PREDICTION DIAMOND PRICE

A diamond is a chunk of coal that is made good under pressure.

รายชื่อสมาชิก กลุ่ม KBub

ชินาธิป มีสวนนิล
รหัสนักศึกษา 63010235

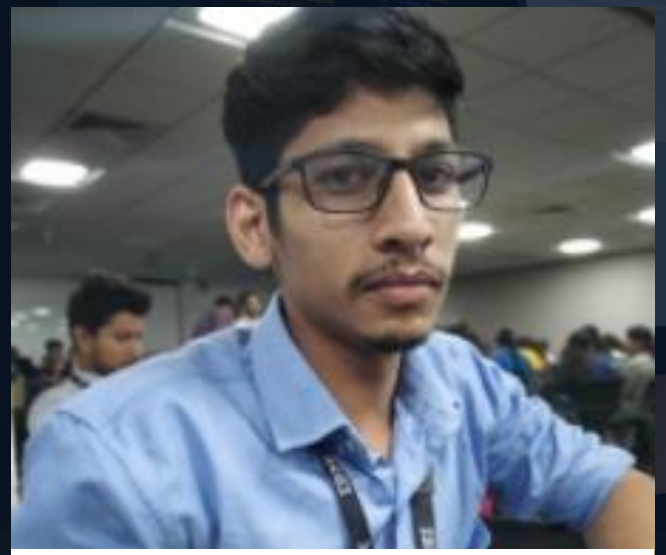
ณัฐพงษ์ นาคสามัคคี
รหัสนักศึกษา 63010326

รายชื่อสมาชิก กลุ่ม พรี่ๆ

นายจิรภัทร แก้วส่งแสง
รหัสนักศึกษา 63010139

นายดิษฐพงษ์ จรัสชัยโรจน์
รหัสนักศึกษา 63010354

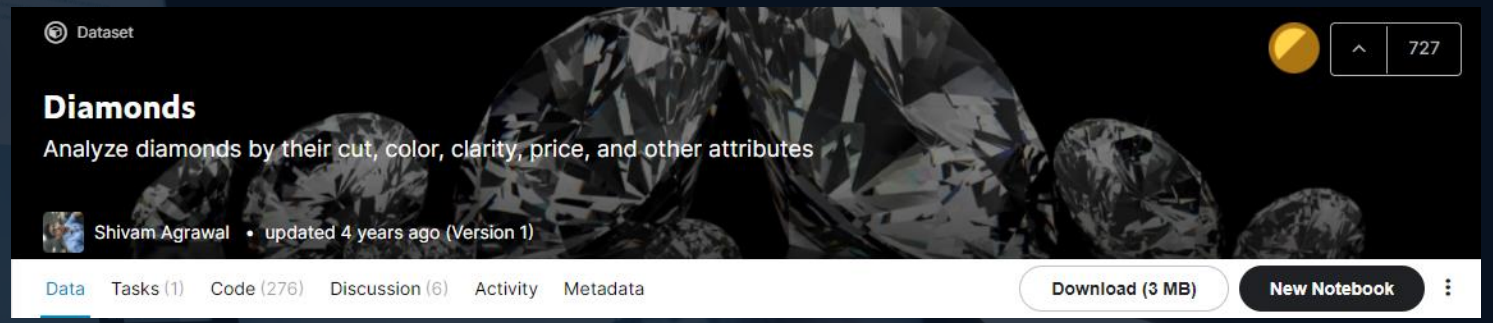
Resources.



Shivam Agrawal

Application Developer at IBM

https://www.kaggle.com/shivam2503/diamonds



	A	B	C	D	E	F	G	H	I	J	K
1		carat	cut	color	clarity	depth	table	price	x	y	z
2	1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
3	2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
4	3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
5	4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
6	5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
7	6	0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48

หลักการทางคณิตศาสตร์

เวกเตอร์

Pearson's Similarity

เมตทริกซ์

Linear Regression

1. นำข้อมูลเข้าสู่โปรแกรม

```python
import pandas as pd
data = pd.read_csv("Diamonds.csv")
data
```

	carat	cut	color	clarity	depth	table	price	x	y	z
0	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...	...
53935	0.72	Ideal	D	SI1	60.8	57.0	2757	5.75	5.76	3.50
53936	0.72	Good	D	SI1	63.1	55.0	2757	5.69	5.75	3.61
53937	0.70	Very Good	D	SI1	62.8	60.0	2757	5.66	5.68	3.56
53938	0.86	Premium	H	SI2	61.0	58.0	2757	6.15	6.12	3.74
53939	0.75	Ideal	D	SI2	62.2	55.0	2757	5.83	5.87	3.64

2. ตรวจสอบข้อมูล

```
1  data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 53940 entries, 0 to 53939
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   carat    53940 non-null  float64
 1   cut      53940 non-null  object
 2   color    53940 non-null  object
 3   clarity  53940 non-null  object
 4   depth    53940 non-null  float64
 5   table    53940 non-null  float64
 6   price    53940 non-null  int64
 7   x        53940 non-null  float64
 8   y        53940 non-null  float64
 9   z        53940 non-null  float64
dtypes: float64(6), int64(1), object(3)
memory usage: 4.1+ MB
```

2. ตรวจสอบข้อมูล

```
1  data
```

	carat	cut	color	clarity	depth	table	price	x	y	z
0	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...	...
53935	0.72	Ideal	D	SI1	60.8	57.0	2757	5.75	5.76	3.50
53936	0.72	Good	D	SI1	63.1	55.0	2757	5.69	5.75	3.61
53937	0.70	Very Good	D	SI1	62.8	60.0	2757	5.66	5.68	3.56
53938	0.86	Premium	H	SI2	61.0	58.0	2757	6.15	6.12	3.74
53939	0.75	Ideal	D	SI2	62.2	55.0	2757	5.83	5.87	3.64

3. Clean Data

เปลี่ยนจาก เชิงคุณภาพ -> เชิงปริมาณ

ข้อมูล Color ก่อนแก้ไข

	carat	cut	color	clarity	depth	table	price	x	y	z
0	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
5	0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
6	0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
7	0.26	Very Good	H	SI1	61.9	55.0	337	4.07	4.11	2.53
8	0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78	2.49
9	0.23	Very Good	H	VS1	59.4	61.0	338	4.00	4.05	2.39

```
new_colors = {"D": 0, "E": 1,"F": 2,"G": 3,"H": 4,"I": 5,"J": 6}
data["color"].replace(new_colors, inplace=True)
data.head()
```

ข้อมูล Color หลังแก้ไข

	carat	cut	color	clarity	depth	table	price	x	y	z
0	0.23	Ideal	1	SI2	61.5	55.0	326	3.95	3.98	2.43
1	0.21	Premium	1	SI1	59.8	61.0	326	3.89	3.84	2.31
2	0.23	Good	1	VS1	56.9	65.0	327	4.05	4.07	2.31
3	0.29	Premium	5	VS2	62.4	58.0	334	4.20	4.23	2.63
4	0.31	Good	6	SI2	63.3	58.0	335	4.34	4.35	2.75
5	0.24	Very Good	6	VVS2	62.8	57.0	336	3.94	3.96	2.48
6	0.24	Very Good	5	VVS1	62.3	57.0	336	3.95	3.98	2.47
7	0.26	Very Good	4	SI1	61.9	55.0	337	4.07	4.11	2.53
8	0.22	Fair	1	VS2	65.1	61.0	337	3.87	3.78	2.49
9	0.23	Very Good	4	VS1	59.4	61.0	338	4.00	4.05	2.39

3. Clean Data

เปลี่ยนจาก เชิงคุณภาพ -> เชิงปริมาณ

ข้อมูลก่อนแก้ไข

	carat	cut	color	clarity	depth	table	price	x	y	z
0	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...	...
53935	0.72	Ideal	D	SI1	60.8	57.0	2757	5.75	5.76	3.50
53936	0.72	Good	D	SI1	63.1	55.0	2757	5.69	5.75	3.61
53937	0.70	Very Good	D	SI1	62.8	60.0	2757	5.66	5.68	3.56
53938	0.86	Premium	H	SI2	61.0	58.0	2757	6.15	6.12	3.74
53939	0.75	Ideal	D	SI2	62.2	55.0	2757	5.83	5.87	3.64

ข้อมูลหลังแก้ไข

	carat	cut	color	clarity	depth	table	price	x	y	z
0	0.23	0	1	7	61.5	55.0	326	3.95	3.98	2.43
1	0.21	1	1	6	59.8	61.0	326	3.89	3.84	2.31
2	0.23	3	1	4	56.9	65.0	327	4.05	4.07	2.31
3	0.29	1	5	5	62.4	58.0	334	4.20	4.23	2.63
4	0.31	3	6	7	63.3	58.0	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...	...
53935	0.72	0	0	6	60.8	57.0	2757	5.75	5.76	3.50
53936	0.72	3	0	6	63.1	55.0	2757	5.69	5.75	3.61
53937	0.70	2	0	6	62.8	60.0	2757	5.66	5.68	3.56
53938	0.86	1	4	7	61.0	58.0	2757	6.15	6.12	3.74
53939	0.75	0	0	7	62.2	55.0	2757	5.83	5.87	3.64

4. ตรวจสอบความสัมพันธ์ของข้อมูล

```
1  data.corr()
```

Correlation

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

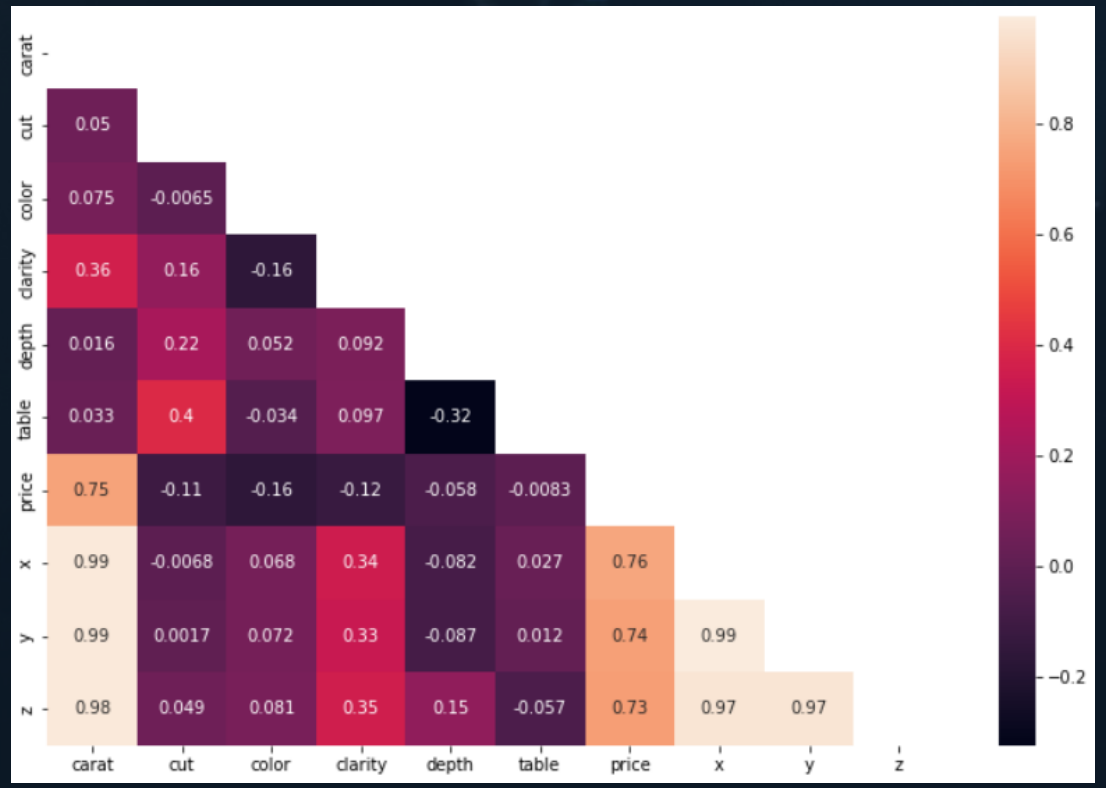
Pearson's Similarity

$$\frac{\sum_{i=1}^{n}(A_i - \overline{A})(B_i - \overline{B})}{\sqrt{\sum_{i=1}^{n}(A_i - \overline{A})^2(B_i - \overline{B})^2}}$$

	carat	cut	color	clarity	depth	table	price	x	y	z
carat	1.000000	0.049582	0.074952	0.355949	0.016252	0.032659	0.747830	0.987340	0.986201	0.984408
cut	0.049582	1.000000	-0.006503	0.159045	0.221849	0.399029	-0.111005	-0.006807	0.001672	0.049201
color	0.074952	-0.006503	1.000000	-0.163526	0.051920	-0.033861	-0.162964	0.068171	0.071632	0.081010
clarity	0.355949	0.159045	-0.163526	1.000000	0.091535	0.097027	-0.124267	0.339305	0.326861	0.352613
depth	0.016252	0.221849	0.051920	0.091535	1.000000	-0.324234	-0.057646	-0.082016	-0.087250	0.152262
table	0.032659	0.399029	-0.033861	0.097027	-0.324234	1.000000	-0.008267	0.027399	0.011792	-0.057083
price	0.747830	-0.111005	-0.162964	-0.124267	-0.057646	-0.008267	1.000000	0.759364	0.737978	0.730446
x	0.987340	-0.006807	0.068171	0.339305	-0.082016	0.027399	0.759364	1.000000	0.992551	0.970394
y	0.986201	0.001672	0.071632	0.326861	-0.087250	0.011792	0.737978	0.992551	1.000000	0.969128
z	0.984408	0.049201	0.081010	0.352613	0.152262	-0.057083	0.730446	0.970394	0.969128	1.000000

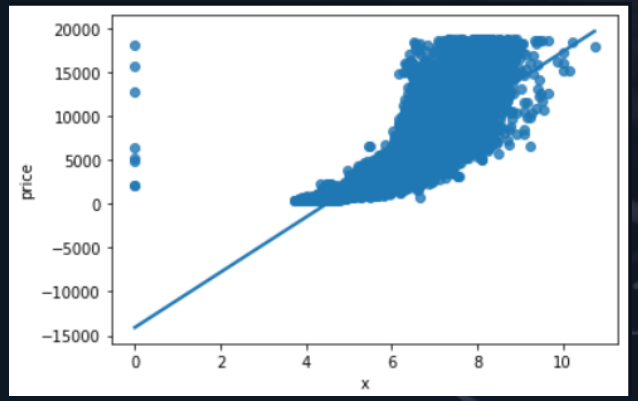
4. ตรวจสอบความสัมพันธ์ของข้อมูล

```
mask = np.triu(np.ones_like(data.corr()))
plt.figure(figsize=(12, 8))
sns.heatmap(data.corr(), annot=True, mask=mask)
```

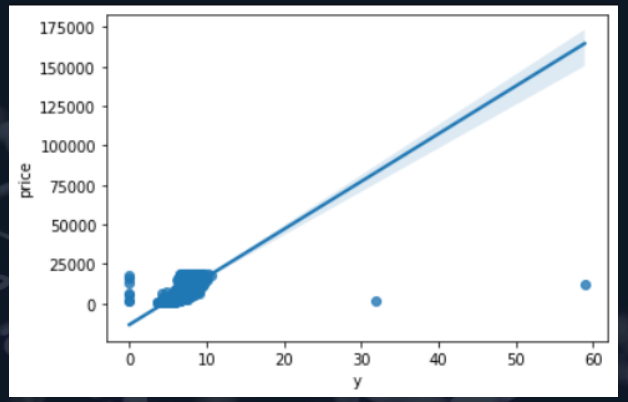


4. นำข้อมูลที่ได้มาทำ regplot

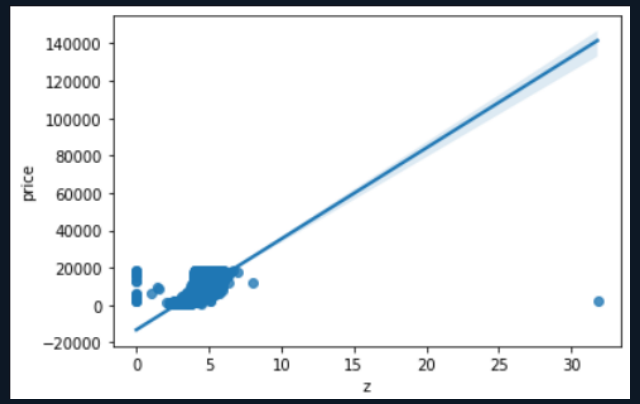
Price - Length



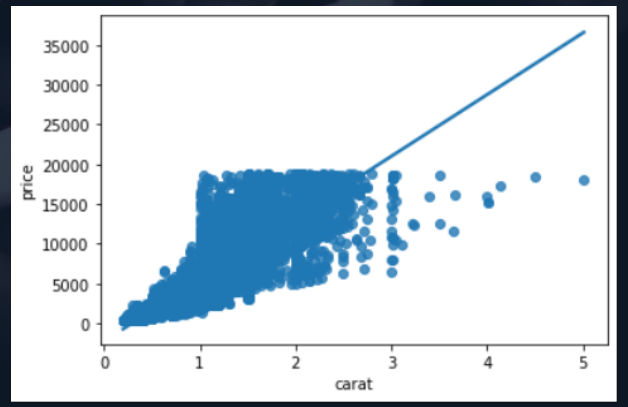
Price - Width



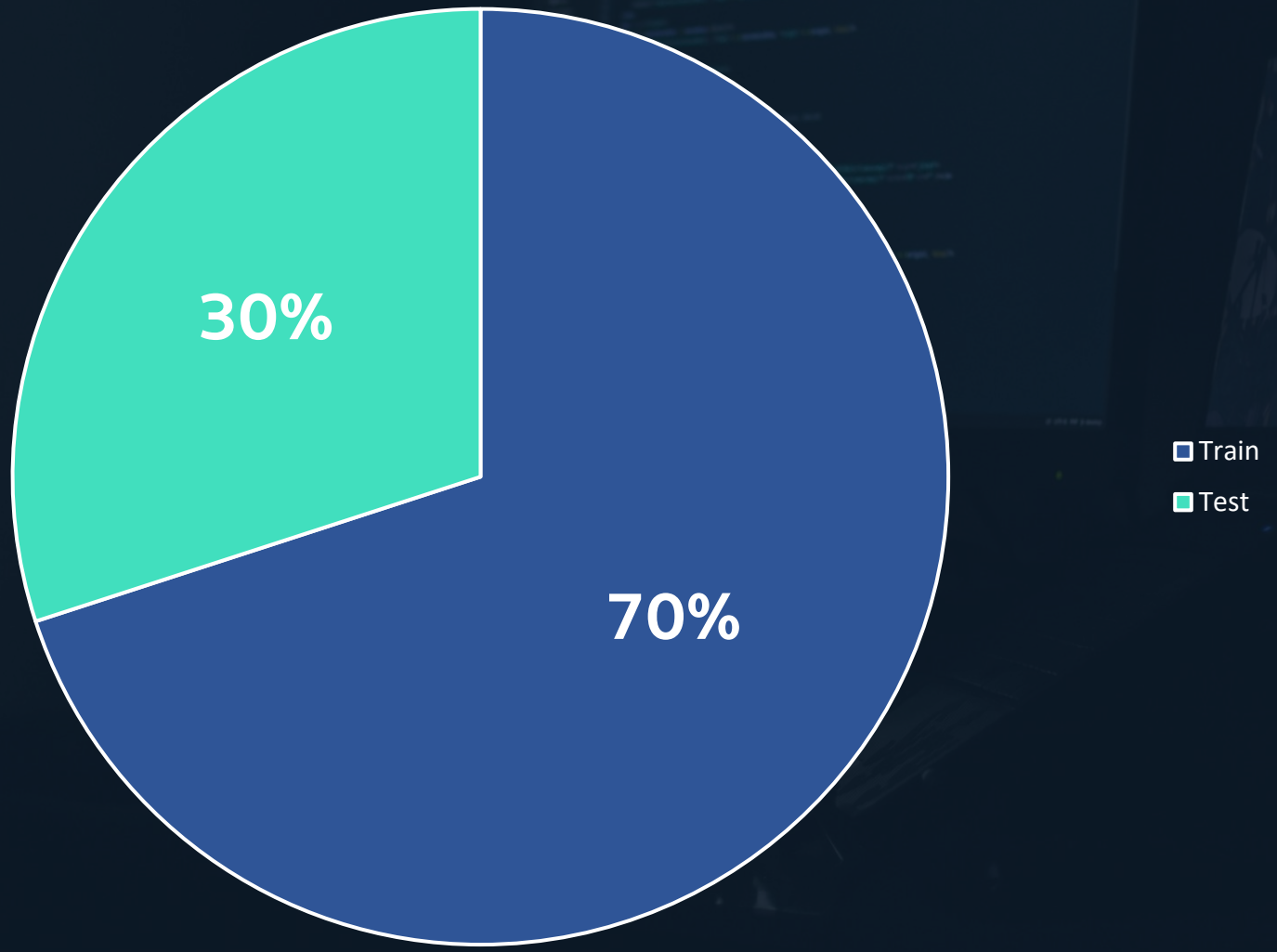
Price - Height



Price - Carat



5. แบ่งข้อมูลสำหรับใช้ Train และ Test ด้วย train_test_split()



6. นำข้อมูลเข้าโมเดล OLS (Ordinary Least Squares)

OLS Regression Results

Dep. Variable:	price	R-squared:	0.909
Model:	OLS	Adj. R-squared:	0.909
Method:	Least Squares	F-statistic:	4.181e+04
Date:	Wed, 17 Nov 2021	Prob (F-statistic):	0.00
Time:	18:15:59	Log-Likelihood:	-3.2134e+05
No. Observations:	37758	AIC:	6.427e+05
Df Residuals:	37748	BIC:	6.428e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	8520.1950	576.800	14.772	0.000	7389.652	9650.738
carat	1.066e+04	60.115	177.394	0.000	1.05e+04	1.08e+04
x	-818.3279	68.793	-11.896	0.000	-953.164	-683.492
y	151.2325	42.080	3.594	0.000	68.755	233.710
z	-237.5618	99.262	-2.393	0.017	-432.118	-43.006
table	-25.6606	3.496	-7.339	0.000	-32.514	-18.807
depth	-56.1960	7.684	-7.314	0.000	-71.257	-41.135
cut	-119.8788	6.743	-17.777	0.000	-133.096	-106.662
color	-327.7898	3.855	-85.032	0.000	-335.345	-320.234
clarity	-505.8815	4.092	-123.613	0.000	-513.903	-497.860

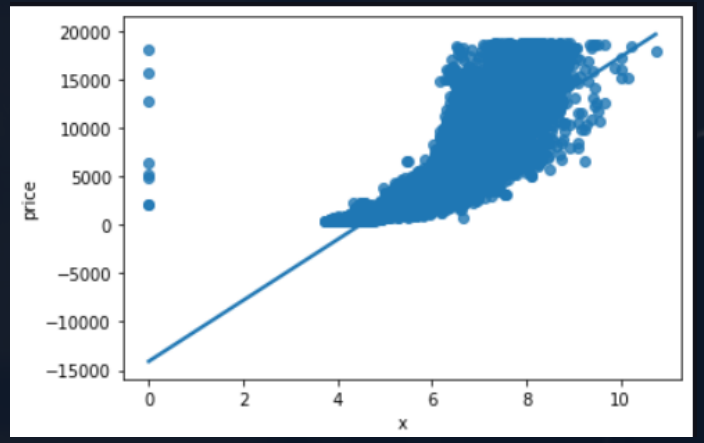
Omnibus:	8459.267	Durbin-Watson:	2.008
Prob(Omnibus):	0.000	Jarque-Bera (JB):	430466.337
Skew:	-0.019	Prob(JB):	0.00
Kurtosis:	19.541	Cond. No.	7.99e+03

	index	price	predict
16162	37823	1002	1157.310209
16163	41780	1247	809.286387
16164	19130	7899	9154.458815
16165	10700	4847	5060.747029
16166	2493	3196	4037.531337
16167	10366	4773	4770.821088
16168	21874	9942	8196.363740
16169	27717	648	-212.472612
16170	41174	1200	1991.792297
16171	3440	3387	3554.031256
16172	21203	9346	8244.925381

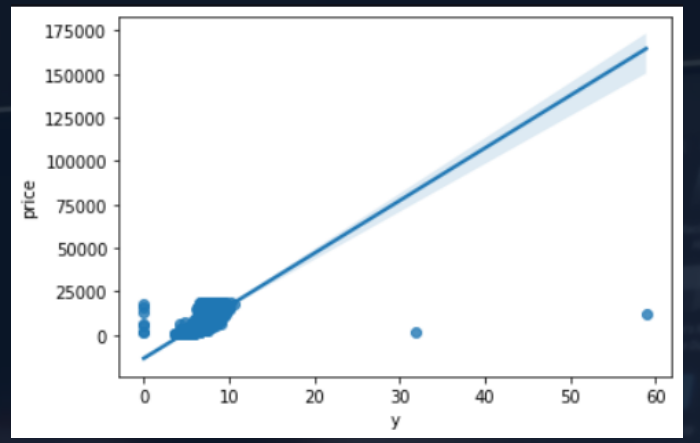
```
summm = 0
for i in predict_data.index:
    d = predict_data["price"][i] - predict_data["predict"][i]
    summm += abs(d)/predict_data["price"][i]

mape=(summm / len(predict_data)) * 100
print("mape =",mape)
```

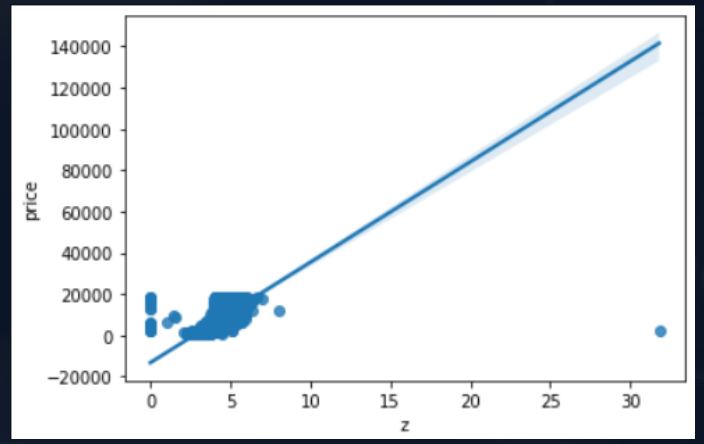
```
mape = 44.12156000705811
```



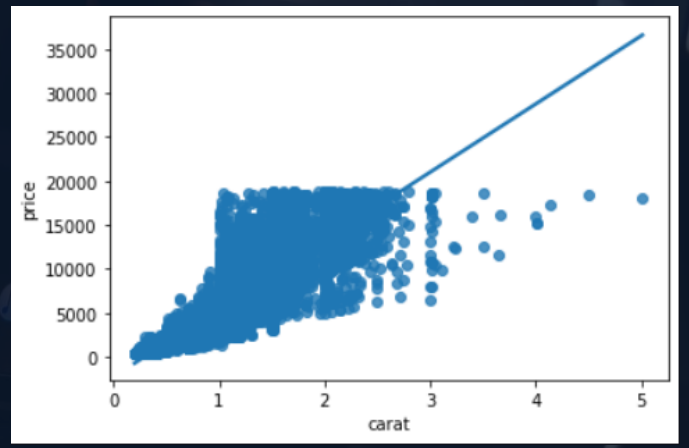
Price - Length



Price - Width



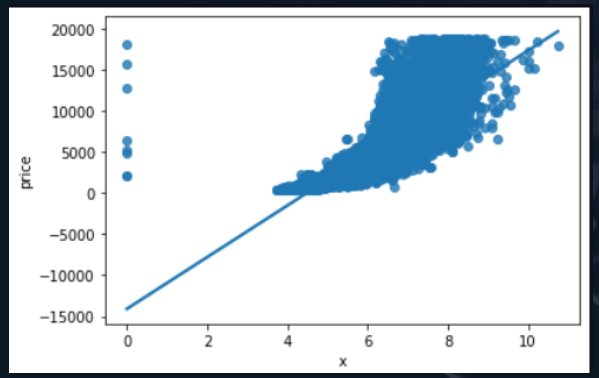
Price - Height

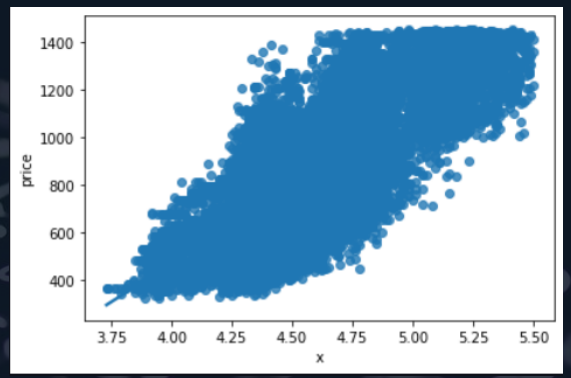


Price - Carat

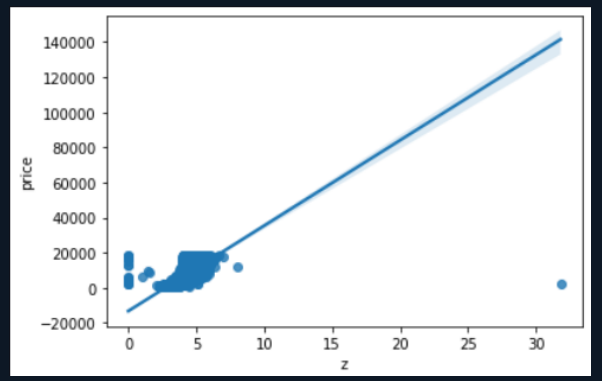
7. นำข้อมูลสุดโต่งออก

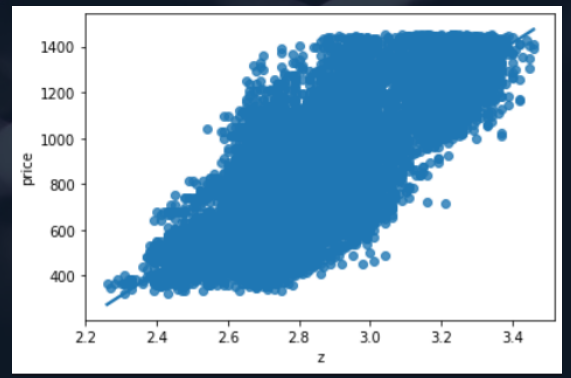
Price - Length





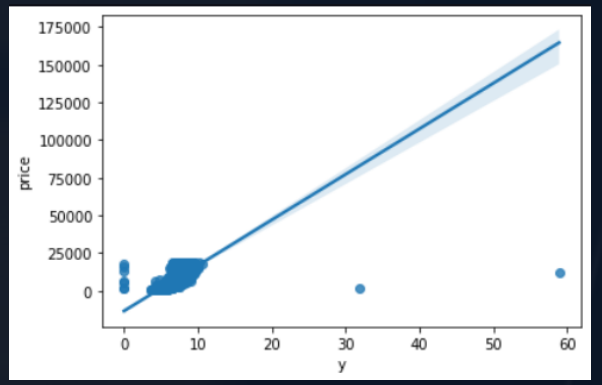
Price - Height

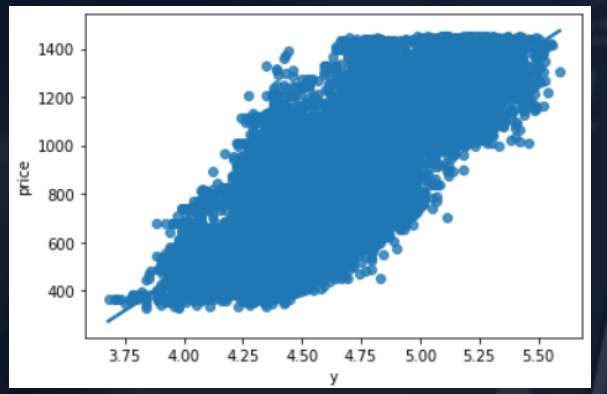




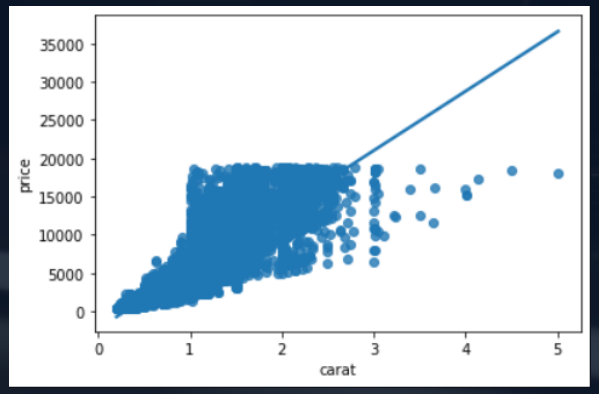
7. นำข้อมูลสุดโต่งออก

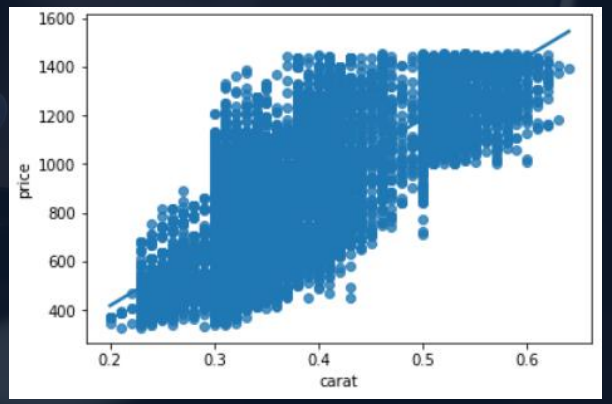
Price – Width





Price - Carat





8. นำข้อมูลเข้าโมเดล OLS (Ordinary Least Squares)

OLS Regression Results

Dep. Variable:	price	R-squared:	0.879
Model:	OLS	Adj. R-squared:	0.879
Method:	Least Squares	F-statistic:	1.077e+04
Date:	Wed, 17 Nov 2021	Prob (F-statistic):	0.00
Time:	20:50:59	Log-Likelihood:	-79262.
No. Observations:	13351	AIC:	1.585e+05
Df Residuals:	13341	BIC:	1.586e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2215.3961	575.603	-3.849	0.000	-3343.659	-1087.133
carat	3877.2509	103.613	37.420	0.000	3674.154	4080.348
x	1721.7128	65.168	26.420	0.000	1593.975	1849.451
y	-1136.8569	64.945	-17.505	0.000	-1264.158	-1009.556
z	-1190.1302	199.100	-5.978	0.000	-1580.395	-799.865
table	-1.7343	0.531	-3.268	0.001	-2.775	-0.694
depth	47.4263	9.178	5.167	0.000	29.436	65.416
cut	-16.5138	0.998	-16.555	0.000	-18.469	-14.558
color	-51.5837	0.506	-101.985	0.000	-52.575	-50.592
clarity	-81.4655	0.522	-156.182	0.000	-82.488	-80.443

Omnibus:	350.292	Durbin-Watson:	1.974
Prob(Omnibus):	0.000	Jarque-Bera (JB):	755.620
Skew:	0.140	Prob(JB):	8.30e-165
Kurtosis:	4.131	Cond. No.	6.52e+04

```
summm = 0
for i in predict_data.index:
    d = predict_data["price"][i] - predict_data["predict"][i]
    summm += abs(d) / predict_data["price"][i]

mape=(summm / len(predict_data))*100
print("mape =",mape)
```

```
mape = 8.65186738879971
```

ตารางข้อมูล และราคาที่ทำนายได้

Length Width Height

	const	carat	x	y	z	table	depth	cut	color	clarity
13441	1.0	0.40	4.80	4.76	2.89	59.0	60.5	1	0	5
5416	1.0	0.31	4.35	4.31	2.67	59.0	61.7	1	0	6
7331	1.0	0.36	4.63	4.58	2.79	56.0	60.6	0	0	6
9141	1.0	0.30	4.28	4.30	2.68	56.0	62.5	0	1	2
8798	1.0	0.42	4.79	4.82	2.96	58.0	61.6	1	4	4
...	...	...	...	...	...	...	...	...	...	...
3894	1.0	0.30	4.29	4.32	2.68	58.0	62.3	1	4	2
14735	1.0	0.33	4.40	4.43	2.74	58.0	62.1	1	0	7
11421	1.0	0.32	4.41	4.38	2.68	59.0	61.0	1	3	1
3643	1.0	0.31	4.42	4.46	2.65	55.0	59.7	2	2	4
12045	1.0	0.50	5.08	5.11	3.16	55.0	62.0	0	5	6

	index	price	predict
0	13441	1050	1091.936728
1	5416	732	717.073696
2	7331	794	912.796608
3	9141	862	891.184598
4	8798	847	929.778020
...	...	...	...
5718	3894	684	701.445828
5719	14735	492	600.211732
5720	11421	952	987.045678
5721	3643	679	646.202061
5722	12045	982	997.712783

https://predictdiamondprice.netlify.app/

แนวทางการพัฒนาต่อ

1. เพิ่ม**ความแม่นยำ**ในการทำนาย
2. **อัพเดท**ข้อมูลให้เป็นปัจจุบัน
3. นำทฤษฎีทางคณิตศาสตร์อื่นๆ มา**ประยุกต์**เพิ่ม
4. เพิ่มประเภทของ**อัญมณี**ที่สามารถทำนายได้

THANK YOU.

Presented By.

Jirapat Kaewsongsang	63010139
Chinathip Meesuannil	63010235
Nuttapong Naksamukkee	63010326
Ditthaphong Jaratchairot	63010354