Exploring Predictors Related to Diabetes

Seth Galluzzi, Haley Egan, JD Pinto, and Sydney Masterson

AIM

The aim of this project was to gain an understanding of different variables that relate to diabetes.

RATIONALE

- We used visualizations to explore relationships between variables.
- We analyzed the prevalence of diabetes within African American communities in central Virginia.

THE DATA

- Faraway package in R: *diabetes*.
- 403 objects and 19 variables.
- Numeric continuous variables: cholesterol, age, weight, and high density lipoprotein.
- Categorical variables: gender, location, and frame.

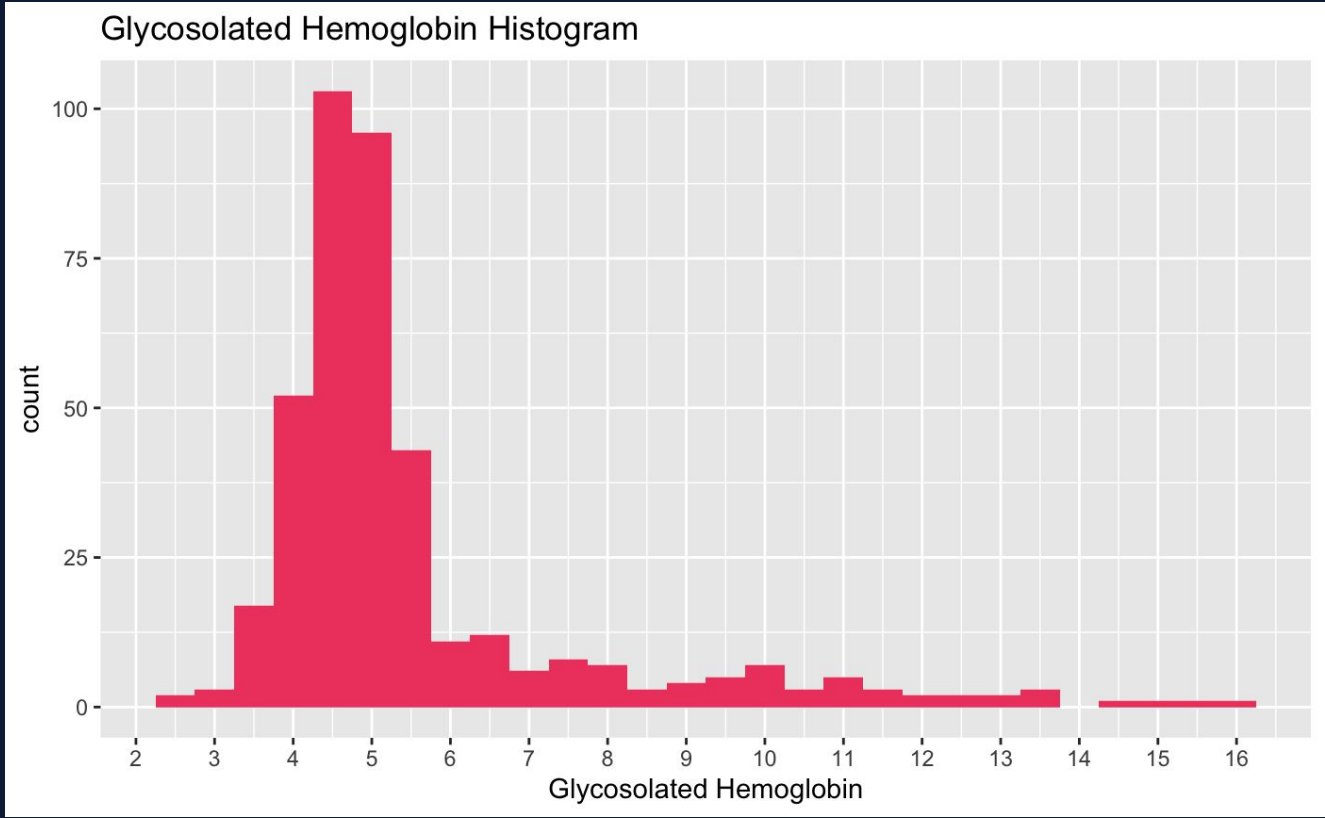
	id	chol	stab.glu	hdl	ratio	glyhb	location	age	gender	height	weight	frame	bp.1s	bp.1d	bp.2s	bp.2d	waist	hip	time.ppn
1	1000	203	82	56	3.6	4.31	Buckingham	46	female	62	121	medium	118	59	NA	NA	29	38	720
2	1001	165	97	24	6.9	4.44	Buckingham	29	female	64	218	large	112	68	NA	NA	46	48	360
3	1002	228	92	37	6.2	4.64	Buckingham	58	female	61	256	large	190	92	185	92	49	57	180
4	1003	78	93	12	6.5	4.63	Buckingham	67	male	67	119	large	110	50	NA	NA	33	38	480
5	1005	249	90	28	8.9	7.72	Buckingham	64	male	68	183	medium	138	80	NA	NA	44	41	300
6	1008	248	94	69	3.6	4.81	Buckingham	34	male	71	190	large	132	86	NA	NA	36	42	195

DATA CLEANING

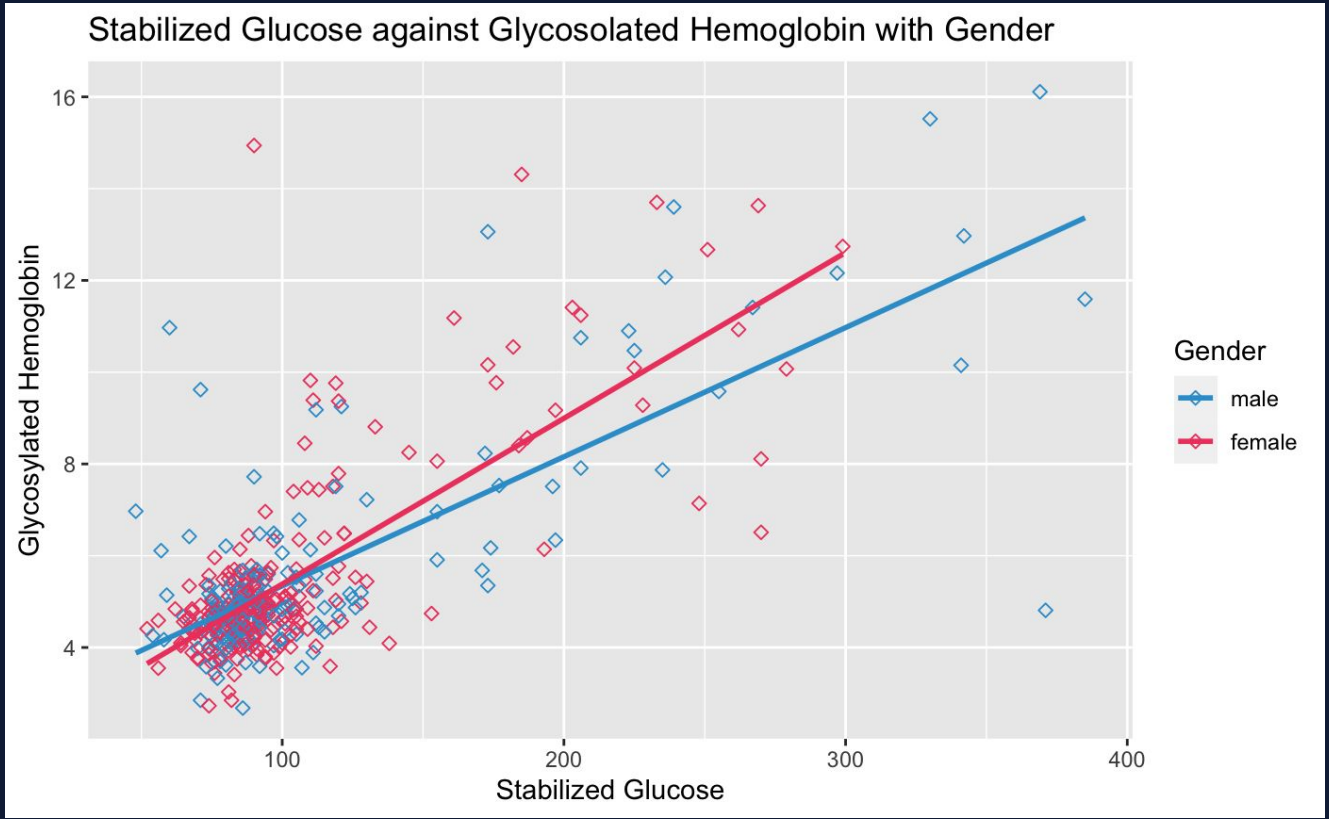
- Dropped: id column, second systolic and second diastolic blood pressure columns.
- Filled the missing values in the remaining columns with the median of each column and added a categorical variable, *diabetes,* for some visualizations.

	chol	stab.glu	hdl	ratio	glyhb	location	age	gender	height	weight	frame	bp.1s	bp.1d	waist	hip	time.ppn
1	203	82	56	3.6	4.31	Buckingham	46	female	62	121	medium	118	59	29	38	720
2	165	97	24	6.9	4.44	Buckingham	29	female	64	218	large	112	68	46	48	360
3	228	92	37	6.2	4.64	Buckingham	58	female	61	256	large	190	92	49	57	180
4	78	93	12	6.5	4.63	Buckingham	67	male	67	119	large	110	50	33	38	480
5	249	90	28	8.9	7.72	Buckingham	64	male	68	183	medium	138	80	44	41	300
6	248	94	69	3.6	4.81	Buckingham	34	male	71	190	large	132	86	36	42	195

EXPLORING VISUALIZATIONS

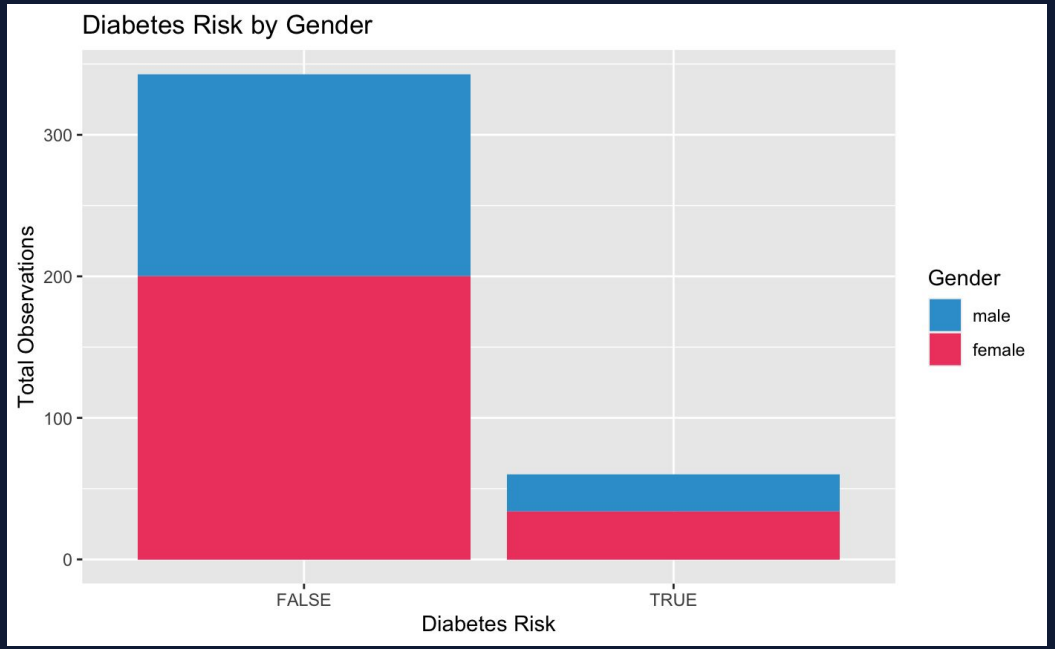


EXPLORING VISUALIZATIONS



EXPLORING VISUALIZATIONS

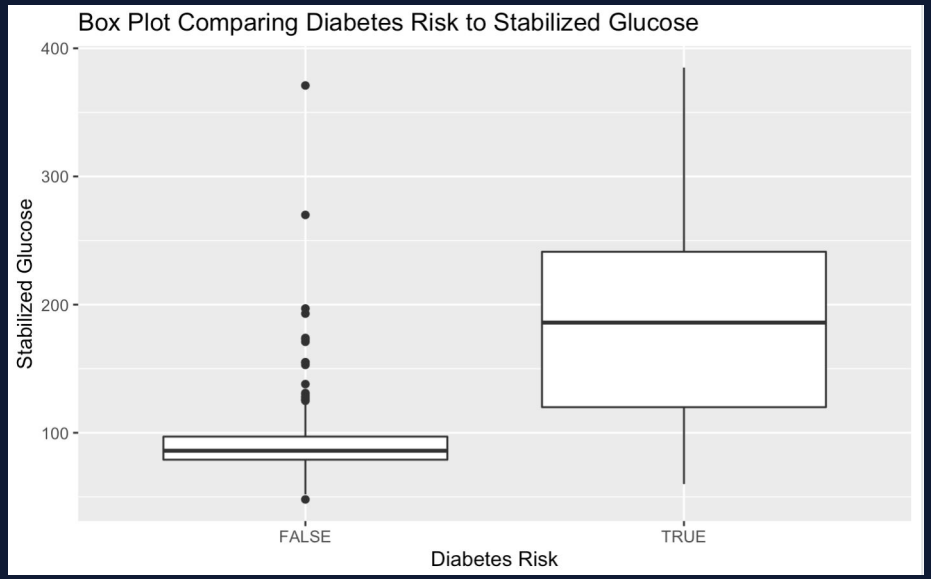
- Response: Glycosylated hemoglobin
- ~14.9% of subjects considered at risk of diabetes

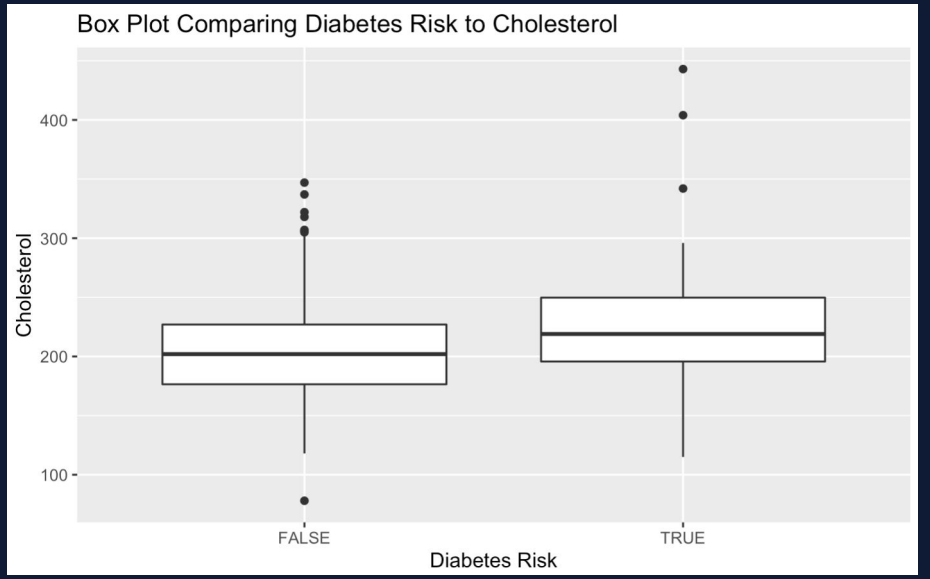


diabetes = TRUE when glycosylated hemoglobin >7

EXPLORING VISUALIZATIONS

The visualizations below illustrate the differences in stabilized glucose and cholesterol of subjects with and without a risk of diabetes.

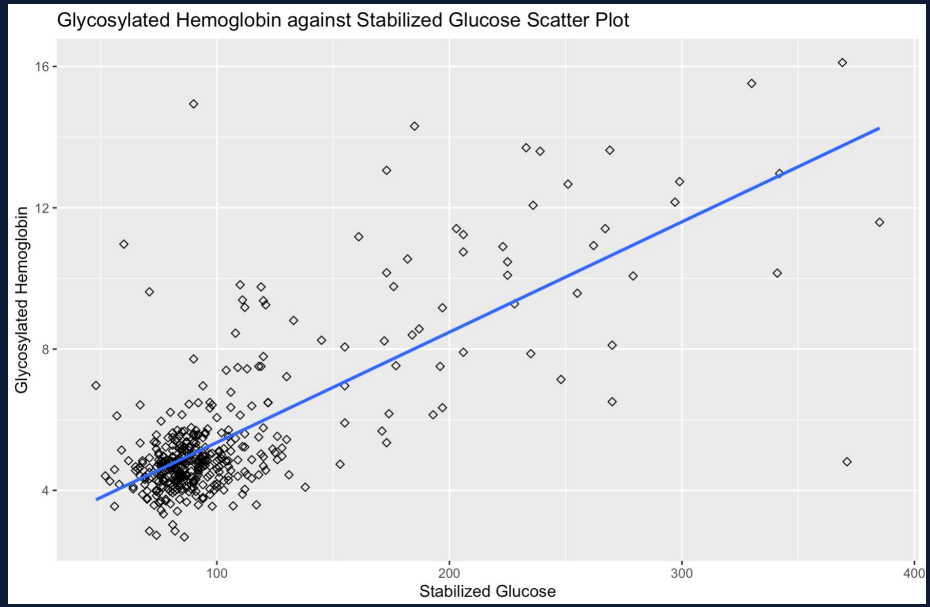




EXPLORING LINEAR REGRESSION

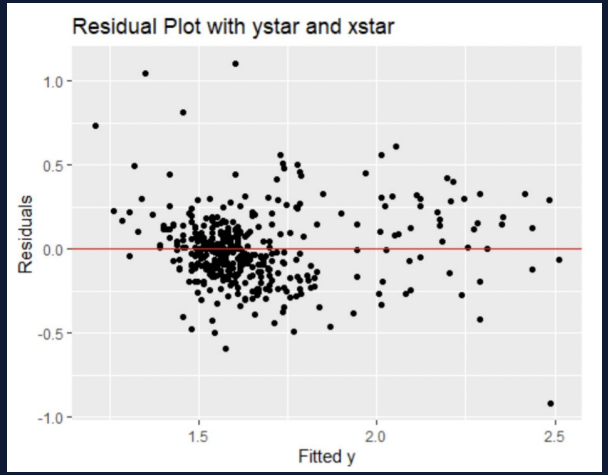
The regression indicates a positive relationship between stabilized glucose and glycosylated hemoglobin.

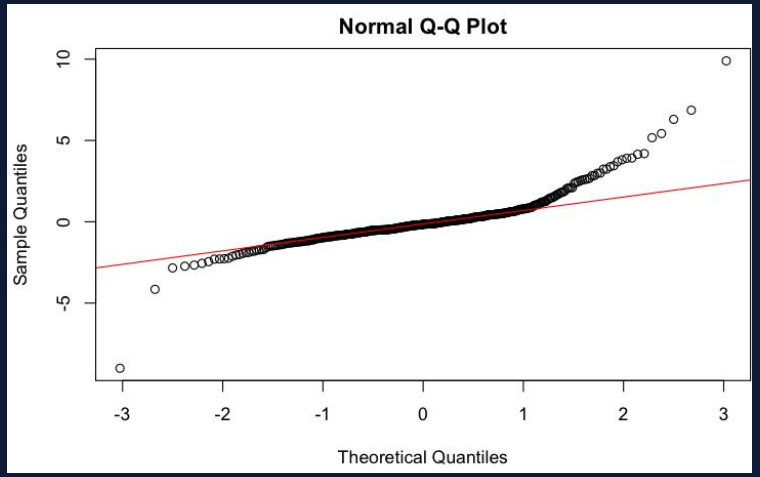
```
Call:
lm(formula = glyhb ~ stab.glu, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0083 -0.6916 -0.1592  0.4255  9.8950

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.235123   0.163877   13.64   <2e-16 ***
stab.glu    0.031221   0.001376   22.69   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.464 on 401 degrees of freedom
Multiple R-squared:  0.5622,    Adjusted R-squared:  0.5611
F-statistic:   515 on 1 and 401 DF,  p-value: < 2.2e-16
```

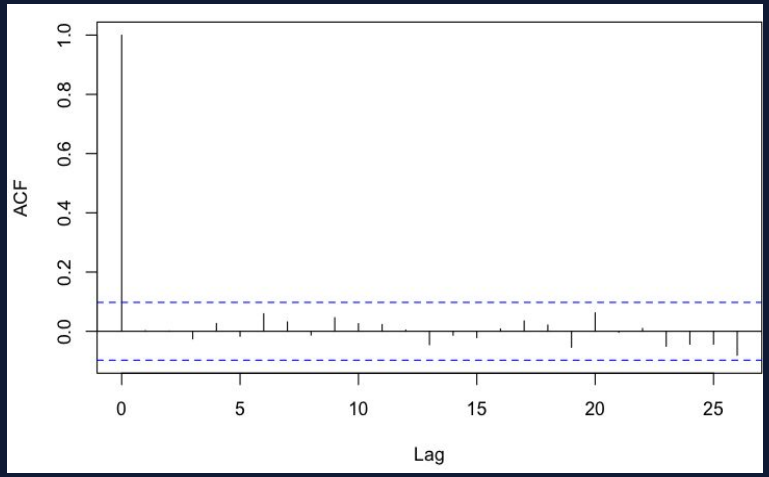


SATISFYING REGRESSION ASSUMPTIONS:

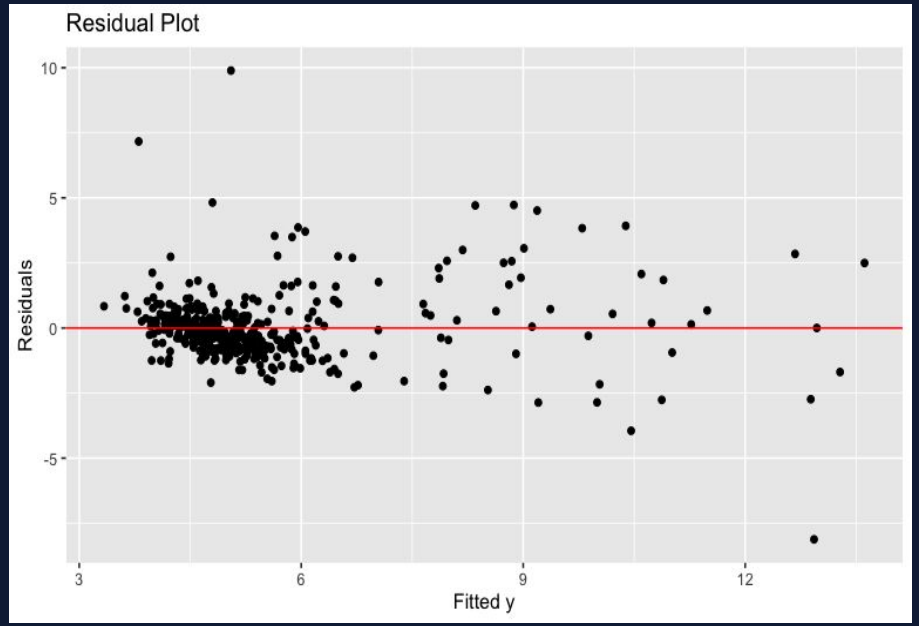
- Linear relationship
- Error terms normally distributed
- Constant variance
- Independent observations

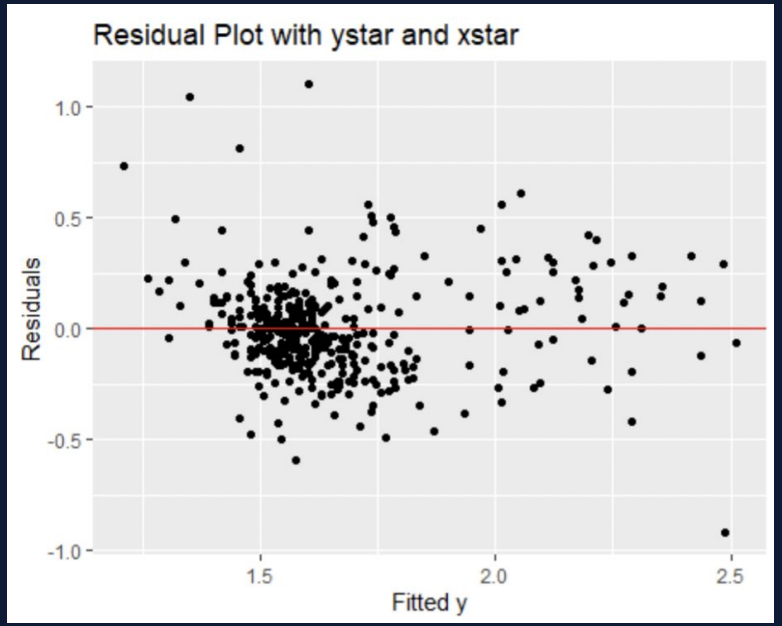






RESIDUAL COMPARISON BEFORE AND AFTER TRANSFORMATIONS





MODEL SELECTION

- After model diagnostics, we chose cholesterol, stabilized glucose, age, postprandial time, and ratio as predictors.

R^2	y = -1.6516 + 0.0034chol + 0.0287stab.glu + 0.1169ratio + 0.016age + 0.3016female + 0.0277height + 0.0006time
Mallow	y = 0.4162 + 0.0034chol + 0.0287stab.glu + 0.1131ratio + 0.0147age + 0.0006time
BIC	y = 0.8668 + 0.0286stab.glu + 0.1533ratio + 0.0163age + 0.0006time
Forward	y = 0.4162 + 0.0287stab.glu + 0.0034chol + 0.0147age + 0.0006time + 0.1131ratio
Backward	y = 0.4162 + 0.0034chol + 0.0287stab.glu + 0.1132ratio + 0.0147age + 0.0006time
Stepwise	y = 0.4162 + 0.0287stab.glu + 0.0034chol + 0.0147age + 0.0006time + 0.1132ratio

Multiple Linear Regression

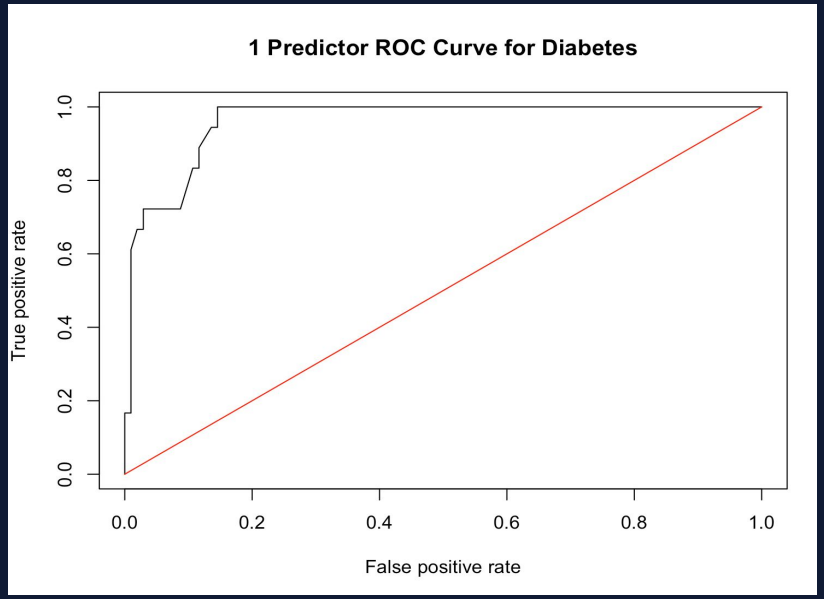
- After comparing two models, it was determined that model 1 would be best suited for our data.

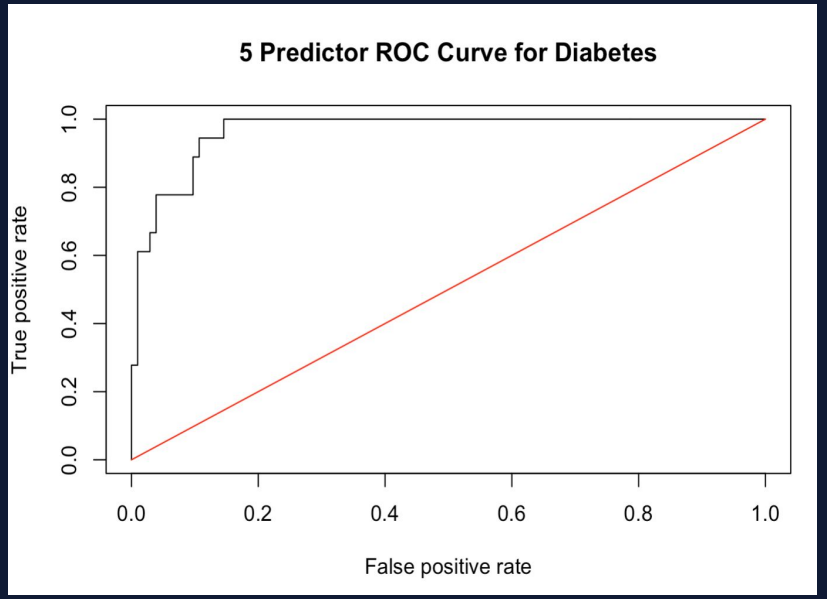
```
Analysis of Variance Table

Model 1: glyhb ~ stab.glu + age + time.ppn + chol + ratio
Model 2: glyhb ~ chol + stab.glu + hdl + ratio + age + height + weight +
    frame + bp.1s + bp.1d + waist + hip + time.ppn
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    397 786.52
2    389 781.70  8    4.8204 0.2998 0.9658
```

```
Call:
lm(formula = glyhb ~ ., data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-8.1591 -0.6610 -0.1499  0.4264  9.9282

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.9752899  2.0101961  -0.485   0.6278
chol         0.0034207  0.0030970   1.105   0.2700
stab.glu     0.0284783  0.0014618  19.481   <2e-16 ***
hdl         -0.0001114  0.0096717  -0.012   0.9908
ratio        0.1071249  0.1083805   0.988   0.3236
age          0.0130169  0.0056457   2.306   0.0217 *
height       0.0096589  0.0232156   0.416   0.6776
weight      -0.0017231  0.0047486  -0.363   0.7169
frame       -0.0816749  0.1206019  -0.677   0.4987
bp.1s        0.0036802  0.0045722   0.805   0.4214
bp.1d       -0.0038652  0.0069898  -0.553   0.5806
waist        0.0116304  0.0286548   0.406   0.6851
hip          0.0165060  0.0299911   0.550   0.5824
time.ppn     0.0005936  0.0002344   2.532   0.0117 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.418 on 389 degrees of freedom
Multiple R-squared:  0.6019,    Adjusted R-squared:  0.5886
F-statistic: 45.24 on 13 and 389 DF,  p-value: < 2.2e-16
```

Comparing ROCs and AUCs

Data indicated a simple linear regression model was nearly as effective in predicting diabetes risk as our 5 predictor model.





ISSUES, NOTES, AND INSIGHTS

- Size of the data set
- Missing values and data not collected
- SLR vs MLR
- Time to explore other variables
- 6.5 vs 7

I can do more, to lower my A1C.

SUMMARY

- We compared visualizations between variables such as glycosylated hemoglobin, stabilized glucose, and diabetes risk.
- Predictors such as age, total cholesterol, and stabilized glucose can impact diabetes risk.
- A linear regression relating stabilized glucose to glycosylated hemoglobin was most appropriate.
- Increase awareness of different predictors related to diabetes.