PREDICTIVE MODELING

PROJECT REPORT

Contents

Problem 1 (Linear Regression)

Problem 2(Logistic Regression and LDA)

Problem 1: Linear Regression

INTRODUCTION

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 DESCRIPTION

Variable Name	Description	Detail	Data Type
Carat	Weight of the cubic zirconia	Carat	Numeric
Cut	Describe cut quality of the cubic zirconia	Quality in increasing order Fair, Good, Very Good, Premium, Ideal	Categorical (Ordinal)
Colour	Colour of the cubic zirconia	D being the worst and J the best	Categorical (Ordinal)
Clarity	Cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes	(In order from Best to Worst, IF = flawless, l1= level 1inclusion) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1	Categorical (Ordinal)
Depth	The Height of a cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter		Numeric
Table	The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter		Numeric
Price	Price of the cubic zirconia	In mm	Numeric
X	Length of the cubic zirconia	In mm	Numeric
Y	Width of the cubic zirconia	In mm	Numeric
Z	Height of the cubic zirconia	In mm	Numeric

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.

Solution.

	count	mean	std	min	25%	50%	75%	max
carat	26967.0	0.798375	0.477745	0.2	0.40	0.70	1.05	4.50
depth	26270.0	61.745147	1.412860	50.8	61.00	61.80	62.50	73.60
table	26967.0	57.456080	2.232068	49.0	56.00	57.00	59.00	79.00
x	26967.0	5.729854	1.128516	0.0	4.71	5.69	6.55	10.23
y	26967.0	5.733569	1.166058	0.0	4.71	5.71	6.54	58.90
z	26967.0	3.538057	0.720624	0.0	2.90	3.52	4.04	31.80
price	26967.0	3939.518115	4024.864666	326.0	945.00	2375.00	5360.00	18818.00

Table 1: Description of the dataset

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

Table 2: Reading the Cubic Zirconia

Summary of the dataset

The data set contains 26967 row and 11 columns. In the given data set there are 2 Integer type features,6 Float type features and 3 Object type features. Where 'price' is the target variable and all other are predictor variable. The first column is an index ("Unnamed: 0") as this only serial no, we can remove it. Except for the column depth, the rest null count is 26967.

EXPLORATORY DATA ANALYSIS
Step 1: Check and remove any duplicates in the dataset
Step 2: Check and treat any missing values in the dataset
Step 3: Outlier Treatment
Step 4: Univariate Analysis
Step 5: Bi-variate Analysis

Step 1: Check and remove any duplicates in the dataset After checking for any duplicate values present in the dataset it is confirmed that there are no duplicates hence it doesn't require treatment to remove duplicates.

```
Number of duplicate rows = 0
(26925, 10)
```

Step 2: Check and treat any missing values in the dataset

```
Out[31]: carat        0
         cut          0
         color        0
         clarity      0
         depth      697
         table        0
         x            0
         y            0
         z            0
         price        0
         dtype: int64
```

Table 3: Missing values in the dataset

Step 3: Outlier Treatment Using the boxplot we confirm and visualise the presence of outliers in the dataset and then proceed to treat the outliers present.

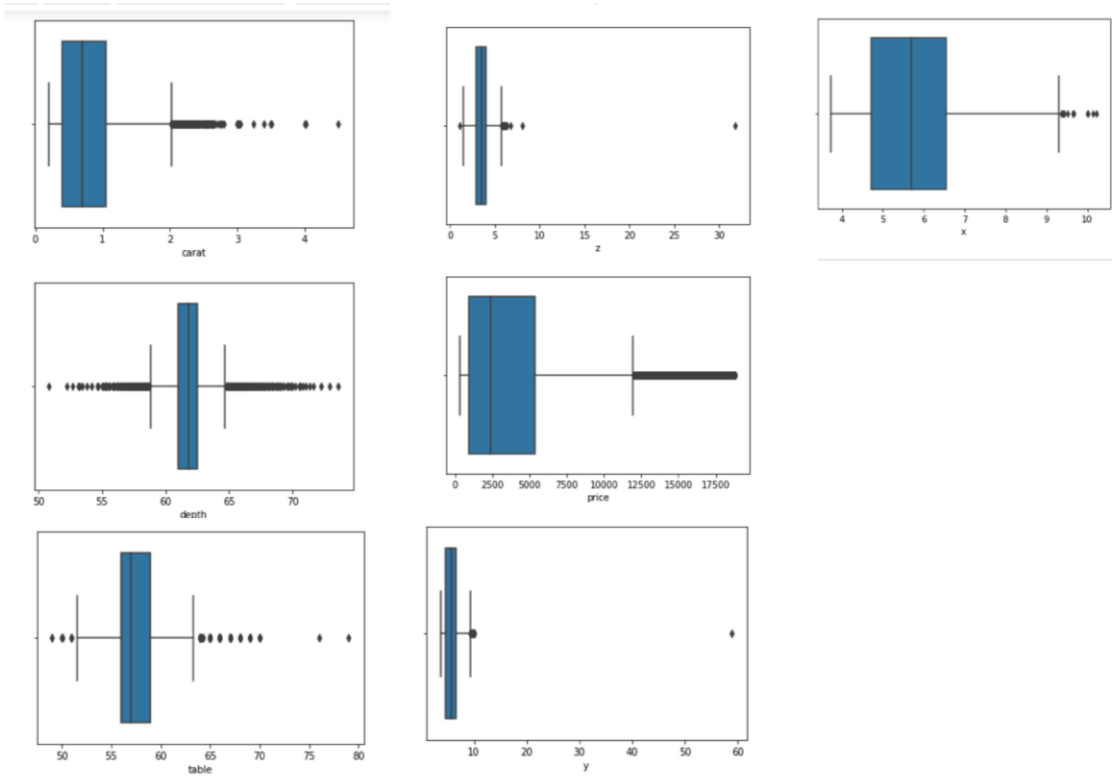


Figure 1: Data visualizing the presence of outliers

Below we see that the outliers have been treated accordingly.

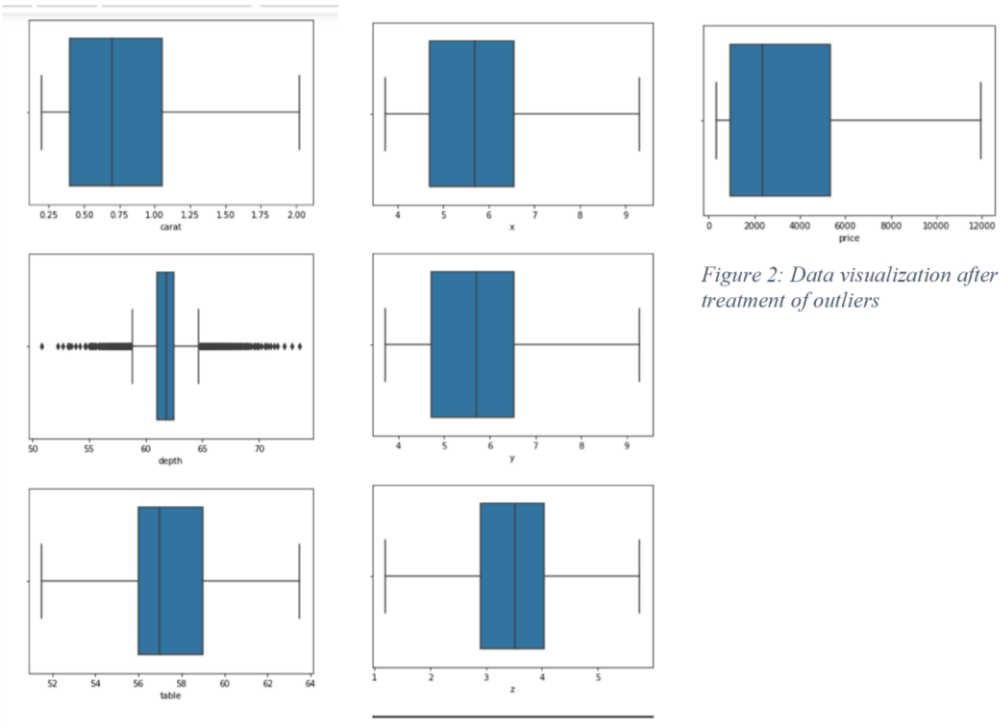
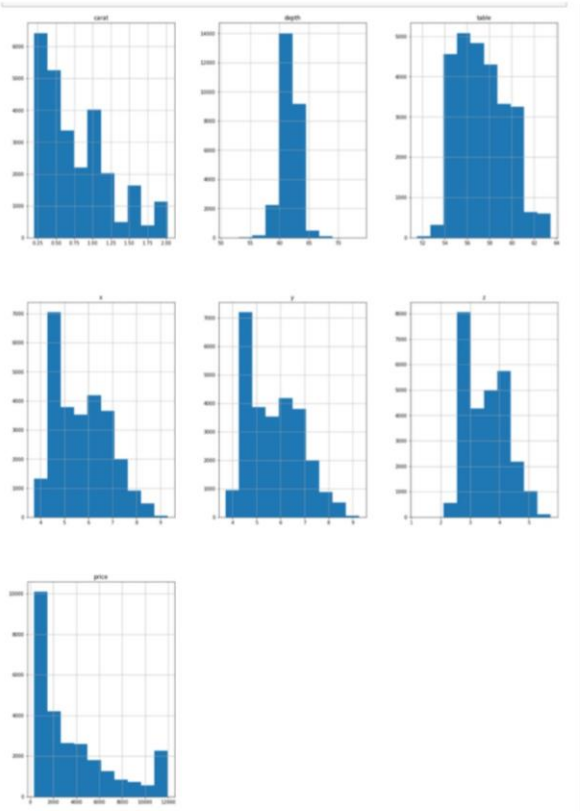


Figure 2: Data visualization after treatment of outliers

Step 4: Univariate Analysis



```
carat    0.917214
depth   -0.025042
table    0.480476
x        0.397696
y        0.394060
z        0.394819
price    1.157121
dtype: float64
```

Figure 3: Univariate Analysis

The dataset indicates that there is significant amount of outliers present in one or few of the variable and skewness is measured for every attributes present and after performing the univariate analysis we can notice that the distribution of some quantitative features like "Carat" and the target feature “Price” are heavily "right-skewed".

Step 5: Bi-variate Analysis

- It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them.
- It can be inferred that most features correlate with the price of Diamond. The notable exception is "depth" which has a negligible correlation (<1%).

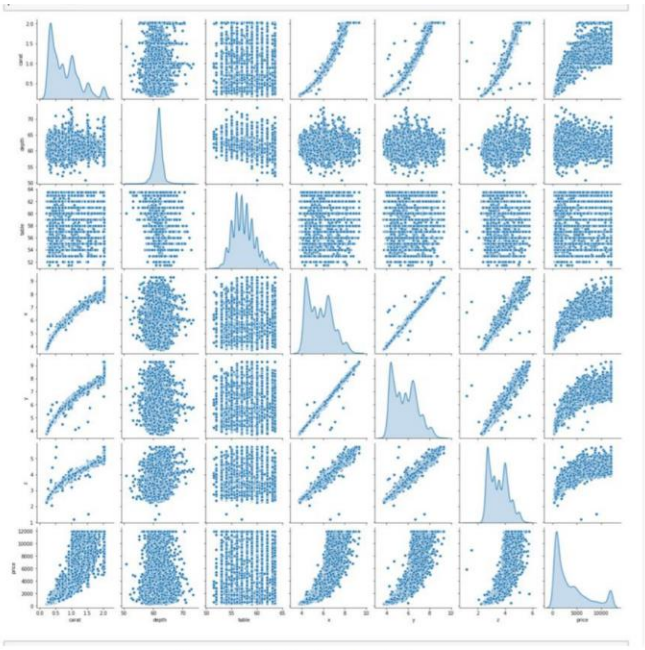


Figure 5: Bi-Variate Analysis

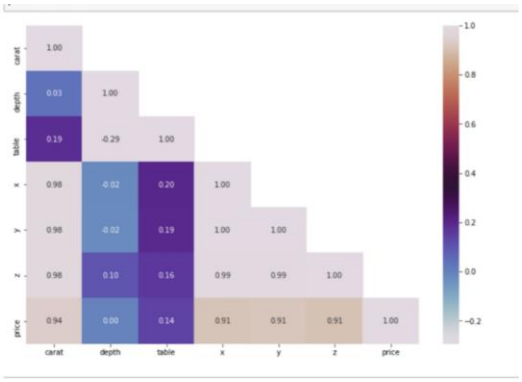


Figure 4: Heat map depicting the correlation between attributes

Above is the Correlation heatmap which helps in graphical representation of **correlation matrix** representing correlation between different variables.

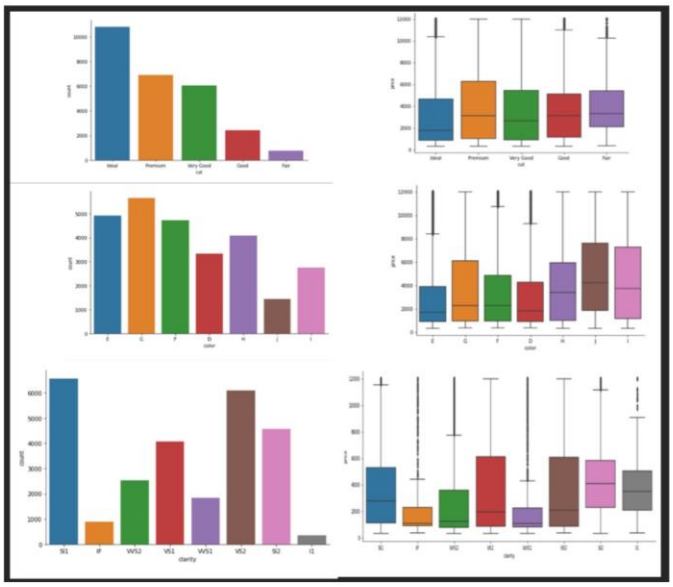


Figure 6: EDA on categorical columns

OBSERVATIONS BASED ON EDA

The inferences drawn from the above Exploratory Data analysis:

Observation-1: 'Price' is the target variable while all others are the predictors. The data set contains 26967 row, 11 column. In the given data set there are 2 Integer type features,6 Float type features. 3 Object type features. Where 'price' is the target variable and all other are predictor variable. The first column is an index ("Unnamed: 0")as this only serial no, we can remove it.

Observation-2: On the given data set the mean and median values does not have much difference. We can observe Min value of "x", "y", "z" are zero this indicates that they are faulty values. As we know dimensionless or 2-dimensional diamonds are not possible. So

we have filter out those as it clearly faulty data entries. There are three object data type 'cut', 'colour' and 'clarity'.

Observation-3: We can observe there are 697 missing value in the depth column. There are some duplicate row present. (33 duplicate rows out of 26958). which is nearly 0.12 % of the total data. So on this case we have dropped the duplicated row.

Observation-4: There are significant amount of outlier present in some variable, the features with datapoint that are far from the rest of dataset which will affect the outcome of our regression model. So we have treat the outlier. We can see that the distribution of some quantitative features like "carat" and the target feature "price" are heavily "right-skewed".

Observation-5: It looks like most features do correlate with the price of Diamond. The notable exception is "depth" which has a negligible correlation (r-s1%). Observation on 'CUT': The Premium Cut on Diamonds are the most Expensive, followed by Very Good Cut.

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.

Solution.

- We start by checking through the dataset for any null values that are present as seen in Figure 8, it shows that there are a total of 697 null values in the depth column.
- Followed by which the median is computed for each attribute so that it can be used to replace the null values that are present in the dataset.
- In below given figure 9 we can see that the null values are replaced by the median that's computed.
- After the removing the null values the shape of the dataset becomes 26925 rows and 10 columns.

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

Figure 8: List of null values present in the dataset

```
carat        0.70
depth       61.80
table       57.00
x            5.69
y            5.70
z            3.52
price     2373.00
dtype: float64
```

Figure 7: Computing the median of the attributes

	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499.0
1	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984.0
2	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289.0
3	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082.0
4	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779.0
...	...	...	...	...	...	...	...	...	...	...
26962	1.11	Premium	G	SI1	62.3	58.0	6.61	6.52	4.09	5408.0
26963	0.33	Ideal	H	IF	61.9	55.0	4.44	4.42	2.74	1114.0
26964	0.51	Premium	E	VS2	61.7	58.0	5.12	5.15	3.17	1656.0
26965	0.27	Very Good	F	VVS2	61.8	56.0	4.19	4.20	2.60	682.0
26966	1.25	Premium	J	SI1	62.0	58.0	6.90	6.88	4.27	5166.0

Figure 9: Imputing median values into the null variables

Is scaling necessary in this case?

No, it is not necessary, we'll get an equivalent solution whether we apply some kind of linear scaling or not. But is recommended for regression techniques as well because it would help gradient descent to converge fast and reach the global minima. When number of features becomes large, it helps in running model quickly else the starting point would be very far from minima, if the scaling is not done in pre-processing.

For now we will process the model without scaling and later we will check the output with scaled data of regression model output.

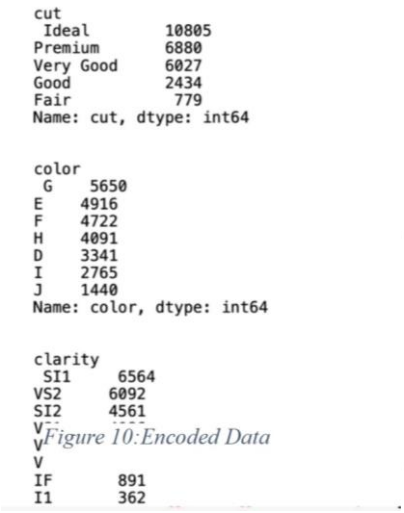
1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train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

Solution.

Train-Test Split:

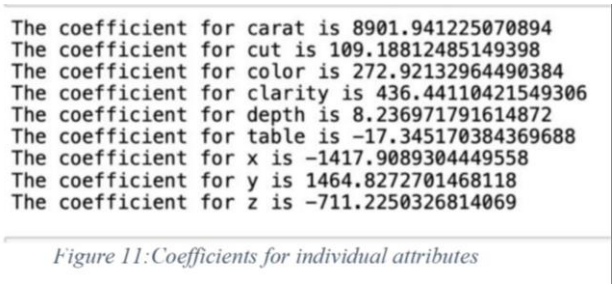
- Copy all the predictor variables into X data frame and copy target into the y data frame. Using the dependent variable we split the X and Y data frames into training set and test set.
- For this we use the Sklearn package and then split X and Y in 70:30 ration and then invoke the linear regression function and find the best fit model on training data.
- The intercept for our model is -3171.9504473076336.

- The intercept (often labelled the constant) is the expected mean values of Y when x=0,and when X is not equal to zero then the intercept has no intrinsic meaning.
- In the present case when the other predictor variable is zero i.e., like carat, cut, color, clarity then C=-3172 (Y = m/X/ m2X2+……+ mnXn + C+e), which means that the price is -3172 which doesn't make any sense so in order to deal with this we have to carry out z-score and make it nearly zero.



```
cut
 Ideal        10805
Premium       6880
Very Good     6027
Good          2434
Fair           779
Name: cut, dtype: int64

color
 G    5650
E    4916
F    4722
H    4091
D    3341
I    2765
J    1440
Name: color, dtype: int64

clarity
 SI1     6564
VS2     6092
SI2     4561
V
V
V
IF       891
I1       362
```

Figure 10:Encoded Data



```
The coefficient for carat is 8901.941225070894
The coefficient for cut is 109.18812485149398
The coefficient for color is 272.92132964490384
The coefficient for clarity is 436.44110421549306
The coefficient for depth is 8.236971791614872
The coefficient for table is -17.345170384369688
The coefficient for x is -1417.9089304449558
The coefficient for y is 1464.8272701468118
The coefficient for z is -711.2250326814069
```

Figure 11:Coefficients for individual attributes

R square on training data : 0.9311935886926559

R square on testing data : 0.931543712584074

- R square is the percentage of the response variable variation that is explained by a linear model and computed by the formula as:

R-square = Explained Variation / Total Variation

- It is always between 0 and 100%, in which 0% indicates that the model explains none of the variability of the response data around its mean and 100% indicates that the model explains all the variability of the response data around its mean.
- In the regression model we can see the R-square value on training and test data respectively as 0.9311935886926559 and - 0.931543712584074.
- The RMSE on training and test data respectively is 907.1312415459143 and 911.8447345328437.
- From the scatter plot, we see that it is a linear and there is very strong correlation present between the predicted y and actual y.
- It also indicates that there's a lot spread which indicates some unexplained variances on the output.
- As the training data & Test data score are almost inline we can conclude that this model is a Right-Fit model.

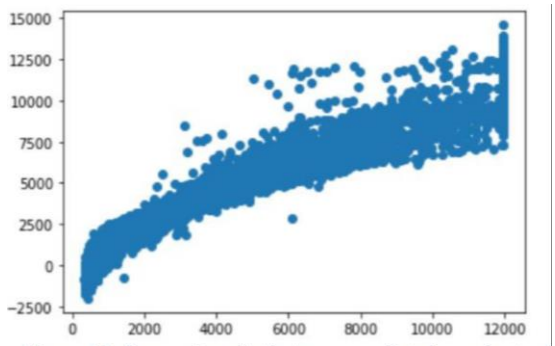


Figure 12: Regression plot between predicted y and actual y

	Training Data	Test Data
R-square	0.9311935886926559	0.931543712584074
RMSE	907.1312415459143	911.8447345328436

Applying z- score stats models

- We initiate the linear Regression function and find the best fit model on the training data and then explore the coefficients for each of the attributes.

```
The coefficient for carat is 1.1837737061779416
The coefficient for cut is 0.03512500065529705
The coefficient for color is 0.1344926928764153
The coefficient for clarity is 0.2080977932562189
The coefficient for depth is 0.003326293718838837
The coefficient for table is -0.010815851633643268
The coefficient for x is -0.459689842412529
The coefficient for y is 0.4716627091792431
The coefficient for z is -0.14249737973827056
```
Figure 13:coefficients of individual attributes after applying z-score

- The intercept for our model is -5.879615251304736e-16 and the co-efficient of determinant is 0.9315051288558229.
- It's observed that by applying z score the intercept has changed from -3171.950447307667 to 5.87961525130473e-16, which tells that the co-efficient has changed and the bias has become nearly zero but the overall accuracy is still the same.

Check Multi-collinearity using VIF

• We can observe very strong multi collinearity present in the data set when ideally it should be within 1 to 5.

```
carat = 121.96543302739589
cut = 10.388738909800333
color = 5.546407587131623
clarity = 5.455999699082339
depth = 1218.3824913329145
table = 878.3985698779234
x = 10744.05623520385
y = 9482.053091580401
z = 3697.5688286012546
```
Figure 14:Checking for Multi-collinearity

Linear Regression using stats models

- Assuming the null hypothesis is true, i.e. price from that universe we have drawn co-efficient for the variable shown above.
- Now we can ask what is the probability of finding this co-efficient in this drawn sample if in the real world the co-efficient is zero. As we see here the overall P value is less than alpha, so rejecting HO and accepting Ha that at least 1 regression co-efficient is not '0'. Here all regression co-efficient are not '0'.
- For example, we can see the p value is showing 0.449 for 'depth' variable, which is much higher than 0.05. That means this dimension is of no use. So we can say that the attribute which are having p value greater than 0.05 are poor predictor for price.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.931
Model:	OLS	Adj. R-squared:	0.931
Method:	Least Squares	F-statistic:	2.833e+04
Date:	Mon, 25 Oct 2021	Prob (F-statistic):	0.00
Time:	15:33:31	Log-Likelihood:	-1.5510e+05
No. Observations:	18847	AIC:	3.102e+05
Df Residuals:	18837	BIC:	3.103e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3171.9504	787.532	-4.028	0.000	-4715.583	-1628.318
carat	8901.9412	82.792	107.521	0.000	8739.661	9064.222
cut	109.1881	7.268	15.024	0.000	94.943	123.433
color	272.9213	4.105	66.478	0.000	264.874	280.968
clarity	436.4411	4.473	97.581	0.000	427.674	445.208
depth	8.2370	10.876	0.757	0.449	-13.080	29.554
table	-17.3452	3.904	-4.443	0.000	-24.998	-9.693
x	-1417.9089	136.590	-10.381	0.000	-1685.637	-1150.181
y	1464.8273	136.068	10.765	0.000	1198.122	1731.533
z	-711.2250	156.187	-4.554	0.000	-1017.366	-405.084

Omnibus:	2652.028	Durbin-Watson:	2.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9642.429
Skew:	0.687	Prob(JB):	0.00
Kurtosis:	6.223	Cond. No.	1.03e+04

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

Figure 15:Regression table

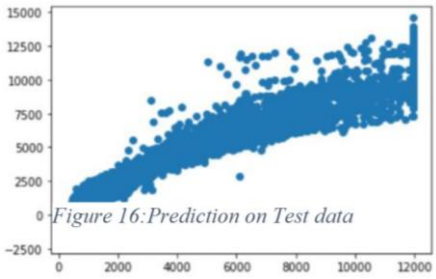


Figure 16:Prediction on Test data

Root Mean Squared Error (Training) ------RMSE: 907.1312415459133

Root Mean Squared Error (lest) ----------RMSE: 911.8447345328433

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

Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

Solution.

Inference:

We can see that the from the linear plot, very strong corelation between the predicted y and actual y. But there are lots of spread. That indicates some kind noise present on the data set i.e. Unexplained variances on the output.

Linear regression Performance Metrics:

Intercept for the model: -3171.950447307667 R square on training data: 0.9311935886926559 R square on testing data: 0.931543712584074 RMSE on Training data: 907.1312415459143 RMSE on Testing data: 911.8447345328436 As the training data & testing data score are almost inline, we can conclude this model is a Right-Fit Model.

Impact of scaling:

We can observe by applying z score the intercept became -5.87961525130473e-16. Earlier it was -3171.950447307667. the co-efficient has changed, the bias became nearly zero but the overall accuracy still same.

Multi collinearity: We can observe there are very strong multi collinearity present in the data set.

From statsmodels: we can see R-squared:0.931 and Adj. R-squared: 0.931 are same. The overall P value is less than alpha.

- Finally we can conclude that Best 5 attributes that are most important are 'Carat', 'Cut', 'colour', clarity' and width i.e. 'y' for predicting the price.
- When 'carat' increases by 1 unit, diamond price increases by 8901.94 units, keeping all other predictors constant.
- When 'cut' increases by 1 unit, diamond price increases by 109.19 units, keeping all other predictors constant.
- When 'colour' increases by 1 unit, diamond price increases by 272.92 units, keeping all other predictors constant.
- When 'clarity' increases by 1 unit, diamond price increases by 436.44 units, keeping all other predictors constant.
- When 'y' increases by 1 unit, diamond price increases by 1464.83 units, keeping all other predictors constant.
- We can see that the p value is 0.449 for depth variable, which is much greater than 0.05. That means this attribute is of no use.
- There are also some negative co-efficient values, we can see the 'X' i.e Length of the cubic zirconia in mm. having negative co-efficient -1417.9089. And the p value is less than 0.05, so can conclude that as higher the length of the stone is a lower profitable stones.
- Similarly for the 'z' variable having negative co-efficient i.e. -711.23. And the p value is less than 0.05, so we can conclude that as higher the 'z' of the stone is a lower profitable stones.

Recommendations:

- The Gem Stones company should consider the features 'Carat', 'Cut', 'colour', 'clarity' and width i.e. 'y' as most important for predicting the price. To distinguish between higher profitable stones and lower profitable stones so as to have better profit share.
- As we can see from the model Higher the widtb('y') of the stone is higher the price.

- So the stones having higher widtb('y') should consider in higher profitable stones. The 'Premium Cut' on Diamonds are the most Expensive, followed by 'Very Good' Cut, these should consider in higher profitable stones.
- The Diamonds clarity with 'VS1' &'VS2' are the most expensive. So these two category also consider in higher profitable stones.
- As we see for 'X' i.e. Length. of the stone, higher the length of the stone is lower the price.
- So higher the Length('x') of the stone are lower is the profitabilim higher the 'z' i.e Height of the stone is, lower the price. This is because if a Diamond's Height is too large Diamond will become 'Dark' in appearance because it will no longer return an Attractive amount of light. That is why.
- Stones with higher 'z' is also are lower in profitability.

Problem 2: Logistic Regression and LDA

INTRODUCTION

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

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.

Solution.

Here I am loading all the necessary library for the model building and reading the head and tail of the dataset to check whether data has been properly fed.

HEAD OF THE DATA

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

Table 18

TAIL OF THE DATA

	Unnamed: 0	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
862	863	no	66900	35	10	1	1	yes
863	864	no	35290	51	9	0	1	yes
864	865	no	25527	41	5	1	0	yes
865	866	yes	44057	35	9	0	2	yes
866	867	yes	22643	42	14	0	0	yes
867	868	no	40030	24	4	2	1	yes
868	869	yes	32137	48	8	0	0	yes
869	870	no	25178	24	6	2	0	yes
870	871	yes	55958	41	10	0	1	yes
871	872	no	74659	51	10	0	0	yes

Table 19

SHAPE OF THE DATASET

(872, 8)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Unnamed: 0         872 non-null    int64
 1   Holliday_Package   872 non-null    object
 2   Salary             872 non-null    int64
 3   age                872 non-null    int64
 4   educ               872 non-null    int64
 5   no_young_children  872 non-null    int64
 6   no_older_children  872 non-null    int64
 7   foreign            872 non-null    object
dtypes: int64(6), object(2)
memory usage: 54.6+ KB
```

Table 20

- We have no null values in the dataset.
- We have integer and object data.

DESCRIBE

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Unnamed: 0	872.0	NaN	NaN	NaN	436.5	251.869014	1.0	218.75	436.5	654.25	872.0
Holliday_Package	872	2	no	471	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	872.0	NaN	NaN	NaN	47729.172018	23418.668531	1322.0	35324.0	41903.5	53469.5	236961.0
age	872.0	NaN	NaN	NaN	39.955275	10.551675	20.0	32.0	39.0	48.0	62.0
educ	872.0	NaN	NaN	NaN	9.307339	3.036259	1.0	8.0	9.0	12.0	21.0
no_young_children	872.0	NaN	NaN	NaN	0.311927	0.61287	0.0	0.0	0.0	0.0	3.0
no_older_children	872.0	NaN	NaN	NaN	0.982798	1.086786	0.0	0.0	1.0	2.0	6.0
foreign	872	2	no	656	NaN	NaN	NaN	NaN	NaN	NaN	NaN

The data that we have is of integer and continuous data, here the holiday package is our target variable .

Salary, age, educ and number young children, number older children of employee have the went to foreign, those are the given attributes we have to cross examine and help the company predict weather the person will opt for holiday package or not.

NULL VALUES

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

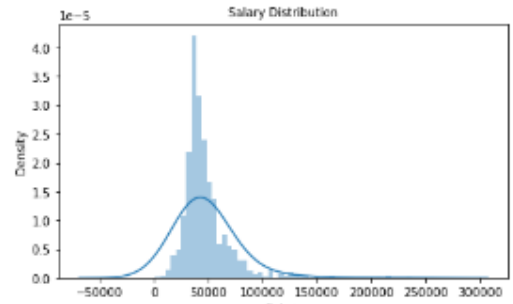
There are no null values in the dataset
CHECK FOR DUPLICATES IN THE GIVEN DATASET
Number of duplicate rows = 0

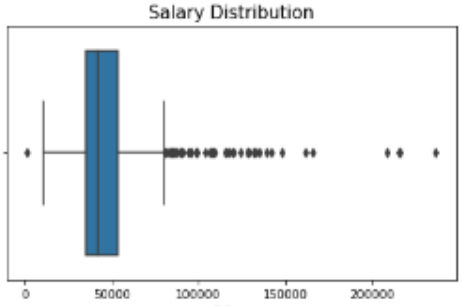
Unique values for categorical variables

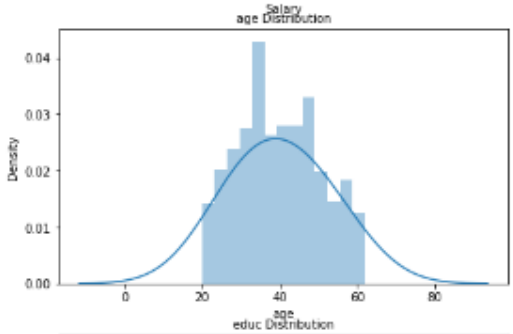
```
HOLLIDAY_PACKAGE :  2
yes    401
no     471
Name: Holliday_Package, dtype: int64


FOREIGN :  2
yes    216
no     656
Name: foreign, dtype: int64
```

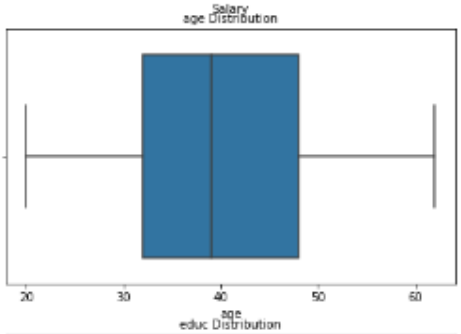
Percentage of employees that are interested in the holiday package 45.9%

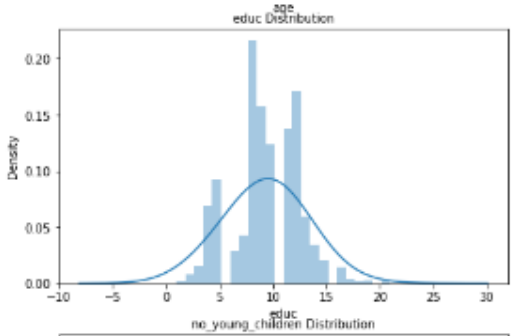
UNIVARIATE ANALYSIS

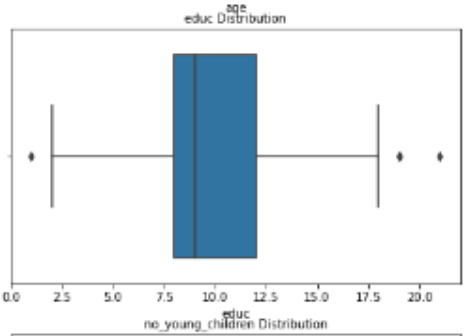


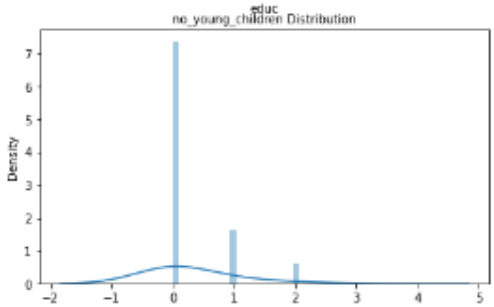


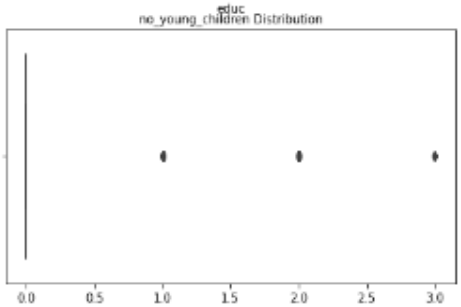


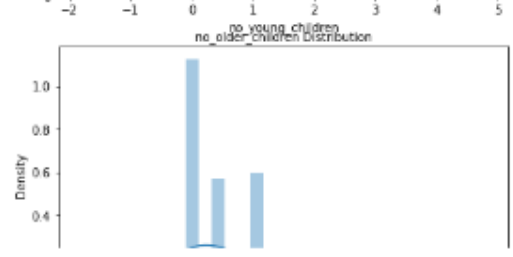


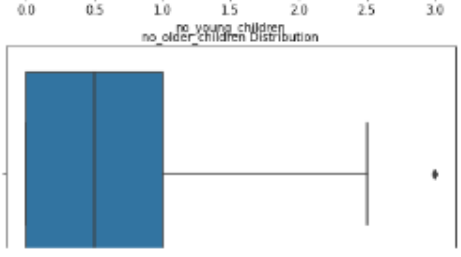












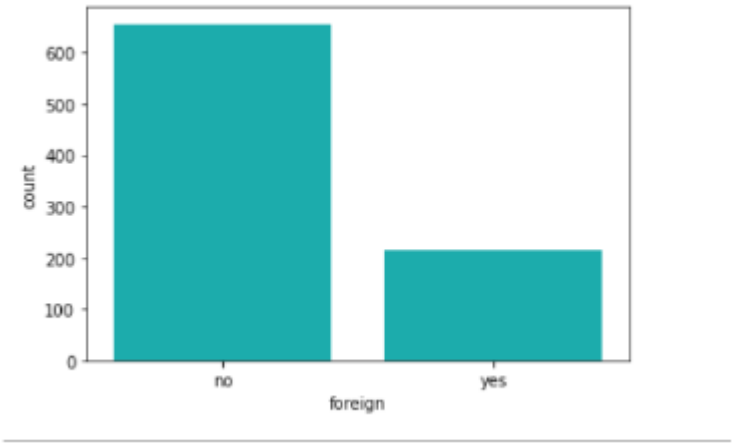
SKEWNESS

Unnamed: 0	0.000000
Salary	3.103216
age	0.146412
educ	-0.045501
no_young_children	1.946515
no_older_children	0.953951

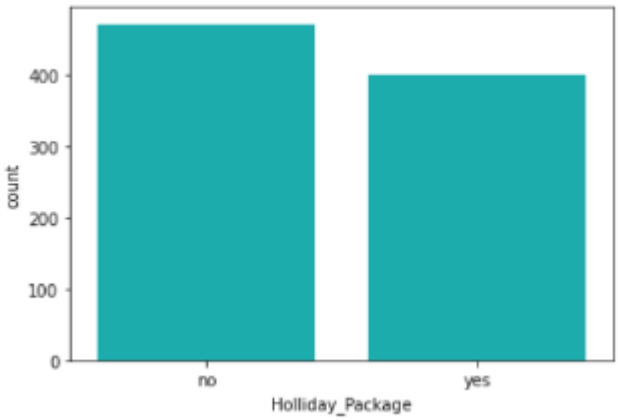
dtype: float64

- We can see that most of the distribution are right skewe except for educ
- Salary distribution has the max no of outliers
- There are some outliers in educ , no of young children and no. of older children

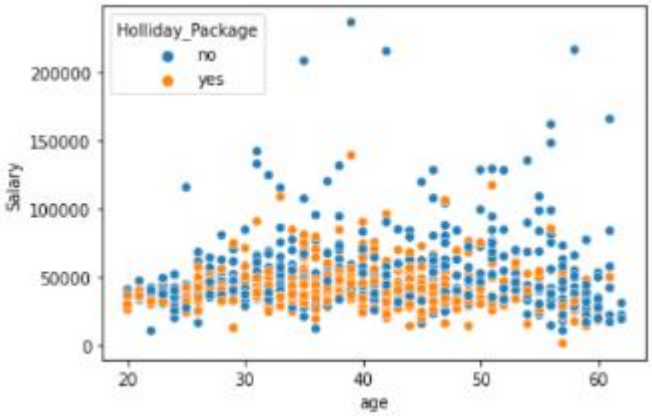
CATOGORICAL UNIVARIATE ANALYSIS

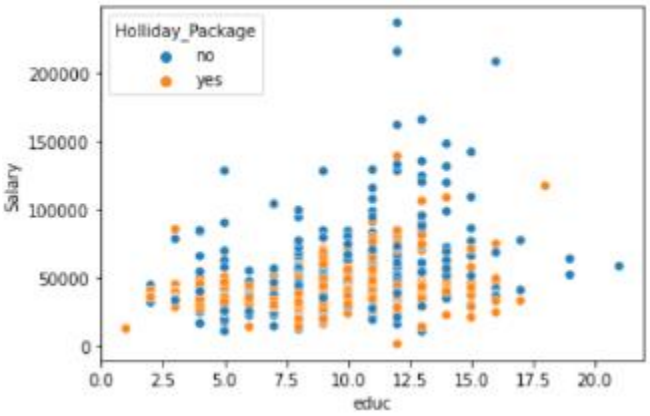


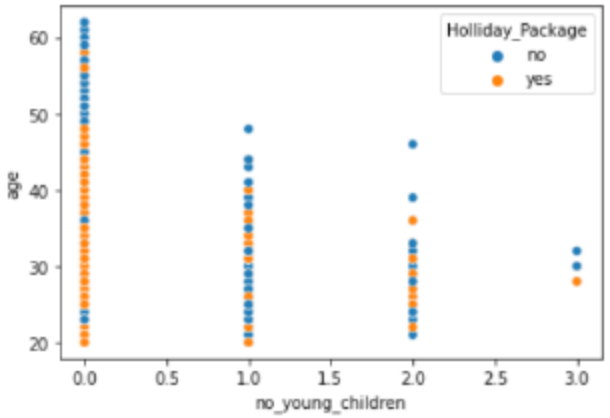
Maximum of the employees don't prefer to go to foreign

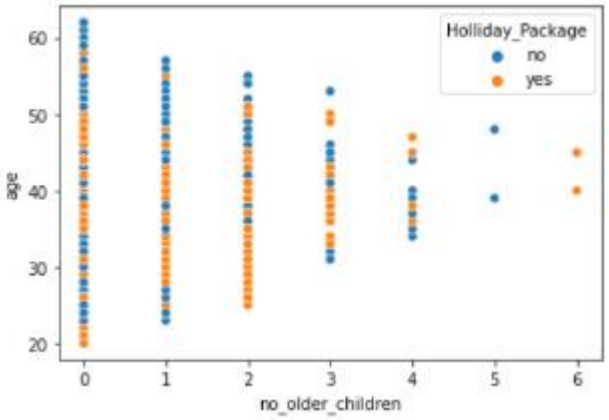


The employees who prefer holiday package are slightly less than who don't.


Holliday_Package
no
yes
Salary
200000
150000
100000
50000
0
20
30
40
50
60
age

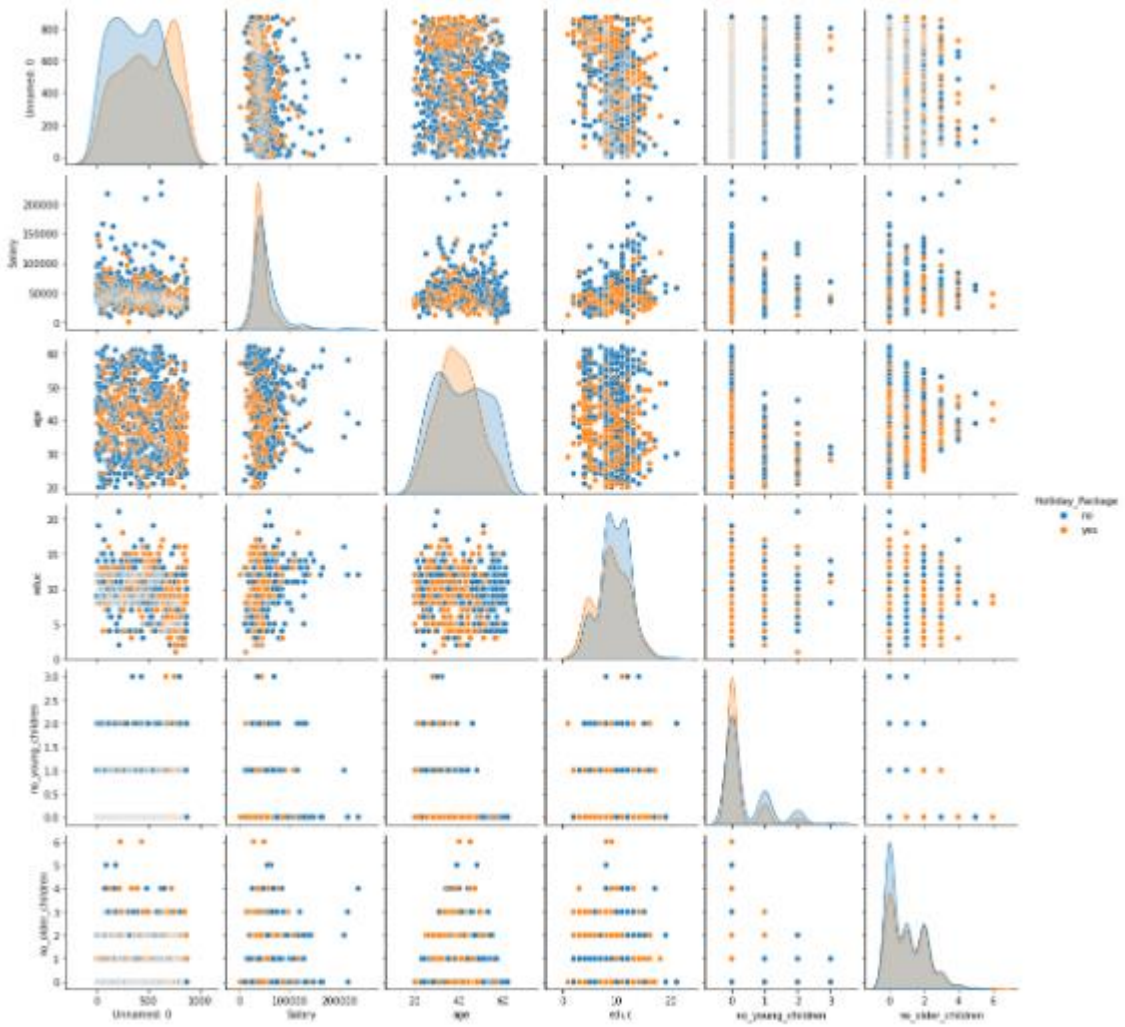

Holliday_Package
no
yes
Salary
200000
150000
100000
50000
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
educ


Holliday_Package
no
yes
age
60
50
40
30
20
0.0
0.5
1.0
1.5
2.0
2.5
3.0
no_young_children


Holliday_Package
no
yes
age
60
50
40
30
20
0
1
2
3
4
5
6
no_older_children

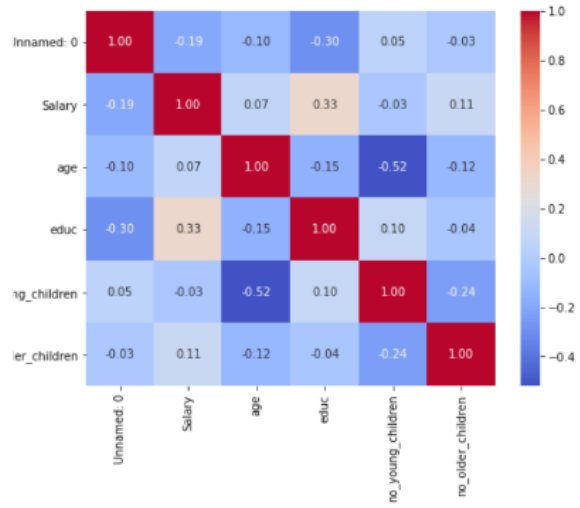
- As we can observe people with salaries below 150000 prefer holiday package.
- Employee age over 50 to 60 have seems to be not taking the holiday package, whereas in the age 30 to 50 and salary less than 50000 people have opted more for holiday package

BIVARITE ANALYIS
DATA DISTRIBUTION



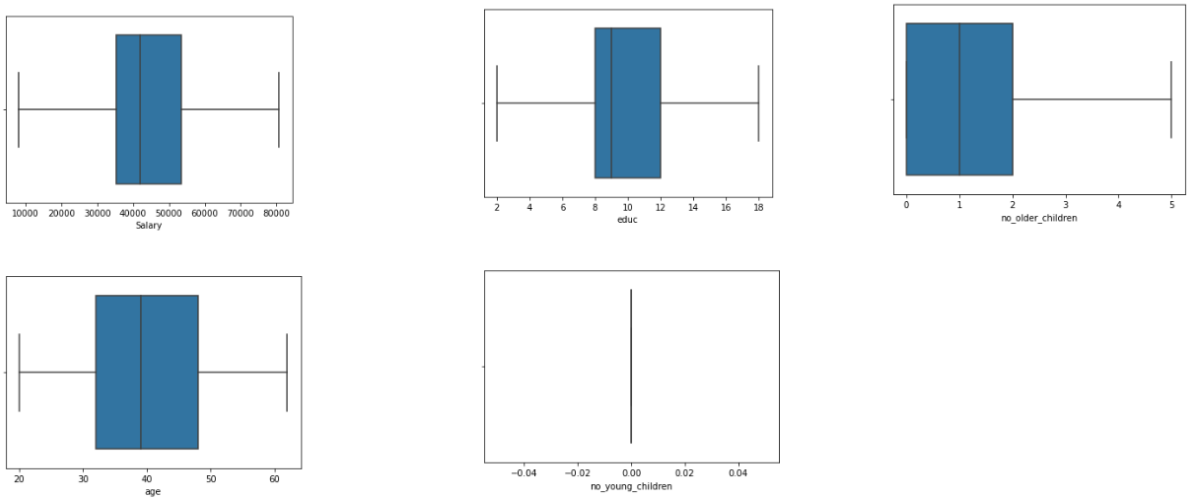
There is hardly any correlation between the data, the data seems to be normal. There is no huge difference in the data distribution among the holiday package, I don't see any clear two different distributions in the dataset provided.

CHECKING FOR CORRELATION



There is hardly any correlation between the data so no collinearity

1. AFTER TREATING OUTLIERS DATA LOOKS LIKE THIS



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).

Solution.

Encoding the data(having string variables)

Head of the dataset

	Salary	age	educ	no_young_children	no_older_children	Holliday_Package_yes	foreign_yes
0	48412.0	30.0	8.0	0.0	1.0	0	0
1	37207.0	45.0	8.0	0.0	1.0	1	0
2	58022.0	46.0	9.0	0.0	0.0	0	0
3	66503.0	31.0	11.0	0.0	0.0	0	0
4	66734.0	44.0	12.0	0.0	2.0	0	0

Here we have done ONE HOT ENCODING to create dummy variables and we can see all values for foreign_yes are 0.
Better results are predicted by logistic regression model if encoding is done.
Train/ Test split
We will split the data in 70/30 ratio

```
# Copy all the predictor variables into X dataframe
X = data.drop('Holliday_Package_yes', axis=1)

# Copy target into the y dataframe.
y = data['Holliday_Package_yes']
```

```
# Split X and y into training and test set in 70:30 ratio
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30 , random_state=1,stratify=y)
```

Applying Logistic Regression
Applying GridSearchCV for Logistic Regression

```
{'penalty': 'l2', 'solver': 'liblinear', 'tol': 1e-06}

LogisticRegression(max_iter=100000, n_jobs=2, solver='liblinear', tol=1e-06)
```

The grid search method is used for logistic regression to find the optimal solving and the parameters for solving.
We have found the parameters using grid search such as penalty=12 , solver: liblinear , tolerance=1e-06

Prediction on the training set

ytrain_predict = best_model.predict(X_train)
ytest_predict = best_model.predict(X_test)

Getting the probabilities on the test set

	0	1
0	0.636523	0.363477
1	0.576651	0.423349
2	0.650835	0.349165
3	0.568064	0.431936
4	0.536356	0.463644

Performance Metrics will be discussed in 2.3

LDA (linear discriminant analysis)
DATASET HEAD

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	no	48412.0	30.0	8.0	0.0	1.0	no
1	yes	37207.0	45.0	8.0	0.0	1.0	no
2	no	58022.0	46.0	9.0	0.0	0.0	no
3	no	66503.0	31.0	11.0	0.0	0.0	no
4	no	66734.0	44.0	12.0	0.0	2.0	no

DATASET HEAD AFTER DATA PROCESSING

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	0	48412.0	30.0	8.0	0.0	1.0	0
1	1	37207.0	45.0	8.0	0.0	1.0	0
2	0	58022.0	46.0	9.0	0.0	0.0	0
3	0	66503.0	31.0	11.0	0.0	0.0	0
4	0	66734.0	44.0	12.0	0.0	2.0	0

Build LDA Model

```
#Build LDA Model
clf = LinearDiscriminantAnalysis()
model=clf.fit(X_train,Y_train)
```

```
# Training Data Class Prediction with a cut-off value of 0.5
pred_class_train = model.predict(X_train)

# Test Data Class Prediction with a cut-off value of 0.5
pred_class_test = model.predict(X_test)
```

PROBABILITY PREDICTION

```
# Training Data Probability Prediction
pred_prob_train = model.predict_proba(X_train)

# Test Data Probability Prediction
pred_prob_test = model.predict_proba(X_test)
```

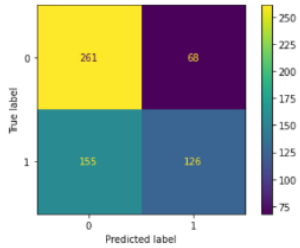
Performance Metrics will be discussed in 2.3

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.

Solution.

PEFORMANCE METRICS FOR LINEAR REGRESSION
Confusion matrix on the training data

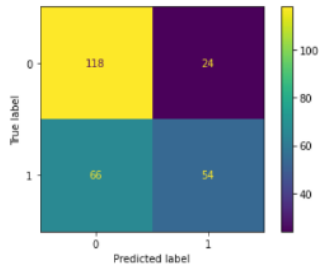
	precision	recall	f1-score	support
0	0.63	0.79	0.70	329
1	0.65	0.45	0.53	281
accuracy			0.63	610
macro avg	0.64	0.62	0.62	610
weighted avg	0.64	0.63	0.62	610



Here we see that precision for 1 is 0.63 , recall is 0.45 accuracy is 0.63 and f1 score is 0.63

Confusion matrix on the test data

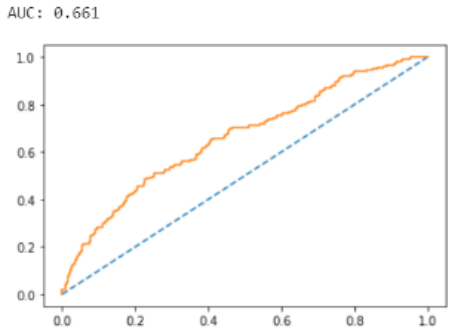
	precision	recall	f1-score	support
0	0.64	0.83	0.72	142
1	0.69	0.45	0.55	120
accuracy			0.66	262
macro avg	0.67	0.64	0.63	262
weighted avg	0.66	0.66	0.64	262



Here we see that precision for 1 is 0.69 , recall is 0.45 accuracy is 0.66 and f1 score is 0.55

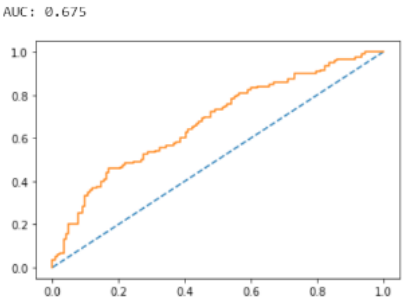
Accuracy - Training Data
0.6344262295081967

AUC and ROC for the training data



Accuracy - Test Data
0.6564885496183206

AUC and ROC for the testing data



Metrics for train data

lr_train_precision 0.65

lr_train_recall 0.45

lr_train_f1 0.53

Metrics for test data

lr_test_precision 0.69

lr_test_recall 0.45

lr_test_f1 0.55

PERFORMANCE METRICS FOR LDA(linear discriminant analysis) MODEL SCORE

0.6327868852459017

CLASSIFICATION REPORT TRAIN DATA

	precision	recall	f1-score	support
0	0.62	0.80	0.70	329
1	0.65	0.44	0.52	281
accuracy			0.63	610
macro avg	0.64	0.62	0.61	610
weighted avg	0.64	0.63	0.62	610

Here we see that precision for 1 is 0.65 , recall is 0.44 accuracy is 0.63 and f1 score is 0.52

confusion_matrix for train data
array([[263, 66],
[158, 123]]

Model score for test data
0.6564885496183206

Classification report for test data

	precision	recall	f1-score	support
0	0.64	0.83	0.72	142
1	0.69	0.45	0.55	120
accuracy			0.66	262
macro avg	0.67	0.64	0.63	262
weighted avg	0.66	0.66	0.64	262

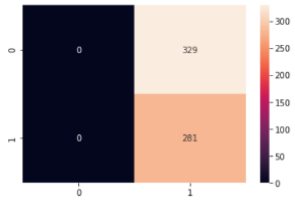
Confusion matrix for test data
array([[118, 24],
[66, 54]]

CHANGING THE CUTT OFF VALUE TO CHECK OPTIMAL VALUE THAT GIVES BETTER ACCURACY AND F1 SCORE

0.1

Accuracy Score 0.4607
F1 Score 0.6308

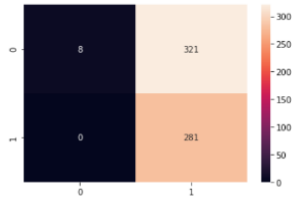
Confusion Matrix



0.2

Accuracy Score 0.4738
F1 Score 0.6365

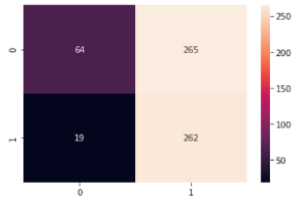
Confusion Matrix



0.3

Accuracy Score 0.5344
F1 Score 0.6485

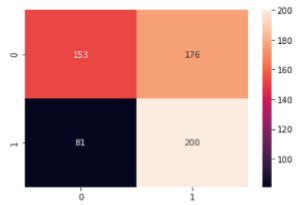
Confusion Matrix



0.4

Accuracy Score 0.5787
F1 Score 0.6088

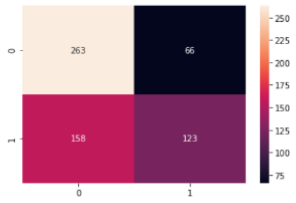
Confusion Matrix



0.5

Accuracy Score 0.6328
F1 Score 0.5234

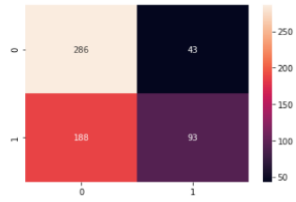
Confusion Matrix



0.6

Accuracy Score 0.6213
F1 Score 0.446

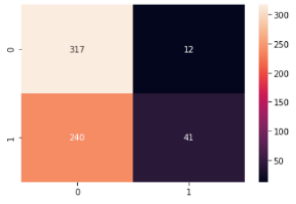
Confusion Matrix



0.7

Accuracy Score 0.5869
F1 Score 0.2455

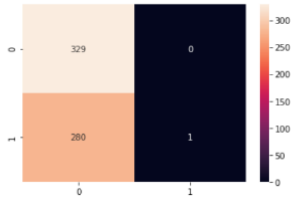
Confusion Matrix



0.8

Accuracy Score 0.541
F1 Score 0.0071

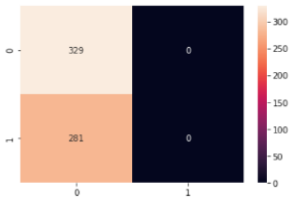
Confusion Matrix



0.9

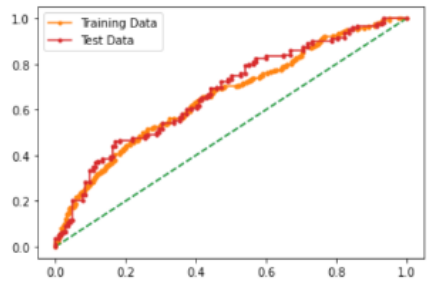
Accuracy Score 0.5393
F1 Score 0.0

Confusion Matrix



AUC and ROC for the training data

AUC for the Training Data: 0.661
AUC for the Test Data: 0.675



	LR Train	LR Test	LDA Train	LDA Test
Accuracy	0.63	0.66	0.63	0.66
AUC	0.66	0.68	0.66	0.68
Recall	0.45	0.45	0.44	0.45
Precision	0.65	0.69	0.65	0.69
F1 Score	0.53	0.55	0.52	0.55

Comparing both these models, we find both results are same, but LDA works better when there is category target variable.
As we can see the results for AUC/ROC for both the models are almost equivalent to each other So it is very difficult to differentiate between the two . The scores are also almost at par with each other . Both the models are working perfectly at par with each other.
Since LDA works better with categorical values so we will pick it in this situation.

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

Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

Solution.

So we had been given a problem where we had to find out whether the employees will opt for a holiday package or not .
We looked in the data using logistic regression and LDA.
We found out that the results using both the methods is same. Predictions were done using both the models.

While doing EDA we found out that

- Most of the employees who are above 50 don't opt for holiday packages. It seems like they are not interested in holiday packages at all .

- Employees who are in the age gap of 30 to 50 opt for holiday packages .It seems like young people believe I spending on holiday packages so age here plays a very important role in deciding whether they will opt for package or not

- Also people who have salary less than 50000 opt for holiday packages . So salary is also a deciding factor for the holiday package.

- Education also plays an important role in deciding the holiday packages .

- To improve our customer base we need to look into those factors

Recommendations

As we already have the customer base who are of the age of 30 to 50 so we need to look for the options and target the older people and the people who are earning more than 150000.
- As we know most of the people who are older prefer to visit religious places so it would be better if we target those places and provide them with packages where they can visit religious places.

- We can also look into the family dynamics of the people of the older people , if the older people have elder children e.g 30 to 40 they can use the holiday packages so the deal should include the family package .

- People who earn more than 150000 don't spend much on the holiday packages , they tend to go for lavish holidays and we can provide them with customized packages according to

their wish , such as fancy hotels , longer vacations , personal cars during the holiday to attract such employees .

- Plus such people who earn more than 150000 we can provide them extra facilities according to their own wishes at the moment.

In this project we started with EDA , descriptive statistics and did null value condition check, we performed Univariate and Bivariate Analysis. did exploratory data analysis ,we treated outliers then we moved on to Logistic regression . We encoded the data (having string values) for Modelling. We split data into train and test (70:30) and finally we applied Logistic Regression and LDA (linear discriminant analysis).