Project: Predictive Modeling

Name- Manisha Rout
PGP-DSBA Online
Mar’22

Date: 22.05.2022

Table of Content

List of Figures

List of Tables

Problem 1.1

Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, duplicate values). Perform Univariate and Bivariate Analysis.

- The given dataset has 26967 rows and 10 columns. There are 7 attributes are of numeric data type (carat', 'depth', 'table', 'x', 'y', 'z', 'price') and 3 attributes are of object data type.
- The dataset has 697 missing values in the depth attribute. There are 34 duplicate instances present in the dataset. The index number '6215' is having 0 values for x (Length), y(width) and z(height) attributes which seems odd as it should have at least certain non-zero positive value.
- Outliers are present in all the numeric features which can be seen from the boxplot.
- Few anomalies present in the dataset, for instance The index number '6215' is having 0 values for x (Length), y(width) and z(height) attributes which seems odd as it should have at least certain non-zero positive value.
- The dataset requires few feature engineering before proceeding for model building.
- For feature engineering duplicate instances have been deleted. After deleting the instances, the dataset has 26933 rows and 10 columns.
- Outlier treatment has not been done on the original dataset since:
    - the outliers carry important information for the prediction, for instance in real life scenario a high carat cubic zirconia will have high price. Hence, we have not removed the outliers in the original data set.

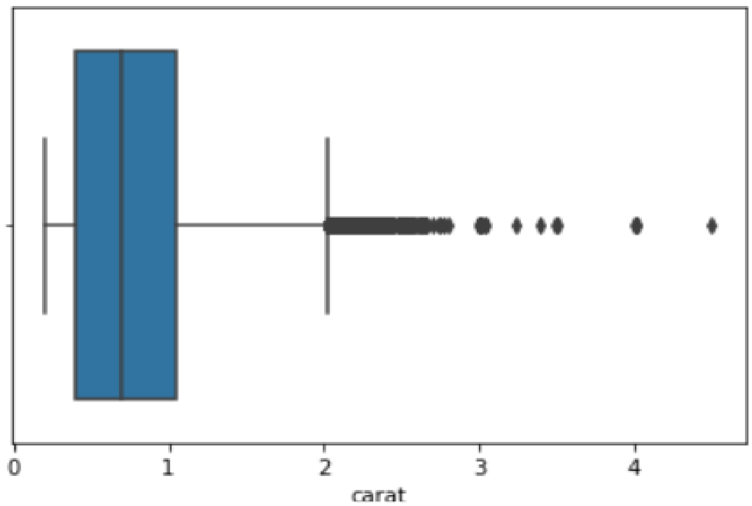
  However, to check the model performance, we have removed outliers in the cloned version of original dataset.

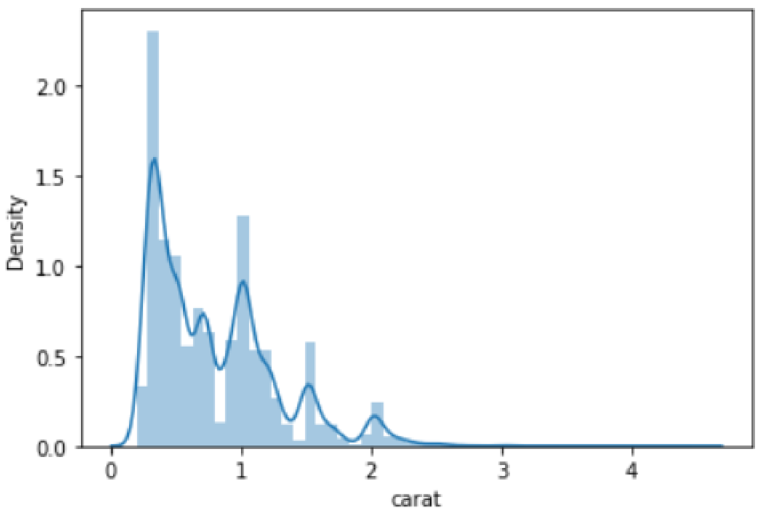
<u>Univariate analysis for Numerical Attributes</u>

1. Carat: Carat weight of the cubic zirconia

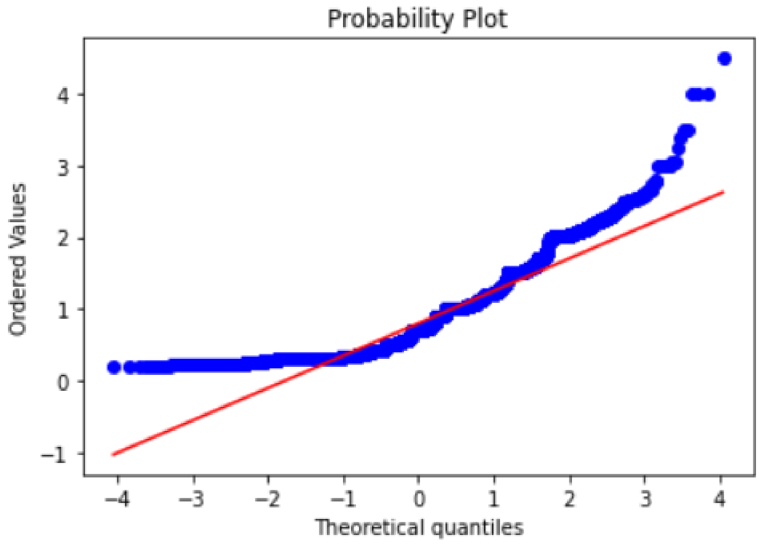
- Carat weight of the cubic zirconia ranges from 0.2 to 4.5
- Average carat weight of the cubic zirconia is 0.79
- The mean is greater than median, the distribution is not normal which is evident from the boxplot and probability plot.
- Skewness of the carat attribution is 1.1 indicating a right tailed distribution, positively skewed
- Outliers are present for this attribution which is evident from the box plot

```
Description ofcarat
..................
count    26933.000000
mean         0.798010
std          0.477237
min          0.200000
25%          0.400000
50%          0.700000
75%          1.050000
max          4.500000
Name: carat, dtype: float64
```





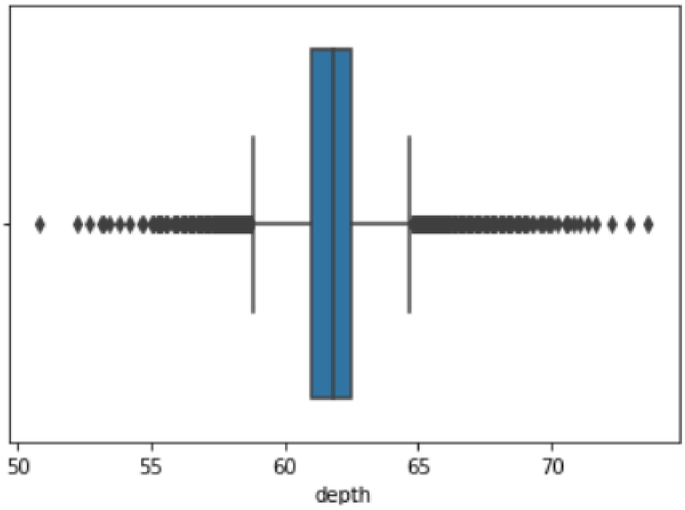
carat:

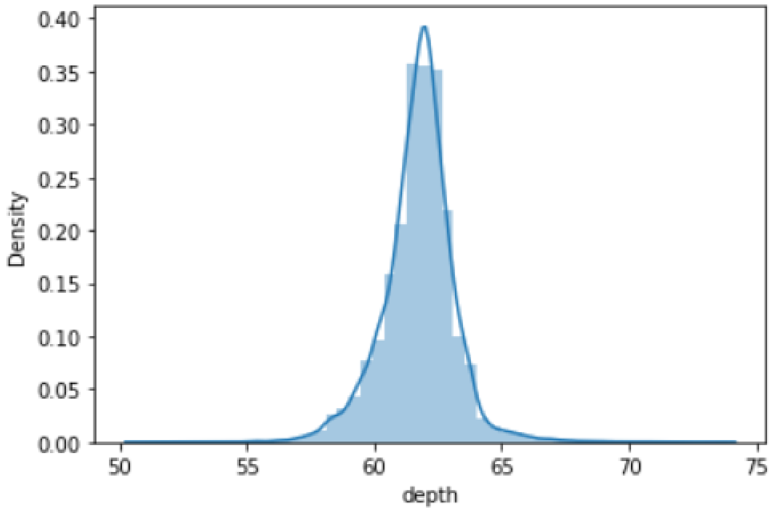


2. Depth: The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.

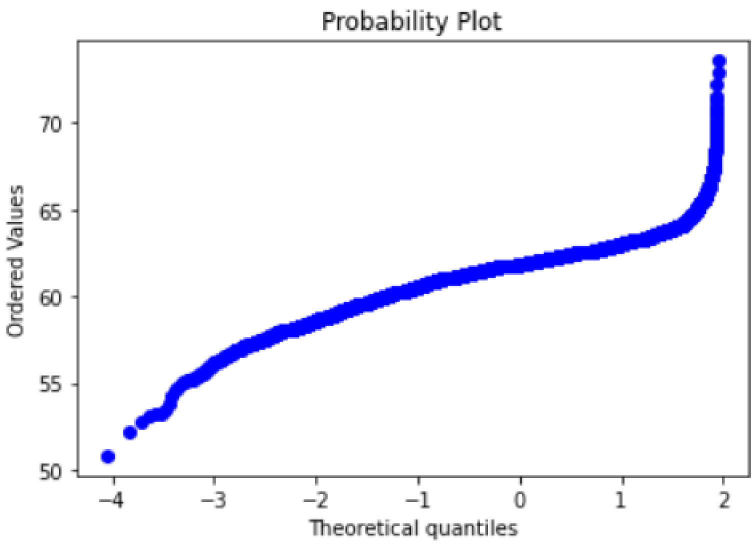
- Depth of cubic zirconia ranges from 51 to 74
- Average depth of cubic zirconia is 62
- The mean is almost equal to median, the distribution is almost normal which is evident from the boxplot and probability plot
- Skewness of the depth attribution is -0.02 indicating a moderately left tailed distribution, negatively skewed.
- Outliers are present for this attribution which is evident from the box plot

```
Description ofdepth
..................
count    26236.000000
mean        61.745285
std          1.412243
min         50.800000
25%         61.000000
50%         61.800000
75%         62.500000
max         73.600000
Name: depth, dtype: float64
```





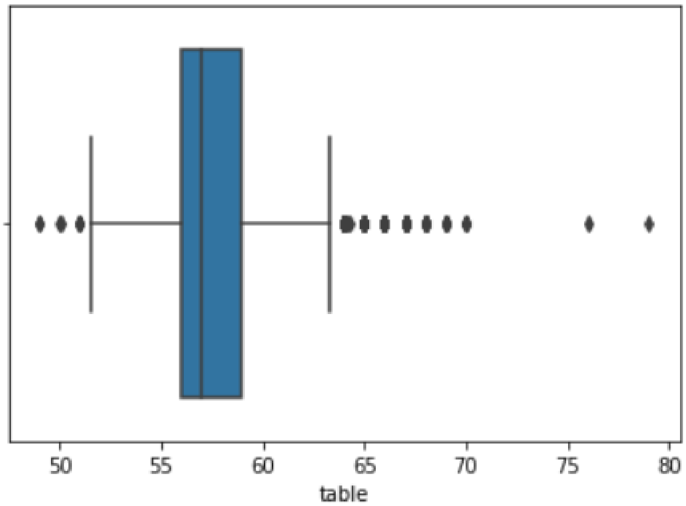
depth:



3. Table: The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.

- Table ranges from 49 to 79
- Average of Table in cubic zirconia is 57
- The mean is almost equal to median, the distribution is almost normal which is evident from the boxplot and probability plot
- Skewness of the depth attribution is 0.76 indicating a moderately right tailed distribution, positively skewed.
- Outliers are present for this attribution which is evident from the box plot

```
Description oftable
..................
count    26933.000000
mean        57.455950
std          2.232156
min         49.000000
25%         56.000000
50%         57.000000
75%         59.000000
max         79.000000
Name: table, dtype: float64
```



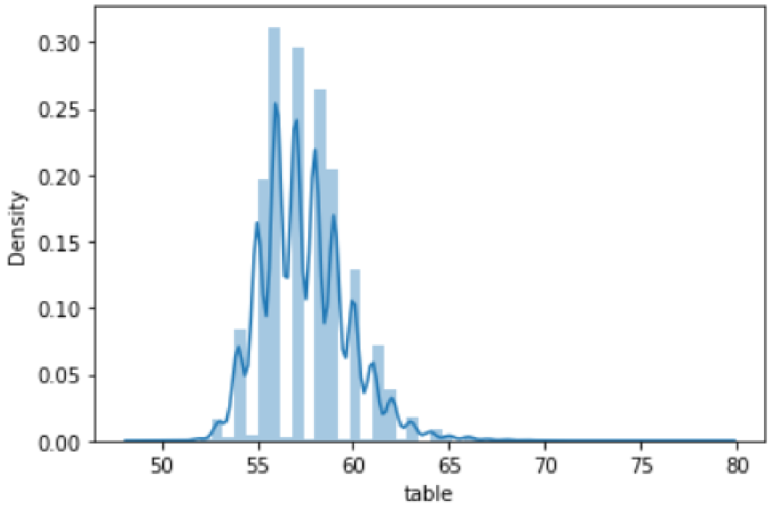
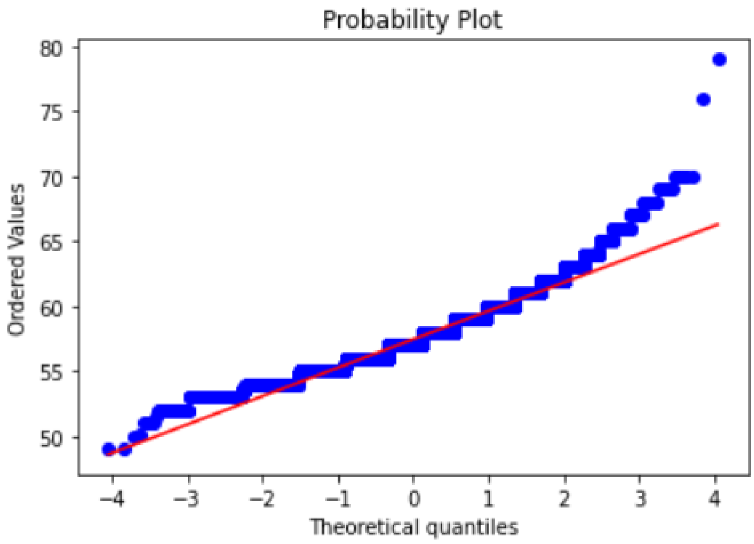


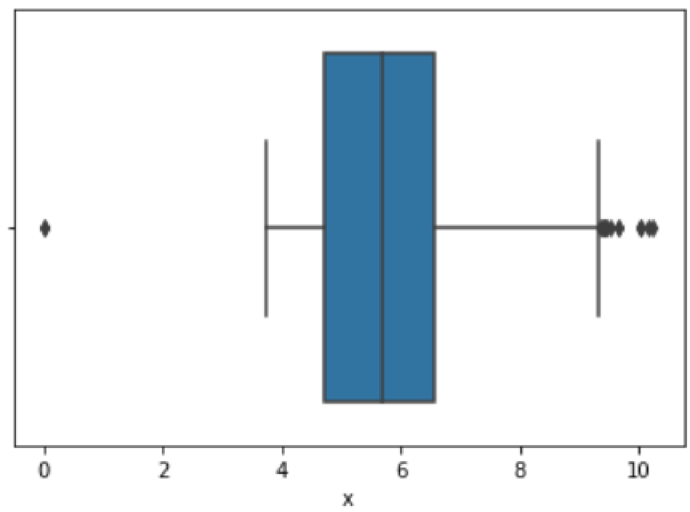
table:

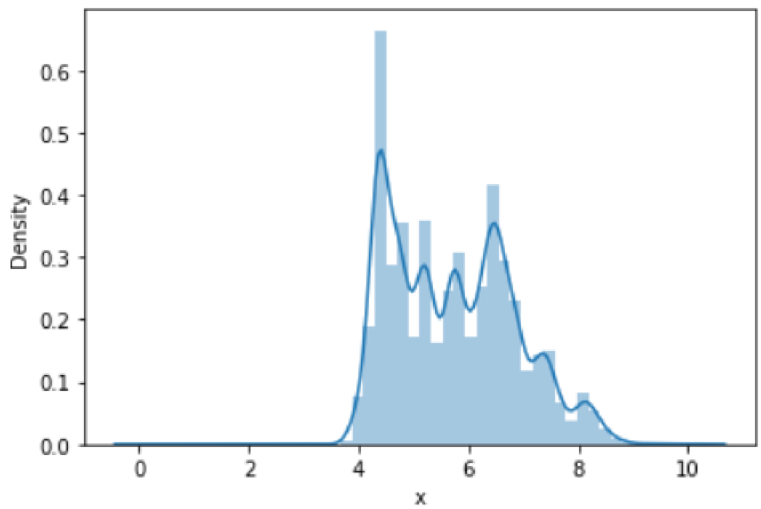


4. X: Length of cubic zirconia in mm.

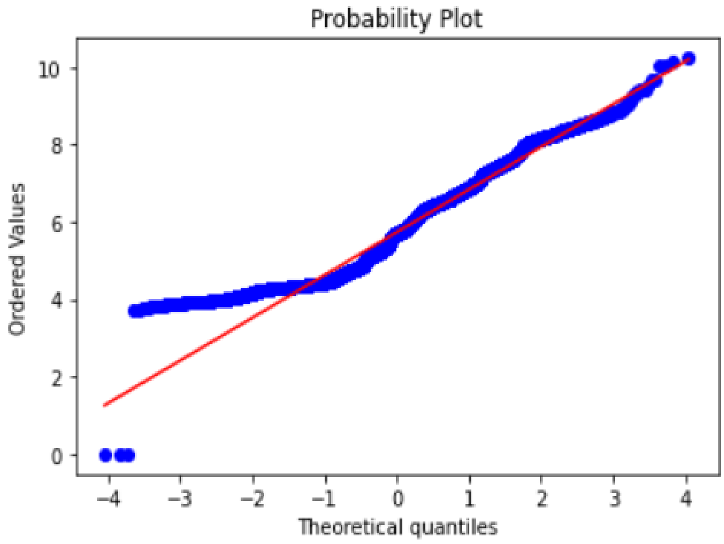
- Length of cubic zirconia ranges from 0.00mm to 10mm
- The minimum length of cubic Zirconia seems odd as it should have at least certain non-zero positive value.
- Average of length of cubic zirconia is 6
- The mean is almost equal to median, the distribution is almost normal which is evident from the boxplot and probability plot
- Skewness of the x attribution is 0.4 indicates a moderately right tailed distribution, positively skewed.
- Outliers are present for this attribution which is evident from the box plot

```
Description ofx
..................
count    26933.000000
mean         5.729346
std          1.127367
min          0.000000
25%          4.710000
50%          5.690000
75%          6.550000
max         10.230000
Name: x, dtype: float64
```





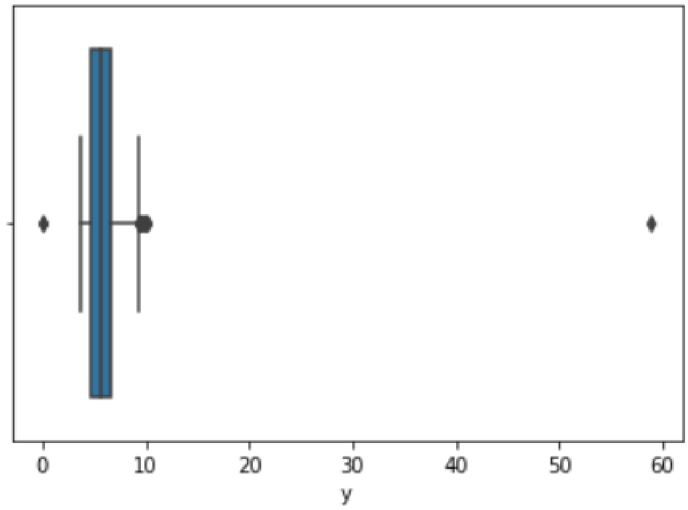
x:

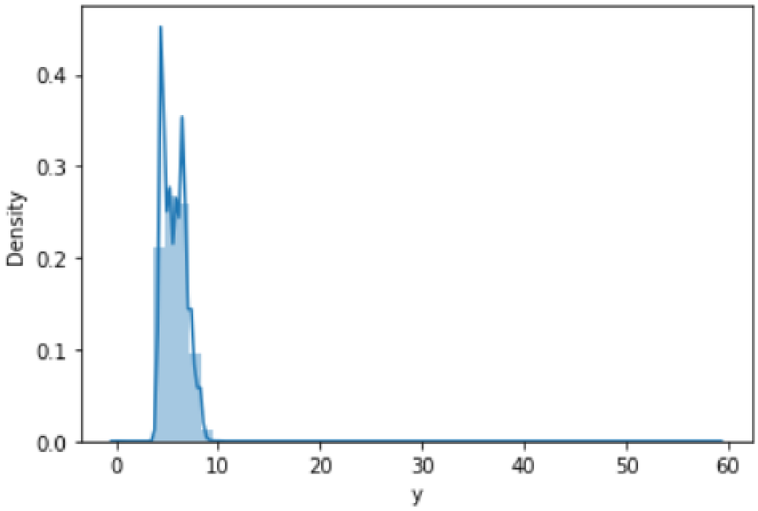


5. Y: Width of the cubic zirconia in mm

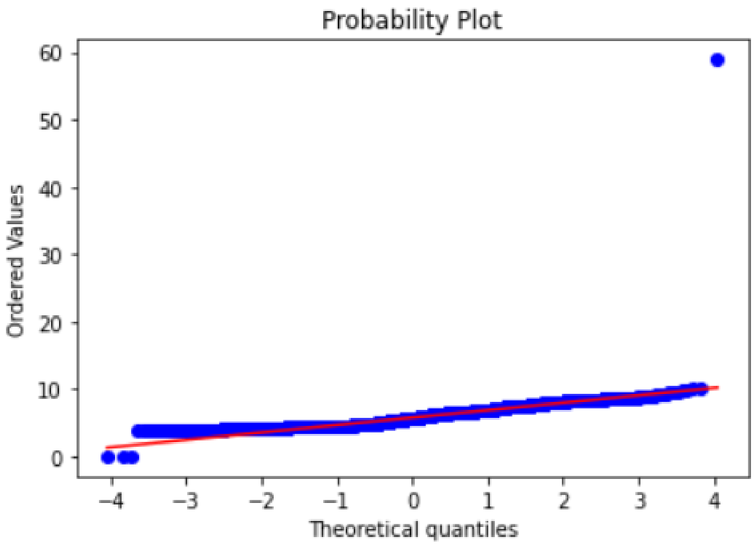
- Width of the cubic zirconia ranges from 0 to 59
- The minimum width of cubic Zirconia seems odd as it should have at least certain non-zero positive value.
- Average of width of cubic zirconia is 6
- The mean is almost equal to median, the distribution is almost normal which is evident from the boxplot and probability plot
- Skewness of the depth attribution is 3.8 indicating a right tailed distribution, positively skewed.
- Outliers are present for this attribution which is evident from the box plot

```
Description ofy
..................
count    26933.000000
mean         5.733102
std          1.165037
min          0.000000
25%          4.710000
50%          5.700000
75%          6.540000
max         58.900000
Name: y, dtype: float64
```





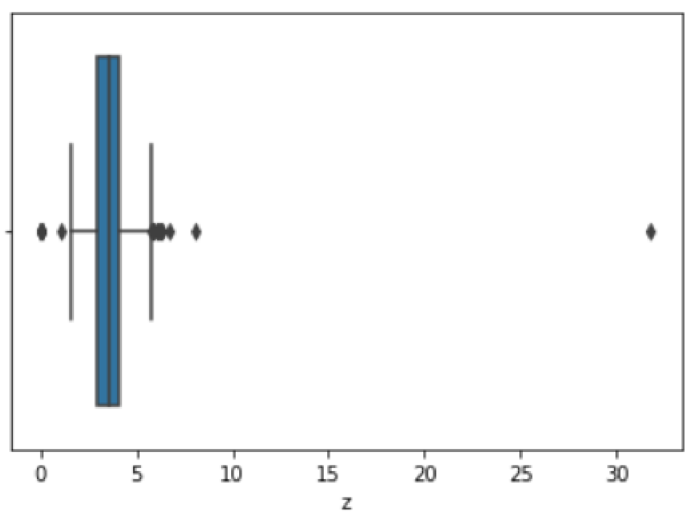
y:

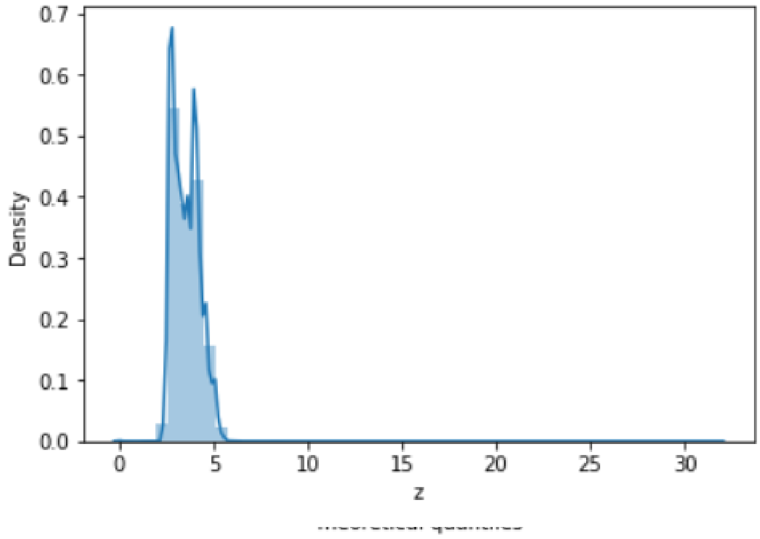


6. Z: Height of the cubic zirconia in mm

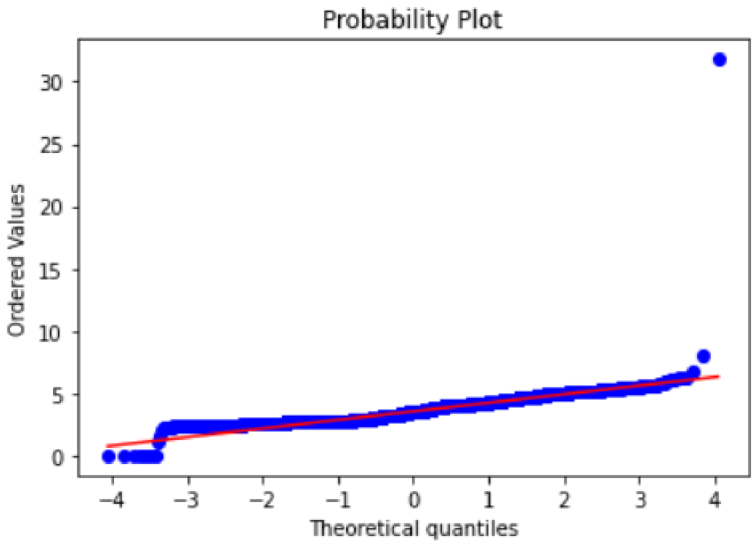
- Height of the cubic zirconia ranges from 0 to 32
- The minimum height of cubic Zirconia seems odd as it should have at least certain non-zero positive value.
- Average of width of cubic zirconia is 4
- The mean is almost equal to median, the distribution is almost normal which is evident from the boxplot and probability plot
- Skewness of the depth attribution is 2.8 indicating a right tailed distribution, positively skewed.
- Outliers are present for this attribution which is evident from the box plot

```
Description ofz
..................
count    26933.000000
mean         3.537769
std          0.719964
min          0.000000
25%          2.900000
50%          3.520000
75%          4.040000
max         31.800000
Name: z, dtype: float64
```





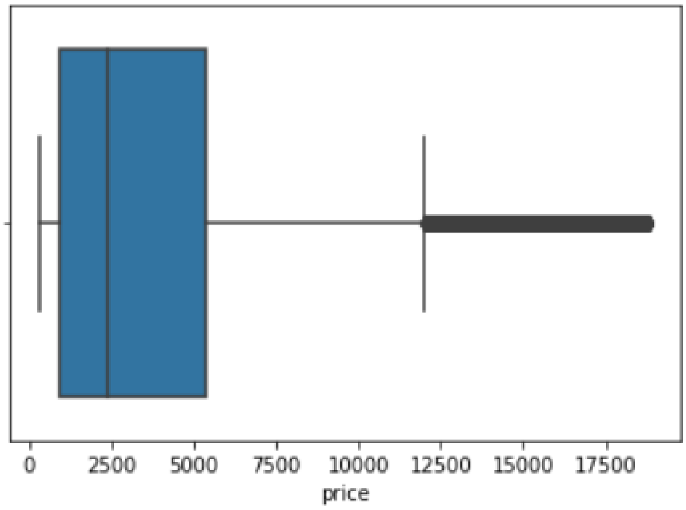
z :

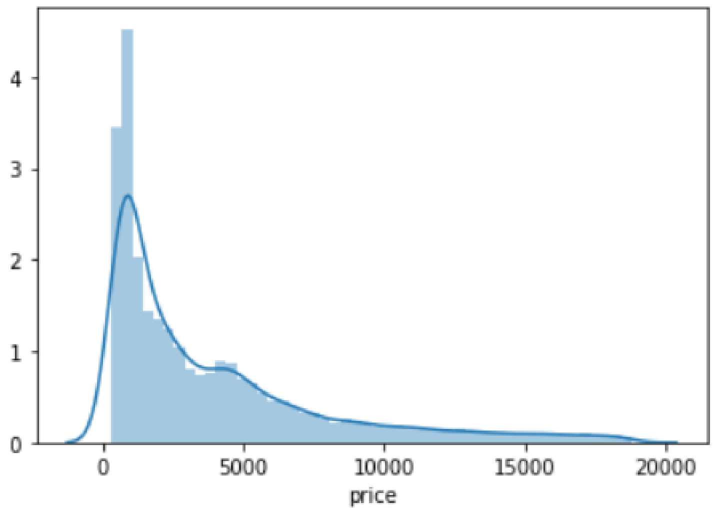


7. **Price: The Price of the cubic zirconia.**

- Price ranges from 326 to 18818
- Average of Table in cubic zirconia is 3937
- The mean is significantly higher than the median, the distribution is skewed which is evident from the boxplot and probability plot
- Skewness of the price attribution is 1.6 indicating a moderately right tailed distribution, positively skewed.
- Outliers are present for this attribution which is evident from the box plot

```
count    26933.000000
mean      3937.526120
std       4022.551862
min        326.000000
25%        945.000000
50%       2375.000000
75%       5356.000000
max      18818.000000
Name: price, dtype: float64
```



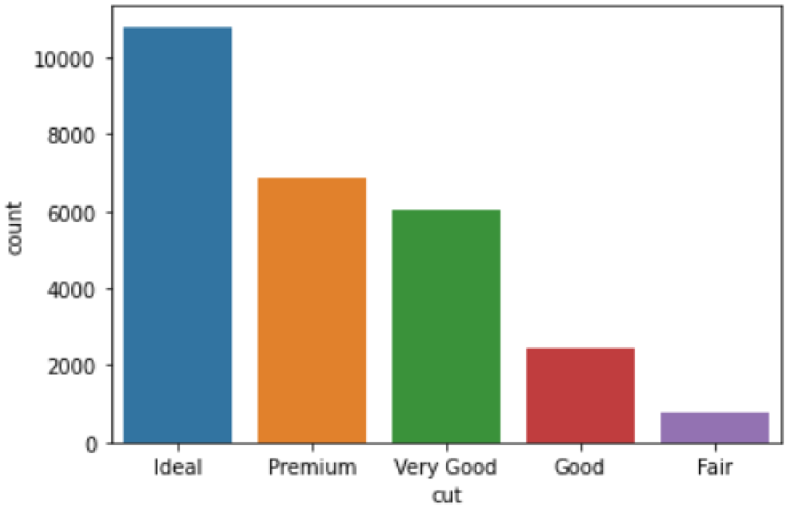


<u>Univariate analysis for Categorical Attributes</u>

1. **Cut: Describe the cut quality of the cubic zirconia**

- There are 5 types of cut available in cubic zirconia
- Ideal cut has maximum instances and fair cut has minimum

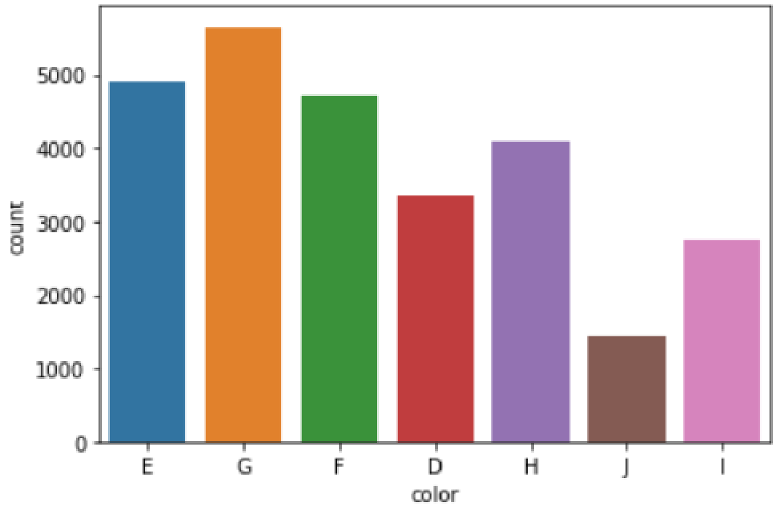
```
Description ofcut
..................
count      26933
unique         5
top        Ideal
freq       10805
Name: cut, dtype: object
```



2. Colour: Describe the colour of the cubic zirconia

- There are 7 types of color codes available in cubic zirconia
- Code G color has maximum instances and code j has minimum

```
Description ofcolor
..................
count      26933
unique         7
top            G
freq        5653
Name: color, dtype: object
Description ofclarity
```

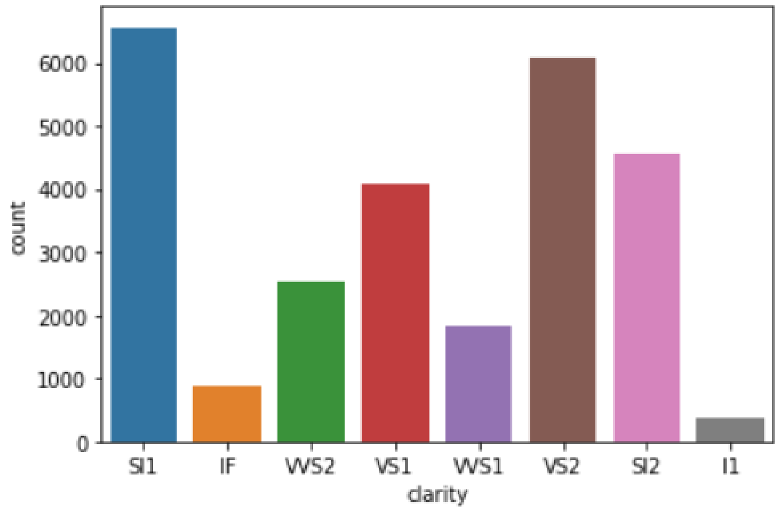


3. Clarity: Clarity refers to the absence of the Inclusions and Blemishes.

- There are 8 different types of clarity available in cubic zirconia
- Code SI1 clarity has maximum instances and code I1 has minimum

```
Description ofclarity
..................
count      26933
unique         8
top          SI1
freq        6565
```

. . . .



Multivariate-Bivariate analysis

Heat map shows the correlation between different numeric attributes by assigning numbers as well as colours and Pair plot gives a graphical representation of correlation between different numeric attributes.

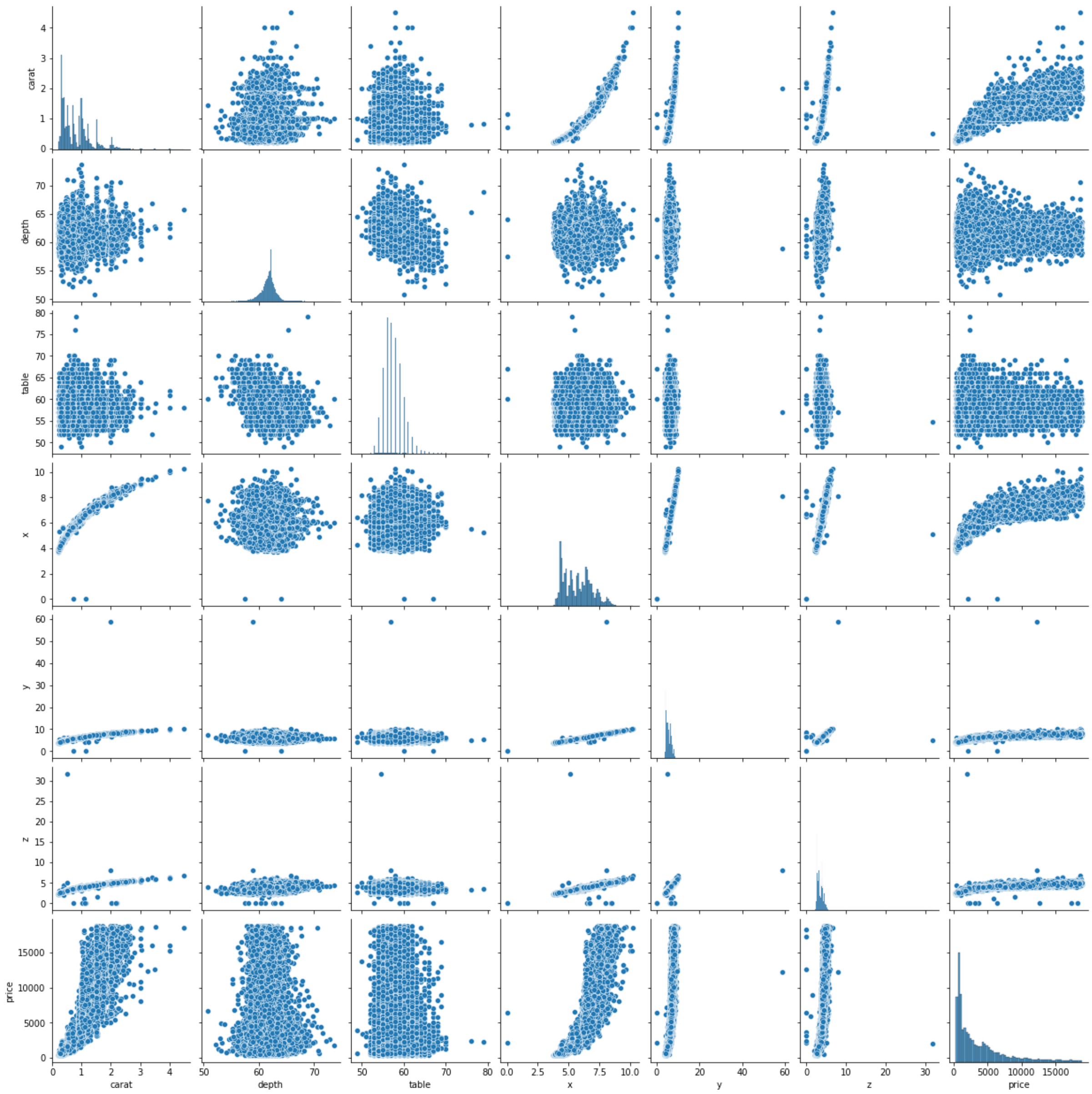


Figure 1: Pairplot of numeric attributes

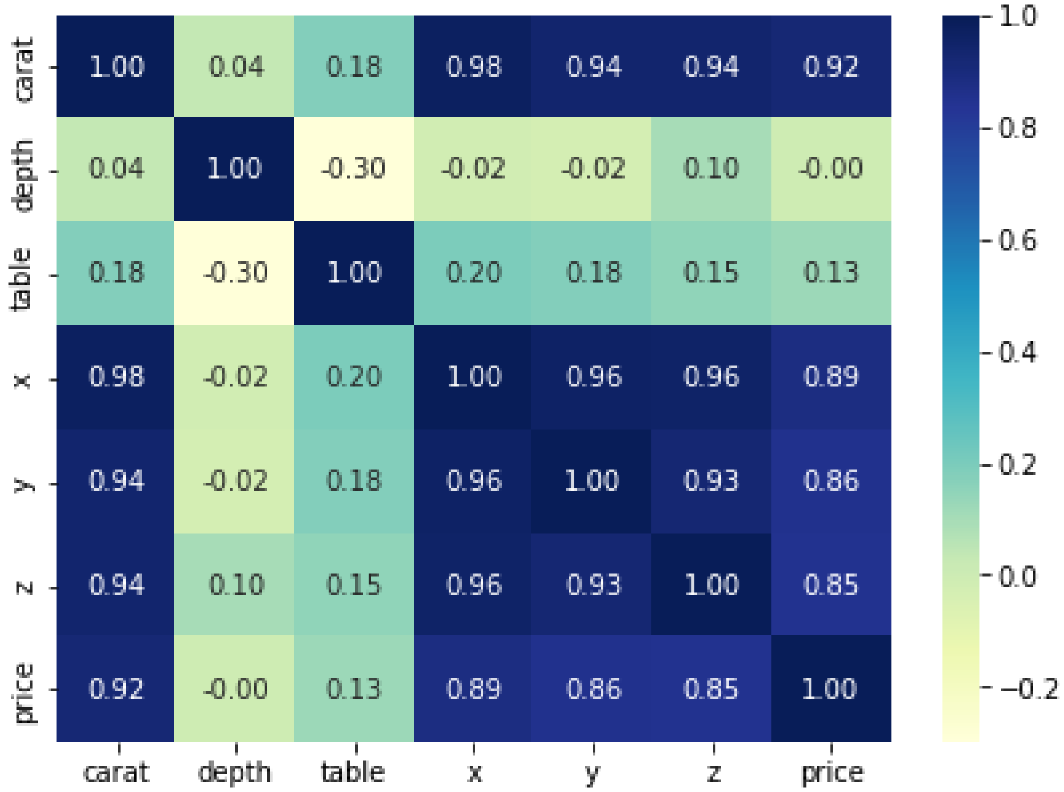
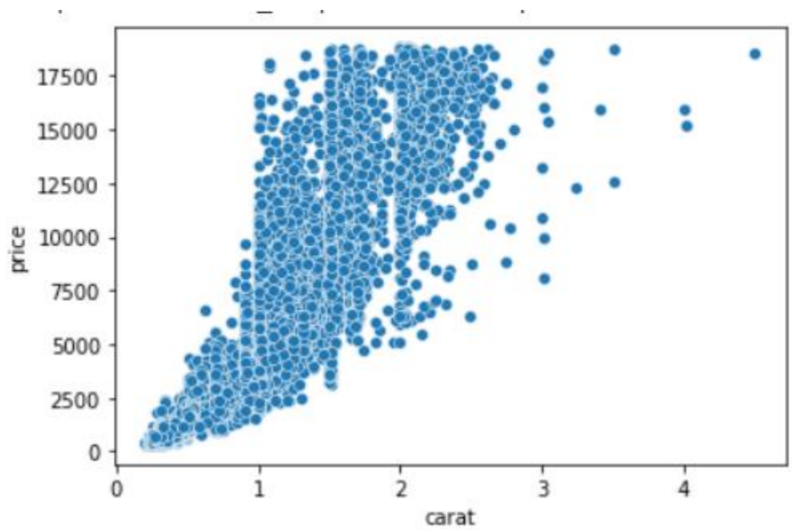


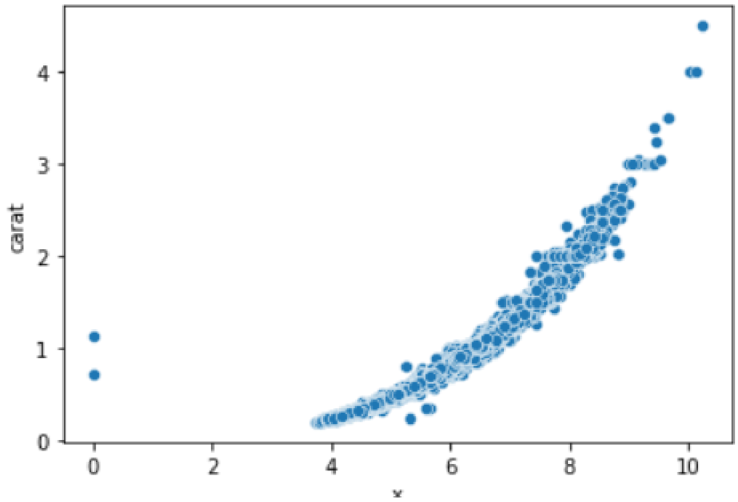
Figure 2: Heat map of numeric attributes

Bivariate analysis for numeric attributes

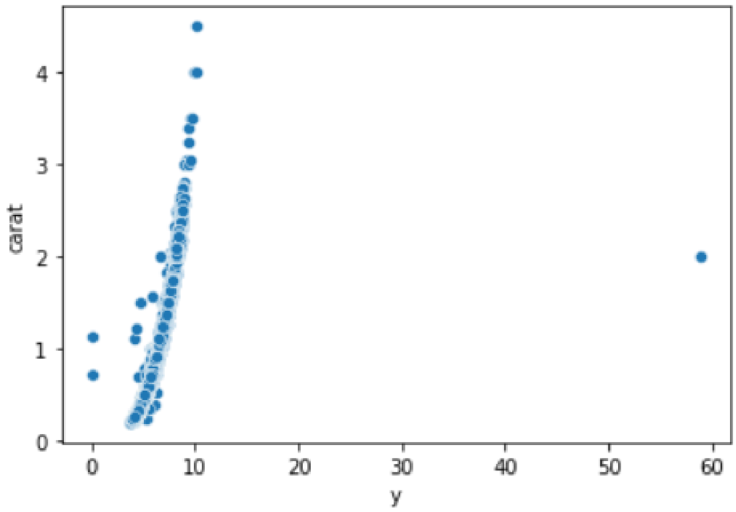
- The observations are as follows:
    - There is a very strong positive correlation (0.92) between the carat and price exists; which infers that as the carat weight of cubic zirconia increases, the price of cubic zirconia also increases. In short, carat is a potential attribute among all other attributes to predict the price of cubic zirconia.



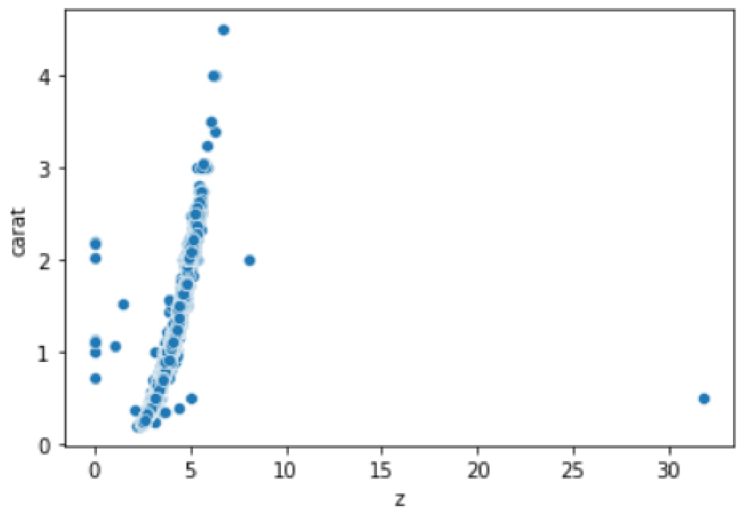
- There is a very strong positive correlation (0.98) between the carat and Length of the cubic zirconia in mm exists; which infers that as the carat weight of cubic zirconia increases, length of cubic zirconia in mm also increases. In short, a strong correlation may cause some impact on model performance



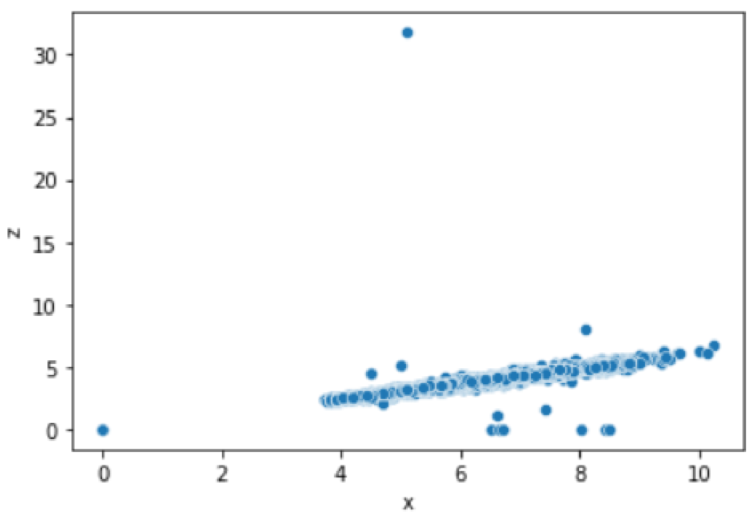
- There is a very strong positive correlation (0.94) between the carat and width of the cubic zirconia in mm exists; which infers that as the carat weight of cubic zirconia increases, width of cubic zirconia in mm also increases. In short, a strong correlation may cause some impact on model performance



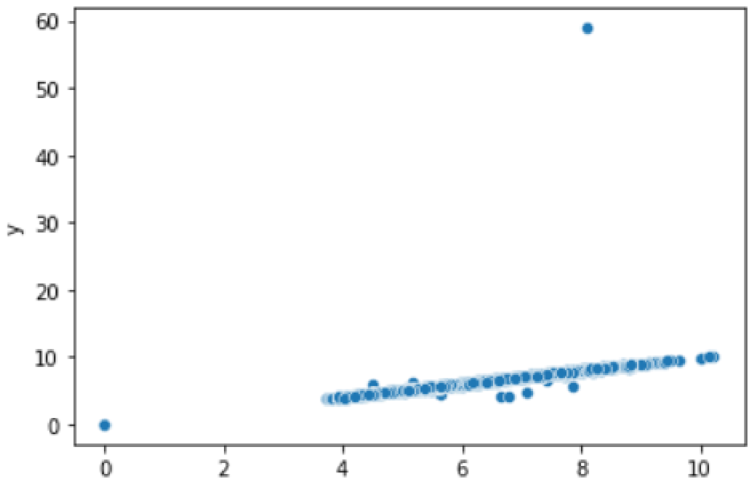
- There is a very strong positive correlation (0.94) between the carat and height of the cubic zirconia in mm exists; which infers that as the carat weight of cubic zirconia increases, height of cubic zirconia in mm also increases. In short, a strong correlation may cause some impact on model performance



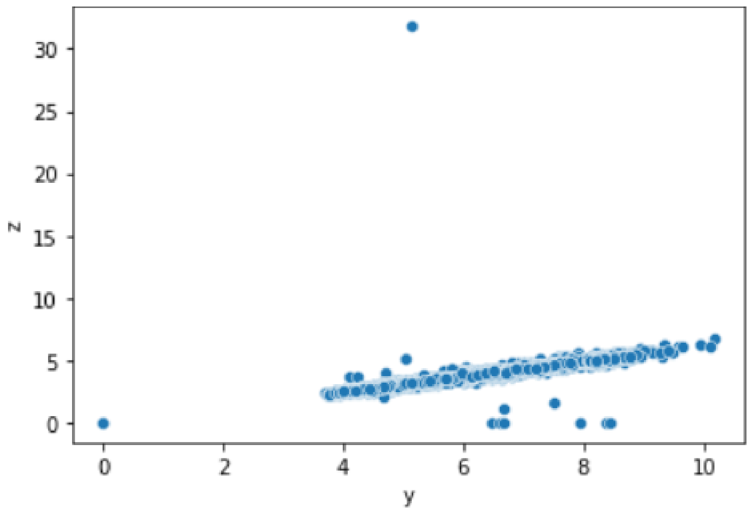
- There is a very strong positive correlation (0.96) between the length and height of the cubic zirconia in mm exists; which infers that as the length of cubic zirconia increases, height of cubic zirconia in mm also increases. In short, a strong correlation may cause some impact on model performance



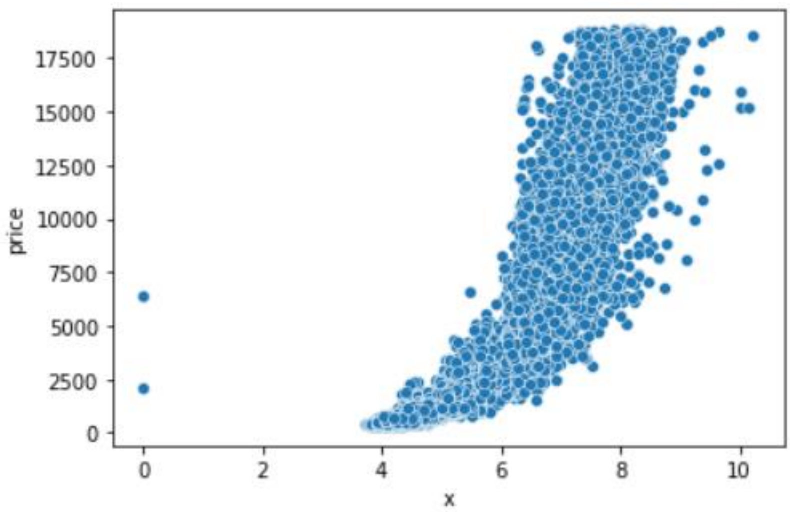
- There is a very strong positive correlation (0.96) between the length and width of the cubic zirconia in mm exists; which infers that as the length of cubic zirconia increases, width of cubic zirconia in mm also increases. In short, a strong correlation may cause some impact on model performance



- There is a very strong positive correlation (0.93) between the width and height of the cubic zirconia in mm exists; which infers that as the width of cubic zirconia increases, height of cubic zirconia in mm also increases. In short, a strong correlation may cause some impact on model performance



- There is a very strong positive correlation (0.89) between the length of cubic zirconia and price; which infers that as the length of cubic zirconia increases, the price of cubic zirconia also increases. In short, length of cubic zirconia is a potential attribute among all other attributes to predict the price of cubic zirconia



- There is a very strong positive correlation (0.86) between the width and price; which infers that as the width of cubic zirconia increases, the price of cubic zirconia also increases. As we have already seen the presence of outliers present in the data, we have checked two scatter plot one with outliers and another without outliers. Inference is, width is a potential attribute among all other attributes to predict the price of cubic zirconia.

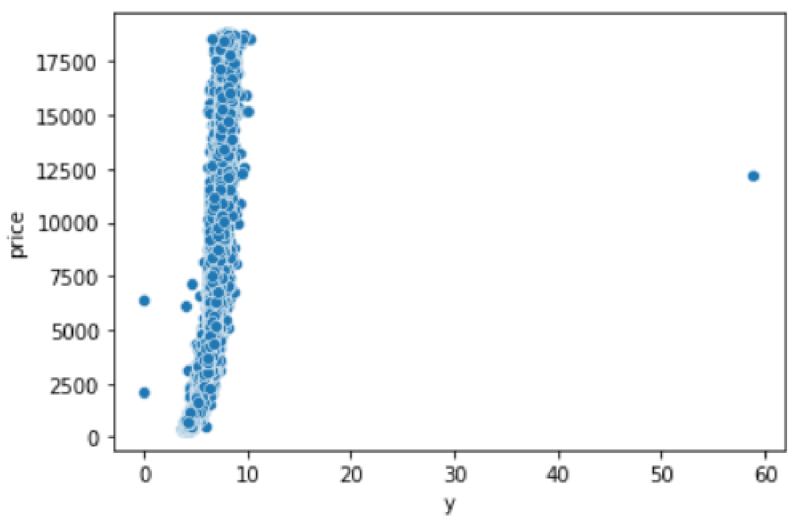


Figure 3: Scatterplot Without the outliers treatment

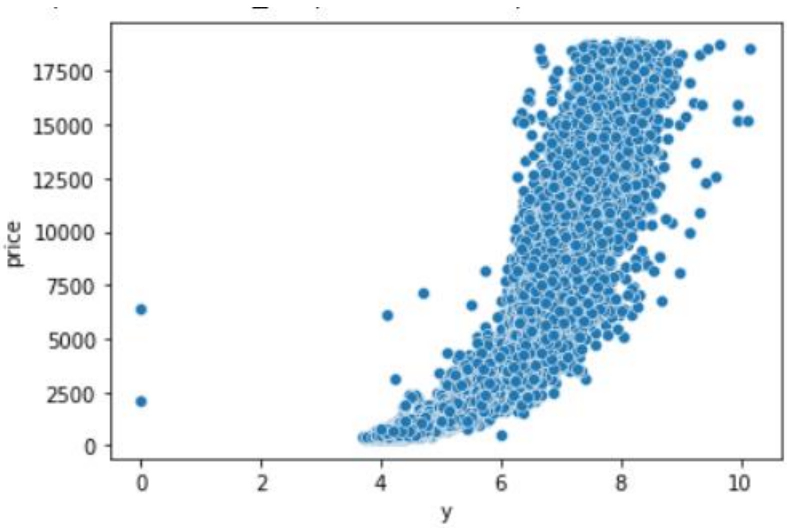


Figure 4: Scatterplot with outliers' treatment

- There is a very strong positive correlation (0.85) between the height and price; which infers that as the height of cubic zirconia increases, the price of cubic zirconia also increases. As we have already seen the presence of outliers in the data, we have checked two scatter plot one with outliers and another without outlier. Inference is, height of cubic zirconia is a potential attribute among all other attributes to predict the price of cubic zirconia

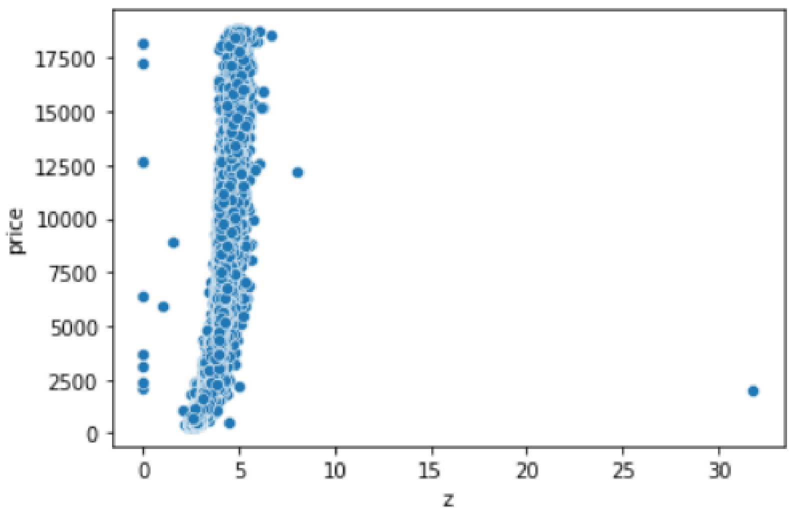


Figure 5: Scatterplot Without the outliers treatment

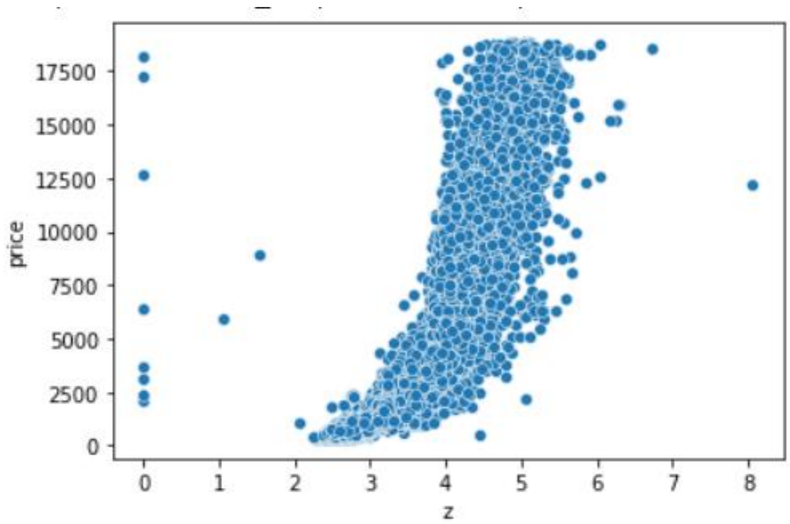


Figure 6: Scatter plot with outliers treatment

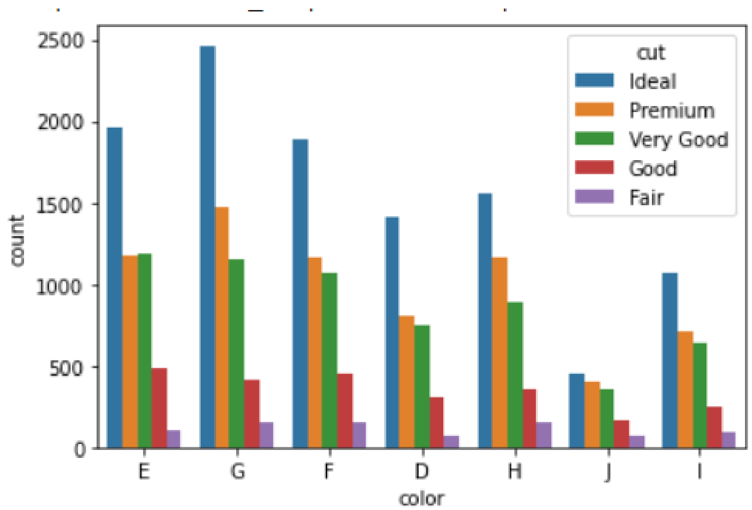
We have done outlier treatment for ‘z’ and ‘y’ attributes:

- Because there is strong correlation between the above attribute with price individually, which we have verified from correlation map
- However, the scatterplot between ‘z’- ‘price’ and ‘y’-‘price’ without the outliers treatment are not clearly observed
- Hence, the outlier treatment

- **Bivariate analysis of all the categorical attributes**

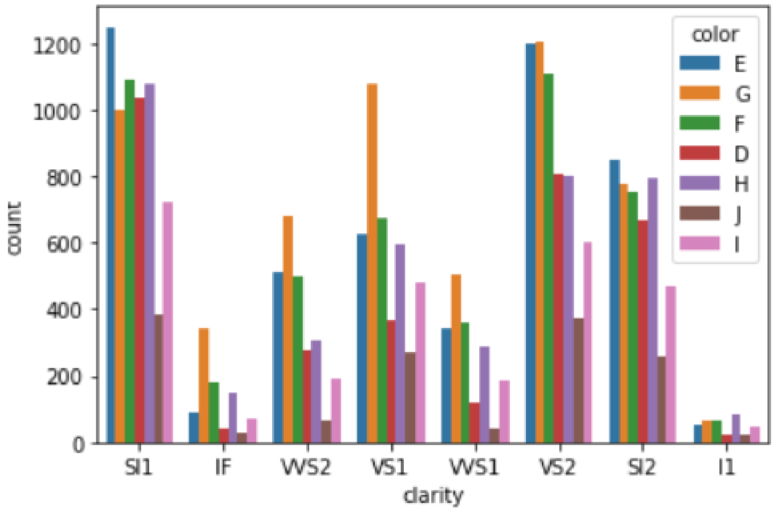
Color, Clarity, Cut

- Ideal cut with color code G has the maximum instances in cubic zirconia whereas fair cut with color J has minimum instances in cubic zirconia



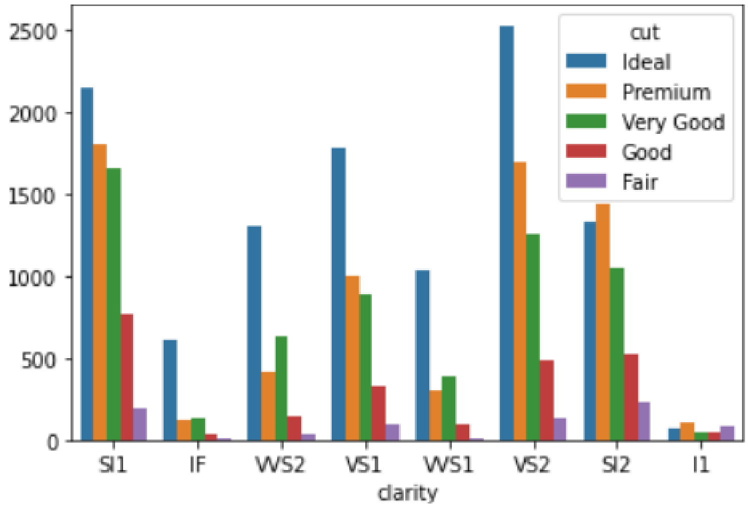
cut	Fair	Good	Ideal	Premium	Very Good	All
color						
D	74	311	1409	806	741	3341
E	100	490	1966	1174	1186	4916
F	148	453	1891	1164	1067	4723
G	147	418	2463	1471	1154	5653
H	149	351	1550	1159	886	4095
I	94	252	1073	707	639	2765
J	68	160	453	405	354	1440
All	780	2435	10805	6886	6027	26933

- Clarity SI1 with color code E has the maximum instances whereas, clarity I1 with color code J has minimum instances in cubic zirconia



color	D	E	F	G	H	I	J	All
clarity								
I1	25	53	67	68	82	48	21	364
IF	38	87	182	340	149	69	26	891
SI1	1039	1249	1088	998	1081	724	386	6565
SI2	669	849	750	778	793	467	258	4564
VS1	369	625	672	1076	593	480	272	4087
VS2	804	1202	1106	1205	803	600	373	6093
VVS1	121	342	360	507	288	183	38	1839
VVS2	276	509	498	681	306	194	66	2530
All	3341	4916	4723	5653	4095	2765	1440	26933

- Ideal cut with color code VS2 has maximum instances whereas, fair cut with color code IF has minimum instances in cubic zirconia



cut clarity	Fair	Good	Ideal	Premium	Very Good	All
I1	89	50	74	108	43	364
IF	4	30	610	115	132	891
SI1	193	764	2146	1809	1653	6565
SI2	224	527	1324	1443	1046	4564
VS1	93	330	1781	996	887	4087
VS2	129	491	2527	1693	1253	6093
VVS1	10	100	1036	307	386	1839
VVS2	38	143	1307	415	627	2530
All	780	2435	10805	6886	6027	26933

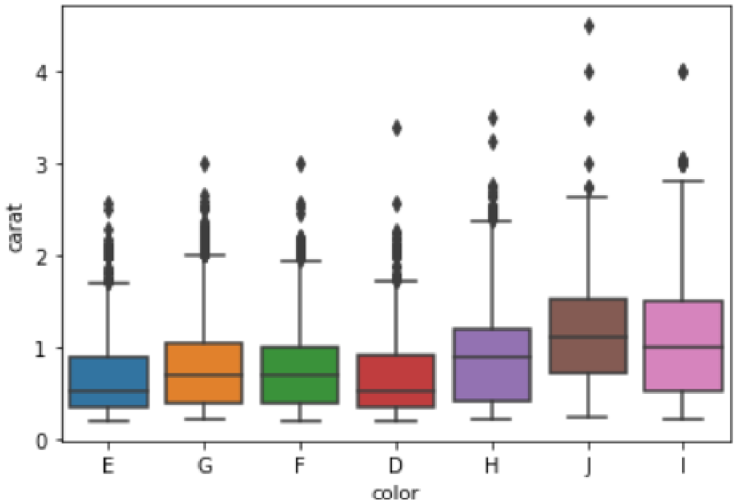
- **Bivariate analysis of all the num-cat attributes**

  Carat with Other categorical attributes

  - Color code J has maximum and color code E has minimum mean carat weight in cubic zirconia

```
Mean of carat for color
color
D    0.658515
E    0.656019
F    0.731139
G    0.770520
H    0.910464
I    1.033515
J    1.161653
Name: carat, dtype: float64
```

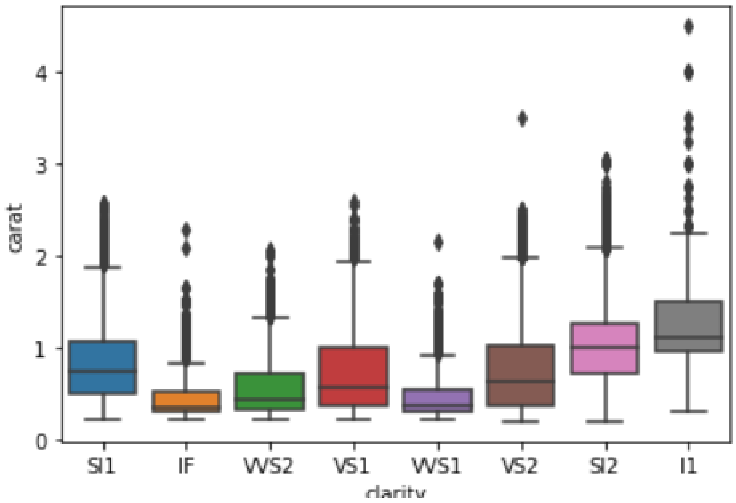
```
Plot of carat vs color
........................
```



- Clarity code I1 has maximum and clarity code IF has minimum mean carat weight in cubic zirconia

```
Mean of carat for clarity
clarity
I1      1.278132
IF      0.495443
SI1     0.849601
SI2     1.082358
VS1     0.726643
VS2     0.767939
VVS1    0.499929
VVS2    0.593047
Name: carat, dtype: float64
```

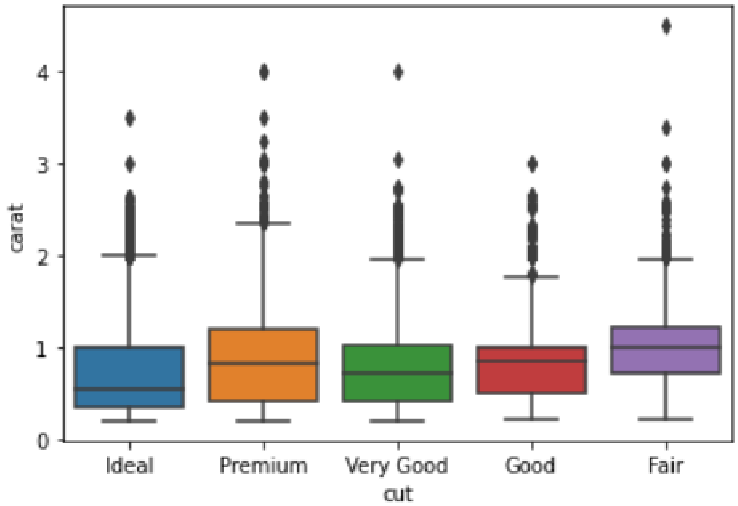
```
Plot of carat vs clarity
........................
```



- Fair cut has maximum and Ideal cut has minimum mean carat weight in cubic zirconia

```
Mean of carat for cut
cut
Fair         1.062000
Good         0.848953
Ideal        0.701430
Premium      0.888360
Very Good    0.813182
Name: carat, dtype: float64
```

```
Plot of carat vs cut
........................
```

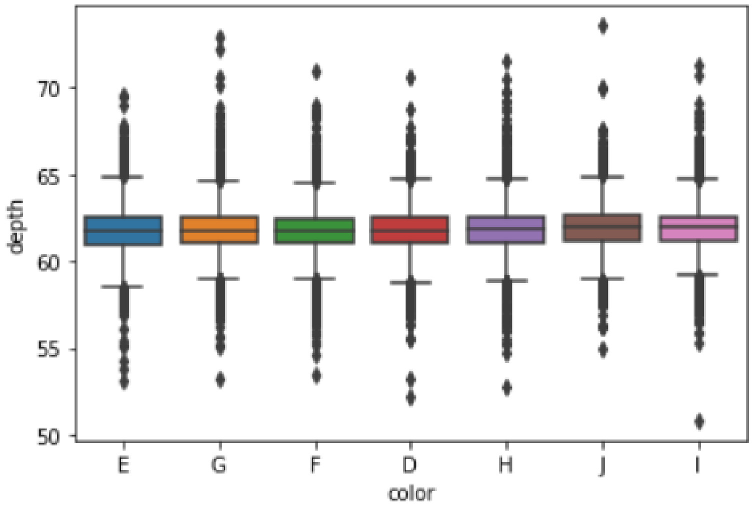


Depth with Other categorical attributes

- Color code J has maximum and color code E has minimum mean depth in cubic zirconia

```
Mean of depth for color
color
D    61.703553
E    61.657575
F    61.676400
G    61.744611
H    61.827937
I    61.869101
J    61.901001
Name: depth, dtype: float64
```

```
Plot of depth vs color
.........................
```

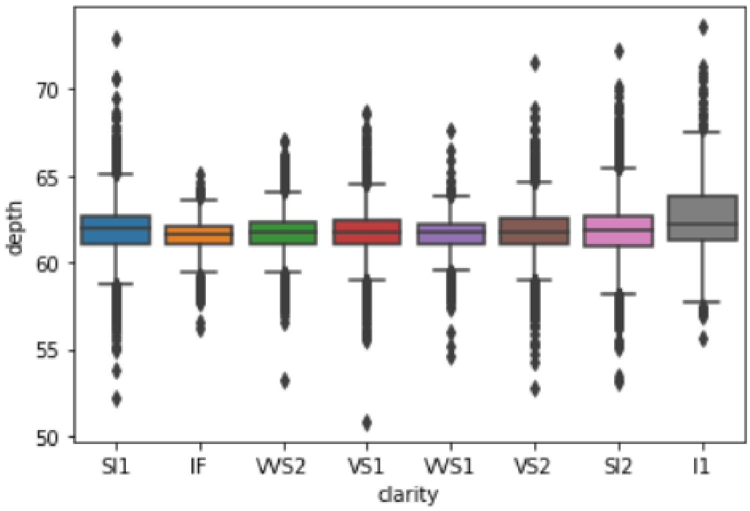


- Clarity code I1 has maximum and clarity code IF has minimum mean depth in cubic zirconia

color

```
Mean of depth for clarity
clarity
I1      62.630791
IF      61.499656
SI1     61.854342
SI2     61.775293
VS1     61.661029
VS2     61.719902
VVS1    61.624344
VVS2    61.653188
Name: depth, dtype: float64
```

```
Plot of depth vs clarity
.........................
```



- Fair cut has maximum and premium cut has minimum mean depth in cubic zirconia

```
Mean of depth for cut
cut
Fair          63.944312
Good          62.373948
Ideal         61.705363
Premium       61.267598
Very Good     61.823932
Name: depth, dtype: float64
```

```
Plot of depth vs cut
.........................
```

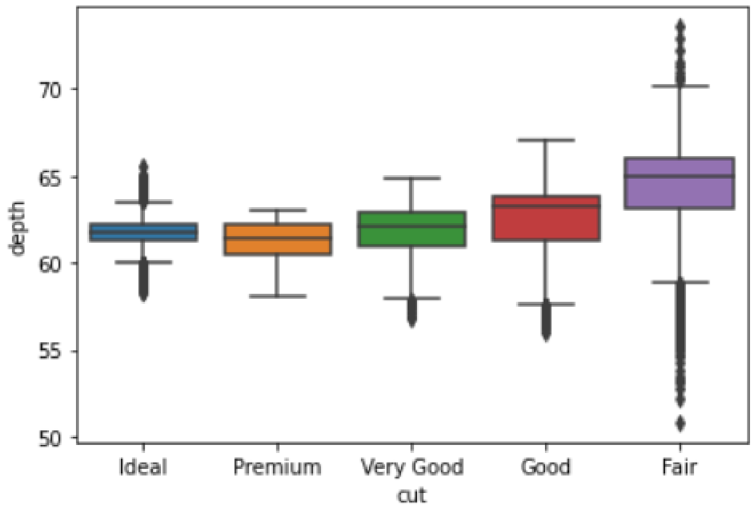
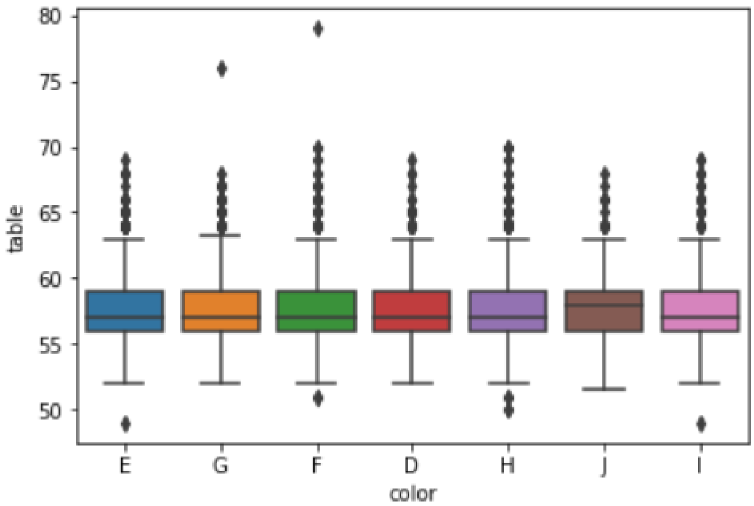


Table with Other categorical attributes

- Color code J has maximum and color code E has minimum mean table in cubic zirconia

```
Mean of table for color
color
D     57.374828
E     57.516843
F     57.439318
G     57.304617
H     57.484420
I     57.565533
J     57.793542
Name: table, dtype: float64
```

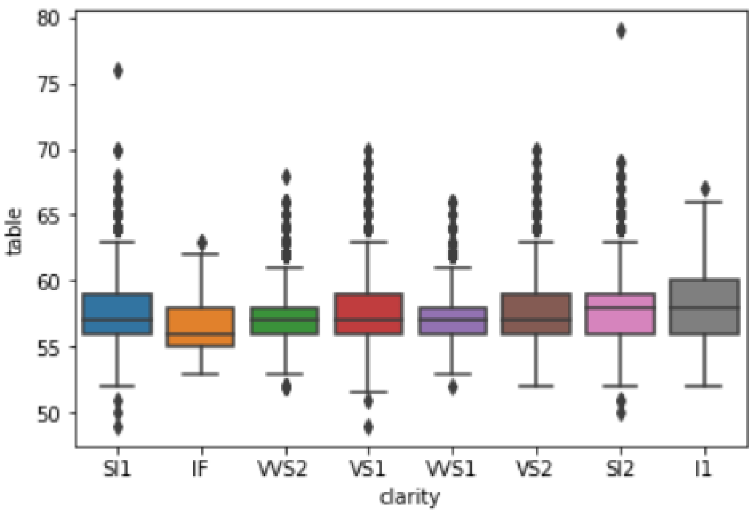
```
Plot of table vs color
........................
```



- Clarity code I1 has maximum and clarity code IF has minimum mean table in cubic zirconia

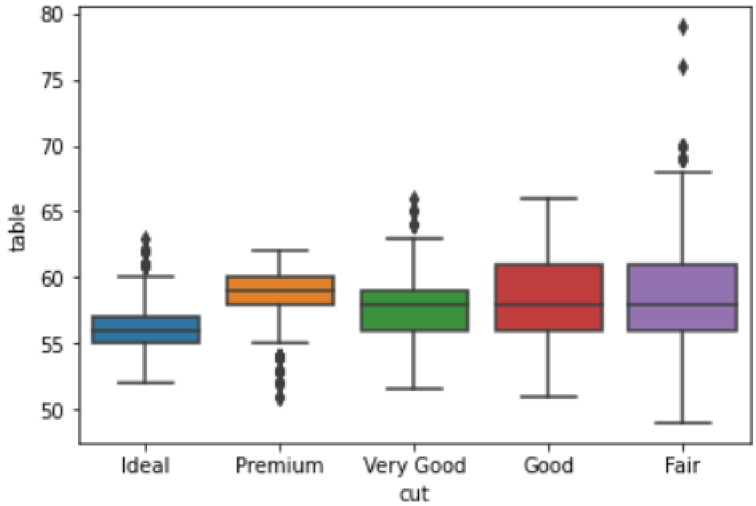
```
Mean of table for clarity
clarity
I1      58.376923
IF      56.449270
SI1     57.637106
SI2     57.912007
VS1     57.322021
VS2     57.429805
VVS1    56.910984
VVS2    57.060632
Name: table, dtype: float64
```

```
Plot of table vs clarity
........................
```



- Fair cut has maximum and Ideal cut has minimum mean table in cubic zirconia

```
Mean of table for cut
cut
Fair         59.300513
Good         58.703860
Ideal        55.956205
Premium      58.714406
Very Good    57.963929
Name: table, dtype: float64

Plot of table vs cut
........................
```

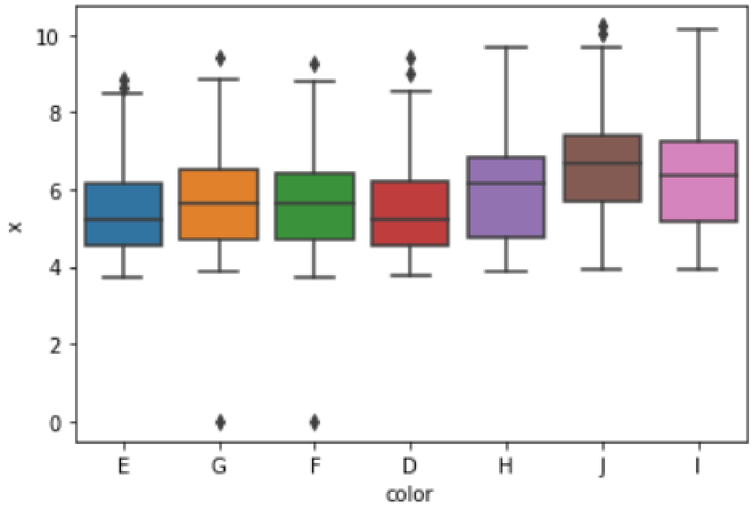


Length ('x') with Other categorical attributes

- Color code J has maximum and color code E has minimum mean length in cubic zirconia

```
Mean of x for color
color
D    5.414385
E    5.403961
F    5.598562
G    5.678289
H    5.979648
I    6.236796
J    6.514146
Name: x, dtype: float64
```

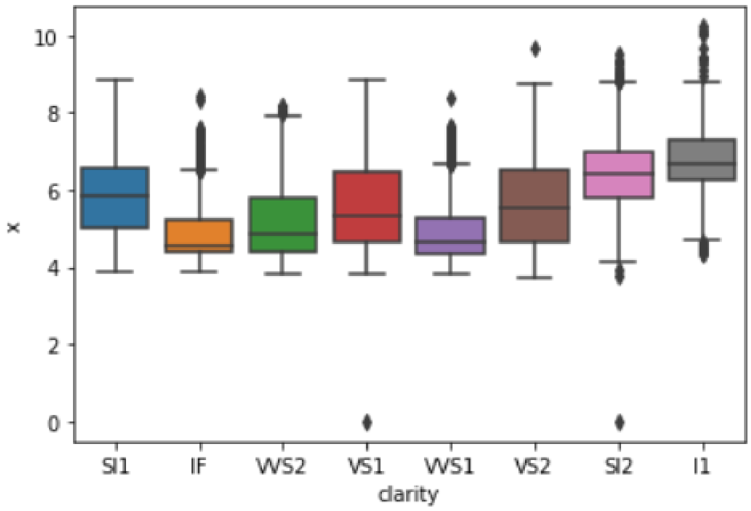
```
Plot of table vs color
........................
```



- Clarity code I1 has maximum and clarity code IF has minimum mean length in cubic zirconia

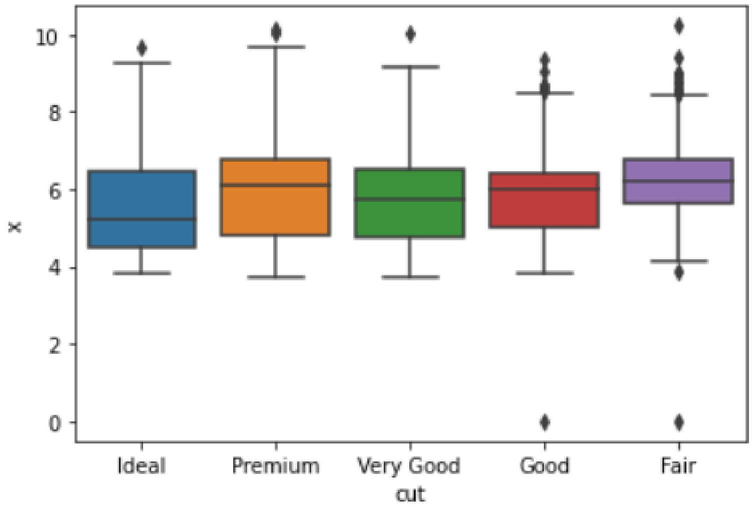
```
Mean of x for clarity
clarity
I1      6.758132
IF      4.943962
SI1     5.884967
SI2     6.411873
VS1     5.567127
VS2     5.665168
VVS1    4.946900
VVS2    5.208213
Name: x, dtype: float64
```

```
Plot of table vs clarity
........................
```



- Fair cut has maximum and Ideal cut has minimum mean length in cubic zirconia

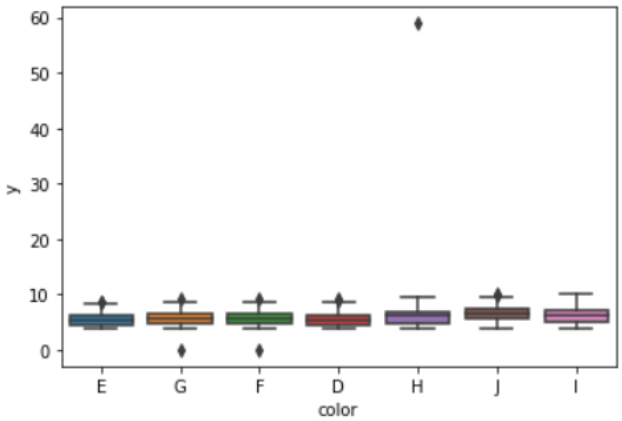
```
Mean of x for cut
cut
Fair         6.284244
Good         5.841326
Ideal        5.500229
Premium      5.966265
Very Good    5.752359
Name: x, dtype: float64

Plot of table vs cut
.........................
```



Width ('y') with Other categorical attributes

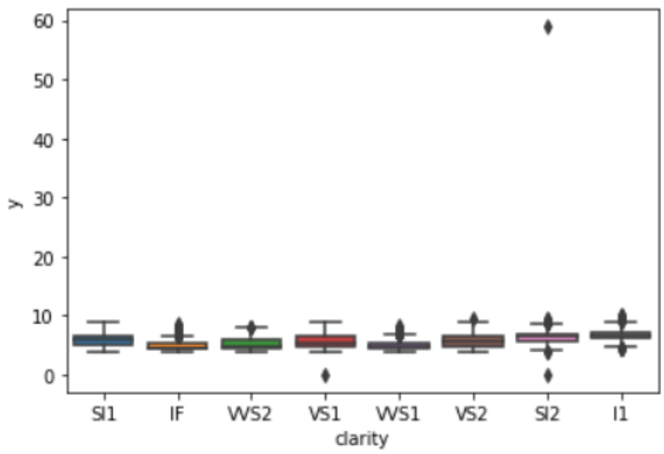
- Color code J has maximum and color code E has minimum mean width in cubic zirconia

```
Mean of y for color
color
D    5.419129
E    5.409329
F    5.602494
G    5.680258
H    5.987057
I    6.236604
J    6.513729
Name: y, dtype: float64

Plot of y vs color
.........................
```



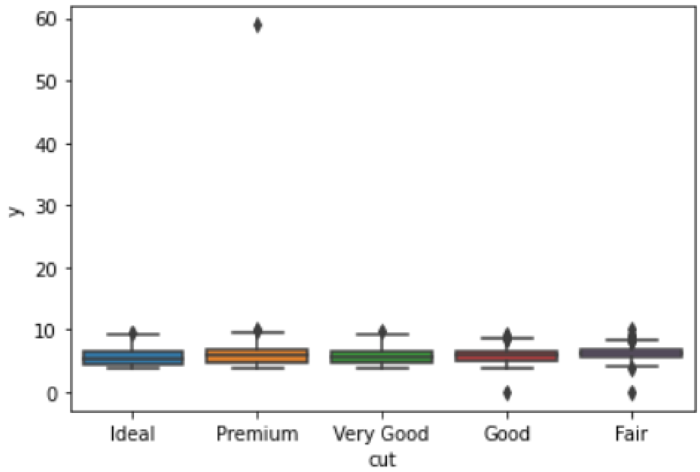
- Clarity code I1 has maximum and clarity code VVS1 has minimum mean width in cubic zirconia

```
Mean of y for clarity
clarity
I1      6.708379
IF      4.965230
SI1     5.885100
SI2     6.413149
VS1     5.572190
VS2     5.666368
VVS1    4.962501
VVS2    5.222810
Name: y, dtype: float64

Plot of y vs clarity
.........................
```



- Fair cut has maximum and Ideal cut has minimum mean width in cubic zirconia

```
Mean of y for cut
cut
Fair         6.216179
Good         5.856033
Ideal        5.511296
Premium      5.940520
Very Good    5.781583
Name: y, dtype: float64

Plot of y vs cut
.........................
```

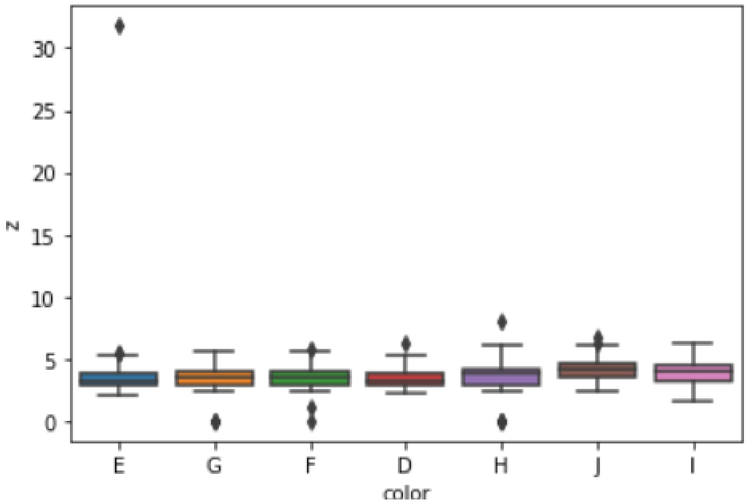


Height ('z') with Other categorical attributes

- Color code J has maximum and color code E has minimum mean height in cubic zirconia

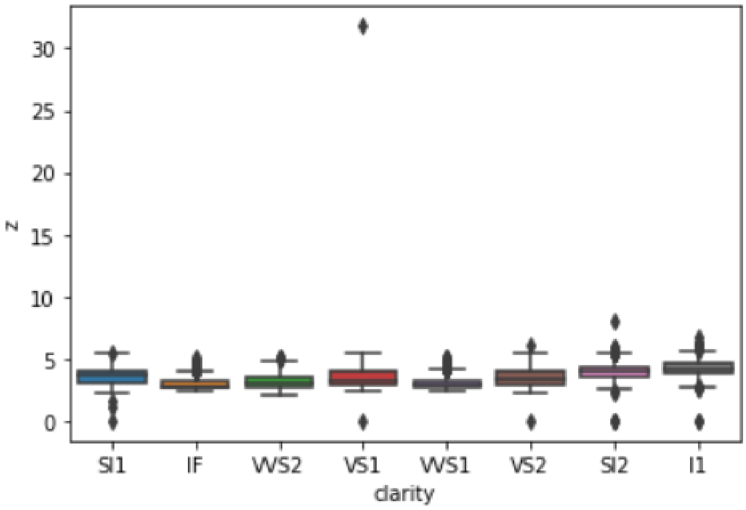
```
Mean of z for color
color
D    3.341152
E    3.338973
F    3.453242
G    3.505128
H    3.691353
I    3.855732
J    4.030708
Name: z, dtype: float64
```

```
Plot of z vs color
.........................
```



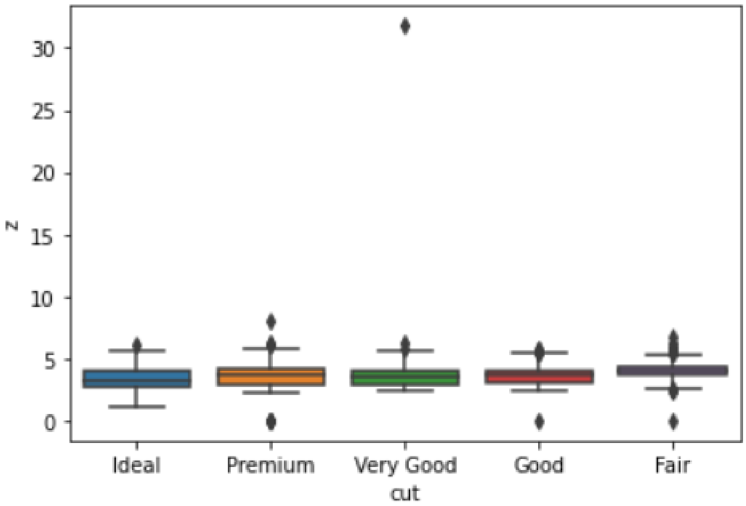
- Clarity code I1 has maximum and clarity code IF has minimum mean height in cubic zirconia

```
Mean of z for clarity
clarity
I1      4.194313
IF      3.045567
SI1     3.637467
SI2     3.955703
VS1     3.440763
VS2     3.495383
VVS1    3.053861
VVS2    3.214542
Name: z, dtype: float64

Plot of z vs clarity
.........................
```



- Fair cut has maximum and Ideal cut has minimum mean carat height in cubic zirconia

```
Mean of z for cut
cut
Fair         3.993013
Good         3.644678
Ideal        3.396558
Premium      3.642084
Very Good    3.569637
Name: z, dtype: float64

Plot of z vs cut
.........................
```

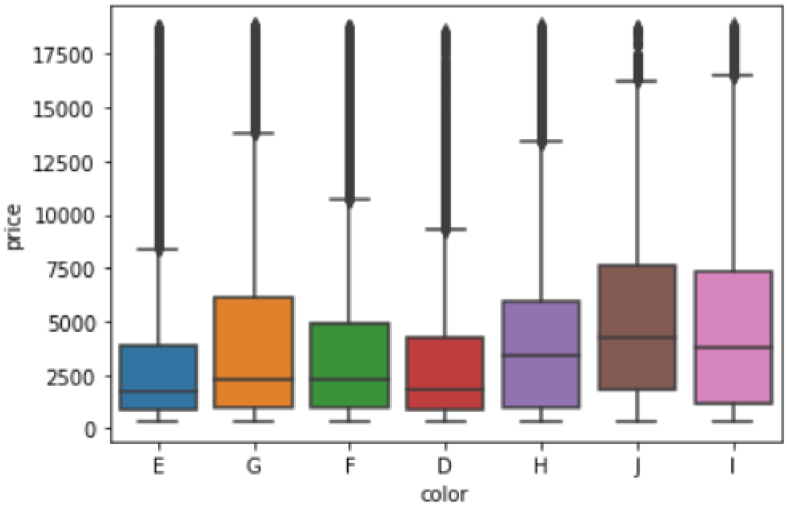


Price with Other categorical attributes

- Color code J has maximum and color code E has minimum mean price in cubic zirconia

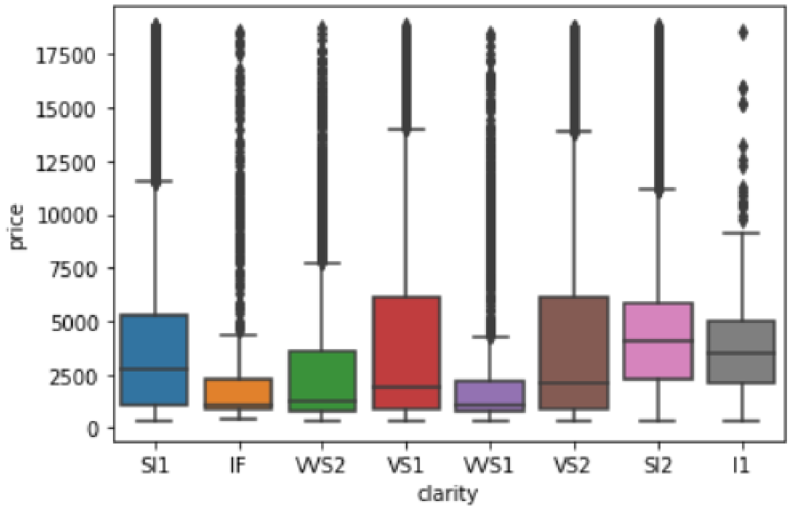
```
Mean of price for color
color
D    3184.827597
E    3073.940399
F    3699.944527
G    4005.046170
H    4477.932112
I    5124.816637
J    5329.706250
Name: price, dtype: float64
```

```
Plot of price vs color
.........................
```



- Clarity code SI2 has maximum and clarity code VVS1 has minimum mean price in cubic zirconia

```
Mean of price for clarity
clarity
I1       3908.750000
IF       2739.534231
SI1      3998.635644
SI2      5088.869413
VS1      3838.752386
VS2      3965.496964
VVS1     2502.874388
VVS2     3263.042688
Name: price, dtype: float64

Plot of price vs clarity
.........................
```

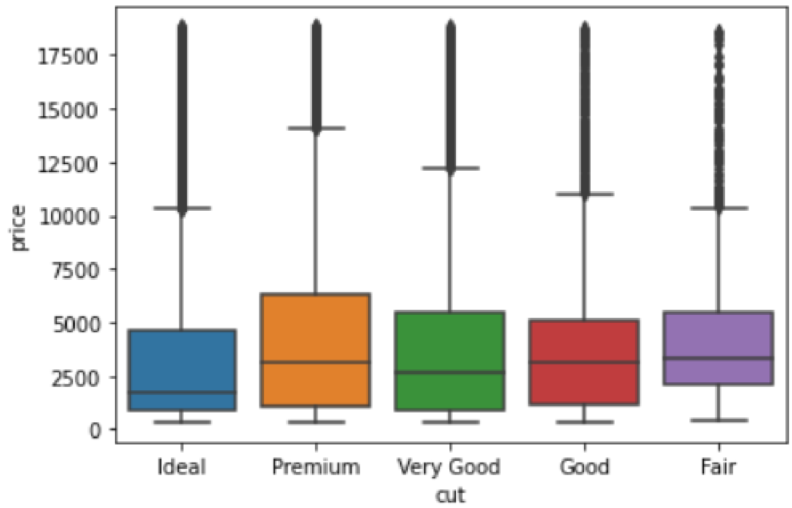


- Fair cut has maximum and Ideal cut has minimum mean price in cubic zirconia

```
Mean of price for cut
cut
Fair          4568.096154
Good          3926.336756
Ideal         3454.820639
Premium       4544.558525
Very Good     4032.267961
Name: price, dtype: float64
```

Plot of price vs cut

....



Problem 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? Check for the possibility of combining the sub levels of a ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning.

Null Value Imputation

- There are 697 NAN values present in the 'depth' attribute. From the correlation matrix we have seen that 'depth' attributes has very low correlation with the price attribute.
- Also, we have seen that 'depth' attribute has approx. normally distributed. Thus, we can use the mean or median values to impute the Null values present in the variable.
- Median value is used for imputation for 'depth' attribute.

Significance of '0' values and its presence in the dataset

- From the data description we have seen anomalies in few attributes. Independent attributes 'X', 'Y', 'Z' are physical measurements. Hence a 0 value seems nonrational
- Moreover, in real life scenario price for 0 value will not give a relevant outcome
- Also, we have removed a duplicate instance that had '0' values for 'x', 'y' and 'z'.
- We have observed that there are two instances where 'x', 'y' and 'z' are all 0. And six other instances where only z is 0, which is 0.03% of the dataset.

Hence, they have been dropped from the original dataset and the dataset shape has changed to (26925,10)

<u>Combining the sub levels of an ordinal variables</u>

- From the frequency table we have seen that cut type 'Premium' and 'Very Good' cut type have similar instances. Also, from the num-cat analysis we have observed that they show similar instance.
- We have checked the correlation matrix after converting the categorical variable into numerical variable
- It can be seen from the correlation matrix that after combining the ordinal sub-level of 'cut' attribute magnitude of correlation changed between the price and cut attributes.
- Hence, we have combined them into one sub-level and saved it another cloned dataset of original dataset
- Further, we have analyzed the performance of the sub-level combined cloned dataset in a model building

	carat	cut
carat	1.000000	-0.139908
cut	-0.139908	1.000000
color	0.293760	-0.026809
clarity	0.354786	-0.183331
depth	0.035070	-0.212275
table	0.181511	-0.443049
x	0.977908	-0.132945
y	0.942378	-0.127480
z	0.946774	-0.154347
price	0.922400	-0.059806

Figure 7: Correlation matrix before combining the sub-level

	carat	cut
carat	1.000000	-0.166600
cut	-0.166600	1.000000
color	0.293760	-0.035903
clarity	0.354786	-0.210420
depth	0.035070	-0.189732
table	0.181511	-0.498909
x	0.977908	-0.162478
y	0.942378	-0.150097
z	0.946774	-0.176546
price	0.922400	-0.077190

Figure 8: Correlation matrix before combining the sub-level

Problem 1.3

Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from stats model. Create multiple models and check the performance of Predictions on Train and Test sets using Square, RMSE & Adj Square. Compare these models and select the best one with appropriate reasoning.

- To find the most optimized model we have done multiple models based on different treatments, factors etc.
- For better understanding we have divided our multiple models into two broad categories A & B
- Category A contains models with outliers treatment whereas, category B contains models without outliers treatment

Category A With Outlier treatment (OT)

Model – 1: Only OT is done

Model- 2: OT + Scaling is done

Model- 3: OT + combining the sub levels of ‘cut’ variable is done

Model- 4: OT+ dropping of ‘depth’ variable is done

Model- 5: OT+ Variance Inflation factor is done

Model-6 : OT+ Scaling+ Variance Inflation factor is done

Without Outlier treatment (No OT)

Category B Without Outlier treatment (OT)

Model- 7: No OT is done

Model- 8: No OT + Scaling is done

Model- 9: No OT+ Variance Inflation factor is done

Model- 10: No OT+ Scaling+ Variance Inflation factor is done

Cat-A:

- Outliers may affect the linear regression models; we have 1st checked the total percentage of outliers in the data set. We observed as follows:
    - 'carat' has all the outliers on the upper side, around 2.44% of the total data points present.
    - 'depth' has outliers on either side, a total of around 5.24% of the total data points present.
    - 'table' has outliers on either side, a total of around 1.18% of the total data points present.
    - 'x', 'y' and 'z' have few outliers on either side, a total of around 0.05%, 0.05% and 0.08% of the total data points present.
    -
- Further, we have replaced the outliers are capped by the appropriate whisker values of the corresponding variables. There are no more outliers as measured by IQR method.
- Categorical value has been encoded and converted into numeric data type
- We have performed the model building by using both scikit learn and statsmodel

Model -1

- Model 1 has been treated with outlier Treatment
- Sub-level of ordinal categories are not combined
- The intercept for model-1 is 9146, which means in in present case when the other predictor variables are zero i.e., like carat, cut, colour, clarity all are zero then the C=-3. (Y = m1X1 + m2X2+ …. + mnXn + C + e) that means price is 9146. which is meaningless. We can do Z score or scaling the data and make it nearly zero.
- R^2 and Adjusted R^2 for model are nearly equal (0.92)
- Hence, we have assumed all the independent variables play a significant role in the prediction of price and should be kept for better prediction
- RMSE for this model is 1156
- Prob (F-statistics) is 0 (less than alpha) for the model, hence rejected the null hypothesis
- We have observed the p(t) with respect to 'depth' attributes is 0.471 and its coefficient magnitude is much higher than 0, which is slippery point in this model
- We have not treated the multicollinearity hence, cannot rely on the coefficient.

```
Dep. Variable:                  price   R-squared:                       0.917
Model:                            OLS   Adj. R-squared:                  0.917
Method:                 Least Squares   F-statistic:                 2.301e+04
Date:                Fri, 20 May 2022   Prob (F-statistic):               0.00
Time:                        18:38:08   Log-Likelihood:            -1.5969e+05
No. Observations:               18847   AIC:                         3.194e+05
Df Residuals:                   18837   BIC:                         3.195e+05
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   9145.7570   1017.388      8.989      0.000    7151.586    1.11e+04
carat       1.379e+04    105.369    130.908      0.000    1.36e+04     1.4e+04
cut          130.8333      9.333     14.019      0.000     112.540     149.127
color       -327.3451      5.239    -62.481      0.000    -337.614    -317.076
clarity     -480.1245      5.706    -84.147      0.000    -491.308    -468.941
depth        -10.2118     14.157     -0.721      0.471     -37.960      17.536
table        -33.6326      4.994     -6.734      0.000     -43.422     -23.843
x          -2473.5489    172.779    -14.316      0.000   -2812.211   -2134.887
y           1550.4375    169.810      9.130      0.000    1217.594    1883.281
z          -1677.6340    177.497     -9.452      0.000   -2025.544   -1329.724
==============================================================================
Omnibus:                     3548.194   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            33949.762
Skew:                           0.629   Prob(JB):                         0.00
Kurtosis:                       9.454   Cond. No.                     1.03e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correct
ly specified.
[2] The condition number is large, 1.03e+04. This might indicate that there ar
e
strong multicollinearity or other numerical problems.
```

Figure 9: Summary table model-1

Model -2

- Model 2 has been treated with outlier Treatment and Scaled with z-score scaling
- Sub-level of ordinal categories is not combined
- The intercept for model-2 is -3.209e-17, which means in in present case when the other predictor variables are zero i.e., like carat, cut, color, clarity all are zero then the C=-3. (Y = m1X1 + m2X2+ …. + mnXn + C + e) that means price tends to 0. Which seems rational
- R^2 and Adjusted R^2 for model are nearly equal (0.92)
- RMSE for this model is 0.29
- Prob (F-statistics) is 0 (less than alpha) for the model, hence rejected the null hypothesis
- We have observed the p(t) with respect to ‘depth’ attributes is 0.471 and its coefficient magnitude tends to 0, hence we have assumed it does not play any significant role in price prediction
- We have not treated the multicollinearity hence, cannot rely on the coefficient.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.917
Model:                            OLS   Adj. R-squared:                  0.917
Method:                 Least Squares   F-statistic:                 2.301e+04
Date:                Fri, 20 May 2022   Prob (F-statistic):               0.00
Time:                        18:43:11   Log-Likelihood:                -3332.8
No. Observations:               18847   AIC:                             6686.
Df Residuals:                   18837   BIC:                             6764.
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -3.209e-17      0.002  -1.53e-14      1.000      -0.004       0.004
carat          1.5826      0.012    130.908      0.000       1.559       1.606
cut            0.0363      0.003     14.019      0.000       0.031       0.041
color         -0.1392      0.002    -62.481      0.000      -0.144      -0.135
clarity       -0.1975      0.002    -84.147      0.000      -0.202      -0.193
depth         -0.0031      0.004     -0.721      0.471      -0.012       0.005
table         -0.0181      0.003     -6.734      0.000      -0.023      -0.013
x             -0.6919      0.048    -14.316      0.000      -0.787      -0.597
y              0.4307      0.047      9.130      0.000       0.338       0.523
z             -0.2900      0.031     -9.452      0.000      -0.350      -0.230
==============================================================================
Omnibus:                     3548.194   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            33949.762
Skew:                           0.629   Prob(JB):                         0.00
Kurtosis:                       9.454   Cond. No.                         62.9
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correct
ly specified.
```

Model -3

- Model 3 has been treated with outlier Treatment and the sub-level value of ordinal category have also been combined
- The intercept for model-3 is 9296, which means in in present case when the other predictor variables are zero i.e., like carat, cut, color, clarity all are zero then the C=-3. (Y = m1X1 + m2X2+ …. + mnXn + C + e) that means price is 9296. which is meaningless.
- R^2 and Adjusted R^2 for model are nearly equal (0.92)
- Hence, we have assumed all the independent variables play a significant role in the prediction of price and should be kept for better prediction
- Prob (F-statistics) is 0 (less than alpha) for the model, hence rejected the null hypothesis
- RMSE for this model is 1157
- We have observed the p(t) with respect to ‘depth’ attributes is 0.2511 and its coefficient magnitude is much higher than 0, which is slippery point in this model
- We have not treated the multicollinearity hence, cannot rely on the coefficient.

```
==============================================================================
Dep. Variable:                  price   R-squared:                       0.917
Model:                            OLS   Adj. R-squared:                  0.917
Method:                 Least Squares   F-statistic:                 2.300e+04
Date:                Fri, 20 May 2022   Prob (F-statistic):               0.00
Time:                        16:29:01   Log-Likelihood:            -1.5969e+05
No. Observations:               18847   AIC:                         3.194e+05
Df Residuals:                   18837   BIC:                         3.195e+05
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   9296.3743   1014.429      9.164      0.000    7308.003    1.13e+04
carat        1.38e+04    105.372    130.947      0.000    1.36e+04     1.4e+04
cut          202.0082     14.510     13.922      0.000     173.567     230.449
color       -326.9693      5.240    -62.404      0.000    -337.239    -316.699
clarity     -479.6436      5.710    -84.005      0.000    -490.835    -468.452
depth        -16.1612     14.073     -1.148      0.251     -43.745      11.423
table        -31.0770      5.104     -6.089      0.000     -41.081     -21.073
x          -2121.4499    171.844    -12.345      0.000   -2458.279   -1784.620
y           1168.0355    169.027      6.910      0.000     836.728    1499.343
z          -1631.4311    177.581     -9.187      0.000   -1979.506   -1283.356
==============================================================================
Omnibus:                     3554.509   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            33736.640
Skew:                           0.633   Prob(JB):                         0.00
Kurtosis:                       9.431   Cond. No.                     1.03e+04
==============================================================================
```

Model -4

- Model 4 has been treated with outlier Treatment and depth attributes has been dropped
- Sub-level of ordinal categories is not combined
- The intercept for model-4 is 8453, which means in in present case when the other predictor variables are zero i.e., like carat, cut, color, clarity all are zero then the C=-3. (Y = m1X1 + m2X2+ …. + mnXn + C + e) that means price is 8453. which is meaningless.
- R^2 and Adjusted R^2 for model are nearly equal (0.92)
- Hence, we have assumed all the independent variables play a significant role in the prediction of price and should be kept for better prediction
- RMSE for this model is 1156
- Prob (F-statistics) is 0 (less than alpha) for the model, hence rejected the null hypothesis
- We have not treated the multicollinearity hence, cannot rely on the coefficient

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.917
Model:                            OLS   Adj. R-squared:                  0.917
Method:                 Least Squares   F-statistic:                 2.588e+04
Date:                Sat, 21 May 2022   Prob (F-statistic):               0.00
Time:                        04:57:47   Log-Likelihood:            -1.5969e+05
No. Observations:               18847   AIC:                         3.194e+05
Df Residuals:                   18838   BIC:                         3.195e+05
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   8453.3571    337.242     25.066      0.000    7792.332    9114.382
carat       1.379e+04    105.018    131.287      0.000     1.36e+04     1.4e+04
cut          132.3327      9.098     14.545      0.000     114.499     150.166
color       -327.4516      5.237    -62.527      0.000    -337.716    -317.187
clarity     -480.3010      5.700    -84.256      0.000    -491.474    -469.128
table        -32.7856      4.854     -6.754      0.000     -42.301     -23.271
x          -2446.0024    168.504    -14.516      0.000   -2776.285   -2115.720
y           1589.5151    160.935      9.877      0.000    1274.068    1904.963
z          -1781.4102    103.962    -17.135      0.000   -1985.186   -1577.635
==============================================================================
Omnibus:                     3549.783   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            33990.190
Skew:                           0.629   Prob(JB):                         0.00
Kurtosis:                       9.458   Cond. No.                     2.41e+03
==============================================================================
```

Model -5

- Model 5 has been treated with outlier Treatment and multicollinearity checked with variation inflation factor and further treated to reduce multicollinearity
- Sub-level of ordinal categories is not combined
- The intercept for model-5 is -1.476e+04, which means in in present case when the other predictor variables are zero i.e., like carat, cut, color, clarity all are zero then the C=-3. (Y = m1X1 + m2X2+ …. + mnXn + C + e) that means price tends to -1.476e+04. which seems nonrational
- R^2 and Adjusted R^2 for model are nearly equal (0.91)
- Hence, we have assumed all the independent variables play a significant role in the prediction of price and should be kept for better prediction
- RMSE for this model is 1225
- We have observed the p(t) with respect to ‘depth’ attribute is 0.075. its coefficient magnitude is nearly 0. Its coefficient is much higher than 0, which is a slippery point in this model
- Prob (F-statistics) is 0 (less than alpha) for the model, hence rejected the null hypothesis

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.064
Model:                            OLS   Adj. R-squared:                  0.063
Method:                 Least Squares   F-statistic:                     255.6
Date:                Sat, 21 May 2022   Prob (F-statistic):          3.64e-265
Time:                        04:57:48   Log-Likelihood:            -1.8248e+05
No. Observations:               18847   AIC:                         3.650e+05
Df Residuals:                   18841   BIC:                         3.650e+05
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -1.476e+04   2269.746     -6.502      0.000   -1.92e+04   -1.03e+04
cut          115.7042     31.049      3.727      0.000      54.845     176.563
color        407.9856     16.628     24.535      0.000     375.392     440.579
clarity      328.6280     17.568     18.706      0.000     294.194     363.062
depth         47.4988     26.641      1.783      0.075      -4.720      99.718
table        227.3681     16.359     13.899      0.000     195.303     259.433
==============================================================================
Omnibus:                     5401.033   Durbin-Watson:                   1.990
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            12698.778
Skew:                           1.629   Prob(JB):                         0.00
Kurtosis:                       5.357   Cond. No.                     6.79e+03
==============================================================================
```

Model -6

- Model 6 has been treated with outlier Treatment and multicollinearity checked with variation inflation factor and further treated to reduce multicollinearity
- It has been scaled with Z-score
- Sub-level of ordinal categories is not combined
- The intercept for model-6 is -1.988e-17, which means in in present case when the other predictor variables are zero i.e., like carat, cut, color, clarity all are zero then the C=-3. (Y = m1X1 + m2X2+ …. + mnXn + C + e) that means price tends to 0. which seems rational
- R^2 and Adjusted R^2 for model are nearly equal (0.91)
- RMSE for this model is 0.28
- We have observed the p(t) with respect to 'table' attribute is 0.310. its coefficient magnitude is nearly 0. Hence, we have assumed table attribute is less significant for price prediction
- Prob (F-statistics) is 0 (less than alpha) for the model, hence rejected the null hypothesis

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.841
Model:                            OLS   Adj. R-squared:                  0.841
Method:                 Least Squares   F-statistic:                 1.656e+04
Date:                Sat, 21 May 2022   Prob (F-statistic):               0.00
Time:                        04:57:48   Log-Likelihood:                -9436.7
No. Observations:               18847   AIC:                         1.889e+04
Df Residuals:                   18840   BIC:                         1.894e+04
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -1.988e-17      0.003  -6.84e-15      1.000      -0.006       0.006
cut            0.0435      0.004     12.233      0.000       0.037       0.050
color         -0.1091      0.003    -35.633      0.000      -0.115      -0.103
clarity       -0.2236      0.003    -69.698      0.000      -0.230      -0.217
depth          0.0497      0.003     14.868      0.000       0.043       0.056
table          0.0037      0.004      1.015      0.310      -0.003       0.011
y              1.0067      0.003    303.086      0.000       1.000       1.013
==============================================================================
Omnibus:                     4345.601   Durbin-Watson:                   2.003
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            10299.292
Skew:                           1.290   Prob(JB):                         0.00
Kurtosis:                       5.541   Cond. No.                         2.23
==============================================================================
```

Cat-B:

- All the models in category B are not treated with the outlier treatment since:
    - In this particular problem outliers may have some importance in price prediction
    - For instance, in the correlation matrix, we have seen carat has a very strong positive correlation with price. Hence, higher the carat weight higher will be price. If we remove outliers, we will be unable to predict the price for the high carat weight
    - Dataset has been encoded and converted into numeric datatype

Model -7

- Model -7 has not been treated with outlier Treatment
- Sub-level of ordinal categories is not combined
- The intercept for model-7 is 1.047e+04, which means in in present case when the other predictor variables are zero i.e., like carat, cut, color, clarity all are zero then the C=-3. (Y = m1X1 + m2X2+ …. + mnXn + C + e) that means price tends to 1.047e+04. which seems nonrational
- R^2 and Adjusted R^2 for model are nearly equal (0.91), Hence, we have assumed all the independent variables play a significant role in the prediction of price and should be kept for better prediction

- RMSE for this model is 1218
- Prob (F-statistics) is 0 (less than alpha) for the model, hence rejected the null hypothesis
- We have observed the p(t) with respect to 'y' and 'z' attributes is 0.780 and 0.312 respectively; However, its coefficient magnitude is much higher than 0, which is slippery point in this model
- We have not treated the multicollinearity hence, cannot rely on the coefficient

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.908
Model:                            OLS   Adj. R-squared:                  0.908
Method:                 Least Squares   F-statistic:                 2.065e+04
Date:                Sat, 21 May 2022   Prob (F-statistic):               0.00
Time:                        10:35:25   Log-Likelihood:            -1.6062e+05
No. Observations:               18847   AIC:                         3.213e+05
Df Residuals:                   18837   BIC:                         3.213e+05
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   1.047e+04    711.141     14.721      0.000    9074.799    1.19e+04
carat       1.105e+04     93.359    118.409      0.000    1.09e+04    1.12e+04
cut          107.3618      9.738     11.025      0.000      88.274     126.450
color       -329.6089      5.507    -59.855      0.000    -340.403    -318.815
clarity     -502.9605      5.957    -84.432      0.000    -514.637    -491.284
depth        -84.3959      7.862    -10.734      0.000     -99.807     -68.985
table        -35.5932      5.011     -7.103      0.000     -45.415     -25.771
x           -951.9872     50.865    -18.716      0.000   -1051.687    -852.287
y              6.6726     23.895      0.279      0.780     -40.164      53.509
z            -42.1737     41.726     -1.011      0.312    -123.960      39.613
==============================================================================
Omnibus:                     4196.790   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           205176.201
Skew:                          -0.059   Prob(JB):                         0.00
Kurtosis:                      19.164   Cond. No.                     6.84e+03
==============================================================================
```

Model -8

- Model-8 has not been treated with outlier Treatment
- It has been scaled with z-score
- Sub-level of ordinal categories is not combined
- The intercept for model-7 is-3.209e-17, which means in present case when the other predictor variables are zero i.e., like carat, cut, color, clarity all are zero then the C=-3. (Y = m1X1 + m2X2+ …. + mnXn + C + e) that means price tends to 0. which seems rational
- R^2 and Adjusted R^2 for model are nearly equal (0.91)
- RMSE for this model is 0.30
- Prob (F-statistics) is 0 (less than alpha) for the model, hence rejected the null hypothesis

- We have observed the p(t) with respect to 'y' and 'z' attributes is 0.780 and 0.312 respectively; its coefficient magnitude is nearly 0. Which implies that these two attributes do not play important roles in price prediction.
- We have not treated the multicollinearity hence, cannot rely on the coefficient

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.908
Model:                            OLS   Adj. R-squared:                  0.908
Method:                 Least Squares   F-statistic:                 2.065e+04
Date:                Sat, 21 May 2022   Prob (F-statistic):               0.00
Time:                        10:35:38   Log-Likelihood:                -4260.0
No. Observations:               18847   AIC:                             8540.
Df Residuals:                   18837   BIC:                             8618.
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -3.209e-17      0.002  -1.45e-14      1.000      -0.004       0.004
carat          1.3100      0.011    118.409      0.000       1.288       1.332
cut            0.0298      0.003     11.025      0.000       0.025       0.035
color         -0.1401      0.002    -59.855      0.000      -0.145      -0.136
clarity       -0.2069      0.002    -84.432      0.000      -0.212      -0.202
depth         -0.0294      0.003    -10.734      0.000      -0.035      -0.024
table         -0.0198      0.003     -7.103      0.000      -0.025      -0.014
x             -0.2664      0.014    -18.716      0.000      -0.294      -0.239
y              0.0020      0.007      0.279      0.780      -0.012       0.016
z             -0.0076      0.008     -1.011      0.312      -0.022       0.007
==============================================================================
Omnibus:                     4196.790   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           205176.201
Skew:                          -0.059   Prob(JB):                         0.00
Kurtosis:                      19.164   Cond. No.                         15.9
==============================================================================
```

Model -9

- Model 9 has not been treated with outlier Treatment and multicollinearity checked with variation inflation factor and further treated to reduce multicollinearity
- Sub-level of ordinal categories is not combined
- The intercept for model-9 is 4149, which means in in present case when the other predictor variables are zero i.e., like carat, cut, color, clarity all are zero then the C=-3. (Y = m1X1 + m2X2+ …. + mnXn + C + e) that means price tends to 4149 which seems nonrational.
- R^2 and Adjusted R^2 for model are nearly equal (0.91)
- Hence, we have assumed all the independent variables play a significant role in the prediction of price and should be kept for better prediction

- RMSE for this model is 1236
- Prob (F-statistics) is 0 (less than alpha) for the model, hence rejected the null hypothesis

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                  price   R-squared:                       0.905
Model:                            OLS   Adj. R-squared:                  0.905
Method:                 Least Squares   F-statistic:                 2.993e+04
Date:                Sat, 21 May 2022   Prob (F-statistic):               0.00
Time:                        10:38:26   Log-Likelihood:            -1.6091e+05
No. Observations:               18847   AIC:                         3.218e+05
Df Residuals:                   18840   BIC:                         3.219e+05
Df Model:                           6                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   4149.3290    655.314      6.332      0.000    2864.855    5433.803
cut          111.6249      9.884     11.293      0.000      92.250     130.999
color       -324.5602      5.589    -58.068      0.000    -335.516    -313.605
clarity     -526.2608      5.972    -88.121      0.000    -537.966    -514.555
depth        -43.4857      7.391     -5.884      0.000     -57.973     -28.999
table        -34.4307      5.084     -6.772      0.000     -44.396     -24.466
carat       8830.9522     21.603    408.783      0.000    8788.608    8873.296
==============================================================================
Omnibus:                     3816.891   Durbin-Watson:                   1.996
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            71040.120
Skew:                           0.474   Prob(JB):                         0.00
Kurtosis:                      12.464   Cond. No.                     6.16e+03
==============================================================================
```

Model -10

- Model 9 has not been treated with outlier Treatment and multicollinearity checked with variation inflation factor and further treated to reduce multicollinearity
- Data has been scaled
- Sub-level of ordinal categories is not combined
- The intercept for model-10 is -3.209e-17, which means in in present case when the other predictor variables are zero i.e., like carat, cut, color, clarity all are zero then the C=-3. (Y = m1X1 + m2X2+ …. + mnXn + C + e) that means price tends to 0, which seems rational.
- R^2 and Adjusted R^2 for model are nearly equal (0.91)
- Hence, we have assumed all the independent variables play a significant role in the prediction of price and should be kept for better prediction
- RMSE for this model is 0.31
- Prob (F-statistics) is 0 (less than alpha) for the model, hence rejected the null hypothesis

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.905
Model:                            OLS   Adj. R-squared:                  0.905
Method:                 Least Squares   F-statistic:                 2.993e+04
Date:                Sat, 21 May 2022   Prob (F-statistic):               0.00
Time:                        10:39:31   Log-Likelihood:                -4555.3
No. Observations:               18847   AIC:                             9125.
Df Residuals:                   18840   BIC:                             9180.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -3.209e-17      0.002  -1.43e-14      1.000      -0.004       0.004
cut            0.0310      0.003     11.293      0.000       0.026       0.036
color         -0.1380      0.002    -58.068      0.000      -0.143      -0.133
clarity       -0.2165      0.002    -88.121      0.000      -0.221      -0.212
depth         -0.0152      0.003     -5.884      0.000      -0.020      -0.010
table         -0.0192      0.003     -6.772      0.000      -0.025      -0.014
carat          1.0465      0.003    408.783      0.000       1.041       1.051
==============================================================================
Omnibus:                     3816.891   Durbin-Watson:                   1.996
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            71040.120
Skew:                           0.474   Prob(JB):                         0.00
Kurtosis:                      12.464   Cond. No.                         2.24
==============================================================================
```

Observation

Model Name	R^2	Adjusted R^2	RMSE	Intercept
Model- 1	0.92	0.92	1156	9146
Model- 2	0.92	0.92	0.29	-3.209e-17
Model- 3	0.92	0.92	1157	9296
Model- 4	0.92	0.92	1156	8453
Model- 5	0.91	0.91	1225	-1.476e+04
Model- 6	0.91	0.91	0.28	-1.988e-17
Model- 7	0.91	0.91	1218	1.047e+04
Model- 8	0.91	0.91	0.30	-3.209e-17
Model- 9	0.91	0.91	1236	4149
Model- 10	0.91	0.91	0.31	-3.209e-17

Figure 10: Comparison table of all the models

With Outlier treatment (OT)

Model – 1: Only OT is done

Model- 2: OT + Scaling is done

Model- 3: OT + combining the sub levels of ‘cut’ variable is done

Model- 4: OT+ dropping of ‘depth’ variable is done

Model- 5: OT+ Variance Inflation factor is done
Model-6 : OT+ Scaling+ Variance Inflation factor is done

Without Outlier treatment (No OT)
Model- 7: No OT is done
Model- 8: No OT + Scaling is done
Model- 9: No OT+ Variance Inflation factor is done
Model- 10: No OT+ Scaling+ Variance Inflation factor is done

- It is evident from the table that R^2 and Adjusted R^2 is similar in model (1, 2, 3, 4) and similar in model (5,6,7,8,9,10)
- The intercept for unscaled models is much higher than 0 which means in in present case when the other predictor variables are zero i.e., like carat, cut, color, clarity all are zero then the C=-3. (Y = m1X1 + m2X2+ …. + mnXn + C + e) that means price tends to have higher number, which is unlikely
- The intercept for scaled models is tends to 0 which means in in present case when the other predictor variables are zero i.e., like carat, cut, color, clarity all are zero then the C=-3. (Y = m1X1 + m2X2+ …. + mnXn + C + e) that means price tends to have higher number, which is very likely
- RMSE for all the scaled models are within the range of 0.28 to 0.31
- P(f-statistics) is less than alpha value (0.05) hence, we have rejected the null hypothesis
- Since, we have seen in model-1 and model -3 there is no significant difference, we proceed with the dataset without combining the sub-level in ordinal value. We have encoded them before making the model

Selection of Model

- Lower the RMSE, better is the performance
- We have seen models with outliers treatment shown better performance than without outliers treatment. However, we have found that the percentage of outliers in certain attributes is significantly higher (In carat 2.24%) and the outliers are not anomalies as well. Hence, we are proceeding with the models which have outliers
- Model 7 and 8 gives similar performance, However, multicollinearity has not been treated in these two models. Hence, we cannot rely on the coefficient of these two models
- Model 9 is treated with the multicollinearity but it has not been scaled hence the intercept value is quite high, which is very unlikely
- Model 10 seems to be the suitable model, performance wise it's R^2 and adjusted R^2 values are same, which implies all the independent variables plays significant values in model selection, RMSE value is 0.31 which shows that the model can relatively predict the price accurately. The intercept value tends to 0, that means when all the independt variables are 0 the dependent variable also tends to 0.
- Hence, we are selecting Meld-10 as our final model for price prediction in zirconia csv.
- Important features are= Carat, Clarity, Cut, Color, Table, Depth

Final equation for Model 10

Price = 1.047* Carat + 0.031* Cut + -0.217 * Clarity + -0.138 * colour + -0.019* table + -0.015* depth + (-3.209e-17)

Problem 1.4

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

Insights from EDA:

- 'carat': From the scatter plot with price, it is observed that overall price seems to increase with 'carat'. This is also confirmed by the high positive correlation value, which is approximately 0.92. This means that the price of the cubic zirconia diamond seems to increase with carat weight of the cubic zirconia.
- 'x': From the scatter plot with price, it is observed that price has non linearly increasing relationship with 'x'. This also supported by the high positive correlation value of approximately 0.88. This means that the price of the cubic zirconia diamond seems to increase with length of the cubic zirconia diamond.
- 'y': From the scatter plot with price, it is observed that price has non linearly increasing relationship with 'y'. This also supported by the high positive correlation value of approximately 0.86. This means that the price of the cubic zirconia diamond seems to increase with width of the cubic zirconia diamond.
- 'z': From the scatter plot with price, it is observed that price has non linearly increasing relationship with 'z'. This also supported by the high positive correlation value of approximately 0.85. This means that the price of the cubic zirconia diamond seems to increase with height of the cubic zirconia diamond.
- 'depth' & 'table': From the scatter plots with price, it is observed that neither 'depth' nor 'table' seem to have any particular increasing or decreasing trend with 'price'. This also supported by low correlation values of -0.003 and 0.13 respectively with 'price'. This means that for the given range of 'depth' and 'table' there are both low and high prices present.
- 'cut', 'color' & 'clarity': It is observed that in general the mean prices of higher quality cut, color or clarity are having fewer mean prices, this contradiction arises because the mean values of 'carat', 'x', 'y' and 'z' are higher for the diamonds with lower quality cut, color or clarity. The effect of 'carat', 'x', 'y' and 'z' more than compensates the effect of quality of diamond based on cut, color or clarity.

Recommendations:

- It is observed that carat and Clarity are important factors in pricing of the diamonds irrespective of cut, color, table, dimension. If carat weight is more, then the prices are higher on average. Thus, the diamond manufactures may focus on increasing the carat weight and clarity such that the diamonds could be priced higher.
- Since the quality of table and depth are not affecting the price as much as carat weight and clarity, i.e., prices more than compensates the effect of quality of diamond based on table and depth. The manufacturers may give less attention in acquiring higher quality of color, cut which are cost and time intensive.
- There is manufacturing cost associated with making depth and table within the ideal range such that light reflection of the diamond is perfect. But since it is observed that certain diamonds not within the ideal range of table and depth are still priced higher owing to the

dimensions and carat weight of the diamond, thus the manufacturer may not pay extra attention to the precision of these parameters and thereby leading to reduction in costs.

Problem 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.

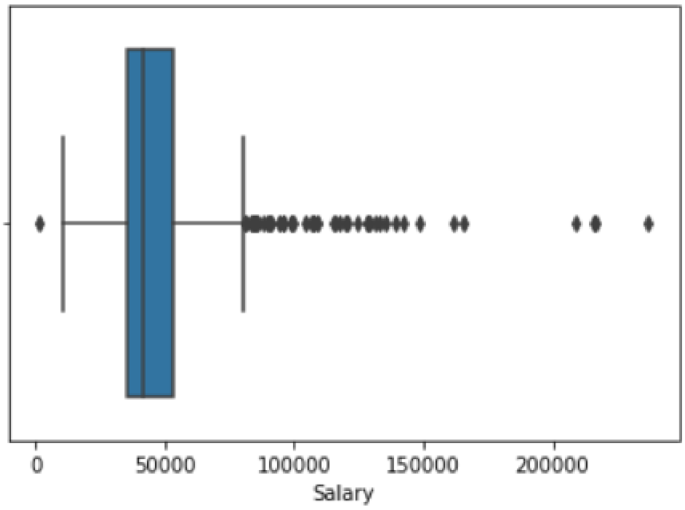
- The given dataset has 872 rows and 7 columns. There are 5 attributes are of numeric data type and 2 attributes are of object data type.
- The dataset has no missing values. There are no duplicate.
- No bad data has been present
- Outliers are present in all the numeric features which can be seen from the boxplot.
- logistic Regression models are not much impacted due to the presence of outliers because the sigmoid function tapers the outliers. Hence, we have not treated the outliers

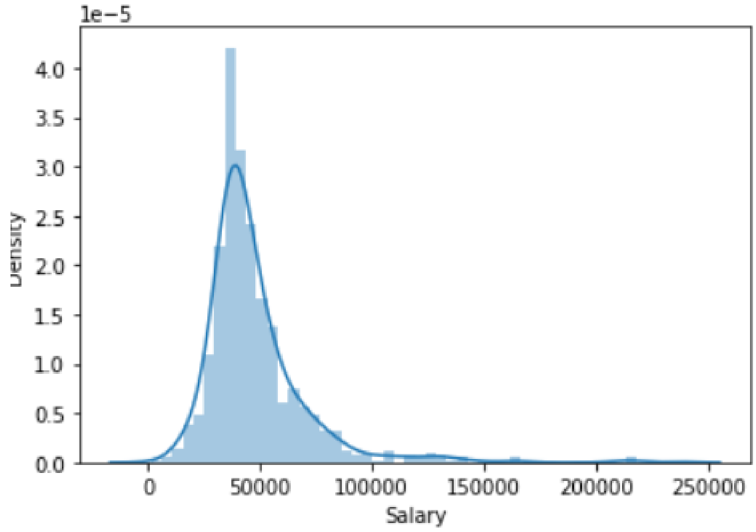
Univariate analysis of numeric Variables

Salary: Employee Salary

- Employee Salary ranges from 1322 to 236961
- Average employee salary is 41903
- The mean is greater than median, the distribution is not normal which is evident from the boxplot and probability plot.
- Skewness of the salary attribution is 3.1 indicating a right tailed distribution, positively skewed
- Outliers are present for this attribution which is evident from the box plot

```
Description ofSalary
..................
count        872.000000
mean       47729.172018
std        23418.668531
min         1322.000000
25%        35324.000000
50%        41903.500000
75%        53469.500000
max       236961.000000
Name: Salary, dtype: float64
```

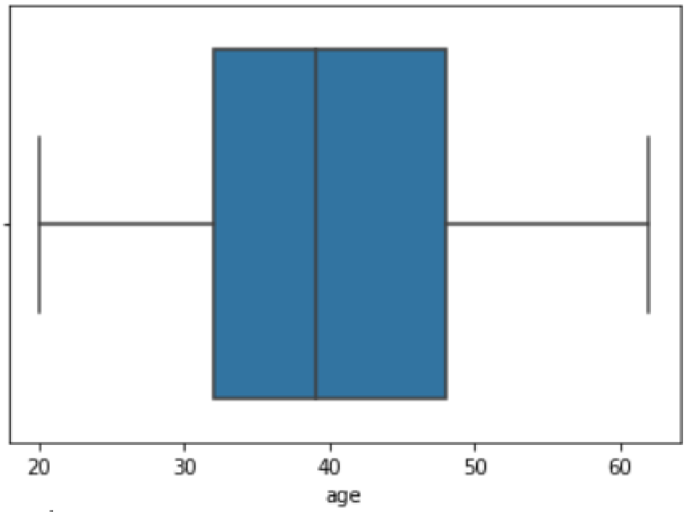


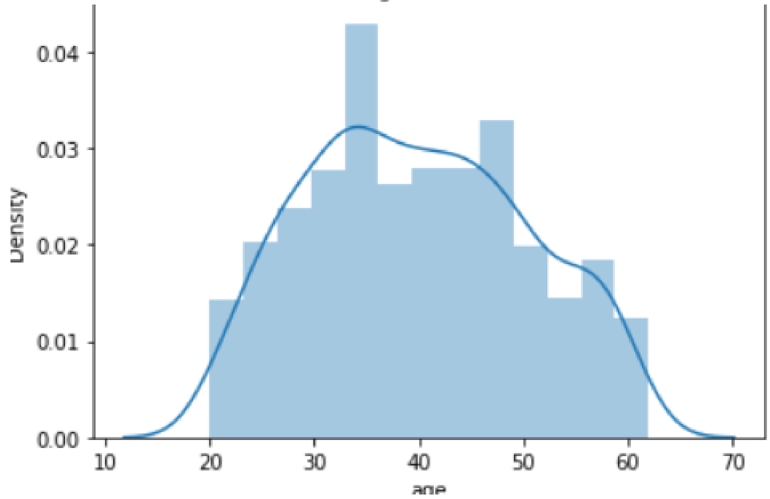


Age: Age in years

- Age ranges from 20 to 62
- Average age is 40
- The mean is greater than median, the distribution is not normal which is evident from the boxplot
- Skewness of the age attribution is 0.4 indicating a right tailed distribution, positively skewed
- Outliers are present for this attribution which is evident from the box plot

```
Description ofage
..................
count    872.000000
mean      39.955275
std       10.551675
min       20.000000
25%       32.000000
50%       39.000000
75%       48.000000
max       62.000000
Name: age, dtype: float64
```

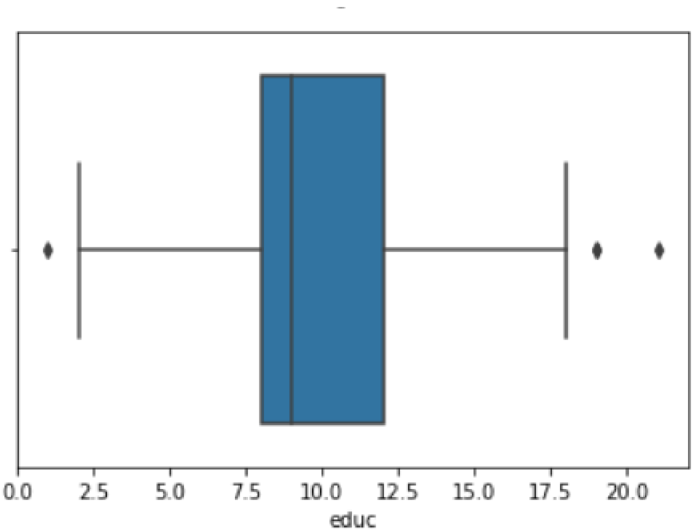


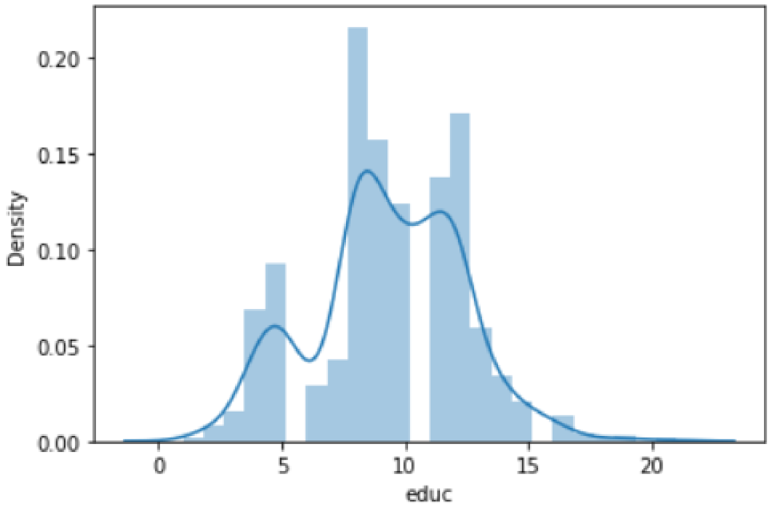


Educ: Years of formal education

- Years of formal education ranges from 1 to 21
- Average Years of formal education is 9
- The mean is nearly equal to median, the distribution is almost normal which is evident from the boxplot
- Skewness of the years of formal education attribution is indicating a left tailed distribution, positively skewed
- Outliers are present for this attribution which is evident from the box plot

```
Description ofeduc
..................
count    872.000000
mean       9.307339
std        3.036259
min        1.000000
25%        8.000000
50%        9.000000
75%       12.000000
max       21.000000
Name: educ, dtype: float64
```

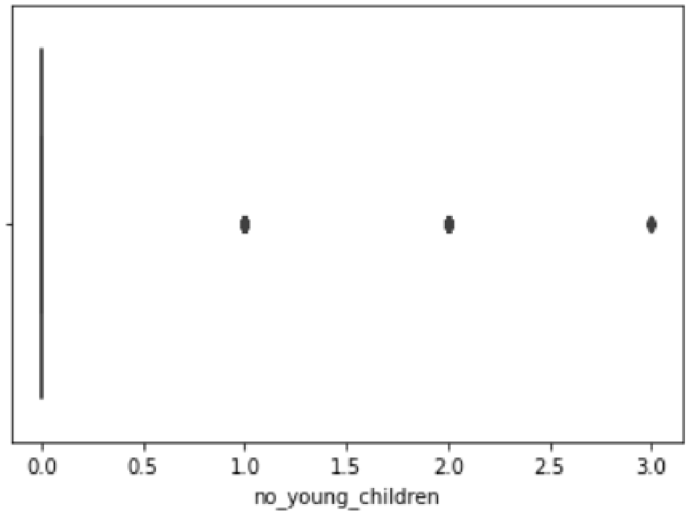


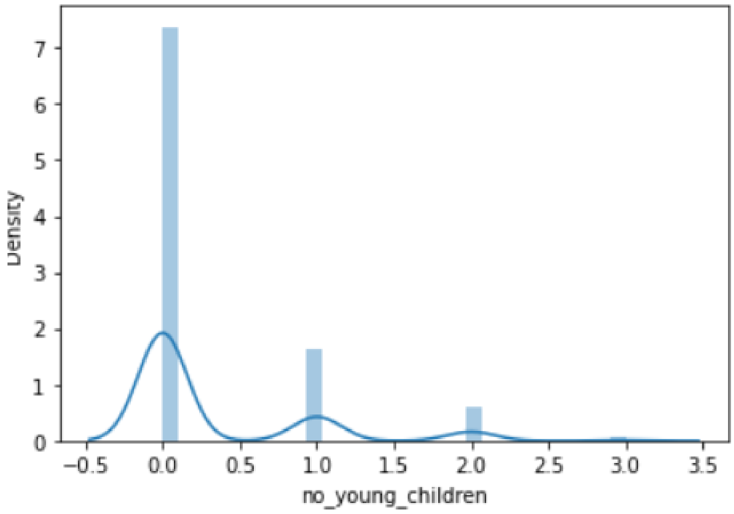


No_young_children: The number of young children (younger than 7 years)

- The number of young children (younger than 7 years ranges from 0.00 to 3
- Average of number of young children (younger than 7 years is 0
- The mean is greater than median, the distribution is not normal
- It looks like clusters formed for each unique value

```
Description ofno_young_children
..................
count     872.000000
mean        0.311927
std         0.612870
min         0.000000
25%         0.000000
50%         0.000000
75%         0.000000
max         3.000000
Name: no_young_children, dtype: float64
```

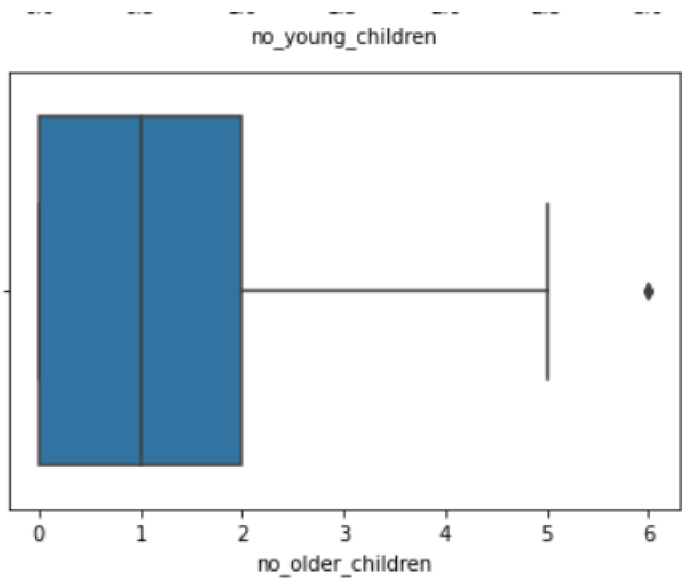


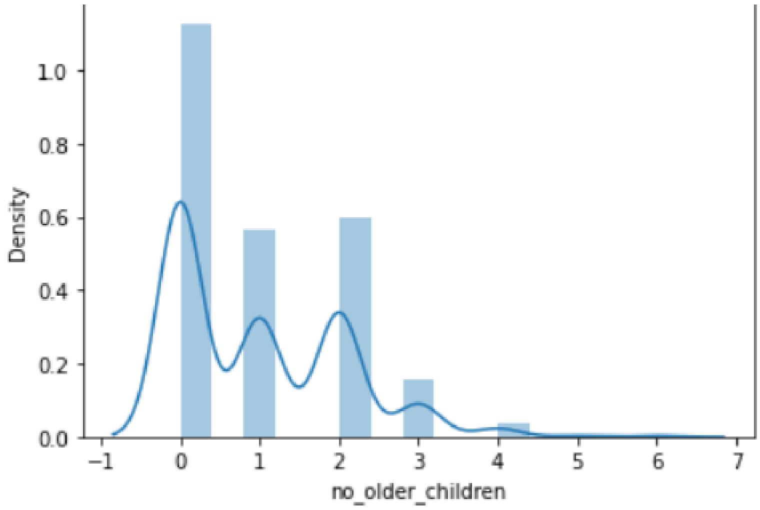


No_older_children: number of older children

- The number of older children ranges from 0.00 to 3
- Average of number of older children is 1
- The mean is almost equal to median, the distribution is not normal

```
Description ofno_older_children
..................
count    872.000000
mean       0.982798
std        1.086786
min        0.000000
25%        0.000000
50%        1.000000
75%        2.000000
max        6.000000
Name: no_older_children, dtype: float64
```





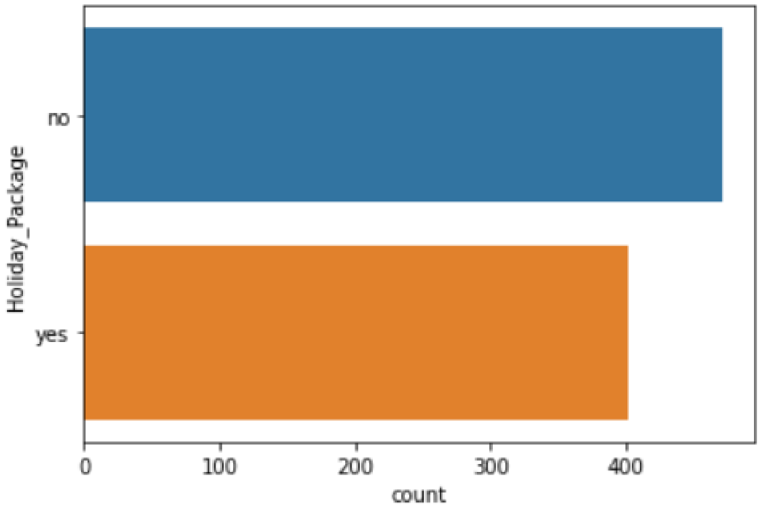
Further, after observing the box plot of No_older_children and No_younger_children, we have found that they have very few unique values compared to other numeric attribute. After performing the unique value function, we found that. The number of unique values are only 4 and 7 in 'no_young_children' and 'no_older_children'. Hence analysing them with the categorical variables can give good insights.

Univariate analysis for Categorical Attributes

Holiday Package: Opted for Holiday Package yes/no

- There are 2 types, People either opted for yes or no
- It has approximately 54% as people who didn't opt for holiday package and 46% of the people opted for the holiday package. Thus, the dataset is nearly balanced, imbalance effect in classification model is not expected.

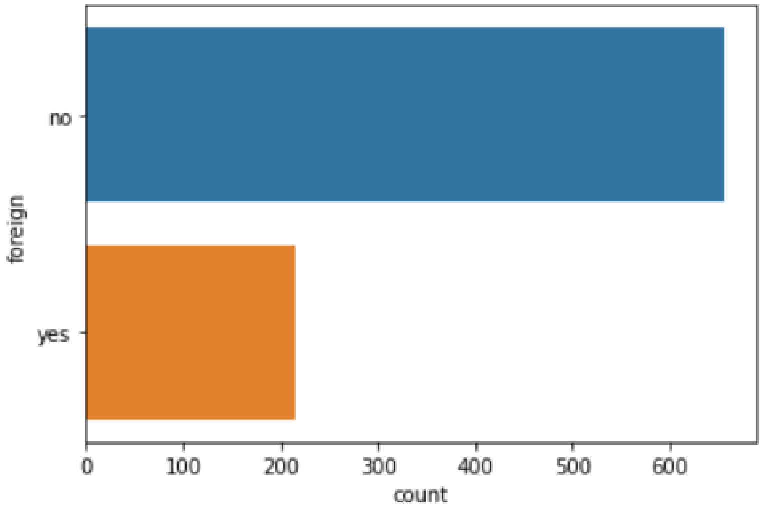
```
Percentage Value counts of Holiday_Package
------------------------
no      0.540138
yes     0.459862
Name: Holiday_Package, dtype: float64

Frequency Distribution of Holiday_Package
------------------------
```



foreign: Foreigners Yes/No

- There are 2 types, either a foreigner employee or a non- foreigner employee
- It has approximately 75% people who are no- foreigners and 25% of the people are foreigners.

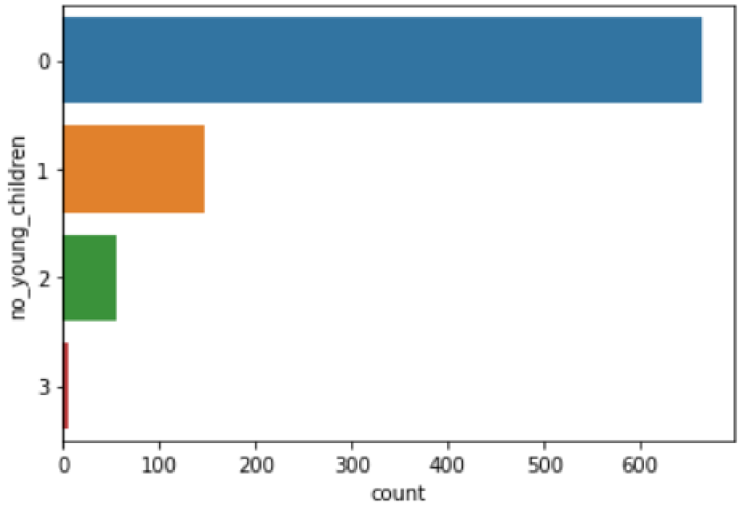
```
Percentage Value counts of foreign
------------------------
no      0.752294
yes     0.247706
Name: foreign, dtype: float64

Frequency Distribution of foreign
------------------------
```



No_young_children: The number of young children (younger than 7 years)

- Number of children under this attribute ranges from 0 to 3
- Most of the employees (76%) have no children below 7 years. The number of employees with 3 number of children below 7 years is very less (0.05%).

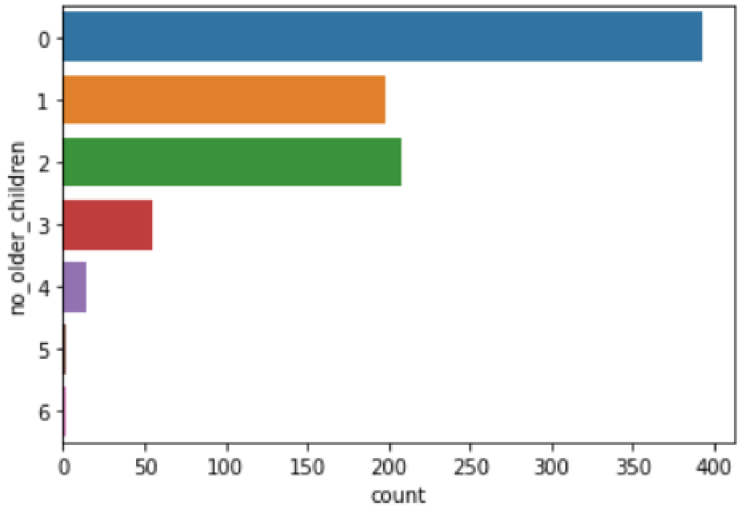
```
Percentage Value counts of no_young_children
------------------------
0     0.762615
1     0.168578
2     0.063073
3     0.005734
Name: no_young_children, dtype: float64

Frequency Distribution of no_young_children
------------------------
```



No_older_children: The number of older children

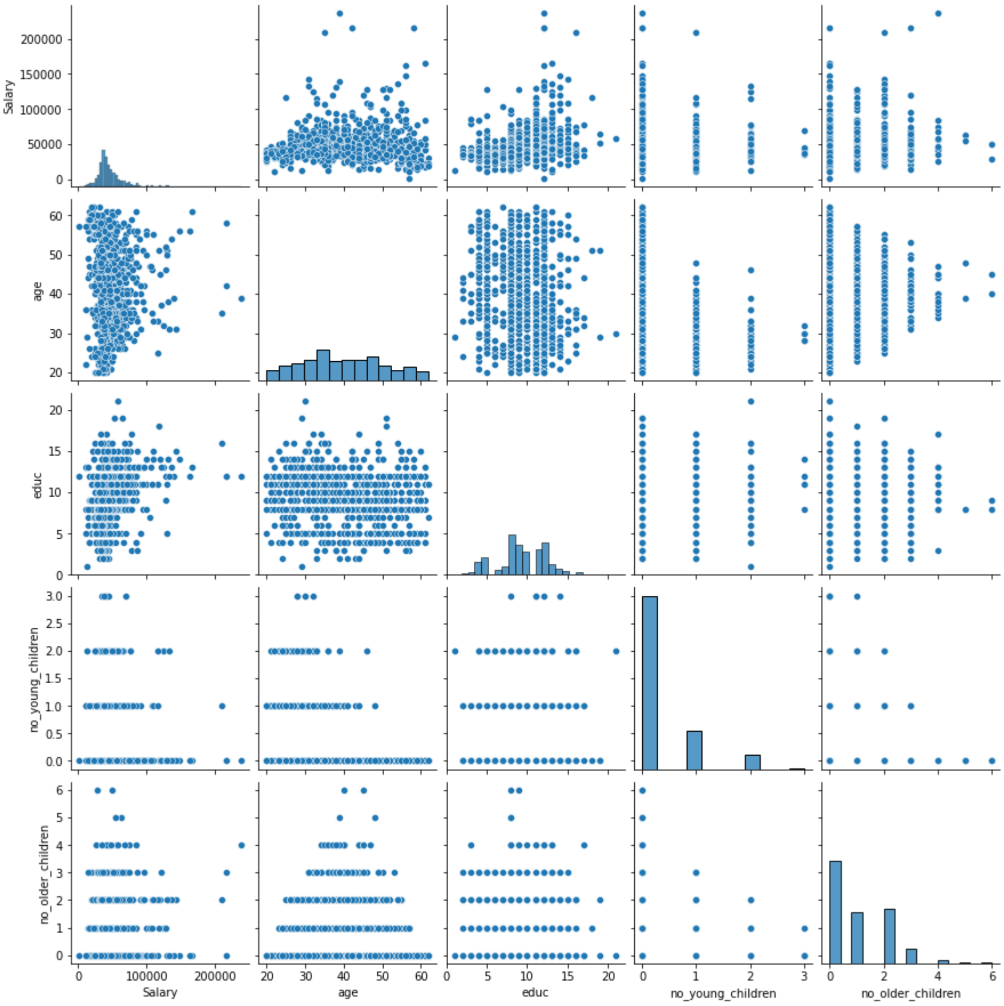
- Number of children under this attribute ranges from 0 to 6
- Most of the employees (45%). The number of employees with 6 number of children below (0.002%).

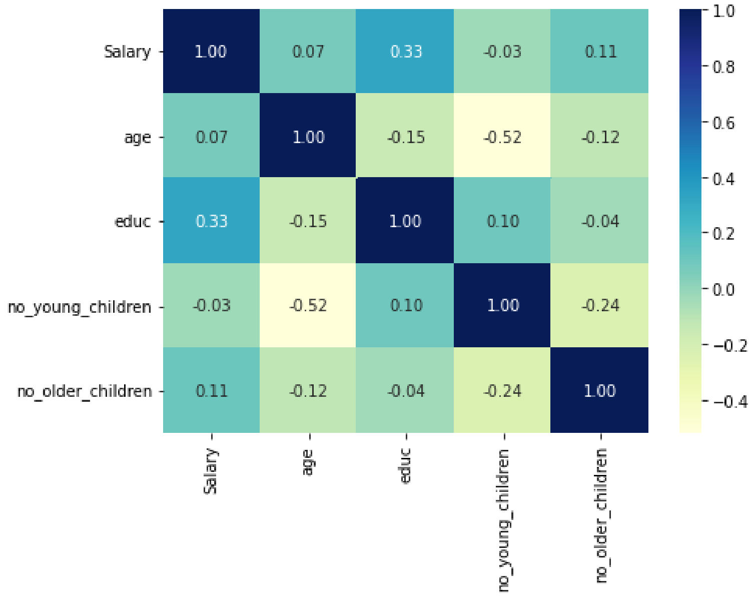
```
Percentage Value counts of no_older_children
------------------------
0    0.450688
2    0.238532
1    0.227064
3    0.063073
4    0.016055
5    0.002294
6    0.002294
Name: no_older_children, dtype: float64

Frequency Distribution of no_older_children
------------------------
```



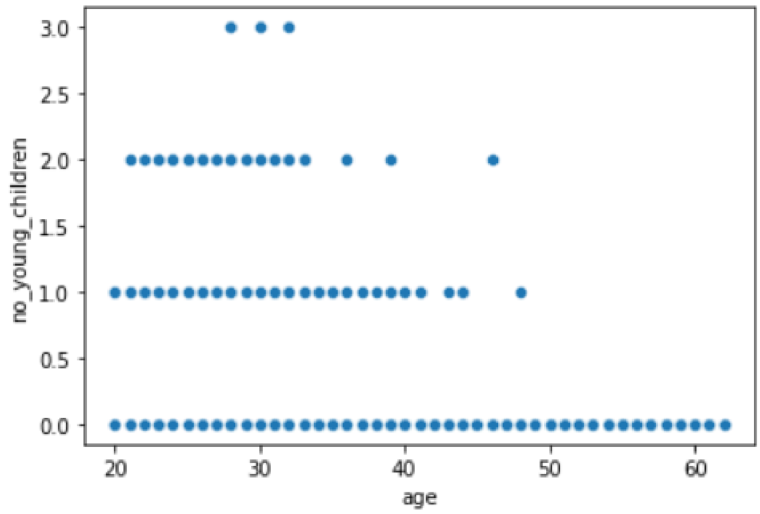
Multivariate-Bivariate analysis

Heat map shows the correlation between different numeric attributes by assigning numbers as well as colors and Pair plot gives a graphical representation of correlation between different numeric attributes


Salary
age
educ
no_young_children
no_older_children


Salary
age
educ
no_young_children
no_older_children
1.00
0.07
0.33
-0.03
0.11
-0.15
-0.52
-0.12
0.10
-0.04
-0.24

- There is no high correlation between the variables.
- The maximum correlation is observed between 'age' and 'no_young_children'. It is a negative correlation of -0.52.
- It can also be observed that 'Salary' does not have high positive correlation with 'educ' or 'age'.

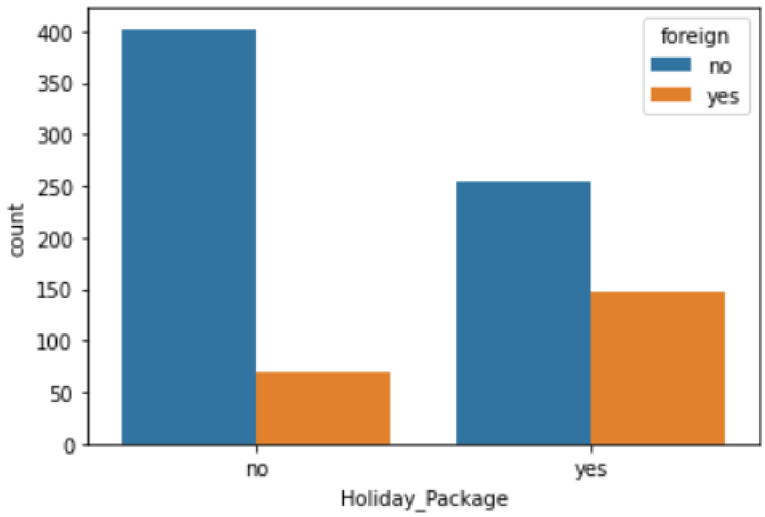


<u>Bivariate analysis for Categorical attributes</u>

Holiday Package and other categorical variables:

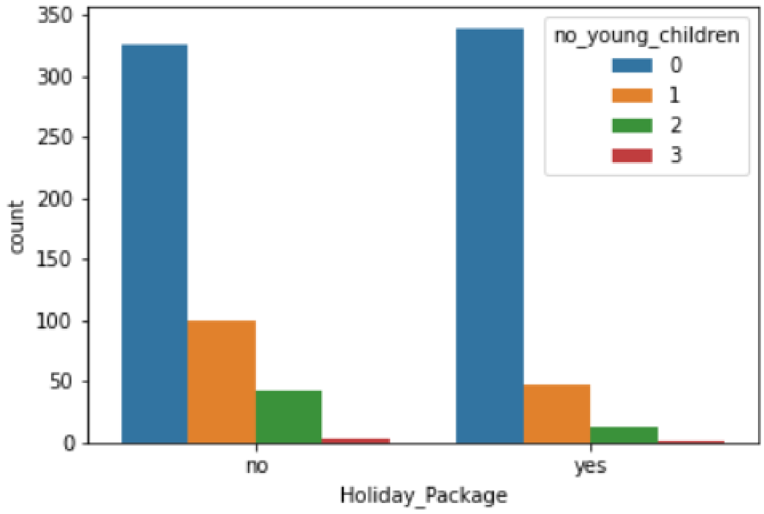
- Non-foreigner employees mostly opted for no Holiday package (402) whereas foreigner employees mostly opted for Holiday package (147)

foreign	no	yes	All
Holiday_Package			
no	402	69	471
yes	254	147	401
All	656	216	872



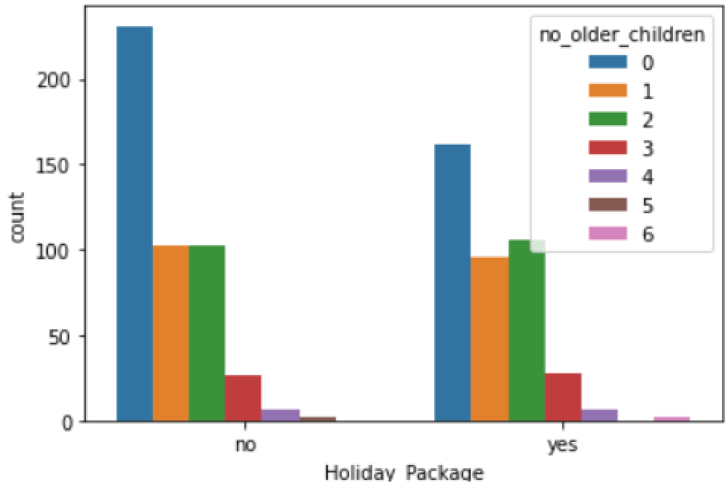
- Employees having no young children opted for holiday package. Remaining Employees having 1 or 2 or 3 young children prefers no holiday package

no_young_children	0	1	2	3	All
Holiday_Package					
no	326	100	42	3	471
yes	339	47	13	2	401
All	665	147	55	5	872



- Employees having no (0) or 1 or 5 old children opted for no holiday package. Employees having 2 or 3 children opted for holiday package. Employees having 4 children have equal instance

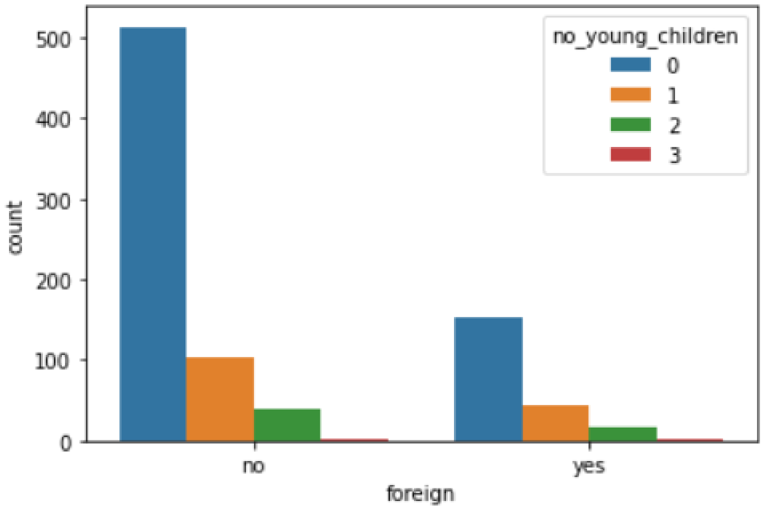
no_older_children	0	1	2	3	4	5	6	All
Holiday_Package								
no	231	102	102	27	7	2	0	471
yes	162	96	106	28	7	0	2	401
All	393	198	208	55	14	2	2	872



<u>Foreign employees and other categorical variables:</u>

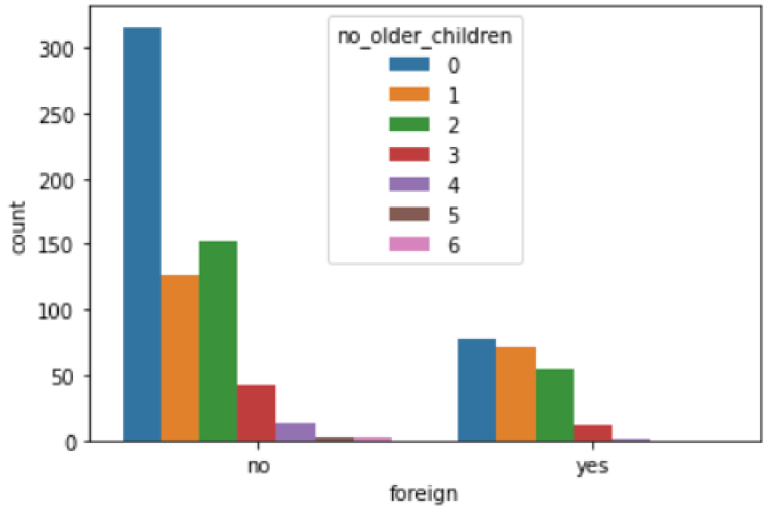
- Maximum foreigner employees have no young children and only 3 foreigner employees have 3 children
- Maximum non- foreigner employees have no young children and only 2 of them have young children

no_young_children	0	1	2	3	All
foreign					
no	513	103	38	2	656
yes	152	44	17	3	216
All	665	147	55	5	872



- Maximum foreigner employees have no old children and none of them have 5 or 6 children
- Maximum non- foreigner employees have no children and 2 out of them have 5 children and 2 out of them have 6 children

no_older_children	0	1	2	3	4	5	6	All
foreign								
no	316	127	153	43	13	2	2	656
yes	77	71	55	12	1	0	0	216
All	393	198	208	55	14	2	2	872

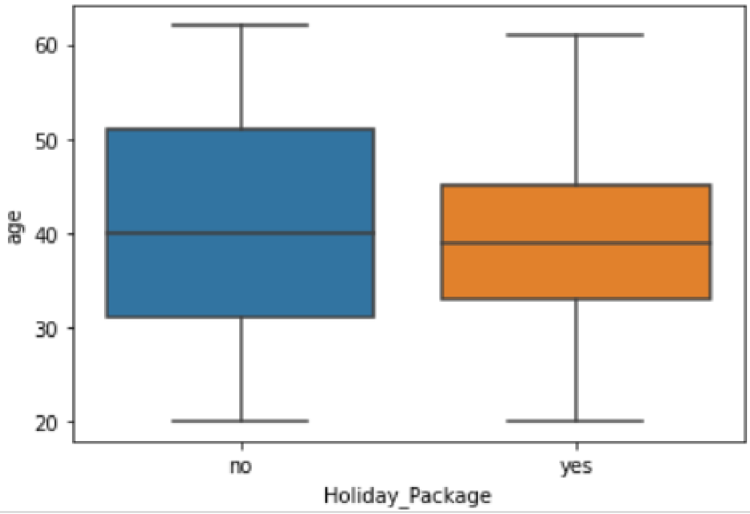


Bivariate Analysis in num-cat attributes

Age with one categorical

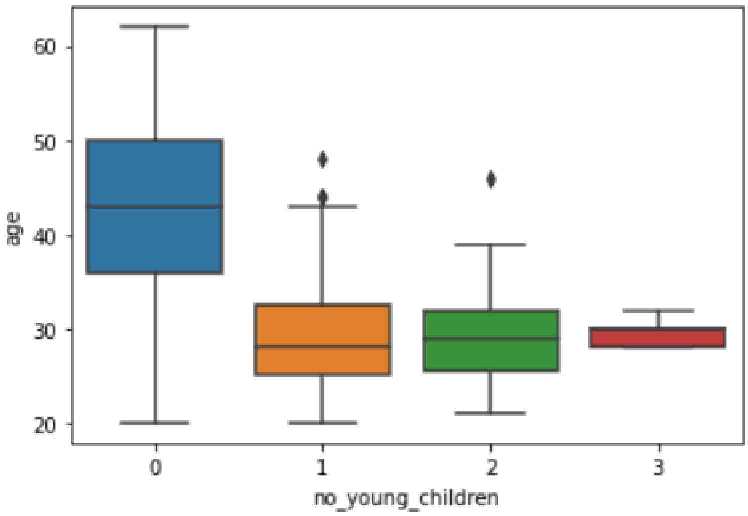
- Mean age of People who opted for holiday package is 40 whereas mean age for people who don’t opt for holiday package is 38.9. Average age for both yes and no for holiday package is nearly same

```
Mean of age for Holiday_Package
Holiday_Package
no      40.853503
yes     38.900249
Name: age, dtype: float64

Plot of age vs Holiday_Package
-----------------------------------------------------
```



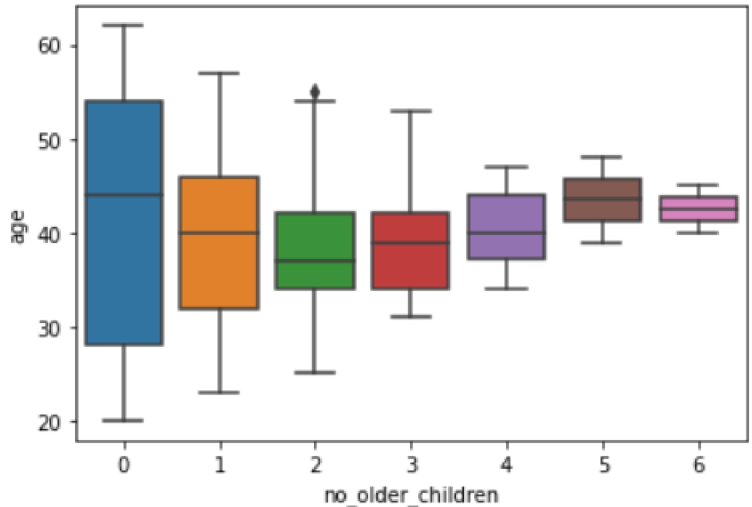
- Employees with no young children have maximum mean age 43, whereas employees with 1 or 2 or 3 young children have nearly equal mean age I.e., 29

```
Mean of age for no_young_children
no_young_children
0     43.296241
1     29.265306
2     29.072727
3     29.600000
Name: age, dtype: float64

Plot of age vs no_young_children
-----------------------------------------------------
```



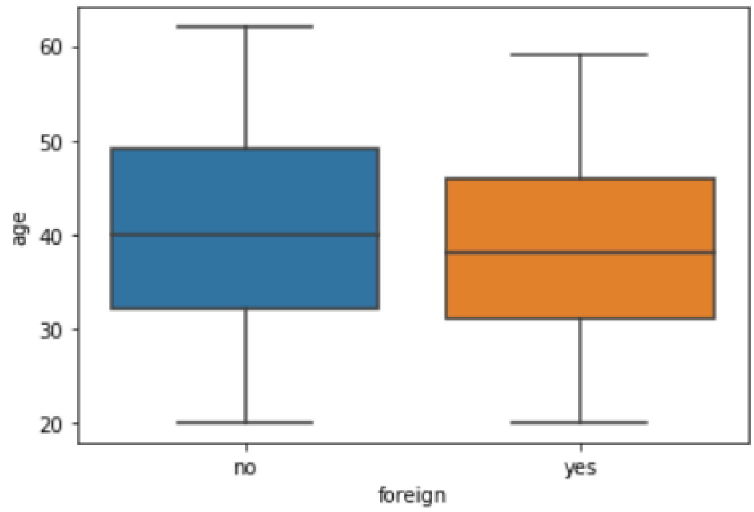
- Employees with 5 old children have maximum mean age 43, whereas employees with 2 old children have minimum mean age of 37

```
Mean of age for no_older_children
no_older_children
0    41.615776
1    39.161616
2    37.798077
3    38.800000
4    40.285714
5    43.500000
6    42.500000
Name: age, dtype: float64

Plot of age vs no_older_children
--------------------------------------------------
```



- Employees who are foreigners have maximum mean age 40; whereas, non-foreigners have minimum mean age

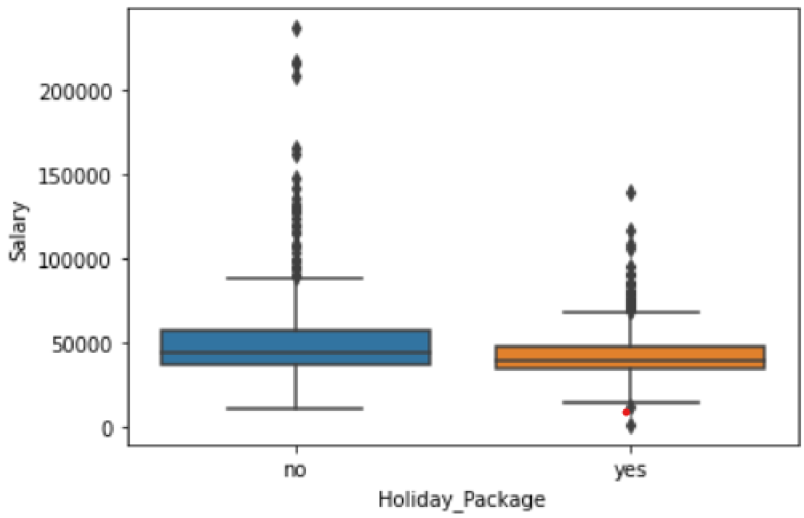
```
Mean of age for foreign
foreign
no      40.603659
yes     37.986111
Name: age, dtype: float64

Plot of age vs foreign
--------------------------------------------------
```



Salary with one categorical variable

- Mean Salary of People who don't opted for holiday package is maximum whereas mean age for people who opt for holiday package is minimum.

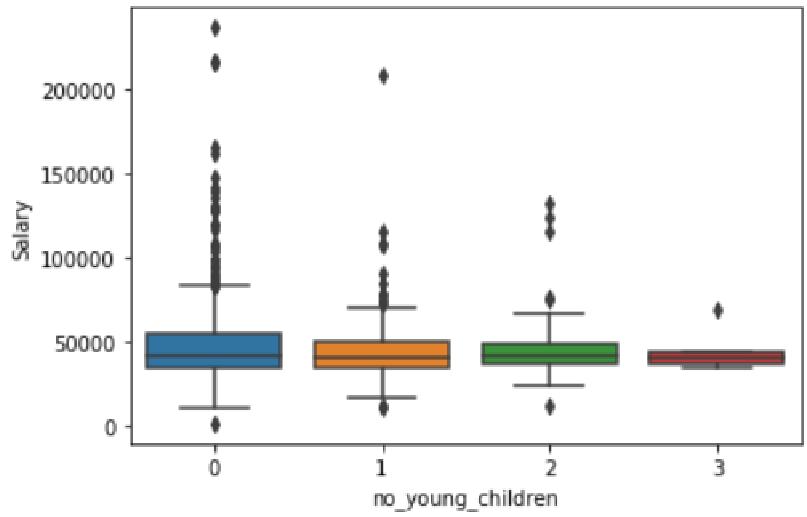
```
Plot of Salry vs Holiday_Package
----------------------------------
```




```
Mean of Salary for no_young_children
no_young_children
0     48210.348872
1     45810.176871
2     47275.854545
3     45137.600000
```

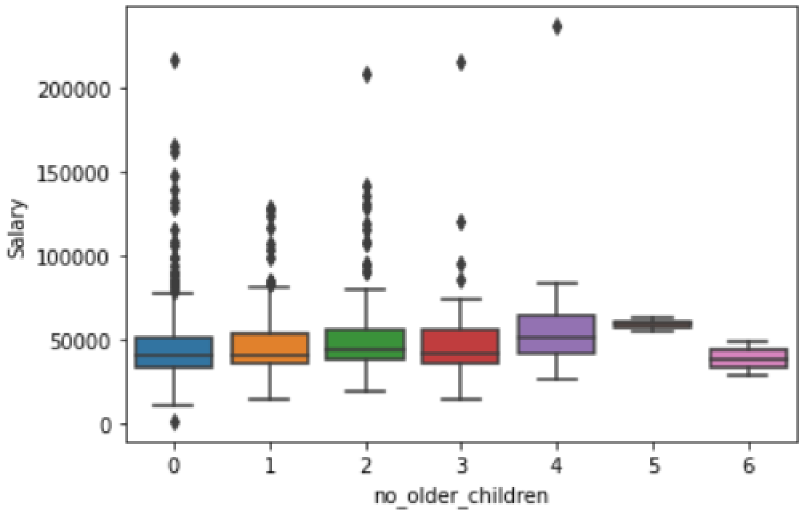
- Employees with no young children have maximum mean salary, whereas employees with 3 young children have minimum mean salary

```
Mean of Salary for no_young_children
no_young_children
0    48210.348872
1    45810.176871
2    47275.854545
3    45137.600000
Name: Salary, dtype: float64

Plot of Salry vs no_young_children
----------------------------------
```



- Employees with 4 old children have maximum mean salary, whereas employees with 6 old children have minimum mean salary

```
Plot of Salry vs no_older_children
----------------------------------
```

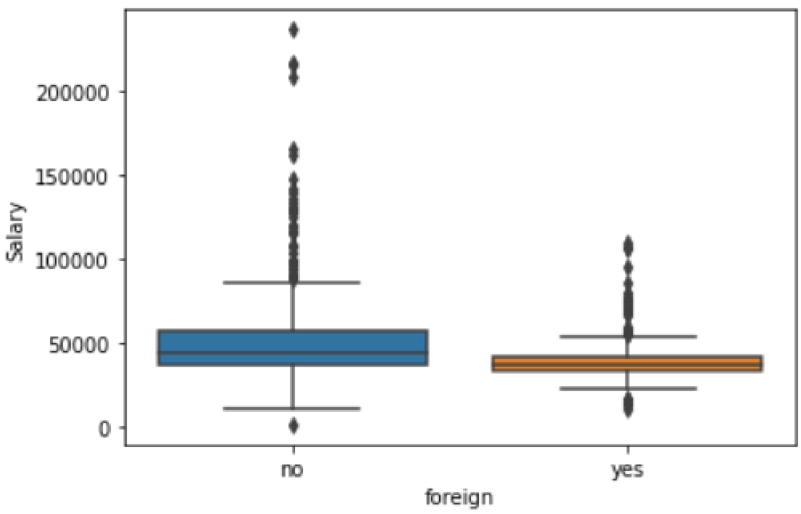



```
Mean of Salary for foreign
foreign
no      50429.248476
yes     39528.939815
Name: Salary, dtype: float64
```

- Employees who are foreigners have maximum mean salary; whereas, non-foreigners have minimum mean salary


```
Mean of Salary for foreign
foreign
no      50429.248476
yes     39528.939815
Name: Salary, dtype: float64
```

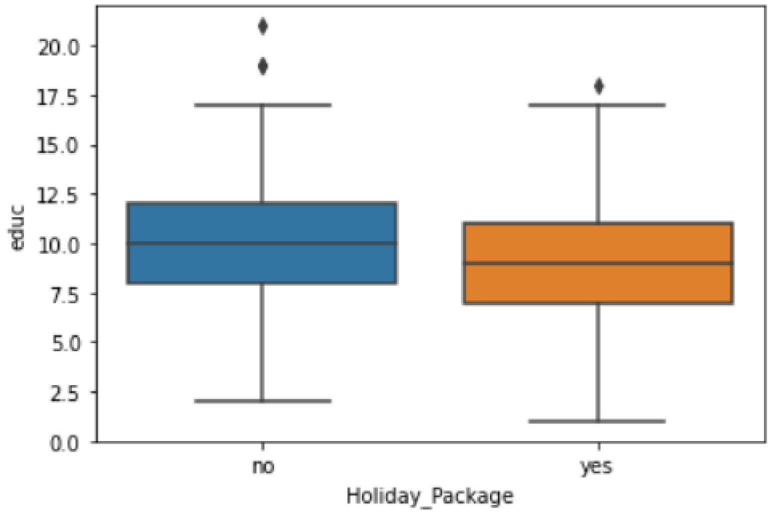

```
Plot of Salry vs foreign
----------------------------------
```



Education with one categorical variable

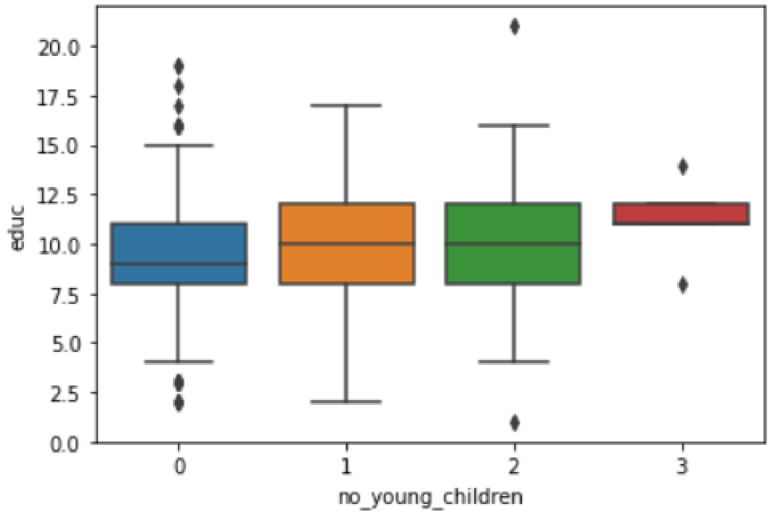
- People who don't opted for holiday package have maximum mean number of years in education whereas people who opt for holiday package have minimum mean number of years in education. Average number of years in education for both yes and no for holiday package is nearly same.

```
Mean of educ for Holiday_Package
Holiday_Package
no      9.594480
yes     8.970075
Name: educ, dtype: float64

Plot of educ vs Holiday_Package
----------------------------------
```



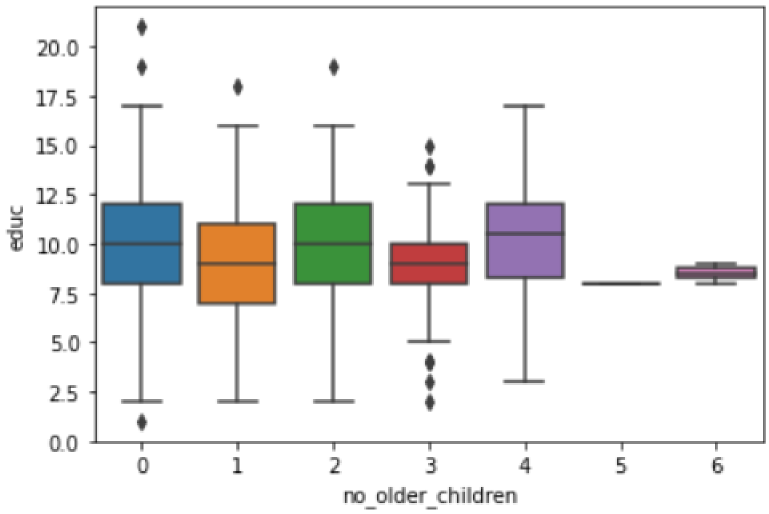
- Employees with 3 young children have maximum mean number of years in education whereas, employees with no or 1 child have almost same mean number of years in education

```
Mean of educ for no_young_children
no_young_children
0     9.144361
1     9.761905
2     9.890909
3    11.200000
Name: educ, dtype: float64

Plot of educ vs no_young_children
----------------------------------
```



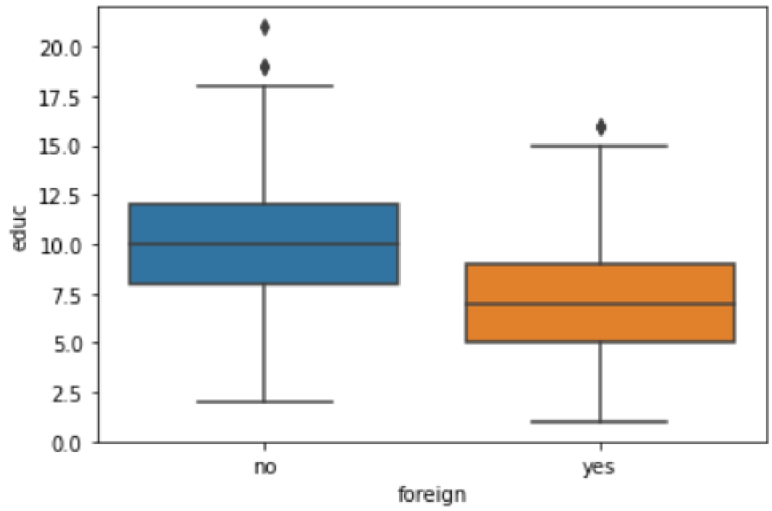
- Employees with 4 old children have maximum mean number of years in education whereas, employees with no or 1 or 2 or 3 or 5 or 6 children have almost same mean number of years in education

```
Mean of educ for no_older_children
no_older_children
0     9.544529
1     8.792929
2     9.461538
3     8.709091
4    10.285714
5     8.000000
6     8.500000
Name: educ, dtype: float64

Plot of educ vs no_older_children
----------------------------------
```



- Foreigner Employees have maximum mean number of years in education whereas, non-foreigner employee has minimum mean number of years

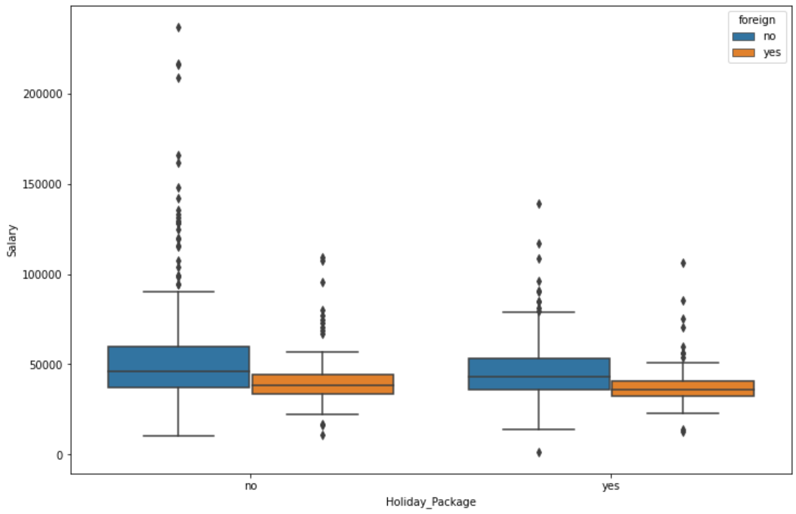
```
Mean of educ for foreign
foreign
no      10.038110
yes      7.087963
Name: educ, dtype: float64

Plot of educ vs foreign
----------------------------------
```



<u>Multivariate analysis cat-num-cat</u>

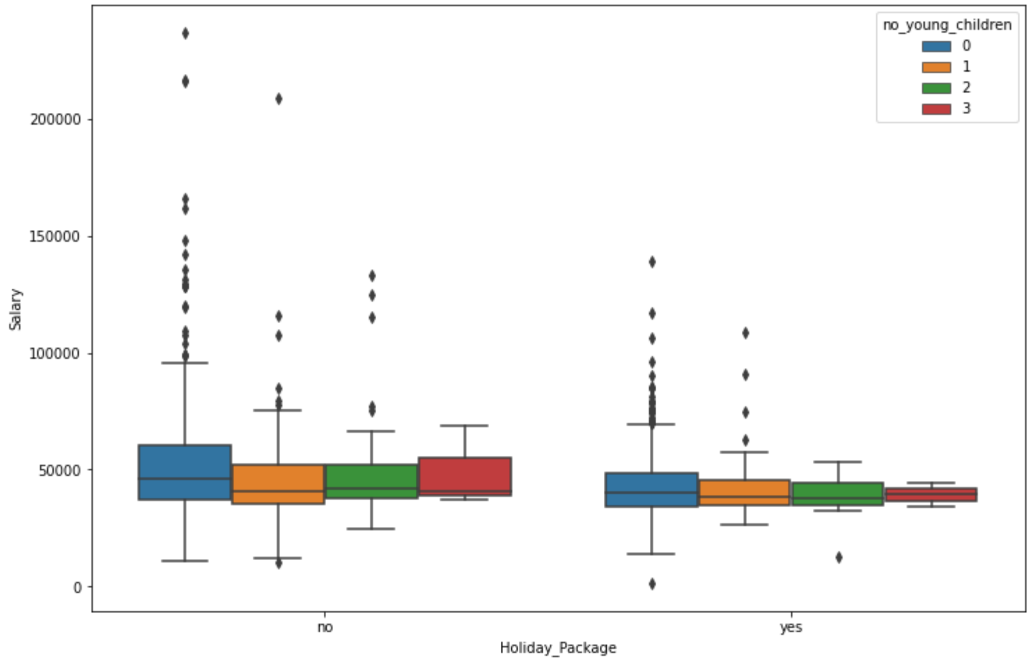
Salary with two categorical attributes

- Employees who are not foreigners and don't opt for holiday package have maximum mean salary compare to foreigners who also don't opt for holiday package. Employees who are not foreigners and opt for holiday package have maximum mean salary compare to foreigners who opt for holiday package.

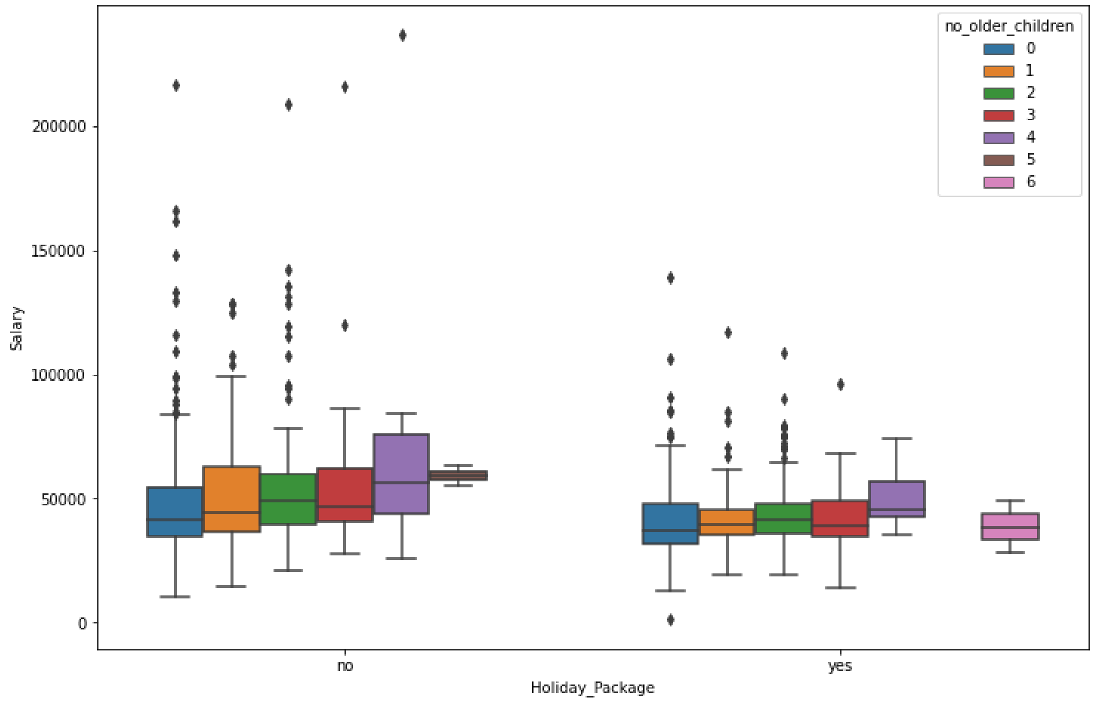


- Employees who don't opt for holiday package and have no young children have maximum mean of salary and with 3 no of young children have minimum mean salary.

Employees who opt for holiday package and have no young children have maximum mean of salary and other employees with 1 or 2 or 3 young children have minimum mean salary.

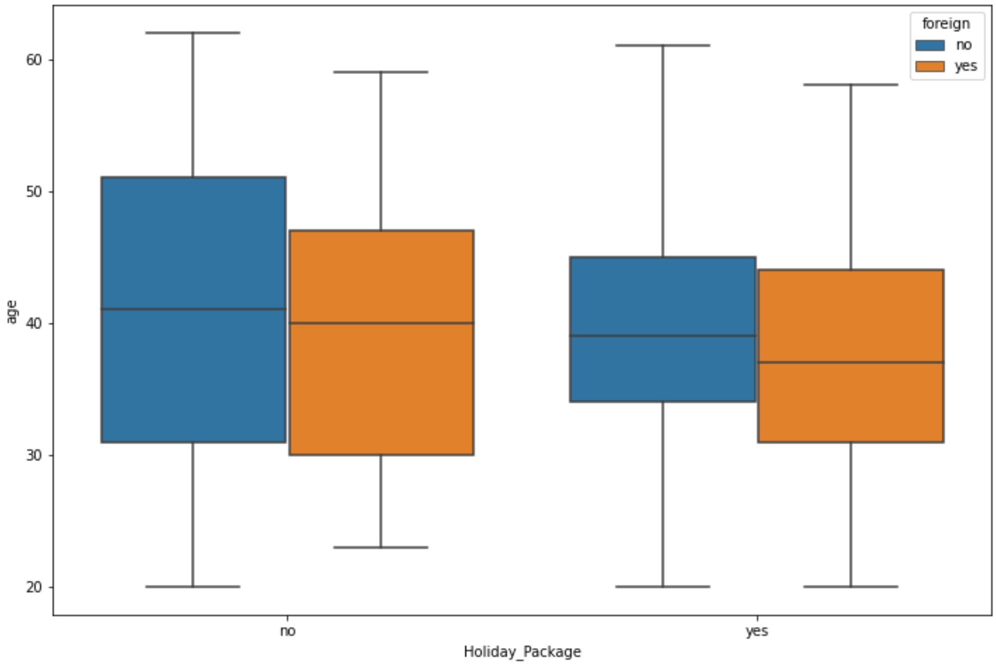


- Employees who don't opt for holiday package and have 5 old children have maximum mean of salary and with no old children have minimum mean salary. Employees who opt for holiday package and have 4 old children have maximum mean of salary and other employees with 0or 1 or 2 or 3 or 6 old children have minimum mean salary. Employees with 5 old children don't opt for holiday package at all.

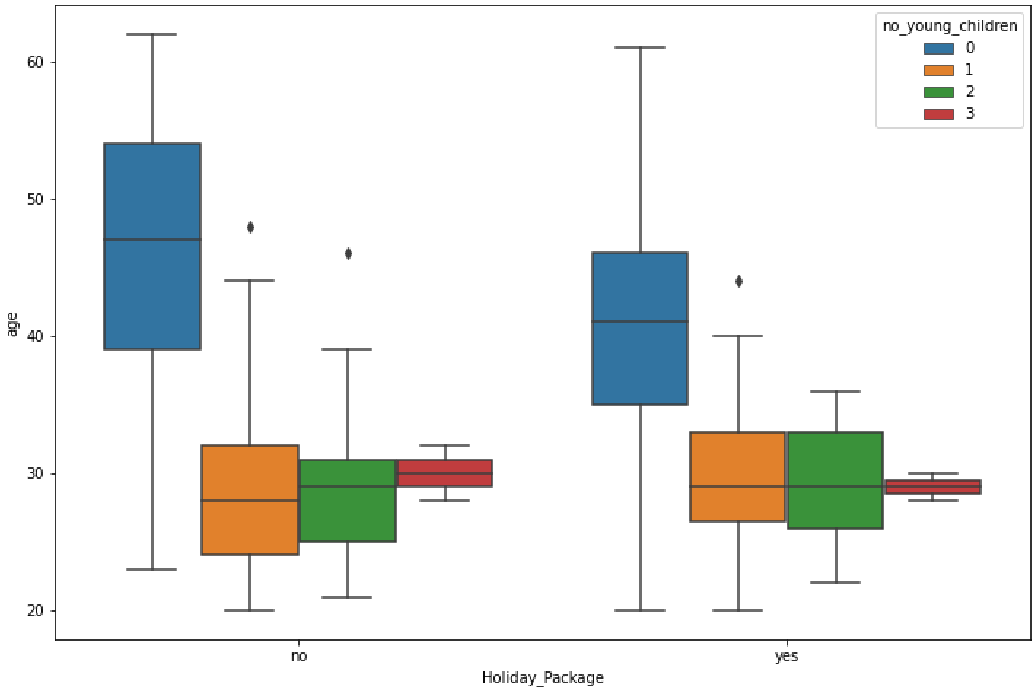


Age with two categorical attributes

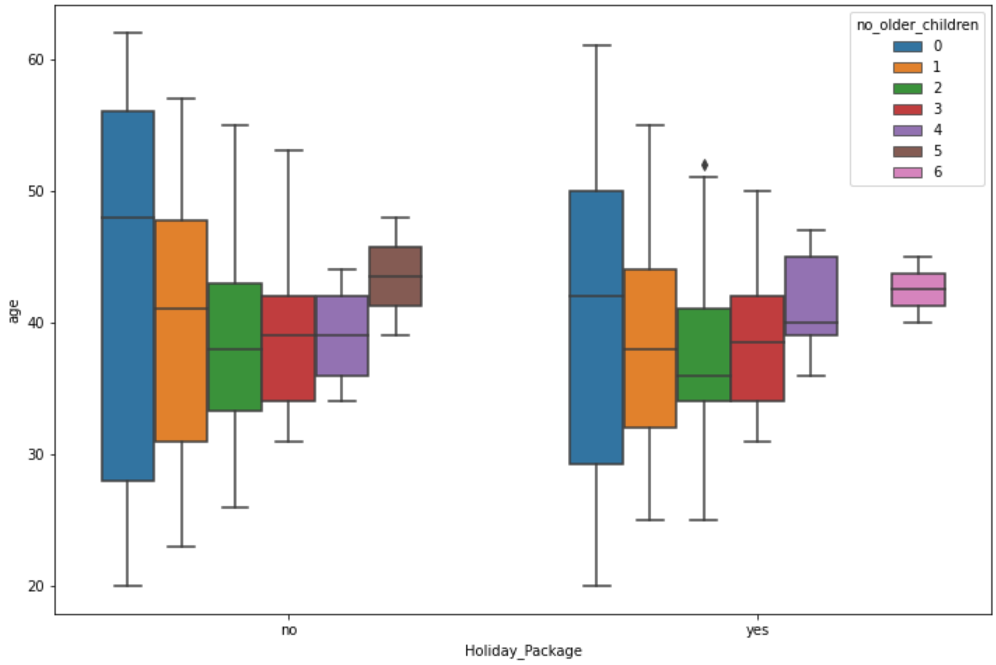
- Employees who are not foreigners and don't opt for holiday package have maximum mean age compare to foreigners who also don't opt for holiday package. Employees who are not foreigners and opt for holiday package have maximum mean salary compare to non- foreigners who opt for holiday package.



- Employees who don't opt for holiday package and have no young children have maximum mean of age and with 1 no of young children have minimum mean age. Employees who opt for holiday package and have no young children have maximum mean of age and other employees with 1 or 2 or 3 young children have minimum mean age.

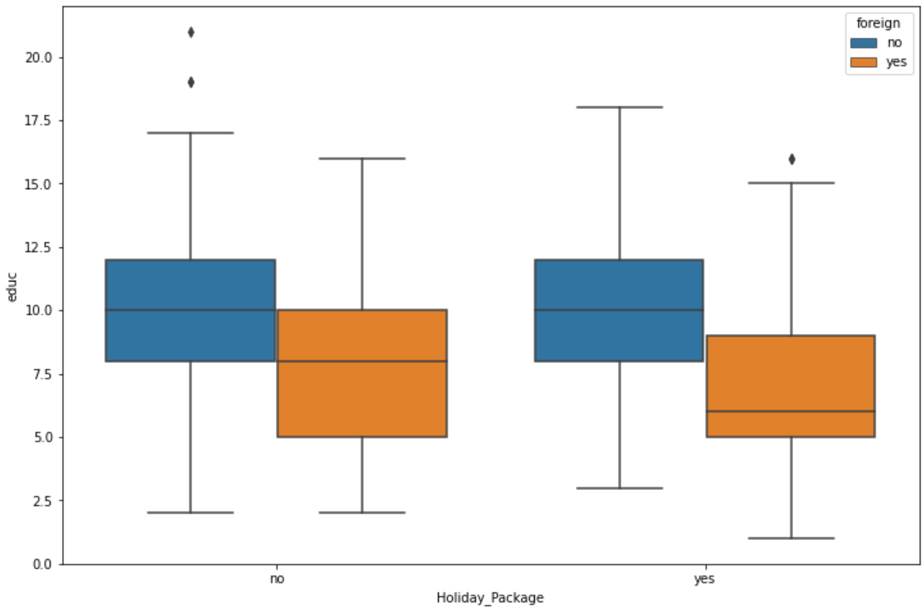


- Employees who don't opt for holiday package and have no old children have maximum mean of age and with 2 old children have minimum mean age. Employees who opt for holiday package and have no old children have maximum mean age and with 2 old children have age.

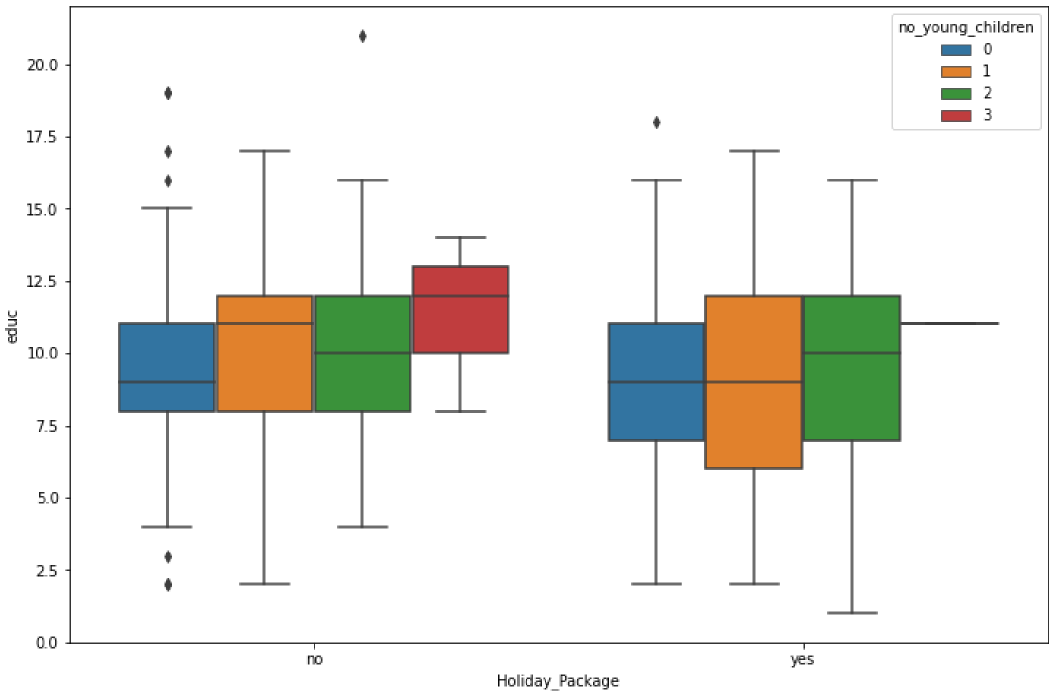


Educ with two categorical attributes

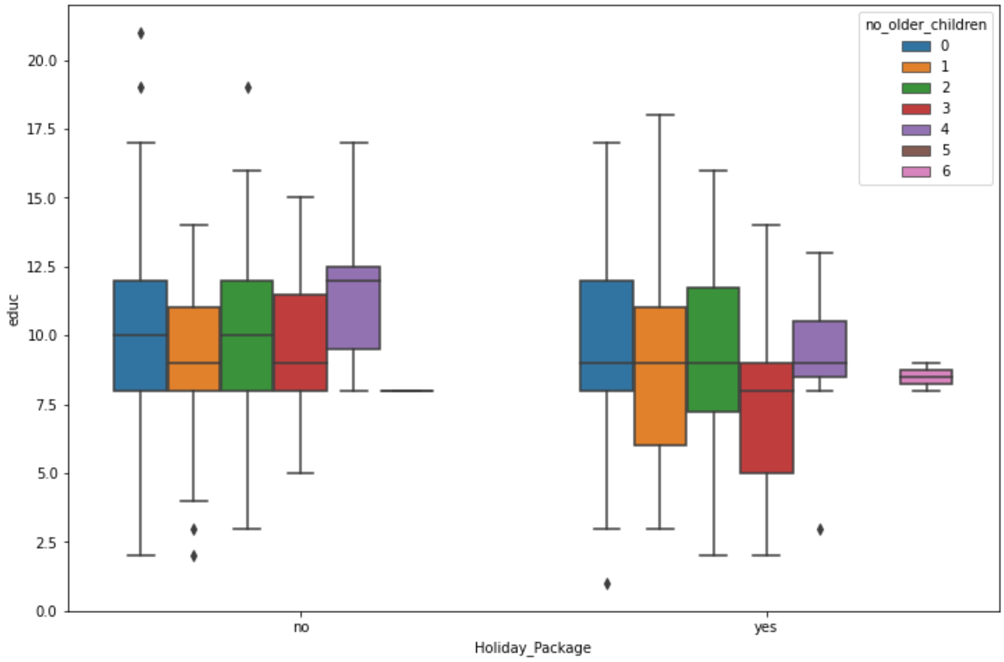
- Employees who are not foreigners and don't opt for holiday package have maximum mean number of years of education compare to foreigners who also don't opt for holiday package. Employees who are not foreigners and opt for holiday package have maximum mean number of years of education compare to non- foreigners who opt for holiday package.



- Employees who don’t opt for holiday package and have 3 young children have maximum mean of number of years of education and with no young children have minimum mean number of years of education. Employees who opt for holiday package have 3 young children have maximum mean of number of years of education and other employees with no or 1 or 2 young children have minimum mean number of years of education.



- Employees who don’t opt for holiday and have 4 old children have maximum mean of number of years of education and with 1 or 3 old children have minimum mean number of years of education. Employees who opt for holiday and have 3 old children have minimum mean number of years of education and



Problem 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).

Logistic Regression

- For Logistic regression model building all the data should be in the form of numerical data type. In the given dataset there are 2 attributes are of object data type presents in the data; therefore, they are converted into categorical type with codes.
- From Sklearn packages train_test_split was imported and splitting of data is being done in 70:30 ratio (70% for train and 30% for test). Random state has been set to 1
- The shape of the data is as follows

```
X_train (610, 6)
X_test (262, 6)
y_train (610, 1)
y_test (262, 1)
Total observations is 872
```

- Logistic Regression imported from Sklearn model package.
- For Model Building we have combined 4 different combinations of Penalty and Solver

Combination 1:

penalty: 'l2'
, solver: 'newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'

- ➢ For better model we have performed grid search with different set of values
- ➢ Cross validation used as 10

- Classification report for the particular model also generated. Accuracy obtained for train sample is 0.68 and for test sample is 0.67.
- AUC score for the train set is 0.742 and test set is 0.705

Combination 2:

Penalty: 'l1'
Solver: 'liblinear', 'saga'

- For better model we have performed grid search with different set of values
- Cross validation used as 10
- Classification report for the particular model also generated. Accuracy obtained for train sample is 0.53 and for test sample is 0.55.
- AUC score for the train set is 0.572 and test set is 0.629

Combination 3:

Penalty: 'elasticnet'
Solver: 'saga'

- For better model we have performed grid search with different set of values
- Cross validation used as 10
- Classification report for the particular model also generated. Accuracy obtained for train sample is 0.54 and for test sample is 0.55.
- AUC score for the train set is 0.590 and test set is 0.634

Combination 4:

Penalty: 'elasticnet'
Solver: 'saga', 'newton-cg', 'lbfgs', 'sag'

- For better model we have performed grid search with different set of values
- Cross validation used as 10
- Classification report for the particular model also generated. Accuracy obtained for train sample is 0.54 and for test sample is 0.55.
- AUC score for the train set is 0.590 and test set is 0.634

Linear Discriminant Analysis

- For LDA model building all the data should be in the form of numerical data type. In the given dataset there are 2 attributes are of object data type presents in the data; therefore, they are converted into categorical type with codes.
- From Sklearn packages train_test_split was imported and splitting of data is being done in 70:30 ratio (70% for train and 30% for test). Random state has been set to 1
- The shape of the data is as follows

```
X_train (610, 7)
X_test (262, 7)
y_train (610, 1)
y_test (262, 1)
Total observations is 872
```

- LDA imported from Sklearn discriminant analysis.
- For Model Building we have made 2 different model with respect to solver

Combination 1:

‘tol’: 0.1
Solver: 'liblinear', 'saga’

- For better model we have performed grid search with different set of values
- Cross validation used as 10
- Classification report for the particular model also generated. Accuracy obtained for train sample is 0.67 and for test sample is 0.63.
- AUC score for the train set is 0.75 and test set is 0.70

Combination 2:

‘tol’: 0.1
Solver: 'lsqr’

- For better model we have performed grid search with different set of values
- Cross validation used as 10
- Classification report for the particular model also generated. Accuracy obtained for train sample is 0.68 and for test sample is 0.69.

Problem 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.

Logistic Regression

Combination 1

```
**Classification Report Training Data:**
              precision    recall  f1-score   support

           0       0.70      0.70      0.70       326
           1       0.65      0.65      0.65       284

    accuracy                           0.68       610
   macro avg       0.68      0.68      0.68       610
weighted avg       0.68      0.68      0.68       610

AUC Score Training Data: 0.742
**Classification Report Testing Data:**
              precision    recall  f1-score   support

           0       0.72      0.65      0.68       145
           1       0.61      0.69      0.65       117

    accuracy                           0.67       262
   macro avg       0.67      0.67      0.67       262
weighted avg       0.67      0.67      0.67       262

AUC Score Testing Data: 0.705
```

Figure 11: CLassification report for Combination 1

```
True Negatives:  227
False Positives:  99
False Negatives:  98
True Positives:  186
```

Figure 12: Confusion Matrix for train data set

```
True Negatives:  94
False Positives:  51
False Negatives:  36
True Positives:  81
```

Figure 13: Confusion Matrix for test data set

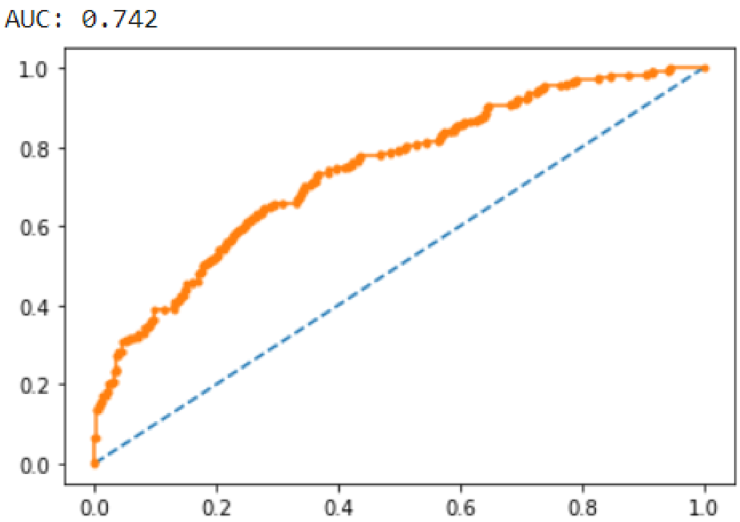


Figure 14: AUC curve train

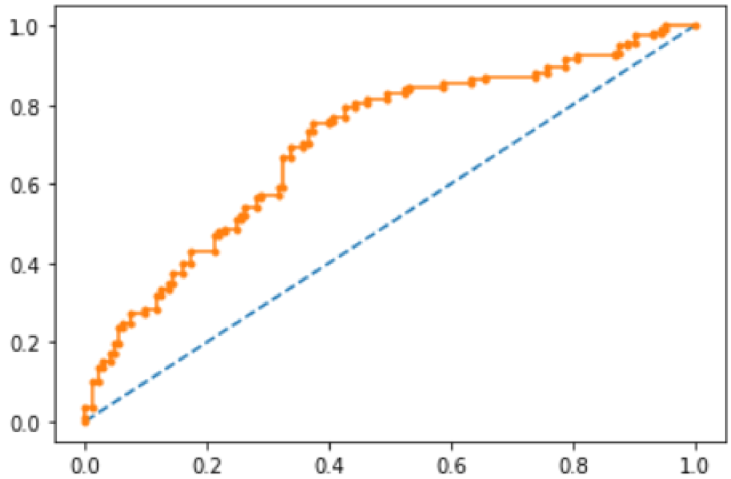


Figure 15: AUC curve test

<u>Inference:</u>

- Accuracy for the train set was found to be 0.67 and for test set 0.68
- Precision for Holiday Package 'Yes' in the train set are found to be 0.65 and for test 0.61, In the test set, it implies from the confusion matrix that 51 instances are false positives
- Recall for claim status 'Yes' in the train set was found to be 0.65 and for test 0.69. This implies 0.31 were wrongly claimed as 'No'. From the confusion matrix of test set we can see that 36 instances are false negatives.
- The accuracy and precision values are almost similar for both the training and test data set which implies no overfitting and underfitting happened in the model
- Precision metrics plays a very important role for this particular business problem. Since there are 51 false positives present, it could lead to a negative implication to Travel agency.
- Recall metrics also have an implication to the business. Since, there are 186 false negatives present in, it could lead to have a negative impression on travel agency

- Area under the curve o training data is 74% and on test data is 70% which seems good. AUC graph for both the test and train dataset are not flat which implies a good performance model
- Overall, it is a decent model can be used for prediction.

<u>Combination 2</u>

Classification Report Training Data:

	precision	recall	f1-score	support
0	0.53	0.95	0.68	326
1	0.41	0.04	0.07	284
accuracy			0.53	610
macro avg	0.47	0.49	0.38	610
weighted avg	0.47	0.53	0.40	610

AUC Score Training Data: 0.572

Classification Report Testing Data:

	precision	recall	f1-score	support
0	0.55	0.98	0.70	145
1	0.25	0.01	0.02	117
accuracy			0.55	262
macro avg	0.40	0.49	0.36	262
weighted avg	0.42	0.55	0.40	262

AUC Score Testing Data: 0.629

Figure 16: Classification Report for combination 2

True Negatives: 227
False Positives: 99
False Negatives: 98
True Positives: 186

Figure 17: Confusion matrix for train set

True Negatives: 94
False Positives: 51
False Negatives: 36
True Positives: 81

Figure 18: Confusion matrix for test set

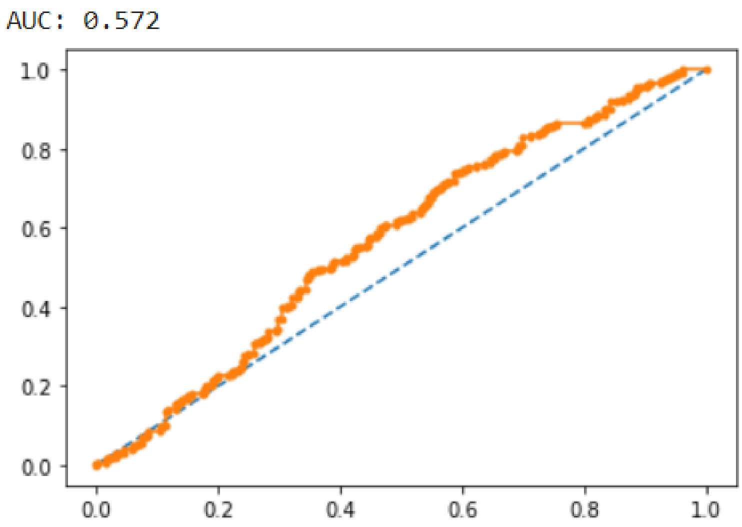


Figure 19: ROC AUC curve train

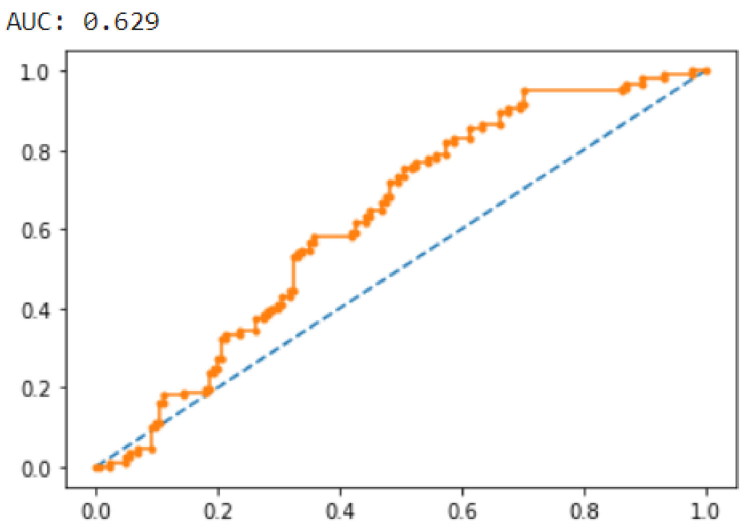


Figure 20: ROC AUC curve test

<u>Inference:</u>

- Accuracy for the train set was found to be 0.53 and for test set 0.55
- Precision for Holiday Package 'Yes' in the train set are found to be 0.41 and for test 0.25, In the test set, it implies from the confusion matrix that 51 instances are false positives
- Recall for claim status 'Yes' in the train set was found to be 0.04and for test 0.95. This implies 0.05 were wrongly claimed as 'No'. From the confusion matrix of test set we can see that 36 instances are false negatives.
- Recall value is very less
- The accuracy and precision values are almost similar for both the training and test data set which implies no overfitting and underfitting happened in the model
- Area under the curve of training data is 57% and on test data is 63% which does not seem good. AUC graph for both the test and train dataset are nearly flat which implies average performance model

- Overall, it is not a decent model cannot be used for prediction

Combination 3

```
**Classification Report Training Data:**
              precision    recall  f1-score   support

           0       0.54      1.00      0.70       326
           1       1.00      0.00      0.01       284

    accuracy                           0.54       610
   macro avg       0.77      0.50      0.35       610
weighted avg       0.75      0.54      0.38       610

AUC Score Training Data: 0.590
**Classification Report Testing Data:**
              precision    recall  f1-score   support

           0       0.55      1.00      0.71       145
           1       0.00      0.00      0.00       117

    accuracy                           0.55       262
   macro avg       0.28      0.50      0.36       262
weighted avg       0.31      0.55      0.39       262

AUC Score Testing Data: 0.634
```

Figure 21: Classification Report for combination 3

```
True Negatives:  326
False Positives:  0
False Negatives:  283
True Positives:  1
```

Figure 22: Confusion matrix for train dataset

```
True Negatives:  145
False Positives:  0
False Negatives:  117
True Positives:  0
```

Figure 23: Confusion matrix for test dataset

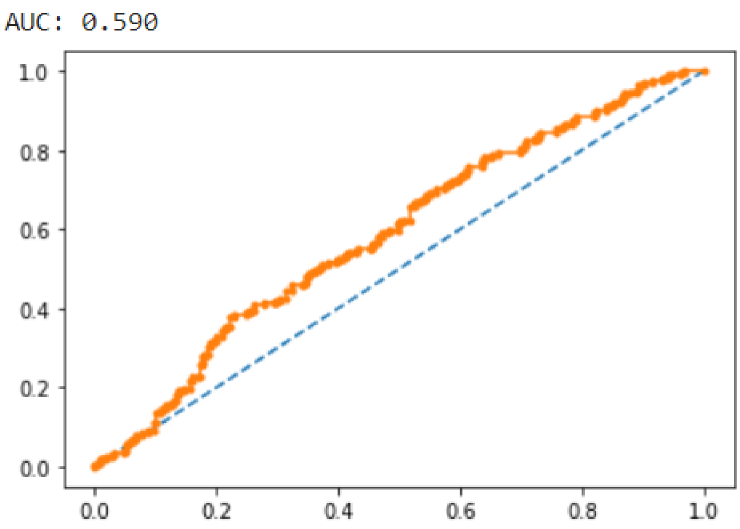


Figure 24: ROC AUC curve for the train data set

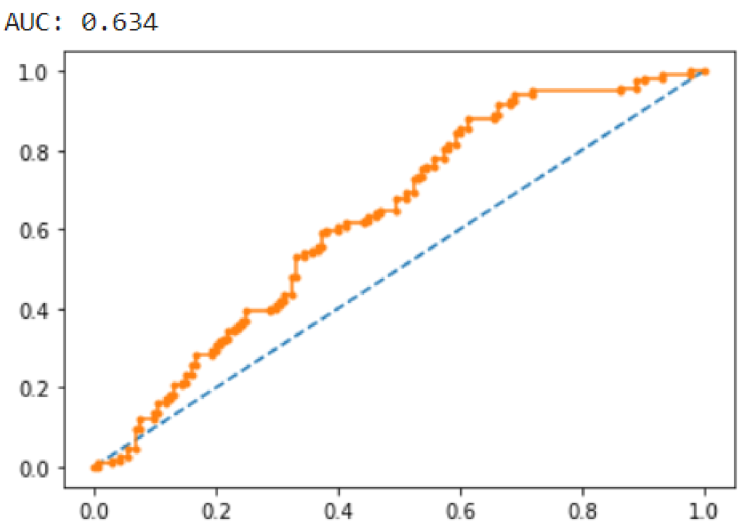


Figure 25: ROC AUC curve for the test data set

Inference:

- Accuracy for the train set was found to be 0.54 and for test set 0.55
- Precision for Holiday Package 'Yes' in the train set are found to be 1 and for test 0.54, In the test set, it implies from the confusion matrix that there are no instances of false positives
- Recall for claim status 'Yes' in the train set was found to be 0 and for test 0.0. However, From the confusion matrix of test set we can see that 283 instances are false negatives. Which is a slippery point for this model
- Recall value is 0
- The accuracy and precision values are almost similar for both the training and test data set which implies no overfitting and underfitting happened in the model

- Area under the curve of training data is 59% and on test data is 63% which does not seem good. AUC graph for both the test and train dataset are nearly flat which implies average performance model
- Overall, it is not a decent model and cannot be used for prediction

Combination 4

```
**Classification Report Training Data:**
              precision    recall  f1-score   support

           0       0.70      0.70      0.70       326
           1       0.66      0.65      0.65       284

    accuracy                           0.68       610
   macro avg       0.68      0.68      0.68       610
weighted avg       0.68      0.68      0.68       610

AUC Score Training Data: 0.743
**Classification Report Testing Data:**
              precision    recall  f1-score   support

           0       0.72      0.67      0.70       145
           1       0.62      0.68      0.65       117

    accuracy                           0.68       262
   macro avg       0.67      0.68      0.67       262
weighted avg       0.68      0.68      0.68       262

AUC Score Testing Data: 0.704
```

Figure 26: Classification Report for combination 4

```
True Negatives:  229
False Positives:  97
False Negatives:  99
True Positives:  185
```

Figure 27: Confusion matrix for train set

```
True Negatives:  97
False Positives:  48
False Negatives:  37
True Positives:  80
```

Figure 28: Confusion matrix for test set

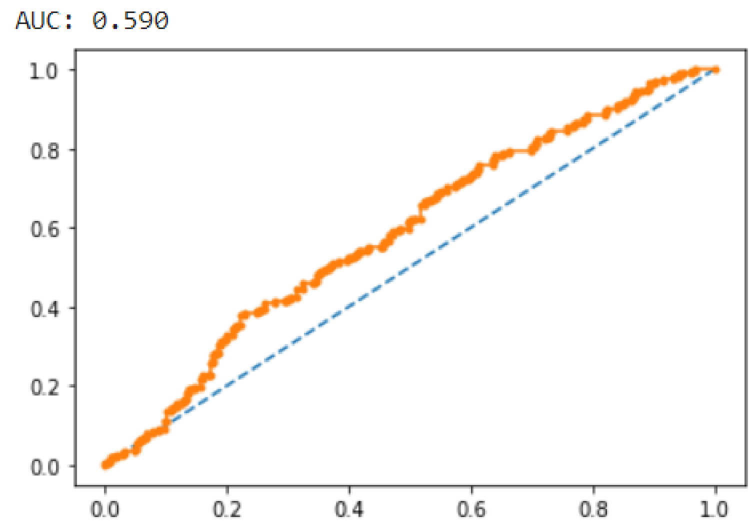


Figure 29: ROC AUC curve for train set

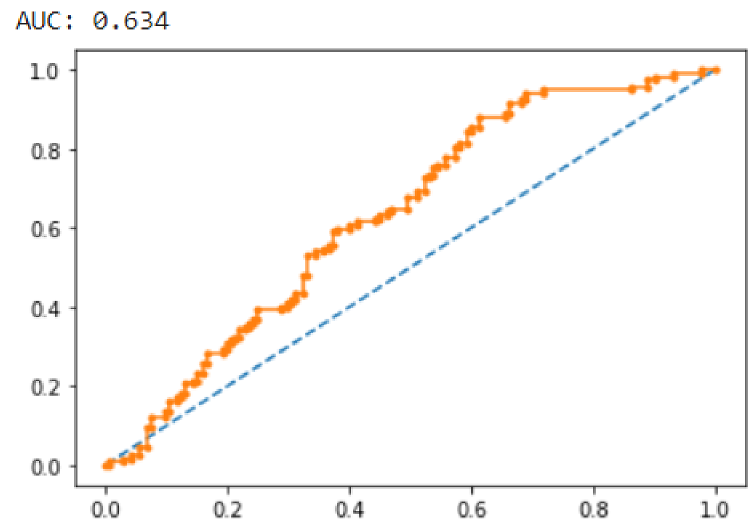


Figure 30: ROC AUC curve for test set

<u>Inference:</u>

- Accuracy for the train set was found to be 0.68 and for test set 0.68
- Precision for Holiday Package 'Yes' in the train set are found to be 0.66 and for test 0.62, In the test set, it implies from the confusion matrix that 51 instances are false positives
- Recall for claim status 'Yes' in the train set was found to be 0.65 and for test 0.68. This implies 0.32 were wrongly claimed as 'No'. From the confusion matrix of test set we can see that 37 instances are false negatives.
- The accuracy and precision values are almost similar for both the training and test data set which implies no overfitting and underfitting happened in the model

- Precision metrics plays a very important role for this particular business problem. Since there are 48 false positives present, it could lead to a negative implication to Travel agency.
- Recall metrics also have an implication to the business. Since, there are 37 false negatives present in, it could lead to have a negative impression on the travel agency
- Area under the curve o training data is 59% and on test data is 64% which seems good. AUC graph for both the test and train dataset are not flat which implies a good performance model
- Overall, it is a decent model can be used for prediction

<u>LDA</u>

<u>Combination 1</u>

	precision	recall	f1-score	support
0	0.67	0.78	0.72	326
1	0.69	0.55	0.61	284
accuracy			0.67	610
macro avg	0.68	0.67	0.67	610
weighted avg	0.68	0.67	0.67	610

AUC Score Training Data: 0.746

Classification Report Testing Data:

	precision	recall	f1-score	support
0	0.66	0.68	0.67	145
1	0.58	0.56	0.57	117
accuracy			0.63	262
macro avg	0.62	0.62	0.62	262
weighted avg	0.62	0.63	0.63	262

AUC Score Testing Data: 0.697

Figure 31: Confusion matrix for combination 1 in LDA

True Negatives: 254
False Positives: 72
False Negatives: 127
True Positives: 157

Figure 32: Confusion matrix for train set

```
True Negatives:  98
False Positives:  47
False Negatives:  51
True Positives:  66
```

Figure 33: Confusion Matrix for test set

AUC: 0.746

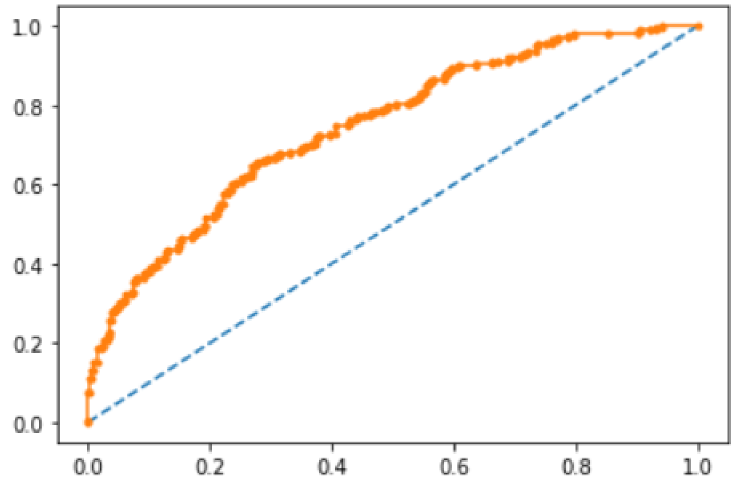


Figure 34: ROC AUC curve for combination 1

AUC: 0.697

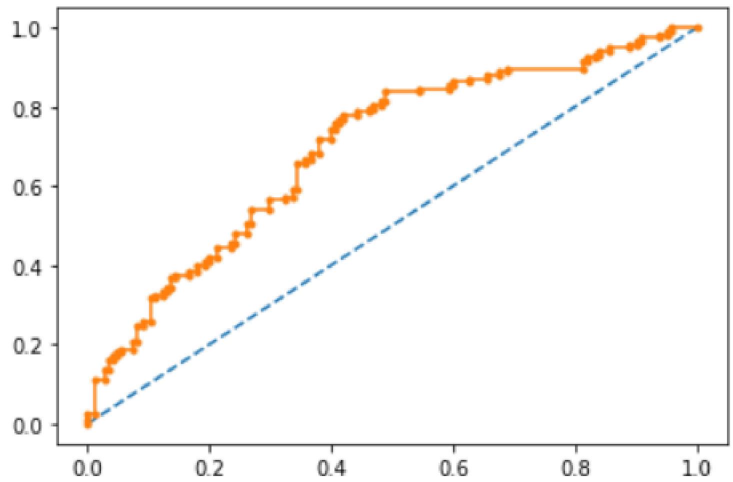


Figure 35: ROC AUC curve for combination 2

<u>Inference:</u>

- Accuracy for the train set was found to be 0.67 and for test set 0.63
- Precision for Holiday Package 'Yes' in the train set are found to be 0.69 and for test 0.58, In the test set, it implies from the confusion matrix that 47 instances are false positives

- Recall for claim status 'Yes' in the train set was found to be 0.55 and for test 0.56. This implies 0.44 were wrongly claimed as 'No'. From the confusion matrix of test set we can see that 51 instances are false negatives.
- The accuracy and precision values are almost similar for both the training and test data set which implies no overfitting and underfitting happened in the model
- Precision metrics plays a very important role for this particular business problem. Since there are 48 false positives present, it could lead to a negative implication to Travel agency.
- Recall metrics also have an implication to the business. Since, there are 51false negatives present in, it could lead to have a negative impression on the travel agency
- Area under the curve o training data is 75% and on test data is 65% which seems good. AUC graph for both the test and train dataset are not flat which implies a good performance model
- Overall, it is a decent model can be used for prediction

Combination 2

```
              precision    recall  f1-score   support

           0       0.67      0.79      0.72       326
           1       0.69      0.55      0.61       284

    accuracy                           0.68       610
   macro avg       0.68      0.67      0.67       610
weighted avg       0.68      0.68      0.67       610

AUC Score Training Data: 0.745
**Classification Report Testing Data:**
              precision    recall  f1-score   support

           0       0.68      0.70      0.69       145
           1       0.62      0.60      0.61       117

    accuracy                           0.66       262
   macro avg       0.65      0.65      0.65       262
weighted avg       0.66      0.66      0.66       262

AUC Score Testing Data: 0.703
```

Figure 36: Classification report for combination 2

```
True Negatives:  256
False Positives:  70
False Negatives:  127
True Positives:  157
```

Figure 37: Confusion matrix for train

```
True Negatives:  102
False Positives:  43
False Negatives:  47
True Positives:  70
```

Figure 38: confusion matrix for test

AUC: 0.745

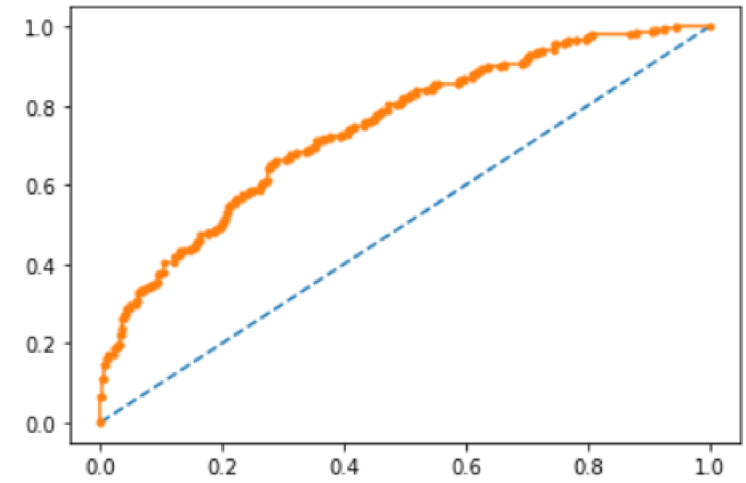


Figure 39: ROC AUC curve for train

AUC: 0.703

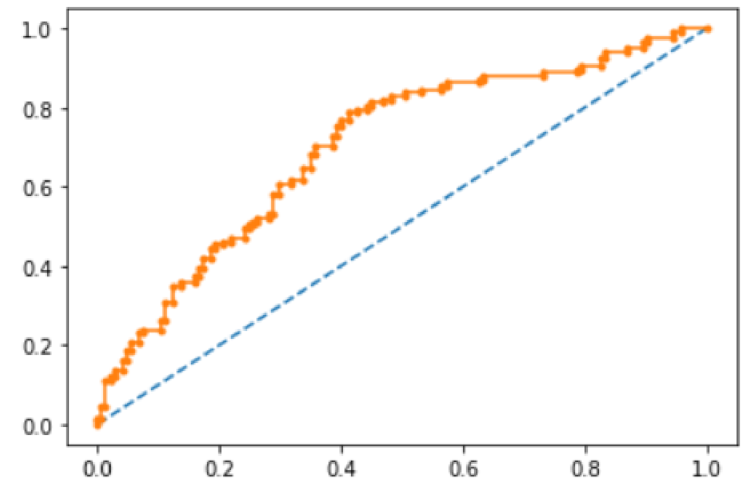


Figure 40: ROC AUC curve for test

<u>Inference:</u>

- Accuracy for the train set was found to be 0.68 and for test set 0.66
- Precision for Holiday Package 'Yes' in the train set are found to be 0.69 and for test 0.62, In the test set, it implies from the confusion matrix that 47 instances are false positives
- Recall for claim status 'Yes' in the train set was found to be 0.55 and for test 0.60. This implies 0.40 were wrongly claimed as 'No'. From the confusion matrix of test set we can see that 47 instances are false negatives.

- The accuracy and precision values are almost similar for both the training and test data set which implies no overfitting and underfitting happened in the model
- Precision metrics plays a very important role for this particular business problem. Since there are 47 false positives present, it could lead to a negative implication to Travel agency.
- Recall metrics also have an implication to the business. Since, there are 43 false negatives present in, it could lead to have a negative impression on the travel agency
- Area under the curve o training data is 75% and on test data is 70% which seems good. AUC graph for both the test and train dataset are not flat which implies a good performance model
- Overall, it is a decent model can be used for prediction

Observation

LRC: Logistic Regression combination

LDA: Linear Discriminant analysis

	Accuracy		Precision		Recall		ROC_AUC score		F1 score	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
LRC-1	0.68	0.67	0.65	0.61	0.65	0.69	0.74	0.70	0.62	0.65
LRC-2	0.57	0.63	0.41	0.25	0.01	0.04	0.58	0.63	0.61	0.62
LRC-3	0.55	0.54	1	1	0.0	0.00	0.59	0.63	0.01	0.00
LRC-4	0.68	0.68	0.66	0.62	0.65	0.68	0.59	0.63	0.65	0.65
LDA-1	0.67	0.63	0.69	0.58	0.55	0.56	0.75	0.70	0.61	0.57
LDA-2	0.68	0.66	0.69	0.62	0.55	0.60	0.75	0.70	0.61	0.61

Figure 41:Comparison table for all the models

Conclusion

Final Model selection in Logistic Regression

- Accuracy score for LRC-2 and LRC-3 are very low and precision value is too high and recall value is too low. ROC _AUC score is too low for both the models. And their area under graph is nearly flat. F1 score is evidently low in in both of them. Hence, we are not considering them as our final model among all other models in Logistic Regression.
- It is evident from the table that accuracy metrics are better for the LRC1, LRC4 models among all other models in Logistic Regression model
- Between LRC-1 And LRC-2 accuracy is similar, Precision metrics is similar, recall metrics is similar and F1 score metrics is similar. However, the ROC_AUC score is much better in LRC-1. Hence, we are considering LRC-1 as our final model among Logistic Regression models

Final Model selection in LDA

- Between LDA-1 And LDA-2 the ROC_AUC score is same. LDA-2 has better accuracy score, Precision score and Recall Score. Hence, we are considering LDA-2 as our final model between the two models in LDA

Final Model selection between Logistic Regression Model and LDA Model

	Accuracy		Precision		Recall		ROC_AUC score		F1 score	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
LRC-1	0.68	0.67	0.65	0.61	0.65	0.69	0.74	0.70	0.62	0.65
LDA-2	0.68	0.66	0.69	0.62	0.55	0.60	0.75	0.70	0.61	0.61

- Accuracy, Precision and ROC_AUC score metrics is similar in both the final models.
- Recall is greater in LRC-1 (logistic regression final model) compare to LDA-1 (LDA final model)
- F1 score is greater in LRC-1 (logistic regression final model) compare to LDA-1 (LDA final model)
- Hence, we are choosing LRC-1 (logistic regression final model) as our final model for prediction in the given model.

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

Important insights from the data set

- The average salary of employees who opted for holiday package is less.
- Among the employees who are not foreigners approximately only 38.7% opted for Holiday package whereas among the employees who are foreigners approximately 68.06% of the employees opted for Holiday package.
- Employees who have no young children below 7 years have nearly equal percentage of employees opting or not opting for the Holiday package. But employees who have 1 or 2 children below 7 years of age have much lesser percentage opting for holiday package.
- Employees who have no children above 7 years of age have lesser percentage opting for holiday package.
- The average age of employees with no children below 7 years who have opted for the holiday package is more than 5 years younger than employees who have not opted.
- Employees with 1 child have much lesser percentage of people opting for Holiday package.
- The average age of employees with no children who have opted for the holiday package is around 10 years younger than employees who have not opted. It could be observed from swarm plot that the employees above 50 have higher number of people not opting for the package.

Recommendation

- Since percentage of employees who are not foreigners opting for Holiday packages is less, this means the Tour agency needs to make packages that attract the local employees. May be the packages at present are focused more in within the country tours and hence the local employees are less interested in these packages. So international holiday packages could be more attractive to the local employees.

- Since the percentage of foreign employees have higher percentage opting for the packages. It may be because this section of employees is interested in seeing and knowing the foreign country. Hence packages designed at more culturally satisfying experiences could further bolster the package selection rate among the foreign employees.
- Since the average salary of the employees not opting for the holiday package is high, it may suggest that the packages are not luxurious enough for the high earners. So, more luxury-oriented packages with a higher price cap could be designed for this section of employees.
- Since percentage of employees with 1 or 2 children below 7 years of age opting for holiday package is less. The Tour agency needs to understand the reason behind this situation and see if the packages designed are comfortable enough for employees with children of this age like baby care facilities. Including features specific to children of this age group like amusement park tours could further make the packages attractive for the parents.
- Employees with no children who are aged above 50 have higher count of employees not opting for holiday packages. So, packages aiming at more comfort and relaxation-oriented features could be designed for this section.