Sana'a University Faculty of Computer and Information Technology

**Group Project Documentation**

**Data Mining Lab**

Data Set- **Boston Housing** (The Boston House-price Data, n.d.)

Submitted by:

|  |  |
| --- | --- |
| **Name** | **Student Number** |
| Hamza Hamood Al-Saqaf | 21\_16\_4174 |
| Osama Abdel Haq Al-Yahyawi | 21\_16\_4079 |
| Osama Mohammad Al-harsh | 21\_16\_4145 |
| Hamza Abbas Ishaq | 21\_16\_4193 |

Submitted to: Dr.Nora Al Hababi

Submission Date: 2024-01-25

INDEX

[PART A :- Overview 3](#_Toc157038251)

[PART B:- Data Pre-processing 5](#_Toc157038252)

[Part D:- linear regression 11](#_Toc157038253)

[REFERENCES 13](#_Toc157038254)

# **PART A :- Overview**

The Boston Housing Dataset

The Boston Housing Dataset is a derived from information collected by the U.S. Census Service concerning housing in the area of [Boston MA](http://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html). The following describes the dataset columns:

|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | Potential Problems/Considerations |
| index | int | None |
| CRIM | float | have outliers |
| ZN | int | None |
| INDUS | float | Missing values, potential data quality issues |
| CHAS | int | Missing values, potential data quality issues |
| Rating | float | Missing values, potential data quality issues |
| NOX | float | None |
| RM | float | have outliers |
| AGE | float | None |
| DIS | float | Missing values, potential data quality issues |
| RAD | int | Missing values, |
| TAX | float | None |
| PTRATIO | float | Missing values, potential data quality issues |
| B | float | have outliers |
| LSTAT | float | None |
| MEDV | float | None |

* **CRIM** - per capita crime rate by town
* **ZN** - proportion of residential land zoned for lots over 25,000 sq.ft.
* **INDUS** - proportion of non-retail business acres per town.
* **CHAS** - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
* **NOX** - nitric oxides concentration (parts per 10 million)
* **RM** - average number of rooms per dwelling
* **AGE** - proportion of owner-occupied units built prior to 1940
* **DIS** - weighted distances to five Boston employment centres
* **RAD** - index of accessibility to radial highways
* **TAX** - full-value property-tax rate per $10,000
* **PTRATIO** - pupil-teacher ratio by town
* **B** - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
* **LSTAT** - % lower status of the population
* **MEDV** - Median value of owner-occupied homes in $1000's

Outliers are the values that do not fit the data type or range of the data. Outliers can be of different data types and we can detect and treat outliers using the methods mentioned below. An everyday example of an outlier can be the salary disparity between an employee of an organisation and the CEO of the same organisation.

Outliers (Cause) -

* Data-entry errors - These errors include decimal misplacements, value misplacement, and errors that occur during data entry.
* Measurement errors - These errors occur when there is a discrepancy in the measurement process of the data. These values may not always be wrong but induce errors in models. (Shmueli et al., 2019)
* Missing Values - The errors may occur due to misinterpretation of data by humans, equipment error, losing of samples while recording values, and others.

Outliers (Review) -

* Box Plot - Outliers can easily be visualized with the help of box plots. A box plot divides and maps the data into 4 quartiles (0-25%, 25%-50%, 50%-75% and 75%-100%).
* Sorting Values -It is used to anticipate the locations of unsorted data components within a sorted sequence. We can sort a column in python with the help of the sort\_values() function. (Shmueli et al., 2019)
* Statistical Methods - quantile(), z score
* isnull() - We can find null values with the isnull() command
* isnan() - We can find NaN values with the isnan() command
* Manually setting a range - Instead of using statistical methods to find the min-max limits, we can manually

Outliers (Treatment) -

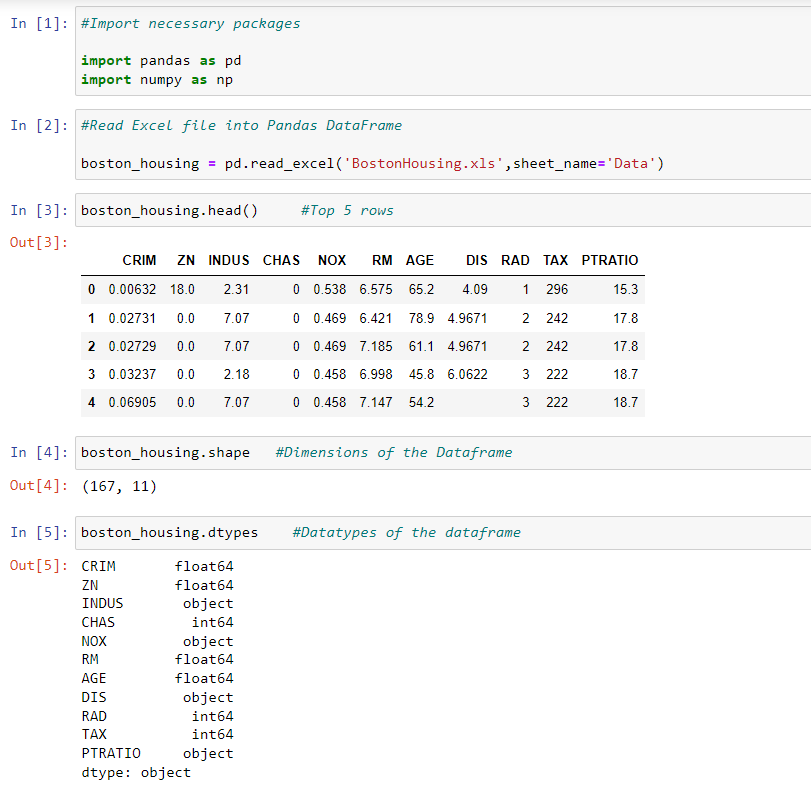
* Outlier trimming: This is not the most effective method for treating outliers since we can lose a lot of data by using this method and can have an impact on the dataset.
* Mean, Median imputation: The process of replacing missing/NaN values.
* Outlier capping: It is used to set a limit above or below a particular value for the field. (Goyal, 2022)

# 

# **PART B:- Data Pre-processing**

Here, we will be using the Boston Housing Dataset. (The Boston House-price Data, n.d.)

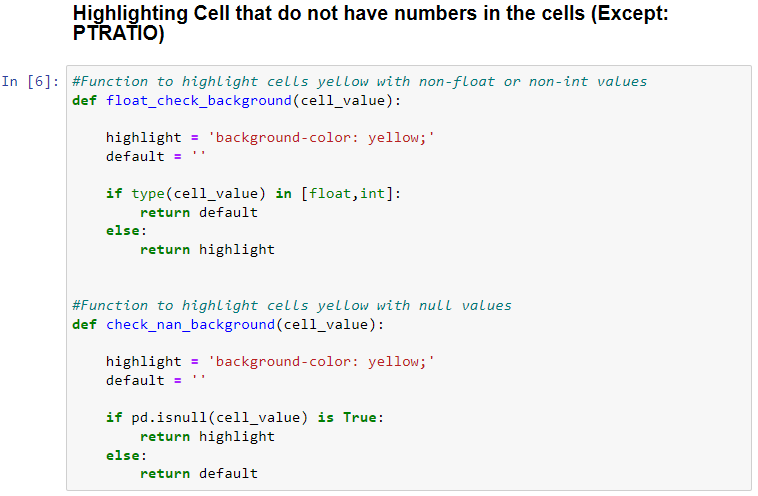
**Importing excel file to Pandas Data Frame**



Here we import the necessary packages and read the excel file using Pandas Data Frame.

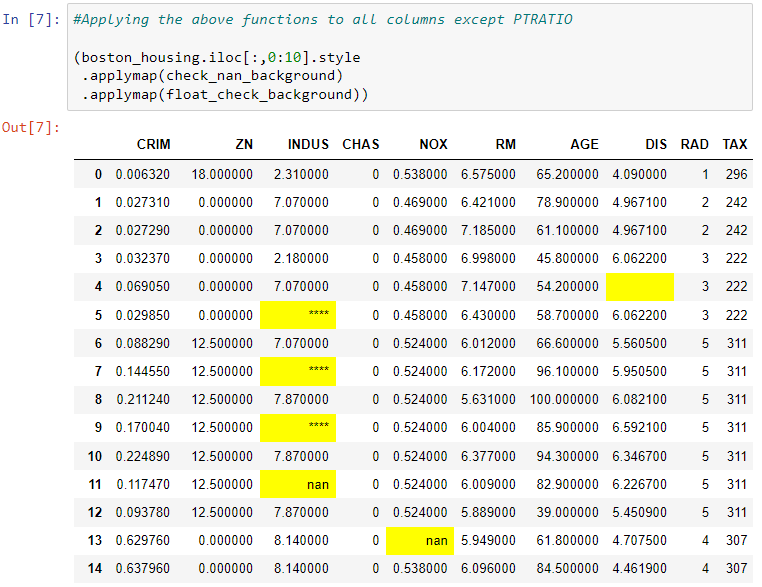
**B.1: Handling Missing Data**

**Creating Functions to highlight missing and wrong values with a yellow background**

****

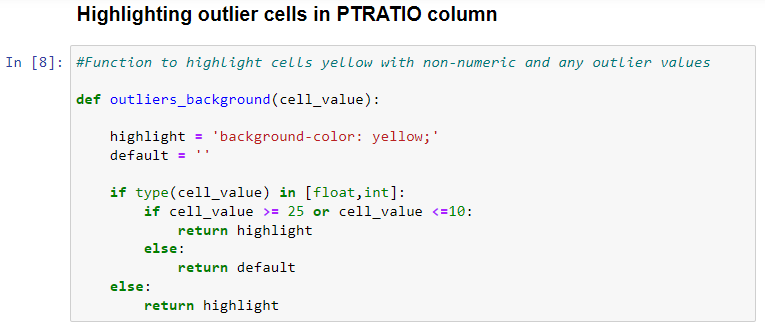
The above function float\_check\_background returns yellow background to the cells that do not have any number in them along with that check\_nan\_background returns yellow background to the cells with null or NaN values with the code referenced from (Felipe, 2022):

In the image below we run the above two functions to apply that function to all the columns except ‘PTRATIO’

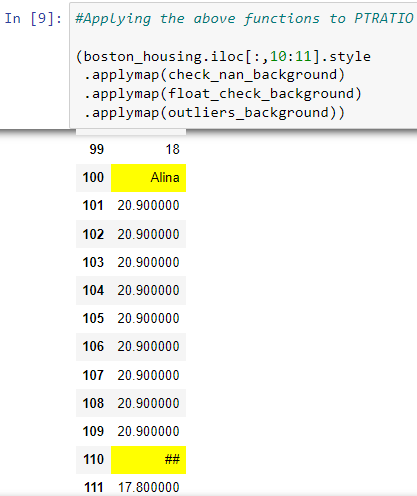
****

**B.2:** **Find possible outliers**

**Creating a function to highlight outliers with a yellow background in the PTRATIO column**



In the above function, we have set the limits for the PTRATIO column, where the values below 10 and above 25 will be highlighted in yellow. (Felipe, 2022)



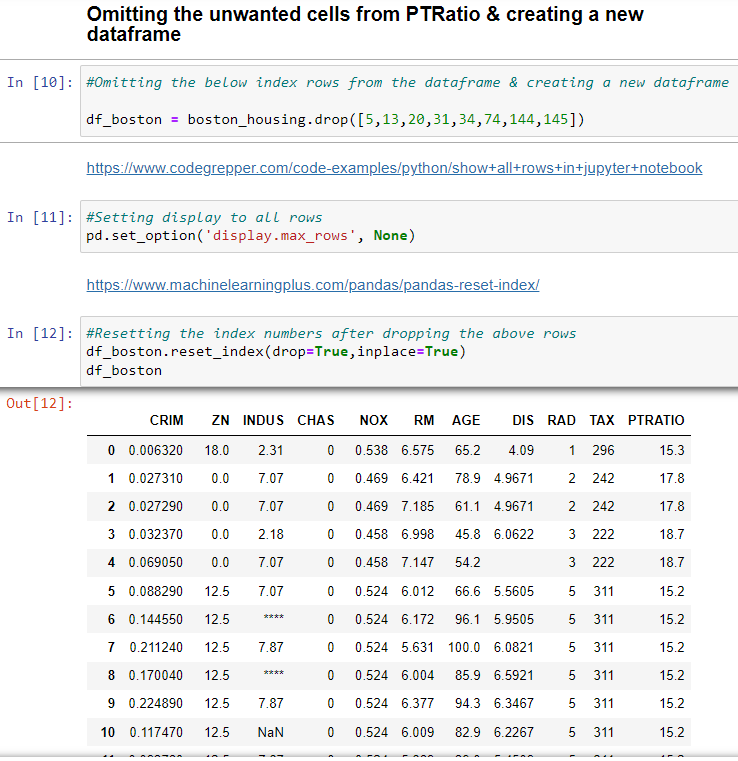
Here we were able to identify different outliers in the PTRATIO column

(a) **Typing non-numeric value:** Identified in index 100, 110, 117

(b) **Shift in decimal place while data entry error:** Identified in index 54

(c) **Genuine case of an outlier:** Identified in index 5, 13, 20, 31, 34, 74, 144, 145

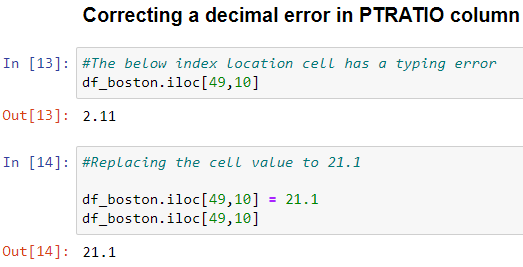
**B.3: Omission and imputation on python**



In the above image, we drop the rows that we identified as ‘genuine outliers’ and would not have an impact on the data if removed. To display all the rows we used the pandas function set\_index.

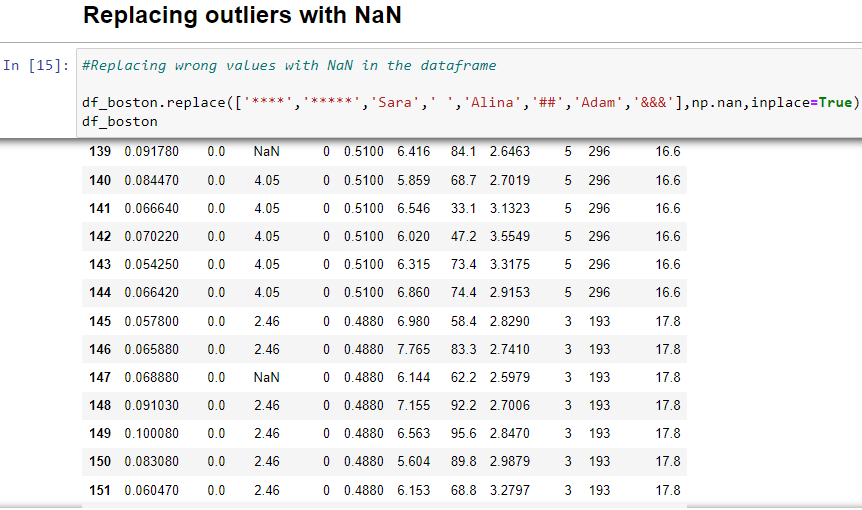
Once the rows are removed we lose the index of those columns as well so we reset the index of the whole data frame to make further analysis easier.(MachineLearningPlus, 2022)

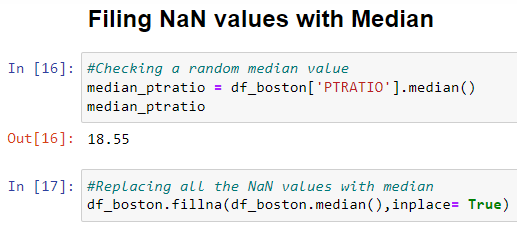
In the below image we change the value of the cell we identified as a decimal typing error to 21.1 using iloc from pandas data frame



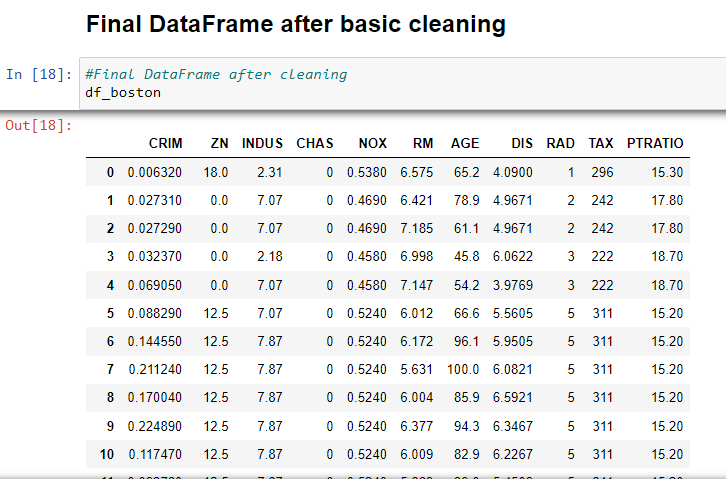
Next, we identified all the other outliers and replaced them with NaN for them to be imputed by the median of the columns.

**B.4: Substituting the missing Data by NaN**





**FINAL RESULT - CLEAN DATASET**



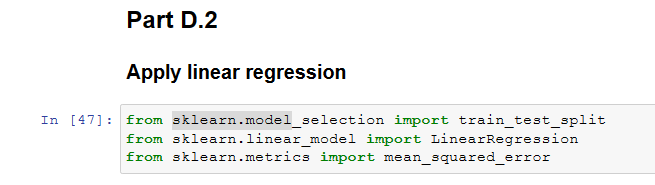
# **Part D:- linear regression**

To apply linear regression in the Boston Housing dataset, we need to choose a dependent variable (target) and an independent variable (feature).

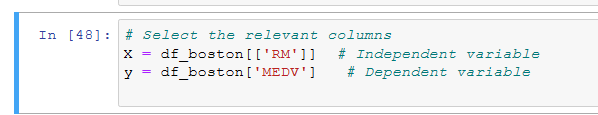
Let's say we want to predict the median value of owner-occupied homes (MEDV) based on the average number of rooms per dwelling (RM).

Here is a simple example for applying linear regression in our data set :

Step 1-: we need to import the the sklearn model



Step 2-: we need to Select the relevant columns Independent variable and Dependent variable from the cleaned data.



Step 3-: Initialize the linear regression model

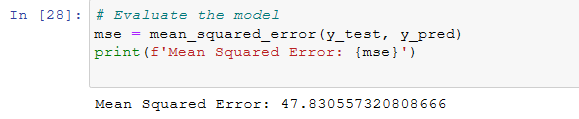


Step 4-: we Fit the model on the training data

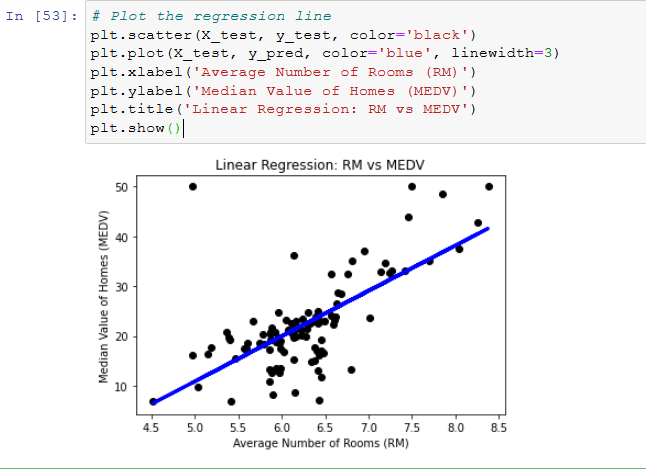
Step 5-: Make predictions on the data



Step 6:- Evaluate the model



step 7:- Plot the regression line



# **REFERENCES**

Felipe. (2022, May 7). *Pandas dataframe examples: Styling cells and conditional formatting*. queirozf.com. Retrieved September 19, 2022, from https://queirozf.com/entries/pandas-dataframe-examples-styling-cells-and-conditional-formatting

Frantic Fowl. (2020, February 25). *Show all rows in Jupyter notebook code example*. Show all rows in jupyter notebook Code Example. Retrieved September 19, 2022, from https://www.codegrepper.com/code-examples/python/show+all+rows+in+jupyter+notebook

Goyal, C. (2022, August 25). *How to detect and remove outliers: Outlier Detection and Removal*. Analytics Vidhya. Retrieved September 19, 2022, from https://www.analyticsvidhya.com/blog/2021/05/feature-engineering-how-to-detect-and-remove-outliers-with-python-code/

MachineLearningPlus. (2022, March 8). *Pandas reset index - how to reset the index and convert the index to a column?* Machine Learning Plus. Retrieved September 19, 2022, from https://www.machinelearningplus.com/pandas/pandas-reset-index/

Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2020). *Data mining for Business Analytics: Concepts, techniques and applications in Python*. John Wiley & Sons, Inc.

*The Boston house-price data*. (n.d.). http://lib.stat.cmu.edu/