Documentation and Data Management Plan Julian Patrick Stoerr

File storage & Backup

- Files will be stored on Github in a Repository titled "DSA103_DataProject_SJ"
- The Repository will be kept up to date over the duration of the project and Branches will be create for further data exploration exceeding the extent of the project.
- Version control will be done using Github

Documentation Procedure

- Documentation will be provided under the folder documentation in the repository at "DSA103_DataProject_SJ/documentation"
- Basic information of the analyse Research Paper is found in the folder data at "DSA103_DataProject_SJ/data" which contains the original README file as well as the data used in the paper and project.
- Protocols are documented in the Jupyter notebooks used for analysis, where each step of data processing and exploration is described directly alongside the corresponding code. Additional contextual information such as dataset description, data source, and environment specifications are provided in the repository documentation and README files.
- The EDA (Experimental Data Analysis) will be conducted in a jupyter notebook/ .ipynb file which will contain code and a step-by-step guide on how to run the code and explore the data in the same manner. It will be compared to the original procedure used in the paper and deviations from original protocol will be clearly marked.

Data Integrity and Traceability of data and analysis

- Datasets from the paper will be left unchanged and all changes will be done during data wrangling and if saved be marked as distinctly different. The used python environment will be frozen and saved as "environment.yml" so reproducibility using the same libraries is ensured.
- The raw data will be kept separately from changed files and if saved, stored in the modified_data folder "DSA103_DataProject_SJ/modified_data". During data wrangling it will be specified when and what is saved so data is not mixed during EDA and further analysis.
- Issues in the code can be marked for revision using the built in Git Issue system for collaboration and reproduction of analysis and data wrangling

FAIR Principle

- **F**indable: Data will be clearly labeled and organised in a structured and clearly labeled repository with documentation provided and description of the project in the README file.
- **A**ccessible: Data are stored in a publicly accessible GitHub repository and are available in standard machine-readable formats.
- **I**nteroperable: Variables are clearly labeled and filenames and data types will be in interoperable formats such as .csv
- **R**eproducible: Data reusability is ensured using clear documentation, frozen environment and usage instructions.

CARE Principle

- **C**ollective Belief: Results are shared in an accessible manner to ensure public availability of methods and analysis
- **A**uthority of Control: Data governance respects contributor authority, and data usage adheres to consent agreements.
- **R**esponsibility: Data are used responsibly and in alignment with research agreements and visions
- **E**thics: Ethical standards of research and education were applied in conducting the project