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.

- 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.
  PharmacyCharge field is presenting data inaccuracy as well presenting symbols together with scientific decimal number, and some errors in some fields, for this test I have removed those error and symbols presented in numeric data as part of data cleansing.
  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.

Strategies:

- ✓ To prevent from making human errors on dates adding a constraint or mandatory fields is necessary to alert the user who is entering the data in admission that DOB value can't be greater than the date of admission date.

- ✓ To remove duplication data normalization can fix this problem or working with disctinctcount measures for powerbi reports. 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.

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 feature chosen provides 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.

   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 and average admissions for each month over the
last two years.
    Include the month and year in the results. **/

SELECT
YearAdmission,
MonthAdmission,
DaysMonth,
COUNT(episode_id) as TotalPerMonth,
COUNT(episode_id)/DaysMonth as AverageAdmissionperday
FROM
(SELECT episode_id,AdmissionDate,YEAR(AdmissionDate) as YearAdmission,
MONTH(AdmissionDate) as MonthAdmission,DAY(EOMONTH(AdmissionDate)) as DaysMonth FROM
RansayData.[dbo].[Data Insights - Synthetic Dataset]) AS TableFechas
GROUP BY YearAdmission,MonthAdmission,DaysMonth
```

	YearAdmission	MonthAdmission	DaysMonth	TotalPerMonth	AverageAdmissionperday
1	2024	1	31	1260	40
2	2024	3	31	1287	41
3	2022	12	31	1244	40
4	2023	2	28	1166	41
5	2022	11	30	1245	41
6	2024	2	29	1171	40
7	2022	9	30	1236	41
8	2023	6	30	1267	42
9	2023	12	31	1257	40
10	2023	11	30	1187	39
11	2023	4	30	1252	41
12	2022	8	31	1215	39
13	2023	5	31	1272	41
14	2024	6	30	1286	42
15	2022	10	31	1238	39
16	2023	9	30	1189	39
17	2023	7	31	1287	41
18	2023	1	31	1252	40
19	2023	3	31	1259	40
20	2023	8	31	1233	39

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.
      Use percentiles to describe the distribution individually**/

SELECT PrincipalDiagnosis,Sex,SUM(TotalCharges) as TotalCharges,PERCENT_RANK()
OVER(ORDER BY SUM(TotalCharges)) As PercentRank FROM
(SELECT
      PrincipalDiagnosis, Sex,

ISNULL(ccu_charges,0)+ISNULL(ICU_Charge,0)+ISNULL(TheatreCharge,0)+ISNULL(Prosthesis
Charge,0)+ISNULL(OtherCharges,0)+ISNULL(BundledCharges,0) as TotalCharges
      FROM [RansayData].[dbo].[Data Insights - Synthetic Dataset]) AS TCharges
WHERE Sex='F'
GROUP BY PrincipalDiagnosis,Sex
--ORDER BY PercentRank,PrincipalDiagnosis,Sex
UNION
SELECT PrincipalDiagnosis,Sex,SUM(TotalCharges) as TotalCharges,PERCENT_RANK()
OVER(ORDER BY SUM(TotalCharges)) As PercentRank FROM
(SELECT
      PrincipalDiagnosis, Sex,

ISNULL(ccu_charges,0)+ISNULL(ICU_Charge,0)+ISNULL(TheatreCharge,0)+ISNULL(Prosthesis
Charge,0)+ISNULL(OtherCharges,0)+ISNULL(BundledCharges,0) as TotalCharges
      FROM [RansayData].[dbo].[Data Insights - Synthetic Dataset]) AS TCharges
WHERE Sex='M'
GROUP BY PrincipalDiagnosis,Sex
   --ORDER BY PercentRank,PrincipalDiagnosis,Sex
```

	PrincipalDiagnosis	Sex	TotalCharges	PercentRank
1	A02.1	F	0	0
2	A05.7	F	0	0
3	A08.0	F	0	0
4	A25.3	F	0	0
5	A28.8	F	0	0
6	A36.1	F	0	0
7	A37.7	F	0	0
8	A43.6	F	0	0
9	A52.1	F	0	0
10	A60.5	F	0	0
11	A86.3	F	0	0
12	B13.6	F	0	0
13	B22.6	F	0	0
14	B46.3	F	0	0
15	B68.8	F	0	0
16	B73.0	F	0	0
17	B79.1	F	0	0
18	B79.2	F	0	0
19	B89.6	F	0	0
20	B99.2	F	0	0