Lily Nguyen

Professor John Ladd

Introduction to Data Analytics 101

**Final Project Codebook and Documentation**

**#General Information about the data:**

* Data name: Student Alcohol Consumption
* Data source: Kaggle (<https://www.kaggle.com/datasets>)
* Original data collectors: Paulo Cortez (University of Minho, Portugal)
* Data size: 41.98 kB

**#Special permissions:**

The data is freely available (CC0: Public Domain Dedication)

**#Explanation:**

The dataset provides us with information to analyze the relationship between the alcohol consumption and the students’ performance on their life, and how the amount of alcohol consumption affects their grades. My goal is to prove that the relationships in students’ lives play an important role in how much they drink; therefore, the amount of alcohol they consumed can alter their final grades in school greatly. The whole dataset consists of thirty-three variables, but I will be focusing on nine variables in this project: *G3* – final grade, *Dalc* – workday alcohol consumption, *Walc* – weekend alcohol consumption, *failures* – past class failures, *famrel* – quality of family relationships, *famsup* – family educational support, *romantic* – with a romantic relationship, *goout* – going out with friends, and *health* – current health status.

I have posed five questions in order to do a full analysis on this thesis. First of all, we are going to investigate the relationship between the family relationships & support and the amount of alcohol students consumed. We will make a new variable representing the total amount of alcohol consumed and use it along with *famrel* and *famsup*. Secondly, we are going to look into the connection between how much students drink and their relationships with friends and romantic partners. Variables *romantic*, *G3* will be used along with the total alcohol consumption. Next, we will compare the grades of students who drink on workdays and those who do not by using the variables *Dalc* and *G3*. After that, we are going to examine whether this thesis is true or not: “the more students drink, the worse their health get, which will later lead to lower final grades”. The final question is to study the relationship between the number of classes students failed and their total amount of alcohol they consumed.

*“student-mat.csv”*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Column name** | **Variable definition** | **Units** | **Data type** | **Variable Codes and definitions** | **Missing value codes** |
| G3 | Represents the final grade (numeric: from 0 to 20) | No units needed | integer | NA | No missing values |
| Dalc | Represents the amount of workday alcohol consumption (numeric: from 1-‘very low’ to 5-‘very high’) | No units needed | integer | NA | No missing values |
| Walc | Represents the amount of weekend alcohol consumption (numeric: from 1-‘very low’ to 5-‘very high’) | No units needed | integer | NA | No missing values |
| failures | Represents the number of past class failures (numeric: from 0 to 3) | No units needed | integer | NA | No missing values |
| famrel | Represents the quality of family relationships (numeric: from 1-‘very bad’ to 5-‘excellent’) | No units needed | integer | NA | No missing values |
| health | Represents the current health status (numeric: from 1-‘very bad’ to 5-‘very good’) | No units needed | integer | NA | No missing values |
| famsup | Represents whether the student receives family educational support or not (binary: yes/ no) | No units needed | character | NA | No missing values |
| romantic | Represents whether the students is in a romantic relationship or not (binary: yes/ no) | No units needed | character | NA | No missing values |
| goout | Represents the frequency of students going out with their friends (numeric: from 1-‘very low’ to 5-‘very high’) | No units needed | integer | NA | No missing values |