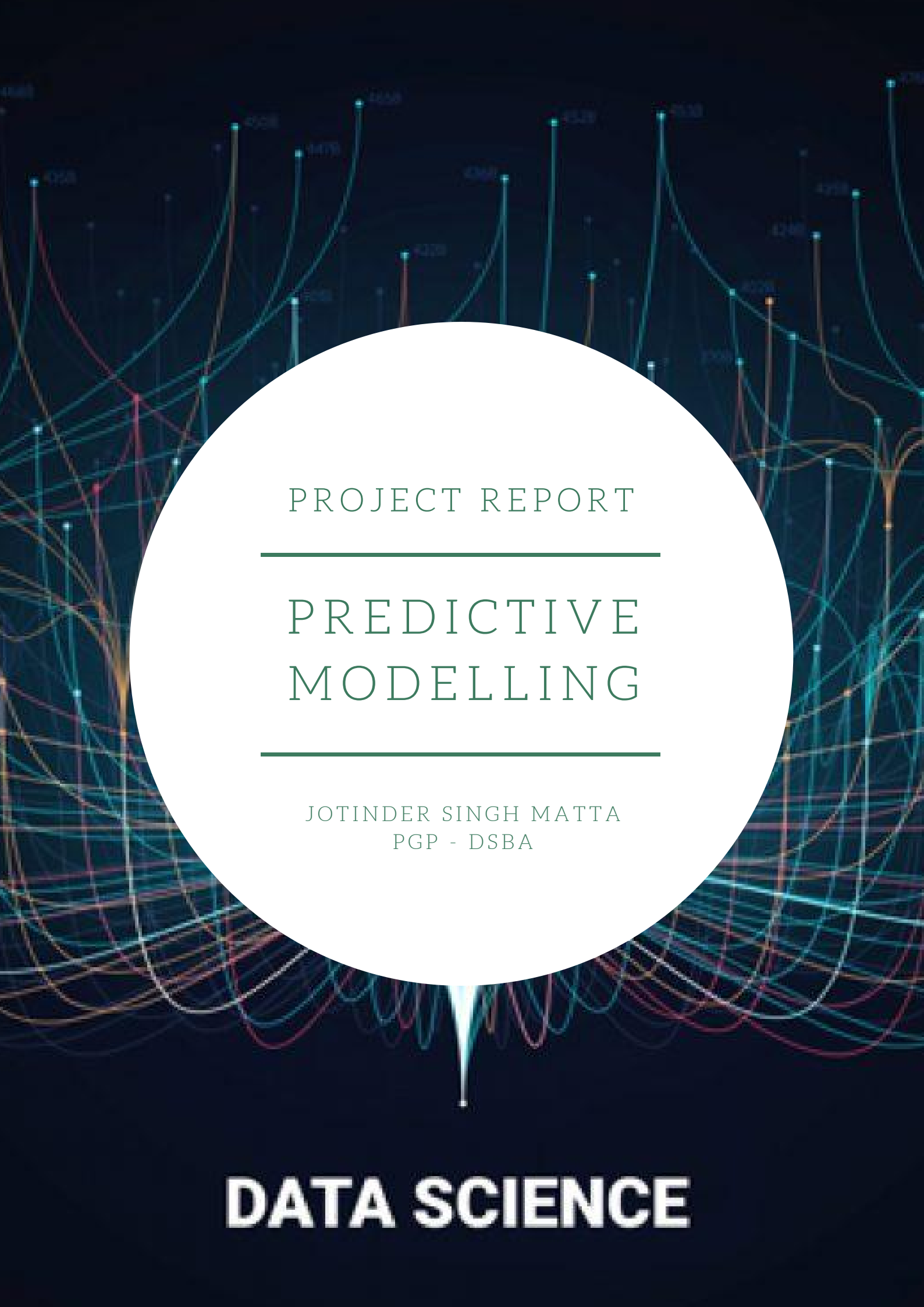
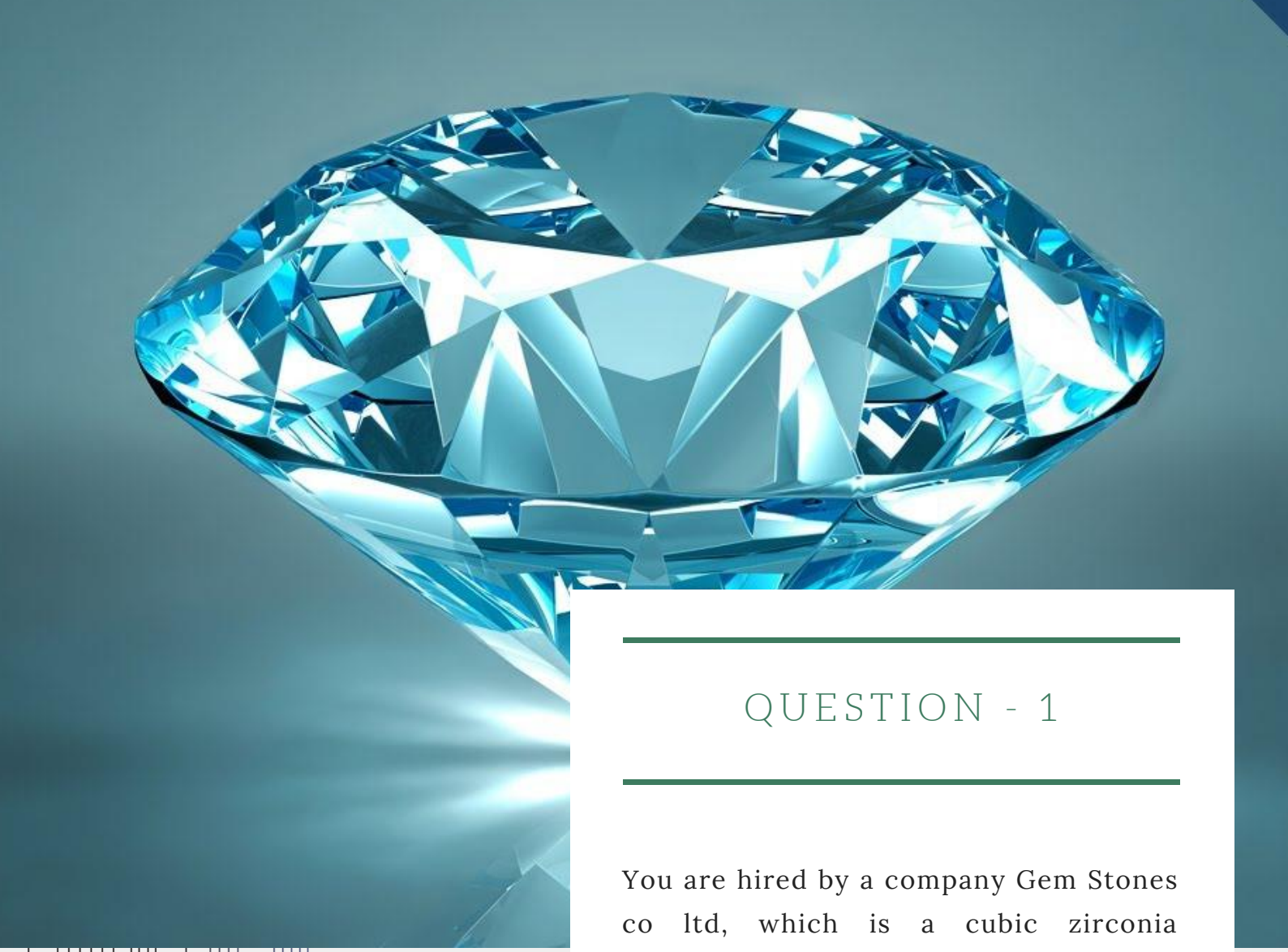


PROJECT REPORT

PREDICTIVE MODELLING

JOTINDER SINGH MATTA
PGP - DSBA



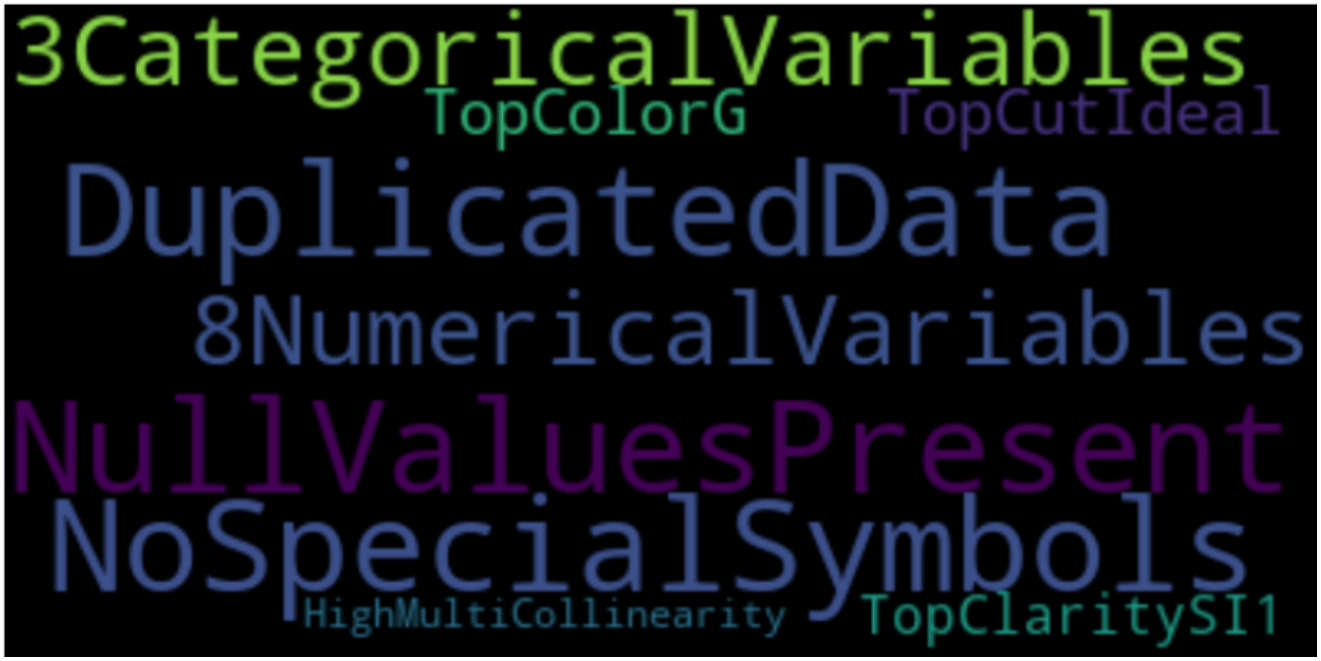
QUESTION - 1

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.

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.

The diamond data set shared with us had 26,967 records. Below are some of the salient features of this data set.



Data set contained, irrelevant column named "Unnamed: 0" which was the same as the index column, hence this was dropped.

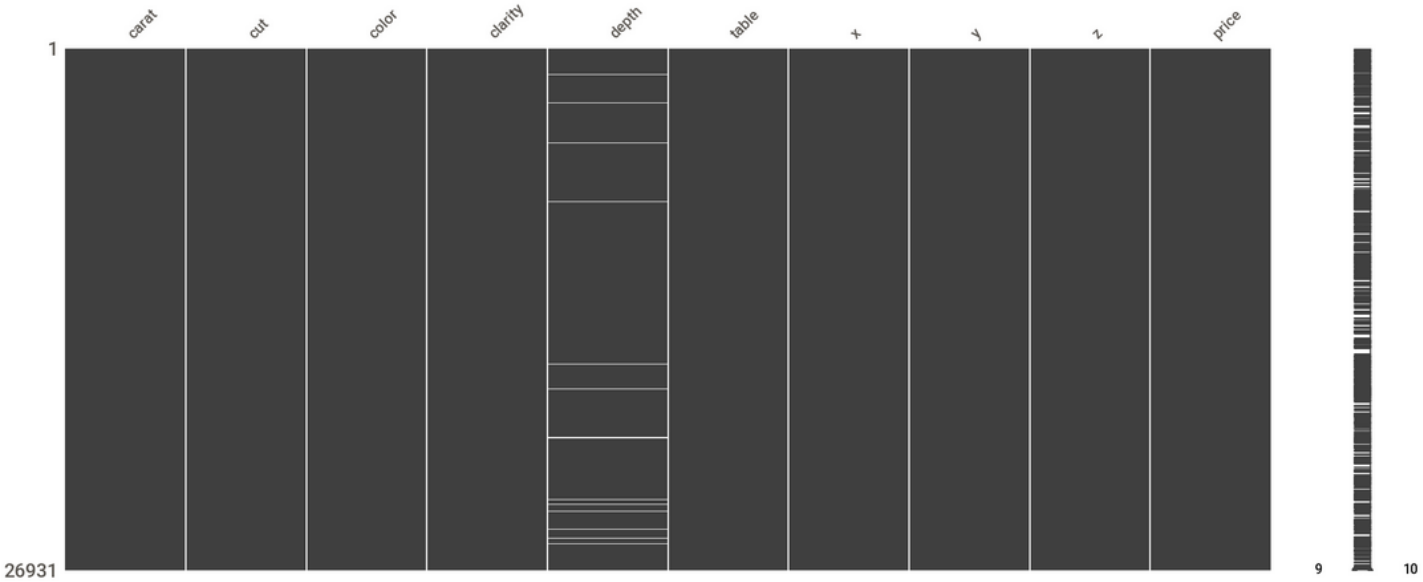
Total of 11 columns, containing 3 categorical and 8 numerical columns.

There were total of 8 rows where the values of either one or all of 'x','y' or 'z' values were 0, these too were dropped from the data set.

Data is mostly right skewed with an exception of "depth" variable, which was slightly left skewed.

Price is the dependent variable in this data set.

NULL VALUES PRESENT



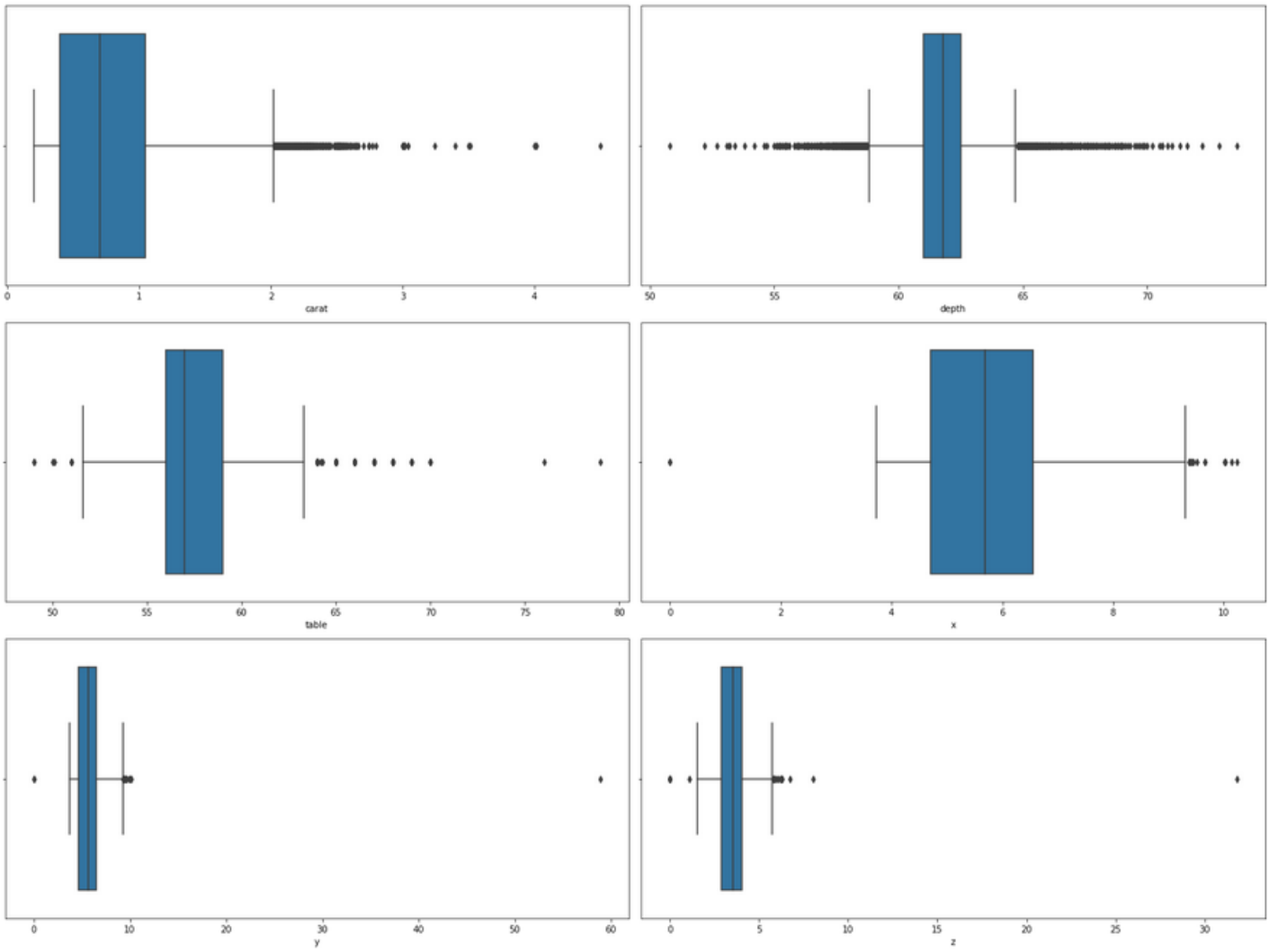
697 null values were present in the data set. These were imputed using the SimpleImputer from sklearn's library, using the median strategy.

5 POINT SUMMARY

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
count	26967.000000	26967.000000	26967	26967	26967	26270.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000
unique	NaN	NaN	5	7	8	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	Ideal	G	SI1	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	10816	5661	6571	NaN	NaN	NaN	NaN	NaN	NaN
mean	13484.000000	0.798375	NaN	NaN	NaN	61.745147	57.456080	5.729854	5.733569	3.538057	3939.518115
std	7784.846691	0.477745	NaN	NaN	NaN	1.412860	2.232068	1.128516	1.166058	0.720624	4024.864666
min	1.000000	0.200000	NaN	NaN	NaN	50.800000	49.000000	0.000000	0.000000	0.000000	326.000000
25%	6742.500000	0.400000	NaN	NaN	NaN	61.000000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	13484.000000	0.700000	NaN	NaN	NaN	61.800000	57.000000	5.690000	5.710000	3.520000	2375.000000
75%	20225.500000	1.050000	NaN	NaN	NaN	62.500000	59.000000	6.550000	6.540000	4.040000	5360.000000
max	26967.000000	4.500000	NaN	NaN	NaN	73.600000	79.000000	10.230000	58.900000	31.800000	18818.000000

0 values were observed for x,y,z variables, which is an anomaly since a 3D object cannot have any of its x,y,z dimensions as 0. This was fixed, by dropping such records having 0 values.

OUTLIER TREATMENT



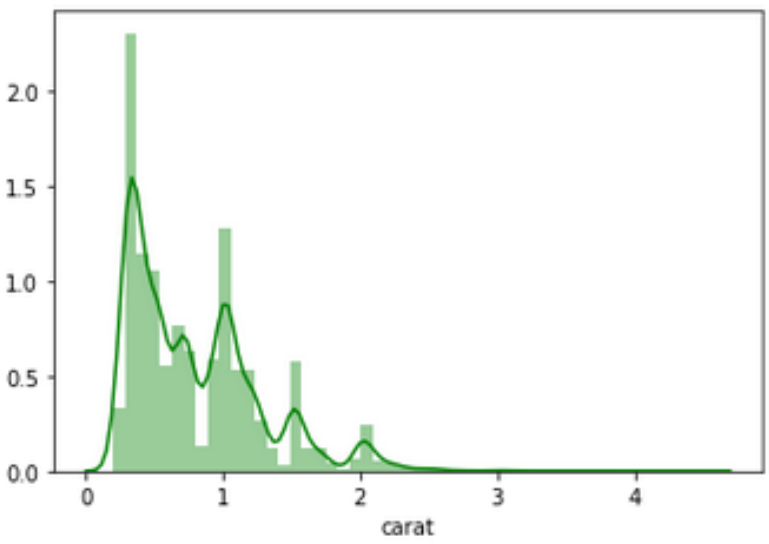
When we checked for over all outlier data, we could see a total of 4022 outliers in the entire data set, across all the columns. This formed a significant chunk of the entire data set.

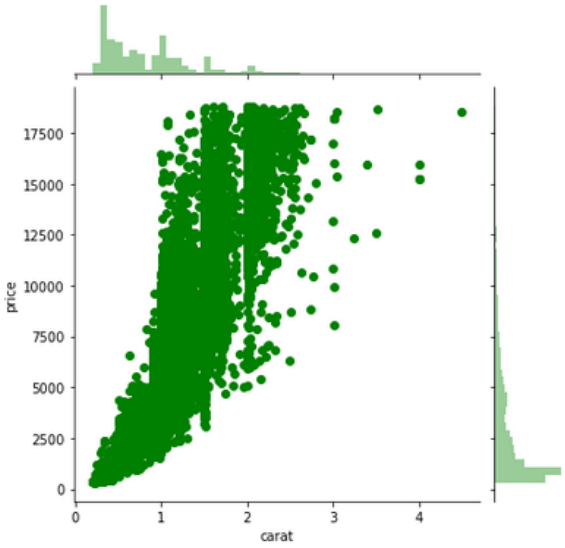
In fact this was 15% of the entire data set. Moreover when we checked for actual data, most of it seemed to be valid data. Altering such a huge amount of data using whatever strategy would have meant changing the essence of the entire data set.

Hence we opted to treat only the extreme outliers and leave the rest as it is. We found only 2 extreme outliers, 1 in column 'y' and 1 in 'z'. These rows were dropped.

UNIVARIATE & BIVARIATE ANALYSIS

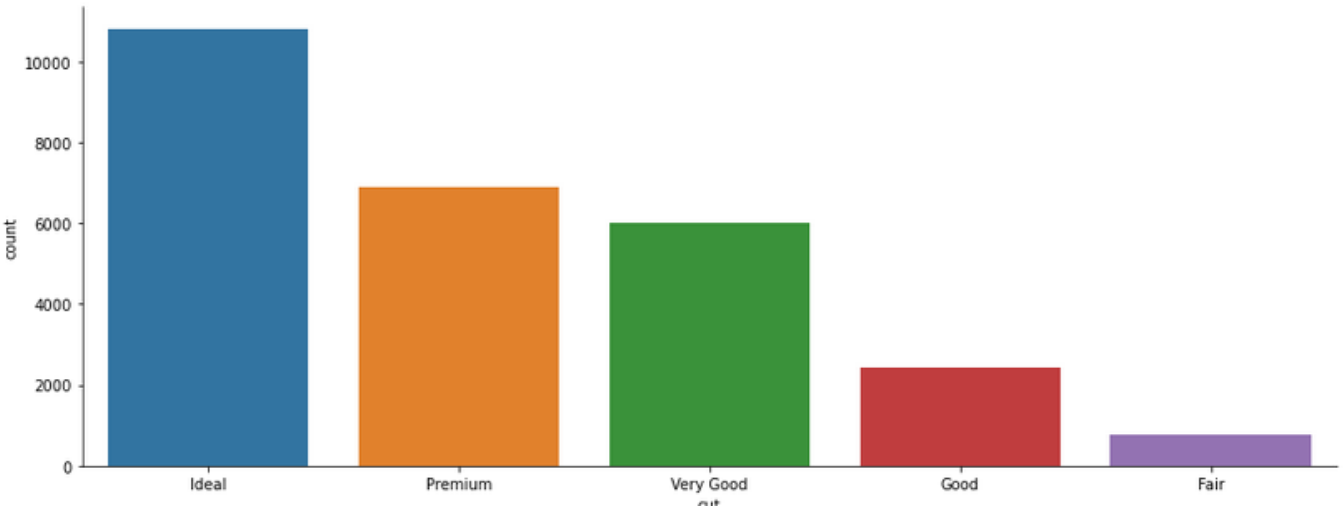
- ***Carat***

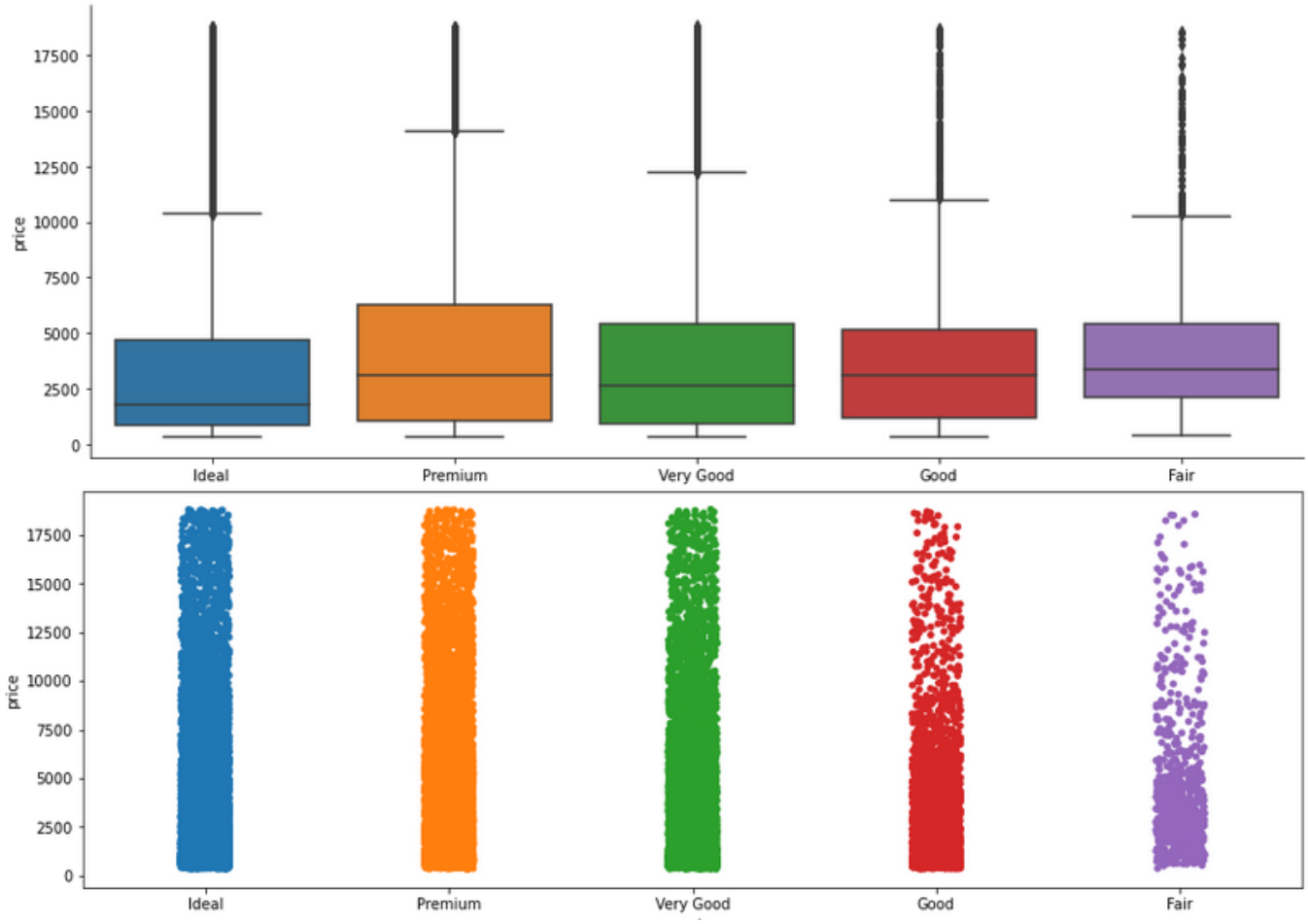




Carat variable is highly right skewed, as we can see from the above distplot. Also the correlation between price and carat is very high and both are directly correlated, i.e. when carat is increased price also increases.

- ***Cut***





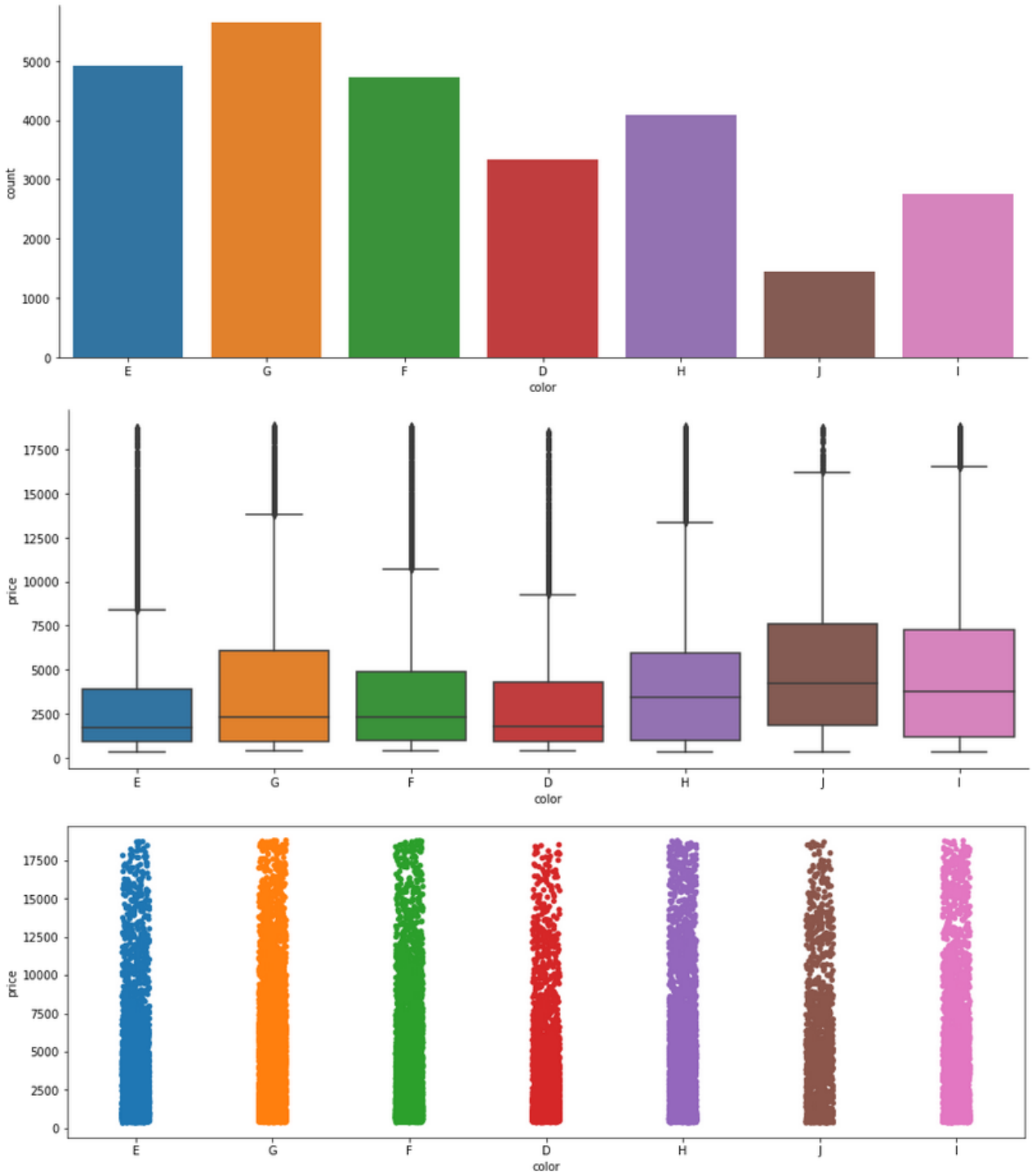
We can see the most common kind of cut is "Ideal" which is also the best possible cut. However upon closer inspection we see, even though "Ideal" is the best cut, its median price is lower than the others.

In fact the lowest quality cut i.e. "Fair" is having the highest median price. Also the number of diamonds with "Fair" cut are lesser in number, yet the median price is higher than the others.

What could be the possible reason for this anomaly as it clearly seems to be wrong.

We will discuss on this later in this section and find out the reason for this behaviour.

- ***Color***

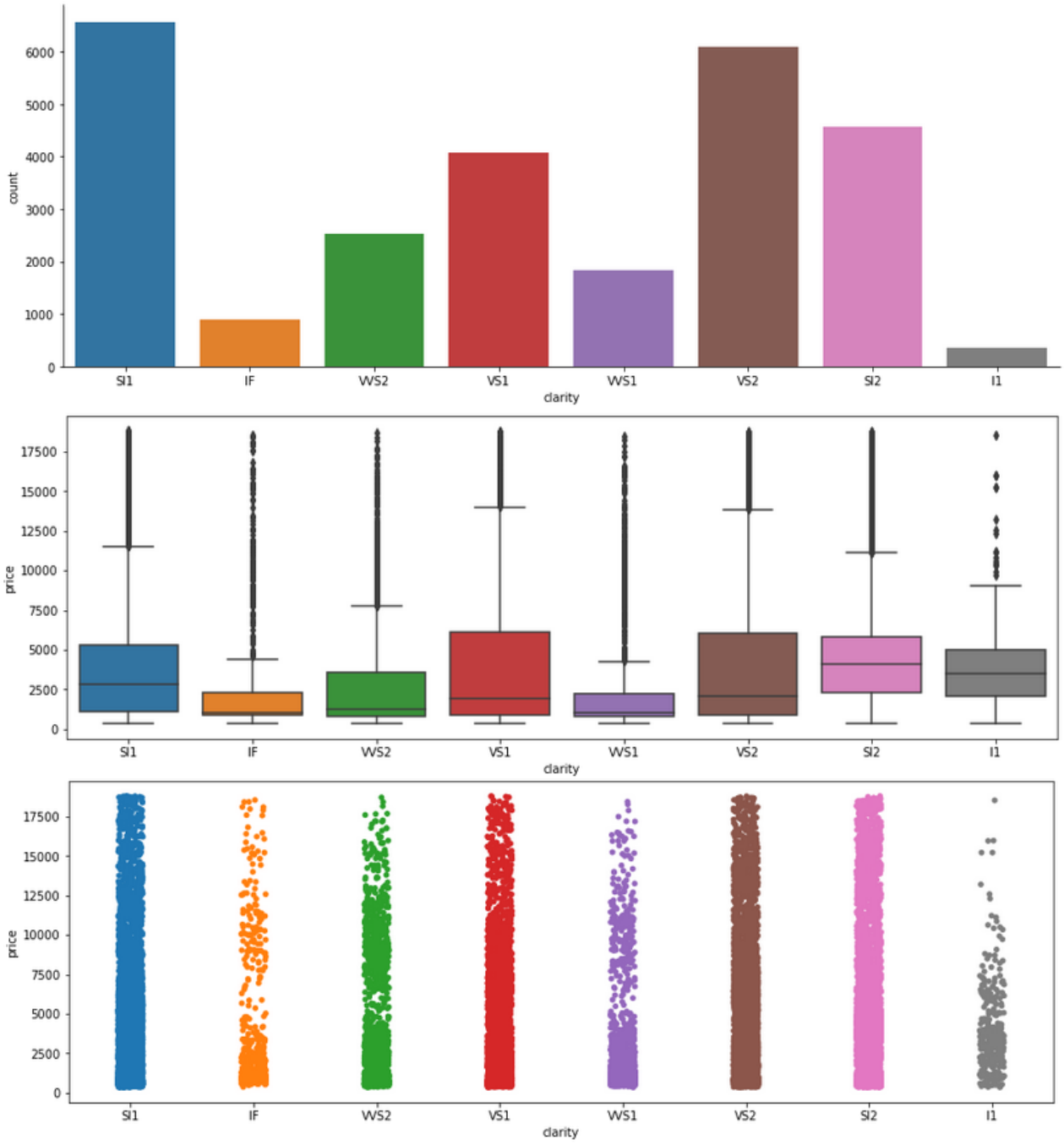


Diamonds with color code G are the most abundant.

Here too we can see the best quality color diamonds are not getting the best price, however the ones which are comparatively lower quality in terms of color are getting a better price.

Again, we will check this anomaly later.

- ***Clarity***

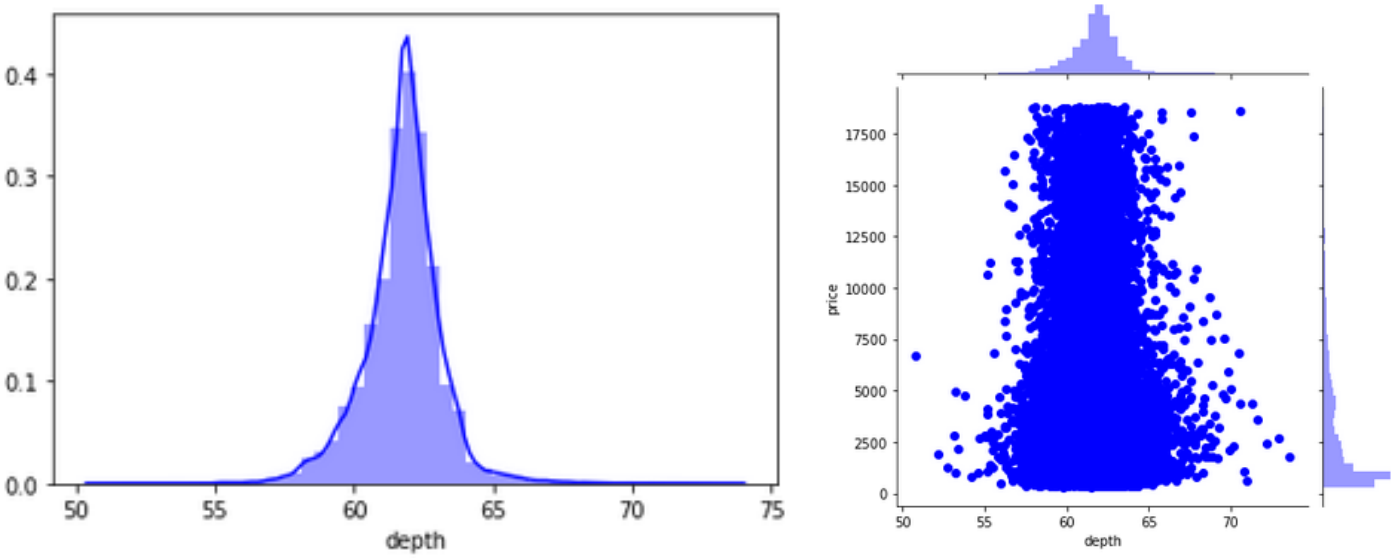


Diamonds with clarity code SI1 are the most abundant.

Here too we can see the best quality diamonds in terms of clarity are not getting the best price, however the ones which are comparatively lower quality in terms of clarity are getting a better price.

There seems to be a pattern emerging. We will take a look at this shortly.

- ***Depth***



Depth variable is slightly left skewed. It seems to be almost normally distributed. Also there is no clear correlation between depth and price variables.

- ***Table***

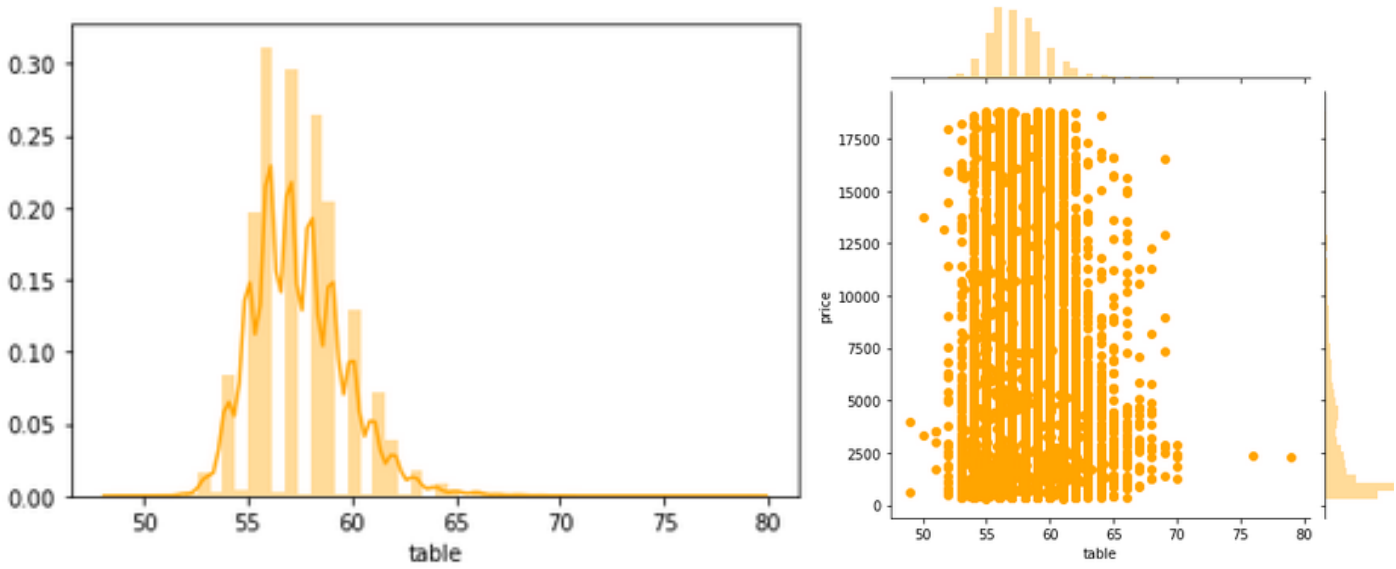
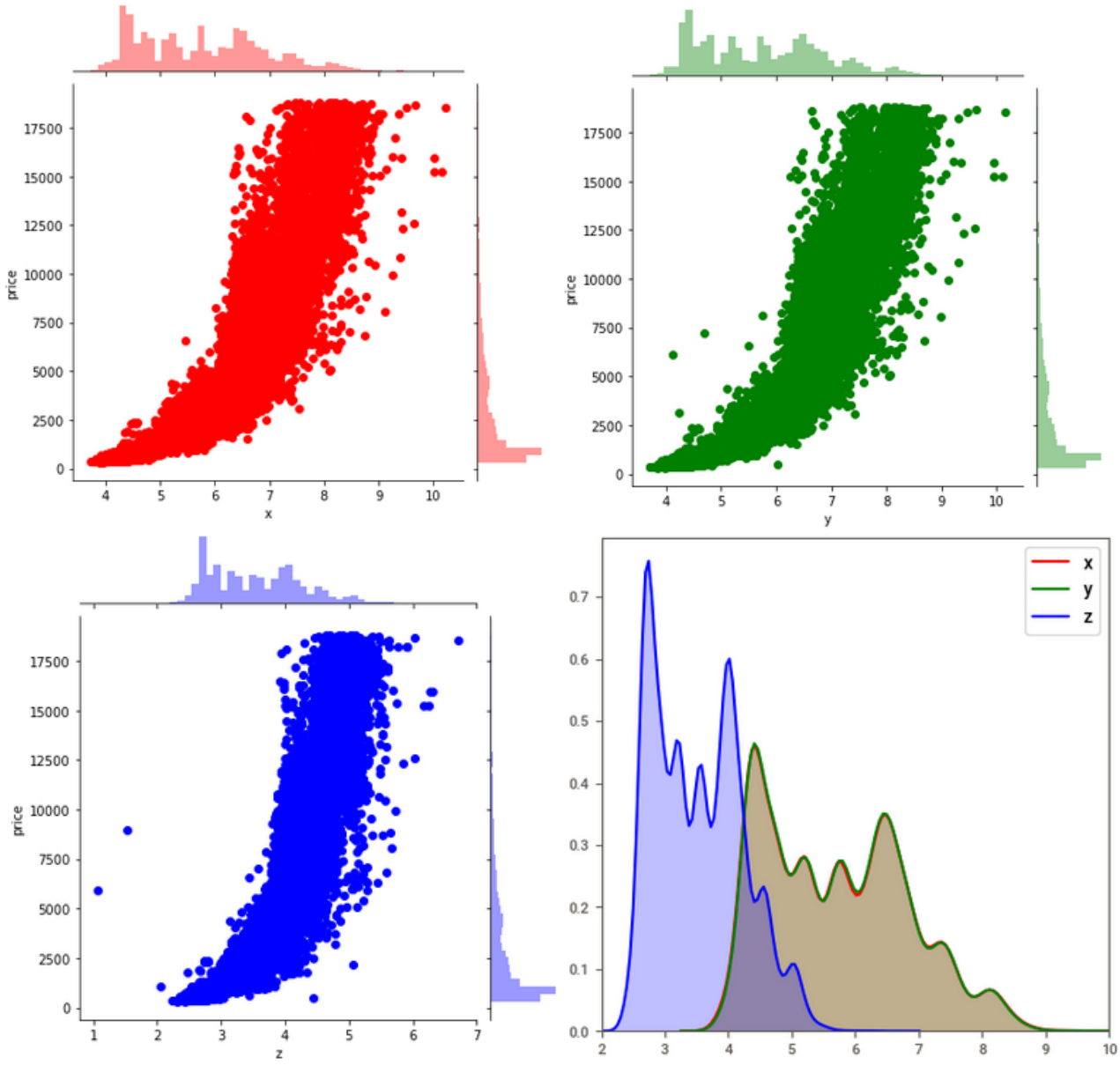


Table variable is slightly right skewed and is almost normal in nature. However there seems to be no strong correlation between the table variable and the price variable.

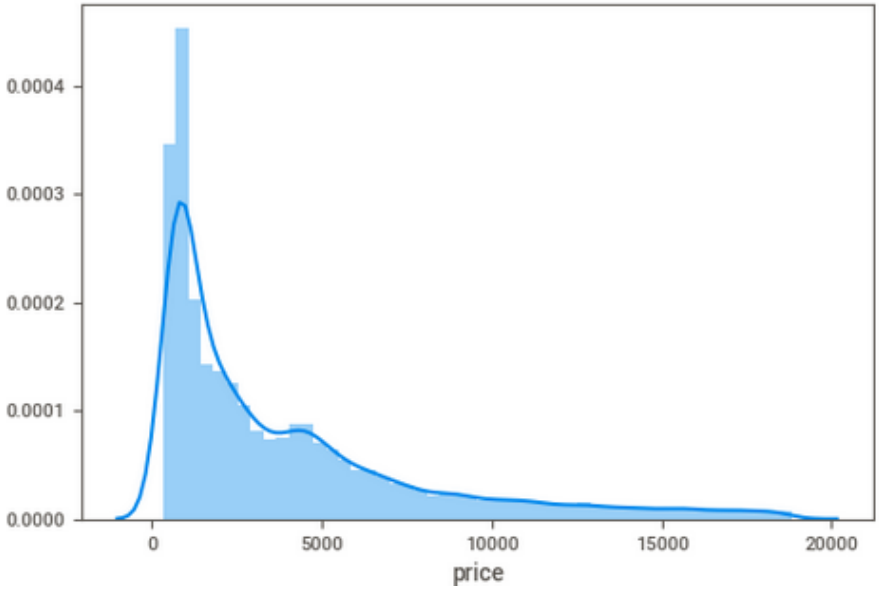
- ***x,y,z***



We can see all 3 variables - x,y & z are having almost the same distribution w.r.t the price variable. All of them seem to be highly correlated to price variable. Also the correlation seems to be direct in nature i.e. an increase in either of x,y, or z variables will lead to increase in price variable.

Looking at the last plot, we can see the distribution of x,y & z are almost identical, with the only difference being that the z variable values are a little on the lower side compared to x & y which are identical.

- ***Price***



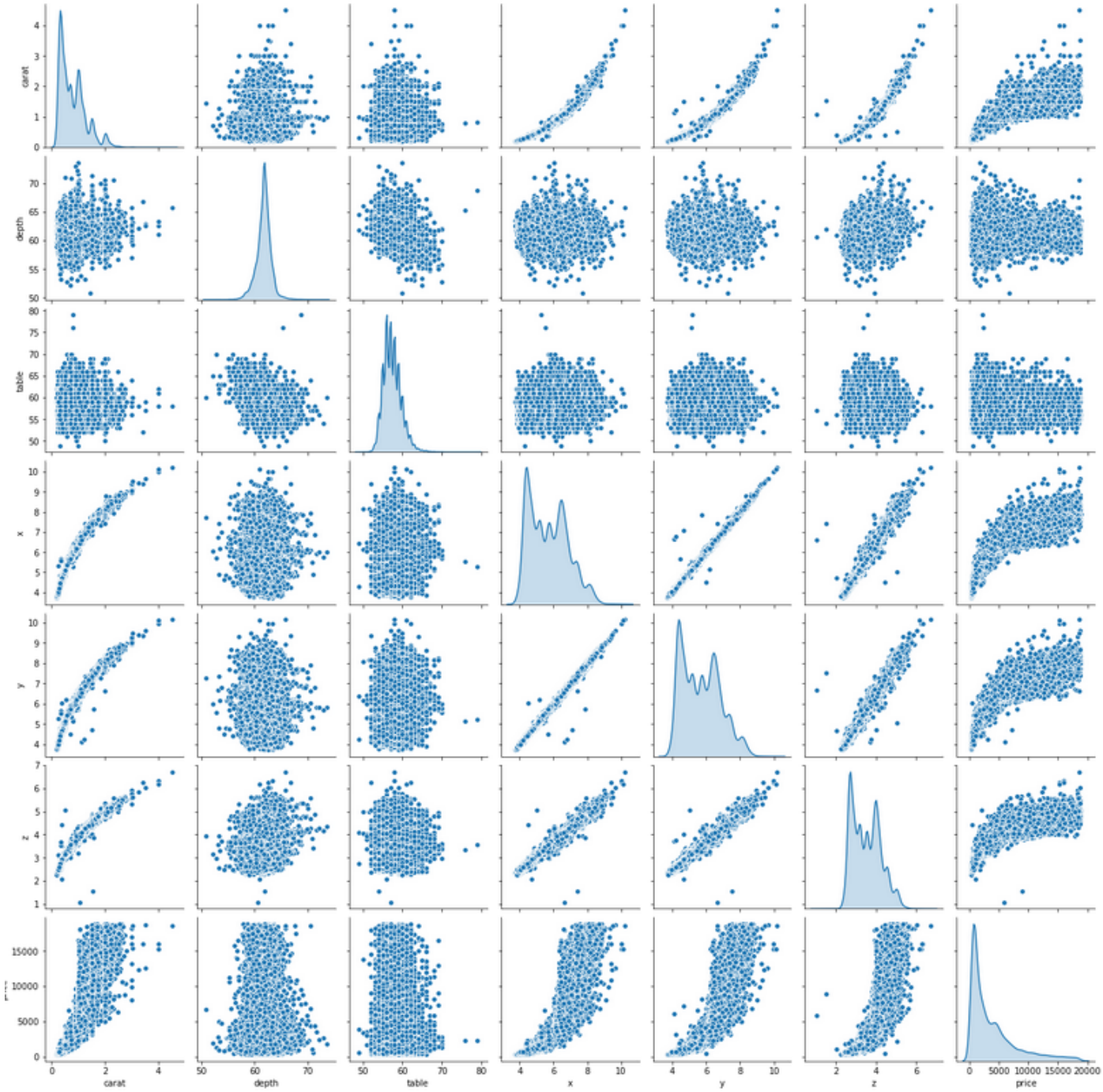
We can see the price variable is highly right skewed. Most of the prices lie between the range of 300 to 5000, with a few exception where the price go as high as 20,000.

- ***Skewness***

	Amount of Skewness
carat	1.114812
depth	-0.028314
table	0.764795
x	0.401915
y	0.397254
z	0.399903
price	1.619163

On the whole the data is mostly right skewed, with carat and price being highly right skewed, while the depth variable is slightly left skewed.

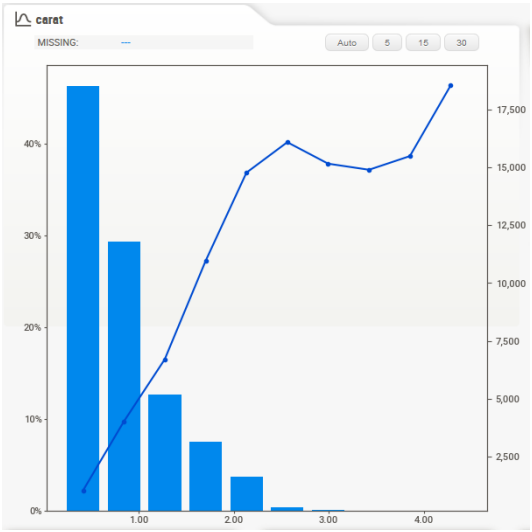
- ***Pairplot***

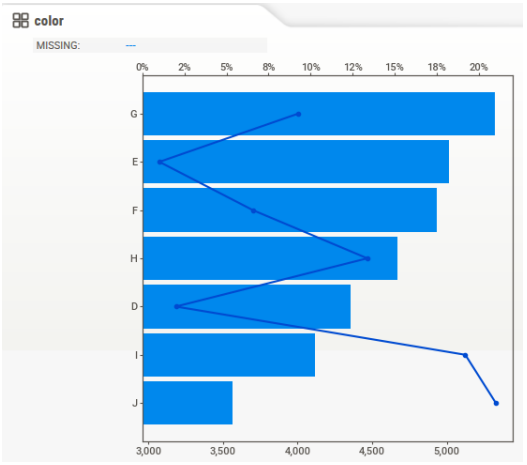


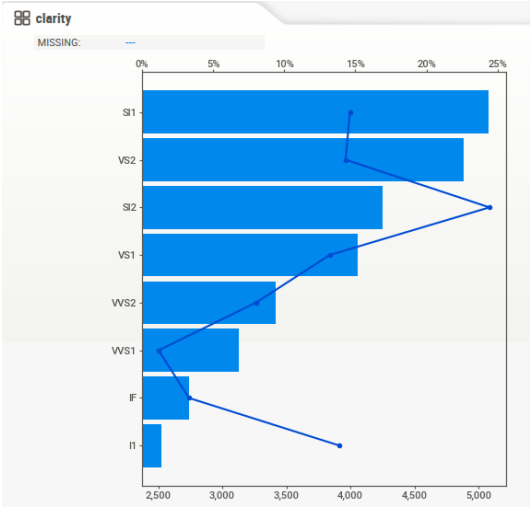
Looking at the above pair plot, we can see very correlation between x,y,z themselves and also with the price variable. Carat also seems to be highly correlated to price variable. Other variables do not paint such clear pictures.

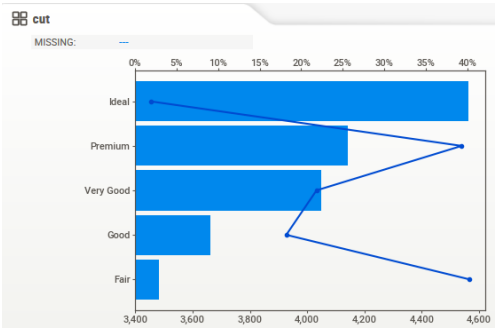
Correlation heat map will further clarify on this.

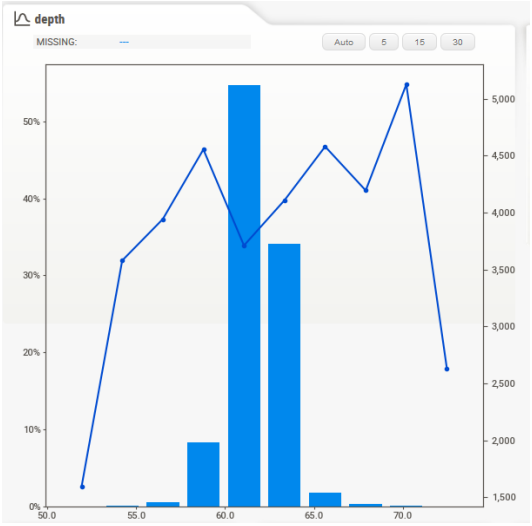
- ***SweetViz Report***

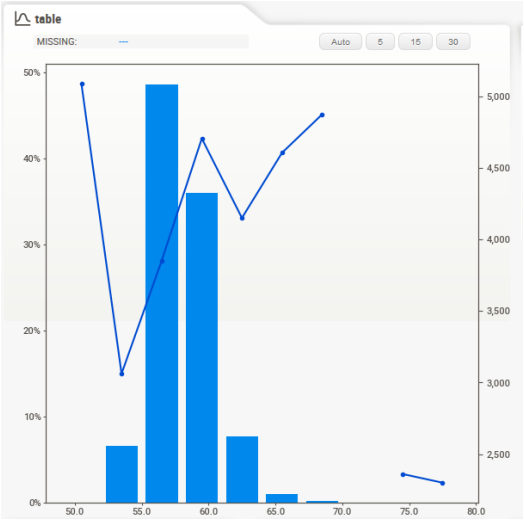


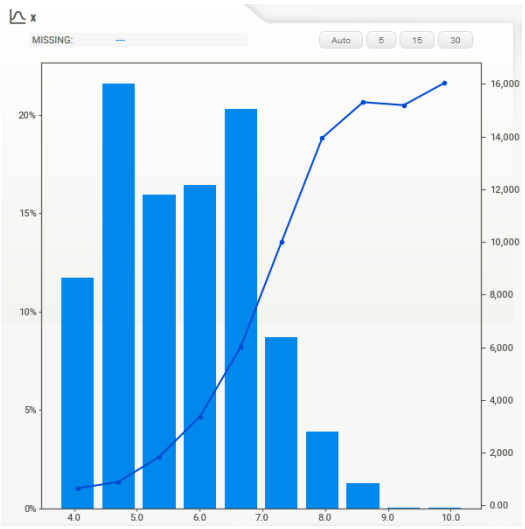


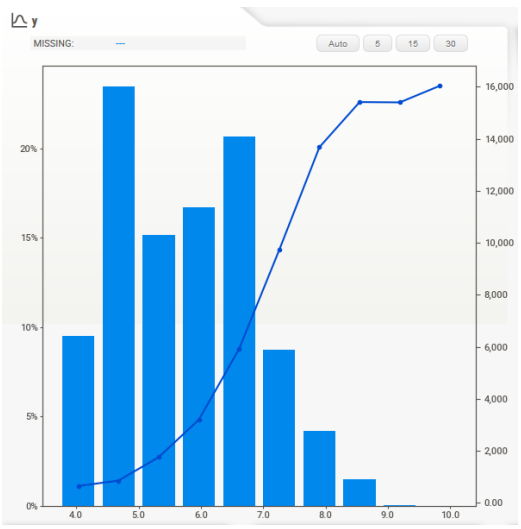


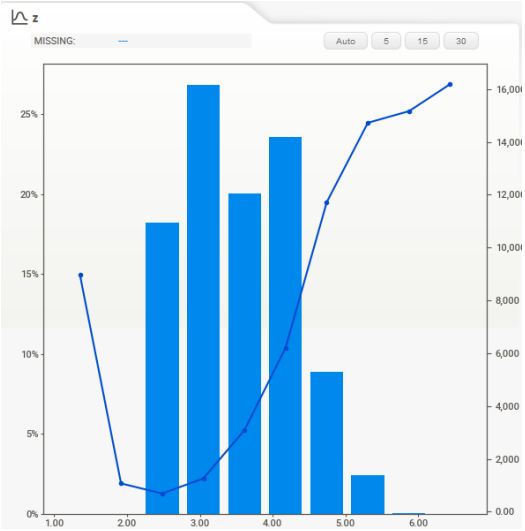












Carat - As the carat increases, price increases.

Cut - Irregular pattern observed, best quality not getting highest price, as observed earlier as well.

Color - Irregular pattern observed.

Clarity - Irregular pattern observed.

Depth - As the depth increases price also increases to some extent, however post 70.0 an increase in depth leads to sharp decline in price.

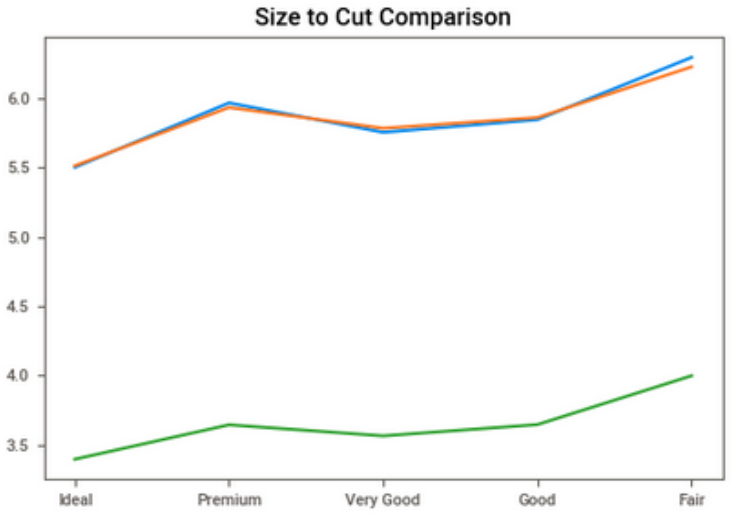
Table - Irregular pattern observed.

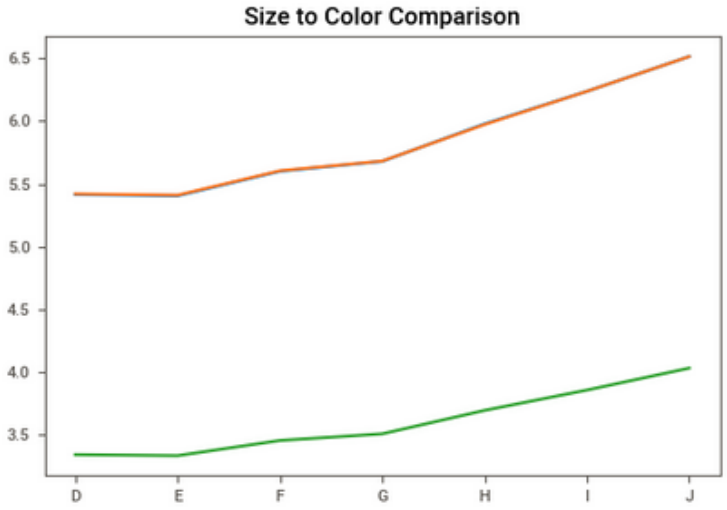
x,y,z - Similar pattern is observed in all 3 variables, i.e. an increase in either of x,y or z leads to an increase in price variable.

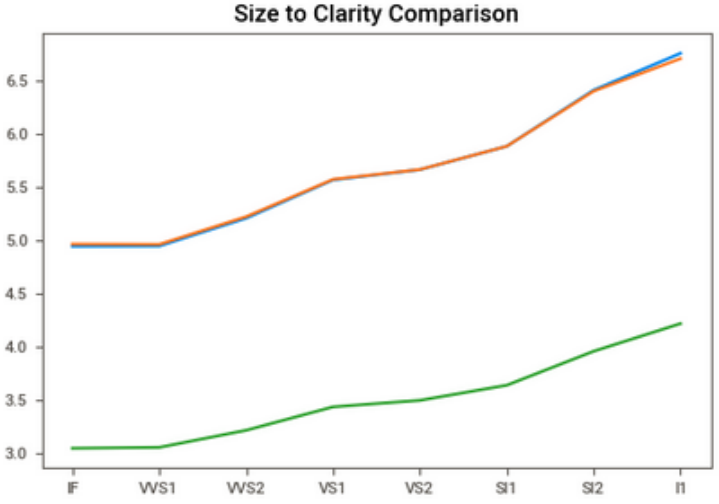
- ***Addressing the Anomaly***

We have earlier seen that diamonds having the best cut,color and clarity are fetching lower price than the diamonds have lower quality cut, color and clarity. Why is that so? This does not seem to make sense.

To find the reason we compare size with cut, color and clarity variables, as we already know size (x,y,z) is so far the biggest factor to determine price.







Moving from best to worst, i.e. from left to right in each of the three plots, we can see a clear trend.

Comparing cut with size we can see Ideal cut diamonds are smaller in size compared to Fair cut diamonds.

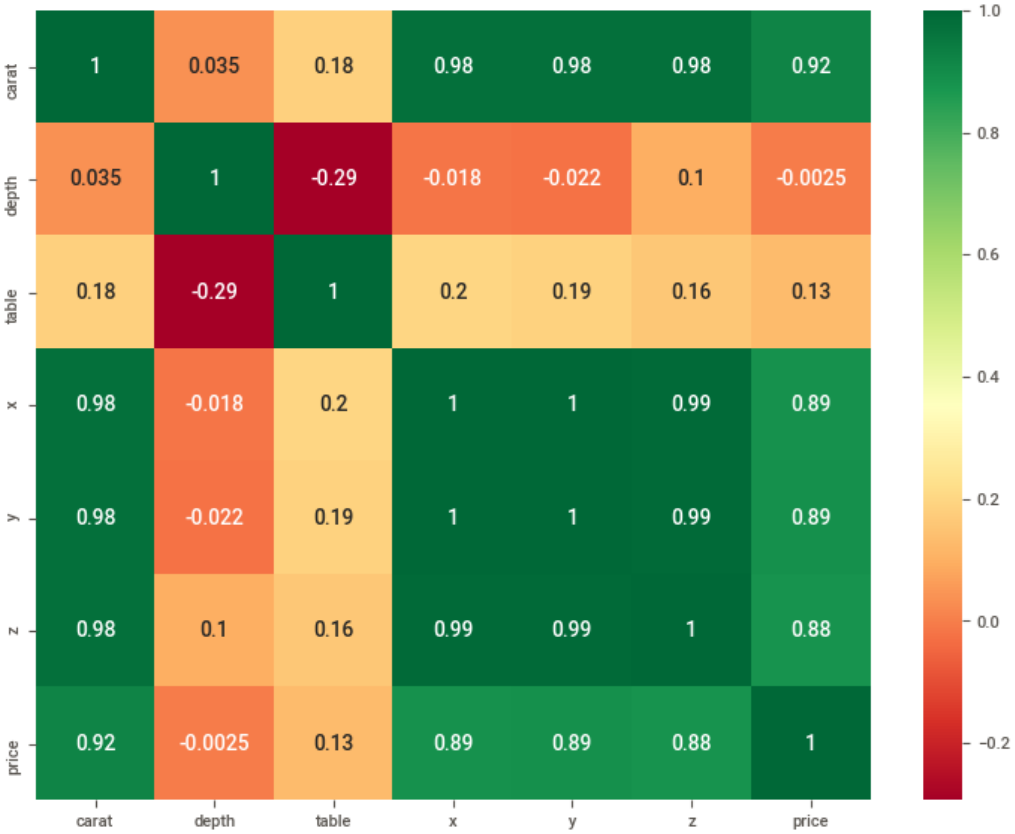
Comparing Color with size we can see best quality colored diamonds are smaller in size compared to lower quality colored diamonds.

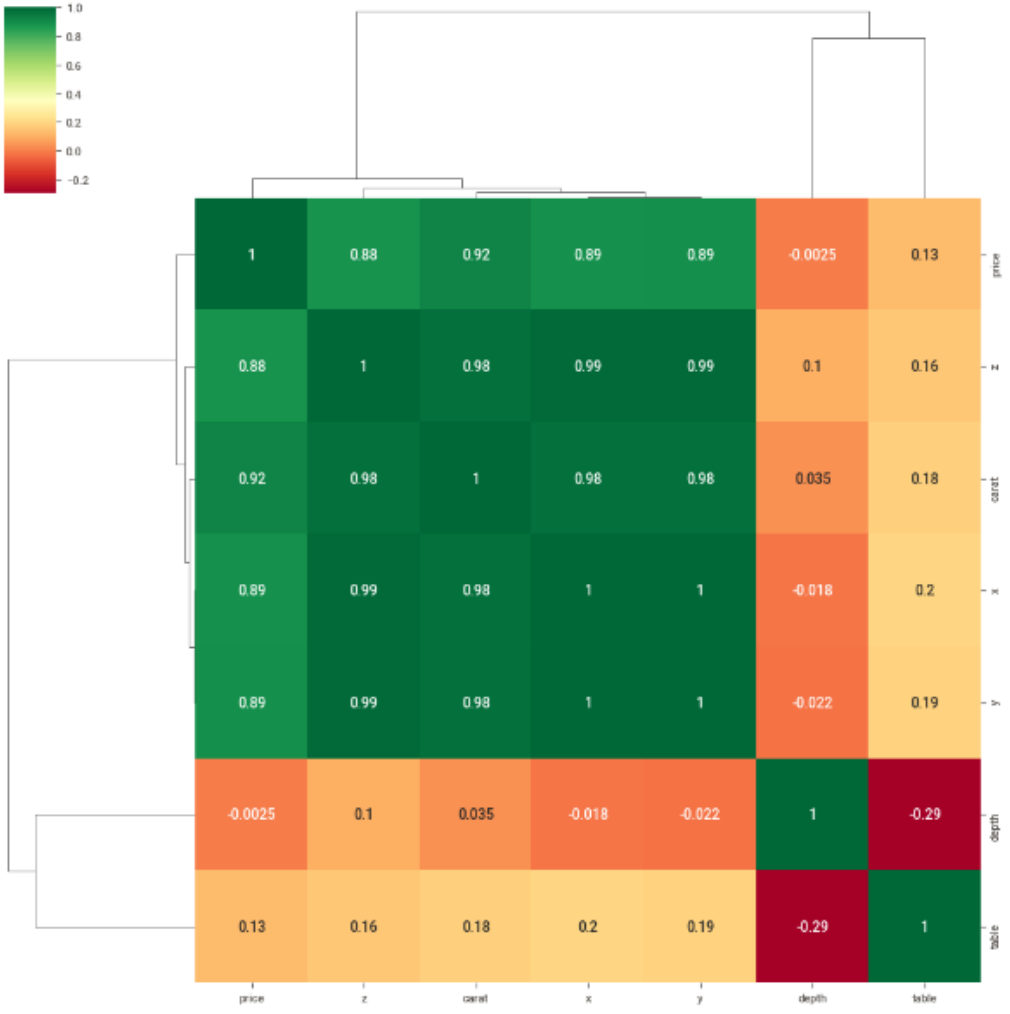
Similarly lower clarity diamonds are bigger in size and hence command higher price, while higher clarity diamonds are smaller in size and hence command lower price.

We can conclude that size is the more important variable compared cut, color and clarity.

We also observed that in general diamonds which are better in quality (i.e. color , cut and clarity) are generally smaller in size. Lower quality diamonds are bigger in size.

- ***Heatmap / Clustermap***





- ***Heatmap / Clustermap***

We cam see there is very high correlation between x,y,z and price. Also high correlation is observed between carat and price variable.

Depth and table variables have weak correlation with all other variables, while they have negative correlation with each other.

We can observe the high correlation represented by the bright green square shown on the top left of the cluster map, while the negative correlations between depth and table are represented by the red boxes on the bottom ride hand side of the cluster map.

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?

A total of 697 null values were found in depth variable. These were imputed using Sklearn's SimpleImputer using median value as the strategy.

Also we found a total of 8 records where value in either of x,y & z were 0. In practice these would not be valid values, as no 3D object can have any of the x , y or z axis values to be zero.

Since these values are invalid from practical usage perspective, we chose to drop all the 8 rows.

At the end of imputing the 697 records in depth variable, dropping 8 rows due to 0 values in x,y,& z variables and also removing the two extreme outliers which was mentioned in the sections above, we were left with a total of 26923 rows which contained valid data.

Scaling was required on this data set as **the variables had different measurement units**, some had units in mm while some are weight variables having different measurement unit. Hence we had to scale the variables.

We tried multiple approaches to scale the data here.

Approach 1 - Scale only numerical variables, keeping categorical variables untouched.

Approach 2 - Scale all variables, including categorical.

Approach 3 - Scale only independent variables, keeping dependent variable untouched.

We then compared the various parameters like rsquare, AIC, BIC, fstat and rmse to zero down on the final appraoch.

- ## *Approach 1 - Scale only numerical variables*

In this approach we only scaled the numerical variables, leaving the categorical variables as it is in their ordinally encoded format.

After applying linear regression using sklearn as well as statsmodel we found below:-

Train Data:-
AIC - 8780
BIC - 8843
RSquare - 0.907
Fstat - 21680
RMSE - 0.303

Test Data:-
AIC - 4071
BIC - 4127
RSquare - 0.907
Fstat - 11210
RMSE - 0.3112

- ## *Approach 2 - Scale all variables*

In this approach we scaled all the variables, including all categorical as well as numerical. Dependent as well as independent.

After applying linear regression using sklearn as well as statsmodel we found below:-

Train Data:-
AIC - 8780
BIC - 8843
RSquare - 0.907
Fstat - 21680
RMSE - 0.303

Test Data:-
AIC - 4071
BIC - 4127
RSquare - 0.907
Fstat - 11210
RMSE - 0.3112

- ***Approach 3 - Scale only independent variables***

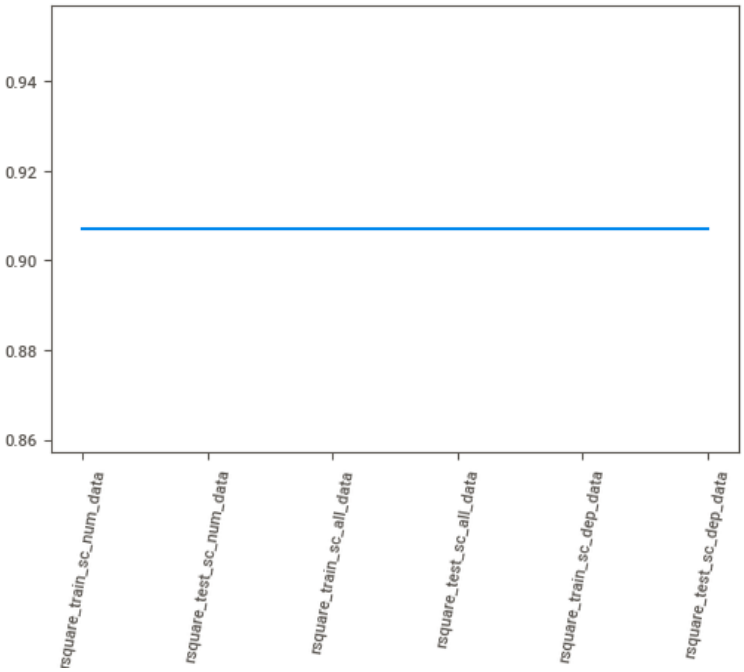
In this approach we only scaled the independent variables, leaving the dependent variable i.e. price as it is.

After applying linear regression using sklearn as well as statsmodel we found below:-

Train Data:-
AIC - 312400
BIC - 138000
RSquare - 0.907
Fstat - 21680
RMSE - 1220

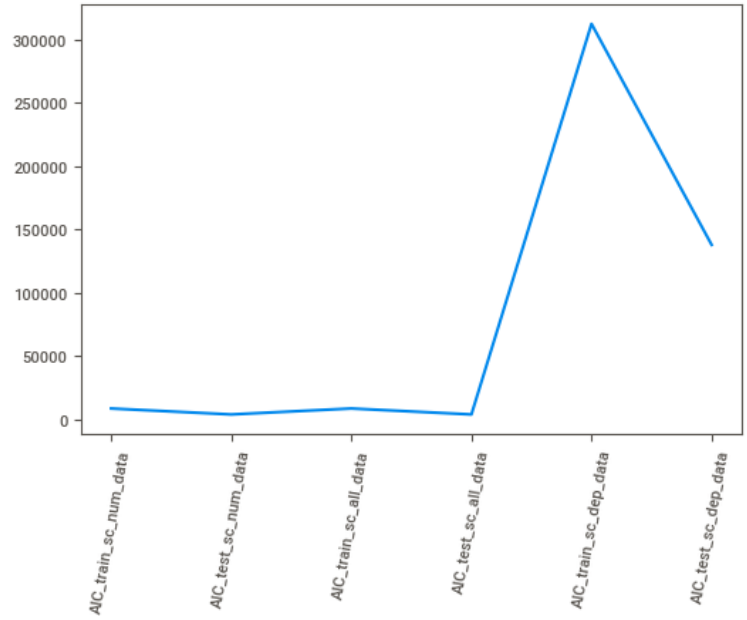
Test Data:-
AIC - 312400
BIC - 138100
RSquare - 0.907
Fstat - 11210
RMSE - 1244

- ***RSquare Comparison***



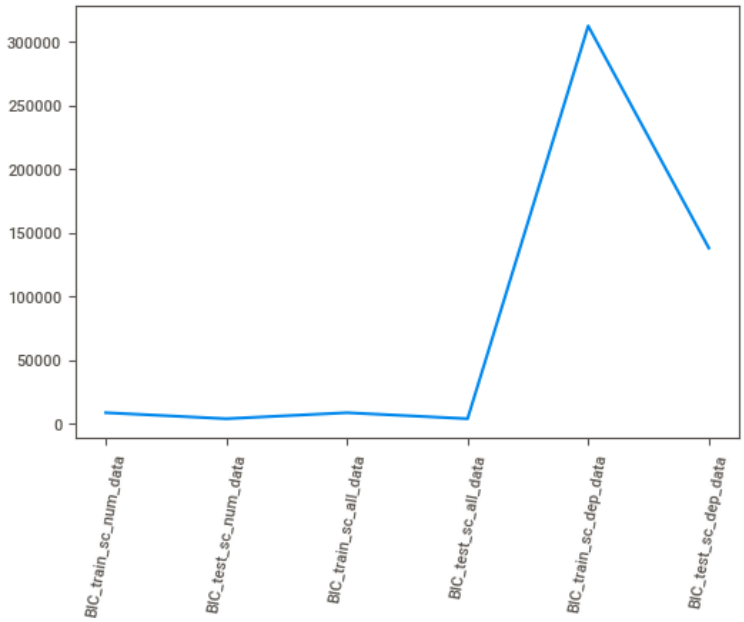
RSquare remained constant through out the different approaches.

- ***AIC Comparison***



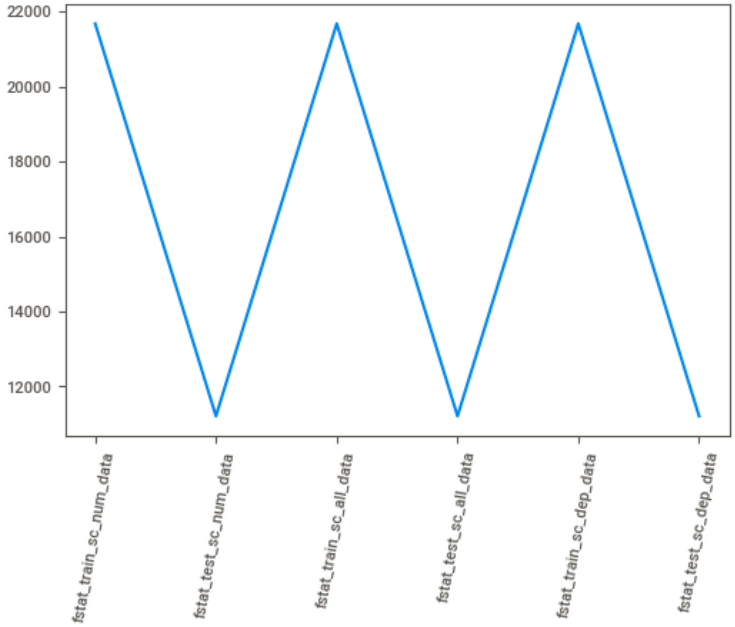
AIC values for approach 1 & 2 were identical while for approach 3, they were on the higher side.

- ***BIC Comparison***



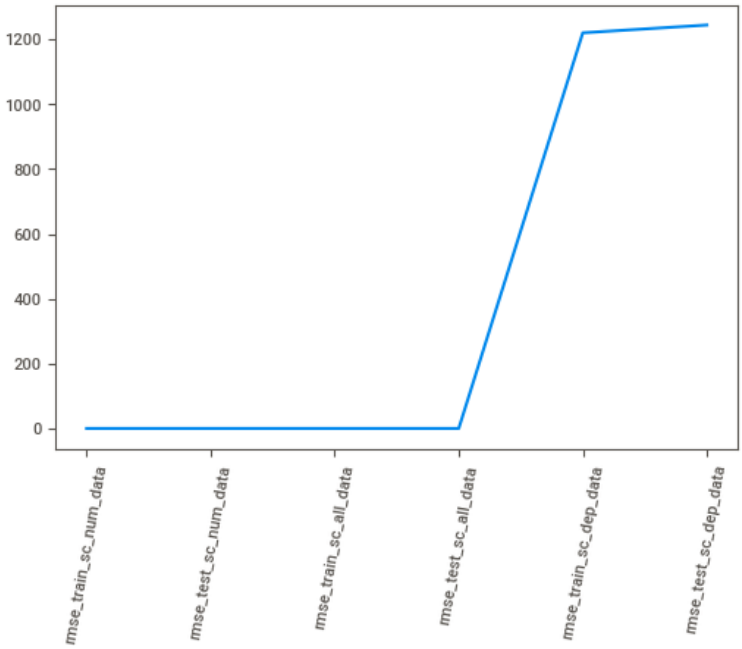
BIC values for approach 1 & 2 were identical while for approach 3, they were on the higher side.

- ***FStat Comparison***



FStat remained constant through out the different approaches.

- ***RMSE Comparison***



RMSE values for approach 1 & 2 were identical while for approach 3, they were on the higher side.

- ***Conclusion***

In approach 3, we could see very high values for AIC, BIC, Fstat and RMSE. Lower AIC/BIC values are typically preferred in the industry, hence this approach was rejected.

Between approach 1 and approach 2, almost everything was identical from AIC, BIC, RMSE, Fstat to rsquare values. However the intercept value was 0 in approach 2 while we had a small negative value.

Hence we finally selected approach 2 since it gave best values for all the parameters and for simplicity of scaling all the variables.

Refer the graphs below for comparison of the above mentioned parameters.

1.3

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

Categorical variables - Cut, Color & Clarity having string values were encoded manually using map function keeping the ordinality in mind. Series map function was passed a dictionary containing the corresponding mappings.

Data was then split into the ratio of 70:30 as per the project requirement. Random state of 22 was used to split the values.

Linear regression model was then applied using various approaches as mentioned earlier. Various performance metrics like AIC/ BIC/ Fstat/ RSquare/ RMSE etc were captured. In this section we will discuss on the values obtained in the finalized approach i.e. approach 2.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.907
Model:	OLS	Adj. R-squared:	0.907
Method:	Least Squares	F-statistic:	2.618e+04
Date:	Sun, 30 Aug 2020	Prob (F-statistic):	0.00
Time:	12:31:49	Log-Likelihood:	-4382.0
No. Observations:	18846	AIC:	8780.
Df Residuals:	18838	BIC:	8843.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4.792e-17	0.002	-2.15e-14	1.000	-0.004	0.004
carat	0.1887	0.055	3.437	0.001	0.081	0.296
cut	0.0290	0.003	10.735	0.000	0.024	0.034
color	0.1377	0.002	58.469	0.000	0.133	0.142
clarity	0.2154	0.002	88.377	0.000	0.211	0.220
depth	0.0041	0.003	1.458	0.145	-0.001	0.010
table	-0.0030	0.003	-1.026	0.305	-0.009	0.003
volume	0.8561	0.055	15.647	0.000	0.749	0.963

Omnibus:	3790.441	Durbin-Watson:	1.983
Prob(Omnibus):	0.000	Jarque-Bera (JB):	68893.823
Skew:	0.474	Prob(JB):	0.00
Kurtosis:	12.318	Cond. No.	54.6

OLS Regression Summary report for training data is as above.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.907
Model:	OLS	Adj. R-squared:	0.907
Method:	Least Squares	F-statistic:	1.121e+04
Date:	Sun, 30 Aug 2020	Prob (F-statistic):	0.00
Time:	12:31:50	Log-Likelihood:	-2027.6
No. Observations:	8077	AIC:	4071.
Df Residuals:	8069	BIC:	4127.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0035	0.003	1.006	0.315	-0.003	0.010
carat	0.1714	0.093	1.841	0.066	-0.011	0.354
cut	0.0366	0.004	8.565	0.000	0.028	0.045
color	0.1395	0.004	38.079	0.000	0.132	0.147
clarity	0.2152	0.004	56.382	0.000	0.208	0.223
depth	7.246e-05	0.004	0.016	0.987	-0.009	0.009
table	-0.0029	0.005	-0.647	0.518	-0.012	0.006
volume	0.8817	0.093	9.517	0.000	0.700	1.063

Omnibus:	1574.802	Durbin-Watson:	2.020
Prob(Omnibus):	0.000	Jarque-Bera (JB):	15869.588
Skew:	0.641	Prob(JB):	0.00
Kurtosis:	9.746	Cond. No.	60.0

OLS Regression summary report for test data is as shown above.

A compilation of all the critical performance metrics is shown below for both training as well as testing data set. It includes performance metrics like RSqaure, RMSE, AIC, BIC & Fstat.

	Train	Test
RSqaure	0.9070	0.9070
AIC	8780.0000	4071.0000
BIC	8843.0000	4127.0000
FStat	21680.0000	11210.0000
RMSE	0.3053	0.3112

A Rsquare value of 0.9070 indicates 90.70% of overall variance was captured by this linear regression model, which is very good.

1.4

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

The final linear regression equation that we get for the test data from the model that we created is as below.

Price = 0.1714*carat + 0.0366*cut + 0.1395*color + 0.2152*clarity + 0.00007246*depth - 0.0029*table + 0.8817*volume

Now since the p-values for **depth** and **table** are greater than 0.05, hence we can conclude that their coefficients do not represent our universe and are a mere statistical fluke and hence cannot be considered for making business inferences.

We consider all the remaining coefficients.

Also since we had standardized our data before applying the regression model, our interpretation should be of the below form.

1) Carat - A coefficient of 0.1714 explains that one standard deviation increase in the carat variable on an average causes 0.1714 standard deviation increase in the log odds of the **Price** variable.

2) Cut - A coefficient of 0.0366 explains that one standard deviation increase in the cut variable on an average causes 0.0366 standard deviation increase in the log odds of the **Price** variable.

3) Color - A coefficient of 0.1395 explains that one standard deviation increase in the color variable on an average causes 0.1395 standard deviation increase in the log odds of the **Price** variable.

4) Clarity - A coefficient of 0.2152 explains that one standard deviation increase in the clarity variable on an average causes 0.2152 standard deviation increase in the log odds of the Price variable.

5) Volume - A coefficient of 0.8817 explains that one standard deviation increase in the volume variable on an average causes 0.8817 standard deviation increase in the log odds of the Price variable.

Now when we rank the coefficients from the linear equation we get below.

	Co-efficients
Volume	0.8817
Clarity	0.2152
Carat	0.1714
Color	0.1395
Cut	0.0366

Looking at the above we can clearly see the engineered column x,y,z has the most direct impact on the price variable. Hence it would be safe to say now with statistical proofs that variables x,y & z have the most direct impact on the price variable. In other words an increase in x,y, or z variable is most likely to cause increase in price variable.

Second most important variable is the clarity, followed by Carat, Color and then finally the cut variable.

In order to increase the profit margins it would be advisable to the organization to deal more in larger stones i.e. to concentrate more on the size of the diamond than on the quality of the diamonds.

If quality is to be a factor then clarity needs to be given the preference over all other quality parameters in order to maximize the profits.

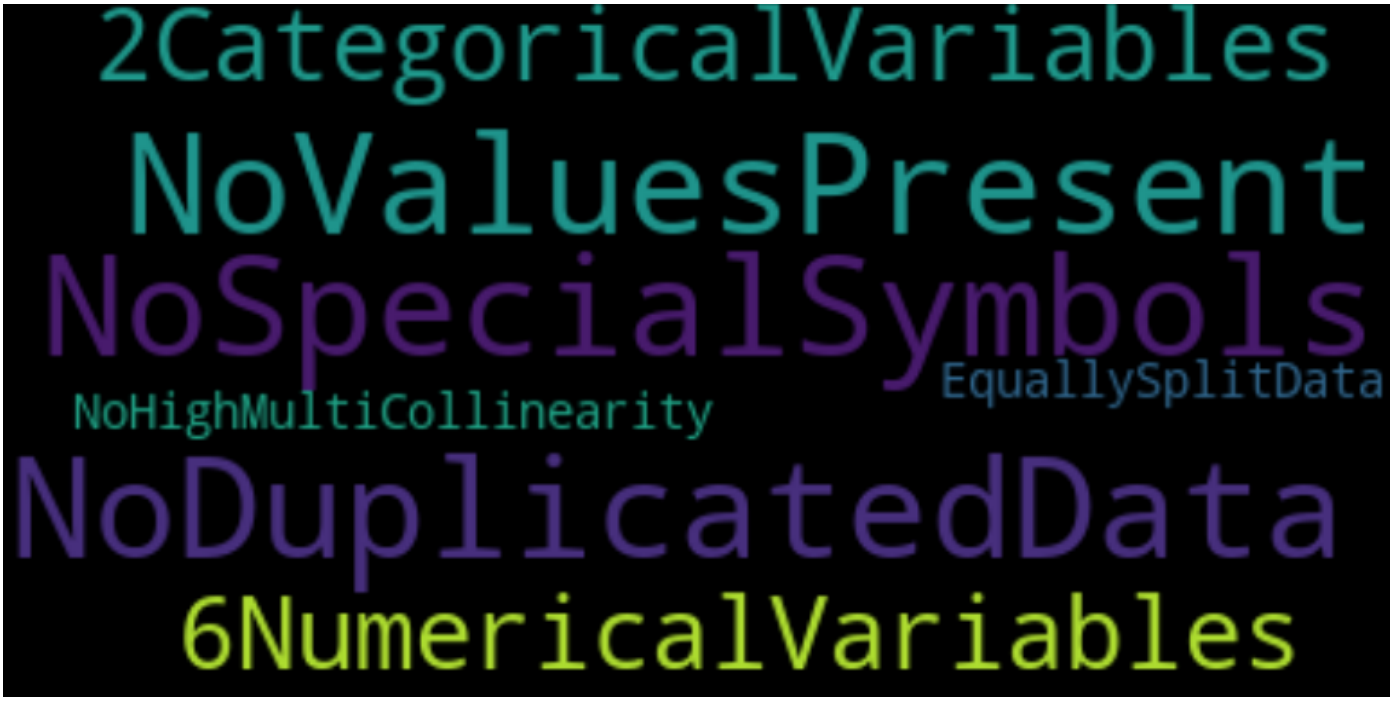
QUESTION - 2

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.

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.

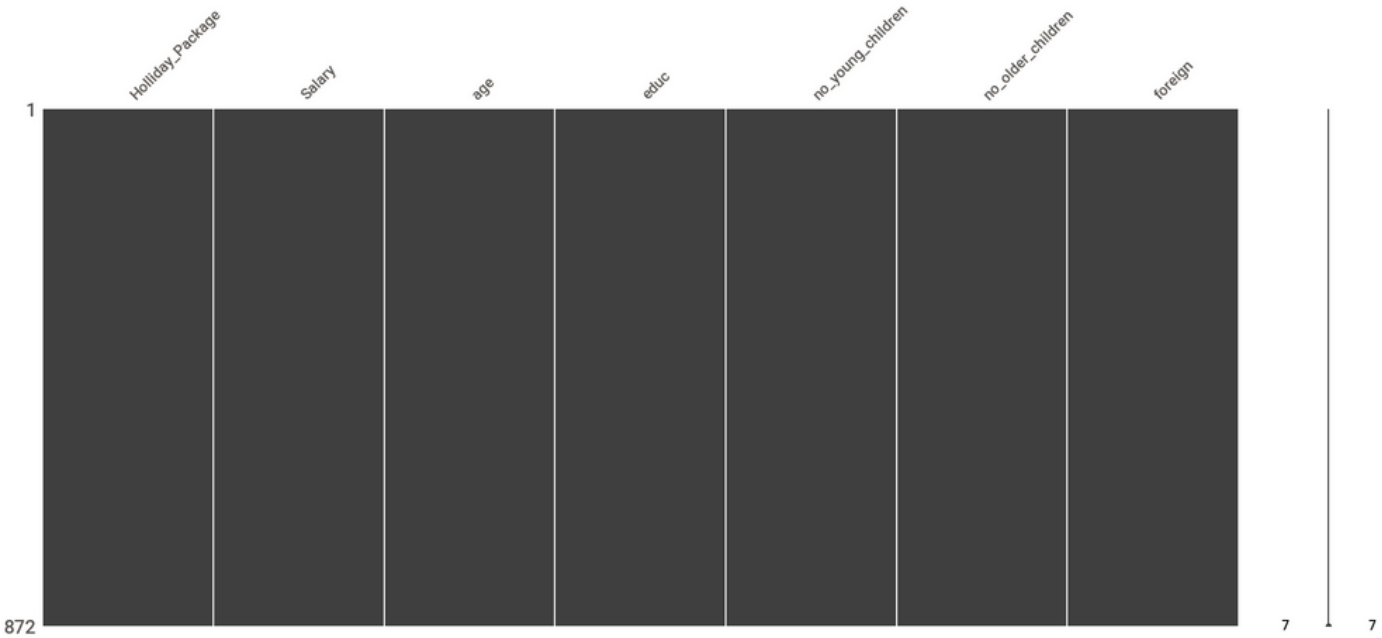
The holiday package data set shared with us had 872 records. Below are some of the salient features of this data set.



There are a total of 8 variables present in the data set. 2 columns out of which are categorical in nature containing values yes & no, while the other 6 are numerical in nature.

```
<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
```

NO NULL VALUES PRESENT



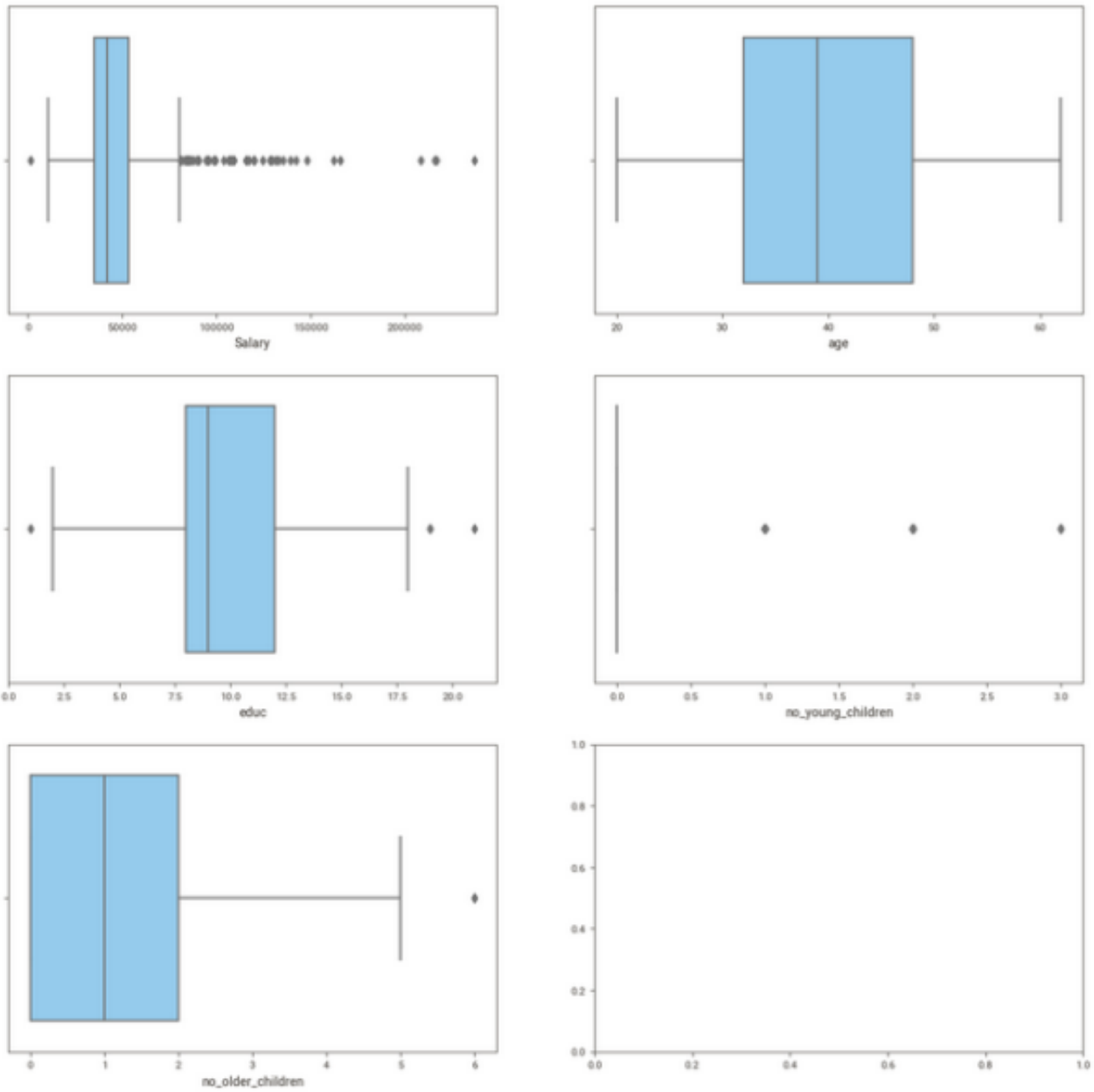
No Null values are present in the data set, as can be seen from the figure above.

5 POINT SUMMARY

	Unnamed: 0	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
count	872.000000	872	872.000000	872.000000	872.000000	872.000000	872.000000	872
unique	NaN	2	NaN	NaN	NaN	NaN	NaN	2
top	NaN	no	NaN	NaN	NaN	NaN	NaN	no
freq	NaN	471	NaN	NaN	NaN	NaN	NaN	656
mean	436.500000	NaN	47729.172018	39.955275	9.307339	0.311927	0.982798	NaN
std	251.869014	NaN	23418.668531	10.551675	3.036259	0.612870	1.086786	NaN
min	1.000000	NaN	1322.000000	20.000000	1.000000	0.000000	0.000000	NaN
25%	218.750000	NaN	35324.000000	32.000000	8.000000	0.000000	0.000000	NaN
50%	436.500000	NaN	41903.500000	39.000000	9.000000	0.000000	1.000000	NaN
75%	654.250000	NaN	53469.500000	48.000000	12.000000	0.000000	2.000000	NaN
max	872.000000	NaN	236961.000000	62.000000	21.000000	3.000000	6.000000	NaN

We can see the mean and median are comparable for Salary, age and educ variables indicating they are more or less normal in nature. 0 values in no_young_children & no_older_children are perfectly acceptable values, hence not treated.

OUTLIER TREATMENT

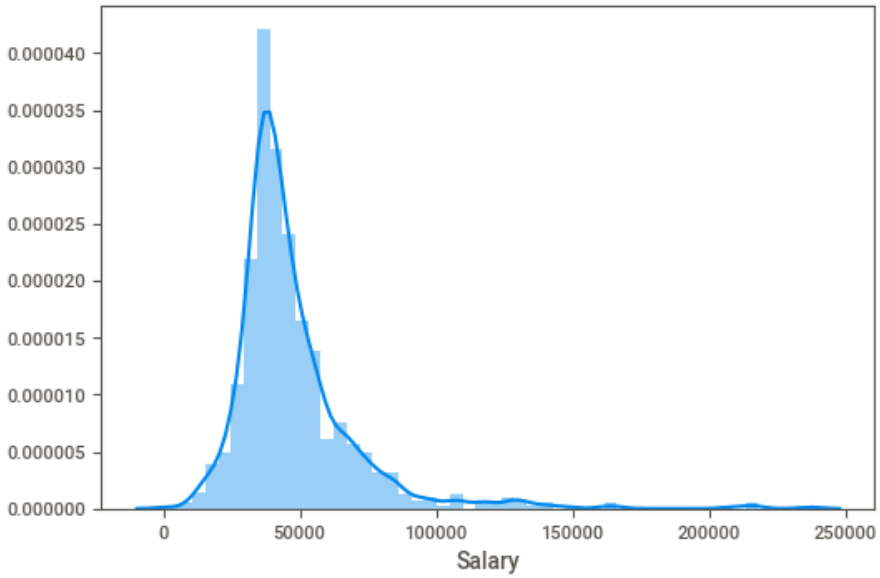


No major outlier data was identified in the data set. Salary column had a total of 57 outliers however on analyzing the data, we could see the data is genuine and should not be treated as it would change the essence of the data.

No significant outliers found in the data set, hence outlier treatment is not undertaken in this data set.

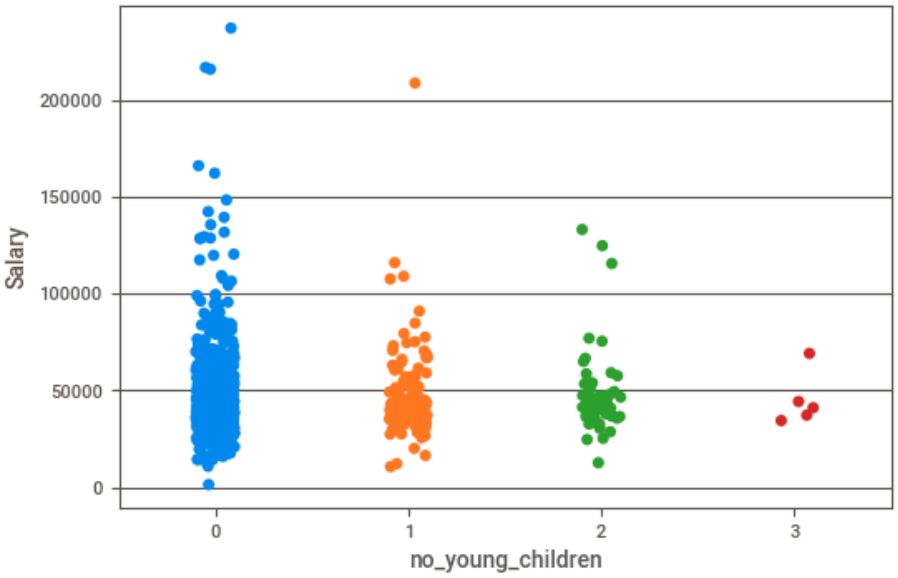
UNIVARIATE & BIVARIATE ANALYSIS

- ***Salary***



we can see most of the people have salaries ranging from a few thousands to up to 1,00,000. While there are a few people who have salaries going up to a maximum of 2,36,000 approximately. Salary variable is highly right skewed.

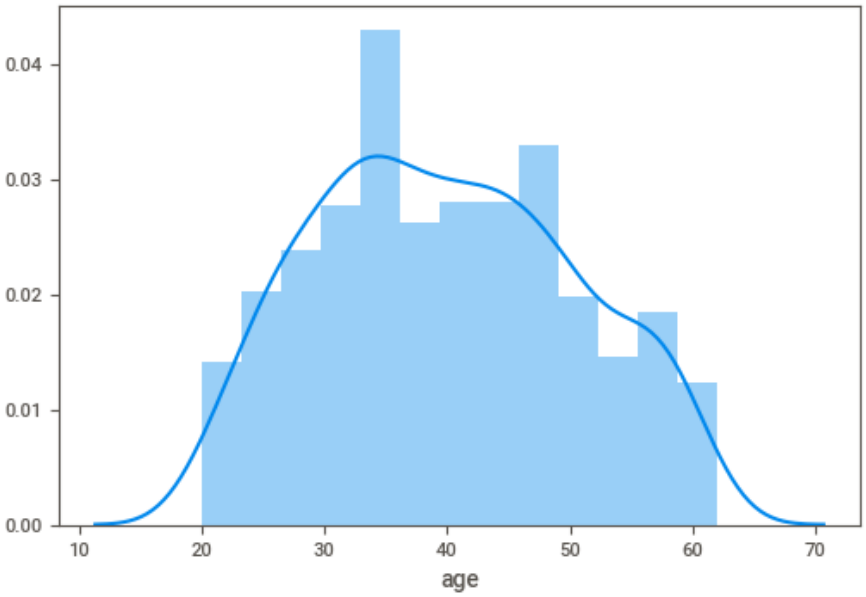
Interestingly not a single person having salary above 1,50,000 opted for the holiday package as per the figure above.



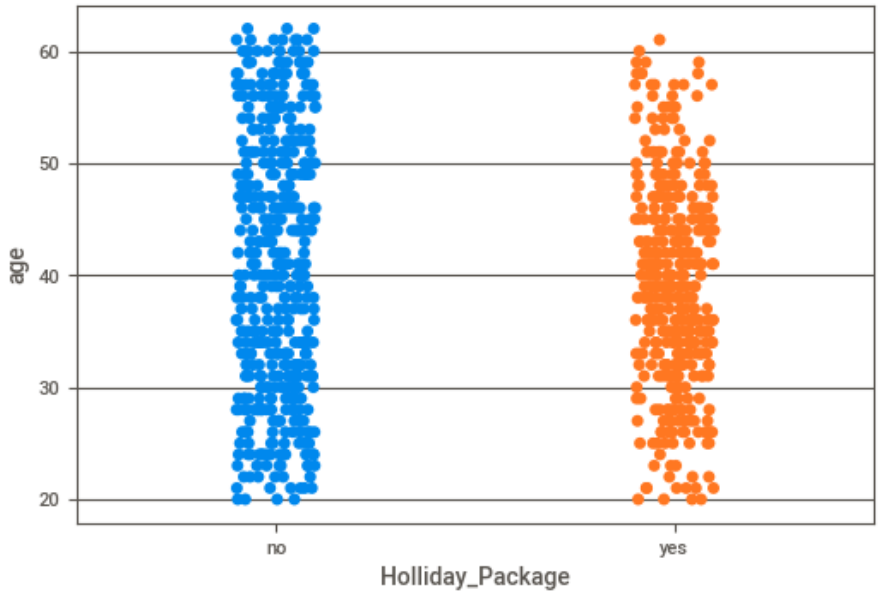
From the data set we can see employees having less number of young children tend to have higher salary compared to the employees who had more number of young children.

No one above the salary of 1,50,000 had more than 1 young children, probably because these employees are senior in age and have more older children than younger children, which also explains their higher salaries.

- ***Age***

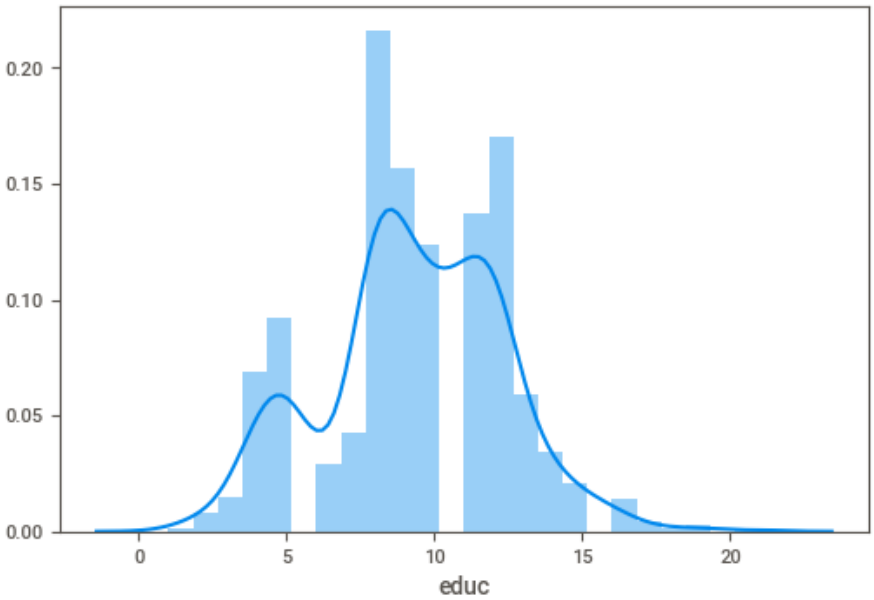


Age variable is more or less normally distributed. age variable is slightly right skewed.

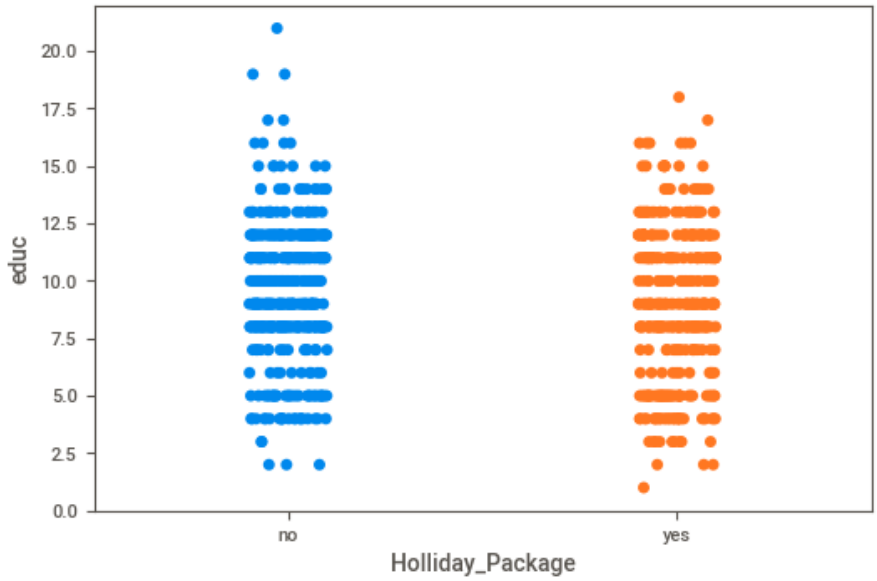


People opting in and out of the holiday package are almost equally distributed amongst the different age groups.

- ***Educ***

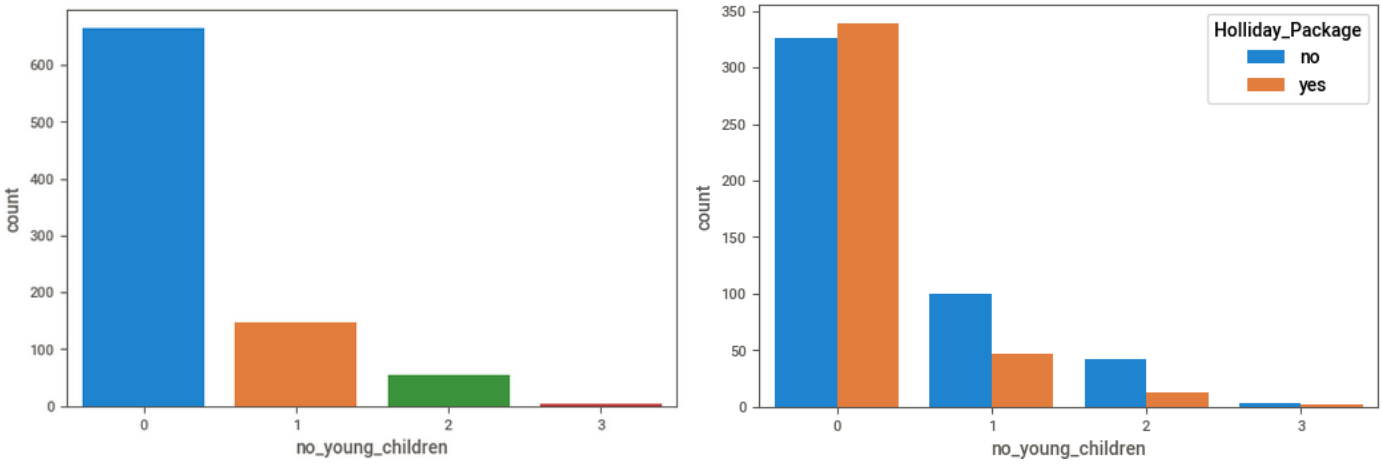


The educ variable almost normally distributed. However it is slightly left skewed in nature.



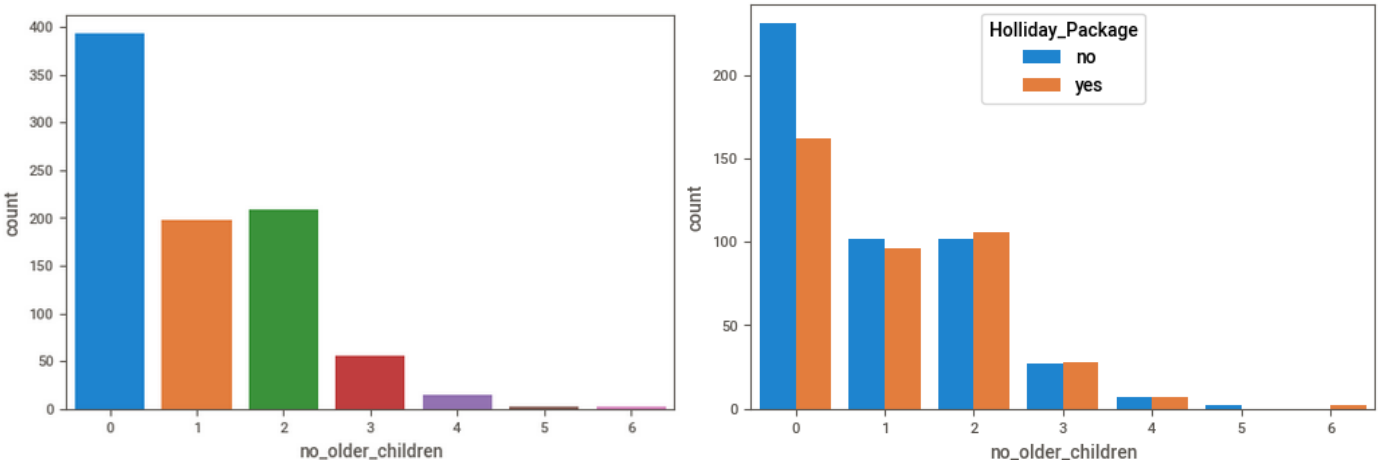
Also no significant difference in the education level of the people opting in and out of the holiday package.

- ***no_young_children***



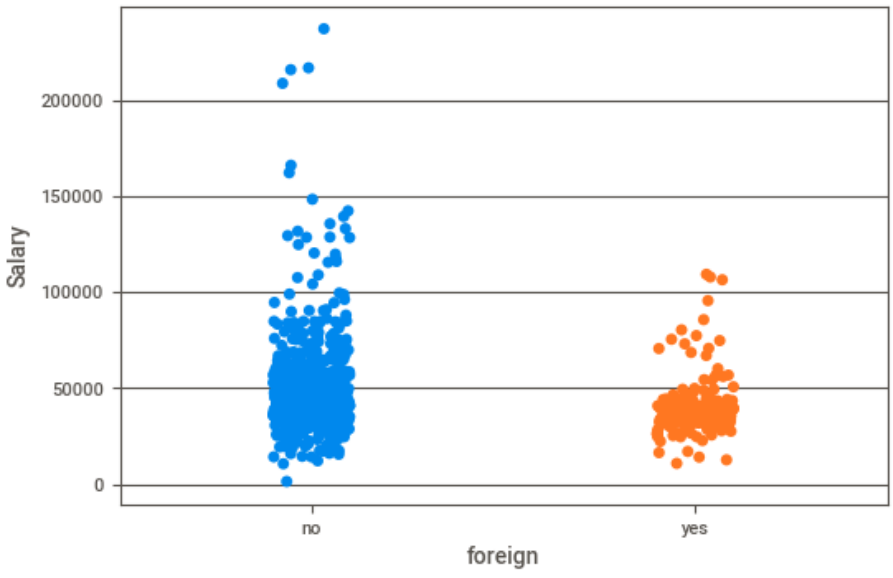
From the above figure we can clearly see that employees having no young children are more likely to opt for the holiday package than the employees having 1 or more young children.

- ***no_older_children***

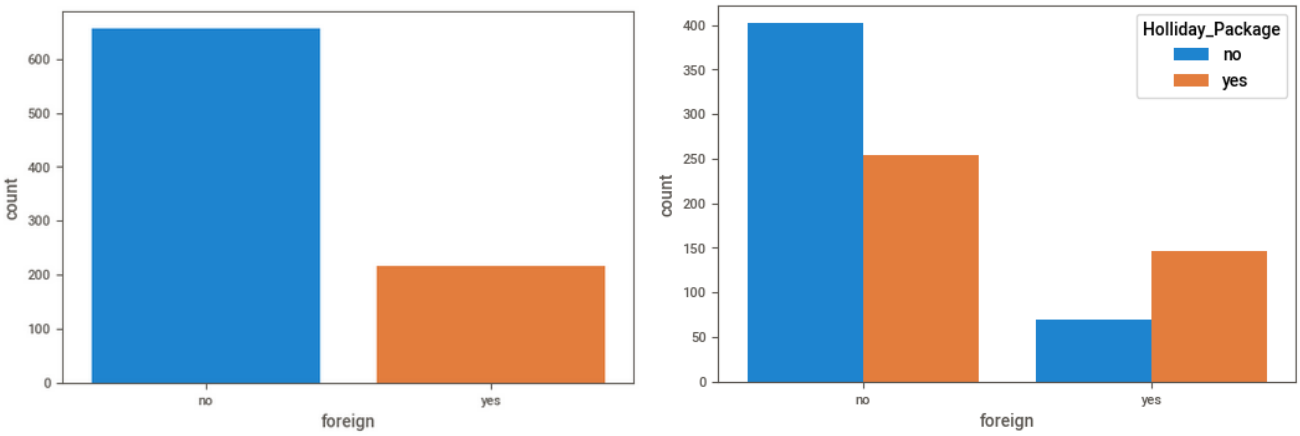


From the above figure, we can conclude that people having no older children are very likely to opt out of the holiday package, however employees having 1 or more older children are almost equally likely to opt in or out of the holiday package.

- ***foreign***



From the given data set, we can see the employees of foreign origin command lesser salary than their Indian / native counter parts.



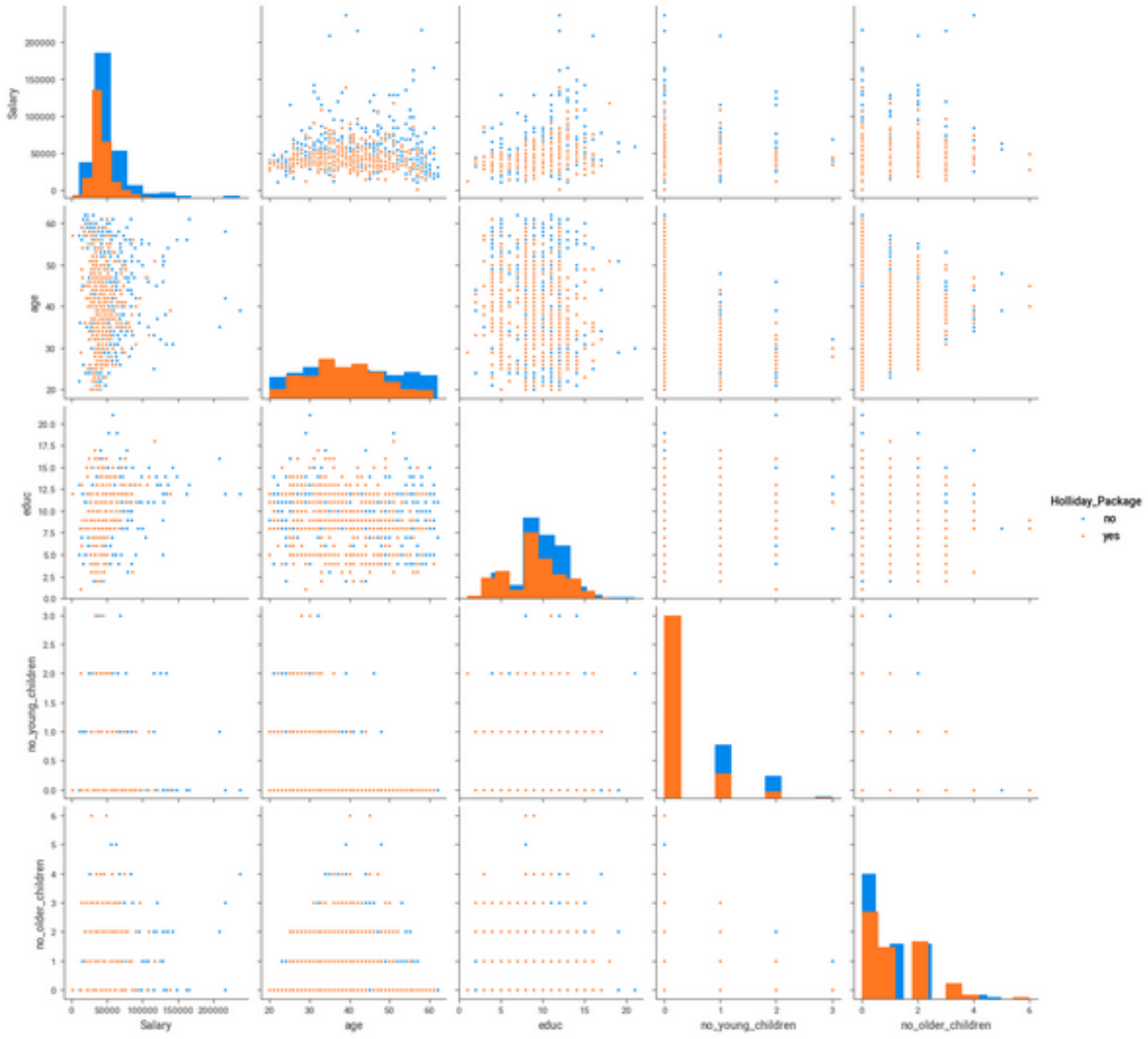
Given the above figure, we can conclude that employees of foreign origin are more likely to opt for the holiday package compared to their indian counter parts.

- ***Skewness***

	Amount of Skewness
Salary	3.097875
age	0.146160
educ	-0.045423
no_young_children	1.943165
no_older_children	0.952310

On the whole the data is mostly right skewed, with Salary being the most right skewed followed by no_young_children, while the educ variable is slightly left skeweed.

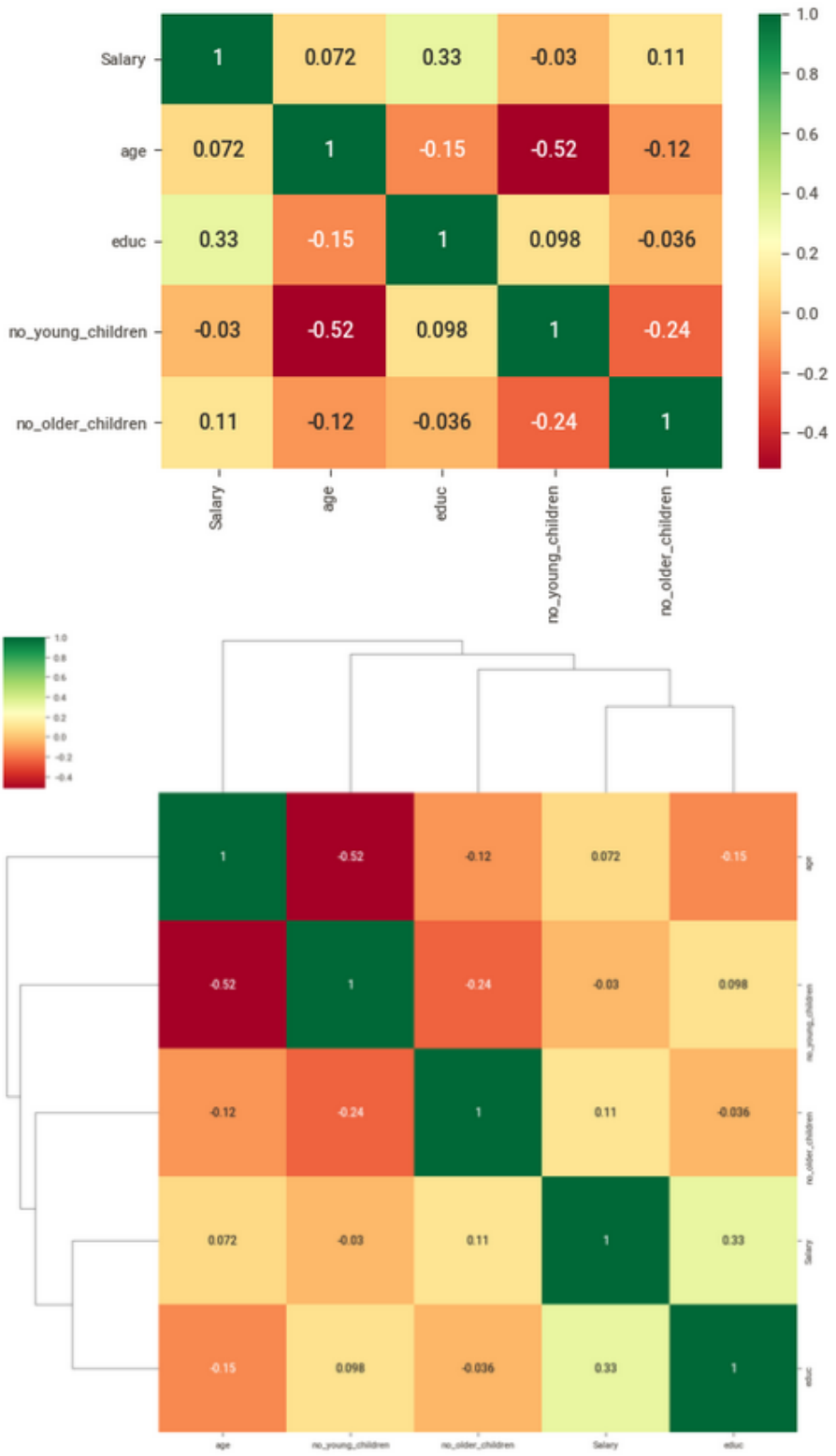
- ***Pair Plot***



Looking at the above pair plot, we can see no clear correlations are observed amongst any of the columns in the data set.

Correlations between the columns will be further looked at in detail when we check the heat map and cluster maps.

- ***Heatmap / Clustermap***



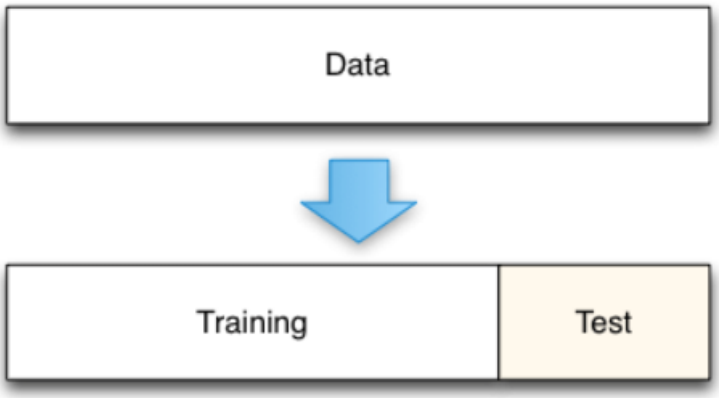
No clear correlations are present in the data set, as is evident from the above shown heat map and cluster maps. Only significant correlation is the negative correlation between age and no_young_children, i.e. as age increases no_young_children decreases which makes logical sense.

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).

Data was **not scaled** in accordance with the project notes.

There are two categorical columns in the data set namely Holliday_Package and foreign, both of them have binary data in the form of yes and no. We encoded these values to 0 & 1 using Sklearn's **LabelEncoder**.



Post this we split the data into test and training set using sklearn's train_test_split method. We could see the yes/no ratio was 46:54% for the dependent variable i.e. Holliday_Package.

Even though the data set was pretty evenly split around the target variable classes, still we used the **stratify** argument while splitting our data set. Using stratify enabled us to ensure the split of yes/no is the same in the training as well as testing set , the way it was in original data set.

Not using stratify could lead to situation where either the train or the test set might become biased towards one of the dependent variable classes - yes/no in this case.

First the logistic regression model was applied on the data set, which yielded an accuracy of approximately 67% for both training as well as testing set.

We performed multiple **GridSearchCV** iterations for parameters like 'penalty', 'solver', 'tol' & 'max_iter'. Cross validation parameter was set to 3 during all the iteration that we undertook.

Below is the best grid of parameters that we obtained after multiple iterations.

```
grid_search.best_params_
```

```
{'max_iter': 325, 'penalty': 'l1', 'solver': 'liblinear', 'tol': 1e-05}
```

```
best_grid = grid_search.best_estimator_
best_grid
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=325,
                   multi_class='auto', n_jobs=None, penalty='l1',
                   random_state=None, solver='liblinear', tol=1e-05, verbose=0,
                   warm_start=False)
```

Then we proceeded ahead to apply the Linear Discriminant Analysis (LDA) on the same data set.The accuracy obtained using LDA was approximately 65% for test as well as training set.

We performed **GridSearchCV** for solver and tol parameters. Cross validation parameter was set to 3.

Below is the best grid of parameters that we obtained.

```
grid_search.best_params_
```

```
{'solver': 'lsqr', 'tol': 0.0001}
```

```
grid_search.fit(X_train, train_labels)
```

```
GridSearchCV(cv=3, error_score=nan,
             estimator=LinearDiscriminantAnalysis(n_components=None,
                                                  priors=None, shrinkage=None,
                                                  solver='svd',
                                                  store_covariance=False,
                                                  tol=0.0001),
             iid='deprecated', n_jobs=None,
             param_grid={'solver': ['lsqr', 'eigen'],
                         'tol': [0.0001, 0.0002, 0.0003]},
             pre_dispatch='2*n_jobs', refit=True, return_train_score=False,
             scoring=None, verbose=0)
```

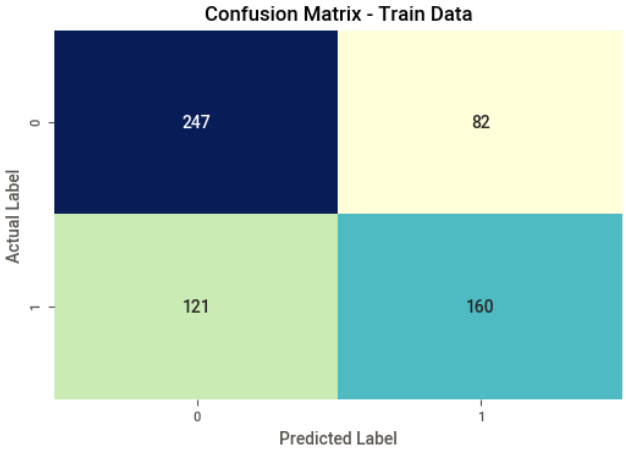
2.3

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.

Performance metrics for Logistic Regression Model

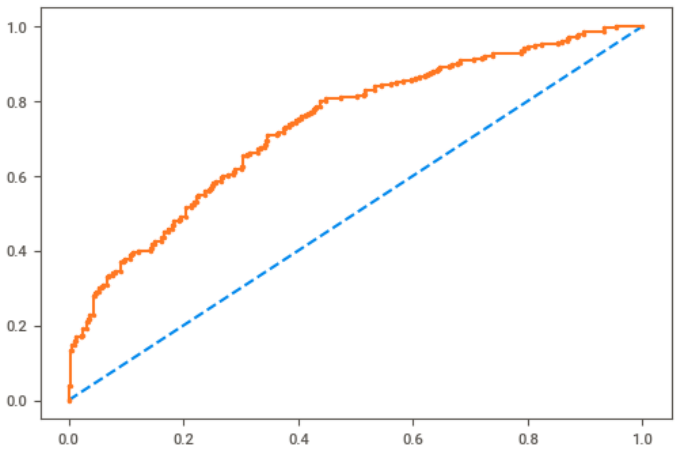
Train Data

Accuracy Score (Train Data) - 66.72%



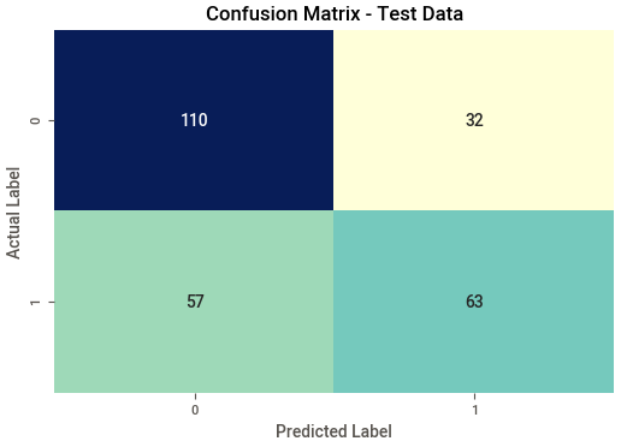
	precision	recall	f1-score	support
0	0.67	0.75	0.71	329
1	0.66	0.57	0.61	281
accuracy			0.67	610
macro avg	0.67	0.66	0.66	610
weighted avg	0.67	0.67	0.66	610

AUC: 0.735



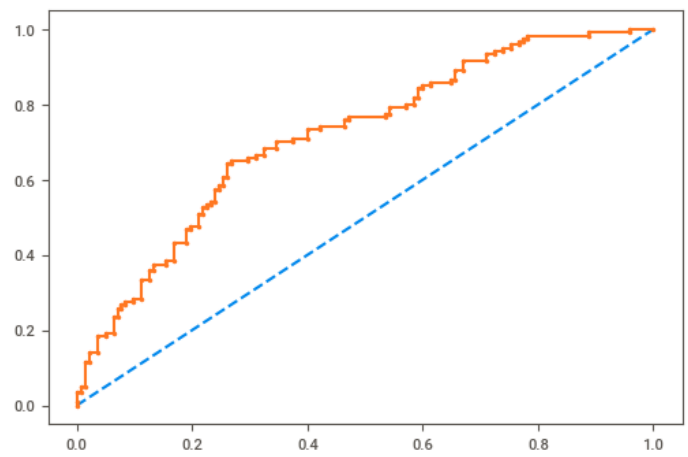
Test Data

Accuracy Score (Test Data) - 66.03%



	precision	recall	f1-score	support
0	0.66	0.77	0.71	142
1	0.66	0.53	0.59	120
accuracy			0.66	262
macro avg	0.66	0.65	0.65	262
weighted avg	0.66	0.66	0.65	262

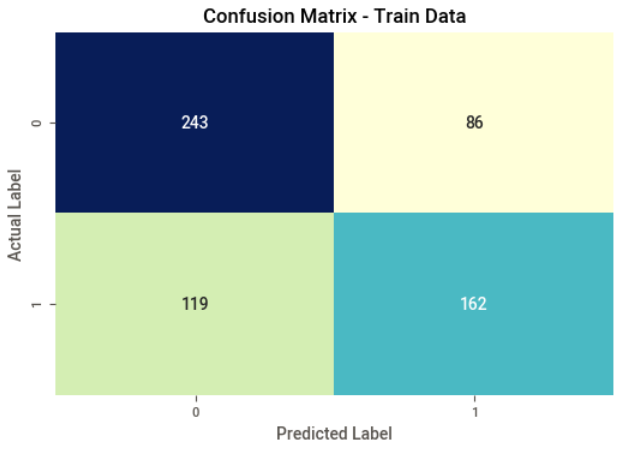
AUC: 0.718



Performance metrics for Linear Discriminant Analysis Model

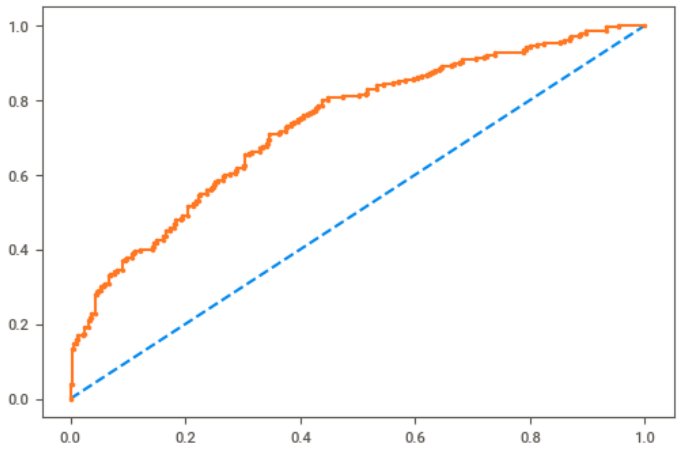
Train Data

Accuracy Score (Train Data) - 66.39%



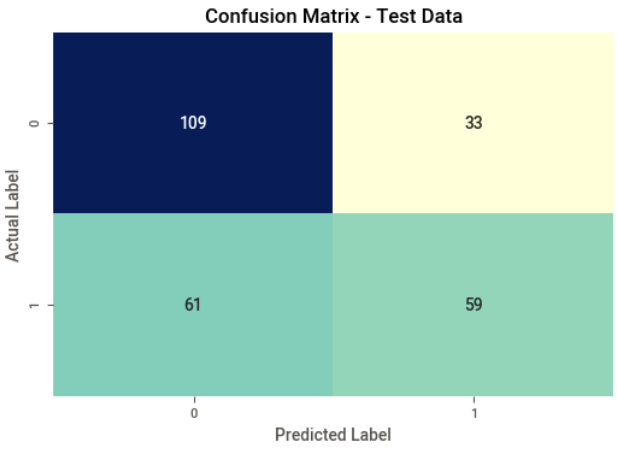
	precision	recall	f1-score	support
0	0.67	0.74	0.70	329
1	0.65	0.58	0.61	281
accuracy			0.66	610
macro avg	0.66	0.66	0.66	610
weighted avg	0.66	0.66	0.66	610

AUC: 0.735



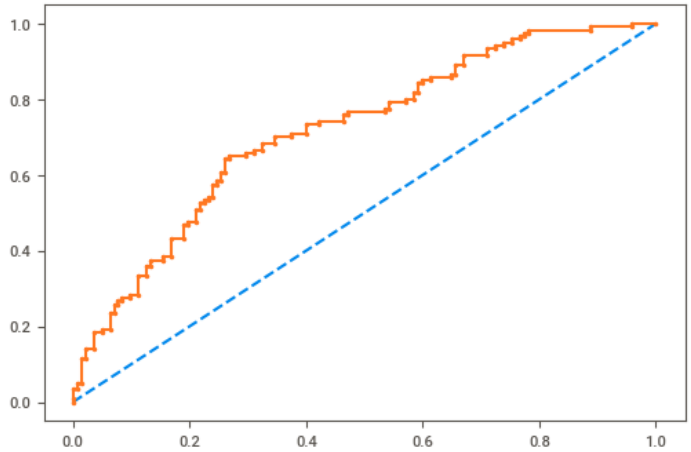
Test Data

Accuracy Score (Test Data) - 64.12%



	precision	recall	f1-score	support
0	0.64	0.77	0.70	142
1	0.64	0.49	0.56	120
accuracy			0.64	262
macro avg	0.64	0.63	0.63	262
weighted avg	0.64	0.64	0.63	262

AUC: 0.718

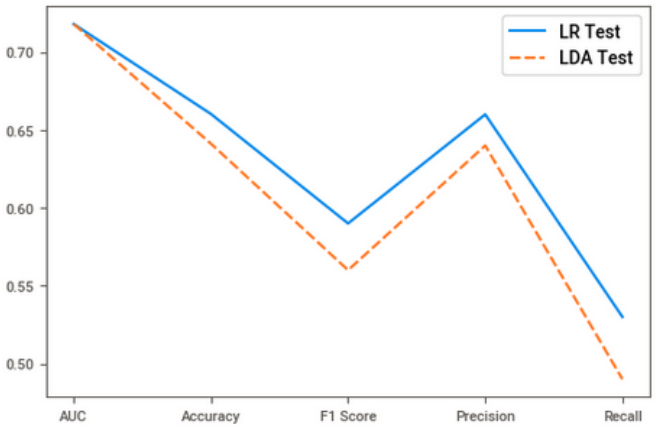


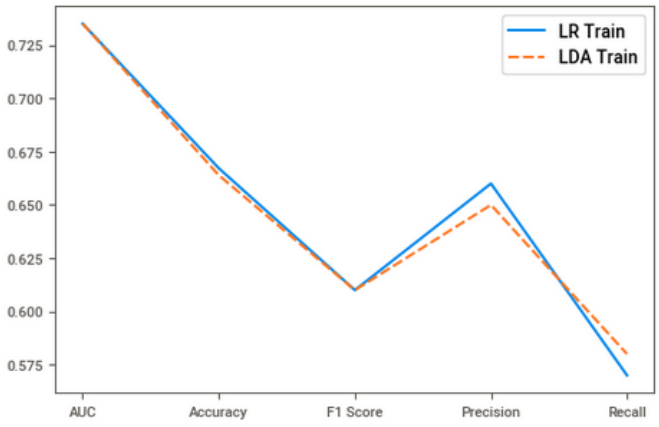
We also tried using custom threshold for classification split on both Logistic regression as well as LDA, the best split was found at 0.4 for both LR and LDA, however the accuracy was still better using the models default threshold values, hence we decided to use the same. Refer notebook for details.

We built two supervised machine learning models namely logistic regression and linear discriminant analysis. Various performance metrics were captured and compared below.

Taking a collective look at the performance metrics for these models before making a final selection of the model to be used.

	LR Test	LR Train	LDA Test	LDA Train
Accuracy	0.66	0.67	0.64	0.66
AUC	0.72	0.74	0.72	0.74
Recall	0.53	0.57	0.49	0.58
Precision	0.66	0.66	0.64	0.65
F1 Score	0.59	0.61	0.56	0.61





For both train as well as test set, we can see the Logistic regression model as out performed LDA on almost every front.

Hence we can say logistic regression is more optimized than LDA in this case and we finalize logistic regression as our model of choice.

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

The models built using the given data are not very accurate hence the first and foremost suggestion for the travel company would be to provide more data so an comprehensive analysis can be undertaken. However since the suggestions are sought from the given data, we will provide the insights which we were able to gather from this small set of data.

Given the predictions based on employee detail the travel company can better position their products and promotions to specifically target the individuals who are are more likely to opt for the holiday package.

Looking at the above analysis the travel company can be given below recommendations:-

1. One of the factors that has a major influence on whether or not an employee will opt for the holiday package is the nationality. Since our analysis had revealed that foreigner employees are much more likely to opt for the holiday package, the company should put more efforts is selling the package to such employees.
2. Concentrate on employees having less number of young children, as we have seen lesser the number of young children, higher is the likelihood to opt for the holiday package.
3. Education is also another factor that can be looked at, we can see employees having less than 8 years and more than 12 years of formal education are more likely to opt for the holiday package.
4. Employees having 1 or more older children are more likely to opt for the package compared to the rest. Hence they can be provided with better offers.
5. Employees having an average salary of less than 1,00,000 are more likely to opt for the holiday package, hence they should be focused upon by the travel company.

These are some of the employees on which the company can focus and provide targeted promotions.