Department of Mathematics
College of Arts and Sciences, Howard University
MATH 014 – Introduction to Data Science
FINAL PROJECT – GUIDELINE to explore the Data

Student Name:

Student ID: **Due date: 12-03-2025**

Topic: Diabetes Prediction Data Set
(https://www.kaggle.com/datasets/marshalpatel3558/diabetes-prediction-dataset)
This project explores biometric, lifestyle, and hereditary factors influencing diabetes using a predictive health dataset. Students will clean the data, perform exploratory data analysis, visualize trends, and identify key risk profiles using Python-based tools.

DATA UNDERSTANING AND CLEANING

Objective: Understand the dataset and clean it for analysis.

- Understand the dataset by examining:
  - the first few rows
  - datatypes and structure
  - columns of interest
  - null values
- Cleaning Requirements:
  - Drop any unnamed or index columns
  - Convert appropriate columns (e.g., "Sex", "Smoking_Status") to categorical types.
  - Check for duplicates and remove them if necessary.
  - Convert columns like date to a proper datetime format (if you have any).
  - Ensure all numeric fields are non-negative and in valid ranges.
  - Rename any long or ambiguous column headers for clarity.
- Additional Cleaning Steps (if required):
  - Derive new categorical columns such as BMI Category, Age Group, or Glucose Control Status.
  - Create risk indicators like High Blood Pressure Flag (based on BP values).

Documentation:

- Clearly explain every transformation or imputation performed.
- Justify why a particular approach was taken.

Note: You can include additional cleaning if required and make sure to highlight those. Explicitly highlight which approach was taken and why.

EXPLORATORY DATA ANALYSIS

Perform the following initial analysis:

- Understand structure and summary statistics.
- Count unique values for categorical fields.
- Explore distributions, group-wise means, and correlations between risk indicators.
- Optional: Create lifestyle score or metabolic syndrome score.

DATA VISUALIZATION AND INTERPRETATION

Include a **Heatmap to understand the correlation analysis** along with the **following core visualizations:**

Visualization	**Purpose**	**Tools Suggested**
Family History vs. Gender	Compare hereditary risk by gender	Seaborn countplot (seaborn/matplotlib)
BMI by Ethnicity	Compare obesity trends by ethnic group	Box plot (seaborn/matplotlib)
Fasting Blood Glucose	Understand glucose distribution and skewness	Histogram (seaborn/matplotlib)
HbA1c by Activity Level	Show glucose control variation by lifestyle	Violin Plot (seaborn/matplotlib)
BMI vs. Waist Circumference	Visualize obesity indicator correlation	Scatter Plot (seaborn/matplotlib)
Correlation Heatmap	Identify correlations between biometric metrics	Heat Map

Each plot should be accompanied by:

- A proper title, axis labels, and legend.
- ***Interpretation summarizing patterns or anomalies.***
- Customize fonts and visual themes for readability

INSIGHTS AND GENERALIZATIONS

Summarize:

- Which cities/countries are most polluted and why
- Which pollutants are most associated with high AQI
- Geographic regions that need attention
- Limitations of the dataset (e.g., missing location info)
- Recommendations or next steps if this data were used in policy

OPTIONAL RESEARCH QUESTIONS

If you choose to answer an optional research question, please restate the question in your report before analyzing it:

- Which countries have the most hazardous AQI levels?
- Does a high PM2.5 level always mean high AQI?
- How does the pollutant composition differ across cities?
- Are there clusters of cities by pollution profiles?
- Can you spot outliers (e.g., low AQI despite high NO2)?

EXPECTED OUTPUT

By the end of the project:

- Provide a clean dataset.
- Generate required visualizations.
- **Include at least 3 additional questions/visuals of your choice. These can use:**
  - Animated plots
  - Advanced interactivity via Plotly
  - (e.g., Age vs Glucose, stacked bar of Smoking vs Alcohol).

Submit:

- Report (.docx or .pdf) – 100 points
- Jupyter Notebook - 100 points (*firstname_lastname_final.ipynb* + *firstname_lastname_final.html*)
- Presentation (*firstname_lastname_final.pptx*) – 50 points

GENERAL INSTRUCTIONS FOR THE JUPYTER NOTEBOOK (CODING PART)

1. **Title of the Project:**
   - The title should be different from the original dataset title.
2. **Code Formatting:**
   - Structure your report clearly with headings for each section of the project (e.g., Data Understanding and Cleaning, Global Trends Analysis, etc.).
   - Provide a brief explanation of the methods and visualizations in each section.
3. **Code Clarity:**
   - Include well-commented code for every analysis or visualization performed.
   - Use meaningful variable names to improve code readability.
4. **Visualizations:**
   - All visualizations must include appropriate titles, axis labels, and legends where necessary.
   - Ensure plots are clear and easy to interpret (e.g., avoid overlapping labels).
   - Save all graphs and include those in the Project Report and Presentations.
5. **Data Cleaning:**
   - Clearly describe the steps taken to clean the data, including how missing values and outliers were handled.
   - If any assumptions were made (e.g., imputing values), explain them briefly in your report.
6. **Insights:**
   - For every question or analysis, include a short summary of your findings.
   - Highlight trends, correlations, or anomalies discovered during the analysis.
7. **Submission Requirements:**
   - Submit this coding part as a part of the project as a Jupyter Notebook (.ipynb) file, an exported HTML report and cleaned csv file.
   - Ensure your notebook runs without errors from start to finish.
8. **Academic Integrity:**
   - Plagiarism or copying code from others will result in penalties. Ensure your submission reflects your work.
   - Collaborate for discussions but submit independent work.
9. **Additions:**
   - Add additional analyses or visualizations if they provide meaningful insights.
   - Include a section titled Additional Analysis/Findings.