Scatter Plots

How We Visualize Correlation Between Variables

Plotting Datasets

- We've imported the CSV into a dataframe.
- Now what?

- We can plot the dataset to look for patterns between **variables**.

Q & A: What are the variables in the CSV shown?

```
GRE,GPA,Gender
316,3.4,M
308,3.1,M
327,3.7,F
310,3.33,F
305,3.45,M
322,3.18,F
316,3.25,M
300,3.4,F
310,3.6,F
```

Plotting Datasets

- We've imported the CSV into a dataframe.
- Now what?

- We can plot the dataset to look for patterns between **variables**. These are called **correlations**.

Q & A: What are the variables in the CSV shown?

GRE, GPA, Gender

```
GRE,GPA,Gender
316,3.4,M
308,3.1,M
327,3.7,F
310,3.33,F
305,3.45,M
322,3.18,F
316,3.25,M
300,3.4,F
310,3.6,F
```

Detecting Correlations

- Datasets usually come out of research studies which have a goal.
- Remember rows are **observations**!

Q & A: What could have been the original goal for the shown dataset?

```
GRE,GPA,Gender
316,3.4,M
308,3.1,M
327,3.7,F
310,3.33,F
305,3.45,M
322,3.18,F
316,3.25,M
300,3.4,F
310,3.6,F
```

Detecting Correlations

- Datasets usually come out of research studies which have a goal.
- Remember rows are **observations**!

Q & A: What could have been the original goal for the shown dataset?

Does a student's GPA depend on GRE or Gender?

```
GRE,GPA,Gender
316,3.4,M
308,3.1,M
327,3.7,F
310,3.33,F
305,3.45,M
322,3.18,F
316,3.25,M
300,3.4,F
310,3.6,F
```

Detecting Correlations

- To determine correlation, classify variables as **independent** or **dependent** based on the goal.
- **Independent variables** are what **dependent variables** depend on.
- For “Does a student’s GPA depend on GRE or Gender?”:
  - Dependent: GPA
  - Independent: GRE and Gender

```
GRE,GPA,Gender
316,3.4,M
308,3.1,M
327,3.7,F
310,3.33,F
305,3.45,M
322,3.18,F
316,3.25,M
300,3.4,F
310,3.6,F
```

Detecting Correlations

"Does a student's GPA depend on GRE or Gender?"

- Possible charts:
  - GPA vs GRE
  - GPA vs Gender

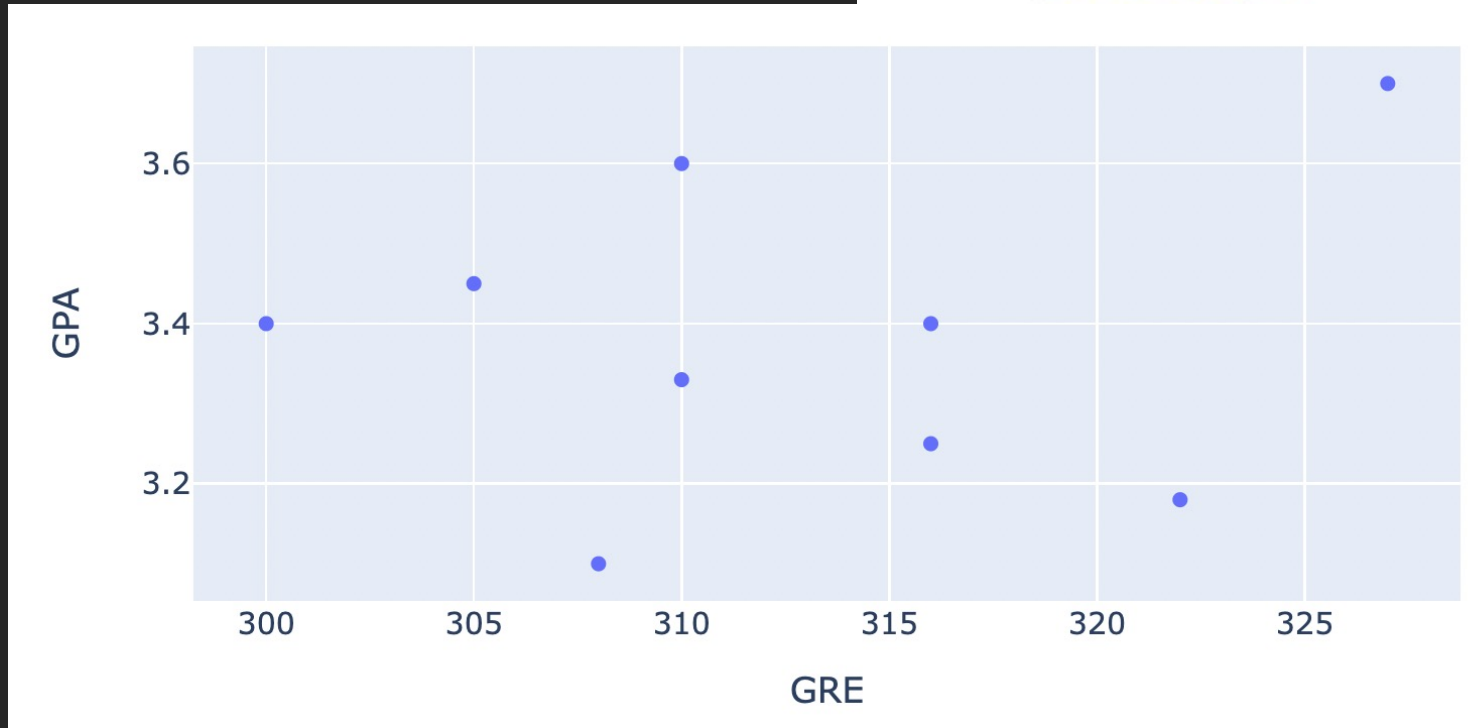
- When you have two variables with numeric values, you can make a **scatter plot.**

```
GRE,GPA,Gender
316,3.4,M
308,3.1,M
327,3.7,F
310,3.33,F
305,3.45,M
322,3.18,F
316,3.25,M
300,3.4,F
310,3.6,F
```

Scatter Plots

- Plot each **datapoint**
- **Independent variable** on x-axis
- **Dependent variable** on y-axis
- (x, y) is now (GRE, GPA)
- Let's find (316, 3.4)
- Look! Points are not in order on graph

```
GRE,GPA,Gender
316,3.4,M
308,3.1,M
327,3.7,F
310,3.33,F
305,3.45,M
322,3.18,F
```



How-to: Design Scatter Plots

Given dataset and goal/question:

1. Identify variables relevant to goal
2. Classify variables as independent or dependent
3. Set up the axes and plot each datapoint
4. Observe the distribution of points and determine correlation

Step 4 is hard. We will cover it in a later session.

How-to: Make Scatter Plots in Jupyter & Blockly

Step 1: Read CSV Data into Pandas Dataframe

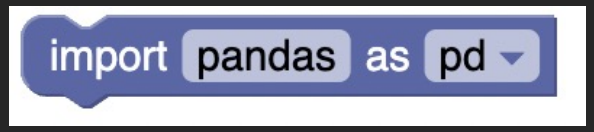
This step creates a Dataframe and stores the content of the dataset into a variable so that we can use this in later steps.

- Substep: Import pandas Library

  Python:

```
import pandas as pd
```

  Blockly:



To read from a csv file, first we will import the pandas library. It has a read_csv function which we will use to automatically parse the csv file and load it into the notebook.

How-to: Make Scatter Plots in Jupyter & Blockly (Step 1 Cont.)

- Substep: Read CSV data and Save in Variable

Python:

```
df = pd.read_csv('datasets/age_height.csv')
```

Blockly:

The read_csv function requires us to supply the relative path to the csv file. It returns a Pandas Dataframe object which we will store in a variable df so we can use it later in the notebook.

How-to: Make Scatter Plots in Jupyter & Blockly (Step 1 Cont.)

- Substep: Display Dataframe Contents

Python:

```
df
```

Blockly:

Output:

	Height	Age	Gender
0	151.765	63.0	male
1	139.700	63.0	female
2	136.525	65.0	female
3	156.845	41.0	male
4	145.415	51.0	female
...	...	...	...
539	145.415	17.0	male
540	162.560	31.0	male
541	156.210	21.0	female
542	71.120	0.0	male
543	158.750	68.0	male

544 rows × 3 columns

Calling the variable in a cell by itself will print the contents of the dataframe to the screen so we can confirm the dataset was imported correctly.

How-to: Make Scatter Plots in Jupyter & Blockly

Step 2: Generate Plotly Scatter Plot

This step uses the content of the dataframe to generate a scatter plot.

- Substep: Import plotly.express Library

  Python:

```
import plotly.express as px
```

  Blockly:

To make a scatter plot, first we will import plotly.express library. It has a scatter function which we will use to make the scatter plot.

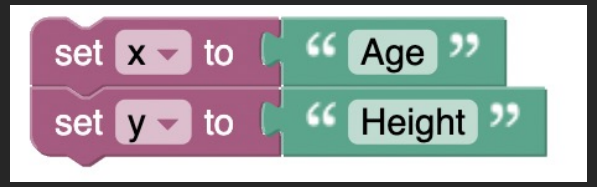
How-to: Make Scatter Plots in Jupyter & Blockly (Step 2 Cont.)

- Substep: Set Columns as x and y

  Python:

```
x = 'Age'
y = 'Height'
```

  Blockly:



The scatter function requires us to supply the names of the columns for the independent (x) and dependent (y) variables. Here “Height” is dependent on “Age” so we will set a variable x to “Age” and a variable y to “Height.”

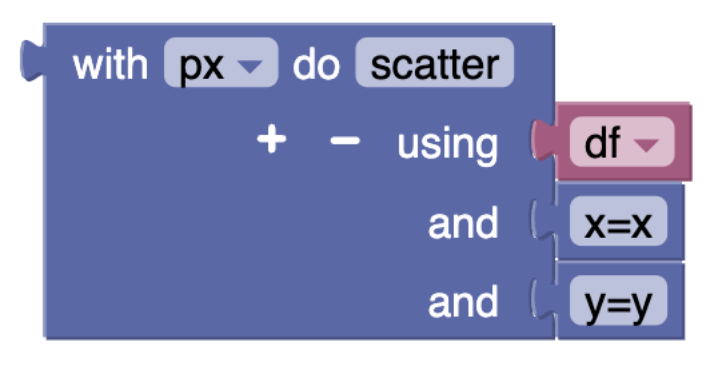
How-to: Make Scatter Plots in Jupyter & Blockly (Step 2 Cont.)

- Substep: Generate scatter plot:

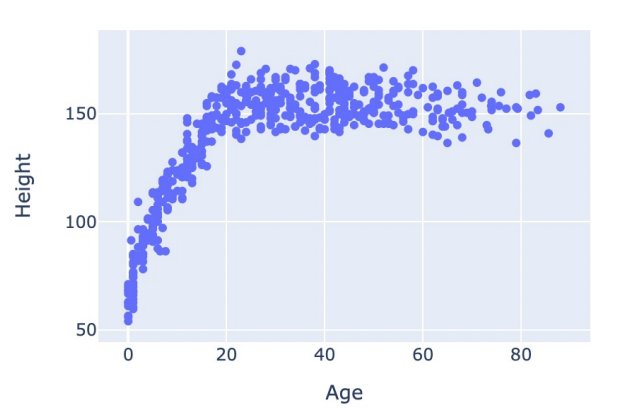
Python:

```
px.scatter(df,x=x,y=y)
```

Blockly:



Output:



The scatter function requires us to supply the dataframe variable and column names for the x axis and y axis which we previously stored in variables x and y.
This function returns a scatter plot.

Reference Notebook

- scatterplots_ex.ipynb

Summary

- Plotting datasets
- Independent and dependent variables
- Scatter Plots
- Making scatter plots with Blockly