



Digital Archives and Methods

Danish Prison Development

Kristoffer Segerstrøm, Emil Gert Hansen
and Lukas Barnewitz Benner

Group Name: 30, Group Number:13



Indholdsfortegnelse

1.0 INTRODUCTION/HYPOTHESIS	2
1.1 BACKGROUND AND MOTIVATION	2
1.2 HYPOTHESIS	3
2.0 METHODS	4
2.1 RESEARCH DESIGN	4
2.2 SOFTWARE FRAMEWORK	4
2.3 DATA ACQUISITION AND PROCESSING	4
3.0 FINDINGS	6
3.1 PRESENTATION AND ILLUSTRATION	6
4.0 DISCUSSION/CRITICAL EVALUATION	11
4.1 SIGNIFICANCE OF FINDINGS	11
4.2 EVALUATION	16
4.3 CONSIDER UTILITY/REPRESENTATIVENESS	17
4.4 LESSONS LEARNT	18
5.0 CONCLUSIONS AND RECOMMENDATIONS	20
AUTHOR CONTRIBUTIONS	21
.....	21
REFERENCES	22
TABLE 1 – SOFTWARE METADATA	24
TABLE 2 – DATA METADATA	25
6.0 PORTFOLIO	26

1.0 Introduction/hypothesis

This research intends to investigate the changes in imprisonment over time in Denmark and the social and political influences that may have affected these changes. The main research question is: *"How has the use of imprisonment in Denmark changed over time, and what social or political factors may have influenced these changes? Are there differences in trends across gender, age, or type of offense?"* We will analyse these dimensions to illustrate how the use of imprisonment has changed within Denmark and what larger social forces have influenced such changes.

1.1 Background and motivation

The Danish Penal Code has experienced a comprehensive and ongoing evolution since its initial consolidation in Christian V's Danish Code (Danske Lov) of 1683, reflecting changes in legal principles, societal norms, and legislative reforms over time. The law was characterized by very harsh punishment, even harsher than those that had been applied in earlier times. It was not until the late 18th century that a shift toward milder punishment began to emerge. This development is evident in the Theft Ordinance (Tyveriforordningen) of 1789 and in reforms introduced around the 1840s. In 1866, a civil penal code was enacted, in which the theory of retribution became the fundamental principle. Offenders were to be punished because the punishment was seen at the criminal's rightful compensation for their actions. However, the severity of punishment was significantly reduced compared to earlier legislation. In 1905, the possibility of suspended sentences was introduced, and a new Penal Code was adopted in 1930, coming into force in 1933. This legislation abolished the death penalty as well as other corporal punishment. Although the death penalty was formally abolished, only four executions had been carried out in Denmark since the 1866 Penal Code, with the last taking place in 1892. The death penalty was reintroduced during the German occupation and in the immediate post-war period, targeting those who had committed serious crimes in service of the occupying forces. However, we have chosen not to include this in our analysis, as it occurred under a state of emergency and in the aftermath of the war. Including it would not provide an accurate representation of

Denmark in the period leading up to these events.¹ After the war, there was a strong focus on resocialization as a means of preventing individuals from ending up in prison. However, this approach began to change in the 1990s. A new mindset emerged, emphasizing harsher punishment as a deterrent to criminal behaviour. For the first time since the Danish Code (Danske Lov), this marked a shift toward increasing, rather than reducing, the severity of sentences.² Due to these changes in the Danish legal system, this paper will focus on examining whether these measures have led to a significant change in the Danish conviction and prison system. The paper will also consider which later initiatives were introduced and their possible impact.

1.2 Hypothesis

This project is hypothesized to show that the use of prison in Denmark has changed over time. We hypothesize that the general Danish justice system has evolved from being more about punishment to focusing more on helping people coming back from a life of crime (rehabilitation). We also think that the way people are punished depends on factors like gender, age, and type of crime. Our hypothesis also states our belief that this happens because of social attitudes and political decisions. The goal of our project is to show how these differences affect punishment in the Danish justice system.

¹ Gyldendal og Politikens Danmarkshistorie. "Forbrydelse og straf." *Lex.dk*. Accessed May 21, 2025. https://gyldendalogpolitikensdanmarkshistorie.lex.dk/Forbrydelse_og_straf.

² Advokatsamfundet. "Man kan jo ikke blive ved med at kriminalisere og skærpe straffene." *Advokatsamfundet.dk*. Published April 2, 2024. Accessed May 19, 2025. <https://www.advokatsamfundet.dk/nyheder-medier/nyheder/2024/man-kan-jo-ikke-blive-ved-med-at-kriminalisere-og-skaerpe-straffene/>.

2.0 Methods

2.1 Research design

This study is a quantitative, nomothetic analysis of the development of convictions and imprisonments in Denmark, based on a historical perspective. We want to identify patterns and trends over time, from 1900 to 2024. The data that we have used, is from the website Danmarks Statistik.

2.2 Software framework

We wrote the code for this project on a MacBook Air with an Apple M3 chip, 16GB RAM, running macOS Sequoia (Version 15.4.1). I worked within a virtualized environment running Ubuntu 20.04 (Focal) and used the desktop version of R (4.4.3) and RStudio.

2.3 Data Acquisition and Processing

The study is based on data from *Årbøger for Kriminalstatistik* (Statistical Yearbooks of Criminal Justice), where we found yearbooks that report on incarceration rates and criminal offenses, disaggregated by age and gender.³ To observe the long-term trends, we selected data points from six years — 1900, 1925, 1950, 1975, 2000, and 2024 — providing consistent 25-year intervals. This allowed us to see whether incarceration and crime rates have increased or decreased across time.

The raw figures were first entered manually into Excel spreadsheets and then imported into R-Studio for processing and analysis. In R-Studio, we conducted all data filtering, transformation, and visualization. Column names were standardized, and individuals were grouped into broader age categories to allow for cleaner comparative analysis. For gendered visualizations, we chose consistent colour codes, using *lightcoral* for female and *steelblue* for male.

³ Danmarks Statistik, “STRAF40: Criminal Offences by Type, Time, and Sex,” *Statistikbanken.dk*, accessed May 15, 2025, <https://www.statistikbanken.dk/STRAF40>.

To control for population growth over time, we retrieved annual population estimates for Denmark from 1900 to 2024. The population increased from approximately 2.4 million in 1900 to nearly 6 million in 2024. To make crime and incarceration data comparable across time, we normalized all values per 100,000 inhabitants using the formula:

$$\text{Number per 100,000 inhabitants} = \left(\frac{\text{Number}}{\text{Total population}} \right) \times 100,000$$

Working the data in this way was important to help us understand whether changes in imprisonment patterns reflected actual shifts in policy or social behaviour, rather than just demographic growth. The workflow itself combined self-written data, structured spreadsheet formatting, and consistent charts in R-Studio.

3.0 Findings

3.1 Presentation and illustration



Figure 1 shows the number of people incarcerated in Danish prisons per 100,000 inhabitants. Lines 151-189 in R-file.

Figure 1 shows the number of incarcerated people in Denmark from 1900 to 2024 per 100,000 inhabitants, and it shows a large growth in imprisonment until the year 2000. The plot was made by combining the male and female datasets into one single dataset, as the Excel-file is a combination of statistics between male and female imprisonments. By doing this we allow an overall view upon how the prison trend has evolved to take place⁴.

⁴ James Long and Paul Teetor, *R Cookbook*, 2nd ed. (Sebastopol, CA: O'Reilly Media, 2021), chap. 5.20, 6.6, and 10.13.

Imprisonments per 100,000 Inhabitants Over Time by Gender

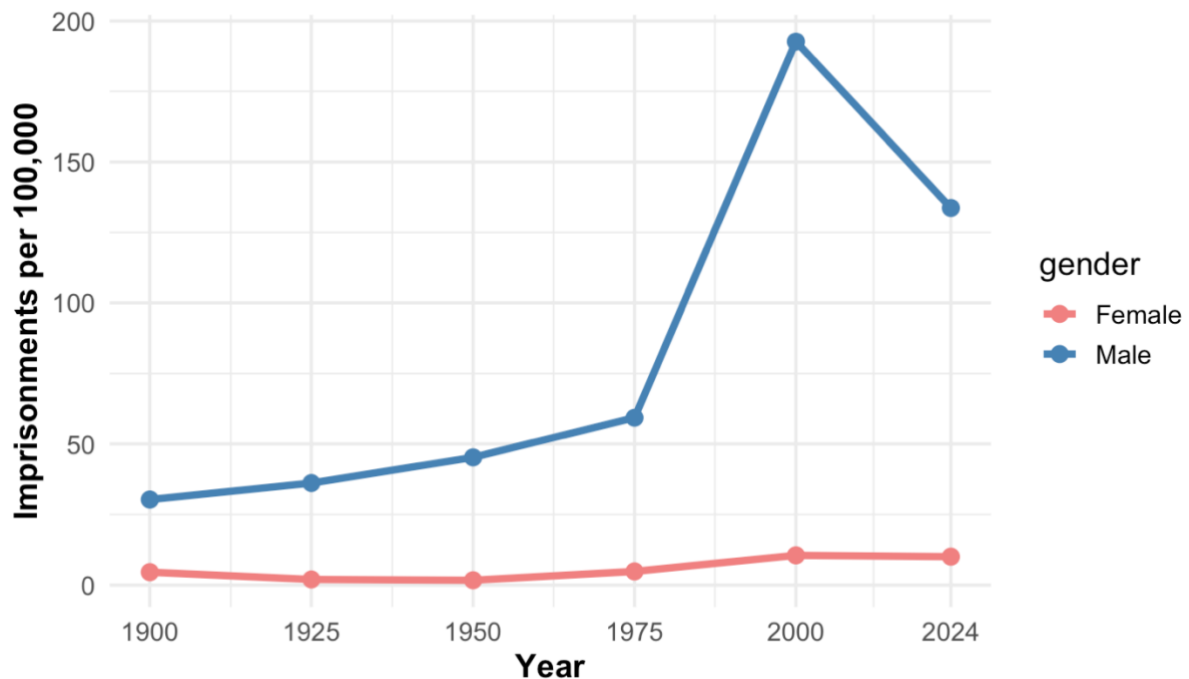


Figure 1: Imprisonment per 100,000 Inhabitants Over Time by Gender. Lines 193-202 in R-file.

Figure 2 shows the number of imprisonments per 100,000 inhabitants in Denmark, separated by gender across the selected years. Males dominate the chart; it shows that from year 1925 the male imprisonment rate increases significantly and peaks in the year 2000. The male chart then falls by nearly 75 in 2024 but remains higher than earlier years. Looking at the female imprisonment rate, it's clear that it's very stable throughout the years with a small increase from 1975 to 2000. The data is grouped by gender to show if there are any patterns of imbalance between the genders. Had we just shown the table, it would have been hard to notice the visible differences between the sexes, which the charts show⁵. We have also created an animation of this graph, which has been linked to GitHub.

⁵ James Long and Paul Teetor, *R Cookbook*, 2nd ed. (Sebastopol, CA: O'Reilly Media, 2021), chap. 6.6, 10.2, and 10.5.



Figure 2: *Convicted Persons per 100,000 Inhabitants Over Time*. Lines 56-77 in R-file.

We made a chart showing the rise and fall of the convictions of 100,000 inhabitants. Figure 3 shows the convicted persons per 100,000 inhabitants over time. It shows a massive rise from the year 1900-2000. When we made the chart, we bared in mind that it was important for us to get all our data into one timeseries so that it was possible to make the comparisons between the years, when looking at the chart. This helped us to visualize the patterns throughout the 20th and 21st centuries.⁶

⁶ James Long and Paul Teetor, *R Cookbook*, 2nd ed. (Sebastopol, CA: O'Reilly Media, 2021), chap. 5.20 and 10.13.

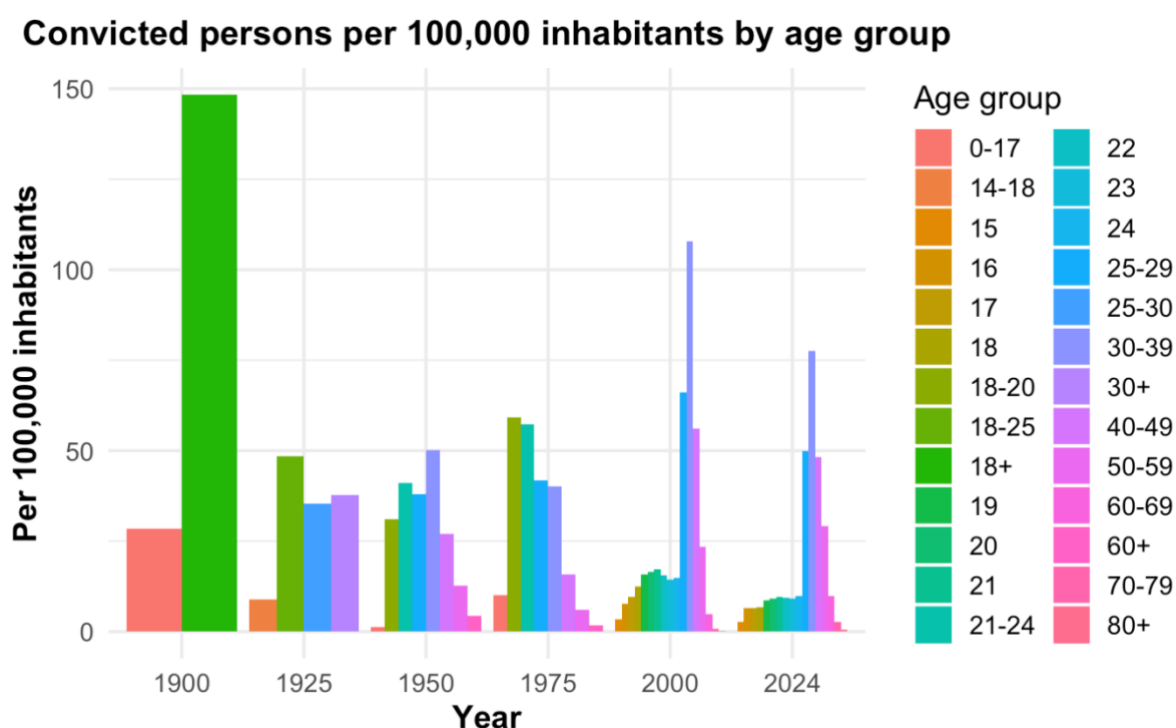


Figure 4: Convicted Persons per 100,000 Inhabitants by Age Group. Lines 112-146.

Figure 4 presents conviction rates in multiple age groups. Overall, adults were convicted more frequently than people under 18 in all years. For this visualization, the data needed to be consistently formatted in order to compare different age groupings next to each other. We preserved just six years to maintain a clear comparison⁷. The investigation is based on quantitative data from *Årbøger for Kriminalstatistik*, (as are the other graphs and plots) we focused on the conviction rates per 100,000 inhabitants based on age in Denmark. In the year 1900, the 18+ chart is very high, this is due to the data not being available to show ages from above years of plus 18. The convicted was only written as two types of ages which is 0-17 and 18+, so looking at 18+, it could be any age from 18 and above. For a reference, it is important to include *Figure 1*, because that figure shows that the year 1900 is the second lowest year of convictions per 100,000 inhabitants. Therefor *Figure 1* helps guide the data to be understood by visualizing clearer and better.

⁷ James Long and Paul Teetor, *R Cookbook*, 2nd ed. (Sebastopol, CA: O'Reilly Media, 2021), chap. 10.10.

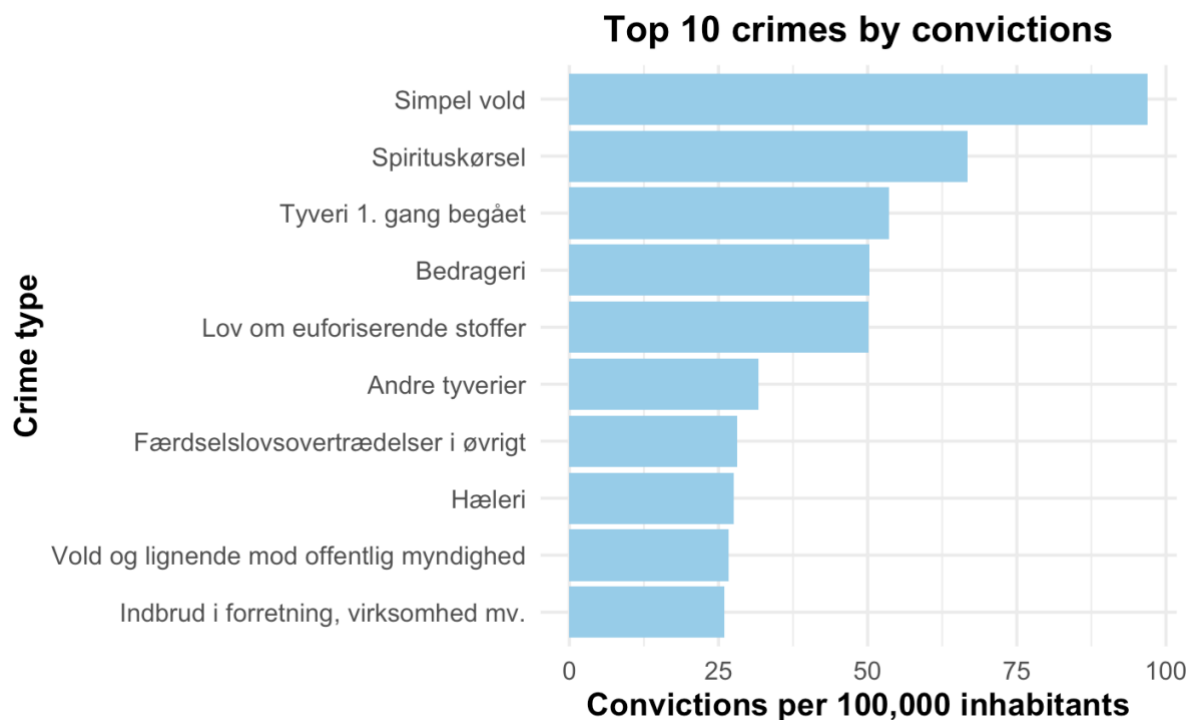


Figure 5: Top 10 Crimes by Convictions. Lin 84-108.

Figure 5 shows the 10 most occurred offense types (in total convictions per 100,000 inhabitants). Some of the top crimes are, as one would expect, theft and violence. We made this plot to expose which of the categories have been the most common throughout the years. We filtered the top 10 (since we don't want to plot all of them at once) to better the visualization of the chart. This is a popular mechanism to ease information overload.⁸ This graph is also a testament to how difficult it can be to render data from multiple years, as we noticed some of the crimes from the start of the 20th century are either decriminalized or called something else. This mean that we must use this graph more as a general idea of the time back then and today, instead of a direct source to how it was/is.

⁸James Long and Paul Teetor, *R Cookbook*, 2nd ed. (Sebastopol, CA: O'Reilly Media, 2021), chap. 6.6 and 10.10.

4.0 Discussion/Critical Evaluation

4.1 Significance of findings

As mentioned, our study reveals a steady increase in the number of inmates over time, along with a significant disparity between male and female prisoners. Starting with the development in inmate numbers, we observe that the figure (Figure 1) was quite low around the year 1900. This can be attributed to the reform of the Penal Code in 1866, which reduced the sentencing range for several offenses. Consequently, the number of prisoners declined, and various reform initiatives were introduced leading up to the turn of the century.⁹ These factors help explain the low prison population in Denmark around 1900. However, by 1925, the number of inmates had increased. This rise is consistent with the legal philosophy embedded in the 1866 Penal Code, which emphasized that a criminal act should be punished as just retribution. A significant shift occurred with the enactment of the 1930 Penal Code, which introduced a new approach to criminal punishment. As previously noted, the focus moved away from pure retribution and toward the resocialization of offenders. This shift is evident in Section 10 of the code, which deals with the treatment of children and young offenders.¹⁰ In the following decades, further reforms were implemented. Several offenses were decriminalized, as seen in the legalization of pornography and the reduction of penalties for certain crimes.¹¹ From 1980 onward, however, sentencing policies became increasingly strict. This trend is clearly visible in the period between 1980 to 2000, during which 43 amendment acts were passed, resulting in 369 legislative changes. In the subsequent period from 2001 to 2022, this development continued with 134 amendment acts leading to a total of 622 legislative changes.¹² This tightening of sentencing is clearly reflected in our graph, which shows a marked increase in the number of inmates from 1975 to 2000. The rise corresponds with a more punitive justice system and the implementation of harsher penalties. Interestingly, from 2000 to 2024, we observe a decline in

⁹ "Danmarks Arresthuse." *Historie-online.dk*. Accessed May 21, 2025. <https://www.historie-online.dk/boger/danmarks-arresthuse>.

¹⁰ Justitsministeriet. *Lov om almindelig borgerlig Straffelov (Straffeloven), 1930*. *Retsinformation.dk*. Accessed May 21, 2025. <https://www.retsinformation.dk/eli/lta/1930/127#P10>.

¹¹ Gyldendal og Politikens Danmarkshistorie. "Forbrydelse og straf." *Lex.dk*. Accessed May 22, 2025. https://gyldendalogpolitikensdanmarkshistorie.lex.dk/Forbrydelse_og_straf.

¹² Advokatsamfundet. "Man kan jo ikke blive ved med at kriminalisere og skærpe straffene." *Advokatsamfundet.dk*. April 2, 2024. Accessed May 20, 2025. <https://www.advokatsamfundet.dk/nyheder-medier/nyheder/2024/man-kan-jo-ikke-blive-ved-med-at-kriminalisere-og-skaerpe-straffene/>.

inmate numbers. It is therefore worth questioning whether crime has simply become less ‘worth committing’ due to changes in societal conditions and lower potential gains.

As shown in Figure 2, men significantly outnumber women in Danish prisons. But what are the reasons behind this disparity? According to Torben Bechmann Jensen, Head of Studies at the Department of Psychology at the University of Copenhagen, several factors contribute to this imbalance. He argues that men are, first and foremost, more likely to commit crimes simply because they have more opportunity to do so. As he puts it: *“Kvinder kan også finde på at ende i et slagsmål, men gruppen af personer, som er svagere end den kvinde, som udøver volden, er mindre end gruppen af personer, der er mindre end den mand, som udøver volden. Da vi af natur helst vil slå med nogle, der er svagere end os selv, begår kvinder mindre voldelig kriminalitet end mænd”*. Another factor Jensen emphasizes is the biological differences between men and women. Men’s greater physical strength often makes their acts of violence more severe, which increases the likelihood of being reported and punished. In short, physical differences between the sexes play a significant role in crime statistics. Beyond biology, social and cultural factors are also central. Both Jensen and Ann-Karina Eske Henriksen, a postdoctoral researcher at the Department of Sociology and Social Work at Aalborg University, point to gendered upbringing as a key contributor. Traditional gender roles often raise boys to be providers and risk-takers, while girls are encouraged to care for others and stay within the domestic sphere. These social expectations influence behavior from a young age. Criminal behavior among boys often begins in adolescence, with petty crimes like vandalism or theft, frequently committed while spending unsupervised time on the streets. Girls, by contrast, are more often expected to stay at home and are subject to stricter social control, reducing their exposure to criminal environments. When looking at general behavioral patterns, male behavior tends to align more closely with traits associated with criminality. Masculine characteristics such as physical strength, courage, and risk-taking are often socially reinforced – and these traits also underpin many types of criminal activity. In this sense, one can argue that masculinity is more compatible with criminality than femininity. Finally, there are notable differences in how men and women respond to economic hardship, such as in the context of substance abuse. Men are more likely to commit crimes such as theft, robbery, or assault to obtain money, whereas women may turn to prostitution – a form of survival that, while

problematic, is less criminalized than violent offenses. This further helps explain why men are overrepresented in prison statistics.¹³

When looking at conviction rates, our data reveals a noticeable increase in the incarceration of men, particularly between 1975 to 2000. In contrast, the number of women has remained stable during the same period. This indicates that changes in penal practices over time between the 20th and 21st centuries have had a stronger impact on men. These findings align with our research that gender plays a key role in how punishment is applied. Both gender and age figures show that young men are overrepresented among the convicted individuals. A clear gender disparity continues to stand out in the data. According to Lennart Frandsen, only about 5% of prisoners are women – roughly 170 at any given time.¹⁴ This imbalance is further illustrated in Figure 2, which shows that male imprisonment rates are far higher than the female rate.

Age also plays a crucial role in our analysis, as it deepens out the understanding of which age group is the most occurring. Figure 4 shows conviction rate per 100,000 by age group. It shows that in earlier decades the younger individuals make up for a big portion of the convicted. Since 1975, however there has been a shift, with the 30-39 age group becoming the dominant one. This could reflect improved registration policies and better classification of age groups in criminal statistics. Looking back from 1900 to 1925, we see a decline in the "18+" category, likely because age tracking became more detailed. The 18–25 age group shows the highest conviction rates, with 25–39 in second place. Those under 18 consistently make up the smallest group. By 1950 and 1975, we begin to see more detailed age categories like 14–18, 18–25, and 25–30, which points to a shift in how age is recorded. In 2000, there's a sharp increase in convictions among the 30–39 age group, making it the highest year group on record.

For a better understanding of social factors, class and age, a background analysis conducted for the book "Det danske klassesamfund – et socialt Danmarksportræt" examined convictions among 18 to 59-year-olds. It showed that youth crime is most common among those from lower-class families and least common among upper- and upper-middle-class families. In 2007, 3.5% of 15–20-year-olds from the lower class were convicted, compared to just 0.7% from the

¹³ Videnskab.dk, "Hvorfor er mænd mere kriminelle end kvinder?" *Videnskab.dk*, accessed May 17, 2025, <https://videnskab.dk/kultur-samfund/hvorfor-er-maend-mere-kriminelle-end-kvinder/>.

¹⁴ Folketingets Ombudsmand, "Kvinder i fængsel," *Beretning fra Ombudsmanden 2011*, accessed May 20, 2025, <https://www.ombudsmanden.dk/find-viden/beretninger-og-andre-publikationer/artikler-offentliggjort-i-ombudsmandens-beretning/artikler-i-fob-2011/kvinder-i-faengsel>.

upper class. Working-class youth had a conviction rate of 1.6%, slightly above the national average of 1.5%. Among 24–27-year-olds, the trend continues: 2.1% from the lower class were convicted, compared to 0.5% from the upper class. This clearly shows a link between social class and criminal convictions. This confirms that socioeconomic status is a major factor influencing the risk of encountering the criminal justice system.¹⁵

This trend is backed up by Gunvor Christensen, who holds a PhD in Sociology¹⁶, who points out that one in five ethnic Danish children living in disadvantaged areas has a father who has received a sentence at some point. Christensen also emphasizes that families in these areas are often more socially vulnerable. Importantly, this vulnerability isn't limited to ethnic minorities—many ethnic Danish families are also affected, particularly those with lower education levels or outside the workforce. This highlights the influence of social conditions like income, education, and housing on criminal behaviour.¹⁷

Here it's important to include the "Strain Theory", developed by sociologist Robert K. Merton. The theory explains how societal structures can pressure individuals to commit crimes. Merton argued that society promotes cultural goals such as wealth and success but does not provide equal access to the legitimate means for achieving them. This imbalance creates strain, which may lead some to pursue deviant paths. Merton identified five responses to strain: conformity, innovation, ritualism, retreatism, and rebellion. For example, someone may accept the goal of financial success but resort to theft (innovation) when legal means are blocked. The theory highlights how inequality can foster criminal behaviour within disadvantaged groups.¹⁸

Tying it together with the crimes that are committed, we look at Figure 5, as it shows the ten most frequently committed crimes in Denmark. Looking at the top of the list, it shows "Simpel vold", (Simple assault) with close to 100 per 100,000 inhabitants. This suggests that interpersonal violence remains the most consistent and prominent cause of penal action in

¹⁵"Kriminalitet i de sociale klasser", Jonas Schytz Juul, Samira Nawa og Andreas Mølgaard, 29. oktober 2012. Side 1-6.

¹⁶"Gunvor Christensen," *Bystrategisk Udsyn*, accessed May 18, 2025, <https://bystrategiskudsyn.dk/presenters/gunvor-christensen/?returnUrl=/>.

¹⁷VIVE – The Danish Center for Social Science Research. "Udsat: Vi overser, hvor hårdt uligheden rammer etnisk danske børnefamilier i udsatte boligområder." *Politiken*, February 23, 2019. Accessed May 17, 2025. <https://www.vive.dk/da/nyheder-og-debat/udsat-vi-overser-hvor-haardt-uligheden-rammer-etnisk-danske-boernefamilier-i-udsatte-boligomraader-mxj1bkx4/>.

¹⁸Josephine Campbell. "Strain Theory (Sociology)." 2024. Accessed May 23, 2025. <https://www.ebsco.com/research-starters/sociology/strain-theory-sociology>.

Denmark. The provision is typically applied in bar fights, street violence, which involves acts of punches, kicks to the body, throwing with objects and/or headbutts. The injuries are usually of limited severity.¹⁹ Danmarks Statistik also confirms that “simpel vold” (simple assault) is the most common type of conviction in Denmark, with over 30,000 cases of violence reported in 2022. These are offenses where men dominate both as offenders and as prisoners.²⁰ “Spirituskørsel” (Drunk driving) is the second most frequent offense with little over 60 convictions per capita. Drunk driving became a criminal offense in Denmark in 1976 and therefor is among the newest.²¹ Rådet For Sikker Trafik claims that 9 out of 10 of convicted are men.²² “Tyveri 1. gang begået” (First time theft) comes in a close third. These categories reflect more behavioural or situational offenses, which we believe are crimes which are often tied to routine checks or socio-economic factors, such as addiction, poverty, or opportunistic actions.

Rasmus Munksgaard, an Assistant Professor at the Department of Social Work at Aalborg University, argues that theft often stems from necessity. People steal to secure their next meal. He further suggests that during periods of high inflation, the demand for stolen goods increases, as they are perceived to be cheaper than items sold in stores.²³

Further down the list are crimes like *andre tyverier* (other types of theft), *færdelslovsovertrædelser i øvrigt* (other traffic offenses), *hæleri* (handling stolen goods), *vold mod offentlig myndighed* (violence against public authorities), and *indbrud i forretning mv.* (business burglary). These are more routine crimes with generally lower conviction rates.

As mentioned earlier, crimes are being punished harsher and harsher. A research group from Aalborg University studied trends in Danish criminal law and found a significant rise in legal activity: faster case handling, tougher sentences, and new categories of offenses. Professor of Criminal Law Birgit Feldtmann notes: “*Det er helt tydeligt, at der efter årtusindskiftet ses en*

¹⁹ Stage Advokatfirma. “Hvad er straffen for vold?” *Stage.dk*. Accessed May 23, 2025.
<https://www.stage.dk/vaerd-at-videre/hvad-er-straffen-for-vold/>.

²⁰ Danmarks Statistik. “Flere voldsdomme og færre bøder.” *Danmarks Statistik*. Accessed May 23, 2025.
<https://www.dst.dk/da/Statistik/nyheder-analyser-publ/nyt/NytHtml?cid=40254>.

²¹ Rådet for Sikker Trafik. “Promillegrænser.” *Sikkertrafik.dk*. Accessed May 23, 2025.
<https://sikkertrafik.dk/rad-og-viden/bil/spirituskørsel/promillegrænser/>.

²² Rådet for Sikker Trafik, “Hvem kører spritkørsel?” *Sikkertrafik.dk*, accessed May 20, 2025,
<https://sikkertrafik.dk/rad-og-viden/bil/spirituskørsel/hvem-korer-spritkørsel/>.

²³ DR, “Antallet af tyverier stiger: Folk stjæler som aldrig før,” *DR.dk*, accessed May 20, 2025,
<https://www.dr.dk/nyheder/regionale/syd/antallet-af-tyverier-stiger-folk-stjaeler-som-aldrig-foer>.

*tendens til, at flere handlinger straffes, flere personer straffes, og straffen bliver hårdere. Denne udvikling tog særlig fart i 2010'erne, og det ser umiddelbart ikke ud til, at dette ændrer sig foreløbigt, når vi kigger på vores resultater.”*²⁴ This also came to light, as we were writing the assignment. The Danish government announced on May 22nd, 2025, that they would introduce a new penal reform aimed at increasing penalties for aggravated violence and rape, while also expanding the country's prison capacity. According to Minister of Justice Peter Hummelgaard, the reform is intended to ensure that serious violent crimes are met with tougher and more tangible consequences. Among the core initiatives are the doubling of sentences for aggravated violence, as well as a 50 percent increase in penalties for aggravated rape and for violence against a partner or child. At the same time, the reform proposes a major expansion of prison capacity. By 2036, more than 2,000 new prison spaces are to be created, including a new high-security prison and a new women's prison. The government thus aims to send a clear message that severe offenses in Denmark will result in stricter punishment.²⁵

4.2 Evaluation

Our investigation succeeded in meeting its primary goal, which was to analyze how the use of imprisonment in Denmark has evolved over time with focus on gender, age, and offense type. Our assumptions were that the legislation would become stricter throughout the course of the years 1900 to 2024. We assumed that the male offender rate would be the highest among the genders, which was proven by the data. We did not expect the female conviction rate to be high, which the data showed. We were surprised by the female incarceration rate, because it was very stable throughout the 124 years.

The age assumption of the convicted did surprise us a bit. We anticipated that the age group 25-30 would be the highest conviction group. The data showed that the age group committing the most crime was the age group 30-39 years (*see Figure 4*).

²⁴ Dreyers Fond, “Mere Straf,” *Dreyersfond.dk*, accessed May 17, 2025, <https://dreyersfond.dk/okay-portfolio/mere-straf/>.

²⁵ DR, “Regeringen vil fordoble straf for grov vold og øge fængselspladser,” *DR.dk*, accessed May 22, 2025, <https://www.dr.dk/nyheder/politik/regeringen-vil-fordoble-straf-grov-vold-og-oege-faengselspladser>.

Regarding the most committed crimes (*see Figure 5*), we assumed that simple violence (simpel vold) and theft (tyveri) would be among the highest crimes by conviction. We were proven right, however we were surprised to see that drunk driving (spirituskørsel) would be as high as number two, since it just became illegal in 1976.

Overall, our results align with historical sources showing a general increase in legislation in Denmark in the late 20th to early 21st century. The data taken from Danmarks Statistik was reliable but had gaps in the year 1900. By looking at the annual report of conviction rates, we were able to determine whether males or females were most represented.

4.3 Consider Utility/Representativeness

When it comes to our analysis of the number of people in prison, the data from the *Årbøger for Kriminalstatistik* is highly comprehensive, as this type of data is particularly easy to measure and import, since it typically consists of a single numerical value assigned to each year (though divided by gender). However, it becomes somewhat more difficult to draw clear conclusions from our conviction data, as this is where the data collection from the *Årbøger for Kriminalstatistik* becomes more blurred. From 1900 to 2024, numerous new laws have been introduced, while others have been repealed or decriminalised, resulting in different categorizations of convictions each year. This makes it difficult to fully conclude from our Top 10 table, for example. Still, the data provides insight into general societal trends across the years. For instance, that violence consistently accounts for a large share of convictions, and that new laws significantly influence penal policy (including drunk driving, *see Figure 5*).

As Helle Strandgaard Jensen points out in her article *Digital Archival Literacy for (all) Historians*, digital archives often prioritise popular and easily accessible material, which can distort how historical sources are represented over time.²⁶ This presents a practical challenge when working with sources like the *Årbøger for Kriminalstatistik*, where categorical

²⁶ Helle Strandgaard Jensen (2020): Digital Archival Literacy for (all) Historians, Media History, DOI: 10.1080/13688804.2020.1779047, page 4, line 27-29. (This is paraphrased and not a direct quote).

definitions or institutional registration practices may have changed over the course of the period studied.

4.4 Lessons learnt

Working with historical data on convictions and imprisonments from 1900 to 2024 gave us some surprising results. One of the biggest surprises was how common the 30–39 age group has become in the 21st century in crime statistics. Most people would probably expect the 25–30 age group to be the most common, but Figure 4 clearly shows a shift. This made us think about how age groups have changed over time, and how that makes it harder to compare old and new data. Another surprising finding was the high number of convictions for drunk driving, which is shown as the second most common type of crime. We had expected violence and theft to be the most common, but the high number of drunk driving cases shows that new laws and more police control might have had a bigger effect than we thought. Our method used data every 25 years. This helped us see big trends over time, but it also had some problems. The big gaps between years probably made us miss smaller changes or short-term effects from politics or social changes. Using data every 5 or 10 years could give a more detailed picture, especially in active periods like the 1980s and 2010s. We also saw how difficult it is to work with old historical data like the *Årbøger for Kriminalstatistik*. The way crimes are counted and grouped has changed a lot over time. This means we couldn't always compare crimes directly from one year to another. For example, when a law changes or a crime becomes legal, it becomes harder to determine whether a decline in convictions is due to changed behaviour in society or if it's just new definitions. As Helle Strandgaard Jensen says in her article *Digital Archival Literacy for (all) Historians*, digital archives often focus on material that is popular or easy to find. This can affect what kind of patterns we see in the data.²⁷ We noticed this in our work with the yearbooks. In the future, we think researchers could improve their studies by:

- More frequent data sampling (e.g., every 5 or 10 years),
- Inclusion of data on regions or social classes (if possible),

²⁷ Helle Strandgaard Jensen (2020): Digital Archival Literacy for (all) Historians, Media History, DOI: 10.1080/13688804.2020.1779047, page 4, line 27-29. (This is paraphrased and not a direct quote).

- Qualitative context from legal texts or media coverage,
- Comparative analysis with other countries to examine whether Denmark's development is unique or part of a broader trend.

The project has given us a much deeper understanding of the challenges involved in working with historical data in digital form. Not only in terms of access and data cleaning, but also in terms of critically analysing what the data represent and what may potentially be missing.

5.0 Conclusions and recommendations

Based on this project, we can conclude that the use of imprisonment in Denmark has changed a lot over the last 124 years. This has happened because of changes in criminal policy, political movements, and social developments. In the early 20th century, the focus was on giving harsh punishments. Later, in the middle of the century, the focus shifted more towards rehabilitation. But from the 1980s until today, there has been a move back towards a stricter way of thinking about prisons and punishment.

One of our most important findings is the dramatic increase in incarceration rates from 1900 to 2000, which later has a decline towards 2024. This development is closely linked to increased law-making activity, particularly in the 2010s, and rising political support for tougher crime policies. Our data also shows strong gender differences. Men have been more affected by imprisonment than women. Likewise, the 30–39 age group has emerged as the most convicted in recent decades. This is a shift that challenges our first assumptions and highlights the need to examine changing age classifications over time. Offense types such as simple violence, drunk driving, and theft dominate conviction statistics. This statistic points to the strong influence that legal changes and enforcement strategies have on crime, as drunk driving is the second most committed crime. Working with long term digital data has really shown the importance of critical source evaluation. Changes in data categories, legislation, and reporting standards make direct comparisons across decades really difficult. Still, this has taught us how digital archives both reveal and somehow change historical patterns. In future research, we recommend smaller intervals in years (perhaps every 5–10 years), inclusion of regional and class-based data, and studies that show crime statistics in other countries. These steps would help making whether the Danish development is unique or part of broader international trends much more clear.

Prior to conducting our analysis, we hypothesized that the use of imprisonment in Denmark would reflect a historical shift from punitive approaches toward a stronger focus on rehabilitation, and that both gender and age would significantly influence sentencing patterns. Our findings support this hypothesis: we observed clear transitions in penal policy over time, and a pronounced disparity between men and women, consistent with our expectations.

Author contributions

Kristoffer Segerstrøm, Lukas Benner, and Emil Hansen worked together on the overall idea and structure of the project. We also shared the work of researching background information and reviewing relevant literature. All three of us took part in collecting, cleaning, and analysing the data using R. Everyone also helped make the graphs and explore the dataset. Section 3.0 *Findings* was mainly written by Kristoffer Segerstrøm, with help from Lukas Benner and Emil Hansen. Sections 4.0 *Discussion/Critical Evaluation*, 4.1, and 4.2 were written by Lukas Benner and Emil Hansen, with help from Kristoffer Segerstrøm. Sections 4.3 and 4.4 were written by Kristoffer Segerstrøm, with support from Lukas Benner and Emil Hansen. The conclusion was written together by all three of us.

All authors have agreed upon this contribution declaration.

In this project, we used ChatGPT to translate our text from Danish to English. We also consulted ChatGPT in relation to our coding, particularly in connection with the animation.

References

Advokatsamfundet. "Man kan jo ikke blive ved med at kriminalisere og skærpe straffene."

2024. <https://www.advokatsamfundet.dk/nyheder-medier/nyheder/2024/man-kan-jo-ikke-blive-ved-med-at-kriminalisere-og-skaerpe-straeffene/>.

Danmarks Radio (DR). "Antallet af tyverier stiger: Folk stjæler som aldrig før."

<https://www.dr.dk/nyheder/regionale/syd/antallet-af-tyverier-stiger-folk-stjaeler-som-aldrig-foer>.

Danmarks Radio (DR). "Regeringen vil fordoble straf for grov vold og øge fængselspladser."

<https://www.dr.dk/nyheder/politik/regeringen-vil-fordoble-straf-grov-vold-og-oegel-faengselspladser>.

Danmarks Statistik. "Flere voldsdomme og færre bøder." Nyt fra Danmarks Statistik.

<https://www.dst.dk/da/Statistik/nyheder-analyser-publ/nyt/NytHtml?cid=40254>.

Danmarks Statistik. "STRAF40: Strafferetlige afgørelser efter lovovertrædelsens art."

<https://www.statistikbanken.dk/STRAF40>.

Dreyers Fond. "Mere straf?" <https://dreyersfond.dk/okay-portfolio/mere-straf/>.

Folketingets Ombudsmand. "Kvinder i fængsel." <https://www.ombudsmanden.dk/find-viden/beretninger-og-andre-publikationer/artikler-offentliggjort-i-ombudsmandens-beretning/artikler-i-fob-2011/kvinder-i-faengsel>.

Gyldendal og Politikens Danmarkshistorie. "Forbrydelse og straf." Lex.dk.

https://gyldendalogpolitikensdanmarkshistorie.lex.dk/Forbrydelse_og_straf.

Jensen, Helle Strandgaard. "Digital Archival Literacy for (All) Historians.", Media History,

2020. <https://doi.org/10.1080/13688804.2020.1779047>.

Justitsministeriet. *Straffeloven af 1930*. Retsinformation.

<https://www.retsinformation.dk/eli/lt/1930/127#P10>.

Juul, Jonas Schytz, Samira Nawa, and Andreas Mølgaard. "Kriminalitet i de sociale klasser."

October 29, 2012. https://www.ae.dk/files/dokumenter/analyse/ae_kriminalitet-i-de-sociale-klasser.pdf

Long, James, and Paul Teetor, "R Cookbook". 2nd ed. Sebastopol, CA: O'Reilly Media, 2021.

Sikkertrafik.dk. "Hvem kører spritkørsel?" <https://sikkertrafik.dk/rad-og-viden/bil/spirituskørsel/hvem-korer-spritkørsel/>.

Stage Advokatfirma. "Hvad er straffen for vold?" <https://www.stage.dk/vaerd-at-vide/hvad-er-straffen-for-vold/>.

VIVE. "Udsat: Vi overser hvor hårdt uligheden rammer etnisk danske børnefamilier i udsatte boligområder." Politiken, February 23, 2019. <https://www.vive.dk/da/nyheder-og-debat/udsat-vi-overser-hvor-haardt-uligheden-rammer-etnisk-danske-boernefamilier-i-udsatte-boligomraader-mxj1bkx4>.

Videnskab.dk. "Hvorfor er mænd mere kriminelle end kvinder?" <https://videnskab.dk/kultur-samfund/hvorfor-er-maend-mere-kriminelle-end-kvinder/>.

"Danmarks Arresthuse." Historie-online.dk. <https://www.historie-online.dk/boger/danmarks-arresthuse>.

Bystrategisk Udsyn – Landbyggefonden. <https://bystrategiskudsyn.dk/presenters/gunvor-christensen/?returnUrl=/>

Josephine Campbell. "Strain Theory (Sociology)." 2024. <https://www.ebsco.com/research-starters/sociology/strain-theory-sociology>.

Rådet for Sikker Trafik. "Promillegrænser." *Sikkertrafik.dk*. <https://sikkertrafik.dk/rad-og-viden/bil/spirituskørsel/promillegraenser/>.

Table 1 – Software metadata


<i>Software metadata description</i>	<i>Please fill in this column</i>
<i>Permanent link to Github repository where you put your script(s), R project, and data</i>	https://github.com/kris296m/Final-Project-.git
<i>Software License</i>	GNU GENERAL PUBLIC LICENSE, Version 3 (29 June 2007)
<i>Data License</i>	<i>Data from Danmarks Statistik, available under public domain</i>
<i>Software versions, Installation requirements & dependencies for software not used in class</i>	R version 4.4.2 (2024-10-31), ImageMagick version ≥ 7.0 , (gganimate)(av).
<i>If available Link to software documentation for special software (only relevant if you go outside the scope of class)</i>	<p>https://imagemagick.org/script/download.php and https://gganimate.com/ to make animation.</p> <p><i>Animation Figure 2 -</i></p>  <p><i>The animation is uploaded in GitHub and can be seen there.</i></p> <p>https://github.com/kris296m/Final-Project-.git</p>
<i>Support email for questions</i>	Au766002@uni.au.dk

Table 2 – Data metadata

<i>Metadata description</i>	<i>Please fill in this column</i>
1900.xlsx	Conviction statistics in Denmark for the year 1900. Source: Statistics Denmark. Extracted in May 2025. Columns include: Year, Gender, Crime type, Count, and Convictions per 100,000 inhabitants.
1925.xlsx	Conviction statistics in Denmark for the year 1925. Source: Statistics Denmark. Extracted in May 2025. Same structure as D1.
1950.xlsx	Conviction statistics in Denmark for the year 1950. Source: Statistics Denmark. Extracted in May 2025. Same structure as D1.
1975.xlsx	Conviction statistics in Denmark for the year 1975. Source: Statistics Denmark. Extracted in May 2025. Same structure as D1.
2000.xlsx	Conviction statistics in Denmark for the year 2000. Source: Statistics Denmark. Extracted in May 2025. Additional column(s) may be present.
2024.xlsx	Conviction statistics in Denmark for the year 2024. Source: Statistics Denmark. Extracted in May 2025. Additional column(s) may be present.
Domfældte alder.xlsx	Conviction statistics by age group in Denmark. Source: Statistics Denmark. Extracted in May 2025. Columns include: Year, Age group, Count, and Convictions per 100,000 inhabitants.
Fængslinger i alt.xlsx	Incarceration statistics in Denmark by year and gender. Source: Statistics Denmark. Extracted in May 2025. Columns: Year, Gender, Count, and Convictions per 100,000 inhabitants.

6.0 Portfolio

On the next page, we have listed our portfolio. The HTML-files and Rmd-files can be seen in the GitHub repository. The link is in *Table 1* and here - <https://github.com/kris296m/Final-Project-.git>

Week 8 assignment

1. What regular expressions do you use to extract all the dates in this blurb: <http://bit.ly/regexexercise2> and to put them into the following format YYYY-MM-DD ?

When we want to convert dates to the same form, we first need to mark all the dates.

We can use these commands to mark the dates in the text:

```
\d+.\d+.\d+
```

or

```
\d{1,2}.\d{1,2}.\s?\d+
```

When we have marked the dates, we put a bracket around every group:

```
(\d{1,2}).(\d{1,2}).\s?(\d+)
```

Then go to function in the right colon and find substitution.

In the textbox that opens with substitution write:

```
$2-$1-$3
```

This follows the order in which the date is shown. Here the date is formed as number 2, 1 is the month and 3 represents the year.. This can be written in any wanted order.

The task in regex.

<https://regex101.com/r/83vEpt/1>

2. Write a regular expression to convert the stopwordlist (list of most frequent Danish words) from Voyant in <http://bit.ly/regexexercise3> into a neat stopword list for R (which comprises "words" separated by commas, such as <http://bit.ly/regexexercise4>). Then take the stopwordlist from R <http://bit.ly/regexexercise4> and convert it into a Voyant list (words on separate line without interpunction)

Stoplist from Voyant to R

Write ([a-z0-9æøå]+) in the textbox followed by \n:

Gruppe: Kristoffer Segerstrøm, Emil Gert Hansen, Lukas Benner

`([a-z0-9æøå]+)\n`

The `\n` makes the text go from colon to text.

Go to function and substitution. Write the following

`"$1"`

Then you have a stoplist for R

The task in Regex:

<https://regex101.com/r/WQtOTA/1>

Stoplist from R to Voyant

Write the following in the textbox:

`\("[a-z0-9æøåüé.]+)(.,)`

`\"` and `(.,)` deletes the characters.

Go to function and substitution. Write the following

`$1\n`

The `\n` makes the text go from text to colon.

Then you have a stoplist for Voyant.

The task in Regex:

<https://regex101.com/r/zTCekK/1>

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

No, OpenRefine does not alter the raw data when you sort or filter it. Sorting and filtering in OpenRefine are non-destructive operations that only change the way data is displayed but do not modify the underlying dataset.

4. Fix the [interviews dataset](#) in OpenRefine enough to answer this question: "Which two months are reported as the most water-deprived/dryest by the interviewed farmer households?"

The two most water-deprived months are October and September.

This conclusion is made by using the command in OpenRefine, where we choose the colon "months_lack_water".

We choose "facet" and then "transformation"

Custom text transform on column months_no_water

Expression Language General Refine Expression Language (GREL) ▾

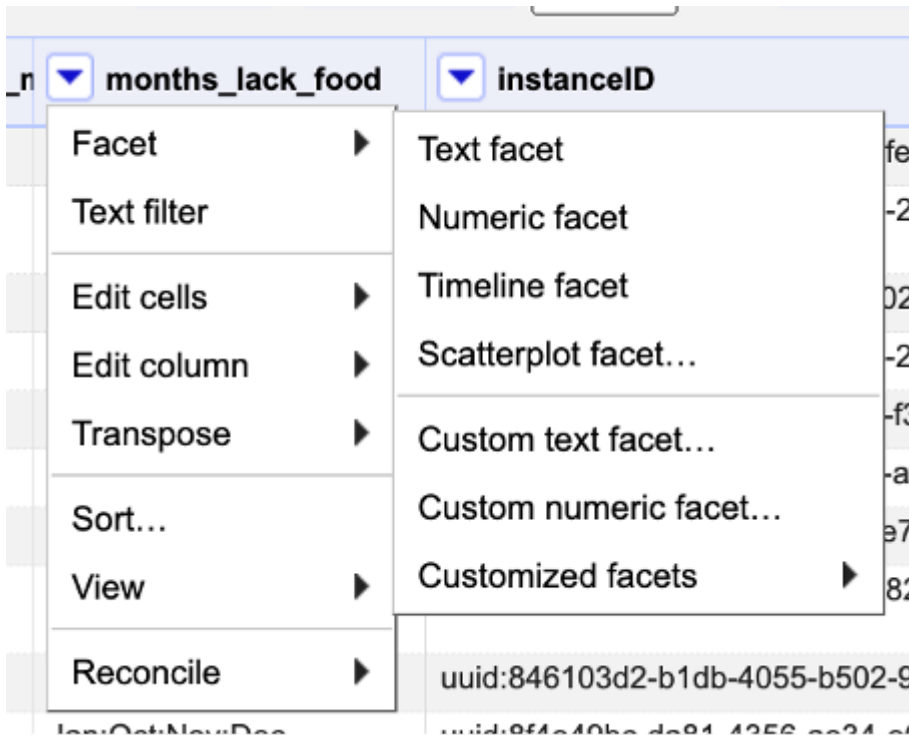
```
value.replace("[",  
").replace("]", "").replace("","").replace("  
", "")
```

No syntax error.

Preview History Starred Help

row	value	value.replace("[", "").replace ...
1.	NULL	NULL
2.	Aug;Sept	Aug;Sept
3.	NULL	NULL
4.	NULL	NULL
5.	NULL	NULL
6.	NULL	NULL

Then we choose "facet" and then "custom text facet"



In “Custom text facet” we write “value.split(“;”)

Custom facet on column months_no_water

Expression Language General Refine Expression Language (GREL) ▾

value.split(";") No syntax error.

Preview [History](#) [Starred](#) [Help](#)

row	value	value.split(";")
1.	NULL	["NULL"]
2.	Aug;Sept	["Aug", "Sept"]
3.	NULL	["NULL"]
4.	NULL	["NULL"]
5.	NULL	["NULL"]
6.	NULL	["NULL"]

We now get a facet, where we can see a list of the months ranged and how many that experienced lack of water these months.



months_no_water		change
11 choices Sort by: name count		
Apr	1	
Aug	33	
Dec	11	
Jan	2	
July	2	
June	1	
May	1	include
Nov	51	
NULL	45	
Oct	74	
Sept	70	
Facet by choice counts		

From the list we can see that in October there were 74 that lacked food and 70 in September.

- 5. Real-Data Challenge: What are the 10 most frequent occupations "erhverv" among unmarried men or women of 20-30 years in [1801 Aarhus](#) census dataset? (hint: first select either men or women to shrink the dataset to a manageable size, then filter by age, and then use merging to cut the erhverv variation ruthlessly.)**

We chose to focus on the women.

Women:

First, we sort by gender:

Facet → text facet

Then we sort by age:

Edit cells → Common transforms → To number

Facet → Numeric facet

Then age can then be adjusted to 20-30

To separate the married and unmarried from each other we create a text filter under “civilstand” and write “ugift” in the facet, which appears.

civilstand

invertreset

ugift

☐ case sensitive☐ regular expression

Then we make a facet of “erhverv” and cluster the words that can be clustered.

First with the methods of “Key collision” and keying function “Metaphone3” where we re-merge and cluster

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 collisionKeying function: Metaphone3Manage clustering functions

☐ Auto-update31 clusters found

Merge?	Values in cluster	New cell value	Cluster size	Row count
<input type="checkbox"/>	<div>inderste og væverske (5 rows)</div> <div>inderste og væverpige (2 rows)</div> <div>inderste og fattige</div> <div>inderste og vanfer</div> <div>inderste og væver</div>	inderste og væverske	5	10
<input type="checkbox"/>	<div>vanvlig og nyder almisse</div> <div>vanvlig og nyder almisse</div> <div>vanvlig og nyder almissekom af sognet</div> <div>vanvlig og nyder hospitalshold</div>	vanvlig og nyder almisse	4	4
<input type="checkbox"/>	<div>inderste og nyder almisse af sognet (3 rows)</div> <div>inderste og nærer sig ved haandgjerning (2 rows)</div> <div>inderst og nærer sig af haand arbejde</div> <div>inderste og nyder almisse</div>	inderste og nyder almisse af :	4	7
<input type="checkbox"/>	<div>ernærer sig ved at spinde (5 rows)</div> <div>ernærer sin ved samling (2 rows)</div>	ernærer sig ved at spinde	4	10

Select allDeselect all

Export clustersMerge selected & re-clusterMerge selected & CloseClose

Choices in cluster

Rows in cluster

Average length of choices

Length variance of choices

Afterwards we re-merge and cluster with “Nearest neighbor” and a radius within 2 and block chars 2

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: Nearest neighborDistance function: LevenshteinManage clustering functions

☐ Auto-update33 clusters found

Radius: 2Block chars: 2

Merge?	Values in cluster	New cell value	Cluster size	Row count
<input type="checkbox"/>	<div>væver (37 rows)</div> <div>væver</div> <div>væver</div>	væver	3	39
<input type="checkbox"/>	<div>tjener for pige (2 rows)</div> <div>tjener som pige</div> <div>tjener som pigen</div>	tjener for pige	3	4
<input type="checkbox"/>	<div>vanfer (13 rows)</div> <div>vanføer (2 rows)</div> <div>wanfer (2 rows)</div>	vanfer	3	17
<input type="checkbox"/>	<div>væverpige (3 rows)</div> <div>wæverpige (2 rows)</div> <div>væpige</div>	væverpige	3	6
<input type="checkbox"/>	<div>inderste (11 rows)</div> <div>binderske</div>	inderste	3	13

Select allDeselect all

Export clustersMerge selected & re-clusterMerge selected & CloseClose

Choices in cluster

Rows in cluster

Average length of choices

Length variance of choices

The rest we edit manually and are sorted roughly into bigger groups seen below.



This shows the top 10 jobs for women, where maid is the main occupation.



This is the overview of the facets.

Danish Kings

Kristoffer Segerstrøm

2025-03-05

The task here is to load your Danish Monarchs csv into R using the **tidyverse** toolkit, calculate and explore the kings' duration of reign with pipes `%>%` in **dplyr** and plot it over time.

Load the kings

Make sure to first create an **.Rproj** workspace with a **data/** folder where you place either your own dataset or the provided **kings.csv** dataset.

1. Look at the dataset that are you loading and check what its columns are separated by? (hint: open it in plain text editor to see)
2. Create a **kings** object in R with the different functions below and inspect the different outputs.

- `read.csv()`
- `read_csv()`
- `read.csv2()`
- `read_csv2()`

FILL IN THE CODE BELOW and review the outputs

```
kings1 <- read.csv("data/Danish_kings")
kings2 <- read_csv("data/Danish_kings")
kings3 <- read.csv2("data/Danish_kings")
kings4 <- read_csv2("data/Danish_kings")
```

Answer: 1. Which of these functions is a **tidyverse** function? Read data with it below into a **kings** object
`read_csv()` and `read_csv2()` is a part of the tidyverse-package (specifically **readr**). `read.csv()` and `read.csv2()` belongs to base R.

2. What is the result of running `class()` on the **kings** object created with a tidyverse function.
`class(kings4)`

[1] "tbl_df" "tbl" "data.frame"

3. How many columns does the object have when created with these different functions?

```
kings1 <- read_csv2("data/Danish_kings")
kings2 <- read_csv2("data/Danish_kings")
kings3 <- read_csv2("data/Danish_kings")
kings4 <- read_csv2("data/Danish_kings")
```

There is 11 columns

4. Show the dataset so that we can see how R interprets each column

```
glimpse(Danish_kings.csv) View(Danish_kings.csv)
```

COMPLETE THE BLANKS BELOW WITH YOUR CODE, then turn the 'eval' flag in this chunk to TRUE.

```
kings <- Danish_kings

class(kings)

glimpse(kings)

View(kings)
```

Calculate the duration of reign for all the kings in your table

You can calculate the duration of reign in years with `mutate` function by subtracting the equivalents of your `startReign` from `endReign` columns and writing the result to a new column called `duration`. But first you need to check a few things:

- Is your data messy? Fix it before re-importing to R
- Do your start and end of reign columns contain NAs? Choose the right strategy to deal with them: `na.omit()`, `na.rm=TRUE`, `!is.na()`

Create a new column called `duration` in the `kings` dataset, utilizing the `mutate()` function from `tidyverse`. Check with your group to brainstorm the options.

The code I used

```
kings <- kings %>%
  filter(!is.na(start_reign) & !is.na(end_reign))
kings <- kings %>%
  mutate(duration = end_reign - start_reign)
glimpse(kings)
```

Calculate the average duration of reign for all rulers

Do you remember how to calculate an average on a vector object? If not, review the last two lessons and remember that a column is basically a vector. So you need to subset your `kings` dataset to the `duration` column. If you subset it as a vector you can calculate average on it with `mean()` base-R function. If you subset it as a tibble, you can calculate average on it with `summarize()` `tidyverse` function. Try both ways!

- You first need to know how to select the relevant `duration` column. What are your options?
- Is your selected `duration` column a tibble or a vector? The `mean()` function can only be run on a vector. The `summarize()` function works on a tibble.
- Are you getting an error that there are characters in your column? Coerce your data to numbers with `as.numeric()`.
- Remember to handle NAs: `mean(X, na.rm=TRUE)`

The code I used

```
kings %>%
  summarise(avg_duration = mean(duration, na.rm = TRUE))
```

The average duration of reign for all rulers was 20.2 years

How many and which kings enjoyed a longer-than-average duration of reign?

You have calculated the average duration above. Use it now to `filter()` the `duration` column in `kings` dataset. Display the result and also count the resulting rows with `count()`

The code I used

```
long_reign_kings <- kings %>%  
  filter(duration > average_duration)  
long_reign_kings %>% count()  
print(long_reign_kings)
```

24 kings enjoyed a longer-than-average duration of reign

How many days did the three longest-ruling monarchs rule?

- Sort kings by reign `duration` in the descending order. Select the three longest-ruling monarchs with the `slice()` function
- Use `mutate()` to create `Days` column where you calculate the total number of days they ruled
- BONUS: consider the transition year (with 366 days) in your calculation!

#The code I used

```
top_3_kings <- kings %>%  
  arrange(desc(duration)) %>%  
  slice(1:3) %>%  
  mutate(days = duration * 365.25)  
print(top_3_kings)  
glimpse(top_3_kings)  
21915.00+18993.00+15705.75
```

The answer is 56613 days

And to submit this rmarkdown, knit it into html. But first, clean up the code chunks, adjust the date, rename the author and change the `eval=FALSE` flag to `eval=TRUE` so your script actually generates an output. Well done!

Assignment_week_12

Kristoffer Segerstroem 2025-03-21

```
knitr::opts_chunk$set(echo = TRUE,
                      warning = TRUE,
                      message = TRUE)

library(tidyverse)

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.4      v readr      2.1.5
## v forcats    1.0.0      v stringr   1.5.1
## v ggplot2    3.5.2      v tibble    3.2.1
## v lubridate  1.9.4      v tidyr     1.3.1
## v purrr      1.0.4
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

# For text mining:
library(pdftools) # Used to extract text from PDF files

## Using poppler version 23.04.0

library(tidytext) # Facilitates text analysis by working with words in a 'tidy' format
library(textdata) # Contains various sentiment dictionaries
library(ggwordcloud) # Used to create word clouds
```

Get the Game of Thrones text:

```
got_path <- "/Users/lars/Desktop/data/got.pdf"
got_text <- pdf_text(got_path) # Extracts text from the PDF file as a vector of strings (one per page)
```

Some wrangling:

```
got_df <- data.frame(got_text) %>%
  mutate(text_full = str_split(got_text, pattern = '\\\\n')) %>% # Splits the text by line breaks
  unnest(text_full) %>% # 'Unnests' the listed text so each line becomes a row in the dataframe
  mutate(text_full = str_trim(text_full)) # Removes leading and trailing spaces from each line
```

Get the tokens (individual words) in tidy format

```
got_tokens <- got_df %>%  
  unnest_tokens(word, text_full) # Splits the text into individual words (tokens), so each row contains
```

```
got_tokens
```

```
## # A tibble: 297,814 x 2  
##   got_text word  
##   <chr>   <chr>  
## 1 "      A GAME OF THRONES\n\n\n      Book One of A So~ a  
## 2 "      A GAME OF THRONES\n\n\n      Book One of A So~ game  
## 3 "      A GAME OF THRONES\n\n\n      Book One of A So~ of  
## 4 "      A GAME OF THRONES\n\n\n      Book One of A So~ thro~  
## 5 "      A GAME OF THRONES\n\n\n      Book One of A So~ book  
## 6 "      A GAME OF THRONES\n\n\n      Book One of A So~ one  
## 7 "      A GAME OF THRONES\n\n\n      Book One of A So~ of  
## 8 "      A GAME OF THRONES\n\n\n      Book One of A So~ a  
## 9 "      A GAME OF THRONES\n\n\n      Book One of A So~ song  
## 10 "     A GAME OF THRONES\n\n\n      Book One of A So~ of  
## # i 297,804 more rows
```

Remove stop words:

```
got_stop <- got_tokens %>%  
  anti_join(stop_words) %>% # Removes stop words (commonly used words like 'the', 'and', 'of') that do  
  select(-got_text) # Removes the original text column as it is no longer needed
```

```
## Joining with 'by = join_by(word)'
```

Count the words

```
got_stop %>%  
  count(word) %>%  
  arrange(-n)
```

```
## # A tibble: 11,295 x 2  
##   word      n  
##   <chr> <int>  
## 1 lord   1341  
## 2 ser    1023  
## 3 jon     787  
## 4 ned     743  
## 5 tyrion  591  
## 6 eyes    567  
## 7 hand    567
```

```
## 8 king      542
## 9 father    512
## 10 told     504
## # i 11,285 more rows
```

Word cloud of GoT words

```
got_top100 <- got_stop %>%  
  count(word) %>% # Counts the occurrences of each word  
  arrange(-n) %>% # Sorts words by frequency  
  head(100) # Keeps only the 100 most frequent words  
  
got_cloud <- ggplot(data = got_top100, aes(label = word)) +  
  geom_text_wordcloud() + # Visualizes the words in a word cloud  
  theme_minimal()  
  
got_cloud
```

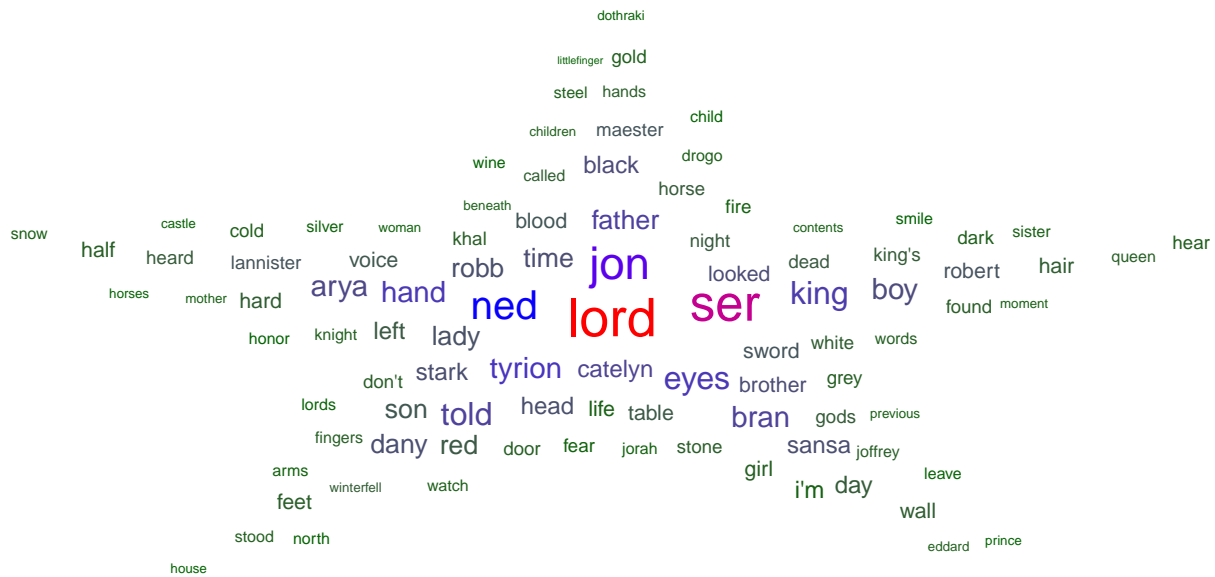


```
# Let's make the word cloud af star
```

```
ggplot(data = got_top100, aes(label = word, size = n)) +  
  geom_text_wordcloud_area(aes(color = n), shape = "star") +  
  scale_size_area(max_size = 12) +
```



```
scale_color_gradientn(colors = c("darkgreen","blue","red")) +
theme_minimal()
```



Sentiment analysis with afinn:

“afinn”: Words ranked from -5 (very negative) to +5 (very positive)

```
got_afinn <- got_stop %>%
  inner_join(get_sentiments("afinn")) # Matches words with sentiment scores from the AFINN lexicon (-5
```

Joining with ‘by = join_by(word)’

The negative words

```
get_sentiments(lexicon = "afinn")
```

```
## # A tibble: 2,477 x 2
##   word      value
##   <chr>     <dbl>
```

```
## 1 abandon      -2
## 2 abandoned    -2
## 3 abandons      -2
## 4 abducted     -2
## 5 abduction    -2
## 6 abductions    -2
## 7 abhor         -3
## 8 abhorred      -3
## 9 abhorrent     -3
## 10 abhors       -3
## # i 2,467 more rows
```

```
# Note: may be prompted to download (yes)
```

The positive words

```
library(tidytext)
afinn <- get_sentiments("afinn")
afinn_pos <- afinn %>% filter(value > 0) #finds the sentiments with a greater value than 0
head(afinn_pos)
```

```
## # A tibble: 6 x 2
##   word      value
##   <chr>    <dbl>
## 1 abilities      2
## 2 ability        2
## 3 aboard          1
## 4 absolve        2
## 5 absolved       2
## 6 absolves       2
```

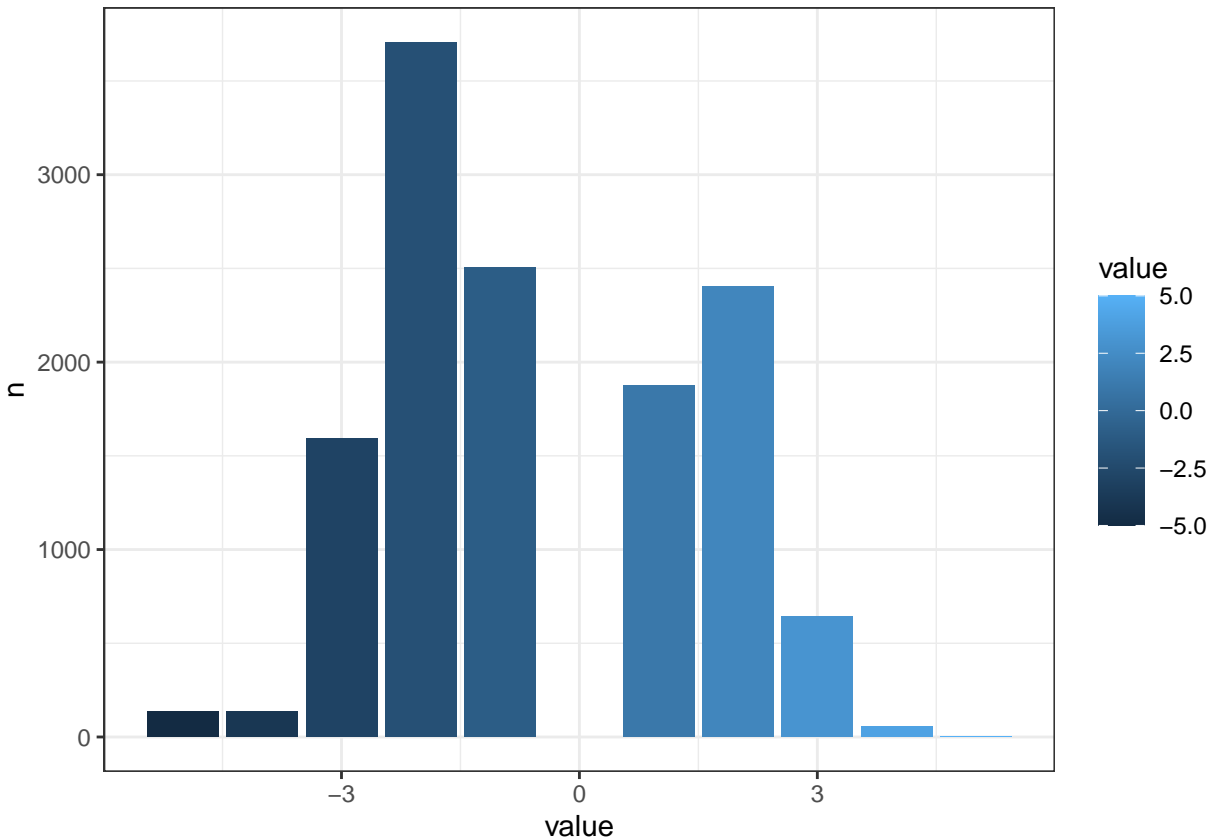
Word association with NRC

```
get_sentiments(lexicon = "nrc")
```

```
## # A tibble: 13,872 x 2
##   word      sentiment
##   <chr>    <chr>
## 1 abacus    trust
## 2 abandon   fear
## 3 abandon   negative
## 4 abandon   sadness
## 5 abandoned anger
## 6 abandoned fear
## 7 abandoned negative
## 8 abandoned sadness
## 9 abandonment anger
## 10 abandonment fear
## # i 13,862 more rows
```

Plot sentiment scores:

```
got_afinn_hist <- got_afinn %>%  
  count(value) # Counts the number of words for each sentiment score  
  
ggplot(data = got_afinn_hist, aes(x = value, y = n)) +  
  geom_col(aes(fill = value)) + # Visualizes sentiment scores with a bar chart  
  theme_bw()
```



Investigate which words have a sentiment score of 2 (quite positive)

```
got_afinn2 <- got_afinn %>%  
  filter(value == 2)  
got_afinn2 %>%  
  distinct(word)
```

```
## # A tibble: 201 x 1  
##   word  
##   <chr>  
## 1 smile  
## 2 fine  
## 3 glory  
## 4 hope  
## 5 smiled  
## 6 care
```

```
## 7 strength
## 8 peaceful
## 9 honor
## 10 carefully
## # i 191 more rows
```

#These commandoes isolates the 2-score words

Finding the unique 2-score words

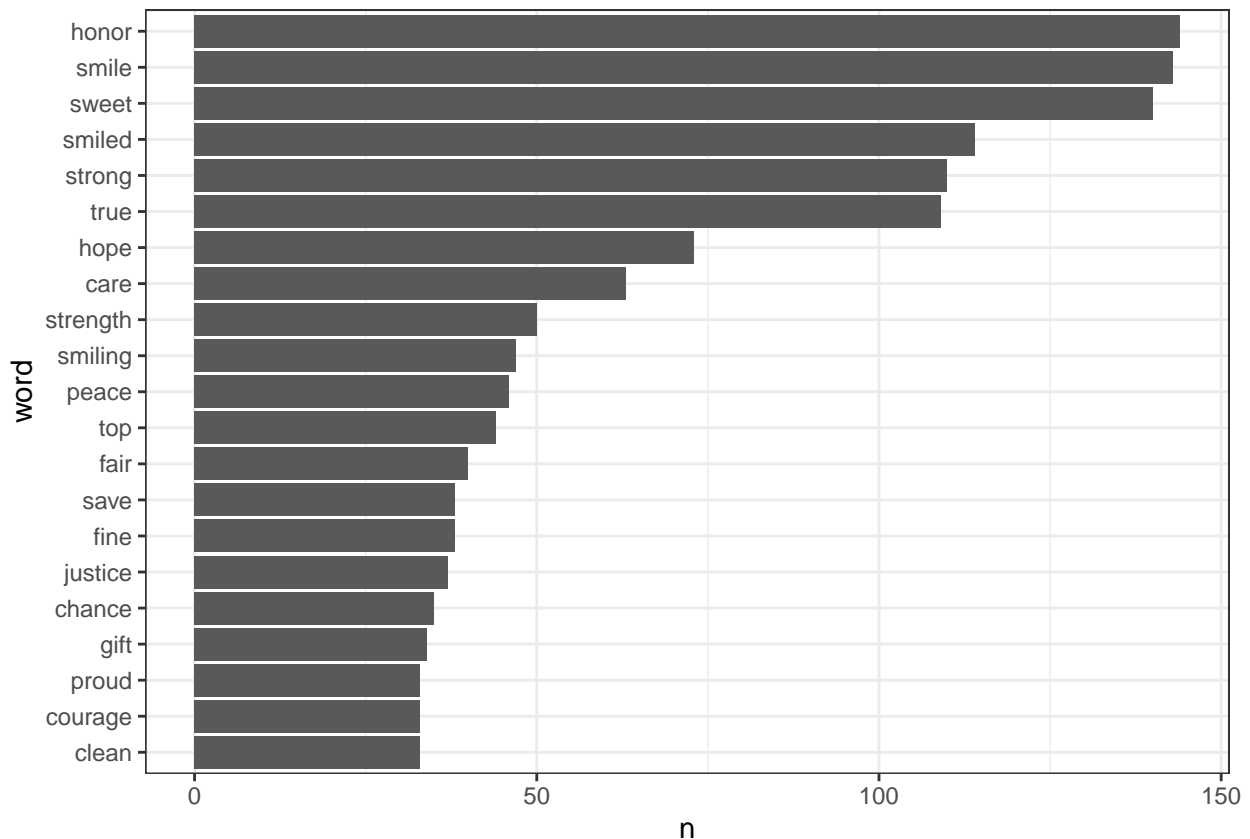
`unique(got_afinn2$word)` *#finds the unique words*

## [1] "smile"	"fine"	"glory"	"hope"
## [5] "smiled"	"care"	"strength"	"peaceful"
## [9] "honor"	"carefully"	"slick"	"top"
## [13] "gained"	"comfort"	"sweet"	"courage"
## [17] "daring"	"elegant"	"justice"	"heroes"
## [21] "fair"	"strong"	"brave"	"solid"
## [25] "proud"	"mercy"	"rescue"	"swift"
## [29] "smiling"	"true"	"noble"	"saved"
## [33] "gift"	"treasures"	"favorite"	"clean"
## [37] "rich"	"fearless"	"fortunate"	"likes"
## [41] "earnest"	"generous"	"chances"	"smiles"
## [45] "hug"	"kiss"	"approved"	"fond"
## [49] "honored"	"consent"	"peace"	"powerful"
## [53] "worthy"	"humor"	"entertaining"	"save"
## [57] "sincerely"	"festive"	"careful"	"stronger"
## [61] "bold"	"eager"	"favored"	"warmth"
## [65] "pardon"	"pardons"	"healthy"	"loving"
## [69] "chance"	"thoughtful"	"enjoy"	"privileged"
## [73] "positively"	"stout"	"encouragement"	"stable"
## [77] "smarter"	"ease"	"ambitious"	"improvement"
## [81] "hopeful"	"hopes"	"relieved"	"helping"
## [85] "cares"	"importance"	"favor"	"tender"
## [89] "welcomed"	"treasure"	"spirited"	"secured"
## [93] "courtesy"	"calm"	"resolved"	"courageous"
## [97] "comfortable"	"sympathy"	"reassuring"	"resolute"
## [101] "brisk"	"appeased"	"enjoying"	"hoping"
## [105] "intricate"	"rescued"	"glorious"	"adventures"
## [109] "friendly"	"astonished"	"reward"	"trusted"
## [113] "honest"	"clever"	"dear"	"favors"
## [117] "determined"	"strengthen"	"approval"	"slicker"
## [121] "sincere"	"jokes"	"joke"	"smartest"
## [125] "favorites"	"hero"	"adventure"	"abilities"
## [129] "strongest"	"courteous"	"exasperated"	"enjoys"
## [133] "rewarded"	"cherished"	"comforting"	"robust"
## [137] "cherish"	"sympathetic"	"surviving"	"cheered"
## [141] "worth"	"boldly"	"acquitted"	"unstoppable"
## [145] "cheer"	"fervent"	"applause"	"cheers"
## [149] "proudly"	"compassionate"	"bless"	"success"
## [153] "supported"	"kinder"	"improved"	"defender"

```
## [157] "tranquil"      "helpful"      "hail"         "tops"
## [161] "thankful"      "calmed"       "sunshine"     "opportunity"
## [165] "inspiration"   "survived"     "gain"         "freedom"
## [169] "growth"        "futile"       "swiftly"      "satisfied"
## [173] "congratulations" "confident"    "pardoned"     "energetic"
## [177] "esteemed"      "benefit"      "secure"       "accomplished"
## [181] "support"       "rewarding"    "ability"      "jovial"
## [185] "cheering"      "hailed"       "playful"      "confidence"
## [189] "consents"      "bargain"      "encouraged"   "relieving"
## [193] "accomplish"    "resolve"      "cleaner"      "prominent"
## [197] "serene"        "defenders"    "strengthened" "wealthy"
## [201] "revive"
```

```
got_afinn2_n <- got_afinn2 %>%
  count(word, sort = TRUE) %>%
  slice_max(n, n = 20) %>%
  mutate(word = fct_reorder(factor(word), n)) #set up for the plot

ggplot(data = got_afinn2_n, aes(x = word, y = n)) +
  geom_col() +
  coord_flip() +
  theme_bw()
```



I asked chatGPT how to show less words, and it came up with slice_max(n, n = 20) %>%

Let's find the median and mean of the sentiment of the words

```
got_summary <- got_afinn %>%
  summarize(
    mean_score = mean(value),
    median_score = median(value)
  )

print(got_summary)
```

```
## # A tibble: 1 x 2
##   mean_score median_score
##   <dbl>         <dbl>
## 1    -0.542          -1
```

The words in the GoT.pdf are not quite as positive, as they have a median score of -1

NRC lexicon for sentiment analysis

```
got_nrc <- got_stop %>%
  inner_join(get_sentiments("nrc")) # Matches words with the NRC lexicon, which categorizes words into

## Joining with 'by = join_by(word)'

## Warning in inner_join(., get_sentiments("nrc")): Detected an unexpected many-to-many relationship between
## i Row 148 of 'x' matches multiple rows in 'y'.
## i Row 9803 of 'y' matches multiple rows in 'x'.
## i If a many-to-many relationship is expected, set 'relationship =
##   "many-to-many"' to silence this warning.
```

Before we do the sentiment analysis, I will find out, which words are excluded

```
got_exclude <- got_stop %>%
  anti_join(get_sentiments("nrc")) #finds the excluded words
```

```
## Joining with 'by = join_by(word)'
```

```
got_exclude_n <- got_exclude %>%
  count(word, sort = TRUE) #counts the excluded words

head(got_exclude_n) #shows the result
```

```
## # A tibble: 6 x 2
##   word      n
##   <chr>  <int>
```

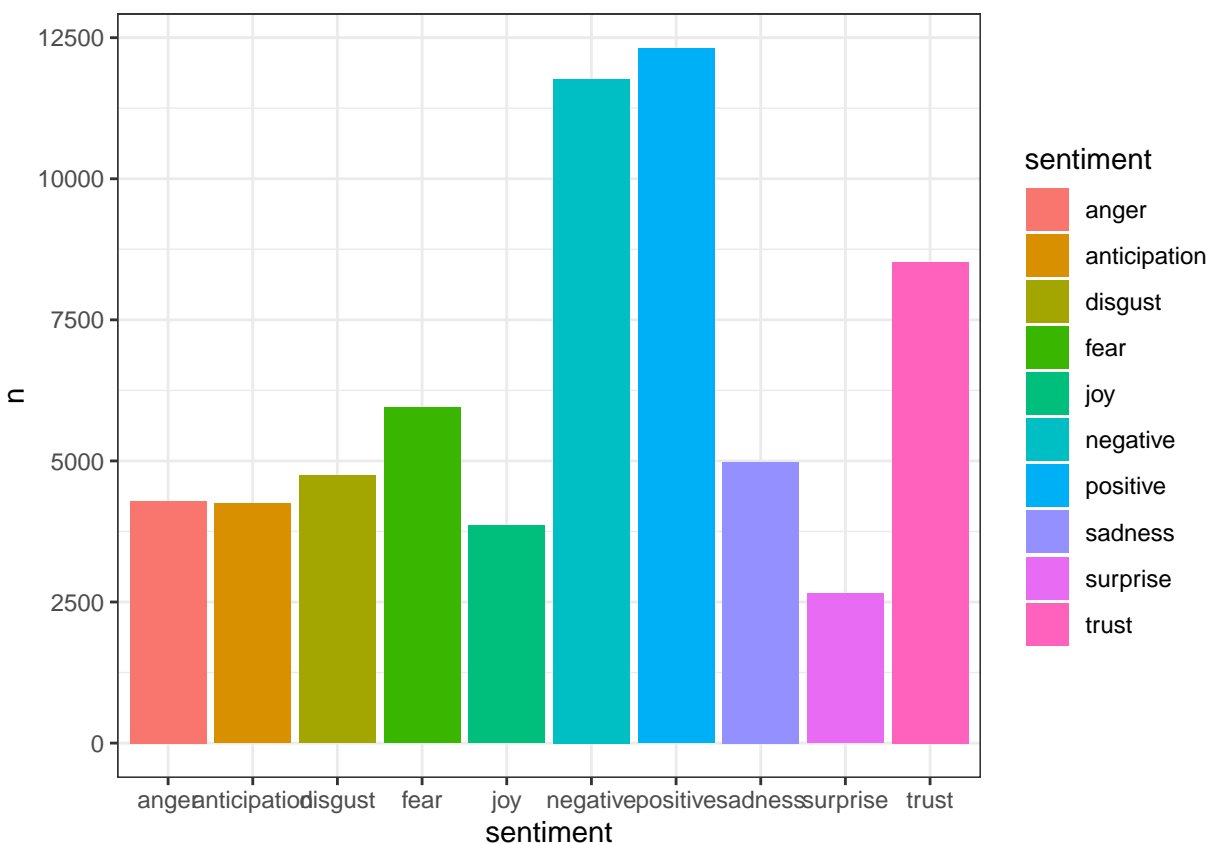
```
## 1 ser      1023
## 2 jon       787
## 3 ned       743
## 4 tyrion    591
## 5 eyes      567
## 6 hand      567
```

Above are the excluded words

Now we can continue analysing the sentiment

```
got_nrc_n <- got_nrc %>%
  count(sentiment, sort = TRUE) # Counts how many words belong to each sentiment category

ggplot(data = got_nrc_n, aes(x = sentiment, y = n, fill = sentiment)) +
  geom_col() + # Visualizes the results in a bar chart
  theme_bw()
```



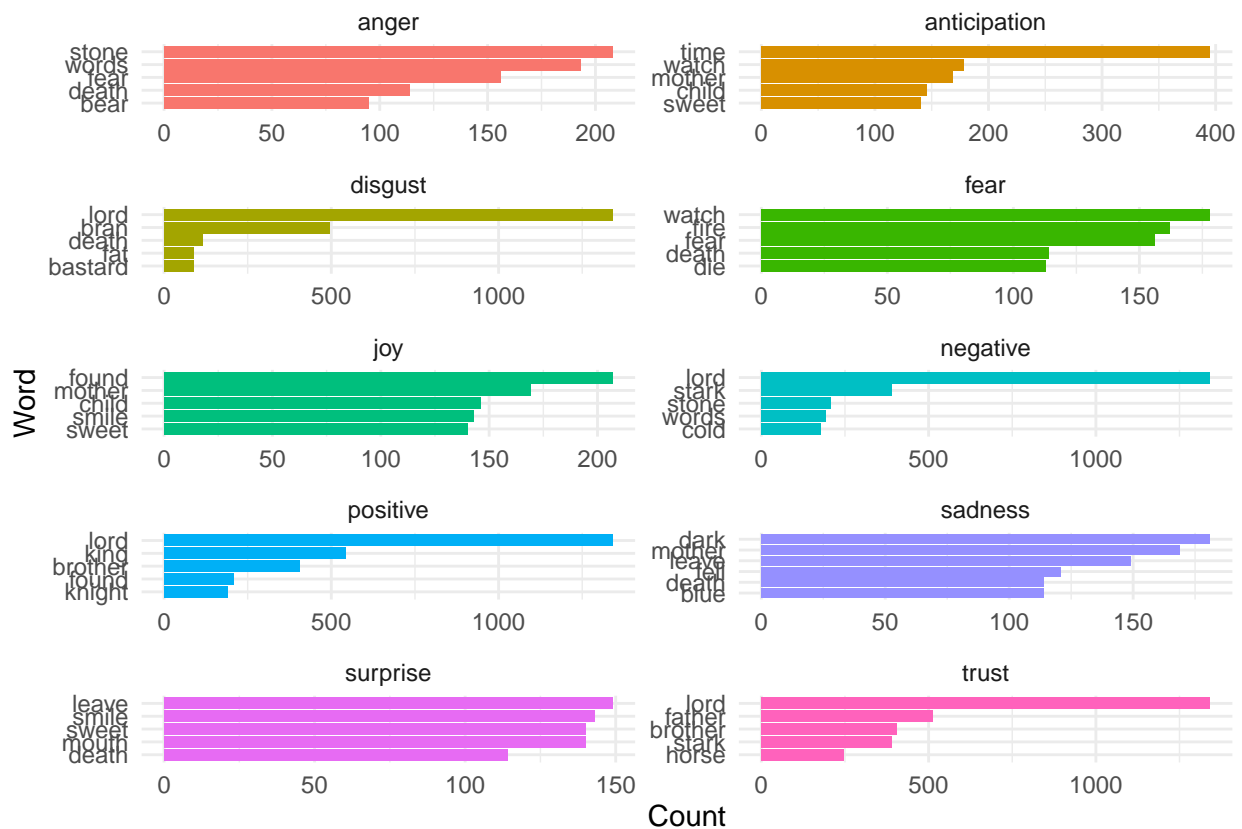
Creating a bar plot of the top words per sentiment category

```
got_nrc_n5 <- got_nrc %>%
  count(word, sentiment, sort = TRUE) %>% # Counts occurrences of each word categorized by sentiment
  group_by(sentiment) %>% # Groups by sentiment category
  top_n(5) %>% # Selects the top 5 most frequent words for each sentiment
  ungroup() # Removes grouping to allow further independent operations
```

Selecting by n

```
# Create a bar plot of the top words per sentiment category
got_nrc_gg <- ggplot(data = got_nrc_n5, aes(x = reorder(word, n), y = n, fill = sentiment)) +
  geom_col(show.legend = FALSE) + # Creates a bar plot without legend
  facet_wrap(~sentiment, ncol = 2, scales = "free") + # Creates separate panels for each sentiment
  coord_flip() + # Rotates the bar chart for better readability
  theme_minimal() + # Applies a minimalistic theme
  labs(x = "Word", y = "Count") # Labels the axes

# Show the plot
got_nrc_gg
```



I notice that the word “lord” is in many of the charts...

```
conf <- get_sentiments(lexicon = "nrc") %>%
  filter(word == "lord")
```



```
# Yep, check it out:  
conf
```

```
## # A tibble: 4 x 2  
##   word    sentiment  
##   <chr> <chr>  
## 1 lord    disgust  
## 2 lord    negative  
## 3 lord    positive  
## 4 lord    trust
```

It was true

Answering the task

My task

Taking this script as a point of departure, apply sentiment analysis on the Game of Thrones. You will find a pdf in the data folder. What are the most common meaningful words and what emotions do you expect will dominate this volume? Are there any terms that are similarly ambiguous to the ‘confidence’ above?

My answer

Using this script, we applied sentiment analysis to Game of Thrones. The most common meaningful words likely include character names, titles (e.g., “king,” “lord”), and thematic words such as “battle” or “death.”

In terms of emotions, we expect a dominance of fear, anger, and trust, as the book revolves around political intrigue, betrayal, and loyalty.

An ambiguous term similar to “confidence” is “lord.” It appears frequently but does not inherently convey a positive or negative sentiment—it depends on context.

Part 1 (Approx. 250 words):

Write 250 words explaining how Schriver and Jensen (2022) apply their overall arguments regarding the relationship between digital archives and historical research to the specific case they examine.

I deres artikel fra 2022 *En ny udfordring til historisk metode* argumenterer Schriver og Jensen for, at digitaliseringen af arkiver betydeligt transformerer historisk forskning ved både at udvide tilgængeligheden af kilder og introducere nye metodologiske udfordringer. De understreger, at digitale arkiver ændrer historikerens forhold til fortiden, især gennem søgefunktioner og metadata-strukturer, som medierer og til tider skaber uklarhed omkring arkivmaterialet. For at illustrere disse bredere påstande undersøger forfatterne et specifikt tilfælde: det digitale arkiv af den danske avis *Aarhus Stiftstidende*. Ved at analysere, hvordan søgeord og digital infrastruktur påvirker, hvilke kilder historikere finder og bruger, viser Schriver og Jensen, at digitalisering ikke er en neutral proces. Tværtimod former den historiske fortællinger ved at fremhæve visse materialer fremfor andre, afhængigt af valg af søgeord og fejl i OCR (optisk tegngenkendelse). Case-studiet viser, hvordan digital søgbarhed både kan styrke og begrænse forskerne. På den ene side giver det store digitale arkiv mulighed for nye typer kvantitative analyser og bredere kildeopdagelse. På den anden side kan det skabe en falsk følelse af omfattende dækning, samtidig med at det marginaliserer kilder, der er dårligt digitaliseret eller indekseret. Forfatterne opfordrer derfor historikere til kritisk at reflektere over, hvordan de digitale værktøjer, de bruger, påvirker deres forskningsspørgsmål, kildevalg og fortolkninger.

Schriver og Jensen konkluderer, at selvom digitale arkiver åbner spændende muligheder, kræver de også en revideret historisk metodologi - en, der forbliver opmærksom på de usynlige arkitekturer i det digitale og de bias, de måtte introducere.

Part 2 (750–1000 words, cohesive and well-edited):

Write 750–1000 words addressing the questions

Digitale ressourcer - Frihedsmuseet og NæstvedArkiverne

Denne opgave vil undersøge digitaliseringen af historiske arkiver med fokus på fotoarkivet fra Frihedsmuseet og NæstvedArkiverne. Den analyserer, hvordan digitale arkiver påvirker historisk forskning, med fokus på metadata, kildekritik og de udfordringer, der opstår ved fortolkning af digitaliserede materialer.

Frihedsmuseets fotoarkiv (Nationalmuseet) er en statslig institution, der modtager offentlig finansiering og har et nationalt kulturarvsfokus. Digitaliseret billeder fra Besættelsen understøtter formidlingen af Danmarks historie. Nationalmuseet prioriterer digitaliseret materiale med høj offentlig interesse og kulturel værdi, men med mindre fokus på forskningsrelevans.

NæstvedArkiverne var oprindeligt selvejende, men er i dag en kommunal institution. Deres digitaliseret af præget projektf finansiering, hvilket medfører ujævn dækning og selektion af samlinger.

Kilder i Frihedsmuseet og NæstvedArkiverne

For at finde relevante kilder benyttede vi to arkiver: Frihedsmuseet og NæstvedArkiverne. På Frihedsmuseet søgte vi på “Næstved” og “modstandsbevægelse”, hvilket gav os otte fotografiske poster. En søgning på “Næstved” og “Besættelsen” gav ingen resultater, hvilket betyder søgningen skal præciseres. Vi angav en tidsperiode fra 1939 til 1945 fandt vi flere billeder relateret til Næstved, som f.eks. et billede af skudhullerne i Næstved Kaserne fra den 29. august 1943 (Augustoprøret).



I NæstvedArkiverne søgte vi på ordet “Næstved”, hvilket førte til samlingen Besættelsestiden i Næstved 1940–1945. Under denne samling fandt vi det illegale julenummer af De Frie Danske fra 1943. Bladet giver et unikt indblik i hverdagen under Besættelsen, fordi det er en samtidig kilde skabt af modstandsbevægelsen selv. Det giver direkte adgang til modstandsfolkets egne formuleringer, prioriteringer og stemningsbilleder på et tidspunkt, hvor censur og kontrol prægede det offentlige rum.



Som Helle Strandgaard Jensen skriver i sin artikel *Digital Archival Literacy For (All) Historians*, “Metadata is in its essence, data about data. It can be what was usually registered in a finding aid (e.g., date, creator, location). But it can also be much more than that including vere detailed subject descriptions spread out over categories¹”. Denne forståelse af metadata er vigtig for at vurdere, hvordan informationen er organiseret og præsenteret i arkiverne, for at kunne lave en nuanceret analyse af kilderne, som vi så på hjemmesiderne under vores kildesøgning.

¹ Jensen, Helle Strandgaard. “Digital Archival Literacy for (All) Historians.” *Media History* 27, (April 3, 2021): side 8. <https://doi.org/10.1080/13688804.2020.1779047>

Informationen i kilderne og arkivskaberne

Peter Jepsen er fotografen bag billedet af skudhullerne på Næstved Kaserne fra Augustoprøret. Dermed er Jepsen arkivskaber, da han er ophavsmand til kildematerialet. Frihedsmuseet, som er en del af Nationalmuseet, fungerer som arkivinstitution, da de har organiseret, bevaret og digitaliseret materialet og har gjort det offentlig. Billedet er en del af et større arkiv, som er organiseret efter temaer som "modstandsbevægelsen". Der er tre øvrige billeder i arkivet er relateret til Augustoprøret, som gør dem vigtige kilder til at forstå den historiske kontekst. Selvom billederne er knyttet til en specifik dato, er der ikke nødvendigvis en sammenhæng, f.eks. et fotografisk projekt, der binder billederne sammen. I NæstvedArkiverne fandt vi kilden "De frie Danske - illegalt blad julenummer 1943", som stammer fra modstandsbevægelsen, der producerede de illegale blade under Besættelsen, hermed bliver skribenterne af avisen til arkivskaberne. NæstvedArkiverne er arkivinstitutionen, da de opbevarer, digitaliserer og tilgængeliggør materialet. Kilden giver et indblik i modstandens aktiviteter i Danmark og er en vigtig beretning om modstandens arbejde i en tid med censur og undertrykkelse. Pomerantz forklarer også, hvor stor en vigtighed, metadataen spiller i opgaver som denne "*Ressource discovery relies on good metadata like this²*". Citatet understøtter vigtigheden af metadata i arkivarbejde, da korrekt metadata gør det muligt at opdage og tilgå kilder effektivt. I NæstvedArkiverne, er det arkivinstitutionen, som organiserer kilderne, hvilket gør dem lette at finde og anvende i videre forskning. Dette ses også ved vores brug af søgeordene "Næstved" eller "Modstandsbevægelse". Her bliver søgningen efter nålen nemmere. Der bliver filtreret irrelevant stof fra og lagt fokus på, hvad vi gerne vil finde frem til, hvilket er et indblik i besættelsestiden i Næstved. Vi ved, vi vil finde frem til noget om besættelsestiden, derfor er der også valgt årstallene (1940-1945), hvor besættelsen fandt sted.

Det "gemte" data

Det er muligt at finde oplysninger om relateret materiale fra samme arkivskaber, selvom materialet ikke er digitaliseret. Arkivinstitutioner opretholder typisk detaljerede kataloger, der omfatter både digitaliseret og ikke-digitaliseret arkivalier. Selv når selve arkivalierne ikke er tilgængeligt online, fungerer metadata som et "kort", der guider forskere gennem arkivets samlinger. Som Pomerantz påpeger, forenkler metadata komplekse informationsrum og gør det lettere at finde. Oplysninger om digitaliseringsprocessen er ofte tilgængelige via arkivets hjemmeside. Dette kan inkludere detaljer

² Pomerantz, Jeffrey. "Introduction." In *Metadata, 1–18. Essential Knowledge. The MIT Press, 2015. Side 17.*
<http://www.jstor.org/stable/j.ctt1pv8904.5>.

om arbejdsgangen, f.eks. om arkivet anvender OCR til søgning i arkivaliet, maskinlæring til automatisk tagging eller crowdsourcing for at inddrage offentligheden i at transskribere eller annotere arkivalier. Nationalmuseet skriver selv i deres forskningsberetning fra 2023: "*Det er Nationalmuseets opgave at "belyse Danmarks og verdens kulturer og deres indbyrdes afhængighed i et nationalt og internationalt perspektiv" (jf bekendtgørelse af museumsloven fra 2014), og det gør vi bl a med udgangspunkt i de enestående samlinger og arkivalier, som vi forvalter, og gennem en tværfaglig forskningspraksis, der inddrager humanistiske, samfundsvidenskabelige og naturvidenskabelige perspektiver*"³". Arkiverne anvender tekniske specifikationer som TIFF, JPEG og PDF for at sikre bevaringen. De følger også etablerede metadata-standards som Dublin Core og METS/ALTO for at sikre konsistens på tværs af samlinger. Citeres en digital kilde, skal det angives, om der er tale om en digital version, og inkluderes oplysninger om, hvor og hvornår den er tilgængelig, samt oplysninger om det oprindelige dokument.

Kildekritisk vurdering - er det muligt?

Vi føler os i stand til at anvende kildekritik på de digitale kilder fra både Frihedsmuseet og NæstvedArkiverne, men vi er også opmærksom på udfordringerne, der følger med digitaliseringen. Som Schrøder & Jensen påpeger, "*I forlængelse heraf har en række historikere argumenteret for, at kritisk-metodisk refleksion over kulturarvsdigitaliseringen og historikeres brug af digitale arkiver er forudsætningen for en metodisk valid forskning*"⁴". Dette understreger vigtigheden af at reflektere, hvordan arkiverne er digitaliseret og organiseret, da metadata og arkivstrukturen påvirker, hvilke kilder vi har adgang til, og hvordan vi fortolker dem.

Selvom vi har haft nem adgang til kilderne, skal vi være opmærksomme på, at digitalisering kan medføre tab af kontekst. Arkivskaberens beslutning om, hvilke materialer der digitaliseres, kan påvirke vores forståelse af kilderne og den historiske kontekst. Det kræver derfor en kritisk tilgang til både de digitale arkivalier og de metadata, der knytter sig til dem.

³ Nationalmuseets Forskningsberetning 2023, tilgængelig 9/4 12:00, https://natmus.dk/fileadmin/user_upload/Editor/natmus/forsking/dokumenter/Forskningsberetninger/Forskningsberetning_2023.pdf, side 7.

⁴ Schrøder, Astrid Ølgaard Christensen, og Helle Strandgaard Jensen. "Arkivets Digitalisering. En Ny Udfordring Til Historisk Metode?" *Temp. Tidsskrift For Historie* Årg. 13, Nr. 25 (2022). Side 6.

Part 3 (250 words, cohesive and well-edited):

Based on the course sessions (including hands-on experiences) and your answers above, discuss how the digitalization process and digital sources relate to historical research and methodologies. Is it, for instance, but not mandatory, possible to apply core historical re- search methods to the digitalization process (why? - Why not?)

Digitalisering har haft en betydelig indvirkning på historisk forskning og metodologi, og har både åbnet op for muligheder og præsenteret udfordringer. Processen med at digitalisere kilder gør det muligt for historikere at udvide adgangen til primære materialer, hvilket giver adgang til et bredere udvalg af dokumenter. Denne demokratisering af historiske ressourcer har gjort det muligt for forskere verden over at få adgang til materialer, de ellers måske ikke ville have haft adgang til, især i fjerntliggende eller begrænsede arkiver. En af de største fordele ved digitalisering er den effektivitet, den bringer til forskningsprocessen. Digitale kilder kan indekseres, søges i og analyseres hurtigere end fysiske kilder, hvilket gør det muligt for historikere at identificere mønstre og forbindelser, som måske ville have været svære at få øje på i traditionel forskning. Dette er især relevant i forbindelse med storskala-projekter, der involverer store mængder data, som for eksempel studier af økonomisk historie eller sociale tendenser over lange perioder. Men digitaliseringsprocessen rejser også vigtige spørgsmål om historiske metodologier. Mens traditionelle metoder som kildekritik og kontekstuel analyse fortsat er essentielle, kan processen med at digitalisere kilder komplicere disse metoder. For eksempel kan digitalisering ændre konteksten eller det originale materiale, og måden en kilde præsenteres på online kan påvirke, hvordan den bliver fortolket. I nogle tilfælde kan metadata eller digitale repræsentationer forvrænge den oprindelige betydning af en kilde, hvilket skaber behov for nye kritiske tilgange. Derfor, selvom de grundlæggende historiske forskningsmetoder kan anvendes på digitale kilder, må historikere være opmærksomme på de unikke udfordringer, som digitalisering medfører. Det kræver, at traditionelle metoder tilpasses for at sikre, at integriteten af den historiske analyse bevares.

Litteraturliste:

- Jensen, Helle Strandgaard. "Digital Archival Literacy for (All) Historians." *Media History* 27, no. 2 (April 3, 2021): 251–65. <https://doi.org/10.1080/13688804.2020.1779047>
- Pomerantz, Jeffrey. "Introduction." In *Metadata*, 1–18. Essential Knowledge. The MIT Press, 2015. <http://www.jstor.org/stable/j.ctt1pv8904.5>
- Schriver, Astrid Ølgaard Christensen, and Helle Strandgaard Jensen. "Arkivets Digitalisering. En Ny Udfordring Til Historisk Metode?" *Temp. Tidsskrift For Historie Årg.* 13, no. Nr. 25 (2022): 5–27.
- Nationalmuseets hjemmeside, <https://natmus.dk/historisk-viden/forskning/>, tilgået 9/4 kl. 11:15
- NæstvedArkiverne, *De Frie Danske – Illegalt blad julenummer 1943*, <https://online.flippingbook.com/view/881125/>, tilgået 9/4 kl. 11:30
- Nationalmuseet Online Samlinger, <https://samlinger.natmus.dk/assetbrowse?collection=FHM&keyword=n%C3%A6stved,modstandsbev%C3%A6gelse&yearfrom=1939>, tilgået 9/4 kl. 11:50

Prison and Conviction Analysis

Kristoffer Segerstrøm, Lukas Benner and Emil Hansen

Loading and combining Excel datasets

We load multiple Excel files (from 1900 to 2024) and combine them into one dataset for analysis. We used the `file.choose()` to find out where the right excel-file was, hence the long file name.

```
file_1900 <- read_excel("/Users/lars/Desktop/Excel til eksamen /1900.xlsx")
file_1925 <- read_excel("/Users/lars/Desktop/Excel til eksamen /1925.xlsx")
file_1950 <- read_excel("/Users/lars/Desktop/Excel til eksamen /1950.xlsx")
file_1975 <- read_excel("/Users/lars/Desktop/Excel til eksamen /1975.xlsx")
file_2000 <- read_excel("/Users/lars/Desktop/Excel til eksamen /2000.xlsx")
file_2024 <- read_excel("/Users/lars/Desktop/Excel til eksamen /2024.xlsx")

# Combining them into a single data frame
combined_data <- bind_rows(file_1900, file_1925, file_1950, file_1975, file_2000, file_2024)

# Previewing the data
head(combined_data)
```

```
## # A tibble: 6 x 7
##   År Køn  Forbrydelser      Antal pr. 100.000 indbygge~1 Forbrydelse Alder
##   <dbl> <chr> <chr>          <dbl>          <dbl> <chr>      <chr>
## 1  1900 Mand Statsforbrydelser      0            0      <NA>    <NA>
## 2  1900 Mand Forbrydelser mod o~    201          8.20    <NA>    <NA>
## 3  1900 Mand Forbrydelser i emb~     4          0.163    <NA>    <NA>
## 4  1900 Mand Mened                  1          0.0408    <NA>    <NA>
## 5  1900 Mand Falsk forklaring f~    23          0.939    <NA>    <NA>
## 6  1900 Mand Forbrydelser mht. ~     0            0      <NA>    <NA>
## # i abbreviated name: 1: 'pr. 100.000 indbyggere'
```

Data preparation

We make sure that “Year” and “Convictions per 100k” are numeric and remove rows with missing values via these codes:

```
combined_data <- combined_data %>%
  mutate(
    År = as.numeric(År),
    pr_100k = as.numeric(`pr. 100.000 indbyggere`)
  ) %>%
  filter(!is.na(pr_100k))
```

Convictions over time

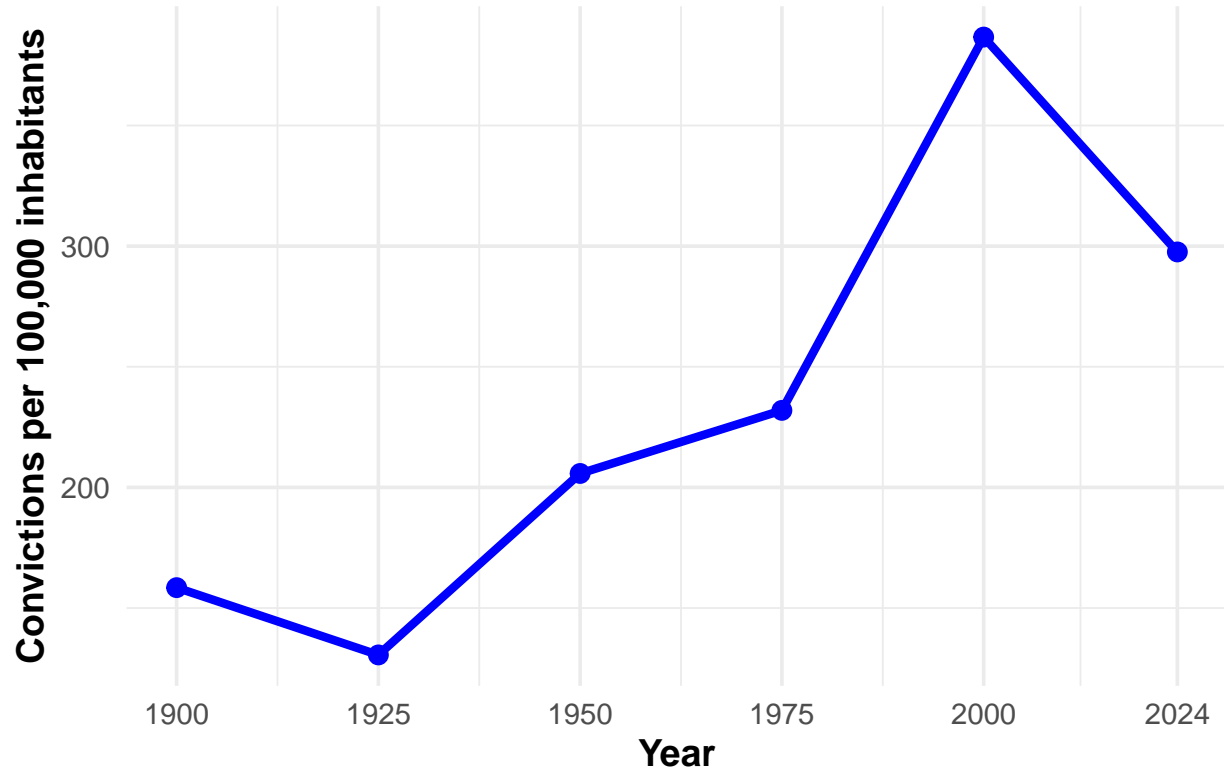
We summarize total convictions per 100,000 inhabitants per year.

```
yearly_data <- combined_data %>%
  group_by(År) %>%
  summarise(total_convicts_per_100k = sum(pr_100k, na.rm = TRUE))

# We plot the data via ggplot
# We have used the plot.title and axis.title to make our graphs and layout nicer to look at
# then we used the scale_x_continuous() to make sure we have the right data in the x-axes of the chart
ggplot(yearly_data, aes(x = År, y = total_convicts_per_100k)) +
  geom_line(color = "blue", size = 1.5) + geom_point(color = "blue", size = 3) +
  labs(
    title = "Convicted persons over time",
    x = "Year",
    y = "Convictions per 100,000 inhabitants"
  ) +
  scale_x_continuous(breaks = c(1900, 1925, 1950, 1975, 2000, 2024)) + theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", size = 16, hjust = 0.5),
    axis.title = element_text(face = "bold")
  )
)
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

Convicted persons over time



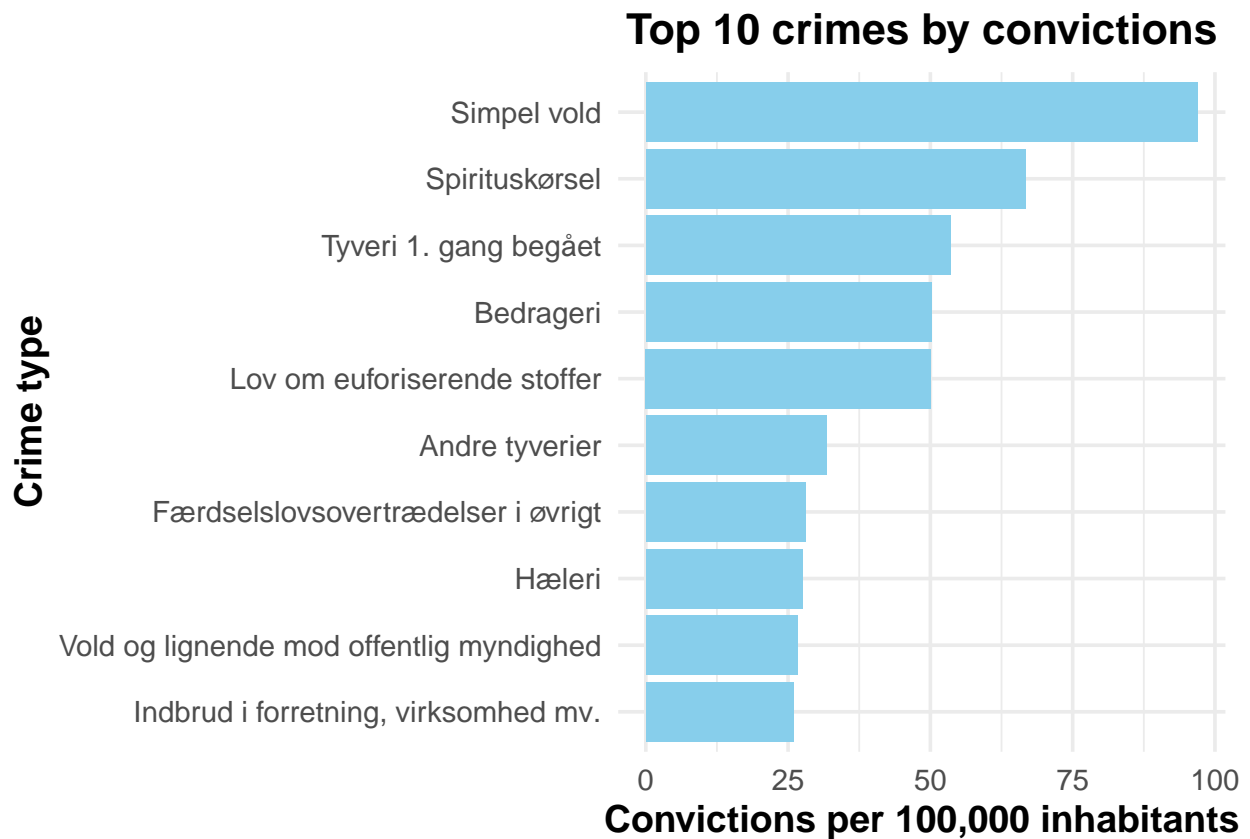
Top 10 crime categories

We identify and visualize the 10 crime types with the highest conviction rates.

```
# We will now group the combined data by crime type ("Forbrydelser"),
# then we will calculate the total conviction rate per 100,000 inhabitants for each crime
crime_data <- combined_data %>%
  group_by(Forbrydelser) %>%
  summarise(total_per_100k = sum(pr_100k, na.rm = TRUE)) %>%
  filter(!is.na(Forbrydelser), Forbrydelser != "NA") %>%
  slice_max(total_per_100k, n = 10)
#The slice_max() is used to make the top 10

# With this ggplot we are creating a horizontal bar chart showing the top 10 crime types by conviction
ggplot(crime_data, aes(x = reorder(Forbrydelser, total_per_100k), y = total_per_100k)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  coord_flip() +
  labs(
    title = "Top 10 crimes by convictions",
    x = "Crime type",
    y = "Convictions per 100,000 inhabitants"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", size = 16, hjust = 0.5),
```

```
axis.title = element_text(face = "bold")
)
```



Convicted by age

```
# We load the data
age_data <- read_excel("/Users/lars/Downloads/Domfældte alder.xlsx")

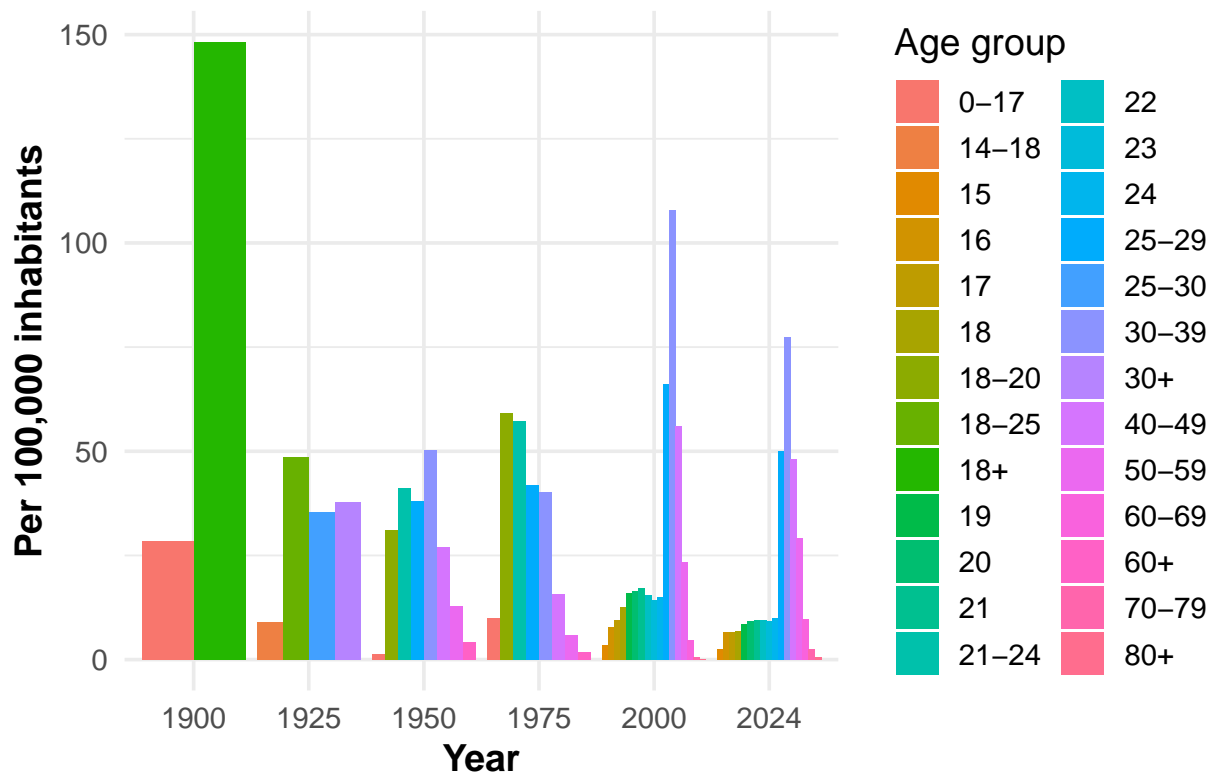
# We use the trimws() commando to clean the column names
names(age_data) <- trimws(names(age_data))

# We rename columns for consistency
age_data <- age_data %>%
  rename(
    year = Year,
    age_group = Age,
    convicted = Convicted,
    rate_per_100k = `Per 100,000 inhabitants`
  )

# We filter the data to only relevant years
age_data <- age_data %>%
  filter(year %in% c(1900, 1925, 1950, 1975, 2000, 2024))
```

```
# then we use the ggplot to make the image
ggplot(age_data, aes(x = factor(year), y = rate_per_100k, fill = age_group)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(
    title = "Convicted persons per 100,000 inhabitants by age group",
    x = "Year",
    y = "Per 100,000 inhabitants",
    fill = "Age group"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", size = 16, hjust = 0.5),
    axis.title = element_text(face = "bold")
  )
)
```

victed persons per 100,000 inhabitants by age group



Number of incarcerated people

```
# We load the data
incarceration_data <- read_excel("/Users/lars/Desktop/Excel til eksamen /Fængslinger i alt.xlsx")

# We use the trimws() again to once again to clean and remove extra invisible characters (whitespace) f
names(incarceration_data) <- trimws(names(incarceration_data))
```

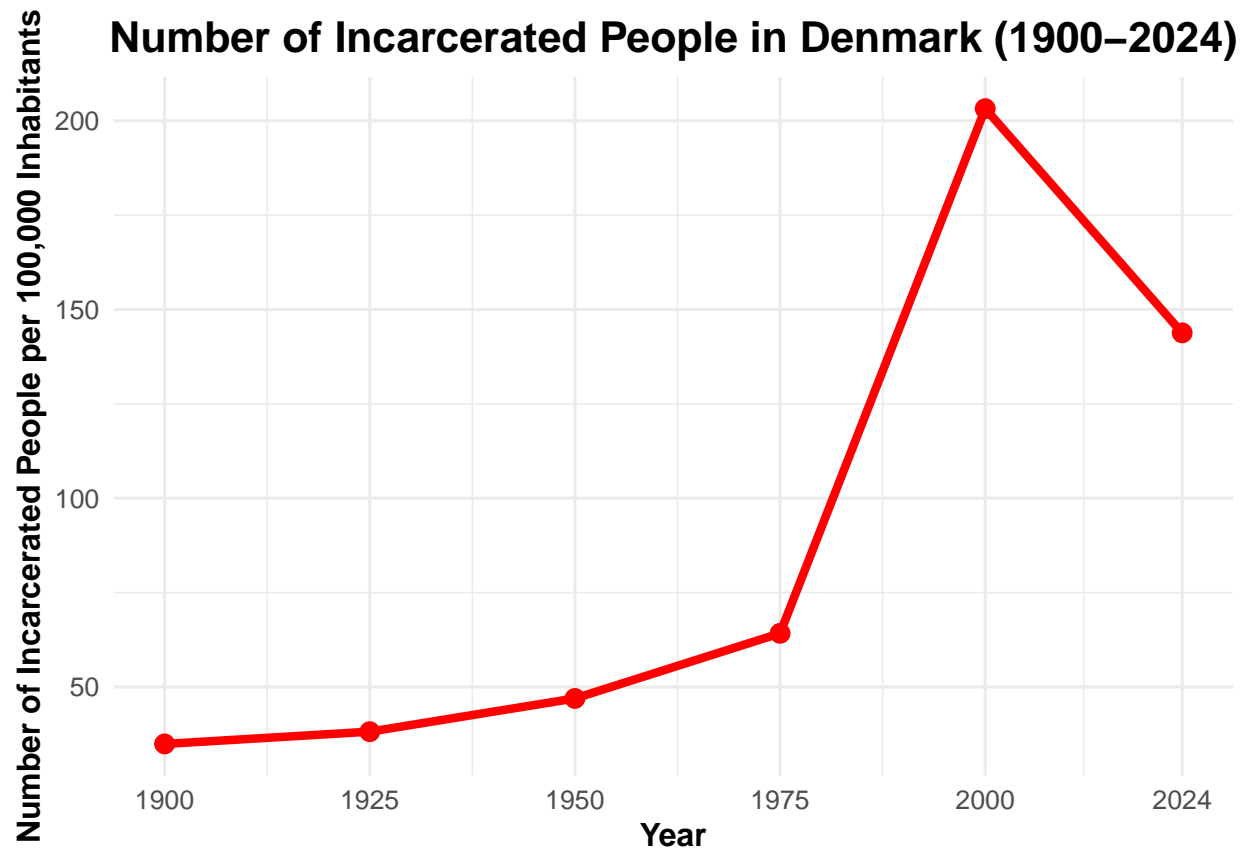
```

# We rename the columns to shorter, consistent, and English-friendly names:
incarceration_data <- incarceration_data %>%
  rename(
    year = Year,
    gender = Gender,
    count = Count,
    rate_per_100k = `Per 100,000 inhabitants`
  )

# then we summarise total incarceration rate (both genders) per year
summary_data <- incarceration_data %>%
  group_by(year) %>%
  summarise(
    total_rate_per_100k = sum(rate_per_100k)
  )

# then we create the plot
ggplot(summary_data, aes(x = year, y = total_rate_per_100k)) +
  geom_line(color = "red", linewidth = 1.5) +
  geom_point(color = "red", size = 3) + scale_x_continuous(breaks = c(1900, 1925, 1950, 1975, 2000, 2024))
labs(
  title = "Number of Incarcerated People in Denmark (1900-2024)",
  x = "Year",
  y = "Number of Incarcerated People per 100,000 Inhabitants"
) +
theme_minimal(base_size = 12) +
theme(
  plot.title = element_text(face = "bold", size = 16, hjust = 0.5),
  axis.title = element_text(face = "bold")
)

```



Number of incarcerated people divided by gender

```
# We read the excel data
incarceration_data <- read_excel("/Users/lars/Desktop/Excel til eksamen /Fængslinger i alt.xlsx")

# then we clean the column names
names(incarceration_data) <- trimws(names(incarceration_data))

# We rename the data for consistency
incarceration_data <- incarceration_data %>%
  rename(
    year = Year,
    gender = Gender,
    count = Count,
    rate_per_100k = `Per 100,000 inhabitants`
  )

# then we summarise the average rate per gender per year
summary_data <- incarceration_data %>%
  group_by(year, gender) %>%
  summarise(rate = sum(rate_per_100k), .groups = "drop")

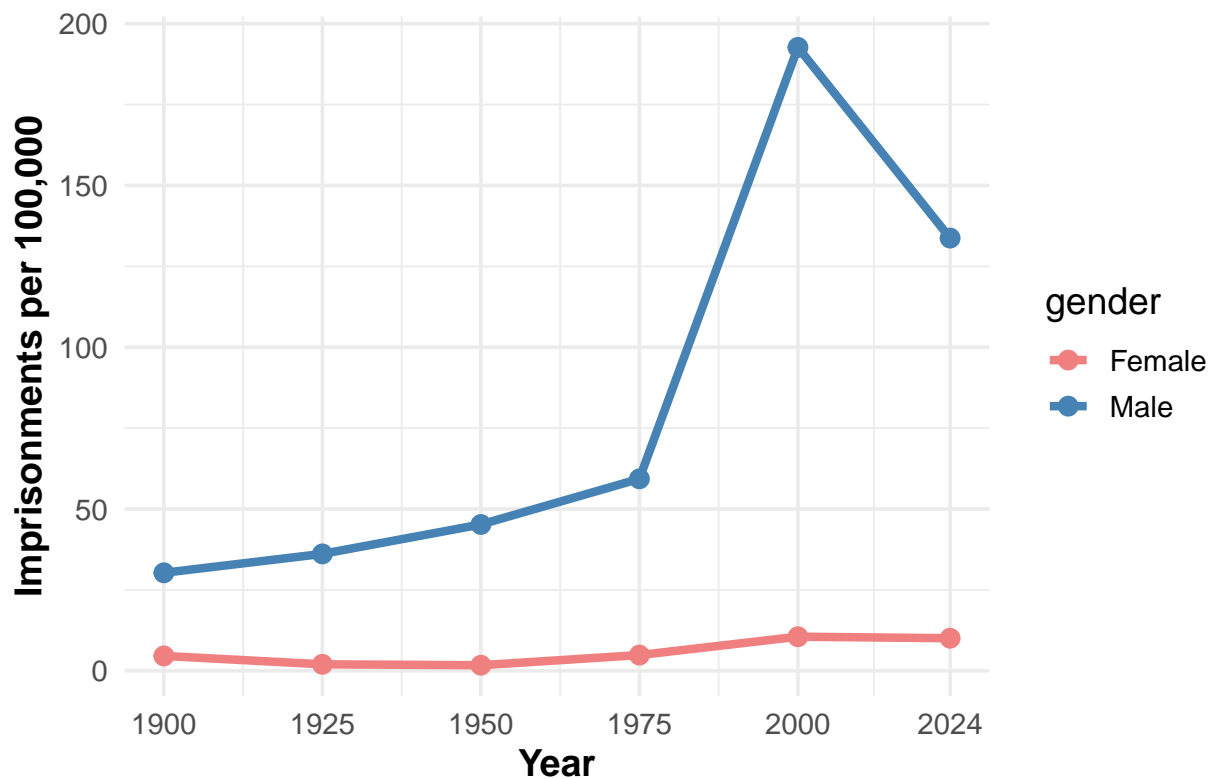
# At last we create the plot
ggplot(summary_data, aes(x = year, y = rate, color = gender)) +
```

```

geom_line(linewidth = 1.5) +
geom_point(size = 3) +
scale_x_continuous(breaks = c(1900, 1925, 1950, 1975, 2000, 2024)) +
labs(
  title = "Imprisonments per 100,000 Inhabitants Over Time by Gender",
  x = "Year",
  y = "Imprisonments per 100,000"
) +
theme_minimal(base_size = 14) +
theme(
  plot.title = element_text(face = "bold", size = 16, hjust = 0.5),
  axis.title = element_text(face = "bold")
) +
scale_color_manual(
  values = c("Male" = "steelblue", "Female" = "lightcoral")
)

```

prisonments per 100,000 Inhabitants Over Time by Gender



Animation of the gender plot

As a bonus, we have made the gender chart as a video.

```

# We start of by loading and cleaning the data
incarceration_data <- read_excel("Fængslinger i alt.xlsx", sheet = "Data-dam")

```



```

names(incarceration_data) <- trimws(names(incarceration_data))

incarceration_data <- incarceration_data %>%
  rename(
    year = Year,
    gender = Gender,
    count = Count,
    rate_per_100k = `Per 100,000 inhabitants`
  )

# then we summarise rate by gender and year
gender_summary <- incarceration_data %>%
  group_by(year, gender) %>%
  summarise(rate = sum(rate_per_100k), .groups = "drop") %>%
  filter(year %in% c(1900, 1925, 1950, 1975, 2000, 2024))

# Now we use transition_reveal to create an animated plot
p <- ggplot(gender_summary, aes(x = year, y = rate, color = gender, group = gender)) +
  geom_line(size = 1.5) +
  geom_point(size = 3) +
  scale_color_manual(values = c("Male" = "steelblue", "Female" = "lightcoral")) +
  scale_x_continuous(breaks = c(1900, 1925, 1950, 1975, 2000, 2024), limits = c(1900, 2024)) +
  labs(
    title = 'Imprisonments per 100,000 by Gender (1900-2024)',
    x = 'Year',
    y = 'Imprisonments per 100,000',
    color = 'Gender'
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", size = 16, hjust = 0.5),
    axis.title = element_text(face = "bold")
  ) +
  transition_reveal(year)

# then we use these commands to make the animation
animate(p, fps = 10, width = 800, height = 500, renderer = av_renderer("imprisonments_by_gender.mp4"))

```

The animation created in this box will be in the “Bilag” section and in github. We were forced to write `eval = FALSE` to be able to save the file as a HTML.