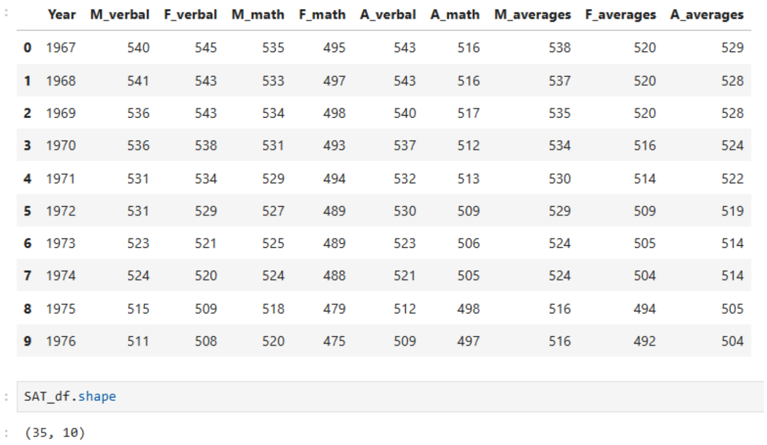
Dataset Summary

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).

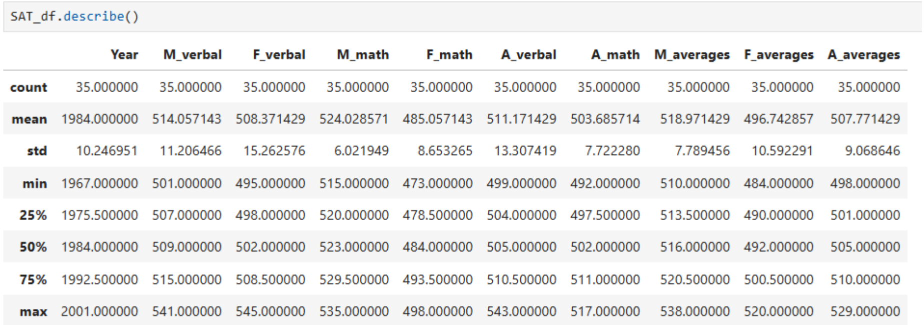


	Year	M_verbal	F_verbal	M_math	F_math	A_verbal	A_math	M_averages	F_averages	A_averages
0	1967	540	545	535	495	543	516	538	520	529
1	1968	541	543	533	497	543	516	537	520	528
2	1969	536	543	534	498	540	517	535	520	528
3	1970	536	538	531	493	537	512	534	516	524
4	1971	531	534	529	494	532	513	530	514	522
5	1972	531	529	527	489	530	509	529	509	519
6	1973	523	521	525	489	523	506	524	505	514
7	1974	524	520	524	488	521	505	524	504	514
8	1975	515	509	518	479	512	498	516	494	505
9	1976	511	508	520	475	509	497	516	492	504

```
SAT_df.shape
```

```
(35, 10)
```

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.



```
SAT_df.describe()
```

	Year	M_verbal	F_verbal	M_math	F_math	A_verbal	A_math	M_averages	F_averages	A_averages
count	35.000000	35.000000	35.000000	35.000000	35.000000	35.000000	35.000000	35.000000	35.000000	35.000000
mean	1984.000000	514.057143	508.371429	524.028571	485.057143	511.171429	503.685714	518.971429	496.742857	507.771429
std	10.246951	11.206466	15.262576	6.021949	8.653265	13.307419	7.722280	7.789456	10.592291	9.068646
min	1967.000000	501.000000	495.000000	515.000000	473.000000	499.000000	492.000000	510.000000	484.000000	498.000000
25%	1975.500000	507.000000	498.000000	520.000000	478.500000	504.000000	497.500000	513.500000	490.000000	501.000000
50%	1984.000000	509.000000	502.000000	523.000000	484.000000	505.000000	502.000000	516.000000	492.000000	505.000000
75%	1992.500000	515.000000	508.500000	529.500000	493.500000	510.500000	511.000000	520.500000	500.500000	510.000000
max	2001.000000	541.000000	545.000000	535.000000	498.000000	543.000000	517.000000	538.000000	520.000000	529.000000

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?

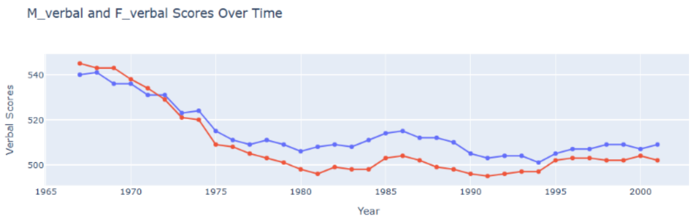
Actions taken for data cleaning and feature engineering

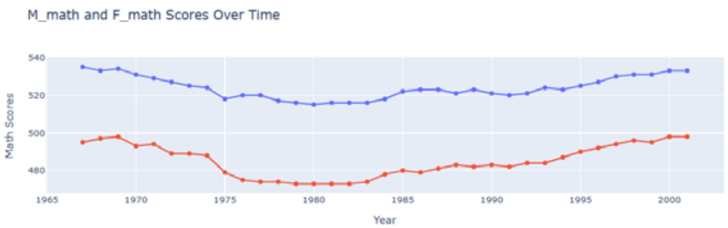
Data cleaning: There were no missing or duplicated entries found in the data set.

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.

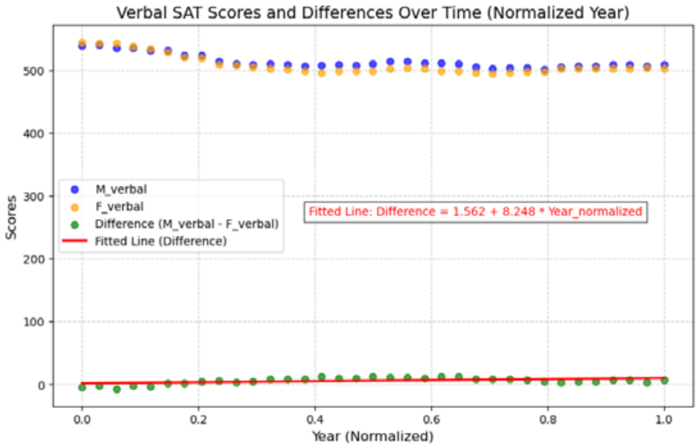
Preliminary investigations

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.

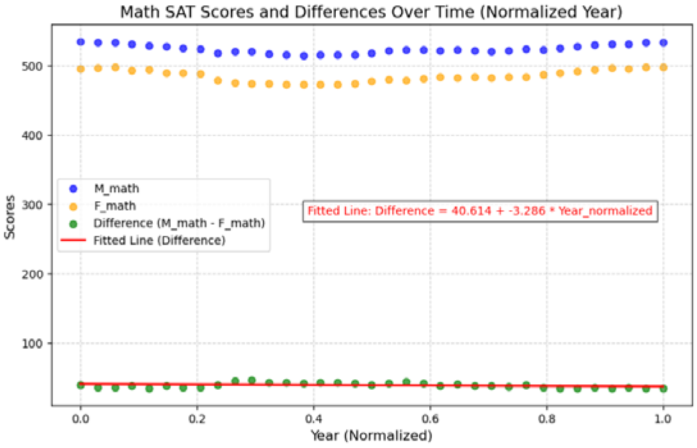




The crossover-point in the verbal scores suggested that it might be interesting to investigate how the gap between male and female verbal scores changed over time. A linear regression was performed on the difference values by year and yielded an R^2 value of 0.240 with a conditional number of 3.90 x 10^5. 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 R^2 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.



```
Linear Regression Details:
Slope: 8.248
Intercept: 1.562
R-squared: 0.240
Skewness of Differences: -0.702
Conditional Number: 4.271e+00
```



```
Linear Regression Details:
Slope: -3.286
Intercept: 40.614
R-squared: 0.087
Skewness of Differences: 0.396
Conditional Number: 4.271e+00
```

Additionally, the average math, verbal and combined scores grouped by decade were also investigated as part of the initial EDA. Interestingly, the boxplot for the average verbal score by decade suggested an exponential decay relationship in the data. In order to increase the number of data points available for fitting analysis, the data was also broken down into three-year chunks. A boxplot of the verbal scores for all students grouped in 3-year chunks also seemed to follow an exponential decay, albeit with less data smoothing due to the shorter time span.

Three hypotheses about the data:

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

- **H_0:** There is no significant difference between the average math and verbal SAT scores for male and female students over the given time period.
- **H_1:** 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 verbal SAT scores for all students with time

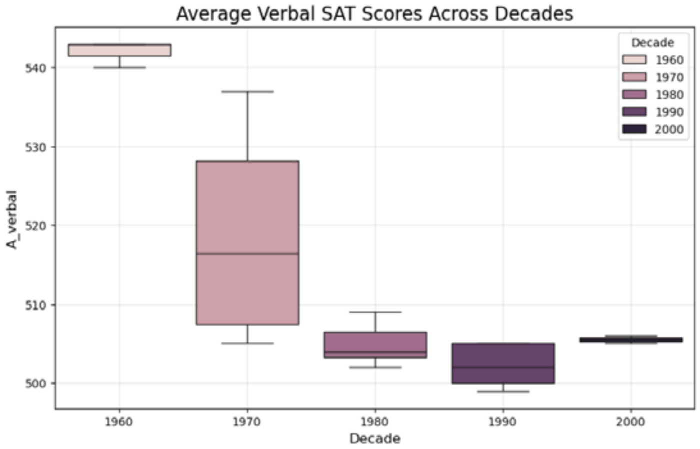
- **H_0:** The verbal SAT scores for all students do not change significantly over time.
- **H_1:** The verbal SAT scores for all students show a significant trend (increase or decrease) over time.

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

- **H_0:** The average verbal SAT scores are equal to the average math SAT scores for all students over the given time period.
- **H_1:** The average verbal SAT scores are significantly different from the average math SAT scores for all students.

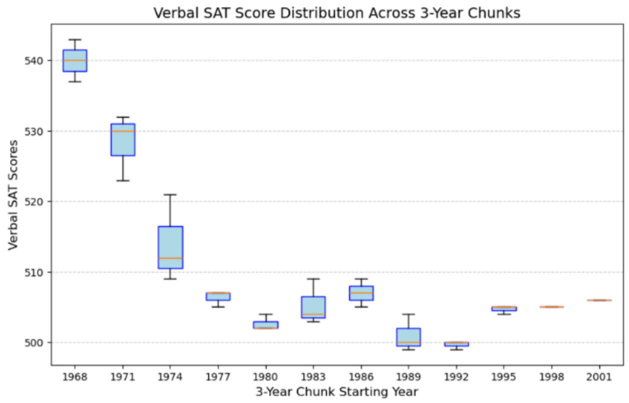
Results of formal hypothesis test

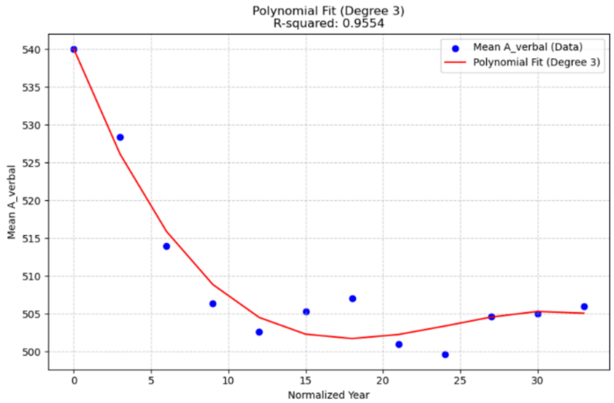
Due to the observed trend in the verbal SAT scores during EDA (see plot below), hypothesis 2 was chosen for additional statistical analysis.



Prior to fitting the data, a second box plot using a smaller time span of three years for the grouping of the data was prepared. As can be seen from the plot below, there is still an exponential-like decay in the grouped data. However, using the smaller data groupings also introduced enough noise that it was not possible to fit the data with an exponential curve.

As an alternative, several levels of polynomial fit were tried with the verbal SAT scores in three-year chunks. Cross-validation and alternative R^2 values were computed for each order, and it was determined that the 3rd order polynomial fit provided the best fit without introducing overfitting. (See below for statistics and plot.)

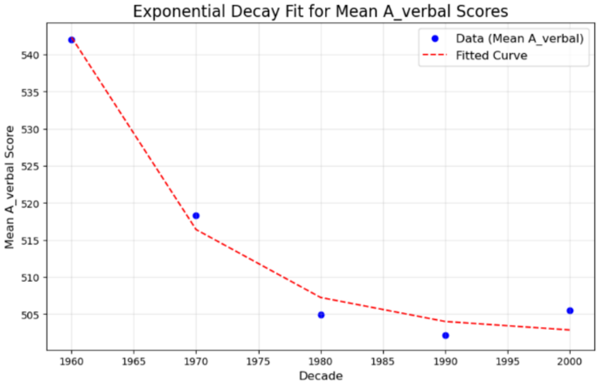




Polynomial order	Initial R^2 value	Adjusted R^2 value	Cross validation	
			Train RMSE	Test RMSE
Degree = 2	0.9013	0.8793	3.5073	5.6323
Degree = 3	0.9554	0.9387	2.8191	0.2889
Degree = 4	0.9630	0.9419	1.3083	13.4953

While the initial R^2 value for the 4th degree was much higher than the other two orders, the fact that there is little difference in the adjusted R^2 values for the 3rd and 4th degree fits, and the small RMSE for the 3rd order upon cross validation, supports using the 3rd order polynomial fit for the best fit of the verbal SAT scores for all students grouped in three year increments.

Because it was visually interesting, the verbal SAT score data grouped by decades was also investigated. It was found that while a polynomial fit gave slightly better R^2 values, the RMSE for the test data was higher than the training data for each order of fitting.



Exponential fit: initial R^2= 0.9815, adjusted R^2 = 0.9630

Polynomial order	Initial R^2 value	Adjusted R^2 value	Cross validation	
			Train RMSE	Test RMSE
Degree = 2	0.9982	0.9965	0.6293	0.700
Degree = 3	0.9996	0.9983	0.000	1.400
Degree = 4	1.000	NaN	0.000	0.4098

As a final step, the F-statistic and the probability of the F-statistic were determined for both the third order polynomial fit with data grouped in 3-year chunks, and for the exponential fit when the data was grouped by decade.

Fit	F-statistic	Probability of F-statistic
Exponential across decades	5.524	0.100
3rd order polynomial fit in 3-year chunks	57.14	9.55×10^{-6}

The probability of F-statistic for the exponential fit is greater than 0.05, which indicates that the observed trend is not significant. The 3rd order polynomial fit using 3-year chunks, while visually similar to the exponential fit, has a probability of F-statistic much smaller than 0.05. The small value indicates that there is a significant change in verbal SAT scores for all students. Therefore, H_o for hypothesis 2 is rejected.

Additional steps for data analysis with this dataset:

- Complete polynomial fits of the math scores for all students over three year chunks of time to determine if the same trend is observed.
- Determine if there is a significant difference between SAT scores for males and females (hypothesis 1)
- Determine if there is a significant difference between math and verbal scores for all students (hypothesis 3)

Evaluation of data set

While this data set was rather small, it was of high quality in that it did not have any missing or duplicated values. The data set also included useful comparisons by breaking the SAT verbal and math scores down by gender, verbal and math scores for all students, as well as the averaged score broken down by gender and for all students. However, it lacks the following additional information:

- Number of students that took the SAT by year
- Proportion of total student population that took the SAT by year
- Demographic information about the students that could impact their performance on the SAT (e.g. highest education level of parents, socio-economic status, area of residence in the US, etc.)

Additional information that would have aided in the analysis is a timeline of significant educational reform efforts that took place after the release of “A Nation at Risk.” Additionally, information about changes to how the SAT was written and scored between the years of 1967 – 2001 would also help identify contributors to score changes outside of student performance.