Inlämningsuppgifter R programmering och statistik

Eva Hegnar

November 2021

Generell information

Inlämning sker i form utav zippade R-filer, en R-fil för varje deluppgift, på PingPong. Deadline 2021-12-10 kl 23:55. Den zippade mappen namnger ni med namn på båda i gruppen och R. Ni jobbar i grupp om två och två efter indelning från Eva.

Färdigställ koden så att Eva kan testköra filerna och få ut rätt svar. Misslyckas ni med detta får ni **omedelbar retur** (inte synonymt med underkänt, men tidsödslande för alla), så var säkra på att ni ger riktiga svar.

Stort lycka till och kämpa på!

1 Tidy data

Inkluderar given data *music.csv*
I den här uppgiften befinner vi oss i ett parallellt universum där olika typer utav metal-musik faktiskt uppskattas. Givet är data för ett par schyssta band - tyvärr är datan lite rörig.

	artist	ranks	title	post_metal_rank	black_metal_rank	death_metal_rank	length
1	Alcest	207	L'île des morts	22	*NA*	*NA*	7:49
2	Alcest	204	Les jardins de minuit	23	*NA*	*NA*	5:46
3	Dödsrit	212	Ändlösa ådror	*NA*	3	*NA*	10:32
4	Dödsrit	213	A drowning voice	*NA*	2	*NA*	9:15
5	Dödsrit	205	Aura	*NA*	7	*NA*	8:17
6	Aktiv Dödshjälp	208	Helrutten värld	*NA*	*NA*	176	3:13
7	Aktiv Dödshjälp	211	Stressad, rädd och förbannad	*NA*	*NA*	120	13:37
8	Aktiv Dödshjälp	202	Hatets legionärer	*NA*	*NA*	32	4:15
9	Ulcerate	206	Stare into death and be still	*NA*	*NA*	3	7:32
10	Ulcerate	210	Drawn into the next void	*NA*	*NA*	2	6:46
11	Ulcerate	209	Exhale the ash	*NA*	*NA*	1	6:32
12	Mgla	201	Exercises in futility IV	*NA*	1	*NA*	6:13
13	Mgla	203	With hearts towards none IV	*NA*	8	*NA*	5:22

Figur 1: Givet dataset med ranks och ranking för respektive genre. Trots att Kind av Plini är den bästa låten i världen finns den inte i datasetet.

Uppgift:

- Flytta så att *ranks* blir den första kolonnen följt av allt annat.
- För att få en kolumn utan NA-värden slå samman *post_metal_rank*, *black_metal_rank* och *death_metal_rank* till en ny kolumn, *genre_rankings*. Ta alltså inte bort kolonnerna, bara skapa en ny kolonn.
- Dela upp kolonnen *length* till *minutes* och *seconds*.

- Gruppera sedan på *artist* och svara på vilken artist som har högst genomsnittsrank från kolonnen *ranks* för sina låtar.
- Döp om *genre_rankings* till *genres*. Byt sen namn på varje kolumn från *x_y_rank* till *x y rank* (t.ex. så att *black_metal_rank* blir till *black metal rank*.
- Ni ska få ut artist, mean och diff i en artistgruppering genom att använda group_by och summary. Diff är en funktion ni får skapa själv som tar in *ranks* för en artist och ger tillbaks skillnaden mellan högsta och minsta rank ifrån *ranks*.

2 Approximera pi med sampling

Inkluderar inga filer
Genom att sampla värden för en x- och y-koordinat mellan 0 och 1 kommer vi att göra en approximation av π!

Uppgift:

Ta reda på hur ni samplar från en uniform distribution mellan 0 och 1 (dvs vilket decimaltal som helst mellan 0 och 1 kan dras med samma sannolikhet).

Ni kan behöva två variabler:

- *points_within_circle*
- *points_outside_circle*

Slumpa punkter med en x- och en y-koordinat. För varje punkt, avgör ifall dess avstånd till origo, [0,0], är mindre än 1. Om det är mindre än ett ökar ni på *points_within_circle*, är det utanför ökar ni på *points_outside_circle*. Kriteriet ni letar efter är alltså:

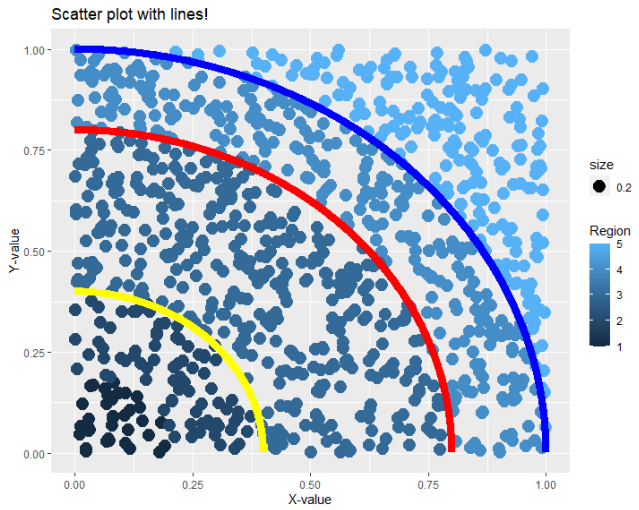
$$\sqrt{x^2 + y^2} \leq 1 \tag{1}$$

Gör detta för $nr\_of\_iterations = 10, 100, 1000$ och 10000 samplade punkter. Utred uttrycket

$$\pi - 4 * \frac{points\_within\_circle}{nr\_of\_iterations} \tag{2}$$

för samtliga antal mätpunkter. Säger det dig något?

Gör efter detta en scatter plot av era samplade värden.
Rita även upp cirkelbågar för avstånden 0.4, 0.8 och 1.0 ifrån origo, [0,0].



Figur 2: Visuellt exempel. Vill ni experimentera med themes och färger får ni förstås göra det också!

3 Hitta mördaren

Inkluderar given data *telemastdata.csv*
Ett hemskt mord har begåtts! Data scientisten Batman Batmansson har hittats brutalt mördad och polisen står handfallen. Som tur är finns följande information att tillgå:

- 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.

Uppgift:

Ge polisen max 6 huvudmisstänkta som de behöver utreda givet informationen.

4 Regression i R

Inkluderar given data *LungCap.csv*
Studien om lungkapacitet fortsätter! Det har tillkommit nya mått och mätningar. Använd den linjära regression vi gått igenom och lyft fram de 3 viktigaste parametrarna för att avgöra lungkapacitet. Glöm inte att inspektera datan - en förklaringsgrad R^2 under 0.75 är **inte** godtagbart!

Hint: Ni kan mycket väl tvingas att fatta ett beslut angående vissa mätvärden i datasetet. Motivera kort hur ni hanterar det och varför det är godtagbart.

Uppgift:

Använd den linjära regression vi gått igenom för att lyfta fram de 3 parametrar som har bäst signifikans för modellen och nå en förklaringsgrad R^2 över 0.75.