Insights on My Drive to Embrace Data Thinking: A Journey from Realization to Motivation

Siim Pähn

Edited by GPT-4

In my current role, I often find myself dealing with large volumes of data derived from diverse sources, including both raw machine data and human inputs. The task entails refining, interpreting, and suitably placing this data.

It's becoming increasingly clear to me that data is the currency of the future, often said to be now more valuable than oil. The magnitude of this insight fuels my motivation to delve deeper into the realm of data analysis. It's clear that the leading organizations across finance, technology, and business will be those who can harness and leverage data effectively. Using this data, we can create products and services that significantly enhance the quality of life for the masses.

My primary objective is to hone my data thinking skills. Data thinking isn't solely about technical proficiency; it also involves creativity, curiosity, and a willingness to question existing norms and paradigms.

📀 + 🤔 The Research Question and Context: A Look at the Covid-19 Data from Estonia

In absence of a predefined research question, I sought inspiration from available datasets. My search led me to an intriguing dataset: Covid-19 data from Estonia.

The dataset consists of an array of information related to the Covid-19 pandemic within Estonia, including test ids, gender, age, county, test results, date, and the timestamps for when the tests were taken and results logged. The data is sourced from the Estonian Health Board and is publicly available on Kaggle.

However, there are uncertainties surrounding the data's completeness and accuracy. It's unclear whether the data fully accounts for all cases, particularly asymptomatic or mild ones, or whether it includes information on the prevalence of Covid-19 variants within Estonia.

As for decision-making, the onus is on me to interpret this data meaningfully and extract intriguing insights from it.

🧐💰 Stakes of the Research Question

Since this is anonymized public data, the stakes aren't exceptionally high. The most significant stake would be the impact on my academic grade.

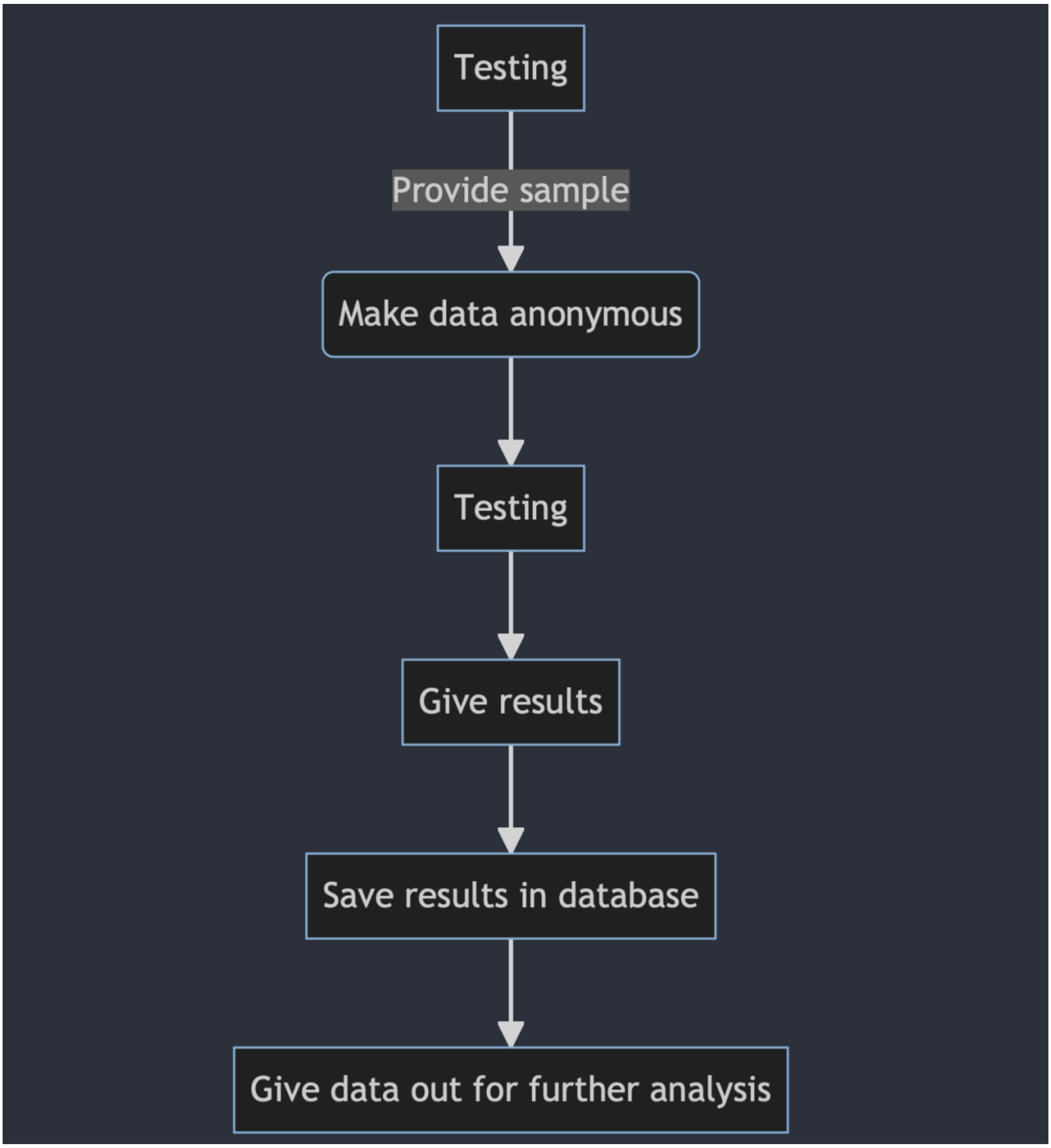
📀 + 🤔 The Data Context and Flowchart

The dataset, 'opendata_covid19_test_results.csv', contains anonymized Covid-19 test results from Estonia, released by the Republic of Estonia Health Board. The data presents opportunities for analyzing the virus's spread among different age groups and counties, potentially informing early warning systems.

The data is anonymized to ensure privacy, and thus, it lacks context regarding testing conditions or underlying medical conditions of tested individuals.

From testing to providing samples, making the data anonymous, and saving results in the database for further analysis, the following flowchart delineates the data journey:

```
Flowchart TD
A[Testing] -->|Provide sample| B(Make data anonymous)
B --> C[Testing of results]
C -->D[Give results]
D -->E[Save results in database]
E -->F[Give data out for further analysis]
```



📐🖍 Mathematical Model and Visualization (in LaTeX or https://typst.app/)

Did not have time to complete due to skill issues with VSC

📀 + 🤔 Assessment of the Analysis

Did not have time to complete due to skill issues with VSC

📀 + 🤔 Reflection on the Journey

Embracing data thinking has been an enlightening but challenging journey, encompassing a spectrum of social, emotional, technical, and legal aspects.

In the learning process, I've grappled with tools like Github and Visual Studio Code. While their use is vital, I felt that a more comprehensive introduction would have made navigation and problem-solving less daunting. This sentiment seemed to echo within the class, as evidenced by dwindling lecture attendance.

Additionally, the introduction of new assignment submission methods, accompanied by inconsistent timeframes and lack of clarity, added to the complexity. The sudden inclusion of bootcamp participants in our communication channels also necessitated a complete muting of notifications to avoid message overload.

These challenges often led to me postponing this work, as the discomfort deterred me from diving into it wholeheartedly. However, this experience has highlighted potential areas for improvement, which I hope will enhance the learning process for future cohorts.

📈 Interactive Altair Visualization

📀 Metadata for the Dataset

The dataset consists of 9 columns with 3,277,016 unique ID values. The gender distribution is 54% female, 46% male, with 11,873 unspecified. No single age group dominates the data, and 94% of entries are from Estonian residents, with 48% residing in Harju county. Regarding test results, 83% were negative. The data spans from March 10th, 2020, to April 11th, 2022.

📓 Code Snippets in Python, SQL (duckdb), dbt

Did not have time to complete

💬 Use of Large Language Models and a Gist Link

For this project, I used GPT-4 to enhance the conciseness and clarity of my text. I also attempted to use GPT-4 for coding but achieved limited success.

🤔🧘🙏 Feedback and Acknowledgments

I extend my gratitude for introducing me to a variety of AI tools that I've begun incorporating into my daily life. This journey has been insightful, enriching my understanding of the expansive landscape of data thinking.