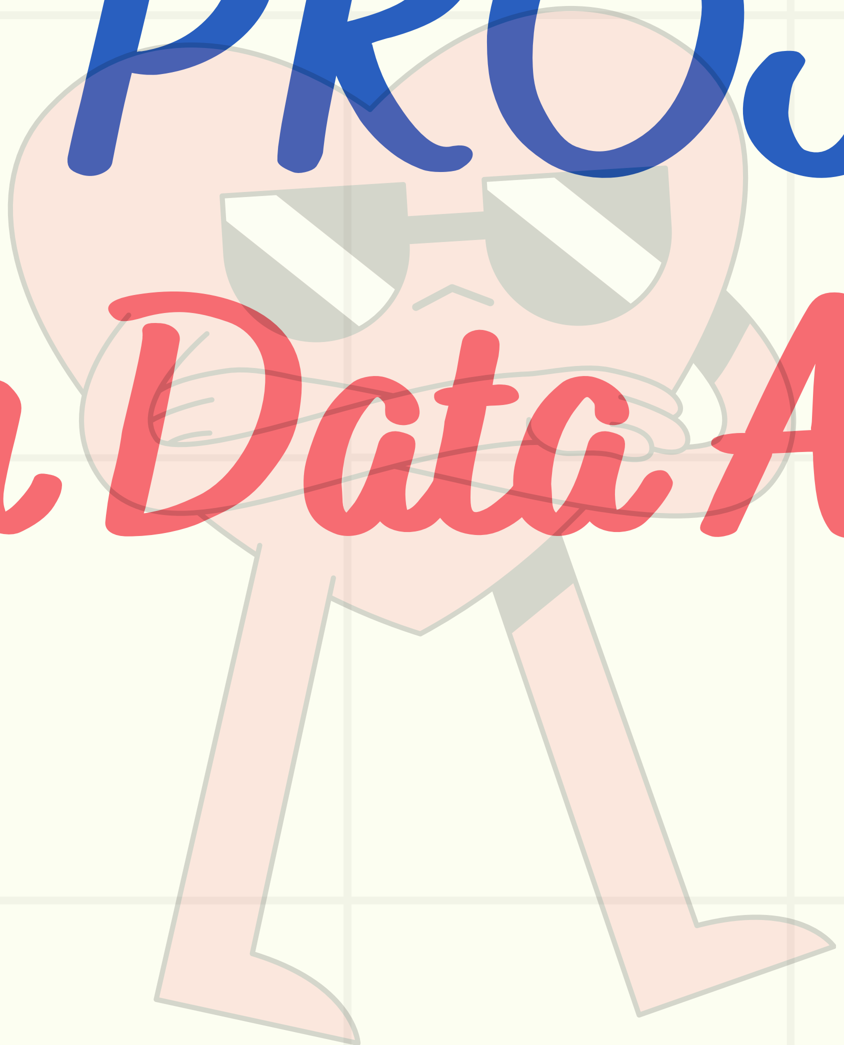
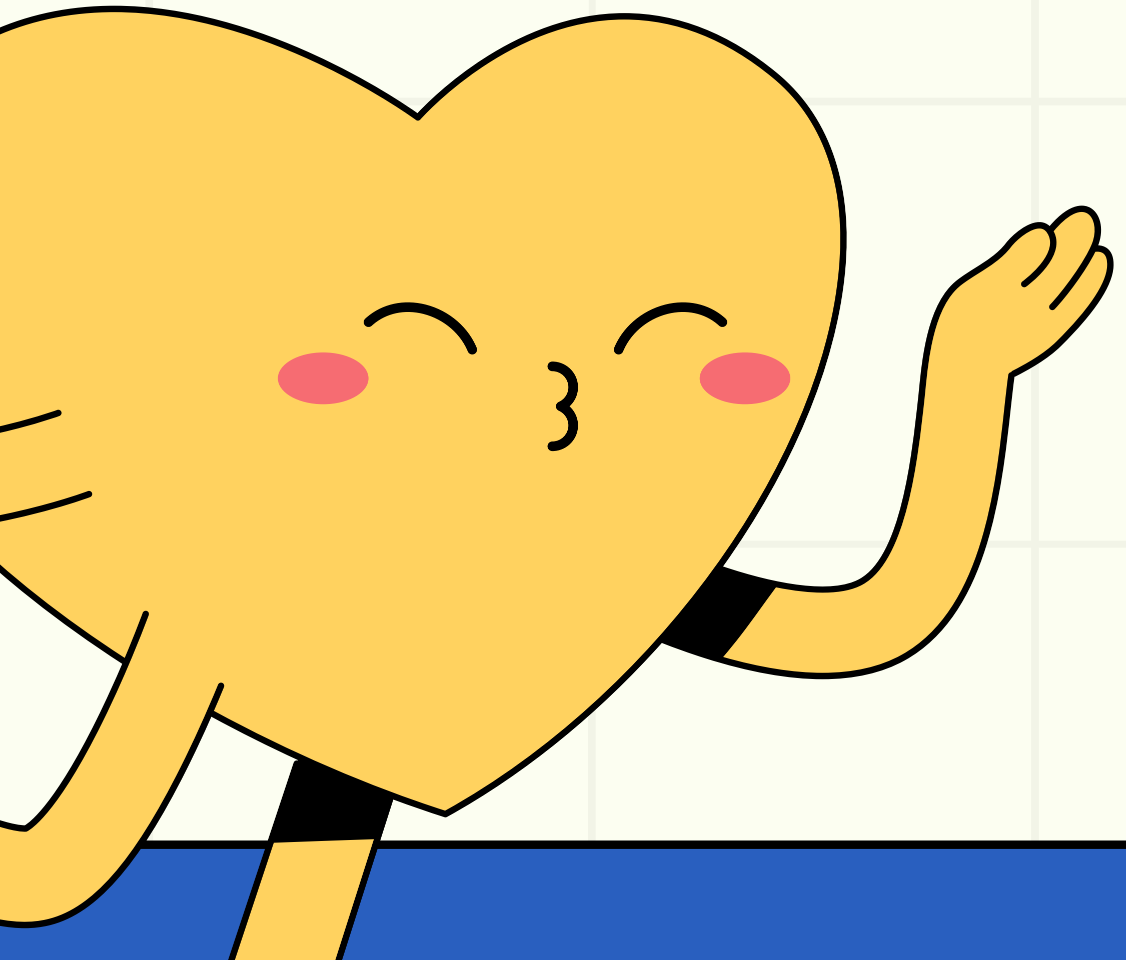
Kelompok 5

MINI PROJECT

Bayesian Data Analysis





Aldo Oktavianus -2702234081
ArielDhipta Tarliman -2702234636
William - 2702225373

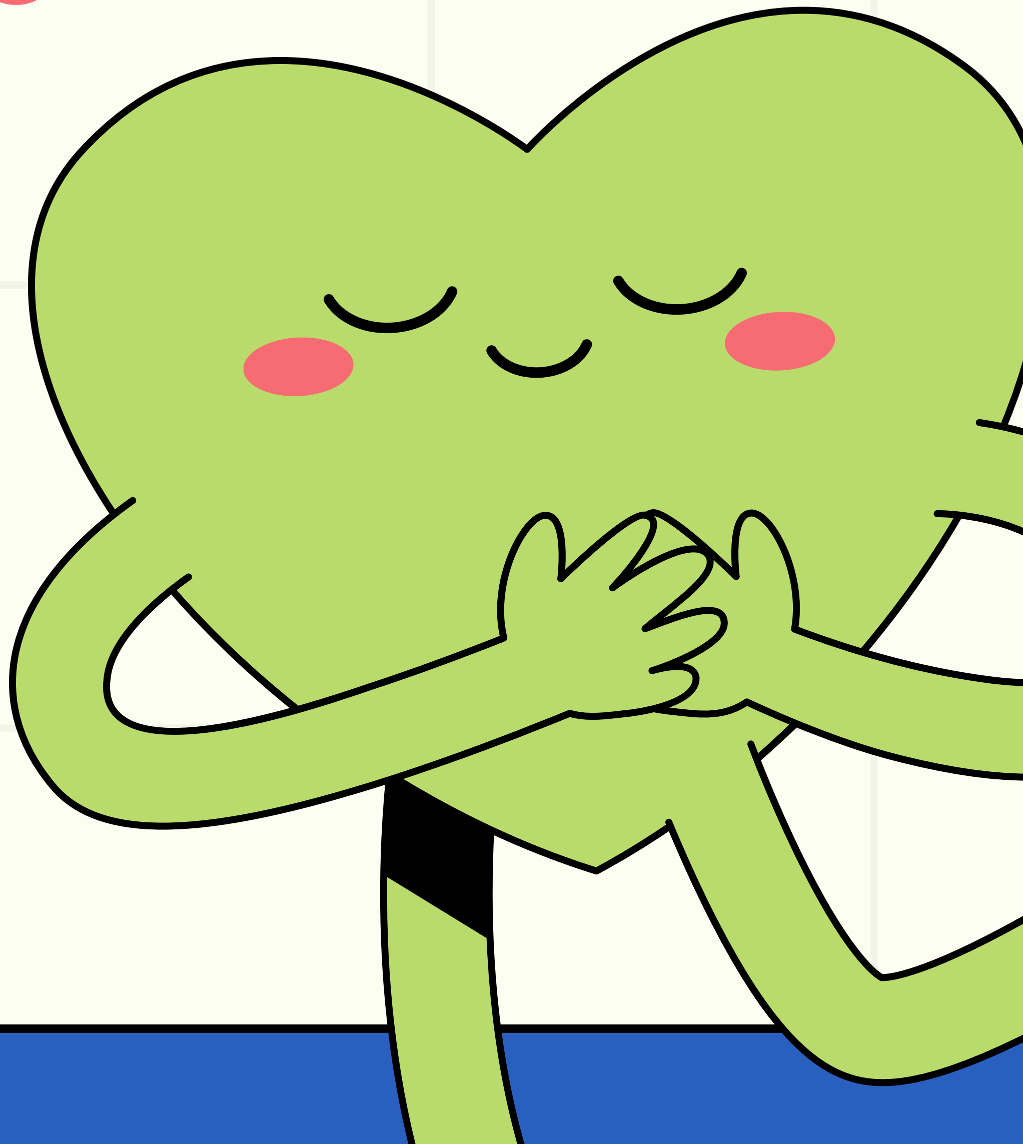
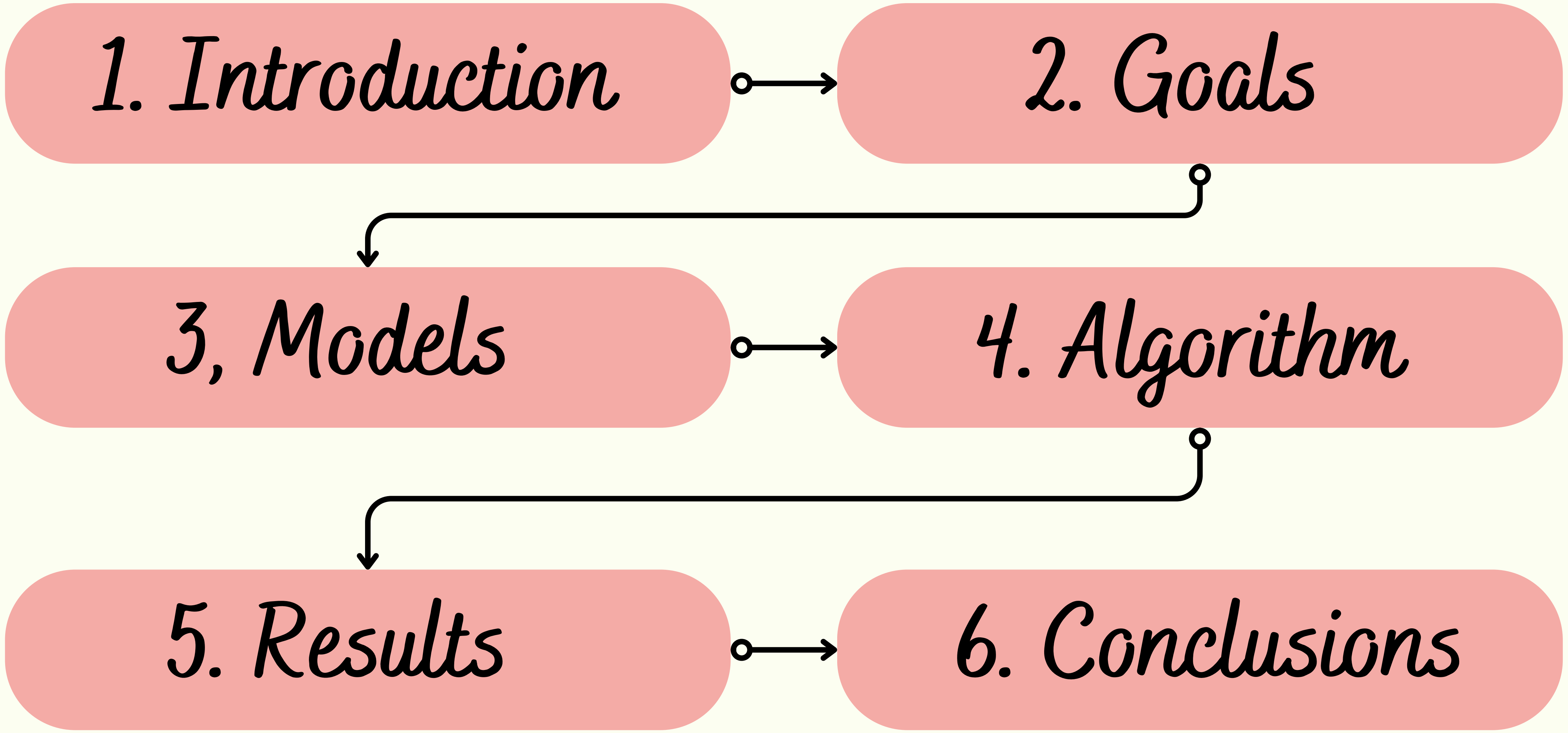


TABLE CONTENTS



1. Introduction

2. Goals

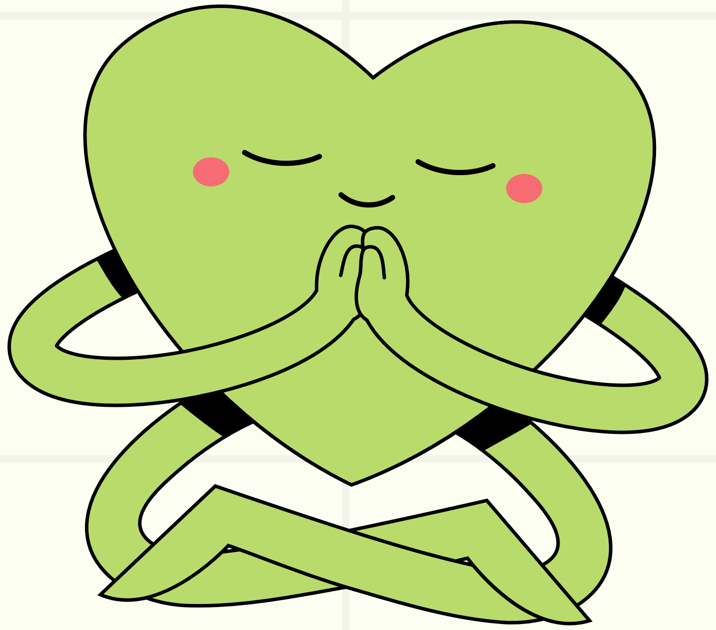
3, Models

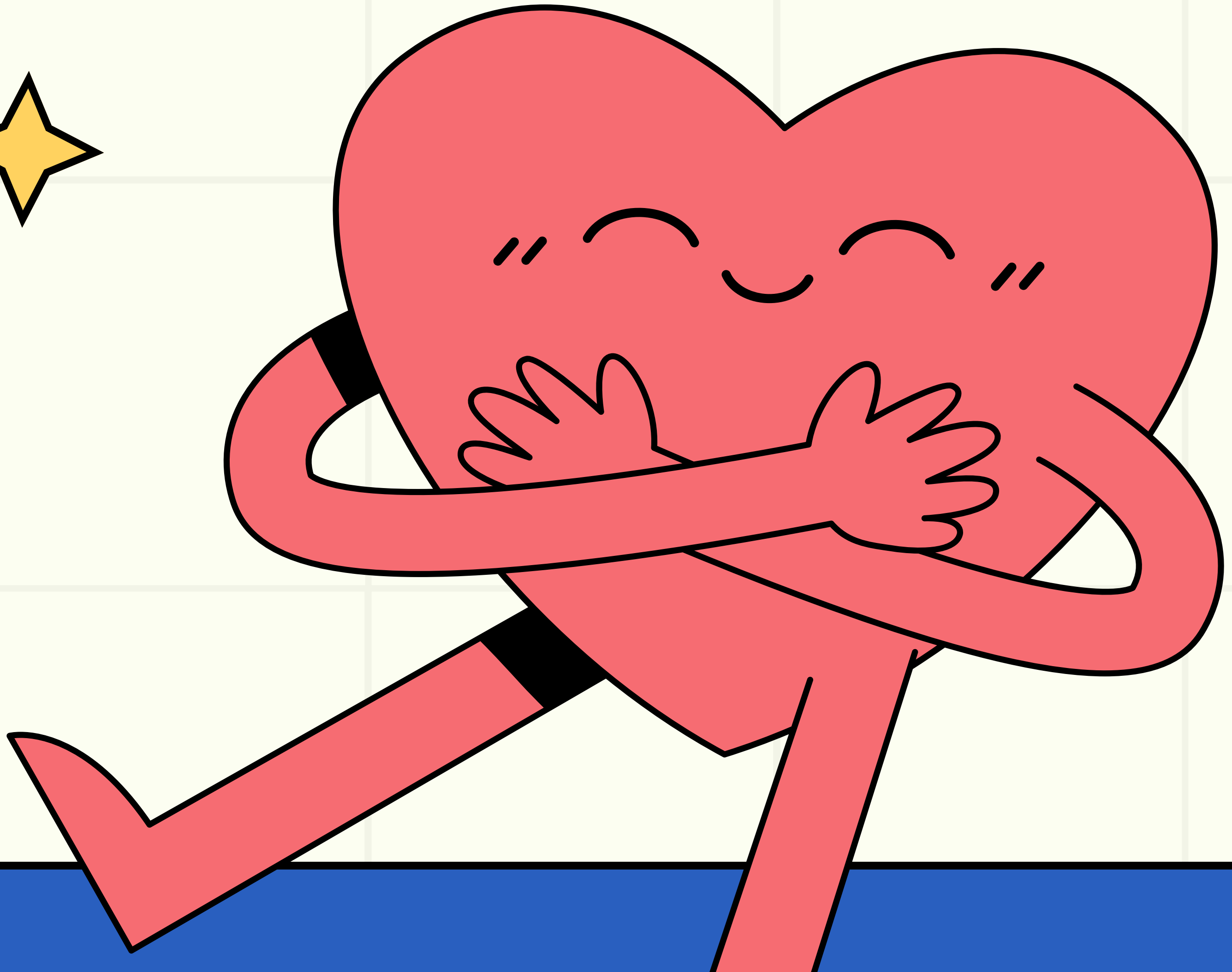
4. Algorithm

5. Results

6. Conclusions

INTRODUCTION



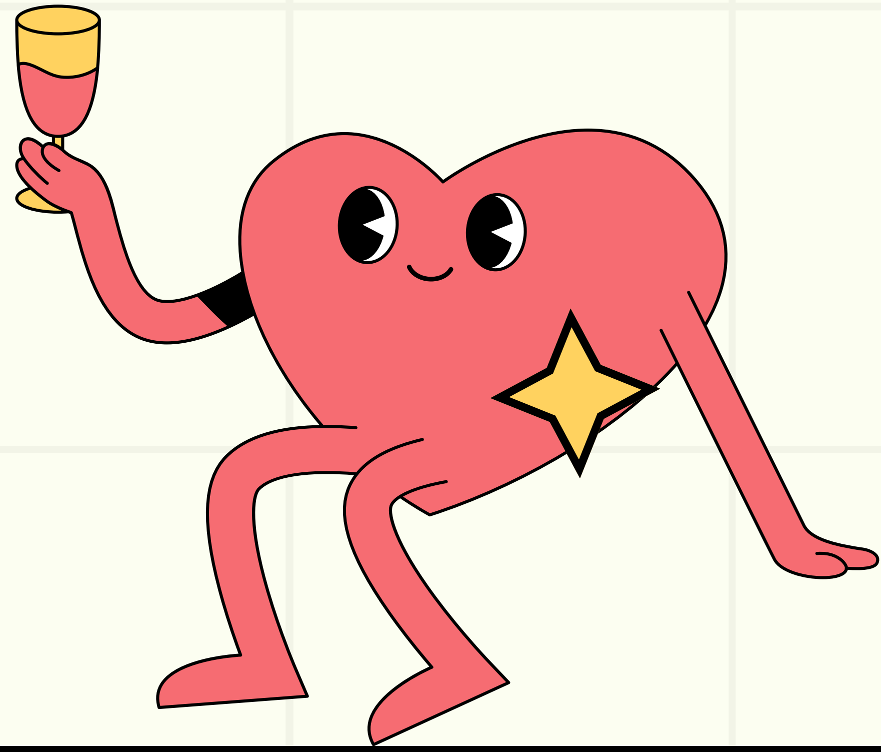


Stroke is a medical condition characterized by a sudden disruption of blood flow to the brain, leading to cell damage. It is a leading cause of death and disability worldwide, with various risk factors such as hypertension, heart disease, age, and lifestyle choices. Early identification of stroke risk factors is crucial for prevention and improving patient outcomes.

Dataset Overview

The dataset used in this project comprises 5,110 rows and 12 columns which consist:

- **id**: Unique identifier for each patient.
- **gender**: Gender of the patient (Male/Female).
- **age**: Age of the patient.
- **hypertension**: Presence of hypertension (0 = No, 1 = Yes).
- **heart_disease**: Presence of heart disease (0 = No, 1 = Yes).
- **ever_married**: Marital status (Yes/No).
- **work_type**: Type of occupation (e.g., Private, Self-employed, Govt job).
- **Residence_type**: Living environment (Urban/Rural).
- **avg_glucose_level**: Average glucose level in the blood.
- **bmi**: Body Mass Index.
- **smoking_status**: Smoking habits (e.g., never smoked, formerly smoked, smokes).
- **stroke**: Stroke occurrence (1 = Yes, 0 = No).



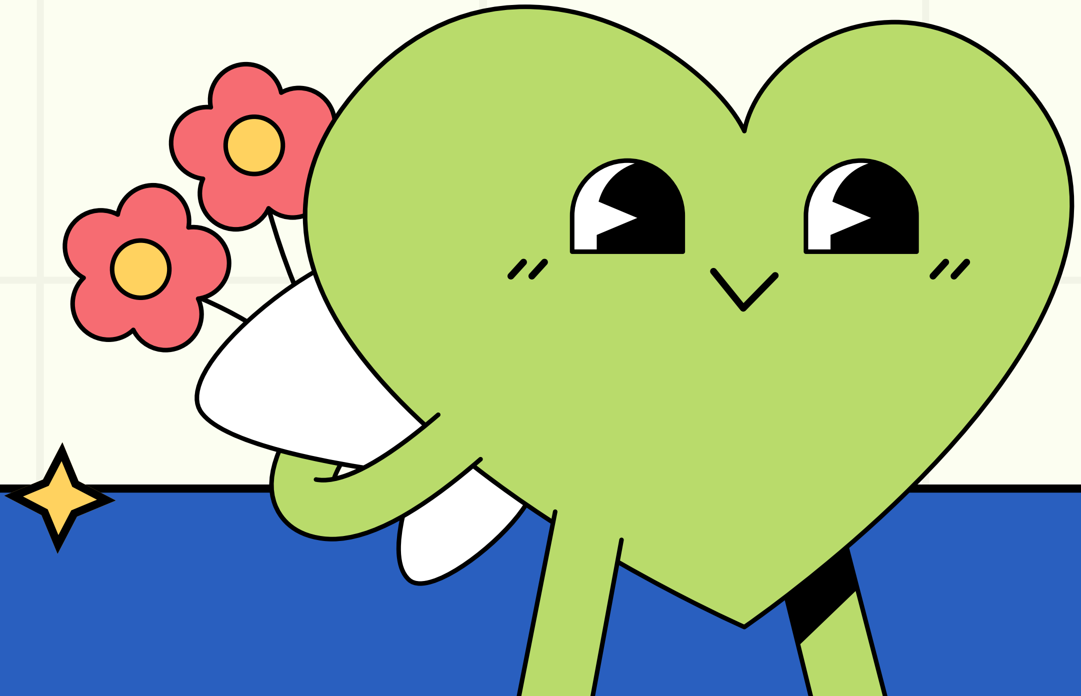
Models

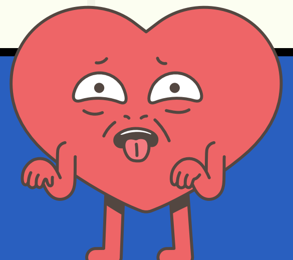
01

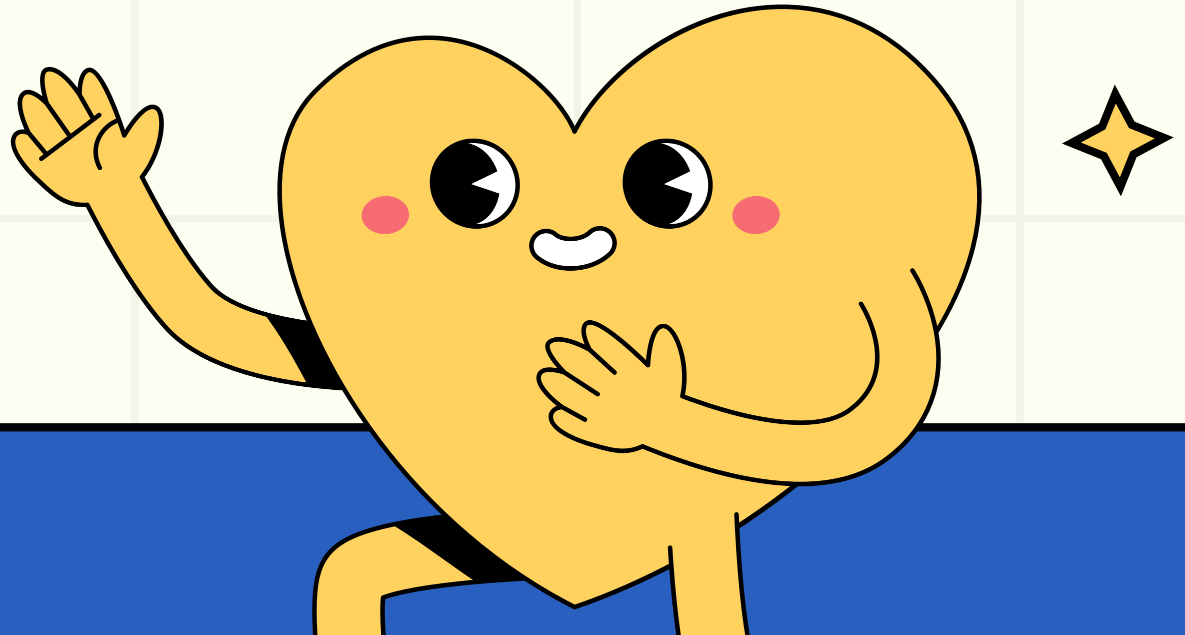
BAYESSIAN LOGISTIC REGRESSION with ALL FEATURES

02

BAYESSIAN LOGISTIC REGRESSION with SELECTED FEATURES



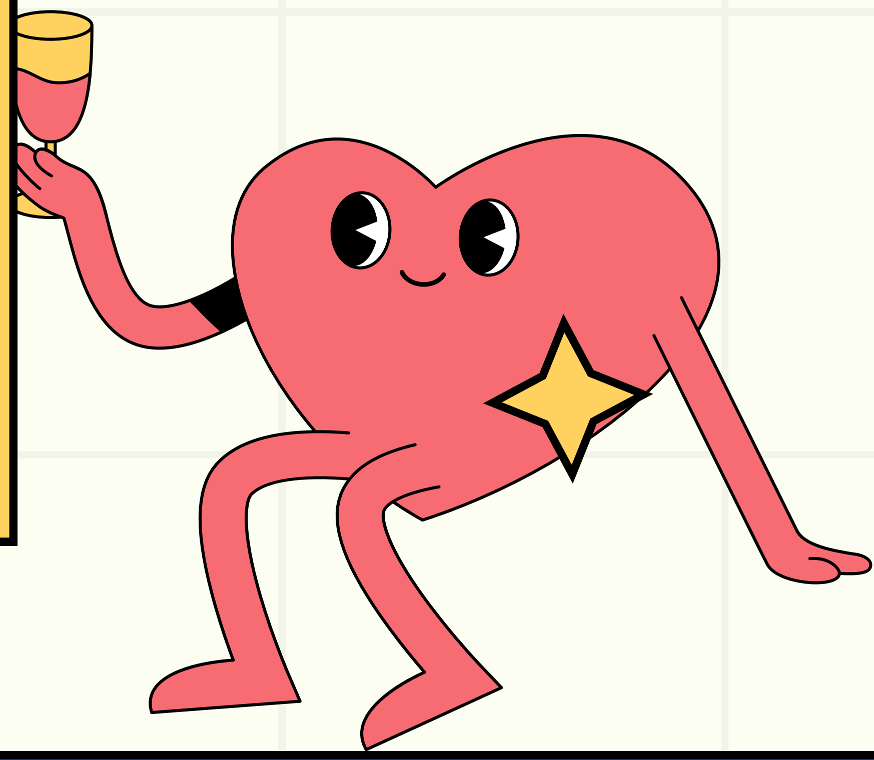




ENCODING

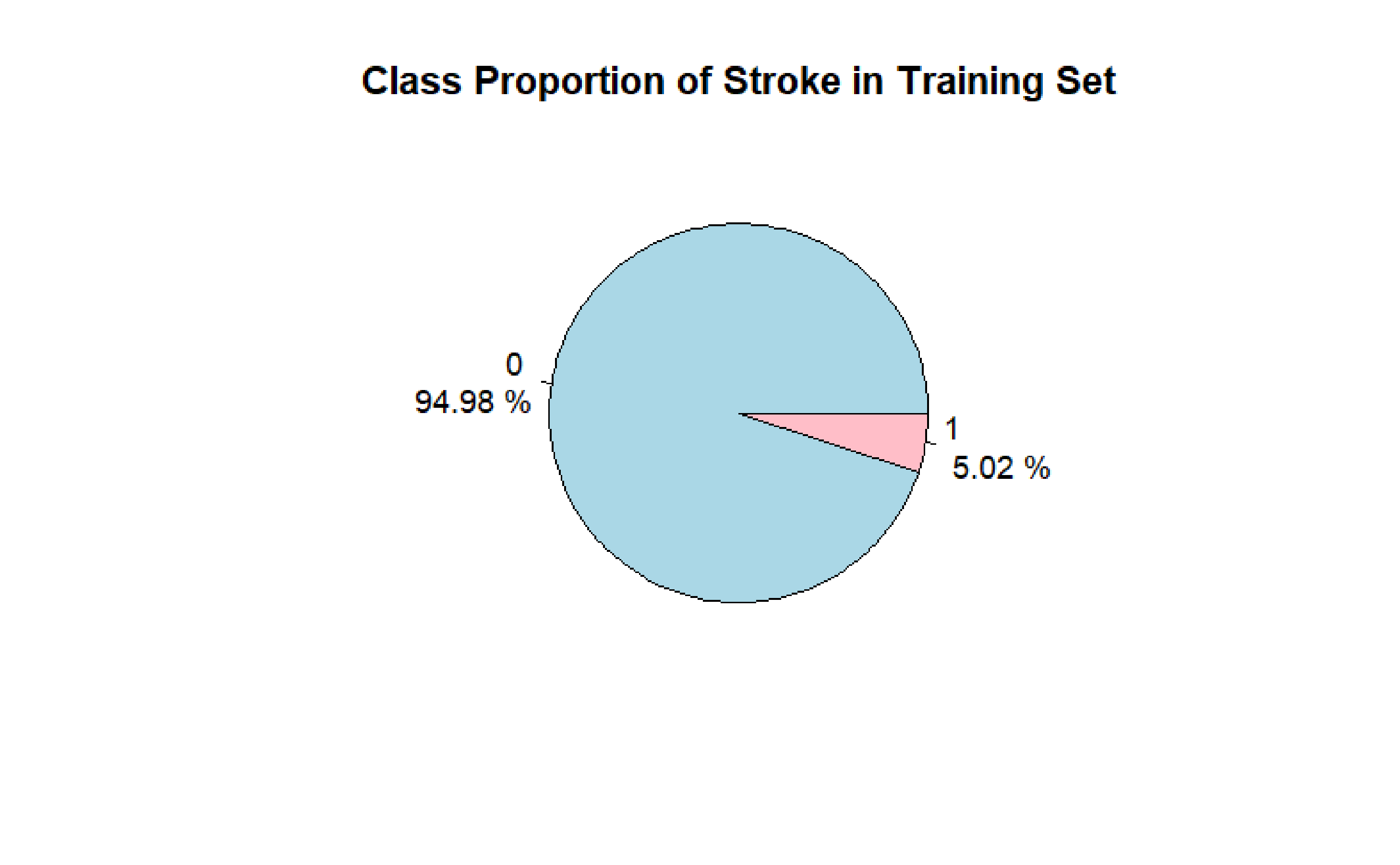
label encoding:

- gender
- ever_married
- work_type
- smoking_statys
- residence_type

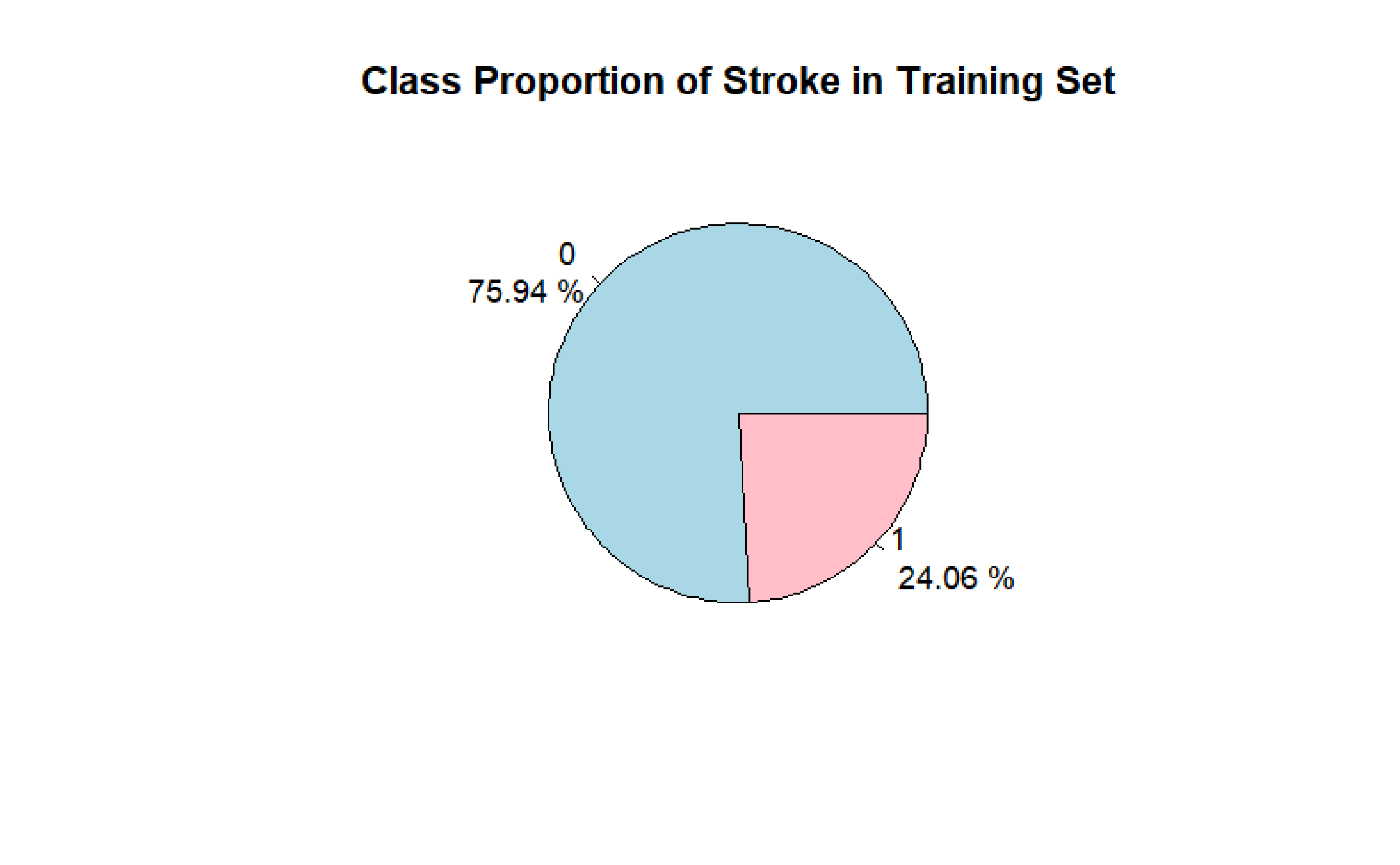


Our Assumption

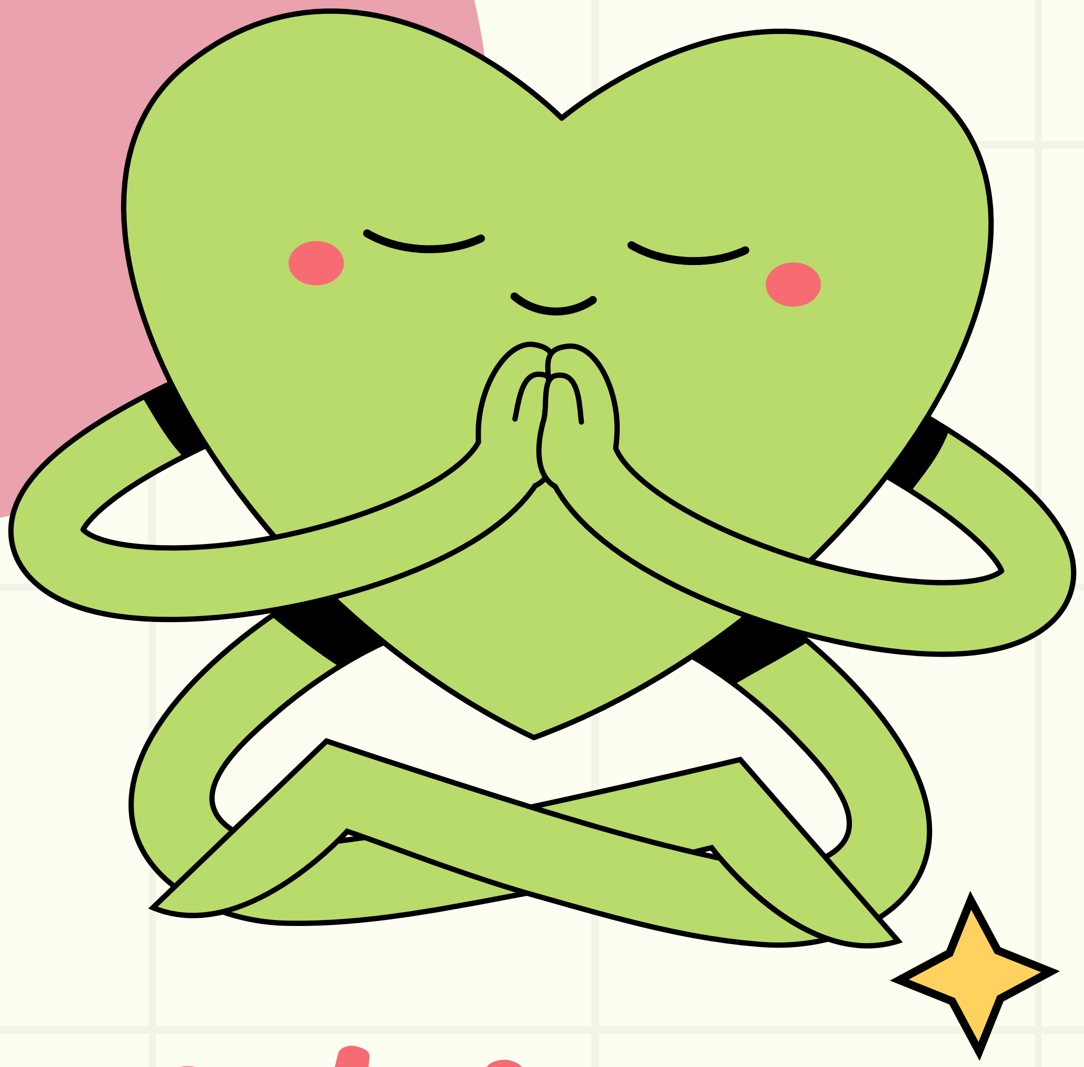
Severe Class Imbalance



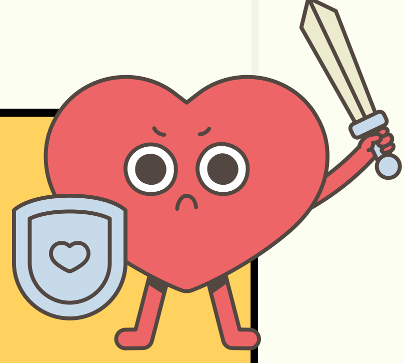
Before SMOTE



After SMOTE



Model 1
All Features



LIKELIHOOD:

Bernoulli Distribution

Y[i] ~ dbern(pi[i])

Reasonings:

- Binary output
- Compatible with logit relationship
- Fits independent events

PRIOR:

Normal Distribution

- Coeficients (β[j]): β[j]   dnorm(0,0.01)
- Intercept (α): α   dnorm(0,0.01)

Reasonings:

- Uninformative priors
- Flexibility

Model 2: Selected Features

LIKELIHOOD:

Bernoulli Distribution

Y[i] ~ dbern(pi[i])

Reasonings:

- Binary output
- Compatible with logit relationship
- Fits independent events

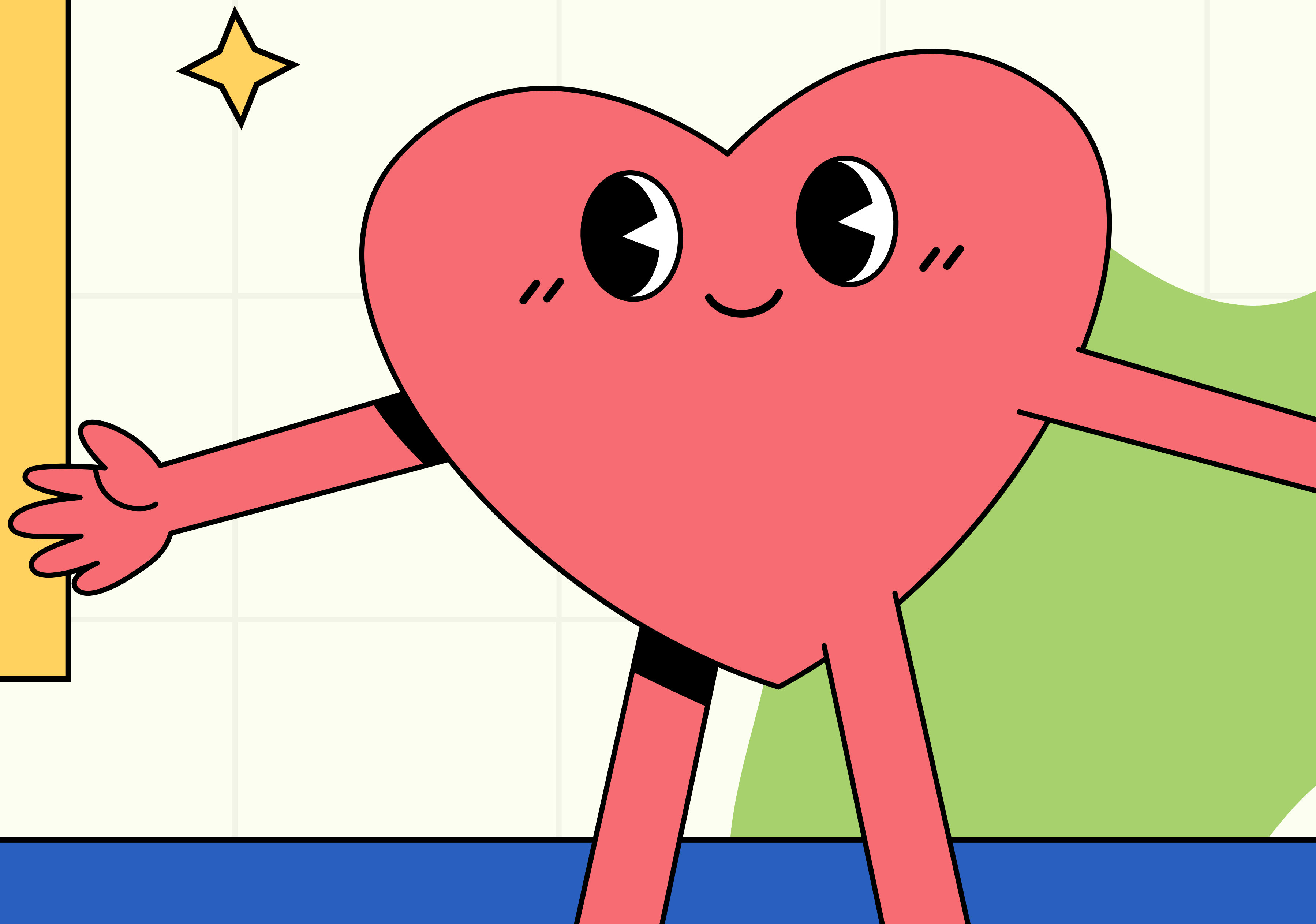
PRIOR:

Normal Distribution

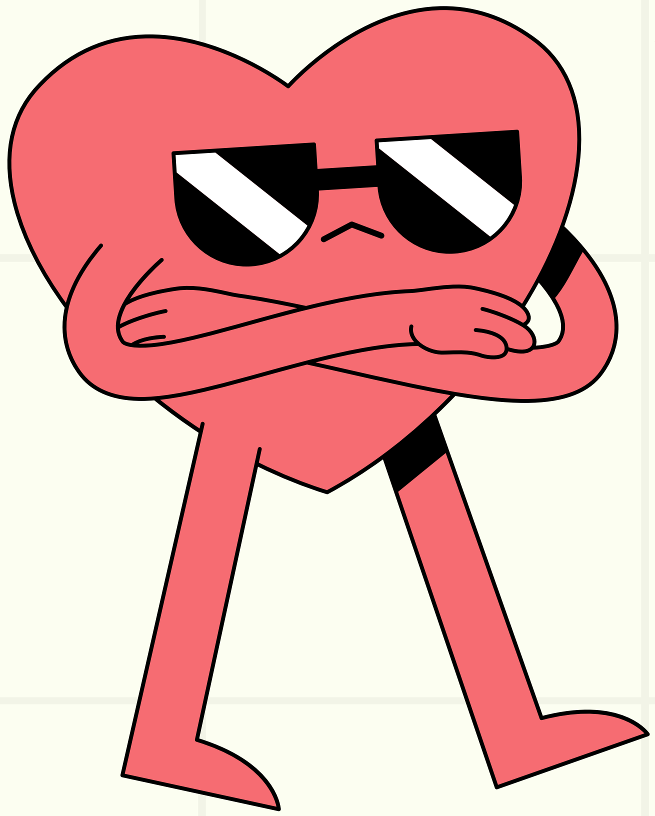
- Coeficients (β[j]): β[j] dnorm(0,0.01)
- Intercept (α): α dnorm(0,0.01)

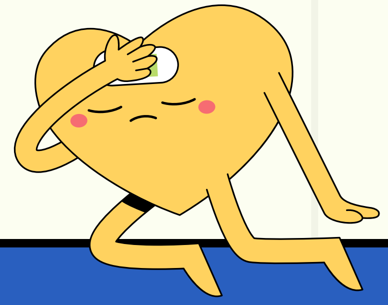
Reasonings:

- Uninformative priors
- Scalability

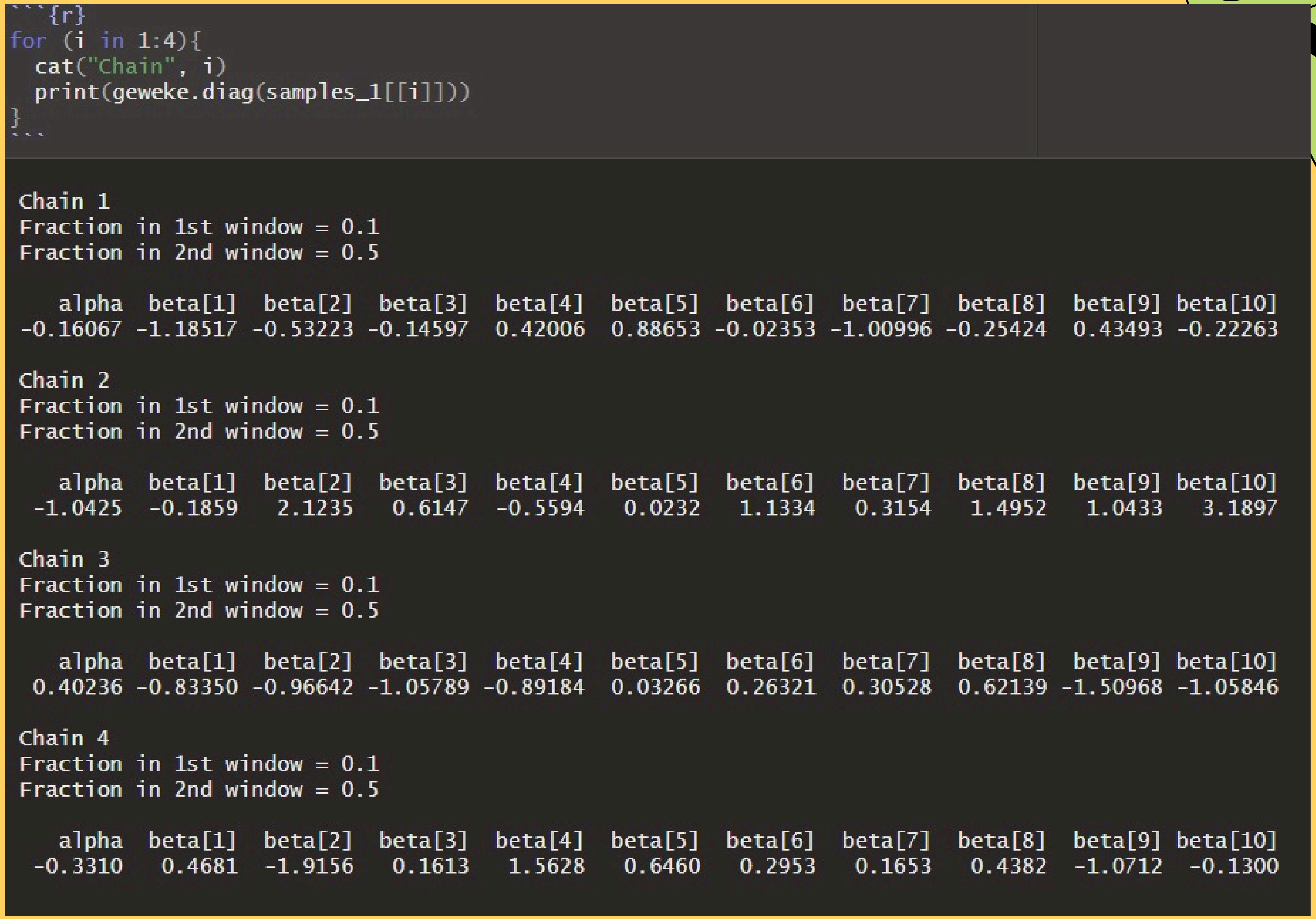


Convergence Diagnostic Model #1





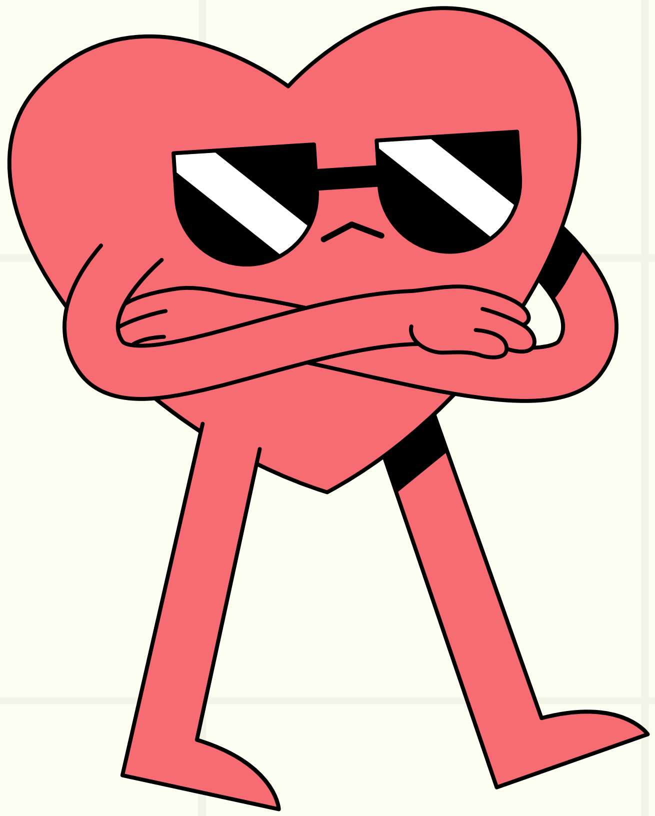
Geweke



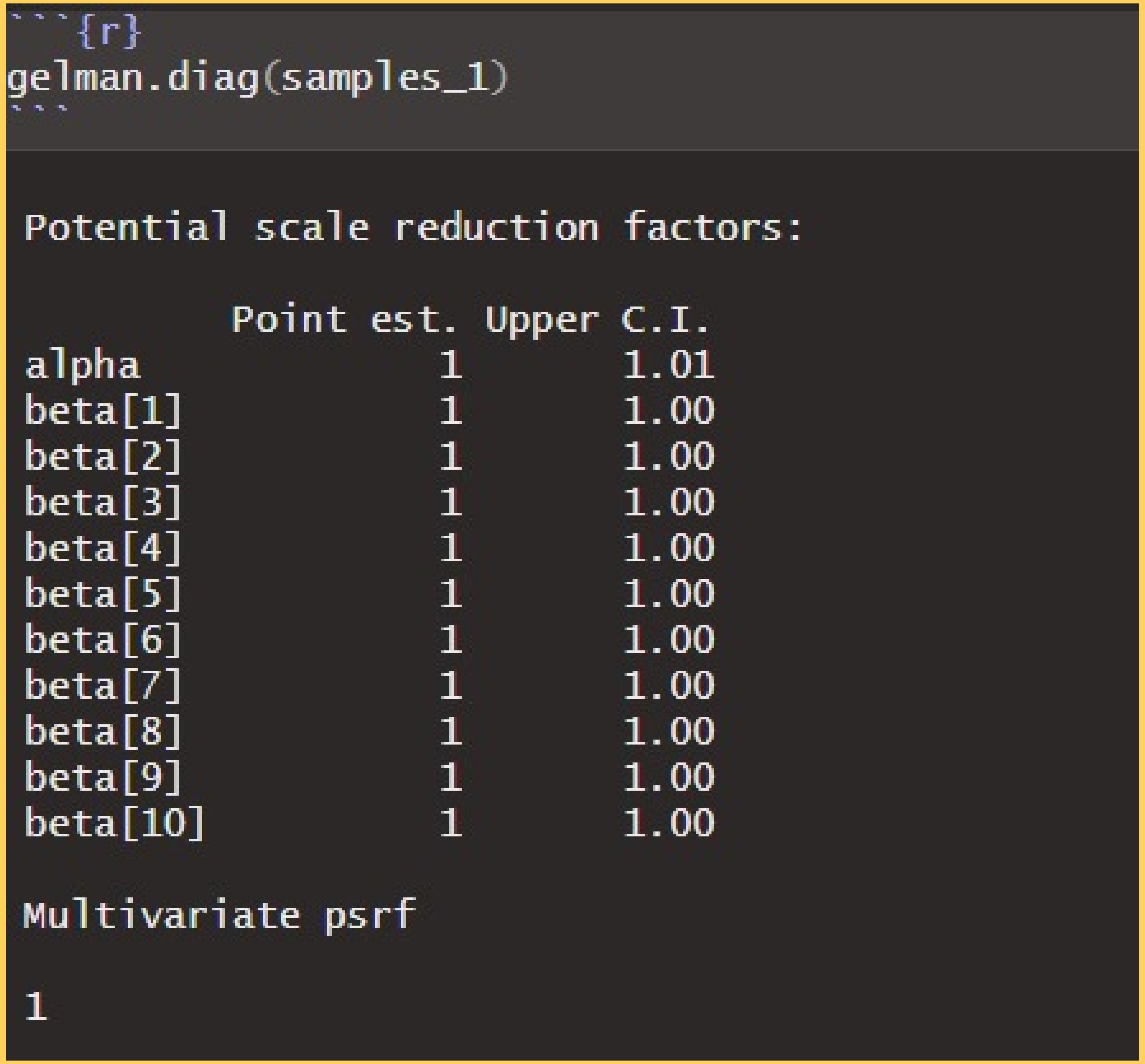
~~~
```{r}
for (i in 1:4){
  cat("Chain", i)
  print(geweke.diag(samples_1[[i]]))
}
```
~~~

```
Chain 1
Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

   alpha  beta[1]  beta[2]  beta[3]  beta[4]  beta[5]  beta[6]  beta[7]  beta[8]  beta[9] beta[10]
-0.16067 -1.18517 -0.53223 -0.14597  0.42006  0.88653 -0.02353 -1.00996 -0.25424  0.43493 -0.22263

Chain 2
Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

   alpha  beta[1]  beta[2]  beta[3]  beta[4]  beta[5]  beta[6]  beta[7]  beta[8]  beta[9] beta[10]
 -1.0425  -0.1859   2.1235   0.6147  -0.5594   0.0232   1.1334   0.3154   1.4952   1.0433   3.1897

Chain 3
Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

   alpha  beta[1]  beta[2]  beta[3]  beta[4]  beta[5]  beta[6]  beta[7]  beta[8]  beta[9] beta[10]
 0.40236 -0.83350 -0.96642 -1.05789 -0.89184  0.03266  0.26321  0.30528  0.62139 -1.50968 -1.05846

Chain 4
Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

   alpha  beta[1]  beta[2]  beta[3]  beta[4]  beta[5]  beta[6]  beta[7]  beta[8]  beta[9] beta[10]
 -0.3310   0.4681  -1.9156   0.1613   1.5628   0.6460   0.2953   0.1653   0.4382  -1.0712  -0.1300
```

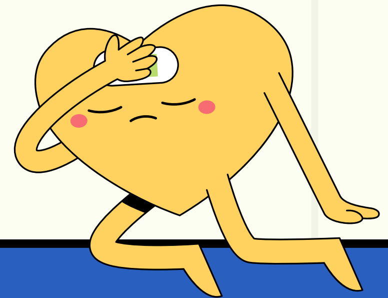
Convergence Diagnostic
Model #1



Gelman Rubin

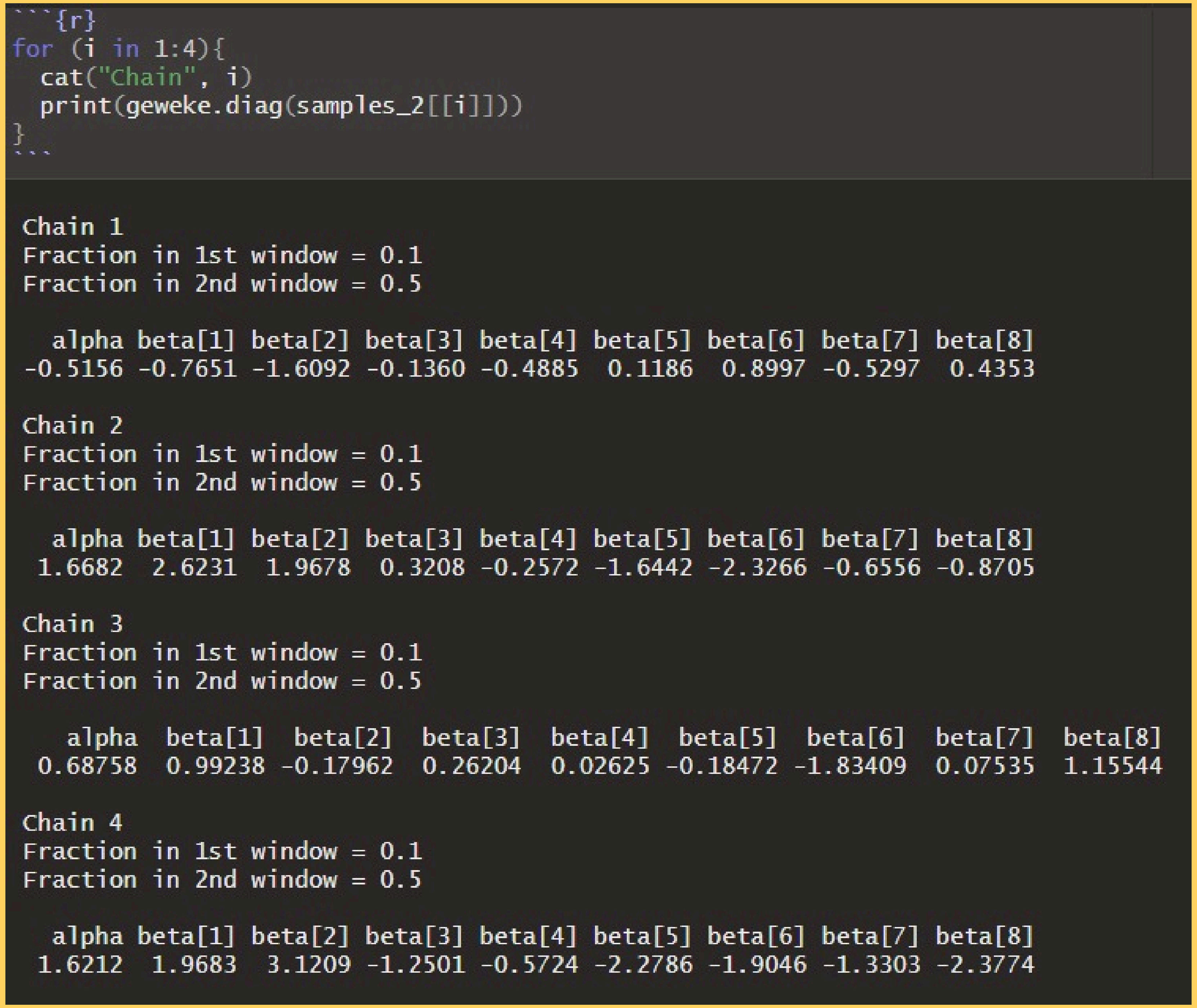


~~~
```{r}
gelman.diag(samples_1)
```

Potential scale reduction factors:

          Point est. Upper C.I.
alpha              1       1.01
beta[1]            1       1.00
beta[2]            1       1.00
beta[3]            1       1.00
beta[4]            1       1.00
beta[5]            1       1.00
beta[6]            1       1.00
beta[7]            1       1.00
beta[8]            1       1.00
beta[9]            1       1.00
beta[10]           1       1.00

Multivariate psrf

1
~~~



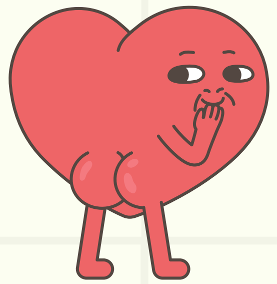
Convergence Diagnostic Model #2

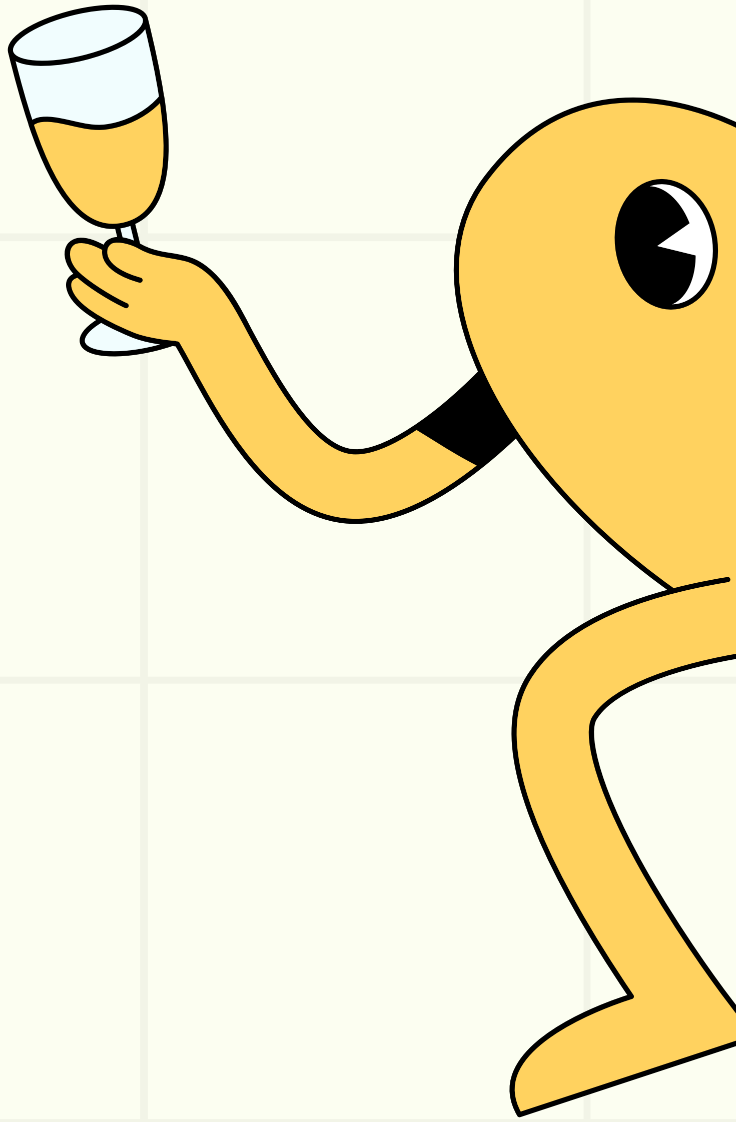
Geweke

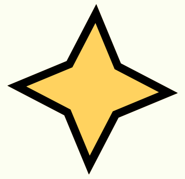


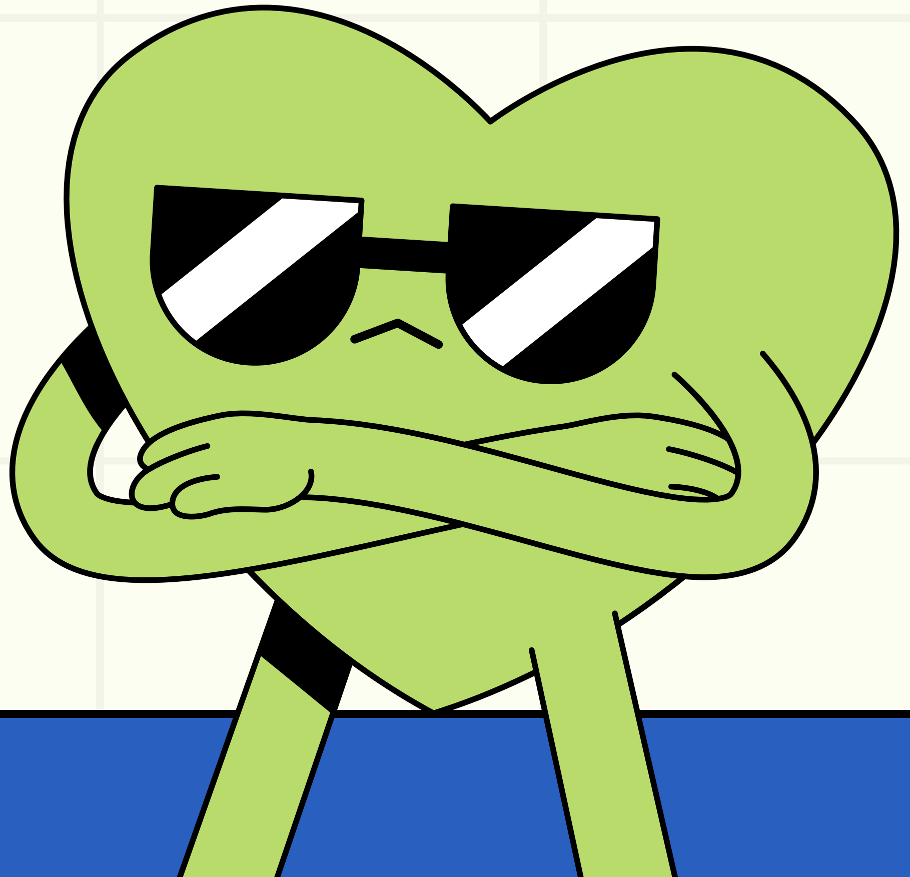
~~~
```{r}
for (i in 1:4){
  cat("Chain", i)
  print(geweke.diag(samples_2[[i]]))
}
```
~~~

```
Chain 1
Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

  alpha beta[1] beta[2] beta[3] beta[4] beta[5] beta[6] beta[7] beta[8] 
-0.5156 -0.7651 -1.6092 -0.1360 -0.4885  0.1186  0.8997 -0.5297  0.4353 

Chain 2
Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

  alpha beta[1] beta[2] beta[3] beta[4] beta[5] beta[6] beta[7] beta[8] 
 1.6682  2.6231  1.9678  0.3208 -0.2572 -1.6442 -2.3266 -0.6556 -0.8705 

Chain 3
Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

   alpha  beta[1]  beta[2]  beta[3]  beta[4]  beta[5]  beta[6]  beta[7]  beta[8] 
 0.68758  0.99238 -0.17962  0.26204  0.02625 -0.18472 -1.83409  0.07535  1.15544 

Chain 4
Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

  alpha beta[1] beta[2] beta[3] beta[4] beta[5] beta[6] beta[7] beta[8] 
 1.6212  1.9683  3.1209 -1.2501 -0.5724 -2.2786 -1.9046 -1.3303 -2.3774 
```

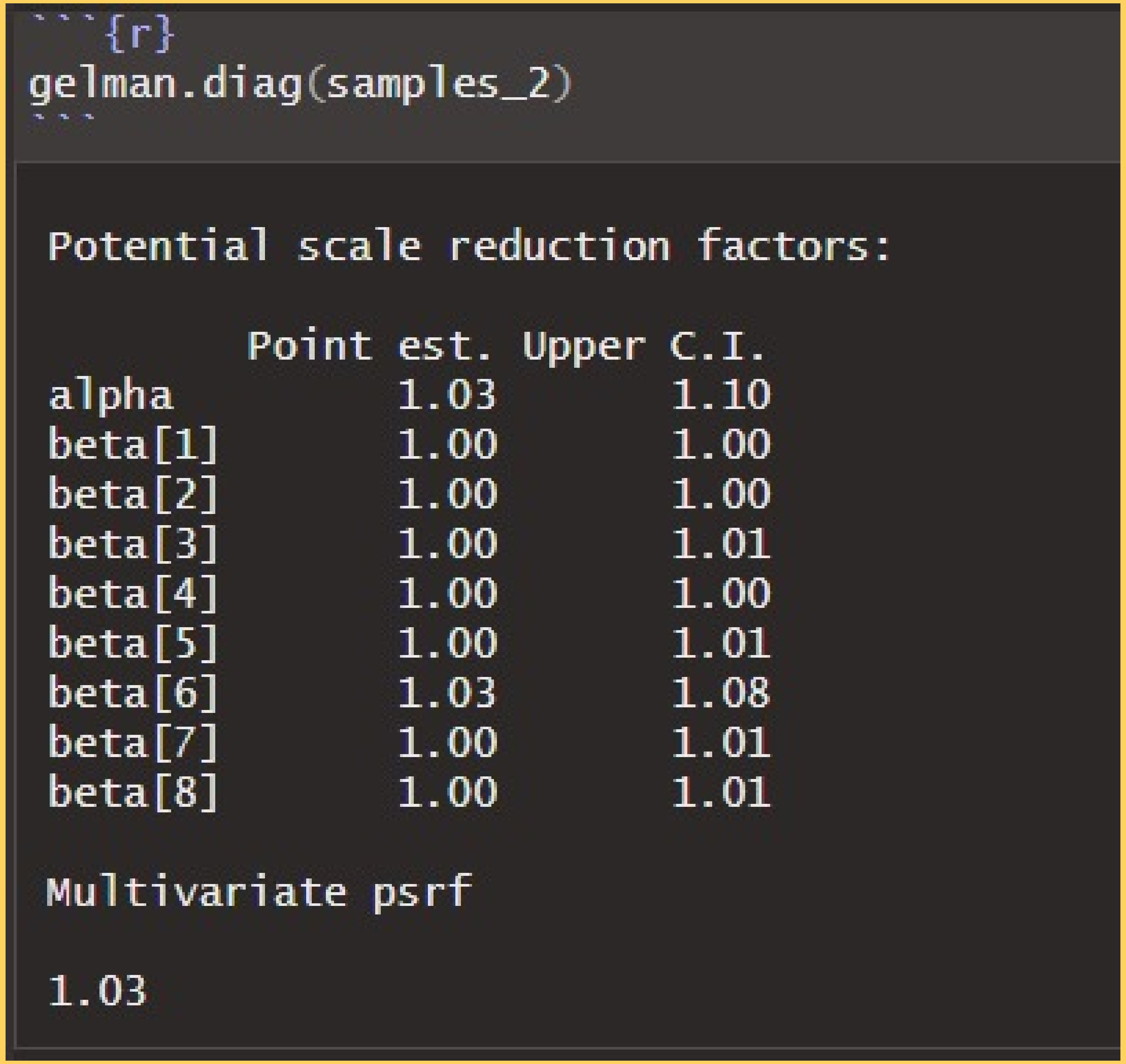








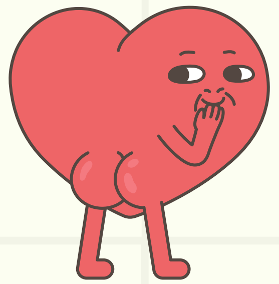
Gelman Rubin

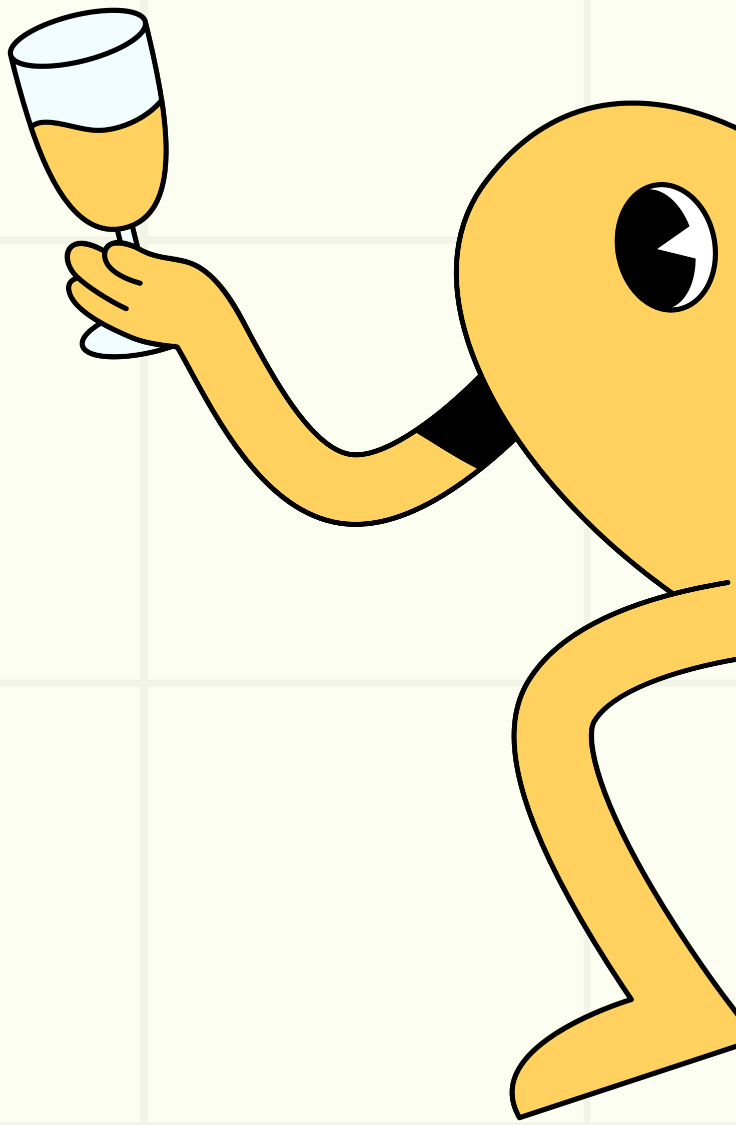


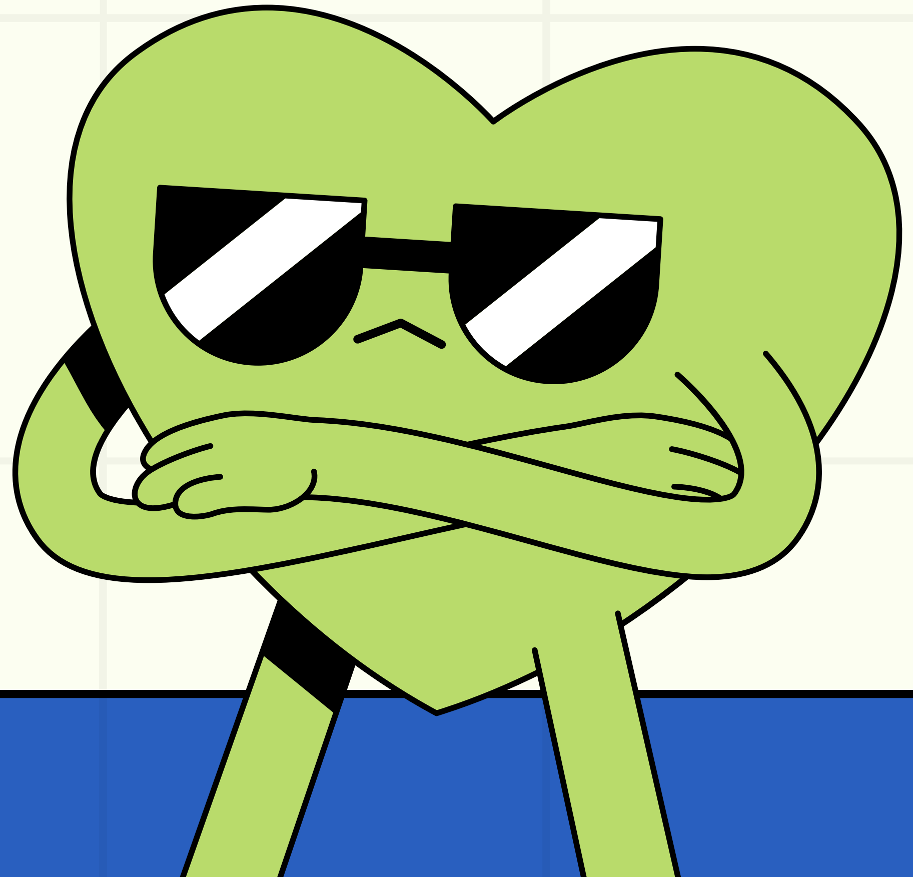
~~~
```{r}
gelman.diag(samples_2)
```
~~~

```
Potential scale reduction factors:

         Point est. Upper C.I.
alpha          1.03       1.10
beta[1]        1.00       1.00
beta[2]        1.00       1.00
beta[3]        1.00       1.01
beta[4]        1.00       1.00
beta[5]        1.00       1.01
beta[6]        1.03       1.08
beta[7]        1.00       1.01
beta[8]        1.00       1.01

Multivariate psrf

1.03
```

Convergence Diagnostic Model #2

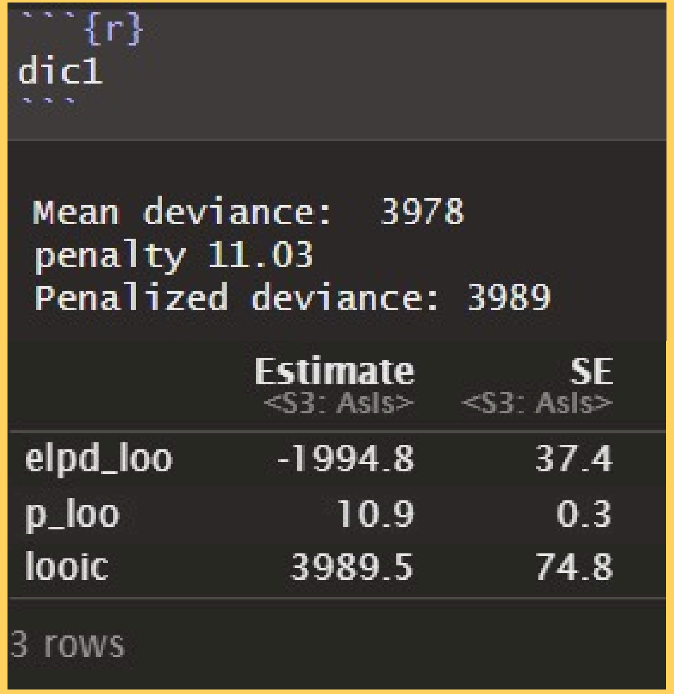






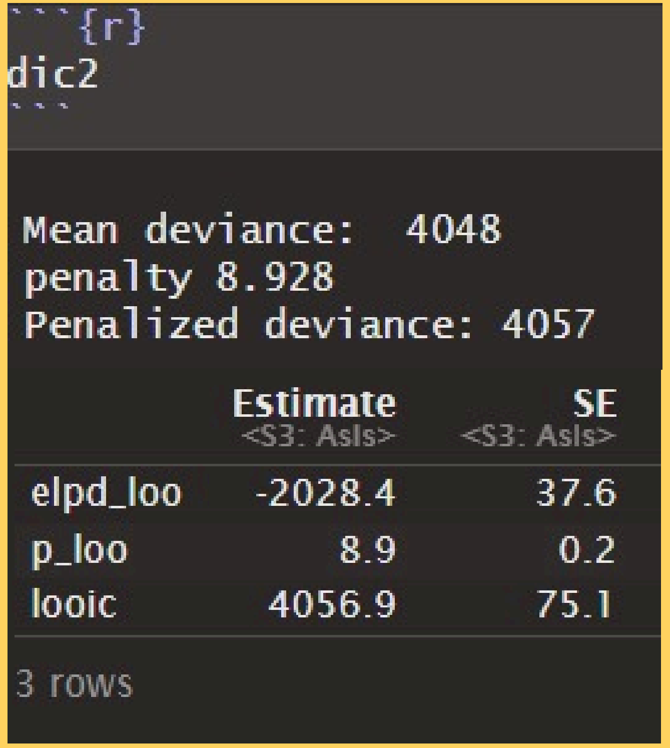
Model Comparison

DIC & WAIC (Model 1)



~~~
```{r}
dic1
```

 Mean deviance:  3978
 penalty 11.03
 Penalized deviance: 3989

                Estimate          SE
              <S3: AsIs>  <S3: AsIs>
elpd_loo         -1994.8        37.4
p_loo               10.9         0.3
looic             3989.5        74.8

3 rows
~~~

DIC & WAIC (Model 2)



~~~
```{r}
dic2
```

Mean deviance:  4048
penalty 8.928
Penalized deviance: 4057

              Estimate          SE
            <S3: AsIs>  <S3: AsIs>
elpd_loo       -2028.4        37.6
p_loo              8.9         0.2
looic           4056.9        75.1

3 rows
~~~

Deviance Information Criterion (DIC)

Mean Deviance: goodness of fit of the model

Penalty: complexity of the model

Penalized Deviance: combination of fit and penalty

Model 1 is preferable overall because it has a lower DIC and better data fit (lower mean deviance), despite being slightly more complex than Model 2.

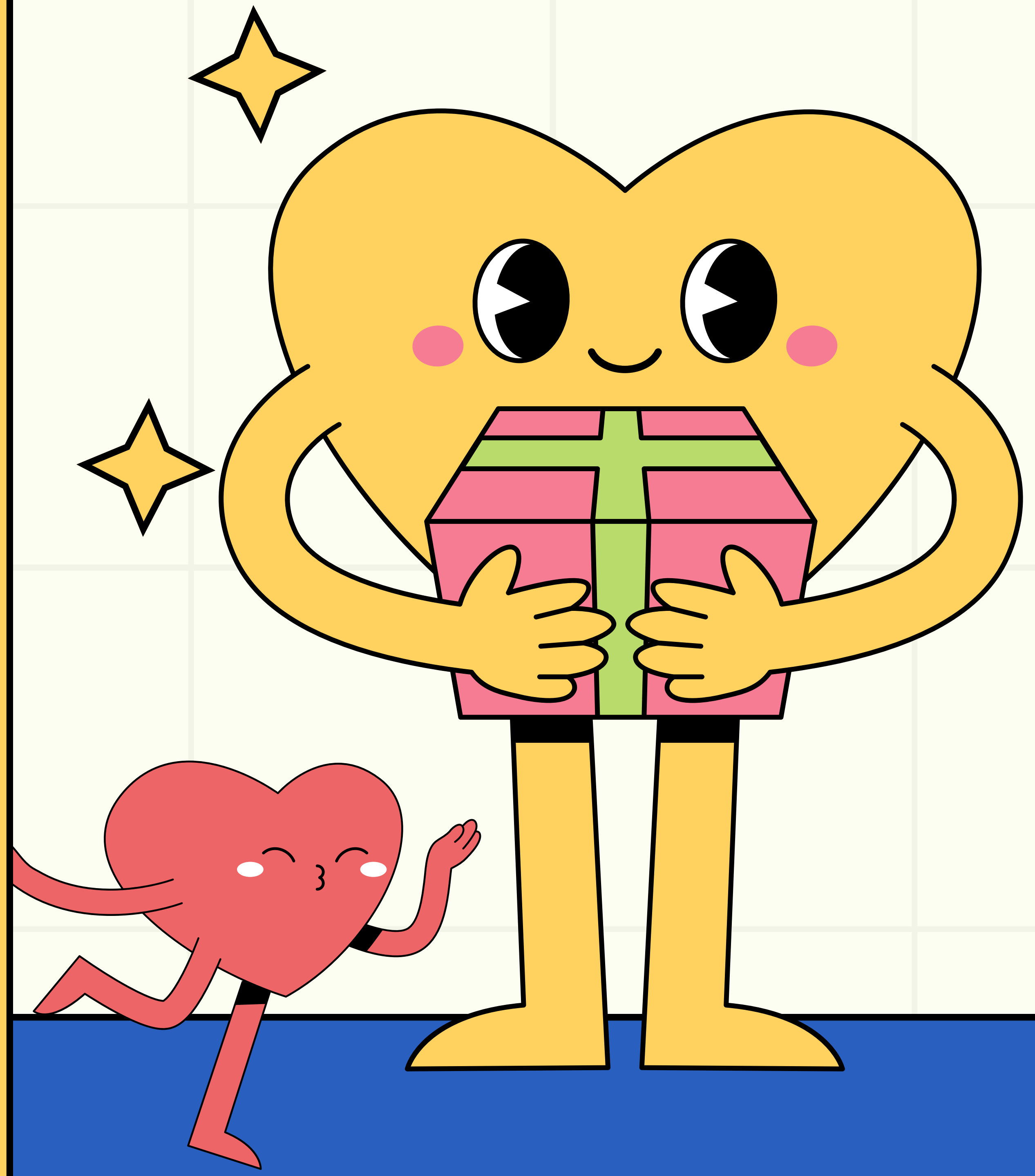
Watanabe-Akaike Information Criterion (WAIC)

Expected Log Predictive Density: measures model predicts unseen data

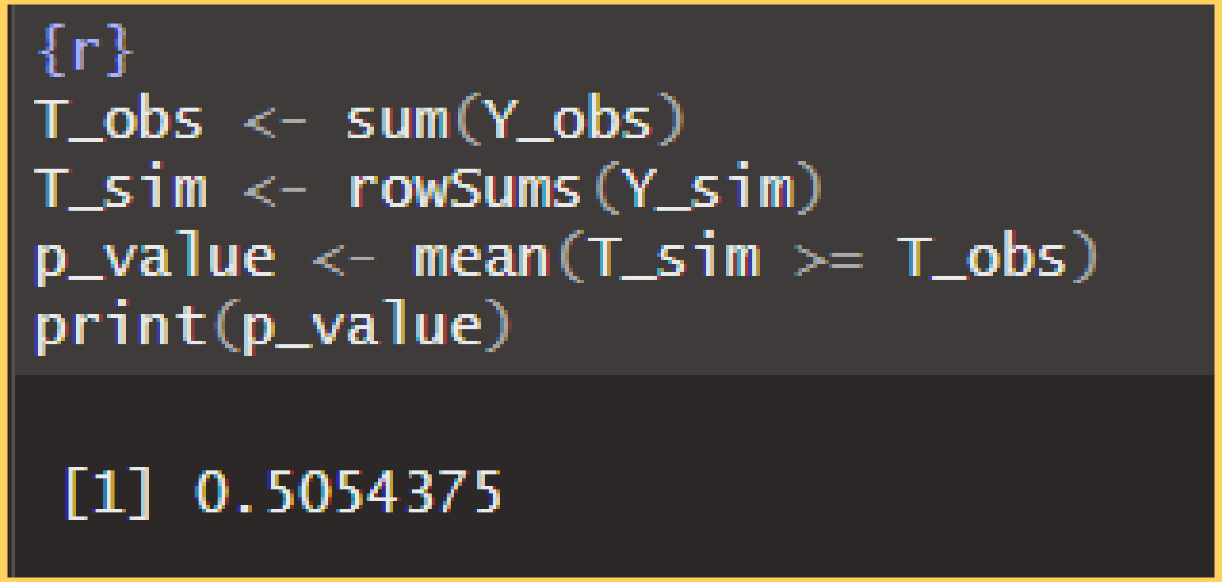
Penalized: model's complexity

Lower Overall Information Criterion: balances between model fit and compexity

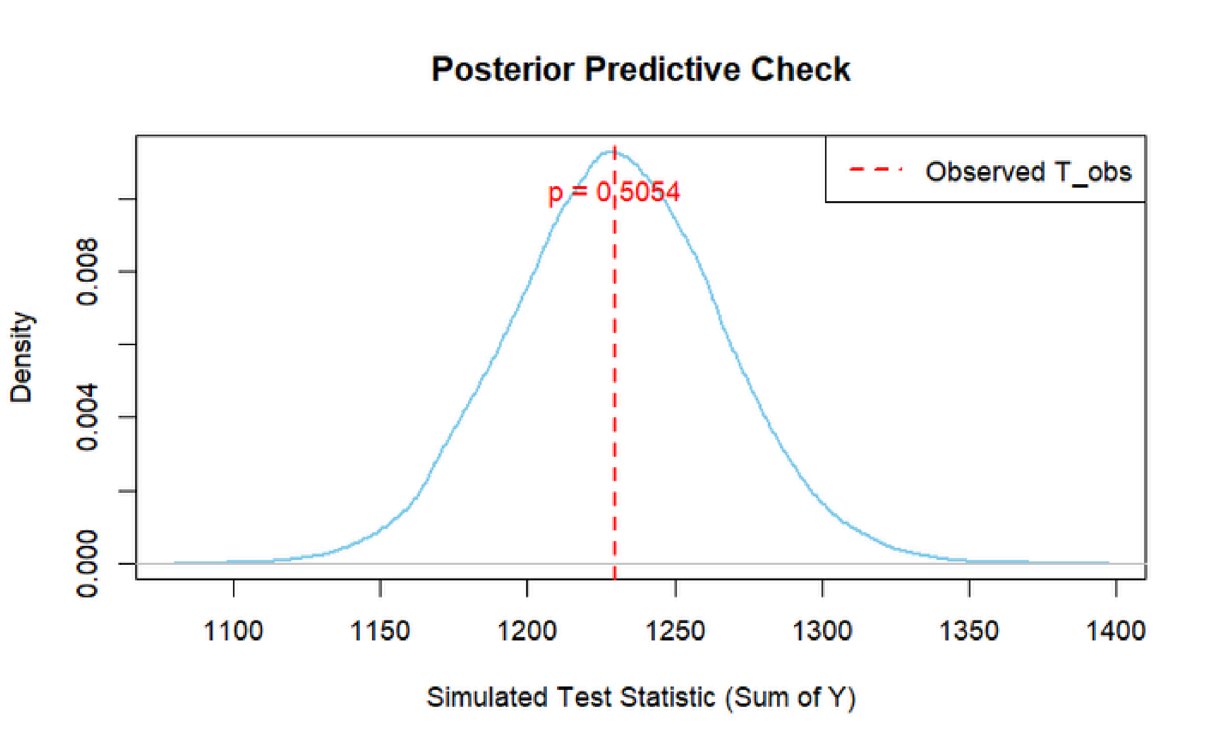
Model 1 is still preferable because has better predictive (higher) and lower overall information criterion (lower)



Posterior Predictive Checks



```
{r}
T_obs <- sum(Y_obs)
T_sim <- rowSums(Y_sim)
p_value <- mean(T_sim >= T_obs)
print(p_value)

 [1] 0.5054375
```

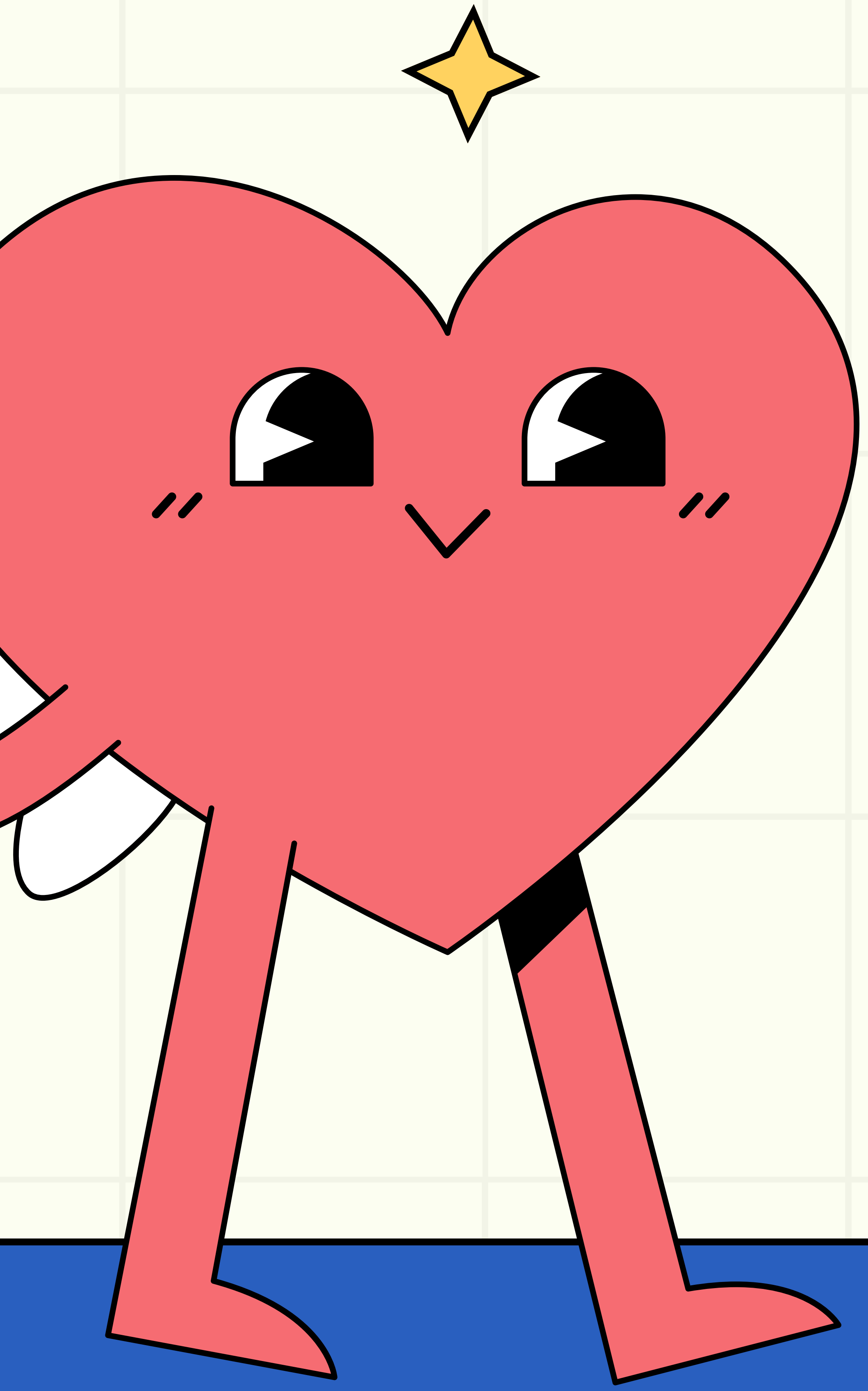


Posterior Predictive Checks

P Value:
The posterior predictive check suggests that the Bayesian logistic regression model provides a good fit to the observed data.

The Bayesian p-value of 0.5054 reflects that about 50.54% of the simulated test statistics are equal to or greater than the observed value

This balance suggests that the model captures the variability in the data well, with no evidence of systematic overestimation or underestimation. Overall, the results indicate that the model is appropriate for describing the observed data.



Thankyou Very Much