Name: _____
Score: _____ / _____

Reading Quiz 8

Part 1

In "Tidy Data", Hadley Wickham (author of R Studio and the tidyverse packages) defines general characteristics of analytic-ready data and sets these as the goals for any data wrangling work.

Which of the following describes a tidy dataset?

A.
The World Bank reports data for all countries from 2000 through 2020 for the following three indicators: the population, percent of the population under 30 years old, and gross domestic product (GDP). There's one row for every country. There's one column for the country name, one for indicator name, and one for each year. The first three values of the column named "2000" report the population, percent of the population under 30 years old, and GDP for Afghanistan, for example.

B.
An epidemiological research lab conducts a study that traces the contacts of people who've been diagnosed with COVID-19. They interview 1000 people and categorize them as asymptomatic, mildly symptomatic, or severely symptomatic. They also categorize each person by the number of contacts they had during the week of the study: fewer than 10, 11-20, 21-30, and greater than 30. The data contain the cross-tabulation of these two categorizations. There are three rows: one for asymptomatic, one for mildly symptomatic, and one for severely symptomatic. There are four columns, labeled: "<11", "[11,20]", "[21,30]", ">30". The cells contain the count of people who have each pair of categorizations.

C.
A financial firm reports on the performance of its stocks over the course of a month. The data contain one row for each stock on each day in the month during which the stock market was open. There is one column for the stock symbol, one column for the day, and one for the closing stock price.

D.
A new student club records the votes of its members during club meetings. During the course of the semester, 50 total votes are taken and club membership increases from 25 to 130. The data contain one row for every club member, and one column for every vote from vote 1 to vote 50. The cells contain 1 if the member voted yes on a particular vote, 0 if the member voted no, and is NA if the member had not yet joined the club when the vote was taken.

D.
A new student club records the votes of its members during club meetings. During the course of the semester, 50 total votes are taken and club membership increases from 25 to 130. The data contain one row for every club member, and one column for every vote from vote 1 to vote 50. The cells contain 1 if the member voted yes on a particular vote, 0 if the member voted no, and is NA if the member had not yet joined the club when the vote was taken.

A real-estate database contains a table with the current values of all properties in Virginia. I collapse the data to have one row per ZIP code, along with the average property value in each ZIP code. I then create a standardized version of average property value by subtracting the mean of this column from every value and dividing by the standard deviation. Finally I put the rows in order from the highest to the lowest value of the standardized average property value.

This workflow used all of the following operations EXCEPT:

A.
transformation

B.
sorting

C.
aggregation

D.
filtering

Which of the following are ways in which Wickham sees the field of data cleaning developing in the future?

A.
Building tools that are easier and more intuitive for users from a cognitive perspective by connecting with the research on human-computer interaction

B.
The development of data cleaning methods and tools for data from multidimensional arrays, such as from images and videos

C.
The creation of tools that automate the process of data cleaning entirely and choose an approach to optimize speed and efficient memory usage

D.
All of the above

What, according to Terrizzano, Schwarz, Roth, and Colino, is the difference between data lake and a curated data lake?

A.
Curated data lakes have been placed into a relational database schema and can be queried using SQL, while data lakes can only be queried using a NoSQL language.

B.
A curated data lake performs data wrangling on the contents of a data lake to abide by licenses, remove duplicate data, transform data to make it ready for analysis, and employ modern archiving principles for long-term storage.

C.
A curated data lake is a subset of datasets within a larger data lake that represent the most useful data sources for business applications.

D.
A curated data lake is a product available for purchase from IBM that provides faster performance on queries relative to other data lakes.

What, according to Terrizzano, Schwarz, Roth, and Colino, is a data governance process?

A.
The steps to add security layers to a data lake to prevent unauthorized users from accessing the data.

B.
The procedure by which a responsible business should consult relevant stakeholders when designing a data lake.

C.
The collection of automated scripts that convert raw data in the data lake to transformed and clean datasets that are ready to be used in analyses.

D.
The steps necessary to bring all the information on licensing, restrictions, and terms of use for all of the third-party data in the data lake together and clearly express them in one document.

Why do Terrizzano, Schwarz, Roth, and Colino argue that careful records must be kept on the ways in which raw data have been edited while generating the analysis-ready versions of the datasets?

A.
These records are part of the data grooming process and illustrate how the data existed in their raw form.

B.
These records are part of the process of provisioning data and provide individual users the steps necessary to run analyses on the data on their personal computers.

C.
These records are part of the process of describing data by showing how key descriptive statistics change as a result of the edits.

D.
These records are part of the process of preserving data and ensure that the data will continue to remain relevant indefinitely.

What is a lambda function, as it is used in the pandas examples that McKinney discusses?

A.
A function that provides a faster way to perform pandas functions by calling the pandas base code directly.

B.
A function that performs elementwise operations by looping over the individual elements of a column.

C.
A function that performs columnwise operations by looping over the individual columns in a dataframe.

D.
A function that allows a user to use a pandas function on specific datapoints in the dataframe, instead of having to apply functions to entire rows or columns.

Which of the following sets of operations describes a split-apply-combine procedure as described by McKinney?

A.
A single dataframe is divided into many individual series - one for each column. A function is used to calculate a summary statistic for each series. Then these statistics are brought together in a single series.

B.
A function is used to calculate a statistic using just the first row of a dataframe, then the next row is included in the function to update the result. The procedure loops across all the rows until all rows are taken into consideration.

C.
A single dataframe is divided into many dataframes based on rows that share the same value of a categorical feature. Every dataframe is collapsed to one row, using a function to summarize specific columns. Then the dataframes are appended back together in a single dataframe.

D.
After taking steps to clean a raw dataset, the cleaned version is compared to the original raw version to catch discrepancies that may indicate possible errors in the data cleaning process.

According to McKinney, which of the following is true about a dataframe column with the category data type as opposed to a dataframe column with the object data type?

A.
object type columns have values that are displayed as strings when viewing dataframes, and category type columns have values that are displayed as numbers when viewing dataframes.

B.
object type columns have categories that do not have a meaningful order, and category type columns have categories that can always be placed in a meaningful order.

C.
category type columns take up less memory than object type columns and operations run faster on category types than on object types.

D.
The category and object data types are equivalent as both can be converted to underlying numeric codes with the .cat.codes attribute.

What, according to McKinney, does .pipe() do?

A.
It provides a way to combine several lines of code into one command by assuming that the output of one function comprises the input of the next function.

B.
It automatically generates documentation of steps that comprise the pipeline from the raw data to the cleaned data.

C.
It allows a user to write their own, original methods for a pandas dataframe.

D.
It enhances data security by requiring a user name and password in order to use functions that manipulate a dataframe.