

Teoriorienteret metode

Homework week 37

1. Create a **tidy** spreadsheet/table listing the names of Danish monarchs with their birth- and death-date and start and end of their reign. They should be sortable by year of birth. Suitable source website is for example [here](#), but you can also use another source, provided you reference it. (Collaboration is welcome. Remember to attach this spreadsheet to Brightspace submission)

[SE VEDHÆFTET EXCEL ARK](#)

Link:

https://1drv.ms/x/s!ArmiJDUiWX_7zB7GBdxSe1667hOA

| Name | Birth | Death | Start_Reign | End_Reign |
|---------------------|-------|-------|-------------|-----------|
| Gorm den gamle | NA | 958 | 936 | 958 |
| Harald Blåtand | NA | 986 | 958 | 986 |
| Svend Tveskæg | NA | 1014 | 986 | 1014 |
| Harald 2 | NA | 1018 | 1014 | 1018 |
| Knud den Store | 995 | 1035 | 1018 | 1035 |
| Hardeknud | 1020 | 1042 | 1035 | 1042 |
| Magnus den Gode | 1024 | 1047 | 1042 | 1047 |
| Svend Estridsen | NA | 1074 | 1047 | 1074 |
| Harald Hen | NA | 1080 | 1074 | 1080 |
| Knud den Hellige | NA | 1086 | 1080 | 1086 |
| Oluf Hunger | NA | 1095 | 1086 | 1095 |
| Erik Ejegod | 1056 | 1103 | 1095 | 1103 |
| Niels | NA | 1134 | 1104 | 1134 |
| Erik Emune | NA | 1137 | 1134 | 1137 |
| Erik Lam | NA | 1146 | 1137 | 1146 |
| Svend Knud Valdemar | NA | NA | 1146 | 1157 |
| Valdemar den Store | 1131 | 1182 | 1157 | 1182 |
| Knud 4 | 1163 | 1202 | 1182 | 1202 |
| Valdemar Sej | 1170 | 1241 | 1202 | 1241 |
| Erik Plovpenning | 1216 | 1250 | 1241 | 1250 |
| Abel | 1218 | 1252 | 1250 | 1252 |
| Christoffer | 1219 | 1259 | 1252 | 1259 |
| Erik Klipping | 1249 | 1286 | 1259 | 1286 |
| Erik Menved | 1274 | 1319 | 1286 | 1319 |
| Christoffer 2 | 1276 | 1332 | 1319 | 1326 |
| Valdemar 3 | 1315 | 1364 | 1326 | 1329 |
| Christoffer | 1276 | 1332 | 1329 | 1332 |
| Valdemar Atterdag | 1320 | 1375 | 1340 | 1375 |

2. Does OpenRefine alter the raw data during sorting and filtering?

When you use OpenRefine for sorting and filtering data, it does not alter the raw data itself. Instead, it provides a way to view, manipulate, and work with your data in a user-friendly interface without making permanent changes to the original data source.

Data remains intact in its original form when you import it to OpenRefine. You can sort the data based on specific columns or criteria within OpenRefine, and this sorting operation only affects the way data is displayed within the application. It does not change the original order or structure of the data in your source file. You can apply filters to your data to show only specific rows that meet certain criteria.

| 53 rows | | | | | |
|----------|---------------------|-------|---------|------------------------------------|-----------|
| Show as: | | rows | records | Show: 5 10 25 50 100 500 1000 rows | |
| All | Name | Birth | Death | Start_Reign | End_Reign |
| 1. | Gorm den gamle | NA | 958 | 936 | 958 |
| 2. | Harald Blåtand | NA | 986 | 958 | 986 |
| 3. | Svend Tveskæg | NA | 1014 | 986 | 1014 |
| 4. | Harald 2 | NA | 1018 | 1014 | 1018 |
| 5. | Knud den Store | 995 | 1035 | 1018 | 1035 |
| 6. | Hardeknud | 1020 | 1042 | 1035 | 1042 |
| 7. | Magnus den Gode | 1024 | 1047 | 1042 | 1047 |
| 8. | Svend Estridsen | NA | 1074 | 1047 | 1074 |
| 9. | Harald Hen | NA | 1080 | 1074 | 1080 |
| 10. | Knud den Hellige | NA | 1086 | 1080 | 1086 |
| 11. | Oluf Hunger | NA | 1095 | 1086 | 1095 |
| 12. | Erik Ejegod | 1056 | 1103 | 1095 | 1103 |
| 13. | Niels | NA | 1134 | 1104 | 1134 |
| 14. | Erik Emune | NA | 1137 | 1134 | 1137 |
| 15. | Erik Lam | NA | 1146 | 1137 | 1146 |
| 16. | Svend Knud Valdemar | NA | NA | 1146 | 1157 |
| 17. | Valdemar den Store | 1131 | 1182 | 1157 | 1182 |
| 18. | Knud 4 | 1163 | 1202 | 1182 | 1202 |
| 19. | Valdemar Sej | 1170 | 1241 | 1202 | 1241 |
| 20. | Erik Plovpenning | 1216 | 1250 | 1241 | 1250 |
| 21. | Abel | 1218 | 1252 | 1250 | 1252 |
| 22. | Christoffer | 1219 | 1259 | 1252 | 1259 |
| 23. | Erik Klipping | 1249 | 1286 | 1259 | 1286 |
| 24. | Erik Menved | 1274 | 1319 | 1286 | 1319 |
| 25. | Christoffer 2 | 1276 | 1332 | 1319 | 1326 |
| 26. | Valdemar 3 | 1315 | 1364 | 1326 | 1329 |
| 27. | Christoffer | 1276 | 1332 | 1329 | 1332 |
| 28. | Valdemar Atterdag | 1320 | 1375 | 1340 | 1375 |

Sorting and filtering in OpenRefine are non-destructive, and it doesn't permanently remove any data from your source. It only controls what you see in the application's interface. The original data remains unchanged unless you explicitly apply a transformation to modify it.

So OpenRefine is a valuable tool for data cleaning and preparation without risking accidental data loss or permanent changes to your source data.

3. Fix the interviews dataset in OpenRefine enough to answer this question: "Which two months are reported as the most water-deprived/driest by the interviewed farmer households?"

October and September. For my result see the steps + screenshots below.

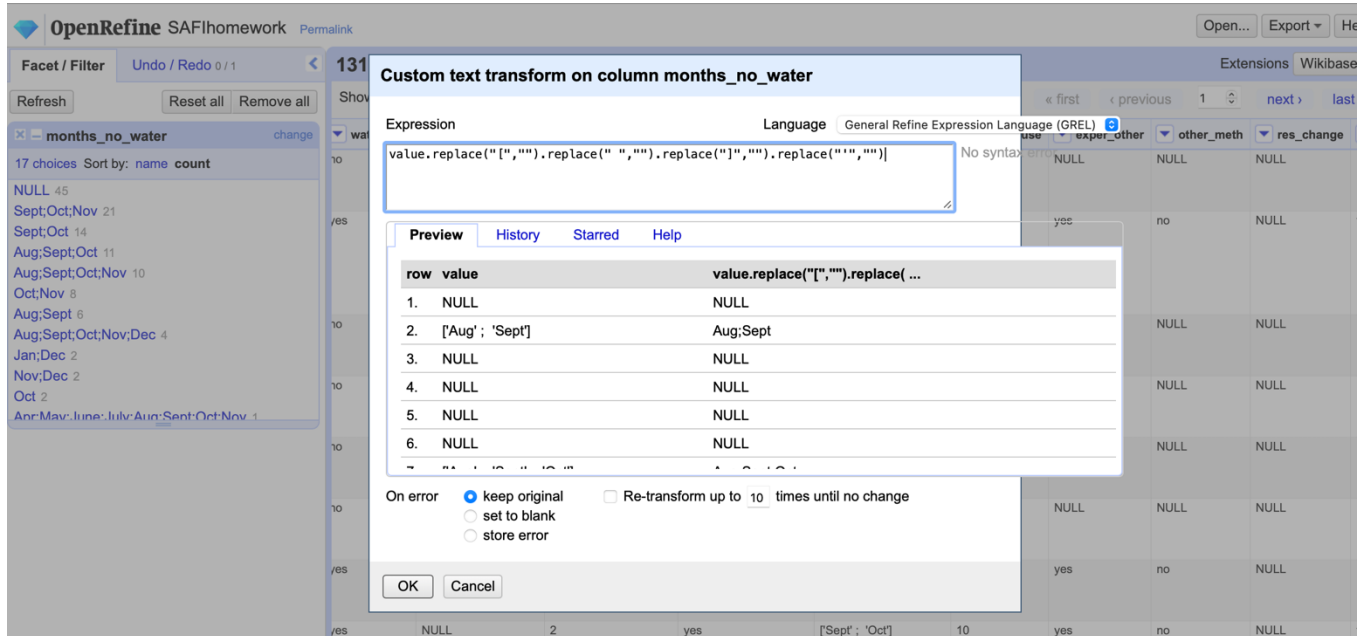
Steps:

1. I Imported the data to OpenRefine.
2. I found the variable “month_no_water”.
3. I clicked on the name and then clicked on “Facet”.
4. I clicked on “Text facet”.
5. So, I got the Facet/Filter box on the left.
6. I clicked on “count” on the facet box (se screenshot – red circle).
7. The box tells me that “Sept, Oct, Nov” is the most water-deprived months.

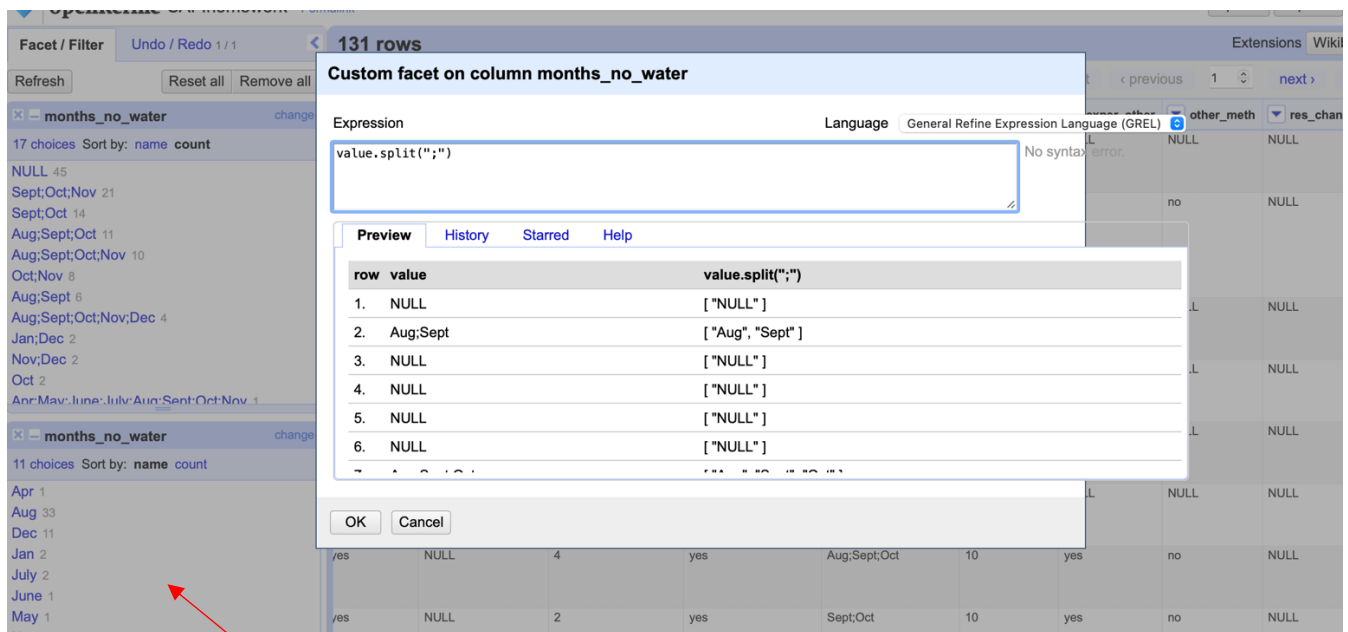
The screenshot shows the OpenRefine interface with the 'interviews dataset' loaded. The 'months_no_water' column is selected, and a text facet is applied. The facet box on the left shows 17 choices, with 'count' circled in red. The main table displays 131 rows of data. The columns are: count, water_use, no_group_count, yes_group_count, no_enough_water, months_no_water, period_use, exper_other, other_meth, and res_chang. The data shows various combinations of water use and months reported as water-deprived.

| count | water_use | no_group_count | yes_group_count | no_enough_water | months_no_water | period_use | exper_other | other_meth | res_chang |
|-------|-----------|----------------|-----------------|-----------------|------------------------|------------|-------------|------------|-----------|
| 2 | no | 2 | NULL | NULL | NULL | NULL | NULL | NULL | NULL |
| 3 | yes | NULL | 3 | yes | ['Aug'; 'Sept'] | 2 | yes | no | NULL |
| 1 | no | 1 | NULL | NULL | NULL | NULL | NULL | NULL | NULL |
| 3 | no | 3 | NULL | NULL | NULL | NULL | NULL | NULL | NULL |
| 2 | no | 2 | NULL | NULL | NULL | NULL | NULL | NULL | NULL |
| 1 | no | 1 | NULL | NULL | NULL | NULL | NULL | NULL | NULL |
| 4 | yes | NULL | 4 | yes | ['Aug'; 'Sept'; 'Oct'] | 10 | yes | no | NULL |
| 2 | yes | NULL | 2 | yes | ['Sept'; 'Oct'] | 10 | yes | no | NULL |

8. I wanted a more specific answer to the question, so I customized the facet to split the month groups, so the months became separated. I clicked “edit cells”, and then “transform”. I used regular expression to replace first. See screenshots below →



9. Then I split the months, by clicking “Facet” → “Custom text facet” →



10. A new text facet appeared on the left, where I could find my result by count.

11. Result. October (74) and September (70) are the most water-deprived month. (see screenshot below)



4. **OPTIONAL Real-Data-Challenge:** What are the 10 most frequent occupations (erhverv) among unmarried men and women in 1801 Aarhus? (hint: some expert judgement interpretation is necessary. As an inspiration, [check out this chapter Making a living outside marriage from the Swedish Gender and Work project of Maria Agren](#))

I used OpenRefine – and copied the URLs to open the data.

Get data from

[This Computer](#)

Web Addresses (URLs)

[Clipboard](#)

[Database](#)

[Google Data](#)

Enter one or more web addresses (URLs) pointing to data to download:

[Add another URL](#) [Next »](#)

The data looked like this:

OpenRefine AarhusErhverv Permalink

44559 rows

Show as: rows records Show: 5 10 25 50 100 500 1000 rows

Facet / Filter Undo / Redo

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? [Watch these screencasts](#)

| | All | ft | so | am | ld | lo | lo | by | fa | fn | en | ko | fa | al | ci |
|-----|------|-----|--------|----|----|-----|-----|----|----------|-------------|-------|---------------|----|-------|----|
| | | | | | | | | | | | | | | | |
| 1. | 1801 | Air | Aarhus | 1 | 1 | Air | Bye | 1 | Niels | Spand | mand | Huusbonde | 50 | gift | |
| 2. | 1801 | Air | Aarhus | 2 | 1 | Air | Bye | 1 | Mar | Møller | kvind | Hans Kone | 53 | gift | |
| 3. | 1801 | Air | Aarhus | 3 | 1 | Air | Bye | 1 | Niels | Sørensen | mand | Jensskarl | 50 | gift | |
| 4. | 1801 | Air | Aarhus | 4 | 1 | Air | Bye | 1 | Niels | Jørgensen | mand | Jensskarl | 67 | gift | |
| 5. | 1801 | Air | Aarhus | 5 | 1 | Air | Bye | 1 | Jørgen | Ellesen | mand | Jensskarl | 49 | gift | |
| 6. | 1801 | Air | Aarhus | 6 | 1 | Air | Bye | 1 | Peder | Ekkliden | mand | Jensskarl | 22 | ugift | |
| 7. | 1801 | Air | Aarhus | 7 | 1 | Air | Bye | 1 | Jens | Ekkliden | mand | Jensskarl | 18 | ugift | |
| 8. | 1801 | Air | Aarhus | 8 | 1 | Air | Bye | 1 | Rasmus | Jørgensen | mand | Jensskarl | 14 | ugift | |
| 9. | 1801 | Air | Aarhus | 9 | 1 | Air | Bye | 1 | Johannes | Abrahamsen | mand | Jensskarl | 11 | ugift | |
| 10. | 1801 | Air | Aarhus | 10 | 1 | Air | Bye | 1 | Else | Jensskarl | kvind | Jensskarl | 27 | ugift | |
| 11. | 1801 | Air | Aarhus | 11 | 1 | Air | Bye | 1 | Rig | Jensskarl | kvind | Jensskarl | 23 | ugift | |
| 12. | 1801 | Air | Aarhus | 12 | 1 | Air | Bye | 1 | Dorthe | Nielsdatter | kvind | Jensskarl | 22 | ugift | |
| 13. | 1801 | Air | Aarhus | 13 | 1 | Air | Bye | 1 | Inger | Jensskarl | kvind | Jensskarl | 20 | ugift | |
| 14. | 1801 | Air | Aarhus | 14 | 2 | Air | Bye | 2 | Bo | Nielsdatter | kvind | Madmoder | 42 | enke | |
| 15. | 1801 | Air | Aarhus | 15 | 2 | Air | Bye | 2 | Søren | Rosenmeyer | mand | hendes søn af | 16 | ugift | |
| 16. | 1801 | Air | Aarhus | 16 | 2 | Air | Bye | 2 | Niels | Rosenmeyer | mand | hendes søn af | 14 | ugift | |
| 17. | 1801 | Air | Aarhus | 17 | 2 | Air | Bye | 2 | Mar | Rosenmeyer | kvind | hendes søn af | 12 | ugift | |
| 18. | 1801 | Air | Aarhus | 18 | 2 | Air | Bye | 2 | Anne | Sophie | kvind | hendes søn af | 9 | ugift | |

I clicked on "include" in the text facet on the "civilstand" I wanted. In this case "ugift".

OpenRefine AarhusErhverv Permalink

Facet / Filter Undo / Redo 1/1 25440 matching rows (50034 total)

Refresh Reset all Remove all Show as: rows records Show: 5 10 25 50 100 500 1000 rows « first < previous

ugift 25440 include exclude

gift 16617

enke 1531

enkemand 694

skilt 6

separeret 5

(blank) 5741

Facet by choice counts

| mt | id | loknr | lokalitet | bygning | famn | fnavn | enavn | koen | famstand | alder | civilstand |
|----|----|-------|-----------|---------|------|---------------|---------------|--------|---------------|-------|------------|
| 6 | 1 | | Alrøe Bye | | 1 | Peder | Eskildsen | mand | tjenestekarl | 22 | ugift |
| 7 | 1 | | Alrøe Bye | | 1 | Jens | Eskildsen | mand | tjenestedreng | 18 | ugift |
| 8 | 1 | | Alrøe Bye | | 1 | Rasmus | Jørgensen | mand | tjenestedreng | 14 | ugift |
| 9 | 1 | | Alrøe Bye | | 1 | Johannes | Abrahamsen | mand | tjenestedreng | 11 | ugift |
| 10 | 1 | | Alrøe Bye | | 1 | Eise | Jeppesdatter | kvinde | tjenestepige | 27 | ugift |
| 11 | 1 | | Alrøe Bye | | 1 | Inger | Jespersdatter | kvinde | tjenestepige | 23 | ugift |
| 12 | 1 | | Alrøe Bye | | 1 | Dorthe Sophie | Nielsdatter | kvinde | tjenestepige | 22 | ugift |
| 13 | 1 | | Alrøe Bye | | 1 | Inger | Jørgensdatter | kvinde | tjenestepige | 20 | ugift |

Then I open a new text facet for "erhverv" and get a lot of mess, which needed to be cleaned. First, I separated people with several different "erhverv" by using: "edit cells" → "split multi-valued cells".

Split multi-valued cells

How to split multi-valued cells

☒ by separator

Separator ☐ regular expression

☐ by field lengths

List of integers separated by commas, e.g., 5, 7, 15

☐ by transition from lowercase to uppercase

[11Abc, Def22]

☐ Reverse splitting order

[11A, bcD, ef22]

☐ by transition from numbers to letters

[11, AbcDef22]

☐ Reverse splitting order

[11AbcDef, 22]

OK Cancel

And then I just kept cleaning... with multiple tools. Ex. "Cluster and edit".

Cluster and edit column "erhverv"

Find groups of different cell values that might be other representations of the same thing. For example, "New York" and "new york" likely refer to the same concept and just differ by capitalization, and "Gödel" and "Godel" probably refer to the same person. Find out more...

Method Key collision Keying function Fingerprint 111 clusters found

| Cluster size | Row count | Values in cluster | Merge? | New cell value |
|--------------|-----------|---|--------------------------|--------------------------|
| 5 | 16 | <ul style="list-style-type: none">nyder Almissee af Sognet (10 rows)Nyder Almissee af Sognet (2 rows)nyder almissee af sognet (2 rows)Nyder Almissee af Sognet,nyder almissee af Sognet | <input type="checkbox"/> | nyder Almissee af Sognet |
| 3 | 104 | <ul style="list-style-type: none">nationalsoldat (62 rows)Nationalsoldat (37 rows)National-Soldat (5 rows) | <input type="checkbox"/> | nationalsoldat |
| 3 | 5 | <ul style="list-style-type: none">Landrecrut (3 rows)Land-Recrutlandrecrut | <input type="checkbox"/> | Landrecrut |
| 3 | 12 | <ul style="list-style-type: none">lever af sine midler (10 rows)Lever af sine Midlerlever af sine Midler | <input type="checkbox"/> | lever af sine midler |
| 3 | 61 | <ul style="list-style-type: none">Landsoldat (27 rows)landsoldat (24 rows)Land-Soldat (10 rows) | <input type="checkbox"/> | Landsoldat |

Select all Deselect all Export clusters Merge selected & re-cluster Merge selected & Close Close

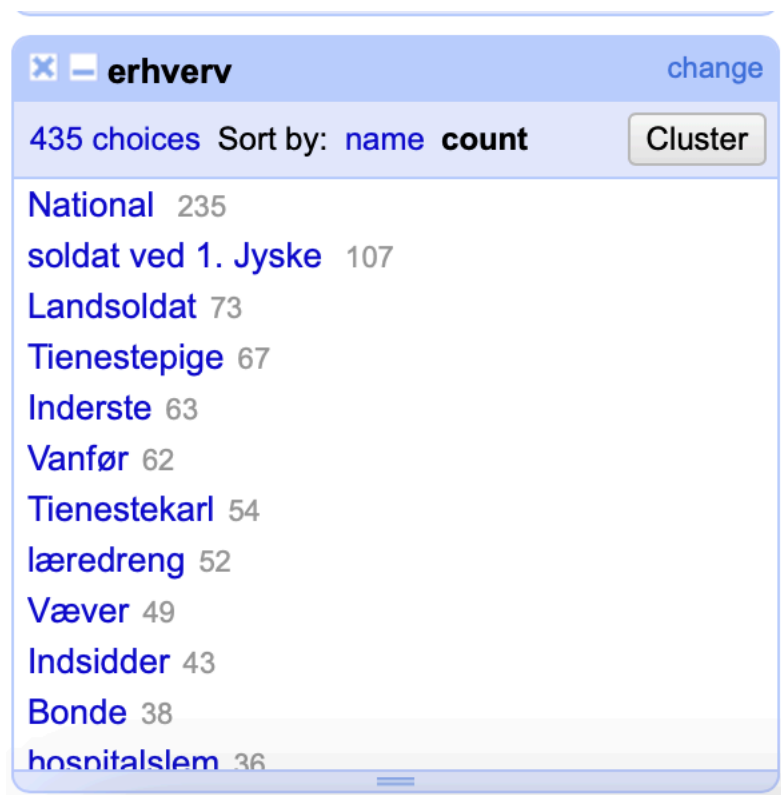
Choices in cluster 2 — 5

Rows in cluster 0 — 110

Average length of choices 3 — 33

Length variance of choices 0 — 1

After more time cleaning, I got all the possible “erhverv” down to 435 choices. So, this is what my result was after some time... I didn’t get further with cleaning.



The screenshot shows a web application window with a blue header bar containing the title 'erhverv' and a 'change' link. Below the header, there is a section with '435 choices' and a 'Sort by:' dropdown menu set to 'name', with a 'count' column header. A 'Cluster' button is located to the right of the sort options. The main content area displays a list of professions and their corresponding counts, sorted alphabetically by name. The list includes: National (235), soldat ved 1. Jyske (107), Landsoldat (73), Tienestepige (67), Inderste (63), Vanfør (62), Tienestekarl (54), læredreng (52), Væver (49), Indsidder (43), Bonde (38), and hosnitalslem (36). The list is partially obscured by a scroll bar on the right.

| | count |
|---------------------|-------|
| National | 235 |
| soldat ved 1. Jyske | 107 |
| Landsoldat | 73 |
| Tienestepige | 67 |
| Inderste | 63 |
| Vanfør | 62 |
| Tienestekarl | 54 |
| læredreng | 52 |
| Væver | 49 |
| Indsidder | 43 |
| Bonde | 38 |
| hosnitalslem | 36 |