Final Project for DATA 606 – Statistical Methods for Data Science

1. **General info**

   The students could form groups and each group consists of at most 4 people. The project is evaluated based on a presentation and a written report. The presentation is to be done in the last two classes. The report is to be uploaded to the dropbox within one week following the last class.

   As the project counts for 50% of the final grade, the detailed split of the evaluation is: 10% of the final grade for the presentation, 40% of the final grade for the report.

2. **Project description**

   In this project, the groups could select the datasets which interest them most from the approved data sources (most of them are free online). The dataset should not be analyzed before (such that the results and codes could be downloaded directly), otherwise the originality of your work will be doubted and the project grade will be adversely affected.

   At the very beginning, the analysts should list several problems which to be studied in the project. The methodologies, which are covered in this course, should be applied as much as possible to the selected datasets. Some conclusions should be drawn from the analysis at the end.

   Students are also encouraged to apply the previously learned techniques (from, e.g., DATA 602 and 603) to the selected datasets. The visualized results are asset.

3. **Report format**

   The project report should be written using R Markdown, Latex or Word. The code, as well as the generated results, should be inserted. The format is quite flexible, the following components are suggested: *background of the problem (or motivation)*, *introduction of the dataset*, *the methodology part and results*, *comparison* and *conclusion*.

   It is also suggested that the lead writer include at the end of the report the detailed job division among the group members.

4. **Some data sources**

   - The Kaggle: `https://www.kaggle.com/datasets`
   - Yelp open dataset: `https://www.yelp.com/dataset`
   - UNICEF Dataset: `https://data.unicef.org/`
   - U.S. Census Bureau: `https://www.census.gov/data.html`
   - Statistics Canada: `https://www.statcan.gc.ca/eng/start`

- European Union Open Data Portal: `https://data.europa.eu/euodp/en/data/`

5. **Timeline**

   - By the end of WEEK ONE, groups should be formed. The students of each group should be emailed to the instructor directly by the group leader.
   - By the end of WEEK TWO, the datasets should be selected. The selected dataset and its source should be sent to the instructor by the group leader.
   - By the end of WEEK FIVE, some preliminary analysis results should be done. For the methodologies covered late in this term (e.g. in week five), a proposal could be given.
   - In WEEK SIX, group-based in-class presentations start. Each presentation will be evaluated by all the other groups as well as the instructor. Again, for some analysis could not be done by that time, a proposal should be given.
   - By the end of WEEK SEVEN, the written report should be all uploaded to the dropbox through D2L.

6. **Friendly reminder**

   As this course is a 6-week course, everything goes fast. PLEASE follow the timeline.