I LEARN ACADEMY
Digital Egypt Pioneers

Heart Disease Data Analysis Proposal

Prepared by:

Nora Raafat
Abdelrahman Gamal
Eman Ayman

Graduation Project Proposal

ProjectTitle:

Heart Disease Analysis

Group Information:

Group Name: ONL3_DAT1_G4

Advisor: Dr. Ahmed Abdelatife

Team Members:

- Nora Raafat
- Abdel Rahman Gamal
- Eman Ayman

1. Introduction and Problem Statement

Cardiovascular Diseases (CVDs) remain the leading cause of mortality globally. Effective public health strategies and clinical intervention rely heavily on accurately identifying high-risk populations. While machine learning offers predictive capabilities, a fundamental understanding of **which risk factors are most prevalent, highly correlated, and statistically significant** within a specific dataset is crucial for generating actionable public health recommendations.

The Problem: The provided dataset contains a wealth of health and lifestyle information, but lacks a clear descriptive analysis outlining the specific profile of individuals at risk for Coronary Heart Disease (CHD) and stroke.

Project Goal: To conduct a comprehensive Exploratory Data Analysis (EDA) and apply robust statistical tests to identify, quantify, and visualize the most significant demographic, clinical, and lifestyle risk factors associated with heart disease and stroke in the dataset. This will culminate in data-backed recommendations for healthcare policy or patient awareness campaigns.

2. Dataset Analysis (heart_disease.csv)

The project will utilize the provided **heart_disease.csv dataset,** focusing on descriptive statistics and the relationship between features and the Heart_ stroke target variable.

Feature Type	Example Features	Role in Analysis
Demographic/Lifestyle	Gender, age, education, currentSmoker, cigsPerDay	Analyze prevalence across different groups.
Medical History	BPMeds, prevalentStroke, prevalentHyp, diabetes	Measure association strength with the target event.
Clinical Measurements	totChol, sysBP, diaBP, BMI, heartRate, glucose	Determine critical thresholds and statistical differences between 'Yes' and 'No' groups.
Target Variable	Heart_ stroke (Binary: Yes/No)	The primary variable for comparative statistical testing.

3. Project Objectives

- Clean and preprocess the dataset using **Python**.
- Conduct **Exploratory Data Analysis (EDA)** using **Python (Pandas, Matplotlib, Seaborn)** to identify trends and correlations.
- Create interactive dashboards in **Power BI** for data visualization and insights presentation.
- Present the results in a structured report showing data-driven insights into heart disease factors.

Key research questions

- Which demographic and lifestyle features (age, gender, smoking, education) associate most strongly with Heart_ stroke occurrences?
- How do clinical indicators (BMI, systolic/diastolic BP, total cholesterol, glucose, heart rate) differ between stroke vs non-stroke groups?
- Which patient subgroups show elevated risk (age groups, smokers, hypertensive patients, diabetics)?

- What simple scoring/aggregation (SQL queries + visual thresholds) can support early identification of at-risk groups for screening/intervention?

---

5. Methodology & tools (by phase)

Phase 1 — Data ingestion & cleaning (Python / Pandas)

- Load dataset and produce data dictionary (variable names, types, descriptions).
- Fix column names and types (e.g., convert Heart_ stroke to consistent categorical encoding).
- Handle missing values with domain-aware strategies:
    - o Categorical missing (education) → consider Unknown / imputation by mode within age groups.
    - o Numeric missing (cigsPerDay, BPMeds) → impute with median or conditional median (e.g., smokers only).
- Detect & treat outliers for continuous variables (winsorize or trim for visualization; keep original values in a cleaned copy).
- Output: cleaned CSV and Python notebook documenting all steps.

Phase 2 — Exploratory Data Analysis (Python)

- Univariate analysis: distributions and summary stats for numeric features (mean, median, IQR) and categorical counts.
- Bivariate analysis against Heart_ stroke: t-tests / non-parametric tests for numeric variables; chi-square for categorical variables.
- Correlation matrix (numeric) and heatmap (visual only — descriptive).
- Grouped aggregations: average BMI, sysBP, glucose by target and by age bucket / gender / smoker status.
- Output: EDA notebook with plots and interpretation.

Phase 3 — Interactive dashboards (Power BI + Tableau)

- Dashboard (Power BI — Decision-maker view):
- **KPI cards:** total patients, stroke count, stroke rate (%), avg age, %smokers.
- **Filters:** age group, gender, smoker, education, diabetes, hypertension.
- **Visuals:** stacked bar (stroke by age group), scatter (BMI vs glucose with marker for stroke), distribution histograms, heatmap for correlation summary.
- **Drill-through pages** for individual risk profiles and aggregated SQL-backed summaries.
- Charts in (Tableau — Story & insights):

- Guided story: Overview → High-risk subgroups → Recommended interventions.
- Exportable charts and printable PDF story.
- Provide .pbix and .twbx deliverables with embedded data extracts.

Phase 4 — Reporting & presentation

- Executive summary (1–2 pages): top insights & recommended actions.
- Technical appendix: notebooks, SQL scripts, dataset dictionary, dashboard user guide.
  Final presentation (PowerPoint) for DEPI.

---

6. Expected Outcomes

- A fully cleaned and structured dataset ready for reporting.
- Descriptive statistics and correlation analysis between major health features.
- Interactive dashboards providing visual insights into heart disease patterns.
- A comprehensive report summarizing methods, findings, and recommendations.

---

7. Project Timeline

Task	**Duration**
Data Cleaning & Preparation (power Query & Python)	Week 1
Data Analysis (Python)	Week 2
Dashboard Development (Tableau & Power BI)	Week 3
Final Report & Presentation	Week 4

---

8. Deliverables

- Cleaned dataset CSV + data dictionary.
- Python notebooks (cleaning + EDA) with comments and visual outputs.
- Power BI (.pbix) dashboard and Tableau (.twbx) storyboard.
- Executive summary report (Word/PDF) + full technical appendix.
- Final presentation deck (PowerPoint) and a 10–12 minute demo script.