ELEC-E7130 Assignment 5. Data Analysis

Esa Hyytiä Markus Peuhkuri Tran Thien Thi
Weixuan Jiang César Iván Olvera Espinosa Yu Fu

Prerequisites

1. To complete this assignment, students require a basic understanding of Python or R including data import, data processing and visualization, and data inference.

   If you are not very familiar with using Python to plot, you can

   a. Take a look at Matplotlib library and matplotlib for data science in Python.

   b. Also use any other tools you are familiar with for analyzing

Learning outcomes

At the end of this assignment, students should be able to

1. Understand the techniques available in data analysis and visualization.
2. Know what numbers describe characteristics best
3. Make graphs that represent information in an easy-to-understand way.
4. Analyze the data set from different perspectives and characteristics such as correlation, stability, trend, seasonality or stationarity.

Introduction

This assignment contains five tasks very helpful to analyze data sets from different perspectives. Please read all instructions before starting.

- Task 1: Understanding different plots
- Task 2: Plot data
- Task 3: Link loads
- Task 4: Pairs plot
- Task 5: Understanding time series concepts

All these exercises can be done using Python or any other available software (such as R, Matlab) as long as the results are consistent and correct.

Recommendations:

- Take a look to the lecture "Data Analysis" on lecture notes section or there are several sources (books, articles, so on) on internet regarding data visualization and analysis that can be useful as a guide, such as the books called An Introduction to Statistical Learning Fundamentals of Data Visualization or Data Analytics.
- According to the chosen tool (Python or R), take a look at the *different cheat sheets* available on internet related to it as well as the libraries/packages (pandas, matplotlib) including the most useful information related to syntax, functions, variables, conditions, formulas, and more.

All the data files required in this exercise are found from `/work/courses/unix/T/ELEC/E7130/general/r-data` directory. Path is also as `RDATA` environment variable if you have sourced the `use.sh` file.

Note: You may type the command `kinit` before accessing to the directory to avoid issues related to the permissions.

```
$ source /work/courses/unix/T/ELEC/E7130/general/use.sh
$ cd $RDATA
$ ls
...
```

Task 1: Understanding different plots

In the first task, **explain the following plots briefly**. For example, what the *y-axis/x-axis represent*, the appropriate *usage scenarios*, and the *limitations* of each plot.

1. Autocorrelation plot
2. Boxplot
3. Lag plot
4. Parallel plot
5. Scatter Matrix Plot

Report, task 1:

- Discussion on different plots.

Task 2: Plot data

In this task, graph various kinds of plots in linear scale and logarithmic scale, and then analyze them.

Download the file `flows.txt` contains values of flow lengths in bytes captured from a network in order to study the flow length variable using your favorite software.

Provide **concise answers to the following sections**.

1. Plot the flow data using:

   - Scatterplot (Number of observations will reside on X-axis)
   - Histogram (Using a suitable number of bins)
   - Boxplot
   - Empirical CDF of the variable

   Note: Provide the plots including the commands or functions used to plot the data in your report.

2. Describe the distributions **choosing variables**. In terms of summary data, it means the expression variable indicates the measure of central tendency of a distribution, such as *mean*, *median*, *mode*, *max*, *min*, etc.

   - First, choose *the first variable* and explain the reason
   - Then, *the second variable* and explain the reason
   - Finally, *the third variable* and explain the reason

   Note: Provide the commands used to get the results as well as *explain the reasons for your selections* based on the information you gathered during the previous section.

3. **Replot data using logarithmic values** and explain why and when it is more suitable to use the logarithmic values?

Finally, **make conclusions** about whether there are best methods to describe the data and why, and briefly explain what the behavior of the flow data is based on the methods used.

Report, task 2

- Provide different plots of `flows.txt`
- Summarize the distribution using various numbers of variables
- Provide different plots of `flows.txt` using logarithmic values
- Conclusions based on the flow information

  Tips:

  - Useful Python functions include `plt.scatter()`, `plt.hist()`, `plt.boxplot()`, `ecdf()`.
  - Useful R functions include `plot()`, `hist()`, `boxplot()`, `ecdf()`, `log()`.

Task 3: Link loads

For the task 3, produce different kinds of plots that could be useful for analyzing network data such as stability and correlation.

Download the files `linkload-*X*.txt` which contain link loads information (in bits per second) of different links in intervals of one second.

1. **Plot the data of each link** through:

- Time plot
- Lag plot (lag-1)
- Correlogram (i.e. autocorrelation plot)

2. **Inspect the data results**, especially for stability and whether previous values contribute to the present value (short and long-range memory)

3. **Explain your own understanding of each data set** (i.e. each link)

   Tips:

   - Useful Python functions could be `plot()`, `lag_plot()`, `autocorrelation_plot()`.
   - Useful R functions could be `lag.plot()`, `acf()`.

Report, task 3

- Plots according to instructions
- Data inspection results
- Conclusions of each data set.

Task 4: Pairs plot

In the case of this task, graph a pairs plot for each one of the variables contained in the data set to verify the correlation and relation between them.

Download the `bytes.csv` dataset contains time series data of 4 relevant columns: transmitted bytes, received bytes, transmitted packets, and received packets.

1. **Plot the pairs plot** for such values.

2. **Answer the following questions**:

- Which variables correlate most to each other?
- Let's assume that you decide to remove one particular column to reduce the computation load of data handling. Based on the pairs plot, what would the column be, and why?

  Tips:

- Useful Python function could be `scatter_matrix()`.
- Useful R function could be `pairs()`.

Report, task 4

- Pair plots and analysis.
- Answer to the questions.

Task 5: Understanding time series concepts

For this task, visualize the data set by creating a time series plot, which helps in understanding the patterns and trends over time.

Download the `querytime.csv` dataset which depicts the query time to a distant website with a server located in Belgium.

1. Plot the time series.
2. By observing and analyzing the plot, **answer the following questions**:

- Is there any trend or seasonality?
- Is the time series stationary?

Report, task 5

- Time series plot.
- Answer both questions with reasoning.

Grading standard

To pass this course, you need to achieve at least 15 points in this assignment. And if you submit the assignment late, you can get a maximum of 15 points.

You can get up to 30 points for this assignment:

Task 1

- Explain the different types of plots required for the exercise. (5p)

Task 2

- According to the requirements of the task, plot using original value and the logarithmic values separately. (8p)
- Use different numbers of numbers to represent the distribution. (3p)
- Summarize based on the exercises done before. (1p)
- Explain the behavior of the data set (1p)

Task 3

- Draw plots for each link as required. (4p)

- Analyze data based on plots. (1p)
- Summarize the four data sets. (1p)

Task 4

- Plot a pairing plot as required. (1p)
- Answer the questions raised in the exercise. (2p)

Task 5

- Plot the time series. (1p)
- Answer two questions based on your own understanding. (2p)

The quality of the report (bonus 2p)

The instruction of assignment

For the assignment, your submission must contain (Please don't contain original data in your submission):

- A zip file that includes your codes and scripts.
- A PDF file as your report.

Regarding the report, your report must have:

- A cover page indicating your name, student ID and your e-mail address.
- The report should include a description of measurements, a summary of the results and conclusions based on the results.
- An explanation of each problem, explain how you solved it and why you did it.