

**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:**  2019308

**LECTURER:** Dr. Muhammad Iqbal

Content

[Problem description: 4](#_Toc121619882)

[Introduction and Objective. 4](#_Toc121619883)

[Dataset Chosen reasoning 4](#_Toc121619884)

[Motivation and Challenges 4](#_Toc121619885)

[- Present the challenges the dataset shows initially 4](#_Toc121619886)

[- Identify Categorical discrete and continuous data 4](#_Toc121619887)

[- Questions about the data quality 5](#_Toc121619888)

[- Show pre-conceptions about the data and how we intend to explore it. 5](#_Toc121619889)

[- Questions for Acquiring knowledge of data. 5](#_Toc121619890)

[Description and explanation of the techniques we used for the data analysis. 5](#_Toc121619891)

[- Data Statistical parameters 5](#_Toc121619892)

[- Question parameters and tendencies. 5](#_Toc121619893)

[-Apply normalizations/standardization Scalar on numerical data. 5](#_Toc121619894)

[-Line, scatter and Heatmaps for correlation between features 5](#_Toc121619895)

[- Data exploratory Analysis; 6](#_Toc121619896)

[- Apply dummy encoding to categorical categories 6](#_Toc121619897)

[- Apply PCA with a chosen number of components 6](#_Toc121619898)

[- What is the purpose of Dimentionality reduction? 6](#_Toc121619899)

[- Data understanding 6](#_Toc121619900)

[Conclusion. 6](#_Toc121619901)

[References. 6](#_Toc121619902)

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.

## - Identifying data type

There are a total of 41 attributes in the dataset, configured in Nominal and continuous attributes.  
The census collected mostly nominal attributes as it’s related to a status defined by social constructs and not numbers; the list follows:

|  |  |
| --- | --- |
| **Continuous** | |
| Age | num persons worked for employer |
| capital gains | total person income |
| capital losses | taxable income amount |
| dividends from stocks | weeks worked in year |
| instance weight | Income |

**LIST OF CONTINUOUS ATTRIBUTES**

The above attributes were represented as numerical in the analysed dataset, and can be analysed in a numerical way without compromising the meaning of the data.

**LIST OF CATEGORICAL ATTRIBUTES**

|  |  |  |
| --- | --- | --- |
| **Categorical** | | |
| class of worker | region of previous residence | Hispanic origin |
| industry code | state of previous residence | Sex |
| occupation code | detailed household and family stat | member of a labour union |
| education | migration code-change in MSA | reason for unemployment |
| veterans benefits | migration code-change in REG | full or part time employment stat |
| enrolled in edu inst last wk | migration code-move within REG | federal income tax liability |
| marital status | live in this house 1 year ago | country of birth mother |
| major industry code | migration prev res in sunbelt | country of birth self |
| major occupation code | family members under 18 | Citizenship |
| Race | country of birth father | own business or self employed |
|  |  | Year |

The above attributes are represented as text, even though some of this text are numbers, these numbers do not represent value and cannot be treated as holding value.   
As specified prior, the actual dataset contains a big number of nominal data:

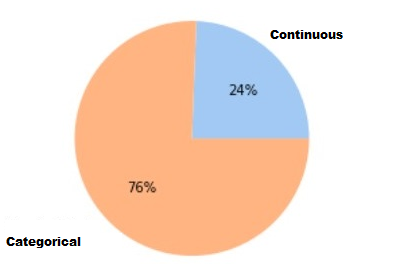
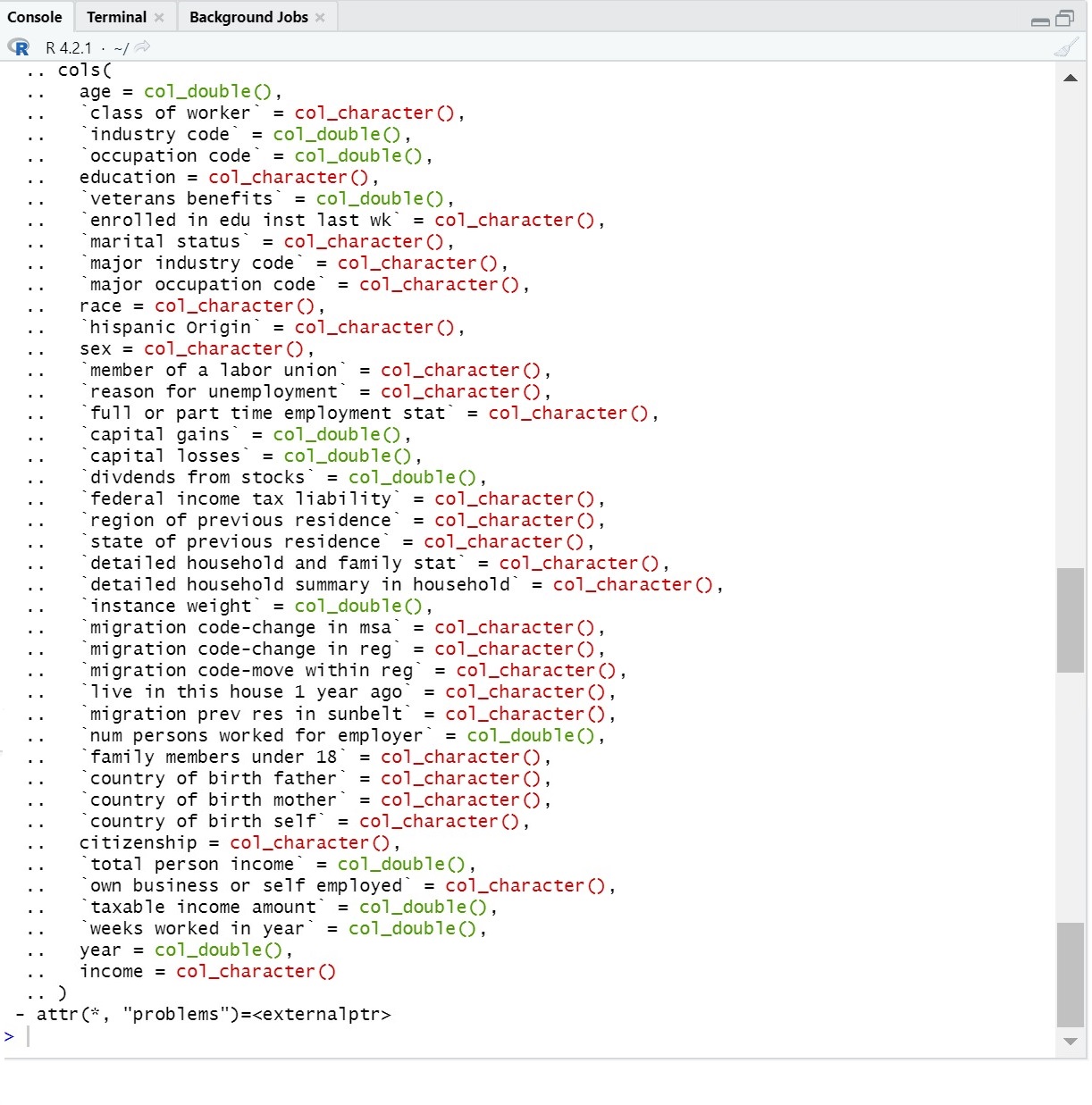


Figure - Ratio between Categorical and Continuous attributes



Some explanation

## - Questions about the data quality

“Not in universe” value shows very often for many cases that doesn't really apply; such as employment for many different ages where people have education and working age; and other cases where the category of work is 'not in universe' even though the person is self employed.

cases like these show that during the data collection many people were not consistent with their answers and this will contribute to the inaccuracy of the data analysed; some of this cases can be mitigated by cleaning the data, but it’s impossible to really know if the data was cleaned properly and the data entered was accurate and trustworthy; this induces the question of which attributes are expected to be not true, and which are mostly certainly true;

examples of this could be the age or gender can be believable much more than taxable income or if the person is not in the working universe; due to social constructs some people will be induced to lie when they know something can be lost if the information is made public; this will make the dataset not very accurate for each scenario.

## - Show pre-conceptions about the data and how we intend to explore it.

What are the impressions we have from the data initially? How can we prepare it to verify it?

## - Questions for Acquiring knowledge of data.

Reason the questions we want to make around the relations between columns and its values;  
how can we make nominal data analysed in a numeric way;

Etc

# Description and explanation of the techniques we used for the data analysis.

Add a subtitle for each technique and shortly explain how it works and how it was applied.

## - Data Statistical parameters

(mean, median, minimum, maximum, deviation)

### - Question parameters and tendencies.

What parameters show which tendencies? How can we challenge it?  
What does that means for the data?  
Does that consists with the main conceptions?

## -Apply normalizations/standardization Scalar on numerical data.

- How can we explore the data with these normalizations?  
- Does that changes any understanding of the data?

## -Line, scatter and Heatmaps for correlation between features

- Reasoning for the correlations

- What can we learn from these plots?

## - Data exploratory Analysis;

Identification of sub-groups of features and new learning’s we explored.

What new was learned from the Analysis done so far?  
 Is something not explored? Why can't we explore this?  
 Are the new findings believable or is the data compromised? how?

## - Apply dummy encoding to categorical categories

- How does this work and what's the reasoning behind it?  
- What new can be learned about the dataset?  
- Pros and cons from how the applied Dummy encoding can be interpreted and how relevant is it for the data

## - Apply PCA with a chosen number of components

- The profile of the components extracted  
- The correlation between the components is real?  
- The distance between items have an explanation?  
- The profile of the analysis answers a question? Which one?

## - What is the purpose of Dimentionality reduction?

- What is Dimensionality reduction? (PCA)  
- When Dimensionality reduction should be used?  
- What we learned using it?

# - Data understanding

- What can be learned after all the data exploration done?  
-Was the data plenty explored?  
 - How valid this new knowledge is? does the data demands any correction or more data?  
- how can the information learned from the data be used?

## Conclusion.

brief conclusion

## References.

ref