Projekt: Programmering i R

March 1, 2023

Som en del av kursen i R-programmering ska ni göra en rapport i Rmarkdown. Projektet är uppdelat i två delar. Den första delen handlar om att läsa in och bearbeta data från externa datakällor och beskriva dessa data.

I den andra delen av projektet ska mer utförlig analys genomföras samt bearbeta och analysera denna data vidare.

För båda delarna gäller att:

- 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]_part[del av projektet]_miniproject.pdf`.
  Exempel på inlämning av projekt del 1 är följande **två** filer:
  - `joswi71_manma97_part1_miniproject.Rmd` och
  - `joswi71_manma97_part1_miniproject.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.

[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 Del I: Deskriptiv analys

Den första delen av projektet är att samla in datamaterial och beskriva materialet kortfattat i en första del av rapporten.

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 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 månader). 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 Inlämning av del I

Den första inlämningsuppgiften handlar om att läsa in i R och beskriva de material ni valt med R-markdown. Notera att kommuner ska vara observationer i ert datamaterial med kommunala data, alltså en kommun motsvarar en rad i era data.frames/matriser med data.

Ni ska beskriva era material i text samt sammanfatta de variabler ni valt med de beskrivande statistiska mått. Ta fram beskrivande statistik för **alla** variabler i data. Beroende på hur data ser ut så kan det vara medelvärden, medianer frekvenstabeller mm. Ni kan göra relevanta transformationer av era variabler om ni vill, tex göra en numeriska variabel till en binär och räkna med andelar eller dela in kommunerna i stora, medelstora och små när det gäller befolkning. Men se till att ha minst två variabler som är numeriska/frekvenser.

Rapporten ska var **ordnad** och **strukturerad**, med lämpliga rubriker. Tänk på:

- Tabeller ska vara "riktiga" tabeller (med ex. `kable()`), inte utskrifter i R-kod. Avrunda till ett lämpligt antal decimaler i tabellerna.
- Se till att plottarna ni har ser snygga ut och har lämpliga axeltexter mm.
- **Inga output från R console/varningar/meddelanden/felmeddelanden ska visas i dokumentet.**

Ni ska ha med:

- Kort inledning
  - Beskrivning av alla era variabler och eventuella transformationer av dessa.

	Variabel 1	Variabel 2	Variabel 3	Variabel 4	Totalt antal invånare
Medelvärde					
Median					
Standardavvikelse					
Minimum					
Maximum					

Table 2:

  - 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.
  - Vissa upp data för **5** kommuner och alla era variabler i en tabell. Ta inte med fler kommuner i tabellen, alla kommuner ska självklart vara med i datasetet när ni gör er statistik nedan.

- Beskrivande statistik av alla variabler och en kort tolkning av statistiken i en eller flera tabeller. Ta inte med all rådata i en tabell.

  - Ni ska ha åtminstone ha med: medelvärde, median, standardavvikelse, minimum och maximum. Om någon variabel är kategorisk så kan ni ha en frekvenstabell för den variabeln istället. Så ni ska ha med en tabell likt den som finns i tabell 2.

- De plottar som beskrivs nedan.

- Skriv en kort kommentar/tolkning till alla era tabeller och plottar.

Följande saker ska ni göra med data med basgrafiken i R:

1. Ni ska minst ha ett histogram, barplot eller boxplot per variabel i kommundata, ni måste inte göra det för "Totalt antal invånare". Skriv en kort kommentar till varje plot. Gör inte en barplot där varje kommun har en egen stapel, alltså ingen barplot med 290 staplar. Har ni variabler som är kontinuerliga eller frekvenser, då passar histogram eller boxplot bra.

2. Minst en scatter plot mellan två av era variabler. Skriv en kort kommentar[3].

3. En tidsseriegraf/linjediagram för tidseriedata. Skriv en kort kommentar. Se till att det är lämplig tidskala på x-axeln (se "Linjediagram" i datorlaboration 5 för tips).

Ibland innehåller data några väldigt stora eller väldigt små värden (kallas outliers), och dessa kan göra att det är svårt att göra snygga plottar. Det är ok att ta bort några sådana värden i era plottar för kommundata, men skriv då att ni har gjort det, varför det var nödvändigt och vilka kommuner som ni tog bort.

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

- I denna del ska samtliga grafer vara skapade med basgrafiken i R.
- Tabeller ska vara "riktiga" tabeller (med ex. `kable()` i paketet `knitr`), inte utskrifter av R-kod. Avrunda till ett lämpligt antal decimaler i tabellerna.
- **Inga output från R console/varningar/meddelanden/felmeddelanden ska visas i dokumentet.**

[3]Se här för tips på hur scatter plots kan tolkas.

2 Del II: Analys (ej uppdaterad)

I den första delen av minprojektet har ni valt ut och beskrivit olika variabler. Nu ska vi fortsätta detta arbete med analyser av materialen. Ni som grupp kommer att ha en del frihet i hur ni utför datanalysen som beskrivs nedan. Det ni ska göra är att bearbeta data, några enkla analyser och olika grafer i `ggplot2`. **Obs!!! alla plottar i del 2 ska vara med `ggplot2`.**

2.1 Inlämning del II

Den fulla rapporten ska lämnas in som en fullt reproducerbar **Rmd**-fil och som ett **PDF**-dokument i LISAM. Nedan framgår exakt vilka analyser som ska genomföras. Ta **inte** med del 1 av projektet i Rmd-filen och i PDF för del 2.

- Rapporten ska var ordnad och strukturerad, med lämpliga rubriker.
- 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.**

2.1.1 Dataanalys av kommundata

- Skriv en kort inledning där ni beskriver era variabler. Kan vara samma som i del 1. Om ni gör nya transformationer av variablerna i del 2 måste de beskrivas också.
- Ta med från del 1: 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:

1. Alla variabler som är relaterade till folkmängd på något sätt ska normaliseras med hjälp av totalt antal invånare i varje kommun/län. 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. Alla plottar ska använda de normaliserade variablerna. I uppgift 5 och 6 får ni välja om ni vill ha de normaliserade eller ej normaliserade variablerna.
2. 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.

3. 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.

4. 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[4].

5. Beräkna korrelationer mellan de två variabler som ni använde i scatter plott 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.

6. Hypotestest: Gör minst ett hypotestest (ej korrelationstest här), där ni ställer upp en nollhypotes och sen testar om ni kan förkasta den. Ni väljer själva vilken nollhypotes ni vill anävnda. Beroende på hur er data ser ut så kan det vara ett t-test, ett χ^2-test eller test av andelar[5]. Har ni inte några kategoriska variabler kan ni använda funktionen `cut()`. Ni får själva välja vilken nollhypotes ni vill testa. Presentera relevant information från testet/testen i en eller flera tabeller. Beskriv kort hur ni tolkar resultatet. *Exempel:* Ni har variabeln medelålder i kommunerna. Ni vill testa om medelvärdet för medelåldern är signifikant skilld från 40 år. Låt μ vara medelvärdet för medelåldern. Ni sätter då upp hypoteserna

$$H_0 : \mu - 40 = 0$$
$$H_a : \mu - 40 \neq 0$$

   och testar sedan om ni kan förkasta H_0(nollhypotesen).

7. Skapa en kategorisk variabel baserat på totalt antal invånare: Utgå från medianen, och låt alla kommuner/län som är mindre än (eller lika med) medianen vara en grupp och låt låt alla kommuner/län som är större än medianen vara en grupp. Kalla denna kategoriska variabel för `pop_grupp`. Detta ger er två grupper av observationer.

8. Mer plottar. Ni ska nu följande plottar som beror på variabeln `pop_grupp`.

[4]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.

[5]Se kurshemsidan för referenser på hur olika test kan göras i R.

(a) Gör en scatterplot där färgen på observationerna ska bero på variabeln `pop_grupp`. T.ex. om ni gör en scatterplot så har alla punkterna olika färger beroende på vilken grupp de tillhör. Beskriv i text vad ni drar för slutsats.

(b) Gör minst ett histogram/barplot/boxplot som är grupperat på `pop_grupp`, där grupperna har olika färger. Beskriv i text vad ni drar för slutsats.

(c) Gör minst en scatterplot/histogram/barplot/boxplot som är uppdelad i två plottar med `facet_grid()` eller `facet_wrap()`. Uppdelningen ska bero på variabeln `pop_grupp`. Beskriv i text vad ni drar för slutsats.

2.1.2 Dataanalys av tidseriedata

Låt `Y` 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.

2. Beräkna medelvärden per månad och spara dessa i `month_means`. Presentera dessa i en tabell och 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. Använd funktionen `summary()` för att fram beskrivande statistik för varje år (det ska vara minst tio år i data). Presentera statistiken i en tabell och 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 `new_Y`. Addera medelvärdet för *hela* tidserien `Y` till `new_Y` för att ge `new_Y` rätt skala. Se nedan.

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

5. Gör en linjeplot mellan `new_Y` och tid i `ggplot2`. Lägg också till `Y` i samma graf som jämförelse.

6. Ni ska nu beräkna glindande medelvärde av den typ som ni gjorde i labb 3. Använd er funktion `my_moving_average()` från tidigare labb, eller annan likvärdig funktion i ett R-paket och beräkna `moving_average_Y`. Lägg till variabel i samma graf som ovan. Totalt ska grafen ha tre linjer i olika färger. Det ska framgå i en legend eller i texten vilken färg som är vilken linje. Det ska tydligt framgå i vilket värde på n (längden på det glidande medelvärdet) som ni använder. Notera att `moving_average_Y` kommer att vara $n-1$ element kortare jämfört med `Y` och `new_Y`, och ska således börja lite längre fram på x-axeln.

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.

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.**

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