Take a look at the data in your project dataset and note the data types and their values.  Then answer the following questions on 1 single-spaced page:

1. Find the five most important attributes in your dataset.  What makes them important?  What are their data types?

In our group assignment of alcohol consumption, there are two datasets of student-mat.csv (Math course) and student-por.csv (Portuguese language course), where we had currently considered one dataset of student-por.csv (Portuguese language course) and we had identified 5 important attributes for alcohol consumption from which identify the inference and conclusion of the ratio of alcohol consumption. The attributes we identified are **age, famrel, dalc, walc, health**. Let we display the content in tabular format to make it more explanatory.

|  |  |  |
| --- | --- | --- |
| S.no | Attributes | Datatypes |
| 1 | **age** - student's age | Integer(Numeric) |
| 2 | **famrel** - quality of family relationships | Integer(Categorical) |
| 3 | **Dalc** - workday alcohol consumption | Integer(Categorical) |
| 4 | **Walc** - weekend alcohol consumption | Integer(Categorical) |
| 5 | **health** - current health status | Integer(Categorical) |

Fig: Selected Attributes with Datatypes  
  
  
Why these attributes are important?  
  
Every Attribute which we had selected would be creating a greater impact on inferencing a good result where many people like government agencies , schools get benefited which looking with results, and make a precautionary steps to decrease alcohol consumption.   
  
Age- This is one the place where we can keep restriction, and identify from what age the people had started alcohol consumption, and increase the bar of age.

Dalc- On being workday how much alcohol consumption would be taken place and where irrespective of age, and where the revenue generated to the government can be identified.

famrel- Quality of a family relation would also impact a greater alcohol consumption in which a child would be taken from parents behavior.

Walc- keeping same rules of weekdays how much it is said to be for weekends.

Health- Seeing alcohol consumption rate, what can be the health status of the people can be inferred, so that we can have a good society can be build from this.

1. For these five attributes, use R to determine mean, median, mode, standard deviation, and Interquartile Range (IQR).  If the attributes are nominal, convert them into numeric format to do this analysis.  For what purpose are these measurements important?

As per the attributes, we find that there are no data value missing, but taken in consideration, if we think age value is missing, we would like numeric value to be added with median , and if it is a nominal value of education ,we would be adding a grade of graduation to 16 because as per graduation the minimum requirement is 16, so like that we can make additional amendments to it and change nominal to numeric value. These Measurements are been important for to get all more clear results with inference of data , and make a more clear output with it.  
  
  
**Source Code:**   
  
## *function for mode*

getmode <- function(v) {

uniqv <- unique(v)

uniqv[which.max(tabulate(match(v, uniqv)))]

}

## *detail of age attribute*

summary(student\_por$age)

modeage <-getmode(student\_por$age)

sd(student\_por$age)

## *detail of quality of family relationship attribute*

summary(student\_por$famrel)

modefamrel <-getmode(student\_por$famrel)

sd(student\_por$famrel)

## *detail of weekdays alcohol consumption attribute*

summary(student\_por$Dalc)

modeDalc <-getmode(student\_por$Dalc)

sd(student\_por$Dalc)

## *detail of weekend alcohol consumption attribute*

summary(student\_por$Walc)

modeWalc <-getmode(student\_por$Walc)

sd(student\_por$Walc)

## *detail of health attribute*

summary(student\_por$health)

modehealth <-getmode(student\_por$health)

sd(student\_por$health)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **S.no** | **Attribute** | **Mean** | **Median** | **Mode** | **Standard deviation** | **Interquartile Range** |
| 1 | Age | 16.74 | 17 | 17 | 1.218138 | 2 |
| 2 | Famrel | 3.931 | 4 | 4 | 0.9557169 | 1 |
| 3 | Dalc | 1.502 | 1 | 1 | 0.9248344 | 1 |
| 4 | Walc | 2.28 | 2 | 1 | 1.28438 | 2 |
| 5 | Health | 3.536 | 4 | 5 | 1.446259 | 3 |

3. If the attributes have missing or implausible values, what is going to be your strategy for fil ling them in? Include your R code and its output in this submission.

As per the attributes, we find that there are no data value missing, but taken in consideration, if we think age value is missing, we would like numeric value to be added with median, and if it is a nominal value of education, we would be adding a grade of graduation to 16 because as per graduation the minimum requirement is 16, so like that we can make additional amendments to it and change nominal to numeric value. These Measurements are been important for to get all more clear results with inference of data , and make a more clear output with it.  
  
  
**Source Code:**   
  
anyNA(student\_por)  
  
Output:  
  
[1] False