W200 Python Fundamentals

Project 2, Team 3: Jeff Braun, Siobhan Harrington, Andres Kodaka

Introduction

Quality Education is a critical success factor in the twenty-first century. Most countries are actively pursuing policy to ensure the best outcome for future generations. In this spirit, the OECD (Organization for Economic Co-operation and Development) created the PISA Project.

> *"The Programme for International Student Assessment (PISA) is a triennial international survey which aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students.*
>
> *In 2015 over half a million students, representing 28 million 15-year-olds in 72 countries and economies, took the internationally agreed two-hour test. Students were assessed in science, mathematics, reading, collaborative problem solving and financial literacy."*
>
> http://www.oecd.org/pisa/aboutpisa/

The test also includes an extensive 1 hour long survey where students can share data on factors such as who helps them with their homework, how many bedrooms are in their house, and what they do in their free time. OECD is generous with this data and enables end users to query data through online tools or the full dataset can be downloaded into a SAS or SPSS file.

Our Project

While many articles have been written about the 2015 PISA test, we set out to dig into the data and find out if it could answer some of our questions:

1. How do demographics impact a student's score? Does gender play a role across various cultures like it seems to play in the USA? Are boys really better at math and science while they lag behind girls in reading? While the test was administered to 15 year olds, does an "almost 16" year old have a competitive advantage over someone almost a year younger?
2. How does income/socio-economic status impact a student's score? With more resources, do students do better? Do poor students in rich countries fare the same as poor students in poor countries?
3. How do culture and family impact student score? Do students become more motivated with a very supportive family? Do scores naturally follow?

Our hypothesis is that certain of the demographic, economic and cultural factors surveyed will show a strong correlation to the differences in student performance.

Our Process

We first downloaded the files from the OECD/PISA website in the SAS7bdat format (http://www.oecd.org/pisa/data/2015database). Thanks to the miracles of google and developer Jared Hobbs, we were able to use python to translate this file into a .csv. (https://pypi.python.org/pypi/sas7bdat). This script took some time, but it was worth it.

We then were able to load the data successfully into dataframes within python. The columns and data itself required some reformatting, but since this data has been set up to be widely distributed - it is fairly clean and well documented with metadata. The surveys themselves were very helpful in understanding what individual variables contained.

The PISA project tests students in math, science, and reading. They break these broad subjects into 10 different categories. Each student is then assigned thirty "plausible" scores. Not every student is tested in every one for the 30 categories, so some student scores are estimated based on responses to questions that were asked and answered. PISA maintains that their scores are accurate in the aggregate and are within a 3 point margin of accuracy at the individual level. The organization does not make a final score available. Since our findings would not be influencing public policy or opinion, we calculated the final math, science, and reading scores using a simple mean of the 10 sub-scores within those broader topic areas. We found that when we averaged our simple results across all the countries, we were within the margin of error as compared with each country's published result.

We then assigned deciles for each of the subject areas, so we could understand how individual students performed in each subject area when compared with their peers around the world.

The data contains 519,334 rows of data with 922 fields representing over 70 countries. We quickly realized that our project should focus on a subset of variables and countries. We selected 20 countries randomly and validated that our countries were geographically and economically diverse. Unfortunately, there seems to be some self-selection, so developing countries are not as well represented as we had hoped. The countries we used were:

- **Europe**: Lithuania, Estonia, United Kingdom, Finland, Luxembourg, Netherlands, Denmark, Ireland, Poland, Greece
- **The Americas:** USA, Chile, Peru, Dominican Republic, Argentina (portions of the country tested)
- **Asia:** Korea, Japan, Macao
- **Middle East:** Qatar
- **Africa:** Tunisa

After selecting for just the above countries, our data set had 130,518 records.

Our Findings

Demographics: Gender

Gender is a typical area of interest in the educational field. Much research has indicated that boys and girls learn differently. The classic stereotype is that girls excel in language arts, while boys excel in math and science. Our question was, is this true in every culture?

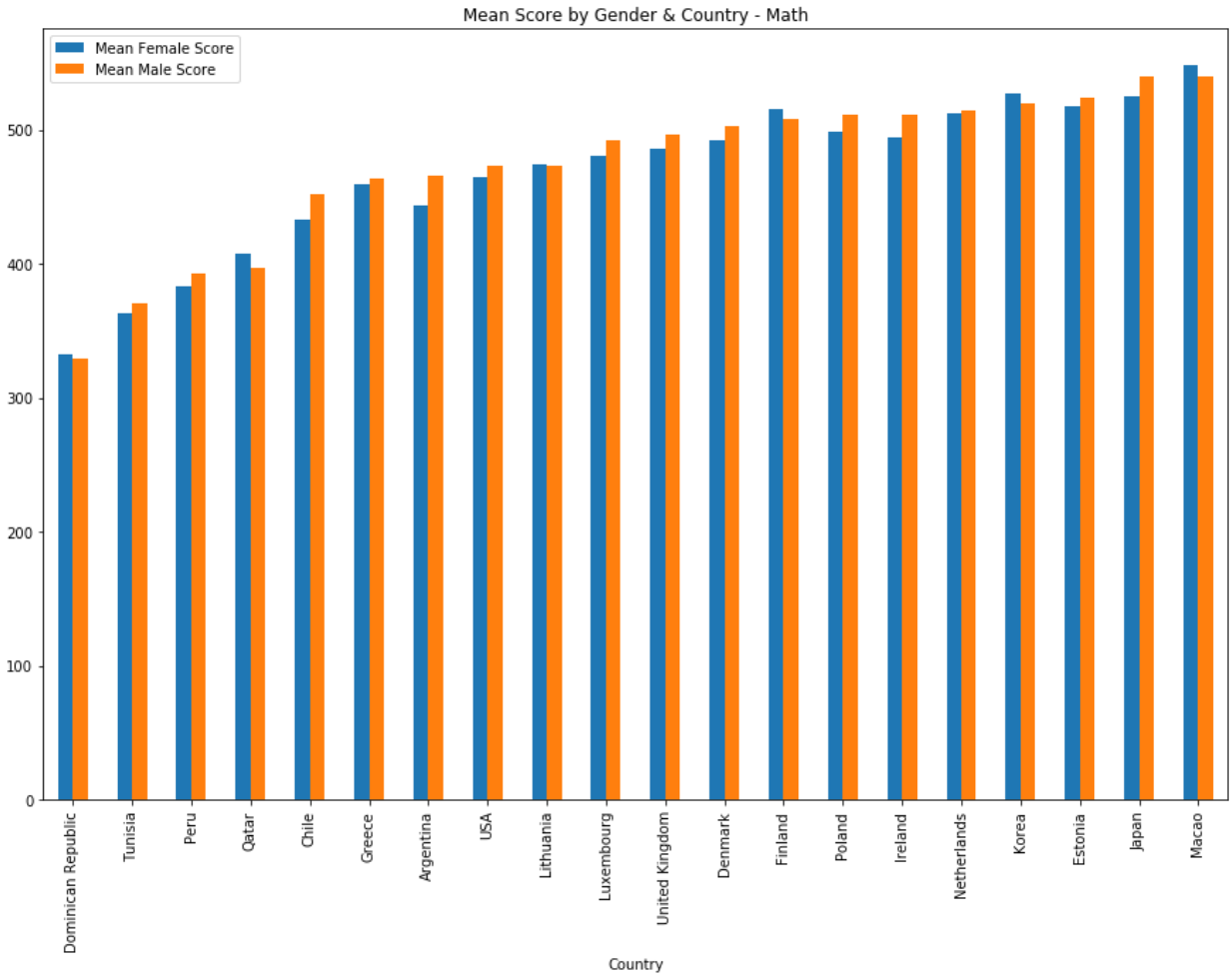
Gender is self-reported, but seemed fairly clean in the full dataset. The data is stored in ST004D01T. 1 indicates Female and 2 indicates Male. There was a leading space included in the data, but we suspected that came from the SAS -> CSV translation. Valid value counts are:

Gender Indicator	Count
1.0 (Female)	260,245
2.0 (Male)	259,089

OECD works hard to ensure that the test is taken by a representative sample of students. Our data validation showed that the samples were not always perfectly even by gender, but they came relatively close. Tunisia is the most off balance with 2696 females taking the test compared 2406 males.

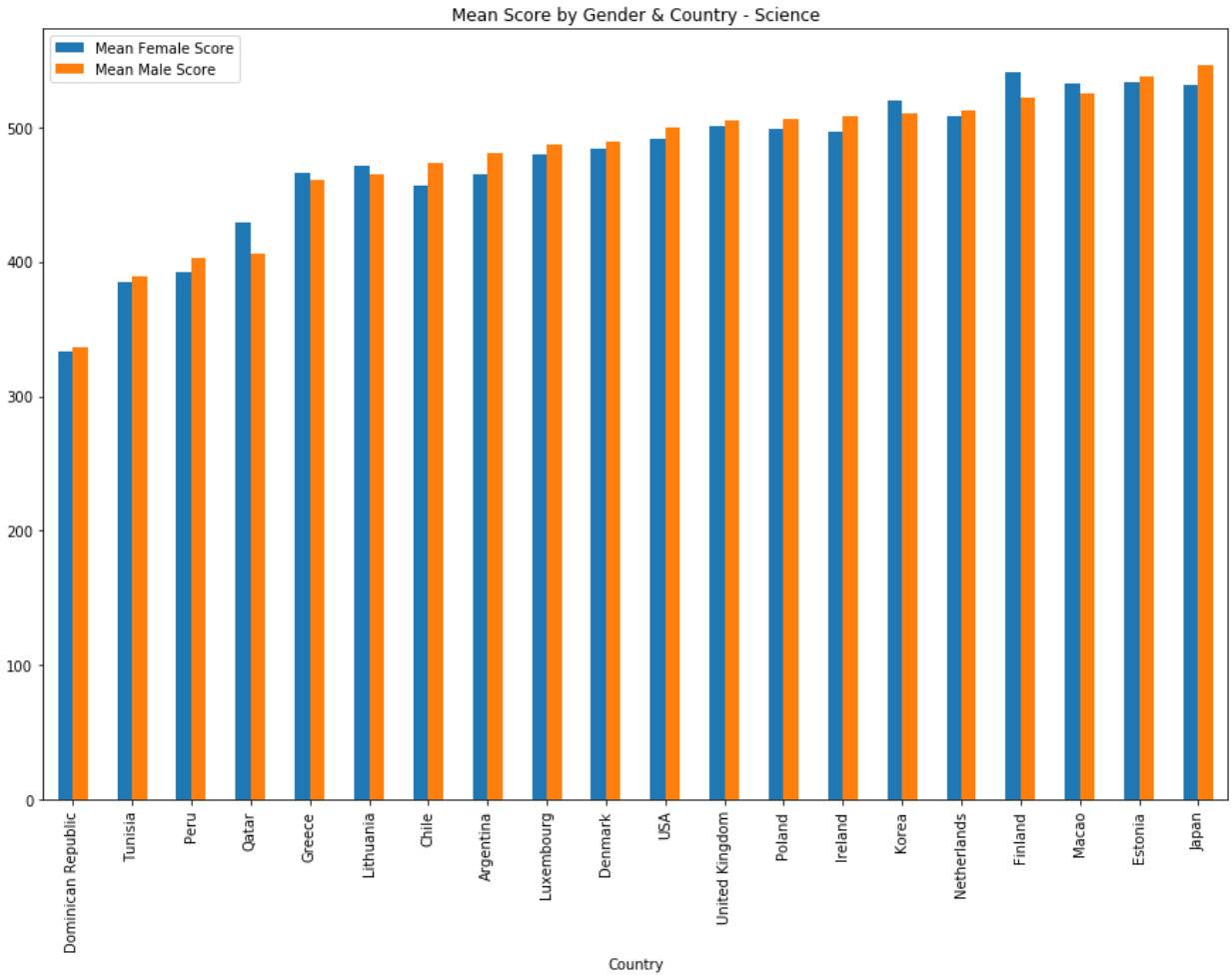
The original plan was to show counts of top twenty percent performers broken out by country and gender. However, when we discovered the male:female ratio was not consistent among countries, our team decided to show mean scores by gender and country instead. Fortunately, the data tells the same story both ways.

Math Results



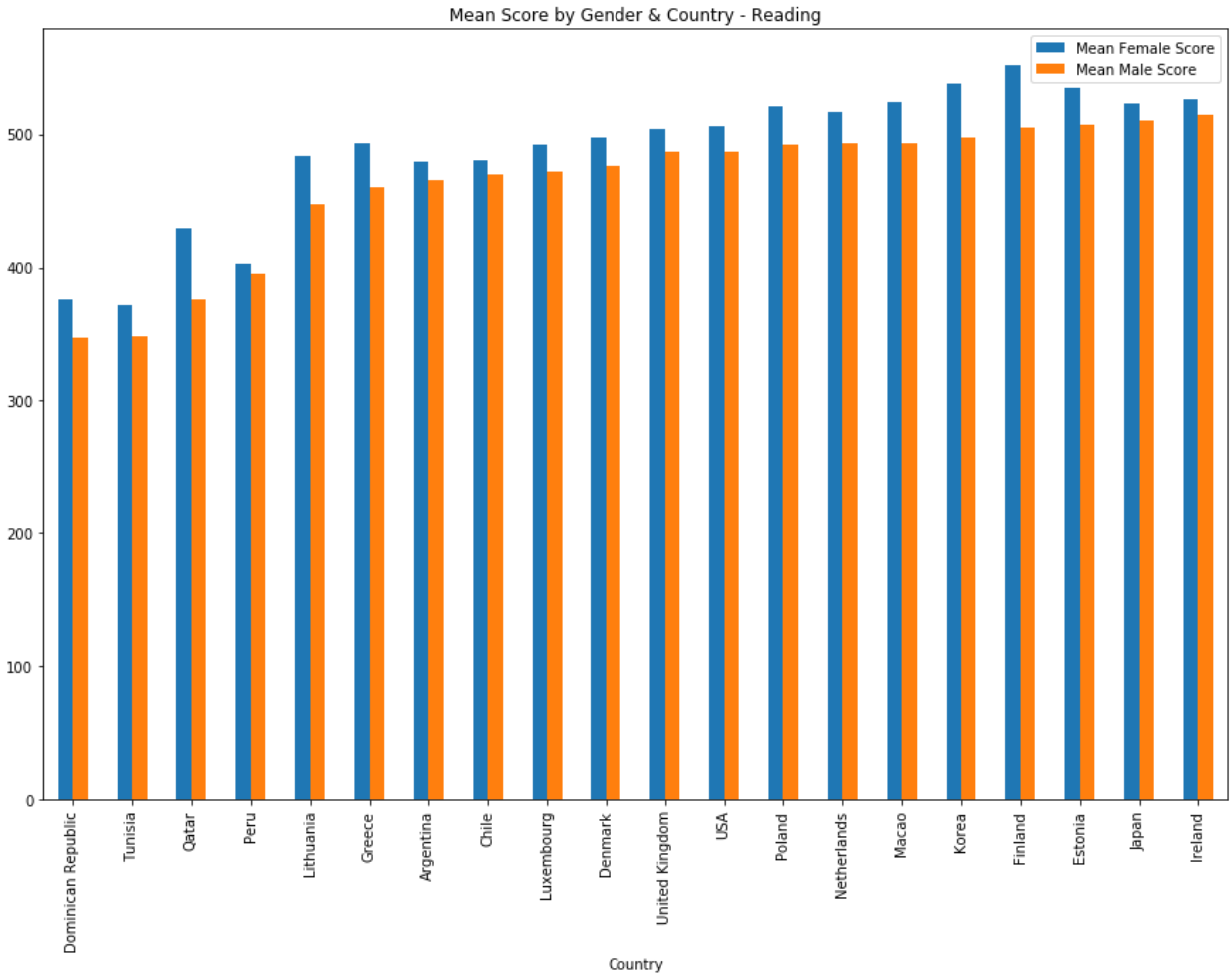
While most countries do show higher math performance for males compared to females, Macau, China, Qatar, Finland, Dominican Republic, and Korea are the exception. What do these countries do differently than the rest of our sample population to ensure females are not falling behind in STEM? We hoped our analysis of some socio-economic and cultural factors provided clues to the answer to this question.

Science Results



Again, several countries show they have cracked the code with STEM mastery in their female student population. Something noteworthy is that for all these countries except Finland and Macao, China- the top twenty percent of performers were mostly male. This means that the male population must have a wider standard deviation of Science Test scores.

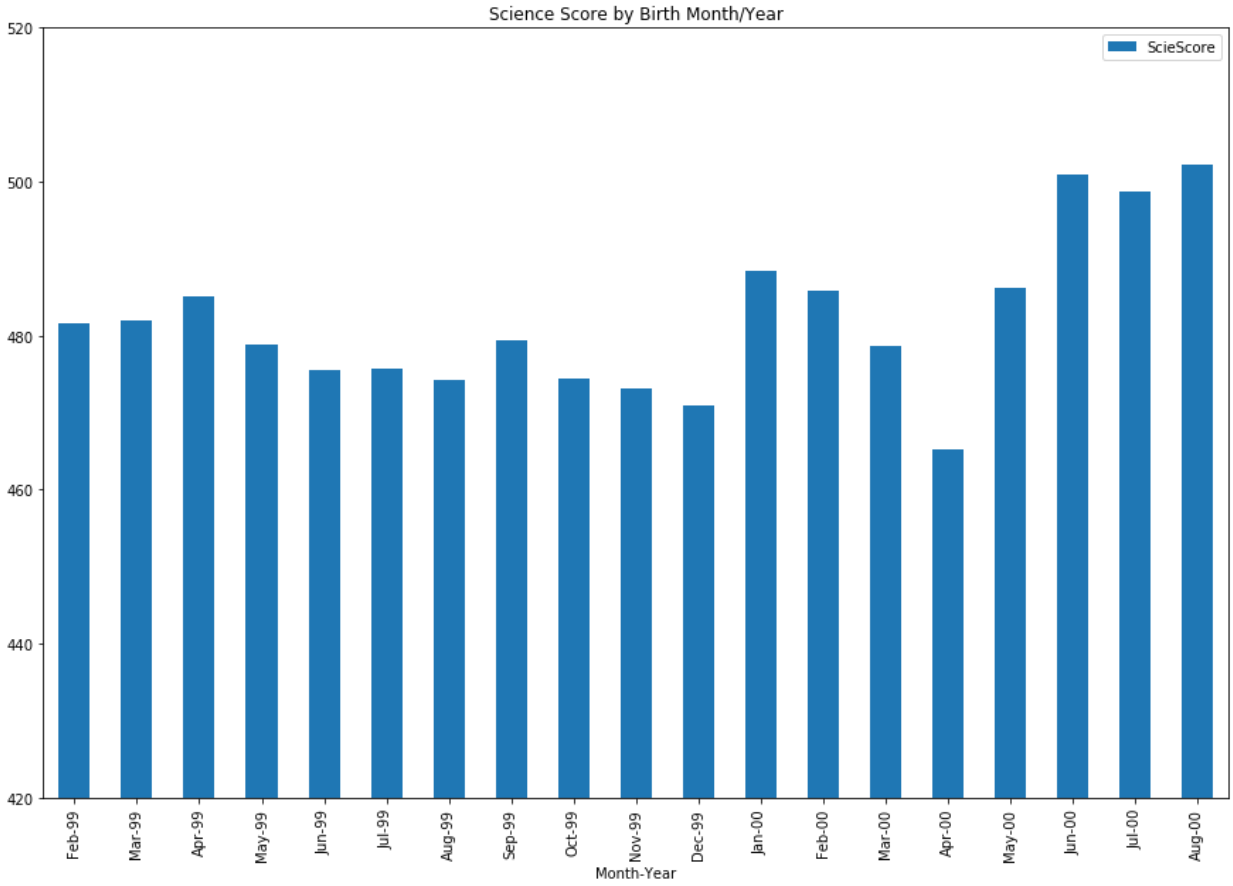
Reading Results: Girls just want to have books



Trying to level the playing field for boys in reading seems like a more challenging problem. There is not a single country in our sample where the boys outscored the girls in the reading subject area. Out of our sample countries, boys made up just 43% of top twenty scores, and this graph depicts higher mean scores for females in every single country.

Demographics: Age

Parents across the USA have been holding their children back in the hopes that the older they are, the better they will perform. This may be true in the younger years, but the data in the PISA test did not bear this out.



Socio-Economic Factors

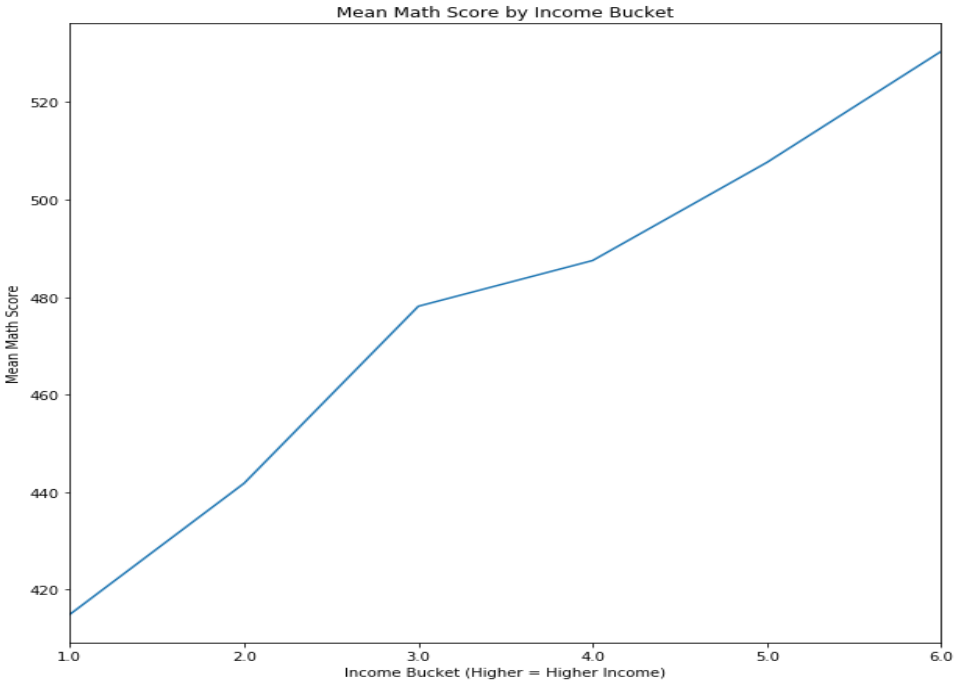
Socio-economic factors potentially play a role in the scores of the students tested. If we consider, for example, that students with higher economic status might have more resources at their disposal to succeed in their education. To take a deeper dive into this hypothesis, we extracted questions from the PISA that directly indicate higher income, as well as questions that could be used as a proxy to a higher economic status. The parents of the students tested were asked what their income was, what their level of education was, and whether they hired tutors for their children. In this portion of the analysis, we also consider questions about the child's home. For example, we analyze the questions that ask whether the child has a desk, a room to study in, a computer, and an Internet connection. Finally, we consider questions to the student about whether they have to work after school, either at home or for pay.

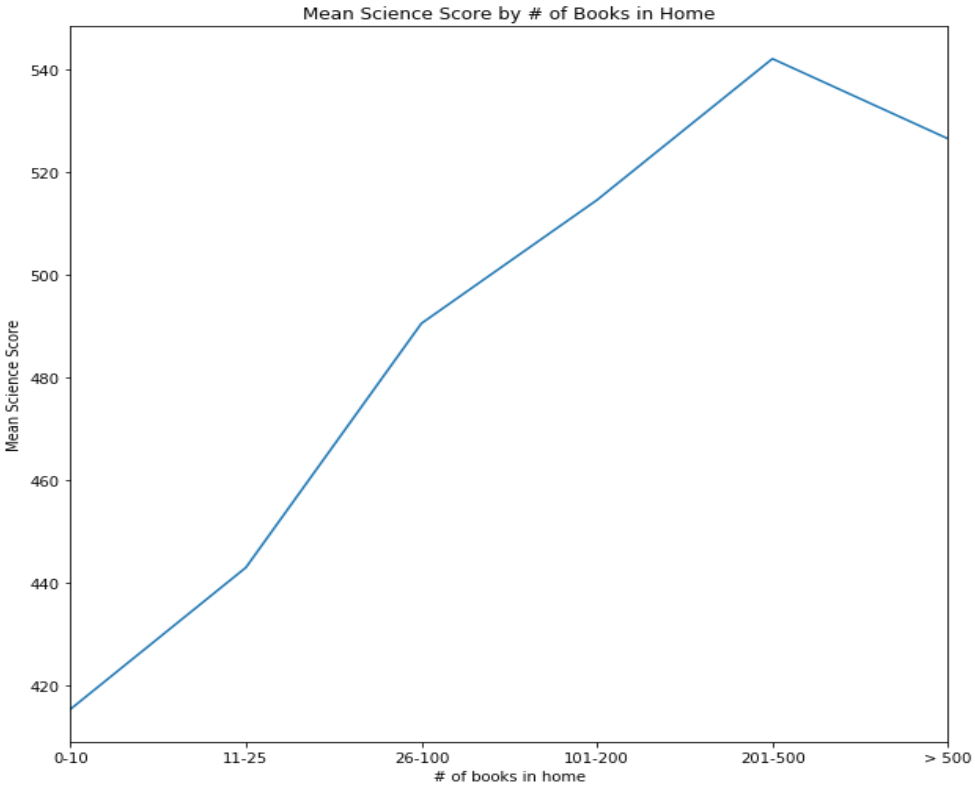
Like the other questions we analyzed, the socio-economic questions had to be cleaned before analysis. Primarily, many variables had to be stripped of leading spaces in each row. Some variables also had additional single quotes that had to be taken off before any analysis could take place. Since this dataset was originally written and saved in a format readable in SAS, we think many of these issues occurred during the translation of this data into Python. We also made the decision to strip the dataset from answers that contained no or null responses. In

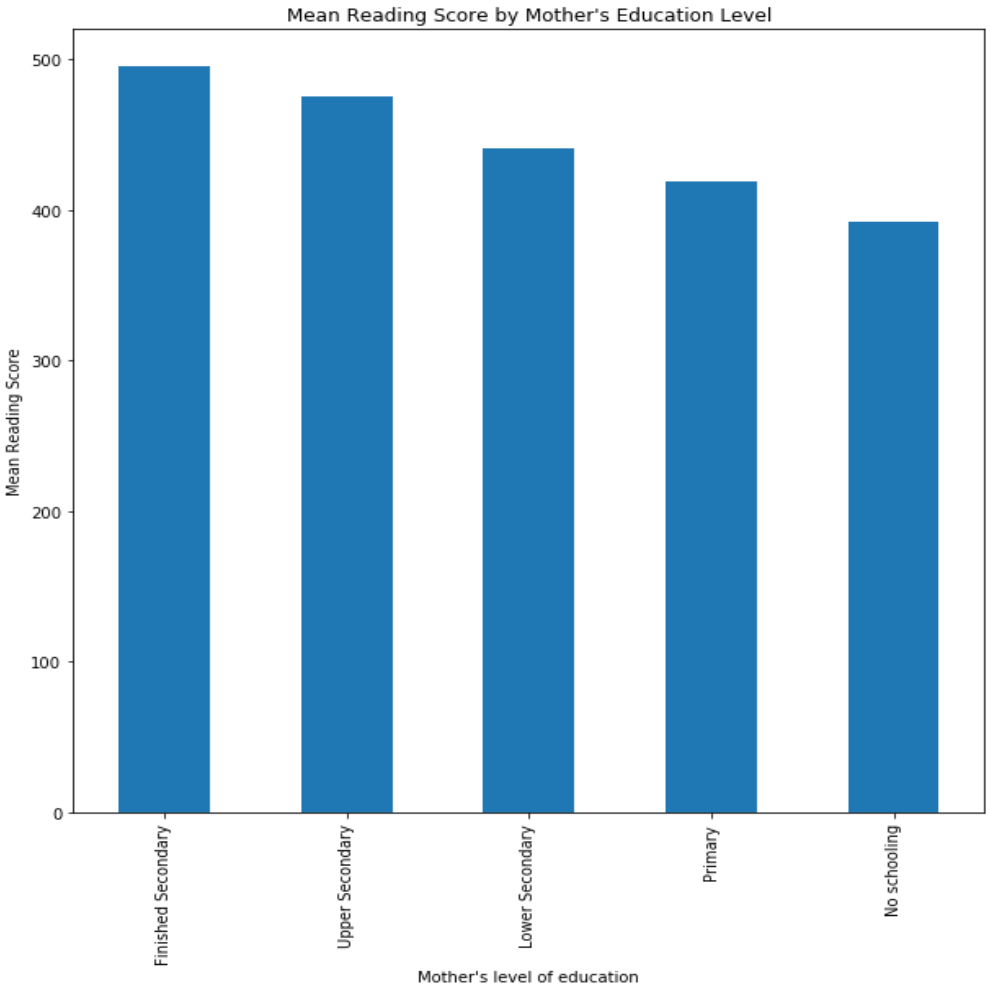
addition to the cleaning of the dataset, we recorded some variables (i.e. from country codes to country names) so that they were easier to understand.

An initial inquiry into the socio-economic questions turned out some expected and unexpected results. The questions relating to whether the students had a desk or a quiet study area resulted in a negative correlation to the Math, Reading, and Science scores obtained by the children. The number of books the children had in their household, however, has a modest positive correlation of 0.4, 0.4, 0.41, for their Math, Reading, and Science scores, respectively. As expected, there was also a positive correlation on the level of income (Math: 0.4, Reading: 0.41, Science: 0.41) and level of education of the parents with test scores (Math: .25, Reading: .27, Science: .26). The higher the income bucket in which the parents belonged, the higher scores were for students. Similar results were seen the higher the parents' spending was on tutors. The final unexpected result was that students who responded that they worked for pay after school actually had higher scores than students who didn't work for pay after school.

Socioeconomic questions:
ST005Q01TA – What is the <highest level of schooling> completed by your mother?
ST007Q01TA – What is the <highest level of schooling> completed by your father?
ST011Q01TA – In your home: A desk to study at
ST011Q02TA – In your home: A room of your own
ST011Q03TA – In your home: A quiet place to study
ST011Q04TA – In your home: A computer you can use for school work
ST011Q06TA – In your home: A link to the Internet
ST012Q09NA – How many in your home: Musical instruments (e.g. guitar, piano)
ST013Q01TA – How many books are there in your home?
ST078Q09NA – After leaving school did you: Work in the household or take care of other family members
ST078Q10NA – After leaving school did you: Work for pay
PA042Q01TA – What is your annual household income?
PA003Q02TA – Eat <the main meal> with my child around a table.
PA041Q01TA – In the last twelve months, about how much would you have paid to educational providers for services?


Mean Math Score by Income Bucket
Mean Math Score
520
500
480
460
440
420
1.0
2.0
3.0
4.0
5.0
6.0
Income Bucket (Higher = Higher Income)


Mean Science Score by # of Books in Home
Mean Science Score
540
520
500
480
460
440
420
0-10
11-25
26-100
101-200
201-500
> 500
# of books in home



Cultural Factors

When processing the cultural factors variables, the data had to be cleaned to ensure extraneous spaces were stripped, just as it had to be done for the demographic data. Also, examination of the value counts for some of the continuous variables pointed out potential errors and inconsistencies. For example, the variables for number of required class periods per week in science, math, and reading had observations as high as 40 periods. The variable for the total number of required periods (all classes) per week had observations exceeding 100.

Clearly, this was erroneous data. We excluded the outlier observations, but only for for purposes of analyzing the cultural factors in question. Observations were ignored for required number of periods of math, science or reading exceeding 15 periods, and observations were ignored for total required periods exceeding 50. The the vast majority of observations remained after excluding this data. We also dropped all observations with Null values for cultural factor analyses.

The cultural factor questions aligned in groups:

Motivation:

ST119Q01NA – I want top grades
ST119Q02NA – I want to select from among the top schools
ST113Q04TA – Many things I learn in science will help get a job
ST062Q01TA - # times I skipped school in the last 2 weeks
ST062Q02TA - # times I skipped some class in the last 2 weeks
ST062Q03TA - # times I arrived late in the last 2 weeks

Workload/class time:

ST059Q01TA - # or required periods per week in native language
ST059Q02TA - # of required periods per week in math
ST059Q03TA - # of required periods per week in science
ST060Q01NA – total # of class periods required per week
ST061Q01NA – average number of minutes in a class period
ST071Q01NA - # of additional hours per week attended in science
ST031Q01NA - # of days per week a physical education class is attended

School environment:

ST097Q01TA – How often do students not listen to what the teacher is saying?
ST097Q02TA – How often is there noise and disorder in the classroom?
ST097Q03TA – How often does the teacher have to wait for students to quiet down?
ST097Q04TA – How often are students unable to work well in the classroom?
ST097Q05TA – How often do students not start working for a long period of time in class?

Parent engagement:

PA003Q01TA – In your family, do you discuss how well your child is doing in school?
PA003Q03TA – In your family, do you spend time just talking with your child?
PA003Q05NA – In your family, do you ask how your child is doing in science class?

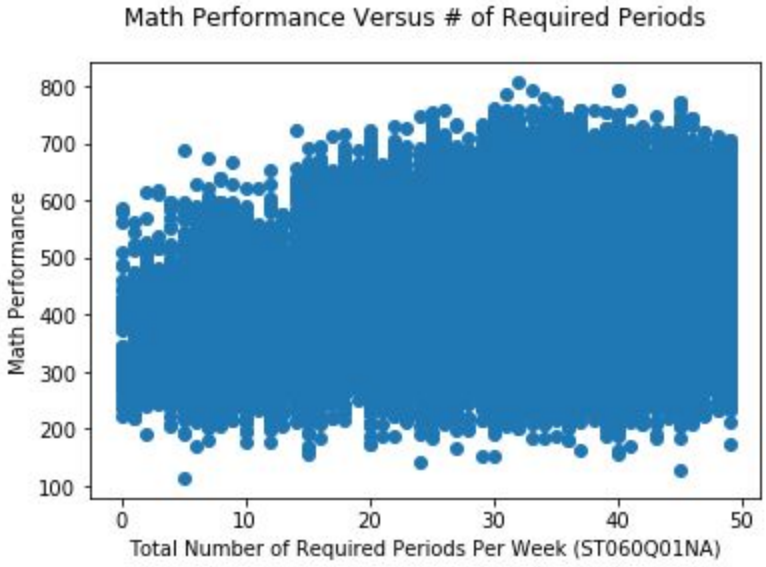
We performed correlation analyses (df.corr()) for each of the above factors against the math, reading and science scores. We wanted to see if there was a strong correlation between any of the factors and student performance.

The results showed very little correlation between these factors and math, reading or science performance. Most correlation factors were below .15.

The only factors that showed modest correlation were:

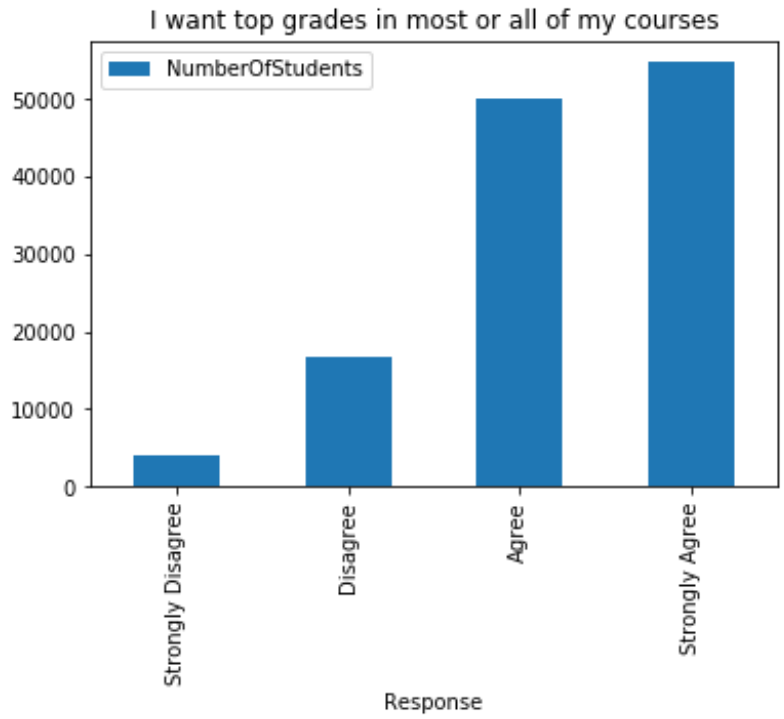
> ST060Q01NA – total # of class periods required per week (Math correlation .22)
> ST062Q01TA - # times I skipped school in the last 2 weeks (Math -.25; Reading -.23; Science -.23)
> ST071Q01NA - # of additional hours per week attended in science (Math -.25; Reading -.27; Science -.25)

Scatter plots confirm weak correlation. For example:



The correlations, though weak, for total # of class periods and for # of times I skipped class make sense. The correlations for # of additional hours per week did not, at first, make sense. On further thought, we reasoned that students taking extra classes outside of their regular school schedule may be students requiring tutoring to improve performance.

One possible reason for the lack of strong correlation between the cultural factors and student performance is the presence of self-serving bias in student and parent survey responses. For example, student responses to the question: To what extent do you disagree or agree with the following statements about yourself? I want top grades in most or all of my courses. (Strongly disagree, Disagree, Agree, Strongly Agree)



If we were to continue this investigation further, we would see if performing the analyses by gender resulted in stronger correlations. Running correlations by country may also show stronger correlations.

Conclusion

We set out to examine demographic, socioeconomic, and cultural patterns in the PISA student performance data. We saw that the data confirmed the commonly held belief that female students are better at reading than males. The data tended to support the belief that males are better than females in math and science, but not for all countries examined. Females clearly performed better than males in math and science in a few key countries.

Our analysis did not show a noticeable impact of age on performance. Students who recently turned 15 years old did not necessarily perform worse than those who were close to age 16.

Some socio-economic factors showed moderate correlation to performance. For example, the highest education level attained by parents does correlate well with the educational performance of their children.

Cultural factors showed weak correlation at best to student performance, and then only for a very few factors. Our conjecture is that self-serving bias may be influencing student and parent responses on their respective surveys, resulting in poor data.

We did not have opportunity to examine all of the 900+ factors in the PISA data. We are confident that analysis of all of the factors would likely reveal multiple strong and intriguing

correlations. Also, deeper analysis of the factors we did examine might also show interesting results. For example, examining the socio-economic and cultural factors by country and by gender would be interesting.