

# KOLEJ PROFESIONAL MARA BERANANG DIPLOMA IN COMPUTER SCEINCE

|  |  |  |
| --- | --- | --- |
| **COURSE NAME** | : | DATA ANALYSIS & VISUALIZATION |
| **COURSE CODE** | : | CSC 2823 |
| **SESSION** | : | 1 2024/2025 |
| **TYPE OF ASSESSMENT** | : | FINAL PROJECT |
| **DURATION** | : | 13/06/2024 – 3/07/2024 |

**CLO 3 Prepare data visualization for effective presentation using computer software tool. (A4, PLO6)**

**INSTRUCTION TO CANDIDATES:**

1. Students are required to answer ALL questions.
2. Students need to submit the assignment report in hardcopy.

|  |  |
| --- | --- |
| **Personal Details** | |
| **Name** | MUHAMMAD EQMAL BIN NOOR RANIZ |
| **I/D Number** | BCS2211-026 |
| **Class** | DCS 5B |
| **Lecturer** | * **PN. MAWARWIDURI AB HALIK** |

|  |  |
| --- | --- |
| **Task** | **Marks** |
| 1 |  |
| 2 |  |
| 3 |  |
| 4 |  |
| 5 |  |
| **Total** | **/ 50** |

# PROJECT SCENARIO

You are a data analyst for VISME IT Sdn Bhd. Your new task is creating an insightful and impactful data visualization project based on a dataset related to an issue addressed by the World Health Organization (WHO). The aim of this project is to uncover meaningful patterns, trends, and insights that can help inform public health decisions or raise awareness about a critical health issue.

Your preparation for the presentation is based on the main tasks given below:

1. Dataset Selection:

Visit the provided link to access a collection of datasets. These datasets are curated from various sources and are related to global health issues that the WHO is concerned with.

Browse through the datasets and select one that interests you and is relevant to a current or historical health issue addressed by the WHO.

Link: <https://www.who.int/data/gho/> Example of dataset:

* + Global Health Estimates by WHO
  + Global Health Observatory (GHO) data
  + WHO COVID-19 Dashboard data
  + WHO Noncommunicable Diseases Data

1. Project Objectives and Reasoning:

Objectives: Clearly state the objectives of your data visualization project. What specific questions are you aiming to answer? What insights are you hoping to reveal?

Reason for Selection: Explain why you selected this particular dataset and health issue. Discuss the potential impact of visualizing this data and how it could benefit public health awareness or policy-maker.

1. Clean and preprocess the collected data to ensure accuracy and consistency. Handle missing or erroneous data points and format the data for analysis
2. Analyze the data to extract meaningful insights and trends. Identify key statistics and patterns by using different types of variable(s) that can effectively communicate the gravity of the awareness issues to the exhibition attendees.
3. Create visually appealing and informative data visualizations to convey the identified insights. Utilize charts, graphs, maps, and other visualization techniques to make the information easily understandable. The visualization should use different types of variables and suitable presentation approaches.

**Assessment Rubrics:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Exploratory Data Analysis (EDA)** | | | | | | | |
|  | Attribute | Task | **2** | **4** | **6** | **8** | Marks |
| 1 | Data Selection | | | |  | | |
|  | Gather | Topic Selection | * The dataset is minimally relevant or not relevant to any WHO   health issue. The connectio n to the WHO's  concerns is unclear or non- existent.   * The dataset has significan t gaps and incomplet e data, covering few   aspects of | * The dataset has some relevance to a WHO health issue, but the connection is weak or not immediatel y clear.   More explanatio n is needed to establish relevance.   * The dataset has some potential for analysis and visualizatio n but may   be limited | * The dataset is relevant to a WHO health issue, with a mostly clear connection. Some aspects of the relevance might need further clarification. * The dataset covers most aspects of the health issue with generally complete data but may have minor gaps. * The dataset has good | * The selected dataset is highly relevant to a critical health issue addressed by the WHO. * The dataset is comprehensiv e, covering all necessary aspects of the health issue with complete and detailed data. * The dataset has high potential for analysis and visualization, with rich data that can yield meaningful insights and trends. * Able to |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | the health issue. | in scope or detail and incomplete  . | potential for analysis and visualization  , with enough data to generate useful  insights. | identify target audience appropriately |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 2 | Project objectives and reasoning | | | | | Marks |
|  | Gather | Project Objectives and Reasoning | **3** | **6** |  |  |
| * Objectives are clear but could be more specific or detailed. * Provide minimal and unclear discussion on the potential impact of visualizing the selected dataset and how it could benefit public health awareness or policy- maker. | * Objectives are:   + clearly stated   + specific and,   + well-defined. * Provide sufficient and clear discussion on the potential impact of visualizing the selected dataset and how it could benefit public health awareness or policy-maker. |  |  |
| 3 | Clean and preprocess the collected data. | | | | |  |
|  | Gather |  | **2** | **4** | **6** |  |
| * Provide a screen shot of the steps to treat any missing values or quantities of zero in the dataset with some steps not accurately done. * No explanations of the step provided | * Provide a screen shot of the correct steps to treat any missing values or quantities of zero in the dataset. * Provide brief explanation of the steps | * Provide a screen shot of the correct steps to treat any missing values or quantities of zero in the dataset. * Provide detailed explanation of the steps |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data Visualization** | | | | | | |
| 4 | Analyze the data | | | | |  |
|  |  |  | **2** | **4** | **6** |  |
|  | Reproduce and Process Information | Conduct univariate analysis(1 variables) by providing an appropriate chart or plot with appropriate observations | * Able to provide **ONE** category of univariate analysis (1 mark) * Able to display visual with a suitable selection of a plot or graph for each category (1 mark) | * Able to provide more than one category of univariate analysis (2 marks) * Able to display visual on categories with a suitable selection of a plot or graph for each category (2 marks) | * Able to provide more than one category of univariate analysis (2 marks) * Able to display visual on categories with a suitable selection of a plot or graph for each category (2 marks) * Provide a clear and concise finding explanation based on the analysis conducted for each category (2 marks) |  |
| Conduct bivariate analysis(2 variables) by  providing an appropriate chart or plot with appropriate observations. | * Able to provide **ONE** category of bivariate analysis (1 marks) * Able to display visual on categories with a suitable selection of a plot or graph for each category (1 marks) | * Able to provide more than one category of bivariate analysis (2 marks) * Able to display visual on categories with a suitable selection of a plot or graph for each category (2 marks) | * Able to provide more than one category of bivariate analysis (2 marks) * Able to display visual on categories with a suitable selection of a plot or graph for each category (2 marks) * Provide a clear and concise finding explanation based on   the analysis conducted |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | for each category (2 marks). |  |
| 5 | Create visually appealing and informative data visualizations | | | | |  |
|  |  |  | **2** | **4** | **6** |  |
|  | Convey | Summarize the findings | * Able to highlight some of the important findings. * No interpretation of result provided. | * Able to highlight some of the important findings. * The interpretation of results is briefly explained but inappropriate, based on the analysis of the data. | * Able to highlight most of the important findings. * The interpretation of results is briefly explained and appropriate, based on the analysis of the   data. |  |
| Use suitable types of presentation  for the non- technical audience | * Provide unsuitable types of   presentations for the targeted audience.   * Use ggplot 2 library to create the graphic. | * Provide suitable types of presentations for the targeted audience. * Use the ggplot2 library to create the graphic. * The created visualisation is simple and only cater to two(2) of the following:   + Colors and contrast   + size   + theme   + scale   + shape | * Provide suitable types of presentations for the targeted audience. * Use the ggplot2 library to create the graphic. * The visualisation produced is complete and caters to four(4) of the following:   + Colors and contrast   + size   + theme   + scale   + shape |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | * label and legend | * label and legend |  |
|  | Convey | Customize graph in creative ways | * The visualisation produced cover only **one(1)** of the aspects of below:   + Good quality with clear and   accurate   * + Have aesthetic appeal with visually pleasing   + easy for the audience to understand. | * The visualisation produced **two(2)** of relevant aspects of below :   + Good quality with clear and accurate   + Have aesthetic appeal with visually pleasing   + easy for the audience to understand. | * The visualisation produced cover all relevant aspects of below   :   * + Good quality with clear and accurate   + Have aesthetic appeal with visually pleasing   + easy for the audience to understand. |  |
| **Total Marks Earned** | | | | | | **/50** |
| **Total Percentage (40%)** | | | | | |  |

***TABLE OF CONTENT***

**1.0 INTRODUCTION**

**2.0 DATASET SELECTION**

**2.1 NAME DATASET**

**2.2 EXPLAIN ABOUT DATASET**

**2.3 SOURCE**

**2.4 STRING DATASET**

**2.5 TARGET AUDIENCE**

**3.0 OBJECTIVE**

**3.1 IMPACT OF VISUALIZING**

**4.0 DATA CLEANING**

**5.0 ANALYZE DATA**

**5.1 UNIVARIATE**

**5.1.1 - CATEGORICAL**

**5.1.2 - CONTINUOUS**

**5.2 BIVARIATE**

**5.2.1 - CATEGORICAL VS CONTINUOUS**

**5.2.2 - CONTINUOUS VS CONTINUOUS**

**6.0 DATA VISUALIZATION**

**6.1 TITLE OF DATA**

**7.0 APPENDICES**

**8.0 REFERENCES**

***1.0 Introduction***

Diabetes is a long-term metabolic illness marked by high blood glucose (blood sugar) levels. Serious damage to the heart, blood vessels, eyes, kidneys, and nerves can result from diabetes over time. The most prevalent kind, type 2 diabetes, usually affects adults and is brought on by insufficient or resistant insulin production in the body. Type 2 diabetes has been far more common over the last three decades in all nations, regardless of wealth. Diabetes type 1 is a chronic illness in which the pancreas generates little or no insulin on its own. It was formerly referred to as juvenile diabetes or insulin-dependent diabetes. Having access to reasonably priced therapy, such as insulin, is essential for the survival of those with diabetes. By 2025, there is a goal set globally to stop the rise in diabetes and obesity.   
  
Approximately 422 million individuals globally suffer from diabetes, with the majority residing in low- and middle-income nations. The disease is directly responsible for 1.5 million fatalities annually. Over the past few decades, there has been a steady rise in both the number of cases and the incidence of diabetes.

**2.0 Dataset Selection**

**2.1 Selected Data Name:** Diabetes Prediction

**2.2 Explanation Dataset:**

The "Diabetes Prediction" dataset, which focuses on predicting diabetes using a variety of health markers, is extremely pertinent to solving a serious health concern identified by the World Health Organization (WHO).  
With complete and detailed data covering all area of diabetes prediction, the "Diabetes Prediction" dataset is very extensive. Its rich data may provide significant trends and insights into the field, and its analysis and visualization potential are strong. To appropriately identify diabetes, this analysis aims to analyze a variety of health-related characteristics and their interconnections. These variables include things like blood sugar level, age, gender, body mass index (BMI), blood pressure, heart disease, history of smoking, and HbA1c level. This thorough analysis will lay the groundwork for future study while also offering insights into the patterns and trends in diabetes risk. Studies can be conducted to better understand the ways in which these factors interact to affect the incidence and course of diabetes, information that is vital for enhancing patient care and results in this increasingly important field of medicine.

**2.3 Source:** [🍬Diabetes : EDA |🌲Random Forest🌲 + HP⚙️ (kaggle.com)](https://www.kaggle.com/code/tumpanjawat/diabetes-eda-random-forest-hp/input)

**2.4 String Dataset:**

A screenshot of a computer code

Description automatically generated

**2.5 Target Audience:** Physicians and the Medical Field

For medical practitioners, the "Diabetes Prediction" dataset is a priceless resource. It provides thorough and precise information covering every facet required for diabetes prediction. Healthcare practitioners will be able to spot trends and gain valuable insights from this dataset through analysis and visualization. By using this data, medical professionals can increase the accuracy of their diagnoses, develop better preventative and treatment plans, and gain a better understanding of the diabetes predictors. This dataset can also be used by public health experts to guide initiatives and policies that manage and lower the prevalence of diabetes in the general population. Furthermore, this type of data could also educate the general population.

**3.0 Project Objectives and Reasoning**

**Objectives:**

To identify trends, patterns, and correlations within the diabetes dataset, which can aid in a deeper understanding of the prevalence and risk factors associated with diabetes, thereby contributing to more targeted and effective interventions.

To create detailed visualizations that display key metrics, such as the incidence and prevalence of diabetes across different population segments, demographic breakdowns including age, gender, and ethnicity, geographical distribution highlighting areas with higher or lower rates of diabetes, and temporal trends showing how these metrics change over time.

To utilize these visualizations to support healthcare professionals, researchers, and policymakers in their efforts to combat diabetes by providing them with clear, accessible, and actionable insights drawn from comprehensive data analysis.

**3.1 Impact of Visualizing:**

**I. Public Awareness**

Campaigns for public awareness are essential for educating communities on the risks, consequences, and prevalence of diabetes. Through data analysis, public health officials can efficiently focus their programs by identifying geographic hotspots and demographic trends associated with diabetes prevalence. By promoting early screening and improved lifestyle choices, this awareness can lower the incidence of diabetes-related complications and improve their management.

**II. Early Detection and Intervention**

By using data analysis to identify diabetes risk factors early, healthcare providers can take prompt action. Healthcare providers can provide focused screening programs, lifestyle modifications, and preventative therapies by identifying those at high risk based on predictive algorithms and historical data analysis. By taking a proactive stance, healthcare expenses related to controlling advanced stages of diabetes are decreased while also improving health outcomes.

**III. Customized Patient Management**

By offering insights into unique patient profiles and treatment outcomes, data analysis facilitates personalized patient management. Healthcare providers can customize treatment programs to match the specific needs of each patient by assessing patient data, including demographics, genetic predispositions, and lifestyle factors. By concentrating interventions where they are most successful, this tailored approach maximizes the use of healthcare resources while improving treatment adherence and patient satisfaction.

**4.0 Data Cleaning**

**First Step:** Approach Data

Start by looking at the rows and variables in the dataset. To get a basic idea of the dataset's structure and content, pay attention to the first cases in the beginning. The groundwork for more data exploration and analysis is laid in this step.

****

**A computer screen shot of numbers

Description automatically generated**

The first step is to carefully review the first examples and the structure of the dataset. This aids in our comprehension of its structure, early detection of abnormalities such as missing data or outliers, and early identification of its content and possible trends. This fundamental finding directs more in-depth investigation and data preparation, guaranteeing data quality and influencing efficient decision-making procedures.

**Second Step:** Check Missing Value

In this stage, we will evaluate each variable's data type and look for any missing values in the dataset. By doing this, we can make sure the data is clean and prepared for analysis, which will help us to find any gaps or inconsistencies that could lower the calibre of the insights.

A close-up of a computer screen

Description automatically generated

A number of numbers in a row

Description automatically generated with medium confidence

This Stage Is Executed to make sure the data is accurate and full by going over the data types and looking for any missing values in the dataset. To ensure data integrity and reliability throughout our study, it is imperative that we identify any areas where the data may be incomplete or formatted incorrectly. This approach assists us in doing just that. It makes sure that the analysis and interpretation that come after are built upon a strong platform of accurate and fully structured data, which improves the overall calibre and validity of our conclusions.

**Third Step:** Profiling Data Input

This Third Step is, thoroughly profiling and classifying incoming data is the main priority. This entails closely analysing the type, origin, and structure of the data that is being received. This stage guarantees accurate and effective processing and analysis by creating comprehensive profiles that include data kinds, quality measures, and expected trends. To verify data integrity, spot abnormalities, and get ready for the next steps of the data lifecycle, profiling data intake is crucial.

****

**A screenshot of a computer

Description automatically generated**

Profiling data input involves assessing the type, format, and quality of incoming data. It helps identify patterns, anomalies, and ensures data integrity. This step prepares data for accurate processing, enhancing its quality for informed decision-making and strategic initiatives.

**Fourth Step:** Ordering Data by Percentage of Zero

At this point, the data is arranged according to the percentage of zeros in each dataset. By providing a clear understanding of the distribution and frequency of zero values, this type of data sorting makes it easier to do further research and get new insights into the characteristics of the dataset and how it may be used for decision-making.

****

**A screenshot of a computer

Description automatically generated**

Sorting data by the percentage of zeros helps us identify sparse data points, assess data quality issues like missing values, and make informed decisions about how to analyze and interpret the dataset effectively.

**Fifth Step:** Removing Variables with High Number of Zeros

Identifying and removing variables from a dataset that show a notable predominance of zero values could be referred to as the "Fifth Step: Removing Variables with High Number Of Zeros" procedure. By minimizing any noise or bias that these factors may cause, this process seeks to improve the quality of the data, preparing the dataset for more precise analysis or modelling.

****

**A close up of text

Description automatically generated**

This stage improves the quality of the data by locating and eliminating variables that have a high percentage of zero values. By lowering any noise and bias in the dataset, this increases the accuracy of our research and guarantees more trustworthy outcomes from further data processing or modelling endeavours.

**Sixth step**: Keep all column except the ones present in ‘vars\_to\_remove’

In this stage, every column in the dataset will be kept, except for those that are marked with an asterisk (\*) in the 'vars\_to\_remove' list. By doing this, you can make sure that only the variables you want to analyse or process further are kept, in line with project specifications and data management objectives.

****

**A screenshot of a computer code

Description automatically generated**

Here, we make sure that our dataset keeps all its columns—aside from the ones listed in 'vars\_to\_remove'. By eliminating extraneous factors and concentrating on pertinent data, we can efficiently streamline our analytical and processing endeavours maintain a simplified dataset that concentrates on pertinent variables by retaining all columns other than those in 'vars\_to\_remove'. By making the data easier to understand and analyse, this simplification guarantees that the most relevant information is used to generate our insights.

**5.0 Analyze Data**

**5.1 Univariate**

The study of a single variable to identify trends and characteristics is known as univariate analysis. It entails utilizing measures of dispersion (range, variance, and standard deviation) and central tendency (mean, median, and mode) to summarize the data. Histograms, box plots, and bar charts are examples of visual aids that are used to show the distribution and point out anomalies. This analysis also includes comprehending the probability distribution and, if necessary, implementing data changes. Before conducting more intricate studies, univariate analysis offers crucial insights into the fundamental characteristics of the data.

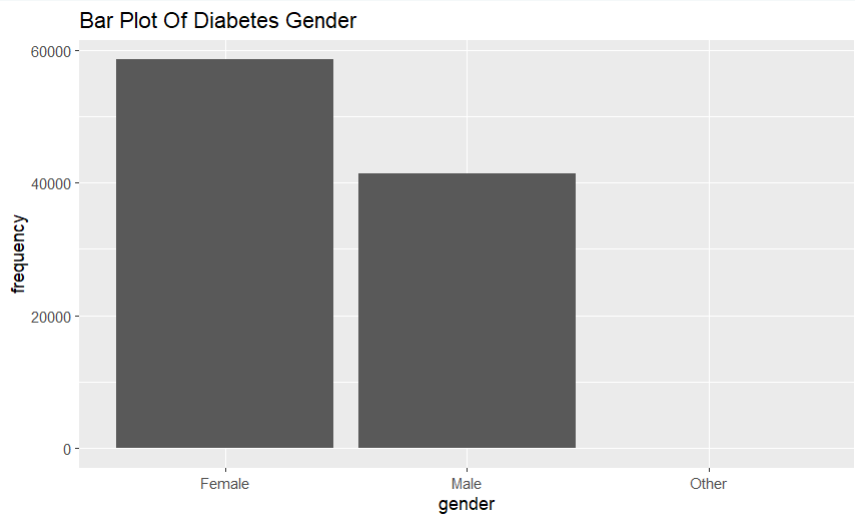
**5.1.1 - Categorical**

In a categorical univariate analysis, we concentrate on the frequency of each gender category within the dataset to characterize the gender distribution.

The various gender categories (such as male, female) are shown on the x-axis, while the frequency which shows the quantity of instances or counts of each gender category is plotted on the y-axis.

For example, this data can be shown graphically using a bar chart. Every bar signifies a gender category, and the bar's height reflects how frequently that category occurs. We can easily determine which gender categories are prominent in the sample thanks to this straightforward yet effective presentation. To display the proportion of each gender category in a more logical and comparable way, a pie chart might alternatively be utilized.

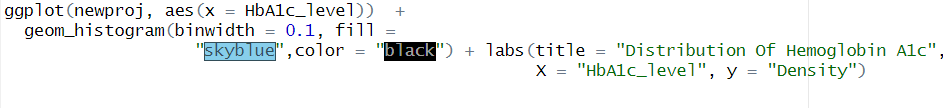
We can comprehend the gender distribution of the population under investigation by looking at this univariate description, which might be important for a variety of studies and decision-making procedures.

****

In summary, women outnumbered men in this dataset when it came to gender dominance.so the majority of the data comes from the perspective of women and mostly of them got diabetes.

**5.1.2 - Continues**

The relationship between hemoglobin A1c levels and density (frequency) can be understood through a histogram or density plot. These visuals show how hemoglobin A1c values are distributed across the dataset. The x-axis represents the range of hemoglobin A1c levels, while the y-axis shows the frequency or density of observations within each range. This helps identify typical levels, patterns, and potential outliers, providing insights into blood sugar control and health status within the population studied.

****

**A graph of a number of blue vertical lines

Description automatically generated with medium confidence**

In conclusion, we can see that most of the individuals in this data set have high blood sugar levels, diabetes that reaches level 6 or above, and prediabetes, which is characterized by blood sugar levels of level 5. The only person with a normal blood sugar level who is on level 4 or lower

**5.2 Bivariate**

Bivariate data analysis involves examining the relationship between two variables. It focuses on understanding how changes in one variable are associated with changes in another. By plotting these variables on a graph, such as a scatter plot, patterns can emerge that show whether the variables move together in a predictable way (positive correlation), move oppositely (negative correlation), or show no apparent relationship. This analysis helps in uncovering connections and dependencies between variables, providing insights into their interactions and potential impacts on outcomes of interest.

**5.2.1 - Categorical Vs Continuous**

This boxplot illustrates gender differences in blood sugar levels (HbA1c) within a diabetes dataset. For females, most have blood sugar levels within a normal range, though some have much higher levels, indicating a need for additional care. Males show a similar pattern, with most controlling their blood sugar within a specific range but a few with higher levels requiring extra attention. The "Other" gender category exhibits more variability in blood sugar levels, indicating a broader range of blood sugar management within this group. ****

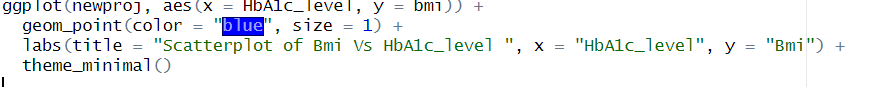
**A graph of different colored squares

Description automatically generated**

Overall, while average blood sugar levels are consistent across all gender categories, the "Other" category shows greater variability. Additionally, some individuals in both the male and female groups have higher levels and may need further support in managing their diabetes

**5.2.2 - Continuous Vs Continuous**

The association between average blood sugar levels and body weight over a period of several months is seen in the scatter plot. It demonstrates that blood sugar levels vary widely across people with different body weights. Higher or lower blood sugar levels do not reliably correlate with body weight; there is no clear trend to support this theory. Rather, people with comparable blood sugar levels can have notably disparate body weights. Common observations are indicated by the majority of data points clustering inside a particular range of blood sugar values. Body weights vary greatly within this range, showing that blood sugar levels can be identical in individuals of varied body weights. There are exceptions, such as people who weigh a lot more than average, and who have different blood sugar levels than other people in the sample.

****

**A graph of blue lines

Description automatically generated**

Overall, the plot highlights the intricate interaction between blood sugar levels and body weight, highlighting the need for individualized health assessments as opposed to broad generalizations based only on blood sugar level or weight categories.

**6.0 Data Visualization**

**Coding**

**A screenshot of a computer code

Description automatically generated**

**Visualization**

**A graph of diabetes level

Description automatically generated with medium confidence**

**6.1 Title of Data**

**Average Blood Glucose Level (Sugar Level) by Gender and Diabetes Status**

**Description:**

This study delves into the average blood glucose levels across different genders and their diabetes statuses, aiming to uncover meaningful insights for healthcare management.

**Findings:**

In analysing the data, significant variations in average blood glucose levels were observed between genders and across different diabetes statuses. Generally, males tended to exhibit slightly higher average glucose levels than females across all categories of diabetes status—whether diabetic or non-diabetic.

The impact of diabetes status on blood glucose levels was pronounced, with diabetic individuals consistently showing higher average glucose levels compared to non-diabetic individuals. This disparity was particularly noticeable in females, where diabetic females demonstrated the highest average glucose levels among all groups studied.

The interaction between gender and diabetes status further highlighted critical nuances. Specifically, diabetic females exhibited a distinct profile with elevated glucose levels, suggesting a need for gender-specific approaches in diabetes management.

**Interpretation:**

These findings underscore the importance of tailored healthcare interventions based on gender and diabetes status. Addressing the observed variations could lead to more effective diabetes management strategies and potentially improved health outcomes for patients. This narrative format aims to convey the key findings and their implications clearly and contextually.

**7.0 Appendices**

A screenshot of a computer

Description automatically generated

**A close-up of a finger with a needle

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**8.0 References**

 World Health Organization (WHO). (n.d.). Global Health Estimates. Retrieved from <https://www.who.int/data/gho/>

 Kaggle. (n.d.). Diabetes: EDA | Random Forest + Hyperparameter Tuning. Retrieved from <https://www.kaggle.com/>

 Mayo Clinic Blood Sugar Test and Glucose Tolerance. Retrieved From

<https://www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-20371451#:~:text=A%20fasting%20blood%20sugar%20level%20from%20100%20to%20125%20mg,this%20test%2C%20you%20fast%20overnight>.

 Explotary Data Analysis (EDA)-Update.pdf

[Exploratory Data Analysis (EDA)-update.pdf (sharepoint.com)](https://beranang.sharepoint.com/sites/CSC2823DCS5BSESI120242025/Class%20Materials/NOTES/WEEK%208-10/Exploratory%20Data%20Analysis%20(EDA)-update.pdf?wdsle=0&CT=1720033374145&OR=ItemsView&wdOrigin=TEAMSFILE.FILEBROWSER.DOCUMENTLIBRARY)