Aflevering P1

Øvelser med data om passagerer ombord på Titanic, der sank i 1912

For at kunne bestemme hvilke resultater der kunne være interessant at trække ud, er der set på datasættet i sin helhed. Data består af deskriptive data som køn, alder, og hvilken klasse passagererne er rejst på. De beskriver lidt over en fjerdedel af del samlede antal passagerer.

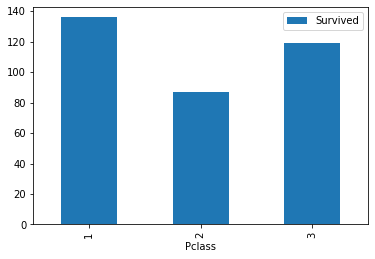
Ovenstående data er let aflæselige, men der er også en kategori som medrejsende søskende/forældre og børn/ægtefælle som er lidt sværere at have med at have med at gøre med i forhold til at trække nogle meningsfulde statistikker ud. Prisen på rejsen ser også ud til at svinge voldsomt fra passager til passager, så uden yderligere information om, hvordan de er udregnet, er det også en kategori, der er svær at have med at gøre analytisk.

Dernæst er der kigget på hvilke slags data, som data sættet består af, for at se på hvordan man kan bruge hvilke kolonner, og hvordan de kan behandles, ved hjælp af en simpel vise kommando. Sættet der hentes ind som en dataframe består af tal som intergers(hele tal), floats(kommatal) og strings som objects.

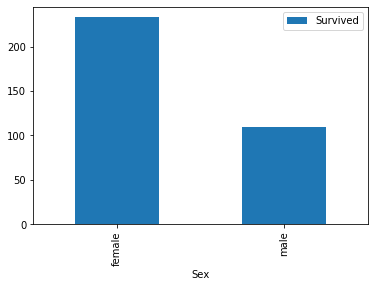
Derefter er der brugt simple funktioner, der ligger i pandas for at kunne besvare spørgsmål om størrelsen på datasættet, spørgsmål om den yngste og ældste passager samt middelværdi af alder på de 887 passagerer i sættet.

For at besvare spørgsmålet om der ens efternavne i sættet, er der først forsøgt at isolere efternavne ved at splitte objektet navn, for derefter at køre en funktion, der viser om der er duplikater i sættet. Analysen viser, at der findes duplikater, men da objekterne var uens af længde, og det derfor var vanskeligt at isolere efternavnet, er analysen ikke helt korrekt udført.

Til sidst viser en pivot tabel forskellen på antal overlevende passagerer på de tre klasser. For at visualisere tabellen og gøre resultatet mere umiddelbart af aflæse, er tabellen herefter lavet til en barchart med funktionen plot.bar():

￼￼

Og til allersidst er der lavet en visning af antal kvinder kontra mænd der overlevede.



Konklusion ￼￼

De tre små analyser til sidst viser lidt om hvad der er muligt at foretage sig i et python-skript med ganske lidt øvelse. Skulle man arbejde videre med tallene ville man skulle se lidt mere på hvilke analyser der ville give mening statistisk set. Antal kvinder og mænd i den sidste analyse ville for eksempel med fordel kunne regnes ud til procent.

Datasættet i sig selv har også nogle svagheder med uigennemsigtige tal og uklare definitioner.