Grand Project on New York School, SAT Test Investigation

Data Wrangling Project

Note: Plagiarism, or copying from each other, is strictly prohibited. Any instances of plagiarism will result in both parties receiving a grade of zero. Please refrain from engaging in this unethical behavior to ensure a fair evaluation.

***Instruction*: Ensure that your work is properly labeled with the corresponding question numbers. Any extraneous code will be omitted. Submission should be in PDF format; other file types such as Word or Python will not be accepted.**

Annually, American high school students undergo the standardized Scholastic Assessment Test (SAT), designed to assess their proficiency in literacy, numeracy, and writing skills. Comprising three sections — reading, math, and writing — each segment carries a maximum score of 800 points. These assessments bear significant importance for both students and colleges, serving as a crucial factor in the college admissions process.

The analysis of school performance holds substantial relevance for various stakeholders, including education professionals, policymakers, researchers, government entities, and parents making informed decisions about their children's educational institutions.

As part of this analytical endeavor, you have been provided with a set of datasets named as below

1. sat_results.csv
2. demographics.csv
3. survey_all.txt
4. class_size.csv
5. survey_d75.txt
6. hs_directory.csv
7. graduation.csv
8. ap_2010.csv

Your task is to provide the Python code against each question with a description

1. Read all CSV files in different data frames using pandas

2. Read text files using pandas with specific parameters such as delimiter= "\t", encoding = "latin-1", and store in all_survey and d75_survey.

3. Concatenate both survey data frames in new survey data fame using axis = 0.

4. update the name of the column from dbn to DBN in the survey data frame

5. update the name of the column from dbn to DBN in the survey hs_directory dataset

6. consider class_size dataset, and add a new column named padded_CSD. Currently, CSD has unique 1 to 32 values. The single digit number should be padded with 0 like 1 become 01. All padded and non-padded data will be stored in new columns named padded_CSD.

7. concatenate the padded_CSD and SCHOOL_CODE and store it in a new column named DBN.

8. consider the dataset name sat_results, and convert the “SAT Critical Reading Avg. Score", "SAT Math Avg. Score", and "SAT Writing Avg. Score” from the sting to numeric.

9. Create a new column named "sat_score" in the sat_results dataset, storing the sum of the scores from the three aforementioned columns. “SAT Critical Reading Avg. Score", "SAT Math Avg. Score", "SAT Writing Avg. Score”

10. Extract longitude and latitude values from the "Location 1" column in the hs_directory dataset and store them in new columns named "lat" and "lon" using regular expressions

11. Select data from the class_size dataset where GRADE is '09-12' and PROGRAM TYPE is 'GEN ED', and overwrite the original dataset.

12. Filter the demographics dataset to include only data where schoolyear is 20112012 and graduation is 'graduation', and overwrite the original data frame.

13. Extract data from the graduation dataset where Demographic is 'Total Cohort' and Cohort is '2006', and overwrite the original data frame.

14. Check for missing values in each dataframe and handle them appropriately, providing a description of the approach taken.

15. Merge the sat_results and ap_2010 datasets using a left join on the 'DBN' column and store the result in a combined data frame.

16. Merge the combined dataframe with the graduation dataframe using a left join on the 'DBN' column and store the result in the combined data frame.

17. Continue merging the remaining dataframe with the combined dataframe using inner joins, one by one, on the 'DBN' column and storing the result in the combined dataframe..

18. Check for null values in the combined dataframe and fill them accordingly, providing a description of the approach taken.

19. Identify and handle outliers in the columns 'SAT Critical Reading Avg. Score,' 'SAT Math Avg. Score,' and 'SAT Writing Avg. Score' by replacing them with NaN values if necessary.

20. Create a scatter plot of total_enrollment vs sat_score and analyze the observed pattern

21. Determine the school with the highest overall result and provide reasoning for its achievement

22. Identify the school with the minimum enrollment and explain the factors contributing to this situation