**Summary of Findings on Patient Readmissions updated**

1. **Key Statistics & Distributions**

* **Age Distribution:** Uniform across the dataset, with no strong trends linked to readmission. Average Age (Train): 53.299 raw data statistic

Average Days in Hospital (Train): 7.3966 raw data statistic.

* **Comorbidity Score:** Distinct peaks suggest common levels of comorbidity among patients.
* **Readmission Rates:** Highly imbalanced dataset; only 18.8% of patients were readmitted.

**2. Correlations & Relationships Between Features**

* **Primary Diagnoses with Highest Readmission Rates:**
  1. Kidney Disease
  2. Diabetes
  3. COPD
  4. Heart Disease
  5. Hypertension
* **Feature Correlations:** No strong linear correlation between individual features and readmission.
* **Readmission vs. Other Features:** No clear trends in scatter plots, suggesting a non-linear relationship.
* **Clusters in Data:** Certain groups emerge based on days in hospital and comorbidity scores, which may indicate high-risk patient categories.

**3. Model Performance & Challenges**

* **Severe Class Imbalance:**
  + The model predicts almost all patients as non-readmitted, leading to high accuracy but poor recall for readmitted patients.
  + **First Logistic Regression Model:**
    - Accuracy: 82.6%
    - Completely failed to predict readmissions (Recall for Class 1 = 0.00)
* **Random Forest Model Performance (After SMOTE Resampling & Hyperparameter Tuning):**
  + Accuracy: 78.9%
  + Readmission Recall: 3% → Still very low, meaning most actual readmitted cases are missed.
* **Alternative Logistic Regression Model (With Class Balancing & SMOTE):**
  + Accuracy: 49.8%
  + Readmission Recall: 38% → Identifies more readmitted patients but at the cost of overall accuracy.

**4. Next Steps for Improvement**

1. **Further Improve Readmission Recall:**
   * Adjust SMOTE sampling strategy to avoid excessive synthetic data.
   * Tune Random Forest parameters (lower min\_samples\_split, adjust class\_weight).
2. **Feature Engineering for Better Predictions:**
   * Introduce interaction terms (e.g., age × comorbidity\_score).
   * Try non-linear transformations (e.g., polynomial features).
3. **Explore More Advanced Models:**
   * Gradient Boosting Models (XGBoost, LightGBM) → Handle imbalanced data better.
   * Neural Networks for capturing complex relationships.

**24.1 Findings:**

**Business Understanding of the Problem**

The objective of this analysis is to predict hospital readmissions within 30 days using a synthetic healthcare dataset. Hospital readmissions are costly and represent poor health outcomes for patients. Identifying patients at high risk of readmission allows healthcare providers to take preventive measures, reduce healthcare costs, and improve patient outcomes. By leveraging machine learning techniques, the aim is to predict the likelihood of readmission and uncover the key factors driving this risk.

**Data Cleaning and Preprocessing**

To ensure the accuracy and reliability of the models, the dataset underwent extensive cleaning and preprocessing, which included:

* Handling missing values.
* Converting categorical variables into appropriate formats (e.g., one-hot encoding for categorical features).
* Normalizing continuous variables to ensure consistent scale across the features.
* Addressing the class imbalance using techniques like oversampling to improve the model's ability to predict the minority class (readmitted).

**Descriptive and Inferential Statistics**

Descriptive statistics revealed that the dataset was highly imbalanced, with the majority of patients not being readmitted. The following key insights were derived:

* The average age of patients in the dataset was **53** years, with a higher proportion of older patients at risk for readmission.
* The comorbidity score, which accounts for additional health complications, showed that patients with higher scores had an increased likelihood of readmission.
* The average length of hospital stay was **7** days, with longer stays correlating with higher readmission risk.

Inferential statistics, such as correlation analysis and hypothesis testing, further reinforced these insights and highlighted significant relationships between patient characteristics and readmission likelihood.

**Key Findings**

After testing multiple machine learning models, the **Random Forest** model was identified as the most effective for predicting readmissions. Below are the key findings regarding feature importance:

1. **Strong Features for Predicting Readmission**:
   * **Age**: Older patients are at higher risk of readmission.
   * **Comorbidity Score**: Patients with more complex medical histories are more likely to be readmitted.
   * **Days in Hospital**: Patients with longer hospital stays tend to have more severe conditions and are more likely to be readmitted.
   * **Number of Procedures**: A higher number of procedures increases readmission likelihood, suggesting these patients may require more intensive treatment or recovery.
2. **SHAP Analysis**: In addition to feature importance, SHAP values provided deeper insights into the factors influencing the model's predictions. These features had a substantial impact on the model's output:
   * **Discharge to Rehabilitation Facility**: Patients discharged to rehabilitation facilities showed a higher likelihood of readmission due to complex care needs post-discharge.
   * **Kidney Disease**: Patients with kidney disease exhibited significantly higher readmission risks, highlighting the chronic nature and complications of the condition.
   * **Discharge to Home Healthcare**: Discharge to home healthcare was associated with higher readmission rates, emphasizing the need for intensive home care after discharge.
   * **Nursing Facility Discharge**: Similar to rehab facility discharges, patients transferred to nursing facilities were at an increased risk of readmission.
   * **Chronic Conditions (Diabetes, Hypertension, and Heart Disease)**: These conditions were strongly linked to higher readmission rates, stressing the need for effective management of chronic diseases to reduce hospital readmissions.

**Actionable Insights for the Healthcare provider:**

Based on the findings, the following actionable steps can be recommended for healthcare providers:

* **Targeted Post-Discharge Care**: Focus on patients who are older, have higher comorbidity scores, or have longer hospital stays for more intensive post-discharge monitoring and care.
* **Care Coordination for Chronic Conditions**: Prioritize patients with kidney disease, diabetes, hypertension, and heart disease for comprehensive follow-up care to manage these chronic conditions effectively.
* **Enhanced Discharge Planning**: Implement tailored discharge plans for patients discharged to rehabilitation facilities, home healthcare, or nursing facilities to ensure adequate care and minimize the risk of readmission.
* **Predictive Alerts**: Develop predictive systems using the identified features (age, comorbidity score, procedures, etc.) to alert healthcare teams about patients at high risk of readmission, enabling proactive interventions.

**Next Steps and Recommendations**

* **Model Refinement**: Experiment with additional techniques such as ensemble models or deep learning to further improve model performance.
* **More Data**: The model could benefit from incorporating more data points, including time-sequenced data such as patient history, to capture long-term trends in readmission risk.
* **Implementation in Real-World Settings**: Consider integrating the predictive model into a hospital's electronic health record system to provide real-time alerts to healthcare providers when patients are at high risk of readmission.

Key Data Visuals

