Python Institute

PCED-30-02

PCED – Certified Entry-Level Data Analyst with Python

For More Information – Visit link below:

https://www.examsempire.com/

Product Version

1. **Up to Date products, reliable and verified.**
2. **Questions and Answers in PDF Format.**

Latest Version: 6.0

Question: 1

You are analyzing a dataset of customer purchase history. You notice that several entries have ages listed as 'N/A'. Which of the following actions would be most appropriate for handling these missing age values in the context of building a predictive model for customer churn?

A. Remove the rows with 'N/A' age values entirely.
B. Replace 'N/A' with the average age of all customers in the dataset.
C. Replace 'N/A' with the median age of all customers in the dataset.
D. Create a new category for 'N/A' ages and treat it as a separate value.
E. Impute the missing ages using a machine learning model trained on the available age data.

Answer: E

Explanation:
Imputing missing age values using a machine learning model is the most appropriate approach in this scenario. This approach takes advantage of the existing patterns in the data to estimate the missing values, preserving more information than simply removing the rows or replacing with a single value. Removing rows may lead to data loss, and replacing with average or median may introduce bias. Creating a separate category for 'N/A' may not be informative in this context. Therefore, using a machine learning model for imputation offers the best balance of data integrity and model accuracy.

Question: 2

Consider the following Python code snippet for data cleaning. What is the primary purpose of the operation? "'python import pandas as pd df = pd.DataFrame({'A': [1 , 2, None, 4, 5], 'B': [None, 2, 3, 4, 5]}) inplace=True) print(df)'''

A. Replace missing values with the last non-null value in the column.
B. Replace missing values with the average of the column.
C. Replace missing values with the most frequent value in the column.
D. Replace missing values with the next non-null value in the column (forward fill).
E. Remove rows containing missing values from the DataFrame.

Answer: D

Explanation:
The operation in Pandas performs forward filling. This means that it propagates the last valid observation forward to fill the missing values in the column. In the code snippet provided, the missing value in the first row of column 'B' will be replaced with the value 2 from the second row. Similarly, the missing value in the third row of column 'A' will be replaced with the value 2 from the second row. This

technique is useful when there is a trend or a temporal aspect to the data, and it is reasonable to assume that the missing values would be similar to the last valid observation.

Question: 3

You are analyzing a dataset with a column representing product prices. You identify that some prices are recorded as strings, such as '$10.99' instead of numerical values. Which of the following approaches would be the most suitable for converting these string values into numerical values while maintaining data integrity?

A. Use the method to remove the dollar sign and then convert the string to a float.
B. Use the 'str.strip()' method to remove any leading or trailing spaces and then convert the string to a float.
C. Use the ‘str.split()' method to split the string at the decimal point, convert each part to a float, and then combine them into a single float.
D. Use the function with the ‘errors='coerce'S parameter to attempt conversion and set invalid values to NaN.
E. Use e ‘str.replace()' method to replace the dollar sign with an empty string and then convert the string to a float, handling any errors with a ‘try-except block.

Answer: A,E

Explanation:
Both option A and option E are suitable approaches for converting string values into numerical values while preserving data integrity. Option A utilizes the ‘str.replace()' method to directly remove the dollar sign and then converts the string to a float. This approach is straightforward and efficient. Option E also uses the ‘str.replace()' method to remove the dollar sign and then converts the string to a float, but it includes a 'try- except' block to handle any potential errors during the conversion process. This approach is more robust and ensures that any invalid or unexpected values are handled gracefully. Option B is not the most suitable approach since it only removes spaces, not the dollar sign. Option C is complex and may not be necessary for this specific scenario. Option D will convert valid values to numeric but will leave invalid values as NaN, which may not be desirable in this case.

Question: 4

You are analyzing a dataset of customer transactions for an e-commerce company. You discover several orders with exceptionally high purchase amounts, far exceeding the typical average. These orders are likely outliers. Which of the following data collection methods would be most helpful in determining the validity of these outliers and their potential impact on your analysis?

A. Random sampling of customer transactions to assess overall distribution.
B. Collecting additional data points about the outlier orders, such as customer demographics, order details, and payment information.
C. Using a statistical test to determine if the outliers are statistically significant.
D. Removing the outliers from the dataset and re-running your analysis to see if the results change.
E. Contacting the customers who placed the outlier orders to confirm the legitimacy of the transactions.

Answer: B,E

Explanation:
Collecting additional data about the outlier orders is crucial to understand their context and validity. This might involve customer demographics, order details, payment information, and even reaching out to the customers to verify the transactions. Option A is less helpful as random sampling doesn't focus on the outliers. While statistical tests (Option C) can identify outliers, they don't necessarily explain their origin. Removing outliers (Option D) should be a last resort, as it can potentially bias your analysis.

Question: 5

Consider the following Python code snippet used for data cleaning. Identify the code's purpose and the potential problem it might encounter when dealing with real-world data.

```
import pandas as pd

data = {'age': [25, 30, 'unknown', 35, 40]}
df = pd.DataFrame(data)
df['age'] = df['age'].fillna(df['age'].mean())
print(df)
```

A. The code replaces missing values in the 'age' column with the average age. It might encounter issues if the 'unknown' value cannot be converted to a numerical type.
B. The code aims to impute missing values using the mean. It might encounter issues if the 'age' column contains outliers, skewing the mean and leading to inaccurate imputations.
C. The code replaces missing values with the mean, which can lead to inaccurate imputation if the data is not normally distributed.
D. The code calculates the mean age and replaces missing values with it. This approach is generally robust and handles outliers well.
E. The code replaces missing values with the mean. It might encounter issues if the 'age' column has a high number of missing values, making the mean unreliable.

Answer: A,B,C,E

Explanation:
This code attempts to replace missing values in the 'age' column with the mean. However, there are several potential problems: 1. Non-numerical value: The code assumes all values in the 'age' column can be converted to a numeric type, which isn't true for 'unknown'. This would raise an error. 2. Outliers : If the data contains outliers, the mean will be skewed, leading to inaccurate imputations. This can be especially problematic for variables like age, where outliers can significantly impact the average. 3. Data Distribution : Replacing missing values with the mean is generally best suited for normally distributed data. If the distribution is skewed, the mean might not be a representative value. 4. High Percentage of Missing Values : If the 'age' column has a significant number of missing values, the mean might not be a

reliable representation of the typical value. In such cases, using more sophisticated imputation methods like k-nearest neighbors or regression might be better.

Question: 6

You are analyzing a dataset of customer feedback ratings for a product. The ratings range from 1 to 5, with 5 being the highest. You notice a significant number of ratings clustered around 4 and 5, with very few ratings below 3. This indicates a potential bias in the data'. Which of the following techniques could be applied to address this bias and get a more accurate representation of customer sentiment?

A. Remove the outlier ratings below 3, as they likely represent negative reviews and can skew the analysis.
B. Normalize the ratings using a standard scaling method, such as Min-Max scaling, to adjust the range and distribution.
C. Collect additional data points, such as customer reviews and comments, to gain a deeper understanding of the reasons behind the high ratings.
D. Apply a transformation technique, such as logarithmic transformation, to compress the range of high ratings and emphasize the lower ratings.
E. Use a robust statistical method, such as the median instead of the mean, to calculate the average rating and minimize the impact of outliers.

Answer: C,D,E

Explanation:
The scenario describes a potential bias towards positive feedback. Here's why the chosen options are the most suitable: - C: Collect additional data : Gathering customer reviews and comments can provide context for the high ratings. This helps you understand if the ratings reflect genuine satisfaction or other factors like marketing bias or incentives. - D: Logarithmic Transformation : A logarithmic transformation can compress the range of high ratings and emphasize the lower ratings. This can help to create a more balanced distribution and provide a better representation of the overall sentiment. - E: Robust Statistics : Using the median instead of the mean can help minimize the impact of outliers (the high ratings) and provide a more accurate representation of the typical customer sentiment. Option A is incorrect because removing outliers can lead to data loss and distort the analysis. Option B, normalization, would rescale the data but wouldn't address the underlying bias.

Question: 7

You are analyzing a dataset of customer purchase history. You notice a few entries with ages over 150 years old. What is the most likely cause of this data error and how should it be addressed?

A. Data entry error, replace with the median age
B. System malfunction, remove the entries as they are unusable
C. Data inconsistency, flag the outliers and investigate further
D. Data corruption, replace the values with a random number within the acceptable range
E. Data transformation error, apply a log transformation to normalize the data

Answer: C

Explanation:
The most likely cause of such extreme outliers is data inconsistency. It's crucial to investigate further to understand the reason for the error. Simply removing or replacing the data without investigation might lead to losing valuable information. Flagging the outliers and performing a deeper analysis is the best approach in this scenario.

Question: 8

You have a dataset with several columns containing categorical variables like 'Gender', 'Marital Status', and 'Education'. Which Python library is most suitable for encoding these categorical variables into numerical ones for use in machine learning algorithms, and why?

A. NumPy, as it provides efficient array operations for handling numerical data.
B. Pandas, as it offers powerful data manipulation capabilities including dummy variable creation.
C. Scikit-learn, as it contains a variety of machine learning algorithms for data analysis.
D. Matplotlib, as it provides functions for visualizing data, helping to identify categorical patterns.
E. Seaborn, as it simplifies data visualization with attractive statistical plots.

Answer: B

Explanation:
Pandas is the most suitable library for encoding categorical variables. It offers functions like 'get_dummies' which efficiently create dummy variables (one-hot encoding) for categorical features. This transformation is crucial for most machine learning algorithms that require numerical input.

Question: 9

You are working with a dataset that contains missing values represented as 'NaN'. You want to fill these missing values with the mean of the respective column. Which Python code snippet will achieve this effectively? (Select all that apply)

A.
```
import pandas as pd
data = pd.read_csv('data.csv')
data.fillna(data.mean(), inplace=True)
```
B.
```
import numpy as np
data = np.loadtxt('data.csv', delimiter=',')
data = np.nan_to_num(data, nan=np.mean(data, axis=0))
```
C.

```
import pandas as pd
data = pd.read_csv('data.csv')
for col in data.columns:
    data[col].fillna(data[col].mean(), inplace=True)
```

D.

```
import numpy as np
data = np.loadtxt('data.csv', delimiter=',')
for i in range(data.shape[1]):
    data[:, i] = np.nan_to_num(data[:, i], nan=np.mean(data[:, i]))
```

E.

```
import pandas as pd
data = pd.read_csv('data.csv')
data = data.apply(lambda x: x.fillna(x.mean()))
```

Answer: A,C,E

Explanation:
All options A, C, and E will correctly fill the missing values with the column mean using different approaches: A: Using the 'fillna' method with the 'mean()' function directly on the DataFrame. C: Iterating through each column and filling missing values with the column mean. E: Applying a lambda function to the DataFrame, filling missing values with the mean of each column. Option B and D use NumPy, which is not the most suitable for handling missing values in DataFrames. Pandas provides more specialized methods for data manipulation.

Question: 10

You are analyzing customer data for an e-commerce company. The dataset includes 'Age' (numerical) and 'Preferred_Category' (categorical). You discover an outlier in the 'Age' column: a customer with an age of 200 years. Which of the following approaches would be most appropriate for handling this outlier in this specific context?

A. Remove the outlier record as it is highly likely an error.
B. Replace the outlier with the mean age of all customers.
C. Replace the outlier with the median age of all customers.
D. Cap the age value at a reasonable maximum, like 120 years.
E. Investigate further to determine if the outlier is valid. If it's not, remove it.

Answer: E

Explanation:
In this scenario, it's crucial to investigate the outlier before taking any action. A customer with an age of 200 years is highly unlikely, suggesting a potential data entry error. Option E is the most suitable approach because it encourages a deeper examination of the data point. Option A, while tempting,

might lead to the loss of potentially valuable data if the outlier is valid. Replacing with mean or median (Options B and C) might distort the distribution of age. Option D could introduce bias by artificially capping the data.

Thank You for Trying Our Product

Special 16 USD Discount Coupon: NSZUBG3X

Email: support@examsempire.com

Check our Customer Testimonials and ratings available on every product page.

Visit our website.

https://examsempire.com/