Mickey Oliver Johansson, mvx394 - PORTFOLIO – Opgave 1, Titanic

Opgave 1 & 2

I den første opgave er der blevet bedt om at idenficere datatyper og manglende data. I opgave 2 skal data-sættet beskrives. Dette er blevet gjort ved at printe indbyggede funktioner i henhold til ”data”. For eksempel kan man ved brug af ”data.dtypes” finde ud af at der bliver brugt ”integer”, ”float” samt ”object” i datasættet. ”Integer”, en numerisk værdi uden decimaler, er brugt i overlevende eller ikke, prisklasse og antallet af familiemedlemmer m.m. ombord. ”Float”, en numerisk værdi med decimaler, brugt i alder og billetpriser. ”Object”, når kolonnen indeholder ”strings” – tekst, er brugt i navne og køn. Man kan også finde ud af navnene på de forskellige kolonner, ved at bruge ”data.columns”, og derved finde frem til hvilke informationer man har at gøre med i datasættet. Disse er følgende: ‘Survived', 'Pclass', 'Name', 'Sex', 'Age', 'Siblings/Spouses Aboard', 'Parents/Children Aboard', 'Fare'. Hvad angår manglende data, kan man se på kolonnen ’Fare’. Her er der en del med værdien 0. Dette kan tyde på manglende data, men hvis man undersøger nogle af navnene nærmere, kan man finde ud af at mange af dem bl.a. var ansatte på Titanic, hvilket højst sandsynligt betyder at de ikke skulle betale for en billet.

Opgave 3

Her skal man beregne på diverse data for at uddrage relevant information. For at løse dette, er der blevet brugt ”statistics” modulet og dets ”mean” og ”median” funktioner. Disse hjalp med at hurtigt udregne gennemsnitsalder, medianalder, gennemsnitspris samt medianpris. For at optælle hvor mange mænd og kvinder der var ombord på Titanic, er der blevet lavet en funktion der indeholder et ”for Loop” og et ”if-statement”. Denne lægger én til en tæller tilhørende mænd hver gang der forekommer ”male” i kolonnen og én til en tæller tilhørende kvinder resten af gangene. Der er blevet lavet en lignede funktion til at optælle antallet af overlevende, dog kunne man også, som vist i koden, blot have lavet en ”sum” af kolonnen ”Survived” for at få det samme resultat.

Opgave 4

For at finde ud af om der var folk med samme efternavn, samt hvilke efternavne der var flest der havde, er der blevet lavet yderligere en funktion. Denne indeholder et indlejret ”for Loop” der først uddrager det sidste navn i hvert navn (efternavnet) og derefter bliver det navn lagt ind i en liste. Efterfølgende bliver ”Counter” funktionen fra ”collections” modulet brugt til at optælle hvor mange gange efternavnene forekommer i data-sættet. Slutteligt bliver funktionen ”most_common” brugt til at vise de, i dette tilfælde, 8 mest forekommende efternavne.

Opgave 5

For at finde hvor mange der rejste på hver prisklasse, er der blevet brugt en for Loop, der minder om de andre der er blevet nævnt. Loopet gennemgår listen ”Pclass” og lægger én til enten ”first”, ”second” eller ”third”-variablen alt efter hvilken værdi der forekommer. Efterfølgende bliver "groupby" funktionen brugt for at dele dataen op i forhold til ”Pclass” og ”Survived” kolonnerne. Funktionen "value_counts" bliver herefter brugt for at vise antallet af gange disse værdier forekommer. Til sidst kan man ud fra resultatet af dette, se at det var prisklasse 3 der havde 368 omkomne og derved flest.