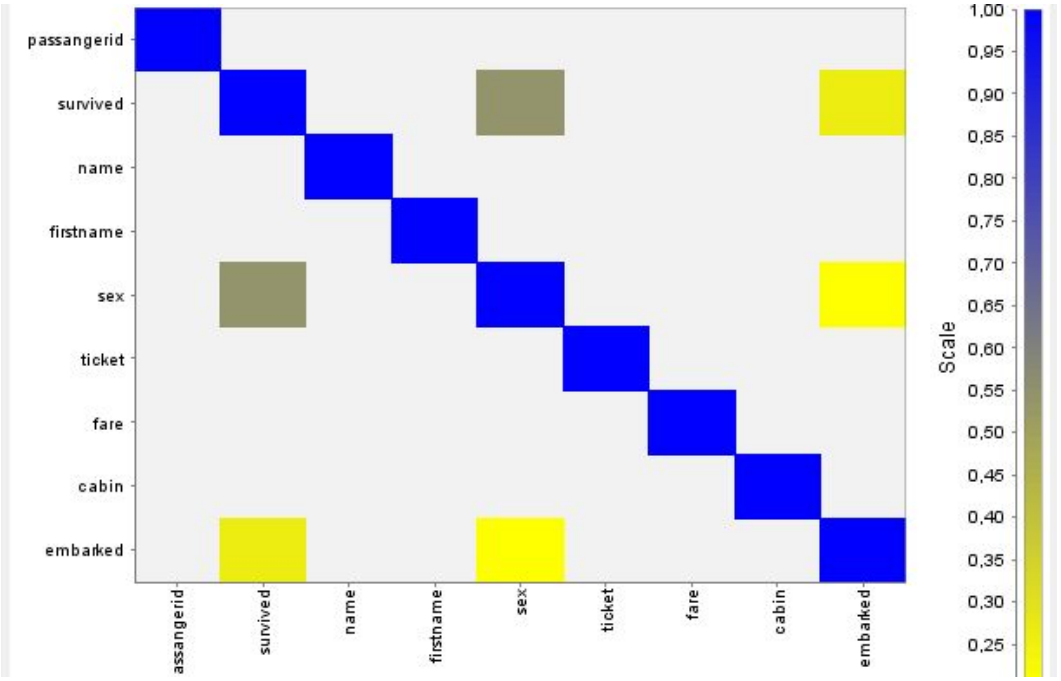
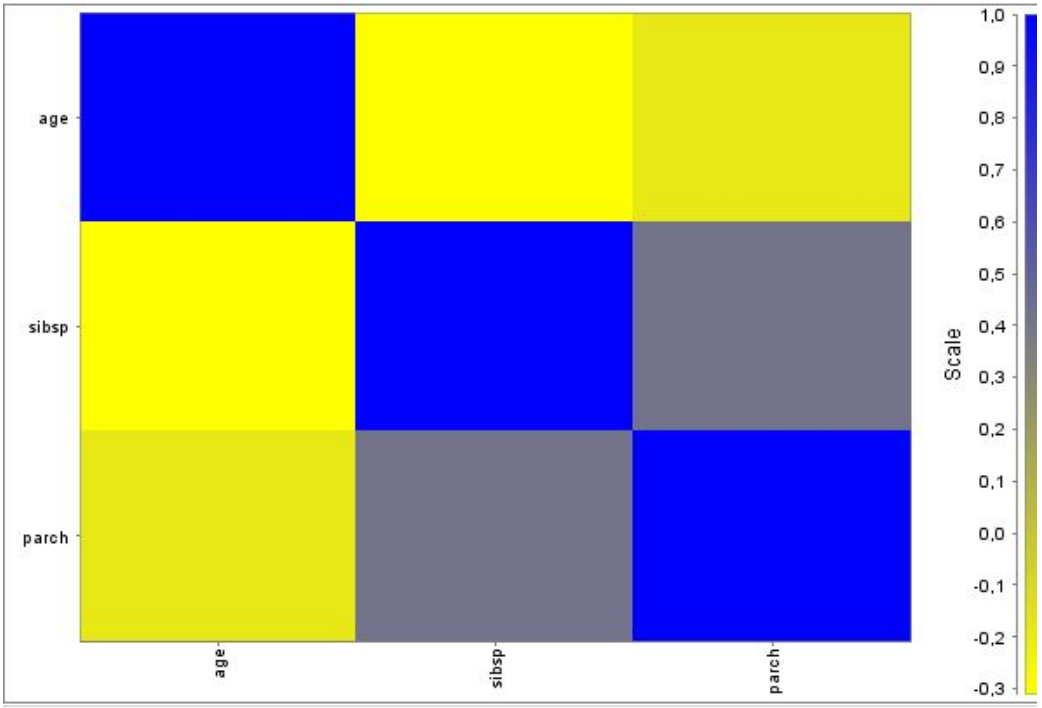
KNIME: Korrelationsmatrix

Zur Erstellung der Korrelationsmatrix wurden die Daten in metrische Merkmale und nominale/ordinale Merkmale unterteilt. Für jede dieser zwei Merkmalskategorien wurde eine extra Korrelationsmatrix erstellt, da je nach Merkmalsart die Berechnung ausgewählt wird.



Bei den nominalen Merkmalen konnten nur survived, embarked und sex betrachtet werden, da die anderen Merkmale zu viele einzigartige Merkmalsausprägungen aufweisen. Der Schwellenwert wurde dafür nicht verändert. Man erkennt, dass das Geschlecht einen Einfluss auf das Merkmal “survived” hat.



Bei den metrischen Merkmalen erkennt man, dass das Merkmal sibsp und parch stark positiv korrelieren. Das Merkmal sibsp und age korrelieren leicht negativ.

KNIME: Neigen die Familienangehörigen zusammen zu versinken oder zu schwimmen?

1. Definition von Familie
    a. Familien bestehen aus mindestens zwei Familienmitgliedern, die sich auf dem Schiff befinden
    b. Familien lassen sich über den Familiennamen (Nachname) und die Familiengröße (sibsp+parch+1) definieren
2. Aggregieren der Familien und aufsummieren der Spalte “survived”
    a. → Familien, bei denen die aufsummierte “survived” Spalte 0 oder der Familiengröße entspricht, sind entweder zusammen gestorben oder haben zusammen überlebt
3. Zählen dieser Familien

Row ID	S zusam...	I Count(...
Row0	getrennt	110
Row1	zusGestorben	74
Row2	zusUeberlebet	18

(Tabelle 1)

    a. 92 Familien haben entweder zusammen überlebt oder sind alle gestorben
    b. → 92/(110+92) = 45% der Familien haben entweder zusammen überlebt oder sind zusammen gestorben
    c. Von diesen 92 Familien sind 74 zusammen gestorben und nur 18 Familien haben zusammen überlebt
    d. 74/92 = ca. 80 % dieser Familien, sind zusammen gestorben
4. Hängt das mit der Familiengröße zusammen?
    a.

Row ID	I family	D Mean(s...
Row0	2	0.553
Row1	3	0.578
Row2	4	0.724
Row3	5	0.2
Row4	6	0.136
Row5	7	0.333
Row6	8	0
Row7	11	0

(Tabelle 2)

    b. Je größer die Familie (>4 Mitglieder), umso geringer ist die Überlebenschance eines Familienmitglieds (Nicht auf die Gesamtfamilie bezogen)

Row ID	family	zusam...	Count(...
Row0	2	getrennt	57
Row1	2	zusGestorben	44
Row2	2	zusUeberlebet	15
Row3	3	getrennt	38
Row4	3	zusGestorben	20
Row5	3	zusUeberlebet	1
Row6	4	getrennt	9
Row7	4	zusGestorben	2
Row8	4	zusUeberlebet	2
Row9	5	getrennt	2
Row10	5	zusGestorben	3
Row11	6	getrennt	2
Row12	6	zusGestorben	3
Row13	7	getrennt	2
Row14	8	zusGestorben	1
Row15	11	zusGestorben	1

(Tabelle 3)

c. Nur Familien mit wenig Mitglieder (<5) haben zusammen überlegt. In 15/18 Fällen war die Familiengröße gleich 2.
d. Je größer die Familie desto größer die Chance zusammen zu sterben bspw. bei Familiengröße 8 und 11 = 100%

Knime, extra Aufgabe

Frage: Hat die Wahl der Kabine einen Einfluss auf die Überlebensrate der Passagiere?

Sämtliche Freitext in Knime sind auf Englisch gehalten, da das Tool auch auf Englisch ist erleichterte das die Bearbeitung.

1. Daten zusammenfassen (wurde nur modelliert, nicht aktiv genutzt, da wir nichts predicted haben)
    - Testdaten können mit einer beliebigen Prediction-Datei zusammengefasst werden, nennen wir das Ergebnis Test_full
        - die Spalten werden in Benennung und Reihenfolge an die Trainingsdaten angepasst
    - anschließend werden Test_full und die Trainingsdaten zusammengefasst

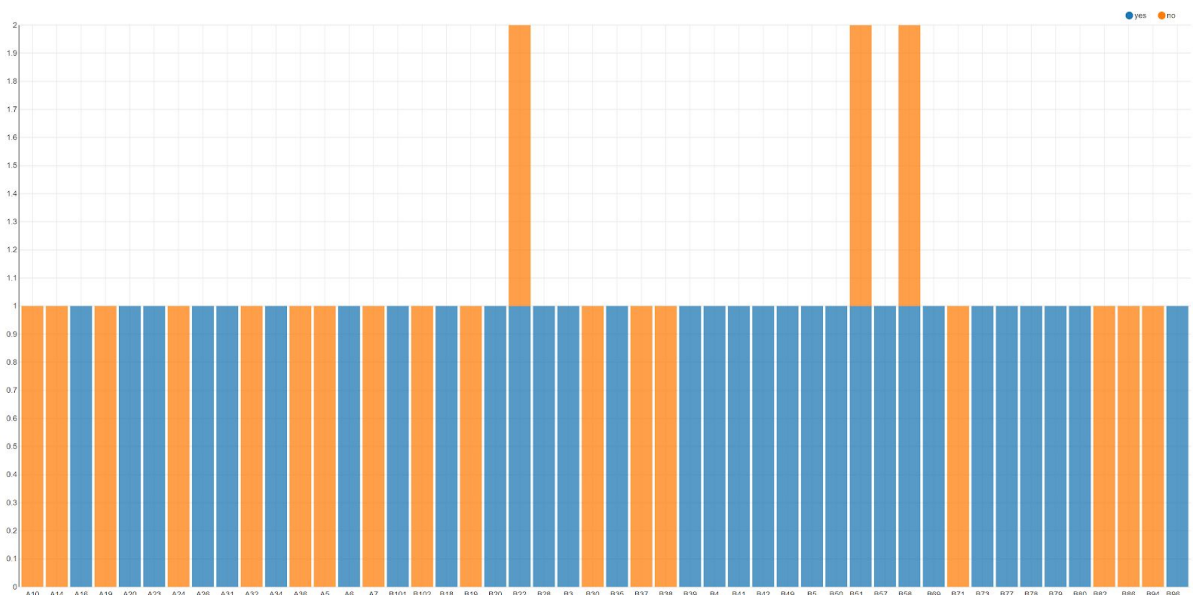
2. Erste Analyse: Wie viele Passagiere gibt es überhaupt mit Kabinennummer und wieviele davon haben überlebt, Ergebnisse:

Survived	Count*...	Missing ...
no	68	481
yes	136	206

- 204 Datensätze mit Kabinennummer (66,66% davon sind Überlebende)
- 687 Datensätze ohne Kabinennummer (29,99% davon sind Überlebende)
- Das Verhältnis zwischen Überlebenden mit und ohne Kabinennummer ist ca. 2,2 : 1

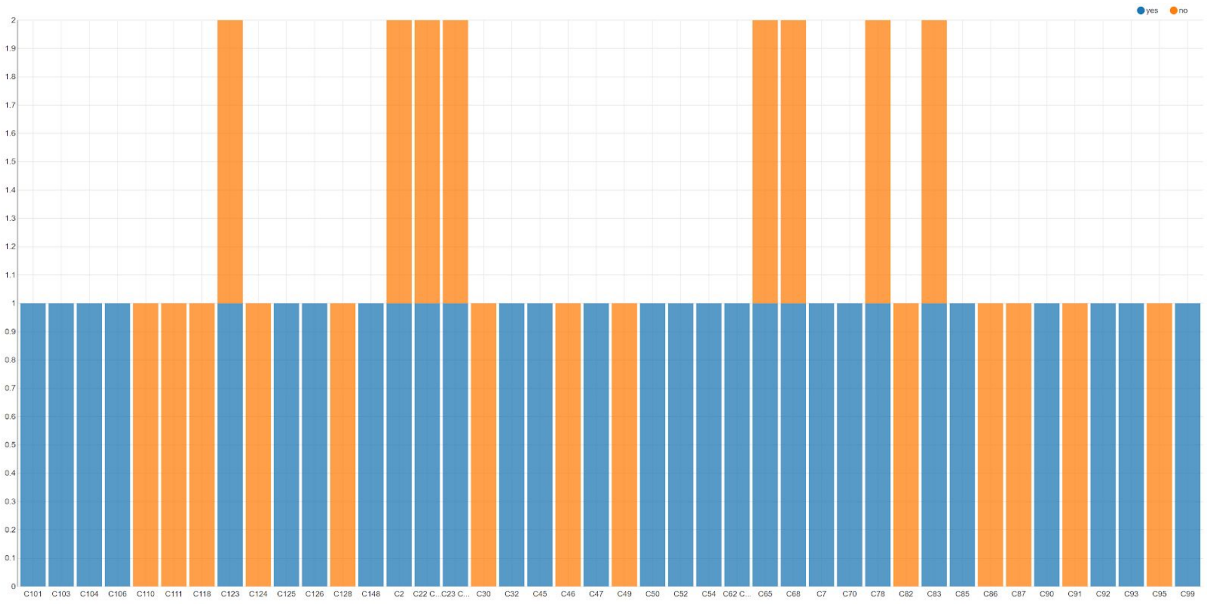
- 342 Überlebende gibt es (39,77% davon haben eine Kabinennummer)
- 549 Verstorbene gibt es (12,39% davon habe eine Kabinennummer)
- Das Verhältnis zwischen Kabinennummer der Überlebenden und Verstorbenen ist ca. 3,2 : 1
- Der enorme Unterschied zwischen den Überlebensraten und den Kabinenraten lässt folgende Hypothese (1) zu: Die Überlebenden wurden im Nachgang bei der Datenerhebung nach Ihrer Kabinenummer und denen ihrer Mitreisenden/Nachbarn gefragt. Und ob letztere überlebt haben.

3. Überprüfung von Hypothese 1:
    - zuerst mussten die Daten aus Kabinenebene aggregiert werden. Dazu wurde nur geschaut ob es überlebende oder verstorbene je Kabine gab. Die jeweilige Anzahl spielt zur Bestätigung der Hypothese keine Rolle. Da pro Kabine/Nachbarkabine nur mindestens eine Person überlebt haben muss, um die Hypothese zu bestötigen.
    - Da es zu viele Kabinen sind um ein übersichtliches Diagramm zu erstellen, wurde die Kabinen in drei Gruppen geteilt.
    - Kabinen A & B



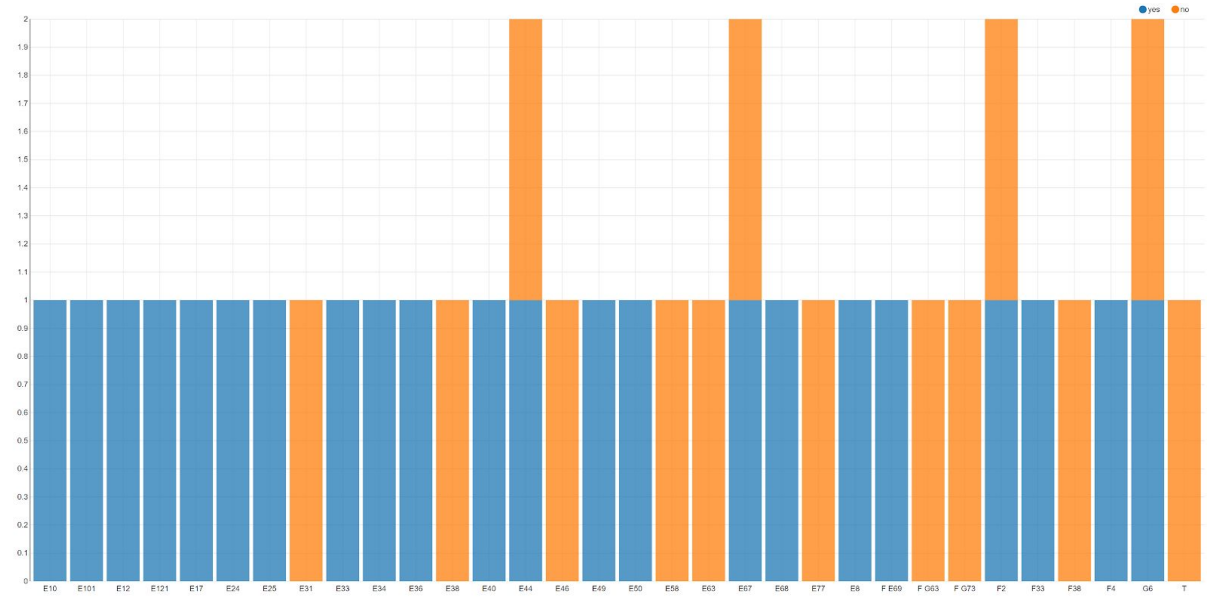
        - die blauen Balken stellen Überlebende da. Die Kabinen sind alphabetisch sortiert, wodurch fast alle benachbarten Kabinen auch nebeneinander dargestellt sind. Man sieht, dass nahezu jeder orange Balken einen blauen(überlebenden) Nachbarn hat. Teilweise haben die Überlebenden auch die Verstorbenen aus Ihren eigenen Kabinen benannt. → Hypothese 1 trifft für Kabinen A & B zu.

    - Kabinen C & D



- hier haben die orangen Balken fast immer blaue Nachbarn. Also haben auch hier die Überleben Informationen zu den Verstorbenen aus den Nachbarkabinen und ihren verstorbenen Mitbewohnern geliefert. → Hypothese 1 trifft auch für Kabinen C & D zu.

- alle weiteren Kabinen

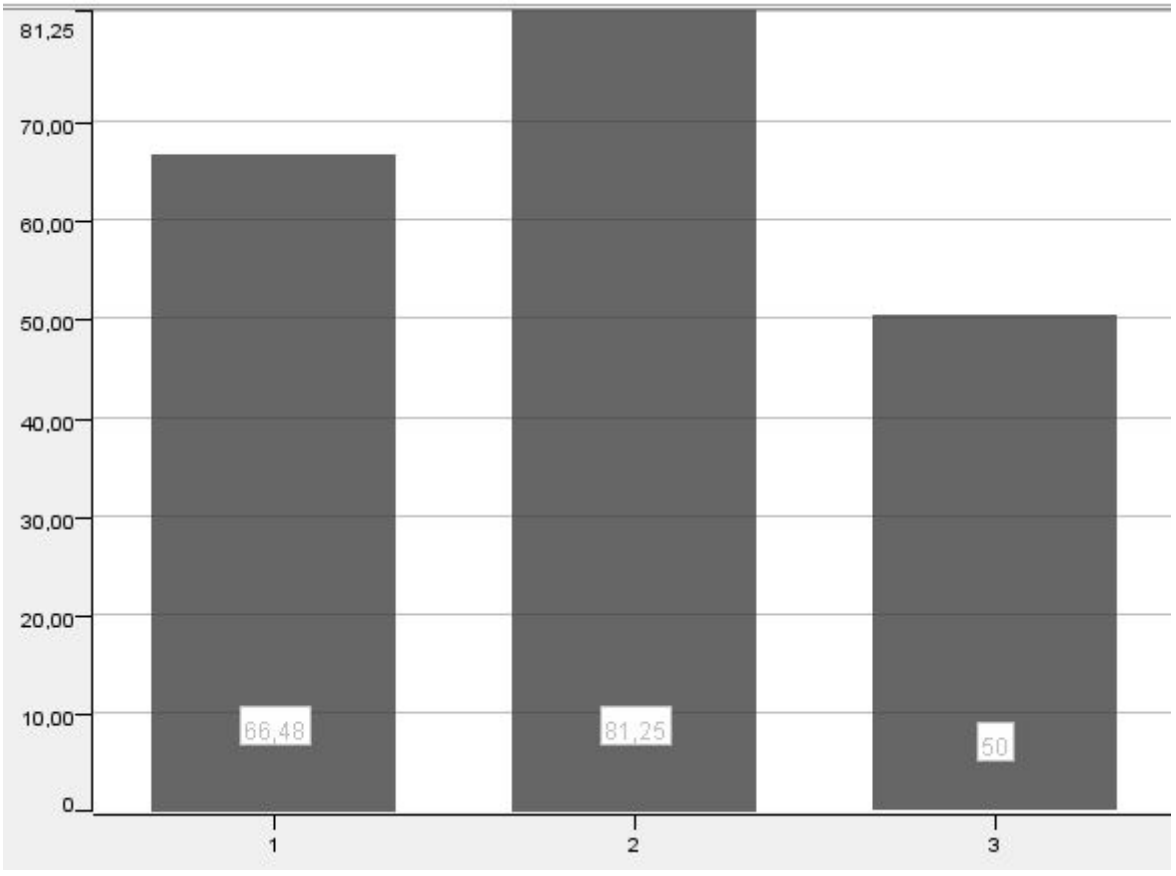


- auch hier haben die Orangen Balken immer überlebende Nachbarn. Die Unschärfe wird u.A. an der Datenbeschaffenheit liegen, da bei einigen Kabinen die laufende Nummer fehlt. Vielleicht war hier nur noch das Deck aber nicht mehr die genaue Nummer bekannt. Die einzige Ausnahme stellt Kabine T dar. Diese könnte ein VIP-Kabine sein, bspw. für den Kapitän, wo von vornherein klar war, dass er die Kabine bezieht. Und vom Spielfilm Titanic wissen wir alle dass der Kapitän stirbt :-)
- also auch hier trifft Hypothese 1 auf alle weiteren Kabinen zu.

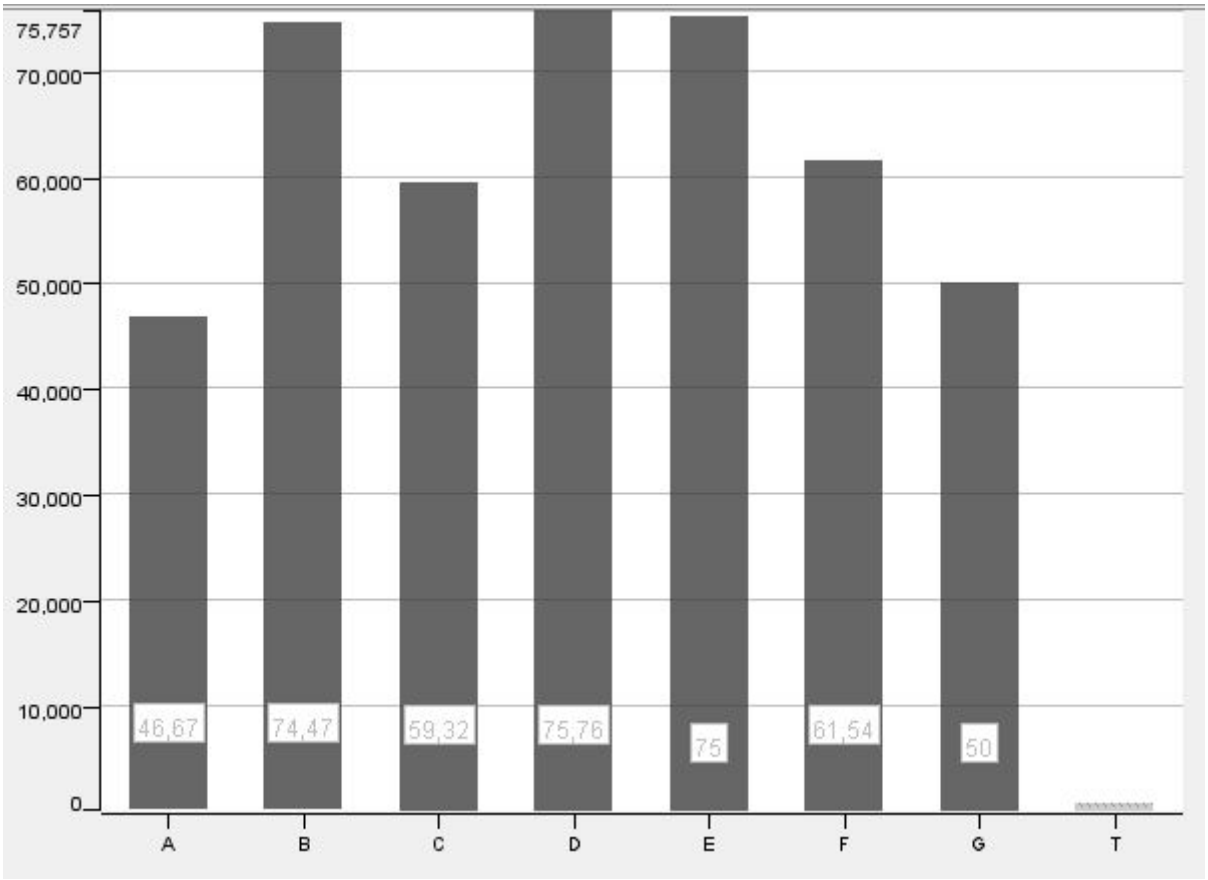
- Dieses Ergebnis zeigt, dass mit den aktuellen Daten keine ausreichende Aussage zu den Überlebenschancen bezogen auf die Kabine getroffen werden kann. Denn es haben ja nur Überlebende und deren Nachbarn/Mitbewohner eine Kabinennummer zugeordnet. Wenn also ein ganze “Batterie” an Kabinen keine Überlebenden hat würden diese Kabinen hier auch nicht auftauchen. Man müsste also eine Liste über alle vorhanden Kabinen haben, und könnte dann darauf schließen, dass alle fehlenden Kabinen eine “schlechte” Überlebenschance mit sich brachten.
- Mögliche Ursache für diese Auffälligkeit: Die Überlebenden haben vermutlich Vermisstenanzeigen bei der Rederei aufgegeben, dadurch sind die Kabinennummern der vermissten/verstorbenen Zimmergenossen & Nachbarn ebenfalls erfasst worden.

4. Dennoch soll überprüft werden, ob die Kabinen (Decks) und die Klasse einen Einfluss auf die Überlebensrate der Passagiere hatten. - Hypothese 2

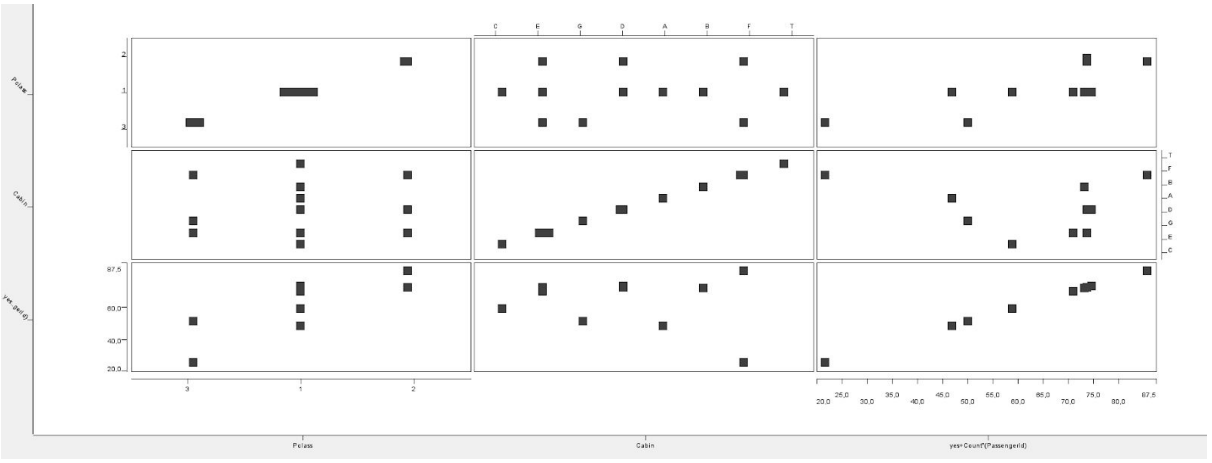
- Die Klasse hat einen Merkliche Einfluss auf die Überlebensrate, wie folgendes Diagramm zeigt.



- Die Kabine (Deck) hat auch einen merklichen Einfluss auf die Überlebensrate, Siehe folgende Grafik:



- in der Kombination von Klasse und Kabine sieht man, dass auch unterschiedliche Überlebensraten bestehen, garde in der Kabinenebene “F”, Siehe folgende Grafik:



- besonders gute Überlebenschancen hatte man demnach in der 1. Klasse auf Deck B; oder in der 2 Klasse in Kabine F; oder in allen Klassen auf Deck E; oder in der 1. & 2. Klasse auf Deck D.

Zusammenfassen kann durchaus eine Zusammenahng zur Überlebenschancen in den einzelnen Kabinen und Klassen getroffen werden. Wenngleich die Daten der Kabinen für eine Vorhersage der Überlebensrate auf Grund der vielen fehlenden Kabinennummern nur einen kleine bis mittlere Korrelation bieten. Die Klasse hingegen ist gut für eine vorhersage

geeignet, da sie fast immer vorhanden ist (war hier aber nicht im Scope, daher auch keine Berechnung).

Sehr interessant war die Hypothese, dass die vorhanden Kabinennummern vermutlich durch Befragung der Überlebenden erhoben wurden. Diese haben vermutlich ihre eigenen Kabinennummern und die Ihrer Mitreisenden und Nachbarn beigesteuert. Wer hätte das gedacht? :-)