Trends in Student Interest in STEM and Non-STEM Subjects in Armenia

CS108 Statistics
Final Project Report

Student
Gor Arzanyan

Professor
Dr. Ashot Abrahamyan

American University of Armenia

May 6, 2025

INTRODUCTION

Understanding the educational landscape is essential for developing effective educational policies and strategies. In Armenia, much of our current understanding of the educational system relies largely on assumptions and limited data, primarily derived from university entrance exams and school graduation assessments. However, these traditional metrics frequently overlook the crucial aspect of genuine student interest, which can provide significant insights into the effectiveness of educational strategies, relevance of course content, effectiveness of teaching methodologies, and overall quality of course design.

This research aims to fill this gap by investigating trends in Armenian students' interests in STEM (Science, Technology, Engineering, Mathematics) and non-STEM subjects, examining how these interests differ across demographic groups and regions. Participation in extracurricular activities and students' willingness to seek additional knowledge beyond standard curricula are used as practical indicators of interest. In this study, participation in subject-specific Olympiads, sourced from the official Armenian Olympiad platform (olymp.am), is employed as a reliable measure of student interest, with rigorous data filtering applied to mitigate potential biases.

This research specifically addresses the following questions:

- Does student participation in extracurricular academic activities significantly differ across regions?
- Are Armenian students significantly more interested in STEM subjects compared to non-STEM subjects?
- Does student interest in STEM and non-STEM subjects differ significantly between students living in different regions?

The findings from this study may lay foundations for future analyses of preferences among Armenian students, enabling stakeholders to identify emerging trends and address potential disparities among different student groups.

The report proceeds with a detailed description of the data, followed by exploratory analysis, methodological explanations, presentation of results, and a concluding discussion that contextualizes findings within Armenia's educational landscape.

DATA DESCRIPTION

The analysis focuses on understanding student interests in various subjects differentiating between STEM and non-STEM courses, with the target population being all Armenian students across different regions and grade levels. The records of students participating in school Olympiads across Armenia are used for the study which is sourced from olymp.am. The collected dataset consists of the first stage of Olympiad participation records (school qualification stage) in the most recent complete year (2025). **The decision to use first-stage Olympiad data ensures that the analysis reflects student interest across a broad spectrum, without bias from subsequent stages that filter participants based on academic performance.** This approach includes every student who signed up, regardless of score, which also minimizes the bias related to specific schools and groups.

To compare STEM versus non-STEM interests, the subjects with the most consistent participation and popularity history were chosen from each category:

- **STEM:** Mathematics, Physics, Informatics
- **Non-STEM:** History, Armenian Language, English

Also, additional statistics was extracted from armstat.am.

Information on data preparation and preprocessing is provided in **Appendix A**.

The complete summary of the dataset is provided in **Appendix B**.

EXPLORATORY DATA ANALYSIS

The extracted dataset of the above-mentioned subjects includes a total of **59,911** datapoints. Key variables in the dataset are:

Region: Geographic region of the student (e.g., Yerevan, Ararat, Lori, etc.).

Gender: male or female.

Subject: The subject in which the student participated

Grade Level: The grade of the participating student.

The distribution of participants by region and gender in 2025 is shown in Table 1 and Fig.1. The largest number of participants come from Yerevan, followed by Kotayk and Lori. The gender distribution is highly unbalanced with girls outnumbering boys in all regions, but from the bar chart in Fig.1 it is clear that **in Yerevan the difference is less significant than in other regions.** However, as shown in Fig.2, when analyzing gender distributions in region's separately for STEM and non-STEM subjects, a different trend emerges. The gender disparity becomes almost insignificant in STEM subjects, and in certain regions, boys even outnumber girls. But the gap is significant for non-STEM subjects.

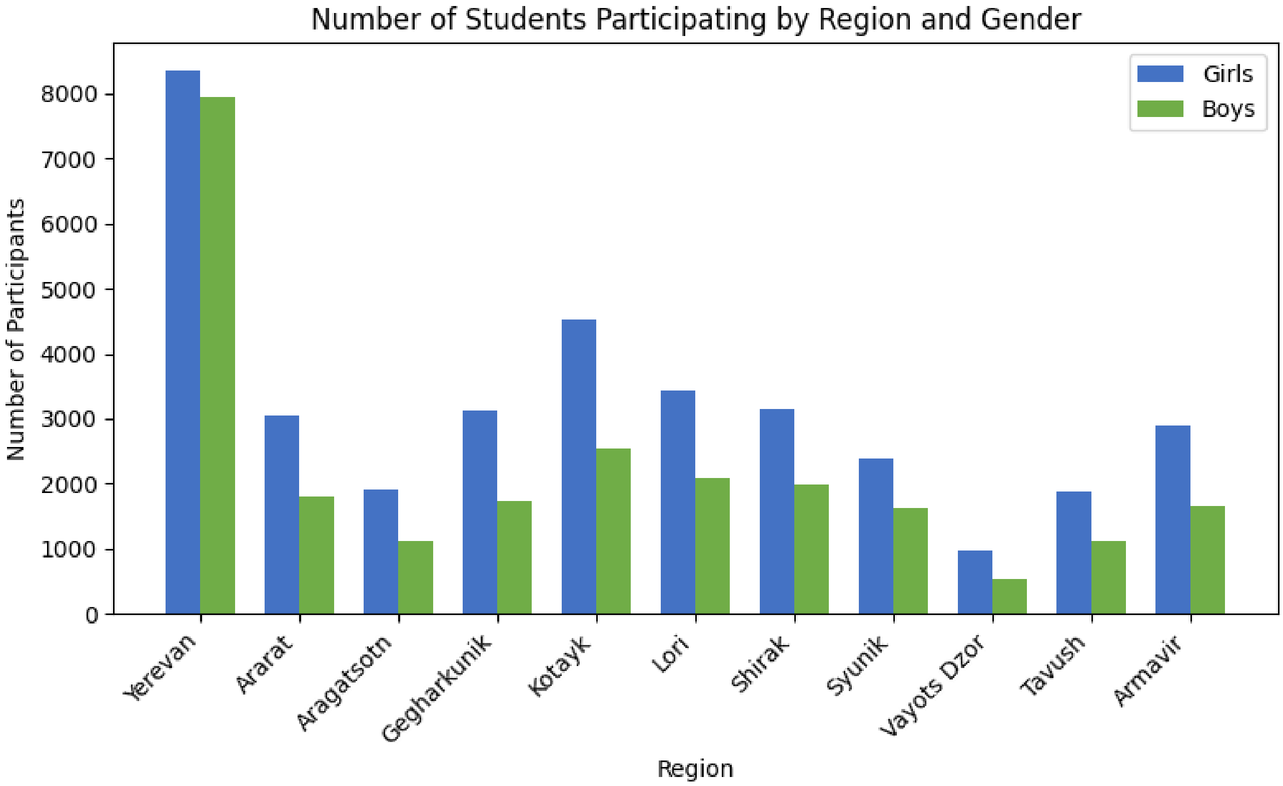

Figure 1: Distribution of participants by regions.

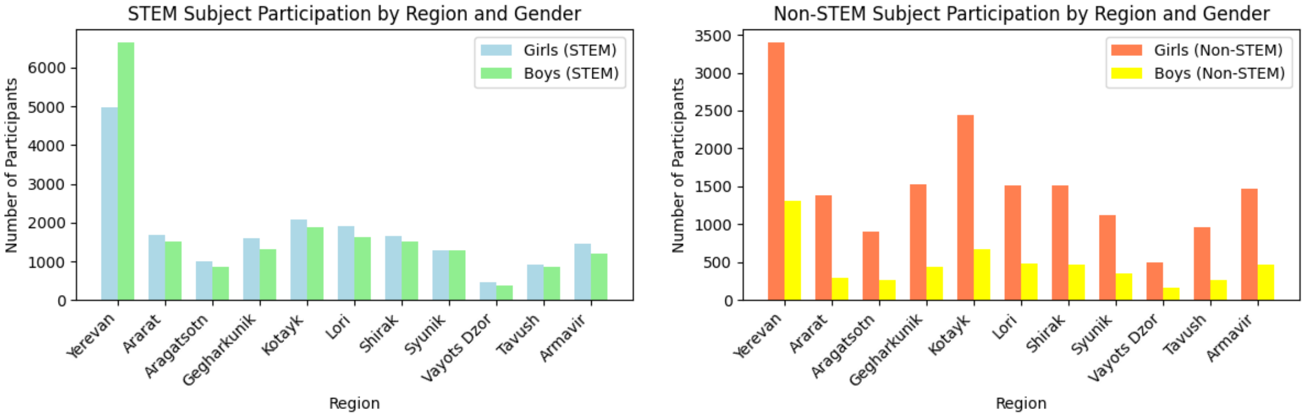

Figure 2: Distribution of participants by regions and gender regarding STEM/non-STEM.

Region	Girls	Boys	Total
Yerevan	8358	7952	16320
Ararat	3063	1800	4863
Aragatsotn	1913	1125	3038
Gegharkunik	3126	1736	4862
Kotayk	4527	2546	7073
Lori	3420	2096	5516
Shirak	3159	1990	5149
Syunik	2390	1624	4014
Vayots Dzor	973	544	1517
Tavush	1889	1125	3014
Armavir	2908	1647	4555
Total	**35726**	**24185**	**59911**

Table 1: Participants by region and gender (2025)

Subject	Count
Armenian Language	9125
Chemistry	2525
History	5188
English	7576
Physics	3493
Mathematics	32004
Total	**59911**

Table 2: Number of participants by subject (2025)

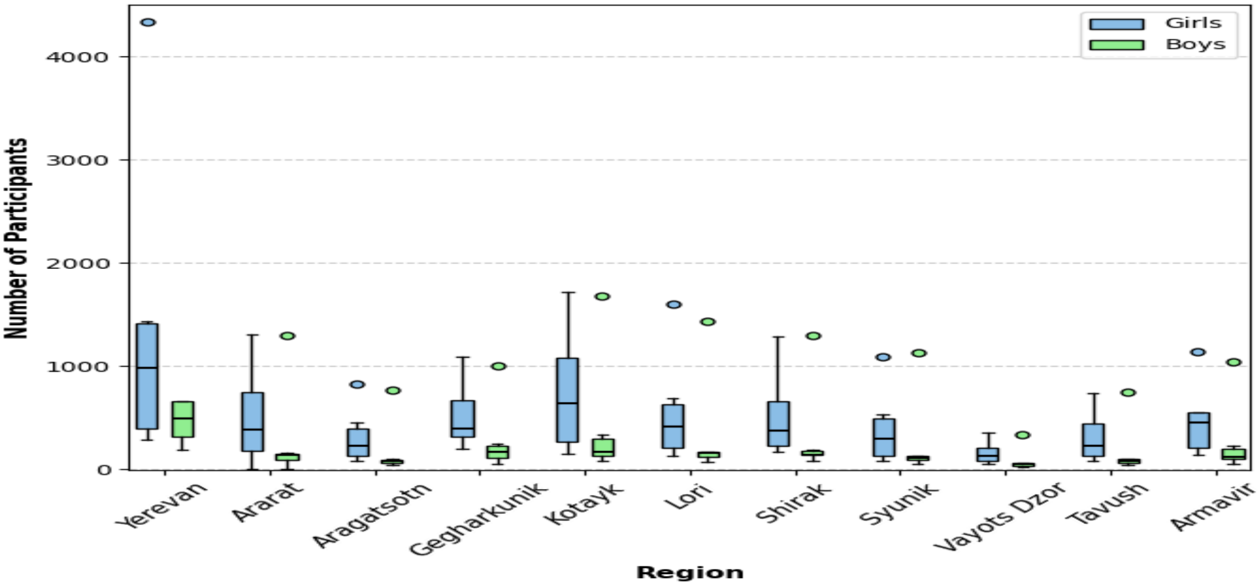

Figure 3: Boxplot of participation numbers based on region and gender.

The box-plot in Fig. 3 illustrates the distribution of participation numbers across different regions and subjects, categorized by gender. Upon comparing the box plot with the actual participation numbers, it is evident that most of the outliers are associated with the Mathematics Olympiad. Indeed, in terms of overall participation, Mathematics has the highest number of participants, followed by Armenian Language and English. This distribution of subject participation is also shown in Table 2.

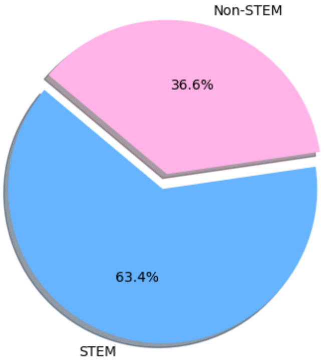

Figure 4: STEM vs non-STEM in Armenia

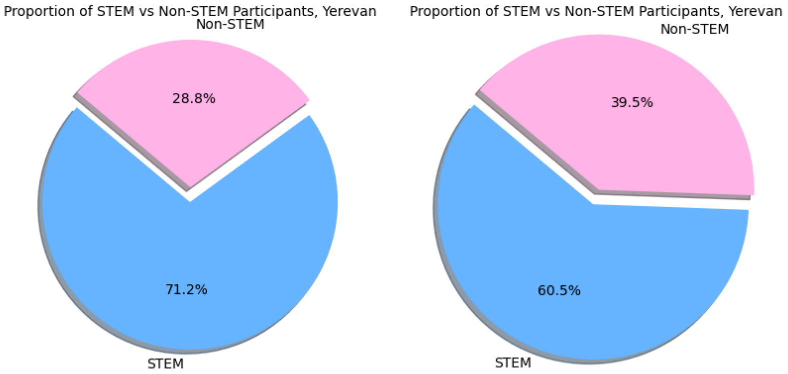

Figure 5: STEM vs non-STEM in Yerevan(left), regions(right)

By considering the STEM and non-STEM participation trends in general, we can see that both in Yerevan (Fig.5 left) and in regions (Fig.6 Right) STEM subjects are more popular among Olympiad participants, but in regions more students are prone to participate in Non-STEM subject Olympiads.

ANALYSIS METHODOLOGY AND RESULTS

In this section, the methodologies applied to each research question are explained and the obtained results are discussed. The implementations of the methods used and the details of the steps performed can be found on https://github.com/GorArzanyanAUA/OlympAnalyzer.

Q1: Does student participation in extracurricular academic activities significantly differ across regions?

Method: Chi-square Goodness-of-Fit Test (Pearson's χ^2 test)

We are analyzing a single categorical variable *region*, with multiple mutually exclusive values representing the origin of each student (e.g., Yerevan, Kotayk, Lori, etc.). Since we have a large number of observations in each region, the Chi-square Goodness-of-Fit test is appropriate for assessing whether participation levels differ significantly across regions.

This test compares the observed distribution of student participation across regions to an expected distribution derived from regional student population sizes.

We define the hypotheses as:

- **Null Hypothesis (H_0):** The distribution of Olympiad participation matches the distribution of the total student population across regions.
- **Alternative Hypothesis (H_1):** The distribution of Olympiad participation significantly differs from the expected distribution based on student population.

Chi-square Statistic:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed count for region , E_i is the expected count based on population proportion, and is n the number of regions.

The resulting χ^2 is compared with $\chi^2_{10,0.95}$ to decide whether to reject H_0 or not.

	Yerevan	Aragatsotn	Ararat	Gegharkunik	Kotayk	Lori	Shirak	Syunik	Vayots Dzor	Tavush	Armavir
Count	144 166	19 357	42 084	28 998	43 658	31 907	32 448	17 934	6 998	17 384	40 013
Proportion	0.339	0.046	0.099	0.068	0.103	0.075	0.076	0.042	0.016	0.041	0.094
Total											**424 927**

Table 3: Olympiad participation counts and proportions by region (2024).

The regional student population data shown in Table 3 is sourced from armstat.am.

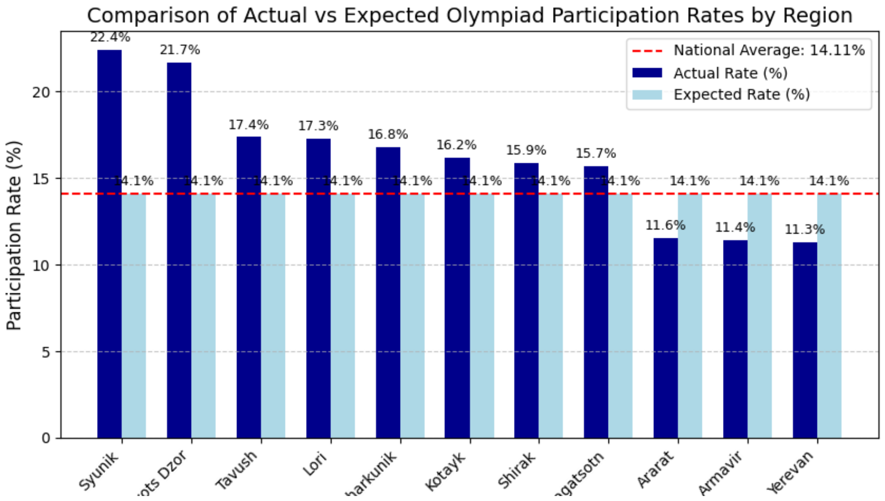

Figure 6: Observed and expected participation rates per region.

Results

Table 4 presents the observed (O_i) and expected (E_i) counts of Olympiad participants by regions in 2025, for a total of $N = 59\,911$ participants. The expected values are calculated using the true proportions shown in Table 3.

	Yerevan	Aragatsotn	Ararat	Gegharkunik	Kotayk	Lori	Shirak	Syunik	Vayots Dzor	Tavush	Armavir
O_i	16 310	3 038	4 863	4 862	7 073	5 516	5 149	4 014	1 517	3 014	4 555
E_i	18 026	2 420	5 262	3 626	5 459	3 989	4 057	2 242	875	2 174	5 003

Table 4: Observed and expected counts of Olympiad participants by region, $N = 59\,911$.

Chi-square test:

$$\chi^2 = 3103.35 \quad \text{which is significantly larger than} \quad \chi^2_{10,0.95} = 18.307$$

Since our calculated χ^2 value exceeds the critical value, we reject the null hypothesis (H_0) in favor of the alternative hypothesis (H_1) that the distribution of Olympiad participation significantly differs across regions.

Interpretation: The results indicate that Olympiad participation significantly differs across regions. The participation proportions do not align with the population proportions per region. Specifically, Fig. 6 further indicates that Syunik, Vayots Dzor, Tavush, and a few other regions exhibit significantly higher participation than expected, while regions such as Ararat, Armavir and Yerevan show participation rates below what would be expected based on their student populations. This finding is particularly interesting for Yerevan, and it can be further studied to understand what factors such as high competition or quality of public educational system in Yerevan is responsible for such situation.

Q2: Are Armenian students significantly more interested in STEM subjects compared to non-STEM subjects?

Method: Z-test for Population Proportion

We consider the student population with two mutually exclusive categories: STEM and non-STEM. Each student has a single category of choice and it can be modeled as a **Bernoulli random variable**, making the dataset suitable for modeling with the **Binomial distribution**. We **assume** independence in each student's subject choice. The large sample size justifies the normal approximation to the binomial distribution.

The appropriate statistical test for this scenario is the Z-test for Population Proportion. The test evaluates whether the observed proportion of students interested in STEM significantly differs from a hypothesized $p_0 = 0.5$, indicating no preference.

- **Null Hypothesis** (H_0): Students show no significant preference towards STEM or non-STEM subjects; $H_0 : p = 0.5$:
- **Alternative Hypothesis** (H_1): Students show a significant preference towards STEM or non-STEM; $H_1 : p \neq 0.5$:

Z-test Statistic: The test statistic is calculated as follows:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- $\hat{p}$ is the observed proportion of students interested in STEM.
- p_0 is the hypothesized proportion (0.5).
- n is the total number of students.

For a given significance level α, we reject H_0 if:

$$|Z| > Z_{1-\frac{\alpha}{2}}$$

Results

- **Z-test Statistic**: $Z = 65.9116$
- **P-value**: 0.000

Since the p-value is exceedingly small ($p < 0.05$), we reject the null hypothesis, concluding that there is a significant preference among Armenian students for STEM subjects over non-STEM subjects.

The sample proportion of students interested in STEM was calculated as $\hat{p} = 0.6346$. The **standard error (SE)** is 0.0020, and the **margin of error** was found to be 0.0038.

The **95% confidence interval** for the population proportion is $(0.6308, 0.6385)$.

Interpretation: The results of both the Z-test and the confidence interval estimation strongly indicate that Armenian students have a significant preference for STEM subjects, with the proportion of students interested in STEM being much greater than the 50% threshold of no preference. The confidence interval further confirms that this proportion lies between 63.08% and 63.85%, offering a clear view of the magnitude of the preference towards STEM.

Q3: Does student interest in STEM and non-STEM subjects differ significantly between students living in different regions?

Method: Chi-square Test of Independence

We analyze two categorical variables—*region* a nd *subject interest* (STEM vs non-STEM). Each student's participation is one observation in an $r \times 2$ contingency table. With large expected counts in each cell, the Chi-square Test of Independence is appropriate to assess whether subject interest is associated with region.

Hypotheses:

- **Null Hypothesis (H_0):** Subject interest is independent of region; the distribution of STEM vs non-STEM participation is the same across all regions.
- **Alternative Hypothesis (H_1):** Subject interest depends on region; the proportion of STEM vs non-STEM participation differs across regions.

Chi-square Statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where:

- O_{ij} is the observed count in the cell for region i and subject category j.
- $E_{ij} = \frac{(\text{row}_i \text{ total}) \times (\text{col}_j \text{ total})}{N}$, with N the overall sample size.

The obtained χ^2 value is compared with the value of the $\chi^2_{r-1,1-\alpha}$. If the obtained value is bigger, then we reject H_0.

Results

First, the contingency table was constructed.

Region	Yerevan	Ararat	Aragatsotn	Gegharkunik	Kotayk	Lori	Shirak	Syunik	Vayots Dzor	Tavush	Armavir	Col.T
STEM	11605	3182	1866	2904	3959	3513	3169	2552	857	1789	2626	38022
Non-STEM	4705	1681	1172	1958	3114	2003	1980	1462	660	1225	1929	21889
Row Total	16310	4863	3038	4862	7073	5516	5149	4014	1517	3014	4555	59911

Table 5: Observed Counts for STEM and Non-STEM Participation by Region

Region	Yerevan	Ararat	Aragatsotn	Gegharkunik	Kotayk	Lori	Shirak	Syunik	Vayots Dzor	Tavush	Armavir	Col.T
Expected STEM	10351.00	3086.26	1928.04	3085.63	4488.82	3500.68	3267.77	2547.45	962.75	1912.81	2890.79	38022
Expected Non-STEM	5958.99	1776.74	1109.96	1776.37	2584.18	2015.32	1881.23	1466.55	554.25	1101.19	1664.21	21889
Row Total	16310	4863	3038	4862	7073	5516	5149	4014	1517	3014	4555	59911

Table 6: Expected Counts for STEM and Non-STEM Participation by Region

Chi-square Statistic: $\chi^2 = 758.24$; p = 0.000

Since the p-value is exceedingly small ($p < 0.05$), we reject the null hypothesis, indicating a significant relationship between region and subject interest (STEM vs. non-STEM).

From the table, it is clear that the observed counts for STEM and non-STEM subjects in each region differ from the expected counts, which further supports the rejection of the null hypothesis. This result indicates that the distribution of subject interest (STEM vs. non-STEM) varies significantly across different regions.

Thus, we conclude that regional factors play a significant role in shaping student interest in STEM and non-STEM subjects in Armenia.

CONCLUSION

This report aimed to explore the trends in student interest in STEM and non-STEM subjects across various regions in Armenia, using data sourced from the Armenian Olympiad platform. We addressed three central research questions and employed statistical methodologies to derive meaningful insights from the data.

1. **Regional Differences in Extracurricular Participation:** The Chi-square Goodness-of-Fit test revealed that student participation in extracurricular academic activities varies significantly across regions in Armenia. This indicates that participation is not equally distributed and may be influenced by regional factors such as educational infrastructure, access to resources, and socio-economic conditions.

2. **STEM vs Non-STEM Interest:** The Z-test for Population Proportion showed a significant preference among Armenian students for STEM subjects over non-STEM subjects. The data confirmed that a larger proportion of students are inclined towards STEM, which reflects the emphasis on science, technology, engineering, and mathematics in Armenia's societal trends.

3. **Regional Differences in STEM vs Non-STEM Interest:** The Chi-square Test of Independence indicated a significant relationship between region and student interest in STEM vs non-STEM subjects. It was evident that regional factors play a substantial role in shaping students' preferences for STEM or non-STEM subjects, with certain regions showing a higher propensity for STEM participation while others leaned more towards non-STEM subjects.

While these findings provide a clearer picture of student interests in Armenia, it is evident that there are additional insights to be gained by integrating other indicators besides Olympiad participation, by considering factors, such as socio-economic status, school quality, and system effectiveness. Furthermore, analyzing data across multiple years would allow for a better understanding of evolving trends and the impact of policy changes on student interests.

This research lays the groundwork for future studies that could explore deeper into the factors influencing these trends, providing a more comprehensive view of how regional, demographic, and educational factors shape the academic interests of students in Armenia.

Bibliography

[1] National Statistical Service of the Republic of Armenia. *Statistical Yearbook of Armenia 2023*. Available at: `https://www.armstat.am/en/`

[2] Armenian Olympiad. *Olympiad Platform*. Available at: `https://www.olymp.am/`

[3] Gor Arzanyan. *OlympAnalyzer Repository*. Available at: `https://github.com/GorArzanyanAUA/OlympAnalyzer`

[4] Ross, S. M. (2010). *Introductory Statistics*, 3rd Edition. Pearson Education. ISBN: 978-0321629111.

APPENDIX A: DATA EXTRACTION AND PROCESSING

The dataset used in this study was compiled by scraping data from the official Armenian Olympiad website, olymp.am. The original data was presented in PDF format, from which tabular data was extracted using automated tools. The extracted tables contained the following fields: student ID, surname, given name, father's name, grade, region, school, and score.

Gender was not explicitly included in the original data. To enrich the dataset with gender information, unique given names were extracted, and a large language model (LLM) was used to classify each name as either male or female based on Armenian cultural and linguistic context.

Some names may have been ambiguously classified, and data entries that were incomplete or improperly extracted during the PDF parsing stage were discarded to maintain the integrity of the dataset. The resulting cleaned and enriched dataset provided the basis for the subsequent statistical analysis and visualizations.

Source code used for the preprocessing can be found here:

https://github.com/GorArzanyanAUA/OlympAnalyzer

Appendix B: Dataset summary

Region	**Tot**	**Girls**	**Boys**	**STEM**	**Non-S**	**Arm**	**Chem**	**Hist**	**Eng**	**Phys**	**Math**	**G-STEM**	**B-STEM**	**G-NonS**	**B-NonS**
Yerevan	16310	8358	7952	11605	4705	1674	478	937	2094	986	10141	6354	5249	3004	1701
Ararat	4863	3063	1800	3182	1681	956	282	0	725	302	2598	1544	1638	1719	1286
Aragatsotn	3038	1913	1125	1866	1172	542	137	333	297	142	1587	1024	842	889	283
Gegharkunik	4862	3126	1736	2904	1958	882	251	662	414	562	2091	1285	1220	1841	1171
Kotayk	7073	4527	2546	3959	3114	1055	296	570	1489	271	3392	1922	1462	1584	1530
Lori	5516	3420	2096	3513	2003	834	203	541	628	271	3039	1733	1780	1092	475
Shirak	5149	3159	1990	3169	1980	940	272	542	498	312	2585	1517	1648	1283	697
Syunik	4014	2390	1624	2552	1462	653	159	297	512	176	2217	1186	1066	1065	396
Vayots Dzor	1517	973	544	857	660	300	89	191	169	81	687	434	453	333	327
Tavush	3014	1889	1125	1789	1225	622	149	331	272	153	1487	806	681	734	491
Armavir	4555	2908	1647	2626	1929	667	209	784	478	237	2180	1305	1121	1228	701