****

**Term:** Spring 2025

**Course Title:**  Statistics for Data Analysts (DSE 501)

Project Report on

**Academic Performance Through the Lens of Socio-Demographic Data**

Submitted to: **Dr. Rong Pan**

Submitted by:

|  | **Name** | **ASU ID** | **Email ID** |
| --- | --- | --- | --- |
| **1** | **Amit Ashok Manjarly** | **1233931935** | **amanjarl@asu.edu** |
| **2** | **Chetan Jitendra** | **1233602398** | **cjitendr@asu.edu** |
| **3** | **Isha Kaushik** | **1233665929** | **ikaushi4@asu.edu** |
| **4** | **Kushal Keshav Ravipati** | **1234047167** | **kravipa1@asu.edu** |

**TABLE OF CONTENTS**

| **TOPIC** | **PAGE NO.** |
| --- | --- |
| ABSTRACT | **2** |
| KEYWORDS | **2** |
| OBJECTIVE | **2** |
| INTRODUCTION | **2** |
| DATASET COLLECTION AND CHARACTERISTICS | **3** |
| PROBLEMS FOUND WITH DATASET AND HOW WE TACKLED THEM | **3** |
| EXPLORATORY DATA ANALYSIS | **3** |
| HYPOTHESIS OF INTEREST | **6** |
| METHODOLOGY OF HYPOTHESIS TEST - 1 using Pearson’s Chi-Square Test | **6** |
| METHODOLOGY OF HYPOTHESIS TEST - 2 using Likelihood Ratio Test | **9** |
| METHODOLOGY OF HYPOTHESIS - 3 using Linear Regression | **11** |
| METHODOLOGY OF HYPOTHESIS - 4 using Chi-Square Test of Independence | **13** |
| CONCLUSION | **17** |
| RECOMMENDATIONS FOR BUSINESS STRATEGIES | **18** |
| LIMITATIONS & FUTURE SCOPE | **18** |
| REFERENCES | **18** |

**Abstract**

This study investigates the key factors influencing student performance in Math, Reading, and Writing. It examines the impact of socio-demographic variables, educational supports, and external factors like test preparation courses and lunch programs on academic outcomes. The analysis utilizes regression analysis to develop a predictive model for student performance, aiming to provide insights for educators and policymakers to enhance academic success.

**Keywords**

Student Performance, Socio-demographic Factors, Educational Supports, Test Preparation, Lunch Program, Regression Analysis, Academic Outcomes

**Objective**

The primary objectives of this study are to identify key factors influencing student performance in Math, Reading, and Writing, distinguishing between socio-demographic variables and educational supports, while also assessing the impact of external factors such as test preparatory courses and lunch types on academic outcomes to determine their significance. Additionally, the study aims to develop a predictive model using regression analysis, trained on existing data, to forecast student performance based on the identified factors, enabling a deeper understanding of the variables that most critically shape academic success.

**Introduction**

Student academic performance is a complex outcome influenced by a multitude of factors. These factors can range from socioeconomic status and parental education background to the availability of school resources and the effectiveness of test preparation. Understanding the interplay of these variables is crucial for identifying and addressing performance gaps, and for developing targeted interventions to improve educational outcomes. This study utilizes a dataset containing student exam scores in Math, Reading, and Writing, along with demographic and background information, to analyze the relative impact of these factors on student achievement.

**Dataset Collection and Characteristics**

The dataset includes detailed records of student exam scores and various influencing factors. The key variables are:

* **Scores:** **Math, Reading, and Writing** - These are the numerical scores obtained by students in each subject.
* **Demographics:** **Race/Ethnicity** - This categorical variable represents the racial or ethnic background of the students.
* **Parental Education: Highest level of education of parents** - This categorical variable indicates the highest level of education attained by the students' parents.
* **Economic Factors:** **Availability of standard or free/reduced-price lunch** - This categorical variable represents the type of lunch program the student participates in, indicating their economic background.
* **Test Preparation: Completed the test preparatory course or not** - This binary variable indicates whether the student completed a test preparation course.

The dataset provides a comprehensive view of factors influencing student performance, allowing for the analysis of how demographic, socioeconomic, and educational variables relate to academic outcomes."

**Problems Found With the Dataset and How We Tackled Them**

During our initial study of the dataset, we discovered that there were no missing values, which simplified the preprocessing procedure. However, the dataset was unbalanced in some categorical variables, like gender and lunch type. Additionally, several categorical variables had to be encoded prior to statistical analysis.

Data types were cleansed and checked to ensure they were appropriate for test applications. Categorical encoding and normalization were used where needed when developing the model.

**Exploratory Data Analysis**

The dataset consists of 1,000 student records and includes both categorical and numerical variables related to academic performance and socio-demographic background.

We began by exploring the structure and summary statistics of the data. There were no missing values, and all data types were appropriate for analysis.

**Summary of numerical data revealed**:

* **Average scores**: Math (66.09), Reading (69.17), and Writing (68.05)
* **Standard deviations** were moderate, indicating varied performance across students.
* Scores ranged from 0 to 100, with some students having extremely low performance, particularly in math.

|  | **Math Score** | **Reading Score** | **Writing Score** |
| --- | --- | --- | --- |
| **Count** | 1000 | 1000 | 1000 |
| **Mean** | 66.089 | 69.169 | 68.054 |
| **Std** | 15.1631 | 14.6002 | 15.1957 |
| **Min** | 0 | 17 | 10 |
| **25%** | 57 | 59 | 57.75 |
| **50%** | 66 | 70 | 69 |
| **75%** | 77 | 79 | 79 |
| **Max** | 100 | 100 | 100 |

**Table 1: Statistics for Math, Reading, and Writing scores of 1000 students**

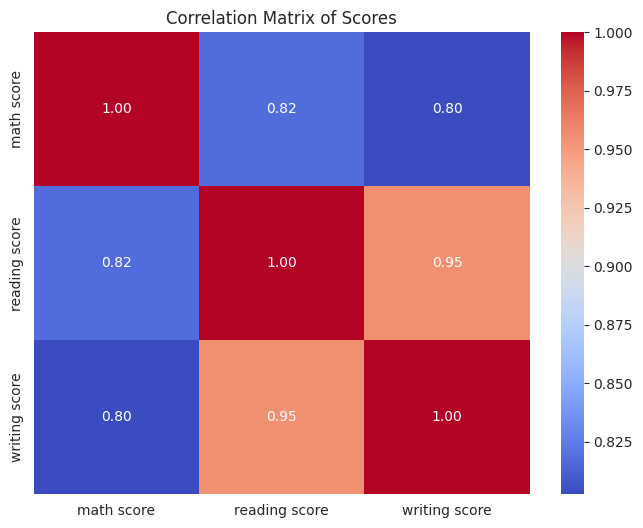
**Categorical variables** showed:

* A majority of students were **female (518)** and from **group C (319)**
* Most common parental education: *some college*
* 64.5% received **standard lunch**, and 64.2% did **not complete** the test preparation course

|  | **Gender** | **Race/Ethnicity** | **Parental Level of Education** | **Lunch** | **Test Preparation Course** |
| --- | --- | --- | --- | --- | --- |
| **Count** | 1000 | 1000 | 1000 | 1000 | 1000 |
| **Unique** | 2 | 5 | 6 | 2 | 2 |
| **Top** | female | Group C | Some College | Standard | None |
| **Freq** | 518 | 319 | 226 | 645 | 642 |

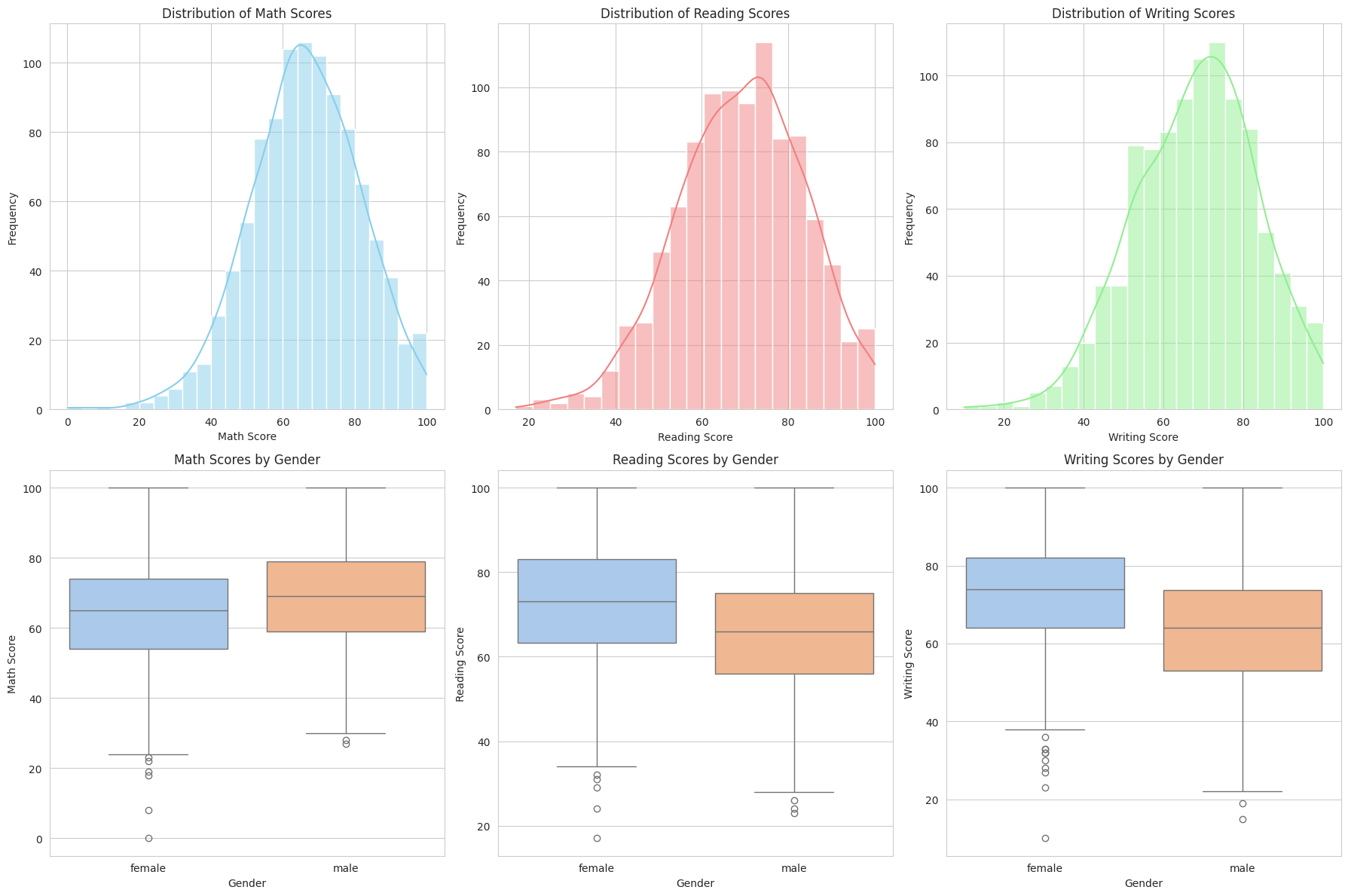
**Table 2 : Demographic Characteristics of Students**

We visualized the relationships between variables to understand patterns in academic performance. The **correlation matrix** showed a strong positive correlation between reading and writing scores, and moderate correlations between math and the other two subjects.



**Figure 1: Correlation Matrix of Student Scores**

Boxplots were created to visualize the variation in scores by **gender**. These plots indicated subtle differences in median scores, which were explored further in hypothesis testing.



**Figure 2: Distribution of Math, Reading, and Writing Scores by Gender**

**Hypothesis of Interest**

**Hypothesis 1: Influence of Socio-Demographic Factors on Academic Performance**

This hypothesis investigates the extent to which socio-demographic factors such as parental education level, gender, and race/ethnicity influence students' academic performance in Math, Reading, and Writing. The underlying premise is that students who have more educated parents would get more academic support, which would boost their performance. In addition, historical evidence and cultural patterns indicate that performance differences across genders and ethnic groups might exist, which calls for statistical verification. The purpose of this analysis is to measure these associations and assess their importance using suitable group comparison tests.

**Objective**

To evaluate whether **gender** significantly impacts student performance in Math, and whether there is any association between gender and test preparation course completion. The hypothesis assumes that gender may influence both academic performance and the likelihood of undertaking preparatory academic support.

**1.** **Gender vs Math Score**

**Variables**

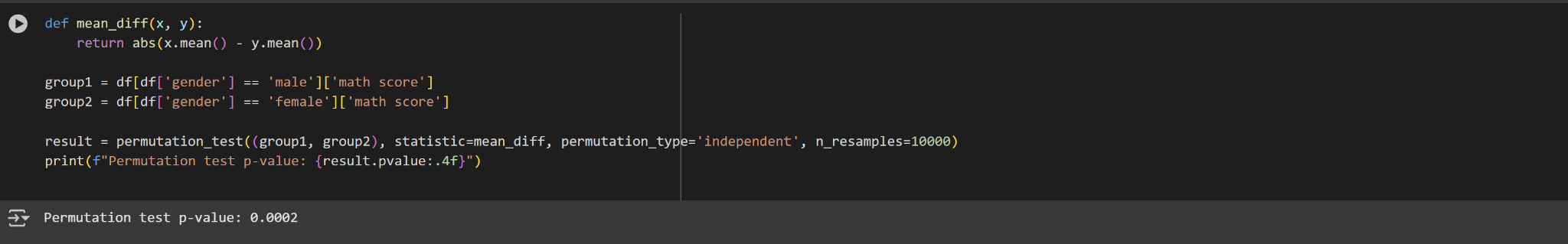
* Independent: *Gender*
* Dependent: *Math Score*

**Methodology**

* Created two groups based on gender.
* The means were compared using a permutation test with 10,000 resamples to assess whether observed differences could be due to random variation.
* A custom function was used to calculate the absolute difference in group means.

**Decision**

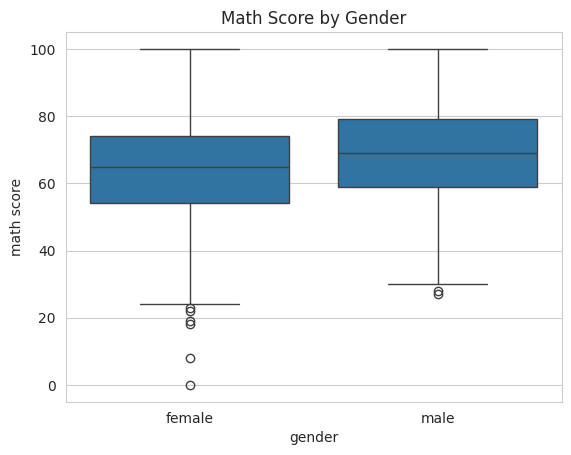
The p-value from the permutation test was reported as **< 0.05**, indicating a **statistically significant** difference in math scores between genders. The code block is shown below:



**Figure 3: Permutation Test Results for Gender Difference in Math Scores**

**Interpretation:**

The statistical analysis suggests that gender has a measurable impact on student performance in certain subjects, specifically Math. The below plot claims the understanding



**Figure 4: Box Plot of Math Scores by Gender**

**2. Gender vs Test Preparation Course Participation**

**Test Used**

* Pearson’s Chi-Square Test

**Variables**

* Independent: *Gender (Male, Female)*
* Dependent: *Test Preparation Course (Completed / None)*

**Methodology**

* A contingency table was created to summarize the counts of students who completed or did not complete the test preparation course, grouped by gender.
* A Chi-Square test of independence was applied to determine whether gender and test preparation course participation were significantly associated.
* The test was implemented using the *‘chi2\_contingency()’* function from *‘scipy.stats’*

**Decision**

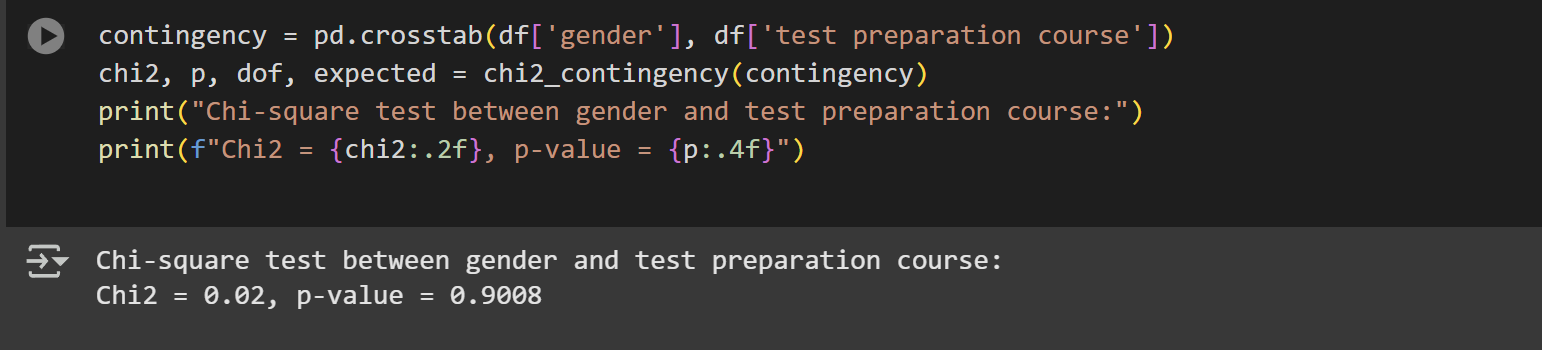
The test yielded the following results:

* **Chi-Square Statistic:** 0.02
* **P-value:** 0.9008
* **Degree of Freedom:** 1

Since the p-value is much greater than 0.05, we fail to reject the null hypothesis.

**Interpretation**

* There is no statistically significant relationship between a student's gender and their participation in the test preparation course.
* The observed variation in participation rates is likely due to random chance, and gender does not appear to influence a student’s decision or opportunity to take the course.

****

**Figure 5: Chi-Square Test for Association Between Gender and Test Preparation Course Participation**

**Hypothesis 2: Students with standard lunch perform better than those with free/reduced lunch**.

**Objective**

To investigate whether economic background, as inferred through the school lunch program status (standard vs free/reduced), significantly influences academic performance. It is hypothesized that students receiving standard lunch, likely from more economically stable households, will perform better in math than their peers receiving subsidized lunches.

**Test Used**

* Likelihood Ratio Test (OLS Regression)

**Variables**

* Independent: *Lunch Program (Standard vs Free/Reduced)*
* Dependent: *Math Score*
* Control Variables: *Gender, Test Preparation Course*

**Methodology**

We developed two linear regression models:

* A full model with lunch, gender, and test preparation course as predictors
* A reduced model with only gender and lunch as predictors.

The Likelihood Ratio Test was applied to compare the models and evaluate whether adding test preparation to the model (while already including lunch) significantly improved the prediction of math scores.

In the full model, we examined the effect of lunch program status specifically through the *‘c(lunch)’* and *‘[T.standard]’* coefficients

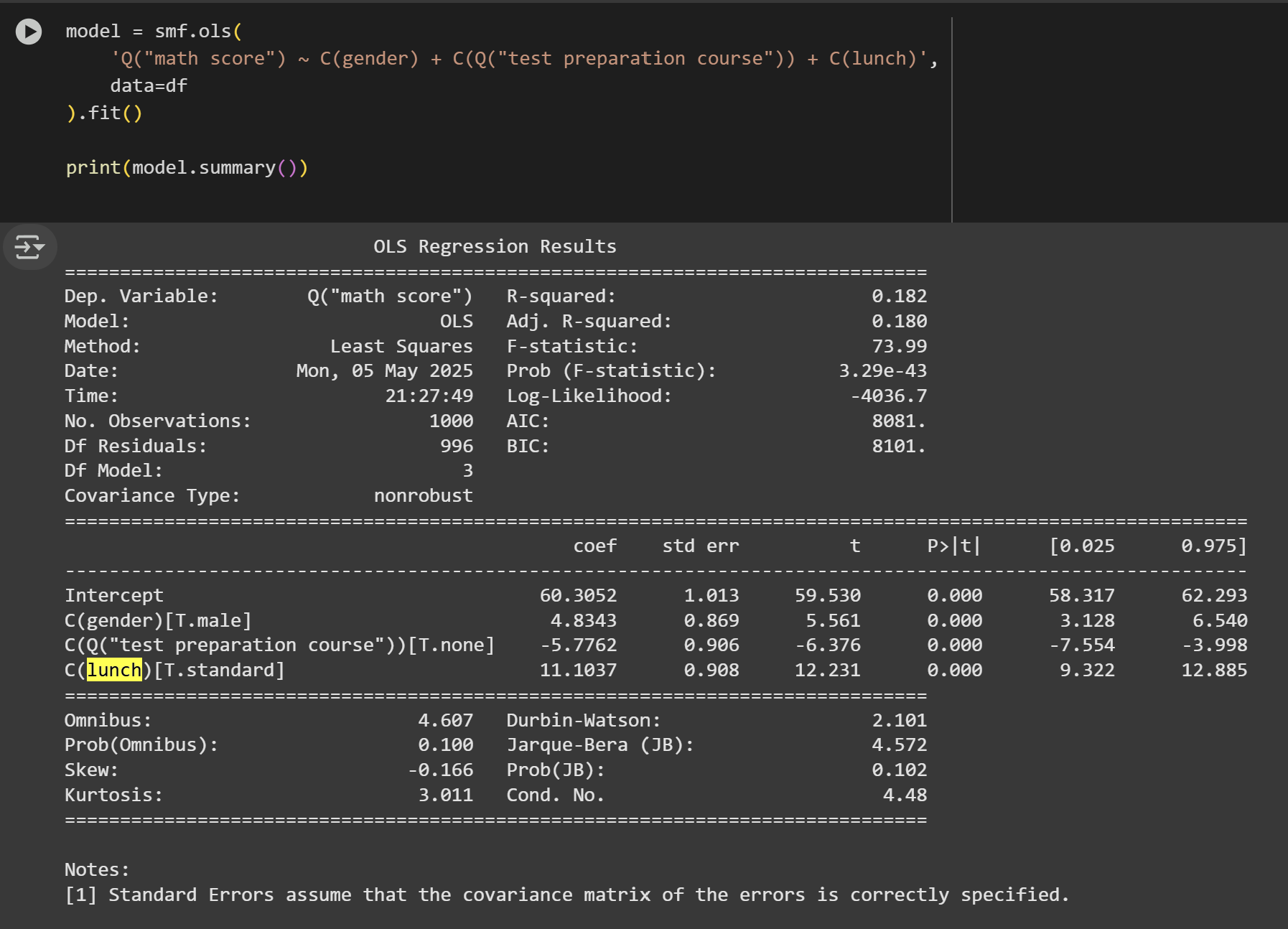
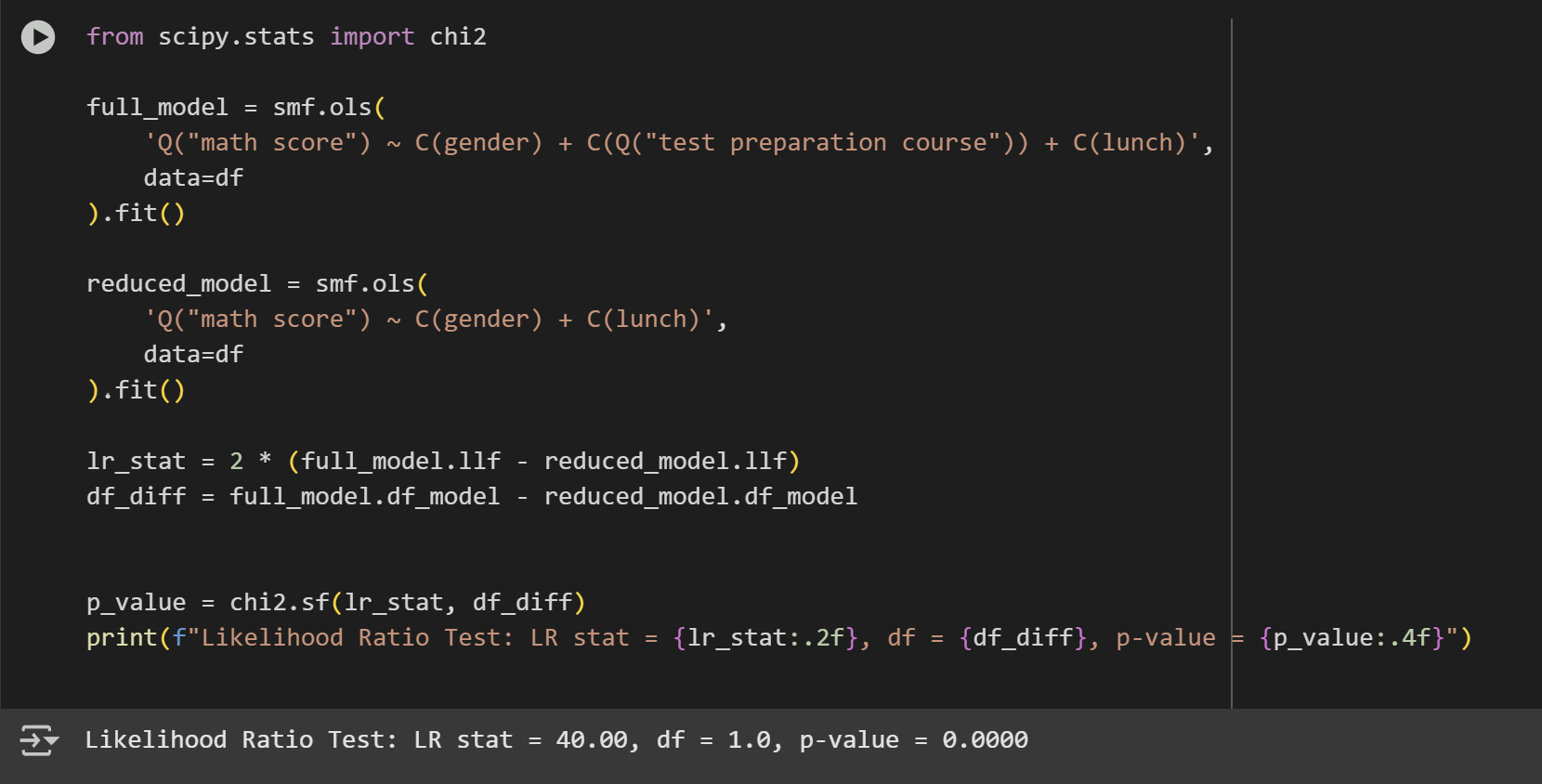
**Decision**

* **Likelihood Ratio Statistic:** 40.00
* **P-value:** 0.0000
* **Degree of Freedom:** 1
* **Regression Coefficient (Lunch = Standard):** +11.10
* **P-value (from regression):** < 0.0001

We reject the null hypothesis, concluding that lunch program status has a statistically significant effect on math performance.

**Interpretation**

* Students who received standard lunch scored, on average, 11 points higher in math compared to those on free/reduced lunch.
* This confirms the hypothesis that economic advantage, represented by access to standard lunch, is associated with better academic performance.
* The test supports that economic background is a strong and significant predictor of student success.



**Figure 6 & 7: Chi-Square Test for Association Between Gender and Test Preparation Course Participation**

**Hypothesis 3: Participation in Test Preparation Courses and Performance**

**Objective**

To evaluate whether completing a test preparation course significantly improves a student’s academic performance. The hypothesis assumes that preparatory support offers academic advantage, regardless of demographic background.

**Test Used**

* Linear Regression (OLS)

**Variables**

* Independent: *Test Preparation Course (Completed / None)*
* Dependent: *Math Score*
* Control Variables: *Gender, Lunch Type*

**Methodology**

* An Ordinary Least Squares (OLS) regression model was used to estimate the impact of completing a test preparation course on student performance.
* The categorical variable *‘test preparation course’* was included in the model, where the group labeled *‘none’* (students who did not complete the course) was compared against the baseline group *‘completed’*
* The model also adjusted for other variables like gender and lunch status, isolating the effect of test preparation on scores.

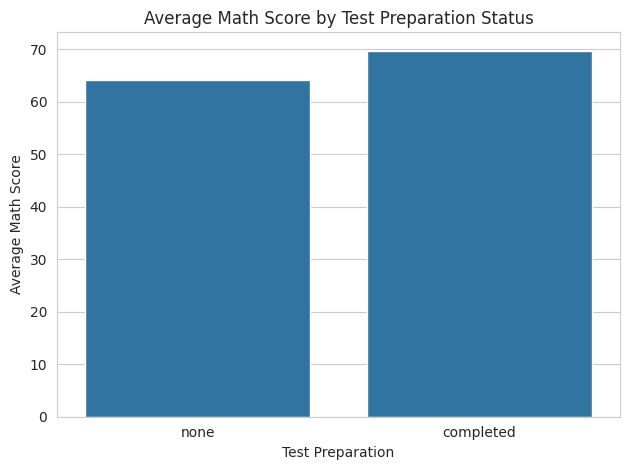
**Decision**

* **Coefficient for students who did not complete the course:** −5.7762
* **p-value:** 0.0000

Since the p-value is well below 0.05, we reject the null hypothesis, indicating a significant difference in scores based on test prep participation.

**Interpretation**

* The regression results show that students who did not complete the test preparation course scored, on average, 5.78 points lower in math compared to those who completed it.
* This difference is statistically significant, confirming that test preparation courses are effective in improving student outcomes.
* The findings support the value of such programs in boosting academic achievement.



**Figure 8: Average Math Score by Test Preparation Status**

**Hypothesis 4: Interrelation between Parental Education and Test Preparation Participation**

**Objective**

To evaluate whether a student’s parental level of education influences their likelihood of completing a test preparation course. The hypothesis assumes that students with more educated parents may be more likely to engage in structured academic support programs.

**Test Used**

* Chi-Square Test of Independence

**Variables**

* Independent: *Parental Level of Education (6 categories)*
* Dependent: *Test Preparation Course (Completed / None)*

**Methodology**

* A contingency table was created between *‘parental level of education’* and *‘test preparation course’* participation.
* A Chi-Square test of independence was applied using *‘chi2\_contingency()’* from the *‘scipy.stats’* module
* This test evaluates whether the distribution of course participation varies significantly across different education levels of parents.

**Decision**

* **Chi-Square Statistic:** 9.5441
* **Degrees of Freedom:** 5
* **p-value:** 0.0892

Since the p-value is greater than 0.05, we fail to reject the null hypothesis.

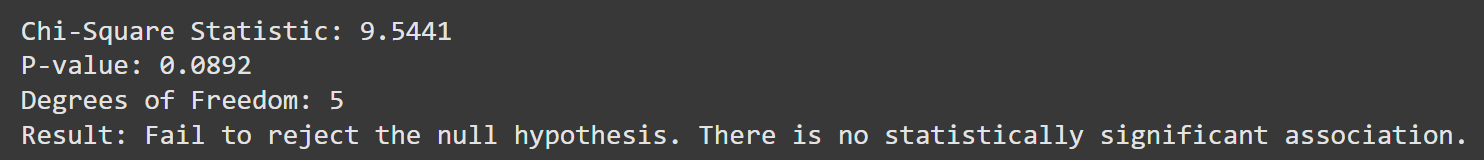
**Interpretation**

* The test indicates no statistically significant association between parental education level and a student’s likelihood of completing the test preparation course.
* While certain education groups may appear more represented, this variation is not strong enough to conclude a true dependency.
* This suggests that factors beyond parental education, such as school culture or personal motivation, may have a greater role in course participation.

| **Parental Level of Education** | **Completed** | **None** |
| --- | --- | --- |
| **Associate’s Degree** | 82 | 140 |
| **Bachelor’s Degree** | 46 | 72 |
| **High School** | 56 | 140 |
| **Master’s Degree** | 20 | 39 |
| **Some College** | 77 | 149 |
| **Some High School** | 77 | 102 |

**Table 3: The Contingency Table of parental education and test preparation**

Results of the Chi-Square test:



**Figure 9: Chi-Square Test Results for Association Analysis**

**Regression-Based Prediction of Student Scores**

In addition to hypothesis testing, we developed multiple linear regression models to predict Math, Reading, and Writing scores using all relevant background information. The goal of these models was to see how effectively demographic and support-related variables (such as gender, race/ethnicity, parental education, lunch style, and test preparation) might predict academic achievement together.

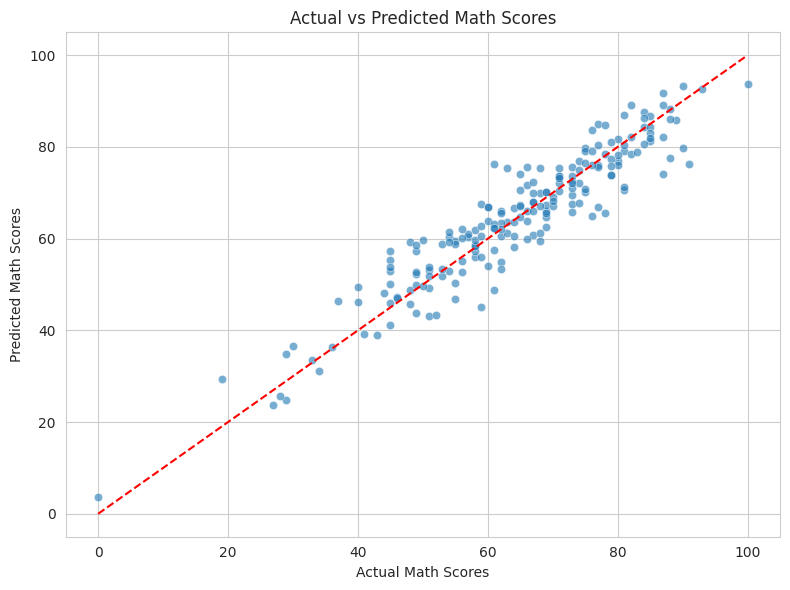
For categorical data, we employed one-hot encoding and then a linear regression pipeline. The dataset was divided into training and test sets (80/20 split). Models were evaluated using R² Score, Root Mean Squared Error (RMSE), and Mean Squared Error (MSE).

**1. Math Score Prediction**

The model predicted math scores with strong accuracy

**Performance Metrics:**

* **R² Score:** 0.880
* **RMSE:** 5.394
* **MSE:** 29.095



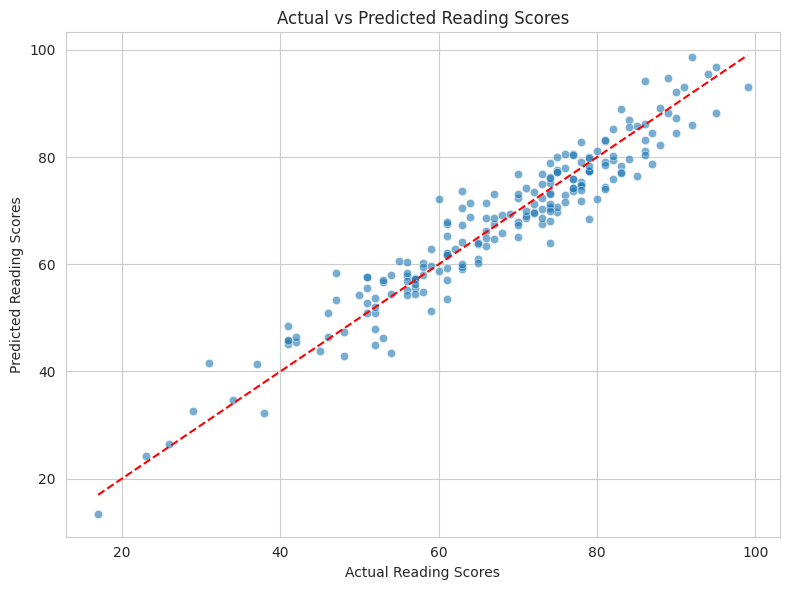
**Figure 10: Actual vs. Predicted Math Scores**

**2. Reading Score Prediction**

The reading score model performed significantly better.

**Performance Metrics:**

* **R² Score:** 0.918
* **RMSE:** 4.309
* **MSE:** 18.566

****

**Figure 11: Actual vs. Predicted Reading Scores**

**3. Writing Score Prediction**

The model for writing scores performed best among the three

**Performance Metrics:**

* **R² Score:** 0.938
* **RMSE:** 3.861
* **MSE:** 14.911



**Figure 10: Actual vs. Predicted Writing Scores**

**Conclusion**

The study used regression modeling and statistical hypothesis testing to investigate how different demographic and socioeconomic characteristics affect students' academic performance.

The findings indicated that gender, test preparation engagement, and economic background (by meal status) all have a substantial impact on performance, particularly math scores. Students who took a test preparation course, had regular meals, or belonged to specific demographic groupings did better.

The investigation did not find a statistically significant direct relationship between education level and test preparation participation or performance, even though parental education demonstrated intuitive linkages to academic assistance.

Overall, the findings support targeted interventions for underperforming student groups, particularly those based on economic and preparation variables, to improve academic performance.

**Limitations and Future Scope**

**Dataset Scope:** The dataset's 1,000-entry limit and concentration on American students may have limited its applicability to other demographics or situations abroad.

**Unexplored Interactions:** The current analysis did not thoroughly examine certain interactions, such as the combined impact of race and test preparation or gender and lunch.

**Future Research:** Trends over time may be shown by a more thorough investigation employing longitudinal or more detailed datasets. Predictive accuracy may also be increased by including qualitative data (such as instructor influence and student motivation).

**Recommendations**

**Encourage Test Preparation Programs:** Schools have to think about requiring or rewarding test preparation programs because of their significant benefits, particularly for children from disadvantaged backgrounds.

**Address Economic Disparities:** The need for support services, such as tutoring and mentorship programs, for students from low-income families is underscored by the strong association found between the kind of lunch program and performance.

**Utilize Data to Customize Learning:** By utilizing demographic information, educational institutions can create specialized academic programs that target particular disparities in gender, ethnicity, or resource availability.

**Encourage Parental participation:** Promoting parental participation may have a positive impact on student results, even if the parents' educational attainment was not statistically significant on its own.

**References**

1. [Al Husaini, Yousuf Nasser Said, and Nur Syufiza Ahmad Shukor. "Factors affecting students’ academic performance: A review." *Social Science Journal* 12.6 (2022): 284-296.](https://www.researchgate.net/profile/Yousuf-Al-Husaini/publication/367360842_Factors_Affecting_Students'_Academic_Performance_A_review/links/63cf73e5d7e5841e0bf04e59/Factors-Affecting-Students-Academic-Performance-A-review.pdf)
2. [Mushtaq, Irfan, and Shabana Nawaz Khan. "Factors affecting students’ academic performance." *Global journal of management and business research* 12.9 (2012): 17-22.](https://d1wqtxts1xzle7.cloudfront.net/49686973/3-Factors-Affecting-Students-Academic-libre.pdf?1476794991=&response-content-disposition=inline%3B+filename%3DGlobal_Journal_of_Management_and_Busines.pdf&Expires=1746755915&Signature=KLezoVqlas~jogX7vAy0EeXEUyzokgrfz~LG0jUjzjPkNUuWyHThKF0HKGibAjsspjaCt06KKIyBYpV4B1DyfD9lP3H7rU5d2DE38q8VYgkRF2AhrOrDDCGNUoaMpUFN5SCspo-Y8mIlg3nl5F6GQ6XOkrMZl9FF6Pc2L2NkaT5xAoX~GiPTWuX7maCBEgnslVe3IEqiqf0UQn5AP3KD3UINQRmGbM037EyAgDxUB4Qo0di5nrpLeMga0Z71RIi2iva-1q0j~IoSzYJm30t2a63VIFeKmihY~IUfvSenzKsHfyBc9mz~RrpNfIlo3zlXBQPMe6Hq5ENK3Vpkgp52rA__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA)
3. [All of Statistics - A Concise Course in Statistical Inference](https://egrcc.github.io/docs/math/all-of-statistics.pdf)
4. [Salem, Raneem O., et al. "Academic and socio-demographic factors influencing students’ performance in a new Saudi medical school." *Medical teacher* 35.sup1 (2013): S83-S](https://www.tandfonline.com/doi/abs/10.3109/0142159X.2013.765551)
5. [Alhajraf, Nout M., and Aishah M. Alasfour. "The impact of demographic and academic characteristics on academic performance." *International Business Research* 7.4 (2014): 92.](https://d1wqtxts1xzle7.cloudfront.net/72348302/1f70cbed39cf74bd5f40c2e21fa0940ec083-libre.pdf?1635146030=&response-content-disposition=inline%3B+filename%3DThe_Impact_of_Demographic_and_Academic_C.pdf&Expires=1746756225&Signature=cv8QSL4bQPh5eV97zy8TY6N4G5sQA6K6YAYB6C56MtlFqhay85BzYxE5sE1GlKJq4H3m0863Sfv0IdOyg03TfjuDiGdppoiwSl~zyrLBZv~djne5uLIx7yjUz7C4pvgECIqt1AcfGEvTWDIKNPP13znUOREvNpv68NewJCDDPdD~vSAFmHZ3ed42U6CqUP5QYBNF9r42EG0qRBUvTrpIfkKI7rCNPsi5-HO17oB8PYGcF7BUIqLPoxmIiK4nX-m9Wxst--1BgfrvFPNitMzqO2ZQTxTocI1H1ktLybnDHXeK5xsrBpirR-aad~7g8ibRtaknxAm8aKTr86ILFgFm0Q__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA)