Irakli Matcharashvili

**Data Analysis project sample**

This data analysis project involves a comprehensive exploration of factors influencing student academic success and dropout rates using a dataset from the UCI Machine Learning Repository[[1]](#footnote-1). The study focuses on variables such as Marital Status, Application Mode, Previous Qualification, Previous Qualification Grade, Daytime/Evening Attendance, Age at Enrollment, and Gender, with the objective of understanding their impact on the educational outcomes (Dropout, Enrolled, Graduate) of students. Descriptive statistics will provide an overview of these variables. The project also includes inferential statistical analyses:

1. a Chi-Square Test of Independence to examine the relationship between marital status and academic success,
2. a One-Way ANOVA to investigate differences in previous qualification grades among different student outcomes,
3. a One-Way ANOVA to investigate differences in previous qualification grades among marital statuses.

These analyses aim to uncover significant patterns and associations that can offer insights into the factors contributing to students' academic trajectories.

**Descriptive Statistics**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| Marital Status | Integer | 1 – single 2 – married 3 – widowed 4 – divorced 5 – facto union 6 – legally separated |
| Application mode | Integer | 1 - 1st phase - general contingent 2 - Ordinance No. 612/93 5 - 1st phase - special contingent (Azores Island) 7 - Holders of other higher courses 10 - Ordinance No. 854-B/99 15 - International student (bachelor) 16 - 1st phase - special contingent (Madeira Island) 17 - 2nd phase - general contingent 18 - 3rd phase - general contingent 26 - Ordinance No. 533-A/99, item b2) (Different Plan) 27 - Ordinance No. 533-A/99, item b3 (Other Institution) 39 - Over 23 years old 42 - Transfer 43 - Change of course 44 - Technological specialization diploma holders 51 - Change of institution/course 53 - Short cycle diploma holders 57 - Change of institution/course (International) |
| Previous qualification | Integer | 1 - Secondary education 2 - Higher education - bachelor's degree 3 - Higher education - degree 4 - Higher education - master's 5 - Higher education - doctorate 6 - Frequency of higher education 9 - 12th year of schooling - not completed 10 - 11th year of schooling - not completed 12 - Other - 11th year of schooling 14 - 10th year of schooling 15 - 10th year of schooling - not completed 19 - Basic education 3rd cycle (9th/10th/11th year) or equiv. 38 - Basic education 2nd cycle (6th/7th/8th year) or equiv. 39 - Technological specialization course 40 - Higher education - degree (1st cycle) 42 - Professional higher technical course 43 - Higher education - master (2nd cycle) |
| Previous qualification (grade) | Continuous | Grade of previous qualification (between 0 and 200) |
| Daytime/evening attendance | Integer | 1 – daytime 0 - evening |
| Age at enrollment | Integer | Age of student at enrollment |
| Gender | Integer | 1 – male 0 – female |
| Target | Categorical – Factor w three levels | Target. The problem is formulated as a three category classification task (dropout, enrolled, and graduate) at the end of the normal duration of the course |

**Frequencies and Percentages for marital status, application mode, previous qualification, Daytime/evening attendance, Gender, Target (outcome)**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Frequency** | **Percentage** |
| **Marital Status** |  |  |
| Single | 892 | 89.2 % |
| Married | 80 | 8.0 % |
| Widowed | 3 | 0.3 % |
| Divorced | 23 | 2.3 % |
| Facto Union | 1 | 0.1 % |
| Legally Separated | 1 | 0.1 % |
| **Application Mode** |  |  |
| 1st phase - general contingent | 391 | 39.1 % |
| 1st phase - special contingent (Azores Island) | 5 | 0.5 % |
| Holders of other higher courses | 28 | 2.8 % |
| International student (bachelor) | 7 | 0.7 % |
| 1st phase - special contingent (Madeira Island) | 6 | 0.6 % |
| 2nd phase - general contingent | 200 | 20.0 % |
| 3rd phase - general contingent | 28 | 2.8 % |
| Over 23 years old | 183 | 18.3 % |
| Transfer | 18 | 1.8 % |
| Change of course | 57 | 5.7% |
| Technological specialization diploma holders | 51 | 5.1 % |
| Change of institution/course | 19 | 1.9 % |
| Short cycle diploma holders | 7 | 0.7% |
| **Previous Qualification** |  |  |
| Secondary education | 831 | 83.1 % |
| Higher education - bachelor's degree | 1 | 0.1 % |
| Higher education - degree | 29 | 2.9 % |
| Higher education - master's | 3 | 0.3 % |
| Frequency of higher education | 6 | 0.6 % |
| 12th year of schooling - not completed | 5 | 0.5 % |
| 11th year of schooling - not completed | 2 | 0.2 % |
| Other - 11th year of schooling | 7 | 0.7% |
| 10th year of schooling - not completed | 1 | 0.1 % |
| Basic education 3rd cycle (9th/10th/11th year) or equiv. | 39 | 3.9 % |
| Technological specialization course | 55 | 5.5 % |
| Higher education - degree (1st cycle) | 12 | 1.2 % |
| Professional higher technical course | 7 | 0.7 % |
| Higher education - master (2nd cycle) | 2 | 0.2 % |
| **Daytime/Evening Attendance** |  |  |
| Evening | 108 | 10.8 % |
| Daytime | 892 | 89.2 % |
| **Gender** |  |  |
| Male | 255 | 25.5 % |
| Female | 745 | 74.5 % |
| **Target** |  |  |
| Dropout | 292 | 29.2 % |
| Enrolled | 175 | 17.5 % |
| Graduate | 533 | 53.3 % |

A graph of a graph of a graph

Description automatically generated with medium confidence

Figure 1 Frequency Distribution of Student Outcomes

**Means, Medians, and Standard Deviations for Previous Qualification Grade and Age at Enrollment**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **M** | **SD** | **Median** |
| **Previous Qualification Grade** | 132.4846 | 13.07 | 133 |
| **Age at Enrollment** | 23.23 | 7.76 | 20 |

**APA Write Up**

The sample included 1000 students with varying marital statuses. The most common marital status was 'Single' (represented numerically as '1'), comprising a significant majority of the sample 892 (89.2%). Other marital statuses were considerably less frequent. Students were categorized based on their attendance type. The majority of the students 892 (89.2%) attended daytime classes, while the remaining 108 (10.8%) attended evening classes. As for the gender (biological sex) of the students of these, 745 (74.5 %) were identified as female, and 255 (25.5 %) as male. The outcome variable classified the students into three categories: Dropout, Enrolled, and Graduate. Among the students, 'Graduate' was the most common outcome, accounting for 533 (53.3%) of the sample.

The average grade for previous qualifications among the students was 132.48 (*SD* = 13.07). The median grade was slightly higher at 133. The mean age at which students enrolled was 23.23 years (*SD* = 7.76). The median age was 20 years.

**Inferential Statistics**

1. a Chi-Square Test of Independence to examine the relationship between marital status and academic success.

**Step 1**: State the hypothesis

H0: Marital status and Academic Success are independent of each other in the dataset.

H1: Marital status and Academic Success are dependent on each other in the dataset.

**Step 2:** Set the criteria for decision

**Step 3**: Conduct the Test and Check Assumptions

We’re going to use Chi Square for independence.

Chi-Square Statistic: 17.216

df = 10

p-value = 0.07

A screenshot of a computer screen

Description automatically generated

Figure 2 Expected VS Observed Frequencies

A screenshot of a computer

Description automatically generated

Figure 3 Residuals

Residual for Enrolled Married students is < -2 and Divorced Dropouts are noteworthy.

**Step 4:** Make a Decision

Since the P-value 0.07is greater than the alpha level of 0.05, we retain the null hypothesis.

This implies that there is not enough evidence to conclude that there is a significant relationship between marital status and academic success in the sample data.

**APA Write up**A Chi-Square Test of Independence was performed to investigate the relationship between marital status and academic success among students. The results of the Chi-Square test, χ² (10, *N* = 1000) = 17.22, revealed that the association between marital status and academic success was not statistically significant, p = .07. This indicates that within this dataset, the frequency of students' academic outcomes—whether they dropped out, remained enrolled, or graduated—does not significantly differ based on their marital status. The distribution of academic success appears to be independent of the marital status of the students in this sample.

1. a One-Way ANOVA to investigate differences in previous qualification grades among different Student Outcomes

**Step 1:** State the hypothesis:

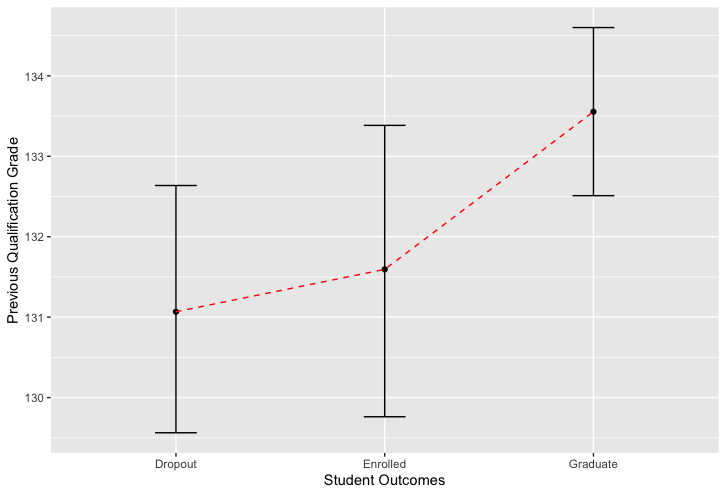
H0: There are no differences in previous qualification grades among the different student outcomes.

H1: There are significant differences in previous qualification grades among the different student outcomes.

**Step 2**: Set the criteria for decision.

= 0.05

**Step 3**: Check assumptions and perform the test



* Normality test (Shapiro-Wilk)

Dropout: p<0.001

Enrolled p< 0.01

Graduate p< 0.001

The normality test shows violations for all three groups. This suggests that the data for each student outcome group may not be normally distributed.

* Homogeneity of Variances (Levene's Test):

Levene statistic: 1.8724 p = 0.1543

The p-value is greater than 0.05, meaning no significant differences in variances between groups, meeting the homogeneity of variances assumption.

* ANOVA

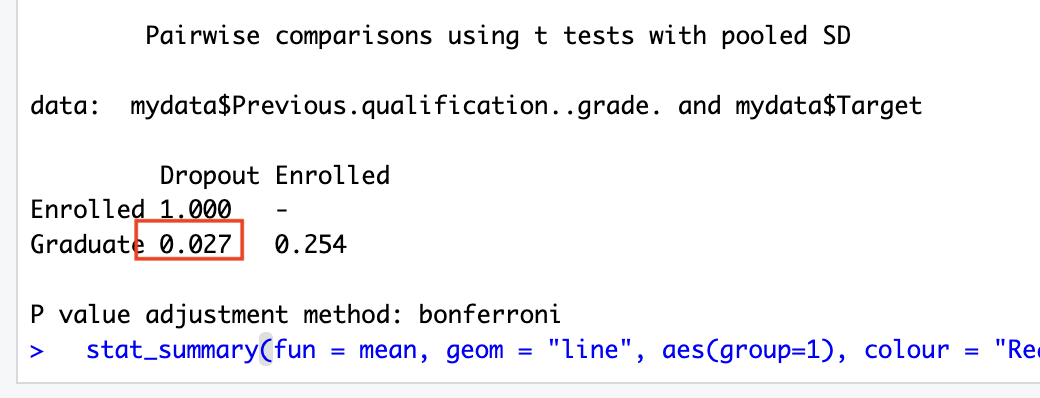
F value: 3.932 p = 0.02

**Step 4**: Make a decision.

Given the p-value (.020) is less than the alpha level of .05, **we reject the null hypothesis**.

This suggests that there are significant differences in previous qualification grades among the different student outcomes.

However, it is important to note, that the normality assumption was violated. This could affect the validity of the ANOVA results. In this case we should follow up Bonferroni test.



The Bonferroni test results indicate significant differences in Previous Qualification Grades between Graduated and Dropout students (p=0.027).

A screenshot of a computer

Description automatically generated

The Tukey test results corroborate Bonferroni findings: suggesting that graduates had significantly higher previous qualification grades compared to students who dropped out.

**APA Write up.**

A One-Way ANOVA was conducted to examine the differences in previous qualification grades among different student outcomes (Dropout, Enrolled, Graduate). The analysis revealed significant differences between these groups, *F*(2, 997) = 3.93, *p* < .05. The effect size was small, with *ω²* = 0.00583, indicating that a small proportion of the variance in previous qualification grades was explained by student outcomes. Post-hoc analyses using the Bonferroni adjustment method indicated a significant difference in previous qualification grades between students who dropped out (*M* = 131.07, *SD* = 12.72) and those who graduated (*M =* 133.55*, SD =* 13.27). However, no significant differences were found between students who dropped out and those who remained enrolled (*M* = 131.59, *SD* = 12.81), or between enrolled students and graduates. Further post-hoc comparisons using the Tukey HSD test corroborated these findings, showing a significant difference in grades only between the Dropout and Graduate groups (*p* = .024). The assumptions of the ANOVA were partially met; the homogeneity of variances was not violated as indicated by Levene’s test (*F* = 1.87, *p* = .154), but the normality assumption was violated for each group (Shapiro-Wilk *p* < .05). Given the violation of normality, interpretations of these results should be made with caution.

1. a One-Way ANOVA to investigate differences in previous qualification grades among different Marital Status.

**Step 1**: State the hypothesis.

H0: There is no difference in previous qualification grades among different marital statuses.

H1: There is a difference in previous qualification grades among different marital statuses.

**Step 2**: Set the criteria

We should use ANOVA since it is suitable for comparing means across more than two groups.

= 0.05

**Step 3**: Check Assumptions

A graph with lines and dots

Description automatically generated

* Normality test (Shapiro-Wilk)

Since widowed, facto union and legally separated are less than 3 I will check other groups for normality.

Single: p<0.001

Married: p< 0.001

Divorced: p< 0.001

The normality test shows violations for all three groups.

* Homogeneity of Variances (Levene's Test):

Levene’s statistic: 3.398 p = 0.005

Since the p-value is below 0.05, we’re not meeting the homogeneity of variances assumption.

* ANOVA

F-value: 0.605, p = 0.696

p>0.05, which means there is no statistically significant difference in previous qualification grades across different marital status groups.

**Step 4:** Make a decision

p>0.05 we retain the null hypothesis.

**APA Write up**

A One-Way ANOVA was conducted to examine the differences in previous qualification grades among students with different marital statuses. The analysis did not reveal significant differences in grades based on marital status; *F*(5, 994) = 0.605 *p* > .05. This outcome indicates that marital status does not significantly affect the previous qualification grades among the students in this sample. In the absence of significant findings from the ANOVA, post-hoc comparisons were not conducted.

1. https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success [↑](#footnote-ref-1)