Data Analysis Projects

ST8002-Project

Elaine

2023-12-22

Contents

1 Project 1

1.1 Background

In examining a dataset featuring exam scores and corresponding hours of study, this analysis endeavors to unravel the intricate relationship between academic performance and study efforts. The background underscores the critical role of study time in influencing exam outcomes and aims to delineate patterns that contribute to student success. The central questions to be addressed revolve around understanding the correlation between hours of study and exam scores. Specifically, the analysis seeks to determine if there exists a statistically significant relationship between the two variables and whether variations in study time can reliably predict variations in exam performance. Additionally, the investigation aims to identify any potential threshold of study hours associated with optimal academic achievement. This exploration into the dynamics of study habits and academic success holds implications for educational strategies and offers insights into factors contributing to student performance.

1.2 Data

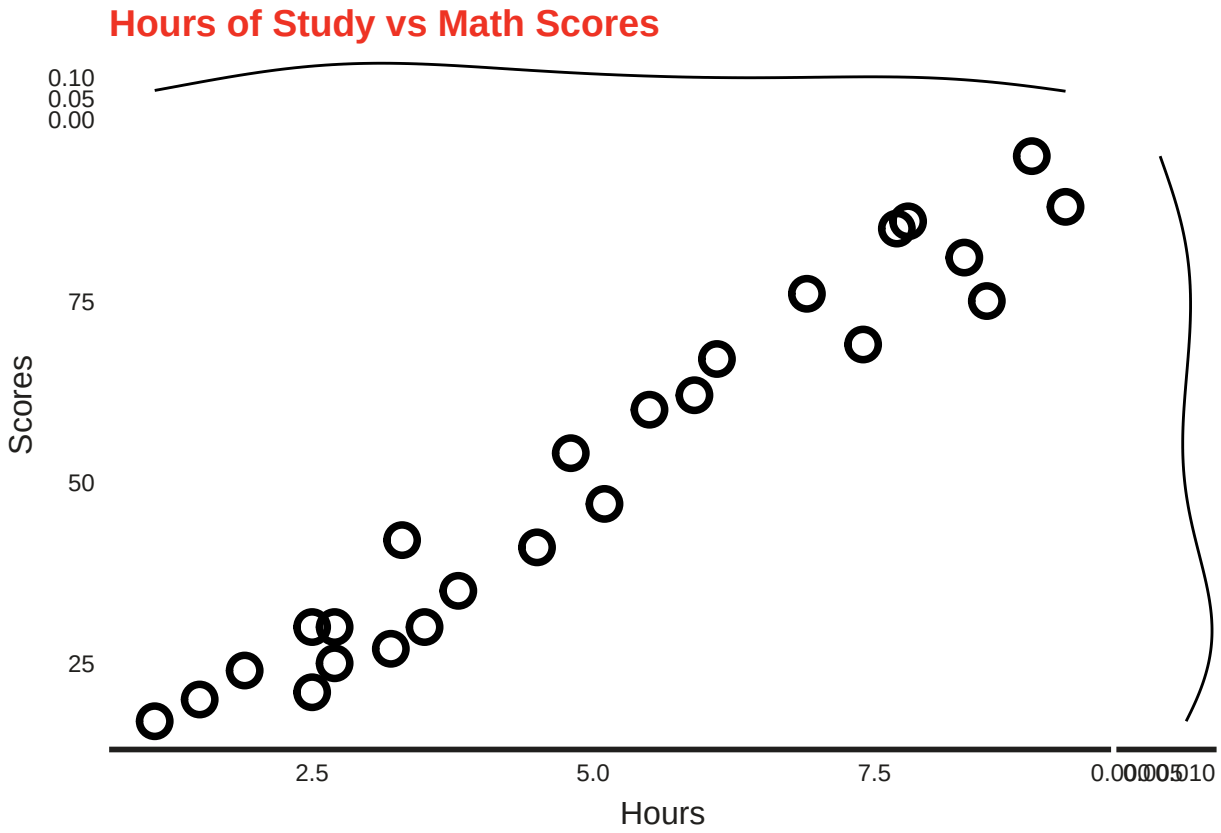
In this project, we analyse data from Kaggle that consists of maths scores and the hours of study [1].

Hours	Scores
2.5	21
5.1	47
3.2	27
8.5	75
3.5	30
1.5	20

1.3 Data Exploration

I start by doing a plot of the hours of study and the test scores.

[1]The data is available on this link, https://www.kaggle.com/datasets/samira1992/student-scores-simple-dataset



We see a substantial positive correlation (+0.976). This is the reason the plot is upward sloping and steep. But is it significant? To find out, we run a correlation test.

1.4 Hypothesis Test

1.4.1 Hours vs Test Scores: Correlation Test

We run the correlation test has the following assumptions.

- Both variables are on an interval or ratio level of measurement.
- Data from both variables follow normal distributions.
- Your data have no outliers.
- Your data is from a random or representative sample.

We test the following hypotheses.

H0: True correlation is equal to 0

H1: True correlation is not equal to 0

```
##
##  Pearson's product-moment correlation
##
## data:  Hours and Scores
## t = 22, df = 23, p-value <0.0000000000000002
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
##  0.946 0.990
## sample estimates:
##   cor
## 0.976
```

The p-value is negligible. We reject the null hypothesis and go with the alternative hypothesis. True correlation is not equal to 0, and hence there is a statistically significant relationship between exam scores and hours of study.

1.5 Conclusion

In this analysis, we examine the relationship between hours of study and test scores. We find that the correlation between the two variables is significant. Hours of study have a strong relationship with hours of study.

2 Project 2

2.1 Background

This analysis aims to examine the associations between math scores and demographic/educational variables, including gender, race/ethnicity, parental education, lunch status, and completion of a test preparation course. The objective is to uncover patterns and disparities that may elucidate the impact of these factors on mathematical achievement. The study recognizes the multifaceted nature of educational outcomes and seeks to inform educational policies by identifying areas where targeted support or resources can enhance academic success among diverse student populations.

2.2 Data

The data is available on this link. I pick the data that has 1000 observations [2].

specifically, the data 8 variables and 1000 observations regarding exams scores. The following are the variables in the data.

[2]See the data source here, http://roycekimmons.com/tools/generated_data/exams

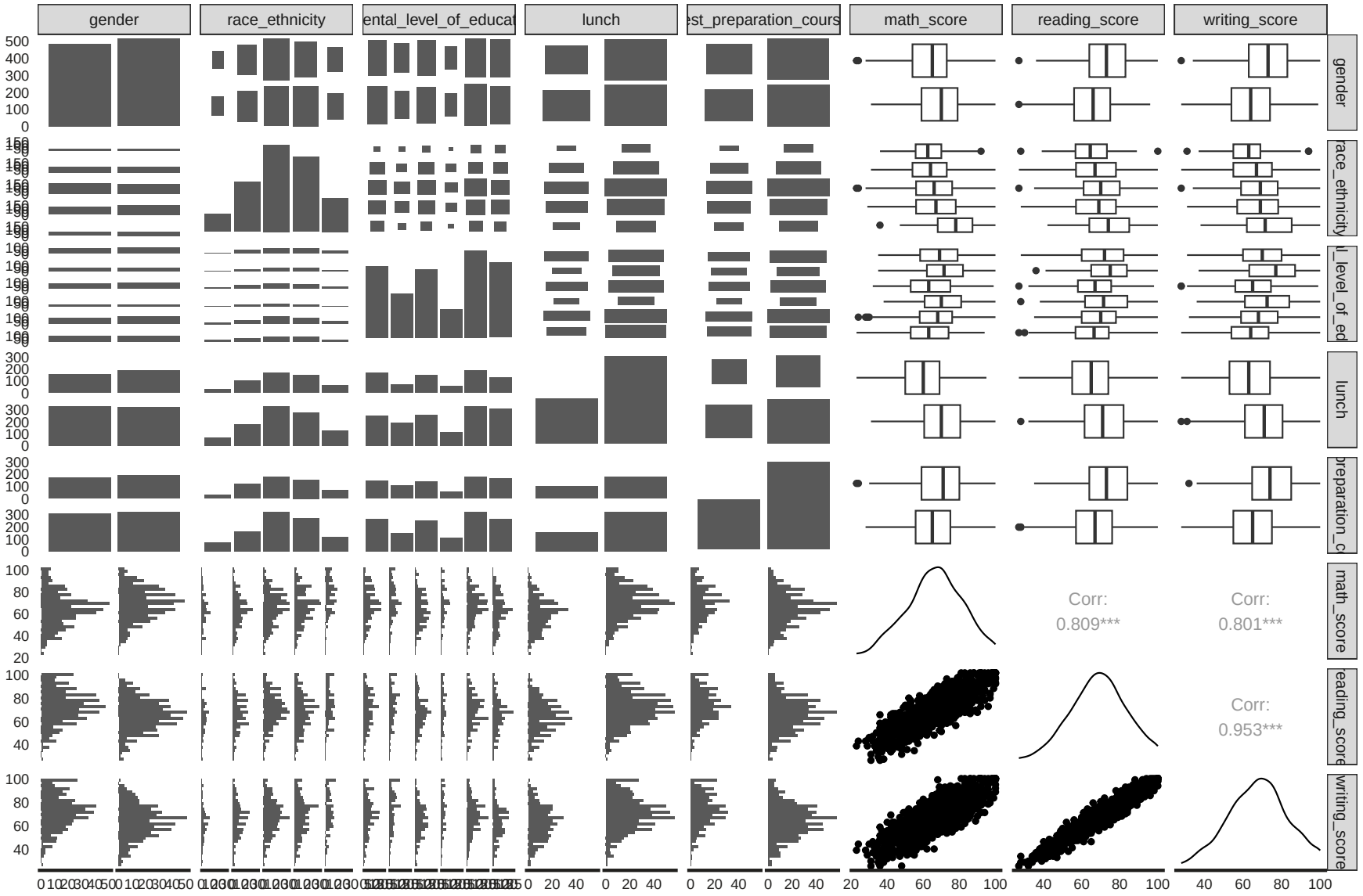


Figure 1: Pairs Plot

I also examine the differences in maths scores and parents education. We see that parents with low levels of education tend to have kids with extremely low scores. The average score does not appear to be affected as these parents also tend to have kids with high math scores.

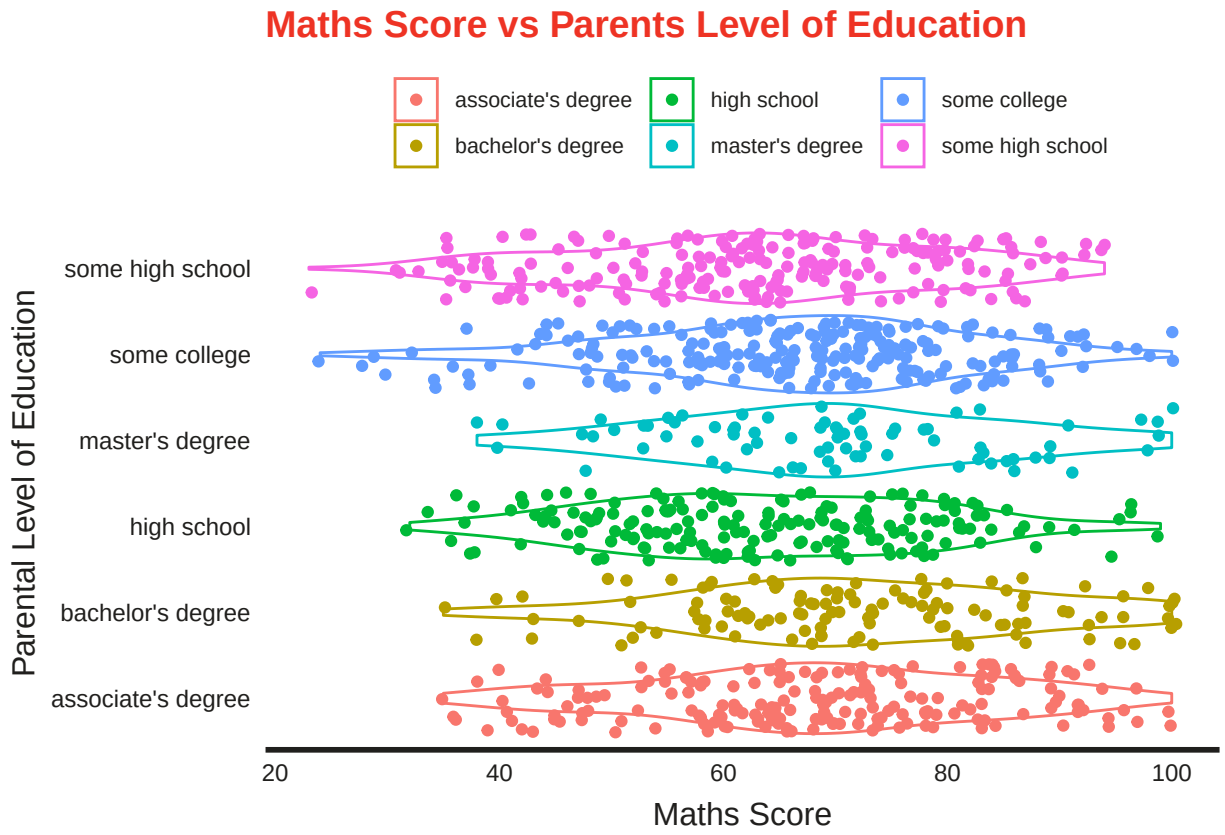


Figure 2: Maths Score vs Parents Level of Education

2.2.1 Race and Test Preparation: Chi-Square Test

To get a sense of people that take test preparation scores, I prepare a table. We see a clear difference in the number of people that take test preparation courses.

race_ethnicity	test_preparation_course	n
group A	completed	20
group E	completed	47
group A	none	48
group B	completed	78
group E	none	81
group D	completed	103
group B	none	111
group C	completed	116
group D	none	181
group C	none	215

To test whether this difference is significant, we do a chi-square test. Chi-Square test is a statistical method which used to determine if two categorical variables have a significant correlation between them. The two variables are selected from the same population. Furthermore, these variables are then categorised as Male/Female, Red/Green, Yes/No etc.

We test the following hypotheses:

H0: There is NO significant difference in the rate of taking test preparation scores by race.

H1: There is a significant difference in the rate of taking test preparation scores by race.

We now run the test.

```
##
##  Pearson's Chi-squared test
##
## data:  table(exams$test_preparation_course, exams$race_ethnicity)
## X-squared = 4, df = 4, p-value = 0.5
```

The p-value of 0.5 means that we go with the null hypothesis. There is no significant difference in the rates that different races and ethnicity take test preparation scores.

2.2.2 Are Maths Scores Above/Below Average? One Sample T-test

In this section, we test the hypothesis that the average grade in the maths test was greater than 66.6 (the average score for the maths test is 66.617). We run a one sample t-test that has the following assumptions.

- The dependent variable must be continuous (interval/ratio).
- The observations are independent of one another.
- The dependent variable should be approximately normally distributed.
- The dependent variable should not contain any outliers.

I start by testing for normality using the Shapiro-Wilk test. The test has the following hypothesis (Razali, Wah, et al. 2011).

H0: The data are not significantly different from normal.

H1: The data are significantly different from normal.

From the output, the p-value < 0.05 implying that the distribution of the data are significantly different from normal distribution. In other words, we can assume the normality.

```
##
##  Shapiro-Wilk normality test
##
## data:  exams$math_score
## W = 1, p-value = 0.0006
```

We run an outlier test that picks values that are greater than 1.5 times the IQR. We see that there are two values 24, and 23. To overcome the presence of outliers, we take the logarithms

of the variables.

```
## [1] 24 23
```

In both cases, we solve the problem by taking the logarithm of the maths score. Now we test the hypothesis that the mean score is greater than 4.17 (the log of 66.6 is 4.17).

We test the following hypothesis.

H0: The average score in the maths test is NOT greater than 4.17.

H1: The average score in the maths test is greater than 4.17.

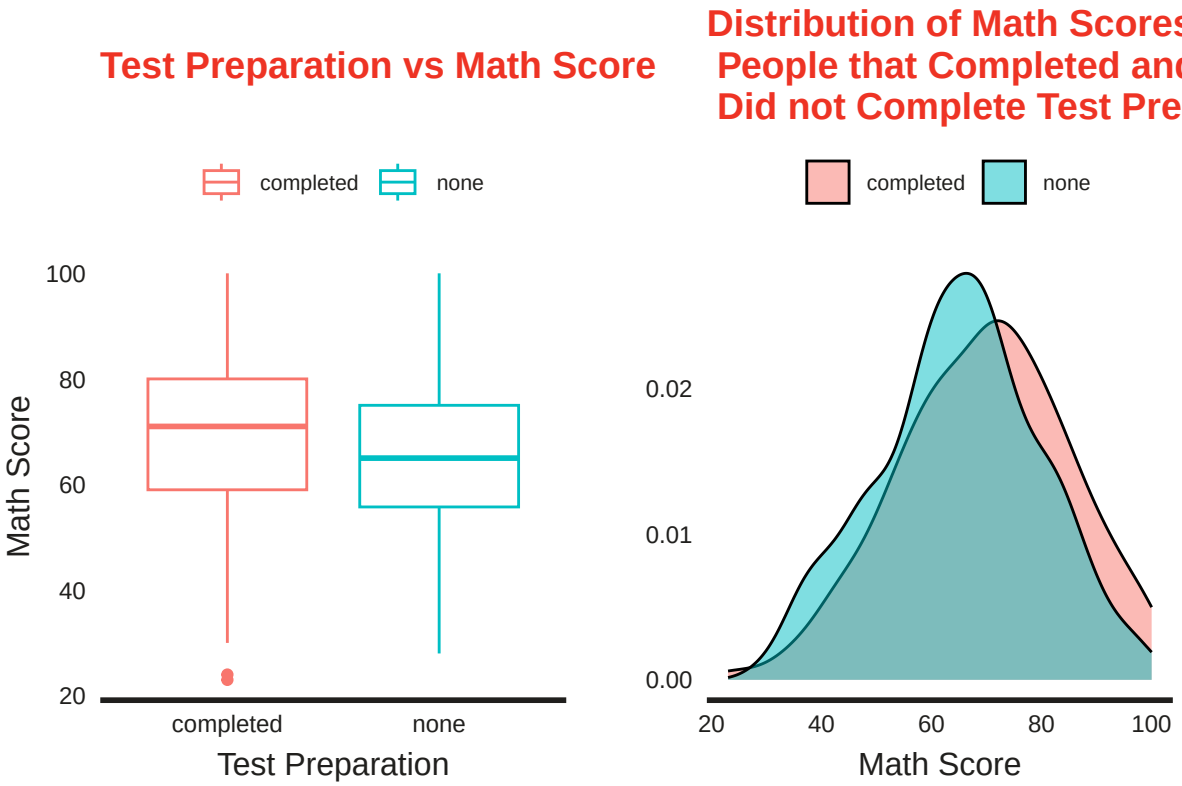
```
##
##  One Sample t-test
##
## data:  .
## t = 0.08, df = 999, p-value = 0.5
## alternative hypothesis: true mean is greater than 4.17
## 95 percent confidence interval:
##  4.16  Inf
## sample estimates:
## mean of x
##      4.17
```

```
## [1] 4.17
```

Going by the p-value of 0.5, we accept the null hypothesis and do away with the alternative. The mean test of the maths score is not greater than 4.17 (score of 66.6).

2.2.3 Test Preparation and Math Scores (T-test)

In this section, we test whether students who take test preparation courses do better in maths than those students that do not. There are 364 students that take the preparation course and 636 students that do not. Let us see the visualization of the scores for these groups



We see from the graph that there is a difference in the median maths score, with the group that completed the test having a higher median math score. But is this difference statistically significant? We use a t-test, that has the following assumptions.

- The data are numeric.
- Observations are independent of one another (that is, the sample is a simple random sample and each individual within the population has an equal chance of being selected)
- The sample mean is normally distributed
- Equal variances between groups.

The visualization above shows close to a normal distribution and roughly equal variances (238 vs 216), as shown below;

```
## # A tibble: 1 x 1
##   var_with_prep
##           <dbl>
## 1          238.
```

```
## # A tibble: 1 x 1
##   `var_no_prep <- var(math_score)`
##                              <dbl>
## 1                             216.
```

We test the following hypotheses.

H0: There is no statistically significant difference in average test scores between students that take the test preparation score and those that do not.

H1: There is a statistically significant difference in average test scores between students that take the test preparation score and those that do not.

This is a 2-tailed test. I run a t-test below.

```
##
##  Welch Two Sample t-test
##
## data:  math_score by test_preparation_course
## t = 5, df = 726, p-value = 0.000007
## alternative hypothesis: true difference in means between group completed and group no
## 95 percent confidence interval:
##  2.55 6.47
## sample estimates:
## mean in group completed      mean in group none
##                    69.5                    65.0
```

We see that the p-value of 0.000007, meaning that we reject the null hypothesis and go with the alternative. Test preparation scores seem to improve the average math score for the students.

2.2.4 Maths Score vs Race: Analysis of Variance

Is there a relationship between race and ethnicity and the average math scores? We test this assertion using the analysis of variance (ANOVA). There are three primary assumptions in ANOVA:

- The responses for each factor level have a normal population distribution.
- These distributions have the same variance.
- The data are independent

Specifically, we test the following hypotheses.

H0: There is no significant difference in mean math scores between difference race and ethnic groups.

H1: There is a significant difference in mean math scores between difference race and ethnic groups.

We first look at the visual depicting math scores against race and ethnicity. We see a clear difference, especially race Group E has a much higher maths score. In our test, we want to see whether this difference is statistically significant. In the chart showing the distribution of the math score by race, there is no significant departure from the normal bell shape and hence the assumption of normality is realistic.

```
##                  Df Sum Sq Mean Sq F value              Pr(>F)
## race_ethnicity    4  18376    4594    21.7 <0.0000000000000002 ***
## Residuals       995 210307     211
```

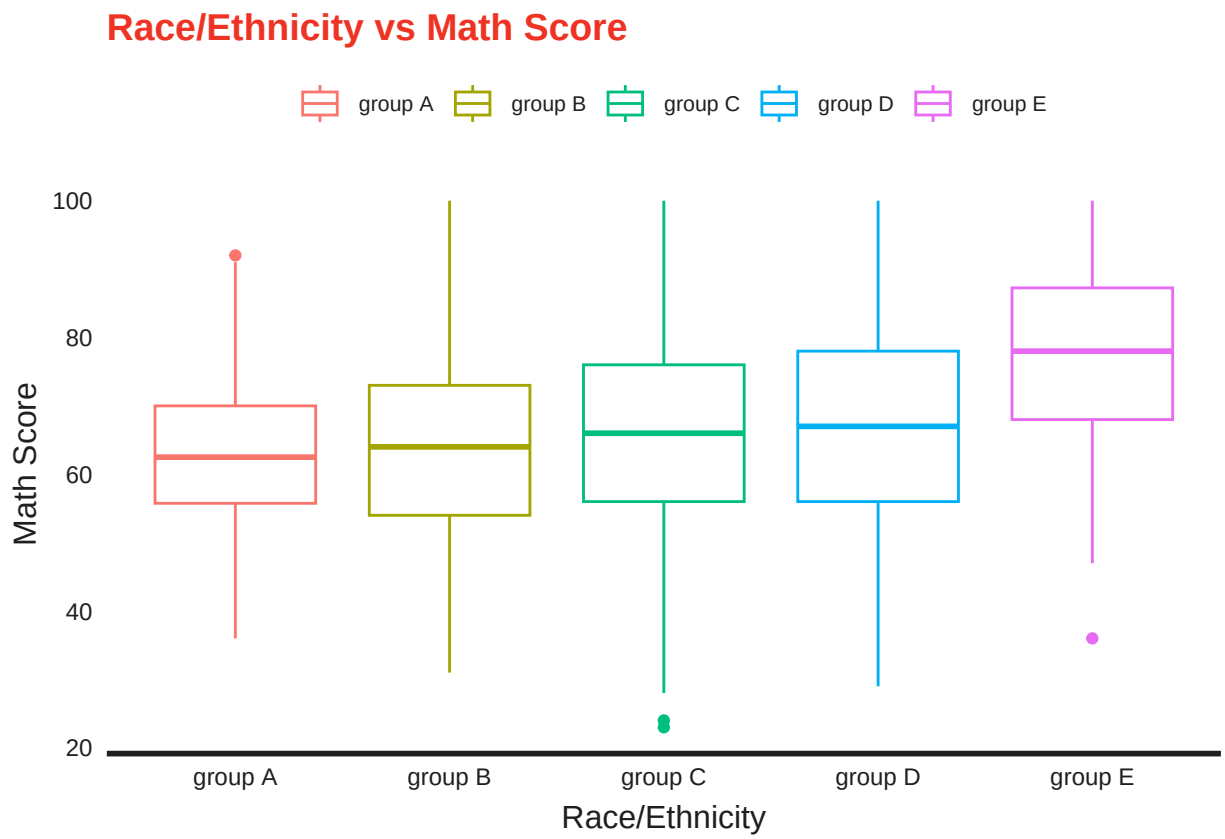


Figure 3: Box Plot of Math Scores by Race

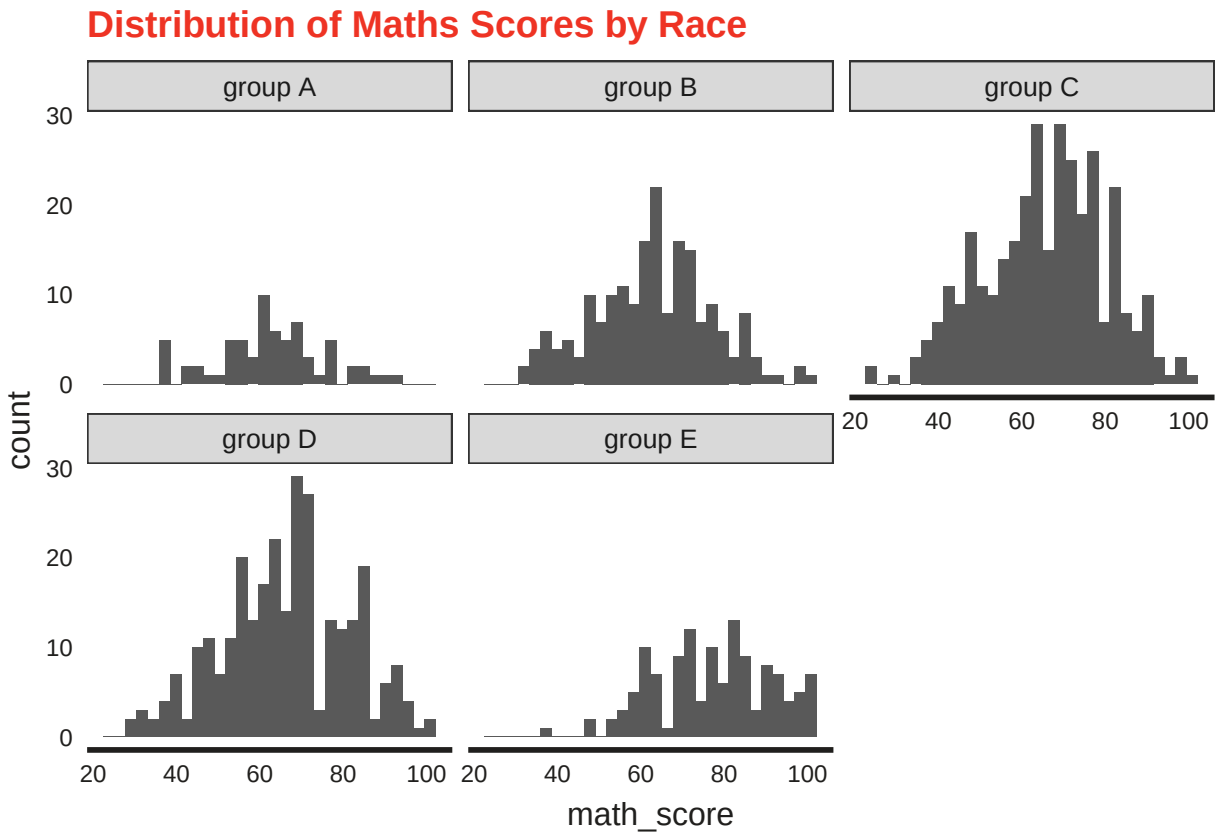


Figure 4: Distribution of Maths Scores by Race

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see a p-value of close to zero. Hence, we reject the null hypothesis and go with the alternative hypothesis. There is a significant difference in mean math scores between difference race and ethnic groups.

2.3 Regression Analysis and the F-test

In this section, I examine the factors that have a significant relationship with maths scores in test using regression analysis.

> The null hypotheses are that each of the coefficients is zero. The alternative hypotheses are that the coefficients are significantly different from zero.

We see from the regression table that kids with a parent with an associate degree have a higher test scores on average. Likewise, kids with high scores in reading and writing tend to have higher test scores, on average, compared to kids with low test scores in reading and writing. Kids on a standard meal have higher test scores on average compared to kids on a free meal. This could reflect the socio-economic background of the kids. In terms of explanatory power, the model has an R^2 of 0.85. This means that the model can explain 85% of the variation in maths scores.

The **F-statistic** is also important in this case. Being significant, we see that the model explains the variation in maths scores better than a model with no independent variables. The null hypothesis is that the model does no better than a null model- a model with no variables. The alternative hypothesis states that the model is significantly better than a null model.

The model diagnostics in the figure below shows how well the model fits the data. The goal of the posterior predictive check is to drive intuitions about the qualitative manner in which the model succeeds or fails, and about what sort of novel model formulation might better capture the trends in the data (Kruschke 2014). We see that the model fits the data pretty well. Again, in panel 2, we see that the relationship between dependent and independent variables is approximately linear. Panel 3 shows that the model does not have a serious heteroscedasticity issue. Again, there are no extreme values that could skew the regression (see the influential observations panel). There was very high correlation between the reading and writing score which raised the risk of multi-collinearity (and the resultant unstable coefficients). Hence I drop the writing score. Finally, the residuals appear relatively flat and within the line, meaning that the model could be useful for making predictions.

Table 3

	Dependent variable:
	math_score
test_preparation_coursenone	2.080***
	(0.432)
parental_level_of_educationbachelor's degree	0.215
	(0.746)
parental_level_of_educationhigh school	−1.460**
	(0.660)
parental_level_of_educationmaster's degree	−0.270
	(0.861)
parental_level_of_educationsome college	−1.080*
	(0.620)
parental_level_of_educationsome high school	−1.220*
	(0.650)
gendermale	12.000***
	(0.418)
reading_score	0.917***
	(0.015)
lunchstandard	4.760***
	(0.438)
Constant	−6.570***
	(1.310)
Observations	1,000
R^2	0.826
Adjusted R^2	0.824
Residual Std. Error	6.340 (df = 990)
F Statistic	522.000*** (df = 9; 990)
Note:	*p<0.1; **p<0.05; ***p<0.01

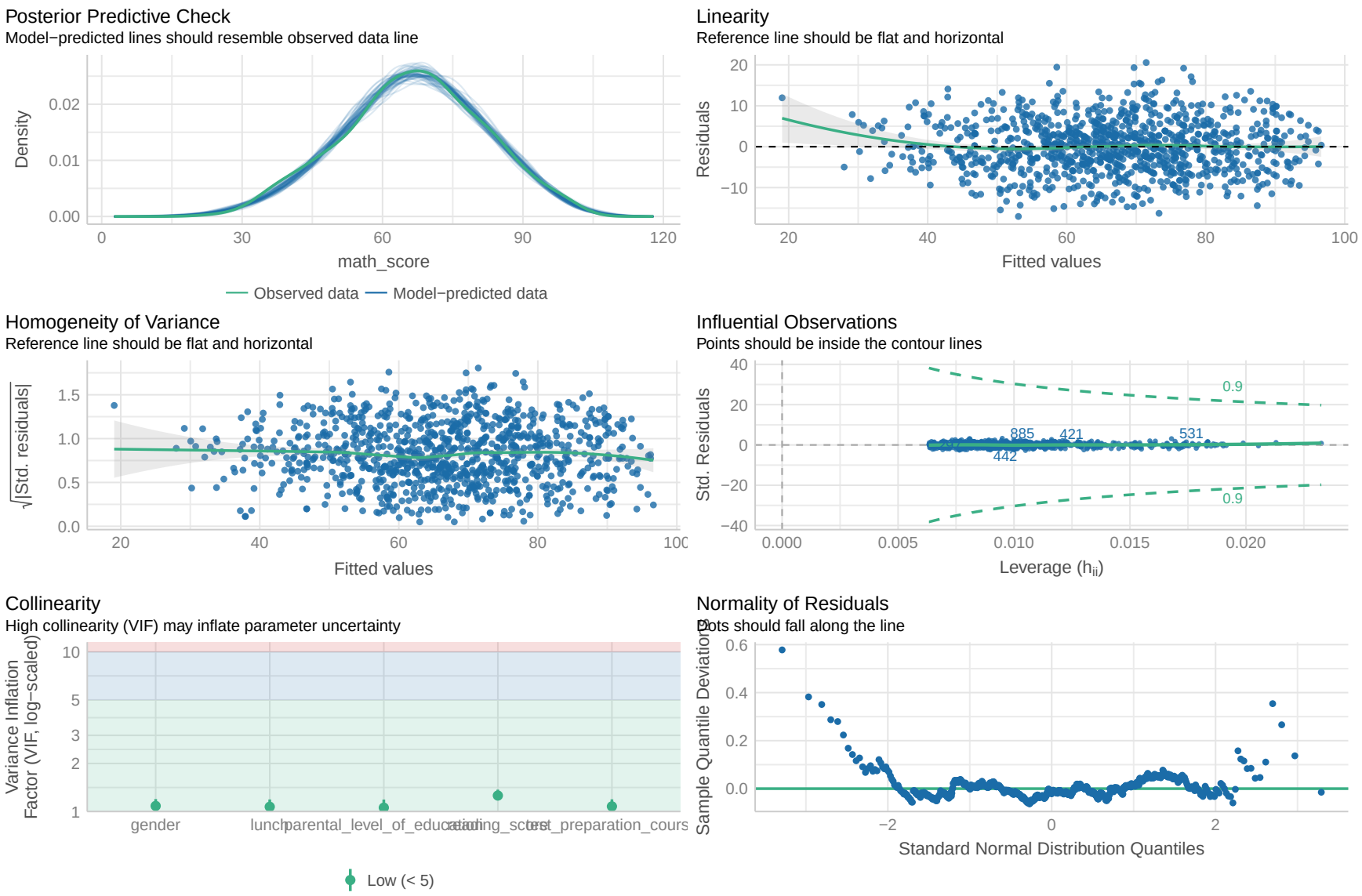


Figure 5: Model Diagnostics

2.4 Multivariate Analysis of Variance

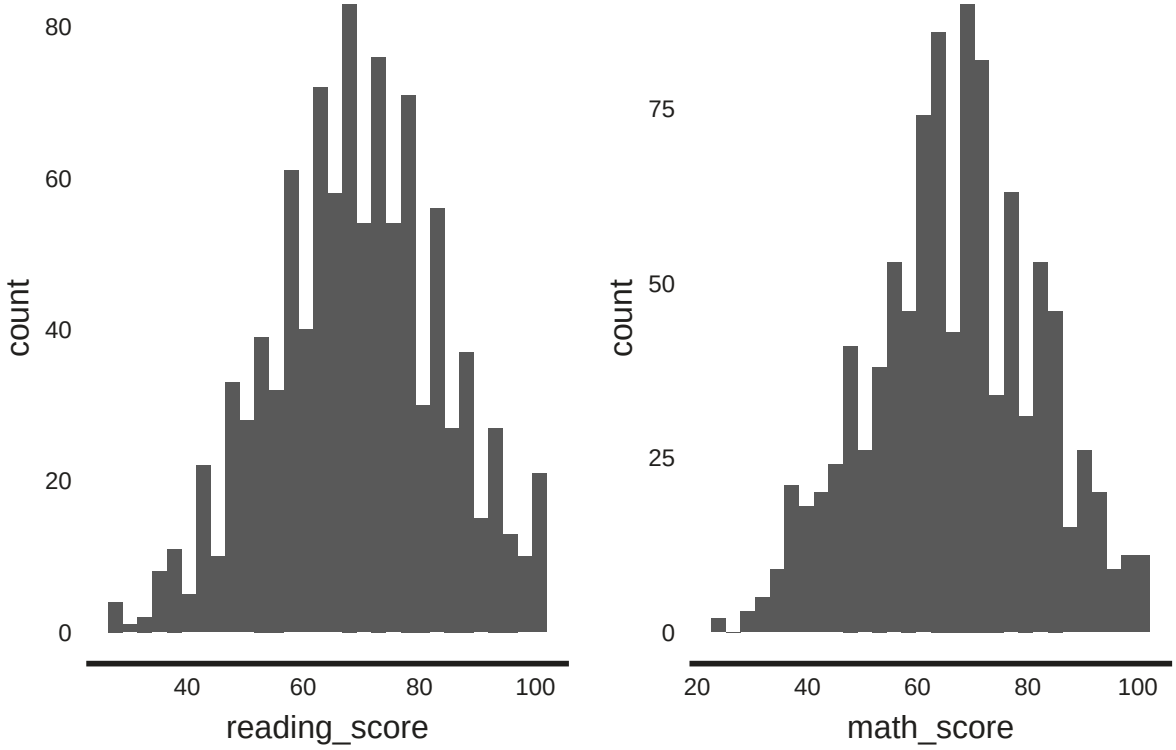
In this case, we look at a situation where we give two treatments (test preparation and lunch) to two groups of students, and we are interested in the math score and reading score of the students. In that case, the math score and reading score of students are two dependent variables, and our hypothesis is that both together are affected by the difference in treatment (test preparation vs provision of lunch). A multivariate analysis of variance could be used to test this hypothesis (Moser and Stevens 1992).

The MANOVA test can be used in certain conditions:

- The dependent variables should be normally distribute within groups (multivariate normality).
- Homogeneity of variances across the range of predictors.
- Linearity between all pairs of dependent variables, all pairs of covariates, and all dependent variable-covariate pairs in each cell

We test for each condition:

The data appears normally distributed among reading and math scores.



To test for the homogeneity of variance we conduct the F-test for homogeneity of variance.

The statistical hypotheses are:

Null hypothesis (H0): the variances of the two groups are equal.

Alternative hypothesis (H1): the variances are different.

We use the variance test for homogeneity of variance.

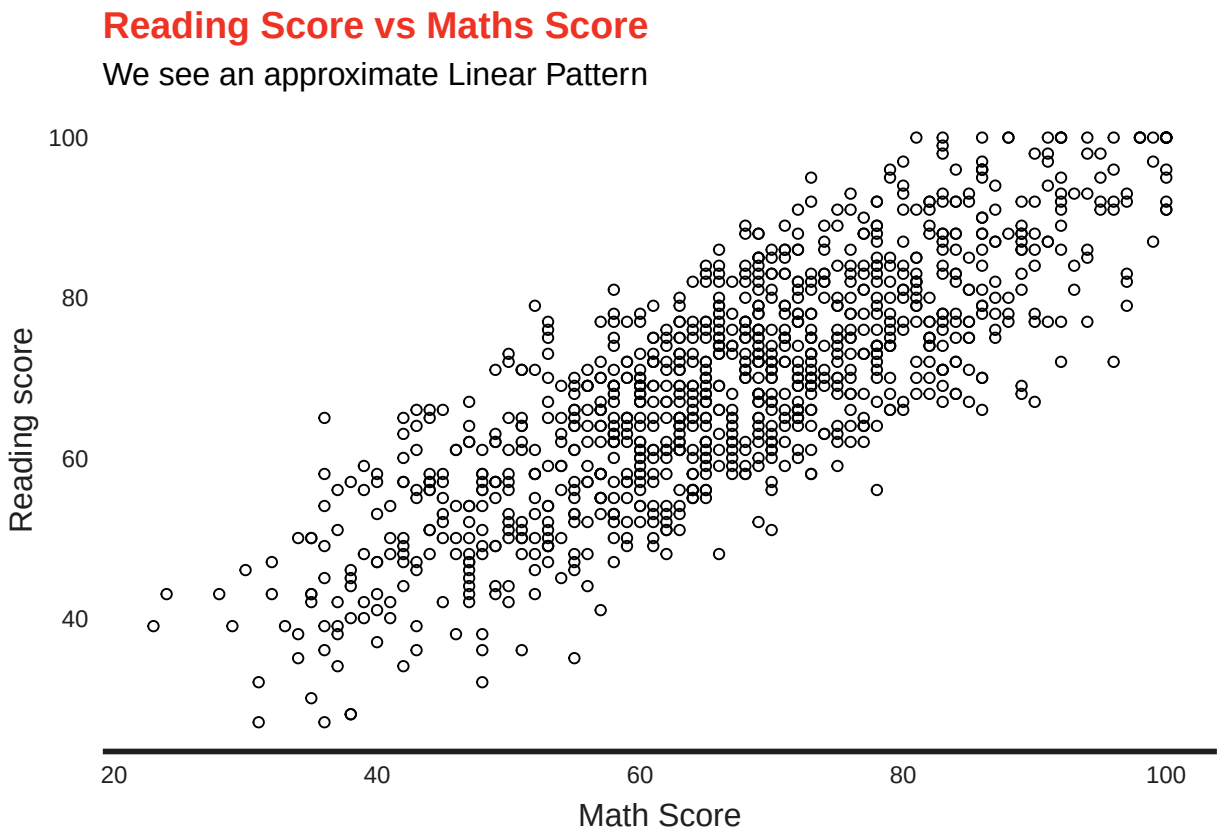
We start with the test whether the two groups- the group that had a test preparation course and the one that did not have equal variance for the maths score. We see that the p-value is 0.5. Hence, we accept the null hypothesis that the two variances are the same.

```
##
##  F test to compare two variances
##
## data:  math_score by test_preparation_course
## F = 1, num df = 363, denom df = 635, p-value = 0.3
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.92 1.33
## sample estimates:
## ratio of variances
##                1.1
```

Next, test whether the two groups- the group that had a test preparation course and the one that did not have equal variance for the reading score. We see that the p-value is 0.5. Hence, we accept the null hypothesis that the two variances are the same.

```
##
##  F test to compare two variances
##
## data:  reading_score by test_preparation_course
## F = 0.9, num df = 363, denom df = 635, p-value = 0.5
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.787 1.134
## sample estimates:
## ratio of variances
##              0.943
```

lastly, we check for approximate linearity between the reading score and maths score. We see an approximate linear pattern from the graph below.



2.5 Conclusion

In this analysis, I have run a series of statistical tests examining the relationship between tests in reading, writing, and maths and a host of factors that are presumed to have a relationship with the former. We find that test preparation is significantly associated with test scores. Gender and reading scores also have a significant relationship with maths scores. The findings have an implication for test preparation.

3 Project 3

3.1 Background

In the context of cognitive studies, understanding the impact of aids on test recall rates is crucial for informing educational practices and cognitive support strategies. This research delves into the comparison of test recall rates between individuals who utilize aids during examinations and those who do not. The motivation behind this inquiry stems from the need to ascertain whether the presence of aids significantly influences recall performance. To address this question, a statistical approach, the Mann-Whitney U test, is employed to rigorously examine any potential differences in recall rates between the two groups. The findings of this investigation bear significance for educational interventions and the design of assessment environments, shedding light on the efficacy of aids in the context of test recall.

3.2 Data

The data is divided into two groups: the retrieval group (R) and the non-retrieval group (N). The retrieval group recalled information without any assistance, whereas the non-retrieval group used aids for recall. In both scenarios, both groups underwent testing after 5 days and then again after 35 days.

retrieval_participant	group	recall_day_5	recall_day_35
A	R	100	25
B	R	100	86
C	R	80	57
D	R	65	46
E	R	90	71
F	R	75	54

3.3 Data Exploration

I start by visualizing the missing data. We see that our data has 6% missing, mostly students that were absent in taking one or both the tests. We shall drop this data during analysis.

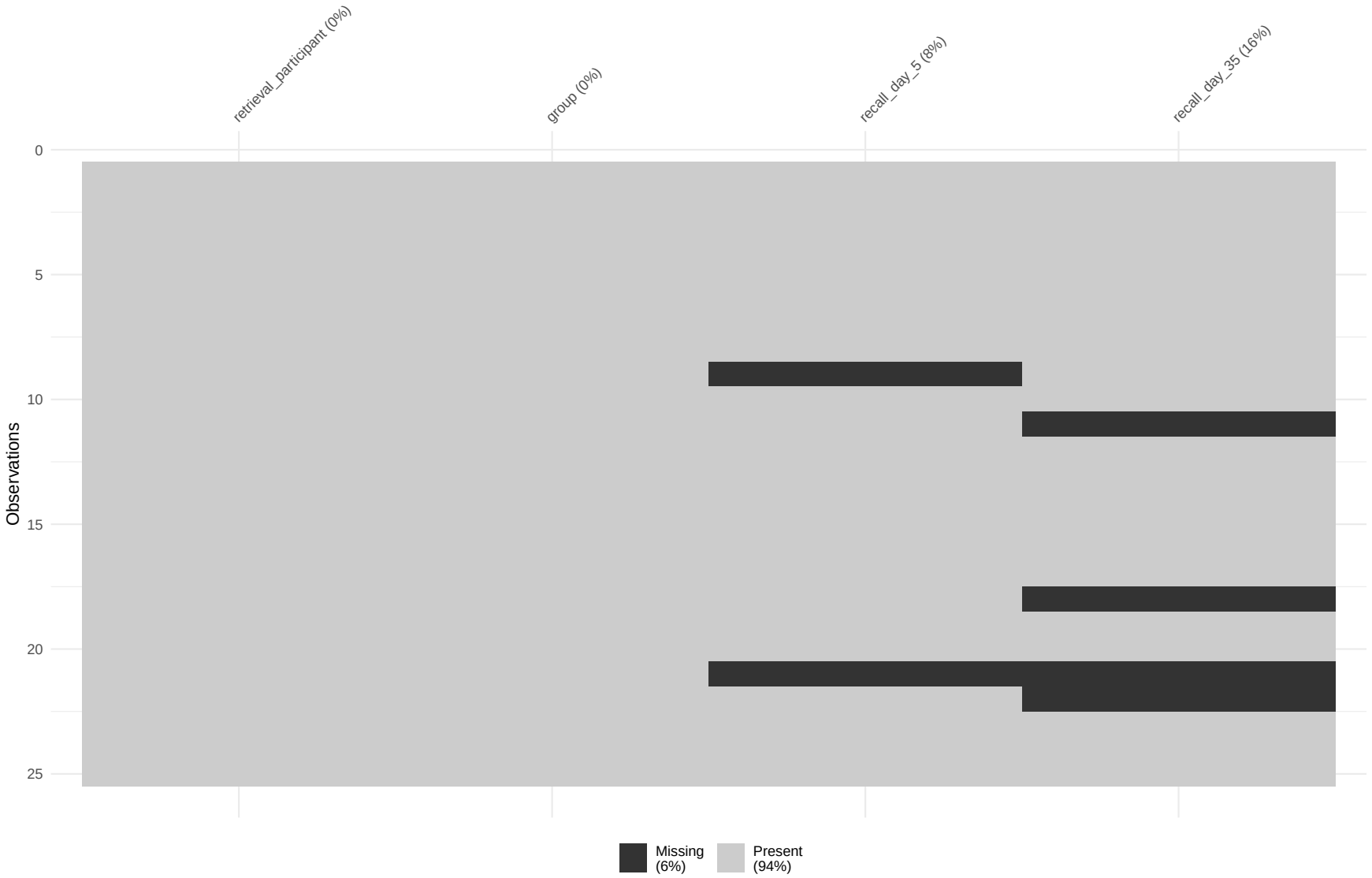
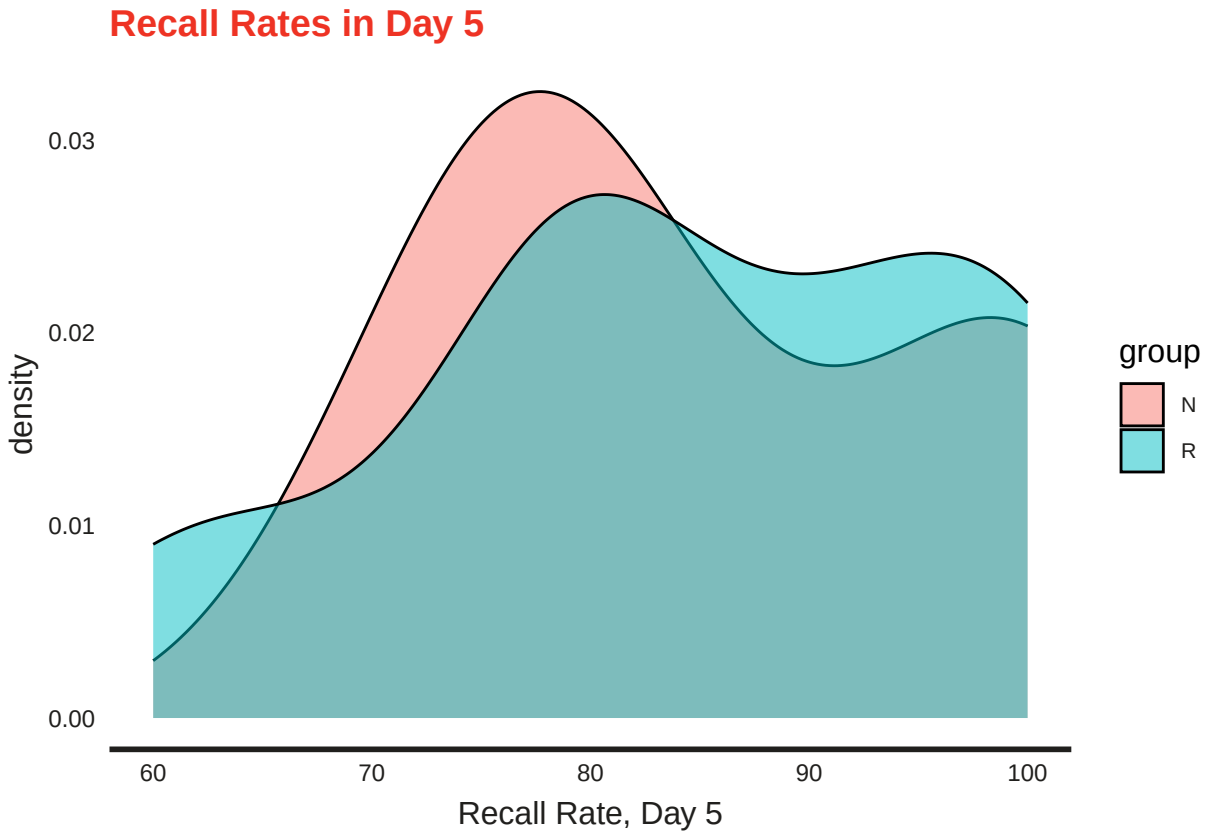


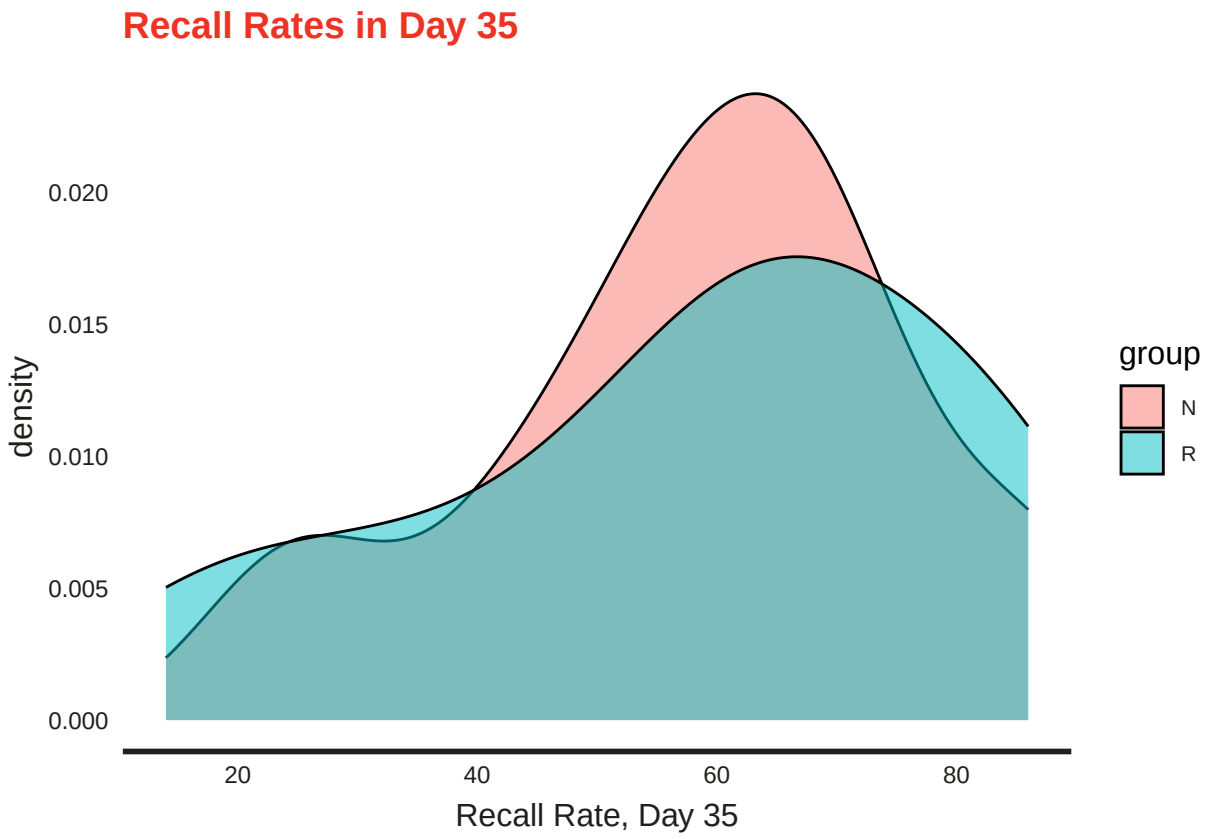
Figure 6: Missing Data

The summary provided outlines the recall rates on day 5 and day 35 for both the retrieval and non-retrieval groups. It is evident that the retrieval group outperforms the non-retrieval group across all measurement metrics. Additionally, it is noteworthy that the retrieval group displays a greater variation (standard deviation) in the recall rates. The question is whether the observed difference is statistically significant.

skim_type	skim_variable	group	n_missing	complete_rate	numeric.mean	numeric.sd	numeric
numeric	recall_day_5	N	1	0.909	85.0	11.5	
numeric	recall_day_5	R	1	0.929	84.2	13.0	
numeric	recall_day_35	N	3	0.727	58.5	18.6	
numeric	recall_day_35	R	1	0.929	57.9	22.5	

We examine the distribution of the recall rates on day 5 and day 35, respectively. We find that the data are not normally distributed.





3.4 Hypothesis Test

Due to the non-normal distribution of the data, the Mann-Whitney U-test, also known as the Wilcoxon rank-sum test, is employed for comparing differences between two independent samples when their distributions aren't normal, and the sample sizes are small ($n < 30$). This nonparametric test serves as an alternative to the two-sample independent t-test.

Assumptions for the Mann-Whitney U Test include having a continuous variable, a non-Normal distribution of data, similar data shapes across groups, independence of the two samples, and a sufficient sample size (typically more than 5 observations in each group). In our case, all conditions are met, as the scores are approximately comparable between the retrieval and non-retrieval groups (McKnight and Najab 2010).

Moving on to the hypothesis test, the null hypothesis (H0) posits no difference in recall rates between the two groups, while the alternative hypothesis (HA) suggests a significant difference. Analyzing recall rates on day 5, the obtained p-value of 1 indicates no significant difference in recall rates between the retrieval and non-retrieval groups.

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  retrieval and non_ret
## W = 264, p-value = 1
## alternative hypothesis: true location shift is not equal to 0
```

We conduct the same test for recall rates in day 35. We also fins no evidence of a difference in recall rates in day 35.

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  retrieval35 and non_ret35
## W = 220, p-value = 1
## alternative hypothesis: true location shift is not equal to 0
```

3.5 Conclusion

In this study, we investigated the disparity in test recall rates between individuals utilizing an aid and those not employing any assistance. Utilizing a statistical analysis method known as the Mann-Whitney U test, we discerned that there is no statistically significant difference between the recall rates of the two groups.

References

Kruschke, John. 2014. *Doing Bayesian Data Analysis: A Tutorial with r, JAGS, and Stan*. Academic Press.

McKnight, Patrick E, and Julius Najab. 2010. "Mann-Whitney u Test." *The Corsini Encyclopedia of Psychology*, 1–1.

Moser, Barry K, and Gary R Stevens. 1992. "Homogeneity of Variance in the Two-Sample Means Test." *The American Statistician* 46 (1): 19–21.

Razali, Nornadiah Mohd, Yap Bee Wah, et al. 2011. "Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests." *Journal of Statistical Modeling and Analytics* 2 (1): 21–33.