We are presented with a dataset that contains the scores of 1,000 students in 3 main subjects of focus: reading, writing, and math. The observations are completely independent, and represent a sample of high school students. Further information gathered about each student includes their: gender, race/ethnicity, parental level of education, and whether they took any test preparation course. The objective of this analysis is to explore how these factors influence students' grades and to derive actionable insights and recommendations based on the findings. A sample is shown in Figure 1 below.

StudentsPerformance

gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
female	group B	bachelor's degree	standard	none	72	72	74
female	group C	some college	standard	completed	69	90	88
female	group B	master's degree	standard	none	90	95	93
male	group A	associate's degree	free/reduced	none	47	57	44
male	group C	some college	standard	none	76	78	75
female	group B	associate's degree	standard	none	71	83	78
female	group B	some college	standard	completed	88	95	92
male	group B	some college	free/reduced	none	40	43	39

Figure 1: A Sample of the Students Performance Dataset

The data was collected anonymously to protect the privacy of the students. All three score columns range from 0 to 100. Also, notice how the race/ethnicity column is encoded with group numbers to reduce bias and adhere to ethical guidelines. The data provided is well-structured and clean and doesn’t need any additional preparation. Additional concerns may arise due to the insufficiency of the gathered field to fully model the data, but this will be discussed later.

Several hypothesis tests are set up to better identify our approach to the analysis, and understand the underlying value. The methods used for the rest of this report will rely on Bayesian Computational Statistics, and the programming language used to perform the computations and visualize the results is R. Several similar scripts are used to carry out the analysis, one for each hypothesis, and the library used across all of them is rstan.

We start by first investigating whether there are significant differences in the scores of each of the 3 subjects. A Bayesian model is set with non-informative priors (meaning that we’re

not assuming any previous knowledge about our data). The non-informative prior is a uniform distribution with the scores in the allowed range. Non-informative priors are also used for the rest of the hypotheses, since we don't know any information from before. Alternatively, we could assume that a certain subject has higher scores than another, or that a certain subject has a mean in a certain range. This wouldn't be necessarily useful, if not more harmful to our results, if we're not sure about our assumptions (especially that the number of observations is small).

The parameter of interest here is the mean of the scores for each subject. Again, this is also the case for all the other hypothesis. The likelihood function used for our parameter given the data is a normal distribution. This makes sense since the mean is expected to follow a distribution that has a peak around a particular value at the middle of the range, where it is highly probable, and the probability is expected to decrease as we move outwards towards the edges. The result of the analysis is shown in Figure 2.

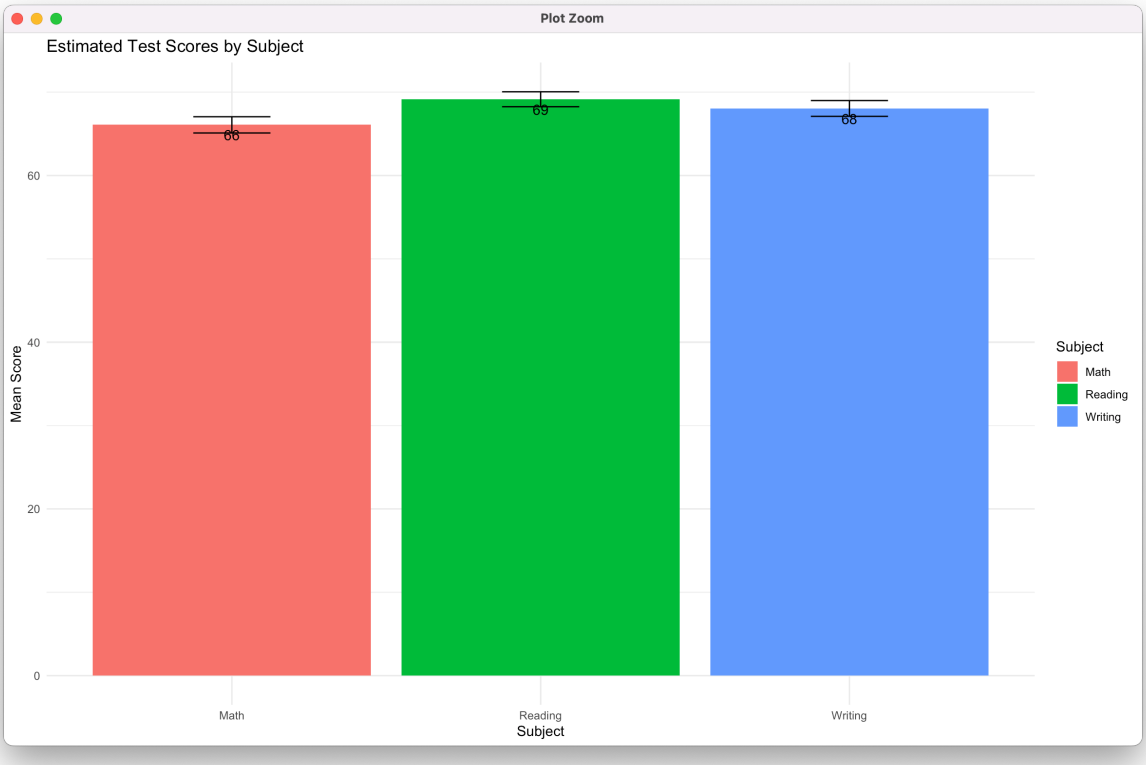

Figure 2: A Bar Chart Showing the Mean Scores of Each Subject

The estimated mean score is denoted by the number on each column, and the error bars (vertical black indicator on each column) represent a 95% credibility interval of the results; that is, there is a 95% chance that the mean falls within the range indicated by the error bar. The scores are 66, 69 and 68 for math, reading and writing respectively. The scores seem very similar and close, and this can be further emphasized by the credibility interval (the ranges are overlapping). We hence conclude that there are no significant differences between the scores of the three subjects.

The following four hypothesis relate the mean score of each subject to each of the other factors in the dataset. That is, we are interested in seeing how each of: the gender of the students, their parental education level, their race/ethnicity, and whether they took a test preparation course affect the scores. The same order of appearance of columns in Figure 1 is followed here for consistency. Similarly, the same approach is adopted where non-informative priors (uniform distribution of the mean ranging from 0 to 100) and a normal distribution likelihood of the parameters given the data are used.

Figure 3 shows the estimated mean scores vs gender for each subject. It is seen from the plot that there are significant differences between the scores of males and females in each of the three subjects. Even when taking the credibility intervals into consideration, there appears to be at least a 5 point difference in scores for each subject. Females appear to be performing better in reading and writing, while males appear to be performing better in math. It isn't possible to directly know the cause of the results, and it is difficult to make any assumptions not knowing what the source of the data is. For example, this could come from an area where gender roles are a heavy influencer on the culture, which makes males more focused on STEM career pathways and females more focused on arts and humanities pathways. It could also be a result of more focus on a certain gender group in some subjects. In any case, it is apparent that some adjustments are needed to eliminate gender bias and ensure that all students are receiving equal care and attention, as well as equal opportunities to the same pathways.

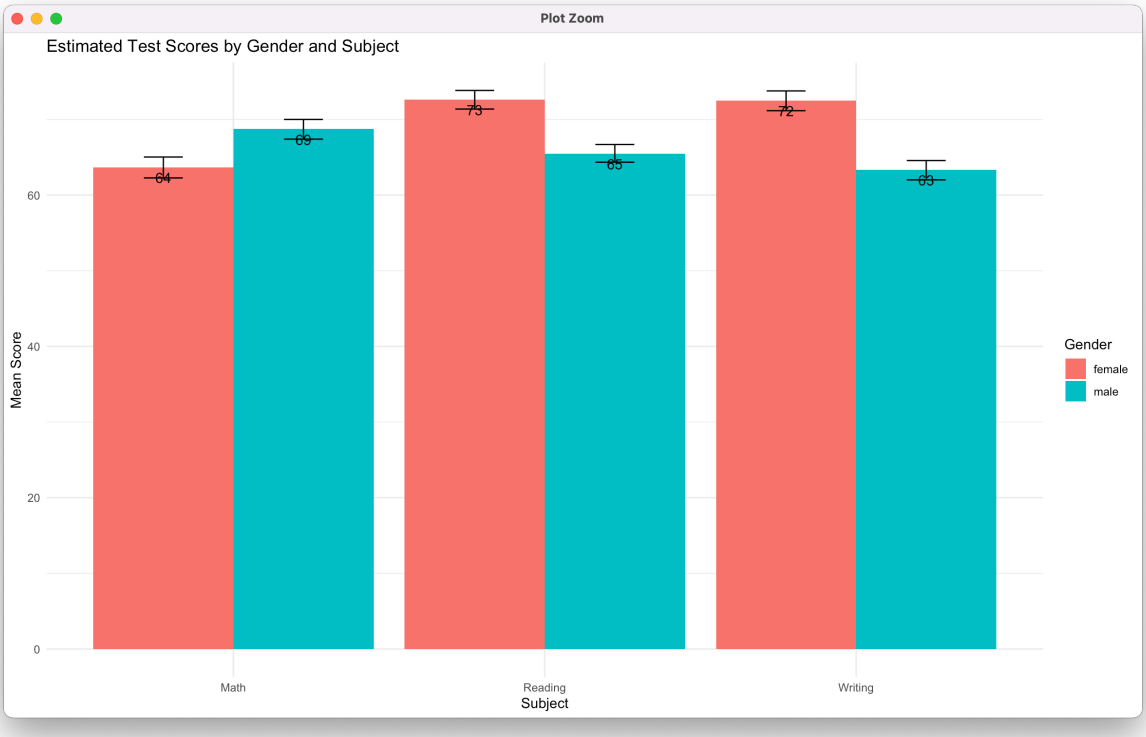

Figure 3: A Bar Chart Showing the Mean Scores of Each Subject by Gender

Figure 4 shows the estimated mean scores vs race/ethnicity groups for each subject. It is very obvious that there is a general trend where the academic performance of groups rises as we go from A to E; that is group E performs better than group D, and group D performs better than group C, and so on. It is also seen that certain groups (groups A and E for example) have wider credibility intervals than others. This could indicate the lower number of samples for that particular group making the data less reliable. Still, there seems to be a general trend across all subjects that certain groups are less competent than others. As before, it is difficult to identify the cause as it could be anything depending on the source of the data including, but not limited to: the effect of stereotyping on students, underrepresentation of certain groups, traditions and beliefs, or income level group of parents. Certain actions such as engaging families and communities, raising awareness of education for personal growth, promoting equity and inclusion, and certain policy changes could all be possible ways to address this problem.

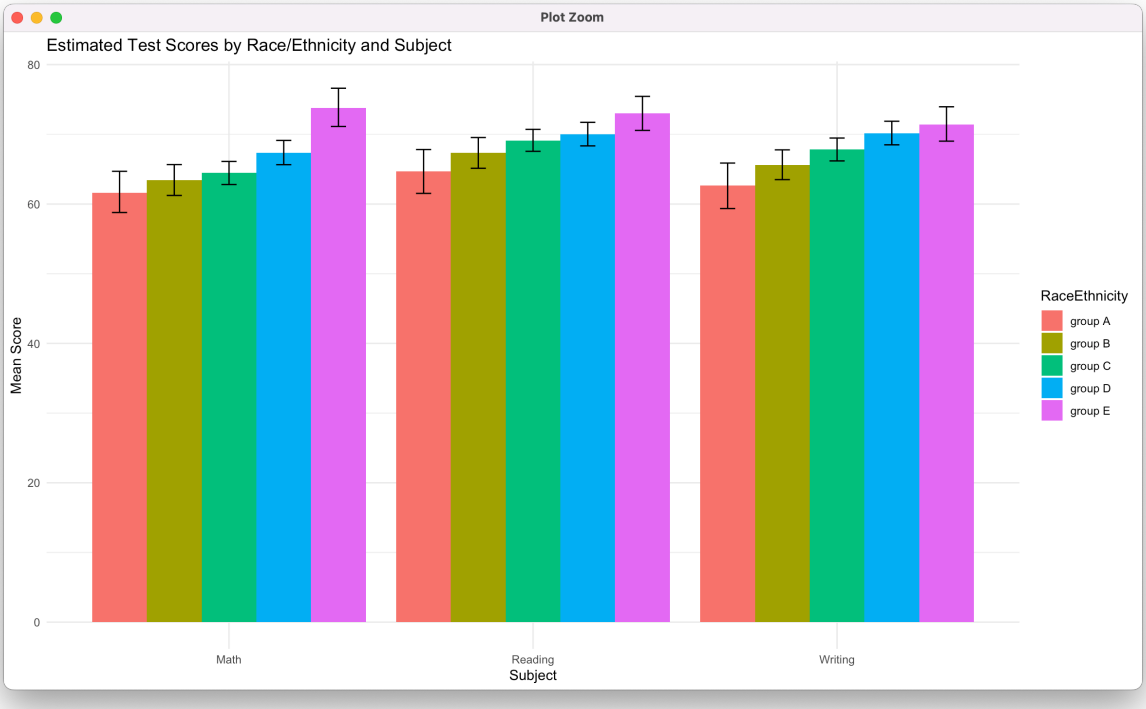

Figure 4: A Bar Chart Showing the Mean Scores of Each Subject by Race/Ethnicity Groups

Figure 5 shows the estimated mean scores vs parental educational level for each subject (their highest degree obtained). Perhaps this one shows very predictable results. The bars are ordered from left to right according to rising level of education, starting from some high school and up until Master's degree. It appears that the higher the level of education a student's parents achieve, the better it influences their results. One thing to note on the plot, though, is that the credibility interval for the Master's degree bar seems to include a wider range than others making it more difficult to decide whether a parent with a Master's actually influences their child's performance vs a parent with a Bachelor's. As a matter of fact, the same applies to nearly any particular degree level and its preceding degree level. For example, it is probably safe to say that someone with a Master's degree is more likely to influence the score of their child than someone with an Associate's degree, but it is not very guaranteed that someone with a Bachelor's will more likely influence the score of their child than someone with an Associate's. The same applies to all levels. One unexpected thing to note is that there seems to be some possible advantage to children whose parents responded "some high school" over parents who responded "high school". It is not a certain statement, but it could be true in some scenario where parents who dropped out of high school seem to be more focused on their children's educations than those who didn't. This hypothesis tells us that it is a good practice for schools to give more attention to children whose parents didn't receive higher education degrees as this will help bring more equality to the equation. It is also useful to raise awareness on the importance education with an extra focus on these students to ensure that their views of educations match those whose parents are already stricter in that sense. Another possible scenario is that the students whose parents achieved a relatively lower education level are unable to receive as much support in studying their subjects, in which case it would be a good idea to offer additional services such as tutoring to help compensate this deficiency if the schools' resources can allow it. To address this, further data is needed to identify whether this is actually the case. Surveys could bridge this gap.

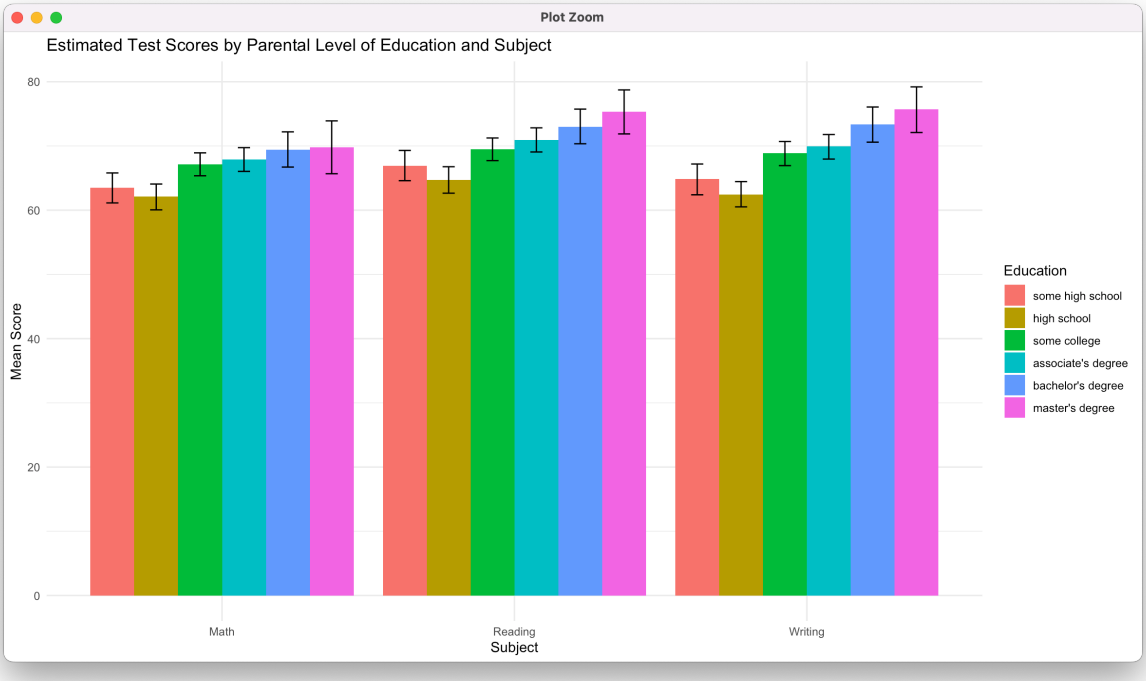

Figure 5: A Bar Chart Showing the Mean Scores of Each Subject by Parental Educational Level

Finally, Figure 6 shows the estimated mean scores vs status of test preparation course for each subject. Again, this one shows predictable results. Put simply, students who take a test preparation course are more likely to get higher scores than students who didn't. This is also proven by the credibility intervals. The estimated extra points that students score having taken the test are different for each subject, but ranges from 6 to 10 points, which is a considerable amount. When it comes to actions, this one is probably on the students and their parents to act on. Schools can only recommend taking the courses, but it is up to the students to take responsibility and complete the course.

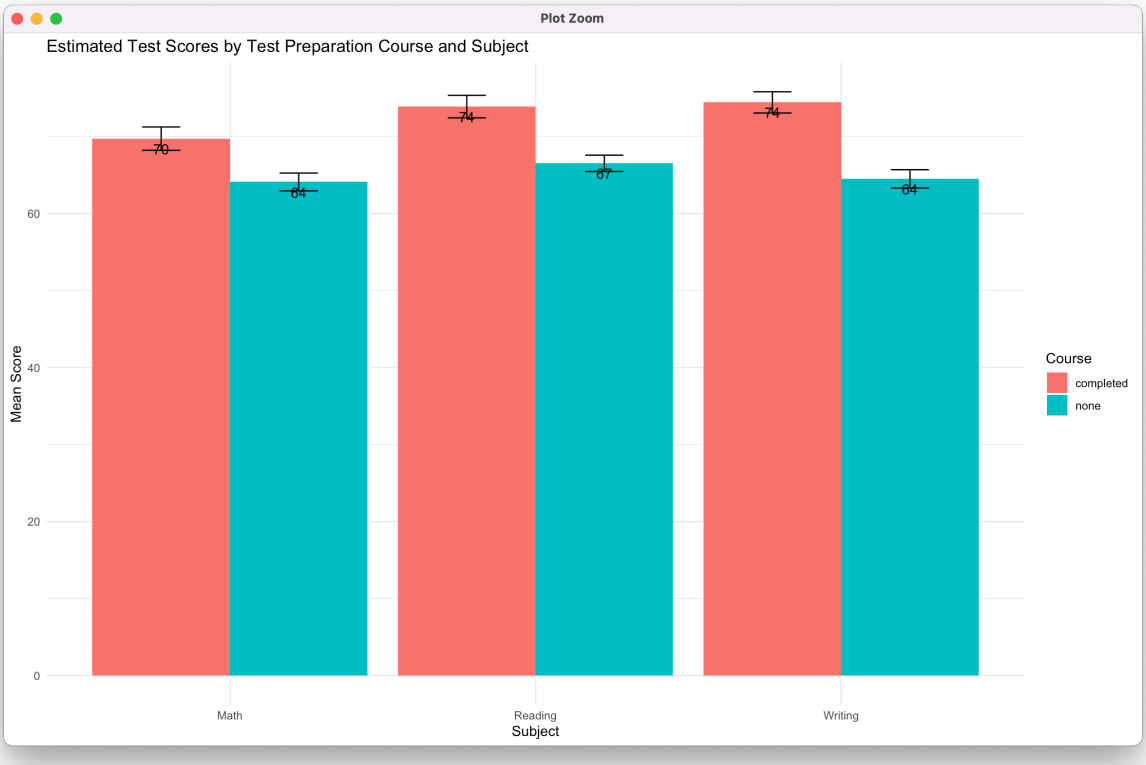

Figure 6: A Bar Chart Showing the Mean Scores of Each Subject by Test Preparation Course Status

Some final thoughts on this analysis are also very important to take into consideration when working towards taking decisions to achieve change. The dataset is very small, representing only 1,000 students. Given that there are millions of high schoolers every year, it would be very unwise to make any decision on a global scale using this data. It could be used effectively, however, for a particular neighborhood or particular high school since this data will be more reflective of this one entity. Therefore, the dataset might be overfitting to the data or observations we have and might not be well generalizing for a larger scale. This could also be seen in some of the figures where plots show overlapping credibility intervals, indicating that the true/unbiased estimate might be in any of the wide range provided by the error bars. In some cases, such as deciding whether a parent's Master's degree is likely to influence a higher performance on their children could not be done. The same could be said about some racial/

ethnic groups with respect to others (groups B vs C for example, since the results are not statistically significant). In others, such as deciding whether the test preparation course affects the scores, reliable conclusions could be made. Also, in areas where some issues were identified, it is also important to gather additional data to determine the root causes of those issues so that the correct actions could be done. It would've been more helpful as well to know where the data came from, as this could tell us further information that couldn't be inferred by the data alone. Regarding the analysis results on R, it could be seen by running the scripts that the values of Rhat are all ones for all of the posterior parameters indicating convergence, so the number of chains and iterations were picked appropriately. Regarding the approach taken for this problem, which is using Bayesian Statistics, it could be argued that it adds unnecessary complications, and it would be easier to use a frequentist approach. After all, Bayesian Statistics are more helpful in scenarios where we have some prior information or assumptions about our parameters, which wasn't the case in this problem where non-informative priors were chosen. Therefore, this is exactly equivalent to grouping data by certain columns and calculating the average on the numerical (score) columns with an additional complexity of setting up models, parameters, likelihood functions, etc. It was verified that performing analysis from the frequentist perspective, following the aforementioned method, yields the exact same results. One final note to add here is that the column lunch was ignored in the analysis because there is no reason to believe that it affects the results in any meaningful way. Since the number of attributes and number of rows are very small anyways, the lunch column was left in the dataset since it is very unlikely to affect any performance.