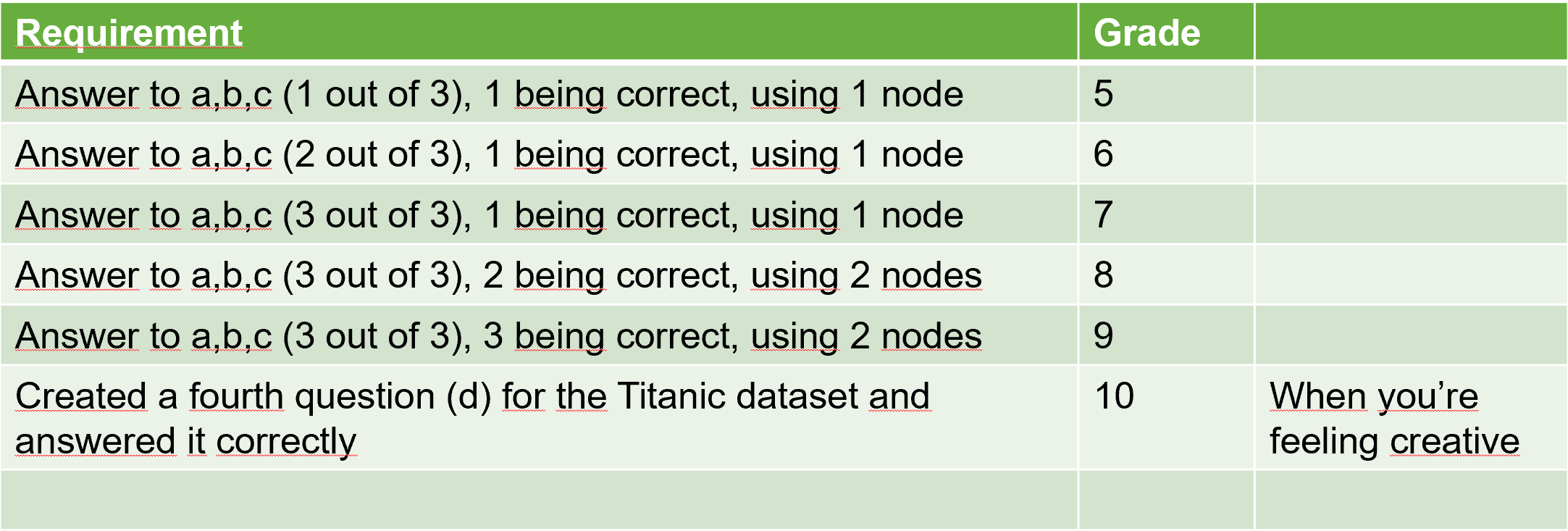
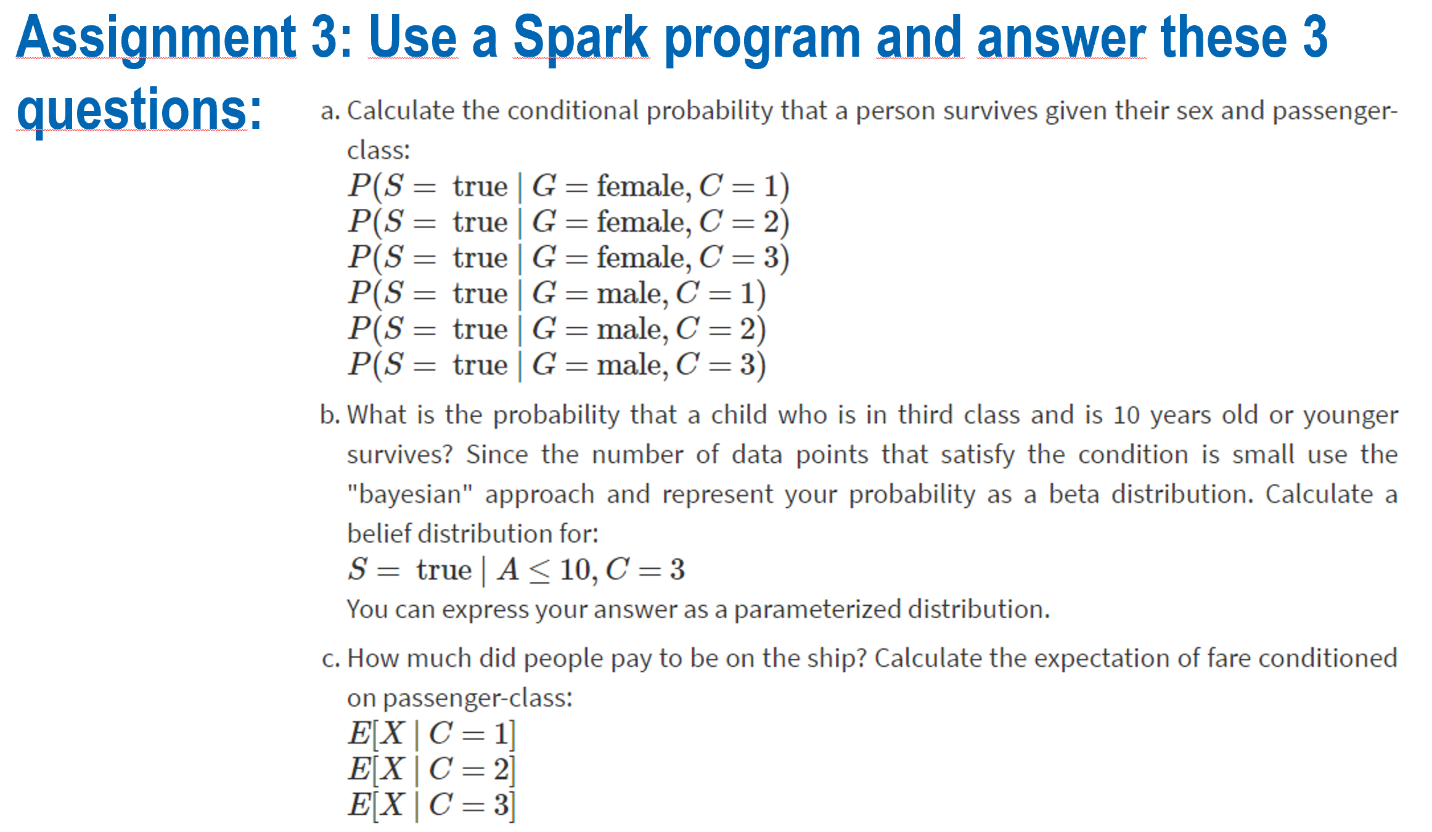
# PDP assignment 3



Prerequisities:

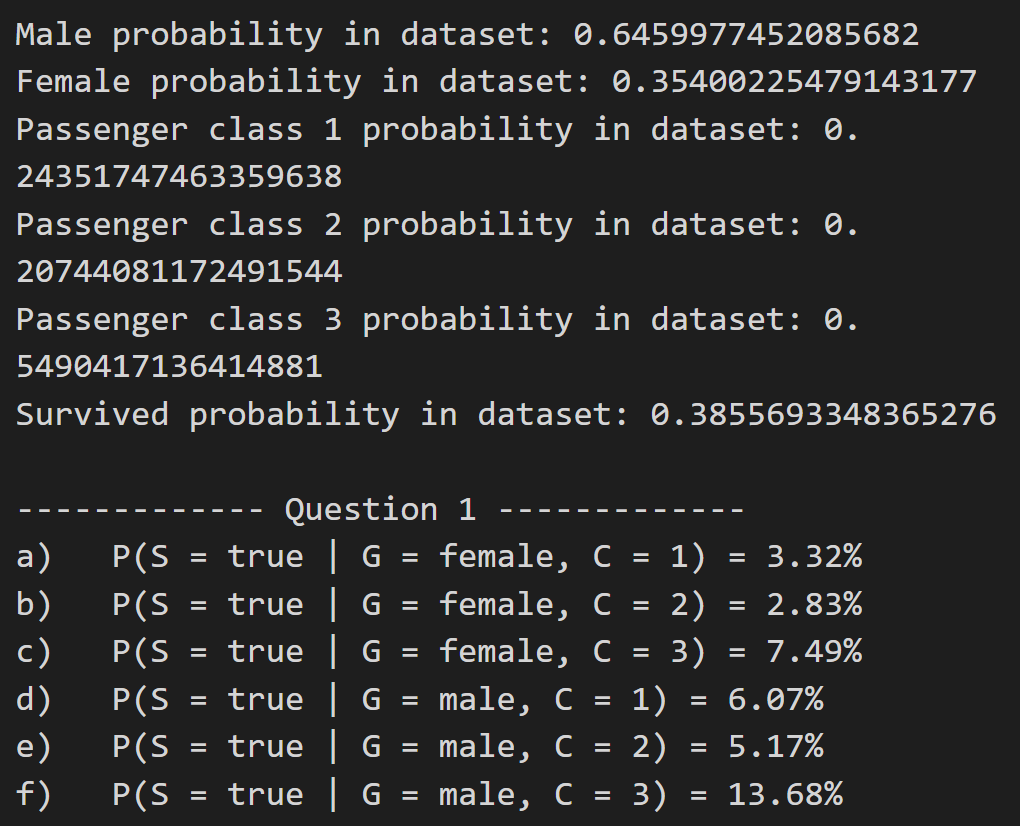
1. Installeer een Python code-editor als Anaconda (Jupyter Notebooks) of VS Code
2. Download de volgende packages (conda/pip install [package])
   1. PySpark
   2. FindSpark
   3. Pandas
   4. Numpy
3. Download Spark 3.0.0 (<https://spark.apache.org/downloads.html>)

*Uitleg code in script. Zie onderstaande URL:* <https://github.com/LennartdeWaart/Hadoop/blob/master/Assignment_3/script.py>

**Resultaat**

*Opdracht 1*

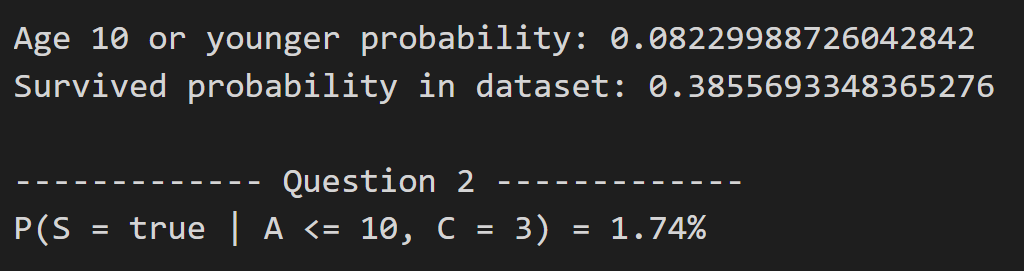
Om de kans te berekenen dat bijvoorbeeld een vrouw uit passagiersklasse 1 overleeft dienen eerst de kansen van de individueel betrokken variabelen te worden berekent (hetzij geslacht, passagiersklasse en of een persoon wel of niet overleeft). Om de antwoorden te krijgen op de gevraagde kans formules dienen de verschillende kansberekeningen van de invoeringen van de variabelen te worden vermenigvuldigt. Het resultaat is als volgt:



**Figuur 1: Resultaat opdracht 1**

*Opdracht 2*

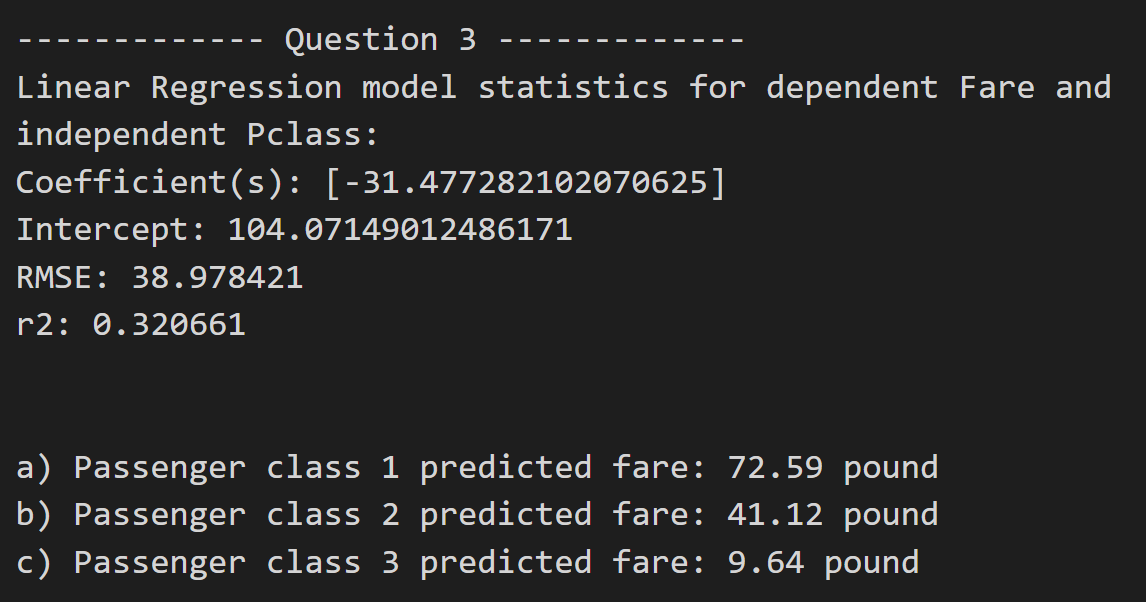
Om een antwoord te krijgen op de gevraagde kans formule dat een kind van 10 jaar of jonger in de 3e passagiersklasse overleeft dient de voorkomenskans van een kind van 10 jaar of jonger in de dataset te worden berekent. Door deze kansberekening te vermenigvuldigen met de kans dat een persoon overleeft kan een antwoord worden gegeven op de vraag. Het resultaat is als volgt:



**Figuur 2: Resultaat opdracht 2**

*Opdracht 3*

Om te voorspellen hoe duur een kaartje was voor de verschillende passagiersklassen dient een Linear Regression model te worden gemaakt. Dit wordt gedaan op basis van de kolommen “Fare” en “Pclass”, waarbij Fare de afhankelijke variabele (label) is en Pclass de onafhankelijke variabele (features). Na het creëren van het regressiemodel kan een formule worden opgesteld om de kosten van een kaartje te voorspellen op basis van de passagiersklasse door gebruik te maken van het format y = ax + b. De formule wordt als volgt: fare = intercept + coefficient \* pclass. Bij het invullen van een passagiersklasse (1, 2 of 3) kan nu de kosten van een kaartje worden voorspelt. Het resultaat is als volgt:



**Figuur 3: Resultaat opdracht 3**