

**Open Data Science**

Opgave 1, September 24. 2019

Københavns Universitet

*Kommunikation og IT, efteråret 2017*

Lavet af

        Christoffer Emil Alsing – lsk576

**Opgave1:**

1. Hvilke typer data kan i identificere?

*I opgave et åbnede jeg csv filen i en text editor og kiggede på indholdet*

Csv filen indeholder 3 forskellige datatyper. ”Strings” i form af navnene på passagererne, ”Integers” i form af klassifikation af klasse, nr. på søskende etc. Til sidst er der også en enkelt ”Float” datatype i form af prisen for en billet.

1. Mangler der data?

I relation til at løse opgaven mangler der ikke data.

**Opgave2:**

Ved kalde af de indbyggede funktioner: .shape og .ndim kan jeg bestemme hvor stor data størrelsen er (887) og hvor mange dimensioner data sættet har som er 2:

**Opgave3:**

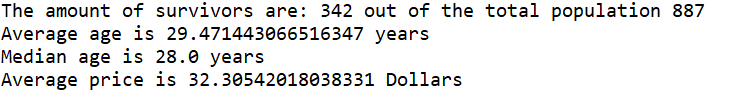
Til denne opgave valgte jeg først og fremmest at lægge størrelsen af datasettet ind i en variabel.

For at bestemme hvor mange overlevende der var anvende jeg funktionen .sum() til at returnere sum værdien i ”Survived” kolonnen.

Dernæst analyserede jeg den gennemsnitlige alder ved at bruge .sum() til at beregne summen af aldrene og dividere den med den totale længde af datasættet.

Til at beregne medianen af datasættet anvende jeg den indbyggede funktion .median().

Til sidst var jeg interesseret i at hvad den gennemsnitlige pris for en billet var (på kryds af alle billet klasserne) og dertil anvendte jeg igen .sum() og dividerede med længden.



*Udskrift fra konsollen*

**Opgave 4:**

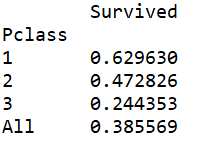
For at kunne bestemme om der var personer med samme efternavn skal jeg først kunne isolere efternavnene og ligge dem ind i en isoleret liste. Derfor overdrager jeg alle elementerne i ”Name” kolonnen over i et array: ”names”. Derefter looper jeg igennem ”names” og anvender split til at dele hvert navn op i tre dele adskilt af mellemrum. Efternavnet vil altid være det sidste navn i den opsplitning. Dog indeholder nogle navne 4 mellemrum kontra de resterende som har 3. Derfor har jeg en ”if” sætning som skubber det opsplittede element på index 3 hvis elementet er 4 i length, og index 2 hvis elementet er 3 i length.

Til sidste anvender jeg den indbyggede funktion set til at finde ud af men der er gengangere i mit ”set” af data. Det returnerede bliver:



**Opgave 5**

For at kunne succesfuldt lave en pivot tabel anvender jeg pandas’s indbyggede funktion .pivot\_table(). Jeg indsætter ”Pclass” som mit index, og ”Survived” som mit ”values”: Derefter bliver dette printet:



Ud fra dette kan jeg udlede at der var flest omkomne på rejseklasse 3 eftersom der er færrest overlevende på denne rejse klasse.