College of Business

Master of Science in Business Analytics

Capstone Project – Fall 2025

Predictive Modeling for Hospital Readmission Reduction

Group Members:

Kwame Boateng Akomeah (L30099531)

Tony Lordson (L30092167)

Anil Kumar Swamy Bandaru (L30094053)

Dev Arora (L30098852)

Supervisor:

Ibrahim Mescioglu, PhD, Chair – Department of Business Analytics

Contents

Abstract

Hospital readmission remains one of the costliest and closely monitored challenges confronted by the healthcare system in United States, with national 30-day all-cause readmission rate of approximately 15%. This project builds an end-to-end analytics framework using 2021 New York SPARCS inpatient data to operationalize risk prediction into hospital workflows. Leveraging Data Management, Business Intelligence, and Machine Learning, the team engineered a proxy readmission flag aligned with national and state rates, designed a star-schema data warehouse and analytics view, and built dashboards to reveal that elderly patients, public insurance holders (especially Medicare), emergency admissions, and high-severity or high-mortality clinical profiles and diagnoses (infectious, respiratory, circulatory) have the greatest readmission risk. Classification models including Logistic Regression, Random Forest, and XGBoost were evaluated, with Random Forest selected as the primary operational model for its strong recall and discrimination, and XGBoost recommended as a complementary option for resource-sensitive scenarios. The project considered healthcare informatics and fairness considerations and, despite data limitations such as the absence of true patient-level 30-day readmission identifiers, provides a practical foundation, guiding targeted transitional care, and supporting ongoing quality-improvement decisions.

1. Introduction

Hospital readmission is among the most critical challenges faced by healthcare systems and providers worldwide. It remains one of the most persistent, costly, and closely monitored challenges confronted by the healthcare system in United States, with national 30-day all-cause readmission rates between 13–15% (Jiang, et al., 2023; Definitive Healthcare, 2025). This means billions in additional hospital stays and $15 billion annually in avoidable healthcare spending across the country (Bradley et.al., 2013).

Hospital readmission occurs when a patient who has been discharged returns to the hospital within a short period. A 30-day all-cause readmission is defined as the number of readmissions for all conditions which occur within 30 days (Jiang, et al., 2023). It is considered one of the most critical measures of quality inpatient care, patient safety, post-discharge outpatient care, and overall hospital performance. Therefore, high readmission rates not only strain hospital capacity and increase operational costs but also signals potential gaps in care delivery and coordination, discharge planning, patient education, and broader socioeconomic factors that influence health outcomes. For instance, a patient treated for pneumonia risks being readmitted within a week if there was inadequate discharge planning or follow-up, or vague instructions about medication, lifestyle changes, and diet.

Accordingly, reducing 30-day readmission rates remains both a clinical and strategic priority for healthcare systems and providers to enhance patient outcomes, optimize resource utilization, reduce operational costs and overall health spending. The Centers for Medicare & Medicaid Services (CMS) with the Hospital Readmissions Reduction Program (HRRP) in their efforts to reduce 30-day avoidable hospital readmissions introduced a penalty of up to 3% reduction in hospital inpatient Medicare payments.

1.1 Problem Statement

High hospital readmission rates are particularly pressing in big US states like New York and Illinois, where readmission patterns mirror national trends but have more cost and resource repercussions. Studies by CMS identifies Massachusetts, Florida, Illinois, Nevada, and West Virginia as the top 5 states recording high readmissions rates, just above 15% (Definitive Healthcare, 2025). These rates are mainly driven by factors such as the presences of some of the largest hospital in the US, limited access to primary care, fragmented post-discharge support, and social determinants of health such as housing instability, transportation barriers, and community-level health inequities (Laurent, n.d., and Definitive Healthcare, 2025).

This highlights both the scale of the problem and the opportunity for data-driven decision-making to improve quality care while reducing avoidable hospital costs. Leveraging publicly available datasets, literature, and industry best practices, this project seeks to develop comprehensive insights into the factors associated with 30-day hospital readmissions and propose a predictive model for early detection of high-risk patients.

1.2 Project Objectives

Integrating data analytics, machine learning and business decision-making, the goal of this project is to develop a comprehensive analytical framework that enables hospitals to:

a. Uncover key risk drivers, trends, and patterns from reports and dashboards.
b. Detect high-risk patients early using predictive modelling, and
c. Make strategic decisions to reduce 30-day hospital readmission rates.

This end-to-end analytics project aligns with all core competencies of Business Analytics, including Quantitative Methods, Business Intelligence and Data Analysis, Data Warehousing and Database Management, Visualizing Information, and Data Mining for Business Decisions.

2. Data Description

The data used for this project is the Hospital Inpatient Discharges (SPARCS De-Identified) dataset, a publicly available inpatient discharge data collected from hospitals across the state of New York in 2021. This dataset is widely used for research in healthcare analytics, policy evaluation, and quality improvement. Each record represents an individual hospital stay, including demographic, clinical, and administrative details. The dataset provides a comprehensive foundation for data-driven analysis and insights into hospital operations and performance, patient outcomes, and readmission reduction strategies. The key variables in the dataset include:

a. **Patient Demographics:** Age group, Gender, Ethnicity, Race, and Payment Typology. These are useful to understand population diversity and patient profiles.
b. **Facility:** Facility code, Hospital name, County, Service area, and ZIP code. These are useful to understand hospital and case distribution and performance.
c. **Admission/Discharge features:** Length of stay, Type of admission, and Patient disposition. These are useful to understand hospital utilization and patient flow.
d. **Clinical features:** Diagnostics, Procedures, Comorbidity indicators (Mortality risk and Severity rate), and Major Diagnostic Category (MDC). These are useful for evidence-based clinical and data-driven decision-making.

2.1 Data Preparation

The dataset was cleaned, cleared of missing values and duplicates, and encoded when necessary. It contained 33 columns, of which variables such as Operating Certificate Number, Facility Name, Birth Weight and description of drugs, diagnosis, and procedure not relevant for analysis were dropped. It is significant to note that only these features irrelevant for analysis contained missing values, as a result dropping them maintained data quality and integrity. Additional features such as Readmitted (Target Variable), Clinical Score, and Combined Score (Clinical and Frequency) were engineered to aid in effective healthcare analytics.

2.2 Deriving Target Variable

The dataset provides rich demographics and clinical information but lacked a readmission flag, therefore, the target variable was derived. Due to data privacy and encryption, the dataset also lacked patient identifiers, admission/discharge dates and history, or other temporal identities needed to compute true 30-day readmission events. In view of these constraints, direct readmission measurement is not achievable.

To address this limitation while maintaining a more realistic measurement, a proxy readmission target variable was derived using risk factors that are clinically validated by CMS as readmission risk factors (CMS, 2025). These risk factors were then calibrated to mirror the national and New York average all-cause 30-day hospital readmission rate of approximately 15% (Horwitz et al., 2014; Jencks et al., 2009; Chollet et al., 2011). This enables us to derive a

statistical estimate of readmission likelihood for healthcare analytics and predictive modeling. The steps involved in deriving the target variable include:

i. **Clinical Readmission Risk Score:** The key clinical predictors and weights used to generate clinical risk scores include Severity rate (40-50%), Mortality risk (23-35%), Medical and Surgical classifications (10-20%), and Major Diagnostic Category (mdc) code (5-10%). The Clinical Score Formula employed is:

$$Clinical\ Score\ = 0.45(Severity) + 0.35(Mortality) + 0.15(Surg) + 0.05(mdc)$$

ii. **Frequency Score:** To compute readmissions based on frequency of similar attributes at hospital facilities, frequency scores were estimated using facility_id, diagnosis_code, and mdc_code.

iii. **Combined Readmission Likelihood Score:** Combined scores encompassing clinical risk scores and frequency scores were created, with clinical features having higher weights than facility frequency. This is to balance the combined score since clinical factors are the main drivers of readmission and facilities with high frequency of similar patient attributes may have higher readmissions.

$$Combined\ Score = 0.75(Clinical\ Score) + 0.25(Frequency\ Score)$$

iv. **Calibrating Combined Score:** To derive an all-causes 30-day readmission rate that mirrors the national and New York readmission rate, the top 15% combined scores were sorted and labeled as likely readmissions.

This approach employed to derive the target variable provides a statistical and clinical estimate of readmission likelihood which aligns with CMS principles, New York estimates, and hospital readmission literature.

3. Methodology

This project employed Agile project management, emphasizing iterative progress, adaptability, and close collaboration throughout the process. The team followed an end-to-end analytical lifecycle, integrating all core competencies of Business Analytics to support data-driven healthcare decision-making. The approach ensured flexibility, rapid feedback, and continuous improvement, echoing best practices in modern healthcare analytics projects. The analytical lifecycle encompassed the following phases:

1. Data Preparation

- Data Cleaning: Addressed missing values, outliers, naming conventions, and inconsistencies to ensure high quality data.
- Feature Engineering: Developed key healthcare analytic fields such as the target variable and risk scores.
- Encoding and Scaling: Applied appropriate transformations to categorical and numerical attributes.
- Exploratory Data Analysis: Uncovered trends and relationships essential for further analytics and modeling.

2. Quantitative Analysis

- Hypothesis Building: Created testable predictions to assess key drivers of patient readmission from CMS and hospital readmission literature.
- Correlation Analysis: Identified relationships among patient features, risk scores, and readmission risk.
- Risk Stratification: Used descriptive statistics to inform patient segment and risk levels.

3. Data Warehousing & Database Management

- Star Schema/Entity-Relationship Diagram (ERD) Development: Created an optimized schema to support analytical queries and reporting.
- Fact and Dimension Tables: Built robust tables for hospital admissions and dimensions such as patient, clinical, facility, comorbidity, and payment typology.
- SQL Views: Created views to present data in simplified and reusable blocks for further queries and business intelligence.

4. Data Mining & Machine Learning

- Classification: Applied a suite of classifiers (Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting).
- Model Performance Evaluation: Used Accuracy, AUC, Precision, Recall, and ROC curves for model assessment.
- Interpretability: Leveraged feature importance to interpret the impact of features.

5. Business Intelligence & Visualization

- Dashboards: Delivered interactive dashboards that summarize hospital readmission and risk factors for key stakeholders.
- Interpretability: Leveraged key influencers, top segments, and scenario analysis to interpret and forecast the impact of features.

This agile approach ensured a robust, flexible, and business-oriented solution tailored to the complexities of business and data analytics project management.

4. Findings and Analysis

This section comprises of all core competencies of Business Analytics, as well as Healthcare Informatics.

4.1 Quantitative Methods for Business

This Descriptive analytics seeks to uncover and understand patterns in the dataset, particularly inpatient and 30-day hospital readmission dynamics in New York. This section utilizes business and statistical techniques, as well as data visualizations to provide a comprehensive overview of interactions in the dataset.

4.1.1 Hypothesis Building

For an informed data-driven framework and success criteria for decision-making, these Hypotheses were constructed based on CMS specifications and clinical literature.

- **H1: Patients with higher Severity Rate and Mortality Risk have higher 30-day Readmission Likelihood.**
  According to CMS and clinical literature, high severity rates and mortality risk are major risk factors positively correlated with risk of acuity post discharge and hence readmission.
- **H2: Public Insurance patients (Medicare, Medicaid, etc.) have higher 30-day Readmission Likelihood.**
  Demographic and socioeconomic factors such as age, access to care and follow-up, and discharge planning efficiency have an influence on readmission rates across the various health insurance services.
- **H3: Patients with Major Diagnostic Categories (MDC) such as Circulatory and Respiratory conditions have higher 30-day Readmission Likelihood.**
  According to CMS, major risk factors for readmission include chronic conditions such as Acute myocardial infarction (AMI), Chronic obstructive pulmonary disease (COPD), Heart failure (HF), Pneumonia, Coronary artery bypass graft (CABG) surgery, Elective primary total hip arthroplasty and/or total knee arthroplasty (THA/TKA)

4.1.2 Descriptive Analysis

The descriptive statistics below show that there are very few high-risk patients in the data.

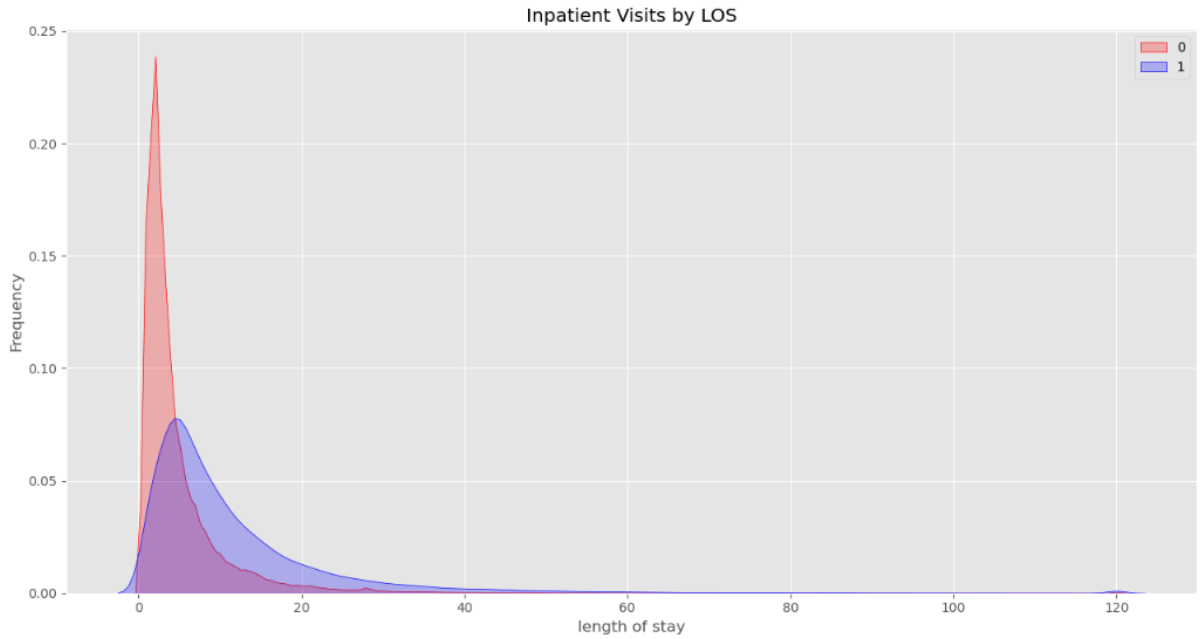
- most patient fall within Minor to Moderate severity with mean and median severity of 2.2 and 2 respectfully.
- most patient fall within Minor to Moderate mortality risk with mean and median mortality of 1.92 and 1 respectively.
- length of stay (LOS) is highly skewed due to the few extremely long LOS outliers (e.g. 120 days) compared to a mean and median of 6 and 3 days respectively. This highlights an uneven hospital utilization, as most patients stay fewer than a week in hospitals.

- clinical (mean: 0.54, median: 0.49) and combined (mean: 0.53, median: 0.51) scores are normally distribution, validating that most patients exhibit moderate aggregated health risks.

Table 1: Descriptive Statistics

	severity_code	mortality_risk	length_of_stay	clinical_score	combined_score
count	1521588.00	1521588.00	1521588.00	1521588.00	1521588.00
mean	2.20	1.92	6.23	0.54	0.53
std	1.01	1.10	9.09	0.22	0.18
min	1.00	1.00	1.00	0.20	0.15
25%	1.00	1.00	2.00	0.38	0.39
50%	2.00	1.00	3.00	0.49	0.51
75%	3.00	3.00	7.00	0.70	0.64
max	4.00	4.00	120.00	1.00	0.98

Figure 1: Distribution of LOS



In essence, while most patients discharged experienced mild to moderate health risks, there are also those with major and extreme severity and mortality cases who are likely to be readmitted. Moreover, most patients stayed less than a week in the hospitals which is consistent with the national average of 5.5 days since 2023 (CareSet, 2025). A short LOS can be an indicator for efficient care delivery and processes, however extremely short LOS could

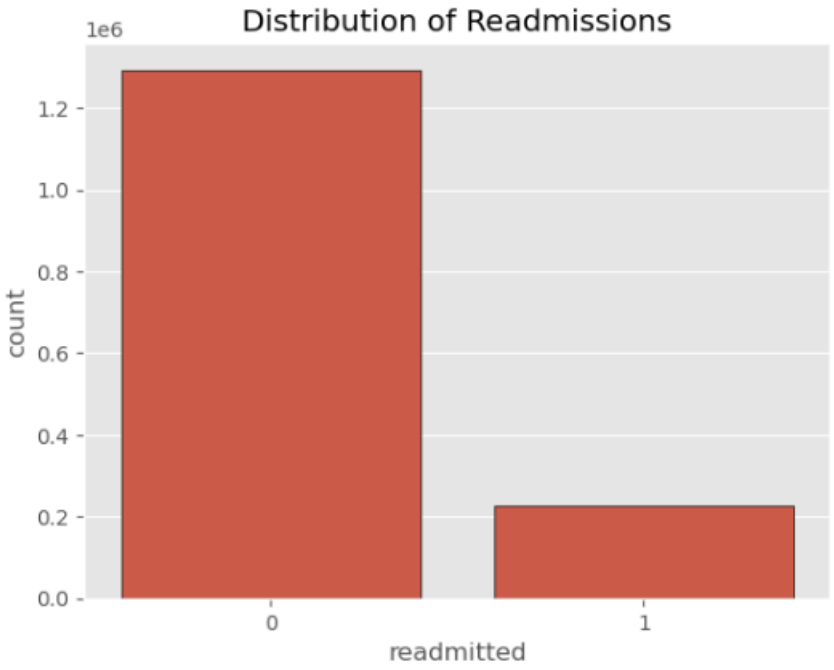
also signal premature discharge, leading to risk of acuity or readmission. On the other hand, long LOS can signal inefficient care delivery and processes, or complications, while extremely long LOS can be associated with necessary complex care such as rehabilitation.

4.1.3 Exploratory Data Analysis

a. Distribution of Target Variable

The Distribution of Readmissions shows a strong class imbalance in the target variable, highlighting that majority of patients discharged (approx. 1.3 million) were not readmitted (0), with approx. 220K readmissions (1). This imbalance in the data signals that predictive algorithms such as tree learning algorithms and performance metrics such as precision, recall, and ROC-AUC may be more useful in the analysis since they are less sensitive to class imbalance and give more robust picture of model performance.

Figure 2: Distribution of Target Variable

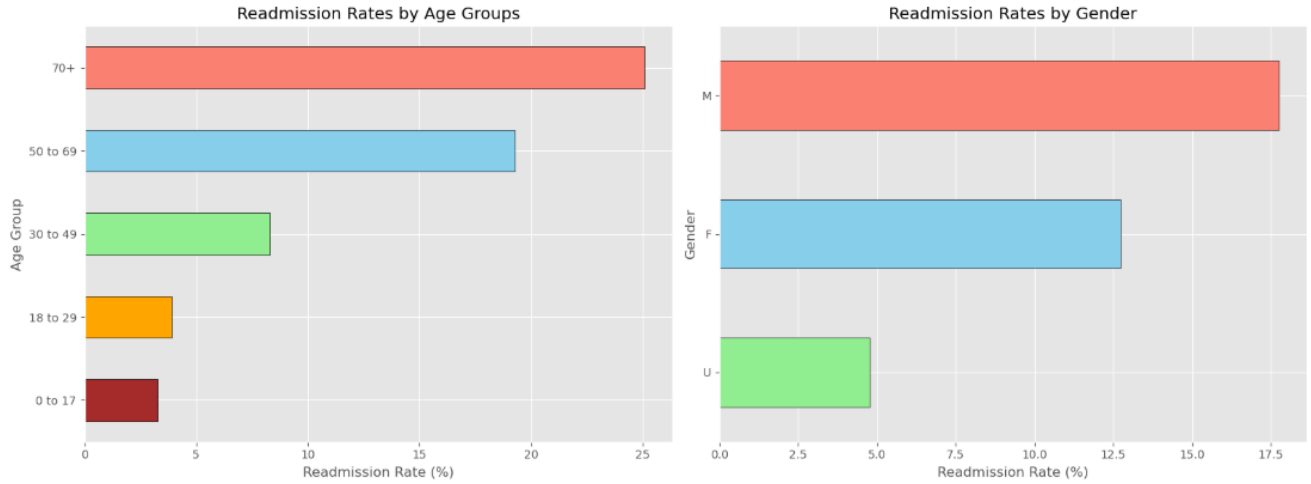


b. Readmission by Age Group and Gender

The readmission rates by age group and gender show how readmission rates vary across patient demographics such as age group and gender.

- Rates by age group highlight that the elderly (70+) have the highest readmission rate, followed by 50 to 69, with middle age and younger groups having the lowest risk of readmissions. This suggests that readmission risks increase with age and as a result, transitional care interventions should be targeted towards the elderly.
- Rates by gender indicate gender differences (although minimal) in readmission risk, with males having the higher risk of readmissions.

Figure 3: Readmission by Age Group and Gender



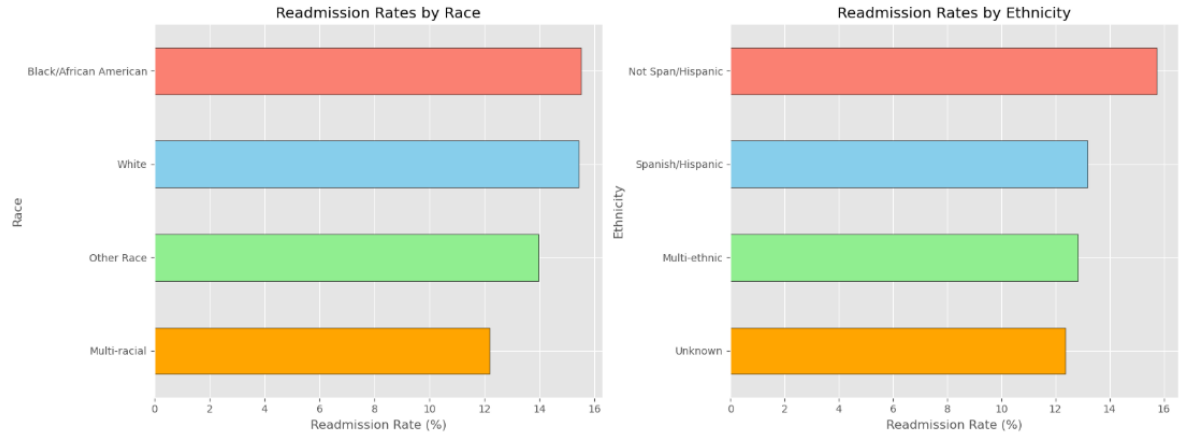
c. Readmission by Race and Ethnicity

The readmission rates by race and ethnicity show how readmission rates vary across patient demographics such as race and ethnicity.

- Rates by race indicate that Black/African American and White patients are the race at most risk of readmission (15-16%), with Multi-racial patients showing the lower risk rates (around 12%).
- Rates by ethnicity show that Not Spanish/Hispanic patients have the highest readmission rate (about 16%), with patients with unknown ethnicity having the lowest (about 12%).

These patterns indicate slight variations in readmission risk across race and ethnicity, suggesting that readmission interventions should be implemented across all racial and ethnic populations.

Figure 4: Readmission by race and Ethnicity

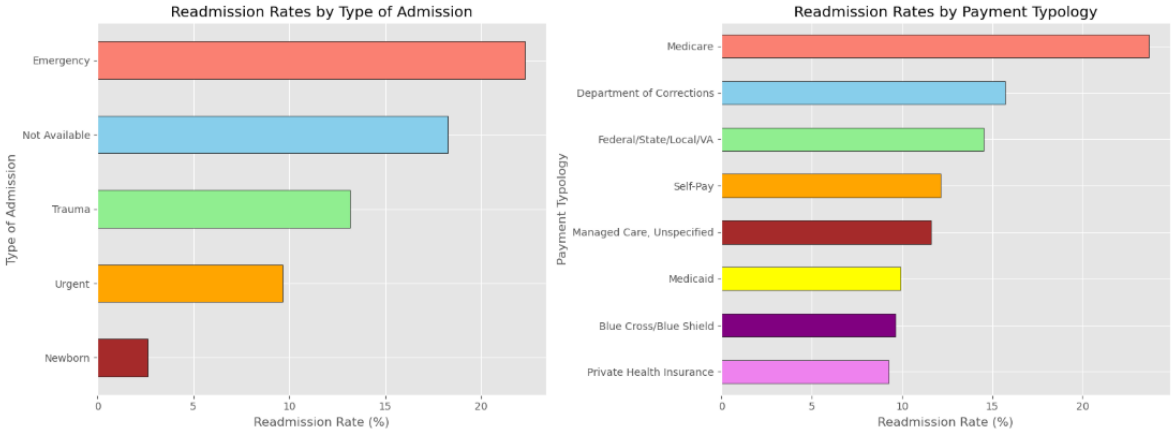


d. Readmission by Admission Type and Payment Typology

The readmission rates by admission type and payment typology show how readmission rates vary across type of inpatient visit and insurance coverage.

- Rates by admission type indicate that inpatient visits that were Emergency cases were at the highest risk of being readmitted (approx. 27%), with Newborns having the least likelihood of readmission (about 2%).
- Rates by Payment Typology validate ***H2*** such that public health insurance holders including Medicare (approx. 27%), Department of Corrections (approx. 16%), and Federal/State/Local/VA (approx. 14%) are at a higher risk compared to private insurance patients (approx. 9%).

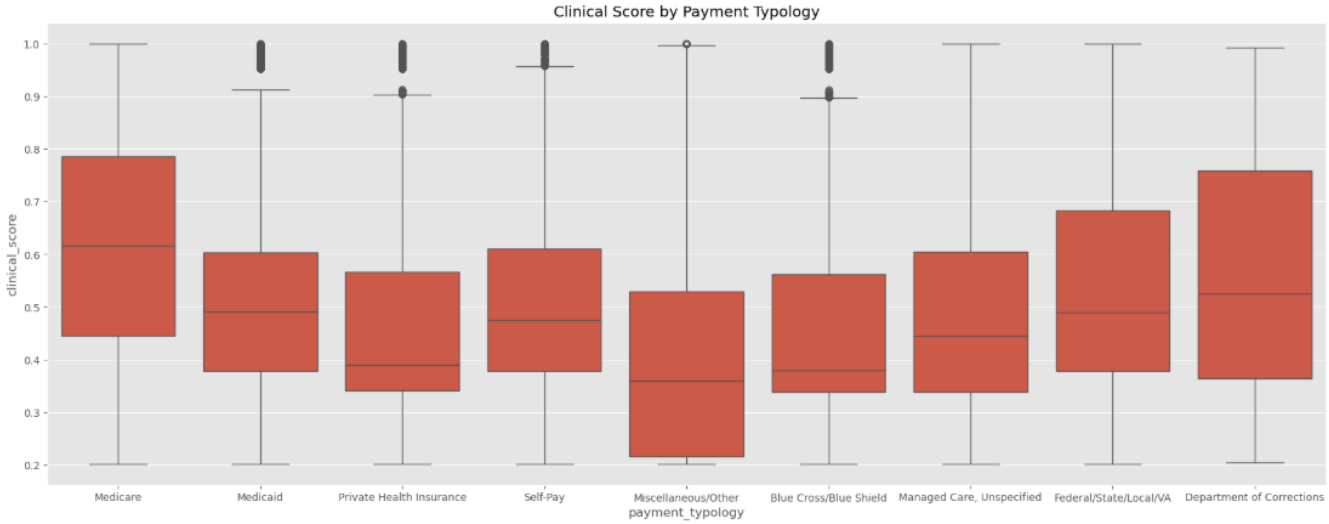
Figure 5: Readmission by Admission Type and Payment Typology



e. Clinical Score by Payment Typology

The boxplot shows that Medicare patients have the highest clinical risk (approx. 0.62) while Private Health Insurance, Blue Cross/Blue Sheild, and Miscellaneous/Other patients (approx. 0.39-0.35) are at the lowest clinical risk. There are a few outliers across all payment typologies, which may represent the major and extreme severity and mortality risks, as well as extremely long LOS cases possibly driving readmission risk. This finding also validates ***H2*** such that Medicare, Medicaid, Department of Corrections, and Federal/State/Local/VA coverages are at a higher clinical risk compared to private insurance patients.

Figure 6: Clinical Score by Payment Typology



f. Aggregate analysis of combined score by payment typology

The aggregate analysis of combined scores by payment typology also validates that public insurance holders including Medicare and Medicaid patients are at high aggregate risk of acuity and readmission.

Table 2: Aggregate Summary of Combined Score by Payment Typology

Payment Typology	Mean	Median	std	Count
Blue Cross/Blue Shield	0.48	0.46	0.17	169809.00
Department of Corrections	0.50	0.47	0.19	745.00
Federal/State/Local/VA	0.51	0.49	0.19	12697.00
Managed Care, Unspecified	0.48	0.46	0.18	18813.00
Medicaid	0.51	0.50	0.16	473583.00
Medicare	0.58	0.57	0.19	571416.00
Miscellaneous/Other	0.42	0.38	0.16	16689.00
Private Health Insurance	0.49	0.48	0.17	239840.00
Self-Pay	0.51	0.49	0.18	17996.00

g. Top Major Diagnostic Categories (MDC)

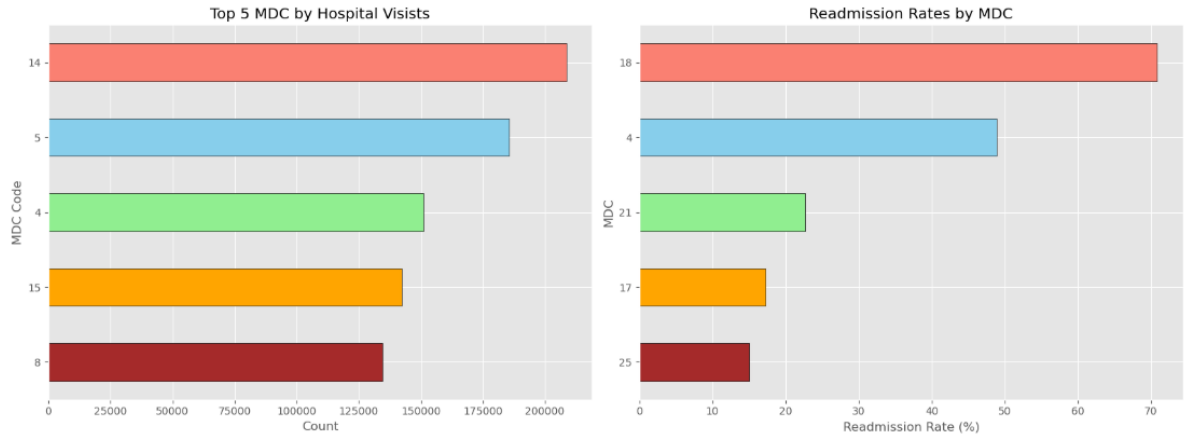
The readmission rates by MDC and count of MDC show how readmissions vary across the various MDC.

- Top 5 MDCs by hospital visits show the 5 MDC which had the highest number inpatient visits. They include:
    - pregnancy and childbirth (14), circulatory (5), respiratory (4), newborn (15), musculoskeletal system and connective tissue disease (8)

- Rates by MDC indicate that infectious and parasitic diseases (18) amount to a significantly highest risk of readmission (approx. 70%), suggesting that it is crucial to channel clinical interventions and transitional care management towards these diseases. It is then followed by:
    - respiratory (4), injuries, poisonings and toxic effects of drugs (21), myeloproliferative diseases and poorly differentiated neoplasm (17), and multiple significant trauma (25)

Comparing these charts emphasized that high volume of visits by MDC does not always signal higher readmission risk, and vice versa. Only respiratory conditions have both a higher number of visits and a higher readmission risk, validating ***H3*** that patients with respiratory conditions have a higher likelihood of being readmitted within 30 days.

Figure 7: Top 5 MDC by Count and Readmission Rate



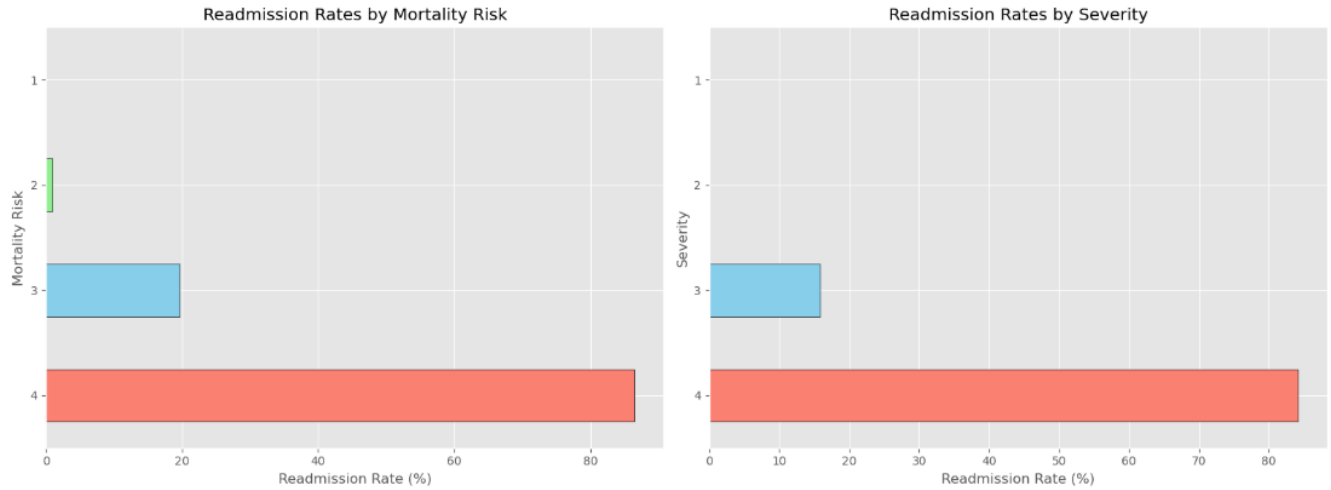
h. Readmission by Mortality Risk and Severity

The readmission rates by Mortality risk and Severity show how readmissions vary across the various acuity rates.

- Rates by mortality risk show that Extreme (4) mortality risk has very high risk of readmission, followed by Major (3) risks, with Moderate (2) and Minor (1) showing very low to no readmission risks. This suggests that patients assessed as having highest mortality risk are extremely likely to be readmitted.
- Similar to mortality risk, rates by severity show that Extreme (4) severity has very high risk of readmission, followed by Major (3) risks, with Moderate (2) and Minor (1) showing no readmission risks. This suggests that patients with the most severe conditions are extremely likely to be readmitted.

This finding validates ***H1*** such that the higher the mortality risk and severity, the higher the likelihood of 30-day readmission risk, confirming that severity and mortality risk increase risk of acuity and are strong predictors of readmission.

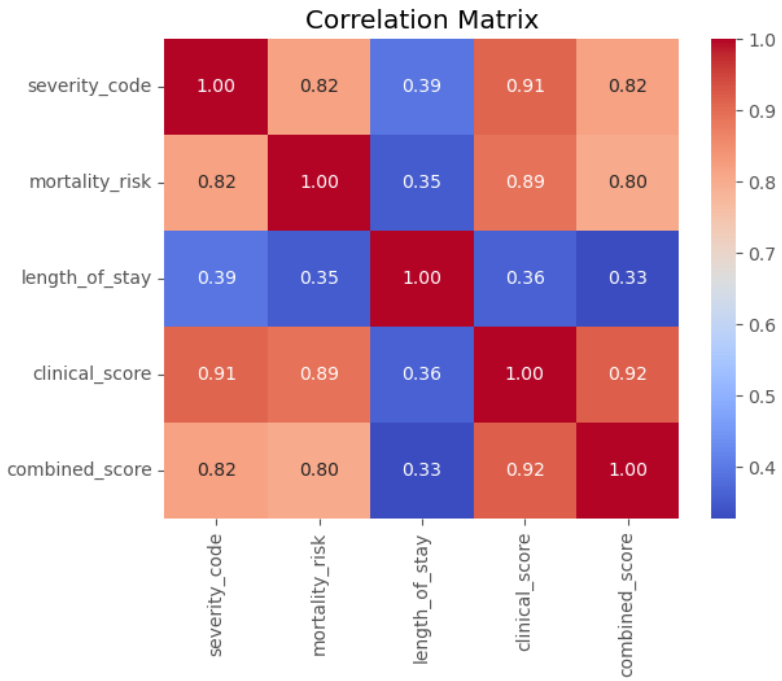
Figure 8: Readmission by Mortality Risk and Severity



i. Correlation Matrix

The correlation matrix shows a strong correlation between the comorbidity predictors (r=0.82), since severity rates and mortality risk are both clinical risk factors. Moreover, the clinical score is strongly correlated with severity rate (r = 0.91) and mortality risk (r = 0.89) since both features were employed to derive the clinical score. Similarly, combined score is strongly correlated with clinical score (r=0.92) since it was also used to derive the combined score, and with a higher weight than frequency score. These strong correlations suggest a multicollinearity concern; a limitation raised from deriving the target variable with these features. On the other hand, LOS is moderately correlated with all the measures, highlighting that other external factors may also influence how long patients stay in the hospital.

Figure 9: Correlation Matrix



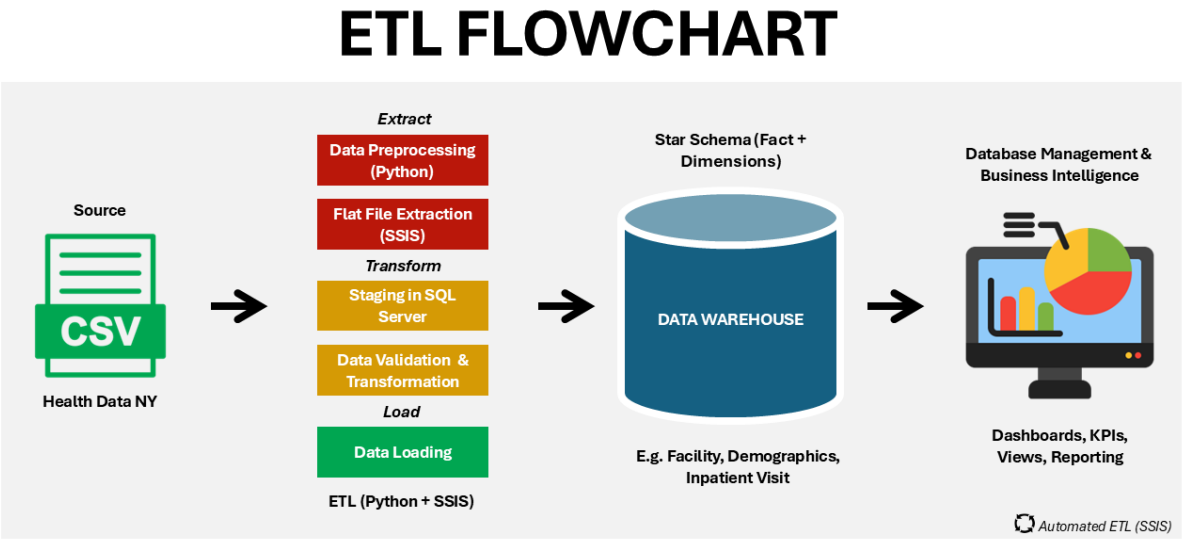
The findings of this quantitative analysis demonstrate a strong statistical support for hypotheses and justify including these predictors in the predictive modeling. They also suggest that reduction programs should be targeted when necessary to achieve the maximum benefits from interventions.

- ***H1:*** Patients with higher severity and mortality risk show drastically high likelihood of readmission.
- ***H2:*** Public insurance holders such as Medicare patients showed higher clinical risks and acuity compared with privately insured patients.
- ***H3:*** Patients with circulatory and respiratory conditions (among other conditions) experience risk of acuity and higher likelihood of readmission.

4.2 Business Data Warehousing

This section involves creating a data repository to support further Database and Business Intelligence analytics. Using Python together with Microsoft SQL Server and Visual Studio, a data storage environment was created locally in MS SQL Server.

Figure 10: ETL Pipeline Flowchart



The steps involved in creating the Inpatient data warehouse include:

a. **Automated ETL Pipeline with SSIS:**
   - Extracted dataset from Health Data NY, preprocessed in python, and extracted as flat file.
   - Utilized SQL Server Integration Services (SSIS) to automate Extract, Transform, and Load (ETL) from flat file in Visual Studio.

b. **Staging and Validation:**

- Loaded raw data into SQL Server staging table for initial validation.
- Applied transformations and type conversions for star schema population.

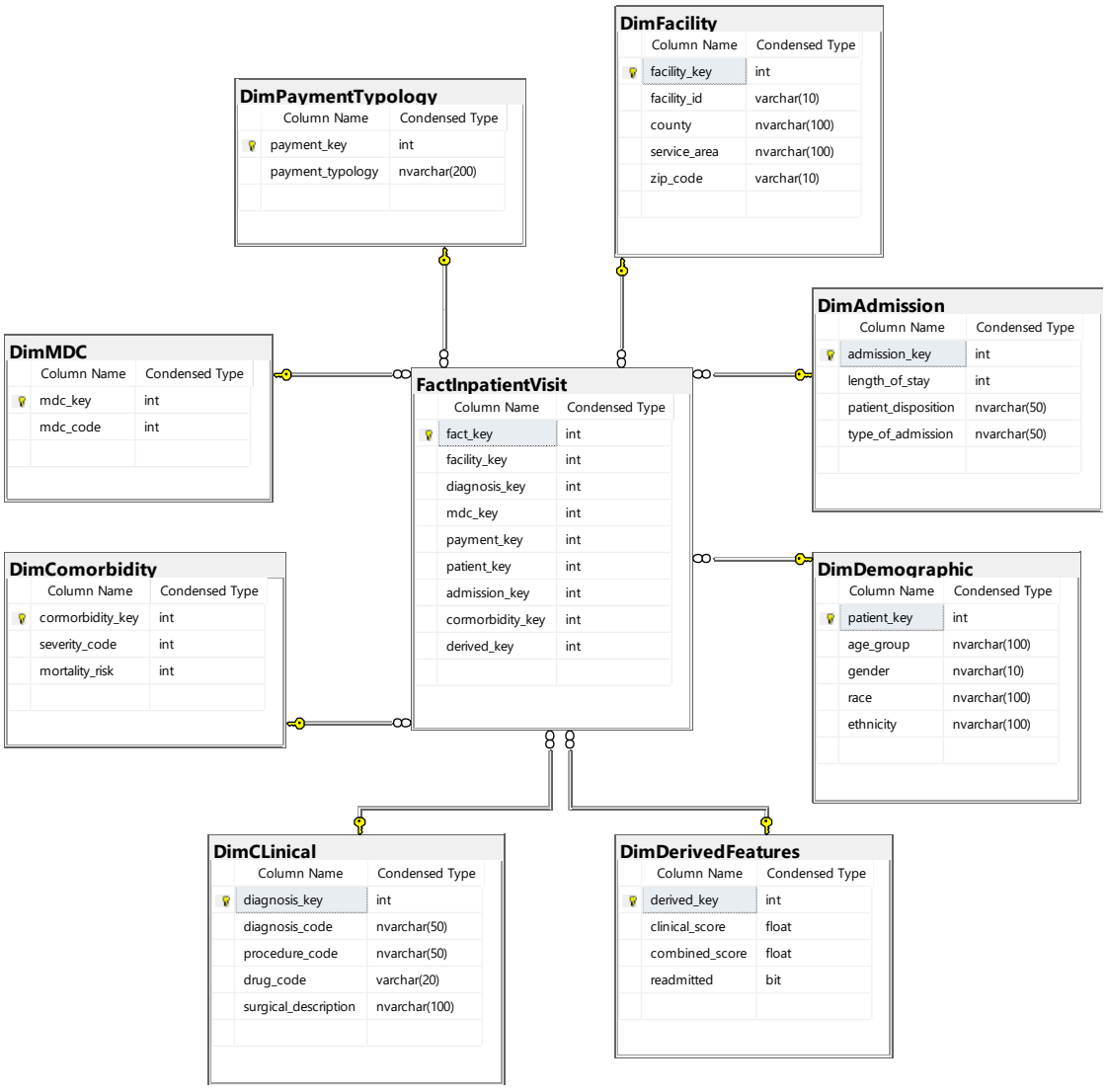
c. Star Schema Architecture:

- Designed a fact table with surrogate keys for robust dimensional linking.
- Created dimension tables for attributes, including any additional relevant categories.

d. Data Integration and Loading:

- Loaded staging data to dimensions and fact table, ensuring efficient warehouse structure and data integrity.
- Developed the star schema model to optimize query performance and support scalable reporting.

Figure 11: Star Schema/ ER Diagram



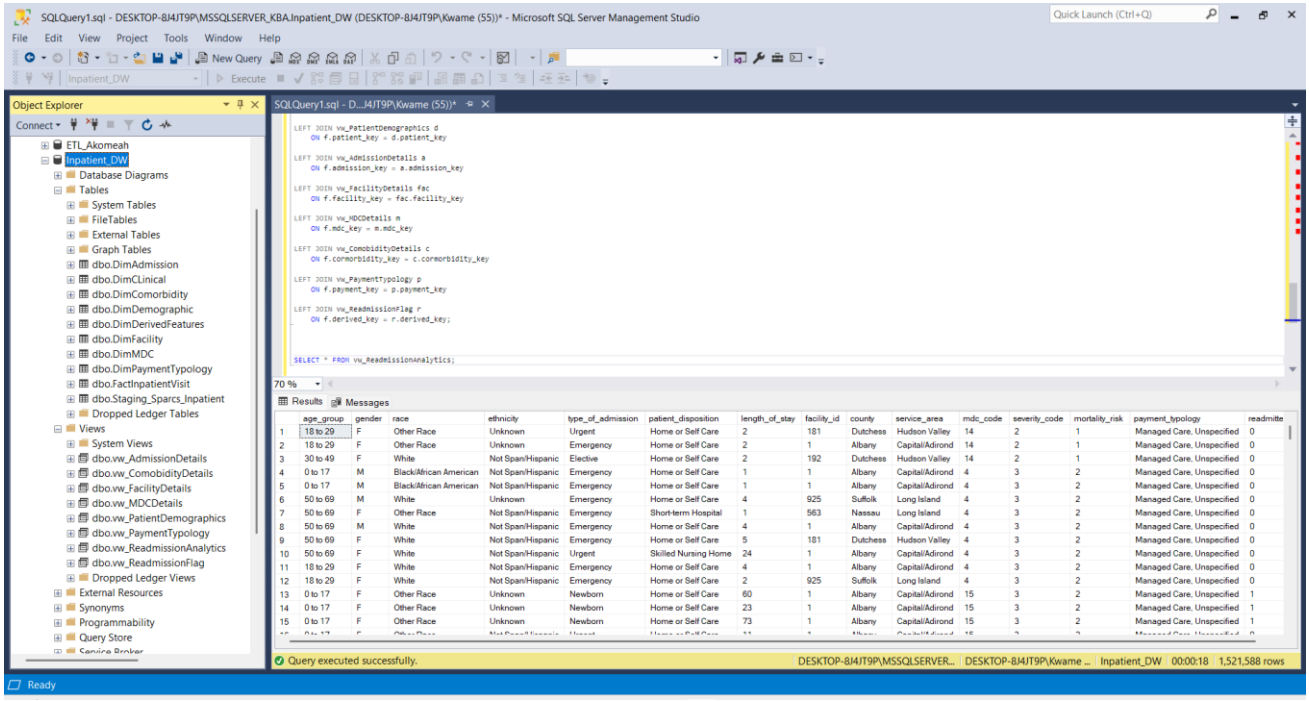
The Star Schema/ER Diagram shows the schema structure of the hospital readmissions data warehouse. Each dimension connects to the fact table in the center through surrogate keys, allowing for efficient analytical queries across the attributes. This star schema structure ensures data integrity and creation of reusable SQL views for business intelligence, reporting, and predictive modeling.

4.3 Database Management

In this section, SQL Server was utilized to implement consistent data types, views and aggregate queries. Views were implemented to improve query execution time, reduce analytical workload, and create reusable, standardized business definitions for readmission analytics.

To prepare the dataset for Business Intelligence and further analytics, a consolidated view *vw_ReadmissionAnalytics* was created, transforming the star-schema structure into an analytics-ready table made of relevant dimensions. The final analytics view provides a unified dataset containing patient demographics, facility information, admission details, MDCs, comorbidity risk levels, and the readmission target variable.

Figure 12: Inpatient Analytics Views

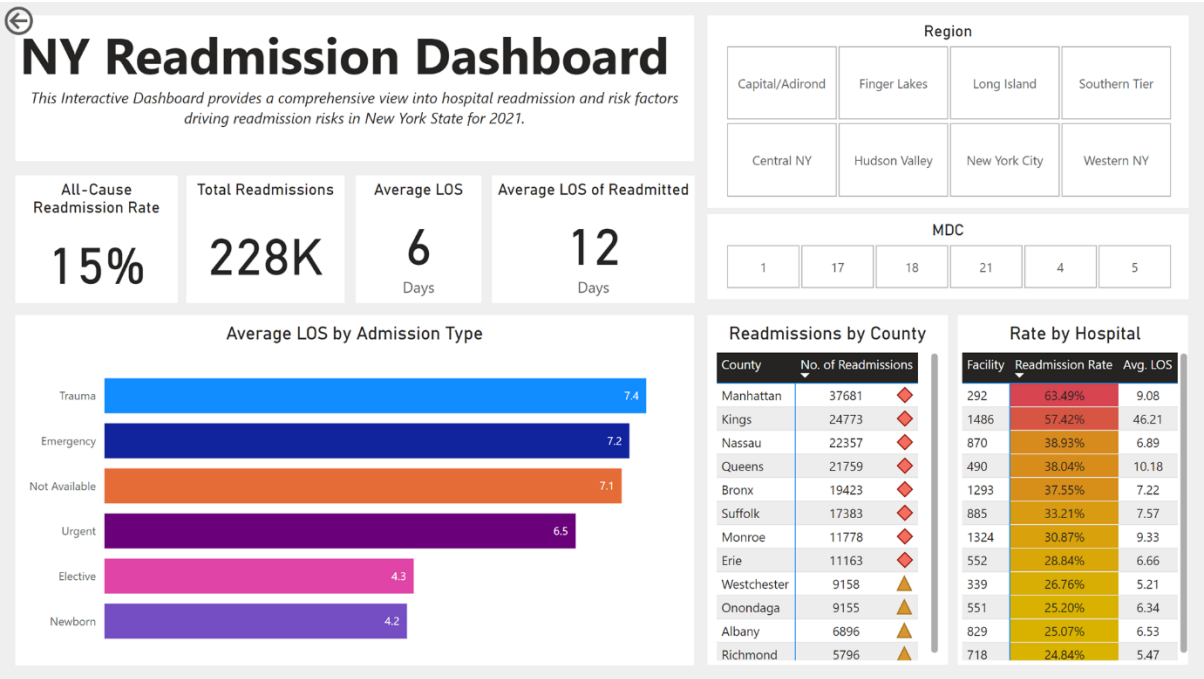


4.4 Business Intelligence and Data Analysis

4.4.1 Business Intelligence with Power BI

This Descriptive and Diagnostic analytics employs interactive Power BI dashboards to understand the nature and patterns of readmissions, while delving deeper into the key influencers driving readmission rates in the state of New York. This BI section transforms raw inpatient data into actionable insights by summarizing key metrics such as readmission rates, LOS, and clinical risk factors, allowing stakeholders to explore patterns across regions, facilities, MDC, and coverage. The use of Key Influencer and scenario analyses further support data-driven decision-making by showing the impact and forecast of top predictors of readmissions and how targeted interventions can reduce readmission rates and improve patient outcomes.

Figure 13: Executive Dashboard



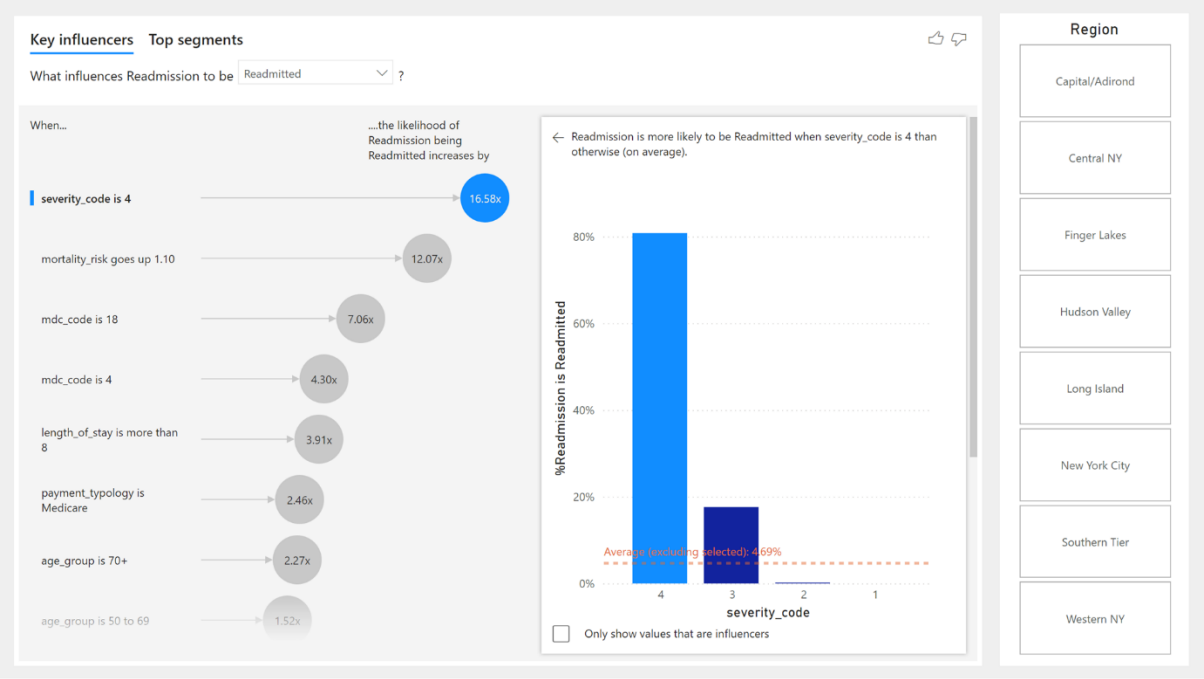
a. Executive Dashboard

The Executive Dashboard page provides a high-level operational view of hospital readmission in the state of New York, answering the question: How big is the readmission problem in the state of New York, where is it most prevalent, and which facilities and admission types are most affected? The top KPIs include:

- The All-cause 30-day readmission rate of 15%.
- The Total number of readmissions (228K).
- The Average LOS for all inpatient visits (6 days).
- The Average LOS for readmitted patients of 12 days which shows that readmitted patients stay about twice as long as the average admission.

- The Average LOS by Admission Type shows that Trauma and Emergency patients have the highest LOS, while Elective and Newborns have the shortest stay. This highlights that unscheduled inpatient visits are associated with higher hospital utilization.
- From the Readmissions by County table, Manhattan, Kings, Nassau, and Queens have the most readmissions in the state of New York and need targeted intervention and outreach programs.
- The Readmission Rates by Hospital table shows facilities struggling with readmissions, providing insights for benchmarking and performance improvement measures.

Figure 14: Top Drivers of Readmissions



b. Key Influencers

The Top Risk Drivers page provides insights into the major predictors driving readmission risk. This diagnostic analytics ensures that healthcare providers understand why high-risk patients were flagged, which factors most strongly drive readmissions, and where to target interventions. The Key Influencers visual answers the question: What factors influence patients to be Readmitted?

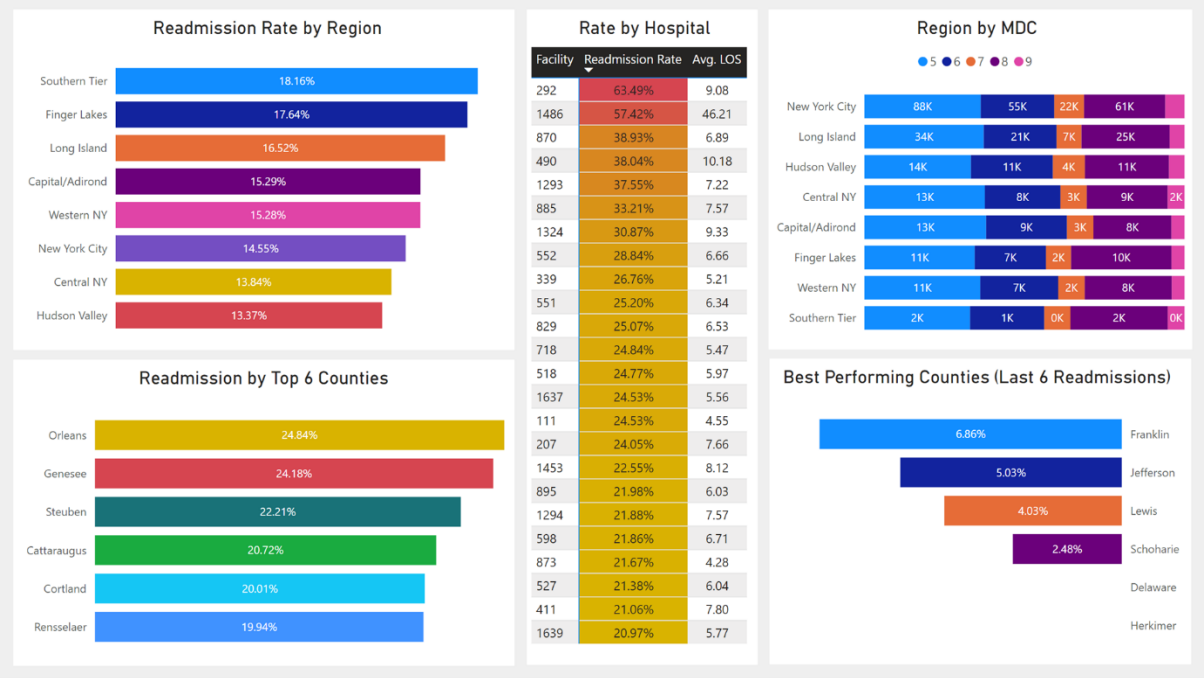
- Listing the most contributing factors of readmission, the visual shows that Extreme severity (4) cases are the most leading predictor for readmissions, validating ***H1***.
- This is followed by Major (3) and Extreme (4) mortality risk groups, (infectious & parasitic diseases (18), and respiratory system diseases (4)) MDC, LOS longer than the average 6 days, Medicare coverage, as well as elderly age groups (70+ and 50 to 69), validating ***H1, H2,*** and ***H3***.
- The bar chart highlights the impact of high acuity such that the readmission rate of extreme severity cases is significantly higher than the average rate (red dotted line).

c. Service Area

The Geographic page provides insights into readmission rates across service areas(regions), counties, and facilities, suggesting where quality and outreach efforts may need more focus.

- Readmission Rate by Region shows that Southern Tier and Finger Lakes have the highest rates, while Central NY and Hudson Valley have the lowest rates, lower than the state and national rates.
- The facility heatmap table lists hospitals by their readmission performance (readmission rate and average LOS), aiding in targeting performance improvements.
- Readmission by Top 6 Counties shows that Orleans and Genesee are among the counties contributing the highest readmission rates and driving the state rate.

Figure 15: Geographic Variations of Readmissions



- The Best Performing Counties on the other hand show the least readmissions, with Herkimer and Delaware having no readmissions. This is useful for performance benchmarking and emulating best practices.
- Region by MDC shows the volume of readmissions by Major Diagnostic Category within each region. It highlights that circulatory system diseases (5) and musculoskeletal system and connective tissue diseases (8) are the key MDCs driving regional differences.

d. Demographics

The Demographics page provides insights into readmission rates across various patient demographics such as payment typology, age groups, gender, ethnicity and race, highlighting

social and demographic groups at higher risk of readmission. It answers the question: Who is being readmitted and under which insurance coverage?

- Readmission Rate by Payment Typology shows that Medicare has the highest rates, validating ***H2*** and suggesting elderly public insurance holders may need more disease management and special insurance programs.
- Readmission Rate by Age Group also emphasized that elderly age groups (70+ and 50 to 69) are at the highest risk of readmission.
- Readmission Rate by Gender show a slight difference between rates of males (18%) and females (13%).
- Readmission Rate by Ethnicity and Race show that Non-Hispanics and Whites are at the highest risk for readmission while multi-ethic and multi-racial patients recorded the lowest risk.

Figure 16: Payment Typology and Patient Demographics

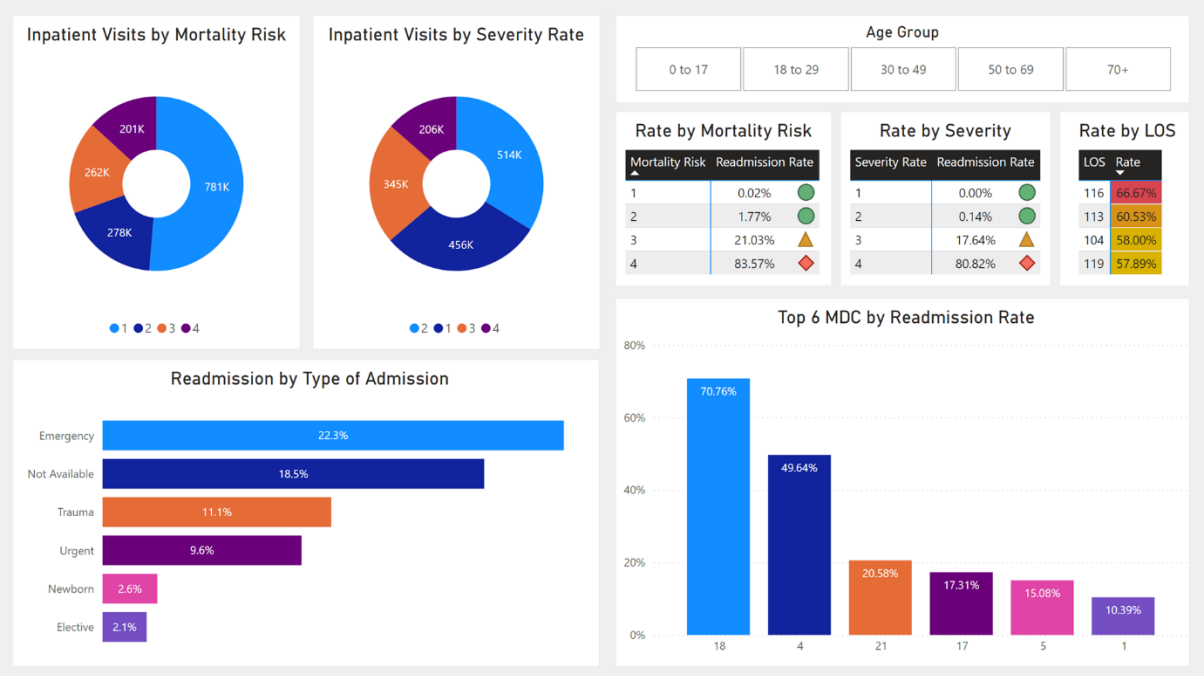
e. Risk Factors

The Clinical page provides insights into how clinical risk factors such as mortality risk, severity, MDC and admission type influence readmission rates.

- Inpatient Visits by Mortality Risk shows the distribution of admissions across the mortality risk categories (minor (1), moderate (2), major (3), extreme (4)). It highlights that most of the patients were admitted with minor cases.
- Inpatient Visits by Severity Rate also show that majority of patients were admitted for less complex cases (moderate (2) and minor (1)).

- However, the Rate by Mortality Risk and Severity tables show that patients with extreme (4) cases were at significantly higher risk of readmission while less complex cases had almost no risk of readmission.
- The Rate by LOS also suggests that patients who stayed extremely long in hospitals were at the highest risk of being readmitted.
- The Top MDC by Readmission rate shows that patients with infectious and parasitic diseases (18) have the highest risk of being admitted. Other top 5 MDC at risk of readmission include respiratory system diseases (4), injuries, poisonings and toxic effects of drugs (21), myeloproliferative diseases and poorly differentiated neoplasm (17), circulatory system diseases (5), and nervous system diseases (1). This analysis validates ***H3***, suggesting which major diagnostic areas or chronic conditions are most problematic and candidates for targeted programs.

Figure 17: Clinical Risk Factors



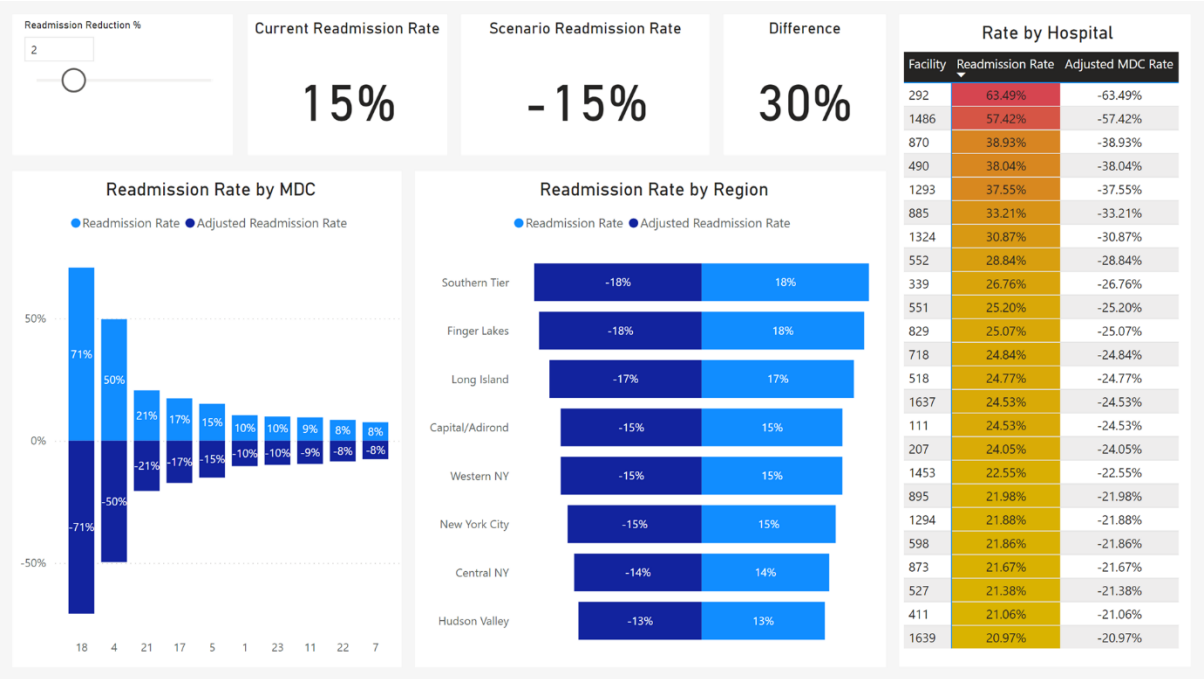
- Readmission by Type of Admission highlights that inpatient visits that are emergencies are at high risk of being readmitted, with Elective cases having the least risk of readmission. This indicates that transitional care and chronic disease management improvements may have a big impact on readmission rates.

f. Scenario Analysis

The Scenario page provides insights into how much readmissions could be reduced by improving care for MDC. It answers the questions: What if we improve MDC care and planning and cut their readmissions by 15%? How would this improvement be distributed across facilities, regions, and diagnosis?

- Scenario Readmission Rate highlights the adjusted readmission rate in MDC, reflecting the potential gain from the intervention (30% reduction compared with the baseline rate).
- Readmission Rate by MDC shows that the top 6 MDC with high readmission rates would see the largest drops in readmissions after this improvement.
- Readmission Rate by Region also shows how each service area would benefit from this scenario.
- Rate by Hospital also shows which hospitals improve greatly or still remain high risk, supporting targeted intervention planning.

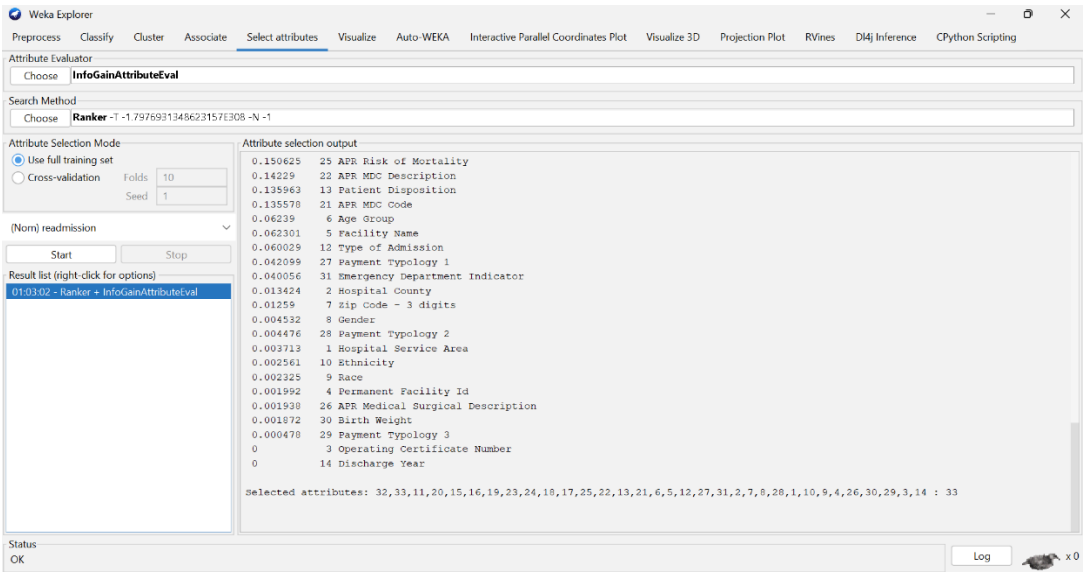
Figure 18: Scenario/What-if Analysis



4.4.2 Data Analysis with Weka

This Predictive analytics section employs Weka for data mining and predictive modeling to identify the key drivers of 30-day hospital readmissions. The analysis only used about 11,000 rows of the inpatient data, a constraint by Weka on handling large datasets. Findings from attribute selection techniques are consistent with ***H1*** and ***H3,*** highlighting Mortality Risk, Patient Disposition, and MDC as the top predictive features for 30-day readmission.

Figure 19: Attribute Selection



Extreme mortality risk, together with other clinical features (major and moderate mortality risk, MDC) emerged as the strongest predictor for readmission in the Logistic Regression. Employing Decision Tree, decision paths showed that higher Severity, higher Mortality Risk, and Emergency Admissions increased the likelihood of readmission, achieving an accuracy of 98.77%, precision (1) of 75%, recall (1) of 64%, and ROC-AUC of 0.90.

Figure 20: J48 Decision Tree

```
|   |   |   |   |   Race = Black/African American
|   |   |   |   |   |   APR MDC Code <= 23: 0 (4.0)
|   |   |   |   |   |   APR MDC Code > 23: 1 (2.0)
|   |   |   |   |   Race = Multi-racial: 1 (1.0)

Number of Leaves  :     64

Size of the tree :      92


Time taken to build model: 0.16 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       10922               98.7701 %
Incorrectly Classified Instances       136                1.2299 %
Kappa statistic                          0.6861
Mean absolute error                      0.0161
Root mean squared error                  0.1009
Relative absolute error                 38.2494 %
Root relative squared error             69.5137 %
Total Number of Instances            11058

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.995    0.357    0.992      0.995    0.994      0.688    0.896     0.994     0
                 0.643    0.005    0.750      0.643    0.692      0.688    0.896     0.655     1
Weighted Avg.    0.988    0.350    0.987      0.988    0.987      0.688    0.896     0.986

=== Confusion Matrix ===

     a     b   <-- classified as
 10769    51 |     a = 0
    85   153 |     b = 1
```

Random Forest on the other hand emerged as the most robust model, with an accuracy of 98.90%, precision (1) of 80%, recall (1) of 66%, and ROC-AUC of 0.990.

Figure 21: Random Forest

```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 1.26 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       10937               98.9058 %
Incorrectly Classified Instances       121                1.0942 %
Kappa statistic                          0.7163
Mean absolute error                      0.0166
Root mean squared error                  0.088
Relative absolute error                 39.2489 %
Root relative squared error             60.6434 %
Total Number of Instances            11058

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.996     0.340     0.993      0.996    0.994      0.720    0.990     1.000     0
               0.660     0.004     0.797      0.660    0.722      0.720    0.990     0.828     1
Weighted Avg.  0.989     0.333     0.988      0.989    0.989      0.720    0.990     0.996

=== Confusion Matrix ===

     a     b   <-- classified as
 10780    40 |     a = 0
    81   157 |     b = 1
```

Overall, machine learning algorithms employed confirmed that Severity and Mortality Risk strongly predict 30-day hospital readmissions, while emergency admissions and MDC are also associated with high readmission risk. Despite the constraint of Weka to use only 10,000 records out of the full 1.5 million in the dataset, the data analysis demonstrates that Random Forest outperforms other models, highlights clinical factors as primary readmission drivers, and provides a solid foundation for advanced data mining and predictive modeling with advanced tools such as Python.

4.5 Visualizing Information

This section emphasizes how effective data visualization and storytelling principles serve as important tools for reporting, turning complex raw inpatient data into clear and actionable insight for healthcare systems and stakeholders.

This project employs visual design elements that combine clear narrative structure with charts, that is, using hierarchy, color, size, and layout to present what matters most to audience. Each metric is measured and presented with appropriate charts, for instance bar charts for comparisons across service areas or demographics, and KPI cards for headline indicators. The interactive nature of dashboards seeks to balance user engagement and objectivity, employing filtering and navigation capabilities, consistent color schemes, and intuitive, domain-focused visuals tailored to clinical, operational, and leadership stakeholders in healthcare for data-driven decision-making.

4.5.1 High level monitoring with KPI cards

KPI cards used in BI provides concise summary of key metrics such as all-cause 30-day readmission rate, total number of readmissions, and average length of stay. At a glance, these headline indicators give stakeholders insights on overall operational performance, as well as providing context for the subsequent dashboards.

4.5.2 Uncovering patterns with tables, bar and donut charts

Bar graphs (horizontal and vertical) and donut charts used in exploration analysis and BI enable the display of various trends and relationships between readmission rate and categorical features. Leveraging bars to compare features allows end users to easily identify patterns such as high-risk factors, underperforming or best performing facilities and areas, or disparities in demographics and payment typology. Similarly, tables along with conditional formatting (icons and heatmap) are highly effective at identifying hotspots and benchmark facilities, counties, and admission types. This aids users to uncover hurdles and tailor interventions towards quality improvement.

4.5.3 Exploring patterns with navigations and filters

Slicers (Regional, MDC, Age Group, and Readmission Reduction %) allow end users to filter dashboards in real time. In the scenario analysis, the Readmission Reduction % (what-if) slider turns the dashboard from a descriptive tool into a forecasting and planning tool for operational and strategic decision-making. Additionally, the drill down, up, and through capabilities in dashboards allows users to explore patterns from both granular and macro point of view, providing insights into statewide performance and supporting intervention planning.

4.5.4 Examining distributions with boxplots

Boxplots are essential visualizations that provide graphical representation of the distribution, central tendency, and spread of data. It summarizes key descriptive characteristics such as minimum, interquartile range (first, second, third quartiles), median, maximum, and outliers present in dataset. Leveraging boxplots, end users can easily understand the distribution of each payment typology in relation to the combined score, enabling evidence-based strategies such as special coverage programs for publicly insured patients.

4.5.5 Explaining drivers with Key Influencers visuals

The Key Influencers visual is an AI powered visual that shows factors that most strongly impact readmission likelihood. This diagnostic analytics tool aids clinicians and stakeholders in understanding why high-risk patients were flagged and which targeted interventions will have the most impact.

4.5.6 Maintaining clarity with color schemes

Color is an essential part of visualization that ensures clarity and quickly communicates differences and relevant information to end users. Using color schemes, palettes, and hues in exploration analysis and BI helps stakeholders to easily distinguish between risk levels, compare categories, and spot anomalies. Conditional formatting using hues guides the attention of audience to hotspots such as high-risk mortality, severity, LOS, counties, and facilities that need intervention. Color palettes are also useful for differentiating between categories such as age group, admission type, and payment typology.

Overall, these visualization and storytelling elements help paint a complete picture of the state of New York's hospital readmissions from facility, demographic, operational, and clinical perspectives. It allows stakeholders to drill down on problem areas to understand risk factors and simulate the effects of targeted interventions. This dashboard serves as a tool to support evidence-based decisions aimed at reducing avoidable 30-day readmissions and improving patient outcomes.

4.6 Data Mining for Business Decisions

This Predictive analytics employs machine learning techniques to develop predictive models for early detection of patients at high risk of 30-day hospital readmission. Three classification algorithms were implemented, namely, Logistic Regression, Random Forest, and Extreme Gradient Boost (XGBoost). These models were employed based on their increasing levels of model robustness and evaluated using performance metrics, confusion matrices, feature importance plots, and cross-validated performance.

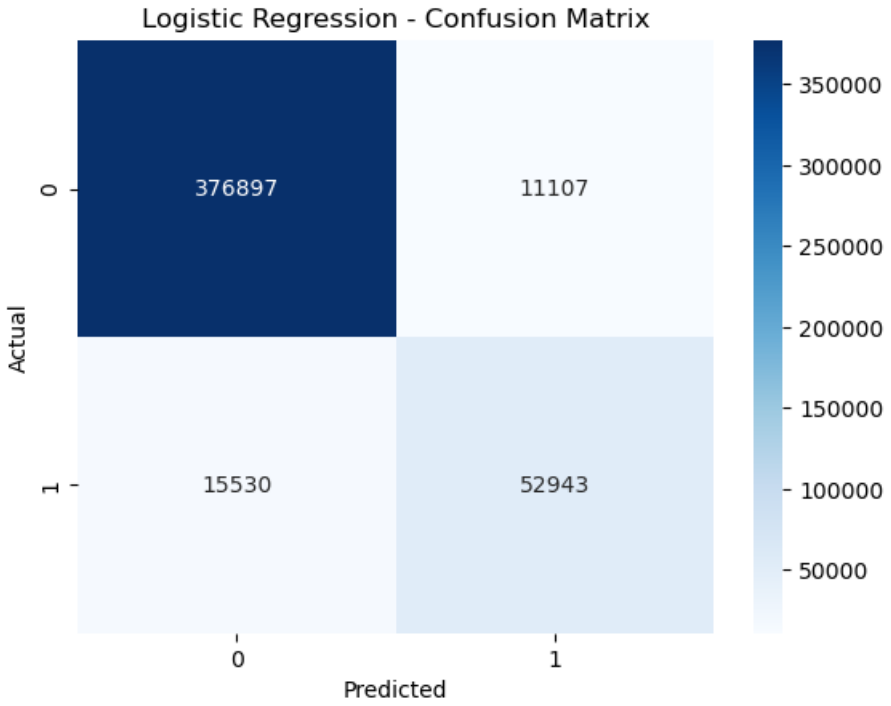
4.6.1 Logistic Regression (Baseline)

Logistic Regression was selected as the baseline model due to its interpretability and historical prominence in clinical risk prediction modeling. The model achieved:

- **Accuracy:** 91.8%
- **ROC-AUC:** 0.955
- **Precision (Class 1):** 61%
- **Recall (Class 1):** 71%
- **F1-score (Class 1):** 0.80

Logistic Regression model was effective at capturing a large majority of true readmissions (52,946); however, it was expected to underperform compared to tree-based models due to its ineffectiveness for multi-dimensional data and non-linear relationships.

Figure 22: Confusion Matrix of Logistic Regression



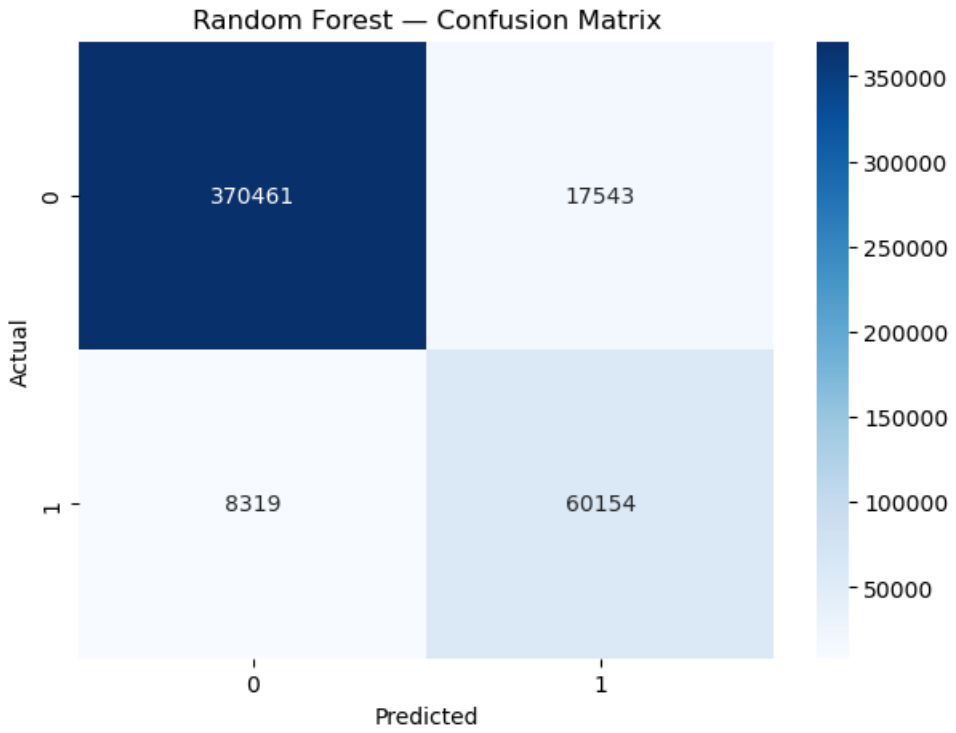
4.6.2 Random Forest (Choice Model)

Ensemble tree learning algorithms are highly effective at handling non-linear relationships and multi-dimensional data such as this dataset. Random Forest improved the predictive capability significantly as shown by its performance metrics:

- **Accuracy:** 94.0%
- **ROC-AUC:** 0.982
- **Precision (Class 1):** 77%
- **Recall (Class 1):** 88%
- **F1-score (Class 1):** 0.82

Random Forest demonstrated more robust discrimination (ROC-AUC: 0.982), highest sensitivity in predicting true readmissions (60,154), and lowest False Negatives (8319) as evident in its highest Recall score, however at the expense of highest False Positives (17,543). Random Forest is the chosen model for deployment since our operational priority is to flag as many risks of readmissions as possible, and missing a readmission (FN) is a greater risk compared to resource management. K-Fold cross-validated showed no to minimal variance (mean AUC = 0.9824), suggesting strong generalizability and robustness of the model.

Figure 23: Confusion Matrix of Random Forest



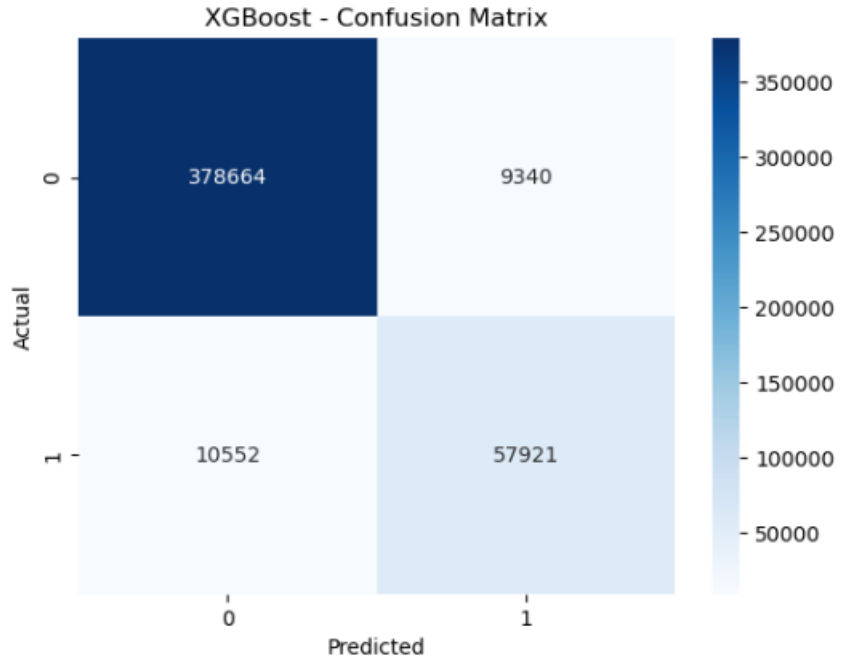
4.6.3 Gradient Boosting, XGBoost (Best Performance)

XGBoost algorithm showed the strongest predictive capability, outperforming the Random Forest model. Gradient boosting builds multiple decision trees sequentially, optimizing errors, and allowing the model to capture more complex relationships and subtle risk patterns automatically. It achieved the highest accuracy, along with other performance metrics across all the models:

- **Accuracy:** 95.64%
- **ROC-AUC:** 0.989
- **Precision (Class 1):** 86%
- **Recall (Class 1):** 85%
- **F1-score (Class 1):** 0.85

XGBoost demonstrated the strongest reduction in False Positives (9,340), representing a reduction in false alarms and an efficient model for resource management. On the other hand, False Negatives (10,554) were slightly higher compared to Random Forest (8,319). This is particularly significant in this framework as missing high-risk patients (FN) can lead to adverse outcomes. The model is highly efficient at ranking patients by their various readmission levels, as evident in its highest ROC-AUC of 0.989. K-Fold cross-validation also showed no to minimal variance (mean AUC = 0.9889) further validating the model's robustness and stability.

Figure 24: Confusion Matrix of XGBoost



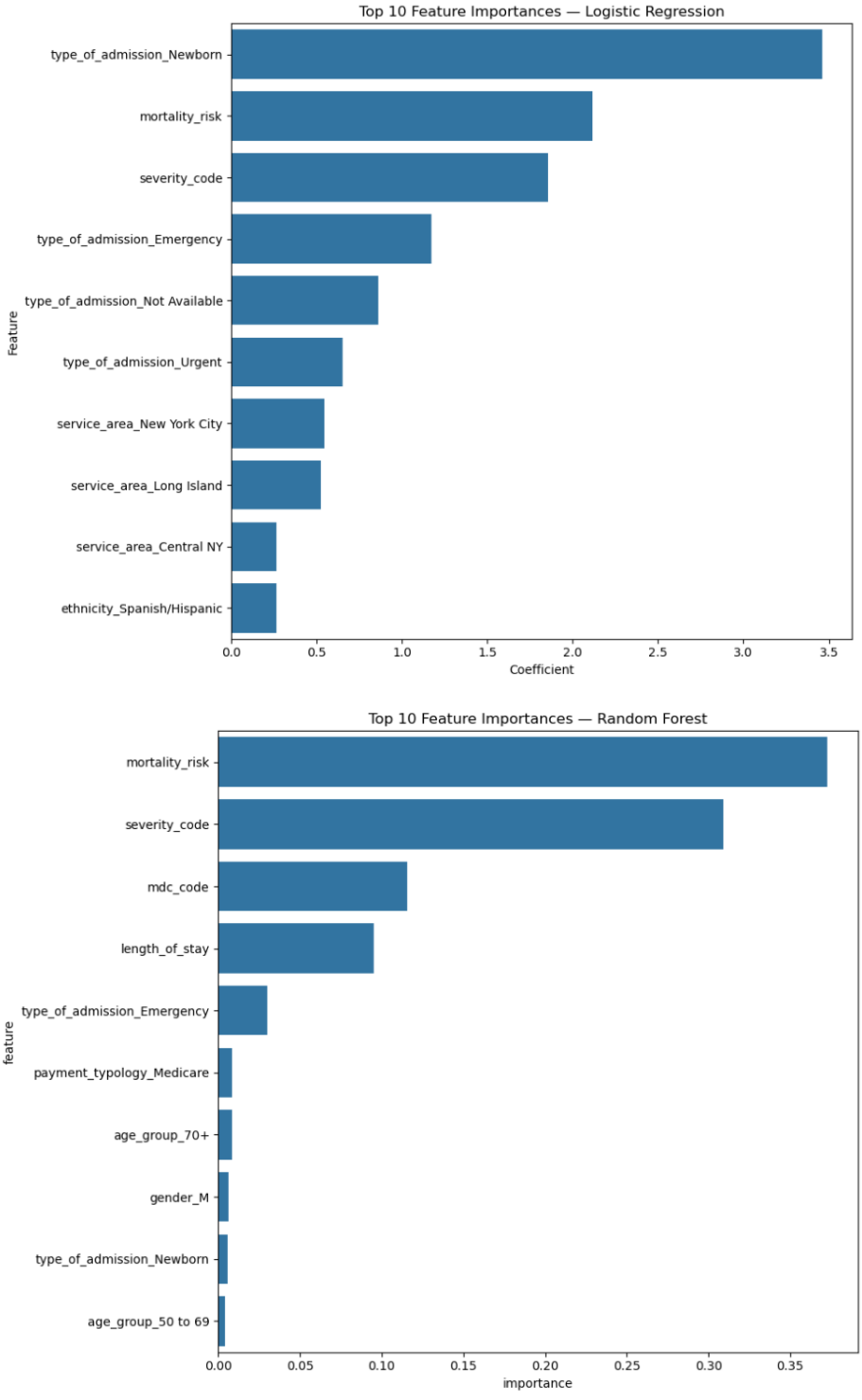
4.6.4 Feature Importance Analysis

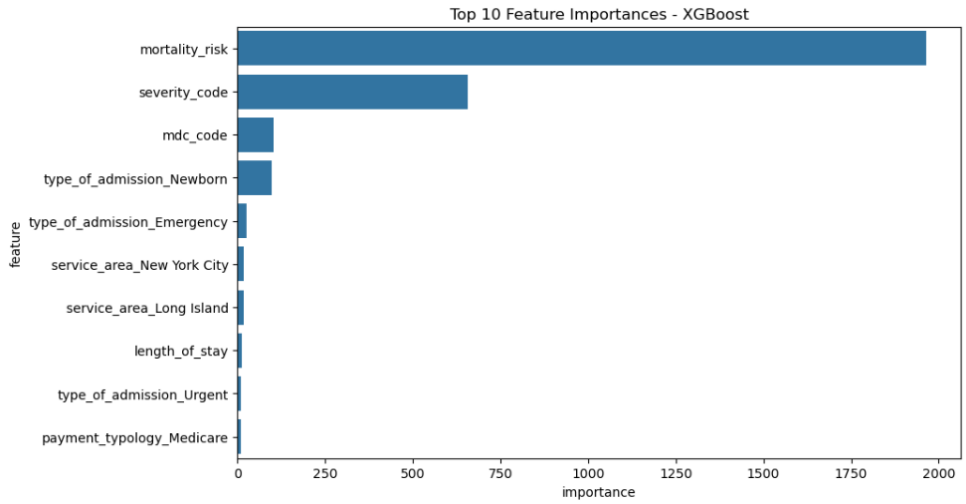
The feature importance analysis strongly validates ***H1*** and ***H3*** that clinical factors such as mortality risk, severity rate, and MDC, are strong predictors of hospital readmission. This aligns with CMS and empirical readmission risk factors. Features such as LOS, admission types (Emergency and Newborn), demographics, and facility attributes contribute relatively minimal risk to hospital readmission.

Payment typology (Medicare) emerged as a predictor (although relatively minimal) across Random Forest and XGBoost, validating ***H2*** that patients with public health insurance, mainly Medicare, are at risk of 30-day hospital readmission. Age groups (70+ and 50 to 69) also showed effects in Random Forest. This is relevant to note as Medicare is a public insurance program for people 65 and older, also validating ***H2***.

Overall, the implemented models showed a clear improvement in predictive capabilities with Random Forest emerging as the best suited model for predicting 30-day hospital readmission risk. Although it flags more false alarm high-risk patients than necessary, this is acceptable as these patients would benefit from preventive interventions, improving their health outcomes. The high recall and ROC-AUC capabilities to effectively identify high-risk readmission patients makes the model a valuable tool for hospitals, healthcare providers, and health systems to reduce readmissions, optimize care coordination, and improve overall patient outcomes.

Figure 25: Feature importance plots for Logistic Regression, Random Forest, and XGBoost





4.7 Introduction to Healthcare Informatics

The use of advanced analytical techniques to predict and reduce hospital readmissions has become increasingly relevant. This project leveraged core Health Informatics (HI) principles to manage health data for efficient support of clinical decision-making and overall improvement in health outcomes. This facilitated aligning data analytics and machine techniques with established standards for data integrity, privacy, security, and fairness.

In conformity with Health Insurance Portability and Accountability Act (HIPAA) of protecting sensitive Protected Health Information, the data source encrypted patient identifier information, however, there remain PHI such as age group, ZIP code, race, ethnicity, facility identifiers, and detailed diagnosis and procedure information.

Data privacy and HIPAA rules applied:

- The analysis followed the minimum necessary rule by using only fields required for predictive modeling, de-identifying facility and diagnosis identifiers.
- Direct identifiers (names, phone numbers, full addresses, medical record numbers, and other HIPAA identifiers) were encrypted from source data and hence, absent from dataset employed.
- Under the HIPAA Privacy Rule, readmission status, severity, mortality risk, and length of stay are all treated as PHI because they describe an individual's health status and utilization.
- In line with the Security Rule, role-based permissions and secure connections were used to protect PHI.

Fairness and bias in predictive modeling:

- Because the dataset includes age group, race, ethnicity, payment typology, severity, and ZIP code, there is a real risk that predictive models could encode or amplify existing disparities rather than just reflect clinical need. Older adults, certain racial or ethnic groups, and publicly insured patients may exhibit higher observed readmission rates due to social determinants of health and access barriers, not because they are inherently higher risk.
- To address this, the modeling workflow emphasized fairness-aware practices: weights were assigned in deriving targeting variable, protected attributes were monitored during modeling; feature-importance methods were used to understand how variables such as severity, mortality risk, and payment typology influenced predictions. Where an attribute was not clinically necessary for prediction, its role in the model was excluded to minimize potential bias.
- These practices reflect emerging HI guidance that machine learning tools in healthcare must be both accurate and equitable, supporting clinicians and administrators without unfairly targeting demographic or socioeconomic groups.

Overall, the project demonstrates how Health Informatics concepts (HIPAA-aligned privacy safeguards, secure handling of PHI, and fairness-awareness) can be integrated into a hospital readmission analytics framework. By combining de-identified, minimally necessary data with interpretable models and bias checks, the work supports data-driven decision-making while respecting patient rights and promoting equitable care.

5. Conclusion and Recommendations

This Predictive Modeling for Hospital Readmission Reduction project successfully demonstrates how predictive analytics can address the most pressing challenges in healthcare, reducing preventable 30-day hospital readmissions. By integrating statistical methods, machine learning, business intelligence, data warehousing, as well as domain knowledge, this project provides a robust framework for hospitals to:

a. Identify high-risk patients early before discharge to develop targeted intervention programs,
b. Understand which factors contribute most to readmissions, and
c. Enhance performance monitoring using Business Intelligence dashboards and reports.

Logistic Regression serving as the baseline model, provided strong discrimination (ROC-AUC 0.955) and good recall, confirming its continued value for interpretable clinical risk scoring. However, Random Forest algorithm delivered stronger performance, achieving higher accuracy (94.0%), higher precision (0.77), robust discrimination (ROC-AUC 0.982), and highest recall (0.88), correctly identifying the largest number of true readmissions and the least false negatives. XGBoost further improved overall accuracy and ROC-AUC (95.64%, 0.989) and reduced false positives, showing superior efficiency in ranking patients by readmission risk. Feature importance and key influencer analyses validated ***H1, H2,*** and ***H3*** by consistently emphasizing severity, mortality risk, and MDC as the dominant predictors, with elderly age groups and Medicare coverage contributing significantly as well.

The following are key insights and recommendations gathered from this framework to ensure effective implementation of predictive modeling and successful reduction of 30-day avoidable hospital readmissions.

1. **Model deployment and governance:**

   - Deploy Random Forest as the operational model into clinical workflow to provide real-time risk scores and flag high risk patient before or at the point of discharge. XGBoost model can also be used together with Random Forest for resource management and hospital utilization initiatives, due to its robustness in reducing false positives or false alarm cases.
   - Leverage real-time data to implement ongoing model and framework monitoring and optimization, tracking recall and precision over time. Dashboards can also be extended to include facility level intervention tracking, measuring the impact of interventions across the state.

2. **Clinical and operational interventions:**

   - Tailor Transitional Care Management (TCM) services to high-risk patients with high severity and mortality risks, emergency cases, elderly and Medicare patients, as well as patients with chronic conditions (respiratory, circulatory, etc.). These TCM

services may include follow-up visits/calls, care coordination, medication reconciliation, patient education, and public health initiatives.

- Strengthen discharge planning and post-discharge processes for high-risk patients, such as medication reconciliation and clear instructions for follow-ups and medications.
- Target quality care improvement initiatives to facility level and service areas where readmission rates are highest, using the best performing facilities or regions as performance benchmark.
- Collaborate with case management, Medicare and Medicaid partners to implement care coverage programs for the elderly and publicly insured patients at risk of acuity and readmissions.

The findings of this end-to-end analytics framework support CMS clinical risk factors and studies, as well as literature on employing machine learning, complex analytics, and multi-dimensional data to predict and reduce 30-day hospital readmission.

6. Limitations

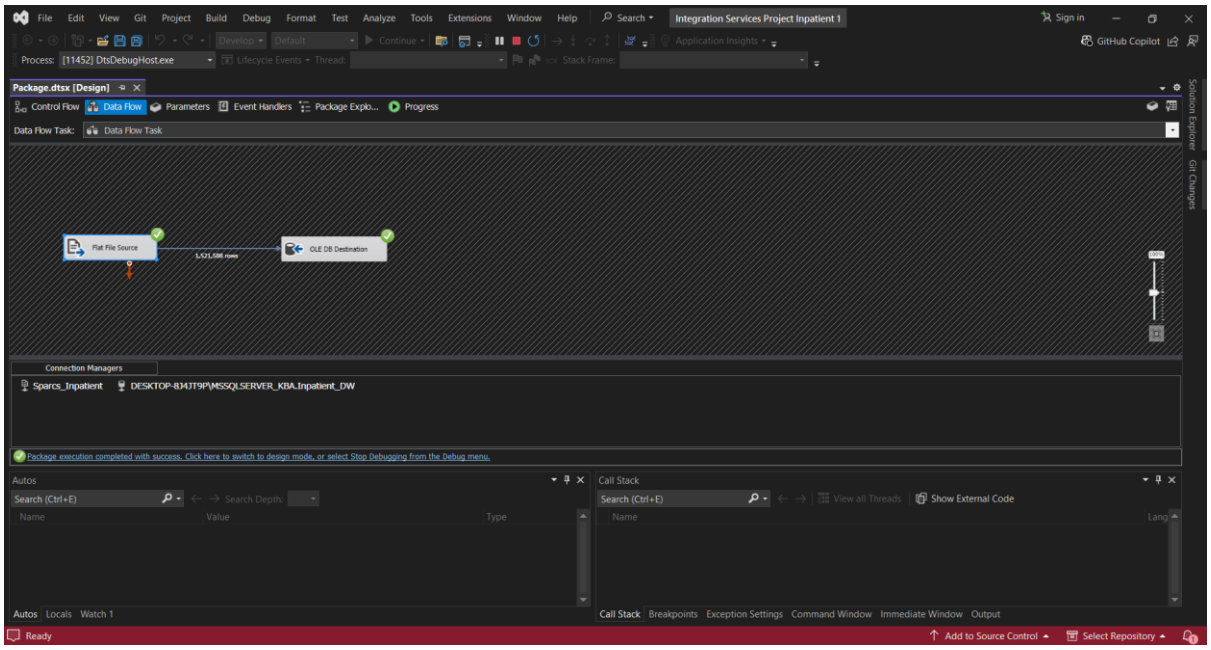
Constraints encountered with the dataset weakened how comprehensively and confidently the models and findings can be used to guide clinical decisions. The dataset does not include key features that enable true 30-day readmission measurement, comprehensive inpatient admission and readmission trends and dynamics. The limitations faced with this project include:

a. Lack of rich clinical variables such as lab results, imaging, medication adherence, and history of diagnosis and medication which present the full medical picture of patients.
b. Lacked patient specific longitudinal features such as patient ID, admission history, and admission/discharge dates which limit ability to verify true 30-day readmissions and contextualize patient dynamics.
c. Lack of readmission target variable necessitated deriving target variable with available features. This approach is prone to multicollinearity, information leakage, and omission of variable bias.
d. Models were trained on historical, not real-time data streams which may affect performance once integrated with real-time data or workflow since case mix and hospital processes change over time.
e. Weka only allowed for 11,000 records of data which limited efforts to validate the predictive capabilities of the models using two different tools and techniques.

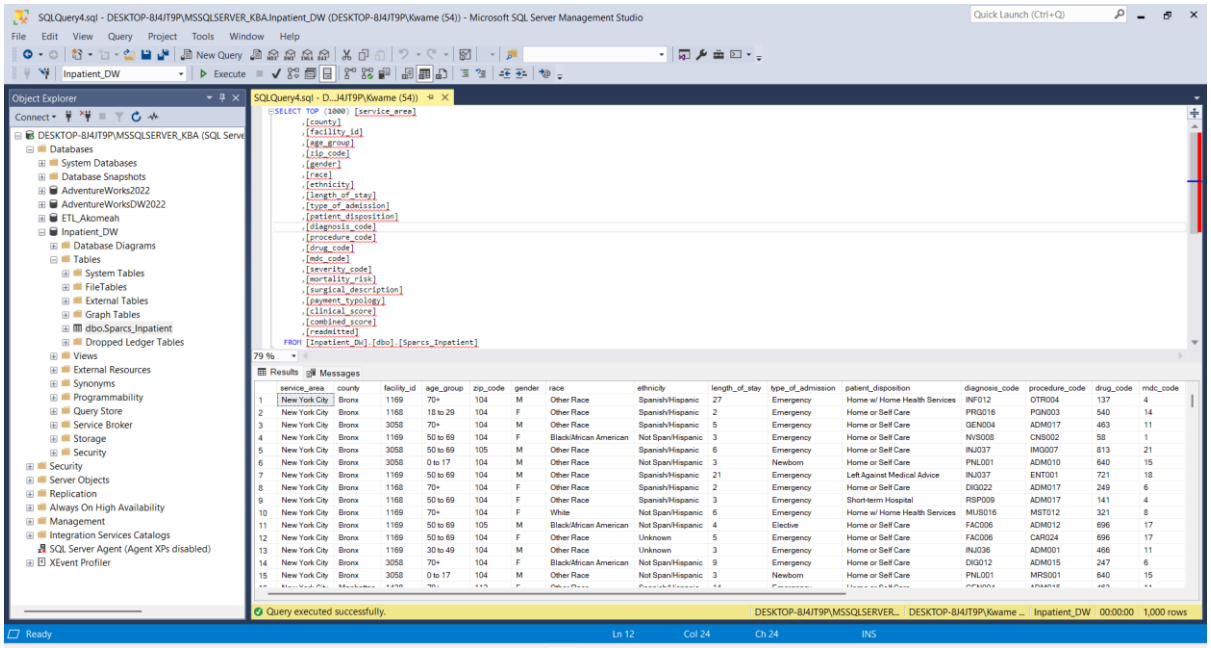
Despite these limitations, the project demonstrates that a robust end to end analytics and modeling framework, supported by comprehensive data, machine learning techniques, BI dashboards, prospective validation, and governance, can provide a powerful foundation for data-driven generalization and strategies to reduce 30-day hospital readmissions and improve patient outcomes.

7. Appendix

Appendix 1: DW: ETL Pipeline (SSIS) screenshots



Appendix 2: Data Warehousing - Staging Table



8. References:

1. Bradley, E. H., Curry, L., Horwitz, L. I., Sipsma, H., Wang, Y., Walsh, M. N., Goldmann, D., White, N., Piña, I. L., & Krumholz, H. M. (2013). Hospital strategies associated with 30-day readmission rates for patients with heart failure. *Circulation. Cardiovascular quality and outcomes*, *6*(4), 444–450. https://doi.org/10.1161/CIRCOUTCOMES.111.000101

2. CareSet AI Researcher. (2025, September 8). *Understanding average hospital length of stay: Key insights for managers*. CareSet. Retrieved [date you accessed it], from https://careset.com/understanding-average-hospital-length-of-stay-key-insights-for-managers/

3. Centers for Medicare & Medicaid Services (CMS). (2025). *Hospital Readmissions Reduction Program (HRRP)*. https://www.cms.gov/medicare/payment/prospective-payment-systems/acute-inpatient-pps/hospital-readmissions-reduction-program-hrrp

4. Chollet, D., Barrett, A., & Lake, T. (2011). *Reducing hospital readmissions in New York State: A simulation analysis of alternative payment incentives*. Mathematica Policy Research. New York Health Foundation. https://nyhealthfoundation.org/resource/reducing-hospital-readmissions-in-new-york-state-simulation-analysis/

5. Definitive Healthcare. (2025). *Average hospital readmission rate by state*. Definitive Healthcare. Retrieved (November 7, 2025), from https://www.definitivehc.com/resources/healthcare-insights/average-hospital-readmission-state

6. Horwitz, L. I., Partovian, C., Lin, Z., Grady, J. N., Herrin, J., Conover, M., Montague, J., Dillaway, C., Bartczak, K., Suter, L. G., Ross, J. S., Bernheim, S. M., Krumholz, H. M., & Drye, E. E. (2014). Development and use of an administrative claims measure for profiling hospital-wide performance on 30-day unplanned readmission. *Annals of internal medicine*, *161*(10 Suppl), S66–S75. https://doi.org/10.7326/M13-3000

7. Jencks, S. F., Williams, M. V., & Coleman, E. A. (2009). Rehospitalizations among patients in the Medicare fee-for-service program. *The New England journal of medicine*, *360*(14), 1418–1428. https://doi.org/10.1056/NEJMsa0803563

8. Jiang, H. J., & Hensche, M. K. (2023). *Characteristics of 30-day all-cause hospital readmissions, 2016–2020* (HCUP Statistical Brief #304). Agency for Healthcare Research and Quality. https://hcup-us.ahrq.gov/reports/statbriefs/sb304-readmissions-2016-2020.jsp

9. Laurent, A. (n.d.). *Hospital readmission rates by state: US data & analysis*. IntuitionLabs. Retrieved (November 7, 2025), from https://intuitionlabs.ai/articles/hospital-readmission-rates-by-state