Part 1: Data Exploration and Preparation

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

E.g. Inconsistencies in data types, unusual values, or outliers.

The provided data file in xlsx format was reviewed. Below are the key data quality issues observed:

- Numerical columns like "AccommodationCharge", "CCU_Charges", "ICU_Charge", "TheatreCharge", and "PharmacyCharge" are stored as characters and have high number of significant figures. On the other hand, columns like "TheatreCharge" and "ProsthesisCharge", "OtherCharges", and "BundledCharges" generally have two digits after decimal point. Assuming charges are dollar values, the higher number of significant figures may not be relevant for business insights. To address this, values are rounded to 2 places of decimal and missing values are replaced with zero for these columns. Replacement with zero value is done as a missing value likely indicates no charge incurred.

For instance, for "AccommodationCharge", due to no missing values, no significant change is observed in distribution due to rounding. For "CCU_Charges", change in distribution is shown below in Figure 1 when "CCU_Charges_Filled" is created to insert zero for missing values.

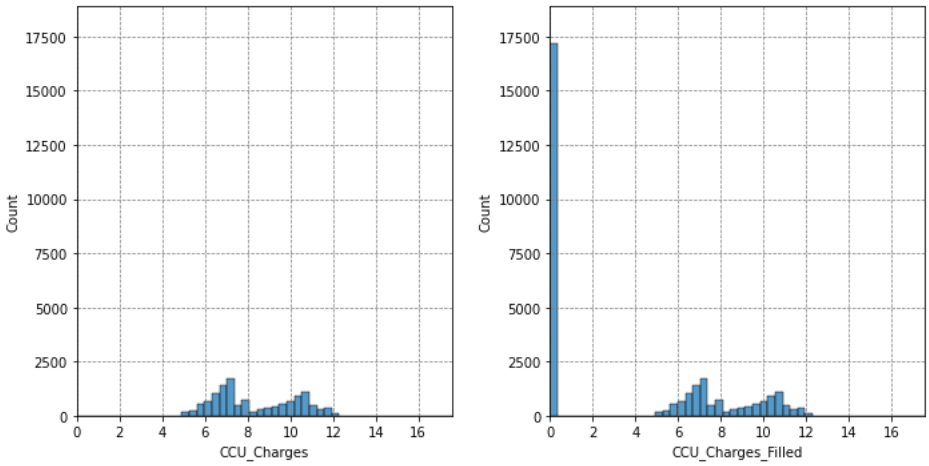


Fig 1: Distribution of 'CCU_Charges" before and after replacing missing values with zero

```
Missing data before:
 17167
Missing data after:
 0
Summary before cleaning
count    12833.000000
mean         8.344885
std          1.904278
min          5.019066
25%          6.808693
50%          7.765532
75%         10.201633
max         17.563012
Name: CCU_Charges, dtype: float64

Summary after cleaning
count    30000.000000
mean         3.569638
std          4.312475
min          0.000000
```

```
25%          0.000000
50%          0.000000
75%          7.240000
max         17.560000
Name: CCU_Charges_Filled, dtype: float64
```
Fig 2: Summary Statistics for columns “CCU_Charges” and “CCU_Charges_Filled”

- “PharmacyCharge” is order of magnitude higher than any other charge which is unrealistic and shows outliers compared to other charges. The column also contains “ERROR” values which are non-numeric. “ERROR” or missing values were replaced with zero to create “PharmacyCharge_Filled” to analyse the data. The data appears to be scaled as the data becomes comparable to other columns pertaining to hospital charges when a logarithmic transform is applied on this column to create “PharmacyCharge_Scaled”. More information may be required about the data generating process as to why these values are outliers. In absence of additional information, the “PharmacyCharge” derived columns are not used for further analysis as the high values will render the analysis inconclusive. Distribution and summary statistics for columns are shown in Figures 3 and 4, respectively.

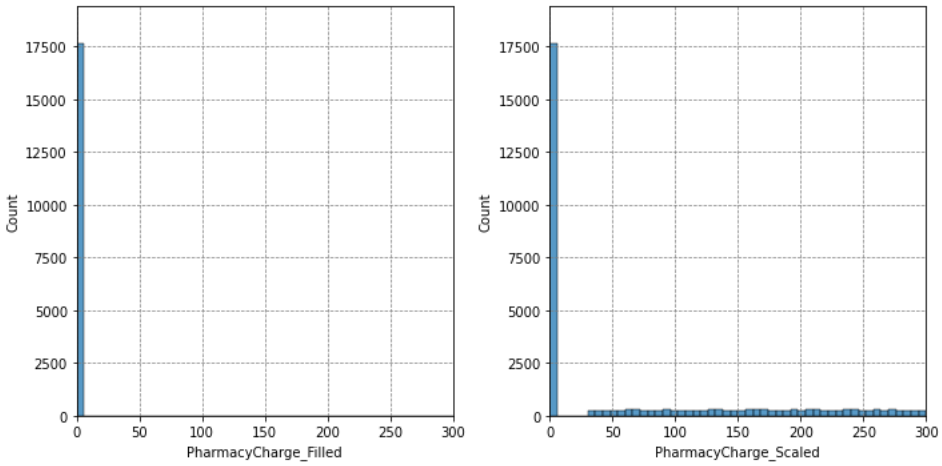


Fig 3: Distribution of “PharmacyCharge_Filled” before and after log scaling

```
Missing data before:
 0
Missing data after:
 0

Summary before cleaning
count     3.000000e+04
mean     2.947087e+127
std      5.582371e+128
min       0.000000e+00
25%       0.000000e+00
50%       0.000000e+00
75%       4.395469e+59
max      2.130293e+130
Name: PharmacyCharge_Filled, dtype: float64
Summary after cleaning
count    30000.000000
mean        68.227829
std         95.511689
min          0.000000
25%          0.000000
50%          0.000000
75%        137.333080
max        300.092322
Name: PharmacyCharge_Scaled, dtype: float64
```
Fig 4: Summary Statistics for columns “PharmacyCharge_Filled” and “PharmacyCharge_Scaled”

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.

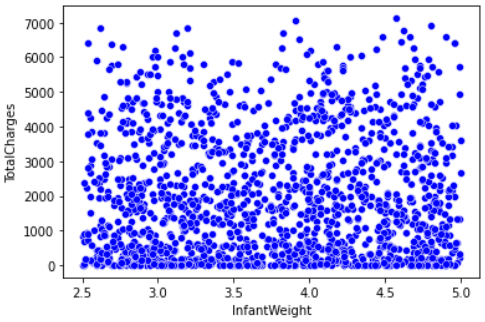
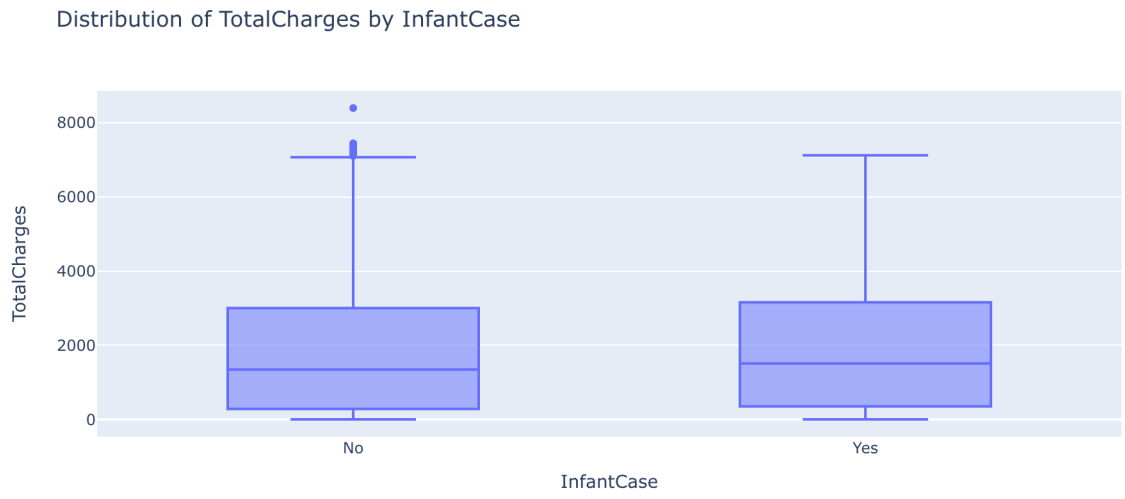


Fig 5a: Scatter plot of "TotalCharges" column with values for "InfantWeight" column



	InfantCase	Median_Duration	Median_Charge	Count_Cases
0	No	373.840000	1349.34	28475
1	Yes	365.241111	1510.89	1525

Fig 5b: Comparison of "TotalCharges" column with values for "InfantCase" column

The "InfantWeight" attribute was analysed to evaluate its impact on "TotalCharges". A scatter plot between both attributes does not show any identifiable correlation (Figure 5a). However, presence of a value in "InfantWeight" column does have an impact as a higher value for "TotalCharges" which seems logical as specialist childcare is likely to incur additional charges. Hence, "InfantCase" binary feature is created using the "InfantWeight" column to indicate whether infant weight has been recorded. The "InfantCase" column contains Yes/No values. On comparison of distribution of "TotalCharges" column (calculated to sum all cost incurred during the episode excluding pharmacy charges), presence of "InfantWeight" indicates a higher value for "TotalCharges" as shown in boxplot in Figure 5b with a higher median and upper quartile for "Yes" values. The median value for "Yes" values is 1510.89 whereas the median value for "No" values is 1349.34.

Evaluation of statistical significance of impact of "InfantCase" on "TotalCharges" is done with t-test and Shapiro-Wilk test (Figure 6). Results shows that for TotalCharges, there is significant difference due to presence of Infant Case as we get a p-value of 0.06 which is lower than an acceptable significance level of 0.1. However, Shapiro-Wilk test indicates data is not normally distributed which is an assumption for validity of t-test. For LengthofStay_hours,

there is no significant difference due to presence of Infant Case. However, Shapiro-Wilk test indicates data is not normally distributed which is an assumption for validity of t-test.

Based on the t-test conducted, “InfantCase” can be used to analyse “TotalCharges”. Although, t-test results do not reliably indicate a statistically significant outcome, “InfantCase” can be used as a valuable input for any predictive models for aiming to predict total charge incurred as a function of other input columns.

```
Comparison for TotalCharges
T-statistic value:  1.8426750503868634
P-Value:  0.06538633008446323
Shapiro-Wilk Test for Group Yes: ShapiroResult(statistic=0.9000214338302612,
pvalue=2.0093349361204517e-30)
Shapiro-Wilk Test for Group No: ShapiroResult(statistic=0.8845122456550598, pvalue=0.0)
Comparison for LengthofStay_hours
T-statistic value:  -0.49960829083498354
P-Value:  0.6173545831556745
Shapiro-Wilk Test for Group Yes: ShapiroResult(statistic=0.9579210877418518,
pvalue=1.4272931999355735e-20)
Shapiro-Wilk Test for Group No: ShapiroResult(statistic=0.9555138349533081, pvalue=0.0)
```

Fig 6: Summary of t-test done to analyse effect of “InfantCase” on “TotalCharges” values

Part 2: Data Analysis and Visualisation

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 charge. Visualise these trends using appropriate charts or graphs and describe the results.

Figures 7a and 7b show the AR-DRGs contributing to highest and lowest median values of "TotalCharges". The bar charts show that groups like I65B, K11B, O63A, U60Z, and B42C have the highest median charges while groups X40B, Z64B, D61A, E76A, and D14B have the lowest median charges.

On further investigation, the AR-DRGs accruing highest median charges are found to be associated with cases of "Rehabilitative" or "Palliative" discharge. For these cases, "BundledCharges" tend to have the highest contribution to total charges. This may be caused by these AR-DRGs being more likely to be needing urgent or critical care with prosthesis or rehabilitation needs which contribute to the higher charges incurred. These are being exhibited in terms of higher values for "BundledCharges" and "ProsthesisCharge". On the other hand, DRGs accruing lowest median charges are more likely to be associated with lower values of "ProsthesisCharge" and lower value of "BundledCharges". This may be caused by these groups being associated with patients with requirements of elective surgeries or non-critical care.

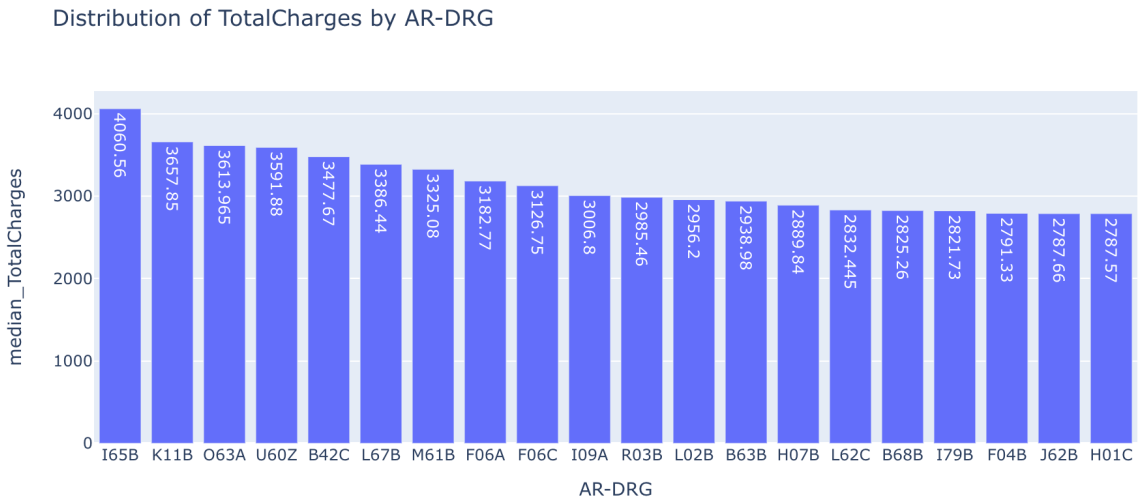


Fig 7a: Comparison of "TotalCharges" to identify AR-DRGs with highest median values of "TotalCharges"

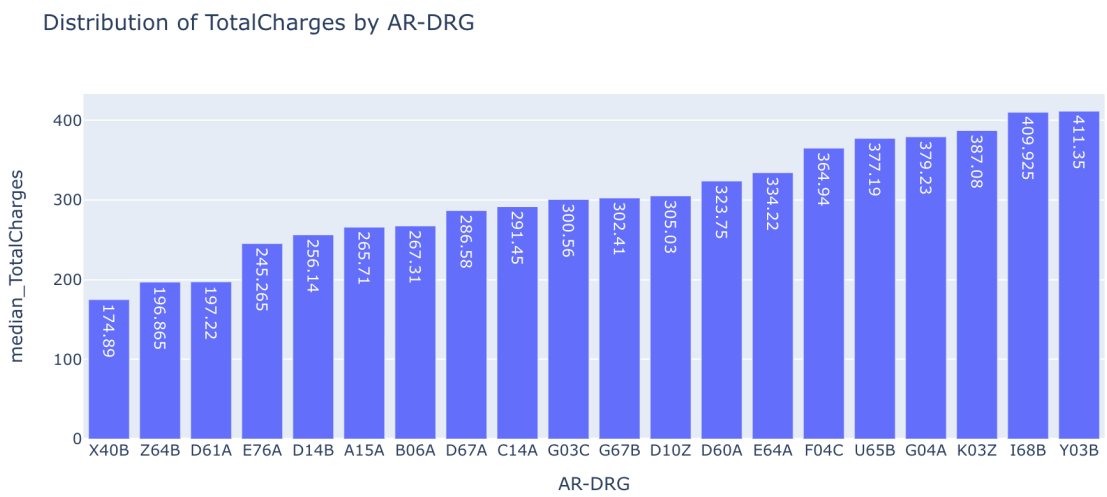


Fig 7b: Comparison of “TotalCharges” to identify AR-DRGs with lowest median values of “TotalCharges”

Part 3

4. Write an SQL query to calculate the total and average admissions for each month over the last two years. Include the month and year in the results.

The provided excel file was directly loaded for querying. Below query in SQLite includes required steps for data processing.

```
WITH calc_date_parts AS (
SELECT
CAST(substr(AdmissionDate,instr(AdmissionDate,  '/')+1,instr(substr(AdmissionDate,instr(AdmissionDate,  '/')+1,7) ,'/')-1)  AS  INTEGER)AS
Month,
CAST(substr(AdmissionDate,instr(AdmissionDate,  '/')+ instr(substr(AdmissionDate,instr(AdmissionDate,  '/')+1,7) ,'/')+1,4)  As  INTEGER)AS
Year,
COUNT(episode_id) as Total_Admissions
FROM "Data Insights - Synthetic Dataset"
GROUP BY substr(AdmissionDate,instr(AdmissionDate, '/')+1,instr(substr(AdmissionDate,instr(AdmissionDate, '/')+1,7) ,'/')-1),
substr(AdmissionDate,instr(AdmissionDate, '/')+ instr(substr(AdmissionDate,instr(AdmissionDate, '/')+1,7) ,'/')+1,4)
ORDER BY
CAST(substr(AdmissionDate,instr(AdmissionDate, '/')+ instr(substr(AdmissionDate,instr(AdmissionDate, '/')+1,7) ,'/')+1,4)  AS INTEGER) ASC,
CAST(substr(AdmissionDate,instr(AdmissionDate, '/')+1,instr(substr(AdmissionDate,instr(AdmissionDate, '/')+1,7) ,'/')-1) AS INTEGER) ASC
),
calc_date_column AS (
SELECT Month, Year,
date(Year||'-'||printf('%02d',Month)||'-01') AS Date,
strftime('%d', date(date(Year||'-'||printf('%02d',Month)||'-01'), 'start of month', '+1 month', '-1 day')) AS days_in_month,
Total_Admissions
FROM calc_date_parts
)
SELECT Month, Year, Total_Admissions,
ROUND(Total_Admissions/days_in_month,2) as Avg_Admissions
FROM calc_date_column;
```

5. Write an SQL query to analyse the distribution of TotalCharges by PrincipalDiagnosis and Sex. Use percentiles to describe the distribution.

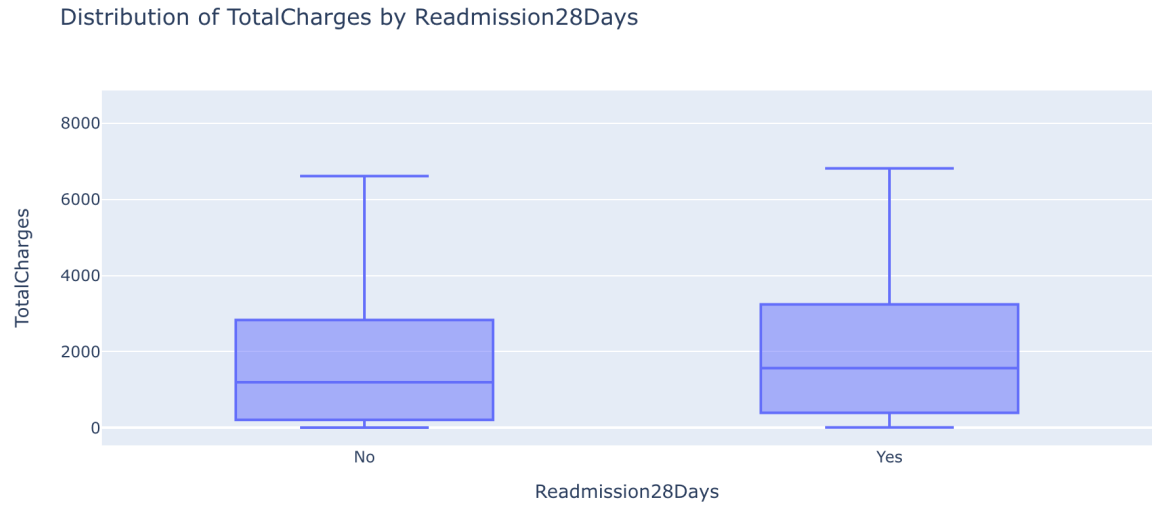
The provided excel file was directly loaded for querying. Below query in SQLite includes required steps for data processing. Below query summarises the PrincipalDiagnosis and Sex values which correspond to quantile values of TotalCharges. For individual combinations of PrincipalDiagnosis and Sex, the number of cases were generally small which means identifying quantiles does not have provide any significant business insight.

```
WITH calc_charges AS(
SELECT PrincipalDiagnosis, Sex,
CASE WHEN AccommodationCharge IS NULL THEN 0.0 ELSE CAST(AccommodationCharge AS REAL) END  As 'AccCharge' ,
CASE WHEN CCU_Charges IS NULL THEN 0.0 ELSE CAST(CCU_Charges AS REAL) END  As 'CCU_Charges' ,
CASE WHEN ICU_Charge IS NULL THEN 0.0 ELSE CAST(ICU_Charge AS REAL) END  As 'ICU_Charge' ,
CASE WHEN TheatreCharge IS NULL THEN 0.0 ELSE CAST(TheatreCharge AS REAL) END  As 'TheatreCharge' ,
CASE WHEN ProsthesisCharge IS NULL THEN 0.0 ELSE CAST(ProsthesisCharge AS REAL) END  As 'ProsthesisCharge' ,
CASE WHEN OtherCharges IS NULL THEN 0.0 ELSE CAST(OtherCharges  AS REAL) END  As 'OtherCharges' ,
CASE WHEN BundledCharges IS NULL THEN 0.0 ELSE CAST(BundledCharges AS REAL) END  As 'BundledCharges'
FROM  "Data Insights - Synthetic Dataset"
),
calc_total AS(
SELECT PrincipalDiagnosis, Sex,
AccCharge + CCU_Charges+ ICU_Charge + TheatreCharge + ProsthesisCharge + OtherCharges +  BundledCharges  As Total_Charges
FROM calc_charges),
calc_percentile AS(
SELECT PrincipalDiagnosis, Sex, AVG(Total_Charges) AS Mean_Total_Charges,
NTILE(100) OVER (ORDER BY AVG(Total_Charges) ) AS 'Percentile'
FROM calc_total
GROUP BY PrincipalDiagnosis, Sex
ORDER BY Percentile)

SELECT
Percentile,
MIN(Mean_Total_Charges),
PrincipalDiagnosis,
Sex
FROM calc_percentile
WHERE Percentile IN (1,25, 50, 75,100)
GROUP BY PrincipalDiagnosis,Sex
ORDER BY Percentile,MIN(Mean_Total_Charges), PrincipalDiagnosis,Sex;
```

Part 4: Strategic Insights and Recommendations

6. 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.



	Readmission28Days	Median_Duration	Median_Charge	Count_Cases
0	No	392.108889	1198.70	487
1	Yes	397.816944	1565.62	524

Fig 8a: Comparison of "TotalCharges" to identify impact of "Readmission28Days" column

```
Comparison for TotalCharges
T-statistic value:  2.018870439379388
P-Value:  0.04376478929991035
Shapiro-Wilk Test for Group Yes: ShapiroResult(statistic=0.9053027033805847,
pvalue=1.5931930386051066e-17)
Shapiro-Wilk Test for Group No: ShapiroResult(statistic=0.8692502975463867,
pvalue=8.748911624388646e-20)
Comparison for LengthofStay_hours
T-statistic value:  0.55522826239376
P-Value:  0.5788616088548534
Shapiro-Wilk Test for Group Yes: ShapiroResult(statistic=0.9497412443161011,
pvalue=2.3645590693788243e-12)
Shapiro-Wilk Test for Group No: ShapiroResult(statistic=0.9519455432891846,
pvalue=1.7404381755237175e-11)
```

Fig 8b: Summary of t-test done to analyse effect of "Readmission28Days" on "TotalCharges"

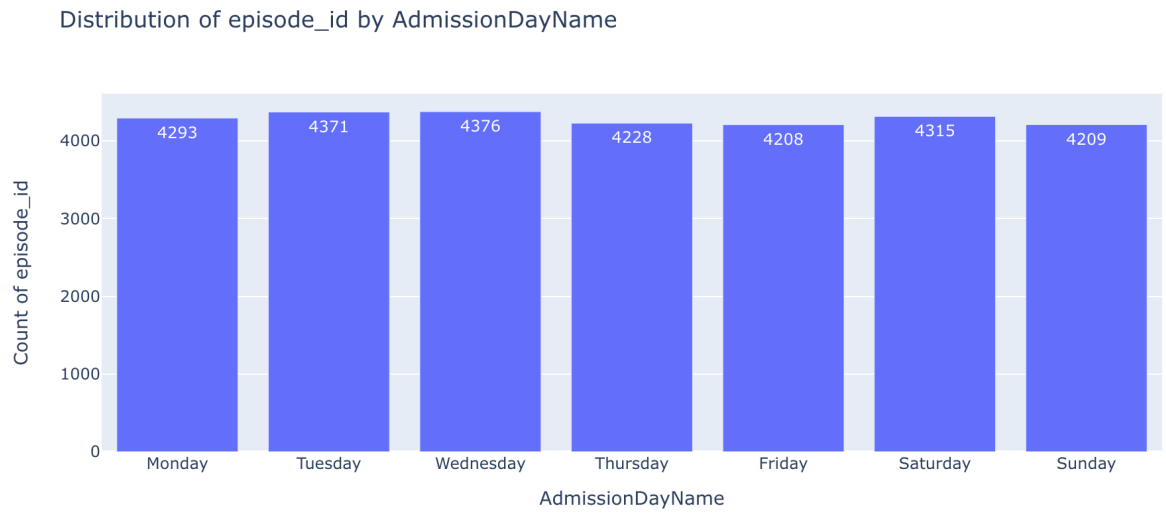
The effect of values in "Readmission28Days" column is analysed on total charge and length of duration. It is assumed that the column indicates whether the patient was readmitted for care after an initial care period of 28 days or more. Figure 8a shows that for "Yes" values, the median charge incurred is higher compared to "No" values although a significant difference in not seen in the median duration of stay in hours. On conducting the t-test for statistical significance (Figure 8b), the difference in total charges is found to be significant although the normal distribution assumption may not be valid.

The observation can lead to following conclusions:

a) This indicates a high degree of confidence is shown by the patients on the staff and facilities as they are ready to come back for care.
b) This also provides an opportunity for improvement. Investigation may be warranted to identify causes as to why the patients need to return for care and if the diagnosis procedure could be improved to ensure that patients do not need to come back for readmission. This will allow the facility to take more efficient care of the patients.

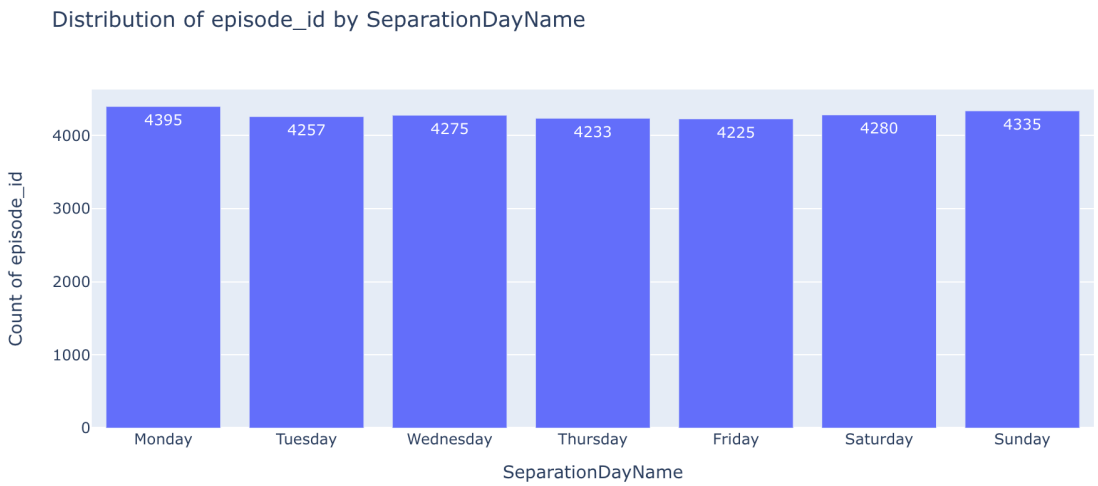
Additionally, the number of episodes recorded for admission and separation over the days for a week were analysed on Figures 9a and 9b, respectively. This indicates that the highest number of median admissions are recorded on Tuesday, Wednesday and Saturday while comparing admissions on days of a week. The highest number of median admissions is on Wednesday is 4,376 which is around 4% higher than lowest number of median admissions recorded on Sundays as 4,209. Similarly, for separations, the highest number of median separations are recorded on Sundays and Mondays while comparing separations on days of a week. The highest number of median separations are recorded on Mondays as 4,395 which is around 4% higher than lowest number of median separations recorded on Fridays to be 4,225 over the period of analysis.

The analysis can be used for workforce and resource planning during the week as additional manpower and resources will need to be diverted to plan for expected admissions or separations over the course of a week.



	AdmissionDayName	Median_Duration	Median_Charge	Count_Cases
0	Monday	374.925278	1363.470	4293
1	Tuesday	371.552778	1331.090	4371
2	Wednesday	372.462500	1379.635	4376
3	Thursday	372.557083	1353.330	4228
4	Friday	380.125278	1394.470	4208
5	Saturday	374.308889	1333.280	4315
6	Sunday	367.547500	1364.020	4209

Fig 9a: Comparison of number of episodes recorded for admission by day of week



	SeparationDayName	Median_Duration	Median_Charge	Count_Cases
0	Monday	385.171944	1295.96	4395
1	Tuesday	363.428611	1329.45	4257
2	Wednesday	364.997778	1405.59	4275
3	Thursday	361.652778	1350.01	4233
4	Friday	384.652222	1408.13	4225
5	Saturday	366.749306	1340.96	4280
6	Sunday	380.476667	1382.81	4335

Fig 9b: Comparison of number of episodes recorded for separation by day of week