PISA dataset

Udacity EDA course
Project by Boris Livshits

What determine PISA results?

What we want to know?

We are going to find some parameters and characteristics of student that could be useful in estimation/prediction of their PISA results. And, visa versa, we want to understand what is important to achieve high PISA results.

Dataset overview:

PISA Dataset contains results of PISA 2012 assesments or students of 65 economies in math, reading and science. Dataset contains full list of answers, estimates and results of assesment. A lot of questions are in form of categorital data (i.e. "Agree"/"Disagree" and so on), as result - about half of records have missing answers for about 30-40% of questions.

Point of interest in dataset

PISA dataset is rather big - about 485000 records with more than 630 variables, so we need to choose a set of variables for further investigation.

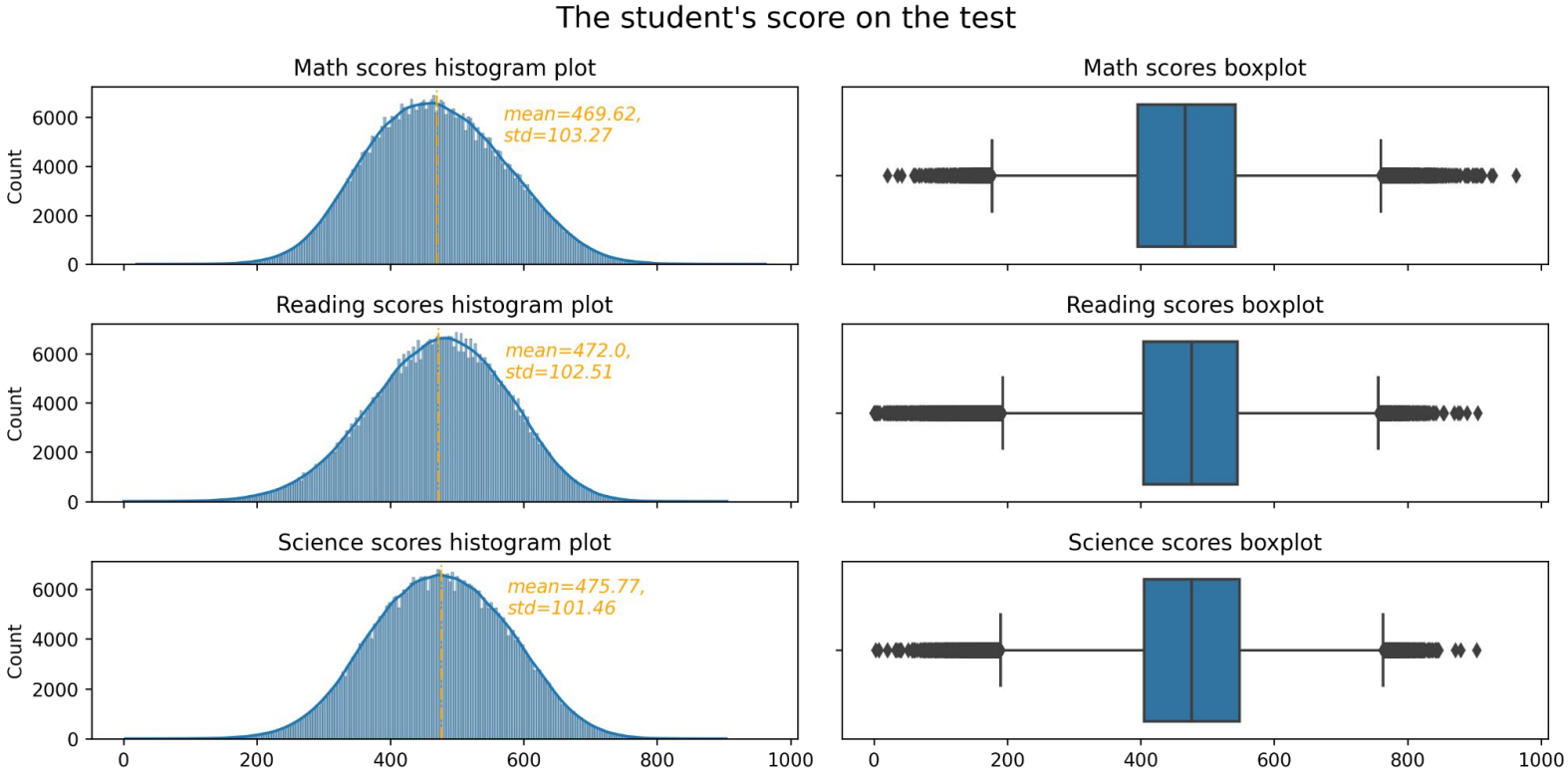
After consideration and some research PISA test design and questions we choose:

As test results - PV1MATH, PV1READ and PV1SCIE - test plausible values #1 in math, read and science, respectively.

As possible dependent factors:

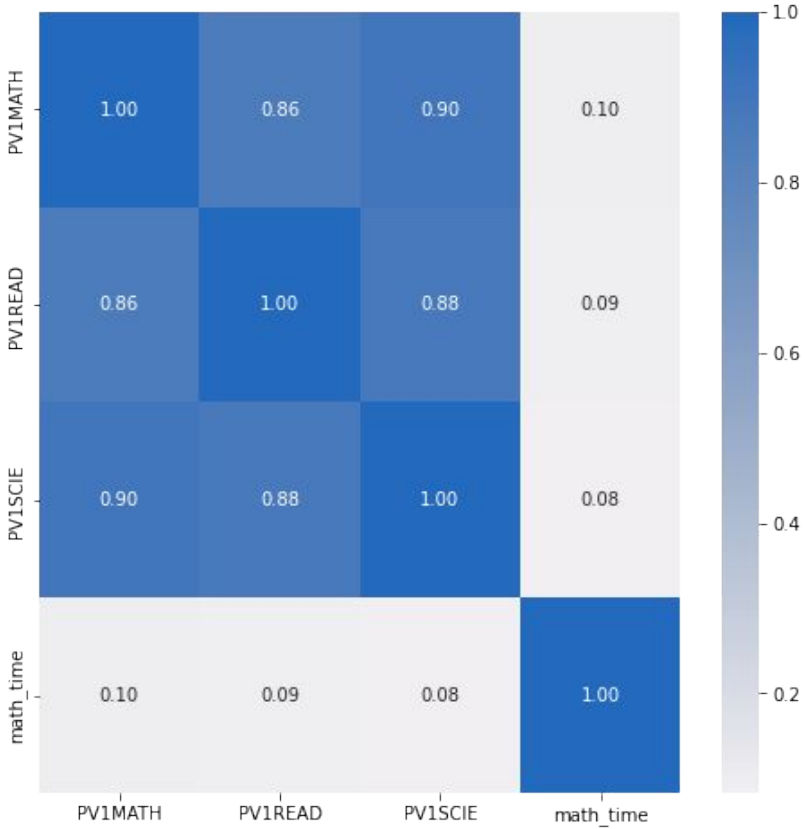
- *ST62Q01* (knowing the concept of exponential function)
- *IC04Q01* (age at first time access the Internet)
- *ST28Q01* (how many books at home)
- And we create new variable - *math_time* (minutes of math lessons in school per week) as a product of 'ST69Q02' - duration of math lesson and ST70Q02 - how many math lessons are in week. By the way - we remove all records from dataset with total duration more than 10 hours per week and less than 30 minutes per week.

Test results distribution

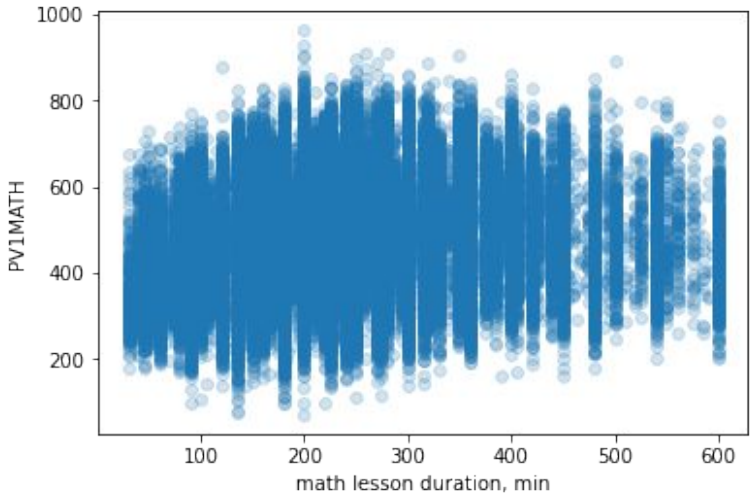


- most important - all three scores is normally distributed
- reading scores looks slightly left-skewed, while math scores - slightly right skewed.

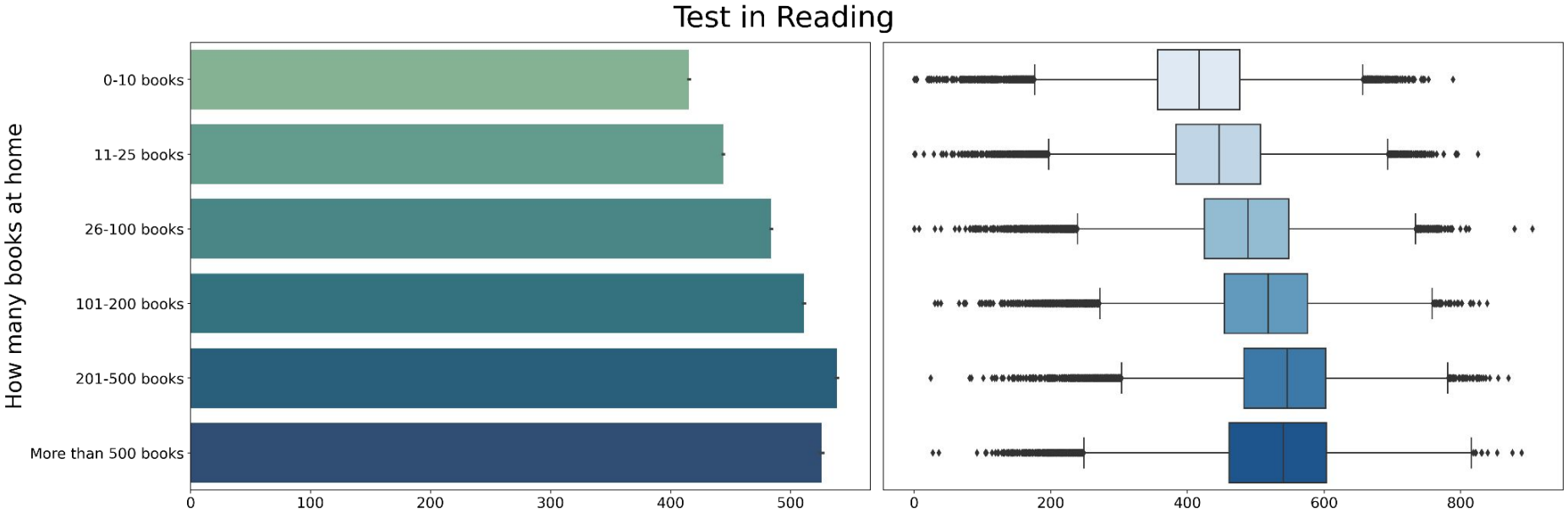
How correlate test results and math lesson duration?



We see that test results are very correlated with each other
However, math lessons duration seems to be insignificant in terms of correlation.
Scater plot of test result in math and lessons duration doesn't provide any intuition or insights.



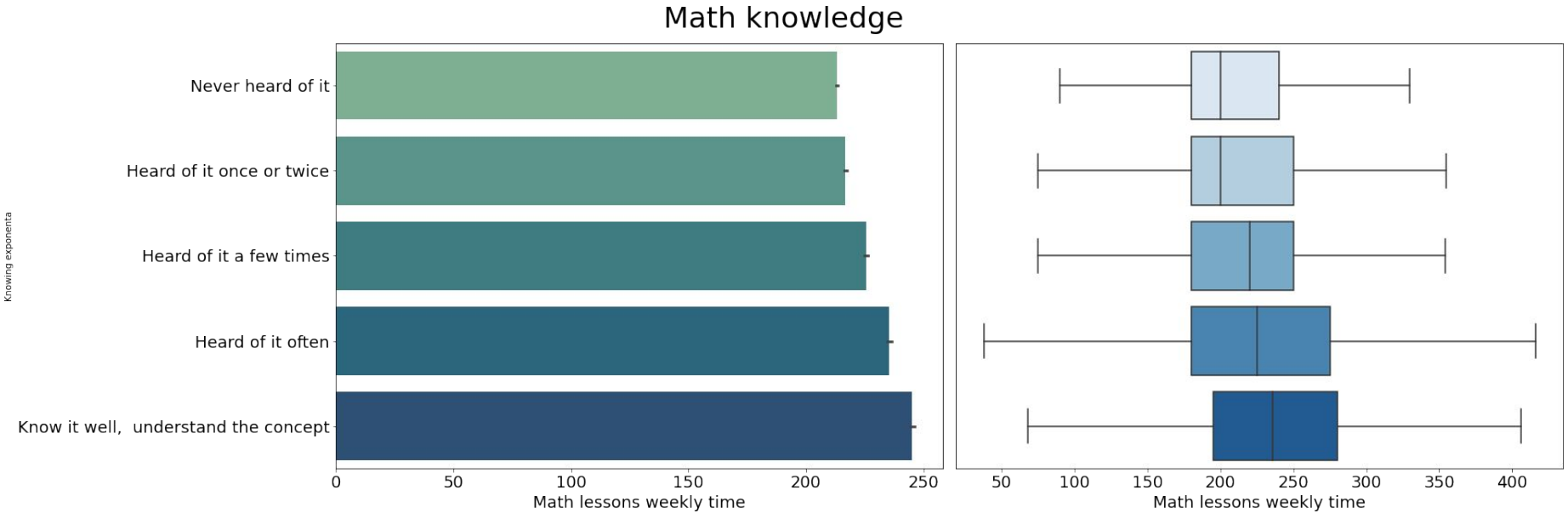
Books at home and Reading Test



Group with 200-500 books at home show on average higher results than the group with 500+ books at home. There can be several explanations:

1. it maybe difficult to estimate how many books you have, so, maybe some students, didn't provide accurate answers.
2. quantity doesn't mean quality - if you have books at home it doesn't necessary mean that you are reading them.

Math knowledge and duration of math lessons

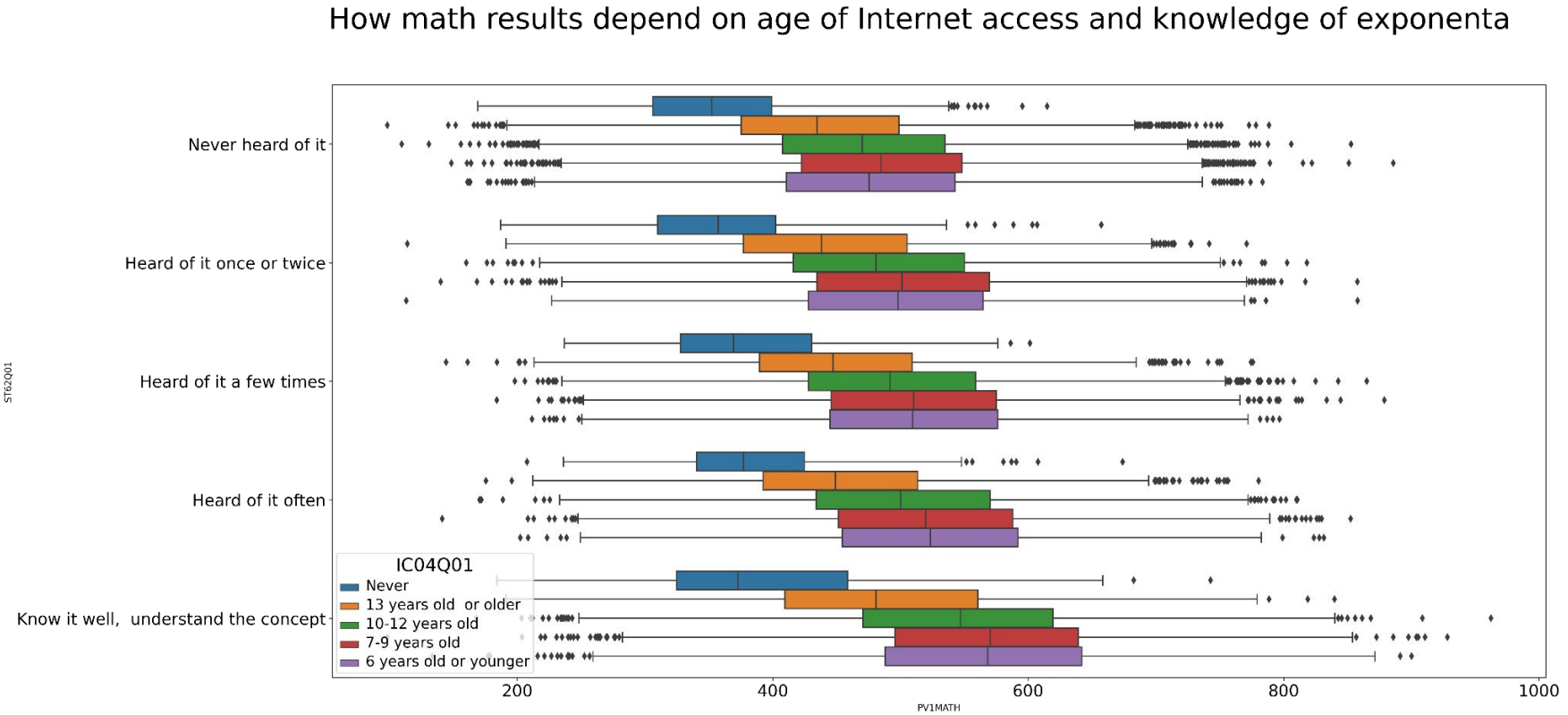


Without outliers - we have some kind of expected results. Than more you learn - than more you know.

However, we have to mention two points:

- There are some outliers in dataset - for example max time of lessons in group “never heard of exp” is top value for all records in dataset.
- Correlation coefficient between duration of lesson and test results is about 0.1, which means no correlation

How do math results depend on age of Internet access and knowledge of exponenta



Brief conclusions:

1. Based on our choice of variables the most important for our purposes is “age of access to the Internet”
2. However, it seems that difference between groups with age of access lower than 12 is not significant. But students who got access at 13+ years or doesn’t get access - their results differ very much. But, probably, is example of correlation and not of causality. Probably, there is some other reasons that influence both - learning (and, as result, PISA test outcomes) and age of access to the Internet.
3. Lessons duration seems to be unimportant - but, possible, the reason is that lesson duration is very similar for most of observations, that’s why it is statistically unimportant.
4. Going further it is interesting to look on other variables that can be important in explaining later age of access and economical welfare of family and society of student.