Jinglun Jiang STA130 HW1

September 7, 2024

[ ]:
```
# Step 1: Import necessary libraries
import pandas as pd

# Step 2: Load the dataset
url = 'https://raw.githubusercontent.com/centralpark/nyc-squirrels/master/
 ↪squirrel_data.csv'
squirrels_df = pd.read_csv(url)

# Step 3: Display the first few rows of the dataset
squirrels_df.head()

# Step 4: Count missing values for each column
missing_data = squirrels_df.isnull().sum()

# Display the result
print("Missing data count per column:\n", missing_data)

# Step 5: Total missing values in the entire dataset
total_missing = missing_data.sum()
print(f"\nTotal missing values in the dataset: {total_missing}")
# ChatGPT Record at:https://chatgpt.com/share/
 ↪f16166be-d085-49d4-b962-240cf3abc92d
```

[ ]:
```
#Observation is an observed individual from the testing groups whose all␣
 ↪characteristics are recorded, like a single squirrel in the central park␣
 ↪being observed, while variable is the shared characteristic that may differ␣
 ↪in different observations, such as the color of the squirrel under␣
 ↪observation.
import pandas as pd

# Load your dataset (assuming CSV file)
df = pd.read_csv('nyc_squirrels.csv')

# Summary of numerical columns
print(df.describe())#describing the dataset in columns
```

```
#There will be no difference in the reported value of ciunt and the number of␣
 ↪columns analyzed if there is non-numeric values in the dataset, as the␣
 ↪method df,shape analyzes the columns without considering their values, while␣
 ↪the same things will decrease if there are missing data in the numeric␣
 ↪variables, as df.describe focus on the form of data.
#An attribute is a stored value of a certain object or target that could not be␣
 ↪called by other instances, while a method is a pre-finished series of events␣
 ↪or orders that could be called amd executed by other instances.
#count is the number of non-missing variables in each column,mean is the␣
 ↪average of the values, std is the standard deviation, or the values that the␣
 ↪variable varies from the mean that could show the variability of the␣
 ↪dataset, 25%,50% and 75% refer to the first, second and third quartiles of␣
 ↪the dataset that shows its 25% th, median and 75%th data. Max is the maximum␣
 ↪value of the dataset.
#When there are significantly more missing value in rows than in columns, using␣
 ↪df.dropna()is better because it removes all incomplete rows while remain the␣
 ↪data structure, which del df['col'] could not.
#On the other hand, when more missing variables distribute as columns, using␣
 ↪del df['col'] works better.
#Applying del df['col'] first could eliminate the missing variable columns,␣
 ↪which could avoid unnecessary row elimination and narrow down the dataset to␣
 ↪increase efficiency.
#Link to chatGPT conversation:https://chatgpt.com/share/
 ↪b77bc4fb-a3dc-4117-9065-8473a746261d
```

[ ]:

[ ]:
```
# Display the number of missing values in each column before cleaning
print("Before cleaning:")
print(df.isnull().sum())
# Drop rows with missing values
df_cleaned = df.dropna()
# Display the number of missing values in each column after cleaning
print("After cleaning:")
print(df_cleaned.isnull().sum())
#Before cleaning:
#X_col1    5
#X_col2    0
#X_col3    3
#dtype: int64
#After cleaning:
#X_col1    0
#X_col2    0
#X_col3    0
#dtype: int64
```

```
#conversation records at: https://chatgpt.com/share/
 ↪6b8ff56f-16ce-4dc2-a8e3-0ee94aed87a3
```

[ ]:
```
#They first create a dataset with respect to unique value col1, then repeat the␣
 ↪process with value col2, and describe the data with the methods mentions in␣
 ↪previous questions.
#The data is first sorted by the value "class" shown as col1, then age shown as␣
 ↪col2, and finally described using the mean and standard deviation, the␣
 ↪quartiles and the maximum.
#Because the missing data still occupies a position and disrupts the overall␣
 ↪structure of the data, which dilutes the final result by increasing the data␣
 ↪size.
#It's easier to solve the problems with GOT as it filters most of the useless␣
 ↪information that has nothing to do with the issue shown on search engines␣
 ↪before it shows the solutions to me.
#Yes
```