Delft University of Technology

FACULTY OF TECHNOLOGY, POLICY AND MANAGEMENT
ENGINEERING SYSTEM AND SERVICES SECTION
ICT TRACK

SEN163A – Fundamentals of Data Analytics
Formative Assignment - Data consistency

Jacopo De Stefani - Joao Pizani Flor

Based on the material by T.Fiebig

February 11, 2022

Learning Objectives

- Identify fundamental pitfalls inherent to data analysis and identify whether they exist in case-studies;
- Propose and develop suitable data mining pipelines for different case-studies;
- Analyze a given dataset to answer a given research question

Disclaimer

All characters and other entities appearing in this work are fictitious. Any resemblance to real persons or other real-life entities is purely coincidental.

Introduction

You are working in a data science team in the fraud department of Groote Nationale Investeer Bank (GNI Bank). You got tipped off by FIOD that there is something fishy going on in your banking system. Management provided you with a recent excerpt of transaction logs, and it is your task to find out what is going on.

You have the following tasks:

a. Describe the dataset you received, in words and with supporting visualizations.
b. Check the dataset for consistency, and document any inconsistency you find, as well as reasons for these inconsistencies.
c. Identify all fraudulent activity within the dataset, and provide supporting evidence and visualizations.
d. Document your method and code used during the investigation.

The data sets that should be used:

- **transaction_data.db:** An SQLite3 database holding transaction data
- **read_db.py:** A small script printing out the dataset in CSV form

You will report your findings through a **Jupyter Notebook**, including both your code and considerations on your findings. The use of figures, formulas, tables and pseudo-code to support your analysis is strongly encouraged. In case you used specific python libraries, you will need to include these libraries in a `requirements.txt` file. A template for the **Jupyter Notebook** and an example of `requirements.txt` file are available on the assignment page.

Evaluation criteria

The final grade for this assignment will be calculated based on the following criteria:

- **Quality of the report** - 35%
  - Reasonable formatting of the document and used citation appropriately
  - Use of proper English (typos, grammar)
  - Code script deliverable
  - Code quality
  - Problem Description
  - Dataset Description
  - Limitations
  - Use of figures and tables to summarize results.
  - Conclusion/Action recommendations
- **Identification of the problems in the database** - 65%

Rules for the assignment delivery

To be read carefully !

1. The assignment must be developed in groups of 4 students.
2. The assignment must include the **name** and **student id** of all the students.
3. The assignment must be submitted in **Brightspace** as a **Zip file** containing:
   - A jupyter notebook (`.ipynb`) containing your code and the discussion of the results
   - *(If you used additional libraries)* A file `requirements.txt` containing, one per line, the additional employed libraries to solve the assignment.
4. You have to follow the following constraints:
   - Upload of a file `Group_X.zip` on the Brightspace page of the course, where `X` should be replaced by your actual group number./
   - Date: **Wednesday 24 February 2022**
   - Time: **Before 18:00**

   After this deadline the assignment will be considered as late and **no feedback will be provided**.
5. Use the feedback you will receive to improve your work for the final version due on **12 April 2022 18:00**.
6. **Knock-off criteria:**
   - Missing names and id on the document/document name.
   - Code not executable: Missing libraries - Errors.