

Portfolio 1 – Open Data Science

Mushtaba Osmani (wgq323)

Anslag med mellemrum: 2.955

**1) Åben filen i en tekst-editor og se på indholdet.**

**a. Hvilke data typer kan i identificere? (tekst / tal / typer af tal mv)**

Der identificeres følgende data typer, ”*Survived*” og ”*Sex*” data-typen er en boolean værdi da man arbejder med to svar, om passageren overlevede og i forhold til køn om hvorvidt det er en mand eller kvinde. Datatyperne ”*Pclass*”, ”*Siblings*/*Spouses* *Aboard* og *Parents*/*Children* *Aboard* er integers da de indeholder hele tal uden decimal. Derudover er ”Age” og ”Fare” bestående af både integers og floats da de indeholder hele tal og tal med decimal. ”*Name*” identificeres som værende en string.

**b. Mangler der data?**

Der mangler data i form af at man kun har 887 ud af 2224 passagerer og mandskab ombord.

På nogle er ”*Fare*” værdien 0, dette kan være som følge af missing data eller at passageren har rejst gratis som for eksempel sekretær eller ansat. Dette kunne være blevet løst ved at lave et andet data type ved at skelne mellem mandskab og passagerer.

**2) Ved at bruge panda – api skal i importere data til en dataframe**

Vi importerer først pandas og derefter lægger vi datasættet under navnet df\_titanic. Herfra består pandas af en række indbygget funktioner som man kan benytte sig af, såsom ”shape” og ”size” der identificer mængden af kolonner og rækker. ”d.types” finder alle data typer i datasættet

**3) Deskriptiv statistik i pandas**

Igen består pandas og spyder af en række relevante indbygget funktioner som man kan benytte i deskriptiv statistik. Vi bruger ”sum” for at finde mængden af overlevende og ”mean” for gennemsnitsalderen og gennemsnitsprisen for en billet. Derudover har vi importeret ”Counter” for at optælle mængden af mænd og kvinder ombord. Sluttende har vi med en for-loop og betingelser optalt hvor mange der har tilhørt i de forskellige prisklasser.

**4) Findes der personer med samme efternavn**

For at vurdere dette henter man først listen ”Name” i datasættet. Derefter laver vi en ny tom liste. For at hente udeluekkende efternavne laver vi en for-loop der henter det sidste ord i ”Name” da det altid er efternavnet. De efternavne bliver indsat i vores nye liste med ”.append”. som bliver gjort en nested for-loop, da den kører i vores eksisterende for-loop. Derefter bruger vi counter til at tælle alle efternavne og deres hyppighed. Sluttende bruger vi ”most\_common(3)” til at finde de 3 mest hyppige efternavne, som er Andersson, Sage og Johnson. Dermed bekræftes det at der findes personer med samme efternavn.

**5) Pivot tabeller i prisklasserne**

For får fordelingen af de tre prisklasser i en pivot tabel bruger vi funktionen ”.value\_counts” der tæller de tre prisklasser og passagerer tilhørende i deres respektive klasser.

For at vurdere hvilken prisklasse der havde flest omkomne, bruger vi ”groupby” med ”P class” og ”Survived” og derefter en ”.value\_counts”. Her får vi en klar pivot tabel med prisklasserne og omkommelsestal ved form af 1 og 0 hvor 0 er omkomne. Prisklasse tre havde flest omkomne med 119.