Cybersecurity Group Assignment 1 Report

CMPT 318 - Fall 2018

Group 8

Aarish Kapila	akapila
Che Jung (Kent) Lee	cjl27
Karan Sharma	ksa95
Razvan Andrei Cretu	rcretu
Yernur Nursultanov	ynursult

Problem

In this assignment, we had to analyze a dataset containing 521860 rows and 9 columns by performing data preprocessing and cleaning to employ several statistical functions.

Methodology and Assumptions

We decided to utilize the DescTools library for geometric mean and mode calculations. This library contains miscellaneous basic statistic functions to speed up the process of initial descriptive analysis of datasets. Although we also found relatively short code snippets online to calculate the geometric mean and mode, we chose this library because its API closely resembled the API for built-in functions like mean and median. For all other statistical operations, we used built-in functions.

When writing our script, the following assumptions were made:

- The given dataset does not require normalization.
- NA values can be omitted from the given dataset.
- Outputs are rounded to two decimal places.
- The local time zone can be used for datetime conversion.
- Minimum and maximum values are required individually for both weekdays and weekends data frames, which are the subsets of the originally given dataset.

Results and Explanation

From Table 1 below, we can infer that the Global Active Power (GAP) distribution has a larger spread than the Global Reactive Power (GRP) distribution, as a high standard deviation suggests that the data is widely spread and a low standard deviation means that the data are clustered closely around the mean.

Table 1: Summary statistics

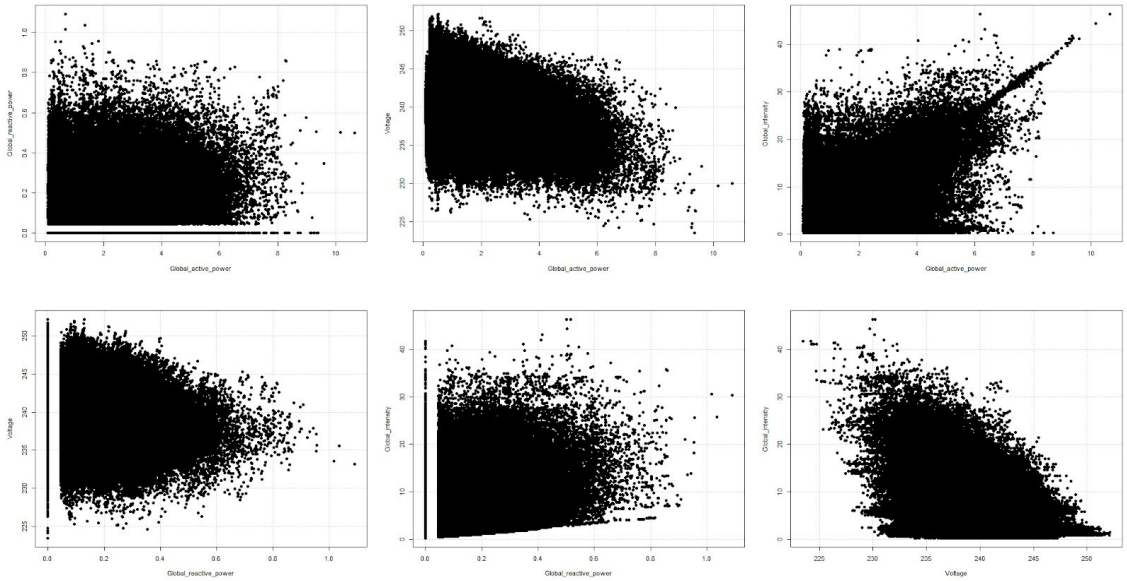
	Global Active Power	Global Reactive Power
Arithmetic mean	1.25	0.12
Geometric mean	0.84	0.00
Median	0.80	0.11
Mode	0.22	0.00
Standard deviation	1.13	0.11

Active Power refers to the power that gets used by appliances attached to meters, whereas Reactive Power is the power which oscillates in a circuit without being consumed (often called “useless” power). As Table 1 shows, both the arithmetic and geometric means of the GAP are greater than those of the GRP, which is intuitive because we expect the actual power consumed to be generally larger than and more variable than the available unused power in an efficient system.

Table 2: Correlation matrix

	Global Active Power	Global Reactive Power	Voltage	Global Intensity
Global Active Power	1	0.16	-0.35	0.73
Global Reactive Power	0.16	1	-0.15	0.28
Voltage	-0.35	-0.15	1	-0.49
Global Intensity	0.73	0.28	-0.49	1

Based on the correlation matrix in Table 2, the strongest correlation is +0.73 between the Global Active Power and Global Intensity. In order to take a closer look at these correlations, we added scatter plots below between all pairs of the 4 features under investigation.



As indicated above, this dataset has a strong positive linear relationship between Global Active Power and Global Intensity. The enlarged scatter plot of the two features is illustrated below, where the shape of a straight line can be seen. The other pairs, however, possess weak to moderate correlations, suggesting that the variables are independent of each other.

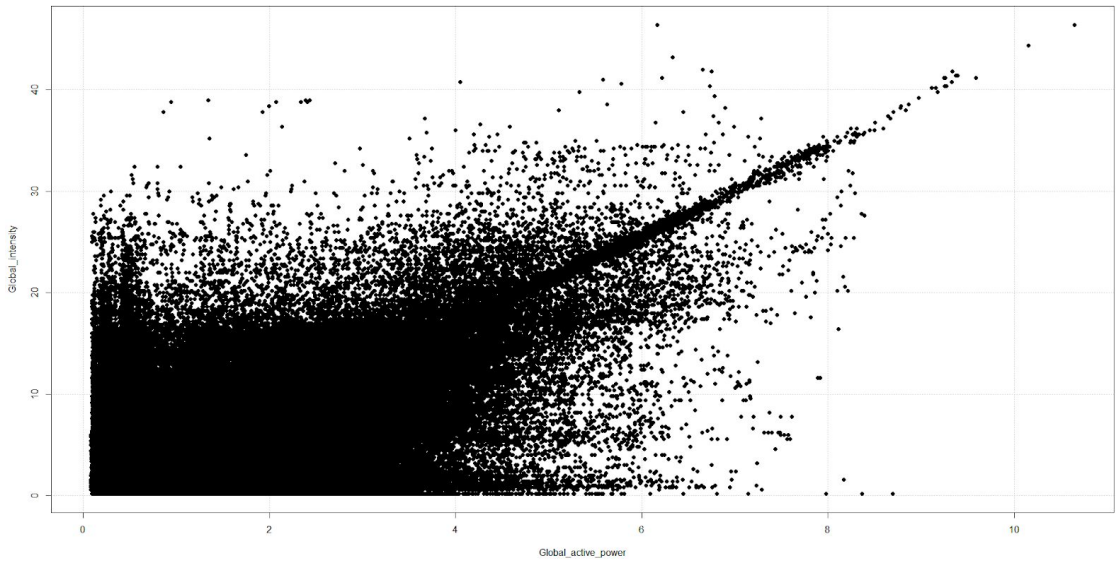


Table 3a: Range on weekdays		
	MIN	MAX
Global Active Power	0.08	9.59
Global Reactive Power	0.00	1.09

Table 3b: Range on weekends		
	MIN	MAX
Global Active Power	0.10	10.65
Global Reactive Power	0.00	0.86

In Table 3a and 3b, we see that the power usage is higher on weekends as opposed to weekdays. This is a reasonable and expected outcome since there are more people staying at home during weekends, resulting in an increase in appliance usage and thus a higher overall power consumption.

Conclusion

Over the course of this assignment, we learned the following regarding R:

- The built-in correlation function already has a pairwise comparison. If parameter y is not given, correlations of columns of x will be calculated.
- The default time zone for POSIXlt is the user's local time zone.
- The date format in the given dataset is reversed, i.e. dd/mm/YYYY instead of YYYY/mm/dd. The latter one is the default input format for POSIX dates.

In summary, this assignment helped us develop a familiarity with writing R code, and installing packages with RStudio. In addition to DescTools, we discovered various other interesting packages that could enable better productivity, such as the dplyr package, which allows for efficient data manipulation and can certainly be used for future work.