powered by
bytewise.

www.bytewiseltd.com

ML/DL Track - Task 2

Created	@June 17, 2024 5:46 PM
Reviewed	☐

Data Analysis Report on NYC Schools' SAT Scores

1. Basic Information about the Dataset

To begin, we imported the dataset and performed basic exploratory data analysis (EDA) to understand the structure and contents of the data. Here are the details:

```
import pandas as pd

df = pd.read_csv('schools.csv')

# Basic information about the dataset
print(df.shape)       # Output: (375, 7)
print(df.columns)     # Output: Index(['school_name', 'boroug
h', 'building_code', 'average_math', 'average_reading', 'aver
age_writing', 'percent_tested'], dtype='object')
print(df.dtypes)      # Output: object (3), int64 (3), float6
4 (1)
print(df.describe())  # Summary statistics
print(df.info())      # More detailed summary including non-n
```

```
ull counts
print(df.head(5))     # First 5 rows of the dataframe
```

2. Verification of SAT Scores

To ensure that the average math scores are within a valid range, we checked the minimum and maximum values:

```
# Verifying the score range
print(df['average_math'].min())  # Output: minimum average ma
th score
print(df['average_math'].max())  # Output: maximum average ma
th score
```

The checks confirmed that the scores are under 800 as expected.

3. Filtering High-Performing Schools in Math

Next, we identified schools with average math scores above 80% of 800 (i.e., 640). We sorted these schools in descending order to highlight the top performers:

```
# Filter and sort schools with high math scores
best_math_schools = df[df['average_math'] > 0.8 * 800]
best_math_schools = best_math_schools[['school_name', 'averag
e_math']].sort_values('average_math', ascending=False)
print(best_math_schools.head(5))
```

4. Handling Missing Data

We noticed that there were 20 null values in the `percent_tested` column. These missing values were filled with the mean of the column:

```
# Filling null values
df['percent_tested'] = df['percent_tested'].fillna(df['percen
t_tested'].mean())
print(df.info())  # Verify that null values are handled
```

5. Top 10 Performing Schools Based on Combined SAT Scores

We calculated the total SAT scores by summing the average math, reading, and writing scores. Then, we identified the top 10 schools with the highest combined SAT scores:

```
# Calculate total SAT score and identify top 10 schools
df['total_SAT'] = df['average_math'] + df['average_reading']
+ df['average_writing']
top_10_schools = df[['school_name', 'total_SAT']].sort_values
('total_SAT', ascending=False).head(10)
print(top_10_schools)
```

6. Borough with the Largest Standard Deviation in Combined SAT Scores

To determine which borough had the most variability in SAT scores, we calculated the standard deviation of the total SAT scores for each borough and identified the one with the highest value:

```
# Calculate statistics by borough
boroughs = df.groupby("borough")["total_SAT"].agg(["count",
"mean", "std"]).round(2)

# Identify the borough with the largest standard deviation
largest_std_dev = boroughs[boroughs["std"] == boroughs["st
```

```
d"].max()]
print(largest_std_dev)
```

Summary of Findings

1. **Dataset Overview**: The dataset contains information on 375 schools with columns related to average SAT scores in math, reading, and writing, as well as the percentage of students tested.
2. **SAT Score Verification**: The average math scores are within the expected range (0 to 800).
3. **High-Performing Math Schools**: The top schools with average math scores above 640 were identified and sorted.
4. **Missing Data Handling**: The missing values in the `percent_tested` column were successfully filled with the mean of the column.
5. **Top 10 Schools by Total SAT Score**: The top 10 schools with the highest combined SAT scores were identified and listed.
6. **Borough with Largest SAT Score Variability**: Manhattan was found to have the largest standard deviation in combined SAT scores, indicating the greatest variability.

Results

- **Top 10 Schools by Total SAT Score**:

```
       school_name                                  total
_SAT
0  Stuyvesant High School                             21
87
1  Bronx High School of Science                       20
88
2  Staten Island Technical High School                20
70
```

```
3  High School of American Studies at Lehman College    20
31
4  High School for Dual Language and Asian Studies      19
97
5  Townsend Harris High School                          19
92
6  Brooklyn Technical High School                       19
76
7  Queens High School for the Sciences at York College  19
70
8  Bard High School Early College                       19
39
9  The Bronx High School for Medical Science            19
27
```

- **Borough with Largest Standard Deviation in SAT Scores**:

```
    borough  count    mean     std
0  Manhattan     89  1340.13  230.29
```

This report provides a comprehensive overview of the analysis performed on the NYC schools' SAT scores dataset, highlighting key insights and findings.