FINAL PROJECT: CEREBRAL STROKE RISK ANALYSIS

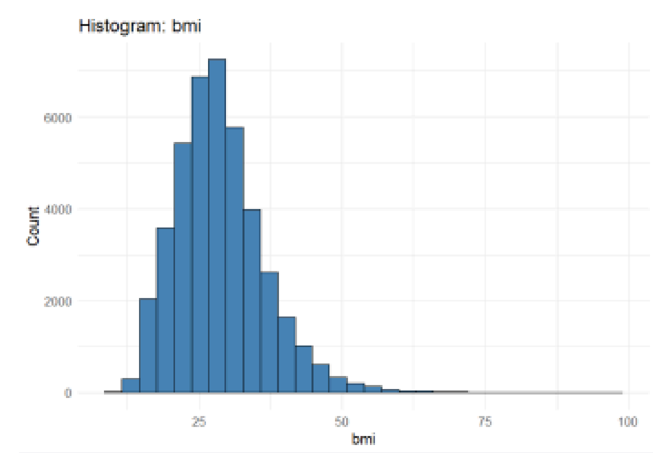
Submitted by: A S Sushmitha Urs
Course Number: 70954
Instructor Name: Amin Karimpour

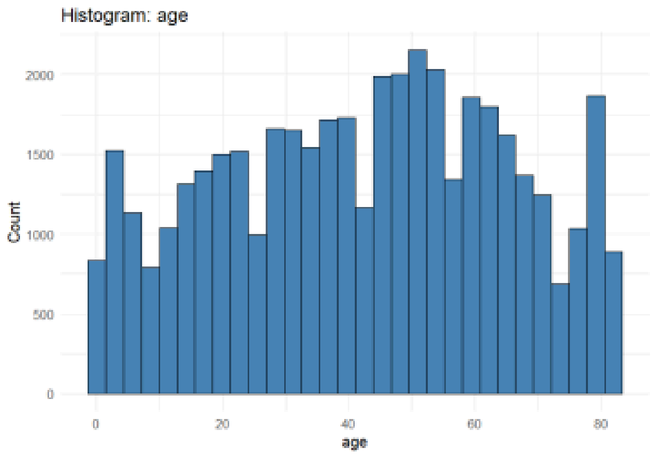
Introduction

Cerebral Stroke is a leading cause of death and long-term disability worldwide, driven by multiple risk factors such as hypertension, obesity, smoking, and age. Understanding these interactions is crucial for prevention and early detection. This study uses exploratory data analysis (EDA), hypothesis testing, and regression modeling to identify key predictors and patterns influencing stroke risk.

The dataset includes 43,400 observations and 12 variables, combining demographic (gender, age, marital status, residence), lifestyle (smoking status, work type), and medical (hypertension, heart disease, glucose, BMI) factors.Data preprocessing involved converting binary fields to Yes/No categories, imputing missing values (median for numeric, mode for categorical), winsorizing outliers, and removing duplicates

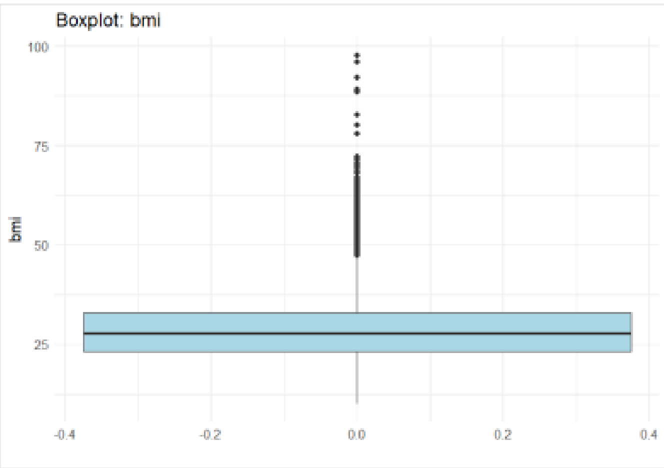
Exploratory Data Analysis

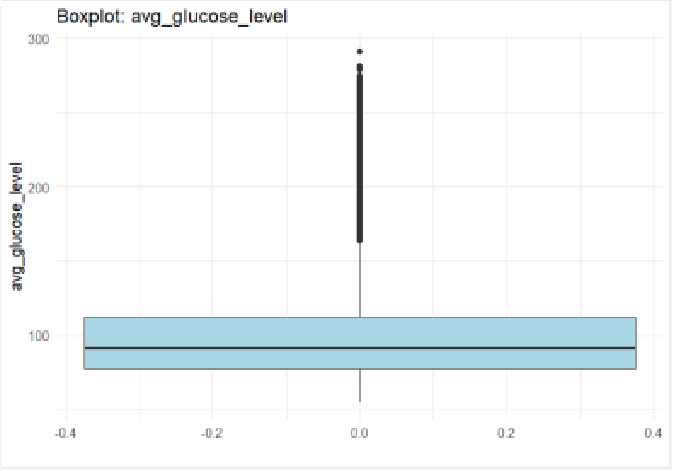




Histogram of BMI

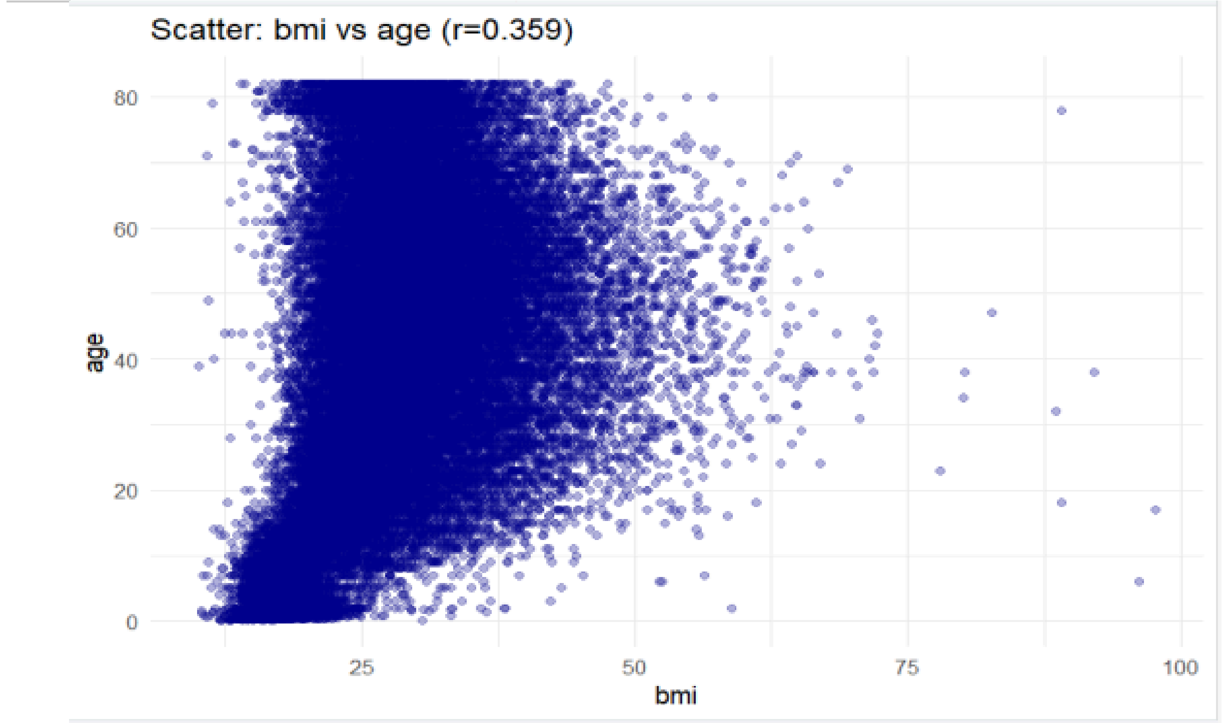
Histogram of Age

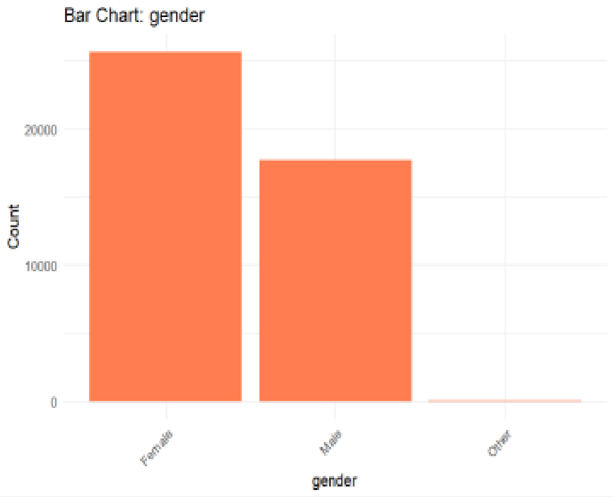


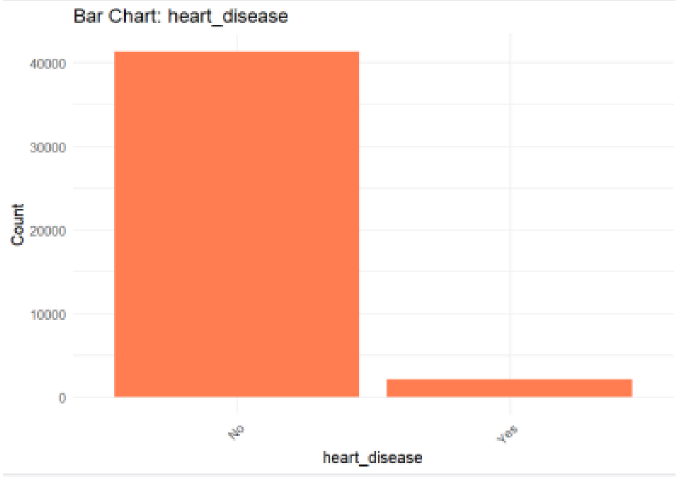


Boxplot of BMI

Boxplot of glucose level







Bar Chart of Gender vs Count

Bar chart heart disease vs Count

Hpothesis Testing

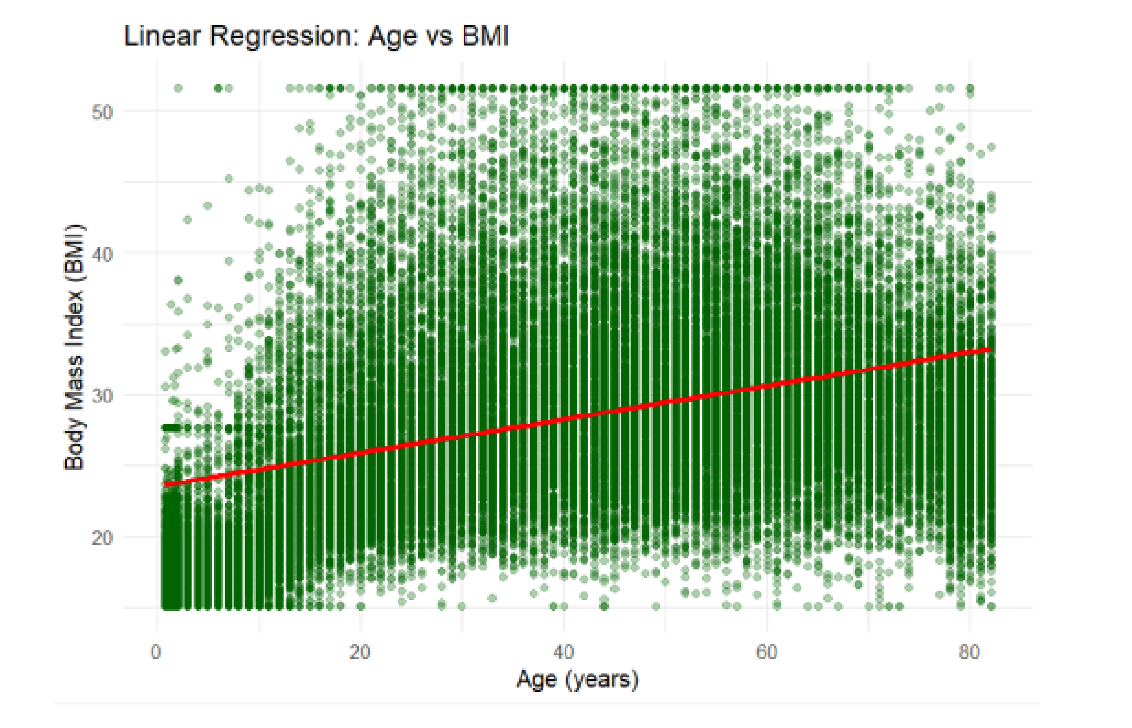
Test No.	Hypothesis Test	Type	p-value	Decision	Interpretation
1	BMI vs Population Mean (26)	One-sample t-test	< 0.001	Reject H_0	The average BMI (28.53) is significantly higher than 26.
2	Glucose Level by Hypertension	Two-sample t-test	< 0.001	Reject H_0	Glucose levels are significantly higher among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

To examine linear relationships among Age, BMI, and Average Glucose Level

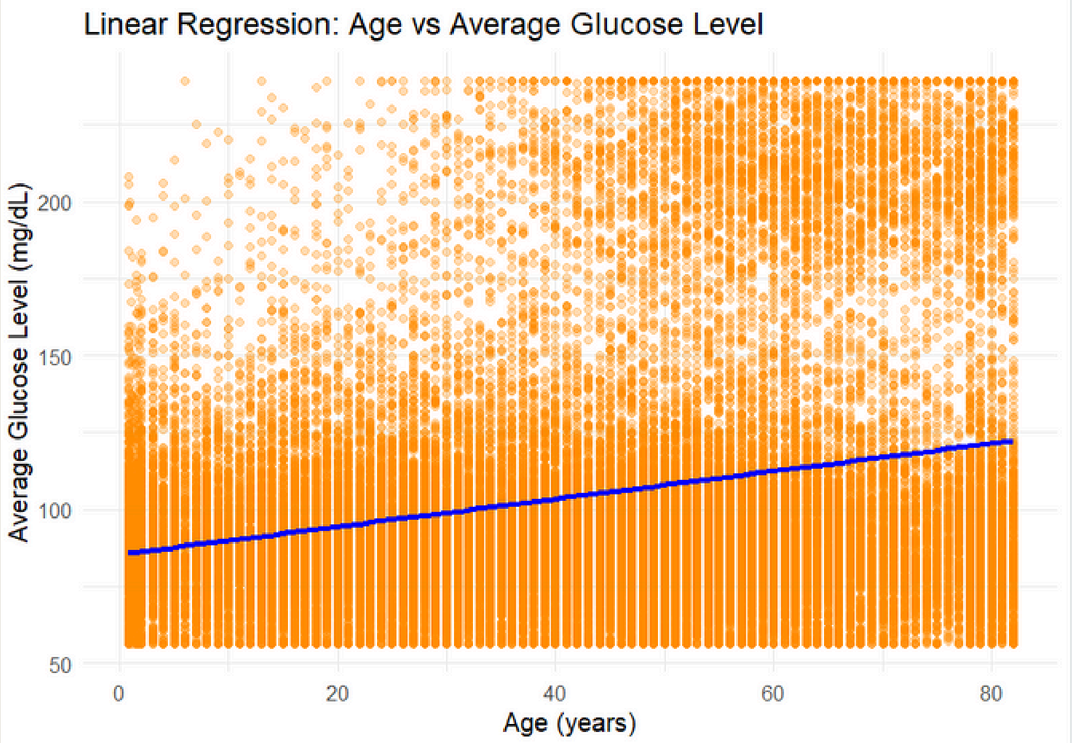
BMI vs Age

- BMI increases slightly with age.
- Indicates a positive association ($\beta_1 > 0$).
- R^2 suggests moderate variance explained.



Average Glucose ~ Age

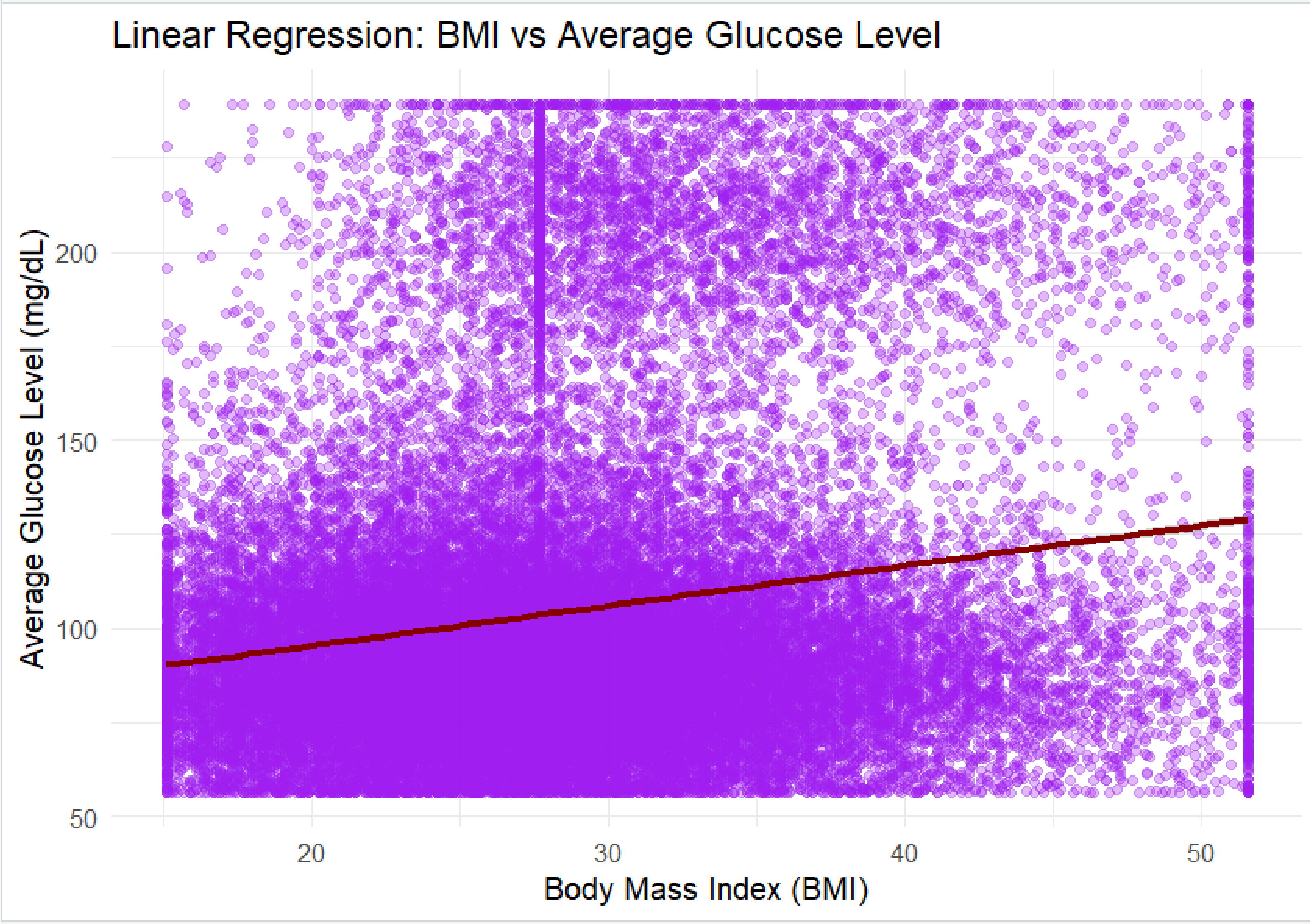
- Glucose tends to rise with age, though the relationship is weaker.
- Age is a mild predictor of glucose level.



Regression Analysis

Average Glucose ~ BMI

- Higher BMI is linked to higher glucose levels.
- BMI shows a stronger relationship with glucose than age does.



Conclusion

This final project integrates exploratory data analysis (EDA), hypothesis testing, and regression modeling to evaluate key factors influencing stroke risk. The findings reveal statistically significant relationships among variables such as age, BMI, glucose level, hypertension, and smoking status, emphasizing the multifactorial nature of stroke. These insights highlight the importance of early risk identification and preventive healthcare measures. Future work could involve developing logistic regression or multivariable predictive models to improve stroke prediction accuracy and guide targeted public health interventions.

Thank You