Business Report – Predictive Modelling

Problem 1: Linear Regression

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.

Head of Data

	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
1	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
2	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
4	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779

Tail of Data

	carat	cut	color	clarity	depth	table	x	y	z	price
26962	1.11	Premium	G	SI1	62.3	58.0	6.61	6.52	4.09	5408
26963	0.33	Ideal	H	IF	61.9	55.0	4.44	4.42	2.74	1114
26964	0.51	Premium	E	VS2	61.7	58.0	5.12	5.15	3.17	1656
26965	0.27	Very Good	F	VVS2	61.8	56.0	4.19	4.20	2.60	682
26966	1.25	Premium	J	SI1	62.0	58.0	6.90	6.88	4.27	5166

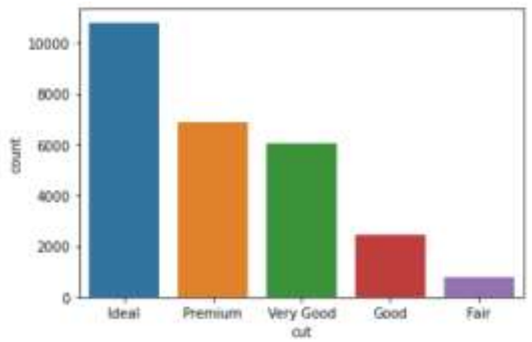
```
Total number of null values before imputing is 697

Shape of the data is (26967, 10)
```

Central tendency report

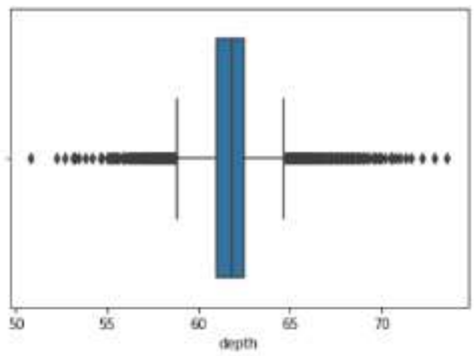
	carat	depth	table	x	y	z	price
count	2.696700e+04	26967.000000	2.696700e+04	2.696700e+04	2.696700e+04	2.696700e+04	2.696700e+04
mean	-1.614017e-16	0.002824	-2.982727e-15	5.350331e-16	-8.057238e-16	-2.124932e-16	-2.910285e-17
std	1.000019e+00	0.874109	1.000019e+00	1.000019e+00	1.000019e+00	1.000019e+00	1.000019e+00
min	-1.252522e+00	-1.969594	-3.788521e+00	-5.077427e+00	-4.917146e+00	-4.909807e+00	-8.978153e-01
25%	-8.338809e-01	-0.463659	-6.523577e-01	-9.037285e-01	-8.778193e-01	-8.854401e-01	-7.440185e-01
50%	-2.059198e-01	0.038319	-2.043343e-01	-3.531563e-02	-2.021276e-02	-2.505828e-02	-3.887204e-01
75%	5.267015e-01	0.540298	6.917124e-01	7.267610e-01	6.916007e-01	6.965523e-01	3.529332e-01
max	7.748254e+00	2.046232	9.652179e+00	3.987740e+00	4.559588e+01	3.921946e+01	3.696710e+00

Univariate/Bivariate Analysis

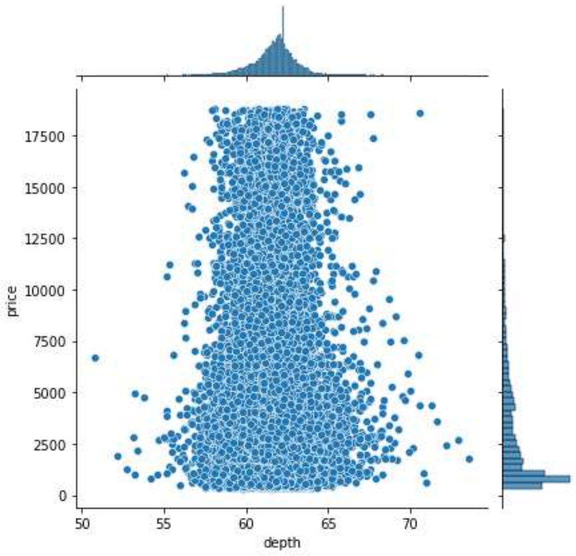
Cut Vs Price



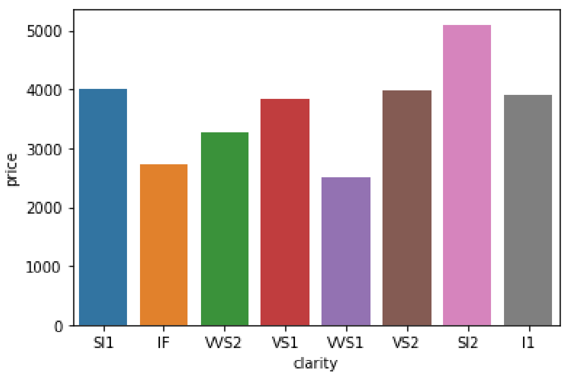
Distribution of depth



Depth vs price



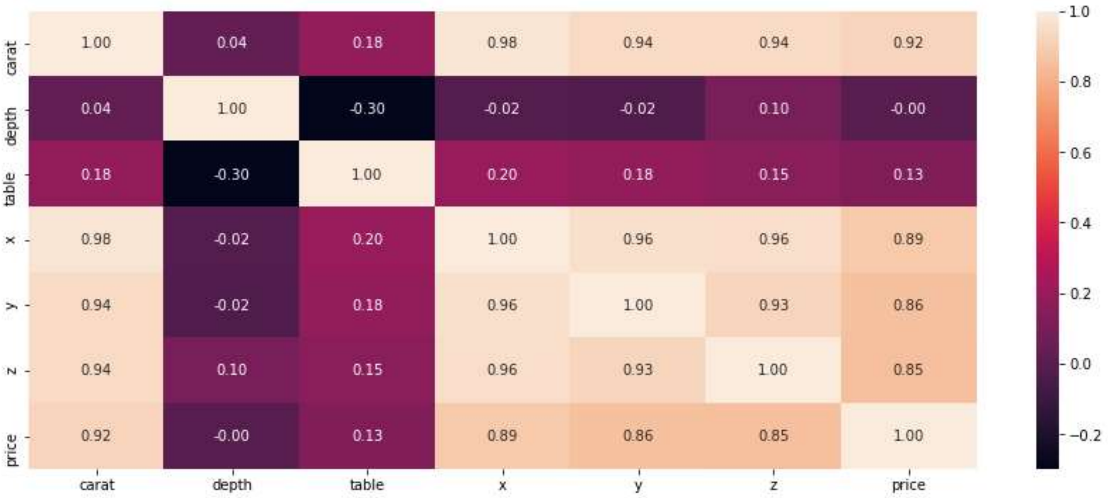
Clarity vs price



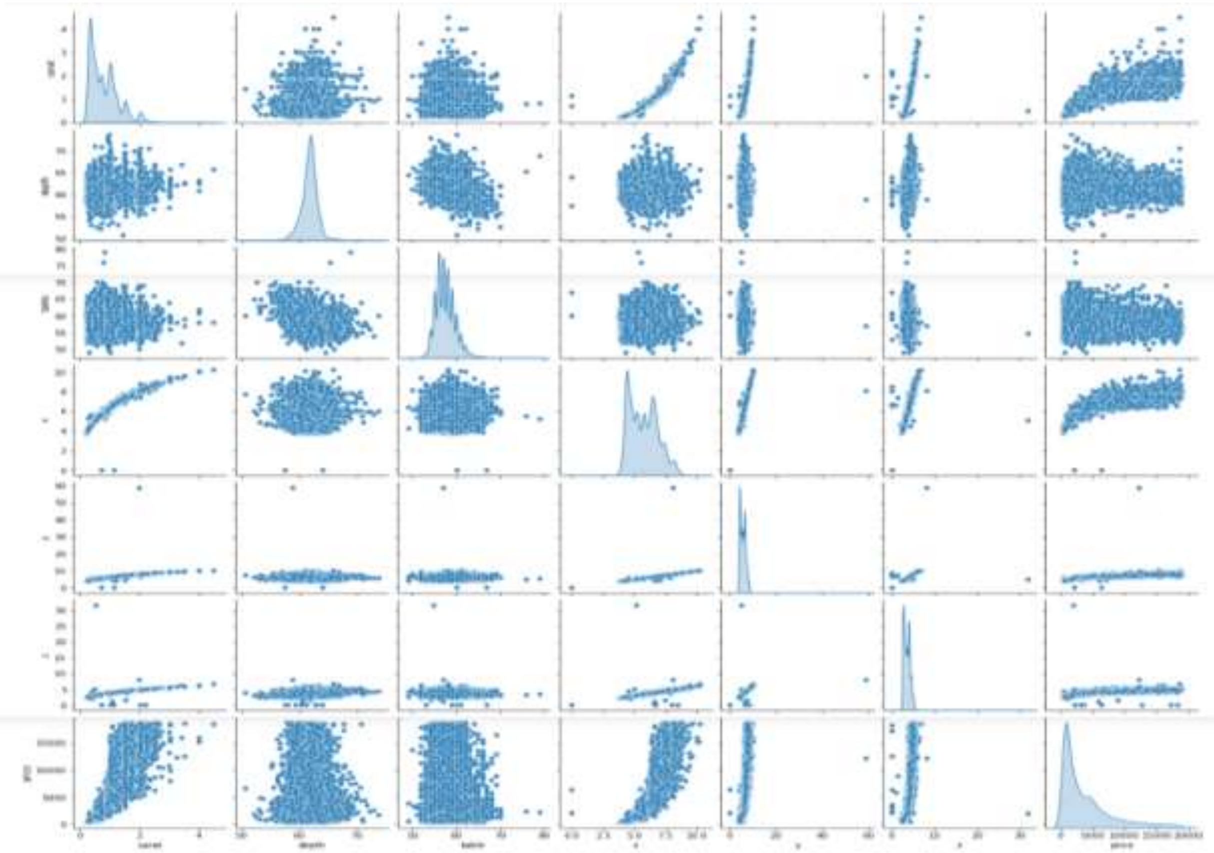
Total number of null values after imputing is 0

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 10 columns):
 #   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     26967 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)
memory usage: 2.1 MB
```

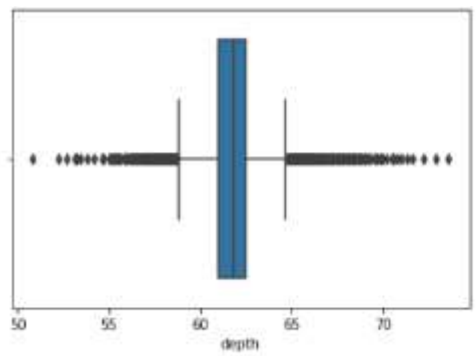
Correlation heatmap



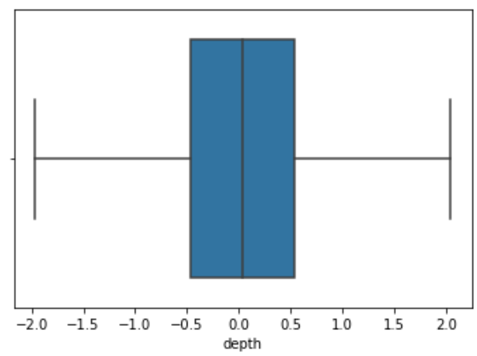
Multivariate analysis



Getting rid of the outliers is one of the important factors to improve the accuracy of the model



Depth with outliers



Depth without outliers

Encoding the data with dummy variables

	carat	depth	table	x	y	z	price	cut_Good	cut_Ideal	cut_Premium	...	color_H	color_I	color_J	clarity_IF	clarity
0	-1.043201	0.253453	0.243689	-1.293628	-1.238014	-1.218491	-0.854832	0	1	0	...	0	0	0	0	
1	-0.980405	-0.678792	0.243689	-1.160708	-1.092221	-1.162983	-0.734329	0	0	1	...	0	0	0	1	
2	0.212721	0.325164	1.139736	0.274832	0.331406	0.335747	0.583753	0	0	0	...	0	0	0	0	
3	-0.792017	-0.105103	-0.652358	-0.806254	-0.800635	-0.802177	-0.709979	0	1	0	...	0	0	0	0	
4	-1.022269	-0.965637	0.691712	-1.222737	-1.117949	-1.232368	-0.785263	0	1	0	...	0	0	0	0	

5 rows × 24 columns

One of each of the dummy variables are dropped to handle the errors created by dummy variable trap

Splitting the dataset into X and Y and further into training and testing using sklearn train_test_split function

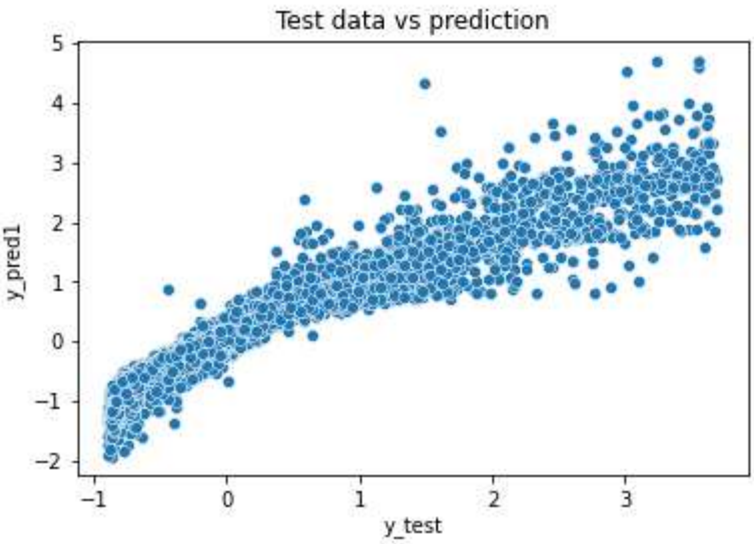
```
The intercept for our model is -1.0476440983159856
```

The coefficient of different independent variables and how much weight they have with the

```
The coefficient for carat is 1.3318892647093354
The coefficient for depth is -0.023662334068712997
The coefficient for table is -0.015054537193578875
The coefficient for x is -0.27503858859539776
The coefficient for y is -0.0013422741196084775
The coefficient for z is -0.009026606739446537
The coefficient for cut_Good is 0.13089872669610342
The coefficient for cut_Ideal is 0.19875969281511321
The coefficient for cut_Premium is 0.1720664221401289
The coefficient for cut_Very Good is 0.16522492442311376
The coefficient for color_E is -0.04967608391111707
The coefficient for color_F is -0.07133803044245939
The coefficient for color_G is -0.11726718578313688
The coefficient for color_H is -0.24059275084306256
The coefficient for color_I is -0.3731419095684082
The coefficient for color_J is -0.5925525929298248
The coefficient for clarity_IF is 1.343275822959704
The coefficient for clarity_SI1 is 0.946121345910691
The coefficient for clarity_SI2 is 0.7094812896129413
The coefficient for clarity_VS1 is 1.169911833351272
The coefficient for clarity_VS2 is 1.1008665625864102
The coefficient for clarity_VVS1 is 1.282674393917035
The coefficient for clarity_VVS2 is 1.2673058443982976
```

Score of the model with different data

```
Training data:
The score of the model for the training set is 0.9202211425206703

Testing data:
The score of the model for the testing set is 0.922955289149439
```



The vif shows the multicollinearity between the data

```
carat ---> 23.609689686089972
depth ---> 1.5528616344755648
table ---> 1.742089008617014
x ---> 45.37075412284442
y ---> 13.95113993620259
z ---> 14.032118182295342
cut_Good ---> 3.496437266066383
cut_Ideal ---> 14.421302501125455
cut_Premium ---> 8.65306433277053
cut_Very Good ---> 7.639882842322413
color_E ---> 2.366943708157424
color_F ---> 2.325147465557399
color_G ---> 2.6637696754813316
color_H ---> 2.1984851525474296
color_I ---> 1.8712033946889979
color_J ---> 1.487179428763507
clarity_IF ---> 2.1948239325098506
clarity_SI1 ---> 8.832052203355904
clarity_SI2 ---> 6.265109423058851
clarity_VS1 ---> 6.041387644134596
clarity_VS2 ---> 8.417973451681862
clarity_VVS1 ---> 3.4213578260726014
clarity_VVS2 ---> 4.18218568061231
```

5 best attributes that are important are

- carat
- clarity_IF
- clarity_VVS1
- clarity_VVS2
- clarity_VS1
- clarity_VS2

Problem 2: Logistic Regression and LDA

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.

LOGISTIC REGRESSION

Head of the data

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	no	48412	30	8	1	1	no
1	yes	37207	45	8	0	1	no
2	no	58022	46	9	0	0	no
3	no	66503	31	11	2	0	no
4	no	66734	44	12	0	2	no

Tail of the data

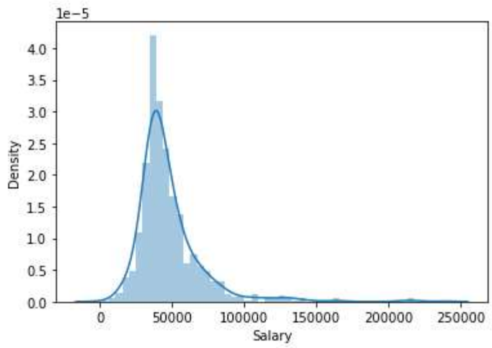
	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
867	no	40030	24	4	2	1	yes
868	yes	32137	48	8	0	0	yes
869	no	25178	24	6	2	0	yes
870	yes	55958	41	10	0	1	yes
871	no	74659	51	10	0	0	yes

Central tendency report

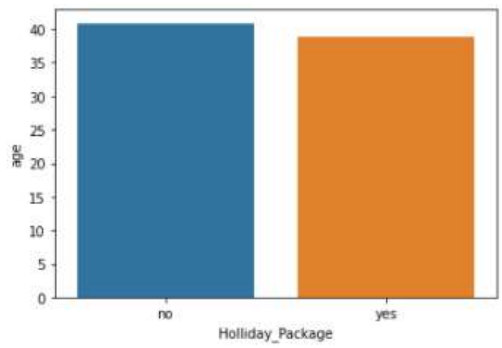
	count	mean	std	min	25%	50%	75%	max
Salary	872.0	47729.172018	23418.668531	1322.0	35324.0	41903.5	53469.5	236961.0
age	872.0	39.955275	10.551675	20.0	32.0	39.0	48.0	62.0
educ	872.0	9.307339	3.036259	1.0	8.0	9.0	12.0	21.0
no_young_children	872.0	0.311927	0.612870	0.0	0.0	0.0	0.0	3.0
no_older_children	872.0	0.982798	1.086786	0.0	0.0	1.0	2.0	6.0

Univariate/Bivariate Analysis

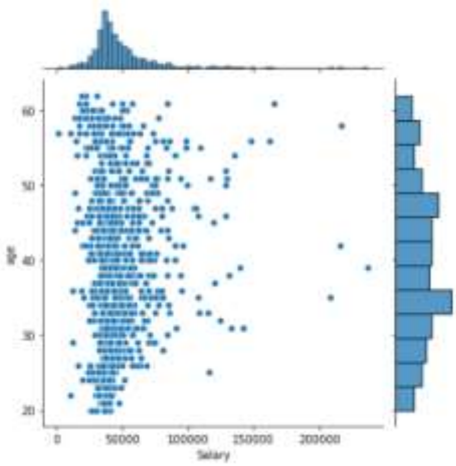
Salary density



Age vs Holiday Package

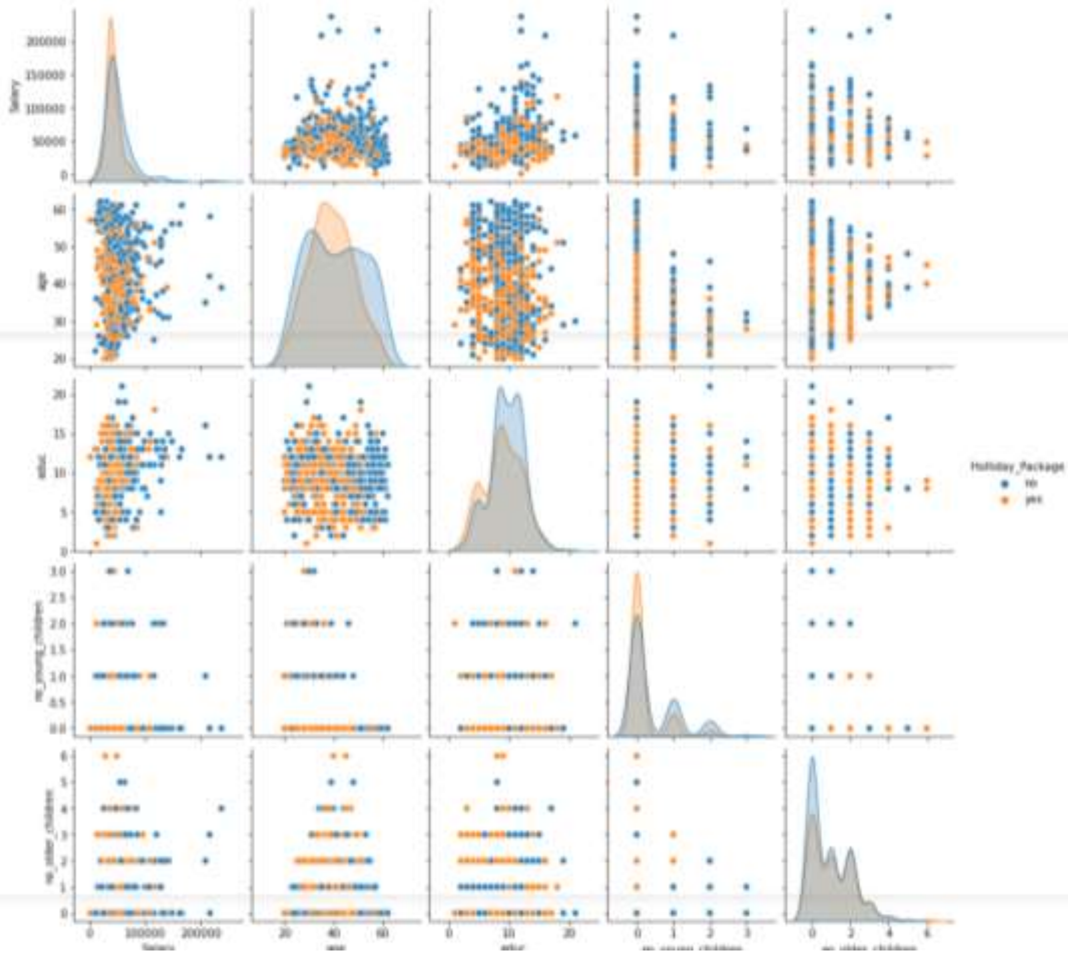


Age vs Salary



Correlation heatmap

Multivariate Analysis



Encoding the data with dummy variables

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

One of each of the dummy variables are dropped to handle the errors created by dummy variable trap

The coefficient of the model

```
The coefficient for Salary is -0.4018215927578848
The coefficient for age is -0.5626369116901058
The coefficient for educ is 0.22962820068946965
```

```
The coefficient for no_young_children is -0.9230301031161489
The coefficient for no_older_children is -0.0557768401319321
The coefficient for foreign_yes is 0.673796137429403
```

Using sklearn splitting the data into X and Y and further into training and testing data

```
The shape of X train split data (610, 6)
The shape of Y train split data (610,)
The shape of X test split data (262, 6)
The shape of Y test split data (262,)
```

Calculating score of the regression model for the training and test data

```
The score of the logistic model on training data is 0.680327868852459
The score of the logistic model on testing data is 0.6374045801526718
The accuracy of the predicted logistic model 0.6374045801526718
```

Confusion matrix

	precision	recall	f1-score	support
0	0.66	0.70	0.68	145
1	0.60	0.56	0.58	117
accuracy			0.64	262
macro avg	0.63	0.63	0.63	262
weighted avg	0.64	0.64	0.64	262

LDA – Linear discriminant Analysis

Head of the data

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	no	48412	30	8	1	1	no
1	yes	37207	45	8	0	1	no
2	no	58022	46	9	0	0	no
3	no	66503	31	11	2	0	no
4	no	66734	44	12	0	2	no

Tail of the data

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
867	no	40030	24	4	2	1	yes
868	yes	32137	48	8	0	0	yes
869	no	25178	24	6	2	0	yes
870	yes	55958	41	10	0	1	yes
871	no	74659	51	10	0	0	yes

Encoding the data

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	no	48412	30	8	1	1	0
1	yes	37207	45	8	0	1	0
2	no	58022	46	9	0	0	0
3	no	66503	31	11	2	0	0
4	no	66734	44	12	0	2	0

One column of each encoded value is dropped to handle dummy trap variable

Using sklearn splitting the data into X and Y and further into training and testing data

```
The shape of X train split data (697, 6)
The shape of Y train split data (697,)
The shape of X test split data (175, 6)
```

```
The shape of Y test split data (175,)
```

Calculating score of the regression model for the training and test data

```
The score of the LDA model on training data is 0.6628407460545194
The score of the LDA model on testing data is 0.6628571428571428
The accuracy of the predicted model 0.6628571428571428
```

Confusion matrix

	precision	recall	f1-score	support
no	0.67	0.73	0.70	94
yes	0.65	0.58	0.61	81
accuracy			0.66	175
macro avg	0.66	0.66	0.66	175
weighted avg	0.66	0.66	0.66	175

From the above, LDA has better precision , accuracy and f1-score and is cleary a better model for this than logistic regression

The most important factos affecting the holiday package choosers are

- The number of young children
- Foreigner Yes/No
- Age of the employee