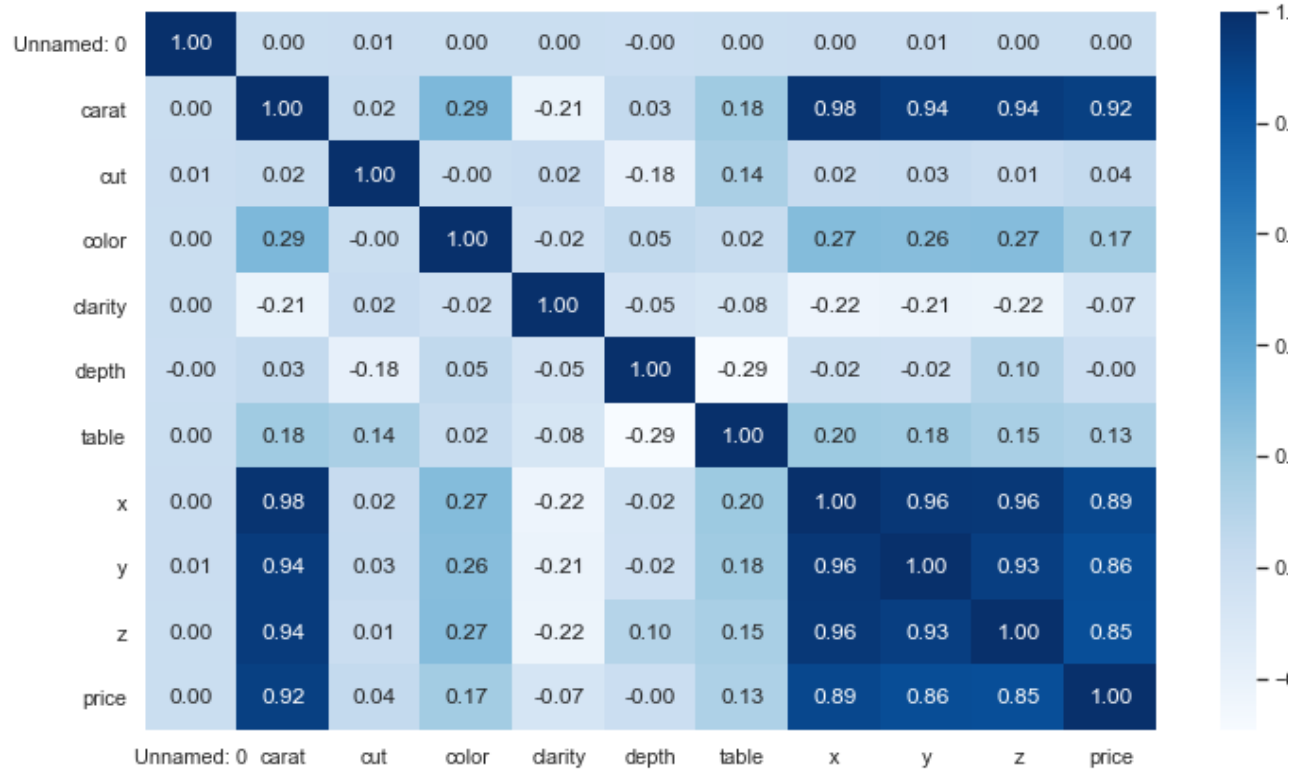

```
Description of z
----------------------------------------
count    26967.000000
mean         3.538057
std          0.720624
min          0.000000
25%          2.900000
50%          3.520000
75%          4.040000
max         31.800000
Name: z, dtype: float64 Distribution of z
----------------------------------------
```






```
Description of price
-----------------------------------------------
count    26967.000000
mean      3939.518115
std       4024.864666
min        326.000000
25%        945.000000
50%       2375.000000
75%       5360.000000
max      18818.000000
Name: price, dtype: float64 Distribution of price
-----------------------------------------------
```

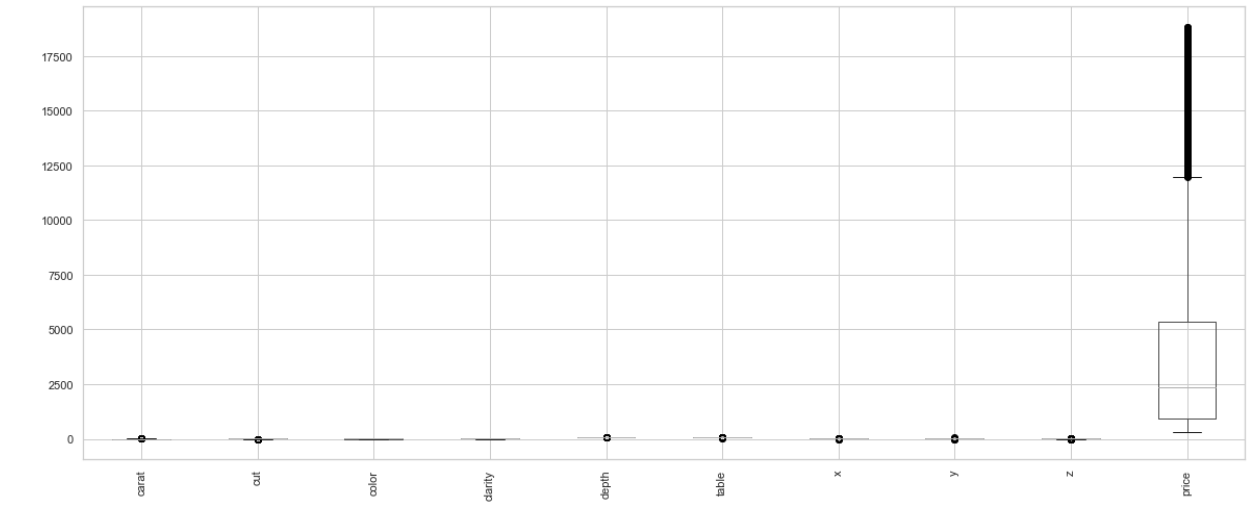

Unnamed: 0
carat
cut
color
clarity
depth
table
x
y
z
price

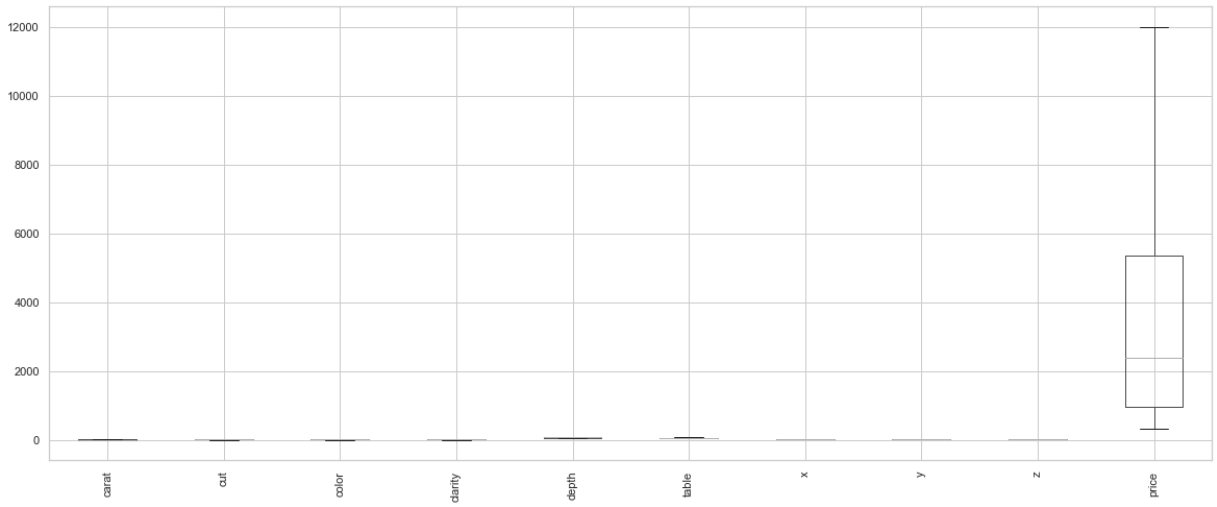


Assumptions of Linear Regression: The regression has five key assumptions:

- Linear relationship
- Multivariate normality
- No or little multicollinearity
- No auto-correlation
- Homoscedasticity

We are going to treat the outliers by capping upper range and lower range for better results. The below box plot shows the datapoints before and after we treat the outliers.





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?

We have imputed null values with mean values in depth column. We encoded cut, clarity, color with numerical values before we fit the model. Standardization is the process of putting different variables on the same scale. In regression analysis, there are some scenarios where it is crucial to standardize your independent variables or risk obtaining misleading results. We have performed zscore scaling to obtain better results.

We need to perform Feature Scaling when we are dealing with Gradient Descent Based algorithms (Linear and Logistic Regression). Changing the

scale of the variable will lead to a corresponding change in the scale of the coefficients and standard errors, but no change in the significance or interpretation.

```
Unnamed: 0    0
carat         0
cut           0
color         0
clarity       0
depth         0
table         0
x             0
y             0
z             0
price         0
dtype: int64
```

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.

After performing the linear regression we obtain the following OLS Regression Results:

	training	testing
R square	0.9103119017894044	0.9076826092045978
RMSE	0.2994797125192217	0.30383777052137917

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.910
Model:                            OLS   Adj. R-squared:                  0.910
Method:                 Least Squares   F-statistic:                 2.128e+04
Date:                Sun, 29 Aug 2021   Prob (F-statistic):               0.00
Time:                        16:52:45   Log-Likelihood:                -4024.9
No. Observations:               18876   AIC:                             8070.
Df Residuals:                   18866   BIC:                             8148.
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   1.214e-17      0.002   5.57e-15      1.000      -0.004       0.004
carat          1.2201      0.012    100.216      0.000       1.196       1.244
cut            0.0100      0.002      4.456      0.000       0.006       0.014
color         -0.1117      0.002    -48.791      0.000      -0.116      -0.107
clarity        0.1232      0.002     54.761      0.000       0.119       0.128
depth         -0.0328      0.003     -9.820      0.000      -0.039      -0.026
table         -0.0464      0.002    -19.496      0.000      -0.051      -0.042
x             -0.6896      0.048    -14.401      0.000      -0.783      -0.596
y              0.5229      0.048     10.950      0.000       0.429       0.617
z             -0.0524      0.020     -2.606      0.009      -0.092      -0.013
==============================================================================
Omnibus:                     4841.288   Durbin-Watson:                   1.984
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            24187.094
Skew:                           1.153   Prob(JB):                         0.00
Kurtosis:                       8.043   Cond. No.                         61.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

The final Linear Regression equation is

price = b0 + b1 * instant_bookable[T.True] + b2 * carat + b3 * cut + b4 * color + b5 * clarity + b6 * depth + b7 * table + b8 * x + b9 * y + b10 * z

price = (0.0) * Intercept + (0) * instant_bookable[T.True] + (1.22) * carat + (0.01) * cut + (-0.11) * color + (0.12) * clarity + (-0.03) * depth + (-0.05) * table + (-0.69) * x + (0.52) * y + (-0.05) * z

When carat increases by 1 unit, price increases by 1.22 units, keeping all other predictors constant.
similarly, when cut increases by 1 unit, price increases by 0.01 units, keeping all other predictors constant.

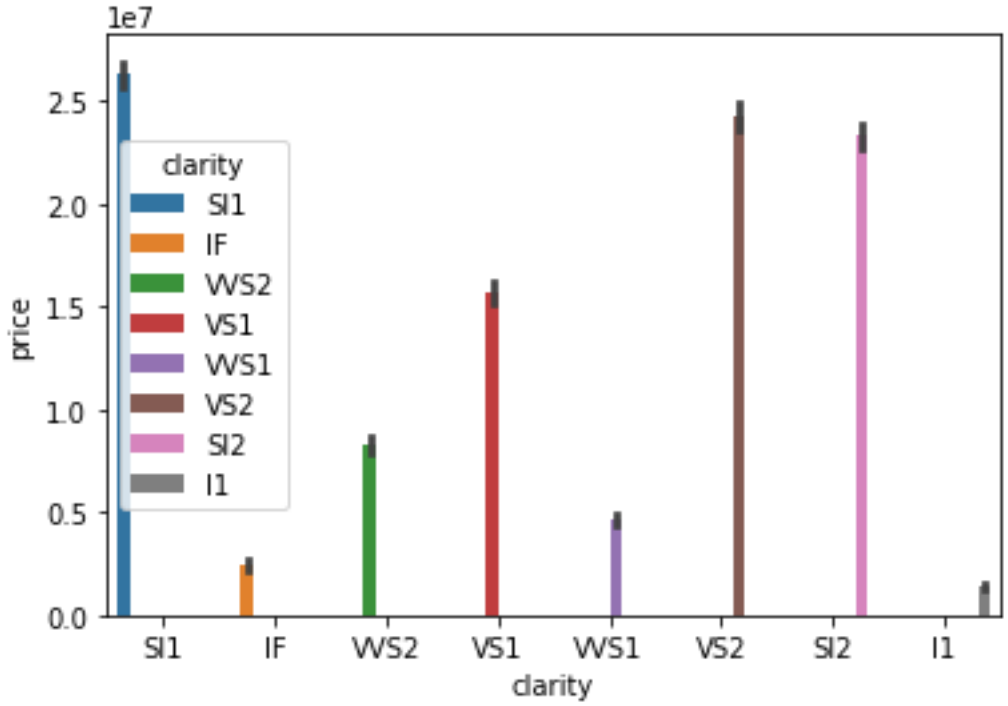
There are also some negative co-efficient values, for instance, color has its corresponding co-efficient as -0.11. This implies, when there is change in color , the price decreases by 0.11 units, keeping all other predictors constant.

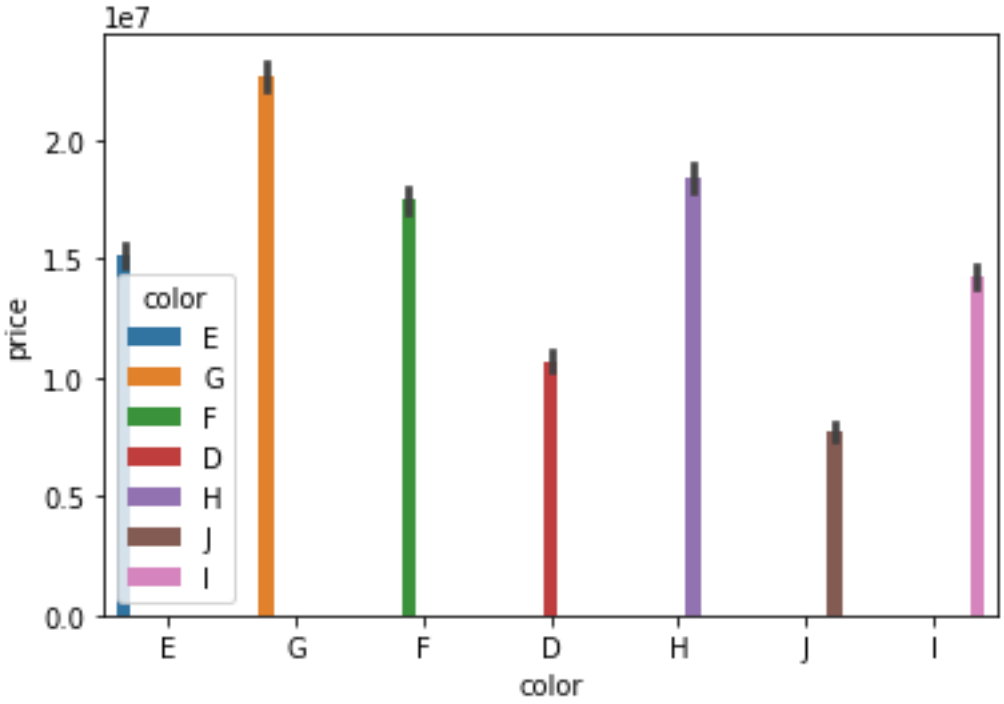
1.4 Inference: Basis on these predictions, what are the business insights and recommendations.

After applying linear regression, we have arrived at the equation to predict the price for the stone on the bases of the details given in the dataset.

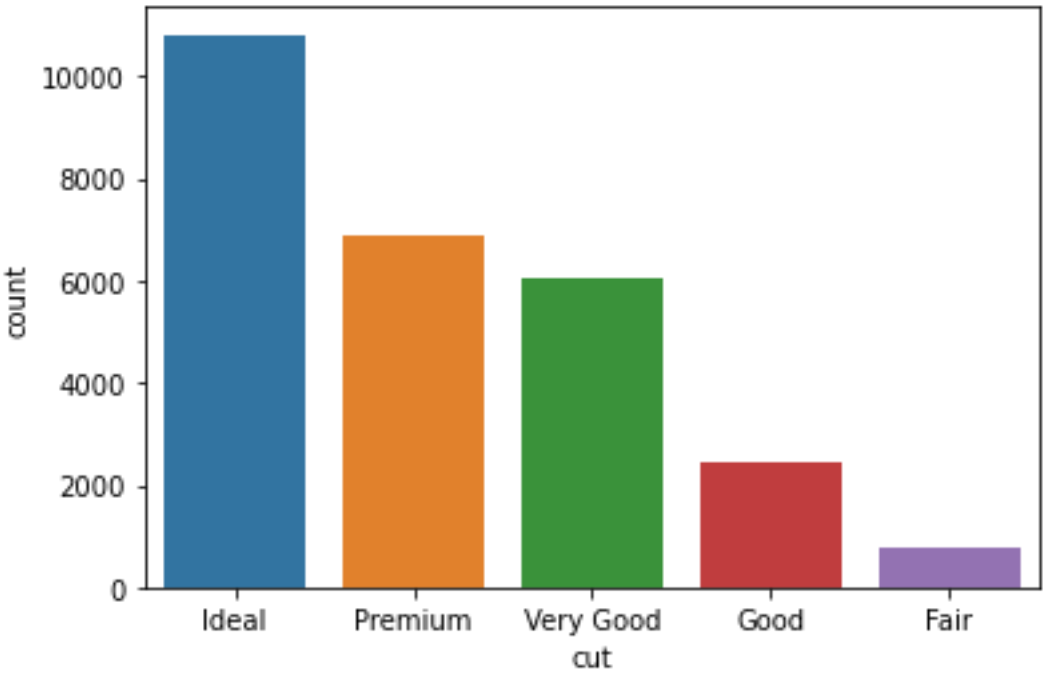
We can summarize that the factors influencing the price most are carat, clarity, Width of the cubic zirconia in mm, color, Length of the cubic zirconia in mm.

In terms of clarity top performing SI1, VS2, SI2 are dominating compared to other gems, the company should stock only these gems.





Though J is the best color and D is the worst color, the frequently bought color is G,H,F the gem company can spend its resources procuring these colors and stop investing on other colors.



From the above graph, clearly Ideal Cut has generated the most sale while Fair cut has generated the least sale. The gem company can certainly leverage this data to produce more of ideal cuts to generate more profit.

Also, according to the depth histogram, 60 to 65mm depth is the most preferable. Since these gems are desirable the company can invest on these gems. According to table histogram, 55 to 60 mm Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter is desirable.

Speaking of length, width and height according to the histogram used in univariate analysis it ranges from 4mm to 7mm, 5mm to 10mm and 3mm to 5mm respectively. Since customers are only spending on gem stones in the above range it would to wise to hold stock only of these precious stones.

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

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.

The given data can be summarized as follows, there are 872 rows and 8 columns:

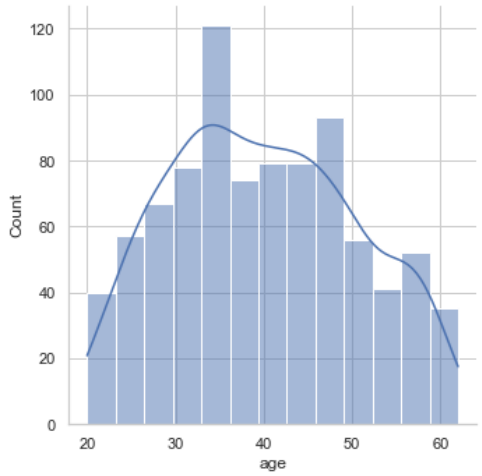
	Unnamed: 0	Salary	age	educ	no_young_children	no_older_children
count	872.000000	872.000000	872.000000	872.000000	872.000000	872.000000
mean	436.500000	47729.172018	39.955275	9.307339	0.311927	0.982798
std	251.869014	23418.668531	10.551675	3.036259	0.612870	1.086786
min	1.000000	1322.000000	20.000000	1.000000	0.000000	0.000000
25%	218.750000	35324.000000	32.000000	8.000000	0.000000	0.000000
50%	436.500000	41903.500000	39.000000	9.000000	0.000000	1.000000
75%	654.250000	53469.500000	48.000000	12.000000	0.000000	2.000000
max	872.000000	236961.000000	62.000000	21.000000	3.000000	6.000000

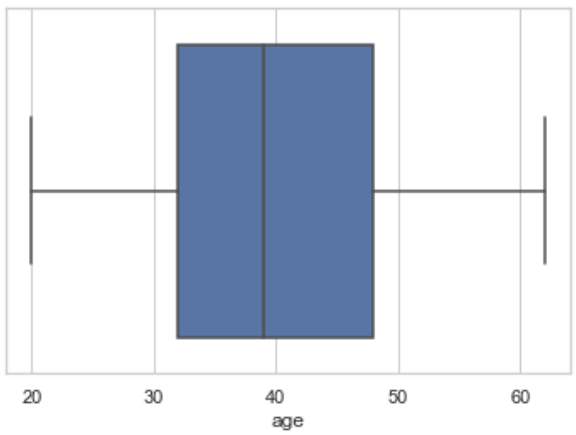
	Unnamed: 0	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	1	no	48412	30	8	1	1	no
1	2	yes	37207	45	8	0	1	no
2	3	no	58022	46	9	0	0	no
3	4	no	66503	31	11	2	0	no
4	5	no	66734	44	12	0	2	no

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Unnamed: 0         872 non-null    int64
 1   Holliday_Package   872 non-null    object
 2   Salary             872 non-null    int64
 3   age                872 non-null    int64
 4   educ               872 non-null    int64
 5   no_young_children  872 non-null    int64
 6   no_older_children  872 non-null    int64
 7   foreign            872 non-null    object
dtypes: int64(6), object(2)
memory usage: 54.6+ KB
```

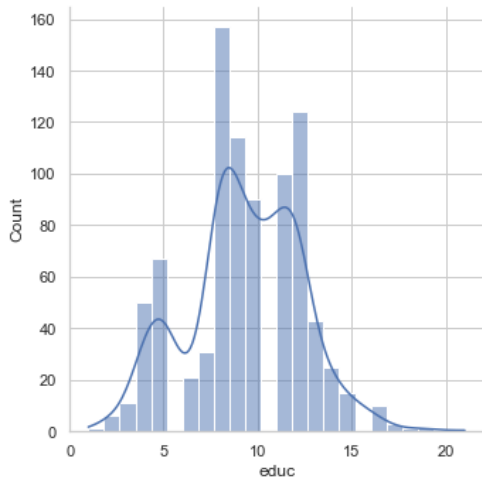
```
Unnamed: 0           0
Holliday_Package     0
Salary               0
age                  0
educ                 0
no_young_children    0
no_older_children    0
foreign              0
dtype: int64
```

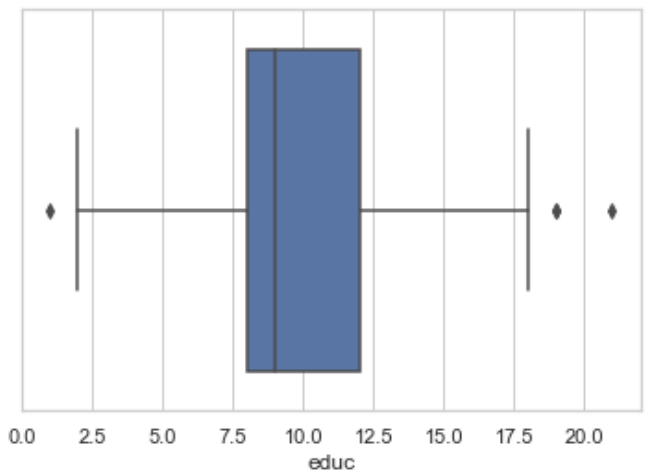
After checking for null values, we do not find any null values in the dataset. Let us now perform multi varite analysis:



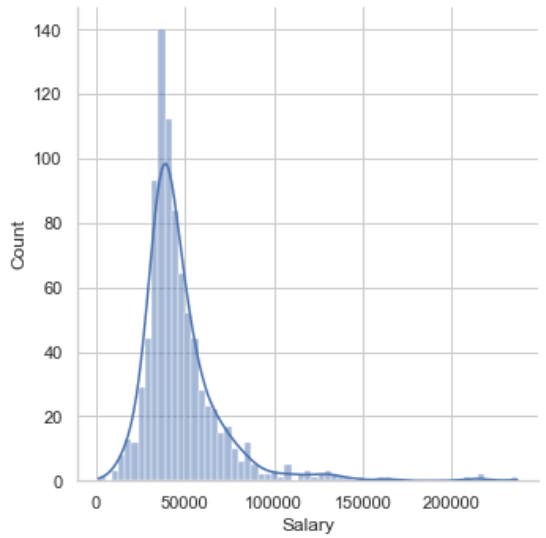


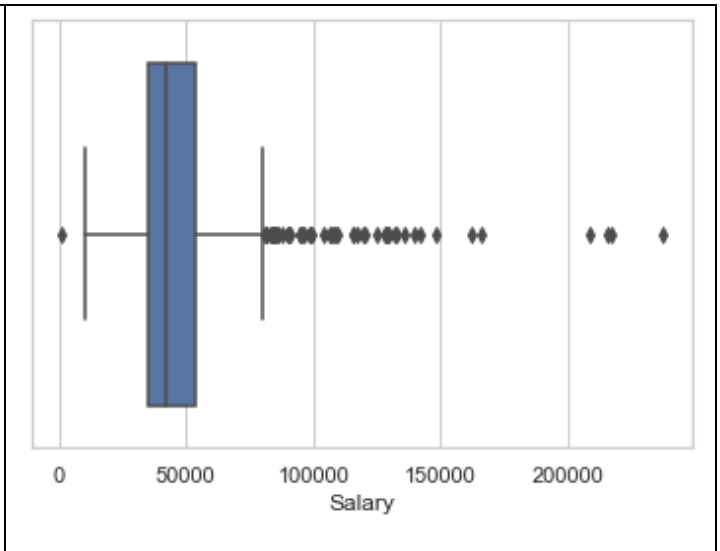

```
Description of age
-----------------------------------------------
count    872.000000
mean      39.955275
std       10.551675
min       20.000000
25%       32.000000
50%       39.000000
75%       48.000000
max       62.000000
Name: age, dtype: float64 Distribution of age
-----------------------------------------------
```



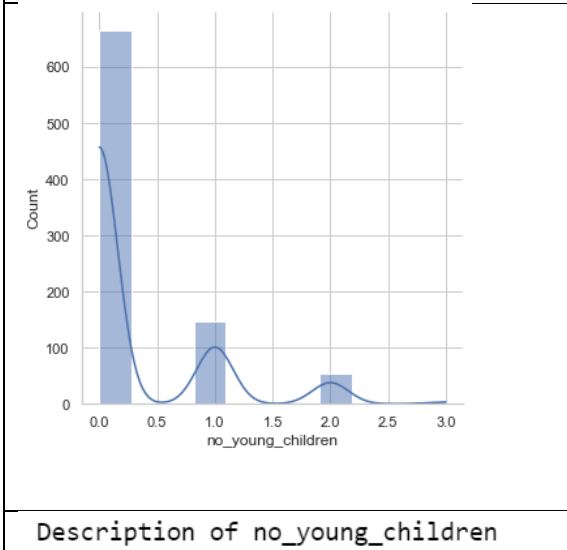


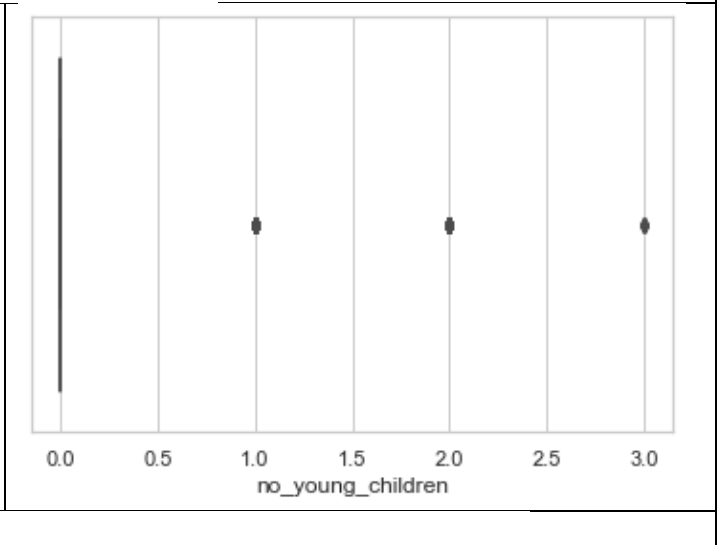

```
Description of educ
-------------------------------------------------
count    872.000000
mean       9.307339
std        3.036259
min        1.000000
25%        8.000000
50%        9.000000
75%       12.000000
max       21.000000
Name: educ, dtype: float64 Distribution of educ
-------------------------------------------------
```



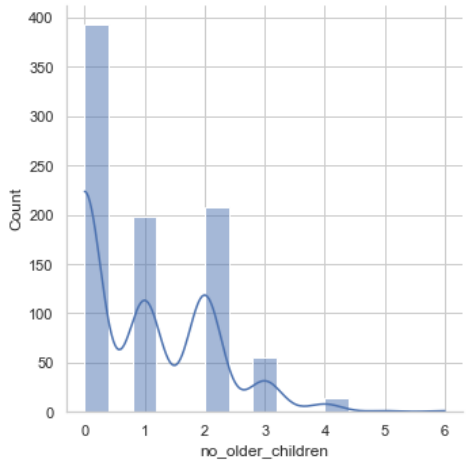


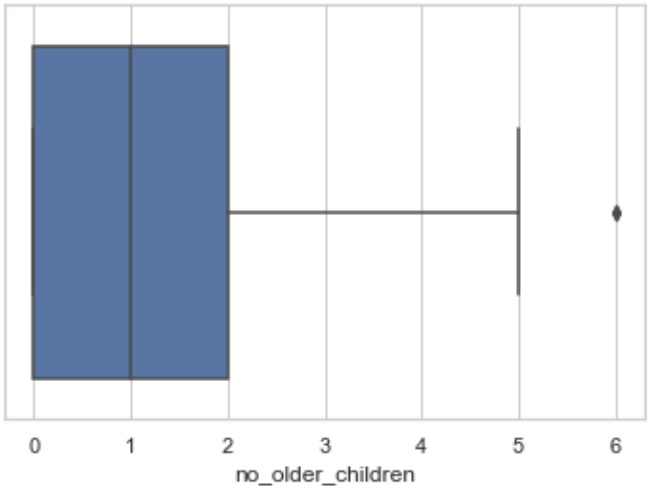

```
Description of Salary
-----------------------------------------------------
count       872.000000
mean      47729.172018
std       23418.668531
min        1322.000000
25%       35324.000000
50%       41903.500000
75%       53469.500000
max      236961.000000
Name: Salary, dtype: float64 Distribution of Salary
-----------------------------------------------------
```



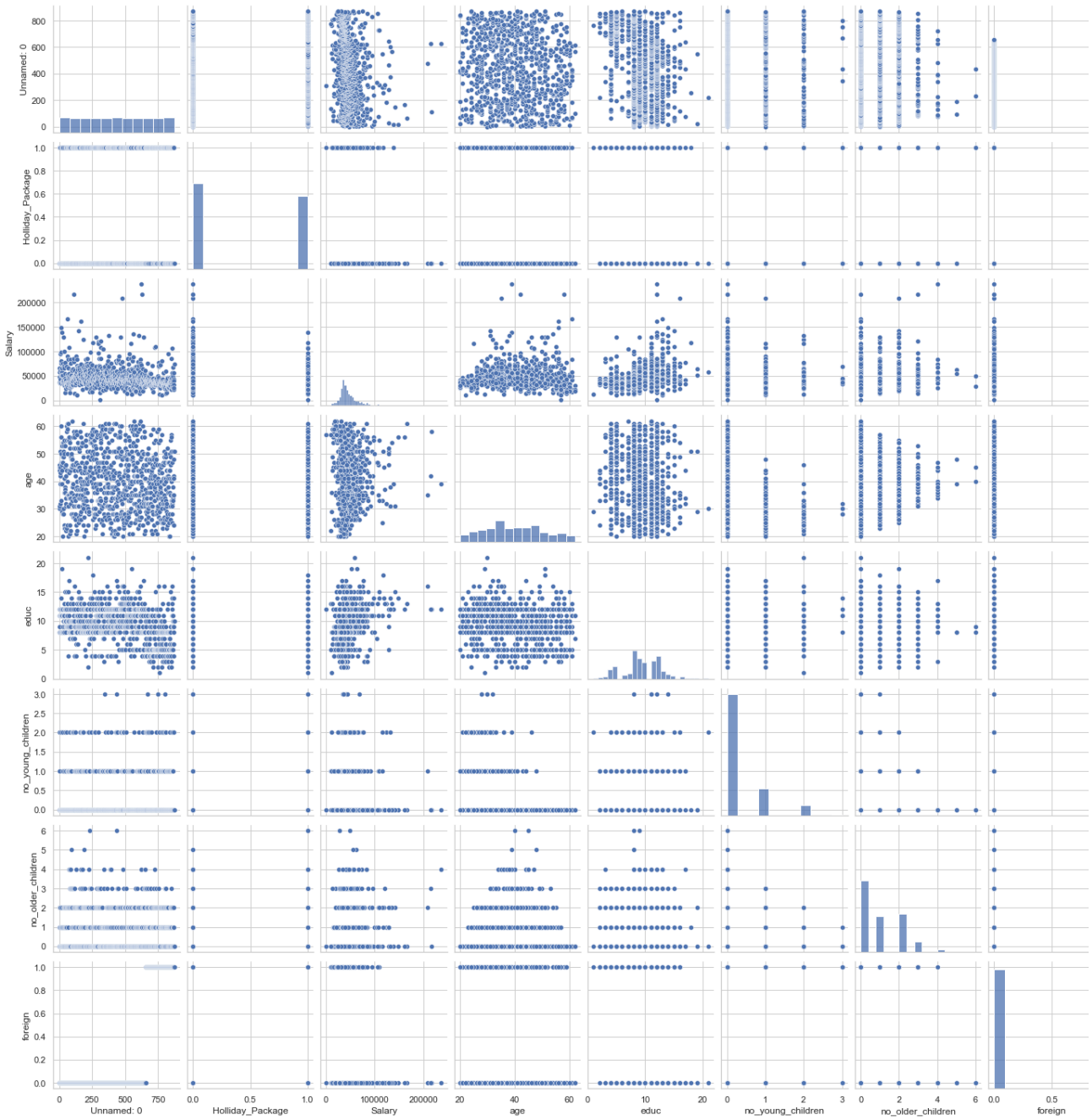



```
Description of no_young_children
-----------------------------------------------------------------------
count    872.000000
mean       0.311927
std        0.612870
min        0.000000
25%        0.000000
50%        0.000000
75%        0.000000
max        3.000000
Name: no_young_children, dtype: float64 Distribution of no_young_children
-----------------------------------------------------------------------
```





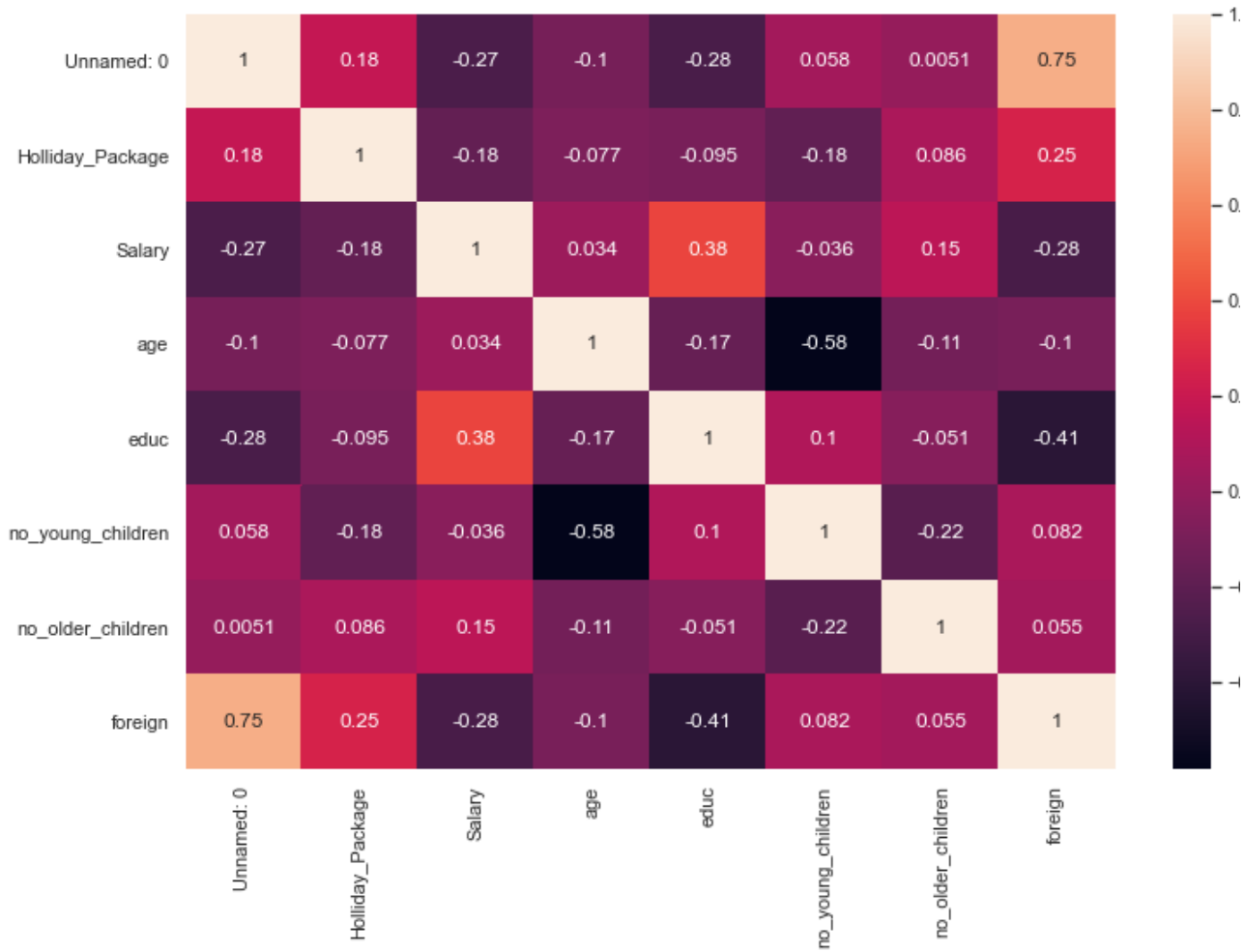

```
Description of no_older_children
-------------------------------------------------------------------------------
count    872.000000
mean       0.982798
std        1.086786
min        0.000000
25%        0.000000
50%        1.000000
75%        2.000000
max        6.000000
Name: no_older_children, dtype: float64 Distribution of no_older_children
-------------------------------------------------------------------------------
```



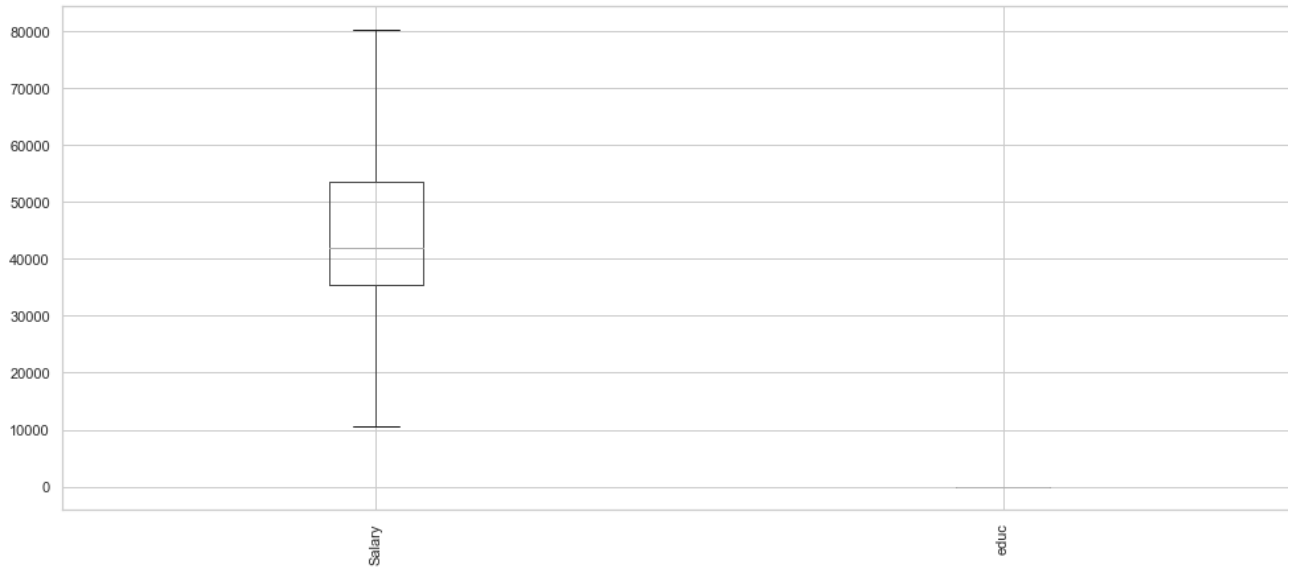
The 6 Assumptions of Logistic Regression:

- The Response Variable is Binary
- The Observations are Independent
- There is No Multicollinearity Among Explanatory Variables
- There are No Extreme Outliers

- There is a Linear Relationship Between Explanatory Variables and the Logit of the Response Variable
- The Sample Size is Sufficiently Large



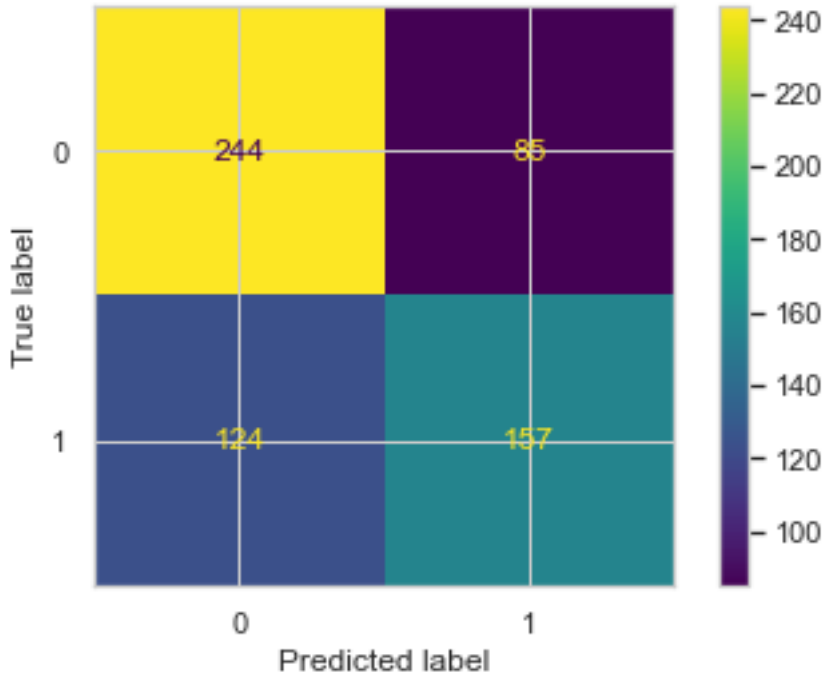
We have capped salary and educ columns with upper range and lower range, for better accuracy



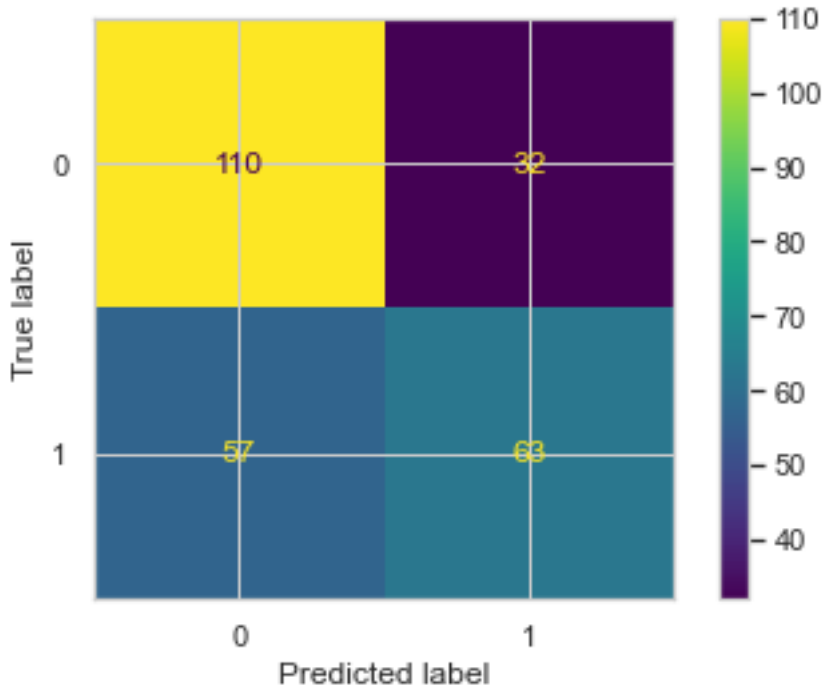
2.2 Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis).

After applying logistic Regression and linear discriminant analysis, we have obtained the following confusion matrix and classification report:

	precision	recall	f1-score	support
0	0.66	0.74	0.70	329
1	0.65	0.56	0.60	281
accuracy			0.66	610
macro avg	0.66	0.65	0.65	610
weighted avg	0.66	0.66	0.65	610



	precision	recall	f1-score	support
0	0.54	1.00	0.70	142
1	0.00	0.00	0.00	120
accuracy			0.54	262
macro avg	0.27	0.50	0.35	262
weighted avg	0.29	0.54	0.38	262

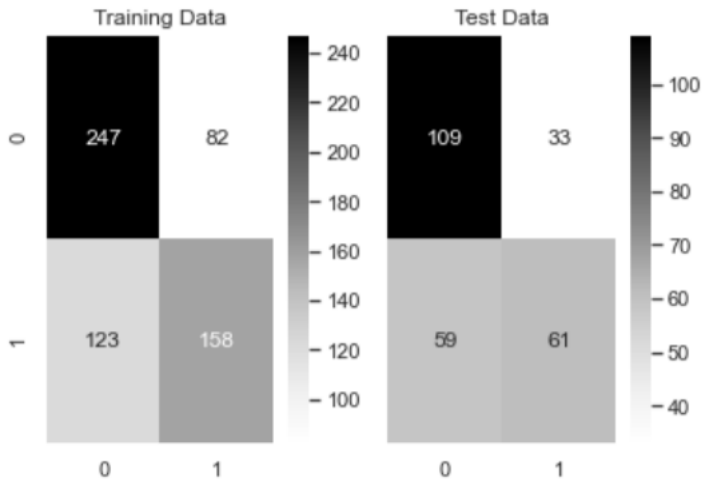


Classification Report of the training data:

	precision	recall	f1-score	support
0	0.67	0.75	0.71	329
1	0.66	0.56	0.61	281
accuracy			0.66	610
macro avg	0.66	0.66	0.66	610
weighted avg	0.66	0.66	0.66	610

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.65	0.77	0.70	142
1	0.65	0.51	0.57	120
accuracy			0.65	262
macro avg	0.65	0.64	0.64	262
weighted avg	0.65	0.65	0.64	262



Training and Test set results are almost similar, and with the overall measures high, the model is a good model. Agency code is the most important variable for predicting claim status. This is a robust model, since train and test accuracy difference is within 10%

Speaking of accuracy according to industry standards,

If your 'X' value is between 60% and 70%, it's a poor model.

If your 'X' value is between 70% and 80%, you've got a good model.

If your 'X' value is between 80% and 90%, you have an excellent model.

If your 'X' value is between 90% and 100%, it's a probably an overfitting case.

The accuracy for all 3 models for test and train ranges between 60% and 70%, so we can say we have got poor model.

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.

We need to evaluate the model that we have built and validate how good (or bad) it is, so you can then decide on whether to implement it. That's where the AUC-ROC curve comes in.

The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the 'signal' from the 'noise'. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

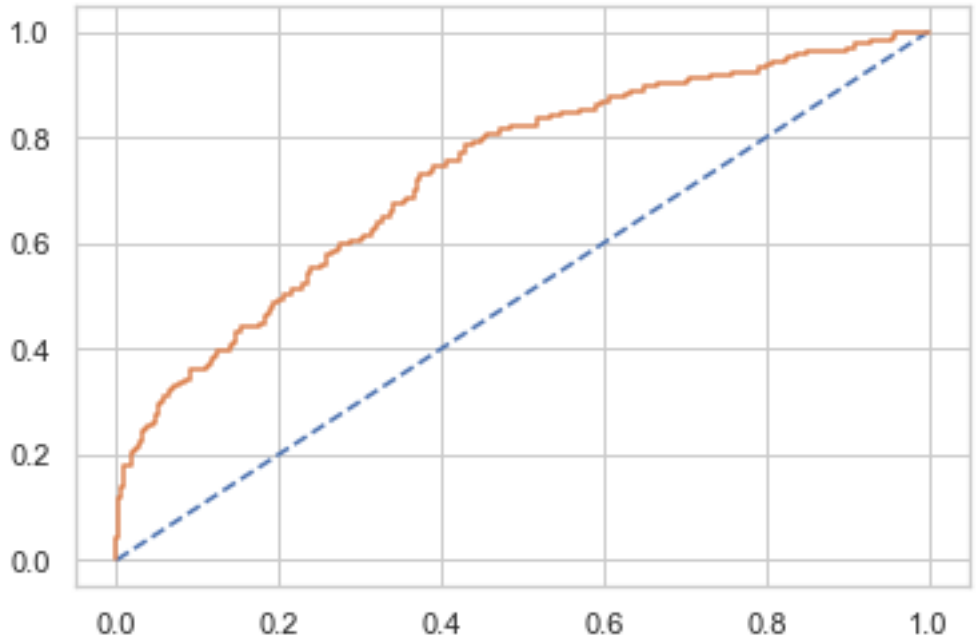
The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

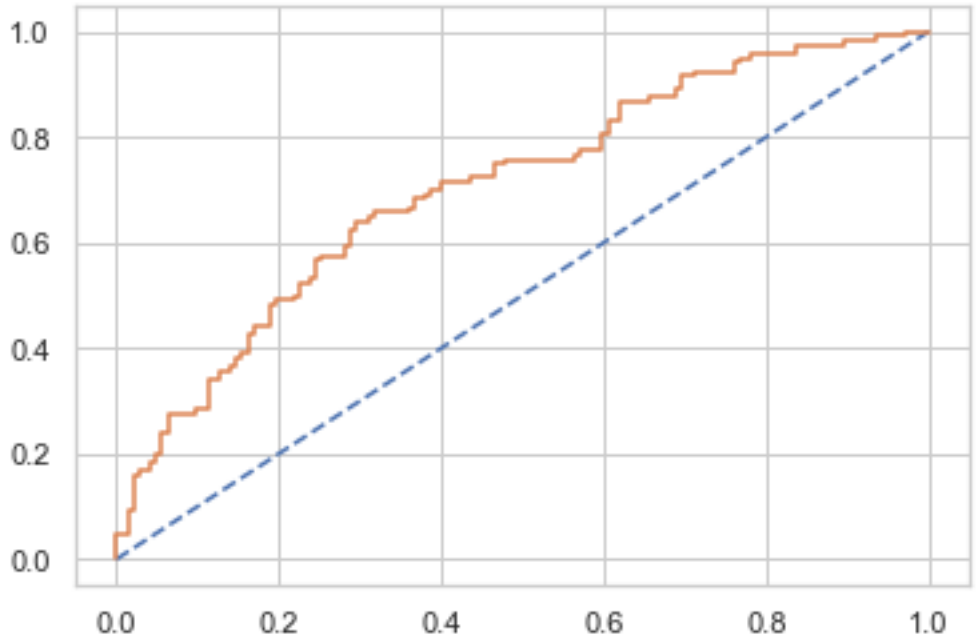
To evaluate our models, we are going to use accuracy, sensitivity, specificity, precision, F1 score.

1. **Accuracy**: from the table we can see the formula for accuracy which is crucial because the goal of our models is to build accurate models with minimum errors
2. **Sensitivity/recall/true positive rate:** this calculates out of all true, how many were predicted positively. A higher TPR and a lower FNR is desirable since we want to correctly classify the positive class.
3. **Specificity/true negative rate:** out of all the true, how many predicted negatively.
4. **Precision**: calculates out of all predicted positive, how many are true
5. **F1 score**: The harmonic mean of precision and recall gives a score call f1 score which is a measure of performance of the model's classification ability.

   F1 score = 2 * (precision * recall)/ (precision + recall)

   F1 score is considered a better indicator of the classifier's performance than the regular accuracy measure.





LDA:

AUC for the Training Data: 0.729
AUC for the Test Data: 0.704

	Logistic regression	Linear discriminant analysis
accuracy	0.66	0.66
AUC	0.731	0.729
recall	0.56	0.56
precision	0.65	0.66

AUC is an effective way to summarize the overall accuracy of the test. It takes values from 0 to 1, where a value of 0 indicates a perfectly inaccurate test and a value of 1 reflects a perfectly accurate test. AUC can be computed using the trapezoidal rule. In general, an AUC of 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding.

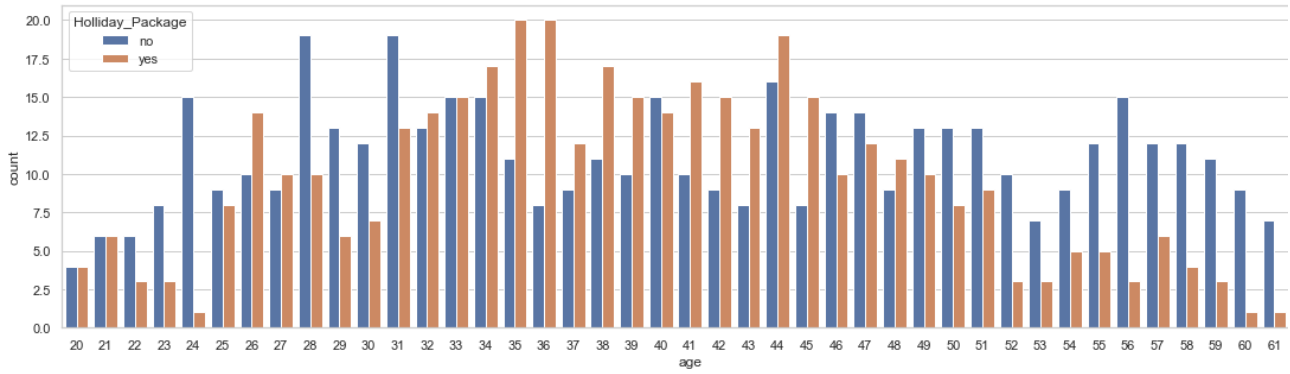
When we compare both models, AUC is acceptable and the results are very similar with LDA seeming to perform slightly better.

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

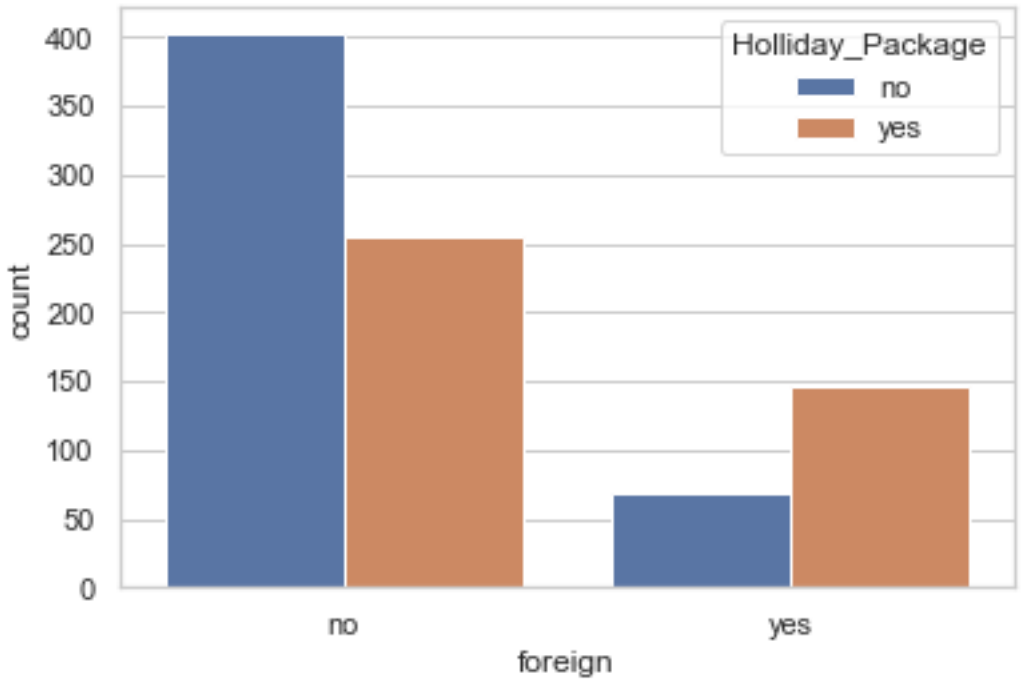
We have performed logistic regression and linear discriminant analysis to predict whether an employee will opt for the package or not on the basis of the information given in the data set.

The model we built is able to predict with 66% accuracy. Let us know look at various factors affecting the propensity to buy holiday package.

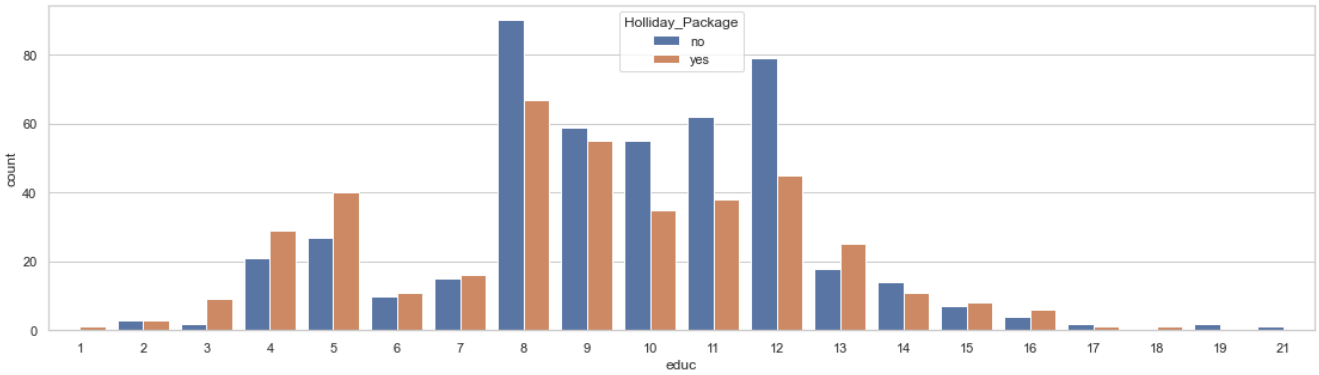
When we look at the predictor variables, we have got age, salary, years of formal education, number of younger and older children, if they are foreign or not?



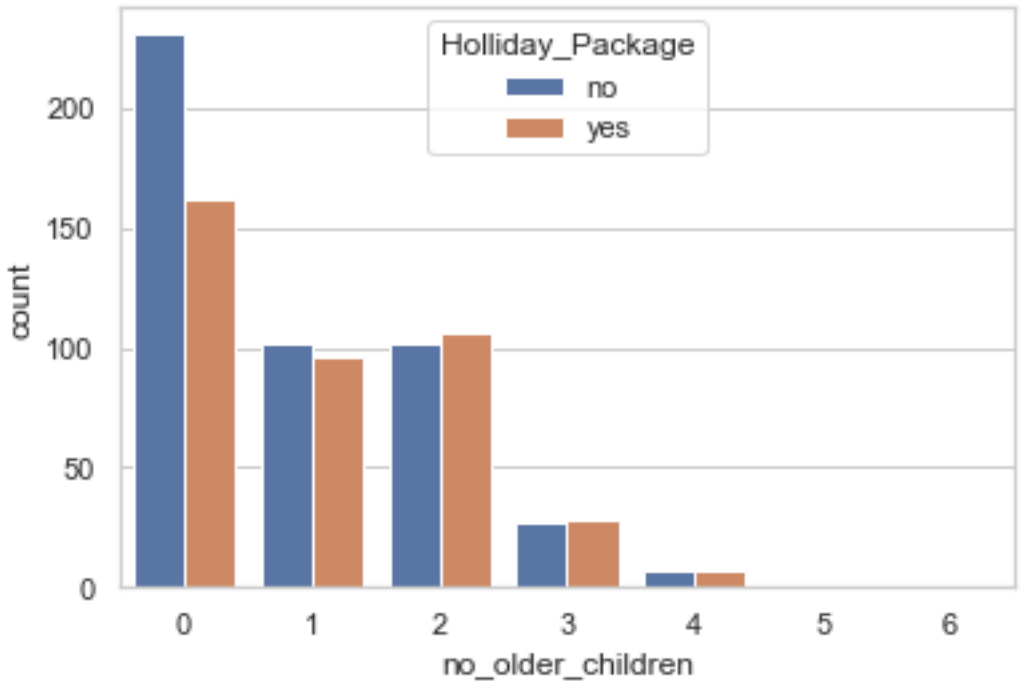
We can notice from this age graph that age group 32 to 45 yrs are more likely to sign up for holiday package, the company can leverage this and focus on target audience in this age group.

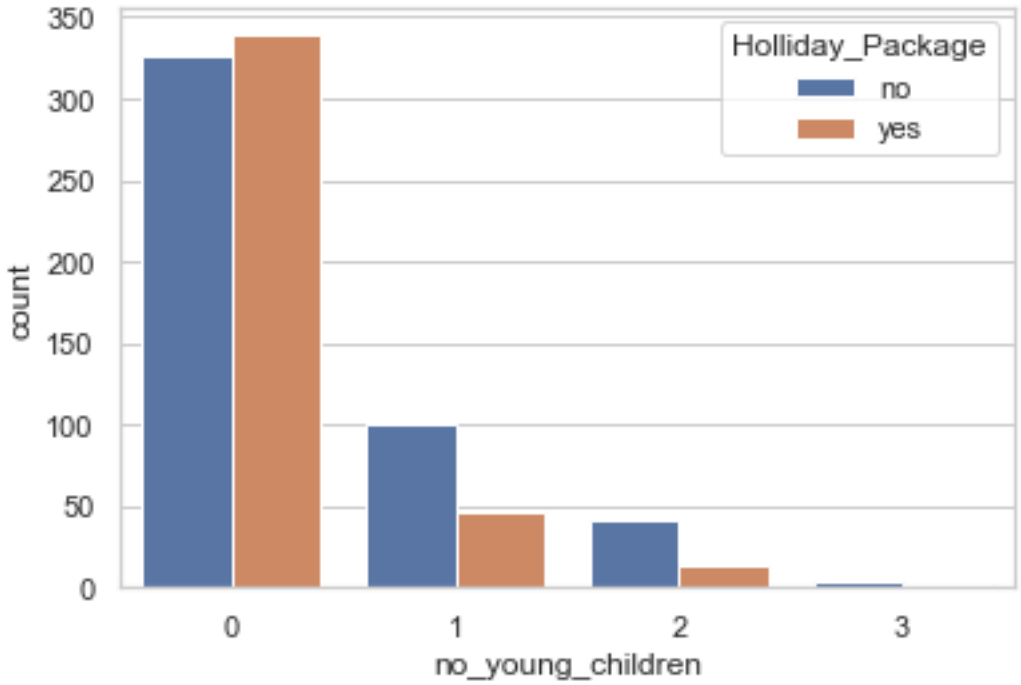


We can also notice that foreigners are more inclined to sign up for the holiday package which means the tour and travel agency which deals in selling holiday packages can target the foreigners to increase sales.



This graph is shedding light on how Customers with less than 8 Years of formal education are more likely to sign up for package which the agency can leverage to target the correct audience.





When we deep dive into data for parents with younger and older children, we can gain some amazing insight into this pattern. Clearly parents with no younger children and parents with more older children have signed up because it is difficult to travel with small children