**Final description**

**Data Analysis, INF 323**

Use this dataset <https://archive.ics.uci.edu/dataset/320/student+performance> (student\_por.csv)

1. What is the distribution of students' ages in the dataset?
2. How many students belong to each school (GP or MS)?
3. What is the gender distribution of students?
4. What is the distribution of students' travel times to school?
5. How do the first period grades (G1) vary with study time (studytime)?
6. Is there a correlation between students' weekly study time (studytime) and their final grades (G3)?
7. How do students' absences (absences) relate to their final grades (G3)?
8. Are there differences in final grades (G3) between students living in urban (U) and rural (R) areas?
9. What is the relationship between family size (famsize) and the quality of family relationships (famrel)?
10. Does the presence of romantic relationships (romantic) affect students' alcohol consumption (Dalc and Walc)?
11. How does the mother's education level (Medu) correlate with the father's education level (Fedu)?
12. Are there differences in students' final grades (G3) based on their parents' cohabitation status (Pstatus)?
13. Create a histogram of students' final grades (G3) to visualize the grade distribution.
14. Generate a scatter plot to show the relationship between students' age and their first period grades (G1).
15. Create a bar chart to compare the average final grades (G3) of students with and without extra educational support (schoolsup).
16. How do final grades (G3) in the math course compare to final grades in the Portuguese course for students who belong to both datasets?
17. Create a side-by-side box plot to compare the distribution of final grades (G3) between the math and Portuguese courses.
18. Is there a significant difference in the average final grades (G3) between male and female students? Conduct a two-sample t-test and visualize the results.
19. Can you create a new variable that categorizes students into age groups (e.g., 15-17, 18-20, 21-22)? How does this grouping affect the analysis of other variables, such as study time or final math grades (G3)?
20. Apply a mathematical transformation, such as logarithm or square root, to the number of school absences (absences). How does this transformation impact the distribution of absences and its relationship with final math grades (G3)?
21. Create a new binary variable that indicates whether a student has above-average weekly study time (studytime). How does this modified variable relate to the final math grades (G3)?
22. Apply feature scaling (e.g., Min-Max scaling or standardization) to numeric variables like age, absences, and study time. How does this scaling affect the relationships between these variables and math grades (G3)?
23. Convert the categorical variables (e.g., "reason" and "Mjob") into numeric format using label encoding or one-hot encoding. How does this transformation make the data suitable for analysis, and what insights can you gain?
24. Combine multiple variables (e.g., mother's education and father's education) to create a composite metric representing the overall parental education level. How does this new metric correlate with students' final math grades (G3)?
25. Calculate the average weekly study time for students from urban (address = 'U') and rural (address = 'R') areas. Are there differences in study time between these two groups?
26. For ordinal variables like the quality of family relationships (famrel), assign meaningful labels to the numerical values (e.g., 'very bad,' 'bad,' 'neutral,' 'good,' 'excellent'). How does this transformation make the data more interpretable?
27. Apply custom aggregation functions to summarize the data, such as calculating the range of ages within different schools or determining the percentage of students with Internet access (internet = 'yes') by gender. What insights do these custom aggregations provide?
28. If relevant, consider applying date-related functions to variables, such as determining the day of the week for which students have the most absences. How does this transformation reveal patterns related to attendance?
29. Calculate the median number of school absences (absences) for students with and without extra educational support (schoolsup).
30. Calculate the percentage of students who want to take higher education (higher) for each level of father's education (Fedu).
31. Calculate the correlation between travel time (traveltime) and final grades (G3).
32. Calculate the weighted average of final grades (G3) using study time (studytime) as weights.
33. Find the student with the highest weekend alcohol consumption (Walc).
34. Replace missing values in the 'guardian' column with 'unknown'.
35. Fill missing values in the 'romantic' column with the most common value.
36. Create a pivot table to find the maximum and minimum study times for each 'reason' for choosing the school.
37. Check if any student has 'teacher' as both mother's and father's job.
38. Replace 'at\_home' in the 'Mjob' and 'Fjob' columns with 'homemaker'.
39. Melt the dataset to convert the 'Mjob' and 'Fjob' columns into a single column 'ParentJob' while preserving other columns
40. Create a custom function that assigns a letter grade (A, B, C, D, or F) based on the final grade (G3) and apply it to a new column.
41. Create a time series plot showing the trend in weekly study time (studytime) over time for a specific student.
42. Create a new DataFrame that combines data from the Math and Portuguese courses for students who appear in both datasets.
43. Calculate and list the top 5 students with the highest final grades (G3) in the 'GP' school.
44. Create a bar chart showing the distribution of students' travel times (traveltime) in the 'MS' school.
45. Compute the mean age of students who have extra-curricular activities (activities) and those who don't.
46. Group the data by 'sex' and 'address,' and find the median number of school absences for each group.
47. Calculate the percentage of students who receive extra educational support (schoolsup) in the 'GP' school.
48. Create a scatter plot of 'G1' versus 'G3' for male students from the 'MS' school.
49. Identify students with a unique combination of 'Mjob' and 'Fjob' that appears only once in the dataset.
50. Calculate the average final grade (G3) for students from 'GP' and 'MS' schools in each 'studytime' category.

**Deadline for submission:**

* **10 December 23:59**
* **Submit here** [**https://docs.google.com/forms/d/e/1FAIpQLSfY7qnL9l-qYE1AyRKHlq43-VwMQe59DzxuNQ6O3AQaIYihlg/viewform?usp=sf\_link**](https://docs.google.com/forms/d/e/1FAIpQLSfY7qnL9l-qYE1AyRKHlq43-VwMQe59DzxuNQ6O3AQaIYihlg/viewform?usp=sf_link)
* **Tutorial on how to submit** [**https://docs.google.com/document/d/1bgosZNzRVFHVdi5jz9BfIklljqf\_5iKjT7aGbjrrNdo/edit?usp=sharing**](https://docs.google.com/document/d/1bgosZNzRVFHVdi5jz9BfIklljqf_5iKjT7aGbjrrNdo/edit?usp=sharing)

**Notes:**

* It is individual based work
* Plagiarism is strictly disallowed
* Report should have at least 20 pages with 12pt times new roman and single line spacing
* Not signing attendance list in time would result in retake

**Report.pdf structure:**

Enumerate and for each question, write explanation of answer and screenshot as image.

Example:

1. Question 1.  
   This code does X. The result returns Y.

Screenshot (size is usually ¼ of A4)

1. Question 2 …

**Code.pdf structure:**

Pdf format of jupyter notebook where all snippets of code are enumerated as sections.