**BY USING THE SMART PRINCIPLES**

**Dataset Evaluation**

1. Nature of the Dataset

ˉ What is the subject matter of your dataset?

The dataset pertains to Europeans VOC personnel that have signed on in the Netherlands. It includes their names, places of origin (usually, but not necessarily their birthplace), job and career trajectory. It also includes information about the(ir) voyage, the amount of days in service, and the reason for not signing another contract.

ˉ Where does it originate from, and who is responsible for its creation?

The dataset originates from archives of the VOC from 1600 until 1794. The data comes from the VOC pay ledgers that recorded European persons who endeavored on a VOC journey from the Netherlands.

These archives were then digitised by historian Ton van Velzen between 2000 and 2012, who manually transcripted them and created the VOC-opvarenden database. Because the VOC pay ledgers were individual contracts, the original dataset was organised as individual contracts/voyages.

This organisation made it difficult to identify individuals and their career paths. Between 2017 and 2024, a collaborative project between historians, data scientists and institutions enriched and restructured the dataset. Individuals were disambiguated and place names were standardised, and ranks were categorised.

ˉ What is the size of the dataset, and what variables or fields does it contain?

The dataset contains 460.452 disambiguated persons, for a total of 774.200 — one of the largest historical datasets ever created.

ˉ In what technical format is it stored? Can you work with this format, or do you need to convert it? If yes, how?

The data is stored in a .csv format, which is perfectly suited to be manipulated with R.

**Formulating a Research Question**

2. Dataset Suitability for Research Questions

ˉ What types of research questions can this dataset realistically help you answer, and for which ones is it unsuitable?

The dataset could give us answers about the places of origin for the VOC personnel, or about their career trajectories. We could also look at contract length or the duration of a voyage, and if there are some interesting things that stand out about that. According to Worden (2009), fleets launched around Easter would be at risk of getting stuck in the doldrums, which could lengthen the journey. Another interesting question could pertain to the amount of debtors aboard. Did these people sign on again? And were they still in debt when they did? This could give us insight into why people would sign up for risky journeys across the world.

ˉ Refine your research question to ensure it can be answered using the dataset. Keep in mind that your research question should adhere to the SMART principles

Is there a correlation to an individual's initial rank category (specifically, 'Military', 'Sea', or 'Trade') within the Dutch East India Company during the eighteenth century and their mortality risk, and if there is, what does this say about the working conditions of the Dutch East India Company?

**3. Reflecting on Data Limitations**

-Consider the broader debates surrounding the dataset. Are there limitations such as missing data, biases, or ethical concerns related to how the data was collected or who it represents? Reflect on how these issues might affect the validity of your research question and findings

One large bias in the dataset, that even the authors point out, is the ‘Dutch bias’. VOC clerks were more familiar with the Dutch language, probably resulting in less variation for the names and places of origin of Dutch onsigners. When creating the dataset from the archive, transcription was done by Dutch people, leading, again probably leading to fewer errors and spelling variation. During the standardisation process by Petram et al. (2024), resources with a focus on the Netherlands were used by Dutch data curators, strengthening the bias.

Place of origin of persons have been assigned by the use of some standardisation rules. Though logical assumptions were made, they remain just this — assumptions. Larger settlements or settlements that had yielded more VOC personnel were favoured, skewing the data and potentially underrepresenting some communities and individuals.

During the creation of the dataset, the creators were conservative during the linking of different contracts. This reduced false positives (Type 1 errors) but increased false negatives (Type 2). Because of this, careers may look shorter than they were in reality.

**Implications:** if individuals with common names were presumed to be from core regions, when they were not, then these individuals would seem to have only one voyage, ending in death. Conservative linking will also contribute to that. Although this may not skew mortality rates amongst ranks, mortality rates might have been lower than they appear to be in this dataset.

There is more information about lower ranking persons, for they were more common. More data usually leads to more reliable outcomes when using statistical analysis. Officers were rarer than sailors, so each missing death has disproportionate weight. The standardisation process does seem to skew this further; five ambiguous persons out of a relatively small pool of officers is proportionally larger than five ambiguous sailors will be.

**Planning the Data Analysis**

4. Data Consistency and Completeness

ˉ Are the data consistent enough to provide reliable answers to your research question? If not, how can you address these issues—through data pre-processing, transformation, or obtaining additional data?

The ranks and deaths of the VOC personnel are consistent and do not contain missing variables. There are other missing variables, such as place of origin. The places of origin seem to be one of the more inconsistent variables, due to the conservative linking of individuals.

ˉ Can the research question be answered directly, or will it require a step-by-step process? Outline these steps clearly.

The research question should be able to be answered fairly easily when the “reason for the end of contract” is turned into a binary variable: deceased or other.   
 QUESTION: Are we leaving out the variables *murdered* and *death penalty* in the category deceased or are we including them?

**5. Data Manipulation**

To prepare the dataset for analysis, think about the following interventions:

ˉ Selection: Are there parts of the dataset you can disregard because they don't contribute to your research question?

Yes, we will only be using the contract information and ranks dataset and joining them. Because of the large size of these datasets, we will probably be removing a lot of variables that are not necessary for answering our research question, both to spare our computers and to make the reading of the dataset easier for us.

ˉ Modeling: Do you need to reformat or restructure the data to make it suitable for your analysis? This might involve discarding data or converting it into a usable format.

ˉ Normalization: Consider how you will standardize data (e.g., dates, names, terminology). What variations are important, and which should be eliminated? What are the potential consequences of not normalizing certain elements?

ˉ Linking: Should your dataset be combined with other datasets or external sources to provide a more comprehensive analysis?

So yes but its from the sama dataset

ˉ Classification: If the dataset is complex, can it be categorized into formal classifications to simplify the analysis? How can these classifications be defined, and should they be applied manually or through automated processes?

**Further Notes for the Assignment**

ˉ This guide centres on evaluating an existing dataset and crafting a SMART research question. It does not address the process of selecting a dataset or formulating research questions from the ground up.

ˉ Keep in mind that a dataset often holds the potential to answer far more questions than initially apparent. Focus on one clear, well-defined question to avoid unnecessary complexity.

ˉ Use these prompts as a tool for critical engagement with your dataset. They are intended not as a checklist, but as a means to encourage deep and systematic thinking about the connection between your dataset and your research

**Source:**

Wezel, F. C., & Ruef, M. (2017). Agents with Principles: The Control of Labor in the Dutch East India Company, 1700 to 1796. *American Sociological Review*, *82*(5), 1009–1036. https://doi.org/10.1177/0003122417718165