Introduction to Programming and Predictive Analytics for Business IFI 8410

Final Project Presentation

Georgia State University | J. MACK ROBINSON COLLEGE OF BUSINESS

Robinson

Outline

- Team introductions
- Business or research question
- Data
- Methods and results
- Limitations and extensions

Team Introductions

- *Tyra Bryant- GSU MBA, 2ND Semester*
- *Chris Drummond- GSU MBA, 4th Semester*

Research Question

- *Which characteristic has the greatest effect on diamond pricing?*
  - Those characteristics include; cut, clarity, color, and carat
  - *Dependent Variable: Y=Price*
  - *We expected clarity to have the biggest effect on diamond price.*

- *How does the analytics from a larger dataset vary from that of the findings of a smaller dataset?*

- *When the larger data set is combined with the smaller dataset, does it confirm or differ from the findings of the smaller dataset?*

Data

```
In [5]: rings.describe()
```

Out[5]:

	carat	color	clarity	cut	channel	store	price
count	425.000000	425.000000	425.000000	425.000000	425.000000	425.000000	425.000000
mean	1.040685	4.312941	6.134118	0.362353	1.609412	9.240000	6355.992941
std	0.421967	1.864122	1.604354	0.481247	0.718952	2.597858	4404.237376
min	0.200000	1.000000	2.000000	0.000000	0.000000	1.000000	497.000000
25%	0.720000	3.000000	5.000000	0.000000	1.000000	10.000000	3430.000000
50%	1.020000	4.000000	6.000000	0.000000	2.000000	10.000000	5476.000000
75%	1.210000	6.000000	7.000000	1.000000	2.000000	11.000000	7792.000000
max	2.480000	9.000000	10.000000	1.000000	2.000000	12.000000	27575.000000

Small Data Summary Statistics

(Rings)

Y Value: Price

Minimum: $497.00

Average: $6355.00

Maximum: $27,575.00

425 rows × 7 columns

Data

In [10]:
```
Drings.describe()
```

Out[10]:

	carat	cut	color	clarity	depth	table	price	x	y	z
count	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000
mean	0.797940	2.553003	2.594197	3.835150	61.749405	57.457184	3932.799722	5.731157	5.734526	3.538734
std	0.474011	1.027708	1.701105	1.724591	1.432621	2.234491	3989.439738	1.121761	1.142135	0.705699
min	0.200000	0.000000	0.000000	0.000000	43.000000	43.000000	326.000000	0.000000	0.000000	0.000000
25%	0.400000	2.000000	1.000000	2.000000	61.000000	56.000000	950.000000	4.710000	4.720000	2.910000
50%	0.700000	2.000000	3.000000	4.000000	61.800000	57.000000	2401.000000	5.700000	5.710000	3.530000
75%	1.040000	3.000000	4.000000	5.000000	62.500000	59.000000	5324.250000	6.540000	6.540000	4.040000
max	5.010000	4.000000	6.000000	7.000000	79.000000	95.000000	18823.000000	10.740000	58.900000	31.800000

- *Big Data Summary Statistics*

  (Rings)

Y Value: Price

Minimum: $326.00

Average: $3932.79

Maximum: $18,823.00

`[53940 rows x 10 columns]>`

Data

Out[392]:

	carat	cut	color	clarity	price
count	54365.000000	54365.000000	54365.000000	54365.000000	54365.000000
mean	0.799837	2.535363	4.478635	3.924198	3953.371783
std	0.474105	1.042801	1.945218	1.679012	4004.661030
min	0.200000	0.000000	1.000000	1.000000	326.000000
25%	0.400000	2.000000	3.000000	2.000000	956.000000
50%	0.700000	2.000000	5.000000	4.000000	2427.000000
75%	1.040000	3.000000	6.000000	5.000000	5364.000000
max	5.010000	4.000000	9.000000	10.000000	27655.000000

Combined Data Summary Statistics (Brings)

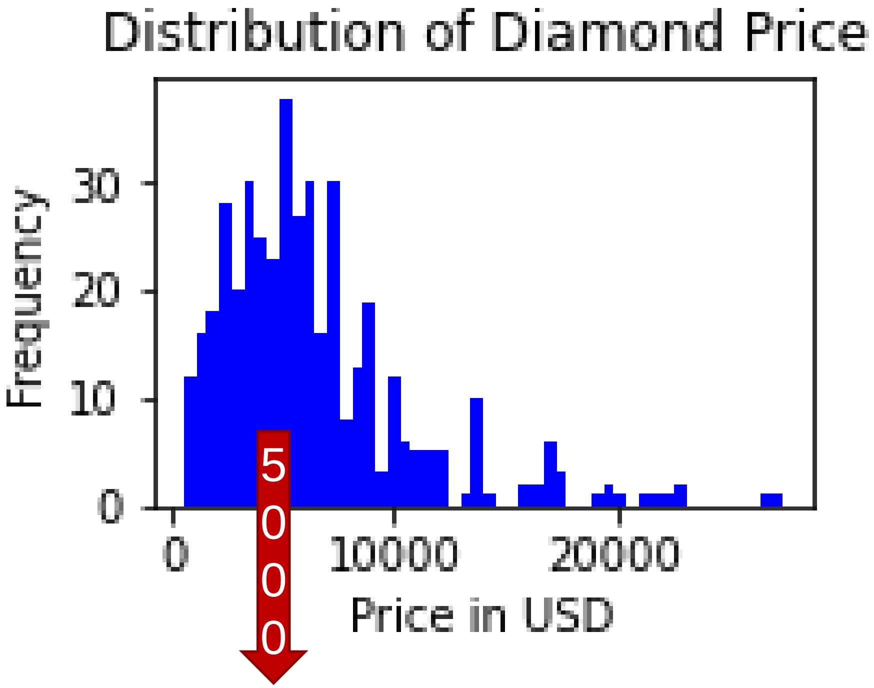
Y Value: Price

Minimum: $326.00

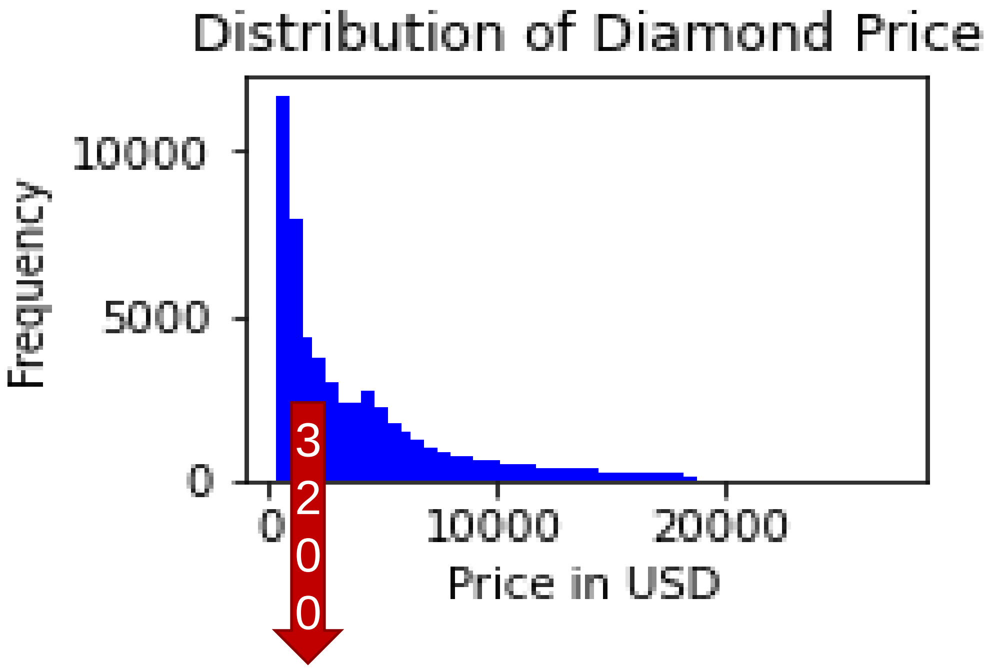
Average: $3,953.37

Maximum: $27,655.00

Carat Price Distribution



Small Dataset



Combined Dataset

Data

In [11]:

```
avg_labels = ['average and below', 'above average']
avg_bins = [-100000, 6355.992941, 100000]
rings['price_average'] = pd.cut(rings['price'], bins=avg_bins, labels=avg_labels)
rings
```

Out[11]:

	carat	color	clarity	cut	channel	store	price	price_average
0	0.826	4	7	1	1	1	7775	above average
1	0.996	5	6	1	1	1	9850	above average
2	1.070	4	7	1	1	1	10950	above average
3	1.070	7	7	0	1	1	7500	above average
4	1.010	8	6	0	1	1	6995	above average
...	...	...	...	...	...	...	...	...
420	0.780	6	8	0	0	12	6699	above average
421	0.375	6	8	0	0	12	1542	average and below
422	0.580	7	6	0	0	12	3389	average and below
423	0.395	4	8	0	0	12	1850	average and below
424	0.390	5	8	0	0	12	1761	average and below

425 rows × 8 columns

Clarity	1	2	3	4	5	6	7	8	9	10	11
	FL	IF	VVS1	VVS2	VS1	VS2	SI1	SI2	I1	I2	I3

Color	1	2	3	4	5	6	7	8	9
	D	E	F	H	I	J	K	L	M

Cut	1	0
	Ideal	Not Ideal

Channel	1	2	0
	Indep.	Internet	Mall

Store	1	2	3	4	5	6	7	8	9	10	11	12
	Good's	Chalm's	Fred M.	R. Hollan.	Ausman's	Univers.	Kay	Zales	Danford	Blue Nile	Ashford	Riddle's

Data

```
Brings['price_average'] = pd.cut(Brings['price'], bins=av
Brings
```

Out[4]:

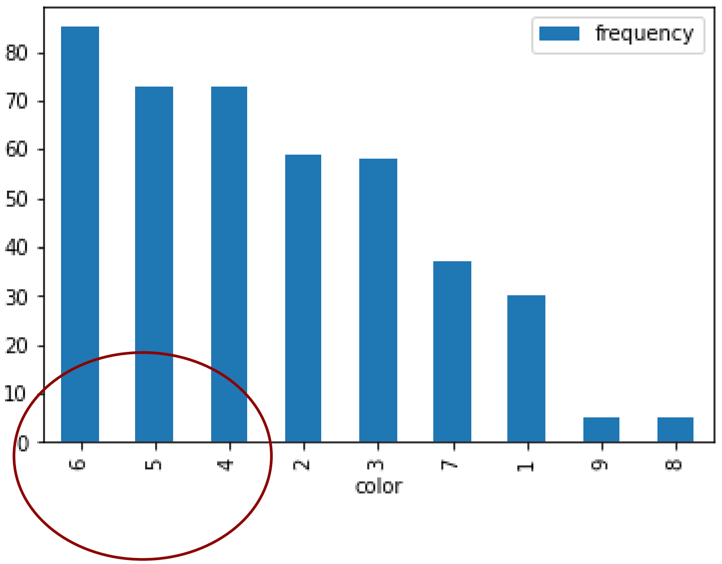
	carat	cut	color	clarity	price	price_average
0	0.36	1	2	4	810	average and below
1	0.40	1	6	5	813	average and below
2	0.31	1	6	5	814	average and below
3	0.34	1	2	5	816	average and below
4	0.41	1	6	3	818	average and below
...	...	...	...	...	...	...
54360	2.00	4	2	5	18759	above average
54361	1.51	4	2	6	18777	above average
54362	2.03	4	6	2	18781	above average
54363	2.00	4	6	3	18803	above average
54364	2.00	4	9	2	18818	above average

54365 rows × 6 columns

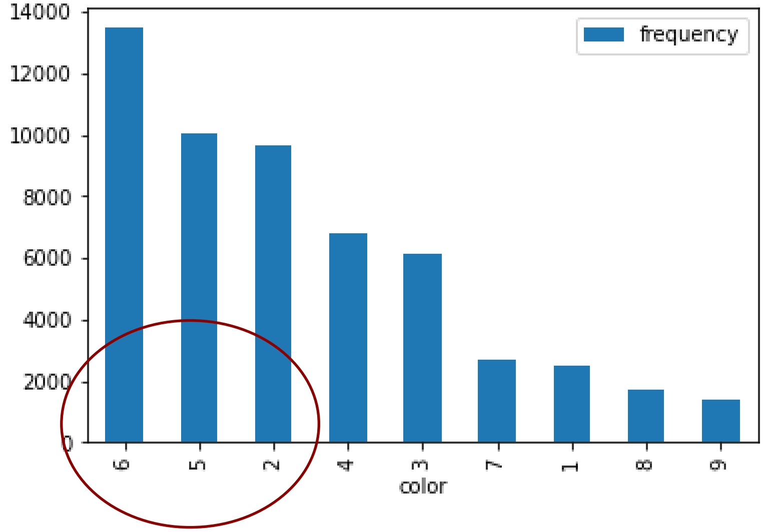
Fair	Good	Ideal	Premium
1	2	3	4

Combined Dataset

Color Frequency

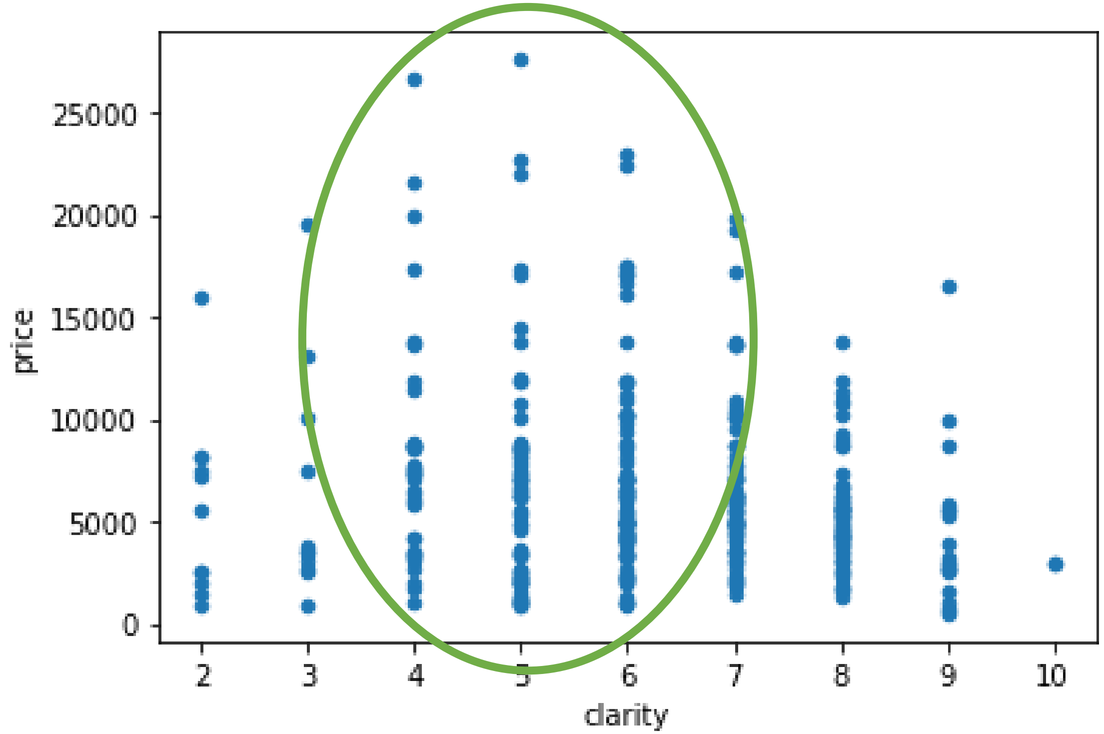


Small Dataset

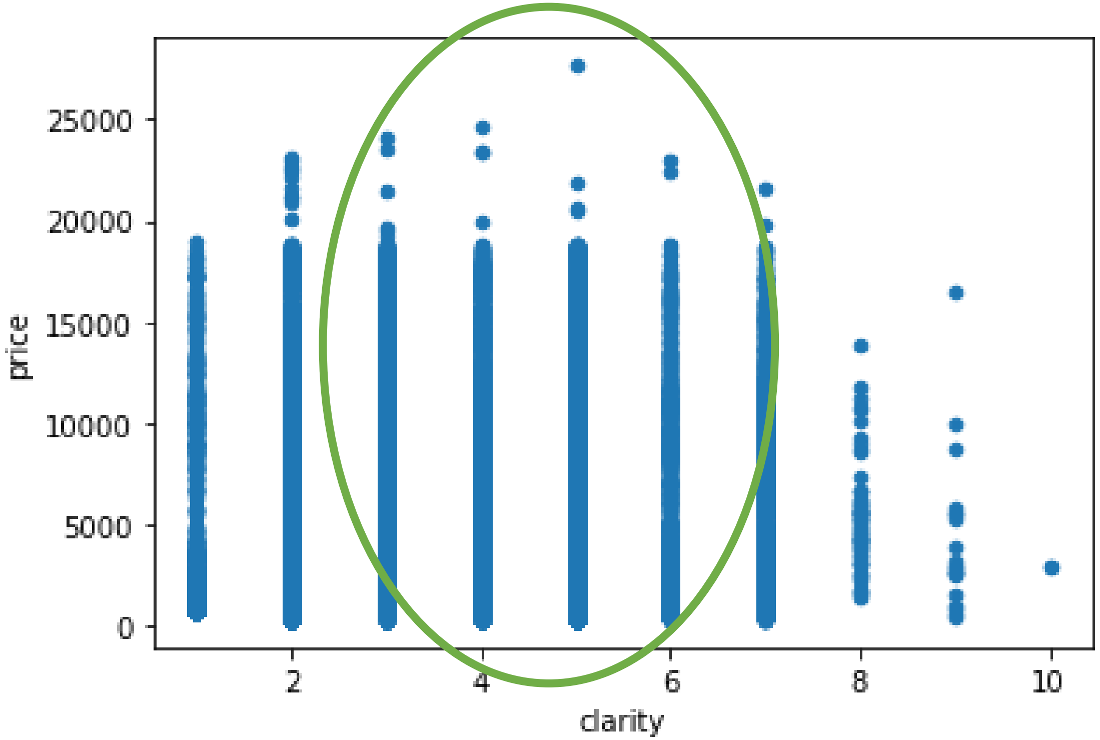


Combined Dataset

Clarity Frequency

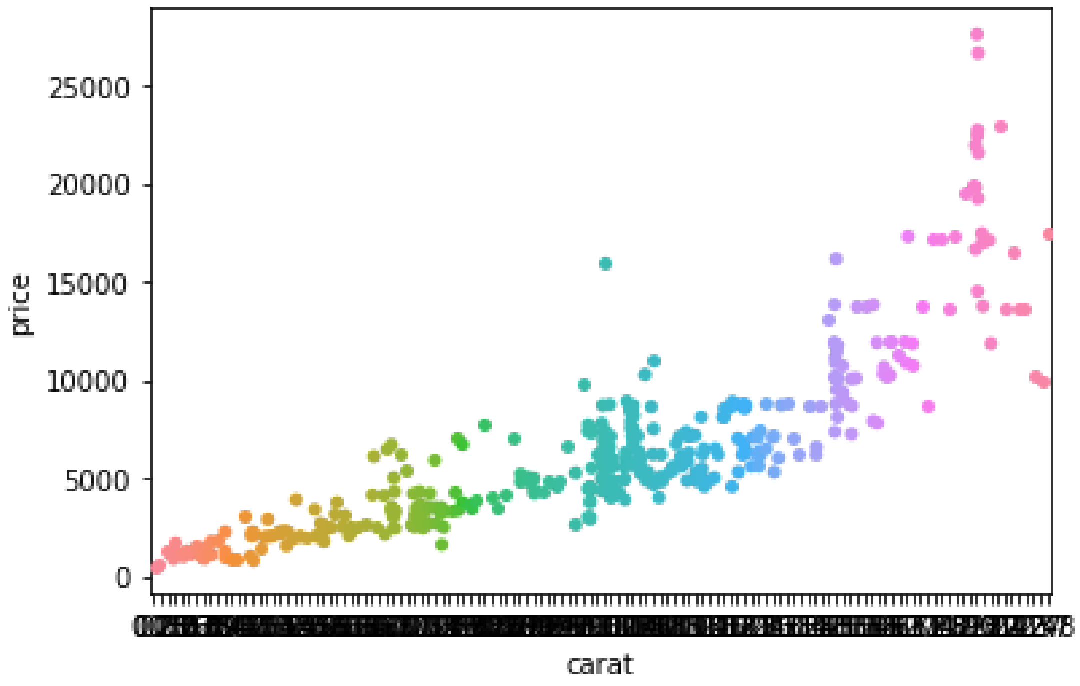


Small Dataset

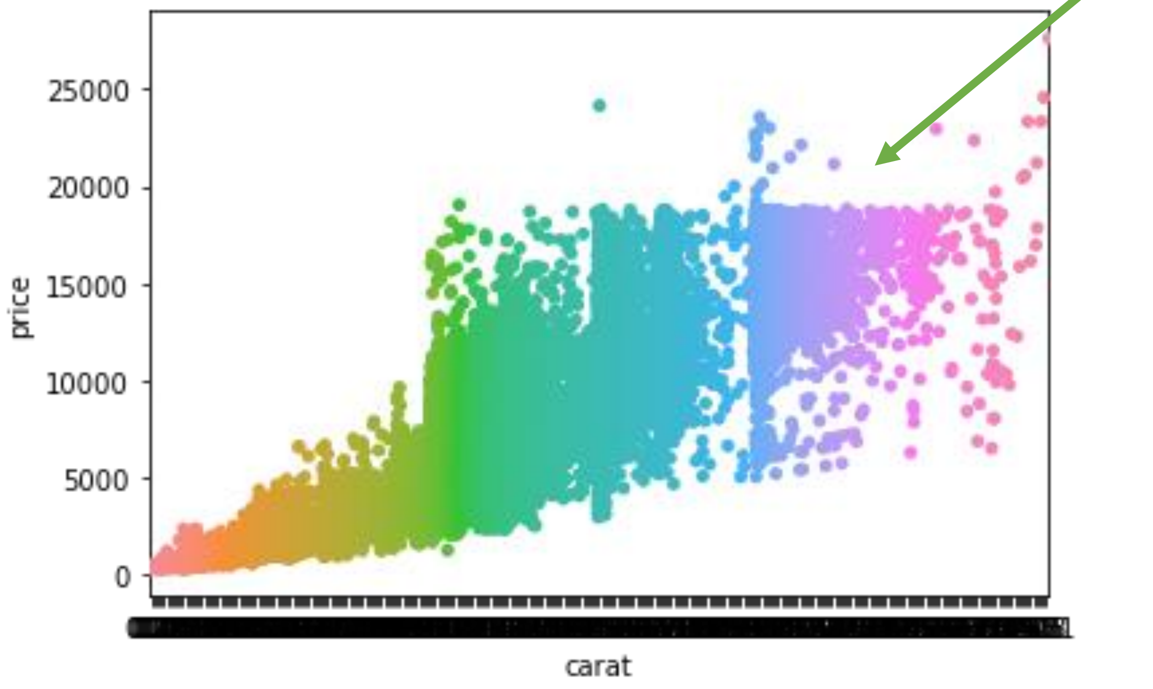


Combined Dataset

Carat Price Distribution

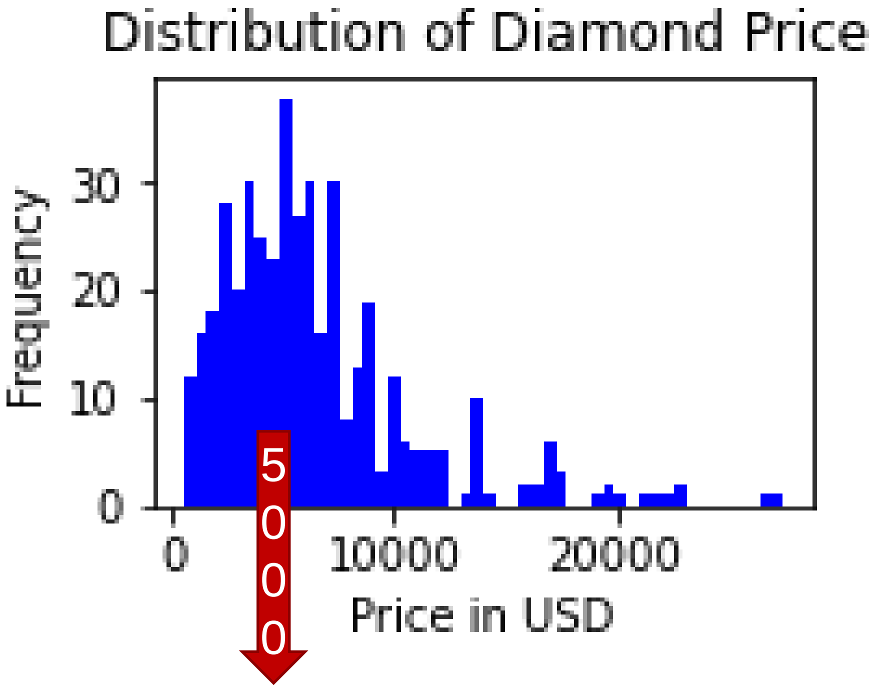


Small Dataset

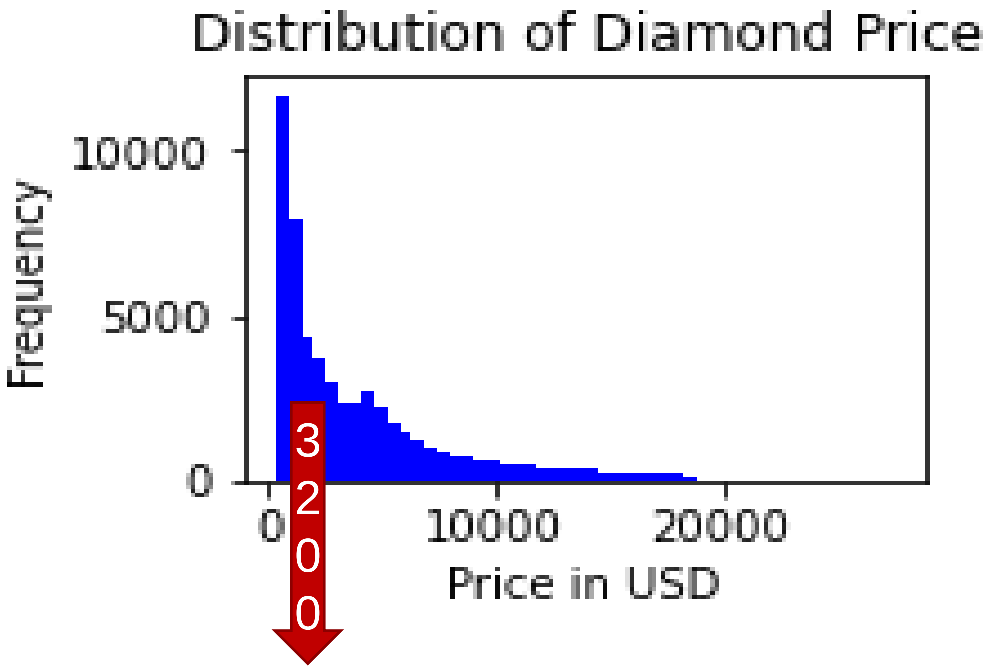


Combined Dataset

Carat Price Distribution



Small Dataset



Combined Dataset

Linear Regression Clarity/Color -Price

Small Dataset

Out[5]:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.026
Model:	OLS	Adj. R-squared:	0.021
Method:	Least Squares	F-statistic:	5.558
Date:	Wed, 21 Apr 2021	Prob (F-statistic):	0.00414
Time:	14:13:39	Log-Likelihood:	-4162.9
No. Observations:	425	AIC:	8332.
Df Residuals:	422	BIC:	8344.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9486.9745	963.300	9.848	0.000	7593.510	1.14e+04
clarity	-384.2163	131.921	-2.912	0.004	-643.520	-124.913
color	-179.4954	113.537	-1.581	0.115	-402.665	43.674

Omnibus:	136.010	Durbin-Watson:	0.327
Prob(Omnibus):	0.000	Jarque-Bera (JB):	363.335
Skew:	1.558	Prob(JB):	1.27e-79
Kurtosis:	6.288	Cond. No.	35.6

Combined Dataset

Out[25]:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.036
Model:	OLS	Adj. R-squared:	0.036
Method:	Least Squares	F-statistic:	1019.
Date:	Wed, 21 Apr 2021	Prob (F-statistic):	0.00
Time:	14:06:51	Log-Likelihood:	-5.2711e+05
No. Observations:	54365	AIC:	1.054e+06
Df Residuals:	54362	BIC:	1.054e+06
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3016.6870	57.493	52.471	0.000	2904.001	3129.373
clarity	-175.3777	10.044	-17.461	0.000	-195.064	-155.691
color	362.8118	8.670	41.849	0.000	345.819	379.804

Omnibus:	14343.573	Durbin-Watson:	0.053
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31496.415
Skew:	1.537	Prob(JB):	0.00
Kurtosis:	5.109	Cond. No.	21.8

Linear Regression Carat/Cut -Price

Small Dataset

Out[6]:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.780
Model:	OLS	Adj. R-squared:	0.779
Method:	Least Squares	F-statistic:	747.2
Date:	Wed, 21 Apr 2021	Prob (F-statistic):	2.17e-139
Time:	14:18:26	Log-Likelihood:	-3846.9
No. Observations:	425	AIC:	7700.
Df Residuals:	422	BIC:	7712.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3635.2801	296.862	-12.246	0.000	-4218.792	-3051.768
carat	9347.7764	243.404	38.404	0.000	8869.342	9826.211
cut	726.3226	213.421	3.403	0.001	306.821	1145.824

Omnibus:	164.976	Durbin-Watson:	1.091
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1166.854
Skew:	1.486	Prob(JB):	4.18e-254
Kurtosis:	10.554	Cond. No.	5.82

Combined Dataset

Out[26]:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.850
Model:	OLS	Adj. R-squared:	0.850
Method:	Least Squares	F-statistic:	1.543e+05
Date:	Wed, 21 Apr 2021	Prob (F-statistic):	0.00
Time:	14:17:09	Log-Likelihood:	-4.7650e+05
No. Observations:	54365	AIC:	9.530e+05
Df Residuals:	54362	BIC:	9.530e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2199.4485	22.181	-99.159	0.000	-2242.924	-2155.974
carat	7791.7999	14.044	554.810	0.000	7764.273	7819.326
cut	-30.1770	7.072	-4.267	0.000	-44.038	-16.316

Omnibus:	14684.838	Durbin-Watson:	0.926
Prob(Omnibus):	0.000	Jarque-Bera (JB):	132973.848
Skew:	1.042	Prob(JB):	0.00
Kurtosis:	10.373	Cond. No.	11.0

Methods and Results

- *Of the methods we've learned, which have you used, and why?*
  - *The most important method used was the linear regression*
- *Why are those methods appropriate to your business or research questions?*
  - *Price is a continuous value*
- *What are your results and interpretation of your results?*
  - *Most of the variables were significant to the price of diamonds*

Limitations and Extensions

- *The original dataset lacked enough information*

- *Given that I was able to combine the smaller dataset to a much larger dataset, I think this is made for a richer dataset and makes the prediction more valuable/valid.*