4月1日

UNIVERSITY OF LONDON

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

(3月11日)

ST2195 Programming for Data Science

Coursework Project (50% of final mark)

The 2009 ASA Statistical Computing and Graphics Data Expo consisted of flight arrival and departure details for all commercial flights on major carriers within the USA, from October 1987 to April 2008. This is a large dataset; there are nearly 120 million records in total, and takes up 1.6 gigabytes of space compressed and 12 gigabytes when uncompressed. The complete dataset along with supplementary information and variable descriptions can be downloaded from the Harvard Dataverse at https://doi.org/10.7910/DVN/HG7NV7 includes more years (better for analyze result)

Choose any subset of (at least two) consecutive years and any of the supplementary information provided by the Harvard Dataverse to answer the following questions using the principles and tools you have learned in this course:

1. When is the best time of day, day of the week, and time of year to fly to minimise delays? → group the data (month, period of time)
2. Do older planes suffer more delays? → correlation → manufacturing / register year
3. How does the number of people flying between different locations change over time? (passenger)
4. Can you detect cascading failures as delays in one airport create delays in others? use matching
5. Use the available variables to construct a model that predicts delays. apply model

(Blk 10)

All questions should be answered using R and Python for all tasks.

Your answers should be provided in a separate structured report of no more than 10 pages. The page limit excludes title, references and table of contents but includes graphics and tables. The report should be in PDF format and also contain adequate explanations for readers not familiar with programming. In addition to the report, you will also be asked to provide your R and Python code in RMarkdown and Jupyter notebooks respectively. All the relevant files will need to be submitted in the designated Atrio submission portal.

Each report should detail all steps you took starting from raw data up to the answer for each question. Any databases you set up, data wrangling/cleaning operations you carry out, and any modelling decisions you make should be clearly described in each structured report. Each report should also include any relevant graphics and tables as part of the answer.

If you are using elements (e.g. code, databases, graphics, etc) from your answer to a previous question to answer the current one, you will need to refer to those elements.

You should also supply the code you used to answer each question, in a way that can be used by someone else to replicate your analyses. You can do this either as separate scripts or separate RMarkdown/Jupyter notebooks per question, clearly indicating (both with comments and in the filename) which question each script refers to.

Both must be used

main = R code in Rmarkdown, Python in Jupyter