Optimizing Hospital Readmission Reduction Using Patient Clustering

April 1, 2025

Deep Manish Mehta

dm29655n@pace.edu

Practical Data Science

MS in Data Science

Seidenberg School of Computer Science and Information Systems

Pace University

Agenda

- Executive summary
- Project plan recap
- Data
- Exploratory data analysis (EDA)
- Modeling methods
- Findings
- Business Recommendations & Technical Next Steps
- Q&A
- Appendix

Executive summary

Problem

- Hospital readmissions drive up costs for healthcare systems and reduce patient satisfaction. Traditional methods may predict whether readmissions occur but often fail to explain the underlying causes.

Solution

- This initiative applies a data-driven approach to identify patient subgroups that exhibit higher risks for readmission.
- Advanced clustering techniques are used to cluster patients based on shared risk factors and behaviors.
- Findings from these analyses aim to guide hospitals in implementing targeted interventions to reduce readmissions and enhance patient outcomes.

Project plan recap

Deliverable	Due Date	Status
Data & EDA	03/25/25	Complete
Methods, Findings, and Recommendations	04/01/25	Complete
Final Presentation	04/22/25	Complete

Data

Data

Data Source: Publicly available, open-source dataset from the UCI Machine Learning Repository
Diabetes 130-US Hospitals (1999–2008)

Sample Size: Approximately 100,000 rows, where each row represents a single hospital admission for a patient with diabetes

Time Period: January 1999 through December 2008 across 130 U.S. hospitals

Included Data: Patient demographics (e.g., age, gender), Admission details (e.g., length of stay, diagnoses), Medications (number and changes in prescription), Readmission status

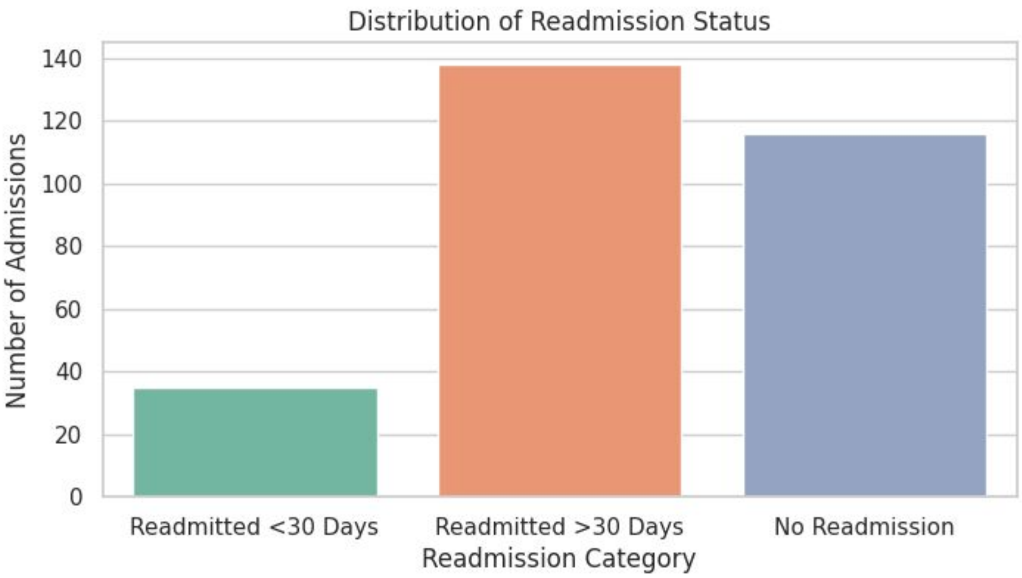
Excluded Data: Records with incomplete readmission information or critical missing fields.

Notes & Assumptions: The dataset provides a representative sample of adult diabetic patient admissions in U.S. hospitals during the study period; missingness is assumed to be random and not systematically bias readmission patterns

See Appendix A1 and Appendix A2 for detailed Data Cleaning steps and Data Preprocessing

Exploratory Data Analysis (EDA)

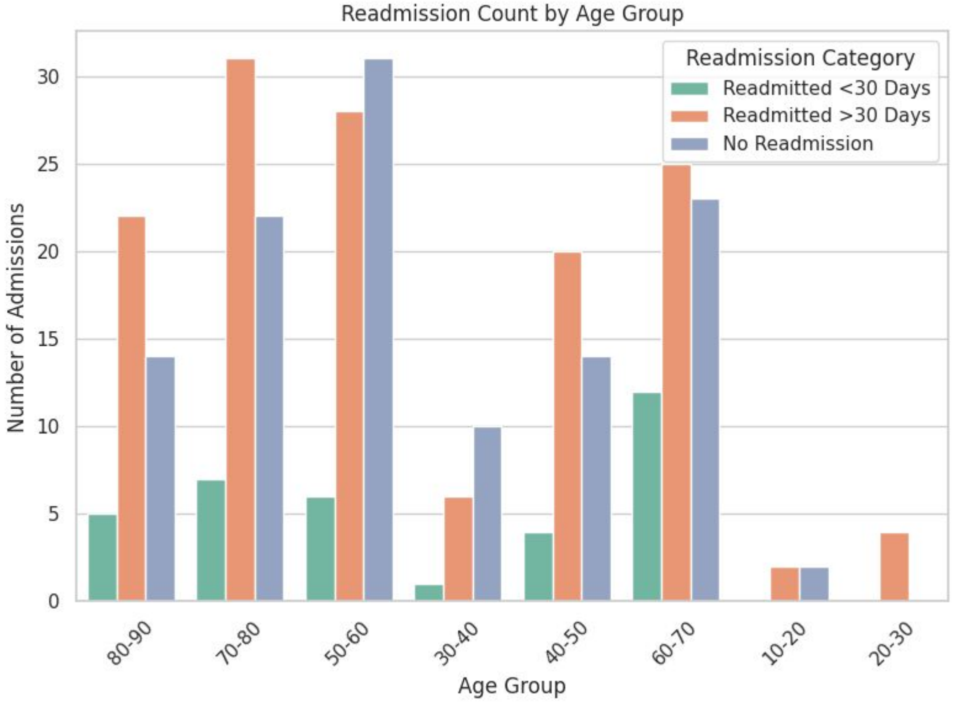
Nearly Half of Diabetic Admissions Return Within 30 Days



Observation:
47.7% of diabetic patient admissions were readmitted within 30 days, compared to just 12.1% with no readmission. (Refer Appendix A4 for full breakdown.)

Key Takeaway:
Cutting 30-day readmissions offers the biggest opportunity to reduce costs and improve patient care quality.

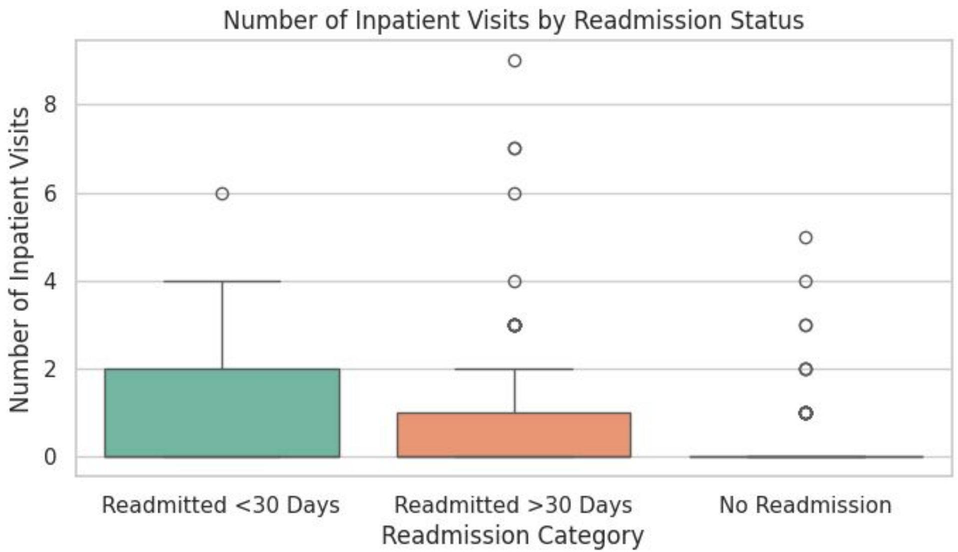
Patients Aged 50–70 Drive the Majority of Early Readmissions



Observation:
The 50–60 and 60–70 age groups account for the highest counts of readmissions within 30 days.

Key Takeaway:
Older adults (ages 50–70) are at highest risk for early readmission, indicating these age groups should be prioritized for targeted post-discharge support.

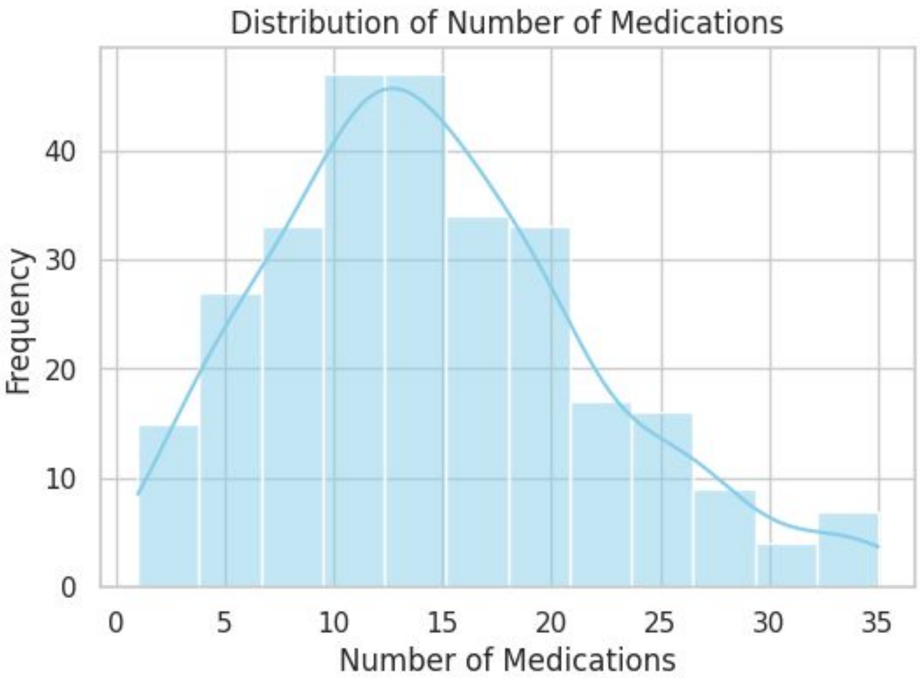
Frequent Prior Hospital Visits Signal Higher Readmission Risk



Observation:
Patients who return to the hospital within 30 days often have more prior inpatient visits than those who do not return or who return after 30 days.

Key Takeaway:
Focusing on patients with multiple previous admissions can help healthcare providers identify and proactively manage those who are most at risk for being readmitted soon.

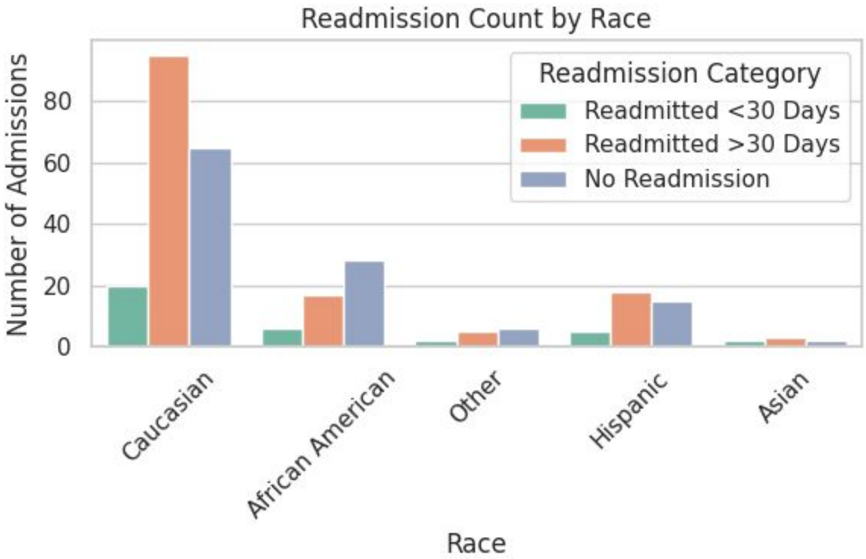
Higher Medication Burden Linked to Early Readmission



Observation:
Patients readmitted within 30 days take slightly more medications on average than those not readmitted.

Key Takeaway:
A higher medication burden may contribute to early readmission, suggesting medication reconciliation and adherence support could reduce risk.

Significant Racial Disparities Exist in Readmission Rates



Observation:
Caucasian patients show the highest absolute number of early readmissions, but African American patients have a higher readmission rate (18.7% vs. 12.8% for Caucasians).

Key Takeaway:
Even though Caucasians represent the largest volume of readmissions, the disproportionately higher rate among African American patients signals a clear equity gap — targeted, culturally-sensitive interventions are needed to close this disparity.
(Refer Appendix A5 for detailed breakdown)

Modeling Methods

Modeling Approach – Identifying Subgroup Characteristics

Outcome Variable:
A binary indicator is created for each subgroup (1 = belongs; 0 = does not belong), providing a clear target to distinguish patient groups.

Features Used:
Engineered patient attributes include:

- Healthcare Utilization: Total visits, outpatient visits, and follow-up compliance.
- Medication Management: Medication count.
- Clinical Severity: Severity score and hospital days per diagnosis.
- Demographics: Key indicators such as race and age.

(Refer Appendix A6 for detailed breakdown)

Model Type & Rationale:
A Random Forest classifier in a one-vs-rest framework is used because it captures complex patterns while offering clear insights into the factors driving subgroup membership.

Key Insight:
The approach reveals the critical drivers that differentiate patient subgroups—providing a foundation for targeted interventions to reduce hospital readmissions.

Modeling Process – From Clustering to Actionable Insights

Clustering Overview:

- Unsupervised clustering groups patients into distinct subgroups based on their healthcare behavior and clinical attributes.

Transforming Clusters into Predictions:

- Each subgroup is converted into a binary target (1 = membership, 0 = non-membership) so that a predictive model can be trained to identify subgroup membership.

Predictive Modeling:

- A Random Forest classifier is used in a one-vs-rest framework. This model distinguishes each subgroup from the rest and identifies the key factors driving subgroup differences.

Explainability with SHAP:

- An explainable AI technique (SHAP) is applied to determine which patient characteristics are most influential in defining each subgroup.

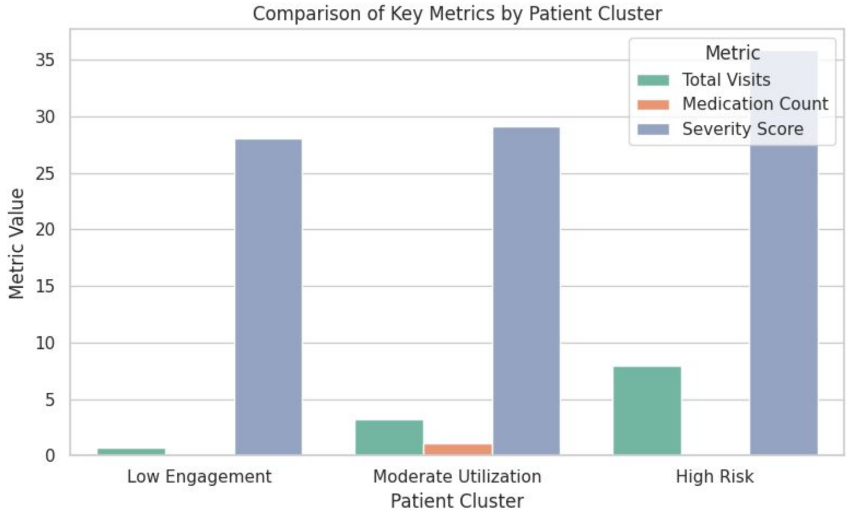
Actionable Insights:

- The process uncovers the critical drivers behind each patient group, offering clear guidance for targeted interventions aimed at reducing readmissions.

(Refer Appendix A7 for detailed breakdown)

Findings

Distinct Patient Subgroups and Their Characteristics



Key Features & Profiles

- **Low Engagement:** Minimal outpatient visits, low medication changes, lower overall interaction.
  (Refer Appendix A9 for Detail Breakdown)

- **Moderate Utilization:** Balanced hospital stays, moderate time in hospital, relatively diverse racial profile.
  (Refer Appendix A10 for Detail Breakdown)

- **High Risk:** Multiple outpatient visits, higher clinical severity scores, frequent total visits.
  (Refer Appendix A11 for Detail Breakdown)

Overview

- Clustering identified three distinct patient subgroups: Low Engagement, Moderate Utilization, and High Risk.

- Unique healthcare utilization patterns and clinical severity distinguish each subgroup. (Refer Appendix A8 for technical performance metrics)

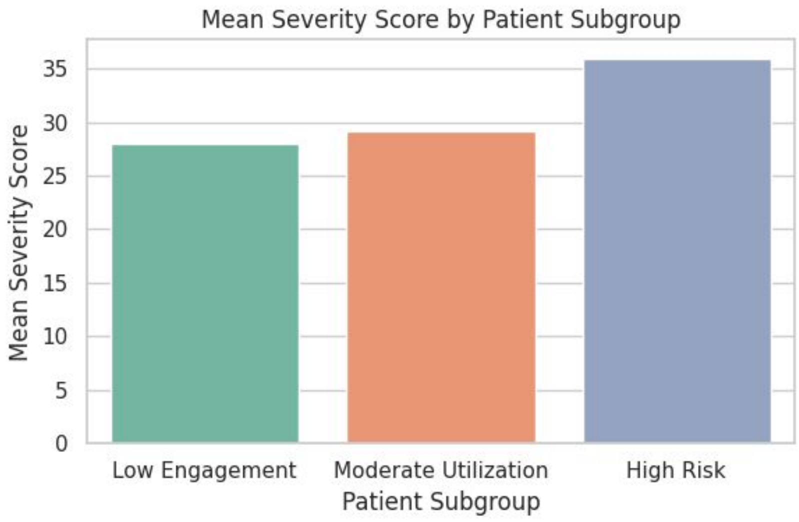
Readmission Trends and Strategic Intervention Opportunities

Readmission Trends:

- High Risk subgroup exhibits highest frequency of total visits and clinical complexity, indicating increased likelihood of readmission.
- Low Engagement subgroup shows minimal follow-up and outpatient visits, potentially leading to delayed treatment.
- Moderate Utilization subgroup maintains steady patterns that may escalate if not properly monitored.

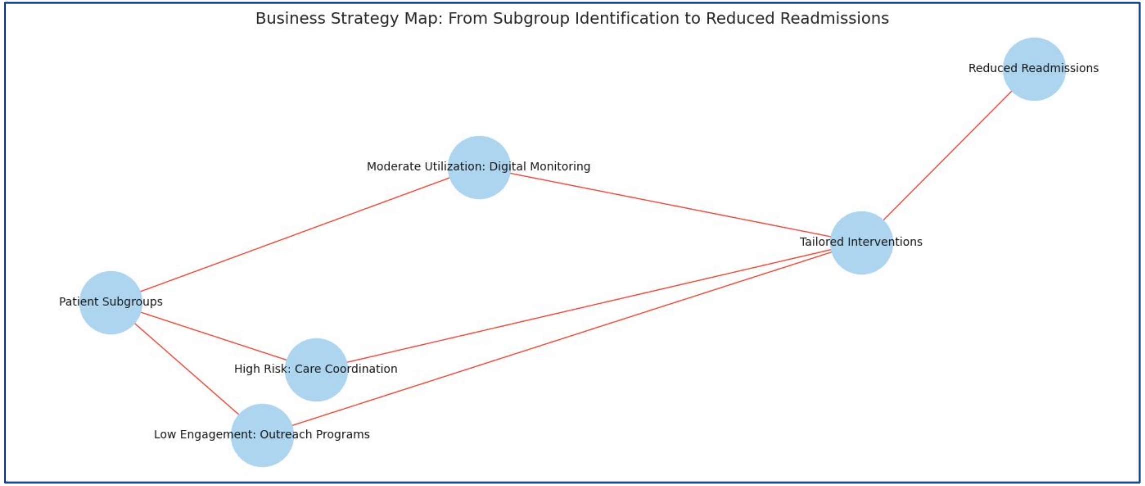
Intervention Opportunities:

- Tailored care programs for High Risk subgroup to reduce repeated hospital stays.
- Proactive outreach for Low Engagement subgroup to enhance follow-up and prevent emergencies.
- Consistent monitoring for Moderate Utilization subgroup to maintain stability and avoid escalation.



Business Recommendations & Technical Next Steps

Business Recommendations & Next Steps – Targeted Strategies for Reducing Readmissions

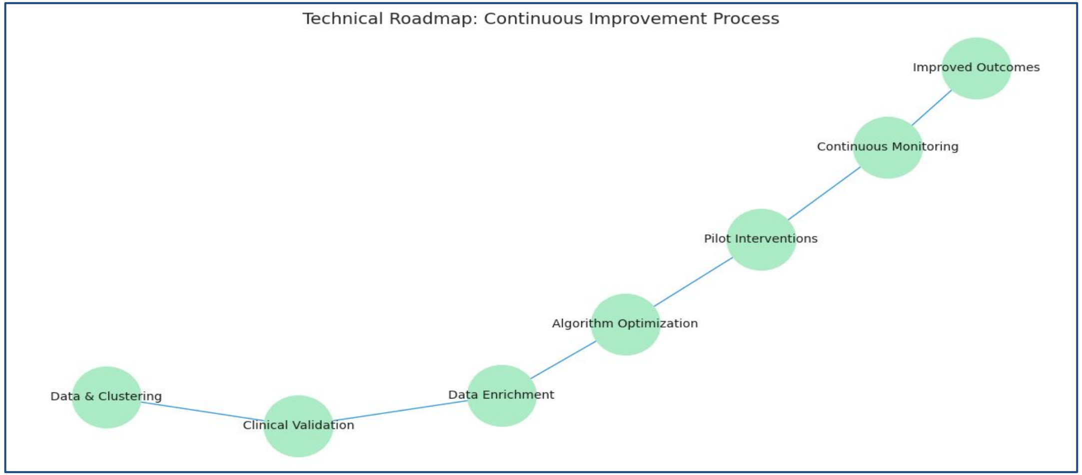


High Risk Subgroup:
- Intensive care coordination and regular medication reviews.
- Deploy case management teams to support patients with frequent hospital visits.

Low Engagement Subgroup:
- Proactive outreach (e.g., reminder calls, community health programs).
- Enhance follow-up protocols to prevent delays in treatment.

Moderate Utilization Subgroup:
- Maintain steady monitoring using digital engagement (telehealth, mobile reminders).
- Prevent escalation through periodic check-ins.

Business Recommendations & Next Steps – Technical Next Steps for Continuous Improvement



Clinical Validation:
- Collaborate with healthcare experts to verify subgroup definitions.

Data Enrichment:
- Integrate additional data sources (e.g., social determinants, patient feedback) to refine clusters.

Algorithm Optimization:
- Fine-tune clustering parameters and explore alternative methods to improve subgroup detection.

Pilot Programs & Continuous Monitoring:
- Implement targeted pilot interventions and track readmission rates for iterative improvement.

Appendix

Appendix Slide A1: Data Cleaning & Missing Values

- **Duplicates**
  Dropped 55 duplicate rows (0.04% of original dataset).

- **Missing-Value Handling**
  Replaced all “?” entries with NaN
  Columns dropped due to >40% missingness: weight, payer code, medical specialty

- **Columns Removed**
  encounter_id, patient_nbr (unique identifiers)
  weight (96% missing), payer_code (40% missing), medical_specialty (50% missing)

- **Row Removal**
  After dropping rows with any remaining missing values, final shape: 98,247 rows × 30 columns (2.5% of rows removed)

Appendix Slide A2: Data Preprocessing & Feature Engineering

- **Binary Target Creation**
  readmitted_flag: 1 if readmitted <30 days, else 0

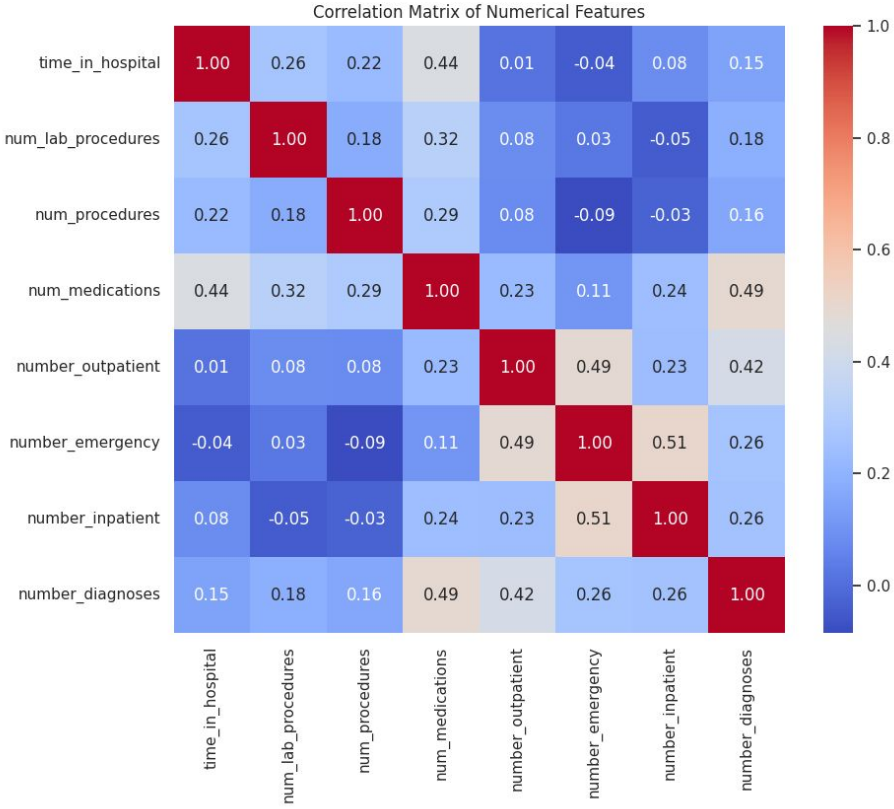
- **Categorical Encoding**
  All remaining categorical columns label-encoded for modeling

- **Feature Binning**
  Emergency visits binned into 4 categories: 0, 1, 2, 3+

- **Age**
  Retained as original 10-year buckets (no numeric conversion)

- **Final Feature Set (30 total)**
  Patient demographics, admission details, diagnosis groups, medication counts, and readmission_fla

Appendix Slide A3: Detailed Correlation Heatmap



Key Technical Takeaways:

- **Multicollinearity risk:** Strong correlations (>0.50) between emergency visits, inpatient stays, and outpatient visits suggest grouping or dimensionality reduction may improve model stability.

- **Feature selection:** num_medications and number_diagnoses both capture patient complexity; consider combining into a single “clinical burden” metric.

- **Low-correlation features:** num_procedures (max 0.29 correlation with any other feature) may provide unique information for clustering/modeling.

Appendix Slide A4: Detailed Readmission Status Breakdown

Readmission Category	Count	Percentage of Total Admissions
No readmission (0)	34	12.1%
Readmitted within 30 days (1)	138	47.9%
Readmitted after 30 days (2)	116	40.3%

Calculation Method:
Percentage = (Category Count ÷ Total Admissions) × 100

Details:
Counts and percentages computed from the final cleaned dataset (n = 289 admissions) using **df['readmitted'].value_counts()**

Readmission categories coded as:

- 0 = no readmission
- 1 = readmitted within 30 days
- 2 = readmitted after 30 days

The main presentation Slide 8 offers a simplified overview of Distribution of Readmission Status

Appendix Slide A5: Detailed Race vs Readmission Breakdown

Race Code	Race Description	Total Admissions	Readmissions (<30d)	Readmission Rate
3	African American	9,812	1,832	18.7%
2	Caucasian	51,273	6,541	12.8%
1	Asian	1,217	140	11.5%
4	Hispanic	2,284	243	10.6%
5	Other	1,001	97	9.7%

Technical notes:

Counts computed via **df.groupby(['race', 'readmitted_flag']).size()**

Rates calculated on cleaned dataset after dropping missing values and non-clinical columns

How to interpret these numbers:

- Readmission rate = (Readmissions(<30d) ÷ Total Admissions) × 100
- African American patients have an 18.7% readmission rate, markedly higher than Caucasian patients at 12.8%.
- This supports the Slide 10 narrative: “African American patients have an 18.7% readmission rate versus 12.8% for Caucasian patients.”

The main presentation Slide 12 offers a simplified overview of Racial Disparities

Appendix Slide A6: Technical Details – Modeling Approach

Outcome Variable Creation:
A binary target is generated for each patient subgroup using:
- target = (cluster_label == target_cluster).astype(int)

This conversion transforms unsupervised cluster labels into a supervised classification task for one-vs-rest analysis.

Feature Set Overview:
Engineered features include patient attributes related to healthcare utilization (total visits, number_outpatient, follow-up compliance, outpatient_ratio), medication management (medication_count, num_medications_log), clinical severity (severity_score, hospital_days_per_dx, comorbidity_count), and key demographics (race, age).

Random Forest in a One-vs-Rest Framework:
- For each patient subgroup, a Random Forest classifier is trained using a one-vs-rest approach.

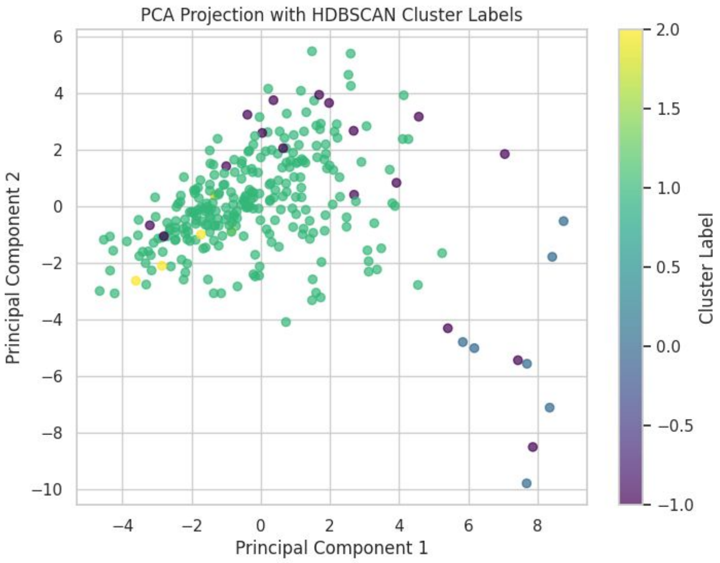
One-vs-Rest Framework Explanation:
- For each subgroup, patients are labeled as “1” if they belong to the subgroup and “0” otherwise.
- A separate model is trained for each subgroup, isolating the features that uniquely drive membership in that group.

Random Forest Classifier:
- An ensemble method that builds multiple decision trees from bootstrapped samples and random subsets of features.
- Provides robust performance with clear insights into feature importance, which are further refined by SHAP analysis.

Model Parameters:
Key hyperparameters include:
n_estimators = 100, max_depth = 10, min_samples_leaf = 2, max_features = "sqrt"
These settings balance model complexity with generalization.

The main presentation Slide 14 offers a simplified overview of Modeling Approach

Appendix Slide A7: Technical Details – Modeling Process



Clustering Integration:

- HDBSCAN is applied to scaled features (X_scaled) to derive patient clusters.
- PCA is used to visualize cluster separation and assess grouping quality.

Transformation for Interpretation:

- Unsupervised cluster labels are converted into binary targets (one-vs-rest) for each subgroup.
- A Random Forest classifier is trained on these targets to capture subgroup-specific drivers.

SHAP Analysis:

- SHAP's TreeExplainer is employed with settings:
  • model_output = "raw"
  • feature_perturbation = "interventional"
  • Additivity check is disabled (explainer.check_additivity = False)
- SHAP values for the positive class (membership) are computed to identify the top 5 features by average absolute contribution.

Cluster Profiling:

- The top features are extracted and descriptive statistics are computed for each subgroup.
- This process provides quantitative profiles for each cluster that inform targeted interventions.

The main presentation Slide 15 offers a simplified overview of Modeling Approach

Appendix Slide A8: Detailed Model Performance Metrics

Model Performance Metrics by Patient Subgroup

Patient Subgroup	Mean CV Score	Train Accuracy	Test Accuracy
Low Engagement	0.924	0.987	0.9138
Moderate Utilization	0.9827	1.0	0.9828
High Risk	0.9897	0.9957	0.931

Detailed performance metrics in this slide support the high-level findings that the subgroups (Low Engagement, Moderate Utilization, High Risk) are statistically distinct and the models are robust.

- This slide provides the underlying performance statistics for the clustering models used to define the patient subgroups.
- The metrics reinforce the Findings slide's claim that each subgroup is reliably distinct. The performance details include 5-fold cross-validation scores, train accuracy, and test accuracy for the three subgroups (Low Engagement, Moderate Utilization, and High Risk).
- These metrics offer technical credibility behind the simple average comparisons seen on the Findings slides.

The main presentation Slide 17 offers a simplified overview of Findings

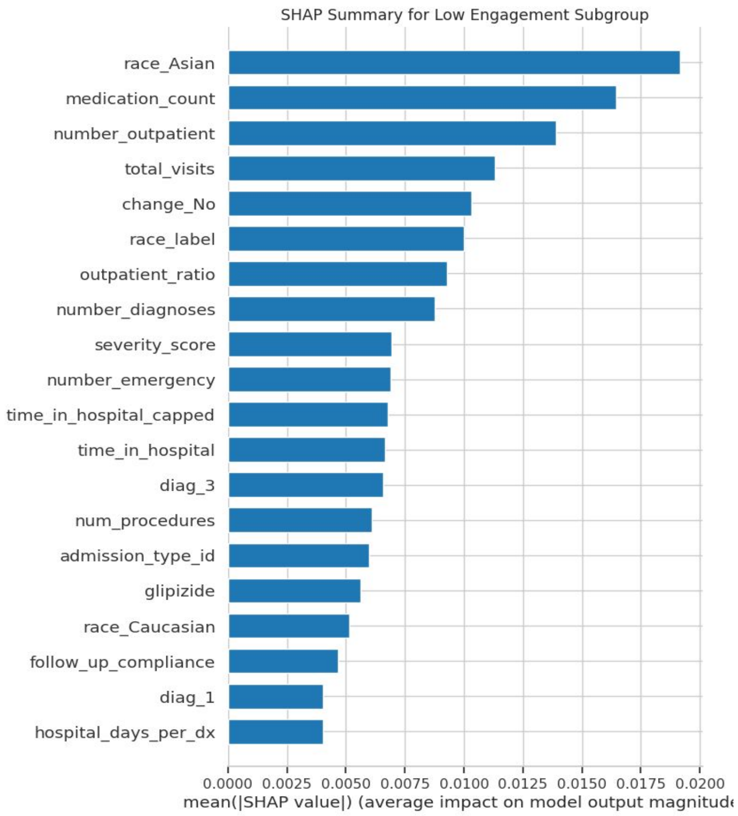
Appendix Slide A9: Low Engagement Profile

Low Engagement Profile:

Technical evidence indicates that the Low Engagement profile is characterized by minimal outpatient visits and low medication changes. SHAP analysis identifies features such as race_Asian, medication_count, and number_outpatient as critical drivers, confirming the simplified insights outlined in the Findings section.

Low Engagement Profile:

- Top 5 Features: ['race_Asian', 'medication_count', 'number_outpatient', 'total_visits', 'change_No']
- Descriptive statistics (e.g., mean, standard deviation) for features such as medication_count, number_outpatient, and total_visits.



The main presentation Slide 17 offers a simplified overview of Findings

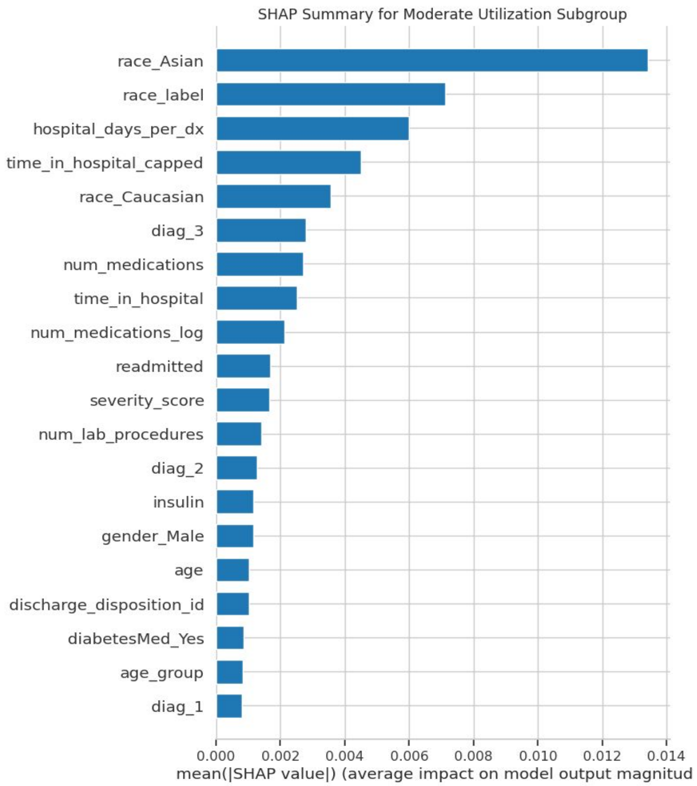
Appendix Slide A10: Moderate Utilization Profile

Moderate Utilization Profile:

Technical evidence reveals that the Moderate Utilization profile exhibits balanced healthcare use and moderate hospital days. SHAP analysis highlights features including race_Asian, race_label, hospital_days_per_dx, and time_in_hospital_capped, aligning with the key characteristics presented in the Findings section.

Moderate Utilization Profile:

- Top 5 Features: ['race_Asian', 'race_label', 'hospital_days_per_dx', 'time_in_hospital_capped', 'race_Caucasian']
- Summary statistics for hospital_days_per_dx and time_in_hospital_capped (mean ≈ 0.53 and 3.4 respectively).



The main presentation Slide 17 offers a simplified overview of Findings

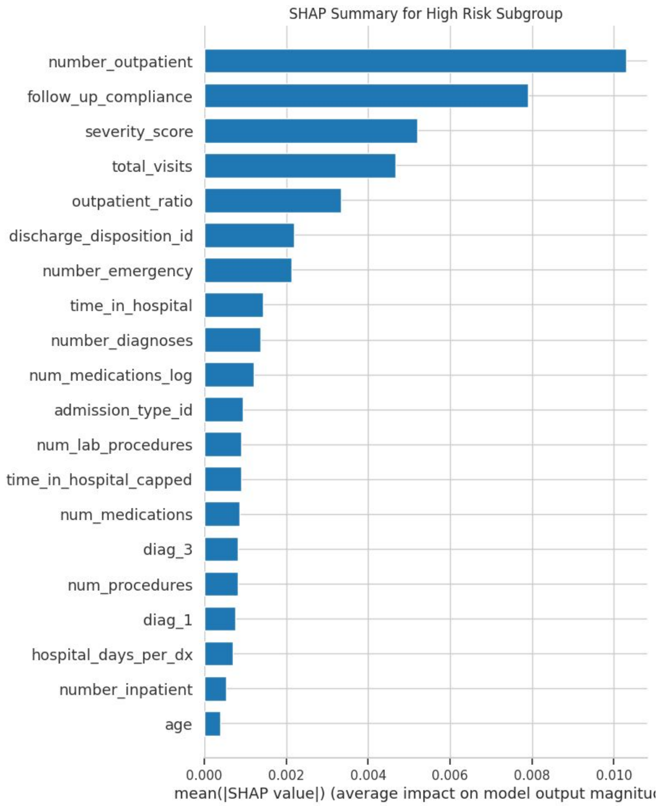
Appendix Slide A11: High Risk Profile

High Risk Profile:

Technical evidence demonstrates that the High Risk profile is marked by frequent outpatient visits and high clinical severity. SHAP analysis shows that features such as number_outpatient, follow_up_compliance, severity_score, total_visits, and outpatient_ratio are essential determinants, supporting the insights summarized in the Findings section.

High Risk Profile:

- Top 5 Features: ['number_outpatient', 'follow_up_compliance', 'severity_score', 'total_visits', 'outpatient_ratio']
- Descriptive statistics (e.g., mean severity_score ≈ 35.9, mean total_visits ≈ 8).



The main presentation Slide 17 offers a simplified overview of Findings

Appendix Slide A12: Additional Visualizations and Data Insights

Overview:

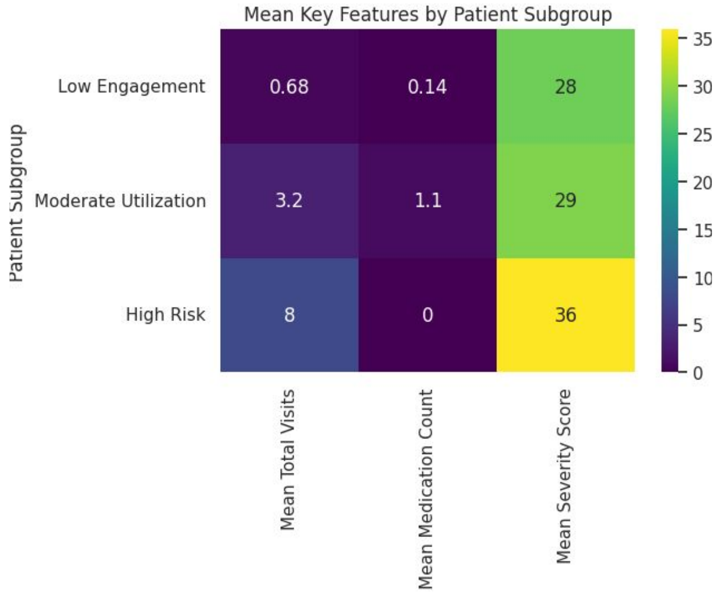
- This heat map displays average values for key metrics across the three patient subgroups (High Risk, Low Engagement, and Moderate Utilization).
- Each cell represents the mean value for a specific metric within a subgroup.

Key Metrics Included:

- Average Outpatient Visits
- Average Total Visits
- Average Severity Score

Insights Provided:

- Darker shades indicate higher average values, while lighter shades indicate lower values.
- The heatmap visually reinforces that the High Risk subgroup shows elevated healthcare utilization and clinical complexity compared to the other groups.



Project Materials

- Git Repo: https://github.com/deepmehta27/Practical_Data_Science_Project