Requirements for reports

1. The questions should be addressed in the same order they appear in the assignment. The text of the question MUST be retained and placed before each answer. The working language is English.

2. The answer to a particular question may take a form of a plot, formula etc followed by a brief explanation and **a conclusion**. All your conclusions MUST be justified numerically, i.e., by some computed quantities, plots, etc. The answers do not need to be lengthy but, again, they MUST be convincing in mathematical and statistical sense, i.e., in terms of some quantitative measures. Note that I pay much attention to the conclusions, so try to make it as clear as possible.

3. Each student MUST use a unique data set. It is your responsibility to make sure that no one else is using the same data. **Check the list on my Google Drive and fill in your name and data you are going to use**. Here is a link to that document https://docs.google.com/document/d/1WOAOzDVsJzTHzIrwQ7ksuYhPUt-beCGamUmRapBMPao/edit?usp=sharing.

4. When submitting your report, the subject of your e-mail MUST be `Data Analysis: your name`. Otherwise, you report may get lost or not be processed properly and on time!

5. The due date on Descriptive Statistics Assignment is September 23, 2018.

6. **Late submission**: 25% off for each week after the due date.

7. The answer to a question MUST contain code in `R` or some other language which can be placed in the appendix of the report.

8. Failures to comply with the above rules may reduce your grade for the assignment.

Lecture slides

Download lecture slides https://www.dropbox.com/s/mpf776oc0tl95jo/IntroAndDescriptive.pdf?dl=0

Data sources for the assignment

You have at least three options:

1. Use `getSymbols` command of `quantmod` package to download prices for some stock or commodity (oil, gold, wheat, etc) from Federal Reserve Economic Data cite http://research.stlouisfed.org/fred2/. You may want to try the following commands if download does not start:

   `options(download.file.method="libcurl")` or `options(download.file.method="wget")` or `options(download.file.method="wininet")`

   - Clearly state in your report what kind of data you are using (daily, monthly etc).

- Check for the missing data and remove the respective entries from the dataset, if any. You may use the following script as an example:
```
getSymbols('GOLDAMGBD228NLBM', src='FRED')
idx <- c(1:nrow(GOLD))[is.na(GOLD)]
GOLD <- GOLD[-idx]
```
See also Section 1.3.3 of [1].
- If you did find the missing data, add a comment on that.

2. Use the built-in datasets provided by packages `UsingR`, `MASS` or `ISwR`. See the summary on p. 24 of [1] for listing and handling the available datasets. You may also look through the Problems in [1] to make a choice.

3. Multivariate data – you may try the following:
   - Go to the JSE archive http://ww2.amstat.org/publications/jse/jse_data_archive.htm.
   - If are unsatisfied with the data from the previous source or these have been already picked up by your classmates, visit https://www.census.gov/data/tables/2015/econ/asm/2015-asm.html or, more generally, https://www.data.gov/. However, some minor research and preprocessing may be needed here to get a meaningful and compact dataset.
   - Suggest your own dataset from some other source. Free sources of data are listed here http://guides.emich.edu/data/free-data. Some research is needed.

   Remember that:
   - There MUST be **at least** 20 observations (the more, the better).
   - There MUST be **at least** 3 variables.
   - You may consider time periods (months, years, etc) as observations, i.e., time series will work.

Assignment on Descriptive Statistics

1. **Univariate data**. Get a univariate dataset from sources 1 or 2 and briefly describe it. Using these data,
   (a) Construct a stem-and-leaf and histogram. Impose the empirical density estimate on the histogram. Discuss the results focusing on the shape of the plots and number of modes.
   (b) Compute the mean and median. Based solely on that, conclude whether the distribution is skewed. Find the proportion of the data which are less than the mean value.
   (c) Compute the 1st and 3rd quartiles, the 90th quantile and the mode. Explain the meaning of the obtained quantities. Find the value that cuts off the top 25% of the data.
   (d) Compute the range, the sample standard deviation and the IQR. Construct the boxplot of the data. Comment on the boxplot including skewness, outliers etc.

(e) Check whether the empirical distribution is normal by examining the QQ-plot.

See examples in [1], Section 2.2.

2. **Bivariate data**. Get the bivariate data by: (i) retrieving prices for two stocks from data source 1 (of equal length and periodicity!) or (ii) finding the appropriate bivariate built-in set 2.

   (a) Create side-by-side boxplots. Compare the centers and spreads.

   (b) Draw the scatter plot. Comment on the possible dependence and presence of outliers.

   (c) Compute Pearson's and Spearman's coefficient of correlation. Interpret and compare their values. Are their values consistent with the scatter plot?

   (d) Add the marginal distributions to the scatter plot. For that purpose, use histogram and box plot.

   (e) Depict the bivariate box plot. Comment on the outliers. Remove the outliers, if any, and re-compute the Pearson correlation coefficient.

   (f) Create the convex hull. Remove the observations lying on the hull and re-compute the correlation coefficient.

   For items 2a-2c, see [1], Section 3. For items 2d-2f, see [2], Section 2.2.

3. **Multivariate data**:

   (a) Pick up a dataset which has three variables (from source 2 or 3) and create the bubble plot. Interpret the result. See [2], Section 2.3.

   (b) Use data source 2 or 3. Create the glyph plot of all observations, Section 2.3. Do any stars look alike?

   (c) Use data source 2 or 3. Create the scatter plot matrix and analyze it. See [2], Section 2.4.

References

[1] J. Verzani, Using R for Introductory Statistics, Second Edition, Chapman & Hall/CRC The R Series, Taylor & Francis, 2014.
URL https://books.google.ru/books?id=O86uAwAAQBAJ

[2] B. Everitt, T. Hothorn, An introduction to applied multivariate analysis with R, Springer, New York, 2011.
URL http://dx.doi.org/10.1007/978-1-4419-9650-3