Computer Science - Core Skills

Semester 1

BSC109224 Group A

Assignment 5

Student: Lucas Madeira Maranho

Student ID: 76990

Lecturer: Bernard Joseph Roche

SUMMARY

Question 1

Before answering the question, I will describe types of data that exist. Data is divided into 2 types: quantitative(numeric) or qualitative(categorical), into a variable, but What is variable?

Variable is any characteristic, number that can be counted or measured and can assume different values. It can be called a data item too. - B.S., Anjana. (2021). Scales of Measurement in Research.

1. Analysis Data and type of Data

I. Quantitative (or numerical):

It is the process to check and analyse the number of a Data. It is used to find average of a population for example, make predictions or do some tests. Bhandari, P. (2023, June 22).

Quantitative is divided into 2 types:

i. Discrete variables:
   Quantitative discrete is integer numbers, which means that is only limited number of values, cannot be broken numbers.

ii. Continuous variables:
   Can take any value or any value between two values, it can be a broken number, for example decimal numbers.

II. Qualitative (or categorical):

Qualitative is basically words but can be images or other media. It is a numerical label arbitrary. For example, you can use your mood: 1 for happy, 2 for sad, 3 for normal. Qualitative is divided into 2 types:

i. Ordinal:
   Ordinal is followed by order or ranking.

ii. Nominal:
   It doesn't follow any order or ranking like ordinal. Consider gender, City, colours.
   365DataScience, 2024. *Data Types: The Complete Guide*. Available at: [Accessed 27 Nov. 2024].

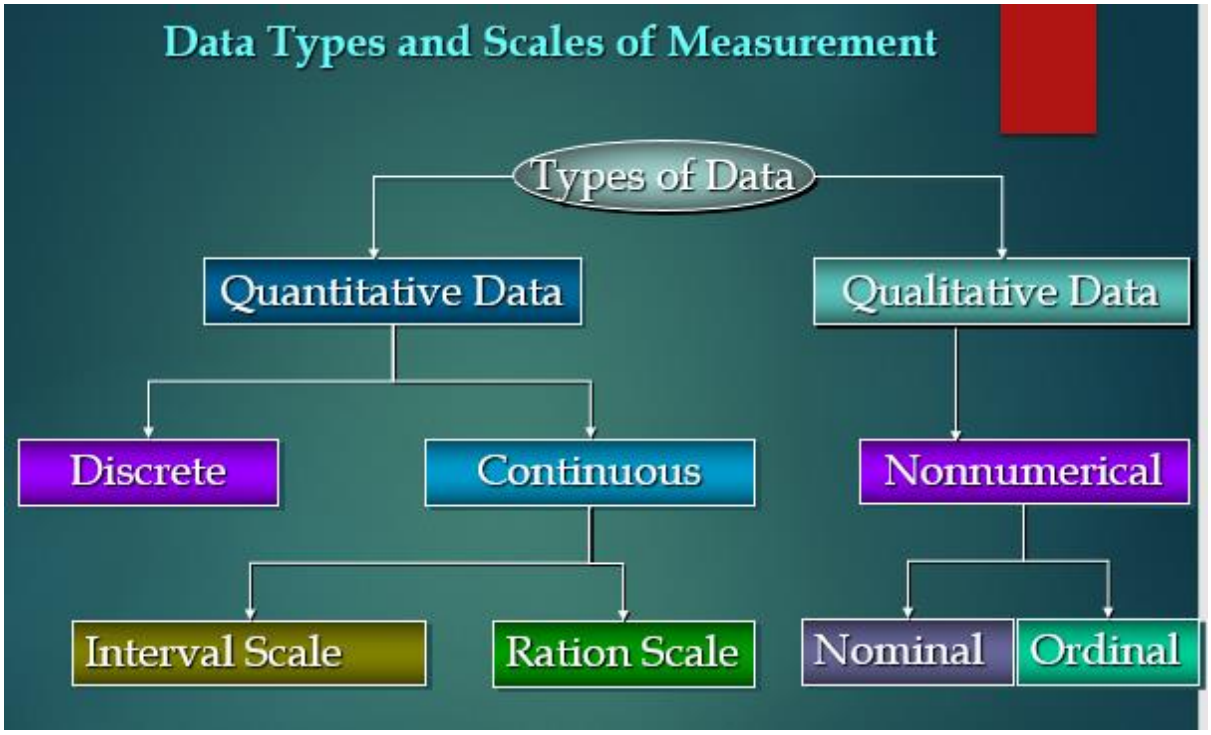

Figure 1 - Data type

2. Measurement Scale

Measurement Scale is how the variables are defined. It is organized and separated in 4 common scales of measurements: Nominal, ordinal, interval and ratio.

The properties are identify, magnitude, equal intervals and minimum value of zero.

I. Interval:
This type of data shows the difference between the variables. It can be added or subtracted but it cannot be multiplied or divided from each other.

II. Ratio:
Include all the properties, it can be nominal, it can have an identity, classified in order, broken down into exact value, like decimal numbers. It can be added, subtracted, multiplied and divided.
UNSW Online, n.d. *Types of Data: Understanding the Four Types of Data*. Available at: [Accessed 27 Nov. 2024].

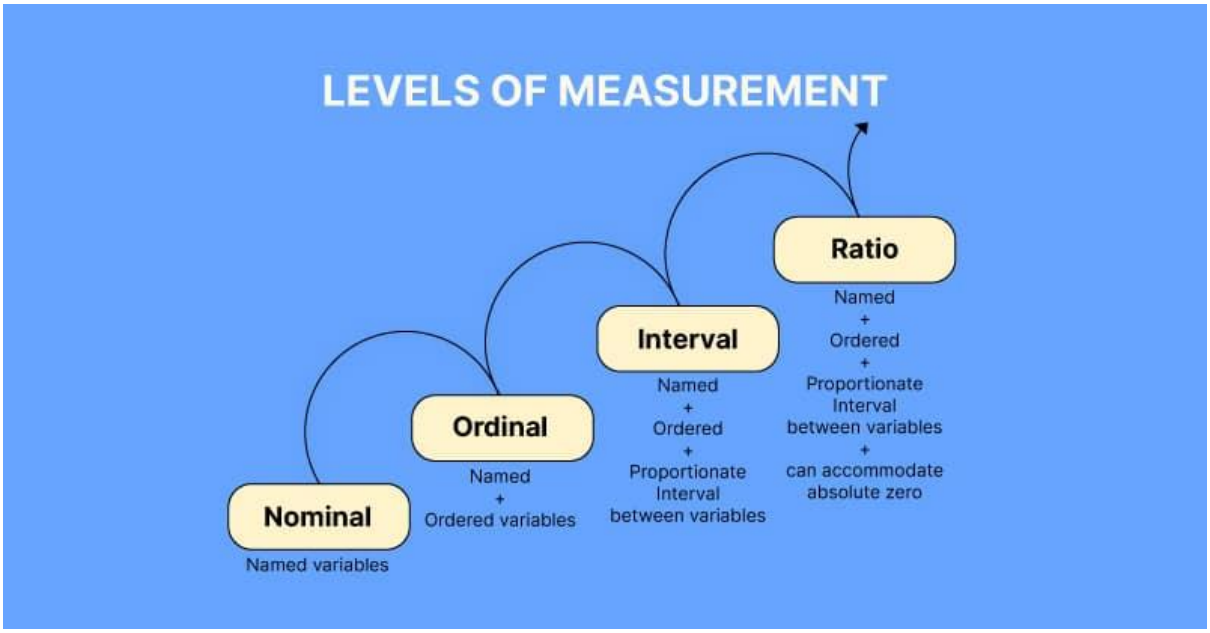

Figure 2 - Measurement levels

Dataset application:

Let's apply it in the dataset:

Column A is an identifier: it is a discrete data.

Column B is a quantitative continuous data.

Column C is biological gender, it is a qualitative Nominal variable.

Column D is height in meters, it is a quantitative continuous ratio.

Column E is how tall people are in letters. It is a qualitative ordinal data.

Column F is nationality, which is qualitative nominal.

Column G is the shoe size, and it is quantitative discrete data.

Column H is their Age, which is a quantitative continuous ratio.

Column J is eye colours, and it is a qualitative nominal data.

3. Observation and Interferences:

The dataset there are not headliners, so let's assume that the dataset can be interpreted as student, grades, gender, height, size classification, nationality, shoes size, age and eye colours, respectively.
It's important to understand that the analysis from the dataset presented was based on assumptions.

4. Population vs sample:

Population is the entire group which you want to get conclusion about, and sample is a specific group where you will collect information from. Sample is always a group less than population. (Published on May 14, 2020, by Pritha Bhandari. Revised on June 21, 2023.)

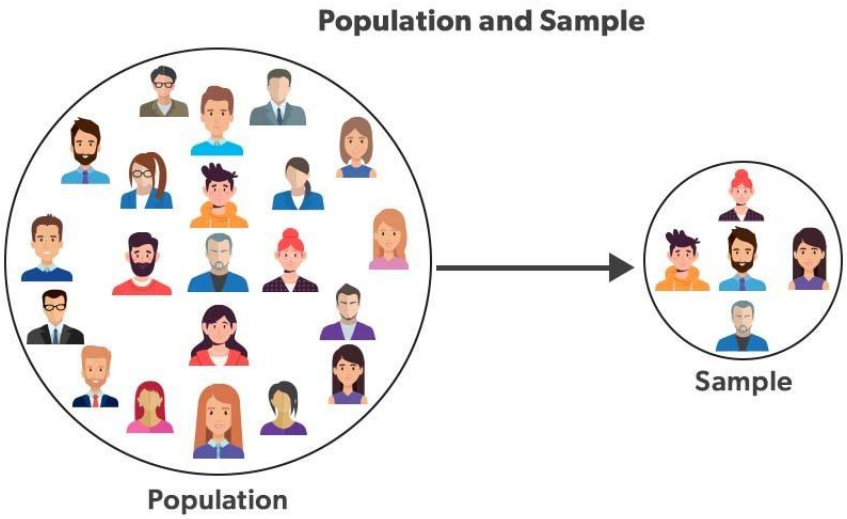

Figure 3 - Population vs sample

Collecting data from a sample:

Sample should be randomly selected and representative of the population, because reduce the risk of sampling bias and guarantee an internal and external validity. (Published on May 14, 2020, by Pritha Bhandari. Revised on June 21, 2023)

Question 2

Part A:

1. Adress Missing Data:

Missing data is a common problem affecting most of the databases in dealing with the analysis, and because of this situation, create a difference in the credibility about the results. There are 3 common situations to deal with missing data:

I. Deletion method(deleting):
It's the simplest way to deal with missing data, you just must discard the row or column that are missing values.
There are 3 ways to do it: complete-case analysis, available-case analysis and weighting method.
   i. Complete-case analysis: all the observations where one missing variable is deleted.
   ii. Available-case analysis: Discards data in the variables that are needed for a specific analysis.
   iii. Weighting-case analysis: it's taken from complete-cases by modelling the missingness to reduce the bias in the available-case.

   Advantages: Simplicity.

   Disadvantages: Depending on the information missing, it can change the integrity of the data.

II. Single Value Imputation method(mean):
In Single imputation, the missing value are added to the data with a predicted value, ignoring the uncertainty and underestimating the variance.
   i. Mean and median: This method is to change the missing value by mean or median variable.

      Advantage: Application will be easier.

      Disadvantage: less variability, causing more estimate errors compared to deletion approaches.

III. Nearest Neighbours:
Refers to finding the similarity or the closest data points in a dataset given with missing value. it works well when it comes related to patterns between them.

Advantages: Easy to understand, it is a multi-class scenario and performs well with small datasets.

Disadvantages: It does not work well with large datasets.

2. Applying the missing value in the data set given:

I. Height, Column D:
In this case there are 2 ways that we can do it, nearest neighbour or mean method:

i. Mean: It will be necessary take the average of all the students: 1.65+1.71+1.82+1.85+1.81 = 1.77.
After it, you add the number 1.77 in the missing value.

ii. Nearest neighbour: It is necessary to observe the data set and check for the similarities from other participants. In this case the nearest neighbour from the student row 3 is the student in the row 2, because there are similarities in gender, nationality, classification size, shoe size, age and eye colour. Just add in the missing value in the Column D the size 1.71

II. Classification size, Column E:
In this situation, it will be used the nearest neighbour to replace the missing value. It can't be used the mean method because there are no numbers to represent the size classification.

i. Nearest neighbour:
As it was done before, the size classification from the student in the row 4 column 5 will be as the same as the student in row 6, they have similarity in gender, nationality, and eye colour. And the height and age are close too.
So, the size classification for student in row 4 will be N in the missing value.

Part B:

1. Nearest Neighbour (NN)

Nearest neighbour analysis is defined as a method to observe the randomness or pattern of a data, comparing the closest neighbour data value to put the missing value. AI generated definition based on: Earth-Science Reviews, 2013

Example from the dataset:

Using the NN to find the missing value in the third row, you must get the closest information from the other participants, in this case the height will be 1.71 because the NN is the participant from the second row that has the same gender, nationality, size classification, shoe size, age and eye colour.

2. K-Nearest Neighbour(K-NN)

It is very similar to Nearest Neighbour, but it makes classifications from group data point.
It can be handled in numerical and categorical data, being flexible with choice in various types of datasets.
For example:
In the fifth row which is missing a value from size classification. K-nearest Neighbour will consider the most similar and get the S, because appears with frequency.

Advantages: easy to implement, adapts easily.
Disadvantages: It doesn't scale well, Overfitting.

3. Relationship between NN and K-NN

Nearest neighbour use one closest data point for classification, while K-nearest neighbours use the majority value of the K closest points.
K-Nearest neighbours are most required to dataset when is more detailed and its necessary more than one opinion to make closer with the reality.

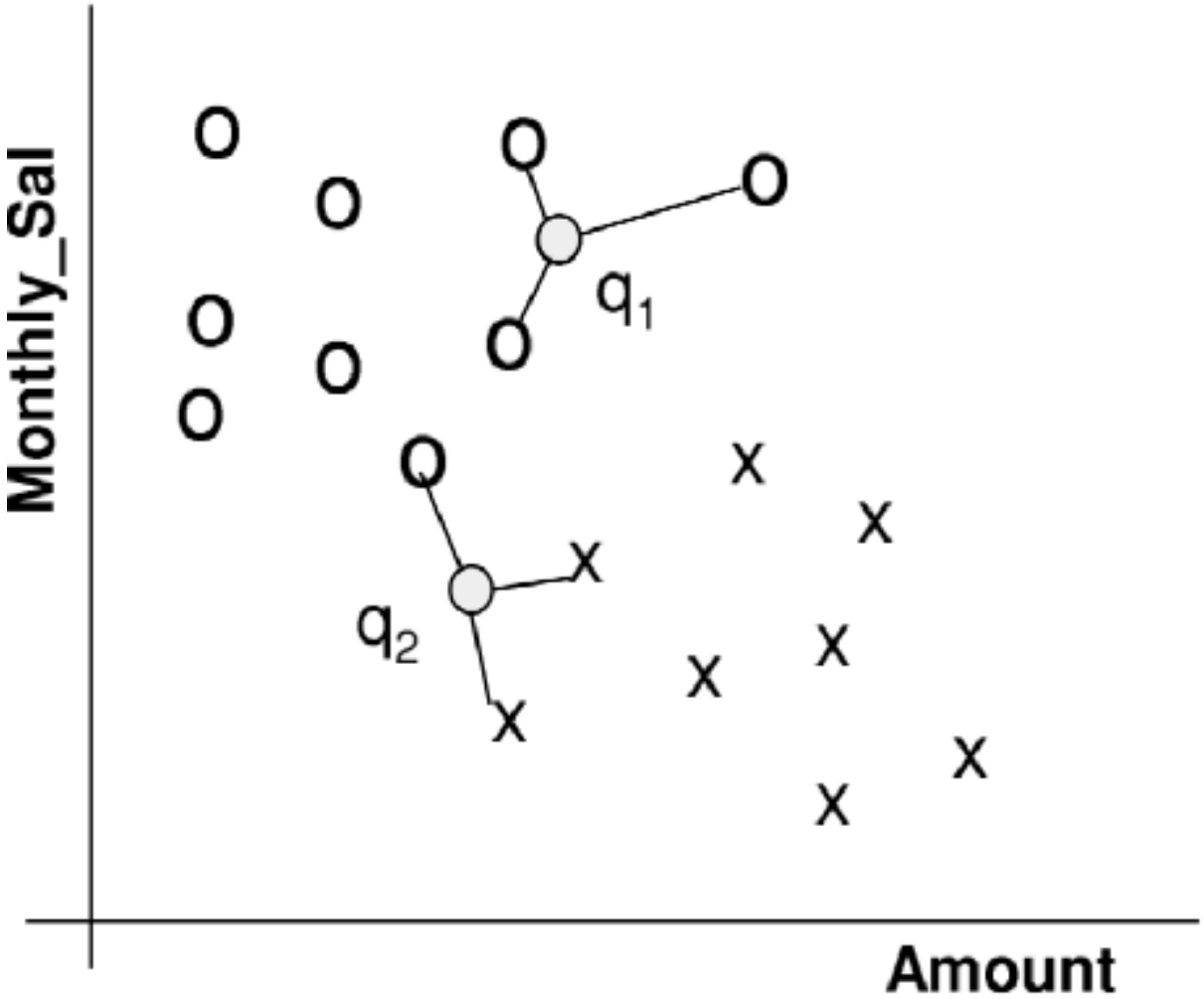

Figure 4 - K-NN and NN

References:

Question 1:

https://365datascience.com/trending/data-types-complete-guide/

https://www.researchgate.net/publication/358149225_Scales_of_Measurement_in_Research#:~:text=Nominal%20scales%20are%20used%20to,true%20zero'%20can%20be%20defined.

https://online.stat.psu.edu/stat800/lesson/1/1.1#:~:text=Quantitative%20variables%20may%20be%20discrete,infinite%20number%20of%20decimal%20places

Figure 1: https://www.analyticsvidhya.com/blog/2021/06/complete-guide-to-data-types-in-statistics-for-data-science/

Figure 2: https://www.voxco.com/blog/measurement-scales/

Figure 3: https://www.geeksforgeeks.org/population-and-sample-statistics/

Question 2:

Part A:

https://link.springer.com/content/pdf/10.1007/978-3-319-43742-2_13?pdf=chapter+toc

https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=9c11e93705e2dea51e7936f846ed2303b46a7a1d

Part B:

https://www.sciencedirect.com/topics/earth-and-planetary-sciences/nearest-neighbor-analysis

https://www.geeksforgeeks.org/k-nearest-neighbours/

https://www.naukri.com/code360/library/knn-vs-k-means

Figure 4: https://www.semanticscholar.org/paper/k-Nearest-Neighbour-Classifiers-A-Tutorial-Cunningham-Delany/3220aa51ae28313c223c961c373b869d20f07b39/figure/0