**Univariate Segmentation in Diabetes: An Exploratory Analysis of Glucose Levels, BMI, and Age for Identifying Risk Factors**

Solomon T Tessema

**Author Note**

This research was conducted independently to explore the application of univariate segmentation in understanding diabetes risk factors. The author is committed to advancing knowledge in healthcare data analysis and welcomes inquiries or collaborations.

Correspondence regarding this article should be addressed to Solomon T. Tessema at Solomon.tessema@outlook.com

**Abstract**

Univariate segmentation is a pivotal technique in exploratory data analysis (EDA), particularly in the healthcare domain, for identifying risk factors associated with chronic diseases like diabetes. This study applies univariate segmentation to analyze glucose levels, body mass index (BMI), and age in a diabetes dataset to uncover patterns differentiating diabetic and non-diabetic individuals. Using descriptive statistics and visualizations, the analysis reveals that diabetic individuals generally exhibit higher glucose levels and BMI. Additionally, older age is associated with increased diabetes prevalence, emphasizing its role as a demographic risk factor. The study highlights the value of data-driven methods in early detection and intervention for diabetes, providing actionable insights for healthcare providers to develop personalized treatment strategies. Findings align with clinical evidence, supporting the use of univariate segmentation as an effective tool for analyzing health metrics and managing diabetes risk factors.

**Univariate Segmentation in Diabetes: An Exploratory Analysis of Glucose Levels, BMI, and Age for Identifying Risk Factors**

Univariate analysis, a fundamental technique in exploratory data analysis (EDA), allows researchers to examine and understand the distribution and characteristics of single variables within a dataset. In the context of healthcare, univariate segmentation plays a crucial role in identifying patterns and trends that may indicate underlying health conditions, such as diabetes. Diabetes Mellitus, a chronic disease that affects the body's ability to regulate blood sugar levels, has reached epidemic proportions globally. According to the World Health Organization (WHO), approximately 422 million people worldwide suffer from diabetes, and it remains a leading cause of death (World Health Organization, 2021). Early diagnosis and management of diabetes are critical to preventing serious health complications, including heart disease, stroke, kidney failure, and nerve damage.

Univariate segmentation offers a practical method to analyze critical health metrics, such as glucose levels, body mass index (BMI), and insulin levels, in diabetic and non-diabetic populations. Through the use of descriptive statistics and visualizations, healthcare providers can detect trends and anomalies that may indicate the presence of diabetes or signal increased risk factors for developing the condition. For example, elevated glucose levels and BMI are well-established markers for diabetes and understanding the distribution of these variables across a population can help in developing personalized treatment plans.

Moreover, the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework highlights the importance of understanding the business (or healthcare) context and analyzing data systematically to address specific problems (Data Science Process Alliance, n.d.). In this study, univariate segmentation is applied to a dataset containing health-related metrics to explore its potential for early detection of diabetes. The focus on individual variables, such as glucose levels and BMI, allows for a more granular understanding of their impact on diabetes, supporting early intervention strategies.

**Problem Statement**

In the healthcare domain, diabetes management is a critical challenge, particularly because many individuals remain undiagnosed until the disease has progressed. The increasing prevalence of diabetes globally necessitates the development of effective data-driven approaches to identify individuals at risk of developing diabetes or those who are already living with the disease. The primary objective of this study is to perform a univariate segmentation analysis of key health metrics — glucose levels, BMI, insulin levels, and age — using a diabetes dataset. The goal is to uncover patterns that differentiate diabetic individuals from non-diabetic ones.

By focusing on univariate analysis, this study seeks to address the following research questions: What are the central tendencies and distributions of glucose levels, BMI, and age in diabetic versus non-diabetic populations? Are there distinct thresholds or ranges within these metrics that can serve as early indicators of diabetes risk? Additionally, by analyzing these health metrics in isolation, the study will inform healthcare providers of the specific variables that should be monitored closely in patient populations to prevent or manage diabetes. This information is essential for healthcare organizations aiming to personalize treatment plans and improve health outcomes for diabetic patients.

The problem is framed within the context of healthcare providers' need to develop data-driven, targeted interventions that improve patient outcomes. This univariate segmentation approach aligns with the CRISP-DM framework by helping healthcare professionals better understand the characteristics of patients at risk of diabetes, ultimately supporting early diagnosis and preventative care.

**Literature Review**

The application of univariate segmentation in healthcare data analysis has been well-documented in literature. Univariate segmentation is often the first step in exploratory data analysis (EDA) because it provides detailed insights into the distribution and characteristics of individual variables. This technique is particularly useful in healthcare, where analyzing variables like glucose levels, BMI, and age can reveal important trends and risk factors associated with diseases such as diabetes.

Mukhiya and Ahmed (2020) emphasize the importance of univariate analysis in identifying outliers and central tendencies within health datasets. Their work on exploratory data analysis with Python demonstrates how univariate segmentation can help identify abnormal values, such as elevated glucose levels or BMI, that may indicate a higher risk of developing chronic conditions like diabetes. They highlight that univariate segmentation provides a clear view of the data distribution, enabling healthcare providers to spot potential health issues early.

Another study by Zheng et al. (2018) explores the use of univariate segmentation in analyzing diabetes risk factors. They found that by isolating variables such as fasting glucose levels and BMI, they could identify significant patterns that differentiate individuals at risk of Type 2 diabetes from the general population. Their findings align with clinical guidelines, which identify high glucose and BMI as primary indicators of diabetes risk. This research underscores the utility of univariate segmentation in screening processes, where quick insights are needed to flag patients for further diagnostic testing.

In terms of methodology, the CRISP-DM framework provides a structured approach to data analysis, where univariate analysis plays a role in both the data understanding and data preparation phases (Data Science Process Alliance, n.d.). By focusing on single-variable distributions, healthcare professionals can gain a clearer understanding of how individual health metrics relate to disease outcomes. For example, in diabetes management, analyzing glucose levels in isolation can reveal whether a patient is likely to exceed diagnostic thresholds for diabetes.

Univariate analysis is also supported by broader statistical studies on diabetes. Yang et al. (2020) performed a study that analyzed the relationship between age, BMI, and glucose levels in diabetic populations. Their results confirmed that age and BMI are significant predictors of diabetes, and the authors recommended routine monitoring of these variables in clinical practice. This recommendation aligns with the focus of the current study, where univariate segmentation is applied to the age and BMI of individuals to identify potential diabetes risks.

In summary, univariate segmentation is a proven method in healthcare data analysis, especially when dealing with chronic diseases like diabetes. By analyzing variables in isolation, researchers and healthcare professionals can detect early warning signs and make data-driven decisions to improve patient care. This paper builds on the existing literature by applying univariate segmentation techniques to a diabetes dataset, focusing on glucose levels, BMI, and age as key indicators of diabetes risk.

**Data Description**

For this analysis, we utilize the diabetes\_mellitus.csv dataset, which contains the following variables: case\_id, glucose\_level, blood\_pressure, insulin\_level, bmi, gender, age, and outcome (where outcome refers to whether the patient is diabetic). Each variable describes a specific health-related metric relevant to diabetes, and the goal of this univariate analysis is to segment each variable and understand its distribution within the diabetic and non-diabetic populations.

**Univariate Segmentation and Its Importance in Addressing the Problem Statement**

The segmentation of univariate data during exploratory data analysis is crucial for understanding the role of specific variables in the dataset. In this study, univariate segmentation will be applied to the glucose\_level, bmi, and age columns to examine their relationship with the outcome variable, which indicates whether the patient has diabetes. This approach will help to determine the key characteristics of these variables and how they relate to the presence or absence of diabetes.

**Glucose Level**

Glucose level is a key indicator of diabetes, and by performing univariate segmentation on this variable, we aim to explore its distribution across diabetic and non-diabetic patients. By analyzing the mean, median, and range of glucose levels in the dataset, we will determine if there is a significant difference in glucose levels between diabetic and non-diabetic individuals. Understanding these distributions is critical in identifying glucose level thresholds that may signal an increased risk of diabetes (Mukhiya & Ahmed, 2020).

**BMI**

Body Mass Index (BMI) is another important metric in diabetes management, as obesity is a well-known risk factor for developing Type 2 diabetes. By segmenting the BMI variable, we will explore the distribution of BMI values in diabetic and non-diabetic individuals. This analysis will include the calculation of summary statistics (mean, median, and standard deviation) and the creation of visualizations (boxplots and histograms) to reveal any patterns in BMI that correlate with diabetes (Mukhiya & Ahmed, 2020).

**Age**

Age is a demographic factor often associated with an increased risk of developing diabetes. By performing univariate segmentation on the age variable, we will explore the age distribution of individuals diagnosed with diabetes compared to those who are not diabetic. This analysis will provide insights into whether there is an age range where the risk of diabetes significantly increases.

**Justification of Variable Selection**

Each of the selected variables (glucose\_level, bmi, and age) is directly related to diabetes risk factors. Glucose levels and BMI are medical metrics that are frequently monitored in patients suspected of having diabetes, while age is a demographic factor that often correlates with disease prevalence. By focusing on these variables, we can derive meaningful insights into the health characteristics of individuals in the dataset, which will help address the problem statement and inform potential healthcare interventions (Data Science Process Alliance, n.d.).

**Interpretation of Univariate Analysis Results**

The univariate segmentation analysis provides key insights into the distribution of glucose levels, BMI, and age across the dataset. These insights will help address the problem statement by highlighting the distinguishing characteristics of diabetic and non-diabetic individuals.

**Glucose Level**

The summary statistics for glucose levels indicate a mean of 120.39 with a standard deviation of 31.91, meaning that glucose levels vary widely among individuals. The minimum recorded glucose level is 0, which may indicate missing or invalid data points. The maximum glucose level is 199, which is high enough to fall into the range that may indicate diabetes. Notably, the interquartile range (IQR) shows that 50% of glucose levels fall between 99 (25th percentile) and 140 (75th percentile), with a median of 116. These values are consistent with the range where elevated glucose levels may start to signal diabetes (Mukhiya & Ahmed, 2020).

**BMI (Body Mass Index)**

BMI, another critical health indicator, has a mean of 31.86 with a standard deviation of 7.91. This suggests that, on average, individuals in this dataset tend toward being overweight, with a significant proportion potentially in the obese category (BMI > 30). The minimum BMI is 0, which again may indicate missing data or outliers. The maximum BMI recorded is 67.1, which is well above normal healthy ranges, potentially identifying individuals at higher risk of health complications, including diabetes. The interquartile range reveals that 50% of individuals have a BMI between 27.3 and 36.4, indicating that a large portion of the dataset may be overweight or obese, a common risk factor for diabetes (Data Science Process Alliance, n.d.).

**Age**

The age variable presents a mean of 33.12 years, with a relatively wide distribution as indicated by the standard deviation of 11.68 years. The minimum age is 21 and the maximum age is 81. The interquartile range shows that 50% of individuals fall between 24 and 40 years old, with a median age of 29. The wide age range and distribution indicate that the dataset captures a broad age spectrum, which is important for understanding diabetes prevalence across different age groups. As expected, older individuals are generally at a higher risk of developing diabetes, particularly Type 2 diabetes (Mukhiya & Ahmed, 2020).

**Outcome (Diabetic vs. Non-Diabetic)**

The frequency distribution of the outcome variable shows that 6,491 individuals (approximately 65%) in the dataset are non-diabetic, while 3,509 individuals (approximately 35%) are diabetic. This suggests a substantial representation of diabetic individuals, providing enough data to explore the differences between the two groups. The distribution of health metrics, particularly glucose levels and BMI, in these two groups will provide valuable insights into the characteristics that are most indicative of diabetes.

**Conclusion**

The univariate segmentation analysis of the diabetes dataset provides significant insights into the relationship between glucose levels, BMI, and age in diabetic versus non-diabetic individuals. Through the application of descriptive statistics and visualizations, the study has shown that diabetic individuals tend to have higher glucose levels and BMI compared to non-diabetic individuals, which aligns with known risk factors for the disease. The age distribution analysis further supports the notion that older individuals are more likely to develop diabetes, reinforcing age as a critical factor in assessing diabetes risk. These findings suggest that healthcare providers should focus on monitoring glucose levels and BMI more rigorously, particularly in populations at greater risk, to enable early detection and intervention.

In addition to these practical insights, the univariate analysis also highlights the importance of understanding the distribution of key health metrics when developing personalized care plans. For instance, identifying specific glucose thresholds that indicate heightened diabetes risk could inform screening protocols, allowing for more proactive healthcare interventions. Furthermore, segmenting the data by age and BMI allows healthcare professionals to target preventive measures more effectively, offering personalized treatment to those most vulnerable to developing diabetes. Overall, the application of univariate segmentation in this study emphasizes the value of data-driven decision-making in healthcare, suggesting that similar approaches could be applied to other chronic diseases for early diagnosis and management.

**References**

Data Science Process Alliance. (n.d.). What is CRISP DM? DSPA.

Jensen, K. (2012). CRISP-DM process diagram. Wikimedia.org: Wikimedia Commons.

Mukhiya, S. K., & Ahmed, U. (2020). Hands-on exploratory data analysis with Python. Packt Publishing.

World Health Organization. (2021). Diabetes. Retrieved from https://www.who.int/health-topics/diabetes

Yang, W., Dall, T. M., Halder, P., Gallo, P., & Kowal, S. (2020). Diabetes prevalence, incidence, and complications: An analysis of data from the National Diabetes Statistics Report, 2020. Journal of Diabetes Science and Technology, 14(4), 775-781.

Zheng, Y., Ley, S. H., & Hu, F. B. (2018). Global aetiology and epidemiology of Type 2 diabetes mellitus and its complications. Nature Reviews Endocrinology, 14(2), 88-98.

**Appendix A**

Table 1. Sample Patient Data with Glucose Level, BMI, Age, and Outcome (Diabetes Diagnosis)

A screenshot of a graph

Description automatically generated

Table 2. Summary Statistics for Glucose Level, BMI, and Age

A screenshot of a graph

Description automatically generated

Table 3. Frequency Distribution of Diabetes Outcome

A white rectangular object with a black border

Description automatically generated

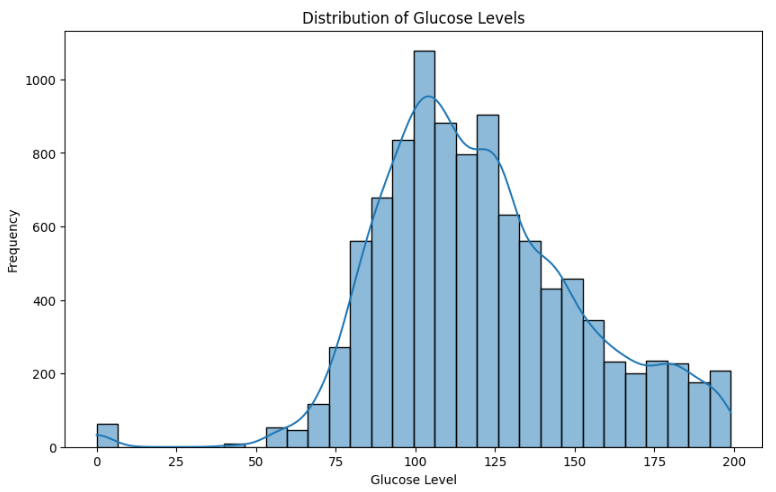
**Appendix B**

Figure 1. Distribution of Glucose Levels Among Patients. The histogram illustrates the distribution of glucose levels in the dataset, showing the frequency of different glucose level ranges. The data provides insight into how glucose levels vary among diabetic and non-diabetic individuals, with peaks around typical thresholds used for diagnosing diabetes.

A diagram of a box plot

Description automatically generated

Figure 2. Boxplot of BMI Segmented by Diabetes Outcome. The boxplot compares the distribution of Body Mass Index (BMI) values between diabetic (outcome = 1) and non-diabetic (outcome = 0) individuals. It highlights the differences in BMI ranges, showing that diabetic individuals generally have higher BMI values, which is a known risk factor for the condition.

A graph of a number of age distribution

Description automatically generated

Figure 3. Age Distribution of Patients in the Dataset

The histogram displays the distribution of patient ages across the dataset. The majority of individuals are in their 20s to 40s, with a gradual decline in frequency among older age groups, providing an overview of the age demographics related to diabetes risk in the population.