**Statistik på CSV-filen Titanic – Sara Liv Broe & Camilla Ejby Pedersen**

Med udgangspunkt i csv filen Titanic, har vi bl.a. set på diverse datatyper, brugt pandas til at importere Titanic som en dataframe, samt fået udtrukket forskellige informationer fra den tilgængelige data. Titanic viser forskellige informationer for 887 passagerer, ud af de i alt 2224 mennesker ombord. Vi får altså ikke data fra alle menneskene ombord, men kun et udsnit. Vi kan ikke vide hvilke informationer der blev registreret, men afhængigt af, hvad man skal bruge oplysningerne til, så kan det være, at mere data ville være brugbart.

Til at starte med åbnede vi filen i en tekst-editor. Dette gjorde vi ved at downloade csv filen og herefter dobbeltklikke på den, hvorefter vi åbnede den i TextEdit.

Titanic består af otte headings:

**Survived**

* Består af integers, 0 og 1, hvor 1 betyder at personen overlevede og 0 betyder det modsatte.

**Pclass** (Passenger class)

* Består ligeledes af integers, 1,2 og 3, og viser hvilken klasse at vedkommende rejste på.

**Name**

* Dette er en string, hvor vi både kan se hvad vedkommende hed, samt kvindernes status (hvis der står Mrs, så var de gift og hvis der står Miss., så var de ikke gift)
* Endvidere kan vi ved nogle også se deres navne fra før de blev gift.

**Sex**

* En string, hvor man kan se passagerens køn, mand (male) eller kvinde (female)

**Age**

* Både intergers, heltal, og floats, kommatal. Her vises passagerens alder, hvis der f.eks. står 0.83, så må man gå ud fra at passageren var mindre end et år, da vi ikke har mulighed for nærmere at undersøge om hvorvidt at kommatallene er en fejl eller ej.

**Siblings/Sp** (søskende/ægtefælle)

* Integers, som viser om passagerens havde enten søskende eller ægtefælle med ombord. Det er ikke muligt at se om der er tale om søskende eller ægtefælle.

**Parents/Ch** (forældre/børn)

* Som i den forrige header er der her tale om integers. Man kan her se om passageren havde enten forældre eller børn med ombord. Det er ikke muligt at afgøre om det er forældre eller børn der er dokumenteret.

**Fare**

* Består af en blandning af integers og floats, hvor man kan se hvor meget at vedkommende betalte for rejsen.

Vi har ved hjælp af Pandas åbnet filen i en dataframe, beskrevet den tilgængelige data, udtrukket og beregnet på data i filen, som f.eks. hvor mange der overlevede og hvad gennemsnitsalderen var for personerne. Vi har set på hvor mange personer der delte efternavn, samt lavet pivot tabeller over hvor mange personer der rejste på hver klasse, samt hvilken klasse der havde flest omkomne.

Alt dette kan ses ved hjælp af koderne som vi skrev med pandas, hvor der også er uddybende kommentarer.