Week 1 Group Assignment

Exploratory Data Analysis and Data Cleaning

This is the first out of three group assessments that you have to complete as part of the +Masters. Each assessment will build on top of each other. In this way, you have the opportunity to create an end-to-end data analysis project, understand the requirements of working on such tasks, and potentially contribute to your technical portfolio.

Task 1: Identify a real-world problem you are interested in solving using data analytics. You should justify why this problem is important to solve, and who would benefit from the solution. In other words, think about your audience or who your stakeholders might be. It is up to you what kind of problem you want to choose. It can be something you are interested, something fun, something that relates to the industry you are most interested in, or all of them.

Note: Before researching datasets, you should think about potential issues that you can encounter. For example, if you work on a healthcare problem, you might not be able to find openly available patient level data.

Task 2: Based on the problem identified, research and find a dataset that can help address this problem. The dataset should be rich enough for complex analysis, including various types of data and, preferably, some missing or dirty data. You might find yourself in the situation where only one dataset might not be enough and that you may need to combine multiple datasets together.

Note: You should explain the data source and the reasons you chose this particular dataset(s). How does it fit your problem?

Task 3: Conduct an in-depth Exploratory Data Analysis (EDA) on your dataset(s), identify key features, any interesting patterns or anomalies, and potential challenges in the data (missing values, outliers, etc.).

Note: You should also give an overview over the data quality.

Task 4: Perform and document the data cleaning process, explaining why you made certain decisions. This should involve handling missing data, dealing with outliers, and ensuring data consistency.