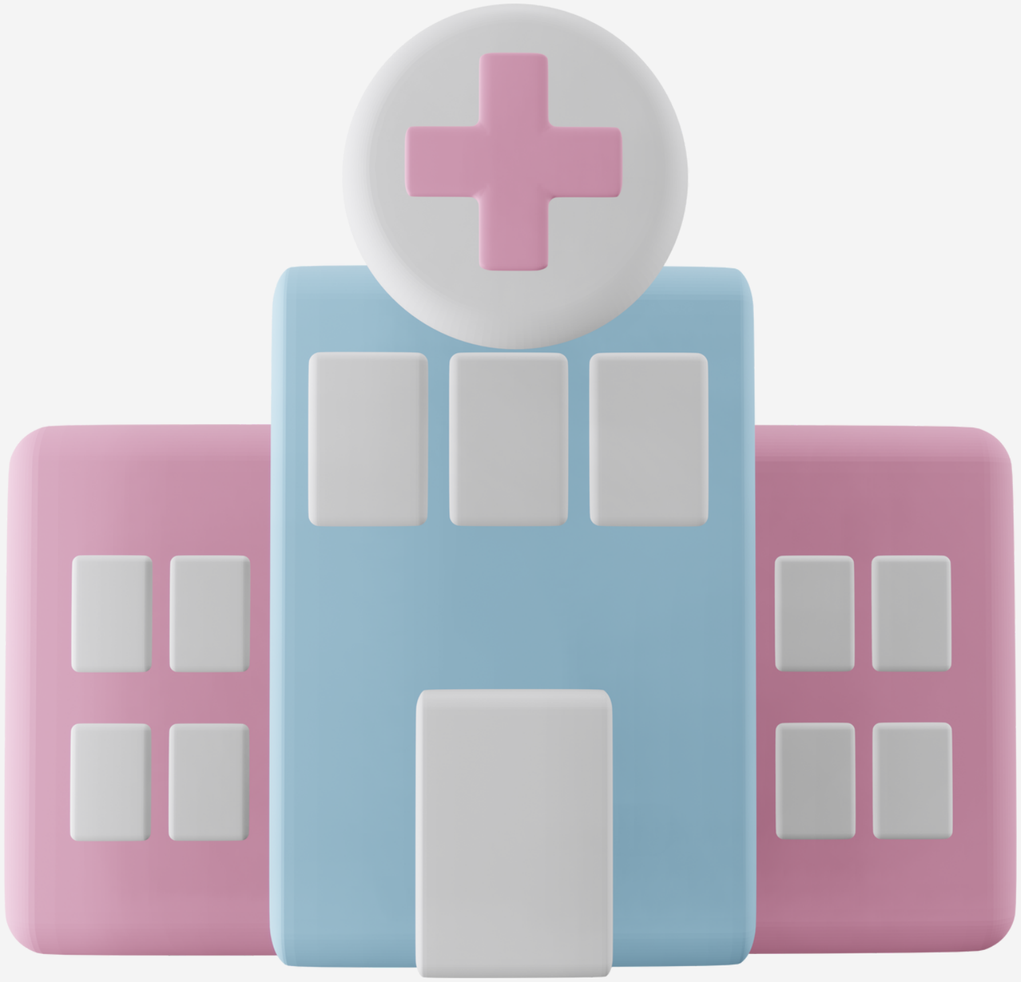


Hospital Length of Stay Prediction

Institute of Data | Capstone Presentation

Byron O'Connell | April 2023

Agenda

01 **Biography**

02 Hospital Sector Overview

03 Business Problem/Question

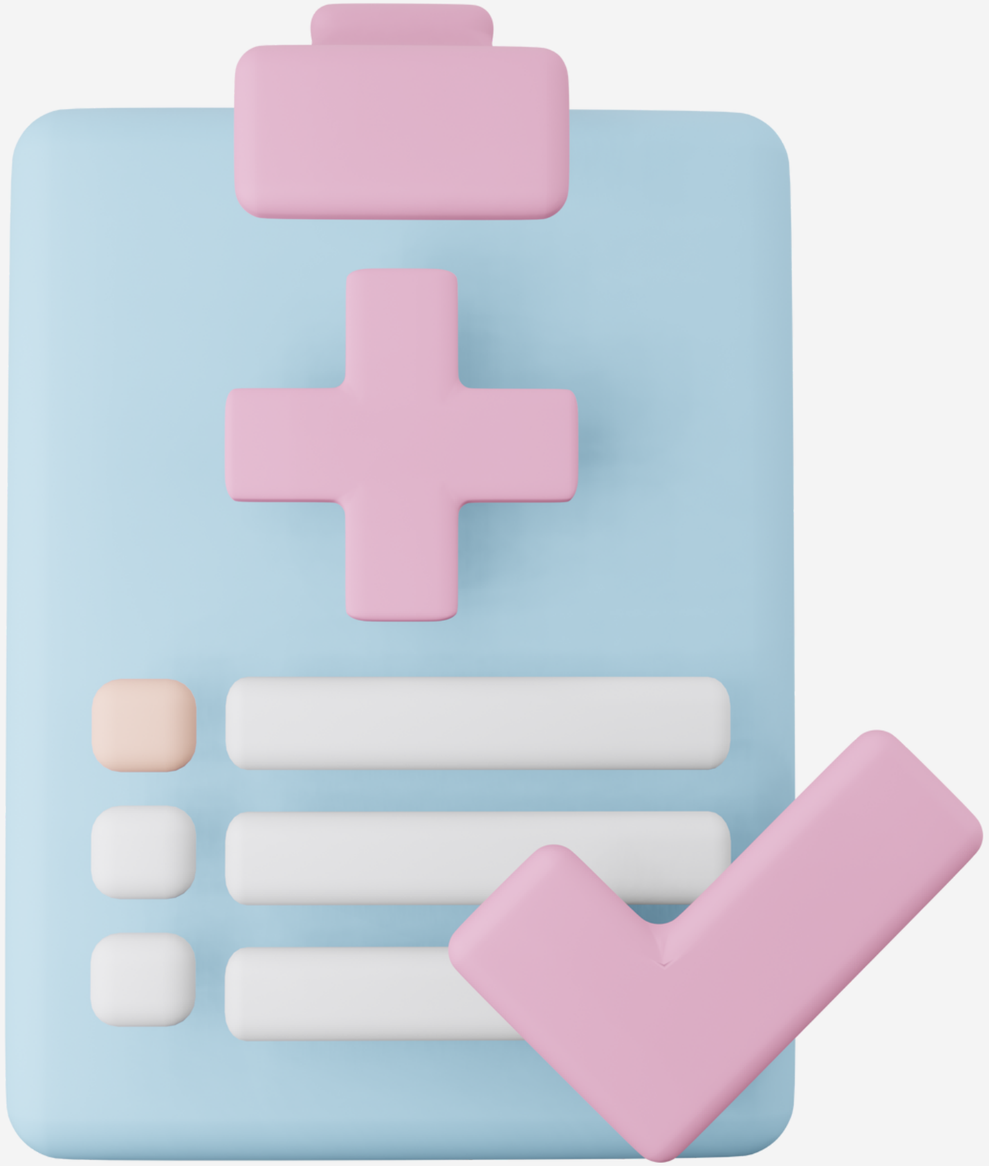
04 Dataset

05 Insights

06 Pipeline

07 Modelling

08 Conclusion



Biography

- NSW Higher School Certificate
- Data Science & Artificial Intelligence Program at the Institute of Data
- 5 years experience in an IT Helpdesk at a tertiary institution
- 2 years experience managing a wine company's e-commerce website

Agenda

01 Biography

02 **Hospital Sector Overview**

03 Business Problem/Question

04 Dataset

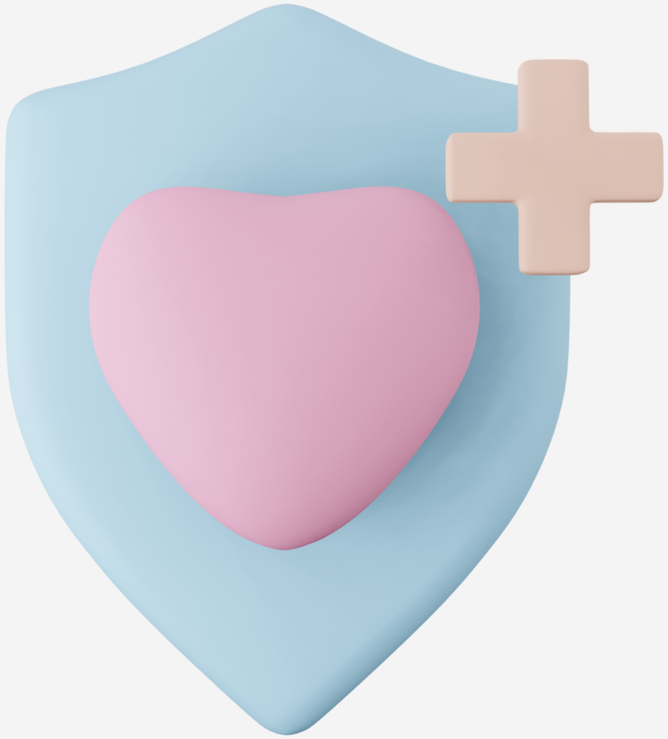
05 Insights

06 Pipeline

07 Modelling

08 Conclusion

Hospital Sector Overview



Summary

In 2021 there were 697 public hospitals and 657 private hospitals in Australia.

On an average day in Australia's hospitals:

- $246 million is spent on hospital services
- 175,000 nurses and 52,000 doctors are employed
- 32,400 people are hospitalised

Spending

In 2021, $89.7 billion was expended at hospitals. 43% was funded by state governments, 37% by the federal government and the remaining 20% by non-government sources (private health insurance, individual payments, etc)

Agenda

01 Biography

02 Hospital Sector Overview

03 **Business Problem/Question**

04 Dataset

05 Insights

06 Pipeline

07 Modelling

08 Conclusion

Business Problem & Question

Patient Length of Stay

Patient length of stay is an important indicator of the efficiency of hospital management. Hospitals have limited resources and staff, requiring efficient use of beds and clinician time.

Problem

With the relatively recent Coronavirus outbreak, the importance of patient length of stay has been heightened.

Stakeholders

Reducing the number of inpatient days results in:

- Decreased risk of infection for the patient
- More efficient bed management for hospitals
- Greater satisfaction for families of the patient.

Business Question

Can we build a model to accurately predict the length of stay for a patient upon receiving a diagnosis from a hospital?

Agenda

01 Biography

02 Hospital Sector Overview

03 Business Problem/Question

04 **Dataset**

05 Insights

06 Pipeline

07 Modelling

08 Conclusion

DATASET

- Contains approx. 2.3 million rows of de-identified patient data
- Includes information such as demographics, diagnoses, treatments, services and costs.
- Target variable: Length of Stay
- Sourced from Kaggle via 2015 New York State Government Health records.

"2015 DE-IDENTIFIED NY INPATIENT DISCHARGE (SPARCS)"

Agenda

01 Biography

02 Hospital Sector Overview

03 Business Problem/Question

04 Dataset

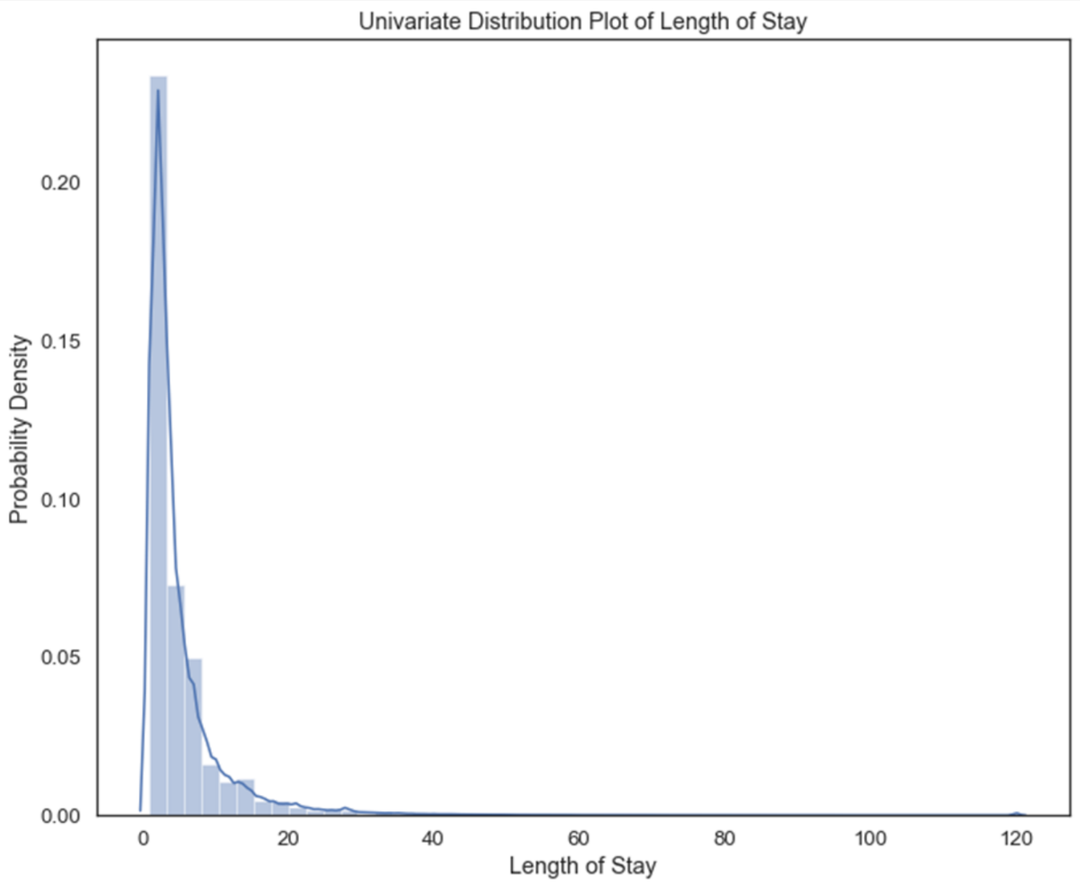
05 **Insights**

06 Pipeline

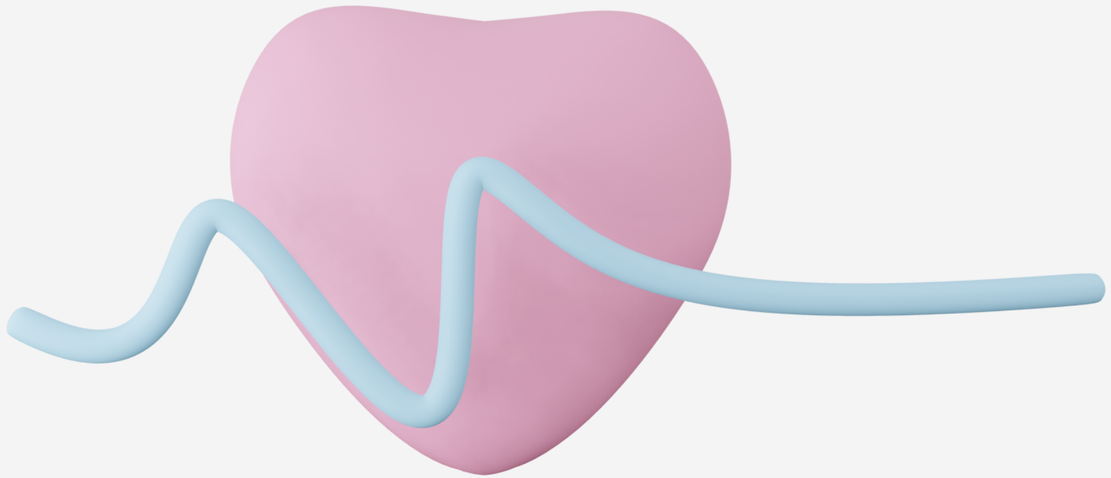
07 Modelling

08 Conclusion

Insights

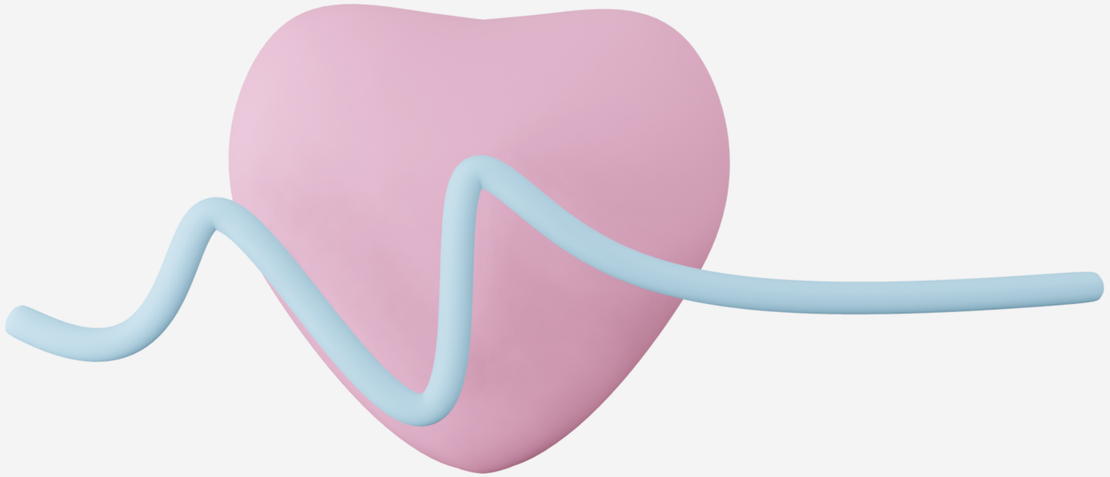


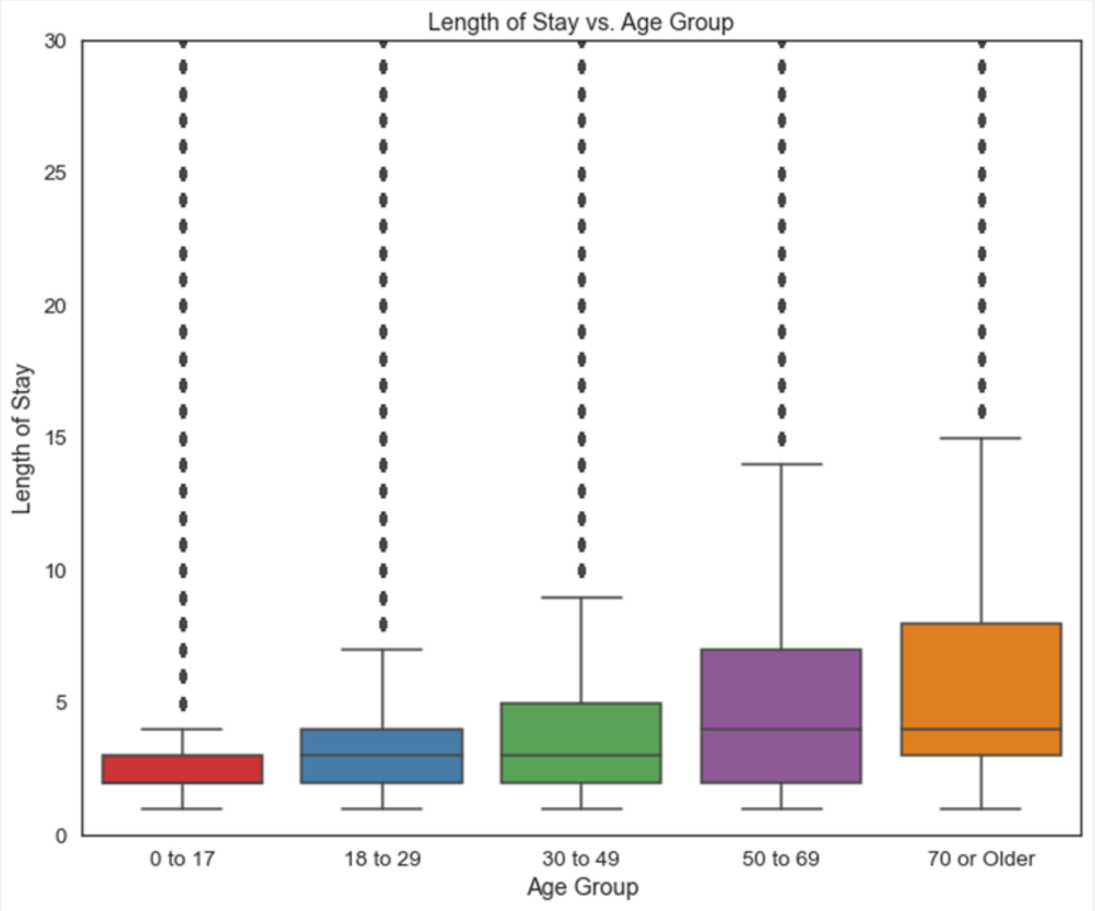
Distribution is very skewed with most patients having a length of stay between 0 to 5 days.



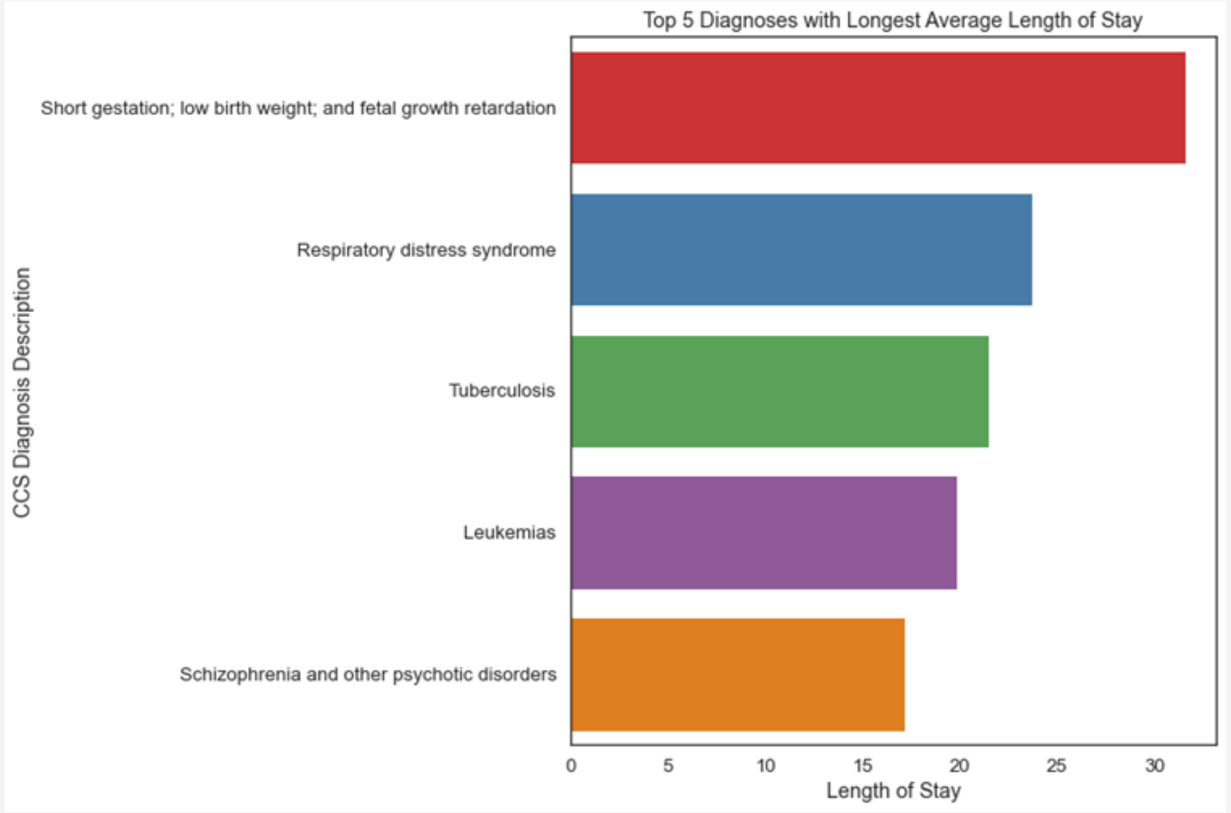
Insights

Clearly shown here is that as age increases, the length of stay distribution also increases.

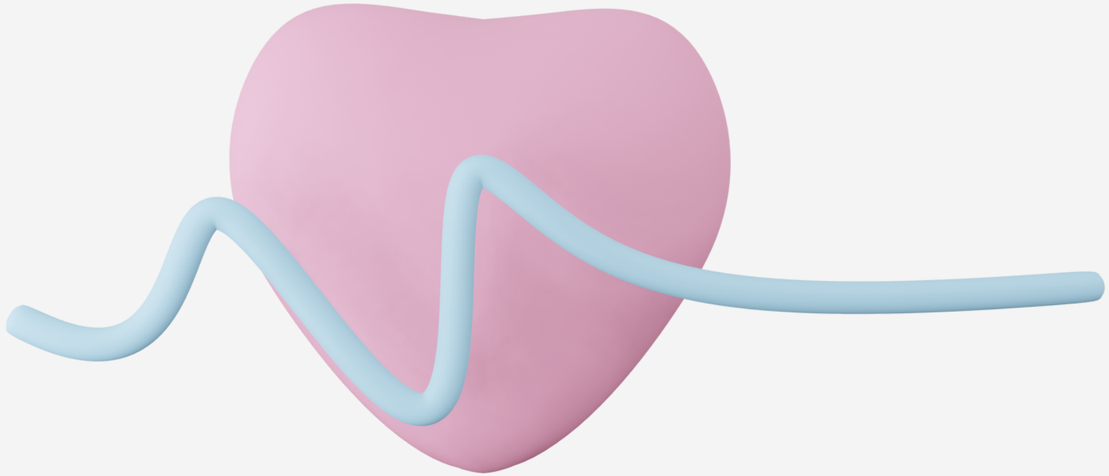




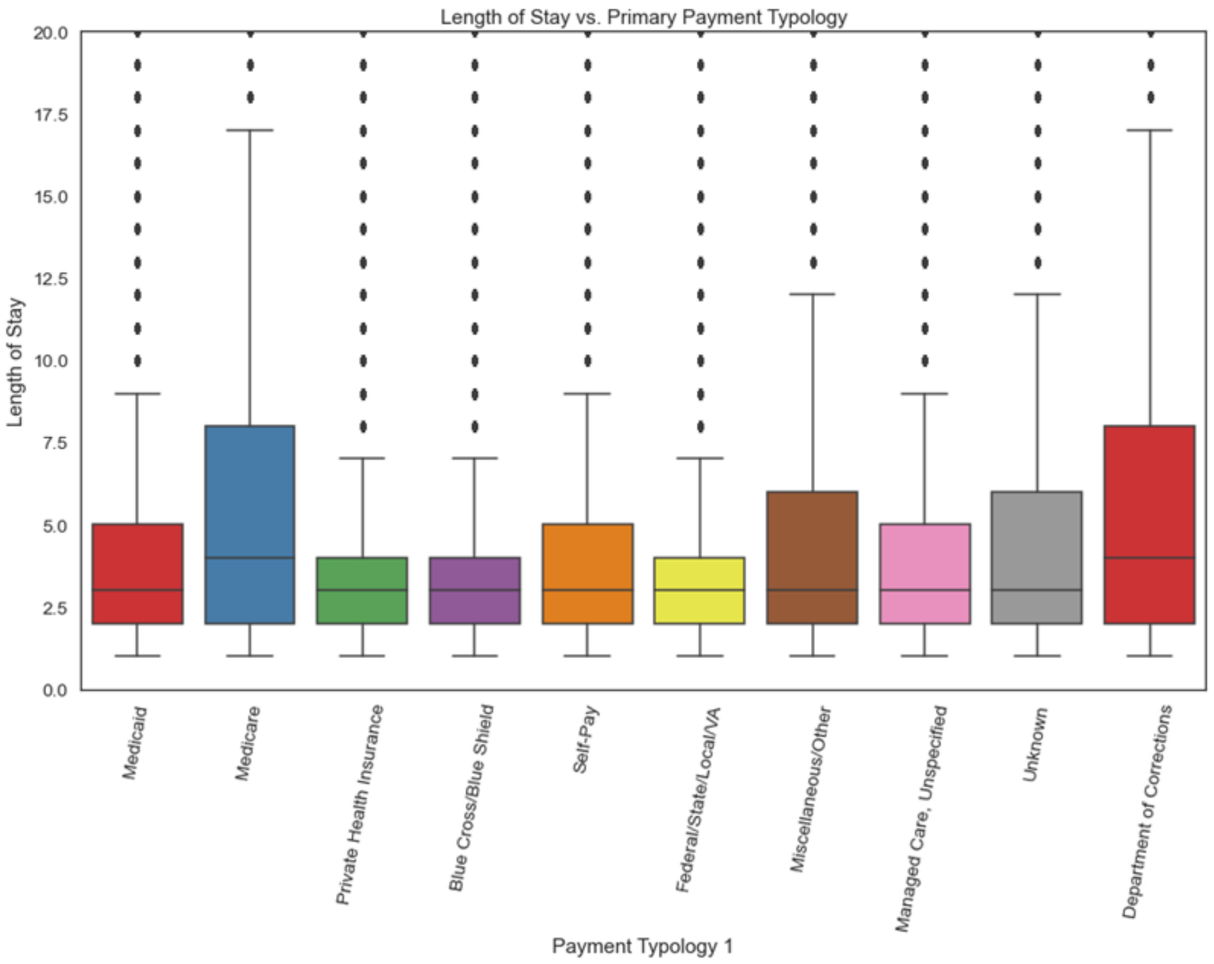
Insights



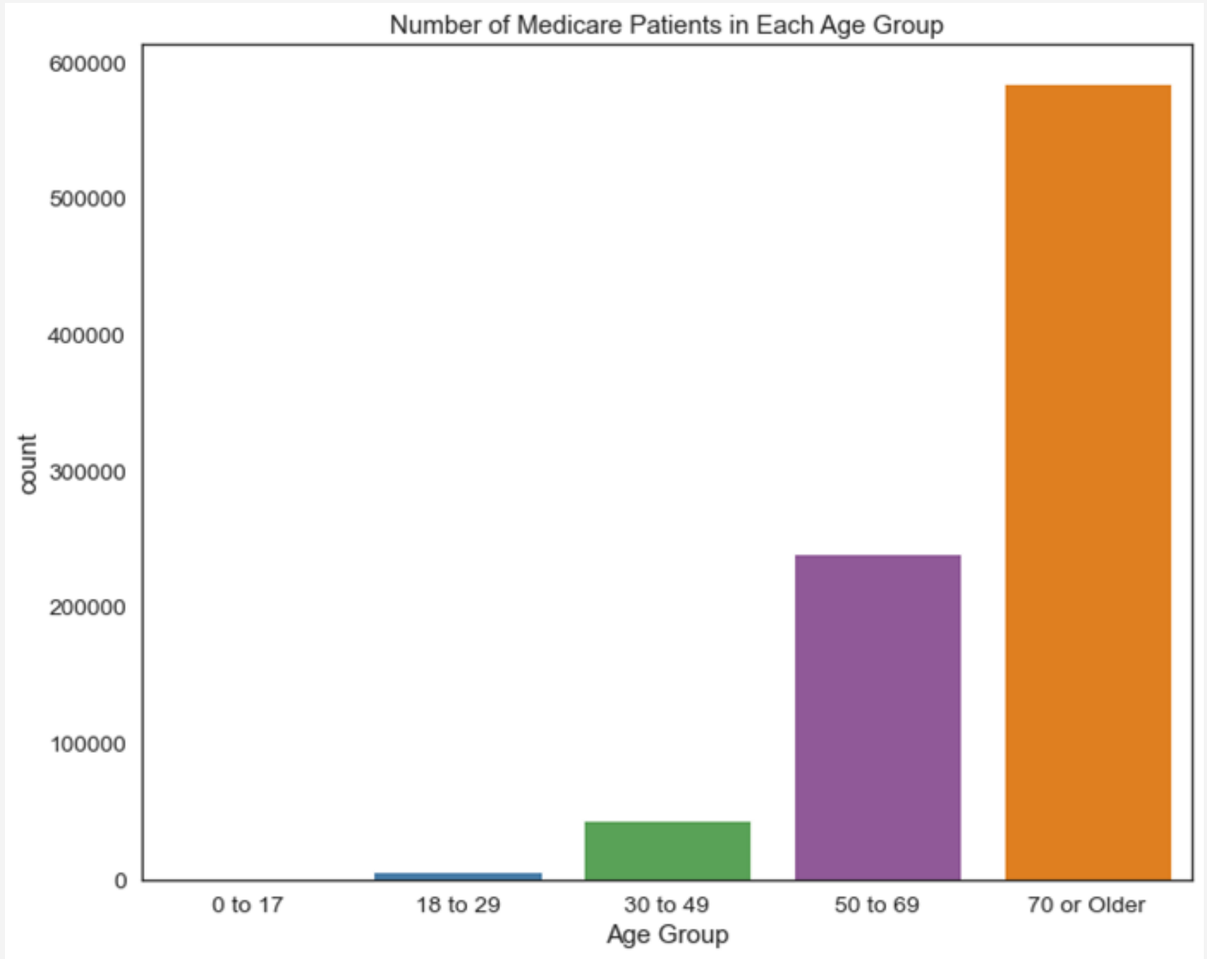
Patients with diagnoses related to birth complications tend to have the longest average length of stay.



Insights



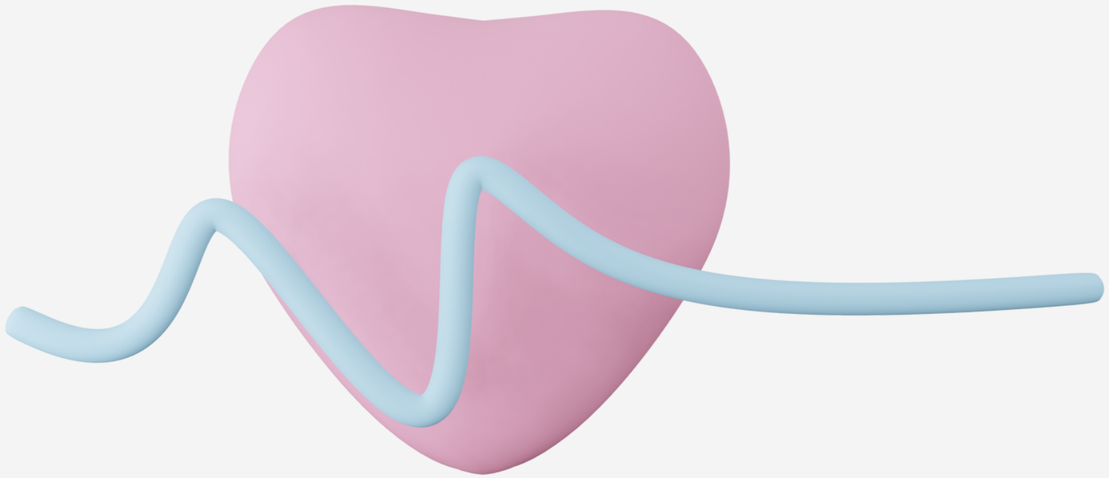
"Medicare" payment typology patients tend to spend longer in hospital than those that use other payment options.

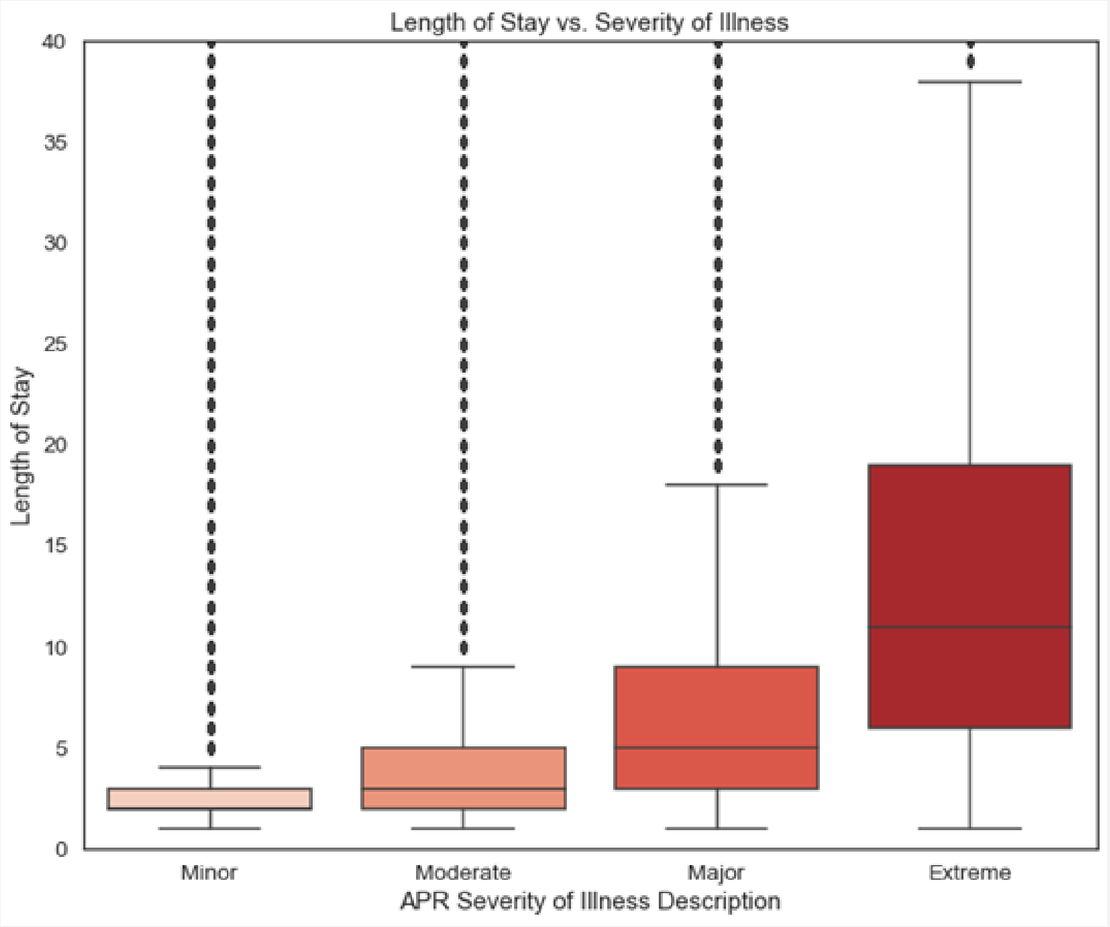


The large majority of Medicare patients are aged 65 and over.

Insights

Extreme illnesses lead to longer lengths of stay at hospital





Agenda

01 Biography

02 Hospital Sector Overview

03 Business Problem/Question

04 Dataset

05 Insights

06 **Pipeline**

07 Modelling

08 Conclusion

Pipeline

Cleaning

Drop columns not useful for predictive modelling

Check missing values

Exploration

Visualisation for insights

Summary statistics

Feature Engineering

Feature encoding

One-hot encoding

Modelling

PCA
Logistic Regression
Balanced class weight parameter
Decision Trees
Random Forest

Interpretation

Application of metrics to business problem
Recommendations

Agenda

01 Biography

02 Hospital Sector Overview

03 Business Problem/Question

04 Dataset

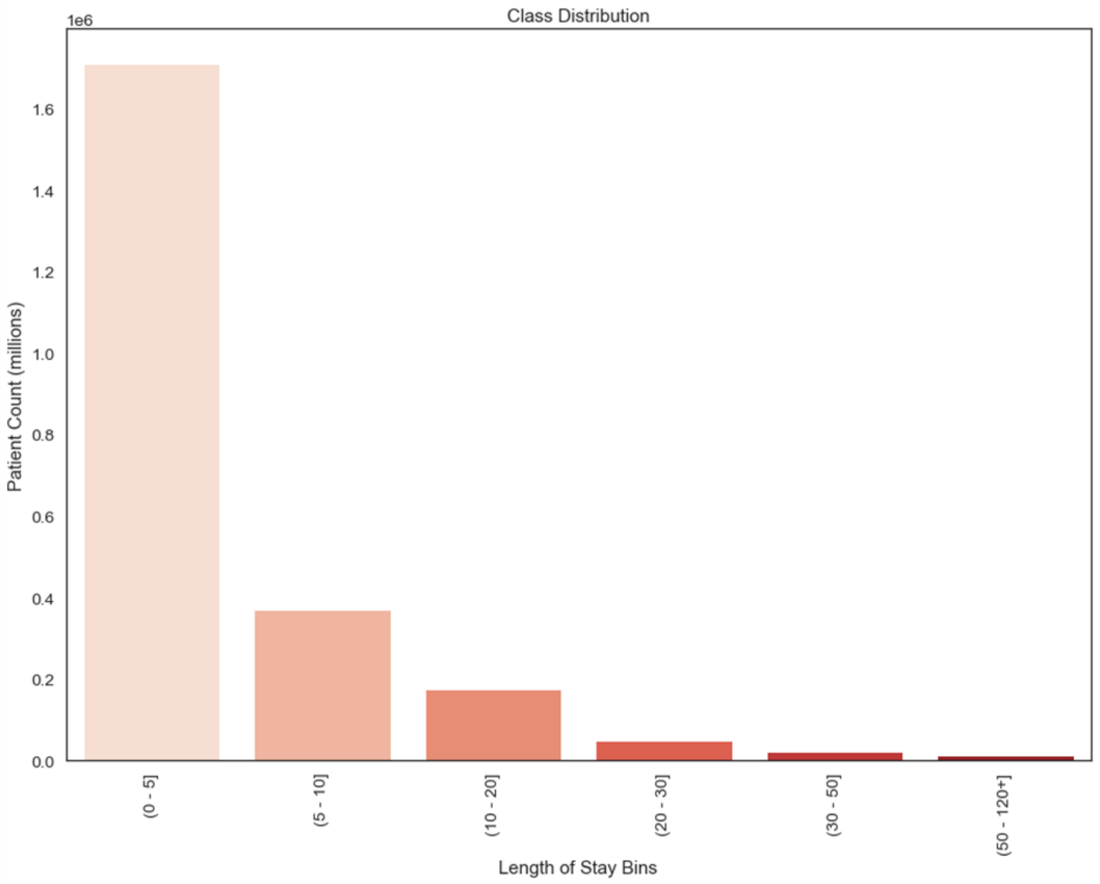
05 Insights

06 Pipeline

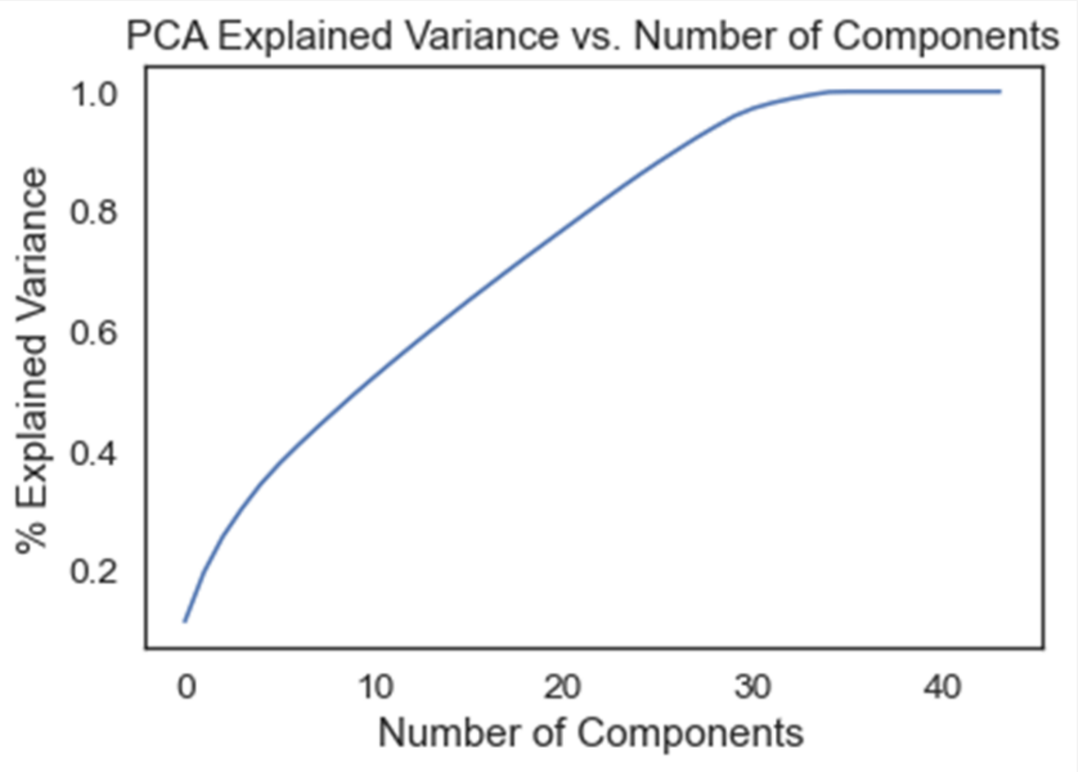
07 **Modelling**

08 Conclusion

Modelling

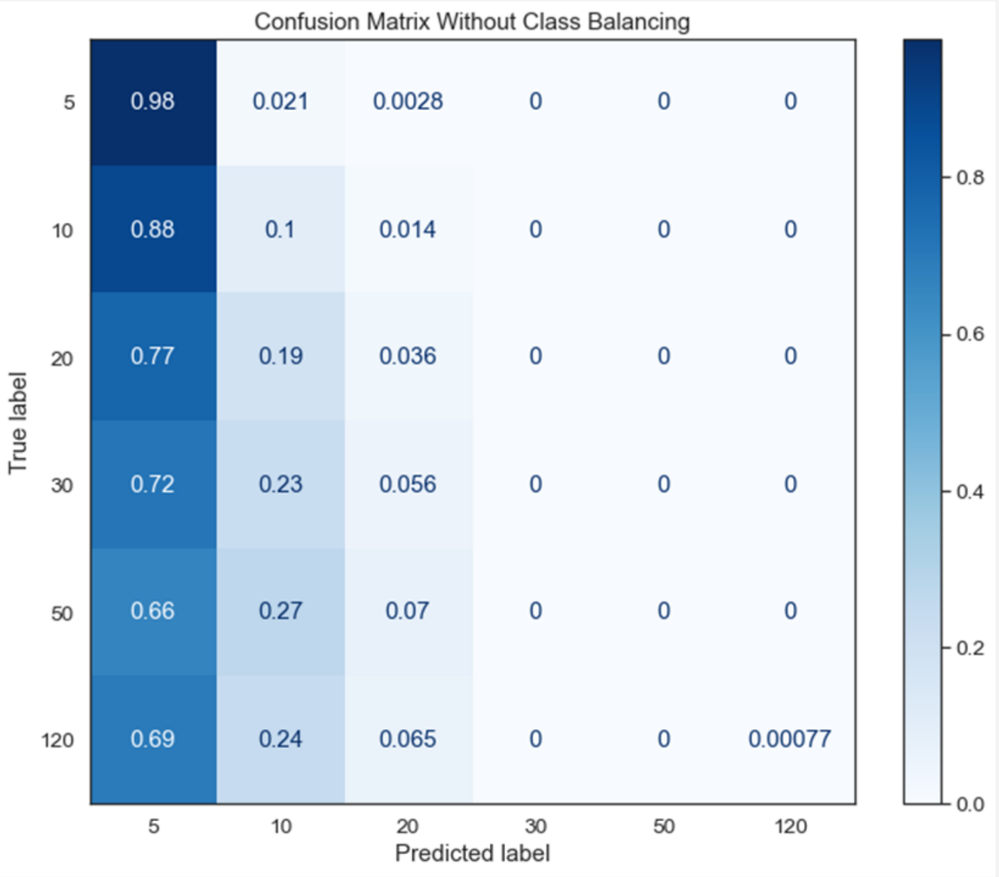


After plotting using new Length of Stay bins, it's clear there is a large class imbalance



As shown here, 29 components is the minimum number required to explain 95% variance in data.

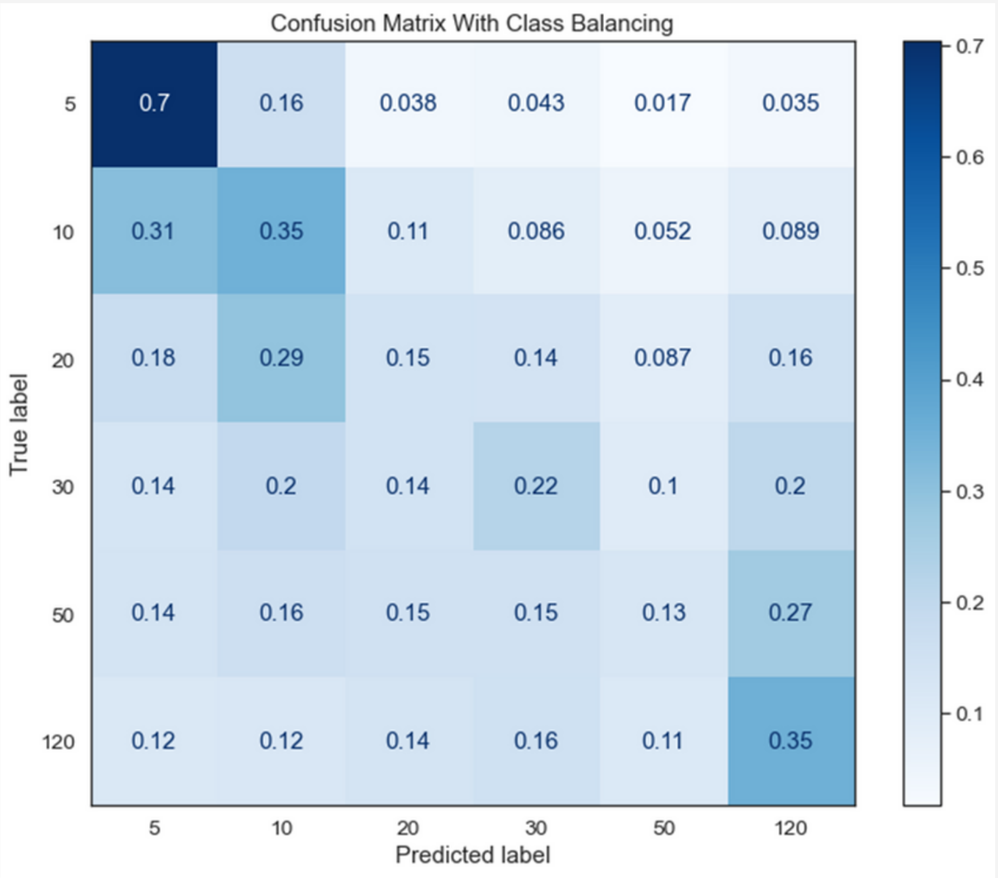
Modelling



Logistic Regression without class balancing

Test Accuracy: 0.73
Train Accuracy: 0.83

	precision	recall	f1-score	support
5	0.89	0.70	0.78	514073
10	0.28	0.35	0.31	110799
20	0.18	0.15	0.16	52600
30	0.07	0.21	0.11	14490
50	0.04	0.12	0.06	7057
120	0.03	0.35	0.05	3648
accuracy			0.59	702667
macro avg	0.25	0.32	0.25	702667
weighted avg	0.71	0.59	0.64	702667



Logistic Regression with balanced class weight parameter

Test Accuracy: 0.59
Train Accuracy: 0.59

Modelling

	precision	recall	f1-score	support
5	0.93	0.71	0.81	514073
10	0.28	0.41	0.33	110799
20	0.23	0.28	0.25	52600
30	0.15	0.26	0.19	14490
50	0.08	0.32	0.12	7057
120	0.09	0.56	0.15	3648
accuracy			0.62	702667
macro avg	0.29	0.42	0.31	702667
weighted avg	0.74	0.62	0.67	702667

Decision Trees with hyperparameter tuning

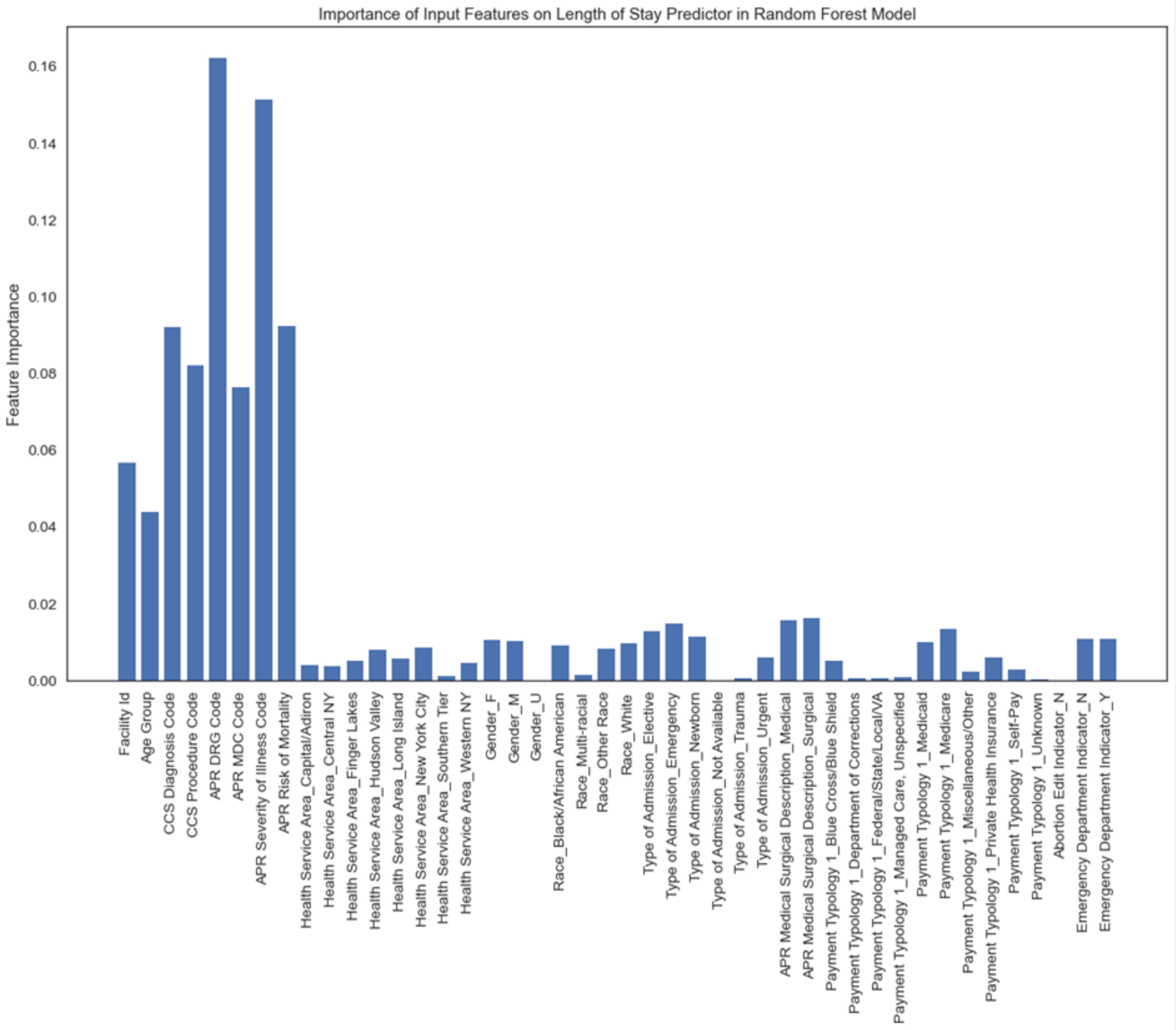
Test Accuracy: 0.62
Train Accuracy: 0.62

	precision	recall	f1-score	support
5	0.92	0.73	0.82	514073
10	0.29	0.48	0.37	110799
20	0.28	0.29	0.28	52600
30	0.18	0.25	0.21	14490
50	0.10	0.30	0.15	7057
120	0.12	0.53	0.19	3648
accuracy			0.65	702667
macro avg	0.32	0.43	0.34	702667
weighted avg	0.75	0.65	0.68	702667

Random Forest

Test Accuracy: 0.67
Train Accuracy: 0.65

Modelling



APR DRG Code and APR Severity of Illness code are the two most important features in predicting a patient's length of stay.

Agenda

01 Biography

02 Hospital Sector Overview

03 Business Problem/Question

04 Dataset

05 Insights

06 Pipeline

07 Modelling

08 **Conclusion**

Conclusion

Summary

I was able to predicate a patient's length of stay with an accuracy of 67% using data only available immediately following diagnosis.

APR DRG Codes and APR Severity of Illness Codes are the two most important features in predicting a patient's length of stay

The Medicare subgroup also has a relatively high importance when compared to other payment types

Key Recommendations

Present model to stakeholders as it's able to improve hospital management and staff/patient well-being.

Additional machine learning models should be explored.

Future models could build on this one and look at the cost associated with a certain length of stay.

Thank you

GitHub Repo

https://github.com/byronoconnell/IOD_Capstone