

Data Cleaning and Filtering

22/03/2025



stratified_history.csv
8.1 MB

Analyze the provided dataset to identify its key characteristics for data-driven decision-making. Upon receiving a specific analytical question, address it strictly using the dataset. Your deliverables must include a clear, concise answer to the question and the Python code that extracts and processes the data used to derive your answer. The dataset should be cleaned by filtering admission_period to include only values between 2006.1 and 2019.2, removing entries where enrollment_type is labeled as "Exempted", and excluding those where situation is "In Progress". The resulting cleaned dataset should be stored in a new DataFrame named df_filtered.



qwen-max-latest

To address your request, I will perform the following steps:

1. **Load the Dataset** : Read the CSV file into a pandas DataFrame.
2. **Filter the Data** : Apply the filtering criteria:
 - admission_period between 2006.1 and 2019.2 (inclusive).
 - Exclude rows where enrollment_type is "Exempted".
 - Exclude rows where situation is "In Progress".
3. **Store the Cleaned Data** : Save the filtered data into a new DataFrame named df_filtered .
4. **Provide Python Code** : Include the code used for these operations.

Here is the Python code to achieve this:

```
python
1 import pandas as pd
2
3 # Step 1: Load the dataset
4 # Assuming the file is in the current working directory
5 file_path = 'stratified_history.csv'
6 df = pd.read_csv(file_path)
7
8 # Step 2: Filter the data based on the specified criteria
9 # Convert admission_period to numeric for proper comparison
10 df['admission_period'] = pd.to_numeric(df['admission_period'], errors='coerce')
11
12 # Apply filters
13 df_filtered = df[
14     (df['admission_period'] >= 2006.1) &
15     (df['admission_period'] <= 2019.2) &
16     (df['enrollment_type'] != 'Exempted') &
17     (df['situation'] != 'In Progress')
18 ]
19
20 # Display the first few rows of the cleaned dataset
21 print(df_filtered.head())
22
23 # Optionally, save the cleaned dataset to a new CSV file for further use
24 df_filtered.to_csv('filtered_dataset.csv', index=False)
```

Key Characteristics of the Cleaned Dataset (df_filtered):

After applying the filters:

- The dataset includes only records with admission_period values between 2006.1 and 2019.2.
- Records with enrollment_type labeled as "Exempted" are excluded.
- Records where situation is "In Progress" are removed.

These steps ensure that the resulting dataset is clean and ready for further analysis or answering specific analytical questions.

Next Steps:

If you have a specific analytical question, please provide it, and I will address it using the cleaned dataset (df_filtered).

Is there a statistically significant difference in grades between students enrolled in Normal



qwen-max-latest

To determine whether there is a statistically significant difference in grades between students enrolled in "Normal" and "Extracurricular" modes, we will:

1. **Filter the Cleaned Dataset** : Use the `df_filtered` DataFrame (already cleaned as per earlier instructions).
2. **Extract Relevant Data** : Focus on the columns `grade` (grades) and `enrollment_mode` (enrollment mode: Normal or Extracurricular).
3. **Perform Statistical Test** : Use an appropriate statistical test to compare the grades of the two groups. Since we are comparing two independent groups, a two-sample t-test is suitable if the data meets the assumptions of normality and equal variances. Otherwise, a non-parametric test like the Mann-Whitney U test can be used.
4. **Interpret Results** : Check the p-value to determine statistical significance (commonly at a threshold of 0.05).

Here is the Python code to perform this analysis:

```
python
1 import pandas as pd
2 from scipy.stats import ttest_ind, mannwhitneyu, shapiro, levene
3
4 # Load the cleaned dataset (if not already loaded)
5 file_path = 'filtered_dataset.csv'
6 df_filtered = pd.read_csv(file_path)
7
8 # Extract grades for Normal and Extracurricular enrollment modes
9 normal_grades = df_filtered[df_filtered['enrollment_mode'] == 'Normal']['grade'].dropna()
10 extracurricular_grades = df_filtered[df_filtered['enrollment_mode'] == 'Extracurricular']['grade'].dropna()
11
12 # Check assumptions for t-test: normality and equal variances
13 def check_assumptions(group1, group2):
14     # Shapiro-Wilk test for normality
15     normality_group1 = shapiro(group1).pvalue
16     normality_group2 = shapiro(group2).pvalue
17     # Levene's test for equal variances
18     equal_variances = levene(group1, group2).pvalue
19     return normality_group1, normality_group2, equal_variances
20
21 normality_normal, normality_extracurricular, equal_variances = check_assumptions(normal_grades,
22                                         extracurricular_grades)
23
24 # Perform the appropriate test
25 if normality_normal > 0.05 and normality_extracurricular > 0.05 and equal_variances > 0.05:
26     # Both groups are normally distributed with equal variances -> use t-test
27     t_stat, p_value = ttest_ind(normal_grades, extracurricular_grades, equal_var=True)
28     test_used = "Two-sample t-test"
29 else:
30     # Non-parametric test (Mann-Whitney U test)
31     t_stat, p_value = mannwhitneyu(normal_grades, extracurricular_grades, alternative='two-sided')
32     test_used = "Mann-Whitney U test"
33
34 # Output the results
35 print(f"Test Used: {test_used}")
36 print(f"T-statistic: {t_stat}")
37 print(f"P-value: {p_value}")
38
39 if p_value < 0.05:
40     print("There is a statistically significant difference in grades between the two enrollment modes.")
41 else:
42     print("There is no statistically significant difference in grades between the two enrollment modes.")
```

Explanation of Steps:

1. **Filtering Data** :
 - We extract grades (`grade`) for students in "Normal" and "Extracurricular" enrollment modes using the `enrollment_mode` column.
 - Missing values (`NaN`) are dropped to ensure valid comparisons.
2. **Assumption Checks** :
 - The Shapiro-Wilk test checks for normality of grade distributions in both groups.
 - Levene's test checks for equality of variances between the two groups.
3. **Statistical Test** :
 - If assumptions for a t-test are met (normality and equal variances), we use `ttest_ind`.
 - Otherwise, we use the non-parametric Mann-Whitney U test (`mannwhitneyu`).
4. **Interpretation** :
 - A p-value less than 0.05 indicates a statistically significant difference between the two groups.

Example Output:

If the analysis yields:

the analysis yields:

```
1 Test Used: Two-sample t-test
2 T-statistic: 2.34
3 P-value: 0.021
4 There is a statistically significant difference in grades between the two enrollment modes.
```

This means that students in "Normal" and "Extracurricular" modes have significantly different grades.

If the p-value is greater than 0.05, it implies no significant difference.

Final Answer:

Run the above code on your dataset to determine whether there is a statistically significant difference in grades between "Normal" and "Extracurricular" enrollment modes. The answer will depend on the computed p-value.

