Ramsay Data Insights
14/10/2024

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.

   Below are the issues I found:

- Data duplication:
  AdmissionProviderID and other data fields present data duplication.

- Data inaccuracy:
  infantWeight column doesn't belong to this dataset analysis I can't find any relationship with age column or other columns.
  PharmacyCharge field presents some data inaccuracy such as symbols and scientific decimal number, "errors" in some fields.
  Noted: I have removed those error and symbols presented in numeric data as part of data cleansing for this test.
  DOB values are bigger than the AdmissionDate by 2 years as couldn't be possible admitting the patient 2 years before being born yet, this could sometimes be a human error.

To enhance the data quality and reliability it is very important to assess the data carefully at the beginning to allow identify missing values, different scales, duplication to ensure the data is clean and clear to assist with more meaningful information and improve the model's performance.

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.

   As the main goal for this test is to provide data analysis skills I have chosen to upload the data information using SQL queries to focus on the data analysis rather than graphical representation, however I have also uploaded a Powerbi file to show some an initial visual representation which can be completed with better visualization for future dashboards if it is required.

   Create a feature representing the *ChargePerDay* (i.e., the total charges divided by *LengthOfStay*).

```
/*this query represents the ChargePerDay and the LenghtOfStay per patient**/

 select *, totalcharge/totaldays as totalchargeperday from
 (select [episode_id], [Postcode], DATEDIFF(day,[AdmissionDate],
[SeparationDate]) as 'totaldays', isnull([AccommodationCharge],0) +
isnull([CCU_Charges],0) + isnull([ICU_Charge],0) +
isnull([PharmacyChargeUPDATED],0) + isnull([BundledCharges],0) +
isnull([TheatreCharge],0) + isnull([ProsthesisCharge],0) +
isnull(CAST([OtherCharges] as float),0) as 'totalcharge' from
[RansayData].[dbo].[Data Insights - Synthetic Dataset]) as TableA
```

	episode_id	Postcode	totaldays	totalcharge	totalchargeperday
1	78773	64012	7	1799.084829198	257.012118456857
2	325	37800	29	388.39959232	13.3930893903448
3	44678	31072	6	3766.964215049	627.827369174833
4	83603	33340	25	105.11569335	4.204627734
5	85397	58698	8	792.026914529	99.003364316125
6	52194	7089	23	495.52774974	21.5446847713043
7	80646	83035	25	1349.11271026	53.9645084104
8	76477	31727	3	549.56835273	183.18945091
9	22611	47484	7	831.227713101	118.746816157286
10	53407	97304	24	135.25426756	5.63559448166667
11	20264	28533	27	25.62584818	0.949105488148148
12	42090	18988	6	10.833652365	1.8056087275
13	51948	53423	17	879.997531512	51.7645606771765
14	12643	46858	11	1238.822092882	112.620190262
15	31747	37070	20	1669.17811077	83.4589055385

Create a binary feature indicating whether the *AdmissionTime* is during office hours (e.g., 9 AM to 5 PM) or outside of office hours.

```
/**This query takes a binary to indicate AdmissionTime result of 0 for “after hours” and 1 for    “office hours” **/

select [episode_id], [Postcode], [AdmissionTime], DATEPART(hour, [AdmissionTime]) as AdmissionHour,
CASE WHEN DATEPART(hour, [AdmissionTime]) >= 9 AND DATEPART(hour, [AdmissionTime]) <= 17 THEN 1 ELSE 0 END AS Officehours
FROM [RansayData].[dbo].[Data Insights - Synthetic Dataset]
```

	episode_id	Postcode	AdmissionTime	AdmissionHour	Officehours
1	78773	64012	02:11:14.0000000	2	0
2	325	37800	19:25:01.0000000	19	0
3	44678	31072	10:06:06.0000000	10	1
4	83603	33340	04:15:55.0000000	4	0
5	85397	58698	02:50:27.0000000	2	0
6	52194	7089	13:21:14.0000000	13	1
7	80646	83035	22:24:27.0000000	22	0
8	76477	31727	13:08:42.0000000	13	1
9	22611	47484	08:09:00.0000000	8	0
10	53407	97304	04:31:00.0000000	4	0
11	20264	28533	15:02:01.0000000	15	1
12	42090	18988	03:04:53.0000000	3	0
13	51948	53423	00:50:05.0000000	0	0
14	12643	46858	00:10:43.0000000	0	0
15	31747	37070	10:52:00.0000000	10	1
16	18801	88586	12:11:09.0000000	12	1
17	87931	65634	18:24:47.0000000	18	0
18	28280	55313	01:28:26.0000000	1	0
19	18925	50602	23:07:19.0000000	23	0
20	83685	45720	14:02:53.0000000	14	1

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. Explore how `TotalCharges` changes over different caretype and modeofseparation.

```
	/**Explore how TotalCharges changes over different caretype and
modeofseparation for DRG**/
	select [episode_id], [Postcode], [AR_DRG], [ModeOfSeparation], [CareType],
	DATEDIFF(day,[AdmissionDate], [SeparationDate]) as 'totaldays',
	isnull([AccommodationCharge],0) + isnull([CCU_Charges],0) +
	isnull([ICU_Charge],0) + isnull([PharmacyChargeUPDATED],0) +
	isnull([BundledCharges],0) + isnull([TheatreCharge],0) +
	isnull([ProsthesisCharge],0) + isnull(CAST([OtherCharges] as float),0) as
	'totalcharge' from [RansayData].[dbo].[Data Insights - Synthetic Dataset]
```

	episode_id	Postcode	AR_DRG	ModeOfSeparation	CareType	totaldays	totalcharge
1	78773	64012	C63A	Other	Inpatient	7	1799.084829198
2	325	37800	P05A	Other	Outpatient	29	388.39959232
3	44678	31072	B03C	Transfer	Emergency	6	3766.964215049
4	83603	33340	B80A	Other	Inpatient	25	105.11569335
5	85397	58698	DRG002	Transfer	Inpatient	8	792.026914529
6	52194	7089	DRG002	Other	Emergency	23	495.52774974
7	80646	83035	G06Z	Transfer	Emergency	25	1349.11271026
8	76477	31727	DRG002	Transfer	Inpatient	3	549.56835273
9	22611	47484	DRG001	Transfer	Emergency	7	831.227713101
10	53407	97304	DRG001	Other	Inpatient	24	135.25426756
11	20264	28533	DRG003	Discharge	Emergency	27	25.62584818
12	42090	18988	DRG001	Discharge	Emergency	6	10.833652365
13	51948	53423	DRG001	Transfer	Inpatient	17	879.997531512
14	12643	46858	K40A	Other	Emergency	11	1238.822092882
15	31747	37070	B78C	Transfer	Outpatient	20	1669.17811077
16	18801	88586	C15B	Other	Inpatient	27	1527.18038271
17	87931	65634	DRG001	Other	Outpatient	23	1629.037830737
18	28280	55313	DRG001	Other	Inpatient	6	1508.98441595
19	18925	50602	DRG001	Other	Outpatient	8	2158.253211574
20	83685	45720	DRG002	Other	Emergency	3	89.27907549

Analyze changes in the *LengthOfStay* over different DRGs.
Identify a trend of interest based on your understanding of the data.

```
/*this query represents the chargeperday and the totaldays, LengthOfStay over
different DRGs**/

select *, totalcharge/totaldays as totalchargeperday from
(select [episode_id], [Postcode], [AR_DRG], DATEDIFF(day,[AdmissionDate],
[SeparationDate]) as 'totaldays', isnull([AccommodationCharge],0) +
isnull([CCU_Charges],0) + isnull([ICU_Charge],0) +
isnull([PharmacyChargeUPDATED],0) + isnull([BundledCharges],0) +
isnull([TheatreCharge],0) + isnull([ProsthesisCharge],0) +
isnull(CAST([OtherCharges] as float),0) as 'totalcharge' from
[RansayData].[dbo].[Data Insights - Synthetic Dataset]) as TableA
```

	episode_id	Postcode	AR_DRG	totaldays	totalcharge	totalchargeperday
1	78773	64012	C63A	7	1799.084829198	257.012118456857
2	325	37800	P05A	29	388.39959232	13.3930893903448
3	44678	31072	B03C	6	3766.964215049	627.827369174833
4	83603	33340	B80A	25	105.11569335	4.204627734
5	85397	58698	DRG002	8	792.026914529	99.003364316125
6	52194	7089	DRG002	23	495.52774974	21.5446847713043
7	80646	83035	G06Z	25	1349.11271026	53.9645084104
8	76477	31727	DRG002	3	549.56835273	183.18945091
9	22611	47484	DRG001	7	831.227713101	118.746816157286
10	53407	97304	DRG001	24	135.25426756	5.63559448166667
11	20264	28533	DRG003	27	25.62584818	0.949105488148148
12	42090	18988	DRG001	6	10.833652365	1.8056087275
13	51948	53423	DRG001	17	879.997531512	51.7645606771765
14	12643	46858	K40A	11	1238.822092882	112.620190262
15	31747	37070	B78C	20	1669.17811077	83.4589055385
16	18801	88586	C15B	27	1527.18038271	56.5622363966667
17	87931	65634	DRG001	23	1629.037830737	70.8277317711739
18	28280	55313	DRG001	6	1508.98441595	251.497402658333
19	18925	50602	DRG001	8	2158.253211574	269.78165144675
20	83685	45720	DRG002	3	89.27907549	29.75969183

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.

```
/** This query calculates the total number of days for each month, finds the total
amount of admission for each month and then finds the average of admission for each
month including Year and the Month**/

WITH My_CTE (MM, YY, Average, Total) as (
SELECT
     MONTH(AdmissionDate) as MM,
     YEAR(AdmissionDate) as YY,
     AVG(CAST(COUNT(DISTINCT Episode_id) AS FLOAT)) OVER (ORDER BY
     MONTH(AdmissionDate),YEAR(AdmissionDate)) AS Average,
     COUNT(Distinct Episode_id) as Total
FROM [RansayData].[dbo].[Data Insights - Synthetic Dataset]
WHERE
     AdmissionDate >= DATEADD(YEAR, -2, GETDATE())
GROUP BY
     MONTH(AdmissionDate),YEAR(AdmissionDate))
       select CONCAT(FORMAT(MM, '00') ,'-' , FORMAT(YY,'0000')) as MMYY,
       FORMAT(Average,'N2') as Average, Total from
My_CTE  order by YY , MM
```

	MMYY	Average	Total
1	10-2022	1,211.12	676
2	11-2022	1,217.89	1235
3	12-2022	1,216.62	1233
4	01-2023	1,243.00	1243
5	02-2023	1,219.67	1159
6	03-2023	1,214.20	1250
7	04-2023	1,227.86	1243
8	05-2023	1,234.89	1263
9	06-2023	1,246.00	1259
10	07-2023	1,250.85	1278
11	08-2023	1,248.93	1217
12	09-2023	1,244.56	1179
13	10-2023	1,216.94	1316
14	11-2023	1,215.80	1176
15	12-2023	1,218.23	1252
16	01-2024	1,250.00	1257
17	02-2024	1,205.25	1162
18	03-2024	1,225.33	1281
19	04-2024	1,231.38	1256
20	05-2024	1,244.70	1333

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

```
/** Query to analyse the distribution of TotalCharges by PrincipalDiagnosis and Sex.


WITH Percentiles AS (
    SELECT
        PrincipalDiagnosis,
        Sex,   ISNULL(AccommodationCharge,0) + ISNULL(CCU_Charges,0) +
ISNULL(ICU_Charge,0) + ISNULL(TheatreCharge,0) +ISNULL(ProsthesisCharge,0) +
ISNULL(OtherCharges,0) + ISNULL(BundledCharges,0)  as Total,
        PERCENTILE_CONT(0.25) WITHIN GROUP (ORDER BY ISNULL(AccommodationCharge,0) +
ISNULL(CCU_Charges,0) + ISNULL(ICU_Charge,0) + ISNULL(TheatreCharge,0)
+ISNULL(ProsthesisCharge,0) + ISNULL(OtherCharges,0) + ISNULL(BundledCharges,0))
OVER (PARTITION BY PrincipalDiagnosis, Sex) AS Percentile_25,
        PERCENTILE_CONT(0.50) WITHIN GROUP (ORDER BY ISNULL(AccommodationCharge,0) +
ISNULL(CCU_Charges,0) + ISNULL(ICU_Charge,0) + ISNULL(TheatreCharge,0)
+ISNULL(ProsthesisCharge,0) + ISNULL(OtherCharges,0) + ISNULL(BundledCharges,0))
OVER (PARTITION BY PrincipalDiagnosis, Sex) AS Median,
        PERCENTILE_CONT(0.75) WITHIN GROUP (ORDER BY ISNULL(AccommodationCharge,0) +
ISNULL(CCU_Charges,0) + ISNULL(ICU_Charge,0) + ISNULL(TheatreCharge,0)
+ISNULL(ProsthesisCharge,0) + ISNULL(OtherCharges,0) + ISNULL(BundledCharges,0))
OVER (PARTITION BY PrincipalDiagnosis, Sex) AS Percentile_75
    FROM
        [RansayData].[dbo].[Data Insights - Synthetic Dataset]
)
SELECT
    PrincipalDiagnosis,
    Sex,
    FORMAT(MIN(Total), 'C', 'en-US')  AS Min_Charges,
    FORMAT(MAX(Total), 'C', 'en-US')  AS Max_Charges,
    FORMAT(AVG(Total), 'C', 'en-US')  AS Average_Total_Charges,
    FORMAT(Percentile_25, 'C', 'en-US') as Percentile_25,
    FORMAT(Median, 'C', 'en-US') as Median ,
    FORMAT(Percentile_75, 'C', 'en-US') as Percentile_75
FROM
     Percentiles
GROUP BY
    PrincipalDiagnosis,
    Sex,
    Percentile_25,
    Median,
    Percentile_75
ORDER BY
    PrincipalDiagnosis,
    Sex;
```

Judith Rios

	PrincipalDiagnosis	Sex	Min_Charges	Max_Charges	Average_Total_Charges	Percentile_25	Median	Percentile_75
	A00.1	F	$1,582.88	$1,582.88	$1,582.88	$1,582.88	$1,582.88	$1,582.88
	A00.2	F	$1,314.93	$1,314.93	$1,314.93	$1,314.93	$1,314.93	$1,314.93
	A00.2	M	$3,385.07	$3,385.07	$3,385.07	$3,385.07	$3,385.07	$3,385.07
	A00.4	F	$550.70	$1,139.79	$845.25	$697.97	$845.25	$992.52
	A00.4	M	$3,011.64	$3,011.64	$3,011.64	$3,011.64	$3,011.64	$3,011.64
	A00.5	F	$3,960.19	$3,960.19	$3,960.19	$3,960.19	$3,960.19	$3,960.19
	A00.6	F	$2,445.66	$2,445.66	$2,445.66	$2,445.66	$2,445.66	$2,445.66
	A00.6	M	$632.05	$632.05	$632.05	$632.05	$632.05	$632.05
	A00.9	F	$1,552.99	$1,552.99	$1,552.99	$1,552.99	$1,552.99	$1,552.99
0	A00.9	M	$1,427.93	$1,427.93	$1,427.93	$1,427.93	$1,427.93	$1,427.93
1	A01.0	M	$4,104.46	$4,104.46	$4,104.46	$4,104.46	$4,104.46	$4,104.46
2	A01.1	F	$1,783.58	$1,783.58	$1,783.58	$1,783.58	$1,783.58	$1,783.58
3	A01.1	M	$3,355.30	$3,355.30	$3,355.30	$3,355.30	$3,355.30	$3,355.30
4	A01.2	F	$213.75	$1,808.76	$1,011.25	$612.50	$1,011.25	$1,410.00
5	A01.5	F	$334.52	$334.52	$334.52	$334.52	$334.52	$334.52
6	A01.5	M	$627.40	$6,104.17	$3,365.78	$1,996.59	$3,365.78	$4,734.98
7	A01.7	F	$232.50	$232.50	$232.50	$232.50	$232.50	$232.50

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.

Strategies:

Data validation technique

- ✓ To prevent from making human errors on dates If that is the reason why is not presenting logical sequence, adding a constraint or mandatory fields to alert the user who is entering the data in admission that DOB value can't be greater than the date of admission date would fix this issue.

Eliminating data redundancy or duplication:

- ✓ There are different techniques to remove duplication from data normalization with SQL data modelling to using disctinctcount measures using Powerbi data visualization tool. Removing duplication with normalization will assist with data storage effectiveness and preventing from deleting a data information when updating current information or deleting a admission information.