Pillar 5: Capstone Project_30-Day Hospital Readmission Prediction

Problem Background

> *"The Centers for Medicare & Medicaid Services (CMS) of the U.S. government pays hospital for Medicare patients (low-income government-insured patients). Thus, they established to define hospital quality metrics and directly ties their payment to the hospital performance."*

> *"One of their quality metrics is Hospital Readmission within 30 days which is a proxy for quality care and a sign of poor discharge planning, care coordination, or follow-up. Because of this, CMS treats avoidable readmissions as costly failures and penalizes hospitals by losing a percentage of total Medicare inpatient payments." "Since safety-net hospitals that serve sicker or older population would be unfairly punished, CMS risk-adjustment exists for age, sex, comorbidities, prior diagnoses, clinical severity, and past utilization."*

> *"Readmission reduction indirectly helps hospitals by keeping bonuses, avoid downstream losses, and improve insurer negotiations."*

Step 1: Problem Understanding & Framing

> *"This project mirrors CMS risk-adjusted readmission modeling* used in U.S. hospital reimbursement, where inaccurate prediction can lead to multi-million-dollar penalties."**

Risk-adjusted Benchmarks Table:

Aspect	CMS Approach
Metric	30-day unplanned readmission
Model type	Risk-adjusted logistic regression
Adjustment factors	Age, comorbidities, clinical history
Financial impact	Up to ~3% Medicate payment reduction
Incentive structure	Penalty-focused

Step 2: Data Collection & Understanding

> *"The data set represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery network."*

> *"It includes over 50 features representing patient and hospital outcomes where different situations could be possibly identified if it will result to <30-day readmission or not."*

Dataset Summary:

Attribute	**Description**
Number of patients	~100,000 hospital encounters
Number of features	50 raw columns
Target variable	30-day readmission status
Class Imbalance	Yes (~11% readmitted within 30 days)

Feature Categories:

Demographic Features	Admission & Discharge Information	Clinical & Diagnosis Features	Medication & Treatment Indicators	Utilization History
`age`	`admission_type`	`num_lab_procedures`	`insulin`	`number_inpatient`
`gender`	`admission_source`	`num_medications`	`metformin`	`number_emergency`
`race`	`discharge_disposition`	`number_diagnoses`	`change`	`number_outpatient`
`payer_code`	`time_in_hospital`	`diag_1, diag_2, diag_3`	`diabetesMed`	

Data Dictionary:

Variable Name	Type	Description	Allowed Values
`encounter_id`	integer	unique identifier for each hospital encounter ...	12522-443867222
`patient_nbr`	integer	unique identifier for each patient	135-189502619
`race`	object	patient's race	AfricanAmerican, Asian, ..
`gender`	object	patient's gender	Female, Male, Unknown
`age`	object	age group of the patient, reported in intervals	[0-10), [10-20), ..

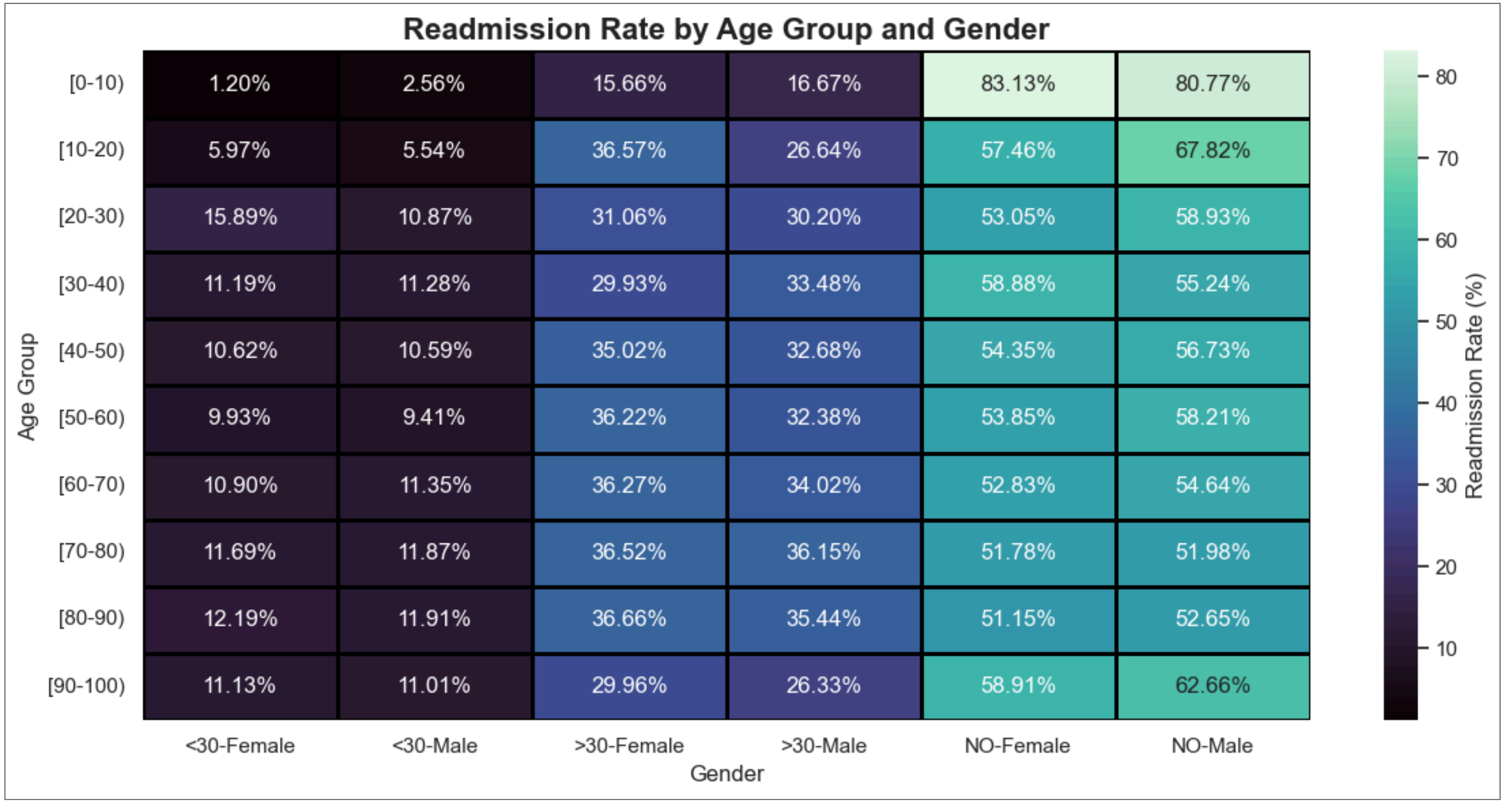
Source: UCI Machine Learning Repository. (n.d.). **https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008**

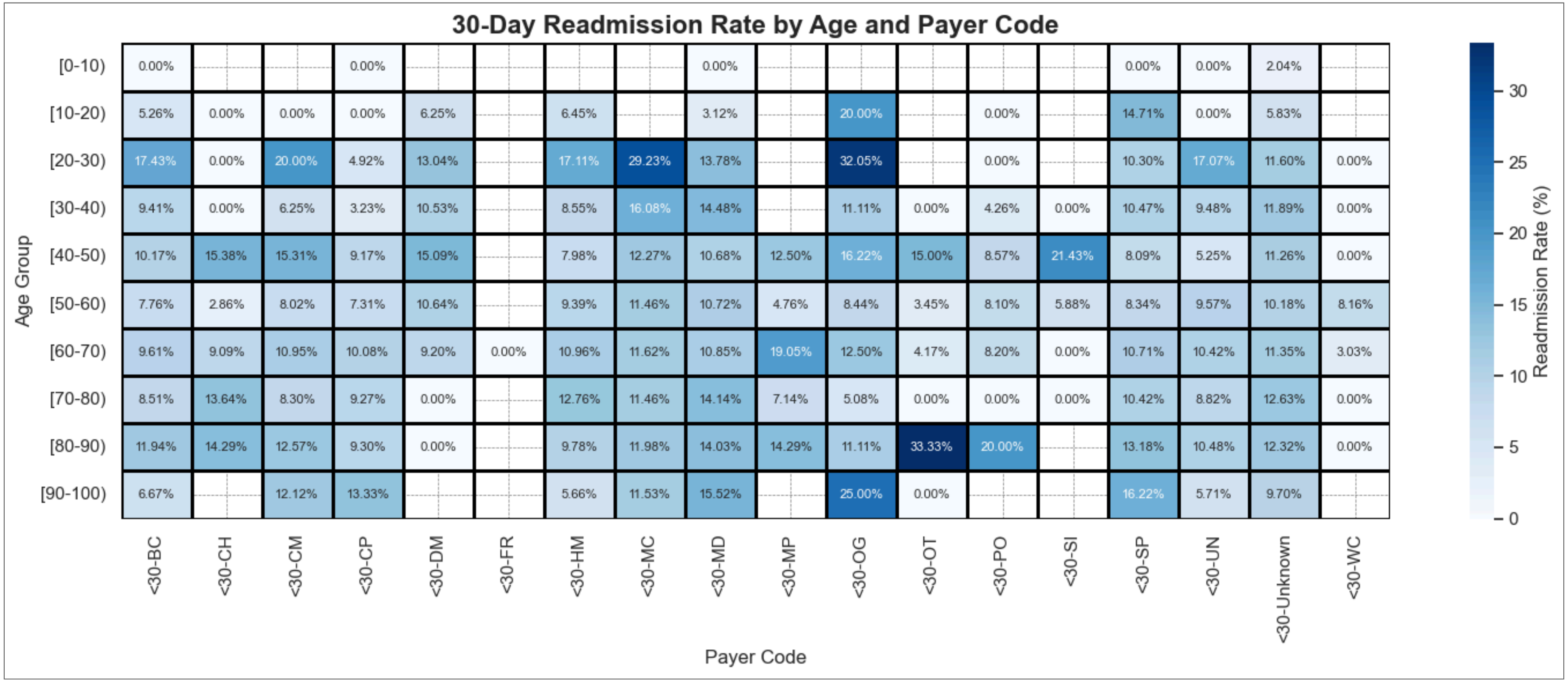
3. Exploratory Data Analysis (EDA)

In alignment with CMS-style readmission metrics, this EDA explored to understand the structure, quality, and clinical context of the dataset.

Demographic Features EDA Questions: *To understand baseline population heterogeneity*

- What percentage of encounters result in 30-day readmission?
- Are there patients with multiple encounters contributing multiple rows?
- Are there differences by gender?
- How does readmission rate vary by age group?
- Are there differences by race/ethnicity?
- Do payer categories show different readmission rates?


Readmission Rate by Age Group and Gender
Age Group
[0-10)
[10-20)
[20-30)
[30-40)
[40-50)
[50-60)
[60-70)
[70-80)
[80-90)
[90-100)
Gender
<30-Female
<30-Male
>30-Female
>30-Male
NO-Female
NO-Male
1.20%
2.56%
15.66%
16.67%
83.13%
80.77%
5.97%
5.54%
36.57%
26.64%
57.46%
67.82%
15.89%
10.87%
31.06%
30.20%
53.05%
58.93%
11.19%
11.28%
29.93%
33.48%
58.88%
55.24%
10.62%
10.59%
35.02%
32.68%
54.35%
56.73%
9.93%
9.41%
36.22%
32.38%
53.85%
58.21%
10.90%
11.35%
36.27%
34.02%
52.83%
54.64%
11.69%
11.87%
36.52%
36.15%
51.78%
51.98%
12.19%
11.91%
36.66%
35.44%
51.15%
52.65%
11.13%
11.01%
29.96%
26.33%
58.91%
62.66%
Readmission Rate (%)
80
70
60
50
40
30
20
10


30-Day Readmission Rate by Age and Payer Code
Age Group
[0-10) 0.00% 0.00% 0.00% 0.00% 0.00% 0.00% 2.04%
[10-20) 5.26% 0.00% 0.00% 0.00% 6.25% 6.45% 3.12% 20.00% 0.00% 14.71% 0.00% 5.83%
[20-30) 17.43% 0.00% 20.00% 4.92% 13.04% 17.11% 29.23% 13.78% 32.05% 0.00% 10.30% 17.07% 11.60% 0.00%
[30-40) 9.41% 0.00% 6.25% 3.23% 10.53% 8.55% 16.08% 14.48% 11.11% 0.00% 4.26% 0.00% 10.47% 9.48% 11.89% 0.00%
[40-50) 10.17% 15.38% 15.31% 9.17% 15.09% 7.98% 12.27% 10.68% 12.50% 16.22% 15.00% 8.57% 21.43% 8.09% 5.25% 11.26% 0.00%
[50-60) 7.76% 2.86% 8.02% 7.31% 10.64% 9.39% 11.46% 10.72% 4.76% 8.44% 3.45% 8.10% 5.88% 8.34% 9.57% 10.18% 8.16%
[60-70) 9.61% 9.09% 10.95% 10.08% 9.20% 0.00% 10.96% 11.62% 10.85% 19.05% 12.50% 4.17% 8.20% 0.00% 10.71% 10.42% 11.35% 3.03%
[70-80) 8.51% 13.64% 8.30% 9.27% 0.00% 12.76% 11.46% 14.14% 7.14% 5.08% 0.00% 0.00% 0.00% 10.42% 8.82% 12.63% 0.00%
[80-90) 11.94% 14.29% 12.57% 9.30% 0.00% 9.78% 11.98% 14.03% 14.29% 11.11% 33.33% 20.00% 13.18% 10.48% 12.32% 0.00%
[90-100) 6.67% 12.12% 13.33% 5.66% 11.53% 15.52% 25.00% 0.00% 16.22% 5.71% 9.70%
<30-BC
<30-CH
<30-CM
<30-CP
<30-DM
<30-FR
<30-HM
<30-MC
<30-MD
<30-MP
<30-OG
<30-OT
<30-PO
<30-SI
<30-SP
<30-UN
<30-Unknown
<30-WC
Payer Code
30
25
20
15
10
5
0
Readmission Rate (%)

Demographic Insights:

Total population

- **23.45%** of the total number of patients have multiple encounters.

Gender Comparison

- Female and Male genders have almost the same <30-day hospital readmission rate (~11%)

Age Comparison

- Sudden increase in <30-day hospitalreadmission rate from ages 10-20 (5.79%) to ages 20-30 (14.24%).
- Gradual decrease in <30-day readmission rate from ages 20-30 (14.24%) to ages 50-60 (9.66%).
- Ages 60 and above plateau to a <30-day readmission rate of ~11% with slight increase to 12% in ages 80-90.

Age & Gender Comparison

- Separating age group into gender shows that the increase in ages 20-30 and 80-90 is largely due to Female increase hospital readmission.
    - 20-30: 15.89% for Female and 10.87% for Male
    - 80-90: 12.19% for Female and 11.91% for Male

Age-Race Comparison

- All races have data record from all age groups that only 1 out of 4 races have <30-day hospital readmission per age group (Caucasian).
- Though Hispanics at the age of 90-100 has the highest <30-day readmission rate of 21.05%, it does not significantly increase the overall age trend.

Payer Code

- The top 5 highest <30-day hospital readmission rate by payer code (OG, SI, MD, MC, DM) are all under Government Insurance that represents the socioeconomically vulnerable populations like the elderly, the disabled and the low-income patients.

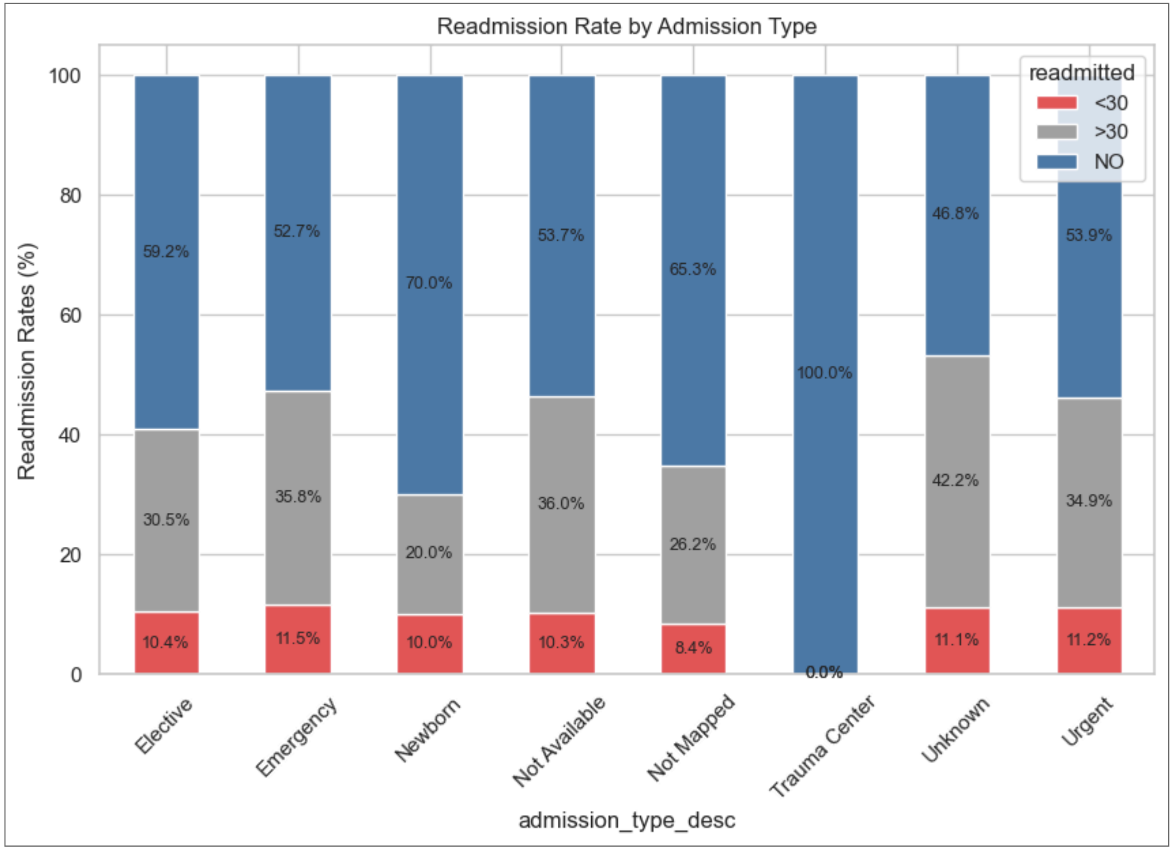
Age & Payer Code Comparison

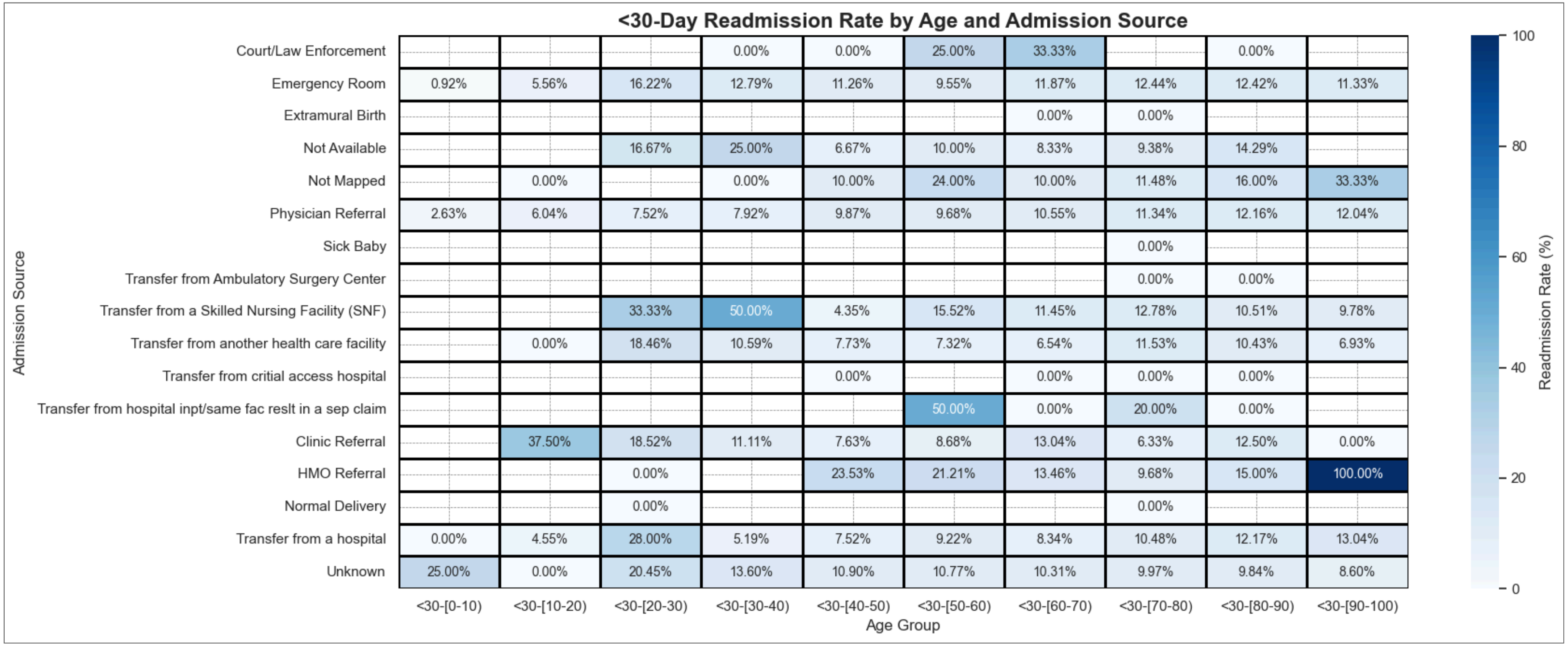
- 9 out of 17 payer code have a sudden increase in ages 20-30.
- It can also be noted that MC (gov't insurance for low-income patients) only started to have a data record by the age 20-30

Admission and Discharge Features EDA Questions:

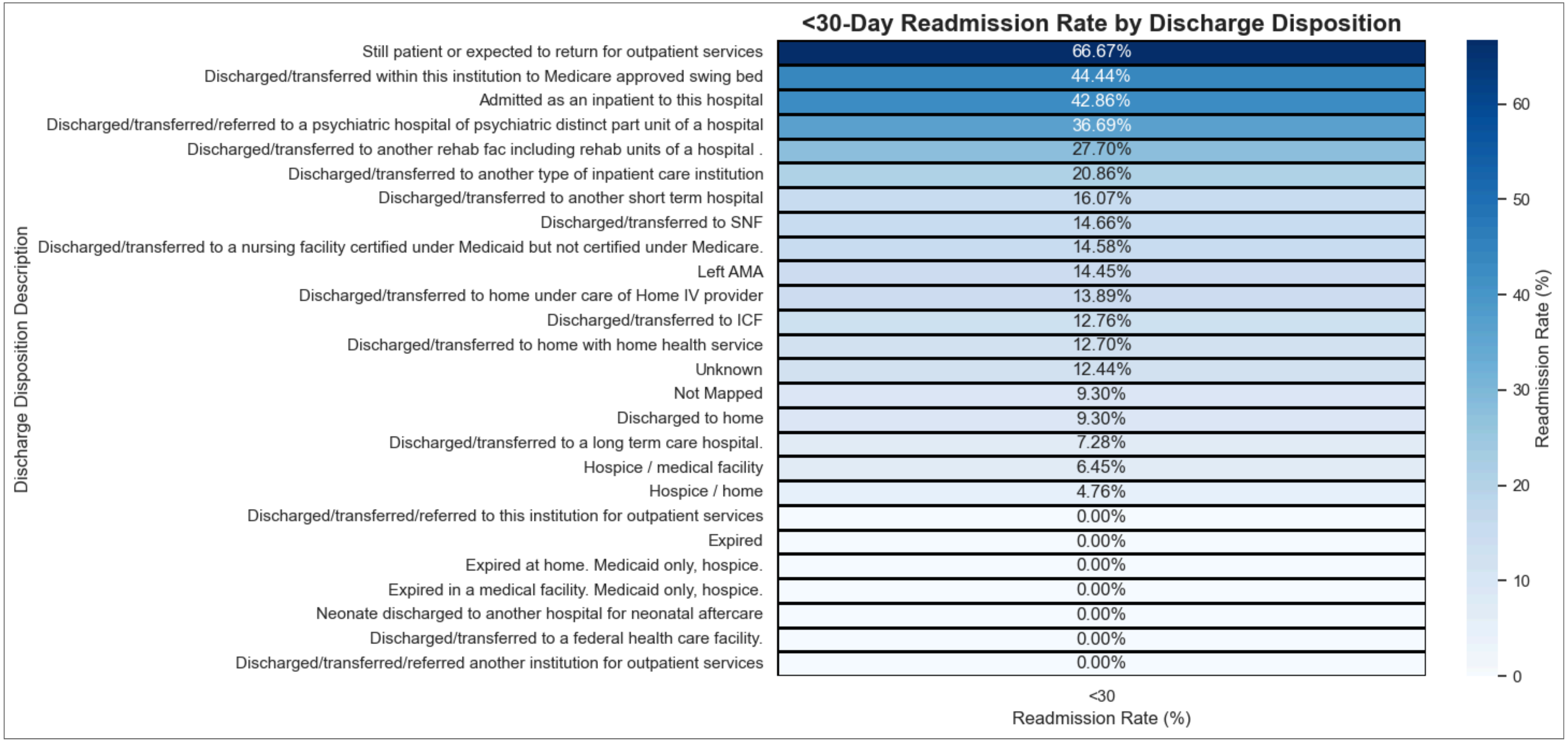
To evaluate care pathways among readmitted cases without assuming causality

- How does readmission vary by admission type (emergency vs elective)?
- Does admission source affect readmission?
- How does discharge disposition relate to readmission?




<30-Day Readmission Rate by Age and Admission Source

Admission Source	<30-[0-10)	<30-[10-20)	<30-[20-30)	<30-[30-40)	<30-[40-50)	<30-[50-60)	<30-[60-70)	<30-[70-80)	<30-[80-90)	<30-[90-100)
Court/Law Enforcement				0.00%	0.00%	25.00%	33.33%		0.00%	
Emergency Room	0.92%	5.56%	16.22%	12.79%	11.26%	9.55%	11.87%	12.44%	12.42%	11.33%
Extramural Birth							0.00%	0.00%		
Not Available			16.67%	25.00%	6.67%	10.00%	8.33%	9.38%	14.29%	
Not Mapped		0.00%		0.00%	10.00%	24.00%	10.00%	11.48%	16.00%	33.33%
Physician Referral	2.63%	6.04%	7.52%	7.92%	9.87%	9.68%	10.55%	11.34%	12.16%	12.04%
Sick Baby								0.00%		
Transfer from Ambulatory Surgery Center								0.00%	0.00%	
Transfer from a Skilled Nursing Facility (SNF)			33.33%	50.00%	4.35%	15.52%	11.45%	12.78%	10.51%	9.78%
Transfer from another health care facility		0.00%	18.46%	10.59%	7.73%	7.32%	6.54%	11.53%	10.43%	6.93%
Transfer from critial access hospital					0.00%		0.00%	0.00%	0.00%	
Transfer from hospital inpt/same fac reslt in a sep claim						50.00%	0.00%	20.00%	0.00%	
Clinic Referral		37.50%	18.52%	11.11%	7.63%	8.68%	13.04%	6.33%	12.50%	0.00%
HMO Referral			0.00%		23.53%	21.21%	13.46%	9.68%	15.00%	100.00%
Normal Delivery			0.00%					0.00%		
Transfer from a hospital	0.00%	4.55%	28.00%	5.19%	7.52%	9.22%	8.34%	10.48%	12.17%	13.04%
Unknown	25.00%	0.00%	20.45%	13.60%	10.90%	10.77%	10.31%	9.97%	9.84%	8.60%


<30-Day Readmission Rate by Discharge Disposition
Discharge Disposition Description
Still patient or expected to return for outpatient services 66.67%
Discharged/transferred within this institution to Medicare approved swing bed 44.44%
Admitted as an inpatient to this hospital 42.86%
Discharged/transferred/referred to a psychiatric hospital of psychiatric distinct part unit of a hospital 36.69%
Discharged/transferred to another rehab fac including rehab units of a hospital . 27.70%
Discharged/transferred to another type of inpatient care institution 20.86%
Discharged/transferred to another short term hospital 16.07%
Discharged/transferred to SNF 14.66%
Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare. 14.58%
Left AMA 14.45%
Discharged/transferred to home under care of Home IV provider 13.89%
Discharged/transferred to ICF 12.76%
Discharged/transferred to home with home health service 12.70%
Unknown 12.44%
Not Mapped 9.30%
Discharged to home 9.30%
Discharged/transferred to a long term care hospital. 7.28%
Hospice / medical facility 6.45%
Hospice / home 4.76%
Discharged/transferred/referred to this institution for outpatient services 0.00%
Expired 0.00%
Expired at home. Medicaid only, hospice. 0.00%
Expired in a medical facility. Medicaid only, hospice. 0.00%
Neonate discharged to another hospital for neonatal aftercare 0.00%
Discharged/transferred to a federal health care facility. 0.00%
Discharged/transferred/referred another institution for outpatient services 0.00%
<30
Readmission Rate (%)
Readmission Rate (%)
0
10
20
30
40
50
60

Admission and Discharge Insights

Admission Type

- Highest <30-day hospital readmission rate for **Emergency** patients with 11.5% closely followed by **Urgent** patients with 11.2%
- **Elective** admissions is at 10.4% readmission rate which I think is considered still close to **Emergency** rate.
- This indicates that planned procedures does not generally mean low admission risk. Severity must still be a major factor.

Admission Source

- Highest <30-day hospital readmission rate for **Transfer from hospital inpt/facility** patients with 16.67% closely followed by **HMO Referral** patients with 15.51%

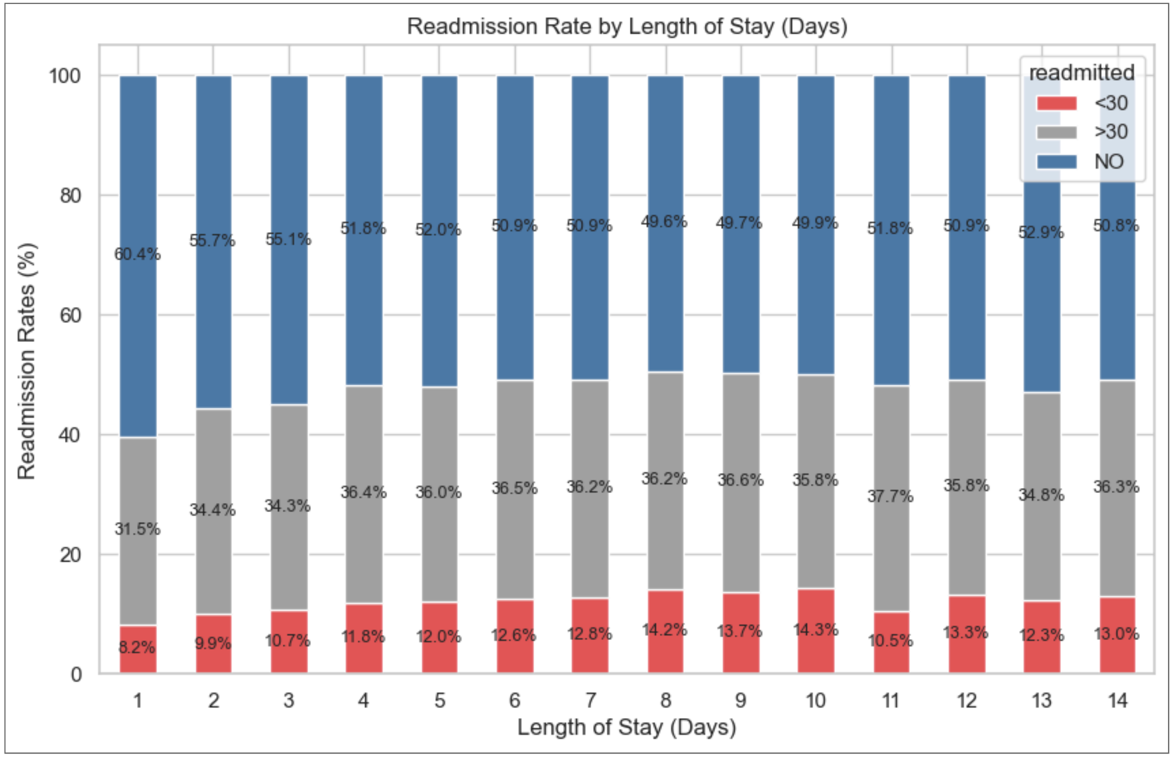
Discharge Disposition

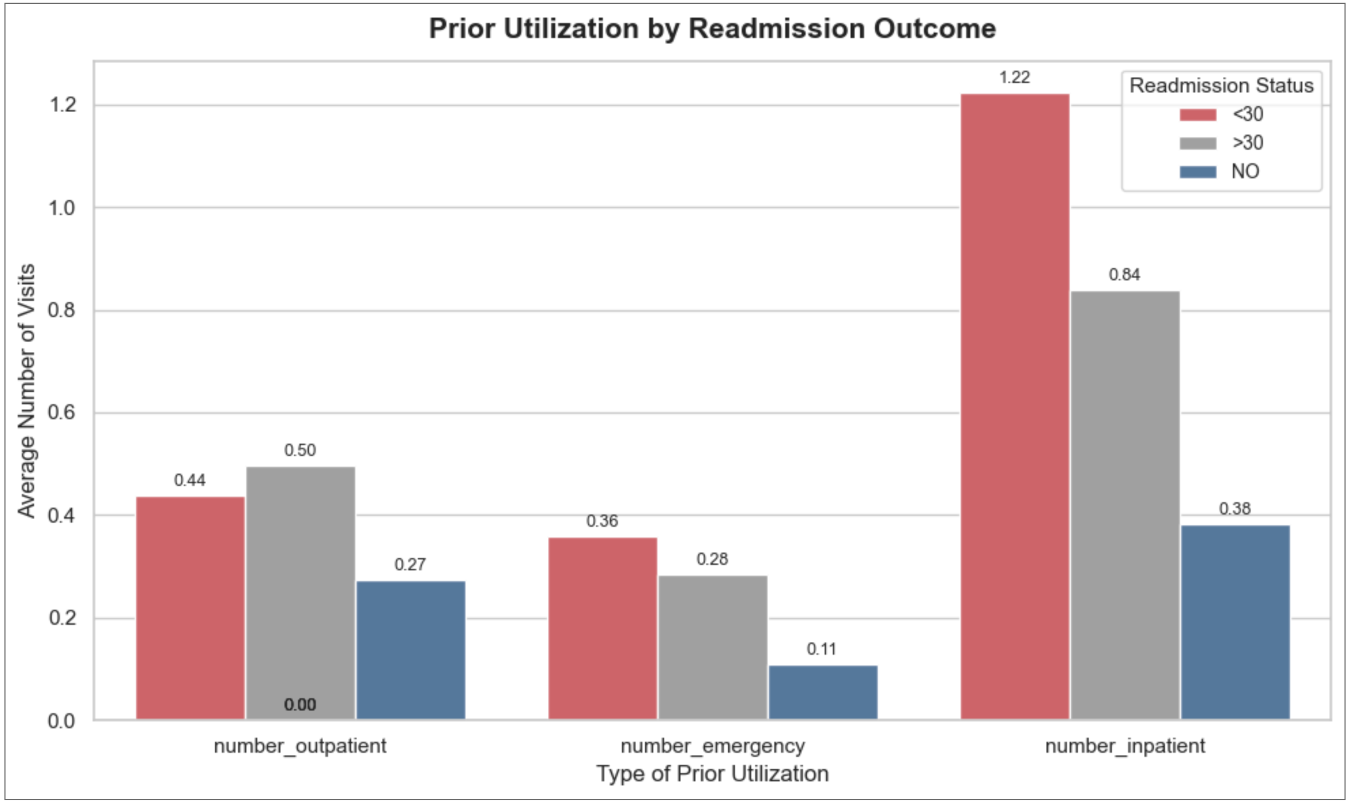
- Highest <30-day hospital readmission rate for **Still patient or expected to return for outpatient services** with 66.67%

Utilization & Severity Features EDA Questions:

To assess operational and healthcare in the year prior to admission

- How does length of stay relate with readmission?
- How does prior utilization relate with readmission?


Readmission Rate by Length of Stay (Days)
Readmission Rates (%)
Length of Stay (Days)
0
20
40
60
80
100
1
2
3
4
5
6
7
8
9
10
11
12
13
14
readmitted
<30
>30
NO
60.4%
55.7%
55.1%
51.8%
52.0%
50.9%
50.9%
49.6%
49.7%
49.9%
51.8%
50.9%
52.9%
50.8%
31.5%
34.4%
34.3%
36.4%
36.0%
36.5%
36.2%
36.2%
36.6%
35.8%
37.7%
35.8%
34.8%
36.3%
8.2%
9.9%
10.7%
11.8%
12.0%
12.6%
12.8%
14.2%
13.7%
14.3%
10.5%
13.3%
12.3%
13.0%


Prior Utilization by Readmission Outcome
Average Number of Visits
Type of Prior Utilization
number_outpatient
number_emergency
number_inpatient
Readmission Status
<30
>30
NO
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.44
0.50
0.27
0.00
0.36
0.28
0.11
1.22
0.84
0.38

Utilization & Severity Insights:

Length of Stay

- Gradual <30-day readmission rate increase as length of stay also increases starting from 1 day with 8.2% up to 7days with 12.8% readmission.
- Sudden increase of <30-day readmission rate after 7 days ranging from 13-14%.

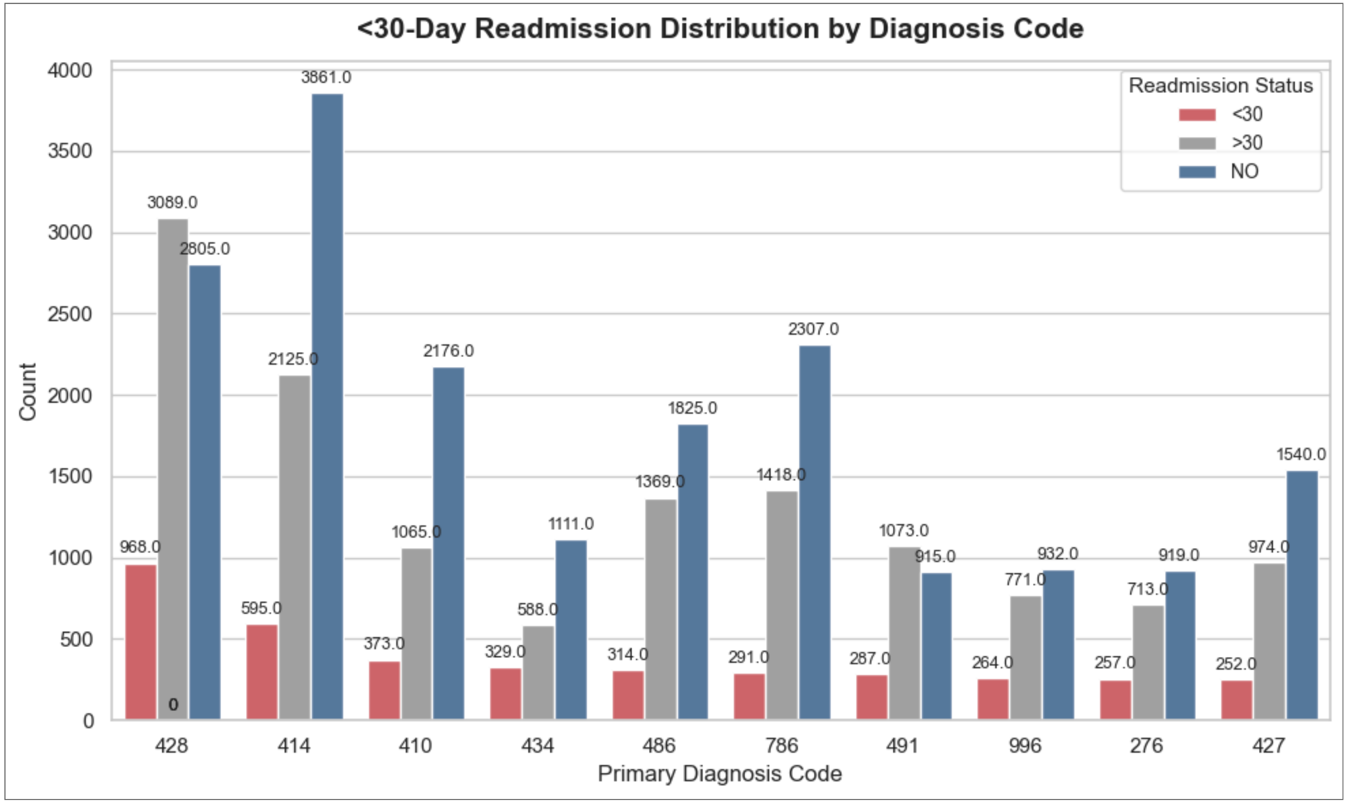
Prior Utilization

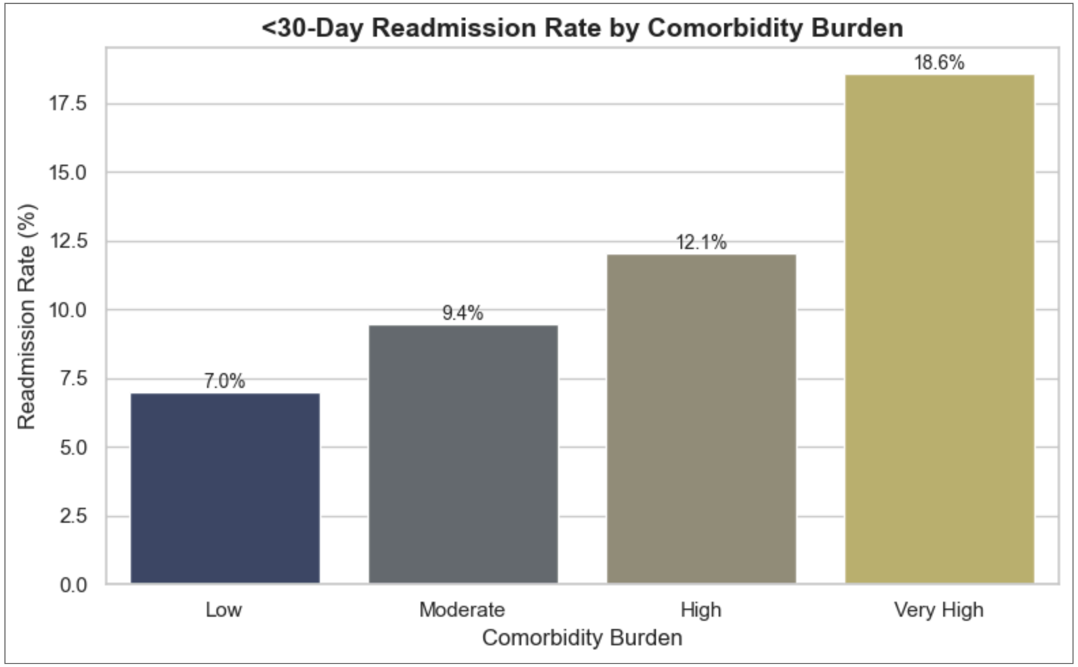
- Highest <30-day hospital readmission average number of visits for `num_inpatients` with 1.22 average followed by `num_outpatients` with 0.44 average.

Diagnoses and Comorbidity Features EDA Questions:

To assess diagnosis data and highlight the need for clinically meaningful abstraction

- Which diagnosis codes are most frequent?
- Are certain diagnosis groups overrepresented among readmitted cases?
- Do chronic conditions cluster together?
- Does comorbidity count increase readmission risk?


<30-Day Readmission Distribution by Diagnosis Code
Readmission Status
<30
>30
NO
Count
4000
3500
3000
2500
2000
1500
1000
500
0
968.0
3089.0
2805.0
0
595.0
2125.0
3861.0
373.0
1065.0
2176.0
329.0
588.0
1111.0
314.0
1369.0
1825.0
291.0
1418.0
2307.0
287.0
1073.0
915.0
264.0
771.0
932.0
257.0
713.0
919.0
252.0
974.0
1540.0
428
414
410
434
486
786
491
996
276
427
Primary Diagnosis Code


<30-Day Readmission Rate by Comorbidity Burden
Readmission Rate (%)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
7.0%
9.4%
12.1%
18.6%
Low
Moderate
High
Very High
Comorbidity Burden

Diagnosis and Comorbidity Insights:

Diagnosis Code

- Highest frequency of diagnosis for 428 representing heart failure followed by diabetes.
- In terms of <30-day hospital readmission count per diagnosis code, **heart failure** is still the highest with 968 counts followed by **other ischemic heart disease** with 595 counts.

Chronic Condition Correlation

- None of the chronic condition pair exceeds |0.7| indicating each chronic condition captures a different aspect of patient risk.

Comorbidity Burden

- Logarithmic increase for <30-day readmission rate as `numb_diagnoses` also increases.

4. Feature Engineering

In alignment with CMS-style readmission metrics, this part transformed raw variables into representations that are domain-specific, meaningful, and usable by models

Check and remove duplicates

Drop unnecessary columns

- `encounter_id`and `patient_nbr` columns are unique identifiers.
- Even if 23.45% of `patient_nbr` have multiple encounters, the corresponding multiple `encounter_id` per `patient_nbr` does not represent the sequence or duration in between readmission.
- `encounder_id` is also not a good data to do seasonality analysis since there is a possibility that `encounter_id` ordering is not chronological ordering.
- In terms of CMS-style readmission, they only evaluate the current encounter and outcome thus the said two columns may be dropped.

Processing Age of Patients

Since the `age` column contains categorical age ranges (e.g., '[0-10)', '[10-20)'), these categories were mapped to their respective average values using a dictionary. After the mapping, the average age of patients were calculated in the dataset using the `.mean()` method.

Processing Length of Stay

The `time_in_hospital` column contains numerical feature from 1 to 14 days. Two additional features will be created where one feature is the binary "long stay" flag since EDA shows risk increase after 7 days. The other feature column is a simple categorical bin for easier SHAP interpretation.

Processing Payer Code

The `payer_code` column contains 18 different categories of insurance payer. These will be grouped into standard healthcare payer categories to ensure reproducibility and alignment with established health services.

Processing Diagnosis Code

`diag_1`, `diag_2`, and `diag_3` will be grouped into clinical categories to reduce dimensionality and prevent overfitting. Chronic disease indicators and comorbidity count will be created to capture disease burden that influences readmission risk.

Processing Admission and Discharge

`admission_type_desc`, `admission_source_desc`, and `discharge_disposition_desc` columns will be transformed into binary indicators capturing emergency, care transitions, and post-discharge dependency to reduce categorical noise.

5. Feature Selection

This part of the project used filter-based approach to identify features statistically associated with 30-day readmission while remaining model-agnostic, reducing overfitting and improving interpretability.

Chi-square test for filtering categorical columns

Since the dataset is high-dimensional and dominated by categorical features, chi-square filter approach is used because it can efficiently evaluates the statistical association between each categorical feature and readmission outcoome independently.

Step 4: Model Implementation

In this part of the project, the following will be conducted:

1. Three models will be implemented to train the model
    - **Regularized Logistic Regression** as baseline model to benchmark more complex machine learning approach
    - **Random Forest** for the nonlinear features relationships and robust to outliers
    - **GradientBoost** as the final model due to its strong performance to model complex nonlinear interactions associated with readmission risk.
2. CMS-style evaluation metric

Scaling of numerical features

This numerical feature scaling is conducted inside the pipeline to prevent data leakage to the test set. RobustScaler* is chosen because it can mitigate the effect of extreme outliers from the utilization features.*

One-hot encoding of categorical features

This encoding is conducted inside the pipeline since to prevent data leakage to the test set and to keep the encoded categoricals untouched by scaling.

> "_Unknown category encountered during transformation were encoded as all zeros to prevent data leakage and ensure model robustness. While this may reduce information for rare categories, it aligns with best practices for healthcare predictive modeling and maintains deployment stability_"

Multicollinearity Checking

This part of the project identidies redundant predictors that could compromise interpretability and generalization

```
Out[118]:
```

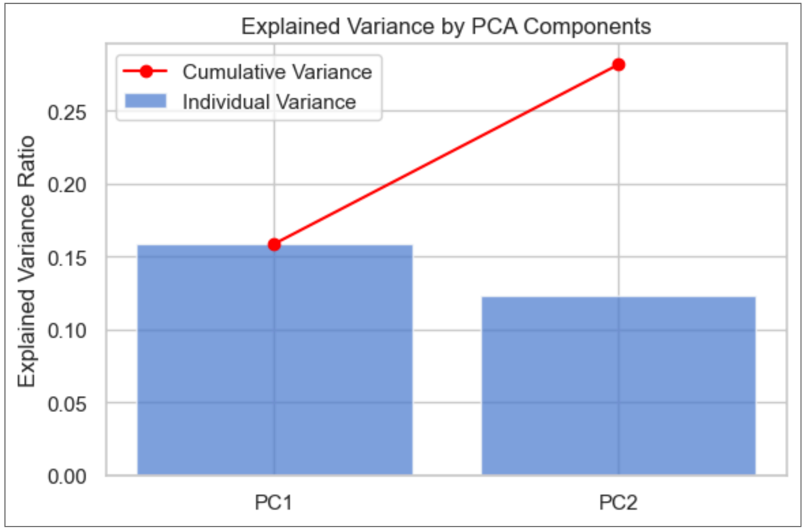
	Feature1	Feature2	corr
860	long_stay_flag	los_group_ord	0.763290
2211	discharged_home	discharged_to_facility	0.718852

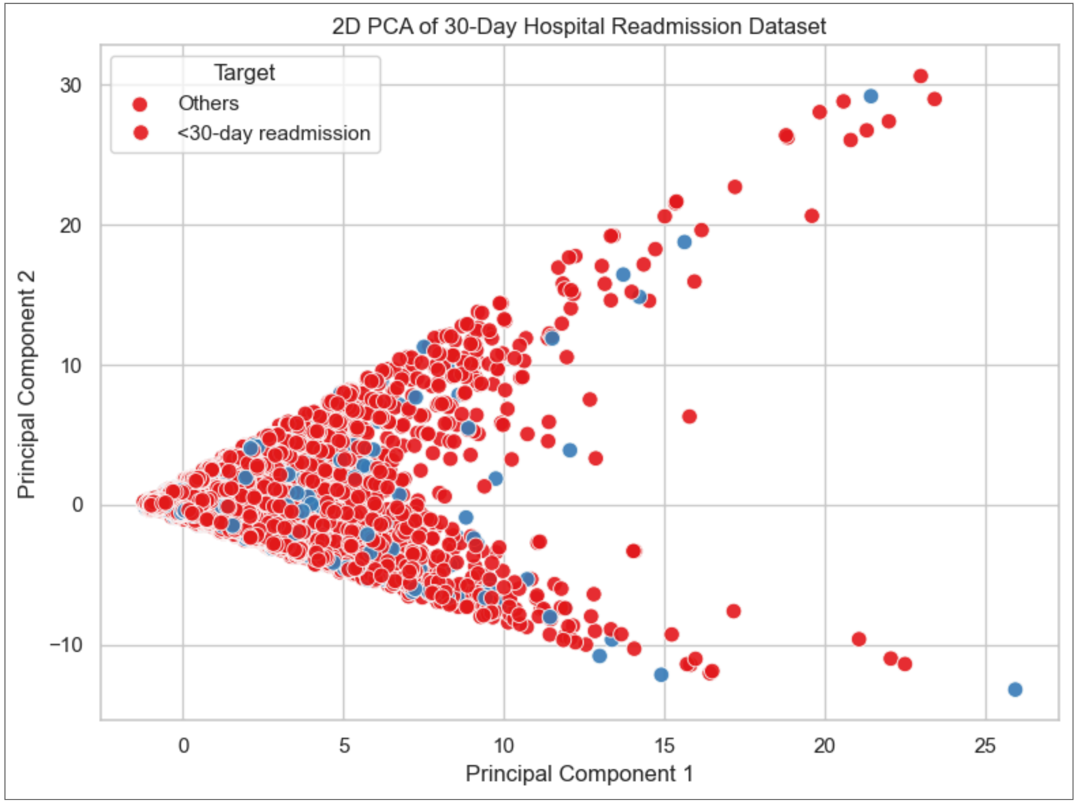
```
Dropped features due to multicollinearity:
['los_group_ord', 'discharged_to_facility']
```

Dimensionality Reduction

This part of the project reduced high-dimensional data from diagnosis codes, medical-specialty, and admission & discharged patterns that resulted to thousands of features.

```
Explained Variance Ratio (2 components): [0.15862313 0.12285977]
Cumulative Variance: [0.15862313 0.28148291]
```




2D PCA of 30-Day Hospital Readmission Dataset
Target
● Others
● <30-day readmission
Principal Component 2
30
20
10
0
−10
Principal Component 1
0
5
10
15
20
25

Train model

```
Best model saved: GradientBoosting (AUC-ROC = 0.6736)
```

Out[127]:

	Model	Recall @ P≥0.80
1	RandomForest	0.00044
2	GradientBoosting	0.00044
0	LogisticRegression	0.00000

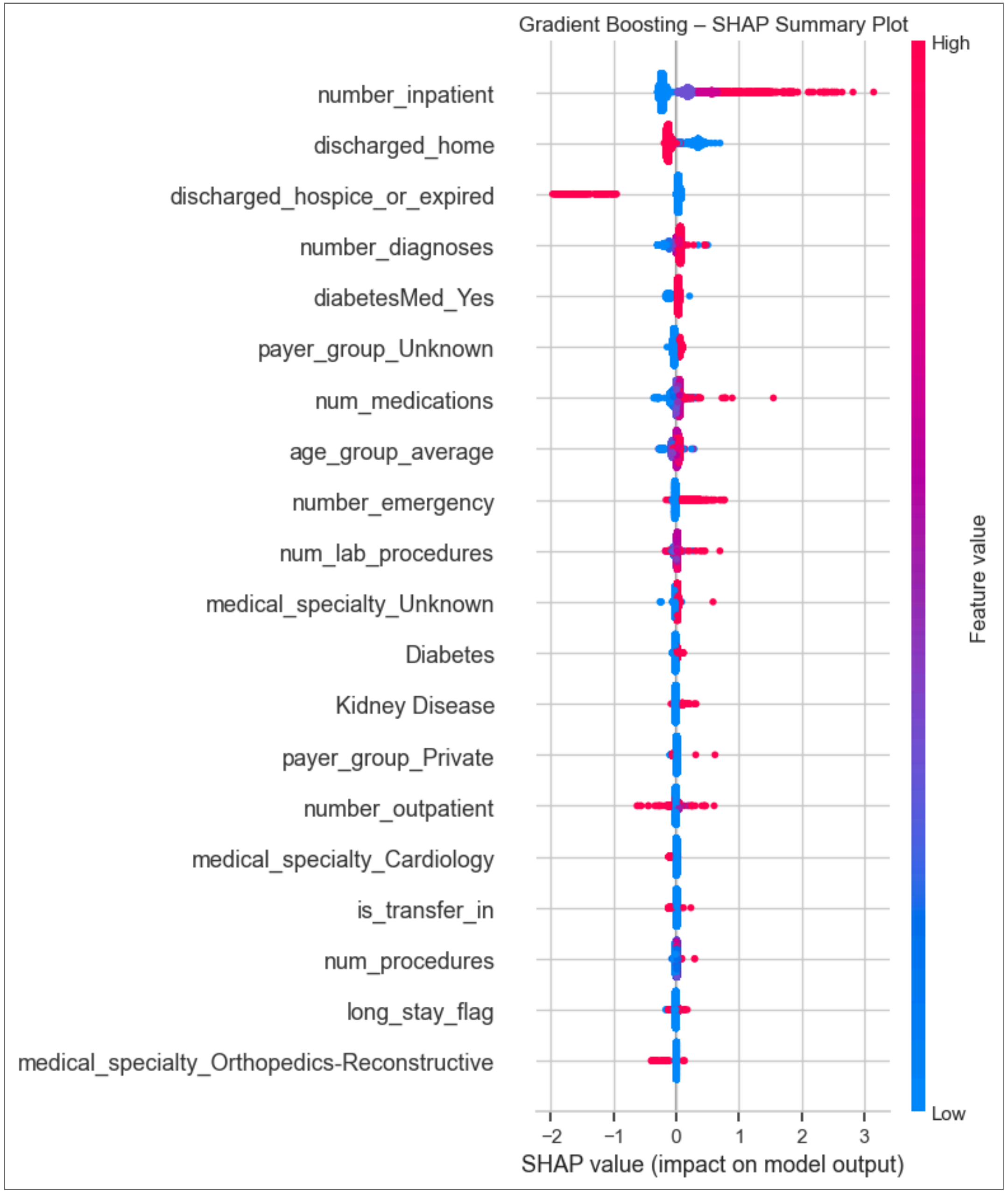
Recall @ fixed precision (important for imbalanced medical datasets)

```
LogisticRegression - Recall @ Precision ≥ 0.8: 0.000
RandomForest - Recall @ Precision ≥ 0.8: 0.000
GradientBoosting - Recall @ Precision ≥ 0.8: 0.000
```

Step 5: Critical Thinking → Ethical AI & Bias Auditing

Model Explainability

SHAP tool is used to explain each feature contribution value to push a specific patient's risk relative to the baseline readmission rate.


Gradient Boosting – SHAP Summary Plot
number_inpatient
discharged_home
discharged_hospice_or_expired
number_diagnoses
diabetesMed_Yes
payer_group_Unknown
num_medications
age_group_average
number_emergency
num_lab_procedures
medical_specialty_Unknown
Diabetes
Kidney Disease
payer_group_Private
number_outpatient
medical_specialty_Cardiology
is_transfer_in
num_procedures
long_stay_flag
medical_specialty_Orthopedics-Reconstructive
−2
−1
0
1
2
3
SHAP value (impact on model output)
High
Low
Feature value

SHAP Summary Plot Interpretation:

- The strongest driver for readmission is `num_inpatient` with high feature value on the right side of baseline means *high inpatient visit = increase risk*
- But since it also has a *wide spread*, this feature *still varies for each patient*.
- Other features that have **positive wide SHAP** spread are:
    - `num_medications`
    - `num_emergency`
    - `num_procedures`
- The features with **negative SHAP** that reduces the probability of readmission:
    - `discharge_home`
    - `discharge_hospice_or_expired`
    - `medical_specialty_Orthopedics-Reconstructive`
    - if the patient has the following feature value above, their readmission risk goes down
- The `num_outpatient` feature both have almost equal **high feature values on the positive and negative side of the baseline** which means that high outpatient visit count sometimes increase risl or decrease risk depending on patient context.

Fairness Metric Interpretation per attributes:

1. `race`: lower performance metrics for `AfricanAmerican` group compared to `Caucasian` group (majority) and 0 performance metrics for the remaining race suggesting that the model is less likely or not likely to predict readmission even when readmission actually occurred indicating **potential under-detection for this group**.
2. `gender`: slightly lower performance metric result for `Male` group compared to `Female` group in terms of **positive prediction and true positive prediction rates**. The false positive rate may be exceedingly lower for Male but it means there are missed true positives, which is a critical concern in clinical risk prediction settings.

Disparate Impact Interpretation per attributes:

- Both `race` and `age` resulted to *disparate value of 0.0* that means there are certain population that do have 0 PR and TPR indicating the **models' non-responsiveness for these populations**.
- While `gender` resulted to a *disparate value of 0.66* indicating that the unprivileged population's PR = 66% PR of the majority population which is **below the 0.80 threshold**.

Overall Disparate Impact Interpretation:

- Severe outcome disparities across `race` and `age` groups, and moderate disparity across `gender`, with results **largely influenced by class imbalance and subgroup underrepresentation**.

Document Limitations

1. Class imbalance
    - class imbalance was addressed by using `class_weight` for Logistic Regression and Random Forest. This is the primary strategy because I want to **keep the clinical realism of the data** instead of creating unrealistic synthetic samples.
    - **No threshold tuning** for GradientBoost was conducted to mitigate class imbalance becuase I used a code that will iterate to the 3 models chosen above.
    - But since all the True Positive Rates are all <0.60, `SMOTE` will be the **second strategy** for recall improvement.
2. Feature Limitations
    - The raw columns of my feature engineered data were dropped for the purpose of reduced redundaync and dimensionality reduction.
    - This is not a standard practice but training the model with thousands of features takes too much time to train and tune and retrain again.
    - I hypothesized that if I created feature engineered columns based on CMS domain-specific, it will have better perfomance compared to other existing hospital readmission models, but it seems to be not the case.
    - This strategy is largely affected by **class imbalance**.
3. Data Leakage
    - I went back and forth if I should put *scaling and encoding before model implementation* (as the capstone_guide and previous lessons suggested) or *putting after splitting the train and test set based on what I read on ResearchGate to prevent data leakage***.
    - It is possible that the performance metric result is caused by this. Either it will overfit or underfit, reinforcing the complexity of the problem.

Future Work

1. Learn to minimize runtime while processing high-dimensional data.
2. Improve skills on how to handle class imbalance.
3. Improve skills on parameter tuning.