DISCOVERY

#14: Scatter Plots

February 18, 2019

Karle Flanagan and Wade Fagen-Ulmschneider

Scatter Plots provide a way to visually represent two columns of numeric data in one visualization. After creating several scatter plots, we will learn different methods to quantify the relationship between the data (ex: **correlation**).

Classical Example: "Diamond Dataset"

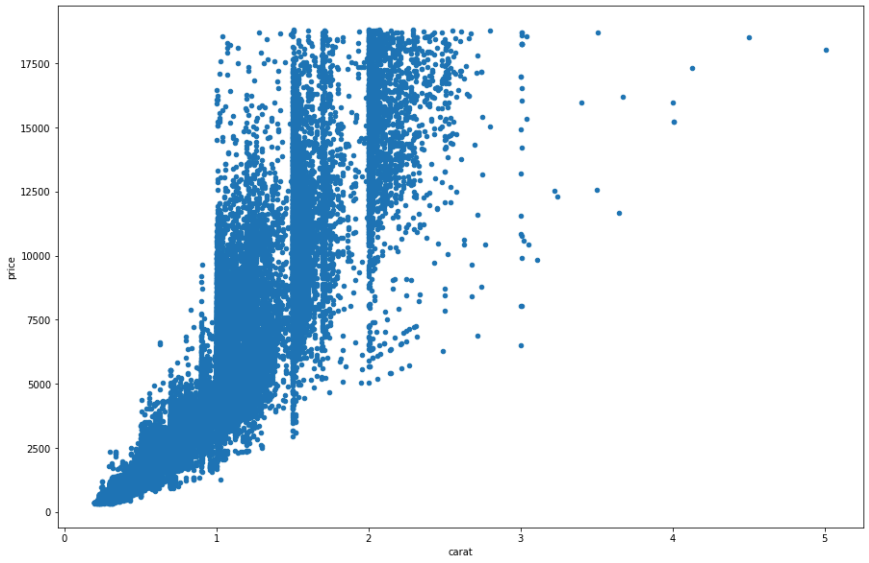
Source: https://ggplot2.tidyverse.org/reference/diamonds.html

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)

Google Sheets Dataset:

https://docs.google.com/spreadsheets/d/1VSYPeI6SCF6JhzqaIcN5PvjH60xI7PrsOROCvfbz2rY/edit?usp=sharing




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

Questions:

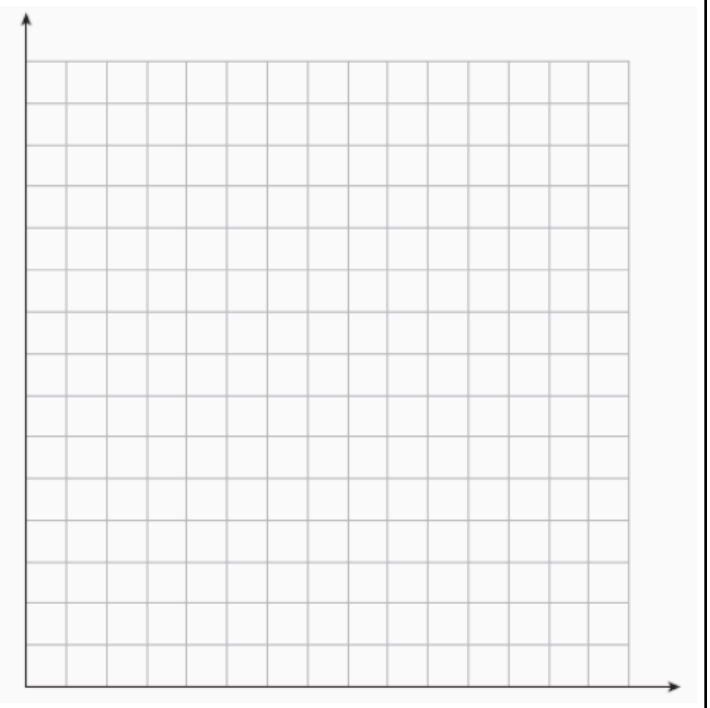
What data is mapped to the x-axis of this scatter plot?

What data is mapped to the y-axis of this scatter plot?

Where is the color of the diamond displayed on this scatter plot?

Illinois Example: Faculty and Student Body Size over 15 Years

The following dataset shows the number of full-time faculty members and the number of students at The University of Illinois over the past 20 years:

Year	FT Faculty	Students	Scatter Plot
2018	3,005	49,339	
2017	2,392	47,826	
2016	2,358	46,496	
2015	2,405	45,375	
2014	2,362	44,754	
2013	2,316	44,580	
2012	2,245	44,197	
2011	2,244	44,067	
2010	2,338	43,601	
2009	2,289	43,522	
2008	2,313	42,914	
2007	2,308	42,102	
2006	2,322	42,093	
2005	2,271	41,269	
2004	2,158	40,322	

Q1: What data will you place on the x-axis?
...what is the **range** of that data?

Q2: What data will you place on the y-axis?
...what is the **range** of that data?

Now create the scatter plot!

Scatter Plots: Python Documentation

Using Python, we can easily create scatter plots with datasets:

Python:	**General Syntax:**
	`df.plot.scatter()`
	
	Options:
	`x = 'Column Name'`
	`y = 'Column Name'`
	
	Example Shown:
	`df = pd.read_csv('diamonds.csv')`
	`df.plot.scatter(x='carat', y='price')`