Miniprojekt: Programmering i R

March 2, 2020

Som en del av kursen i R-programmering ska ni göra en rapport i Rmarkdown. Miniprojektet 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 miniprojektet 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.
- 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. Filerna ska kallas:
  `[liu id 1]_[liu id 2]_part[del av miniprojektet]_miniproject.pdf`.
  Exempel på inlämning av miniprojekt 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.
- **Rmd**-filen ska kunna köras och reproducera era resultat. D.v.s. den ska innehålla all er kod som behövs för analysen.
- **Namn**, **liu-id** och **gruppnummer** ska framgå i början av rapporten.
- Tänk på att kommentera er kod!

1 Del I: Deskriptiv analys

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

Till miniprojektet 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 kommun om ni vill. Beskrivningen nedan utgår från kommunala data. I miniprojekt får ni använda det gamla eller det nya[1] gränssnittet till `pxweb`, så länge ni har kod som fungerar och ni lyckas skapa reproducerbara rapporter. Om ni vill dölja varningar kan ni använda

```
suppressWarnings(
        mitt_data <- get_pxweb_data(url = "en sökväg",
        dims = list("lista med lämpliga element"),clean = TRUE)
)
```

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. Ni rekommenderas att välja totalt antal invånare i kommun som en variabel, då denna kommer att användas i del 2 av projektet.

Tidsserie 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 variabeln som vi är intresserade av och en med tidpunkterna.

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. Ni ska beskriva era material i text samt sammanfatta de variabler ni valt med de beskrivande statistiska mått som ni själva finner lämpliga. Ta fram beskrivande statistik för **alla** variabler i data.

[1] Se här för mer info.

Beroende på hur data ser ut så kan det vara medelvärden, 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.

Rapporten ska var ordnad och strukturerad, med lämpliga rubriker. Ni ska alltså ha med:

- Kort inledning
- Beskrivning av alla era variabler och eventuella transformationer av dessa.
- Beskrivande statistik av alla variabler och en kort tolkning av statistiken. Ni ska ha med minst en "riktig" markdowntabell (inte bara R output i konsolen) med statistik. Ta inte med rådata i en tabell. (**Tips!** `kable()` i paketet `knitr`)
- De plottar som beskrivs nedan.

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

1. Ni ska minst ha ett histogram eller barplot per variabel i kommun-materialen. Skriv en kort kommentar till varje plot.
2. En tidsseriegraf/linjediagram för tidseriematerialet. Skriv en kort kommentar till grafen.

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.
- **Inga output från R console/varningar/meddelanden/felmeddelanden ska visas i dokumentet.**

2 Del II: Analys

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`.

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

Slå samman de era dataset med kommundata så det blir ett dataset som innehåller variablerna från alla dataset. Om ni gör rätt här så ska ni få ett dataset med en variabel över kommun och minst 4 andra variabler. Detta kan göras på olika sätt, ett är att använda funktionen `merge()`. [Här] finns en video för hur ni kan använda `merge()`.

- 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.

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 scatterplot mellan två variabler. Beskriv i text vad ni drar för slutsats.

3. Producera minst ett histogram. Beskriv i text vad ni drar för slutsats.

4. Producera minst en barplot, om ni bara har kontinuerliga variabler kan ni använda `cut()`. Beskriv i text vad ni drar för slutsats.

5. Hypotestest:

   (a) Gör minst ett hypotestest, 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[2]. 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).

   (b) 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. Gör nu minst ett statistiskt test där ni jämför de två grupperna. Ni ska göra testet på någon annan variabel än totalt antal invånare. Ett exempel kan vara att göra ett two-sample t-test där ni testar om medelvärdet för en variabel är olika mellan grupperna. Presentera relevant information från testet/testen i en eller flera tabeller. Beskriv kort hur ni tolkar resultatet.
   Exempel: Ni har variabeln energiförbrukning per invånare i kommunerna. Ni testar då om medelvärdet över energiförbrukning skiljer sig mellan de små och stora kommunerna. Detta kan göras med ett two-sample t-test. Låt $\mu_{små}$ vara medelvärdet över de små kommunerna och låt μ_{stor} vara medelvärdet över de stora kommunerna. Ni sätter då upp hypoteserna

   $$H_0 : \mu_{små} - \mu_{små} = 0$$
   $$H_a : \mu_{små} - \mu_{små} \neq 0$$

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

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

6. Beräkna korrelationer mellan minst två variabler (fler går bra). Gör minst ett hypotestest där ni testar om korrelationen mellan två variabler är noll (=de är linjärt oberoende). Tips: `cor.test()`. Presentera relevant information i en eller flera tabeller. Beskriv kort hur ni tolkar resultatet.

7. Mer plottar. Ni ska nu skapa två till plottar som beror på variabeln `pop_grupp`.

   (a) Gör en scatterplot/histogram/barplot 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 en scatterplot/histogram/barplot som är uppdelad i två plottar med `facet_grid()`. 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?[3] Dra er slutsats och skriv ned den i dokumentet.

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 `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.**

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