**ADITHYA MANIVANNAN**

**PGP – DSBA**

**PROJECT – PREDICTIVE MODELING**

Contents

[**PROBLEM 1:** Linear Regression 3](#_Toc96830178)

[Introduction 3](#_Toc96830179)

[Data Dictionary: 4](#_Toc96830180)

[1.1 Read the data and do exploratory data analysis. Describe the data briefly. Perform Univariate and Bivariate Analysis. 4](#_Toc96830181)

[Sample of the dataset: 5](#_Toc96830182)

[Exploratory Data Analysis: 5](#_Toc96830183)

[5 Point Summary: 5](#_Toc96830184)

[Bivariate Analysis 7](#_Toc96830185)

[Pairplot 8](#_Toc96830186)

[Correlation heatmap: 8](#_Toc96830187)

[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? 9](#_Toc96830188)

[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. 10](#_Toc96830189)

[1.4 Inference: Basis on these predictions, what are the business insights and recommendations. 13](#_Toc96830190)

[**Problem 2:** Logistic Regression and LDA 14](#_Toc96830191)

[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. 15](#_Toc96830192)

[Sample Dataset: 15](#_Toc96830193)

[Exploratory Data Analysis: 15](#_Toc96830194)

[Descriptive Statistics 16](#_Toc96830195)

[Bivariate Analysis 18](#_Toc96830196)

[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). 20](#_Toc96830197)

[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. 21](#_Toc96830198)

[b) Performance Metrics for LDA 23](#_Toc96830199)

[Confusion Matrix For LDA 23](#_Toc96830200)

[2.5 Basis on these predictions, what are the insights and recommendations. 25](#_Toc96830201)

# **PROBLEM 1:** Linear Regression

## Introduction

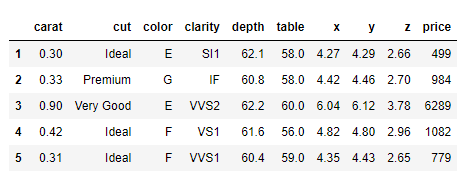
You are hired by a company Gem Stones co ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of almost 27,000 cubic zirconia (which is an inexpensive diamond alternative with many of the same qualities as a diamond). The company is earning different profits on different prize slots. You have to help the company in predicting the price for the stone on the bases of the details given in the dataset so it can distinguish between higher profitable stones and lower profitable stones so as to have better profit share. Also, provide them with the best 5 attributes that are most important.

## Data Dictionary:

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Carat | Carat weight of the cubic zirconia. |
| Cut | Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal. |
| Color | Colour of the cubic zirconia.With D being the worst and J the best. |
| Clarity | Clarity refers to the absence of the Inclusions and Blemishes. (In order from Worst to Best in terms of avg price) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1 |
| Depth | The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter. |
| Table | The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter. |
| Price | the Price of the cubic zirconia. |
| X | Length of the cubic zirconia in mm. |
| Y | Width of the cubic zirconia in mm. |
| Z | Height of the cubic zirconia in mm. |

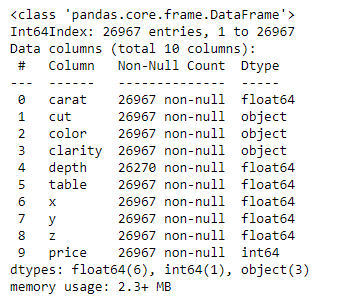
### Read the data and do exploratory data analysis. Describe the data briefly. Perform Univariate and Bivariate Analysis.

### Sample of the dataset:



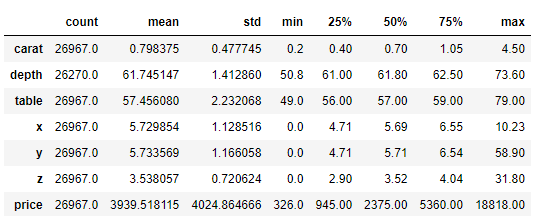
The dataset has 10 different variables that describe the various attributes of the zirconia cube.

### Exploratory Data Analysis:



From the above diagram, we can understand that there are 10 columns with 26967 rows each except for the “depth” column that seems to have missing values. Cut, color and clarity are object data type ie contains strings value. Remining columns are of numerical datatype (either integer or float).

### 5 Point Summary:



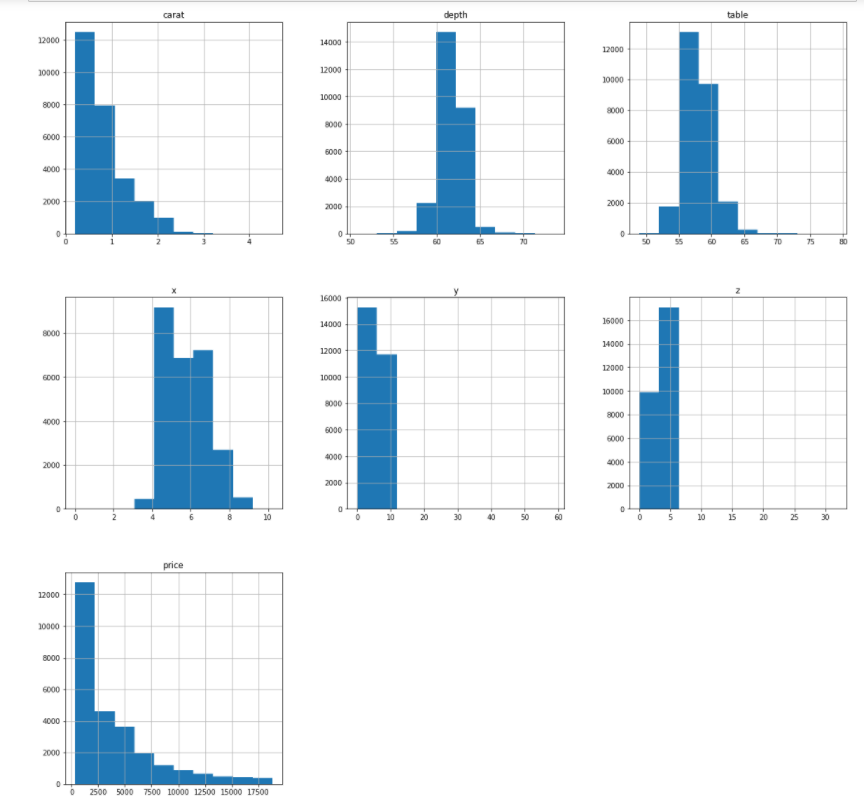
From the 5-point summary, we can understand the following:

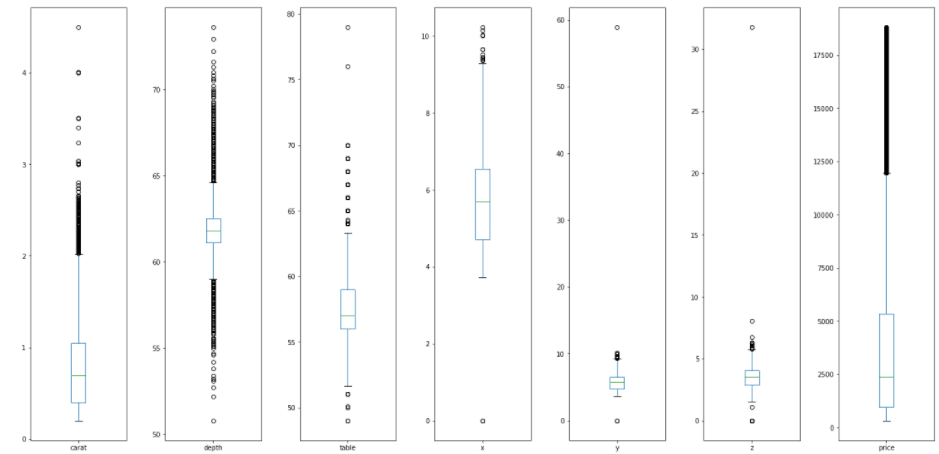
* It looks like all the variables except “price” are normally distributed since the mean and median are almost the same.

All variables (columns) have different scale. For most variables there is a huge difference between the 75thpercentile and maximum value. So, there is chance for outliers.

#### Univariate Analysis:

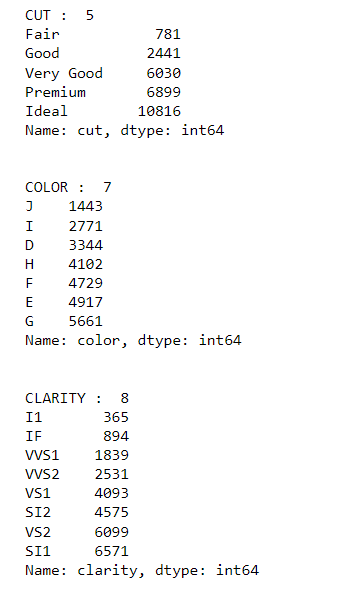
The following plots show the distribution of each variable and that they have different scales.



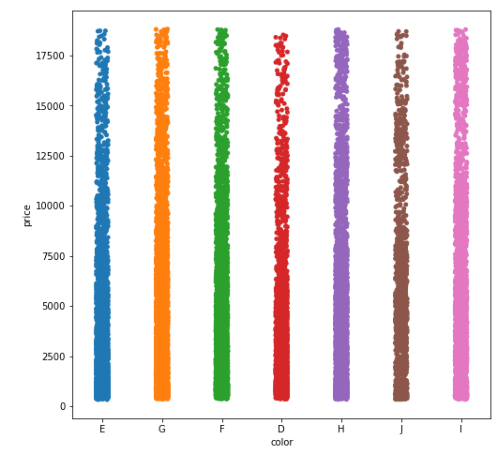


From the boxplot, we can understand that all the variables have their outliers.

We can below see how the distribution of data is within the categorical variables.

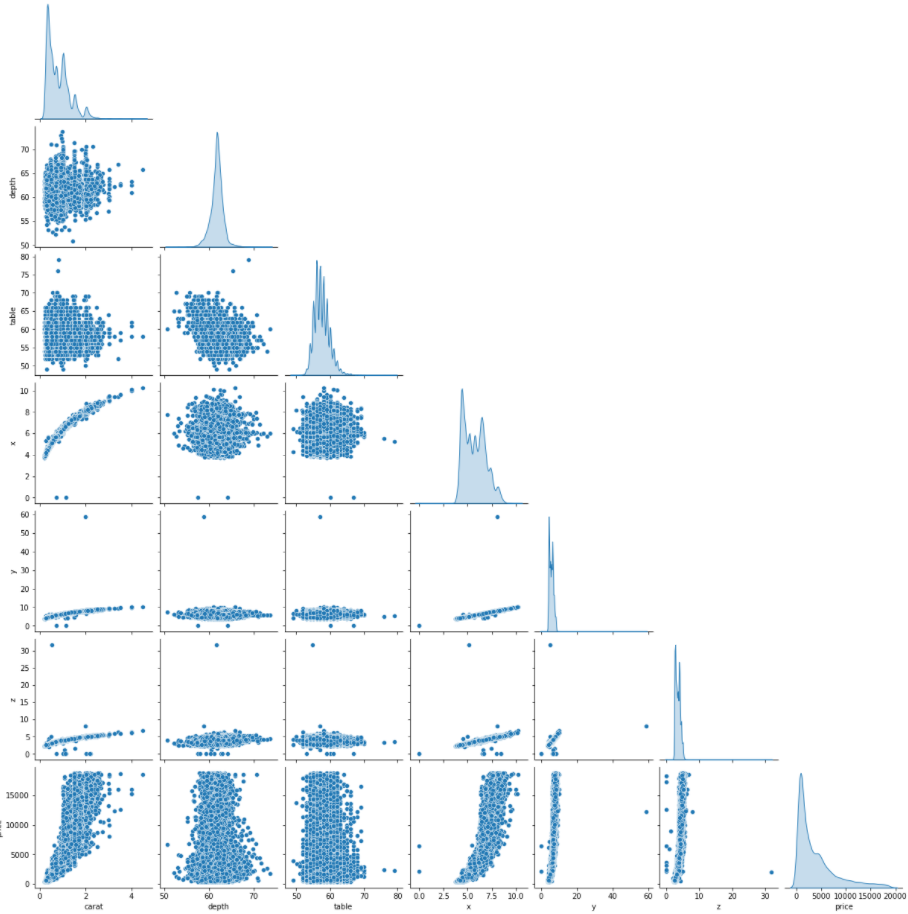


### Bivariate Analysis



There is no significant relation between the color and the price.

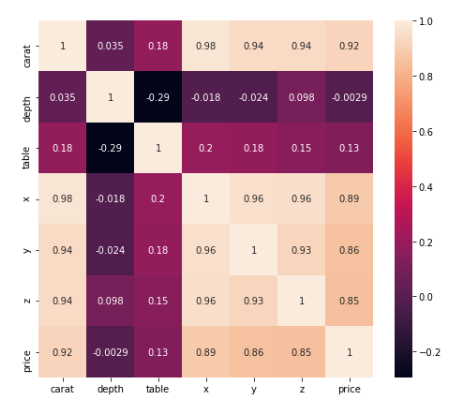
### Pairplot



The pairplot shows that there is a linear relationship between a lot of the variables.

Let us know find the correlation using a correlation heatmap.

### Correlation heatmap:



There is a high numerical correlation between target column price and carat, x, y, z.

There is a high categorical correlation between price and color / clarity.

According to the visualization, there is no much correlation between price and depth.

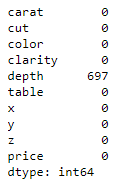
There is a small correlation between price vs table/cut Only the price (target variable) and depth have uniform distribution. Other variable has random distribution with multiple peaks. This may be due to multiple groups in the other variables.

Price (Target variable) Is right Skewed.

Business Insights: This information suggests that as price increases the sales (quantity) reduces. The carat and dimensions (x(length), y (width) and z (height)) plays the huge impact for the pricing than other variables.

## 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.)Does the data seta as any null values?



Depth column has null values.

b.) Does the data seta has zero as values?

No. There is no zero value in the dataset.

c) Do they null values in the dataset has any meaning or do we need to change them or drop them?

As the column depth is least corelated (in fact no correlation) with the target variable, imputing the null values of the depth column will not change the results of prediction much. We will go ahead and impute the null values with the mean of the column.

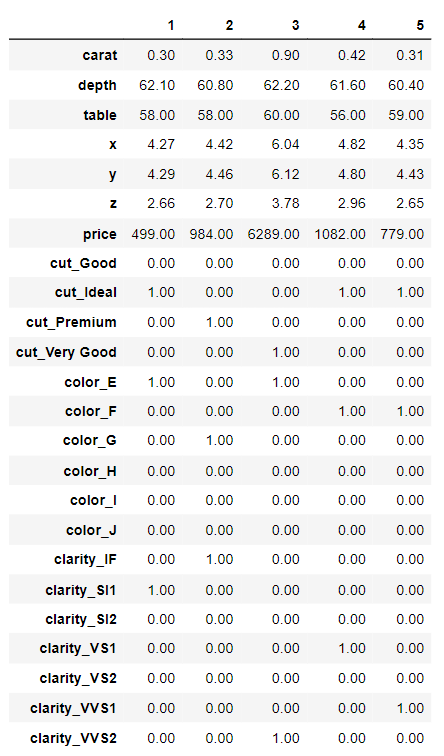
Yes, scaling is recommended. the data columns have different scales .so we need to scale the data set. It will also help us to center the variables and make predators have the value mean 0 (at least near zero). So it will help us to interpret the intercept term as the expected value of price (target value ) when the predictor values are set to their means.

## 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.

Encoding the string values:

I have performed one hot encoding on the object data type variables.

The head of the data set after encoding is as of follows.



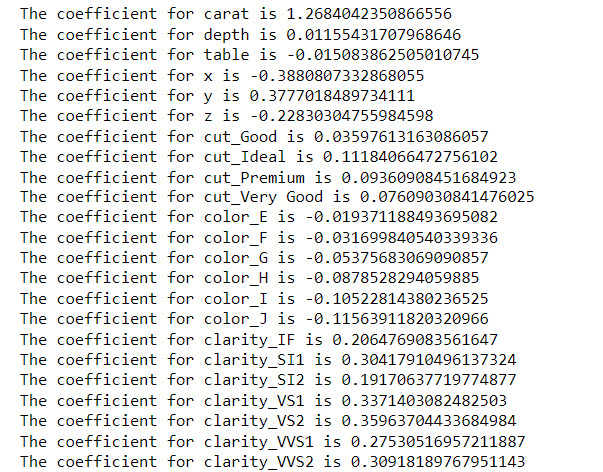
After the encoding, we split the data set into train and test using the sklearn train\_test\_split function in the 70/30 ratio.

We then apply the **z\_score** method and scale the data and pass the scaled data to build the model.

We then perform the linear regression function.



We have successfully built the linear regression model with the following coefficients.



Performance Metrics: Check the performance of Predictions on Train and Test sets using Rsquare, RMSE

Rsquare [Model Score]:

1. Train dataset:



2) Test dataset:



RMSE:

R^2 is not a reliable metric as it always increases with addition of more attributes even if the attributes have no influence on the predicted variable. Instead we use adjusted R^2 which removes the statistical chance that improves R^2 Scikit does not provide a facility for adjusted R^2... so we use statsmodel, a library that gives results similar to what you obtain in R language. This library expects the X and Y to be given in one single dataframe.

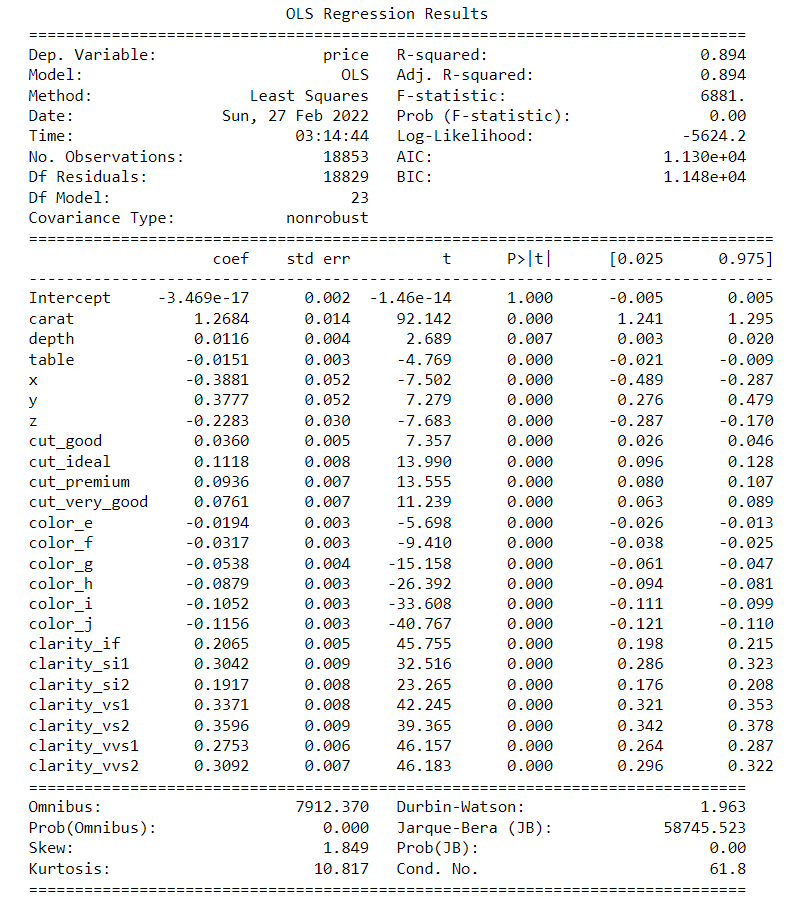
Train Dataset:

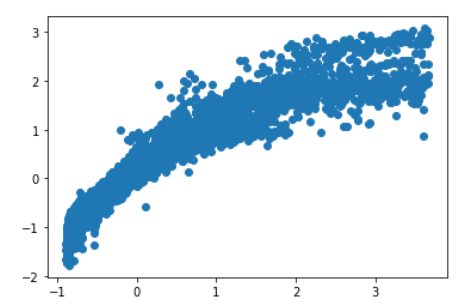


Test Dataset:



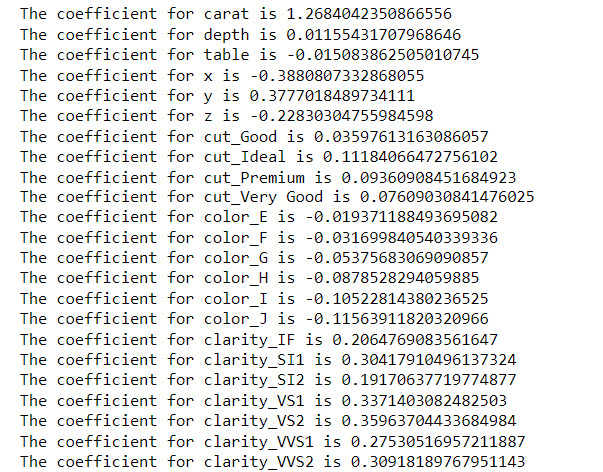
We also use the **statsmodel** method to perform the linear regression model.





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

The coefficient of each variable is :



**Price**= Intercept + (1.2684042351) \* carat + (0.0115543171) \* depth + (-0.0150838625) \* table + (-0.3880807333) \* x + (0.377701849) \* y + (-0.2283030476) \* z + (0.0359761316) \* cut\_good + (0.1118406647) \* cut\_ideal + (0.0936090845) \* cut\_premium + (0.0760903084) \* cut\_very\_good + (-0.0193711885) \* color\_e + (-0.0316998405) \* color\_f + (-0.0537568307) \* color\_g + (-0.0878528294) \* color\_h + (-0.1052281438) \* color\_i + (-0.1156391182) \* color\_j + (0.2064769084) \* clarity\_if + (0.304179105) \* clarity\_si1 + (0.1917063772) \* clarity\_si2 + (0.3371403082) \* clarity\_vs1 + (0.3596370443) \* clarity\_vs2 + (0.2753051696) \* clarity\_vvs1 + (0.3091818977) \* clarity\_vvs2

The most important factors which determines the price of zirconia are Carat, x (length (negative coefficient)), y (width), clarity (some of them have a higher factor than others) and z (height (negative coefficient))

When Carat increases by 1 unit, price of zirconia increases by 1.28 units, keeping all other predictors constant.

When X(Length.) increases by 1 unit, price decreases by 0.388 units, keeping all other predictors constant.

When Y (width) increases by 1 unit, price increases by 0.377 units, keeping all other predictors constant.

When Z(Height.) increases by 1 unit, price decreases by 0.228 units, keeping all other predictors constant.

Color, Cut and Depth have relatively lower importance in factoring the price.

Hence, the recommendation to the business is to focus on increasing the ones that influence price positively such as carat, cut, y and clarity while trying to reduce x,z and color to make the most profits.

The five most important factors are:

Carat , x, y, clarity\_vs2, clarity\_vs1 and clarity\_vvs2.

# **Problem 2:** Logistic Regression and LDA

**Business scenario:**

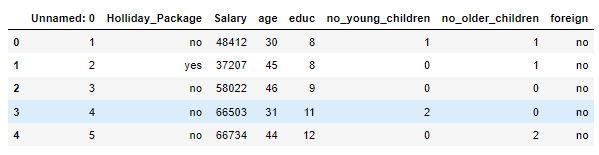
You are hired by a tour and travel agency which deals in selling holiday packages. You are provided details of 872 employees of a company. Among these employees, some opted for the package and some didn't. You have to help the company in predicting whether an employee will opt for the package or not on the basis of the information given in the data set. Also, find out the important factors on the basis of which the company will focus on particular employees to sell their packages

**Data Dictionary:**

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Holiday\_Package | Opted for Holiday Package yes/no? |
| Salary | Employee salary |
| age | Age in years |
| edu | Years of formal education |
| no\_young\_children | The number of young children (younger than 7 years) |
| no\_older\_children | Number of older children |
| foreign | foreigner Yes/No |

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

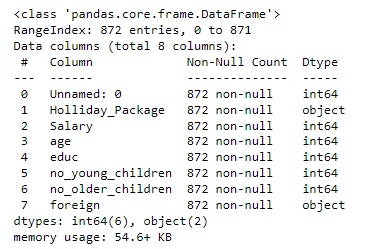
### Sample Dataset:



Dataset has 8 columns. The first column (Unnamed column :0) is of no use for analysis and can be removed. Holliday\_Package will be our target variable.

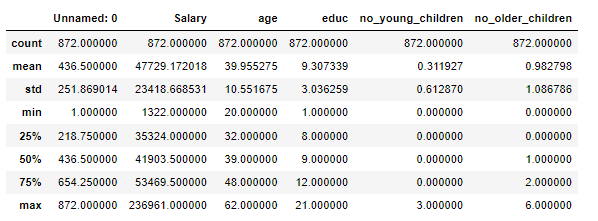
### Exploratory Data Analysis:

#### Information of dataset:



Dataset has no null values. Holiday\_package and foreign columns are of object data type i.e. contains strings value. Remining columns are of numerical datatype (either integer). Dataset has 872 observations (Rows of data)

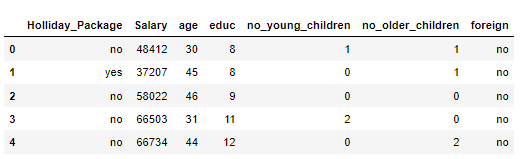
### Descriptive Statistics



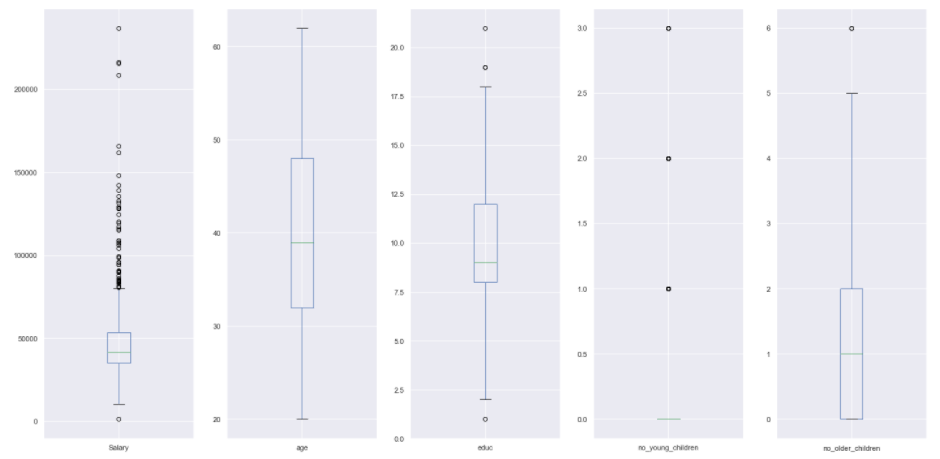
All variables (columns) have different scale. For most variables there is a huge difference between the 75thpercentile and maximum value compared to the 50 percentile and 75 percentiles So, there is chance for outliers.

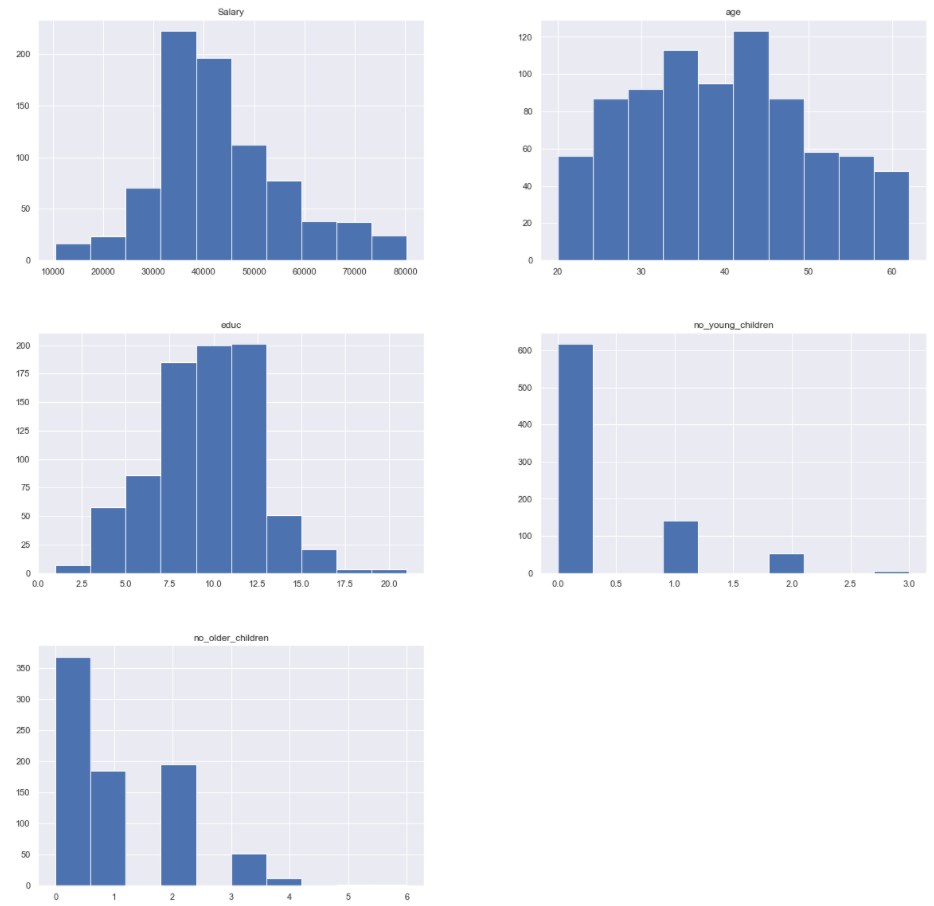
Dataset has no null values.

It also has no duplicates. The “Unnamed: 0” column was dropped.



#### Univariate Analysis



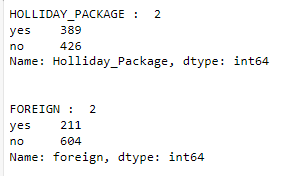


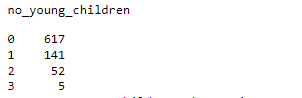
Only Salary and age is continuous variable. Others are categorical. Only age variables does not have outlier while the others do.

It looks like Salary, age and educ are normally distributed as well.

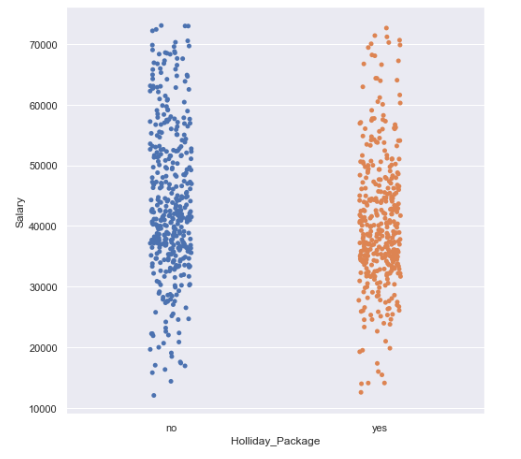
We shall go ahead and treat the outliers present in the salary column.

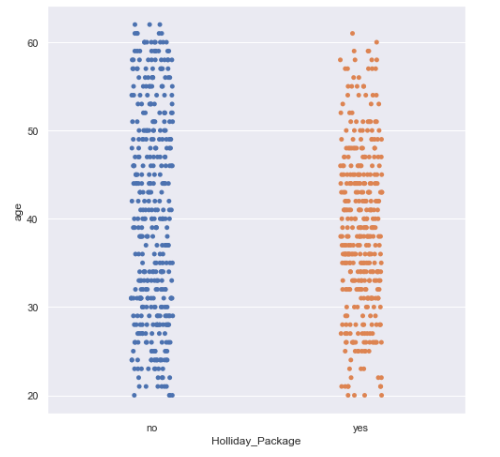
Value count of categorical variables.

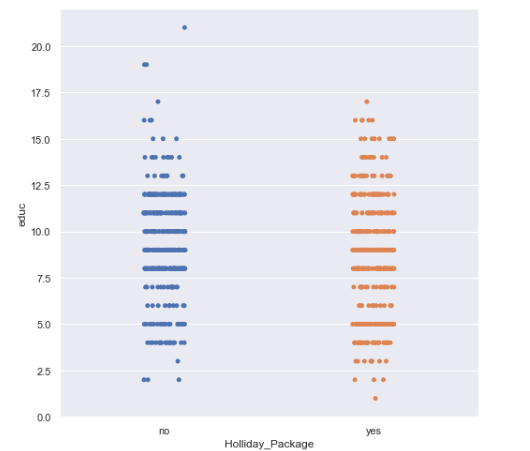




### Bivariate Analysis

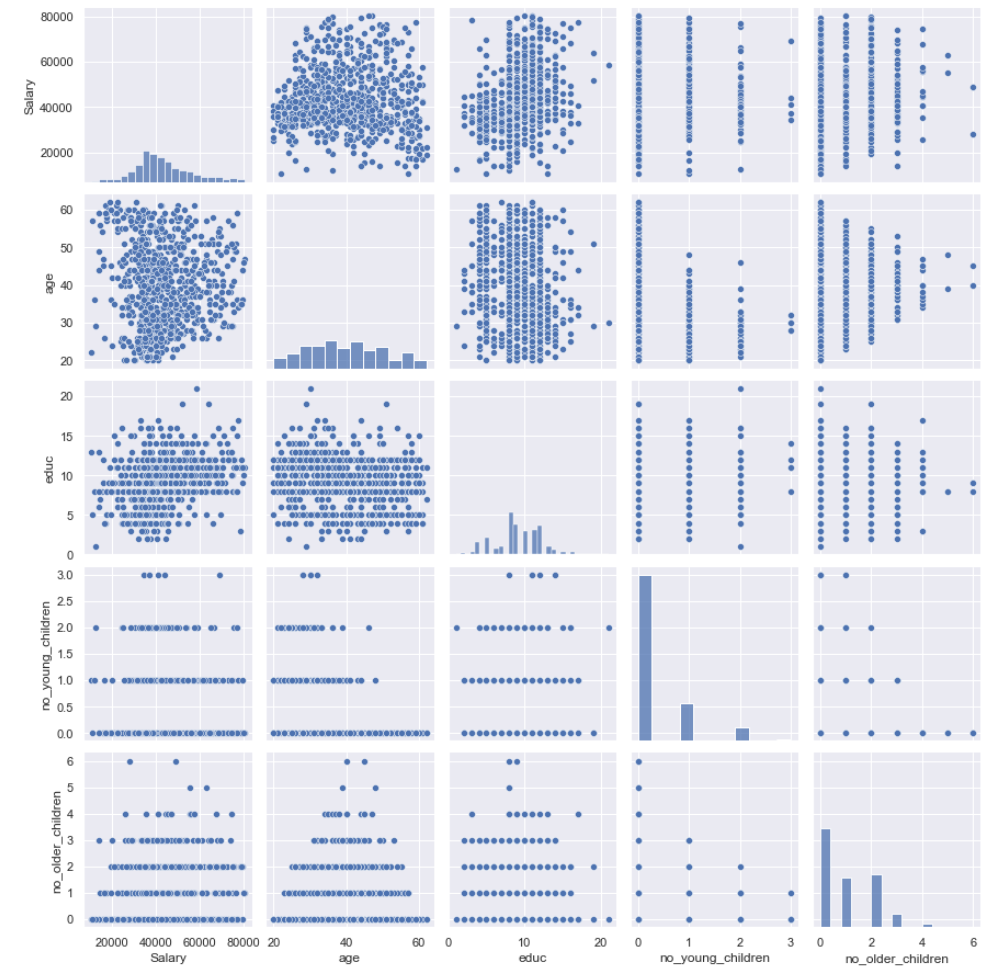






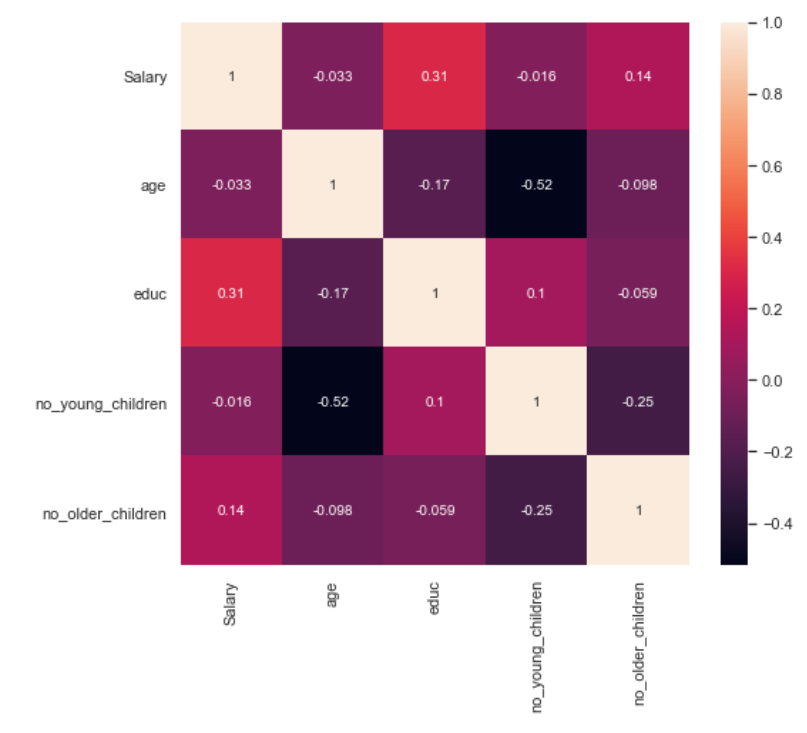
There is high chance for them to take the package if the employee salary is between 30k to 40 k. This suggest that package pricing is average pricing. There Is a higher chance of taking up the package if the employee age is between age of 25 of to 50 years. After 50 years there is a higher chance of telling no. If the employee has no young children then there is huge chance to tell yes.

#### Pairplot



From the above pairplot we can see that there is hardly any multicollinearity between the variables. We can confirm the same using the heatmap.

#### Correlation Heatmap

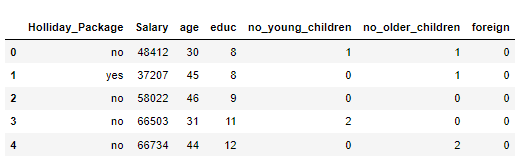


There is hardly any correlation between the variables. Salary and Educ have a positive correlation but they are not strong enough.

## **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).

**Encode data:**

Head of dataset after encoding



Foreign column is encoded to zero (if no) and one (if yes)

To perform Logistic Regression and LDA, the data is successfully split into train and test (70:30) and random state is 1

Data Apply Logistic Regression and Linear discriminant analysis

A.) Logistic Regression



B.) LDA Inference:



Data is encoded, split into training an testing and model for logistic regression and LDA is built

## **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

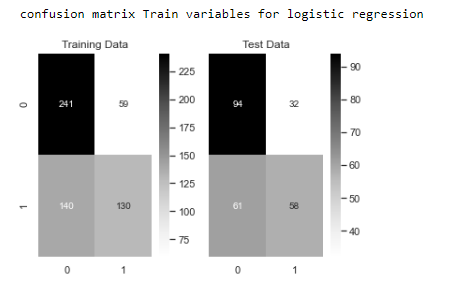
Accuracy score for Training data:



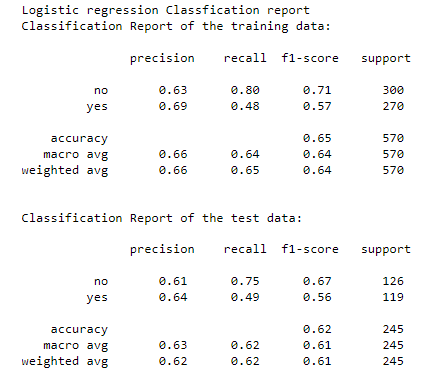
Accuracy score for Test Data:



Confusion Matrix for Logistic Regression:



Classification Report for Logistic Regression:



#### ROC curve and ROC\_AUC score for Logistic Regression

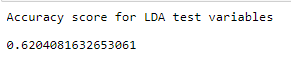


## Performance Metrics for LDA

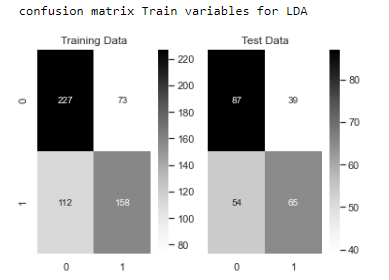
#### Accuracy score for Training data



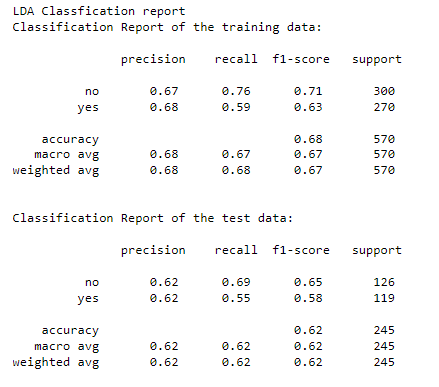
#### Accuracy score for Testing data



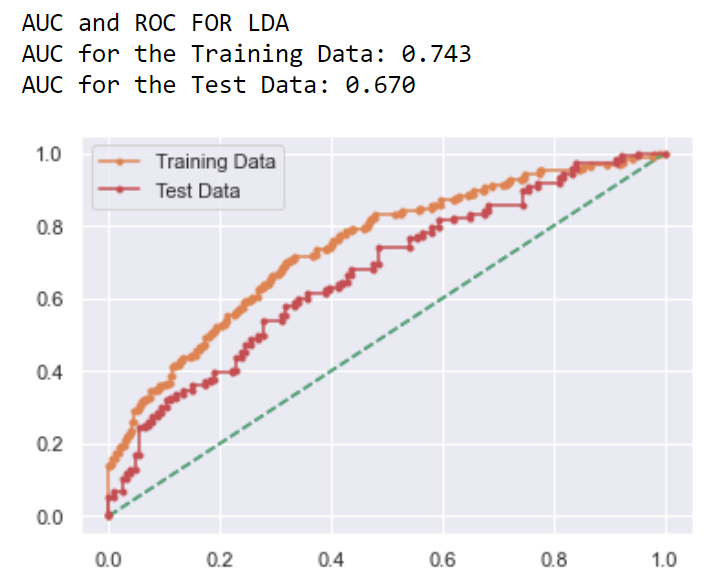
### Confusion Matrix For LDA



#### Classification Report For LDA

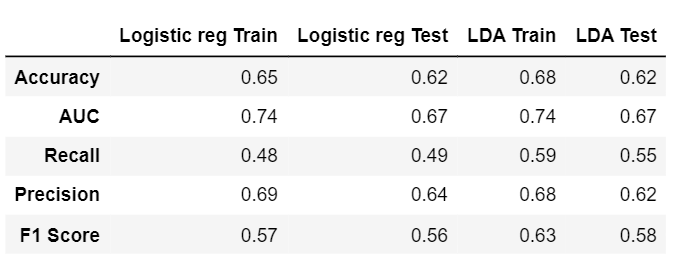


#### AUC and ROC for Linear Discriminant Analysis –



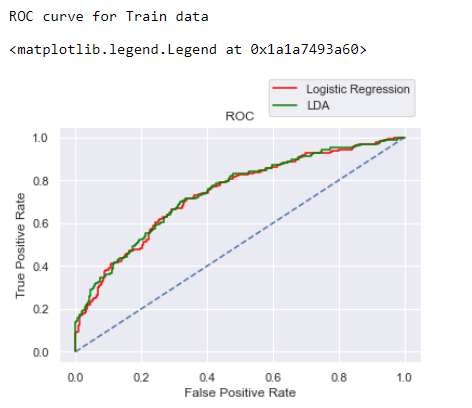
#### Performance Metrics for Final Model: Compare Both the models and write inference which model is best/optimized.

Comparison in Table form:

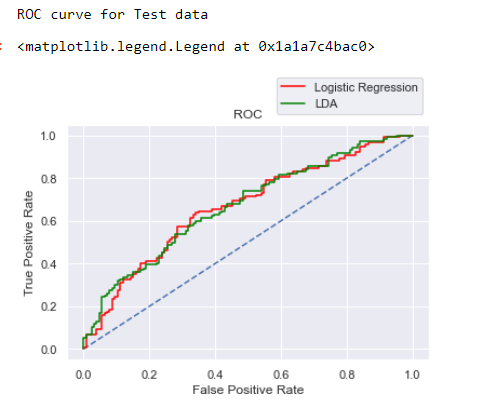


#### ROC curve and ROC\_AUC score comparison

For Train Data:



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.

### **2.5** Basis on these predictions, what are the insights and recommendations.

The Linear discriminant analysis model will be able to predicting whether an employee will opt for the package or not with around 70 percent accuracy.

Business Insights: The important factors which determine whether an employee will opt in for package are Salary, Age, no of young children and foreign. The company must focus on the people who earns between 30 to 40 k and between age 25 to 50 years and if they have no children, there is a huge chance for them to opt in for a package.

**Recommendation**: The greatest number of people who are opting in for the package has a salary of range between 30 to 40 k. It suggests that the package is of average price with medium level facilities. So, if they add some additional luxury packages with facilities like booking in star hotels, luxury cars etc. it may help to increase the sales of packages to a higher income group. The analysis shows that a greater number of foreigners opt in for packages than the non-foreigners. This along with the previous analysis which shows that most of the people are from salary group of 30 to 50k(so it is not expensive package ) suggest that packages provided are either of local sightseeing place or of less interest to the non-foreigners. So, suggest the company to add some more activities or places in their packages.

The analysis shows that data if the employee as no young children, there is more chance of them taking up the package. As count of children increases, the willingness to opt in for a holiday package decreases. So, I suggest the company to provide additional discounts or children attractiveness for the employee who has young children to boost up the chance of them opting in for the package.

------------Thanks------------