

# Live - demo ChatGPT Advanced Data Analysis

Maarten de Borst  
Numbers



Model: DALL-E3 via ChatGPT | Prompt: Historical photograph of the Titanic, a symbol of luxury and engineering, sailing smoothly in a harbor, with the city skyline faintly visible in the background.

— Masterclass ChatGPT

# Maarten de Borst

- > Oprichter Numbers:
  - > Boutique advieskantoor AI, crypto & blockchain
    - > Consulting
    - > Opleidingen
    - > Publicaties
- > Opsporingsambtenaar Politie AA, Cybercrime & FinEC: digitale criminele geldstromen
- > Achtergrond financiële sector (PM Controlling, Accounting & Reporting)



— Masterclass ChatGPT

# ChatGPT

- > Tool voor het genereren van taal?!
- > Maar ook:
  - > Code
  - > Spraak
  - > Afbeeldingen
  - > En data-analyses!



— Masterclass ChatGPT

# Advanced Data Analysis

- > ChatGPT kan álle soorten data verwerken:
  - > Platte tekst
  - > Data (CSV, XLSX etc.)
- > Verwerking in Python



ChatGPT  
Advanced Data Analysis



— Masterclass ChatGPT

# Advanced Data Analysis

- > Data set passagiers Titanic 1912
- > TXT-bestand
- > Wat willen we hieruit halen?

Model: DALL-E3 via ChatGPT | Prompt: Historical photograph of the Titanic, a symbol of luxury and engineering, sailing smoothly in a harbor, with the city skyline faintly visible in the background.



— Masterclass ChatGPT

## Advanced Data Analysis

- > Hoe waarheidsgetrouw was de film nu écht?
- > Wat waren de overlevingskansen van de passagiers?
- > Doelen:
  - > Data - beschrijving
  - > Analyse
  - > Visu als
  - > Management summary
- > De hoofdrolspelers van vandaag...

AB Titanic.txt Document

Beschrijf deze dataset, Gebruik alleen tekst

Om een beschrijving van de dataset te kunnen geven, zal ik eerst de inhoud van het bestand moeten bekijken. Laten we beginnen met het lezen van de eerste paar regels om een idee te krijgen van de structuur van de gegevens.

Finished working Show work

De dataset bevat gegevens over de passagiers van de Titanic. Hier is een overzicht van de variabelen in de dataset:

1. **pclass**: Klasse van de passagier (1 = 1e klas, 2 = 2e klas, 3 = 3e klas)
2. **survived**: Overlevingsstatus (0 = overleden, 1 = overleefd)
3. **name**: Naam van de passagier
4. **sex**: Geslacht van de passagier (mannelijk of vrouwelijk)
5. **age**: Leeftijd van de passagier
6. **sibsp**: Aantal broers/zussen of echtgenoten aan boord
7. **parch**: Aantal ouders/kinderen aan boord
8. **ticket**: Ticketnummer
9. **fare**: Prijs van het ticket
10. **cabin**: Hutnummer
11. **embarked**: Haven van inschepping (C = Cherbourg; Q = Queenstown; S = Southampton)
12. **boat**: Reddingsbootnummer (indien overleefd)
13. **body**: Lichaamsnummer (indien overleden)
14. **home.dest**: Thuisbestemming

De dataset lijkt te zijn gescheiden door puntkomma's (;). Elk veld wordt omringd door dubbele aanhalingstekens als het een tekstveld is, terwijl numerieke velden geen aanhalingstekens hebben. Er zijn ook enkele ontbrekende waarden, aangeduid met "NA".

Regenerate

## — Masterclass ChatGPT

# Advanced Data Analysis

- > We laden de data in en vragen om een algemene beschrijving van de dataset
- > ChatGPT gaat rekenen, bekijkt de inhoud en schrijft deze overzichtelijk voor ons uit
- > De classificatie ziet er aannemelijk uit
- > Maar die zesde variabele – sibsp – is dat wel juist?





---

## Masterclass ChatGPT

# Advanced Data Analysis

- > Laten we het checken
- > Door ChatGPT simpelweg te vragen waar die classificatie vandaan komt
- > ChatGPT combineert de gegevens van de dataset met de algemene kennis die hij tot zijn beschikking heeft

AB Over die variabele sibsp; hoe weet je dit het aantal broers/zussen of echtgenoten aan boord is?

 De variabele "sibsp" staat voor "siblings/spouses", wat in het Engels vertaald kan worden als "broers/zussen en echtgenoten". Deze variabele geeft het aantal broers/zussen en echtgenoten van de passagier aan die aan boord van de Titanic waren.   

De informatie over de betekenis van de variabele "sibsp" en andere variabelen in de dataset komt vaak uit de beschrijving van de dataset of de documentatie die bij de gegevens hoort. In dit geval, omdat ik geen toegang heb tot de documentatie van deze specifieke dataset, baseer ik mijn uitleg op algemene kennis van bekende datasets over de Titanic. Het is mogelijk dat er in dit specifieke geval een andere definitie wordt gebruikt, maar dit zou dan in de bijbehorende documentatie moeten staan.



AB

**You**

Welke variabelen uit deze dataset denk je dat toepasbaar zijn op Jack en Rose uit de titanic-film? denk aan leeftijd, klasse etc

**ChatGPT**

In de film "Titanic" zijn de hoofdpersonen Jack en Rose fictieve personages. Gebaseerd op hun kenmerken in de film, kunnen de volgende variabelen uit een Titanic dataset op hen van toepassing zijn:

### 3. **name (naam):**

1. **p**
  - Jack Dawson en Rose DeWitt Bukater zijn hun volledige namen in de film.

### 4. **sex (geslacht):**

2. **s**
  - Jack: Man.
  - Rose: Vrouw.

### 5. **age (leeftijd):**

- Jack: Zijn exacte leeftijd wordt niet genoemd, maar hij wordt afgebeeld als een jonge man, mogelijk in zijn vroege twintiger jaren.
- Rose: Ze is 17 jaar oud in het verhaal.

### 6. **sibsp (aantal broers/zussen of echtgenoten aan boord):**

- Beide personages hebben geen broers, zussen of echtgenoten aan boord.

## — Masterclass ChatGPT

# Advanced Data Analysis

- > Welke variabelen uit onze set zijn toepasbaar op onze hoofdrolspelers?
- > Laten we het vragen

CU

Oke helder. Vertel me nu de belangrijkste conclusies van deze dataset. Antwoord alleen in tekst



Om de belar  
op de gegev

1. **Algemene Statistieken:**
  - Er zijn in totaal 1.309 passagiersgegevens in de dataset.
  - De gemiddelde leeftijd van de passagiers is ongeveer 29,9 jaar met een standaardafwijking van 14,4 jaar. De jongste passagier was ongeveer 2 maanden oud, terwijl de oudste 80 jaar oud was.
  - De meeste passagiers betaalden een tarief van ongeveer 33,3 eenheden (mogelijk in ponden of dollars), maar er is een brede spreiding in tarieven met sommige passagiers die niets betaalden en anderen die tot 512,3 eenheden betaalden.
2. **Overlevingspercentage:**
  - Van alle passagiers overleefde ongeveer 38,2% de ramp.
3. **Overleving per Klasse:**
  - Passagiers in de 1e klasse hadden de hoogste overlevingskans met ongeveer 61,9% overlevenden.
  - Dit percentage daalt naar 42,96% voor de 2e klasse en verder naar 25,53% voor de 3e klasse.
4. **Overleving op Basis van Geslacht:**
  - Vrouwelijke passagiers hadden een aanzienlijk hogere overlevingskans in vergelijking met mannelijke passagiers. Ongeveer 72,7% van de vrouwen overleefde, terwijl slechts 19,1% van de mannen dat deed.
5. **Leeftijdsverdeling:**
  - De meeste passagiers waren tussen de 21 en 39 jaar oud.
6. **Invloed van Familie:**
  - Passagiers die familieleden aan boord hadden (broers, zussen, ouders of kinderen) hadden een hogere overlevingskans. Ongeveer 50,3% van de passagiers met familie overleefde, vergeleken met 30,3% van degenen zonder familie aan boord.

Samenvattend toont de dataset duidelijke patronen in overlevingskansen op basis van klasse, geslacht en aanwezigheid van familie aan boord. Zoals vaak wordt opgemerkt in analyses van de Titanic-gegevens, hadden vrouwen en passagiers uit hogere klassen een grotere kans om te overleven.

Regenera

## — Masterclass ChatGPT

# Advanced Data Analysis

- > We kunnen nu allerlei vragen gaan bedenken
- > Maar waarom laten we ChatGPT dit niet voor ons doen?
- > En we vragen simpelweg om de belangrijkste conclusies op basis van deze dataset

AB

Super! Wat zijn de belangrijkste cijfers van deze dataset? Neem er tien. Laat zien in een tabel



Hier zijn de tien belangrijkste cijfers van de dataset:



1. **Aantal passagiers:** 1309
2. **Aantal overlevenden:** 500
3. **Aantal mannen:** 843
4. **Aantal vrouwen:** 466
5. **Gemiddelde leeftijd van de passagiers:** 29.88 jaar
6. **Jongste passagier:** 0.17 jaar
7. **Oudste passagier:** 80 jaar
8. **Gemiddelde ticketprijs:** 33.30
9. **Hoogste ticketprijs:** 512.33
10. **Laag**

Deze  
berek

Finist

Hier z

Statistiek	Waarde
Aantal passagiers	1309
Aantal overlevenden	500
Aantal mannen	843
Aantal vrouwen	466
Gemiddelde leeftijd van de passagiers	29.88 jaar
Jongste passagier	0.17 jaar
Oudste passagier	80 jaar
Gemiddelde ticketprijs	33.30
Hoogste ticketprijs	512.33
Laagste ticketprijs	0.00

## — Masterclass ChatGPT

# Advanced Data Analysis

- > Nu iets meer de diepte in
- > Wat zijn de belangrijkste cijfers van deze dataset
- > En genereer daar een tabel bij

AB Kan je deze tabel voor mij exporteren?

Finished working Show work ▾

Ik heb de tabel geëxporteerd naar een CSV-bestand. Je kunt het [hier](#) downloaden.

AB Bij nader inzien wil ik 'm ook als tekstbestand en als afbeelding. Kan je de tabel exporteren als CSV, TXT en JPEG?

Finished working Show work ▾

Ik heb de tabel geëxporteerd als CSV-, TXT- en JPEG-bestanden. Je kunt ze hier downloaden:

- [CSV-bestand](#)
- [TXT-bestand](#)
- [JPEG-bestand](#)

## — Masterclass ChatGPT

# Advanced Data Analysis

- > En als we die tabel later nog eens willen gebruiken, bijvoorbeeld in een presentatie..
- > Dan vragen we ChatGPT gewoon om die te exporteren
- > ChatGPT zet de tabel vervolgens om en exporteert die
- > En als we 'm nu ook als TXT-bestand en afbeelding willen?
- > Dan vragen we dit gewoon

AB

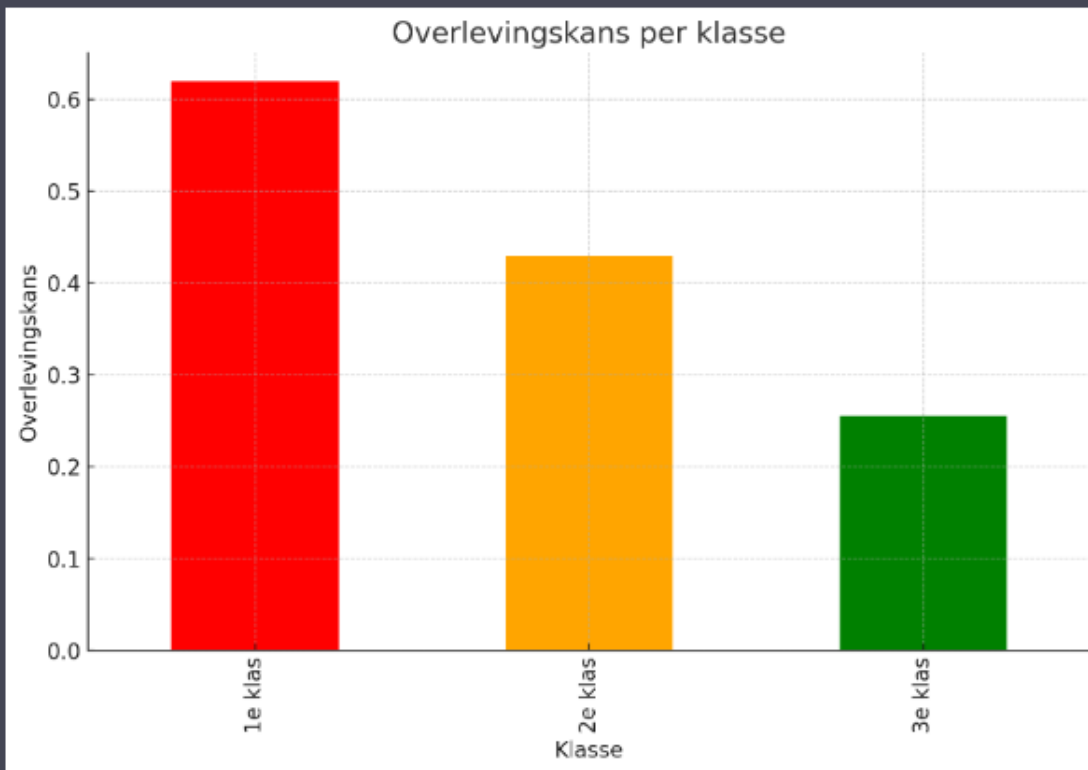
Dankje! Genereer grafieken waar de eerder genoemde vijf conclusies uit blijken



Om de vijf conclusies visueel weer te geven, zal ik de volgende grafieken maken:



1. **Sociaaleconomische status:** Een staafdiagram dat de overlevingskans per klasse toont.
2. **Geslacht en leeftijd:** Een staafdiagram dat de overlevingskans per geslacht toont en een boxplot die de leeftijdsverdeling per geslacht toont.



De eerste grafiek toont de overlevingskans per klasse. Zoals we kunnen zien, hadden passagiers in de eerste klas de hoogste overlevingskans, gevolgd door passagiers in de tweede en derde klas.

## — Masterclass ChatGPT

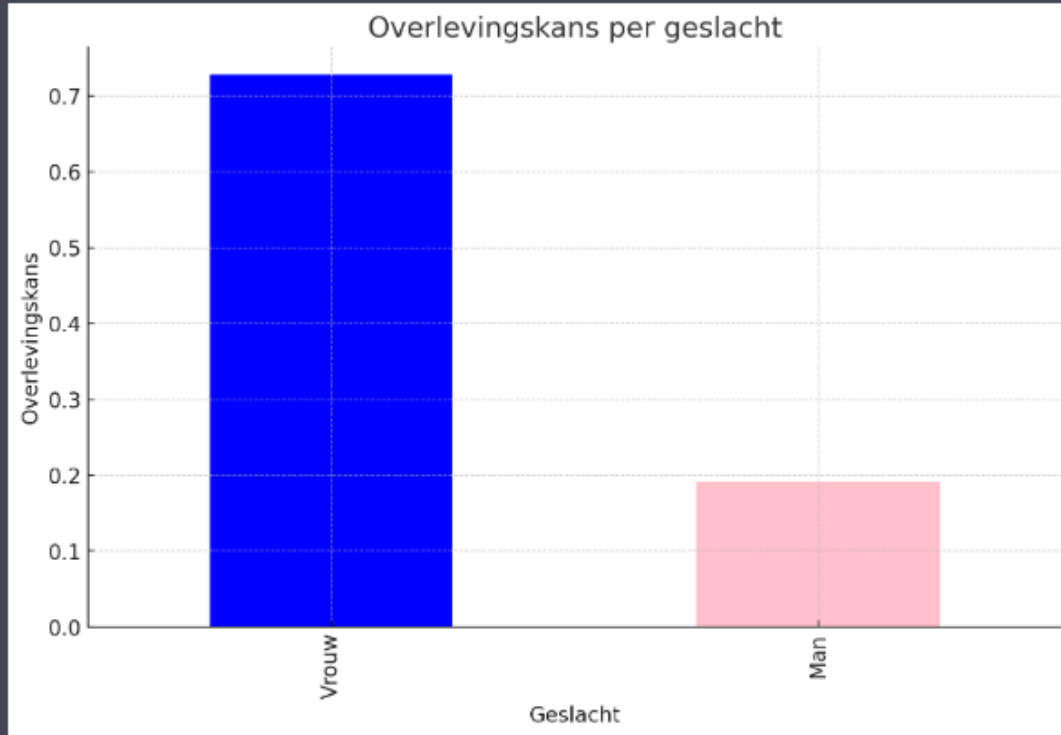
# Advanced Data Analysis

- > Maar uiteraard willen we dit ook visueel
- > En ook daarvoor geldt: je hoeft het maar te vragen
- > En ChatGPT genereert de ene grafiek na de andere...
- > Te beginnen met overlevingskans per klasse

Laten we nu de grafieken maken voor geslacht en leeftijd.

Finished working

Show work ▾



De tweede grafiek toont de overlevingskans per geslacht. Vrouwen hadden een aanzienlijk hogere overlevingskans dan mannen.

De derde grafiek toont de leeftijdsverdeling per geslacht in de vorm van boxplots. We kunnen zien dat de leeftijdsverdeling van mannen en vrouwen vergelijkbaar is, met een vergelijkbare mediaan en spreiding.

— Masterclass ChatGPT

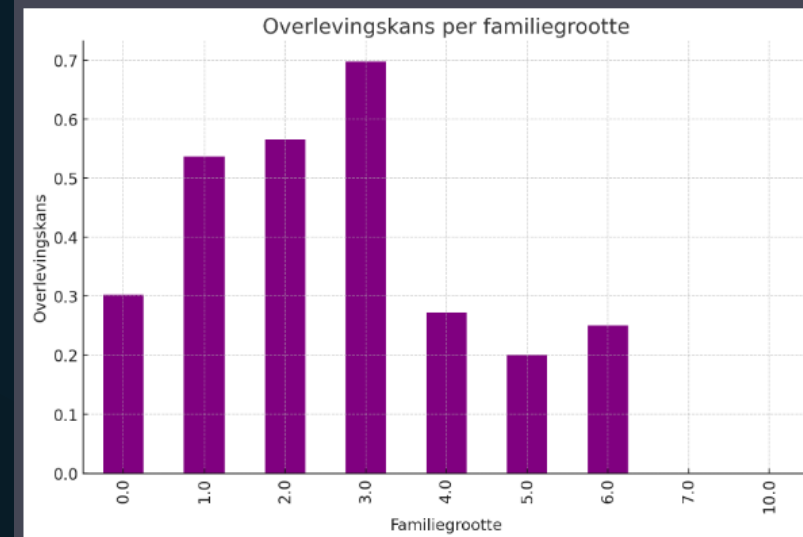
# Advanced Data Analysis

- > Naar geslacht
- > En familie grootte

Laten we nu de grafieken maken voor familie.

Finished working

Show work ▾



De vierde grafiek toont de overlevingskans per familie grootte. We kunnen zien dat passagiers met kleine gezinnen (1-3 familieleden) een hogere overlevingskans hadden dan passagiers zonder familie of met grote gezinnen.

AB

Nu ben ik wel benieuwd naar het aantal overlevenden/doden per leeftijdscategorie. Wat is een duidelijke grafiek om dat te laten zien?

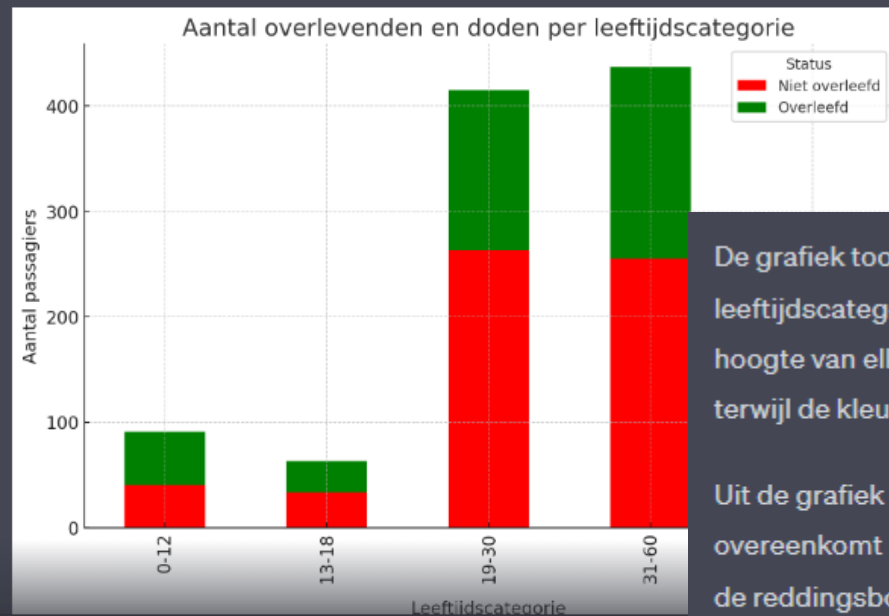


Een goede manier om het aantal overlevenden en doden per leeftijdscategorie te visualiseren is met een gestapelde staafdiagram. In dit type grafiek wordt elke leeftijdscategorie weergegeven als een aparte staaf, waarbij de hoogte van de staaf het totale aantal passagiers in die categorie weergeeft. De staaf wordt vervolgens opgedeeld in segmenten die het aantal overlevenden en doden vertegenwoordigen. Op deze manier kunnen we in één oogopslag zien hoe het aantal overlevenden en doden zich verhoudt binnen elke leeftijdscategorie.

Laten we deze grafiek maken.

Finished working

Show work ▾



De grafiek toont het aantal overlevenden (groen) en doden (rood) per leeftijdscategorie. De leeftijdscategorieën zijn onderverdeeld in vijf groepen: 0-12, 13-18, 19-30, 31-60 en 61-100. De hoogte van elke staaf vertegenwoordigt het totale aantal passagiers in die leeftijdscategorie, terwijl de kleuren aangeven hoeveel van die passagiers hebben overleefd of zijn overleden.

Uit de grafiek blijkt dat de overlevingskans voor jonge kinderen (0-12) relatief hoog was, wat overeenkomt met het algemene patroon dat kinderen voorrang kregen bij het instappen van de reddingsboten. Voor andere leeftijdscategorieën lijkt het overlevingspercentage vergelijkbaar.

Laat me weten als je nog verdere vragen hebt of als er iets anders is dat ik voor je kan doen.

## — Masterclass ChatGPT

# Advanced Data Analysis

- > Hoe zit dat per leeftijd?
- > ChatGPT 'bedenkt' zelf logische categorieën
- > En komt met een voorstel voor de meest logische grafiek
- > En sluit af met een beschrijving van wat we zien

AB

Oke duidelijk. Genereer nu een grafiek die onderscheid maakt tussen overlevenden/doden en de klasse waarin de passagiers zaten



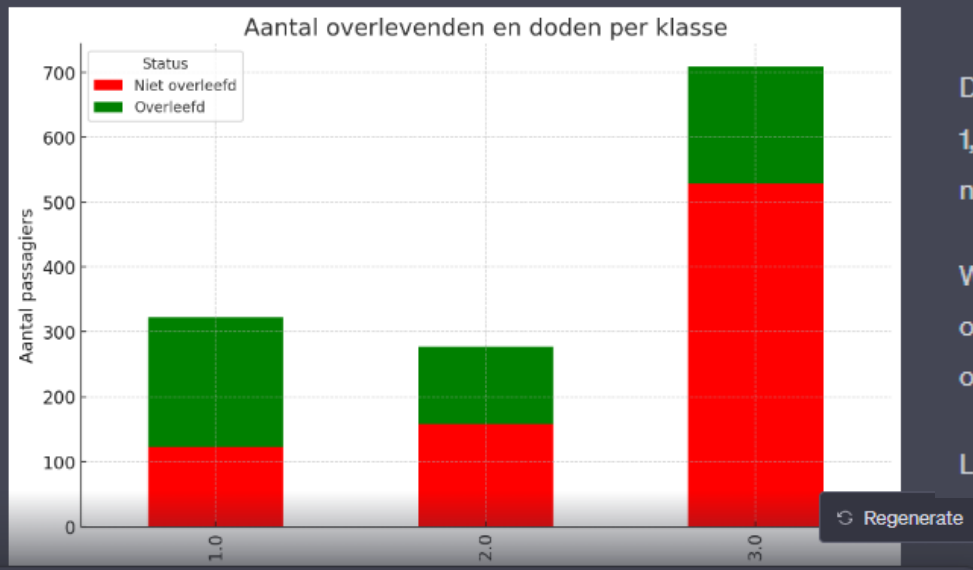
Om de relatie tussen overlevingsstatus en klasse te visualiseren, kunnen we gebruik maken van een gestapelde staafdiagram. In deze grafiek wordt elke klasse weergegeven als een aparte staaf, waarbij de hoogte van de staaf het totale aantal passagiers in die klasse weergeeft. De staaf wordt vervolgens opgedeeld in segmenten die het aantal overlevenden en doden vertegenwoordigen. Op deze manier kunnen we in één oogopslag zien hoe het aantal overlevenden en doden zich verhoudt binnen elke klasse.



Laten we de grafiek maken.

Finished working

Show work



## — Masterclass ChatGPT

# Advanced Data Analysis

- > En welke relatie zit er tussen passagiersklasse en de overlevingskans?
- > We hoeven het maar te vragen..

De grafiek toont het aantal overlevenden en doden per klasse. De klassen zijn aangeduid als 1, 2 en 3 op de x-as, terwijl het aantal passagiers op de y-as staat. De staafjes zijn gekleurd naar overlevingsstatus, waarbij rood staat voor niet-overleefd en groen voor overleefd.

We kunnen zien dat de eerste klasse de hoogste overlevingskans had, met meer overlevenden dan doden. In de tweede en derde klasse waren er meer doden dan overlevenden, waarbij de derde klasse de laagste overlevingskans had.

Laat me weten als je nog verdere vragen hebt of als er iets anders is dat ik voor je kan doen.



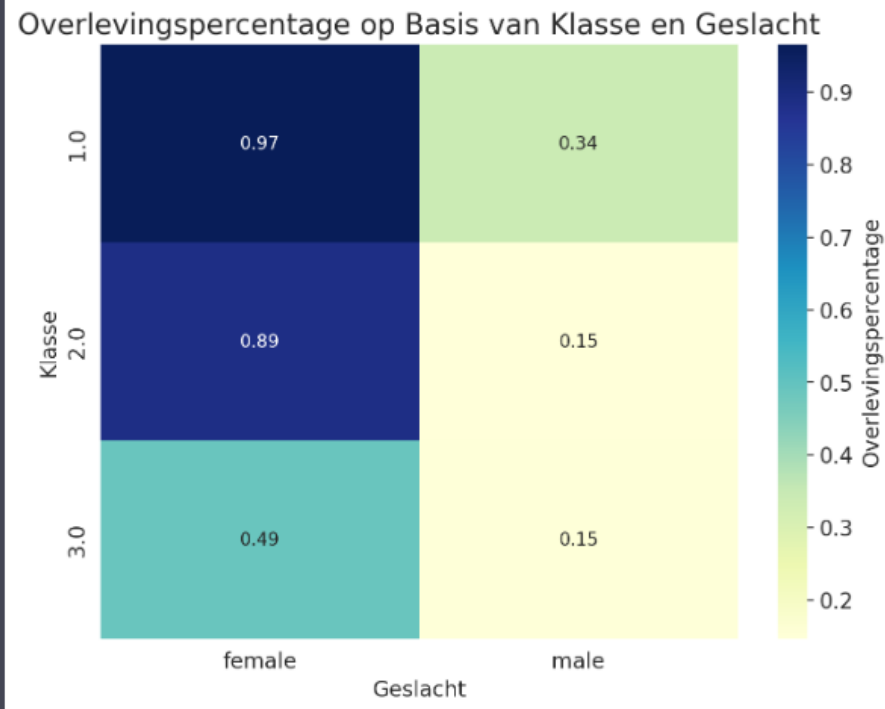
CU

Genereer nu een heatmap van overlevingskansen per klasse en geslacht



Finished working

Show work



Hier is een heatmap die de overlevingspercentages toont op basis van klasse en geslacht:

- De donkerdere kleuren duiden op hogere overlevingspercentages.
- Uit de heatmap blijkt dat vrouwelijke passagiers uit de 1e en 2e klasse de hoogste overlevingskansen hadden, met percentages van respectievelijk ongeveer 96,8% en 89,1%.
- Mannelijke passagiers uit de 3e klasse hadden daarentegen de laagste overlevingskansen met slechts ongeveer 15,2%.

## — Masterclass ChatGPT

# Advanced Data Analysis

- > Maar ChatGPT kan niet alleen grafieken en diagrammen maken
- > Maar is ook goed in boxplots, histogrammen en heatmaps

**You**

Schrijf nu een management summary van maximaal 200 woorden. Start met een korte inleiding over het zinken van de Titanic, vervolgens een aantal belangrijke conclusies uit de gemaakte analyse en sluit af met de statistische overlevingskansen van Jack en Rose

**Data Analysis**

De Titanic, een iconisch passagiersschip, zonk op 15 april 1912 na een botsing met een ijsberg, wat leidde tot één van de meest tragische maritieme rampen in de geschiedenis. Uit analyse van de beschikbare passagiersgegevens blijkt dat de overlevingskansen aanzienlijk verschilden op basis van geslacht, klasse en leeftijd. Vrouwen hadden een veel hogere overlevingskans dan mannen, wat de toenmalige reddingsprotocollen weerspiegelt waarbij vrouwen en kinderen voorrang kregen. Passagiers in de eerste klasse hadden aanzienlijk betere overlevingskansen in vergelijking met die in lagere klassen, wat wijst op sociale en economische ongelijkheden. Leeftijd speelde ook een rol, waarbij jongere passagiers over het algemeen betere overlevingskansen hadden.

Familiebanden bleken ook van invloed te zijn op de overlevingskansen. Passagiers met familieleden aan boord, vooral die met kinderen of echtgenoten, toonden een variërend overlevingspatroon. Economische factoren, zoals het betaalde tarief, waren eveneens indicatoren voor overleving, waarbij hogere tarieven betere overlevingskansen.

In de context van de fictieve personages Jack en Rose uit de film 'Titanic', zou statistisch gezien Rose (een jonge vrouw uit de eerste klasse) een hoge overlevingskans hebben, terwijl Jack (een jonge man uit de derde klasse) aanzienlijk lagere overlevingskansen zou hebben gehad, wat de historische gegevens en sociale normen van die tijd weerspiegelt.

## — Masterclass ChatGPT

# Advanced Data Analysis

> Om 'm a f t e m a k e n : d e m a n a g e m e n t  
s u m m a r y



## — Masterclass ChatGPT

# Advanced Data Analysis

- > En als toetje: een regressiemodel om de overlevingskans te voorspellen
- > De AI-tool neemt je vervolgens mee in de verschillende stappen

AB Kun je een regressiemodel maken om de overlevingskans te voorspellen op basis van onder andere leeftijd, geslacht en klasse?

Hier zijn de resultaten van het logistische regressiemodel dat we hebben gemaakt om de overlevingskans te voorspellen op basis van leeftijd, geslacht en klasse:

1. **Nauwkeurigheid:** Het model heeft een nauwkeurigheid van 73.8%, wat betekent dat het in 73.8% van de gevallen de juiste voorspelling heeft gedaan.
2. **Verwarringsmatrix:**

```
lua
```

Copy code

```
[[96, 24]
 [31, 59]]
```

De verwarringsmatrix toont het aantal true positives (96), false positives (24), false negatives (31), en true negatives (59).

1. **Classificatierapport:**

```
markdown
```




Copy code

	precision	recall	f1-score	support
0.0	0.76	0.80	0.78	120
1.0	0.71	0.66	0.68	90
accuracy			0.74	210
macro avg	0.73	0.73	0.73	210
weighted avg	0.74	0.74	0.74	210

Het classificatierapport toont verschillende evaluatiemetingen, waaronder precisie, recall, f1-score en support voor elk van de twee klassen (0 voor niet overleefd en 1 voor overleefd).

Regenerate



Het logistische regressiemodel dat we hebben gemaakt, heeft de volgende algemene vorm:   


$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)}}$$

Waar:

- $P(Y = 1)$  is de kans dat de passagier overleeft ( $Y=1$ ).
- $\beta_0$  is de intercept van het model.
- $\beta_1, \beta_2, \beta_3$  zijn de coëfficiënten voor de onafhankelijke variabelen  $X_1, X_2, X_3$ .
- $X_1$  is de leeftijd van de passagier.
- $X_2$  is het geslacht van de passagier, waarbij 0 staat voor man en 1 voor vrouw.
- $X_3$  is de klasse waarin de passagier zat (1, 2 of 3).

De waarden van  $\beta_0, \beta_1, \beta_2, \beta_3$  worden bepaald tijdens het trainen van het model op basis van de trainingsdata. Laten we de specifieke waarden van deze coëfficiënten voor ons model ophalen.

Finished working

Show work 

Voor ons specifieke model zijn de waarden van de intercept en de coëfficiënten als volgt:

- Intercept ( $\beta_0$ ): 2.07
- Coëfficiënt voor leeftijd ( $\beta_1$ ): -0.035
- Coëfficiënt voor geslacht ( $\beta_2$ ): 2.57
- Coëfficiënt voor klasse ( $\beta_3$ ): -1.13

Dus de formule van ons model is:

$$P(Y = 1) = \frac{1}{1 + e^{-(2.07 - 0.035 \times X_1 + 2.57 \times X_2 - 1.13 \times X_3)}}$$

Waar:

- $X_1$  is de leeftijd van de passagier.
- $X_2$  is het geslacht van de passagier (0 voor man, 1 voor vrouw).
- $X_3$  is de klasse waarin de passagier zat (1, 2 of 3).

## — Masterclass ChatGPT

# Advanced Data Analysis

> Inclusief de formule van het logistische regressiemodel...



Model: DALL-E3 via ChatGPT | Prompt: Historical photograph of the Titanic, a symbol of luxury and engineering, sailing smoothly in a harbor, with the city skyline faintly visible in the background.

## — Masterclass ChatGPT

# Advanced Data Analysis

- > Dit brengt omgaan met data naar het volgende niveau
- > Één tool voor:
  - > Structureren
  - > Analyseren
  - > Visualiseren
  - > Beschrijven van data
- > Iedereen is een data-analist!

— Masterclass ChatGPT

# Kansen voor de financial



- > Ma ken va n co m p l e x e a n a l y s e s
- > H u l p b i j a u d i t s
- > F o r e c a s t i n g o . b . v . v e r s c h i l l e n d e s c e n a r i o ' s
- > M o n i t o r e n v a n m o g e l i j k e f r a u d e
- > A u t o m a t i s e r e n v a n w e r k p r o c e s s e n
- > B u d g e t t e r i n g
- > B o u w e n v a n r i s i c o m o d e l l e n
- > A n a l y s e r e n v a n f a c t u r e n
- > A u t o m a t i s c h e f i n a n c i a l r e p o r t i n g
- > ...

— Masterclass ChatGPT

Dit is een waanzinnige tijd om  
een financieel te zijn!



— Masterclass ChatGPT

# Da n k!

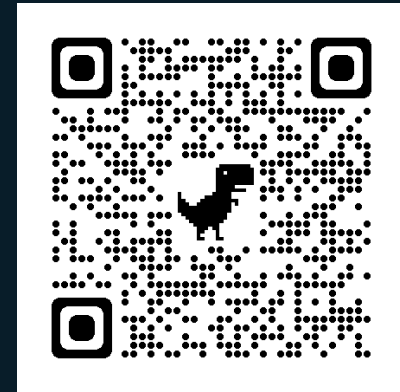
E- m a il:

[maarten@numbers.amsterdam](mailto:maarten@numbers.amsterdam)

Telefoon:

06- 15.53.91.19

Linke d in :





# Numbers leidt financials op

## ChatGPT voor Financials

- > Introductie in AI
- > ChatGPT
- > Bedreigingen voor de organisatie
- > Kansen voor de organisatie
- > Concrete handvatten

## De toekomst van geld

- > Introductie crypto
- > Blockchain-technologie
- > Smart contracts
- > De digitale euro
- > De laatste ontwikkelingen

## De impact van digitale valuta

- > Verdieping toekomst van geld
- > Trends in de markt
- > Impact op de financiële sector
- > Aanstaande regelgeving
- > Criminaliteit

## Crypto door de bril van de financial

- > Introductie crypto
- > Crypto en accountancy
- > Fiscaal-juridische implicaties
- > Wetgeving
- > Fraude en criminaliteit

## In-company op maat

- > Onderwerpen naar keuze
- > In-company voor de hele afdeling of organisatie
- > Fysiek of digitaal
- > Flexibele duur

## Cryptocurrency forensics

- > Introductie crypto
- > Criminaliteit
- > Blockchain-technologie
- > Rechercheren op de blockchain
- > Casuïstiek