Zorg-Los EDA Rapport

Tussentijdse opdracht 1

Sjoerd van Dorp

1032109

BIMCVD10R3

Inhoud

[Technisch 3](#_Toc151479533)

[Informatie behoefte business 4](#_Toc151479534)

[Bronnenlijst 5](#_Toc151479535)

[Bijlagen 6](#_Toc151479536)

# Technisch

1. **Geef aan hoe je de EDA hebt uitgevoerd en met welke tools.**

De EDA is uitgevoerd op een dataset om een inzicht te krijgen in de gegevens. Mijn aanpak was het gebruik van verschillende tools en libraries in Python. Deze libraries zijn:

* numpy voor numerieke bewerkingen
* pandas voor gegevensmanipulatie en -analyse
* seaborn voor datavisualisatie
* matplotlib.pyplot voor het maken van plots
* scipy.stats voor statistische functies

Ik begon met het inlezen van de dataset (train.csv') met behulp van de read\_csv-functie van pandas.

Om een overzicht te krijgen, heb ik de info()-functie gebruikt voor informatie over de kolommen en datatypes. Daarna heb ik statistische waarden afgedrukt met describe().

Daarna heb ik van de kolommen: stay, age, Hospital\_type\_code, Hospital\_region\_code, Department, ward\_type, Ward\_Facility\_Code, Type of Admission en Severity of Illness de waarden omgezet naar numerieke waarden

Ik heb de verdeling van kolom ('Admission\_Deposit') gevisualiseerd met een histogram en een normaal verdelingsplot.

Voor het aanpakken van ontbrekende waarden heb ik het aantal per kolom geïdentificeerd en vervolgens de lege velden ingevuld met de meest modus.

Ik heb het aantal unieke waarden per kolom afgedrukt om inzicht te krijgen in de variabiliteit van de gegevens.

Voor het begrijpen van onderlinge relaties heb ik een correlatiematrix gemaakt en visueel weergegeven met een heatmap.

Ik heb 4 staafdiagrammen van de kolommen: Hospital\_type\_code, Ward\_Type, Bed Grade, Age gemaakt om inzichten te krijgen in de meest voorkomende waarden.

Uiteindelijk heb ik de verwerkte dataset opgeslagen als Zorg-LOS.csv

1. **Geef aan hoe je structureel transformatie en EDA kan uitvoeren (dus niet eenmalige handmatige handelingen).**

Om EDA op een meer systematische en herhaalbare manier uit te voeren, kan er gebruik gemaakt worden van Python-scripts en -pipelines. Hiermee kan je de stappen in een EDA- en transformatieproces gestructureerd doorlopen en deze gemakkelijk herhalen voor verschillende datasets, zonder dat er elke keer handmatig dezelfde handelingen verrichte hoeft te worden.

Een manier om dit te doen is met de AutoViz library in Python.

“AutoViz is a library in Python that can be used to automatically generate visualizations for a given dataset. It can be used to quickly get a visual overview of the data, making it easier to perform exploratory data analysis.” Gupta, R. (2023, 28 maart).

Deze library genereert meerdere visualisaties dat inzichten geeft van de data.

# Informatie behoefte business

1. **Wat zijn de verschillende EDA uitkomsten op gebruikelijke categorieën (bv. lege velden etc.)? Doe hiervoor onderzoek want er zijn er een aantal.**

Ontbrekende Waarden:

1. Bed Grade:

Aantal Ontbrekende Waarden: 113

Behandeling: Het invullen van ontbrekende waarden met de meest voorkomende waarde (modus) of gebruik een methode zoals voorspelling op basis van andere kenmerken.

1. City\_Code\_Patient:

Aantal Ontbrekende Waarden: 6689

Behandeling: Voor deze missende waarden is er geen goeie oplossing.

1. Stay:

Aantal Ontbrekende Waarden: 137057

Behandeling: Het invullen van ontbrekende waarden met de meest voorkomende waarde (modus) of gebruik een methode zoals voorspelling op basis van andere kenmerken.

1. **Beschrijf het data cleaning en preparation proces indien van toepassing.**

Ik begon met het importen van de train dataset en deze op te splitsen tot een train en test set. Door dit bestand op te splitsen kon ik beter resultaat krijgen op de uitgevoerde EDA.

Verder heb ik van de stay, age, Hospital\_type\_code, Hospital\_region\_code, Department, ward\_type, Ward\_Facility\_Code, Type of Admission en Severity of Illness de waarden omgezet naar numerieke waarden om beter een analyse te kunnen maken.

1. **Beschrijf en visualiseer de conclusie en aanbeveling op business, technisch en data niveau**

Verder hebben de bedden niet goed gescoord en zouden deze beter moeten scoren, zie bijlage 1: Staafdiagram Bed Grade.

De grafiek van de leeftijd/age kolom laat zien dat de modus van de leeftijd van de patiënten 41-50 is, wat kan wijzen op een specifieke behoefte aan gezondheidszorgdiensten in deze leeftijdsgroep, wat belangrijk kan zijn voor de planning van gezondheidszorgfaciliteiten en -diensten. Zie bijlage 2: Staafdiagram Age

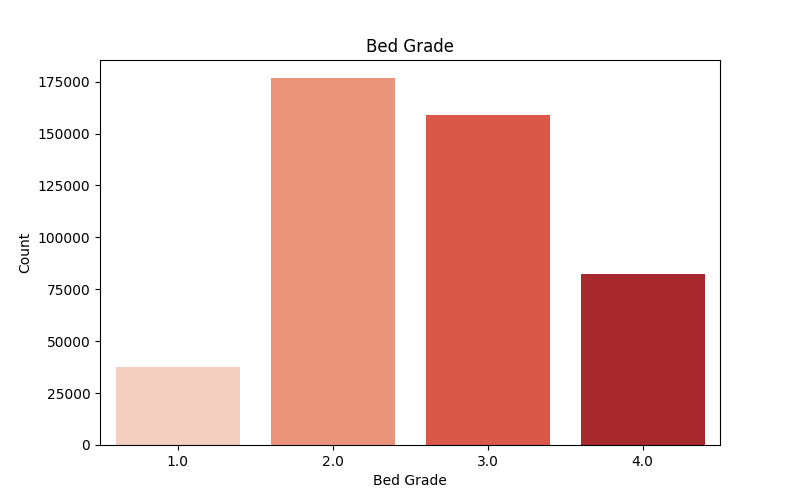
Op de heatmap, zie bijlage 5: Heatmap, is te zien dat er op plekken zoals Visitors with Patien – Admission\_Deposit het nog niet helemaal goed gaat maar op plekken zoals Hospital\_code-City\_Code\_Hospital wel.

In de boxplot is te zien dat de waarden van de vrije kamers en bezoekers per patient dicht bij elkaar liggen, maar er wel een aantal flinke uitschieters te zien zijn waar rekening mee gehouden moet worden. Zie: Bijlage 7: Boxplot.

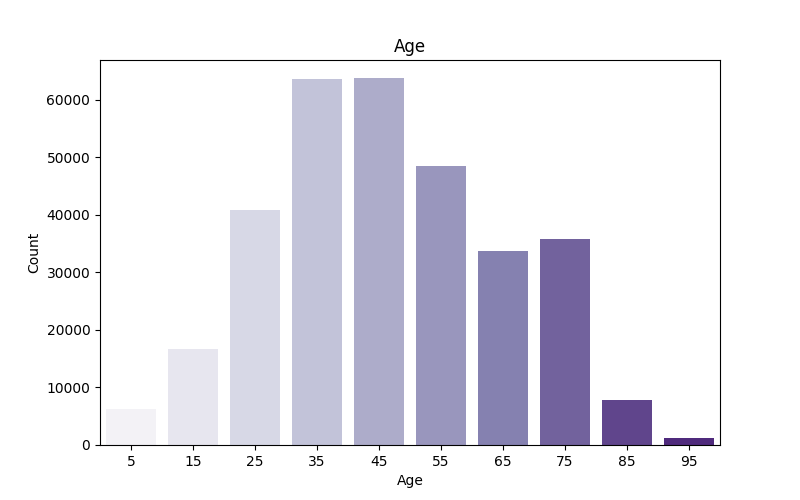
# Bronnenlijst

Gupta, R. (2023, 28 maart). 10 automated EDA tools that will save you hours of work. *Medium*. <https://medium.com/@riteshgupta.ai/10-automated-eda-tools-that-will-save-you-hours-of-work-b3edf3740fc2>

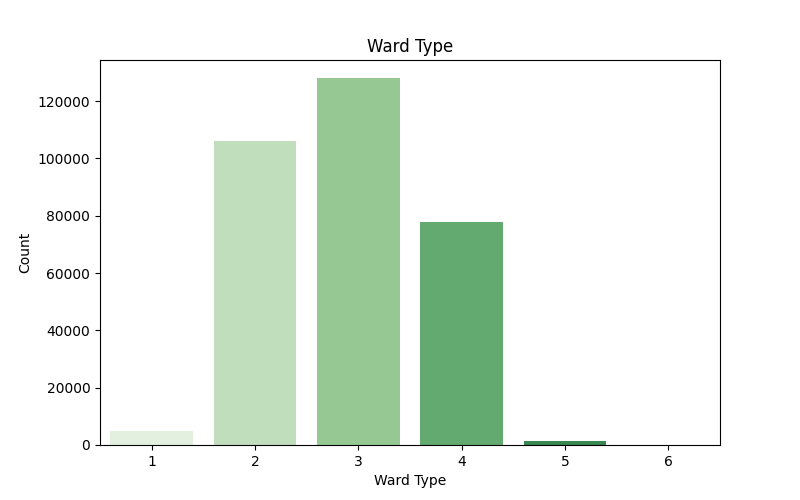
Kidu, T. (2023, 16 oktober). Data analysis with a single line of code using advanced Python libraries: automate your EDA and understand the data quickly. *Medium*. <https://medium.com/@ThomTechSavvy/data-analysis-with-a-single-line-of-code-using-advanced-python-libraries-automate-your-eda-and-e1e3fb2ed110>

Bijlagen

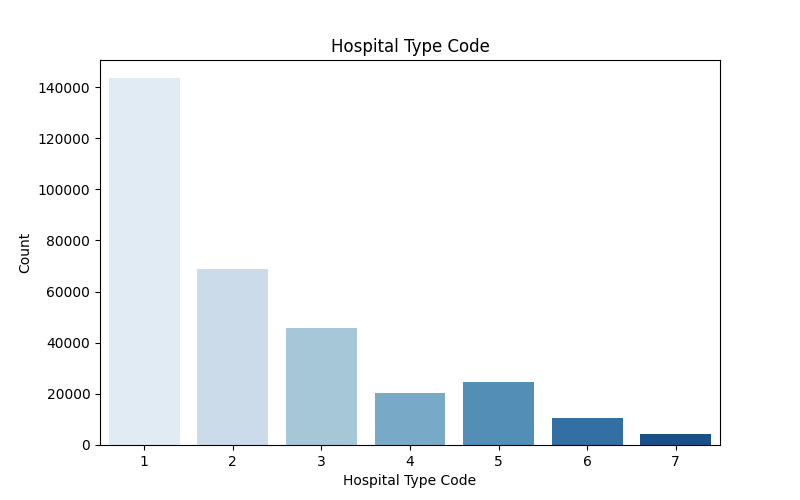
Bijlage 1: Staafdiagram Bed Grade



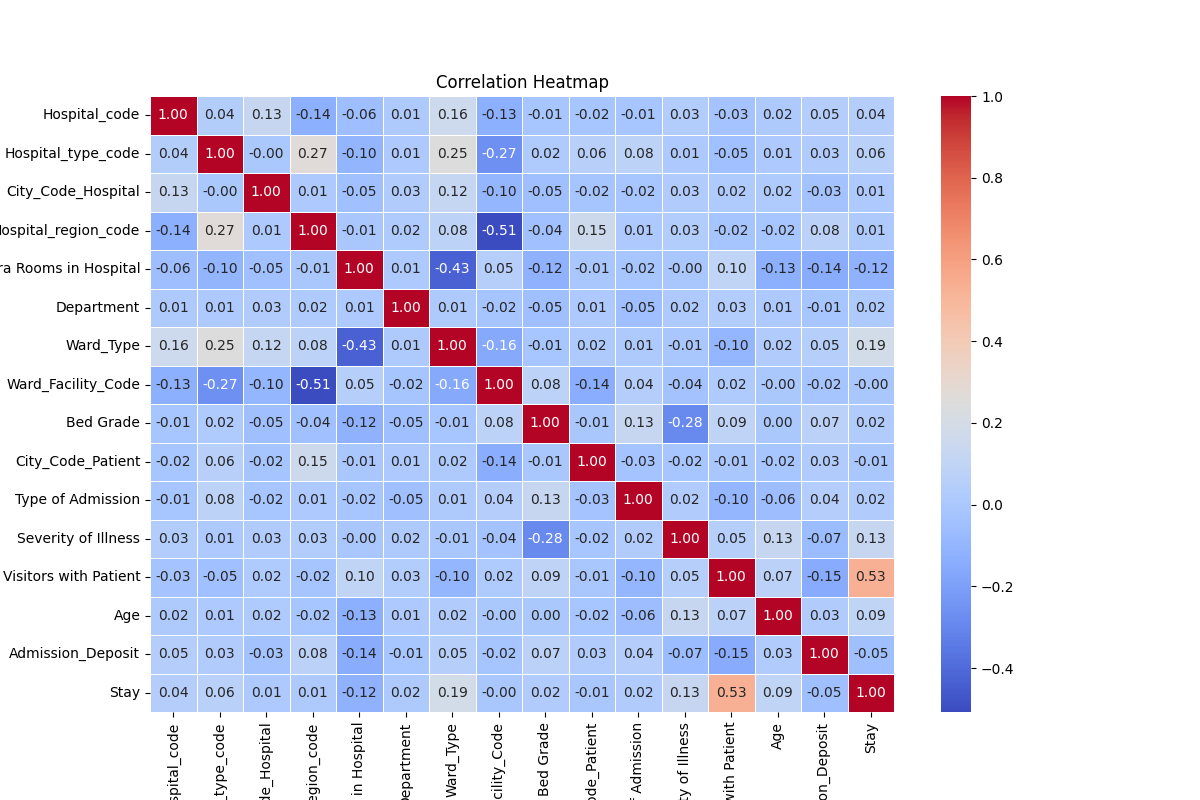
Bijlage 2: Staafdiagram Age



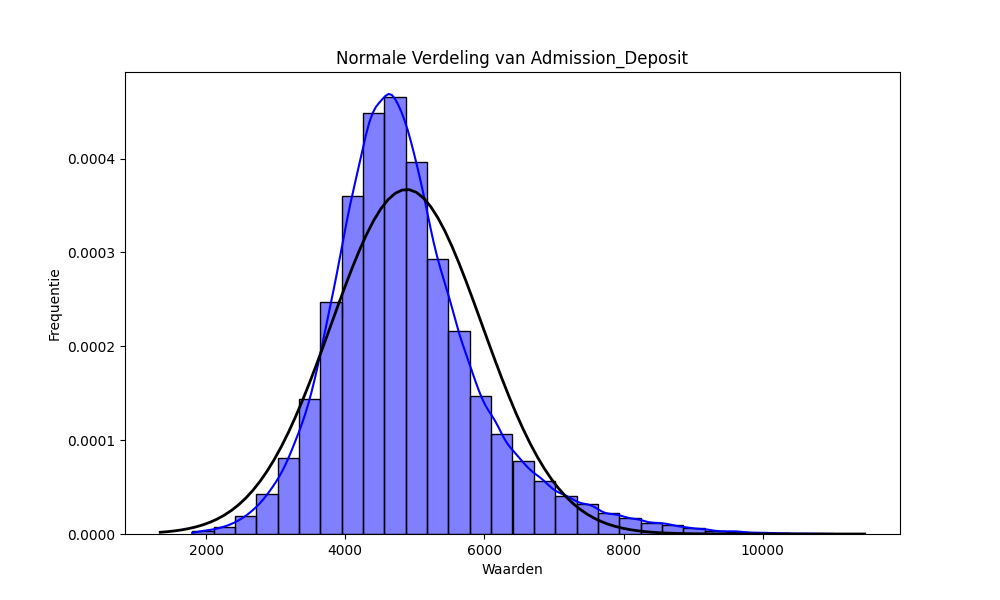
Bijlage 3: Staafdiagram Ward Type



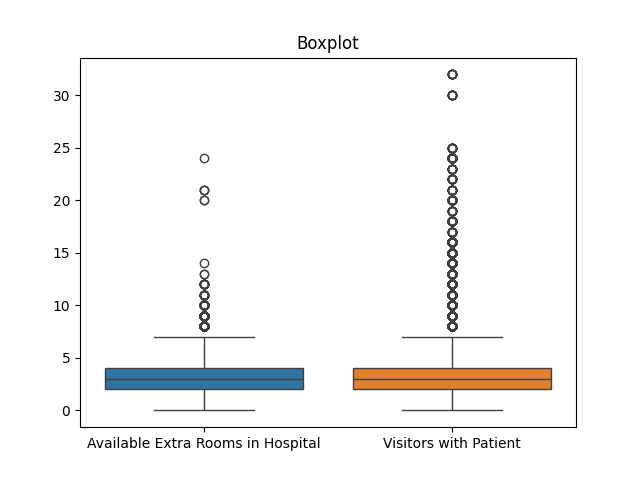
Bijlage 4: Staafdiagram Hospital Type Code



Bijlage 5: Heatmap



Bijlage 6: Normaal verdeling



Bijlage 7: Boxplot