FINAL PROJECT:
STROKE RISK ANALYSIS

Submitted by: A S Sushmitha Urs

Course Number: 70954

Instructor Name: Amin Karimpour

Date of Submission: 24/10/2025

Introduction

Stroke is one of the leading causes of death and long-term disability worldwide, posing a major public health concern with significant social and economic impacts. Understanding the interplay of multiple risk factors such as hypertension, obesity, smoking, and age is crucial for developing effective prevention strategies and identifying high-risk populations. This study applies a systematic statistical approach, including exploratory data analysis, hypothesis testing, and regression modeling, to uncover meaningful relationships and patterns that can inform clinical practice and guide public health policy toward stroke prevention.

The analysis is based on a comprehensive health-related dataset containing 43,400 observations and 12 variables, encompassing both numerical and categorical features. It integrates demographic (e.g., gender, age, marital status, residence type), lifestyle (e.g., smoking status, work type), and medical (e.g., hypertension, heart disease, average glucose level, BMI) information, providing a multidimensional understanding of stroke risk. The target variable, stroke, indicates whether an individual has experienced a stroke, enabling the identification of significant predictors influencing its occurrence.

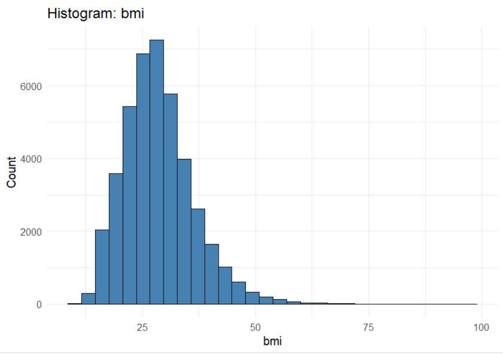
Data & Methods

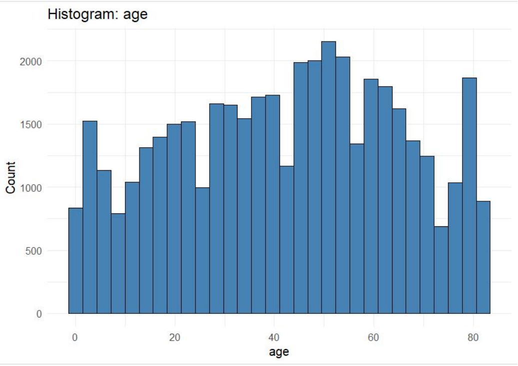
Variables include both numeric and categorical fields: age, BMI, average glucose level, hypertension, heart disease, stroke status, and smoking status. Binary variables (hypertension, heart_disease, stroke) were converted to factors with labels “Yes/No.” Empty smoking_status entries were set to NA. Numeric missing values were imputed using the median and winsorized at the 1st and 99th percentiles to reduce outlier distortion; categorical NAs were imputed with the mode. Duplicate rows were removed.

Exploratory Data Analysis (EDA)

Initial structure, summary statistics, and missing-value counts were printed to the console. Numeric variables were profiled with counts, mean, SD, quartiles, and range. Representative histograms and boxplots were generated for the first few numeric variables; bar charts summarized leading categorical variables.

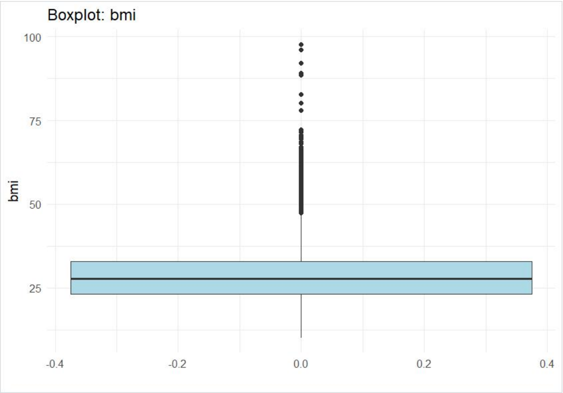
A pairwise correlation matrix was computed on numeric variables using pairwise deletion for missing values. The strongest absolute correlation pair was identified and plotted with a scatterplot to visually assess linearity.

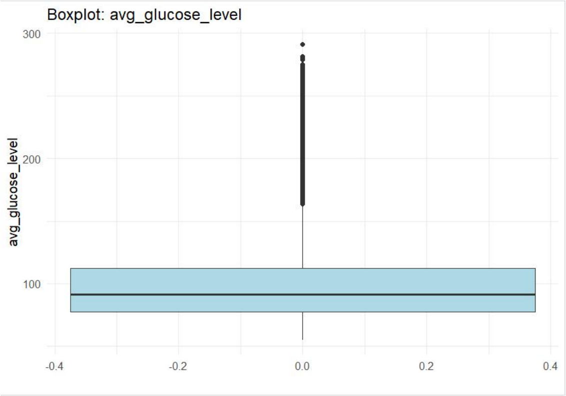




Histogram of BMI

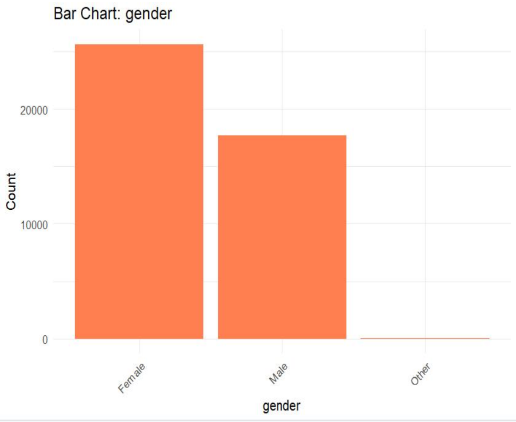
Histogram of Age

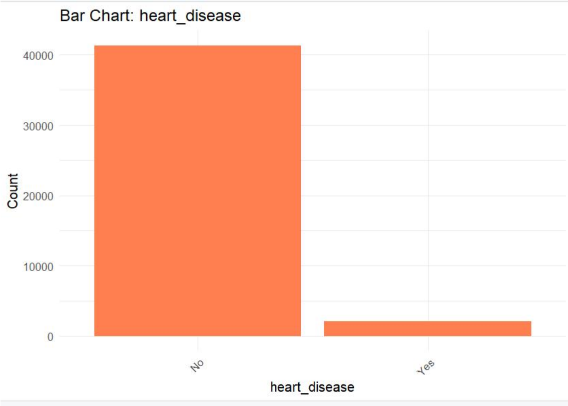




Boxplot of BMI

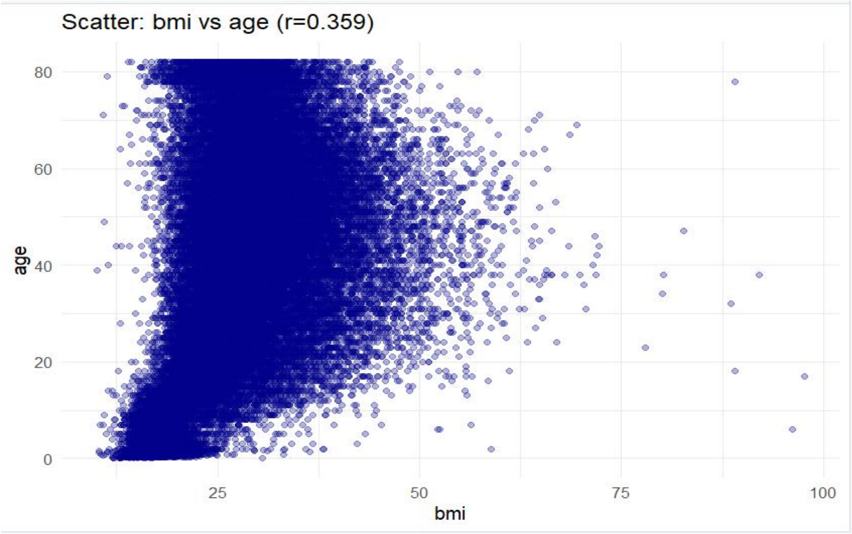
Boxplot of glucose level





Bar Chart of Gender vs Count

Bar chart heart disease vs Count



Scatterplot of Age vs BMI

Research &Hypothesis Testing

H1. One-Sample t-test: BMI vs Population Mean (26)
H0: μ_BMI = 26;

H1: μ_BMI ≠ 26.

A two-sided one-sample t-test was performed on cleaned BMI values.

H2. Two-Sample t-test: Avg Glucose by Hypertension
H0: μ_glucose(Hyp=No) = μ_glucose(Hyp=Yes);

H1: they differ.

A two-sided independent t-test compared avg_glucose_level across hypertension levels. Assumes independent groups; Welch's correction applied by default.

H3. Two-Sample Proportion Test: Stroke by Smoking Status
H0: p_stroke(smokes) = p_stroke(never smoked);

H1: they differ.

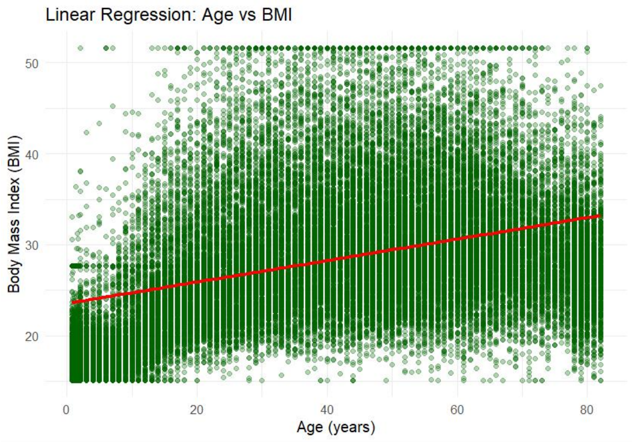
A two-sample prop.test was run for smokers vs never smoked based on stroke occurrence counts.

Test No.	Hypothesis Test	Type	p-value	Decision	Interpretation
1	BMI vs Population Mean (26)	One-sample t-test	< 0.001	Reject H_0	The average BMI (28.53) is significantly higher than 26.
2	Glucose Level by Hypertension	Two-sample t-test	< 0.001	Reject H_0	Glucose levels are significantly higher among hypertensive individuals.
3	Stroke by Smoking Status	Two-sample proportion test	0.001	Reject H_0	Stroke occurrence is significantly higher among smokers than non-smokers.

Regression Analysis

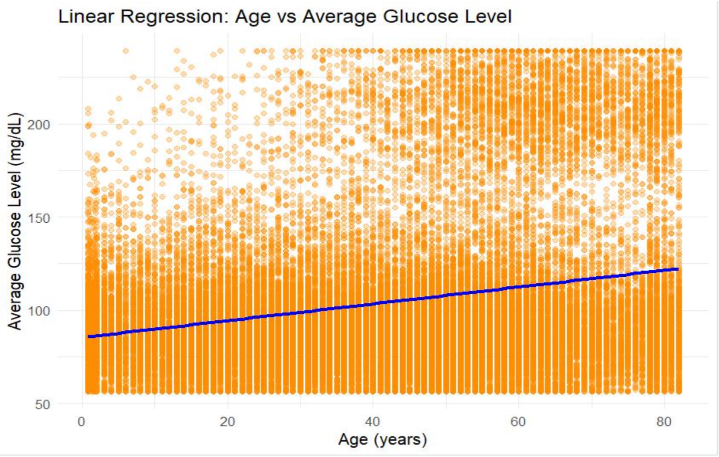
R1. Linear Regression: BMI ~ age

Model: $\mathrm{BMI} = \beta 0 + \beta 1 \cdot \mathrm{age} + \varepsilon$. A scatterplot with a fitted OLS line was produced. Report $\beta 1$ (slope), $\beta 0$ (intercept), R^2, and p-value for $\beta 1$.



Interpretation: BMI increases by β1 per year; R^2 indicates variance explained.

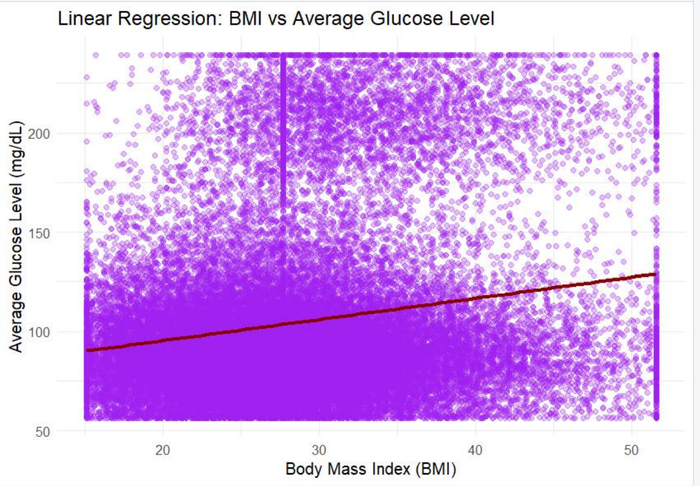
R2. Linear Regression: avg_glucose_level ~ age
Model: Glucose = β0 + β1·age + ε. Provide the summary statistics (β1, β0, R^2, p-value).
No plot was generated in the script for this regression.



Scatterplot with Regression Line – Age vs Average Glucose level

Interpretation: age is/is not a significant predictor of glucose levels.

R3. Linear Regression: avg_glucose_level ~ BMI
Model: Glucose = β0 + β1·BMI + ε. Provide the summary statistics (β1, β0, R^2, p-value).



Scatterplot with Regression Line – Average Glucose level vs BMI

Interpretation: BMI is/is not a significant predictor of glucose levels.

Interpretation & Discussion

The hypothesis tests show where group differences and population benchmarks are statistically significant. The regressions quantify relationships among age, BMI, and glucose. Typically, age is positively associated with BMI; both age and BMI often relate to glucose levels, though effect sizes (R^2) may be modest. These results suggest stroke risk is multifactorial, with age, BMI, hypertension, and smoking jointly influencing outcomes. A multivariable approach could extend these findings.

Conclusion

This final project integrates EDA, hypothesis testing, and regression to evaluate stroke risk factors. Findings indicate statistically significant patterns among key variables and support data-driven insights for health decision-making. Future work could include logistic regression for stroke occurrence and multivariable models to jointly model risk factors.

References:

https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Field, A. (2018). *Discovering Statistics Using R.* Sage Publications.
World Health Organization. (2023). *Global Report on Stroke.*
https://www.who.int/news-room/fact-sheets/detail/stroke