ZEWAIL CITY
UNIVERSTY OF SCIENCE AND TECHNOLOGY

Analyzing Diabates Dataset

Amr Yasser,
Omar Hazem
Mohamed Mourad,
Youssef Mohammad,
Hady Emad

DSAI 307: Statistical Inference

Agenda

01 Introduction

02 Data Quality

03 Exploratory Data Analysis (EDA)

04 Key Questions & Risk Profiles

05 Hypothesis Testing

06 Simulation Study

07 Conclusion

01 Introduction

The primary goal is to analyze:

- Diabetes contributing **factors**
- Health **Trends**

Provided data:

- Adapted from **governmental** sources
- Sized **750**+ rows X **9** features

02 Data Quality

- Unlogical **zeros** handling → nan
- **Duplicates** check
- Mean **Imputation**

03 EDA

Insights

- Imbalance (fig.1)

- **Weak correlation** between *glucose* and *BMI*

- **Direct correlation** between *age* and *glucose* (fig.2)

03 EDA

Insights

- *Glucose* in diabetic patients is **always higher** (fig.2)
- *BMI* is almost **normally distributed** (fig.3)
- *DPF* has **outliers** (fig.4)

fig.1

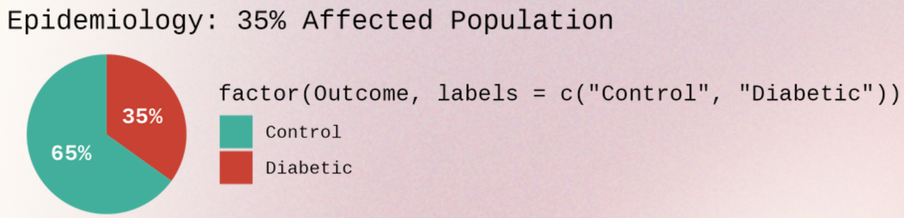

Epidemiology: 35% Affected Population
35%
65%
factor(Outcome, labels = c("Control", "Diabetic"))
Control
Diabetic

fig.2

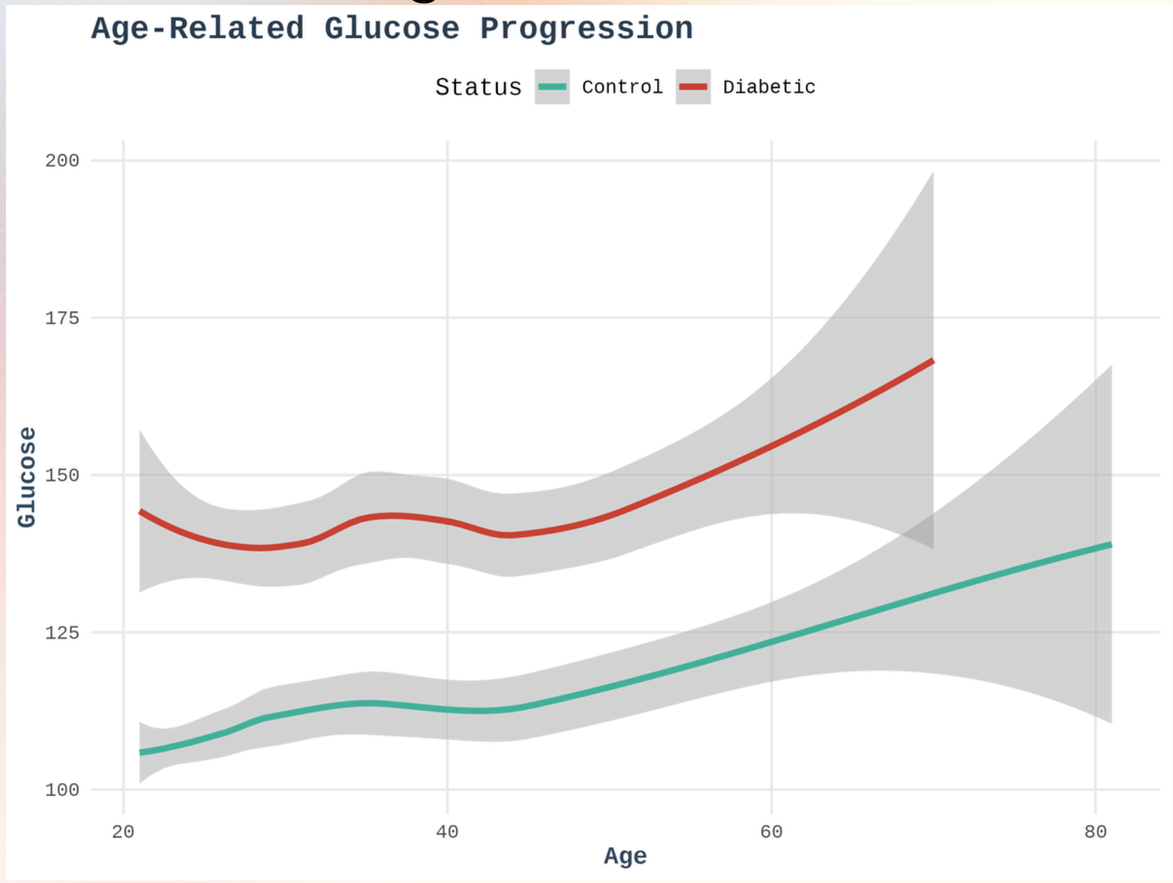

Age-Related Glucose Progression
Status
Control
Diabetic
200
175
150
125
100
Glucose
20
40
60
80
Age

fig.3

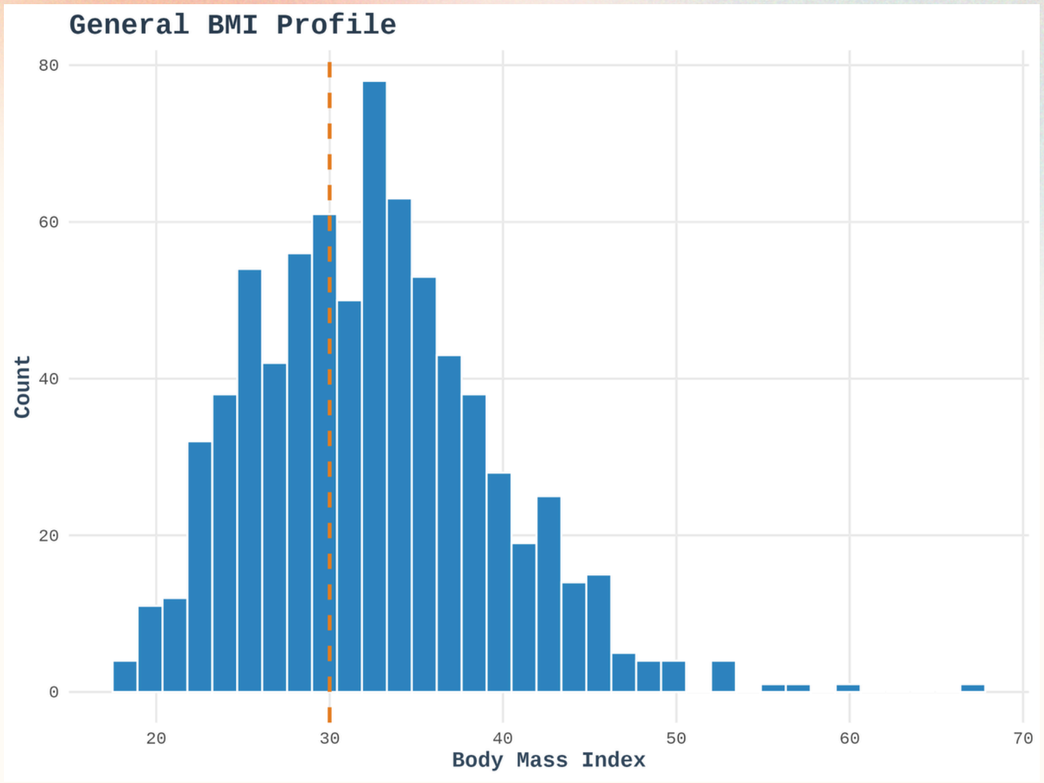


fig.4

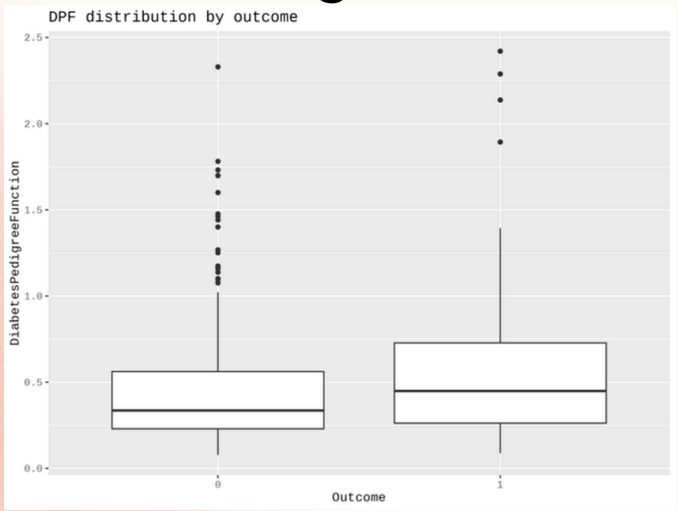
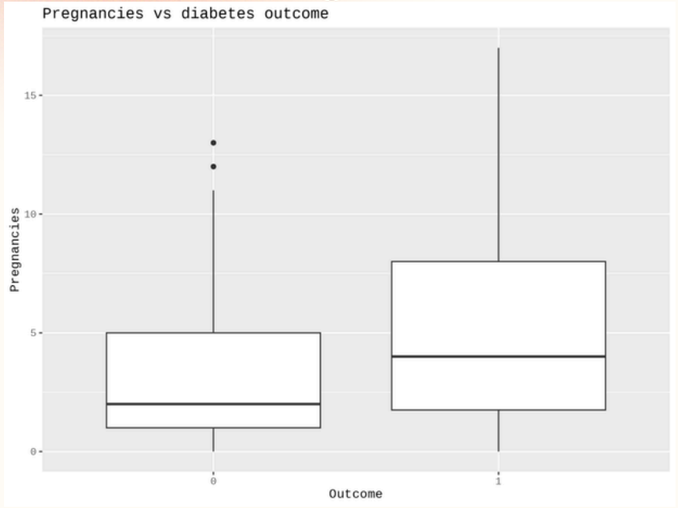


fig.5



04 Key Questions

Q1. Diabetes-Glucose Direct Correlation

Yes. Visualized in fig.2 (Task 1), glucose levels are **higher** in diabetic group. (fig.6)

Q2. Glucose-BMI Direct Correlation

A **positive** correlation exists at **~23%. (**fig. 8)

Q3. Pregnancies-Diabetes Risk Correlation

As shown in Task 8, the incidence of diabetes **increases steadily** with the number of pregnancies recorded. (fig.7)

04 Key Questions

Q4. Age vs Insulin and Glucose

Glucose shows a **strong upward** trend with age. (fig.10)

Insulin shows more **peak dynamics** in **middle-aged** groups.

Q5. Risk Profiles

Healthy:

Age **< 30**, *Glucose* **< 110**, *BMI* **< 30**.

High Risk:

Age **> 35**, *Glucose* **> 140**, *BMI* **> 34**.

fig.6

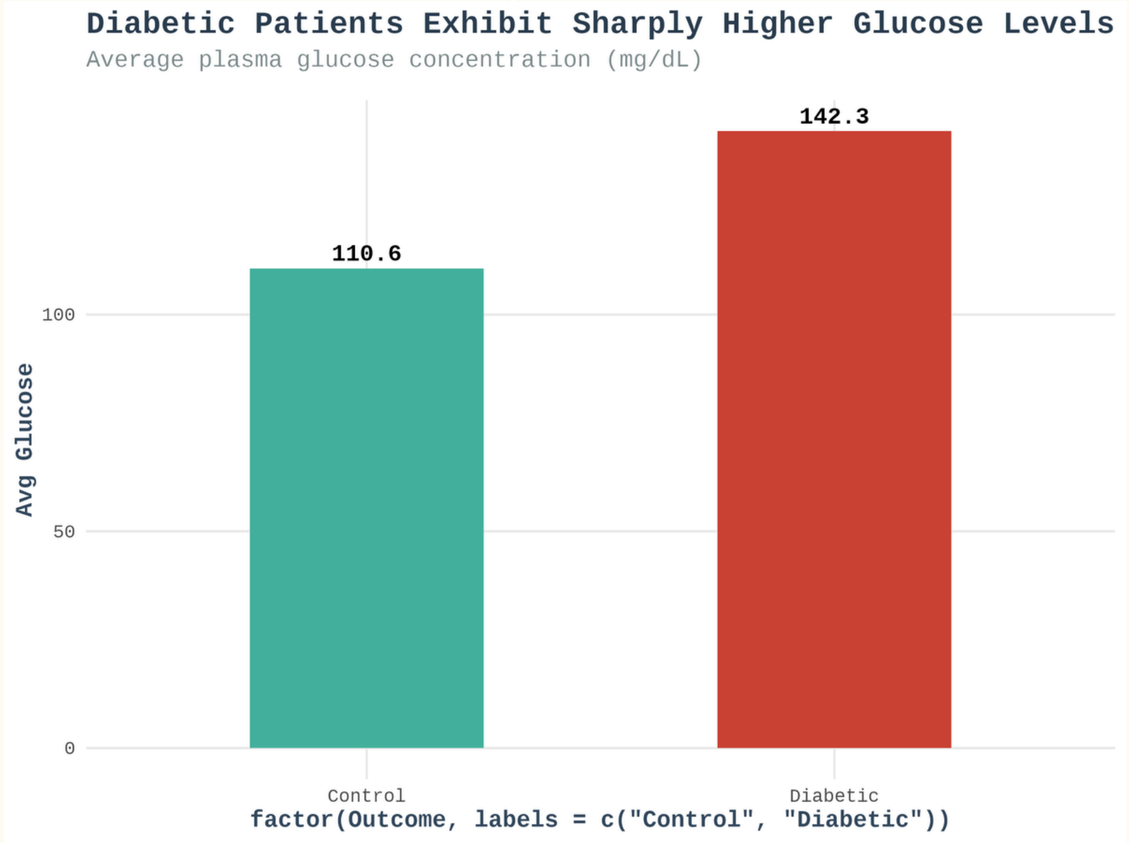

Diabetic Patients Exhibit Sharply Higher Glucose Levels
Average plasma glucose concentration (mg/dL)
142.3
110.6
100
50
0
Avg Glucose
Control
Diabetic
factor(Outcome, labels = c("Control", "Diabetic"))

fig.7

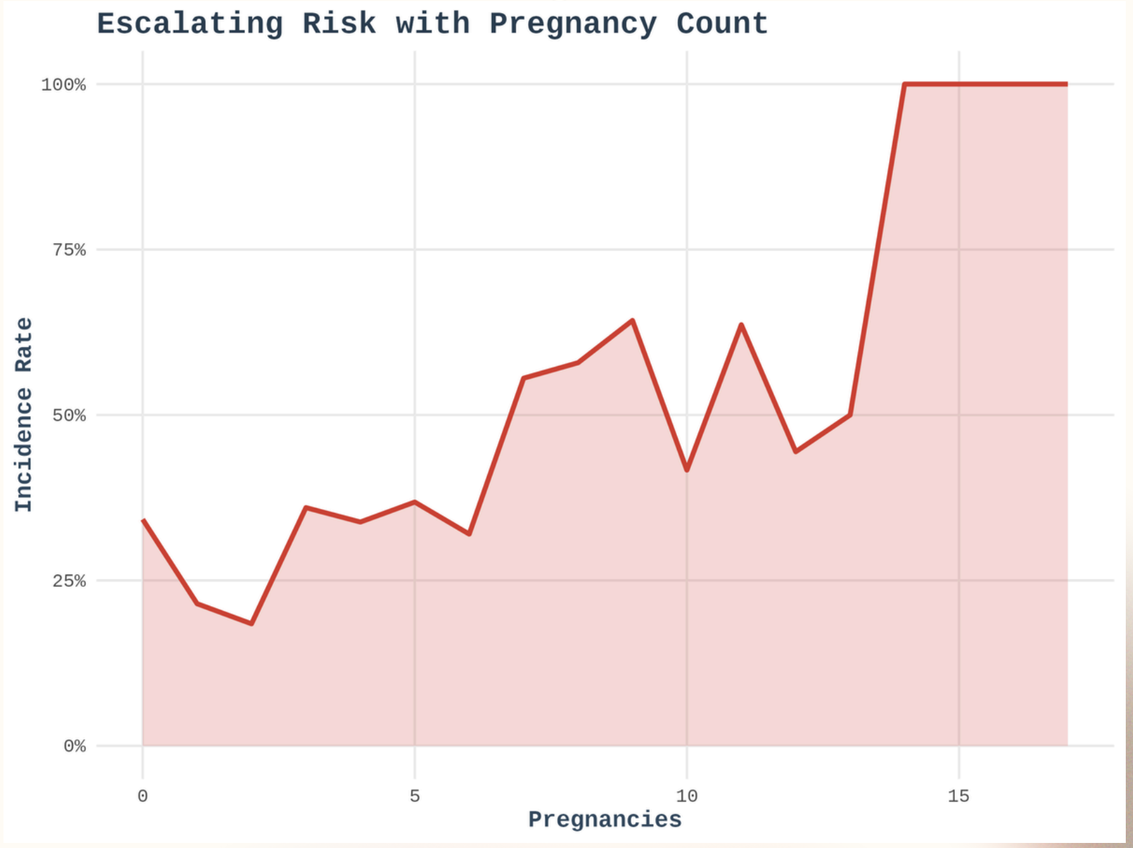

Escalating Risk with Pregnancy Count
100%
75%
50%
25%
0%
Incidence Rate
0
5
10
15
Pregnancies

fig.8

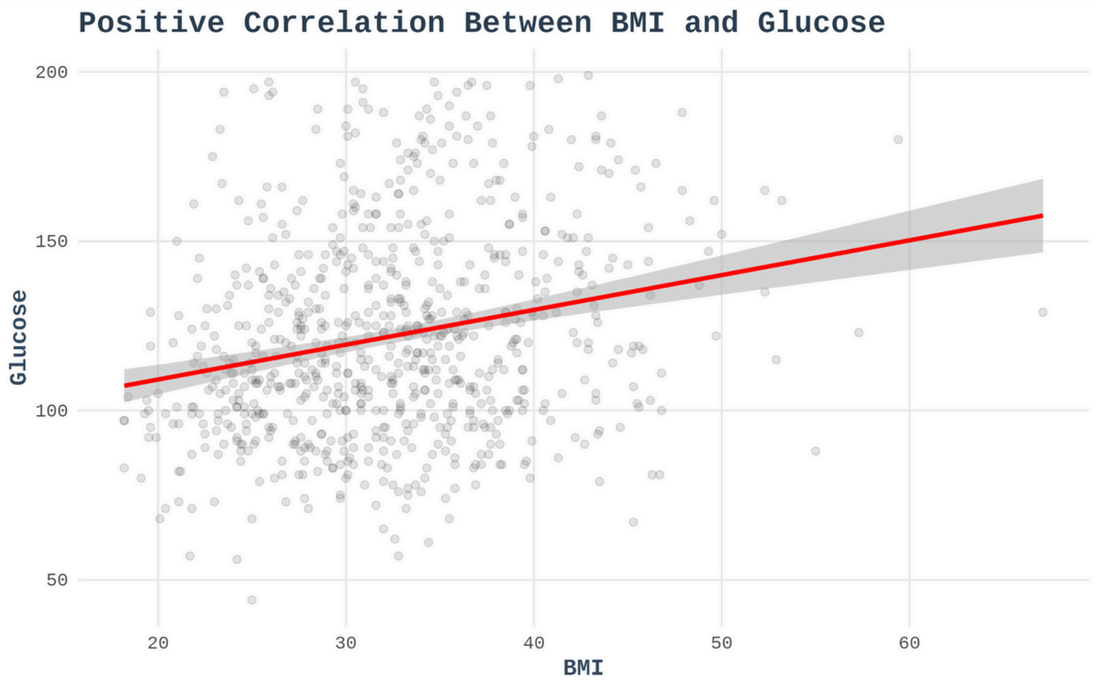

Positive Correlation Between BMI and Glucose
Glucose
200
150
100
50
20
30
40
50
60
BMI

fig.9

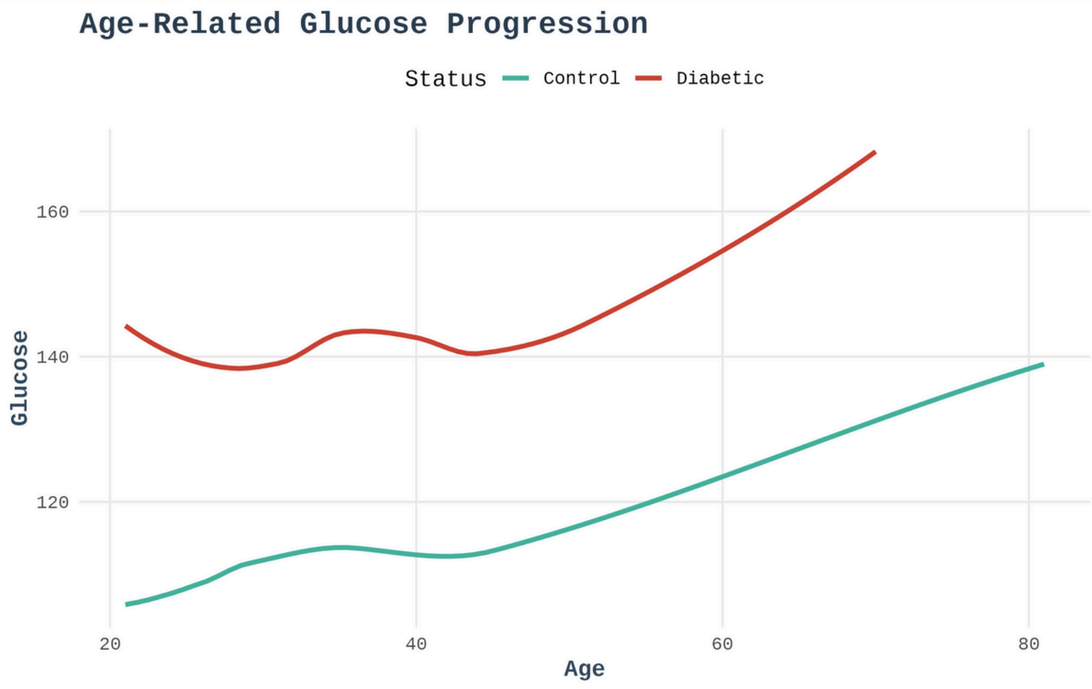

Age-Related Glucose Progression
Status
Control
Diabetic
Glucose
160
140
120
20
40
60
80
Age

fig.10

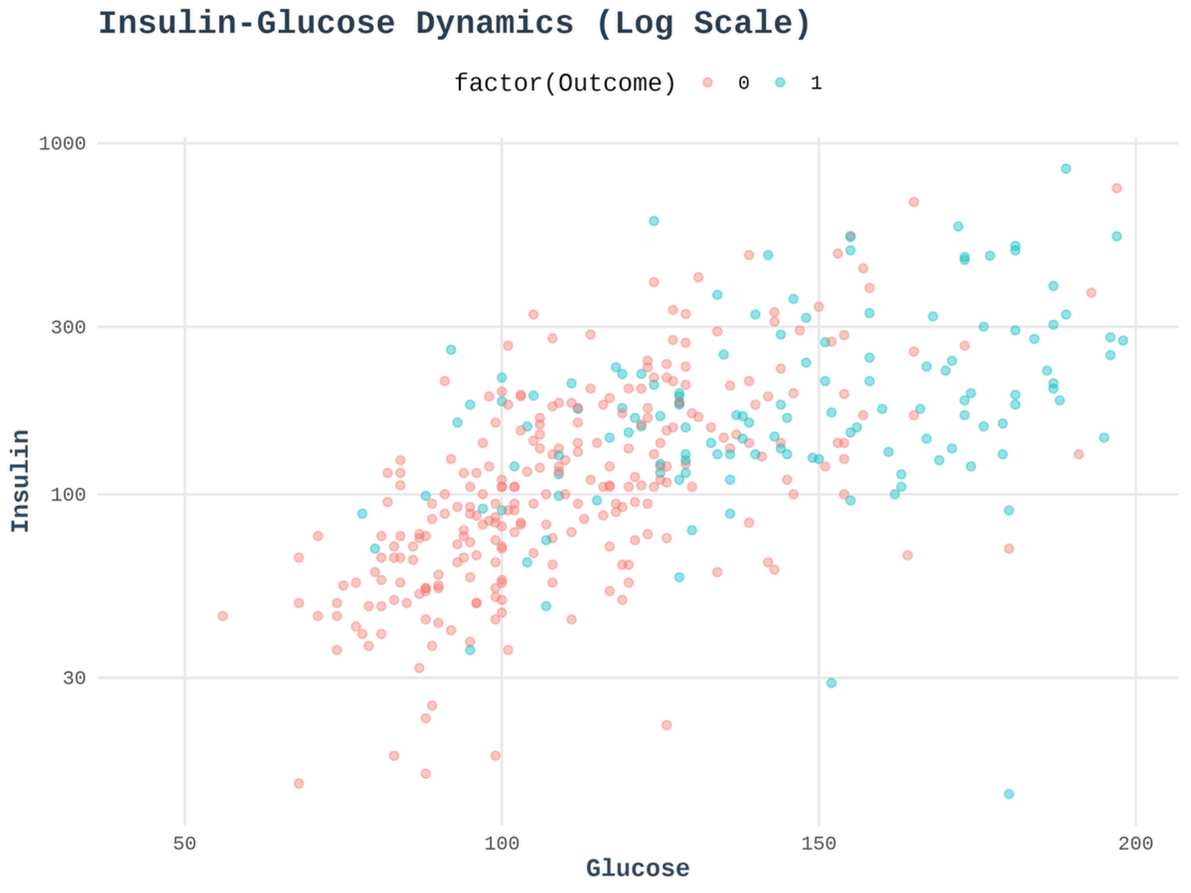
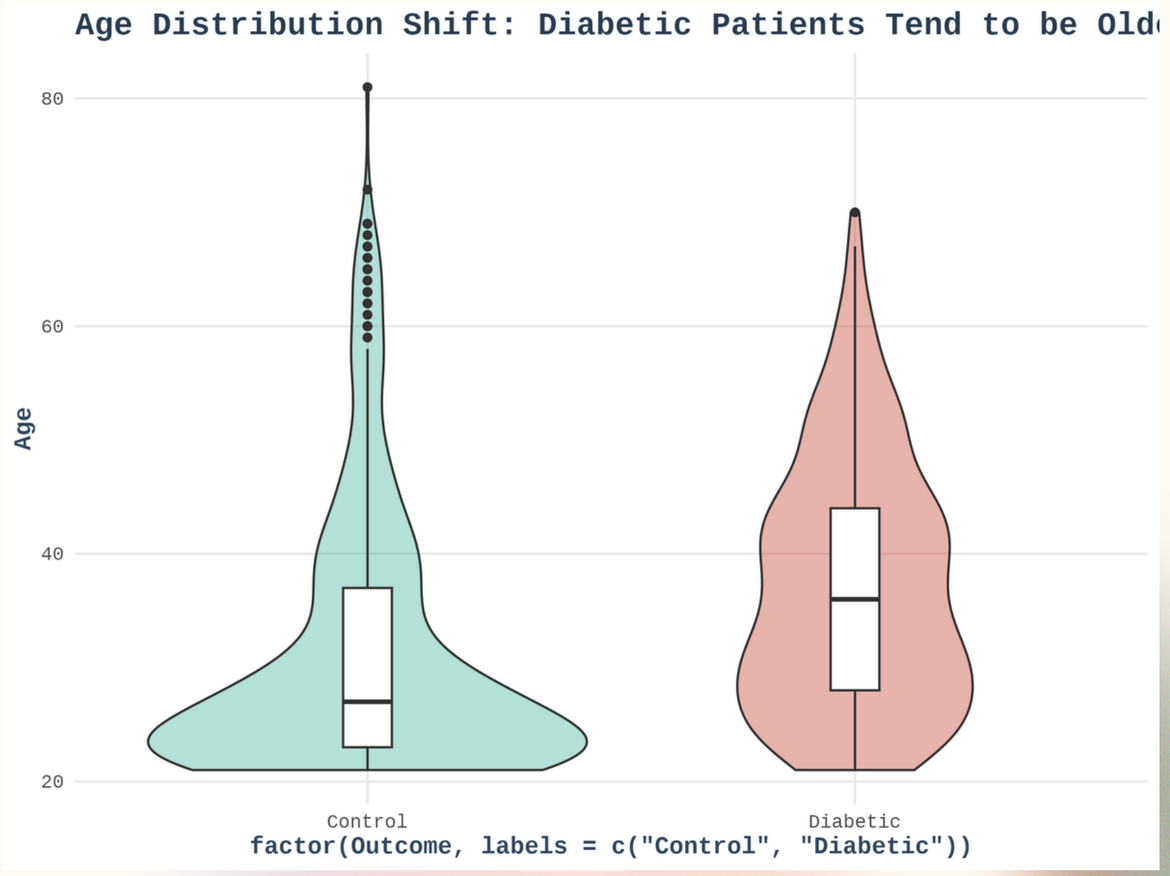

Insulin-Glucose Dynamics (Log Scale)
factor(Outcome) 0 1
1000
300
100
30
Insulin
50
100
150
200
Glucose

fig.11


Age Distribution Shift: Diabetic Patients Tend to be Old
80
60
40
20
Age
Control
Diabetic
factor(Outcome, labels = c("Control", "Diabetic"))

04 Key Questions

Additional **+5 Qs** are present in the report

Supported with visualizations and answered

Many more at:

Full Report and GitHub Repository

05 Hypothesis Testing

Claim 1:

Glucose Differences across groups: diabetic and non-diabetic patients

Hypotheses:

- H0 (Null): Mean *glucose* is **equal** across groups.
- H1 (Alternative): Mean *glucose* is **not equal** across groups.

Results: Using Two-sample T-Test

- **P-value < 0.05** (specifically ~0.0).
- **Rejects null** hypothesis

05 Hypothesis Testing

Claim 2:

***BMI* Differences** across groups: diabetic and non-diabetic patients

Hypotheses:

- H0 (Null): Mean *BMI* is **equal** across groups.
- H1 (Alternative): Mean *BMI* is **not equal** across groups.

Results: Using Two-sample T-Test

- **P-value < 0.05** (specifically ~10^ -16).
- **Rejects null** hypothesis

O6 Simulation Study

- Selected ***BMI*** (normally distributed)
- **population mean** = ~ *32.5*
- Applied 3 experiments

Experiment A: n = 15

- Generated **25** random samples
- Confidence interval of 95%

Results:

- Successfully captured the true population mean.

06 Simulation Study

Experiment B: n = 100

- Generated **25** random samples
- **Observed:** decreased **confidence width**

Results:

- led to narrower, more precise intervals.
- reducing uncertainty.

06 Simulation Study

Experiment C: n = 10

- Generated **20** random samples

Results:

- **Wider** Intervals.
- **More variable accuracy** in containing the true mean.

Thank You !