2025年5月22日 21:29

数据处理：
encounter_id：身份证 无缺失值

patient_nbr：标识患者的编号【数据处理中发现患者可能有再入院的情况--->patient_nbr出现多次 】 无缺失值

race:记录患者种族【Caucasian（白种人）、Asian（亚洲人）、African American（非裔美国人）、Hispanic（西班牙裔）和 other（其他） 一共5种】 缺失值有2273行 直接删了

分类的映射：
0: Caucasian
1: AfricanAmerican
2: Other
3: Asian
4: Hispanic

Gender:性别 没有缺失值 【分为三类 male, female, and unknown/invalid】
数据处理时发现，unknown/invalid这一类只有一行 所以直接删了

分类的映射：
0: Female
1: Male

age：Grouped in 10-year intervals: [0, 10), [10, 20),..., [90, 100)
没有缺失值

分类的映射：
0: [0-10)
1: [10-20)
2: [20-30)
3: [30-40)
4: [40-50)
5: [50-60)
6: [60-70)
7: [70-80)
8: [80-90)
9: [90-100)

weight：缺失值有96433 rows 太多了 直接删除这一列

admission_type_id：记录患者入院类型的整数型标识符
原分类数据：
1 Emergency 急诊
2 Urgent 紧急
3 Elective择期
4 Newborn新生儿
5 Not Available未知
6 NULL空值
7 Trauma Center创伤中心
8 Not Mapped未映射

---

5 Not Available未知
6 NULL空值
这两种合为一种 --> 空值
合并后的这个数据量大概一万条 还是比较大的 而且不知道实际情况没法填充 就先不删除了

8 Not Mapped未映射 一共317条 直接删除

分类的映射：
0: 5 Not Available未知
1: 1 Emergency 急诊
2: 2Urgent 紧急
3: 3Elective择期
4: 4Newborn新生儿
5: 7Trauma Center创伤中心

discharge_disposition_id：出院去向 29种 无缺失值

原数据及映射：

1 Discharged to home
2 Discharged/transferred to another short term hospital
3 Discharged/transferred to SNF
4 Discharged/transferred to ICF
5 Discharged/transferred to another type of inpatient care institution
6 Discharged/transferred to home with home health service
7 Left AMA
8 Discharged/transferred to home under care of Home IV provider
9 Admitted as an inpatient to this hospital
10 Neonate discharged to another hospital for neonatal aftercare
11 Expired
12 Still patient or expected to return for outpatient services
13 Hospice / home
14 Hospice / medical facility
15 Discharged/transferred within this institution to Medicare approved swing bed
16 Discharged/transferred/referred another institution for outpatient services
17 Discharged/transferred/referred to this institution for outpatient services
18 NULL
19 Expired at home. Medicaid only, hospice.
20 Expired in a medical facility. Medicaid only, hospice.
21 Expired, place unknown. Medicaid only, hospice.
22 Discharged/transferred to another rehab fac including rehab units of a hospital .
23 Discharged/transferred to a long term care hospital.
24 Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare.
25 Not Mapped
26 Unknown/Invalid
30 Discharged/transferred to another Type of Health Care Institution not Defined Elsewhere
27 Discharged/transferred to a federal health care facility.
28 Discharged/transferred/referred to a psychiatric hospital of psychiatric distinct part unit of a hospital
29 Discharged/transferred to a Critical Access Hospital (CAH).

上述内容主要是 出院后的事项/情况
主要进行的数据处理：
NULL（18）【3673条数据】、Not Mapped（25）【971条数据】、Unknown/Invalid（26）【没有数据】

21 26 29 这三类是没有数据的

我把
NULL（3673条数据）、Not Mapped（25）【971条数据】、Unknown/Invalid（26）【没有数据】数据都删除了

可以重点关注 Discharged to home【1】 这类患者治好了 Expired【11】这类患者死亡了

分类的映射：
0: 1 Discharged to home
1: 3 Discharged/transferred to SNF
2: 6 Discharged/transferred to home with home health service
3: 2 Discharged/transferred to another short term hospital
4: 5 Discharged/transferred to another type of inpatient care institution
5: 11 Expired
6: 7 Left AMA
7: 10 Neonate discharged to another hospital for neonatal aftercare
8: 4 Discharged/transferred to ICF
9: 14 Hospice / medical facility
10: 8 Discharged/transferred to home under care of Home IV provider
11: 13 Hospice / home
12: 12Still patient or expected to return for outpatient services
13: 16Discharged/transferred/referred another institution for outpatient services
14: 17Discharged/transferred/referred to this institution for outpatient services
15: 22 Discharged/transferred to another rehab fac including rehab units of a hospital .
16: 23Discharged/transferred to a long term care hospital.
17: 9 Admitted as an inpatient to this hospital
18: 20 Expired in a medical facility. Medicaid only, hospice.
19: 15 Discharged/transferred within this institution to Medicare approved swing bed
20: 24 Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare.

21: 28 Discharged/transferred/referred to a psychiatric hospital of psychiatric distinct part unit of a hospital
22: 19 Expired at home. Medicaid only, hospice.
23: 27 Discharged/transferred to a federal health care facility.

admission_source_id：入院来源 无缺失值

原数据：

1 Physician Referral
2 Clinic Referral
3 HMO Referral
4 Transfer from a hospital
5 Transfer from a Skilled Nursing Facility (SNF)
6 Transfer from another health care facility
7 Emergency Room
8 Court/Law Enforcement
9 Not Available
10 Transfer from critial access hospital
11 Normal Delivery
12 Premature Delivery
13 Sick Baby
14 Extramural Birth
15 Not Available
17 NULL
18 Transfer From Another Home Health Agency
19 Readmission to Same Home Health Agency
20 Not Mapped
21 Unknown/Invalid
22 Transfer from hospital inpt/same fac reslt in a sep claim
23 Born inside this hospital
24 Born outside this hospital
25 Transfer from Ambulatory Surgery Center
26 Transfer from Hospice

其中，7，2，4，1，6, 20，5，3, 17，8，9, 14, 10, 22, 11, 25, 13
只有这几类有数据 ，其他分类无数据

由于
9 Not Available
17 NULL
20 Not Mapped
是缺失值，而且Not Available 和 Not Mapped 的数据量非常小 也就100-200左右
所以直接删除了

17 NULL 有6300条 占一个分类位置吧

分类的映射：
0: 7 Emergency Room
1: 2 Clinic Referral
2: 4 Transfer from a hospital
3: 1 Physician Referral
4: 6 Transfer from another health care facility
5: 5 Transfer from a Skilled Nursing Facility (SNF)
6: 3 HMO Referral
7: 17 NULL
8: 8 Court/Law Enforcement
9: 14 Extramural Birth
10: 10 Transfer from critial access hospital
11: 22 Transfer from hospital inpt/same fac reslt in a sep claim
12: 11 Normal Delivery
13: 25 Transfer from Ambulatory Surgery Center
14: 13 Sick Baby

time_in_hospital：在医院呆的时间 无缺失值
1-14天

payer_code：支付码 有缺失值

34613 rows 缺失值 直接把这一列删掉（不太有实际意义）

medical_specialty：患者治病的科室/专业
这个缺失值也非常大，我把他归为一类了，没有删除(如果这类 对模型影响不大的话 不建议使用)

分类的映射：
0: NULL
1: InternalMedicine
2: Family/GeneralPractice
3: Cardiology
4: Surgery-General
5: Orthopedics
6: Gastroenterology
7: Nephrology
8: Psychiatry
9: Orthopedics-Reconstructive
10: Pulmonology
11: Surgery-Neuro
12: Obsterics&Gynecology-GynecologicOnco
13: Pediatrics-CriticalCare
14: Endocrinology
15: Urology
16: Radiology
17: Pediatrics-Endocrinology
18: ObstetricsandGynecology
19: Pediatrics
20: Pediatrics-Hematology-Oncology
21: Surgery-Cardiovascular/Thoracic
22: Anesthesiology-Pediatric
23: Emergency/Trauma
24: Psychology
25: Neurology
26: Hematology/Oncology
27: Psychiatry-Child/Adolescent
28: Surgery-Colon&Rectal
29: Podiatry
30: Pediatrics-Pulmonology
31: Gynecology
32: Oncology
33: Pediatrics-Neurology
34: Surgery-Plastic
35: Surgery-Thoracic
36: Surgery-PlasticwithinHeadandNeck
37: Ophthalmology
38: Surgery-Pediatric
39: Pediatrics-EmergencyMedicine
40: PhysicalMedicineandRehabilitation
41: Otolaryngology
42: InfectiousDiseases
43: Anesthesiology
44: AllergyandImmunology
45: Surgery-Maxillofacial
46: Pediatrics-InfectiousDiseases
47: Pediatrics-AllergyandImmunology
48: Dentistry
49: Surgeon
50: Surgery-Vascular
51: Osteopath
52: Psychiatry-Addictive
53: Surgery-Cardiovascular
54: PhysicianNotFound
55: Hematology
56: Proctology
57: Rheumatology
58: Obstetrics

59: SurgicalSpecialty
60: Radiologist
61: Pathology
62: Dermatology
63: SportsMedicine
64: Speech
65: Hospitalist
66: OutreachServices
67: Cardiology-Pediatric
68: Perinatology
69: Neurophysiology
70: Endocrinology-Metabolism
71: DCPTEAM
72: Resident

num_lab_procedures: 记录 患者在一次就医过程中接受的实验室检查项目数量

这是检查项目数量最多的前五项如下：
项目数量 患者数量
1 3047位
43 2600位
44 2293位
45 2181位
38 2065位
【数值越高，通常意味着诊疗过程越复杂或病情需要更多的检查支持】

没啥可处理的 无缺失值

num_procedures ：记录 患者在一次就医过程中接受的非实验室检查类操作数量
数量分类比较少：
0 43427
1 19304
2 11795
3 8653
6 4513
4 3844
5 2781
最多也就5次检查数量

这里的num_lab_procedures（实验室）和num_procedures（非实验室）没搞明白区别是啥

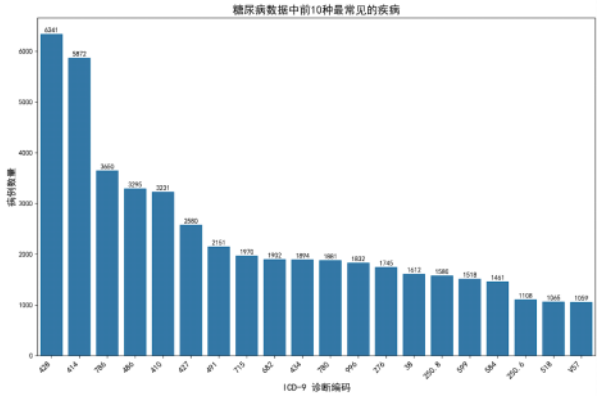
num_medications：住院或诊疗期间使用的不同通用名药物的数量 无缺失值
数量大概是1-81的范围内

number_outpatient：患者在就诊**前一年**的门诊就诊次数 无缺失值
数据中，0次的占比非常大，大概是8万条数据，说明很多患者是初次就诊
最大的门诊就诊次数前一年就诊了42次

number_emergency：患者在就诊前一年的紧急就诊次数 无缺失值
数据中，0次的占比非常大，大概83198条数据，说明很多患者是初次就诊
最大的紧急就诊次数前一年就诊了76次

number_inpatient：患者在遭遇前一年的住院就诊次数 无缺失值
数据中，0次的占比非常大，大概62264条数据，说明很多患者是初次就诊
最大的住院就诊次数前一年就诊了21次

以下三个变量 处理方式相同：
特征值：仅取ICD-9 编码体系的 前三位数字:
diag_1（主要诊断）：反映患者本次就诊的直接原因



根据上图，可以把前几种数量大的病种的编码提取出来，查找它的病名， 单独分析

diag_2（次要诊断 1）：记录与主要诊断并存的其他疾病或状况【可以理解为并发症】
diag_3（次要诊断 2）：进一步补充患者的其他健康问题
三个变量都有缺失值，且缺失值较少，所以直接删除了

三个变量都是分类变量，变为分类变量了（见csv表格）

这里的ICD-9 编码体系前三位数字可以根据
https://www.findacode.com/code-set.php?set=ICD9
这个网址查看具体的基本类型（不建议搜索，建议自己一点一点根据分类往下找，因为搜索的结果不准确）
（比如：250是糖尿病）

number_diagnoses：录入系统的诊断数量　无缺失值

虽然不是分类变量，但是诊断数量只有如下14种：

9	47034
8	9851
5	9738
7	9559
6	9288
4	4885
3	2506
16	40
13	16
10	16
11	11
12	9
15	8
14	7

max_glu_serum：葡萄糖血清检测结果　无缺失值
分类变量：“>200”【糖尿病】、“>300”【严重高血糖】、“正常”和“无”（如果未测量）

一共四个分类：

Norm	2369
>200	1330
>300	1162

剩下未测量的都是“未测量None”（数量巨大，无法删除）

分类的映射：
0: None
1: >300
2: Norm
3: >200

A1Cresult：A1c 检测结果【（糖化血红蛋白检测）】
分类变量：“>8”表示结果大于 8%【血糖控制 较差】，“>7”表示结果大于 7% 但小于 8%【 血糖控制中等偏下】，“正常”表示结果小于 7%，“无”表示未测量。

缺失值也比较大，无法删除

分类的映射:
0: None
1: >7
2: >8
3: Norm

以下是药物变量（这块 网站上有解释错误的地方 以下内容已纠正）：

一共24个变量：metformin repaglinide nateglinide chlorpropamideglimepiride acetohexamide glipizide glyburide tolbutamide pioglitazone rosiglitazone acarbose miglitol troglitazone tolazamide examide sitagliptin insulin glyburide-metformin glipizide-metformin glimepiride-pioglitazone metformin-rosiglitazone metformin-pioglitazone
这24中药物均是治疗糖尿病的药物，都有四类别：
No：没有吃这个药
Steady：吃了，但是剂量一直不变
Up ：吃了，但是剂量增加了
Down：吃了，但是剂量减少了
这些类别用于描述其处方状态或剂量调整情况（吃没吃/吃了多少）

metformin：（二甲双胍）是糖尿病治疗的一线药物 【无缺失值】

分类的映射:
0: No 74729 未处方二甲双胍**(这个类别占大部分-->74729 的患者)**
1: Steady 16783 诊疗期间剂量未改变（维持原方案）
2: Up 938 诊疗期间增加二甲双胍剂量
3: Down 518 诊疗期间减少二甲双胍剂量

repaglinide：(瑞格列奈) 【无缺失值】

分类的映射:
0: No 91480 未处方瑞格列奈**（这个类别占大部分-->91480条数据）**
1: Steady 1339 诊疗期间剂量未改变（维持原方案）
2: Up 107 诊疗期间增加瑞格列奈剂量
3: Down 42 诊疗期间减少瑞格列奈剂量

nateglinide：（那格列奈）【无缺失值】

分类的映射:
0：No 92281
1：Steady 653
2：Up 23
3：Down 11

Chlorpropamide：（氯磺丙脲） 【无缺失值】

分类的映射:
0：No 92907
1：Steady 56
2：Up 4
3：Down 1

glimepiride：（格列美脲） 【无缺失值】

分类的映射
0：No 88129
1：Steady 4373
2：Up 287
3：Down 179

acetohexamide：（乙酰己酰胺） 【无缺失值】
分类的映射
0：No 92967
1：Steady 1
其他分类没有
有效值太少，删掉这一列

glipizide：（格列吡嗪） 【无缺失值】

分类的映射
0：No 81189
1：Steady 10565
2：Up 704
3：Down 510

glyburide： （格列本脲） 【无缺失值】

分类的映射
0：No 83700
1：Steady 8089
2：Up 704
3：Down 475

tolbutamide：（甲苯磺丁脲）【无缺失值】
分类的映射
0：No 92946
1：Steady 22
其他分类没有
有效值太少，删掉这一列

pioglitazone：（吡格列酮） 【无缺失值】

分类的映射
0：No 86124
1：Steady 6524
2：Up 216
3：Down 104

rosiglitazone：（罗格列酮）【无缺失值】

分类的映射
0：No 87108
1：Steady 5621
2：Up 163
3：Down 76

acarbose：（阿卡波糖）【无缺失值】
分类的映射

0：No 92692
1：Steady 265
2：Up 8
3：Down 3

Miglitol:（米格列醇）【无缺失值】
分类的映射
0：No 92933
1：Steady 29
2：Up 2
3：Down 4
有效值太少，删掉这一列

troglitazone：（曲格列酮)【无缺失值】
分类的映射
0：No 92966
1：Steady 2
其他分类没有
有效值太少，删掉这一列

Tolazamide：（托拉他胺）【无缺失值】
分类的映射

0：No　　92941
1：Steady　　27
其他分类没有
有效值太少，删掉这一列

examide：【无缺失值】
分类的映射
0：No　92968
其他分类没有
有效值太少，删掉这一列

citoglipton：（西格列汀）【无缺失值】
Citoglipton：经查阅 应该是写错了　我看论文中的数据--> 应该是sitagliptin【西他列汀】：
所以我把这一列的特征名换了

分类的映射：
0：No　92968
其他分类没有
有效值太少，删掉这一列

insulin：（胰岛素）【无缺失值】---» 这个变量还是挺有意义的 它的数据差距比起其他药物不是很大

```
分类的映射：
0：No　　42484
1：Steady　　28314
2：Up　　10649
3：Down　　11521
```

glyburide-metformin：（格列本脲-二甲双胍）【无缺失值】

```
分类的映射：
0：No　　92305
1：Steady　　655
2：Up　　5
3：Down　　3
```

glipizide-metformin：（格列吡嗪-二甲双胍）【无缺失值】
分类的映射：
0：No　　92955
1：Steady　　13
有效值太少，删掉这一列

glimepiride-pioglitazone：（格列美脲-吡格列酮）【无缺失值】
分类的映射：
0：No　　92967
1：Steady　　1
有效值太少，删掉这一列

metformin-rosiglitazone：（二甲双胍-罗格列酮）【无缺失值】
分类的映射：
0：No　92968
有效值太少，删掉这一列

metformin-pioglitazone：（二甲双胍-吡格列酮）【无缺失值】
分类的映射：
0：No　　92967
1：Steady　　1
有效值太少，删掉这一列

change：指示糖尿病药物（剂量或通用名称）是否有变化。分类：change 和 no change

```
0：Ch    43543
1：No    49425
```

【无缺失值】

diabetesMed：是否开了任何糖尿病药物。分类：yes 和 no 【无缺失值】

```
0：Yes   71931
1：No    21037
```

readmitted：住院再入院天数。【无缺失值】

分类: 如果患者在 30 天内再次入院，则为 <30，如果患者在 30 天外再次入院，则为 >30，如果患者没有再次入院记录，则为否，表示没有再入院记录。

```
0：NO     49404
1：>30    33101
2：<30    10463
```