Profiling Diabetes Readmission Data:

An Exploratory Analysis

Janin Beck, Scott D. Hoover Jr., Brandon Lewis, & Britt Perez

Information System and Business Analytics

Park University

CIS 611: Introduction to Business Analytics

Phil Heppel-Kennard

September 7, 2025

Profiling Diabetes Readmission Data: An Exploratory Analysis

Hospital readmission is a persistent challenge in healthcare. It increases costs and exposes weaknesses in patient management. Diabetes contributes to this problem because complications often require additional hospital visits, and continuity of care shapes long-term outcomes for patients with this condition. A dataset sourced from Kaggle (2017), originally derived from Cerner Corporation's Health Facts database, provides a way to examine inpatient encounters involving individuals diagnosed with diabetes. It spans more than one hundred thousand admissions over a ten-year period across hospitals in the United States. It captures a variation in demographics, diagnoses, and patterns of hospital use. Each admission includes clinical details, utilization measures, and outcomes that record whether patients returned within thirty days, after thirty days, or not at all.

The dataset's breadth makes it valuable for identifying patterns in utilization and readmission, however, its limitations reduce its relevance for modern clinical decision-making. Weight is almost entirely missing, payer code is incomplete, and socioeconomic information is absent. Laboratory measures are reduced to broad categories that conceal meaningful variation. These gaps weaken its strength as a foundation for interventions but illustrate the challenges of working with clinical data outside of controlled research environments. The dataset can inform practice by showing how demographic and clinical factors interact with readmission risk, however, in this case, highlight the challenges of preparing messy clinical data for analysis. This project evaluates the dataset through data types, presence, quality, and character to determine its usefulness for exploring diabetes readmission and to identify its limits as a resource for rigorous statistical analysis.

Data Types

The dataset contains forty-eight variables that fall into three categories: numeric, categorical, and boolean. Numeric variables measure quantities such as age, time in hospital, number of laboratory procedures, and number of medications. Age ranges from 5 to 95 years, time in hospital spans 1 to 14 days, and laboratory procedures extend to 132, which supports descriptive and comparative analysis.

Categorical variables account for most fields. These variables describe qualities or classifications rather than measured amounts. Categorical variables in the dataset include demographic attributes such as race and gender, clinical results such as HbA1c and max_glu_serum, and diagnostic information. Three diagnostic variables, diag_1, diag_2, and diag_3, record primary and secondary conditions using coded values. Admission_type_id and admission_source_id are stored as integers but represent categories because the numeric codes map to categories such as emergency, elective, or referral. The way categorical data are structured effects analysis because statistical procedures vary depending on the type of variable. Hazra and Gogtay (2016) explain in an article on biostatistical methods, that when assumptions of normality do not hold, researchers often rely on nonparametric approaches or transform the data to ensure valid results. This reinforces the need to handle categorical and non-normally distributed variables in the dataset carefully, since applying inappropriate tests would distort findings.

A single boolean variable, diabetesMed, represents whether a patient received diabetes medication during the encounter. Boolean variables simplify analysis by storing only two possible values, such as yes/no, which allows straightforward contrasts between groups. In this

dataset, diabetesMed provides a clear way to separate patients who received treatment from those who did not.

The profiling results highlight variables that add no predictive value. Encounter_id is highly correlated with df_index, and patient_nbr operates as an identifier rather than a clinical feature. Several medication fields, including citoglipton and examide, show no variability because every record holds the same value. Excluding these variables avoids distortion and keeps the analysis focused on meaningful predictors.

Data Presence

Age, time in hospital, and the number of prior inpatient, outpatient, and emergency visits measure patient history and resource use. These fields show how heavily a patient has already interacted with the healthcare system, which strongly relates to the likelihood of readmission. Laboratory measures such as HbA1c and maximum glucose serum categories indicate glycemic control, an established driver of complications. Strack et al. (2014) demonstrated that patients who had HbA1c measured during admission showed lower readmission rates, yet the test was rarely ordered in hospital settings. The dataset contains HbA1c only in broad categories, which means critical variation in blood sugar control is lost. The presence of the variable is useful, but its categorical structure limits how fully it can predict outcomes. Medication fields, including diabetesMed and the directional changes recorded for insulin and oral agents, show treatment choices that affect stability after discharge. Demographic measures such as race and gender allow comparisons across groups to detect disparities in risk.

Several dimensions are missing that limit the dataset's ability to reflect known predictors of readmission. Diagnostic codes are raw and ungrouped, which makes it difficult to interpret

comorbidities with accuracy. Greater clarity about whether conditions were chronic or acute, or pre-existing versus newly identified, would improve the usefulness of these fields. Laboratory results beyond HbA1c and glucose are absent, even though research links other measures such as lipid profiles and kidney function to diabetic complications. Puspitasari and Aliviameita (2018) found a significant relationship between renal function tests and lipid profiles in patients with diabetes in an observational study. This reinforces the importance of including these variables to capture complications that influence outcomes.

Socioeconomic context is also excluded, which blocks analysis of factors outside the hospital setting. Lusk et al. (2022) showed that patients from socioeconomically deprived neighborhoods faced higher 30-day readmission rates for diabetes even after accounting for age, comorbidities, and insurance coverage. This illustrates how the absence of variables tied to neighborhood, income, or insurance status prevents the dataset from capturing influences that extend beyond clinical care. The analysis risks overlooking drivers of readmission that the literature has already established as significant without these factors.

Race includes 2,271 missing entries, a number large enough to equal smaller categories such as Hispanic or Other. This gap weakens the ability to measure disparities in outcomes by race. Diagnosis codes also include missing entries and reduce the precision of any analysis linking comorbidities to readmission. Leaving gaps unaddressed can bias results and mishandling them through careless imputation can create false patterns. Recognizing the missing information and the limits of the data that is available is necessary to ensure the analysis builds on variables that matter and acknowledges the gaps where insight may be lost.

Data Quality

Completeness varies dramatically across variables. Core measures such as age, gender, time in hospital, and readmission outcome are fully populated. Race is nearly complete with only 2% missing, and the three diagnosis fields each contain less than 2% missing values. However, weight is absent for 97% of patients, payer code for over half, and medical specialty for more than half. These gaps hinder the opportunity for important perspectives. Weight directly relates to complications in diabetes management, yet its near absence removes an essential predictor. Deng et al. (2025) found that higher BMI correlates with higher HbA1c levels, showing that obesity significantly impairs glycemic control. The lack of weight data in the dataset makes it impossible to explore this relationship by leaving out an important factor that influences readmission. Insurance coverage, captured by payer code, often determines the continuity of care following discharge, but the field is missing for most patients.

Provider specialty information could help separate outcomes by generalists versus endocrinologists, but its absence leaves the analysis blind to differences in expertise. Hosseinzadeh et al. (2025) found completeness to be the most frequently assessed dimension of health data quality across studies and emphasize that missingness in clinically relevant variables undermines both the accuracy of research findings and the development of targeted interventions.

Demographic fields such as gender, age, and readmission status are consistently coded, which eliminates errors that could fragment the dataset, however, there are validity issues. The gender field includes an “Unknown/Invalid” category that introduces ambiguity in demographic analysis. Diagnostic codes are shortened to three digits of ICD-9, simplifying storage but erasing

distinctions that matter in clinical practice. For example, a three-digit code may indicate "diabetes with complications," while the extended code specifies whether the complication is renal or neurological. Lewis et al. (2023) identifies this issue as a loss of correctness in coding, where simplified entries compromise precision and weaken the ability to tie outcomes to specific conditions.

Every encounter required lab testing and medication administration, which reduces the likelihood of irrelevant or fabricated records. Its multi-institutional coverage over a decade increases reliability by incorporating variation across hospitals. Yet reliability is weakened in variables with extreme missingness. Weight cannot be considered accurate or representative with only 3% of cases recorded, making it misleading for analysis even though it is a central factor in diabetes outcomes. Hosseinzadeh et al. (2025) emphasize that implausible or incomplete fields add noise and diminish reproducibility, which parallels how the near absence of weight reduces the dataset's trustworthiness.

The years 1999 to 2008 came before many of the advances in diabetes care now considered standard. New drug classes such as GLP-1 receptor agonists and SGLT2 inhibitors, along with wider adoption of outpatient disease management, reshape how patients are treated today. Lewis et al. (2023) describes timeliness as a distinct dimension of health data quality and caution that older datasets lose relevance as clinical practice changes. The historical value of the dataset remains, but its age restricts direct application to modern predictive models.

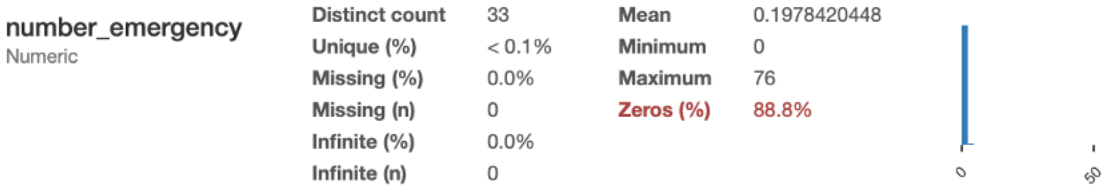
The dataset offers reliable coverage of basic demographics, utilization, and outcomes but falls short in completeness, validity, and timeliness.

Data Character

Variables in the diabetes readmission dataset show their character through how they distribute and vary. Numeric fields reveal asymmetries. The number of emergency visits skews heavily, with most patients recording no visits and a few showing extreme counts (see Figure 1). Skewness of this magnitude can overwhelm statistical models, which often assume more balanced distributions. Adjustments may be needed to reduce the influence of extreme values and make the data easier to analyze.

Figure 1

Summary Statistics and Distribution of Number of Emergency Visits.



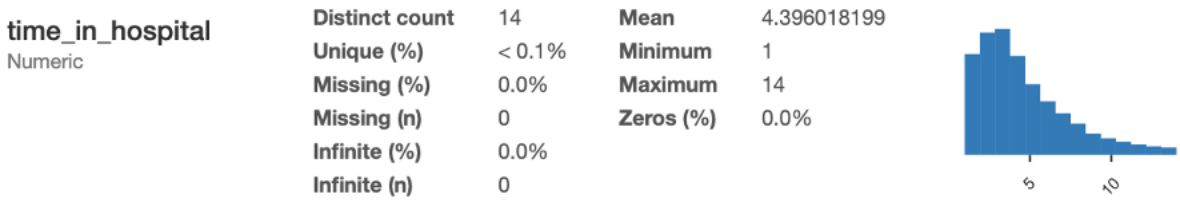
Note. Most patients recorded zero emergency visits, while a small subset reached as high as 76, producing a right-skewed distribution. Data derived from the Diabetes Profiling Report from Kaggle (2017).

Other numeric variables, such as time in hospital, display more constrained ranges but still show long tails. Length of stay ranges from 1 to 14 days, with most admissions clustering around shorter stays and a gradual taper among patients hospitalized for longer periods (see Figure 2). The average stay is about 4.4 days, yet a minority of cases extend toward the maximum. These longer hospitalizations consume disproportionate resources and often indicate higher severity or complications. Their presence in the dataset is important because they may

signal patients at higher risk of returning, even though they represent a small fraction of the population.

Figure 2

Summary Statistics and Distribution of Length of Hospital Stay

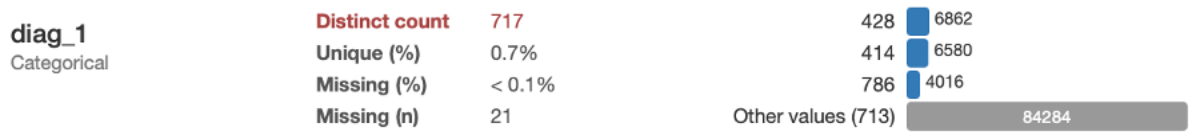


Note. Most admissions lasted fewer than five days, while a smaller subset of patients remained hospitalized up to 14 days, creating a right-skewed distribution. Data derived from the Diabetes Profiling Report from Kaggle (2017).

The dataset's categorical fields add detail but also create complexity. Three diagnostic variables each contain hundreds of unique codes, which fragments the data and complicates interpretation. The primary diagnosis field alone contains 717 unique values, with a few codes such as 428 and 414 appearing frequently while most others occur rarely (see Figure 3). Without grouping these codes into broader categories, analysis risks producing noise rather than uncovering meaningful trends. Organizing diagnoses into systems or chapters aligns them with clinical reasoning and strengthens their value in modeling.

Figure 3

Summary Statistics and Distribution of Primary Diagnosis Codes (diag_1)



Note. The field diag_1 contains 717 unique diagnosis codes, with codes 428 and 414 occurring most often while hundreds of other codes appear only rarely. This distribution demonstrates high cardinality, which complicates analysis without grouping. Data derived from the Diabetes Profiling Report from Kaggle (2017).

Medication variables add complexity because they track whether treatments changed during the hospital stay. Categories such as “Steady,” “Up,” or “Down” show whether a drug remained the same, increased, or decreased in dosage. These categories are ordinal, since they carry a natural order that reflects clinical progression. Treating them as ordinary categories would ignore that order, while encoding them as ordinal values preserves the meaning behind the changes. This illustrates how the character of the data shapes the need for transformation, since the way these variables are structured directly affects how well they capture clinical patterns.

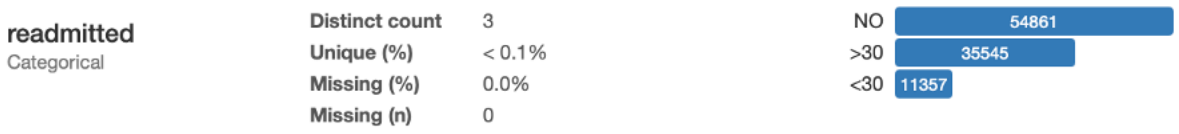
Several fields expose redundancy or constancy that weakens the dataset’s character. Encounter_id and patient_nbr act as identifiers rather than explanatory features. Citoglipton and examide remain constant across all records, providing no variability. Variables that fail to vary cannot contribute to predictive strength and can even mislead algorithms if retained.

Outliers and imbalances should also be considered. A handful of patients record extraordinary values for utilization variables, such as dozens of emergency visits, which skew descriptive statistics and complicate visualization. The outcome variable of readmission is also

imbalanced, with most patients not returning, fewer readmitted after thirty days, and the smallest group readmitted within thirty days (see Figure 4). This imbalance affects how predictive models assign importance to outcomes and creates the need for techniques such as resampling or weighting so that patterns associated with readmission are not overshadowed by the majority class.

Figure 4

Summary Statistics and Distribution of Readmission Outcomes



Note. Most patients were not readmitted, while fewer returned after thirty days and the smallest group returned within thirty days and produced a clear class imbalance. Data derived from the Diabetes Profiling Report from Kaggle (2017).

Numeric fields require transformation to manage skewness and outliers. Categorical fields benefit from grouping or encoding strategies that preserve meaning. Redundant and constant variables need to be discarded to streamline analysis. By recognizing these qualities, the analysis can embrace the dataset's strengths while counteracting distortions and set the stage for models that capture the realities of diabetes readmission more faithfully.

General Reflections

The dataset shows how hospitals once tracked diabetes care, and it invites stories that can be told through careful organization. Patterns in length of stay suggest most admissions were brief, yet a small fraction of patients stayed much longer and consumed disproportionate

resources. This could frame a narrative around which patients drive the heaviest utilization and whether those extended stays translate into lower or higher readmission risk. Emergency visit counts show a similar imbalance, with most patients never using the emergency department and a handful relying on it repeatedly. These patterns highlight how resource use is concentrated in a small group of patients.

Hundreds of unique codes scatter across the dataset and make them unusable in their raw form. Grouping them into broader systems or organ categories could create a framework for asking more coherent questions, such as whether cardiac complications or renal complications play a larger role in readmission. The codes in their current state are more administrative in nature rather than clinical insight. With thoughtful restructuring they could carry narrative weight. Medication variables show directional changes that trace how treatment shifted during admission. These categories have the potential to show responsiveness to treatment, however, their lack of detail weakens their power. A small increase in insulin looks the same as a dramatic change and leaves the analyst with only the outline of a story rather than its substance.

Weight is almost entirely missing, removing an anchor for studying the role of obesity in diabetes outcomes. Insurance coverage is incomplete, and socioeconomic indicators never appear, erasing the chance to connect readmission risk to access or deprivation. Laboratory results, like HbA1c, are reduced to broad categories that blur meaningful variation. These omissions make for shallow data and not suited for clinical decision-making.

Distributions showed skewness that simple averages would have concealed, such as the clustering of emergency visits at zero with a long tail of extreme values. Summary statistics revealed how missingness concentrated in certain variables, exposing the weakness of weight

and payer code fields. The visual and numerical summaries clarified which variables had meaningful variability and which did not.

The descriptive statistics provide a reliable first step for exploratory data analysis by surfacing skewness, imbalance, and missingness in a structured way. They offer orientation rather than validation and guide how to reshape the dataset through grouping diagnosis codes, encoding ordinal medication changes, removing constant variables, and addressing imbalanced outcomes. This dataset is best for teaching what it means to work with messy healthcare data, since it shows that volume does not guarantee insight and that gaps and fragmentation require careful preparation. It is useful for practicing profiling and transformation but cannot serve as a guide for modern policy or clinical practice because it is dated, incomplete, and shallow in certain variables. The dataset provides a foundation for exploring patterns in diabetes readmission, however, it exposes the limits of real-world clinical data and the preparation required before rigorous statistical analysis.

References

Deng, L., Jia, L., Wu, X.-L., & Cheng, M. (2025, February 21). Association Between Body Mass Index and Glycemic Control in Type 2 Diabetes Mellitus: A Cross-Sectional Study. National Library of Medicine. https://pmc.ncbi.nlm.nih.gov/articles/PMC11853989/

Hazra, A., & Gogtay, N. (2016). Biostatistics Series Module 3: Comparing Groups: Numerical Variables. Indian Journal of Dermatology. https://journals.lww.com/ijd/fulltext/2016/61030/biostatistics_series_module_3__comparing_groups_.2.aspx

Hosseinzadeh, E., Afkanpour, M., Momeni, M., & Tabesh, H. (2025, August 9). Data Quality Assessment in Healthcare, Dimensions, Methods and Tools: A Systematic Review. National Library of Medicine. https://pmc.ncbi.nlm.nih.gov/articles/PMC12335082/

Humberto, B. (2017, October 31). Diabetes 130 US Hospitals for Years 1999-2008. Kaggle. https://www.kaggle.com/datasets/brandao/diabetes

Lewis, A. E., Weiskopf, N., Abrams, Z. B., Foraker, R., Lai, A. M., Payne, P. R. O., & Gupta, A. (2023, June 30). Electronic Health Record Data Quality Assessment and Tools: A Systematic Review. Oxford Academic. https://academic.oup.com/jamia/article/30/10/1730/7216383

Lusk, J. B., Hoffman, M. N., Clark, A. G., Bae, J., Corsino, L., & Hammill, B. G. (2022, September 15). Neighborhood Socioeconomic Deprivation and 30-Day Mortality and Readmission for Patients Admitted for Diabetes Management. American Diabetes Association. https://diabetesjournals.org/care/article/45/11/e169/147605/Neighborhood-Socioeconomic-Deprivation-and-30-Day

Puspitasari, & Aliviameita, A. (2018). Relationship Between Renal Function Test Serum and Lipid Profile in Patients with Diabetes Mellitus. IOP Science. https://iopscience.iop.org/article/10.1088/1742-6596/1114/1/012011/meta

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014, April

3). *Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records*. Wiley Online Library. https://onlinelibrary.wiley.com/doi/epdf/10.1155/2014/781670