PISA Data Description and Wrangling

Overview

PISA is a survey of students' skills and knowledge as they approach the end of compulsory education. It is not a conventional school test. Rather than examining how well students have learned the school curriculum, it looks at how well prepared they are for life beyond school.

Around 510,000 students in 65 economies took part in the PISA 2012 assessment of reading, mathematics and science representing about 28 million 15-year-olds globally. Of those economies, 44 took part in an assessment of creative problem solving and 18 in an assessment of financial literacy. For more details see PISA website.

The data and topics of investigation come from the PISA Data Visualization Competition.

Topics suggested by Udacity:

1. The importance of school factors in explaining academic performance.
2. Differences in achievement based on gender, location, or student attitudes.
3. Differences in achievement based on teacher practices and attitudes.
4. Inequalities in academic achievement.

Data wrangling

In this report the PISA 2012 will be used to investigate the differences in achievement in mathematics tests based on location, gender and student attitudes. Keeping these tasks in mind, the data wrangling will proceed as follows:

1. Download the two datafiles 'pisadict2012.csv' (which contains the description of all codes and abbreviations in the main table) and 'pisa2012.csv' (the main datafile, the unzipped csv file is 2.75 GB).
2. Wrangle the dictionary of terms file, keep only those columns that are relevant to this analysis.
3. Use sqlalchemy to extract a managable size Pandas dataframe from the main PISA data file, this is done using tthe methods described in Working with large csv files in Python.
4. Clean some minor issues regarding the countries involved in the study.

In [1]:

```
### import the necessary packages to work with the datasets
import numpy as np
import pandas as pd

from sqlalchemy import create_engine
```

In [2]:

```
### option to display full content of columns in the dataframes
pd.set_option('display.max_colwidth', -1)
```

The dictionary of terms datafile

In [3]:

```
### save the dictionary of terms as pandas dataframe
df_dict=pd.read_csv("pisadict2012.csv", encoding='iso-8859-1')
```

In [4]:

```
### investigate the dataframe
df_dict.sample(4)
```

Out[4]:

	Unnamed: 0	x
585	W_FSTR35	FINAL STUDENT REPLICATE BRR-FAY WEIGHT35
442	HOSTCUL	Acculturation: Host Culture Oriented Strategies
551	W_FSTR1	FINAL STUDENT REPLICATE BRR-FAY WEIGHT1
254	ST89Q05	Attitude toward School - Trying Hard is Important

In [5]:

```
### rename the columns
df_dict.columns = ['Code', 'Description']
```

In [6]:

```
### get more information about the dataframe
df_dict.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 635 entries, 0 to 634
Data columns (total 2 columns):
Code           635 non-null object
Description    635 non-null object
dtypes: object(2)
memory usage: 10.0+ KB
```

Downsizing the list of columns to be used in the analysis:

1. Investigate the available codes in the list, there are several categories of such codes: abbreviations, ST#, IC#, EC#, PV# and W_FSTR#.
2. Look at ST# codes, identification codes from the beginning of the dataframe, and codes that contain the work mathematics (or versions of it) in description.
3. ST codes are divided on categories, keep some of ST3 and all ST4 and ST9 entries.
4. Manually select the codes that make reference to mathematics in description.

In [7]:

```
df_dict.head(20);
```

In [8]:

```
### the codes that contain information about the student
df1 = df_dict.iloc[[0, 1, 2, 3, 6, 7, 11], :]
df1
```

Out[8]:

	Code	Description

0	CNT	Country code 3-character
1	SUBNATIO	Adjudicated sub-region code 7-digit code (3-digit country code + region ID + stratum ID)
2	STRATUM	Stratum ID 7-character (cnt + region ID + original stratum ID)
3	OECD	OECD country
6	STIDSTD	Student ID
7	ST01Q01	International Grade
11	ST04Q01	Gender

In [9]:

```
### investigate the ST0# codes and extract them in a separate dataframe
df_st0=df_dict[df_dict['Code'].str.contains('ST0')]
df_st0
```

Out[9]:

	Code	Description
7	ST01Q01	International Grade
8	ST02Q01	National Study Programme
9	ST03Q01	Birth - Month
10	ST03Q02	Birth -Year
11	ST04Q01	Gender
12	ST05Q01	Attend <ISCED 0>
13	ST06Q01	Age at <ISCED 1>
14	ST07Q01	Repeat - <ISCED 1>
15	ST07Q02	Repeat - <ISCED 2>
16	ST07Q03	Repeat - <ISCED 3>
17	ST08Q01	Truancy - Late for School
18	ST09Q01	Truancy - Skip whole school day

In [10]:

```
### investigate the ST1# codes and extract them in a separate dataframe
df_st1=df_dict[df_dict['Code'].str.contains('ST1')]
df_st1
```

Out[10]:

	Code	Description
19	ST115Q01	Truancy - Skip classes within school day
20	ST11Q01	At Home - Mother
21	ST11Q02	At Home - Father
22	ST11Q03	At Home - Brothers
23	ST11Q04	At Home - Sisters
24	ST11Q05	At Home - Grandparents

25	ST11Q06	At Home - Others
26	ST13Q01	Mother<Highest Schooling>
27	ST14Q01	Mother Qualifications - <ISCED level 6>
28	ST14Q02	Mother Qualifications - <ISCED level 5A>
29	ST14Q03	Mother Qualifications - <ISCED level 5B>
30	ST14Q04	Mother Qualifications - <ISCED level 4>
31	ST15Q01	Mother Current Job Status
32	ST17Q01	Father<Highest Schooling>
33	ST18Q01	Father Qualifications - <ISCED level 6>
34	ST18Q02	Father Qualifications - <ISCED level 5A>
35	ST18Q03	Father Qualifications - <ISCED level 5B>
36	ST18Q04	Father Qualifications - <ISCED level 4>
37	ST19Q01	Father Current Job Status
275	ST101Q01	Problem Route Selection - Read brochure
276	ST101Q02	Problem Route Selection - Study map
277	ST101Q03	Problem Route Selection - Leave it to brother
278	ST101Q05	Problem Route Selection - Just drive
279	ST104Q01	Problem Ticket Machine - Similarities
280	ST104Q04	Problem Ticket Machine - Try buttons
281	ST104Q05	Problem Ticket Machine - Ask for help
282	ST104Q06	Problem Ticket Machine - Find ticket office

In [11]:

```
### investigate the ST2# codes and extract them in a separate dataframe
df_st2=df_dict[df_dict['Code'].str.contains('ST2')]
df_st2
```

Out[11]:

	Code	Description
38	ST20Q01	Country of Birth International - Self
39	ST20Q02	Country of Birth International - Mother
40	ST20Q03	Country of Birth International - Father
41	ST21Q01	Age of arrival in <country of test>
42	ST25Q01	International Language at Home
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
49	ST26Q07	Possessions - literature
50	ST26Q08	Possessions - poetry
51	ST26Q09	Possessions - art
52	ST26Q10	Possessions - textbooks
53	ST26Q11	Possessions - <technical reference books>
54	ST26Q12	Possessions - dictionary
55	ST26Q13	Possessions - dishwasher
56	ST26Q14	Possessions - <DVD>
57	ST26Q15	Possessions - <Country item 1>
58	ST26Q16	Possessions - <Country item 2>
59	ST26Q17	Possessions - <Country item 3>
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
66	ST29Q01	Math Interest - Enjoy Reading
67	ST29Q02	Instrumental Motivation - Worthwhile for Work
68	ST29Q03	Math Interest - Look Forward to Lessons
69	ST29Q04	Math Interest - Enjoy Maths
70	ST29Q05	Instrumental Motivation - Worthwhile for Career Chances
71	ST29Q06	Math Interest - Interested
72	ST29Q07	Instrumental Motivation - Important for Future Study
73	ST29Q08	Instrumental Motivation - Helps to Get a Job
391	ST22Q01	Acculturation - Mother Immigrant (Filter)
392	ST23Q01	Acculturation - Enjoy <Host Culture> Friends
393	ST23Q02	Acculturation - Enjoy <Heritage Culture> Friends
394	ST23Q03	Acculturation - Enjoy <Host Culture> Celebrations
395	ST23Q04	Acculturation - Enjoy <Heritage Culture> Celebrations
396	ST23Q05	Acculturation - Spend Time with <Host Culture> Friends

397	ST23Q06	Acculturation - Spend Time with <Heritage Culture> Friends
398	ST23Q07	Acculturation - Participate in <Host Culture> Celebrations
399	ST23Q08	Acculturation - Participate in <Heritage Culture> Celebrations
400	ST24Q01	Acculturation - Perceived Host-Heritage Cultural Differences - Values
401	ST24Q02	Acculturation - Perceived Host-Heritage Cultural Differences - Mother Treatment
402	ST24Q03	Acculturation - Perceived Host-Heritage Cultural Differences - Teacher Treatment

In [12]:

```
### investigate the ST3# codes and extract them in a separate dataframe
df_st3_all=df_dict[df_dict['Code'].str.contains('ST3')]
df_st3_all
```

Out[12]:

	Code	Description
74	ST35Q01	Subjective Norms -Friends Do Well in Mathematics
75	ST35Q02	Subjective Norms -Friends Work Hard on Mathematics
76	ST35Q03	Subjective Norms - Friends Enjoy Mathematics Tests
77	ST35Q04	Subjective Norms - Parents Believe Studying Mathematics Is Important
78	ST35Q05	Subjective Norms - Parents Believe Mathematics Is Important for Career
79	ST35Q06	Subjective Norms - Parents Like Mathematics
80	ST37Q01	Math Self-Efficacy - Using a <Train Timetable>
81	ST37Q02	Math Self-Efficacy - Calculating TV Discount
82	ST37Q03	Math Self-Efficacy - Calculating Square Metres of Tiles
83	ST37Q04	Math Self-Efficacy - Understanding Graphs in Newspapers
84	ST37Q05	Math Self-Efficacy - Solving Equation 1
85	ST37Q06	Math Self-Efficacy - Distance to Scale
86	ST37Q07	Math Self-Efficacy - Solving Equation 2
87	ST37Q08	Math Self-Efficacy - Calculate Petrol Consumption Rate

In [13]:

```
### only the first six ST3# codes are relevant for this study
### extract them in a separate dataframe
df_st3 = df_st3_all.iloc[0:6, :]
df_st3
```

Out[13]:

	Code	Description
74	ST35Q01	Subjective Norms -Friends Do Well in Mathematics
75	ST35Q02	Subjective Norms -Friends Work Hard on Mathematics
76	ST35Q03	Subjective Norms - Friends Enjoy Mathematics Tests
77	ST35Q04	Subjective Norms - Parents Believe Studying Mathematics Is Important
78	ST35Q05	Subjective Norms - Parents Believe Mathematics Is Important for Career

79	ST35Q06	Subjective Norms - Parents Like Mathematics

In [14]:

```
### investigate the ST4# codes and extract them in a separate dataframe
df_st4=df_dict[df_dict['Code'].str.contains('ST4')]
df_st4
```

Out[14]:

	Code	Description
88	ST42Q01	Math Anxiety - Worry That It Will Be Difficult
89	ST42Q02	Math Self-Concept - Not Good at Maths
90	ST42Q03	Math Anxiety - Get Very Tense
91	ST42Q04	Math Self-Concept- Get Good <Grades>
92	ST42Q05	Math Anxiety - Get Very Nervous
93	ST42Q06	Math Self-Concept - Learn Quickly
94	ST42Q07	Math Self-Concept - One of Best Subjects
95	ST42Q08	Math Anxiety - Feel Helpless
96	ST42Q09	Math Self-Concept - Understand Difficult Work
97	ST42Q10	Math Anxiety - Worry About Getting Poor <Grades>
98	ST43Q01	Perceived Control - Can Succeed with Enough Effort
99	ST43Q02	Perceived Control - Doing Well is Completely Up to Me
100	ST43Q03	Perceived Control - Family Demands and Problems
101	ST43Q04	Perceived Control - Different Teachers
102	ST43Q05	Perceived Control - If I Wanted I Could Perform Well
103	ST43Q06	Perceived Control - Perform Poorly Regardless
104	ST44Q01	Attributions to Failure - Not Good at Maths Problems
105	ST44Q03	Attributions to Failure - Teacher Did Not Explain Well
106	ST44Q04	Attributions to Failure - Bad Guesses
107	ST44Q05	Attributions to Failure - Material Too Hard
108	ST44Q07	Attributions to Failure - Teacher Didnt Get Students Interested
109	ST44Q08	Attributions to Failure - Unlucky
110	ST46Q01	Math Work Ethic - Homework Completed in Time
111	ST46Q02	Math Work Ethic - Work Hard on Homework
112	ST46Q03	Math Work Ethic - Prepared for Exams
113	ST46Q04	Math Work Ethic - Study Hard for Quizzes
114	ST46Q05	Math Work Ethic - Study Until I Understand Everything
115	ST46Q06	Math Work Ethic - Pay Attention in Classes

116	ST46Q07	Math Work Ethic - Listen in Classes
117	ST46Q08	Math Work Ethic - Avoid Distractions When Studying
118	ST46Q09	Math Work Ethic - Keep Work Organized
119	ST48Q01	Math Intentions - Mathematics vs. Language Courses After School
120	ST48Q02	Math Intentions - Mathematics vs. Science Related Major in College
121	ST48Q03	Math Intentions - Study Harder in Mathematics vs. Language Classes
122	ST48Q04	Math Intentions - Take Maximum Number of Mathematics vs. Science Classes
123	ST48Q05	Math Intentions - Pursuing a Career That Involves Mathematics vs. Science
124	ST49Q01	Math Behaviour - Talk about Maths with Friends
125	ST49Q02	Math Behaviour - Help Friends with Maths
126	ST49Q03	Math Behaviour - <Extracurricular> Activity
127	ST49Q04	Math Behaviour - Participate in Competitions
128	ST49Q05	Math Behaviour - Study More Than 2 Extra Hours a Day
129	ST49Q06	Math Behaviour - Play Chess
130	ST49Q07	Math Behaviour - Computer programming
131	ST49Q09	Math Behaviour - Participate in Math Club

In [15]:

```
### investigate the ST5# codes and extract them in a separate dataframe
df_st5=df_dict[df_dict['Code'].str.contains('ST5')]
df_st5
```

Out[15]:

	Code	Description
132	ST53Q01	Learning Strategies- Important Parts vs. Existing Knowledge vs. Learn by Heart
133	ST53Q02	Learning Strategies- Improve Understanding vs. New Ways vs. Memory
134	ST53Q03	Learning Strategies - Other Subjects vs. Learning Goals vs. Rehearse Problems
135	ST53Q04	Learning Strategies - Repeat Examples vs. Everyday Applications vs. More Information
136	ST55Q01	Out of school lessons - <test lang>
137	ST55Q02	Out of school lessons - <maths>
138	ST55Q03	Out of school lessons - <science>
139	ST55Q04	Out of school lessons - other
140	ST57Q01	Out-of-School Study Time - Homework
141	ST57Q02	Out-of-School Study Time - Guided Homework
142	ST57Q03	Out-of-School Study Time - Personal Tutor
143	ST57Q04	Out-of-School Study Time - Commercial Company
144	ST57Q05	Out-of-School Study Time - With Parent

145	ST57Q06	Out-of-School Study Time - Computer

In [16]:

```
### investigate the ST6# codes and extract them in a separate dataframe
df_st6=df_dict[df_dict['Code'].str.contains('ST6')]
df_st6
```

Out[16]:

	Code	Description
146	ST61Q01	Experience with Applied Maths Tasks - Use <Train Timetable>
147	ST61Q02	Experience with Applied Maths Tasks - Calculate Price including Tax
148	ST61Q03	Experience with Applied Maths Tasks - Calculate Square Metres
149	ST61Q04	Experience with Applied Maths Tasks - Understand Scientific Tables
150	ST61Q05	Experience with Pure Maths Tasks - Solve Equation 1
151	ST61Q06	Experience with Applied Maths Tasks - Use a Map to Calculate Distance
152	ST61Q07	Experience with Pure Maths Tasks - Solve Equation 2
153	ST61Q08	Experience with Applied Maths Tasks - Calculate Power Consumption Rate
154	ST61Q09	Experience with Applied Maths Tasks - Solve Equation 3
155	ST62Q01	Familiarity with Math Concepts - Exponential Function
156	ST62Q02	Familiarity with Math Concepts - Divisor
157	ST62Q03	Familiarity with Math Concepts - Quadratic Function
158	ST62Q04	Overclaiming - Proper Number
159	ST62Q06	Familiarity with Math Concepts - Linear Equation
160	ST62Q07	Familiarity with Math Concepts - Vectors
161	ST62Q08	Familiarity with Math Concepts - Complex Number
162	ST62Q09	Familiarity with Math Concepts - Rational Number
163	ST62Q10	Familiarity with Math Concepts - Radicals
164	ST62Q11	Overclaiming - Subjunctive Scaling
165	ST62Q12	Familiarity with Math Concepts - Polygon
166	ST62Q13	Overclaiming - Declarative Fraction
167	ST62Q15	Familiarity with Math Concepts - Congruent Figure
168	ST62Q16	Familiarity with Math Concepts - Cosine
169	ST62Q17	Familiarity with Math Concepts - Arithmetic Mean
170	ST62Q19	Familiarity with Math Concepts - Probability
171	ST69Q01	Min in <class period> - <test lang>
172	ST69Q02	Min in <class period> - <Maths>
173	ST69Q03	Min in <class period> - <Science>

In [17]:

```
### investigate the ST7# codes and extract them in a separate dataframe
df_st7=df_dict[df_dict['Code'].str.contains('ST7')]
df_st7
```

Out[17]:

	Code	Description
174	ST70Q01	No of <class period> p/wk - <test lang>
175	ST70Q02	No of <class period> p/wk - <Maths>
176	ST70Q03	No of <class period> p/wk - <Science>
177	ST71Q01	No of ALL <class period> a week
178	ST72Q01	Class Size - No of Students in <Test Language> Class
179	ST73Q01	OTL - Algebraic Word Problem in Math Lesson
180	ST73Q02	OTL - Algebraic Word Problem in Tests
181	ST74Q01	OTL - Procedural Task in Math Lesson
182	ST74Q02	OTL - Procedural Task in Tests
183	ST75Q01	OTL - Pure Math Reasoning in Math Lesson
184	ST75Q02	OTL - Pure Math Reasoning in Tests
185	ST76Q01	OTL - Applied Math Reasoning in Math Lesson
186	ST76Q02	OTL - Applied Math Reasoning in Tests
187	ST77Q01	Math Teaching - Teacher shows interest
188	ST77Q02	Math Teaching - Extra help
189	ST77Q04	Math Teaching - Teacher helps
190	ST77Q05	Math Teaching - Teacher continues
191	ST77Q06	Math Teaching - Express opinions
192	ST79Q01	Teacher-Directed Instruction - Sets Clear Goals
193	ST79Q02	Teacher-Directed Instruction - Encourages Thinking and Reasoning
194	ST79Q03	Student Orientation - Differentiates Between Students When Giving Tasks
195	ST79Q04	Student Orientation - Assigns Complex Projects
196	ST79Q05	Formative Assessment - Gives Feedback
197	ST79Q06	Teacher-Directed Instruction - Checks Understanding
198	ST79Q07	Student Orientation - Has Students Work in Small Groups
199	ST79Q08	Teacher-Directed Instruction - Summarizes Previous Lessons
200	ST79Q10	Student Orientation - Plans Classroom Activities
201	ST79Q11	Formative Assessment - Gives Feedback on Strengths and Weaknesses
202	ST79Q12	Formative Assessment - Informs about Expectations
203	ST79Q15	Teacher-Directed Instruction - Informs about Learning Goals

204	ST79Q17	Formative Assessment - Tells How to Get Better

In [18]:

```python
### investigate the ST8# codes and extract them in a separate dataframe
df_st8=df_dict[df_dict['Code'].str.contains('ST8')]
df_st8
```

Out[18]:

	Code	Description
205	ST80Q01	Cognitive Activation - Teacher Encourages to Reflect Problems
206	ST80Q04	Cognitive Activation - Gives Problems that Require to Think
207	ST80Q05	Cognitive Activation - Asks to Use Own Procedures
208	ST80Q06	Cognitive Activation - Presents Problems with No Obvious Solutions
209	ST80Q07	Cognitive Activation - Presents Problems in Different Contexts
210	ST80Q08	Cognitive Activation - Helps Learn from Mistakes
211	ST80Q09	Cognitive Activation - Asks for Explanations
212	ST80Q10	Cognitive Activation - Apply What We Learned
213	ST80Q11	Cognitive Activation - Problems with Multiple Solutions
214	ST81Q01	Disciplinary Climate - Students Dont Listen
215	ST81Q02	Disciplinary Climate - Noise and Disorder
216	ST81Q03	Disciplinary Climate - Teacher Has to Wait Until its Quiet
217	ST81Q04	Disciplinary Climate - Students Dont Work Well
218	ST81Q05	Disciplinary Climate - Students Start Working Late
219	ST82Q01	Vignette Teacher Support -Homework Every Other Day/Back in Time
220	ST82Q02	Vignette Teacher Support - Homework Once a Week/Back in Time
221	ST82Q03	Vignette Teacher Support - Homework Once a Week/Not Back in Time
222	ST83Q01	Teacher Support - Lets Us Know We Have to Work Hard
223	ST83Q02	Teacher Support - Provides Extra Help When Needed
224	ST83Q03	Teacher Support - Helps Students with Learning
225	ST83Q04	Teacher Support - Gives Opportunity to Express Opinions
226	ST84Q01	Vignette Classroom Management - Students Frequently Interrupt/Teacher Arrives Early
227	ST84Q02	Vignette Classroom Management - Students Are Calm/Teacher Arrives on Time
228	ST84Q03	Vignette Classroom Management - Students Frequently Interrupt/Teacher Arrives Late
229	ST85Q01	Classroom Management - Students Listen
230	ST85Q02	Classroom Management - Teacher Keeps Class Orderly
231	ST85Q03	Classroom Management - Teacher Starts On Time

232	ST85Q04	Classroom Management - Wait Long to <Quiet Down>
233	ST86Q01	Student-Teacher Relation - Get Along with Teachers
234	ST86Q02	Student-Teacher Relation - Teachers Are Interested
235	ST86Q03	Student-Teacher Relation - Teachers Listen to Students
236	ST86Q04	Student-Teacher Relation - Teachers Help Students
237	ST86Q05	Student-Teacher Relation - Teachers Treat Students Fair
238	ST87Q01	Sense of Belonging - Feel Like Outsider
239	ST87Q02	Sense of Belonging - Make Friends Easily
240	ST87Q03	Sense of Belonging - Belong at School
241	ST87Q04	Sense of Belonging - Feel Awkward at School
242	ST87Q05	Sense of Belonging - Liked by Other Students
243	ST87Q06	Sense of Belonging - Feel Lonely at School
244	ST87Q07	Sense of Belonging - Feel Happy at School
245	ST87Q08	Sense of Belonging - Things Are Ideal at School
246	ST87Q09	Sense of Belonging - Satisfied at School
247	ST88Q01	Attitude towards School - Does Little to Prepare Me for Life
248	ST88Q02	Attitude towards School - Waste of Time
249	ST88Q03	Attitude towards School - Gave Me Confidence
250	ST88Q04	Attitude towards School- Useful for Job
251	ST89Q02	Attitude toward School - Helps to Get a Job
252	ST89Q03	Attitude toward School - Prepare for College
253	ST89Q04	Attitude toward School - Enjoy Good Grades
254	ST89Q05	Attitude toward School - Trying Hard is Important

In [19]:

```
### investigate the ST9# codes and extract them in a separate dataframe
df_st9=df_dict[df_dict['Code'].str.contains('ST9')]
df_st9
```

Out[19]:

	Code	Description
255	ST91Q01	Perceived Control - Can Succeed with Enough Effort
256	ST91Q02	Perceived Control - My Choice Whether I Will Be Good
257	ST91Q03	Perceived Control - Problems Prevent from Putting Effort into School
258	ST91Q04	Perceived Control - Different Teachers Would Make Me Try Harder
259	ST91Q05	Perceived Control - Could Perform Well if I Wanted
260	ST91Q06	Perceived Control - Perform Poor Regardless
261	ST93Q01	Perseverance - Give up easily

262	ST93Q03	Perseverance - Put off difficult problems
263	ST93Q04	Perseverance - Remain interested
264	ST93Q06	Perseverance - Continue to perfection
265	ST93Q07	Perseverance - Exceed expectations
266	ST94Q05	Openness for Problem Solving - Can Handle a Lot of Information
267	ST94Q06	Openness for Problem Solving - Quick to Understand
268	ST94Q09	Openness for Problem Solving - Seek Explanations
269	ST94Q10	Openness for Problem Solving - Can Link Facts
270	ST94Q14	Openness for Problem Solving - Like to Solve Complex Problems
271	ST96Q01	Problem Text Message - Press every button
272	ST96Q02	Problem Text Message - Trace steps
273	ST96Q03	Problem Text Message - Manual
274	ST96Q05	Problem Text Message - Ask a friend

In [20]:

```
### create a list of codes that contain the word M(m)athematic(s) in their
description
### and are not in the ST# category

df_math = df_dict[(df_dict['Description'].str.contains('athematic'))
                  & (~df_dict['Code'].str.contains('ST3'))
                  & (~df_dict['Code'].str.contains('ST4'))]
df_math
```

Out[20]:

	Code	Description
413	ANXMAT	Mathematics Anxiety
419	CLSMAN	Mathematics Teacher's Classroom Management
423	COGACT	Cognitive Activation in Mathematics Lessons
429	EXAPPLM	Experience with Applied Mathematics Tasks at School
430	EXPUREM	Experience with Pure Mathematics Tasks at School
431	FAILMAT	Attributions to Failure in Mathematics
432	FAMCON	Familiarity with Mathematical Concepts
433	FAMCONC	Familiarity with Mathematical Concepts (Signal Detection Adjusted)
452	INSTMOT	Instrumental Motivation for Mathematics
453	INTMAT	Mathematics Interest
461	MATBEH	Mathematics Behaviour
462	MATHEFF	Mathematics Self-Efficacy
463	MATINTFC	Mathematics Intentions

464	MATWKETH	Mathematics Work Ethic
466	MMINS	Learning time (minutes per week)- <Mathematics>
467	MTSUP	Mathematics Teacher's Support
475	SCMAT	Mathematics Self-Concept
478	SUBNORM	Subjective Norms in Mathematics
485	USEMATH	Use of ICT in Mathematic Lessons
491	ANCCLSMAN	Mathematics Teacher's Classroom Management (Anchored)
492	ANCCOGACT	Cognitive Activation in Mathematics Lessons (Anchored)
493	ANCINSTMOT	Instrumental Motivation for Mathematics (Anchored)
494	ANCINTMAT	Mathematics Interest (Anchored)
495	ANCMATWKETH	Mathematics Work Ethic (Anchored)
496	ANCMTSUP	Mathematics Teacher's Support (Anchored)
497	ANCSCMAT	Mathematics Self-Concept (Anchored)
499	ANCSUBNORM	Subjective Norms in Mathematics (Anchored)
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

In [21]:

```
### create a list of codes that contains reference to math
df_math_mat = df_math[(df_math['Code'].str.contains('MAT'))
                      & (~df_math['Code'].str.contains('ANC'))]
df_math_mat
```

Out[21]:

	Code	Description
413	ANXMAT	Mathematics Anxiety
431	FAILMAT	Attributions to Failure in Mathematics
453	INTMAT	Mathematics Interest
461	MATBEH	Mathematics Behaviour
462	MATHEFF	Mathematics Self-Efficacy
463	MATINTFC	Mathematics Intentions
464	MATWKETH	Mathematics Work Ethic
475	SCMAT	Mathematics Self-Concept
485	USEMATH	Use of ICT in Mathematic Lessons
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

In [22]:

```
### use the previous steps to create a dataframe
### that contains the list of column names to be extracted from the main datafile
df_dict_clean=pd.concat([df1, df_st3, df_st4, df_st9, df_math_mat])
df_dict_clean
```

Out[22]:

	Code	Description
0	CNT	Country code 3-character
1	SUBNATIO	Adjudicated sub-region code 7-digit code (3-digit country code + region ID + stratum ID)
2	STRATUM	Stratum ID 7-character (cnt + region ID + original stratum ID)
3	OECD	OECD country
6	STIDSTD	Student ID
7	ST01Q01	International Grade
11	ST04Q01	Gender
74	ST35Q01	Subjective Norms -Friends Do Well in Mathematics
75	ST35Q02	Subjective Norms -Friends Work Hard on Mathematics
76	ST35Q03	Subjective Norms - Friends Enjoy Mathematics Tests
77	ST35Q04	Subjective Norms - Parents Believe Studying Mathematics Is Important
78	ST35Q05	Subjective Norms - Parents Believe Mathematics Is Important for Career
79	ST35Q06	Subjective Norms - Parents Like Mathematics
88	ST42Q01	Math Anxiety - Worry That It Will Be Difficult
89	ST42Q02	Math Self-Concept - Not Good at Maths
90	ST42Q03	Math Anxiety - Get Very Tense
91	ST42Q04	Math Self-Concept- Get Good <Grades>
92	ST42Q05	Math Anxiety - Get Very Nervous
93	ST42Q06	Math Self-Concept - Learn Quickly
94	ST42Q07	Math Self-Concept - One of Best Subjects
95	ST42Q08	Math Anxiety - Feel Helpless
96	ST42Q09	Math Self-Concept - Understand Difficult Work
97	ST42Q10	Math Anxiety - Worry About Getting Poor <Grades>
98	ST43Q01	Perceived Control - Can Succeed with Enough Effort

99	ST43Q02	Perceived Control - Doing Well is Completely Up to Me
100	ST43Q03	Perceived Control - Family Demands and Problems
101	ST43Q04	Perceived Control - Different Teachers
102	ST43Q05	Perceived Control - If I Wanted I Could Perform Well
103	ST43Q06	Perceived Control - Perform Poorly Regardless
104	ST44Q01	Attributions to Failure - Not Good at Maths Problems
...	...	...
259	ST91Q05	Perceived Control - Could Perform Well if I Wanted
260	ST91Q06	Perceived Control - Perform Poor Regardless
261	ST93Q01	Perseverance - Give up easily
262	ST93Q03	Perseverance - Put off difficult problems
263	ST93Q04	Perseverance - Remain interested
264	ST93Q06	Perseverance - Continue to perfection
265	ST93Q07	Perseverance - Exceed expectations
266	ST94Q05	Openness for Problem Solving - Can Handle a Lot of Information
267	ST94Q06	Openness for Problem Solving - Quick to Understand
268	ST94Q09	Openness for Problem Solving - Seek Explanations
269	ST94Q10	Openness for Problem Solving - Can Link Facts
270	ST94Q14	Openness for Problem Solving - Like to Solve Complex Problems
271	ST96Q01	Problem Text Message - Press every button
272	ST96Q02	Problem Text Message - Trace steps
273	ST96Q03	Problem Text Message - Manual
274	ST96Q05	Problem Text Message - Ask a friend
413	ANXMAT	Mathematics Anxiety
431	FAILMAT	Attributions to Failure in Mathematics
453	INTMAT	Mathematics Interest
461	MATBEH	Mathematics Behaviour
462	MATHEFF	Mathematics Self-Efficacy
463	MATINTFC	Mathematics Intentions
464	MATWKETH	Mathematics Work Ethic
475	SCMAT	Mathematics Self-Concept
485	USEMATH	Use of ICT in Mathematic Lessons
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

91 rows × 2 columns

In [23]:
```
### store the selected set of codes as a csv file
df_dict_clean.to_csv('pisadict2012_clean.csv', index=False)
```

In [24]:
```
### write the selected codes to a list and print this list
selected_codes = df_dict_clean['Code'].tolist()
print(','.join(selected_codes))
```

```
CNT,SUBNATIO,STRATUM,OECD,STIDSTD,ST01Q01,ST04Q01,ST35Q01,ST35Q02,ST35Q03,S
T35Q04,ST35Q05,ST35Q06,ST42Q01,ST42Q02,ST42Q03,ST42Q04,ST42Q05,ST42Q06,ST42
Q07,ST42Q08,ST42Q09,ST42Q10,ST43Q01,ST43Q02,ST43Q03,ST43Q04,ST43Q05,ST43Q06
,ST44Q01,ST44Q03,ST44Q04,ST44Q05,ST44Q07,ST44Q08,ST46Q01,ST46Q02,ST46Q03,ST
46Q04,ST46Q05,ST46Q06,ST46Q07,ST46Q08,ST46Q09,ST48Q01,ST48Q02,ST48Q03,ST48Q
04,ST48Q05,ST49Q01,ST49Q02,ST49Q03,ST49Q04,ST49Q05,ST49Q06,ST49Q07,ST49Q09,
ST91Q01,ST91Q02,ST91Q03,ST91Q04,ST91Q05,ST91Q06,ST93Q01,ST93Q03,ST93Q04,ST9
3Q06,ST93Q07,ST94Q05,ST94Q06,ST94Q09,ST94Q10,ST94Q14,ST96Q01,ST96Q02,ST96Q0
3,ST96Q05,ANXMAT,FAILMAT,INTMAT,MATBEH,MATHEFF,MATINTFC,MATWKETH,SCMAT,USEM
ATH,PV1MATH,PV2MATH,PV3MATH,PV4MATH,PV5MATH
```

The PISA2012 main datafile

In [25]:
```
### set up a variable that points to the csv file
pisa = "pisa2012.csv"
```

In [26]:
```
### take a look at the 'head' of the csv file to see what the contents migh
t look like
pd.read_csv(pisa, nrows=5)
```

Out[26]:

	Unnamed: 0	CNT	SUBNATIO	STRATUM	OECD	NC	SCHOOLID	STIDSTD	ST
0	1	Albania	80000	ALB0006	Non-OECD	Albania	1	1	10
1	2	Albania	80000	ALB0006	Non-OECD	Albania	1	2	10
2	3	Albania	80000	ALB0006	Non-OECD	Albania	1	3	9
3	4	Albania	80000	ALB0006	Non-OECD	Albania	1	4	9
4	5	Albania	80000	ALB0006	Non-OECD	Albania	1	5	9

5 rows × 636 columns

In [27]:
```
### create a local sqllite database
csv_dbase = create_engine('sqlite:///csv_dbase.db')
```

In [28]:
```
### iterate through the CSV file in chunks and store the data into sqllite

chunksize = 10000
i = 0
j = 1
for df in pd.read_csv(pisa, chunksize=chunksize,
                     encoding='iso-8859-1', iterator=True, low_memory=Fals
e):
      df = df.rename(columns={c: c.replace(' ', '') for c in df.columns})
      df.index += j
      i+=1
      df.to_sql('table', csv_dbase, if_exists='append')
      j = df.index[-1] + 1
```

In [29]:
```
###create the cleaned dataframe that contains the 'selected_codes' only
df_pisa = pd.read_sql('SELECT CNT,SUBNATIO,STRATUM,OECD,STIDSTD,\
                           ST01Q01,ST04Q01,\
                           ST35Q01,ST35Q02,ST35Q03,ST35Q04,ST35Q05,ST35Q06,
\
                           ST42Q01,ST42Q02,ST42Q03,ST42Q04,ST42Q05,\
                           ST42Q06,ST42Q07,ST42Q08,ST42Q09,ST42Q10,\
                           ST43Q01,ST43Q02,ST43Q03,ST43Q04,ST43Q05,ST43Q06,
\
                           ST44Q01,ST44Q03,ST44Q04,ST44Q05,ST44Q07,ST44Q08,
\
                           ST46Q01,ST46Q02,ST46Q03,ST46Q04,ST46Q05,ST46Q06,
\
                           ST46Q07,ST46Q08,ST46Q09,\
                           ST48Q01,ST48Q02,ST48Q03,ST48Q04,ST48Q05,\
                           ST49Q01,ST49Q02,ST49Q03,ST49Q04,ST49Q05,  \
                           ST49Q06,ST49Q07,ST49Q09,\
                           ST91Q01,ST91Q02,ST91Q03,ST91Q04,ST91Q05,ST91Q06,
\
                           ST93Q01,ST93Q03,ST93Q04,ST93Q06,ST93Q07,\
                           ST94Q05,ST94Q06,ST94Q09,ST94Q10, ST94Q14,\
                           ST96Q01,ST96Q02,ST96Q03,ST96Q05,\
                           ANXMAT,FAILMAT,INTMAT,MATBEH,MATHEFF,\
                           MATINTFC,MATWKETH,SCMAT,USEMATH,\
                           PV1MATH,PV2MATH,PV3MATH,PV4MATH,PV5MATH \
                           FROM "table"', csv_dbase)
```

In [30]:
```
df_pisa.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 485490 entries, 0 to 485489
Data columns (total 91 columns):
CNT         485490 non-null object
SUBNATIO    485490 non-null int64
STRATUM     485490 non-null object
OECD        485490 non-null object
```

```
STIDSTD     485490 non-null int64
ST01Q01     485490 non-null int64
ST04Q01     485490 non-null object
ST35Q01     315860 non-null object
ST35Q02     315315 non-null object
ST35Q03     314873 non-null object
ST35Q04     315160 non-null object
ST35Q05     314843 non-null object
ST35Q06     313389 non-null object
ST42Q01     313855 non-null object
ST42Q02     313502 non-null object
ST42Q03     312176 non-null object
ST42Q04     311980 non-null object
ST42Q05     312624 non-null object
ST42Q06     312327 non-null object
ST42Q07     312583 non-null object
ST42Q08     312456 non-null object
ST42Q09     312223 non-null object
ST42Q10     312853 non-null object
ST43Q01     314971 non-null object
ST43Q02     314182 non-null object
ST43Q03     313494 non-null object
ST43Q04     313420 non-null object
ST43Q05     313228 non-null object
ST43Q06     313470 non-null object
ST44Q01     314119 non-null object
ST44Q03     313405 non-null object
ST44Q04     312645 non-null object
ST44Q05     312996 non-null object
ST44Q07     312970 non-null object
ST44Q08     313374 non-null object
ST46Q01     313898 non-null object
ST46Q02     313567 non-null object
ST46Q03     312994 non-null object
ST46Q04     312997 non-null object
ST46Q05     313043 non-null object
ST46Q06     312900 non-null object
ST46Q07     312854 non-null object
ST46Q08     312989 non-null object
ST46Q09     313040 non-null object
ST48Q01     294410 non-null object
ST48Q02     289827 non-null object
ST48Q03     298479 non-null object
ST48Q04     267716 non-null object
ST48Q05     287992 non-null object
ST49Q01     313495 non-null object
ST49Q02     313025 non-null object
ST49Q03     312168 non-null object
ST49Q04     312378 non-null object
ST49Q05     312582 non-null object
ST49Q06     312571 non-null object
ST49Q07     312425 non-null object
ST49Q09     312752 non-null object
ST91Q01     311430 non-null object
ST91Q02     310396 non-null object
ST91Q03     309826 non-null object
ST91Q04     309398 non-null object
```

```
ST91Q05     309610 non-null object
ST91Q06     309656 non-null object
ST93Q01     312856 non-null object
ST93Q03     312140 non-null object
ST93Q04     311311 non-null object
ST93Q06     312270 non-null object
ST93Q07     312259 non-null object
ST94Q05     312404 non-null object
ST94Q06     312185 non-null object
ST94Q09     311413 non-null object
ST94Q10     311747 non-null object
ST94Q14     312001 non-null object
ST96Q01     311381 non-null object
ST96Q02     311460 non-null object
ST96Q03     311078 non-null object
ST96Q05     311319 non-null object
ANXMAT      314764 non-null float64
FAILMAT     314448 non-null float64
INTMAT      316708 non-null float64
MATBEH      313847 non-null float64
MATHEFF     315948 non-null float64
MATINTFC    301360 non-null float64
MATWKETH    314501 non-null float64
SCMAT       314607 non-null float64
USEMATH     290260 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
dtypes: float64(14), int64(3), object(74)
memory usage: 337.1+ MB
```

In [31]:

```
### list of participating countries
set(df_pisa.CNT)
```

Out[31]:

```
{'Albania',
 'Argentina',
 'Australia',
 'Austria',
 'Belgium',
 'Brazil',
 'Bulgaria',
 'Canada',
 'Chile',
 'China-Shanghai',
 'Chinese Taipei',
 'Colombia',
 'Connecticut (USA)',
 'Costa Rica',
 'Croatia',
 'Czech Republic',
 'Denmark',
 'Estonia',
 'Finland',
 'Florida (USA)',
 'France',
 'Germany',
 'Greece',
 'Hong Kong-China',
 'Hungary',
 'Iceland',
 'Indonesia',
 'Ireland',
 'Israel',
 'Italy',
 'Japan',
 'Jordan',
 'Kazakhstan',
 'Korea',
 'Latvia',
 'Liechtenstein',
 'Lithuania',
 'Luxembourg',
 'Macao-China',
 'Malaysia',
 'Massachusetts (USA)',
 'Mexico',
 'Montenegro',
 'Netherlands',
 'New Zealand',
 'Norway',
 'Perm(Russian Federation)',
 'Peru',
 'Poland',
 'Portugal',
 'Qatar',
 'Romania',
 'Russian Federation',
 'Serbia',
 'Singapore',
 'Slovak Republic',
 'Slovenia',
 'Spain',
 'Sweden',
 'Switzerland',
 'Thailand',
 'Tunisia',
 'Turkey',
 'United Arab Emirates',
 'United Kingdom',
 'United States of America',
 'Uruguay',
 'Vietnam'}
```

In [32]:

```
### replace 'Florida (USA)', 'Connecticut (USA)' and 'Massacusets (USA)'
### with 'United States of America
df_pisa['CNT'].replace('Connecticut (USA)', 'United States of America', in
place=True)
df_pisa['CNT'].replace('Florida (USA)', 'United States of America', inplac
e=True)
df_pisa['CNT'].replace('Massachusetts (USA)', 'United States of America',
inplace=True)
```

In [33]:

```
### replace 'Perm(Russian Federation)' with 'Russian Federation'
df_pisa['CNT'].replace('Perm(Russian Federation)', 'Russian Federation', i
nplace=True)
```

In [34]:

```
### combine 'China-Shangai', 'Hong King - China', 'Macao-China'  as 'China'
df_pisa['CNT'].replace('China-Shanghai', 'China', inplace=True)
df_pisa['CNT'].replace('Hong Kong-China', 'China', inplace=True)
df_pisa['CNT'].replace('Macao-China', 'China', inplace=True)
```

In [35]:

```
### the updated list of countries
set(df_pisa.CNT)
```

Out[35]:

```
{'Albania',
 'Argentina',
 'Australia',
 'Austria',
 'Belgium',
 'Brazil',
 'Bulgaria',
 'Canada',
 'Chile',
 'China',
 'Chinese Taipei',
 'Colombia',
 'Costa Rica',
 'Croatia',
 'Czech Republic',
 'Denmark',
 'Estonia',
 'Finland',
 'France',
 'Germany',
 'Greece',
 'Hungary',
 'Iceland',
 'Indonesia',
 'Ireland',
 'Israel',
 'Italy',
 'Japan',
 'Jordan',
 'Kazakhstan',
 'Korea',
 'Latvia',
 'Liechtenstein',
 'Lithuania',
 'Luxembourg',
 'Malaysia',
 'Mexico',
 'Montenegro',
 'Netherlands',
 'New Zealand',
 'Norway',
 'Peru',
 'Poland',
 'Portugal',
 'Qatar',
 'Romania',
 'Russian Federation',
 'Serbia',
 'Singapore',
 'Slovak Republic',
 'Slovenia',
 'Spain',
 'Sweden',
 'Switzerland',
 'Thailand',
 'Tunisia',
 'Turkey',
 'United Arab Emirates',
 'United Kingdom',
 'United States of America',
 'Uruguay',
 'Vietnam'}
```

In [36]:

```
### the number of participating countries, as defined here
len(set(df_pisa.CNT))
```

Out[36]:

```
62
```

In [37]:

```
### review the cleaned dataframe
df_pisa.head(4)
```

Out[37]:

	CNT	SUBNATIO	STRATUM	OECD	STIDSTD	ST01Q01	ST04Q01	ST35Q01	ST
0	Albania	80000	ALB0006	Non-OECD	1	10	Female	Disagree	Ag
1	Albania	80000	ALB0006	Non-OECD	2	10	Female	Strongly agree	Str ag
2	Albania	80000	ALB0006	Non-OECD	3	9	Female	Strongly agree	Str ag
3	Albania	80000	ALB0006	Non-OECD	4	9	Female	None	No

4 rows × 91 columns

In [38]:

```
### get an overall description of the numerical data
df_pisa.describe()
```

Out[38]:

	SUBNATIO	STIDSTD	ST01Q01	ANXMAT	FAILMAT	IN
count	4.854900e+05	485490.000000	485490.000000	314764.000000	314448.000000	316708.0
mean	4.315457e+06	6134.066201	9.813323	0.152647	-0.013110	0.212424
std	2.524434e+06	6733.144944	3.734726	0.955031	1.029037	1.00471
min	8.000000e+04	1.000000	7.000000	-2.370000	-3.766600	-1.78000
25%	2.030000e+06	1811.000000	9.000000	-0.470000	-0.530000	-0.34000

50%	4.100000e+06	3740.000000	10.000000	0.060000	-0.076000	0.30000
75%	6.880000e+06	7456.000000	10.000000	0.790000	0.640000	0.91000
max	8.580000e+06	33806.000000	96.000000	2.550000	3.906700	2.29000

In [39]:

```
### store the cleaned dataframe as a csv file
df_pisa.to_csv('pisa2012_clean.csv', index=False)
```

In [ ]:

```

```