DEPARTMENT OF
DATA SCIENCE

DS311
Exploratory Data Analysis

Laboratory Activity 3-4

Section 3: Lab Activity

In this lab, you'll apply what we've covered to the student performance data. The dataset we use is from a study of secondary school students from two Portuguese schools, with attributes on student background and grades. We will use the "Math" class subset (395 students) for these exercises.

1. Data Understanding: Load the student performance dataset into R (you can use the code provided in the examples). Examine its structure using functions like `dim()`, `str()`, or `summary()`. How many observations and variables are there? Identify which variables are numerical and which are categorical. (Hint: Variables like `school`, `sex`, etc. are categorical; `age` is numeric; some, like `studytime`, are ordered factors coded as numeric.)

2. Exploring Relationships: Pick at least five numeric variables and compute a correlation matrix for them (for example, you might include the three grade columns `G1, G2, G3` and a couple of others like `studytime`, `absences`, `failures`). Which pair of variables has the highest correlation? Which pair has a strong negative correlation (if any)? Describe what those correlations imply in the context of student performance.

3. Visualizing Multivariate Patterns: Create a scatter plot to visualize the relationship between two interesting variables from the dataset. A good choice is **first period grade (`G1`) vs final grade (`G3`)** as we did above. Plot these using `ggplot2::geom_point()`. Do you notice any pattern or outliers? Now improve the plot by addressing overplotting: for example, use `geom_jitter()` or set `alpha` transparency as demonstrated. How does this help in seeing the data distribution? Briefly describe the trend you observe and any deviations.

(Optional extension: Color the points by a categorical variable such as `sex` (gender) or `schoolsup` (extra school support) to see if the relationship differs by subgroup.)

4. Principal Component Analysis: Perform PCA on a set of numerical variables related to student performance. A suggested set: `c("G1","G2","G3","failures","absences","Medu","Fedu")` (as we used

in the example). Ensure you scale the variables (`prcomp(scale.=TRUE)`). How many principal components have eigenvalues greater than 1? Look at the proportion of variance explained by the first two PCs. Now inspect the PCA loadings (use `pca_result$rotation` or `print(pca_result)`). Based on the loadings, what does PC1 seem to represent? What about PC2? (In other words, which variables heavily influence PC1 vs PC2?)

5. Factor Analysis: Using the same set of variables as in #4, conduct an exploratory factor analysis. Try extracting 2 factors with `factanal(..., factors=2, rotation="varimax")`. Examine the factor loadings. Which variables load strongly on Factor1, and which on Factor2? Name each factor in plain language (e.g., "Factor1: _____ factor, Factor2: _____ factor"). How well does this factor solution make sense, and how does it compare to the PCA results from #4?

Write down your findings and interpretations for each step.