Inlämningsuppgift R programmering och statistik

November 2021

Generell information

Inlämning sker i form utav zippade dokument och R-filer. Godkänt-delen är en textfil av valfritt format tex word, pdf. För VG-delen är det R-filer för varje deluppgift. Lämnas in på PingPong. Deadline 2021-12-17 kl 23:55. Den zippade mappen namnger ni med namn, R och betygsönske, tex *Eva_Hegnar_R_VG.zip*.

Kom ihåg att ange källor i godkänt-delen! Inlämningen är **individuell**.

Betygkriterier

G

- På ett grundläggande sätt kunna redogöra för begrepp inom statistik används inom Data Science
- På ett grundläggande sätt kunna förklara hur statistik används inom Data Science
- På ett grundläggande sätt kunna förklara termer och funktioner inom R

VG

- På ett självständigt sätt kunna producera välskriven kod i språket R
- På ett fördjupat sätt kunna tillämpa statistiska beräkningar i R
- På ett självständigt sätt kunna använda effektiv dataanalys med R

Uppgifter

G

Svara på följande teori-uppgifter för att få godkänt på inlämningen:

1. Varför är det viktigt att städa datan (tidy data)? Nämn två funktioner man kan använda för att städa data och hur de fungerar.
2. Vad är en Dataframe i R och vad är fördelarna med att använda det när man ska jobba med data?
3. Var är funktioner i R och vad är syftet med att skapa de?
4. Vad är skillnaden mellan en for-loop och en while-loop?
5. Ge ett exempel där det passar att använda en if-statement.
6. Inom linjär regression kan metricen R^2 användas. Vad innebär denna?
7. Vad är statistisk signifikans och hur används det?
8. En population är normalfördelad. Vad innebär detta?
9. Vad är ett konfidensintervall och hur används det?
10. Vad innebär det att två variabler är korrelerade?
11. Vad är outliers och hur kan man hantera dem?

VG

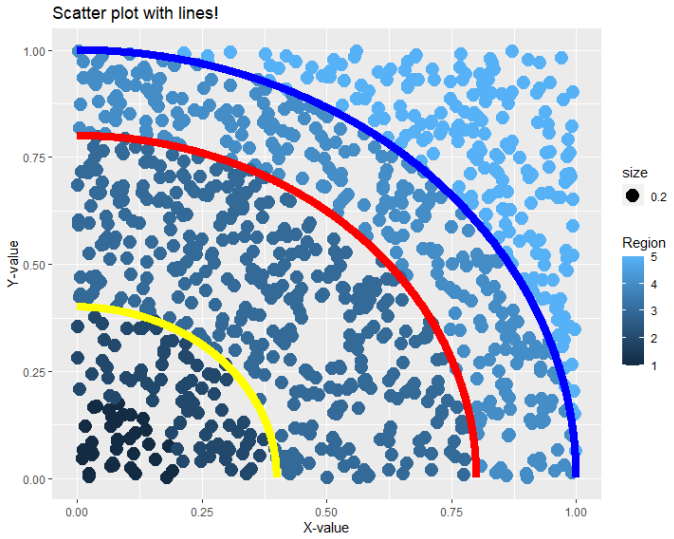
Följande uppgifter bygger på uppgifterna från grupparbetet. Gör båda uppgifterna, i tillägg till frågorna för godkänt, för att uppnå VG.

Approximera pi med sampling

Inkluderar inga filer

Genom att sampla värden för en x- och y-koordinat mellan 0 och 1 har ni gjort en approximation av π.

Ni har rita upp cirkelbågar för avstånden 0.4, 0.8 och 1.0 ifrån origo. Ge nu **punkterna** i dessa intervall olika färger beroende på vilken region de tillhör.



Figur 1: Visuellt exempel. Färger valfritt. Vill du experimentera med themes får du förstås göra det också!

Hitta mördaren

Inkluderar given data *telemastdata.csv* och R-filera *hittaMordaren.r* och *triangulation.r*

Ett hemskt mord har begåtts! Data scientisten Batman Batmansson har hittats brutalt mördad och polisen står handfallen. Du har redan hittat top 6 misstänkta enligt följande information:

- Vi vet inom vilket tidsspann mordet inträffade. Det finns i kolumnen *time0* och ägde rum vid tidspunkt $time0 = 416 \pm 9$ minuter.
- Vi känner till platsen för mordet.
- Ett vittne såg att mördaren talade i en iPhone.
- Vi har data från telemaster som loggar typ av telefon och tidpunkt för sändning och mottagning av telefonsignaler.

Som sista handling innan Batman Batmansson bet i gräset lämnade han lyckligtvis efter sig en trianguleringsalgoritm som han lade i filen *triangulation.r*. Använd de givna funktionerna i triangulation.r för att med hjälp av datan från telemasterna begränsa antalet misstänkta till en enda person! Någon ska alltså bli burad för detta illdåd.

Tips! Att använda mutate här kan vara svårt, prova istället att plocka ut datan du vill ha från dataframen och använda cbind() samt circle_intersection().

Tips 2! Tänk på att datan i telemasterna är av enheten tid, att du letar efter en distans och att du har en given hastighet i R-filen.