Dataset Summary

A csv formatted dataset was downloaded from Kaggle (<https://www.kaggle.com/datasets/fundal/sat-by-year-and-gender-1967-2001?select=SAT_by_Year_Gender_1967_2001.csv>) and converted into a 35 x 10 data frame in a Jupyter notebook. As can be seen from the screenshot of the data frame below, there are a number of different comparisons made between average male (M\_verbal, M\_math) and average female (F\_verbal, F\_math) SAT scores as a function of the testing year. The average verbal and math scores for all students, A\_verbal and A\_math respectively, are also listed as a function of test date. The average of the verbal and math scores are listed in the last three columns for males, females and all students (M\_average, F\_average, A\_average respectively).

A screenshot of a computer

Description automatically generated

The statistics for the data set (shown below) were obtained using SAT\_df.describe(), and SAT\_df.info() indicated that all values in the data set were int64 data types.A screenshot of a computer

Description automatically generated

Initial plan for data exploration

The first step, was to determine if there were any historical factors that could have impacted the SAT scores collected over time. It was found that in 1983, the National Commission on Excellence in Education released its report called, “A Nation at Risk.” The impetus for preparing this report was a significant drop in both math and verbal SAT scores had dropped by 40 – 50 points during the period 1963 to 1980. (Wikipedia, “A Nation at Risk”) This report initiated numerous reform efforts aimed improving the educational outcome in the United States.

In looking at the statistics for the data set, female students have lower average verbal and math scores than males. When the mean values for all students are compared, math scores are lower than verbal scores. These observations combined with the historical context given above suggest three areas of investigation for this data set:

1. Is there a significant difference between male and female verbal and math scores?
2. Is there a relationship between the average math scores and the math scores sorted by gender?
3. How have average SAT scores for all students changed over time?

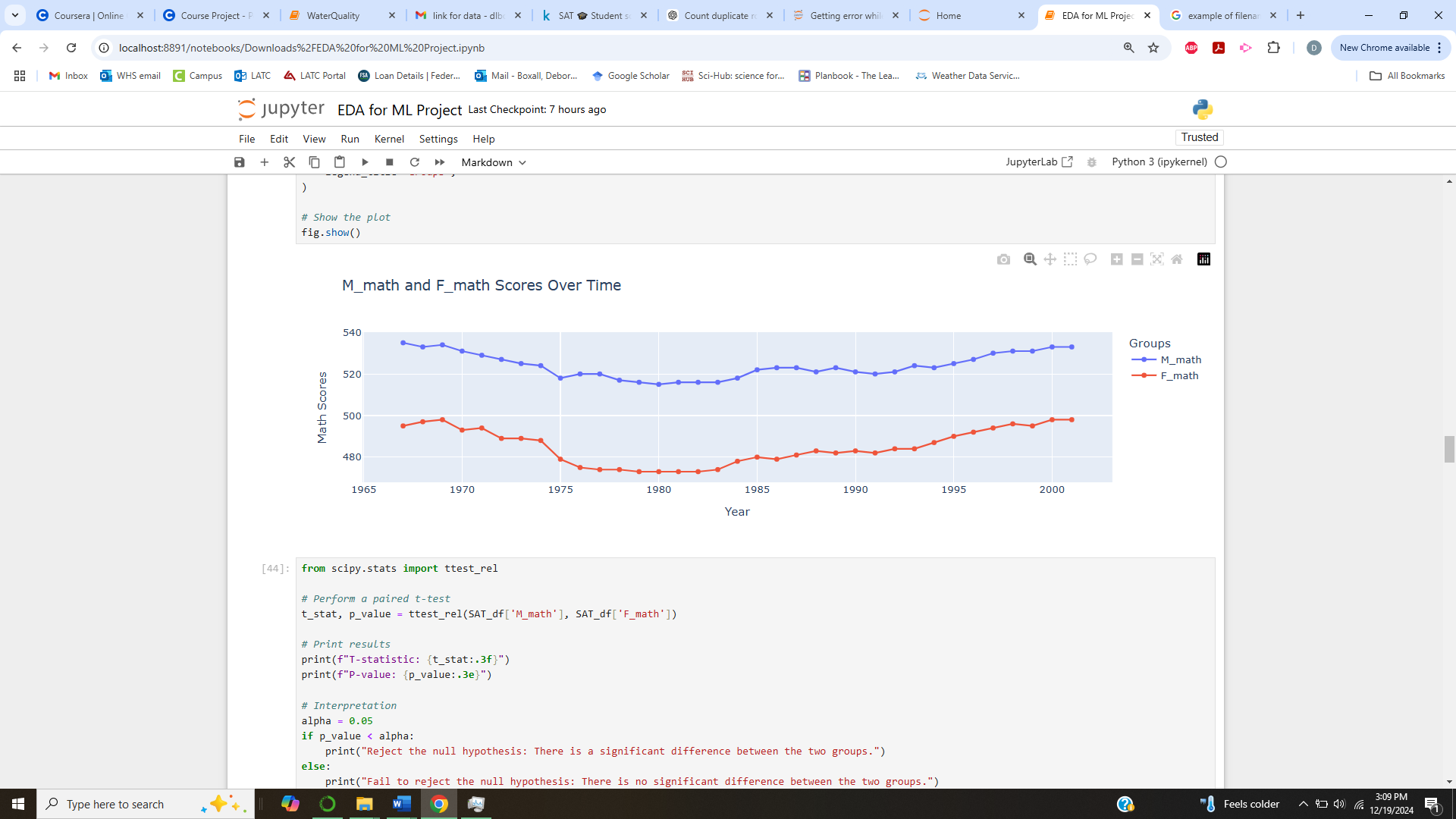
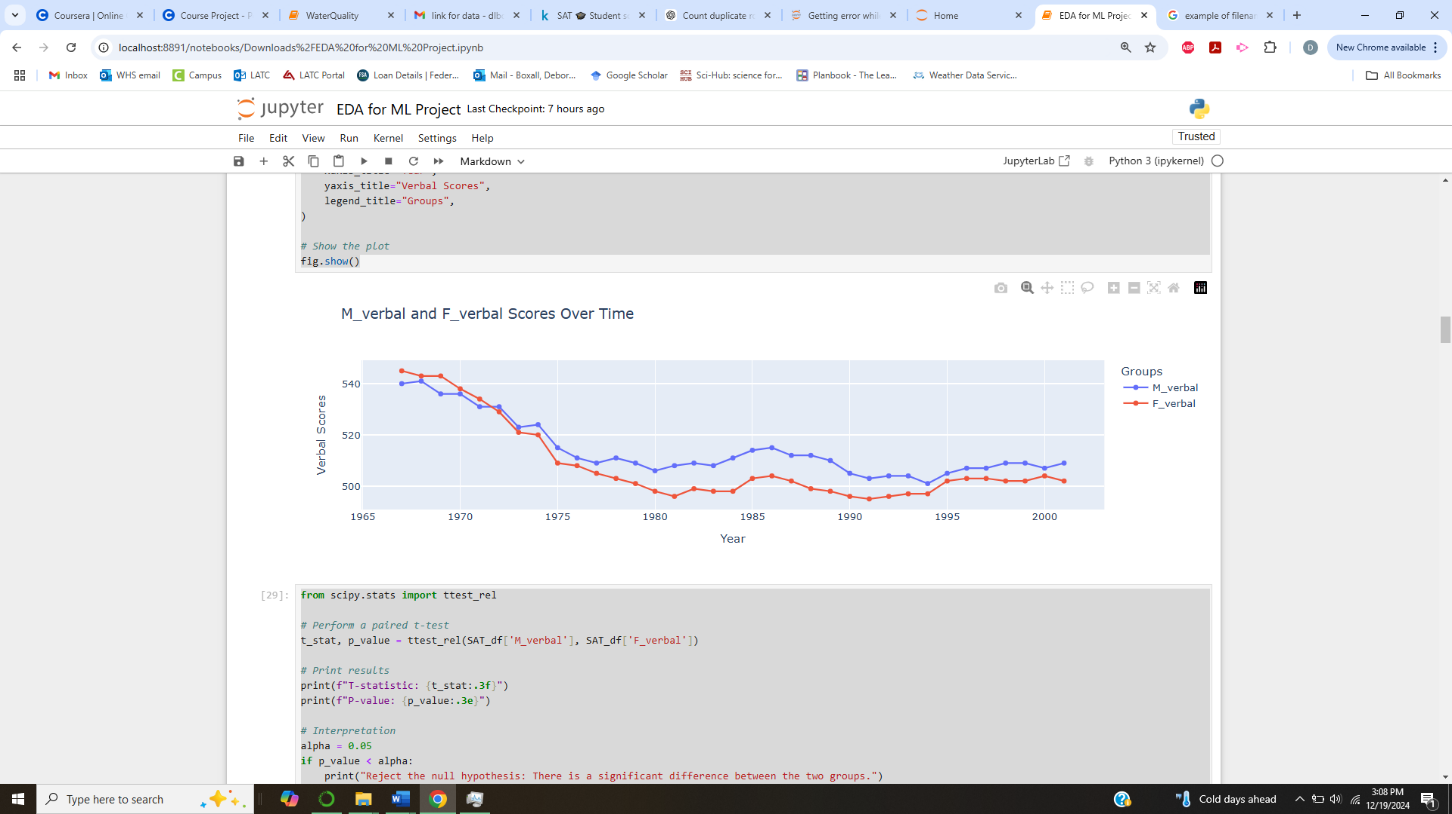
**Actions taken for data cleaning and feature engineering**

Data cleaning: There were no missing or duplicated entries found in the data set.

Feature engineering: The verbal and the math portions of the SAT, and the averaged values are all out of a maximum of 800 points so there was no need to normalize the values.

Preliminary investigations

Two scatter plots were prepared that compared the verbal and math scores of male (blue markers) and female (red markers) students. Up until 1972, female students scored slightly higher than males on the verbal portion of the SAT. After that point, female students consistently scored lower in both the verbal and the math portion of the SAT. In looking at the math scores, there has been a gradual increase in math scores for both genders after 1983, which suggests that the education initiatives may have had a positive impact. The verbal scores also showed a slight increase around the same time frame, but the scores drop off again starting around 1986.



Key findings and insights (synthesizes results of EDA in insightful and actionable manner)

Formulate three hypotheses about data

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedThe crossover-point in the verbal scores suggested that it might be interesting to investigate how the gap between male and female verbal scores changed over time. A linear regression was performed on the difference values by year and yielded an R2 value of 0.240 with a conditional number of 3.90 x 105. The large conditional number indicates either a multiple collinearity in the data, or that some of the features need to be scaled. The values for the x-axis were normalized and the regression was rerun. There was no change in the R2 value but the conditional number was only 4.27. This seems to suggest that the conditional value should be checked as part of linear regressions to determine if the x-axis values need to be normalized.

**Three hypotheses about the data:**

Hypothesis 1:there is no difference in average scores between male and female students

* **H₀:** There is no significant difference between the average math and verbal SAT scores for male and female students over the given time period.
* **H₁:** There is a significant difference between the average math and verbal SAT scores for male and female students.

Hypothesis 2: there is no trend in average SAT scores for all students with time

* **H₀:** The combined SAT scores (math and verbal averages) for all students do not change significantly over time.
* **H₁:** The combined SAT scores for all students show a significant trend (increase or decrease) over time.

Hypothesis 3: there is no difference between the verbal and math SAT scores for all students

* **H₀:** The average verbal SAT scores are equal to the average math SAT scores for all students over the given time period.
* **H₁:** The average verbal SAT scores are significantly different from the average math SAT scores for all students.

A computer screen shot of a computer screen

Description automatically generated