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.

Load dataset

```
data <- read.csv("student-mat.csv", sep=";")
head(data,3)
```

```
##   school sex age address famsize Pstatus Medu Fedu    Mjob    Fjob reason
## 1     GP   F  18       U     GT3       A    4    4 at_home teacher course
## 2     GP   F  17       U     GT3       T    1    1 at_home   other course
## 3     GP   F  15       U     LE3       T    1    1 at_home   other  other
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0       yes     no   no         no
## 2   father          1         2        0        no    yes   no         no
## 3   mother          1         2        3       yes     no  yes         no
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1     yes    yes       no       no      4        3     4    1    1      3
## 2      no    yes      yes       no      5        3     3    1    1      3
## 3     yes    yes      yes       no      4        3     2    2    3      3
##   absences G1 G2 G3
## 1        6  5  6  6
## 2        4  5  5  6
## 3       10  7  8 10
```

1. Data Understanding:

```
dim(data)
```

```
## [1] 395  33
```

- **395 observations (rows)**
- **33 variables(columns)**

```
str(data)
```

```
## 'data.frame':    395 obs. of  33 variables:
##  $ school    : chr  "GP" "GP" "GP" "GP" ...
```

```
##  $ sex       : chr  "F" "F" "F" "F" ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : chr  "U" "U" "U" "U" ...
##  $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus   : chr  "A" "T" "T" "T" ...
##  $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : chr  "at_home" "at_home" "at_home" "health" ...
##  $ Fjob      : chr  "teacher" "other" "other" "services" ...
##  $ reason    : chr  "course" "course" "other" "home" ...
##  $ guardian  : chr  "mother" "father" "mother" "mother" ...
##  $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
##  $ schoolsup : chr  "yes" "no" "yes" "no" ...
##  $ famsup    : chr  "no" "yes" "no" "yes" ...
##  $ paid      : chr  "no" "no" "yes" "yes" ...
##  $ activities: chr  "no" "no" "no" "yes" ...
##  $ nursery   : chr  "yes" "no" "yes" "yes" ...
##  $ higher    : chr  "yes" "yes" "yes" "yes" ...
##  $ internet  : chr  "no" "yes" "yes" "yes" ...
##  $ romantic  : chr  "no" "no" "no" "yes" ...
##  $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : int  6 4 10 2 4 10 0 6 0 0 ...
##  $ G1        : int  5 5 7 15 6 15 12 6 16 14 ...
##  $ G2        : int  6 5 8 14 10 15 12 5 18 15 ...
##  $ G3        : int  6 6 10 15 10 15 11 6 19 15 ...
```

Categorical Variables: `school`, `sex`, `address`, `famsize`, `Pstatus`, `Mjob`, `Fjob`, `reason`, `guardian`, `schoolsup`, `famsup`, `paid`, `activities`, `nursery`, `higher`, `internet`, `romantic`. There are total of **17 categorical variables**.

Numeric variables: `age`, `Medu`, `Fedu`, `traveltime`, `studytime`, `failures`, `famrel`, `freetime`, `goout`, `Dalc`, `Walc`, `health`, `absences`, `G1`, `G2`, `G3`. There are **16 numerical variables**.

2. Exploring Relationships:

```
num_vars <- data[, c("G1", "G2", "G3", "studytime", "absences", "failures")]
cor_matrix <- cor(num_vars, use="complete.obs")
round(cor_matrix, 2)
```

```
##              G1    G2    G3 studytime absences failures
## G1         1.00  0.85  0.80      0.16    -0.03    -0.35
## G2         0.85  1.00  0.90      0.14    -0.03    -0.36
## G3         0.80  0.90  1.00      0.10     0.03    -0.36
## studytime  0.16  0.14  0.10      1.00    -0.06    -0.17
```

```
## absences  -0.03 -0.03  0.03     -0.06     1.00     0.06
## failures  -0.35 -0.36 -0.36     -0.17     0.06     1.00
```

Highest positive correlation: `G2` and `G3` which has a correlation of 0.90.

- Interpretation: Students who perform well in the second period also tend to perform well in the final grade.

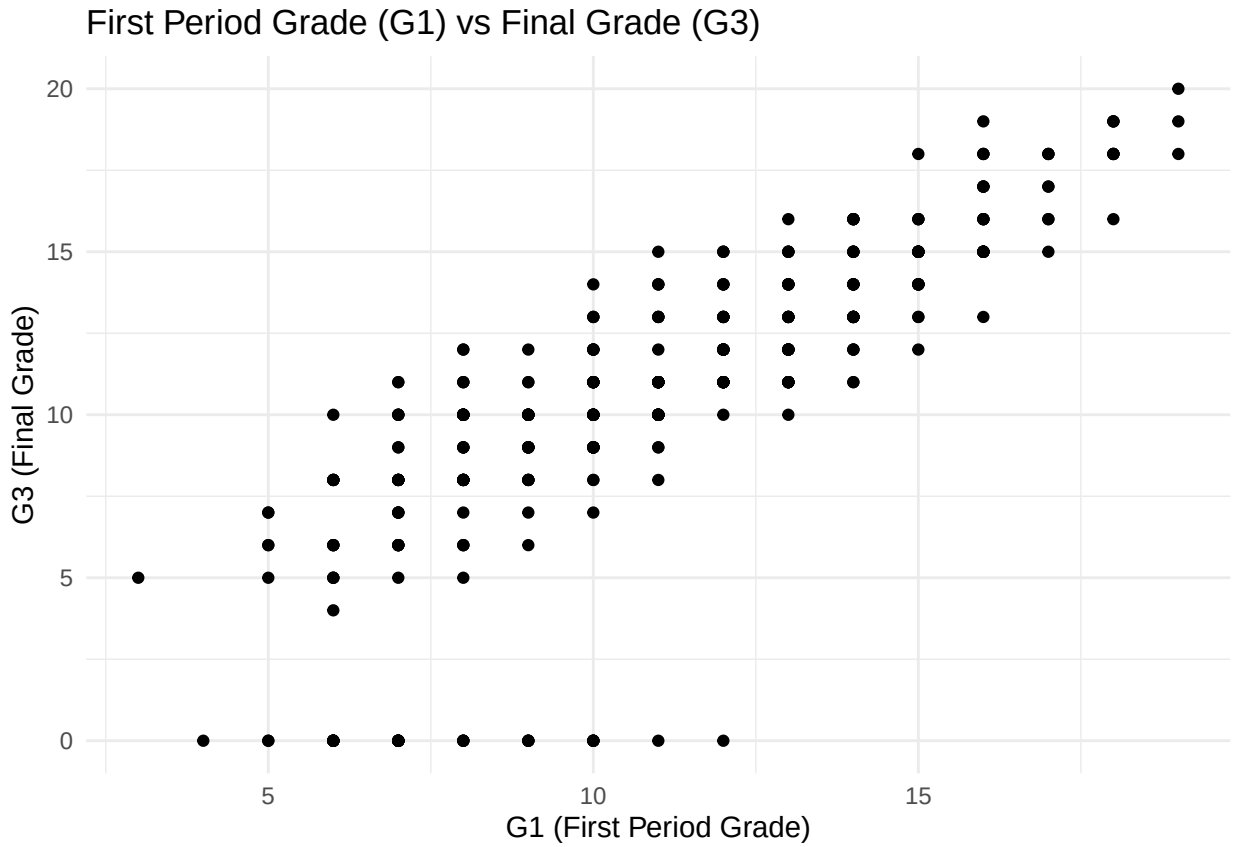
Strongest negative correlation: `failures` and `G1` / `G2` / `G3` which has correlation in between -0.35 to -0.36.

- Interpretation: Students with more previous class failures tend to score lower in all three grading periods.

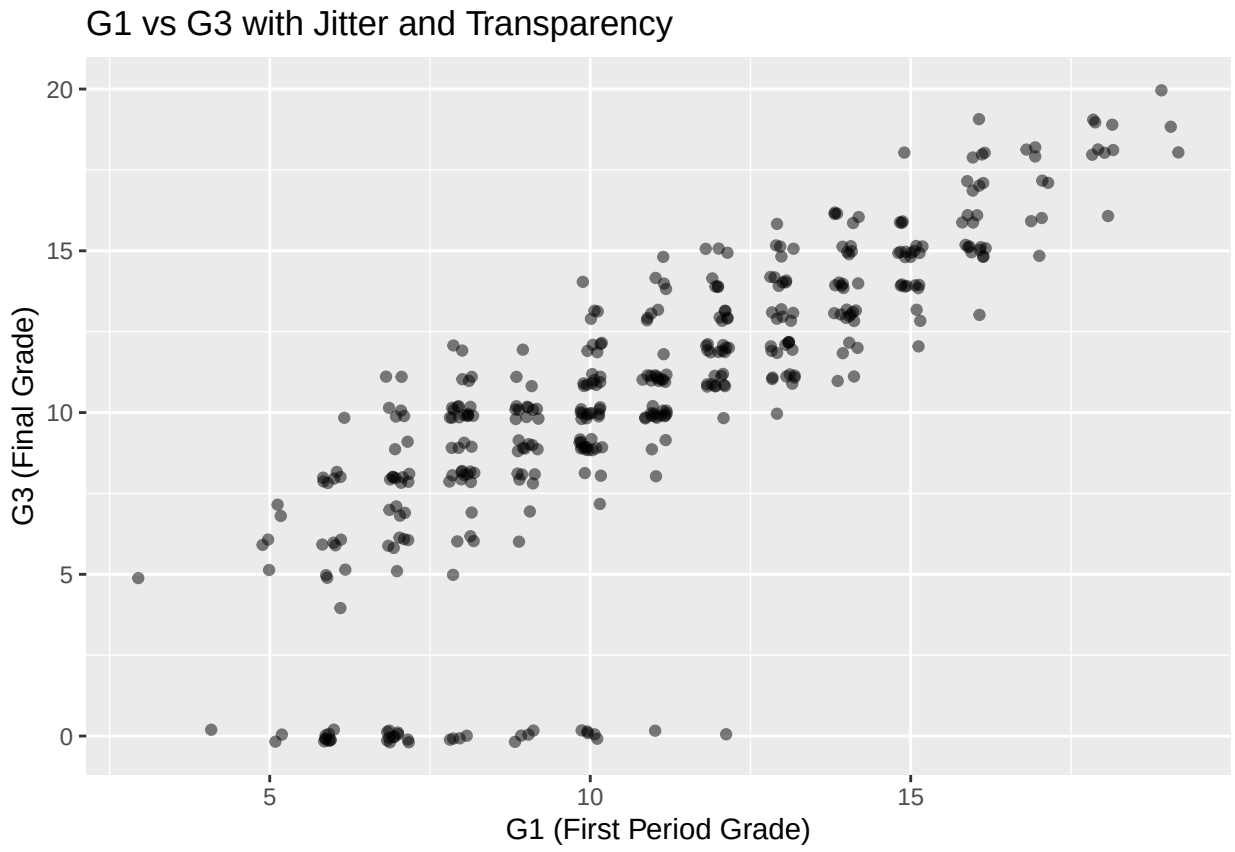
Other observations:

- `studytime` has a weak positive relationship with grades.
- `absences` shows very little or no relationship with grades in this dataset.

3. Visualizing Multivariate Patterns:



- You'll see a strong positive trend: as `G1` increases, `G3` also tends to increase.



- `geom_jitter()` adds small random noise so overlapping points separate slightly.
- This makes **dense clusters and outliers visible**.

Observations:

- **Clear positive relationship:** Higher G1 grades are associated with higher G3 grades.
- **Some outliers:** A few students have low G3 despite high G1, suggesting they dropped performance over time.
- Most points cluster along a diagonal line (roughly G1 to G3), showing consistency in student performance.

4. Principal Component Analysis:

```
vars <- c("G1","G2","G3","failures","absences","Medu","Fedu")
pca_vars <- data[,vars]
pca_result <- prcomp(pca_vars, scale. = TRUE)
summary(pca_result)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.7645 1.2154 1.0094 0.8551 0.61038 0.44885 0.29212
## Proportion of Variance 0.4448 0.2110 0.1456 0.1045 0.05322 0.02878 0.01219
## Cumulative Proportion  0.4448 0.6558 0.8013 0.9058 0.95903 0.98781 1.00000
```

- **3 principal components have eigenvalues > 1 (PC1, PC2, PC3).**

Inspect PCA loadings

```
pca_result$rotation
```

```
##                   PC1        PC2         PC3         PC4         PC5
## G1         0.501062200  0.2149437  0.04243443 -0.13378178 -0.12114035
## G2         0.517805645  0.2402556  0.05972029 -0.14514997  0.01090063
## G3         0.509005737  0.2292847  0.12482182 -0.09608585  0.04710142
## failures -0.315043948  0.1135754  0.28479071 -0.89732287 -0.03492824
## absences -0.001386686 -0.1627197  0.93976809  0.28051607 -0.09039359
## Medu       0.258573793 -0.6251826  0.01457684 -0.19340116  0.70590786
## Fedu       0.235822237 -0.6428644 -0.12070638 -0.17469923 -0.68941335
##                   PC6           PC7
## G1        -0.791064588  0.2063478999
## G2         0.216892772 -0.7761174024
## G3         0.555386276  0.5940170271
## failures   0.007976032  0.0164394850
## absences -0.039706020 -0.0438120871
## Medu      -0.079677682  0.0039212664
## Fedu       0.103593970  0.0009805477
```

- `PC1` represents **academic performance / achievement**
  - High PC1 score = high grades, few failures
- `PC2` represents **parental education / family background**
  - High negative PC2 scores = highly educated parents
  - High positive PC2 scores = lower parental education

There are **three** principal components with eigenvalues greater than 1. The first two components together explain about 65% of the total variance. `PC1` is mainly influenced by grades (`G1, G2, G3`) and negatively by `failures`, representing overall academic performance. `PC2` is mainly associated with `parental education` and `absences`, reflecting a family background and attendance pattern.

5. Factor Analysis:

```
fa_result <- factanal(data[, vars], factors = 2, rotation = "varimax")
fa_result$loadings
```

```
##
## Loadings:
##          Factor1 Factor2
## G1         0.870
## G2         0.979
## G3         0.924
## failures -0.371  -0.224
## absences
```

```
## Medu      0.220   0.707
## Fedu      0.168   0.829
##
##                Factor1 Factor2
## SS loadings      2.785   1.244
## Proportion Var   0.398   0.178
## Cumulative Var   0.398   0.576
```

Factor 1:

- **Academic performance** loads the strongest.
- **Strong positive loadings:** `G1`, `G2`, `G3`

Factor 2:

- **Parental Education** loads the strongest.
- **Strong positive loadings:** `Medu`, `Fedu`

The factor analysis result matches the **first two PCA components almost exactly**, except:

- `absences` **did not load strongly** on either factor, so it is not represented in the 2-factor solution.
- PCA showed absences as its own separate third component.

Factor 1 represents academic performance, with high loadings from `G1, G2, and G3` and a negative loading from `failures`. **Factor 2** represents parental education, with strong loadings from `Medu` and `Fedu.` This two-factor solution makes **clear conceptual sense** and aligns closely with the PCA structure, except that it does not include the absences dimension that PCA identified separately.