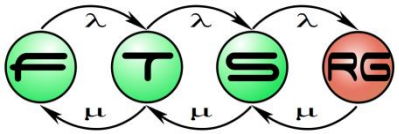
Homework requirements

Visual Analysis of Measurement Data

2019.10.10.

Budapest University of Technology and Economics
Fault Tolerant Systems Research Group



Homework submission

- Content:
  - Input data (separately), at least sample
  - Data description, identification of important variables
    + metrics (next slide)
  - Data model used, reasons for abstraction, faulty outliers identified
  - Scripts/steps for data cleaning
  - Visualization questions
  - Visualization + interactive sample
    - 1D, 2D/pairwise, multiple D
    - Effect/trend/correlation visualization („R2” for the most important metrics), corrgram/heatmap, sankey, …
    - +code/dashboard, packages/install script
  - Evaluation
    - Questions answered? Assumptions correct? Suggestions/further questions?
- Documentation: markdown/notebook
- Submission: form submission: 2019.12.08. 23:59 (Week 13)

Requirements

- Submission at homepage
  - Form submission: 2019.10.20. (Week 6) → 10.24
  - https://forms.gle/UtJ53DFJ7QcHehCX7
  - Selected dataset
    - Size, source, realated applications
    - Basic statistical summary of main variables
  - Main visualization goals/questions
  - Visualization „tpye”
    - Report/dashboard/webapp/embedded/infographics
    - At least one interactive part is needed

Dataset suggestions

- „Smart city”
  - http://iot.ee.surrey.ac.uk:8080/datasets.html
  - https://amsterdamsmartcity.com/projects/dataamsterdamnl
  - https://datasf.org/opendata/
  - https://opendata.cityofnewyork.us/
  - https://data.london.gov.uk/
  - http://www.london.ca/city-hall/open-data/Pages/default.aspx
  - https://chicago.socrata.com/
  - https://carto.com/blog/forty-brilliant-open-data-projects-preparing-smart-cities-2018/
- Public data from kaggle/kdnuggets/…
- Own measurement data

Further readings

- https://cran.r-project.org/web/packages/SmartEDA/vignettes/SmartEDA.html
- https://towardsdatascience.com/tableau-esque-drag-and-drop-gui-visualization-in-r-901ee9f2fe3f
- https://python-graph-gallery.com/
- https://www.r-graph-gallery.com/

Data metrics required

- Basic statistics:
- # of data, N/A ratio, min, max, average, StdDev, Q1,Q2,Q3
- Histogram /barchart (if applicable)

Extensions

- Extract knowledge from data
  - Clustering (k-means)
  - Decision tree (basic ANOVA)
  - Trend analysis

  → Accuracy of results! (e.g. correlation)

Checklist

- Coloring
- Variable ordering
- Variable names (~~AA_X1_avbs_LLL~~)
- Variable type
  - String/enum, int/bool, NA handling
- Plots oriented towards message

Homework presentation

- 5 minutes presentation
  - Ppt/prezi/dashboard
  - Should be submitted by 2012.12.08
- 2019.12.12 12:15-14:00, IL405

BRIEF TECH SUMMARY

Python

- Anaconda (framework)
- Jupyter Notebook (report)
- Dash (web)
- Pandas: dataframe
- Bokeh, matplotlib, seaborn, plotly: visualization
- SciKit: stats

R

- Dataframe. Data.table
- tidyverse, dplyr: data manipulation
- ggplot2: visualization
- Shiny: web
- iplots,ggvis, plotly: interactive

BI tools

- Microsoft Power BI
  - Data manipulation: Power Query (DAX), M, R
  - Visualization: visuals, custom visuals (typescript), R + Python visuals
  - Stats: R, Python, supported custom visuals
- Alternative: general purpuse visualization (D3JS)

Outlook: forecasting tools (practical insights)

- Simple methods can be integrated
  - …careful check of results needed
- Practical summary:

https://grisha.org/blog/2016/01/29/triple-exponential-smoothing-forecasting/

- PowerBI:

https://powerbi.microsoft.com/en-us/blog/introducing-new-forecasting-capabilities-in-power-view-for-office-365/

http://radacad.com/time-series-series-with-power-bi-forecast-with-arima-part-12

- Facebook prophet
- https://facebook.github.io/prophet/docs/quick_start.html