Predictive Modelling

Problem Statement:

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

Data Dictionary:

Variable Name	Description
Carat	Carat weight of the cubic zirconia.
Cut	Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
Color	Colour of the cubic zirconia. With D being the best and J the worst.
Clarity	cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, FL = flawless, I3= level 3 inclusions) FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3
Depth	The Height of a cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.
Table	The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
Price	the Price of the cubic zirconia.
X	Length of the cubic zirconia in mm.
Y	Width of the cubic zirconia in mm.
Z	Height of the cubic zirconia in mm.

Exploratory Data Analysis:

- There are totally 26,967 records in the dataset with 11 columns. The **price** being the dependant variable we need to predict the price.
- In the 11 columns there is a index column named ***Unnamed: 0*** we don't need that column so we are dropping it.
- In the dataset cut, color and clarity are categorical variables

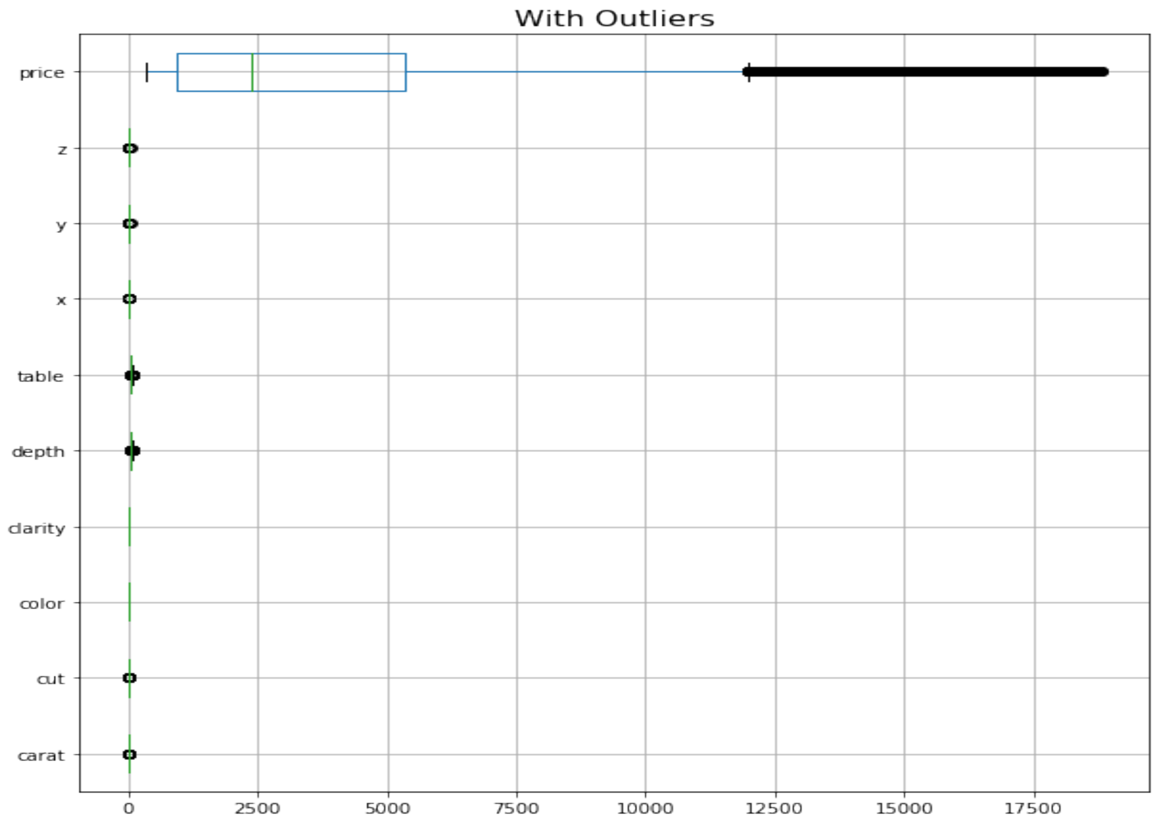
CUT : 5 values	COLOR : 7 Values	CLARITY : 8 Values
• Fair:781 • Good:2441 • Very Good:6030 • Premium:6899 • Ideal:10816	• J:1443 • I:2771 • D:3344 • H:4102 • F:4729 • E:4917 • G:5661	• I1: 365 • IF: 894 • VVS1: 1839 • VVS2: 2531 • VS1: 4093 • SI2: 4575 • VS2: 6099 • SI1: 6571

- There are 697 blank value in depth column we are going to fill that with mean
- We are going to remove the 34 duplicated records
- Below is the description of the data

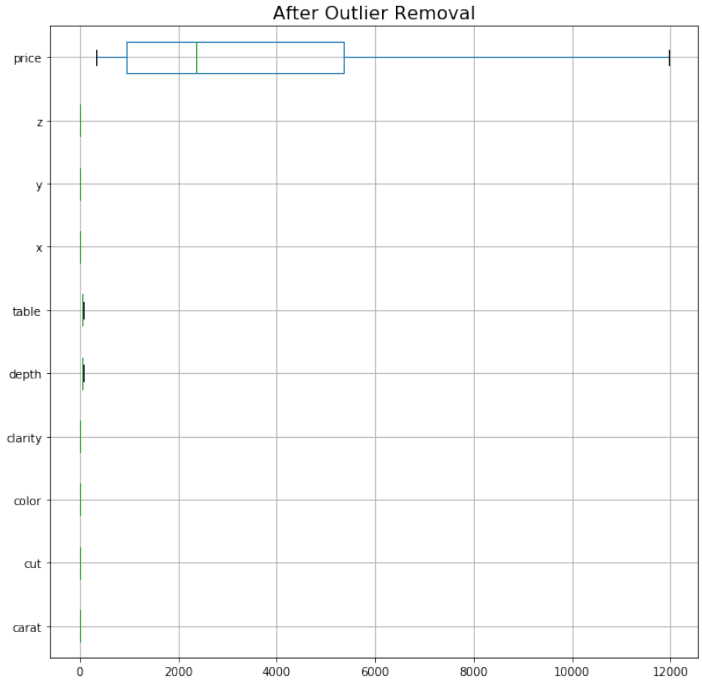
	carat	cut	color	clarity	depth	table	x	y	z	price
count	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000
mean	0.798375	2.554604	2.606111	3.833537	61.745147	57.456080	5.729854	5.733569	3.538057	3939.518115
std	0.477745	1.024243	1.705992	1.724904	1.394481	2.232068	1.128516	1.166058	0.720624	4024.864666
min	0.200000	0.000000	0.000000	0.000000	50.800000	49.000000	0.000000	0.000000	0.000000	326.000000
25%	0.400000	2.000000	1.000000	2.000000	61.100000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	0.700000	2.000000	3.000000	4.000000	61.800000	57.000000	5.690000	5.710000	3.520000	2375.000000
75%	1.050000	3.000000	4.000000	5.000000	62.500000	59.000000	6.550000	6.540000	4.040000	5360.000000
max	4.500000	4.000000	6.000000	7.000000	73.600000	79.000000	10.230000	58.900000	31.800000	18818.000000

- The data's are corelated especially carat is highly corelated with length, width, height of the stone. And also inter corelated with them. And the price is also corelated with X,Y,Z and carat

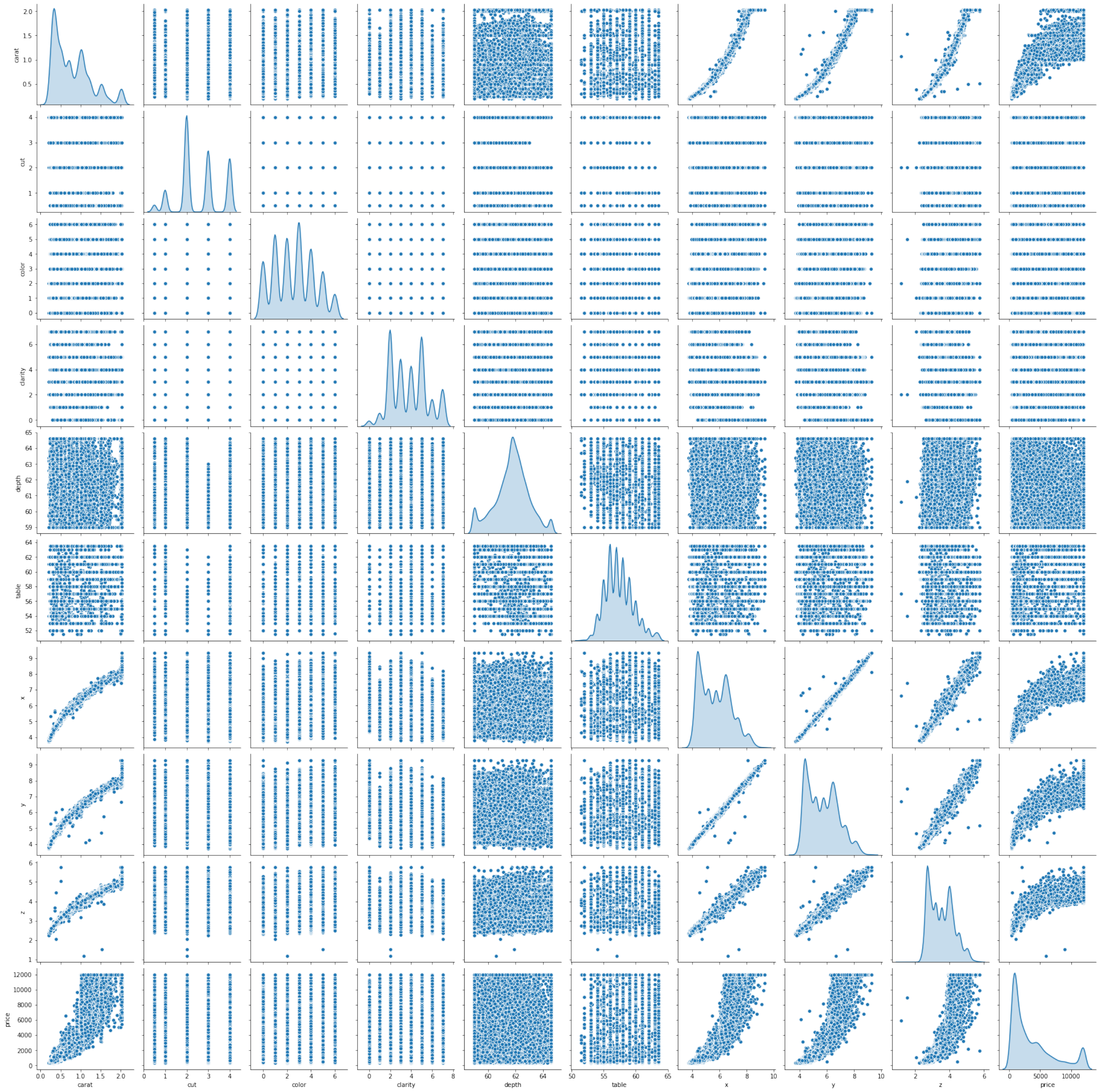
- There are outliers in the data below diagram shows that clearly that price has lot of outliers



- To fix this outliers we are going to calculate the Interquartile range and assign those values to outliers, After fixing the outliers the box plot looks like this



- On analysing the distribution of data with pair plot



- We can clearly see length, width and height are closely distributed with carat and also with price
- We can see the X,Y and Z has values as zero there are 9 records like that we will remove them from dataset.
- Whereas **carat** has values close to *ZERO* but we cannot remove as they are important for the data set

Linear Regression:

- We will encode the CUT, COLOR and CLARITY variables with pandas Categorical Encoder
- Let's split the data 70:30 ratio and 70% will be our training data and 30% will be our test data

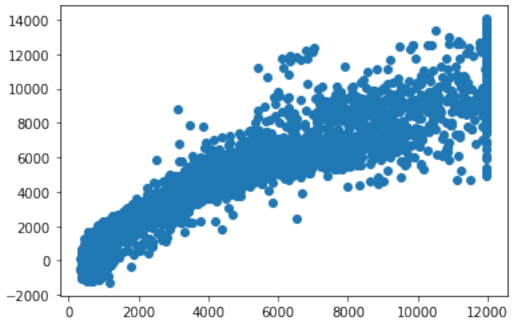
- Let's find the coefficient of our independent variables

```
The coefficient for carat 9140.743990834615
The coefficient for cut 47.4175585836999
The coefficient for color -228.5342863591243
The coefficient for clarity 252.69588872936308
The coefficient for depth -85.16472603152953
The coefficient for table -72.65710002164471
The coefficient for x -1943.3434985964316
The coefficient for y 1508.171116031764
The coefficient for z -357.2460002161683
```

- Post applying the Linear Regression algorithm let's take some metrics
  - *The R square value for training data is "**0.9096509837813009**"*
  - *The R square value for testing data is "**0.9130281960820806**"*
  - *The RMSE on training data is "**1046.8423105652398**"*
  - *The RMSE on testing data is "**1028.4320460001943**"*
  - The OLS Regression Result looks like below

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.910
Model:                            OLS   Adj. R-squared:                  0.910
Method:                 Least Squares   F-statistic:                 3.012e+04
Date:                Fri, 03 Jul 2020   Prob (F-statistic):               0.00
Time:                        23:49:06   Log-Likelihood:            -2.2538e+05
No. Observations:               26933   AIC:                         4.508e+05
Df Residuals:                   26923   BIC:                         4.509e+05
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   9175.6371    609.853     15.046      0.000    7980.294    1.04e+04
carat       9140.7440     77.274    118.290      0.000    8989.284    9292.204
cut           47.4176      6.403      7.406      0.000      34.868      59.967
color       -228.5343      3.908    -58.477      0.000    -236.194    -220.874
clarity      252.6959      3.807     66.377      0.000     245.234     260.158
depth        -85.1647      8.399    -10.140      0.000    -101.627     -68.703
table        -72.6571      3.202    -22.690      0.000     -78.934     -66.381
x          -1943.3435    110.616    -17.568      0.000   -2160.157   -1726.530
y           1508.1711    109.265     13.803      0.000    1294.006    1722.336
z           -357.2460     93.365     -3.826      0.000    -540.247    -174.245
==============================================================================
Omnibus:                     6851.055   Durbin-Watson:                   2.016
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            35171.111
Skew:                           1.136   Prob(JB):                         0.00
Kurtosis:                       8.117   Cond. No.                     8.21e+03
==============================================================================
```

- Once the data is predicted the scatter plot for the predicted item will be looking like this



- The final Linear regression formula will be

price = b0 + b1 * carat + b2 * cut + b3 * color + b4 * clarity + b5 * depth + b6 * table + b7 * x + b8 * y + b9 * z

price = (9175.64) * Intercept + (9140.74) * carat + (47.42) * cut + (-228.53) * color + (252.7) * clarity + (-85.16) * depth + (-72.66) * table + (-1943.34) * x + (1508.17) * y + (-357.25) * z

Conclusion:

- With the above approach we came to the inference that

- When carat increases by 1 unit, price increases by 9140.74 units, keeping all other predictors constant.
  similarly, when clarity increases by 1 unit, price increases by 252.7 units, keeping all other predictors constant.

- There are also some negative co-efficient values, for instance, color has its corresponding co-efficient as -228.53. This implies, when the color is different, the price decreases by 228.53 units, keeping all other predictors constant.
- The attributes which play a vital role in pricing are
  - **Carat** (when carat increases price increases)
  - **Length** (Length of the stone increase price decreases)
  - **Width** (Width of the stone increases price increases)
  - **Clarity** (Clarity of the stone is good price increases)
  - **Color** (Colour of the stone increases price decreases meaning stone should be as colourless as possible)

Problem Statement:

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

Data Dictionary:

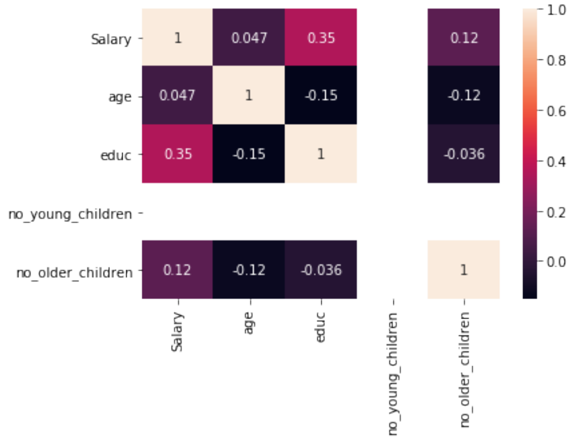
Variable Name	Description
Holiday_Package	Opted for Holiday Package yes/no?
Salary	Employee salary
age	Age in years
edu	Years of formal education
no_young_children	The number of young children (younger than 7 years)
no_older_children	Number of older children
foreign	foreigner Yes/No

Exploratory Data Analysis:

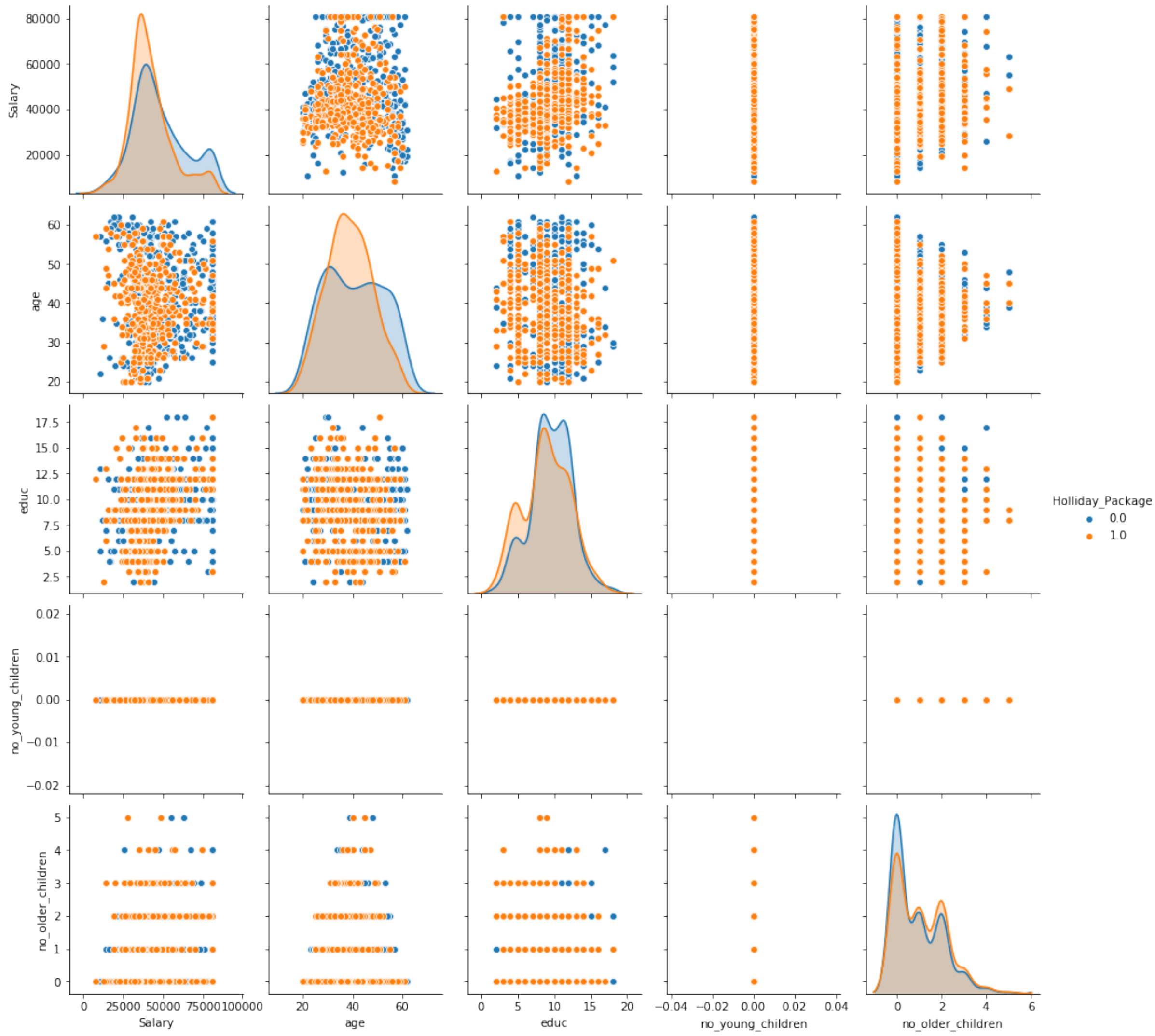
- The dataset has 872 rows and 8 columns. Holiday Package being the dependant variable we need to train our model to predict the value.
- In the 8 columns there is a index column named ***Unnamed: 0*** we don't need that column so we are dropping it.
- The proportion people taking the holiday package and not are 54:45
- The Holiday Package and foreign variables are categorical variable
- There are no duplicate records in our dataset
- There are no null values in our dataset
- The description of the data is as follows

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

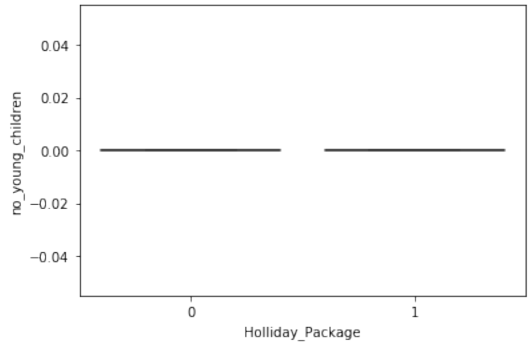
- The data is not correlated we can see that from the heat map

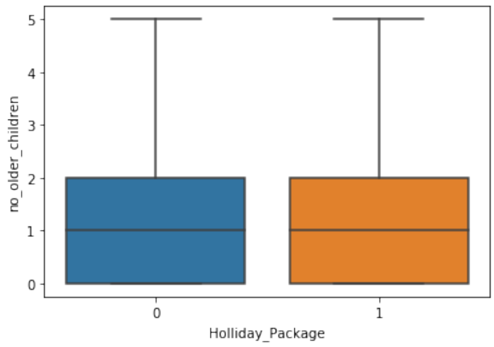


- The pair plot helps us to identify the distribution of the data

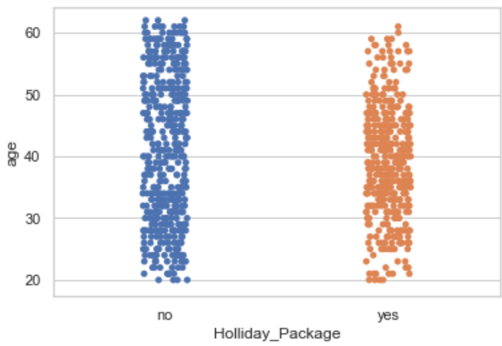


- We can able to identify the no of young children and older children are not affecting the dataset so we can remove that. To support that more lets draw a box plot for Young children vs Holiday package and same with old children

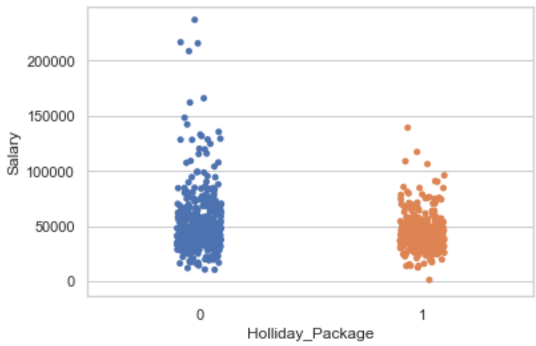




- We can clearly see the mean are same for both who opted for the package and those who don't so these 2 values not going to make impact in our prediction
- When we do a bivariate analysis for Age and Holiday Package we can see people between age 30 and 50 are choosing more holiday packages.



- People who are earning more than 150K are not opting for Holiday Packages



- Let's Split the data 70:30 70% being our training data and 30% being our Test data and apply Logistic Regression and Linear Discriminant Analysis techniques

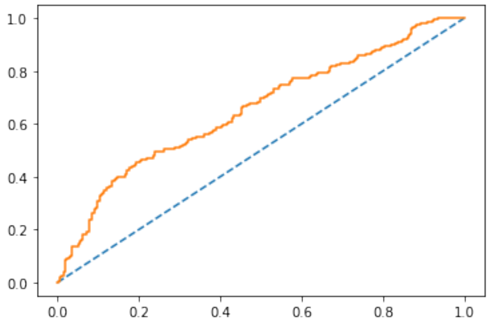
- The metrics for Logistic Regression is
    - The AUC score of train data is 0.65
    - The AUC score of test data is 0.65
    - The Classification Report for the train data is

	precision	recall	f1-score	support
0	0.62	0.83	0.71	326
1	0.68	0.42	0.52	284
accuracy			0.64	610
macro avg	0.65	0.62	0.61	610
weighted avg	0.65	0.64	0.62	610

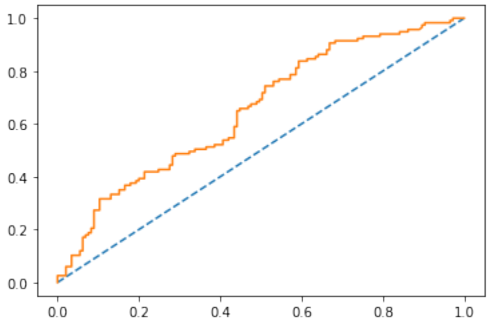
- The Classification Report for Test data is

	precision	recall	f1-score	support
0	0.62	0.75	0.68	145
1	0.58	0.42	0.49	117
accuracy			0.60	262
macro avg	0.60	0.59	0.58	262
weighted avg	0.60	0.60	0.59	262

- The AUC curve for train data is



- The AUC curve for Test data is



- We can say the model did not perform well because the accuracy precision and recall values are poor for both train and test data.

- The metrics for LDA are
    - The classification report of the predicted data looks like

	precision	recall	f1-score	support
0	0.66	0.76	0.71	471
1	0.66	0.54	0.60	401
accuracy			0.66	872
macro avg	0.66	0.65	0.65	872
weighted avg	0.66	0.66	0.66	872

By comparing both the models LDA performed slightly better than logistic regression model as the Precision, Accuracy, Recall,f1-score are all better in LDA compared to Logistic Regression Model.

Conclusion:

With the above modelling we can able to tell that

- If the employee is a foreigner there is a high chance he may opt for Holiday Package
- If employee is educated for more than 17.5 years there is a less chance he will opt for Holiday package
- If the employee age is between 30 and 50 there is a good chance he will opt for Holiday package
- If the employee is earning less than 150K there is a high chance he will opt for Holiday Package.