**Lab assignment 1**

**CIS 492**

**Fabio Hinojosa Jimenez**

**2861151**

**No extra credit**

Table of Contents

[Part 1 3](#_Toc163917184)

[Feature Selection 3](#_Toc163917185)

[Select Attributes 3](#_Toc163917186)

[Determining data properties 3](#_Toc163917187)

[Part 2 4](#_Toc163917188)

[Data Preprocessing and Transformation 4](#_Toc163917189)

[Handling null values 4](#_Toc163917190)

[Random Sampling & Bootstrap Sampling 5](#_Toc163917191)

[Mean/Variance/Standard Deviation 5](#_Toc163917192)

[Boxplot, Histogram, or Z-score 5](#_Toc163917193)

[Identify Outliers 8](#_Toc163917194)

[Normalization & Standardization 8](#_Toc163917195)

[One Hot Encoding 9](#_Toc163917196)

[Final preprocessed data 9](#_Toc163917197)

# Part 1

## Feature Selection

### Select Attributes

For the this part I read csv file and select the features that I think will be necessary. I then save it in a new file called ‘vTargetBuyers’.

A white background with red text

Description automatically generated

### Determining data properties

For this part I go through each feature value, first it checks if it is int or float, from there I check if they are Ratio, Interval, or Ordinal. If the value is not int or float, I used the data property assigned before to determine if the data is discrete or continuous.

A screenshot of a computer code

Description automatically generated

This being the results:

**A screenshot of a computer

Description automatically generated**

# Part 2

## Data Preprocessing and Transformation

### Handling null values

We first import all the necessary libraries, initialize the necessary objects, and start a loop that will go through all of the features.

To start we ignore ‘bike buyer’ and ‘Customer key’

A screenshot of a computer program

Description automatically generated

To handle with null values I replace them with the mean of each data feature

A close-up of a computer code

Description automatically generated

### Random Sampling & Bootstrap Sampling

Performing random sampling and bootstrap sampling by using ‘np.random.choice’, making it around 70% of the original size.



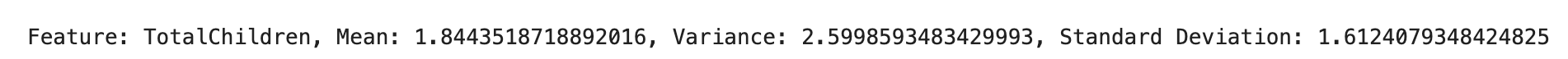
### Mean/Variance/Standard Deviation

Using Pandas package, I get the mean, variance, and Standard deviation.

A white background with red text

Description automatically generated

Results:







### Boxplot, Histogram, or Z-score

Using matplotlib and seaborn, we display the graph for Z-score, histogram, and the boxplot.

A screenshot of a computer code

Description automatically generated

A black and red text

Description automatically generated

Results:

A graph with blue squares

Description automatically generatedA graph of a person with histogram

Description automatically generated

A graph of a graph of data dispersion

Description automatically generatedA diagram of a box plot

Description automatically generated

A diagram of a box

Description automatically generatedA blue and black box plot

Description automatically generated

A graph with blue lines

Description automatically generatedA graph with blue lines

Description automatically generated

A graph of blue dots

Description automatically generated

### Identify Outliers

Identifying outliers, we create two variables called Q1 and Q3 for each quantile. Then we get IQR by subtracting Q1 and Q3. We then get everything else by using the equations.

A close-up of a computer code

Description automatically generated

### Normalization & Standardization

Using scikit-learn’s ‘MinMaxScaler’ and ‘StandardScaler’ I performed normalization and standardization. We changed the data into DataFrame type to have a better visualization.

A screen shot of a computer code

Description automatically generated

Results:

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

### One Hot Encoding

In this last section of the code I added the OneHotEncoding, in all my categorical features.

I create new columns with a name being ‘feature\_category’, the columns are added to the dataset, and the old columns are deleted.

A screenshot of a computer code

Description automatically generated

You can see the results of the OneHotEncoder in the final process data screenshot.

## Final preprocessed data

A screenshot of a table

Description automatically generated

A screenshot of a chart

Description automatically generated