Debora Bimbi
Ironhack Machine
Learning Bootcamp
Third Week

# Diabetes Dataset Analysis

## What health factors matter for diabetes?

Agenda:

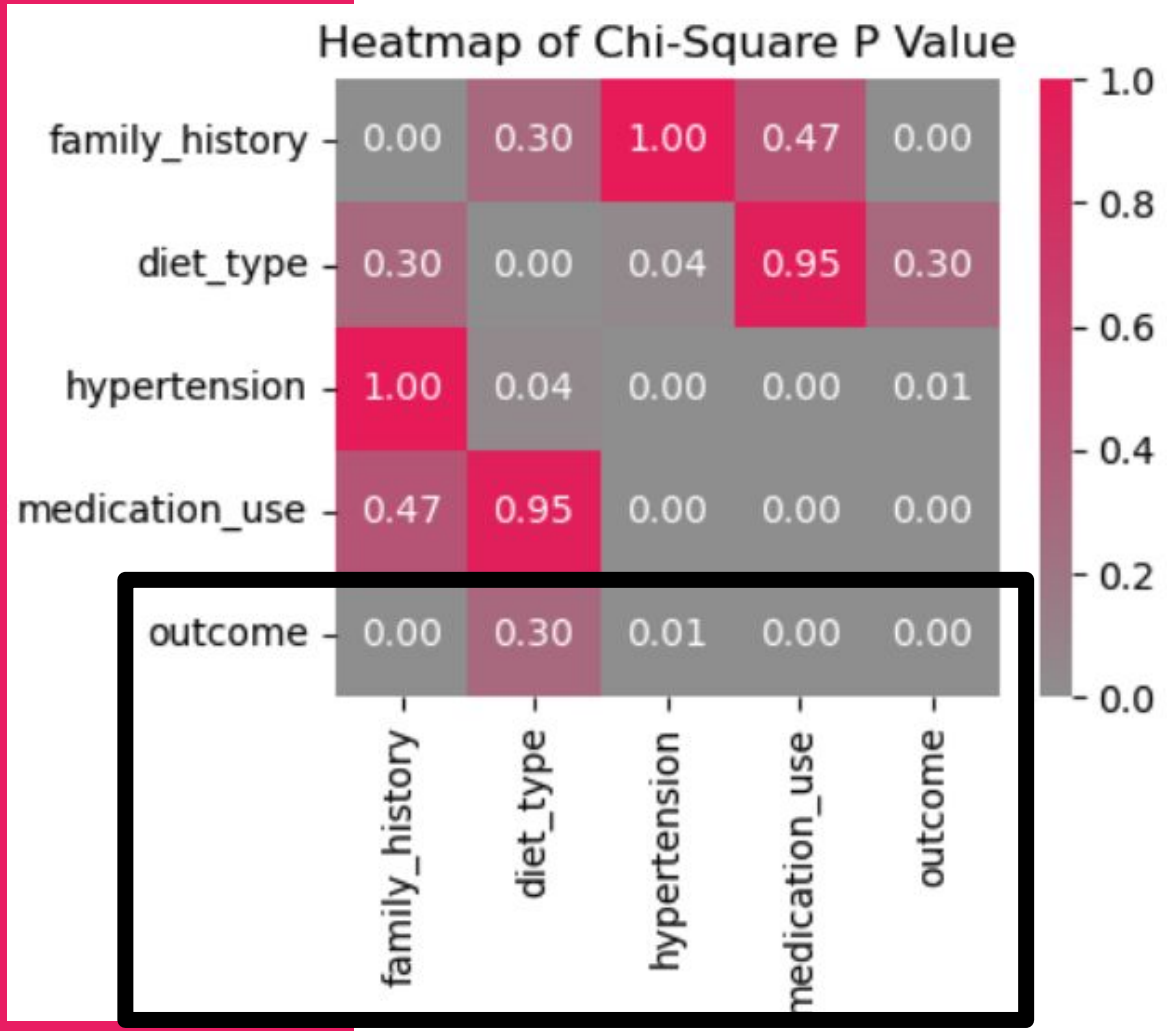
1. Introduce the dataset
2. Categorical Analysis
3. Numerical Analysis
4. Conclusion

ldl
bmi
hip_circumference
waist_circumference
family_history
hb_a1c
whr
diet_type
hypertension
triglycerides
hdl
medication_use
glucose
blood_pressure

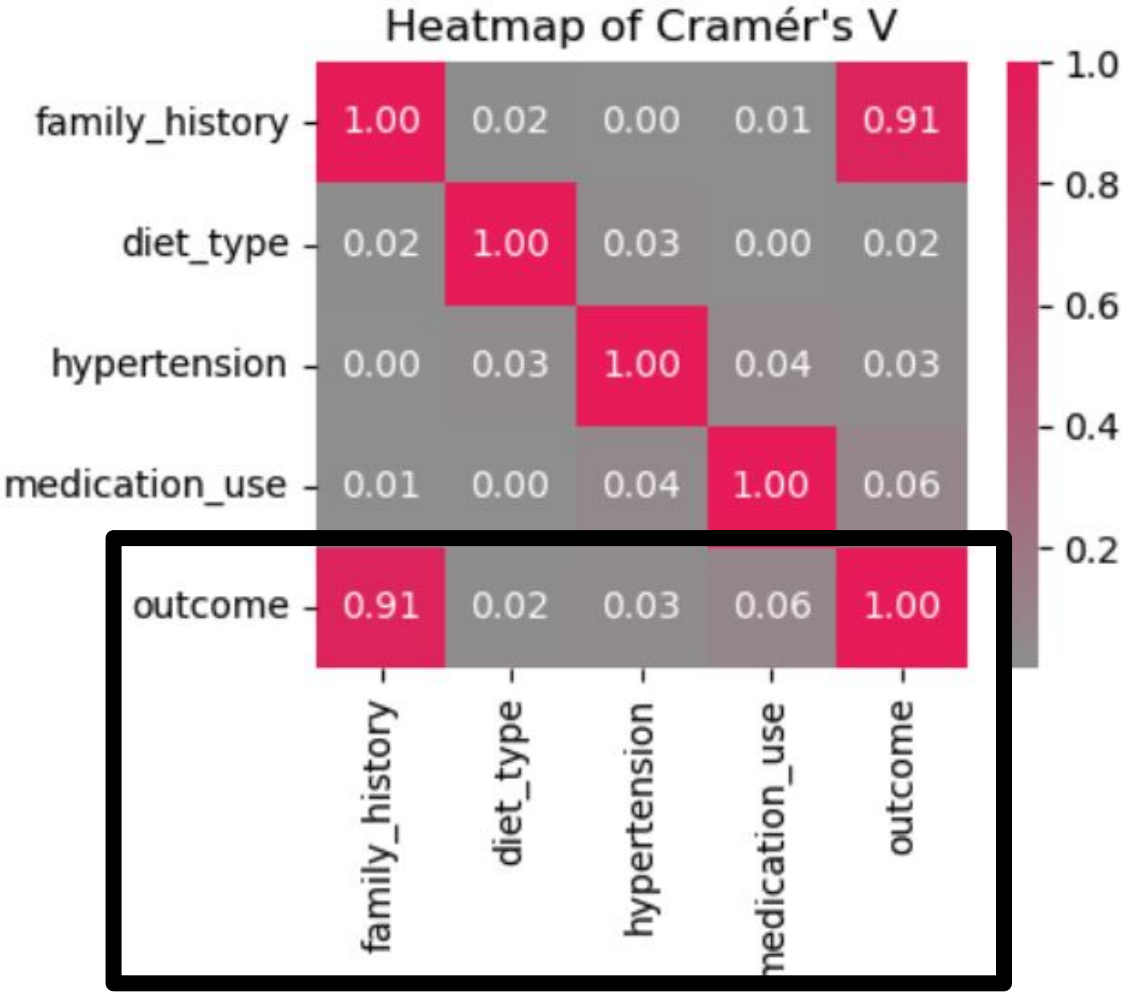
The Dataset:

- from www.kaggle.com
- 9538 patient entries
  - ~34% have diabetes (3282)
  - ~66% have no diabetes (6256)
- 17 columns check health factors, lifestyle and genetic factors for diabetes

# Categorical Analysis [Chi-square p]



# Categorical Analysis [Cramér's V]



family_history
outcome

# Chi-square p value is 0
# Cramér's Value is 0.91

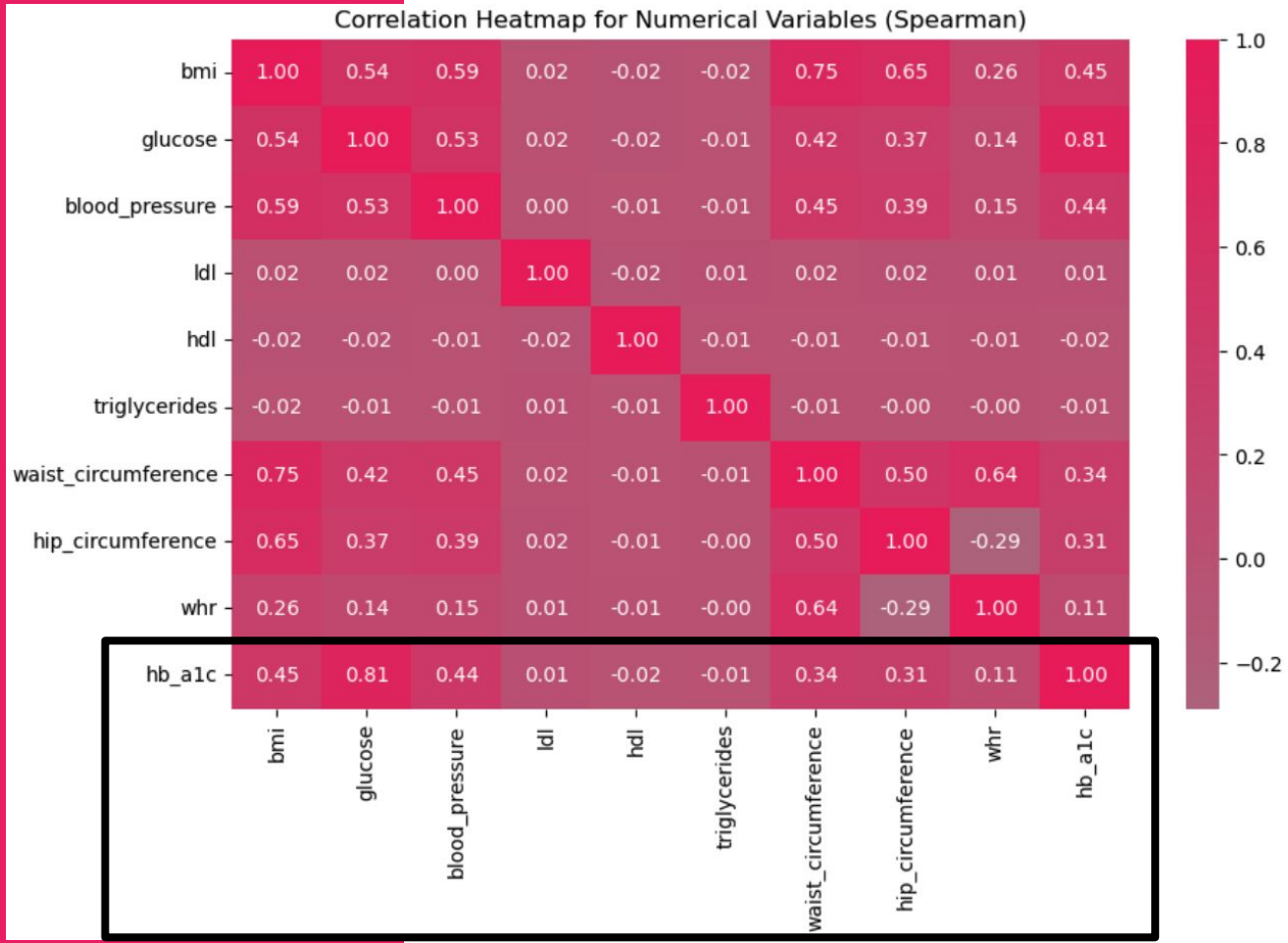
→ We can reject H0 that the two values are independent

→ There is a strong association between both values

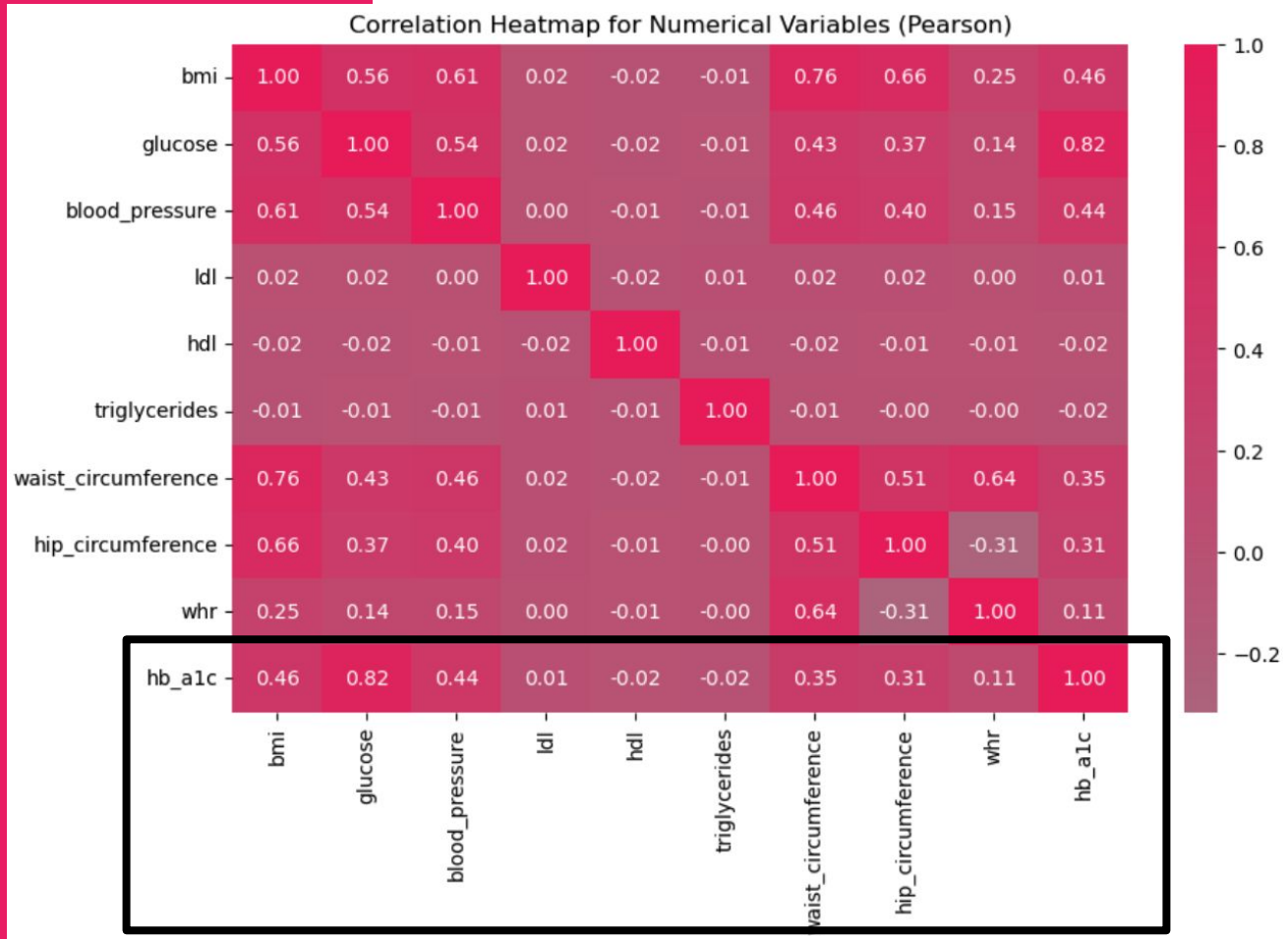
# family_history
# outcome

| Diabetes | Family History | Absolute Frequency | Relative Frequency |
|---|---|---|---|
| False | False | 6256 | 100.0 |
| True | False | 397 | 12.1 |
| | True | 2885 | 87.9 |

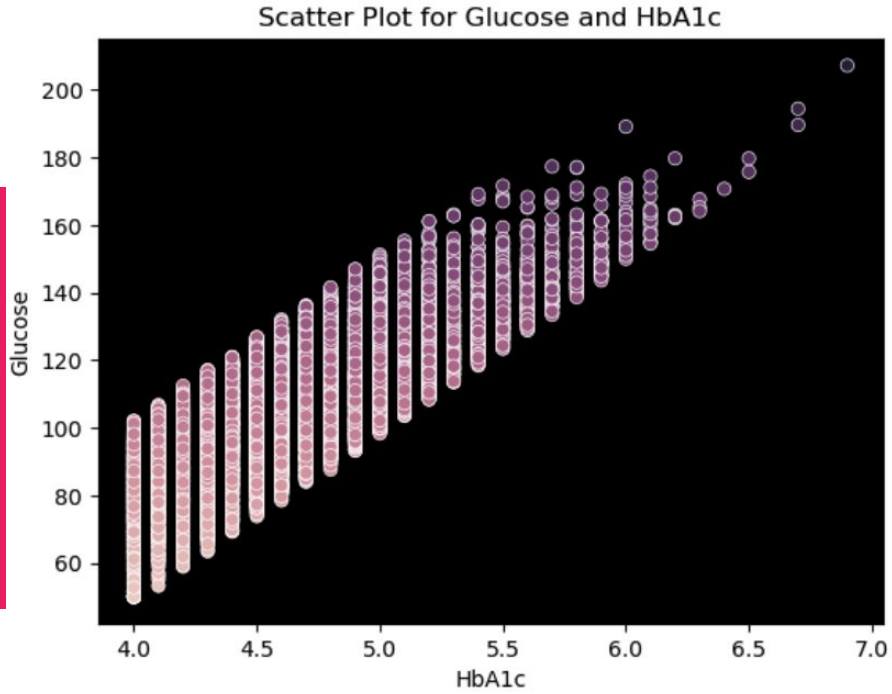
# Numerical Analysis [Spearman]



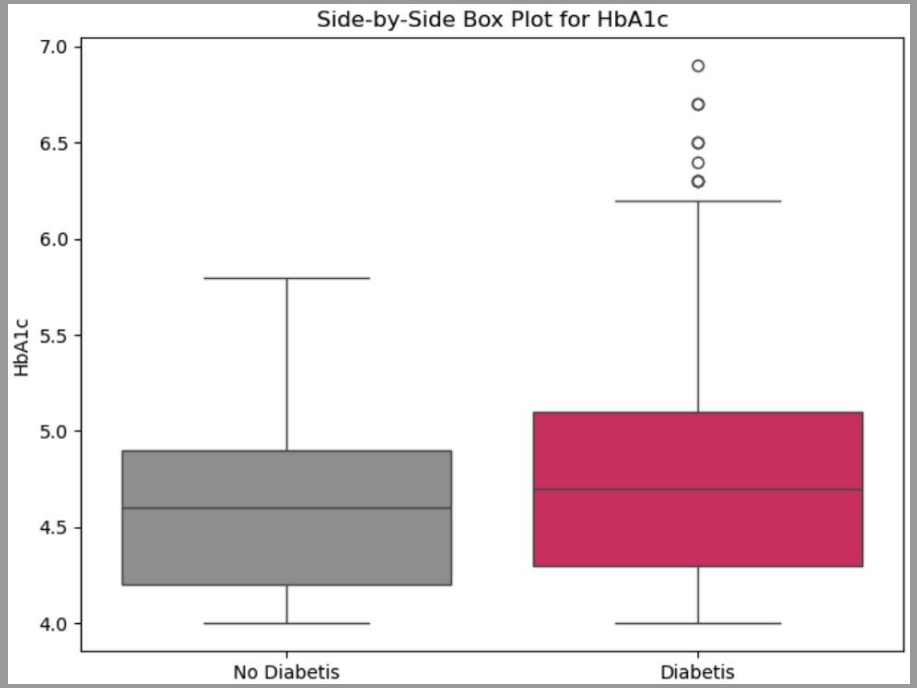
# Numerical Analysis [Pearson]

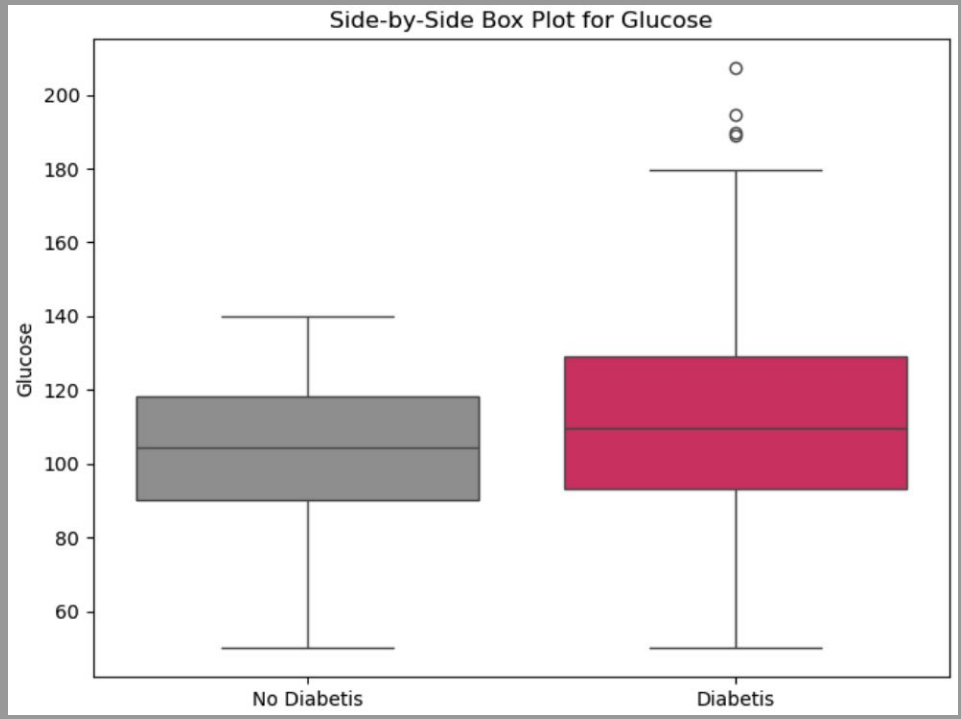


# HbA1c
# Glucose



# HbA1c




# GLUCOSE

# CONCLUSION

1. Only Family History is strongly associated with having Diabetes
2. Only Glucose is correlated with long term blood sugar HbA1c

# My Wish

A more differentiated outcome with Diabetes Type I or II, Pre-Diabetic etc.

A binary column did not allow to collect many results.

# Thank You