Ramsay Data Insights

# Part 1: Data Exploration and Preparation

## Data Issue

1. Missing values: Several columns have a large number of missing values, including:

* CCU\_Charges (17,167 missing values out of 30,000)
* ICU\_Charge (17,138 missing values)
* TheatreCharge (17,164 missing values)
* PharmacyCharge (16,735 missing values)
* ProsthesisCharge, OtherCharges, and BundledCharges have similar levels of missing data.
* UnplannedTheatreVisit, InfantWeight, Readmission28Days, HoursMechVentilation, and PalliativeCareStatus have missing data in nearly all rows (around 28,000–29,000 missing entries).

1. Inconsistent data types: Some columns have inconsistent or unexpected data types:
2. Outliers or impossible values:

PharmacyCharge has extremely large and unlikely values (e.g., 1.0866012197370352e+111), indicating possible data corruption or incorrect formatting.

## Strategies to address these issues:

1. Missing values:

* Handle or impute missing data depending on the use case. For all of the charges, we can impute null values with 0.
* However, for UnplannedTheatreVisit, InfantWeight, Readmission28Days, HoursMechVentilation, and PalliativeCareStatus, there are too many missing rows in those columns. If we impute the missing rows with ‘No’, this will affect our analysis since ` UnplannedTheatreVisit`, ` Readmission28Days`, ` PalliativeCareStatus` are all crucial features to monitor quality of care. My suggestion is to leave all those columns as they are and not include them in the analysis until we can understand why there are so many missing values.

1. Outliers:

* All of values in PharmacyCharge are significantly larges and unrealistic. This suggests that there might be data corruption.
* My suggestion is to leave the columns as it is and not include it in any analysis.

## Create New Feature:

* Length of Stay:
  + The "Length of Stay" feature captures a critical aspect of healthcare data, as it is directly related patient outcomes, and hospital operational efficiency. This feature is often used in healthcare analytics to assess hospital performance, monitor quality indicators and evaluate cost drivers.
* Diagnosis Group
  + There are around 600 principal diagnosis codes in the whole dataset, which is quite challenging and impractical for the analysis. So we will group the principal diagnosis codes into broader groups based on ICD10 Codes. This will simplify the dataset by reducing the number of diagnosis codes.
* TotalCharges:
  + ‘TotalCharges’ is created based on sum of the individual charges for each patient.
  + The **Total Charge** feature gives a complete picture of the overall cost incurred during a patient's treatment.
  + The Total Charge feature helps in identifying key cost drivers. For example, if certain procedures or diagnoses consistently have higher total charges, you can investigate further to find the cause (e.g., complications, prolonged ICU stays, high resource use).

# Part 2: Data Analysis and Visualisation

Our analysis will focus on Total Charges for each patient to identify the key cost drivers. Understanding key cost drivers can help Ramsay Hospital improve operational efficiency and allocate resources to high-cost patients.

## Total Charges:

A graph with blue and grey bars

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated

* The analysis reveals that the top three **DRGs (Diagnosis-Related Groups)** accruing the largest total charges are:
  + DRG002: Total charges of $9,234,722
  + DRG003: Total charges of $9,213,060
  + DRG001: Total charges of $9,029,666

These DRGs far surpass the charges of other groups, with the next highest DRG (L62C) accruing only $89,634 in total charges.

* The analyst also indicates that the top 3 DRGs with the highest average charges are:
  + I65B: Average charges of $3,444
  + L67B: Average charges of $3,287
  + B42C: Average charges of $3246
* Hypotheses for Drivers of Charges:
  + Length of Stay: Those DRGs may involve with patients with more severe conditions that requires more intensive management, leading to higher length of hospital stay, accommodation and theatre charges.
  + Number of patients: The fact that DRG002, DRG003, and DRG001 are among the top three in total charges but not in the top ten for average charges suggests that these DRGs have significantly more patients compared to others.
  + A graph of a number of patients

    Description automatically generatedEmergency or Urgent Care: If these DRGs are related to emergency or urgent admissions, the need for immediate, intensive care can also drive up costs.
* The funnel plot indicates the average cost per patient versus the number of patients in each DRG. It is evident that there are 3 DRGs with almost 5000 patients in each while the other DRGs have less than 100 patients. The high number of patients might be one of the factors that drive up the total cost.
* In fact, a large proportion of patients (almost 50% of patients) are concentrated in just three Diagnosis-Related Groups.
* Other factors that can affect the total cost:



* + A pie chart with different colored circles

    Description automatically generatedBundled Charges are highly correlated with Total Charges, indicating they are significant contributors to overall costs. In fact, Bundled Charges contributes 34.3% of the total charges:
  + Correlation Matrix also shows that Length of Stay Correlated with Accommodation Charges but Weakly with Total Charges
* Total Charges by CareType:

A chart of a chart with a few blue squares

Description automatically generated with medium confidence

* Although the median total charge is similar across different types of care (Inpatient, Outpatient, and Emergency), there are more outliers with higher charges in Emergency care. This suggests that while typical cases in Emergency care may incur similar costs to Inpatient or Outpatient care, there are certain high-cost cases driving up the overall charges in Emergency settings.
* Total Charges by Age Group:

A comparison of a bar chart

Description automatically generated with medium confidence

* Number of patients in age group 80+ is higher than other age groups, which is also the reason for the total charges for this group of patients is slightly higher.

*\* Detailed analysis can be found in the notebook data-exploratory*

## Part 4: Strategic Insights and Recommendations:

### **Bundled Charges Significantly Impact Total Charges**

### **Insight: Bundled Charges are Highly Correlated with Total Charges:**

### Bundled Charges constitute a substantial portion of the Total Charges for patient episodes.

### They are more strongly correlated with Total Charges than individual cost components like Accommodation Charges or Length of Stay (LOS).

Recommendations:

* Conduct a detailed review of the services included in Bundled Charges to identify high-cost items.

1. Concentration of Patients in Specific DRGs

Insight: 50% of Patients are Concentrated in Just Three DRGs

* A significant proportion of patients are under a small number of DRGs, which drive up the cost in those DRGs
* No significant difference in Length of Stay among the top three DRGs, suggesting uniformity in patient stay durations across these groups.
* No significant difference in Total Charges among the top three DRGs, suggesting stability in total charges across these groups.

1. Length of Stay (LOS) is Not Strongly Correlated with Total Charges
   * This suggests that factors other than LOS, such as procedures, diagnosis and bundled services, are primary cost drivers, emphasizing the need for a comprehensive predictive model.

*\* Detailed analysis can be found in the notebook data-exploratory*

## Part 5: Model Development:

Based on the insights discovered in Part 4, a predictive model that would greatly benefit Ramsay Health Care is one that predicts Total Charges for patients at the time of admission. This model would help the organization anticipate high-cost cases and allocate resources more effectively, especially in the highly-concentrated AR-DRG.

This prediction will help us to:

* Anticipate high-cost cases to allocate resources efficiently.
* Improve financial forecasting and budget planning.
* Identify opportunities to reduce unnecessary expenses without compromising patient care.
* Improve planning for staff and equipment based on expected patient cost and resource utilization.
* Develop tailored care plans for patients likely to incur high costs, potentially improving outcomes.

1. Prediction Target:

* Total Charges: The total cost incurred by a patient during their hospital stay.

1. Features

Different kinds of features should be used in the model including patient demographics (age, sex), admission Details: admission type, source of referral, urgency of admission, Clinical Information: AR-DRC code, principal diagnosis (ICD-10 codes), principal procedure, hospital Operational Data: CareType.

1. Data Preprocessing

* Handle Missing Values
* Encode Categorical Variables:
  + Use one-hot encoding for categorical features like DRG codes, diagnoses, procedures, and admission type.
* Normalise numerical features.

1. Model Selection

* Start with baseline model for interpretability.
* Use cross-validation to compare model performance based on the metrics. The model yields the highest accuracy rate is the final model.
* There are a lot of models that can be used to predict continuous variables, which includes:
  1. Linear Regression: We can use this model as a baseline model but might not capture complex relationships.
  2. Tree-Based Modes: Decision Trees: Simple and interpretable but prone to overfitting, Random Forests, Gradient Boosting Machines (e.g., XGBoost, LightGBM)

1. Model Training and Evaluation

* Split the Data into training Set: Typically 70-80% of the data and test Set: The remaining 20-30% for evaluating model performance.
* Since our target variable is continuous, the following metrics can be used:
  1. Mean Absolute Error (MAE): Average absolute difference between predicted and actual charges.
  2. Root Mean Squared Error (RMSE): Penalizes larger errors more than MAE.
  3. R-squared (R²): Proportion of variance explained by the model.
  4. Adjusted R-squared: Adjusts R² for the number of predictors in the model.

1. Model Optimisation

* Ensemble Mode:
  + We can build different models to predict the total cost and have a meta-learner.
  + The meta-learner will learn the predictions of each individual model and yield the final predictions.
* Hyperparameter Tuning:
* Optimize model parameters using grid search or random search.
* Model Interpretation
  + Identify which variables contribute most to the prediction to gain insights into cost drivers.
  + SHAP Values can gives us the top 3 drivers for the predictions for each patient.

1. Other predictions:

* Predict Length of Stay (LOS):
  + Even though LOS isn't strongly correlated with Total Charges, predicting LOS can help with bed management and resource planning.
* Predict Readmission Risk. If there is enough data, we can predict patients who are more likely to be re-admitted within 28 days. This is one of the most important indicators to monitor quality of care. The predictions can guide the interventions to improve patient outcomes and reduce additional costs associated with readmissions.

## Part 6: MLOps:

1. Prepare the Model for Production

* Use **MLFLow** to track the lifecycle of ML models. MLFlow enables us to log metrics, artifacts, parameters, ensuring that every version change is documented and easily retrieved.
* Model Tracking with MLflow: Use MLflow Tracking to log parameters, code versions, metrics, and artifacts from all the experiments. This creates a reproducible record of your model development process.
* Model Packaging: Save the trained model using MLflow's model packaging format. MLflow supports various model flavors (e.g., scikit-learn, TensorFlow, PyTorch), making it easier to serve or deploy the model later.

1. Model Registry with MLflow

* Model Registry in MLflow helps to manage different versions of the model. This allow us to transition models to ‘Staging’, ‘Production’ or ‘Archieved’

1. Containerization:

* Use MLflow's support for **Docker** to containerize the trained model along with its dependencies. This ensures consistency across development, testing, and production environments

1. Deployments:
   * Deploy the model as a RESTful API using MLflow's built-in serving capabilities. This can be done locally or on cloud platforms such as Databrick, Azure Machine Learning, Google Cloud AI Platform or AWS SageMaker.
   * Integrate with platforms like Kubernetes for scalable deployment.
   * Incorporate MLlow into CI/CD pipelines using GitHub Actions
2. Monitoring Model Performance and the System
   * Use Evidently AI to create interactive dashboards that track model performance metrics.
   * We also need to monitor data drift by generating a reports that highlight which features have drifted and the potential impact on model performance
   * Set up pipelines that retrain the model when Evidently AI detects significant data drift or performance drops.
   * Monitor system resources like CPU, memory, and GPU utilization using MLflow.
   * Use OpenTelemetry Framework to have an insight into the machine learning service.
   * Document the deployment process, monitoring setup, and incident response plans. Store this documentation alongside the MLflow projects.