Documentation of the Workflow

Group 2 – Natalia, Rebecca, Boris, Jorge, Roza, Stas

Project Overview:

Our project aims to answer the question: **“To what extent did the geographical origin of VOC workers influence their wages and career opportunities within the company?"**

To approach this, we first had to get up to speed on the inner workings of their VOC , especially how wages were calculated and where employees worked. This basic understanding enabled us to better understand the data and contextualise our findings.

We divided our subquestions into two groups: one being Internal Analysis, that is, within the Netherlands, focusing on how location within the Netherlands affected wages and career development and one about the External comparison which looks at differences between Dutch workers and the rest of Europe.

External Hypotheses:

H1: Workers born in the Netherlands earned more than their counterparts from other European countries.

H2: Workers born in the Netherlands had better chances to advance in rank within the VOC as compared to those from other European countries.

Internal Hypotheses:

H1: Workers from larger cities within the Netherlands earned more than those from smaller towns or rural areas.

H2: Workers from larger cities had better chances to advance in rank within the VOC as compared to those from rural areas.

These hypotheses, alongside relevant literature to contextualise our findings, are discussed in detail in the result session.

Data Acquisition & Methodology

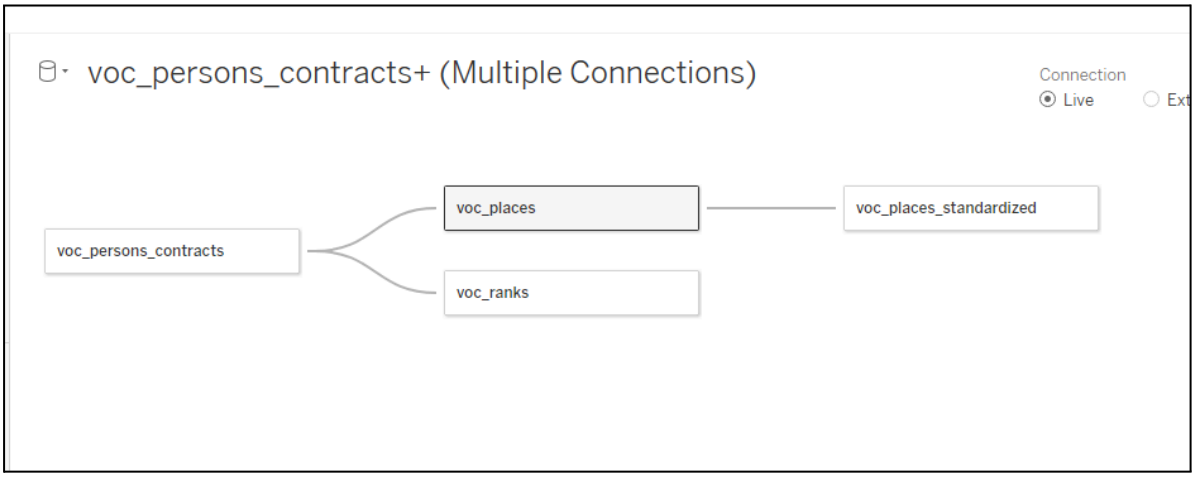
The Dutch East India Company's (VOC) historical records provide the dataset that we employed in our study.

The Excel files were quite huge. After forming our preliminary hypotheses, we came to the conclusion that the tables voc_persons_contracts, voc_ranks, voc_places, and voc_places_standardised contained all pertinent information. In Tableau, we made the decision to join these datasets.

By establishing a many-to-one relationship between Place Id and Place Id, we were able to link the voc_persons_contracts dataset to voc_places. Next, we established a many-to-one relationship between the variables both named Standardised Place Id in order to link voc_places and voc_standardised places. This made it possible for us to view the standardized form of each person's origin.

We linked the voc_persons_contracts dataset to voc_places by creating a many-to-one relationship between the two variables named Rank Id. This made it possible to visualise the median salary of each person.

A schematic that helps visualise the steps done is shown below.



We started by making some visualisations. Each visualisation made us reconsider how to handle the data and faced us with new limitations. We decided that it makes more sense to explain this case-by-case, since the process is not universal for each step, nor is it linear. So

for each visualisation we briefly described why and how we made it, what problems we faced and which insights it provided. Not all of the visualisations yielded interesting results, some led us to dead ends. Still, all of them are documented in the doc Visualisations.

Workflow Steps

At the outset, we investigated several possible correlations, including:

- Birth-in
- A sea-side municipality;
- An inland municipality;
- Class and nationality

Any career opportunities within the VOC for either Dutch or non-Dutch workers.

Our focus, therefore, became the examination of those factors affecting wages. After taking a closer look at our data set, we abandoned our initial plan of linking wage changes to worldwide economic cycles because wages seemed to be less volatile than we had expected, and the scope of such a task was too broad. By contrast, we turned our attention to a geographical point of view, checking whether proximity to the coast or the size of urban settlements lead to different wages. This we had to do instead of focusing on ethnic backgrounds because our dataset covers Indonesian and other Asian workers only incompletely or, at best, applies to the highest ranks of those workers only. Moreover, as confirmed in other studies, most workers were enslaved or otherwise compelled to work. Due to data limitations, we had to narrow it down to NL, and voilà, some new ideas were born, such as the visualisation of average wages in different parts of the Netherlands, and also a comparison of wages for the VOCs with those of all others. The differences between Dutch and non-Dutch staff working within the industry of the VOC might be assumed to be a sub-question, not part of the main research question.

This then became our visual representation to display our findings on the mean wage, represented as a thermal map across various geographies.

Now, for further explanation, an enlarged portion of the globe mainly focused on Europe will be shown for better understanding. This will be followed by a thermal map that will be

magnified only on the intricately detailed part of the Netherlands, which highlights our pertinent data.

We have also attempted a more in-depth analysis of our data through a statistical study using R between average wages within and outside the Netherlands.

However, one important problem we identified, if we wanted to carry out all these analyses, was the redundancy of the data we had, especially in the cases of individuals represented more than once with varying degree or contract matters. To overcome this challenge, we created a calculated field in Tableau by using the formula: { FIXED [Person Cluster Id]: MIN([Vocop Id]) }. This created the "Unique Persons" field that made sure each person was represented once, to give us cleaner analysis.

Regarding the articles and academic research, we could find two supportive documents for our statement, particularly one document that really emphasises the difference of city workers to those in the rural areas.

In conclusion, our research process involved:

1. Team collaboration and lots of planning
2. Dataset exploration and identifying limitations
3. Experimenting with visualisations
4. Formulating and revising our research questions
5. Refining our data interpretation through multiple iterations.

This process ultimately helped us gain a new level of understanding in our approach to the data.

Challenges and Solutions

One of the most important problems was incomplete data for some periods, especially for the early years of VOC. This did not allow us to take the results from the analysis in longitudinal terms all the time. Hence, we tried to narrow our analysis to the periods with the most complete series, particularly for the 18th century, for which more than 90% of the series was complete. Narrowing the scope allowed us to draw more reliable conclusions without having to interpolate the missing data.

Another challenge we had to face was the fact that, due to different contracts, the same individual was represented several times in the dataset; we struggled to identify which were unique people, and which were repeated records. That's what kept us busy for a few days since we were uncertain what the best course of action would be. It was hard to decide whether to analyse all items, dealing with each contract separately, or to focus only on single individuals. If all entries are used, the new recruits will be treated as separate individuals. On the other hand, if only single individuals are used, one of their positions has to be chosen, based on the assumption that the last occupation was the highest in rank while the first was the lowest. Based on that we proceed to create a calculated field in Tableau which counts each person once based on the Person Cluster ID.

Moreover, one issue was the omission of data and zero values regarding place of origin or quality which could distort the analysis. If these null values are included in the views, they could alter the results. In any case, since the value is not present, it is automatically excluded and grouped into the category called "non-value" in Tableau.
We carefully filtered entries with null values when required, especially in graphs that represented the geographic origin. Where this would make a big difference in the valid number of entries in a data set, we have included null values in a separate "non-value" category so as not to distort the representation.

In addition, during the data cleansing phase of our project, spellings were inconsistent and place names had variations in spelling, making geographical analysis of the workers' origins impossible. For example, the same or similar places may be recorded in different ways, which makes linkage of the data quite confusing.
We have, therefore, included the standardization of place names, adding to each Place Id entry the Standardized Place Id. This was imperative because different variations had to fall into one site. In doing this, we reduce the chances of misidentifying any worker's origin to a bare minimum.

Ethical Considerations

In our analysis of wage and rank flow within the VOC, we made the decision to exclude workers originating from Asia. Since the majority of that labour force is omitted from our dataset due to being enslaved and unpaid, and because only the highest ranks represent a widespread system of inequality under colonial control, we have chosen to fully exclude them for our study. Our project's reach is further limited because we are excluding a whole continent when we remove Asian workers from our salary research.
This creates an information gap, which makes any direct and fair comparison between European and Asian workers impossible. However, the inclusion of these few numbers would have resulted in misleading conclusions about the status and standing of Asian workers within the VOC hierarchy.

Results

Having created the different visualisations, we have seen the different possible results of our data.
The data appears to indicate several notable findings.
Pertinent and interesting results are those that look into conditions in the Netherlands. As would be expected, "bigger" is not necessarily "better." It becomes obvious that there is one region within the Netherlands, Zeeland, which has a higher proportion of promotions than any other region in the country. Zeeland is a unique area for perhaps many reasons: its citizens were closer to a recruitment centre; this proximity might have allowed for a wider social network, especially concerning people who also worked at sea.
Moreover, being a region with coastline and several harbours, it is likely that many had competencies in navigation and seafaring.
Finally, the Zeeland Chamber was the second largest in VOC, possibly offering more roles and opportunities than other regions.

In response to the first Internal hypothesis, based on the Heatmap we have created using the research by author Van Zanden, it would appear that people from big cities and from coastal areas were earning more. This could be explained by the fact that people in the countryside often worked in agricultural fields and the prospects of employment in VOC were not so

tempting. These jobs were only taken when there was no available alternative for the people. Furthermore, due to the proximity of residing in big cities to recruitment centres it was also easy to access the job opportunities offered by VOC. For instance, the company has helped decrease poverty and unemployment as it has employed many people involved in jobs offering very low incomes. This led to a shortage of women in Dutch society due to close to a million men who left, mainly for Indies of whom only half returned. Even though it accounted for only 6% of the Dutch merchant ships, the VOC employed a quarter of the Dutch seamen and accounted for a quarter of the miles travelled by Dutch ships(Poeze 1996, p. 381).

Another critical consideration is the European employees in the VOC. Most of these employees showed up as managers in almost every region that was far-flung. In as much as there were really harsh conditions, long distances, military conflicts, and diseases, these jobs were in great demand. However, because there were long-term contracts in addition to poverty in pre-industrial Europe, these jobs were alluring to those who wished to get a fortune, social and economic status advanced, provided they survived adversity (Rei, 2013, p. 28).

Speaking about the second internal hypothesis, the most interesting thing that can be remarked is that most of the cities where workers come from are either big cities or localities from Zeeland Province. If the existence of big cities was not so hard to explain, what really made us curious was why Zeeland registered such a high percentage of promotion.

The graph teaches us that, although one can observe a link between, on the one hand, the amount of workers from a country in the VOC and on the other, the relative amount of promotions, there are a lot of exceptions, and Germany is one of them. We may, however, with great certainty establish that being a Dutch worker was an enormous advantage: a Dutch worker had almost one in three chances of advancing within the VOC.

That brought us to the analysis within the Netherlands, focusing only on the modern borders. We included a filter to include only locations with more than 250 employees, for which the results are shown in the same graph below. Indeed, results show that the more populated areas, like Amsterdam and Rotterdam did indeed provide ample opportunities to climb up in the ranks of the VOC. But what really caught our attention was the region of the Zeeland peninsula, which had, by far, the largest number of promotions.

This can be explained by a number of factors: compensation in the VOC did not involve just fixed wages; in addition, most of the employees received business privileges, presents, and other forms of compensation from local jurisdictions, or financial benefits-of-position, as expressed in the literature(Rei 2013, p. 41). In fact, Rei shows that only 64% of the total remuneration took the form of wages, with the balance arriving from these other sources. Clearly, the more lucrative the position of the VOC worker, the more he benefited from the success of the company.

Instead, by focusing on the External Hypotheses, it is possible to highlight some specific results illustrated in our graphs.
There are several possible explanations why the average wage of Dutch workers is higher than that of foreigners.
Probably, Dutch employees had more "confidence" for holding positions with greater responsibilities. A more plausible theory still supposes that proficiency in the Dutch language was necessary to hold positions of leadership. Generally, Dutch workers were more endowed with their social networks, which might have helped them obtain better job opportunities compared to other European countries. This is further supported by a pie chart representing that Dutch employees have a higher percentage of well-paid jobs as compared to other employees from around the world.

One of the charts indicates that workers from Malaysia, Indonesia, and South Africa received appreciably higher wages compared to their counterparts in northwestern Europe. However, evidence shows that the non-European labour force was paid markedly low remunerations and many of them comprised the forced labour. The foreign workers were usually relegated to low-lining positions and needed to show far greater effort in their line of work to receive pay hikes or promotion orders. This skews our results since most of the Asian workforce was enslaved and undocumented. It also means that information from these countries is disproportionately influenced by the more lucrative job titles, or those held by unoccupied persons or persons of high standing.
(For a better understanding of our findings, we have provided a separate document with all the necessary visualisations and explanations)

Documentation from our meetings:

First meeting 12/09:

After we all sat down together for the first time and critically analysed our data set, we changed our plans in terms of research. We discovered that we do not have individual wages of workers, which makes it difficult to answer our initial question. We decided to change our research question to still examine wages, but using averaged numbers together with position data in our dataset. We thought this would allow us to explore wage dynamics between regions using average data instead of specific wages. So our new RQ became: 'Did place of birth influence a worker's opportunities in terms of salary and role within the VOC?'.

Second Group Meeting 16/09:

We worked on our Tableau files. We solved problems with standardisation and linking data between different databases.
However, a potential problem arose: The relationship between voc_places and voc_places_standardised does not match in one of our Tableau versions.
We started working on a map for visualising the nationalities of VOC employees. We started at several possible visualisation options. Looking at the map in Tableau we noticed that it used modern boundaries, so we wondered if this was a problem. A few questions started to arise such as: What is relevant?

Third meeting, 19/09:

We began to explore what the research pathways might be. We thought about:

- possible correlations between place of birth (coastal or landlocked cities)
- between rank and nationality
- or between the career opportunities of Dutch versus non-Dutch workers within the VOC.

We realised that we should do some research and come up with ideas and insights for the next meeting.

After that, we focused on the research question once again. We asked ourselves whether it was better to use the simpler version of our RQ is therefore: Were there significant differences in the wages and jobs of Dutch versus non-Dutch workers within the VOC?
Or: 'Did the place of birth influence the opportunities a worker had in terms of salary and role within the VOC?'.
But then talking, we decided that since we are not sure how difficult it would be to link the datasets, it would be better to focus on the simpler version of our research question first. After that, if this proves feasible, we will extend it to one of the possible paths.

Fourth Meeting, 30/09 :

We established that the data sets are very compatible for linking, and that it was easier than expected to do so. The relational schema below closely resembles the sketch. Here is what we did:
We linked the dataset voc_persons_contracts to voc_places by creating a many-to-one relationship between Place Id and Place Id. We then linked voc_places to voc places_standardized by creating a many-to-one relationship between the variables both named Place Standardized Id. This allowed us to see the standardized version of each person's place of origin.
We linked the dataset voc_persons_contracts to voc_places by creating a many-to-one relationship between the two variables named Rank Id. This made it possible to see each person's median wage.
Potential problems that raised: too many null values in e.g. wage by country graph - points to?

Update from a group meeting, 2\10 (After consultation with Erika):

After discussing some doubts we had with Erika, we realised that we should have taken another look at the cardinality. Instead of one-to-one as we said before, we set them as if they should all be many-to-one. We still need to verify this. We also need to check one more thing: can the same person be present several times in the dataset? Each row has a different voice id, but some have the same name. Question we asked ourselves: we can check if they come from the same place, but what else? In case the same person appears several times, the cardinality of the relationship from voc_persons_contracts to voc_ranks and to voc_places should be many-to-many.

We thought it was also important to read the document on our datasets carefully, as they could explain their decisions in designing the datasets and answer our questions.
Furthermore, we talked about the fact that we need to analyse where our null values come from and whether they correspond to the missing values in the datasets.

Group meeting, 3/10:

Together we developed some examples of specific sub-questions such as:

- How did wages vary between different regions in the Netherlands?
- By what factors are wages influenced? Can they be influenced by geographical factors, such as coastal or land origins?

We tried to come up with different visualisation approaches:
One idea was a heat map over time to visualise changes in personnel only during the 18th century, as more than 90% of our data relates to this period.
The use of a flow chart (Sankey diagram) to represent the routes was also discussed.

The use of complex visualisations from course documents for more advanced data representation was explored.
Since our dataset covers a 200-year span, it was difficult to compare the population sizes of cities over time. Standardised names make location analysis possible, but assigning exact population data would be complex.
Therefore, other hypotheses came up.
1. the idea of contrasting the richest areas in the Netherlands by mapping ‘average to average’ wages.
2. An expected comparison between VOC wages and general wages outside the VOC over the same period.
3. A hypothesis for the development of an NL-wide wage map

Update from a group meeting, 10/10:

We talked about how we could look at individuals with multiple entries, to see if their average wage has changed over time based on past experience.

A difficulty emerged in deciding whether to analyse all entries or only unique individuals. Ambiguity because we can look at all entries, which allows us to examine the distribution of jobs and wages. However, looking at these entries alone obscures the data on where these individuals come from (we will count the same person every time they are in the database). Therefore, either we use everything to answer the data on average wages, job distribution, etc., or we use the data that counts each person only once, which obscures the average wages and job distribution (it only considers an individual's first entry).

We considered analysing individuals with multiple entries in the dataset to assess whether their average wage has changed over time, expecting wages to increase with experience, and made some preliminary visualisations that we might use in our presentation.

Team contributions and backgrounds:

Team member	Background	Contributions
Natalia Bielecka	Student of MKDA	Project documentation Team organisation Tracking workflow documentation
Roza Bracic	Student of Mathematics	Dataset processing Tableau visualisations Data preprocessing
Boris Braun	Student of Communication & Information Sciences	Literature research Result Analysis and Contextualisation
Rebecca Casalini	Student of Cultural Anthropology and Development Sociology.	Project documentation Ethical concerns Research limitations Literature research
Jorge Crepo Rubio	Student of EBE	Dataset processing Tableau visualisations Data preprocessing
Stanisław Szumiński	Student of MKDA	Literature research Presentations Concluding literature

Sustainability

Looking forward, there are a number of ways in which this research might be expanded on. One crucial step would be an expansion of the data set to include detailed information on more nationalities. In particular, the Asian data could be fleshed out and the lower ranks better represented. In this way, much more can be learned about the varied labour force that supplied the VOC.

Other major improvements would be inclusions of fuller wage information coupled with an in-depth analysis of current relevant economic indicators for modern enterprises. These would lead the way to a more nuanced understanding of trends of wages by offering the links between past trends and present-day realities, hence allowing effective comparative studies to be made.

Eventually, a time-series analysis of career paths might help in assessing how experience may have influenced remuneration and professional development within the VOC. These factors would complement our test and allow a more correct presentation of social and economic mobility in the historical context of the VOC.

Reflection

Our research has confirmed the crucial role that the socio-economic factors of birthplace and citizenship have played in determining professional success for VOC officers and sailors. By combining a range of historical records and using statistical analysis, we were able to establish evident discrepancies in income and career advancement that were previously under-researched.

The progression of our analysis was remarkable, as it unfolded incrementally: each stage of the investigation enabled the identification of new data and elements of the puzzle that had not been contemplated at the outset. Consequently, this process facilitated the construction of a more comprehensive analysis than originally anticipated, owing to the ongoing enhancement of our methodology.

Nevertheless, a number of limitations have influenced our analysis, including the lack of complete information on individual wages and incomplete records for a number of historical periods. An important avenue of future research would therefore be the incorporation of detailed information on individual wages and including local workers employed by the VOC in the colonies, to provide a better-rounded view of the VOC workforce.

References for literature review

Pusztai, G. & Teszelszky, K. (2016). In Dienst van de VOC. Een voorlopige inventarisatie van Hongaren in dienst van de Verenigde Oost-Indische Compagnie (1602 - 1795). *Acta Neerlandica, 12,* 25-108. https://ojs.lib.unideb.hu/actaneer/article/view/10469

Poeze, H. A. (1989). Korte signaleringen. *Bijdragen Tot De Taal- Land- En Volkenkunde / Journal of the Humanities and Social Sciences of Southeast Asia*, *145*(2), 379–392. https://doi.org/10.1163/22134379-90003263

Rei, C. (2013). Careers and wages in the Dutch East India Company. *Cliometrica*, *8*(1), 27–48. https://doi.org/10.1007/s11698-013-0093-3