DIAMOND ATTRIBUTE ASSOCIATION ANALYSIS

Benjamin McDaniel

Western Governors University

INTRODUCTION: SPEAKER QUALIFICATIONS & RELEVANT BACKGROUND

Benjamin McDaniel

MSDA candidate

Masters in Business Administration

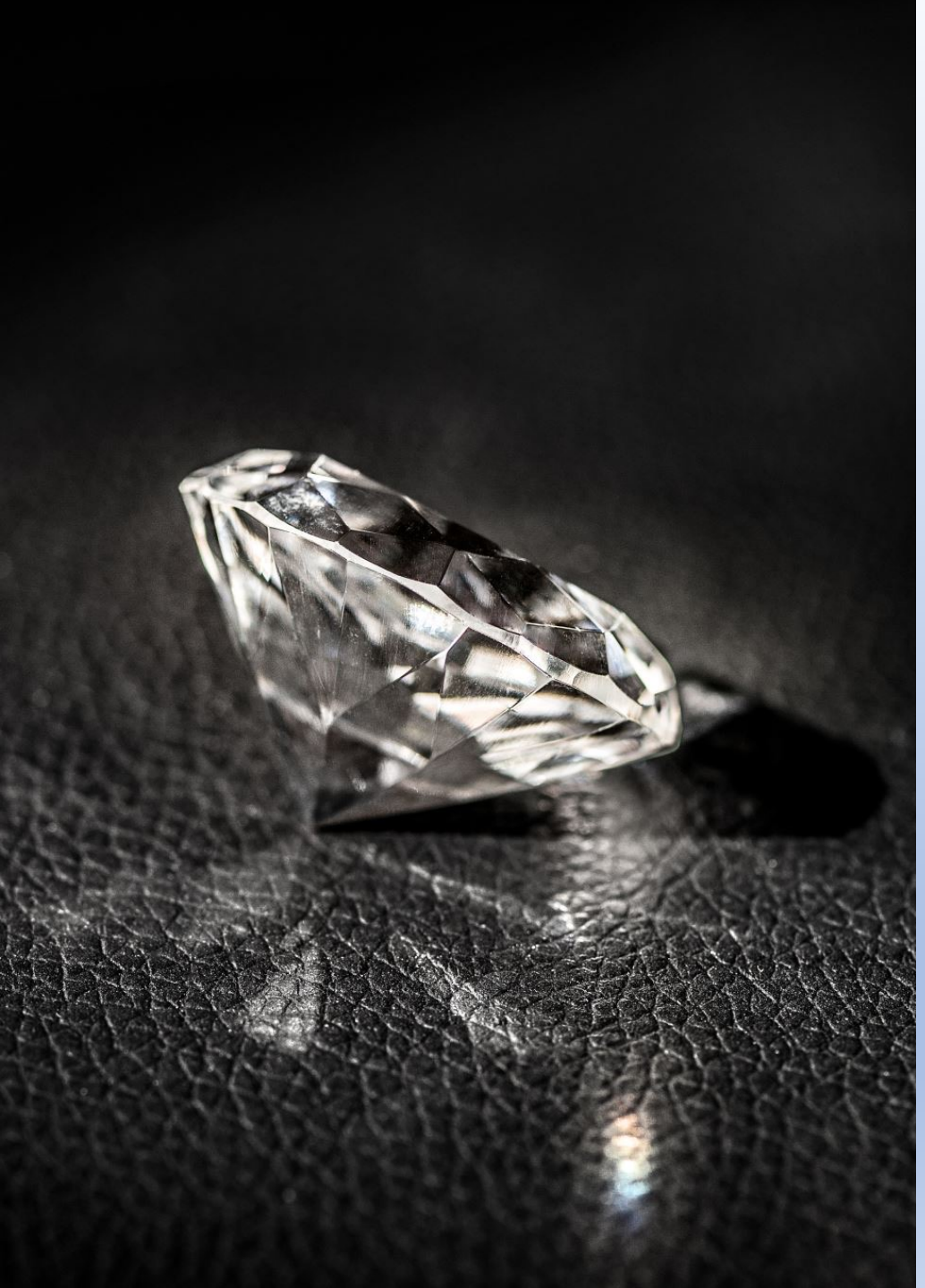
Certified Lean Six Sigma Black Belt

Graduate Gemologist

15+ years of industry experience manufacturing/evaluation

PROBLEM

- Historically the diamond industry has used a set of pre-defined attributes to describe rarity
- Portions of the diamond industry responsible for educating diamond experts on how to measure these attributes
- Industry education holds that there is no single attribute that can be held above others in determining value
- “It depends on your taste and budget, and what you and your partner deem most important”, (GIA.edu, 2019).
- Because budget is mentioned one could infer that there is an association between attributes and the price or value of a diamond



BUSINESS QUESTION:

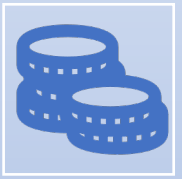
- Is there a statistically significant difference in the effect of diamond attributes on price determination for both natural and synthetic diamonds?

WHY IS THE ANSWER IMPORTANT?

The answer to this question will provide key insight into the effect of diamond attributes on the price of a diamond.

Allows for more informed decision making on purchasing items for inventory resulting in better ROI and or faster inventory turnover

Informs Inventory control on expected values of items coming into stock



Provides a deeper understanding of competitions pricing structure

Analysis will give clear indication of inventory and investment levels split by category that can be used as a benchmark moving forward

HYPOTHESIS:

- Null Hypothesis: There is no statistically significant difference in the effect of diamond attributes on price determination for both natural and synthetic diamonds
- Alternative Hypothesis: There is a statistically significant difference in the effect of diamond attributes on price determination for natural and synthetic diamonds

SUMMARY OF THE DATA ANALYSIS PROCESS

Data Acquisition

Exploratory Data Analysis

Feature Engineering

Data Transformation

T-test

Linear Regression Analysis

Final Determination

DIAMOND DATA

1

The data used in this analysis was provided by (Corral, 2020)

2

Available for use under attribution non-commercial international liscense

3

119,307 observations of diamond attributes and diamond prices

4

Loaded into a PostgreSQL database for initial exploration

DIAMOND ATTRIBUTES OF INTEREST

Price: Retail price in U.S. Dollars

Shape: Diamond industry nomenclature that describes physical shape of the diamond

Carat: Weight of the diamond in carats

Cut: Describes the efficiency of the proportions of the faceted diamond in relation to pre-defined recognized industry standards of such.

Color: The presence of body color of a diamond compared to industry standard labels for the presence of color

Clarity: A designated grade that describes the type, number, location, visibility, and effect of inclusions on the appearance, and durability of a diamond, as compared to an industry standard.

Report: Provides the name of the gemological laboratory that assigned the grades associated with each diamond in the data set.

Type: Indicates that the diamond is either of natural origin or originated in a laboratory that creates diamonds through man made means.

EXPLORATORY ANALYSIS SUMMARY

- PostgreSQL database actions:
- Initial exploration of summary statistics of the data set

	num_shapes bigint	num_colors bigint	num_cuts bigint	num_reports bigint	dia_types bigint
1	10	7	5	4	2

	min_price numeric	max_price numeric	average_price numeric	std_price numeric
1	270.00	1348720.00	3286.8432698835776610	9114.695376398422

EXPLORATORY ANALYSIS SUMMARY

- PostgreSQL database actions: Insights

	shape character varying (10)	min_price numeric	max_price numeric	average_price numeric	std_price numeric
1	Asscher	300.00	303330.00	7966.7230769230769231	17539.39962394
2	Cushion	280.00	792400.00	6622.7833605982706240	16499.46026537
3	Emerald	280.00	1348720.00	3964.2918518518518519	19571.88321595
4	Heart	470.00	150910.00	3507.7147016011644833	7606.994788941461
5	Marquise	290.00	62790.00	1884.6551724137931034	3228.611229820026
6	Oval	270.00	151850.00	3940.4993065187239945	6090.285310945712
7	Pear	280.00	336330.00	2203.7208545710877345	5499.559613975096
8	Princess	420.00	439380.00	3855.0048685491723466	12754.33809481
9	Radiant	310.00	410290.00	5898.5000000000000000	13830.57361564
10	Round	300.00	728890.00	2970.8750000000000000	7212.298229264892

EXPLORATORY ANALYSIS SUMMARY

	shape character varying (10)	highest_price numeric	rank bigint
1	Emerald	1348720.00	1
2	Cushion	792400.00	1
3	Round	728890.00	1
4	Princess	439380.00	1
5	Radiant	410290.00	1
6	Pear	336330.00	1
7	Asscher	303330.00	1
8	Oval	151850.00	1
9	Heart	150910.00	1
10	Marquise	62790.00	1

	color character varying (1)	highest_price numeric	rank bigint
1	D	1348720.00	1
2	F	648290.00	1
3	E	612570.00	1
4	G	563000.00	1
5	H	303330.00	1
6	J	280860.00	1
7	I	241870.00	1

EXPLORATORY ANALYSIS SUMMARY

	cut character varying (11)	highest_price numeric	rank bigint
1	Super Ideal	1348720.00	1
2	Good	410290.00	1
3	Ideal	336330.00	1
4	Very Good	167210.00	1
5	Fair	126030.00	1

	clarity character varying (4)	highest_price numeric	rank bigint
1	FL	1348720.00	1
2	VVS2	728890.00	1
3	VVS1	648290.00	1
4	VS1	563000.00	1
5	IF	437940.00	1
6	VS2	428920.00	1
7	SI2	194660.00	1
8	SI1	120150.00	1

EXPLORATORY ANALYSIS SUMMARY

	report character varying (4)	highest_price numeric	rank bigint
1	GIA	1348720.00	1
2	IGI	145630.00	1
3	GCAL	99040.00	1
4	HRD	41890.00	1

	dia_type character varying (7)	highest_price numeric	rank bigint
1	natural	1348720.00	1
2	lab	141710.00	1

EXPLORATORY ANALYSIS SUMMARY

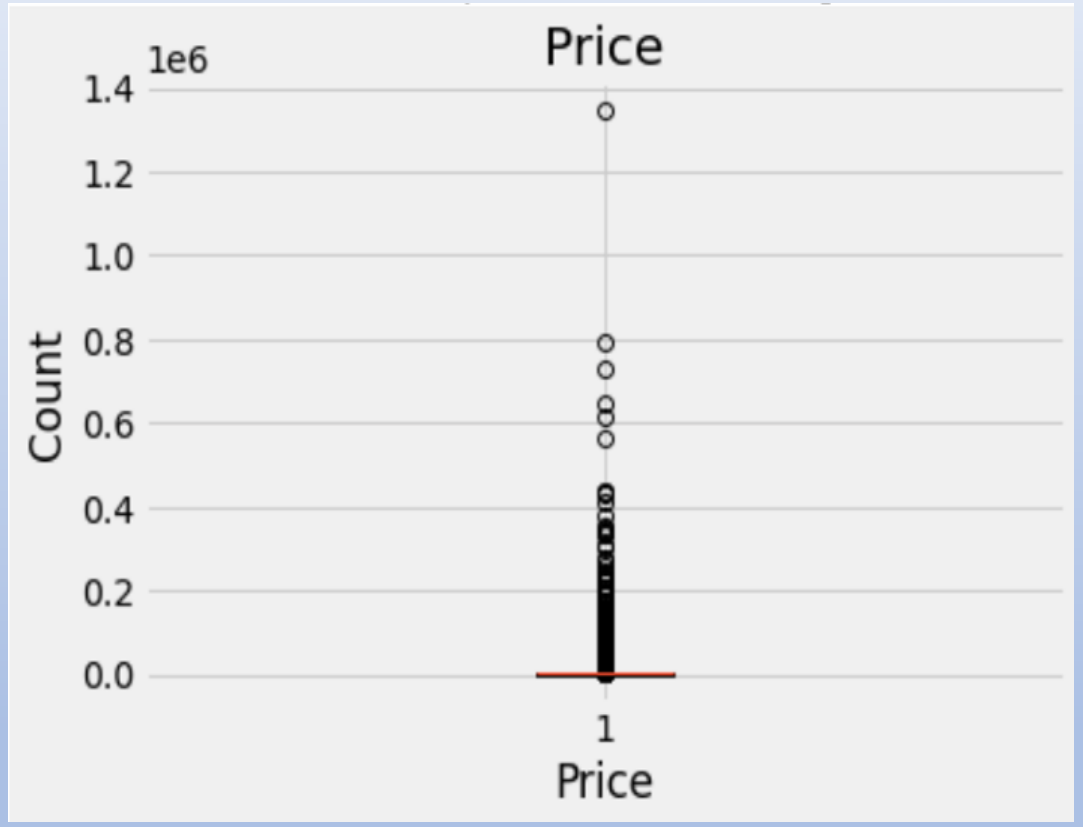
- Addition of price per carat and normalized price features.

	id integer	znorm_price numeric	price numeri	price_per_ct. numeric	shape character va	carat numeri	cut character v	color charact	clarity characte	report characte	dia_type character
1	10029850	-0.30904414832968606205	470.00	1021.74	Oval	0.46	Very Good	F	VS1	IGI	lab
2	9896069	-0.30904414832968606205	470.00	1021.74	Marquise	0.46	Ideal	G	VS1	GCAL	lab
3	9895135	-0.30904414832968606205	470.00	1119.05	Round	0.42	Ideal	G	VS1	GCAL	lab
4	10067472	-0.30904414832968606205	470.00	1021.74	Round	0.46	Ideal	I	SI1	IGI	lab
5	9960447	-0.30904414832968606205	470.00	1093.02	Round	0.43	Ideal	H	VS1	IGI	lab

EXPLORATORY DATA ANALYSIS SUMMARY: PRICE

```
univariate_continuous(df1['price'])
```

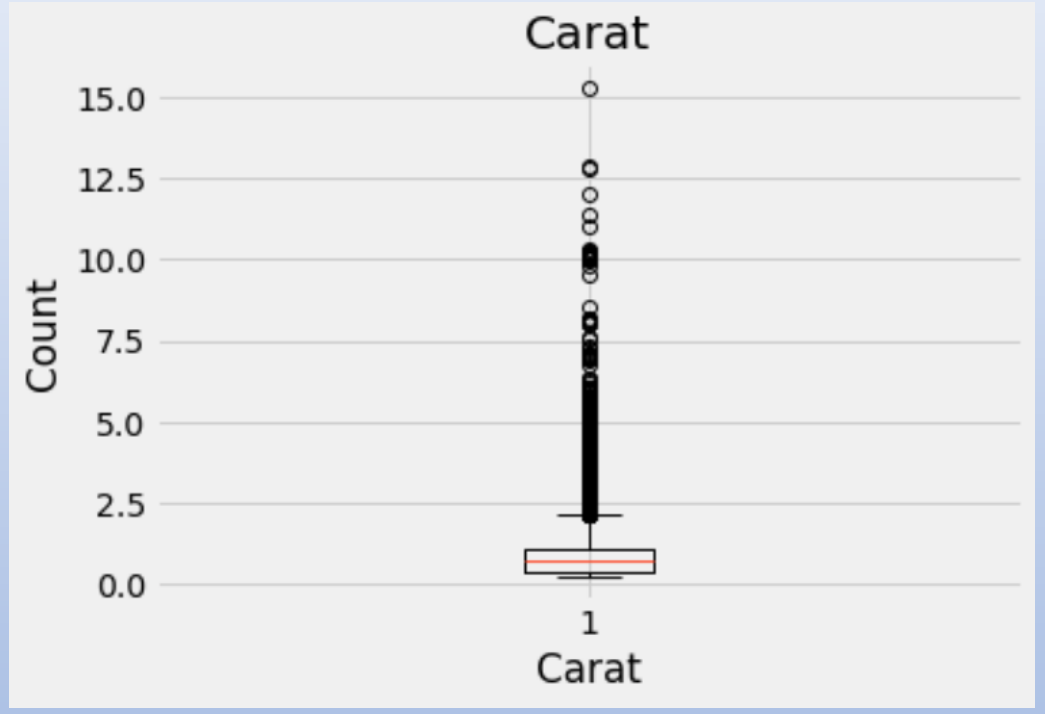
```
'The Mean for Price is: 3286.843'
'The Median for Price is: 1770.0'
'The Mode of Price is: ModeResult(mode=array([780.]), count=array([704]))'
'The Variance for Price is: 83077671.805'
'The Standard Deviation for Price is: 9114.657'
'The Minimum for Price is: 270.0'
'The Maximum for Price is: 1348720.0'
'The Range for Price is: 1348450.0'
'The First Quartile for Price is: 900.0'
'The Second Quartile for Price is: 1770.0'
'The Third Quartile for Price is: 3490.0'
'The IQR for Price is: 2590.0'
'The Skewness of Price is: 54.86'
'The Kurtosis of Price is: 5642.026'
'Outliers have values above 30630.814, and below -24057.127999999997'
'Anderson Darling Test for Normality Results Below: '
'Statistic: 23786.981'
'Indicates Not Gaussian Distribution: Critical Value 0.576 at Significance Level 15.0'
'Indicates Not Gaussian Distribution: Critical Value 0.656 at Significance Level 10.0'
'Indicates Not Gaussian Distribution: Critical Value 0.787 at Significance Level 5.0'
'Indicates Not Gaussian Distribution: Critical Value 0.918 at Significance Level 2.5'
'Indicates Not Gaussian Distribution: Critical Value 1.092 at Significance Level 1.0'
'End of Anderson Darling Test for Normality Results'
```



EXPLORATORY DATA ANALYSIS SUMMARY: CARAT

```
univariate_continuous(df1['carat'])
```

```
'The Mean for Carat is: 0.884'
'The Median for Carat is: 0.7'
'The Mode of Carat is: ModeResult(mode=array([0.3]), count=array([11422]))'
'The Variance for Carat is: 0.45'
'The Standard Deviation for Carat is: 0.671'
'The Minimum for Carat is: 0.25'
'The Maximum for Carat is: 15.32'
'The Range for Carat is: 15.07'
'The First Quartile for Carat is: 0.4'
'The Second Quartile for Carat is: 0.7'
'The Third Quartile for Carat is: 1.1'
'The IQR for Carat is: 0.7'
'The Skewness of Carat is: 2.551'
'The Kurtosis of Carat is: 16.433'
'Outliers have values above 2.897, and below -1.129'
'Anderson Darling Test for Normality Results Below: '
'Statistic: 6086.278'
'Indicates Not Gaussian Distribution: Critical Value 0.576 at Significance Level 15.0'
'Indicates Not Gaussian Distribution: Critical Value 0.656 at Significance Level 10.0'
'Indicates Not Gaussian Distribution: Critical Value 0.787 at Significance Level 5.0'
'Indicates Not Gaussian Distribution: Critical Value 0.918 at Significance Level 2.5'
'Indicates Not Gaussian Distribution: Critical Value 1.092 at Significance Level 1.0'
'End of Anderson Darling Test for Normality Results'
```



EXPLORATORY DATA ANALYSIS SUMMARY

- Initial univariate analysis indicated that the price variable contained outliers and was not of normal distribution. This would influence analysis negatively.
- 636 observations were identified as outliers and removed.

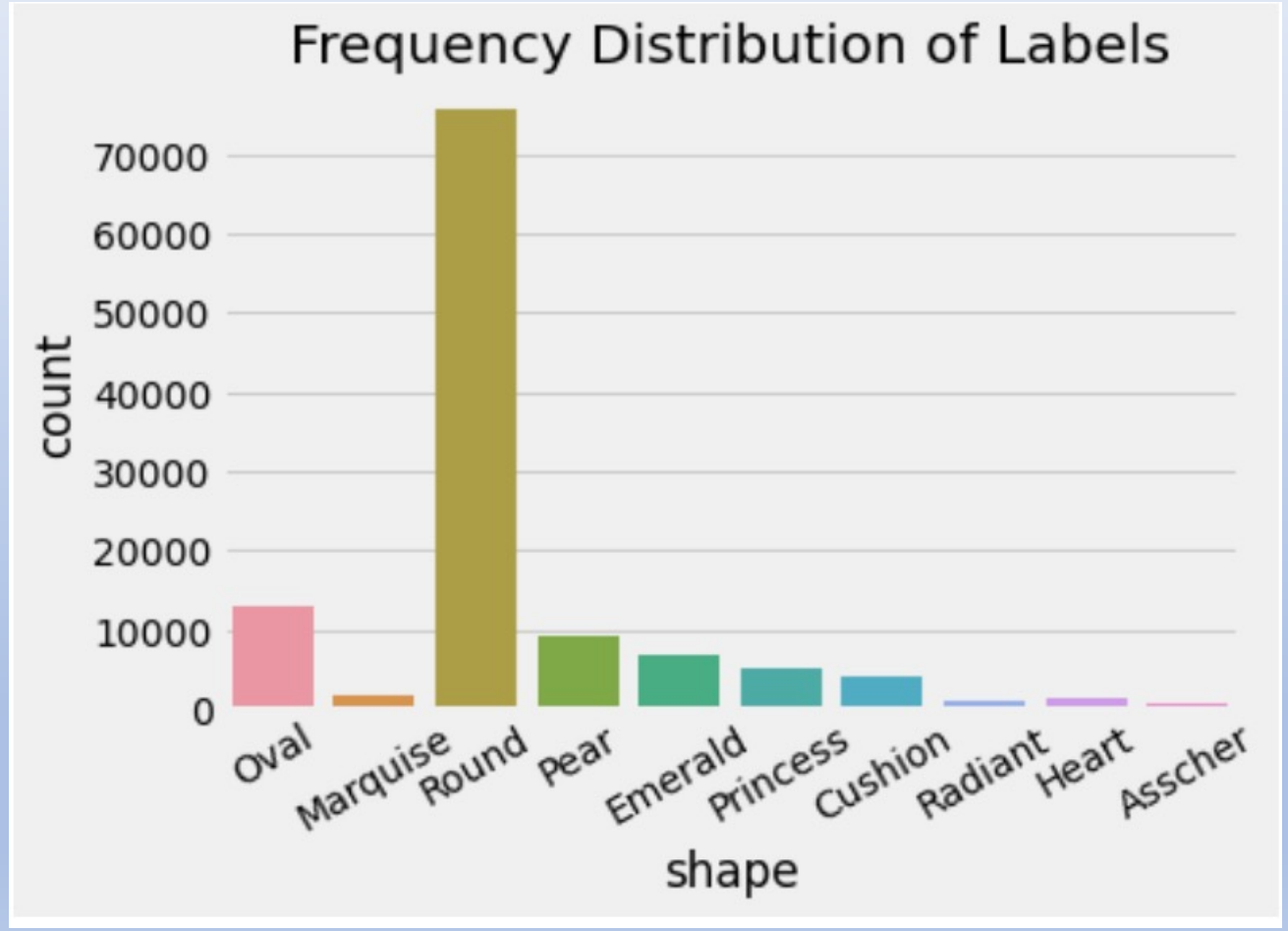
```
univariate_continuous(df1['price'])
```

```
'The Mean for Price is: 2930.626'
'The Median for Price is: 1760.0'
'The Mode of Price is: ModeResult(mode=array([780.]), count=array([704]))'
'The Variance for Price is: 12760748.114'
'The Standard Deviation for Price is: 3572.204'
'The Minimum for Price is: 270.0'
'The Maximum for Price is: 30600.0'
'The Range for Price is: 30330.0'
'The First Quartile for Price is: 900.0'
'The Second Quartile for Price is: 1760.0'
'The Third Quartile for Price is: 3440.0'
'The IQR for Price is: 2540.0'
'The Skewness of Price is: 3.302'
'The Kurtosis of Price is: 14.198'
'Outliers have values above 13647.238000000001, and below -7785.986000000001'
'Anderson Darling Test for Normality Results Below: '
'Statistic: 11810.707'
'Indicates Not Gaussian Distribution: Critical Value 0.576 at Significance Level 15.0'
'Indicates Not Gaussian Distribution: Critical Value 0.656 at Significance Level 10.0'
'Indicates Not Gaussian Distribution: Critical Value 0.787 at Significance Level 5.0'
'Indicates Not Gaussian Distribution: Critical Value 0.918 at Significance Level 2.5'
'Indicates Not Gaussian Distribution: Critical Value 1.092 at Significance Level 1.0'
'End of Anderson Darling Test for Normality Results'
```

EXPLORATORY DATA ANALYSIS SUMMARY: SHAPE

```
univariate_categorical(df1['shape'])
```

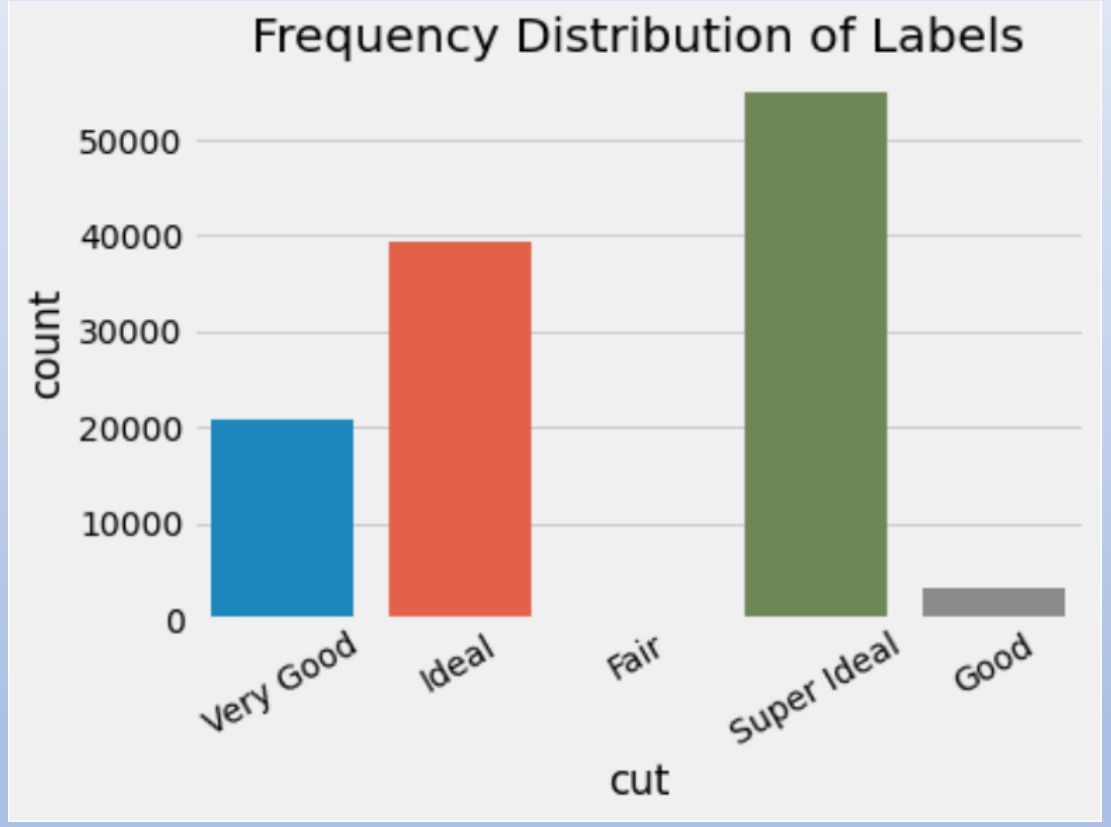
```
'The values of this variable include:'
['Oval',
 'Marquise',
 'Round',
 'Pear',
 'Emerald',
 'Princess',
 'Cushion',
 'Radiant',
 'Heart',
 'Asscher']
'The Frequency of occurance for each label is: '
Round       75739
Oval        12896
Pear         9201
Emerald      6693
Princess     5114
Cushion      4208
Marquise     1736
Heart        1363
Radiant      1092
Asscher       629
Name: shape, dtype: int64
'The table of proportions for this variable is below:'
Round       0.638227
Oval        0.108670
Pear        0.077534
Emerald     0.056400
Princess    0.043094
Cushion     0.035459
Marquise    0.014629
Heart       0.011486
Radiant     0.009202
Asscher     0.005300
Name: shape, dtype: float64
'The most frequent value of this variable is: Round'
```



EXPLORATORY DATA ANALYSIS SUMMARY: CUT

```
univariate_categorical(df1['cut'])
```

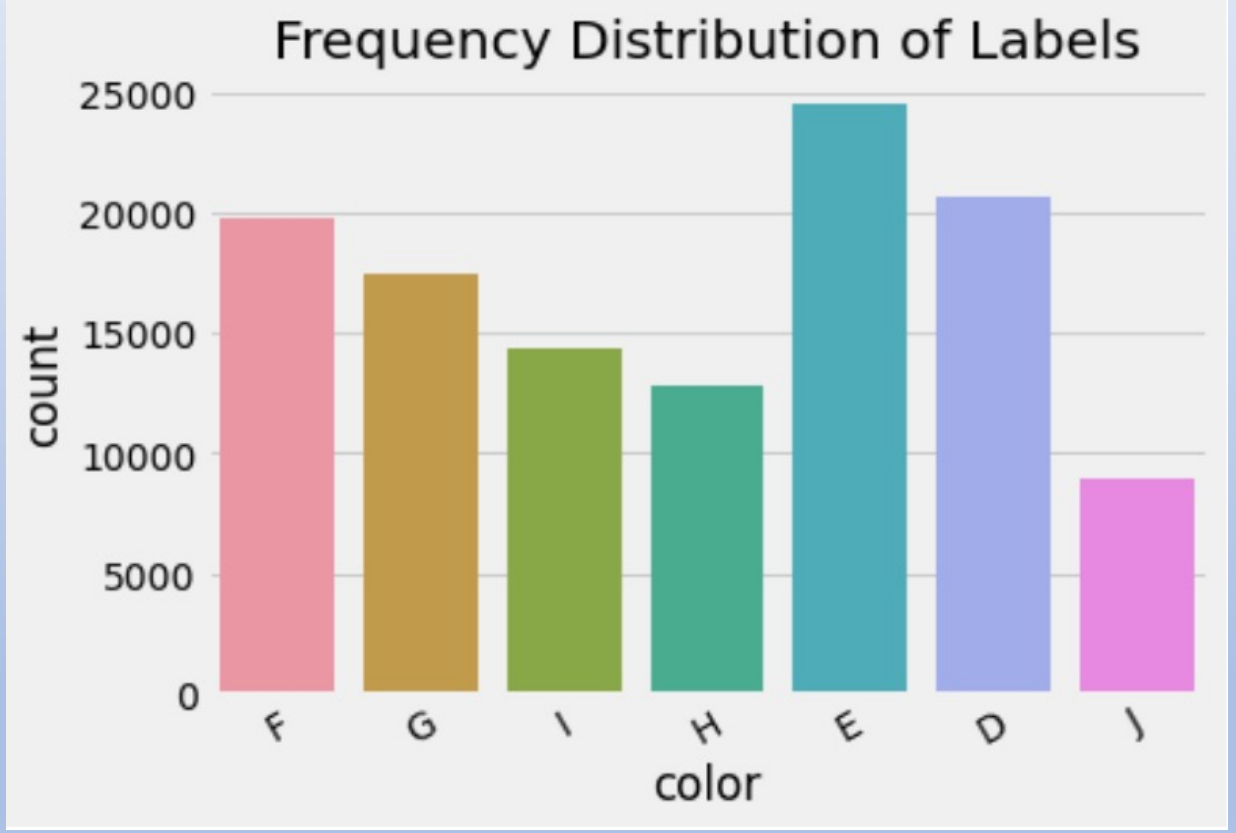
```
'The values of this variable include:'
['Very Good', 'Ideal', 'Fair', 'Super Ideal', 'Good']
'The Frequency of occurance for each label is: '
Super Ideal    54863
Ideal          39288
Very Good      20831
Good            3358
Fair             331
Name: cut, dtype: int64
'The table of proportions for this variable is below:'
Super Ideal    0.462312
Ideal          0.331067
Very Good      0.175536
Good           0.028297
Fair           0.002789
Name: cut, dtype: float64
'The most frequent value of this variable is: Super Ideal'
```



EXPLORATORY DATA ANALYSIS SUMMARY: COLOR

```
univariate_categorical(df1['color'])
```

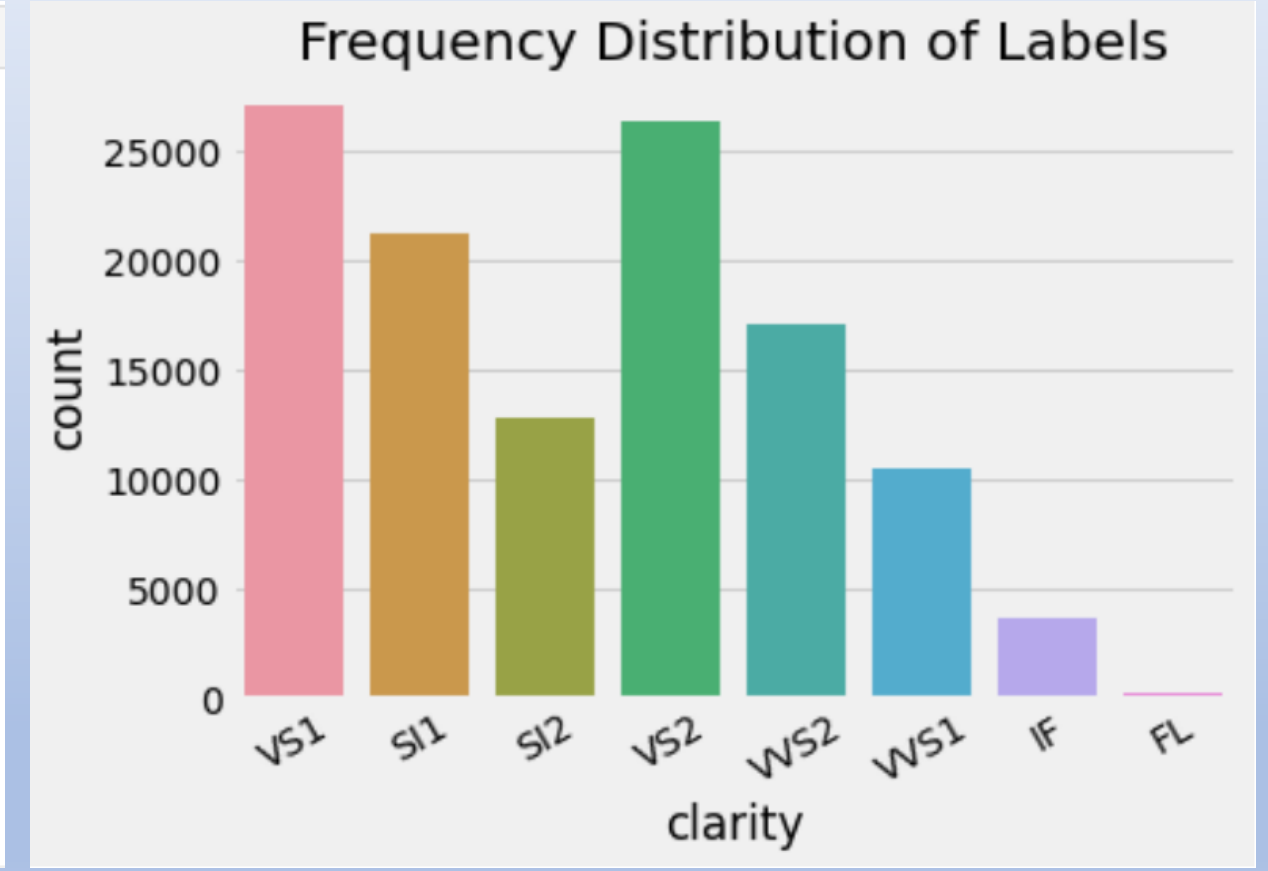
```
'The values of this variable include:'
['F', 'G', 'I', 'H', 'E', 'D', 'J']
'The Frequency of occurance for each label is: '
E    24594
D    20671
F    19745
G    17492
I    14369
H    12817
J     8983
Name: color, dtype: int64
'The table of proportions for this variable is below:'
E    0.207245
D    0.174187
F    0.166384
G    0.147399
I    0.121083
H    0.108004
J    0.075697
Name: color, dtype: float64
'The most frequent value of this variable is: E'
```



EXPLORATORY DATA ANALYSIS SUMMARY: CLARITY

```
univariate_categorical(df1['clarity'])
```

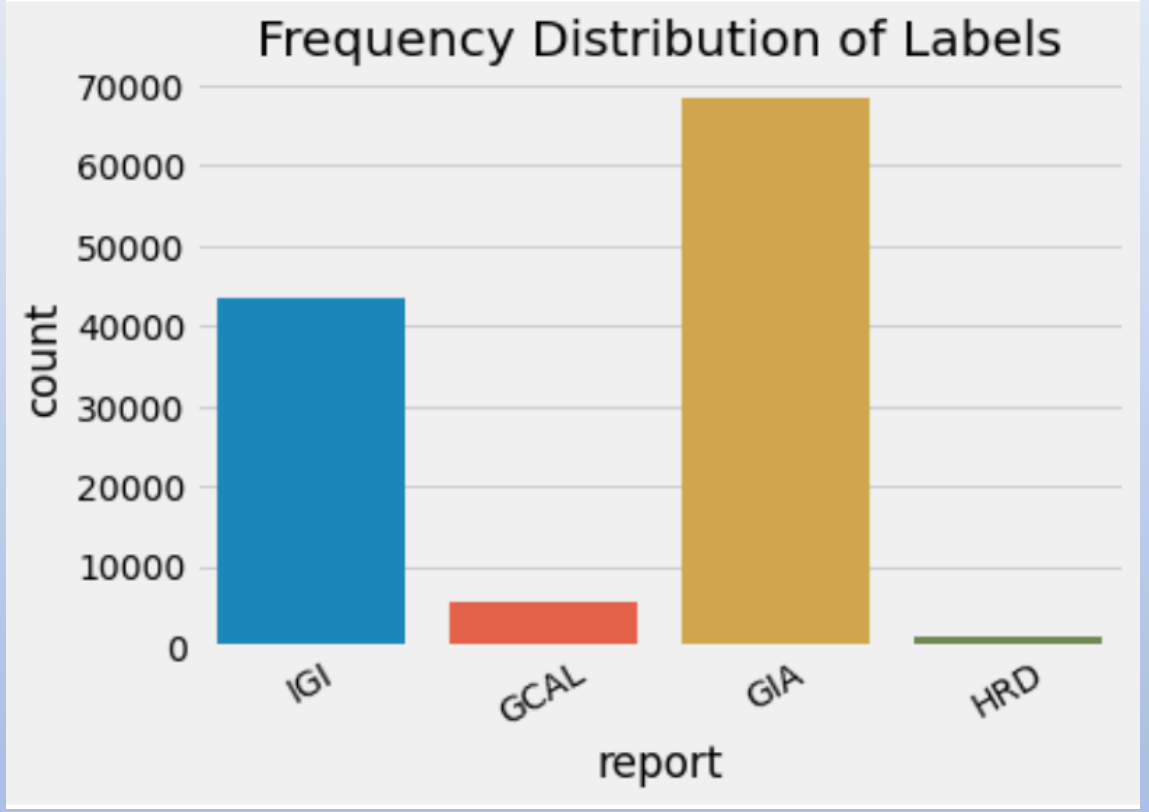
```
'The values of this variable include:'
['VS1', 'SI1', 'SI2', 'VS2', 'VVS2', 'VVS1', 'IF', 'FL']
'The Frequency of occurance for each label is: '
VS1     27120
VS2     26340
SI1     21154
VVS2    17106
SI2     12736
VVS1    10417
IF       3616
FL        182
Name: clarity, dtype: int64
'The table of proportions for this variable is below:'
VS1     0.228531
VS2     0.221958
SI1     0.178258
VVS2    0.144146
SI2     0.107322
VVS1    0.087781
IF      0.030471
FL      0.001534
Name: clarity, dtype: float64
'The most frequent value of this variable is: VS1'
```



EXPLORATORY DATA ANALYSIS SUMMARY: REPORT

```
univariate_categorical(df1['report'])
```

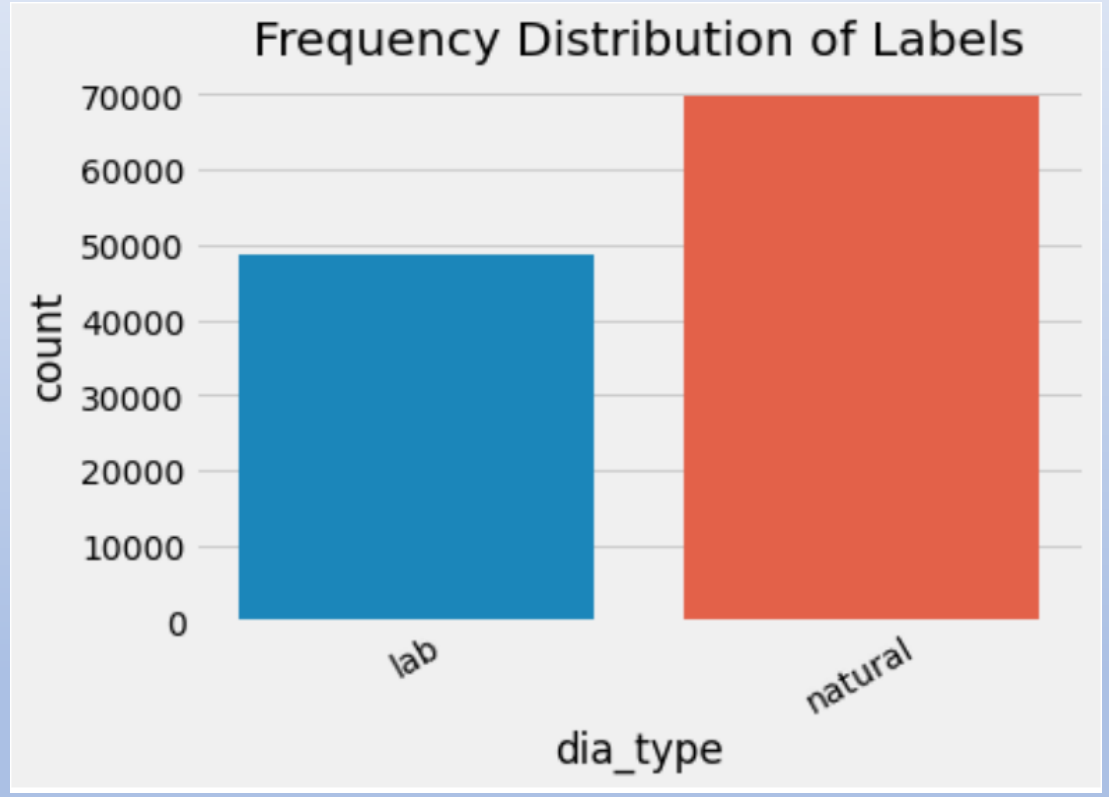
```
'The values of this variable include:'
['IGI', 'GCAL', 'GIA', 'HRD']
'The Frequency of occurance for each label is: '
GIA     68399
IGI     43378
GCAL     5743
HRD      1151
Name: report, dtype: int64
'The table of proportions for this variable is below:'
GIA     0.576375
IGI     0.365532
GCAL    0.048394
HRD     0.009699
Name: report, dtype: float64
'The most frequent value of this variable is: GIA'
```



EXPLORATORY DATA ANALYSIS SUMMARY: TYPE

```
univariate_categorical(df1['dia_type'])
```

```
'The values of this variable include:'
['lab', 'natural']
'The Frequency of occurance for each label is: '
natural    69921
lab        48750
Name: dia_type, dtype: int64
'The table of proportions for this variable is below:'
natural    0.5892
lab        0.4108
Name: dia_type, dtype: float64
'The most frequent value of this variable is: natural'
```

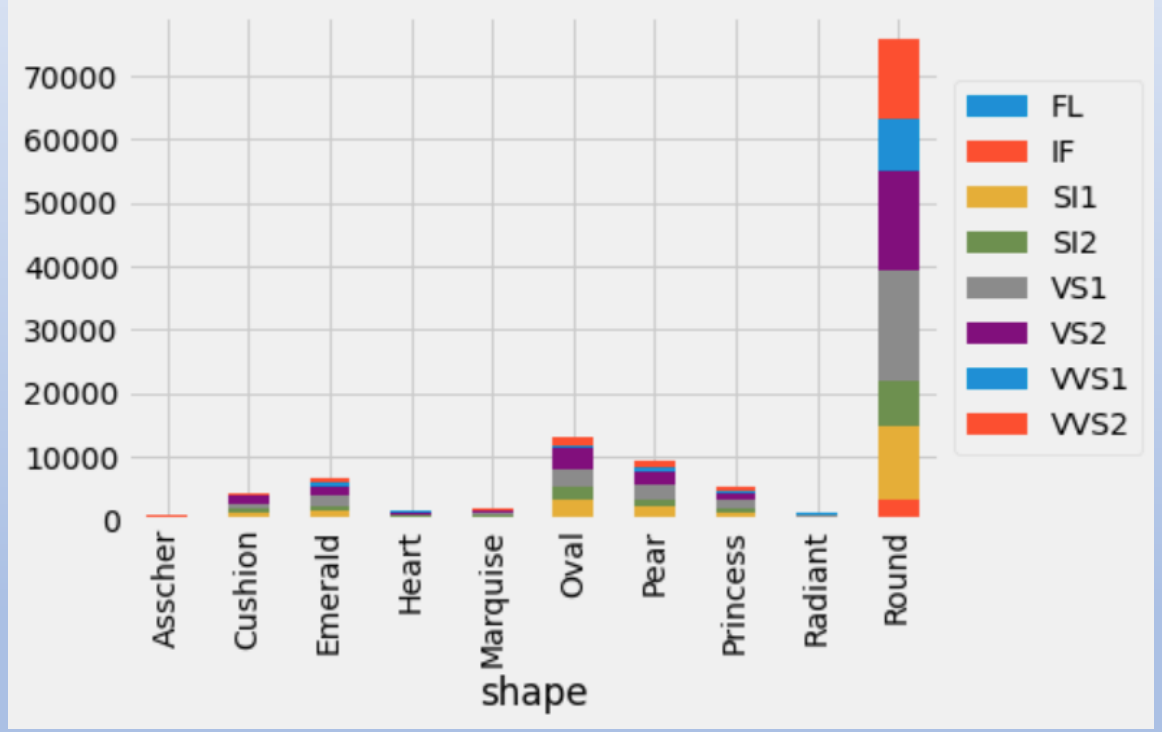


EXPLORATORY DATA ANALYSIS SUMMARY: BIVARIATE

- Bivariate analysis of the variables indicate that there is association between each pair of variables included in the data set.
- This was determined through chi-square test for categorical variable comparison. The resulting statistic and associated p-value indicate to reject the null hypotheisis that the variables are independent of one another.
- An example of the output can be found in the next slide.

EXPLORATORY DATA ANALYSIS SUMMARY: BIVARIATE EXAMPLE OUTPUT SHAPE VS CUT

```
Expected Values
[[2.0000e+00 1.8000e+01 2.0800e+02 2.9100e+02 1.1000e+02]
 [1.2000e+01 1.1900e+02 1.3930e+03 1.9450e+03 7.3900e+02]
 [1.9000e+01 1.8900e+02 2.2160e+03 3.0940e+03 1.1750e+03]
 [4.0000e+00 3.9000e+01 4.5100e+02 6.3000e+02 2.3900e+02]
 [5.0000e+00 4.9000e+01 5.7500e+02 8.0300e+02 3.0500e+02]
 [3.6000e+01 3.6500e+02 4.2690e+03 5.9620e+03 2.2640e+03]
 [2.6000e+01 2.6000e+02 3.0460e+03 4.2540e+03 1.6150e+03]
 [1.4000e+01 1.4500e+02 1.6930e+03 2.3640e+03 8.9800e+02]
 [3.0000e+00 3.1000e+01 3.6200e+02 5.0500e+02 1.9200e+02]
 [2.1100e+02 2.1430e+03 2.5075e+04 3.5015e+04 1.3295e+04]]

Observed values
cut         Fair  Good  Ideal  Super Ideal  Very Good
shape
Asscher        4    45    210           94        276
Cushion       19   284   1524          922       1459
Emerald       22   224   3463         1442       1542
Heart          4    86    681          122        470
Marquise      24   102   1053          191        366
Oval          12   395   6241         3042       3206
Pear          27   560   4934         1002       2678
Princess      22   229   2863         1073        927
Radiant       15   121    428          189        339
Round        182  1312  17891        46786       9568

H0: The two variables are independent.
H1: The two variables are not independent.

The Chi-Square statistic is: 22472.84676982554

The P_value is: 0.0

Determination:
Reject the null hypothesis (H0) that the two variables are independent
```



EXPLORATORY DATA ANALYSIS SUMMARY: BIVARIATE EXAMPLE OUTPUT SHAPE VS CUT

- Contingency table
- Table of proportions
- Marginal proportions by variable
- Expanded findings can be found in (McDaniel, 2022)

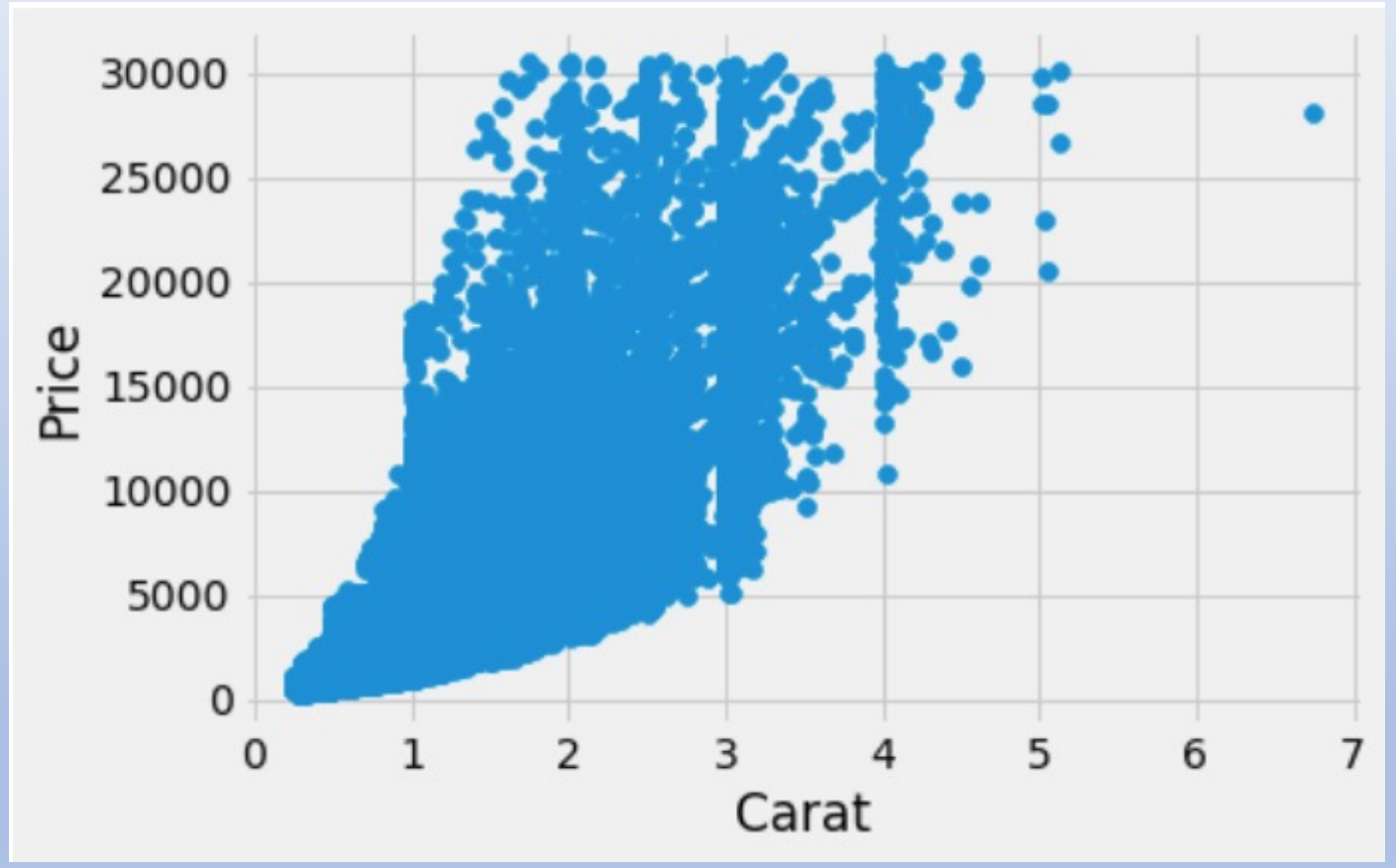
```
bivar_categorical(df1['shape'], df1['cut'])
```

```
Contingency Table
cut        Fair  Good  Ideal  Super Ideal  Very Good
shape
Asscher       4    45    210           94        276
Cushion      19   284   1524          922       1459
Emerald      22   224   3463         1442       1542
Heart         4    86    681          122        470
Marquise     24   102   1053          191        366
Oval         12   395   6241         3042       3206
Pear         27   560   4934         1002       2678
Princess     22   229   2863         1073        927
Radiant      15   121    428          189        339
Round       182  1312  17891        46786       9568

Table of Proportions
cut            Fair      Good     Ideal  Super Ideal  Very Good
shape
Asscher    0.000034  0.000379  0.001770     0.000792   0.002326
Cushion    0.000160  0.002393  0.012842     0.007769   0.012294
Emerald    0.000185  0.001888  0.029182     0.012151   0.012994
Heart      0.000034  0.000725  0.005739     0.001028   0.003961
Marquise   0.000202  0.000860  0.008873     0.001609   0.003084
Oval       0.000101  0.003329  0.052591     0.025634   0.027016
Pear       0.000228  0.004719  0.041577     0.008444   0.022567
Princess   0.000185  0.001930  0.024126     0.009042   0.007812
Radiant    0.000126  0.001020  0.003607     0.001593   0.002857
Round      0.001534  0.011056  0.150761     0.394250   0.080626

The marginal proportion of Cut by type is:  cut
Fair          0.002789
Good          0.028297
Ideal         0.331067
Super Ideal   0.462312
Very Good     0.175536
dtype: float64

The marginal proportion of Shape by type is:  shape
Asscher     0.005300
Cushion     0.035459
Emerald     0.056400
Heart       0.011486
Marquise    0.014629
Oval        0.108670
Pear        0.077534
Princess    0.043094
Radiant     0.009202
Round       0.638227
dtype: float64
```

EXPLORATORY DATA ANALYSIS SUMMARY: BIVARIATE CONTINUOUS PRICE VS CARAT

```
bivar_continuous(df1['price'],df1['carat'])
```

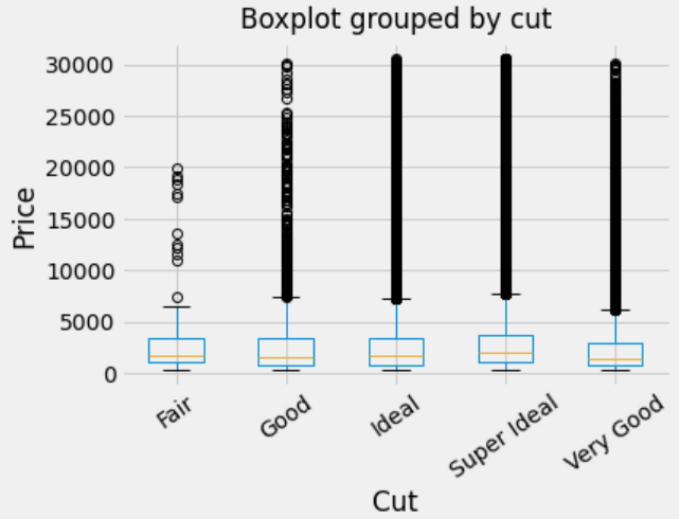
```
Pearson correlation coefficient: 0.7686705668748961
Preason correlation p-value: 0.0

The spearman rank order correlation coefficient is: 0.8314303282367526
The spearman rank order correlation p-value is: 0.0

Covariance Matrix for Price and Carat
[[1.27607481e+07 1.68552556e+03]
 [1.68552556e+03 3.76802736e-01]]

Model summary of simple linear regression:

                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.591
Model:                            OLS   Adj. R-squared:                  0.591
Method:                 Least Squares   F-statistic:                 1.714e+05
Date:                Wed, 16 Mar 2022   Prob (F-statistic):               0.00
Time:                        12:53:55   Log-Likelihood:            -1.0862e+06
No. Observations:              118671   AIC:                         2.172e+06
Df Residuals:                  118669   BIC:                         2.172e+06
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -944.9906     11.474    -82.362      0.000    -967.479    -922.502
carat       4473.2307     10.806    413.971      0.000    4452.052    4494.410
==============================================================================
Omnibus:                    67585.022   Durbin-Watson:                   0.209
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1045311.403
Skew:                           2.432   Prob(JB):                         0.00
Kurtosis:                      16.702   Cond. No.                         3.15
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Parameters:
 const    -944.990607
carat    4473.230694
dtype: float64

The fitted regression equation is:
Price = -944.9906069975741 + 4473.2306937261665 * Carat value
```



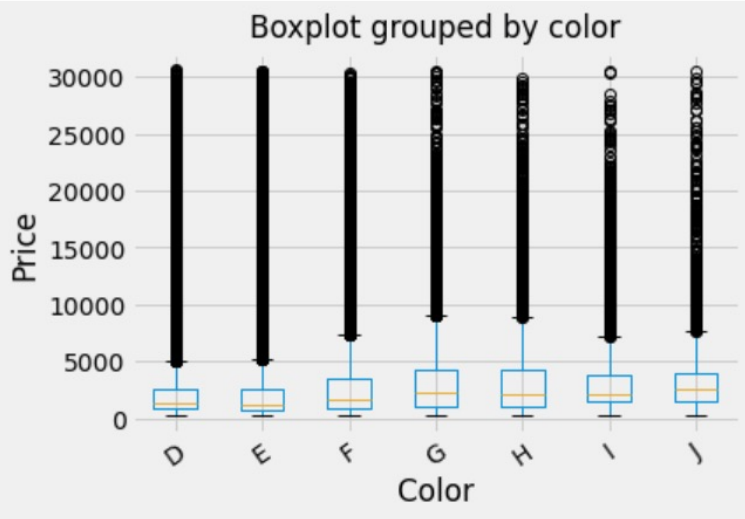
EXPLORATORY DATA ANALYSIS SUMMARY: BIVARIATE CONTINUOUS VS CATEGORICAL VISUALIZATION

- Visualizations of the bivariate relationship between categorical and continuous variable price were generated.
- Each instance indicates a right skewed distribution of price

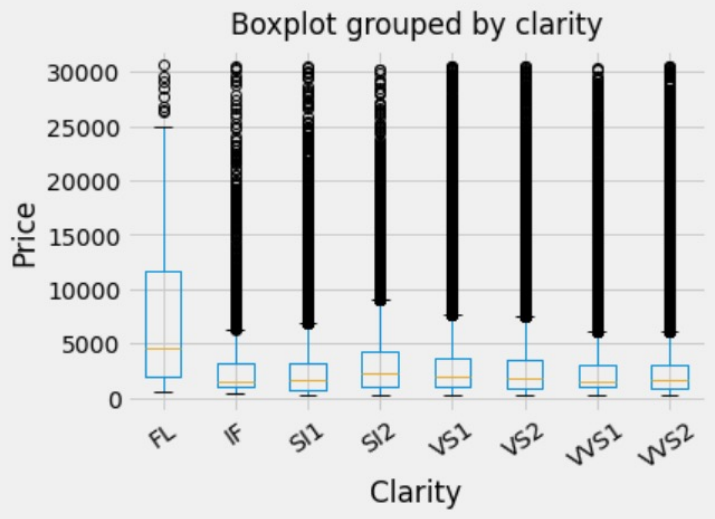
```
bivar_cont_cat(df1['price'], df1['cut'], df1)
```




```
bivar_cont_cat(df1['price'], df1['color'], df1)
```




```
bivar_cont_cat(df1['price'], df1['clarity'], df1)
```

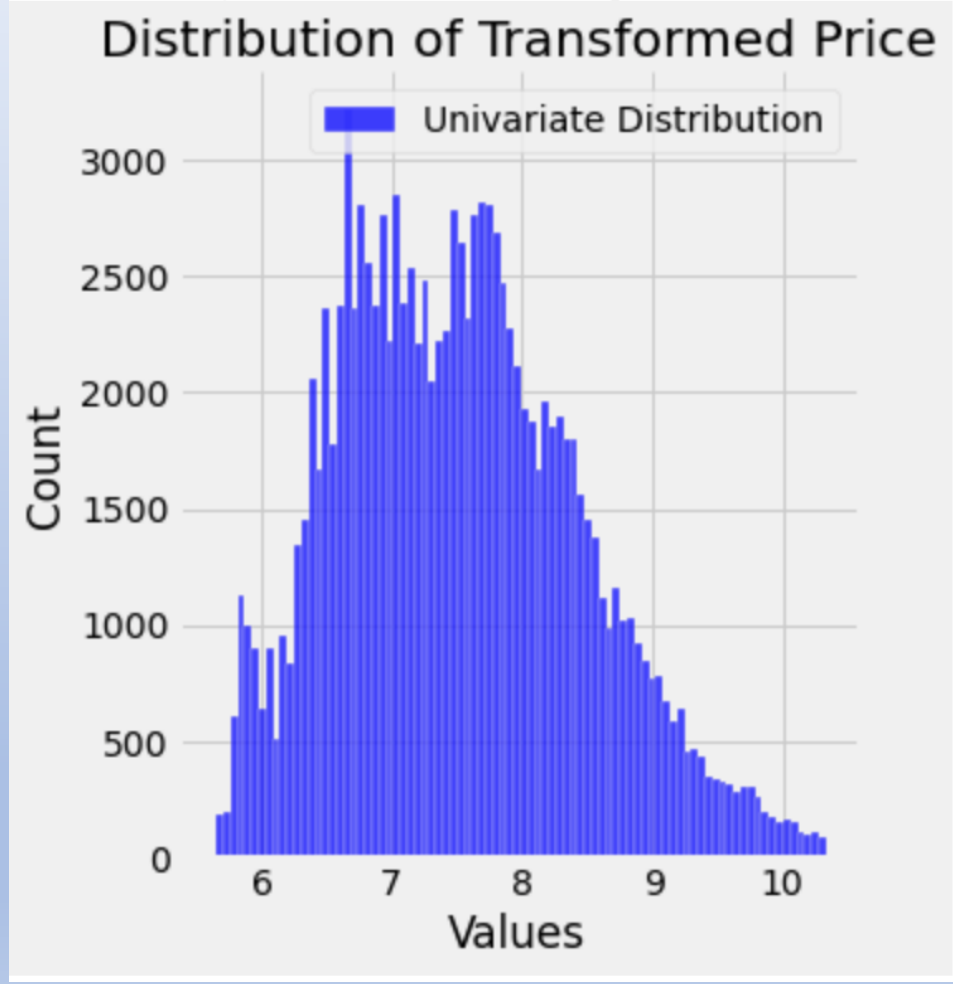


DATA REDUCTION & TRANSFORMATION

- Price variable was transformed to approach a natural distribution and named price_nat_log

```
transform_col(df1, df1['price'],0)
df1.head()
```

```
Natural Log Transformation Complete
```



- Natural and Synthetic diamond populations including all attributes

Dia Type
lab
natural

DIAMOND DASHBOARD
Use the filters across the top of the dashboard to produce different views

Carat
0.25 to 15.32
and Null values

Color
All

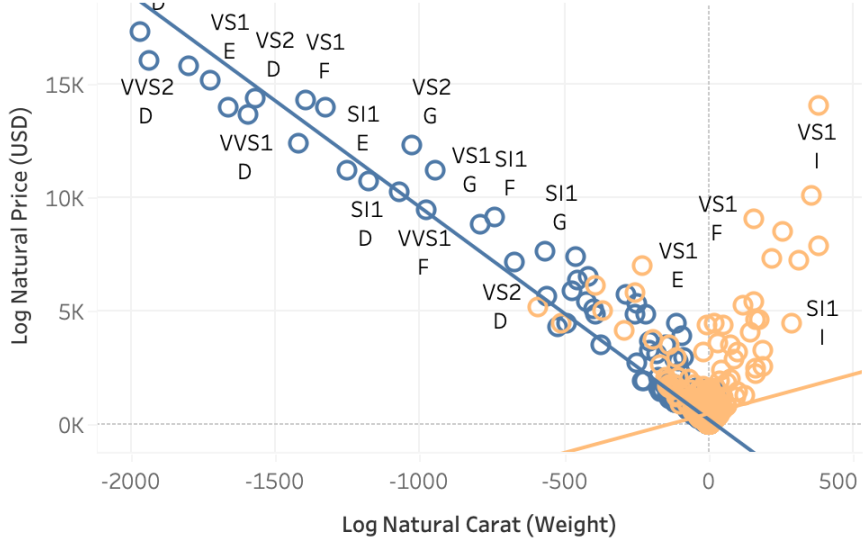
Cut
All

Clarity
All

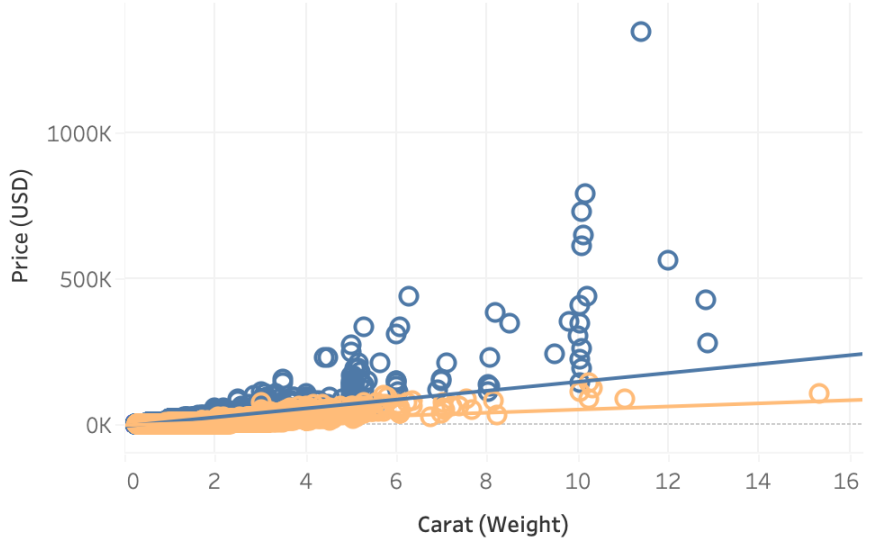
Dia Type
All

Shape
All

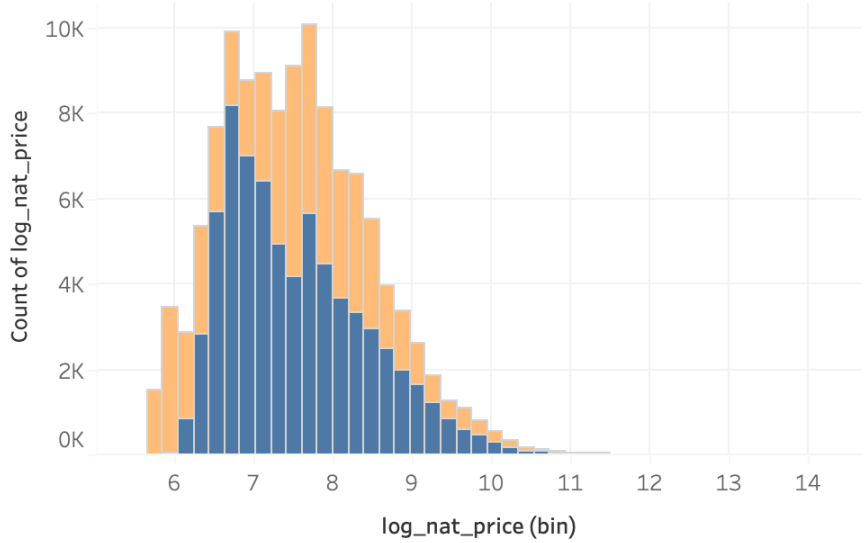
LN Carat vs LN Price



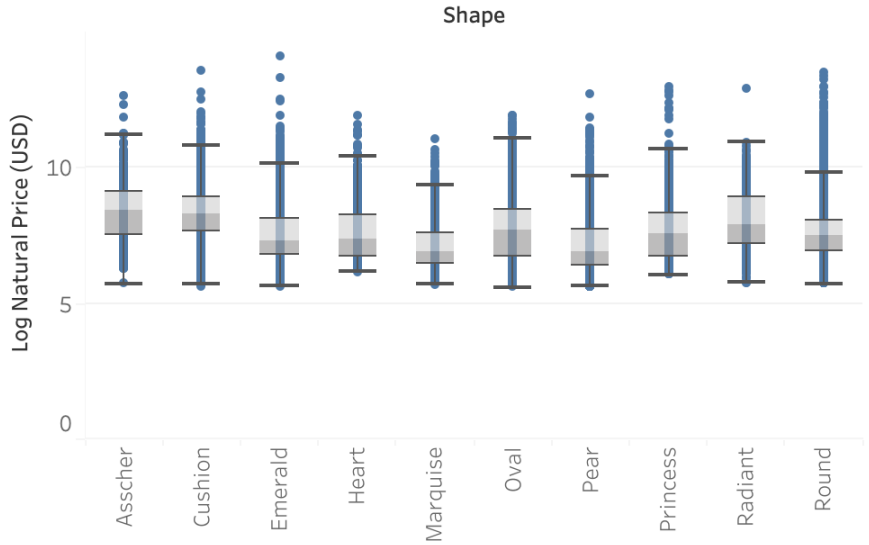
Carat vs Price



Log Nat Price Hist



Boxplots shape vs Log Natural Price



- Natural diamond population from the data set.

Dia Type
natural

DIAMOND DASHBOARD
Use the filters across the top of the dashboard to produce different views

Carat
0.25 to 15.32 and Null values

Color
All

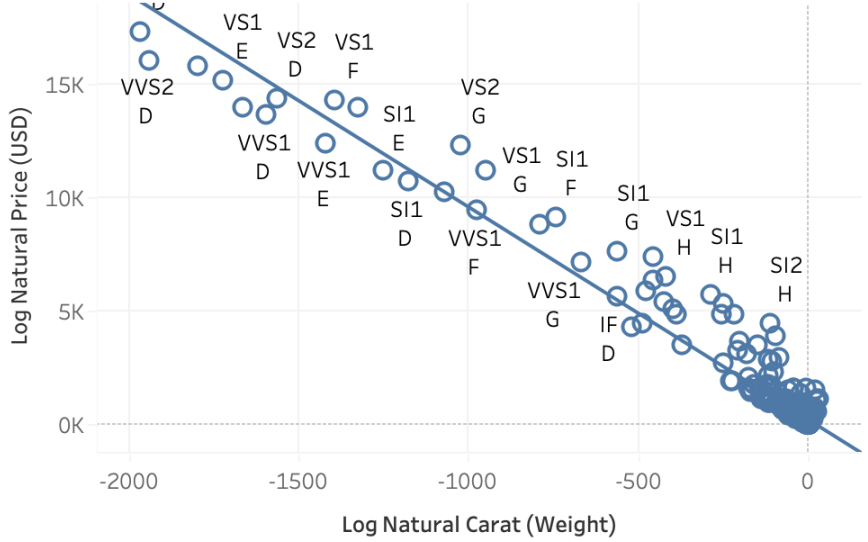
Cut
All

Clarity
All

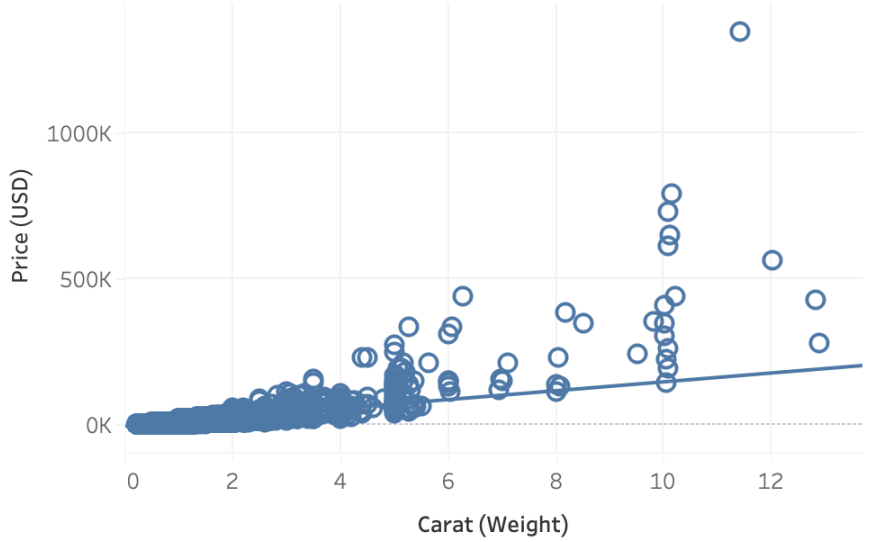
Dia Type
natural

Shape
All

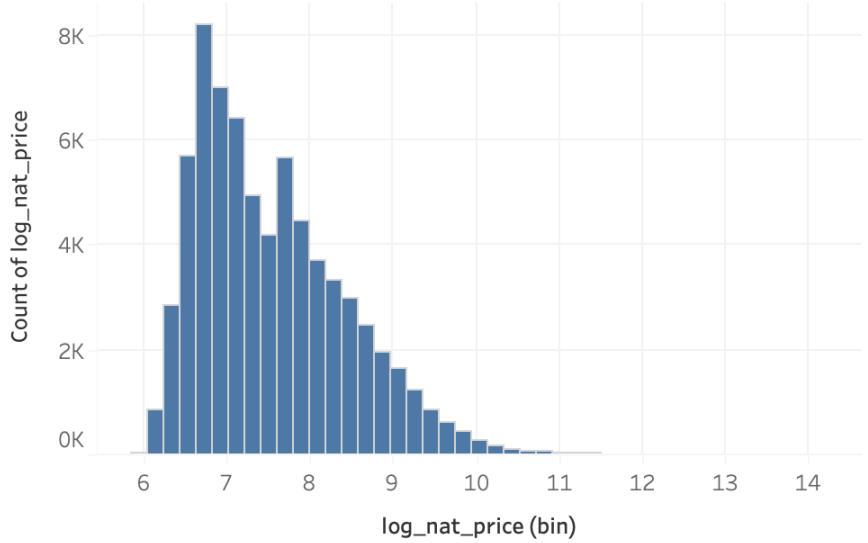
LN Carat vs LN Price



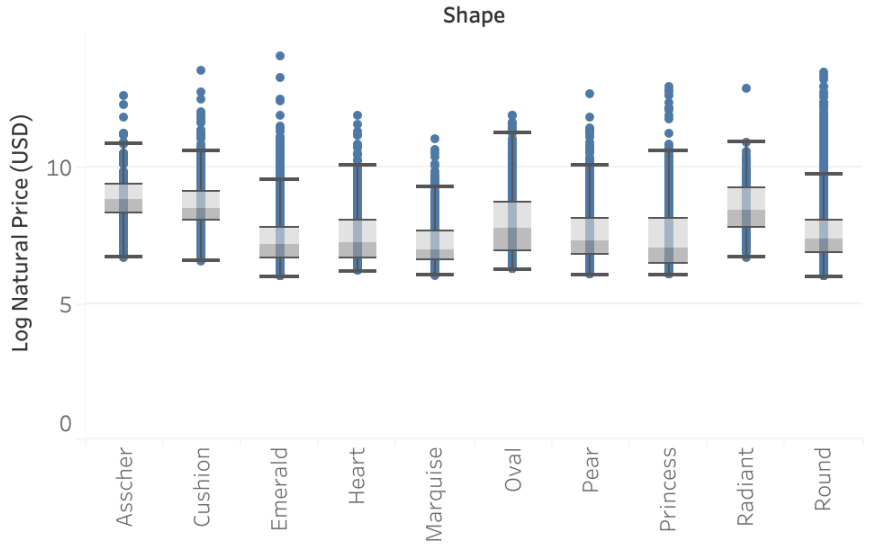
Carat vs Price



Log Nat Price Hist



Boxplots shape vs Log Natural Price



- Synthetic diamond population from the data set

Dia Type
lab

DIAMOND DASHBOARD
Use the filters across the top of the dashboard to produce different views

Carat
0.25 to 15.32 and Null values

Color
All

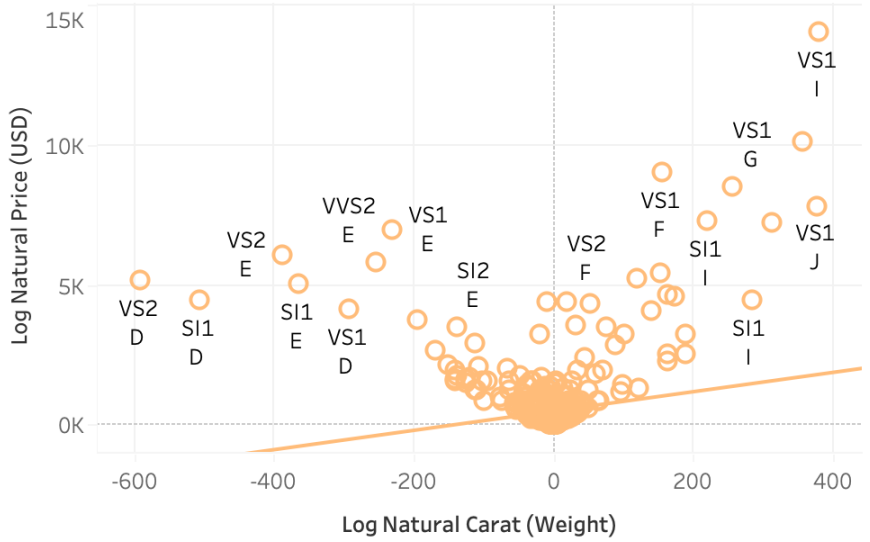
Cut
All

Clarity
All

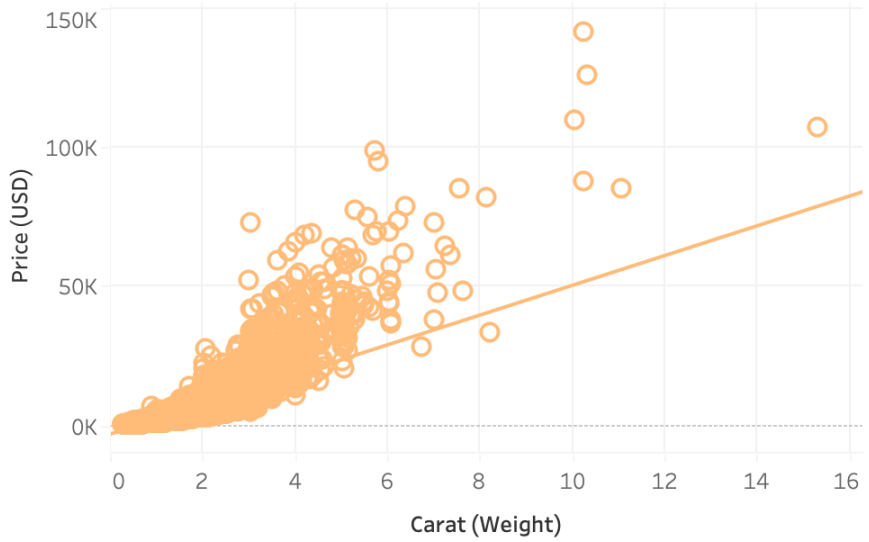
Dia Type
lab

Shape
All

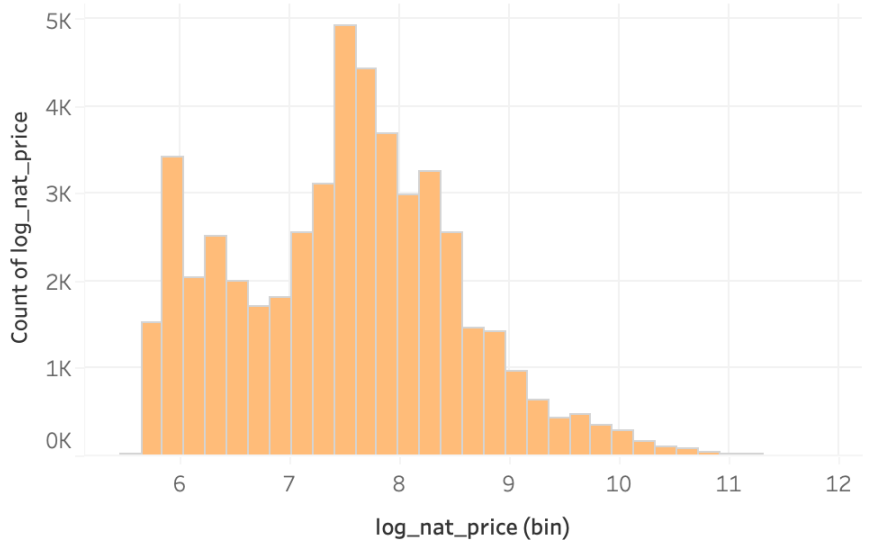
LN Carat vs LN Price



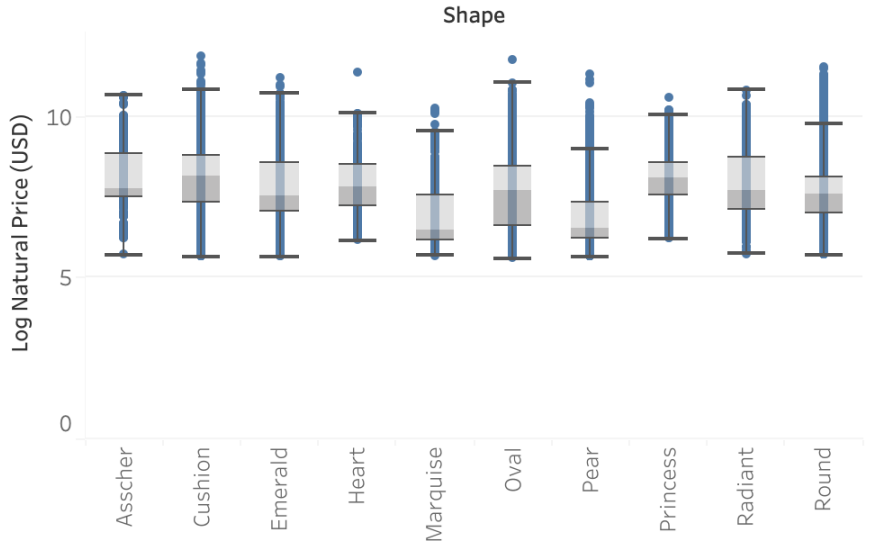
Carat vs Price



Log Nat Price Hist

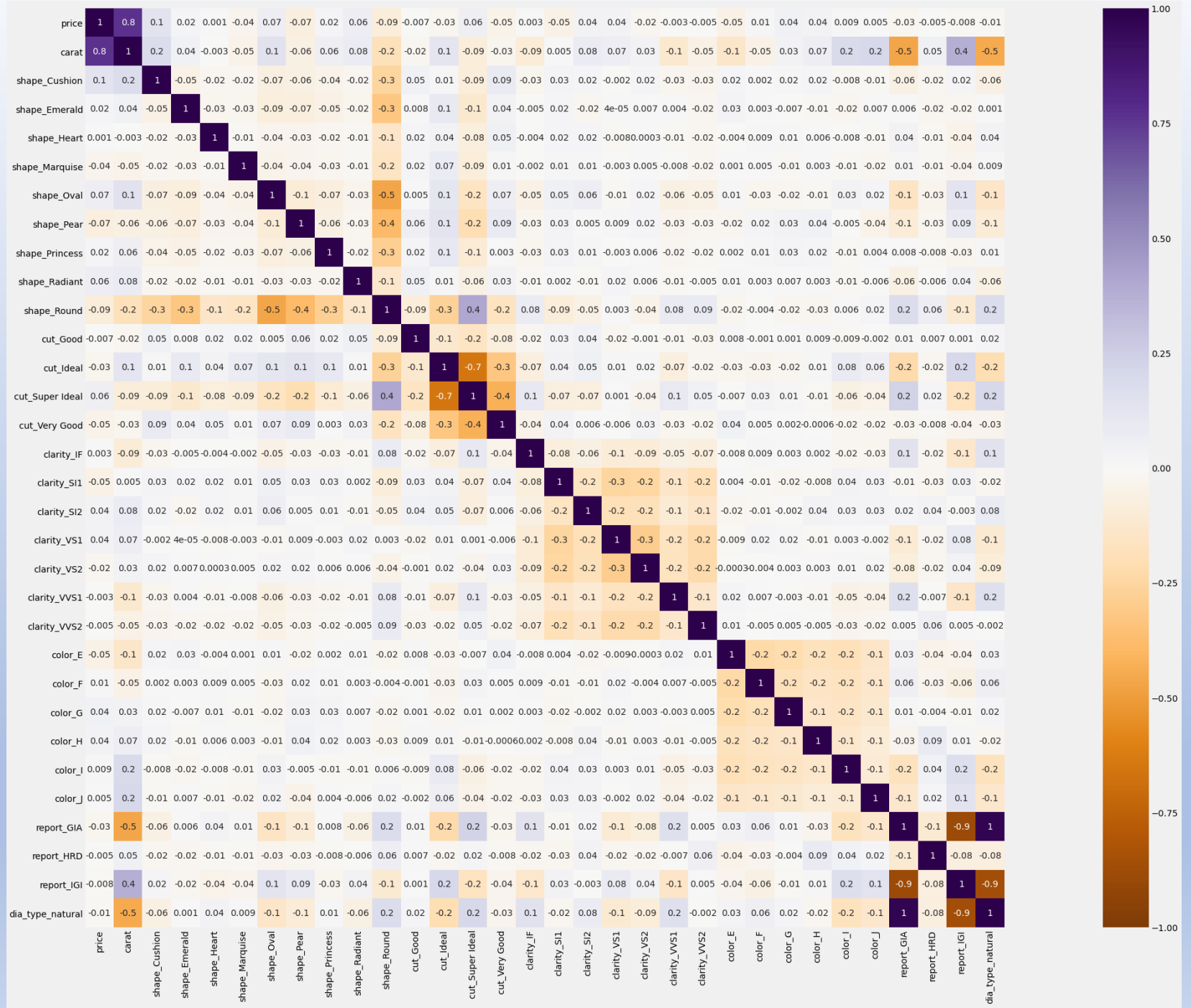


Boxplots shape vs Log Natural Price



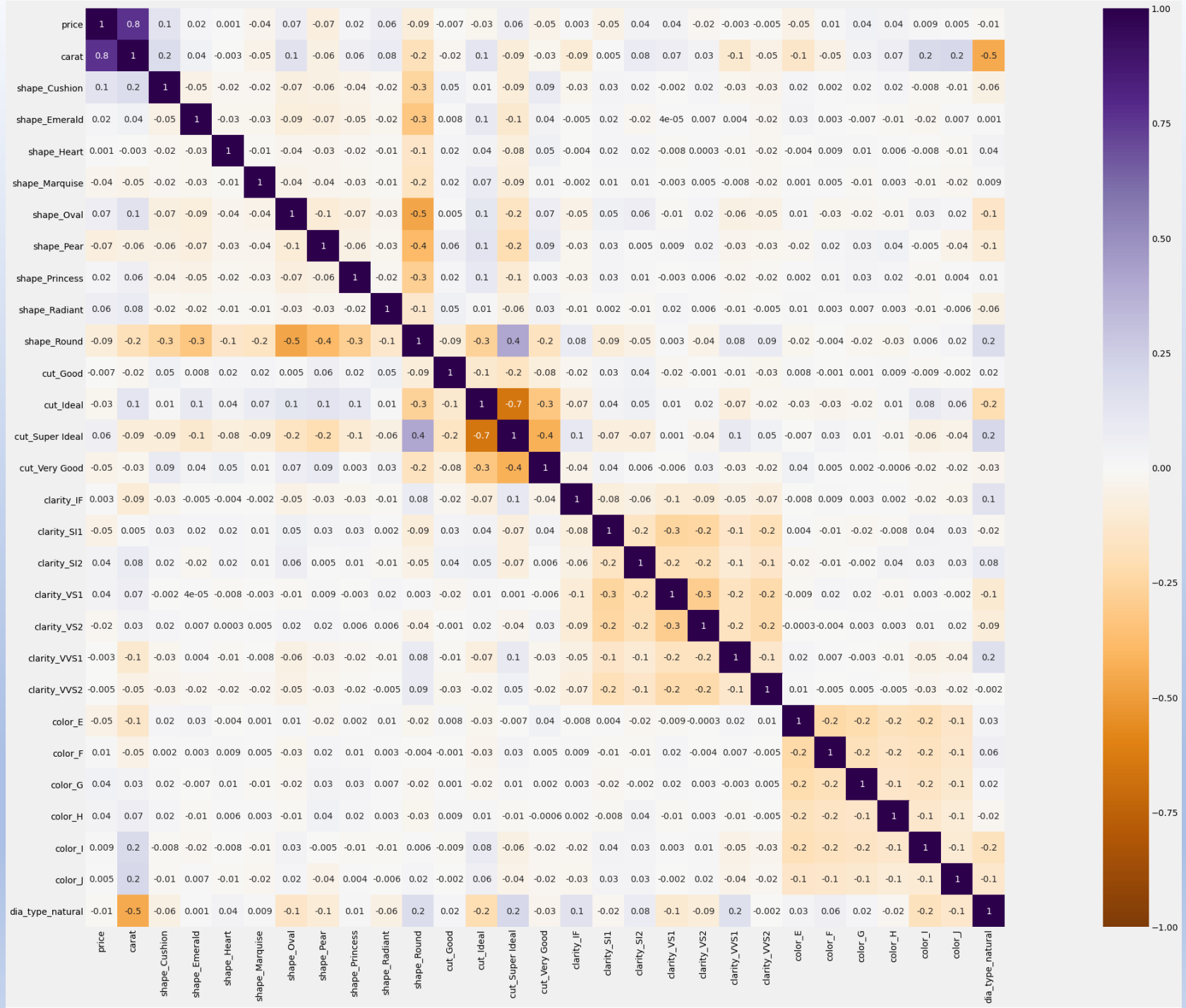
DATA REDUCTION & TRANSFORMATION

- The data was inspected for multicollinearity
- Predictor variables with high pearson-r correlations > 0.7 are removed.
- The report features are identified for removal



DATA REDUCTION & TRANSFORMATION

- The correlation of the reduced data-set



PAIRED T-TESTS

- The data was subset for each t-test
- The first subset parameters can be seen to the right
- A random sample of 30 diamonds from each was selected and an assessment of variance equality and normality of the samples was performed.
- In this case the null hypotheses that there is no statistically significant difference in the means of the two populations is rejected.
- P-value 3.51e-19

Begin T-Test for difference of means transformed price column

Natural vs synthetic populations

H0 There is no statistically significant difference of means between natural and synthetic populations

H1 There is a statistically significant difference in the means natural and synthetic populations

p-val less than 0.05 = statistically significant results reject the null hypothesis

```
#create a subsamle of diamonds lab created between 0.5 and 1ct
lab_over_half_g_ideal_df = df1[(df1['dia_type_natural'] == 0)
            & (df1['shape_Round']==1)
            & (df1['carat'] <= 1)
            & (df1['carat'] > 0.5)
            & (df1['color_G']==1)
            & (df1['cut_Ideal']==1)
            & (df1['clarity_VS1']==1)]
print('lab diamond data shape is: {}'.format(lab_over_half_g_ideal_df.shape))


#create a subsample of diamonds natural between 0.5 and 1ct
nat_over_half_g_ideal_df = df1[(df1['dia_type_natural'] == 1)
            & (df1['carat'] <= 1)
            & (df1['carat'] > 0.5)
            & (df1['shape_Round']==1)
            & (df1['color_G']== 1)
            & (df1['cut_Ideal'] == 1)
            & (df1['clarity_VS1'] == 1)]
print('natural diamond data shape is: {}'.format(nat_over_half_g_ideal_df.shape))
```

```
lab diamond data shape is: (58, 30)
natural diamond data shape is: (30, 30)
```

PAIRED T-TESTS

- The data was subset for each t-test
- The first subset parameters can be seen to the right
- A random sample of 22 diamonds from each was selected and an assessment of variance equality and normality of the samples was performed.
- In this case the null hypotheses that there is no statistically significant difference in the means of the two populations is rejected.
- P-value 3.493e-18
- Wilcoxon signed rank test performed confirms results

```
#subsample lab created diamonds from df1
lab_less_half_g_ideal_df = df1[(df1['dia_type_natural'] == 0)
            & (df1['carat'] <= 0.5)
            & (df1['shape_Round']==1)
            & (df1['color_G']==1)
            & (df1['cut_Ideal'] == 1)
            & (df1['clarity_VS1'] == 1)]
print('lab diamond data shape is: {}'.format(lab_less_half_g_ideal_df.shape))

nat_less_half_g_ideal_df = df1[(df1['dia_type_natural'] == 1)
            & (df1['carat'] <= 0.5)
            & (df1['shape_Round']==1)
            & (df1['color_G']== 1)
            & (df1['cut_Ideal'] == 1)
            & (df1['clarity_VS1'] == 1)]
print('natural diamond data shape is: {}'.format(nat_less_half_g_ideal_df.shape))
```

```
lab diamond data shape is: (22, 30)
natural diamond data shape is: (95, 30)
```

LINEAR REGRESSION ANALYSIS

Data split into training, validation and test sets

Multiple linear regression was performed to determine how each predictor variable aka diamond attribute informs the response variable.

The response variable is the natural log of the price variable.

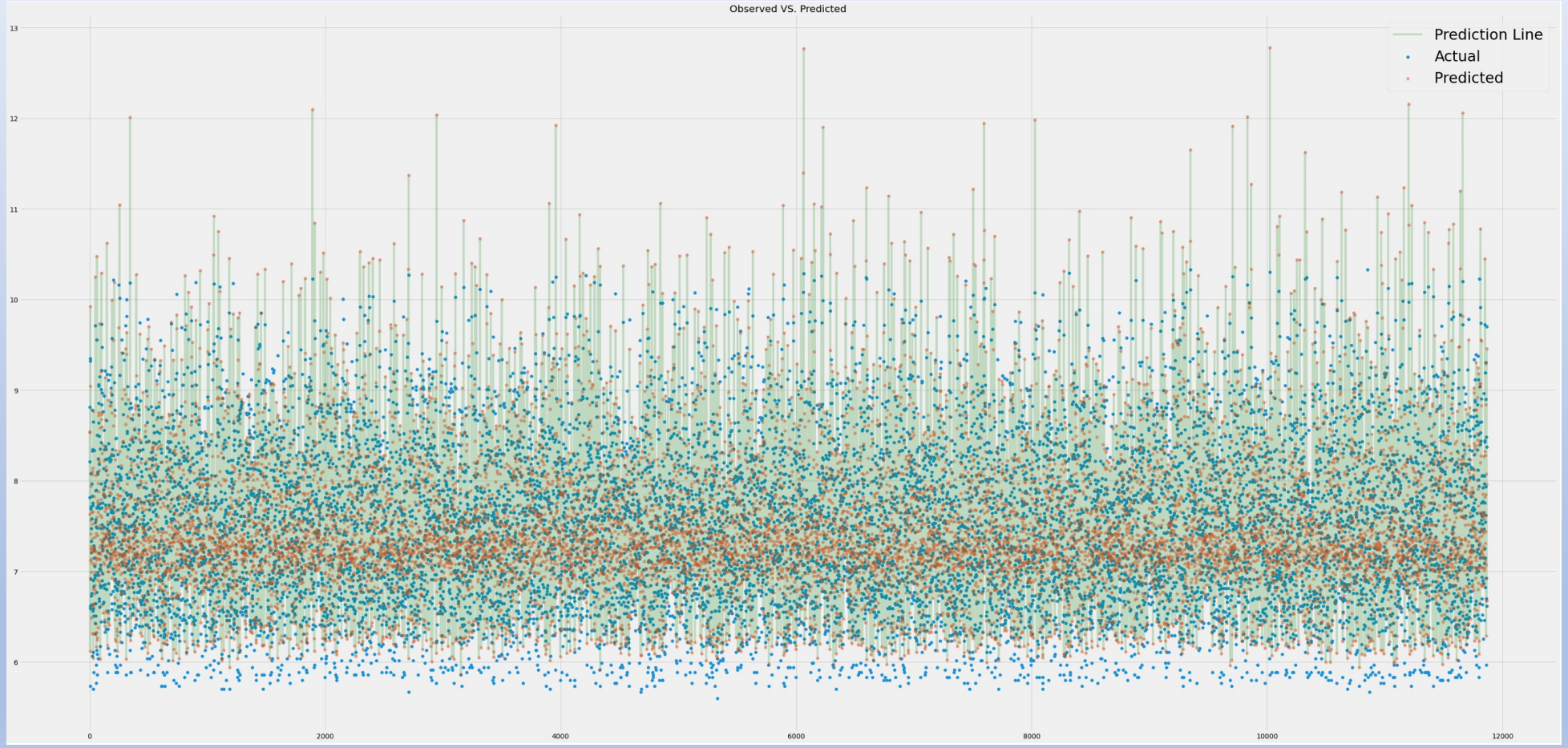
Results were confirmed on the validation set and solidified using the test set

A second regression was performed using the top two informative predictor variables identified in the first analysis and confirmed using validation and test sets.

Assumptions of linear regression were evaluated for each regression performed.

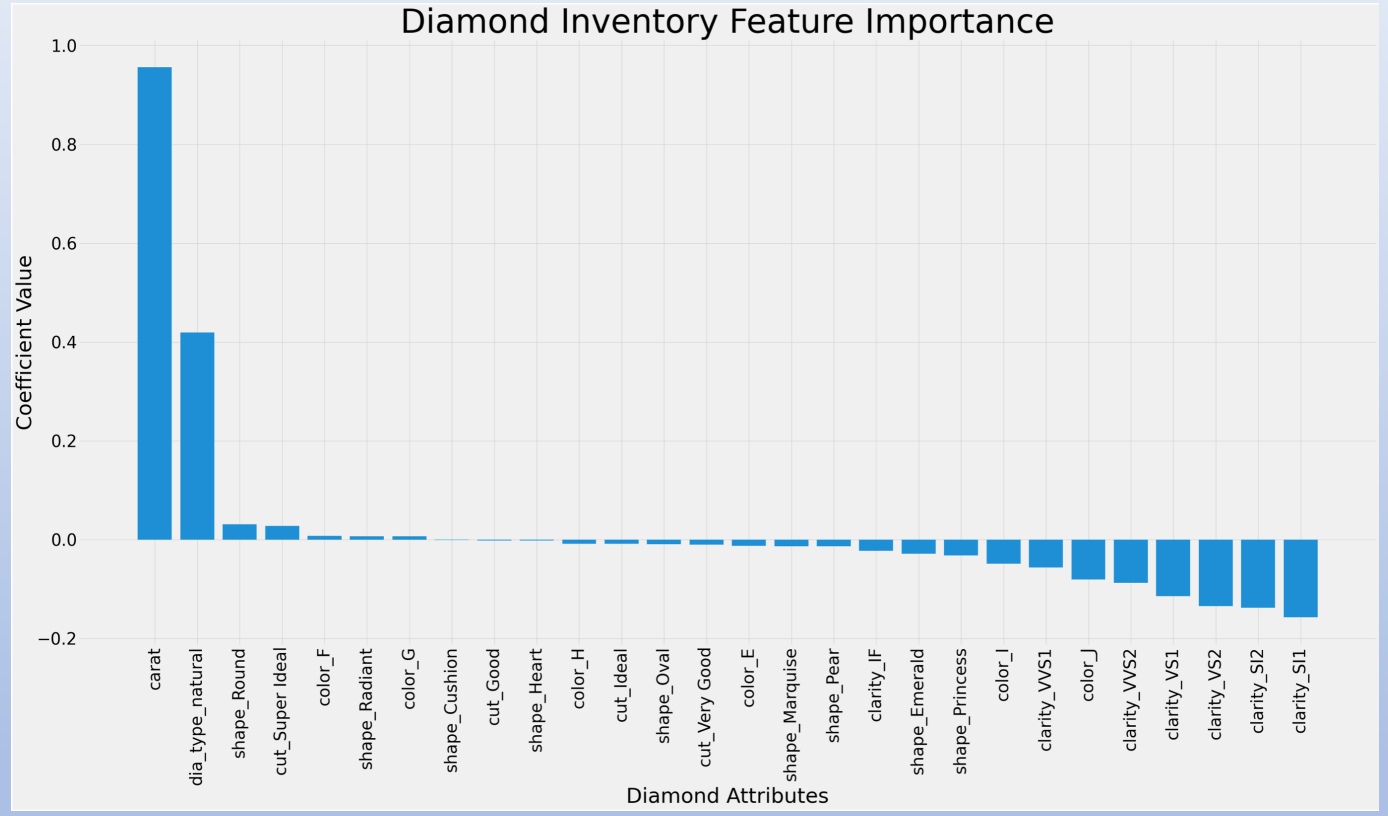
LINEAR REGRESSION 1

- Plot of predictions and observed values for the first regressive analysis best model



LINEAR REGRESSION 1

- The most informative feature in the data set is carat.
- The second most informative feature is the determination of natural origin.
- Best model scores:
- Training = 0.848
- Validation = 0.852
- Test = 0.856
- MSE = 0.124



LINEAR REGRESSION 2

Best model scores:

Training = 0.819

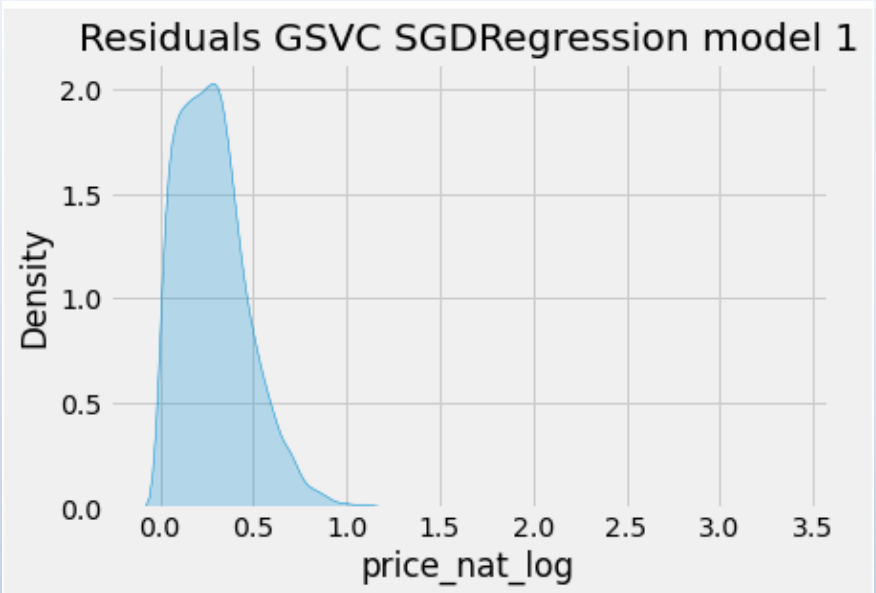
Validation = 0.825

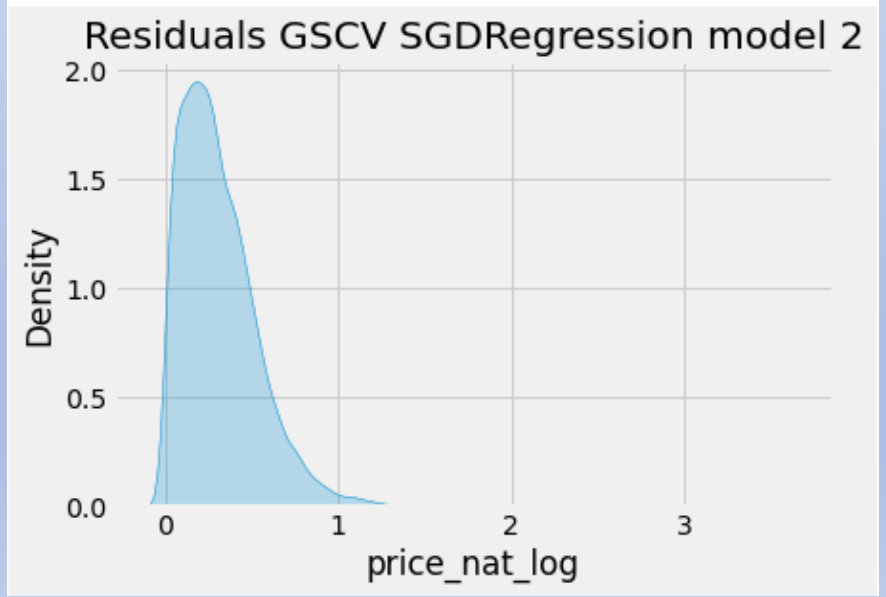
Test = 0.83

MSE = 0.147

EVALUATION OF RESIDUALS AND FITTED VALUES

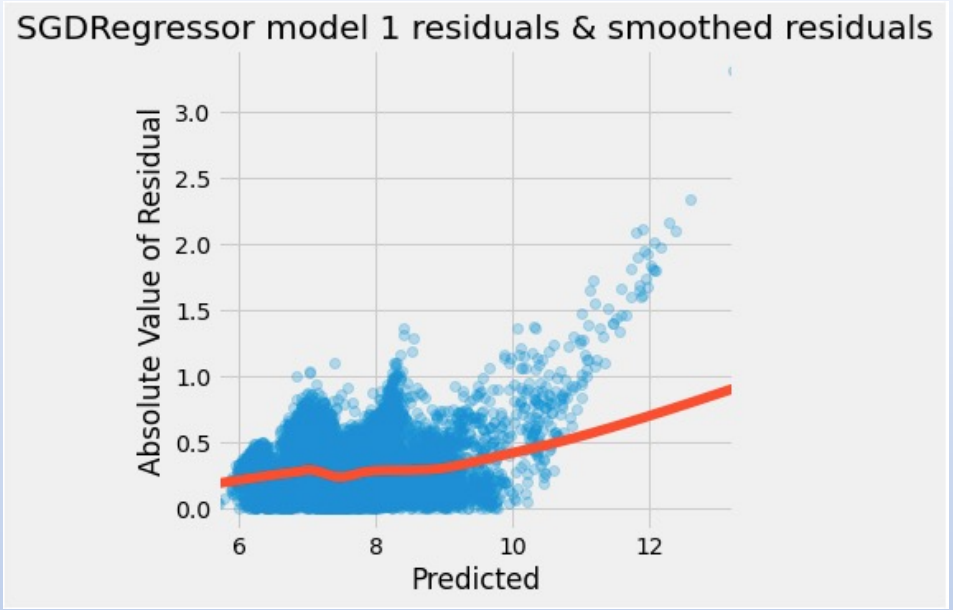
- Residual distributions are approaching normal for each of the regressive models

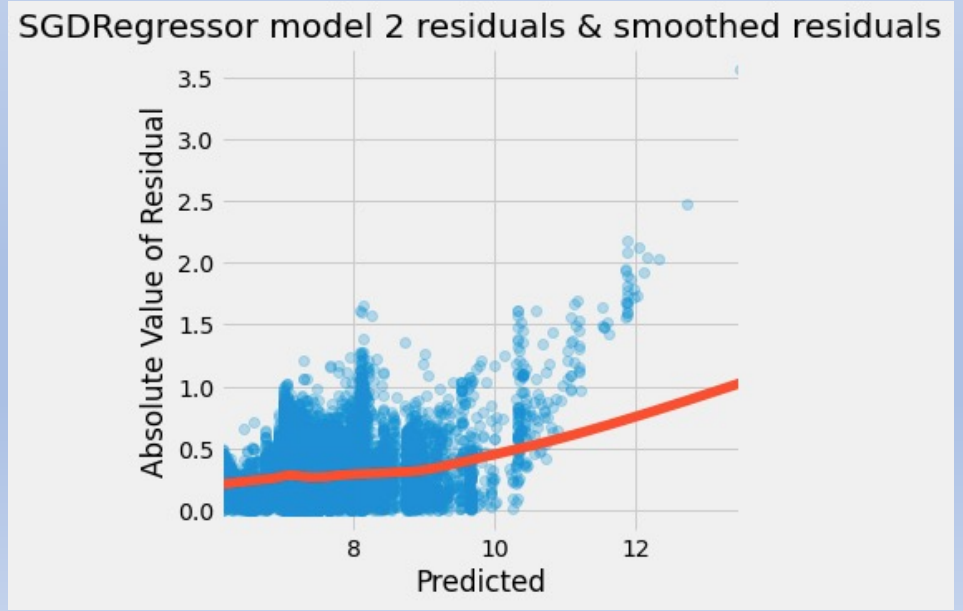




EVALUATION OF RESIDUALS AND FITTED VALUES

- The distribution of fitted values vs residuals shows a non-linear relationship at higher values of diamond price (heteroscedastic error)
- This may be due in part by changing retail markup schedules for diamonds of higher wholesale value





FINAL DETERMINATION OF ANALYSIS

- Reject the null hypothesis that there is no statistically significant difference in the effect of diamond attributes on price determination for both natural and synthetic diamonds
- T-test results
- Linear regression: determination of natural 2nd most informative feature

LIMITATIONS OF ANALYSIS

- Paired t-tests are most informative and reliable when input distributions are all normally distributed.
- Linear regression only capture linear relationships in data if non-linear relationships are present they will not be directly observed
- Heteroscedastic error in the residual vs fitted valures

PROPOSED ACTIONS

Cluster analysis on a larger group of observations split between natural and synthetic populations

Random forest ensemble regression analysis on each cluster defined

Observe similarities and differences in the clustering and predictive accuracy of models on natural and synthetic populations

EXPECTED BENEFITS: THIS ANALYSIS

Determination of the existence of a difference between the value of natural and synthetic diamonds for use in making more efficient purchases for inventory

Assessment of current inventory prices moving forward

Determine more realistic sale prices using the linear model in most price ranges

Use as a competitive analysis tool regarding inventory structure and price offering

EXPECTED BENEFITS: PROPOSED ANALYSIS

Solidify the determination of difference between assigning value to synthetic and natural diamonds

Determine the differences between value related clusters in the data for future use and application on inventories and prospective retail and wholesale offerings in and out.

Use to make more efficient purchases of diamonds by identifying increased gap between price and value increasing ROI

Increased inventory turnover through competitive pricing based on analysis results

Better investment choices for diamond inventory.

REFERENCES

Corral, M. (2020, December 14). Brilliant diamonds. Kaggle. Retrieved March 3, 2022, from https://www.kaggle.com/miguelcorraljr/brilliant-diamonds

GIA.edu. (2019, January 24). *4Cs of Diamond Quality: What's the most important C?* GIA 4Cs. Retrieved March 21, 2022, from https://4cs.gia.edu/en-us/blog/4cs-diamond-quality-most-important-c/

McDaniel, B. (2022, March 18). *Diamond Attribute Association Analysis*. Western Governors University. Retrieved March 21, 2022, from https://tasks.wgu.edu/student/000194874/course/20900008/task/2806/submission

THANK YOU FOR YOUR TIME