Aufgabe 1 – Titanic

Gegeben ist der Datensatz der Passagiere der Titanic. Folgende Informationen können dem Datensatz entnommen werden:

* Hat eine Person das Unglück überlebt (0: nein, 1: ja)
* In welcher Klasse ist die Person gereist (pclass)
* Name des Passagiers
* Geschlecht (sex)
* Alter (age)
* Anzahl der Geschwister/Ehegatten/Ehegattinnen an Board (sibsp)
* Anzahl der Eltern/Kinder an Board (parch)
* Ticketnummer (ticket)
* Fahrpreis (fare)
* Kabinennummer (cabin)
* Ort, an dem Passagier zugestiegen ist (embarked, C: Cherbourg, Q: Queenstown, S: Southampton)

Nutze zuerst Python, um aus dem großen DataFrame mehrere kleine DataFrames mit nur einem Teil der Spalten zu generieren. Erzeuge so folgende DataFrames mit den Inhalten:

1. DataFrame: passengerid, Name und Alter
2. DataFrame: passengerid, survived
3. DataFrame: passengerid, Klasse, Ticketnummer, Fahrpreis, Kabine, embarked
4. DataFrame: passengerid, Geschlecht
5. DataFrame: passengerid, sibsp, parch

Erzeuge außerdem DataFrames, in denen der Überlebensstatus durch Textdaten ausgedrückt wird (z.B. „nicht\_ueberlebt“ statt 0), sowie ein DataFrame, in dem das Kürzel unter embarked mit dem ausgeschriebenen Namen des Ortes verbunden wird (also z.B. C/Cherbourg).

Lade diese neuen DataFrames nun in eine SQL-Datenbank deiner Wahl (und begründe in der Präsentation dann deine Wahl 😉) und verknüpfe die Tabellen entsprechend mit sinnvollen Keys.

Analysiere nun den Datensatz mit SQL und beantworte mitunter folgende Fragen:

1. Wie viele Personen waren an Board?
2. Wie viel Prozent der Frauen/Männer/Kinder haben das Unglück überlebt?
3. Spielte die Klasse, mit der die Passagiere gereist sind eine Rolle für die Überlebenschance?
4. Hatte der Ticketpreis einen Einfluss auf die Überlebenschance?
5. Spielte der Ort, an dem die Passagiere zugestiegen sind, eine Rolle?
6. Haben Personen, die mit Familie gereist sind, eher überlebt?
7. Spielt das Alter eine Rolle?
8. Lass dir drei weitere sinnvolle Fragen einfallen.