PREDICTIVE MODELING PROJECT REPORT

AHMED BASHA K

25-09-2022

Contents

Problem 1 (Linear Regression) ______________________________________

1.1 The very first step of any data analysis assignment is to do the exploratory data analysis (EDA). Once you have understood the nature of all the variables, identified the response and the predictors, apply appropriate methods to determine whether there is any duplicate observation or missing data and whether the variables have symmetric or skewed distribution. Note that data may contain various types of attributes and numerical and/or visual data summarization techniques need to be appropriately decided. Both univariate and bivariate analyses and pre-processing of data are important. Check for outliers and comment on removing or keeping them while model building. Since this is a regression problem, the dependence of the response on the predictors needs to be thoroughly investigated.

1.2 Use the Pre-processed <u>Full Data</u> to develop a model to identify significant predictors. Check whether the proposed model is free of multicollinearity. Apply variable selection method as required. Show all intermediate models leading to the final model. Justify your choice of the final model. Which are the significant predictors?

1.3 Alternatively, if prediction accuracy of the price is the <u>only</u> objective, then you may want to divide the data into a training and a test set, chosen randomly, and use the training set to develop a model and test set to validate your model. Use the models developed in Part (2) to compare accuracy in training and test sets. Compare the final model of Part (2) and the proposed one in Part (3). Which model provides the most accurate prediction? If the model found in Part (2) is different from the proposed model in Part (3), give an explanation.

Problem 2(Logistic Regression and LDA) ______________________________

2.1 The very first step of any data analysis assignment is to do the exploratory data analysis (EDA). Once you have understood the nature of all the variables, identified the response and the predictors, apply appropriate methods to determine whether there is any duplicate observation or missing data and whether the variables have symmetric or skewed distribution. Note that data may contain various types of attributes and numerical and/or visual data summarization techniques need to be appropriately decided. Both univariate and bivariate analyses and pre-processing of data are important. Check for outliers and comment on removing or keeping them while model building. Since this is a regression problem, the dependence of the response on the predictors needs to be thoroughly investigated.

2.2 Use the Pre-processed <u>Full Data</u> to develop a model to identify significant predictors. Check whether the proposed model is free of multicollinearity. Apply variable selection method as required. Show all intermediate models leading to the final model. Justify your choice of the final model. Which are the significant predictors?

2.3 Alternatively, if prediction accuracy of the price is the <u>only</u> objective, then you may want to divide the data into a training and a test set, chosen randomly, and use the training set to develop a model and test set to validate your model. Use the models developed in Part (2) to compare accuracy in training and test sets. Compare the final model of Part (2) and the proposed one in Part (3). Which model provides the most accurate prediction? If the model found in Part (2) is different from the proposed model in Part (3), give an explanation.

Problem 1: Linear Regression

You are hired by a company named Gem Stones Co Ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of approximately 27,000 pieces of cubic zirconia (which is an inexpensive synthesized diamond alternative with similar qualities of a diamond).

Your objective is to accurately predict prices of the zircon pieces. Since the company profits at a different rate at different price levels, for revenue management, it is important that prices are predicted as accurately as possible. At the same time, it is important to understand which of the predictors are more important in determining the price.

The data dictionary is given below.

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.
Colour	Colour of the cubic zirconia. With D being the best and J the worst.
Clarity	Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst in terms of avg price) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, I1
Depth	The Height of 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.

The data is given in the File " cubic_zirconia.csv" As shown below.

	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
...	...	...	...	...	...	...	...	...	...	...	...
26962	26963	1.11	Premium	G	SI1	62.3	58.0	6.61	6.52	4.09	5408
26963	26964	0.33	Ideal	H	IF	61.9	55.0	4.44	4.42	2.74	1114
26964	26965	0.51	Premium	E	VS2	61.7	58.0	5.12	5.15	3.17	1656
26965	26966	0.27	Very Good	F	VVS2	61.8	56.0	4.19	4.20	2.60	682
26966	26967	1.25	Premium	J	SI1	62.0	58.0	6.90	6.88	4.27	5166

26967 rows × 11 columns

1.Exploratory Data Analysis for

The very first step of any data analysis assignment is to do the exploratory data analysis (EDA). Once you have understood the nature of all the variables, identified the response and the predictors, apply appropriate methods to determine whether there is any duplicate observation or missing data and whether the variables have a symmetric or skewed distribution. Note that data may contain various types of attributes and numerical and/or visual data summarization techniques need to be appropriately decided. Both univariate and bivariate analyses and pre-processing of data are important. Check for outliers and comment on removing or keeping them while model building. Since this is a regression problem, the dependence of the response on the predictors needs to be thoroughly investigated.

Exploratory Data Analysis:

HEAD

	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

SHAPE OF THE DATA:

- Number of rows: **26967**
- Number of columns: **10**

INFO OF THE DATA:

```
<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     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)
memory usage: 2.1+ MB
```

Descriptive Statistics:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
carat	26967.0	NaN	NaN	NaN	0.798375	0.477745	0.2	0.4	0.7	1.05	4.5
cut	26967	5	Ideal	10816	NaN	NaN	NaN	NaN	NaN	NaN	NaN
color	26967	7	G	5661	NaN	NaN	NaN	NaN	NaN	NaN	NaN
clarity	26967	8	SI1	6571	NaN	NaN	NaN	NaN	NaN	NaN	NaN
depth	26270.0	NaN	NaN	NaN	61.745147	1.41286	50.8	61.0	61.8	62.5	73.6
table	26967.0	NaN	NaN	NaN	57.45608	2.232068	49.0	56.0	57.0	59.0	79.0

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
x	26967.0	NaN	NaN	NaN	5.729854	1.128516	0.0	4.71	5.69	6.55	10.23
y	26967.0	NaN	NaN	NaN	5.733569	1.166058	0.0	4.71	5.71	6.54	58.9
z	26967.0	NaN	NaN	NaN	3.538057	0.720624	0.0	2.9	3.52	4.04	31.8
price	26967.0	NaN	NaN	NaN	3939.518115	4024.864666	326.0	945.0	2375.0	5360.0	18818.0

Unique value:

```
CUT :  5
Fair           781
Good          2441
Very Good     6030
Premium       6899
Ideal        10816
Name: cut, dtype: int64
```

```
COLOR :  7
J    1443
I    2771
D    3344
H    4102
F    4729
E    4917
G    5661
Name: color, dtype: int64
```

```
CLARITY :  8
I1       365
IF       894
VVS1    1839
VVS2    2531
VS1     4093
SI2     4575
VS2     6099
SI1     6571
Name: clarity, dtype: int64
```

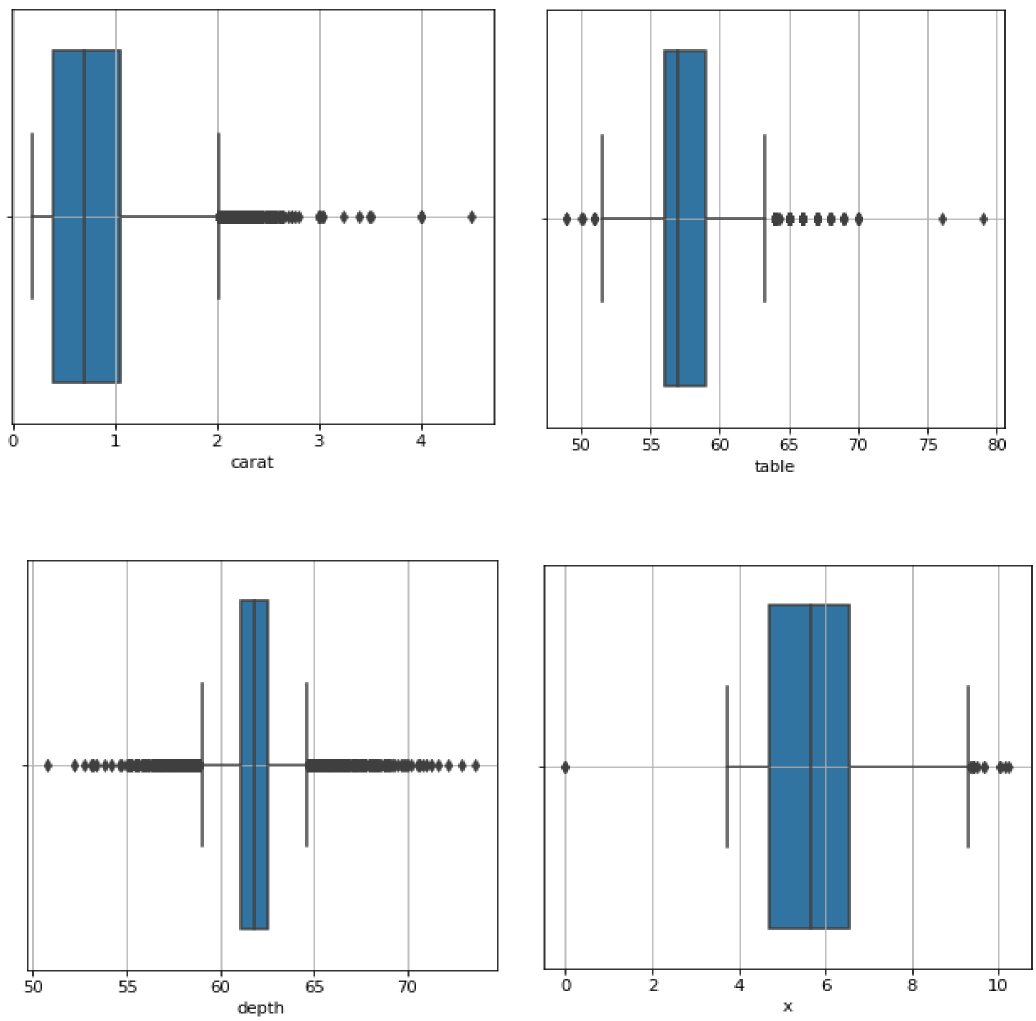
Missing values:

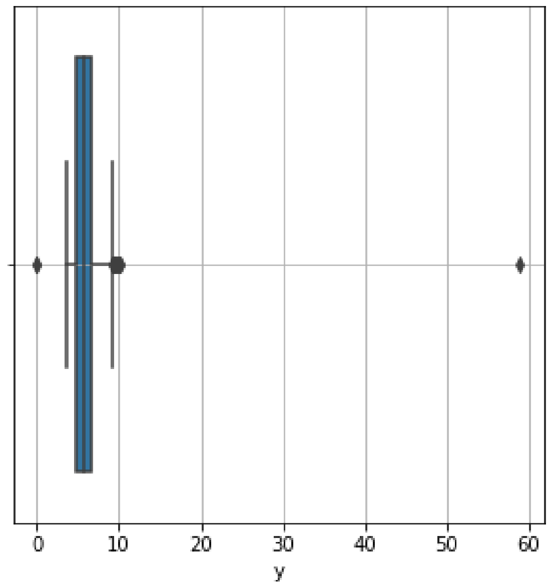
```
carat        0
cut          0
color        0
clarity      0
depth      697
table        0
x            0
y            0
z            0
price        0
dtype: int64
```

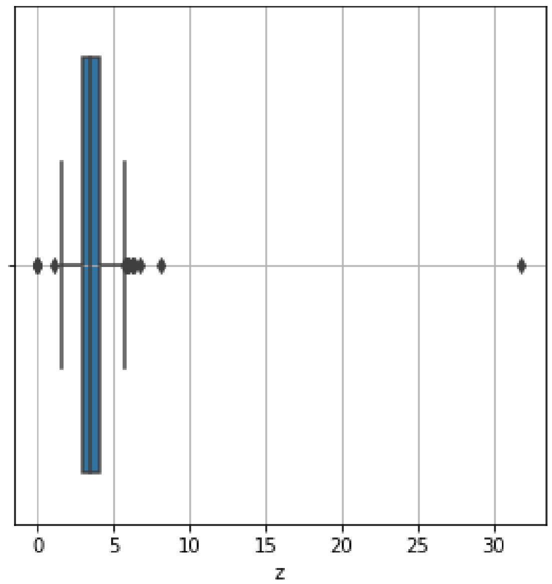
Duplicates values:

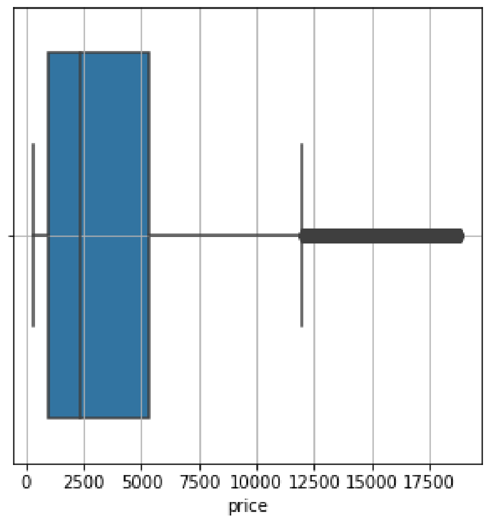
34

Univariate Analysis of Continuous and Categorical variables:

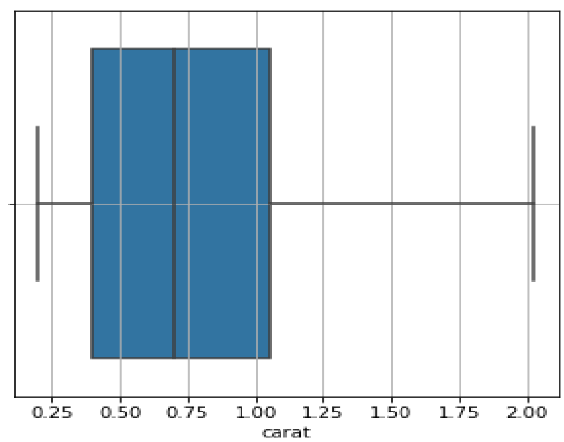


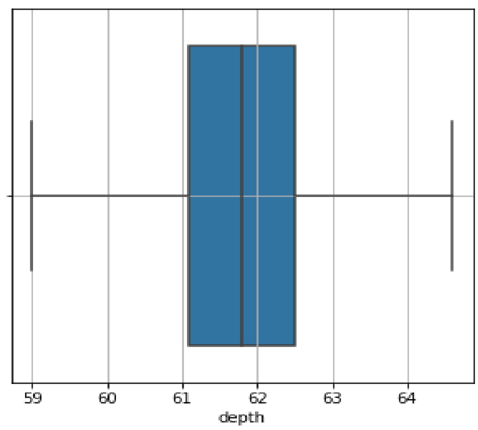


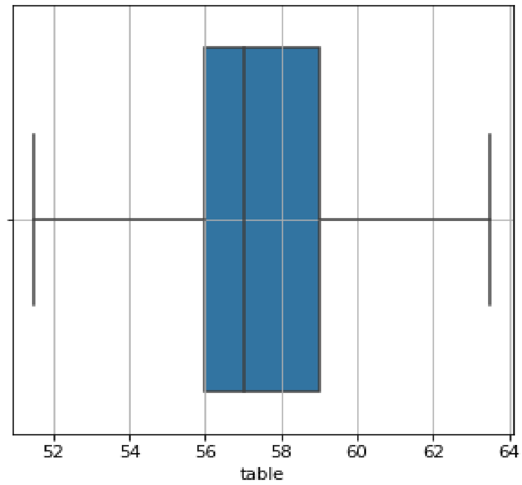


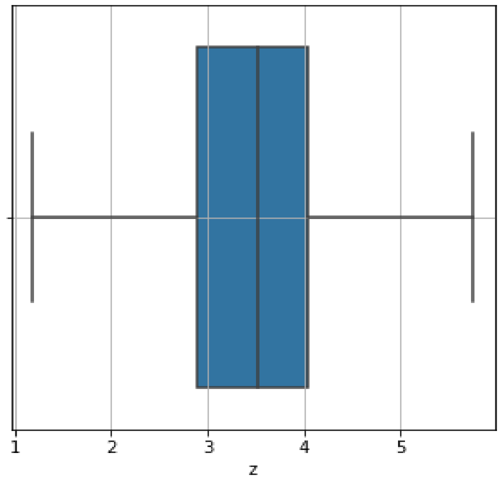


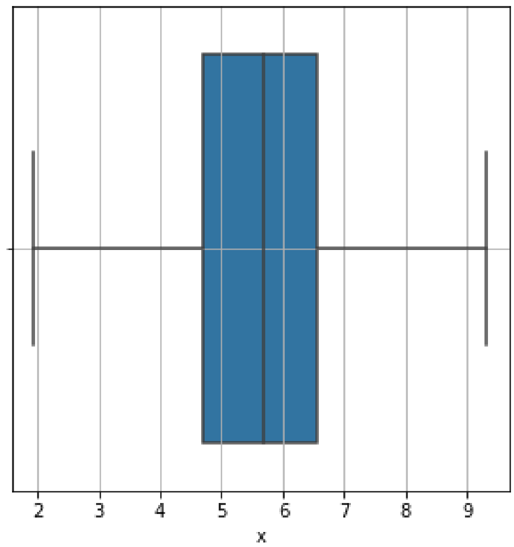
<u>After treating the outliers:</u>

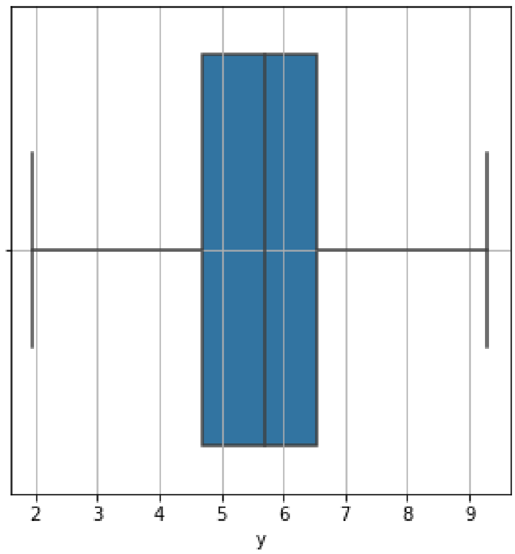


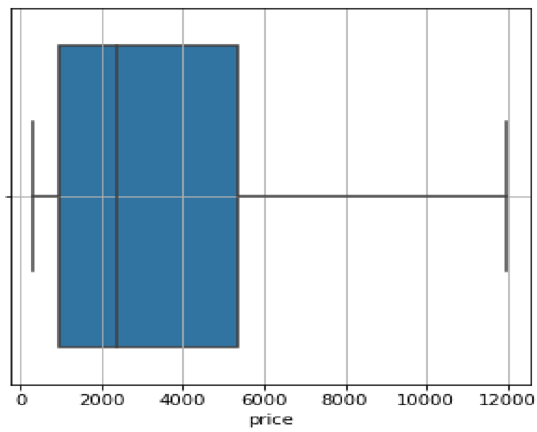



52
54
56
58
60
62
64
table

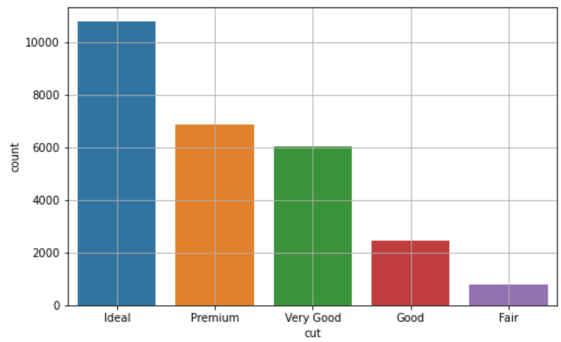

1
2
3
4
5
z

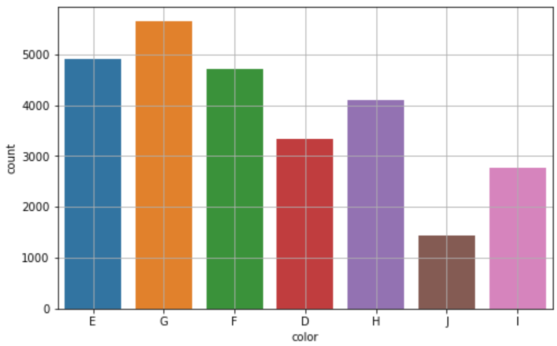

2
3
4
5
6
7
8
9
x

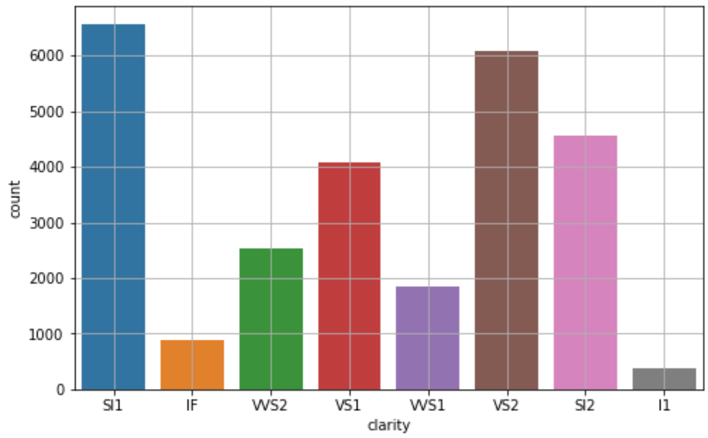

2
3
4
5
6
7
8
9
y


0
2000
4000
6000
8000
10000
12000
price

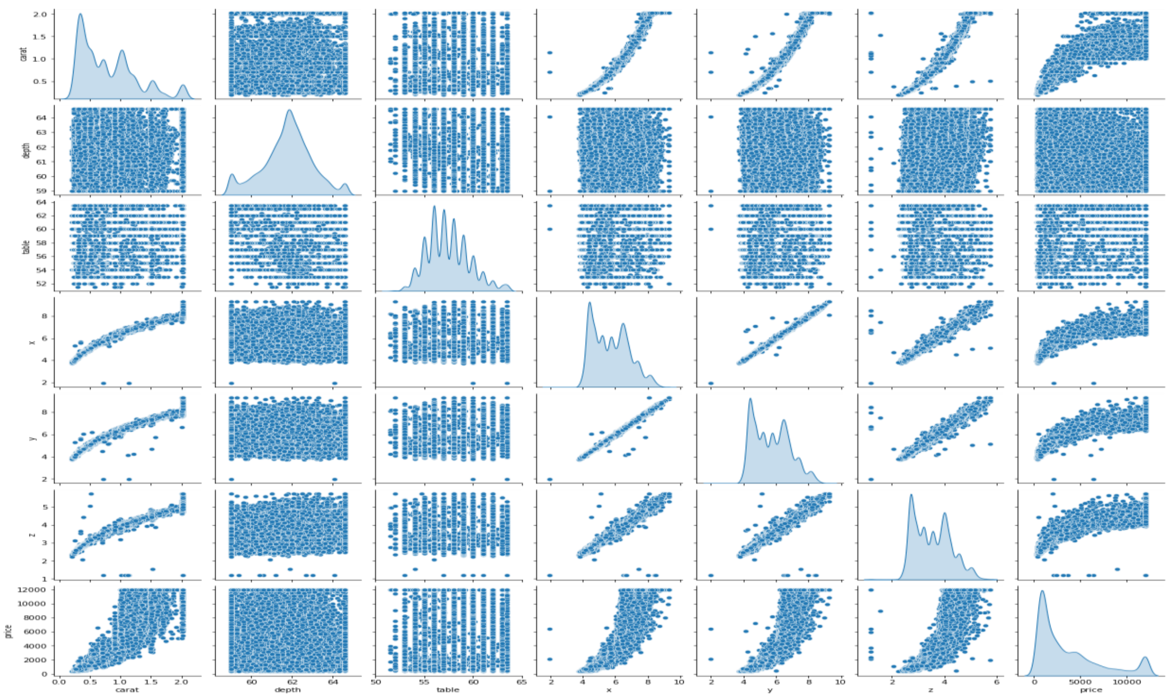
Checking the spread of the data using count plot for the categorical variables.







<u>Bivariate Analysis:</u>

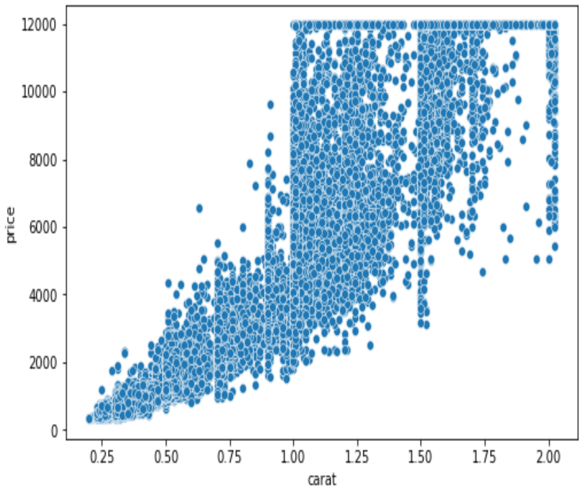


Correlation:

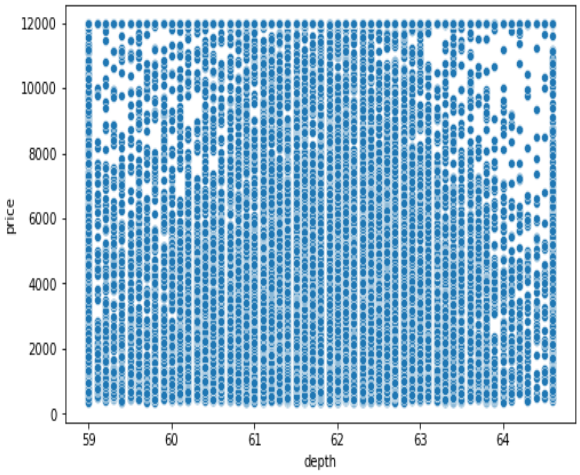
	carat	depth	table	x	y	z	price
carat	1.000000	0.029433	0.187143	0.982387	0.981464	0.977508	0.936762
depth	0.029433	1.000000	-0.289357	-0.019848	-0.022884	0.095253	-0.001060
table	0.187143	-0.289357	1.000000	0.199061	0.193428	0.159380	0.137880
x	0.982387	-0.019848	0.199061	1.000000	0.998491	0.988168	0.912933
y	0.981464	-0.022884	0.193428	0.998491	1.000000	0.987841	0.914361
z	0.977508	0.095253	0.159380	0.988168	0.987841	1.000000	0.905866
price	0.936762	-0.001060	0.137880	0.912933	0.914361	0.905866	1.000000

Comparison:

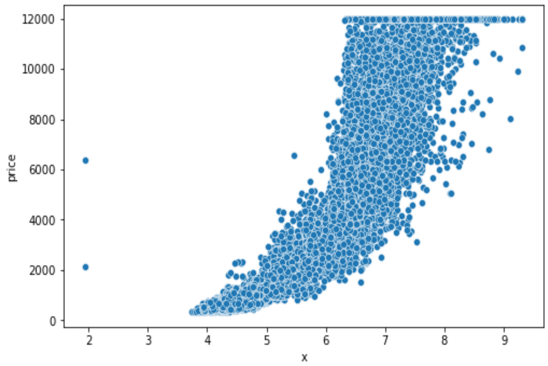
Carat vs Price



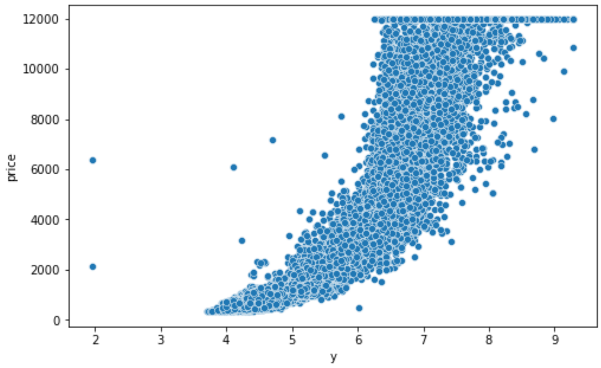
Depth vs Price



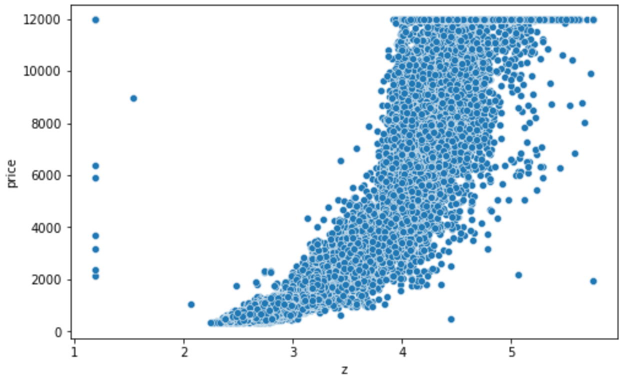
X vs Price



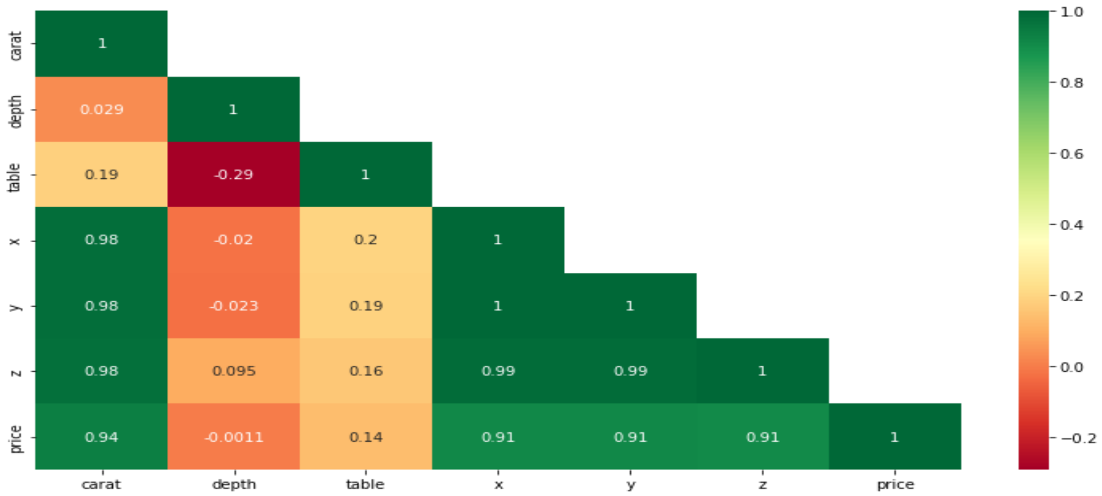
Y vs Price



Z vs Price



<u>Correlation plot:</u>



The matrix clearly shows the presence of multi collinearity in the dataset.

Conclusion of EDA:

- Price – This variable gives the continuous output with the price. This will be our Target Variable.
- Carat, depth, table, x, y, z variables are numerical or continuous variables.
- Cut, Clarity and colour are categorical variables.
- our study which leaves the shape of the dataset with 26967 rows & 10 Columns.
- Only in-depth 697missing values are present which we will impute by its median values.

2.Build various iterations of the Linear Regression model using appropriate variable selection techniques for the full data.

Use Full Data to develop a model to identify significant predictors. Check whether the proposed model is free of multicollinearity. Apply variable selection method as required. Show all intermediate models leading to the final model. Justify your choice of the final model. Which are the significant predictors?

Getting unique counts of all Objects.

```
cut
 Ideal        10805
Premium       6886
Very Good     6027
Good          2435
Fair           780
Name: cut, dtype: int64


color
 G    5653
E    4916
F    4723
H    4095
D    3341
I    2765
J    1440
Name: color, dtype: int64


clarity
 SI1     6565
VS2     6093
SI2     4564
VS1     4087
VVS2    2530
VVS1    1839
IF       891
I1       364
Name: clarity, dtype: int64
```

Converting objects to categorical codes:

	carat	cut	color	depth	table	x	y	z	price	clarity_0	clarity_1	clarity_2	clarity_3	clarity_4
0	0.30	2	1	62.1	58.0	4.27	4.29	2.66	499.0	0	0	0	1	0
1	0.33	2	2	60.8	58.0	4.42	4.46	2.70	984.0	1	0	0	0	0
2	0.90	1	1	62.2	60.0	6.04	6.12	3.78	6289.0	0	1	0	0	0
3	0.42	2	2	61.6	56.0	4.82	4.80	2.96	1082.0	0	0	1	0	0
4	0.31	2	2	60.4	59.0	4.35	4.43	2.65	779.0	0	1	0	0	0

Check for Multicollinearity:

```
carat  VIF =  331.65
cut  VIF =  331.65
color  VIF =  331.65
clarity_0  VIF =  331.65
clarity_1  VIF =  331.65
clarity_2  VIF =  331.65
clarity_3  VIF =  331.65
clarity_4  VIF =  331.65
depth  VIF =  331.65
table  VIF =  331.65
x  VIF =  331.65
y  VIF =  331.65
z  VIF =  331.65
price  VIF =  331.65
```

Building a base model with all the features:

OLS Regression Results			
Dep. Variable:	price	**R-squared:**	0.930
Model:	OLS	**Adj. R-squared:**	0.930
Method:	Least Squares	**F-statistic:**	2.979e+04
Date:	Sat, 24 Sep 2022	**Prob (F-statistic):**	0.00
Time:	23:30:09	**Log-Likelihood:**	-2.2195e+05
No. Observations:	26933	**AIC:**	4.439e+05
Df Residuals:	26920	**BIC:**	4.440e+05
Df Model:	12		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	1256.6028	480.359	2.616	0.009	315.073	2198.132
carat	8728.4221	68.185	128.011	0.000	8594.776	8862.068
cut	188.6361	11.831	15.944	0.000	165.447	211.825
color	-386.9550	5.492	-70.452	0.000	-397.721	-376.189
clarity_0	1642.9248	98.277	16.717	0.000	1450.296	1835.554
clarity_1	1408.1692	96.554	14.584	0.000	1218.919	1597.420
clarity_2	805.0389	97.147	8.287	0.000	614.625	995.453
clarity_3	-157.4673	97.925	-1.608	0.108	-349.406	34.472
clarity_4	-2442.0629	105.584	-23.129	0.000	-2649.013	-2235.113
depth	-14.8724	7.634	-1.948	0.051	-29.836	0.092
table	-25.5486	3.032	-8.426	0.000	-31.492	-19.605

x	-1496.6947	99.029	-15.114	0.000	-1690.796	-1302.593
y	1302.7690	97.637	13.343	0.000	1111.396	1494.142
z	-282.6023	82.225	-3.437	0.001	-443.767	-121.437

Omnibus:	5531.122	**Durbin-Watson:**	2.015
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	16723.084
Skew:	1.065	**Prob(JB):**	0.00
Kurtosis:	6.220	**Cond. No.**	1.39e+17

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 9.99e-27. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Building 2nd iteration removing 'y' as p-value>0.05:

OLS Regression Results			
Dep. Variable:	price	**R-squared:**	0.930
Model:	OLS	**Adj. R-squared:**	0.929
Method:	Least Squares	**F-statistic:**	3.227e+04
Date:	Sat, 24 Sep 2022	**Prob (F-statistic):**	0.00
Time:	23:30:10	**Log-Likelihood:**	-2.2204e+05
No. Observations:	26933	**AIC:**	4.441e+05
Df Residuals:	26921	**BIC:**	4.442e+05

Df Model:	11		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	3226.9665	458.599	7.037	0.000	2328.089	4125.843
carat	8781.2335	68.293	128.581	0.000	8647.375	8915.092
cut	159.8091	11.670	13.694	0.000	136.935	182.683
color	-387.7263	5.510	-70.365	0.000	-398.527	-376.926
clarity_0	2057.5587	93.541	21.996	0.000	1874.213	2240.904
clarity_1	1818.0238	91.838	19.796	0.000	1638.017	1998.031
clarity_2	1207.1576	92.658	13.028	0.000	1025.544	1388.771
clarity_3	241.2205	93.562	2.578	0.010	57.835	424.606
clarity_4	-2096.9942	102.704	-20.418	0.000	-2298.299	-1895.689
depth	-44.3476	7.332	-6.049	0.000	-58.718	-29.977
table	-32.2164	3.001	-10.737	0.000	-38.098	-26.335
x	-396.0711	54.976	-7.204	0.000	-503.827	-288.316
z	-0.3495	79.718	-0.004	0.997	-156.601	155.902

Omnibus:	5467.329	**Durbin-Watson:**	2.015
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	16861.885
Skew:	1.046	**Prob(JB):**	0.00
Kurtosis:	6.263	**Cond. No.**	1.42e+17

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 9.6e-27. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Building 3nd iteration removing 'Z' as p-value>0.05:

OLS Regression Results			
Dep. Variable:	price	**R-squared:**	0.930
Model:	OLS	**Adj. R-squared:**	0.929
Method:	Least Squares	**F-statistic:**	3.550e+04
Date:	Sat, 24 Sep 2022	**Prob (F-statistic):**	0.00
Time:	23:30:11	**Log-Likelihood:**	-2.2204e+05
No. Observations:	26933	**AIC:**	4.441e+05
Df Residuals:	26922	**BIC:**	4.442e+05
Df Model:	10		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	3228.0094	392.058	8.234	0.000	2459.556	3996.463
carat	8781.2151	68.163	128.827	0.000	8647.612	8914.818
cut	159.8126	11.642	13.728	0.000	136.994	182.631

color	-387.7260	5.510	-70.369	0.000	-398.526	-376.926
clarity_0	2057.7662	80.687	25.503	0.000	1899.615	2215.918
clarity_1	1818.2317	78.642	23.121	0.000	1664.090	1972.373
clarity_2	1207.3660	79.528	15.182	0.000	1051.487	1363.245
clarity_3	241.4295	80.502	2.999	0.003	83.642	399.217
clarity_4	-2096.7840	90.831	-23.084	0.000	-2274.818	-1918.751
depth	-44.3694	5.392	-8.228	0.000	-54.939	-33.800
table	-32.2155	2.993	-10.763	0.000	-38.082	-26.348
x	-396.2783	28.108	-14.098	0.000	-451.372	-341.185

Omnibus:	5467.320	**Durbin-Watson:**	2.015
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	16861.833
Skew:	1.046	**Prob(JB):**	0.00
Kurtosis:	6.263	**Cond. No.**	1.41e+17

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 9.63e-27. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

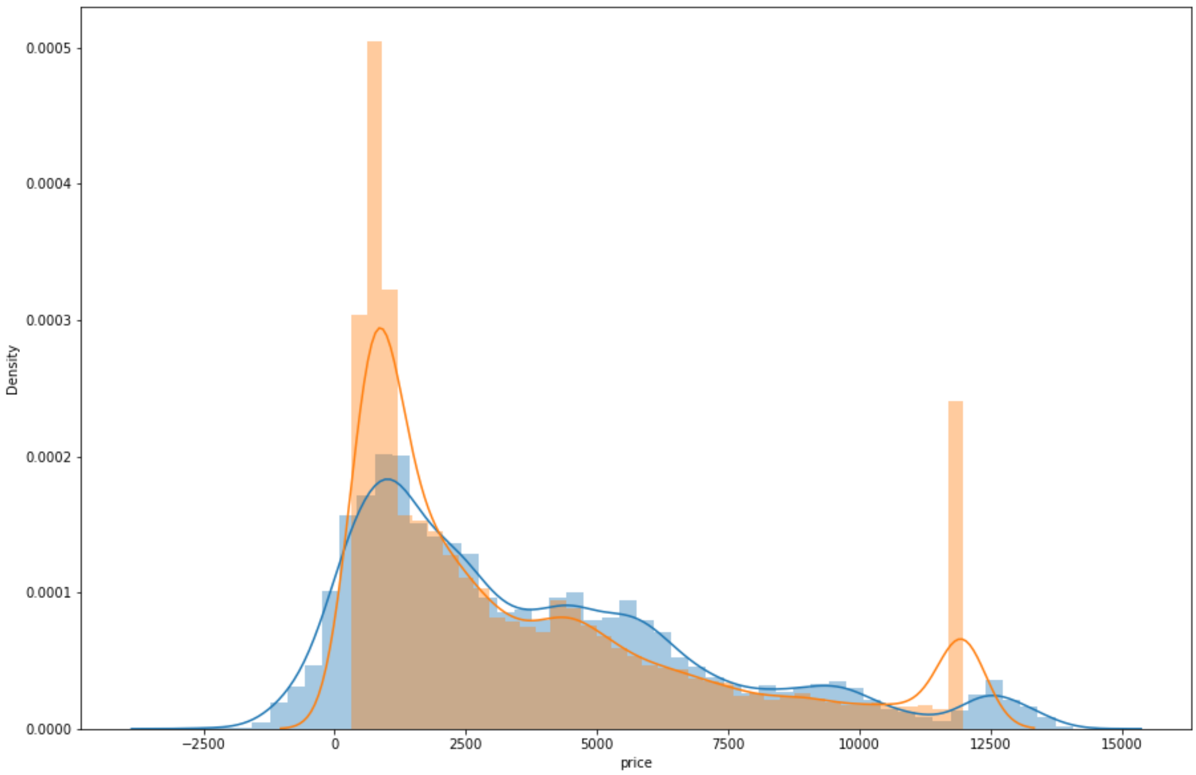
Re-check for Multicollinearity:

```
cut  VIF =  1.01
color  VIF =  1.07
clarity_0  VIF =  1.02
clarity_1  VIF =  1.06
clarity_2  VIF =  1.01
clarity_3  VIF =  1.08
clarity_4  VIF =  1.01
depth  VIF =  1.0
table  VIF =  1.04
x  VIF =  inf
```

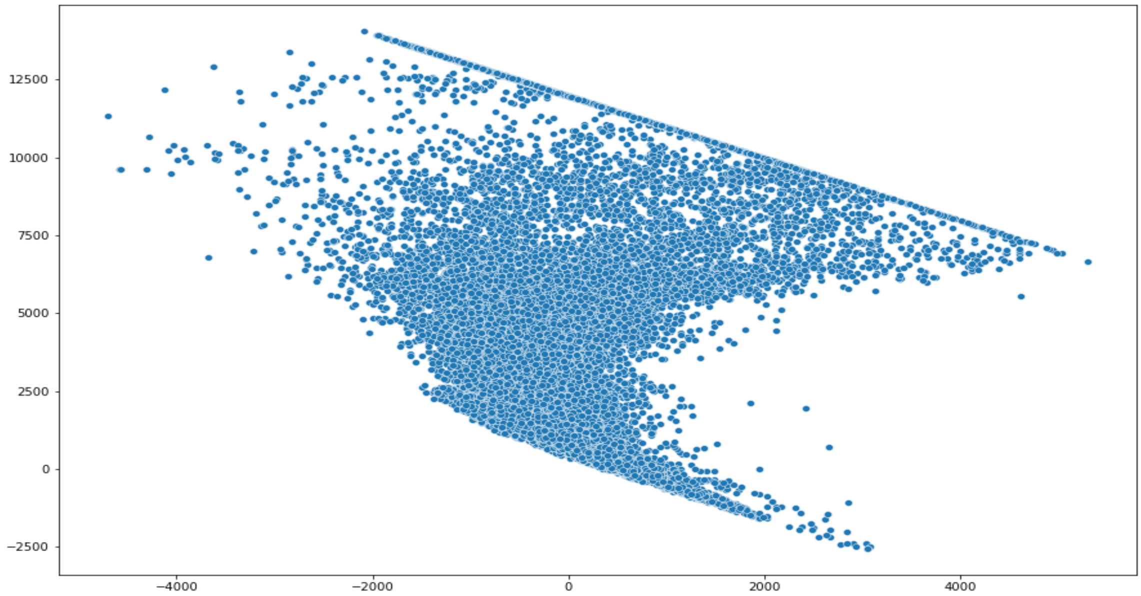
Using the last best model:

```
0        -280.241006
1        1410.044509
2        5635.197209
3        1220.377845
4        1008.157509
             ...
26962    5508.652147
26963    1062.233010
26964    2210.661885
26965     595.030098
26966    5860.960279
Length: 26933, dtype: float64
```

Distplot:



Linear Relationship b/w Dependent and Independent Variables:



3.Split the data into training (70%) and test (30%). Build the various iterations of the Linear Regression models on the training data and use those models to predict on the test data using appropriate model evaluation metrics.

If prediction accuracy of the price is the only objective, then you may want to divide the data into a training and a test set, chosen randomly, and use the training set to develop a model and test set to validate your model. Use the models developed in Part (II) to compare accuracy in training and test sets. Compare the final model of Part (II) and the proposed one in Part (III). Which model provides the most accurate prediction? If the model found in Part (II) is different from the proposed model in Part (III), give an explanation.

Best Model vs Base Model

Base Model building using sklearn Linear Regression:

- After Training Data Prediction:

```
Training Data RMSE of model base: 923.42
```

- After Test Data Prediction:

```
Test Data RMSE of model base: 905.12
```

	RMSE Training Data	RMSE Test Data
Base Model	**923.42**	**905.12**

Best Model building using sklearn Linear Regression:

- After Training Data Prediction

`Training Data RMSE of model best:` **926.17**

- After Test Data Prediction:

`Test Data RMSE of model best:` **908.81**

	RMSE Training Data	RMSE Test Data
Best Model	**926.17**	**908.81**

Best Model vs Base Model

	RMSE Training Data	RMSE Test Data
Base Model	**923.42**	**905.12**
Best Model	**926.17**	**908.81**

Problem 2: Logistic Regression

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

The data is given in the File " Holiday_Package.csv" As shown below.

	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
...	...	...	...	...	...	...	...	...
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

872 rows × 8 columns

2.1Exploratory Data Analysis for Problem 2

The very first step of any data analysis assignment is to do the exploratory data analysis (EDA). Once you have understood the nature of all the variables, especially identified the response and the predictors, apply appropriate methods to determine whether there is any duplicate observation or missing data and whether the variables have a symmetric or skewed distribution. Note that data may contain various types of attributes and numerical and/or visual data summarization techniques need to be appropriately decided. Both univariate and bivariate analyses and pre-processing of data are important. Check for outliers and comment on removing or keeping them while model building. For this is a classification problem, the dependence of the response on the predictors needs to be investigated.

<u>EDA:</u>

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

After removing the **Unnamed: 0** column:

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	0	48412.0	30	8	1	1	no
1	1	37207.0	45	8	0	1	no
2	0	58022.0	46	9	0	0	no
3	0	66503.0	31	11	2	0	no
4	0	66734.0	44	12	0	2	no

Shape of the data:

- Number of rows: **872**
- Number of columns: **7**

Information of the data:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Holliday_Package   872 non-null    int8   
 1   Salary             872 non-null    float64
 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: float64(1), int64(4), int8(1), object(1)
memory usage: 41.9+ KB
```

Descriptive Statistics:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Holliday_Package	872.0	NaN	NaN	NaN	0.459862	0.498672	0.0	0.0	0.0	1.0	1.0
Salary	872.0	NaN	NaN	NaN	45608.336869	15699.745151	8105.75	35324.0	41903.5	53469.5	80687.75
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

Missing Values:

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

Duplicates:

Number of duplicate rows = 1

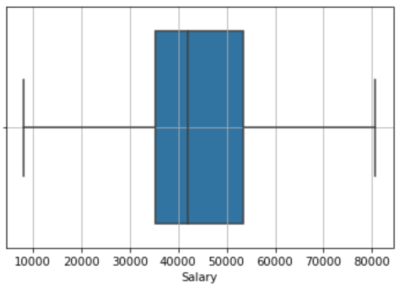
unique counts of all Objects:

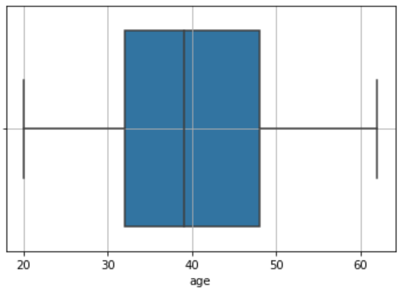
foreign

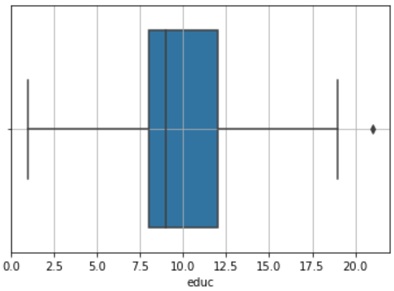
```
no      656
yes     216
Name: foreign, dtype: int64
```

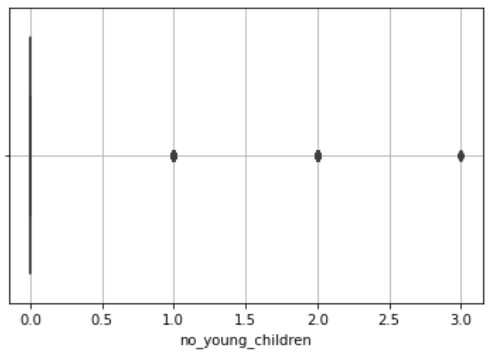
Univariate Analysis:

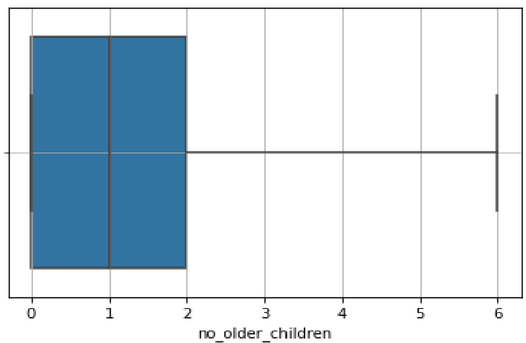
Boxplot



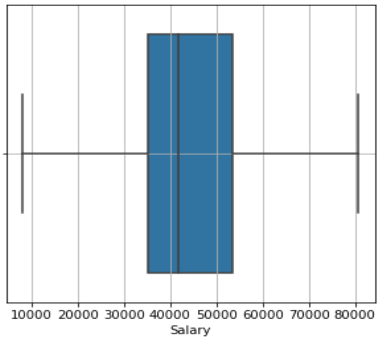


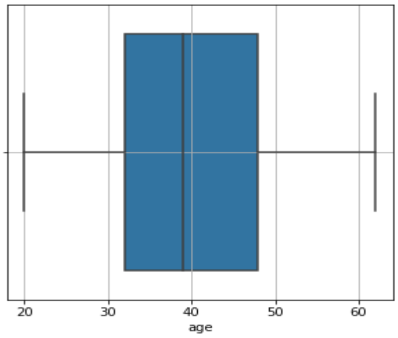


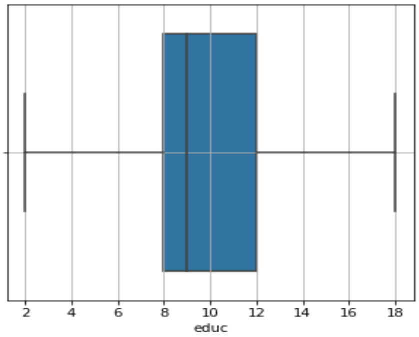


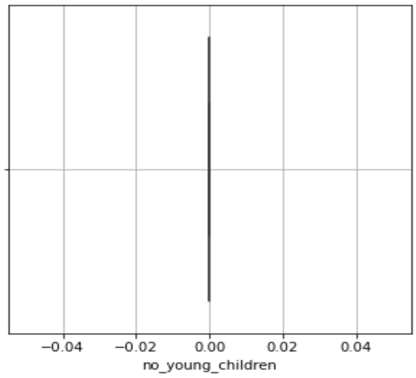


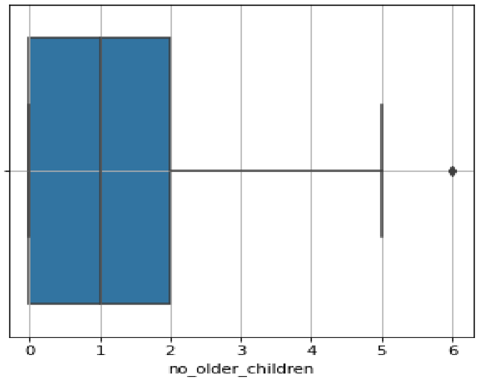
Treating the outliers:



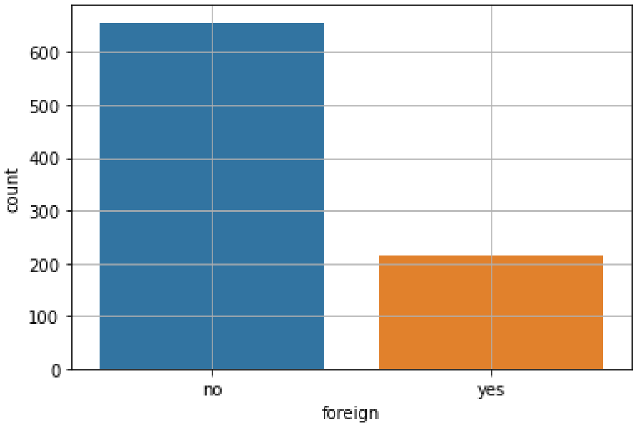


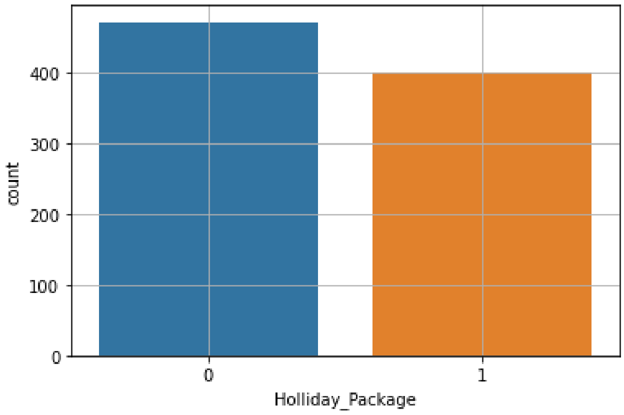






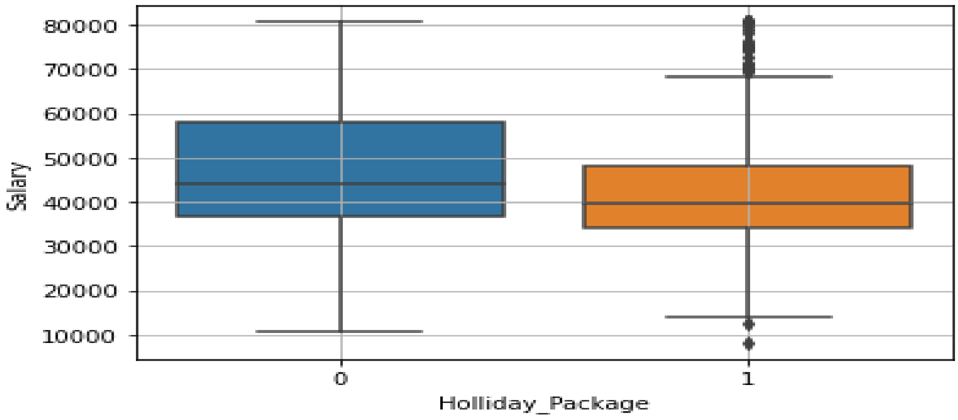
Count plot



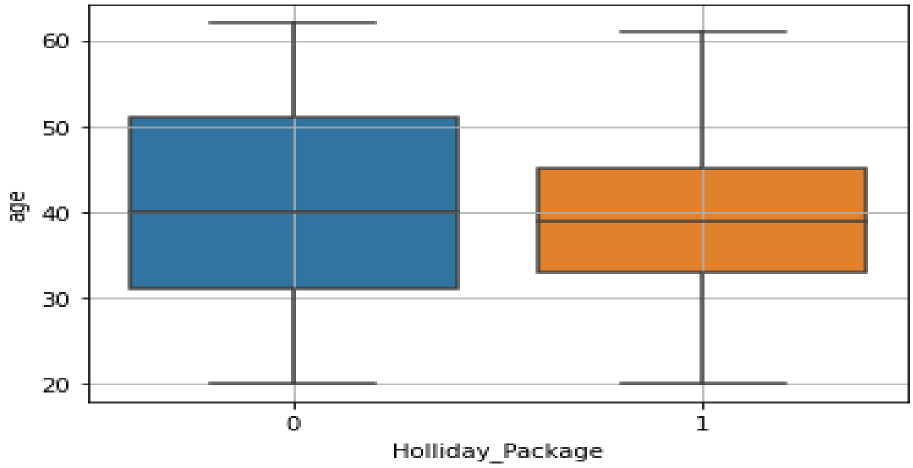


Bivariate Analysis:

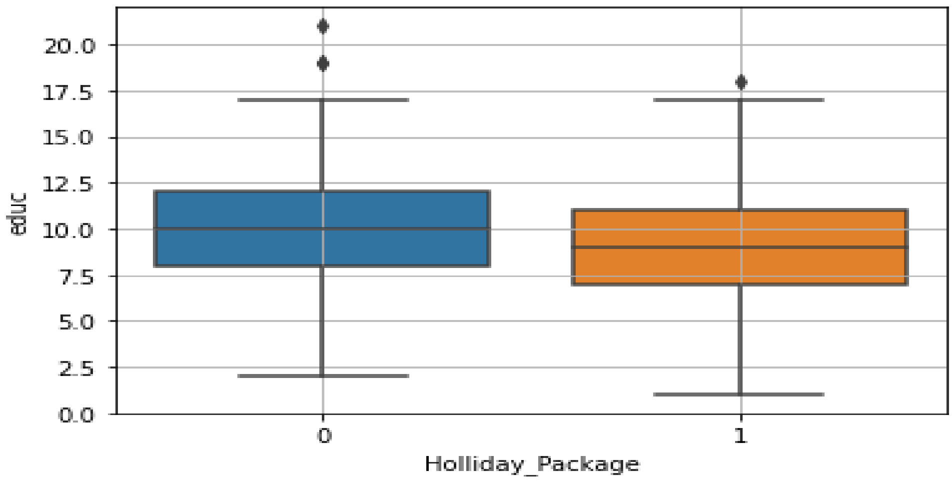
Salary VS Holiday Package



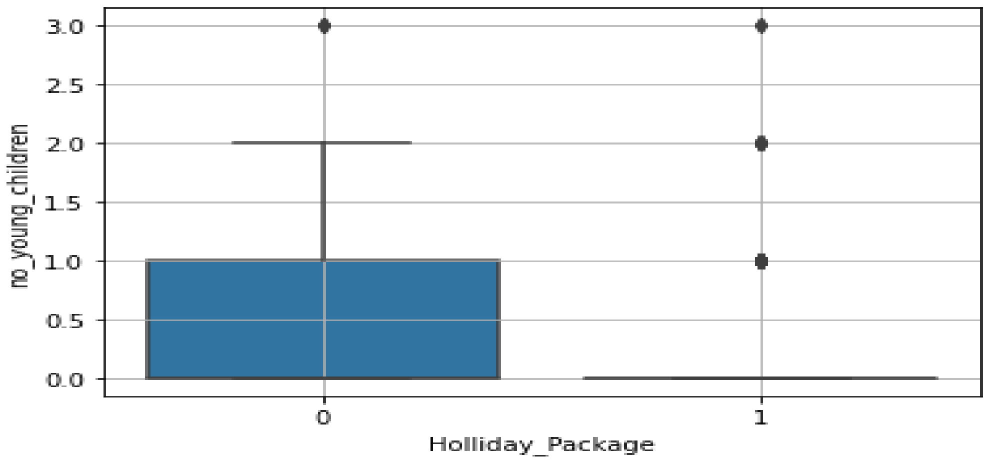
Age VS Holiday Package



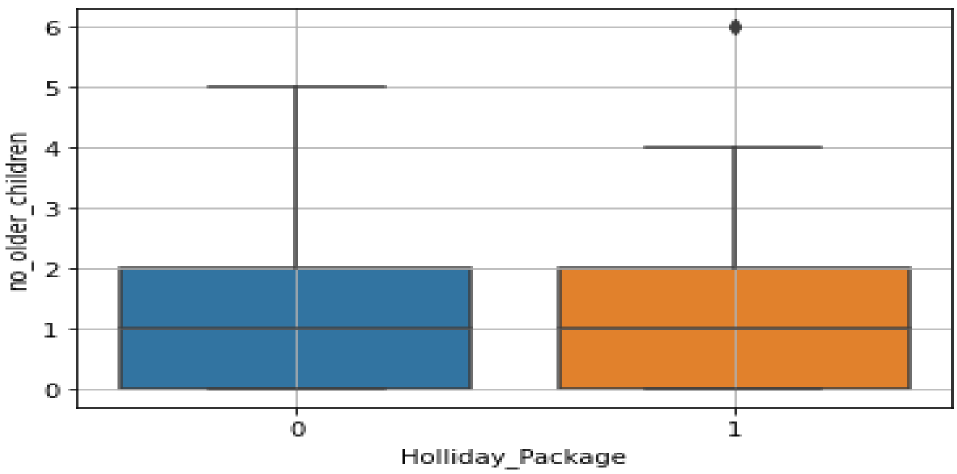
Educe VS Holiday Package



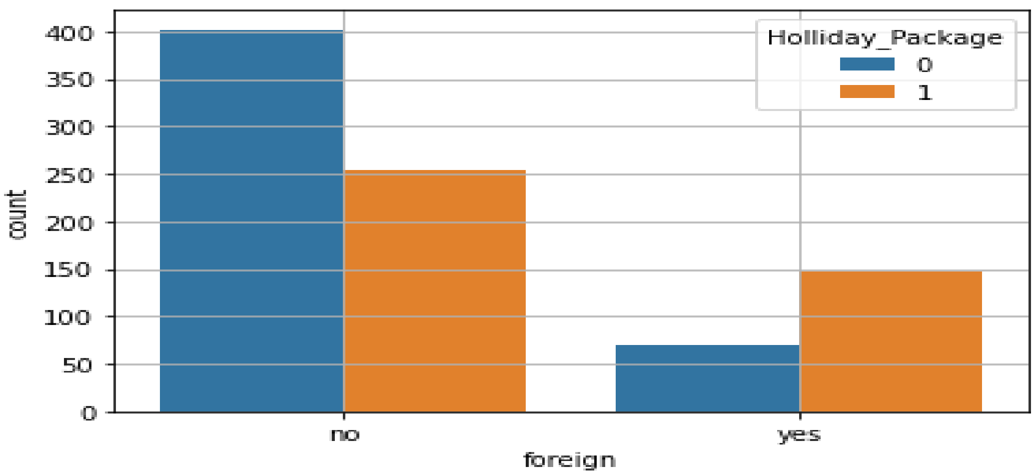
No young children Vs Holiday Package



No older children Vs Holiday Package



Foreign Vs Holiday Package



Converting the Target Variable into Categorical:

```
0    471
1    401
Name: Holliday Package, dtype: int64
```

Information of the data:

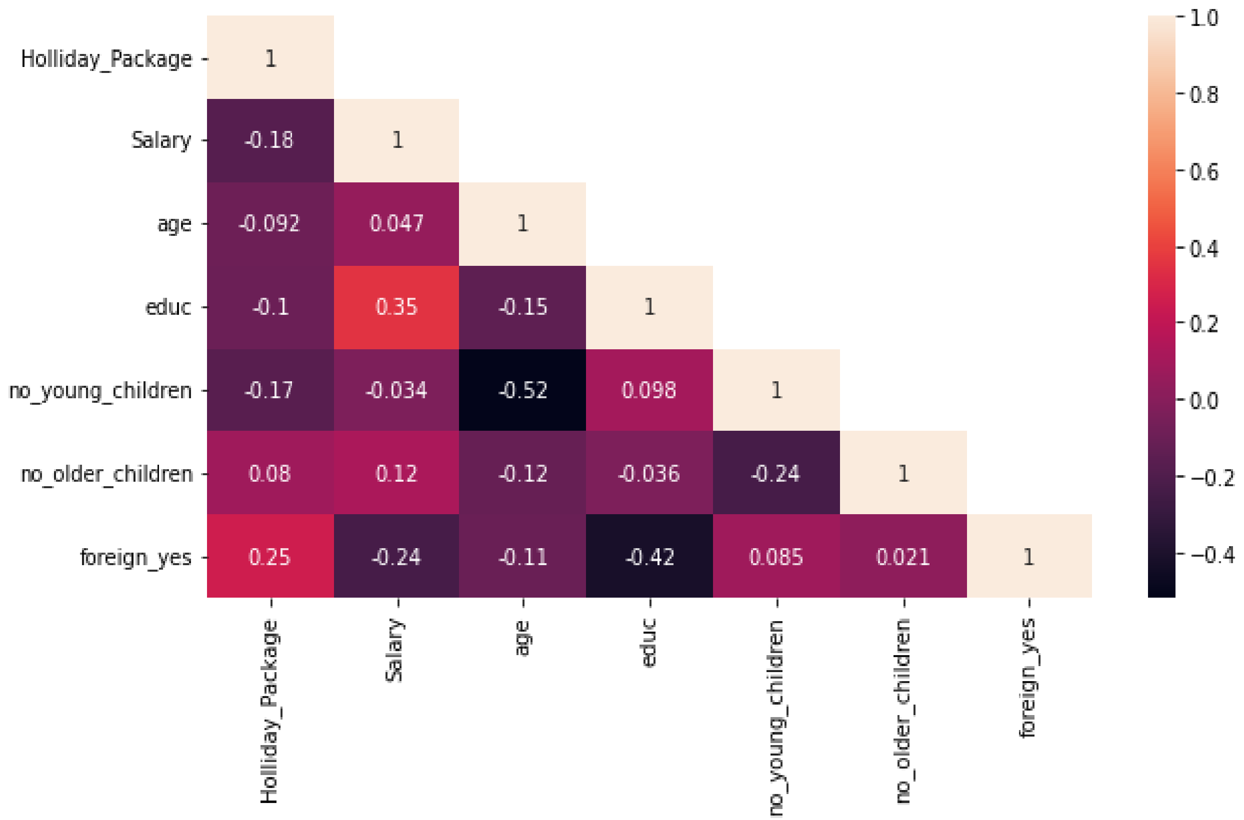
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Holliday_Package   872 non-null    int8   
 1   Salary             872 non-null    float64
 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: float64(1), int64(4), int8(1), object(1)
memory usage: 41.9+ KB
```

Creating the dummy variables for foregin variable:

	Holliday Package	Salary	age	educe	no_young_children	no_older_children	foreign yes
0	0	48412.0	30	8	1	1	0
1	1	37207.0	45	8	0	1	0
2	0	58022.0	46	9	0	0	0
3	0	66503.0	31	11	2	0	0
4	0	66734.0	44	12	0	2	0

Correlations:

	Holliday Package	Salary	age	educe	no_young_children	no_older_children	foreign_yes
Holliday Package	1.000000	-0.180214	-0.092311	-0.102552	-0.173115	0.080286	0.254096
Salary	-0.180214	1.000000	0.047029	0.352726	-0.034360	0.121993	-0.239387
age	-0.092311	0.047029	1.000000	-0.149294	-0.519093	-0.116205	-0.107148
educe	-0.102552	0.352726	-0.149294	1.000000	0.098350	-0.036321	-0.419678
no_young_children	-0.173115	-0.034360	-0.519093	0.098350	1.000000	-0.238428	0.085111
no_older_children	0.080286	0.121993	-0.116205	-0.036321	-0.238428	1.000000	0.021317
foreign_yes	0.254096	-0.239387	-0.107148	-0.419678	0.085111	0.021317	1.000000

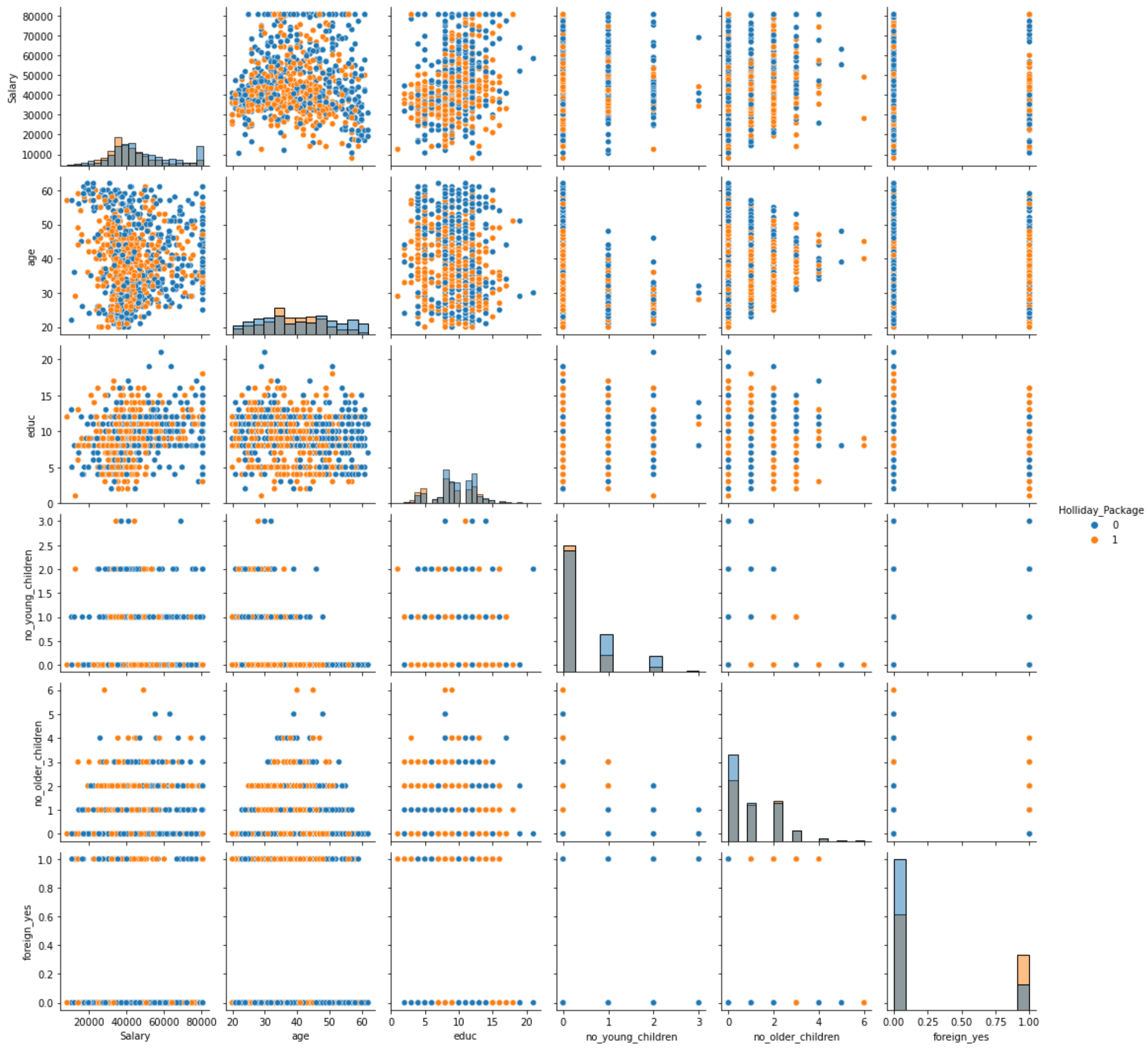


There is hardly any correlation between the variables.

Descriptive Statistics:

	count	mean	std	min	25%	50%	75%	max
Holliday_Package	872.0	0.459862	0.498672	0.00	0.0	0.0	1.0	1.00
Salary	872.0	45608.336869	15699.745151	8105.75	35324.0	41903.5	53469.5	80687.75
age	872.0	39.955275	10.551675	20.00	32.0	39.0	48.0	62.00
educ	872.0	9.307339	3.036259	1.00	8.0	9.0	12.0	21.00
no_young_children	872.0	0.311927	0.612870	0.00	0.0	0.0	0.0	3.00
no_older_children	872.0	0.982798	1.086786	0.00	0.0	1.0	2.0	6.00
foreign_yes	872.0	0.247706	0.431928	0.00	0.0	0.0	0.0	1.00

Pair plot using SNS.



2.2Build various iterations of the Logistic Regression model using appropriate variable selection techniques for the full data. Compare values of model selection criteria for proposed models. Compare as many criteria as you feel are suitable.

Use Full Data to develop a logistic regression model to identify significant predictors. Check whether the proposed model is free of multicollinearity. Apply variable selection method as required. Show all intermediate models leading to the final model. Justify your choice of the final model. Which are the significant predictors?

<u>Model 1</u>

Logit Regression Results			
Dep. Variable:	Holliday_Package	**No. Observations:**	872
Model:	Logit	**Df Residuals:**	865
Method:	MLE	**Df Model:**	6
Date:	Sat, 24 Sep 2022	**Pseudo R-squ.:**	0.1244
Time:	18:20:55	**Log-Likelihood:**	-526.78
converged:	True	**LL-Null:**	-601.61
Covariance Type:	nonrobust	**LLR p-value:**	9.138e-30

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	2.5432	0.559	4.550	0.000	1.448	3.639
Salary	-2.088e-05	5.26e-06	-3.970	0.000	-3.12e-05	-1.06e-05
age	-0.0496	0.009	-5.491	0.000	-0.067	-0.032
educ	0.0342	0.029	1.172	0.241	-0.023	0.091
no_young_children	-1.3287	0.180	-7.386	0.000	-1.681	-0.976
no_older_children	-0.0251	0.074	-0.341	0.733	-0.169	0.119
foreign_yes	1.3037	0.200	6.519	0.000	0.912	1.696

Check for multicollinearity in the predictor variables using Variance Inflation Factor (VIF).

Check for Multicollinearity:

```
Holliday Package VIF = 1.19
Salary VIF = 1.22
age VIF = 1.62
educe VIF = 1.41
no_young_children VIF = 1.69
no_older_children VIF = 1.19
foreign_yes VIF = 1.34
```

Model 2

Note: Threshold value considered is VIF < 1.5

Logit Regression Results			
Dep. Variable:	Holliday Package	**No. Observations:**	872
Model:	Logit	**Df Residuals:**	866
Method:	MLE	**Df Model:**	5
Date:	Sat, 24 Sep 2022	**Pseudo R-squ.:**	0.09790
Time:	18:28:17	**Log-Likelihood:**	-542.72
converged:	True	**LL-Null:**	-601.61
Covariance Type:	nonrobust	**LLR p-value:**	9.214e-24

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	0.0723	0.323	0.224	0.823	-0.561	0.706
Salary	-2.325e-05	5.16e-06	-4.503	0.000	-3.34e-05	-1.31e-05
educ	0.0654	0.028	2.312	0.021	0.010	0.121
no_young_children	-0.7949	0.140	-5.675	0.000	-1.069	-0.520
no_older_children	0.1029	0.068	1.502	0.133	-0.031	0.237
foreign_yes	1.3914	0.197	7.079	0.000	1.006	1.777

Model 3

Note: Threshold value considered is VIF < 1.5

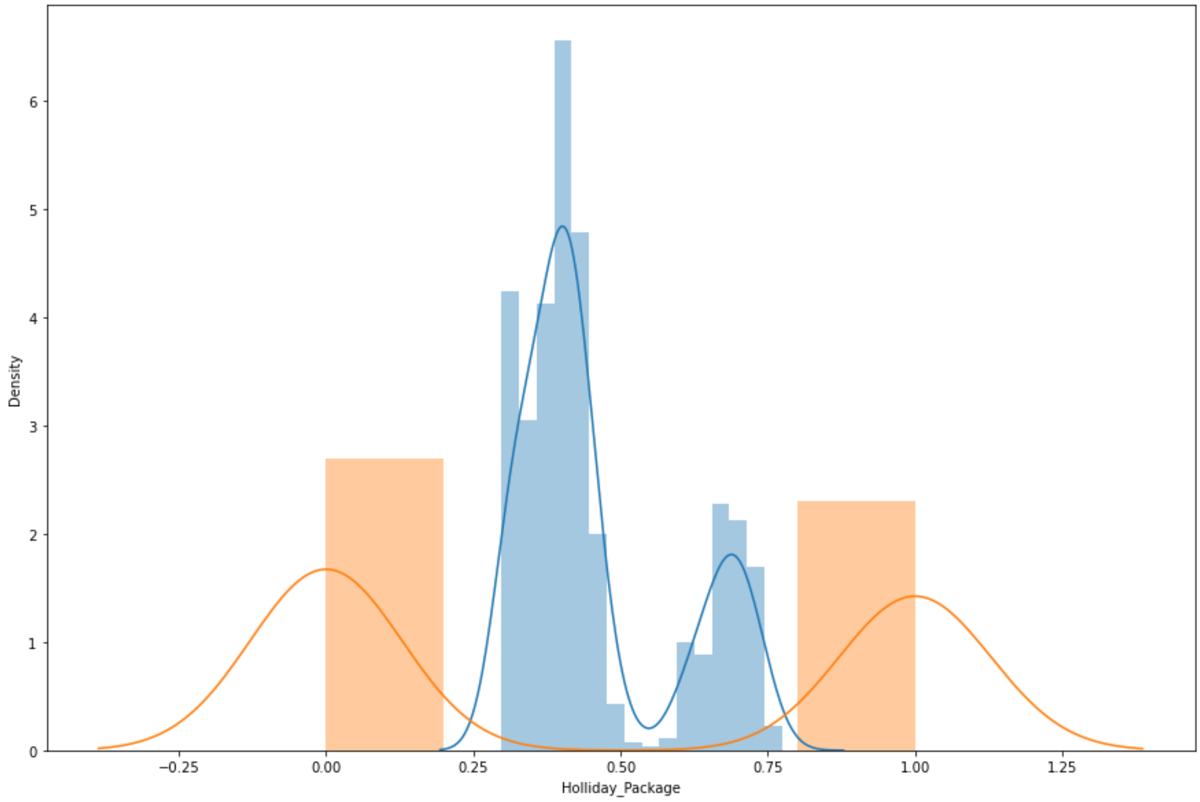
Logit Regression Results			
Dep. Variable:	Holliday Package	**No. Observations:**	872
Model:	Logit	**Df Residuals:**	867
Method:	MLE	**Df Model:**	4
Date:	Sat, 24 Sep 2022	**Pseudo R-squ.:**	0.09602
Time:	18:30:07	**Log-Likelihood:**	-543.85
converged:	True	**LL-Null:**	-601.61
Covariance Type:	nonrobust	**LLR p-value:**	4.807e-24

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	0.1467	0.319	0.459	0.646	-0.479	0.773
Salary	-2.213e-05	5.09e-06	-4.346	0.000	-3.21e-05	-1.21e-05
educ	0.0639	0.028	2.265	0.023	0.009	0.119
no_young_children	-0.8367	0.137	-6.093	0.000	-1.106	-0.568
foreign_yes	1.4043	0.196	7.148	0.000	1.019	1.789

Check for Multicollinearity:

```
Holliday_Package  VIF =  1.14
Salary  VIF =  1.19
educ  VIF =  1.36
no_young_children  VIF =  1.08
foreign_yes  VIF =  1.33
```

Using the best model:



2.3 Split the data into training (70%) and test (30%). Build the various iterations of the Logistic Regression models on the training data and use those models to predict on the test data using appropriate model evaluation metrics.

If prediction accuracy of the full scholarship is the only objective, then you may want to divide the data into a training and a test set, chosen randomly, and use the training set to develop a model and test set to validate your model. Use the models developed in Part (II) to compare accuracy in training and test sets. Compare the final model of Part (II) and the proposed one in Part (III). Which model provides the most accurate prediction? If the model found in Part (II) is different from the proposed model in Part (III), give an explanation.

Shape of Train data:

- Number of rows: **610**
- Number of columns: **7**

Shape of Test data:

- Number of rows: **262**
- Number of columns: **7**

value counts of Holliday Package:

Train data:

```
0    0.539344
1    0.460656
Name: Holliday_Package, dtype: float64
```

Test data:

```
0    0.541985
1    0.458015
Name: Holliday_Package, dtype: float64
```

Build the models 1,2 and 3 on the training data, check the accuracy score of each of the models on the training data and use those models to predict the classes and the corresponding probabilities on the test data.

Model 1 - Building the model on the Training Data and checking the Accuracy score on the training data.

Accuracy Score of Model 1: 0.6672131147540984

Model 2 - Building the model on the Training Data and checking the Accuracy score on the training data.

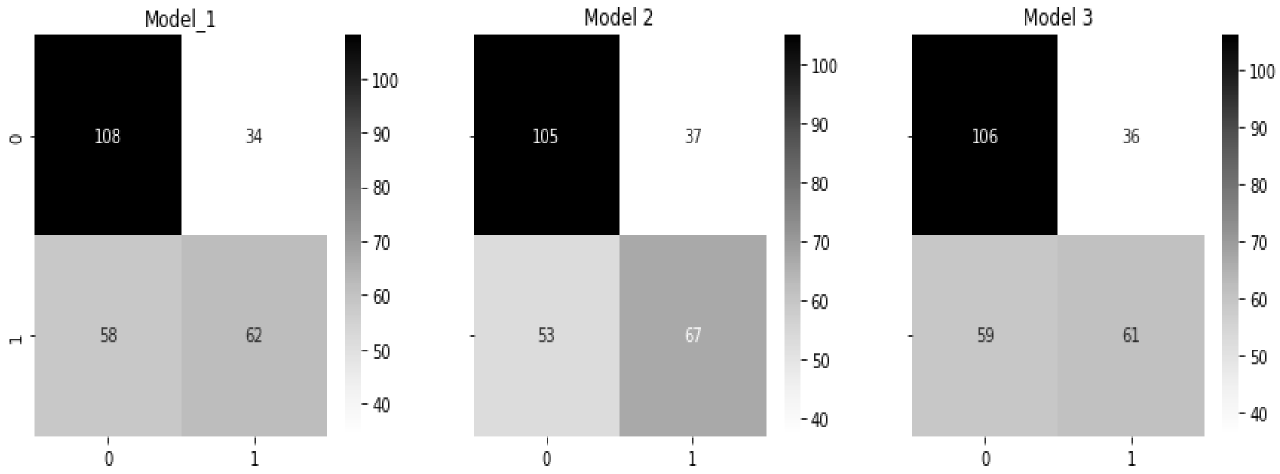
Accuracy Score of Model 2: 0.6491803278688525

Model 3 - Building the model on the Training Data and checking the Accuracy score on the training data.

Accuracy Score of Model 3: 0.6344262295081967

Evaluate the three models on the test data using the various statistics of the confusion matrix:

Confusion Matrix summary statistics Evaluation on the Test Data.



confusion matrix:

Model 1

```
True Negative: 108
False Positives: 34
False Negatives: 58
True Positives: 62
```

Model 2

```
True Negative: 105
False Positives: 37
False Negatives: 53
True Positives: 67
```

Model 3

```
True Negative: 106
False Positives: 36
False Negatives: 59
True Positives: 61
```

Classification report:

Model 1

	precision	recall	f1-score	support
0	0.65	0.76	0.70	142
1	0.65	0.52	0.57	120
accuracy			0.65	262
macro avg	0.65	0.64	0.64	262
weighted avg	0.65	0.65	0.64	262

Model 2

	precision	recall	f1-score	support
0	0.66	0.74	0.70	142
1	0.64	0.56	0.60	120
accuracy			0.66	262
macro avg	0.65	0.65	0.65	262
weighted avg	0.66	0.66	0.65	262

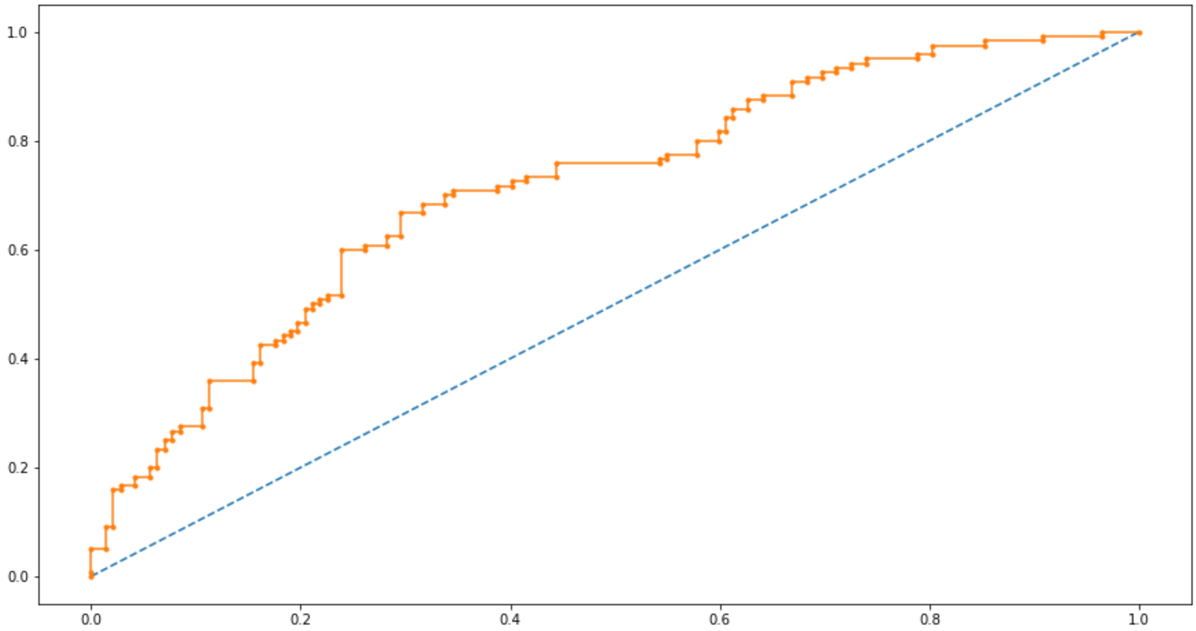
Model 3

	precision	recall	f1-score	support
0	0.64	0.75	0.69	142
1	0.63	0.51	0.56	120
accuracy			0.64	262
macro avg	0.64	0.63	0.63	262
weighted avg	0.64	0.64	0.63	262

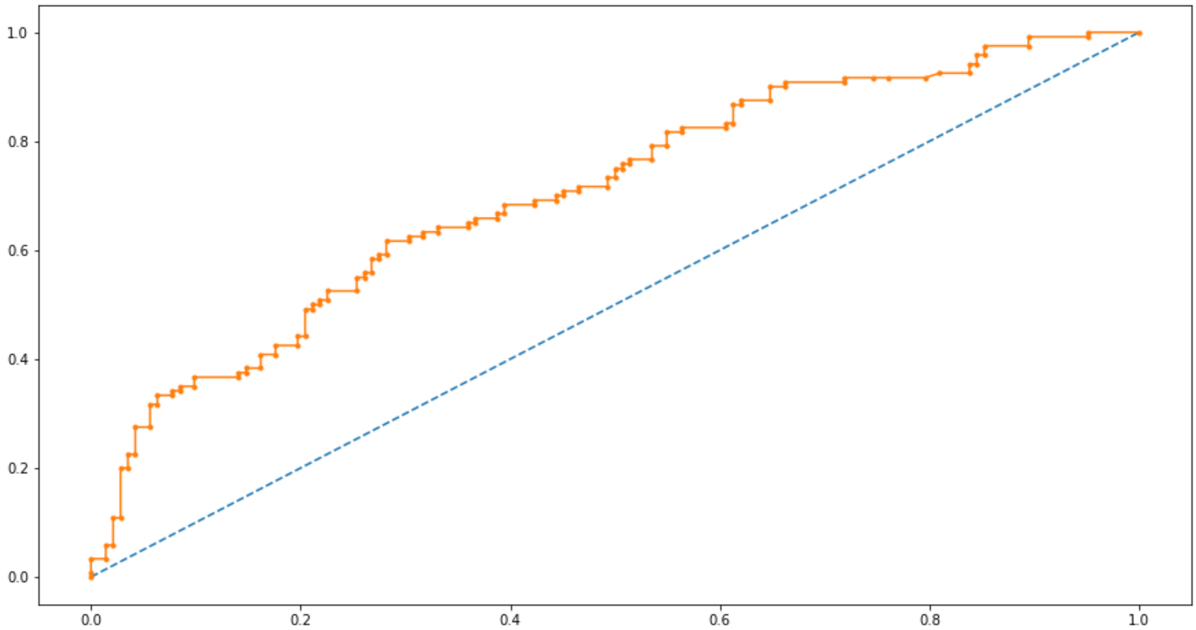
Check the summary statistics of the AUC-ROC curve for all the three Logistic Regression Models built. This is for the test data.

AUC and ROC:

Model 1 = 0.71496



Model 2 = 0.70628



<u>Model 3</u> = `0.70120`