WORKFLOW DOCUMENTATION GROUP 1

LISANNE KOEK, SAMANTHA MOOIJ, SIERRA SCOTT

PROJECT OVERVIEW

Our project utilizes the *Charting Lives and Careers: Enriched Data About the Dutch East India Company's Eighteenth-Century European Workforce* dataset (Petram et al., 2024). We used this newly processed data on VOC personnel to examine the relationship between an individual's rank category ('Military', 'Sea', 'Medical', 'Ship', 'Trade', or 'Other') and their recorded risk of mortality during the eighteenth century. Furthermore, we aimed to explore what any observed patterns about this relationship might suggest about how work, (financial) hierarchy and status were organized within the Dutch East India Company.

Research Question: *Is there a relationship between an individual's initial rank category within the eighteenth-century Dutch East India Company (VOC) and their recorded risk of mortality, and if there is, what might this relationship reveal about the organization of work, hierarchy, and inequality within the VOC during this period?*

Main Objectives:

- Quantify and compare mortality rates across different rank categories within the VOC to identify any significant disparities,
- Investigate the medium wages associated with different rank categories within the VOC, assessing how these wage differences correlate with mortality rates and overall working conditions,
- Engage with existing historical literature on the VOC and its corporate practices to build a theoretical framework for interpreting the data,
- Assess how the hierarchical structure of the VOC influenced decision-making processes regarding employee welfare and safety, and its implications for mortality outcomes.

Thesis Statement: *Mortality patterns across initial rank categories in the eighteenth-century Dutch East India Company (VOC) can highlight how differences in working conditions and occupational status shaped vulnerability among employees, revealing broader systems of inequality and control within the Company's labor organization.*

DATA ACQUISITION

The datasets used were sourced from Zenodo, which is an open, general-purpose digital repository catered for researchers to share and preserve all types of research outputs. The original dataset contains information on 460,452 individuals employed by the Dutch East India Company (Verenigde Oost-Indische Compagnie, VOC) in the seventeenth and especially the eighteenth centuries, developed from 774,200 muster records in the ‘VOC-opvarenden’ collection.

These records originate from VOC archives and were originally digitised by Ton van Velzen and his team between 2000 and 2012, who manually transcripted them and created the VOC-opvarenden database. This database was turned into datasets between 2017 and 2024 during a collaborative project between historians, data scientists and institutions. This process is documented in Petram et al. (2024).

To curate the enriched datasets, which are split into 9 files, the original data has been enhanced through the disambiguation of individual records. We utilized 2 out of the 9 enriched datasets for our project: the “voc_persons_contracts” dataset and the “voc_ranks” dataset. The 9 enriched datasets were published to Zenodo in February of 2024.

To address our research question, we preprocessed the “voc_persons_contract” dataset by adding an additional variable to the dataset, ***“rank_group.”*** To curate the “rank_group” variable, we organized the ranks into categories in accordance with the rank categories included in the “voc_ranks” dataset. Also during the pre-processing stage, we further pre-processed the “voc_persons_contract” dataset by creating another variable, ***“died.”*** This variable was created to capture all Dutch VOC workers who had their contract terminated due to being deceased. The original dataset also captured other ways in which a Dutch VOC worker’s contract was terminated that involved death, such as death by shipwreck, age and death by murder. However, these additional causes of death proved to be insignificant, so for clarity we decided to only capture Dutch VOC workers whose contract was terminated due to being exclusively deceased. Therefore, the final datasets used for this project were the “voc_ranks” enriched dataset in which no pre-processing was necessary, and the “voc_persons_contract” enriched dataset with two additional variables added: “rank_group” and “died.”

METHODOLOGY

Task division:

In this section, we explain how we approached task division in our project. Each team member focused on tasks that aligned with their strengths and background, but most parts of the process (especially decision-making, literature research, and the presentation) were done collaboratively to keep the project consistent and coherent throughout. An overview of the work packages and how they were divided is presented below.

- **Research Framing:**
  - Goal: Finalize research framing
  - Lead: Lisanne

    - Support: Samantha and Sierra
    - Tasks:
        - Conduct literary research on the VOC
        - Formulate the research question and thesis statement

- **Data Collection & Preprocessing:**
    - Goal: Prepare dataset for analysis
    - Lead: Sierra
    - Tasks:
        - Preprocess the data using R

- **Data Analysis & Visualization:**
    - Goal: Provide analytical insights and visualizations
    - Lead: Sierra
    - Support: Samantha
    - Tasks:
        - Conduct data analysis in R
        - Generate visualizations to support findings

- **Literature & Contextual Interpretation:**
    - Goal: Interpret results in context of existing literature
    - Lead: Samantha
    - Support: Lisanne
    - Tasks:
        - Connect quantitative findings to historical context

    - **Presentation:**
    - Goal: Final group presentation
    - Co-leads: All three (Sierra, Samantha, Lisanne)
    - Tasks:
        - Write text that provides an overview of the project for the presentation
        - Build slides and visual assets
        - Present in class

Justification for Chosen Methodologies:
To collect, pre-process and analyze our datasets, we chose to utilize the R coding language because multiple group members were most familiar and comfortable working with data via R. Similarly, we chose to use Github for sharing due to team familiarity with that platform.

WORKFLOW STEPS

Throughout the course, our project underwent four main stages:

1) Literature review

- Conducted a general literature review on our topic, and initially chose to examine mortality within the Dutch VOC. After an extensive literature review, we defined the initial target of our research that would inform our research question: examine mortality differences across rank groups in the Dutch VOC.

  Tools
- Web searches, notes in a text document (no special software required).

  Key decisions & rationale
- Decided to broaden the scope of our research question by assessing most vulnerable groups in Dutch VOC to gauge inequalities within the Dutch VOC's hierarchical system. This lead to additional research examining both mortality and wage distribution between rank groups
- Decided that examining wage difference across rank groups would provide additional insight into which VOC works were the most vulnerable in the Dutch VOC's hierarchical system

2) Data collection & import

- Downloaded the enriched VOC datasets from Zenodo and added them to the project repository via Github. We selected two of the nine datasets that aligned best with the research question: the persons/contracts dataset (records of individual contracts, including descriptions of when and why contracts were terminated) and the ranks dataset (rank descriptions, including wage information for each worker).

  Tools
- R for data processing, Git/GitHub for version control and sharing.

  Key decisions & rationale
- Used the persons/contracts dataset because it contains the contract outcomes (death, transfer, etc.) and the ranks dataset to map rank terms into clearer categories and examine wage information.

  Alternatives considered
- We considered using all nine enriched datasets, but this approach was not chosen because most other files did not add much to the mortality question and would increase complexity and data-joining effort.

3) Data Pre-Processing

- Once both datasets were loaded into R, an initial missingness check was conducted on both datasets. Fortunately, we found minimal missingness within our target variables for each dataset, so we proceeded with pre-processing
- **“Voc_persons_contract” dataset:**
  - Collapsed very granular rank strings into a small set of meaningful rank groups (we limited to ≤ 8 groups to keep interpretation tractable). We chose to use the same rank groups included in the “voc_ranks” dataset. Below is a brief description of each rank group:
    - ***Medical:*** referred to all medical professionals
    - ***Military:*** referred to all military workers
    - ***Ship:*** referred to all ship workers who held official/managerial positions on a ship
    - ***Sea:*** referred to a wider range of ship workers who didn’t exclusively hold official/managerial positions
    - ***Trade:*** referred to workers who were exclusively labelled as a type of merchant
    - ***Other:*** referred to commercial workers who held specialized positions and weren’t exclusively labelled as merchants
    - ***NA:*** referred to workers whose rank was not specified; while there were minimal workers whose rank was not specified, this category was kept to ensure consistency in rank group labelling across both datasets
  - Created a binary mortality indicator based on the reason in which a VOC worker’s contract was terminated (1 = contract termination due to being deceased; 0 = contract terminated for a reason other than being deceased).This re-labels many textual reasons for contract termination into a (died / not died) variable.
- **“Voc_ranks” dataset:**
  - No additional pre-processing was necessary for this dataset

Tools

- R for data processing, Git/GitHub for version control and sharing.

Key decisions & rationale

- Two new variables were added to the “voc_persons_contract” dataset: “rank_group”, in which we assigned workers to one of seven rank categories, and “died”, a binary variable which indicated whether or not a worker’s contract was terminated due to being deceased. We referred to the “voc_ranks” dataset to create the “rank_group” variable.
- Creating these two new variables made it simpler to create visualizations that better addressed our research question

Alternatives considered

- For the “voc_persons_contract” dataset, we considered augmenting the “Trade” rank group to include all workers who held commercial occupations, so that this rank group wouldn’t include exclusively merchants. We decided against this approach because it complicated the “Other” rank group, and to maintain consistency, we thought it would be best to keep the rank groups consistent in both the “voc_persons_contract” and “voc_ranks datasets

- When creating the “died” variable for the “voc_persons_contract” dataset, we initially chose to assign the label, “died”, to any VOC worker who had their contract terminated due to being deceased, shipwrecked or murdered, because these reasons each involved death. However, we decided against this approach because there was an insignificant number of contract terminations due to being shipwrecked or murdered, and we didn’t want to complicate the “died” variable. Therefore, we assigned the “died” label to any worker who had their contract terminated solely due to being deceased.

4) Data Analysis

- Once both datasets were pre-processed, we sought to create a visualization that illustrated the proportion of workers in each rank group who had their contract terminated due to being deceased, and a visualization that illustrated the wage distribution for each rank group to address our research question

Tools

- R was used to create all visualizations.

Key decisions & rationale

- Created a stacked bar plot to visualize the proportion of deaths across rank groups
- Created a visualization containing multiple boxplots to illustrate the wage distribution across rank groups

Alternatives considered

- Because the “Other” rank group referred to a wide range of workers that held specialized positions compared to the “Trade” rank group, we considered creating two stacked bar plots: one that contained the “Other” category, and one that did not contain that category. However, we decided against this approach because it would introduce redundancy on our project presentation, and we dedicated to orally explain the key differences between the “Other” and “Trade” categories rather than create two separate visualizations
- To visualize mortality, we constructed several types of bar plots to illustrate the proportion of deaths for each rank group. However, we decided to utilize a stacked bar plot because it was the most readable and visually appealing

CHALLENGES AND SOLUTIONS

Most challenges during our project came in the form of choices. For a lot of these choices, we followed decisions made by Petram et al. (2024), partly due to time constraints. One of the challenges was the question if we should use only the records of ambiguated persons. Because of the “Dutch Bias” inherent to our dataset, and the favouring of larger places of origin, using only ambiguated persons could artificially decrease the chance of death. However, due to the conservative linking used in the enrichment process, including disambiguated persons could artificially increase the mortality rate. After exploring these options in RStudio, we decided to include both ambiguated and disambiguated persons. While the numbers vary, the relational relationship between rank and mortality wasn't affected, so we decided to follow Petram et al. (2024).

We also thought about including place of origin alongside or instead of wage data. Literature suggests that non-native Dutch personnel were excluded and treated worse than Dutch personnel, possibly affecting their chance of death. For example, Worden (2009), notes that a lot of military personnel originated from inland Germany — and according to our data analysis, military personnel were also most likely to die.This might be another form of hierarchy present on VOC ships.

However, the placenames are one of the variables most affected by Dutch Bias and all of them are in Dutch, limiting accessibility and interpretability for international researchers. We decided to exclude place of origin from our analysis, because the complexity of working with this variable would be very hard to do during our limited time frame, but this is definitely an avenue for further research.

We also encountered difficulties with how ranks were defined in the dataset. The "other" category contains a lot of occupations, which weren't at all congruent with each other. Many of these jobs were land-based, but these people still appeared in the muster records, as they had to travel by sea to perform their duties in the colonies. It isn't clear whether their deaths occurred during or as a result of the voyage or while stationed abroad. Their wages were also quite inconsistent, as visible in the amount of outliers in the boxplot and the standard deviation within that category.

Finally, our last challenge concerned missing and incomplete data, which is not unusual when working with historical data. It is interesting that while our group considered the fact that the dataset only includes European personnel that signed on in the Netherlands as part of the dataset description, the other group saw this as a limitation. Apart from missing records concerning non-European personnel, not all administrative records were digitised, and some voyages were only partially recorded due to missing pay ledgers. This incompleteness affects the representativeness of the dataset. In addition, not all median wages were available in the dataset, and this missing information made it more difficult to compare economic status between ranks and how this may have contributed to mortality risk.

ETHICAL CONSIDERATIONS

The VOC was a trading company during colonial times, subjugating and enslaving native people for a profit. Our dataset contains data on European VOC personnel, people who introduced and benefitted from this system, while documentation concerning the lives of the people who lived under this system are often not even digitised (Zaagsma, 2022). However, from our research it has become clear that inequality was not only perpetuated by the VOC in foreign countries; poverty, vulnerability and structural inequality in Europe were also exploited. Above that, not all

Europeans were working on ships voluntarily — many were debtors forced to work off debts, criminals sentenced to serve on ships, or were otherwise coerced (Petram et al., 2024; Worden, 2009).

Projects such as GLOBALISE are also doing important work with VOC archives to research the VOC's violence and exploitation during this time. Lodewijk Petram, co-author of our dataset, has also worked with this archive. Studying *all* aspects of VOC is necessary to acknowledge colonial pasts and decolonise the production of knowledge.

Other ethical concerns are privacy and bias. Privacy is not a big concern: all identities are historical, and during our research the names of the VOC personnel were not relevant. Bias however, was a big issue. A lot of choices were made during the creation of the dataset. There is of course the previously mentioned "Dutch Bias", caused by Dutch VOC clerks, archivists and researchers that are more familiar with Dutch names and places, leading to overrepresentation of Dutch personnel and better data reliability for them This can make Dutch careers seem longer, making them appear less likely to perish than they were in reality. The conservative linking method used by Petram et al. (2024) also introduces bias, making careers appear shorter than they were in reality and artificially increasing death rates. We addressed these biases by being transparent about them.

RESULTS

Our analysis examined the relationship between workers' rank categories within the eighteenth-century Dutch East India Company (VOC) and their likelihood of mortality, with the goal of understanding how occupational hierarchy and working conditions reflected broader systems of inequality within the Company. Using the processed VOC "persons_contracts" and "ranks" datasets, we created summary visualizations in **R** to compare mortality proportions and wage distributions across consolidated rank groups.

Mortality Patterns by Rank Group

The results revealed notable differences in mortality across rank categories. Deaths were most frequent among individuals classified in the ***Military, Trade,*** and ***Other*** rank groups. The "Other" category largely consisted of specialized or hybrid occupations that did not fit cleanly into administrative, merchant-exclusive or seafaring classifications, often reflecting lower-status or less protected forms of labor. By contrast, ***Medical*** professionals and ***Ship*** officials exhibited the lowest mortality proportions. These findings suggest that workers performing on-board administrative or health-related duties were better protected from the dangers of manual or military labor, whether through improved living conditions, higher access to resources, or occupational privilege.

Wage Inequality and Vulnerability

An examination of wage data provided further evidence of structural inequality. Workers in the ***Military*** category (who also faced the highest mortality rates) earned the lowest average wages within the Company. In contrast, ***Ship*** and ***Medical*** professionals, whose mortality rates were some of the lowest, received comparatively higher pay. Rank groups categorized as ***Trade*** and ***Other*** displayed the ***largest wage variability****,* suggesting that within these categories, income was highly unequal and may have depended on specific skills, connections, or contract terms rather than standardized compensation.

Interpreting Structural Inequality within the Dutch VOC

Taken together, the mortality and wage analyses reveal a consistent pattern of systemic inequality within the VOC's labor organization. Those occupying lower-status or physically demanding roles faced both higher mortality risk and lower compensation. This pattern aligns with historical literature describing how the VOC's bureaucratic and early-capitalist structure depended on maintaining rigid hierarchies of poverty, dependence, and discipline. The Company's labor organization thus reproduced class divisions through occupational stratification at sea and outposts on land. Thus, this arrangement in which economic exploitation translated into heightened bodily vulnerability is reflective of the severe consequences of implementing a caste system driven by capitalistic goals.

Our quantitative results complement qualitative historical accounts of the VOC, demonstrating that inequality was not only ideological but materially embedded in workers' survival and compensation. Mortality risk within the Company mirrored the economic hierarchies that sustained its global operations, illustrating how class and occupation shaped life-and-death outcomes for early modern workers.

DOCUMENTATIONS AND SUSTAINABILITY

All datasets and code used in this project are fully documented. The enriched *VOC-opvarenden database* (Petram et al., 2024), is openly accessible via Zenodo for reproduction and citation purposes (DOI: 10.5281/zenodo.10599528). Metadata, including codebooks containing variable definitions,"cleaning" procedures, and data provenance are included.

Our data, literature review and scripts used for data processing, analysis and visualisation were written in R and are available on our public GitHub repository (https://github.com/SamanthaGTHB/Intro-to-DH-SH---Group-1-VOC) to ensure reproducibility and transparency. The repository also contains our project report, outlining our research question, methodology and analytical decisions.The results of this project provide a foundation for more in-depth research into mortality risks and socio-economic hierarchies among VOC personnel.

Regarding sustainability, we aimed to make our research as open and reusable as possible. GitHub provides version control, so data and scripts are recoverable, and code in R can be reproduced from this repository. The permanent Zenodo DOI ensures the long-term preservation and citability of the dataset. The permanent Zenodo DOI ensures the long-term preservation and citability of the dataset. These platforms safeguard transparency and sustain our research outcomes.

REFLECTION

Workflow effectiveness:

During our project, we encountered several obstacles that affected our workflow and made it less smooth than we initially hoped. At times, we became stuck on a particular step and had to revisit previous stages to resolve the issue. As previously explained, the most significant challenge we faced was the inconsistency and occasional unreliability of the data, which initially slowed our progress as we searched for effective solutions. However, our commitment to clear communication and teamwork was key to overcoming these problems. By breaking down each step of the process and collaboratively discussing our challenges and possible solutions, we managed to navigate the difficulties without excessive frustration.

Suggestions for future research:

To expand on our current project we have thought of some suggestions for future research that can contribute to an even richer analysis of how various factors shaped labor practices and inequalities in the eighteenth century.

- Examining the racial and ethnic composition of each rank group within the VOC, analyzing how these demographics influenced social hierarchies and inclusion. This research could also explore potential connections between race, ethnicity, and wage averages, shedding light on economic disparities and their implications for employee welfare.
- Conducting a comparative analysis between the VOC workforce and those of merchant companies from other countries, such as China and England. This approach could reveal differences in organizational structures, labor hierarchies, and risk management practices across cultural and national contexts. By examining these differences, the research can highlight how broader economic, political, and social factors shaped labor practices and identify patterns of inequality and inclusion within multinational trading networks.

REFERENCES

Petram, L., Koolen, M., Wevers, M., & Van Lottum, J. (2024). Charting Lives and Careers: Enriched Data About the Dutch East India Company's Eighteenth-Century European Workforce. *Journal Of Open Humanities Data*, *10*. https://doi.org/10.5334/johd.210

Worden, N. (2009). 'Below the Line the Devil Reigns': Death and Dissent aboard a VOC Vessel. *South African Historical Journal*, *61*(4), 702–730. https://doi.org/10.1080/02582470903500384

Zaagsma, G. (2022). Digital History and the Politics of Digitization. *Digital Scholarship in The Humanities*, *38*(2), 830–851. https://doi.org/10.1093/llc/fqac050