LINKÖPINGS UNIVERSITET

Visualisering 3

732G48 Introduktion till statistik och dataanalys

Adrian Mansur, Thong Vinh Phat, Viet Tien Trinh, Duy Thai Pham

Höstterminen 2023

Innehåll

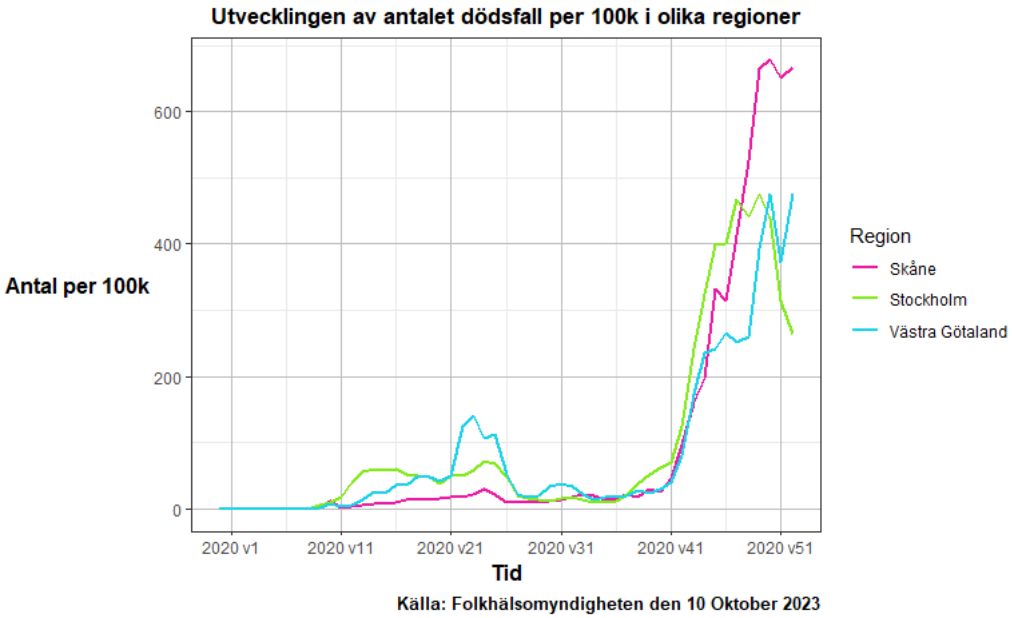
1. Inledning

I denna rapport ska gruppen göra det sista visualisering momentet. Under detta moment två nya visualiseringsmetoder som kartdiagram och 3D-spridningsdiagram. Övriga uppgifter handlar om självständig datamaterialhantering och såväl som visualisering.

2. Uppgifterna

2.1

Figur 1: Linjediagram med tre tidserier "Skåne, Stockholm, Västra Götaland", visar hur skillnad och likheter om antal dödsfall per 100k.

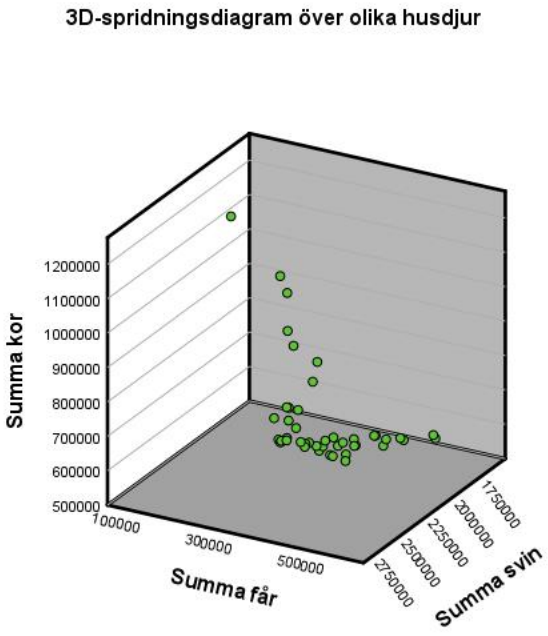


Ovanför beskrivs en linjediagram från R Studio med tre tidsserier av antalet dödsfall per 100 000 i olika svenska regioner "Skåne, Stockholm och Västra Götaland" från 2020_veckan1 till 2020_veckan53. I grafen inkluderades 2 ökningar, den första är från veckan 10 till veckan 24 då alla regioner ökades av dödsfall och man kan hitta att inom de tre regionerna har Västra Götaland högst antal av dödsfall då det var nästan 150 i veckan 23, näst mest är Stockholm ca 60 i veckan 24 och Skåne ca 25 i veckan 24. Sedan antalet av dödsfall minskades det gradvis i tre regioner och håller sig låg tills veckan 40 då antalet ökades igen och det ökade mycket kraftig jämfört med första ökningen. Vid andra ökning var Skåne den regionen som hade flest antal dödsfall och nått nästan 700 i vecka 50, största antalet av dödsfall i Stockholm och Västra Götaland i andra ökning är nästan 500 i vecka 49 och 50. Vid ungefär vecka 25 och 40 hade alla tre regioner mindre än 100 dödsfall med olika dödsfallnivåer, det var relativt lågt dödsfall jämfört med andra perioder inom grafen.

2.2

I denna uppgift skapas ett 3D-spridningsdiagram med hjälp av SPSS som visar samband mellan olika husdjur från det datamaterial som kommer från SCB:s Statistikdatabas. I datamaterialet som finns inlagd i Excel finns det 4 kolumner, en kolumn visar olika år från 1961 till 2007 och de övriga visar summa kor, summa får respektive summa svin. Om man tittar på datamaterialet horisontellt, så ser man hur summa kor, summa får och summa svin förhåller sig till varandra under det specifika året. Detta är just vad olika punkter i 3D-spridningsdiagrammet i figuren 2 visar.

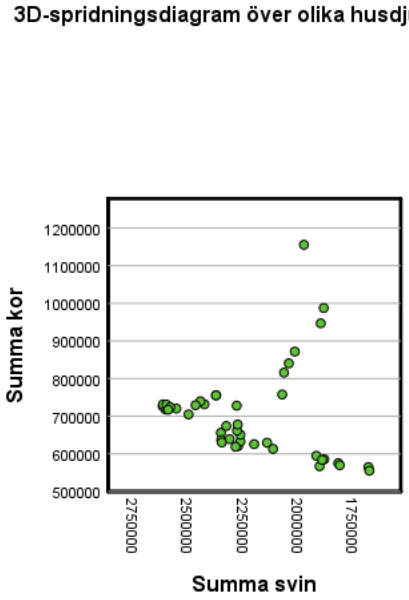
Figur 2: 3D-spridningsdiagram över olika husdjur.



Källa: SCB

I SPSS går det att rotera 3D-spridningsdiagram och vi får därmed flera olika vinklar som mer tydligt visar sambandet mellan olika husdjur i figur 3, 4 och 5.

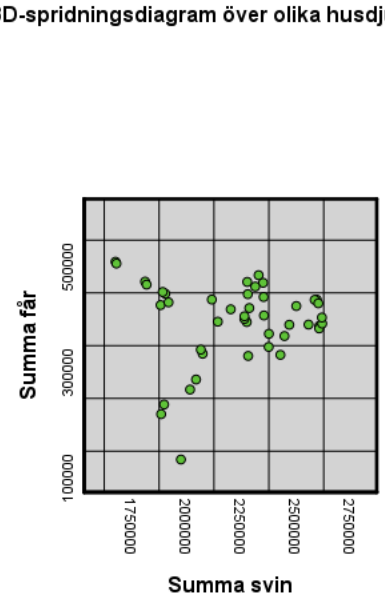
Figur 3: Sambandet mellan summa kor och summa svin utifrån 3D-spridningsdiagram över olika husdjur.



Källa: SCB

I figur 3 visar det summa kor och summa svin. Det ser ut som att det finns ett linjärt samband mellan de två. Med stora värden på summa svin ligger summa kor mellan 700 000 och 800 000. När summa svin avtar, så avtar summa kor också gradvis. Detta tyder alltså på ett positivt samband, där stora värden på summa svin följer med stora värden på summa kor och vice versa. Dock är detta samband ganska svagt. Trots att man kan se en positiv korrelation, så är punkterna generellt ganska utspridda och det finns dessutom flera punkter som helt avviker från det positiva linjära sambandet, även kallad för extremvärden. Slutligen skulle vi säga att det finns ett positivt linjärt samband mellan summa kor och summa svin men sambandet är ganska svagt och har även flera extremvärden.

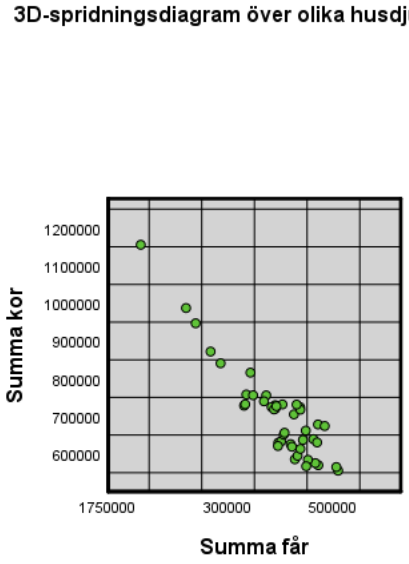
Figur 4: Sambandet mellan summa får och summa svin utifrån 3D-spridningsdiagram över olika husdjur.



Källa: SCB

I figur 4 visar det summa får och summa svin. Det ser ut som att det inte finns ett samband mellan de två. Från vänster, när värden på summa får ligger runt 500 000 och summa svin ligger nära 1 750 000, verkar det vara ett negativt linjärt samband. Med ökande värden på summa svin verkar det som att värden av summa får minskar men om man går vidare till höger är punkterna mycket utspridda, vilket tyder på ett svagt samband. Trots att en negativ korrelation verkar gälla från vänstersidan, så är punkterna generellt utspridda och det finns dessutom flera punkter som helt avviker från det negativa linjära sambandet. Därmed finns det inte något samband mellan summa får och summa kor.

Figur 5: Sambandet mellan summa kor och summa får utifrån 3D-spridningsdiagram över olika husdjur.



Källa: SCB

I figur 5 visar det summa kor och summa svin. Det ser ut som att det finns ett linjärt samband mellan de två variablerna. Med minskande värden på summa får, så minskar summa kor också och när summa får ökar, så ökar summa kor samtidigt. Detta tyder på ett positivt samband, där små värden på den ena variabeln hänger ihop med små värden på den andra och vice versa. Med en tydligt positiv korrelation och att punkterna generellt är inte jätteutspridda, så är detta samband starkt. Det finns inte heller extremvärden. Slutsatsen är att det finns ett starkt positivt linjärt samband mellan summa kor och summa får.

2.3

Tabell 1: Data på utbildningsnivå i Östergötland från 2008-2022. Antalet eftergymnasial utbildning med 3 år eller mer.

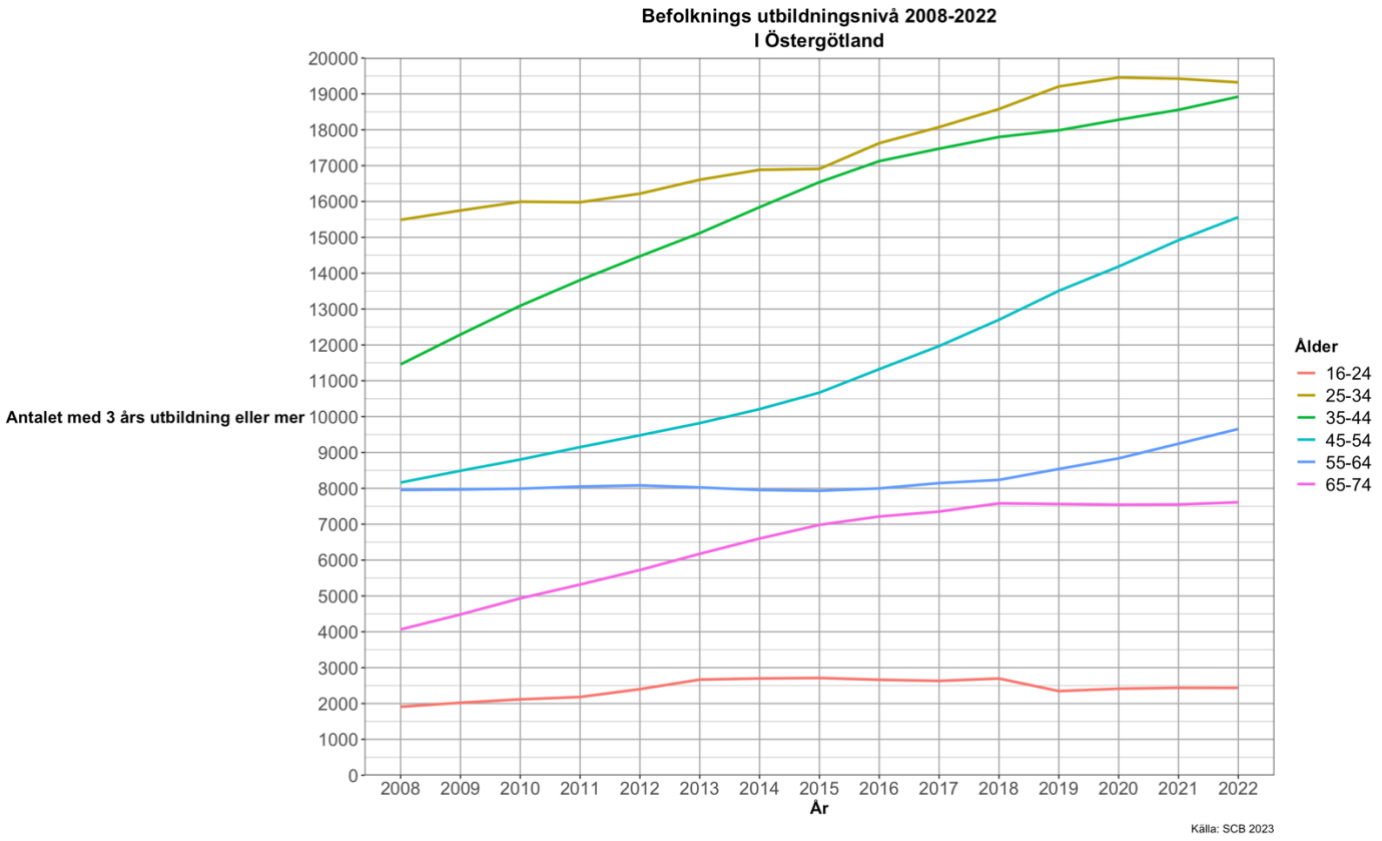
	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
16-24 år	**1910**	**2022**	**2116**	**2182**	**2401**	**2667**	**2699**	**2712**	**2662**	**2631**	**2697**	**2347**	**2413**	**2438**	**2435**
25-34 år	**15488**	**15749**	**15994**	**15978**	**16218**	**16607**	**16883**	**16909**	**17626**	**18073**	**18577**	**19207**	**19457**	**19425**	**19322**
35-44 år	**11457**	**12289**	**13092**	**13808**	**14473**	**15119**	**15841**	**16539**	**17125**	**17472**	**17799**	**17985**	**18279**	**18555**	**18922**
45-54 år	**8162**	**8490**	**8804**	**9150**	**9480**	**9819**	**10210**	**10670**	**11324**	**11968**	**12701**	**13508**	**14186**	**14919**	**15562**
55-64 år	**7954**	**7968**	**7991**	**8050**	**8082**	**8026**	**7954**	**7933**	**7998**	**8147**	**8236**	**8540**	**8836**	**9245**	**9656**
65-74 år	**4066**	**4483**	**4932**	**5318**	**5723**	**6174**	**6601**	**6981**	**7215**	**7352**	**7582**	**7563**	**7544**	**7551**	**7611**

Uppgiftens datamaterial är hämtat från Statistikdatabasen / Utbildning och forskning / Befolkningens utbildning / Befolkningens utbildning / Befolkning 16-74 år efter region, utbildningsnivå, ålder och kön / År 1985 – 2022. Samtliga variabler och rader som står i tabell 1 ovan kommer att visualiseras.

Alla kategorier som till exempel kön och alla årtal som anses vara innan 2008 ingår inte med den slutligen datamaterial som senare visualiseras. Syftet med en minimerade datamaterial är för att det ska anpassa till den senare visualiseringsmetoden. Målet från början var att undersöka populationen från Östergötland från 16-74 år och deras utbildningsnivå. Eftersom det finns många kategorier för vad en utbildningsnivå[1] kan vara, bestämde gruppen att bara undersöka ”Antalet med eftergymnasial utbildning med 3 år eller mer.

[1] Definition av utbildningsnivå enligt SCB ”En utbildningsnivå innehåller såväl utbildningar som avslutats med examen som utbildningar som avslutats utan examen, forskarutbildning undantaget där endast examinerade ingår.” (SCB u.å.)

Figur 6: Befolknings utbildningsnivå från 2008 till 2022 i Östergötland. Antalet med eftergymnasial utbildning med 3 år eller mer.



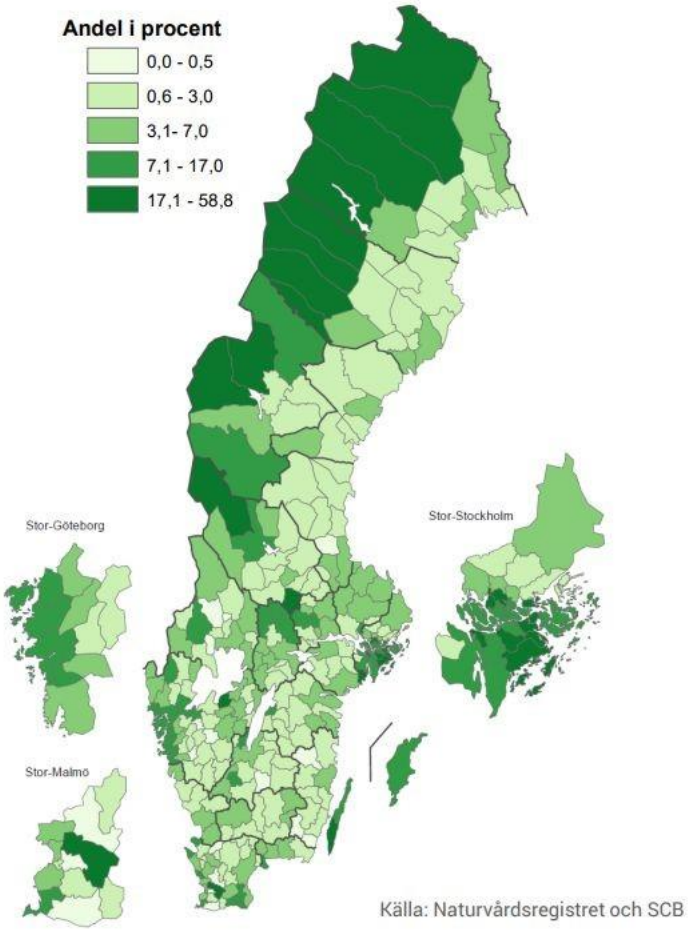
Figur 6 ovan är flera tidsserier spridningsdiagram på befolknings utbildningsnivå från 2008-2022 i Östergötland, uppdelat i 6 åldersgrupper. På y-axel står antalet som i fortsättningsvis efter sin gymnasieexamen pluggar vidare på en utbildning som är 3 år eller mer. Vidare på x-axel står årtalet från 2008 till 2022, denna kategori är valfritt och starttiden är slumputvalda, dvs. 2008. Den första åldersgruppen 16-24 har en svag ökning sedan starten, den största ökningen för denna åldersgrupp är från 2011 till 2013, från och med 2013 har åldersgruppen ett jämnt antal med de som har 3 års utbildning eller mer. Från 2018 till 2019 finns en minskning och det fortsätter med att ha ett jämnt antal fram tills 2022. Vidare har åldersgruppen 25-34 en hög antal redan från början och har ett jämnt ökning varje två år. Den största ökningen för denna grupp är från 2015 till 2019, sedan blir det jämt för antalet person med 3 års utbildning eller mer. Ytterligare åldersgruppen 35-44 har en stor ökning sedan starten i 2008 fram till 2016. Efter 2016 har denna åldersgrupp en mindre årlig ökningen. Åldersgruppen 45-54 ökar årligen men den kraftigare ökningen sker i 2015 och trenden fortsätter till 2022. Sedan 2008 har åldersgruppen 55-64 ett jämnt antal 8000 personer som efter gymnasieexamen har 3 års utbildning eller mer. Fram till mitten av 2016 är då ökningen börjar sker. Stora ökningen för denna åldersgrupp börjar i 2018 och trenden fortsätter till 2022. Den sista åldersgruppen 65-74 har från början av 2008 en ökning fram till 2016 då trenden börjar lugna ner, och från 2018 till mitten av 2021 satte ett stop för denna trend, då antalet ligger jämn på 7000.

2.4

Alternativ 2

Från
https://www.scb.se/contentassets/9335da8f14fa4cb3bce584a756f428bb/mi0603_2022a01_scb-bilder_skyddadnatur.pdf
Vi valde ett kartdiagram som nedan, detta kartdiagram visar andel skyddar landareal per kommun, 2022-12-31 i Sverige. Vi kan se att detta kartdiagram sortera alla områden eller kommun genom färgens ljushet nämligen är gröna, med ljusaste gröna betyder att den område har andel skydda landareal är från 0,0% - 0,5%, nästa är från 0,6% - 3,0% med liten mörkare gröna och så vidare till 3,1% - 7,0%, 7,1% - 17,0% och mörkaste gröna är 17,1% - 58,8%. Vi tycker att detta kartdiagram visualiserade dess data väl bra, vi kan lätta hitta flesta område i Sverige har andel skyddar landareal ligger mellan 0,6% - 3,0% och 3,1% - 7,0% och de område ligger främst i ort med storstäder där finns många människor bor. Med område som har andel skydda landareal större än 7 nämligen från 7,1% - 17,0% och 17,1% - 58,8% ligger främst i norr. Dessutom visar detta kartdiagram också andel skydda landareal i tre olika kommuner är “Stor-Göteborg”, “Stor-Stockholm” och “Stor-Malmö”.
Det finns många olika visualiseringsmetoder att man använder för att uppnå samma syfte som stapeldiagram eller spridningsdiagram, men med dess två metoder finns de både ett problem är variabelvärdet är för många då de inte passa att utföra och de absolut skulle ger en sämre bild av data, använder ett kartdiagram är bäst metod man kan använder för att beskriva data av denna typ.

Figur 7: Kartdiagram beskriva andel skydda landareal per kommun, 2022-22-31 i Sverige.



3. Använda hjälpmedel

Uppgift 1:
R studio

Uppgift 2:
SPSS

Uppgift 3:
SCB (2023)
R Studio

Uppgift 4:
SCB (2022)

4. Lärdomar, problem och övriga kommentarer

I denna laboration har vi tränat att skapa olika sorters diagram via SPSS och R Studio och sedan analyserar dessa. Vi fick även importera valfritt datamaterial från SCB, bearbeta det och sedan använda en lämplig visualiseringsmetod. På uppgift 4 har vi fått välja ett kartdiagram och analyserar detta. Hur man skriver en rapport utifrån all information som vi har kommit fram har vi tränat mer på. Vi fortsätter förbättra vårt samarbete, som alltid delar vi upp uppgifter i olika delar och varje person ansvarar för en del och syntetiserar slutligen all information om uppgifter tillsammans. Generellt sett hade vi inte många svårigheter när vi utförde denna laboration, det var bara lite svårt att kommunicera med varandra när vi inte var i labbet och att analysera när ett samband inte är tydligt.

5. Bilaga

R Studio kod: Uppgift 1

```
#Vi skapa en vektor med etiketter utifrån Tid_etiketter i datamaterial, används funktion seq () för att
skapa en sekvens av heltal. Sekvensen börja med 1 till högst 159 och öka med 10 i varje steg.
etiketter = VIS3_COVID_19_Okt2023$Tid_etiketter [seq(1,159, by = 10)]
etiketter
#Använder paket ggplot2
library(ggplot2)
#Använder data “VIS3_COVID_19_Okt2023” för att skapa en linjediagram med x axel är “Tid_index”
#och y axel är “Antal_fall_per_100 000_inv” och linjestorlek är 1.0.
ggplot(data = VIS3_COVID_19_Okt2023) + aes(x = Tid_index,
                                y = Antal_fall_per_100 000_inv,
                                group = Region,
                                color = Region) +
 geom_line(linewidth = 1.0) +
#Anger etiketter “Tid” och tilldelar sekvens för x-axeln.
 scale_x_continuous(labels = etiketter, breaks = seq(1,159, by = 10)) +
#Anger tema.
 theme_bw() +
#Ändrar egenskaper för text och stödlinjer.
 labs(x = "Tid",
     y = "Antal per 100k",
     title = "Utvecklingen av antalet dödsfall per 100k i olika regioner",
     caption = "Källa: Folkhälsomyndigheten den 10 oktober 2023") +
 theme(plot.title = element_text(face = "bold",
                          size = 13,
                          hjust = 0.5),
       plot.caption = element_text(face = "bold",
                           size = 10),
       axis.title.x = element_text(face = "bold",
                           size = 12),
       axis.title.y = element_text(face = "bold",
                           size = 12,
                           angle = 0,
                           vjust = 0.5),
       panel.grid.major.x = element_line(color = "gray"),
       panel.grid.major.y = element_line(color = "gray")) +
#Anger färger för respektive tidsserie.
 scale_color_manual(values = c("#ea1ea4","#84ea1e","#1ed1ea"))
```

```
library(ggplot2)

#Datamaterial passar bäst med flera tidsserier linjediagram.
#X är årtal. Y är antalet som efter sin gymnasieexamen pluggar vidare dessa utbildningar som har 3 år eller mer. Det finns kön som ett val, men i denna visualisering är vi mer intresserade i åldersgrupper.
ggplot(Utbildning_2) +
  geom_line(aes(x = År, y = tre_år_eller_mer, group = Ålder, color = Ålder),
            size = 1) +
#För att förtydliga siffrorna på y-axel samt slipper de onödiga tal har vi använt kommand nedan.
    scale_y_continuous(breaks = seq(from = 0, to = 20000, by = 1000),
                       limits = c(0, 20000),
                       expand = c(0,0))+

#Här är lika viktig eftersom vi vill att de stödlinjerna, textstorlekar och positioner ska synas bra och står på rätt plats.
 theme_bw() +
  theme(panel.grid.major.y = element_line(color = "darkgray"),
        panel.grid.minor.y = element_line(color = "gray"),
        panel.grid.major.x = element_line(color = "darkgray"),
        panel.grid.minor.x = element_line(color = "gray"),
        axis.title.y = element_text(angle = 0, vjust = 0.5, face = "bold", size = 13),
        axis.title.x = element_text(face = "bold", size = 13),
        axis.title = element_text(size = 15),
        axis.text.y = element_text(size = 14),
        axis.text.x = element_text(size = 14),
        legend.text = element_text(size = 14),
        legend.title = element_text(size = 13, face = "bold"),
        plot.title = element_text(hjust = 0.5, size = 15, face = "bold"),
        plot.subtitle = element_text(hjust = 0.5, size = 15, face = "bold")) +
#Till sist är det bara att justera rubriken, lägga till källa och namnge x och y axel. Något som vi hade inte gjort innan var subtitle, bara för att vi vill vara noggrant med att denna diagram är riktat åt populationen som befinner/befann i just Östergötland.
  labs(title = "Befolknings utbildningsnivå 2008-2022",
       subtitle = "I Östergötland",
       caption = "Källa: SCB 2023",
       y = "Antalet med 3 års utbildning eller mer")
```