

# Student Performance Dataset Documentation

2025-12-27

## Team

Melisa Cihan, Hrusheekesh Sawarkar, Kezia Fernandes, Raju Ahmed

## 1. Introduction

This document provides a short documentation on the student performance dataset used for the project. The dataset was obtained from Machine Learning Repository:

**Source:** *Retail Store Sales (Dirty) – For Data Cleaning* URL: <https://archive.ics.uci.edu/dataset/320/student+performance>

The goal of this documentation is to summarize the dataset, describe its variables and types and provide a reproducible R script used to load the data.

## 2. Dataset Description

The dataset contains student performance data in secondary education of two Portuguese schools. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). However, we will only analyze and focus on the dataset for Mathematics (mat). Since this dataset contains too many variables we will only consider a subset of these variables, which are displayed in the following.

```
# required packages
library(readr)
library(dplyr)

# Use read_csv2() for semicolon-separated values
data <- read_csv2("/Users/melisacihan/Desktop/Statistical Computing/StatsProj/Student_Performance/stude

## i Using ',',.'" as decimal and ',.' as grouping mark. Use 'read_delim()' for more control.

## # Rows: 395 Columns: 33
## -- Column specification -----
## Delimiter: ","
## chr (17): school, sex, address, famsize, Pstatus, Mjob, Fjob, reason, guardi...
## dbl (16): age, Medu, Fedu, travelttime, studytime, failures, famrel, freetime...
##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.
```

### 3. Summary of the Dataset

#### 3.1 Number of Rows and Columns

```
dim(data)  
  
## [1] 395 33
```

#### 3.2 Summary of Variables

```
# Select only the variables of interest for better readability  
data_selected <- data %>%  
  select(sex, age, Medu, traveltime, studytime, failures, paid,  
         activities, higher, internet, famrel, absences, G3)  
  
summary(data_selected)  
  
##      sex           age          Medu        traveltime  
##  Length:395      Min.   :15.0    Min.   :0.000  Min.   :1.000  
##  Class :character 1st Qu.:16.0   1st Qu.:2.000  1st Qu.:1.000  
##  Mode  :character Median :17.0   Median :3.000   Median :1.000  
##                Mean   :16.7   Mean   :2.749   Mean   :1.448  
##                3rd Qu.:18.0   3rd Qu.:4.000   3rd Qu.:2.000  
##                Max.   :22.0   Max.   :4.000   Max.   :4.000  
##      studytime     failures       paid        activities  
##  Min.   :1.000   Min.   :0.0000  Length:395   Length:395  
##  1st Qu.:1.000  1st Qu.:0.0000  Class :character  Class :character  
##  Median :2.000  Median :0.0000  Mode  :character  Mode  :character  
##  Mean   :2.035  Mean   :0.3342  
##  3rd Qu.:2.000  3rd Qu.:0.0000  
##  Max.   :4.000  Max.   :3.0000  
##      higher        internet       famrel       absences  
##  Length:395      Length:395      Min.   :1.000  Min.   : 0.000  
##  Class :character Class :character  1st Qu.:4.000  1st Qu.: 0.000  
##  Mode  :character Mode  :character  Median :4.000   Median : 4.000  
##                Mean   :3.944   Mean   : 5.709  
##                3rd Qu.:5.000   3rd Qu.: 8.000  
##                Max.   :5.000   Max.   :75.000  
##      G3  
##  Min.   : 0.00  
##  1st Qu.: 8.00  
##  Median :11.00  
##  Mean   :10.42  
##  3rd Qu.:14.00  
##  Max.   :20.00
```

dimension of selected subset

```
dim(data_selected)
```

```
## [1] 395 13
```

### 3.3 Data Types of Each Variable

```
glimpse(data_selected)
```

```
## Rows: 395
## Columns: 13
## $ sex      <chr> "F", "F", "F", "F", "M", "M", "M", "M", "F", "F", ~
## $ age       <dbl> 18, 17, 15, 15, 16, 16, 16, 17, 15, 15, 15, 15, 15, 15, ~
## $ Medu      <dbl> 4, 1, 1, 4, 3, 4, 2, 4, 3, 3, 4, 2, 4, 4, 2, 4, 4, 3, 3, 4, ~
## $ traveltim <dbl> 2, 1, 1, 1, 1, 1, 2, 1, 1, 3, 1, 2, 1, 1, 3, 1, 1, 1, ~
## $ studytim  <dbl> 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 3, 1, 2, 3, 1, 3, 2, 1, 1, ~
## $ failures   <dbl> 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, ~
## $ paid        <chr> "no", "no", "yes", "yes", "yes", "yes", "no", "no", "yes", ~
## $ activities  <chr> "no", "no", "no", "yes", "no", "yes", "no", "no", "no", "ye~
## $ higher     <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "ye~
## $ internet    <chr> "no", "yes", "yes", "yes", "no", "yes", "yes", "no", "yes", ~
## $ famrel      <dbl> 4, 5, 4, 3, 4, 5, 4, 4, 5, 3, 5, 4, 5, 4, 4, 3, 5, 5, 3, ~
## $ absences    <dbl> 6, 4, 10, 2, 4, 10, 0, 6, 0, 0, 4, 2, 2, 0, 4, 6, 4, 16, ~
## $ G3          <dbl> 6, 6, 10, 15, 10, 15, 11, 6, 19, 15, 9, 12, 14, 11, 16, 14, ~
```

## 4. Variable Description

Variable Name	Data Type	Scale of Measure	Description
sex	Binary	Nominal	Student's sex (F - female, M - male)
age	Numeric	Numeric discrete	Student's age (ranges from 15 to 22 years)
Medu	Categorical	Ordinal	Mother's education level (0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education, 4 - higher education)
traveltim	Categorical	Ordinal	Home to school travel time (1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, 4 - >1 hour)
studytim	Categorical	Ordinal	Weekly study time (1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, 4 - >10 hours)

Variable Name	Data Type	Scale of Measure	Description
<b>failures</b>	Numeric	Discrete	Number of past class failures (numeric value, where 4 represents 3 or more failures)
<b>paid</b>	Binary	Nominal	Extra paid classes within the course subject (yes or no)
<b>activities</b>	Binary	Nominal	Participation in extra-curricular activities (yes or no)
<b>higher</b>	Binary	Nominal	Wants to pursue higher education (yes or no)
<b>internet</b>	Binary	Nominal	Internet access at home (yes or no)
<b>famrel</b>	Categorical	Ordinal	Quality of family relationships (scale from 1 - very bad to 5 - excellent)
<b>absences</b>	Numeric	Discrete	Number of school absences (ranges from 0 to 93)
<b>G3</b>	Numeric	Continuous*	Final grade (ranges from 0 to 20)

## 5. Example Data (First Few Observations)

```
head(data_selected, 5)

## # A tibble: 5 x 13
##   sex     age  Medu traveltime studytime failures paid  activities higher
##   <chr>    <dbl> <dbl>      <dbl>      <dbl>    <dbl> <chr> <chr>       <chr>
## 1 F        18     4          2          2        0 no    no      yes
## 2 F        17     1          1          2        0 no    no      yes
## 3 F        15     1          1          2        3 yes   no      yes
## 4 F        15     4          1          3        0 yes   yes     yes
## 5 F        16     3          1          2        0 yes   no      yes
## # i 4 more variables: internet <chr>, famrel <dbl>, absences <dbl>, G3 <dbl>
```

## 6. Conclusion

The dataset's focus on secondary education enables intuitive analysis of socioeconomic drivers behind academic achievement. Furthermore, the data structure supports meaningful bivariate analysis and linear regression modeling to predict student performance. As already mentioned, since this dataset contains too many variables we will only consider a subset of these variables.

## 7. Limitation

This dataset does not contain any continuous variables. Even variables like traveltimes are displayed as categorical variables, since this is generally the case in social sciences. However, for some of the analyses we would like to use the variable G3 (final grade, ranges from 0-20) as it were a continuous variable, even though it displays discrete integer values.