

**TITLE:** Data Exploration and Preparation – Group C.A

**STUDENT NAME:** Alisson Alves De Moura.  
**STUDENT NUMBER:**  2019142

**STUDENT NAME:** Marcos Vinicius Mariano Rodrigues  
**STUDENT NUMBER:**  2019146

**STUDENT NAME:** Valter Brito  
**STUDENT NUMBER:**  2019xxx

**LECTURER:** Dr. Muhammad Iqbal

Content

[Problem description: 3](#_Toc121618070)

[Introduction and Objective. 3](#_Toc121618071)

[Dataset Chosen reasoning 3](#_Toc121618072)

[Motivation and Challenges 3](#_Toc121618073)

[- Present the challenges the dataset shows initially 3](#_Toc121618074)

Thank you!

Problem description:

This area will explain what will be the problem, what we intend to achieve as a solution for it

## Introduction and Objective.

Explain in details what type of techniques we want to show, knowledge to be applied to achieve an objective; then explain the objective in details such as how the dataset will be completely analysed and so on.

## Dataset Chosen reasoning

The dataset needs to have certain characteristics and also be of challenge to the students in order to create solutions for a good analysis and reasoning of data

# Motivation and Challenges

Explain the area will show the challenges and the analysis of the dataset and how we intent to overcome these limitations

## - Present the challenges the dataset shows initially

The main challenge to understand this data is that the major number of data is not numerical.

Also, plotting non-numerical data is challenging as it demands an accurate reasoning of facts to make reasonable and efficient plotting; some of the data won't make sense at first before reasoning other aspects of the dataset, this can be an challenge aggravated by the mentioned above as interpretation of the data when it’s not numerical is subject to human error.

The dataset contains important data from children and adults; once they are categorised in the same dataset with all columns applying to them, sometimes a column won't make sense for one of them such as "veteran benefits" considering that children don't fight wars.

Sub-datasets can be created to separately run analysis and join together when terms apply for all, such as gender to overcome this limitation; some questions for the dataset are important to be analysed, such as why children value for education don't apply for all minor ages?

Cleaning an dataset with so many attributes is always an delicate action once valuable data can be lost and change how an accurate measurement of factor can look like when an comparison of other aspects where there is no missing data in the attributes analysed.

## - Identify Categorical discrete and continuous data