

**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 Lopes de Brito  
**STUDENT NUMBER:**  2019308

**LECTURER:** Dr. Muhammad Iqbal

Content

[Problem description: 4](#_Toc121656494)

[Introduction and Objective. 4](#_Toc121656495)

[Dataset Chosen reasoning 4](#_Toc121656496)

[Motivation and Challenges 5](#_Toc121656497)

[- Challenges the dataset shows initially 5](#_Toc121656498)

[- Identifying data type 5](#_Toc121656499)

[- Questions about the data quality 7](#_Toc121656500)

[- Show pre-conceptions about the data and how we intend to explore it. 8](#_Toc121656501)

[- Questions for Acquiring knowledge of data. 8](#_Toc121656502)

[Description and explanation data analysis techniques 8](#_Toc121656503)

[- Data Statistical parameters 8](#_Toc121656504)

[- Question parameters and tendencies. 9](#_Toc121656505)

[-Apply normalizations/standardization scalar on numerical data. 9](#_Toc121656506)

[-Line, scatter and Heat Maps for correlation between features 9](#_Toc121656507)

[- Data exploratory Analysis; 9](#_Toc121656508)

[- Apply dummy encoding to categorical categories 9](#_Toc121656509)

[- Apply PCA with a chosen number of components 9](#_Toc121656510)

[- Data understanding 10](#_Toc121656511)

[Conclusion. 10](#_Toc121656512)

[References. 10](#_Toc121656513)

Thank you!

Problem description:

This assessment provides the student group the challenge to use tools for data preparation to elaborate exploration of a dataset in way to acquire knowledge of it to answer questions regarding to what the data represents, what to expect in the future, the robustness the data shows and how the new knowledge can be used.

For this some steps were created in order to go through the dissection of the chosen dataset as shown in the index and as it follows;

## Introduction and Objective.

The dataset was chosen initially following a few mandatory rules regarding the subject the dataset should represent, the number of instances and attributes the dataset should have available after cleaned.  
After these criteria were met, the dataset should have understandable data to be made reasoning from, the chosen dataset for this was census taken in the USA that meets economic and housing status from citizens.

With this dataset, we will be able to utilise methods to visualize statistical parameters such as median, standard deviation, etc. Enabling the students to understand tendencies and patterns and correlations; Apply normalizations and/or standardization techniques to explore the data in different perspectives; Plot and analyse lines, scatters and heat maps to verify correlations for further understanding of the data; perform data exploratory analysis in order to identify subgroups and acquire new learning from the data; Apply a dummy encoding to categorical categories and verify what new can be learn of the data in these circumstances; Verify if PCA is applicable and if positive verify correlation between its components.

All these techniques will help the students and the reader to acquire data understanding and new information the dataset holds within; possibly creating new questions and suggestion on how further data should be collected and addressing society problems with the knowledge learned after the data being comprehended.

## Dataset Chosen reasoning

With the challenge of analysing a specific problem within the areas as health, housing, agriculture, economy and environment the chosen dataset covers economy as the civilian status and housing for the last year; these two aspects are explored within 41 different attributes inside one or both universes in order to provide an good scope in the relation between both areas and knowledge between the relation within and among such areas.  
  
The dataset is a subset from an even bigger dataset; this decision was made due to the collaboration work type chosen being using GitHub to update changes made by group members in real-time; creating an environment where all member can work together at the same time.  
GitHub has a limitation of 100MB upload by commit, and the original dataset would be bigger than this limit.

This dataset contains more than 90,000 instances that will mitigate the problem of outstanding values that can be considered noise in the data analysis; this mitigates also mistakes done in data cleaning and other data-filters that can change the information the plotting of the data show. The dataset also will provide some interesting challenges, once most of its data is Categorical and for this needs to be comprehended and analysed from close as to avoid mistakes from interpretation and plot reasoning

# Motivation and Challenges

This topic will discuss the main challenges to overcome with this assessment and ask questions and pre-assumptions the data give as impression in a first glance before being prepared and transformed for proper exploration.

Firstly the focus will go on the initial challenges the dataset provides and how can this impact the data exploration.  
Then the quality of the data will be discussed as to if it provides liabilities in the data analysis and what could be done for mitigating bad data.  
With a crude understanding of the data, some assumptions will be done in order to create a basic understanding of the data where some questions can be asked from.  
Furthermore a few questions will be elaborated in order to analyse the data as to acquire some knowledge of it so better constructed questions can be asked out of it in order to make good reasoning of the data.

The main motivation is the proper analysis of this dataset will lead to better understanding of how data was collected, what are the important data in the dataset, how to even collect data in a more efficient way and improve even more data understanding in order to achieve solid knowledge.

## - 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 a dataset with so many attributes is always a 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 Categorical and Continuous attributes.  
The census collected mostly Categorical attributes as it’s related to a status defined by social constructs and doesn’t hold numeric value; the lists 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:

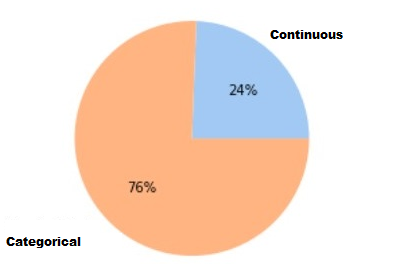


Figure - Ratio between Categorical and Continuous attributes

With this knowledge, our dataset was treated as to hold the right information as number or string with the code as follows:

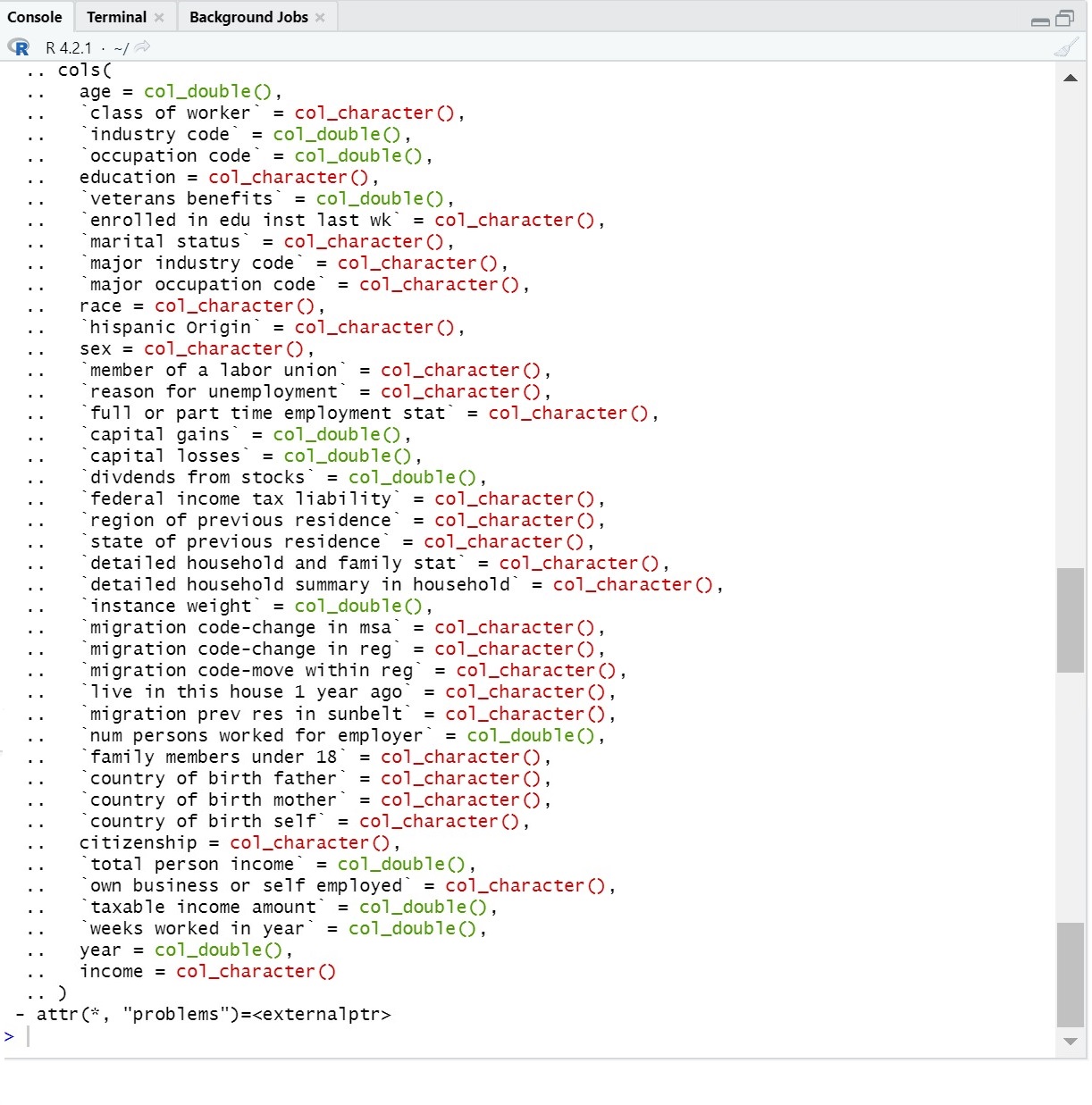


Figure - Data categorization

## - 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?

The data seems to have many aspects that don’t apply for many of the

## - 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 data analysis techniques

To acquire the desirable knowledge from the data, it’s important to use some important techniques for data analysis as to learn the relation between attributes and treat the numerical and categorical data as to learn knowledge from it.

The techniques covered will be presented furthermore.

## - Data Statistical parameters

Calculating statistical parameters is paragon to data analysis and not only for numerical data as it might indicate at first; Some categorical data can be prepared as to hold some actual value to it making it possible to b analysed with statistical tools;

Some of the main statistical parameters we will be analysing in this data are:

* **Mean:** Mean is the sum of all values divided by the number of values, with this it’s possible to find the average value among all the instances; this helps the reader to understand data tendencies and is an important step to find other useful information that depends entirely on the average.
* **Median:** Median is the immediate middle value among all the instances; this value also is important to see the data distribution among the dataset, as the middle value being above average would show a tendency for values to be over the average in the future for instance.
* **Minimum & Maximum:** These values will inform the greatest and smallest values in the dataset.
* **Deviation:** Deviation will show the difference between values and the average, The standard deviation shows the average difference between the values; This is important to see the value spread;

### - Question parameters and tendencies.

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

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

Normalization is the term used for when data entries is organized in order to ensure they appear similar across instances making it easier to group and analyse information.  
Standardization is the term used for converting data to a common format, making it possible to process and analyse data.

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

## -Line, scatter and Heat Maps 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

Dummy encoding is a method designated to transform categorical variables into binary variables where it can be analysed and processed with techniques it wasn’t possible to apply prior as the type of data changes.

- How does this work and what's the reasoning behind it?  
- What new can be learned about the dataset?  
  
When a model has many dummy variables, the consistency of the information can be challenged as not being trustworthy; this also can be an liability for explain big finding in a dataset once most of the information wasn't actually collected in first place; it’s interesting to use it with caution.

## - Apply PCA with a chosen number of components

Principal Component Analysis (PCA) uses machine learning to accomplish reduction of variables for a predictive model with better performance; considering that when there is much data to be analysed, processing all this data will take off performance data from machines that will have to process a huge amount of data before it can make decisions. To mitigate this problem, using machine learning it is possible to reduce the dimensionality of the data without compromising much the efficiency of the data lost causes, making sure the performance wouldn't be a liability in the final process.

- 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 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.

sisense.com (no date) - *Data standardization*, *Sisense*. Sisense. Available at: https://www.sisense.com/glossary/data-standardization/ (Accessed: December 08, 2022).

Saxena, S. (2020) *What is categorical data: Categorical Data Encoding methods*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/#h2\_8 (Accessed: December 8, 2022).

Brownlee, J. (2020) *Principal component analysis for dimensionality reduction in python*, *MachineLearningMastery.com*. machinelearningmastery.com. Available at: https://machinelearningmastery.com/principal-components-analysis-for-dimensionality-reduction-in-python/#:~:text=for%20Dimensionality%20Reduction-,Dimensionality%20Reduction%20and%20PCA,to%20predict%20the%20target%20variable. (Accessed: December 10, 2022).