M6-02: Coefficient Coefficient in Python

Part of the "Towards Machine Learning" Learning Badge

Video Walkthrough: ***https://discovery.cs.illinois.edu/m6-02/***

Classical Dataset: "The Diamond Dataset"

Source: https://ggplot2.tidyverse.org/reference/diamonds.html

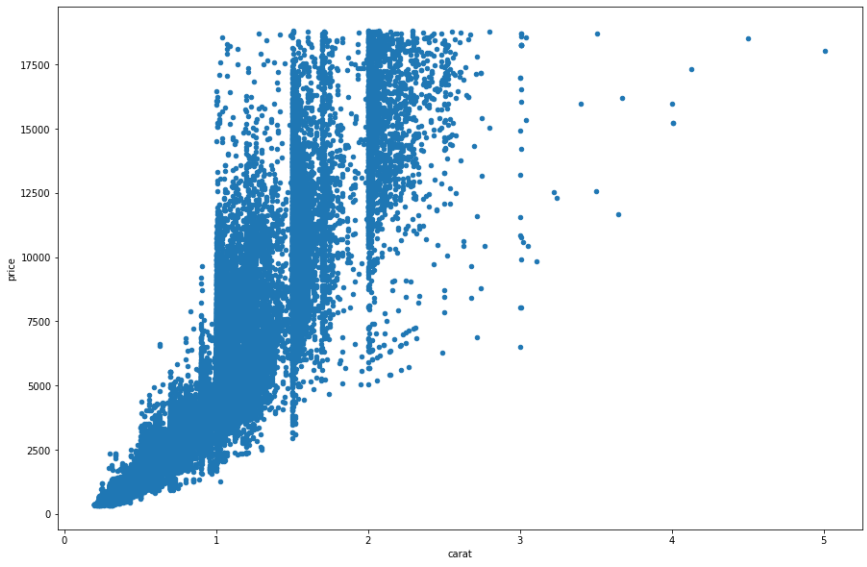
Description: A dataset containing the prices and other attributes of almost 54,000 diamonds. The dataset includes ten different columns of data:

- **price**, price in US dollars ($326 - $18,823)
- **carat**, weight of the diamond (0.2 - 5.01)
- **cut**, quality of the cut (Fair, Good, Very Good, Premium, Ideal)
- **color**, diamond colour, from J (worst) to D (best)
- **clarity**, a measurement of how clear the diamond is (I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
- **x**, length in mm (0 - 10.74)
- **y**, width in mm (0 - 58.9)
- **z**, depth in mm (0 - 31.8)
- **depth**, total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
- **table**, width of top of diamond relative to widest point (43--95)
- Available: https://waf.cs.illinois.edu/discovery/diamonds.csv

Exploratory Data Analysis

With such a large dataset, it's worth exploring some trends in this dataset. I'm specifically interested in the relationship between the size (carat) and price of diamonds:

Scatter plot of the "Diamond Dataset"




```
df.plot.scatter(x='carat', y='price')
```

Puzzle #1: What data is mapped to the x-axis of this scatter plot?

Puzzle #2: What data is mapped to the y-axis of this scatter plot?

Puzzle #3: Where is the color of the diamond displayed on this scatter plot?

M6-02: Coefficient Coefficient in Python

Part of the "Towards Machine Learning" Learning Badge

Video Walkthrough: *https://discovery.cs.illinois.edu/m6-02/*

Correlation Coefficient in Python

In Python, the following code will display the correlation coefficient for every numeric column (variable) in a DataFrame:

Python:

```
df = pd.read_csv(
  "https://waf.cs.illinois.edu/discovery/diamonds.csv")

df.corr()
```

Result:

	carat	depth	table	price	x	y	z
carat	1.00	0.03	0.18	0.92	0.98	0.95	0.95
depth	0.03	1.00	-0.26	-0.01	-0.03	-0.03	-0.09
table	0.18	-0.26	1.00	0.13	0.20	0.18	0.15
price	0.92	-0.01	0.13	1.00	0.88	0.87	0.86
x	0.98	-0.03	0.20	0.88	1.00	0.97	0.97
y	0.95	-0.03	0.18	0.87	0.97	1.00	0.95
z	0.95	-0.09	0.15	0.86	0.97	0.95	1.00

Observations:

1. This table has a special name:

2. What is special about the **main diagonal** of this matrix?

3. What does a correlation coefficient of **1.00** *(r=1)* mean?

4. What is **always true** about the **upper triangular** region and the **lower triangular** region?

5. The correlation coefficient between **carat** and **price** is _____. What does this tell us?

6. The correlation coefficient between **table** and **depth** is _____. What does this tell us?

7. The correlation coefficient between **depth** and **price** is _____. What does this tell us?