# Database Systems II Project

Mendelova Univerzita v Brně

Pavel Kostarev

11 June, 2020

Students’ performance in studying and factors that influence it.
Comparing High School and University studying performance

Submitted to: Ing. Jan Přichystal, Ph.D.
Ing. Pavel Turčínek, Ph.D.
Provozně ekonomická fakulta

# Contents

## Intro

This project was created from 2 datasets. One of them I found on Kaggle.com and another one we created a year ago for statistic purposes at my Home university

# 1 Dataset information

## 1.1 Math students' performance from Kaggle

"This dataset contains the final scores of students at the end of a math programs with several features that might or might not impact the future outcome of these students."

Here is example of first three rows and all the columns that file student-mat.csv contains

| school | sex | age | address | famsize | Pstatus | Medu | Fedu | Mjob | Fjob | reason |
|---|---|---|---|---|---|---|---|---|---|---|
| GP | F | 18 | U | GT3 | A | 4 | 4 | at_home | teacher | course |
| GP | F | 17 | U | GT3 | T | 1 | 1 | at_home | other | course |
| GP | F | 15 | U | LE3 | T | 1 | 1 | at_home | other | other |
| | | | | | | | | | | |
| guardian | traveltime | studytime | failures | schoolsup | famsup | paid | activities | nursery | higher | internet |
| mother | 2 | 2 | 0 | yes | no | no | no | yes | yes | no |
| father | 1 | 2 | 0 | no | yes | no | no | no | yes | yes |
| mother | 1 | 2 | 3 | yes | no | yes | no | yes | yes | yes |
| | | | | | | | | | | |
| romantic | famrel | freetime | goout | Dalc | Walc | health | absences | G1 | G2 | G3 |
| no | 4 | 3 | 4 | 1 | 1 | 3 | 6 | 5 | 6 | 6 |
| no | 5 | 3 | 3 | 1 | 1 | 3 | 4 | 5 | 5 | 6 |
| no | 4 | 3 | 2 | 2 | 3 | 3 | 10 | 7 | 8 | 10 |

Number of rows = 396

Number of columns = 33

1. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

2. sex - student's sex (binary: 'F' - female or 'M' - male)

3. age - student's age (numeric: from 15 to 22)

4. **address** - student's home address type (binary: 'U' - urban or 'R' - rural)

5. **famsize** - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

6. **Pstatus** - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

7. **Medu** - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 → “ 5th to 9th grade, 3 → secondary education or 4 → higher education”)

8. **Fedu** - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 → “ 5th to 9th grade, 3 → secondary education or 4 → higher education ”)

9. **Mjob** - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other')

10. **Fjob** - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other')

11. **reason** - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')

12. **guardian** - student's guardian (nominal: 'mother', 'father' or 'other')

13. **traveltime** - home to school travel time (numeric: 1 - 1 hour)

14. **studytime** - weekly study time (numeric: 1 - 10 hours)

15. **failures** - number of past class failures (numeric: n if 1

16. **schoolsup** - extra educational support (binary: yes or no)

17. **famsup** - family educational support (binary: yes or no)

18. **paid** - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

19. **activities** - extra-curricular activities (binary: yes or no)

20. **nursery** - attended nursery school (binary: yes or no)

21. **higher** - wants to take higher education (binary: yes or no)

22. **internet** - Internet access at home (binary: yes or no)

23. **romantic** - with a romantic relationship (binary: yes or no)

24. **famrel** - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

25. **freetime** - free time after school (numeric: from 1 - very low to 5 - very high)

26. **goout** - going out with friends (numeric: from 1 - very low to 5 - very high)

27. **Dalc** - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

28. **Walc** - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

29. **health** - current health status (numeric: from 1 - very bad to 5 - very good)

30. **absences** - number of school absences (numeric: from 0 to 93)

**These grades are related with the course, subject - Mathematics**

31. G1 - first period grade (numeric: from 0 to 20)
32. G2 - second period grade (numeric: from 0 to 20)
33. G3 - final grade (numeric: from 0 to 20, output target)

## 1.2 Dataset “Ankieta2019” from my Home University, “West Pomeranian University of Technology in Szczecin, Poland”

We used this dataset at our Applied mathematics and statistics course. Translated from Polish.

| Sex | Weight | Height | Place of living | High School | ECTS | GPA | No. of hours | No. of OS | System | Age | F.education | No hrs code |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| M | 84 | 194 | Dormitory | Adv math | 24 | 3.6 | 4 | 1 | Windows 10 | 7 | FT | (0,4] |
| M | 69.5 | 180 | Rental housing | Adv math | 24 | 2.9 | 5 | 1 | Other | 7 | FT | (4,8] |
| M | 70 | 186 | Rental housing | Tech college | 18 | 2.1 | 4 | 1 | Windows 10 | 6 | FT | (0,4] |

Number of rows = 91
Number of columns = 13

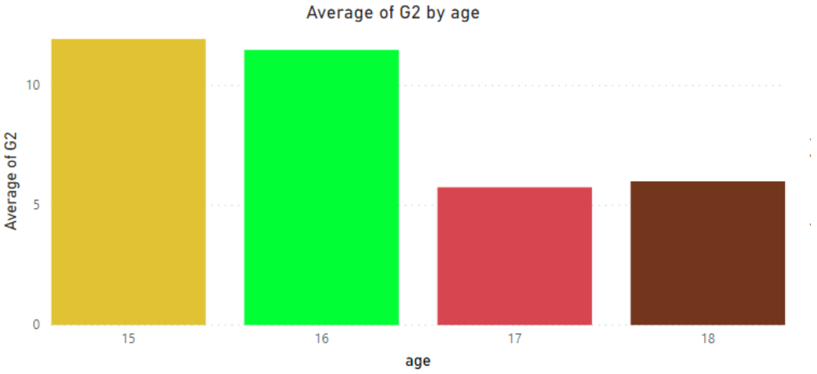
Attributes (blue colored attributes were included into the project)

1. → **Sex** (binary: M or F)
2. → **Weight** (decimal number)
3. → **Height** (whole number)
4. → **Place of living** - actual place of student's living (nominal: Dormitory, Rental Housing or Apartment) [used]
5. → **High School** - which school student attended before entering the university (nominal: adv math, basic math, tech college, other)[used]
6. → **ECTS** - number of credits received after the 1st semester (whole number)[used]
7. → **GPA** - the average value of the student's grades after the 1st semester (decimal number)[used]
8. → **No. Of hours** - spending in front of the PC (whole number)
9. → **No. Of Operation Systems** (whole number)
10. → **System** (nominal: windows 10, windows 8 or older, other)
11. → **Age of PC** (whole number)
12. → **Form of education** (binary: FT - full time, C - correspondence learning) [used]
13. → **No. Of hours spent coding** (interval)

# 2 KPI and their visual representation in Power BI

## 2.1 KPI 1 - Age / grade dependency

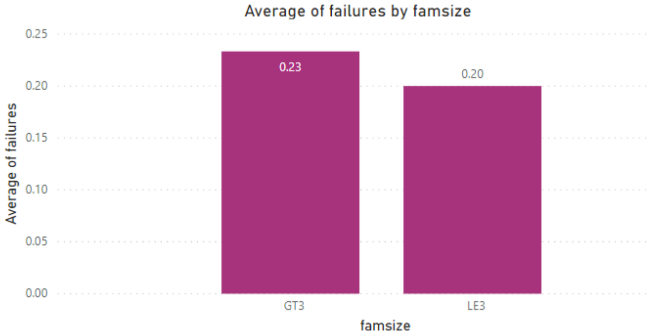
How depend the results of the second period grade on the age of students?



→Interpretation: analyzing the plot of students' average grades for the second period exams depending on the age we can say that younger students have better average results.

## 2.2 KPI 2 - Family size / failure dependency

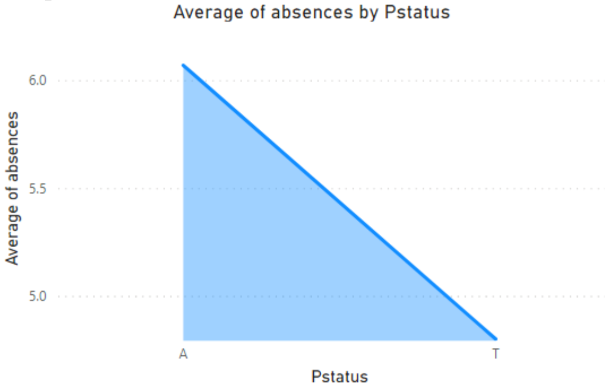
Does the size of family influence the average number of exam failures?



→Interpretation: analyzing the plot of the data of students' failures depending on the family size we can claim that the bigger family the higher average number of students' failures.

## 2.3 KPI 3 - Parents status / absences dependency

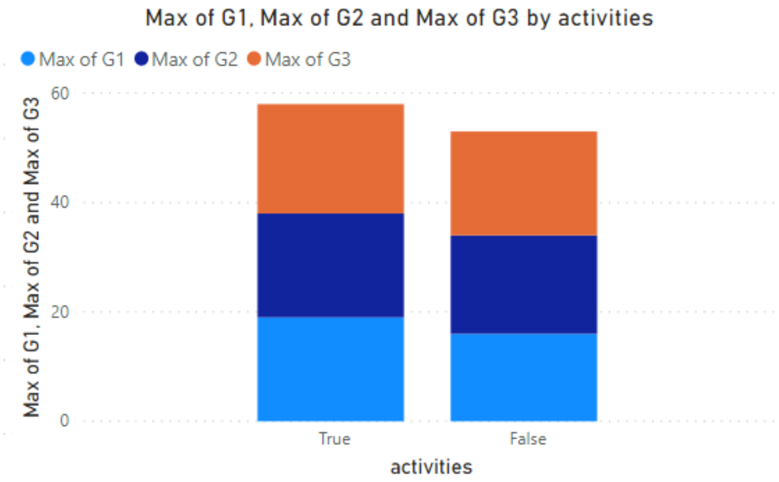
Can parent status influence the number of absences at school?



→Interpretation: analyzing the plot of the data of students with maximum first, second and last period exam result we can see that the students with add. activities always have better results in all the exam periods.

## 2.4 KPI 4 - Grade / activities dependency

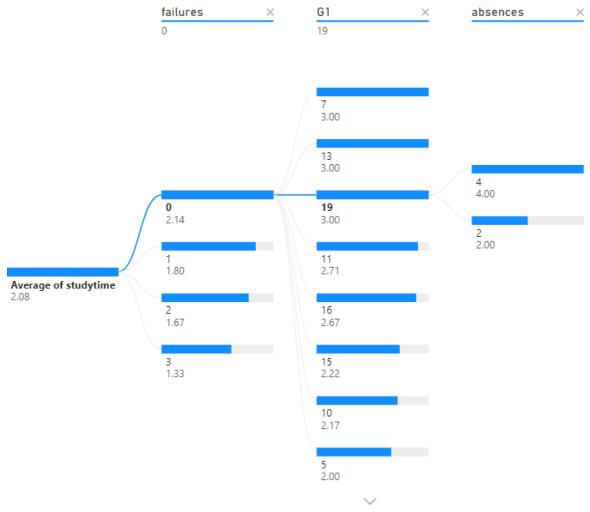
What is the maximum grades for students who have additional activities and for those who do not



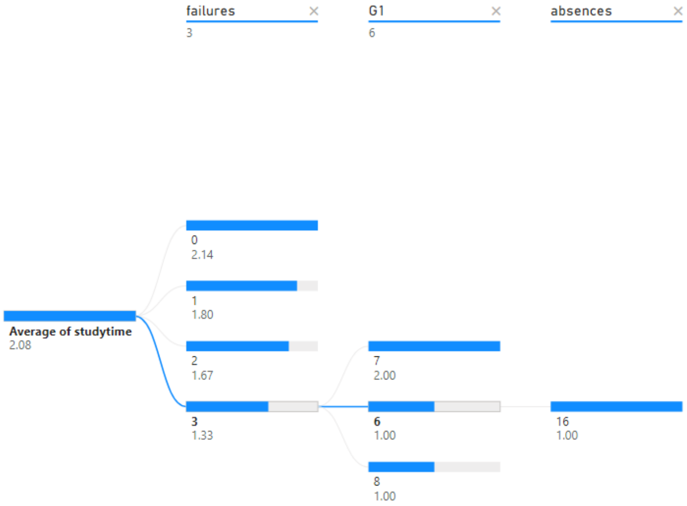
→Interpretation: analyzing the plot of the data of students with maximum first, second and last period exam result we can see that the students with add. activities always have better results in all the exam periods.

## 2.5 KPI 5 - What kind of students study the best?

How failures, absences and study time influence the grade result all together?



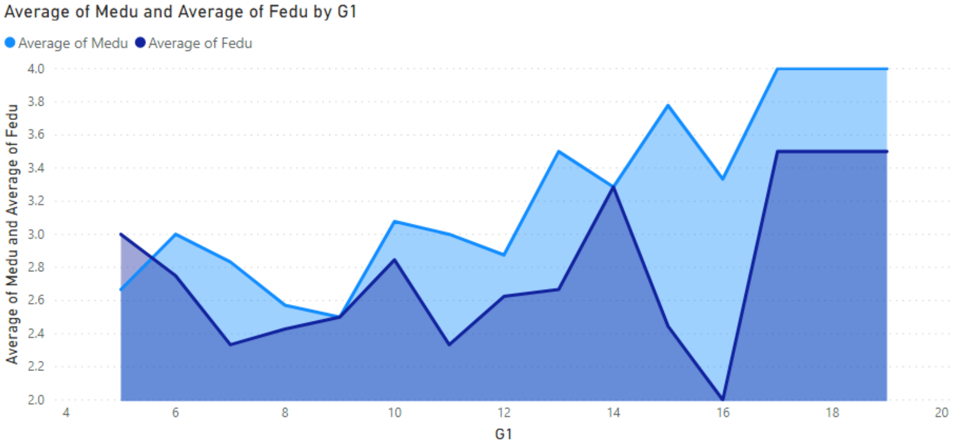
→Interpretation: analyzing the plot of data of students with 0 failures and average of study time 2.14 with the highest exam result at the first period we see two students who had few absences, i.e. 4 and 2



→Interpretation: analyzing the plot of the data of students with 3 failures (max possible) and average of study time 1.33 with the lowest exam result at the first period we see one student who had a lot of absences.

## 2.6 KPI 6 - Is parents' education so important?

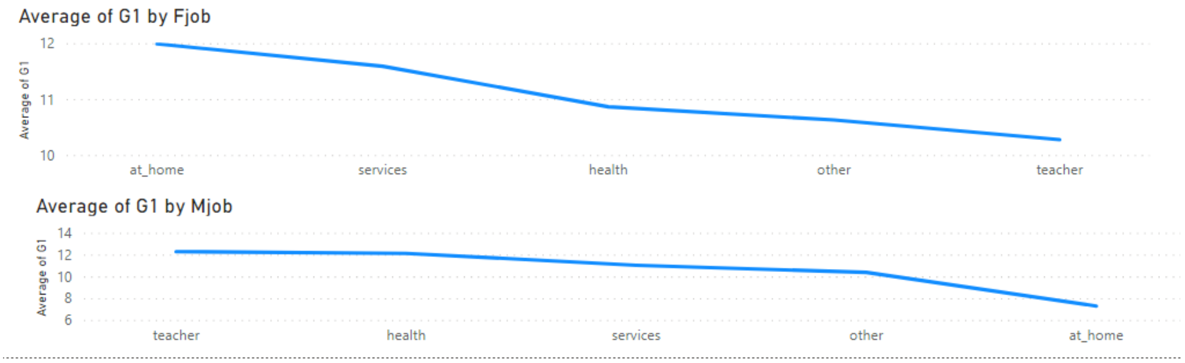
How can parents' education influence the students' performance?



→Interpretation: Students' mothers from this dataset more often have better education than students' fathers and the more educated parents are, the higher grade of a student is.

## 2.7 KPI 7 Is parents' job important?

Does the students' performance depend on what kind of job parents do?



→Interpretation: Students whose mother is a teacher and father works from home have the best first period grade results, but vice versa if mother works at home and father is a teacher it produces the lowest grade results.

# 3 Transforming the data

## 3.1 Reducing the dataset

I deleted too many rows from the Kaggle dataset to make the two datasets equal in the number of rows. I also deleted unnecessary column from the initial dataset.

## 3.2 Changing the data types

I changed the values of 'yes' to TRUE and 'no' to FALSE in the tables, making the cells binary.

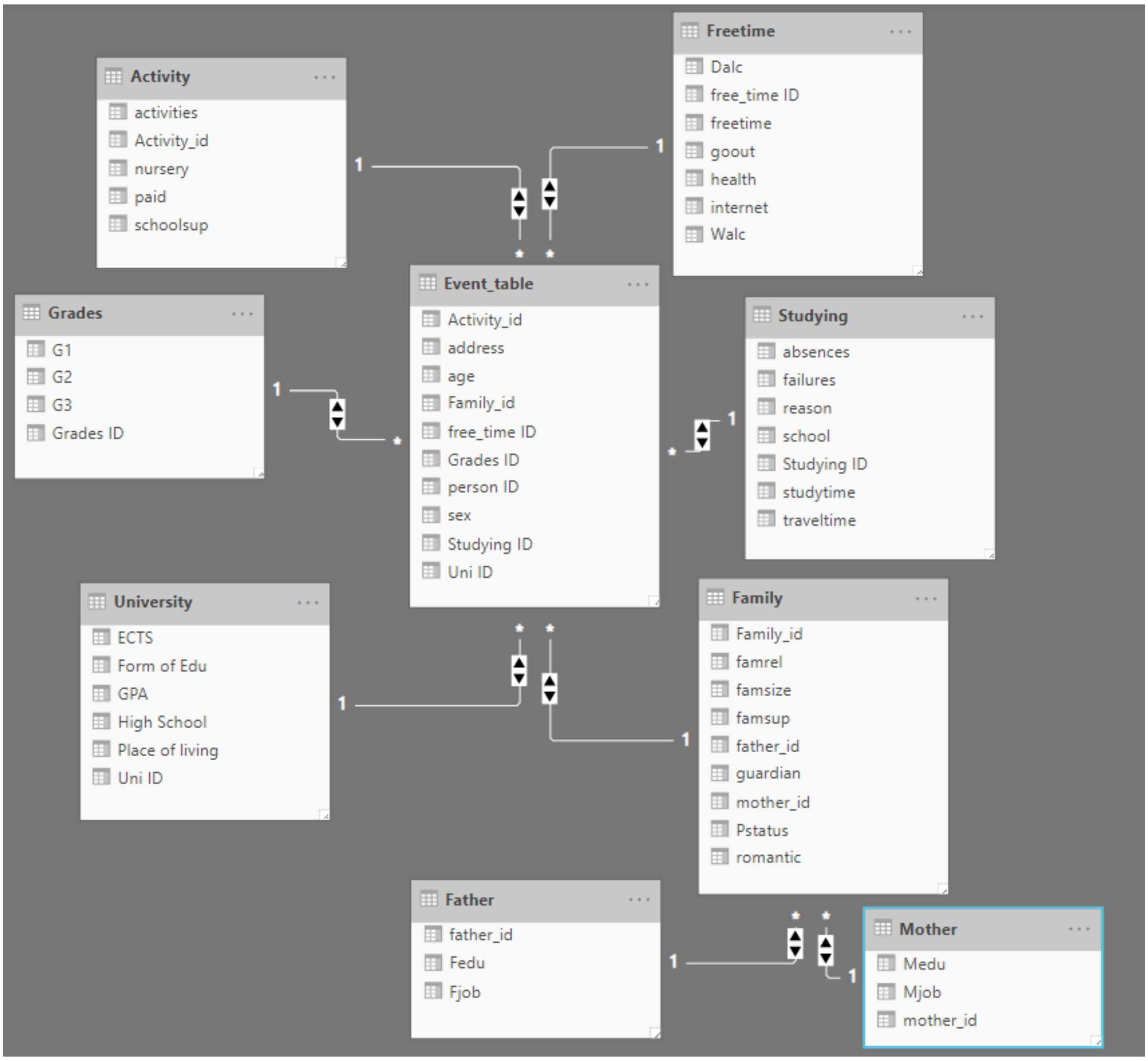
## 3.3 Snowflake scheme

I divided the table into several tables that have common attributes to describe a particular student connected with the help of foreign keys (unique ID of a particular table connected) with the Event table. All the tables have many-to-one relationships with the Event table.

## 3.4 Merging two datasets

I added the attributes that could be interesting to analyze and produce new correlations and dependencies with the data, having some conclusion. The second dataset can help us to imagine what grades or points would math students have if they would decide to study Informatics.

## 3.5 Model



## 3.6 Model description

Event table

| | | |
|---|---|---|
| person ID | int | Primary key |
| address | varchar | U/R |
| age | int | 18 |
| sex | varchar | M/F |
| Activity_id | int | Foreign key |
| Studying ID | int | Foreign key |
| Uni ID | int | Foreign key |
| Grades ID | int | Foreign key |
| Family_id | int | Foreign key |
| free_time ID | int | Foreign key |

Activity table

| Activity_id | int | Primary key |
|---|---|---|
| activities | binary | TRUE/FALSE |
| nursery | binary | TRUE/FALSE |
| paid | binary | TRUE/FALSE |
| schoolsup | binary | TRUE/FALSE |

**Free time table**

| free_time ID | int | Primary key |
|---|---|---|
| freetime | int | Values 1:5 |
| goout | int | Values 1:5 |
| Dalc | int | Values 1:5 |
| Walc | int | Values 1:5 |
| health | int | Values 1:5 |
| internet | binary | TRUE/FALSE |

**Studying table**

| Studying ID | int | Primary key |
|---|---|---|
| school | varchar | U/R |
| studytime | int | 18 |
| failures | varchar | M/F |
| absences | int | Foreign key |
| reason | int | Foreign key |
| traveltime | int | Foreign key |

**Family table**

| Family_id | int | Primary key |
|---|---|---|
| famsize | varchar | GT3/LE3 |
| Pstatus | varchar | A/T |
| famsup | binary | TRUE/FALSE |
| romantic | binary | TRUE/FALSE |
| famrel | int | Values 1:5 |
| guardian | varchar | mother/father |
| mother_id | int | Foreign key |
| father_id | int | Foreign key |

**Father table**

| father_id | int | Primary key |
|---|---|---|
| Fedu | int | Values 1:5 |
| Fjob | varchar | at_home/services/health/other/teacher |

**Mother table**

| mother_id | int | Primary key |
|---|---|---|
| Medu | int | Values 1:5 |

| | | |
|---|---|---|
| Mjob | varchar | at_home/services/health/other/teacher |

Grades table

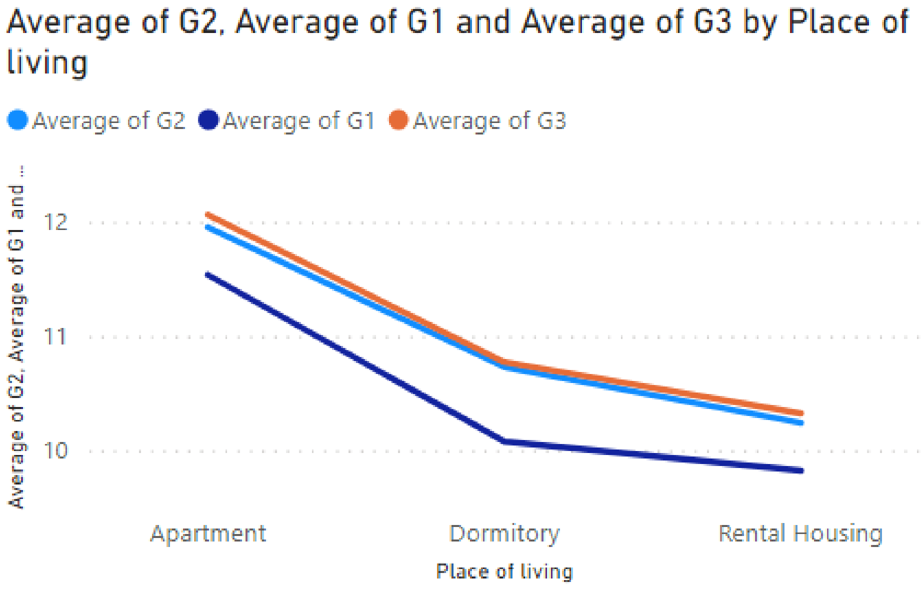
| Grades ID | int | Primary key |
|---|---|---|
| G1 | int | Values 0:20 |
| G2 | int | Values 0:20 |
| G3 | int | Values 0:20 |

University table

| Uni ID | int | Primary key |
|---|---|---|
| Place of living | varchar | Apartment, Rental Housing, Dormitory |
| GPA | double | 4.3 |
| Form of Edu | varchar | FT/C |
| ECTS | int | 25 |
| High School | varchar | Adv math, basic math, tech college, other |

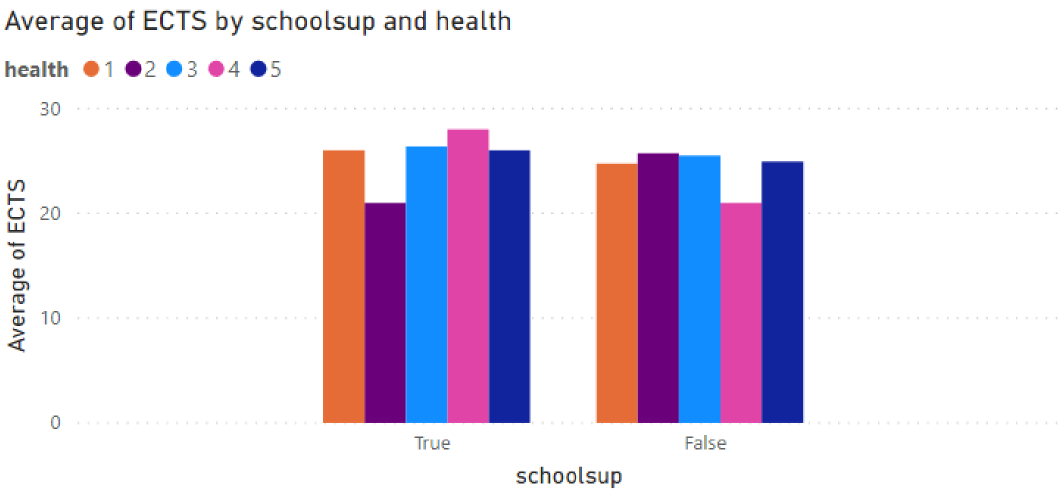
# 4 Visualization of new information that we get

## 4.1 Where probably will live students according to the all three grades results?



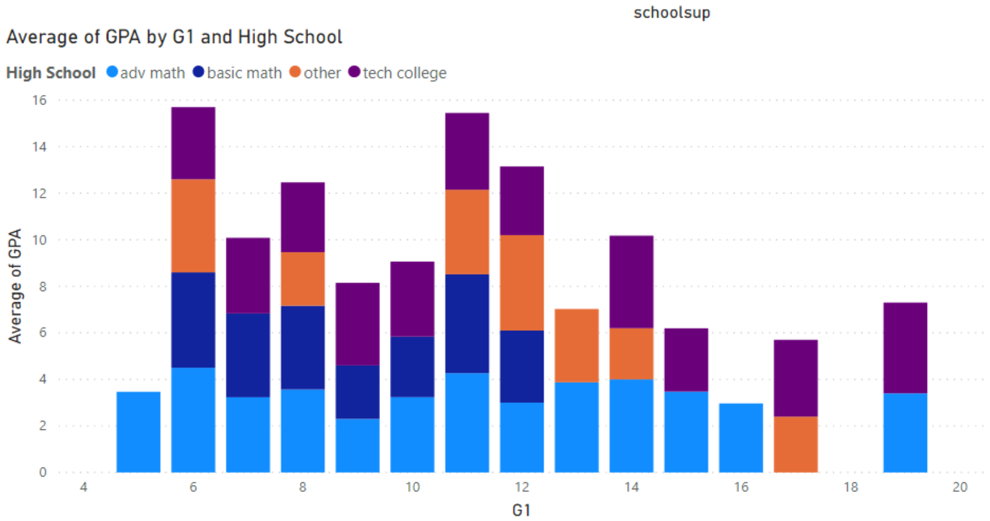
→Interpretation: analyzing the plot we can come to the conclusion that students who had the highest grades at school usually live in the apartment and students with the lowest ones rent a house.

## 4.2 How health at schooltime and school support will influence the number of ECTS credits at the University?



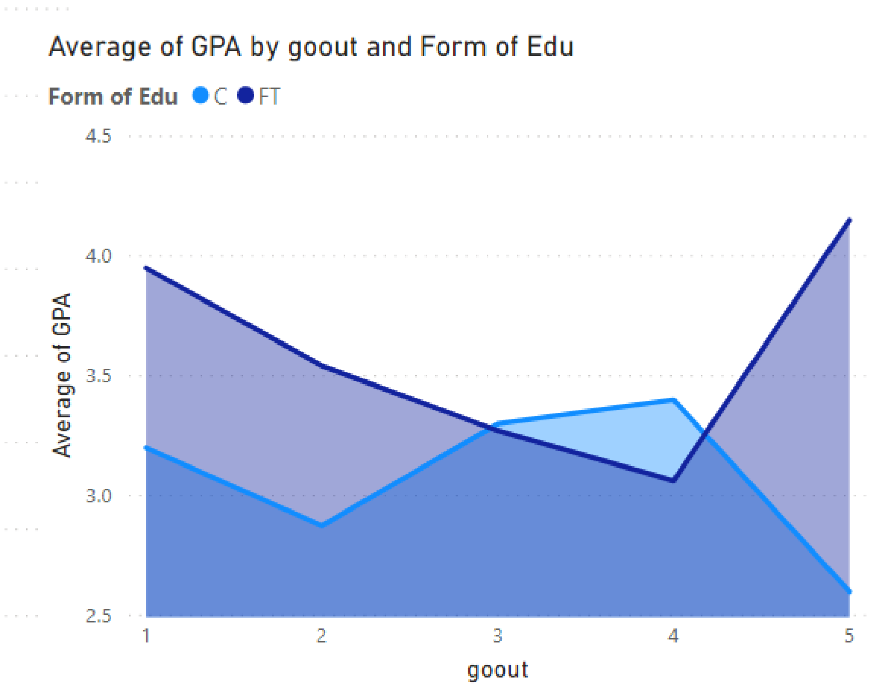
→Interpretation: analyzing the plot of data of students it can be said that students who had some extra educational support and have average or perfect state of health are more likely to have higher number of ECTS credits studying at the university.

## 4.3 Will background education and first period exam grade play role at the university GPA?



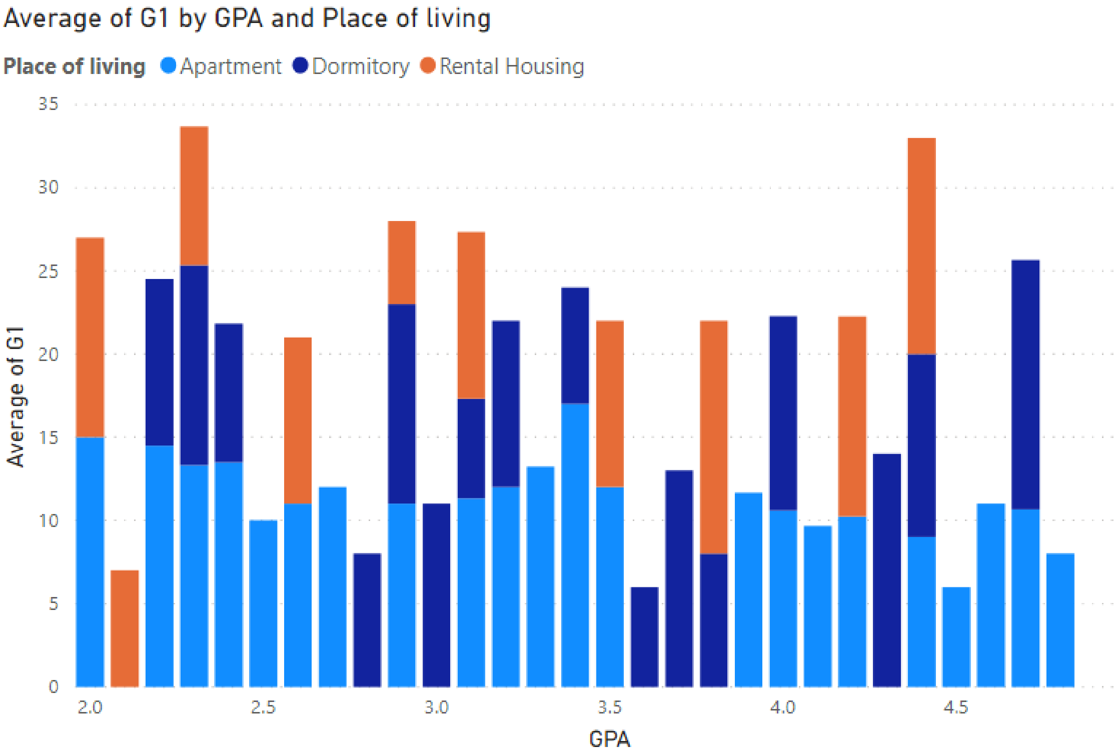
→Interpretation: students who had advanced math course at the high school and who had good results from the Math exam at the school are more likely to have higher GPA at the university. Although it does not often refer to the technical college students.

## 4.4 How influence students' going out on the average GPA results for full-time and correspondence learning students?



→Interpretation: students who were going out more during the school time do not have problems with studying analyzing their university GPA if their form of studying is Full studying, but if they are attending correspondence course it is opposite.

## 4.5 Does differ GPA values from the place of living and what grade these students at school had? Do they study better now?



→Interpretation: there are many students who had higher exam grades at schooltime, but do not have enough good GPA at the university. Place of living doesn't have strong influence on it, but still students living in an apartment have higher GPA.

# 5 Pros and Cons of the data analysis.

## 5.1 What are the disadvantages of the model?

The datasets have very different origin and probably not all the information could be referring to real life situations. It would be good to compare datasets from one place or have datasets from many different places to generalize the results.

# 6 References

- https://www.kaggle.com/janiobachmann/math-students
- P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of

5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

- Dataset “Ankieta2019” was prepared by my professors from the questionnaire that students of the Informatics Department were participating: dr **Małgorzata Machowska-Szewczyk**, dr **Joanna Banaś, West Pomeranian University of Techonology, Szczecin, Poland**
- The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, Third Edition Published by John Wiley & Sons, Inc. 10475 Crosspoint Boulevard Indianapolis, IN 46256 www.wiley.com Copyright © 2013 by Ralph Kimball and Margy Ross Published by John Wiley & Sons, Inc., Indianapolis, Indiana
- https://docs.microsoft.com/en-us/power-bi/guided-learning/
- https://docs.microsoft.com/en-us/power-bi/connect-data/desktop-quickstart-connect-to-data
- https://docs.microsoft.com/en-us/power-bi/transform-model/desktop-analytics-pane
- https://www.youtube.com/watch?v=nPhtDVlRvSo