M2-03: Grouping Data in Python

Part of the "Exploratory Data Analysis" Learning Badge

Video Walkthrough: *https://discovery.cs.illinois.edu/m2-03/*

Aggregation of Groups

Nearly all analysis we will perform on groups will require us to **aggregate** the groups together into a new DataFrame, containing **summaries** of each group.

Three aggregations are extremely common in data science:

Type	Python Code	Every variable is summarized for each group, producing the following summary:
Count	`df.groupby("Subject")` `  .agg(  ).reset_index()`	Produces a **count** of the number of rows that contain data. Each row will only ever contribute one to the count -- the actual value of the data does not matter, just the existence of data.
Sum	`df.groupby("Subject")` `  .agg(  ).reset_index()`	Produces the **sum** of all observations for each variable within the group.
Mean (Average)	`df.groupby("Subject")` `  .agg(  ).reset_index()`	Produces the **average (mean) value** of all observations within each group for each variable within the dataset.
Other Common Aggregations: `"median", "min", "max", "std", "var", ...and more!`		

Let's discovery happens when we aggregate by "Subject", using the GPA dataset:

Type	Result	Interpretation
Count		
Sum		
Mean (Average)		