Assignment – Part B: Big Data Analysis

COMP SCI 7209 – Big Data Analysis and Project

Student Name: Prashant Shrestha

Dataset: ER Wait Time Dataset (Creative Commons, 2025)

Introduction and Restatement of Research Question

In today's healthcare systems, long waiting times in Emergency Departments (Eds) is one of the critical issues affecting both patient results and satisfaction. With increasing pressure on public health services, understanding the root causes of these delays is most important. Building upon our earlier exploration in Part A, we revisited the refined research question:

Which hospital-related issues have the major effect on emergency room waiting times, and how can this information help reduce waiting times and optimize patient experience?

Data Overview and Key Features

The dataset comprises **5,000 patient entries** over a simulated year, reflecting visit-level records including patient urgency, time of arrival, hospital characteristics and outcome metrics. The data was preprocessed to:

- Fix typo errors (e.g., “St. Maryâ€™s” corrected to “St. Mary’s”).
- Standardise column names by using lowercase and underscores.
- Convert date fields to datetime format and extract useful temporal features (day, hour, month).
- Encode categorical variables such as *urgency_level* and *season*. (GeeksforGeeks, 2025)

Feature Engineering Summary

To prepare the dataset for predictive modelling, a range of feature engineering techniques were applied to enhance the quality and relevance of input variables:

- **Date Time conversion**:
  The original visit_date column was converted into a datetime object, from which new variables such as hour, day_of_week and month were extracted. This allowed us to capture temporal patterns in emergency department activity and align them with operational capacity. (GeeksforGeeks, 2025)

- **Categorical Encoding**:
  Categorical variables including urgency_level, time_of_day and season were label-encoded to convert them into numerical form suitable for regression and classification models. These transformations ensured compatibility with algorithms that require numeric inputs.

- **Operational Indicators**:
  Features such as nurse-to-patient_ratio and specialist_availability were preserved as-is due to their intrinsic numeric structure. However, they were evaluated for correlation and possible interaction effects during feature selection.

- **Multicollinearity Consideration**:
  Variables like time_to_registration, time_to_triage and time_to_medical_professional showed very high correlation with the target variable (total_wait_time). These relationships were acknowledged and dimensionality reduction techniques such as **Lasso regression** or **Principal Component Analysis (PCA)** are being considered to mitigate multicollinearity in future modelling. (J Munro, 2006)

Key Attributes Examined:

- **Operational**: nurse-to-patient_ratio, specialist_availability, bed_count
- **Time-Related**: time_of_day, day_of_week, season
- **Outcome**: total_wait_time_(min), patient_satisfaction

Univariate Analysis

To gain a foundational understanding, we examined the distribution of both categorical and numerical variables:

- **Urgency Level**: Most visits were categorised as *Medium*, followed by *Critical* and *High and Low* cases even though the difference is minimal (Appendix Figure 6).
- **Time of Day**: Visits peaked during the *afternoon* and *evening*, suggesting periods of higher load (Appendix Figure 7).
- **Nurse-to-Patient Ratio** displayed multimodal distribution with major peaks a 3.0 and 4.0 indicating hospitals follow set staffing models, such as 3:1 or 4:1 ratio (Appendix Figure 8).

Visuals Used:

- Count plots for categorical features
- Histograms for numerical distributions

Bivariate Analysis and Pattern Discovery

1. Wait Time by Urgency Level

A **boxplot** revealed a positive correlation between urgency and reduced wait time. *Critical* patients had the shortest median wait, which aligns with triage prioritisation as shown in Figure 1.

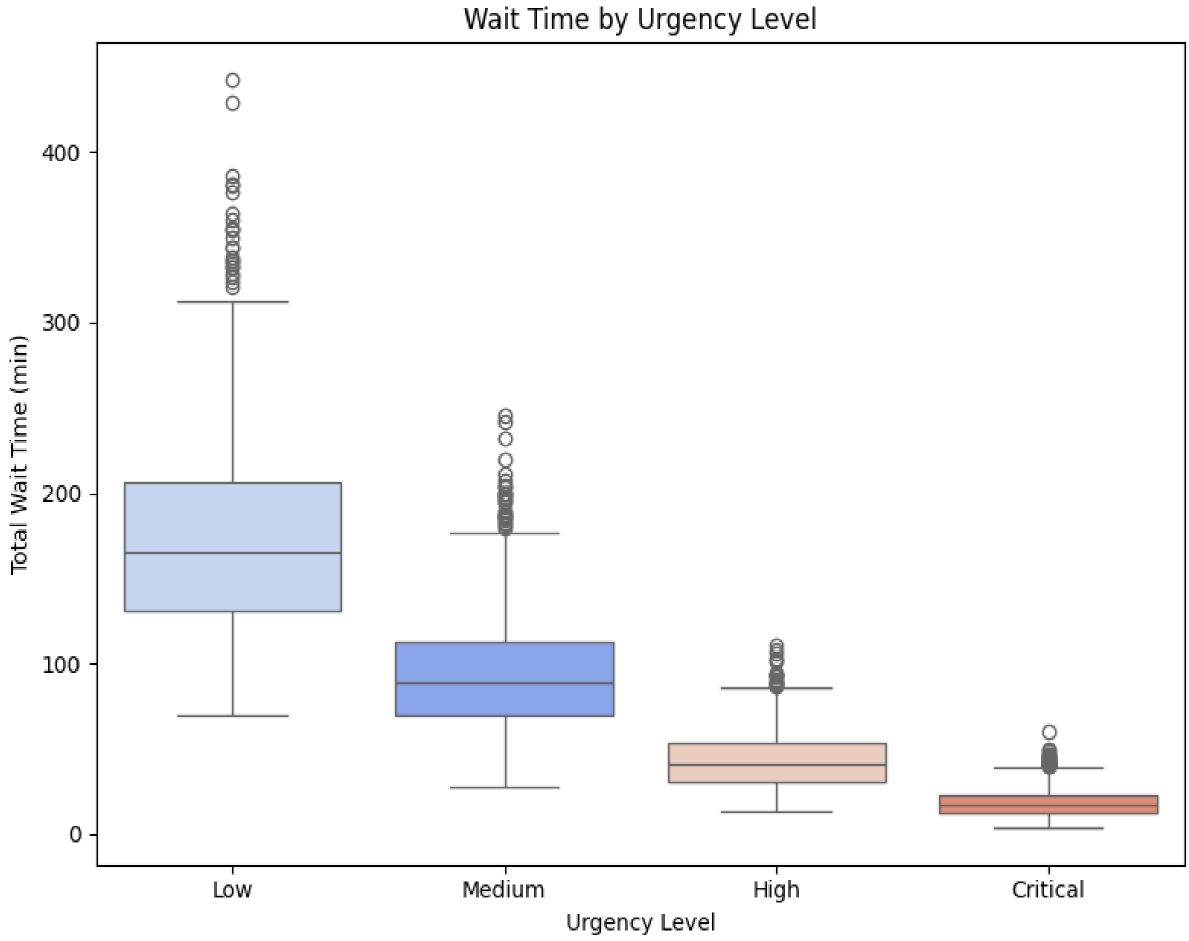

Figure 1: Wait Time by Urgency Level

2. Time of Day vs Wait Time

A **bar chart (Figure 2)** showed that wait times were longest during the *afternoon* (14:00 - 18:00), with early morning visits experiencing significantly shorter delays. This may lead to resource fatigue or staff shortages later in the day.

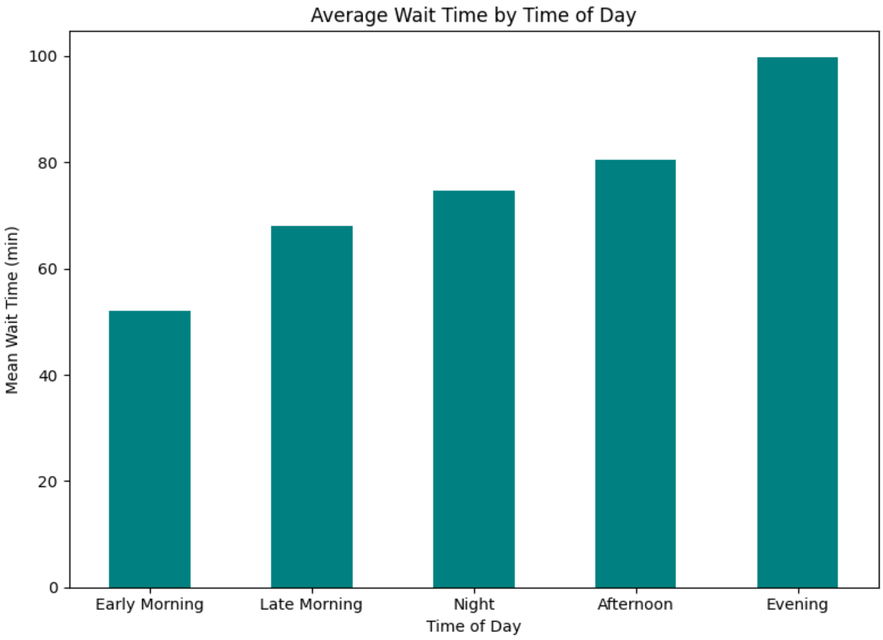

Figure 2: Average Wait Time by Time of Day

3. Day of Week Trends

Using a **line chart (Figure 3)**, we observed that *Mondays* consistently had the highest average wait times, whereas *weekends* showed moderate to lower delays-likely due to reduced patient inflow.

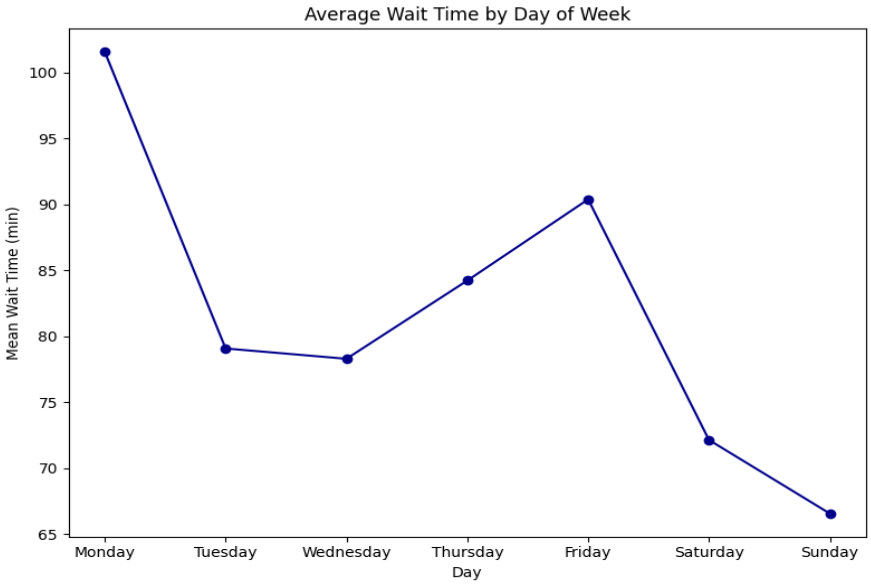

Figure 3: Average Wait Time by Day of Week (Mike Yi, 2025)

Multivariate Exploration

Seasonal and Regional Impact

A **grouped bar chart (Figure 4)** comparing average wait time across **seasons and regions** uncovered interesting disparities:

- Winter and Autumn generally recorded longer delays, especially in rural regions, possibly due to flu seasons or reduced staffing.

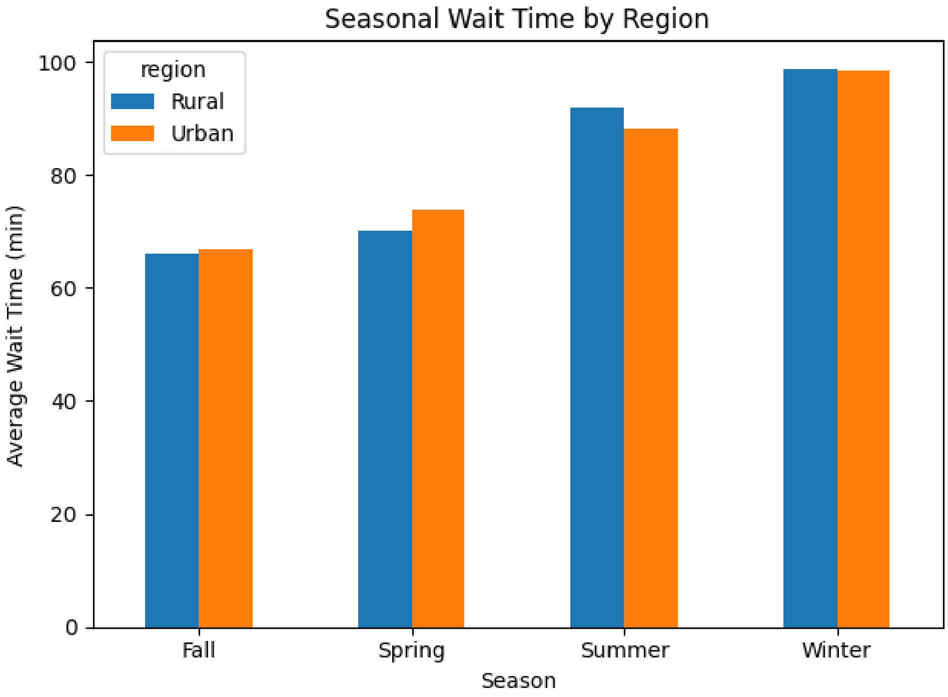

Figure 4: Seasonal Wait Time by Region (Mike Yi, 2025)

Correlation Heatmap

A **heatmap (Figure 5)** of numerical features highlighted:

Operational Delays and Wait Times

- total_wait_time_(min) has:
    - Strong correlation with:
        - time_to_registration (**r = 0.92**)
        - time_to_triage (**r = 0.95**)
        - time_to_medical_professional (**r = 0.98**)

      suggests each stage of the patient's journey adds cumulatively to total wait time.

Patient Satisfaction

- Strong **negative correlation** with:
  - total_wait_time (**r = –0.87**)
  - time_to_medical_professional (**r = –0.86**)
  - nurse-to-patient_ratio (**r = –0.74**)

  indicates that longer waits and higher nurse loads reduce patient satisfaction.

Staffing and Capacity

- nurse-to-patient_ratio is:
  - Moderately correlated with longer wait times (**r = 0.69**)
  - Negatively correlated with satisfaction (**r = –0.74**)

Temporal Features

- hour and month show negligible correlation with patient outcomes or operational metrics (|r| < 0.05)

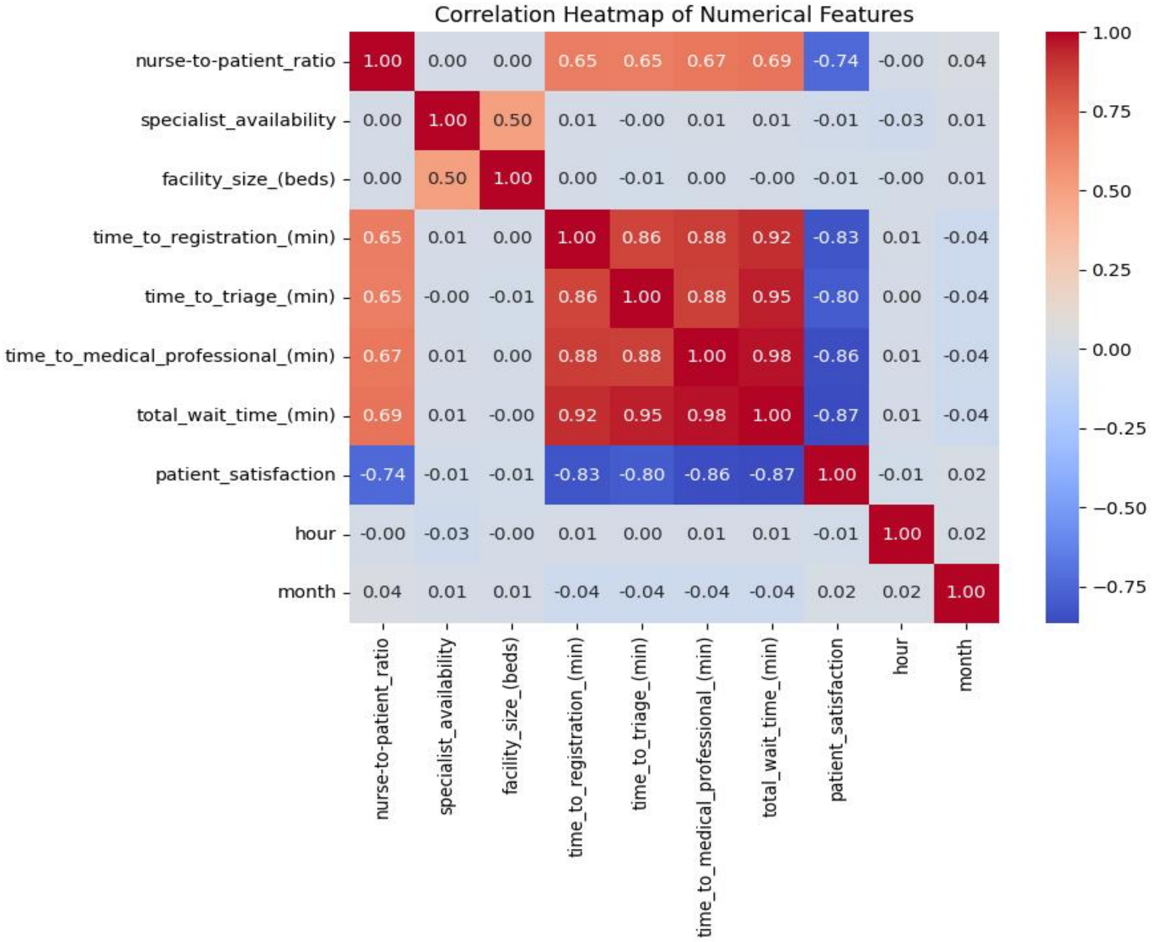

Figure 5: Correlation Heatmap

Refining the Research Scope

Based on the exploratory analysis, the research question was sharpened:

What are the most influential hospital-level factors (e.g., staffing ratios, specialist availability) that contribute to extended ER wait times, particularly across different urgency levels and time slots?

We also plan to investigate:

- Whether increasing nurse-to-patient ratios consistently reduce delays across all urgency categories.
- If staffing policies should be seasonally adjusted to match patient demand patterns.

Data Limitations and Next Steps

While the dataset offers a rich synthetic view, several limitations must be acknowledged:

- **Simulated Nature**: Though grounded in real-world patterns, actual hospital variability may differ.
- **Missing Patient Demographics**: Age, gender, or pre-existing conditions could improve the robustness of predictive modelling.
- **Overlapping Features**: Some fields like time_to_triage and time_to_medical_professional are interdependent and may introduce multicollinearity.

To address these:

- We'll test for multicollinearity during modelling.
- If required, feature selection techniques such as Lasso or PCA will be applied.

Conclusion and Modelling Plan

This stage has highlighted key operational inefficiencies and patterns in emergency room delays. The next phase will involve:

- **Predictive Modelling**: Linear regression, decision trees, or ensemble models to forecast total wait time.
- **Feature Importance Analysis**: Determine which operational and temporal factors contribute most.
- **Visualisation**: Continued use of Seaborn and Matplotlib for interpretability.

References

Creative Commons, 2025. *ER Wait Time.* [Online]
Available at: https://www.kaggle.com/datasets/rivalytics/er-wait-time?resource=download&select=ER+Wait+Time+Data+Overview.txt
[Accessed June 2025].

GeeksforGeeks, 2025. *Data Preprocessing in Data Mining.* [Online]
Available at: https://www.geeksforgeeks.org/dbms/data-preprocessing-in-data-mining/
[Accessed June 2025].

J Munro, S. M. J. N., 2006. *Effectiveness of measures to reduce emergency department waiting times: a natural experiment.* [Online]
Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC2564124/
[Accessed June 2025].

Mike Yi, M. R., 2025. *How to choose the right data visualization.* [Online]
Available at: https://www.atlassian.com/data/charts/how-to-choose-data-visualization
[Accessed July 2025].

Appendix

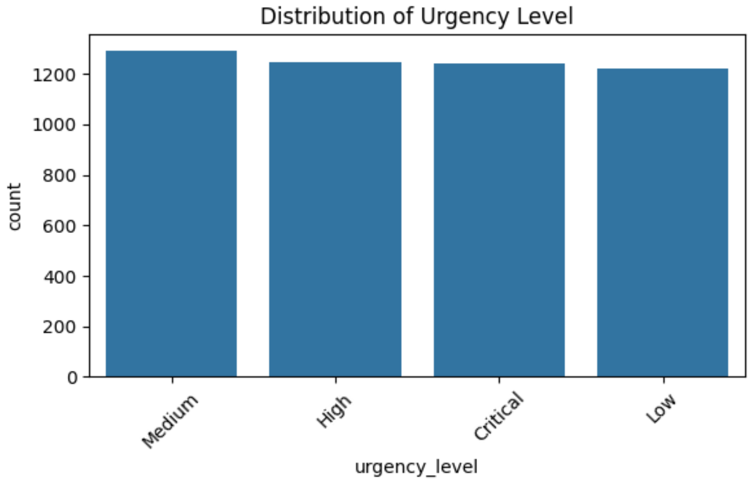

Figure 7: Distribution of Urgency Level

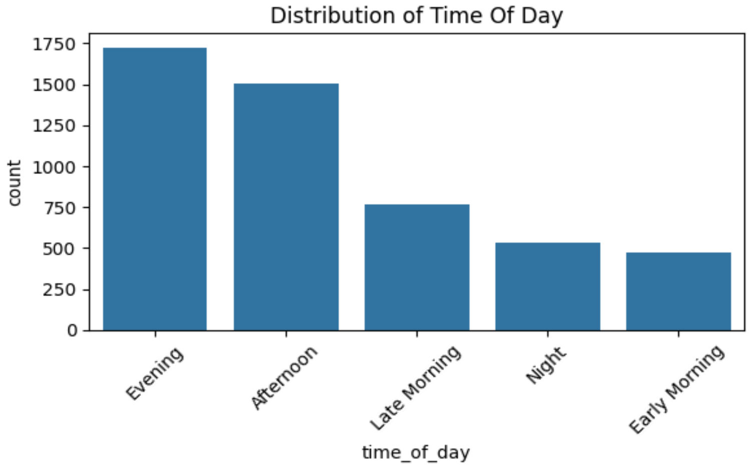

Figure 8: Distribution of Time of Day

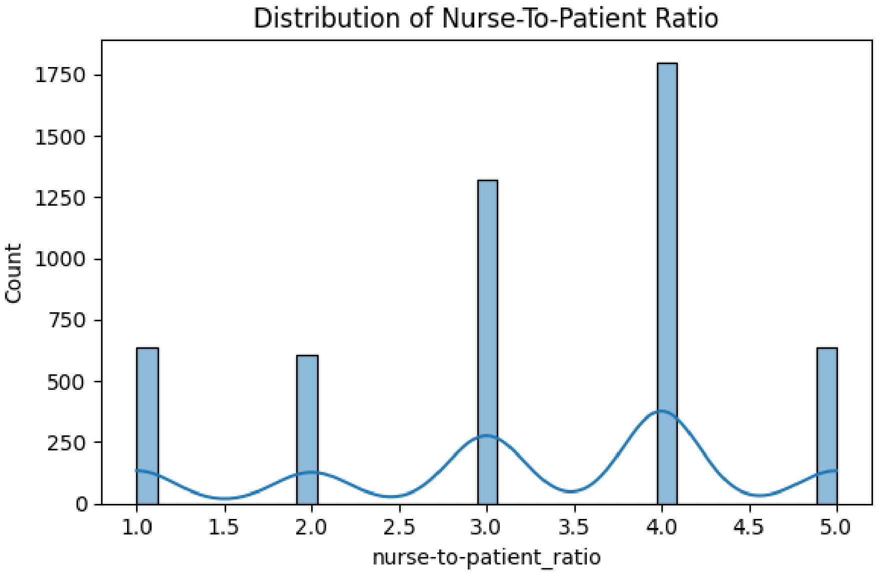

Figure 6: Distribution of Nurse-to-patient Ratio

GitHub Repository link:

https://github.com/preheriaa/Big-Data-Analysis-and-Project.git