STQD6324-Pan
homework

This article is used to discuss DATA LIFE CYCLE, propose, collect, assure, describe, submit, preserve, discover, integrate, analyze, publish.

1. Propose part

Establishing a Data Management Plan at proposal stage in the Data and Research Life Cycle facilitates a structured work with data and saves time later on.

My plan is to find the data, do some processing, get the processed data, visualize it, save the results, upload the results, analyze, and then submit the results.

2. Collect part

Data provided by the linked author:

https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Fwww.kaggle.com%2Fdatasets%2Fakshaydattatraykhare%2Fdiabetes-dataset

For detailed information about the data, you can view the link or the colab I edited.

3. Assure

Assure refers to quality control and assurance. Assurance encompasses all those activities which ensure the reliability of data. High quality data are a key element for research and impact replicability of results. I cleaned the data, removing missing values and the like.

Outliers may represent data contamination, a violation of the assumptions of the study, or failure of the instrumentation. I did a simple removal of outliers.

4. Describe

Additional information about data is called metadata. Metadata describe all aspects of data (e.g. who, why, what, when and where) that would allow one to understand the physical format, content and context of the data, as well as possibly how to acquire, use and cite the data. Data provided by the linked author.

5. Submit

Submission is the transfer of data to a curated environment. This is usually an archive, a data centra, a repository, or a collection. Then I just save the data in GitHub and Google Cloud. Submission to a curated environment ensures safe long term storage and makes data discoverable for other researchers (in the team or outside).

6. Preserve

Digital data are fragile. Hardware fails, software becomes obsolete, files are subject to bit rot which produces bit errors.

Backup of data is not the same as preservation. Backups are short-term recovery solutions whereas preservation includes measures taken for long term storage and archiving. I store it in the cloud, and my stored files can be accessed even if I change devices.

7. Discover

Discovering data means to search and find data collected by other researchers. This data can be used for different purposes like long-time analysis, modelling or comparative studies. Mine is for submitting assignments to get grades.

8. Integrate

Integration is the merging of multiple datasets from different sources, like your recently collected data with former data from other owners, resulting in a new, bigger dataset. I don't have Integration, I just used a dataset.

9. Analyse

Analysis comprises the actions used to derive and understand information from data. My analysis should be data visualization. For details, see the colab file I submitted. Simple analysis is also in the text part of my code, which is the interpretation of the histogram and scatter plot.

10. Publish

The last step of the Data Life Cycle deals with the publication of data, especially with datasets linked with the publication of related academic papers. My assignment will not be published.