PREDICTIVE MODELING:

Click Behavior on Advertisement

GROUP 3:

Sherly Vaneza
2702222163

Miecel Alicia Angel J
2702327601

Natasha Kayla Cahyadi
2702235891

contents.

introduction.

Advertisement effectiveness is a key area of interest in marketing and business analytics. Advertisements play a crucial role in driving **consumer behavior, increasing brand awareness, and boosting sales.** Accurate prediction of advertisement clicks is essential for optimizing marketing strategies and enhancing the return on investment. By understanding the factors that influence user engagement, businesses can tailor their campaigns more effectively, ensuring they reach the right audience with the right message.

dataset overview.

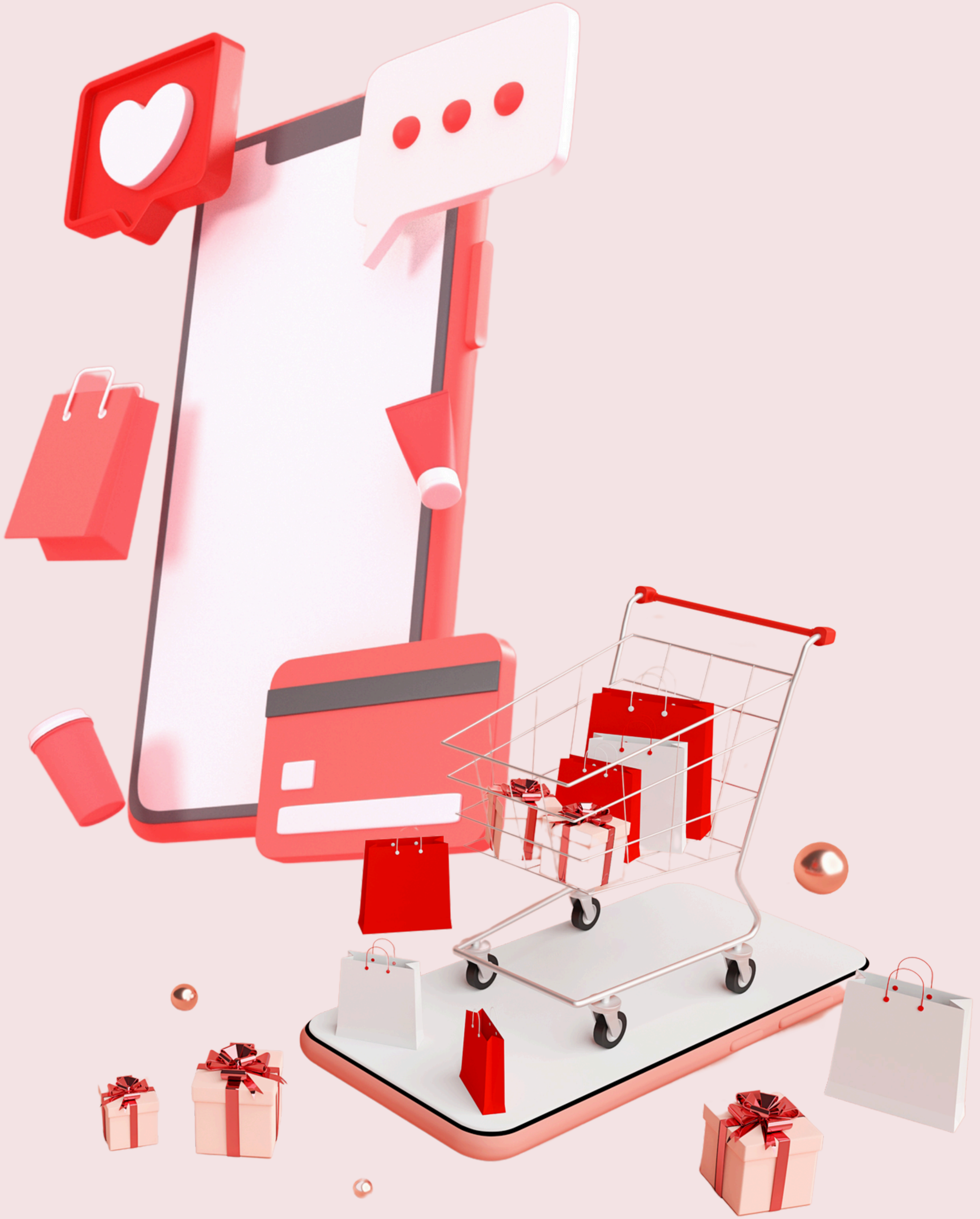
Dataset Used: Dataset Advertisement - Clicked on Ad

This dataset contains **1,000 rows and 10 columns**, providing data about user behavior on an advertising platform. **Below is the summary of the structure and content:**

1. **Daily Time Spent on Site (float):** The average time (in minutes) a user spends on the site daily.
2. **Age (int):** User's age in years.
3. **Area Income (float):** Average income of the user's geographical area.
4. **Daily Internet Usage (float):** Average daily internet usage (in minutes).
5. **Ad Topic Line (object):** The headline of the ad the user sees.
6. **City (object):** The city of the user.
7. **Male (int):** Gender of the user (1 for male, 0 for female).
8. **Country (object):** The country of the user.
9. **Timestamp (object):** The time when the user clicked or did not click on the ad.
10. **Clicked on Ad (int):** Target variable indicating whether the user clicked on the ad (1 for yes, 0 for no).

project goal.

The goal of this project is to develop a probabilistic model using Bayesian methods to predict **whether a user will click on an advertisement.** By analyzing the relationship between user attributes (such as age, daily time spent on site, area income, and daily internet usage) and the likelihood of clicking on ads, the project aims to provide insights into user behavior and improve targeted advertising strategies. The ultimate objective is to quantify uncertainties and interpret the influence of various features on ad-clicking behavior through a Bayesian approach.



data preparation.

Checking Duplicates: *No duplicates value in the dataset.*

```
{r}
sum(duplicated(df))

 [1] 0
```

Checking Missing Values: *No missing values in the dataset.*

```
{r}
colSums(is.na(df))

Daily.Time.Spent.on.Site                      Age              Area.Income
                       0                        0                        0
      Daily.Internet.Usage            Ad.Topic.Line                     City
                       0                        0                        0
                    Male                  Country                Timestamp
                       0                        0                        0
             Clicked.on.Ad
                       0
```

data preparation.

Renaming Column ‘Male’ to ‘Gender’: *To make the column more representative and inclusive.*

```
{r}
# Rename column Male to Gender [0 = Female, 1 = Male]
df <- df %>% rename(Gender = Male)
```

Deleting High Cardinality Categorical Columns: *To make the model simpler, reduce overfitting, and improve computational efficiency.*

```
{r}
df$Ad.Topic.Line <- NULL
df$City <- NULL
df$Timestamp <- NULL
df$Country <- NULL
```

models.

1 LOGISTIC REGRESSION - INFORMATIVE PRIORS

A widely used statistical technique for **modeling binary outcomes.**

The informative prior is derived from **the observed data in the dataset,** reflecting the estimated probabilities and uncertainties for the parameters.

2 PROBIT REGRESSION - UNINFORMATIVE PRIORS

Assumes that the underlying **distribution of the data follows a normal distribution,** with a less informative prior that doesn't rely on historical data.

algorithm - Logistic Regression

LIKELIHOOD :

Bernoulli Distribution - Y[i] ~ dbern(p[i])

- Models binary outcomes (0 or 1).
- Suitable for probabilities modeled with a logit function.
- Assumes independence between observations.

PRIORS :

Normal Distribution - beta[j] ~ dnorm(mu[j], tau[j])

- Computationally efficient and flexible for parameter updates.
- Incorporates prior knowledge from historical data to improve model accuracy.

algorithm - Probit Regression

LIKELIHOOD :

Bernoulli Distribution - Y[i] ~ dbern(p[i])

- Models binary outcomes (0 or 1).
- Suitable for probabilities modeled with a logit function.
- Assumes independence between observations.

PRIORS :

Normal Distribution - beta[j] ~ dnorm(0, 0.01)

- Computationally efficient and flexible for parameter updates.
- Incorporates prior knowledge from historical data to improve model accuracy.

Convergence Diagnostic - Logistic Regression

ESS:

	beta[1]	beta[2]	beta[3]	beta[4]	beta[5]	beta[6]
Convergence	Convergence	Convergence	Convergence	Convergence	Convergence	Convergence

AUTOCORRELATION:

Parameter	Absolute Mean Autocorrelation	Convergence Category
beta[1] (Intercept)	0.3681	Moderate
beta[2] (Age)	0.2592	Moderate
beta[3] (Area Income)	0.2328	Moderate
beta[4] (Daily Internet Usage)	0.2502	Moderate
beta[5] (Gender)	0.1887	High
beta[6] (Daily Time on Site)	0.0498	High

Convergence Diagnostic - Logistic Regression

GELMAN DIAGNOSTICS:

	beta[1]	beta[2]	beta[3]	beta[4]	beta[5]	beta[6]
Convergence	Convergence	Convergence	Convergence	Convergence	Convergence	Convergence

GEWEKE DIAGNOTICS:

	beta[1]	beta[2]	beta[3]	beta[4]	beta[5]	beta[6]
Convergence	Convergence	Convergence	Convergence	Convergence	Convergence	Convergence

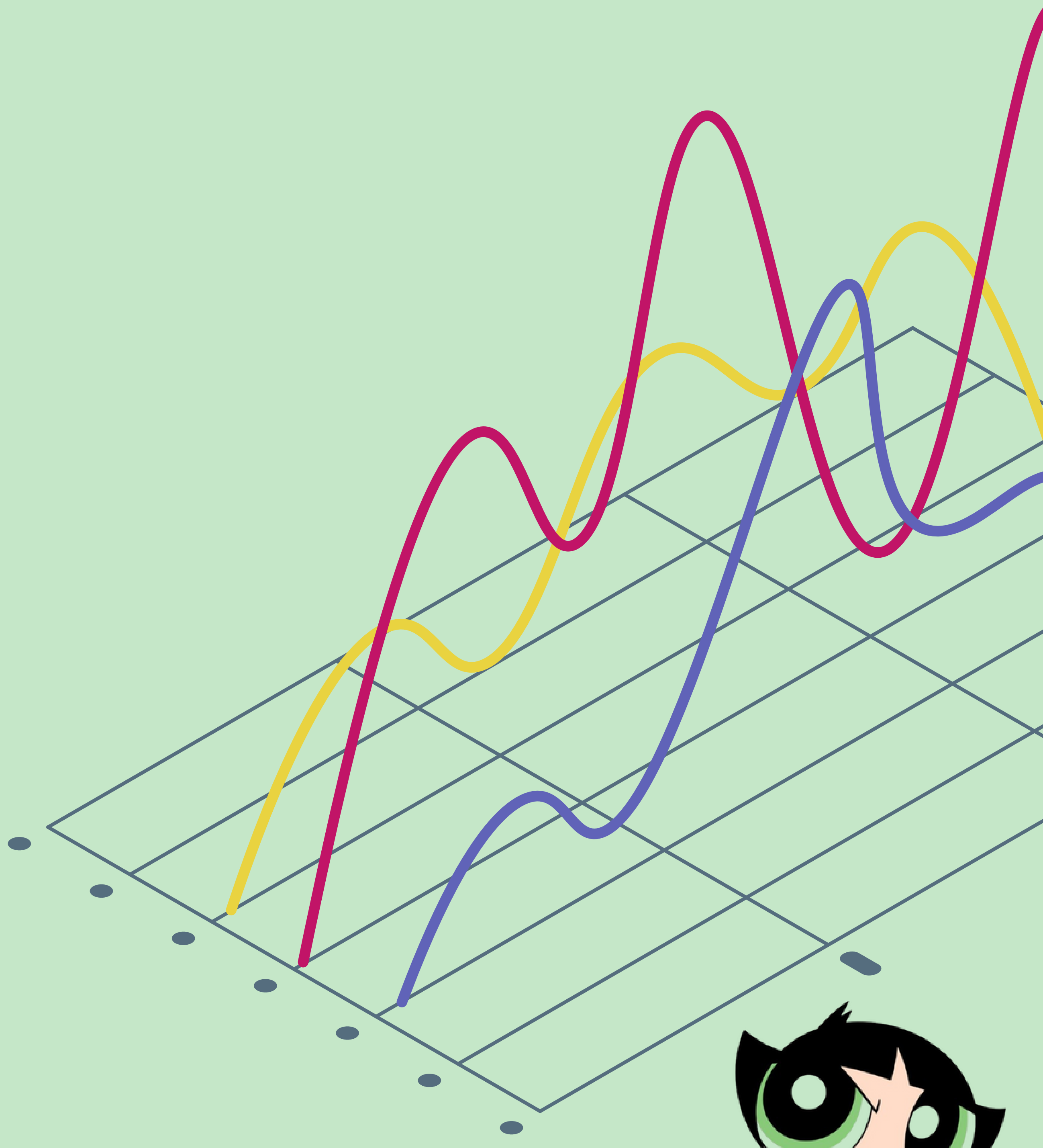
Convergence Diagnostic - Probit Regression

ESS:

	beta[1]	beta[2]	beta[3]	beta[4]	beta[5]	beta[6]
Convergence	Convergence	Convergence	Convergence	Convergence	Convergence	Convergence

AUTOCORRELATION:

Parameter	Absolute Mean Autocorrelation	Convergence Category
beta[1] (Intercept)	0.417	Poor
beta[2] (Age)	0.360	Moderate
beta[3] (Area Income)	0.315	Moderate
beta[4] (Daily Internet Usage)	0.375	Moderate
beta[5] (Gender)	0.317	Moderate
beta[6] (Daily Time on Site)	0.055	High



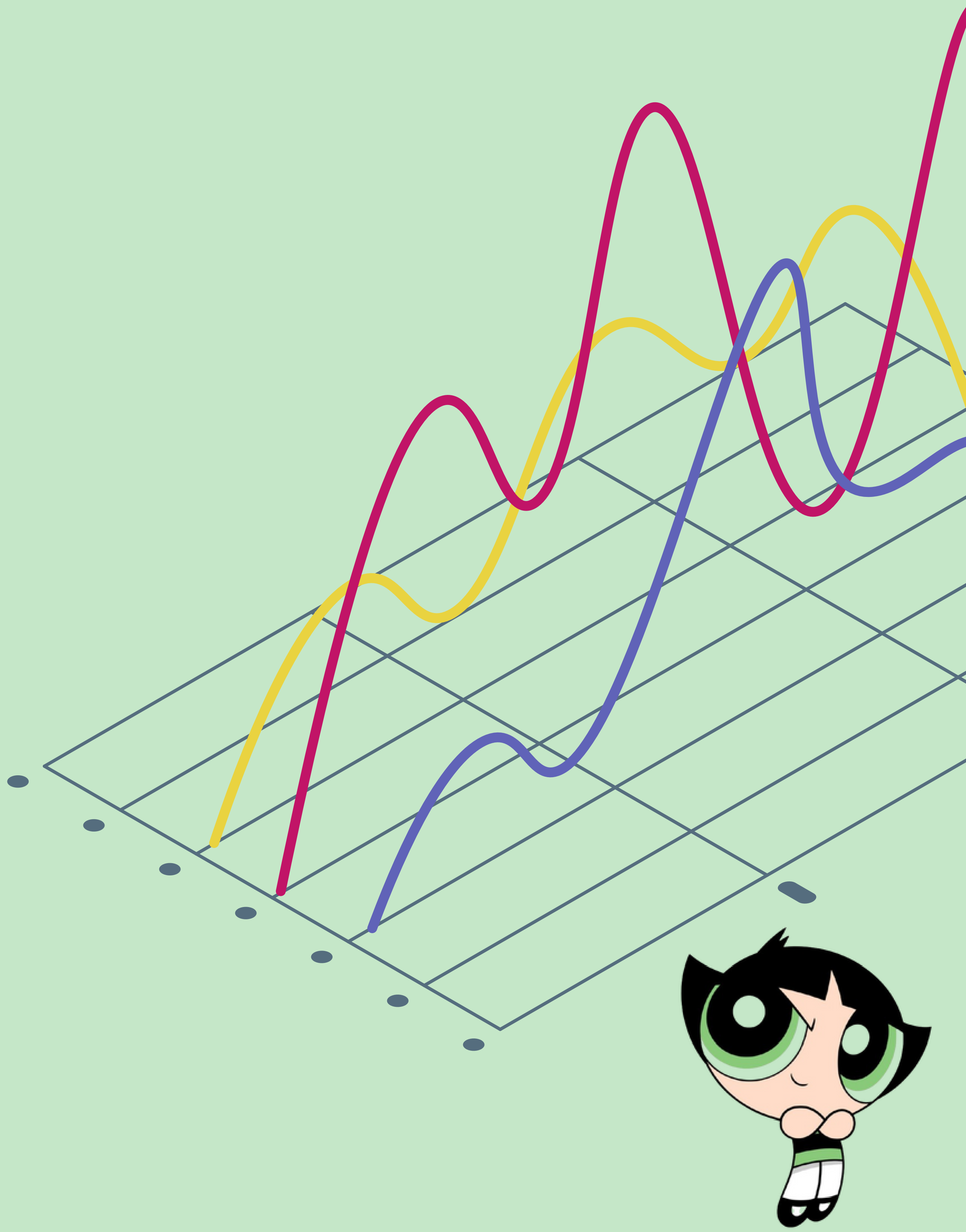
Convergence Diagnostic - Probit Regression

GELMAN DIAGNOSTICS:

	beta[1]	beta[2]	beta[3]	beta[4]	beta[5]	beta[6]
Convergence	Convergence	Convergence	Convergence	Convergence	Convergence	Convergence

GEWEKE DIAGNOTICS:

	beta[1]	beta[2]	beta[3]	beta[4]	beta[5]	beta[6]
Convergence	Convergence	Convergence	Convergence	Convergence	Convergence	Convergence



RESULTS.

DIC MODEL 1:

```
Mean deviance:  192.3
penalty 5.65
Penalized deviance: 198
```

DIC MODEL 2:

```
Mean deviance:  191.5
penalty 6.065
Penalized deviance: 197.6
```

When comparing the two models, Model 2 demonstrates a better overall performance with a lower DIC value (197.6) compared to Model 1 (197.9). While Model 2 is slightly more complex, its better fit to the data justifies the additional complexity. **Therefore, for the advertising dataset, Model 2 is the preferred choice as it offers a more optimal balance between data fit and model complexity.**

RESULTS.

WAIC MODEL 1:

	Estimate	SE
elpd_waic	-4.3	4.2
p_waic	0.3	0.0
waic	8.6	8.4

WAIC MODEL 2:

	Estimate	SE
elpd_waic	-2.4	2.8
p_waic	0.1	0.0
waic	4.8	5.6

The second model has a lower WAIC (4.8) compared to the first model (8.6), indicating better predictive performance. Although both models have relatively high standard errors (8.4 for the first and 5.6 for the second), **the notable difference in WAIC suggests that the second model is more likely to generalize well to new data.**

RESULTS.

Both models closely match the observed mean, **indicating that they fit the data well.** However, Model 2 is slightly closer to the observed mean, suggesting a marginally better alignment with the observed data. Although the difference is minimal, **Model 2 demonstrates a more accurate fit and is therefore the preferred choice for the advertisement dataset.**

POSTERIOR PREDICTIVE CHECKS [MODEL 1]:

Min Y	Max Y	Mean Y
0.0000	1.0000	0.4982

POSTERIOR PREDICTIVE CHECKS [MODEL 2]:

Min Y	Max Y	Mean Y
0.000	1.000	0.505

Conclusion.

Referring to the combined evaluation using **DIC and WAIC,** Model 2 emerges as the preferred choice for the advertising dataset as it achieves a better balance between data fit, complexity, and predictive performance. Additionally, **both models closely match the observed mean**, indicating a good fit with the data. However, Model 2 aligns slightly closer to the observed mean, suggesting a marginally better representation of the observed data. While the difference is minor, **Model 2 demonstrates superior accuracy and is thus the optimal choice for the advertising dataset.**

THANK YOU! 😄