Raed K Alotaibi

raedkma@gmail.com

Abstract

[Draw your reader in with an engaging abstract. It is typically a short summary of the document.   
When you’re ready to add your content, just click here and start typing.]

Analysis of Diabetic Population Claims Data

A report

**Technical Exercise Overview**

Attached to this communication is a dataset representing a sample of diabetic population claims. This dataset includes demographic information, ICD-10 codes, Charlson comorbidity index scores, BMI, a de-identified unique member ID, and a de-identified policy number.

The purpose of this exercise is to evaluate the applicant's proficiency in accurately analyzing the provided data. The applicant should demonstrate their ability to perform a thorough analysis and then articulate their insights and conclusions in the form of an analytical report. The report should address the following points:

* Preprocess the raw data to prepare it for analysis.
* Conduct descriptive analytics to summarize the data.
* Demonstrate any quality or plausibility concerns.
* Identify any significant patterns or trends from the analysis.
* Apply advanced analytical techniques such as predictive modeling, clustering, or machine learning methods.
* Propose strategies to address any identified gaps, and share insights and reflections based on the analysis.
* Describe the methodology and tools used in the analysis.

This exercise is an opportunity to showcase analytical capabilities and strategic thinking.

Okay lets now structure the analysis report. It should contain the following sections:

* Executive summary
* Introduction to the project [aims, objectives including demonstrating the applicants analytic and technological capabilities (we will include a skills matrix the would be reflected in the steps of the report). –
* Preprocessing steps
* Descriptive analytics summarizing the data including
  + [data set size,
  + column description,
  + column types,
  + numeric and categorical summaries including
  + missing values and duplicate rows] –
* Data quality and plausibility with suggested changes to enhance the data
  + [ e.g. member code is not a unique id with reasons,
  + no time stamps to indicate the trend in outcome or other measures like BMI,
  + claim type with an "I" is a duplicate with an "O" value row] –
* Feature engineering and creation of a feature store prior to advanced analytics –
* Patterns and trends analysis
* Advanced analytics: [
  + hypothesis generation and testing,
  + clustering analysis to identify similar grouping,
  + other machine learning]
* Strategies to address gaps, insights and reflections
* Describe the methodology and tools used:
  + Project setup and version control
  + Data processing & Analytics:
  + Visualization
  + Machine learning
  + AI & LLM: DeepSeek, Agentic flow, Langflow for architecture
  + Front end: streamlit
* Future Directions & Enterprise -level suggestions

Contents

[Analysis of Diabetic Population Claims Data 4](#_Toc189035668)

[1. Executive Summary 4](#_Toc189035669)

[2. Introduction 5](#_Toc189035670)

[2.1 Project Aim 5](#_Toc189035671)

[2.2 Objectives 5](#_Toc189035672)

[2.3 Skills Matrix 5](#_Toc189035673)

[3. Preprocessing Steps 6](#_Toc189035674)

[3.1 Data Cleaning 6](#_Toc189035675)

[3.2 Data Transformation 6](#_Toc189035676)

[4. Descriptive Analytics 7](#_Toc189035677)

[4.1 Dataset Summary 7](#_Toc189035678)

[4.2 Numeric Data Summary 7](#_Toc189035679)

[4.3 Categorical Data Summary 7](#_Toc189035680)

[5. Data Quality and Plausibility Assessment 8](#_Toc189035681)

[Suggested Enhancements: 8](#_Toc189035682)

[6. Feature Engineering & Feature Store Creation 9](#_Toc189035683)

[7. Patterns and Trends Analysis 10](#_Toc189035684)

[8. Advanced Analytics 11](#_Toc189035685)

[8.1 Hypothesis Generation & Testing 11](#_Toc189035686)

[8.2 Clustering Analysis 11](#_Toc189035687)

[8.3 Predictive Modeling 11](#_Toc189035688)

[9. Addressing Gaps, Insights, and Reflections 12](#_Toc189035689)

[10. Methodology & Tools Used 13](#_Toc189035690)

[11. AI Capabilities & Agentic Flows 14](#_Toc189035691)

[12. Future Directions & Enterprise-Level Suggestions 15](#_Toc189035692)

# **Analysis of Diabetic Population Claims Data**

## **1. Executive Summary**

This report presents an in-depth analysis of a diabetic population claims dataset to assess data quality, identify key trends, and apply advanced analytical techniques. The report outlines preprocessing steps, descriptive analytics, feature engineering, machine learning applications, and strategic recommendations for improving data usability and deriving actionable insights. The findings will support better decision-making in healthcare claims management.

## **2. Introduction**

### **2.1 Project Aim**

The primary goal of this project is to analyze claims data for a diabetic population, highlighting key trends, data quality issues, and predictive insights. The analysis will demonstrate expertise in data processing, feature engineering, and advanced analytics, ultimately supporting enhanced claims processing and risk assessment.

### **2.2 Objectives**

* **Demonstrate analytical and technological capabilities** through systematic data exploration and predictive modeling.
* **Enhance data quality** by identifying inconsistencies and proposing corrective measures.
* **Uncover significant patterns and trends** in the diabetic population's claims.
* **Develop machine learning models** to generate predictive insights.
* **Recommend strategies** to improve data integrity and analytics adoption in enterprise settings.

### **2.3 Skills Matrix**

|  | **Application in Report** |
| --- | --- |
| Data Cleaning | Handling missing values, duplicates, and plausibility checks |
| Descriptive Analytics | Summarizing dataset attributes, distribution, and trends |
| Data Quality Assessment | Identifying inconsistencies and proposing enhancements |
| Feature Engineering | Creating structured and meaningful features for modeling |
| Advanced Analytics | Applying clustering, predictive modeling, and hypothesis testing |
| AI & LLM Integration | Exploring retrieval-based AI insights for structured and unstructured data |

## **3. Preprocessing Steps**

### **3.1 Data Cleaning**

* Identified and handled missing values.
* Removed duplicate records in claim types and demographic attributes.
* Standardized categorical values (e.g., claim type inconsistencies like “I” vs. “O”).
* Verified uniqueness of Member ID and Policy Number.

### **3.2 Data Transformation**

* Converted categorical variables into suitable formats for analysis.
* Engineered new time-based features (if timestamps were available).
* Created additional risk stratification categories from Charlson comorbidity index scores.

## **4. Descriptive Analytics**

### **4.1 Dataset Summary**

* **Dataset size:** [number of rows, columns]
* **Column descriptions and types**
* **Missing values and completeness assessment**
* **Duplicate row analysis**

### **4.2 Numeric Data Summary**

* Statistical measures (mean, median, standard deviation, min-max range) for BMI, age, and other numeric fields.
* Distribution plots for key numerical attributes.

### **4.3 Categorical Data Summary**

* Frequency distributions for claim types, ICD-10 codes, and policy numbers.
* Cross-tabulations to examine relationships between categories.

## **5. Data Quality and Plausibility Assessment**

* **Member Code Verification:** Checking if it is truly a unique identifier.
* **Time Series Gaps:** Assessing whether timestamps are present for trend analysis.
* **Data Completeness:** Evaluating missing BMI values and inconsistencies in ICD-10 code formatting.
* **Claim Type Consistency:** Identifying duplicate claims with different coding formats.
* **Charlson Comorbidity Index Checks:** Ensuring correct computation and validity.

### **Suggested Enhancements:**

* Assigning standardized unique IDs if needed.
* Introducing temporal tracking of claims to analyze trends.
* Standardizing ICD-10 coding for better classification.

## **6. Feature Engineering & Feature Store Creation**

* **New Features for Predictive Modeling**
  + Risk stratification using Charlson comorbidity index.
  + BMI categorization (underweight, normal, overweight, obese).
  + Claim frequency indicators per member.
* **Feature Encoding**
  + One-hot encoding for categorical features.
  + Dimensionality reduction where applicable.
* **Storing Engineered Features for Reusability**
  + Creation of a structured feature store for machine learning applications.

## **7. Patterns and Trends Analysis**

* **Distribution of Claim Types across Age Groups.**
* **BMI vs. Claim Frequency:** Identifying whether obesity increases claims likelihood.
* **Charlson Score & ICD-10 Trends:** High-risk conditions and frequent diagnoses.
* **Policy Number Analysis:** Potential clustering of members under specific policies.

## **8. Advanced Analytics**

### **8.1 Hypothesis Generation & Testing**

* Does BMI significantly influence the frequency of claims?
* Are certain ICD-10 codes associated with higher costs?
* Is there a seasonal trend in claims data (if timestamps exist)?

### **8.2 Clustering Analysis**

* Grouping members by claim frequency and comorbidity risk.
* Identifying outlier patterns in claim behavior.

### **8.3 Predictive Modeling**

* **Classification Models** (e.g., Decision Trees, Logistic Regression) to predict high-risk claimants.
* **Regression Models** to estimate claim costs.

## **9. Addressing Gaps, Insights, and Reflections**

* **Data Enhancement Needs:**
  + Improving unique identifiers.
  + Introducing timestamps for trend tracking.
  + Standardizing claim types.
* **Strategic Recommendations:**
  + Implementing real-time anomaly detection for fraudulent claims.
  + Developing AI-driven dashboards for policy risk assessments.
* **Challenges and Considerations:**
  + Data completeness and bias in missing records.
  + The necessity for longitudinal data for better insights.

## **10. Methodology & Tools Used**

* **Data Processing & Analytics:** Python (pandas, NumPy, scikit-learn), SQL
* **Visualization:** Matplotlib, Seaborn, PowerBI
* **Machine Learning:** Clustering (K-Means, DBSCAN), Predictive Modeling (Logistic Regression, Decision Trees)
* **AI & LLM Integration:** Exploring retrieval-based insights using generative AI to enhance structured and unstructured data insights.

## **11. AI Capabilities & Agentic Flows**

* **Retrieval-Augmented Generation (RAG) for Claims Analysis:**
  + Implementing LLMs to extract structured and unstructured insights.
  + Enhancing claims decision-making through AI-assisted summarization.
* **Agentic Flows for Automated Analytics:**
  + Developing AI workflows to classify claims into risk categories.
  + Automating feature engineering using generative AI models.

## **12. Future Directions & Enterprise-Level Suggestions**

* **Enterprise-Level AI Strategy:**
  + Implementing predictive risk assessment in claims processing.
  + Integrating external health data sources for deeper insights.
  + Developing an AI-powered anomaly detection system for fraud detection.
* **Data Governance & Security Considerations:**
  + Ensuring privacy compliance in claims data.
  + Standardizing data pipelines for enhanced efficiency.