Group Project

Project A: Data Source Quality

As a student of, or a new professional working in, data science, you will not always be collecting new primary data. It's just as important to be able to locate, critically evaluate, and properly clean existing sources of secondary data. (Collecting and Preparing Data will cover the topic of data collection and cleaning in more detail.)

Some reputable government data sources are:

Data.gov (https://openstax.org/r/datagov)

Bureau of Labor Statistics (BLS) (https://openstax.org/r/blsgov1)

National Oceanic and Atmospheric Administration (NOAA) (https://openstax.org/r/noaagov)

Some reputable nongovernment data sources are:

Kaggle (https://openstax.org/r/kaggle1)

Statista (https://openstax.org/r/statista)

Pew Research Center (https://openstax.org/r/pewresearch)

Using the suggested sources or similar-quality sources that you research on the Internet, find two to three datasets about the field or industry in which you intend to work. (You might also try to determine whether similar data sets are available at the national, state/province, and local/city levels.) In a group, formulate a specific, typical policy issue or business decision that managers in these organizations might make. For the datasets you found, compare and contrast their size, collection methods, types of data, update frequency and recency, and relevance to the decision question you have identified. Project B: Data Visualization

Using one of the data sources mentioned in the previous project, find a dataset that interests you. Download it as a CSV file. Use Python to read in the CSV file as a Pandas DataFrame. As a group, think of a specific question that might be addressed using this dataset, discuss which features of the data seem most important to answer your question, and then use the Python libraries Pandas and Matplotlib to select the features and make graphs that might help to answer your question about the data. Note, you will learn many sophisticated techniques for doing data analysis in later chapters, but for this project, you should stick to simply isolating some data and visualizing it using the tools present in this chapter. Write a brief report on your findings. Project C: Privacy, Ethics, and Bias

Identify at least one example from recent current events or news articles that is related to each of the

following themes (starting references given in parentheses):

a. Privacy concerns related to data collection (See the Protecting Personal Privacy (https://openstax.org/r/ gaog) website of the U.S. Government Accountability Office.)

b. Ethics concerns related to data collection, including fair use of copyrighted materials (See the U.S.

Copyright Office guidelines (https://openstax.org/r/fairuse).)

c. Bias concerns related to data collection (See the National Cancer Institute (NCI) article (https://openstax.org/r/bias) on data bias.)

Suppose that you are part of a data science team working for an organization on data collection for a major project or product. Discuss as a team how the issues of privacy, ethics, and equity (avoiding bias) could be addressed, depending on your position in the organization and the type of project or product.

Chapter Review

1. Select the incorrect step and goal pair of the data science cycle.

   a. Data collection: collect the data so that you have something for analysis.

   b. Data preparation: have the collected data stored in a server as is so that you can start the analysis.

   c. Data analysis: analyze the prepared data to retrieve some meaningful insights.

   d. Data reporting: present the data in an effective way so that you can highlight the insights found from the analysis.

2. Which of the following best describes the evolution of data management in the data science process?

   a. Initially, data was stored locally on individual computers, but with the advent of cloud-based systems, data is now stored on designated servers outside of local storage.

   b. Data management has remained static over time, with most data scientists continuing to store and process data locally on individual computers.

   c. The need for data management arose as a result of structured data becoming unmanageable, leading to the development of cloud-based systems for data storage.

d. Data management systems have primarily focused on analysis rather than processing, resulting in the development of modern data warehousing solutions.

3. Which of the following best exemplifies the interdisciplinary nature of data science in various fields?

a. A historian traveling to Italy to study ancient manuscripts to uncover historical insights about the

Roman Empire

b. A mathematician solving complex equations to model physical phenomena

c. A biologist analyzing a large dataset of genetic sequences to gain insights about the genetic basis of diseases

d. A chemist synthesizing new compounds in a laboratory  Critical Thinking

1. For each dataset (https://openstax.org/r/spreadsheetsd1), list the attributes.

a. Spotify dataset

b. CancerDoc dataset

2. For each dataset (https://openstax.org/r/spreadsheetsd1), define the type of the data based on following criteria and explain why:

- Numeric vs. categorical
- If it is numeric, continuous vs. discrete; if it is categorical, nominal vs. ordinal

a. “artist_count” attribute of Spotify dataset

b. “mode” attribute of Spotify dataset

1 • Quantitative Problems

c. “key” attribute of Spotify dataset

d. the second column in CancerDoc dataset

3. For each dataset (https://openstax.org/r/spreadsheetsd1), identify the type of the dataset—structured vs.

unstructured. Explain why. a. Spotify dataset

b. CancerDoc dataset

4. For each dataset (https://openstax.org/r/spreadsheetsd1), list the first data entry. a. Spotify dataset

b. CancerDoc dataset

5. Open the WikiHow dataset (ch1-wikiHow.json (https://openstax.org/r/filed)) and list the attributes of the dataset.

6. Draw scatterplot between bpm (x-axis) and danceability (y-axis) of the Spotify dataset (https://openstax.org/r/filed) using: a. Python Matplotlib

b. A spreadsheet program such as MS Excel or Google Sheets (Hint: Search "Scatterplot" on Help.)

7. Regenerate the scatterplot of the Spotify dataset (https://openstax.org/r/filed), but with a custom title and x-/y-axis label. The title should be "BPM vs. Danceability." The x-axis label should be titled "bpm" and range from the minimum to the maximum bpm value. The y-axis label should be titled "danceability" and range from the minimum to the maximum Danceability value.

   a. Python Matplotlib (Hint: DataFrame.min() and DataFrame.max() methods return min and max values of the DataFrame. You can call these methods upon a specific column of a DataFrame as well. For example, if a DataFrame is named df and has a column named "col1", df["col1"].min() will return the minimum value of the "col1" column of df.)

   b. A spreadsheet program such as MS Excel or Google Sheets (Hint: Calculate the minimum and maximum value of each column somewhere else first, then simply use the value when editing the scatterplot.)

8. Based on the Spotify dataset (https://openstax.org/r/spreadsheet4), filter the following using Python Pandas:

   a. Tracks whose artist is Taylor Swift

   b. Tracks that were sung by Taylor Swift and released earlier than 2020

Quantitative Problems

1. Based on the Spotify dataset (https://openstax.org/r/spreadsheet4), calculate the average bpm of the songs released in 2023 using: a. Python Pandas

b. A spreadsheet program such as MS Excel or Google Sheets (Hint: The formula AVERAGE() computes the average across the cells specified in the parentheses. For example, within Excel, typing in the command "=AVERAGE(A1:A10)" in any empty cell will calculate the numeric average for the contents of cells A1 through A10. Search "AVERAGE function" on Help as well.)