markdown-version

Group 1

2024-04-28

Pima Indians Diabetes Dataset

Dataset: Pima Indians Diabetes Dataset

Variables

- pregnancies (int)

- plasma glucose (int)

- blood pressure (int)

- skin thickness (int)

- BMI (dec)

- diabetes pedigree function (dec)

- age (int)

Outcome

- binary value indicating diabetes diagnosis within a 5-year period.

The diagnostic criteria defined as 2 hour post-load plasma glucose at 200 (mg/dl)

Data Cleaning: Eliminating Null/NA records

```r
knitr::kable(head(pima))
```

pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age	diabetes
6	148	72	35	NA	33.6	0.627	50	pos
1	85	66	29	NA	26.6	0.351	31	neg
8	183	64	NA	NA	23.3	0.672	32	pos
1	89	66	23	94	28.1	0.167	21	neg
0	137	40	35	168	43.1	2.288	33	pos
5	116	74	NA	NA	25.6	0.201	30	neg

Pima Dataset

```r
knitr::kable(head(pima_clean))
```

	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age	outcome
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
7	3	78	50	32	88	31.0	0.248	26	1
9	2	197	70	45	543	30.5	0.158	53	1
14	1	189	60	23	846	30.1	0.398	59	1
15	5	166	72	19	175	25.8	0.587	51	1

Cleaned Pima Dataset: Summary

```r
knitr::kable(summary(pima_clean))
```

	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age	outcome
	Min. : 0.000	Min. : 56.0	Min. : 24.00	Min. : 7.00	Min. : 14.00	Min. :18.20	Min. :0.0850	Min. :21.00	Min. :0.0000
	1st Qu.: 1.000	1st Qu.: 99.0	1st Qu.: 62.00	1st Qu.:21.00	1st Qu.: 76.75	1st Qu.:28.40	1st Qu.:0.2697	1st Qu.:23.00	1st Qu.:0.0000
	Median : 2.000	Median :119.0	Median : 70.00	Median :29.00	Median :125.50	Median :33.20	Median :0.4495	Median :27.00	Median :0.0000
	Mean : 3.301	Mean :122.6	Mean : 70.66	Mean :29.15	Mean :156.06	Mean :33.09	Mean :0.5230	Mean :30.86	Mean :0.3316
	3rd Qu.: 5.000	3rd Qu.:143.0	3rd Qu.: 78.00	3rd Qu.:37.00	3rd Qu.:190.00	3rd Qu.:37.10	3rd Qu.:0.6870	3rd Qu.:36.00	3rd Qu.:1.0000
	Max. :17.000	Max. :198.0	Max. :110.00	Max. :63.00	Max. :846.00	Max. :67.10	Max. :2.4200	Max. :81.00	Max. :1.0000

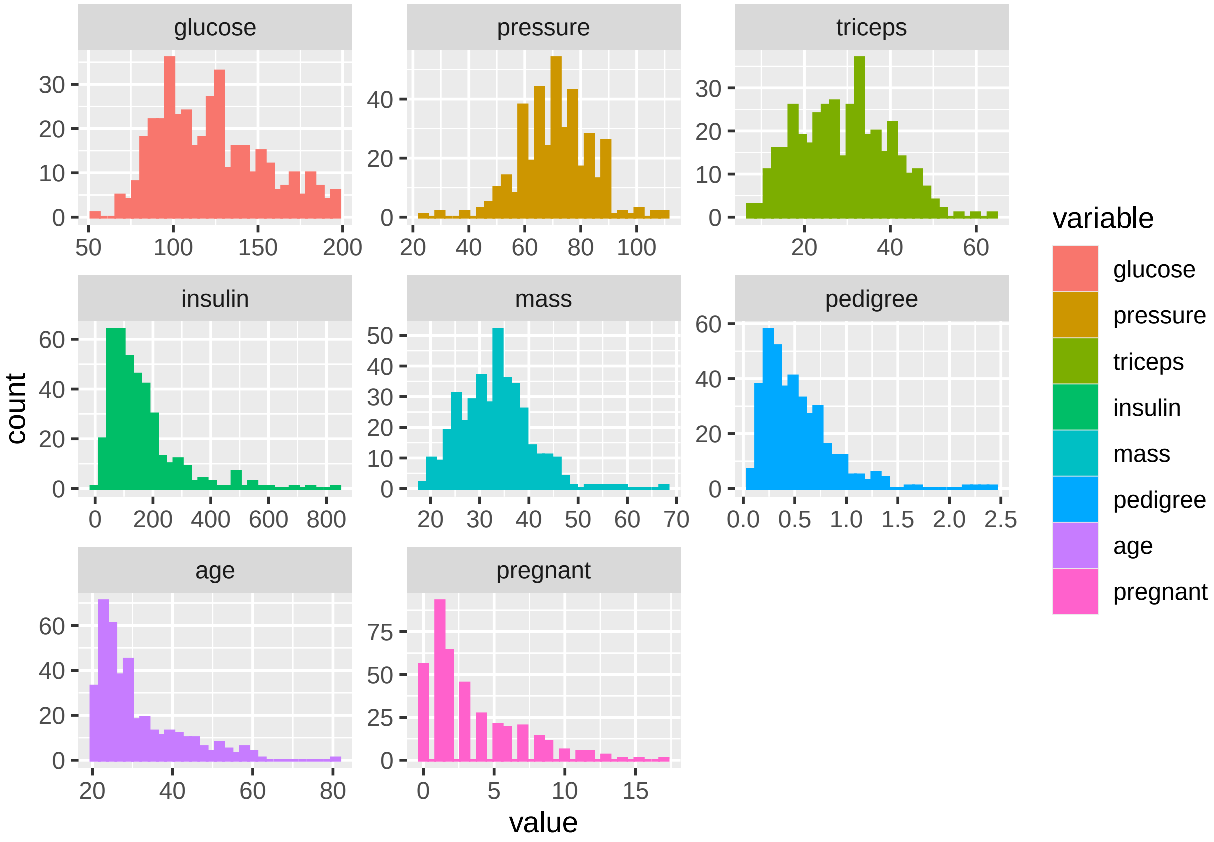
Explantory Variables

```r
bycols <- colnames(pima_clean)

melted_continuous <- reshape2::melt(data=pima_clean[c("glucose", "pressure", "triceps", "insulin", "mas


# Plot histograms for continuous variables
ggplot(melted_continuous, aes(x = value, fill = variable, color = variable)) +
  geom_histogram() +
  facet_wrap(~variable, scales = "free")
```

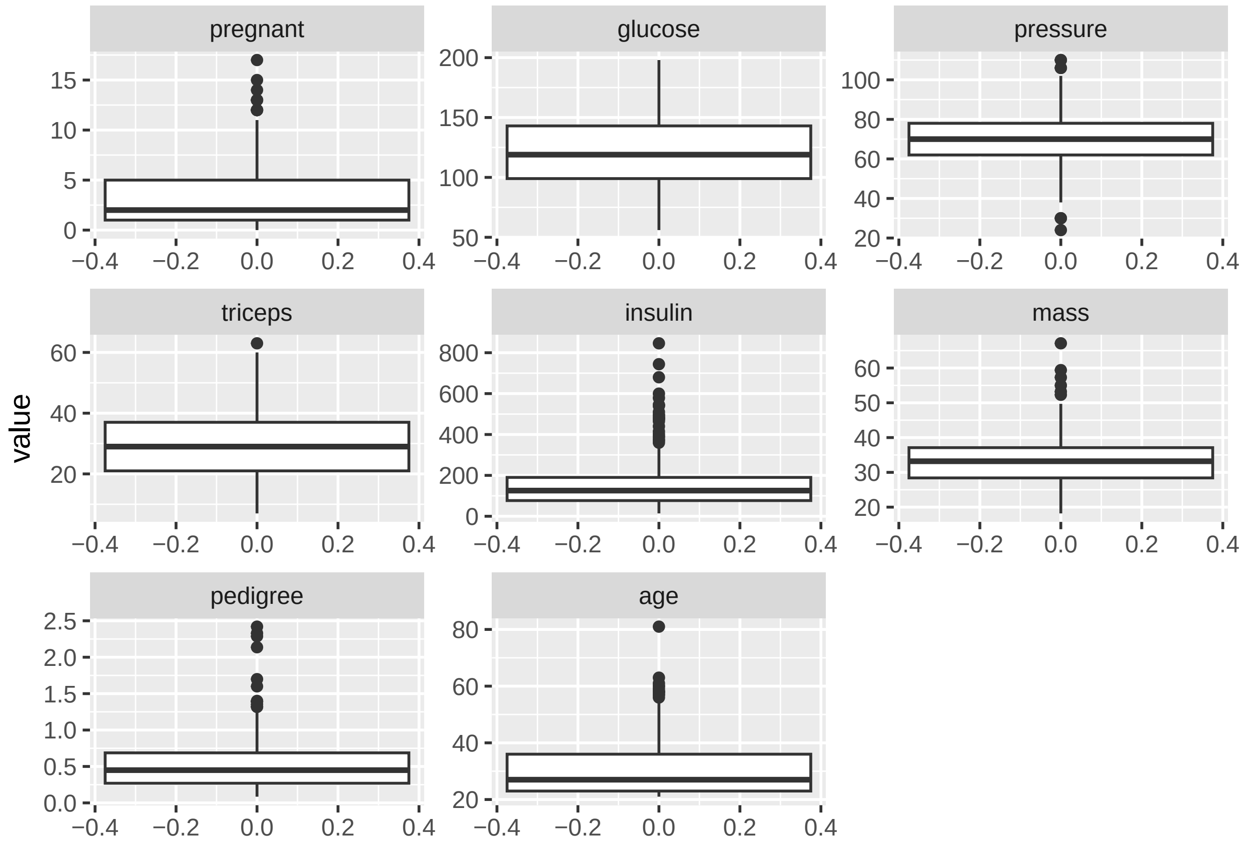
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



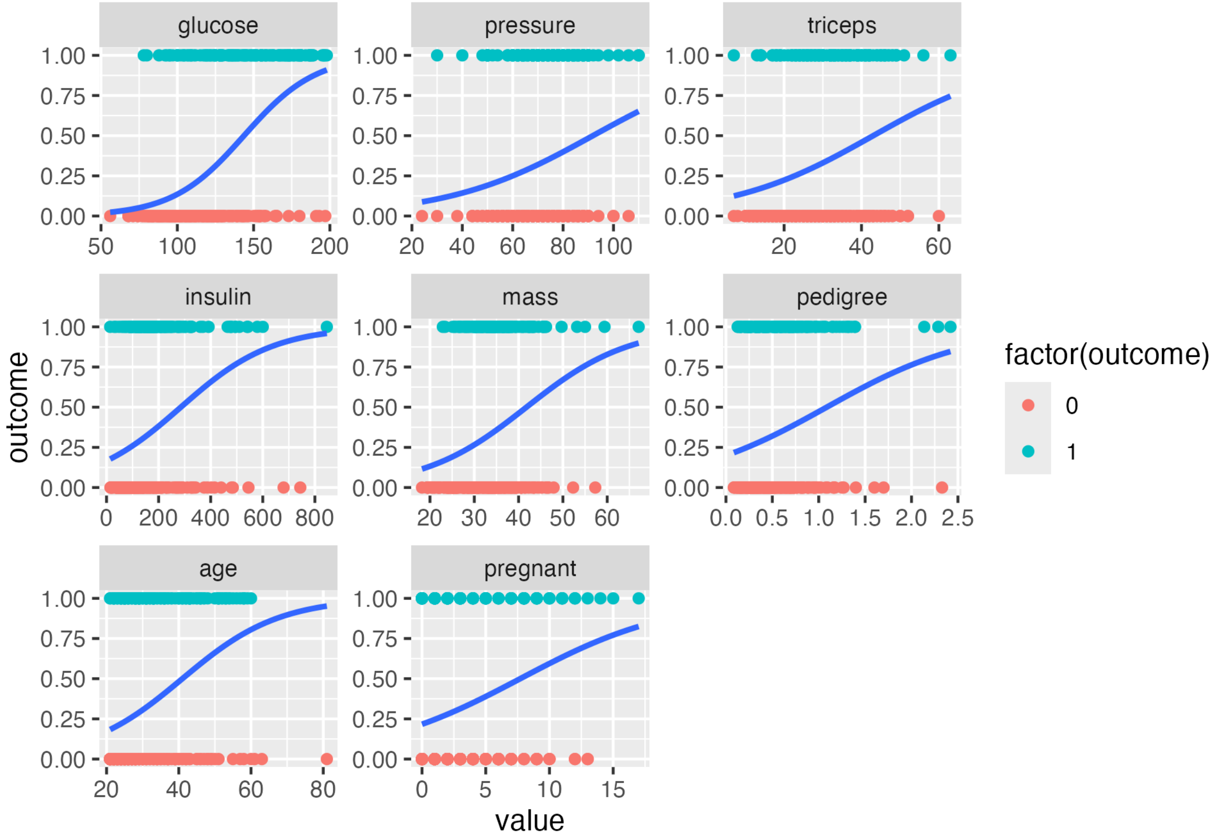
Identifying Outliers

```
pima_clean %>%
  select(-c(outcome)) %>%
  reshape2::melt() %>%
  ggplot(aes(y=value)) +
  geom_boxplot() +
  # geom_histogram() +
  facet_wrap(~variable, scales = "free")
```

```
## No id variables; using all as measure variables
```

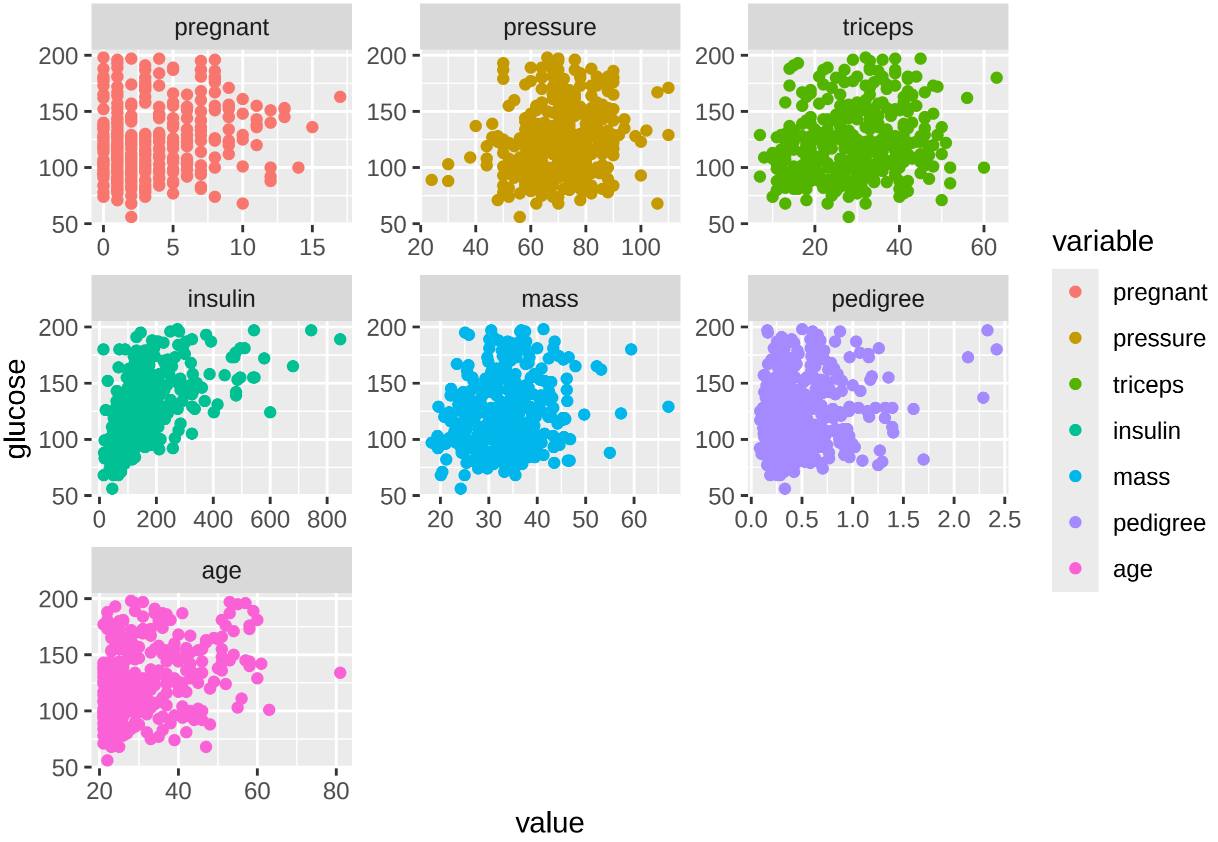

pregnant
glucose
pressure
triceps
insulin
mass
pedigree
age
value
0
5
10
15
50
100
150
200
20
40
60
80
100
20
40
60
0
200
400
600
800
20
30
40
50
60
0.0
0.5
1.0
1.5
2.0
2.5
20
40
60
80
−0.4
−0.2
0.0
0.2
0.4

Plotted against outcome



Plotted against Blood Glucose

```
pima_clean %>%
  dplyr::select(-c(outcome)) %>%
  reshape2::melt(id.vars = "glucose") %>%
  ggplot(aes(x = value, y = glucose, color = variable)) +
  geom_point() +
  facet_wrap(~variable, scales = "free")
```



Plotting Correlations within the Dataset: Pearson's correlation

```r
cormat <- round(cor(pima_clean[,-9]), 2)
knitr::kable(head(cormat))
```

	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age
pregnant	1.00	0.20	0.21	0.09	0.08	-0.03	0.01	0.68
glucose	0.20	1.00	0.21	0.20	0.58	0.21	0.14	0.34
pressure	0.21	0.21	1.00	0.23	0.10	0.30	-0.02	0.30
triceps	0.09	0.20	0.23	1.00	0.18	0.66	0.16	0.17
insulin	0.08	0.58	0.10	0.18	1.00	0.23	0.14	0.22
mass	-0.03	0.21	0.30	0.66	0.23	1.00	0.16	0.07

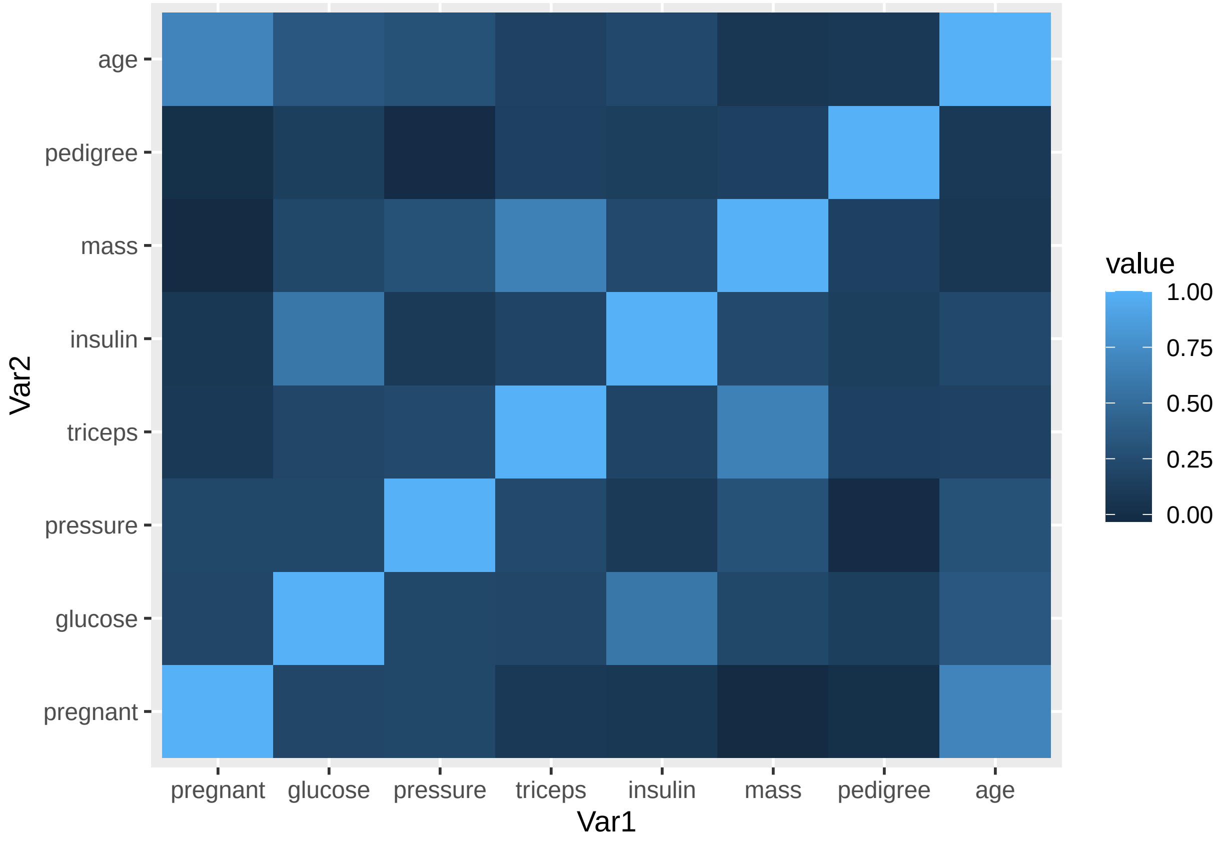
Plotting Correlations within the Dataset: Spearman correlation

```r
cormat_sp <- round(cor(pima_clean[,-9], method='spearman'), 2)
knitr::kable(head(cormat_sp))
```

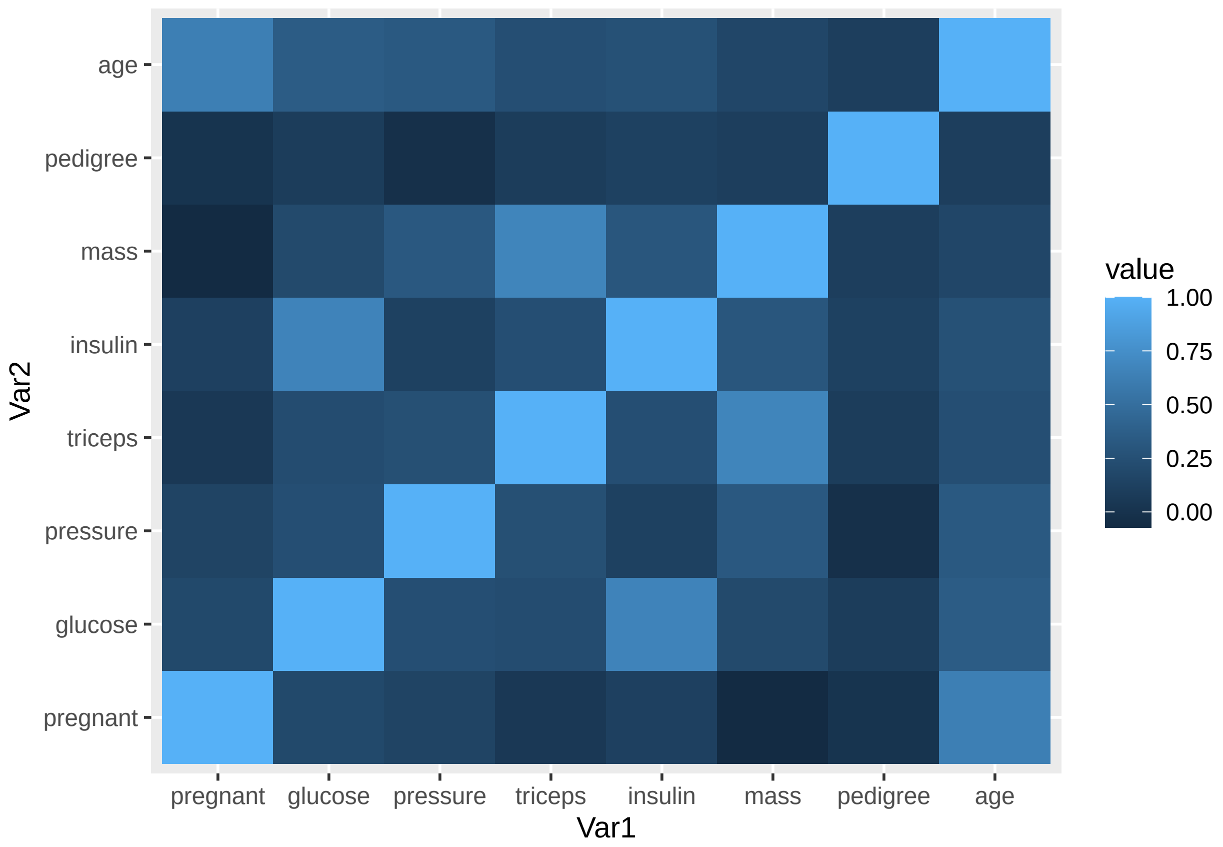
	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age
pregnant	1.00	0.19	0.15	0.05	0.12	-0.07	0.01	0.63
glucose	0.19	1.00	0.24	0.22	0.66	0.20	0.09	0.35
pressure	0.15	0.24	1.00	0.25	0.13	0.32	-0.02	0.33
triceps	0.05	0.22	0.25	1.00	0.24	0.67	0.09	0.24
insulin	0.12	0.66	0.13	0.24	1.00	0.30	0.13	0.26
mass	-0.07	0.20	0.32	0.67	0.30	1.00	0.10	0.17

Correlation Heatmaps: Pearsons vs Spearman

```r
melted_cormat <- reshape2::melt(cormat)
ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile()
```



```r
melted_cormat_sp <- reshape2::melt(cormat_sp)
ggplot(data = melted_cormat_sp, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile()
```



Regression Analysis: Univariate Logistic Regression

```
summary(pima_clean)
```

```
##     pregnant         glucose         pressure         triceps     
##  Min.   : 0.000   Min.   : 56.0   Min.   : 24.00   Min.   : 7.00  
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.:21.00  
##  Median : 2.000   Median :119.0   Median : 70.00   Median :29.00  
##  Mean   : 3.301   Mean   :122.6   Mean   : 70.66   Mean   :29.15  
##  3rd Qu.: 5.000   3rd Qu.:143.0   3rd Qu.: 78.00   3rd Qu.:37.00  
##  Max.   :17.000   Max.   :198.0   Max.   :110.00   Max.   :63.00  
##     insulin            mass          pedigree           age       
##  Min.   : 14.00   Min.   :18.20   Min.   :0.0850   Min.   :21.00  
##  1st Qu.: 76.75   1st Qu.:28.40   1st Qu.:0.2697   1st Qu.:23.00  
##  Median :125.50   Median :33.20   Median :0.4495   Median :27.00  
##  Mean   :156.06   Mean   :33.09   Mean   :0.5230   Mean   :30.86  
##  3rd Qu.:190.00   3rd Qu.:37.10   3rd Qu.:0.6870   3rd Qu.:36.00  
##  Max.   :846.00   Max.   :67.10   Max.   :2.4200   Max.   :81.00  
##     outcome      
##  Min.   :0.0000  
##  1st Qu.:0.0000  
##  Median :0.0000  
##  Mean   :0.3316  
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

```
Indicators <- pima_clean[, c("pregnant", "glucose", "pressure", "triceps", "insulin", "mass", "pedigree

models <- list()
model_summaries <- list()
# Iterate over the columns of 'Indicators' dataframe
for (col in colnames(Indicators)) {
  # Fit a logistic regression model for each predictor variable
  form = as.formula(paste("outcome ~", col))
  models[[col]] <- glm(formula=form, family = binomial(link = "logit"), data=pima_clean)
  # Storing summary of each model in the list
  model_summaries[[col]] <- summary(models[[col]])
}

for (col in names(model_summaries)) {
  print(paste("Summary for", col, "predictor:"))
  print(model_summaries[[col]])
}
```

```
## [1] "Summary for pregnant predictor:"
##
## Call:
## glm(formula = form, family = binomial(link = "logit"), data = pima_clean)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.28480    0.16671  -7.707 1.29e-14 ***
## pregnant     0.16674    0.03443   4.843 1.28e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 473.03  on 390  degrees of freedom
## AIC: 477.03
##
## Number of Fisher Scoring iterations: 4
##
## [1] "Summary for glucose predictor:"
##
## Call:
## glm(formula = form, family = binomial(link = "logit"), data = pima_clean)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.095521   0.629787  -9.679   <2e-16 ***
## glucose      0.042421   0.004761   8.911   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 386.67  on 390  degrees of freedom
## AIC: 390.67
##
## Number of Fisher Scoring iterations: 4
##
## [1] "Summary for pressure predictor:"
##
## Call:
## glm(formula = form, family = binomial(link = "logit"), data = pima_clean)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.168012   0.676646  -4.682 2.84e-06 ***
## pressure     0.034492   0.009233   3.736 0.000187 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 483.16  on 390  degrees of freedom
## AIC: 487.16
##
## Number of Fisher Scoring iterations: 4
##
## [1] "Summary for triceps predictor:"
##
## Call:
## glm(formula = form, family = binomial(link = "logit"), data = pima_clean)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.32588    0.35765  -6.503 7.86e-11 ***
## triceps      0.05404    0.01101   4.910 9.09e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 471.96  on 390  degrees of freedom
## AIC: 475.96
##
## Number of Fisher Scoring iterations: 4
##
## [1] "Summary for insulin predictor:"
##
## Call:
## glm(formula = form, family = binomial(link = "logit"), data = pima_clean)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.612947   0.203687  -7.919 2.40e-15 ***
## insulin      0.005653   0.001058   5.345 9.04e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 462.92  on 390  degrees of freedom
## AIC: 466.92
##
## Number of Fisher Scoring iterations: 4
##
## [1] "Summary for mass predictor:"
##
## Call:
## glm(formula = form, family = binomial(link = "logit"), data = pima_clean)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.60614    0.59173  -6.094 1.10e-09 ***
## mass         0.08633    0.01705   5.062 4.14e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 469.03  on 390  degrees of freedom
## AIC: 473.03
##
## Number of Fisher Scoring iterations: 4
##
## [1] "Summary for pedigree predictor:"
##
## Call:
## glm(formula = form, family = binomial(link = "logit"), data = pima_clean)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.3912     0.2104  -6.613 3.76e-11 ***
## pedigree      1.2809     0.3289   3.895 9.82e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 481.37  on 390  degrees of freedom
## AIC: 485.37
##
## Number of Fisher Scoring iterations: 4
##
## [1] "Summary for age predictor:"
```

```
##
## Call:
## glm(formula = form, family = binomial(link = "logit"), data = pima_clean)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.05823    0.38793  -7.884 3.18e-15 ***
## age          0.07461    0.01165   6.405 1.50e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 450.67  on 390  degrees of freedom
## AIC: 454.67
##
## Number of Fisher Scoring iterations: 4
```

Regression Analysis: Multivariate Logistic Regression

```
all_vars <- glm(outcome ~ ., family = binomial(link = "logit"), data=pima_clean)
summary(all_vars) #slightly smaller .. so not contributing more than our 3 strongest values 361
```

```
##
## Call:
## glm(formula = outcome ~ ., family = binomial(link = "logit"),
##     data = pima_clean)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.004e+01  1.218e+00  -8.246  < 2e-16 ***
## pregnant     8.216e-02  5.543e-02   1.482  0.13825
## glucose      3.827e-02  5.768e-03   6.635 3.24e-11 ***
## pressure    -1.420e-03  1.183e-02  -0.120  0.90446
## triceps      1.122e-02  1.708e-02   0.657  0.51128
## insulin     -8.253e-04  1.306e-03  -0.632  0.52757
## mass         7.054e-02  2.734e-02   2.580  0.00989 **
## pedigree     1.141e+00  4.274e-01   2.669  0.00760 **
## age          3.395e-02  1.838e-02   1.847  0.06474 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 344.02  on 383  degrees of freedom
## AIC: 362.02
##
## Number of Fisher Scoring iterations: 5
```

```
strong_vars <- glm(outcome ~ glucose + age + mass + pedigree, family = binomial(link = "logit"), data=pi
summary(strong_vars)
```

```
##
## Call:
## glm(formula = outcome ~ glucose + age + mass + pedigree, family = binomial(link = "logit"),
##     data = pima_clean)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.092018   1.080251  -9.342  < 2e-16 ***
## glucose       0.036189   0.004982   7.264 3.76e-13 ***
## age           0.053012   0.013439   3.945 8.00e-05 ***
## mass          0.074449   0.020267   3.673 0.000239 ***
## pedigree      1.087129   0.419408   2.592 0.009541 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 347.23  on 387  degrees of freedom
## AIC: 357.23
##
## Number of Fisher Scoring iterations: 5
```

Comparing Model Outcome: AIC - Akaike's Information Criterion

For logistic regression, we use AIC to compare model fit, similar to adjusted R^2 where we see a penalty for additional parameters which don't contribute to the model. A lower AIC indicates a more parsimonious model.

$$\text{k = number of parameters}\text{LL = log-likelihood}AIC = 2*(k-LL)$$

Model Comparison using AIC: Which model performs best?

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Tue, Apr 30, 2024 - 21:35:27

Best fit model analysis: Log-odds Ratios

```
## Waiting for profiling to be done...
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Tue, Apr 30, 2024 - 21:35:27

In the context of our analysis, we apply to the Bonferroni correction, adjusting the significance threshold by dividing it by the number of comparisons being made.

Even with this new threshold, the p-values for all four variables remain significant.

Table 6:

	Dependent variable:									
	outcome									
	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age	All vars	Strong vars
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
pregnant	0.167***								0.082	
	(0.034)								(0.055)	
glucose		0.042***							0.038***	0.036***
		(0.005)							(0.006)	(0.005)
pressure			0.034***						−0.001	
			(0.009)						(0.012)	
triceps				0.054***					0.011	
				(0.011)					(0.017)	
insulin					0.006***				−0.001	
					(0.001)				(0.001)	
mass						0.086***			0.071***	0.074***
						(0.017)			(0.027)	(0.020)
pedigree							1.281***		1.141***	1.087***
							(0.329)		(0.427)	(0.419)
age								0.075***	0.034*	0.053***
								(0.012)	(0.018)	(0.013)
Constant	−1.285***	−6.096***	−3.168***	−2.326***	−1.613***	−3.606***	−1.391***	−3.058***	−10.041***	−10.092***
	(0.167)	(0.630)	(0.677)	(0.358)	(0.204)	(0.592)	(0.210)	(0.388)	(1.218)	(1.080)
Observations	392	392	392	392	392	392	392	392	392	392
Log Likelihood	−236.517	−193.333	−241.579	−235.978	−231.459	−234.516	−240.686	−225.334	−172.011	−173.617
Akaike Inf. Crit.	477.035	390.666	487.159	475.956	466.917	473.031	485.372	454.668	362.021	357.235
Note:										*p<0.1; **p<0.05; ***p<0.01

Table 7:

	Dependent variable:
	outcome
	Strong vars
glucose	1.037 (1.027, 1.047)
	t = 7.264
	p = 0.000
age	1.054 (1.028, 1.083)
	t = 3.945
	p = 0.0001
mass	1.077 (1.036, 1.122)
	t = 3.673
	p = 0.0003
pedigree	2.966 (1.327, 6.871)
	t = 2.592
	p = 0.010
Constant	0.00004 (0.00000, 0.0003)
	t = −9.342
	p = 0.000
Observations	392
Log Likelihood	−173.617
Akaike Inf. Crit.	357.235
Note:	*p<0.1; **p<0.05; ***p<0.01

Conclusions

The model with lowest AIC contained the four strongest performing parameters

- Blood Glucose
- Mass
- Age
- Pedigree

Limitations

- The study findings might not be generalisable to a larger population.
- Null/Missing data limited the amount of viable observations.

References