**Assignment 1**

By Harsh Sharma

**Part II. Data Exploration**

In Part II, you will have hands-on data exploration. You can use any machine learning API or tool, or any program language. Answer to each of the following questions and also submit your program codes used for Part II.

1. **Check there are any missing values in the dataset. If any, report it.**

**Solution:** No, there are no missing values in the dataset.

1. Report the summary statistics for the TAX feature (full-value property-tax rate per $10,000) with **(1)** minimum, maximum, and range, **(2)** mean and media, **(3)** variance and standard deviation, **(4)** 1st quartile and 3rd quartile, **(5)** inter-quartile range, **(6)**12th percentile.

Solution:

Summary Statistics for TAX feature:

Minimum: 187.0

Maximum: 711.0

Range: 524.0

Mean: 408.2371541501976

Median: 330.0

Variance: 28404.75948812273

Standard Deviation: 168.53711605495903

1st Quartile: 279.0

3rd Quartile: 666.0

Inter-quartile Range: 387.0

12th Percentile: 243.6

1. Show (1) a **histogram** for each numerical feature including the target feature, MEDV. **(2)** Is there any features which show “bimodal” distributions?

Solution:

The histograms of each numerical feature are:

A graph of a graph

Description automatically generated

A graph of a number of individuals

Description automatically generated

A graph of blue bars

Description automatically generated

A graph of blue bars

Description automatically generated

A graph of a graph

Description automatically generated

A graph of age and age

Description automatically generated

A graph of dis

Description automatically generated

A graph of a graph

Description automatically generated

A graph of tax

Description automatically generated

A graph of a number of columns

Description automatically generated with medium confidence

A graph of a number of people

Description automatically generated with medium confidence

A graph of a number of bars

Description automatically generated with medium confidence

A graph of a patient's height

Description automatically generated with medium confidence

**3 (2)** Is there any features which show “bimodal” distributions?

Upon visually inspecting the histograms I could find these features to be bimodal features: TAX, RAD. Also, INDUS could be one.

1. Show (1) a **scatter plot matrix** of numeric features from the dataset to check for correlation between features. (2) Which feature pairs show positive correlation? (3) Which feature pairs show negative correlation?

A grid of blue lines

Description automatically generated

Feature pairs with positive correlation:

[('CRIM', 'RAD'), ('CRIM', 'TAX'), ('ZN', 'DIS'), ('INDUS', 'NOX'), ('INDUS', 'AGE'), ('INDUS', 'RAD'), ('INDUS', 'TAX'), ('INDUS', 'LSTAT'), ('NOX', 'INDUS'), ('NOX', 'AGE'), ('NOX', 'RAD'), ('NOX', 'TAX'), ('NOX', 'LSTAT'), ('RM', 'MEDV'), ('AGE', 'INDUS'), ('AGE', 'NOX'), ('AGE', 'TAX'), ('AGE', 'LSTAT'), ('DIS', 'ZN'), ('RAD', 'CRIM'), ('RAD', 'INDUS'), ('RAD', 'NOX'), ('RAD', 'TAX'), ('TAX', 'CRIM'), ('TAX', 'INDUS'), ('TAX', 'NOX'), ('TAX', 'AGE'), ('TAX', 'RAD'), ('TAX', 'LSTAT'), ('LSTAT', 'INDUS'), ('LSTAT', 'NOX'), ('LSTAT', 'AGE'), ('LSTAT', 'TAX'), ('MEDV', 'RM')]

Feature pairs with negative correlation:

[('ZN', 'INDUS'), ('ZN', 'NOX'), ('ZN', 'AGE'), ('INDUS', 'ZN'), ('INDUS', 'DIS'), ('NOX', 'ZN'), ('NOX', 'DIS'), ('RM', 'LSTAT'), ('AGE', 'ZN'), ('AGE', 'DIS'), ('DIS', 'INDUS'), ('DIS', 'NOX'), ('DIS', 'AGE'), ('DIS', 'TAX'), ('TAX', 'DIS'), ('PTRATIO', 'MEDV'), ('LSTAT', 'RM'), ('LSTAT', 'MEDV'), ('MEDV', 'PTRATIO'), ('MEDV', 'LSTAT')]

**5.** Conduction a **heatmap** which shows correlation values for every feature pairs.

A screenshot of a graph

Description automatically generated

**6.** Conduct **standardization** on all numeric features except the target feature, MEDV*.*

Standardized Dataset:

This contains only the first 7 values of the entire dataset. You can refer to jupyter notebook which I’ve shared to see the entire standardized dataset.

CRIM ZN INDUS NOX RM AGE DIS

0 -0.419782 0.284830 -1.287909 -0.144217 0.413672 -0.120013 0.140214 \

1 -0.417339 -0.487722 -0.593381 -0.740262 0.194274 0.367166 0.557160

2 -0.417342 -0.487722 -0.593381 -0.740262 1.282714 -0.265812 0.557160

3 -0.416750 -0.487722 -1.306878 -0.835284 1.016303 -0.809889 1.077737

4 -0.412482 -0.487722 -1.306878 -0.835284 1.228577 -0.511180 1.077737

5 -0.417044 -0.487722 -1.306878 -0.835284 0.207096 -0.351157 1.077737

6 -0.410243 0.048772 -0.476654 -0.265154 -0.388411 -0.070229 0.839244

RAD TAX PTRATIO B LSTAT MEDV MEDV CHAS

0 -0.982843 -0.666608 -1.459000 0.441052 -1.075562 0.159686 24.0 0

1 -0.867883 -0.987329 -0.303094 0.441052 -0.492439 -0.101524 21.6 0

2 -0.867883 -0.987329 -0.303094 0.396427 -1.208727 1.324247 34.7 0

3 -0.752922 -1.106115 0.113032 0.416163 -1.361517 1.182758 33.4 0

4 -0.752922 -1.106115 0.113032 0.441052 -1.026501 1.487503 36.2 0

5 -0.752922 -1.106115 0.113032 0.410571 -1.043322 0.671222 28.7 0

6 -0.523001 -0.577519 -1.505237 0.426798 -0.031268 0.039964 22.9 0