Diamonds Data Exploration

Dataset

>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 the creative problem solving and 18 in an assessment of financial literacy.

Summary of Findings

In the exploration, I found that there is a strong and rather linear relationship between the academic performance of the students in all three subjects (math, science, and reading) and the social status of the family and the country they are living in.

Since almost none of the plots in my investigation had outliers, no transformation was needed.

Almost whatever I found out by exploring the dataset was exactly what I had expected except for two issues.

I have depicted the relationship between each feature (such as occupation, education, family structure, preschooling, etc.) with the average academic performance of each student. Then, I tried to prove that we can consider the social status index of the families as an indicator of all those main factors (occupation, education, family structure) and, thus, comparing the performance with an only the social status index would be sufficient.

As I expected, the higher the social status index the better the performance of the student.

Two unexpected interesting issues were about countries with rather low social status index but high performances like China and VietNam. The other interesting thing I found was the great impact of preschooling on learning the language better even for adolescents.

Key Insights for Presentation

For the presentation, I focus on just the influence of the four main features on students' academic performance.

I start with the parents' education level, followed by the parent's job earnings, then plot the relationship between family structure and students' performance.

I've made sure to use different color palettes for each quality variable to make sure it is clear that they're different between plots.