REPORT - PREDICTIVE MODELING

By: Tanushri Das

Table of Contents:

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.

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.

Ans)

We have been provided with details of cubic zirconia manufactured by Gem Stones co ltd. The various attributes taken into account have been listed below with their description.

Sr.No	Variable Name	Description
1	Carat	Carat weight of the cubic zirconia.
2	Cut	Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
3	Color	Color of the cubic zirconia. With D being the best and J the worst.
4	Clarity	Cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, FL = flawless, I3= level 3 inclusions) FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3
5	Depth	The Height of a cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.
6	Table	The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
7	Z	Height of the cubic zirconia in mm.
8	X	Length of the cubic zirconia in mm.
9	Y	Width of the cubic zirconia in mm.
10	**Price (Target)**	**The Price of the cubic zirconia.**

Table 1: Data Description

The purpose is to build a model which can predict the price of the cubic zirconia gemstone given the details of independent variables.

EDA:

The dataset provided has 10 attributes and 26967 observations. The basic information of the dataset is as follows:

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

Table 2: First five rows of dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 10 columns):
 #   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
dtypes: float64(6), int64(1), object(3)
memory usage: 2.1+ MB
```

Figure 1: Information on dataset

Cut, color and clarity are categorical variables, which are ordinal in nature as described in Table 1 and the rest are numerical. The five point summary in tabular form is as follows:

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

Table 3: Five point summary

The mean and median values are pretty close to each other which may mean the columns are normally distributed, but could be observed further using their distribution. We see that the minimum values of x, y and z are 0. The rows having such values are given below:

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

Table 4: Data having 0 dimensional values

These are practically impossible for a 3D object, so we might need to remove these rows during data preprocessing.

By looking at the number of unique values and their value counts (line 9), we see that there are no incorrect values present. Null values are only present in the depth column.
The distribution of depth is as follows:

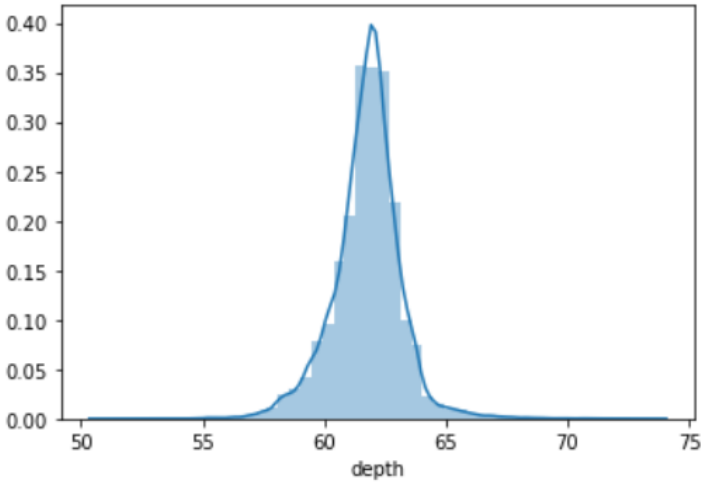

Figure 2: Distribution of Depth column

Since it is normally distributed, we can impute the missing values using the mean of the column.

There are 33 duplicate values in the dataset. Since it is highly unlikely that two gemstones have exactly same values of all variables, we will consider them as genuine duplicates and remove them while processing data.

Univariate Analysis:

After imputing the null values and removing the duplicates, we proceed with the Univariate analysis. We are now left with 26925 observations.

Looking at the boxplot to check if outliers are present in the data or not.

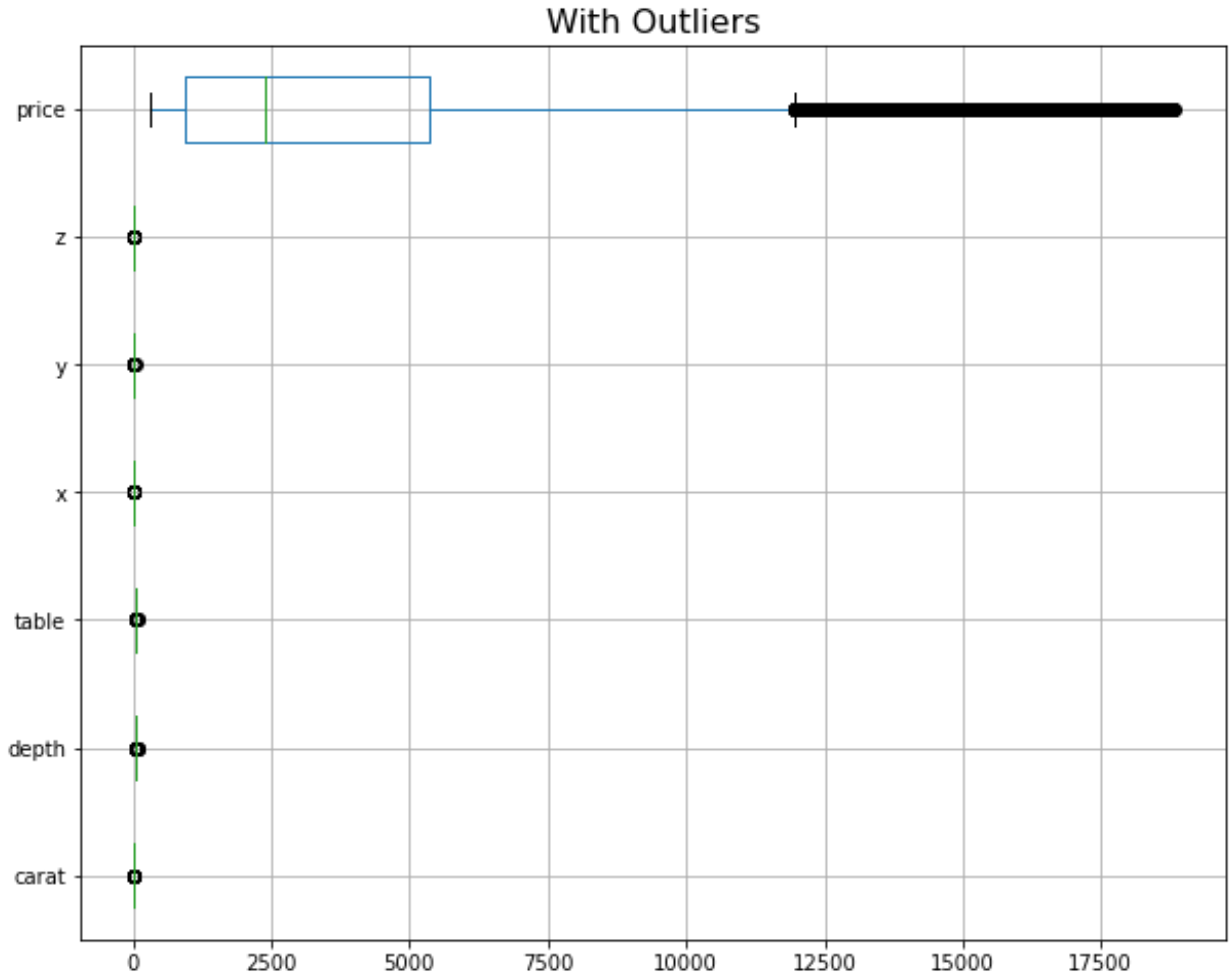

Figure 3: Boxplot with outliers

We can clearly see that data is at different scales and units. Price is an amount and others, carat is weight and others are dimensional measurements. They all contain outliers within them. To have a closer look and visualize outliers in others, we remove the price and again plot the boxplot.

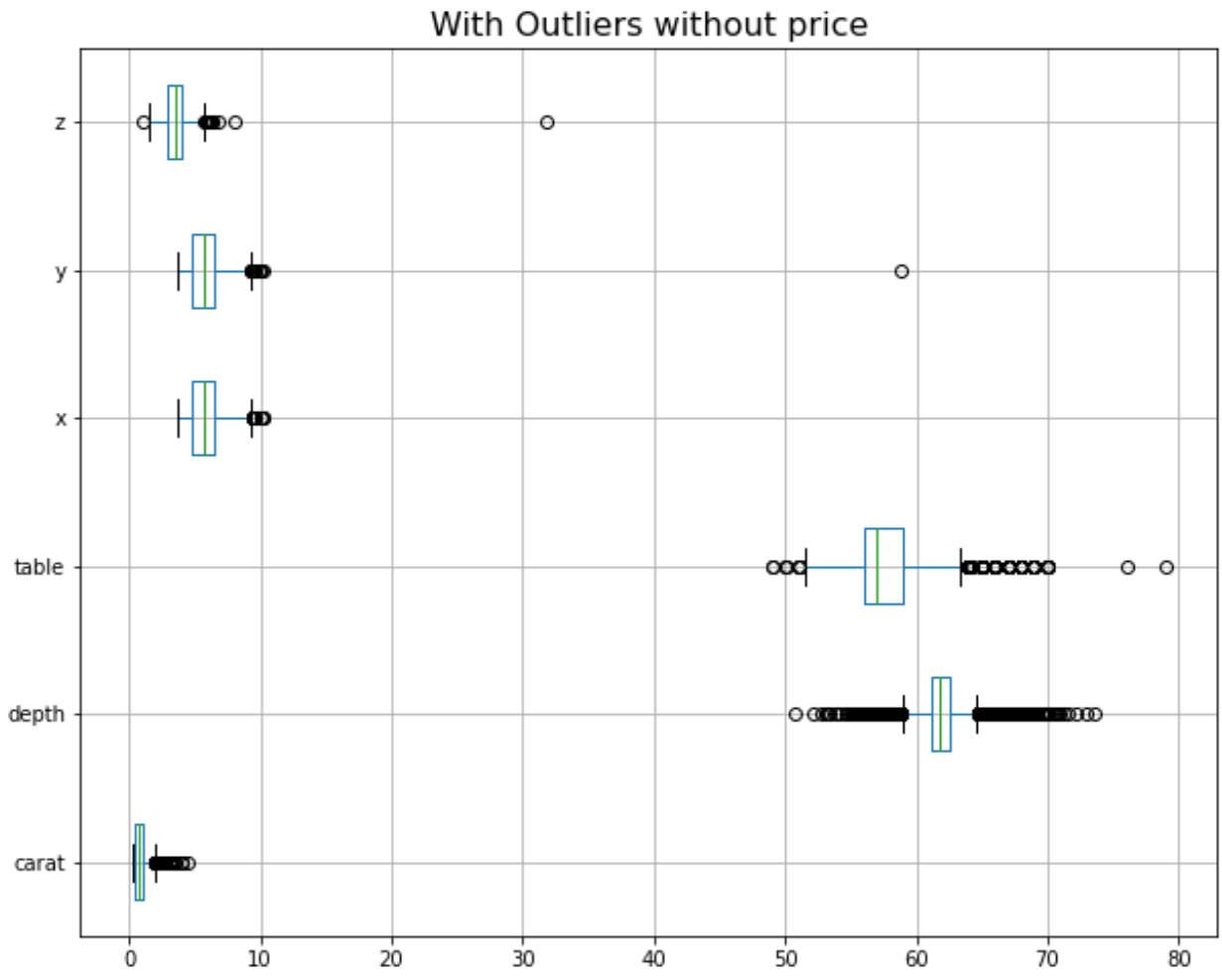

Figure 4: Boxplot with outliers without price

Each column contains certain number of outliers. Outliers in regression are observations that fall far from the "cloud" of points. These points are especially important because they can have a strong influence on the least squares line. Points that fall horizontally away from the center of the cloud tend to pull harder on the line, so we call them points with high leverage.

If your data set includes an influential point, here are some things to consider.

- An influential point may represent bad data, possibly the result of measurement error. If possible, check the validity of the data point.
- Compare the decisions that would be made based on regression equations defined with and without the influential point. If the equations lead to contrary decisions, use caution.

Though several models are robust enough to handle the outliers, it is better to be treated before modeling, so that the best fit line is gravitated towards the extreme points.

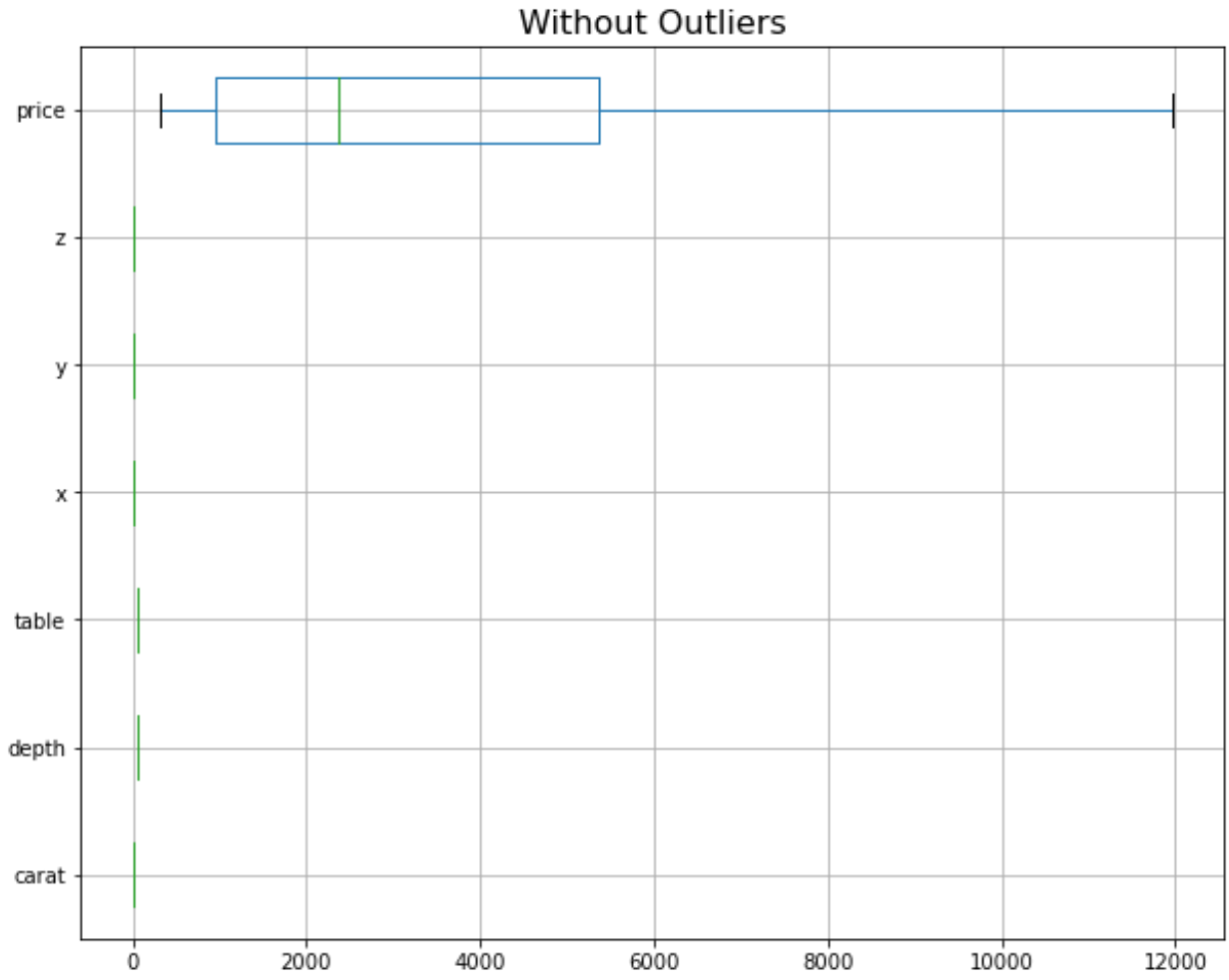

Figure 5: Boxplot without outliers

Bivariate Analysis:

Checking the scatter plots using pairplot for data with outliers:

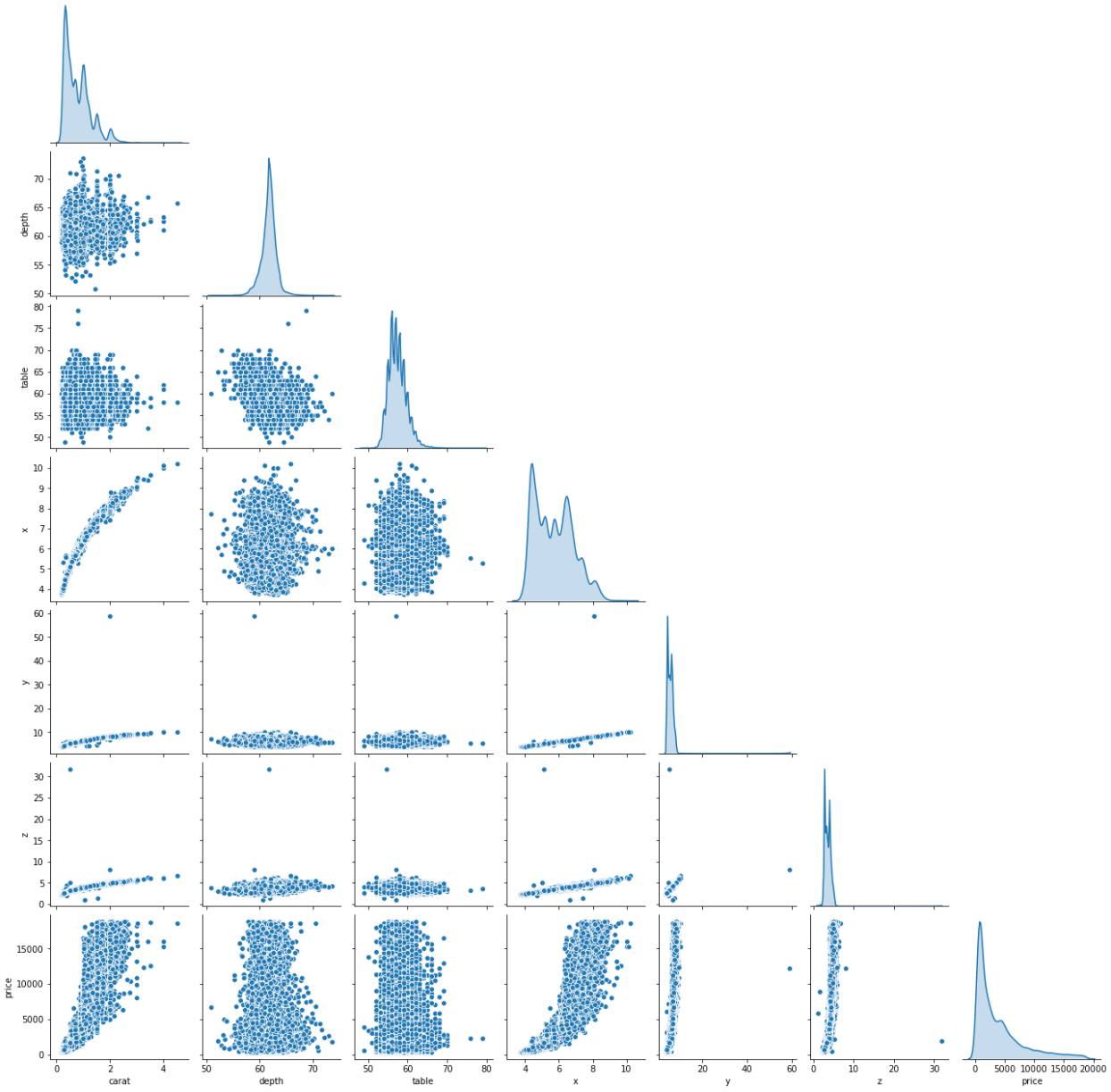

Figure 6: Pairplot with outliers

A linear relationship of the dimensions x, y and z can be seen with other variables. Also outliers are clearly visible in the scatter plots too. We would look at pairplot of dataset without outliers for further examination.

Checking the scatter plots using pairplot for data without outliers:

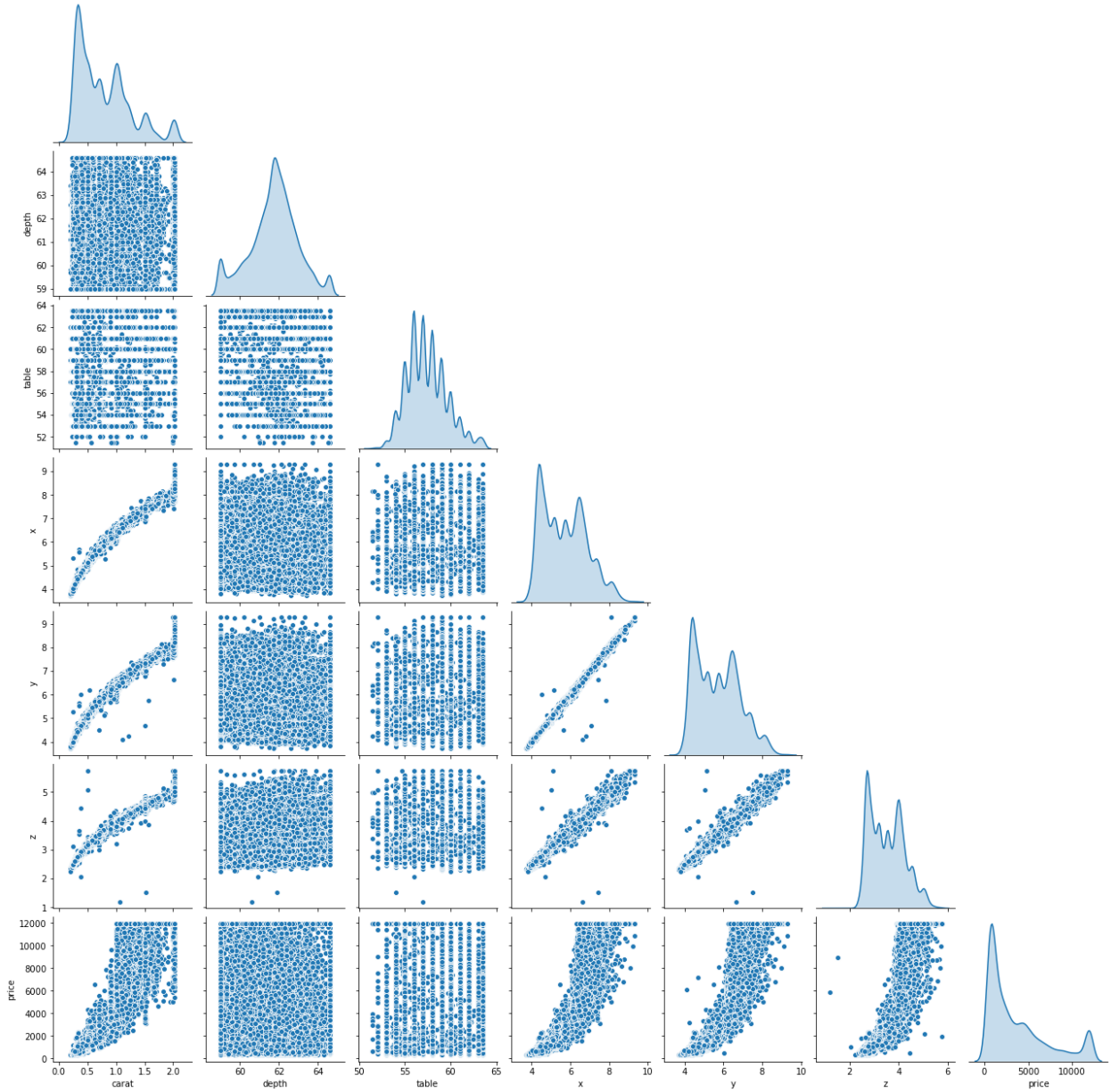

Figure 7: Pairplot without outliers

Here the linear relationship among the independent variables and between independent and dependent variables is more prominent. There is a positive linear relationship of price with carat and dimensions of the cubic zirconia stone.

Checking the correlation using heatmap for data with outliers:

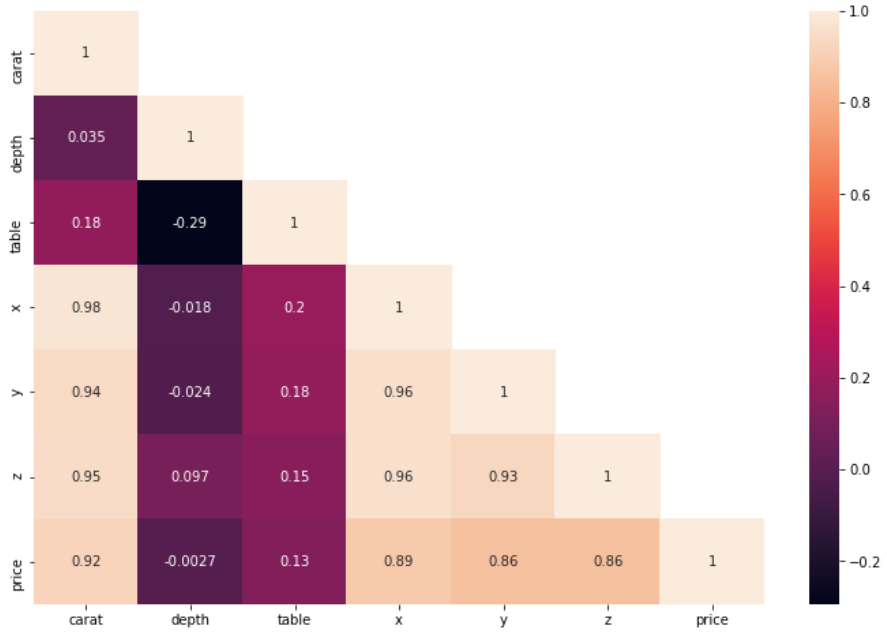

Figure 8: Correlation plot with outliers

Checking the correlation using heatmap for data without outliers:

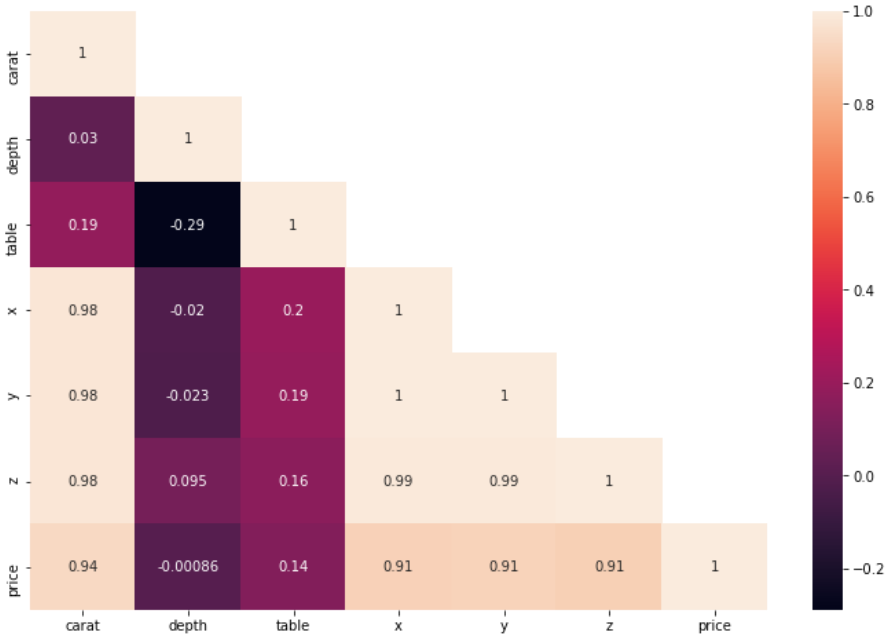

Figure 9: Correlation plot without outliers

The positive correlation observed from the pairplots is further confirmed with high correlation coefficients of price with carat and x, y and z dimensions.

The following EDA plots were used to derive further insights from the data. They would be explained in detail in the last question. Briefly going through them here.
Following is a point plot of cut, color and clarity with the target price. We can see which cut, color and clarity were priced high.

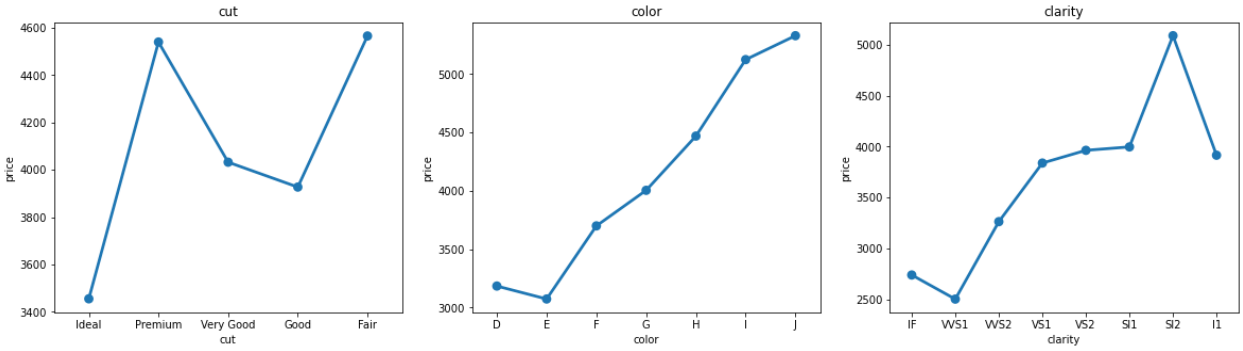

Figure 10: Pointplot of categorical variables with price

The following jitter plots show the frequency of the gemstone in the particular category in a price range.

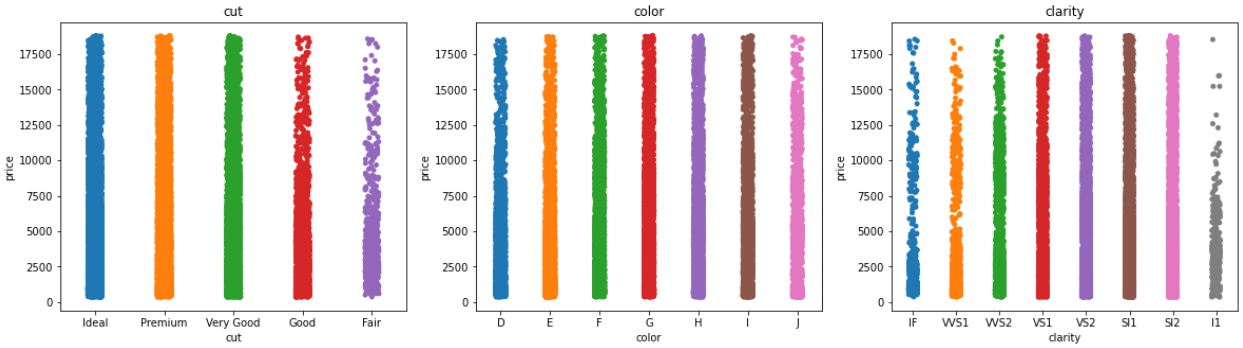

Figure 11: Stripplot of categorical variables with price

Following is the plot of number of gemstones in each subcategory of categorical variables

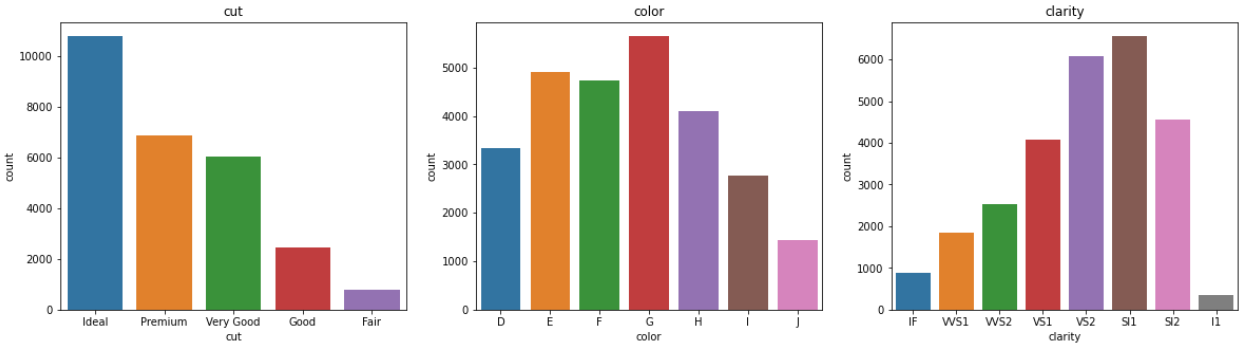

Figure 12: Countplot of categorical variables

Following is a plot of categorical variables with price where the x variable is color, each column is the clarity level and shows distribution for various cuts.

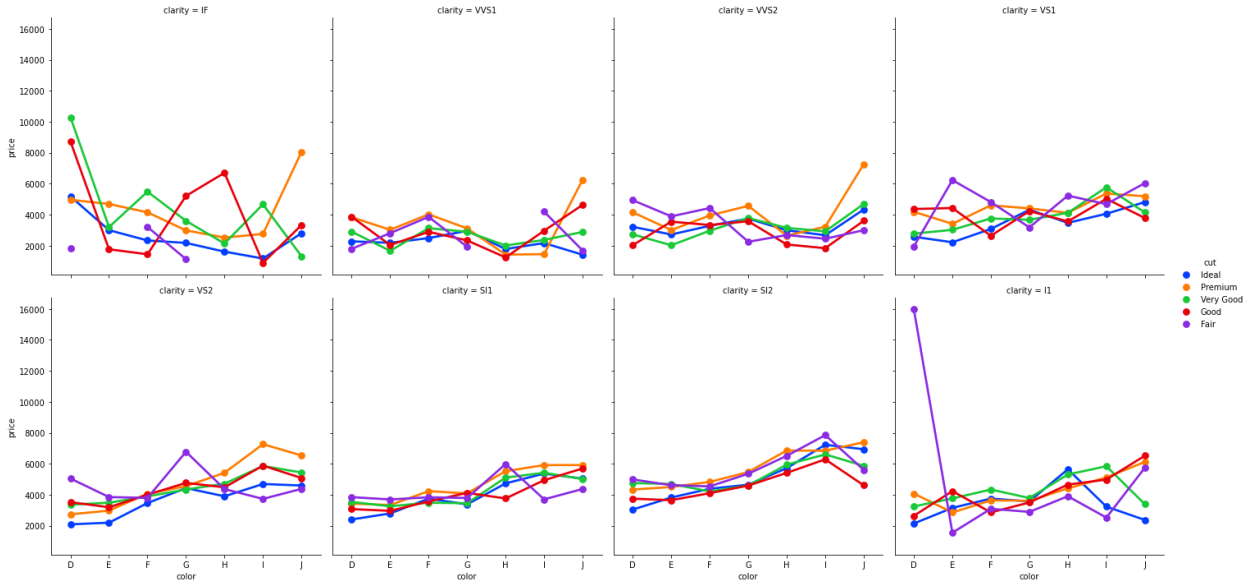

Figure 13: Factorplot of categorical variables

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?

Ans)

As seen from Table 4, there are some observations which have dimensions as 0. Even if one dimension becomes zero, it becomes a 2D object which is practically absurd. So these observations don't make any sense. There are only 9 such observations, hence we can drop these rows for correct modeling purposes. We imputed the null values with mean values as explained in the previous question.

Linear regression uses the ordinary least squares method (OLS) to form model which uses gradient descent as an optimization technique. Take a look at the formula for gradient descent below:

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

The presence of feature value X in the formula will affect the step size of the gradient descent. The difference in ranges of features will cause different step sizes for each feature. To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model.

Having features on a similar scale can help the gradient descent converge more quickly towards the minima. We can speed up ***gradient descent*** *by* ***scaling****. This is because θ will descend quickly on small*

ranges and slowly on large ranges, and so will oscillate inefficiently down to the optimum when the variables are very uneven.

The point of normalization is to change your observations so that they can be described as a normal distribution. A much better multivariate regression performance is achieved from the Gradient Descent algorithm if mean normalization is applied to the training set. And since the dataset is at different scales and units; like price and dimension, it will be good to apply scaling in this case.

Although point to be noted that scaling does not change RMSE, R-squared value, Adjusted R-squared value, p-value of coefficients. In regression analysis, we can calculate **importance of variables** by ranking independent variables based on the descending order of absolute value of standardized coefficient.
Scaling can be interpreted as a means of giving the same importance to each feature. Also scaling results in bringing the intercepts to 0. Intercept is the initial value which is predicted if there are no input variables provided. Sometimes this value turns out to be not meaningful hence scaling is employed which centers the data bringing the intercept close to 0.

1.3 Encode the data (having string values) for Modeling. 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.
Ans)

Manual label encoding was performed on the categorical data as they are ordinal in nature and their order was business specific. The best ones were given the highest scores and worst ones the least scores. They can be viewed in line 31 of notebook.

2 models were built, one with scaled data and one with scaled data without outliers to compare their performances. Using the linear regression model the RMSE and accuracy of the model was found as follows:

	Train RMSE	Test RMSE	Training Score	Test Score	Intercept
LR Scaled	1215.798402	1217.784216	0.907986	0.909617	-2.319767e-16
LR Scaled No Outliers	1195.589058	1194.259958	0.931228	0.931626	-6.543022e-16

Table 5: RMSE and Accuracy of models

From RMSE we can infer that the actual value of target variable (in this case price), is likely to be +/- the RMSE value. So lesser the value, the better. We can see from the results that the model without outliers has performed very well in terms of accuracy and RMSE score. Scaling results in bringing the intercept close to 0 which can be seen here.

To obtain a detailed summary of data statsmodels module was used to create a summary of model.

Scaled data

OLS Regression Results

Dep. Variable:	price	R-squared:	0.908
Model:	OLS	Adj. R-squared:	0.908
Method:	Least Squares	F-statistic:	2.065e+04
Date:	Thu, 27 Aug 2020	Prob (F-statistic):	0.00
Time:	12:32:05	Log-Likelihood:	-4260.0
No. Observations:	18847	AIC:	8540.
Df Residuals:	18837	BIC:	8618.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.988e-17	0.002	-9e-15	1.000	-0.004	0.004
carat	1.3100	0.011	118.409	0.000	1.288	1.332
cut	0.0298	0.003	11.025	0.000	0.025	0.035
color	0.1401	0.002	59.855	0.000	0.136	0.145
clarity	0.2069	0.002	84.432	0.000	0.202	0.212
depth	-0.0294	0.003	-10.734	0.000	-0.035	-0.024
table	-0.0198	0.003	-7.103	0.000	-0.025	-0.014
x	-0.2664	0.014	-18.716	0.000	-0.294	-0.239
y	0.0020	0.007	0.279	0.780	-0.012	0.016
z	-0.0076	0.008	-1.011	0.312	-0.022	0.007

Omnibus:	4196.790	Durbin-Watson:	1.993
Prob(Omnibus):	0.000	Jarque-Bera (JB):	205176.193
Skew:	-0.059	Prob(JB):	0.00
Kurtosis:	19.164	Cond. No.	15.9

Figure 14: Summary of scaled data

Scaled data without outliers

OLS Regression Results

Dep. Variable:	price	R-squared:	0.931
Model:	OLS	Adj. R-squared:	0.931
Method:	Least Squares	F-statistic:	2.834e+04
Date:	Thu, 27 Aug 2020	Prob (F-statistic):	0.00
Time:	12:32:12	Log-Likelihood:	-1516.4
No. Observations:	18847	AIC:	3053.
Df Residuals:	18837	BIC:	3131.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.504e-18	0.002	-7.87e-16	1.000	-0.004	0.004
carat	1.1818	0.011	107.645	0.000	1.160	1.203
cut	0.0364	0.002	15.494	0.000	0.032	0.041
color	0.1346	0.002	66.557	0.000	0.131	0.139
clarity	0.2083	0.002	97.723	0.000	0.204	0.212
depth	0.0124	0.004	3.181	0.001	0.005	0.020
table	-0.0094	0.002	-3.854	0.000	-0.014	-0.005
x	-0.4374	0.044	-9.966	0.000	-0.523	-0.351
y	0.5027	0.043	11.734	0.000	0.419	0.587
z	-0.1941	0.028	-6.967	0.000	-0.249	-0.140

Omnibus:	2652.211	Durbin-Watson:	2.004
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9565.281
Skew:	0.690	Prob(JB):	0.00
Kurtosis:	6.206	Cond. No.	62.9

Figure 15: Summary of scaled data without outliers

From the summary of scaled data we observe that p values for y and z were higher than our confidence level of 0.05, which indicate that they are not essential and could be dropped. Also owing to high correlation between the dimensional variables, a new column called volume is created which is the multiplication of these variables and the original columns are dropped. The summary of this model is as follows:

Model with volume column and removing x, y and z dimensions

OLS Regression Results

Dep. Variable:	price	R-squared:	0.932
Model:	OLS	Adj. R-squared:	0.932
Method:	Least Squares	F-statistic:	1.590e+04
Date:	Thu, 27 Aug 2020	Prob (F-statistic):	0.00
Time:	19:03:24	Log-Likelihood:	-643.40
No. Observations:	8078	AIC:	1303.
Df Residuals:	8070	BIC:	1359.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0113	0.003	3.761	0.000	0.005	0.017
carat	1.1173	0.006	197.832	0.000	1.106	1.128
cut	0.0322	0.004	9.031	0.000	0.025	0.039
color	0.1343	0.003	43.716	0.000	0.128	0.140
clarity	0.2101	0.003	65.452	0.000	0.204	0.216
depth	-0.0106	0.003	-3.139	0.002	-0.017	-0.004
table	-0.0102	0.004	-2.775	0.006	-0.017	-0.003
Volume	-0.0277	0.002	-14.368	0.000	-0.031	-0.024

Omnibus:	815.422	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3378.859
Skew:	0.434	Prob(JB):	0.00
Kurtosis:	6.047	Cond. No.	6.26

Figure 16: Summary of model with volume column and removing x, y and z dimensions

The meaning of few of the criteria used for analysis is as follows:

- **P>|t|:** P-value that the null-hypothesis that the coefficient = 0 is true. If it is less than the confidence level, often 0.05, it indicates that there is a statistically significant relationship between the term and the response. Here all variables are significant.
- **R-squared:** The coefficient of determination. A statistical measure of how well the regression line approximates the real data points
- **Adj. R-squared:** It reflects the fit of the model. R-squared values range from 0 to 1, where a higher value generally indicates a better fit, assuming certain conditions are met. The higher the value, the better the explanability of the model. The Adjusted R-Squared value is always a bit lower than the Multiple R-Squared value because it reflects model complexity (the number of variables) as it relates to the data, and consequently is a more accurate measure of model performance.
- **F-Stat/ Prob (F-Statistic)**: It is a statistical test that compares the fit of the intercept-only model with your model. In simple works, if P value for the F-Stat (here **Prob (F-Statistic)**) is less than your significance level, one can reject the null hypothesis that a intercept-only model is better, meaning your model is better.

- **Correlation Coefficient:** The Pearson correlation coefficient is also an indicator of the extent and strength of the linear relationship between the two variables.
- **Omnibus/Prob(Omnibus)**: A test of the skewness and kurtosis of the residual. **The Prob (Omnibus)** performs a statistical test indicating the probability that the residuals are normally distributed.
- **Skew** – a measure of data symmetry.
- **Kurtosis:** A measure of "peakiness", or curvature of the data. Higher peaks lead to greater Kurtosis. Greater Kurtosis can be interpreted as a tighter clustering of residuals around zero, implying a better model with few outliers.
- **Durbin-Watson:** Tests for homoscedasticity. Ideal value between 1 and 2.
- **Jarque-Bera (JB)/Prob(JB):** Like the Omnibus test, tests both skew and kurtosis.
- **Condition Number** – This test measures the sensitivity of a function's output as compared to its input. When we have multicollinearity, we can expect much higher fluctuations to small changes in the data; hence, we hope to see a relatively small number, something below 30.

The output above shows that, when the other variables remain constant, if we compare two gemstones whose attributes differ by one unit, the applicant with higher attribute value will, on average, have the coefficient times units higher price. Using the P >|t| result, we can infer that the variables p-value is less than 0.05 are significant variables. The lowest condition number was obtained in the third model, as it treated the issue of multicollinearity in data. The AIC value is also least in the data which is preferred in industries. It slightly improved the r-squared and adjusted r-squared values. The value of coefficients indicates the importance of variables.

The price variable can be predicted using the following equation:

Price = (0.01) * Intercept + (1.12) * carat + (0.03) * cut + (0.13) * color + (0.21) * clarity + (-0.01) * depth + (-0.01) * table + (-0.03) * Volume

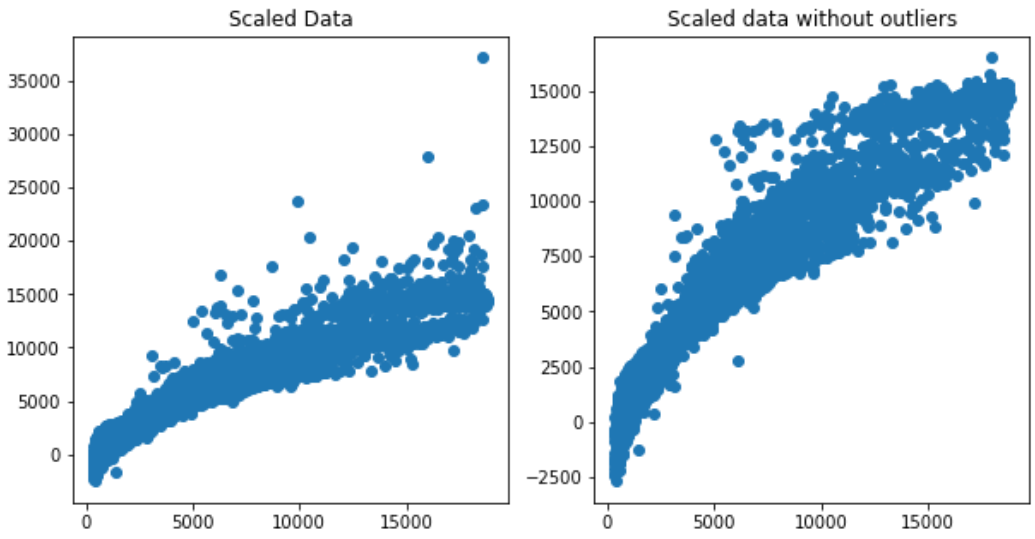

Figure 17: Plot between predicted and actual values

The predicted and actual values are following a direct linear relationship which is expected and desired. The plot is better for data without outliers.

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

Ans)

From question 1.2, we had performed the following EDA on categorical variables.

Following is a point plot of cut, color and clarity with the target price. We can see which cut, color and clarity were priced high.

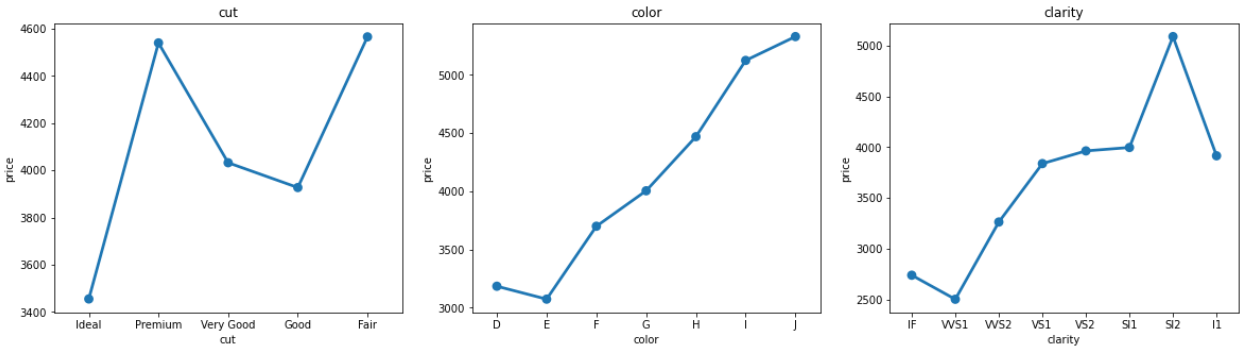

Figure 18: Pointplot of categorical variables with price

The cuts having higher prices are premium ones and surprisingly the fair cut also has higher prices. The colors **I, J** followed by H, G F have higher prices. And the clarity of **S12** was the priciest with **VS1, VS2, S11 and I1** having same price range.

The least priced cut was **ideal**, colors were **E and D**, and clarity level was **VVS1**. Since the company earns profits based on price slots, they would earn higher profits for higher priced gemstones. The company should manufacture more stones with higher price stone characteristics.

The following jitter plots show the frequency of the gemstone in the particular category in a price range.

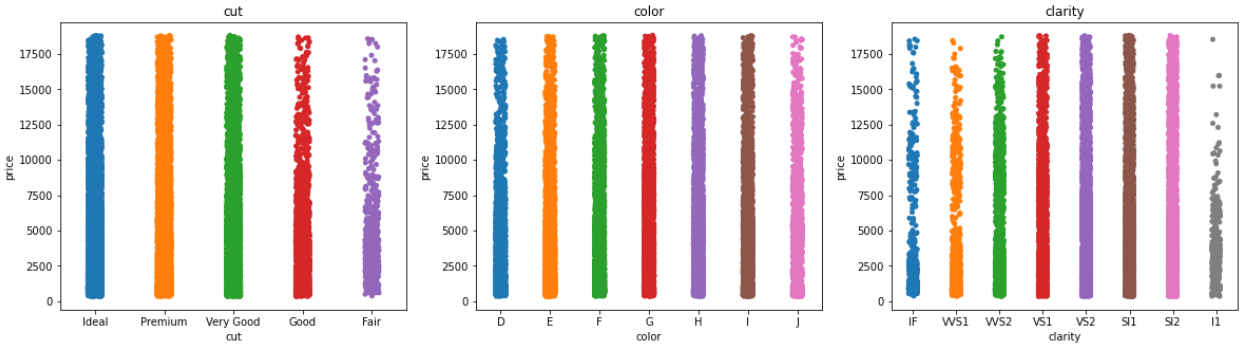

Figure 19: Stripplot of categorical variables with price

Following is the plot of number of gemstones in each subcategory of categorical variables

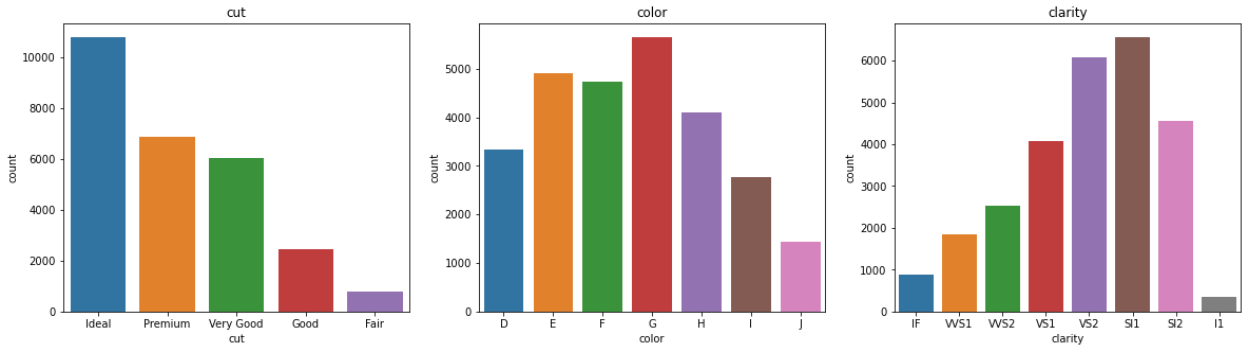

Figure 20: Countplot of categorical variables

The ideal cut stones are highest in number, but they are sold at way cheaper rates, as compared to other cuts. Their higher production would not help company to increase their profits, instead the production capital could be used in manufacturing premium and fair cut diamonds that could yield higher profits. Similar strategy can be applied to color and clarity.

Following is a plot of categorical variables with price where the x variable is color, each column is the clarity level and shows distribution for various cuts.

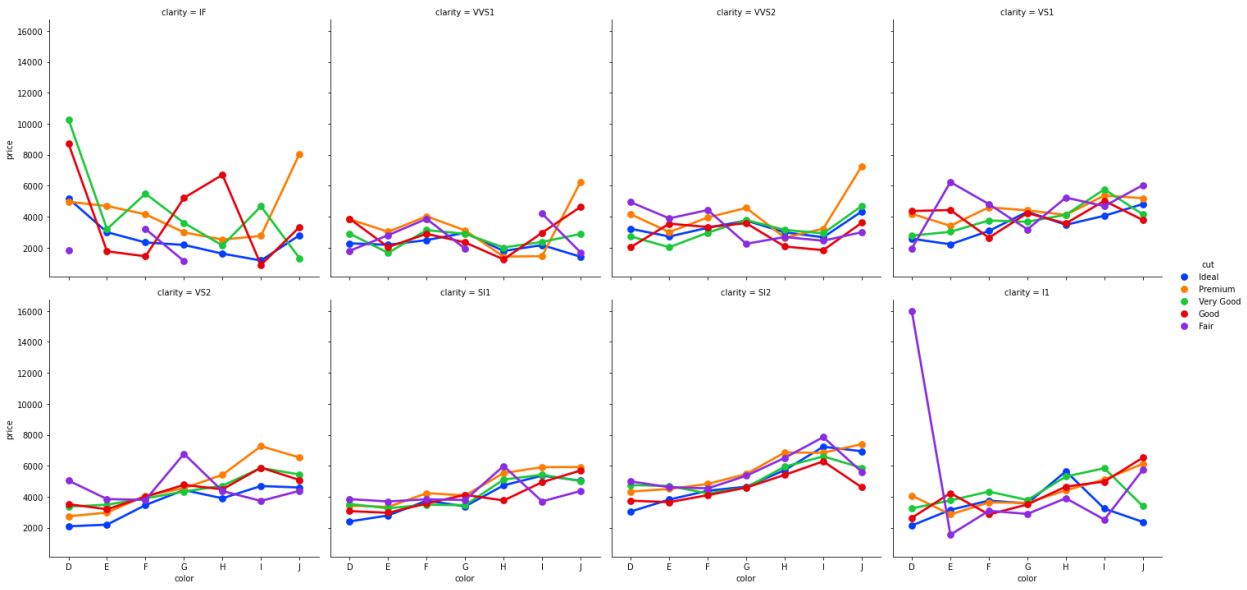

Figure 21: Factorplot of categorical variables with price

Looking at most desirable clarity S12, we can observe that the fair cut stones are priced higher with I and J colors preferred highest in accordance with our above observations. The fair stones are performing well in all cases in general. This observation can be exploited by company since fair cut stones would be cheaper and easier to produce as compared to ideal/premium cut stones. There was one stone with best color and fair cut with highest price. As per linear equation, the most important parameter is **carat**, followed by **clarity**, **color, cut** and **volume**.

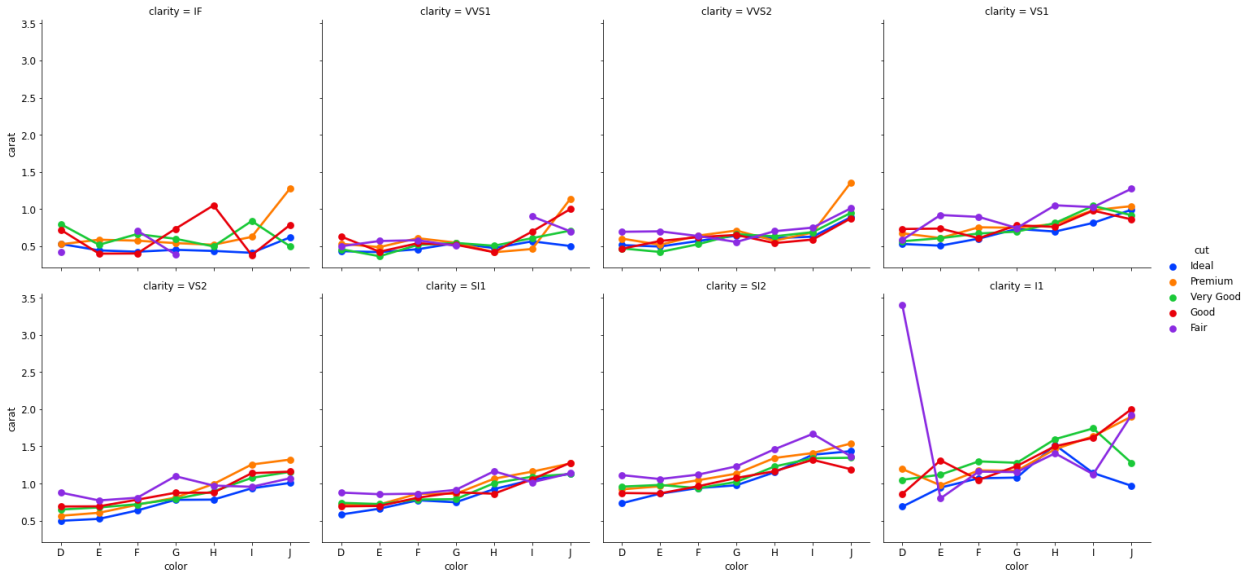

Figure 22: Factorplot of categorical variables with carat

Higher the carat, higher the prices. The IF clarity stones with D color and lower carats still have highest prices, which seems to be a unique exception and business can look into this.

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.

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.

Ans)

We have been provided with employee details of a company. A tours and travel agency needs help in predicting if an employee would opt for a holiday package in future. The predictive model is to be built using given data. The dataset contains details of 872 employees and 7 attributes, its details are given below.

Sr.No	Variable Name	Description
1	Holiday_Package	Opted for Holiday Package yes/no?
2	Salary	Employee salary
3	age	Age in years
4	edu	Years of formal education
5	no_young_children	The number of young children (younger than 7 years)
6	no_older_children	Number of older children
7	foreign	foreigner Yes/No

The data profile is as follows:

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

Table 5: data profile

We check the basic information.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Holliday_Package   872 non-null    object
 1   Salary             872 non-null    int64
 2   age                872 non-null    int64
 3   educ               872 non-null    int64
 4   no_young_children  872 non-null    int64
 5   no_older_children  872 non-null    int64
 6   foreign            872 non-null    object
dtypes: int64(5), object(2)
memory usage: 47.8+ KB
```

Figure 23: Information of data

There are two object variables and others are numerical. They seem to be correct datatypes for respective columns. No null values are present in the dataset.

The five point summary of

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

Table 6: Description of data

More than 75% of people in no_young_children column have no children. And more than 75% of people have below 2 older children. The age and education seem to be normally distributed, will check further through distribution plot. Salary of employees seems to be skewed as mean is greater than median. Let us look at the proportion of target variable. This is a balanced dataset with not much difference in the numbers of target variable. There are 401 (54%) employees who had opted for holiday packages and 471 (46%) who didn't. Also there are no duplicates present in the dataset.

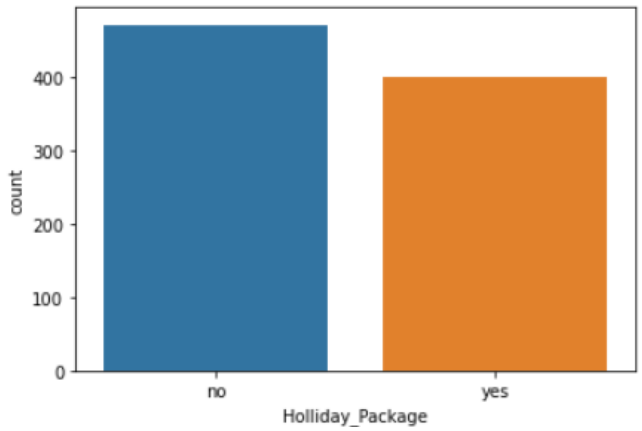

Figure 24: Countplot of target variable

Univariate Analysis:

Let us check for outliers using boxplot.

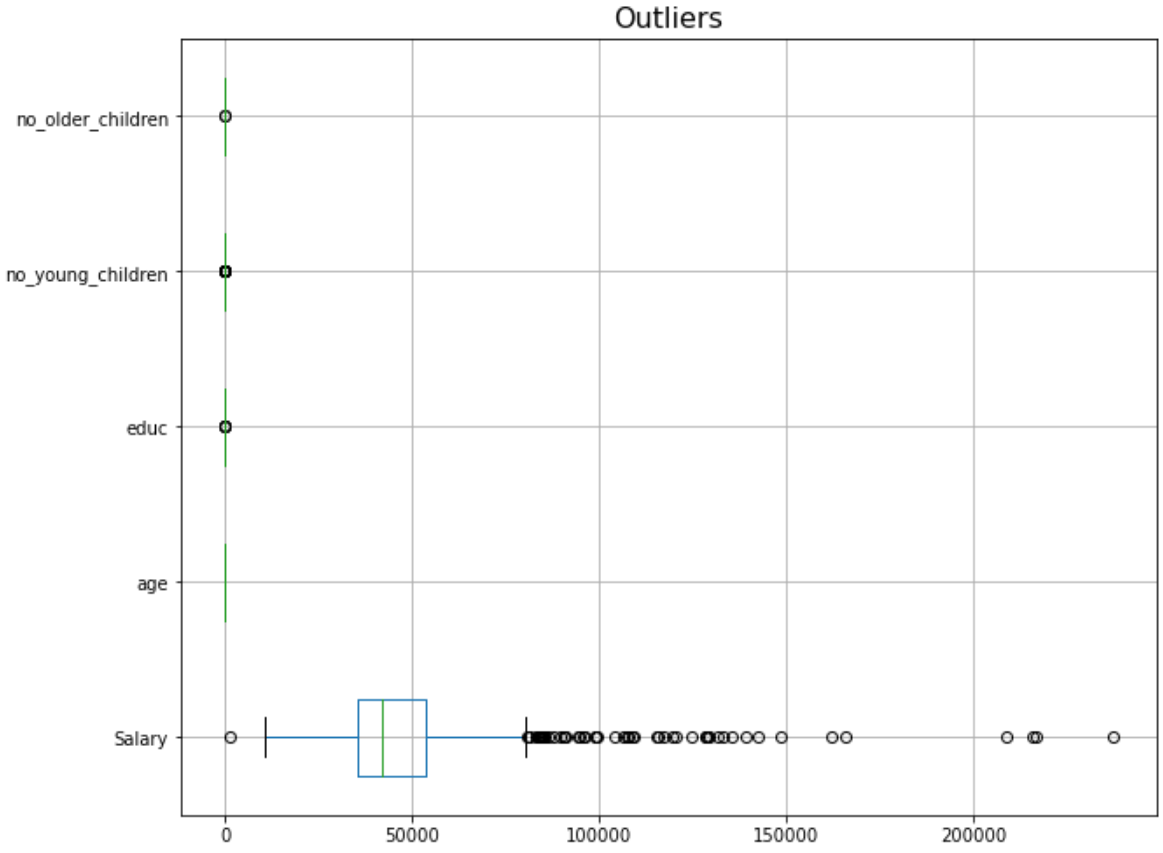

Figure 25: Boxplot with salary

We see that there seem to be lot of outliers in salary column and some in others. Let's have a closer look at the plots without salary.

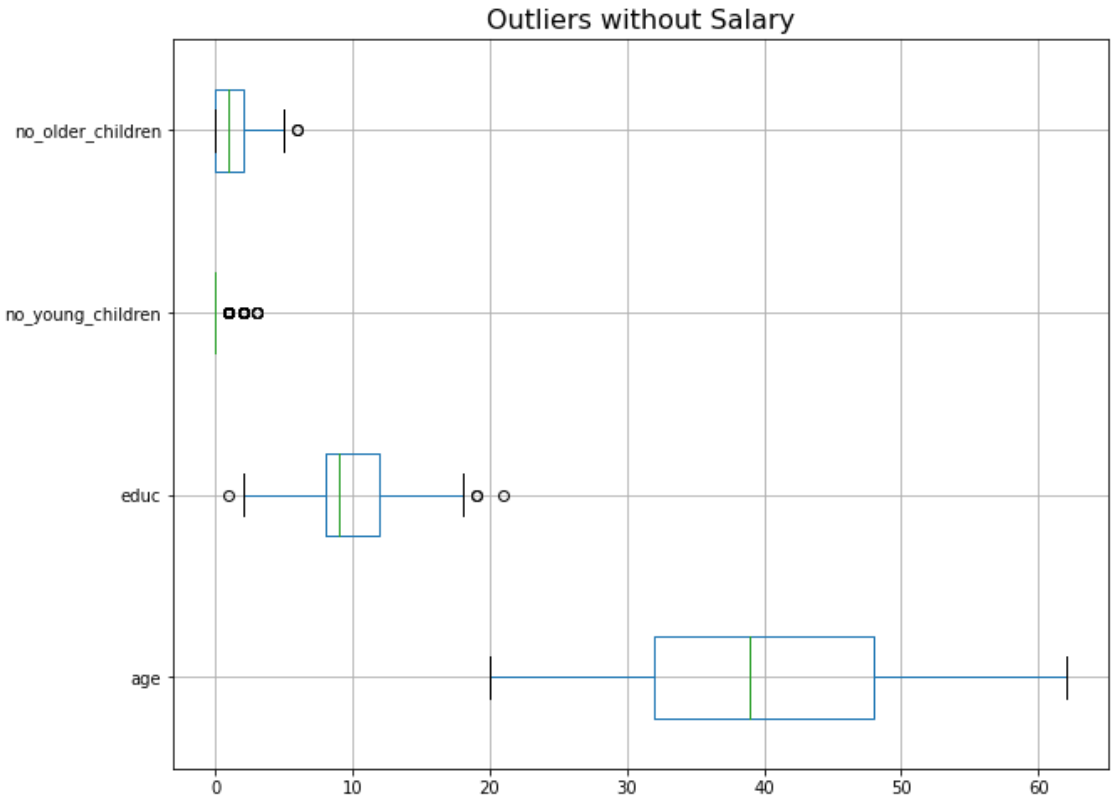

Figure 26: Boxplot without salary

Age has no outliers, other columns like no_young_children, no_older_children and educ are numbers (counts) having 2 to 3 values as outliers, but these are valid data points. Coming to salary, there are 57 (6%) values which are outliers which is not much compared to size of dataset. Hence in this case outlier treatment doesn't seem necessary and we can proceed with model building without it.

Bivariate Analysis:

We can have a look at the correlation among the numerical values and their distribution using the given pairplot. We can infer that

- Employees having lesser salary were more likely to opt for holiday package
- Employees with 1 young children are most likely to opt for holiday package
- Employees with no older children are most likely to opt for holiday package
- Employees with age above 50 are more likely to opt for holiday package

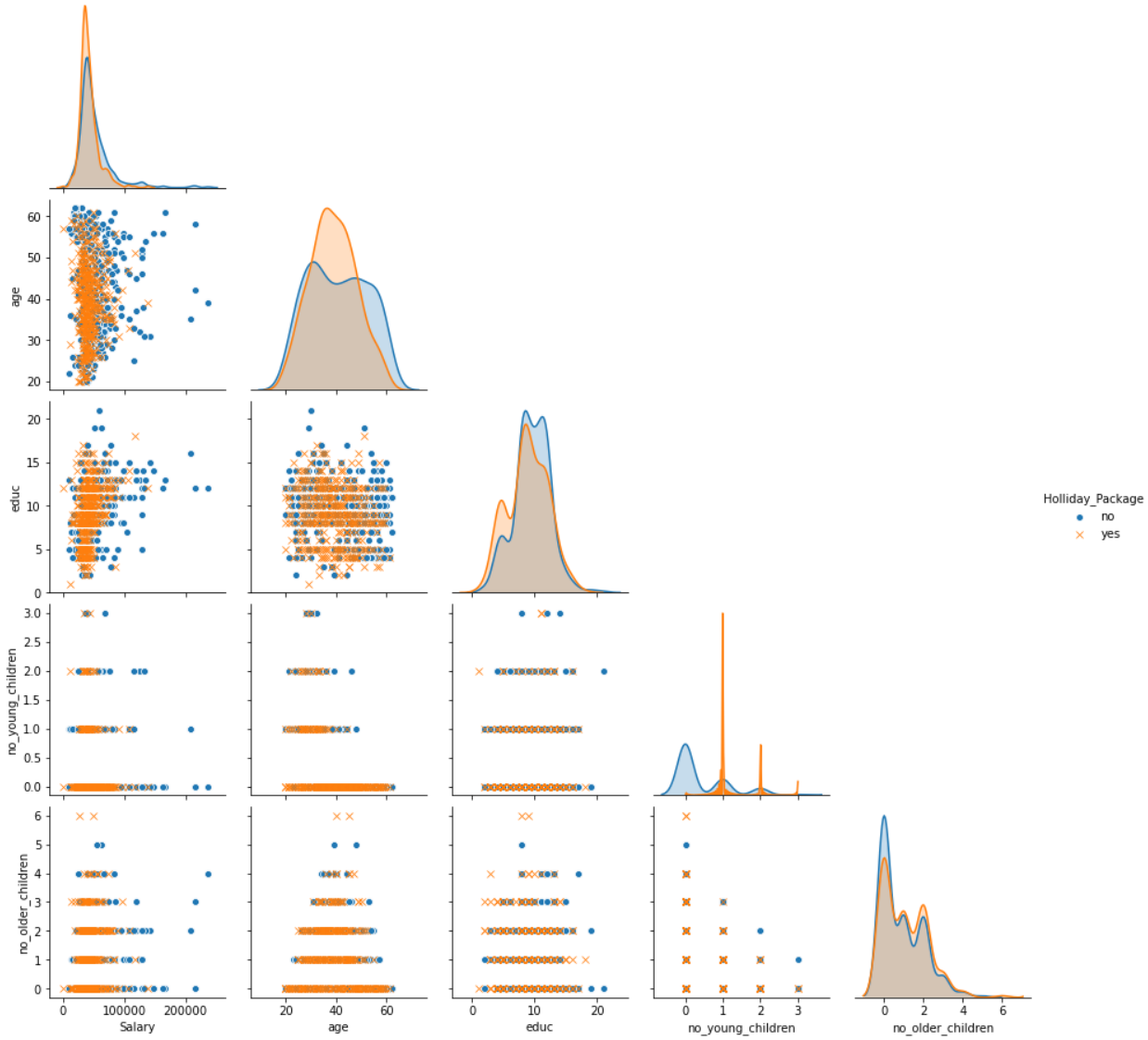

Figure 27: Pairplot

From the below correlation plot we can infer that there is no such correlation among variables except educ and salary which is understandable as educational background does impact the salary of an employee in a company.

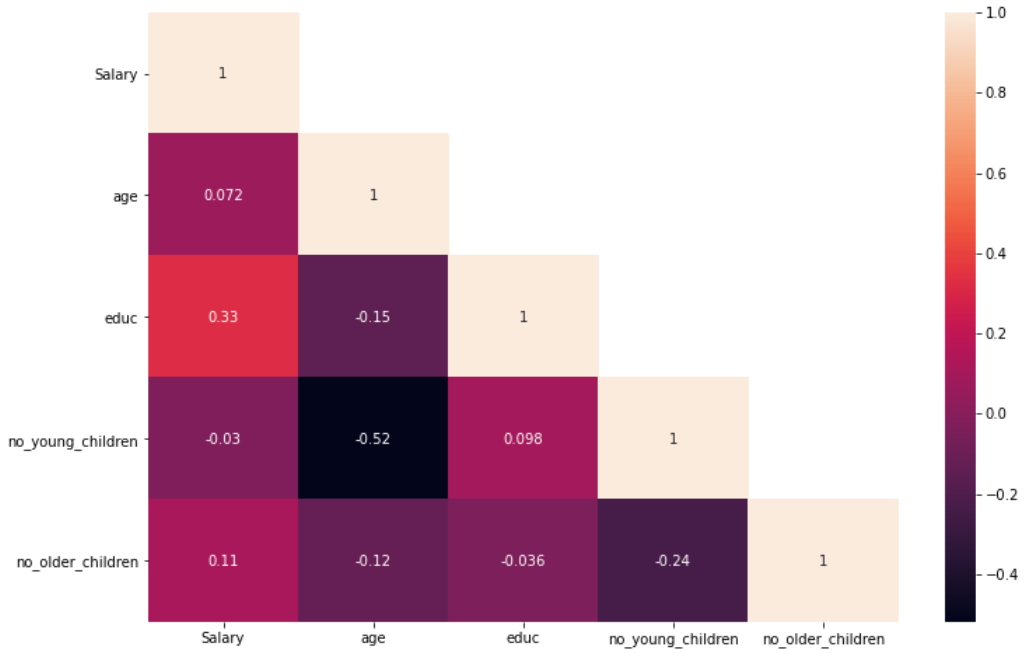

Figure 28: Correlation plot

EDA

We have included a total children attribute to see if number of children has any impact on target regardless of their age.

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

Table 7: Data with total children column added

By plotting each column w.r.t. their decision to opt for a holiday package, we see that

- Age, educ and total children have more or less no impact on the decision to opt for a holiday package
- Employees with lesser salaries and number of younger children are more likely to opt for holiday package
- Employees with more older children opting for holiday pakage are higher

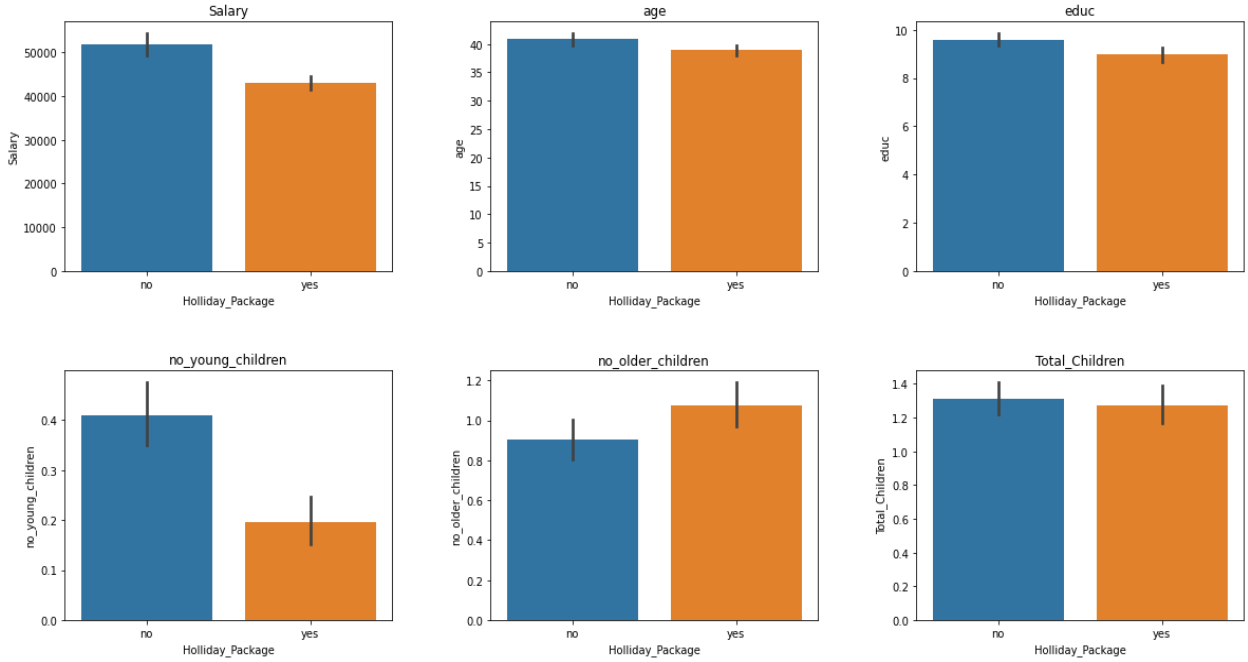

Figure 29: Barplot of dependent and independent variables

This can be further assessed from the below count plots:

- There are many employees with no young children and have opted for holiday package, people with 1, 2 or 3 children opting are lesser
- Although the number of foreigners are less, they are more likely to opt for the holiday package

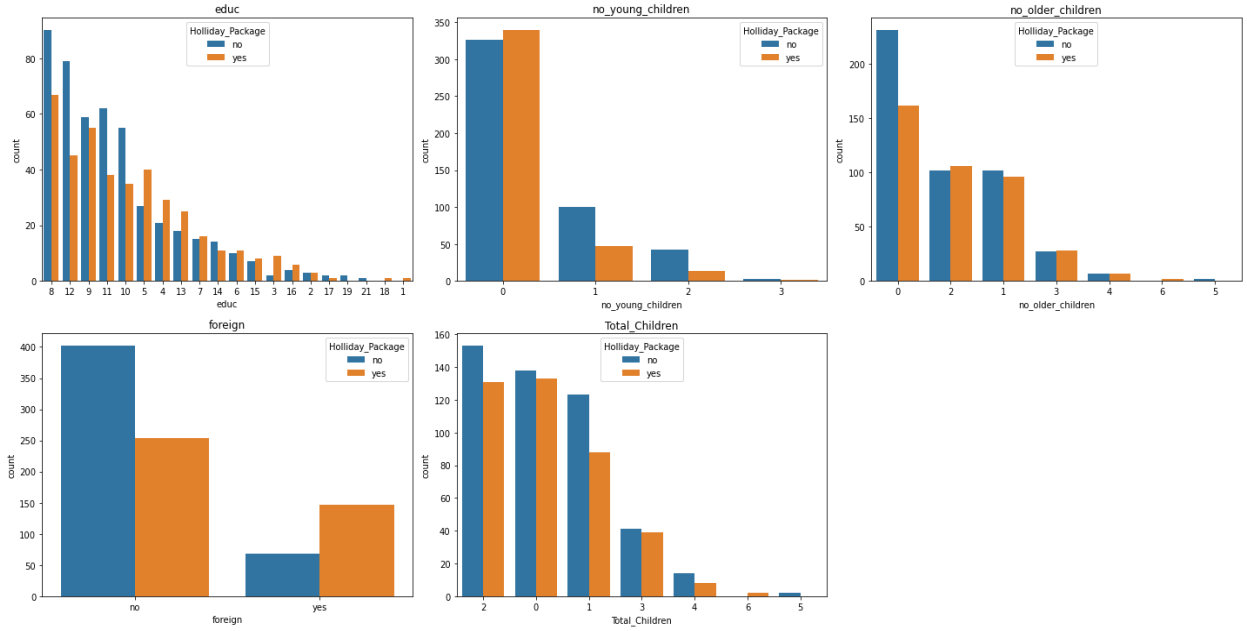

Figure 30: Countplot independent variables

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

Ans)

For encoding, one hot encoding was performed on the foreign variable and label encoding on the target variable. The train and test data was split in 70:30 ratio. The steps can be viewed from **line -.**

First a base model without any parameters was built for logistic regression and the accuracy for train and test data for it was coming as 53% and 54% respectively. In order to increase the performance GridSearch was applied using the following parameters:

- **Penalty**: Used to specify the norm used in the penalization. The 'newton-cg', 'sag' and 'lbfgs' solvers support only l2 penalties. 'elasticnet' is only supported by the 'saga' solver. If 'none' (not supported by the liblinear solver), no regularization is applied.
- **Solver**: Algorithm to use in the optimization problem.
- **Tol**: Tolerance for stopping criteria.
- **multi_class**: If the option chosen is 'ovr', then a binary problem is fit for each label. For 'multinomial' the loss minimised is the multinomial loss fit across the entire probability distribution, *even when the data is binary*. 'multinomial' is unavailable when solver='liblinear'. 'auto' selects 'ovr' if the data is binary, or if solver='liblinear', and otherwise selects 'multinomial'.

The accuracy of model improved to 67% and 65% for train and test data respectively.

Then LDA model was built. The accuracy for training and testing data for it was 66% and 64% for train and test data respectively.

The best threshold values for both the models were checked from **line -** . The best values which enhanced further enhanced the performances of both the models came as 0.4. The performance metrics will be mentioned in the following question.

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.

Ans)

The code can be referred from line 95 in notebook.

1. **<u>Logistic Regression:</u>**

Accuracy:

Train set: 0.67
Test set: 0.65

Since the difference in accuracies of train and test is almost similar, they are converging and can be deemed as a good fit model.

Classification Report:

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.67	0.74	0.71	329
1	0.66	0.58	0.62	281
accuracy			0.67	610
macro avg	0.67	0.66	0.66	610
weighted avg	0.67	0.67	0.66	610

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.65	0.77	0.71	142
1	0.65	0.52	0.58	120
accuracy			0.65	262
macro avg	0.65	0.64	0.64	262
weighted avg	0.65	0.65	0.65	262

Figure 31: Classification report of LR model

The recall and F1-score for test data is not good.

Confusion matrix:

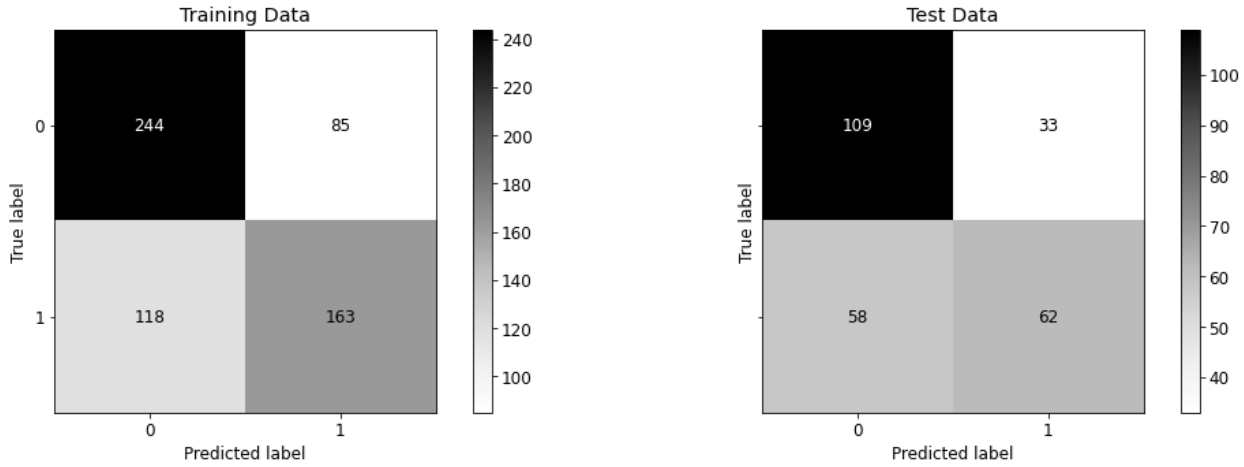

Figure 32: Confusion matrix of LR model

ROC Curve:

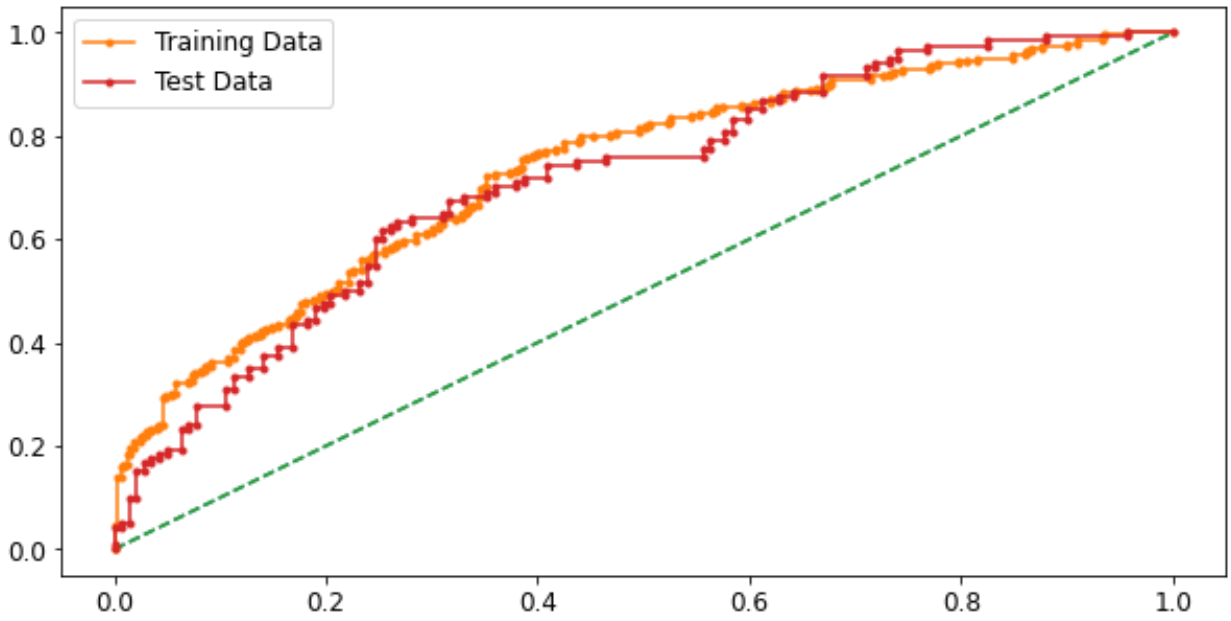

Figure 33: ROC Curve of LR model

AUC:

AUC for the Train set: 0.735

AUC for the Test set: 0.717

The training data performs slightly better than test data.

2. <u>LDA:</u>

Accuracy:

Train set: 0.66

Test set: 0.64

Classification Report:

Classification Report of the training data:

	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

Classification Report of the test data:

	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

Figure 34: Classification report of LDA model

Confusion matrix:

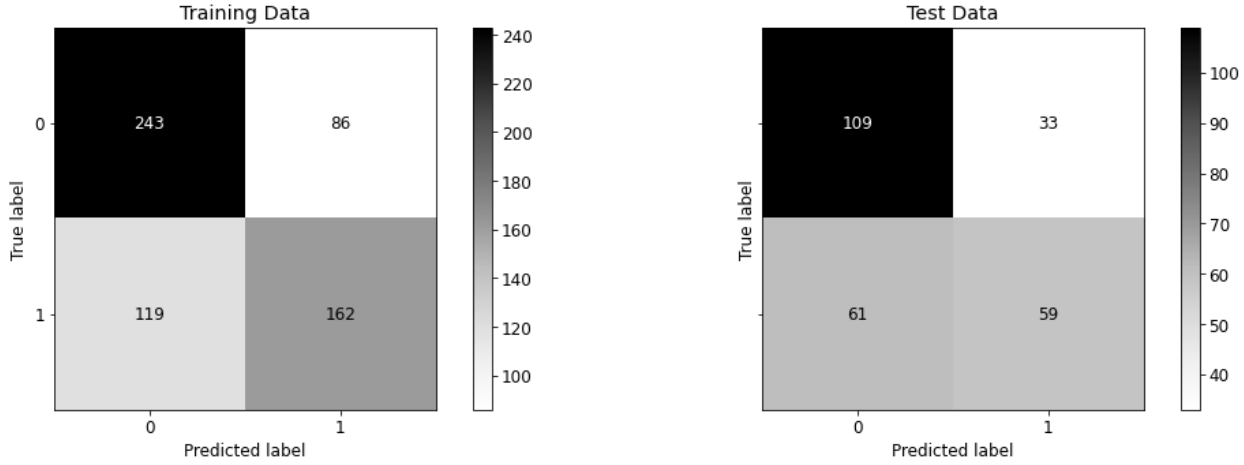

Figure 35: Confusion matrix of LDA model

ROC Curve:

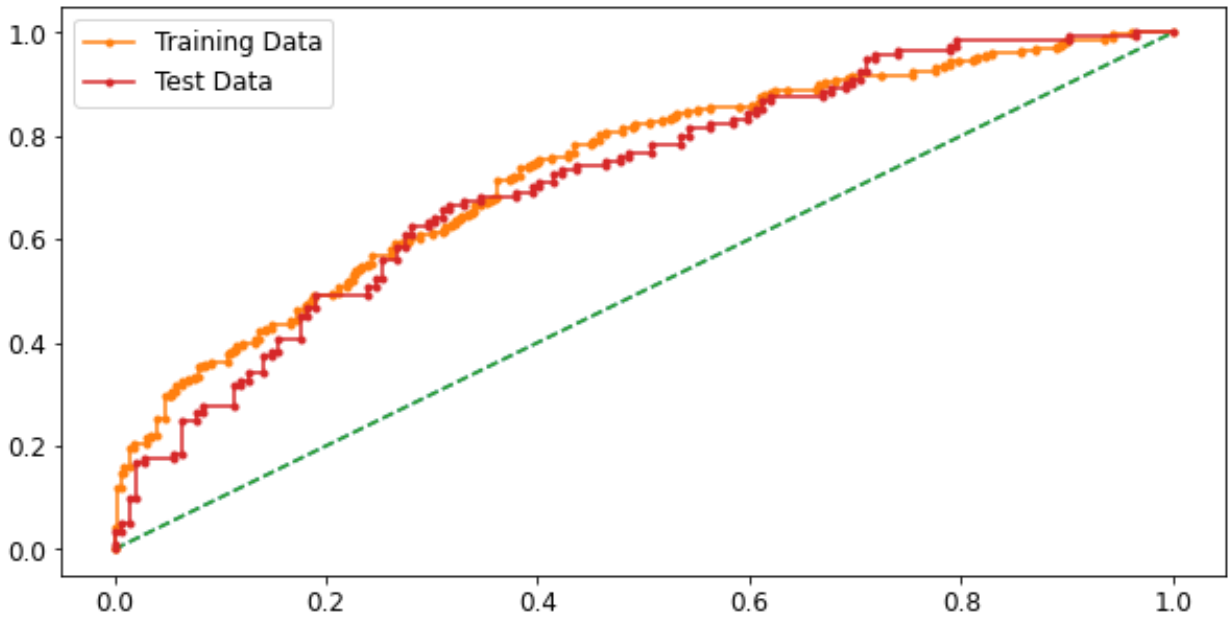

Figure 36: ROC Curve of LDA model

AUC:

AUC for the Training Data: 0.733

AUC for the Test Data: 0.714

The training data performs slightly better than test data.

Comparison of Both models:

	LR Train	LR Test	LDA Train	LDA Test
Accuracy	0.67	0.65	0.66	0.64
AUC	0.74	0.72	0.73	0.71
Recall	0.58	0.52	0.58	0.49
Precision	0.66	0.65	0.65	0.64
F1 Score	0.62	0.58	0.61	0.56

Table 8: Comparison of metrics of LR and LDA models

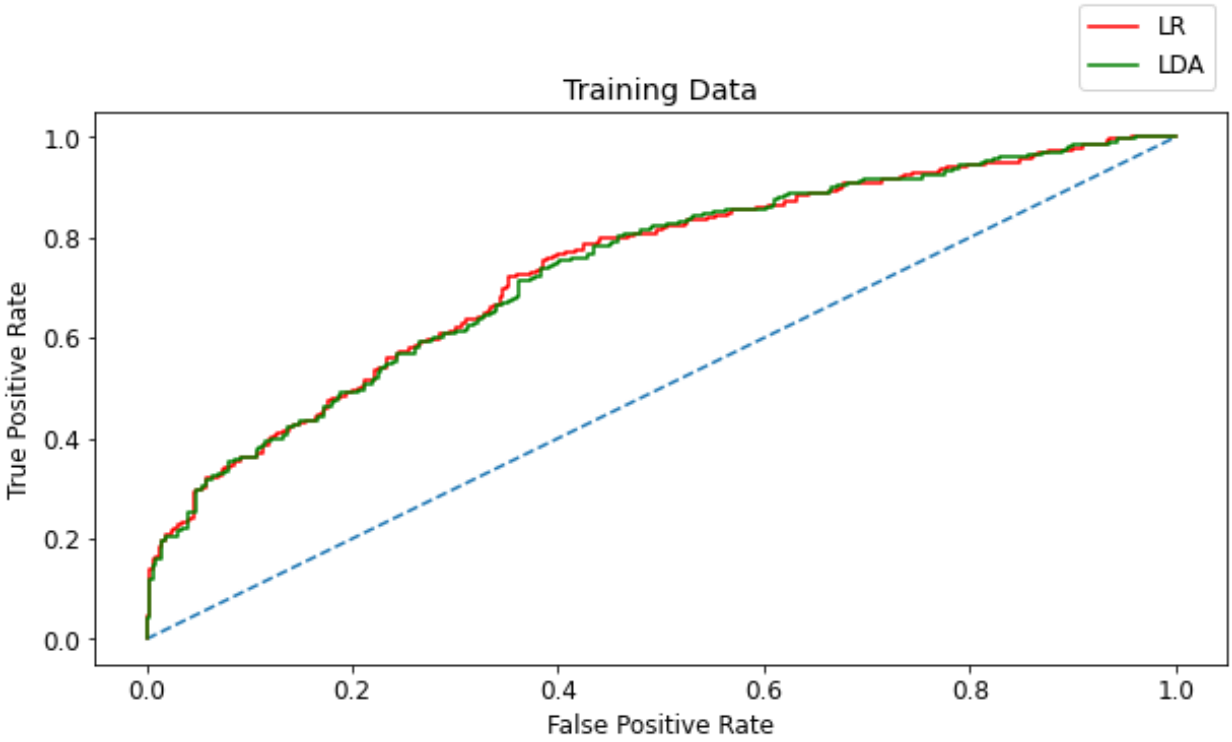

Figure 37: Comparison of ROC curve of LR and LDA model for training data

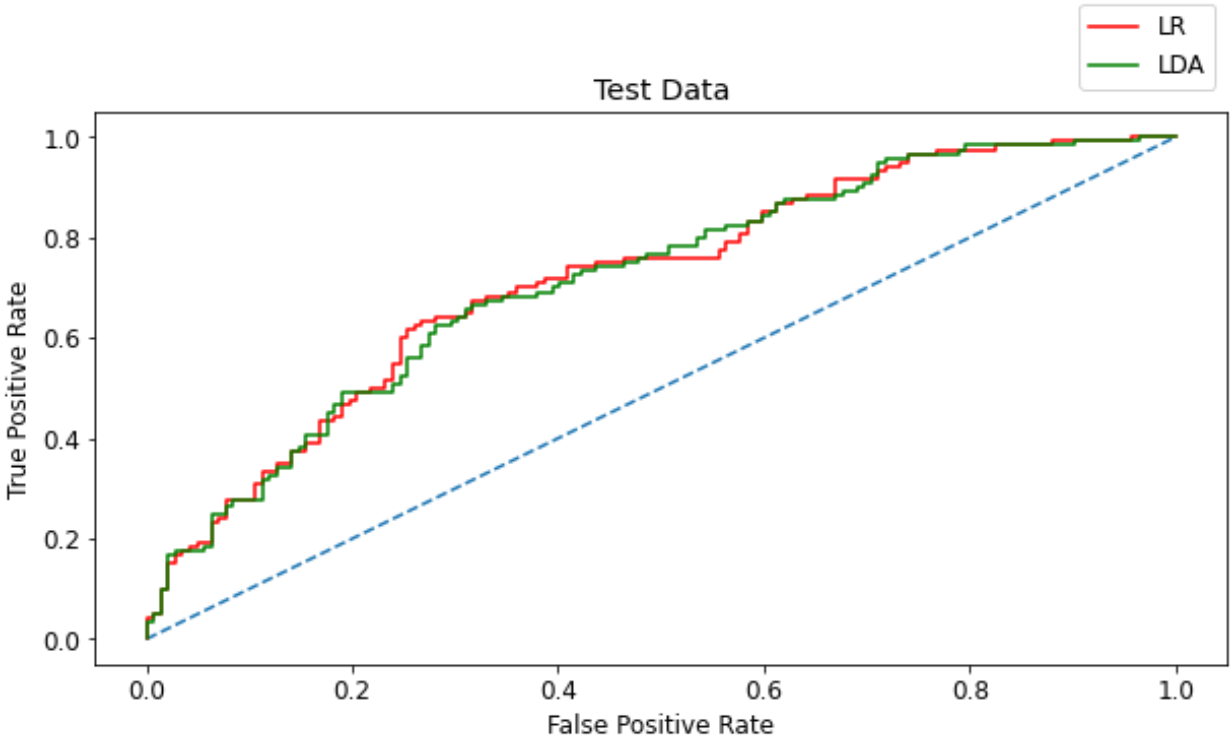

Figure 38: Comparison of ROC curve of LR and LDA model for testing data

Here it can be observed that the logistic regression has performed slightly better than LDA model in terms of above parameters. But still the values of recall all F1 score are not satisfactory.

In an attempt to further enhance the performance of both models, different threshold values were tried to find the best accuracy and F1-score values, the code can be referred from line 106.

The default threshold taken by model is 0.5. It was found that threshold value of **0.4** in this case yielded much better results for both the models.

The performance metrics for them is as follows:

1. Logistic Regression

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.75	0.60	0.67	329
1	0.62	0.77	0.68	281
accuracy			0.68	610
macro avg	0.68	0.68	0.68	610
weighted avg	0.69	0.68	0.67	610

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.72	0.56	0.63	142
1	0.59	0.75	0.66	120
accuracy			0.65	262
macro avg	0.66	0.65	0.64	262
weighted avg	0.66	0.65	0.64	262

Figure 39: Classification report of optimized LR model

Confusion matrix:

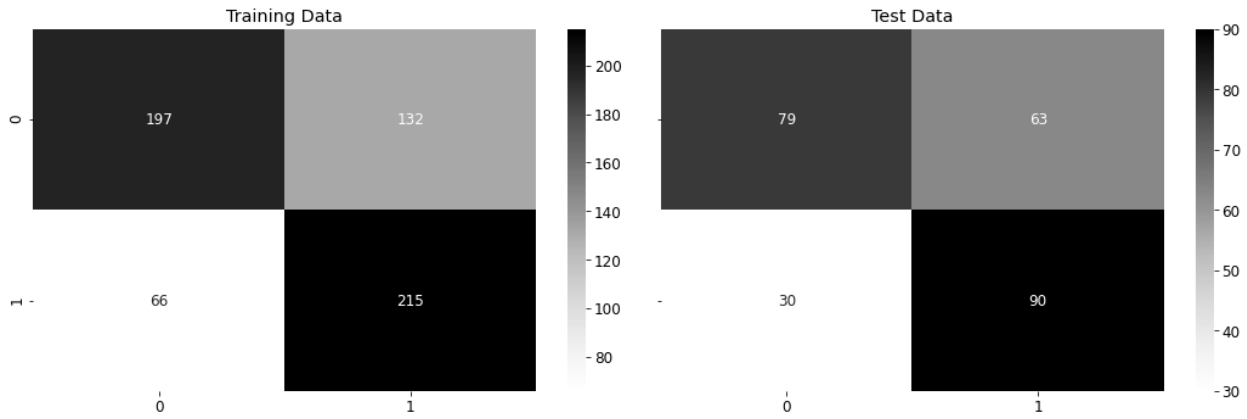

Figure 40: Confusion matrix of optimized LR model

2. LDA

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.74	0.59	0.65	329
1	0.61	0.76	0.68	281
accuracy			0.67	610
macro avg	0.67	0.67	0.67	610
weighted avg	0.68	0.67	0.66	610

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.71	0.58	0.64	142
1	0.59	0.72	0.65	120
accuracy			0.65	262
macro avg	0.65	0.65	0.64	262
weighted avg	0.66	0.65	0.64	262

Figure 41: Classification report of optimized LDA model

Confusion matrix:

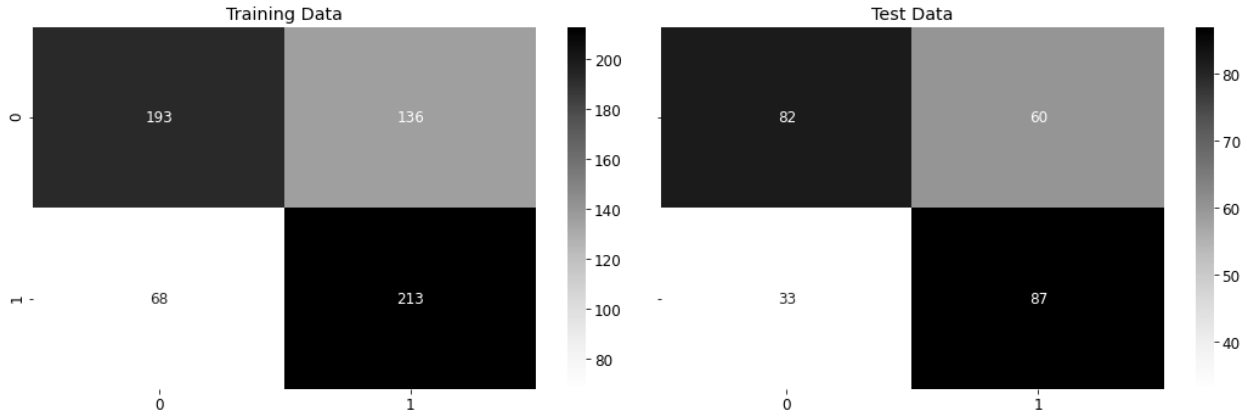

Figure 42: Confusion matrix of optimized LDA model

We can see that although the accuracy is same, the recall and F1-score has tremendously improved for both the models.

	LR Train (Recall)	LR Test (Recall)	LDA Train (Recall)	LDA Test (Recall)	LR Train (F1 score)	LR Test (F1 score)	LDA Train (F1 score)	LDA Test (F1 score)
Threshold 0.5	0.58	0.52	0.58	0.49	0.62	0.58	0.61	0.56
Threshold 0.4	0.77	0.75	0.76	0.72	0.68	0.66	0.68	0.65

Table 9: Comparison of both models with different threshold values

The F1 score of LR and LDA models are almost same, LR better by 0.01 units, and the recall of LR test data is better than LDA model. Both the models have performed almost equally well after optimization. In this case study the company is interested in knowing which employee opts for the holiday package, in that case recall values play more important role hence although both models can be considered here, I would give preference to the logistic regression model. Not only for its better recall values but also better explanability of the importance of variables.

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

Ans)

From the EDA performed in question 2.1, we had observed the following things:
The following were observed from the pairplot:

- Employees having lesser salary were more likely to opt for holiday package
- Employees with 1 young children are most likely to opt for holiday package
- Employees with no older children are most likely to opt for holiday package
- Employees with age above 50 are more likely to opt for holiday package

By plotting each column w.r.t. their decision to opt for a holiday package, we see that

- Age, educ and total children have more or less no impact on the decision to opt for a holiday package
- Employees with lesser salaries and number of younger children are more likely to opt for holiday package
- Employees with more older children opting for holiday package are higher

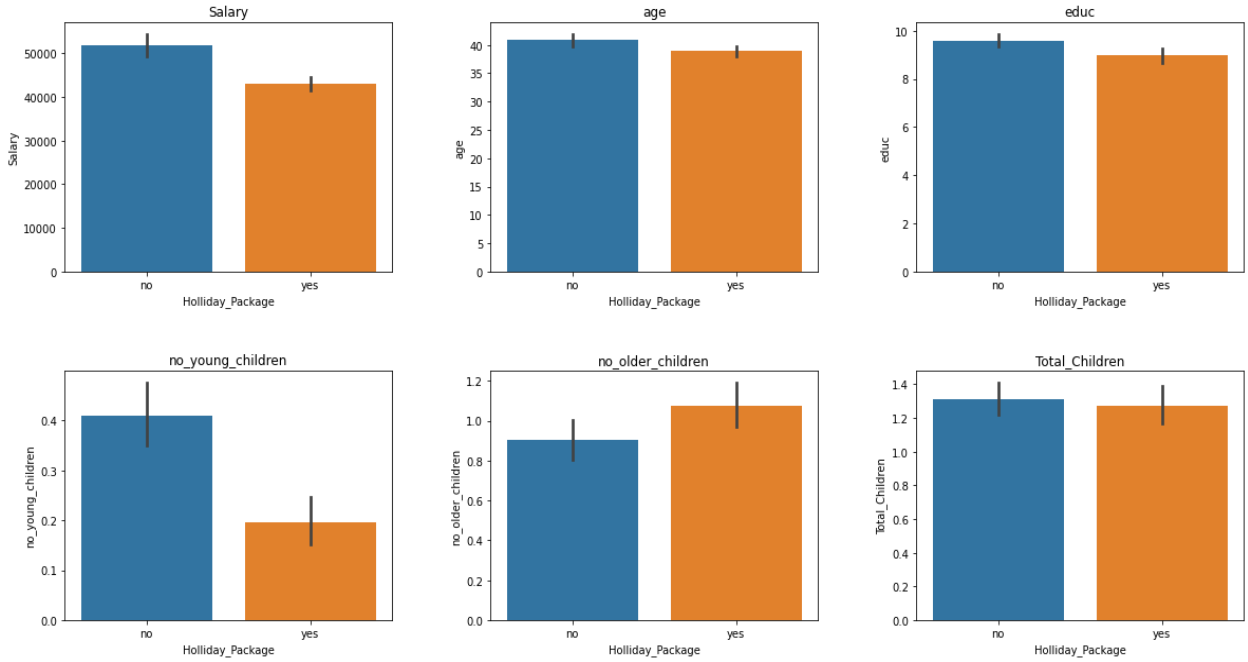

Figure 43: Barplot of dependent and independent variables

This can be further assessed from the below count plots:

- There are many employees with no young children and have opted for holiday package, people with 1, 2 or 3 children opting are lesser
- Although the number of foreigners are less, they are more likely to opt for the holiday package

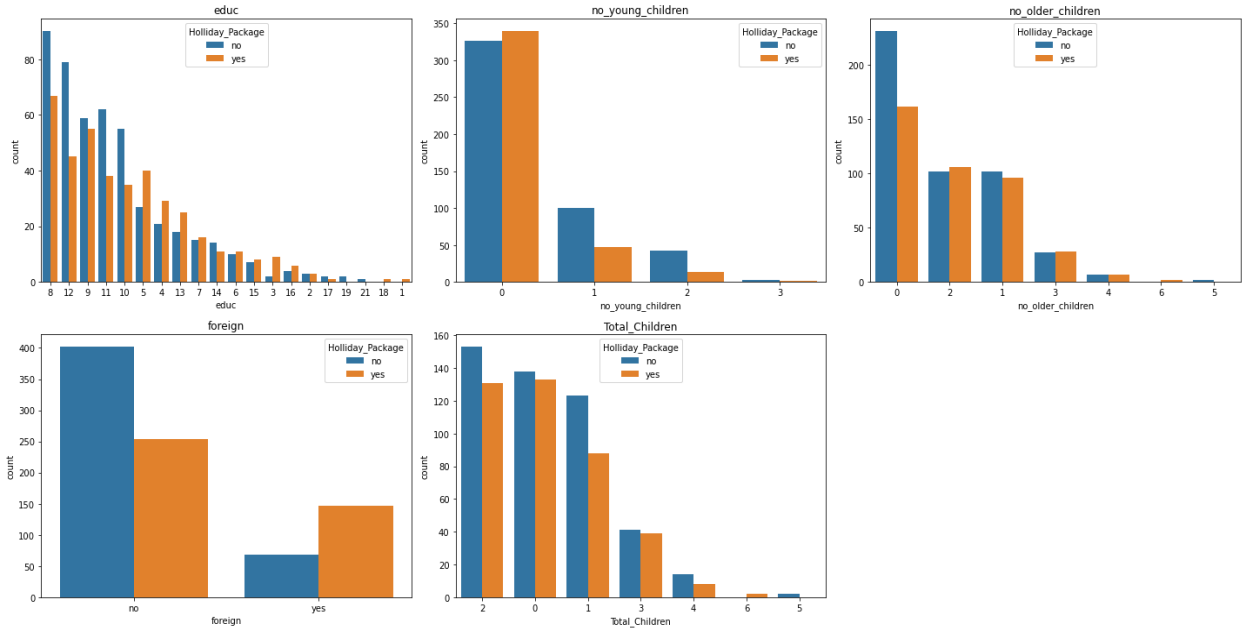

Figure 44: Countplot independent variables

The people with higher salaries can afford to custom arrange their tour and their luxury lies in these custom arrangements. Whereas employees with lower salaries might not travel much but want to travel and are okay with the arrangements provided by travel agencies, so are more willing to pay for the tour packages.
Keeping this in mind the company could have different tour packages for different earning employees so that higher salaried employees could be enticed to opt for holiday packages.

Having younger children makes it difficult to manage children during holidays, so people usually prefer to go on holidays when they are of a manageable age; this trend could also be observed in the given set of employees. The company could think of modifying tour packages, like creating special packages for families with and without children, and include children friendly options like visit to theme park, accessories or emergency support for child care etc so that people are encouraged to go on holidays even with children.

Foreigners are always explorers by nature. They love to explore the place which is foreign to them, and as they might not familiar with the communication systems, they would prefer to take holiday packages as they don't need to take care of logistics, worry if they are being scammed by locals, they just need to enjoy the country. The company also indicates that although the number of foreigners is lower, they more of them have opted for holiday packages. So the company should market their packages to foreigners so that it seems foreigner/tourist friendly and they get a feel of safe and worry free travel experience.