Introduction

We expect to have four components in our project to analyze the international datasets from PISA and World Bank Education data. Our data inquiry project is to depict countries' test performances on PISA and investigate contextual factors relating to student and sociocultural characteristics that can explain gender differences in test performances across countries.

Component 1: Data retrieving: clarifying data type, files, and restrictions for use.

Download and install 'savReaderWriter', which can read sav files and create their pandas dataframes.

- ○ Input:
    - ■ CY6_MS_CMB_STU_QQQ.sav: Dataset of student level that includes educational features such as country ID, school ID, gender, age, family information, etc.
    - ■ CY6_MS_CMB_SCH_QQQ.sav: Dataset of school level that includes educational features such as country ID, projector number, study room number, educational programs, etc.
- ○ To merge 3-level PISA data files: 'pandas', 'csv', 'savReaderWriter' packages are required in this component. We first read files into pandas dataframe. Each student belongs to a school and each school belongs to a country. Therefore, we can make two dictionaries mapping from students to schools and from schools to countries respectively. In addition, we can link PISA file with World Bank file using 'country_label' feature.
- ○ Output: Dataframe_PISA, dataframe_WB
- ○ Interactions: the output will be used to do data cleaning.

Component 2: Data cleaning

PISA data cleanup

PISA 2015 raw data contains 921 variables that have numeric and string values. After the dataset is retrieved, readable in csv file, and merged, we will use a function that takes `pandas` dataframe and clean it according to interested variables identified to explain the issue in students' gender performance differences. For missing data, PISA does not impute missing information for questionnaire variables. PISA variables contains missing data information that are not of our interests of analysis. Thus PISA dataset contains four kinds of missing data codes that are used across all countries.

- Load the data into python using a pandas DataFrame
- PISA missing data cleaning: to employ listwise deletion on four types: "Nonresponse", "missing or invalid", "not applicable" and "not reached", which in four different cases are coded as 5-9. 95-99, 995-999 and 9995-9999 in PISA raw data.
- Perform a quick check up: `isnull().values.any()`
- One method is listwise deletion, using dframe.dropna(inplace = True). Alternatively we can specify missing values in python using na_values, fillna().
- Output: dataframe_PISA

World Bank data cleanup

World Bank Education Equality data contains series names after selection. Country names are mapped with PISA data and added country ID. From merging PISA and WB datasets, a function has been designed to link the World Bank and PISA dataframe (input) to the one that contains country-level columns from both (output). In order to do that, we looked at WB dataset for country-level aggregated data provided by EdStats_Indicators query data (link below). Since WB data has multiple years and each country has one value correspondingly, we can use previous years data to replace any NaN value with mean or mode of the data, fillna can be used with row means, such as `dframe.fillna(value = dframe.mean(), inplace = True)`.
http://databank.worldbank.org/Data/indicator/UIS.LPP.AG15T99?id=c755d342&report_name=EdStats_Indicators_Report&populartype=series#

- Input: dataframe_PISA, dataframe_WB
- Output: dataframe_coun (merged file on country-level)

Component 3: Data processing
With cleaned data, following functions are used for further data analysis.

- ○ Usage: Functions of data analysis, modeling, and testing
- ○ Function1(country_subject): Input a variable, Output table descriptives including min, mean, mode, max, etc
- ○ Hierarchical level modeling for explaining gender differences in student performance

Function 1:

country_subject(country, subject):
Input: Str country
Str subject
Output: tuple (differences of two genders, 10% score, mean, 90% score)

Given a country name and a subject, return this country's four scores in this subject as a tuple.

Function 2:

countries_subject(subject):
Input: Str subject
Output: map{key country: value score-tuple)

Given a subject, return a map whose keys are different countries, and value are score-tuples.

Function 3:

countries_variable(variable):
Input: Str variable
Output: map{key country: value value of the variable}

Given a variable, return a map whose keys are different countries, and value are the country’s value of the variable.

Component 4: Visualization:
We expect to use Bokeh package as our visualization tool. It's a Python interactive visualization library that can create interactive plots, dashboards, and applications.
The Python packages that will be utilized for visualization:

- Bokeh
- Output
  - Visualization, tables, graphs (design specification)
  - Variables correlation visual

The visualization will consist of several plots of the Educational Test data. These visuals give reference for further analysis that can be done based on the given database. The visualization includes the following:

- An overall observation of male and female test performance by using box plot.
- Descriptives of measures of contextual factors across countries and interaction terms. For example, we'd like to know which course subjects are male and female students good at respectively for each country. They will be visualized on a map.
- (if appropriate) Hierarchical linear modeling plot and respective interactional terms among the countries which have highest and lowest gender difference scores, we'd like to know what factors are contributing to the gap between average difference score of two genders.

Interactions
The components listed above are specifications needed for users to read data easily, perform exemplar analysis and view visualizations. The output of data retrieval is a single database of PISA three-level data and WB country data, which can be read into python `pandas` for cleaning process. The output of data cleaning component leads to the culmination of data descriptives and in-depth analysis into relationships of gender difference and cultural variables. The analysis is focused on correlation analysis and multilevel linear modeling after scraping, merging and cleaning. The output will feed into visualization of findings.
Parents
Enthusiastic parents who care about the quality of children's education would look at the comprehensive and international data by using our analysis tool to find out what's the best decision to make when it comes to children's education. Questions they hold may concern the school quality and type, such as public and private school performance, and associated school cultural characteristics. Global parents may want to know the overall quality of education by test performance, in an interest or decision of migration.

Schools , educators, teachers, and practitioners
Teachers may want to know the practices of better-performing schools and associated instructive efficacy for ideas of instruction improvement and strategies. By selecting variables on teacher-level and students' performance data they could see correlations of intended variables. They may also want to know the better picture of education across countries or pay attention to gender inequality and what social and cultural factors interweaved into it.

Education researchers and research organizations

Researchers may want to investigate research questions such as gender equality in education and educational performances across countries. The National Center of Education Statistics provided by federal grants seek to understand education data like this to make data-informed policy reforms, incentives, implementations to improve U.S. education quality.

Testing organizations, companies and administration

Testing organizations and companies need test statistics for feedbacks and insights on item development, measurement, and data-oriented decisions to make tests and test items more equitable, especially differential item functionally between different genders. They also need to use the data for international comparisons of test performances.

Project plan

At this stage, we have outlined the following project tasks based on priorities.

- Data retrieving and cleaning: including data type, file, restrictions for use, --due W6
    - Data linkage of different files
    - Data cleaning
- Data functions and package creation: --due W7
    - statistic models creation
    - Package selection
- Testing and pylint styles --due W8
    - Unit testing
    - Code quality and commenting -PEP8, PyLint;
    - Integrity checks
    - Project structure checks
- Data output --due W9
    - Visualization development, tables, graphs
    - (design specification and function specification updates)