December 17, 2018

eda (Exploratory data analysis)

for student performance in exams

# Introduction

This is an extensive Exploratory Data Analysis for Student Performance in Exams, which is a dataset downloaded from [Kaggle](https://www.kaggle.com/spscientist/students-performance-in-exams), with the powers of Python , Pandas and Seaborn.

The data comes from one file:

* StudentsPerformance.csv: Marks secured by the students in college

Example research questions:

* How effective is the test preparation course?
* Which major factors contribute to test outcomes?
* What would be the best way to improve student scores on each test?
* What patterns and interactions in the data can you find? Let me know in the comments section below.

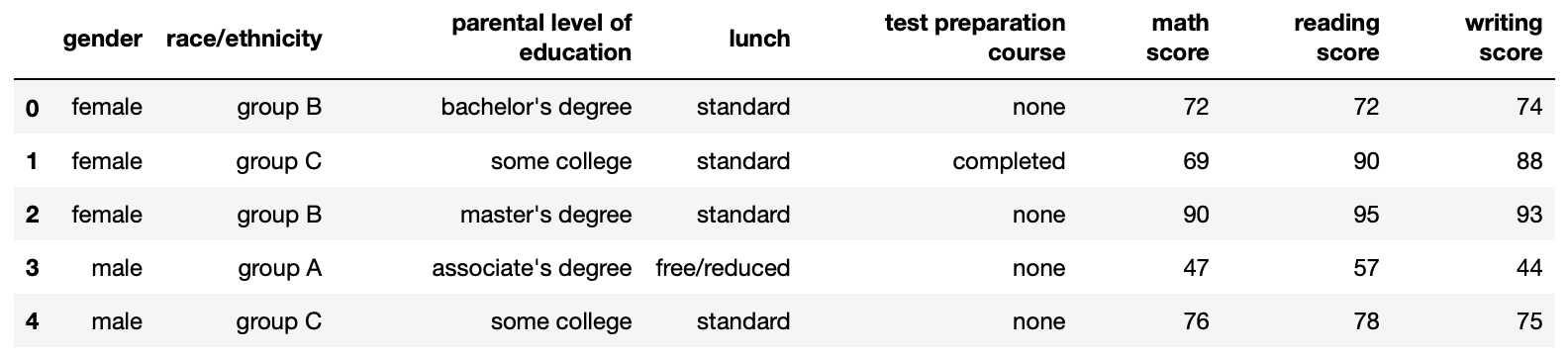
# Preparations

We load the file for general data wrangling and general data visualization with Pandas and Seaborn.

# File structure and content

In StudentsPerformance.csv, there are 1000 different students and each of them has eight features.

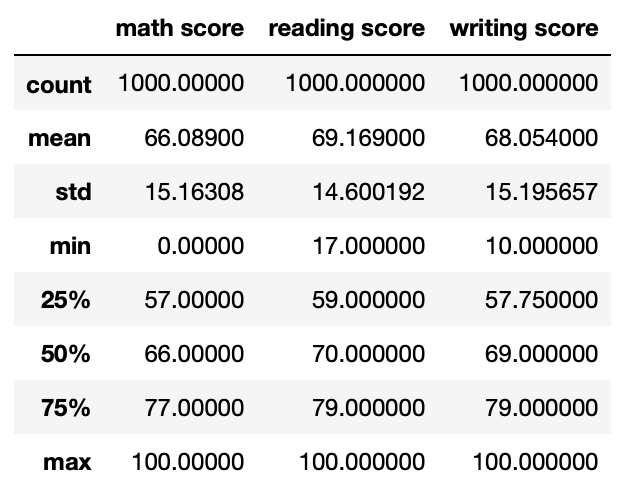
Examples of the first five data:



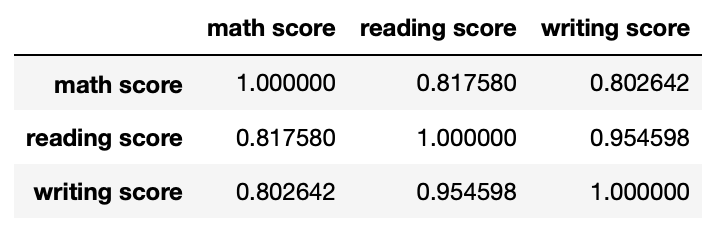
The eight features include five categorical features and three numerical features.

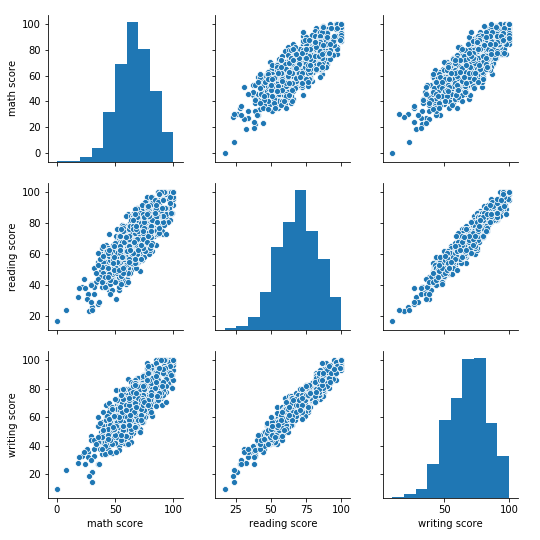
* Categorical features and their values:
* gender: ['female' 'male']
* race/ethnicity: ['group A' 'group B' 'group C' 'group D' 'group E']
* parental level of education: ["associate's degree" "bachelor's degree" 'high school' "master's degree" 'some college' 'some high school']
* lunch: ['free/reduced' 'standard']
* test preparation course: ['completed' 'none']
* Numerical features:
  + math score: 0~100
  + reading score: 0~100
  + writing score: 0~100

Take a first look at the numerical data statsistics (detailed explanation to be continued).



Correlations between numerical data:





# Indivisual feature visualization

## Gender

## There are 518 female and 482 male students in the dataset. The following table shows…

## In general, male students perform better in math, while female students excel in reading and writing exams…

## 

## Race/ethnicity

## The amount of data for each ethnic group is not balanced.

## Group A: 89 students

## Group B: 190 students

## Group C: 319 students

## Group D: 262 students

## Group E: 140 students

## 

## We observe apparent trend between median scores and the student’s ethnicity.

## 

## Parental level of education

## 

## Lunch

## Median scores grouped by “lunch”:

## 

## 

## Test preparation course

## 

# feature relations

## Race/ethnicity vs. Household income

## 