# Projekt: Programmering i R

April 16, 2024

Som en del av kursen i R-programmering ska ni göra en rapport i Rmarkdown. Det handlar om att läsa in, presentera och bearbeta data.

- R-markdown ska användas. En mall kan ni hitta här.
- Undvik att använda **å**,**ä** eller **ö** i variabelnamn i er R-kod.
- Spara er fil i UTF-8 kodning. I Rstudio gör ni: "File" → "Save with Encoding" välj UTF-8 och klicka OK.
- Ha en god kodstil och kommentera er kod. Se datorlaboration 4 för detaljer.
- Rapporterna ska lämnas in som både **PDF** och **.Rmd**-fil. Om ni har problem att skapa PDF så går det bra att lämna in som **HTML**-fil. Notera att PDF är att föredra. Det är ok att skapa en HTML som ni sedan sparar/skriver ut som PDF[1]. Filerna ska kallas:
  `[liu id 1]_[liu id 2]_project.pdf`.
  Exempel på inlämning av projekt är följande **två** filer:
  - `joswi71_manma97_project.Rmd` och
  - `joswi71_manma97_project.pdf`.
- Samtliga material ska laddas in i R från webben som **externa datakällor**. Vill ni använda ett eget material får ni lägga upp det öppet på github, dropbox, google docs eller dylikt och läsa in det därifrån i R. Syftet är att rapporten ska vara helt reproducerbar och kunna återskapas på godtycklig dator.
- **Inga output från R console/varningar/meddelanden/felmeddelanden ska visas i dokumentet.** Antingen skapar ni tabeller (med `kable()`) eller grafer. T.ex. kan ni ange message=FALSE, warning=FALSE i chunk options när ni skapar chunks med R-kod.
- Ni ska använda ggplot2 för alla plottar.

[1] Detta går att göra i de flesta webbläsare, välj skriv ut och sen skriv till pdf.

- **Rmd**-filen ska kunna köras och reproducera era resultat på godtycklig dator. D.v.s. den ska innehålla all er kod som behövs för att ladda ner data, era beräkningar och er rapporttext.
- **Namn**, **liu-id** och **gruppnummer** ska framgå i början av rapporten.
- **Tänk på att kommentera er kod och ha god kodstil!**

# 1 Projektinstruktioner

Till projektet behöver två olika typer av datamaterial, ett material med kommunala data och ett material som innehåller en tidsserie. Det är okej att välja data på lännivå eller regionnivå istället för kommunnivå om ni vill. Beskrivningen nedan utgår från kommunala data. I projekt ska ni använda pxweb[2], i filen ska ni ha kod som fungerar och ger reproducerbara rapporter. Om ni vill dölja varningar kan ni använda

```
suppressWarnings({
        # min kod här
})
```

Tänk på att välja material ni själva tycker är intressant!

**Kommunala data** Ni ska ladda ner kommunala data, där ni i slutändan har minst 4 variabler på kommunnivå (d.s.v. för alla 290 kommuner) Ett exempel skulle kunna vara antal arbetslösa i varje kommun. Spara er data i en eller flera data.frames. Totalt ska dataseten ska ha **minst 4 variabler** utöver kommunnamn. Ni väljer själv vilka variabler som ska ingå och vilka områden data ska komma ifrån. Tanken är att i ska göra enklare analyser och grafer som baseras på dessa variabler. Utöver dessa 4 variabler så ska ni också ladda ner totalt antal invånare i kommun som en variabel.

När ni har valt ut era variabler ska ni ha en data.frame där **varje rad motsvarar en kommun** och där det finns minst 5 kolumner med variabler. Kommuner är alltså **observationer** i era analyser. Kolumnerna motsvarar era variabler. Notera att många av de variabler som finns på SCB:s databas är frekvenser, exempel: antal arbetslösa i varje kommun. Tabell 1 visar ett exempel med hur data ska vara strukturerat.

**Tidsseriedata** Hitta ett dataset som innehåller en **tidserie**, det innebär att det finns en variabel som har observerats över tiden. Kravet är att data ska innehålla data på **månadsnivå** och innehålla data från **minst 10 år** (120 månder). Här ska ni alltså hitta en variabel som observerats under minst 120 tidpunkter, men fler går bra. Data ska alltså innehålla två kolumner, en med

[2]Se här för mer info.

| Radnamn | Variabel 1 | Variabel 2 | Variabel 3 | Variabel 4 | Totalt antal invånare |
|---|---|---|---|---|---|
| Linköping | | | | | |
| Norrköping | | | | | |
| Mjölby | | | | | |
| Motala | | | | | |
| ⋮ | | | | | |

Table 1:

variabeln som vi är intresserade av och en med tidpunkterna. Här finns en lista över några olika tidserier som används tidigare år.

**Obs!** Tidsperioden ska vara fix, d.v.s ex. jan 2005 - jan 2015. Detta innebär att ni måste ange ett fixt tidsintervall när ni laddar ner data med `pxweb`. Om ni laddar ner data en månad senare ska ni erhålla samma data med samma kod. Om ni laddar ner data från SCB/pxweb så ska ni **inte** ange "`*`" på tiden.

## 1.1 Beskrivning av data

### 1.1.1 Dataanalys av kommundata

- Skriv en kort inledning där ni beskriver era variabler. Om ni gör nya transformationer av variablerna i del 2 måste de beskrivas också.
- Vissa upp data för 5 kommuner och alla era variabler i en tabell. Ta inte med fler kommuner. Ni väljer själv vilka kommuner ni visar i tabellen.

Följande saker ska ni göra/ta med:

- Alla variabler som är relaterade till folkmängd på något sätt ska normaliseras med hjälp av totalt antal invånare i varje kommun. Detta eftersom det oftast är intressant att kolla på andelar istället för absoluta antal. T.ex. andelen arbetslösa i en kommun istället för antalet arbetslösa. I de fall då det är relevant att normalisera en variabel, då ska ni använda den normaliserade variabeln i plottar mm. Vissa variabler är inte relaterade till folkmängd, exempel "Antal höns" eller "Medelålder", sådana variabler behöver inte normaliseras.
  - Exempel: antalet arbetslösa/totalt antal invånare = andelen arbetslösa, gör denna beräkning för varje kommun. För många variabler som har små andeler så passar det att skapa variabler av typen "antal per 10 000/100 000 invånare". Då räknar vi ut det som: (antal/totalt antal invånare)*10 000.

1. Producera minst en barplot, om ni bara har kontinuerliga variabler kan ni använda `cut()`. Beskriv i text vad ni drar för slutsats. Notera! Gör inte en plot där varje kommun har en egen stapel, alltså ingen barplot med 290 staplar.

2. Producera minst ett histogram. Lägg till vertikala linjer för följande punkter på x-axeln: medianen, första kvartilen och tredje kvartilen. Beskriv i text vad ni drar för slutsats. Tips: `geom_vline()` och här.

3. Producera minst en scatterplot mellan två variabler. Lägg till en regressionslinje med `stat_smooth(method="lm", se=FALSE)`. Beskriv i text vad ni drar för slutsats[3].

4. Beräkna korrelationer mellan de två variabler som ni använde i scatter plot i steget ovan. Gör ett hypotestest där ni testar om korrelationen mellan dessa två variabler är noll (=de är linjärt oberoende). Ni ska alltså använda hypoteserna:

$$H_0 : cor\,(x_1, x_2) = 0$$
$$H_a : cor\,(x_1, x_2) \neq 0$$

   Tips: `cor.test()`. Presentera relevant information i en eller flera tabeller. Beskriv kort hur ni tolkar resultatet. Ni ska alltså presentera både den skattade korrelationen och information från testet i rapporten. Ni får testa korrelationen mellan fler variabler om ni vill. Presentera även ett konfidensintervall för den skattade korrelationen.

5. Skapa två kategoriska variabler utifrån era ursprungliga variabler. Om ni redan har kategoriska variabler för era kommuner så kan ni använda dessa. Dessa (nya) variabler ska användas i olika plottar. Se till att beskriva era nya variabler.

6. Mer plottar:

   (a) Gör en scatterplot där färgen på observationerna ska bero på en av era kategoriska variabler. Beskriv i text vad ni drar för slutsats.

   (b) Gör minst ett histogram/boxplot som är grupperat på en av era kategoriska variabler. Beskriv i text vad ni drar för slutsats.

   (c) Gör ett grupperad barplot där ni använder båda era kategoriska variabler. Beskriv i text vad ni drar för slutsats.

   (d) Gör minst en scatterplot/histogram/barplot/boxplot som är uppdelad i minst två plottar med `facet_grid()` eller `facet_wrap()` baserat på era kategoriska variabler. Beskriv i text vad ni drar för slutsats.

### 1.1.2 Dataanalys av tidseriedata

Låt `Y`[4] vara er variabel i tidsseriematerialet. Utför nu följande:

1. Gör en linjeplot mellan `Y` och er tidsvariabel. Skalan på x-axeln ska vara en lämplig tidsskala. Detta gäller alla tidseriegrafer som baseras på `Y`. Skriv en kort kommentar.

[3] För tips på tolkning: Se kap 13.2.5 i kursboken, speciellt figur 13-5. Se även här för tips på hur scatter plots kan tolkas.

[4] `Y` är ett arbetsnamn, er variabel måst inte heta så i rapporten

2. Beräkna medelvärden per månad och spara dessa i `month_means`. Presentera dessa i en tabell och som en lämplig graf. Skriv en kort kommentar. Ni ska alltså beräkna ett medelvärde för alla värden för januari och sen upprepa detta för alla månader. **Tips!** `aggregate()`

3. Gör grupperade boxplots för `Y`, där grupperingen baseras på år. Ni ska alltså ha en boxplot för varje år, och de ska ligga sida vid sida i samma plot. Skriv en kort kommentar.

4. Subtrahera månadsmedelvärden från `Y`, så ni tar bort säsongsvariationen i data. Månadsmedelvärdet för januari ska subtraheras från alla januarivärden i data, och likadant för de andra månaderna. Spara den nya tidserie som `Z`. Addera medelvärdet för *hela* tidserien `Y` till `Z` för att ge `Z` rätt skala. Se nedan.

```
Z<-Z+mean(Y)
```

5. Gör en linjeplot mellan `Z` och tid i `ggplot2`. Lägg också till `Y` i samma graf som jämförelse, men med annan färg. Ange tydligt med en legend eller i text vilken linje som är `Y` och `Z`.

6. Gör samma plot som i steg 1., men lägg till en regressionslinje med `geom_smooth(method=''lm'')`[5].

7. Verkar det finnas någon trend i data? Dvs ökar/minskar data med tiden, eller är data konstant över tid. Finns det någon säsongsvariation i data?[6] Dra er slutsats och skriv ned den i dokumentet.

[5] En regressionslinje kan hjälpa oss att se om det finns någon tydlig minskande eller ökande *linjär* trend. Om linjen är nära att vara horisontell så tolkar vi det som att det inte finns någon tydlig linjär trend.

[6] Exempel på säsongsvariation: December har ofta ett mycket högre värde än övriga månader, sommarhalvåret har alltid lägre värden.

Lämna in rapporten både som en fullt reproducerbar **Rmd**-fil och som **PDF/HTML** i LISAM. Tänk på följande:

- I denna del ska samtliga grafer vara skapade med `ggplot2`.
- Tabeller ska vara "riktiga" tabeller (med ex. `kable()`), inte utskrifter i R-kod. All statistik, information från statistiska test, korrelationer etc ska presenteras i markdowntabeller eller med inline-kod. Avrunda till ett lämpligt antal decimaler i tabellerna.
- **Inga output från R console/varningar/meddelanden/felmeddelanden ska visas i dokumentet.**