Business Report

Diamonds & Holidays

Data analysis of cubic zirconia and holiday package datasets

Linear regression, Logistic regression, LDA

Filed by: Aditya Rishi

For: Predictive modelling project

Date: August 1, 2021

INDEX

Tables

Charts

PROBLEM 1 :

LINEAR REGRESSION

Executive Summary

Cubic-zirconia manufacturer Gem Stones Company Limited hired us to look at a dataset of prices and other attributes of almost 27,000 specimens of cubic zirconia, an inexpensive diamond alternative with many of the same qualities as an actual precious stone. The company earns different profits on different prize slots. Our task is to help the company predict the price for the stone on the bases of the details given in the dataset, so the maker can distinguish between higher profitable and lower profitable stones so as to have a better profit share. The company also wants the 5 attributes that are most important.

Figure 1. Faceted cubic zirconia, the largest is 9 millimetres across. Courtesy of the Ceres Corporation, Waltham, Massachusetts, and MSB Industries, New York.

Introduction

Cubic zirconia (CZ) is the cubic crystalline form of zirconium dioxide (ZrO_2). The synthesized material is hard and colourless usually, but may be made in a variety of different colours. It should not be confused with zircon, which is a zirconium silicate ($ZrSiO_4$). It is sometimes called cubic zirconium erroneously.

Soon after it was first marketed in 1976, colorless cubic zirconia became the dominant diamond imitation, with current production of approximately 60 million carats per year. Although cubic zirconia was discovered as a natural mineral in 1937, crystals usable for faceting were first produced ill 1969 and it was not until a practical sltull-melting technique was developed in the USSR in1972 that commercial production became feasible. Cubic zirconia was discovered as a natural mineral in 1937, when two German mineralogists, von Staclzelberg and Chudoba (1937), were examining a highly metamict zircon given to them by B. W. Anderson. The zircon contained some tiny crystals which they identified by X-ray diffraction as the cubic form of zirconium oxide (or zirconia), a compound known as baddeleyite when in the monoclinic form.

So little did von Staclzelberg and Chudoba think of this discovery that they did not even assign a name to the new mineral. As a result, it is known to this day by its scientific name, cubic zirconia, and the prefix synthetic, although proper, is not included usually.

This same material had already been used for many years as a ceramic composition for high-temperature industrial and scientific purposes; because of an exceptionally high melting point, "stabilized zirconia" ceramics can be used at temperatures up to 2540°C (4604°F) and are very resistant to most chemical substances. Typically, this kind of stabilized zirconia consists of 96% ZrOs (zirconia) and 4% CaO (lime), although MgO (magnesia) or Y203 (yttria) also can be used in place of the CaO.

It may not be all that soon that a potential successor to cubic zirconia will arrive. The optical constants of cubic zirconia are sufficiently close to diamond in a material of adequate hardness and wearability that a large improvement cannot be expected. This was hardly true of any of the previous diamond imitations. Second, the cost involved in developing and marketing a new synthetic or imitation is never small and, given the existence of a highly satisfactory material in the marketplace, may not be justifiable in terms of the potential returns.

Figure 2. Coloured cubic zirconia, rough, the largest piece is 2.5 inches (6.5 cm) long. Courtesy of the Ceres Corporation, Waltham, MA.

Distinction from Diamond

The distinction is obvious to the trained eye. In a loose stone, the high specific gravity is apparent. Flatness of faces and sharpness of edges are not fool-proof criteria, and girdles apparently showing "naturals" have been observed on this stone.

Size, Weight, and Shape

Cubic zirconia stones may be sold by size, by weight, or by equivalent diamond weight (the last is not always so specified).

Linear regression

The purpose of a regression model is to determine a linear function between the X and Y variables that best describes the relationship between the two variables. In linear regression, it's assumed that Y can be calculated from some combination of the input variables. The relationship between the input variables (X) and the target variables (Y) can be portrayed by drawing a line through the points in the graph. The line represents the function that best describes the relationship between X and Y (for example, for every time X increases by 3, Y increases by 2).
The goal is to find an optimal "regression line", or the line/function that best fits the data.

Lines are typically represented by the equation: Y = m*X + b. X refers to the dependent variable while Y is the independent variable.

Meanwhile, m is the slope of the line, as defined by the "rise" over the "run". Machine learning practitioners represent the famous slope-line equation a little differently, using this equation instead:

y(x) = w0 + w1 * x

In this equation, y is the target variable while "w" is the model's parameters and the input is "x". So, the equation is read as: "The function that gives Y, depending on X, is equal to the parameters of the model multiplied by the features". The parameters of the model are adjusted during training to get the best-fit regression line.

Ordinary least squares linear regression using sklearn.linear_model.LinearRegression

sci-kit learn library's LinearRegression function in Python fits a linear model with coefficients w = (w1, …, wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.

A regression can also be done with multiple features. In the case of "multiple linear regression", the equation is extended by the number of variables found within the dataset. In other words, while the equation for regular linear regression is y(x) = w0 + w1 * x, the equation for multiple linear regression would be y(x) = w0 + w1x1 plus the weights and inputs for the various features. If we represent the total number of weights and features as w(n)x(n), then we could represent the formula like this:

y(x) = w0 + w1x1 + w2x2 + … + w(n)x(n)

A cost function is used to measure how close the assumed Y values are to the actual Y values when given a particular weight value. The cost function for linear regression is mean squared error, which just takes the average (squared) error between the predicted value and the true value for all of the various data points in the dataset. The cost function is used to calculate a cost, which captures the difference between the predicted target value and the true target value. If the fit line is far from the data points, the cost will be higher, while the cost will become smaller the closer the line gets to capturing the true relationships between variables. The weights of the model are then adjusted until the weight configuration that produces the smallest amount of error is found.

Things to remember

1. "Regression" predicts a real number, generally.

2. The method models data with linear combination of the explanatory variables, hence called "linear".

3. A linear combination is an expression where one or more variables are scaled by a constant factor and added together.

4. In the case of simplest linear regression with a single explanatory variable, the linear combination used in linear regression can be expressed as:

Dependent variable value = (weight ∗ independent variable) + constant

5. It is the straight line in the scatter plot of the variables

Data Dictionary:	
Variable Name	**Description**
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 best and J the worst.
Clarity	cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, FL = flawless, I1= level 1 inclusion) IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1
Depth	The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.
Table	The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
Price	the Price of the cubic zirconia.
X	Length of the cubic zirconia in mm.
Y	Width of the cubic zirconia in mm.
Z	Height of the cubic zirconia in mm.

1.1. EXPLORATORY DATA ANALYSIS

•The dataset contains 26,967 observations and ten variables, including the dependent variable 'price' of the zirconia stones

• There are three categorical ordinal variables (cut, color and clarity), which represent the respective quality of the stone from lower to higher grade

• All other six variables are continuous numeric types, where x, y and z are cubic stones' three dimensions

• Statistical summary shows average price of a zirconia stone to be about 3,940, and a price range of 326 to 18,818, showing outliers and a skewed distribution

• There are 697 null values in the variable 'depth' which are imputed with the mean of the variable

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
0	1	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	2	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
2	3	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	4	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
4	5	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779
5	6	1.02	Ideal	D	VS2	61.5	56.0	6.46	6.49	3.99	9502
6	7	1.01	Good	H	SI1	63.7	60.0	6.35	6.30	4.03	4836
7	8	0.50	Premium	E	SI1	61.5	62.0	5.09	5.06	3.12	1415
8	9	1.21	Good	H	SI1	63.8	64.0	6.72	6.63	4.26	5407
9	10	0.35	Ideal	F	VS2	60.5	57.0	4.52	4.60	2.76	706

Table 1: The cubic zirconia dataset

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

Table 2: Data info

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
carat	26967	NaN	NaN	NaN	0.798375	0.477745	0.2	0.4	0.7	1.05	4.5
cut	26967	5	Ideal	10816	NaN	NaN	NaN	NaN	NaN	NaN	NaN
color	26967	7	G	5661	NaN	NaN	NaN	NaN	NaN	NaN	NaN
clarity	26967	8	SI1	6571	NaN	NaN	NaN	NaN	NaN	NaN	NaN
depth	26270	NaN	NaN	NaN	61.7451	1.41286	50.8	61	61.8	62.5	73.6
table	26967	NaN	NaN	NaN	57.4561	2.23207	49	56	57	59	79
x	26967	NaN	NaN	NaN	5.72985	1.12852	0	4.71	5.69	6.55	10.23
y	26967	NaN	NaN	NaN	5.73357	1.16606	0	4.71	5.71	6.54	58.9
z	26967	NaN	NaN	NaN	3.53806	0.720624	0	2.9	3.52	4.04	31.8
price	26967	NaN	NaN	NaN	3939.52	4024.86	326	945	2375	5360	18818

Table 3: Data description, five-point summary

• The dimension variables, x, y & z, shows a min value of zero, which are dropped as shape of the stone can't be zero and there are only 3 such observations.

• The categorical quality variables have a proportion as shown in the pie charts. There are 10,816 stones in the 'Ideal' category, which is of the highest grade, 5661 stones are graded 'G' in color quality and 6571 stones are observed as 'SI1' in clarity grading.

• All the continuous variables have significant number of outliers and different scales of values. Outliers are removed so these don't pull the regression line towards them.

```
CUT :  5
Fair            781
Good           2441
Very Good      6030
Premium        6899
Ideal         10816
Name: cut, dtype: int64


COLOR :  7
J    1443
I    2771
D    3344
H    4102
F    4729
E    4917
G    5661
Name: color, dtype: int64


CLARITY :  8
I1        365
IF        894
VVS1     1839
VVS2     2531
VS1      4093
SI2      4575
VS2      6099
SI1      6571
Name: clarity, dtype: int64
```

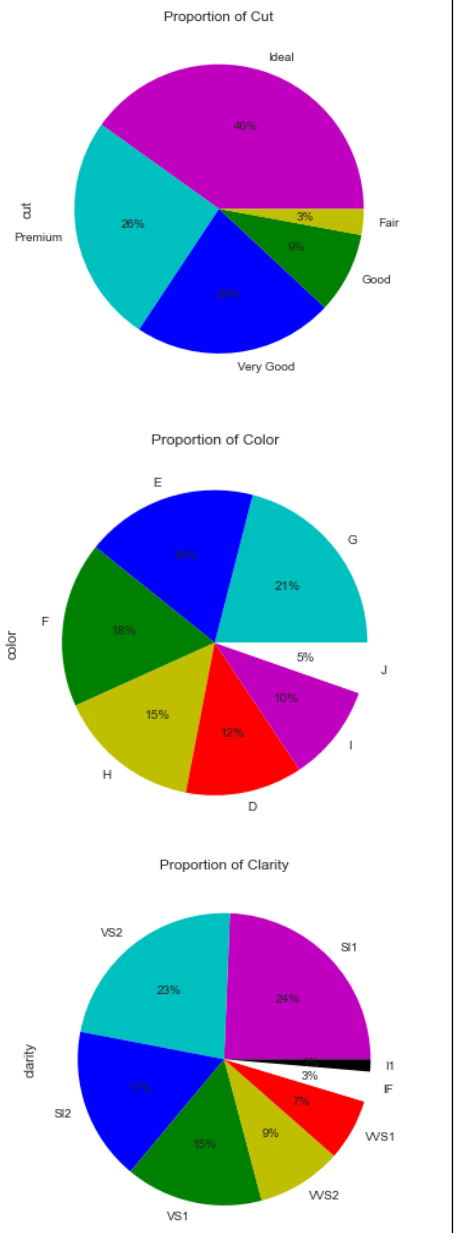


Table 4, Figure 3: Count, categories, proportions in categorical variables

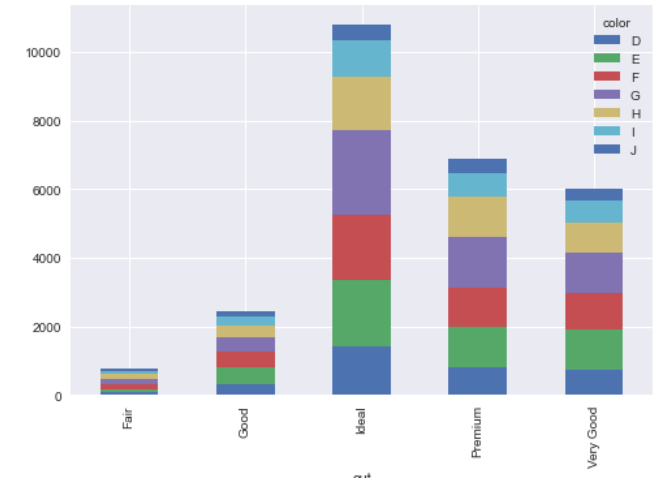


Figure 4, Table 5: Crosstab to show proportion of colour in every cut

color	D	E	F	G	H	I	J
cut							
Fair	74	100	148	147	150	94	68
Good	311	491	454	419	352	253	161
Ideal	1409	1966	1893	2470	1552	1073	453
Premium	808	1174	1167	1471	1161	711	407
Very Good	742	1186	1067	1154	887	640	354

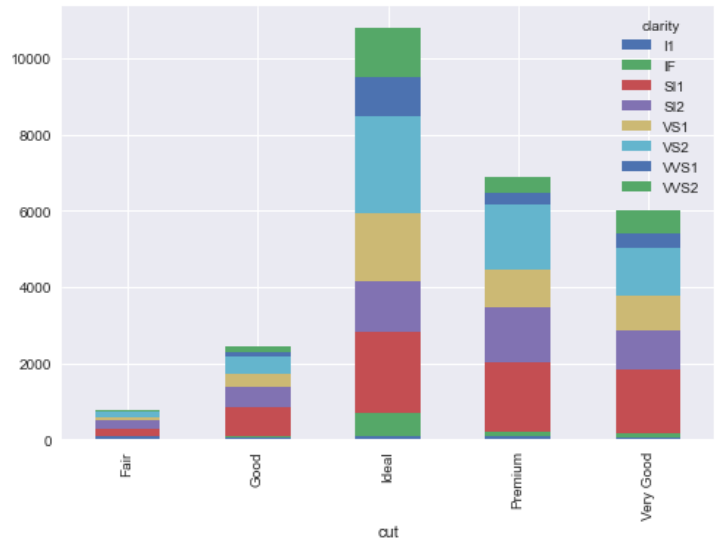


Figure 5, Table 6: Crosstab to show proportion of clarity in every cut

clarity	I1	IF	SI1	SI2	VS1	VS2	VVS1	VVS2
cut								
Fair	89	4	193	225	93	129	10	38
Good	51	30	765	530	331	491	100	143
Ideal	74	613	2150	1324	1784	2528	1036	1307
Premium	108	115	1809	1449	998	1697	307	416
Very Good	43	132	1654	1047	887	1254	386	627

Analysis of the target/dependent/predicted variable

Figure 6: Price with outliers

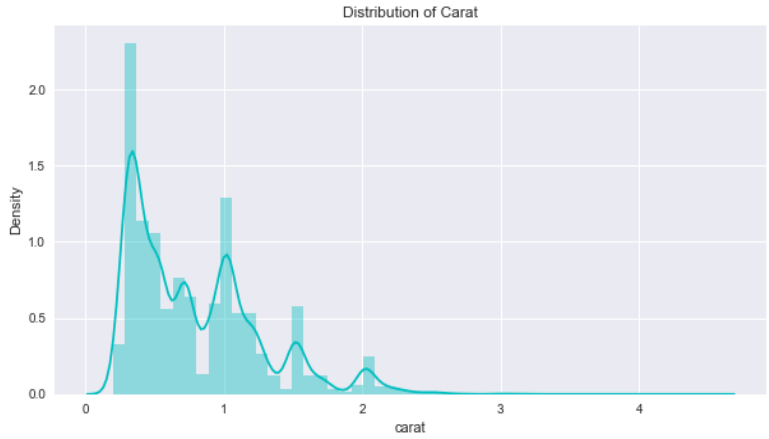
Figure 7: Distribution of price

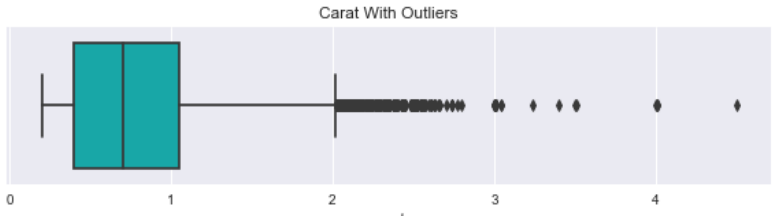
From further plots, we can infer that the response variable 'price' has a collinear relationship with variables such as 'carat', 'x', 'y' and 'z', whereas there is no significant relation seen with 'depth' and 'table'

• From the heatmap of Pearson's correlation of the continuous variables, we can infer that the response variable 'price' is highly correlated with 'carat' and the dimension variables (x, y & z)

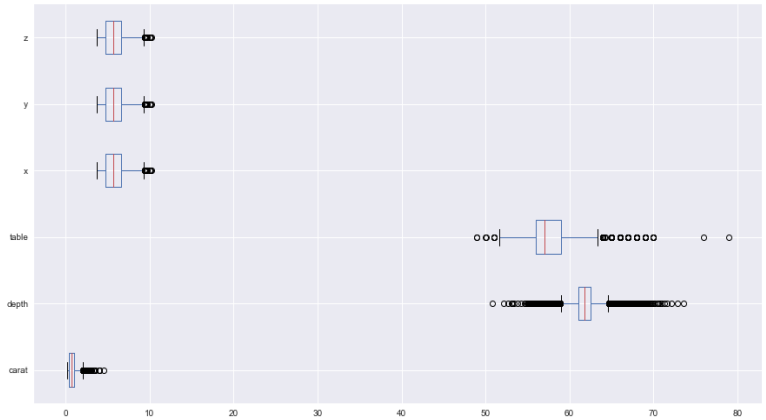
• From the pairplot and the heatmap we can also see the collinearity of 'carat' is very high with the dimension variables, which indicates that 'carat' is influenced by the size of the stone

Predictor/independent variables





Outliers



Outliers removed

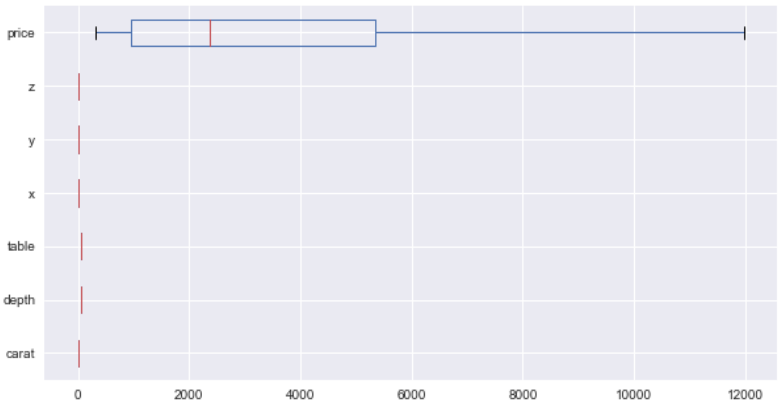
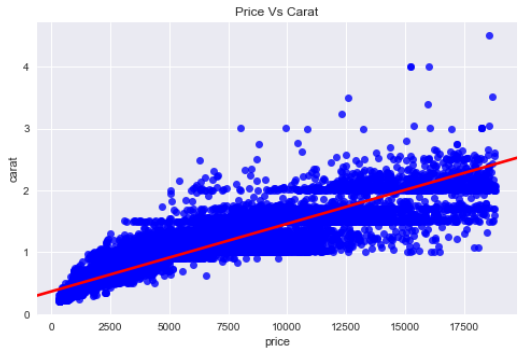
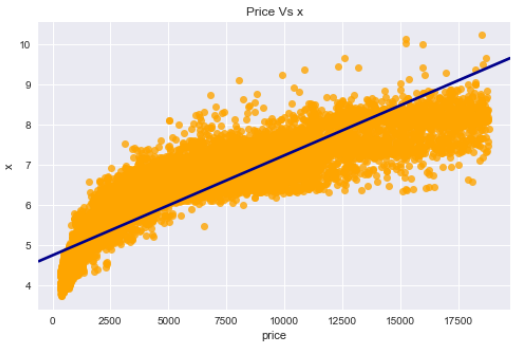


Fig 8, 9, 10, 11: Carat distribution, boxplot, outliers, treated

Before removing outliers, we checked the linear distribution of price against independent variables. (Figure 12 to 16)





From the pairplot, we can see the distribution of the continuous variables and the collinearity between them • The collinearity and correlation of x, y and z, which indicate that for a given observation the values of x, y and z will be same, which is because the stones are cubic in shape

• We can retain one of the dimension variable and drop the rest for training the model. We will use x and volume of the stone in iteration 2
• Outliers are removed from the dataset shown in boxplot. Boxplots of continuous variables after removing outliers were also made

• After imputing the null values and removing zero values in dimensions, there are 47 duplicate records which can be dropped, since the number is insignificant considering the total number of observations

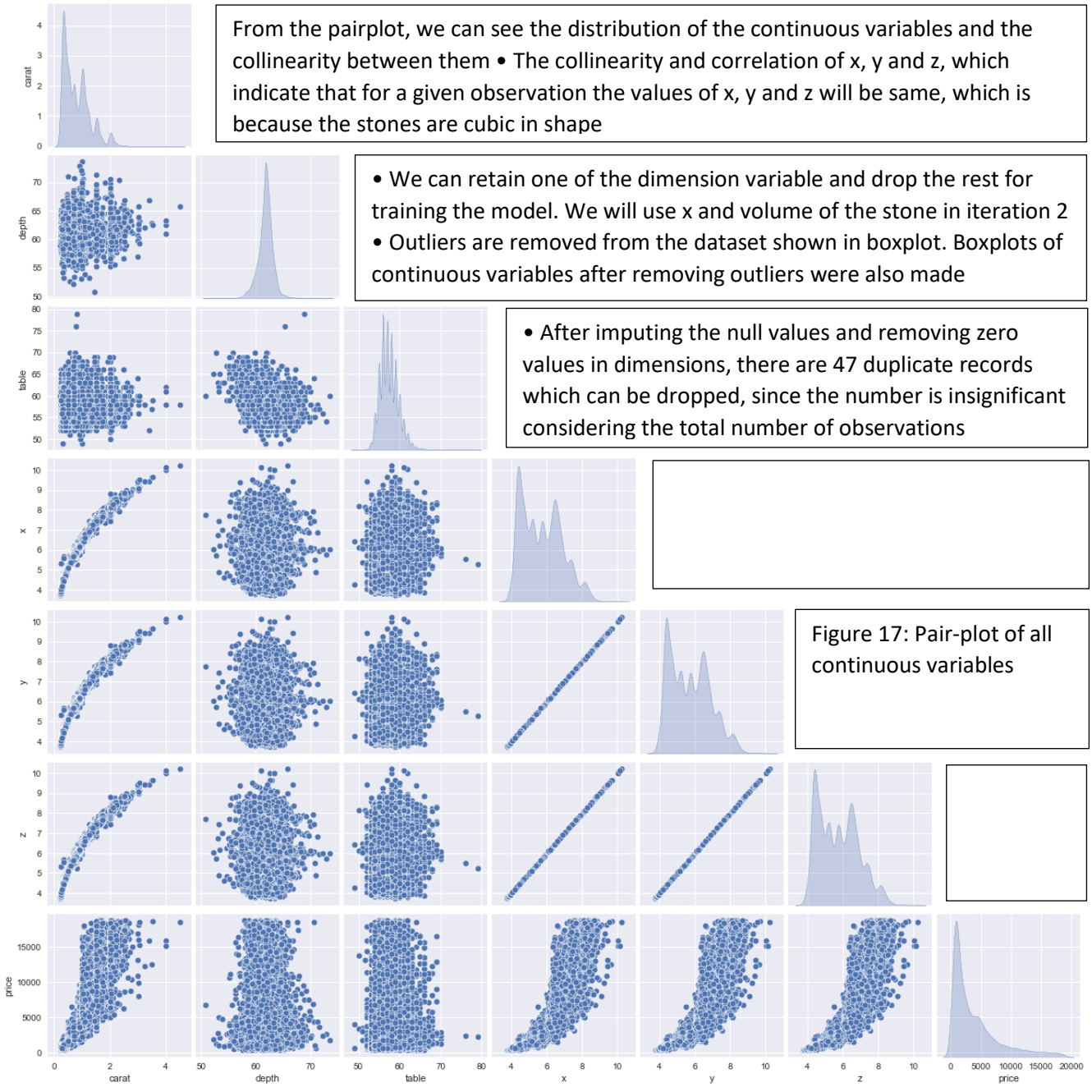


Figure 17: Pair-plot of all continuous variables

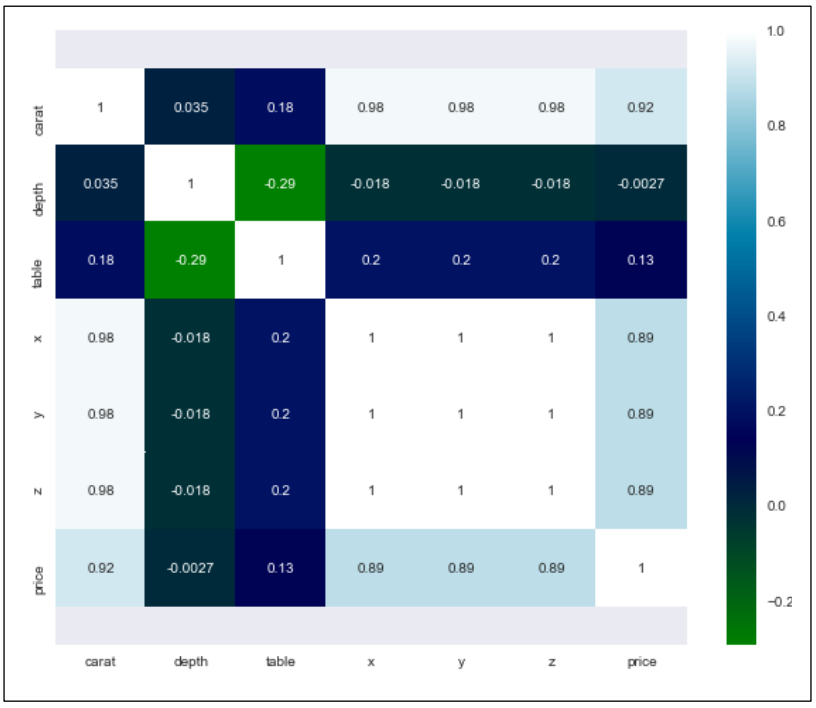


Fig 18: Heatmap. Table 7: Minimum mean and maximum price based on cut of cubic zirconia

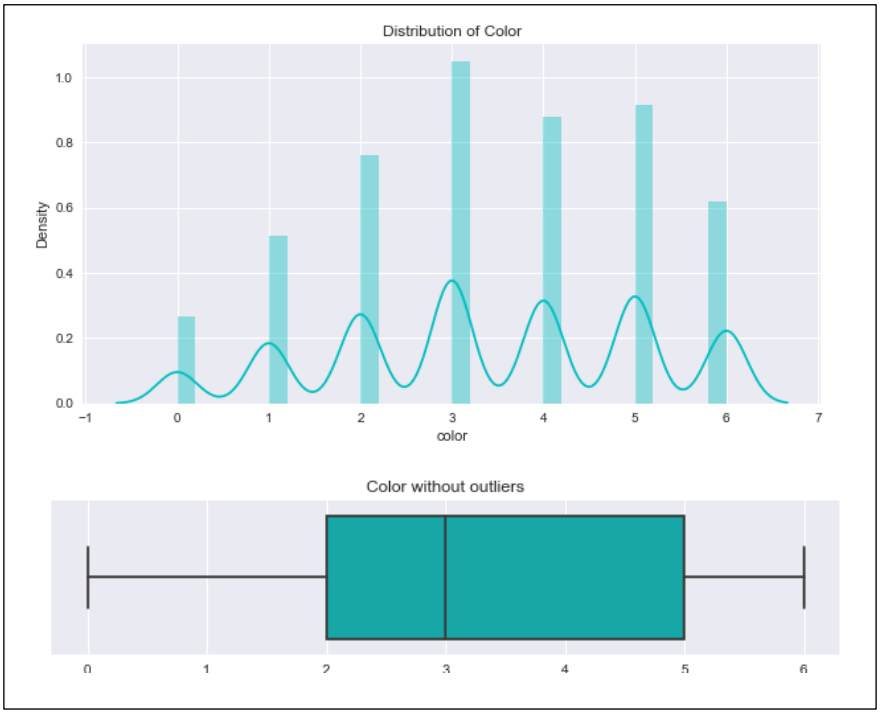
	price		
	min	mean	max
cut			
0	369	4565.768935	18574
1	335	3927.074774	18707
2	336	4032.581812	18818
3	326	4545.127814	18795
4	326	3457.579264	18804

color	price		
	min	mean	max
0	335	5329.706250	18701
1	336	5124.816637	18795
2	337	4478.887152	18795
3	361	4008.067493	18818
4	357	3701.517585	18791
5	326	3074.382706	18731
6	357	3186.841222	18526

clarity	price		
	min	mean	max
2	345	3908.750000	18531
3	326	5090.164182	18804
4	326	3998.861974	18818
5	357	3968.153909	18791
6	338	3840.991426	18795
7	336	3264.062475	18718
8	336	2502.874388	18445
9	369	2743.687289	18552

Table 8, 9: Minimum mean and maximum price based on color and clarity of cubic zirconia

More distributions of independent variables



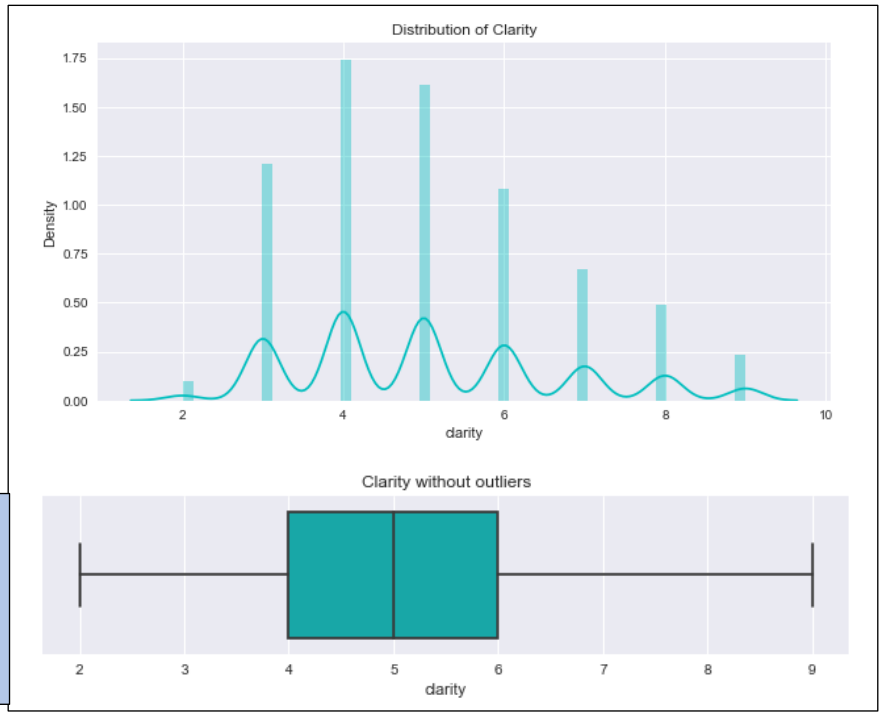
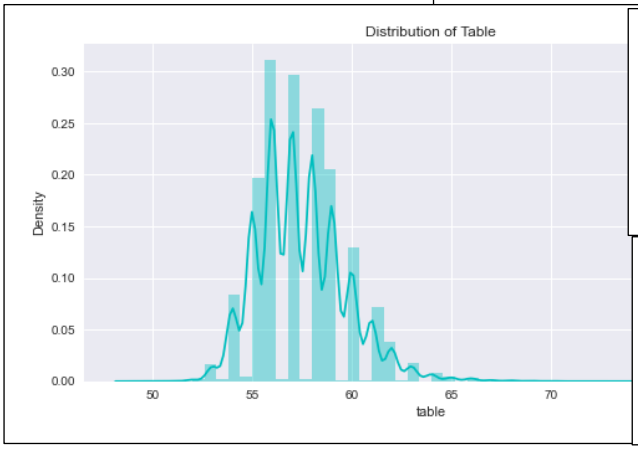


Figure 19 to 25: Distribution of colour, clarity, and table



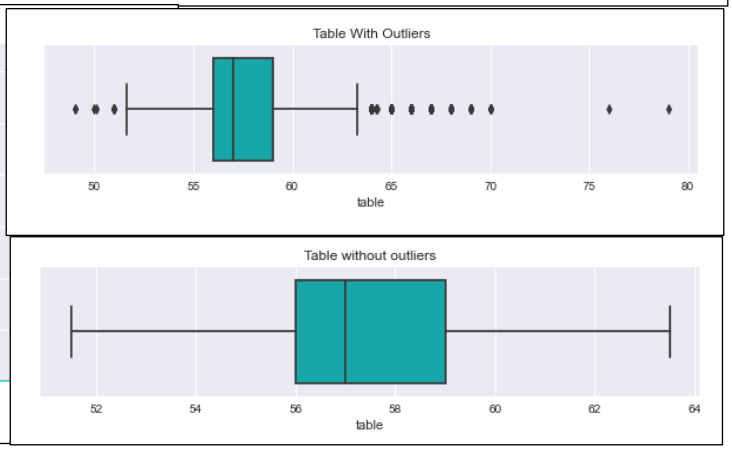
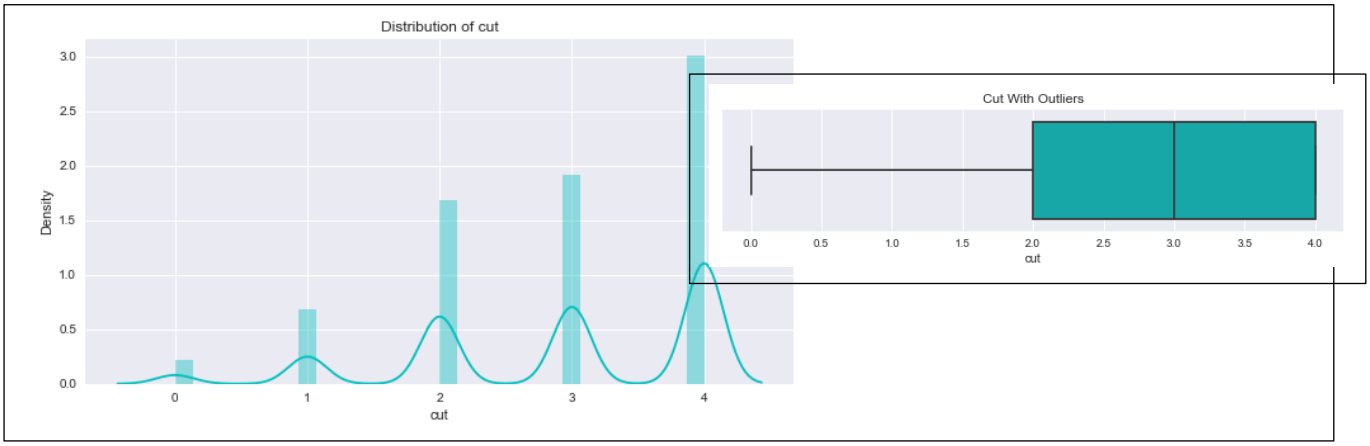
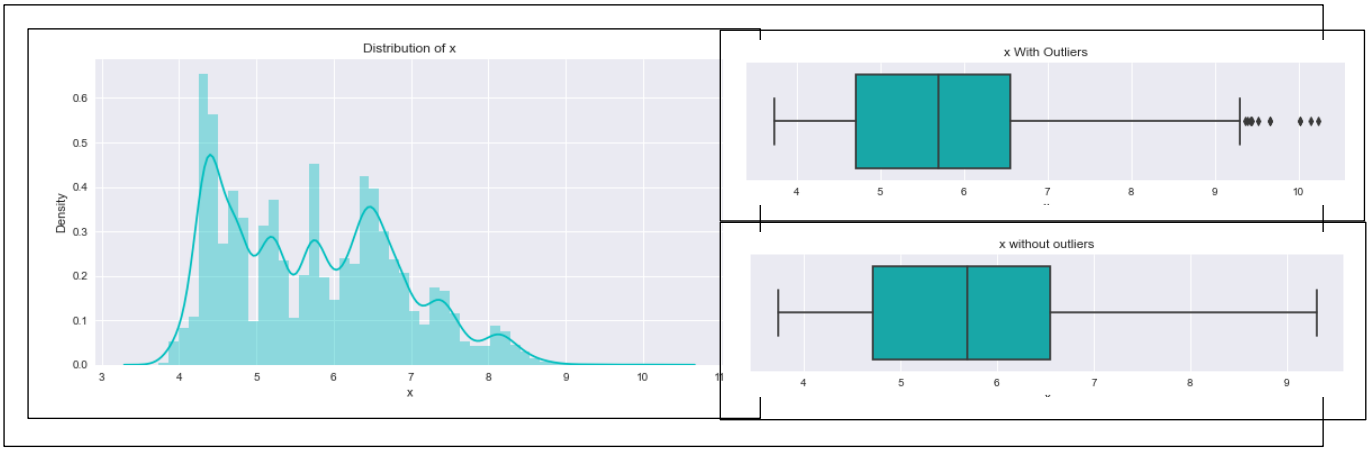
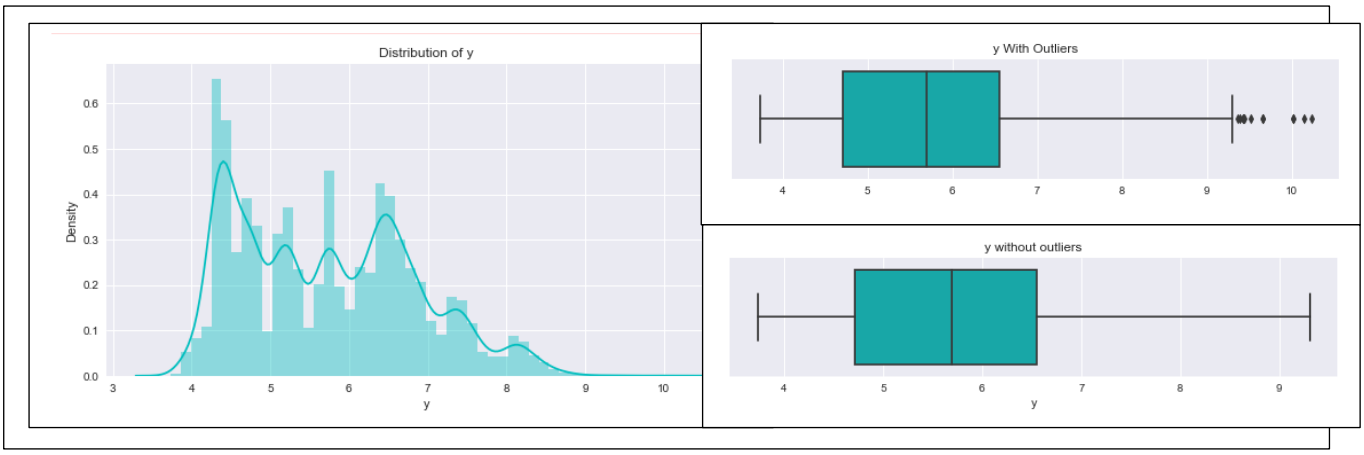
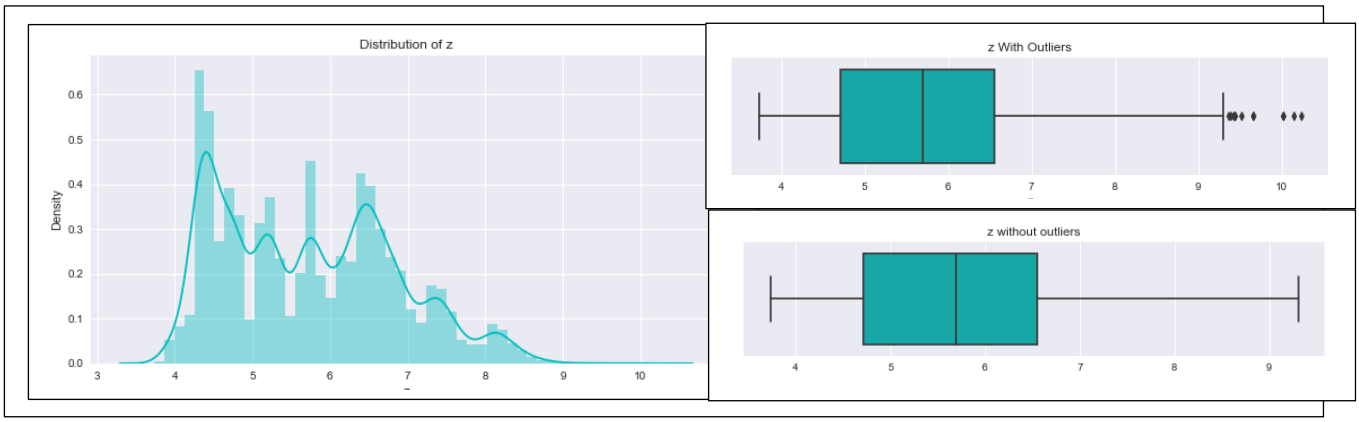


Figure 26 to 35: Distributions of cut, x, y, and z









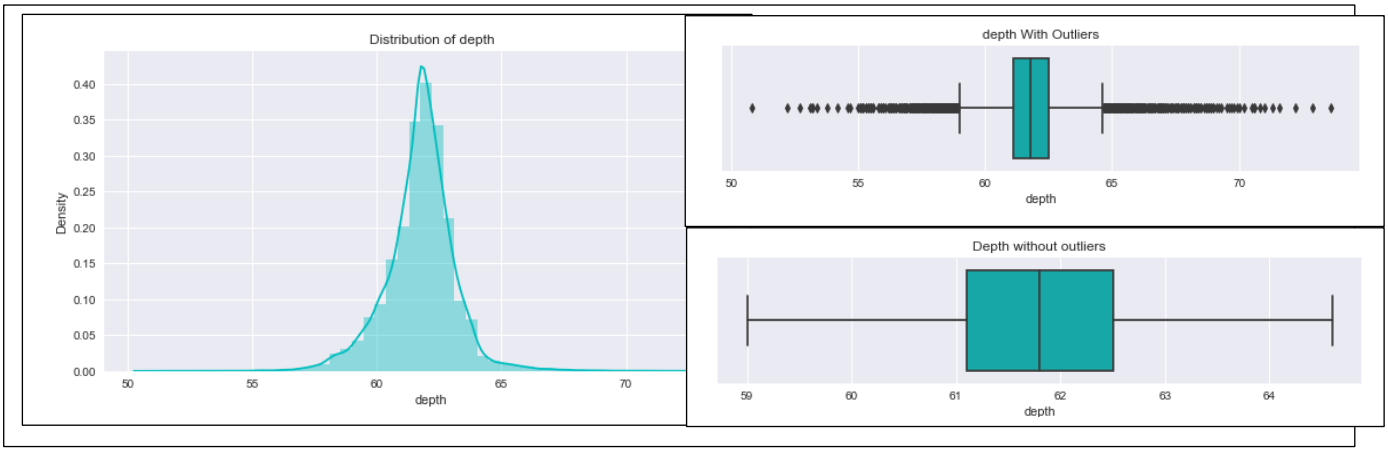
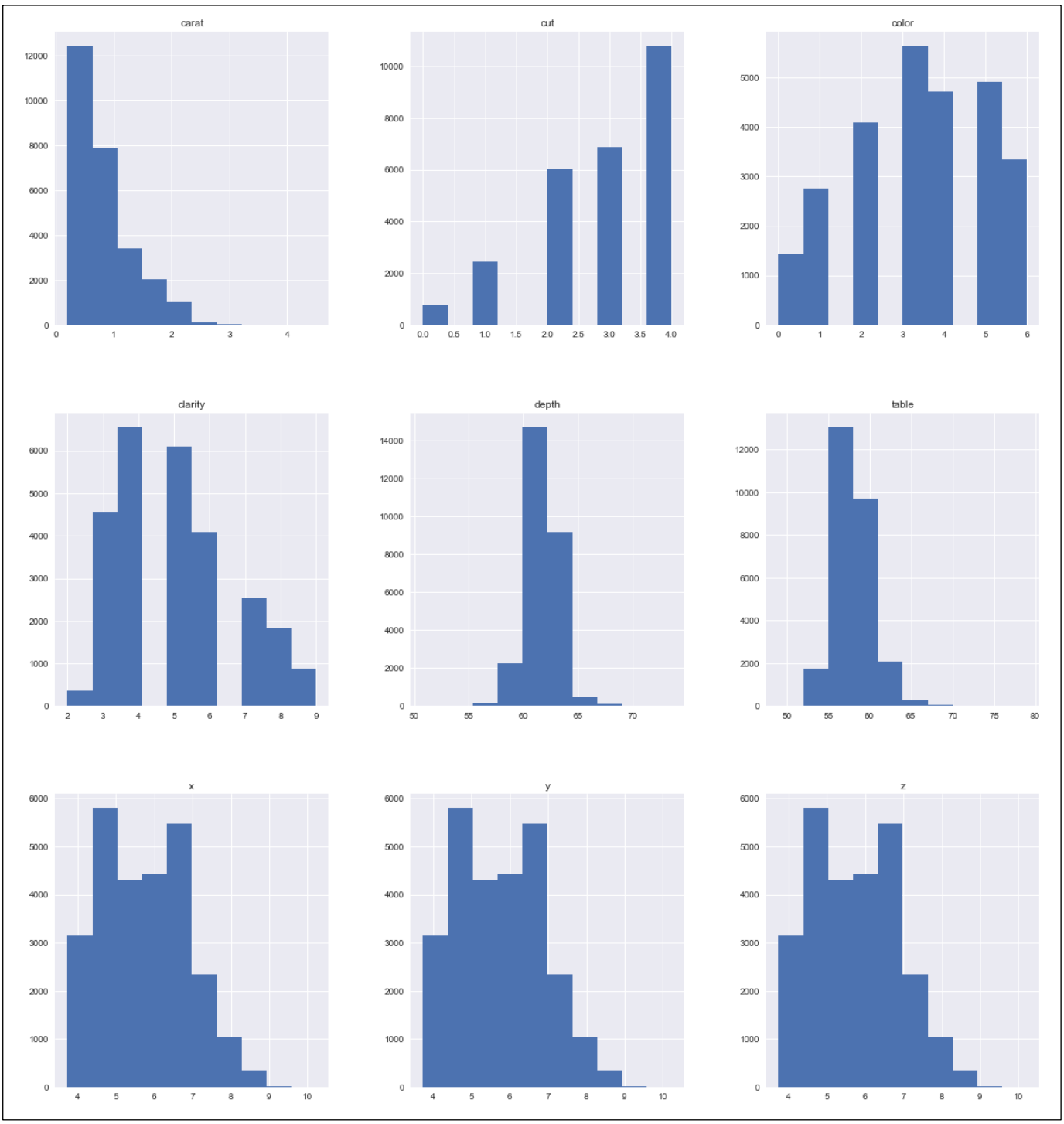


Figure 36 to 38: Distribution of depth. Figure 39 (Below) Histograms of 9 variables



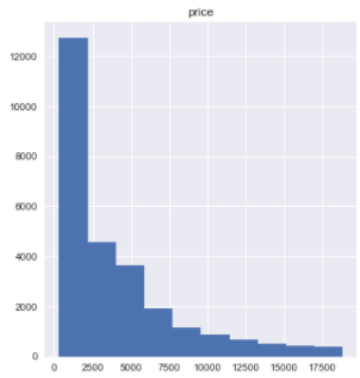
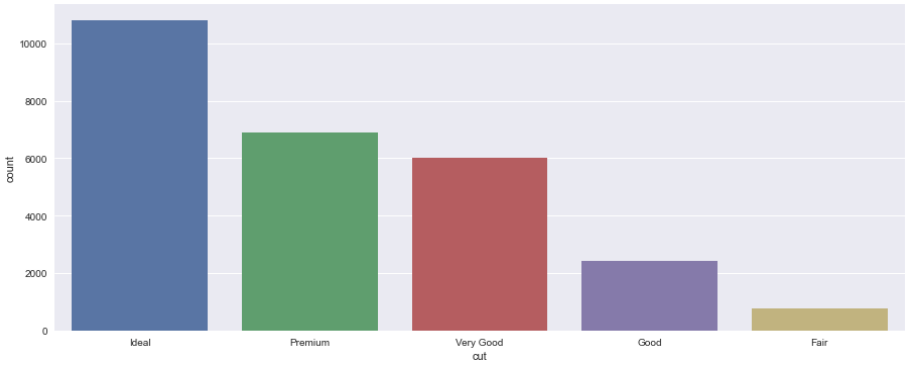


Figure 40: histogram for price: Highly skewed



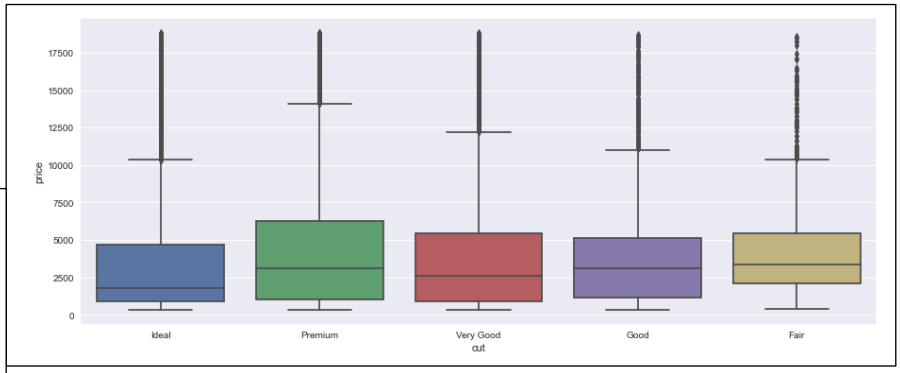
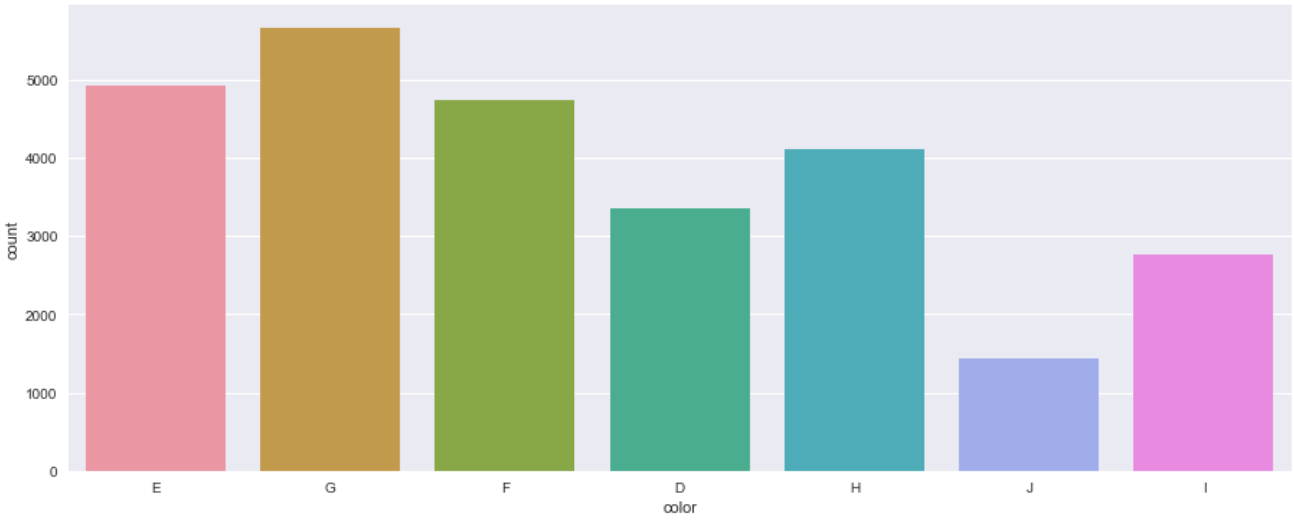


Figure 41 and 42: Observations on Cut. The premium-cut diamonds are the most expensive, followed by those with very good cut.



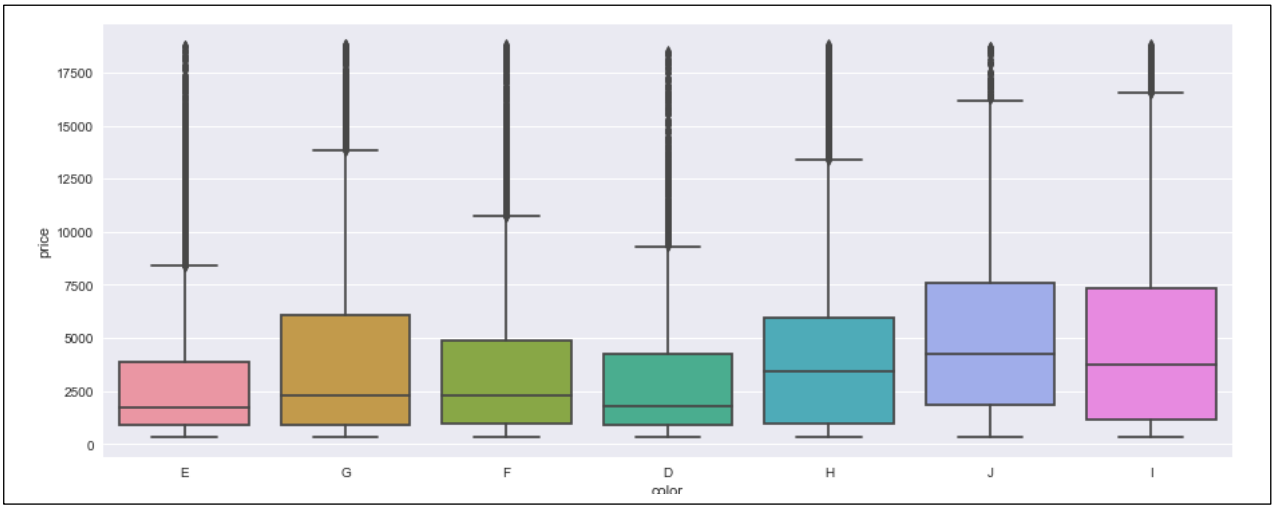
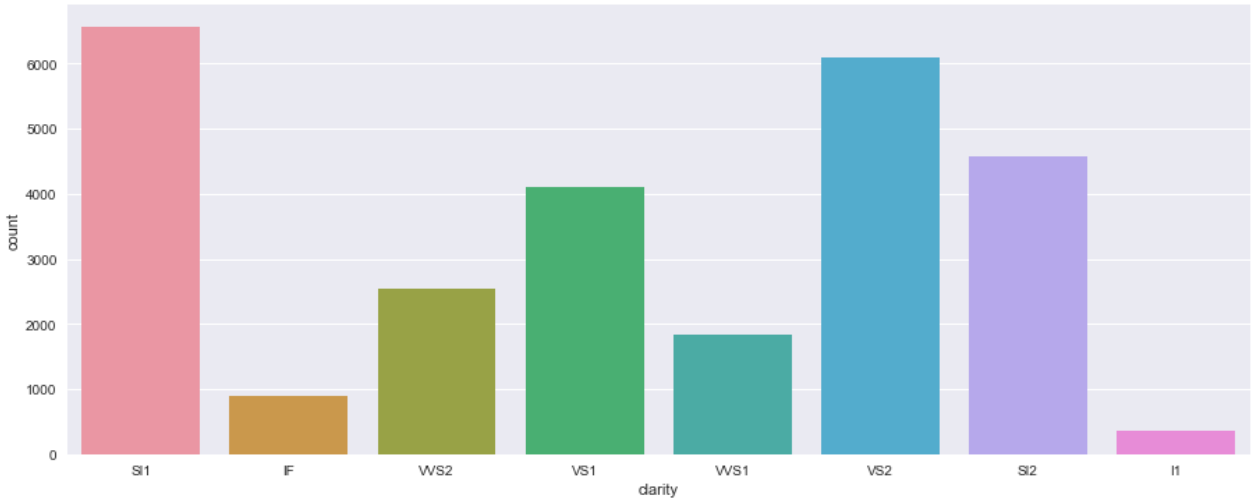


Figure 43 and 44: Observations on color. The ‘J’ or worst-color diamonds are the most expensive, followed by ‘I’, suggesting some industrial usage.



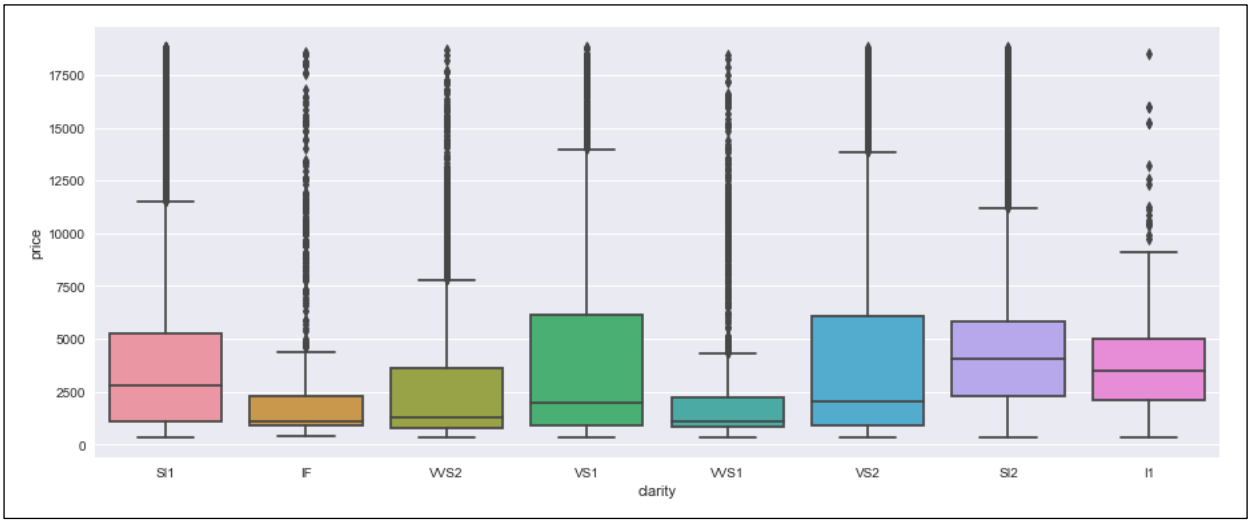


Figure 45 and 46: Observations on clarity. The diamonds with VS1 & VS2 clarity are the most expensive.

The inferences drawn from the exploratory data analysis or EDA:-

Observation-1:
(1).'Price' is the target variable while all others are the predictors. (2). The data set contains 26967 rows and 11 columns. (3). In the given dataset, there are 2 integer-type features 6 float-type features, and 3 object-type features, where 'price' is the target variable and all other are predictor variable. (4). The first column ("Unnamed: 0") is an index of some sort or only a bunch of serial numbers, so we removed it.

Observation-2: (1). In the given dataset, the mean and median values do not differ a lot. (2). We can observe the minimum values of "x", "y", "z" to be zero. These are faulty values, as dimensionless or 2-dimensional diamonds are not possible. So, we have to filter out those obvious faulty data entries. (3). There are three object data types: 'cut', 'color' and 'clarity'.

Observation-3: There are 697 missing values in the depth column. There are some duplicate rows (34 out of 26,958 or 0.12 % of the total data), so in this case, we can drop the duplicated rows safely.

Observation-4: There are significant number of outliers, points that are far from the rest of the dataset, which will affect the outcome of our regression model. So, we have to treat the outliers. The distribution of some quantitative features such as "carat" and the target feature "price" are "right-skewed" heavily.

Observation-5: Most features seem to correlate with the price of the diamond. The notable exception is "depth", which has a negligble correlation (~1%)

1.2. DATA PRE-PROCESSING

• **Null values:** There are 697 null value observations in the variable 'depth' which are imputed with the mean of the variable

• **Zero values:** The dimension variables, x, y & z shows a min value of zero, which are dropped as shape of the stone can't be zero and there are only 3 such observations

• **Ordinal Encoding:** All the three categoric ordinal variables such as cut, color and clarity, represents the respective quality of the stone from lower to higher order

• Ordinal encoding is applied in this case as the observation with higher quality category must take precedence over the lower ones in training the model

• The variables are encoded from zero to higher values in the increasing order of the respective quality attribute

• **Scaling the data:** The different factors in the given dataset is in different scales, which can potentially be standardized.

• We are **going to use two methods** such as LinearRegression function from sklearn and OLS function from statsmodels in this business case. Both of which use 'ordinary least square' method which may not be influenced by scaling • Scaling would be effective if the algorithm used gradient descent method to converge faster

• Scaling using Z score has been applied in this business case to evaluate the impact of scaling and **no significant improvement is found with scaling**

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
carat	26967	NaN	NaN	NaN	0.798375	0.477745	0.2	0.4	0.7	1.05	4.5
cut	26967	5	Ideal	10816	NaN	NaN	NaN	NaN	NaN	NaN	NaN
color	26967	7	G	5661	NaN	NaN	NaN	NaN	NaN	NaN	NaN
clarity	26967	8	SI1	6571	NaN	NaN	NaN	NaN	NaN	NaN	NaN
depth	26270	NaN	NaN	NaN	61.7451	1.41286	50.8	61	61.8	62.5	73.6
table	26967	NaN	NaN	NaN	57.4561	2.23207	49	56	57	59	79
x	26967	NaN	NaN	NaN	5.72985	1.12852	0	4.71	5.69	6.55	10.23
y	26967	NaN	NaN	NaN	5.73357	1.16606	0	4.71	5.71	6.54	58.9
z	26967	NaN	NaN	NaN	3.53806	0.720624	0	2.9	3.52	4.04	31.8
price	26967	NaN	NaN	NaN	3939.52	4024.86	326	945	2375	5360	18818

Table 10: Zero values of dimension variables x, y & z

```
carat        0
cut          0
color        0
clarity      0
depth      697
table        0
x            0
y            0
z            0
price        0
dtype: int64
```

Table 11: Null values in variable 'depth'

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

Table 12: Null values imputed with mean

```
Number of duplicate rows = 34
```

```
Shape before replacing zeros and removing all missing values:  (26967, 10
Shape after replacing zeros and removing all missing values:  (26964, 10)
```

```
Number of duplicate rows = 47
```

```
Shape before removing duplicate values (26964, 10)
Shape after removing duplicate values (26917, 10)
```

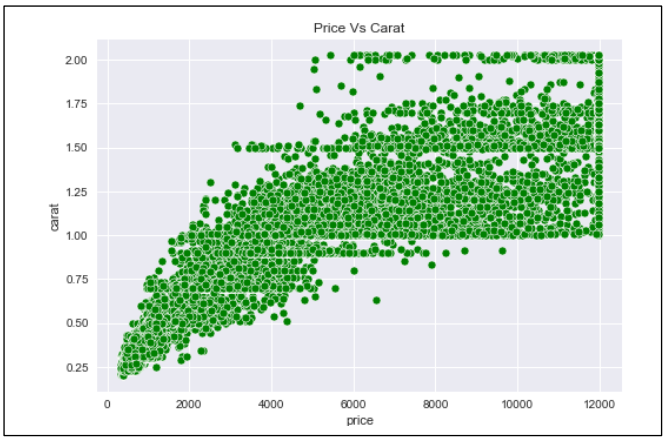


Figure 47: 'Price vs carat' scatter plot

Figure 48: 'Price vs x' scatter plot

	count	mean	std	min	25%	50%	75%	max
carat	18841.0	-1.394305e-16	1.000027	-1.279710	-0.848277	-0.201127	0.553881	2.657119
cut	18841.0	-9.571335e-16	1.000027	-2.620303	-0.821169	0.078398	0.977964	0.977964
color	18841.0	-3.684166e-16	1.000027	-1.995467	-0.820537	-0.233072	0.941859	1.529324
clarity	18841.0	1.387116e-16	1.000027	-1.860740	-0.638135	-0.026832	0.584470	2.418378
x	18841.0	-4.254922e-16	1.000027	-1.771287	-0.902872	-0.034457	0.727622	3.173362
volume	18841.0	-2.678065e-16	1.000027	-1.261872	-0.843421	-0.209022	0.561104	4.745797
price	18841.0	-2.006722e-16	1.000027	-0.980376	-0.802275	-0.392271	0.463993	2.372026

Table 13: Dataset after scaling using Z score. This is just for experiment, otherwise scaling not necessary

The necessity and futility of scaling in regression

Scaling or standardizing the features around the centre with mean equal to 0 and standard deviation equal to 1 is important when we compare measurements that have different units. Variables that are measured at different scales do not contribute equally to the analysis and might end up creating a bias.

For example, A variable that ranges between 0 and 1000 will outweigh a variable that ranges between 0 and 1. Using these variables without standardization will give the variable with the larger range a weight of 1,000 in the analysis. Transforming the data to comparable scales can prevent this problem.

In this dataset, all the variable are on different scales. Price is in 1000s of dollars and depth and table are in 100s unit, while carat is in 10s. So, one is tempted to scale or standardise the data to allow each variable to be compared on a common scale.

With data measured in different "units" or on different scales (as here with different means and variances), scaling is an important data processing step if the results are to be meaningful or not dominated by the variables that have large variances.

But scaling in this case is not necessary. We'll get an equivalent solution whether we apply some kind of linear scaling or not. It's demonstrated. But in regression techniques, it can help gradient descent to converge faster and reach the global minima. When the features are too many, it helps run the model quickly, else the starting point would be very far from minima, if the scaling is not done in pre-processing. So, it's work best on gradient descent model, while here we use OLS (ordinary least squares).

So, linear transformations of the response is never necessary but it may, however, aid us in interpreting our model.

1.3. LINEAR REGRESSION MODELLING

• Ordinal Encoding: All the three categorical ordinal variables such as cut, color and clarity, represents the respective quality of the stone from lower to higher order

• Ordinal encoding is applied in this case as the observation with higher quality category must take precedence over the lower ones

• The variables are encoded from zero to higher values in the increasing order of quality attribute

• Split data: The data was split in 70:30 ratio for train and test data sets. The response variable 'price' as y and predictor variables as X in train and test data sets were created

• Modelling: For this exercise, two linear regression methods from sklearn and statsmodels packages has been applied.

• OLS method from statsmodel has been applied to get a R like summary of the model

• Multiple iterations of modelling has been applied to improve the coefficient of determination (R2) and to improve upon basic assumptions of linear regressions

• VIF values of the variables were used to validate the multicollinearity and few variables were eliminated for further iterations

• Model performance: Model performance measures such as R2 and Root Mean of Squared Errors (RMSE) is used to validate the linear regression model

• The predicted labels and the test labels are also plotted on a scatterplot to validate the predictions

• The assumptions of linear regression such as normality of residuals and homoscedasticity of residuals were also validated using probability plot and residual plot respectively

• Iteration 2 was selected as final model based on the performance measures

```
cut_dict = {'Fair': 0,
            'Good': 1,
            'Very Good': 2,
            'Premium': 3,
            'Ideal':4
           }
df.cut = df.cut.map(cut_dict)
df.cut.astype(str).astype(int)

color_dict = {'D': 6,'E': 5,'F': 4,'G': 3,'H': 2,'I': 1,'J': 0}
df.color = df.color.map(color_dict)
df.color.astype(str).astype(int)

clarity_dict = {'FL': 10,
                'IF': 9,
                'VVS1': 8,
                'VVS2': 7,
                'VS1': 6,
                'VS2': 5,
                'SI1': 4,
                'SI2': 3,
                'I1': 2,
                'I2': 1,
                'I3': 0}
```

Table 14: All three categorical ordinal variables encoded based on their lower to higher ranks

```
carat     26917 non-null  float64
cut       26917 non-null  int64
color     26917 non-null  int64
clarity   26917 non-null  int64
depth     26917 non-null  float64
table     26917 non-null  float64
x         26917 non-null  float64
y         26917 non-null  float64
z         26917 non-null  float64
price     26917 non-null  int64
```

Table 15: Data types after encoding

cut	color	clarity
4	2	9
3	3	5
2	1	5
4	3	5
3	2	5
4	3	7
2	4	6
4	1	3
3	3	4
4	3	3

Table 16: The categorical values after encoding

```
X_train.shape
(18841, 9)

X_test.shape
(8076, 9)

y_train.shape
(18841, 1)

y_test.shape
(8076, 1)
```

Summary of the final model

• In iteration 2, the square root of the response variable 'price' was used to normalise the response variable for training the model.

• To validate against test dataset the predictions using the trained model was squared.

• VIF values of x, y and z came out to be infinity as the values are same for each observation

• Volume feature was also introduced from the x, y and z factors. As x, y and z values are same and multicollinear to each other, only 'x' was used in the final model

• The overall P value is less than alpha (0.05), so rejecting the null hypothesis and accepting the alternate hypothesis, that at least one of the predictor variables is influencing the response variable

• That is, at least one regression co-efficient is not zero. Here all regression co-efficient are not zero

• The P value of all the predictor variables is less than alpha, which means that all the variables are statistically significant in deciding the response variable

• The model can explain 96.3%of the variance of price in train as both the R- squared and adjusted R-squared values are 0.963

• In test, the model can explain 93.8% of variation as the R-squared came out to be 0.938, making the model slightly overfit

OLS Regression Results

Dep. Variable:	price	R-squared:	0.963
Model:	OLS	Adj. R-squared:	0.963
Method:	Least Squares	F-statistic:	8.234e+04
Date:	Sat, 31 Jul 2021	Prob (F-statistic):	0.00
Time:	03:38:18	Log-Likelihood:	-57661.
No. Observations:	18841	AIC:	1.153e+05
Df Residuals:	18834	BIC:	1.154e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-89.0803	0.770	-115.694	0.000	-90.590	-87.571
carat	49.6690	0.773	64.252	0.000	48.154	51.184
cut	0.7375	0.035	20.949	0.000	0.668	0.806
color	1.9873	0.023	85.000	0.000	1.941	2.033
clarity	3.2577	0.025	128.093	0.000	3.208	3.308
volume	-0.1158	0.003	-43.574	0.000	-0.121	-0.111
x	18.0779	0.187	96.461	0.000	17.711	18.445

Omnibus:	1141.599	Durbin-Watson:	1.982
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3942.711
Skew:	0.240	Prob(JB):	0.00
Kurtosis:	5.189	Cond. No.	5.88e+03

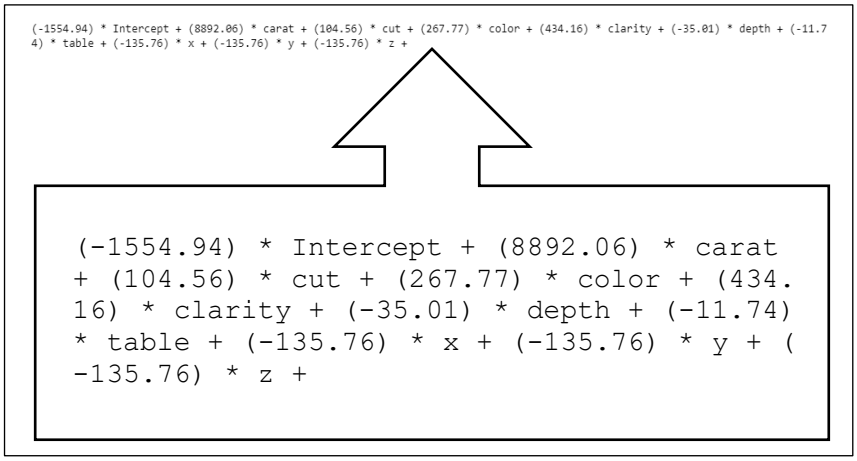
Table 17: The final model

Performance measures from three iterations

	R^2 Train	R^2 Test	RMSE Train	RMSE Test
Iteration 1	0.932	0.927	905.05	931.92
Iteration 2*	0.963	0.938	5.16	816.75
Iteration 3~	0.932	0.921	0.2606	0.2704

* In iteration 2, the response variable was transformed using square root to train the model.
~ In iteration 3, the dataset was scaled using Z score and OLS method applied.

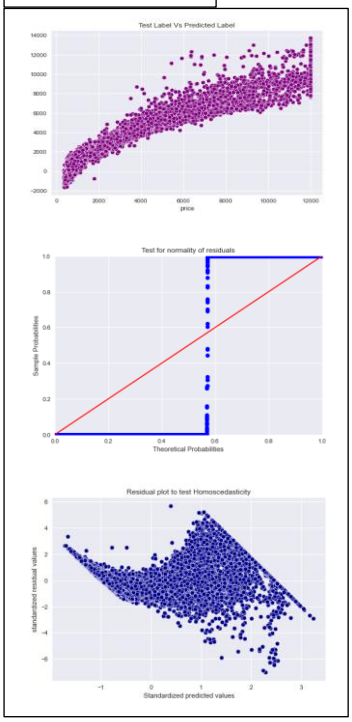
Table 18: Performance measure from three iterations



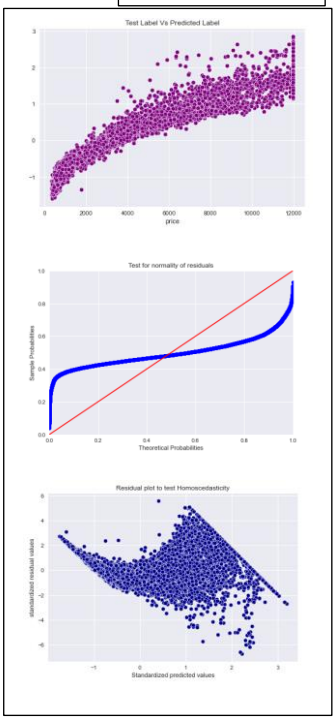
Final linear regression equation

LINEAR REGRESSION MODELLING

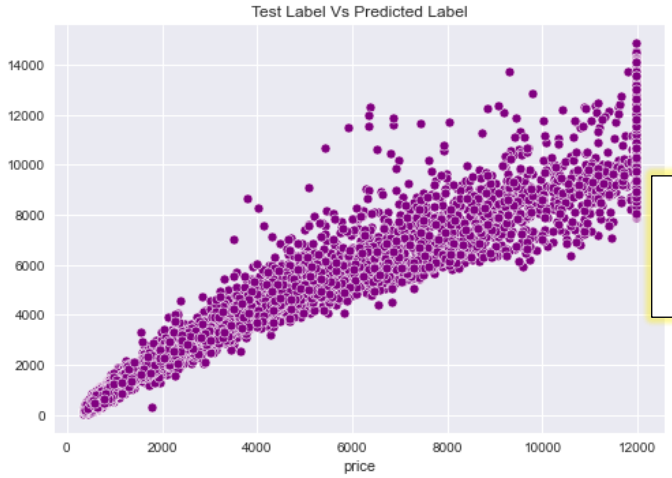
Iteration 1



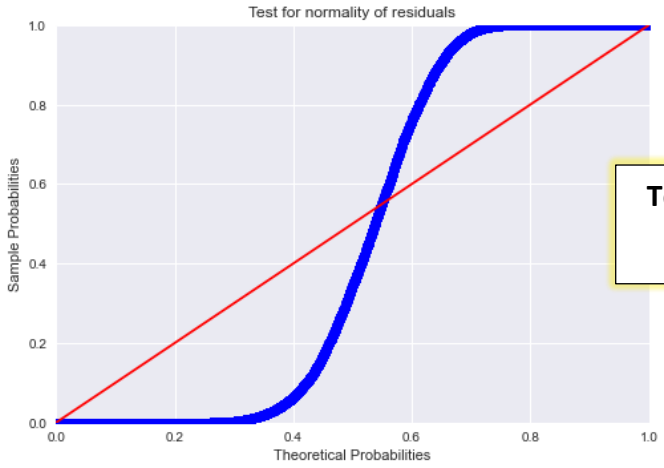
Iteration 3



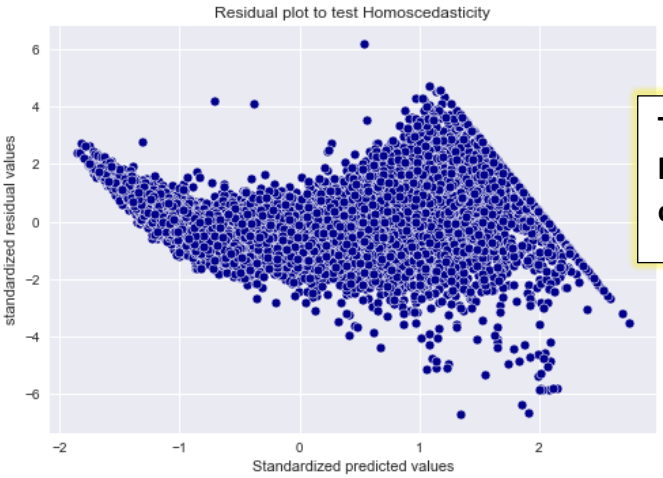
Final model



Test label Vs Predicted label for price



Test of normality of residuals



Test of homoscedasticity of residuals

Figure 49 to 51: Comparison of the three linear regression models (the 3 iterations)

1.4. INSIGHTS AND RECOMMENDATIONS

• The regression coefficient of the predictor variable 'carat' is found to be highest (8892.06), which means that for 1 unit increase in carat, the price of a Zirconia stone will increase by 8892 units

• 'Clarity' is found to be the next most important influencer in deciding the price with the coefficient of 434.16

• 'Color' and 'cut' are the next influencers, with coefficient values of 267.77 and 104.56 respectively

• The factors 'depth' and 'table' are found to be insignificant in deciding the price of Zirconia stones, also VIF score is also high

• Additional factor, volume was introduced in iteration 2 and 3 from the dimension variables, x, y and z

• In the final model, only 'x' was used as dimension and it was found to be the second most significant influencer in deciding the price

• In all the three iterations, the test for homoscedasticity was done by plotting the residuals and it was found that in all the three cases it follows a funnel shape pattern. It indicates that the residuals are heteroscedastic in nature, which violates one of the assumptions. It means scope for further improvement and would recommend further features to be added to improve the prediction of the response variable 'price'

• The recommendation for the company to identify higher profitable stones is to classify them based on the carat, dimension, clarity, color and cut in the respective order of significance

```
The coefficient for carat is 8892.062993696514
The coefficient for cut is 104.55511505651437
The coefficient for color is 267.76628197735033
The coefficient for clarity is 434.1590655419247
The coefficient for depth is -35.00943288069535
The coefficient for table is -11.74099667010334
The coefficient for x is -135.7609753524804
The coefficient for y is -135.76097535248027
The coefficient for z is -135.76097535248047
```

Table 19: Regression coefficients from iteration 1

Intercept	-89.080297
carat	49.669038
cut	0.737454
color	1.987285
clarity	3.257709
volume	-0.115795
x	18.077889

Table 20: Regression coefficients after feature engineering in final model

	column	VIF
0	carat	107.749383
1	cut	8.843211
2	color	5.535496
3	clarity	12.451131
4	depth	574.800463
5	table	577.205320
6	x	inf
7	y	inf
8	z	inf

VIF values of variables

Best 5 factors in deciding the price

1. Carat
2. Dimension (x/y/z)
3. Clarity
4. Color
5. Cut

PROBLEM 2 :
LOGISTIC REGRESSION & LDA

Executive summary

A tour and travel agency that deals in selling holiday packages hired us and provided us with the details of 872 employees of a company, among whom, some had opted for the package and some hadn't. Our task is to help the company predict whether an employee will opt for the package or not, on the basis of the information given in the dataset. Also, we must find out the important factors on the basis of which the company will focus on particular employees to sell their packages.

Maybe they took a company-paid holiday but how many of the other company employees will? People sit on the beach facing the Teign estuary and the holiday resort of Teignmouth in Shaldon, Devon, England. Visiting the fishing village of Shaldon a small cluster of mainly Georgian houses and shops at the mouth of the River Teign, is like stepping back into a bygone era. It features simple pleasures that hark back to analog, unplugged summer days: a book and a picnic blanket, a bucket and spade, fish and chips.

—AP Photo/Tony Hicks

Introduction

An employer-paid holiday is paid time off that allows an employee to observe a holiday if they choose.

Typically, employer-paid holidays are part of a larger compensation package that also includes other paid time off, such as vacation days and sick days.

Employer-paid holidays are days off with pay given to employees and traditionally associated with federally observed holidays. They are not required by law.

The employers are not bound by law to pay employees for time not worked, such as paid vacations, paid holidays, and paid sick leave. These are determined by the employer or negotiated by the employee union. Increasingly, competitively paid holidays and other time-off benefits are becoming crucial to an employer's ability to attract the best employees who have skills that are critical for the operation of the business.

Paid holidays may be negotiated by employees who have an employment contract. Senior-level employees with a contract are likely to have come from positions in other organizations where their seniority gave them the maximum paid holidays and vacation time.

Senior-level employees are unlikely to settle for less time off when accepting a new position. In fact, if an employer doesn't offer equivalent paid vacation and holiday time, it may be a deal-breaker for the prospect—even if they don't typically use all of the time available.

Employer-paid vacation time can differ between exempt and non-exempt employees. Exempt employees are expected to work whatever hours are necessary to complete a job. They may have more flexibility in their schedule and less supervision, as well. These employees are exempt from the overtime provisions. Travel agencies do have an unpredictable time selling holiday package to employee. It’s always a suspense how many will subscribe to the scheme. But if past could predict the future, the companies can cultivate frequent travellers indeed. Machine learning can help.

Travel agencies take a hard look at their data. Travel consultant Phillip Koinis sits behind the desk at Oxford Travel agency in Sydney, Australia. —Stefica Nicol Bikes/Reuters

Logistic regression and LDA

Based on the historical data, accurate estimation of future response is an important problem. However, so far only the case of a continuous response has been considered.

In reality response can be continuous or categorical. When an applicant comes to a bank for loan, the bank asks for a lot of information on the applicant; such as, age, salary, marital status, educational background and qualification, any other existing loan, if yes, then monthly payment, number of children and many other facts.

Based on this information the bank takes a decision regarding whether the loan may be approved or not. In this case, approval is the response which may take only two values Approved or Rejected. Clearly this is a case where the response is binary. The standard linear regression modelling will not work here. Binary response is a very common phenomenon. In financial domain credit card default or loan default is a major source of uncertainty. Credit card companies would like to predict, given a person's earlier behaviour, whether s/he would be able to pay the minimum amount next month. E-commerce companies would like to predict hether a prospective buyer would close a deal with them, given her browsing pattern.

Algorithms are developed to classify an email as spam and direct it away from inbox. In all such cases, and in many more, response has only two levels, denoted by Yes (Success) or No (Failure). In this business challenge, we deal with binary response only and the technique developed will be logistic regression. Buy no means, response needs to be restricted to two levels nor logistic model is the only method of binary prediction. In healthcare domain, given a patient's risk profile, a physician would like to categorize him as low risk, medium risk or high risk.

Response here has multiple ordinal levels. Several extensions of logistic regression have been mentioned in Section 4.

Logistic regression may also be taken as a classification algorithm, since it assigns a class label to each observation. The underlying technique of classification is based on regression and a misclassification probability can also be calculated.

Logistic regression is a modelling technique for a binary response. Unlike linear regression, the value of the response is not predicted. If the two response levels are taken as success and failure, probability (of success) is calculated based on the values of the predictors.

LDA or linear discriminant analysis uses linear combinations of independent variables to predict the class in the response variable of a given observation. LDA assumes that the independent variables (p) are normally distributed and there is equal variance / covariance for the classes.

LDA is popular, because it can be used for both classification and dimensionality reduction. When these assumptions are satisfied, LDA creates a linear decision boundary. Research claim that the LDA performs well when these assumptions are violated.

"Linear discriminant analysis frequently achieves good performances in the tasks of face and object recognition, even though the assumptions of common covariance matrix among groups and normality are often violated (Duda, et al., 2001)." (Tao Li, et al., 2006). LDA is based upon the concept of searching for a linear combination of predictor variables that best separates the classes of the target variable.

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. DESCRIPTIVE STATISTICS

	Unnamed: 0	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	1	no	48412	30	8	1	1	no
1	2	yes	37207	45	8	0	1	no
2	3	no	58022	46	9	0	0	no
3	4	no	66503	31	11	2	0	no
4	5	no	66734	44	12	0	2	no
5	6	yes	61590	42	12	0	1	no
6	7	no	94344	51	8	0	0	no
7	8	yes	35987	32	8	0	2	no
8	9	no	41140	39	12	0	0	no
9	10	no	35826	43	11	0	2	no

Table 21: The holiday package dataset

• The dataset contains 872 observations and 7 variables. The dependent variable is Holiday_Package, indicating whether an employee has opted for a holiday or not

• Salary, age, education, number of young children and number of older children are the continuous independent variables. The categoric variable foreign indicate whether an employee is a foreign national or not

• The average salary of an employee is around 47729.2, maximum is 236961 and minimum 1322. The range suggests outliers

• The average age of an employee is around 40 years, the youngest being 20 and the oldest 62.

• On an average the employees had about 10 years of formal education. The highest being 21 years and the lowest at 1 year

• The proportion of foreign nationals in the company is 25%

• The proportion of the target variable 'Holiday_Package' shows that 46% of the worker took the package, while 54% didn't.

• There are no null values in the dataset • After removing the outliers from the continuous variables, it is noticed that all the observations for no_of_young_children are in zero values. Thus, the variable has been removed before further analysis

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Holliday_Package	872	2	no	471	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	872	NaN	NaN	NaN	47729.2	23418.7	1322	35324	41903.5	53469.5	236961
age	872	NaN	NaN	NaN	39.9553	10.5517	20	32	39	48	62
educ	872	NaN	NaN	NaN	9.30734	3.03626	1	8	9	12	21
no_young_children	872	NaN	NaN	NaN	0.311927	0.61287	0	0	0	0	3
no_older_children	872	NaN	NaN	NaN	0.982798	1.08679	0	0	1	2	6
foreign	872	2	no	656	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 22: Descriptive statistical summary of the holiday package dataset

```
Holliday_Package
no     471
yes    401
```

```
foreign
no     656
yes    216
```

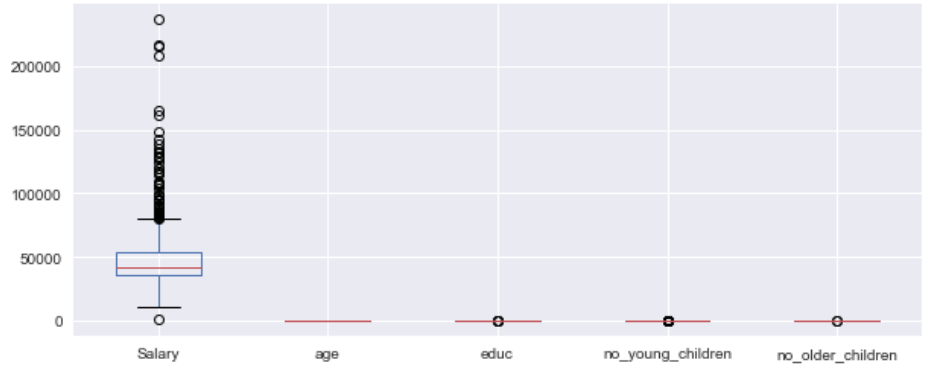
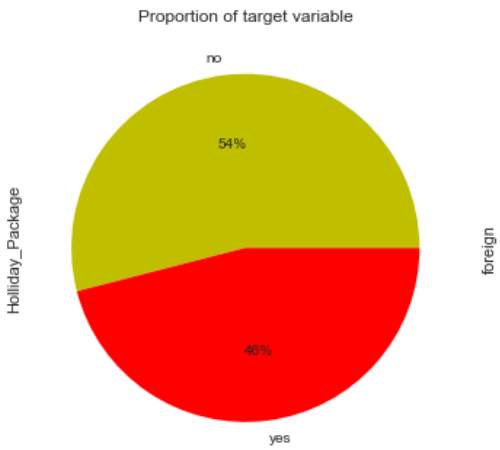


Figure 53: Outliers before being removed



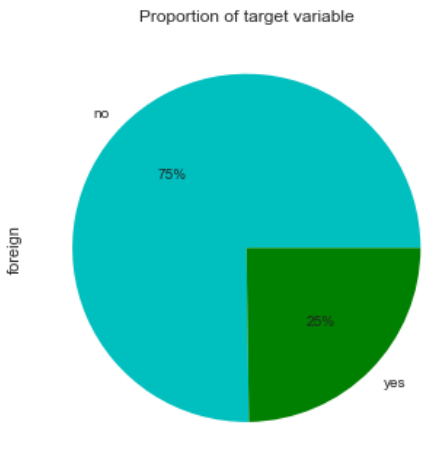


Figure 52: Read these pie charts to know how many people took the package and how many of the takers are foreign nationals

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

Table 23: No null values in the dataset

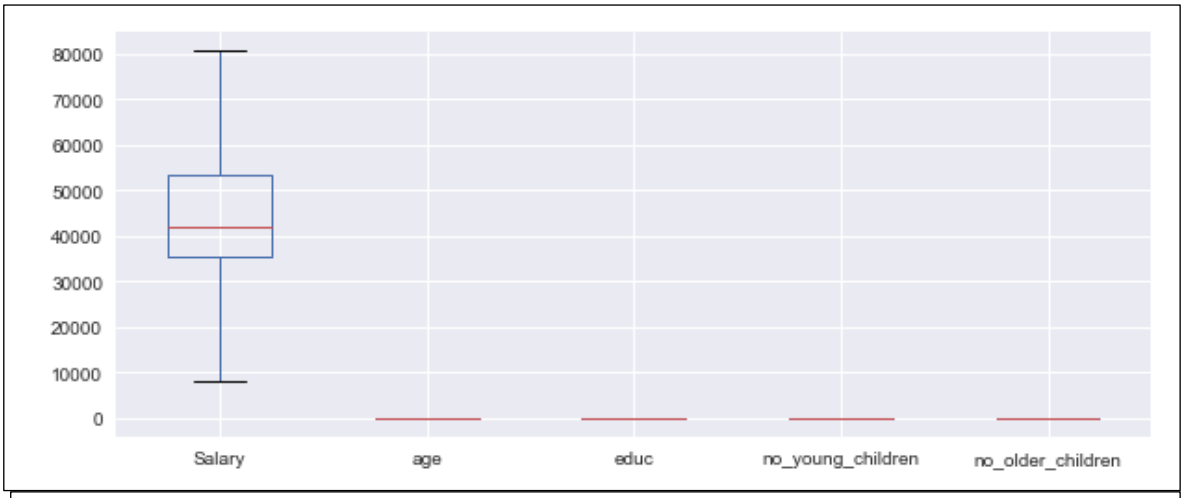
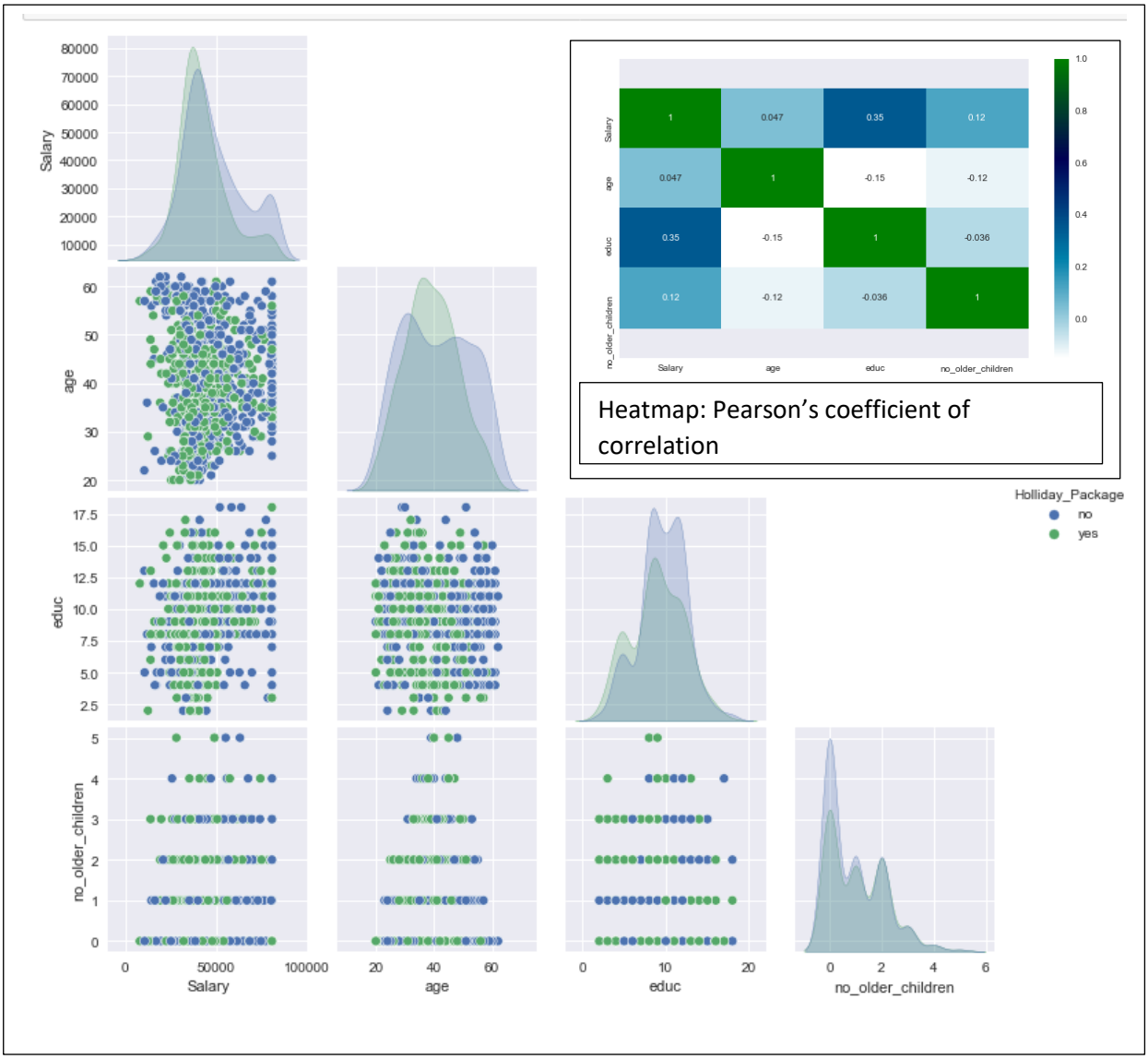


Figure 54: Outliers removed from the holiday package dataset



Heatmap: Pearson's coefficient of correlation

• There is no significant correlation found between the independent and target variable • Only salary and age shows some level of differentiation in their distribution with respect to an employee's decision to holiday package or not

• Education and no of older children aren't good predictors for the target variable as their distributions overlaps each other • Employees of foreign nationality are more likely to opt for a holiday package, whereas more local employees opted not to take a holiday package

Fig 54

• The pair plot and heatmap of the variables are as in the previous slide

• Jointplots of holiday package Vs other dependent variables such as salary, age, years of education and no of children above 7 years has been plotted

• From the plots, we can infer that there is no significant correlation between salary, age, years of education and no of older children with an employee opting for a holiday package or not

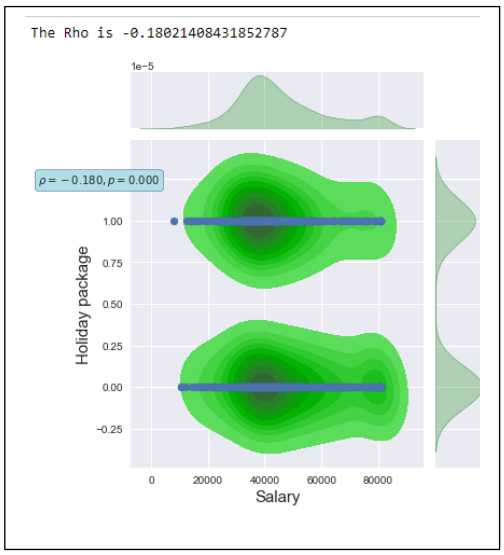
• We can infer that among those who opted for a holiday package, their density is higher around 30000 and 50000 range of salary. But we can't find any relation between those who opted for a holiday package or not

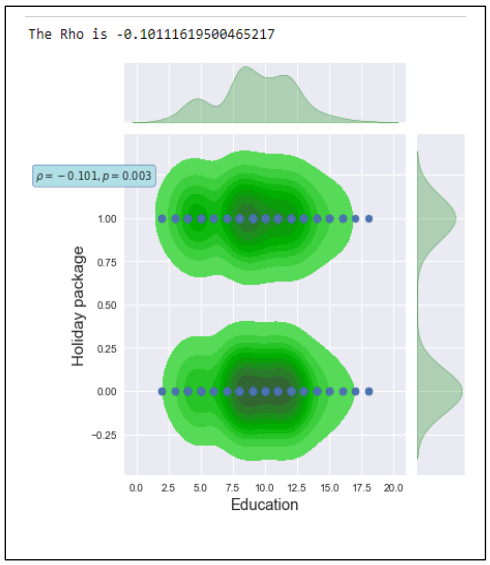
• Similarly age does not have any significant correlation with an employee's decision to opt for a holiday package. However there is a higher density between the age group of 30 to 50 years in opting for a package

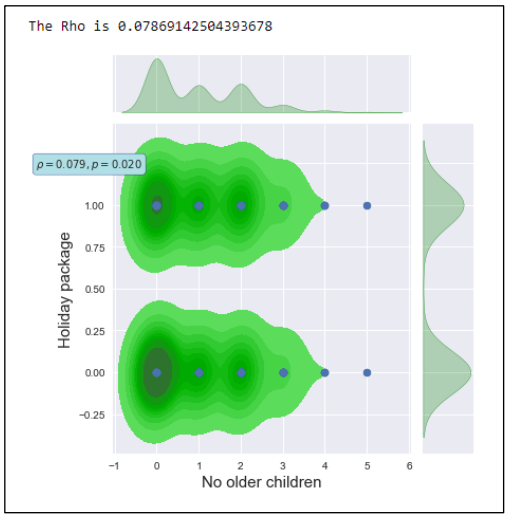
• The years of formal education is also a poor predictor of whether an employee would opt for a holiday package or not

• The no of older children an employee has, is also turned out to be a very poor predictor as the distribution on status of holiday package is completely overlapping

• Though none of the predictor variables is helpful in deciding the status of the target variable, the salary and age may be able to provide some differentiation







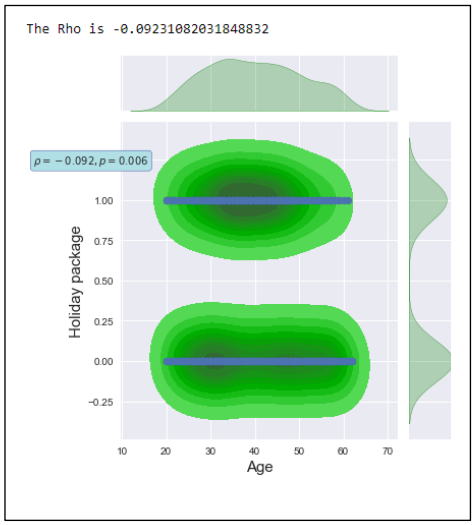


Fig 55 to 58: Jointplots of holiday package versus several dependent variables

MORE EXPLORATORY DATA ANALYSIS

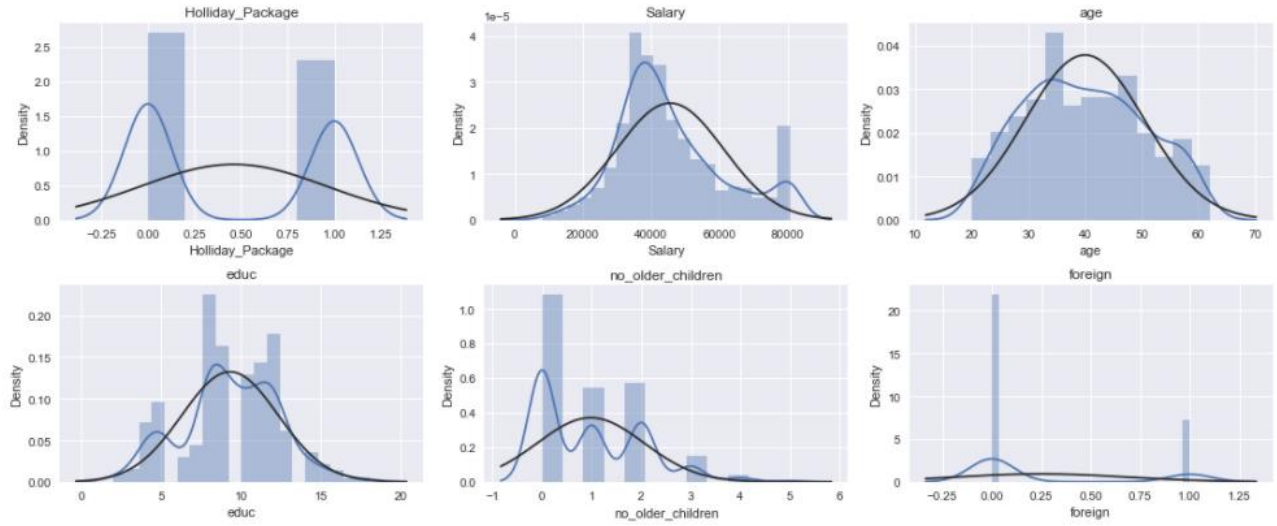


Figure 59 to 64: The distribution of various variables of the holiday package dataset. There are several bimodal and multimodal variables. Normal distribution curve is plotted in black for reference

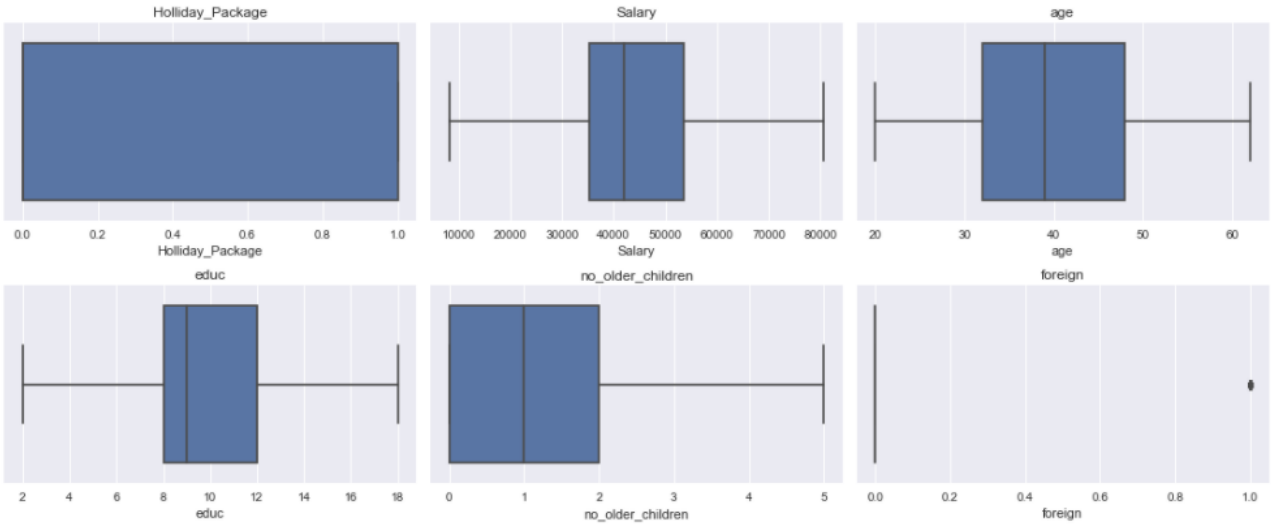
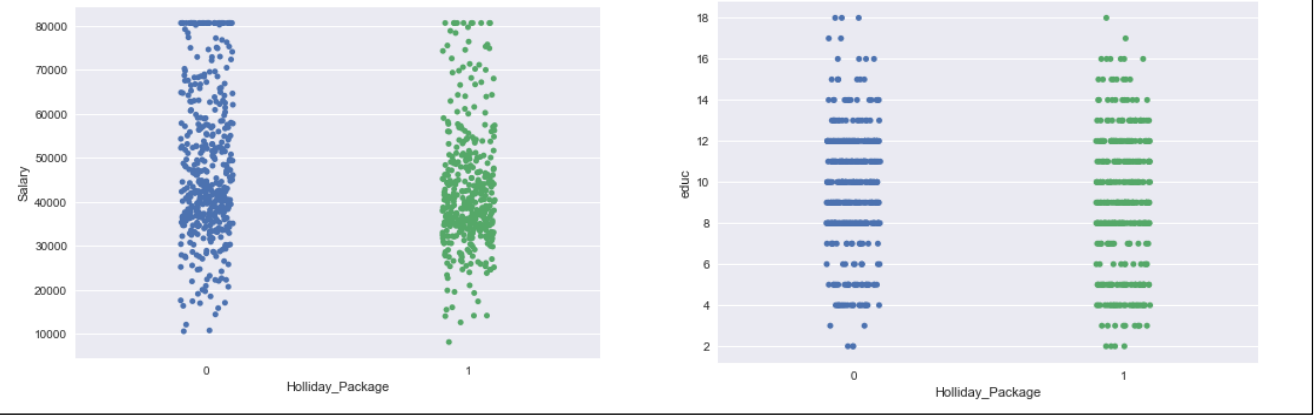
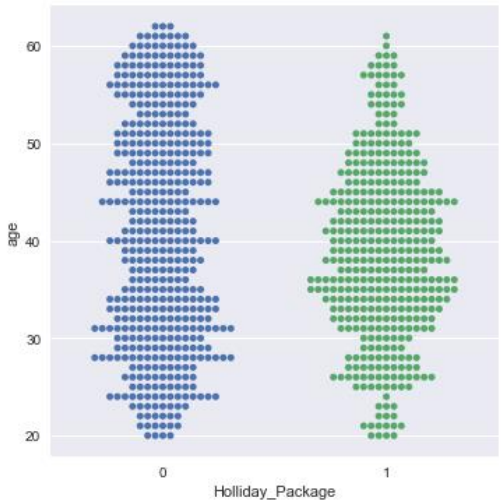


Figure 64 to 69: Boxplots of the variables of the holiday package dataset. Salary is concentrated, while a bulk of the employee force is below 50. Below (Fig 70 & 71): Bulk of pay and education is in mid ranges





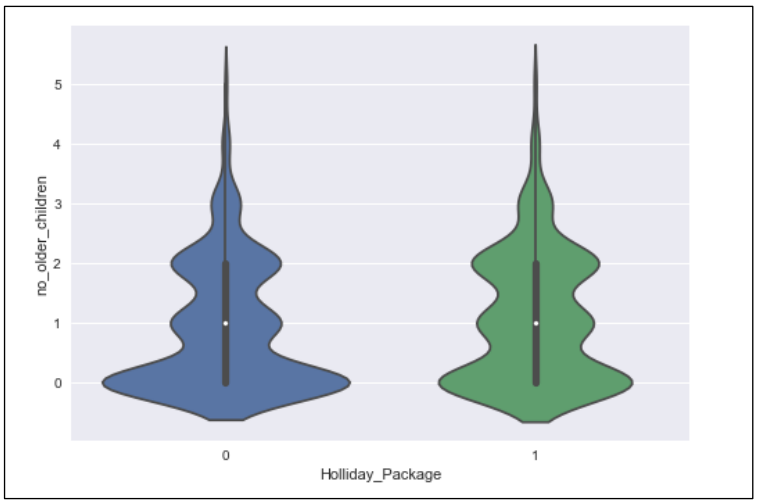


Figure 72 and 73: Holiday package versus age and no older children. Both bottom heavy when it comes to density

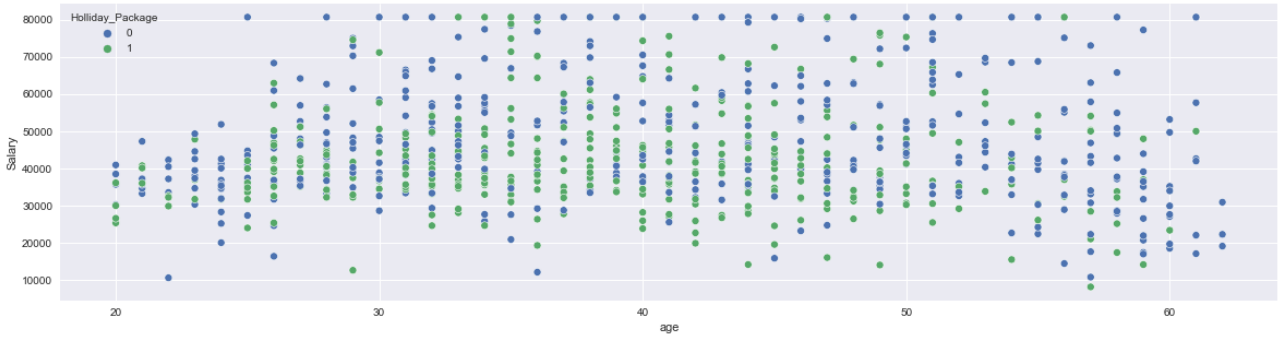


Figure 74: Age versus salary, hued on holiday package

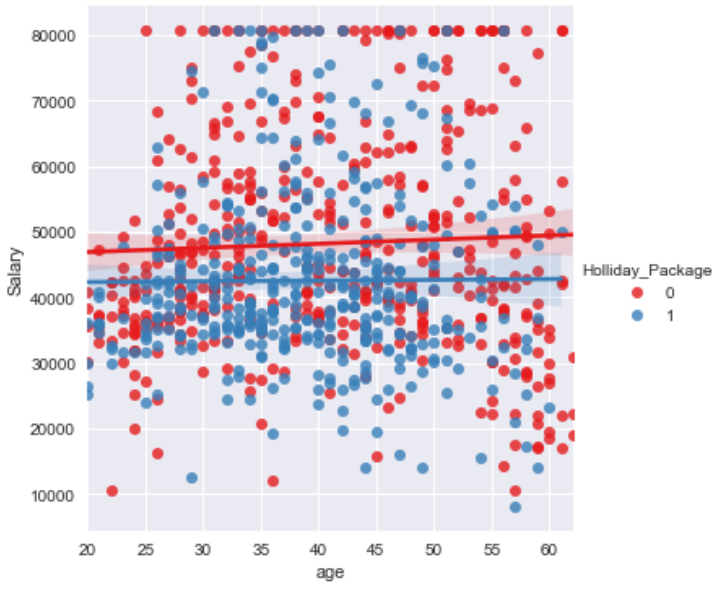


Fig 75: An lmplot of the same thing as above

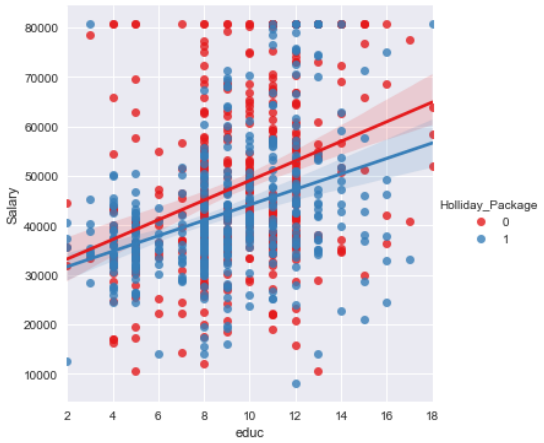
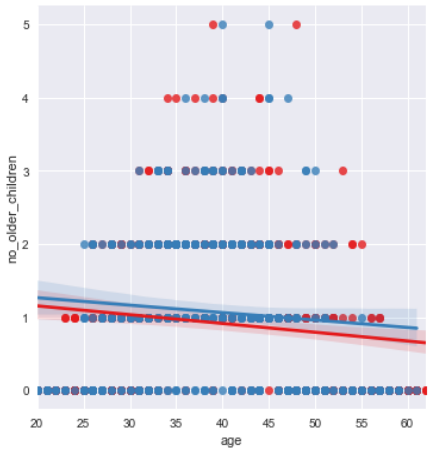


Fig 76: Education & salary. Red line for those who refused the holiday package



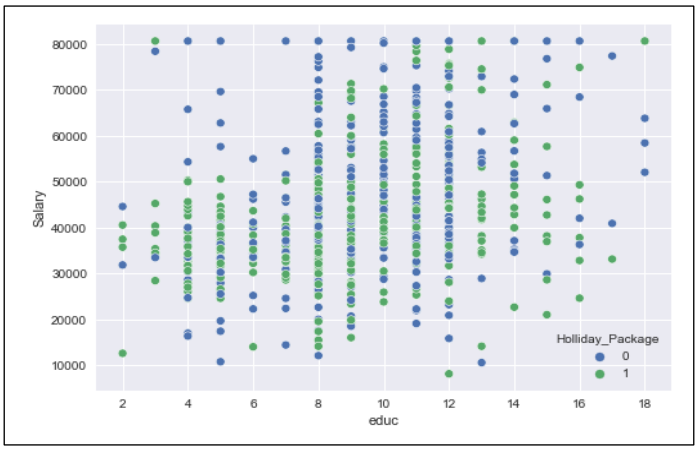


Fig 77 and 78: Two graphs hued on holiday package: Age versus no older children, and education versus salary. Data not well separated. Time for LDA to get into action

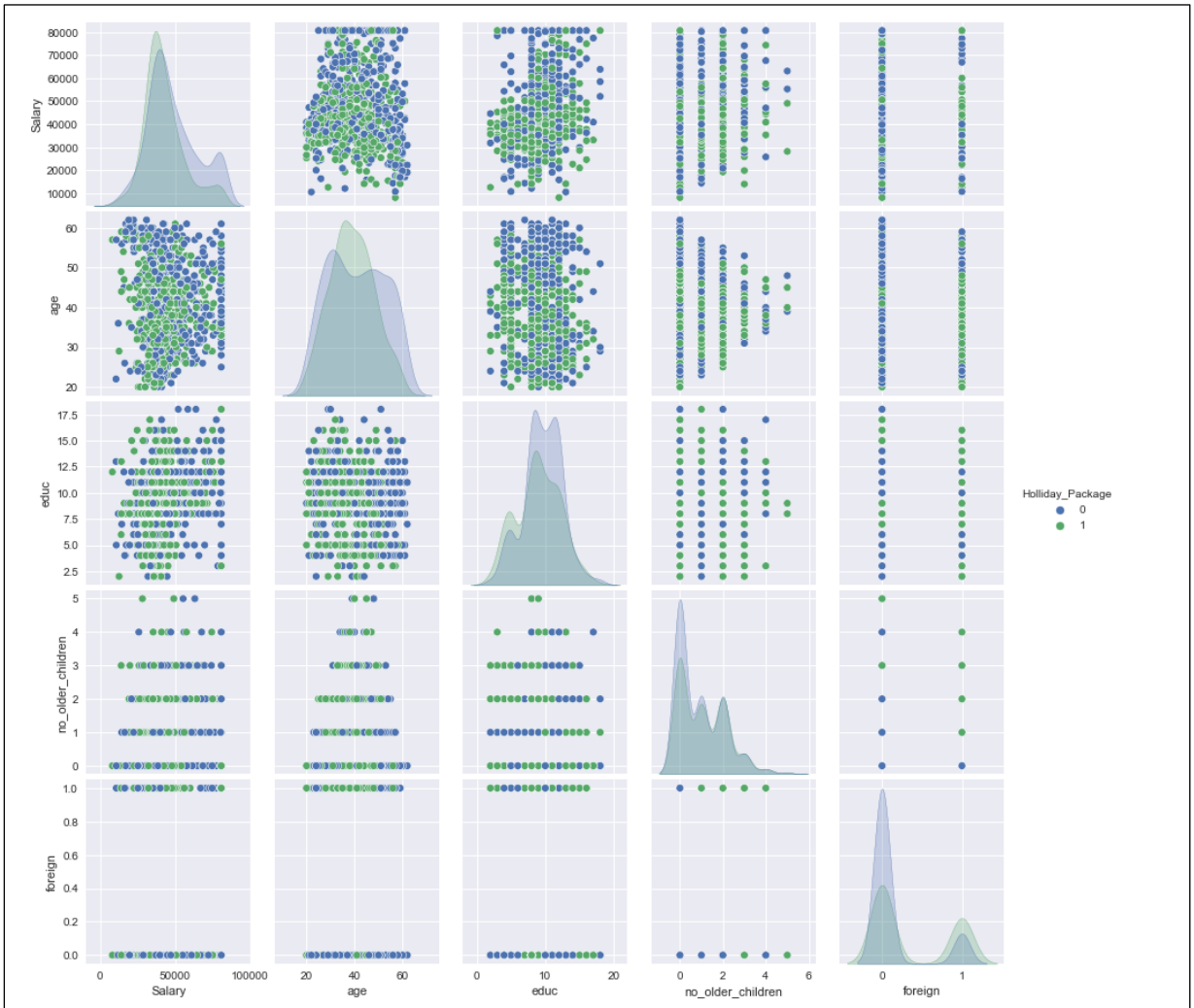
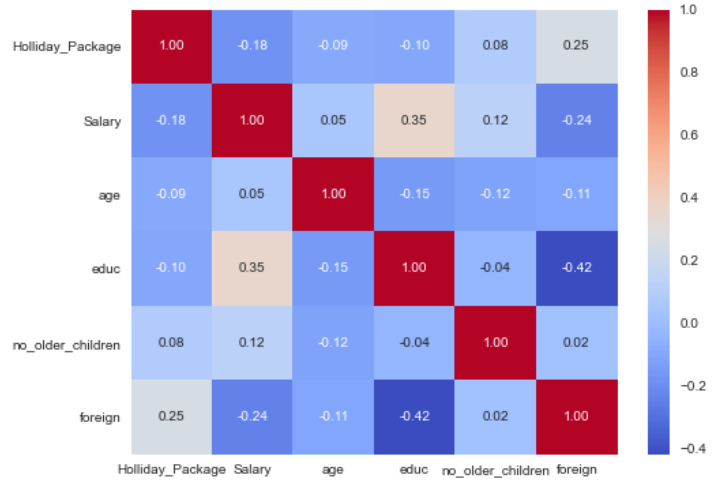


Figure 79: A pairplot hued on holiday package. A lot of overlap



	Feature 1	Feature 2	Correlation Coefficient
5	Holliday_Package	Holliday_Package	1.000000
8	Holliday_Package	foreign	0.254096
13	Holliday_Package	no_older_children	0.079477
20	Holliday_Package	age	-0.063780
23	Holliday_Package	educ	-0.081569
31	Holliday_Package	Salary	-0.146646

Figure 80, Table 23: Checking for correlation of the features with target variable holiday package. None found.

2.2. PREDICTIVE MODELLING

```
RangeIndex: 872 entries, 0 to 871
Data columns (total 6 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Holliday_Package   872 non-null    int8
 1   Salary             872 non-null    float64
 2   age                872 non-null    float64
 3   educ               872 non-null    float64
 4   no_older_children  872 non-null    float64
 5   foreign            872 non-null    int8
```

```
LogisticRegression(max_iter=10000, n_jobs=2, penalty='none', solver='newton-cg')
```

Logistic regression model fitted

	Holliday_Package	Salary	age	educ	no_older_children	foreign
0	0	48412.00	30.0	8.0	1.0	0
1	1	37207.00	45.0	8.0	1.0	0
2	0	58022.00	46.0	9.0	0.0	0
3	0	66503.00	31.0	11.0	0.0	0
4	0	66734.00	44.0	12.0	2.0	0
5	1	61590.00	42.0	12.0	1.0	0
6	0	80687.75	51.0	8.0	0.0	0
7	1	35987.00	32.0	8.0	2.0	0
8	0	41140.00	39.0	12.0	0.0	0
9	0	35826.00	43.0	11.0	2.0	0
10	0	42643.00	45.0	11.0	2.0	0
11	0	35157.00	60.0	12.0	0.0	0
12	0	75327.00	33.0	11.0	0.0	0
13	0	80687.75	56.0	14.0	0.0	0
14	0	80687.75	56.0	11.0	0.0	0
15	0	80297.00	47.0	11.0	1.0	0
16	0	52117.00	50.0	8.0	0.0	0
17	1	80687.75	39.0	12.0	0.0	0
18	0	62858.00	47.0	8.0	1.0	0
19	1	57400.00	53.0	11.0	0.0	0

Table 24 and 25: The data types (top) and the dataset after encoding

	0	1
0	0.328956	0.671044
1	0.620871	0.379129
2	0.520804	0.479196
3	0.529495	0.470505
4	0.502566	0.497434
5	0.362219	0.637781
6	0.611121	0.388879
7	0.224005	0.775995
8	0.344641	0.655359
9	0.601561	0.398439

Table 26: Predicted probabilities for logit model

```
{'C': 0.004281332398719396,
 'penalty': 'l2',
 'solver':
 'newton-cg'}
```

Best parameters

	0	1
0	0.571193	0.428807
1	0.499505	0.500495
2	0.507600	0.492400
3	0.547663	0.452337
4	0.431135	0.568865
5	0.489791	0.510209
6	0.538316	0.461684
7	0.381410	0.618590
8	0.502573	0.497427
9	0.520557	0.479443

Table 27: Predicted probabilities for LDA model

```
GridSearchCV(cv=10, estimator=LogisticRegression(),
             param_grid={'C': array([1.00000000e-03, 2.06913808e-03, 4.28133240e-03, 8.85866790e-03,
       1.83298071e-02, 3.79269019e-02, 7.84759970e-02, 1.62377674e-01,
       3.35981829e-01, 6.95192796e-01, 1.43844989e+00, 2.97635144e+00,
       6.15848211e+00, 1.27427499e+01, 2.63665090e+01, 5.45559478e+01,
       1.12883789e+02, 2.33572147e+02, 4.83293024e+02, 1.00000000e+03]),
                         'penalty': ['l2', 'none'], 'solver': ['newton-cg']},
             scoring='recall', verbose=True)
```

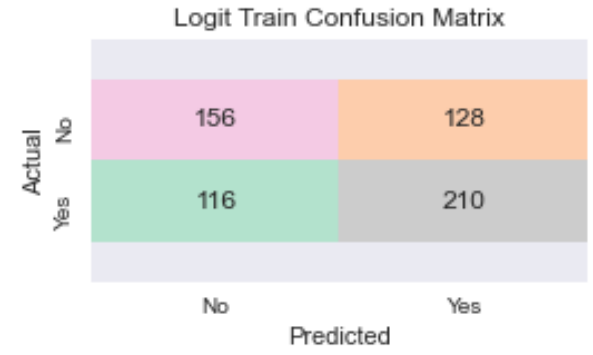
Logistic regression modelling using grid search

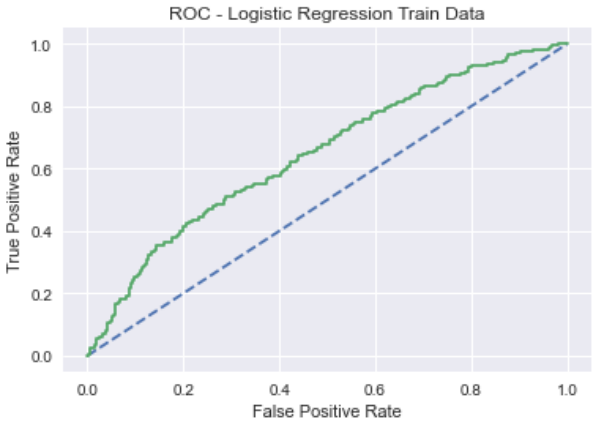
2.3. PERFORMANCE METRICS

LOGISTIC REGRESSION

Train

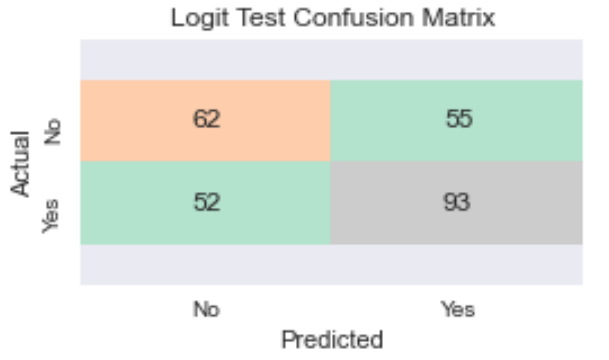
	precision	recall	f1-score	support
0	0.621302	0.644172	0.632530	326.0
1	0.573529	0.549296	0.561151	284.0
accuracy	0.600000	0.600000	0.600000	0.6
macro avg	0.597416	0.596734	0.596841	610.0
weighted avg	0.599060	0.600000	0.599298	610.0





Test

	precision	recall	f1-score	support
0	0.628378	0.641379	0.634812	145.000000
1	0.543860	0.529915	0.536797	117.000000
accuracy	0.591603	0.591603	0.591603	0.591603
macro avg	0.586119	0.585647	0.585804	262.000000
weighted avg	0.590635	0.591603	0.591042	262.000000



The categoric variables Holiday_Package and foreign are encoded and converted to integers

• The target positive case 'yes' for Holiday_Package is 1 and the negative case 'No' is 0

• The data is split in 70:30 ratio into train and test datasets

• LogisticRegression() method from sklearn package is used to build the model

• To optimise the hyper-parameters, GridSearchCV is used

• LinearDiscriminantAnalysis() from sklearn package is used for building the LDA Model

• The classification report and confusion metrics for the models are used

• The LDA model produced an accuracy of 65% in train and 60% in test

• The precision for the positive target case in train is 68% and in test it is 58%

• As recall is considered as the most significant performance measure for this business case, the recall rate for the positive target case in train is 49% and in test it is 39%

• The confusion metrices for both train and test are given

• AUC for train is 0.67 and for test it is 0.65

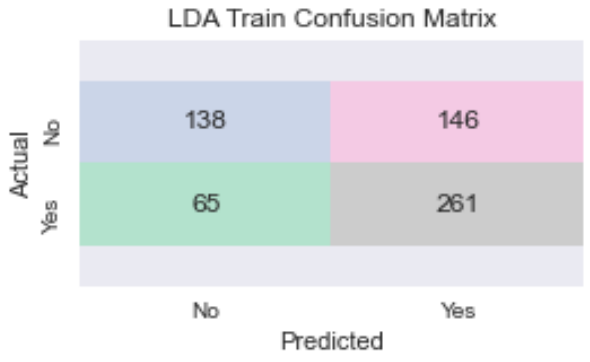
• The recall rate came out to be too low and the model is found to be overfitting in terms of accuracy

• The derived model is not reliable to predict the target cases

LINEAR DISCRIMINANT ANALYSIS

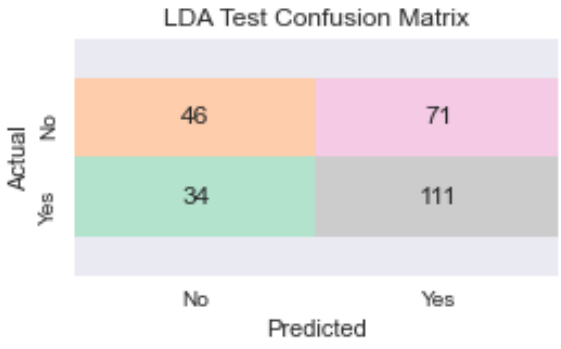
Train

	precision	recall	f1-score	support
0	0.641278	0.800613	0.712142	326.000000
1	0.679803	0.485915	0.566735	284.000000
accuracy	0.654098	0.654098	0.654098	0.654098
macro avg	0.660540	0.643264	0.639438	610.000000
weighted avg	0.659214	0.654098	0.644444	610.000000



Test

	precision	recall	f1-score	support
0	0.609890	0.765517	0.678899	145.000000
1	0.575000	0.393162	0.467005	117.000000
accuracy	0.599237	0.599237	0.599237	0.599237
macro avg	0.592445	0.579340	0.572952	262.000000
weighted avg	0.594309	0.599237	0.584275	262.000000



• The LDA model produced an accuracy of 65% in train and 60% in test

• The precision for the positive target case in train is 68% and in test it is 58%

• As recall is considered as the most significant performance measure for this business case, the recall rate for the positive target case in train is 49% and in test it is 39%

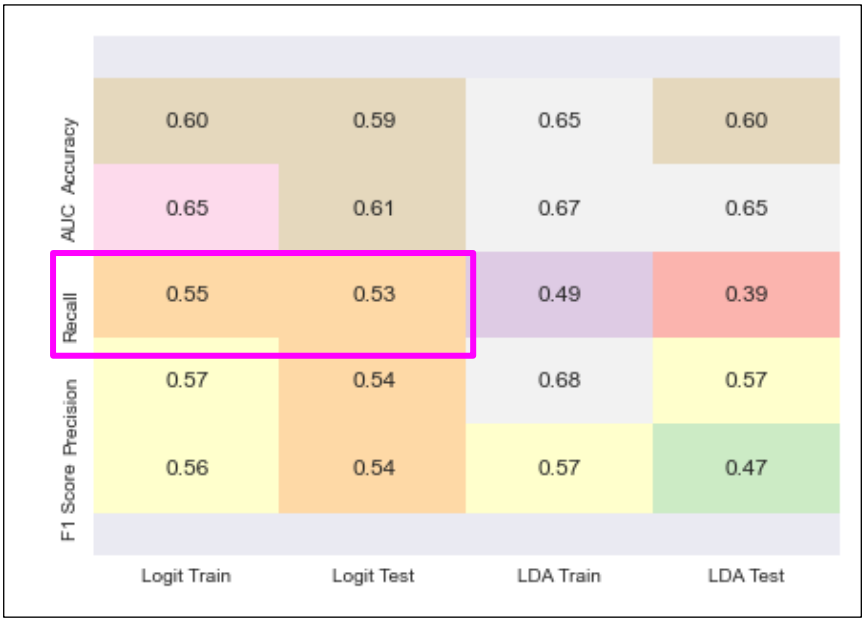
• The confusion metrices for both the train and test is given here

• AUC for train is 0.67 and for test it is 0.65

• The recall rate came out to be too low and the model is found to be overfitting in terms of accuracy

• The derived model is not reliable to predict the target cases

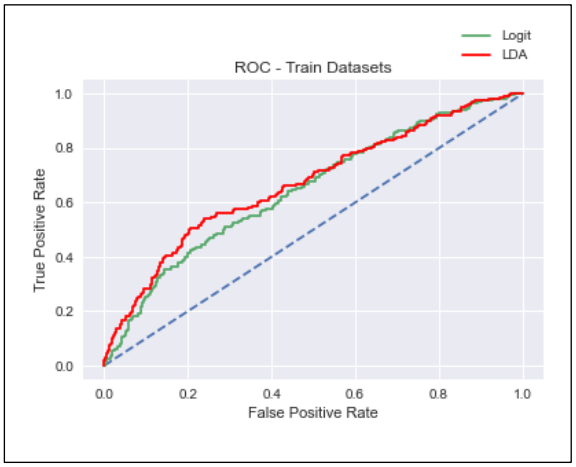
FINAL MODEL COMPARISON

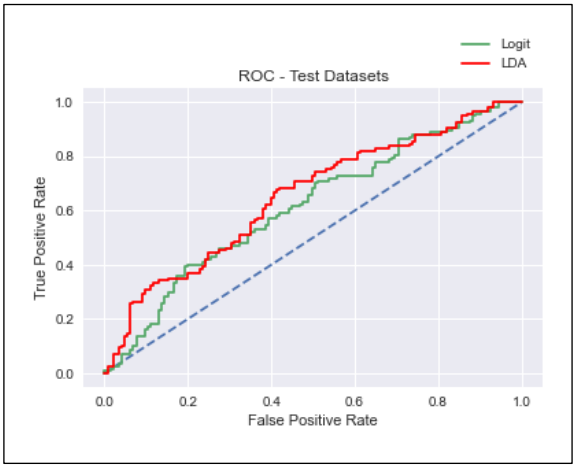


• The accuracy of the LDA model is higher than that of the Logit model, 60% and 59% respectively for the test dataset

• The precision for the test dataset is also higher in the LDA model than the Logit model, 57% and 54% respectively

• The AUC score on the test dataset from the LDA model is significantly higher than the Logit model, 0.65 and 0.61 respectively





• However, considering the business case the ability to recall the positive target case is important than other performance measures

• The recall score is higher from the logit model at 53% for the test dataset, were as the recall score from the LDA model is only 39% which is unacceptable

• The overall F1 score is also higher in the Logit model at 54% for test data. Whereas it is only 47% for test data in the LDA Model

• The ROC curve shows a better coverage in the LDA model, but based on the recall and F1 score, Logit model is the final choice for this business case

2.4. INSIGHTS & RECOMMENDATIONS

• None of the regression models gave us an optimistic model for predicting the target variable, which is because none of the given predictor variables could differentiate between the positive and negative target cases

• The resulted predictions are closer to the 'toss of a coin' than to a reliable that foretells the target response variable

• A recall rate of 53% indicates that the model has been able to recollect only half of the actual cases of employees who opted for a holiday package

• The travel company can introduce tailored packages based on age groups to attract employees at different age levels to opt for respective packages

• Value add-ons could be considered based on number of children the employee has and the age of the children.

• Foreign nationals can be encouraged with targeted add-ons and specific packages based on their interest, which will vary from that of locals

• We would like to recommend the travel company to identify new features that can correlate the pattern of opting for a holiday package or not positively for training and building a better model

• Features such as marital status, gender, interest categories in travel (adventure, beach side, hill stations, theme parks) etc could be considered as additional features

—Courtesy Anirban Bora, The Times of India