

Communicate_Data_Findings

September 18, 2020

1 Project Overview

This project has two parts that demonstrate the importance and value of data visualization techniques in the data analysis process. In the first part, you will use Python visualization libraries to systematically explore a selected dataset, starting from plots of single variables and building up to plots of multiple variables. In the second part, you will produce a short presentation that illustrates interesting properties, trends, and relationships that you discovered in your selected dataset. The primary method of conveying your findings will be through transforming your exploratory visualizations from the first part into polished, explanatory visualizations.

1.1 Table of Contents

- Section ??
- Section ??
- Section ??
- Section ??

Why this project?

Data visualization is an important skill that is used in many parts of the data analysis process.

Exploratory data visualization generally occurs during and after the data wrangling process, and is the main method that you will use to understand the patterns and relationships present in your data. This understanding will help you approach any statistical analyses and will help you build conclusions and findings. This process might also illuminate additional data cleaning tasks to be performed.

Explanatory data visualization techniques are used after generating your findings, and are used to help communicate your results to others. Understanding design considerations will make sure that your message is clear and effective. In addition to being a good producer of visualizations, going through this project will also help you be a good consumer of visualizations that are presented to you by others.

For this project I choose to analyse the results of the OECD **Programme for International Student Assessment (PISA)** in 2012.

From OECD website: *PISA is an international study that was launched by the OECD in 1997, first administered in 2000 and now covers over 80 countries. Every 3 years the PISA survey provides comparative data on 15-year-olds' performance in reading, mathematics, and science. In addition, each cycle explores a distinct "innovative domain" such as Collaborative Problem Solving (PISA 2015) and Global Competence (PISA 2018). The results have informed education policy discussions at the national*

and global level since its inception.

<https://www.oecd.org/pisa/aboutpisa/pisa-based-test-for-schools-faq.htm>

The PISA goals are:

- Empower school leaders and teachers by providing them with evidence-based analysis of their students' performance.
- Measure students' knowledge, skills and competencies that will equip them for success in education and the world of work.
- Provide valuable information on the learning climate within a school, students' socioeconomic background and motivation for learning.
- Help schools measure a wider range of 21st century skills beyond maths, reading and science.
- Provide opportunities for global peer-learning among teachers and school leaders.

Based on the objectives of the PISA, using the data, the following questions can be answered:

1. What is students' performance at schools in different countries (including whether country is a OECD member).
2. What are the characteristics of students participated in PISA 2012:
 - * gender,
 - * age,
 - * whether a student passed the test in the country of birth or not,
 - * international grade and grade compared to modal grade in country.
3. What's a relationship between students performance and highest parental education measured in years as well as mother's and father's highest schooling?
4. Whether there exist a correlation between family wealth (measured in the number of telephones, computers, etc.) and students performance?
5. How do student possessions such as own room and desk, etc. affect his/her performance?
6. Last but not least, whether total time learning and out of school lessons on math, science, and reading affect student performance?

```
[1]: # Import libraries

import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline

pd.set_option('display.max_columns', 700)
```

```
[2]: FOLDER = 'C:/Sasha/udacity/Data Analyst Nanodegree Program/5. Data_
      ↳Visualization/8. Communicate Data Findings'

FILE_NAME = 'pisa2012.csv'
DICT_NAME = 'pisadict2012.csv'

file_path = os.path.join(FOLDER, FILE_NAME)
dict_path = os.path.join(FOLDER, DICT_NAME)
```

```
[3]: DTYPES = {'SCHOOLID': 'str', 'STIDSTD': 'str'}
```

```
[4]: # Load in data
```

```
df = pd.read_csv(file_path, dtype=DTYPES, encoding='ISO-8859-1')
df.head(2)
```

C:\ProgramData\Anaconda3\lib\site-

packages\IPython\core\interactiveshell.py:3057: DtypeWarning: Columns (15,16,17, 21,22,23,24,25,26,30,31,36,37,45,65,123,155,156,157,158,159,160,161,162,163,164, 165,166,167,168,169,170,171,284,285,286,287,288,289,290,291,292,293,294,295,296, 297,298,299,300,301,302,303,307,308,309,310,311,312,313,314,315,316,317,318,319, 320,321,322,323,324,325,326,327,328,329,330,331,332,333,334,335,336,337,338,339, 340,341,342,343,344,345,346,347,348,349,350,351,352,353,354,355,356,357,376,377, 378,379,380,381,382,383,384,385,386,387,388,389,390,391,392,393,394,395,396,397, 398,399,400,401,402,403,475) have mixed types. Specify dtype option on import or set low_memory=False.

```
interactivity=interactivity, compiler=compiler, result=result)
```

```
[4]: Unnamed: 0      CNT  SUBNATIO  STRATUM      OECD      NC SCHOOLID STIDSTD  \
0           1  Albania      80000  ALB0006  Non-OECD  Albania  0000001  00001
1           2  Albania      80000  ALB0006  Non-OECD  Albania  0000001  00002
```

```
      ST01Q01  ST02Q01  ST03Q01  ST03Q02  ST04Q01      ST05Q01  \
0          10         1.0         2      1996  Female      No
1          10         1.0         2      1996  Female  Yes, for more than one year
```

```
      ST06Q01  ST07Q01  ST07Q02  ST07Q03      ST08Q01  ST09Q01  \
0         6.0  No, never  No, never  No, never      None      None
1         7.0  No, never  No, never  No, never  One or two times  None
```

```
      ST115Q01  ST11Q01  ST11Q02  ST11Q03  ST11Q04  ST11Q05  ST11Q06  \
0          1.0      Yes      Yes      Yes      Yes      NaN      NaN
1          1.0      Yes      Yes      NaN      Yes      NaN      NaN
```

```
      ST13Q01  ST14Q01  ST14Q02  ST14Q03  ST14Q04  \
0 <ISCED level 3A>      No      No      No      No
1 <ISCED level 3A>      Yes      Yes      No      No
```

```
      ST15Q01      ST17Q01  ST18Q01  ST18Q02  \
0  Other (e.g. home duties, retired) <ISCED level 3A>      NaN      NaN
1      Working full-time <for pay> <ISCED level 3A>      No      No
```

```
      ST18Q03  ST18Q04      ST19Q01      ST20Q01  \
0      NaN      NaN  Working part-time <for pay>  Country of test
1      No      No  Working full-time <for pay>  Country of test
```

```
      ST20Q02      ST20Q03  ST21Q01      ST25Q01  ST26Q01  \
```

0	Country of test	Country of test	NaN	Language of the test	Yes
1	Country of test	Country of test	NaN	Language of the test	Yes

	ST26Q02	ST26Q03	ST26Q04	ST26Q05	ST26Q06	ST26Q07	ST26Q08	ST26Q09	ST26Q10	\
0	No	Yes	No	No	No	No	Yes	No	Yes	
1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	

	ST26Q11	ST26Q12	ST26Q13	ST26Q14	ST26Q15	ST26Q16	ST26Q17		ST27Q01	\
0	No	Yes	No	Yes	8002	8001	8002		Two	
1	Yes	Yes	Yes	Yes	8001	8001	8002	Three or more		

	ST27Q02		ST27Q03	ST27Q04	ST27Q05		ST28Q01	ST29Q01	\
0	One		None	None	None		0-10 books	Agree	
1	Three or more	Three or more		Two	Two	201-500 books		Disagree	

	ST29Q02	ST29Q03	ST29Q04	ST29Q05	ST29Q06	ST29Q07	\
0	Strongly agree	Agree	Agree	Agree	Agree	Agree	
1	Strongly agree	Disagree	Disagree	Agree	Agree	Disagree	

	ST29Q08		ST35Q01		ST35Q02	ST35Q03	ST35Q04	ST35Q05	\
0	Strongly agree		Disagree		Agree	Disagree	Agree	Agree	
1	Disagree	Strongly agree		Strongly agree	Disagree	Agree		Disagree	

	ST35Q06		ST37Q01		ST37Q02		ST37Q03	\
0	Agree	Not at all confident		Not very confident			Confident	
1	Agree		Confident		Very confident		Very confident	

	ST37Q04		ST37Q05		ST37Q06		ST37Q07	\
0	Confident		Confident	Not at all confident			Confident	
1	Confident	Very confident			Confident	Very confident		

	ST37Q08	ST42Q01	ST42Q02	ST42Q03	ST42Q04	ST42Q05	ST42Q06	\
0	Very confident	Agree	Disagree	Agree	Agree	Agree	Agree	
1	Not very confident	NaN	NaN	NaN	NaN	NaN	NaN	

	ST42Q07	ST42Q08	ST42Q09	ST42Q10		ST43Q01		ST43Q02	\
0	Agree	Disagree	Disagree	Disagree		Agree		Disagree	
1	NaN	NaN	NaN	NaN	Strongly agree		Strongly agree		

	ST43Q03	ST43Q04	ST43Q05	ST43Q06	ST44Q01		ST44Q03	\
0	Disagree	Agree	NaN	Disagree	Likely		Slightly likely	
1	Strongly disagree	Disagree	Agree	Disagree	Likely		Slightly likely	

	ST44Q04		ST44Q05		ST44Q07		ST44Q08	ST46Q01	\
0	Likely		Likely		Likely	Very	Likely	Agree	
1	Slightly likely	Very	Likely	Slightly likely			Likely	Agree	

	ST46Q02	ST46Q03	ST46Q04	ST46Q05	ST46Q06	ST46Q07	\		
0	Agree	Agree	Agree	Agree	Agree	Agree			
1	Agree	Strongly agree	Strongly agree	Strongly agree	Agree	Agree			
	ST46Q08	ST46Q09		ST48Q01			\		
0	Agree	Agree	Courses after school	Test Language					
1	Disagree	Agree	Courses after school	Math					
		ST48Q02		ST48Q03			\		
0	Major in college	Science	Study harder	Test Language					
1	Major in college	Science	Study harder	Math					
		ST48Q04		ST48Q05	ST49Q01	ST49Q02	\		
0	Maximum classes	Science	Pursuing a career	Math	Often	Sometimes			
1	Maximum classes	Science	Pursuing a career	Science	Sometimes	Often			
		ST49Q03	ST49Q04		ST49Q05		\		
0		Sometimes	Sometimes		Sometimes				
1	Always or almost always	Sometimes	Always or almost always						
	ST49Q06	ST49Q07	ST49Q09		ST53Q01		\		
0	Never or rarely	Never or rarely	Never or rarely		NaN				
1	Never or rarely	Never or rarely	Often	relating to known					
	ST53Q02	ST53Q03	ST53Q04				\		
0	NaN	NaN	NaN						
1	Improve understanding	in my sleep	Repeat examples						
			ST55Q01				\		
0			NaN						
1	I do not attend	<out-of-school time lessons>	i...						
			ST55Q02				\		
0			NaN						
1	2 or more but less than 4 hours	a week							
			ST55Q03		ST55Q04	ST57Q01	\		
0			NaN		NaN	NaN			
1	2 or more but less than 4 hours	a week	Less than 2 hours	a week		NaN			
	ST57Q02	ST57Q03	ST57Q04	ST57Q05	ST57Q06	ST61Q01	ST61Q02	ST61Q03	\
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1	NaN	6.0	0.0	0.0	2.0	Rarely	Rarely	Frequently	
	ST61Q04	ST61Q05	ST61Q06	ST61Q07	ST61Q08	ST61Q09			\
0	NaN	NaN	NaN	NaN	NaN	NaN			
1	Sometimes	Frequently	Sometimes	Frequently	Never	Frequently			

				ST62Q01	\			
0				NaN				
1	Know it well,	understand the	concept					
				ST62Q02		ST62Q03	\	
0				NaN		NaN		
1	Know it well,	understand the	concept	Heard of it	once or	twice		
				ST62Q04	\			
0				NaN				
1	Know it well,	understand the	concept					
				ST62Q06	\			
0				NaN				
1	Know it well,	understand the	concept					
				ST62Q07		ST62Q08	\	
0				NaN		NaN		
1	Know it well,	understand the	concept	Never heard of it				
				ST62Q09	\			
0				NaN				
1	Know it well,	understand the	concept					
				ST62Q10		ST62Q11	\	
0				NaN		NaN		
1	Know it well,	understand the	concept	Never heard of it				
				ST62Q12		ST62Q13	\	
0				NaN		NaN		
1	Know it well,	understand the	concept	Heard of it	once or	twice		
				ST62Q15	\			
0				NaN				
1	Know it well,	understand the	concept					
				ST62Q16		ST62Q17	\	
0				NaN		NaN		
1	Know it well,	understand the	concept	Never heard of it				
	ST62Q19	ST69Q01	ST69Q02	ST69Q03	ST70Q01	ST70Q02	ST70Q03	\
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1	Heard of it often	45.0	45.0	45.0	7.0	6.0	2.0	
	ST71Q01	ST72Q01	ST73Q01	ST73Q02	ST74Q01	ST74Q02	ST75Q01	\
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

1	NaN	30.0	Frequently	Sometimes	Frequently	Frequently	Sometimes
	ST75Q02	ST76Q01	ST76Q02	ST77Q01	ST77Q02	ST77Q04	\
0	NaN	NaN	NaN	Every Lesson	Every Lesson	Every Lesson	
1	Sometimes	Sometimes	Sometimes	NaN	NaN	NaN	
	ST77Q05	ST77Q06		ST79Q01	ST79Q02		\
0	Every Lesson	Every Lesson	Never or Hardly Ever	Most Lessons			
1	NaN	NaN		NaN	NaN		
		ST79Q03	ST79Q04	ST79Q05	ST79Q06		\
0	Never or Hardly Ever	Every Lesson	Most Lessons	Every Lesson			
1		NaN	NaN	NaN	NaN		
	ST79Q07	ST79Q08		ST79Q10	ST79Q11		\
0	Every Lesson	Every Lesson	Never or Hardly Ever	Most Lessons			
1	NaN	NaN		NaN	NaN		
	ST79Q12	ST79Q15	ST79Q17		ST80Q01		\
0	Every Lesson	Every Lesson	Every Lesson	Always or almost always			
1	NaN	NaN	NaN		NaN		
	ST80Q04	ST80Q05		ST80Q06			\
0	Sometimes	Never or rarely	Always or almost always				
1	NaN	NaN		NaN			
		ST80Q07		ST80Q08		ST80Q09	\
0	Always or almost always	Always or almost always	Always or almost always				
1		NaN		NaN		NaN	
	ST80Q10	ST80Q11		ST81Q01	ST81Q02		\
0	Often	Often	Never or Hardly Ever	Never or Hardly Ever			
1	NaN	NaN	NaN		NaN		
		ST81Q03		ST81Q04	ST81Q05		\
0	Never or Hardly Ever	Never or Hardly Ever	Never or Hardly Ever				
1		NaN		NaN	NaN		
	ST82Q01		ST82Q02		ST82Q03	ST83Q01	\
0	Strongly disagree	Strongly disagree	Strongly disagree	Strongly disagree			
1		NaN	NaN	NaN		NaN	
	ST83Q02	ST83Q03	ST83Q04	ST84Q01	ST84Q02	ST84Q03	ST85Q01
0	Agree	Agree	Agree	Strongly agree	Strongly agree	Disagree	Agree
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
		ST85Q02	ST85Q03	ST85Q04	ST86Q01	ST86Q02	ST86Q03
							\

0	Strongly disagree	Disagree	Agree	Agree	Strongly disagree	Agree
1	NaN	NaN	NaN	NaN	NaN	NaN

	ST86Q04	ST86Q05	ST87Q01	ST87Q02	ST87Q03	ST87Q04	\
0	Agree	Disagree	Agree	Agree	Strongly disagree	Strongly agree	
1	NaN	NaN	NaN	NaN	NaN	NaN	

	ST87Q05	ST87Q06	ST87Q07	ST87Q08	ST87Q09	\
0	Strongly agree	Strongly disagree	Agree	Strongly disagree	Agree	
1	NaN	NaN	NaN	NaN	NaN	

	ST88Q01	ST88Q02	ST88Q03	ST88Q04	ST89Q02	\
0	Agree	Strongly agree	Strongly disagree	Strongly disagree	Agree	
1	NaN	NaN	NaN	NaN	NaN	

	ST89Q03	ST89Q04	ST89Q05	ST91Q01	\
0	Strongly agree	Strongly agree	Strongly agree	Strongly agree	
1	NaN	NaN	NaN	NaN	

	ST91Q02	ST91Q03	ST91Q04	ST91Q05	\
0	Strongly agree	Strongly agree	Strongly disagree	Disagree	
1	NaN	NaN	NaN	NaN	

	ST91Q06	ST93Q01	ST93Q03	\
0	Strongly disagree	Very much like me	Very much like me	
1	NaN	Not at all like me	Not at all like me	

	ST93Q04	ST93Q06	ST93Q07	ST94Q05	\
0	Very much like me	Somewhat like me	Very much like me	Somewhat like me	
1	Mostly like me	Somewhat like me	Very much like me	Somewhat like me	

	ST94Q06	ST94Q09	ST94Q10	ST94Q14	\
0	Mostly like me	Mostly like me	Mostly like me	Somewhat like me	
1	Not much like me	Not much like me	Mostly like me	Not much like me	

	ST96Q01	ST96Q02	ST96Q03	\
0	definitely do this	definitely do this	definitely do this	
1	probably not do this	probably do this	probably not do this	

	ST96Q05	ST101Q01	ST101Q02	ST101Q03	ST101Q05	ST104Q01	\
0	definitely do this	4.0	2.0	1.0	1.0	1.0	
1	probably do this	1.0	2.0	3.0	2.0	2.0	

	ST104Q04	ST104Q05	ST104Q06	IC01Q01	IC01Q02	IC01Q03	IC01Q04	IC01Q05	\
0	2.0	1.0	1.0	NaN	NaN	NaN	NaN	NaN	
1	3.0	1.0	1.0	NaN	NaN	NaN	NaN	NaN	

	IC01Q06	IC01Q07	IC01Q08	IC01Q09	IC01Q10	IC01Q11	IC02Q01	IC02Q02	IC02Q03	\
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

	IC02Q04	IC02Q05	IC02Q06	IC02Q07	IC03Q01	IC04Q01	IC05Q01	IC06Q01	IC07Q01	\
0	NaN	NaN	NaN	NaN	NaN	NaN	99	99	99	
1	NaN	NaN	NaN	NaN	NaN	NaN	99	99	99	

	IC08Q01	IC08Q02	IC08Q03	IC08Q04	IC08Q05	IC08Q06	IC08Q07	IC08Q08	IC08Q09	\
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

	IC08Q11	IC09Q01	IC09Q02	IC09Q03	IC09Q04	IC09Q05	IC09Q06	IC09Q07	IC10Q01	\
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

	IC10Q02	IC10Q03	IC10Q04	IC10Q05	IC10Q06	IC10Q07	IC10Q08	IC10Q09	IC11Q01	\
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

	IC11Q02	IC11Q03	IC11Q04	IC11Q05	IC11Q06	IC11Q07	IC22Q01	IC22Q02	IC22Q04	\
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

	IC22Q06	IC22Q07	IC22Q08	EC01Q01	EC02Q01	EC03Q01	EC03Q02	EC03Q03	EC03Q04	\
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

	EC03Q05	EC03Q06	EC03Q07	EC03Q08	EC03Q09	EC03Q10	EC04Q01A	EC04Q01B	\
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

	EC04Q01C	EC04Q02A	EC04Q02B	EC04Q02C	EC04Q03A	EC04Q03B	EC04Q03C	\
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

	EC04Q04A	EC04Q04B	EC04Q04C	EC04Q05A	EC04Q05B	EC04Q05C	EC04Q06A	\
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

	EC04Q06B	EC04Q06C	EC05Q01	EC06Q01	EC07Q01	EC07Q02	EC07Q03	EC07Q04	EC07Q05	\
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

	EC08Q01	EC08Q02	EC08Q03	EC08Q04	EC09Q03	EC10Q01	EC11Q02	EC11Q03	EC12Q01	\
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

	ST22Q01	ST23Q01	ST23Q02	ST23Q03	ST23Q04	ST23Q05	ST23Q06	ST23Q07	ST23Q08	\
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

	ST24Q01	ST24Q02	ST24Q03		CLCUSE1	CLCUSE301	CLCUSE302	DEFFORT	\
0	NaN	NaN	NaN	A Simple calculator		99	99	99	
1	NaN	NaN	NaN	A Simple calculator		99	99	99	

	QUESTID	BOOKID		EASY	AGE	GRADE	\
0	StQ Form B	booklet 7	Standard set of booklets	16.17	0.0		
1	StQ Form A	booklet 9	Standard set of booklets	16.17	0.0		

	PROGN	ANXMAT	ATSCHL	ATTLNACT	BELONG	\
0	Albania: Upper secondary education	0.32	-2.31	0.5206	-1.18	
1	Albania: Upper secondary education	NaN	NaN	NaN	NaN	

	BFMJ2	BMMJ1	CLSMAN	COBN_F	COBN_M	COBN_S	COGACT	CULTDIST	CULTPOS	\
0	76.49	79.74	-1.3771	Albania	Albania	Albania	0.6994	NaN	-0.48	
1	15.35	23.47	NaN	Albania	Albania	Albania	NaN	NaN	1.27	

	DISCLIMA	ENTUSE	ESCS	EXAPPLM	EXPUREM	FAILMAT	FAMCON	FAMCONC	\
0	1.85	NaN	NaN	NaN	NaN	0.6400	NaN	NaN	
1	NaN	NaN	NaN	-0.0681	0.7955	0.1524	0.6387	-0.08	

	FAMSTRUC		FISCED	HEDRES	HERITCUL		HISCED	HISEI	\
0	2.0	ISCED 3A, ISCED 4	-1.29	NaN	ISCED 3A, ISCED 4	NaN			
1	2.0	ISCED 3A, ISCED 4	1.12	NaN	ISCED 5A, 6	NaN			

	HOMEPOS	HOMSCH	HOSTCUL	ICTATTNEG	ICTATTPOS	ICTHOME	ICTRES	ICTSCH	\
0	-2.61	NaN	NaN	NaN	NaN	NaN	-3.16	NaN	
1	1.41	NaN	NaN	NaN	NaN	NaN	1.15	NaN	

	IMMIG	INFOCAR	INFOJOB1	INFOJOB2	INSTMOT	INTMAT	ISCEDD		ISCEDL	\
0	Native	NaN	NaN	NaN	0.80	0.91	A	ISCED level 3		
1	Native	NaN	NaN	NaN	-0.39	0.00	A	ISCED level 3		

	ISCEDO	LANGCOMM	LANGN	LANGRPPD	LMINS	MATBEH	MATHEFF	MATINTFC	\
0	General	NaN	Albanian	NaN	NaN	0.6426	-0.77	-0.7332	
1	General	NaN	Albanian	NaN	315.0	1.4702	0.34	-0.2514	

	MATWKETH		MISCED	MMINS	MTSUP	\
0	0.2882	ISCED 3A, ISCED 4	NaN	-0.9508		
1	0.6490	ISCED 5A, 6	270.0	NaN		

	OCOD1	\
0	Building architects	

1 Tailors, dressmakers, furriers and hatters

		OCOD2	OPENPS	OUTHOURS	PARED	PERSEV	\
0	Primary school teachers	0.0521		NaN	12.0	-0.3407	
1	Building construction labourers	-0.9492		8.0	16.0	1.3116	

		REPEAT	SCMAT	SMINS	STUDREL	SUBNORM	TCHBEHFA	\
0	Did not repeat a <grade>	0.41	NaN	-1.04	-0.0455		1.3625	
1	Did not repeat a <grade>	NaN	90.0	NaN	0.6602		NaN	

	TCHBEHSO	TCHBEHTD	TEACHSUP	TESTLANG	TIMEINT	USEMATH	USESCH	WEALTH	\
0	0.9374	0.4297	1.68	Albanian	NaN	NaN	NaN	-2.92	
1	NaN	NaN	NaN	Albanian	NaN	NaN	NaN	0.69	

	ANCATSCHL	ANCATTLNACT	ANCBELONG	ANCCLSMAN	ANCCOGACT	ANCINSTMOT	\
0	-1.8636	-0.6779	-0.7351	-0.7808	-0.0219	-0.1562	
1	NaN	NaN	NaN	NaN	NaN	NaN	

	ANCINTMAT	ANCMATWKETH	ANCMTSUP	ANCSCMAT	ANCSTUDREL	ANCSUBNORM	\
0	0.0486	-0.2199	-0.5983	-0.0807	-0.5901	-0.3346	
1	NaN	NaN	NaN	NaN	NaN	NaN	

	PV1MATH	PV2MATH	PV3MATH	PV4MATH	PV5MATH	PV1MACC	PV2MACC	\
0	406.8469	376.4683	344.5319	321.1637	381.9209	325.8374	324.2795	
1	486.1427	464.3325	453.4273	472.9008	476.0165	325.6816	419.9330	

	PV3MACC	PV4MACC	PV5MACC	PV1MACQ	PV2MACQ	PV3MACQ	PV4MACQ	\
0	279.8800	267.4170	312.5954	409.1837	388.1524	373.3525	389.7102	
1	378.6493	359.9548	384.1019	373.1968	444.0801	456.5431	401.2385	

	PV5MACQ	PV1MACS	PV2MACS	PV3MACS	PV4MACS	PV5MACS	PV1MACU	\
0	415.4152	351.5423	375.6894	341.4161	386.5945	426.3203	396.7207	
1	461.2167	366.9653	459.6588	426.1645	423.0488	443.3011	389.5544	

	PV2MACU	PV3MACU	PV4MACU	PV5MACU	PV1MAPE	PV2MAPE	PV3MAPE	\
0	334.4057	328.9531	339.8582	354.6580	324.2795	345.3108	381.1419	
1	438.6275	417.5962	379.4283	438.6275	440.1854	456.5431	486.9216	

	PV4MAPE	PV5MAPE	PV1MAPF	PV2MAPF	PV3MAPF	PV4MAPF	PV5MAPF	\
0	380.363	346.8687	319.6059	345.3108	360.8895	390.4892	322.7216	
1	458.101	444.0801	411.3647	437.8486	457.3220	454.2063	460.4378	

	PV1MAPI	PV2MAPI	PV3MAPI	PV4MAPI	PV5MAPI	PV1READ	PV2READ	\
0	290.7852	345.3108	326.6163	407.6258	367.1210	249.5762	254.3420	
1	434.7328	448.7537	494.7110	429.2803	434.7328	406.2936	349.8975	

	PV3READ	PV4READ	PV5READ	PV1SCIE	PV2SCIE	PV3SCIE	PV4SCIE	\
--	---------	---------	---------	---------	---------	---------	---------	---

0	406.8496	175.7053	218.5981	341.7009	408.8400	348.2283	367.8105	
1	400.7334	369.7553	396.7618	548.9929	471.5964	471.5964	443.6218	
	PV5SCIE	W_FSTUWT	W_FSTR1	W_FSTR2	W_FSTR3	W_FSTR4	W_FSTR5	W_FSTR6 \
0	392.9877	8.9096	13.1249	13.0829	4.5315	13.0829	13.9235	13.1249
1	454.8116	8.9096	13.1249	13.0829	4.5315	13.0829	13.9235	13.1249
	W_FSTR7	W_FSTR8	W_FSTR9	W_FSTR10	W_FSTR11	W_FSTR12	W_FSTR13 \	
0	13.1249	4.3389	4.3313	13.7954	4.5315	4.3313	13.7954	
1	13.1249	4.3389	4.3313	13.7954	4.5315	4.3313	13.7954	
	W_FSTR14	W_FSTR15	W_FSTR16	W_FSTR17	W_FSTR18	W_FSTR19	W_FSTR20 \	
0	13.9235	4.3389	4.3313	4.5084	4.5084	13.7954	4.5315	
1	13.9235	4.3389	4.3313	4.5084	4.5084	13.7954	4.5315	
	W_FSTR21	W_FSTR22	W_FSTR23	W_FSTR24	W_FSTR25	W_FSTR26	W_FSTR27 \	
0	13.1249	13.0829	4.5315	13.0829	13.9235	13.1249	13.1249	
1	13.1249	13.0829	4.5315	13.0829	13.9235	13.1249	13.1249	
	W_FSTR28	W_FSTR29	W_FSTR30	W_FSTR31	W_FSTR32	W_FSTR33	W_FSTR34 \	
0	4.3389	4.3313	13.7954	4.5315	4.3313	13.7954	13.9235	
1	4.3389	4.3313	13.7954	4.5315	4.3313	13.7954	13.9235	
	W_FSTR35	W_FSTR36	W_FSTR37	W_FSTR38	W_FSTR39	W_FSTR40	W_FSTR41 \	
0	4.3389	4.3313	4.5084	4.5084	13.7954	4.5315	4.5084	
1	4.3389	4.3313	4.5084	4.5084	13.7954	4.5315	4.5084	
	W_FSTR42	W_FSTR43	W_FSTR44	W_FSTR45	W_FSTR46	W_FSTR47	W_FSTR48 \	
0	4.5315	13.0829	4.5315	4.3313	4.5084	4.5084	13.7954	
1	4.5315	13.0829	4.5315	4.3313	4.5084	4.5084	13.7954	
	W_FSTR49	W_FSTR50	W_FSTR51	W_FSTR52	W_FSTR53	W_FSTR54	W_FSTR55 \	
0	13.9235	4.3389	13.0829	13.9235	4.3389	4.3313	13.7954	
1	13.9235	4.3389	13.0829	13.9235	4.3389	4.3313	13.7954	
	W_FSTR56	W_FSTR57	W_FSTR58	W_FSTR59	W_FSTR60	W_FSTR61	W_FSTR62 \	
0	13.9235	13.1249	13.1249	4.3389	13.0829	4.5084	4.5315	
1	13.9235	13.1249	13.1249	4.3389	13.0829	4.5084	4.5315	
	W_FSTR63	W_FSTR64	W_FSTR65	W_FSTR66	W_FSTR67	W_FSTR68	W_FSTR69 \	
0	13.0829	4.5315	4.3313	4.5084	4.5084	13.7954	13.9235	
1	13.0829	4.5315	4.3313	4.5084	4.5084	13.7954	13.9235	
	W_FSTR70	W_FSTR71	W_FSTR72	W_FSTR73	W_FSTR74	W_FSTR75	W_FSTR76 \	
0	4.3389	13.0829	13.9235	4.3389	4.3313	13.7954	13.9235	
1	4.3389	13.0829	13.9235	4.3389	4.3313	13.7954	13.9235	

	W_FSTR77	W_FSTR78	W_FSTR79	W_FSTR80	WVARSTRR	VAR_UNIT	SENWGT_STU	\
0	13.1249	13.1249	4.3389	13.0829	19	1	0.2098	
1	13.1249	13.1249	4.3389	13.0829	19	1	0.2098	

	VER_STU
0	22NOV13
1	22NOV13

Let's first drop unused columns that could slow down our work.

```
[5]: cols_to_keep = ['CNT', 'OECD', 'SCHOOLID', 'STIDSTD', 'AGE', 'ST04Q01',
→ 'ST20Q01', 'ST01Q01', 'GRADE',
→ 'ST11Q01', 'ST11Q02', 'ST11Q05', 'PARED', 'ST13Q01', 'ST17Q01',
→ 'ST27Q01', 'ST27Q02', 'ST27Q03',
→ 'ST27Q04', 'ST27Q05', 'ST28Q01', 'ST26Q01', 'ST26Q02',
→ 'ST26Q03', 'ST26Q04', 'ST26Q05', 'ST26Q06',
→ 'ST55Q01', 'ST55Q02', 'ST55Q03', 'LMINS', 'SMINS', 'MMINS',
→ 'PV1MATH', 'PV2MATH', 'PV3MATH', 'PV4MATH', 'PV5MATH',
→ 'PV1SCIE', 'PV2SCIE', 'PV3SCIE',
→ 'PV4SCIE', 'PV5SCIE', 'PV1READ', 'PV2READ', 'PV3READ',
→ 'PV4READ', 'PV5READ']

df = df[cols_to_keep]
df.head(2)
```

```
[5]:      CNT      OECD SCHOOLID STIDSTD      AGE ST04Q01      ST20Q01 \
0  Albania  Non-OECD  0000001   00001  16.17  Female  Country of test
1  Albania  Non-OECD  0000001   00002  16.17  Female  Country of test

      ST01Q01  GRADE ST11Q01 ST11Q02 ST11Q05  PARED      ST13Q01 \
0         10    0.0    Yes     Yes     NaN   12.0  <ISCED level 3A>
1         10    0.0    Yes     Yes     NaN   16.0  <ISCED level 3A>

      ST17Q01      ST27Q01      ST27Q02      ST27Q03 ST27Q04 \
0  <ISCED level 3A>          Two          One          None  None
1  <ISCED level 3A>  Three or more  Three or more  Three or more  Two

      ST27Q05      ST28Q01 ST26Q01 ST26Q02 ST26Q03 ST26Q04 ST26Q05 ST26Q06 \
0    None    0-10 books    Yes     No     Yes     No     No     No
1    Two  201-500 books    Yes     Yes    Yes     Yes    Yes     Yes

      ST55Q01 \
0          NaN
1  I do not attend <out-of-school time lessons> i...

      ST55Q02 \
0          NaN
1  2 or more but less than 4 hours a week
```

		ST55Q03	LMIN5	SMIN5	MMIN5	PV1MATH	\
0		NaN	NaN	NaN	NaN	406.8469	
1	2 or more but less than 4 hours a week	315.0	90.0	270.0	486.1427		

	PV2MATH	PV3MATH	PV4MATH	PV5MATH	PV1SCIE	PV2SCIE	PV3SCIE	\
0	376.4683	344.5319	321.1637	381.9209	341.7009	408.8400	348.2283	
1	464.3325	453.4273	472.9008	476.0165	548.9929	471.5964	471.5964	

	PV4SCIE	PV5SCIE	PV1READ	PV2READ	PV3READ	PV4READ	PV5READ
0	367.8105	392.9877	249.5762	254.3420	406.8496	175.7053	218.5981
1	443.6218	454.8116	406.2936	349.8975	400.7334	369.7553	396.7618

We can look at meaning of columns using PISA dictionary data.

```
[6]: # Load in dict data

dict_df = pd.read_csv(dict_path, header=None, names=['value', 'meaning'],
                    skiprows=1, encoding='ISO-8859-1')
dict_df.head()
```

```
[6]:      value      meaning
0      CNT      Country code 3-character
1  SUBNATIO  Adjudicated sub-region code 7-digit code (3-di...
2  STRATUM   Stratum ID 7-character (cnt + region ID + orig...
3      OECD      OECD country
4      NC       National Centre 6-digit Code
```

```
[7]: dict_df[dict_df['value'].isin(df.columns)]
```

```
[7]:      value      meaning
0      CNT      Country code 3-character
3      OECD      OECD country
5  SCHOOLID  School ID 7-digit (region ID + stratum ID + 3-...
6  STIDSTD      Student ID
7  ST01Q01      International Grade
11 ST04Q01      Gender
20 ST11Q01      At Home - Mother
21 ST11Q02      At Home - Father
24 ST11Q05      At Home - Grandparents
26 ST13Q01      Mother<Highest Schooling>
32 ST17Q01      Father<Highest Schooling>
38 ST20Q01      Country of Birth International - Self
43 ST26Q01      Possessions - desk
44 ST26Q02      Possessions - own room
45 ST26Q03      Possessions - study place
46 ST26Q04      Possessions - computer
47 ST26Q05      Possessions - software
48 ST26Q06      Possessions - Internet
60 ST27Q01      How many - cellular phones
```

61	ST27Q02	How many - televisions
62	ST27Q03	How many - computers
63	ST27Q04	How many - cars
64	ST27Q05	How many - rooms bath or shower
65	ST28Q01	How many books at home
136	ST55Q01	Out of school lessons - <test lang>
137	ST55Q02	Out of school lessons - <maths>
138	ST55Q03	Out of school lessons - <science>
410	AGE	Age of student
411	GRADE	Grade compared to modal grade in country
460	LMINS	Learning time (minutes per week) - <test lang...>
466	MMINS	Learning time (minutes per week)- <Mathematics>
472	PARED	Highest parental education in years
476	SMINS	Learning time (minutes per week) - <Science>
500	PV1MATH	Plausible value 1 in mathematics
501	PV2MATH	Plausible value 2 in mathematics
502	PV3MATH	Plausible value 3 in mathematics
503	PV4MATH	Plausible value 4 in mathematics
504	PV5MATH	Plausible value 5 in mathematics
540	PV1READ	Plausible value 1 in reading
541	PV2READ	Plausible value 2 in reading
542	PV3READ	Plausible value 3 in reading
543	PV4READ	Plausible value 4 in reading
544	PV5READ	Plausible value 5 in reading
545	PV1SCIE	Plausible value 1 in science
546	PV2SCIE	Plausible value 2 in science
547	PV3SCIE	Plausible value 3 in science
548	PV4SCIE	Plausible value 4 in science
549	PV5SCIE	Plausible value 5 in science

1.1.1 General Properties

```
[8]: print('Shape:', df.shape[0], 'rows and', df.shape[1], 'columns')
```

Shape: 485490 rows and 48 columns

There're 1471 schools with 33806 students in the dataset.

```
[9]: df.SCHOOLID.nunique(), df.STIDSTD.nunique()
```

```
[9]: (1471, 33806)
```

Schools and students identifiers were loaded as integers first time, so their types were specified directly when loading data.

```
[10]: df.dtypes
```

```
[10]: CNT          object
      OECD         object
      SCHOOLID     object
```

STIDSTD	object
AGE	float64
ST04Q01	object
ST20Q01	object
ST01Q01	int64
GRADE	float64
ST11Q01	object
ST11Q02	object
ST11Q05	object
PARED	float64
ST13Q01	object
ST17Q01	object
ST27Q01	object
ST27Q02	object
ST27Q03	object
ST27Q04	object
ST27Q05	object
ST28Q01	object
ST26Q01	object
ST26Q02	object
ST26Q03	object
ST26Q04	object
ST26Q05	object
ST26Q06	object
ST55Q01	object
ST55Q02	object
ST55Q03	object
LMINS	float64
SMINS	float64
MMINS	float64
PV1MATH	float64
PV2MATH	float64
PV3MATH	float64
PV4MATH	float64
PV5MATH	float64
PV1SCIE	float64
PV2SCIE	float64
PV3SCIE	float64
PV4SCIE	float64
PV5SCIE	float64
PV1READ	float64
PV2READ	float64
PV3READ	float64
PV4READ	float64
PV5READ	float64
dtype:	object


```
[11]: # First, I think, "ST55Q.." clumms should be numeric
```

```
df.ST55Q01.unique()
```

```
[11]: array([nan,  
        'I do not attend <out-of-school time lessons> in this subject',  
        'Less than 2 hours a week',  
        '4 or more but less than 6 hours a week',  
        '2 or more but less than 4 hours a week', '6 or more hours a week'],  
       dtype=object)
```

Most rows have some missing values.

```
[12]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 485490 entries, 0 to 485489  
Data columns (total 48 columns):  
CNT          485490 non-null object  
OECD         485490 non-null object  
SCHOOLID     485490 non-null object  
STIDSTD      485490 non-null object  
AGE          485374 non-null float64  
ST04Q01      485490 non-null object  
ST20Q01      476363 non-null object  
ST01Q01      485490 non-null int64  
GRADE        484617 non-null float64  
ST11Q01      460559 non-null object  
ST11Q02      441036 non-null object  
ST11Q05      348180 non-null object  
PARED        473091 non-null float64  
ST13Q01      457979 non-null object  
ST17Q01      443261 non-null object  
ST27Q01      477079 non-null object  
ST27Q02      476548 non-null object  
ST27Q03      473459 non-null object  
ST27Q04      472499 non-null object  
ST27Q05      469643 non-null object  
ST28Q01      473765 non-null object  
ST26Q01      473079 non-null object  
ST26Q02      469693 non-null object  
ST26Q03      472020 non-null object  
ST26Q04      473877 non-null object  
ST26Q05      463178 non-null object  
ST26Q06      473182 non-null object  
ST55Q01      307761 non-null object  
ST55Q02      308171 non-null object  
ST55Q03      306090 non-null object  
LMINS        282866 non-null float64
```

```

SMINS      270914 non-null float64
MMINS      283303 non-null float64
PV1MATH    485490 non-null float64
PV2MATH    485490 non-null float64
PV3MATH    485490 non-null float64
PV4MATH    485490 non-null float64
PV5MATH    485490 non-null float64
PV1SCIE    485490 non-null float64
PV2SCIE    485490 non-null float64
PV3SCIE    485490 non-null float64
PV4SCIE    485490 non-null float64
PV5SCIE    485490 non-null float64
PV1READ    485490 non-null float64
PV2READ    485490 non-null float64
PV3READ    485490 non-null float64
PV4READ    485490 non-null float64
PV5READ    485490 non-null float64
dtypes: float64(21), int64(1), object(26)
memory usage: 177.8+ MB

```

[13]: *# Check missing data*

```

df_missing = df.isnull().sum().sort_values(ascending=False)
df_missing[df_missing > 0] / df.shape[0]

```

[13]:

SMINS	0.441978
LMINS	0.417360
MMINS	0.416460
ST55Q03	0.369524
ST55Q01	0.366082
ST55Q02	0.365237
ST11Q05	0.282828
ST11Q02	0.091565
ST17Q01	0.086982
ST13Q01	0.056666
ST11Q01	0.051352
ST26Q05	0.045958
ST27Q05	0.032641
ST26Q02	0.032538
ST26Q03	0.027745
ST27Q04	0.026759
ST26Q01	0.025564
PARED	0.025539
ST26Q06	0.025352
ST27Q03	0.024781
ST28Q01	0.024151
ST26Q04	0.023920
ST20Q01	0.018800

```
ST27Q02    0.018419
ST27Q01    0.017325
GRADE      0.001798
AGE         0.000239
dtype: float64
```

```
[14]: df[['SMINS', 'LMINS', 'MMINS']].head(10)
```

```
[14]:   SMINS  LMINS  MMINS
0    NaN    NaN    NaN
1   90.0  315.0  270.0
2    NaN  300.0    NaN
3   90.0  135.0  135.0
4    NaN    NaN    NaN
5    NaN    NaN    NaN
6   90.0  135.0  225.0
7    NaN    NaN    NaN
8    NaN    NaN    NaN
9  270.0  240.0   90.0
```

Columns SMINS, LMINS, and MMINS (that's Learning time (minutes per week)) have more than 40% missing values. So, last 7th question could be modified as follows:

Whether absense of data about learning time on math, science, and reading, affect students performance? And if a student report info about total time learning, how these influence each grade in assesment?

The same could be applied to ST55Q03, ST55Q01, and ST55Q02 (Out of school lessons) because of high percent (about 36%) of missingvalues.

Finally, it's also more than a quater missig values in ST11Q05 column (At Home - Grandparents). He we can asume, NaN in ST11Q05 could be also considered as missing data.

```
[15]: df.ST11Q05.unique()
```

```
[15]: array([nan, 'No', 'Yes'], dtype=object)
```

In all other columns share of missing data isn't exceed than 10%. So, for explanatory data analysis, we could ignore this distortions, and drop missing values.

```
[16]: for column in df_missing.index:
        if column not in ('SMINS', 'LMINS', 'MMINS', 'ST55Q03', 'ST55Q01',
        → 'ST55Q02', 'ST11Q05'):
            df[column].fillna(df[column].mode()[0], inplace=True)

# Check NaN
df_missing = df.isnull().sum().sort_values(ascending=False)
df_missing[df_missing > 0] / df.shape[0]
```

```
[16]: SMINS      0.441978
      LMINS      0.417360
      MMINS      0.416460
      ST55Q03    0.369524
```

```
ST55Q01    0.366082
ST55Q02    0.365237
ST11Q05    0.282828
dtype: float64
```

```
[17]: # Check duplicated rows
```

```
df.duplicated().sum() # => there's no duplicated rows
```

```
[17]: 0
```

```
[18]: # Summary statistics of numeric columns
```

```
df.describe()
```

```
[18]:
```

	AGE	ST01Q01	GRADE	PARED	\
count	485490.000000	485490.000000	485490.000000	485490.000000	
mean	15.784234	9.813323	-0.162671	12.969808	
std	0.290203	3.734726	0.655005	3.358615	
min	15.170000	7.000000	-3.000000	3.000000	
25%	15.580000	9.000000	0.000000	12.000000	
50%	15.750000	10.000000	0.000000	13.000000	
75%	16.000000	10.000000	0.000000	16.000000	
max	16.330000	96.000000	3.000000	18.000000	

	LMINS	SMINS	MMINS	PV1MATH	\
count	282866.000000	270914.000000	283303.000000	485490.000000	
mean	219.276636	211.122460	226.007056	469.621653	
std	97.997730	131.368322	97.448421	103.265391	
min	0.000000	0.000000	0.000000	19.792800	
25%	165.000000	120.000000	180.000000	395.318600	
50%	200.000000	180.000000	220.000000	466.201900	
75%	250.000000	270.000000	250.000000	541.057800	
max	2400.000000	2975.000000	3000.000000	962.229300	

	PV2MATH	PV3MATH	PV4MATH	PV5MATH	\
count	485490.000000	485490.000000	485490.000000	485490.000000	
mean	469.648358	469.648930	469.641832	469.695396	
std	103.382077	103.407631	103.392286	103.419170	
min	6.473000	42.226200	24.622200	37.085200	
25%	395.318600	395.240700	395.396500	395.240700	
50%	466.124000	466.201900	466.279800	466.435600	
75%	541.447300	541.291500	541.447300	541.447300	
max	957.010400	935.745400	943.456900	907.625800	

	PV1SCIE	PV2SCIE	PV3SCIE	PV4SCIE	\
count	485490.000000	485490.000000	485490.000000	485490.000000	
mean	475.769824	475.813674	475.851549	475.78524	
std	101.464426	101.514649	101.495072	101.51220	

min	2.648300	2.834800	11.879900	8.42970
25%	404.457300	404.457300	404.550500	404.45730
50%	475.699400	475.606100	475.699400	475.97910
75%	547.780700	547.873900	547.967200	547.78070
max	903.338300	900.540800	867.624000	926.55730

	PV5SCIE	PV1READ	PV2READ	PV3READ \
count	485490.000000	485490.000000	485490.000000	485490.000000
mean	475.820184	472.004640	472.068052	472.022059
std	101.566347	102.505523	102.626198	102.640489
min	17.754600	0.083400	0.703500	0.703500
25%	404.457300	403.600700	403.360100	403.360100
50%	475.885900	475.455000	475.535200	475.455000
75%	547.780700	544.502500	544.503500	544.503500
max	880.958600	904.802600	881.239200	884.447000

	PV4READ	PV5READ
count	485490.000000	485490.000000
mean	471.926562	472.013506
std	102.576066	102.659989
min	4.134400	2.307400
25%	403.354600	403.360100
50%	475.535200	475.535200
75%	544.502500	544.503500
max	881.159000	901.608600

Part I - Data Wrangling and Data Exploration

[19]: *# Check values of object columns , and simplify where possible*

```
cat_cols = df.select_dtypes(include='object').columns
cat_cols = cat_cols.drop(['SCHOOLID', 'STIDSTD', 'CNT'])
print('Number of string columns', cat_cols.shape[0])
cat_cols
```

Number of string columns 23

[19]: Index(['OECD', 'ST04Q01', 'ST20Q01', 'ST11Q01', 'ST11Q02', 'ST11Q05',
 'ST13Q01', 'ST17Q01', 'ST27Q01', 'ST27Q02', 'ST27Q03', 'ST27Q04',
 'ST27Q05', 'ST28Q01', 'ST26Q01', 'ST26Q02', 'ST26Q03', 'ST26Q04',
 'ST26Q05', 'ST26Q06', 'ST55Q01', 'ST55Q02', 'ST55Q03'],
 dtype='object')

There's a lot of categories in CNT column (countries). Look at them separately.

[20]: `for column in cat_cols:`
`print(column, df[column].nunique(), df[column].unique())`

OECD 2 ['Non-OECD' 'OECD']
 ST04Q01 2 ['Female' 'Male']

```

ST20Q01 2 ['Country of test' 'Other country']
ST11Q01 2 ['Yes' 'No']
ST11Q02 2 ['Yes' 'No']
ST11Q05 2 [nan 'No' 'Yes']
ST13Q01 5 ['<ISCED level 3A> ' '<ISCED level 3B, 3C> '
'She did not complete <ISCED level 1> ' '<ISCED level 2> '
'<ISCED level 1> ']
ST17Q01 5 ['<ISCED level 3A> ' '<ISCED level 3B, 3C> ' '<ISCED level 2> '
'He did not complete <ISCED level 1> ' '<ISCED level 1> ']
ST27Q01 4 ['Two' 'Three or more' 'One' 'None']
ST27Q02 4 ['One' 'Three or more' 'Two' 'None']
ST27Q03 4 ['None' 'Three or more' 'Two' 'One']
ST27Q04 4 ['None' 'Two' 'One' 'Three or more']
ST27Q05 4 ['None' 'Two' 'One' 'Three or more']
ST28Q01 6 ['0-10 books ' '201-500 books ' 'More than 500 books' '11-25 books '
'101-200 books ' '26-100 books ']
ST26Q01 2 ['Yes' 'No']
ST26Q02 2 ['No' 'Yes']
ST26Q03 2 ['Yes' 'No']
ST26Q04 2 ['No' 'Yes']
ST26Q05 2 ['No' 'Yes']
ST26Q06 2 ['No' 'Yes']
ST55Q01 5 [nan 'I do not attend <out-of-school time lessons> in this subject'
'Less than 2 hours a week' '4 or more but less than 6 hours a week'
'2 or more but less than 4 hours a week' '6 or more hours a week']
ST55Q02 5 [nan '2 or more but less than 4 hours a week'
'I do not attend <out-of-school time lessons> in this subject'
'Less than 2 hours a week' '6 or more hours a week'
'4 or more but less than 6 hours a week']
ST55Q03 5 [nan '2 or more but less than 4 hours a week'
'4 or more but less than 6 hours a week' 'Less than 2 hours a week'
'I do not attend <out-of-school time lessons> in this subject'
'6 or more hours a week']

```

For column “How many properties are at home?” replace “None” and NaN values with zeros.

```

[21]: cols_with_none = 'ST27Q01', 'ST27Q02', 'ST27Q03', 'ST27Q04', 'ST27Q05'
      for column in cols_with_none:
          df[column] = df[column].replace(['None', np.nan], 'Zero')

[22]: print('Number of unique countries', df['CNT'].nunique())
      df.CNT.unique()

```

Number of unique countries 68

```

[22]: array(['Albania', 'United Arab Emirates', 'Argentina', 'Australia',
'Austria', 'Belgium', 'Bulgaria', 'Brazil', 'Canada',
'Switzerland', 'Chile', 'Colombia', 'Costa Rica', 'Czech Republic',
'Germany', 'Denmark', 'Spain', 'Estonia', 'Finland', 'France',

```

```
'United Kingdom', 'Greece', 'Hong Kong-China', 'Croatia',
'Hungary', 'Indonesia', 'Ireland', 'Iceland', 'Israel', 'Italy',
'Jordan', 'Japan', 'Kazakhstan', 'Korea', 'Liechtenstein',
'Lithuania', 'Luxembourg', 'Latvia', 'Macao-China', 'Mexico',
'Montenegro', 'Malaysia', 'Netherlands', 'Norway', 'New Zealand',
'Peru', 'Poland', 'Portugal', 'Qatar', 'China-Shanghai',
'Perm(Russian Federation)', 'Florida (USA)', 'Connecticut (USA)',
'Massachusetts (USA)', 'Romania', 'Russian Federation',
'Singapore', 'Serbia', 'Slovak Republic', 'Slovenia', 'Sweden',
'Chinese Taipei', 'Thailand', 'Tunisia', 'Turkey', 'Uruguay',
'United States of America', 'Vietnam'], dtype=object)
```

Some change could be made: 1. Hong Kong-China -> Hong Kong 2. China-Shanghai -> China 3. Perm(Russian Federation) -> Russian Federation (since Perm is just a city in RF) 4. Florida (USA) -> United States of America 5. Connecticut (USA) -> United States of America 6. Massachusetts (USA) -> United States of America 7. Chinese Taipei -> Taiwan 8. Macao-China -> Macao

[23]: *# Implement changes in NCT column*

```
df['CNT'] = (df['CNT'].replace('Hong Kong-China', 'Hong Kong')
            .replace('China-Shanghai', 'China')
            .replace('Perm(Russian Federation)', 'Russian_
→Federation')
            .replace('Florida (USA)', 'United States of America')
            .replace('Connecticut (USA)', 'United States of_
→America')
            .replace('Connecticut (USA)', 'United States of_
→America')
            .replace('Massachusetts (USA)', 'United States of_
→America')
            .replace('Chinese Taipei', 'Taiwan')
            .replace('Macao-China', 'Macao'))
```

[24]: *# Check changes*

```
print('Number of unique countries', df['CNT'].nunique())
df['CNT'].unique()
```

Number of unique countries 64

[24]: array(['Albania', 'United Arab Emirates', 'Argentina', 'Australia',
'Austria', 'Belgium', 'Bulgaria', 'Brazil', 'Canada',
'Switzerland', 'Chile', 'Colombia', 'Costa Rica', 'Czech Republic',
'Germany', 'Denmark', 'Spain', 'Estonia', 'Finland', 'France',
'United Kingdom', 'Greece', 'Hong Kong', 'Croatia', 'Hungary',
'Indonesia', 'Ireland', 'Iceland', 'Israel', 'Italy', 'Jordan',
'Japan', 'Kazakhstan', 'Korea', 'Liechtenstein', 'Lithuania',
'Luxembourg', 'Latvia', 'Macao', 'Mexico', 'Montenegro',

```
'Malaysia', 'Netherlands', 'Norway', 'New Zealand', 'Peru',
'Poland', 'Portugal', 'Qatar', 'China', 'Russian Federation',
'United States of America', 'Romania', 'Singapore', 'Serbia',
'Slovak Republic', 'Slovenia', 'Sweden', 'Taiwan', 'Thailand',
'Tunisia', 'Turkey', 'Uruguay', 'Vietnam'], dtype=object)
```

Finally, the independent variables “PV...” - plausible values in math, science, and reading - will be summed and divided by 5 (number of column of each subject).

```
[25]: df['PV_MATH'] = (df.PV1MATH + df.PV2MATH + df.PV3MATH + df.PV4MATH + df.
    ↪PV5MATH) / 5
df['PV_SCIE'] = (df.PV1SCIE + df.PV2SCIE + df.PV3SCIE + df.PV4SCIE + df.
    ↪PV5SCIE) / 5
df['PV_READ'] = (df.PV1READ + df.PV2READ + df.PV3READ + df.PV4READ + df.
    ↪PV5READ) / 5

df[['PV_MATH', 'PV_SCIE', 'PV_READ']].describe()
```

```
[25]:
```

	PV_MATH	PV_SCIE	PV_READ
count	485490.000000	485490.000000	485490.000000
mean	469.651234	475.808094	472.006964
std	100.786610	97.998470	98.863310
min	54.767080	25.158540	6.445400
25%	396.019620	405.762800	405.044200
50%	465.734520	475.512860	475.477980
75%	540.123060	546.381920	542.831195
max	903.107960	857.832900	849.359740

```
[26]: # Drop initial "PV..." columns

df.drop(['PV1MATH', 'PV2MATH', 'PV3MATH', 'PV4MATH', 'PV5MATH', 'PV1SCIE',
    ↪'PV2SCIE', 'PV3SCIE', 'PV4SCIE',
    ↪'PV5SCIE', 'PV1READ', 'PV2READ', 'PV3READ', 'PV4READ', 'PV5READ'],
    ↪axis=1, inplace=True)
```

Part II - Explanatory Data Analysis

```
[27]: df.head(3)
```

```
[27]:
```

	CNT	OECD	SCHOOLID	STIDSTD	AGE	ST04Q01	ST20Q01 \
0	Albania	Non-OECD	0000001	00001	16.17	Female	Country of test
1	Albania	Non-OECD	0000001	00002	16.17	Female	Country of test
2	Albania	Non-OECD	0000001	00003	15.58	Female	Country of test

	ST01Q01	GRADE	ST11Q01	ST11Q02	ST11Q05	PARED	ST13Q01 \
0	10	0.0	Yes	Yes	NaN	12.0	<ISCED level 3A>
1	10	0.0	Yes	Yes	NaN	16.0	<ISCED level 3A>
2	9	-1.0	Yes	Yes	No	16.0	<ISCED level 3B, 3C>

	ST17Q01	ST27Q01	ST27Q02	ST27Q03	ST27Q04 \
0	<ISCED level 3A>	Two	One	Zero	Zero

1	<ISCED level 3A>	Three or more	Three or more	Three or more	Three or more	Two
2	<ISCED level 3A>	Three or more		Two	Two	One

	ST27Q05		ST28Q01	ST26Q01	ST26Q02	ST26Q03	ST26Q04	ST26Q05	\
0	Zero	0-10 books	Yes	No	Yes	No	No		
1	Two	201-500 books	Yes	Yes	Yes	Yes	Yes	Yes	
2	Two	More than 500 books	Yes	Yes	Yes	Yes	Yes	No	

	ST26Q06		ST55Q01	\
0	No		NaN	
1	Yes	I do not attend <out-of-school time lessons> i...		
2	Yes	Less than 2 hours a week		

	ST55Q02	\
0	NaN	
1	2 or more but less than 4 hours a week	
2	2 or more but less than 4 hours a week	

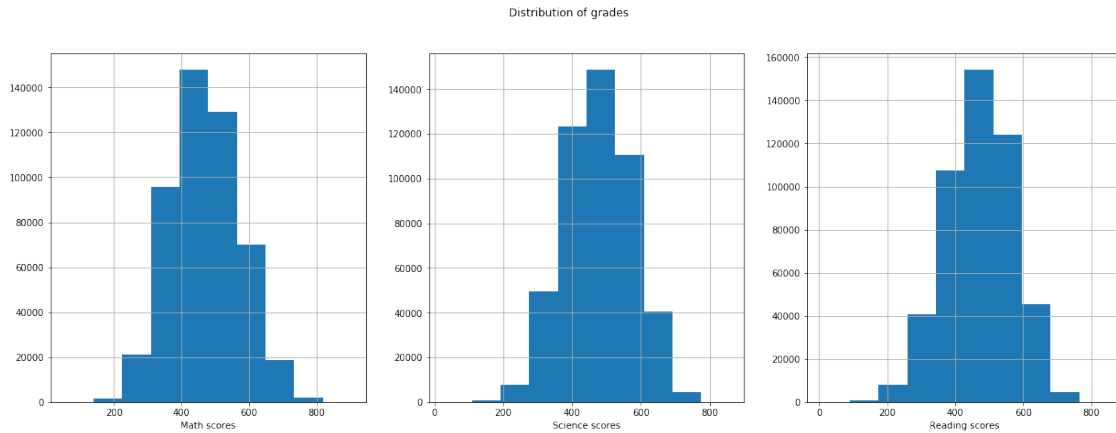
	ST55Q03	LMINS	SMINS	MMINS	PV_MATH	\
0	NaN	NaN	NaN	NaN	366.18634	
1	2 or more but less than 4 hours a week	315.0	90.0	270.0	470.56396	
2	4 or more but less than 6 hours a week	300.0	NaN	NaN	505.53824	

	PV_SCIE	PV_READ
0	371.91348	261.01424
1	478.12382	384.68832
2	486.60946	405.18154

1. What is students' performance at schools in different countries (including whether country is a OECD member).

```
[28]: base_color = sns.color_palette()[0]
```

```
[29]: fig, ax = plt.subplots(nrows=1, ncols=3, figsize=(21,7))
df['PV_MATH'].hist(ax=ax[0])
df['PV_SCIE'].hist(ax=ax[1])
df['PV_READ'].hist(ax=ax[2])
ax[0].set_xlabel('Math scores')
ax[1].set_xlabel('Science scores')
ax[2].set_xlabel('Reading scores')
plt.suptitle('Distribution of grades')
plt.show()
```



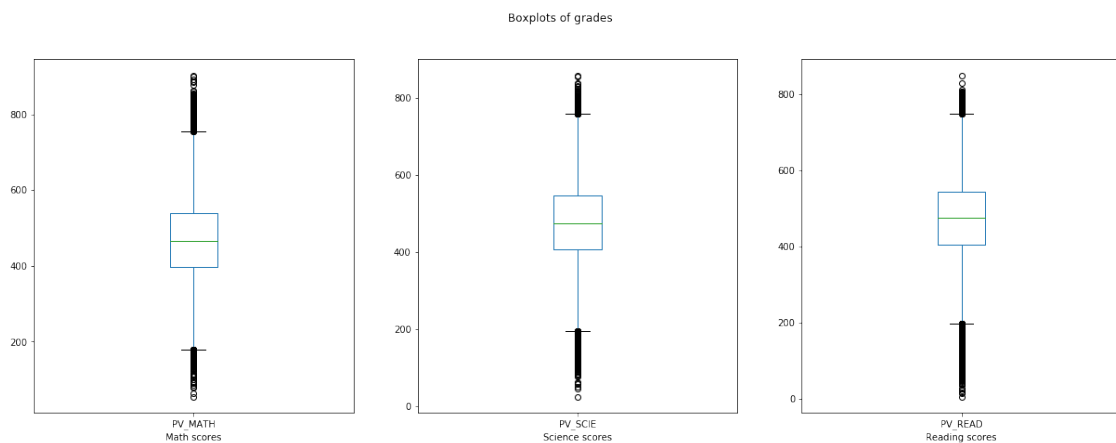
```
[30]: df['PV_MATH'].mean(), df['PV_SCIE'].mean(), df['PV_READ'].mean()
```

```
[30]: (469.65123385442615, 475.80809403002854, 472.0069640898506)
```

If we plot all the grades by subject, then scores in each subject looks normally distributed. Mean scores of science are about 3 points higher than average reading scores. In its turn, average reading scores are about 3 points higher than average math scores. So scores in those 3 subjects are very similar.

So, let's look at their boxplots.

```
[31]: fig, ax = plt.subplots(nrows=1, ncols=3, figsize=(21,7))
df['PV_MATH'].plot(kind='box', ax=ax[0])
df['PV_SCIE'].plot(kind='box', ax=ax[1])
df['PV_READ'].plot(kind='box', ax=ax[2])
ax[0].set_xlabel('Math scores')
ax[1].set_xlabel('Science scores')
ax[2].set_xlabel('Reading scores')
plt.suptitle('Boxplots of grades')
plt.show()
```



In general, there're outliers in every Series of scores. Moreover, math scores have approximately equal tails of outliers, but science and reading scores have outliers with lower scores more, than outliers with higher scores. let's go deeper, and look at students performance in the context of countries, OECD membership, and other columns.

```
[32]: df_plot1 = df.groupby('CNT').agg({'PV_MATH': 'mean', 'PV_SCIE': 'mean',
    → 'PV_READ': 'mean', 'STIDSTD': 'nunique'})

print('Shape', df_plot1.shape)
df_plot1.head()
```

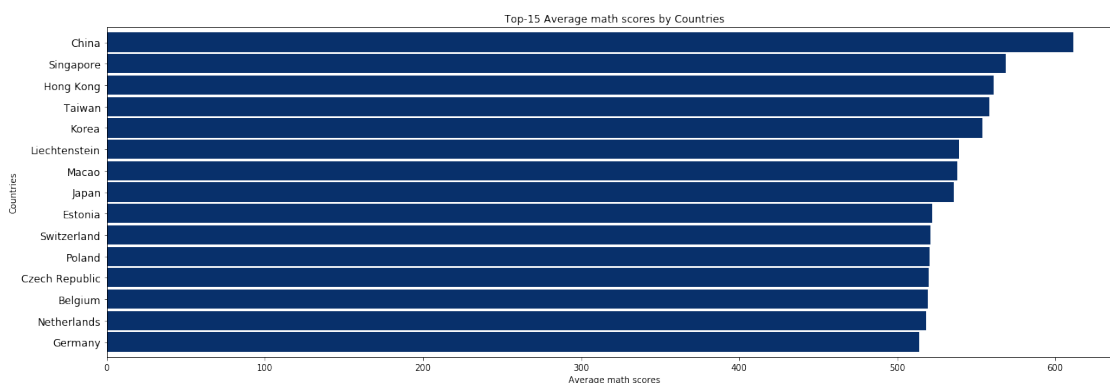
Shape (64, 4)

```
[32]:
```

	PV_MATH	PV_SCIE	PV_READ	STIDSTD
CNT				
Albania	394.878912	398.916529	396.250245	4743
Argentina	395.635711	410.478404	403.596060	5908
Australia	493.268939	511.638212	501.056931	14481
Austria	507.778785	508.036810	491.485551	4755
Belgium	519.668410	510.302595	512.281728	8597

```
[33]: df_plot1[['PV_MATH']].sort_values('PV_MATH').iloc[-15:].plot.
    → barh(figsize=(21,7), width=.9, legend=False,

    → cmap='Blues_r')
plt.yticks(fontsize=12)
plt.xlabel('Average math scores')
plt.ylabel('Countries')
plt.title('Top-15 Average math scores by Countries')
plt.show()
```

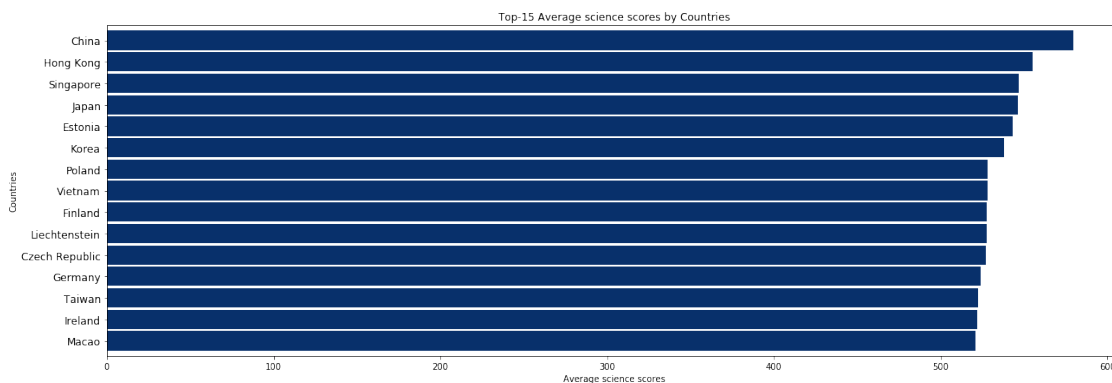


```
[34]: df[df.CNT == 'China'].PV_MATH.mean(), df[df.CNT == 'Peru'].PV_MATH.mean()
```

```
[34]: (611.4389329882152, 367.8596761126761)
```

Except Liechtenstein which is on the 6th position, on average, students from Asia countries receive the highest scores on math. China, Singapore, Hong Kong, Taiwan, and Korea are in Top-5. Macao and Japan follow immediately behind Liechtenstein. Chinese students receive on average 611 points. In comparison, in Peru average math scores are equal 368. This's 1.7 times less than in China.

```
[35]: df_plot1[['PV_SCIE']].sort_values('PV_SCIE').iloc[-15:].plot.  
      →barh(figsize=(21,7), width=.9, legend=False,  
  
      →cmap='Blues_r')  
plt.yticks(fontsize=12)  
plt.xlabel('Average science scores')  
plt.ylabel('Countries')  
plt.title('Top-15 Average science scores by Countries')  
plt.show()
```

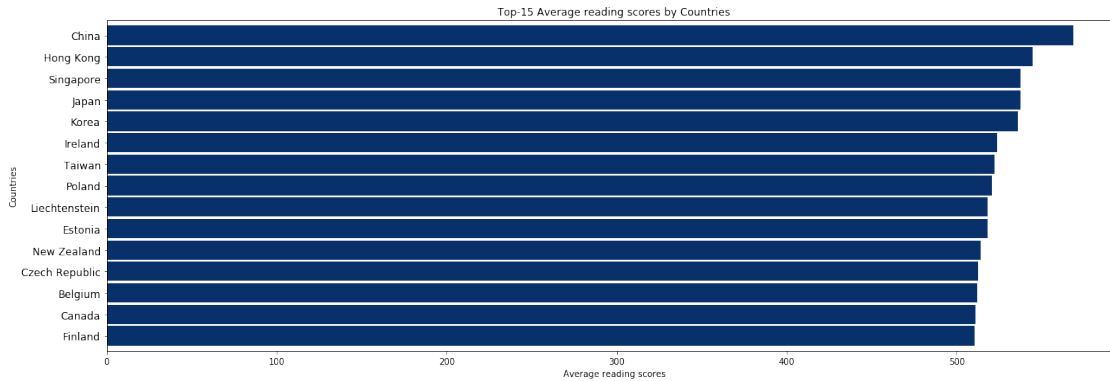


```
[36]: (df[df.CNT == 'China'].PV_SCIE.mean(), df[df.CNT == 'Hong Kong'].PV_SCIE.mean(),  
      df[df.CNT == 'Singapore'].PV_SCIE.mean())
```

```
[36]: (579.5565404481328, 554.9864334004274, 546.8229195961078)
```

Average science scores are less than math scores by about 6 points. And this is becoming noticeable for countries with the highest average scores in science. China, Gang Kong and Singapore are also in the Top-3 with an average score of 547 to 579. For China, this difference is 32 points or 5.2%.

```
[37]: df_plot1[['PV_READ']].sort_values('PV_READ').iloc[-15:].plot.  
      →barh(figsize=(21,7), width=.9, legend=False,  
  
      →cmap='Blues_r')  
plt.yticks(fontsize=12)  
plt.xlabel('Average reading scores')  
plt.ylabel('Countries')  
plt.title('Top-15 Average reading scores by Countries')  
plt.show()
```



```
[38]: df[df.CNT == 'China'].PV_READ.mean()
```

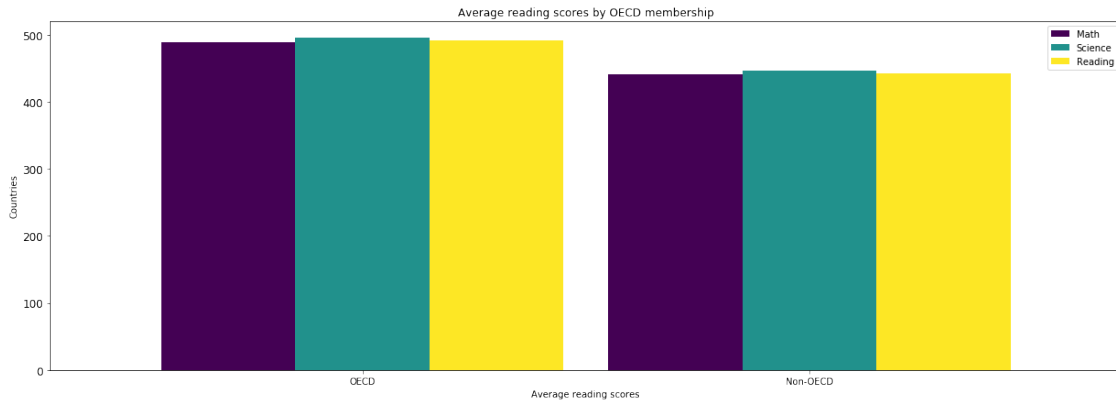
```
[38]: 568.6292328568668
```

For average reading scores, China, Hong Kong, Singapore, Japan, Korea and Taiwan continue to be the leaders with a maximum average of 569 points for China. This average score is the lowest for China in three subjects, possibly also because English is not a native language for a large population of the country.

```
[39]: df_plot2 = df.groupby('OECD').agg({'PV_MATH': 'mean', 'PV_SCIE': 'mean', 'PV_READ': 'mean', 'STIDSTD': 'nunique'})
print(df_plot2)
```

	PV_MATH	PV_SCIE	PV_READ	STIDSTD
OECD				
Non-OECD	440.509684	446.101570	442.803610	19204
OECD	488.401237	494.921608	490.796733	33806

```
[40]: df_plot2[['PV_MATH', 'PV_SCIE', 'PV_READ']].sort_values('PV_READ',
    →ascending=False).plot.bar(
    →figsize=(21,7), width=.9, legend=True, cmap='viridis',
    →rot=0)
plt.legend(['Math', 'Science', 'Reading'])
plt.yticks(fontsize=12)
plt.xlabel('Average reading scores')
plt.ylabel('Countries')
plt.title('Average reading scores by OECD membership')
plt.show()
```

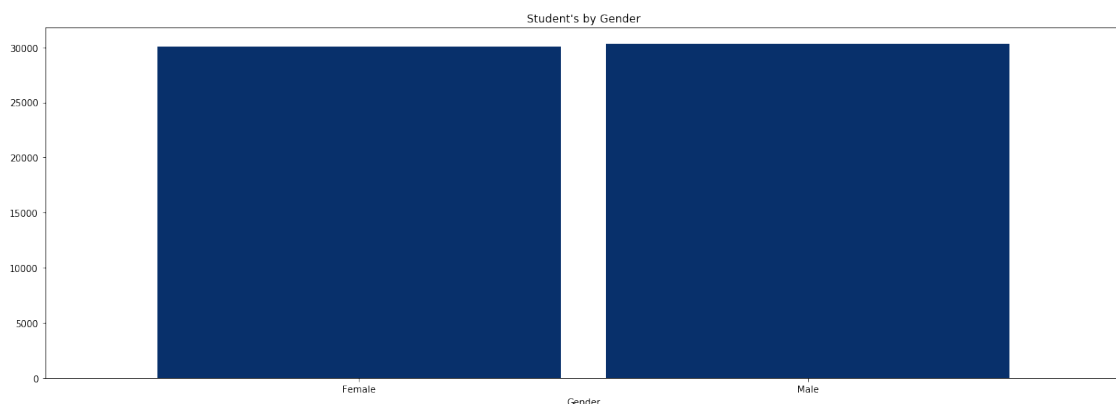


The difference is noticeable in all three subjects at once: average scores in mathematics, science and reading are higher in OSCE countries than in non-OSCE countries. The difference is about 48 points for each subject.

2. What are the characteristics of students participated in PICA 2012:

- gender,
- age,
- whether a student passed the test in the country of birth or not,
- international grade and grade compared to modal grade in country.

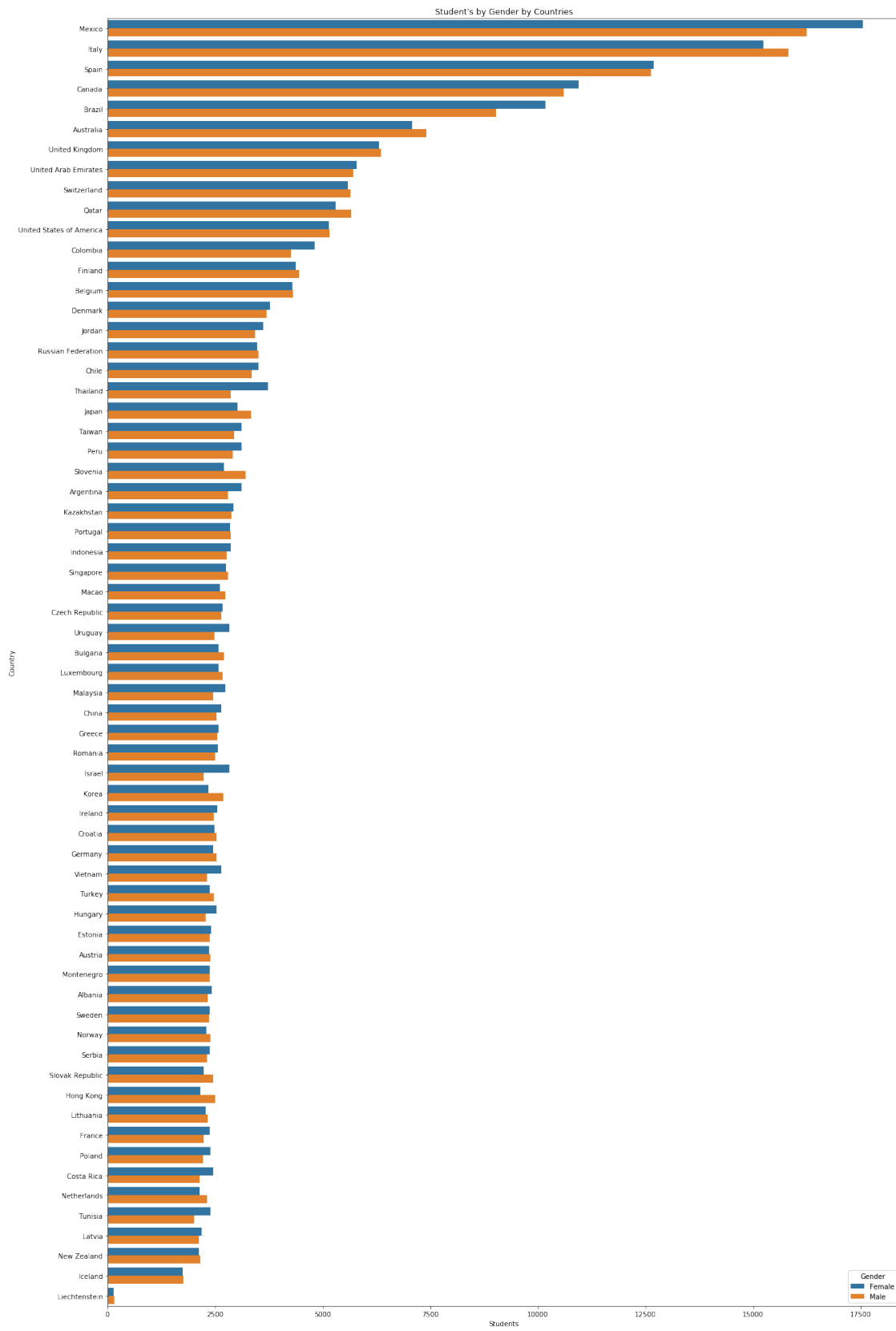
```
[41]: df.groupby('ST04Q01').STIDSTD.nunique().sort_index().plot.bar(figsize=(21,7),
    ↳width=.9, legend=False,
    ↳cmap='Blues_r', rot=0)
plt.xlabel('Gender')
plt.title('Student\'s by Gender')
plt.show()
```



Number of female students is little more (by 0.6%) than number of male students. Let's look at gender by countries.

The largest number of students are in Mexico, Italy, Spain, Canada and Brazil. Except for Italy, the number of the females is greater than that of the male. In Brazil, there are 8% fewer males than females. The number of students in Mexico is 1.8 times higher than in Brazil, which is in 5th place, the number of males is 2.8 times less and the number of females is 2.7.

```
[42]: fig, ax = plt.subplots(figsize=(21,35))
df.sort_values(by=['ST04Q01', 'CNT'])
sns.countplot(data=df, y='CNT', hue='ST04Q01', ax=ax, order=df['CNT'].
    →value_counts().index)
ax.legend(title='Gender', loc='best')
plt.xlabel('Students')
plt.ylabel('Country')
plt.title('Student\'s by Gender by Countries')
plt.show()
```

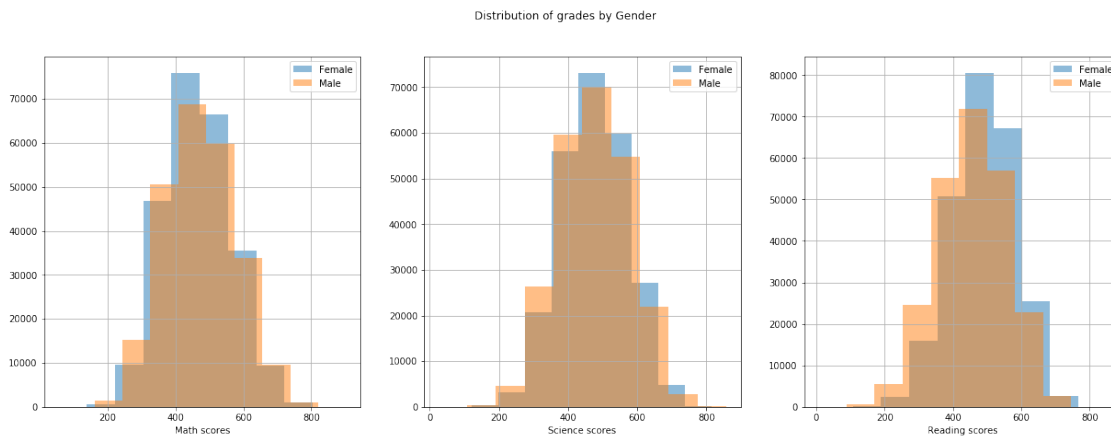



```
[43]: fig, ax = plt.subplots(nrows=1, ncols=3, figsize=(21,7))
df[df.ST04Q01 == 'Female']['PV_MATH'].hist(ax=ax[0], alpha=.5, label='Female')
df[df.ST04Q01 == 'Male']['PV_MATH'].hist(ax=ax[0], alpha=.5, label='Male')
ax[0].set_xlabel('Math scores')
ax[0].legend()

df[df.ST04Q01 == 'Female']['PV_SCIE'].hist(ax=ax[1], alpha=.5, label='Female')
df[df.ST04Q01 == 'Male']['PV_SCIE'].hist(ax=ax[1], alpha=.5, label='Male')
ax[1].set_xlabel('Science scores')
ax[1].legend()

df[df.ST04Q01 == 'Female']['PV_READ'].hist(ax=ax[2], alpha=.5, label='Female')
df[df.ST04Q01 == 'Male']['PV_READ'].hist(ax=ax[2], alpha=.5, label='Male')
ax[2].set_xlabel('Reading scores')
ax[2].legend()

plt.suptitle('Distribution of grades by Gender')
plt.show()
```



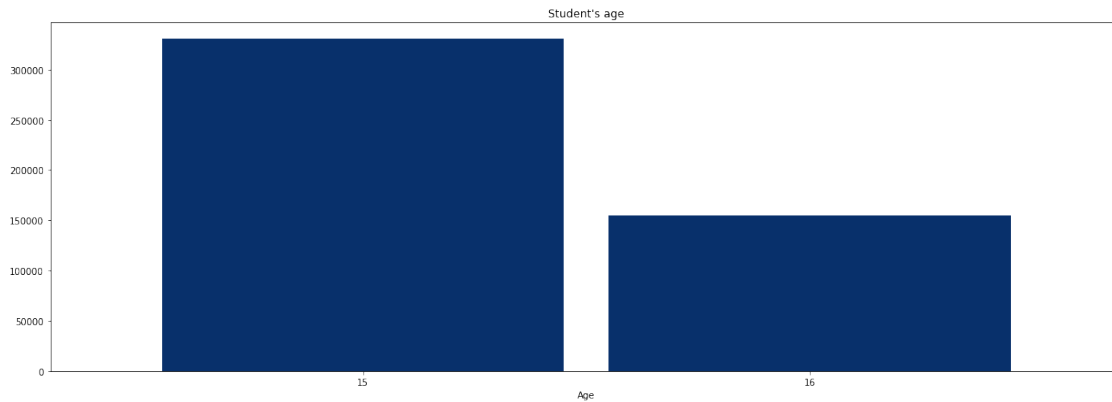
Distribution of males and females math and science score are distributed approximately normal. However, there's slight difference of reading scores: female have slightly higher grades than males.

Since student age is between 15 and 16 year old, and number of students who are 15 years old are twice larger than students who are 16 years old, there would be interesting to compare whether there's some biases due to the different age.

```
[44]: df.AGE.astype(int).value_counts().sort_index().plot.bar(figsize=(21,7), width=.
      →9, legend=False,

      →cmap='Blues_r', rot=0)
plt.xlabel('Age')
plt.title('Student\'s age')
```

```
plt.show()
```



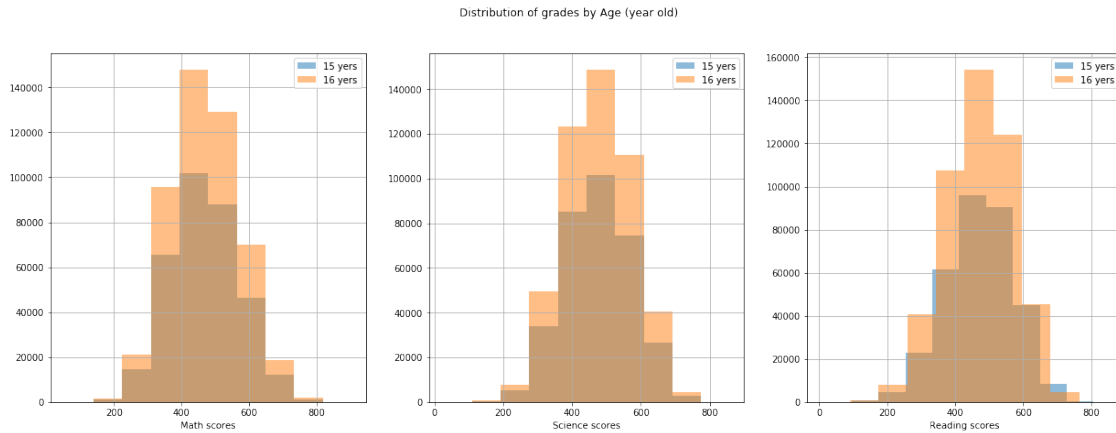
Distribution of scores of students from 15 and 16 years old groups is distributed normally, and I think, there's no significant difference between these students.

```
[45]: fig, ax = plt.subplots(nrows=1, ncols=3, figsize=(21,7))
df[df.AGE < 16]['PV_MATH'].hist(ax=ax[0], alpha=.5, label='15 yers')
df[df.AGE > 15]['PV_MATH'].hist(ax=ax[0], alpha=.5, label='16 yers')
ax[0].set_xlabel('Math scores')
ax[0].legend()

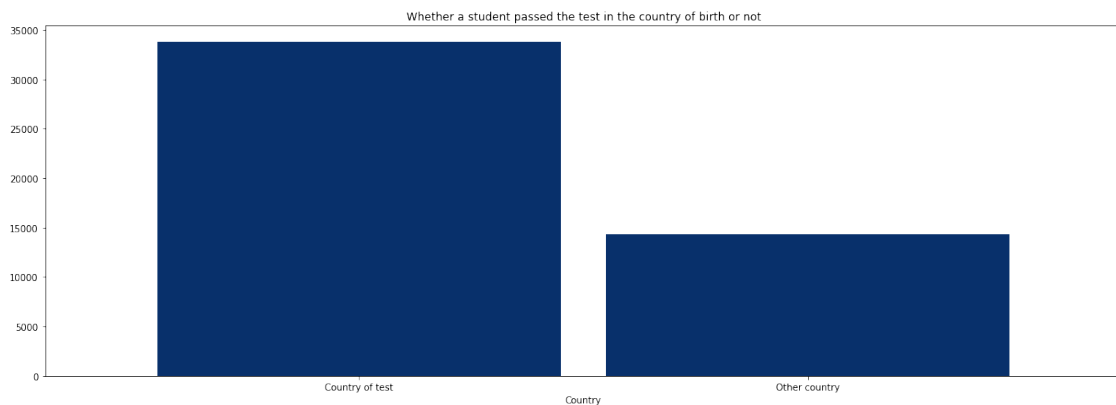
df[df.AGE < 16]['PV_SCIE'].hist(ax=ax[1], alpha=.5, label='15 yers')
df[df.AGE > 15]['PV_SCIE'].hist(ax=ax[1], alpha=.5, label='16 yers')
ax[1].set_xlabel('Science scores')
ax[1].legend()

df[df.AGE < 16]['PV_READ'].hist(ax=ax[2], alpha=.5, label='15 yers')
df[df.AGE > 15]['PV_READ'].hist(ax=ax[2], alpha=.5, label='16 yers')
ax[2].set_xlabel('Reading scores')
ax[2].legend()

plt.suptitle('Distribution of grades by Age (year old)')
plt.show()
```



```
[46]: df.groupby('ST20Q01').STIDSTD.nunique().sort_index().plot.bar(figsize=(21,7),
    width=.9, legend=False,
    cmap='Blues_r', rot=0)
plt.xlabel('Country')
plt.title('Whether a student passed the test in the country of birth or not')
plt.show()
```



```
[47]: # Whether a student passed the test in the country of birth or not

fig, ax = plt.subplots(nrows=1, ncols=3, figsize=(21,7))
df[df.ST20Q01 == 'Country of test']['PV_MATH'].hist(ax=ax[0], alpha=.5,
    label='Country of test')
df[df.ST20Q01 == 'Other country']['PV_MATH'].hist(ax=ax[0], alpha=.5,
    label='Other country')
ax[0].set_xlabel('Math scores')
ax[0].legend()
```

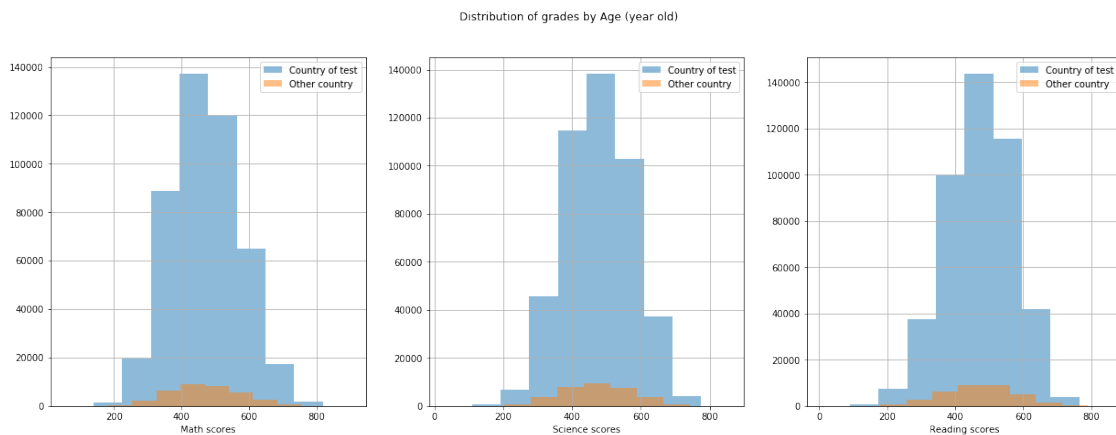
```

df[df.ST20Q01 == 'Country of test']['PV_SCIE'].hist(ax=ax[1], alpha=.5,
→label='Country of test')
df[df.ST20Q01 == 'Other country']['PV_SCIE'].hist(ax=ax[1], alpha=.5,
→label='Other country')
ax[1].set_xlabel('Science scores')
ax[1].legend()

df[df.ST20Q01 == 'Country of test']['PV_READ'].hist(ax=ax[2], alpha=.5,
→label='Country of test')
df[df.ST20Q01 == 'Other country']['PV_READ'].hist(ax=ax[2], alpha=.5,
→label='Other country')
ax[2].set_xlabel('Reading scores')
ax[2].legend()

plt.suptitle('Distribution of grades by Age (year old)')
plt.show()

```



An average international grade of students is 9.8 points, and on the same time, the mean grade compared to modal grade in country is equal -0.16 points.

Among all 64 countries represented in the dataset, students from Canada, Italy, Mexico, and Spain have the highest average international rate.

[48]: *# International grade and grade compared to modal grade in country*

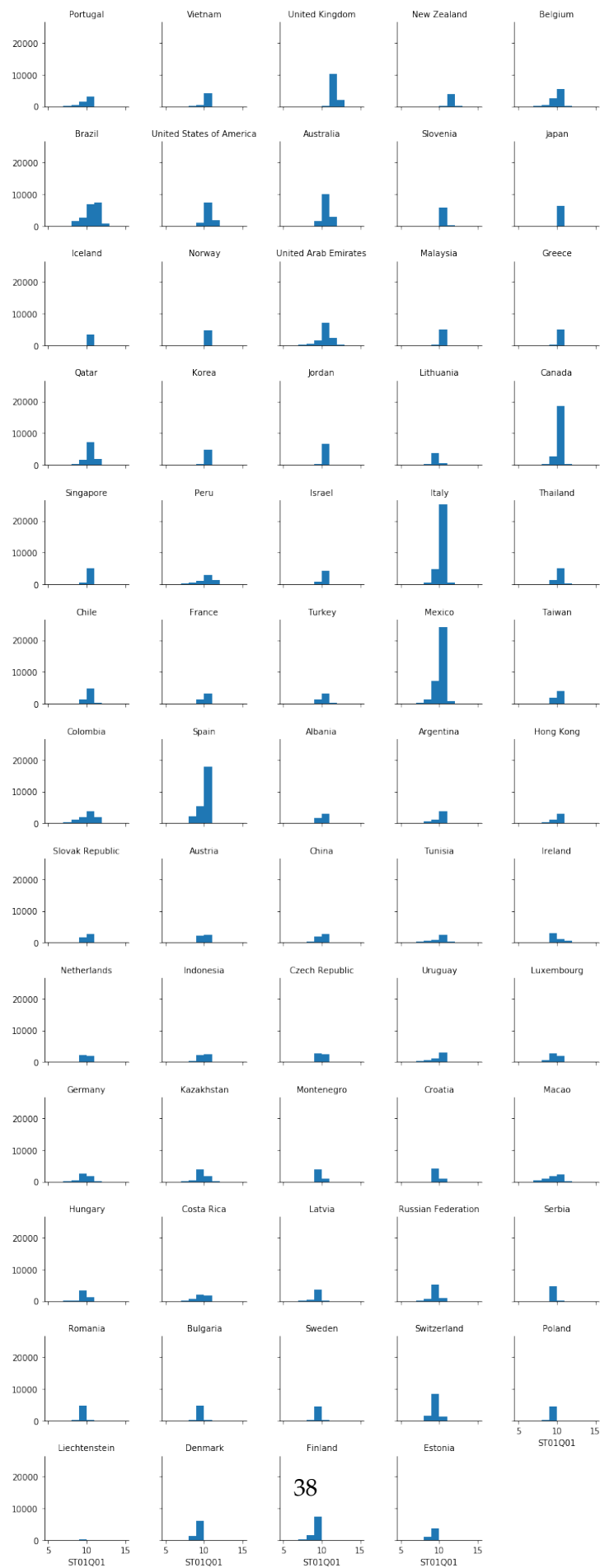
```
df[['ST01Q01', 'GRADE']].mean()
```

```
[48]: ST01Q01    9.813323
      GRADE    -0.162671
      dtype: float64
```

```
[49]: group_means = df.groupby(['CNT']).mean()
      group_order = group_means.sort_values(['ST01Q01'], ascending = False).index

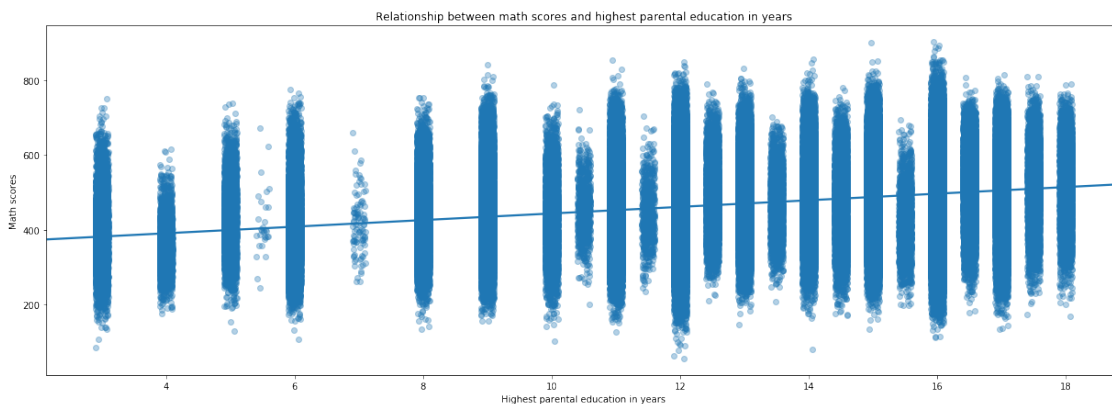
      g = sns.FacetGrid(data = df, col = 'CNT', col_wrap = 5, height = 2,
```

```
        col_order = group_order)
g.map(plt.hist, 'ST01Q01', bins = np.arange(5, 15+1, 1))
g.set_titles('{col_name}')
plt.show()
```



3. What's a relationship between students performance and highest parental education measured in years as well as mother's and father's highest schooling?

```
[50]: plt.subplots(figsize=(21,7))
sns.regplot(df['PARED'], df['PV_MATH'], fit_reg=True, x_jitter=0.1, y_jitter=0.
→1, scatter_kws={'alpha': 1/3})
plt.xlabel('Highest parental education in years')
plt.ylabel('Math scores')
plt.title('Relationship between math scores and highest parental education in_
→years')
plt.show()
```



There exist a positive weak relationship between highest parental education in years and students math scores. To check whether this relationship is significant, linear regression can be fitted to determine if increase in parental education affects increases students math scores.

4. Whether there exist a correlation between family wealth (measured in the number of telephones, computers, etc.) and students performance?

```
[51]: # How many - computers (here we'll look at 2 extreme - no computer vs. 3 or_
→more computers)

fig, ax = plt.subplots(nrows=1, ncols=3, figsize=(21,7))
df[df.ST27Q03 == 'Zero']['PV_MATH'].hist(ax=ax[0], alpha=.5, label='Zero')
df[df.ST27Q03 == 'Three or more']['PV_MATH'].hist(ax=ax[0], alpha=.5,
→label='Three or more')
ax[0].set_xlabel('Math scores')
ax[0].legend()

df[df.ST27Q03 == 'Zero']['PV_SCIE'].hist(ax=ax[1], alpha=.5, label='Zero')
```

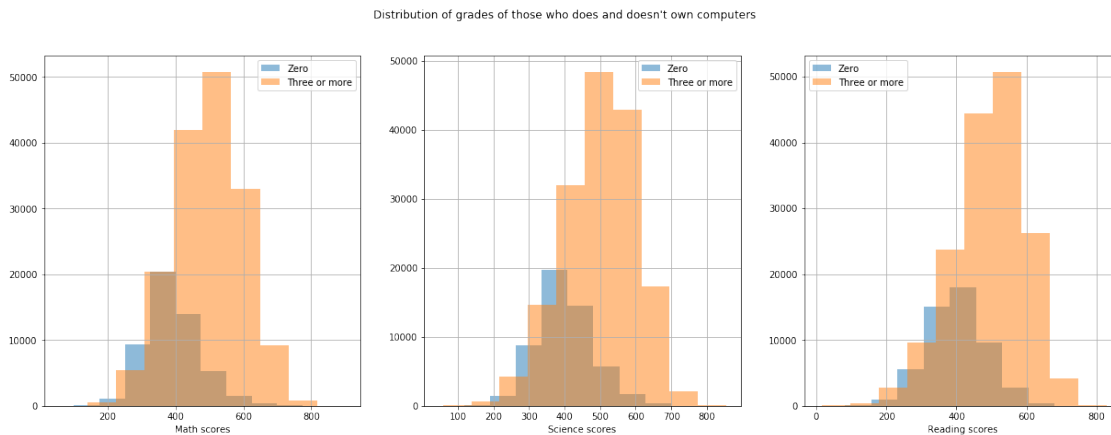
```

df[df.ST27Q03 == 'Three or more']['PV_SCIE'].hist(ax=ax[1], alpha=.5,
    →label='Three or more')
ax[1].set_xlabel('Science scores')
ax[1].legend()

df[df.ST27Q03 == 'Zero']['PV_READ'].hist(ax=ax[2], alpha=.5, label='Zero')
df[df.ST27Q03 == 'Three or more']['PV_READ'].hist(ax=ax[2], alpha=.5,
    →label='Three or more')
ax[2].set_xlabel('Reading scores')
ax[2].legend()

plt.suptitle('Distribution of grades of those who does and doesn\'t own
    →computers')
plt.show()

```



More than half of all students don't have a computer at all. Therefore, we can observe, that distribution of score of those students who doesn't have a computer is skewed to the right for two subjects - mathematics and science.

[52]: *# How many - cars (here we'll look at 2 extreme - no car vs. 3 or more cars)*

```

fig, ax = plt.subplots(nrows=1, ncols=3, figsize=(21,7))
df[df.ST27Q04 == 'Zero']['PV_MATH'].hist(ax=ax[0], alpha=.5, label='Zero')
df[df.ST27Q04 == 'Three or more']['PV_MATH'].hist(ax=ax[0], alpha=.5,
    →label='Three or more')
ax[0].set_xlabel('Math scores')
ax[0].legend()

df[df.ST27Q04 == 'Zero']['PV_SCIE'].hist(ax=ax[1], alpha=.5, label='Zero')
df[df.ST27Q04 == 'Three or more']['PV_SCIE'].hist(ax=ax[1], alpha=.5,
    →label='Three or more')
ax[1].set_xlabel('Science scores')
ax[1].legend()

```

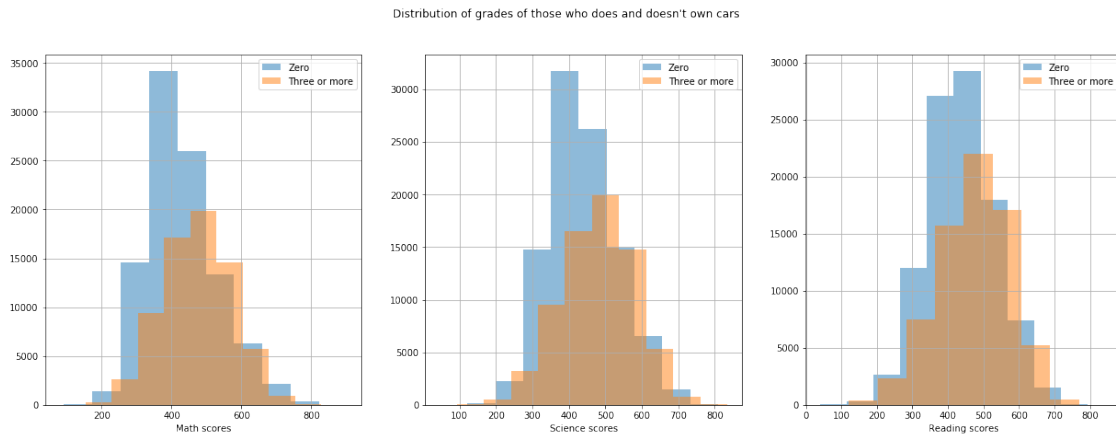


```

df[df.ST27Q04 == 'Zero']['PV_READ'].hist(ax=ax[2], alpha=.5, label='Zero')
df[df.ST27Q04 == 'Three or more']['PV_READ'].hist(ax=ax[2], alpha=.5,
→label='Three or more')
ax[2].set_xlabel('Reading scores')
ax[2].legend()

plt.suptitle('Distribution of grades of those who does and doesn\'t own cars')
plt.show()

```



A half of all students don't have a car in family. And we can observe, that distribution of score of those students who has no car in the family is skewed to the right for all 3 subjects - math, science, and reading.

[53]: *# How many - cellular phones (here we'll look at 2 extreme - no cellular phone
→vs. 3 or more cellular phones)*

```

fig, ax = plt.subplots(nrows=1, ncols=3, figsize=(21,7))
df[df.ST27Q01 == 'Zero']['PV_MATH'].hist(ax=ax[0], alpha=.5, label='Zero')
df[df.ST27Q01 == 'Three or more']['PV_MATH'].hist(ax=ax[0], alpha=.5,
→label='Three or more')
ax[0].set_xlabel('Math scores')
ax[0].legend()

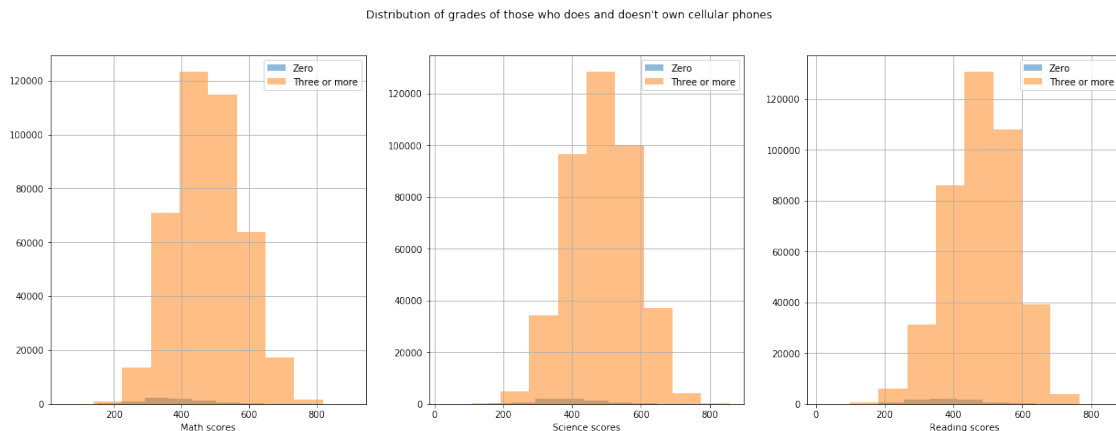
df[df.ST27Q01 == 'Zero']['PV_SCIE'].hist(ax=ax[1], alpha=.5, label='Zero')
df[df.ST27Q01 == 'Three or more']['PV_SCIE'].hist(ax=ax[1], alpha=.5,
→label='Three or more')
ax[1].set_xlabel('Science scores')
ax[1].legend()

df[df.ST27Q01 == 'Zero']['PV_READ'].hist(ax=ax[2], alpha=.5, label='Zero')
df[df.ST27Q01 == 'Three or more']['PV_READ'].hist(ax=ax[2], alpha=.5,
→label='Three or more')

```

```
ax[2].set_xlabel('Reading scores')
ax[2].legend()

plt.suptitle('Distribution of grades of those who does and doesn\'t own
→cellular phones')
plt.show()
```



```
[54]: df[df.ST27Q01 == 'Zero'].shape[0] / df.shape[0]
```

```
[54]: 0.014589383921399
```

Almost every student in the dataset has at least one cellular phone. And it is almost impossible to determine what the distribution of grades looks like for those students who do not have a cell phone, since the number of such guys in the dataset is very small (about 1.5%).

5. How do student possessions such as own room and desk, etc. affect his/her performance?

```
[55]: # Possessions - own room

fig, ax = plt.subplots(nrows=1, ncols=3, figsize=(21,7))
df[df.ST26Q02 == 'Yes']['PV_MATH'].hist(ax=ax[0], alpha=.5, label='Own room',
→color='green')
df[df.ST26Q02 == 'No']['PV_MATH'].hist(ax=ax[0], alpha=.5, label='No room',
→color='black')
ax[0].set_xlabel('Math scores')
ax[0].legend()

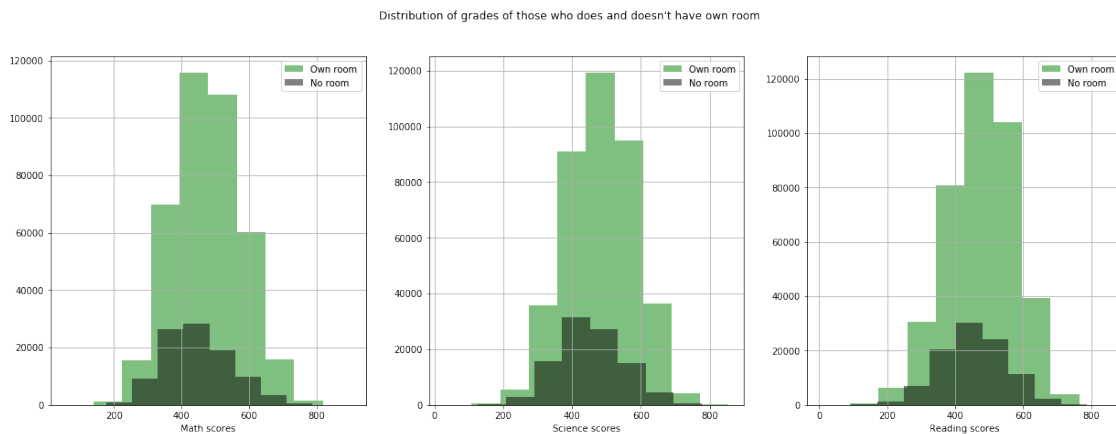
df[df.ST26Q02 == 'Yes']['PV_SCIE'].hist(ax=ax[1], alpha=.5, label='Own room',
→color='green')
df[df.ST26Q02 == 'No']['PV_SCIE'].hist(ax=ax[1], alpha=.5, label='No room',
→color='black')
ax[1].set_xlabel('Science scores')
ax[1].legend()
```

```

df[df.ST26Q02 == 'Yes']['PV_READ'].hist(ax=ax[2], alpha=.5, label='Own room',
    →color='green')
df[df.ST26Q02 == 'No']['PV_READ'].hist(ax=ax[2], alpha=.5, label='No room',
    →color='black')
ax[2].set_xlabel('Reading scores')
ax[2].legend()

plt.suptitle('Distribution of grades of those who does and doesn\'t have own_
    →room')
plt.show()

```



About a quarter of students doesn't have their own rooms. This affects their preparation to the exam. And as the result, the distribution of math and science scores of those students who don't have their own room is skewed to the right

[56]: *# Possessions - has desk*

```

fig, ax = plt.subplots(nrows=1, ncols=3, figsize=(21,7))
df[df.ST26Q01 == 'Yes']['PV_MATH'].hist(ax=ax[0], alpha=.5, label='Has desk',
    →color='green')
df[df.ST26Q01 == 'No']['PV_MATH'].hist(ax=ax[0], alpha=.5, label='No desk',
    →color='black')
ax[0].set_xlabel('Math scores')
ax[0].legend()

df[df.ST26Q01 == 'Yes']['PV_SCIE'].hist(ax=ax[1], alpha=.5, label='Has room',
    →color='green')
df[df.ST26Q01 == 'No']['PV_SCIE'].hist(ax=ax[1], alpha=.5, label='No desk',
    →color='black')
ax[1].set_xlabel('Science scores')
ax[1].legend()

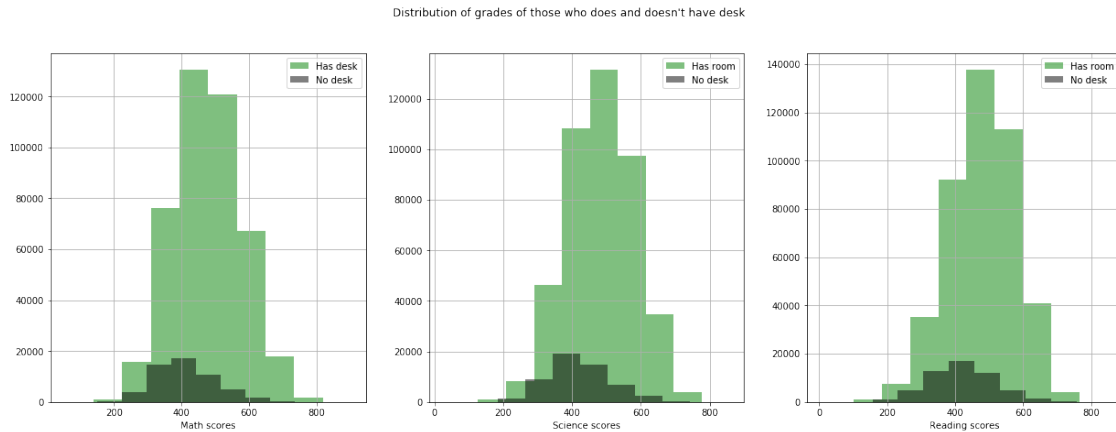
```

```

df[df.ST26Q01 == 'Yes']['PV_READ'].hist(ax=ax[2], alpha=.5, label='Has room',
    color='green')
df[df.ST26Q01 == 'No']['PV_READ'].hist(ax=ax[2], alpha=.5, label='No desk',
    color='black')
ax[2].set_xlabel('Reading scores')
ax[2].legend()

plt.suptitle('Distribution of grades of those who does and doesn\'t have desk')
plt.show()

```



```
[57]: df[df.ST26Q01 == 'No'].shape[0] / df.shape[0]
```

```
[57]: 0.11132258130960473
```

11.1% of students don't have a desk, therefore, on average their math and science scores are lower than scores of students who has a table. Both, the distribution of reading scores of those who have and who doesn't have a desk is normally distributed without any skewedness.

```
[58]: # Possessions - has study place
```

```

fig, ax = plt.subplots(nrows=1, ncols=3, figsize=(21,7))
df[df.ST26Q03 == 'Yes']['PV_MATH'].hist(ax=ax[0], alpha=.5, label='Has study_
    place', color='green')
df[df.ST26Q03 == 'No']['PV_MATH'].hist(ax=ax[0], alpha=.5, label='No study_
    place', color='black')
ax[0].set_xlabel('Math scores')
ax[0].legend()

df[df.ST26Q03 == 'Yes']['PV_SCIE'].hist(ax=ax[1], alpha=.5, label='Has study_
    place', color='green')
df[df.ST26Q03 == 'No']['PV_SCIE'].hist(ax=ax[1], alpha=.5, label='No study_
    place', color='black')
ax[1].set_xlabel('Science scores')
ax[1].legend()

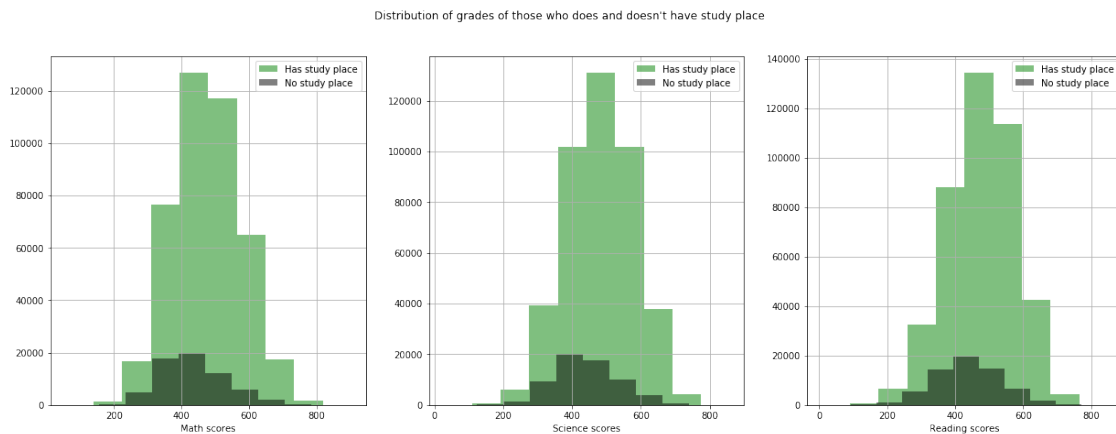
```

```

df[df.ST26Q03 == 'Yes']['PV_READ'].hist(ax=ax[2], alpha=.5, label='Has study_
    place', color='green')
df[df.ST26Q03 == 'No']['PV_READ'].hist(ax=ax[2], alpha=.5, label='No study_
    place', color='black')
ax[2].set_xlabel('Reading scores')
ax[2].legend()

plt.suptitle('Distribution of grades of those who does and doesn\'t have study_
    place')
plt.show()

```



It's really difficult to prepare to the assessment if you don't have study place at home. As a result the distribution of scores of those students who don't have a study place on average receive lower scores on math and science.

[59]: *# Possessions - has computer*

```

fig, ax = plt.subplots(nrows=1, ncols=3, figsize=(21,7))
df[df.ST26Q04 == 'Yes']['PV_MATH'].hist(ax=ax[0], alpha=.5, label='Has_
    computer', color='green')
df[df.ST26Q04 == 'No']['PV_MATH'].hist(ax=ax[0], alpha=.5, label='No computer',_
    color='black')
ax[0].set_xlabel('Math scores')
ax[0].legend()

df[df.ST26Q04 == 'Yes']['PV_SCIE'].hist(ax=ax[1], alpha=.5, label='Has_
    computer', color='green')
df[df.ST26Q04 == 'No']['PV_SCIE'].hist(ax=ax[1], alpha=.5, label='No computer',_
    color='black')
ax[1].set_xlabel('Science scores')
ax[1].legend()

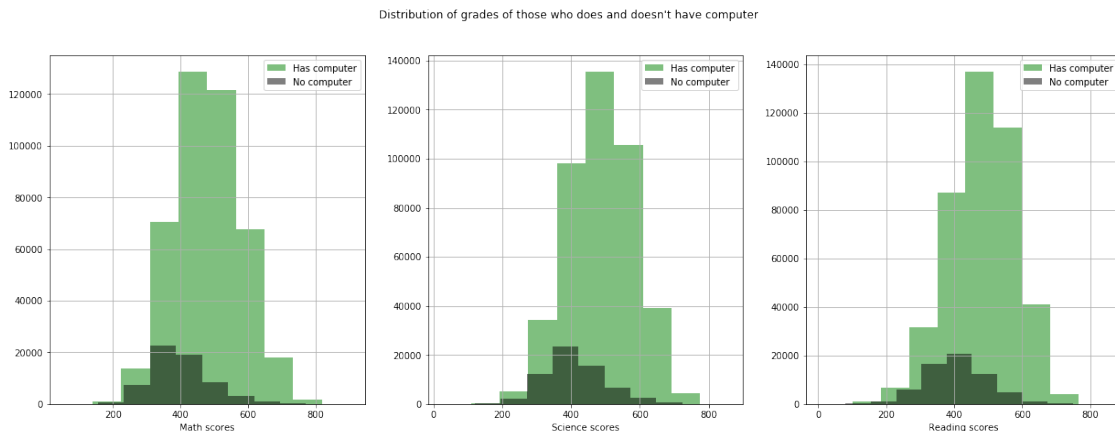
```

```

df[df.ST26Q04 == 'Yes']['PV_READ'].hist(ax=ax[2], alpha=.5, label='Has computer', color='green')
df[df.ST26Q04 == 'No']['PV_READ'].hist(ax=ax[2], alpha=.5, label='No computer', color='black')
ax[2].set_xlabel('Reading scores')
ax[2].legend()

plt.suptitle('Distribution of grades of those who does and doesn\'t have computer')
plt.show()

```



Absence of computer significantly complicates the preparation not only for the exam, but also for the homework. Because for example, not all students have large-screen tablets or smartphones that can partially replace a computer. As a result, the distribution of math scores is significantly skewed to the right. Distributions of reading and science scores are also slightly skewed to the right.

```

[60]: # Possessions - has software

fig, ax = plt.subplots(nrows=1, ncols=3, figsize=(21,7))
df[df.ST26Q05 == 'Yes']['PV_MATH'].hist(ax=ax[0], alpha=.5, label='Has software', color='green')
df[df.ST26Q05 == 'No']['PV_MATH'].hist(ax=ax[0], alpha=.5, label='No software', color='black')
ax[0].set_xlabel('Math scores')
ax[0].legend()

df[df.ST26Q05 == 'Yes']['PV_SCIE'].hist(ax=ax[1], alpha=.5, label='Has software', color='green')
df[df.ST26Q05 == 'No']['PV_SCIE'].hist(ax=ax[1], alpha=.5, label='No software', color='black')
ax[1].set_xlabel('Science scores')
ax[1].legend()

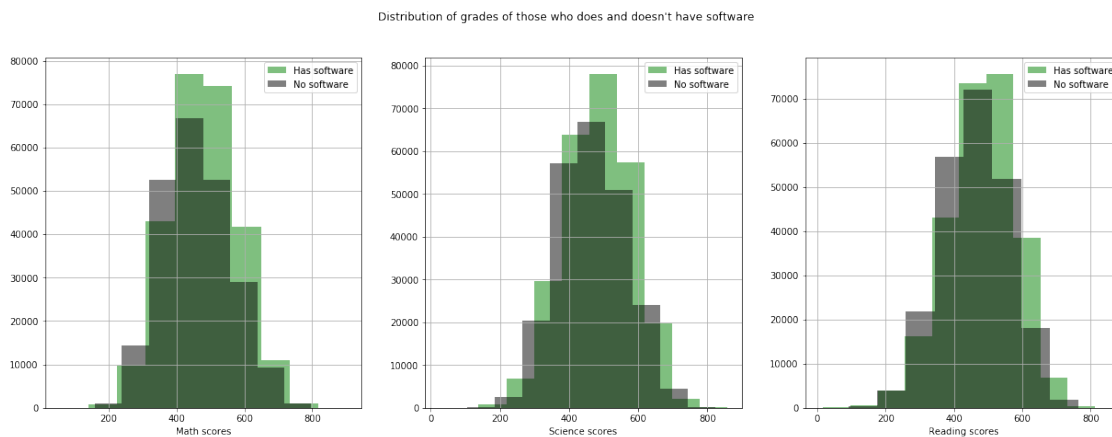
```

```

df[df.ST26Q05 == 'Yes']['PV_READ'].hist(ax=ax[2], alpha=.5, label='Has
    ↳software', color='green')
df[df.ST26Q05 == 'No']['PV_READ'].hist(ax=ax[2], alpha=.5, label='No software',
    ↳color='black')
ax[2].set_xlabel('Reading scores')
ax[2].legend()

plt.suptitle('Distribution of grades of those who does and doesn\'t have
    ↳software')
plt.show()

```



I can assume that the lack of software does not affect the distribution of grades in any way, since not all students pay money for software, thus, the lack of a computer worsens the average grade for the test more significantly.

[61]: *# Possessions - has Internet*

```

fig, ax = plt.subplots(nrows=1, ncols=3, figsize=(21,7))
df[df.ST26Q06 == 'Yes']['PV_MATH'].hist(ax=ax[0], alpha=.5, label='Has
    ↳Internet', color='green')
df[df.ST26Q06 == 'No']['PV_MATH'].hist(ax=ax[0], alpha=.5, label='No Internet',
    ↳color='black')
ax[0].set_xlabel('Math scores')
ax[0].legend()

df[df.ST26Q06 == 'Yes']['PV_SCIE'].hist(ax=ax[1], alpha=.5, label='Has
    ↳Internet', color='green')
df[df.ST26Q06 == 'No']['PV_SCIE'].hist(ax=ax[1], alpha=.5, label='No Internet',
    ↳color='black')
ax[1].set_xlabel('Science scores')
ax[1].legend()

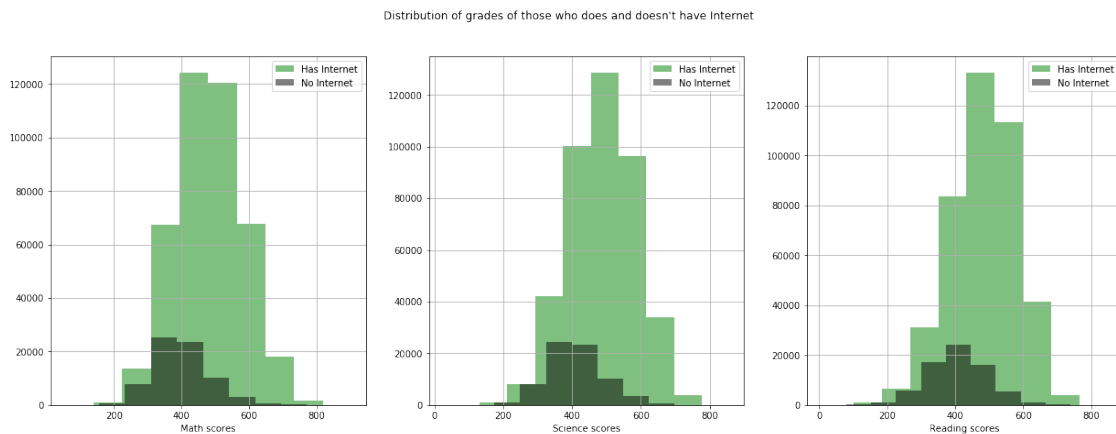
```

```

df[df.ST26Q06 == 'Yes']['PV_READ'].hist(ax=ax[2], alpha=.5, label='Has Internet', color='green')
df[df.ST26Q06 == 'No']['PV_READ'].hist(ax=ax[2], alpha=.5, label='No Internet', color='black')
ax[2].set_xlabel('Reading scores')
ax[2].legend()

plt.suptitle('Distribution of grades of those who does and doesn\'t have Internet')
plt.show()

```



The lack of the Internet affects the distribution of grades in mathematics and science, since the process of obtaining the necessary information is either very slow or not at all. But it is worth noting that the presence of the Internet affects the distribution of reading grades in such a way that for those students who have the Internet, the opportunity to obtain additional information leads to a distortion of the distribution of grades to the left.

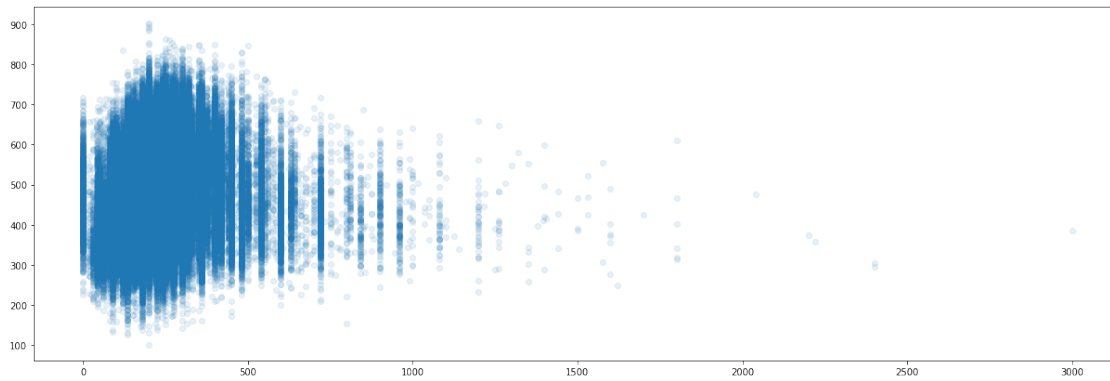
Summing up, I would like to say that for a student who prepares and takes the exam, any help, whether it be a computer, the Internet, a place for preparation, or his own team, positively correlates with higher grades in both mathematics and science and reading. This is a very interesting study that can be done by collecting additional missing data and adding information from other sources, for example, information on income and / or expenses of students' families.

6. Whether total time learning and out of school lessons on math, science, and reading affect student performance?

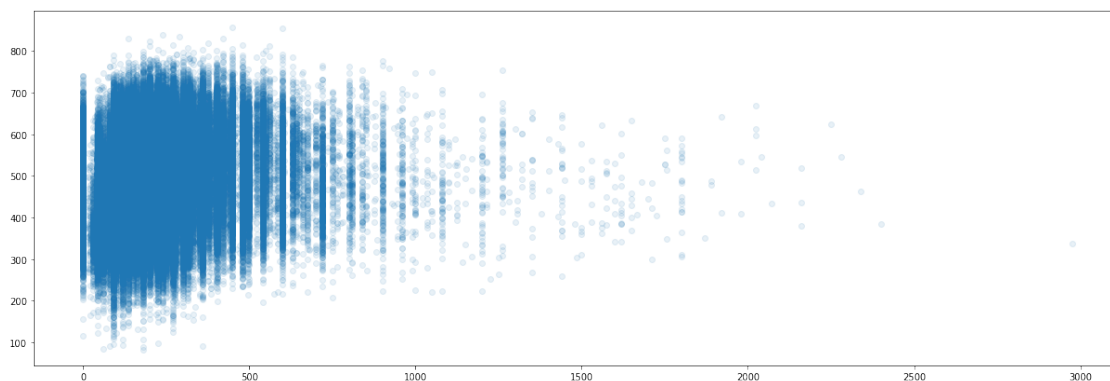
```

[62]: plt.subplots(figsize=(21,7))
plt.scatter(df['MINS'], df['PV_MATH'], alpha=.1, cmap='Blues_r')
plt.xlabel('')
plt.ylabel('')
plt.title('')
plt.show()

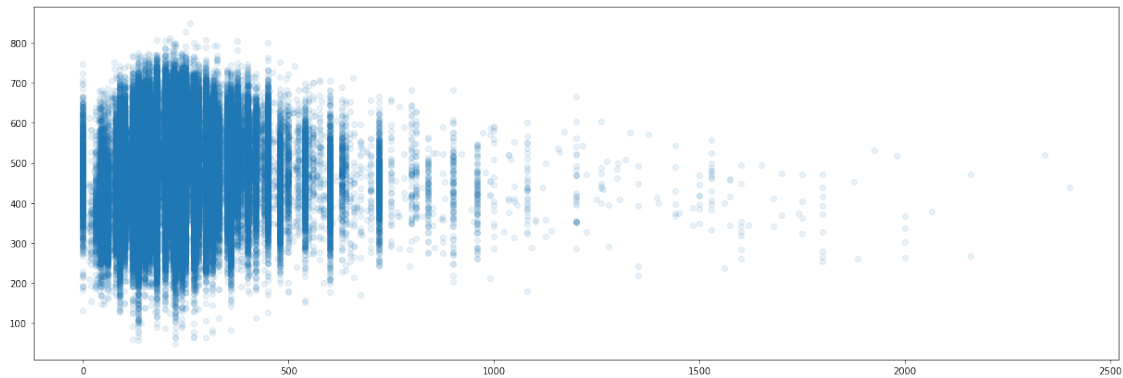
```

```
[63]: plt.subplots(figsize=(21,7))
plt.scatter(df['SMINS'], df['PV_SCIE'], alpha=.1, cmap='Blues_r')
plt.xlabel('')
plt.ylabel('')
plt.title('')
plt.show()
```



```
[64]: plt.subplots(figsize=(21,7))
plt.scatter(df['LMINS'], df['PV_READ'], alpha=.1, cmap='Blues_r')
plt.xlabel('')
plt.ylabel('')
plt.title('')
plt.show()
```



```
[65]: print(df[['MINS', 'PV_MATH']].corr().iloc[0,1])
print(df[['SINS', 'PV_SCIE']].corr().iloc[0,1])
print(df[['LINS', 'PV_READ']].corr().iloc[0,1])
```

```
0.07118616777216413
0.1497322691619717
0.030070040120862448
```

Probably, the answer is obvious to the question whether there is a positive correlation between the number of hours of preparation for a particular subject and the grade for the exam after such preparation. But according to the schedules of preparing students for both mathematics and science and literature, one cannot say that there is a moderated relationship between this action and the result.

In order to understand whether there really is no relationship between the preparation time for the exam and the grade for it. I calculated the Pearson correlation coefficient. And indeed the highest correlation coefficient is 15%.

Part III - Conclusions

1. What is students' performance at schools in different countries (including whether country is a OECD member) If we plot all the grades by subject, then scores in each subject looks normally distributed. Mean scores of science are about 3 points higher than average reading scores. In its turn, average reading scores are about 3 points higher than average math scores. So scores in those 3 subjects are very similar.

So, let's look at their boxplots.

In general, there're outliers in every Series of scores. Moreover, math scores have approximately equal tails of outliers, but science and reading scores have outliers with lower scores more, than outliers with higher scores. let's go deeper, and look at students performance in the context of countries, OECD membership, and other columns.

Except Liechtenstein which is on the 6th position, on average, students from Asia countries receive the highest scores on math. China, Singapore, Hong Kong, Taiwan, and Korea are in Top-5. Macao and Japan follow immediately behind Liechtenstein.

Chinese students receive on average 611 points. In comparison, in Peru average math scores are equal 368. This's 1.7 times less than in China.

Average science scores are less than math scores by about 6 points. And this is becoming

noticeable for countries with the highest average scores in science. China, Hong Kong and Singapore are also in the Top-3 with an average score of 547 to 579. For China, this difference is 32 points or 5.2%.

For average reading scores, China, Hong Kong, Singapore, Japan, Korea and Taiwan continue to be the leaders with a maximum average of 569 points for China. This average score is the lowest for China in three subjects, possibly also because English is not a native language for a large population of the country.

The difference is noticeable in all three subjects at once: average scores in mathematics, science and reading are higher in OSCE countries than in non-OSCE countries. The difference is about 48 points for each subject.

2. What are the characteristics of students participated in PICA 2012:

- gender:

Number of female students is little more (by 0.6%) than number of male students. Let's look at gender by countries. The largest number of students are in Mexico, Italy, Spain, Canada and Brazil. Except for Italy, the number of the females is greater than that of the male. In Brazil, there are 8% fewer males than females. The number of students in Mexico is 1.8 times higher than in Brazil, which is in 5th place, the number of males is 2.8 times less and the number of females is 2.7.

Distribution of males and females math and science score are distributed approximately normal. However, there's slight difference of reading scores: female have slightly higher grades than males.

- age:

Since student age is between 15 and 16 year old, and number of students who are 15 years old are twice larger than students who are 16 years old, there would be interesting to compare whether there's some biases due to the different age.

Distribution of scores of students from 15 and 16 years old groups is distributed normally, and I think, there's no significant difference between these students.

- international grade and grade compared to modal grade in country:

An average international grade of students is 9.8 points, and on the same time, the mean grade compared to modal grade in country is equal -0.16 points.

Among all 64 countries represented in the dataset, students from Canada, Italy, Mexico, and Spain have the highest average international rate.

3. What's a relationship between students performance and highest parental education measured in years as well as mother's and father's highest schooling?

There exist a positive weak relationship between highest parental education in years and students math scores. To check whether this relationship is significant, linear regression can be fitted to determine if increase in parental education affects increases students math scores.

4. Whether there exist a correlation between family wealth (measured in the number of telephones, computers, etc.) and students performance?

More than half of all students don't have a computer at all. Therefore, we can observe, that distribution of score of those students who doesn't have a computer is skewed to the right for two subjects - mathematics and science.

A half of all students don't have a car in family. And we can observe, that distribution of score of those students who has no car in the family is skewed to the right for all 3 subjects - math, science, and reading.

Almost every student in the dataset has at least one cellular phone. And it is almost impossible to determine what the distribution of grades looks like for those students who do not have a cell phone, since the number of such guys in the dataset is very small (about 1.5%).

5. How do student possessions such as own room and desk, etc. affect his/her performance?

About a quarter of students doesn't have their own rooms. This affects their preparation to the exam. And as the result, the distribution of math and science scores of those students who don't have their own room is skewed to the right

11.1% of students don't have a desk, therefore, on average their math and science scores are lower than scores of students who has a table. Both, the distribution of reading scores of those who have and who doesn't have a desk is normally distributed without any skewedness.

It's really difficult to prepare to the assessment if you don't have study place at home. As a result the distribution of scores of those students who don't have a study place on average receive lower scores on math and science.

Absence of computer significantly complicates the preparation not only for the exam, but also for the homework. Because for example, not all students have large-screen tablets or smartphones that can partially replace a computer. As a result, the distribution of math scores is significantly skewed to the right. Distributions of reading and science scores are also slightly skewed to the right.

I can assume that the lack of software does not affect the distribution of grades in any way, since not all students pay money for software, thus, the lack of a computer worsens the average grade for the test more significantly.

Summing up, I would like to say that for a student who prepares and takes the exam, any help, whether it be a computer, the Internet, a place for preparation, or his own team, positively correlates with higher grades in both mathematics and science and reading.

This is a very interesting study that can be done by collecting additional missing data and adding information from other sources, for example, information on income and / or expenses of students' families.

6. Whether total time learning and out of school lessons on math, science, and reading affect student performance?

Probably, the answer is obvious to the question whether there is a positive correlation between the number of hours of preparation for a particular subject and the grade for the exam after such preparation. But according to the schedules of preparing students for both mathematics and science and literature, one cannot say that there is a moderated relationship between this action and the result.

In order to understand whether there really is no relationship between the preparation time for the exam and the grade for it. I calculated the coefficient of Pearson's correlation. And indeed the highest correlation coefficient is 15%.

```
[ ]: ! jupyter nbconvert *.ipynb --to slides --post serve --template output_toggle.  
      ↪tpl
```

```
[ ]:
```