Homework 1 - Descriptive statistics

Principles of Statistical Data Analysis 2014-15

For this project, we are interested in the server activity of the website of the Master in Marketing Analysis (`www.mma.ugent.be`). The file `all_logs.csv` contains data about the MMA website's traffic. Every time a file from the website is requested by the user, the server writes a line in the log file. Each observation records the name of a specific file of the resulting MMA website. Visiting one webpage can mean that multiple files are requested to be loaded. We've provided a random sample of 5000 observations drawn from the complete log file. The following variables have been made available:

- **ip**: the ip address of the user
- **file_name**: the filename that was opened
- **status_code**: the HTTP status code of the requested file (see `http://en.wikipedia.org/wiki/List_of_HTTP_status_codes`)
- **object_size**: the size of the file in bytes
- **origin**: the previous website of the visitor. `'-'` means there is no origin. An url starting with `'http://www.mastat.ugent.be/'` means that the origin is internal.
- **date**: a character string depicting the date the file was opened
- **time**: a character string depicting the time the file was opened

The goal of this project is to describe the variables in the dataset and gain a deeper understanding of the usage pattern of the website over the considered time period.

The descriptive analysis should address the following topics:

1. Provide a concise univariate descriptive analysis of every variable (except the ones indicating time). Think carefully about what aspect of each variable could be of interest and accompany your descriptive analysis with an appropriate graph or numerical summary.
2. Next, focus on the evolution of visits and usage over time. In this vein, you should transform or aggregate the data in such a way that a new variable indicating the usage rate can be derived. Try to first gain a clear insight into the structure of the raw data (i.e. how can you define the variables and observation units) and verify whether and how this structure changes after transforming or aggregating the data!

   In particular, we wish to learn how the rate of usage (the number of requests per time unit) differs over time (e.g. changes over hours within a day, days within a week, between weekdays and weekend...). Make sure to account in your targeted description for any gaps in recording time you might discover. Clearly describe how you have attempted/managed to identify such gaps and how you have dealt with these in further analyses.
3. Finally, consider the evolution over time of a variable of your choice. This analysis should be guided by a specific research question you deem relevant for the creators of the website. Make sure to first clearly formulate this research question.

Some guidelines to get you on the right track...

- Before starting, we strongly recommend you to write a protocol for your analysis approach, which you may (have to) amend later on as you are learning more about the data while moving through the consecutive steps.
- Download the help file `HW1-help.R` from Minerva. This already includes hints on how to recode some of the variables in R. Make sure to download the dataset with your group number from the `data` directory on Minerva.
- Explore the data using functions such as `dim()`, `str()`, `head()` and `summary()`. As you can see, the data will need some work. Try to find an appropriate way for transforming all variables into a suitable `class` in R. After cleaning the data, perform the univariate descriptive analyses.
- Derive the outcome variable(s) of interest that represent(s) a rate of usage: the number of files downloaded per unit of time. Does the usage rate evolve over time, and if so, how does it evolve? Don't forget to consider different time scales (i.e. changes over hours within a day, days within a week, between weekdays and weekend...)! Investigate time periods with deviating results and try to suggest explanations for potential deviations.
- Interpret your results carefully and ask yourself which results you think might need further investigation (e.g., because they might raise new questions).
- Reflect on the advantages and disadvantages of the methods used. Think about ways to improve them and consider (or even resort to) alternatives if needed.

Summarize the **key findings** of your study in a report of **at most three pages** (font size no smaller than 10 pt and margins no smaller than 2.5 cm!) refering to a limited number of tables and figures on the following pages (in the appendix). This should contain the sections 'Problem', 'Methods', 'Results' and 'Conclusions and discussion' in the **style of a paper**.

We expect you to work in **groups of 4 or 5**. Please indicate at the end of the report who did what.

Don't forget to provide a separate file including the R-commands you have used. Make sure that (i) your R-code is **clear** (use indentation and comments to explain what you did) and **reproducible** and that (ii) the information provided in the report is **clear and concise** but sufficiently exhaustive (i.e. avoid forcing the reader to consult the appendix or R-code as much as possible)!

You are expected to post your project report (including appendix; in **.pdf**) and R-code (**.R**) on the Minerva course web site of Principles of Statistical Data Analysis by **October 6** at **24:00**. Please send it via the dropbox to both Johan Steen and Els Goetghebeur. The name of the report and of the code files should start with `groupXX_prinstathw1` (where `XX` stands for your group number). This group number should be repeated at the top of your report, which should also list the names of the group members.

Good luck!