What Factors at School Influence Student Academic Performance at University

Nazim Ibragimov
Department of Computer Science
Suleyman Demirel University
Kaskelen, Kazakhstan
nazim.ibragimov@sdu.edu.kz

Asmina Barkhandinova
Department of Computer Science
Suleyman Demirel University
Kaskelen, Kazakhstan
190103102@stu.sdu.edu.kz

Nurzat Shayakhmetov
Department of Computer Science
Suleyman Demirel University
Kaskelen, Kazakhstan
190107095@stu.sdu.edu.kz

Aruzhan Akkoziyeva
Department of Computer Science
Suleyman Demirel University
Kaskelen, Kazakhstan
190107012@stu.sdu.edu.kz

Sultanmakhmud Bazarbayev
Department of Computer Science
Suleyman Demirel University
Kaskelen, Kazakhstan
190103178@stu.sdu.edu.kz

Zhandos Tangatar
Department of Computer Science
Suleyman Demirel University
Kaskelen, Kazakhstan
190107047@stu.sdu.edu.kz

Abstract—**This research demonstrates the use of some statistical tests popular in Data Science applied in Social Sciences. The usage is shown on the case of investigation of what factors might affect students' performance. Most recent studies focus mostly on short-term results such as performance in high school. This paper examines these factors in the long-term run, taking the factors in schools that might affect a person in the future university. Analytical approach is used to derive most significant factors. The results approve and disapprove some stereotypes that were present before the research. The derived factors give motivation for further investigation of educational success.**

Keywords—education, inequality, factors of success, application of statistical tests, students' performance, university performance.

I. INTRODUCTION

In the last decade the term of 'Data Science' was widely used in business, engineering and science for problem-solving. Indeed, analytical approach and working with Big Data can significantly improve business and industrial performance, by extracting meaningful insights and delivering understanding in what direction the company should aim. However, these tools can also help with conducting researches in Social Sciences, such as Sociology, Political Science, History etc., since most of the researches deal with a big amount of data obtained after some surveys or interviews. Usually, the aim of the surveys is to identify which factors affect people's choices or actions. Even people's behavior or characteristics can be presented in form of numerical vectors and be analyzed through the prism of Data Science to extract these valuable factors.

To show how this approach can help in conducting Sociology research, we decided to conduct our own research about one of the keystones of any nation – education. Nowadays education is viewed as a social elevator, which can in theory equalize all citizens regardless of their origin. Unfortunately, it is not always true, since schools can vary in terms of sponsorship, equipment, location and other factors. In addition to this inequality, there occur many prejudices regarding some type of schools, and according to these opinions the parents decide to which school send their child. Parents select future schools very carefully hoping that this investment will pay off in university and in the child's career. Since the school might affect the child's future, we decided to derive factors in school that affect student's performance at higher educational institutions.

II. DESIGN OF RESEARCH

In order to find out what factors at school affect a student's academic performance at university we created a questionnaire for students. We designed questions based on literature and previous researches adapting the results to our Kazakhstani educational system and our aim of research.

First of all, we asked the students their Grade Point Average (GPA), because GPA is an objective indicator of a student's success. We asked exclusively GPA of the first semester during the freshmen year, since it can also demonstrate a student's ability to adapt to new academic environment.

Since our country, Kazakhstan, is a multinational country, accordingly we have many nations and languages. One of the articles considered is Azattyq Ruhy, 'Kazakhstani Parents Started Giving Their Children to Russian Classes' [1]. Most of the students in the country are taught in two languages. These are Russian and Kazakh languages. There are also students studying other languages like Uighur, Uzbek, Tajik and so on. But they are a minority. The article examines which school graduates are more educated. According to the results of the study, the Russian classes are more educated. This is the result of 2011. We were interested in whether things are the same now. That's why we asked the students in what language they studied.

Since we conducted the research in Kazakhstan, we mainly relied on articles that are related to Kazakhstan. One of these articles is a Forbes article about the international PISA program [3]. This article discusses the results and factors that affect the education of schoolchildren. Here we noticed some interesting facts. Girls scored more points than boys. That is, gender can also influence education. The Internet in schools is mentioned here. About 90 percent of schools are provided with the Internet, but its quality is poor. This means that not all schools have good Internet. We thought that the access to the school Internet can also affect academic performance.

The article by Forbes [3] examines the main achievements of Kazakhstan in PISA and the main reasons for the backlog of Kazakhstani schoolchildren relative to schoolchildren of other countries. PISA touches on the socio-economic status of parents of students. In the latest PISA-2018 study, it was revealed that in Kazakhstan, socially and economically prosperous pupils outperformed disadvantaged pupils in reading by 40 points. We agree that the socio-economic status of parents can affect the educational process of a child. Parents

who earn well, as a rule, have a good education, which may affect the availability of books in the house or the child's greater interest in knowledge. Besides, wealthy parents can hire tutors for their children, pay for sport and other hobbies, which can affect the child's performance. The correlation between children's involvement in extracurricular activities and performance is also noted in 'The Factors Influencing Students' Performance at Universiti Teknologi MARA Kedah, Malaysia' [4]. As a result, all these factors (parents' education and wage, tutors, sports and other hobbies) were considered in the survey.

In the article "The division of income between School districts and Inequality among students Achievement" [5], inequality of American students are considered. Here we are talking about children from high-income and low-income families, their race and difference in sponsorship of schools. The author explains how these factors affect the education of children. Unfortunately, we were not able to include all these factors in our questionnaire because we do not have a social problem by race. But we may have a problem with the difference in education between the village and the city or even between district in the same region. So, we added a question about the place of school.

Our topic is what factors at school affect a student's academic performance at the university, so it is logical to assume that the average grade in school subjects will play at least some role, so it is the first option from our team.

As we know, there are many schools in Kazakhstan such as BIL, NIS, RPhMS, it is interesting how graduates of these schools prove themselves at the university, so the second option is from which of these BIL/NIS/RPhMS schools was a student. In addition to these schools, there are others that are divided into ordinary, gymnasiums and lyceums. There are also private schools, so we added gymnasiums/lyceums/ordinary/private schools as options.

We also assumed that knowledge of English would facilitate the student's academic activities. Therefore, we added the level of English language proficiency to the list of options.

With all these options, we conducted a survey among students of several universities, processed the data and obtained the following results.

III. DATA AND METHODS

74 students took part in our survey, and only 69 results were counted as valid. All of them are students of 2,3 and 4 courses from different universities of Kazakhstan, such as Suleyman Demirel University, Kazakh-British Technical University, Nazarbayev University and Almaty Management University.

As it was mentioned above in research design part, we asked our respondents not only their GPA of first semester, but also their gender, place of living, language of school they attended, form and type of school, whether they were busy with extra lessons, sport or art hobbies, Internet availability at school, their English level by the end of high school, their school performance, parents' income and parents' education.

In the following sections we want to describe our respondents.

Fig. 1 Regarding gender, 46 were male, and 23 were female, which is 66.7% and 33.% respectively. By place of living we have 57 town folks and 12 village residents. More people were studying in Kazakh than in Russian: 37 versus 32. For school type in Fig. 2 we had Gymnasium/lyceum attended 29, ordinary schools - 22, NIS/BIL/RPhMS - 17 and private schools (like Miras or Galaxy) only 1. Which is surprising, since 8 people thought they attended private school, and other 61 attended state schools (see Fig. 3).
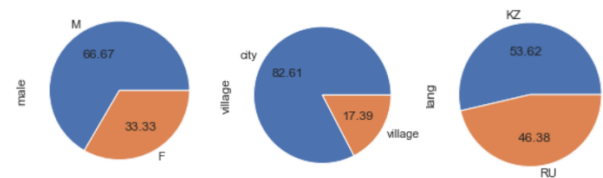

Fig. 1. Gender, place of living and language of school distributions.
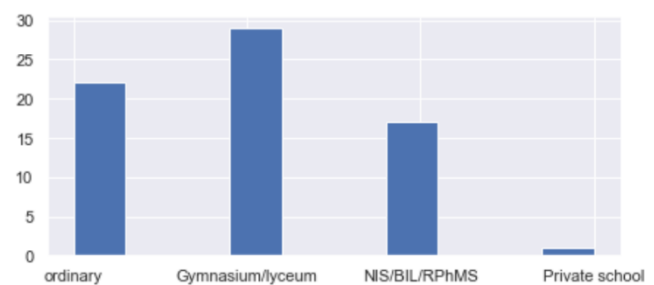

Fig. 2. Type of school distribution.

Fig.3. Type of property of school distribution.

Roughly equal proportion were attending private tutors and were not: 34 for 'yes' and 35 for 'no'. A bigger difference was for sports: 30 people were going to sport clubs whereas 39 were not. In the art section (like painting, theatre, handwork) only 27 were interested, while the other 42 were not. So, as we see, having extra lessons beyond schooling is more popular than having non-educational hobbies.

Surprisingly, 45 people had internet at school, 24 did not.

School performance: 34 and 32 had Good and Excellent performance respectively, only 3 had Satisfactory marks (see Fig. 4).
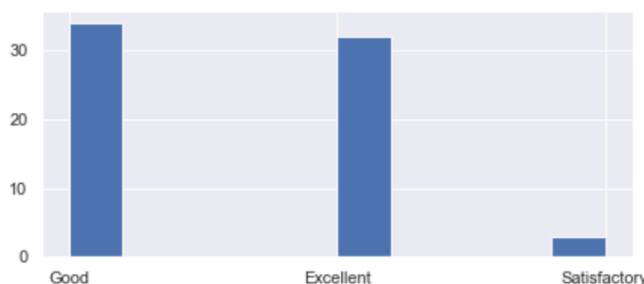
Fig. 4. School performance distribution.

On Fig. 5 we had a very wide distribution in English level: A0-A1 level had 14 students, A2 - 9 people, B1 and B2 both had 19, and C1-C2 - 8.
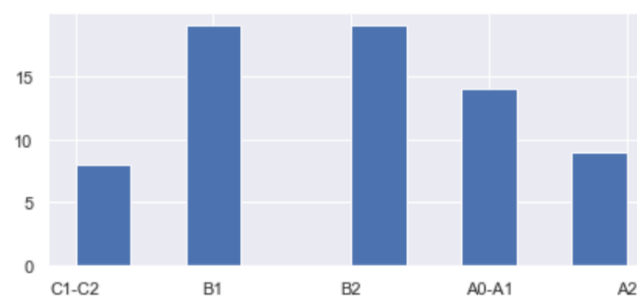

Fig. 5. English level distribution.

Since parents' state could have a huge impact on students, we also asked about their education and average wage (see Fig. 6. and Fig. 7.). Most parents had higher education (60 vs 9) and the most common wage was in the range 200 000 - 300 000 tenge.

Fig. 6. Parents' education: 'yes' if at least one parent had college diploma, 'no' if none of parents attended college.
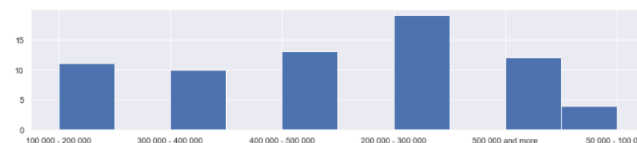

Fig. 7. Parents' income during high school.

Now, after we described all the data we had, we try to estimate which of these characteristics of school past can affect future performance in university.

In statistics we have several tests to estimate which parameters have the most weight on our dependent variable (GPA in university in the first semester).

First test is the P-values test. We will apply the 5% cut-off rule, which means that parameters with p-values less than 0.05 (5%) are more statistically significant.

Here is Table 1 with coefficients and p-values. As we see, such variables as English at B2, English at C1-C2, Excellent and Good performance at high school are statistically significant for GPA in university.

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.9513	0.312	6.248	0.000	1.324	2.579
male	-0.2505	0.186	-1.349	0.183	-0.623	0.123
village	0.1212	0.281	0.431	0.668	-0.444	0.686
lang	-0.1860	0.216	-0.862	0.393	-0.620	0.248
private	-0.0208	0.335	-0.062	0.951	-0.694	0.653
tutors	-0.0531	0.190	-0.280	0.781	-0.435	0.328
sport	0.1114	0.210	0.531	0.598	-0.310	0.533
art	-0.2938	0.205	-1.430	0.159	-0.707	0.119
internet	0.0294	0.196	0.150	0.881	-0.364	0.422
edu_parents	0.1058	0.268	0.395	0.695	-0.433	0.645
wage	-5.199e-07	5.95e-07	-0.874	0.386	-1.72e-06	6.75e-07
type_Gymnasium/lyceum	0.3783	0.203	1.864	0.068	-0.029	0.786
type_NIS/BIL/RPhMS	0.1202	0.251	0.480	0.634	-0.384	0.624
type_Private school	1.0888	0.615	1.770	0.083	-0.147	2.325
type_ordinary	0.3640	0.249	1.459	0.151	-0.137	0.865
english_A0-A1	-0.0659	0.192	-0.344	0.733	-0.451	0.319
english_A2	0.2834	0.290	0.978	0.333	-0.299	0.865
english_B1	0.2779	0.177	1.567	0.124	-0.079	0.634
english_B2	0.5455	0.189	2.886	0.006	0.166	0.925
english_C1-C2	0.9105	0.243	3.741	0.000	0.421	1.400
gpa_school_Excellent	0.8982	0.152	5.905	0.000	0.593	1.204
gpa_school_Good	0.5753	0.170	3.386	0.001	0.234	0.917
gpa_school_Satisfactory	0.4778	0.368	1.300	0.200	-0.261	1.216

Table 1. Coefficients and p-values (P>\|t\|).

Another test, such as Lasso Regression will shrink all insignificant variables to 0, while setting significant ones to proper values greater than 0. On Fig. 8 the Lasso test showed that English at A0-A1 and C1-C2 levels, Excellent performance at high school, art and gender are significant. However, art and gender might be misleading since our data is imbalanced, i.e., we have too few females and too few art people. Also, we see that English level and Excellent performance are still meaningful for future GPA at university.
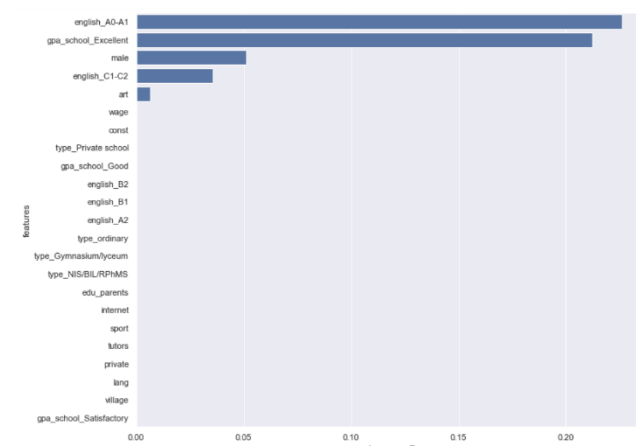

Fig. 8. Lasso Regression test shrinks non-significant values to 0.

Third test is to fit the Linear Regression model, which will try to predict Y (GPA at college) looking at data X (all other metrics). It will have the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i$$

Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable

Fig. 9. Linear Regression equation.

Then we will look at coefficients β's: the bigger the coefficients (the 'weights'), the more significant the X (the metrics like English level and so on).

After fitting the model, we have in Fig. 10 that Private School, English levels at B2, C1-C2, Excellent performance

have more 'weight' on future GPA. Also, the place of living could affect the performance.

	Coefficient
type_Private school	6.009576e-01
english_C1-C2	5.201947e-01
gpa_school_Excellent	2.477412e-01
english_B2	1.552082e-01
village	1.211566e-01
sport	1.114048e-01
edu_parents	1.058424e-01
internet	2.941340e-02
const	0.000000e+00
wage	-5.199016e-07
private	-2.076515e-02
tutors	-5.312671e-02
gpa_school_Good	-7.512335e-02
english_A2	-1.069031e-01
type_Gymnasium/lyceum	-1.095431e-01
english_B1	-1.123827e-01
type_ordinary	-1.238260e-01
gpa_school_Satisfactory	-1.726178e-01

Fig. 10. Linear Regression Coefficient in descending order.

The last test to try is Correlation Matrix. A correlation matrix is simply a table which displays the correlation. The measure is best used in variables that demonstrate a linear relationship between each other. As we see in Table 3, GPA is most correlated with Excellent performance at school, English at C1-C2, A2 and B2 levels, having extra lessons with tutors.

```
gpa                        1.000000
gpa_school_Excellent       0.289497
english_C1-C2              0.200991
tutors                     0.101124
english_A2                 0.078511
english_B2                 0.063493
type_Private school        0.050088
sport                      0.048814
type_Gymnasium/lyceum      0.045345
edu_parents                0.015328
village                    0.012915
type_ordinary              0.002382
english_B1                -0.021995
lang                      -0.022968
wage                      -0.023771
private                   -0.049217
type_NIS/BIL/RPhMS        -0.068411
internet                  -0.084145
art                       -0.155846
male                      -0.169619
gpa_school_Satisfactory   -0.182062
gpa_school_Good           -0.214502
english_A0-A1             -0.271849
```

Table 3. Correlation matrix shows the correlation for GPA with other variables.

To sum up, all 4 tests showed that students with Excellent performance at school, English at a very good level will boost the student in university. Other variables that some tests showed were attending Private schools, tutors, place of living and art hobby.

P-value test	Lasso Regression test	Linear Regression test	Correlation Matrix
English B2 English C1-C2 Excellent at school Good at school	English A0-A1 English C1-C2 Excellent at school Gender Art	English B2 English C1-C2 Excellent at school Private school Place of living	English B2 English C1-C2 English A2 Excellent at school Tutors

Table 4. Comparing the results from different statistical tests.

IV. RESULTS

Excellent students might have better GPAs at college thanks to good time management acquired at school. English at a high level is a good instrument at college with programs in English. Other factors that appeared in tests (Private schools, tutors, place of living) could be caused by parents' income, which we can test too.

Let's look at average GPA among Excellent, Good and Satisfactory marks at school (Fig. 11). As we see, A-students have higher GPA than other groups (3.16 average GPA for Excellent, 2.81 for Good and 2.39 for Satisfactory). So we can conclude that performance at school can affect university performance significantly.
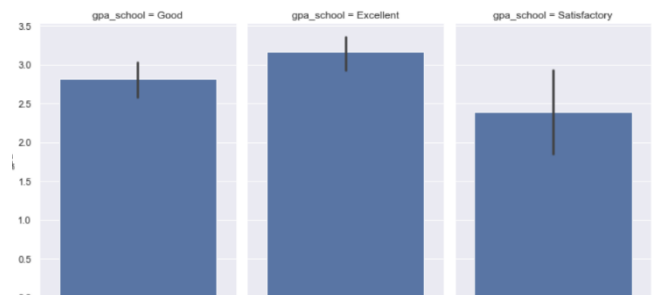

Fig. 11. GPA dependence on school performance.

Let's compare GPA among English levels. On Fig. 12 we again see a strong link between level and GPA: people with level at C1-C2 have higher GPA (3.32) and people with English at level A0-A1 have the lowest GPA (2.6 on average). So, the significance of the English level is proved.
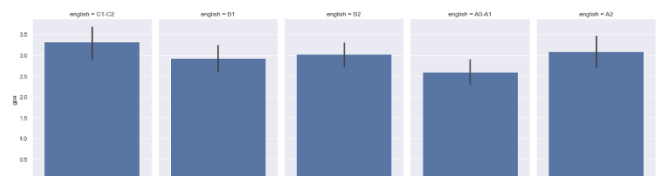

Fig. 12. GPA dependence on English level.

One of the tests also showed the significance of the type of school. We cannot conclude a strong relationship, since we had only 1 respondent who chose 'Private school' type. Moreover, NIS/BIL/RPhMS showed lower GPA than students from ordinary and gymnasium/lyceum schools (Fig. 13).
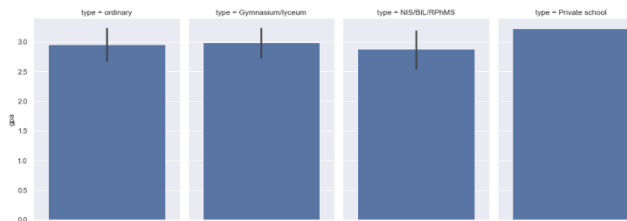
Fig. 13. GPA dependence on school type.

Another test showed the significance of having private tutors. On Fig. 14 tutors really could improve a student's future performance (2.89 vs 3.02 average GPAs).
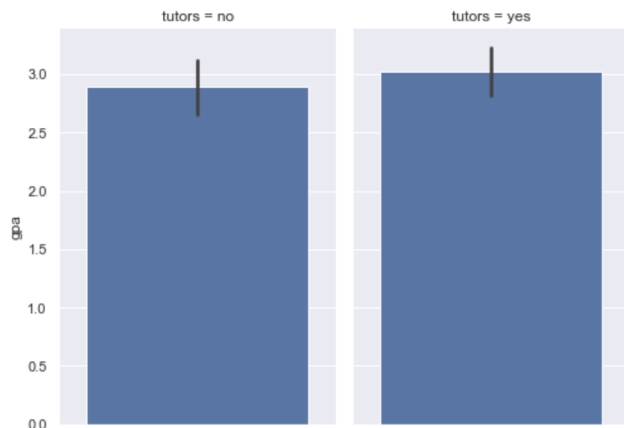

Fig. 14. GPA dependence on having private tutors at school.

Another test showed the significance of gender. Indeed, girls have a higher GPA (3.11) than boys (2.88) according to Fig. 15.
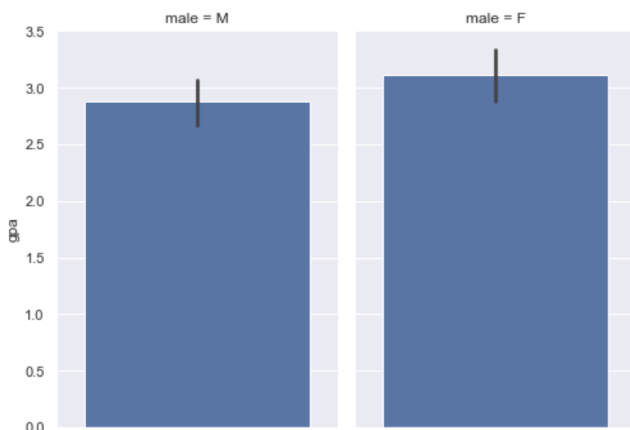

Fig. 15. GPA dependence on gender.

What is interesting, art hobbies were affecting GPA in the opposite way: art people had a lower GPA (2.82) than other people (3.03) on Fig. 16. Maybe art people were spending more time on their professionalism in art than in their college studies.
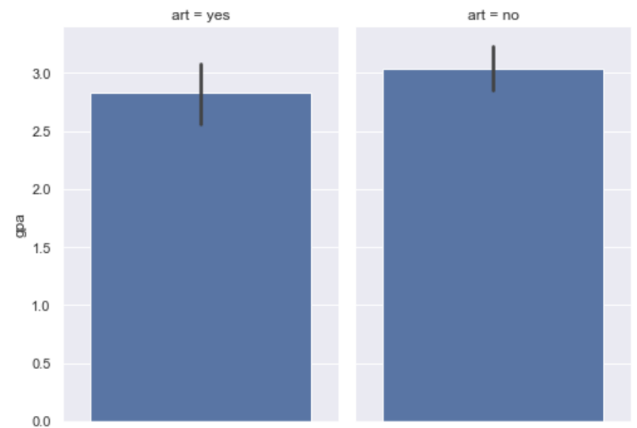

Fig. 16. GPA dependence on attending art classes.

Even though our tests (p-values, Lasso Regression, Linear Regression and Correlation matrix) did not highlight other metrics as significant, we decided to check all metrics to find interesting correlations.

For example, in the beginning of the article we were referring to Azattyq Ruhy article [1], that parents were giving their children to Russian classes rather than to Kazakh, since there are some prejudices about the quality of education in Kazakh language. As we see, average GPAs are roughly on the same level (2.94 and 2.97 for Russian and Kazakh schools on Fig. 17). After further study it can be proved that the language of school does not really affect pupils' future performance.
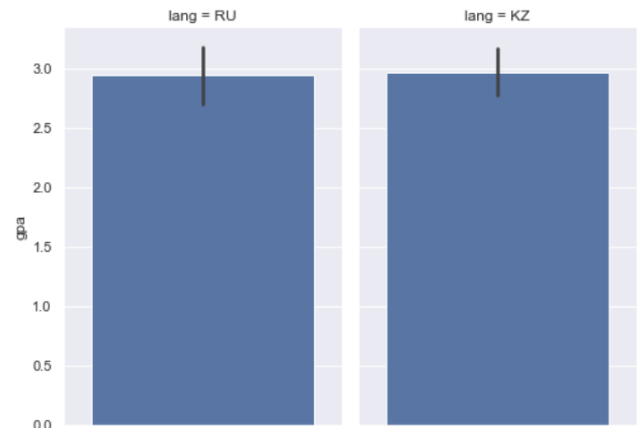

Fig. 17. GPA dependence on the language of studies in school.

We also see that place of living did not cause a big difference in GPA for city and village (2.95 and 2.97 on average on Fig. 18). However, the village has a greater deviation, thus not as stable as the city in GPA.
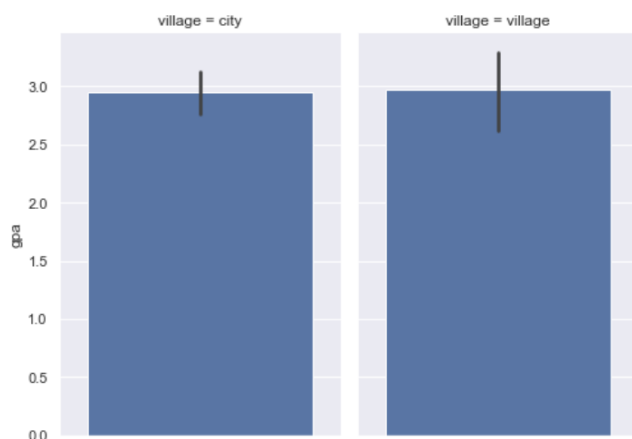
Fig. 18. GPA dependence on place of living.

Another interesting correlation showed sports: people who were participating in sport classes were more successful than those who did not (2.99 vs 2.92 on Fig. 19). Maybe in the further investigation, we will see either a good effect on students' health and thus better grades, or that sport trains discipline in students.
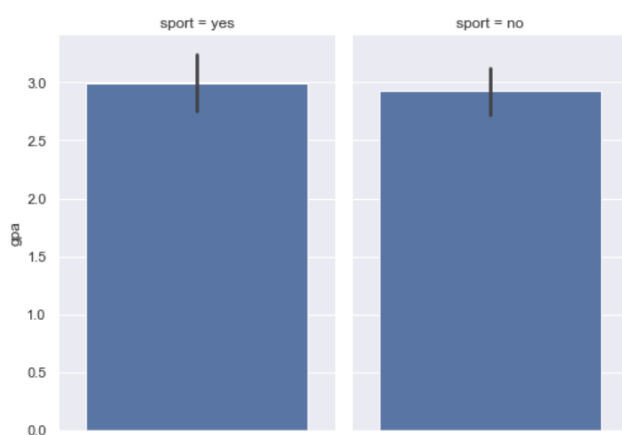

Fig. 19. GPA dependence on sport.

We also cited an article in the beginning [4] that parents' status may affect school performance. As we already noticed, school performance affects college GPA, so we tried to see the correlation between parents' status and GPA at college. As we see on Fig. 20, if at least one parent had higher education, the student's GPA was slightly higher (2.96) than those whose parents did not attend college (2.93). However, GPA is more stable for those whose parents had diplomas, whereas GPA for non-diploma parents fluctuates drastically.
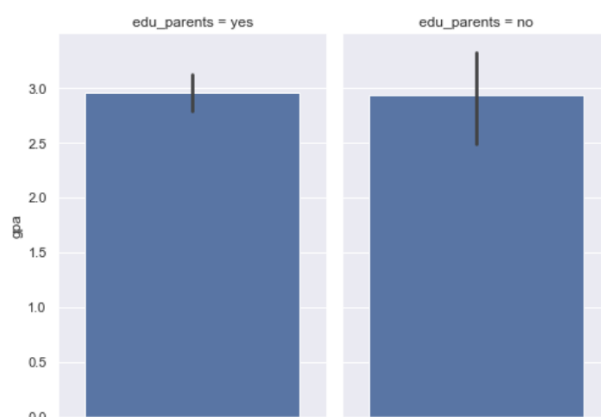

Fig. 20. GPA dependence on parents' education.

As we see on the graph on Fig. 21, students whose parents' income was in the range of 400 000 - 500 000 had GPAs a bit higher than others on average. However, it does not grow linearly: students with house income 500 000 and more perform less on average. So, we cannot conclude a strong link between parents' wages and students' future grades, as we were suspecting at the beginning of the article [4].
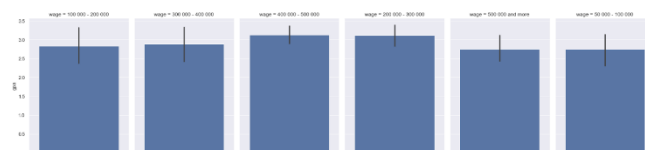

Fig. 21. GPA dependence on parents' income.

V. Conclusion

We used analytical approach to identify which factors might be most significant in the research. Mostly the results of 4 different tests coincided.

The results were interesting, we found out the type of school was not important, that on average all students no matter from what kind of school they came from, have approximately the same GPA. The interesting thing that we found out is that GPA is not dependent on the language of school. Also, we do not find a big difference in GPA of students from village and city, but little preponderance among students who are interested in sport, compared to students who are not interested. Thus, we disapproved some prejudices about the abilities of students from that or another school.

The main factors that impacted the success in college were knowledge of English and performance in school: A-students could adapt better for high load in university, thus achieving better grades.

By gender, we find out that there is little overweight GPA for women. Their interest in art is also slightly overweight to people who are not interested in art.

But in fact, of the salary of parents, we find out that the higher GPA has students whose parents have a salary above 200-300 thousand tenge. So concluded, that assumption in the article [5] that parents' status has an effect on students' success is valid for our situation.

The paper provides a framework how usage of tools from Data Science can provide valuable insights for future investigation. E.g., since school performance affects future student's success, we could continue the research with investigating which features helps A-students to maintain high grades.

REFERENCES

[1] Azattyq Ruhy, 'Kazakhstani Parents Started Giving Their Children to Russian Classes', 2021.

[2] Forbes,'Why Language of Education is One of the Factors of Educational Inequity', 2021.

[3] Forbes, 'Why Kazakhstani Pupils Failed International Examination PISA', 2020.

[4] Norhidayah Ali, 'The Factors Influencing Students' Performance at Universiti Teknologi MARA Kedah, Malaysia', 2009.

[5] Ann Owens, 'Income Segregation between School Districts and Inequality in Students' Achievement', 2018.

[6] CAMPONOTES, 'Factors Influencing National Systems of Education', 2013.

[7] Makhmutova L. R., 'Influencing Factors of Performance of Students in Universities', 2018.

[8] Kosheleva G.V., 'Influencing Factors of Performance of Students in Universities', 2017.

[9] BSTUDY.NET, 'Factors of Success of Students', 2017.

[10] Shmarikhina E. S. 'Investigation of the factors of students' performance', 2018.