HUMBER INSTITUTE OF TECHNOLOGY

AND ADVANCED LEARNING

(HUMBER COLLEGE)

Group Assignment

# COURSE: Machine Learning 1

# Group: G-5

SUBMITTED BY:

|  |  |  |
| --- | --- | --- |
| Last Name | First Name | Student Number |
| Ding | Hui | n01151542 |
| Jain | Ayush | n01519672 |
| Tran | Gia Bao | n01556359 |

SUBMITTED TO: Professor Sarama Shehmir

SUBMISSION DATE: 2/10/2023

**Dataset Overview:**

Boston Housing is a dataset which provides a comprehensive data about housing in Boston, the USA. It consists of 14 effective attributes in it, each providing specific information about the neighborhoods including measurements related to crime rates, environmental factors, demographics, and more. There are 506 data observations which represent a different town or neighborhood in Boston. However, it is not tidy dataset for some missing values and outliers, so we have to do some data wraggling before doing any other analysis.

**Data Manipulation:**

Top of Form

Initially, we read and checked columns and rows. The output shows there is a column called “unnamed 14” that contains NaN in all of it.

A screenshot of a computer

Description automatically generated

We decided to delete it for the next step of omission using the following code:

A screenshot of a computer

Description automatically generated

Next, We generated a heatmap to gain an overview of missing values. Then, we employed a 'highlight null' code to identify them, highlighted in yellow.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

We also found the second way to highlight the null values using this function.

A screenshot of a computer

Description automatically generated

Next, we found the data in column “PTRATIO” contains some outliers. We drew a box plot to detect them:

**A screen shot of a graph

Description automatically generated**

We used IQR to detect outliers. First, we calculated the IQR, which is the range between the first quartile (Q1) and the third quartile (Q3) of the data. Any data point below Q1 - 1.5 \* IQR or above Q3 + 1.5 \* IQR is considered an outlier.

A screenshot of a computer program

Description automatically generated

A table with numbers and letters

Description automatically generated

Then we highlighted the outliers in red to differentiate them from missing values in yellow:

A computer code with text

Description automatically generated

A screenshot of a computer

Description automatically generated

The possible causes of outliers are: (a) Typing non-numeric value. (b) Shift in decimal place while data entry error and (c) Genuine case of outlier. The column is pupil teacher ratio, so we conclude that all these outliers are c type from this feature.

A table of numbers with a number on it

Description automatically generated

After discovering all missing values and outliers, the next step we did was to clean the data. There are two ways to do it, one is omission, the other is imputation.

We initially cleaned data by omission because the missing values are not in large numbers, which would not influence the final analytical result. We first marked all missing values as “NaN”:

**A screenshot of a computer

Description automatically generated**

Then, we used “dropna” to delete all rows which contain missing values.

**A screenshot of a computer code

Description automatically generated**

After performing the omission, there remain 491 rows, and they have been organized in a tidy format suitable for subsequent analysis.

Another method for data cleansing involves replacing missing values with either the mean or median. If the data is normally distributed, we will replace it with mean, and if it is skewed, we will replace it with median. Therefore, we checked normality for each column containing missing values by histogram.

A close-up of a sign

Description automatically generated

A graph of a person

Description automatically generated with medium confidence

A white background with blue and red text

Description automatically generated

A graph with numbers and a bar

Description automatically generated

A white rectangular sign with blue and red text

Description automatically generated

A graph of a number of blue bars

Description automatically generated with medium confidence

A white rectangular sign with blue and red text

Description automatically generated

A graph of blue bars

Description automatically generated

A close up of a text

Description automatically generated

A graph of a number of blue bars

Description automatically generated with medium confidence

A white rectangular sign with blue and red text

Description automatically generated

A graph with blue lines

Description automatically generated

Based on the histograms of these columns, we found that they were not normally distributed, so we filled the missing values with the median.

A close-up of a white background

Description automatically generated

A screenshot of a computer

Description automatically generated

After conducting imputation, we checked the dataset, and it showed no missing value, and it was tidy and well-prepared for any analysis followed up.

**A screenshot of a computer

Description automatically generated**

# References

Boston Housing dataset: <http://lib.st>a[t.cmu.edu/datasets/boston.](http://lib.stat.cmu.edu/datasets/boston)) The dataset has 167 cases.

The data was originally published by Harrison, D. and Rubinfeld, D.L. `Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.

Top of Form