# Problem Description

The Programme for International Student Assessment (PISA) is a test given every three years to 15-year-old students from around the world to evaluate their performance in mathematics, reading, and science. This test provides a quantitative way to compare the performance of students from different parts of the world. In this homework assignment, we will predict the reading scores of students from the United States of America on the 2009 PISA exam.

The datasets contain information about the demographics and schools for American students taking the 2009 exam. The data is collected from [National Center for Education Statistics](https://mydatacafe.indielms.com/teacher_lesson/show/nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2011038).

Each row in the datasets pisa2009train.csv and pisa2009test.csv represents one student taking the exam. The datasets have the following variables:

**grade:**The grade in school of the student (most 15-year-olds in America are in 10th grade)

**male:**Whether the student is male (1/0)

**raceeth:**The race/ethnicity composite of the student

**preschool:**Whether the student attended preschool (1/0)

**expectBachelors:**Whether the student expects to obtain a bachelor's degree (1/0)

motherHS: Whether the student's mother completed high school (1/0)

**motherBachelors:**Whether the student's mother obtained a bachelor's degree (1/0)

**motherWork:**Whether the student's mother has part-time or full-time work (1/0)

**fatherHS:**Whether the student's father completed high school (1/0)

**fatherBachelors:**Whether the student's father obtained a bachelor's degree (1/0)

**fatherWork:**Whether the student's father has part-time or full-time work (1/0)

**selfBornUS:**Whether the student was born in the United States of America (1/0)

**motherBornUS:**Whether the student's mother was born in the United States of America (1/0)

**fatherBornUS:**Whether the student's father was born in the United States of America (1/0)

**englishAtHome:**Whether the student speaks English at home (1/0)

**computerForSchoolwork:**Whether the student has access to a computer for schoolwork (1/0)

read30MinsADay: Whether the student reads for pleasure for 30 minutes/day (1/0)

**minutesPerWeekEnglish:**The number of minutes per week the student spend in English class

**studentsInEnglish:**The number of students in this student's English class at school

**schoolHasLibrary:**Whether this student's school has a library (1/0)

**publicSchool:**Whether this student attends a public school (1/0)

**urban:**Whether this student's school is in an urban area (1/0)

**schoolSize:** The number of students in this student's school

**readingScore:**The student's reading score, on a 1000-point scale

Questions:

1. Load the training and testing sets using the read\_csv() function, and save them as dataframes with the names pisaTrain and pisaTest.   
   How many students are there in the training set?
2. Using groupby on pisaTrain, what is the average reading test score of males?

Of Females?

1. Which variables have missing data in at least one observation in the training set?
2. Linear regression discards observations with missing data, so we will remove all such observations from the training and testing sets. Ideally we should be filling the missing values with plausible information, but for now, lets remove them and move ahead.   
     
   Type the following commands into your notebook to remove observations with any missing value from pisaTrain and pisaTest:   
     
   pisaTrain = train.dropna(axis=0) #axis=0 because we want to delete the rows having NA   
     
   pisaTest = test.dropna(axis=0)   
     
   How many observations are now in the training set?

How many observations are now in the testing set?

1. Which of the following variables is an unordered factor/categorical variable with at least 3 levels?
2. To include unordered categorical variables in a linear regression model, we define one level as the "reference level" and add a binary variable for each of the remaining levels. In this way, a categorical variable with n levels is replaced by n-1 binary variables. The reference level is typically selected to be the most frequently occurring level in the data-set or the first level in alphabetical order. ( We can change from one Level to another as well, Note:- This makes no difference to our solution, it remains the same irrespective of the reference taken)   
     
   As an example, consider the unordered categorical variable variable "color", with levels "red", "green", and "blue". If "green" were the reference level, then we would add binary variables "colorRed" and "colorBlue" to a linear regression problem. All red examples would have colorRed=1 and colorBlue=0. All blue examples would have colorRed=0 and colorBlue=1. All green examples would have colorRed=0 and colorBlue=0.   
     
   Now, consider the variable "raceeth" in our problem, which has levels "American Indian/Alaska Native", "Asian", "Black", "Hispanic", "More than one race", "Native Hawaiian/Other Pacific Islander", and "White".   
     
   If we select White as the reference level, which binary variables will be included in the regression model? (Select all that apply.)
3. Consider again adding our unordered Category race to the regression model with reference level "White".   
     
   For a student who is Asian, which binary variables would be set to 0? All remaining variables will be set to 1.
4. For a student who is white, which binary variables would be set to 0? All remaining variables will be set to 1.
5. Because the race variable takes on text values, we will have to create Dummy Variables before we run Modeling on the data.   
     
   Build a linear regression model (call it lmScore) using the training set to predict readingScore using all the remaining variables. (Including the dummies). Remember to omit one of the dummy variable ( the omitted dummy will be the reference)   
     
   What is the Multiple R-squared value of lmScore on the training set?
6. What is the training-set root-mean squared error (RMSE) of lmScore?
7. Consider two students A and B. They have all variable values the same, except that student A is in grade 11 and student B is in grade 9. What is the predicted reading score of student A minus the predicted reading score of student B?
8. What is the meaning of the coefficient associated with variable raceethAsian?
9. Based on the significance codes, which variables out of these are candidates for removal from the model? Select all that apply. (The factor variable raceeth should only be removed if none of its levels are significant.)
10. Using the "predict" function , use the lmScore model to predict the reading scores of students in pisaTest. Call this list of predictions "predTest". Do not change the variables in the model (for example, do not remove variables that we found were not significant in the previous part of this problem).   
      
    What is the range between the maximum and minimum predicted reading score on the test set?
11. What is the sum of squared errors (SSE) of lmScore on the testing set?   
    \_\_\_\_\_ (give a integer)   
      
    What is the root-mean squared error (RMSE) of lmScore on the testing set?   
    \_\_\_\_\_ (give in 2 decimals)
12. What is the predicted test score used in the baseline model? Remember to compute this value using the training set and not the test set.
13. What is the sum of squared errors of the baseline model on the testing set? HINT: We call the sum of squared errors for the baseline model the total sum of squares (SST).
14. What is the test-set R-squared value of lmScore?

# Solutions:

1. 3663
2. 483.53, 512.94
3. Raceeth, preschool, expectBachelors etc..
4. 2414, 990
5. Raceeth
6. American Indian/Alaska Native, Asian, Black, Hispanic, More than one race, Native Hawaiian/Other Pacific Islander, White
7. Indian/Alaska Native, Black, Hispanic, More than one race, Native Hawaiian/Other Pacific Islander
8. American Indian/Alaska Native, Asian, Black, Hispanic, More than one race, , Native Hawaiian/Other Pacific Islander
9. 0.3
10. 73.36
11. 59.09
12. Predicted difference in the reading score between an Asian student and a American Indian/Alaska Native student who is otherwise identica
13. …
14. 284
15. 5762082, 76.92
16. 517
17. 7802354
18. 0.26