STAT40830 - Adv Data Prog with R (online)

Final Project

Due on Sunday 16th August 2020 11:59pm (IST).

This assessed component corresponds to 60% of the final grade for this module. The project will be marked out of 100 according to the following criteria:

- *20 marks* technical parts of the code;
- *20 marks* coding style (e.g. whether it is easy to read/understand/extend and appropriately commented);
- *20 marks* quality of the graphics (e.g. `ggplot2` usage);
- *20 marks* understanding of the statistical components;
- *10 marks* quality of the interactive graphics (e.g. `ggiraph` and Shiny);
- *10 marks* explanations and exported documents (e.g. RMarkdown)

To complete this assessed component, you should upload on Brightspace:

- an RMarkdown document in html format, containing all the relevant output, an explanation of your work and any code you think is relevant (although this code should be kept concise);
- a zipped file containing your `.Rmd` file and any other relevant scripts or material.

Important
For this assessed component if you have special circumstances causing you to miss the deadline, you will need to follow the official procedure as explained in `http://www.ucd.ie/students/studentdesk/extenuating.html`

You may need some or all of the following R packages:

```
readr
dplyr
magrittr
ggplot2
shiny
```

```
ggiraph
gganimate
reshape2
```

Note:

- when preparing your project, you should follow, as closely as possible, the coding guidelines that have been discussed throughout this module;
- this is a project: differently from a final exam, further reading may become necessary to complete some parts;
- the dataset used contains plenty of missing data. This may result in a number of warnings, which can generally be disregarded. In the case that the missing values make it impossible to complete a task then rows containing NAs may be excluded.

Questions

1. Import the dataset `exo_data.csv` as a tibble. Columns 1, 16, 17, 18, 25 should be characters. Columns 2, 14 should be factors. Column 15 should be integers. The remaining columns should be doubles.
   Note: the file `metadata.txt` contains useful information about this dataset. Also, you may consult `https://en.wikipedia.org/wiki/Exoplanet`
2. Exclude the exoplanets with an unknown method of discovery.
3. Create a graphic which illustrates the relationship between the log-distances from the Sun and the methods of discovery.
4. Create scatterplots of the log-mass versus log-distances, separating by methods of discovery. Hovering with the cursor highlights the point and displays its name, and, if you click, the exoplanet's page on the Open Exoplanet Catalogue will be opened. (paste the id after `http://www.openexoplanetcatalogue.com/planet/`).
5. Rename the radius into `jupiter_radius`, and create a new column called `earth_radius` which is 11.2 times the Jupiter radius.
6. Focus only on the rows where log-earth radius and log-period have no missing values, and perform kmeans with four clusters on these two columns.
7. Add the clustering labels to the dataset through a new factor column called `type`, with levels `rocky`, `hot_jupiters`, `cold_gas_giants`, `others`; similarly to `https://en.wikipedia.org/wiki/Exoplanet#/media/File:ExoplanetPopulations-20170616.png` and produce the scatterplot highlighting these clusters.
8. Use a violin plot to illustrate how these clusters relate to the log-mass of the exoplanet.
9. Transform `r_asc` and `decl` into two new variables that are the same varibales but in values of seconds. Use these as coordinates to represent a celestial map for the exoplanets.

10. Create an animated time series where multiple lines illustrate the evolution over time of the total number of exoplanets discovered for each method up to that year.

11. Create an interactive plot with Shiny where you can select the year (slider widget, with values ≥ 2009) and exoplanet type. Exoplanets appear as points on a scatterplot (log-mass vs log-distance coloured by method) only if they have already been discovered. If type is equal to `all` all types are plotted together.

12. Fit a linear regression model where `log period` is the response variable and the logs of `host_mass`, `host_temp` and `axis` are the covariates (exclude rows that contain at least one missing value). Include an intercept term in the regression model.

13. Include in your RMarkdown document some model summaries and an interpretation of the model you have fit.

14. Embed the Shiny app from (11) in your RMarkdown document.

Please ensure that all documents are uploaded to Brightspace **clearly and on time**. It is students' responsibility in taking care of uploading all assignments. Pending or incomplete submissions after the deadline will be considered late and be penalised according to UCD guidelines, unless proven and valid justification (e.g. doctor's certificate) is provided. **Plagiarism is prohibited**. Please refer to the UCD Plagiarism Policy.

In addition, you must not discuss your answers or attempts on any of the module forums or with your classmates. Questions relating to the assignment can be discussed on the forum, but this will be closely monitored since this work is assessed.

Good luck and enjoy the project.