Activity 3.01: Generating Statistics from a CSV File

Suppose you are working with the Boston Housing price dataset. This dataset isfamous in the machine learning community. Many regression problems can beformulated, and machine learning algorithms can be run on this dataset. You willperform a basic data wrangling activity (including plotting some trends) on thisdataset (.csv file) by reading it as a pandas DataFrame. We will perform a fewstatistical operations on this DataFrame.NoteThe Boston Housing dataset can be found here: https://packt.live/2ZPdvc2.The pandas function for reading a CSV file is read\_csv.

These steps will help you complete this activity:

1. Load the necessary libraries.

2. Read in the Boston Housing dataset (given as a .csv file) from thelocal directory.

3. Check the first 10 records. Find the total number of records.

4. Create a smaller DataFrame with columns that do not include CHAS, NOX, B,and LSTAT:Chas: Charlse River Dummy variableNox: Nitric Oxide concentrationB: Proportion of the population that is African AmericanLSTAT: Percentage of lower-income population

5. Check the last seven records of the new DataFrame you just created.

6. Plot the histograms of all the variables (columns) in the new DataFrame.

7. Plot them all at once using a for loop. Try to add a unique title to the plot.

8. Create a scatter plot of crime rate versus price.

9. Plot log10(crime) versus price.

10. Calculate some useful statistics, such as mean rooms per dwelling, median age, mean distances to five Boston employment centers, and the percentage of houses with a low price (< $20,000). Hint: To calculate the percentage of houses below $20,000, create a pandas series with the PRICE column and directly compare it with 20. You can do this because pandas series is basically a NumPy array and you have seen how to filter NumPy array in the exercises in this chapter.

The output should be as follows:

Mean rooms per dwelling: 6.284634387351788

Median age: 77.5

Mean distances to five Boston employment centers: 3.795042687747034

Percentage of houses with a low price (<$20,000): 41.50197628458498

Activity 4.01: Working with the Adult Income Dataset (UCI)

In this activity, we will detect outliers in the Adult Income Dataset from the UCImachine learning portal <https://packt.live/2N9lRUU>.

You can find a description of the dataset https://packt.live/2N9lRUU. We will use theconcepts we've learned throughout this chapter, such as subsetting, applying user-defined functions, summary statistics, visualizations, boolean indexing, and groupby to find a whole group of outliers in a dataset. We will create a bar plot to plot thisgroup of outliers. Finally, we will merge two datasets by using a common key.

These are the steps that will help you solve this activity:

1. Load the necessary libraries.

2. Read the adult income dataset from the following URL: <https://packt.live/2N9lRUU>.

3. Create a script that will read a text file line by line.

4. Add a name of Income for the response variable to the dataset.

5. Find the missing values.

6. Create a DataFrame with only age, education, and occupation by using sub setting.

7. Plot a histogram of age with a bin size of 20.

8. Create a function to strip the whitespace characters.

9. Use the apply method to apply this function to all the columns with string values, create a new column, copy the values from this new column to the old column, and drop the new column.

10. Find the number of people who are aged between 30 and 50.

11. Group the records based on age and education to find how the mean age is distributed.

12. Group by occupation and show the summary statistics of age. Find which profession has the oldest workers on average and which profession has its largest share of the workforce above the 75th percentile.

13. Use subset and groupBy to find the outliers.

14. Plot the outlier values on a bar chart. It should look something like this:

15. Merge the two DataFrames using common keys to drop duplicate values.