UDVIKLINGS OG FORENKLINGS STYRELSEN

Notat

26. april 2019

Programkontor 1, ICI

Praktisk opgave til nye kandidater

- Cold case: hvem overlevede Titanic?

Denne opgave består af tre dele som kan løses på flere måder. Formålet med opgaven er at vise dine evner i forhold til at tænke logisk og arbejde analytisk med databearbejdning i praksis, samt videreformidling af metode og analyseresultater.
I første del skal du samle det datasæt som du har fået tilsendt. I anden del skal du, på baggrund af det samlede datasæt, besvare en problemstilling. I tredje del skal du formidle dine analyseresultater, samt beskrive de forskellige dataskridt du har taget for at samle og rense datasættet.
Du skal løse opgaven på din egen computer. Den kan løses i Excel, men hvis du har et andet databehandlingsprogram liggende, må du meget gerne bruge det. Du forventes at bruge maks. 3 timer på opgaven. Du forventes derfor heller ikke at levere en færdig analyse til sidst, men blot præsentere og formidle de resultater du kommer frem til.

Dataforberedelse

Du har fået udleveret en liste med informationer om passagererne på Titanic. På grund af rod i maskineriet er listen blevet delt op i to dele, som er gemt i forskellige filformater, og passagerernes navne er endt på en særskilt fil. Passagerlisten skal derfor samles igen, så vi har styr på hvilke passagerer, der var ombord på Titanic.
Hver af de tre filer er gemt i to forskellige filformater, så du kan selv vælge hvilket filformat du vil arbejde med. "Datasæt del1" indeholder oplysninger om 601 passagerer, "datasæt del2" indeholder oplysninger om 712 passagerer og "navne" indeholder navneoplysninger på samtlige passagerer.

I datasættet indgår følgende oplysninger:

- *pclass*: Passenger class (1 = 1st; 2 = 2nd; 3 = 3rd)
- *survival*: A Boolean indicating whether the passenger survived or not (0 = No; 1 = Yes); this is our target
- *name*: A field rich in information as it contains title and family names
- *sex*: male/female
- *age*: Age, asignificant portion of values aremissing
- *sibsp*: Number of siblings/spouses aboard
- *parch*: Number of parents/children aboard
- *ticket*: Ticket number
- *fare*: Passenger fare (British Pound)
- *cabin*: Location of the cabin
- *embarked*: Port of embarkation (C = *Cherbourg*; Q = *Queenstown*; S = *Southampton*)
- *boat*: Lifeboat, many missing values

- *body*: Body Identification Number
- *homedest*: Home/destination
- *case_id*: Identification number

Når du har samlet det fulde datasæt, vil det fremgå, at der har sneget sig nogle gæster ombord i datasættet som ikke hørte til på Titanic. Du skal derfor undersøge og rense datasættet for outliers, inden du kan fortsætte med analysen. Du får nogle hints til at finde dem:

- En af passagererne har betalt overpris.
- Der er en passager på listen, hvis overlevelse har været svær at fastslå.
- En af passagererne blev først født efter Titanic sank.
- En af passagererne er blevet registreret flere gange.

Når du er færdig med at samle og rense datasættet, kan du gå videre til næste opgave, hvor du skal besvare nogle spørgsmål på baggrund af den samlede passagerliste.

Dataanalyse

I denne opgave skal du udvælge og undersøge nogle af de forhold, der kan have haft betydning for sandsynligheden for, hvorvidt passagererne på Titanic overlevede det skæbnesvangre skibsforlis. Du må særligt gerne have fokus på, om passagererne på 1. klasse alle opførte sig som Rose's forlovede Cal Hockley fra filmen Titanic og selv sprang i redningsbådene, eller om der var engelske gentlemen om bord, som gav pladserne til kvinder og børn på 2. og 3. klasse. Du vælger selv, hvordan du vil undersøge og besvare denne problemstilling. Du er velkommen til at bruge regressionsanalyse eller teststatistikker, men kan også fint løse opgaven med krydstabeller. Nedenstående er eksempler på relevante spørgsmål, som du kan vælge at inddrage eller bruge som inspiration til din besvarelse. Det forventes ikke, at du besvarer samtlige spørgsmål.

- Havde kvinder og børn på 3. klasse, større sandsynlighed for at overleve, end mænd på 1. klasse?
- Er der en sammenhæng mellem billetprisen og de personer som overlevede?
- Overlevede samtlige passagerer, som endte i en redningsbåd?
- Havde ugifte kvinder større sandsynlighed for at overleve end gifte kvinder?

Formidling

Du skal som forberedelse til din næste samtale have udarbejdet en præsentation af dine resultater. Du vælger selv formen på din præsentation, det kan for eksempel være som et PowerPoint show eller som et kort notat. Udover en fremstilling af din besvarelse på ovenstående problemstilling, skal din præsentation også indeholde en beskrivelse af, hvordan du har bearbejdet datasættene og renset dem inden analysen.
Det er vigtigt, at du arbejder med formidling, og gerne grafisk fremstilling, i din præsentation, og målretter den et bredere publikum, som ikke har de samme metodiske kundskaber som dig selv.