Pandas I

A Note on Delivery

- This unit's lessons will occur in jupyter notebooks
  - Slides will be an introduction to the lesson (no code, just overview)
  - Then, we'll open a notebook and start coding!

Learning Objectives

After this lesson, you will be able to:

- Use Pandas to read in a dataset.
- Investigate a dataset's integrity.
- Filter, sort, and manipulate DataFrame series.

What is Pandas?

- A group of adorable bears 🐼🐼🐼
- A Python library for data manipulation.

So, Pandas the Library

The Swiss Army Knife of data manipulation!

Pandas:

- Is *the* library for exploratory data analysis (EDA).
- Formats, wrangles, cleans, and prepares our data.

Quick Backstory from 2009:

- A humble open source project for Panel Data (hence “Pandas”) from Wes McKinney.
- A ‘panel’ is the name of the object (in pandas) holding an n-dimensional numpy array
- Don’t let the term fool you, a panel is effectively the same thing as an excel workbook (a collection of sheets)
- A 2-dimensional panel is a Dataframe (rows and columns)
- A 1-dimensional panel is a Series (column)

Exploratory Data Analysis (EDA)

The process of understanding our dataset and producing our first level of insights.

This includes:

- Reading in data: “Import cat population.”
- Checking data types. “Is the population count in integers?”
- Renaming columns: “`cat_breed` is more helpful than `Biological Family`”
- Joining together data: “Join the cat population data with the cat population data.”
- Looking for missing data: “It doesn’t mention corgis.”
- And more!

Today, we will focus on the most ‘mission critical’ elements of EDA.

Quick Review

- Exploratory Data Analysis (EDA) is the process of understanding our dataset, and producing our first level of insights.What does this include?
- Pandas is a prominent Python library used for exploratory data analysis

What dataset are we exploring?

- Adventure Works Cycles!
- We will be using a dataset developed by Microsoft for training purposes in SQL server, known the Adventureworks Cycles 2014OLTP Database.
- It is based on a fictitious company called Adventure Works Cycles (AWC), a multinational manufacturer and seller of bicycles and accessories.
- The company is based in Bothell, Washington, USA and has regional sales offices in several countries.
- We will be looking at a single table from this database, the Production.Product table, which outlines some of the products this company sells.

Discussion: What Could We Examine?

- What are some potential insights you'd like to uncover given the data?
- What if you are examining it from the standpoint of a the business?
- What if you are a potential distributor of their products?

Our Modified Adventure Works Dataset

The full dataset is actually a large, star-schema relational databse.

We will work with a modified dataset.

Key changes:

- Only a single table from this database
- Contains information on products the company makes
    - Such as the product names
    - The product weights, measures
    - And the product prices

Data Integrity

The first thing we check! Assuring our data can be trusted to produce meaningful insights.

Correctly formatted datatypes.

- “Decimals are floats, not strings.”

Missing Data

- i.e. “Why do we only have even days of the month?”

Clean Truth about Dirty Data

- Assessing data integrity isn’t a one-stop step.
- Much like EDA itself, it’s an ongoing process!
- We uncover additional potential problems and anomalies to remedy along the way.

Launch our notebook

We'll work in the Notebook - We're fledgling data scientists!

The `.ipynb` file you will open is called " `intro-to-pandas-i.ipynb` ".

Open it up!

Jump down to `Import`.

Additional Resources

- Pandas documentation
- DataSchool 30-video series (by a former GA instructor!)