Module-03, Python for Data Analysis

Real Data Example(Housing dataset)

Dostdar Ali
Instructor

Data science and Artificial Intelligence
3-Months Course
at
Karakaroum international Univrsity

January 16, 2024

Table of Contents

1. Loading the Housing dataset into a data frame

2. Data Description

3. Exploratory data analysis (EDA)

Loading the Housing dataset into a data frame

- We will load the Housing dataset using the pandas read...csv function, which is fast and versatile and a recommended tool for working with tabular data stored in a plaintext format.
- The features of the 506 examples in the Housing dataset have been taken from the original source that was previously shared on https://archive.ics.uci.edu/ml/ datasets/Housing and summarized below,
- 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 and 0 otherwise)

Loading the Housing dataset into a data frame

- We will load the Housing dataset using the pandas read...csv function, which is fast and versatile and a recommended tool for working with tabular data stored in a plaintext format.
- The features of the 506 examples in the Housing dataset have been taken from the original source that was previously shared on https://archive.ics.uci.edu/ml/ datasets/Housing and summarized below,
- 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 and 0 otherwise)

Loading the Housing dataset into a data frame

- We will load the Housing dataset using the pandas read...csv function, which is fast and versatile and a recommended tool for working with tabular data stored in a plaintext format.
- The features of the 506 examples in the Housing dataset have been taken from the original source that was previously shared on https://archive.ics.uci.edu/ml/ datasets/Housing and summarized below,
- 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 and 0 otherwise)

Loading the Housing dataset into a data frame

- We will load the Housing dataset using the pandas read...csv function, which is fast and versatile and a recommended tool for working with tabular data stored in a plaintext format.
- The features of the 506 examples in the Housing dataset have been taken from the original source that was previously shared on https://archive.ics.uci.edu/ml/ datasets/Housing and summarized below,
- 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 and 0 otherwise)

Loading the Housing dataset into a data frame

- We will load the Housing dataset using the pandas read...csv function, which is fast and versatile and a recommended tool for working with tabular data stored in a plaintext format.
- The features of the 506 examples in the Housing dataset have been taken from the original source that was previously shared on https://archive.ics.uci.edu/ml/ datasets/Housing and summarized below,
- 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 and 0 otherwise)

Loading the Housing dataset into a data frame

- We will load the Housing dataset using the pandas read...csv function, which is fast and versatile and a recommended tool for working with tabular data stored in a plaintext format.
- The features of the 506 examples in the Housing dataset have been taken from the original source that was previously shared on https://archive.ics.uci.edu/ml/ datasets/Housing and summarized below,
- 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 and 0 otherwise)

Data Description

- NOX: Nitric oxide 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 centers
- 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 [people of African American descent] by town
- LSTAT: Percentage of lower status of the population
- MEDV: Median value of owner-occupied homes in 1000s

Data Description

- NOX: Nitric oxide 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 centers
- 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 [people of African American descent] by town
- LSTAT: Percentage of lower status of the population
- MEDV: Median value of owner-occupied homes in 1000s

Data Description

- NOX: Nitric oxide 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 centers
- 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 [people of African American descent] by town
- LSTAT: Percentage of lower status of the population
- MEDV: Median value of owner-occupied homes in 1000s

Data Description

- NOX: Nitric oxide 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 centers
- 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 [people of African American descent] by town
- LSTAT: Percentage of lower status of the population
- MEDV: Median value of owner-occupied homes in 1000s

Data Description

- NOX: Nitric oxide 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 centers
- 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 [people of African American descent] by town
- LSTAT: Percentage of lower status of the population
- MEDV: Median value of owner-occupied homes in 1000s

Data Description

- NOX: Nitric oxide 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 centers
- 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 [people of African American descent] by town
- LSTAT: Percentage of lower status of the population
- MEDV: Median value of owner-occupied homes in 1000s

Data Description

- NOX: Nitric oxide 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 centers
- 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 [people of African American descent] by town
- LSTAT: Percentage of lower status of the population
- MEDV: Median value of owner-occupied homes in 1000s

Data Description

- NOX: Nitric oxide 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 centers
- 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 [people of African American descent] by town
- LSTAT: Percentage of lower status of the population
- MEDV: Median value of owner-occupied homes in 1000s

Data Description

- NOX: Nitric oxide 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 centers
- 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 [people of African American descent] by town
- LSTAT: Percentage of lower status of the population
- MEDV: Median value of owner-occupied homes in 1000s

Data Description

- NOX: Nitric oxide 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 centers
- 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 [people of African American descent] by town
- LSTAT: Percentage of lower status of the population
- MEDV: Median value of owner-occupied homes in 1000s

Exploratory data analysis (EDA)

- Exploratory data analysis (EDA) is an important and recommended first step prior to the training of a machine learning model. In the rest of this section, we will use some simple yet useful techniques from the graphical EDA toolbox that may help us to visually detect the presence of outliers, the distribution of the data, and the relationships between features.
- First, we will create a scatterplot matrix that allows us to visualize the pair-wise correlations between the different features in this dataset in one place. To plot the scatterplot.
- As we can see in the following visuals, the scatterplot matrix provides us with a useful graphical summary of the relationships in a dataset.

Exploratory data analysis (EDA)

- Exploratory data analysis (EDA) is an important and recommended first step prior to the training of a machine learning model. In the rest of this section, we will use some simple yet useful techniques from the graphical EDA toolbox that may help us to visually detect the presence of outliers, the distribution of the data, and the relationships between features.
- First, we will create a scatterplot matrix that allows us to visualize the pair-wise correlations between the different features in this dataset in one place. To plot the scatterplot.
- As we can see in the following visuals, the scatterplot matrix provides us with a useful graphical summary of the relationships in a dataset.

Exploratory data analysis (EDA)

- Exploratory data analysis (EDA) is an important and recommended first step prior to the training of a machine learning model. In the rest of this section, we will use some simple yet useful techniques from the graphical EDA toolbox that may help us to visually detect the presence of outliers, the distribution of the data, and the relationships between features.
- First, we will create a scatterplot matrix that allows us to visualize the pair-wise correlations between the different features in this dataset in one place. To plot the scatterplot.
- As we can see in the following visuals, the scatterplot matrix provides us with a useful graphical summary of the relationships in a dataset.

Great Job
Thank you