Data Science and R – Project Proposal

You will work in your group to perform Exploratory Data Analysis (EDA) and provide useful insights on large datasets.

Your project presentation will take place on **Week 14 (Week starting November 28th)**.

The first stage is to find a **large dataset** to work on. You and your group will propose a dataset that you will investigate.

Guidelines for dataset selection:

- The dataset should have at least **50,000 rows and at least 5 columns**. Data could be spread across several subsets. The bigger the better. If the dataset is slightly smaller in its dimensions we can negotiate this. Some datasets have more columns than rows.
- Have a **topic** of interest in mind, is this something you are passionate about, e.g. sports, news, weather, health, etc.?
- What kind of questions would you or people in general be curious about regarding this dataset? And what kind of analysis or tools can R provide in order to answer those questions? E.g. a large movie database might provide insight into what are the most popular movies by year, decade, genres, and users' reviews for certain movies.
- Look for datasets in csv, tsv, txt or Excel format. It would help that you do NOT have to work with other types of data such as images, sound, xml or json, though it is not impossible to work with xml and json in R with parsers.

The main aim would be to **clean, visualise and perform simple analysis** on this data, e.g. linear modelling, time series analysis, clustering, etc. (the course will try to cover one or two simple analyses). However, the emphasis would be on the cleaning and preparation of the data using EDA. It is important that you find a good dataset to work on.

In the proposal you should provide the following (ONE per group):

0) The group number, members and leader.

1) The dataset that you plan to use and its web link if available

2) How are you going to access it? (is it freely available for download?)

3) What format is it in? (csv, txt, etc.)

4) What are its dimensions (size).

5) What do you plan to find out from this data? (Initial ideas)

Submit your proposal (ONE page max) by midnight **Wednesday, November 2nd, 2022**

Your proposal will need to be **approved** by the instructors before you can proceed.

Here are some sites where you can find datasets. You are not limited to just these sites, you may find your data by any means, even private data is welcome but you will have to get approval before you can proceed.

Kaggle: https://www.kaggle.com/datasets **(data less than 3 months old with NO EDA solutions available)**

UCI Machine Learning repository: https://archive.ics.uci.edu/ml/datasets.php

Korean Public Datasets: data.go.kr

Data World: https://data.world/datasets/data

MicroData Integrated Service (Korean) https://mdis.kostat.go.kr/index.do

Google dataset search: https://datasetsearch.research.google.com/

Machine Learning Dataset: https://www.datasetlist.com/

Previous year's project datasets included: Melon, Kickstarter, YouTube analysis, Seoul Public Bike, Google Play Store, Credit card data, Korea divorce data, Firearm suicide, Transport, Weather and a few others.

If for whatever reason, after a week, you find that the dataset is unsuitable to work with, you may change your dataset by the end of Week 11. **NO changes in dataset will be permitted after November 14th.**