Bayesian Mini Project

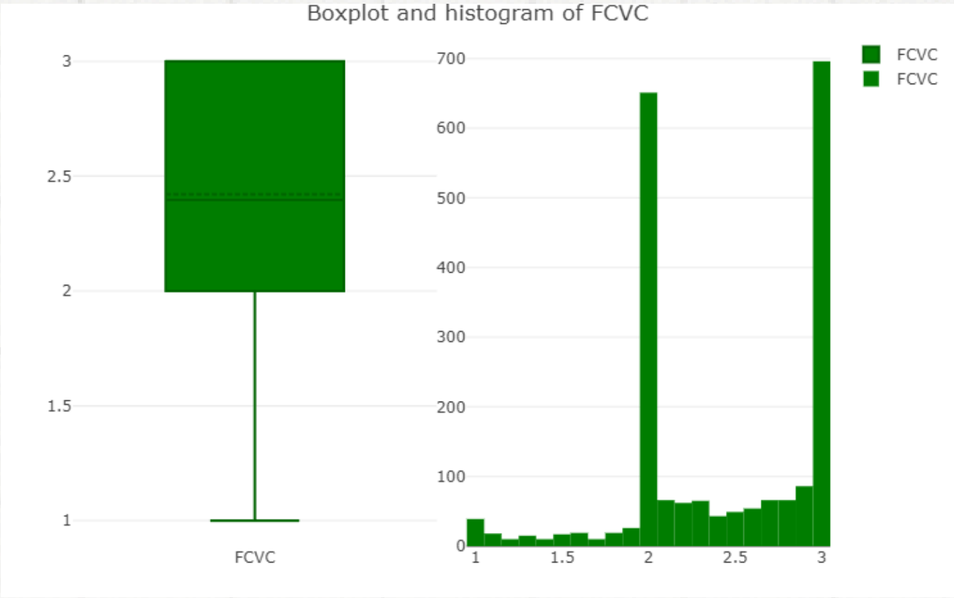
Andreas Immanuel Lukito
Aaron Winston Gho
Rayhan Permana
Vanness Christian Rulianto

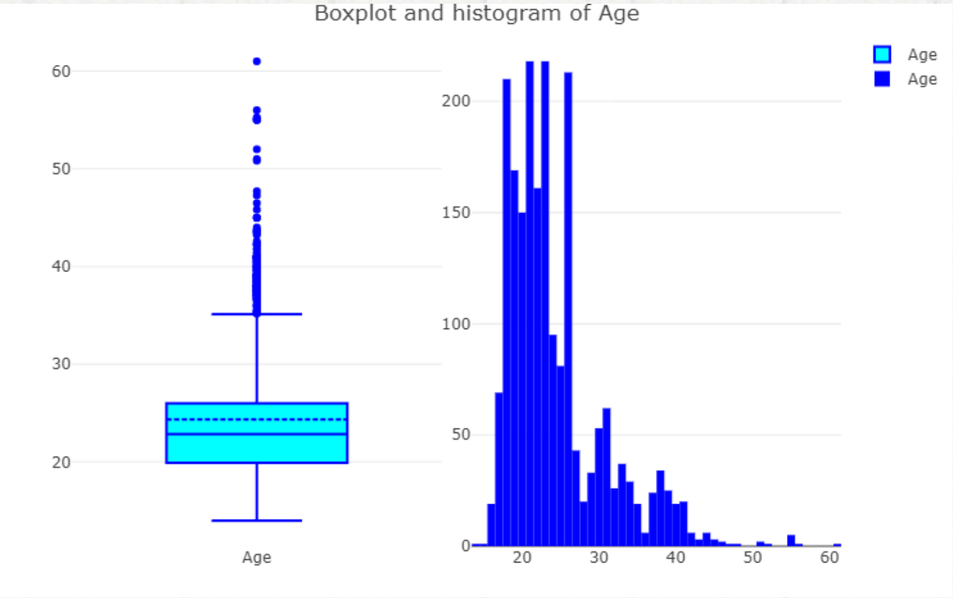
Dataset Detail

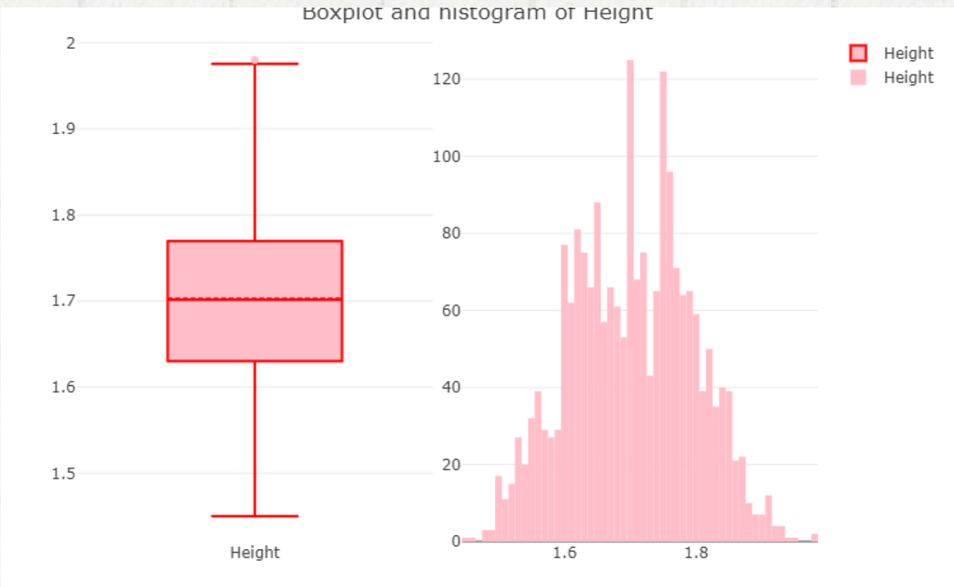
The selected dataset comes from Kaggle and was **created by Marília Prata.** It is titled "Obesity Levels & Lifestyle" and includes information from Mexico, Peru, and Colombia, which can be **used to estimate obesity rates** (Palechor & Manotas, 2019). This dataset consists of 17 columns that contain data about individuals' lifestyles, as well as their age, height, weight, and gender. In this project we will classify a person into 2 class which is obesity and no_obesity.

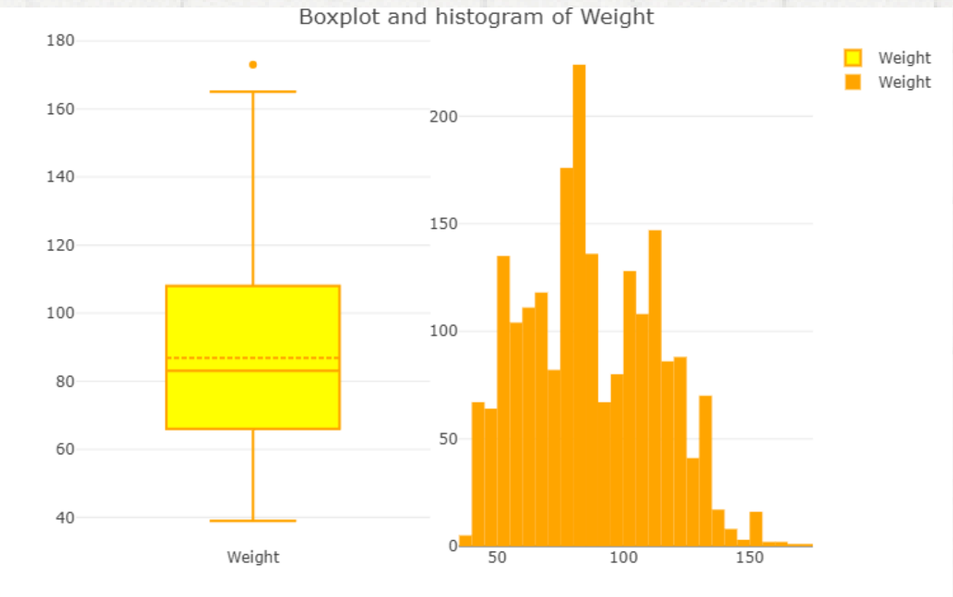
EDA

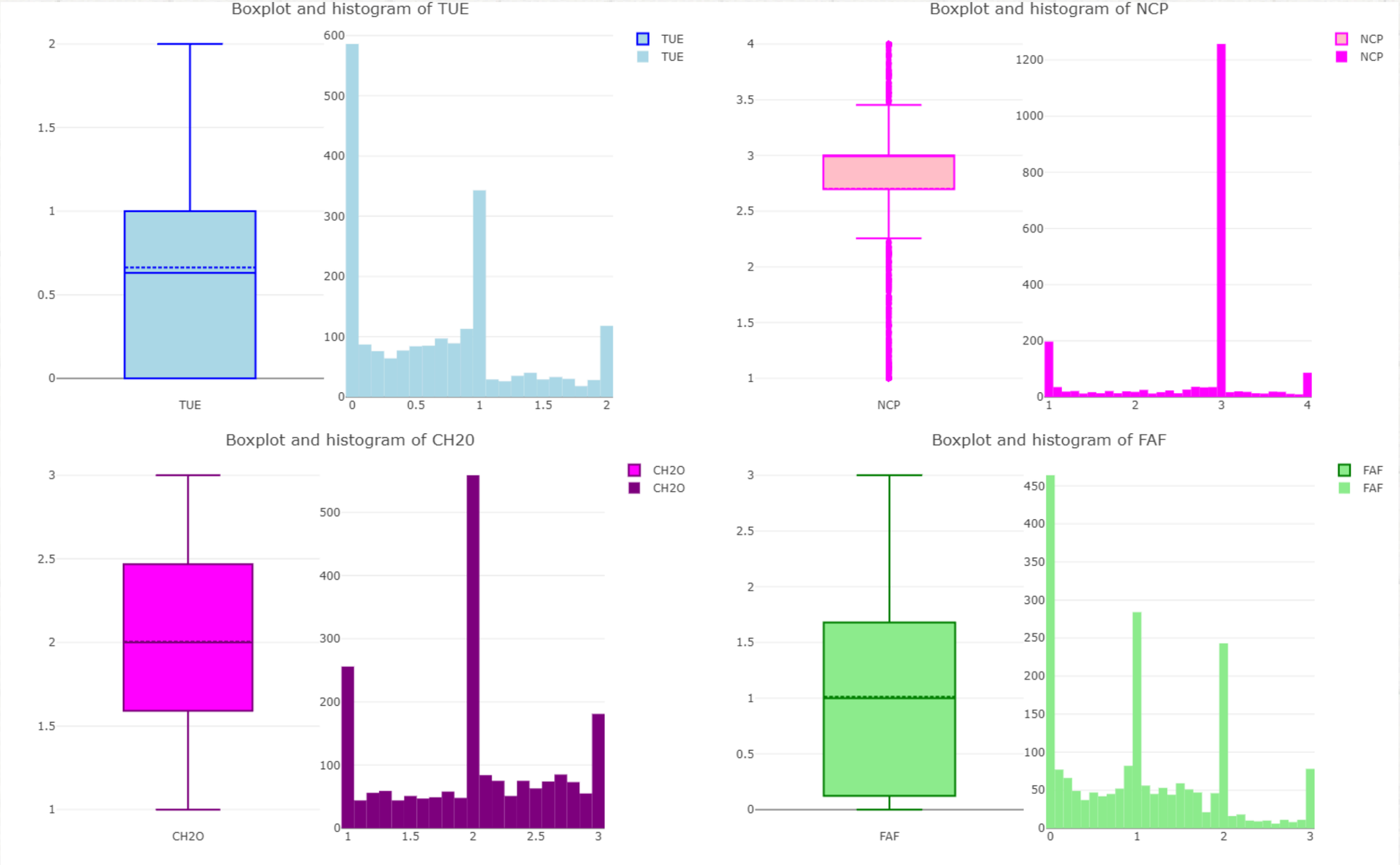
Column Name/Features	Description	Data Type
Gender	The gender of the individual	Categorical(Male, Female)
Age	The age of the individual	Numeric
Height	The height of the individual in meters	Numeric
Weight	The weight of the individual in Kilograms	Numeric
family_history_with_overweight	Historical record of the individual family member's with obesity	Categorical(yes, no)
FAVC (Frequent consumption of high caloric food)	Does the individual consumes high caloric food frequently	Categorical(yes, no)
FCVC (Frequency of consumption of vegetables)	Dies the individual frequently eat vegetables	Categorical(yes, no)
NCP (Number of main meals)	The number of main meals the individual consumes	Numeric
CAEC (Consumption of food between meals)	The frequency of the individual consuming snacks in between main meals	Categorical (No, Sometimes, Frequently, Always)
SMOKE	Smoking status of the individual	Categorical (yes, no)
CH2O (Daily water intake)	The number of water consumed by the individual in Liters	Numeric
SCC (Monitor calories consumption)	Does the individual monitor their calorie consumption	Categorical (yes, no)
FAF (Physical activity frequency)	The number of physical activities the individual did per week	Numeric
TUE (Time using technology devices)	The amount of time the individual spent on technological device in hours	Numeric
CALC (Alcohol consumption)	How frequent the individual drink alcohol	Categorical (no, sometimes, frequently, always)
MTRANS (Transportation used)	The type of transportation used by the individual for daily commute	Categorical
NObeyesdad (Obesity level)	Obesity level category of the individual	Categorical

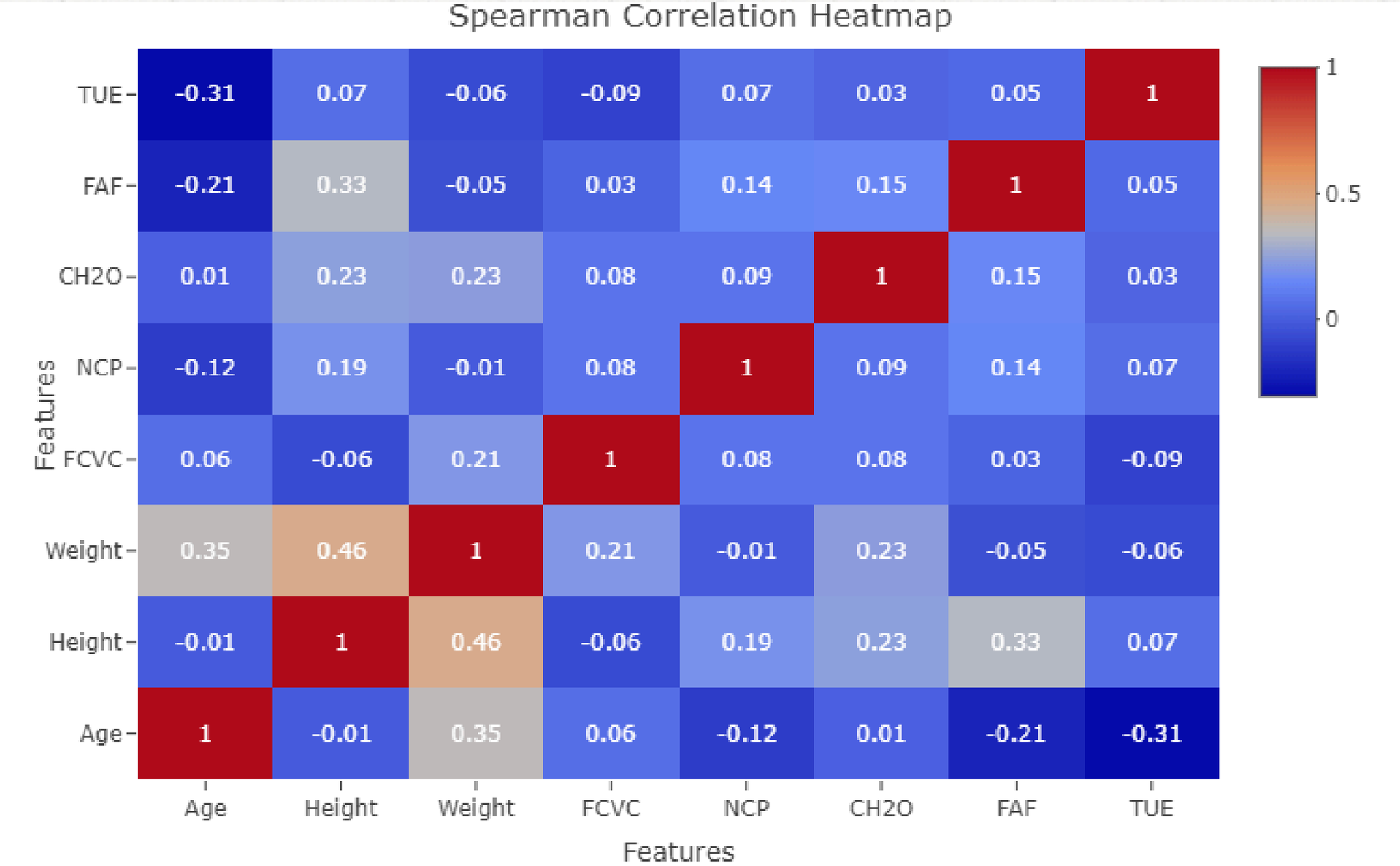

Boxplot and histogram of FCVC
FCVC
FCVC
FCVC

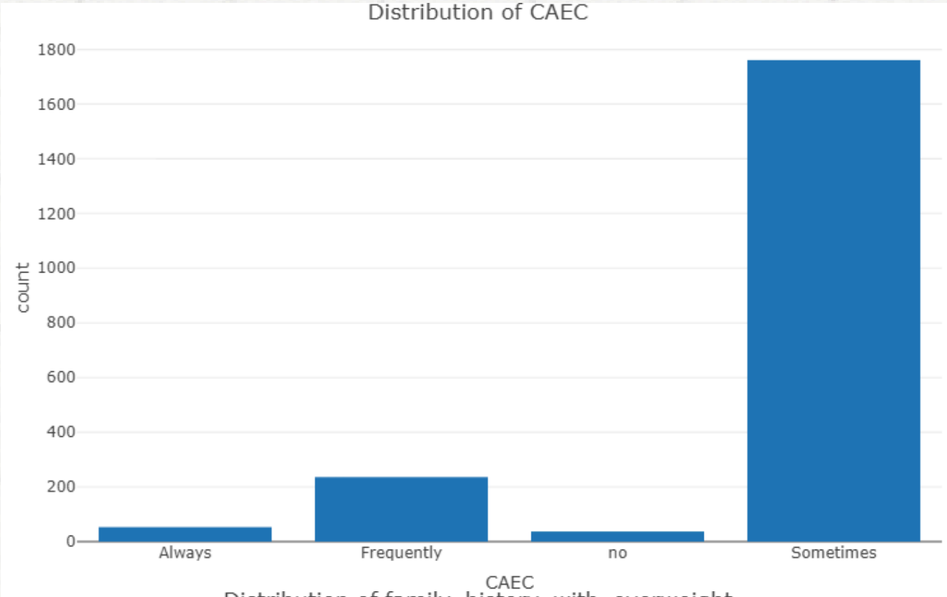

Boxplot and histogram of Age
Age
Age
Age

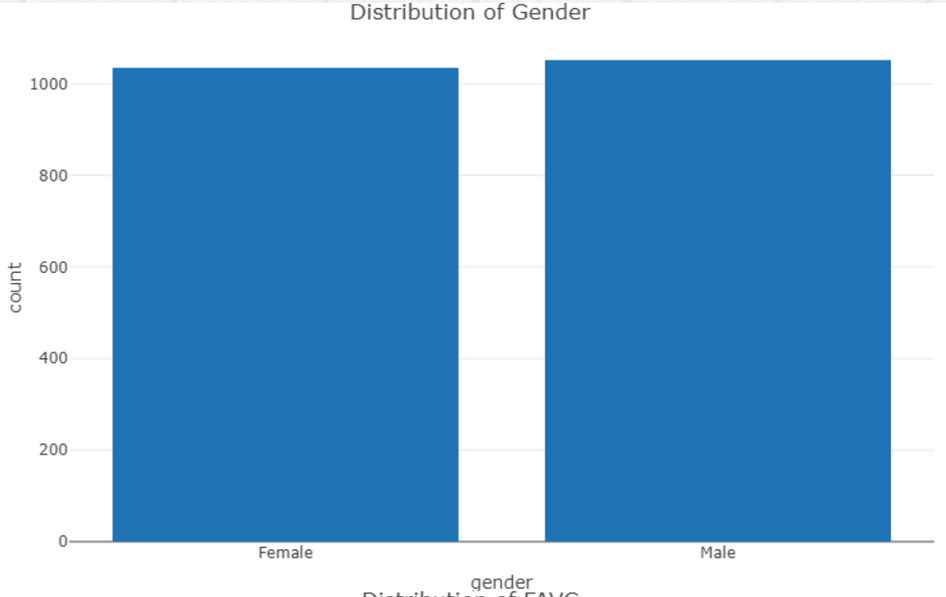

Boxplot and histogram of Height
Height
Height
Height

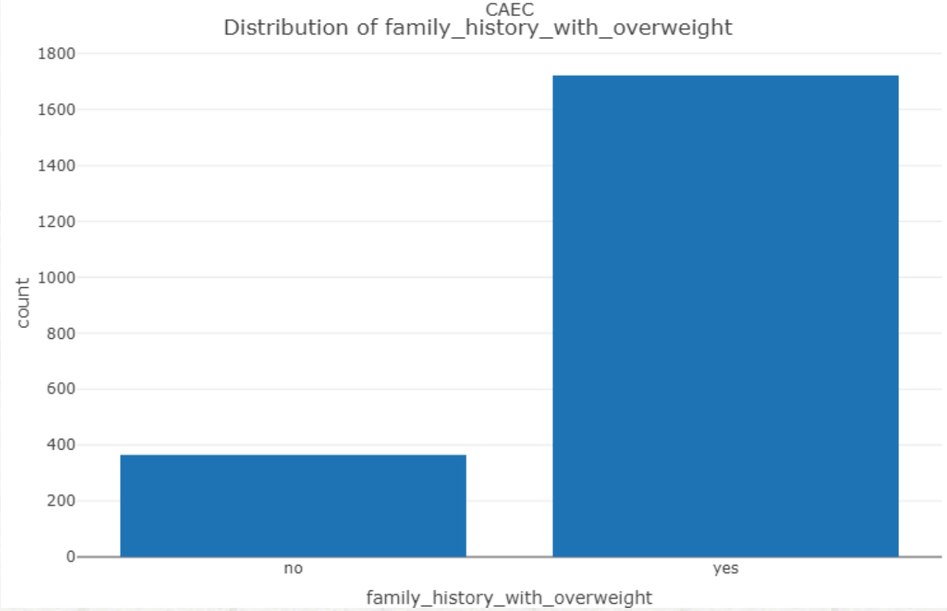

Boxplot and histogram of Weight
Weight
Weight
Weight

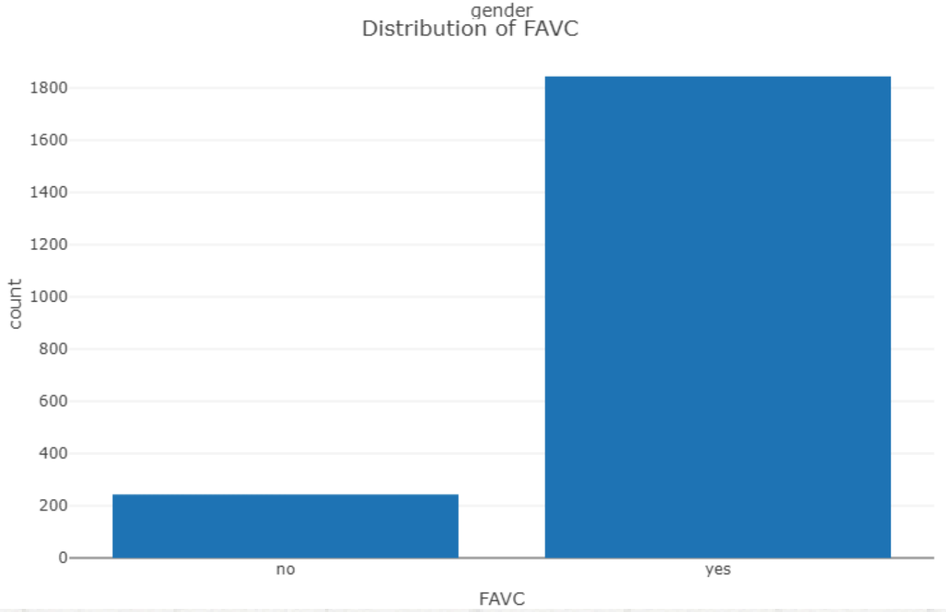

Boxplot and histogram of TUE
TUE
TUE
TUE
Boxplot and histogram of NCP
NCP
NCP
NCP
Boxplot and histogram of CH20
CH2O
CH2O
CH2O
Boxplot and histogram of FAF
FAF
FAF
FAF

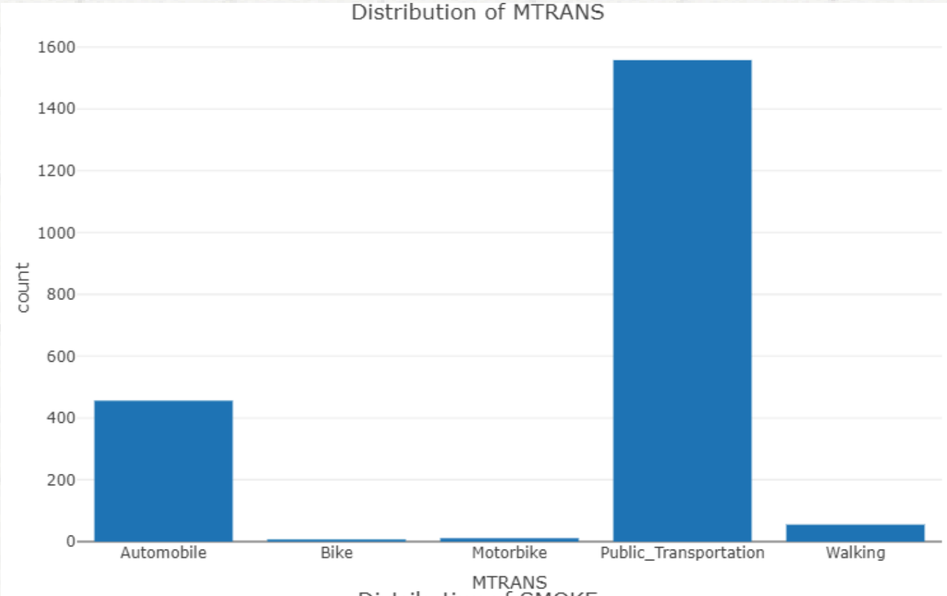

Spearman Correlation Heatmap
TUE -0.31 0.07 -0.06 -0.09 0.07 0.03 0.05 1
FAF -0.21 0.33 -0.05 0.03 0.14 0.15 1 0.05
CH2O 0.01 0.23 0.23 0.08 0.09 1 0.15 0.03
NCP -0.12 0.19 -0.01 0.08 1 0.09 0.14 0.07
FCVC 0.06 -0.06 0.21 1 0.08 0.08 0.03 -0.09
Weight 0.35 0.46 1 0.21 -0.01 0.23 -0.05 -0.06
Height -0.01 1 0.46 -0.06 0.19 0.23 0.33 0.07
Age 1 -0.01 0.35 0.06 -0.12 0.01 -0.21 -0.31
Features
Age Height Weight FCVC NCP CH2O FAF TUE
Features
1
0.5
0

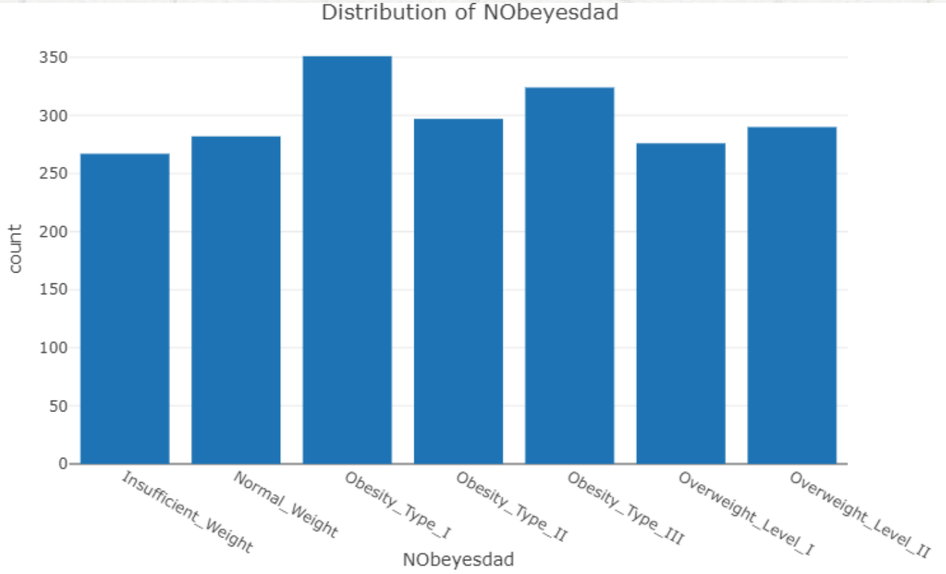

Distribution of CAEC
count
1800
1600
1400
1200
1000
800
600
400
200
0
Always
Frequently
no
Sometimes
CAEC

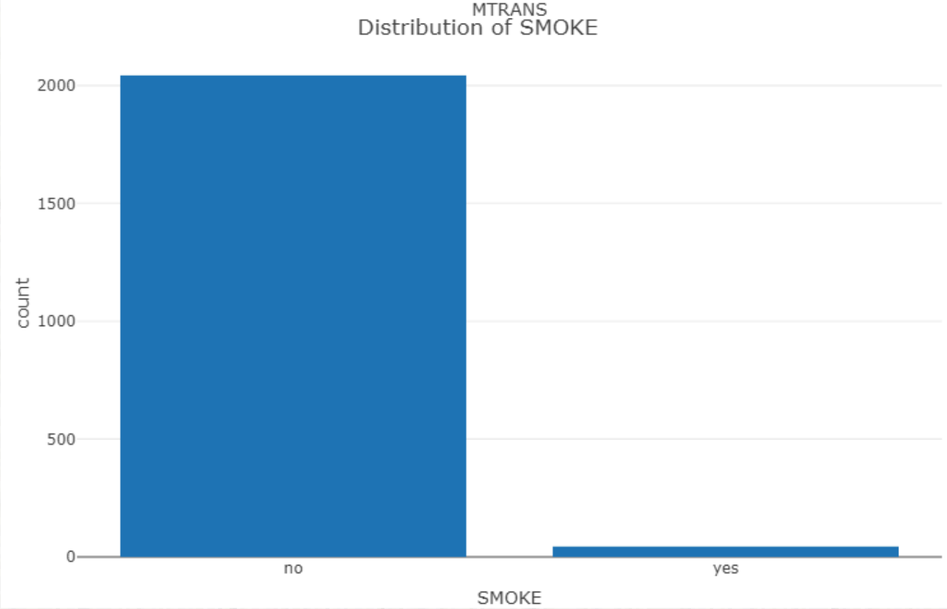

Distribution of Gender
count
1000
800
600
400
200
0
Female
Male
gender

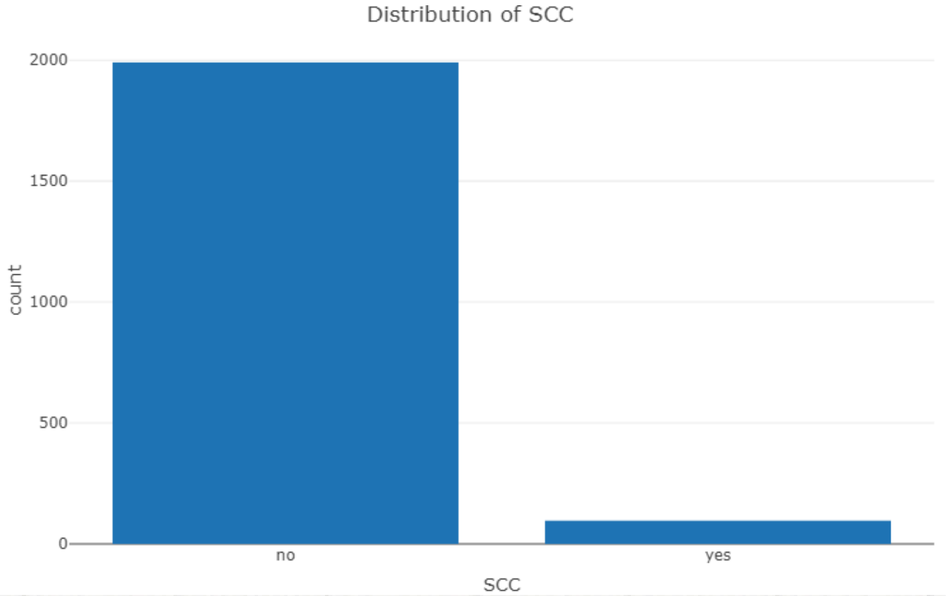

Distribution of family_history_with_overweight
count
1800
1600
1400
1200
1000
800
600
400
200
0
no
yes
family_history_with_overweight


Distribution of FAVC
count
1800
1600
1400
1200
1000
800
600
400
200
0
no
yes
FAVC


Distribution of MTRANS
count
1600
1400
1200
1000
800
600
400
200
0
Automobile
Bike
Motorbike
Public_Transportation
Walking
MTRANS


Distribution of NObeyesdad
count
350
300
250
200
150
100
50
0
Insufficient_Weight
Normal_Weight
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Overweight_Level_I
Overweight_Level_II
NObeyesdad


Distribution of SMOKE
count
2000
1500
1000
500
0
no
yes
SMOKE


Distribution of SCC
count
2000
1500
1000
500
0
no
yes
SCC

Data Preprocessing

```{r}
data$Gender <- as.numeric(as.factor(data$Gender))
data$MTRANS <- as.numeric(as.factor(data$MTRANS))
data$family_history_with_overweight <- as.numeric(as.factor(data$family_history_with_overweight))
data$FAVC <- as.numeric(as.factor(data$FAVC))
data$FCVC <- as.numeric(as.factor(data$FCVC))
data$NCP <- as.numeric(as.factor(data$NCP))
data$CAEC <- as.numeric(as.factor(data$CAEC))
data$SMOKE <- as.numeric(as.factor(data$SMOKE))
data$CH2O <- as.numeric(as.factor(data$CH2O))
data$SCC <- as.numeric(as.factor(data$SCC))
data$FAF <- as.numeric(as.factor(data$FAF))
data$TUE <- as.numeric(as.factor(data$TUE))
data$CALC <- as.numeric(as.factor(data$CALC))
```

Assuming 'NObeyesdad' is the binary target variable (1 = Yes, 0 = No)

```{r}
Y <- as.numeric(data$NObeyesdad == "Yes")  # Convert 'NObeyesdad' to a binary numeric variable
```

Define a smaller set of predictors (e.g., Age, Height, Weight, and Family History)

```{r}
X <- cbind(data$Age, data$Height, data$Weight, data$family_history_with_overweight)
```

Scale the features (standardize to mean 0, sd 1)

```{r}
X <- scale(X)
```

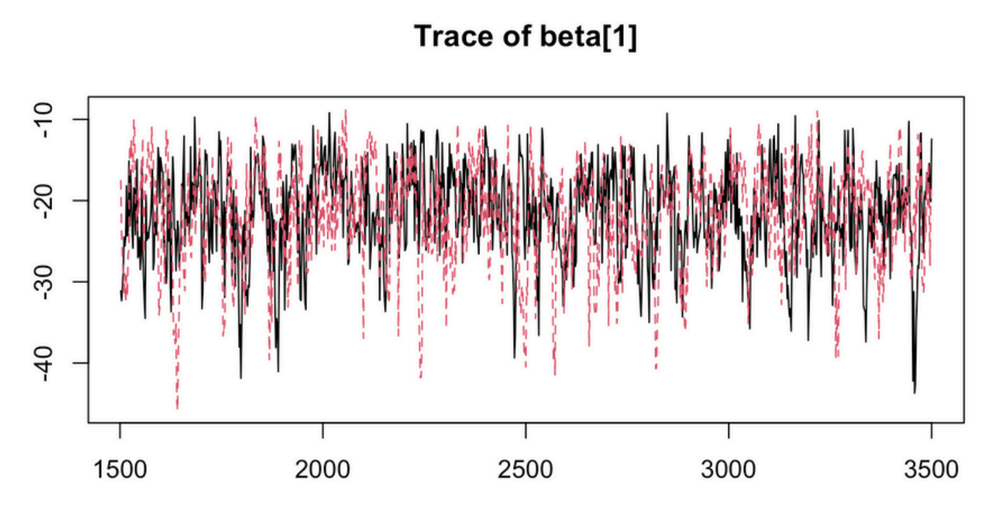
FIT TO MODEL

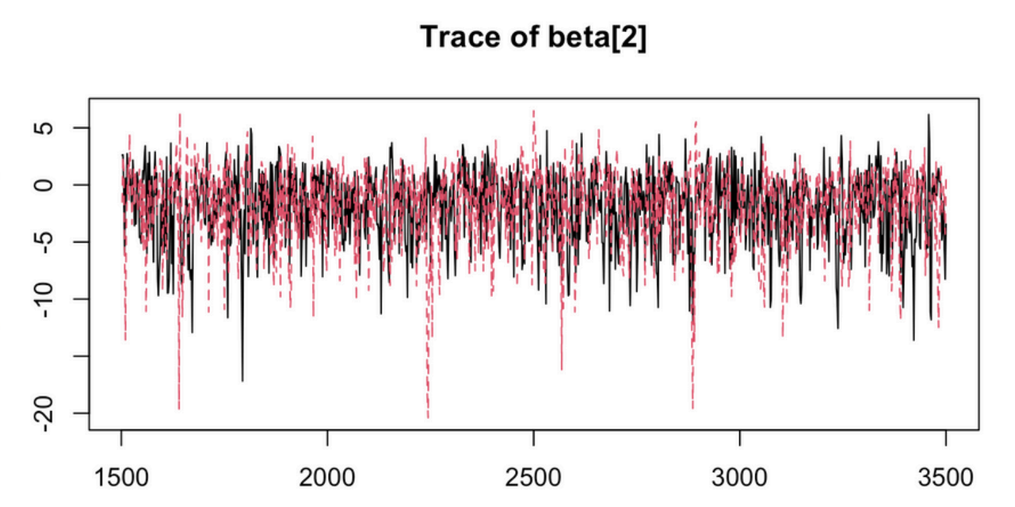
UNINFORMATIVE

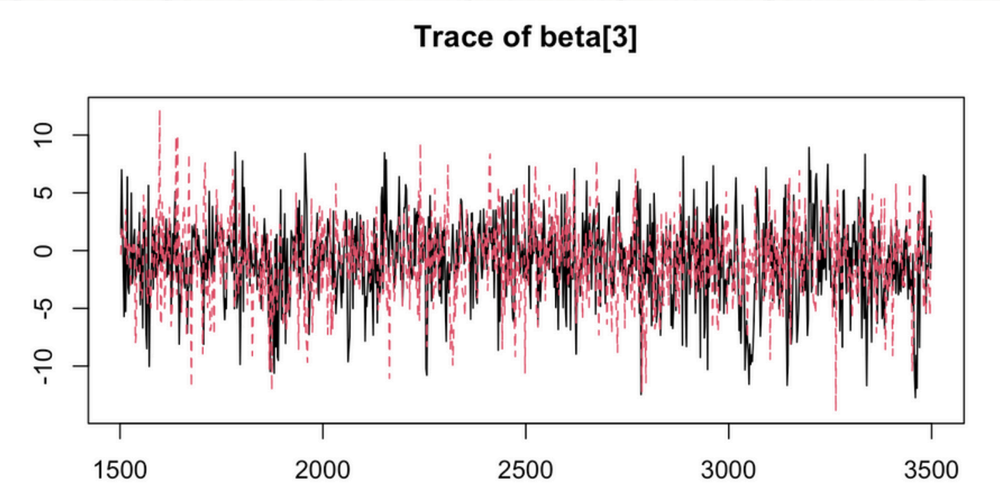
Variable	Mean	SD	Naive SE	Time-series SE
β_1	-21.6475	6.023	0.13467	0.29114
β_2	-1.8980	3.452	0.07719	0.09735
β_3	-0.9318	3.509	0.07847	0.10023
β_4	0.6269	3.880	0.08676	0.10978
β_5	2.9172	5.214	0.11659	0.13183

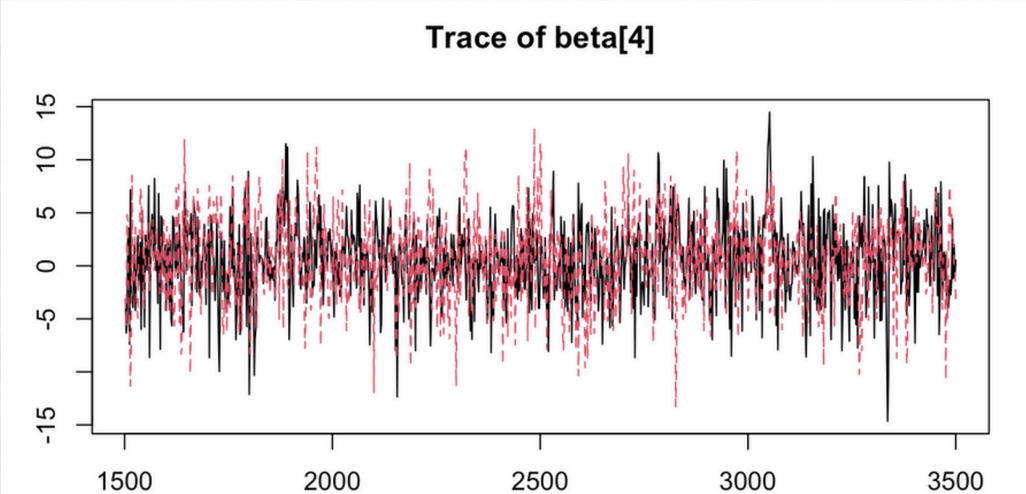
Variable	2.5%	25%	50%	75%	97.5%
β_1	-35.202	-25.3358	-21.0203	-17.2303	-11.892
β_2	-9.695	-3.9881	-1.2733	0.6595	3.273
β_3	-8.511	-3.0983	-0.8471	1.3513	5.703
β_4	-7.392	-1.8630	0.8205	3.0372	7.767
β_5	-5.682	-0.6519	2.0248	5.9281	14.697

UNINFORMATIVE

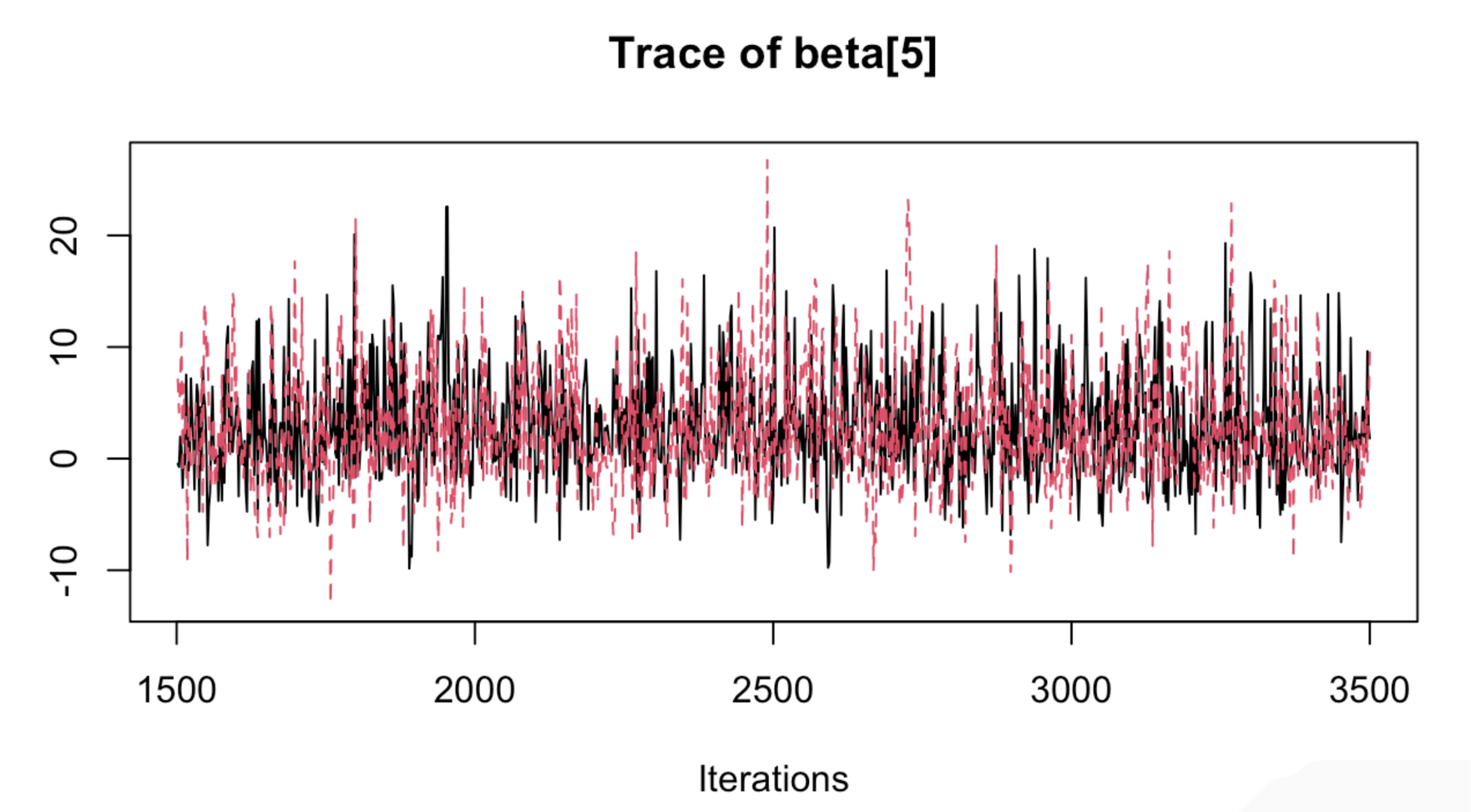




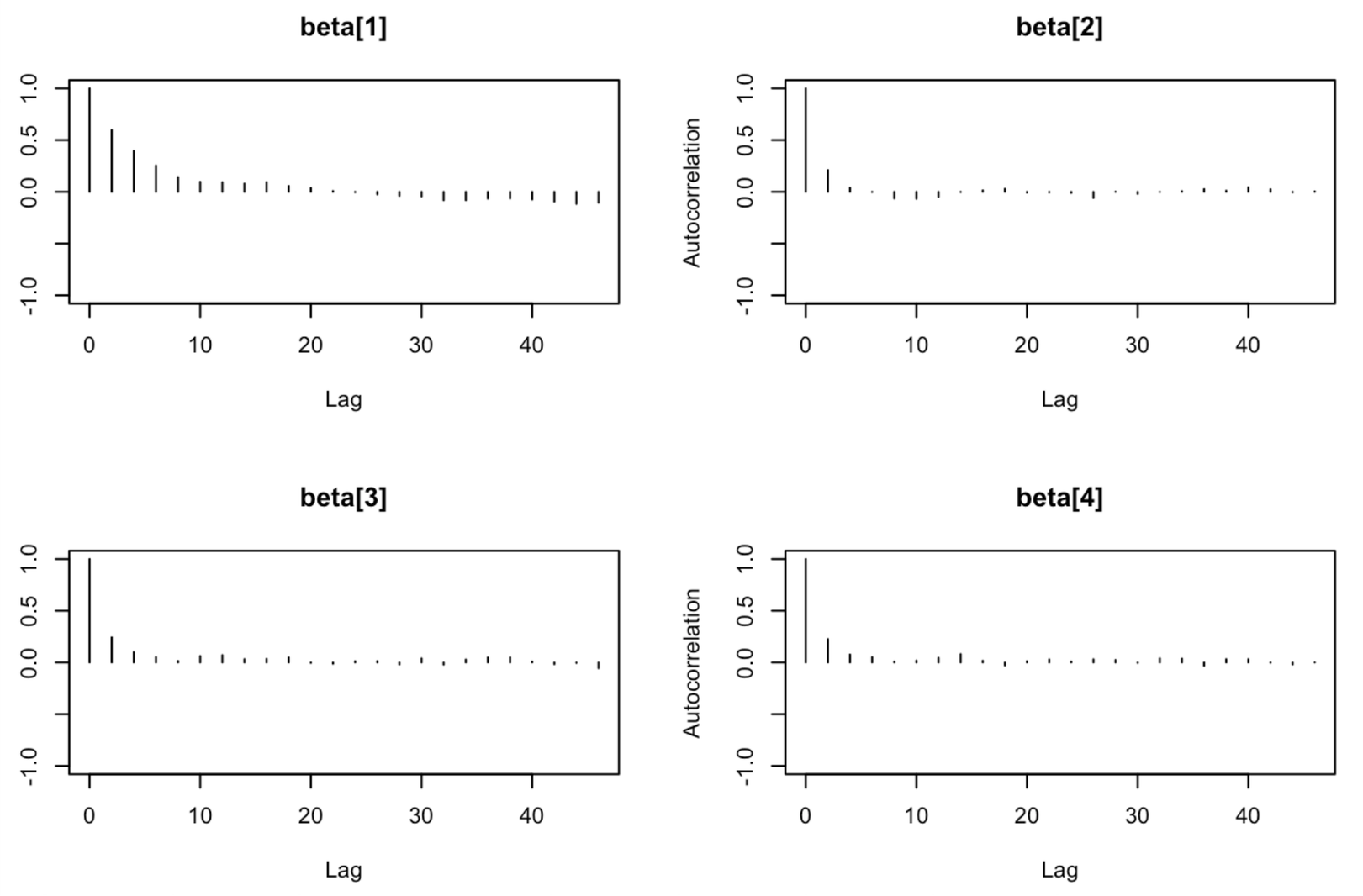


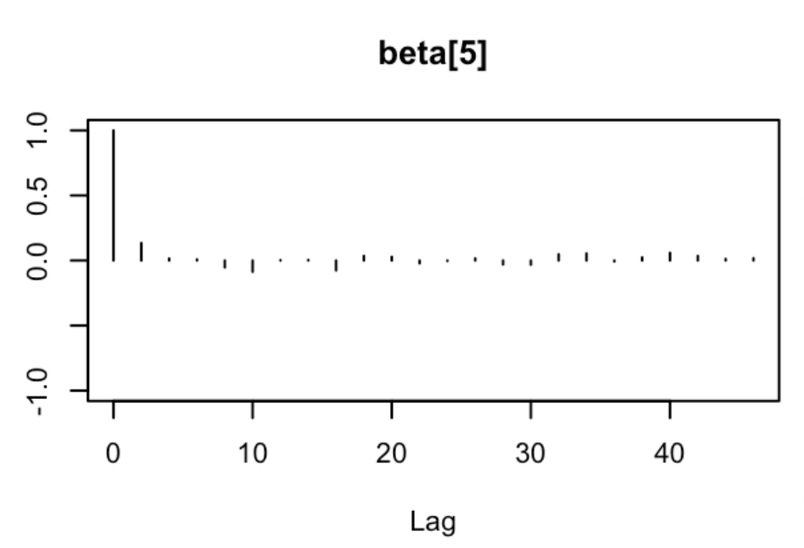


UNINFORMATIVE



UNINFORMATIVE





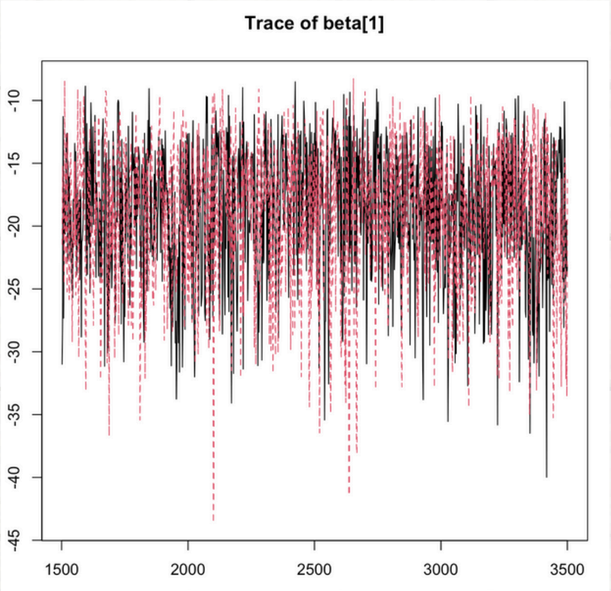
```
> gelman.diag(samps1)
Potential scale reduction factors:

        Point est. Upper C.I.
beta[1]          1       1.01
beta[2]          1       1.00
beta[3]          1       1.00
beta[4]          1       1.00
beta[5]          1       1.00

Multivariate psrf

1
>
> effectiveSize(samps1)
  beta[1]   beta[2]   beta[3]   beta[4]   beta[5]
 429.8745 1261.9960 1227.0033 1250.0577 1641.6912
```

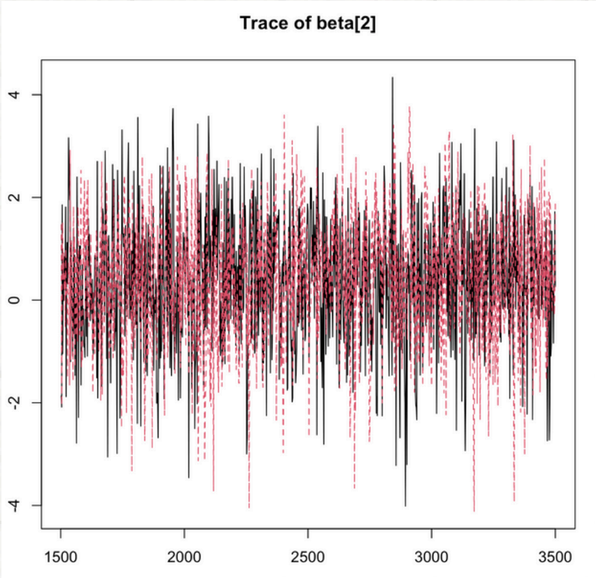
INFORMATIVE

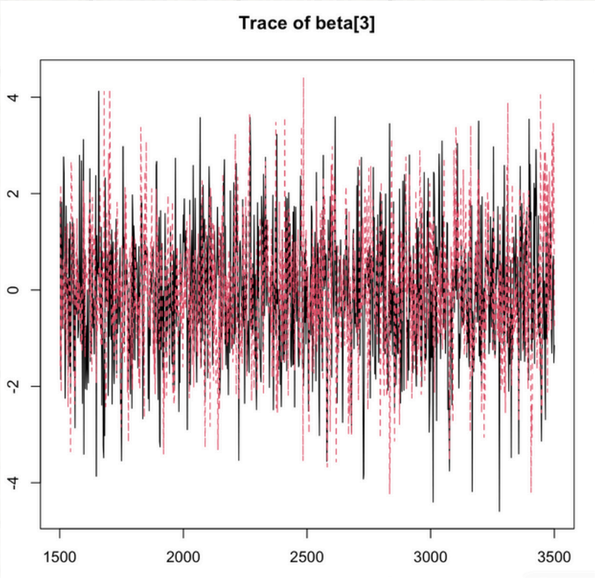
Parameter	2.5%	25%	50%	75%	97.5%
β_1	-31.4999	-22.43204	-18.21699	-14.6715	-10.546
β_2	-2.2413	-0.46562	0.44494	1.2501	2.679
β_3	-2.7496	-0.98394	-0.09101	0.7949	2.748
β_4	-0.0319	1.23784	1.90846	2.5830	3.843
β_5	-1.6678	-0.02485	0.90704	1.8217	3.684

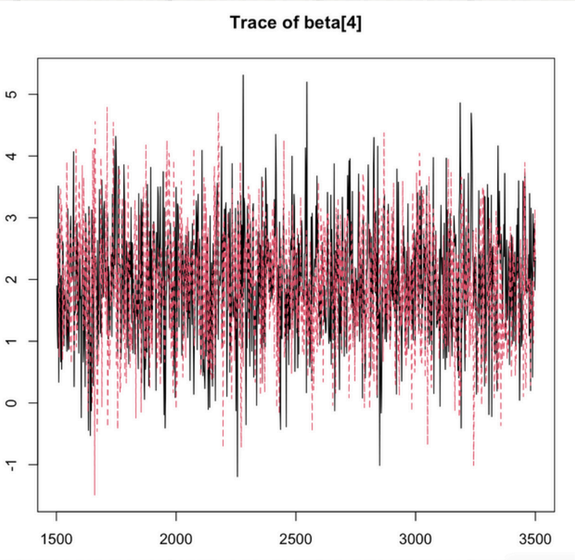
Parameter	Mean	SD	Naive SE	Time-series SE
β_1	-19.04125	5.5967	0.12514	0.15254
β_2	0.37895	1.2735	0.02848	0.03122
β_3	-0.06573	1.3719	0.03068	0.03436
β_4	1.91528	0.9752	0.02181	0.02439
β_5	0.93295	1.3668	0.03056	0.03230

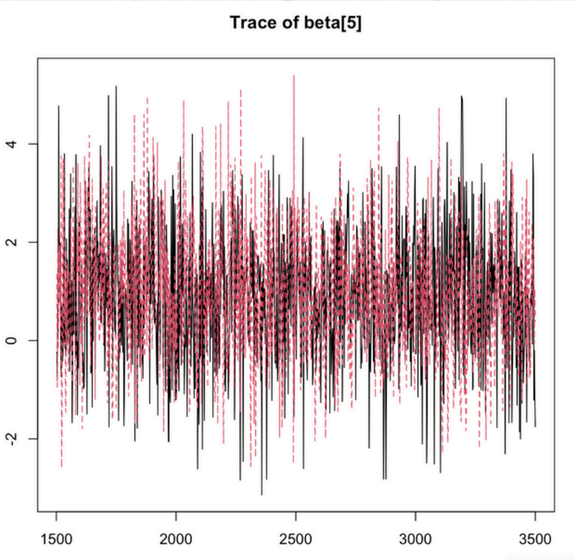
INFORMATIVE







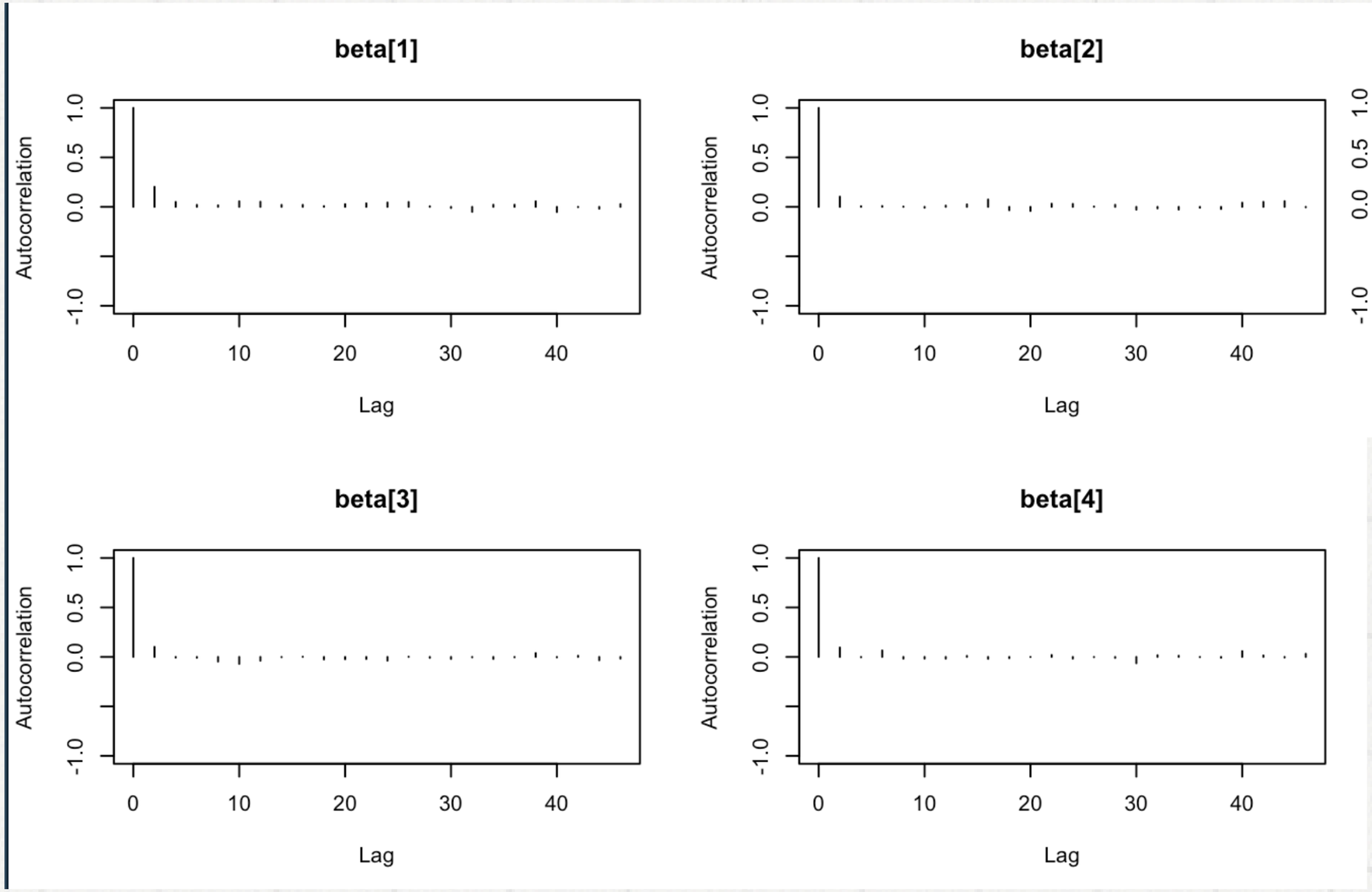




INFORMATIVE

```
> gelman.diag(samps2)
Potential scale reduction factors:

        Point est. Upper C.I.
beta[1]       1.00       1.01
beta[2]       1.00       1.00
beta[3]       1.00       1.00
beta[4]       1.01       1.03
beta[5]       1.00       1.01

Multivariate psrf

1.01
> effectiveSize(samps2)
 beta[1]  beta[2]  beta[3]  beta[4]  beta[5]
1344.966 1665.453 1594.848 1603.987 1790.030
```

INFORMATIVE

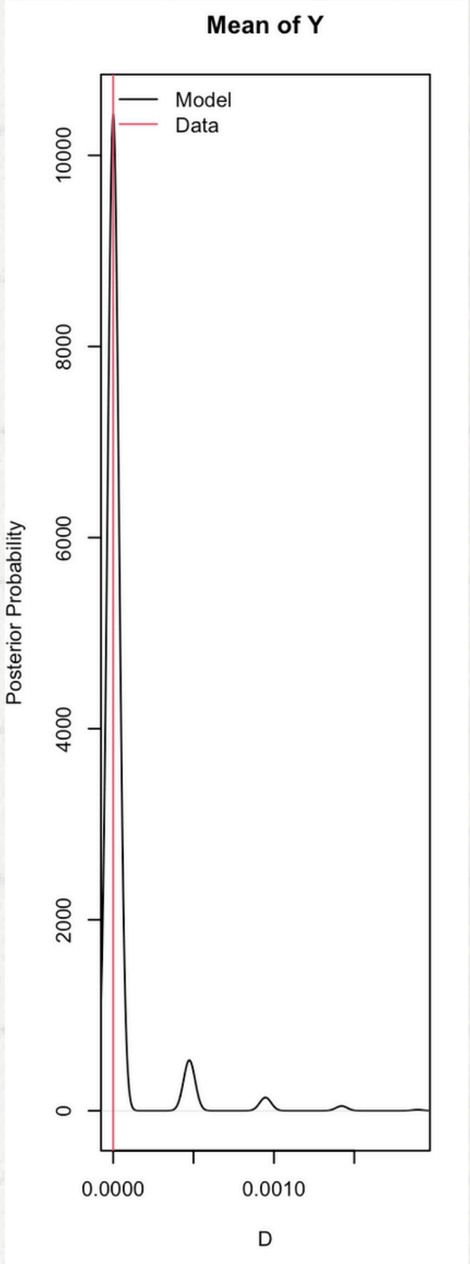


PERBANDINGAN DIC DAN WAIC

Metric	Informative Model	Uninformative Model
WAIC	0.2	0.2
DIC	1.181	4.27

THE BEST MODEL

Age	Positive	Significant	Moderate to Strong
Height	Neutral	Likely Not Significant	Weak (Posterior Mean ~ 0)
Weight	Positive	Significant	Strong (~2)
Family History	Positive	Significant	Moderate (~1)



EVALUATE THE MODEL

The model performs well in predicting the mean of Y, as indicated by the close alignment between the model's predictions and the observed data. This suggests that the model adequately captures the underlying data structure and provides reliable predictions.

Thank You