Titanic dataset

Verslag

Data Science

Witse Meeussen

Thomas De Dobbeleer

Klaas Eelen

Academiejaar 2021-2022

Campus Geel, Kleinhoefstraat 4, BE-2440 Geel

INHOUDSTAFEL

1 OVER DE DATASET

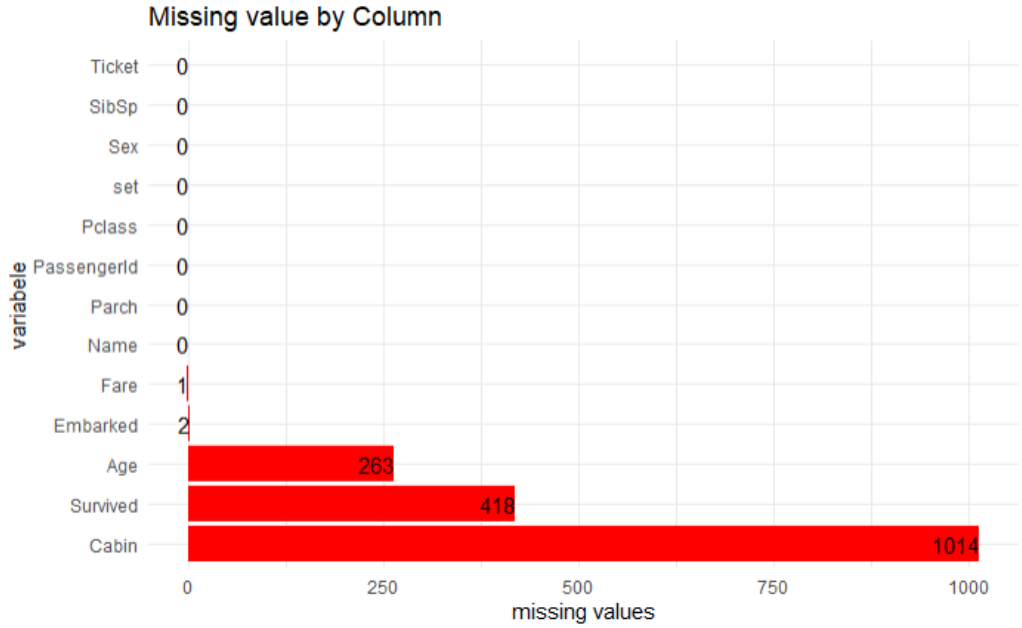
De Titanic dataset bevat data van de passagiers aan boord van de Titanic, het beruchte cruiseschip dat in april van 1912 zonk na een aanvaring met een ijsberg. Deze dataset bevat de info van de passagiers, en of ze de ramp overleefde of niet.
Deze dataset is dan ook bekend in de AI-wereld en wordt veel gebruikt voor het oefenen van machine learning. De opdracht is dan om een model te creëren dat kan voorstellen of een passagier de befaamde ramp heeft overleeft of niet op basis van hun leeftijd, klasse, geslacht, etc. Deze data wordt dan opgesplitst in een training- en testset.

Wij gaan echter geen machine learning maar data analyse op toepassen. Dit wilt dus zeggen dat wij geen modellen gaan maken, maar grafieken. De training/test splitsing hebben we dus ook niet nodig en dus voegen wij deze twee terug samen in één set.

2 ANALYSE

2.1 Missing data

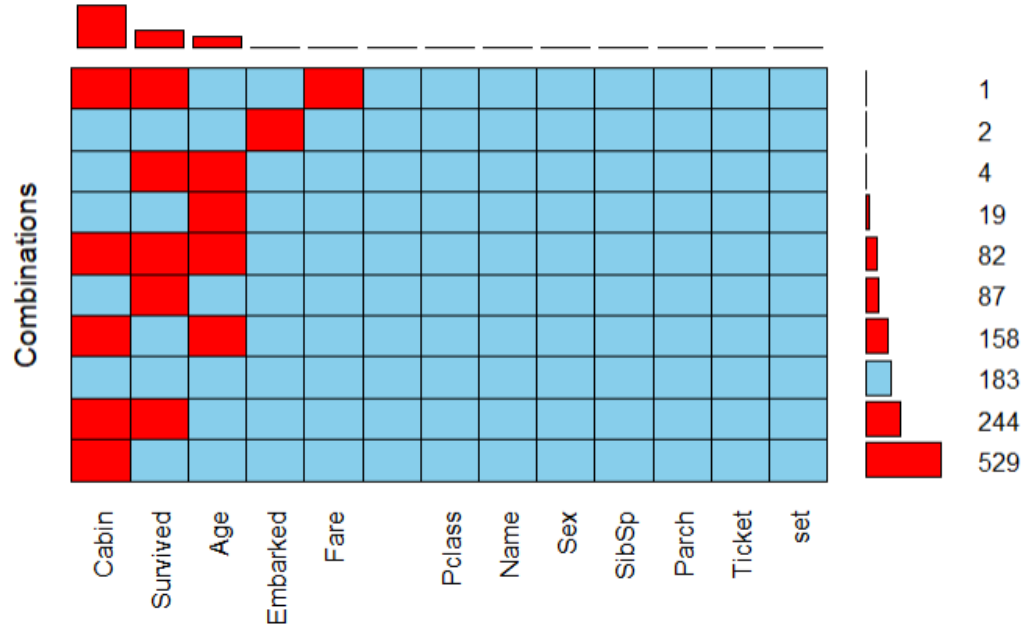
Tijdens het analyseren merkte we dat er hier en daar toch wat datapoints mistte. Na een beetje zoeken hebben we een manier gevonden om deze data te visualiseren. Hieronder kan je duidelijk zien wat er allemaal ontbreekt.



Het meest prominente is de kajuit van elke passagier, slechts iets meer dan 20% is bekend. Dit vormt echter niet echt een probleem bij onze analyse aangezien wij vooral focussen op de overlevingsgraad op basis van geslacht, klasse, grootte van gezin, etc. Stel dat we een vergelijking maakte met overlevingsgraad en locatie op het schip, dan hadden we in de problemen gezeten.

De andere twee waar er veel data ontbreekt zijn de leeftijd en of de passagier wel of niet de ramp overleefd heeft.
De leeftijd hebben we eenvoudig kunnen oplossen door behulp van het gemiddelde te pakken. Vervolgens hebben we de leeftijden opgedeeld in groepen
Of de passagier wel of niet overleefd heeft is echter wel van cruciaal belang. Helaas kunnen we hier weinig aan veranderen dus nemen we dit maar mee.

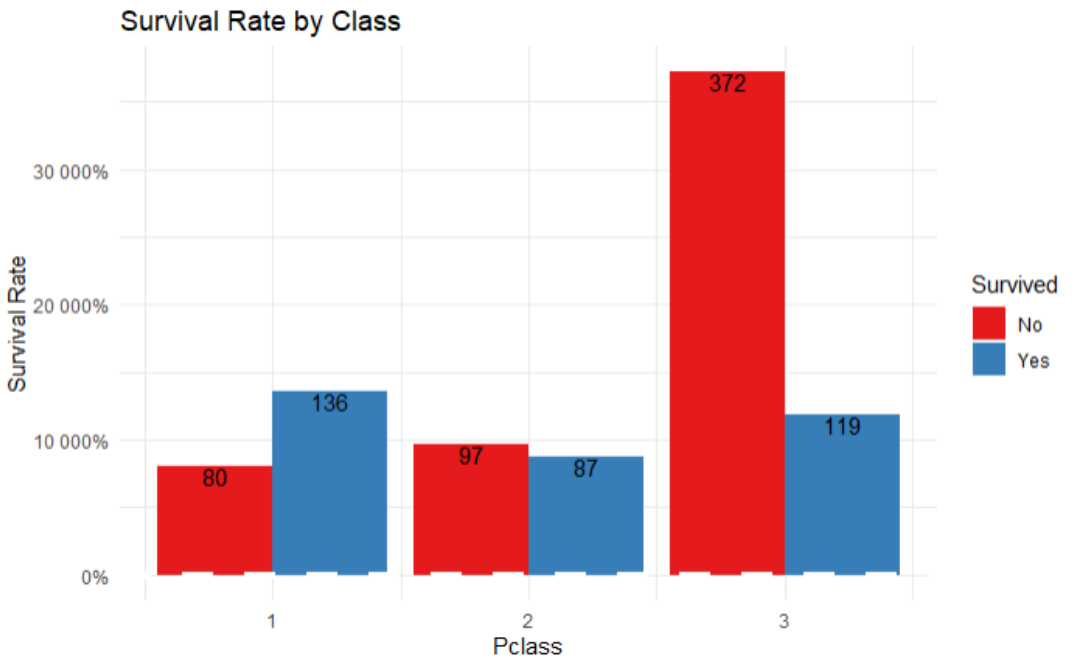
Vervolgens zijn we dieper gaan kijken welke data er exact ontbrak. Hieronder kan je een visualisatie zien van de verschillende combinaties van ontbrekende data in de records. Op de bovenste x-as geeft weer wat de vorige grafiek liet zien. De rechtse y-as geeft weer hoeveel entries er zijn met die exacte combinatie van missing data. Bijvoorbeeld; er zijn 529 entries waar enkel de kajuit ongekend is, 244 waar zowel kajuit alsook of ze het overleefd hebben niet gekend is, 183 passagiers waar we alles van weten, enz. Wat dus wilt zeggen dat de meerderheid van de passagiers wel iets ontbrekend heeft in hun data.



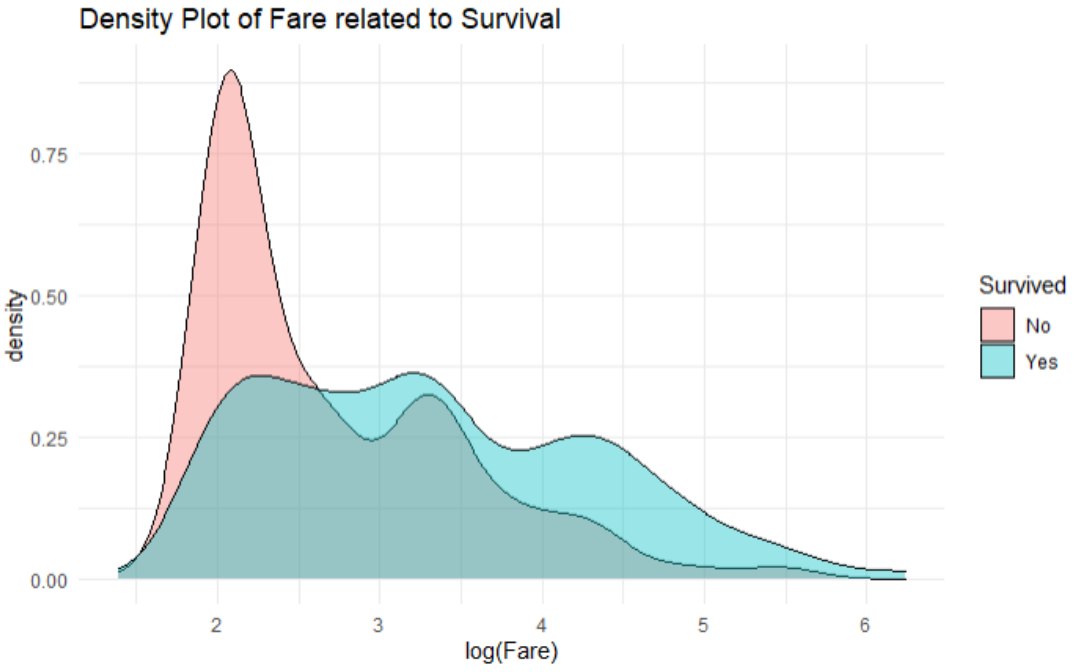
2.2 Klasse & Tarief

Iedereen weet dat de passagiers aan boord van de Titanic verdeeld waren in drie klassen; de Derde Klasse waren de armste, die naar Amerika vertrokken om een beter leven te lijden. De Tweede Klasse waren de middenstand en de Eerste Klasse waren de rijkste mensen, voor wie de reis meer een pleziertripje was dan een manier van transport.

De visualisatie toont duidelijk dat de Eerste Klasse, en dus de allerrijkste, wel degelijk de voorkeur kregen als het om de overlevingslotto ging. Meer dan de helft van deze passagiers overleefde de ramp, terwijl dat de passagiers uit Derde Klasse aanzienlijk onder het gemiddelde zitten. Tenslotte zien we dat het bij mensen uit tweede klasse een muntworp was of ze al dan niet levend naar huis gingen.



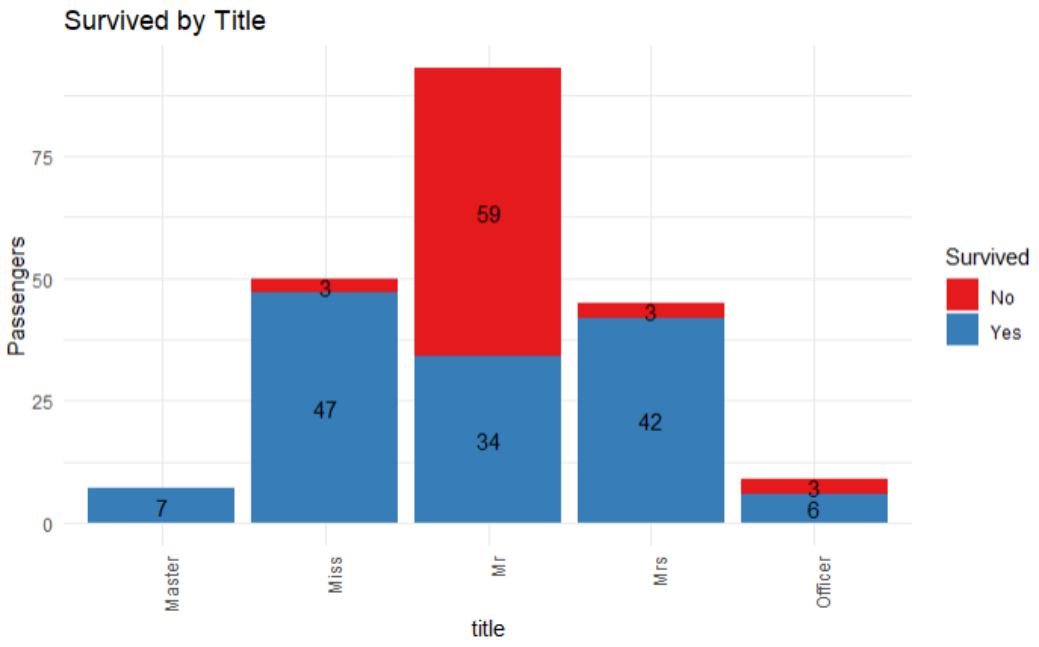
Deze trends zien we ook terug als we de overlevingskans plotten tegenover het tickettarief. Bij de goedkoopste tickets, Derde Klasse dus, zien we dat het merendeel de maagdenreis niet overleefde. Bij de middelste tarieven zijn overlevingskans en dodental vrijwel gelijk, en hoe duurder de tickets, hoe minder doden.



2.3 Titel

Iets handig wat we hebben opgemerkt is dat er in de namen van de passagiers ook hun titel verwerkt is (Mr, Mrs, Dr, Countess, enz.). Met behulp van regular expressions is het ons dan ook gelukt om deze titels af te zonderen en hiermee dan een visualisatie te maken. Omdat er veel verschillende titels waren hebben we deze verdeeld in 5 groepen. Hiermee krijgen we vat op zowel het aantal soort passagier er aanwezig was, alsook hun overlevingskans.

We zien dat het voor vrouwen weinig uitmaakte of je nu gehuwd was of niet, terwijl je als man meer kans had om te overleven als je een hogere status had, die relatie zagen we ook in de vorige relatie.

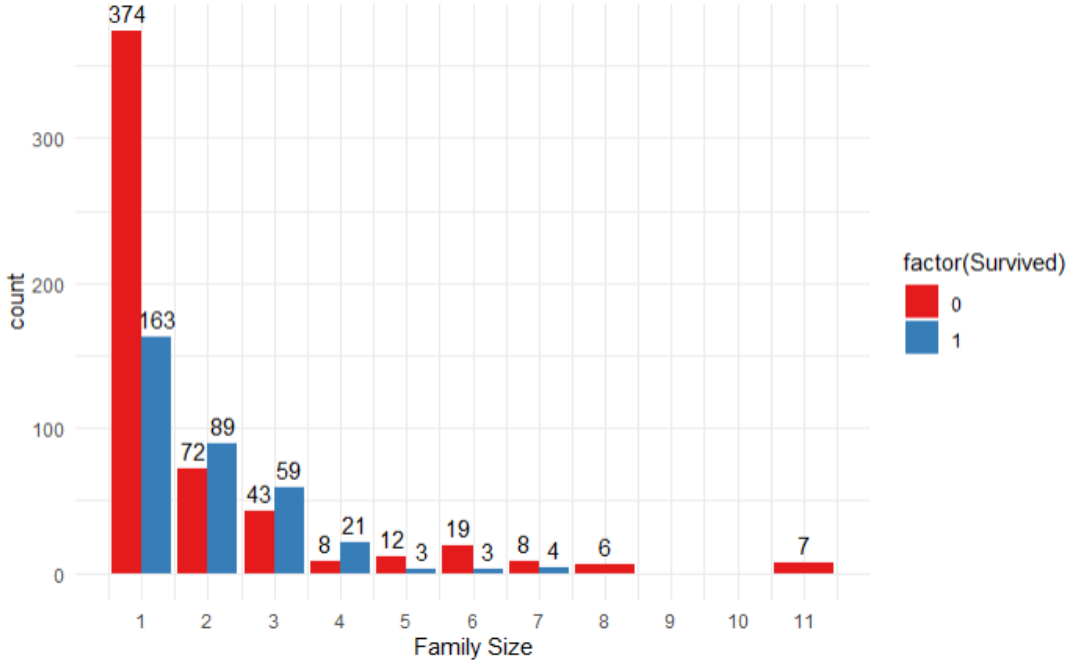


2.4 Gezinsgrootte

Niet elke passagier reisde alleen. We wilden weten of de grootte van het gezin ook een invloed had op je overlevingskans.
Als we de grafiek hieronder bekijken zien we dat als je alleen was, er een grote kans was dat je de ramp niet kan navertellen. Dit kan echter een bias zijn omdat er in totaal gezien gewoon veel meer alleen-reizigers waren dan gezinnen.

De overlevingskans verschuift echter naar de positieve kant bij gezinnen van 2, 3 en 4 personen. Bevat het gezin meer mensen dan dit, is de kans op overleving terug kleiner.

Opvallend is wel dat er geen gezinnen waren met 9 of 10 personen, maar wel enkele van 11, helaas overleefde geen compleet gezin van deze grootte.

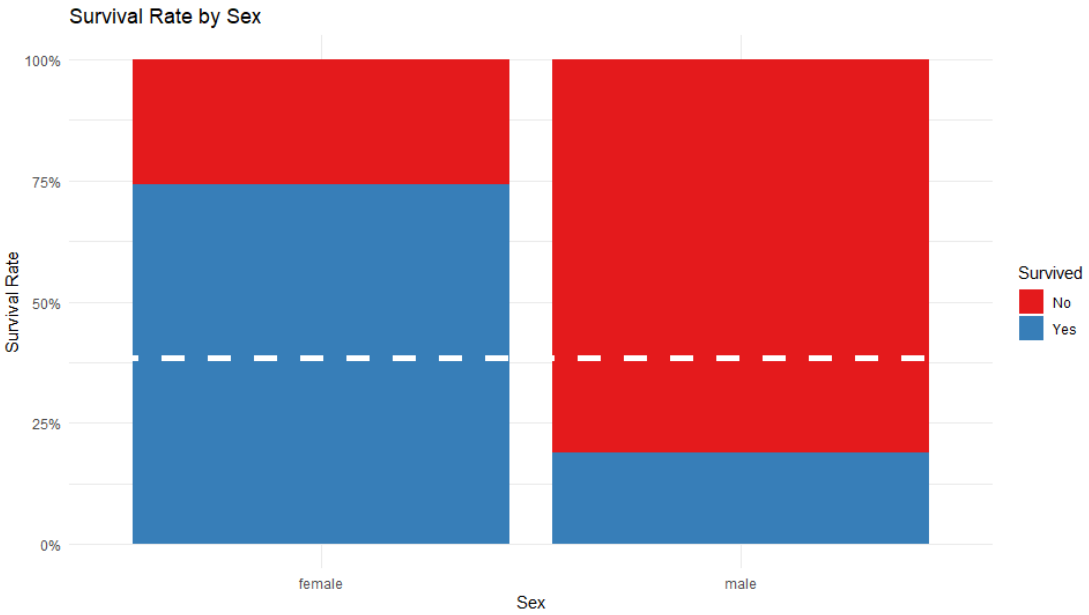


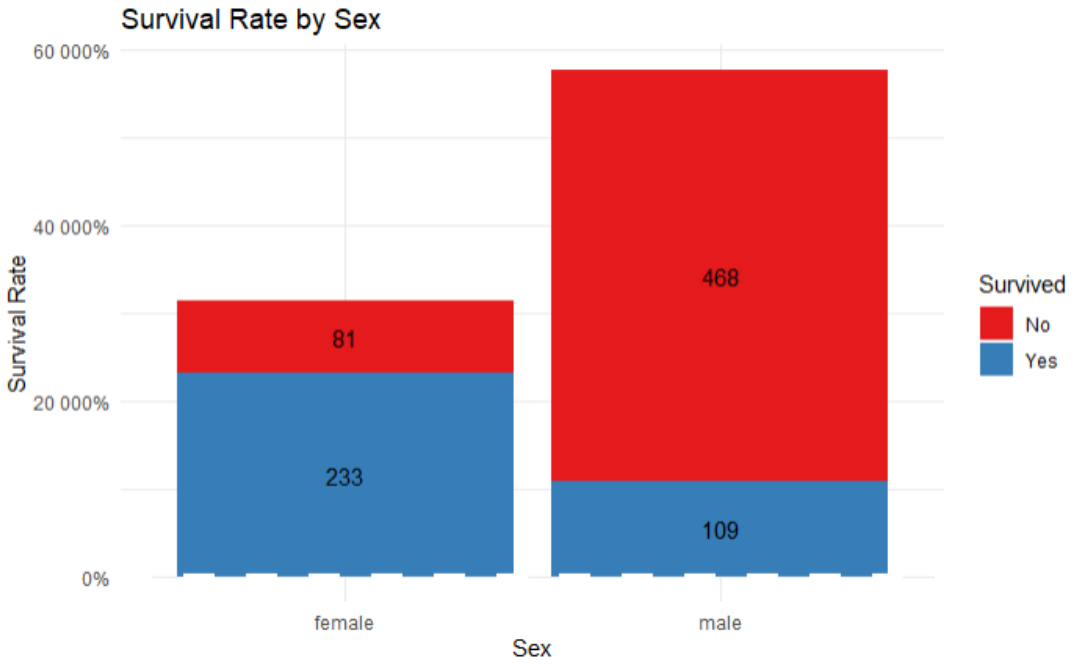
2.5 Geslacht

En last but not least kijken we nog naar het geslacht. Waarschijnlijk de meest onderzochte vergelijking. We hebben deze visualisatie opgedeeld in twee; eentje waar de overlevingsgraad makkelijker te zien is in verhouding en de andere met totale waardes.

Wat we overduidelijk kunnen zien is dat het bekende gezegde; "vrouwen en kinderen eerst" wel meer is dan een filmcliché. Driekwart van de vrouwelijke passagiers overleefde de zeemansdood, terwijl je als man minder dan 25% kans had op overleving.

Ten tweede zien we ook dat er bijna dubbel zoveel mannen aan boord waren dan vrouwen. Dit kan dan weer invloed hebben op de overlevingskansen. Gendergelijkheid was in die tijd nog niet echt ingeburgerd zo te zien.





3 CONCLUSIE

We kunnen concluderen dat over het algemeen de meeste vrouwen en kinderen de ramp hebben overleefd, alsook de bemanningsleden. De mannen waren bij de minderheid als het ging over levend uit de ramp te komen.

De klasse alsook de grootte van de familie had wel degelijk een impact op de overlevingskans. Hoe dit precies komt kunnen we niet uit de data halen.