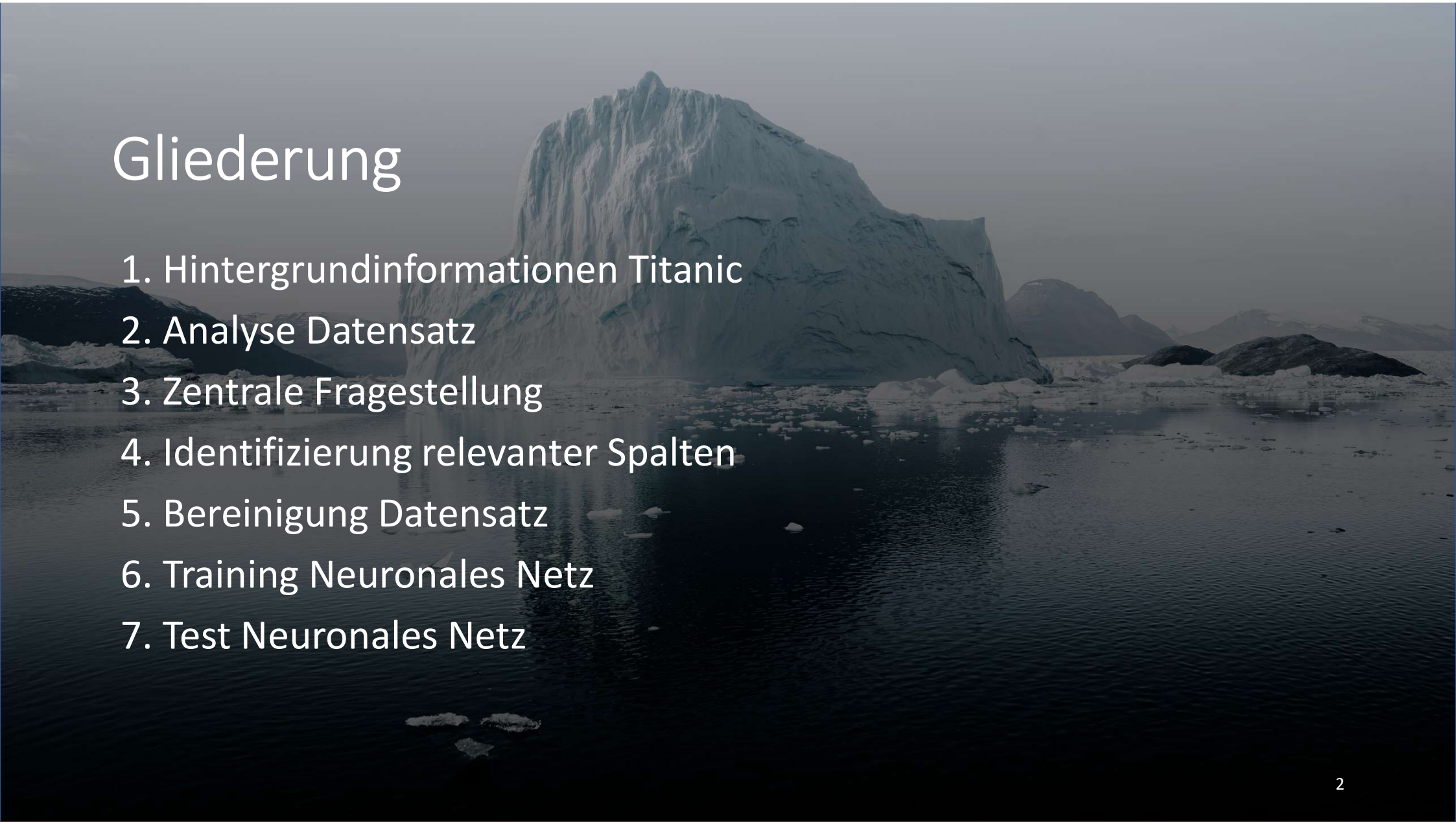
Neuronales Netz
-
Überlebende des Untergangs der Titanic

Gruppe 1

1294343, 4517830, 5223626, 1749577, 6310429

Dozent: Dr. Ralf Höchenberger

Termin: 06.09.21



Gliederung

1. Hintergrundinformationen Titanic
2. Analyse Datensatz
3. Zentrale Fragestellung
4. Identifizierung relevanter Spalten
5. Bereinigung Datensatz
6. Training Neuronales Netz
7. Test Neuronales Netz

1. Hintergrundinformationen Titanic

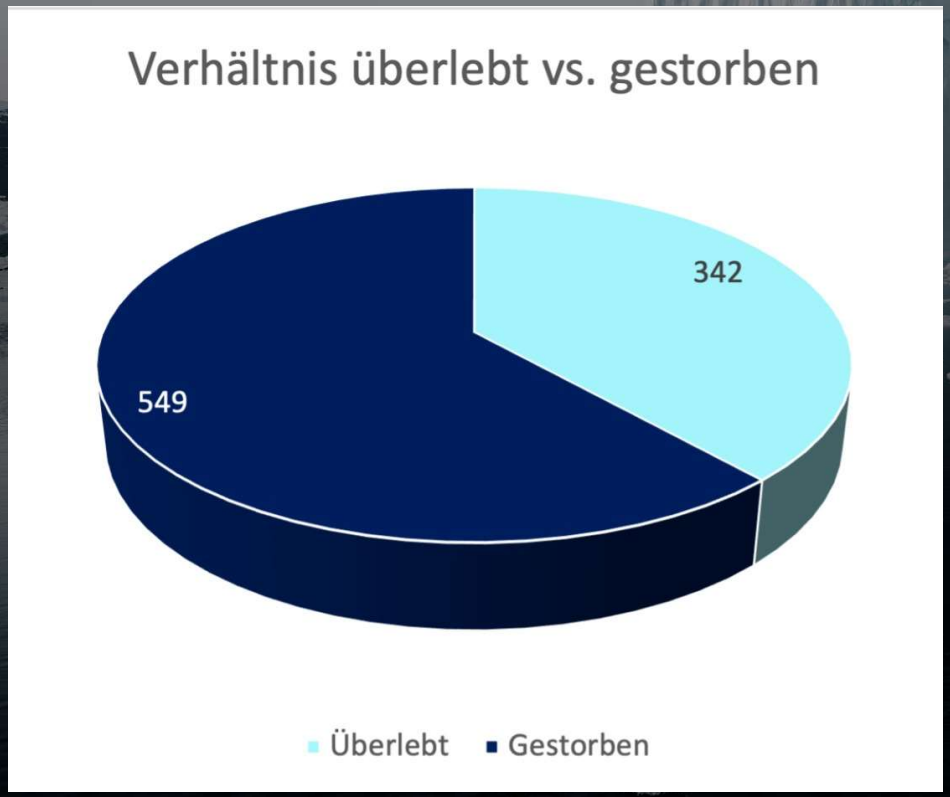
Quelle: https://www.youtube.com/watch?v=pyLGpq--WNc

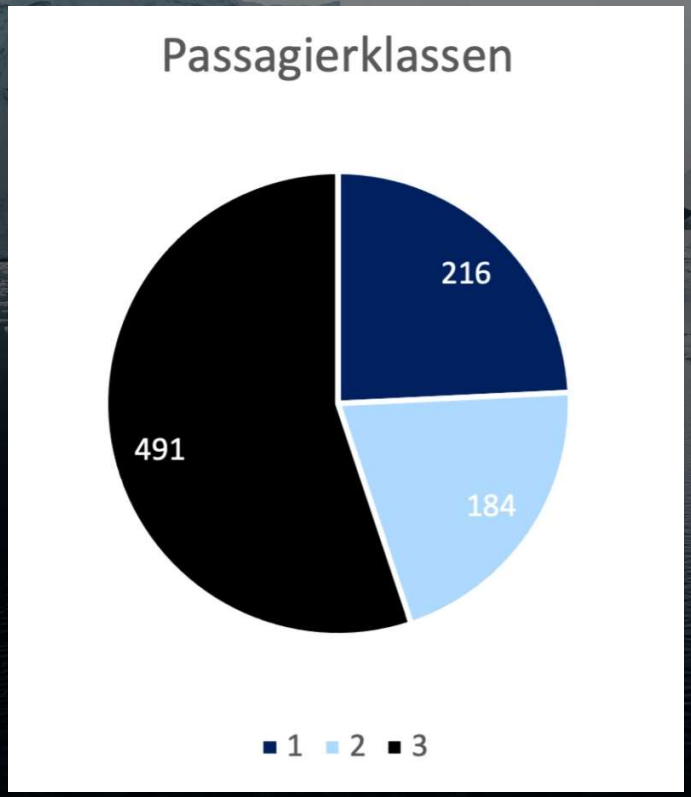
2. Analyse Datensatz – Spalte und Werte

Spaltenname			**Welche Werte nehmen die Variablen an?**
PassengerId		Passagiernummer	1-891
Survived		Überleben	0 = nein, 1 = ja
Pclass	Passenger Class	Passagierklasse	1 (hoch), 2 (mittel), 3 (niedrig)
Name		Name mit Titel	Name mit Titel
Sex		Geschlecht	male, female
Age		Alter	0-80
SipSp	Number of Siblings/Spouses Aboard	Anzahl der Geschwister/Ehepartner an Board	0-8
Parch	Number of Parents/Children Aboard	Anzahl der Eltern/Kinder an Board	0-6
Ticket	Ticket Number	Ticketnummer	zufällige Zahl
Fare	Passenger Fare	Ticketkosten	0-512,33 Dollar
Cabin		Kabinenbezeichnung	A-G für das Deck und danach eine Nummer (G106) (A ist das Höchste und G das Niedrigste)
Embarked	Port of Embarkation	Einschiffungshafen	C = Cherbourg; Q = Queenstown; S = Southampton

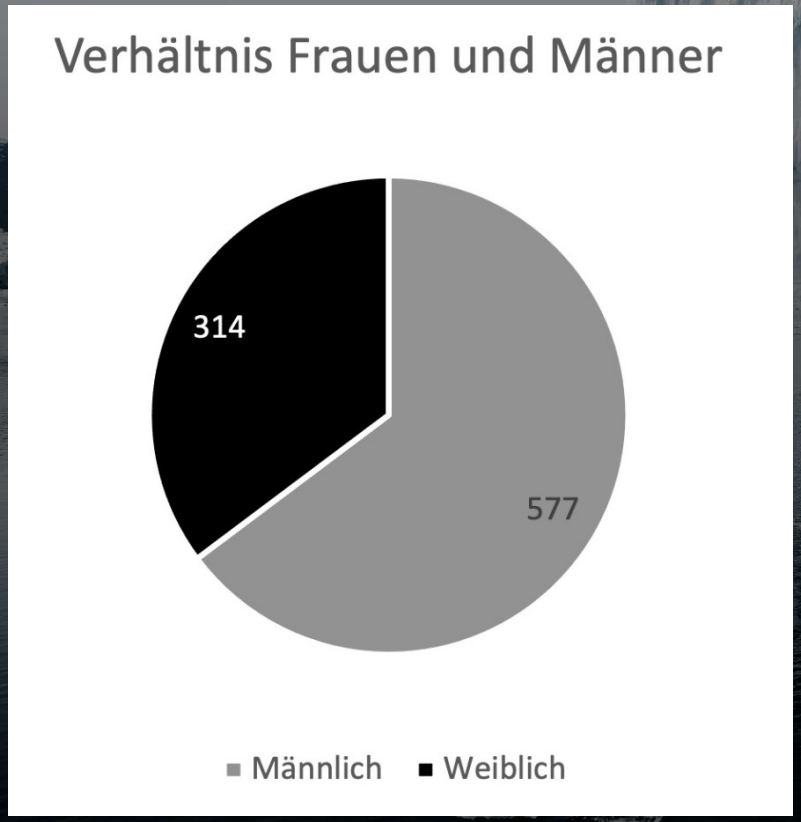
Der Datensatz (training_titanic.csv) umfasst Informationen zu 891 Passagieren.

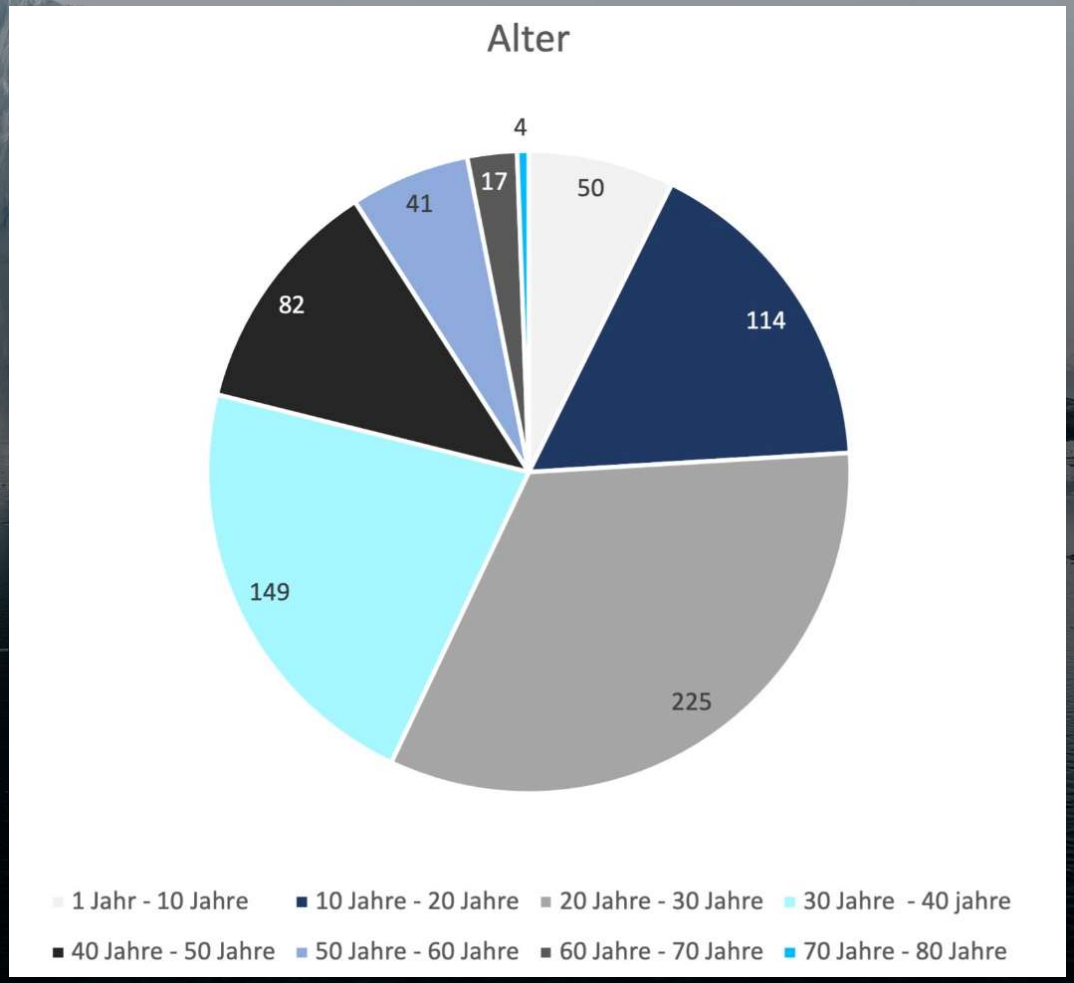
2. Analyse Datensatz – Grafiken



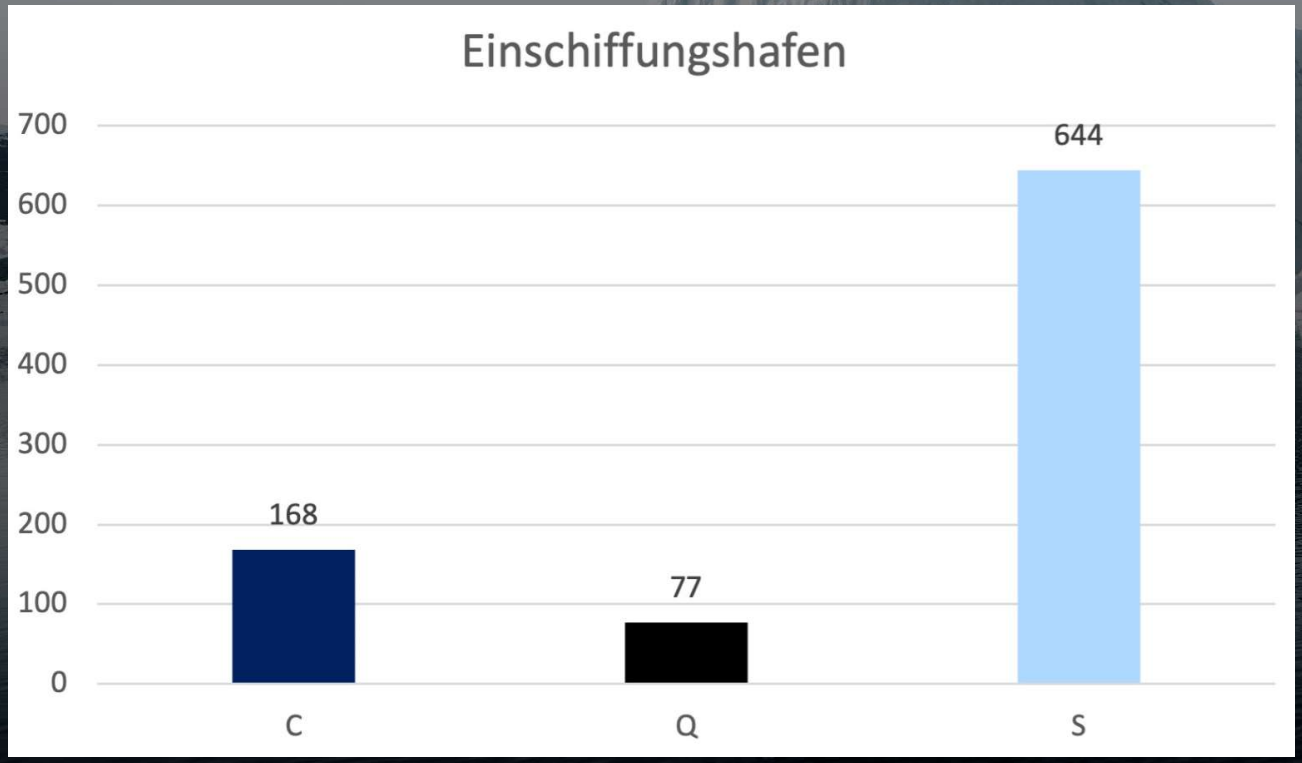


2. Grafiken

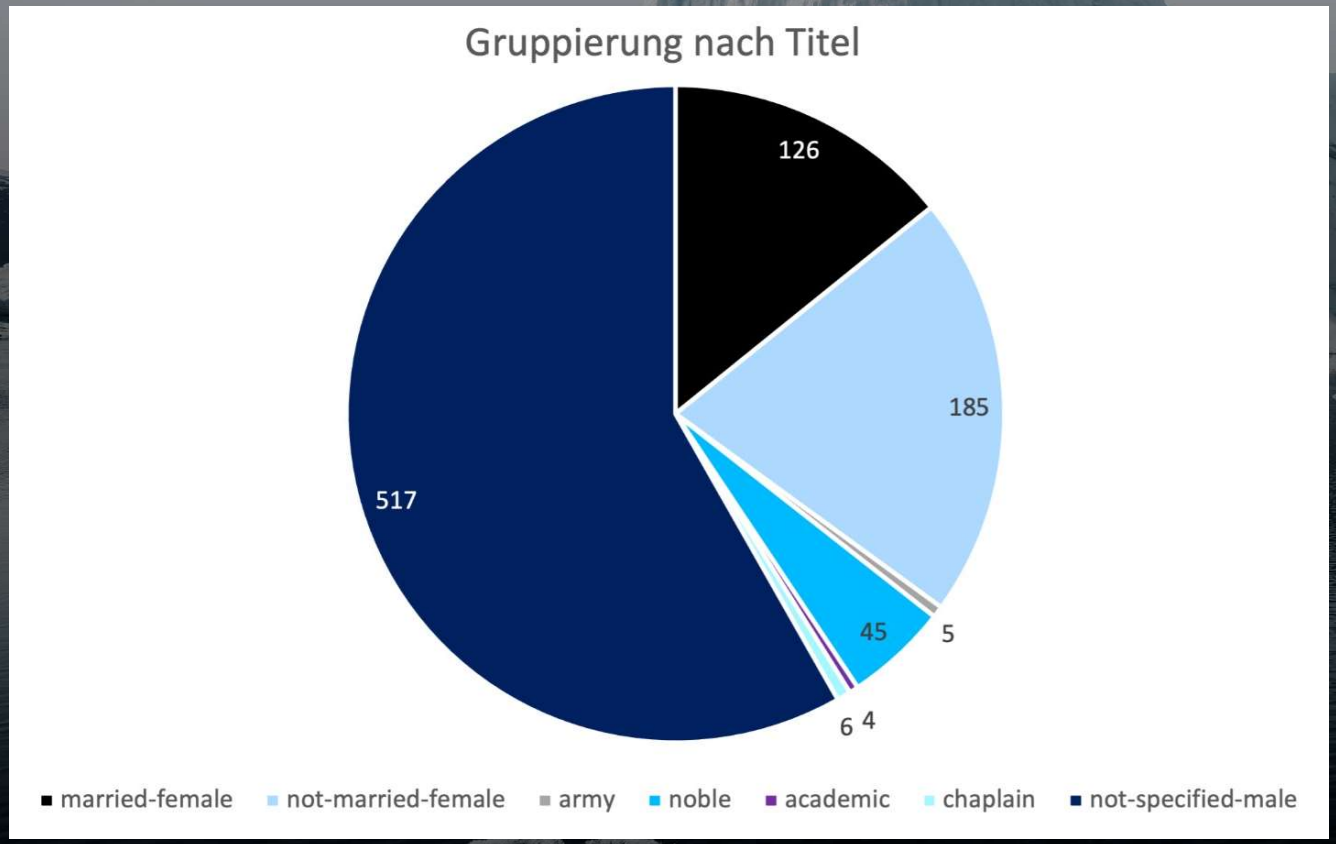

Verhältnis Frauen und Männer
314
577
Männlich
Weiblich


Alter
4
17
50
41
114
82
149
225
1 Jahr - 10 Jahre
10 Jahre - 20 Jahre
20 Jahre - 30 Jahre
30 Jahre - 40 jahre
40 Jahre - 50 Jahre
50 Jahre - 60 Jahre
60 Jahre - 70 Jahre
70 Jahre - 80 Jahre

2. Grafiken


Einschiffungshafen
700
600
500
400
300
200
100
0
168
77
644
C
Q
S

2. Grafiken


Gruppierung nach Titel
126
185
5
45
6
4
517
married-female
not-married-female
army
noble
academic
chaplain
not-specified-male

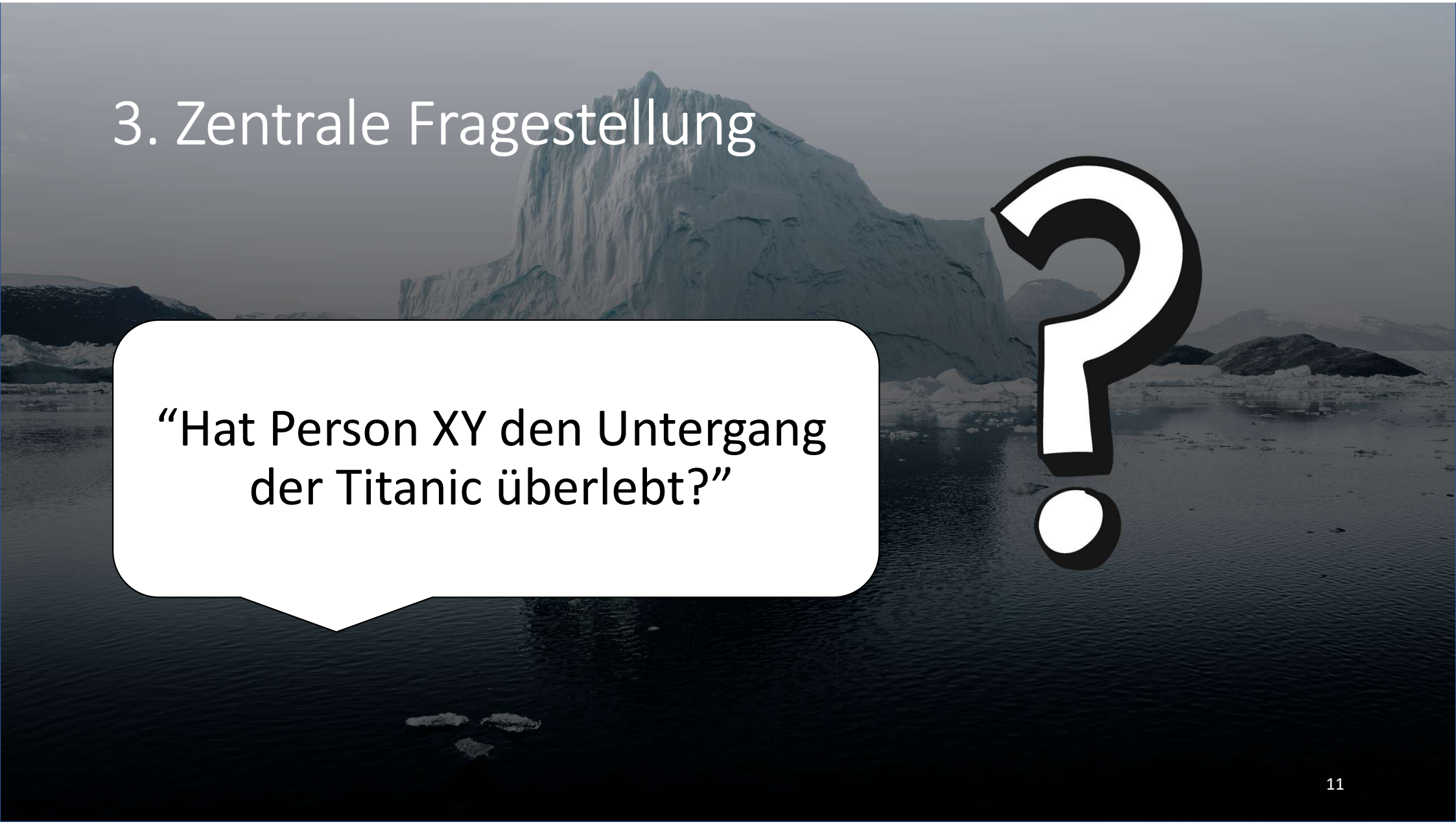
2. Analyse Datensatz – Analyse Korrelationsmatrix

- Allgemein: Wert zwischen -1 und 1
  - 0: überhaupt nicht korreliert -> Werte/Attribute hängen nicht zusammen
  - +1: Direkt Proportional -> Umso größer Wert A umso größer auch Wert B
  - -1: Indirekt Proportional -> Umso größer Wert A, umso kleiner Wert B
- Korrelation zwischen Überlebt und:
  - Pclass: -0.338481 -> je höher die Klasse, desto niedriger die Überlebenschance
  - Fare: 0.257307 -> siehe Pclass: je teurer das Ticket, desto besser (niedriger) die Klasse, desto höher die Überlebenschance
  - Age : -0.077221 -> schwach korreliert, aber: umso jünger, desto eher überlebt
    - Kontext wichtig: wenig Kinder/Jugendliche auf Schiff. Jung heißt eher unter 30 (siehe median)
  - Parch: 0.081629 -> schwach korreliert
    - Aber Tendenz zu: Umso mehr Eltern/Kinder von einem selbst dabei waren, umso eher überlebt man.

2. Analyse Datensatz – Summary Statistics

	Age	SipSp	Parch	Fare
Minimum	0.42	0.0	0.0	0.0
Maximum	80.0	8.0	0.6	512.33
Durchschnitt	29.7	0.52	0.38	32.2
Median	28.0	0.0	0.0	14.45
Varianz	211.019	1.22	0.65	2469.44

3. Zentrale Fragestellung



“Hat Person XY den Untergang der Titanic überlebt?”

4. Identifizierung relevanter Spalten

Spalte	Relevanz	Begründung
PassengerId	/	Fortlaufende Zahlen ohne weiteren Zusammenhang
Survived	relevant	
Pclass	relevant	
Name	/	Ohne Zusammenhang, nur extrahierter Titel relevant
Title	**relevant**	
Sex	relevant	
Age	relevant	20% fehlende Daten, daher keine Berücksichtigung der Spalte
SipSp	relevant	
Parch	relevant	
Ticket	/	Zufällige Nummer
Fare	relevant	
Cabin	/	77% fehlende Daten, daher keine Berücksichtigung der Spalte
Embarked	relevant	

5. Bereinigung Datensatz – Titel extrahieren

1. Alle vorhandenen Titel definieren
2. Neue Spalte mit extrahierten Titeln (aus Namensspalte löschen)
3. Titel kategorisieren
4. Inhalt der Titel-Spalte mit Namen der Kategorien ersetzten (aus Gründen der Übersichtlichkeit)

→ Code

5. Bereinigung Datensatz – Fehlende Datensätze

- Embarked: 0,2 % = Spalte wird berücksichtigt
- Age: 19,86% = Spalte nicht berücksichtigen
- Cabin: 77,1% = Spalte nicht berücksichtigen

	column_name	percent_missing
PassengerId	PassengerId	0.000000
Survived	Survived	0.000000
Pclass	Pclass	0.000000
Name	Name	0.000000
Sex	Sex	0.000000
SibSp	SibSp	0.000000
Parch	Parch	0.000000
Ticket	Ticket	0.000000
Fare	Fare	0.000000
Embarked	Embarked	0.002245
Age	Age	0.198653
Cabin	Cabin	0.771044

5. Bereinigung Datensatz – Sex/Embarked

- Sex:
  - Male = 0
  - Female = 1
  - → Code
- Embarked:
  - S = 1
  - C = 2
  - Q = 3
  - → Code

6. Training Neuronales Netz

Aufteilung des Datensatzes in

- 80% Trainingsdaten und
- 20% Testdaten

→ Code

7. Test Neuronales Netz

→Code

https://github.com/ToKoSoftware/big-data-analytics

https://tomaskostadinov.com/titanic/


Vielen Dank für die
Aufmerksamkeit!