Cultural Data Science 2022
Rikke Uldbæk
202007501

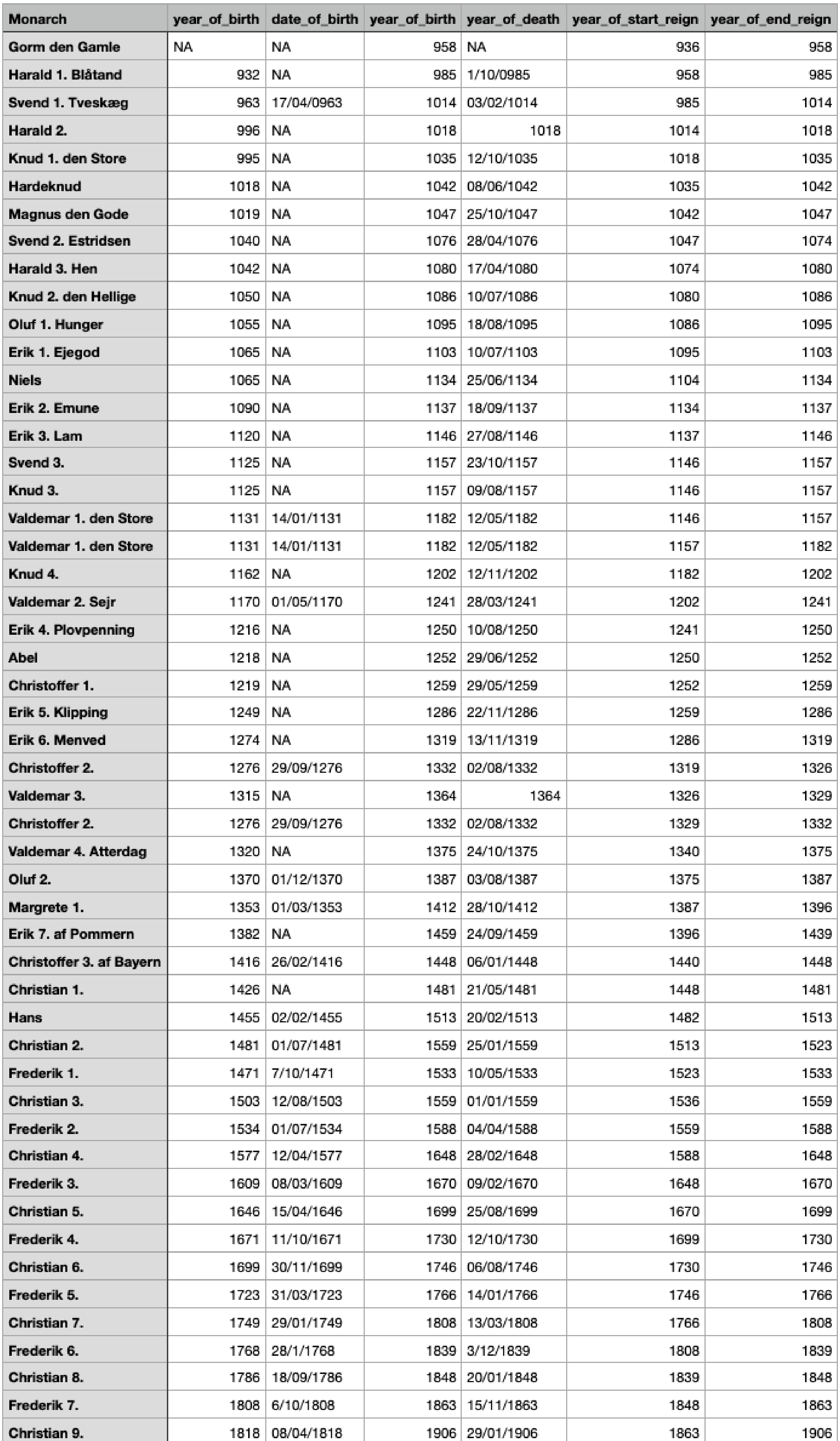
2:W35: Open Refine

DESCRIPTION

Upload your answers to these questions:

1. **Create a spreadsheet listing the names of Danish monarchs with their birth- and death-date and start and end year of reign. Make it *tidy*! They should be sortable by year of birth. Suitable source websites are here and here, but you can also use another source, provided you reference it. (Group collaboration is expected and welcome. Remember to attach this spreadsheet to Brightspace submission)**

Cultural Data Science 2022
Rikke Uldbæk
202007501

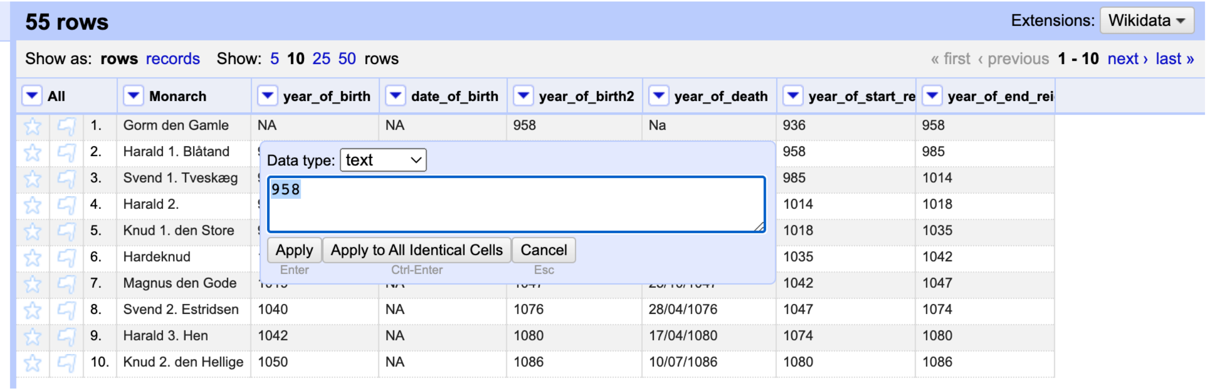


Monarch	year_of_birth	date_of_birth	year_of_birth	year_of_death	year_of_start_reign	year_of_end_reign
Gorm den Gamle	NA	NA	958	NA	936	958
Harald 1. Blåtand	932	NA	985	1/10/0985	958	985
Svend 1. Tveskæg	963	17/04/0963	1014	03/02/1014	985	1014
Harald 2.	996	NA	1018	1018	1014	1018
Knud 1. den Store	995	NA	1035	12/10/1035	1018	1035
Hardeknud	1018	NA	1042	08/06/1042	1035	1042
Magnus den Gode	1019	NA	1047	25/10/1047	1042	1047
Svend 2. Estridsen	1040	NA	1076	28/04/1076	1047	1074
Harald 3. Hen	1042	NA	1080	17/04/1080	1074	1080
Knud 2. den Hellige	1050	NA	1086	10/07/1086	1080	1086
Oluf 1. Hunger	1055	NA	1095	18/08/1095	1086	1095
Erik 1. Ejegod	1065	NA	1103	10/07/1103	1095	1103
Niels	1065	NA	1134	25/06/1134	1104	1134
Erik 2. Emune	1090	NA	1137	18/09/1137	1134	1137
Erik 3. Lam	1120	NA	1146	27/08/1146	1137	1146
Svend 3.	1125	NA	1157	23/10/1157	1146	1157
Knud 3.	1125	NA	1157	09/08/1157	1146	1157
Valdemar 1. den Store	1131	14/01/1131	1182	12/05/1182	1146	1157
Valdemar 1. den Store	1131	14/01/1131	1182	12/05/1182	1157	1182
Knud 4.	1162	NA	1202	12/11/1202	1182	1202
Valdemar 2. Sejr	1170	01/05/1170	1241	28/03/1241	1202	1241
Erik 4. Plovpenning	1216	NA	1250	10/08/1250	1241	1250
Abel	1218	NA	1252	29/06/1252	1250	1252
Christoffer 1.	1219	NA	1259	29/05/1259	1252	1259
Erik 5. Klipping	1249	NA	1286	22/11/1286	1259	1286
Erik 6. Menved	1274	NA	1319	13/11/1319	1286	1319
Christoffer 2.	1276	29/09/1276	1332	02/08/1332	1319	1326
Valdemar 3.	1315	NA	1364	1364	1326	1329
Christoffer 2.	1276	29/09/1276	1332	02/08/1332	1329	1332
Valdemar 4. Atterdag	1320	NA	1375	24/10/1375	1340	1375
Oluf 2.	1370	01/12/1370	1387	03/08/1387	1375	1387
Margrete 1.	1353	01/03/1353	1412	28/10/1412	1387	1396
Erik 7. af Pommern	1382	NA	1459	24/09/1459	1396	1439
Christoffer 3. af Bayern	1416	26/02/1416	1448	06/01/1448	1440	1448
Christian 1.	1426	NA	1481	21/05/1481	1448	1481
Hans	1455	02/02/1455	1513	20/02/1513	1482	1513
Christian 2.	1481	01/07/1481	1559	25/01/1559	1513	1523
Frederik 1.	1471	7/10/1471	1533	10/05/1533	1523	1533
Christian 3.	1503	12/08/1503	1559	01/01/1559	1536	1559
Frederik 2.	1534	01/07/1534	1588	04/04/1588	1559	1588
Christian 4.	1577	12/04/1577	1648	28/02/1648	1588	1648
Frederik 3.	1609	08/03/1609	1670	09/02/1670	1648	1670
Christian 5.	1646	15/04/1646	1699	25/08/1699	1670	1699
Frederik 4.	1671	11/10/1671	1730	12/10/1730	1699	1730
Christian 6.	1699	30/11/1699	1746	06/08/1746	1730	1746
Frederik 5.	1723	31/03/1723	1766	14/01/1766	1746	1766
Christian 7.	1749	29/01/1749	1808	13/03/1808	1766	1808
Frederik 6.	1768	28/1/1768	1839	3/12/1839	1808	1839
Christian 8.	1786	18/09/1786	1848	20/01/1848	1839	1848
Frederik 7.	1808	6/10/1808	1863	15/11/1863	1848	1863
Christian 9.	1818	08/04/1818	1906	29/01/1906	1863	1906

b. CSV will be attached in the submission

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

    a. When opening the raw monarch data in OpenRefine, every observation seems to be made into text. Thus all numerical observations suddenly become converted into text, which disturbs the data organization.



55 rows

Extensions: Wikidata

Show as: **rows** records Show: 5 **10** 25 50 rows « first ‹ previous **1 - 10** next › last »

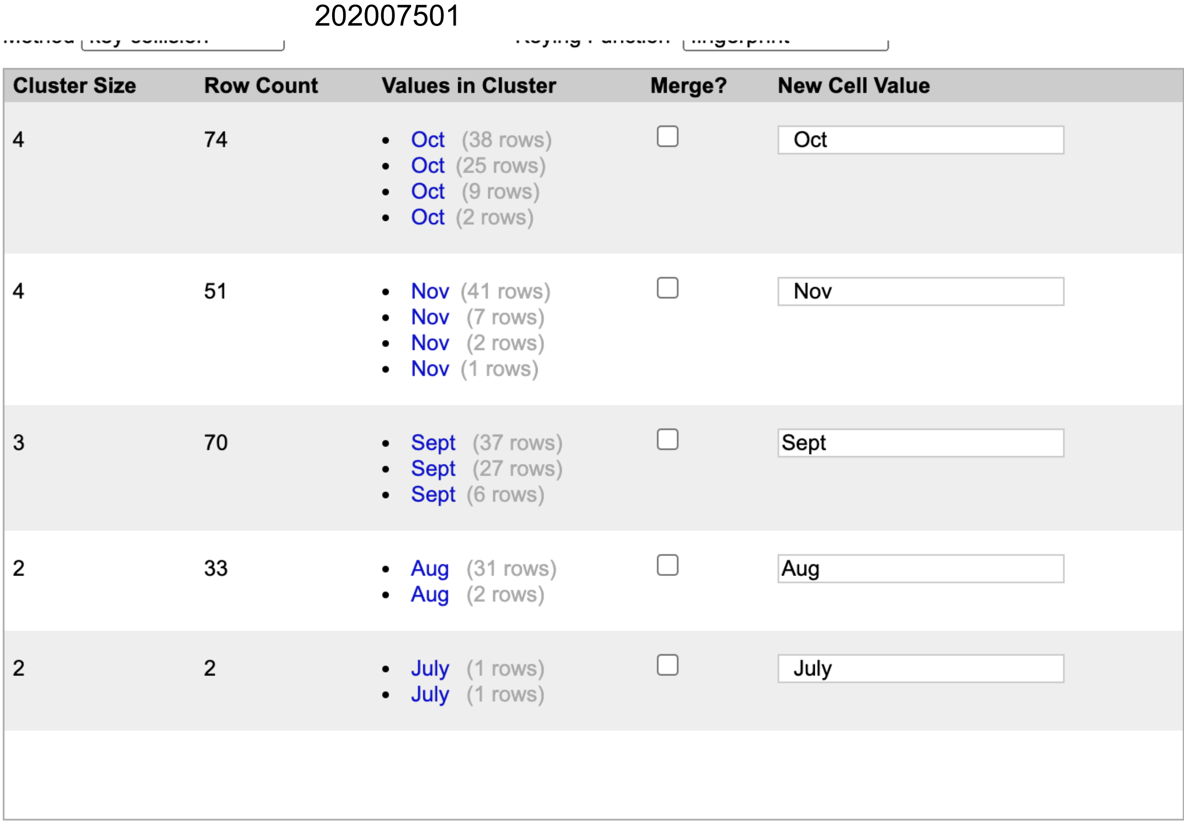
All			Monarch	year_of_birth	date_of_birth	year_of_birth2	year_of_death	year_of_start_re	year_of_end_rei
☆		1.	Gorm den Gamle	NA	NA	958	Na	936	958
☆		2.	Harald 1. Blåtand	[illegible]				958	985
☆		3.	Svend 1. Tveskæg	[illegible]				985	1014
☆		4.	Harald 2.	[illegible]				1014	1018
☆		5.	Knud 1. den Store	[illegible]				1018	1035
☆		6.	Hardeknud	[illegible]				1035	1042
☆		7.	Magnus den Gode	[illegible]	[illegible]	[illegible]	[illegible]	1042	1047
☆		8.	Svend 2. Estridsen	1040	NA	1076	28/04/1076	1047	1074
☆		9.	Harald 3. Hen	1042	NA	1080	17/04/1080	1074	1080
☆		10.	Knud 2. den Hellige	1050	NA	1086	10/07/1086	1080	1086

Data type: text

958

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?"**

    a. Using the filtering functions to remove "['']", i.e, value.replace("['']", ""), I am able to count the most water-deprived months. The two most water deprived months are October and September, as September occurs 70 times in the dataset and October occurs 74 times in the dataset as seen below.

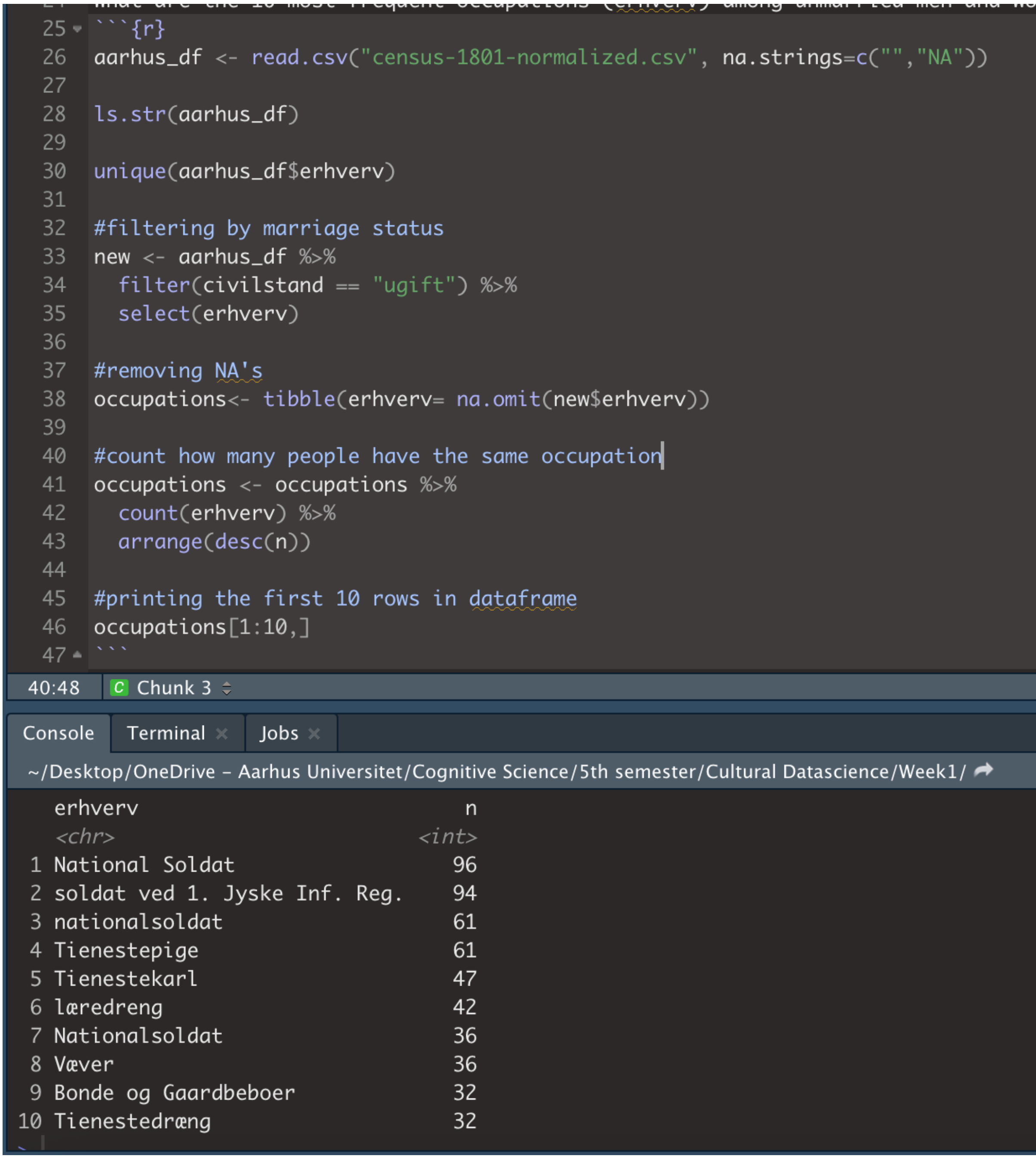


Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
4	74	• Oct (38 rows) • Oct (25 rows) • Oct (9 rows) • Oct (2 rows)	□	Oct
4	51	• Nov (41 rows) • Nov (7 rows) • Nov (2 rows) • Nov (1 rows)	□	Nov
3	70	• Sept (37 rows) • Sept (27 rows) • Sept (6 rows)	□	Sept
2	33	• Aug (31 rows) • Aug (2 rows)	□	Aug
2	2	• July (1 rows) • July (1 rows)	□	July

b.

4. **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, look at the HISCO classification "Historical International Standard of Classification of Occupations" on Dataverse if ambitious)**
    a. Below are listed the top 10 most frequent occupations among unmarried men and women in 1801 Aarhus.
    b. Note that number 1, 2, 3, and 7 of the observations on the list are pretty much the same occupation, however it still implies that the most frequent occupation seems to be soldier..

c.



~~~
What are the 10 most frequent occupations (erhverv) among unmarried men and wo
```{r}
aarhus_df <- read.csv("census-1801-normalized.csv", na.strings=c("","NA"))

ls.str(aarhus_df)

unique(aarhus_df$erhverv)

#filtering by marriage status
new <- aarhus_df %>%
  filter(civilstand == "ugift") %>%
  select(erhverv)

#removing NA's
occupations<- tibble(erhverv= na.omit(new$erhverv))

#count how many people have the same occupation
occupations <- occupations %>%
  count(erhverv) %>%
  arrange(desc(n))

#printing the first 10 rows in dataframe
occupations[1:10,]
```
~~~

~/Desktop/OneDrive - Aarhus Universitet/Cognitive Science/5th semester/Cultural Datascience/Week1/

```
   erhverv                         n
   <chr>                       <int>
 1 National Soldat                96
 2 soldat ved 1. Jyske Inf. Reg.  94
 3 nationalsoldat                 61
 4 Tienestepige                   61
 5 Tienestekarl                   47
 6 læredreng                      42
 7 Nationalsoldat                 36
 8 Væver                          36
 9 Bonde og Gaardbeboer           32
10 Tienestedræng                  32
```