Data Preparation / Cleansing

While in the long term, I hope that this project will evolve into a general methodology which is useful for a variety of situations and conditions, I’ll be constructing the project around a diary of my personal symptoms. The Data for the project was collected via Google sheets, and consisted of a table of incidents (anxiety attacks), and an adjacent table of normal daily activities customized to suit my lifestyle at the time.



When an incident occurred, it was recorded in the first table along with its time, date and severity. The entire day’s schedule was then recorded, including activities which were planned for later in the day. “Controls” consisted of days in which no incident occurred, but for which an entry is made with a Severity of 0, and the events of the day fully documented. In total, roughly 30% of entries were controls.

While this format was effective for end-user data collection, the following changes were implemented in order to be compatible with modelling:

-Empty records dropped

-Times converted to 24 binary fields

-Time and Date data standardized

Data was subsequently exported to CSV via DaxStudio. Some additional truncation was performed in SPSS to increase cleanliness and readability.

Modelling and Exploration

Initial modelling was to be performed in IBM SPSS. Depending on the availability of time and resources, I intend the scope of this project to eventually include implementing the model in R; however SPSS’s exploratory tools significantly reduce the total time cost of the project (despite limiting compatibility).