BAN110: Data preparation and Handling

Group project

Fall 2019

## Introduction

This is a data cleaning group project. The goal of this project is to apply data cleaning tools and techniques learnt throughout the course to check and correct errors in data and prepare data for analysis. The project objectives are as follows:

* Describe the data analytics task for the chosen dataset.
* Examine the dataset characteristics, the type of independent variables and the type of the dependent variable (the target variable if there is).
* Examine the categorical variables: check and correct errors, create new derived variables that serve to answer a specific question, or combine categories (values) of a specific variable when necessary.
* Examine the numerical variables. Check for errors given range of values or any other criteria if possible. Check for missing values, and replace them by imputation when necessary. (note: you may need to treat outliers for a specific variable before applying imputation as the value of the mean will be greatly impacted by the presence of outliers).
* Test for normality of at least one numeric variable (especially the target variable) and suggest suitable transformations when necessary
* Detect outliers based on the standard deviation method or the interquartile range.
* Treat outliers by deletion.

## Datasets requirements:

* The dataset should contain a mix of numeric and categorical attributes.
* The dataset should contain at least five attributes.
* The dataset contains missing values for at least one of the variables.

## Project report layout

1. Dataset and task description
2. Loading data
3. Dataset characteristics
   1. Target variable
      1. If categorical, show the frequency distribution of each of the possible values. Interpret. Is the dataset balanced? Any other comment?
      2. If numerical, show the statistics (min, max, mean) and the shape of the distribution of the target variable through a histogram. In some case, numerical target variables need transformation to make data modeling possible.

### Categorical variables

1. Check and correct errors when necessary.
2. Check and treat missing values through imputation with the mode.
3. Create one or more derived variables. Justify why the derived variable is created? Does it answer a specific question? Does it serve for data modeling? Etc..

### Numerical variables

1. Check (range of values/ less than/larger than) and correct errors by deletion.
2. Check for missing values and correct through imputation with the mean.
3. Check the distribution of one or more numerical variables to decide which method to use for outlier detection.
4. Detect and remove outliers.
5. Test for normality and plot histogram and QQ plots for a variable with a skewed distribution. Apply a transformation and test for normality again with histogram and QQ plot.

## Submission:

Include the following in your submission:

1. Submit a project report in Ms word. Code snippets can be added.
2. Submit a powerpoint presentation showing the chosen dataset, the cleaning steps and any issues you found in the data.
3. Submit a zipped folder containing the dataset folder, sas programs or code in jupyter notebook with the corresponding html file.

## Rubrics:

|  |  |
| --- | --- |
| Rubric | Detail |
| Data analytics Task [5%] | Mention any task associated with the dataset. The UCI machine learning repository provides a task description such as regression, classification, clustering. You are required to provide a context of use of the dataset if this information is not made available by the dataset publisher. |
| Dataset characteristics[5%] | Provide the dataset characteristics:   1. The number of instances 2. The list of variable names and their types 3. Identify the dependent variable/target variable and the independent variable. 4. For the target variable, if categorical, show the frequency table and examine if the dataset is balanced among the different values or not. |
| Categorical variables [30%] | |
| Check and correct errors when necessary [10%] | For each categorical variable,   1. List the values and show the frequency table. 2. Report any error, 3. Correct errors by deletion |
| Check for missing values  Treat missing values [10%] | For each categorical variable,   1. Check for missing values, you may run a code that checks for missing values for all the categorical variables using the informat method (check the lectures for reference). 2. Treat missing values by imputation. With this option, you replace the missing values with the mode. |
| Create one or more derived variables or combine values of a categorical variable. [10%] | 1. Propose suggestions such as:    * Create a new derived variable?    * Break down a variable into two or more to serve answering a specific question. |
| Numerical variables [40%] | |
| Check and correct errors [10%] | For the numerical variables,   1. List the min, max, mean and the quantile table and some other basic statistical measures. 2. Use suitable data cleaning tools to report errors. Errors should be identified and located with the variable id. Do not output a table with valid, invalid, errors without being able to identify the record id. You may use:    * Range of values    * Extreme observations table 3. For errors the result from out of range values, correct by deletion. |
| Check for missing valuesTreat missing values [10%] | For each categorical variable,   1. Check for missing values, you may run a code that checks for missing values for all the numerical variables at a time. 2. Treat missing values:    * By imputation. Justify why. With this option, you may need to detect outliers first and isolate them before you apply the imputation. |
| Detect and remove outliers [10%] | Examine at least 2 numerical variables:   1. Detect outliers. Use a method of your choice.    * If you would like to use the standard deviation method, check if the variable is normally distributed so you can apply this method.    * The interquartile range allows checking for outliers whatever the distribution of the variable is. |
| Test for normality and transformation of distribution [10%] | Detect a variable that is highly skewed.   1. Plot a histogram to show the shape of the distribution of that variable 2. Show the QQ plots of the variable. Comment on the QQ plot. 3. Apply a transformation on the variable. You may try log transformation 4. Test for normality again with histogram and QQ plot. Comment on the QQ plot. |
| Peer evaluation [10%] | |
| Report quality and presentation [10%] | |

## Dataset options:

1. Auto-mpg dataset:

<https://www.kaggle.com/uciml/autompg-dataset>

1. Heart disease dataset

<https://www.kaggle.com/ronitf/heart-disease-uci>

1. Census income dataset

<https://www.kaggle.com/uciml/adult-census-income>

1. Bike sharing dataset

<https://www.kaggle.com/marklvl/bike-sharing-dataset>

1. Suicide rates dataset:

<https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016>

You are free to use any other dataset from the following sources. Please make sure the dataset meets the requirements listed in dataset requirements section.

Kaggle: <https://www.kaggle.com/datasets>