Final Term Project

Data Analysis 2 and Coding with R

MS in Business Analytics,
2020/2021 Fall

Your task is to analyse your research question, based on your chosen dataset. The aim of this assignment is to show your skills in data analysis and coding, learnt during this Fall semester. There are no strict rules - therefore precise grading scheme - for this task. What we are looking in your final term project is the following:

- You show a compact report on an issue - which is easy to read (nicely formatted) and understand (not using complicated technical terms or itemize steps of data analysis).
- The report has a clear message - what we can learn from the data - and there is a proper argument based on learnt statistical tools.
- You can specify properly your research question based on your data.
- You are confident in, what we can learn from the data and what you can not claim based on your dataset.
- You understand the possible challenges of your dataset and can access the uncertainties, which comes from data quality issues. These uncertainties are well articulated when conclusions are made.
- You can show pattern of association between variables, you can transform these variables in order to handle them in a (linear) regression model.
- You understand the nature of your outcome variable and you can use proper model(s) to handle this nature.
  - Within this model you can show you have mastered what we have learnt and can use them in analysis to make your argument more robust.
- You can show how to generalize your results and what are the constraints of this generalization.
- Formatting requirement:
  - pdf and html knitted by rmarkdown.
  - **Length**: max 6 pages of report and then appendix (this can be as long as you wish)
- You shall study the provided materials in details (on moodle: step-by-step for analysis and example for term project) and use them as guidelines.

This assignment is evaluated for both Data Analysis 2: Finding Patterns with Regressions **and** Coding 1: Data Management and Analysis with R.

You need to upload a zip file containing the required file structure and all the files to ceulearning site to Assignment 2. The Readme.md file in the root folder needs to contain your assignment's **public** github repo's url.

Deadline: Sunday, 3 January 2021, 11:55 PM.
Late submission: 1-2 day delay -50%, after that no points.

1 Structure of final term project

1. Create github repo with proper file structure:
   (a) data folder: with raw and clean sub-folders, also contains variables.xlsx and readme.md files.
   (b) codes folder: rmd and if want an .R with a readme which tells in 1-2 sentence which file does what
   (c) docs folder: .html and .pdf generated from .rmd. Here there is no need for readme file.
   (d) out folder *if needed*: contains any output, which generated by the codes: e.g. model comparison

2. **All in your coding files, the data-import commands must be specified by reading the data from your github repo.**
   (a) This is important for us to see whether your code auto-run and we want to avoid possible path-specification problems. You can see examples for this e.g. among Class_12 codes.
   (b) If confidential data → use a simple randomized and unanimous sample and upload that to check if your code runs.

2 Some strong suggestions

- Create rather a newspaper article style of report then an itemized description what you have done.
- Pay attention to format your graphs and tables:
  - Ticks/values of y and x axis (e.g. +scale_y_continuous())
  - Use a unified theme (e.g. theme_bw()) and be consistent during the report to use only one type.
  - Regression tables - name your variables (avoid e.g. ln_gdp_per_capita_sq, rather use $\ln(gppc)^2$.
    * You can use '$' symbol in rmarkdown to make greek letters, powers, ect. This makes your output even fancier.
  - When reporting numbers use 2 digits, if important use max 4 digits.
- Never include code chunks or outputs in your report. Always use some kind of formatting!
- In case of showing a simple regression result, you can still use a simple table e.g. by 'stargazer' or 'texreg'.
- If you use weighted linear regression - there is a different interpretation than simple linear regression.

3 Evaluation

3.1 DA2

Overall you can earn 30 points. For DA2 only pdf/html is going to be evaluated. The main parts that the report should include:

- Executive summary/abstract.

- Introduction: aim of the analysis, what we can learn and possible other reports/analysis connection.
- Data: what is your used data, data quality issues, descriptive statistics and graphs and variable transformation.
- Model:
  - show your main model and what is the interpretation/what we learn from it. If there is two competing model, you can show two of them, but not more.
  - Here you can show the model fit and other technical details - but be brief: only what is important to understand the constraints of your analysis - all the others should be in the appendix.
  - Robustness check: in both causal and prediction analysis you should show some robustness check for your analysis. Use different data, model setup in order to show that your results are robust or not.
- Generalization, external validity and causality:
  - Generalization of your result to the general pattern or population. What we can infer?
  - External validity - analysis for generalization to unknown populations/general patterns. What do you believe about this? IS your results valid for them as well?
  - If causal analysis: how close it is to causality, what would be an omitted variable or how would it be closer to an experimental design?
- Summary - summary of you main results/most important conclusions.

3.2 Coding in R

Overall you can get 40 points.

- Required folder structure with the files as it was asked.
- Rmarkdown file runs and produces the attached pdf/html file(s)
- Readability of the code
- Code does what it is intended/claimed to do in pdf/html
- Visualization - how graphs looks and annotated
- Tables - how tables and regression results are presented and annotated
- Overall formatting of pdf/html output.
- Appendix formatting and readability.