VILNIUS UNIVERSITY

Life Sciences Center

Rimantė ŽEDAVEINYTĖ

Final project of the course “Synthetic Biology”

Analysis of the dataset “Students Performance in Exams”

Vilnius, 2020

Introduction

For my final project of the course “Synthetic Biology” I have chosen a dataset „Students Performance in Exams“ (link: https://www.kaggle.com/spscientist/students-performance-in-exams#StudentsPerformance.csv) from database www.kaggle.com. It shows the grades of 1000 high school students from the US in three exams – Maths, Reading and Writing – together with additional factors that could possibly affect them.

There are no missing values in this dataset as it is fictional and automatically generated every time someone downloads it. As it is created for educational purposes, it may be that some tendencies are kept within the data that are based on some known correlation traits between given categories (however, it was not stated anywhere). Therefore, I found it interesting to check whether some difference in exam scores could be found based on students’ family background and lifestyle.

No information is given about possible biases or principles of generating this dataset, so perhaps the only possible external validation of these findings would be comparison with the real data. This, however, wouldn’t be easy, as it is not stated which grade students, what type of school (private or state) and which part of the US this dataset should describe.

Descriptive analysis

This dataset consists of 1000 independent measurements (rows), each of which gives 8 variables (columns). 5 of these variables are discrete nominal (categorized as “objects”):

- **Gender** (2 groups): male, female;
- **Race/ethnicity** (5 groups): groups A, B, C, D, E (no additional information about them);
- **Parental level of education** (6 groups): bachelor's degree, some college, master's degree, associate's degree, high school, some high school.
- **Lunch** (2 groups): standard; free/reduced;
- **Test preparation course** (2 groups): none, completed.

The remaining 3 variables are continuous (category: “integers”) and gives the numerical results of the taken exams:

- **Math score;**
- **Reading score;**
- **Writing score.**

As there are way more observations than 15, we can do statistical tests with the given data. At first, I aimed to check whether the data generated for three different exams results are distributed normally. As we can see from the boxplot in Figure 1, distributions for all three exams seems normal.

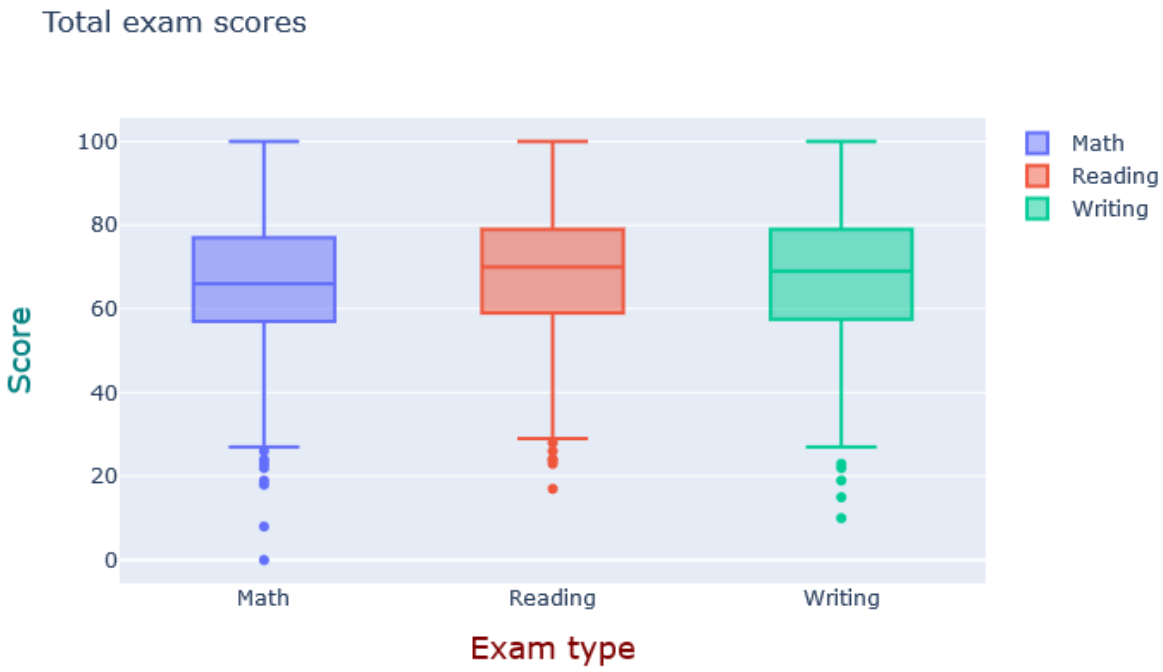

Figure 1 Distribution of exams scores based on all observations.

Looking at the numerical values of these exams scores we can confirm that their scores distribution is not only normal, but very similar, with their mean score ranging between 66.1-69.2, standard deviation being in the range ±14.6-15.2. In all exams maximum of 100 points was reached, minimum differs from 0 points in Math and 17 points in Reading exam (Table 1).

Table 1 Description of the initial data (exam scores).

	math score	**reading score**	**writing score**
count	1000.0	1000.0	1000.0
mean	66.1	69.2	68.1
std	15.2	14.6	15.2
min	0.0	17.0	10.0
25%	57.0	59.0	57.8
50%	66.0	70.0	69.0
75%	77.0	79.0	79.0
max	100.0	100.0	100.0

Inferential analysis

We can look at the correlation between the results of different exams. If we compare them in pairs, we see that the correlation is quite strong in all cases (R^2 being > 0.8, Figure 2). However, the strongest correlation could be seen between reading and writing scores (R^2 = 0.95), which seems logical, as both exams test students' linguistic abilities.

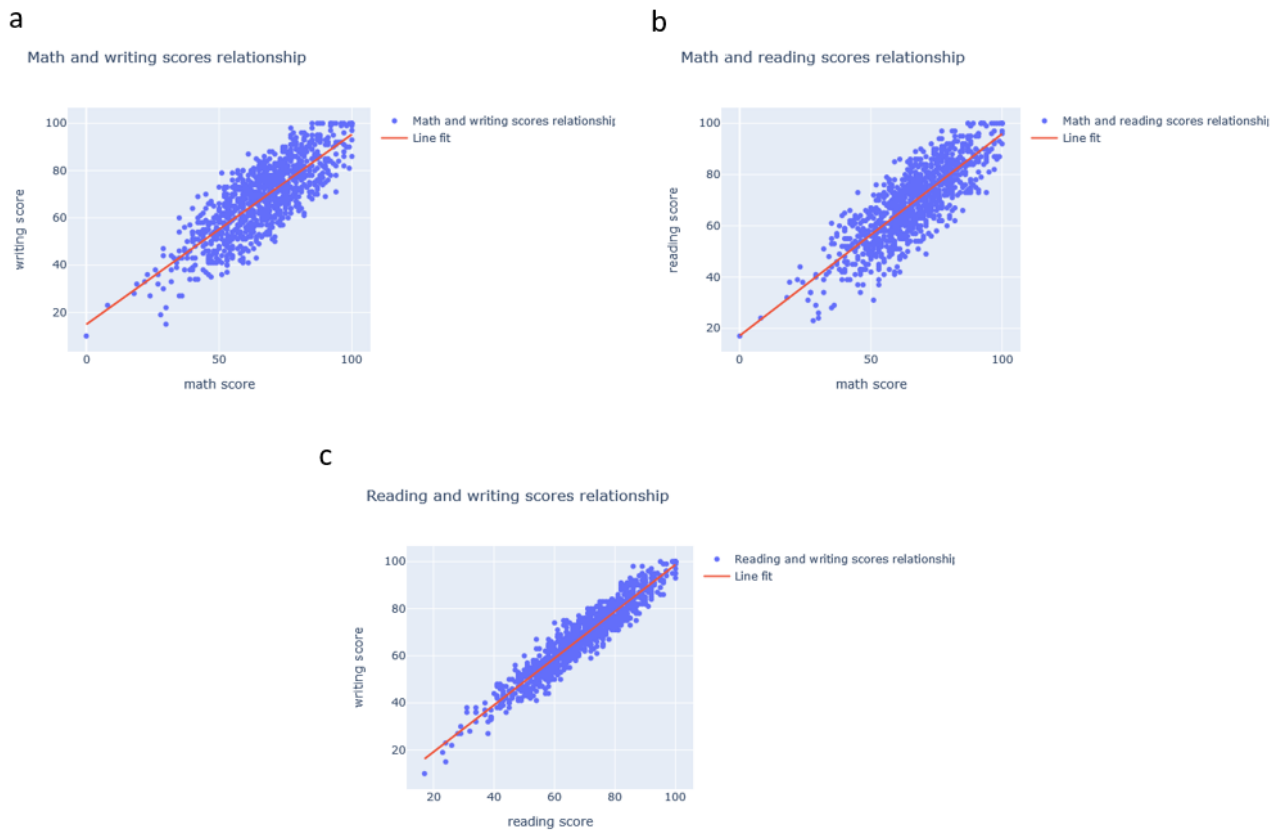

Figure 2 Linear regression plots for all three exams pairs. (a) Correlation between Math and writing scores (Pearson R value is: 0.80264 with a p-value of: 3.37603e-226). (b) Correlation between Math and reading scores (Pearson R value is: 0.81758 with a p-value of: 1.78775e-241). (c) Correlation between Reading and writing scores (Pearson R value is: 0.95460 with a p-value of: 0.00000e+00).

This dataset consists of three categorical variables – "gender", "lunch" and "test preparation course" – that each have only two unique values. Therefore, we can compare the exam results between these groups doing t-tests (and visualise drawing boxplots)

Grouping based on gender shows statistically significant differences between males and females. From the boxplot (Figure 3) we can see that while males are getting higher scores in Math, females do better in linguistics – reading and writing.

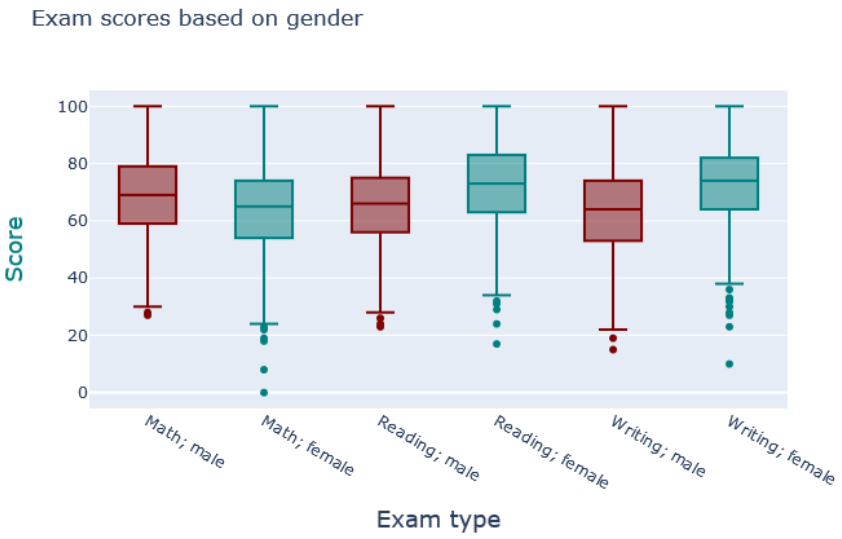

Figure 3 Exam results comparison between males and females.
P-Value for the Math score: 9.120185549328822e-08
P-Value for the Reading score: 4.680538743933289e-15
P-Value for the Writing score: 2.019877706867934e-22

Students can also be grouped based on their lunch type: standard or free/reduced. T-tests show statistical significance for all exams, and from Figure 4 we can say that reduced lunch can be associated with lower achievements during testing in all three subjects.

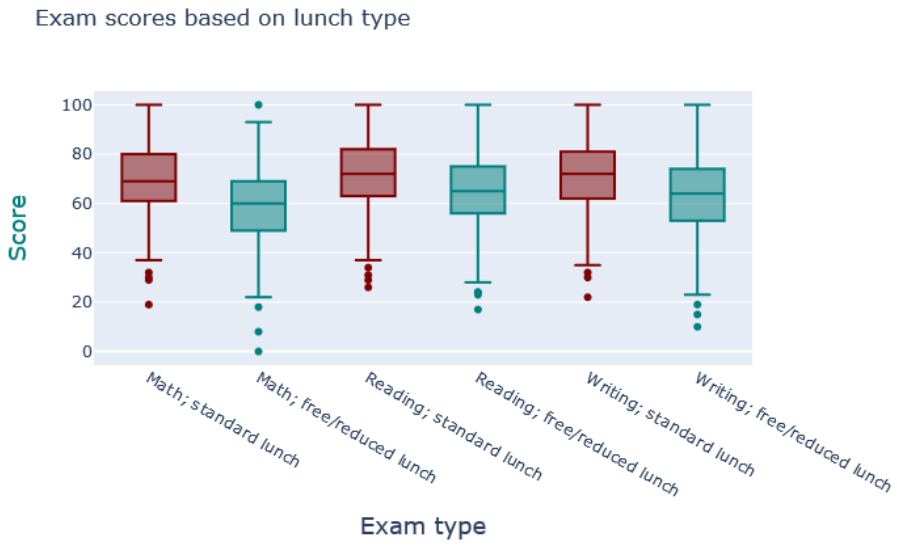

Figure 4 Exams results comparison between students, getting different lunch types: standard or free/reduced.
P-Value for the Math score: 2.4131955993137074e-30
P-Value for the Reading score: 2.0027966545279011e-13
P-Value for the Writing score: 3.186189583166477e-15

There is also information given whether a student completed a preparation course before taking an evaluated test. Again, t-test shows that there is a difference between those who took the preparation course and those who did not. From Figure 5 we can again see, that increase in results after taking the course is visible for all exams, having the biggest effect on writing exam scores.

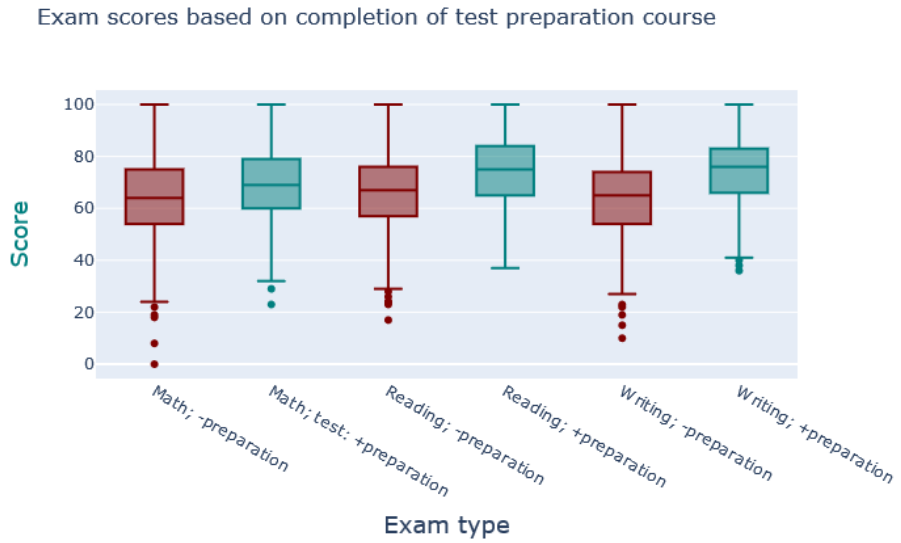

Figure 5 Exams scores based on completion of a preparation course.
P-Value for the Math score: 1.5359134607147415e-08
P-Value for the Reading score: 9.081783336892205e-15
P-Value for the Writing score: 3.68529173524572e-24

The two remaining categorical variables "parental level of education" and "race/ethnicity" have 6 and 5 distinct groups, respectively. As data in these groups have normal distribution (can be seen from histograms in the notebook), analysis workflow for them was the same.

At first, I decided to look at the means and standard deviations for each group in these categories (Table 2). Then I chose an exam, for which the means difference was the biggest. After doing one-way ANOVA I got statistical significance in both cases analysed, so then proceeded to do Tukey test to find which groups show significant differences.

Table 2 Means and standard deviations of all exam scores based on (a) parental level of education and (b) race/ethnicity.

(a)	**math score**		**reading score**		**writing score**	
	mean	**std**	**mean**	**std**	**mean**	**std**
parental level of education						
associate's degree	67.9	15.1	70.9	13.9	69.9	14.3
bachelor's degree	69.4	14.9	73.0	14.3	73.4	14.7
high school	62.1	14.5	64.7	14.1	62.4	14.1
master's degree	69.7	15.2	75.4	13.8	75.7	13.7
some college	67.1	14.3	69.5	14.1	68.8	15.0
some high school	63.5	15.9	66.9	15.5	64.9	15.7

(b)	**math score**		**reading score**		**writing score**	
	mean	**std**	**mean**	**std**	**mean**	**std**
race/ethnicity						
group A	61.6	14.5	64.7	15.5	62.7	15.5
group B	63.5	15.5	67.4	15.2	65.6	15.6
group C	64.5	14.9	69.1	14.0	67.8	15.0
group D	67.4	13.8	70.0	13.9	70.1	14.4
group E	73.8	15.5	73.0	14.9	71.4	15.1

For “parental level of education” I analysed writing exam results. Based on P-Values given by Tukey test (Table 3) we could see that students whose parents did not have any additional education after high school (categories "high school" and "some high school") got lower grades during writing exams compared to almost all other groups.

Table 3 Tukey test for the groups of “parental level of education”.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
associate's degree	bachelor's degree	3.485	0.2988	-1.2997	8.2696	False
associate's degree	high school	-7.4474	0.001	-11.5637	-3.3311	True
associate's degree	master's degree	5.7816	0.0797	-0.3699	11.933	False
associate's degree	some college	-1.0557	0.9	-5.0243	2.9129	False
associate's degree	some high school	-5.0081	0.0095	-9.227	-0.7893	True
bachelor's degree	high school	-10.9324	0.001	-15.8259	-6.0389	True
bachelor's degree	master's degree	2.2966	0.9	-4.3998	8.9931	False
bachelor's degree	some college	-4.5406	0.0728	-9.3105	0.2293	False
bachelor's degree	some high school	-8.4931	0.001	-13.4732	-3.513	True
high school	master's degree	13.229	0.001	6.9925	19.4655	True
high school	some college	6.3917	0.001	2.2925	10.4909	True
high school	some high school	2.4393	0.5827	-1.9027	6.7813	False
master's degree	some college	-6.8373	0.0189	-12.9772	-0.6973	True
master's degree	some high school	-10.7897	0.001	-17.0944	-4.485	True
some college	some high school	-3.9524	0.0792	-8.1546	0.2497	False

For the category “race/ethnicity” I chose to analyse Math scores. We can see (Table 4) that students from group E have significantly higher Math scores when compared to all other groups, as well as students from group D (with the exception when compared to group C students).

Table 4 Tukey test for the groups of “race/ethnicity”.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
group A	group B	1.8234	0.8597	-3.36	7.0068	False
group A	group C	2.8347	0.4966	-2.0028	7.6723	False
group A	group D	5.7334	0.0138	0.7824	10.6844	True
group A	group E	12.1922	0.001	6.7215	17.6629	True
group B	group C	1.0113	0.9	-2.6867	4.7094	False
group B	group D	3.91	0.0441	0.0647	7.7552	True
group B	group E	10.3688	0.001	5.8741	14.8635	True
group C	group D	2.8986	0.1287	-0.4659	6.2632	False
group C	group E	9.3575	0.001	5.2665	13.4485	True
group D	group E	6.4588	0.001	2.2343	10.6834	True

When trying to look at all exams as a whole and doing a PCA, the only category that showed distinct clusters was separation by gender. The initial clustering was seen from bivariate relationships in scatterplot matrices (Figure 6), clusters separating a bit when plotting writing/math and reading/math scores.

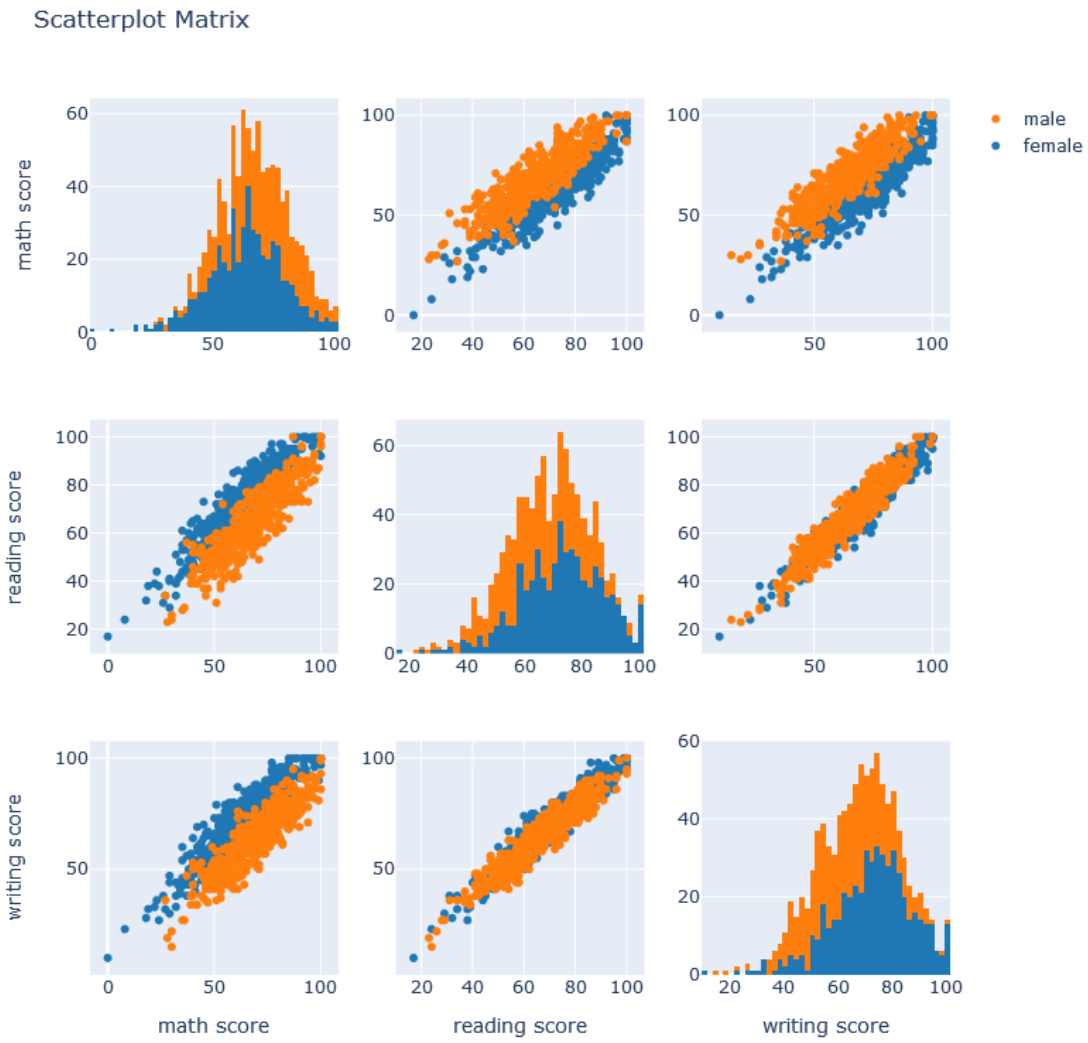

Figure 6 Scatterplots of all exams compared to each other. Male and female students results form clusters when compared in writing/math and reading/math scores.

I also decided to do PCA analysis, so at first I did two dimensional plotting (Figure 7a), then – three dimensional plotting (Figure 7b), which in this case does not show reduced dimensionality, as there were only three continuos variables compared.

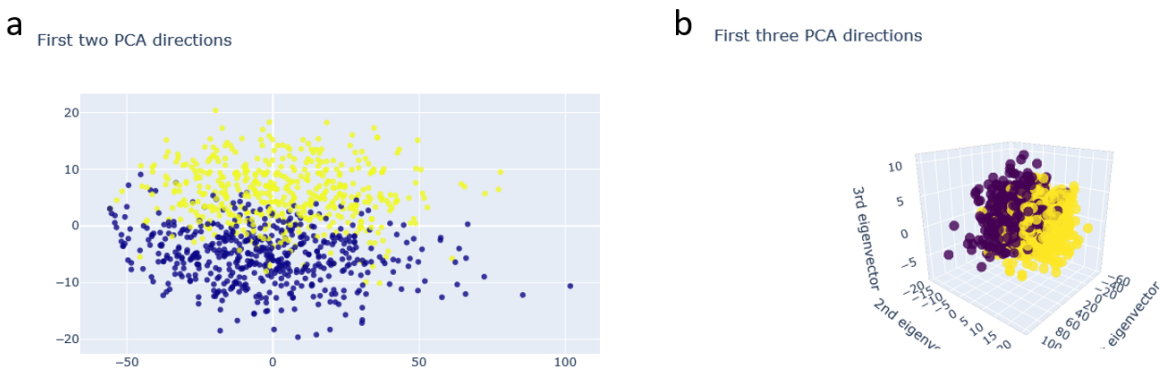

Figure 7 PCA plots for all three exam scores based on gender (a) two-dimensional and (b) three-dimensional.

Conclusions

In this project I have analysed a fictional dataset of 1000 US high school students scores in three exams: Math, writing and reading. There was no missing data in the dataset, all exams results were distributed normally, categorical variables had quite similar amounts of values given. That let me do simple statistical tests, such as t-tests and ANOVA.

After statistical analysis it seems that having better nutrition and additional preparations for exams could help students get higher scores in Math, writing and reading exams. There is also a trend, that males do better in Math and females succeed to get better results in writing and reading tests. Moreover, gender was the main category, separating all students tested (based on PCA plots). When looking at parental education level results in writing exam were statistically significantly lower only for students whose parents had only high school education. However, looking at the means, we can see that the average score increases with every additional level of education. Analysis of a category race/ethnicity in the field of Math scores showed a significant difference for groups E and D students, but as there is no description of these groups, we cannot make any further conclusions. The only interesting thing was that the mean scores showed the biggest variation between ethnical groups in Math, but not in linguistic tests.

When looking at the correlation between all three exams for which the scores were measured, the correlation coefficients R^2 in all cases were > 0.8, showing strong correlation, with the strongest being for reading and writing exams. These results were quite expected, as better students tend to get better results in all subjects and reading and writing are the exams testing similar skills.

The significance of this work is mainly creating a workflow, that could help analyse similar type of data. All these analyses could only give an insight what could have impact for the scores of students results. But as the dataset is fictional and there is no validation how it was composed, no real conclusions could be made from it.