# Analysis of Consumer Buying Patterns on Black Friday

Ruicheng Zhang

**Abstract**

Using a data set of 233,599 transactions, this paper provides a comprehensive analysis of consumer buying patterns during Black Friday. Using several models, including multilevel models, this paper aims to dissect the differences in product consumption tendencies among customer segments in different cities, while considering the impact of Black Friday on customer behavior as well as the impact of product categories. It is hoped that this study provides some insights into how companies can customize their marketing strategies and products for different consumer segments during one of the most important shopping events of the year.

***Keyword:*** *Multilevel model*, *Black Friday*, *Regression*, *Sales Analysis*, *Consumer Buying Patterns*

## 1 Introduction

Black Friday, a pivotal event in the retail calendar, marks a significant surge in consumer spending, offering a unique opportunity to analyze and understand consumer behavior. This project leverages a comprehensive dataset from Kaggle, encompassing 233,599 transactions, to delve into the complexities of consumer purchasing patterns during this peak shopping period.

The crux of this analysis lies in its focus on the relationship between consumer demographics and their propensity to purchase specific products. By examining variables such as age, gender, marital status, type of city, and duration of stay, the project aims to uncover how these factors influence buying decisions across different product categories.

To achieve a nuanced understanding of these dynamics, the project employs advanced analytical techniques. Multilevel modeling is central to this approach, allowing for an in-depth examination of the interaction between individual purchases and broader consumer profiles. Additionally, customer segmentation through clustering and the use of association rule mining to identify product purchasing patterns offer a comprehensive view of consumer behavior.

## 2 Method

### 2.1 Data processing

Table 1: Overview of Variables in Dataset

| Variable | Type | Description |
|---|---|---|
| User ID | Integer | Unique identifier of the user |
| Product ID | Integer | Unique identifier of the product |
| Gender | Integer (Categorical) | Gender of the user (0 for female, 1 for male) |
| Age | Integer (Categorical) | Age group of the user (coded as integers) |
| Occupation | Integer (Categorical) | Occupation code of the user |
| City Category | Integer (Categorical) | Category of the city(1 for A, 2 for B, 3 for C) |
| Stay In Current City Years | Integer (Categorical) | Number of years the user has lived in the current city |
| Marital Status | Integer (Categorical) | Marital status of the user (0 for single, 1 for married) |
| Product Category 1 | Integer | Product category code 1 |
| Product Category 2 | Integer | Product category code 2 (may contain NA for missing values) |
| Product Category 3 | Integer | Product category code 3 (may contain NA for missing values) |
| Purchase | Integer | Purchase amount in dollars |

The dataset contains 550068 rows and 12 columns, the above table is the initial processed data, the data type is mainly integer (int64), there are two columns (Product Category2 and Product Category 3) is a floating point number (float64), Product Category 2 and Product Category 3 columns have missing data. Category 3 columns have missing data.

I have processed each variable accordingly, User ID column: since each data is greater than 100000, it is subtracted from 100000, which doesn't affect the data analysis; Gender column takes 1 for male, 0 for female; Age column contains "0-17" "55+" "26-35" "46-50" "51-55" "36-45" "18- The Age column contains "0-17," "55+," "26-35," "46-50," "51-55," "36-45," and "18- 25," which are coded from 0 to 6 from smallest to largest.Purchase: Purchase amount ranges from 12 to 23,961, the average value is about 9,264, the value is too large, use log-transform to deal with it.

In this paper, Purchase will be selected as the target variable, and next, in order to observe the relationship between each variable and the target variable, a correlation analysis graph is made using r.

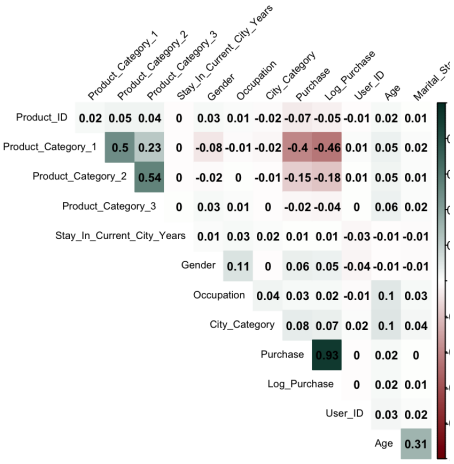



Figure 1: Correlation Plot

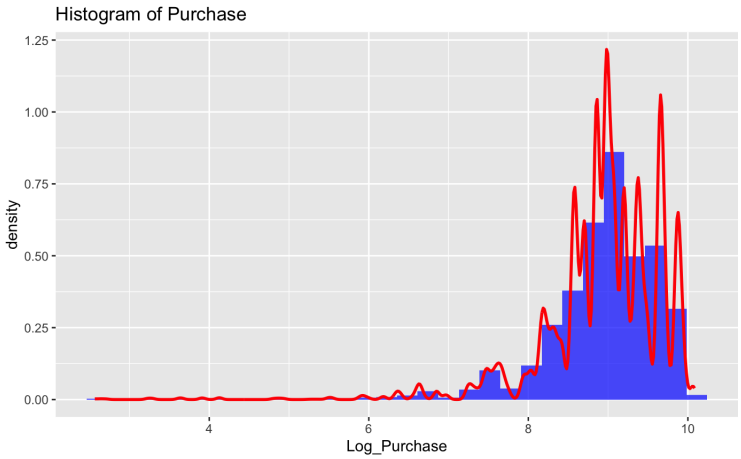
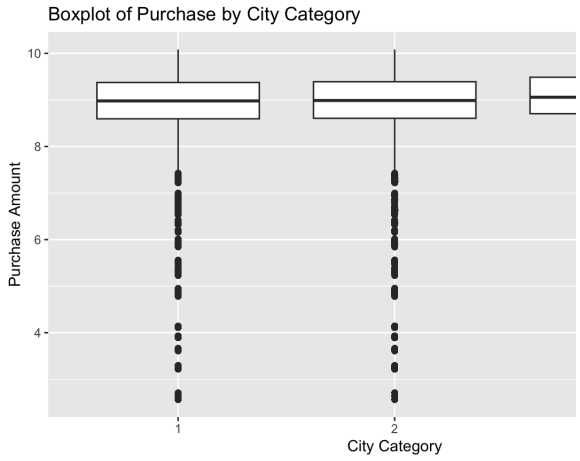



Figure 2: Histogram of Purchase

Based on this correlation chart, delete the variables that have a weak relationship with the target variable: Occupation,Marital Status,Product Category 3,Stay In Current City Years,Age.

This graph shows the histogram and density plot of the logarithmic transformation of the Purchase column. Graphically, the data still does not appear to be perfectly symmetrical and exhibits multiple peaks (multimodal), and the transformed data still does not conform to the standard shape of a normal distribution.

## 2.2 Data analysis



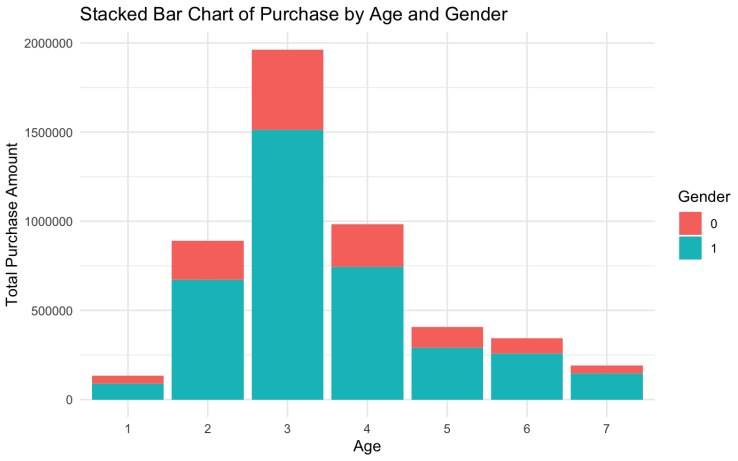




Figure 3: Boxplot of Purchase by City Category

Figure 4: Stacked Bar Chart of Purchase by Age and Gender

The boxplot depicts the distribution of purchase amounts across three city categories, with the median purchase amount represented by the horizontal line within each box. All three categories exhibit a relatively similar median purchase value, suggesting that city category may not be a strong predictor of purchase amount. Notably, category 2 displays a slightly higher interquartile range and more outliers, indicating greater variability in purchase amounts within this city category compared to categories 1 and 3.

The graph on the right illustrates that 36-45 year olds have the highest average spending power, while men have a higher level of purchasing power compared to women.

## 2.3 Modeling

1. Null Model

$$\hat{y} = \beta_0 \tag{1}$$

Build the Null Model,$\hat{y}$: This is the predicted value of the dependent variable Purchase. In your model, it represents the average expected value for Purchase.$\beta_0$: This is the intercept term that represents the average expected value of the dependent variable without considering any of the independent variables. In your Null Model, this is the only parameter that represents the average Purchase of all observations.

The Null Model, featuring only an intercept, indicates an average expected value of Purchase at around 9264, as evidenced by its statistically significant intercept with a large t-value. However, the model's high residual standard error and wide range of residuals suggest limited predictive power and the potential need for a more complex model to better explain the data.

2. No Pooling Model

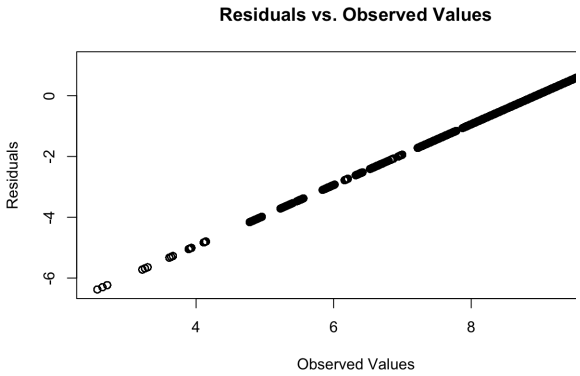



Figure 5: Residual Plot for Null Model

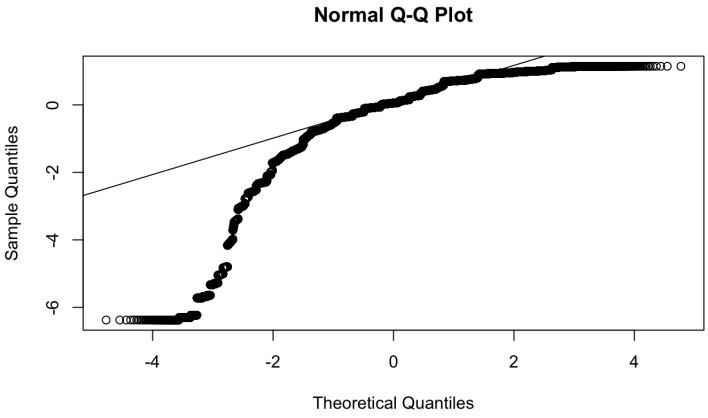



Figure 6: QQ plot

$$\hat{y} = \beta_0 + \beta_1 \times \text{Gender} + \beta_2 \times \text{Product\_Category\_1} + \beta_3 \times \text{Product\_Category\_2} + \beta_4 \times \text{City\_Category} \tag{2}$$

where:
$\hat{y}$: This is the predicted value of the dependent variable (Purchase).
$\beta_0$: Intercept term, representing the average expected value of Purchase when all independent variables are zero.
$\beta_1$: Coefficient for the Gender variable, representing the average impact of Gender on Purchase.
$\beta_2$: Coefficient for the Product_Category_1 variable, representing the average impact of Product_Category_1 on Purchase.
$\beta_3$: Coefficient for the Product_Category_2 variable, representing the average impact of Product_Category_2 on Purchase.
$\beta_4$: Coefficient for the City_Category variable, representing the average impact of different City_Category values on Purchase.

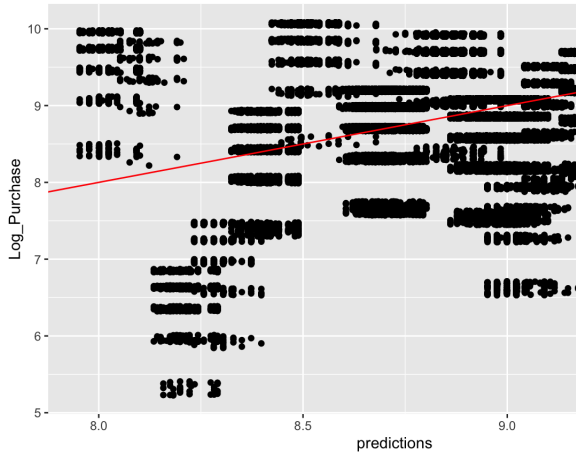



Figure 7: Predictions vs. Actual values

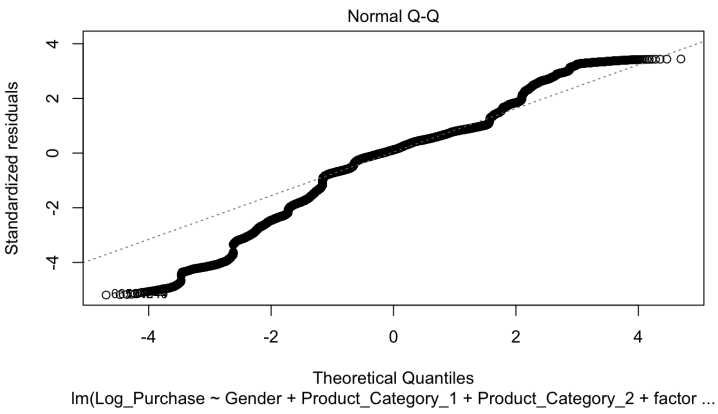



Figure 8: QQ plot

The No Pooling Model demonstrates strong statistical significance across all variables with a high Multiple R-squared value of 0.9958, indicating a good fit to the data. However, the presence of a large residual standard error of 0.5883 on 376424 degrees of freedom.

3. Complete Pooling Model

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{3}$$

where:
$\hat{y}$: The predicted value of the dependent variable (Purchase).
$\beta_0$: Intercept term, representing the average expected value of Purchase when all independent variables are zero.
$\beta_1, \beta_2, \ldots, \beta_k$:Coefficients for the independent variables
$X_1, X_2, \ldots, X_k$,representing the average impact of each variable on Purchase.
$X_1, X_2, \ldots, X_k$: The independent variables included in the model, representing all columns in the dataset other than Purchase.

The model explains approximately 0.2208 of the variance in purchase amounts (as indicated by the R-squared value), suggesting a moderate fit. However, a large residual standard error and wide range of residuals imply that the model might not capture all the variability in the data, and there could be other important predictors or non-linear relationships not accounted for.

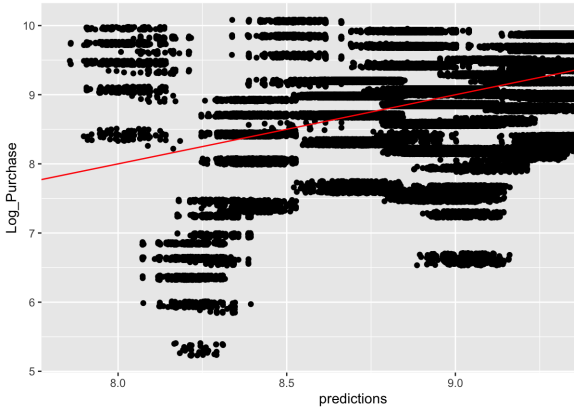



Figure 9: Predictions vs. Actual values

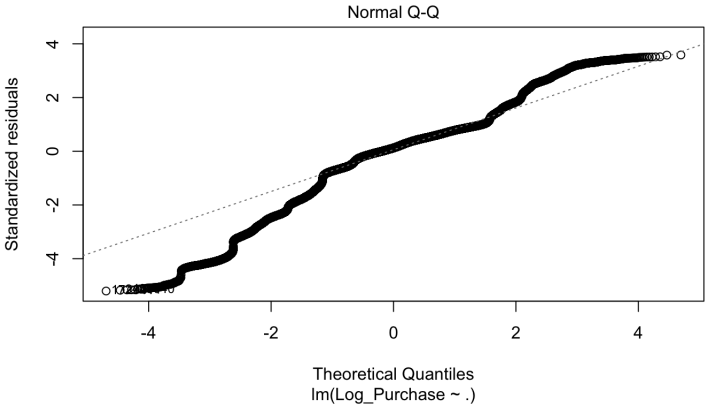



Figure 10: QQ plot

Most predictors are statistically significant, with p-values less than 0.05, indicating a meaningful association with the purchase amount. The negative coefficients for Product ID and Product Category 1 suggest a decrease in purchase amount with an increase in these categories, whereas positive coefficients for Gender, City Category, and Product Category 2 suggest an increase in purchase amounts.

4. Partial Pooling Model

$$\hat{y}_i = \beta_0 + \beta_1 \times \text{Gender}_i + \beta_2 \times \text{Product\_Category\_1}_i + \beta_3 \times \text{Product\_Category\_2}_i + \mu_{j[i]} + \epsilon_i \tag{4}$$

where:
$\hat{y}_i$: The predicted value of the dependent variable (Purchase) for observation i.
$\beta_0$: Intercept term, representing the overall average expected value of Purchase.

$\beta_1, \beta_2, \beta_3$ : Coefficients for the independent variablesGender, Product_Category_1, and Product_Category_2.

$\text{Gender}_i, \text{Product\_Category\_1}_i, \text{Product\_Category\_2}_i$ : The values of the independent variables for observation $i$.

$\mu_{j[i]}$ : Random effect for City_Category $j$ associated with observation $i$.

$\epsilon_i$ : Random error term associated with observation $i$.

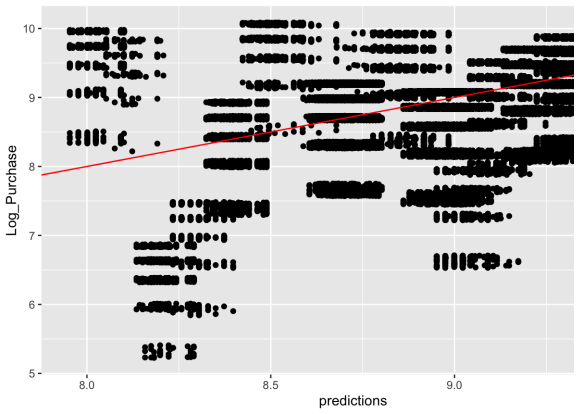



Figure 11: Predictions vs. Actual values

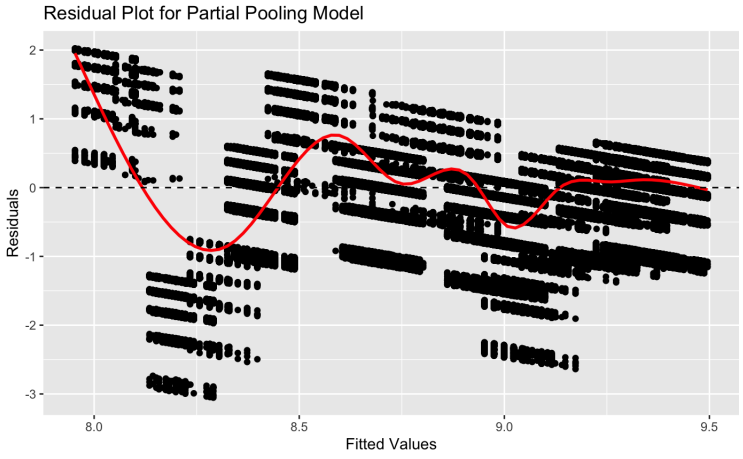



Figure 12: Residual Plot for Partial Pooling Model

Model Structure and Random Effects: This partial pooling model, or multilevel model, includes random intercepts for City Category, indicating that the model accounts for variations in baseline purchase amounts across different city categories. The significant variance in the random intercepts (Std.Dev. = 0.04758) suggests that city categories have substantial heterogeneity in their baseline purchase levels.

Fixed Effects and Model Interpretation: The fixed effects for Gender, Product Category 1, and Product Category 2 are statistically significant, as indicated by their t-values. This model indicates that, controlling for city category differences, Gender and Product Categories still have a notable impact on purchase amounts. The coefficients are consistent with the complete pooling model, indicating similar directional influences on the purchase amount.

Model Fit and Residuals: The model's residuals are (Std.Dev. = 0.58830), suggesting that while the model accounts for city-level variability, there is still considerable unexplained variation at the individual level. The scaled residuals' range indicates that the model's assumptions about normality and homoscedasticity of residuals may not be fully met, which could affect the model's predictions and inference.

5. Normal Linear Regression Model (Robust)

$$\hat{y} = \beta_0 + \beta_1 \times \text{Gender} + \beta_2 \times \text{Product\_Category\_1} + \beta_3 \times \text{Product\_Category\_2} \tag{5}$$

where:
$\hat{y}$: The predicted value of the dependent variable (Purchase).
$\beta_0$: Intercept term, representing the average expected value of Purchase when all independent variables are zero.
$\beta_1$, $\beta_2$, $\beta_3$: Coefficients for the independent variables Gender, Product_Category_1, and Product_Category_2, representing the average impact of each variable on Purchase.

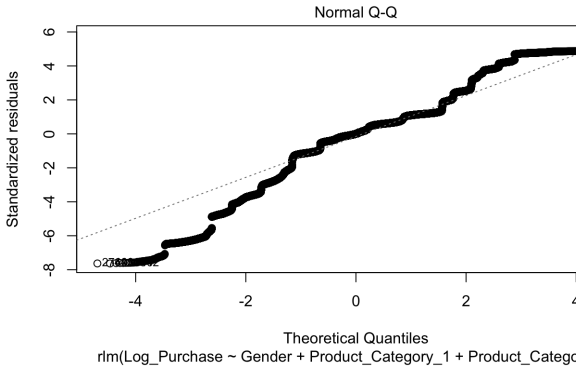



Figure 13: QQ plot

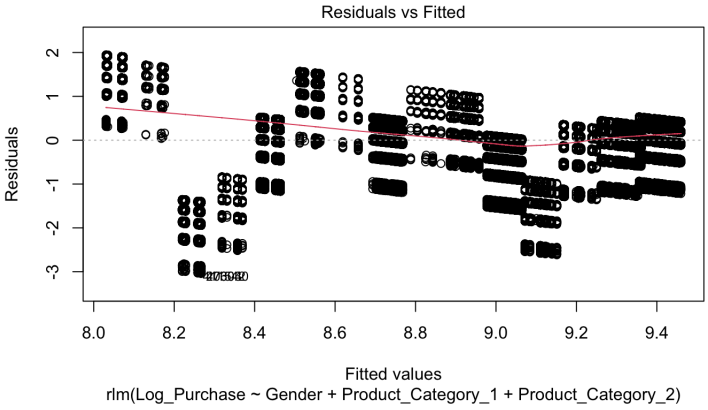



Figure 14: Residual Plot for Robust Model

The model has a residual standard error of 0.3985, which suggests that while the model captures the central tendencies well (as indicated by the significant t-values), there is still considerable unexplained variance. The range of residuals (-3.040917 to 1.948054) also indicates that there may be outliers or extreme values in the purchase data that the model does not adequately account for, typical in real-world data scenarios where robust regression techniques are beneficial.

6. Bayesian Regression Model

Data:

$N$ : Number of observations

$K$ : Number of predictors

$X$ : Matrix of predictors of size $N \times K$

$y$ : Vector of observed values

Parameters: (6)

$\beta$ : Vector of coefficients of size $K$

$\alpha$ : Intercept

$\sigma$ : Standard deviation of errors

Model:

$$y \sim \text{Normal}(X\beta + \alpha, \sigma)$$

Parameter Estimates and Uncertainty: The Bayesian model provides mean estimates and standard errors for the coefficients (beta[1] to beta[6]) and the intercept (alpha), as well as the residual standard deviation (sigma). These values indicate the central tendency and uncertainty of each parameter. For example, beta[3] and beta[4] have relatively large mean values, but their standard errors and the range between the 0.025 and 0.975 percentiles also suggest high uncertainty in these estimates.
Model Convergence Issues: The Rhat values, which should ideally be close to 1, are significantly greater than 1 for many parameters, indicating potential convergence issues in the Bayesian sampling process. This suggests that the model may not

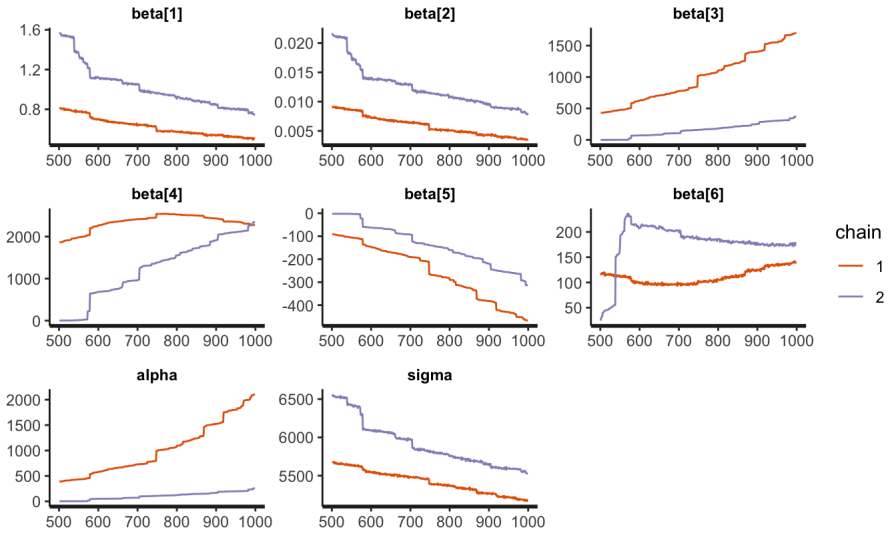



Figure 15: Trace plot

have adequately explored the parameter space, leading to unreliable estimates. The n eff values, representing the effective sample size, are also low for several parameters, further supporting the presence of convergence issues.
In summary, while the Bayesian model provides a probabilistic framework for understanding the relationships in the data, the apparent convergence problems indicated by high Rhat values and low effective sample sizes suggest that the model results may not be reliable, and additional steps are needed to improve the model's convergence and the robustness of its estimates.

# 3 Result

Table 2: Model Comparison

| Model | R-squared | MSE | MAE | p-value |
|---|---|---|---|---|
| Null Model | 0 | 0.5460 | 0.7393 | - |
| No pooling Model | 0.9958 | 0.3472251 | 0.4237728 | 2.2e-16 |
| Complete pooling Model | 0.2208 | 0.3457825 | 0.4226185 | 2.2e-16 |
| Partial pooling Model | 0.218 | 0.3472249 | 0.4237722 | - |
| Robust Model | 0.56069 | 0.3515939 | 0.4188003 | - |

The "No pooling Model" performs best in terms of variance explained, but the lack of p-values for two models may require further investigation to fully assess model validity. The "Robust Model" might be preferred if the data contains outliers or non-normal errors due to its median-based metrics (MAE), which are less sensitive to outliers compared to mean-based metrics (MSE).

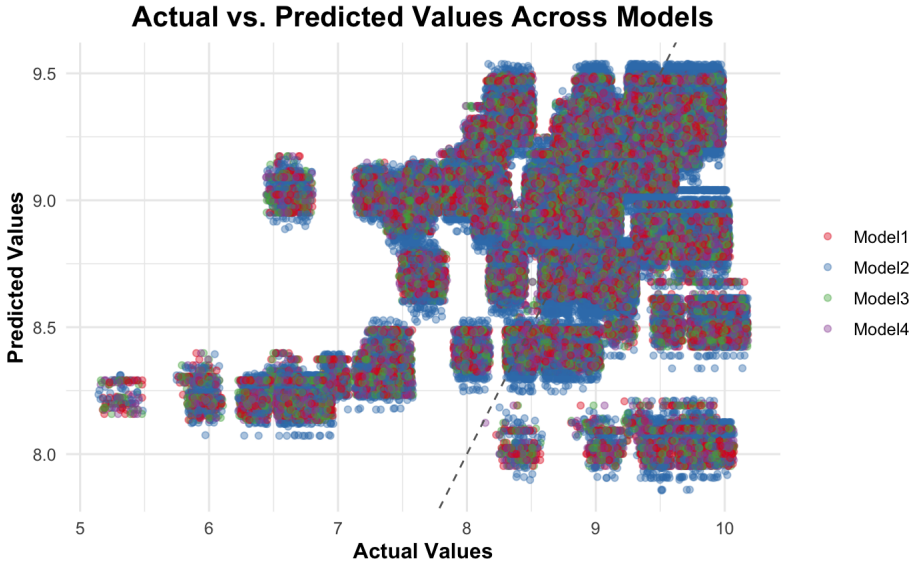



Figure 16: Trace plot

There's a significant overlap among the predictions of all four models, which is especially evident for actual values between approximately 7 and 9. This overlap indicates that the models may be producing similar predictions for many of the data points.The models seem to have varying degrees of prediction ranges. For instance, some models have predictions that extend to the higher end of the scale (near 9.5), while others are more concentrated around the middle (near 8.5).

In conclusion, the "No pooling Model" significantly outperformed other models, as evidenced by an R-squared value of 0.9958 and a highly significant p-value. Despite the high explanatory power of the "No pooling Model", the "Robust Model" exhibited the lowest MAE, suggesting its potential utility in datasets with outliers.

# 4 Discussion

In terms of modeling, the lower MAE of the 'robust model' suggests that it may be better suited to predicting purchasing behavior when dealing with data containing outliers or non-standard distributions. This is particularly important when considering the impact of intrinsic product retail prices on customer purchasing behavior, as extreme values may reflect premium or heavily discounted items.

In terms of data, future research steps also include integrating temporal data to analyze the longitudinal impact of Black Friday on consumer purchasing trends, either by adding data from Black Friday in different years or sales data from usual weekdays.

Meanwhile, on this basis, adding more holiday data from companies in different regions to train the model more effectively has achieved good prediction results.

In conclusion, applying multilevel modeling to the Black Friday dataset reveals an important relationship between consumer characteristics and purchasing behavior. The next step will be to start from the product level, observing multiple factors such as product pricing, product type, and product factory in order to consider other complex factors in the field of consumer behavior.

# References


[1] M. Javed Awan, M. S. Mohd Rahim, H. Nobanee, A. Yasin, and O. I. Khalaf, "A big data approach to black friday sales," *MJ Awan, M. Shafry, H. Nobanee, A. Yasin, OI Khalaf et al.," A big data approach to black friday sales," Intelligent Automation & Soft Computing*, vol. 27, no. 3, pp. 785–797, 2021.

[2] C.-S. M. Wu, P. Patil, and S. Gunaseelan, "Comparison of different machine learning algorithms for multiple regression on black friday sales data," in *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, 2018, pp. 16–20.

[3] S. Kalra, B. Perumal, S. Yadav, and S. J. Narayanan, "Analysing and predicting the purchases done on the day of black friday," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, 2020, pp. 1–8.

# A R Code

```
data$Product_ID <- as.numeric(sub("P", "", data$Product_ID))

data$Gender <- as.numeric(factor(data$Gender))
data$Gender <- ifelse(data$Gender == "2", 1, 0)
data$City_Category <- as.numeric(factor(data$City_Category))
data$Stay_In_Current_City_Years <- as.numeric(sub("\\+", "", data$Stay_In_Current_City_Years))

data$Log_Purchase <- log(data$Purchase + 1)

unique(data$Age)
age_levels <- c("0-17", "18-25", "26-35", "36-45", "46-50", "51-55", "55+")
data$Age <- factor(data$Age, levels = age_levels, labels = 0:6)
data$Age <- as.numeric(data$Age)
data$User_ID <- data$User_ID - 1000000
hist(data$Log_Purchase, main="Histogram of Log-transformed Purchase", xlab="Log(Purchase)", ylab="Frequency")
variable_details <- data.frame(
  Variable = c("User_ID", "Product_ID", "Gender", "Age", "Occupation", "City_Category",
               "Stay_In_Current_City_Years", "Marital_Status", "Product_Category_1",
               "Product_Category_2", "Product_Category_3", "Purchase"),
  Type = c("Integer", "Integer", "Integer (Categorical)", "Integer (Categorical)",
           "Integer (Categorical)", "Integer (Categorical)", "Integer (Categorical)",
           "Integer (Categorical)", "Integer", "Integer", "Integer", "Integer"),
  Description = c("Unique identifier of the user",
                  "Unique identifier of the product",
                  "Gender of the user (0 for female, 1 for male)",
                  "Age group of the user (coded as integers)",
                  "Occupation code of the user",
                  "Category of the city (1 for A, 2 for B, 3 for C )",
                  "Number of years the user has lived in the current city",
                  "Marital status of the user (0 for single, 1 for married)",
                  "Product category code 1",
                  "Product category code 2 (may contain NA for missing values)",
                  "Product category code 3 (may contain NA for missing values)",
                  "Purchase amount in dollars")
)

knitr::kable(variable_details, format = "markdown", caption = "Overview of Variables in Dataset")
ggplot(data, aes(x = Log_Purchase)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = "blue", alpha = 0.7) +
  geom_density(colour = "red", size = 1) +
  labs(title = "Histogram of Purchase")

qqnorm(data$Log_Purchase)
qqline(data$Log_Purchase, col = "red")
cor_matrix <- cor(select_if(data, is.numeric), use = "complete.obs")

corrplot(cor_matrix, method = "color",
         type = "upper", # Display only upper half of the matrix
         order = "hclust", # Order the matrix based on hierarchical clustering
         tl.col = "black", # Text label color
         tl.srt = 45, # Text label rotation
         tl.cex = 0.6, # Text label size
         diag = FALSE, # Don't display the diagonal
         addCoef.col = "black", # Add correlation coefficient on the plot
         number.cex = 0.7, # Size of the correlation coefficient
         cl.cex = 0.7, # Size of the color legend text
         cl.ratio = 0.1, # Ratio of the color legend size
         col = colorRampPalette(c("#7F0000", "white", "#003C30"))(200)) # Custom color palette

ggplot(data, aes(x = as.factor(City_Category), y = Log_Purchase)) +
  geom_boxplot() +
  labs(title = "Boxplot of Purchase by City Category", x = "City Category", y = "Purchase Amount")

ggplot(data, aes(x = Log_Purchase)) +
  geom_density(fill = "blue", alpha = 0.5) +
  labs(title = "Density Plot of Purchase Amount", x = "Purchase Amount", y = "Density")
```

```r
ggplot(data, aes(x = as.factor(Gender), y = Log_Purchase)) +
  geom_violin(trim = FALSE) +
  labs(title = "Violin Plot of Purchase by Gender", x = "Gender", y = "Purchase Amount")
ggplot(data, aes(x = as.factor(Age), y = Log_Purchase, fill = as.factor(Gender))) +
  geom_bar(stat = "summary", fun = "sum", position = "stack") +
  labs(title = "Stacked Bar Chart of Purchase by Age and Gender",
       x = "Age",
       y = "Total Purchase Amount",
       fill = "Gender") +
  theme_minimal()

data <- subset(data, select = -c(Occupation,Marital_Status,Product_Category_3,Stay_In_Current_City_Years,Age

set.seed(123)

train_ratio <- 0.7


train_n <- floor(train_ratio * nrow(data))

train_indices <- sample(seq_len(nrow(data)), size = train_n)

train_set <- data[train_indices, ]
test_set <- data[-train_indices, ]
train_set <- na.omit(train_set)
test_set <- na.omit(test_set)

null_model <- lm(Log_Purchase ~ 1, data = data)
summary(null_model)
plot(data$Log_Purchase, null_model$residuals,
     main = "Residuals vs. Observed Values",
     xlab = "Observed Values",
     ylab = "Residuals")

qqnorm(null_model$residuals)
qqline(null_model$residuals)
npm <- lm(formula = Log_Purchase ~ Gender + Product_Category_1 + Product_Category_2 + factor(City_Category) +
plot(npm, which = 2)
summary(npm)

predictions <- predict(npm, newdata = test_set)


mse <- mean((test_set$Log_Purchase - predictions)^2)
mae <- mean(abs(test_set$Log_Purchase - predictions))
print(mse)
print(mae)
plot(npm, which = 1)
ggplot(test_set, aes(x = predictions, y = Log_Purchase)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, col = "red")

cpm <- lm(formula =  Log_Purchase ~ ., data = data)
summary(cpm)
predictions <- predict(cpm, newdata = test_set)


mse <- mean((test_set$Log_Purchase - predictions)^2)
mae <- mean(abs(test_set$Log_Purchase - predictions))
print(mse)
print(mae)
ggplot(test_set, aes(x = predictions, y = Log_Purchase)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, col = "red")
plot(cpm, which = 1)
plot(cpm, which = 2)
```

```
partial_pooling_model <- lmer(Log_Purchase ~ Gender + Product_Category_1 + Product_Category_2 + (1 | City_Cat
summary(partial_pooling_model)

predictions <- predict(partial_pooling_model, newdata = test_set)


mse <- mean((test_set$Log_Purchase - predictions)^2)
mae <- mean(abs(test_set$Log_Purchase - predictions))
print(mse)
print(mae)
library(lme4)
library(performance)
r2(partial_pooling_model)
ggplot(test_set, aes(x = predictions, y = Log_Purchase)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, col = "red")
residuals <- resid(partial_pooling_model)
fitted_values <- fitted(partial_pooling_model)

ggplot() +
  geom_point(aes(x = fitted_values, y = residuals)) +
  geom_smooth(aes(x = fitted_values, y = residuals), color = "red") +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Fitted Values", y = "Residuals", title = "Residual Plot for Partial Pooling Model")
library(MASS)
robust_model <- rlm(Log_Purchase ~ Gender + Product_Category_1 + Product_Category_2, data = data)

summary(robust_model)

predictions <- predict(partial_pooling_model, newdata = test_set)


mse <- mean((test_set$Log_Purchase - predictions)^2)
mae <- mean(abs(test_set$Log_Purchase - predictions))
print(mse)
print(mae)
ggplot(test_set, aes(x = predictions, y = Log_Purchase)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, col = "red")
plot(robust_model, which = 1)
plot(robust_model, which = 2)

hcode<- "data {
  int<lower=0> N;
  int<lower=0> K;
  matrix[N, K] X;
  vector[N] y;
}
parameters {
  vector[K] beta;
  real alpha;
  real<lower=0> sigma;
}
model {
  y ~ normal(X * beta + alpha, sigma);
}
"
df<- na.omit(data)
stan_data <- list(
  N = nrow(df),
  K = ncol(df[, -which(names(df) == "Purchase")]),
  X = as.matrix(df[, -which(names(df) == "Purchase")]),
  y = df$Purchase
)

fit <- stan(model_code =hcode, data = stan_data, iter = 1000, chains = 2)
summary(fit)
library(rstan)
```

```
library(bayesplot)
traceplot(fit)
dall <- data.frame(
  Actual = test_set$Log_Purchase,
  Model1 = predict(npm, newdata = test_set),
  Model2 = predict(cpm, newdata = test_set),
  Model3 = predict(partial_pooling_model, newdata = test_set),
  Model4 = predict(partial_pooling_model, newdata = test_set)
  )
data_long <- pivot_longer(
  dall,
  cols = starts_with("Model"),
  names_to = "Model",
  values_to = "Prediction"
)

ggplot(data_long, aes(x = Actual, y = Prediction, color = Model)) +
  geom_point(position = position_jitter(width = 0.1, height = 0), alpha = 0.4, size = 1.5) +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "grey40", size = 0.5) +
  scale_color_brewer(palette = "Set1") +
  theme_minimal(base_size = 12) +
  labs(
    title = "Actual vs. Predicted Values Across Models",
    x = "Actual Values",
    y = "Predicted Values"
  ) +
  theme(
    legend.position = "right",
    legend.title = element_blank(),
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    axis.title = element_text(face = "bold")
  )
```