A CASE STUDY ON GEM STONES CO LTD.

1.1)EDA & UNIVARIATE, BIVARIATE ANALYSIS:

- We could see from below information of the dataset there are 3 variables of Object/Category type and rest are continuous . 'depth' variable is missing 697 data points.

```
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   carat    26967 non-null  float64
 1   cut      26967 non-null  object
 2   color    26967 non-null  object
 3   clarity  26967 non-null  object
 4   depth    26270 non-null  float64
 5   table    26967 non-null  float64
 6   x        26967 non-null  float64
 7   y        26967 non-null  float64
 8   z        26967 non-null  float64
 9   price    26967 non-null  int64
dtypes: float64(6), int64(1), object(3)
```

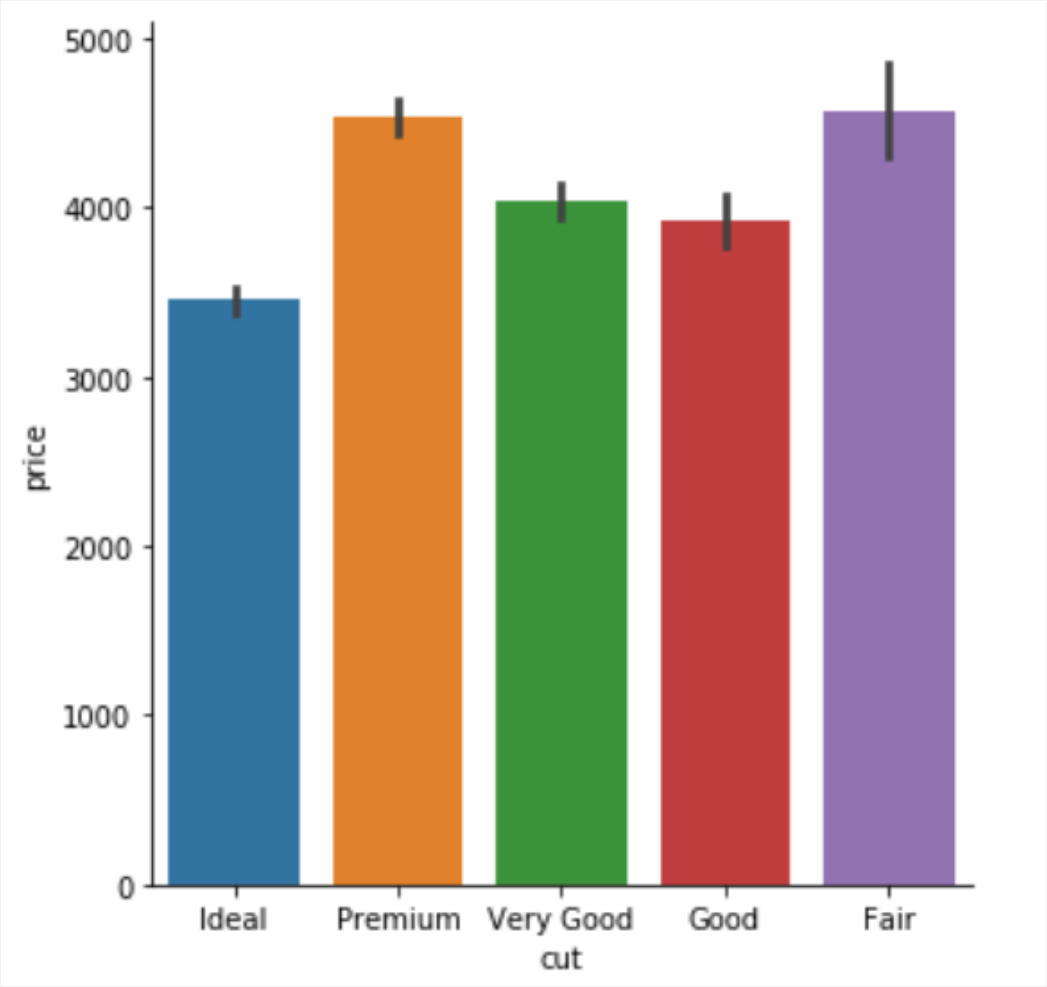
- From below summary of the dataset we could see that variables of different scale .We might need to standardize the variables to bring all of them to the same scale . For example ‘carat’ is within 10s ‘table’ is expressed in percentages.

	carat	depth	table	x	y	z	price
count	26967.000000	26270.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000
mean	0.798375	61.745147	57.456080	5.729854	5.733569	3.538057	3939.518115
std	0.477745	1.412860	2.232068	1.128516	1.166058	0.720624	4024.864666
min	0.200000	50.800000	49.000000	0.000000	0.000000	0.000000	326.000000
25%	0.400000	61.000000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	0.700000	61.800000	57.000000	5.690000	5.710000	3.520000	2375.000000
75%	1.050000	62.500000	59.000000	6.550000	6.540000	4.040000	5360.000000
max	4.500000	73.600000	79.000000	10.230000	58.900000	31.800000	18818.000000

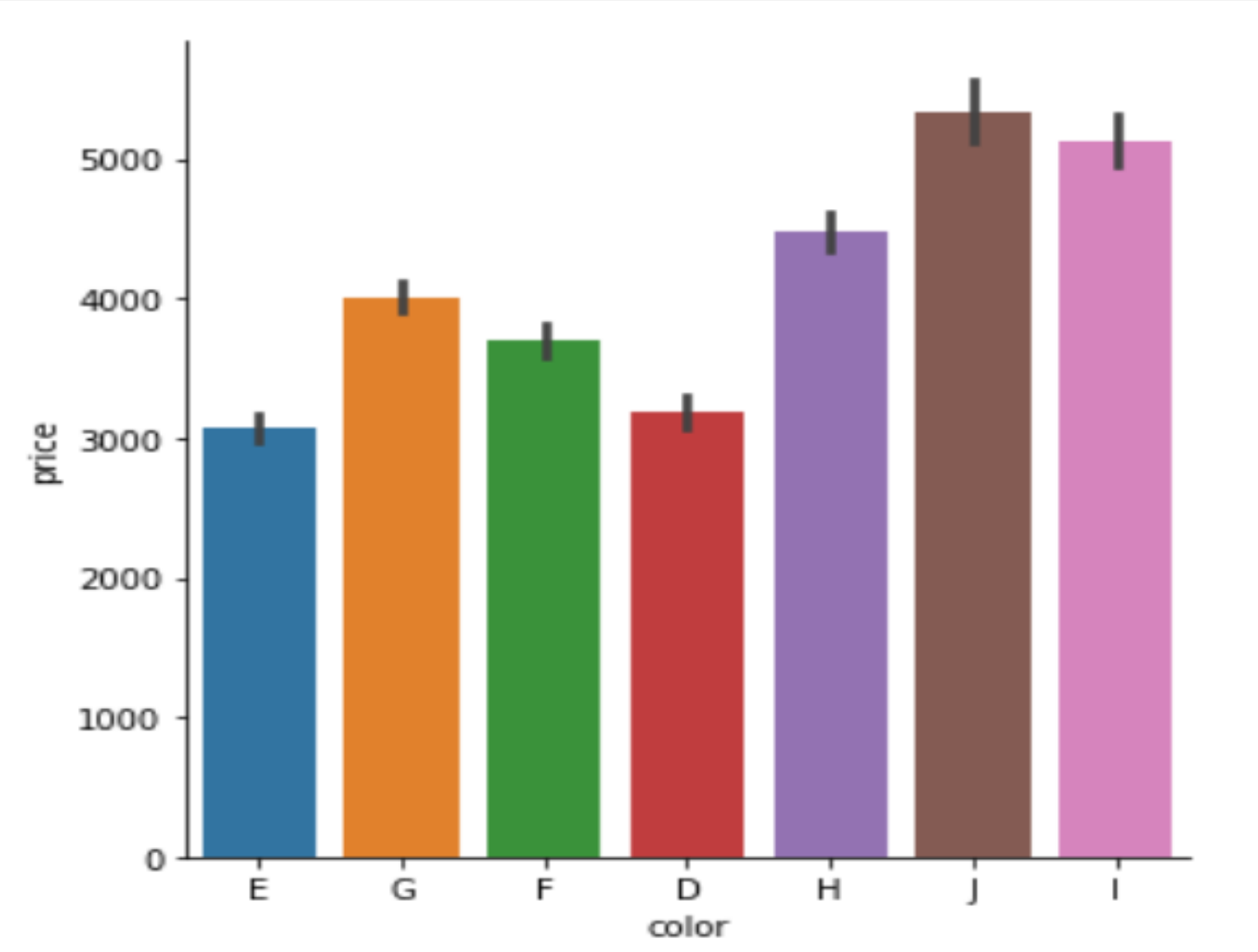
- For categorical variable we could see following has most occurrences : ’ideal’ in column ‘cut’, ’G’ in column ‘color’, ‘SI1’ in column ‘clarity’

	cut	color	clarity
count	26967	26967	26967
unique	5	7	8
top	Ideal	G	SI1
freq	10816	5661	6571

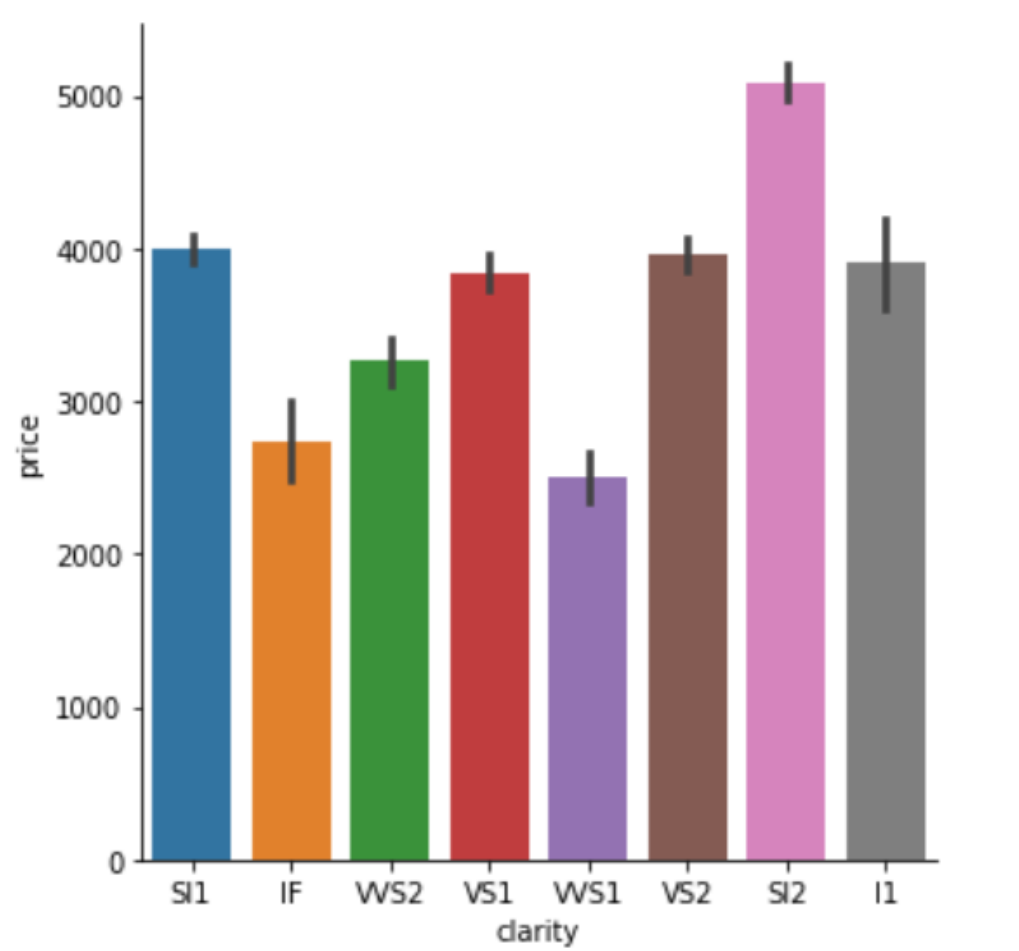
- From below plot depicting ‘price’ vs ‘cut’ quality we could see ‘premium’ quality cut and ‘fair’ quality cut has highest mean price. It is surprising to find that ‘ideal’ quality cut has lower mean price than the ‘fair’ quality cut, since ‘ideal’ represents good quality cut it should have fetched higher price.



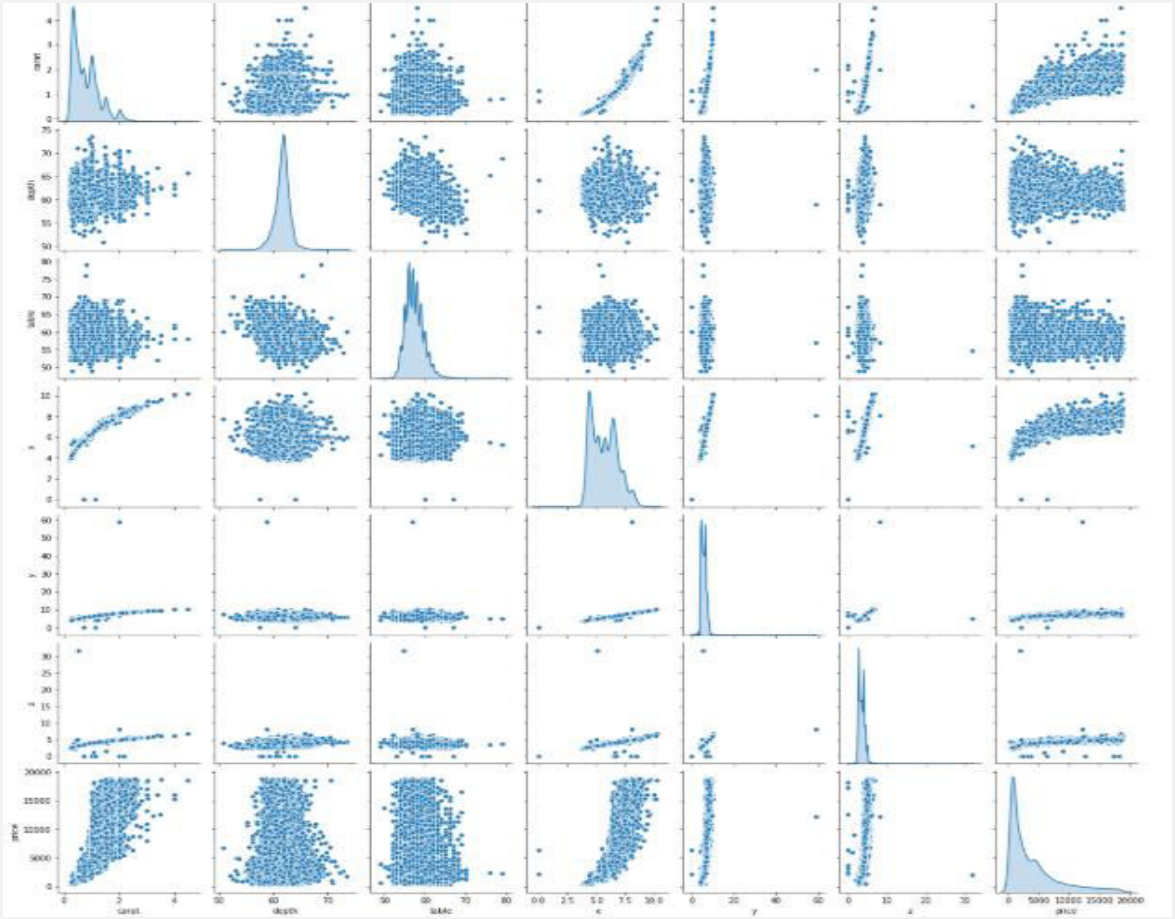
- Bar plot on ‘color’ vs ‘price’ shows mean price is higher for ‘J’(worst of all the colors) followed by ‘I’



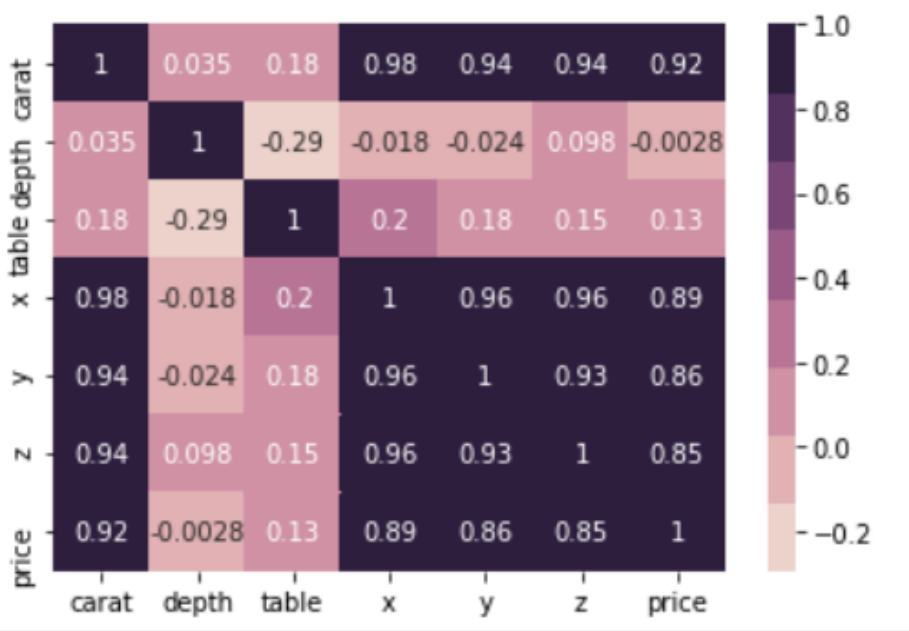
- 'SI1' level is the most popular in sample dataset followed by VS2
- 'SI2' has the highest mean price followed by VS2,VS1 and SI1,again it is surprising even though 'IF' represents close to flawlessness it doesn't show a good mean price(again may be due to the low no of observations in sample dataset)



- From pair plot of variables in the dataset we could see there is some collinearity between few independent variables and moreover there is strong relationship between ‘carat’ and ‘price’. High collinearity between x,y,z and ‘carat’ could be seen from the pair plot

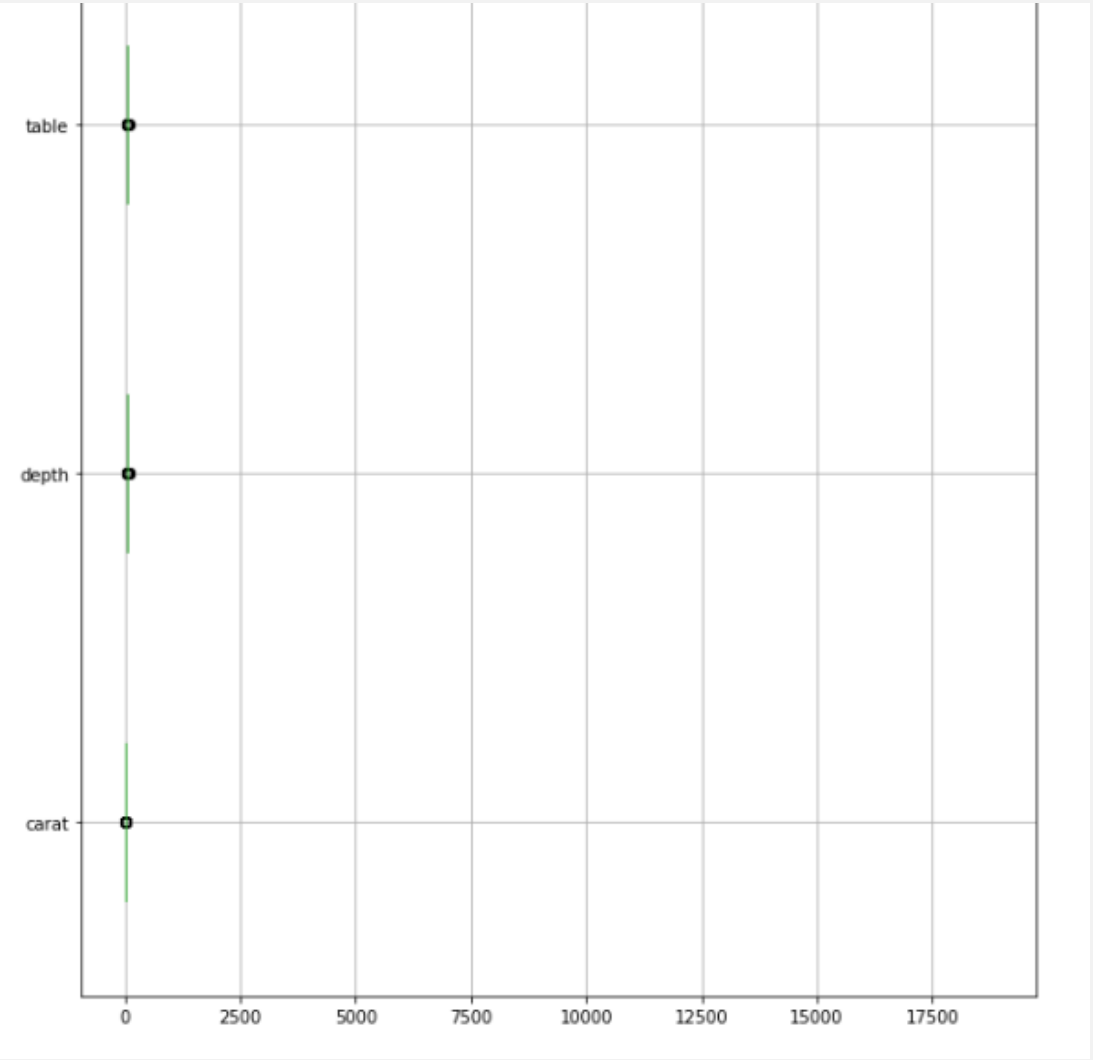


- From the below correlation matrix we could see as expected there is a high correlation between ‘x’,’y’,’z’ and ‘carat’ which implies presence of multi collinearity. Further down the analysis we could remove either set of variables depending on the importance of feature from business perspective:



- From above analysis we could conclude that dimensions 'x','y','z' and 'carat' has high correlation.These 4 attributes indicates dimensions of the Zirconium gem.x,y,z together contributes volume of the gem and carat contributes to weight per cubic centimeter of the Zirconium gem.Since 'carat' column already relates to volume of the gem , we could remove the attributes x,y,z from the dataset before applying the model.

- There are outliers present in data set which needs to treated:



1.2)CHECK NULL VALUES & SCALING

- As mentioned before ‘depth’ column has 697 null values ,here we try to impute the missing value with median value of the ‘depth’ variable

```
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   carat    26967 non-null  float64
 1   cut      26967 non-null  object
 2   color    26967 non-null  object
 3   clarity  26967 non-null  object
 4   depth    26270 non-null  float64
 5   table    26967 non-null  float64
 6   x        26967 non-null  float64
 7   y        26967 non-null  float64
 8   z        26967 non-null  float64
 9   price    26967 non-null  int64
dtypes: float64(6), int64(1), object(3)
```

- There are zeros present in x,y and z column, since these represent dimensions of a gem values cannot be zero. Moreover since x,y and z has correlation with variable ‘carat’, these variables removed before creating the model for the reasons quoted below
- From above analysis we could conclude that dimensions 'x','y','z' and 'carat' has high correlation. These 4 attributes indicates dimensions of the Zirconium gem.x,y,z together contributes volume of the gem and carat contributes to weight per cubic centimeter of the Zirconium gem.Since 'carat' column already relates to volume of the gem , we could remove the attributes x,y,z from the dataset before applying the model.
- Scaling is necessary for the dataset since Linear regression depends on calculation of weights/coefficients for the variables , so if all the feature selected are not brought to same scale it might lead to selection undesired weights which will result in wrong predictions. Also variables present in the dataset are also having different scale which also is a concrete reason to Standardize the variables

- Also for categorical variables since they are of ordinal nature here LabelEncoder is used to assign integer value depending on the rank.

1.3)SPLITTING DATASET & MODEL PERFORMANCE

- Categorical/String type variables are encoded using LableEncoding technique,LabelEncoder is used since all the categorical variables present in the dataset are of type ordinal(rank based).Since LabelEncoder uses numerical value to assign rank for a specific ‘string’ we use method for all the categorical variables.
- Train and test data split into 70:30 and model is built from both LinearRegression in sklearn and smf in statsmodel.
- Training dataset are scaled using StandardScaler and test dataset is transformed from parameters of training dataset.

- We are able to train the model to have R^2 and adjusted R^2 as 0.882 which implies that model is in between overfit and good fit , in other words it represent how much variance is capture by the model with respect to actual/observed data
- Accuracy score of train is 88.21% while accuracy for the test dataset being 88.69%.Here model is performing better by a small margin for test dataset.
- All the independent can be used for final model building since at 5% significance level we could select all the variables for constructing the linear regression equation.
- Most Weightage/imp

OLS Regression Results

Dep. Variable:	price	R-squared:	0.882
Model:	OLS	Adj. R-squared:	0.882
Method:	Least Squares	F-statistic:	2.350e+04
Date:	Sat, 16 May 2020	Prob (F-statistic):	0.00
Time:	19:20:09	Log-Likelihood:	-1.6301e+05
No. Observations:	18853	AIC:	3.260e+05
Df Residuals:	18846	BIC:	3.261e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3934.0075	10.028	392.320	0.000	3914.353	3953.662
carat	3985.1461	10.916	365.085	0.000	3963.750	4006.542
cut	76.5474	10.278	7.448	0.000	56.402	96.693
color	-456.5158	10.502	-43.469	0.000	-477.101	-435.931
clarity	527.8854	10.293	51.287	0.000	507.711	548.060
depth	-131.7312	10.643	-12.378	0.000	-152.592	-110.871
table	-221.6499	10.796	-20.530	0.000	-242.812	-200.488

Omnibus:	5170.694	Durbin-Watson:	1.967
Prob(Omnibus):	0.000	Jarque-Bera (JB):	28213.401
Skew:	1.212	Prob(JB):	0.00
Kurtosis:	8.481	Cond. No.	1.57

- RMSE- the difference between predicted and actuals vary by an amount 1362, which could be considered as somewhat insignificant value considering the value of Zirconium gem.

1.4)INSIGHTS/CONCLUSION/RECOMMENDATIONS

- Utmost Weightage is given to ‘carat’, since carat for a gem is an important attribute since it defines gem grading/profiling.
- Gem Stones Co Ltd can depend on the ‘carat’ attribute of a gem for profit making
- Carat weight is important attribute for price of the Zirconium gems and also reflects its rarity.
- Clarity attribute also contributes to the price of the Zirconium gems.Clarity defines how well the gem is without any imperfections(inclusions or blemishes).
- Cut could also have been a good attribute but here when compared to other attributes cut quality contributes least to the final equation.Maybe the complexity involved in the process of ‘cut’ it would have contributed less compared to other feature.

- Table and Depth together contributes to the the amount of light to be entered into gem to be refracted to give a brilliant glow to the viewer's eyes.
- Depth parameter should n't too high so that all light escapes from the first facet of the gem. It should have an ideal value between 55-67 % depending on the shape of the gem.Here to fetch a good price depth should be low.
- Table is the flat facet which we could see when gem is face up . As the largest facet in the gem it plays important role in giving its brilliant sparkle. Similar to depth it shouldn't have very high value.
- Color also contribute to brilliance of gem,here although we are unable to deduce based on what the color is ranked but we could assume from the coefficient and with general sense that light colors contribute to attractiveness to the gem while dark colors are very much less attractive

THE END

A CASE STUDY ON SELLING HOLIDAY PACKAGES

2.1)EDA & Bivariate, Univariate Analysis

- There are about 872 data points in the data set and none of them has null values. Of the 7 variables 2 are categorical and rest are of type numeric. 'Holiday_Package' is the target variable:

```
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Holliday_Package   872 non-null    object
 1   Salary             872 non-null    int64
 2   age                872 non-null    int64
 3   educ               872 non-null    int64
 4   no_young_children  872 non-null    int64
 5   no_older_children  872 non-null    int64
 6   foreign            872 non-null    object
dtypes: int64(5), object(2)
```

- Summary of dataset tells us that there are no obvious outliers present except for ‘Salary’ column, but again have to set up a box plot to draw conclusion. We could combine ‘**no_young_children’ and ‘no_older_children’ .**

	Salary	age	educ	no_young_children	no_older_children
count	872.000000	872.000000	872.000000	872.000000	872.000000
mean	47729.172018	39.955275	9.307339	0.311927	0.982798
std	23418.668531	10.551675	3.036259	0.612870	1.086786
min	1322.000000	20.000000	1.000000	0.000000	0.000000
25%	35324.000000	32.000000	8.000000	0.000000	0.000000
50%	41903.500000	39.000000	9.000000	0.000000	1.000000
75%	53469.500000	48.000000	12.000000	0.000000	2.000000
max	236961.000000	62.000000	21.000000	3.000000	6.000000

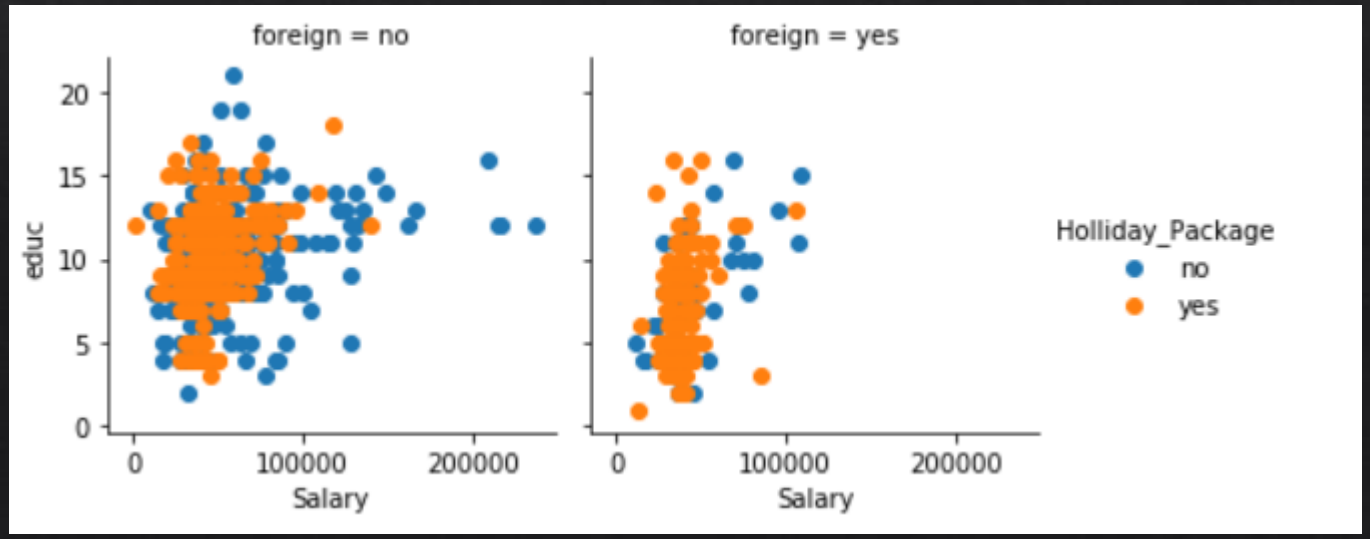
- From given dataset we could see there are more than 50% employees not opting for Holiday Package . Majority of the employees are not foreigners:

	Holliday_Package	foreign
count	872	872
unique	2	2
top	no	no
freq	471	656

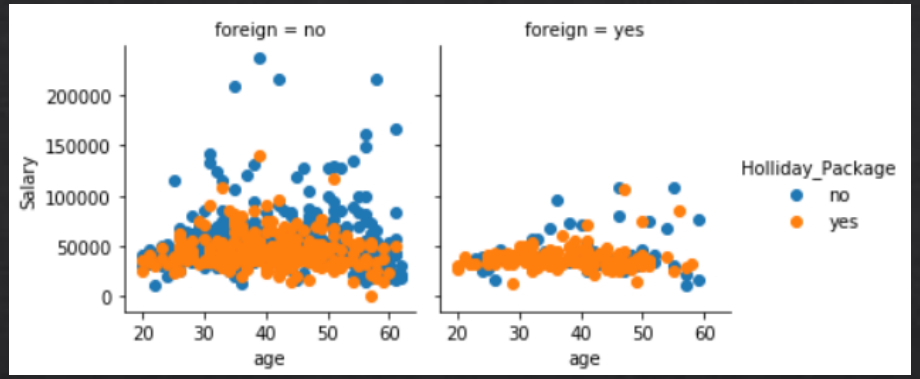
- We could see of the foreigners majority opt for holiday package where as for those who are not foreigners more than half of them do not opt for holiday package:

	foreign	Holliday_Package	Freq
0	no	no	402
1	no	yes	254
2	yes	no	69
3	yes	yes	147

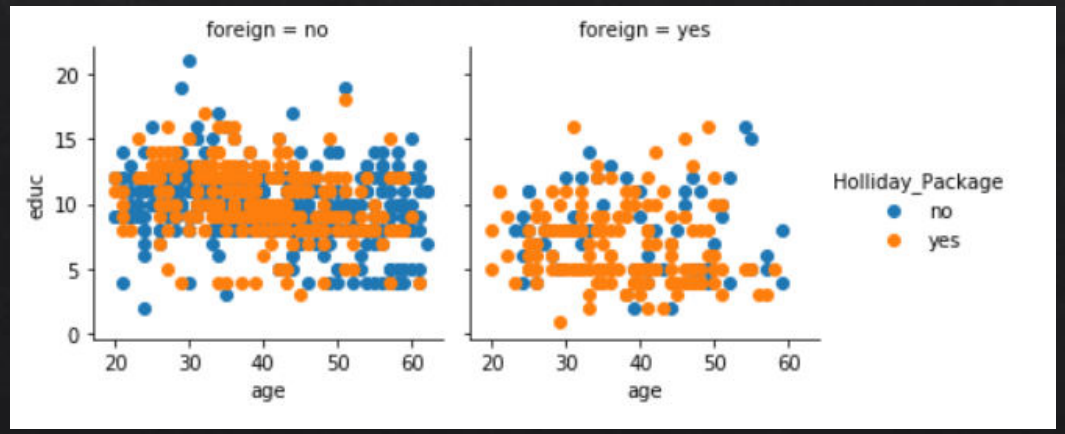
- For both foreigners and residents we could see opting holiday packages are the most for those who are earning between 0-70K and also who has years of education between 0-10.



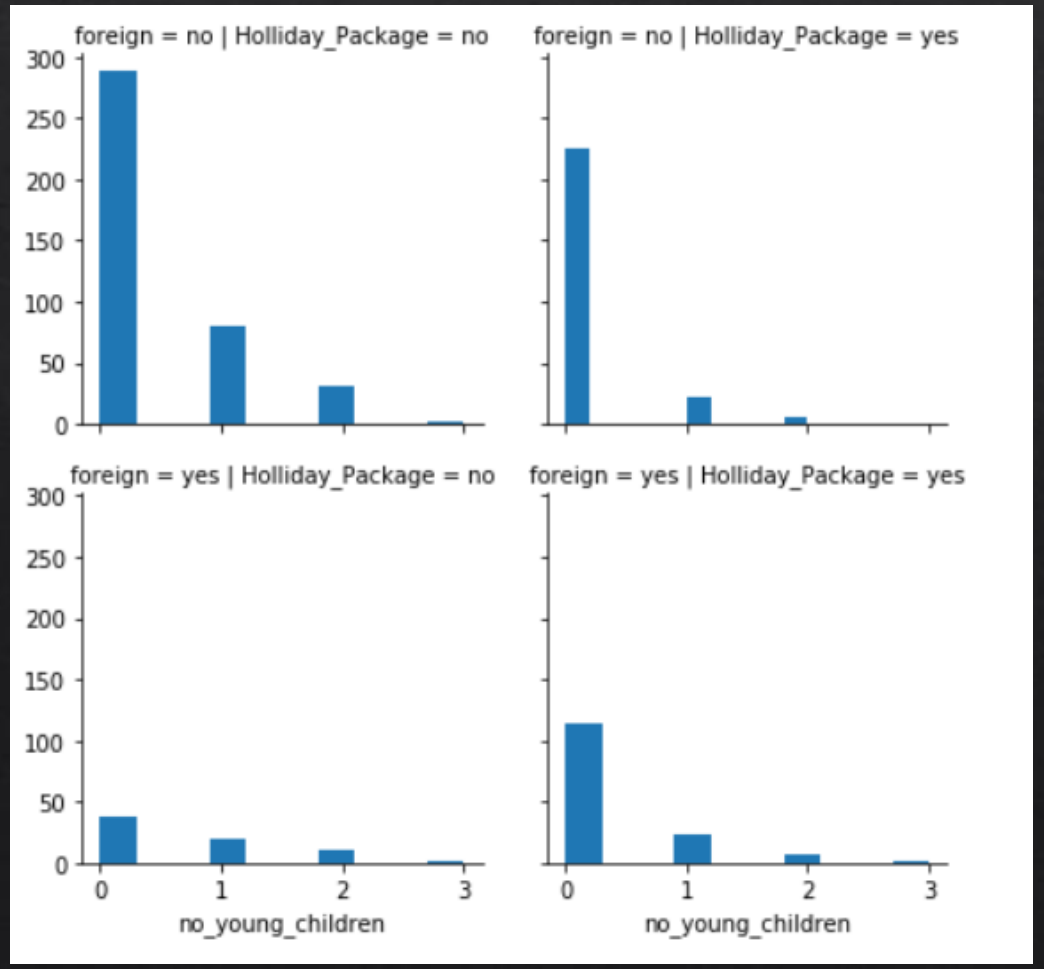
- We could see there heavy concentration of employees opting for package whose age are in range between 30-45. Since most of the employees in the dataset has salary in between 0-1Lac we are unable to find a trend who are earning 1Lac and above.

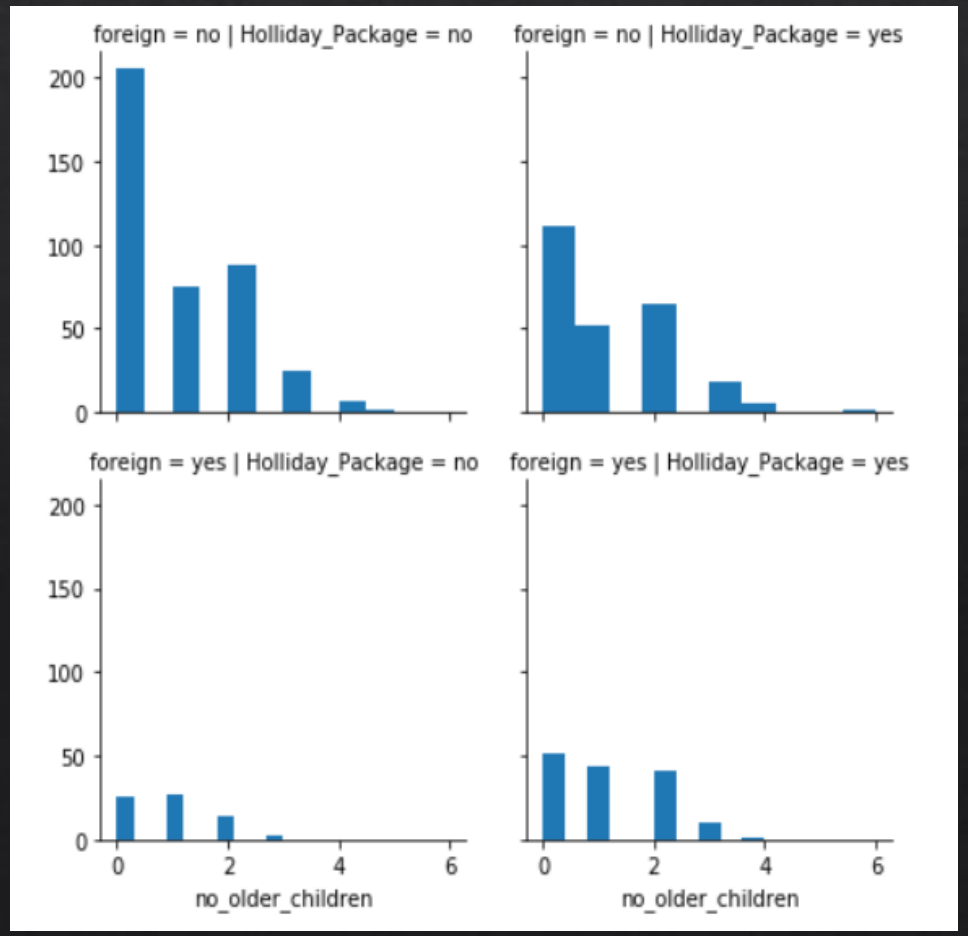


- Foreigners opt for the package irrespective of the age but as we go up the years of the education opting diminishes. While for residents most of the people opted as years of education between 10-15 and age between 30-45.

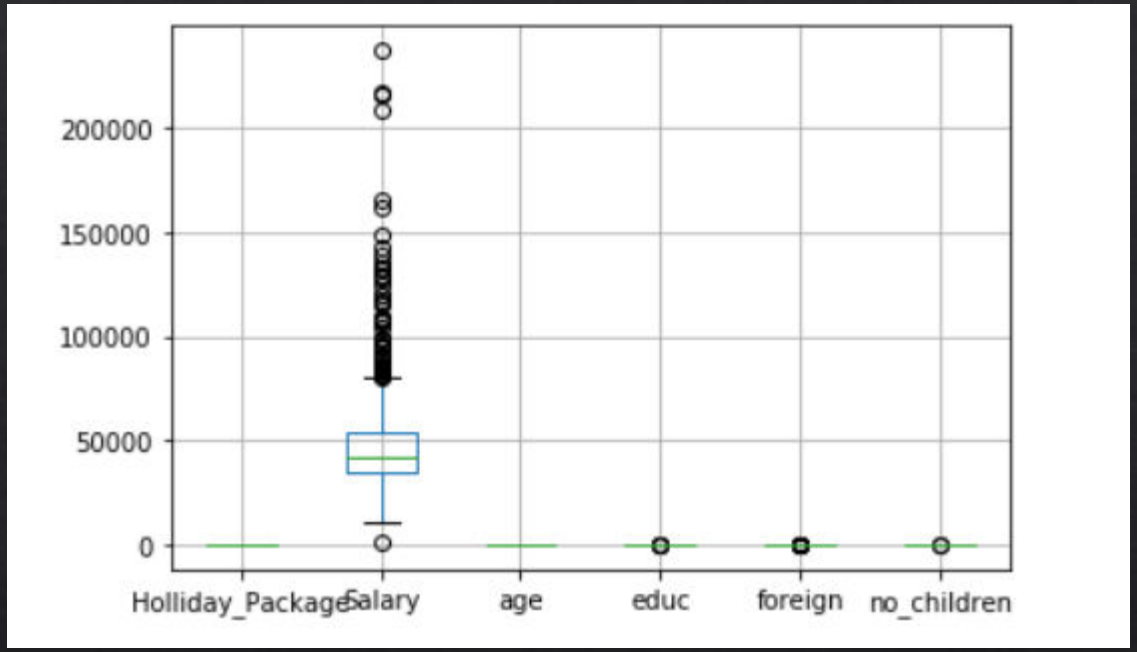


- Generally for an employees having more no: of children there is a decreasing trend for opting the package . Employees who has no children tend to opt for the Holiday Package. Trend is same for younger children and older children we could combine them into a single feature





- Here we avoid treating outliers since we may miss some important information on employees earning higher salaries and having higher education



2.2)Test-Train Split and Encoding

- Please refer python Notebook ‘Problem-2’ for more details . Since nothing to describe here.

2.3)Model Performance and Comparison

- From classification report and confusion matrix we could see LDA model performs a tad better than Logistic regression for train dataset . Again Logistic regression model took a heavy toll on precision/recall parameter . Both model have performed poorly in aspects of accuracy, precision and F1-score but when compared with each other LDA has slight edge over Logistic regression:

**********Classificiation Report and Confusion Matrix**********
**********Logistic Regression**********
Confusion Matrix
[[296 30]
 [261 23]]
Classification Report

	precision	recall	f1-score	support
0	0.53	0.91	0.67	326
1	0.43	0.08	0.14	284
accuracy			0.52	610
macro avg	0.48	0.49	0.40	610
weighted avg	0.49	0.52	0.42	610

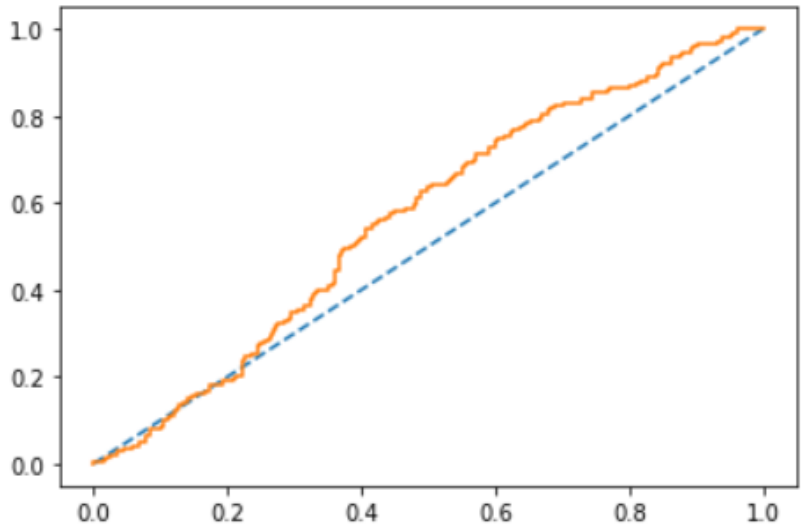
**********LDA**********
Confusion Matrix
[[263 63]
 [155 129]]
Classification Report

	precision	recall	f1-score	support
0	0.63	0.81	0.71	326
1	0.67	0.45	0.54	284
accuracy			0.64	610
macro avg	0.65	0.63	0.62	610
weighted avg	0.65	0.64	0.63	610

- Logistic Regression again performs poorer in terms of AUC –ROC for train dataset .Linear Discriminant analysis has better AUC compared to Logistic regression:

AUC: 0.566

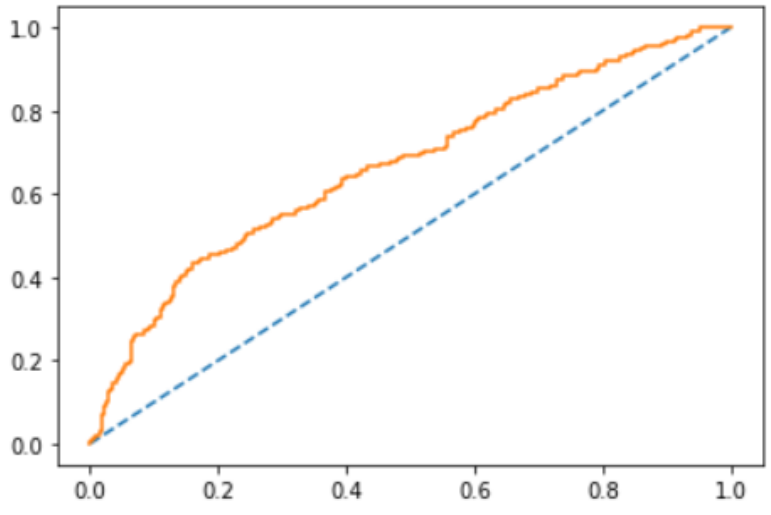
[<matplotlib.lines.Line2D at 0x1f76f20ac88>]



Logistic Regression-Model

AUC: 0.666

[<matplotlib.lines.Line2D at 0x1f7719ad240>]

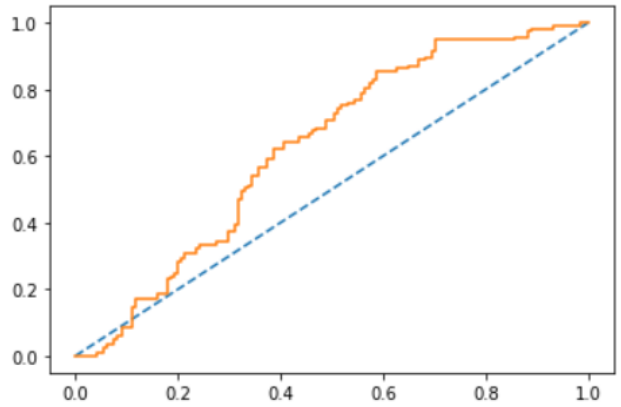


LDA-Model

- Also for the test data LDA performs well compared to Logistic Regression in all

AUC: 0.626

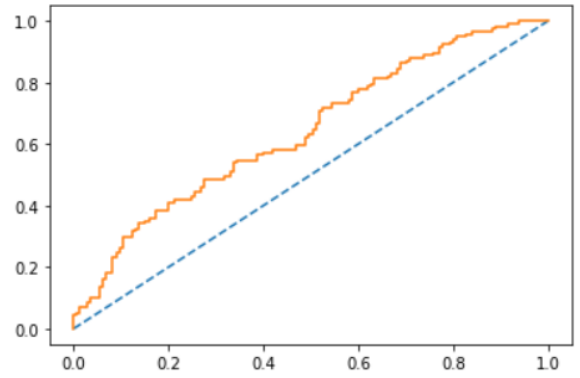
[<matplotlib.lines.Line2D at 0x1f76f17b198>]



Logistic Regression-Test

AUC: 0.642

[<matplotlib.lines.Line2D at 0x1f76f2001d0>]



LDA-Test

**********Classificiation Report and Confusion Matrix**********
**********Logistic Regression**********
Confusion Matrix
[[132 13]
[107 10]]
Classification Report

	precision	recall	f1-score	support
0	0.55	0.91	0.69	145
1	0.43	0.09	0.14	117
accuracy			0.54	262
macro avg	0.49	0.50	0.42	262
weighted avg	0.50	0.54	0.44	262

**********LDA**********
Confusion Matrix
[[108 37]
[66 51]]
Classification Report

	precision	recall	f1-score	support
0	0.62	0.74	0.68	145
1	0.58	0.44	0.50	117
accuracy			0.61	262
macro avg	0.60	0.59	0.59	262
weighted avg	0.60	0.61	0.60	262

- LDA and Logistic Regression performs poor on all parameters but when compared to each other LDA performs better for this dataset.
- LDA could be utilized for making predictions for Holiday_Packages for this particular case.
- Logistic Regression fails miserably in predicting/identifying the actual true positives.
- Logistic Regression shows significant improvement in AUC for test dataset .

2.4)Conclusion/Inferences

- Of the number of foreigners, majority of them opt for Holiday_Packages, while for residents case is opposite.
- Residents could be offered with attractive holiday packages , like Food,Stay and adventure activities could be provided as an add on to the packages. Since residents would have already visited the places around , inclusion of adventure like activities would attract them.
- Foreigners and residents having salary in range between 0-70k showing a greater interest in opting for Holiday Packages.
- Employees between age 20-45 are more enthusiastic in opting for Tour package.
- For people who have years of education between 5-15 shows interest , this demarcation is obvious for residents but not significant in case of foreigners

- Also we could see Employees having no children are opting for the packages more compared to those who at least have a child , which is kind of obvious trend since Employees with children would have more responsibilities, to attract this segment of customers we could include an attractive family packages(Packages which include more of sightseeing activities rather than adventurous activities).

THE END