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**.

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?

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?"

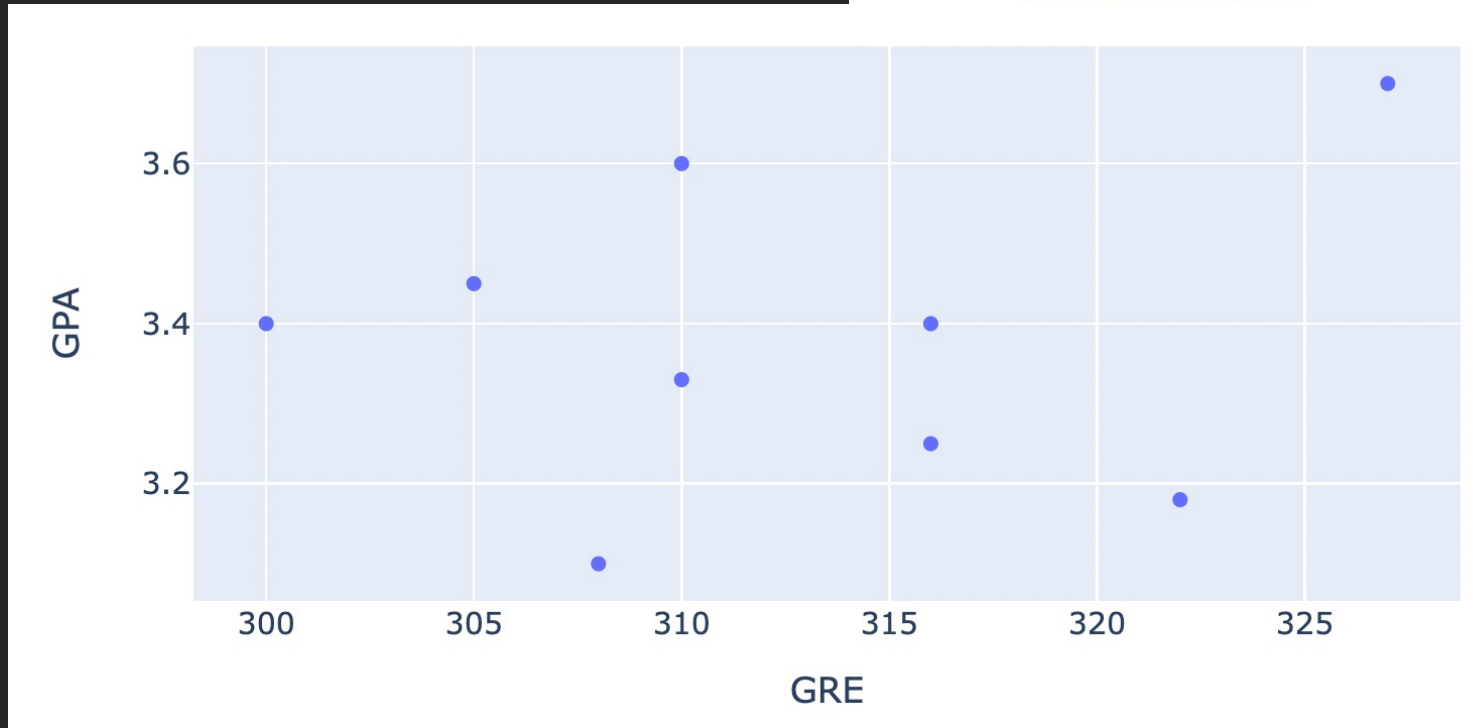
- Possible charts:
  - GPA vs GRE
  - GPA vs Gender
- When you have two variables with numeric values, you can make a **scatter plot**

```
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

1. Read CSV Data into Pandas Dataframe
   - Import Pandas Library
   - Read CSV data and Save in Variable
   - Display Dataframe Contents
2. Generate Plotly Scatter Plot
   - Import Plotly Express Library
   - Set Columns as x and y
   - Set Additional Plot Options (Optional)
   - Generate Chart

Activity 1: Scatterplots E1

1. Open scatterplots_e1.ipynb
2. Answer Pre-Programming Discussion questions
3. In pods, go through programming section and do Try it's
   - Follow the instructions to replicate the example code using blocks for Blockly practice.

Activity 2: Scatterplots E2

1. Open scatterplots_e2.ipynb
2. Answer Pre-Programming Discussion questions
3. In pods, do Programming section
4. Upload screenshot of final plot in Discord

Summary

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