Predictive Modeling

PGPDSBA Online Feb_D 2021

Project - 5

Vaishnavi Karelia

Table of Contents

List of Figures

List of Tables

Problem 1: Linear Regression

A company Gem Stones co ltd, which is a cubic zirconia manufacturer has 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. Assignment is 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	cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, IF = 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. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA). Perform Univariate and Bivariate Analysis.

Reading the data

	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

Table - 1 cubic zirconia - top 5 rows

Initial checks: Null values, Duplicate values, shape, dimension, datatypes etc.

Info of the data

```
----------------
<class 'pandas.core.frame.DataFrame'>
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
None
```

Dimension check

```
---------------
Total number of rows in the dataset = 26967
Total number of columns in the dataset = 11
Total number of elements in the dataset = 296637
```

Missing values check

```
--------------------
There are 697 missing values in the data. Need further checks!
```

Duplicate Values Check

```
----------------------
Number of duplicate rows = 0
There are no duplicate values in the dataset.
```

Inference

There are 26967 rows and 10 columns in this dataset which is 269670 total elements. There are float64(6), int64(1), object(3) <- carat, depth, table, x, y, z as float64, price ad int64 and cut, clarity and color as object datatypes. We will need to encode the categorical columns for our model.
Total missing values in the dataset are 697. From the info, we can see that the missing values are from dept column of the dataset. We will need to further perform statistical analysis for depth to impute the missing values.
Total duplicate values in the dataset are 0.
Total memory usage by the dataset is 2.1+ MB

Checking for duplicate values, if any after dropping the sr no columns (Unnamed: 0)

```
Duplicate Values Check
----------------------
Number of duplicate rows = 34
Need further checks!
```

	carat	cut	color	clarity	depth	table	x	y	z	price
4756	0.35	Premium	J	VS1	62.4	58.0	5.67	5.64	3.53	949
6215	0.71	Good	F	SI2	64.1	60.0	0.00	0.00	0.00	2130
8144	0.33	Ideal	G	VS1	62.1	55.0	4.46	4.43	2.76	854
8919	1.52	Good	E	I1	57.3	58.0	7.53	7.42	4.28	3105
9818	0.35	Ideal	F	VS2	61.4	54.0	4.58	4.54	2.80	906

Table - 2 cubic_zirconia - duplicate entries

Total number of rows are 26933 after removing the duplicate values. Total 34 rows are duplicate i.e., 0.1261% of the original dataset.

Exploratory Data Analysis

Univariate Analysis for numerical variables

	count	mean	std	min	25%	50%	75%	max	cv
carat	26933.0	0.798010	0.477237	0.2	0.40	0.70	1.05	4.50	0.598034
depth	26236.0	61.745285	1.412243	50.8	61.00	61.80	62.50	73.60	0.022872
table	26933.0	57.455950	2.232156	49.0	56.00	57.00	59.00	79.00	0.038850
x	26933.0	5.729346	1.127367	0.0	4.71	5.69	6.55	10.23	0.196771
y	26933.0	5.733102	1.165037	0.0	4.71	5.70	6.54	58.90	0.203212
z	26933.0	3.537769	0.719964	0.0	2.90	3.52	4.04	31.80	0.203508
price	26933.0	3937.526120	4022.551862	326.0	945.00	2375.00	5356.00	18818.00	1.021594

Table - 3 cubic_zirconia – stats for numerical variables

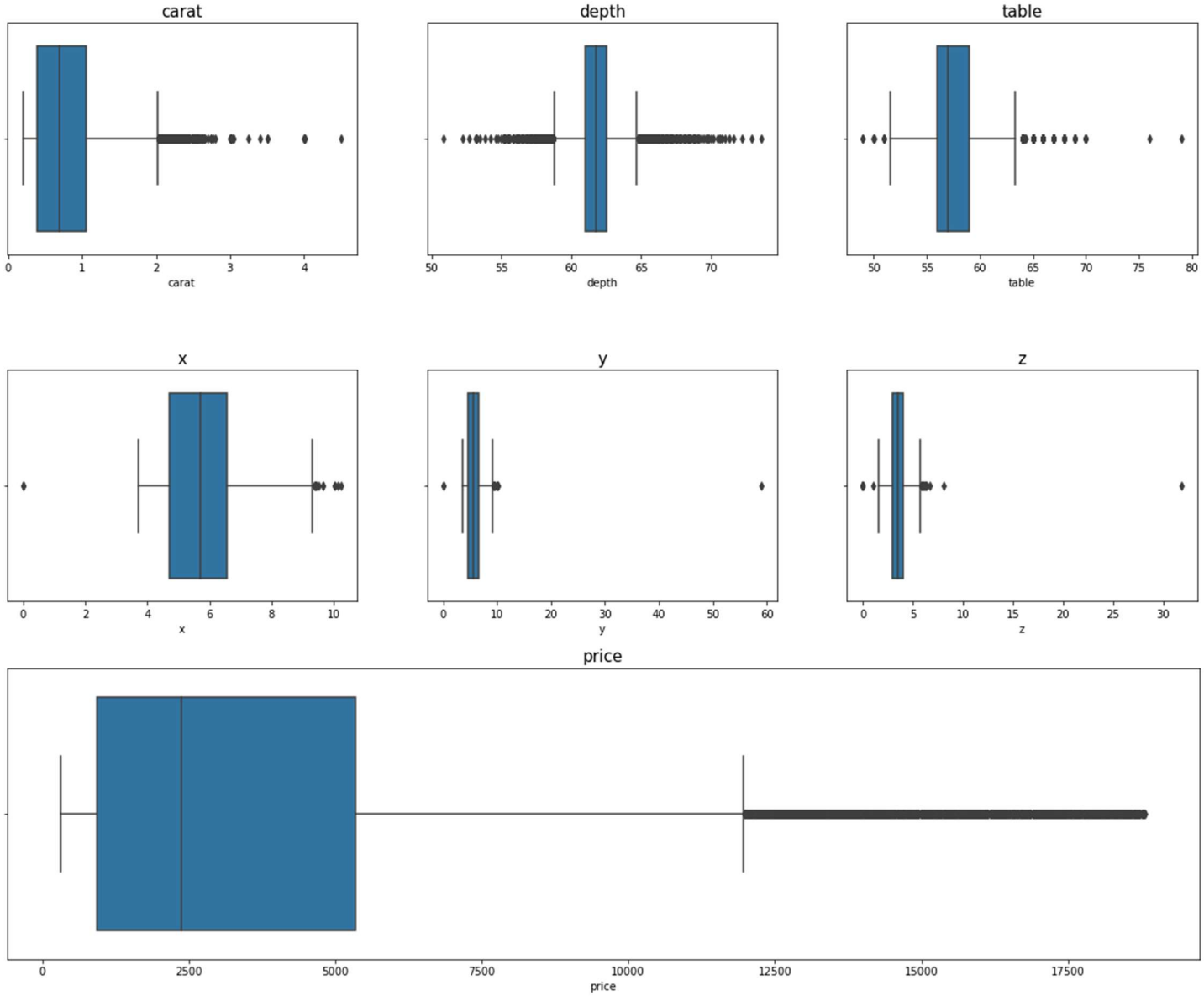


Figure - 1 Univariate Analysis - Numerical

Shapiro-Wilk test for normality

```
------------------------------------------------------------
pvalue for carat column is 0.0,
The data is not normally distributed.

pvalue for depth column is 1.0,
The data is normally distributed.

pvalue for table column is 0.0,
The data is not normally distributed.

pvalue for x column is 0.0,
The data is not normally distributed.

pvalue for y column is 0.0,
The data is not normally distributed.

pvalue for z column is 0.0,
The data is not normally distributed.
```

Detecting outliers using IQR

```
------------------------------------------------------------
```

	Columns	Outliers
0	carat	True
4	depth	True
5	price	True
6	table	True
7	x	True
8	y	True
9	z	True

Table - 4 cubic_zirconia – outliers

Inference

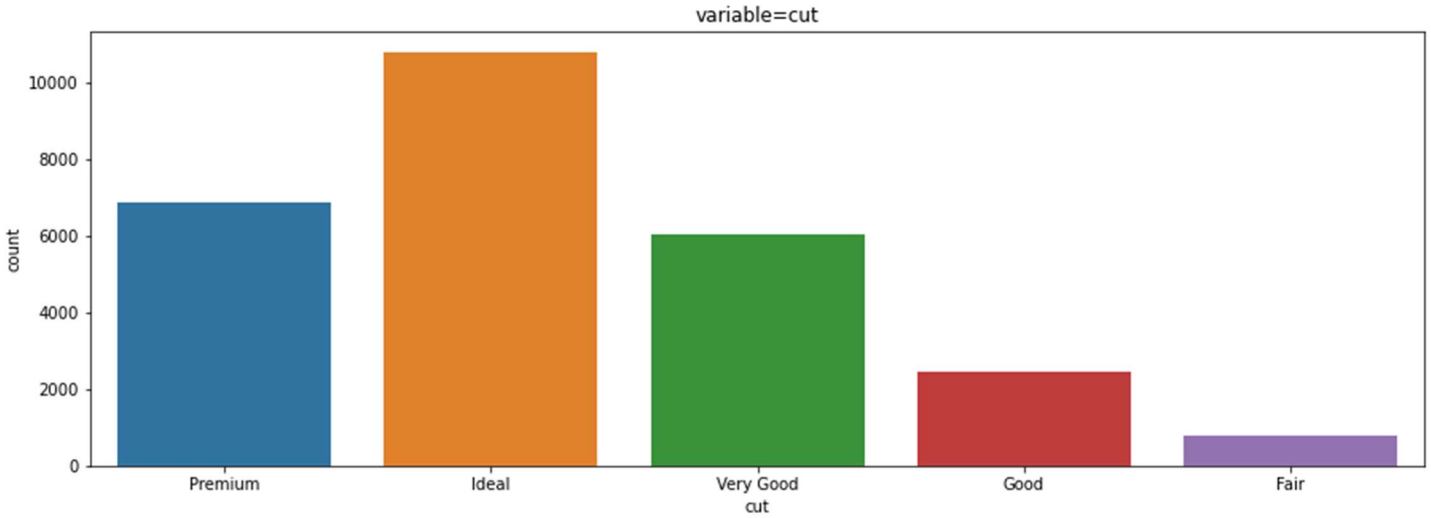
The coefficient of variation (CV) is a statistical measure of the relative dispersion of data points in a data series around the mean.
The coefficient of variation is < 1 for all the variables except the Price. This means that distributions with a coefficient of variation higher than 1 are considered to be high variance whereas those with a CV lower than 1 are considered to be low-variance.
As per the boxplot and outliers check using the IQR method, there are outliers in the dataset. The columns x,y, and z have the minimum value as 0 which needs to be checked further.

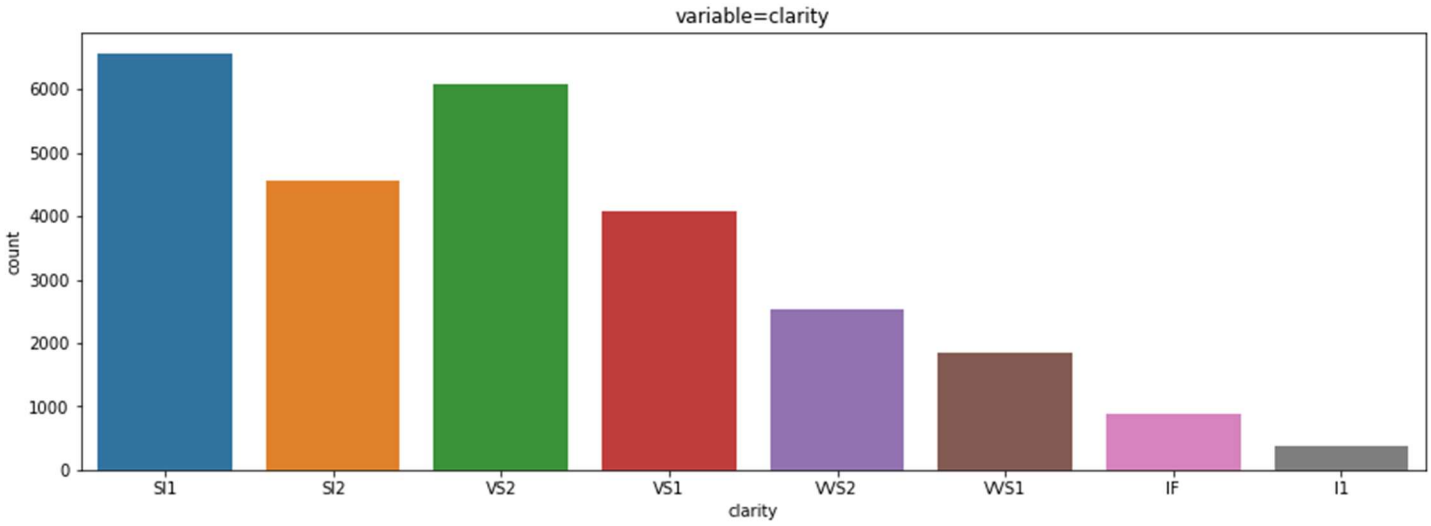
From the Shapiro Wilkin's Test results, the variables except dept are all not normally distributed.

Univariate Analysis for categorical variables

	count	unique	top	freq	%
cut	26933	5	Ideal	10805	40.1181
color	26933	7	G	5653	20.9891
clarity	26933	8	SI1	6565	24.3753

Table - 5 cubic_zirconia - stats for categorical variables





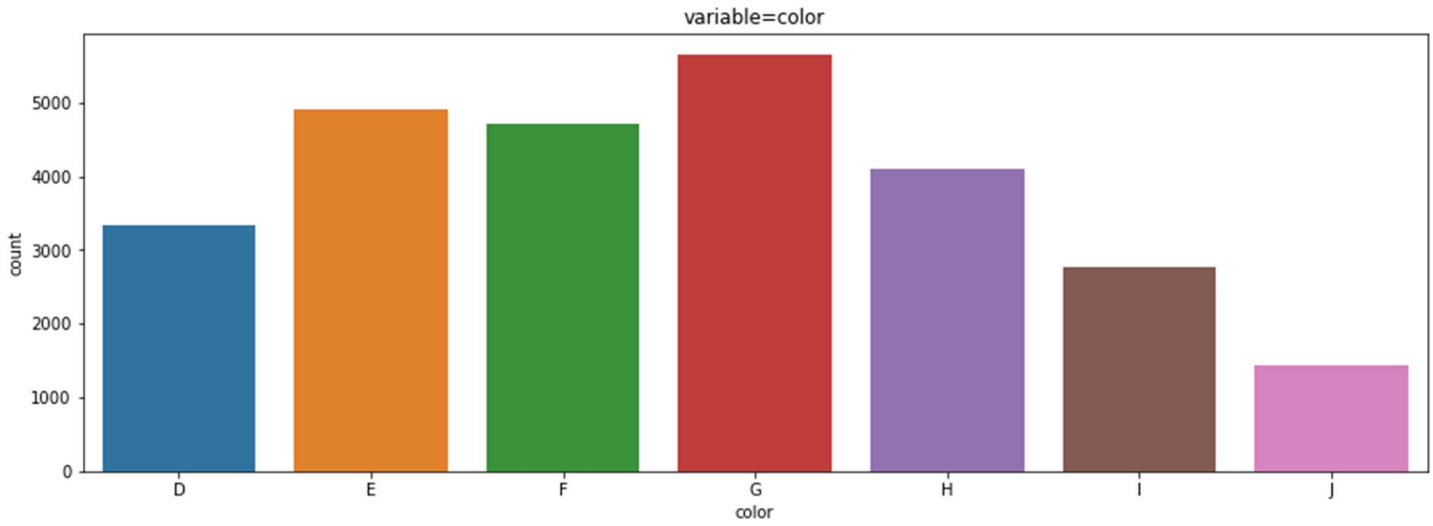


Figure - 2 Univariate Analysis – Categorical

Inference

The quality of a cut ranges from Fair to Ideal. With the 5 unique measures of quality around 10816 i.e., 40.11% of the data are marked as Ideal cut.
The color of the stones with D being the worst and J the best, 5661 of the stones are ranked with G quality which is average. This is 20.99% of the data.
Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, IF = flawless, l1= level 1 inclusion). Most of the stones are marked as SI1 which is 24.37% of the data.

Bivariate Analysis - Numerical and Numerical

`Price` and `carat` are highly correlated. x, y and z which is the dimension of the cubic zirconia also have a significant correlation with the price variable. The depth and table are the only variables which as approx. 0 correlation with the price.

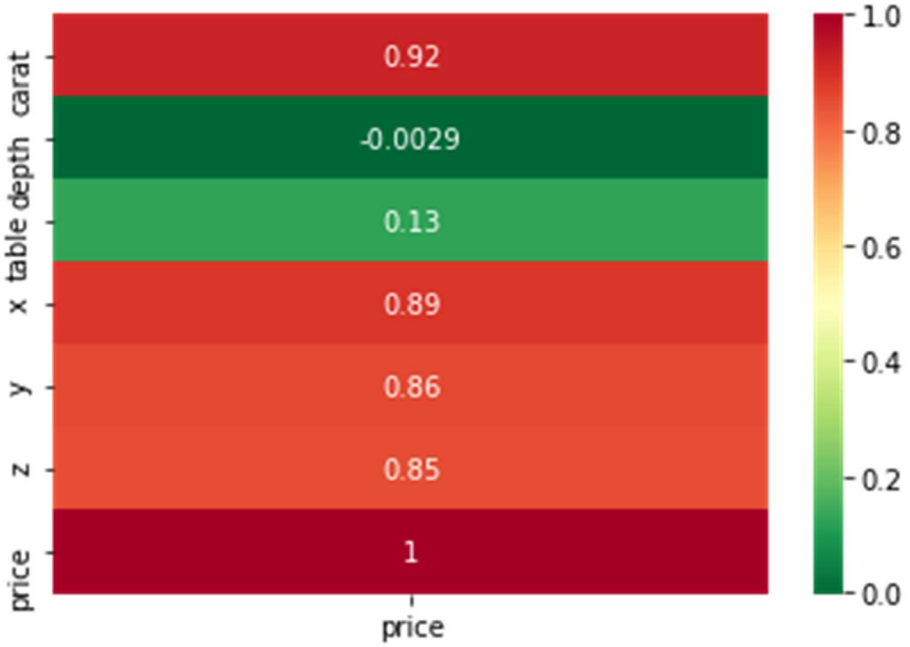
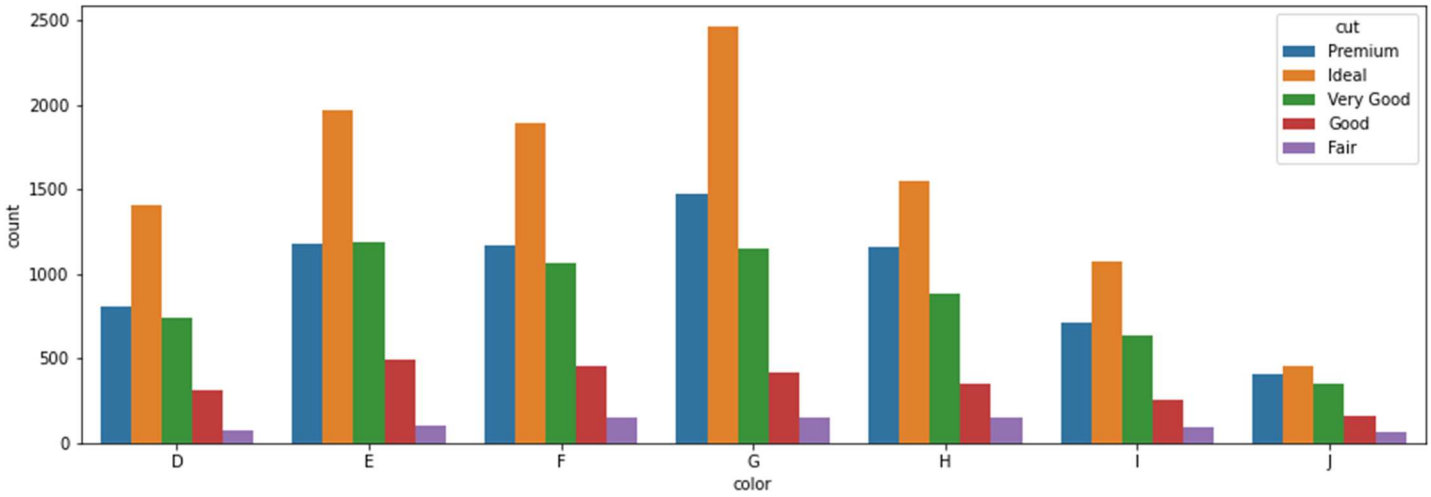


Figure - 3 Bivariate Analysis - Numerical and Numerical

Bivariate Analysis – Categorical and Categorical



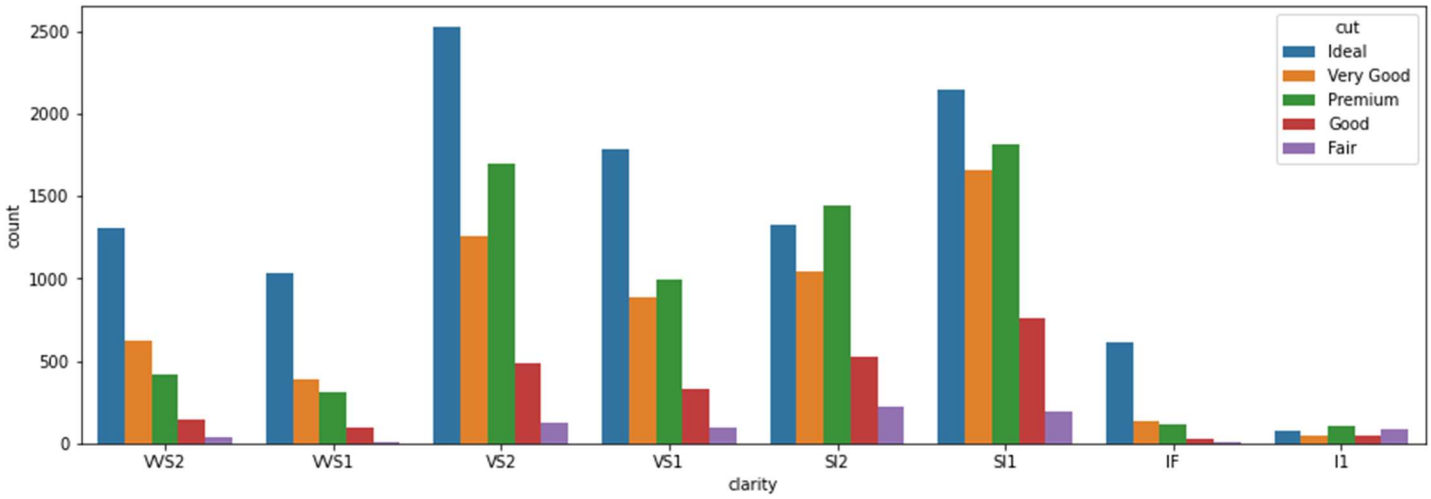


Figure - 4 Bivariate Analysis - Categorical and Categorical

Inferences

Cubic zirconia with ideal cut has highest frequency ranging in all types of color. J which is the best rank in terms of measuring the color has the high frequency of Ideal cut and also for Very Good and Premium cut.
G in color is the average rank which has most of the cut category in Ideal and Premium.
Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, IF = flawless, l1= level 1 inclusion) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1 : VS2 which is the average rank in clarity has most of the cubic zirconia marked as an Ideal or Premium cut.

Bivariate Analysis - Numerical and Categorical

Observation

The price of the cubic zirconia with Premium and Fair cut are higher than the rest.
Colour of the cubic zirconia with D being the worst and J the best. The price shows a positive correlation as the price is low for D rank rises as the rank increases towards J.
In order from Best to Worst, IF = flawless, l1= level 1 inclusion - IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1. The Sl2 has the highest price range. VS1 and SI1 has the second highest price range.

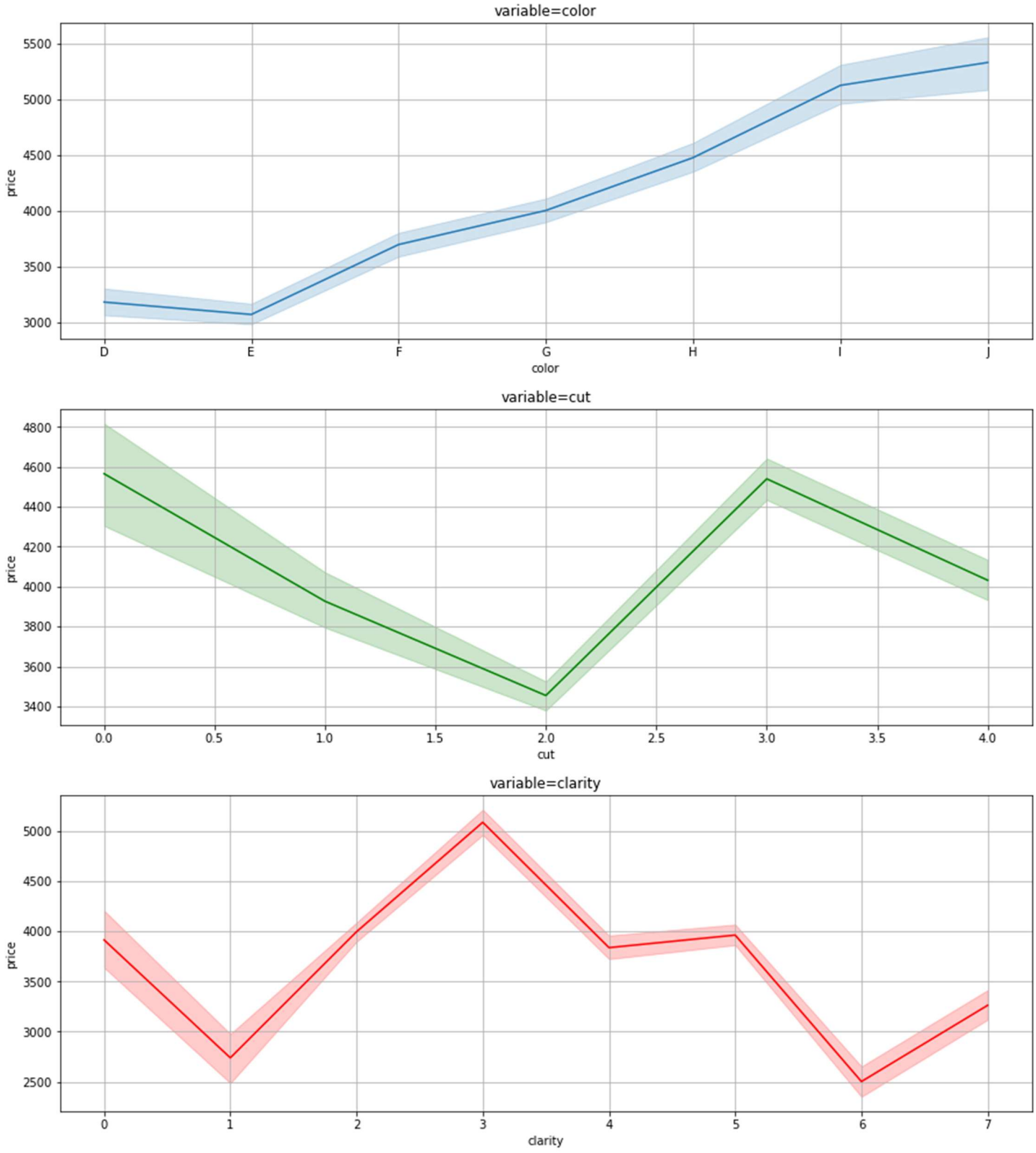


Figure - 5 Bivariate Analysis - Numerical and Categorical

Multivariate Analysis

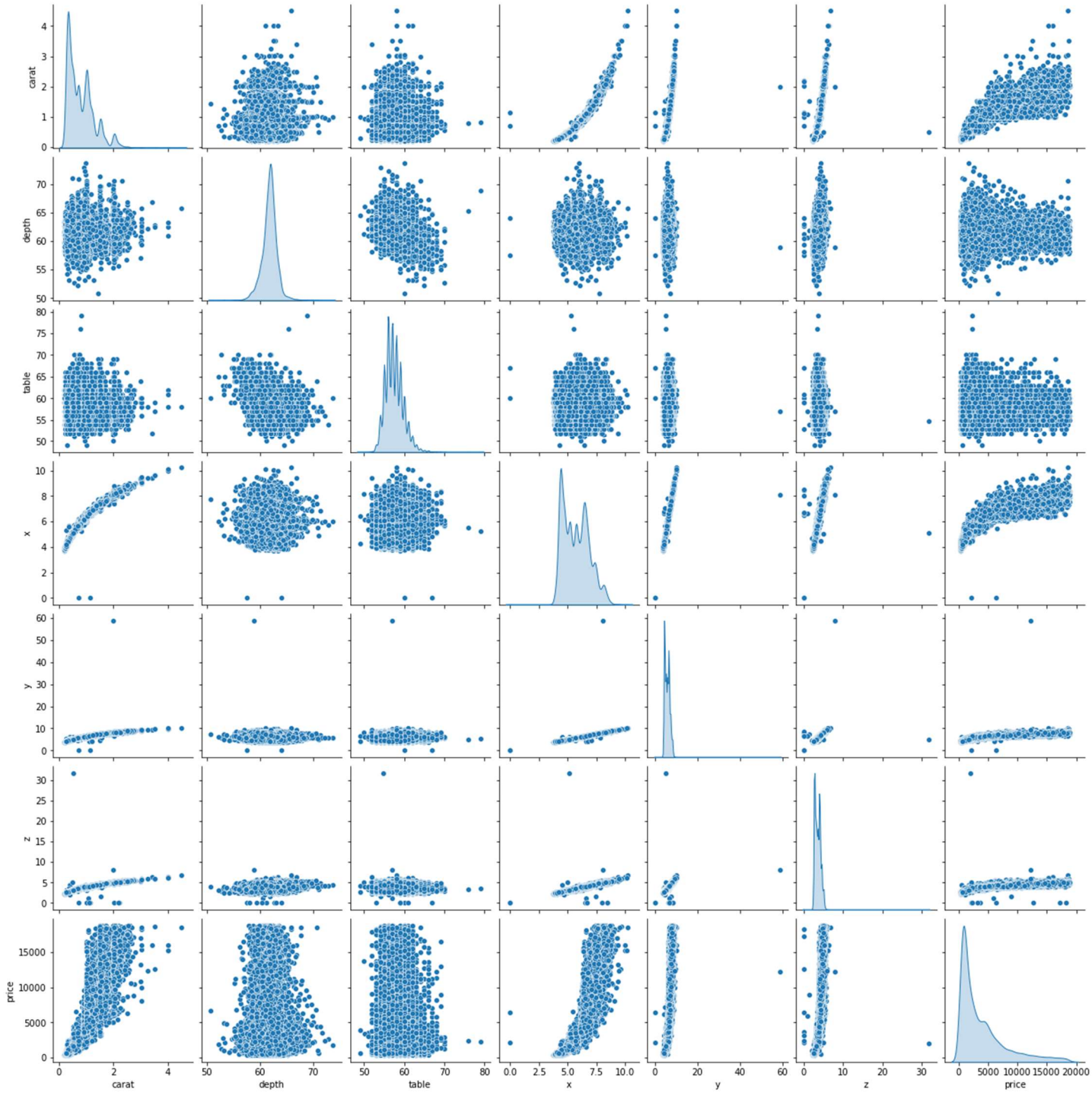


Figure - 6 Multivariate Analysis - pair plot

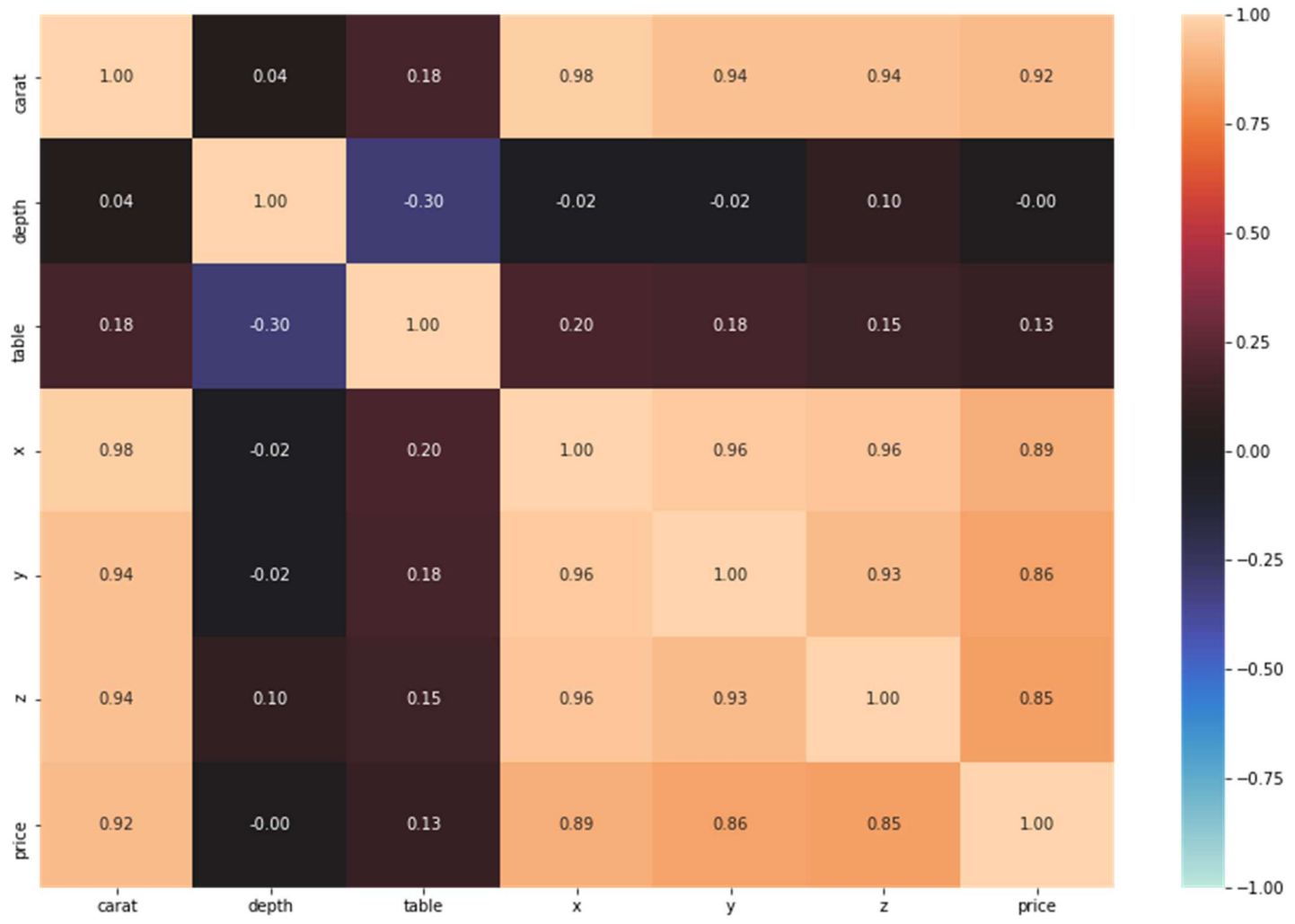


Figure - 7 Multivariate Analysis - heatmap

Inference

The carat, x, y, and z have high correlation with the price variable.
Variable's depth, table have approx. 0 correlation with the price variable.

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?

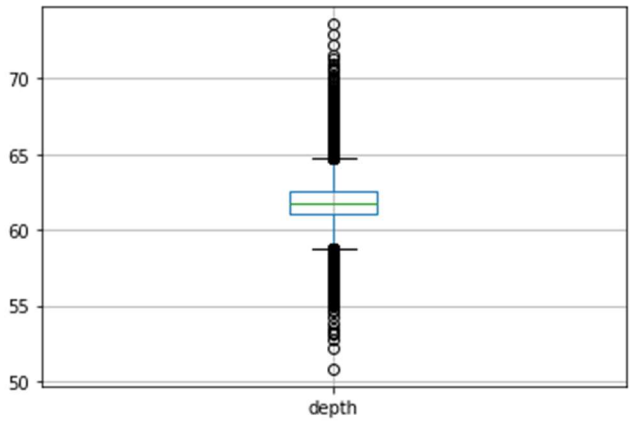
Checking the null values

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

Table - 6 cubic_zirconia – missing values

```
count    26236.000000
mean        61.745285
std          1.412243
min         50.800000
25%         61.000000
50%         61.800000
75%         62.500000
90%         63.300000
95%         63.800000
99%         65.600000
max         73.600000
Name: depth, dtype: float64
```

Table – 7 Stats for cubic_zirconia depth column



Imputing the missing values based on the median or 50th percentile, also known as Q2 (second quantile).

50 percentile ⇨ *61.800000*

Check for values that are zero

	carat	depth	table	x	y	z	price
min	0.2	50.8	49.0	0.0	0.0	0.0	326.0

The values for x, y and z are 0 – since these columns are the dimensions for the cubic zirconia stones, it should not have a zero value.

Table - 6 cubic_zirconia - check for 0 values

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

Table - 7 cubic_zirconia - null values

Total records with either x, y or z is at 0 value or all of them. As the dimensions of the stones cannot be zero, let's drop these 8 records. Total number of rows are 26925 after removing the null values. Total 8 rows are with null values i.e., 0.0297% of the dataset.

Outlier Treatment

```
Skewness in carat = 1.1148092515927273
Skewness in depth = -0.028401722720503546
Skewness in table = 0.7648470275882089
Skewness in x = 0.40198758736963747
Skewness in y = 3.8883904771914586
Skewness in z = 2.6393815543007713
```

Table - 8 Skewness before outlier treatment

```
Skewness in carat = 0.7809486410102108
Skewness in depth = -0.2685426806962745
Skewness in table = 0.3501203974981991
Skewness in x = 0.3894077991186886
Skewness in y = 0.38576851069581153
Skewness in z = 0.38717933491895623
```

Table – 9 Skewness after outlier treatment

Treatment of outliers through Winsorizing

Winsorizing or winsorization is the transformation of statistics by limiting extreme values in the statistical data to reduce the effect of possibly spurious outliers.

Scaling

In regression analysis, we can calculate importance of variables by ranking independent variables based on the descending order of absolute value of

standardized coefficient. It is also helpful to standardize a variable when an interaction is created from two variables that are not centred on 0, as some amount of collinearity will be induced.
Also, it is necessary to standardize variables before using Lasso and Ridge Regression as both have constraints on the size of the coefficients associated to each variable. The result of cantering the variables means that there is no longer an intercept.

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

Encode of the data

Encoding the categorical columns into the codes:

	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.30	2	1	2	62.1	58.0	4.27	4.29	2.66	499
1	0.33	3	3	1	60.8	58.0	4.42	4.46	2.70	984
2	0.90	4	1	7	62.2	60.0	6.04	6.12	3.78	6289
3	0.42	2	2	4	61.6	56.0	4.82	4.80	2.96	1082
4	0.31	2	2	6	60.4	59.0	4.35	4.43	2.65	779

Table - 8 cubic_zirconia Encoded

Scaling the X dataset

	carat	cut	color	clarity	depth	table	x	y	z
count	26,925.0000	26,925.0000	26,925.0000	26,925.0000	26,925.0000	26,925.0000	26,925.0000	26,925.0000	26,925.0000
mean	-0.0000	-0.0000	0.0000	-0.0000	-0.0000	0.0000	-0.0000	0.0000	-0.0000
std	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
min	-1.3194	-2.4945	-1.5269	-2.2231	-2.3617	-2.7788	-1.7785	-1.8109	-2.8908
25%	-0.8686	-0.5418	-0.9408	-1.0637	-0.5539	-0.6733	-0.9064	-0.9146	-0.9177
50%	-0.1926	-0.5418	0.2315	0.0958	0.0487	-0.1948	-0.0342	-0.0273	-0.0248
75%	0.5962	0.4345	0.8177	0.6756	0.6513	0.7622	0.7311	0.7256	0.7241
max	2.7821	1.4109	1.9900	1.8350	2.4592	2.8199	3.1784	3.1636	3.1869

Table - 9 Scaling of the X data (cubic_zirconia)

Splitting the data into 70:30 ratio

```
X_train data shape = (18847, 9)
y_train data shape = (18847,)
X_test data shape = (8078, 9)
y_test data shape = (8078,)
```

Applying the Linear Regression Model from sklearn library

LinearRegression()

Coefficient:

The coefficient for carat is 5328.730260501825

The coefficient for cut is 75.20551056111245

The coefficient for color is -422.0041699920125

The coefficient for clarity is 486.6236780902803

The coefficient for depth is -45.8938876243742

The coefficient for table is -230.57068703638157

The coefficient for x is -2274.8667253801395

The coefficient for y is 2337.4873109209593

The coefficient for z is -1480.5675505967884

Intercept

The intercept for Linear model is 3940.020772707241

Performance Metrics for train data:

MAE *– measures the average magnitude of errors in a set of predictions, without considering their directions.*

$$MAE = \frac{\sum_{i=1}^{n} |y_i - y_i^p|}{n}$$

MSE *– the sum of squared distances between our target variable and predicted values.*

$$MSE = \frac{\sum_{i=1}^{n} (y_i - y_i^p)^2}{n}$$

RMSE *–the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are*

$$RMSE = SD_y \sqrt{(1 - r^2)}$$

R2-square/ R-squared *- statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.*

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

	Model	MAE	MSE	RMSE	R2 Square
0	Linear Regression	957.3434	2,159,238.0703	1,469.4346	0.8660

Table - 10 Linear Regression - train

Performance Metrics for test data:

	Model	MAE	MSE	RMSE	R2 Square
0	Linear Regression	956.3598	2,119,943.8707	1,456.0027	0.8699

Table - 11 Linear Regression – Test

Applying the Lasso Regression from sklearn library

```
Lasso(max_iter=10000, random_state=42)
```

Coefficient

```
The coefficient for carat is 5265.81624995657

The coefficient for cut is 79.73583768368432

The coefficient for color is -420.51371585013266

The coefficient for clarity is 488.14340512964077

The coefficient for depth is -67.07245389077376

The coefficient for table is -233.62466297234403

The coefficient for x is -1607.689245125138

The coefficient for y is 1526.2234812646618
```

```
The coefficient for z is -1273.12498975702
```

Intercept

```
The intercept for Linear model is 3939.9275784025954
```

Applying the Ridge Regression from sklearn library

```
Ridge(random_state=42)
```

Coefficient

```
The coefficient for carat is 5316.424513944125

The coefficient for cut is 75.85943249094453

The coefficient for color is -421.8543898421464

The coefficient for clarity is 486.94206101584916

The coefficient for depth is -47.91956643108293

The coefficient for table is -231.04817594127147

The coefficient for x is -2191.3349562629965

The coefficient for y is 2244.8534253672624

The coefficient for z is -1459.1499522412698
```

Intercept

```
The intercept for Linear model is 3940.0093481305803
```

Applying the Stochastic Gradient Descent Regression from sklearn library

Using the Grid Search CV

```
GridSearchCV(cv=10, estimator=SGDRegressor(max_iter=10000),
             param_grid={'penalty': ['l2', 'elasticnet'],
                         'tol': [0.0001, 1e-05]})
```

Best Params:

```
SGDRegressor(max_iter=10000, penalty='elasticnet', tol=1e-05)
```

Performance in train dataset for all the models

	Model	MAE	MSE	RMSE	R2 Square
0	Linear Regression	957.3434	2,159,238.0703	1,469.4346	0.8660
1	Lasso Regression	957.9525	2,160,967.2883	1,470.0229	0.8659
2	Ridge Regression	957.5281	2,159,263.9269	1,469.4434	0.8660
3	SGD Regression	952.6697	2,169,296.4408	1,472.8532	0.8654

Table - 12 Model comparison-train

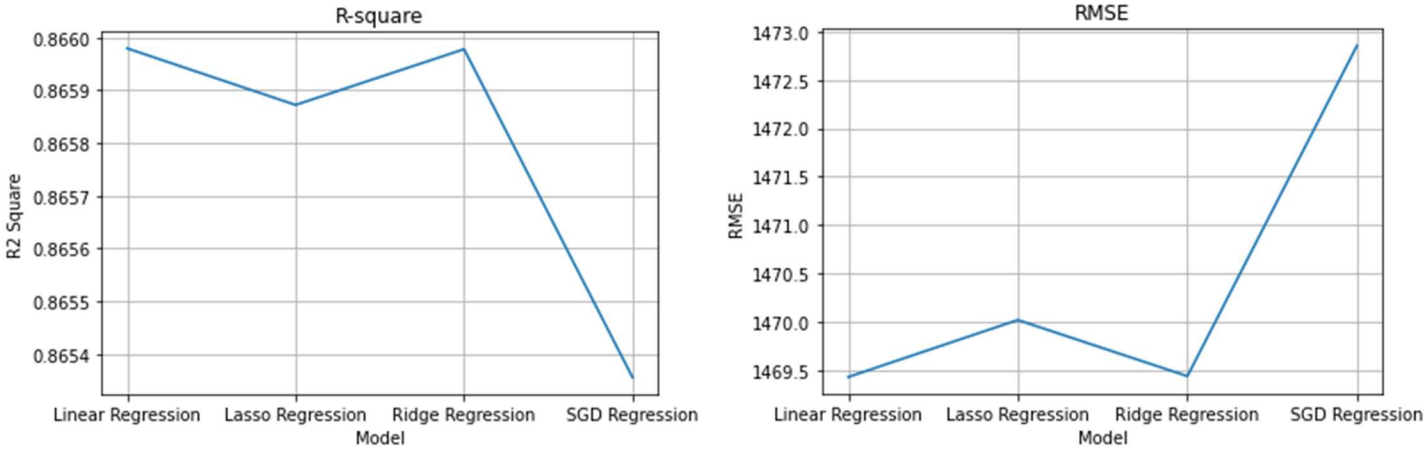
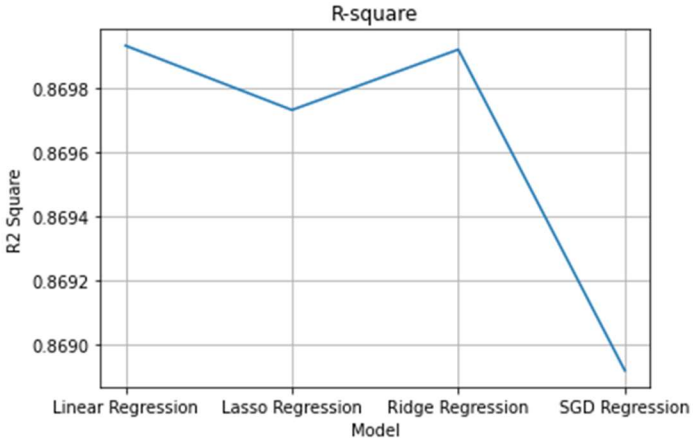


Figure - 8 Performance Metrics - train

Performance in test dataset for all the models

	Model	MAE	MSE	RMSE	R2 Square
0	Linear Regression	956.3598	2,119,943.8707	1,456.0027	0.8699
1	Lasso Regression	956.5790	2,123,209.1835	1,457.1236	0.8697
2	Ridge Regression	956.5374	2,120,141.1982	1,456.0705	0.8699
3	SGD Regression	951.8377	2,136,434.5898	1,461.6547	0.8689

Table - 13 Model Comparison – test



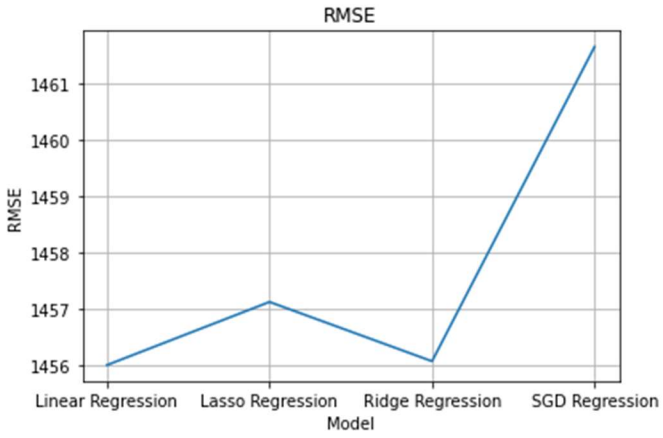


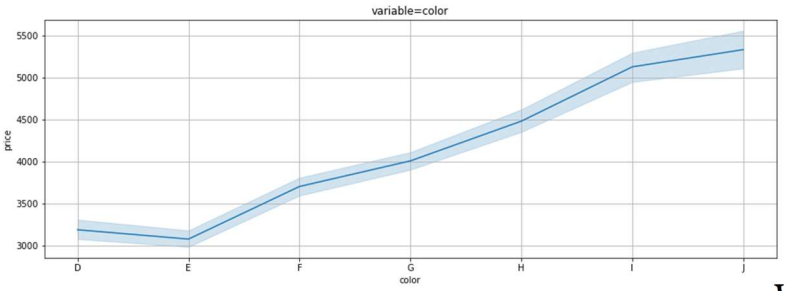
Figure - 9 Performance Metrics – Test

Based on the various model and its performances on train and test dataset, it can be concluded that the best model is Linear Regression model as it has the highest r-squared of 86.99% and RMSE of 1,456.0027 in the test dataset and 86.60% r-squared with RMSE of 1,469.4346 in train dataset.

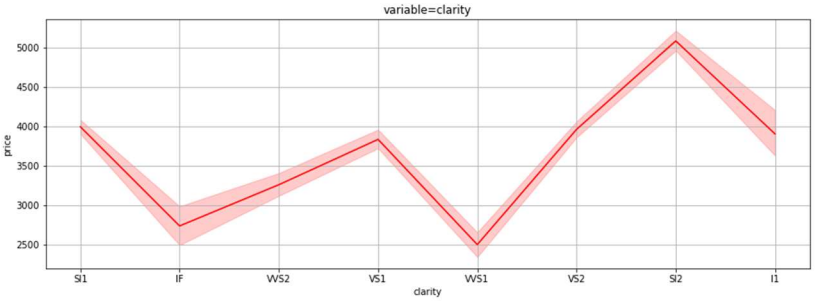
1.4. Inference: Basis on these predictions, what are the business insights and recommendations.

Insights and Recommendation

Based on the observation the price is higher when carat is more than 1.6, cut is marked as 'ideal' or 'premium' and color is more or equal to average rank (>=G rank) and when table is more than 50% and depth more than 60.



In order from Best to Worst, IF = flawless, l1= level 1 inclusion - IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1. The Sl2 has the highest price range. VS1 and Sl1 has the second highest price range.



*Based on **cut, color, clarity, depth and carat** the cubic zirconia (CZ) stones can distinguish between higher profitable stones and lower profitable stones.*

	depth	table
count	2693.000000	2693.000000
mean	61.663089	57.819012
std	1.212051	2.015562
min	59.000000	51.600000
25%	60.900000	56.000000
50%	61.800000	58.000000
75%	62.500000	59.000000
max	64.600000	63.000000

Problem 2: Logistic Regression and LDA

A tour and travel agency which deals in selling holiday packages has provided details of 872 employees of a company. Among these employees, some opted for the package and some didn't. Assignment is 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.

Reading the data

	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

Initial checks

```
Info of the data
----------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Unnamed: 0         872 non-null    int64
 1   Holliday_Package   872 non-null    object
 2   Salary             872 non-null    int64
 3   age                872 non-null    int64
 4   educ               872 non-null    int64
 5   no_young_children  872 non-null    int64
 6   no_older_children  872 non-null    int64
 7   foreign            872 non-null    object
dtypes: int64(6), object(2)
memory usage: 54.6+ KB
None

Dimension check
---------------
Total number of rows in the dataset = 872
Total number of columns in the dataset = 8
Total number of elements in the dataset = 6976

Missing values check
--------------------
There are no missing values in the data.

Duplicate Values Check
----------------------
Number of duplicate rows = 0
There are no duplicate values in the dataset.
```

Inference

There are total of 872 records in the Holiday Package dataset with total 8 columns.
This means that the total elements in the dataset are 6976.
There are no missing or duplicate values in the data however, this needs to be further checked after removing the Unnamed: 0 column.
There are 6 integer columns and 2 object columns.

```
Duplicate Values Check after dropping the Unnamed:0 column
----------------------------------------------------------
Number of duplicate rows = 0
There are no duplicate values in the dataset.
```

Null Value Check

	count	mean	std	min	25%	50%	75%	max	cv
Salary	872.0000	47,729.1720	23,418.6685	1,322.0000	35,324.0000	41,903.5000	53,469.5000	236,961.0000	0.4907
age	872.0000	39.9553	10.5517	20.0000	32.0000	39.0000	48.0000	62.0000	0.2641
educ	872.0000	9.3073	3.0363	1.0000	8.0000	9.0000	12.0000	21.0000	0.3262
no_young_children	872.0000	0.3119	0.6129	0.0000	0.0000	0.0000	0.0000	3.0000	1.9648
no_older_children	872.0000	0.9828	1.0868	0.0000	0.0000	1.0000	2.0000	6.0000	1.1058

The null value is in no young children and no older children, it can be 0 if the employees does not have any children and hence this should not be imputed with any other value.

The numerical columns age and salary can be binned together to create the object variable in classes. By doing so we are creating a dummy variable derived from the existing independent variable:

Stats for age variable

```
count   872.0000
mean     39.9553
std      10.5517
min      20.0000
25%      32.0000
50%      39.0000
75%      48.0000
max      62.0000
Name: age, dtype: float64
<AxesSubplot:>
```

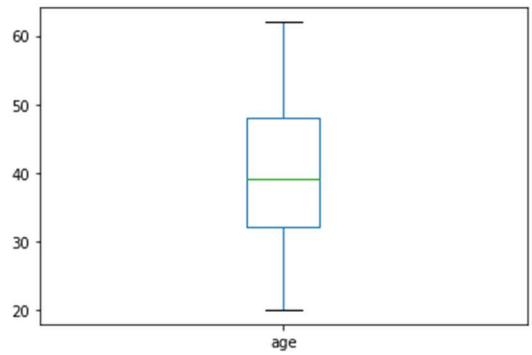


Figure - 10 Holiday data - Boxplot - age

From the age column we can see that the minimum age is 20 and the maximum age is 62. For the age variable we can create a bins for the employees in different age group.

Stats for Salary

```
count       872.0000
mean     47,729.1720
std      23,418.6685
min       1,322.0000
25%      35,324.0000
50%      41,903.5000
75%      53,469.5000
max     236,961.0000
Name: Salary, dtype: float64
<AxesSubplot:>
```

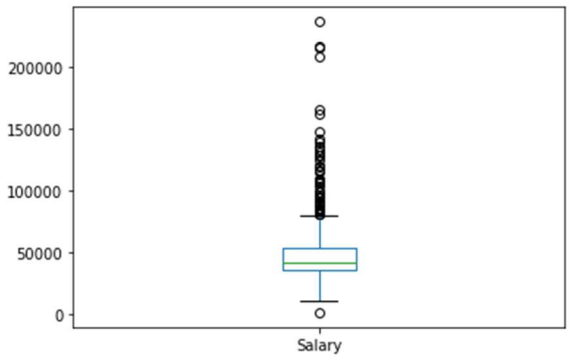


Figure - 11 Holiday data - Boxplot - salary

The minimum salary is 1,322 and the maximum salary of an employee is 236,961. From this data we can create a new variable salary group which bins the salary from various levels.

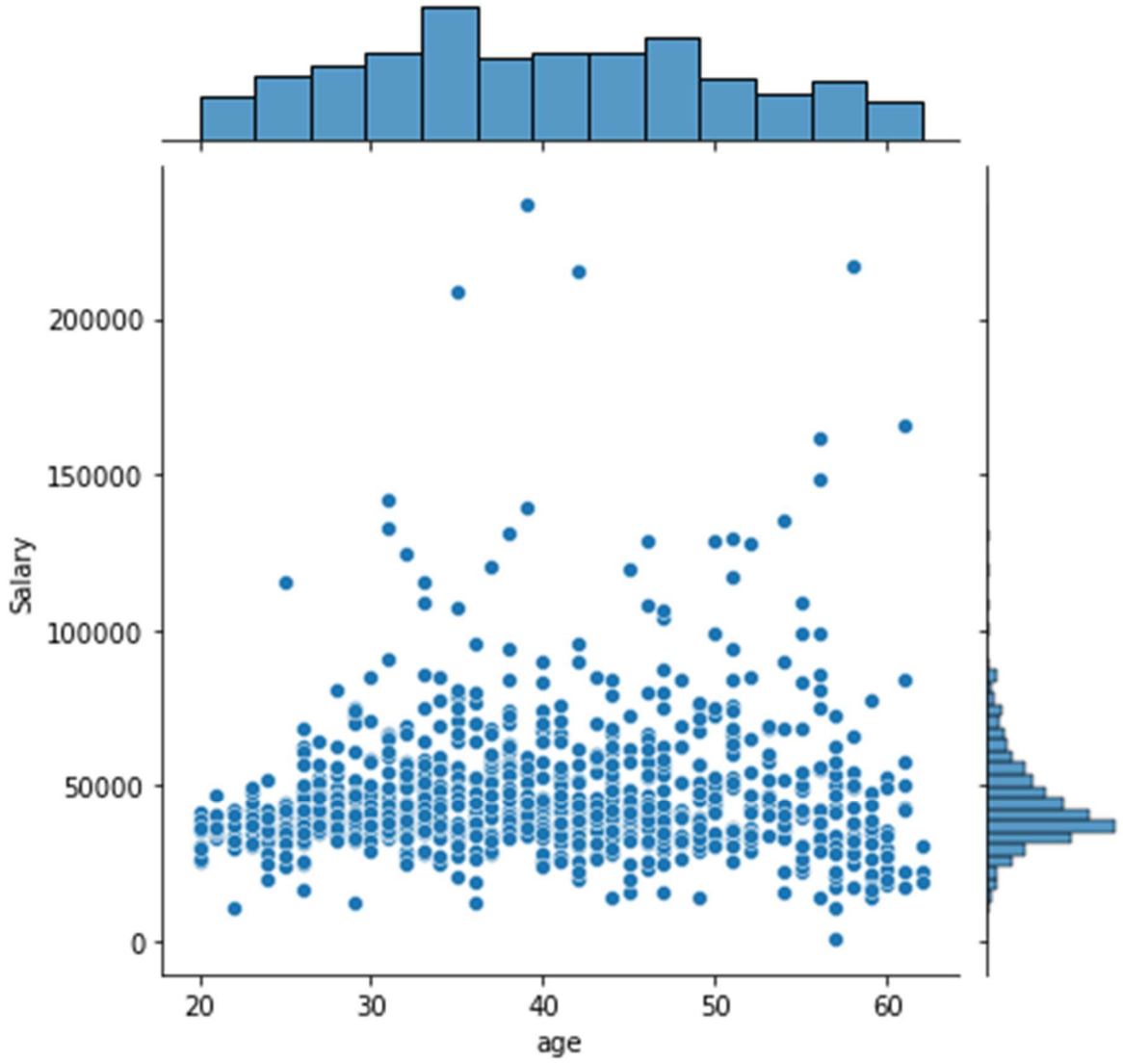


Figure - 12 Holiday data - Salary and age

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign	salary_group	age_group
0	no	48412	30	8	1	1	no	30k-60k	0-30
1	yes	37207	45	8	0	1	no	30k-60k	40-50
2	no	58022	46	9	0	0	no	30k-60k	40-50
3	no	66503	31	11	2	0	no	60k-90k	30-40
4	no	66734	44	12	0	2	no	60k-90k	40-50

Table - 14 Holiday - salary and age group

Exploratory Data Analysis

Univariate Analysis – Numerical

	count	mean	std	min	25%	50%	75%	max	cv
Salary	872.0000	47,729.1720	23,418.6685	1,322.0000	35,324.0000	41,903.5000	53,469.5000	236,961.0000	0.4907
age	872.0000	39.9553	10.5517	20.0000	32.0000	39.0000	48.0000	62.0000	0.2641
educ	872.0000	9.3073	3.0363	1.0000	8.0000	9.0000	12.0000	21.0000	0.3262
no_young_children	872.0000	0.3119	0.6129	0.0000	0.0000	0.0000	0.0000	3.0000	1.9648
no_older_children	872.0000	0.9828	1.0868	0.0000	0.0000	1.0000	2.0000	6.0000	1.1058

Table - 15 Holiday data - Stats for numerical data

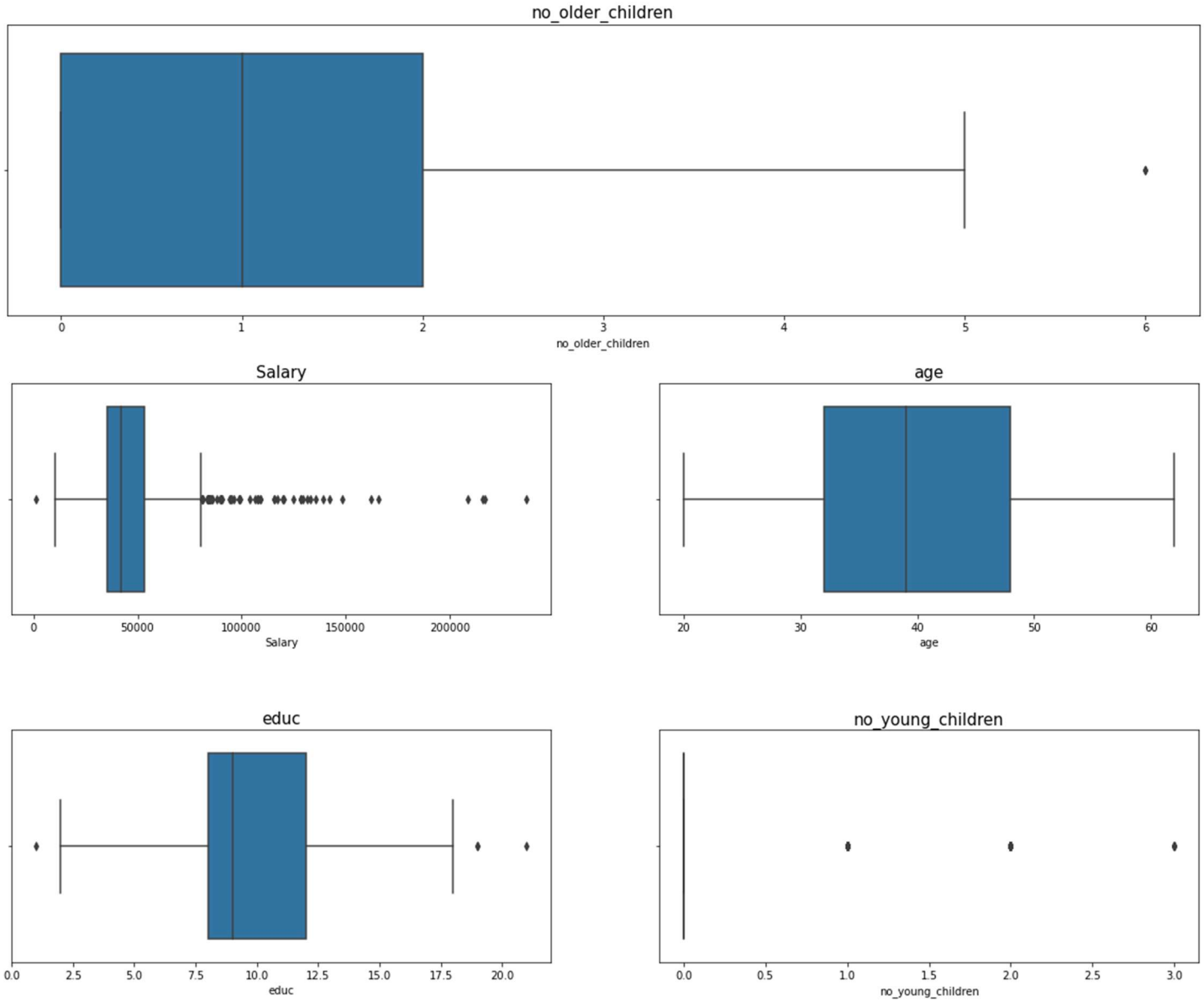


Figure - 13 Holiday - Univariate – Numerical

Shapiro-Wilk test for normality

```
------------------------------------------------------------
pvalue for Salary column is 0.0,
The data is not normally distributed.

pvalue for age column is 0.0,
The data is not normally distributed.

pvalue for educ column is 0.0,
The data is not normally distributed.

pvalue for no_young_children column is 0.0,
The data is not normally distributed.

pvalue for no_older_children column is 0.0,
The data is not normally distributed.
```

Detecting outliers using IQR

```
-------------------------------
```

	Columns	Outliers
1	Salary	True
4	educ	True
6	no_older_children	True
7	no_young_children	True

Table - 16 Holiday - Outliers

Inference

The data is not normally distributed for the variables and there are outliers for Salary, education, no_older_children and no_younger_children.
The coefficient of variance is higher in no_older_children and no_younger_children which means it has high variance.

Univariate Analysis – Categorical

	count	unique	top	freq	%
Holliday_Package	872	2	no	471	54.0138
foreign	872	2	no	656	75.2294
salary_group	872	8	30k-60k	625	71.6743
age_group	872	5	30-40	283	32.4541

Table - 17 Holiday - stats for categorical

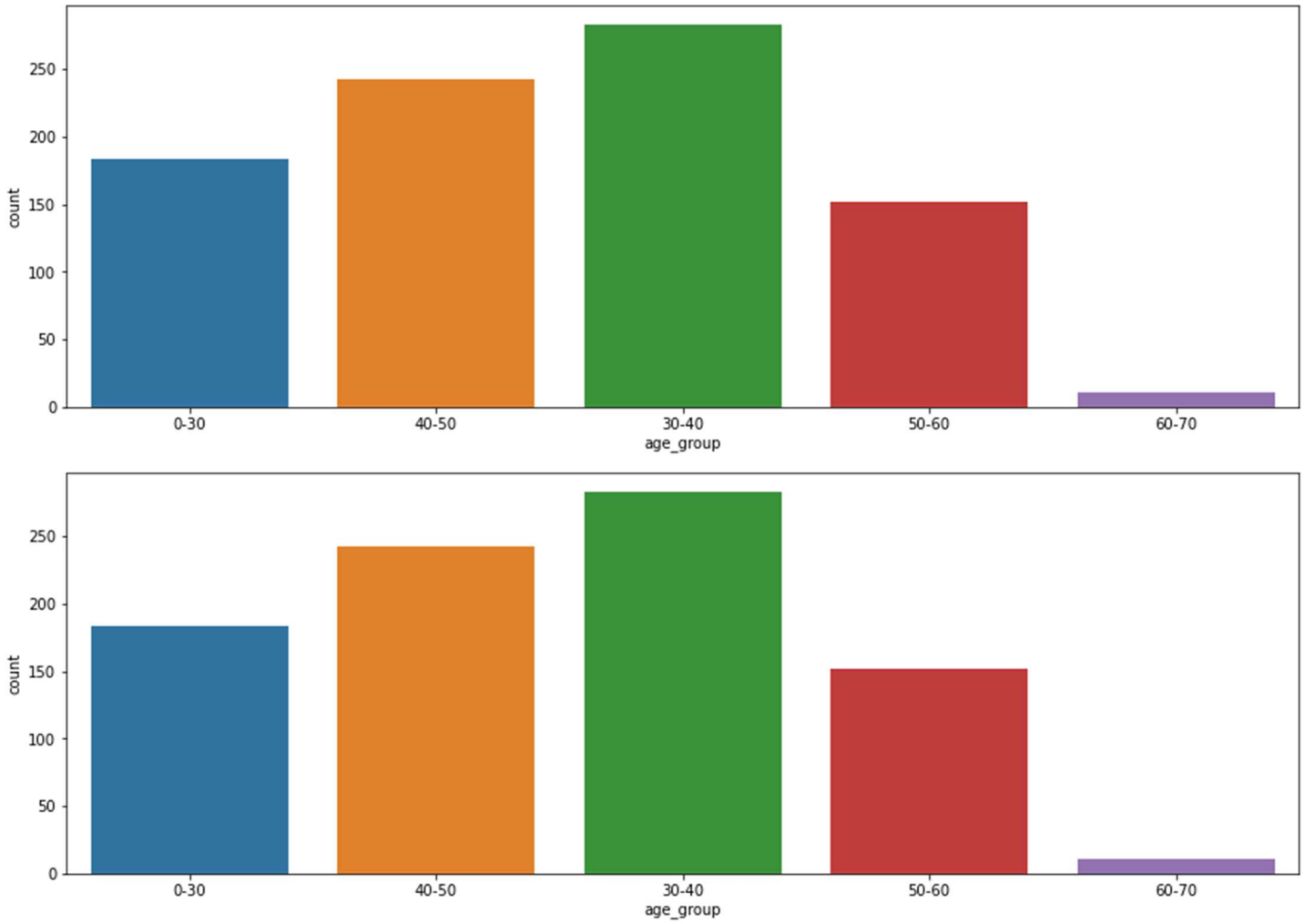


Figure - 14 Holiday - Univariate - Categorical

```
Holliday_Package
--------------
 no     54.0138
yes    45.9862
Name: Holliday_Package, dtype: float64
```

```
foreign
--------------
 no     75.2294
yes    24.7706
Name: foreign, dtype: float64
```

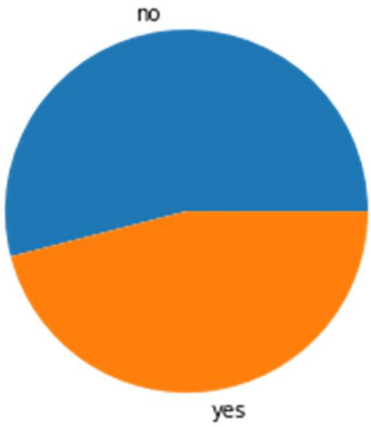


Figure - 16 Holiday_Package

Figure - 15 Foreign Y/N

Observation

Most of the employees lies in the 30-40 age group and most of the employees lies in the 30k to 60k salary bin. Out of total employees 75.22% are residents and not foreigner. Only 45.98% of employees opt for the holiday package.

Bivariate Analysis – Numerical and Numerical

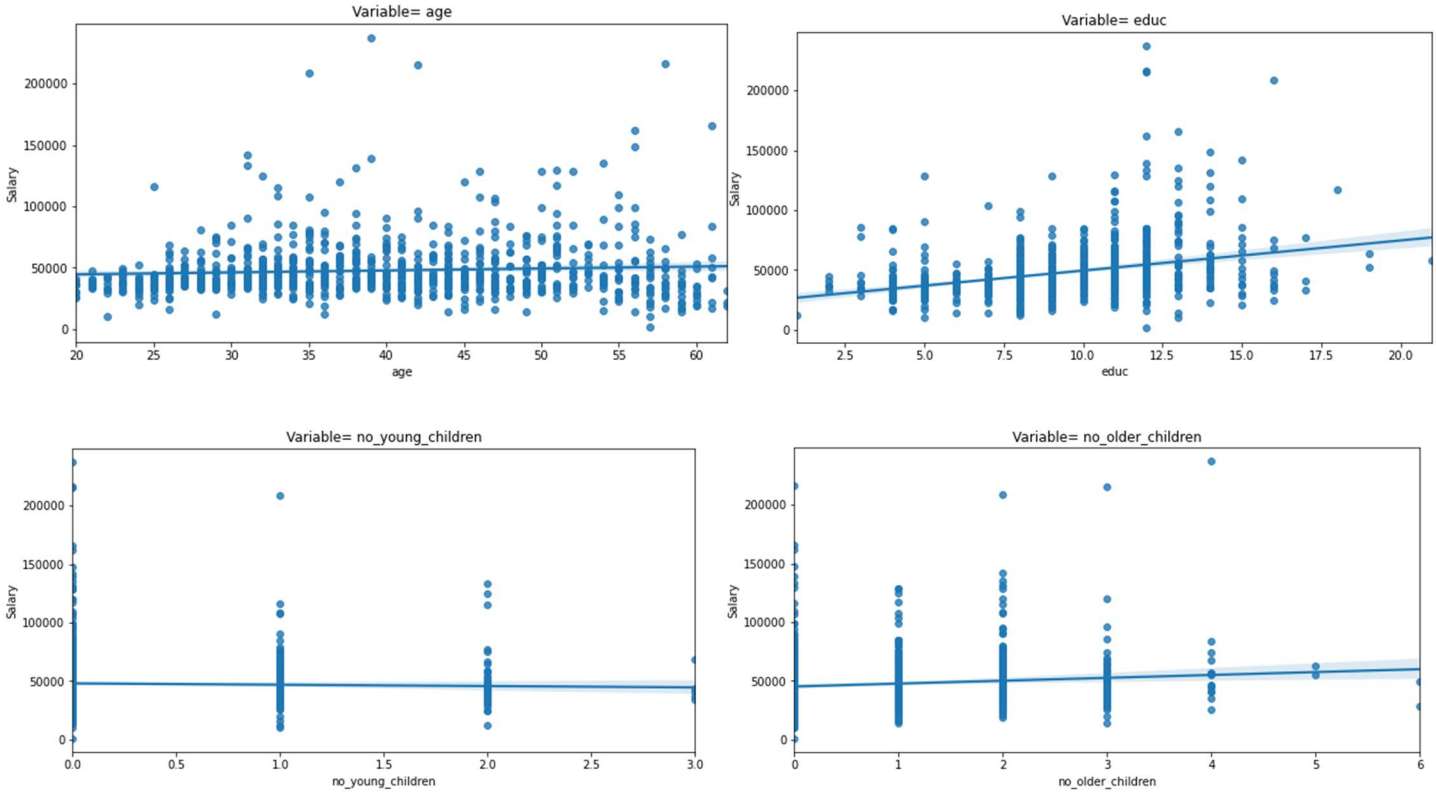
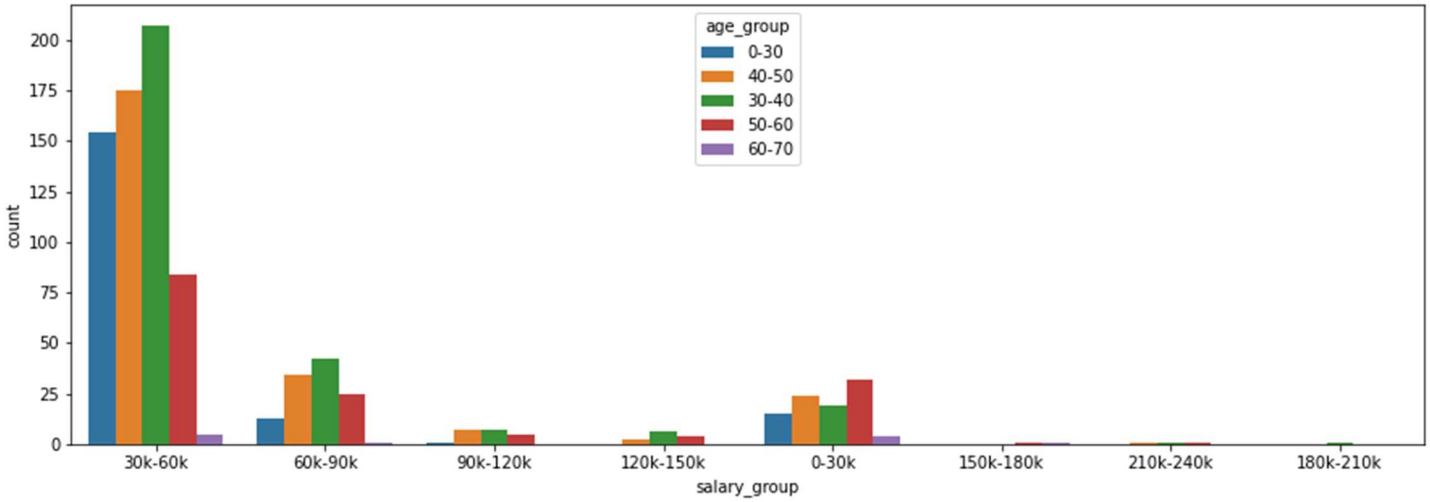
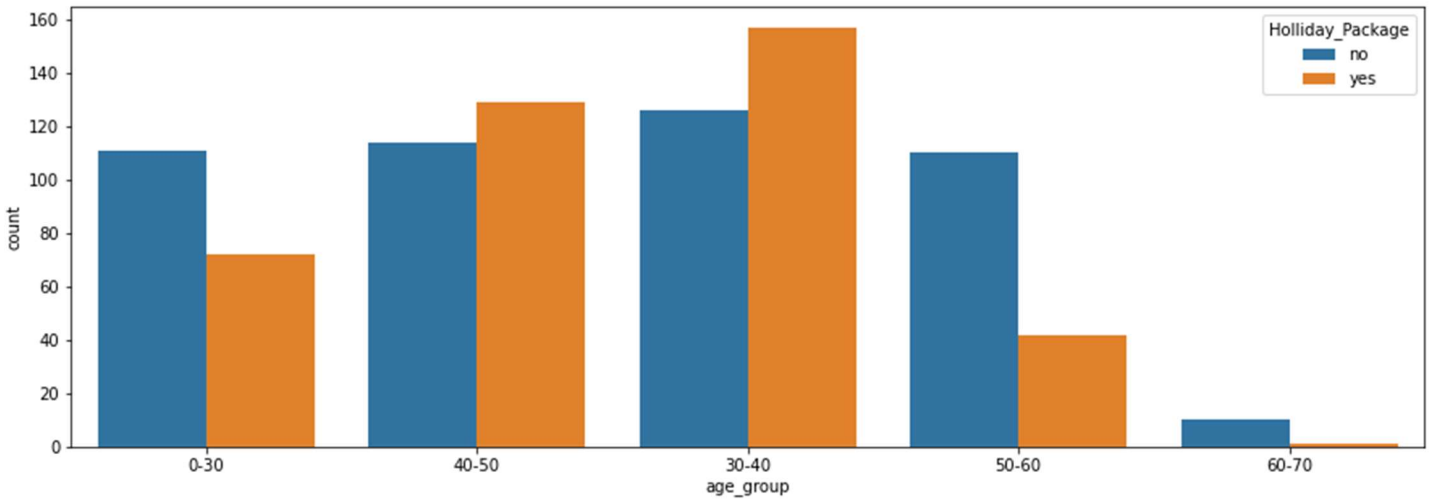
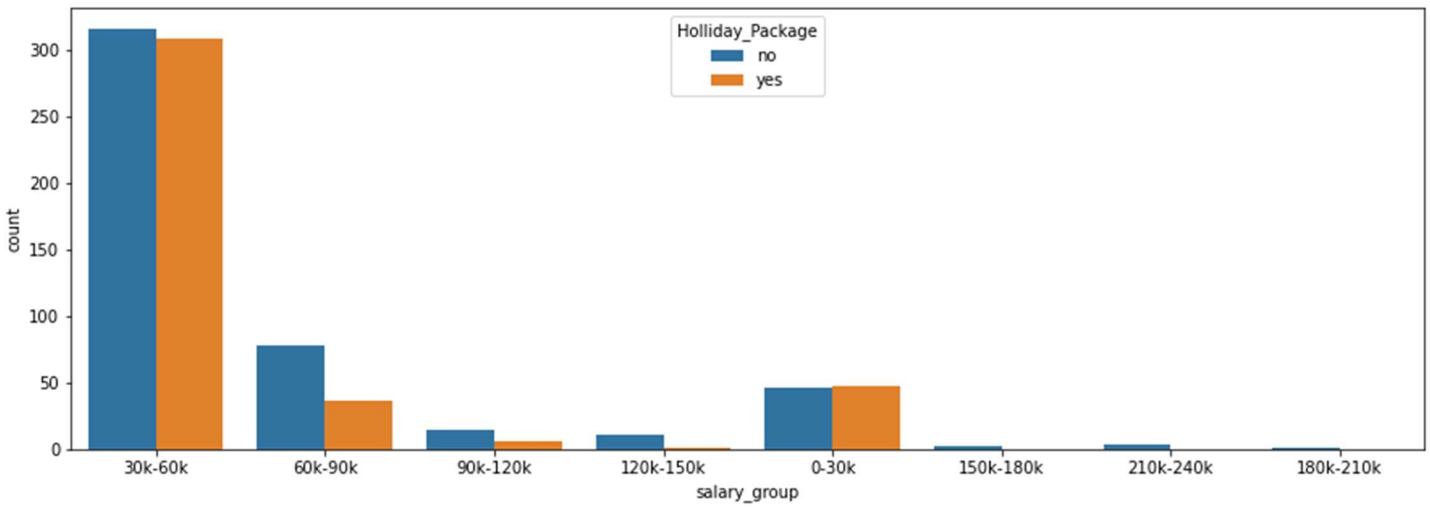


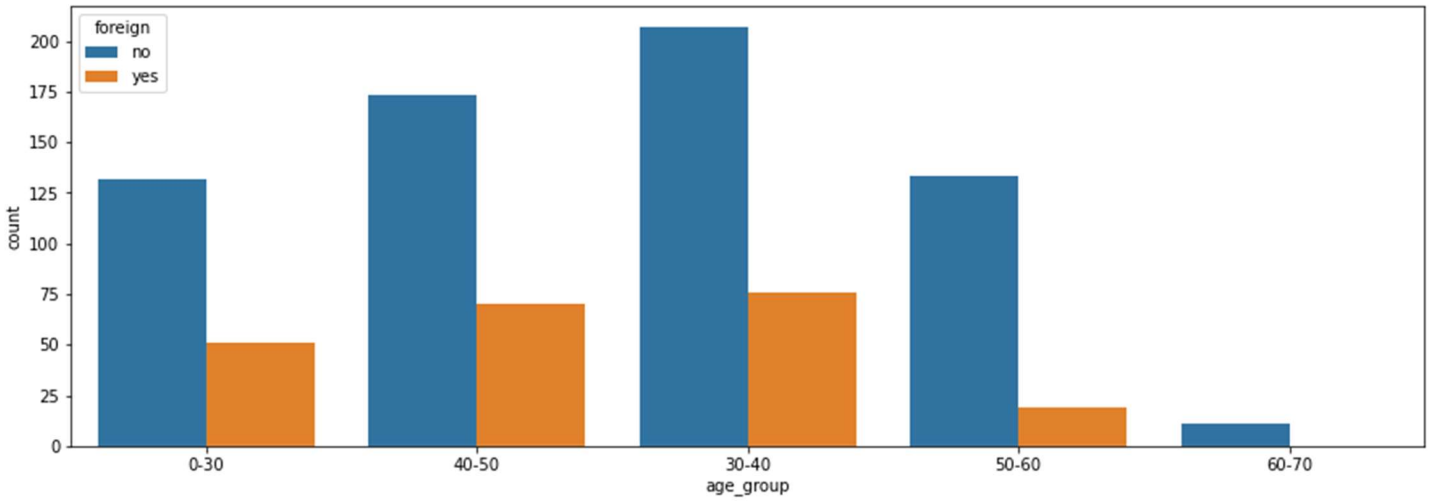
Figure - 17 Holiday - Bivariate - Numerical

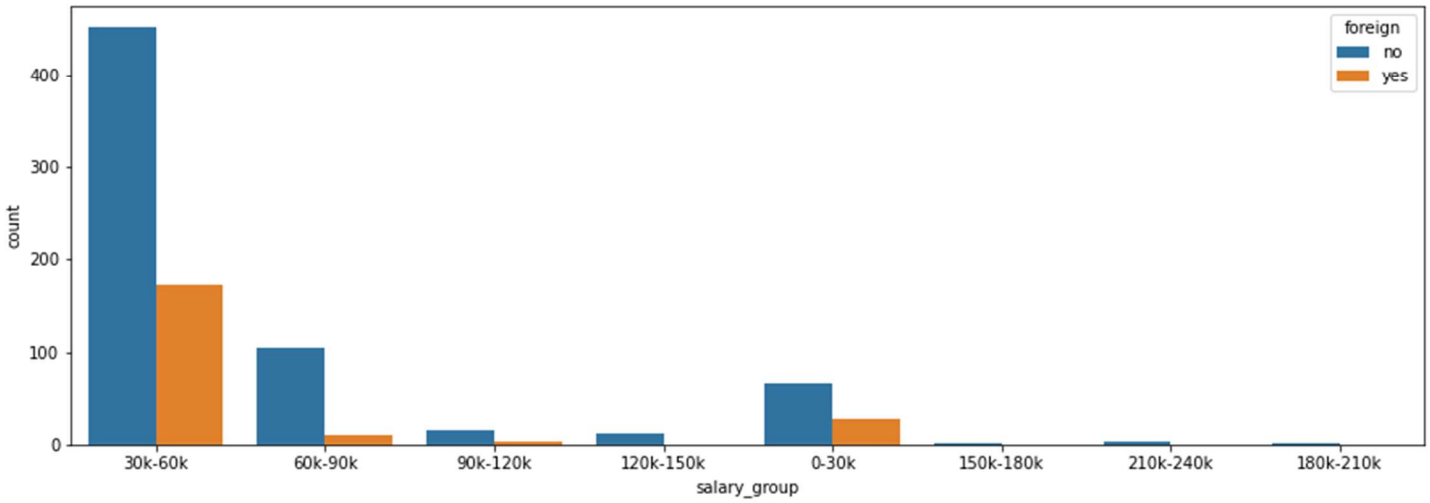
Between numerical variables, salary is more centered around 50,000. There seems to be an upward trend however as the education level increases as well as age. Employees with no older children shows a positive trend in salary while employees with no younger children is almost nonlinear.

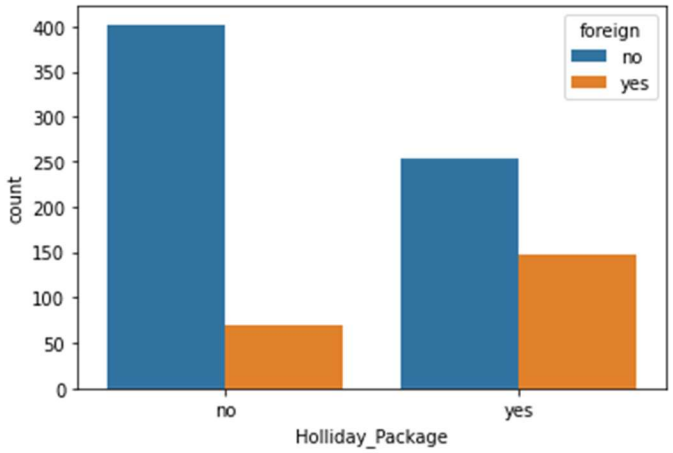












Inference

The employees in the age of 30-40 earns around 30k to 60k. Employees with age 30 to 45+ opt for the holiday package where salary is maximum 60k. As most of the employees are not foreigner the focus should be for the employees who are native.

Figure - 18 Holiday - Bivariate - Categorical

Bivariate – Numerical and Categorical

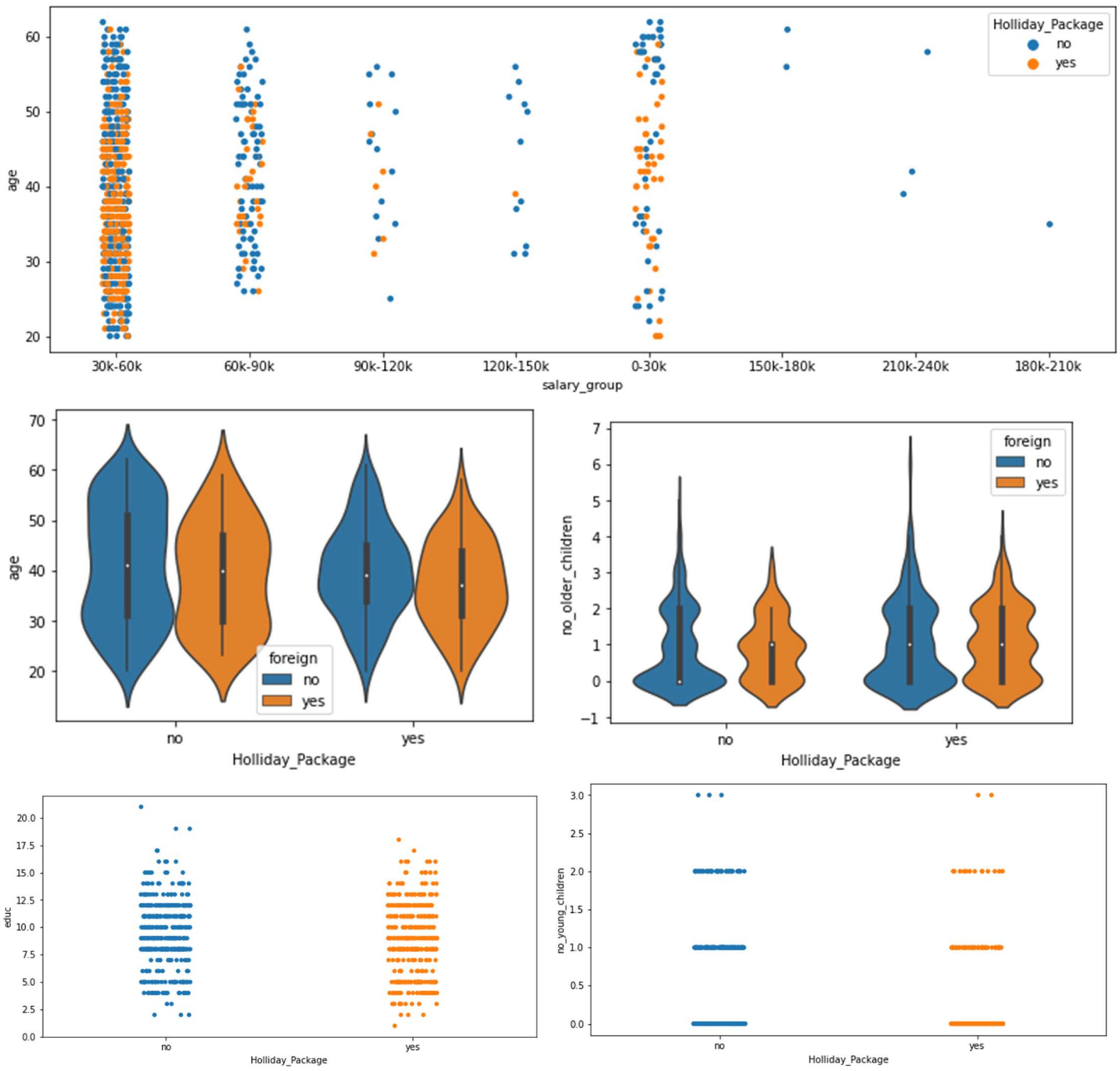
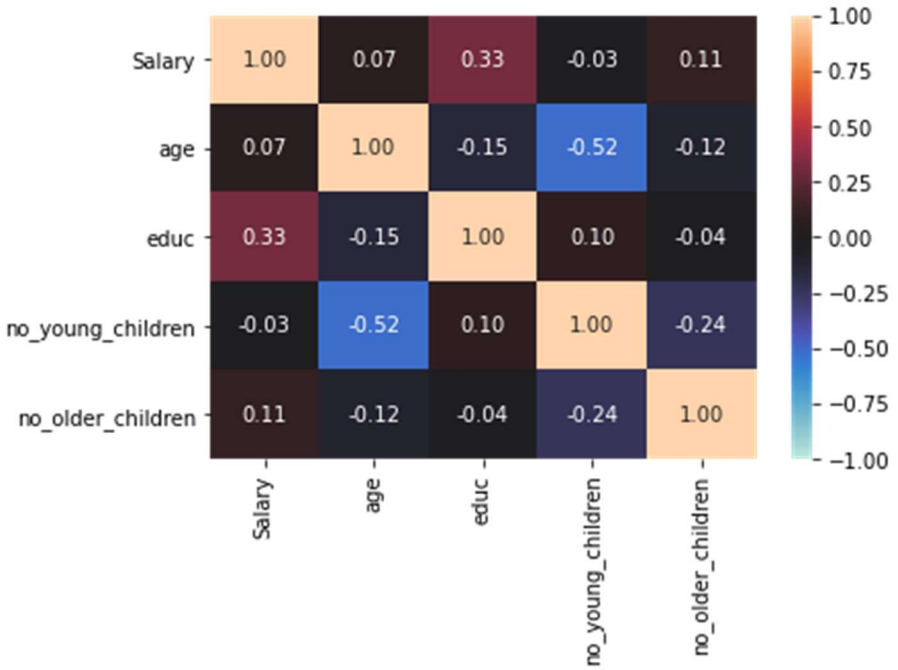


Figure - 19 Holiday - Bivariate

Inference

Employees in the age group of 30-45+ are more likely to opt for the holiday package as there is an increase in salary. Also, the education is an important factor as the salary increases with the education level and chances to opt for a holiday package also increases. Employees with higher no of younger children does not opt for the package.

Multivariate Analysis



Overall, there is no correlation between the variables. However, education and Salary shows a slight positive correlation whereas age and no young children shows a negative correlation.

Figure - 20 Holiday - Multivariate

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).

Encoding the categorical variables

	Salary	age	educ	no_young_children	no_older_children	salary_group	age_group	Holliday_Package_yes	foreign_yes
0	48412	30	8	1	1	5	0	0	0
1	37207	45	8	0	1	5	2	1	0
2	58022	46	9	0	0	5	2	0	0
3	66503	31	11	2	0	6	1	0	0
4	66734	44	12	0	2	6	2	0	0

Table - 18 Holiday – Encoding

The age group and salary group are replaced with the categorical codes. For holiday package and foreign, dummy variables are created where No = 0 and Yes = 1

Train-Test Split into 70:30 ratio

```
X_train data shape = (610, 8)
y_train data shape = (610,)
X_test data shape = (262, 8)
y_test data shape = (262,)
```

Applying Logistic Regression using sklearn library

Using GridSearchCV

```
GridSearchCV(cv=10, estimator=LogisticRegression(),
             param_grid={'max_iter': [10000, 1000], 'penalty': ['l2', 'l1'],
                         'solver': ['liblinear', 'newton-cg'],
                         'tol': [0.0001, 1e-05]},
             scoring='f1', verbose=2)
```

Best Parameter: `LogisticRegression(max_iter=1000, penalty='l1', solver='liblinear', tol=1e-05)`

	No	Yes	Holiday_Package
257	0.2563	0.7437	Yes
258	0.7165	0.2835	No
259	0.4682	0.5318	Yes
260	0.1735	0.8265	Yes
261	0.6176	0.3824	No

Table - 19 Probability – Test

Using the best parameter identified with the grid search cv, the table reflects the probability ratio of an employee opting for a holiday package. The probability is calculated both for yes and no, and the class with higher probability which is also known as odds is the prediction.

$$Probability = \frac{Odds}{1 + Odds}$$

Applying Linear Discriminant Analysis from sklearn library

```
LinearDiscriminantAnalysis(solver='eigen')
```

	Discriminant_Score	Holiday_Package
0	2.8813	No
1	3.0398	No
2	2.2580	Yes
3	2.0668	Yes
4	2.6053	No

Table - 20 Discriminant Score

$$DS = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \ldots + \beta kXk$$

Where:

DS = Discriminant Score
β's = Discriminant weight (coefficients)
X's = Explanatory (Predictor or independent) variables

LDA determines group means and computes, for each individual, the probability of belonging to the different groups.

Coefficients of linear discriminants:

Shows the linear combination of predictor variables that are used to form the LDA decision rule.

	Variable	Component
0	Salary	-0.0000
1	age	-0.0397
2	educ	0.0151
3	no_young_children	-1.2375
4	no_older_children	-0.0339
5	salary_group	0.0207
6	age_group	-0.0595
7	foreign_yes	1.3512

Table - 21 Components

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 in train data

```
Logistic Regression > Train
===========================

Accuracy = 0.6721311475409836

AUC = 0.7241506127702841
```

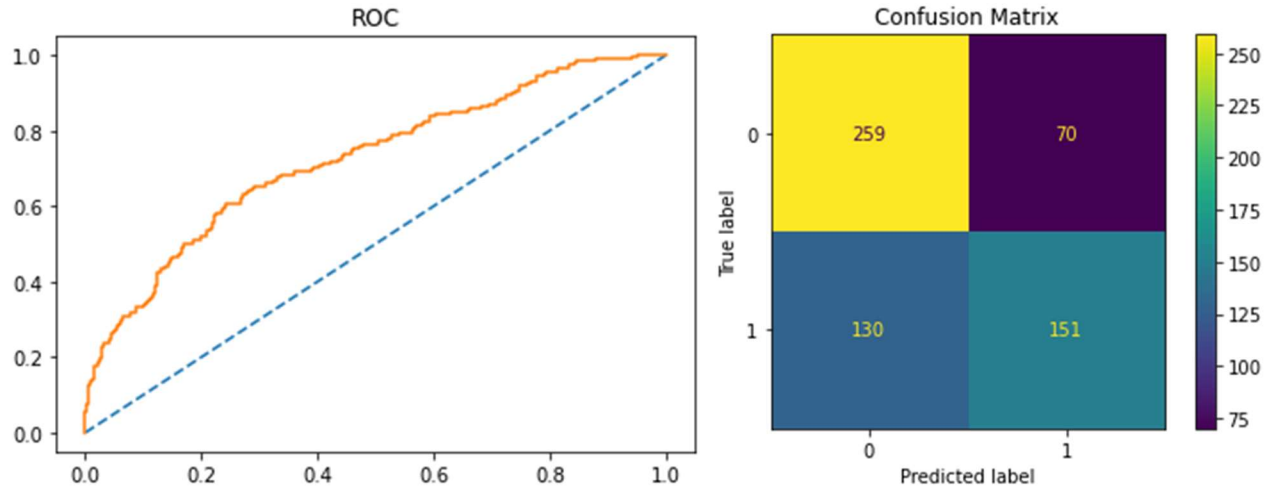


Figure - 21 ROC - Logistic Regression - train

Figure - 22 Confusion Matrix - LogisticRegression - train

classification_report

```
              precision    recall  f1-score   support

           0       0.67      0.79      0.72       329
           1       0.68      0.54      0.60       281

    accuracy                           0.67       610
   macro avg       0.67      0.66      0.66       610
weighted avg       0.67      0.67      0.67       610
```

F1 Score = 0.6015936254980079

Precision_score = 0.6832579185520362

Recall_score = 0.5373665480427047

roc_auc_score = 0.662300295297948

Performance metrics for Logistic Regression in test data

Logistic Regression > Test

==========

Accuracy = 0.6717557251908397

AUC = 0.7417840375586855

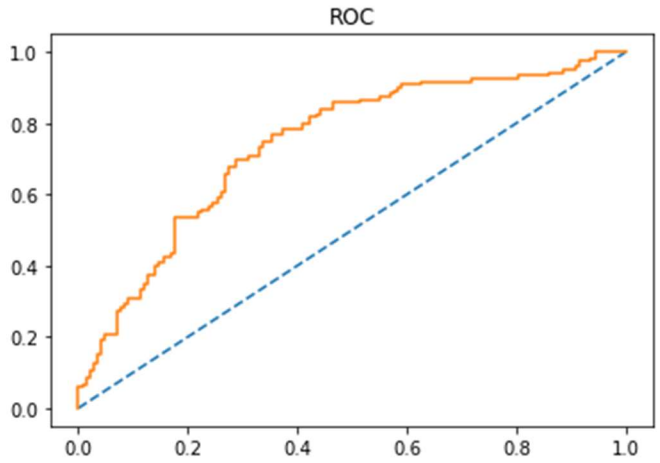


Figure - 24 ROC - LogisticRegression - test

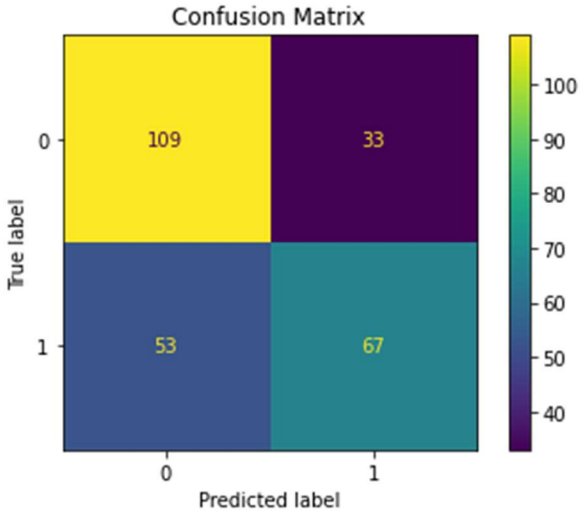


Figure - 23 Confusion matrix - Logistic - test

classification_report

```
              precision    recall  f1-score   support
```

0	0.67	0.77	0.72	142
1	0.67	0.56	0.61	120
accuracy			0.67	262
macro avg	0.67	0.66	0.66	262
weighted avg	0.67	0.67	0.67	262

F1 Score = 0.6090909090909091

Precision_score = 0.67

Recall_score = 0.5583333333333333

roc_auc_score = 0.6629694835680752

Performance metrics Linear Discriminant Analysis

Accuracy_train = 0.660655737704918
Accuracy_test = 0.6679389312977099
AUC for the Training Data: 0.727
AUC for the Test Data: 0.743

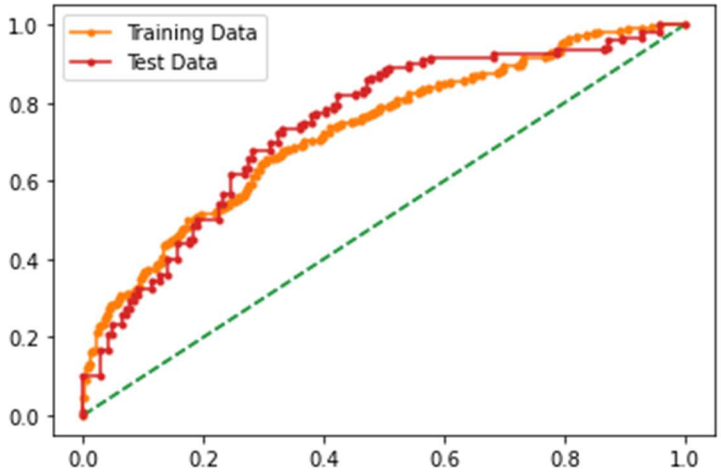


Figure - 25 ROC default LDA

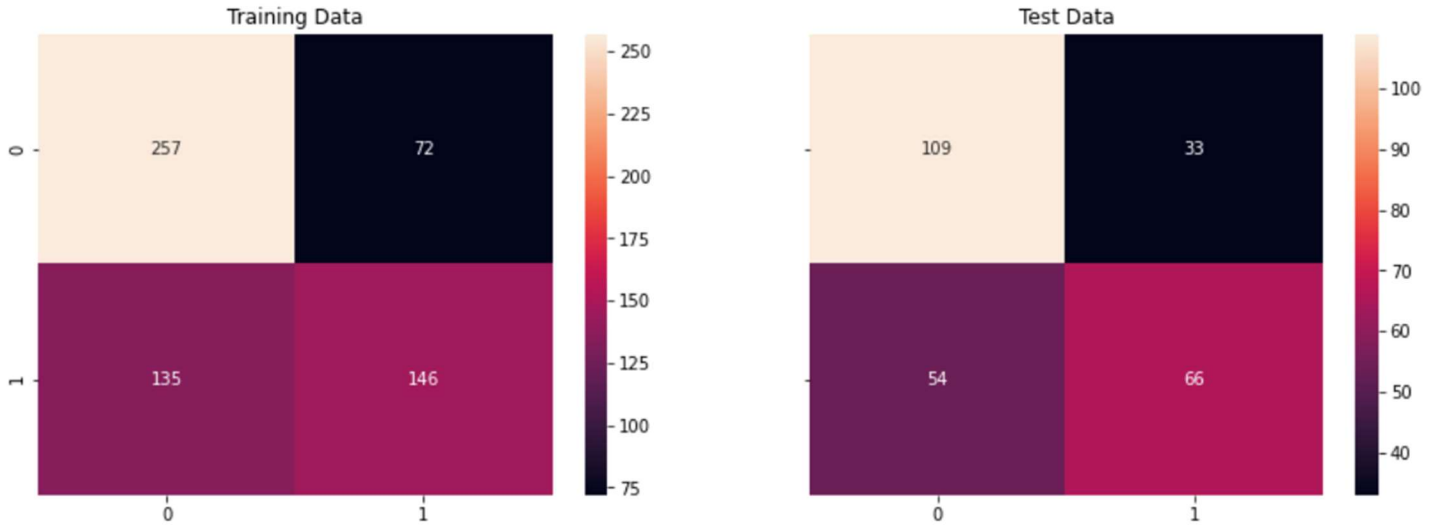


Figure - 26 Default cutoff - LDA

Identifying the best cut off point for predicting the class for LDA

	Probabilities	cutoff_5	cutoff_7
0	0.3804	0	0
1	0.3483	0	0
2	0.5144	1	0
3	0.5560	1	0
4	0.4387	0	0

Table - 22 Probability and cutoff

Performance metrics when the cut off is set to:

```
Cutoff: 0.1     Accuracy Score 0.4836 F1 Score 0.64 Precision 0.4714 recall 0.9964
Cutoff: 0.2     Accuracy Score 0.5148 F1 Score 0.6509 Precision 0.4868 recall 0.9822
Cutoff: 0.3     Accuracy Score 0.5656 F1 Score 0.6581 Precision 0.5162 recall 0.9075
Cutoff: 0.4     Accuracy Score 0.6557 F1 Score 0.6624 Precision 0.6041 recall 0.7331
Cutoff: 0.5     Accuracy Score 0.6607 F1 Score 0.5852 Precision 0.6697 recall 0.5196
Cutoff: 0.6     Accuracy Score 0.6508 F1 Score 0.5058 Precision 0.7267 recall 0.3879
Cutoff: 0.7     Accuracy Score 0.6426 F1 Score 0.4293 Precision 0.8119 recall 0.2918
Cutoff: 0.8     Accuracy Score 0.5918 F1 Score 0.2194 Precision 0.9211 recall 0.1246
Cutoff: 0.9     Accuracy Score 0.5443 F1 Score 0.0211 Precision 1.0 recall 0.0107
```

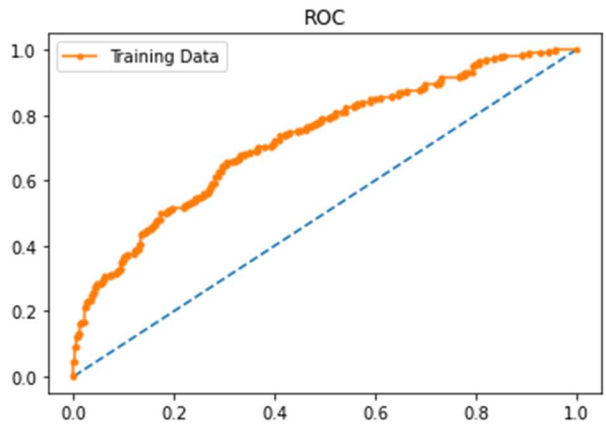
The accuracy score is higher when the cutoff is 0.5 and 0.4, however when we set cut off at 0.4, we also have high F1 score – 0.6624, high recall at 0.7331 and comparatively good precision score which is 0.6041. The accuracy is 0.6557 which is adequate.

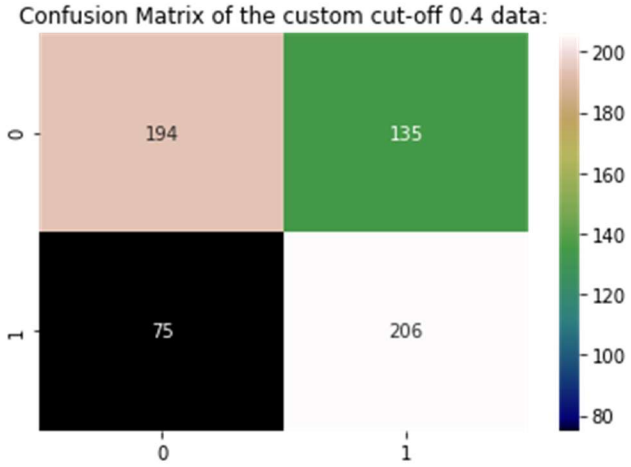
Performance metrics for Linear Discriminant Analysis model in train when cut off is set to 0.4

```
LinearDiscriminantAnalysis > Train
==================================

Accuracy = 0.6557377049180327

AUC = 0.7266168373914267
```





Classification Report of the custom cut-off 0.4:

	precision	recall	f1-score	support
0	0.72	0.59	0.65	329
1	0.60	0.73	0.66	281
accuracy			0.66	610
macro avg	0.66	0.66	0.66	610
weighted avg	0.67	0.66	0.66	610

f1_score 0.662379421221865

precision_score 0.6041055718475073

recall_score 0.7330960854092526

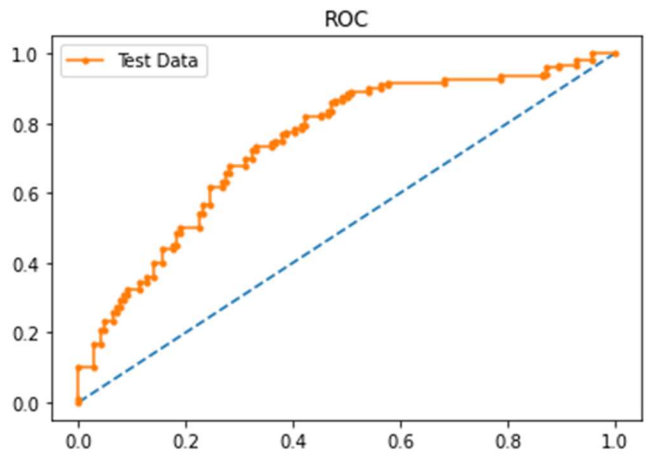
roc_auc_score 0.6613808694523468

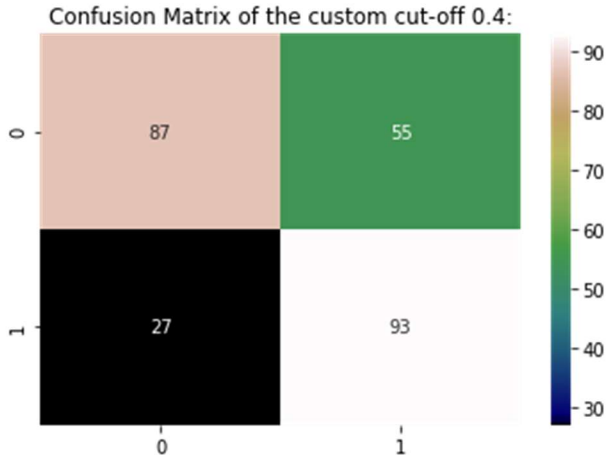
Performance metrics for Linear Discriminant Analysis in test data set when cut off is 0.4

LinearDiscriminantAnalysis > Test
==========

Accuracy = 0.6870229007633588

AUC = 0.743075117370892





Classification Report of the custom cut-off 0.4:

```
              precision    recall  f1-score   support

           0       0.76      0.61      0.68       142
           1       0.63      0.78      0.69       120

    accuracy                           0.69       262
   macro avg       0.70      0.69      0.69       262
weighted avg       0.70      0.69      0.69       262
```

f1_score 0.6940298507462687

precision_score 0.6283783783783784

recall_score 0.775

roc_auc_score 0.6938380281690142

Comparing both the models for best optimized results

Train data

Model	Accuracy	F1 Score	Precision	Recall	ROC_AUC Score
Logistic Regression	0.6721	0.6016	0.6833	0.5374	0.6623
Linear Discriminant Analysis	0.6557	0.6624	0.6041	0.7331	0.6614

Test data

Model	Accuracy	F1 Score	Precision	Recall	ROC_AUC Score
Logistic Regression	0.6718	0.6091	0.67	0.5583	0.6623
Linear Discriminant Analysis	0.687	0.694	0.6284	0.775	0.6938

Table - 23 Comparing models - train and test

Accuracy –How accurately / cleanly does the model classify the data points. Lesser the false predictions, more the accuracy

$$\text{Accuracy} = (TP + TN) / (TP+TN+FP+FN)$$

Recall - How many of the actual True data points are identified as True data points by the model. False Negatives are those data points which should have been identified as True.

$$\text{Recall} = TP/TP+FN$$

Precision –Among the points identified as Positive by the model, how many are really Positive

Precision = TP/TP+FP

F1-Score – The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall.

F1-Score = precision * recall / precision + recall

ROC_AUC Score – The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

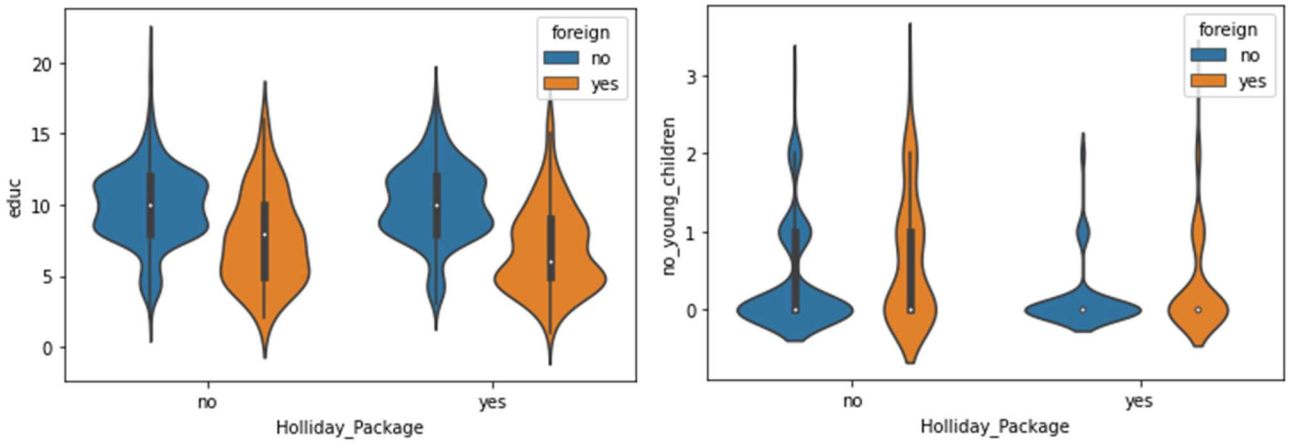
Based on all the evaluation metrics, Logistic Regression model does have a high accuracy (0.6718) however has lower f1 score and recall. Precision for train data is 0.6833 and for test is 0.67 and recall for train data is 0.5374 and test is 0.5583.

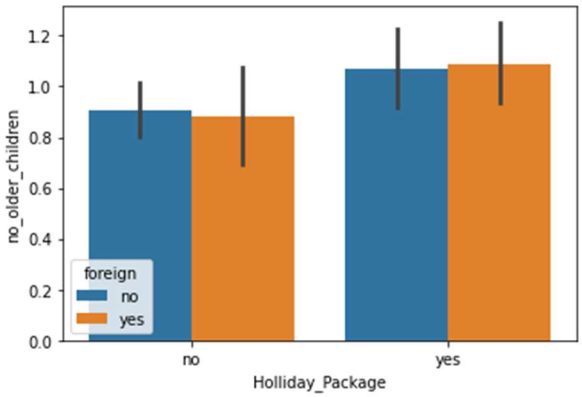
When we evaluate the Linear Discriminant Analysis model, even though the accuracy (0.687) is marginally less than Logistic Regression the f1-score is high in both train (0. 6624) and test data (0.694). The recall and precision are also better than the logistic regression model with recall at 0.7331 in train data and 0.775 in test data and precision as 0.6014 in train data and 0.6284 in test.

The ROC_AUC score for Linear Discriminant Analysis model is 0.6614 in train and 0. 6938 in test.

Concluding that Linear Discriminant Analysis model is a far superior model in terms of quality that Logistic Regression model for this particular case study.

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





Based on the analysis, employees who are between the ***age group of 30 to 50*** *are the ones that opt for the holiday package. Also, the employees with* ***no young children*** *or 0 or* ***at most 1 older child*** *prefer the holiday package. It can also be seen that the employees with number of* ***education years below 10*** *are more likely to plan holidays as their age will be most likely fall in 30 to 50 age group.*