

Final Course Project:

Analyzing Diabetes Dataset



In this project we will use the publicly available diabetes dataset offered through Kaggle [[Diabetes Dataset](https://www.kaggle.com/datasets/hasibur013/diabetes-dataset?resource=download)] to analyze health trends and investigate factors contributing to diabetes onset. You are encouraged to leverage concepts from the book **"Storytelling with Data"** to select and design appropriate visualizations, ensuring clarity and impactful communication of insights.

| **PLEASE NOTE:** You will need to carefully review the diabetes dataset to understand how the data was collected, what it represents, and any limitations or potential biases that might influence your analysis. Additionally, ensure you interpret the columns and data values correctly by studying any available **documentation** or **metadata** associated with the dataset. Pay attention to:   1. **Data Format & Structure:** Familiarize yourself with the dataset layout, including column names and value ranges. 2. **Missing or Anomalous Data:** Identify any missing values or outliers that could impact your results. 3. **Health Context:** Consider the medical implications of the data and its attributes (e.g., glucose levels, BMI) when interpreting results. 4. **Scope & Generalizability:** Note that the dataset is specific to a particular demographic (e.g., female patients) and may not generalize to other populations. |
| --- |

In the next sections, you get to decide which data (or combinations of data) to use in each analysis task.



# PART 1: Exploratory Analysis: [20%]

**All analysis points must be accompanied by your commentary and at least one appropriate visualization.**

**Use the appropriate statistics and plots to investigate the following:**

1. The average glucose levels among patients with and without diabetes.
2. The average age of patients with and without diabetes.
3. The average blood pressure measurements across diabetic and non-diabetic groups.
4. The average BMI of diabetic versus non-diabetic patients.
5. The rate of diabetes among patients in the dataset.
6. The distribution of BMI values among all patients.
7. The distribution of Diabetes Pedigree Function (DPF) values for diabetic and non-diabetic patients.
8. The relationship between the number of pregnancies and diabetes occurrence.
9. The correlation between glucose levels and BMI.
10. The trend of glucose levels with age among diabetic and non-diabetic patients.





# PART 2: Answering Questions: [20%]

**All analysis points must be accompanied by your commentary and at least one appropriate visualization.**

* 1. **Use the appropriate statistics and plots to answer the following questions:**
     1. Are higher glucose levels associated with a greater likelihood of diabetes?
     2. Are patients with high glucose concentrations also likely to have higher BMI values?
     3. Are patients with a higher number of pregnancies at greater risk of developing diabetes?
     4. Are older patients more likely to have higher insulin concentrations and blood glucose levels?
     5. Can you identify common “risk profiles” for diabetic patients based on key metrics (glucose, BMI, age, etc.)?
  2. **Come up with 5 more bivariate/multivariate analysis questions and similarly answer each with appropriate visuals and commentary.**





# PART 3: Hypothesis Testing: [20%]

**Claim: “There is a significant difference in glucose levels between diabetic and non-diabetic patients.”**

* 1. **Formulate a hypothesis test to assess the validity of this claim given the available data:**
     + State the test you will use and justify your choice.
     + Clearly state the hypotheses.
     + Conduct the test and report the result.
     + Make a conclusion as to the validity of the claim, assume a significance level of 0.05.
  2. **Come up with your own claim from the available data and conduct a hypothesis test for it following in the same steps.**



# PART 4: Simulation Task: [15%]

In this part, you will apply confidence interval analysis to your provided dataset to determine the coverage of confidence intervals for relevant population parameters. Specifically, you'll analyze sample statistics and compute 95% confidence intervals for different sample sizes to observe the behavior of confidence intervals in a real-world dataset setting.

**4.1 Take 25 Random Samples of Size 15 from the Dataset**

* + - Select a relevant column from the dataset (e.g., **Glucose**, **BMI**, or **Age**).
    - Generate **25 random samples**, each containing **15 data points**, from the selected column.
    - For each sample, compute the **mean** and then calculate the **95% confidence interval** for the sample mean.
    - From your output, determine the proportion of these 25 confidence intervals that **contain the true population mean** of the selected column (e.g., mean glucose or BMI).

**4.2 Increase the Sample Size to 100**

* + - Now, take **25 random samples**, each with a sample size of **100**, from the same column you selected in part (a)..
    - For each sample, compute the **mean** and calculate the **95% confidence intervals**.
    - Observe the following:
      * Does the **width of the confidence intervals increase or decrease**?
      * Does increasing the sample size result in **more or fewer intervals containing the true population mean**?
    - Compare your findings with actual computations and interpret the results.

### 

**4.3 Take 20 Random Samples of Size 10 from the Dataset**

* + - Finally, generate **20 random samples**, each with a sample size of **10**, from the same selected column in the dataset.
    - For each sample, compute the **mean** and determine the **95% confidence intervals**.
    - Evaluate the coverage proportion and analyze the relationship between sample size and the accuracy of confidence intervals.



# PART 5: Documentation:[10%]

### Notebook Comments and Markdown

The **notebook** should focus purely on technical implementation, with concise Markdown sections to explain **why** certain steps or visualizations are used. It should contain:

1. **Steps for Data Cleaning and Preprocessing**
   1. Include step-by-step processes for data cleaning and transformation, explained briefly in Markdown.
   2. Use comments in code for clarity on specific operations.
2. **Visualizations**
   1. Generate visualizations and explain why each one is necessary.
   2. Use Markdown to clarify the purpose of each visualization (e.g., *"This plot shows the distribution of variable X to identify outliers."*).



1. **Functions (If Any)**
   1. Include any functions directly used in the notebook.
   2. Provide short Markdown explanations of their purpose, inputs, and outputs.
2. **Analysis Steps**
   1. Include code with concise Markdown headings to organize the workflow.
   2. Explain the logic or reasoning for each major step briefly (e.g., *"This section performs feature selection to improve model performance."*).
3. **Conclusions (Technical)**
   1. Summarize technical outcomes and computational results (e.g., metrics, patterns observed).

### PDF Report

The **PDF report** should focus on providing a high-level narrative and insights, excluding technical details like code or step-by-step processes. Include the following:

1. **Introduction**
   * Overview of the dataset, its source, and relevance to the project.
   * Clear problem statement and objectives.
2. **Data Processing and Analysis Steps**
   * Description of the methods used for data collection, cleaning, and preprocessing.
   * Overview of the analytical approach
3. **Challenges, Limitations, and Assumptions**
   * Discussion of challenges faced and solutions implemented.
   * Explanation of limitations in data or methodology and assumptions made.
4. **Results and Visualizations**
   * Include visualizations (charts, graphs) with detailed captions and explanations of their relevance.
   * Provide high-level interpretations of the results.
5. **Conclusion**
   * Summarize the findings and insights derived from the analysis.
   * Suggest future directions or potential improvements.





# PART 6: Presentation:[15%]

For the final presentation you should prepare slides summarizing all your findings with visualizations and brief comments. You’ll be graded on the quality of your visual aids and your ability to present your findings in a brief but compelling manner.



# Deliverables:

* Notebook (**R** Code, Comments, and Markdowns)
* Technical report (pdf)
* Presentation (.pptx Slides)

