1. Identify and describe 2 data quality issues present in the database. Briefly propose strategies to address these issues. Document the steps taken and provide a summary of the data quality improvements

The few data quality issues that stood out and the steps taken to rectify there are:

a. The columns Accommodation Charge, CCU Charges, ICU Charge, Theatre Charge, Prosthesis Charge, other Charges and Bundled Charges have inconsistent data types. Some values were stored in different formats which could lead to potential calculation issues. These data types were standardised to Decimal (18, 10) when data was ingested Microsoft SQL Server Management Studio, I standardized these columns into Decimal (18, 10). These errors can be avoided by implementing input validation and data type constraints during data entry or data import.

b. Out of 30,000 Pharmacy Charge columns 29,706 had anomalies. 12,971 values had scientific notations, 900 rows contained the value 'Error' and 17,635 records have Non-numeric charges. The lowest Pharmacy Charge value is around 10 Billion which indicates that the data quality issue most likely is arising during data entry or data import. I identified and flagged anomalous entries using SQL queries to invalid entries. It is essential to implement input constraints to ensure only valid numeric values are allowed during data entry. It will be beneficial to introduce data validation scripts to verify the data's consistency during import.

c. Bundled Charges and Theatre Charges have significant differences between their lowest and highest charges. However, after calculating the Z-scores, the data fell within the normal ranges indicating no significant outliers.

d. Some records for elderly patients had values in the InfantWeight column. This issue could be resolved by enforcing conditional data entry so the column is only enabled when the patient's age is less than 1 year.

e. When I analyse the number of duplicate episode IDs independently, there is a significant number of duplications. However, when combined with Principal Diagnosis and AR-DRG, the episode IDs are unique. For future calculations, episode ID will be used alongside one of these columns, ensuring that duplications do not impact the results.

f. There are records with invalid dates where the Admission Date is before the Date of Birth. There were several discrepancies where the reported age did not match the calculated age (based on Date of Birth and Admission Date). One provider Admission Provider ID 7045611 reported 7 Invalid Age which is the highest number of such errors. There are 4 Admission Provider IDs with 6 Invalid Age errors. These manual data entry errors can be eliminated if automated age calculation feature based on Admission Date and Date of Birth are introduced.

2. Using the data provided create a feature that could be valuable for analysis or modelling. Explain the rationale behind the feature you created and how they might be useful for analysis

It is critical to understand cost dynamics of patient care for benchmarking performance, identifying outliers and optimising resource allocations. I created a Charges Per Day feature to offer a detailed view of the expenses incurred during patient stays.

The Charges Per day aims to reflect daily cost of patient's stay by distributing the total charges over the length of stay. During data quality assessment(discussed in in 1.b) significant anomalies were found in the Pharmacy Charge column. Given the extent of these anomalies, it would be inappropriate to make assumptions about these values without further information. Including

Pharmacy Charges in the analysis could lead to inaccurate results, so it has been excluded from the per-day calculations.
Prosthesis Charges represent one-off costs associated with specific procedures and are not relevant for all Diagnosis-Related Groups (DRGs). Including these charges in the daily calculation would skew the results, especially for groups that do not require prosthesis. For this reason, Prosthesis Charges were excluded from the DailyTotalCharges metric, ensuring the analysis reflects ongoing daily care expenses rather than on-time costs. As a result two metrics were developed:

1. Daily Total Charges: Excludes both Pharmacy Charges and Prosthesis charges
2. Total Charges: Includes all charges except for Pharmacy Charges

The Charges Per Day feature takes the Length of Stay to ensure accurate cost distribution. In cases where admission date and separation date are the same the Daily Charges are treated as total charges for that admission. The charges per day enables benchmarking of costs across different providers, care types and principal diagnosis. By understanding which types of care incur the highest daily costs, stakeholders can target areas for efficiency improvements. It also helps detect outliers by highlighting extremely high and low daily charges. It also can be used to forecast future expenses, identify drivers of high costs and assess the financial impact of potential changes in care delivery.

The Office Hours provides information on the timing of admissions helping healthcare providers make data-driven decisions about resource allocation and service delivery. In this feature, I not only included whether the admission occurred during office hours but also whether it took place on the weekend. '1' represents admission during office hours and '0' represents out of office hours or on weekends.

3. Using the data provided produce a piece of analysis that describes to Ramsay which DRGs accrue the largest charges and your hypotheses for the drivers of these charges.

The data highlights that DR001 (Craniotomy), DR002(Spinal Procedures), and DRG003 (Vascular Procedures) consistently generate high Total Charges. These procedures require extensive resources, including specialized equipment, skilled surgical staff, and intensive post-operative care. Their complexity and the length of recovery results in higher charges due to increased staff, medication, and equipment costs.

A notable observation emerges when comparing Total Charges and Total Daily Charges between January 2023 and July 2024, the top DRGs by Total Charges were DRG001, DRG002, DRG003 in that order. However, between January 2024 and July 2024, although DRG001 and DRG002 continue to show high Total Daily Charges, DRG003 (Vascular Procedures) surpasses DRG002 in Total Charges. This pattern suggests that vascular procedures often involve high prosthesis-related costs that are recorded as one-off charges. This indicates the importance of managing prosthesis procurements and usage to control cost effectively for these procedures.

The data reveals that emergency care incurs higher charges for procedures like Craniotomy and Spinal Surgeries due to the urgency and complexity involved. These emergencies require immediate allocation of staff and equipment resulting in higher costs. On the other hand, vascular procedures are more prevalent in in-patient care settings, likely due to the extended recovery periods associated with these procedures. The analysis highlights the role out-patient care in certain cases. For example, craniotomy procedures show a high proportion of out-patient discharges, indicating advancements in minimally invasive surgical techniques or rapid discharge protocols that reduce hospital stays.

The analysis shows that between 2022 to 2024 the majority of admissions (65.76%) occurred outside office hours or on weekends, while only 34.24% took place during office hours. Year to date for 2024 shows a similar trend with 66.89% patients being admitted outside business hours and only 33.11% admitted during business hours. The high cost procedures like DRG001, DRG002 and DRG003 are associated with admissions outside office ours. The analysis also indicates that several high-cost admissions occur outside office hours particularly Kidney, urinary tract and respiratory conditions.

4. Based on your analysis, identify two strategic insights that could help Ramsay improve hospital operations or patient care. Justify your insights with evidence from your data analysis.

My analysis shows data inconsistencies such as incorrect or missing infant weight and discrepancies in reported patient age. These issues can occur when staff members are working under pressure make errors during manual entry. Automating specific data fields can significantly reduce these errors. Integrating an auto-fill function for calculating patient age based on admission date and date of birth ensures consistency and eliminates manual errors. Conditional input restrictions to enable infant weight only when patient is below 1 yr of age can improve data quality. This not only reduces burden on the staff but also improves operational efficiency.

The analysis also showed a significant number of high-cost and complex procedures where patients are transferred, discharged or ‘other’. Understanding the transfer and discharge destinations e.g. rehabilitation or palliative care can optimise care and reduce readmission rates. Patients transferred to rehabilitation centres benefit from specialized care that supports recovery while those transferred to palliative care might require additional coordination to manage chronic conditions. Therefore, a comprehensive follow-up program to monitor patient outcomes after transfer can provide valuable insights into care gaps and risks of readmission. Ramsay can explore partnerships with rehab facilities to strengthen care coordination or identify areas where rehabilitation care could be improved.

Further, the analysis showed that 65.76% of admissions occur outside of business hours between 2022 and 2024. The analysis showed that 67.75% YTD admissions for DRG001, DRG002 and DRG003 occurred outside office hours. This pattern suggests Ramsay needs to optimize after-hours operations to accommodate the increasing demand outside business hours. Ramsay can explore strengthening outpatient services for conditions that do not require immediate inpatient care like respiratory ailments. Offering extended outpatient hours for respiratory consultations during peak season (winter) can divert non-urgent cases for emergency admissions.