General Assembly

STUDY GUIDE

SUMMARIZING AND GROUPING

Key Terms and Definitions

» **Some Useful Pandas Attributes and Methods, Listed:**

» **Exploring Data**:

- The **.shape** attribute is used to return the number of rows and columns.
- The **.dtypes** attribute is used to return the type of data stored in each column.

» **Changing Data Types**:

- The **.to_numeric()** Pandas function automatically changes the data type of a**DataFrame** column to a numeric format. It can be used on multiple columns.
- **.astype()** is similar to **.to_numeric()** but allows for non-numeric conversion, albeit with some limitations.

» **Summary Statistics**:

- Pandas provides a rich set of methods for retrieving statistics on the data within a DataFrame, including**mean()**, **.median()**, **.min()**, and **.max()**.
- The **.count()** method is used to return a count of the non-NaN (non-Null) rows in a column.
- **.describe()** returns the count, mean, and standard deviation along with various percentiles.

» **Aggregating Data**:

- **.groupby()** enables the split-apply-combine paradigm by splitting the data, grouping it by a particular column's values, and returning an aggregation (which can be any aggregation supported in Pandas, such as median, sum, min, max, etc.).
- **.pivot_table()** can take in a variable, value, and index to**.groupby()** and **.apply()** aggregate functions to summarize the data.

» **Manipulating Data**:

- **.apply()** allows us to run custom functions or functions from other packages on our Pandas objects.

Guiding Questions

1. If you were analyzing student test scores, what might be some ways to use**.groupby()** to provide insight?
2. What are some functions you might want to create and run using**.apply()**?

Additional Resources

1. DataCamp:
   » Manipulating DataFrames With Pandas. See Section 4, "Grouping Data." Make sure to focus on the topics "Categoricals & Groupby" and "Groupby & Aggregation."
2. Pandas DataFrame Documentation
3. GA Demo Video: Summarizing DataFrames