Principles of Data Science 1st edition Ault Solutions Manual

SKU: 9798385161850-SOLUTIONS

openstax™

Principles of Science

Principles of Data Science 1st edition Ault Solutions Manual

Chapter 1

What Are Data and Data Science?

Chapter Review

[1.1, LO 1.1.1, 1.1.2]

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.

Solution: b. Data preparation: have the collected data stored in a server as is so that you can start the analysis.
Rarely is collected data already in good shape for analysis. Most of the time, collected data needs to be processed to be suitable for the analysis of interest. An example of preparation can be dealing with missing data—removing them or filling them.

[1.1, LO 1.1.3]

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.

Solution: 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.

Principles of Data Science 1st edition Ault Solutions Manual

Data storage evolved from local to cloud-based systems for a variety of reasons, including increasing complexity of data, security concerns, reliability, etc. Option b) fails to recognize the evolution of data storage. Option c) incorrectly focuses on structured data as the sole reason for data storage solutions changing over time. Option d) incorrectly suggests that analysis of data was a driving factor in the evolution of data storage.

[1.2, LO 1.2.1]
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

Solution: c. A biologist analyzing a large dataset of genetic sequences to gain insights about the genetic basis of diseases
Traditionally, biologists would conduct lab experiments to answer questions in their field; however, nowadays data science is being used to analyze large datasets to extract valuable information that can shed light on complex topics such as the genetic basis of diseases. Option a) is incorrect as studying primary sources does not inherently involve data science. Option b) is incorrect as solving equations is not in the domain of data science. Option d) is incorrect as it describes the traditional work of a chemist as a lab scientist.

Critical Thinking

[1.3, LO 1.3.4]
1. For each dataset, list the attributes.

a. Spotify dataset
b. CancerDoc dataset

Solution a: Following is the list of attributes in the Spotify dataset:
track_name, artist(s)_name, artist_count, released_year, released_month, released_day, in_spotify_playlists, in_spotify_charts, streams, in_apple_playlists, in_apple_charts, in_deezer_playlists, in_deezer_charts, in_shazam_charts, bpm, key, mode, danceability_%, valence_%, energy_%, acousticness_%, instrumentalness_%, liveness_%, speechiness_%

Solution b: The CancerDoc dataset has three attributes; however, none of these attributes have a clear name. They are: the column with numeric identifiers (the first column), the column with cancer type (the second column), and the actual text (the third column).

Principles of Data Science 1st edition Ault Solutions Manual

[1.3, LO 1.3.3]

2. For each dataset, 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
   c. "key" attribute of Spotify dataset
   d. the second column in CancerDoc dataset

Solution a: "artist_count" are integers that indicate the number of times that each track was played. Thus, it is a numeric and discrete type of data since the count can only be integers.

Solution b: "mode" has only two string values—"Major" and "Minor"—so it is categorical. There is no notion of ordering, so the data is nominal.

Solution c: "key" is a string attribute that has a finite set of values (e.g., "A", "F", "F#"), so it is categorical data. It can be either nominal or ordinal depending on how a data scientist wants to treat this data. If they want to consider ordering notion with respect to pitch (e.g., G is higher than C), they can argue it is ordinal. If they simply want to treat the different kinds of pitch without any ordering notion, they can argue it is nominal.

Solution d: The second column of the CancerDoc dataset indicates the type of cancer of each entry in a string. There is a finite number of cancers in the dataset (thyroid, colon, and lung) and these categories have no notion of ordering, so it is a categorical and nominal data type.

[1.3, LO 1.3.2]

3. For each dataset, identify the type of the dataset—structured vs. unstructured. Explain why.
   a. Spotify dataset
   b. CancerDoc dataset

Solution a: The Spotify dataset is a structured dataset since each item in the dataset is in a same form.

Solution b: The CancerDoc dataset is an unstructured dataset since the third column is the main information while the first and second columns serve as labels of each entry (i.e., used to distinguish each item in the dataset). The third column is a free-form text, so this dataset is unstructured.

[1.3, LO 1.3.4]

4. For each dataset, list the first data entry.
   a. Spotify dataset
   b. CancerDoc dataset

Principles of Data Science 1st edition Ault Solutions Manual

Solution a: The first entry shows up on the first row on each CSV file unless the top row is a header row that indicates the names of different attributes in the dataset.
The first row on spotify-2023.csv is a header row. Therefore, the first item is located on the second row of the CSV file. That is:

> "Seven (feat. Latto) (Explicit Ver.)", "Latto, Jung Kook", 2, 2023, 7, 14, 553, 147, 141381703, 43, 263, 45, 10, 826, 125, "B", "Major", 80, 89, 83, 31, 0, 8, 4

Solution b: The first entry shows up on the first row on each CSV file unless the top row is a header row that indicates the names of different attributes in the dataset. The first row on CancerDoc.csv is a header row. Therefore the first item is located on the second row of the CSV file. That is:

> 0, "Thyroid_Cancer", "Thyroid surgery in children…"

Note that the last attribute value is abbreviated with … since it is a very long string.

[1.3, LO 1.3.4]
5. Open the WikiHow dataset (ch1-wikiHow.json) and list the attributes of the dataset.

Solution: The ch1-wikiHow.json file has a list of items in an array (i.e., [ ]). Each array has an object (i.e., { }) in which there are nine attributes total. The attributes are: "Time", "URL", "MainTask", "MainTaskSummary", "Steps", "Categories", "Ingredients", "Requirements", and "Tips".
Note that some attributes have data in the form of an array as well. For example, "Steps" is an array of which each element is also an object with three fields—"Headline", "Description", and "Links".

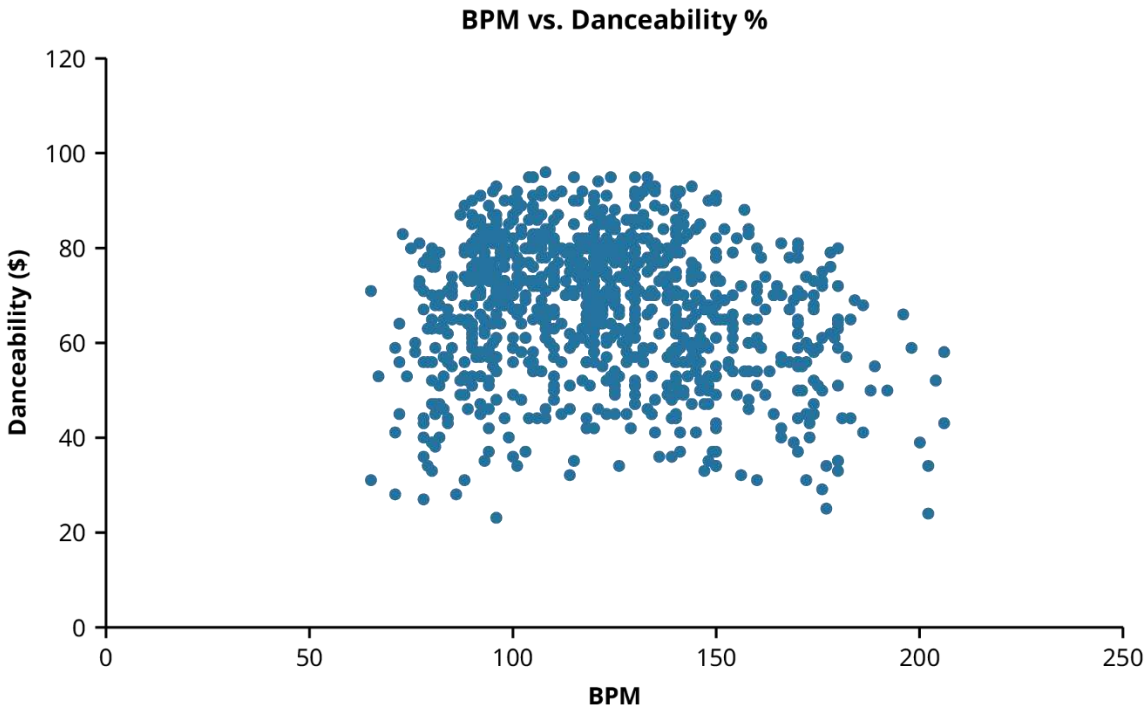
[1.5, LO 1.5.3]
6. Draw scatterplot between bpm (*x*-axis) and danceability (*y*-axis) of the Spotify dataset using:

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

Solution a: The following code draws the same scatterplot. Note the Python-generated plot will not have a title.

```
import matplotlib.pyplot as plt
plt.scatter(data["bpm"], data["danceability_%"]) # draw the
scatterplot
plt.show()
```


Principles of Data Science 1st edition Ault Solutions Manual



Solution b: (This solution is based on MS Excel.) You draw a scatterplot by clicking the scatterplot icon under the Insert tab. Once a default scatterplot shows, click the plot and edit the data range by clicking the Select Data button under the Chart Design tab.

On the Select Data pop-up window, make sure you only show one series (i.e., danceability with respect to bpm) and its *x*-values are BPMs (i.e., $O$2:$O$954) and *y*-values are danceability numbers (i.e., $R$2:$R$954) (see figure below).

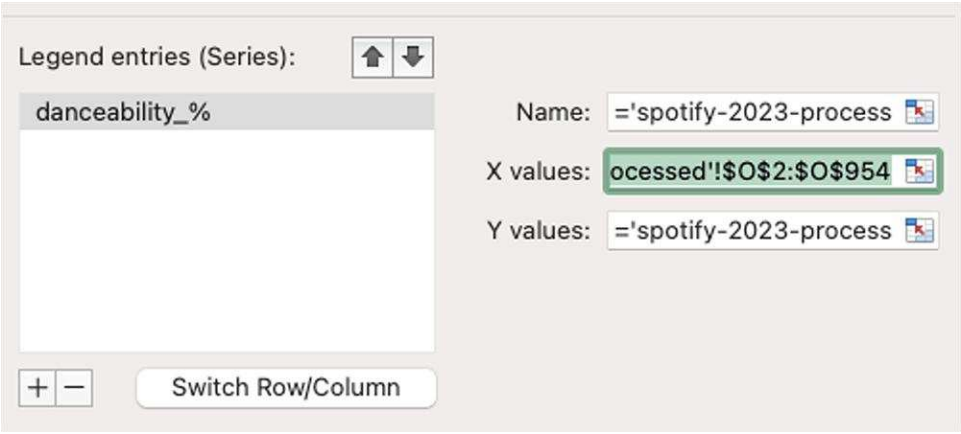


Figure 1.22 Selecting Danceability on Select Data Pop-Up (Used with permission from Microsoft)

The resulting plot appears in the figure below**.**

Principles of Data Science 1st edition Ault Solutions Manual

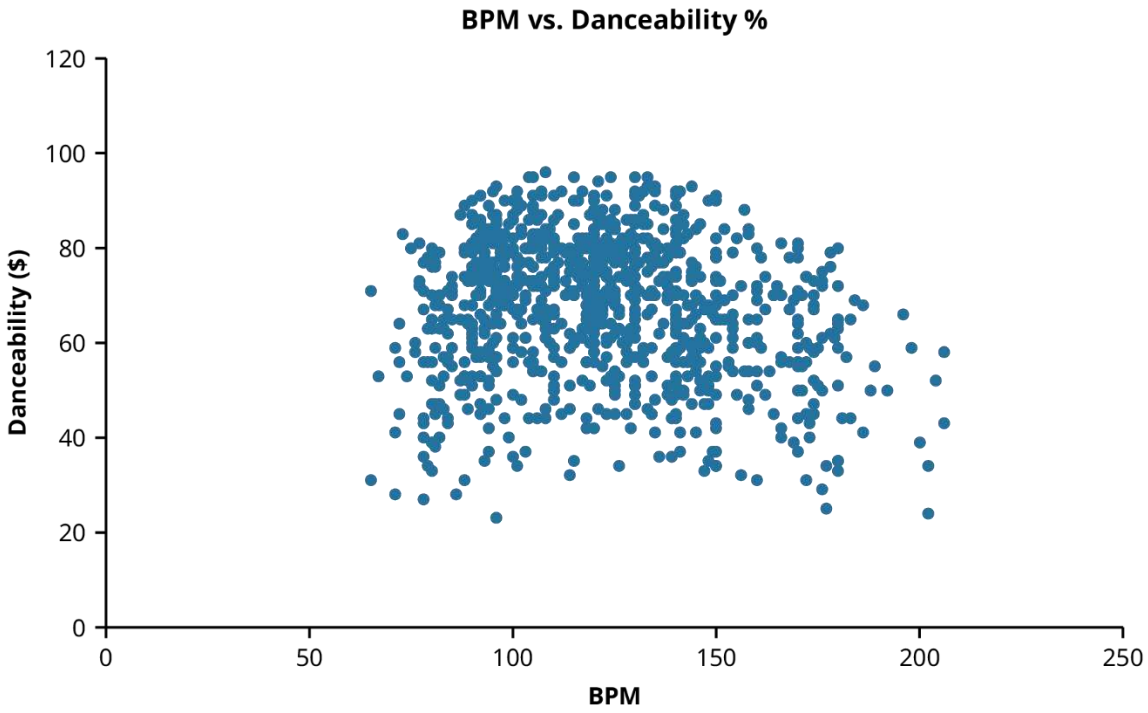


Figure 1.23 Scatterplot of Danceability with Respect to BPM, Drawn with MS Excel

[1.5, LO 1.5.3]

7. Regenerate the scatterplot of the Spotify dataset, 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.)

Solution a: The following code draws the same scatterplot with the custom title and axis labels.

```
import matplotlib.pyplot as plt
plt.scatter(data["bpm"], data["danceability_%"]) # draw the scatterplot
plt.title("BPM vs. Danceability") # set the title

plt.xlabel("BPM") # set the x-axis label
plt.xlim(data["bpm"].min(), data['bpm'].max()) # set the range of the axis
```


Principles of Data Science 1st edition Ault Solutions Manual

```
# set the y-axis label and its range of values
plt.ylabel("Danceability (%)")
plt.ylim(data["danceability_%"].min(), data['danceability_%'].max())

plt.show()
```

Solution b: (This solution is based on MS Excel.) You can edit the chart title by double-clicking the title text. A cursor will show up, and you can edit the title text. The axis labels can be added by clicking Chart Design > Add Chart Element > Axis Titles. Primary Vertical and Primary Horizontal will add a text box for the *x*- and *y*-axes, respectively. You can edit the text boxes by double-clicking them.

To set the range of the values to be related to the minimum and maximum values of the bpm and danceability column, on Excel you need to calculate those values first. You can do so by using =MIN() and =MAX() on each column. Note those values somewhere and use them in the text boxes under Format Axis > Axis Options > Bounds. You can open the Format Axis menu by either 1) double-clicking the axis elements or 2) right-clicking the axis elements and then selecting Format Axis….

The resulting plot appears below**.**

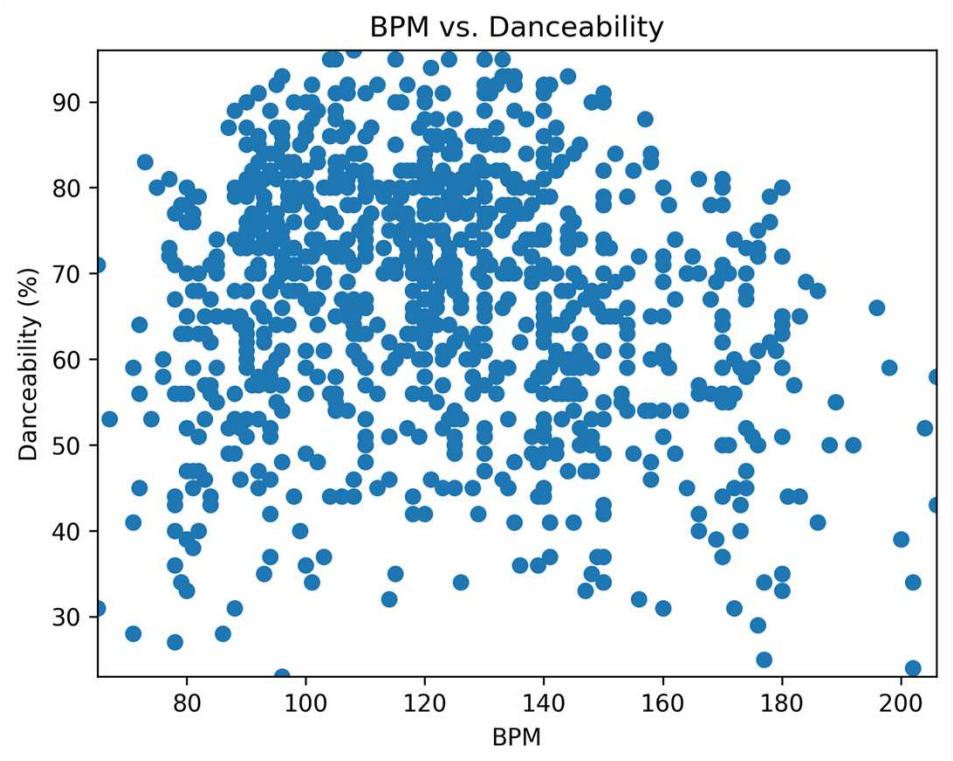


Figure 1.24 Modified Scatterplot with a Custom Title and Axis Labels, Drawn with Python

Principles of Data Science 1st edition Ault Solutions Manual

[1.5, LO 1.5.2]

8. Based on the Spotify dataset, 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

Solution a: `DataFrame.loc[]` is used to filter rows with a specific criteria. The following code filters all the rows whose artist name is Taylor Swift.

```
data.loc[data["artist(s)_name"] == "Taylor Swift"]
```

The result has 34 rows of Taylor Swift tracks.

Solution b: This question uses two filtering criteria, and both need to be met to be filtered in. Thus, connecting both criteria using an AND operator (`&`) will do the job. The following code will show only eight rows.

```
data.loc[(data["artist(s)_name"] == "Taylor Swift") & (data["released_year"]
< 2020)]
```

Quantitative Problems

[1.4] [LO 1.5.1, 1.5.2]

1. Based on the Spotify dataset, 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.)

Solution a: The average (or mean) of "bpm" is 122.54.
To calculate the average using Python `Pandas`, first you need to upload the dataset as a `Pandas` DataFrame. Then `describe()` will compute some basic statistics of the dataset including average. The following code loads the dataset as a `Pandas` DataFrame and computes the average. The average (or mean) of "bpm" is 122.54.

```
data = pd.read_csv("[path to spotify-2023.csv]") # load the
dataset
data.describe() # compute some basic statistics, including
average
```

Solution b: The resulting average is 122.54.
In Excel, the "bpm" attribute is located in column O. The bpm values are listed from row 2 to row 954. Thus, the following Excel formula will calculate the average bpm.

Principles of Data Science 1st edition Ault Solutions Manual

```
=AVERAGE(O2:O954)
```

The resulting average is 122.54.

This file is copyright 2024, Rice University. All Rights Reserved.

Principles of Data Science 1st edition Ault Solutions Manual