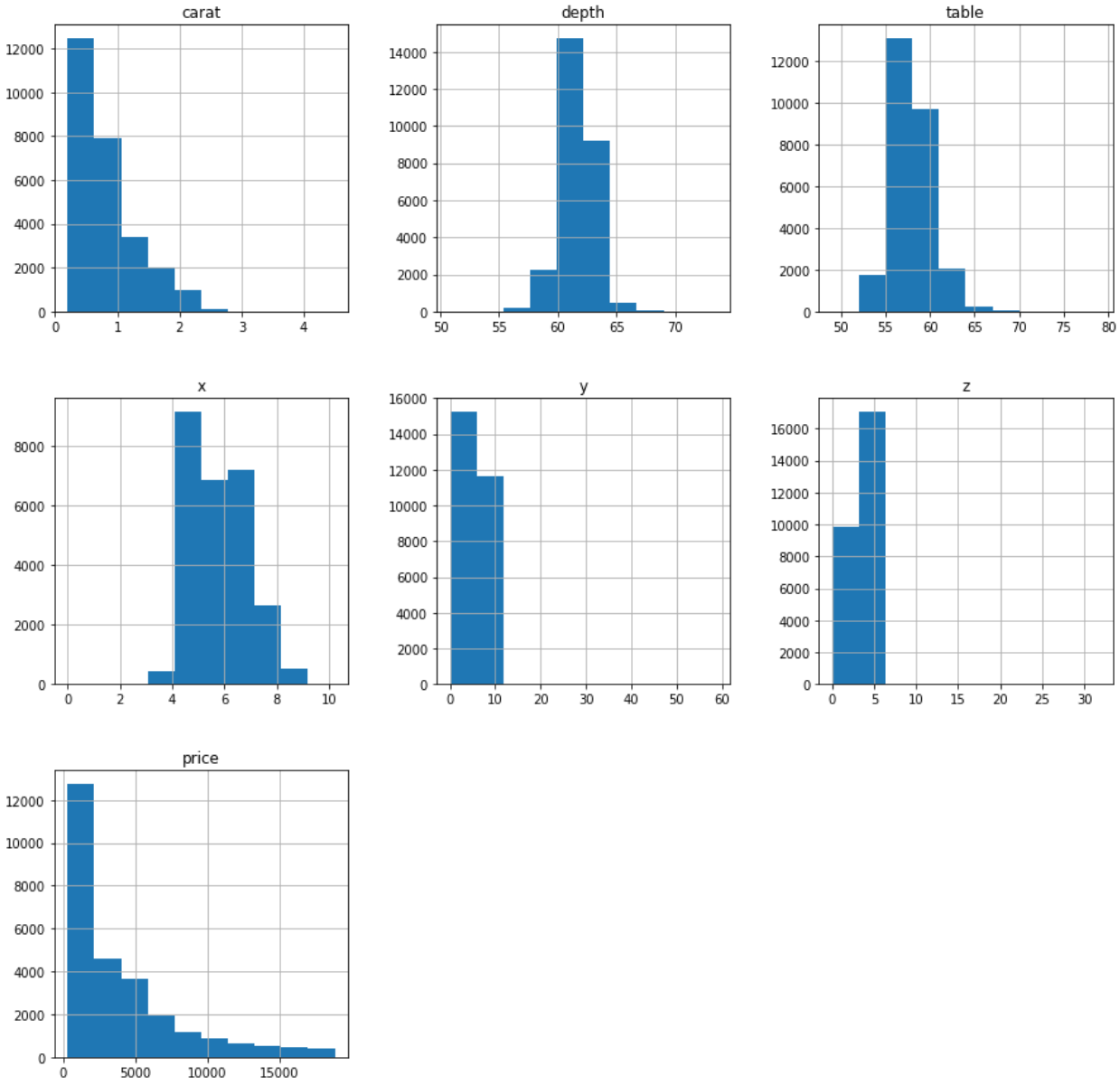
PGP DSBA SEPT'B 2021 BATCH

~ ROHAN CHOURASIA

SOLUTION DOC

1.1)



Variables x, y, and z seems to follow a normal distribution with a few outliers.

Checking Correlation in the data using Heatmap

Observations: High correlation between the different features like carat, x, y, z and price.
Less correlation between table with the other features.
Depth is negatively correlated with most the other features except for carat

Univariate & Bivariate Analysis Getting unique counts of Categorical Variables

Looking at the above unique values for variable “Cut “ we see the ranking given for each unique value like “ Fair, Good, Ideal, Premium, Very Good “

Price Distribution of Cut Variable

For the cut variable we see the most sold is Ideal cut type gems and least sold is Fair cut gems
All cut type gems have outliers with respect to price
Slightly less priced seems to be Ideal type and premium cut type to be slightly more expensive

Price Distribution of Color Variable

For the color variable we see the most sold is G colored gems and least is J colored gems
All color type gems have outliers with respect to price
However, the least priced seems to be E type; J and I colored gems seems to be more expensive

Price Distribution of Clarity Variable

For the clarity variable we see the most sold is SI1 clarity gems and least is I1 clarity gems
All clarity type gems have outliers with respect to price
Slightly less priced seems to be SI1 type; VS2 and SI2 clarity stones seems to be more expensive

Getting unique counts of Numeric Variables

Histograms and Boxplot for each variable to check the data distribution Observations:
Independent Variables
Depth is the only variable which can be considered as normal distribution
Carat, Table, x, y, z these variables have multiple modes with the spread of data
Outliers: Large number of outliers are present in all the variables (Carat, Depth, Table, x, y, z)
Price will be the target variable or dependent variable
It is right skewed with large range of outliers

There are outliers present in all the variables as per the above plot

From above data it is seen that except for carat and price variable, all other variables have mean and median values very close to each other, seems like there is no skewness in these variables. Whereas for carat and price we see some difference in value of mean and median,

which slightly indicates existence of some skewness in the data Treatment of outliers by IQR method Box Plots after outliers' treatment

Checked for data Correlation via heatmap: Heatmap showing correlation between variables

We see strong correlation between Carat, x,y, and z that are demonstrating strong correlation or multicollinearity

Bivariate Analysis : Pair Plot :

Observations: Pair plot allows us to see both distribution of single variable and relationships between two variables.

Price – This variable gives the continuous output with the price of the cubic zirconia stones. This will be our Target Variable. Carat, depth, table, x, y, z variables are numerical or continuous variables. Cut, Clarity and colour are categorical variables.

1.2)
We can see that independent variables such as Carat, x, y, and z have a strong association. All of these variables have a strong relationship with the goal variable, which is price. This suggests that our dataset is suffering from multicollinearity. The price variable has no substantial relationship with depth. Before developing the linear regression model for this case study, I would remove the x, y, and z variables. Similarly, Depth does not appear to influence my variable price, therefore I will remove it from my model-building procedure at some time. In light of the foregoing, I don't believe scaling the data makes sense for this dataset.

1.3)
Model score on the test data is 91.7% Our model is in the right fit zone. We have a good model with us. Our model is neither an underfit nor an overfit model and its working out fine for us.

1.4)
Carat is, as expected, a powerful determinant of the stone's overall price. Clarity refers to the lack of Inclusions and Blemishes, and it has also proven to be a good predictor of pricing. The corporation is using the clarity of stone types IF, VVS 1, VVS 2, and vs1 to set an expensive price cap on the stones. The colour of the stones, such as H, I, and J, will not assist the firm in putting a high price cap on such stones. Instead, the corporation should concentrate on stones in the colours D, E, and F in order to fetch greater prices and boost sales.
This could also signal that the corporation should be exploring for new colour stones, such as transparent stones or a different color/unique colour, to assist boost the pricing. To raise prices, the company should concentrate on the carat and clarity of the stone. More earnings will be generated as a result of ideal customers. Customers can be educated about the importance of a higher carat score and the clarity index through marketing initiatives. Following that, the corporation can create segments and target customers depending on their income/paying capacity, among other factors that can be explored further.

2.1)
Insights: If the employee's pay is between $30k and $40k, there's a good likelihood they'll accept the package.
This indicates that package pricing is average.
If the employee is between the ages of 25 and 50, he or she has a better likelihood of accepting the package.
There's a better possibility of saying no after 50 years.
If the employee does not have any young children, there is a good likelihood that they will say yes.
If the employee is a foreigner, there is a good likelihood that he or she will say yes.

2.3)

Logistic regression Classfication report
Classification Report of the training data:

	precision	recall	f1-score	support
no	0.53	0.90	0.67	326
yes	0.43	0.08	0.14	284
accuracy			0.52	610
macro avg	0.48	0.49	0.40	610
weighted avg	0.48	0.52	0.42	610

Classification Report of the test data:

	precision	recall	f1-score	support
no	0.55	0.89	0.68	145
yes	0.43	0.10	0.17	117
accuracy			0.54	262
macro avg	0.49	0.50	0.42	262
weighted avg	0.50	0.54	0.45	262

AUC and ROC FOR Logistic regression
AUC for the Training Data: 0.568
AUC for the Test Data: 0.627

Accuracy score for LDA train variables
0.6721311475409836

Accuracy score for LDA test variables
0.6412213740458015

LDA Classfication report
Classification Report of the training data:

	precision	recall	f1-score	support
no	0.67	0.77	0.72	326
yes	0.68	0.56	0.61	284
accuracy			0.67	610
macro avg	0.67	0.66	0.66	610
weighted avg	0.67	0.67	0.67	610

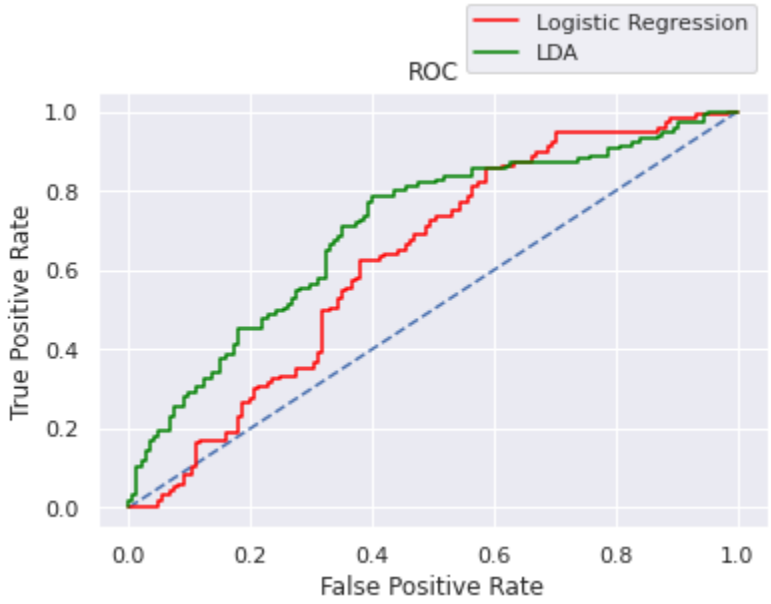
Classification Report of the test data:

	precision	recall	f1-score	support
no	0.66	0.71	0.69	145
yes	0.61	0.56	0.58	117
accuracy			0.64	262
macro avg	0.64	0.63	0.63	262
weighted avg	0.64	0.64	0.64	262

AUC for the Training Data: 0.742
AUC for the Test Data: 0.703

	Logistic reg Train		Logistic reg Test	LDA Train	LDA Test
Accuracy	0.52	0.54	0.67	0.64	
AUC	0.57	0.63	0.74	0.70	
Recall	0.08	0.10	0.56	0.56	
Precision	0.43	0.43	0.68	0.61	
F1 Score	0.14	0.17	0.61	0.58	

ROC curve for Test data



Based on comparing the performance metrics, Linear discriminant analysis (LDA) performs better than the Logistic regression because it has the best recall rate .Even accuracy is more for LDA.So it is the best model .

2.4) The majority of those who have signed up for the package have an income of between $30,000 and $40,000. It implies that the package is of average cost and has medium-level amenities.
So, if they add some additional luxury packages with amenities like star hotel bookings, luxury automobiles, and so on, it may assist to boost package sales to a higher income group.
According to the findings, foreigners are more likely to choose packages than non-foreigners.

This, combined with the preceding study, which shows that the majority of persons earn between $30K and $50K (suggesting that the package is not expensive), suggests that the packages supplied are either of local sightseeing places or of little interest to non-foreigners. As

a result, advise to the company that they include extra activities or destinations in their packages.

According to the analysis, if the employee does not have any young children, there is a higher chance that they will accept the package. As the number of children grows, so does the inclination to purchase a vacation package. So, I propose that the employer offer additional discounts or child-friendly benefits to employees with small children in order to increase the likelihood of them signing up for the package.