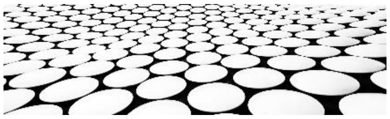
CSI4142 Fundamentals of Data Science



ASSIGNMENT 1

Exploratory Data Analysis

GOALS

The overall goal of this assignment is to become familiar with EDA, more specifically with central tendency and dispersion measures as well as univariate and bivariate visualization tools for both numerical and categorical features.

At the end of this assignment, you will have:

- Reviewed your Python skills, as the project MUST be done in Python
- Explored Kaggle, an amazing resource of challenges and datasets
- Programmed multiple descriptive analysis on 2 datasets
- Documented, in a Jupyter Notebook, everything about your empirical study in a way to make your analysis understandable and reproducible.
- Practiced how to tell the “story” behind data, choosing top insights to highlight and supporting them through data analysis

SUBMISSION INFORMATION

- Deadline:
  - Submission of your notebook: **Tuesday, January 28th, midnight**
  - Grace period (no penalty) is until 12:30… after that it's 20% per day off.
- Groups:
  - You are expected to form groups of 2 and do a single submission per group. You first need to register your group in Brightspace to later be able to do a group submission.
  - If you prefer to work alone, that is accepted (but not encouraged), but the requirements are not changed.
- Where to submit:
  - Your submission must be done in Brightspace in Assignment section (Assignment 1)
- Submission format:
  - Only Jupyter Notebook files will be accepted
  - For the datasets used in your analysis, your code MUST contain a link to the dataset so the TAS do NOT need to download any data.

PLEASE NOTE:

If the corrector cannot run your code, the mark will be zero for the entire assignment. It is your responsibility to test that the code cells in your Notebook are executable. The most frequent problem is that the dataset is not readable (file not found). Please make sure the dataset is either read directly from a public place (a dataset repository) or read from one of your shared repository (accessible to anyone) so the TA can run your notebook without any data download.

TUTORIALS/TECHNOLOGIES

To achieve this assignment, you need to explore different programming environments. As this is a 4th year course, and you all have programming experience, you are mature enough to perform the exploration by yourself. If you "get stuck"… don't hesitate to post your questions on the Brightspace forum in the Assignments section.

To help you, on various aspects of the project, here are some resources:

1. **Python**: The project MUST be done using the Python programming language. There are many tutorials that you can use to familiarize yourself with Python.

    - Python Tutorial for Beginners https://www.youtube.com/watch?v=t8pPdKYpowI
    - Python Full Course for Beginners https://www.youtube.com/watch?v=_uQrJ0TkZlc

2. **Python for Data Science**: Here is a suggestion for enhancing your python skills with data science examples.

    - https://jakevdp.github.io/PythonDataScienceHandbook/
    - https://github.com/jakevdp/PythonDataScienceHandbook?tab=readme-ov-file

3. **Jupyter Notebook**: Your project will have to submitted as a Jupyter Notebook. You can create/run such notebooks using Colab (see point 4).

    - Jupyter Notebook tutorial (Windows) https://www.youtube.com/watch?v=2WL-XTl2QYI
    - Jupyter Notebook tutorial (Mac) https://www.youtube.com/watch?v=HW29067qVWk

4. **Colab** https://colab.research.google.com to access machines allow you to run code without installing anything.

    - Google Colab Tutorial https://www.tutorialspoint.com/google_colab/index.htm

5. **Visualization libraries**

    - Matplotlib : https://matplotlib.org/stable/tutorials/index
    - Seaborn : https://seaborn.pydata.org/tutorial.html

INSTRUCTIONS

1. Choose 2 datasets

The Kaggle site is a very interesting site to explore as it contains datasets for many tasks in data science and AI. You must select, from Kaggle 2 datasets to explore.

For **Dataset 1**, it MUST be one among the 3 suggested below. The 3 suggestions are in 3 different domains: finance, healthcare, and mobile usage.

1. German Credit Dataset
   - Size: 10 columns, 1000 rows
   - Description: Financial data for credit scoring
   - Link
2. Heart Failure Prediction Dataset
   - Size: 12 columns, 918 rows
   - Description: Health-related data for heart failure prediction
   - Link
3. Mobile Device Usage Dataset
   - Size: 11 columns, 700 rows
   - Description: Technology/mobile behavior data
   - Link

For **Dataset 2**, you must select a dataset that is NOT within the 3 suggested sets above. Furthermore, your second dataset MUST be in a domain different from the first dataset. I want you to explore 2 different domains. Another constraint is to find a dataset with a minimum of 10 columns so you have various features to explore.

*** Attention: During the semester, for your assignments, you will need to choose various datasets for different explorations. You will NOT be allowed to reuse a dataset from one assignment to another. So, if you find a few datasets that you think are interesting, save some for your future assignments!*

2. Report the story of each dataset

The purpose of the report (written within a Jupyter Notebook) is to illustrate 10 insights that you found through your analysis for each dataset. In the notebook, you will be able to alternate between text and code, both required.

Your Jupyter Notebook should include:

1. Group number, names and student numbers of group members
2. Introduction to provide the goal of the analysis/report and mention who the audience would be (imagine an audience who would read your report).
3. A description of the two datasets used *(see Dataset description requirement section)*
4. **A set of 10 insights for each dataset**. For EACH ONE:
    a. State the insight in a single sentence.
    b. Show supporting evidence from the data making sure that evidence is as self-explanatory as possible (graph with title, axis descriptions, etc)
    c. Mention what type of analysis was done to arrive at such insight *(Analysis description requirement section)*.
    d. Show how this evidence was obtained (and how to reproduce it – provide code)
5. Conclusion
6. References

3. Dataset description requirements

For the dataset description, make sure you include the following information:

- Dataset name, author, purpose (what was it made for)
- Shape: how many rows and columns
- A list of the features + descriptions (what do they represent and are they categorical or numerical)
- Highlight any hint at redundancy or missing values from the dataset schema

4. Analysis description requirements

Given that we are in an **academic context**, I have additional requirements to make sure that you explore various types of analysis and visualization tools. Therefore, **among your 20 insights (10 for each dataset), you MUST have a diversity of supporting evidence including at least one of each of the analysis (written as r1 to r7) below**:

a) Univariate analysis
   a. Numerical data:
      i. (r1) Simple histogram for visualization of dispersion
   b. Categorical data
      i. (r2) Countplot for a category with multiple values
      ii. (r3) Grouped-Data countplot in which you group some values (and explain how you did the grouping)

b) Bivariate analysis
   a. Categorical/Categorical
      i. (r4) Comparing categories with 2 values
      ii. (r5) Comparing categories with more than 2 values for which you set the order (e.g. increasing counts, or alphabetical order)
   b. Numerical/Numerical
      i. (r6) Use the scatterplot to highlight correlation
   c. Numerical/Categorical
      i. (r7) Split the data by certain categories to explore the numerical distributions

☆☆☆

EVALUATION (50 points)

- Overall effort in the report (5 points)
  - Provided all required sections clearly identified
  - Good cell separation (text, code, results, etc)
  - Tests on various examples easy to perform by the corrector
  - Report detailed enough for reproducibility
- Dataset descriptions (5 points)
  - Descriptions include all required elements
- Insights sentences (10 points)
  - Sentences must be clear, with well-chosen words.
  - Set of insights is varied and interesting.
- Insights supporting evidence (15 points)
  - The choice of charts/graph/table used as supporting evidence for the associated claims are well-chosen and correctly done.
  - Quality of graphs/images, All of them should be self-explanatory, with titles as well as legends for both x and y axis.
  - Having the 7 required analysis as part of the supporting evidence
- Code for reproducing the analysis is clear and easy to run (12 points)
  - Well commented code overall
  - Good choice of variable names to make the code easy to read
  - Code including parameter settings to arrive at self-explanatory graphs
- References (3 points)
  - For any part of your code taken from a web site (even a tutorial site or stackoverflow), you must provide the reference to it.
  - If part of your code was generated using Generative AI, please list the tool as well as the queries performed for each code snippet used.
  - Any theory/algorithms found in books, slides, tutorials that you used should be referenced.

QUESTIONS

- You can ask your questions within the assignment topic of the discussion forum on Brightspace.
- To make sure you get answers on the forum on a regular basis, the 3 TAs will share the days to answer your question. Notice that they are NOT available on weekends.

	Monday	Tuesday	Wednesday	Thursday	Friday
Morning	Gurdarshan	Gurdarshan	Bhavneet		Ángel
Evening	Ángel	Gurdarshan	Ángel	Bhavneet	Bhavneet