Diabetes Data Analysis

Predicting hemoglobin A1c / diabetes diagnosis

Gage Benne

The Data Science Process

A guided analysis framework

1. ASK
2. GET
3. EXPLORE
4. MODEL
5. REPORT

ASK

Problem Statement

Using the CoNVO framework

- **Co**ntext - what is the context?
- **N**eed - what organizational need requires fixing?
- **V**ision - what is required and what does success look like?
- **O**utcome - how will the result work itself back into the organization?

The goal is to produce a model predicting hemoglobin A1c measurements based on various basic health metrics.

Problem statement

GET

The Dataset

diabetes.csv

- Flat file sourced from the Vanderbilt University Department of Biostatistics
- 19 variables
- 403 individuals from 1046 subjects
- Primarily individuals from counties in Virginia

VANDERBILT
UNIVERSITY

Extract, Transform, Load

Cleaning the data

- Remove index
- Remove location
- Missing data
  - Consolidate four blood pressure readings into two
  - Keep rows with missing data

Index

Location

Systolic blood pressure (1st)

Diastolic blood pressure (1st)

Extract, Transform, Load

Cleaning the data

- Remove index
- Remove location
- Missing data
  - Consolidate four blood pressure readings into two
  - Keep rows with missing data

🚫 Index

🚫 Location

🚫 Systolic blood pressure

🚫 Diastolic blood pressure

Variables

Ready for analysis

Total cholesterol

Stabilized glucose

High-density lipoprotein

Cholesterol ratio

Glycosylated hemoglobin

Age

Gender

Height

Weight

Frame

Systolic blood pressure

Diastolic blood pressure

Waist

Hips

Lab postprandial time

Variables

Ready for analysis

Total cholesterol

Stabilized glucose

High-density lipoprotein

Cholesterol ratio

Glycosylated hemoglobin

Age

Gender

Height

Weight

Frame

Systolic blood pressure

Diastolic blood pressure

Waist

Hips

Lab postprandial time

> 6.4%

Positive diagnosis of diabetes

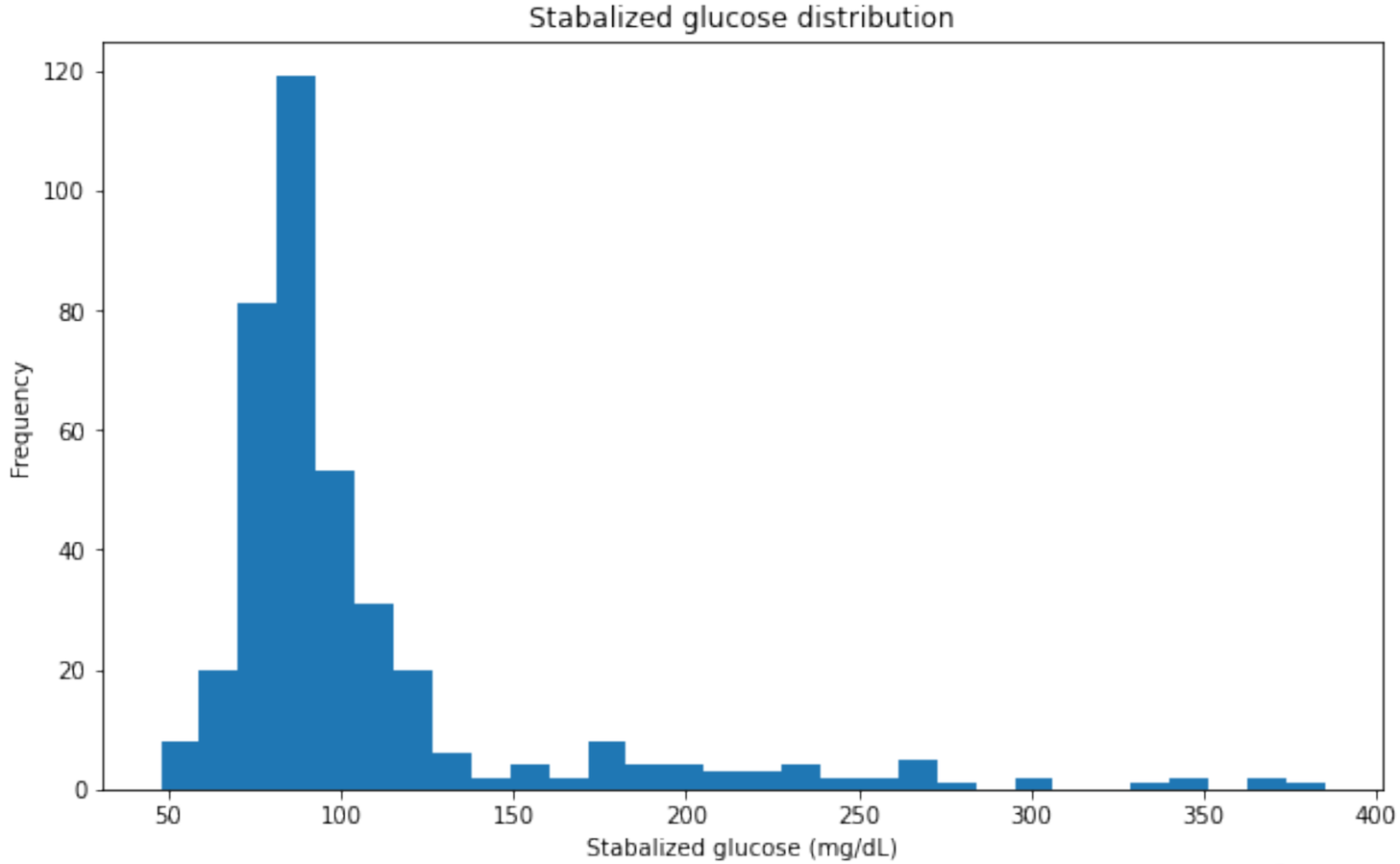
5.7 - 6.4%

pre-diabetes

EXPLORE

Highlights

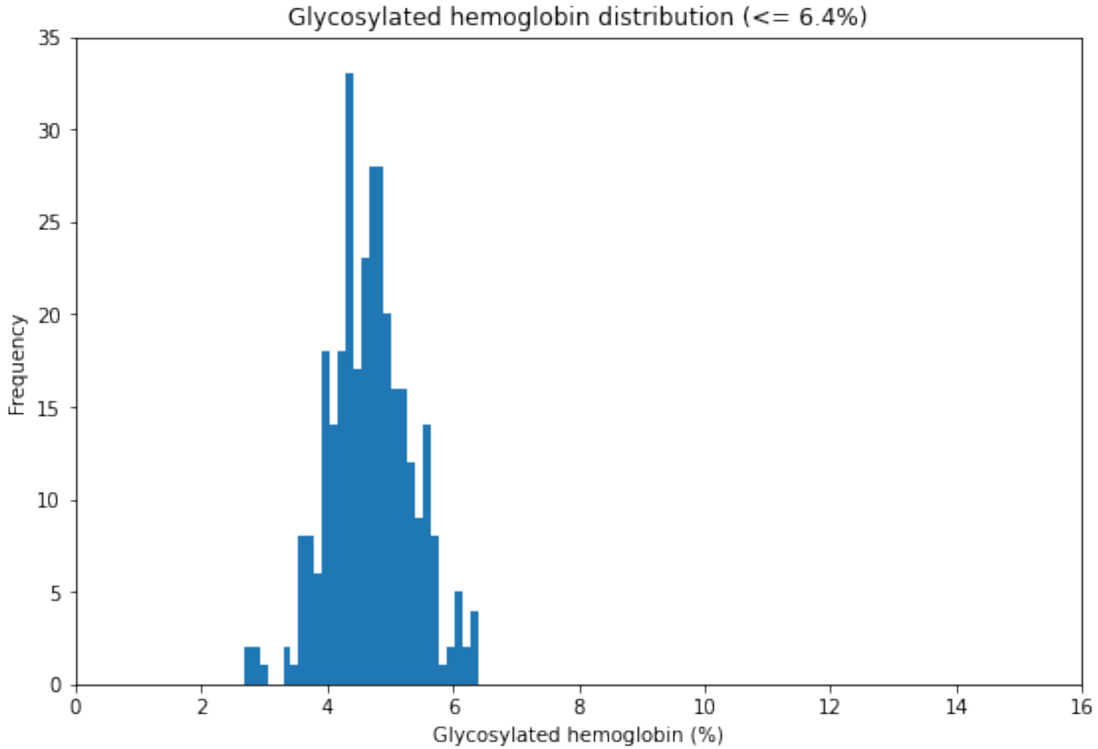
Single variable exploratory data analysis

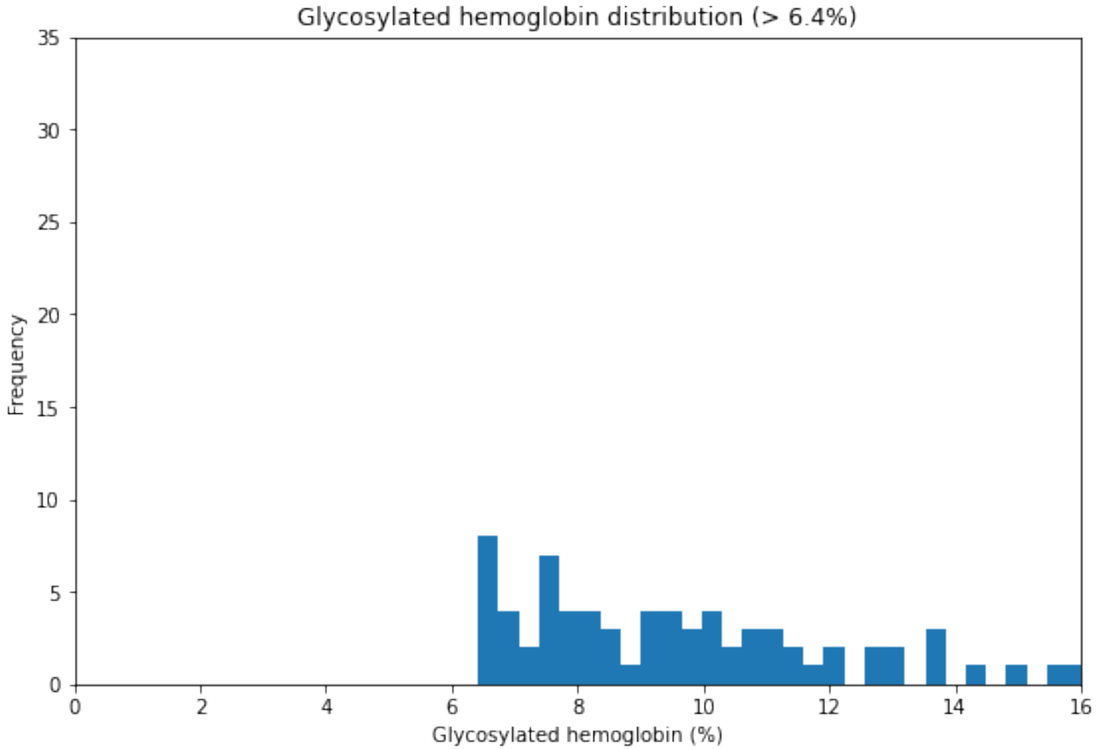


🍭 Stabilized glucose

Highlights

Single variable exploratory data analysis

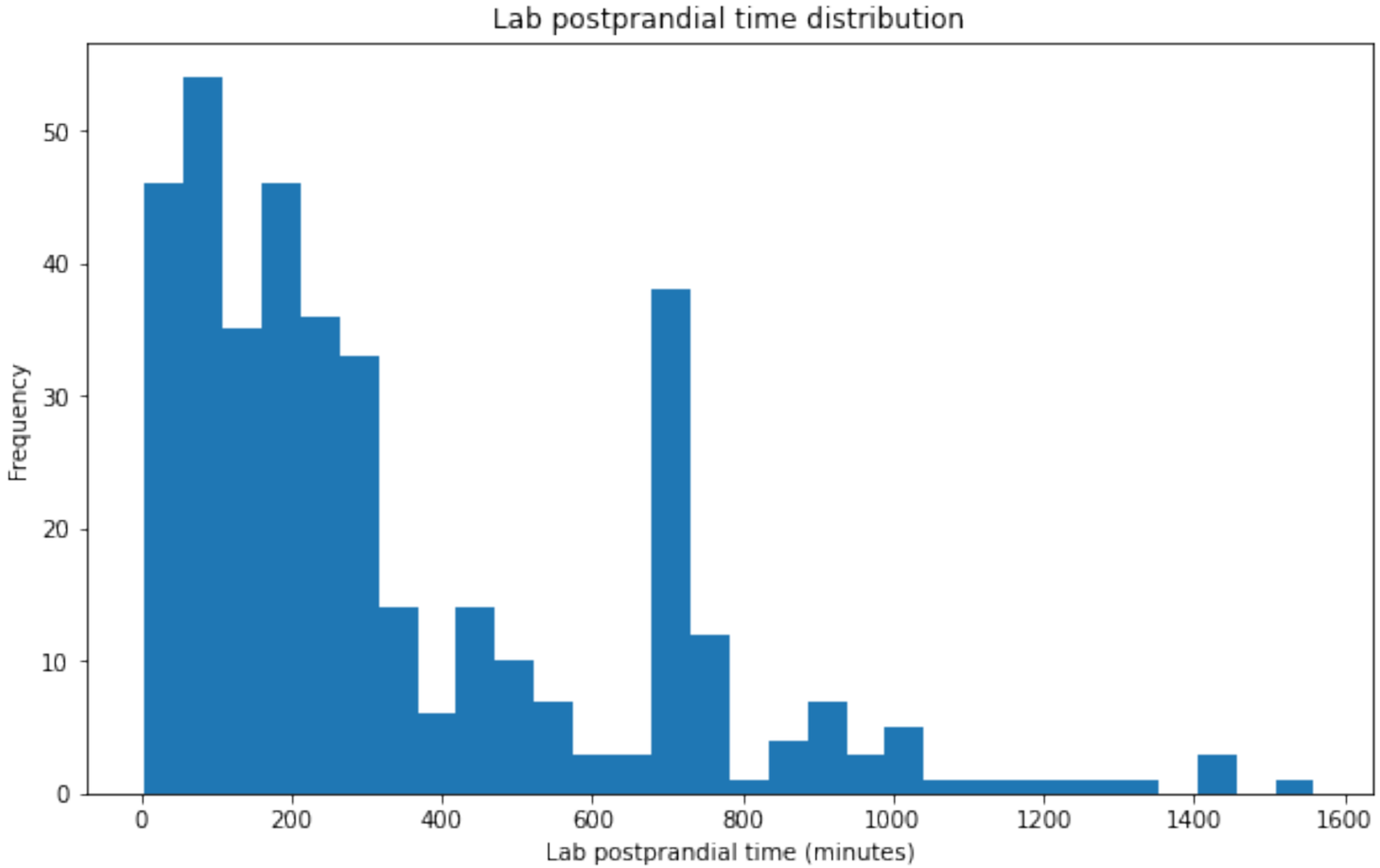




🩸 Glycosylated hemoglobin (by diagnosis)

Highlights

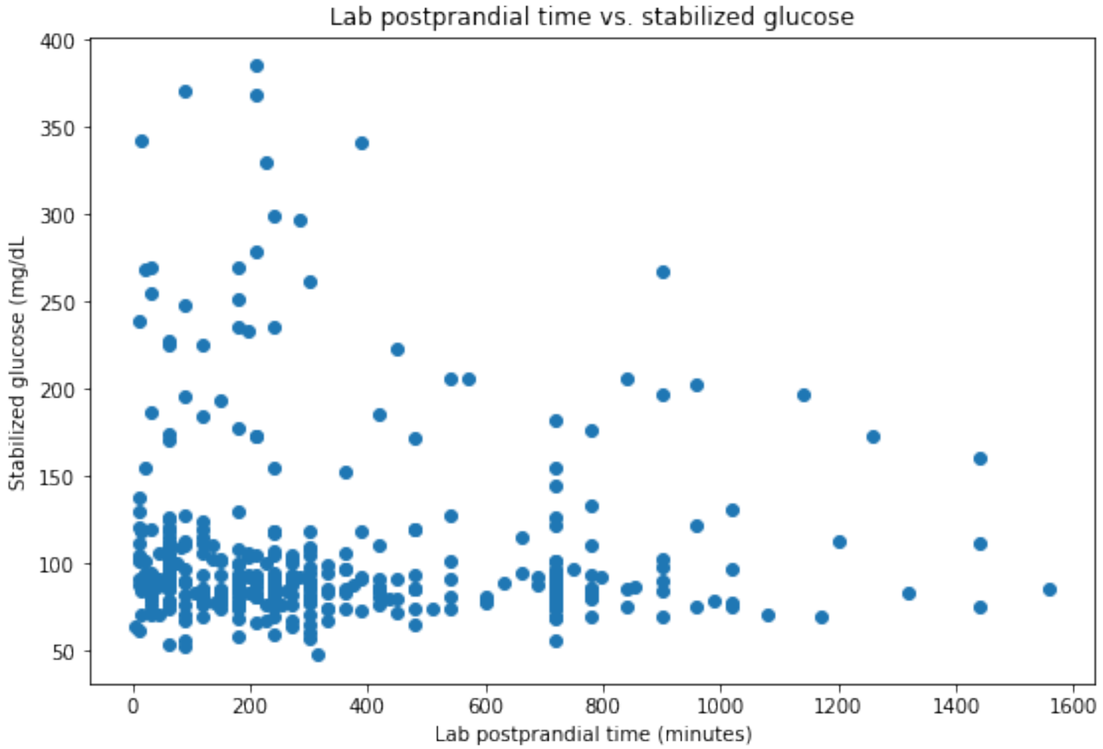
Single variable exploratory data analysis

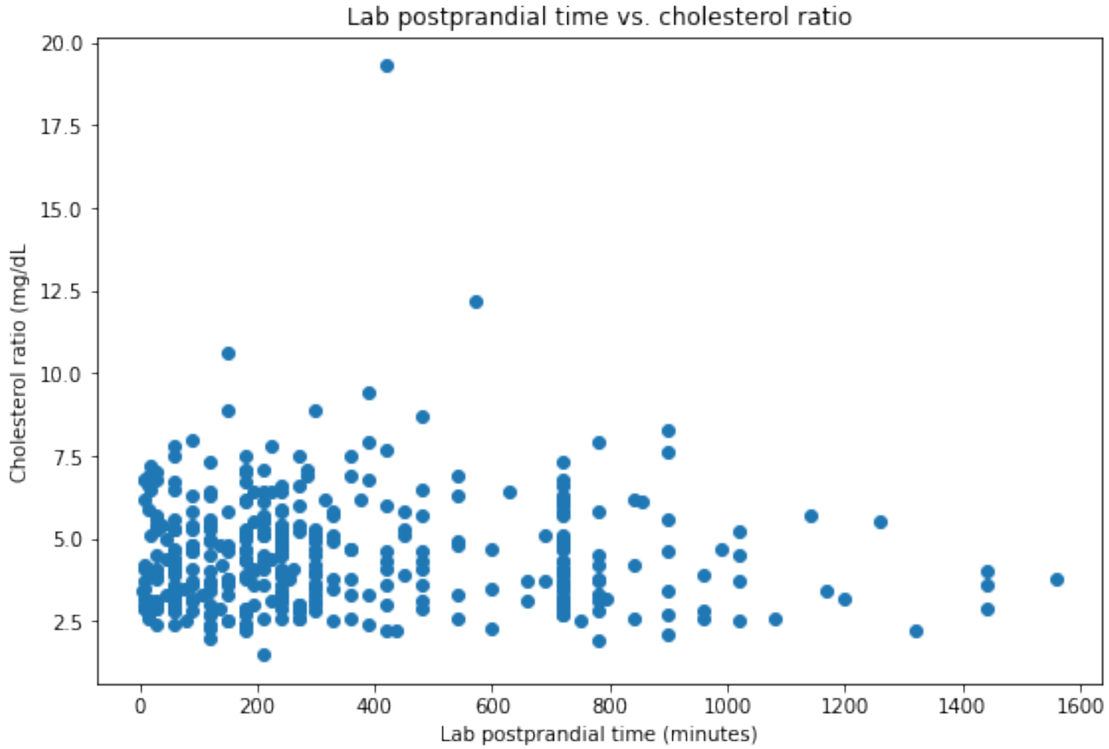


⏰ Lab postprandial time

Highlights

Pairwise variable exploratory data analysis

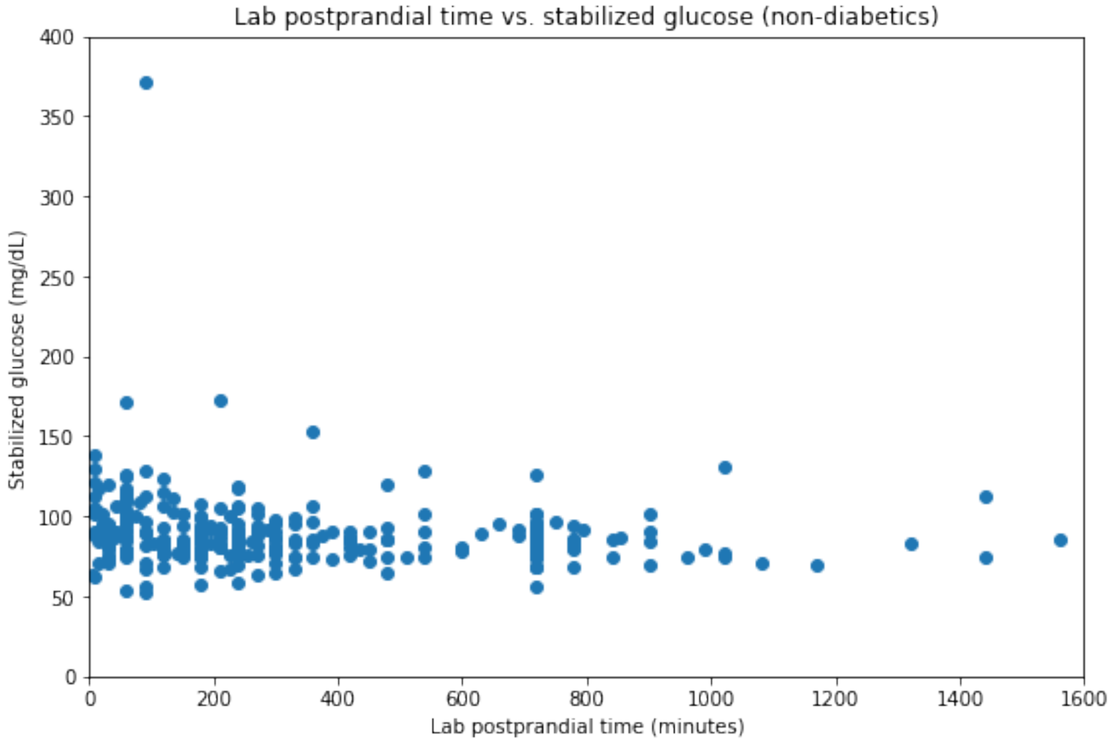


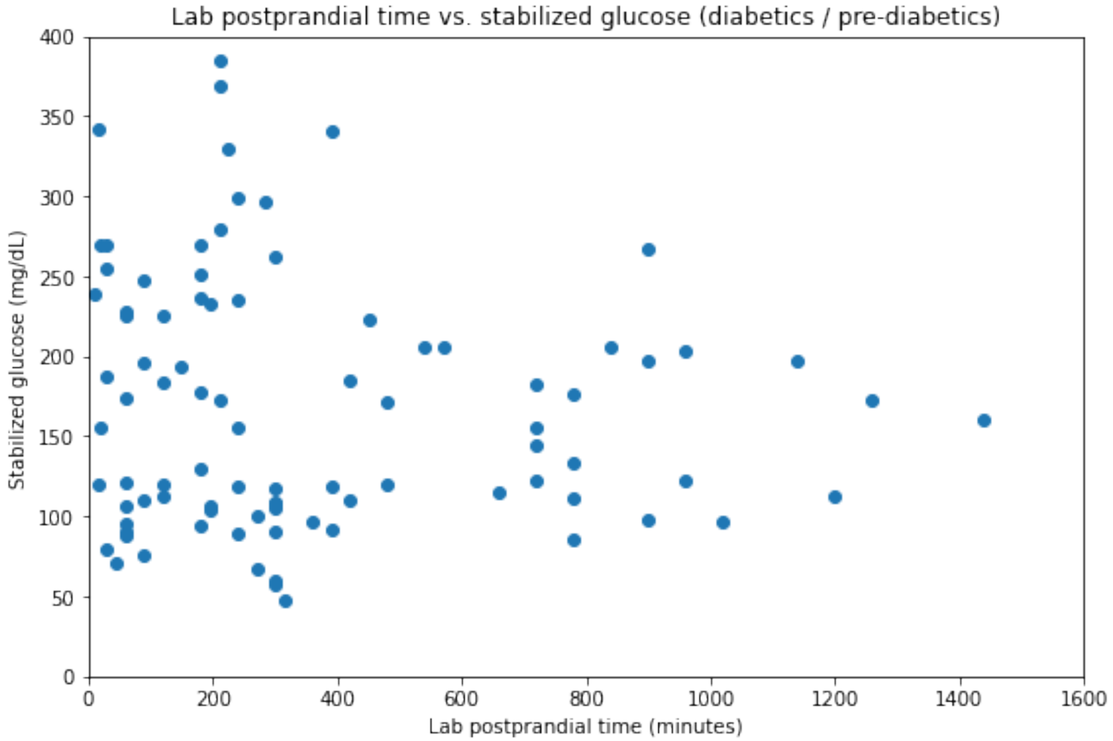


⏰ Lab postprandial time and 🍭 stabilized glucose / 📈 cholesterol ratio

Highlights

Pairwise variable exploratory data analysis

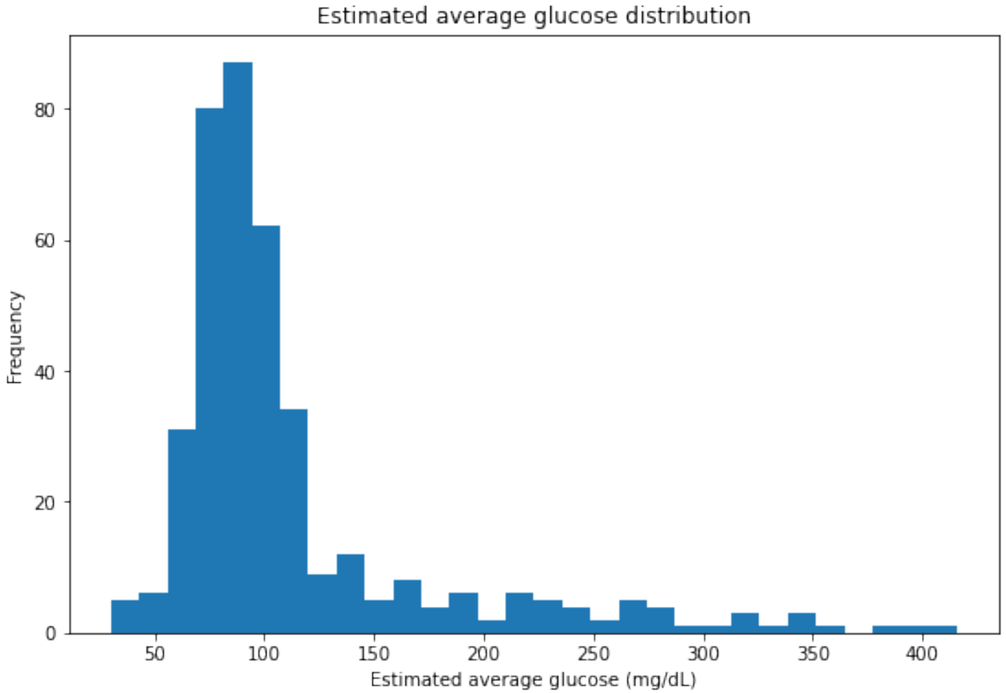


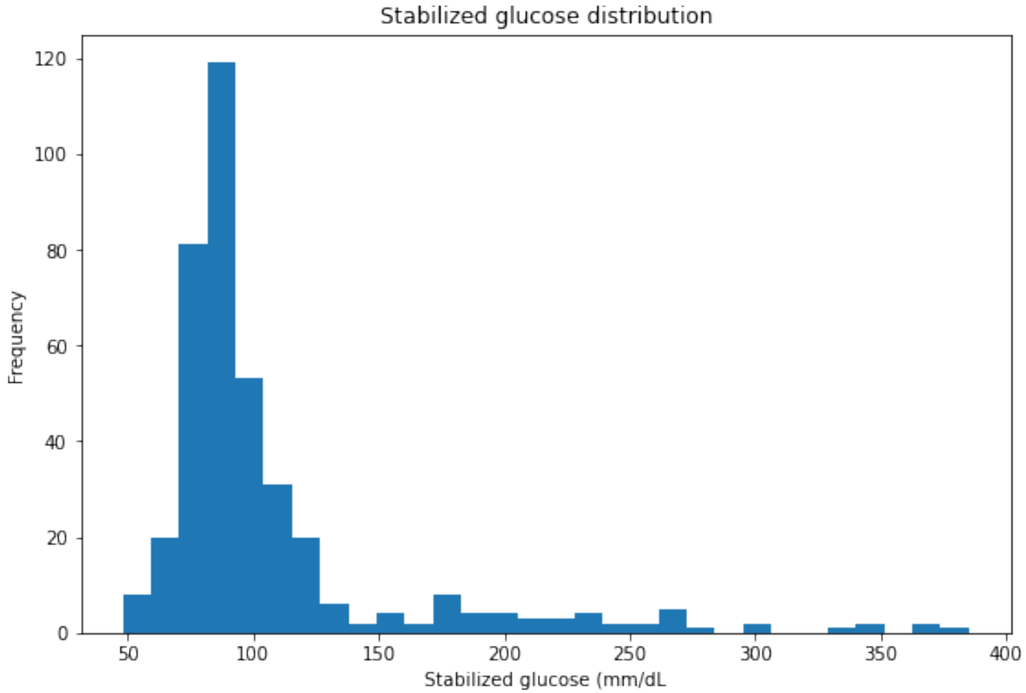


⏰ Lab postprandial time and 🍭 stabilized glucose (by diagnosis)

Highlights

Pairwise variable exploratory data analysis

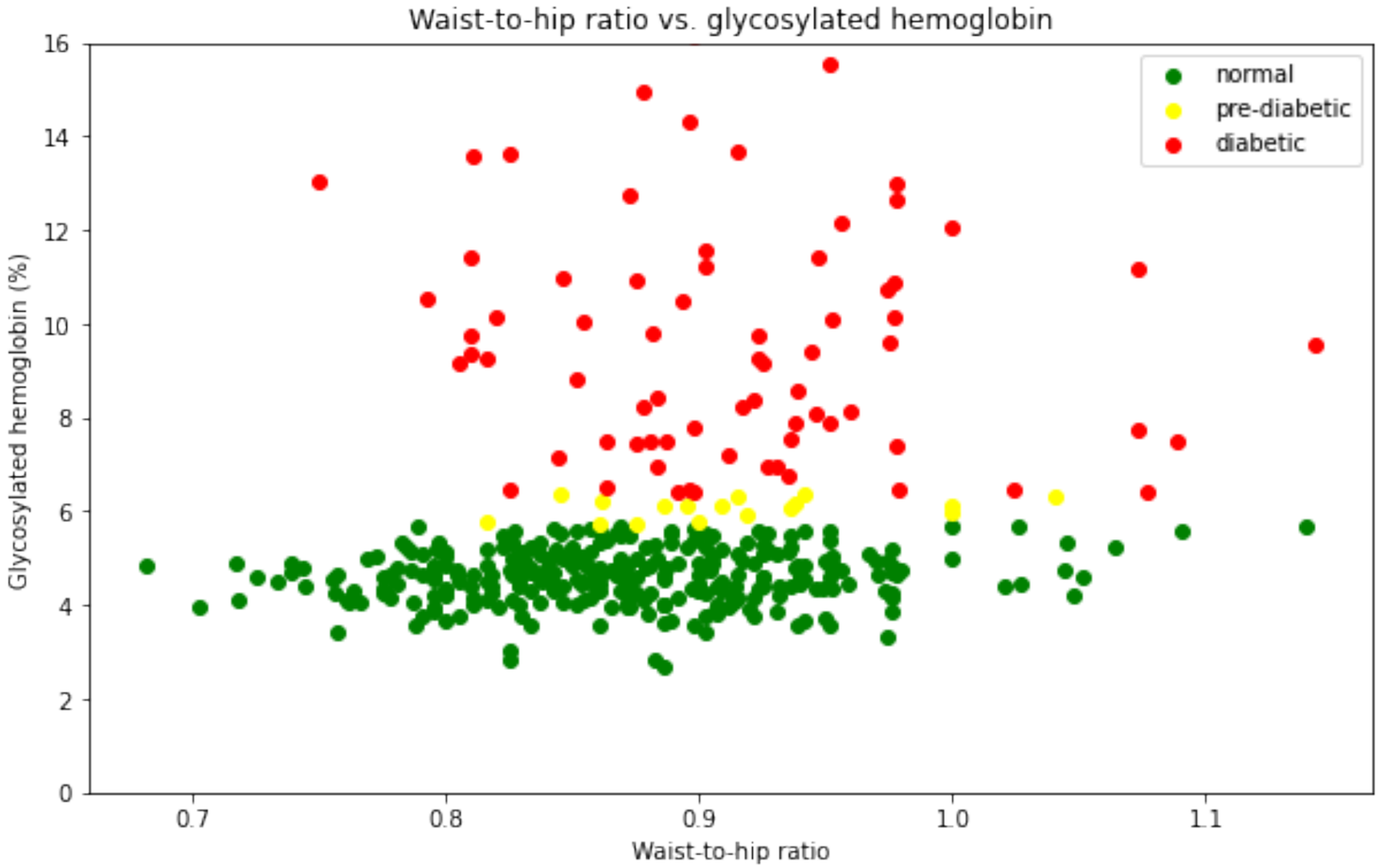




📅 Estimated average glucose and 🍭 stabilized glucose

Highlights

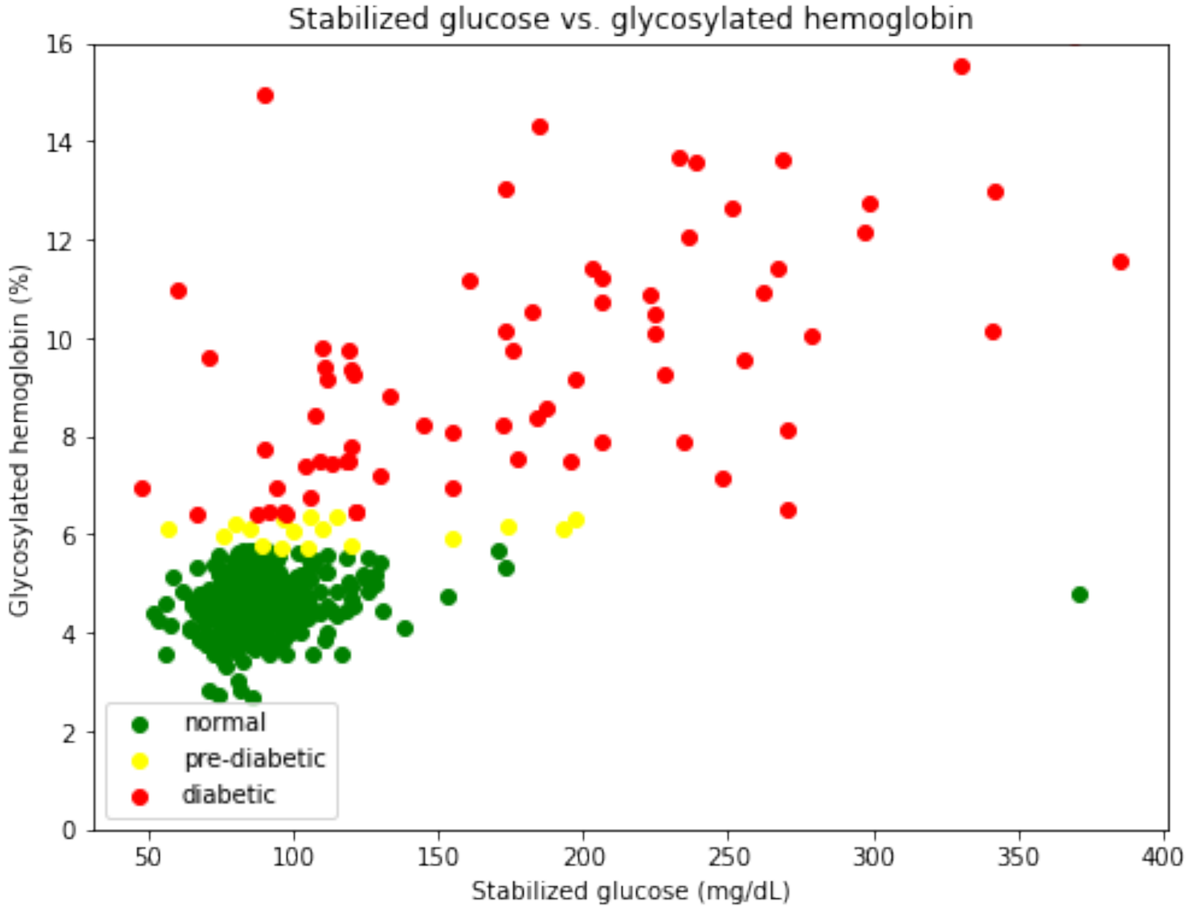
Pairwise variable exploratory data analysis



💃🕺 Waist-to-hip ratio and 🩸 glycosylated hemoglobin

Highlights

Pairwise variable exploratory data analysis



🍭 Stabilized glucose and 🩸 glycosylated hemoglobin

MODEL

Linear Regression Model

Linear Regression Model

The "all in" model

Total cholesterol

Stabilized glucose

High-density lipoprotein

Cholesterol ratio

Glycosylated hemoglobin

Age

Gender

Height

Weight

Frame

Systolic blood pressure

Diastolic blood pressure

Waist

Hips

Lab postprandial time

$R^2 = .60$ $\sigma = 1.43$

Linear Regression Model

Correlation coefficients

.272 .226 Total cholesterol	.741 .520 Stabilized glucose	-.169 -.210 High-density lipoprotein	.355 .310 Cholesterol ratio	 Glycosylated hemoglobin
.332 .427 Age	.048 .069 Gender	.052 .026 Height	.168 .223 Weight	.128 .009 Frame
.164 .250 Systolic blood pressure	.030 .068 Diastolic blood pressure	.248 .226 Waist	.152 .234 Hips	.037 .037 Lab postprandial time

Pearson's **Spearman's**

Linear Regression Model

Reducing number of variables

Total cholesterol

Stabilized glucose

High-density lipoprotein

Cholesterol ratio

Glycosylated hemoglobin

Age

Gender

Height

Weight

Frame

Systolic blood pressure

Diastolic blood pressure

Waist

Hips

Lab postprandial time

Linear Regression Model

Reducing number of variables

Stabilized glucose

Cholesterol ratio

Glycosylated hemoglobin

Age

Height

Weight

Frame

Waist

Hips

Linear Regression Model

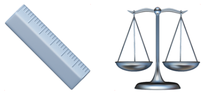
Reducing number of variables

Stabilized glucose

Cholesterol ratio

Glycosylated hemoglobin

Age



BMI

Frame

Waist-to-hip ratio

Linear Regression Model

Improvements

- Reducing number of variables
  glyhb ~ stab_glu + ratio + waist_hip + large + medium + bmi + age

- Domain knowledge, numerical to categorical
  glyhb ~ stab_glu + ratio + waist_hip + large + medium + bmi + age + obese + older + hypertension

- Interaction terms and transformations
  glyhb ~ stab_glu:numeric_diagnosis + ratio + waist_hip + large + medium + numeric_diagnosis + age
  glyhb ~ stab_glu_100 + ratio + waist_hip + large + medium + bmi + age

- Logarithmic transformation
  log_glyhb ~ stab_glu + ratio + waist_hip + large + bmi + age

.59

Mean R^2

.43 - .72

95% credible interval for R^2

glyhb ~ stab_glu + ratio + waist_hip + large + medium + bmi + age

Linear Regression Model

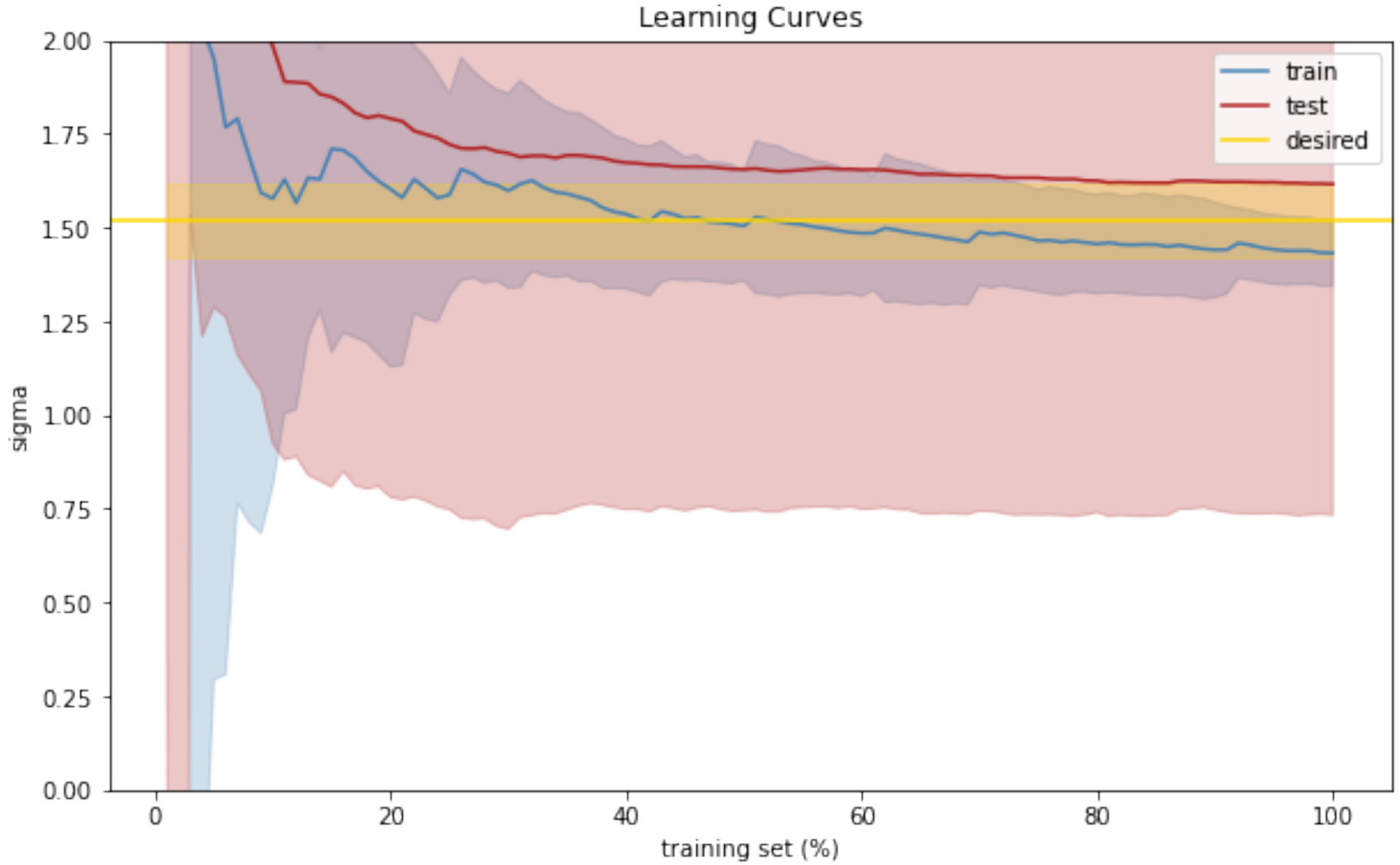
Cross validation

Five rounds of ten-fold cross validation

95% confidence interval for σ	.8936 - 2.4521%
95% confidence interval for R^2	.1338 - .8210
95% confidence interval for *mean* σ	1.3879 - 1.6249%
95% confidence interval for *mean* R^2	.5071 - .6006

Linear Regression Model

Learning curves



REPORT

Predictions

Healthy individual

Stabilized glucose	Waist-to-hip ratio	Cholesterol ratio	Age	Frame	BMI	Glycosylated hemoglobin
75 mg/dL	0.95	3.5	70 years	Small	20	

Predictions

Healthy individual

Stabilized glucose	Waist-to-hip ratio	Cholesterol ratio	Age	Frame	BMI	Glycosylated hemoglobin
75 mm/dL	0.95	3.5	70 years	Small	20	**5.02%**

Predictions

Healthy individual

Stabilized glucose	Waist-to-hip ratio	Cholesterol ratio	Age	Frame	BMI	Glycosylated hemoglobin
200 mm/dL	1.25	4.5	45 years	Large	35	

Predictions

Healthy individual

🍭 Stabilized glucose	💃🕺 Waist-to-hip ratio	📈 Cholesterol ratio	👵🏻 Age	💪 Frame	📏⚖️ BMI	🩸 Glycosylated hemoglobin
200 mm/dL	1.25	4.5	45 years	Large	35	**8.24%**

85.0%

Diagnosis accuracy with model prediction (pre-diabetic / diabetic)

72.7%

Diagnosis accuracy with null model (pre-diabetic / diabetic)

90.3%

Diagnosis accuracy with model prediction (strictly diabetic)

Conclusion

Overall thoughts

- Decent model
- Interesting data exploration and data cleaning
- Underwhelming dataset
- Logistic regression interests