Parental Education and Student Success

Evidence from Polytechnic Institute of Poretalegre

Course: Data 602 Statistical Data Analysis

---

1. Purpose

1.1 Domain:

We are conducting a statistical analysis that evaluates the influence parental education has on a child's success rate in terms of graduation status and academic performance.

1.2 Investigation Goal:

To statistically examine the relationship between parental education and student graduation status, while quantifying how much parental education impacts semester grades and academic outcomes.

1.3 Practical Implications:

Parental education often reflects broader socio-economic and cultural capital. A deeper understanding of its influence on graduation outcomes can help universities design targeted support strategies for first-generation college students and those with limited educational backgrounds.

1.4 Population(s) of Interest:

In this statistical analysis our population of interest are students enrolled at the Polytechnic Institute of Portalegre.

1.5 Variables of Interest:

Our analysis will focus on independent variables like mother's qualification, father's qualification, and students curriculum units performance in the first and second semesters. The target variable indicating the graduation status will serve as the dependent variable. For mothers and father qualifications we will stratify the data into 2 subgroups for different levels of education. These subgroups will be parents with a post secondary education or higher and parents with at most high school education.

1.6 Data Collection Method:

The dataset was developed as part of a project aimed at reducing student dropout and failure in higher education, supported by the Portuguese administrative capacity-building program SATDAP. It was later donated to the UCI Machine Learning Repository and Kaggle, where it has been made widely available for research. Each record corresponds to an individual student and includes information collected at enrollment like demographics, socio-economic status, and previous qualifications along with performance during the first and second semesters including grades, credits. The dataset is

designed to classify students at the end of their normal course duration into categories such as dropout, still enrolled, or graduate.

---

2. Data

2.1 Source of Data:

Student Performance (PIP) dataset, accessed via Kaggle (uploaded by user *mikhail1681*), originally from the UCI Machine Learning Repository. Available at: https://www.kaggle.com/datasets/mikhail1681/student-performance-pip

2.2 Data Characteristics:

In this investigation, we analyze the Overall Marks of students, which represents a continuous numerical variable. Additionally, we consider categorical variables such as the Parents' Qualification Status and the Target variable, which indicates whether a student is a dropout or not. These variables together help us understand how academic performance and parental education levels relate to student retention outcomes.

2.2 Permission/Accessibility:

The dataset is publicly available on Kaggle and is licensed under the **Creative Commons Attribution 4.0 International (CC BY 4.0)** license. This license permits sharing and adaptation for academic and research purposes, provided proper credit is given.

2.3 Time Frame of Data Collection:

The data was collected during the 2005-2006 school year from secondary education students in two Portuguese schools. The study that used this data was later published in **2008.**

---

3. Topic(s) to Investigate

3.1 Focus of Statistical Investigation:

The purpose of this statistical investigation is to determine whether students whose parents have higher qualifications perform better academically than those whose parents have lower qualifications. In addition, we are to examine whether parental qualification level is associated with student dropout status.

3.2 Background/Context:

Parents play a central role in a child's behavioural and emotional development. We ask the question if parents have this same influence academically. This study explores the academic influence a parent's prior education has on their child's success rate. It may be thought that parents with higher schooling may better support their child's education by setting higher expectations, better guidance and greater learning resources. Conversely, students that enter college with parents with lower education may not have these advantages which may cause limitations and additional barriers. By quantifying this relationship in our dataset, we can estimate if parental qualification relates to both academic performance and graduation status. This can provide insight in identifying limitations and design support programs for students at this disadvantage.

---

4. Statistical Methods

4.1 Methods:

In our statistical analysis, we conducted four tests to evaluate whether students with more highly educated parents perform better academically than those whose parents have lower qualifications. The dataset was stratified based on parental education. One group includes students with at least one parent holding a university degree or higher, while the other includes students whose highest parental education is high school. In our analysis, parents with a higher qualification are encoded as 1 whereas parents with lower qualifications are encoded as 0. Since the dataset contains no missing values, no additional preprocessing was required. With the data prepared, we proceed to our statistical analysis.

- Method 1: A one-tailed Two Sample t-test is used to compare the two groups at a 5% significance level ($\alpha = 0.05$). This analysis helps assess whether parental education level is associated with higher student performance, providing insight into the potential influence of parental qualifications on academic achievement.

- Method 2: Bootstrap Method: Using bootstrap method to estimate the mean grade difference between the two groups- parents with a higher education and parents with a lower education

- Method 3: Permutation test: A two-tailed permutation test at the 5% significance level ($\alpha = 0.05$) was conducted to compare mean overall grades between students with higher (1) and lower (0) parental qualifications. This non-parametric approach assesses group differences by repeatedly shuffling labels across 4,999 permutations to build an empirical null distribution, providing a robust evaluation without assuming normality.

- Method 4: A Chi-square test of independence was conducted to examine whether there is a significant relationship between parental qualification level and student dropout status at a 5% significance level ($\alpha = 0.05$).

---

5. Results

5.1 Descriptive Statistics and Preliminary Exploration:

Recall that in this investigation, a value of 0 represents students whose parents do not have higher education, while 1 represents students with at least one parent who does. The summary table presents semester-based academic performance and course load (*mean_enrolled*), along with each group's sample size, standard deviation, and range (min/max).

Parents.qualification <int>	n <int>	mean_first_sem_grade <dbl>	sd_first_sem_grade <dbl>	min_first_sem_grade <dbl>	max_first_sem_grade <dbl>
0	3636	10.65589	4.767200	0	18.000
1	788	10.57129	5.184523	0	18.875

mean_second_sem_grade <dbl>	sd_second_sem_grade <dbl>	min_second_sem_grade <dbl>	max_second_sem_grade <dbl>	mean_enrolled <dbl>	sd_enrolled <dbl>
10.24790	5.159856	0	18.57143	6.295930	2.417132
10.14856	5.442378	0	17.69231	6.153553	2.751093

min_second_sem_grade <dbl>	max_second_sem_grade <dbl>	mean_enrolled <dbl>	sd_enrolled <dbl>	min_enrolled <int>	max_enrolled <int>
0	18.57143	6.295930	2.417132	0	26
0	17.69231	6.153553	2.751093	0	21

Figure 1: Summary Statistics of our dataset

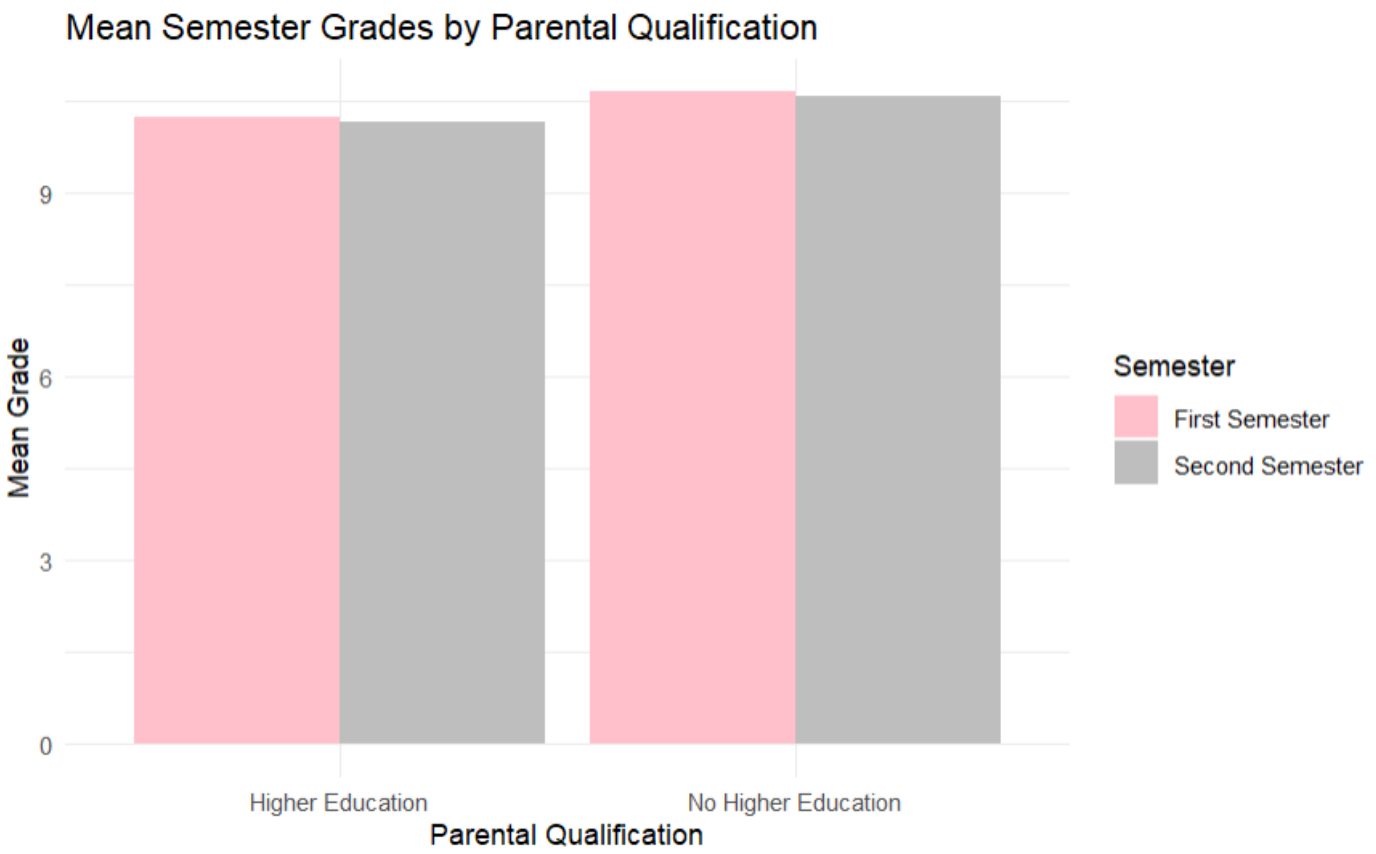


Figure 2: Bar Graph of 1st and 2nd semester grades for students with highly educated parents and less educated parents

The dataset is predominantly composed of students from less educated parental backgrounds (n = 3636), compared to 788 students with higher-educated parents. This means that approximately 82% of our sample comes from the former group. When comparing average grades, both parental groups show nearly identical first-semester means (10.66 vs. 10.57), indicating no meaningful difference in performance. The same pattern holds for second-semester grades, where mean scores remain almost the same. In terms of course enrollment, students with less-educated parents took slightly more courses on average, though both groups enrolled in approximately six units. Hence this classifies as a negligible difference. However we see the standard deviation vary as students with higher parental education indicate that the course load varies. The difference between these groups is minimal, specifically 0.33 units, and this can be considered a normal random variation.

Overall, the descriptive summary suggests that differences between the two groups are small and not practically significant. Further statistical testing will be conducted to explore whether any underlying relationships exist between parental education and student academic outcomes.

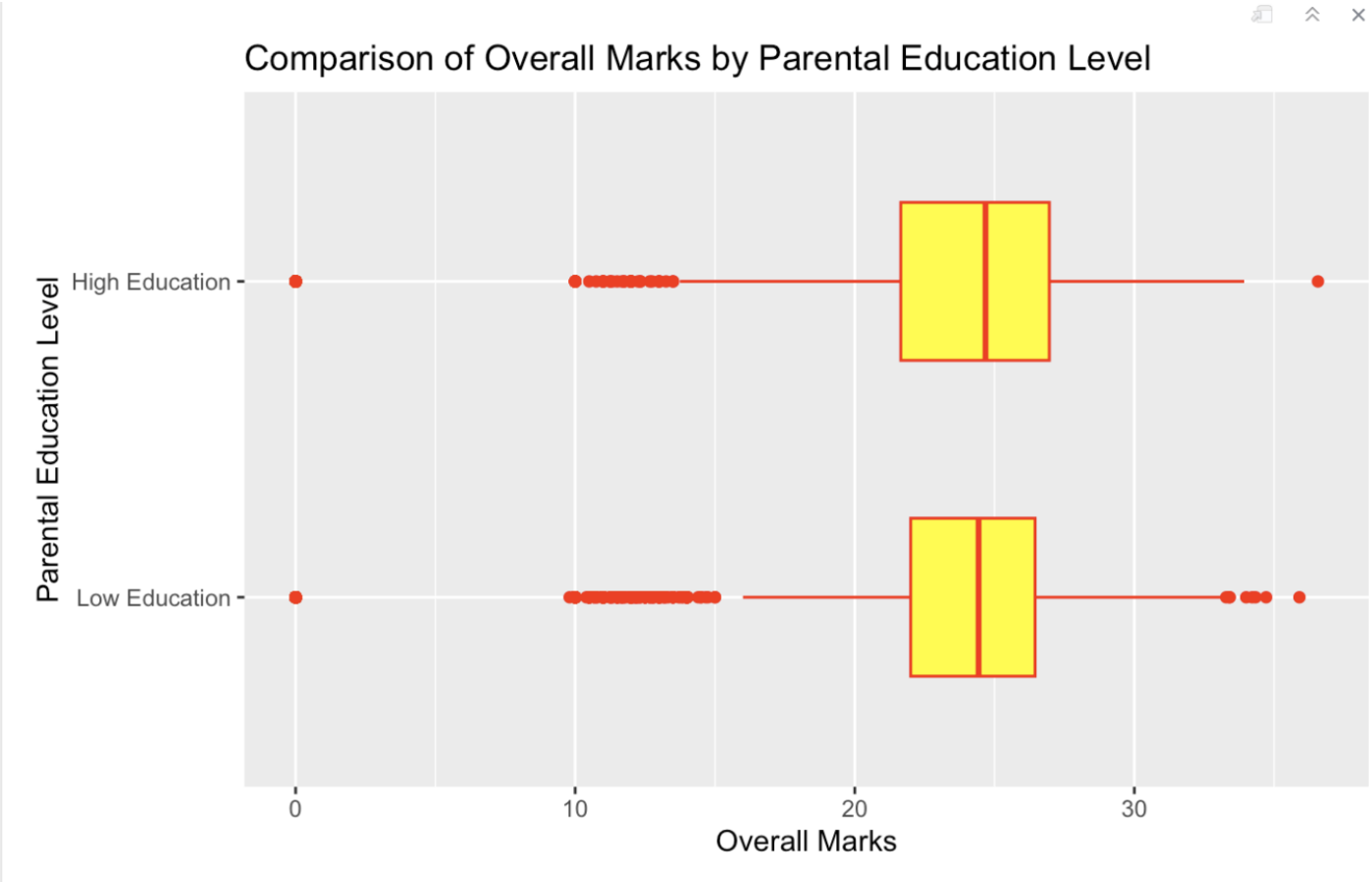


Figure 3: Boxplot of comparing Overall Marks by Parental Education Levels

This boxplot shows the distribution of total student grades (sum of Semester 1 + Semester 2) for two groups, students whose parents have low and high education levels. Statistically, the plot reveals that both groups have very similar medians and interquartile ranges, meaning that the middle 50% of grades overlap almost entirely. The median line for both categories lies near the mid-20s, indicating comparable overall performance. The spread (length of whiskers) and presence of outliers on both ends show that grade variability is roughly equal for both groups. Overall, there is no noticeable shift or separation between the two boxes, suggesting that parental education level does not significantly influence students overall marks in this dataset. Any observed differences appear minor and could be attributed to random variation rather than a systematic effect.

To investigate whether parental qualification level influences student academic performance, we establish the following

hypotheses:

Null hypothesis (H_o):
μ_(Student_marks_Parents_with_HIGH_qualification) ≤ μ_(Student_marks_Parents_with_LOW_qualification)

The mean overall marks of students whose parents have higher qualifications are less than or equal to the mean overall marks of students whose parents have lower qualifications.

Alternative hypothesis (Ha):
μ_(Student_marks_Parents_with_HIGH_qualification) > μ_(Student_marks_Parents_with_LOW_qualification)

The mean overall marks of students whose parents have higher qualifications are greater than the mean overall marks of students whose parents have lower qualifications.

We summarized our observed data on Parent Qualifications using the favstats function, which provides key descriptive statistics (such as mean, median, standard deviation, and range) to support our analysis.

```{r}
favstats( Overall_marks ~ Parents_Qualification, data =students_required_data )
```

Description: df [2 × 10]

Parents_Qualification <chr>	min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>
0	0	22.00000	24.42857	26.45060	35.90476	20.90379	9.513695	3636
1	0	21.64583	24.67669	26.95952	36.56731	20.71984	10.191913	788

2 rows | 1-9 of 10 columns

Figure 4: Summary of stats for Parent Qualification

Upon taking the average of both semesters, the descriptive statistics indicate that students whose parents have lower qualifications achieved overall marks ranging from 0 to 35.90, with a mean of 20.90 and a standard deviation of 9.51. In comparison, students whose parents have higher qualifications recorded marks ranging from 0 to 36.57, with a mean of 20.72 and a standard deviation of 10.19. Both groups exhibit similar quartile values (Q1, median, and Q3) and comparable variability in their scores, suggesting that the distributions of academic performance are generally alike across the two parental education groups.

5.2 T Test:

Given that the dataset consists of two independent groups (students categorized by parental qualification level) and the variable of interest, Overall Marks, is continuous, a *t*-test is the most appropriate statistical method. Specifically, Welch's *t*-test is employed because the two groups have unequal sample sizes (n = 3636 vs n = 788). This test provides a robust comparison of group means without assuming equal population variances, making it well-suited for evaluating whether students with parents of higher qualifications perform significantly better academically than those whose parents have lower qualifications.

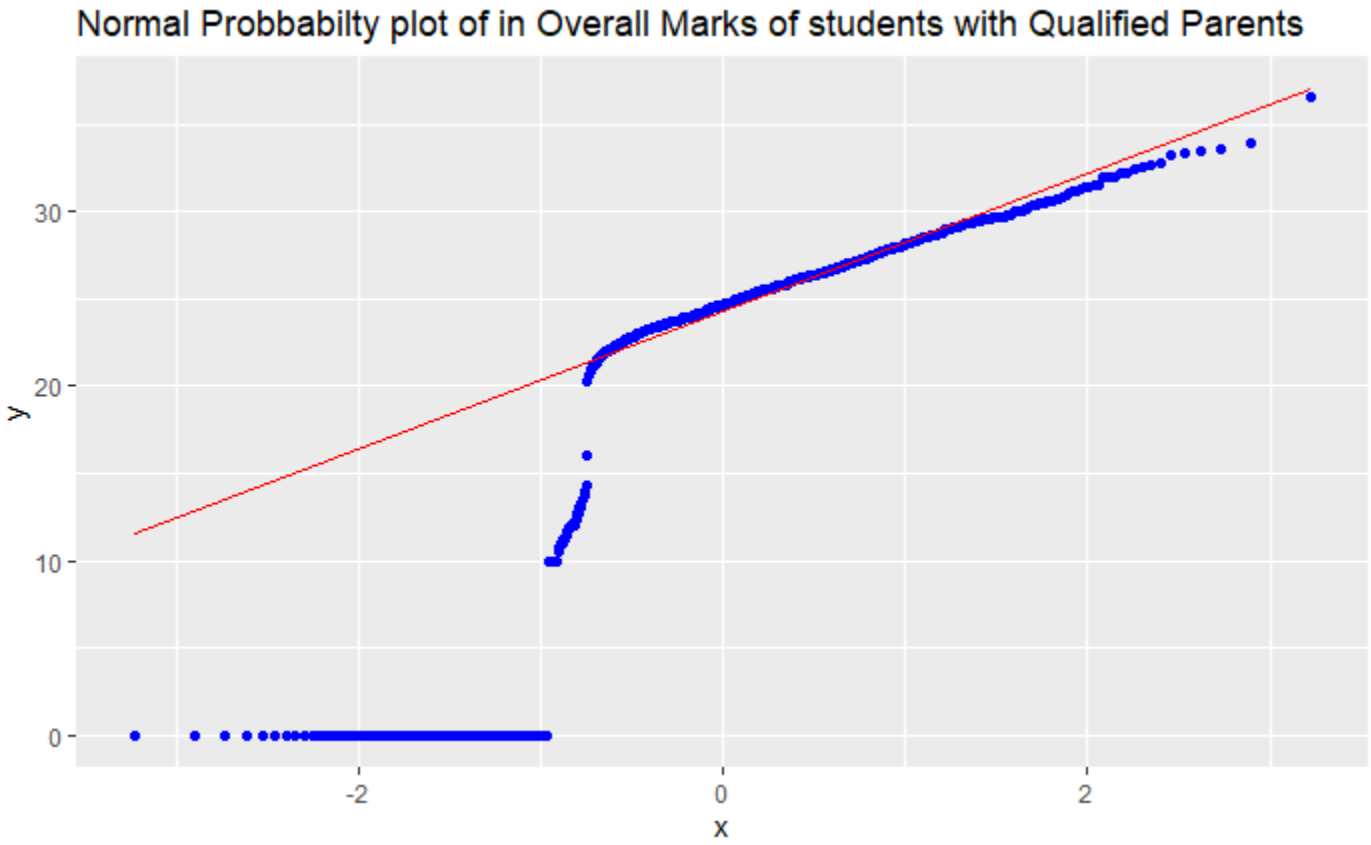


Figure 5: Plot of Overall Marks of Students with Qualified Parents

The normal probability plot displays the distribution of overall marks for students whose parents are highly qualified (coded as 1) against a theoretical normal distribution. In this plot, the blue points represent the actual data values, and the red line represents the expected values under perfect normality. Most points approximately follow the line in the middle range, but noticeable deviations occur at both ends, particularly at the lower tail, where many data points fall below the line. This indicates the presence of non-normality, likely due to several students scoring very low or zero marks. Overall, while the central portion of the data roughly follows a normal trend, the departures at the tails suggest the distribution is slightly skewed and not perfectly normal. However, given the large sample size (n = 788), the t-test remains robust to this mild violation of normality, and the test results are still considered valid.

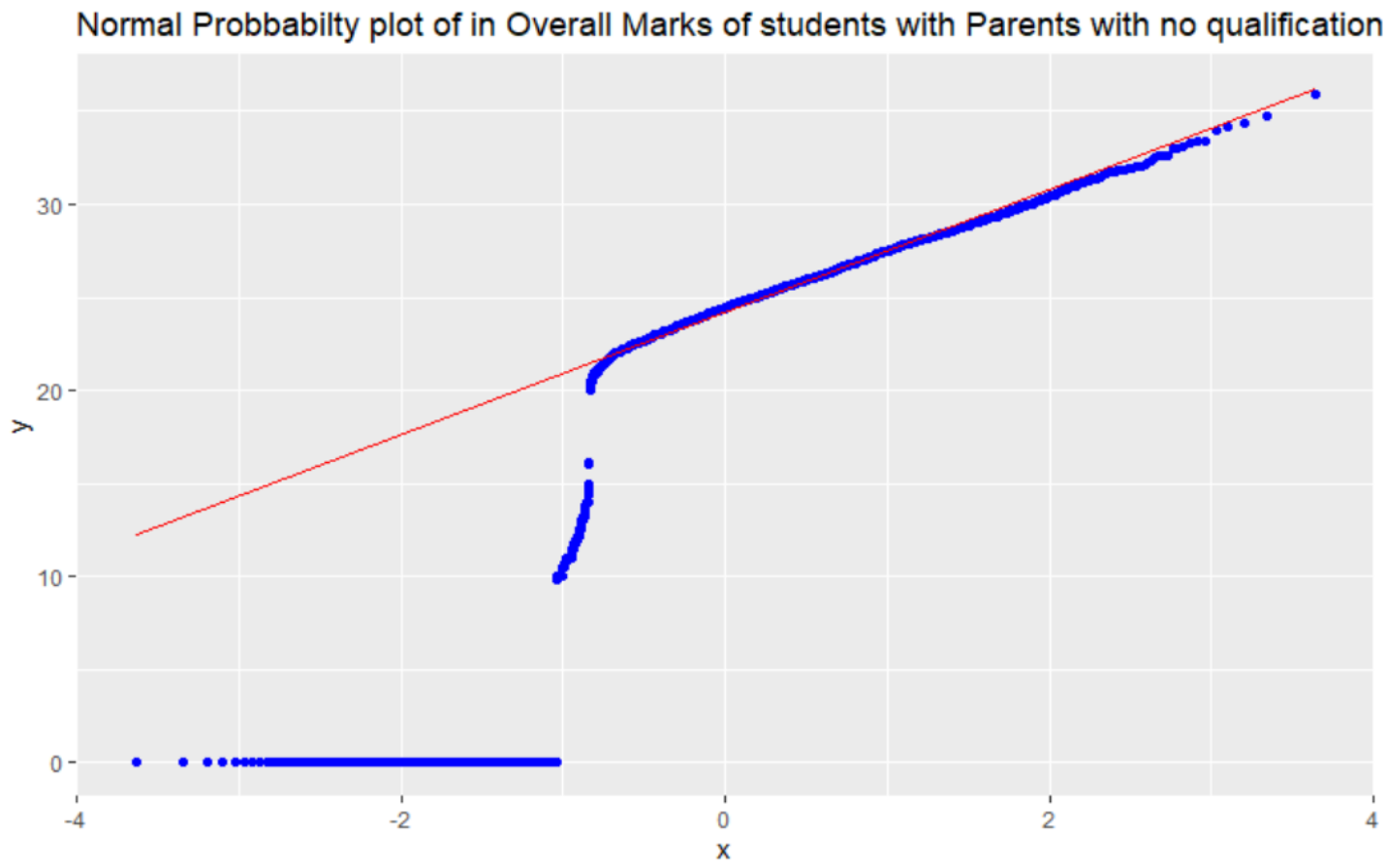


Figure 6: Plot of Overall Marks of Students with Parents with no qualifications

The normal probability plot shows the distribution of overall marks for students with parents who have no formal qualifications compared to a theoretical normal distribution. In this plot, the blue points represent the actual data values, while the red line represents the expected normal values. The data points deviate noticeably from the red line, especially in the lower tail, where many scores are concentrated near zero. This pattern indicates a departure from normality, suggesting that a large number of students in this group performed poorly compared to the rest. Despite this deviation, the large sample size (n = 3636) helps minimize the impact of non-normality on parametric tests. Therefore, while the data are slightly skewed and not perfectly normal, the t-test remains an appropriate and robust choice for comparing means between the two parental qualification groups.

Although the normal probability plots indicate some deviations from normality in the distribution of students' overall marks, particularly at the lower end of the scale, the large sample sizes in both groups (n = 3,636 for parents with no qualifications and n = 788 for parents with qualifications) allow us to proceed with the independent samples *t*-test. According to the Central Limit Theorem, when the sample size is sufficiently large, the sampling distribution of the mean tends to approximate normality regardless of the shape of the original data distribution. Therefore, the *t*-test remains a robust and appropriate method for comparing the mean performance between the two groups, even if the raw data are not perfectly normally distributed.

```r
t.test(group1, group0, alternative = "greater" )
```

```
        Welch Two Sample t-test

data:  group1 and group0
t = -0.46467, df = 1103.8, p-value = 0.6789
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.835645       Inf
sample estimates:
mean of x mean of y
 20.71984  20.90379
```

Figure 7: Statistics results of T Test in R

A one-tailed Welch's two-sample *t*-test was conducted to examine whether students whose parents have higher qualifications achieve higher overall marks than those whose parents have lower qualifications. The test yielded a *t*-statistic of -0.4647 with approximately 1,103.8 degrees of freedom and a *p*-value of 0.6789.

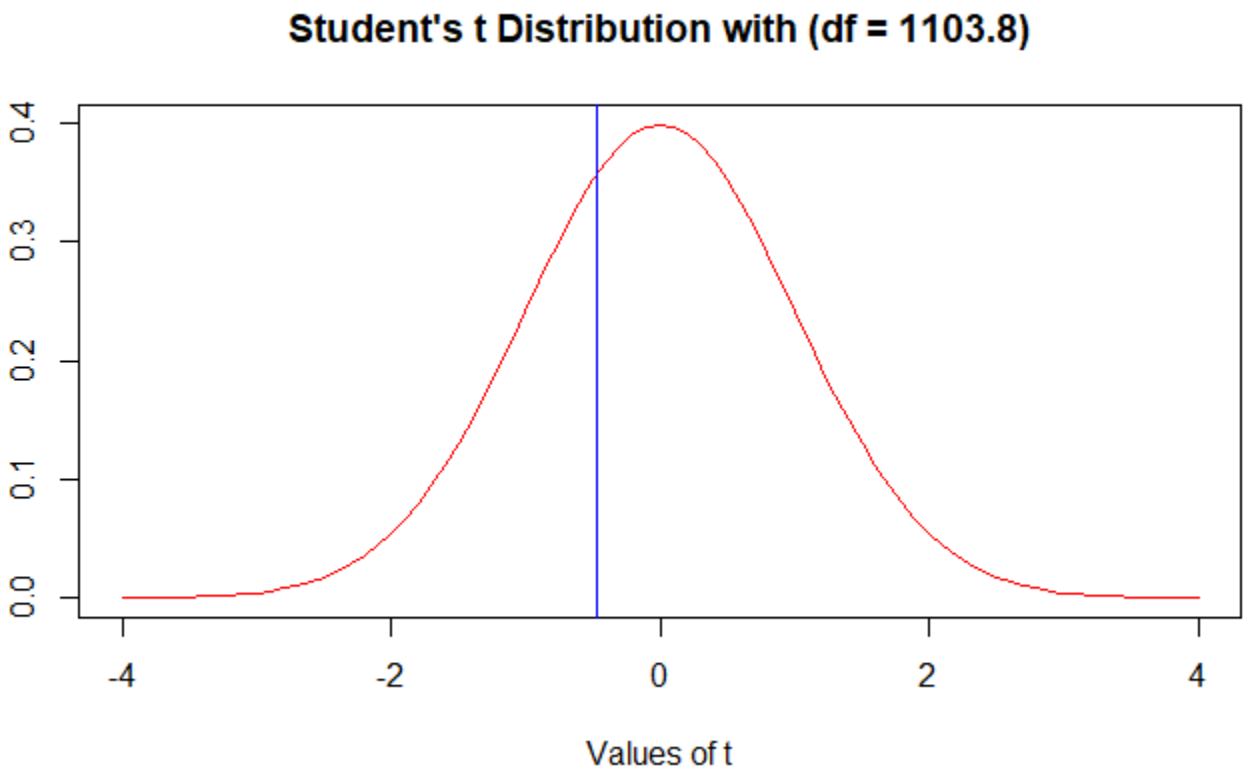


Figure 8: Plot of T Distribution

The t-distribution plot illustrates the sampling distribution of the test statistic with 1103.8 degrees of freedom. The blue vertical line represents the observed test statistic (t = -0.46), which lies very close to the center of the distribution, around zero. This indicates that the observed difference between the two groups means is minimal. The visual clearly signifies that the difference in overall marks between students with higher- and lower-qualified parents is not statistically

significant.

A *p*-value of 0.6789 indicates that, if the null hypothesis were true (students with highly qualified parents do not perform better than those with lower-qualified parents), there would be a 67.89% chance of observing a difference in sample means as extreme or more extreme than the one obtained, in the direction specified by the alternative hypothesis. Since this probability is high, the observed data are consistent with the null hypothesis.

Therefore with 5% significance, we fail to reject the null hypothesis (H_o). This suggests there is no statistically significant evidence to conclude that students with parents of higher qualifications perform better academically than those with lower parental qualifications. In summary, based on this test, the data do not support the claim that higher parental qualifications are associated with higher student marks.

5.3 Bootstrap Estimation:

Unlike classical methods such as the t test, the bootstrap method is distribution free. This means it does not rely on theoretical assumptions about the population distribution regarding normality or equal variances. The Bootstrap Estimation was used to estimate the mean grade difference between two groups: students whose parents have higher education and those whose parents have lower education. The bootstrap estimate of the mean difference was -0.089, indicating that on average students with more highly educated parents scored approximately 0.09 points lower than those with less educated parents. However, this difference is minimal and can be considered negligible in practical terms.

The 95% bootstrap confidence interval, ranging from -0.476 to 0.296, represents the plausible range of mean differences between the two groups. Since the interval includes both positive and negative values, the difference could favor either group. Moreover, because 0 lies within the interval, we cannot conclude that there is a statistically significant difference in mean grades between students based on parental education level.

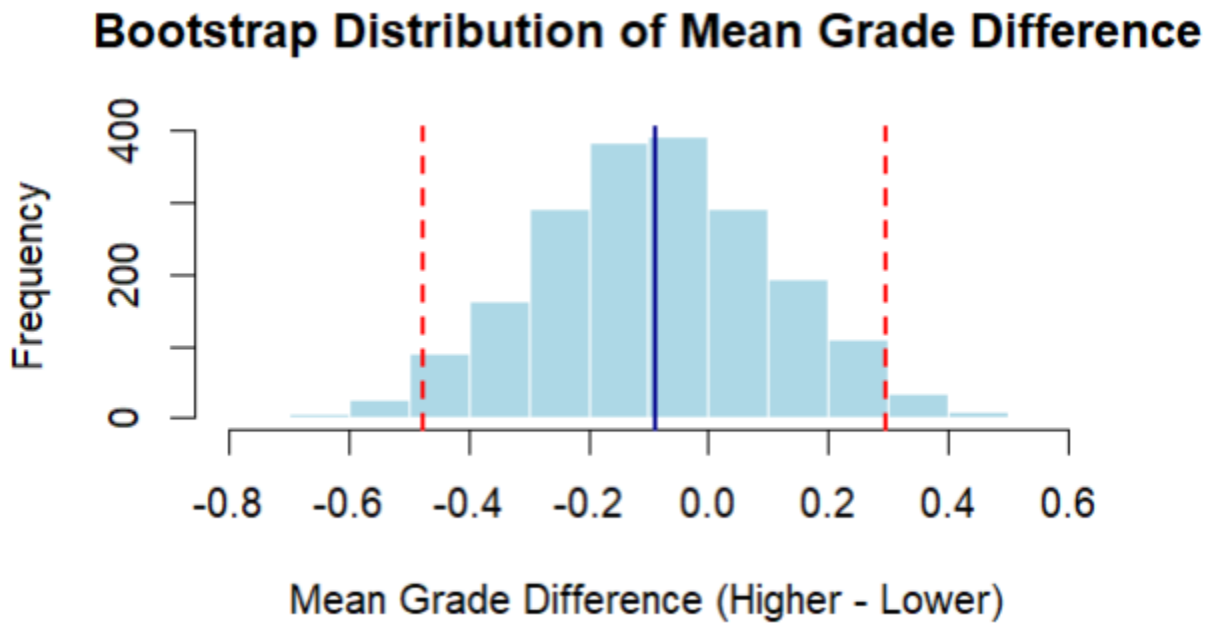


Figure 9: Histogram of Bootstrap Distribution

This histogram illustrates the bootstrap distribution of the mean grade difference between students with highly educated parents and those with less educated parents. The central blue line represents the bootstrap mean difference, approximately -0.09, while the distribution itself is bell-shaped and centered near zero. This indicates that on average there is little to no consistent difference in mean grades between the two groups. The dashed red lines mark the 95% confidence interval bounds at approximately -0.48 and 0.30.

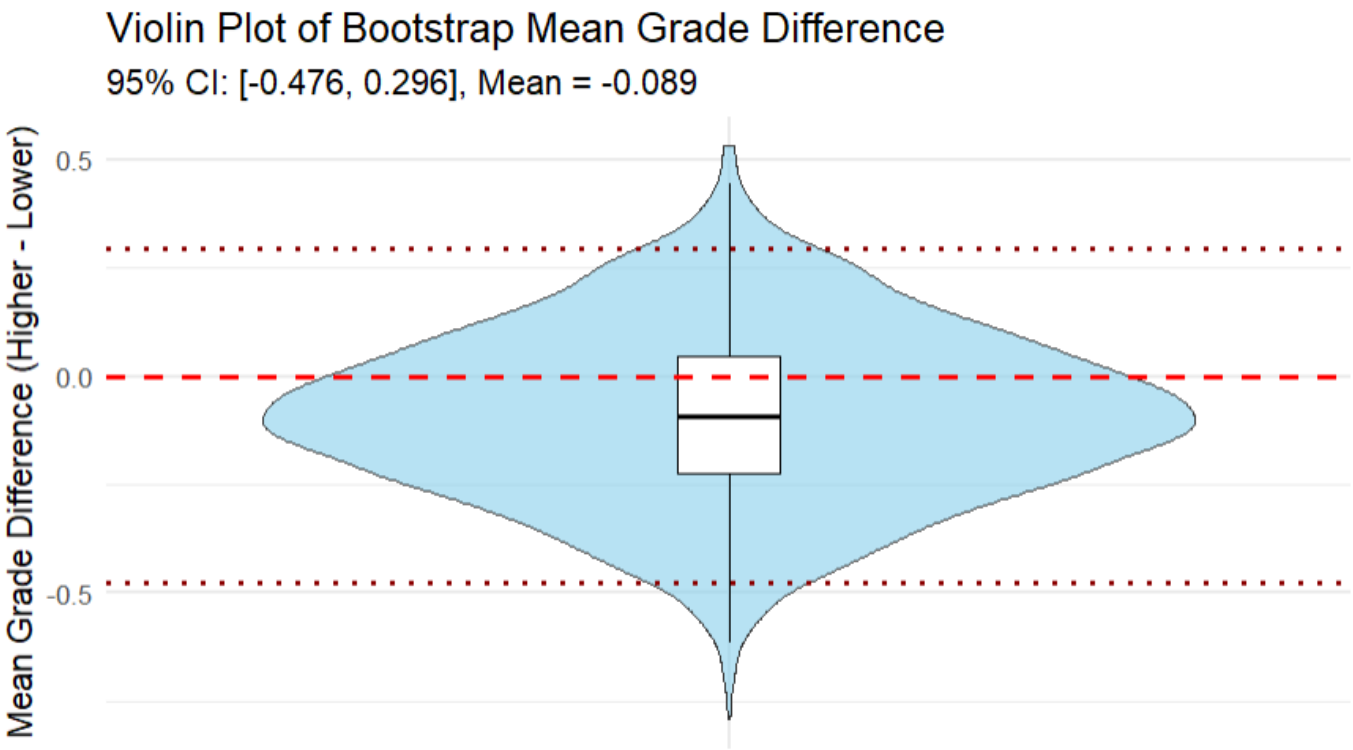


Figure 10: Violin Plot of Bootstrap Mean Grade Difference

This violin plot visualizes the bootstrap distribution of the mean grade difference between students whose parents have a higher education and whose parents do not. The horizontal line showcases the distribution of 2000 resampled mean differences. The shape is roughly symmetric and centered at zero implying that the average difference fluctuates around zero. The white box in the middle indicates the middle of the bootstrap samples with a black line showing the median. The dashed horizontal line at 0 marks the point of no difference. This implies that on average, both groups perform about identically. The two dotted red at the top and bottom represent the 95% bootstrap confidence interval which ranges from approximately -0.48 to 0.30.

5.4 Permutation Test:

Furthermore, a right-tailed permutation test was conducted to assess whether students whose parents have higher educational qualifications achieve higher overall grades than those with lower qualifications.

Null hypothesis (H_0):
$\mu\_(Student\_marks\_Parents\_with\_HIGH\_qualification) \leq \mu\_(Student\_marks\_Parents\_with\_LOW\_qualification)$

The mean overall marks of students whose parents have higher qualifications are less than or equal to the mean overall marks of students whose parents have lower qualifications.

Alternative hypothesis (Ha):
μ_(Student_marks_Parents_with_HIGH_qualification) > μ_(Student_marks_Parents_with_LOW_qualification)

The mean overall marks of students whose parents have higher qualifications are greater than the mean overall marks of students whose parents have lower qualifications.

The test statistic, defined as the difference in mean overall grades was −0.1839, indicating slightly lower average scores for the high-qualification group.

```r
{r}
#Carry out permutation test (5000 permutations)
N = 4999
perm.outcome1 = numeric(N)
n_total <- nrow(student_profile_data)
n_high  <- sum(student_profile_data$`Parents.qualification` == 1)

for(i in 1:N)
{
  index = sample(n_total, n_high, replace = FALSE)
  perm.outcome1[i] =
    mean(student_profile_data$overall_marks[index]) -
    mean(student_profile_data$overall_marks[-index])
}
```

To evaluate the significance of this difference under the assumption that parental education has no effect, 4,999 random label permutations were performed while preserving group sizes. The resulting p-value of 0.692 indicates a 69.2% probability of observing a difference as large or larger (in the direction of the alternative hypothesis) if the null hypothesis were true.

Since this value exceeds the 5% significance level ($\alpha = 0.05$), we fail to reject the null hypothesis (H_o). Therefore, there is no statistically significant evidence that students with higher-qualified parents perform better academically. This suggests that, within this dataset, parental education level does not have a meaningful impact on overall student grades.

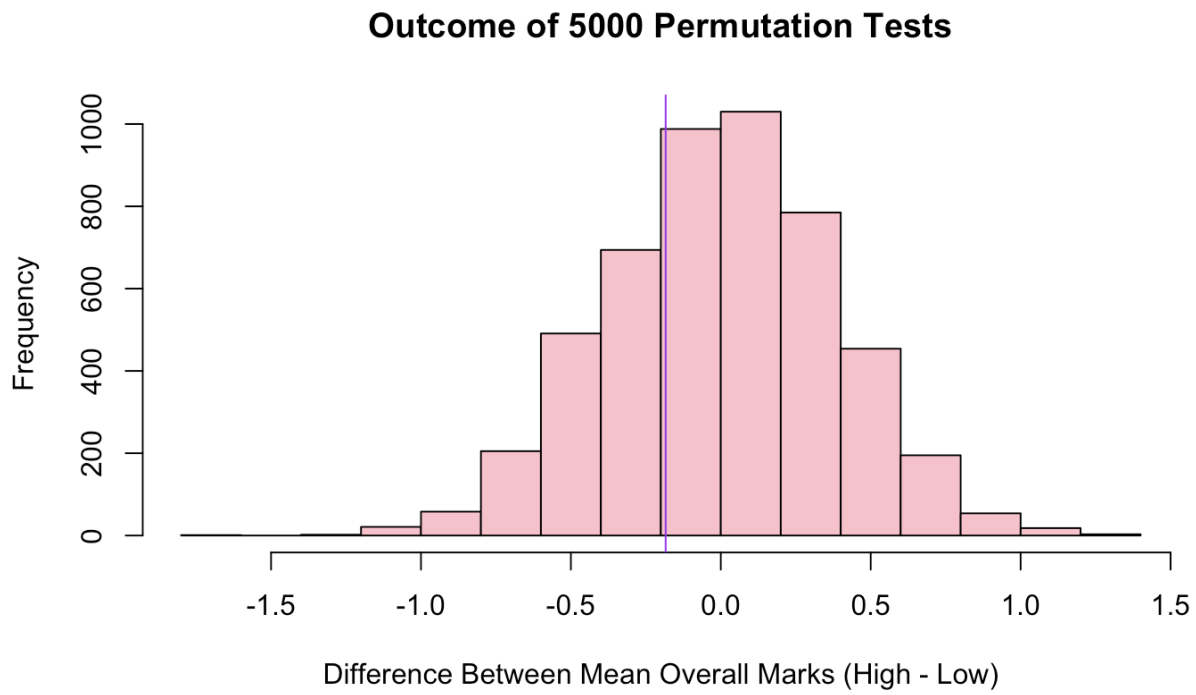


Figure 11: Histograms of Permutation distribution

This histogram displays the permutation distribution of the mean difference in overall marks between students with higher and lower parental qualifications. The distribution is centered around zero, representing the differences expected if parental education has no real effect on student performance. The purple line indicates the observed mean difference (−0.183), which falls well within the main body of the distribution suggesting that such a result is typical under random chance. The high p-value (0.692) further supports this conclusion, indicating no statistically significant evidence that parental qualification level influences overall grades. Thus, the observed variation is consistent with random noise rather than a genuine effect.

5.5 Chi-Square Test of Independence:

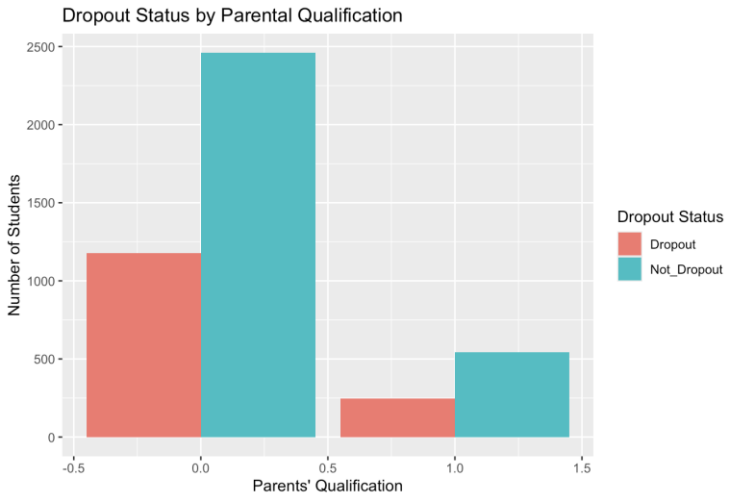


Figure 12: Distribution of student dropout status by parental qualification.

The bar chart displays the relationship between parents' qualification level (low = 0, high = 1) and students' dropout

status. It shows that most students come from the low-qualification group, indicating sample imbalance. Within both groups, the Not Dropout category consistently exceeds the Dropout category, though the gap between them appears similar across qualification levels.

To investigate whether parental qualification level influences student dropout we performed a Chi- Square Test of Independence. Such a test is ideal when determining statistical significance between two categorical variables. However, we must make some assumptions before proceeding. We assume independence of observations, which implies that a student's dropout status does not influence other students. In this dataset, we notice that there is no paired data or repeated data hence this assumption is met. In addition, this method assumes that the expected frequency of each cell in the contingency table is 5 or greater.

Upon the satisfaction of these assumptions, the following is our hypotheses:

Null Hypothesis (H_0):
Parental qualification level and student dropout status are independent.
(There is no association between parents education level and whether a student drops out)

Alternative Hypothesis (H_1):
Parental qualification level and student dropout status are not independent.
(There is an association between parents education level and whether a student drops out)

The test was conducted at a 5% significance level ($\alpha = 0.05$). If the resulting *p*-value is less than 0.05, we would reject the null hypothesis and conclude that parental qualification has a statistically significant association with student dropout status.

```
# Create contingency table
table_parents = table(df$`Parents.qualification`, df$Dropout_status)
table_parents
```

```
  Dropout Not_Dropout
0    1176        2460
1     245         543
```

```
# Perform Chi-square test
chisq.test(table_parents)
```

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  table_parents
X-squared = 0.40986, df = 1, p-value = 0.522
```

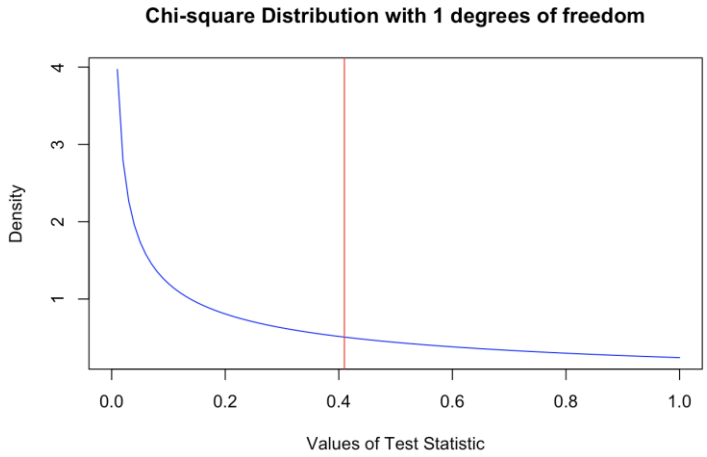


Figure 13: Chi-square distribution with 1 degree of freedom.

The graph shows the Chi-square distribution with 1 degree of freedom. The blue curve represents how the Chi-square test statistic is expected to behave under the null hypothesis. The horizontal axis shows possible values of the test statistic, and the vertical axis represents their corresponding probability density. The red vertical line marks the observed Chi-square value from our test.

Based on our analysis using R code, the Chi-Square Test of Independence produced a test statistic of $\chi^2 = 0.41$ with 1 degree of freedom and a corresponding p-value of 0.522. This value exceeds our significance level of 0.05. Therefore, we fail to reject the null hypothesis. These results indicate that there is no statistically significant association between parental qualification level and student dropout status. In other words, within this dataset, whether or not a student's parents pursued higher education does not appear to influence the student's likelihood of dropping out.

---

6. Conclusion and future work

6.1 Interpretation of Results

The statistical analyses conducted in this study provide consistent results that parental qualification does not have a significant impact on students academic performance or likeliness to dropout. Our t-test revealed a p value of 0.6789, exceeding our 0.05 significance threshold, resulting in the rejection of our null hypothesis. This indicates that there is no meaningful difference in mean grades between students with higher educated parents and those with lower qualified parents. To further validate this finding, the bootstrap and permutation tests were also performed. Both methods produced consistent results with the *t*-test, reinforcing the conclusion that any observed difference in mean scores is likely due to random variation rather than a true effect of parental qualification. The Chi-square test was conducted to examine the relationship between two categorical variables: the parental qualification and student dropout status. Through our test statistic and p values we were able to reject the null hypothesis. The test results showed no significant relationship between the two variables, confirming that dropout status is independent of parental qualification.

While the t-test provided a formal parametric inference, the bootstrap offered a visual and intuitive estimate of uncertainty, and the permutation test confirmed robustness without relying on distributional assumptions. Overall, the results suggest that parental qualification does not have a significant impact on students' overall academic performance in this dataset.

6.2 Limitations

Several limitations may have influenced the findings of this study. First, the parental qualification variable was represented in a binary format (high or low), which may oversimplify the range of educational backgrounds and reduce variability in the data. Second, the study relied on a limited number of variables, focusing mainly on parental qualification and student grades, while other factors such as socioeconomic status, school quality, motivation, and learning environment were not considered. Finally, the analysis assumed independence and approximate normality of the data, which might not hold perfectly in real-world educational data. The permutation test assumes that, under the null hypothesis, all observations are exchangeable between groups. If there are hidden factors or systematic differences socioeconomic background, school quality) that make the groups inherently different, this assumption is violated, and the p-value may not be fully reliable.

6.3 Recommendations and Future Directions

Future research could address these limitations by using more detailed classifications of parental education levels (e.g., high school, undergraduate, postgraduate) and including additional variables such as family income, study habits, and school type to better explain academic performance. Applying advanced statistical models or machine learning techniques could also help capture complex relationships between parental factors and student outcomes. Expanding the sample size and diversity of the dataset would improve generalizability, while longitudinal studies could reveal how parental education influences academic progress over time. Finally, incorporating qualitative data, such as interviews or surveys, could provide deeper insights into how parental support and educational background shape student success.

---

7. References

[List all sources cited in your report, including datasets, articles, books, and websites.]

- Mikhail1681. (n.d.). Student performance (PIP). Kaggle - https://www.kaggle.com/datasets/mikhail1681/student-performance-pip
- Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. UCI Machine Learning Repository - https://archive.ics.uci.edu/dataset/320/student+performance
- Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M. T., & Realinho, V. (2021). Predict Students' Dropout and Academic Success (PIP Dataset). UCI Machine Learning Repository - https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success
- Statistics By Jim. (n.d.). Chi-square test of independence and an example. Retrieved from - https://statisticsbyjim.com/hypothesis-testing/chi-square-test-independence-example/
- Fritz, C., & MacKinnon, D. P. (2007). Confidence Intervals for Effect Sizes: Applying Bootstrap Resampling. Practical Assessment, Research & Evaluation, 12(5), 1-18 - https://files.eric.ed.gov/fulltext/EJ1094191.pdf

---

8. Appendix (if applicable)

To calculate the overall academic performance of each student, a new column named overall_marks was created in the dataset. This column was computed by summing the grades obtained by each student in the first and second semesters using the mutate() function from the dplyr package. This approach allowed us to consolidate semester-wise performance into a single measure, overall_marks, which represents the total marks obtained by a student across both semesters. The resulting overall_marks column was then used in subsequent analyses to compare academic performance between different groups of students.

```r
#creating a new column with the over all marks from first semester and second semester
student_profile_data = student_profile_data %>%
  mutate( overall_marks = Curricular.units.1st.sem..grade. +Curricular.units.2nd.sem..grade.)

student_profile_data$overall_marks
```

```
  [1]  0.00000 27.66667  0.00000 25.82857 25.33333 23.35714 27.64500  0.00000 28.01786 24.90000
 [11] 26.53333 26.42857  0.00000 21.57143 25.25000 13.20000 23.00000 27.85125 24.75000 25.16667
 [21]  0.00000 22.86250 25.71429 25.66071 27.41095 22.60000 24.66071 25.66667 24.33333 26.65000
 [31] 22.42857 26.23333 24.83333 27.04167 25.82857 10.00000  0.00000 22.00000 22.60000 23.66667
 [41] 12.75000 30.60000 28.27086 26.80000  0.00000 25.83333 24.86667 27.28810 24.78571 25.16667
 [51] 27.83333 23.66667 25.00000 30.51429 21.60000 26.63944  0.00000 25.90000 29.87000  0.00000
 [61] 25.35476 24.00000  0.00000 21.40000  0.00000 22.90000  0.00000 25.00000 21.95000 22.80000
 [71]  0.00000 25.72500  0.00000 23.66667 24.66667 21.19048 21.50000 27.00000 29.89841 23.66667
 [81] 24.40000  0.00000 25.46667 23.33333 22.00000 26.16000 22.10000 21.86667 27.80000 10.00000
 [91]  0.00000 27.50000 12.00000 12.66667 24.20833 27.25000 27.00000 28.22778 28.00000 11.96000
[101] 26.27500  0.00000  0.00000 28.86235  0.00000 25.27273  0.00000 27.45000 25.56667 24.66667
[111] 28.85714 24.00000 26.20000 25.23810 25.03333 24.61667 20.75000 21.00000 24.54286 26.60000
[121] 22.00000 22.05000  0.00000 27.00000 26.36667 12.10000 27.94444 26.20000 26.66667 28.47825
[131] 26.98036 25.33333 27.00000 14.00000 23.38095 24.50000 23.70000 29.66667 26.08086 34.71250
[141] 25.33333 24.66667  0.00000 22.83333 25.93625 22.80000  0.00000 22.86667 27.54167 28.04190
[151]  0.00000 25.80000 23.16667 25.13333 22.66667 23.50000 22.40000 24.10714 22.83333 23.60000
[161] 26.16667 25.61667 30.27375 22.50000 28.40000 25.50000  0.00000 22.66667 28.51429 23.33333
```

To focus the analysis on the relevant variables, a new dataset students_required_data was created, containing only the students' first and second semester grades, overall marks, and parental qualification. The rename() function was used to give these columns more intuitive names: semester1, semester2, Overall_marks, and Parents_Qualification. Additionally, the Parents_Qualification column was converted to a character type to facilitate grouping and comparison in later analyses. This streamlined dataset ensures that only the necessary variables are used for statistical testing and visualization:

```r
#taking required data
students_required_data = data.frame(student_profile_data$Parents.qualification,student_profile_data$Curricula
r.units.1st.sem..grade.,student_profile_data$Curricular.units.2nd.sem..grade.,student_profile_data$overall_ma
rks)

students_required_data = students_required_data %>%
  rename(semester1= student_profile_data.Curricular.units.1st.sem..grade.,  semester2 =
student_profile_data.Curricular.units.2nd.sem..grade., Parents_Qualification =
student_profile_data.Parents.qualification ,  Overall_marks = student_profile_data.overall_marks )

students_required_data$Parents_Qualification = as.character(students_required_data$Parents_Qualification)


students_required_data
```

Description: df [4,424 × 4]

Parents_Qualification <chr>	semester1 <dbl>	semester2 <dbl>	Overall_marks <dbl>
0	0.00000	0.00000	0.00000
1	14.00000	13.66667	27.66667
0	0.00000	0.00000	0.00000
0	13.42857	12.40000	25.82857
0	12.33333	13.00000	25.33333
0	11.85714	11.50000	23.35714
0	13.30000	14.34500	27.64500
0	0.00000	0.00000	0.00000
0	13.87500	14.14286	28.01786
0	11.40000	13.50000	24.90000

1-10 of 4,424 rows Previous 1 2 3 4 5 6 ... 100 Next

The dataset was divided into two groups based on parental qualifications: group0 includes students whose parents have lower qualifications (coded as "0"), and group1 includes students whose parents have higher qualifications (coded as "1"). This separation allows for a direct comparison of overall academic performance between the two groups.

```r
group0 <- students_required_data$Overall_marks[students_required_data$Parents_Qualification == "0"]
group1 <- students_required_data$Overall_marks[students_required_data$Parents_Qualification == "1"]
```

Bootstrap Method:

```r
#loading library
library(ggplot2)

# storing and creating a data frame
grades=data.frame(Parents_qualification = c("No Higher Education", "Higher Education"),First_Sem=c(10.65589,
10.57129),Second_Sem=c(10.24790, 10.14856))

grades_long=data.frame(Parents_qualification = rep(grades$Parents_qualification, each = 2),Semester = rep(c("First semester", "Second
semester"), times = 2),Mean_Grade = c(grades$First_Sem, grades$Second_Sem))

#plotting
ggplot(grades_long, aes(x = Parents_qualification, y = Mean_Grade, fill = Semester)) +geom_bar(stat = "identity", position =
position_dodge()) +labs(title = "Mean Semester Grades by Parental Qualification",x = "Parental Qualification", y = "Mean Grade")
+scale_fill_manual(values = c("pink", "grey"))+theme_minimal()
```

```r
#importing libraries
library(dplyr)

#reading csv file
data <- read.csv("modified.csv")

#creating summary statistics using modified Parental Qualifications column
summary_table <- data %>%
  group_by(Parents.qualification) %>%
  summarise(
    n = n(),

    mean_first_sem_grade = mean(Curricular.units.1st.sem..grade., na.rm = TRUE),
    sd_first_sem_grade   = sd(Curricular.units.1st.sem..grade., na.rm = TRUE),
    min_first_sem_grade  = min(Curricular.units.1st.sem..grade., na.rm = TRUE),
    max_first_sem_grade  = max(Curricular.units.1st.sem..grade., na.rm = TRUE),

    mean_second_sem_grade = mean(Curricular.units.2nd.sem..grade., na.rm = TRUE),
    sd_second_sem_grade   = sd(Curricular.units.2nd.sem..grade., na.rm = TRUE),
    min_second_sem_grade  = min(Curricular.units.2nd.sem..grade., na.rm = TRUE),
    max_second_sem_grade  = max(Curricular.units.2nd.sem..grade., na.rm = TRUE),

    mean_enrolled = mean(Curricular.units.1st.sem..enrolled., na.rm = TRUE),
    sd_enrolled   = sd(Curricular.units.1st.sem..enrolled., na.rm = TRUE),
    min_enrolled  = min(Curricular.units.1st.sem..enrolled., na.rm = TRUE),
    max_enrolled  = max(Curricular.units.1st.sem..enrolled., na.rm = TRUE)
  )

print(summary_table)
```

```r
#installing libraries
library(mosaic)
set.seed(123)

#uploading csv
data <- read.csv("modified.csv")

#selecting columns to take mean grade of
data$mean_grade <- rowMeans(data[, c("Curricular.units.1st.sem..grade.","Curricular.units.2nd.sem..grade.")],na.rm = TRUE)


#splitting data based on parental education
higher <- data$mean_grade[data$Parents.qualification == 1]
lower  <- data$mean_grade[data$Parents.qualification == 0]

#defining bootstrap parameters
B <- 2000
boot_diff <- numeric(B)

#perform bootstrap
for (i in 1:B) {
  boot_higher <- sample(higher, length(higher), replace = TRUE)
  boot_lower  <- sample(lower, length(lower), replace = TRUE)
  boot_diff[i] <- mean(boot_higher, na.rm = TRUE) - mean(boot_lower, na.rm = TRUE)
}

#95% bootstrap confidence interval
ci <- quantile(boot_diff, c(0.025, 0.975), na.rm = TRUE)


#printing results
cat("Bootstrap mean difference in average grade (Higher - Lower):",
    round(mean(boot_diff), 3), "\n")
cat("95% Confidence Interval: [",
    round(ci[1], 3), ",", round(ci[2], 3), "]\n")

 #plotting bootstrap distribution
 hist(boot_diff,
      main = "The bootstrap distribution of mean grade difference",
      xlab = "The mean grade difference (Higher - Lower)",
      col = "lightblue", border = "white")
 abline(v = ci, col = "red", lwd = 2, lty = 2)
 abline(v = mean(boot_diff), col = "darkblue", lwd = 2)


 ```
```

PERMUTATION TEST
```

```r
#Loading data
student_profile_data <- read.csv("~/Downloads/data_with_parents_qualification.csv")
head(student_profile_data,5)
student_profile_data_df <- data.frame(student_profile_data)
```

Description: df [5 × 40]

	Marital.status <int>	Application.mode <int>	Application.order <int>	Co... <int>
1	1	17	5	171
2	1	15	1	9254
3	1	1	5	9070
4	1	17	2	9773
5	2	39	1	8014

5 rows | 1-5 of 40 columns

```r
#Combining indicidual mark into a overall column
library(dplyr)
student_profile_data <- student_profile_data_df %>%
  mutate(overall_marks= Curricular.units.1st.sem..grade. + Curricular.units.2nd.sem..grade.)
student_profile_data$overall_marks
```

```
   [1]  0.00000 27.66667  0.00000 25.82857 25.33333 23.35714 27.64500  0.00000 28.01786
 24.90000 26.53333 26.42857  0.00000 21.57143
  [15] 25.25000 13.20000 23.00000 27.85125 24.75000 25.16667  0.00000 22.86250 25.71429
 25.66071 27.41095 22.60000 24.66071 25.66667
  [29] 24.33333 26.65000 22.42857 26.23333 24.83333 27.04167 25.82857 10.00000  0.00000
 22.00000 22.60000 23.66667 12.75000 30.60000
  [43] 28.27086 26.80000  0.00000 25.83333 24.86667 27.28810 24.78571 25.16667 27.83333
```

```r
#favstats calculation
library(mosaic)
favstats(~ overall_marks | `Parents.qualification`, data = student_profile_data)
```

Description: df [2 × 10]

Parents.qualification <chr>	m <dl	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>
0	0	22.00…	24.42…	26.45…	35.90…	20.90…
1	0	21.64…	24.67…	26.95…	36.56…	20.71…

2 rows | 1–7 of 10 columns

```r
#Difference in mean
diff_mean = favstats(~ overall_marks | `Parents.qualification`, data = student_profile_data
)$mean[2] - favstats(~ overall_marks | `Parents.qualification`, data = student_profile_data
)$mean[1]
diff_mean
```

```
[1] -0.1839479
```

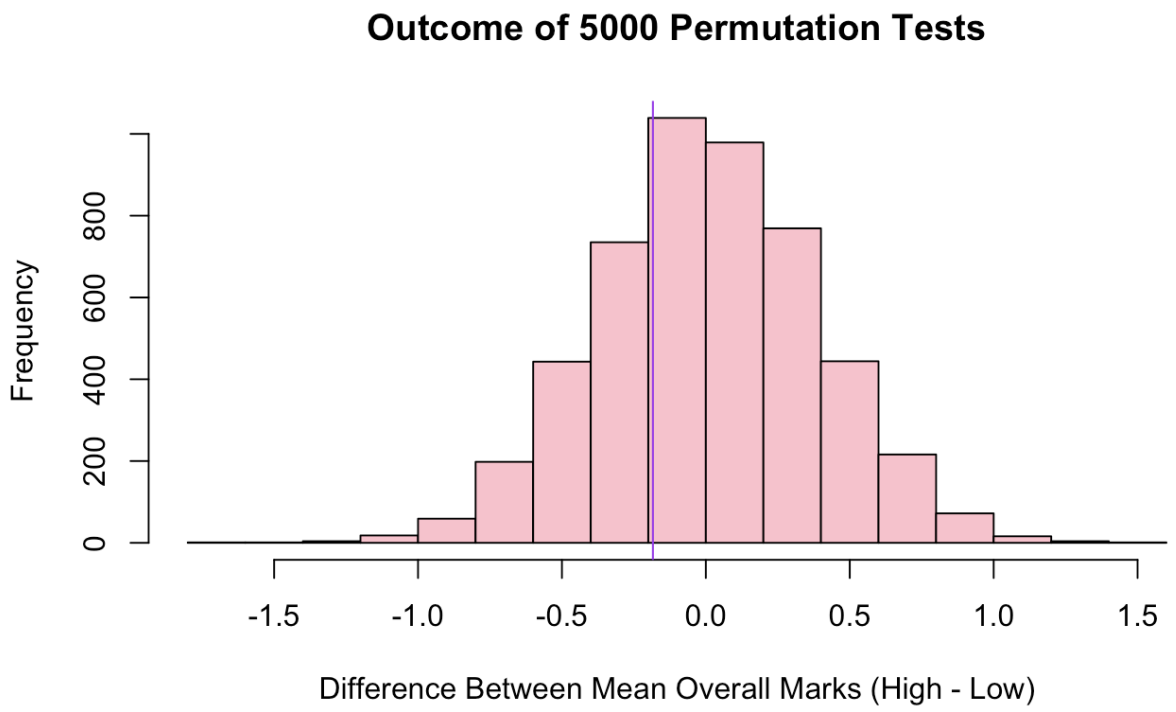
```r
#Carry out permutation test (5000 permutations)
N = 4999
perm.outcome1 = numeric(N)
n_total <- nrow(student_profile_data)
n_high  <- sum(student_profile_data$`Parents.qualification` == 1)

for(i in 1:N)
{
  index = sample(n_total, n_high, replace = FALSE)
  perm.outcome1[i] =
    mean(student_profile_data$overall_marks[index]) -
    mean(student_profile_data$overall_marks[-index])
}
```

```r
#Computing p value
p_right <- (sum(perm.outcome1 >= diff_mean) + 1) / (N + 1)
p_right
```

```
[1] 0.692
```

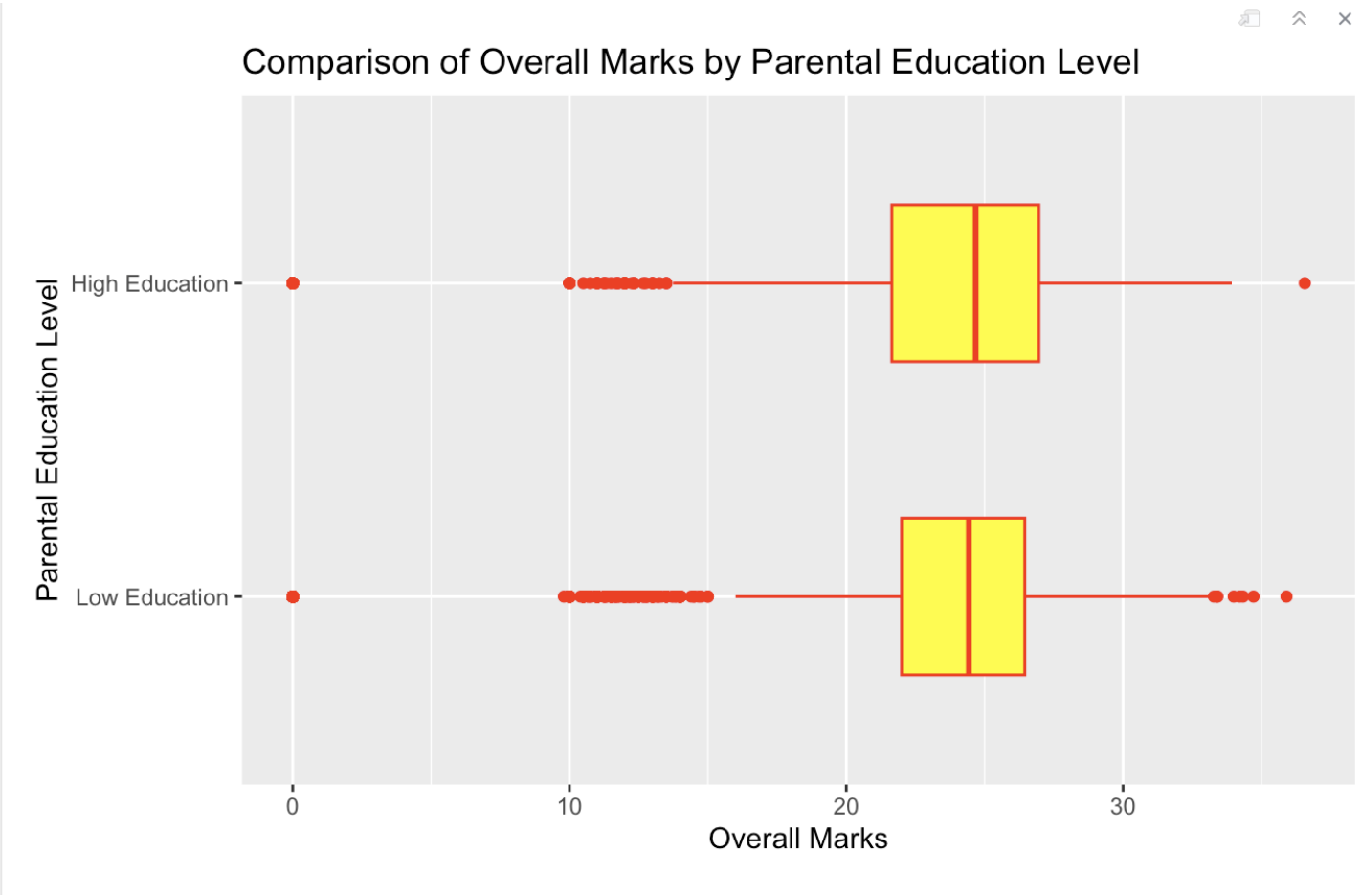
```r
{r}
#Visualize
hist(perm.outcome1,
     xlab = "Difference Between Mean Overall Marks (High - Low)",
     ylab = "Frequency",
     main = "Outcome of 5000 Permutation Tests",
     col = "pink")
abline(v = diff_mean, col = "purple")
```



```r
student_profile_data$Parents.qualification <- factor(
  student_profile_data$Parents.qualification,
  levels = c(0, 1),
  labels = c("Low Education", "High Education")
)

# Create horizontal box plot
library(ggplot2)

ggplot(student_profile_data,
       aes(x = Parents.qualification, y = overall_marks)) +
  geom_boxplot(col = "red", fill = "yellow", width = 0.5) +
  ggtitle("Comparison of Overall Marks by Parental Education Level") +
  xlab("Parental Education Level") +
  ylab("Overall Marks") +
  coord_flip()
```



Chi-square Test

```r
library(ggplot2)

df = read.csv("/Users/sanjay/Downloads/data_with_parents_qualification.csv",
              stringsAsFactors = TRUE)

df$Dropout_status = ifelse(df$Target == "Dropout", "Dropout", "Not_Dropout")

# Create contingency table
table_parents = table(df$`Parents.qualification`, df$Dropout_status)
print(table_parents)

# Chi-square test
chisq.test(table_parents)

chivalues = seq(0, 1, 0.01)
plot(chivalues, dchisq(chivalues,1), xlab="Values of Test Statistic",
     ylab="Density",type="l", col="blue",
     main="Chi-square Distribution with 1 degrees of freedom")
abline(v=0.4098, col='red')

ggplot(df, aes(x = `Parents.qualification`, fill = Dropout_status)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Dropout Status by Parental Qualification",
    x = "Parents' Qualification",
    y = "Number of Students",
    fill = "Dropout Status"
  )
```

Group Members:

1. Ishwarya Rani Murali - 30271950
2. Sanjay Sundar - 30299383
3. Simrat Toor - 30112376
4. Shreyaa Sathesh Kumar - 30290646