Data Prep & EDA Benchmark Assessment

Review Results

Assessment PYTHON-1697674-BENCHMARK

Thinkific User ID:	**195642843**
Full Name:	Onyebuchi, Augustine
Email Address:	augustinesopuluonyebuchi@gmail.com
Respondent ID:	126877
Date Started:	10/24/2024 09:09:36 AM
Date Completed:	10/24/2024 09:23:21 AM
Attempt:	1

Question Topic	Num	Question	
		Respondent's Answer	Correct Answer

Intro to Data Science

Earned 0 of 1 points (0%).

1. The first step of a data science project is to _____.

Respondent's Answer	Correct Answer
(X) gather data	( ) gather data
( ) clean the data	( ) clean the data
( ) model the data	( ) model the data
( ) scope the project	(X) scope the project
( ) I don't know yet	( ) I don't know yet

Explanation: Even though data science has data in the name, data science projects don't start with data. Instead, data scientists start by scoping a project and identifying end users, problems, etc. before diving into actual data.

Related Lecture: Data Science Workflow

Scoping a Project

Earned 0 of 1 points (0%).

2. Which of the following is an example of unsupervised learning?

Respondent's Answer	Correct Answer
( ) Predicting house prices over the next year	( ) Predicting house prices over the next year
( ) Flagging which customers are most likely to cancel their membership	( ) Flagging which customers are most likely to cancel their membership
( ) Identifying the main themes mentioned in customer reviews	(X) Identifying the main themes mentioned in customer reviews
(X) Estimating how many customers will visit your website next week	( ) Estimating how many customers will visit your website next week
( ) I don't know yet	( ) I don't know yet

Explanation: Supervised learning is all about using historical data to make future predictions. When identifying the main themes in customer reviews, no predictions are made. Finding themes is an unsupervised learning problem.

Related Lecture: Supervised vs Unsupervised Learning

Gathering Data

Earned 2 of 4 points (50%).

3. A Pandas DataFrame has which characteristics?

(X) Index starting at 0, each column containing a single data type	(X) Index starting at 0, each column containing a single data type
( ) Index starting at 1, each column containing a single data type	( ) Index starting at 1, each column containing a single data type
( ) Index starting at 0, each column containing multiple data types	( ) Index starting at 0, each column containing multiple data types
( ) Index starting at 1, each column containing multiple data types	( ) Index starting at 1, each column containing multiple data types
( ) I don't know yet	( ) I don't know yet

Explanation: Counting in Python starts at 0 and each column in a DataFrame can only contain a single data type.

Related Lecture: The Pandas DataFrame

4. Which file formats can Python read in with one line of code?

( ) .csv files	( ) .csv files
( ) .xlsx files	( ) .xlsx files
( ) .json files	( ) .json files
(X) All of the above	(X) All of the above
( ) I don't know yet	( ) I don't know yet

Explanation: Using pd.read_csv(), pd.read_excel() and pd.read_json(), you can read those file types into Python with just one line of code.

Related Lecture: Reading Flat Files

5. What will happen when the code above is executed?

( ) The "Instructors" tab will be read into Python	( ) The "Instructors" tab will be read into Python
(X) The "Instructors" tab will be read into Python and saved as a DataFrame	( ) The "Instructors" tab will be read into Python and saved as a DataFrame
( ) The "Courses" tab will be read into Python	(X) The "Courses" tab will be read into Python
( ) The "Courses" tab will be read into Python and saved as a DataFrame	( ) The "Courses" tab will be read into Python and saved as a DataFrame
( ) I don't know yet	( ) I don't know yet

Explanation: Because Python is zero-indexed, reading the 0th tab would mean reading the first tab and reading the 1st tab would mean reading the second tab. This code only reads in the data, but does not save it as a DataFrame.

Related Lecture: Reading Excel Files

6. Which method will return the range of values within each column?

( ) .head()	( ) .head()
( ) .count()	( ) .count()
( ) .describe()	(X) .describe()
(X) .info()	( ) .info()
( ) I don't know yet	( ) I don't know yet

Explanation: The describe method will return summary statistics including the min and max of each column.

Related Lecture: Quickly Exploring a DataFrame

Cleaning Data

Earned 1 of 6 points (17%).

7. What will happen when the code above is executed?

(X) df.Income will change from an object to a numeric data type
( ) df.Income will be set equal to numeric
( ) Both 1 and 2
( ) Neither 1 or 2, you would get an error
( ) I don't know yet

( ) df.Income will change from an object to a numeric data type
( ) df.Income will be set equal to numeric
( ) Both 1 and 2
(X) Neither 1 or 2, you would get an error
( ) I don't know yet

Explanation: You would get an error because text with $ values cannot be converted to numeric data types. You would need to use str.replace() to remove the $ values before converting.

Related Lecture: Converting to Numeric

8. Which of the following is NOT a way that missing data is represented in Python?

( ) np.NaN
( ) pd.NaN
(X) pd.NA
( ) None
( ) I don't know yet

( ) np.NaN
(X) pd.NaN
( ) pd.NA
( ) None
( ) I don't know yet

Explanation: Various ways to represent missing values in Python are Numpy's NaN, Pandas' NA and base Python's None.

Related Lecture: Finding Missing Data

9. Which approach allows you to replace the "New York" value with "NY" in the State column?

(X) .loc[]
( ) np.where()
( ) .map()
( ) All of the above
( ) I don't know yet

( ) .loc[]
( ) np.where()
( ) .map()
(X) All of the above
( ) I don't know yet

Explanation: These are all ways that you can replace a value within a column, .loc[] to replace a specific value, np.where to replace a values based on a conditional and .map() to map a set of values to another set of values.

Related Lecture: Handling Inconsistent Text & Typos

10. What is considered an outlier?

( ) Data that is negative
(X) Data that is greater than three times the mean
( ) Data that is more than 3 standard deviations from the mean
( ) All of the above
( ) I don't know yet

( ) Data that is negative
( ) Data that is greater than three times the mean
(X) Data that is more than 3 standard deviations from the mean
( ) All of the above
( ) I don't know yet

Explanation: A rule of thumb in statistics is that outliers are data points that are more than ~3 standard deivations from the mean.

Related Lecture: Finding Outliers

11. How would you extract the day of the week from the Run Date column?

( ) dayofweek(run_times['Run Date'])
(X) dt.dayofweek(run_times['Run Date'])
( ) run_times['Run Date'].dayofweek
( ) run_times['Run Date'].dt.dayofweek
( ) I don't know yet

( ) dayofweek(run_times['Run Date'])
( ) dt.dayofweek(run_times['Run Date'])
( ) run_times['Run Date'].dayofweek
(X) run_times['Run Date'].dt.dayofweek
( ) I don't know yet

Explanation: To extract the day of the week, you would use a datetime method, dt.dayofweek, and chain it on to the Run Date column.

Related Lecture: Creating DateTime Columns

12. If you applied str[:6] to the run_notes data above, which characters in the text would be returned?

(X) Characters 0-5	(X) Characters 0-5
( ) Characters 0-6	( ) Characters 0-6
( ) Characters 1-5	( ) Characters 1-5
( ) Characters 1-6	( ) Characters 1-6
( ) I don't know yet	( ) I don't know yet

Explanation: Extracting a portion of a string value using str returns the first location (if blank, then 0) up to the last location, non-inclusive. So in this case, it would be characters 0 through 5.

Related Lecture: Creating Text Columns

Exploratory Data Analysis

Earned 2 of 4 points (50%).

13. What does df.groupby('col').head() do?

(X) Groups the data by col and returns the first 5 rows of the results	( ) Groups the data by col and returns the first 5 rows of the results
( ) Groups the columns and returns the first 5 rows of the results	( ) Groups the columns and returns the first 5 rows of the results
( ) Returns the first 5 rows within each group	(X) Returns the first 5 rows within each group
( ) Returns an error	( ) Returns an error
( ) I don't know yet	( ) I don't know yet

Explanation: For each column value, this will return the first 5 rows for each value.

Related Lecture: Grouping

14. Why would you put parentheses around multiple chained methods?

(X) To be able to put each method on a separate line	(X) To be able to put each method on a separate line
( ) To allow the chained code to run	( ) To allow the chained code to run
( ) To create a function from the code	( ) To create a function from the code
( ) To comment the code	( ) To comment the code
( ) I don't know yet	( ) I don't know yet

Explanation: When chaining multiple methods together, if you wrap the code in parentheses, you can place each method on a separate line, which makes reading the code easier.

Related Lecture: Grouping

15. What data is typically right skewed (picture above)?

( ) Student grades	( ) Student grades
( ) Household income	(X) Household income
( ) Human lifespan	( ) Human lifespan
( ) Men and women's heights	( ) Men and women's heights
(X) I don't know yet	( ) I don't know yet

Explanation: Right skewed data is data where there are very few large values, and this example of household income is one of them.

Related Lecture: Distributions

16. Which one of these charts shows a correlation of 0?

( ) Chart 1
(X) Chart 2
( ) Chart 3
( ) None of them
( ) I don't know yet

( ) Chart 1
(X) Chart 2
( ) Chart 3
( ) None of them
( ) I don't know yet

Explanation: The middle chart shows a correlation of 0, meaning there is no relationship between Hours Talking to Friends and their Grade on the Test.

Related Lecture: Correlations

Preparing for Modeling

Earned 2 of 4 points (50%).

17. What things need to be changed in the table above to put it into a model to predict house prices?

( ) Remove the Address column
( ) Change City and Color to numeric values
(X) Both 1 and 2
( ) Neither 1 or 2, the data is ready for modeling
( ) I don't know yet

( ) Remove the Address column
( ) Change City and Color to numeric values
(X) Both 1 and 2
( ) Neither 1 or 2, the data is ready for modeling
( ) I don't know yet

Explanation: All data must be non-null and numeric before inputting it into a model. Unique identifiers like address are not useful for modeling and should be removed before modeling.

Related Lecture: Data Prep for EDA vs Modeling

18. Which of the following could you use to vertically stack two DataFrames?

(X) .append()
( ) .join()
( ) .merge()
( ) .concat()
( ) I don't know yet

( ) .append()
( ) .join()
( ) .merge()
(X) .concat()
( ) I don't know yet

Explanation: pd.concat() allows you to both vertically stack and horizontally combine two DataFrames.

Related Lecture: Creating a Single Table

19. What can you do to a right skewed histogram to make it normally distributed?

(X) Apply a log transformation
( ) Apply a supervised learning algorithm
( ) Add a mirror image of the right skewed data
( ) Nothing, you cannot change skewed data
( ) I don't know yet

(X) Apply a log transformation
( ) Apply a supervised learning algorithm
( ) Add a mirror image of the right skewed data
( ) Nothing, you cannot change skewed data
( ) I don't know yet

Explanation: Log transformations turn skewed data into more normally-distributed data.

Related Lecture: Feature Transformations

20. Why can't you input the above DataFrame directly into a model to predict house prices?

( ) The columns are not all the same data type
( ) The columns all need to be on the same scale
( ) One house price seems much lower than the rest
( ) The zip code column is saying that 60202 is better than 60201
(X) I don't know yet

( ) The columns are not all the same data type
( ) The columns all need to be on the same scale
( ) One house price seems much lower than the rest
(X) The zip code column is saying that 60202 is better than 60201
( ) I don't know yet

Explanation: Sometimes data seems ready to be input into a model because it's numeric, but in cases like this, a higher zip code doesn't necessarily mean a higher / lower house price. In situations like this, you can either create a dummy variable or identify a proxy variable.

Related Lecture: Proxy Variables

Time Used: 00:13:45 **Final Score:** 35%