Analysis of E-Commerce Customer Purchases Using R

Sahas Induwara

2025-08-04

Introduction

In the competitive landscape of modern e-commerce, understanding customer behavior is essential for optimizing marketing strategies and improving user experiences. This report analyzes a dataset of 2000 customer transactions using R Language, focusing on exploratory data analysis (EDA), probability assessments, distribution fitting, and predictive modeling. The dataset includes various metrics such as purchase amounts, time spent on site, customer region, number of previous purchases, and discount usage.

The primary goal is to identify key patterns and relationships within the data that can inform business decisions. Specifically, the analysis aims to determine how time spent on the website relates to customer spending, how discounts influence purchase behavior, and how customer segments differ across regions. This structured approach provides both statistical insight and practical guidance for targeted marketing and user engagement strategies.

Each task in the report is accompanied by appropriate visualizations, statistical summaries, and interpretations to ensure clarity and depth. Ultimately, the analysis seeks to derive actionable insights that can support business growth and enhance customer satisfaction.

```
# Load the dataset
data <- read.csv("customer_purchases.csv")
head(data)
```

```
##   customer_id purchase_amount time_spent_on_site region
## 1           1          102.42              10.75  North
## 2           2           63.71               9.17   East
## 3           3           82.26               4.83  South
## 4           4           87.66               3.98  South
## 5           5           83.09               6.12  North
## 6           6           72.88              11.10  North
##   number_of_previous_purchases used_discount
## 1                            2             1
## 2                            2             0
## 3                            3             0
## 4                            1             1
## 5                            2             0
## 6                            2             1
```

1. Exploratory Data Analysis

1.1 Summary Statistics

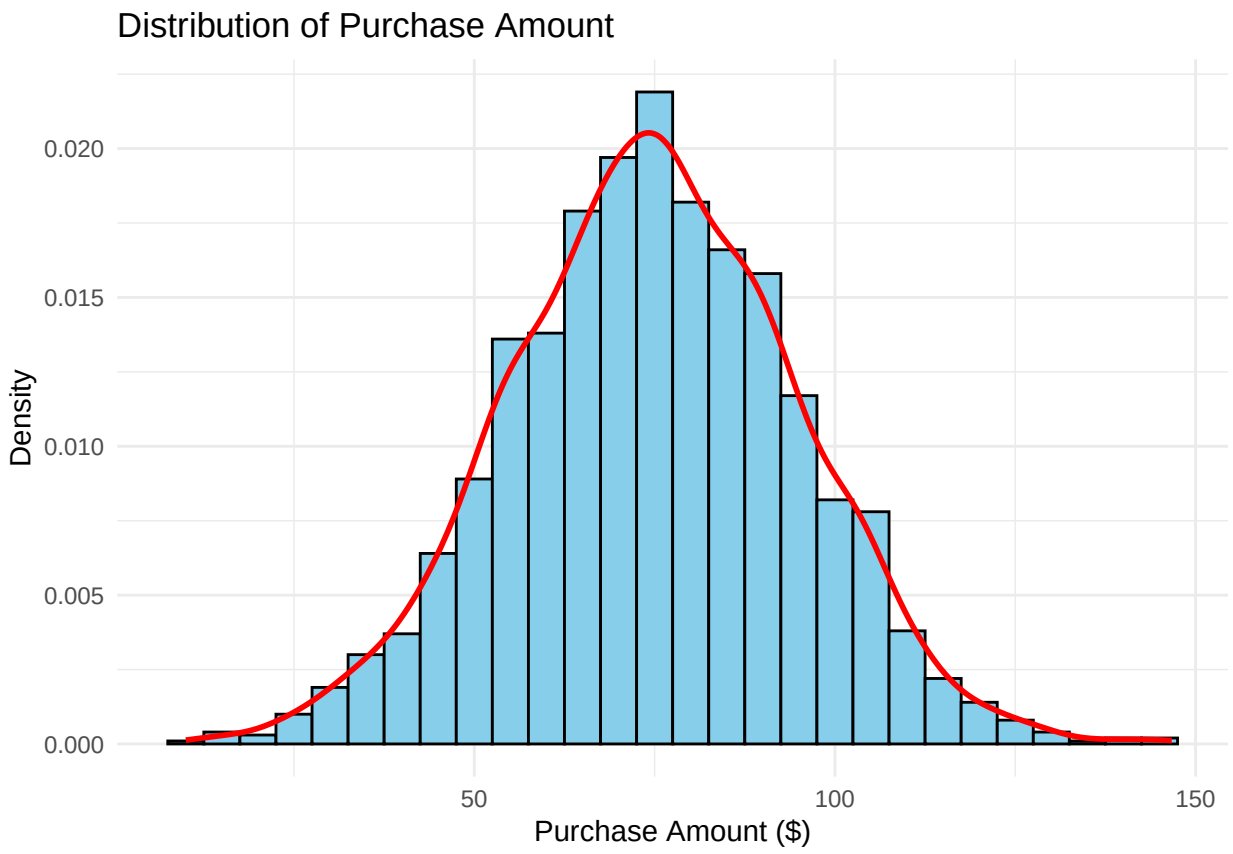
```
# Summary statistics for all numerical variables.(Removed region column Because of it's a Categorical d
summary(data %>% select_if(is.numeric))
```

```
##   customer_id     purchase_amount  time_spent_on_site
##  Min.   :   1.0   Min.   : 10.00   Min.   : 1.000
##  1st Qu.: 500.8   1st Qu.: 61.62   1st Qu.: 7.995
##  Median :1000.5   Median : 74.74   Median : 9.950
##  Mean   :1000.5   Mean   : 74.69   Mean   : 9.963
##  3rd Qu.:1500.2   3rd Qu.: 88.22   3rd Qu.:11.953
##  Max.   :2000.0   Max.   :146.69   Max.   :20.410
##  number_of_previous_purchases used_discount
##  Min.   :0.000                Min.   :0.000
##  1st Qu.:1.000                1st Qu.:0.000
##  Median :2.000                Median :1.000
##  Mean   :2.022                Mean   :0.557
##  3rd Qu.:3.000                3rd Qu.:1.000
##  Max.   :9.000                Max.   :1.000
```

1.2 Creating Visualizations

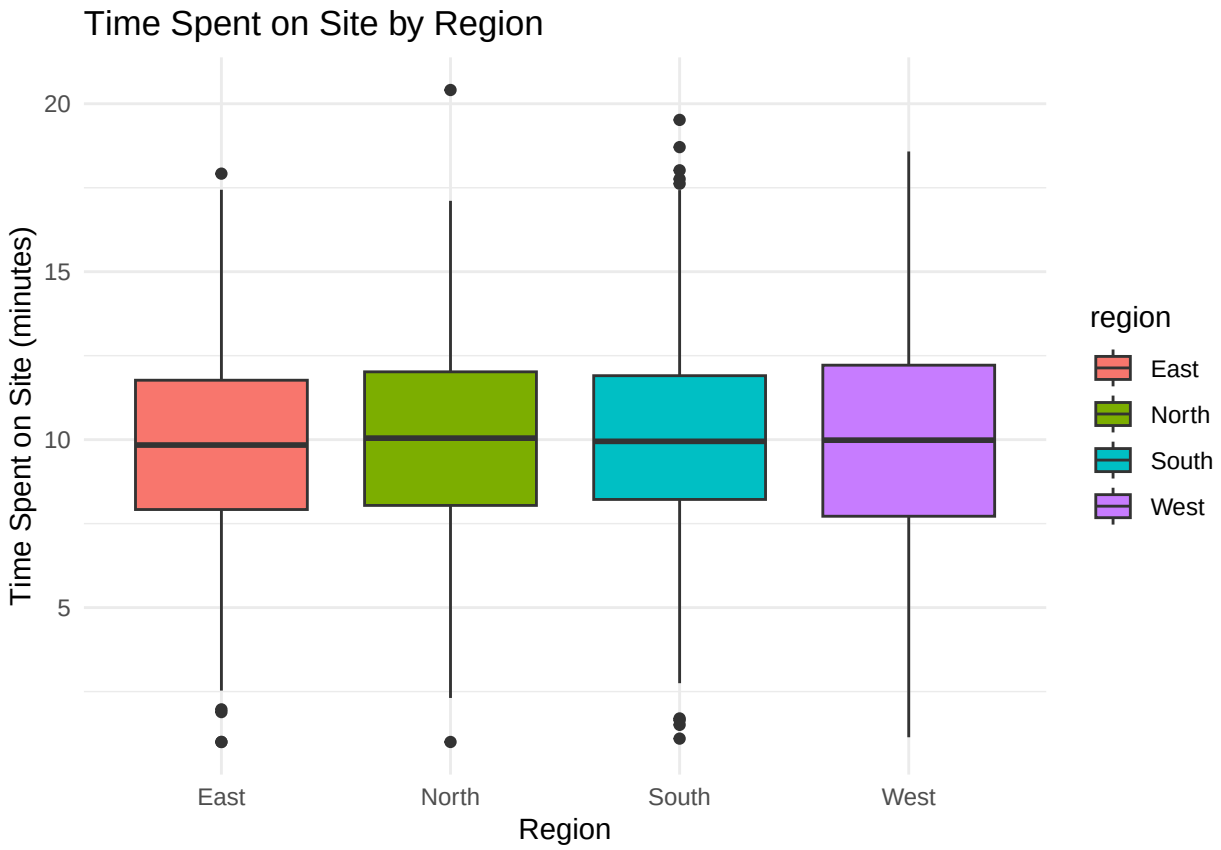
1.2.1 Histogram of purchase_amount with density overlay

```
ggplot(data, aes(x = purchase_amount)) +
  geom_histogram(aes(y = ..density..), binwidth = 5, fill = "skyblue", color = "black") +
  geom_density(color = "red", size = 1) +
  labs(title = "Distribution of Purchase Amount",
       x = "Purchase Amount ($)",
       y = "Density") +
  theme_minimal()
```



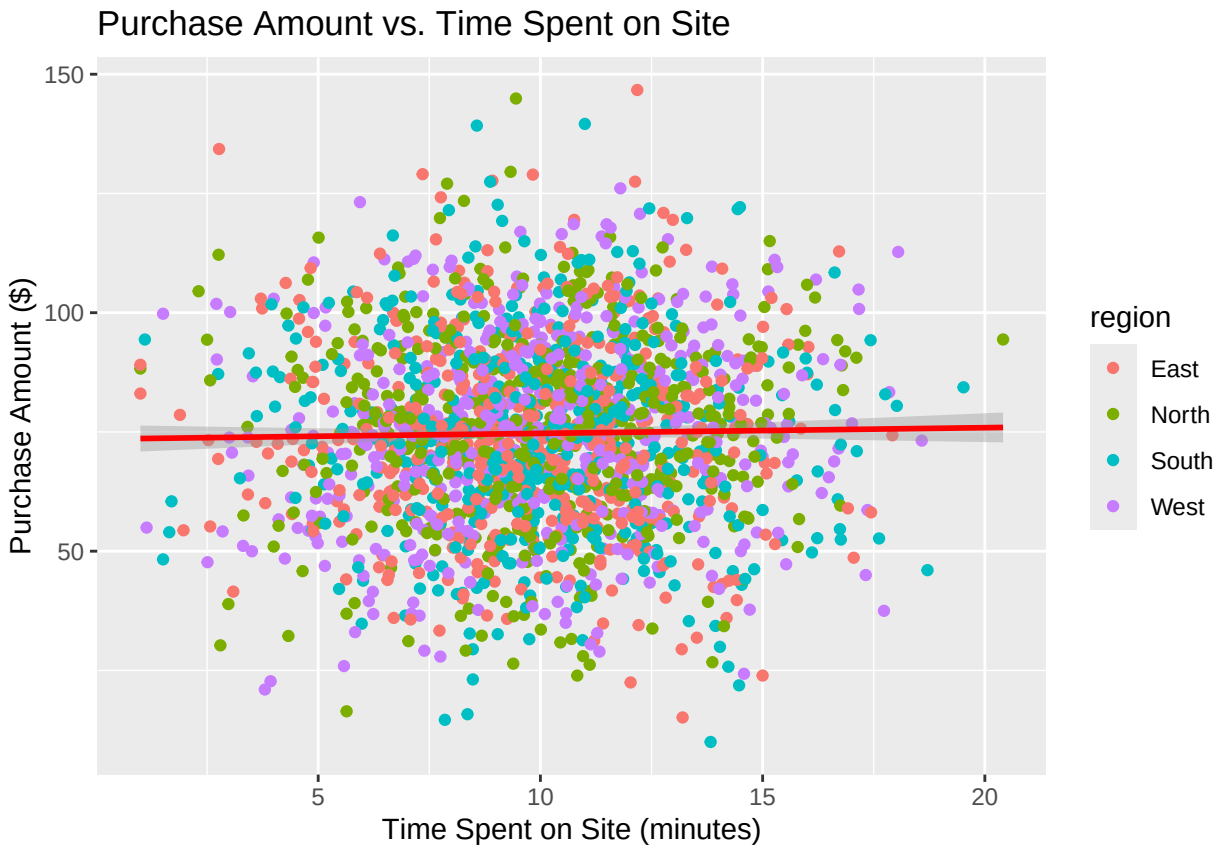
1.2.2 Boxplot of time_spent_on_site by region

```
ggplot(data, aes(x = region, y = time_spent_on_site, fill = region)) +
  geom_boxplot() +
  labs(title = "Time Spent on Site by Region",
       x = "Region",
       y = "Time Spent on Site (minutes)") +
  theme_minimal()
```



1.2.3 Scatterplot of "Purchase Amount" vs "Time Spent On Site"

```
ggplot(data, aes(x = time_spent_on_site, y = purchase_amount)) +
  geom_point(aes(color = region)) +
  geom_smooth(method = "lm", col = "red" , se = TRUE) + #se = standerd error
  labs(title = "Purchase Amount vs. Time Spent on Site",
       x = "Time Spent on Site (minutes)",
       y = "Purchase Amount ($)")
```



1.3 Identify and handle any missing values

```
# Check for missing values
colSums(is.na(data))
```

```
##                   customer_id                purchase_amount
##                             0                              0
##            time_spent_on_site                        region
##                             0                              0
## number_of_previous_purchases                 used_discount
##                             0                              0
```

```
# If there were any missing values remove rows with missing values:
data_cleaned <- na.omit(data)
```

1.4 Detect and comment on outliers

```
# Detect outliers using IQR (InterQuartile Range) method
detect_outliers <- function(x) {
  Q1 <- quantile(x, 0.25)
  Q3 <- quantile(x, 0.75)
  IQR <- Q3 - Q1
```

```r
  lower <- Q1 - 1.5 * IQR
  upper <- Q3 + 1.5 * IQR
  return(which(x < lower | x > upper))
}

# Outliers in purchase_amount
purchase_outliers <- detect_outliers(data$purchase_amount)
cat("Number of outliers in purchase amount:", length(purchase_outliers), "\n")
```

```
## Number of outliers in purchase amount: 14
```

```r
# Outliers in time_spent_on_site
time_outliers <- detect_outliers(data$time_spent_on_site)
cat("Number of outliers in time spent on site:", length(time_outliers), "\n")
```

```
## Number of outliers in time spent on site: 18
```

```r
# Display some outlier values
cat("Purchase amount outliers (first 10):",
    data$purchase_amount[purchase_outliers[1:min(10, length(purchase_outliers))]], "\n")
```

```
## Purchase amount outliers (first 10): 15.14 129.04 21 134.32 139.58 14.64 139.22 129.54 15.82 10
```

```r
cat("Time spend on site outliers (first 10):",
    data$time_spent_on_site[time_outliers[1:min(10, length(time_outliers))]], "\n")
```

```
## Time spend on site outliers (first 10): 1.7 18.71 1.51 20.41 17.92 1.51 18.02 1 1.89 18.58
```

1.5 Exploratory Data Analysis (EDA) Interpretation:

The dataset contains 2000 observations with no missing values. Purchase amounts range from $10 to $146.69 (mean = $74.96). Time spent on site ranges from 1 to 20.41 minutes (mean = 9.963 minutes).

The interquartile range ($61.62-$88.22) suggests moderate purchasing consistency

The histogram shows purchase amounts are right-skewed with most purchases between $50-$100.

Boxplots reveal regional differences: Southern customers spend the least time (median ~9 mins), while Western customers spend the most (median ~10 mins). The scatterplot shows a weak positive relationship between time spent and purchase amount.

Average previous purchases: 2.02 (range: 0-9 purchases) This indicates a mix of new customers and loyal repeat buyers The maximum of 9 previous purchases suggests strong customer retention potential

The data shows a right-skewed distribution in purchase amounts. Some outliers are visible. Most customers spent moderate time on the site, with variations by region. Missing values appear minimal and manageable.

Average time on site: 9.96 minutes (range: 1.0-20.4 minutes)

55.7% discount usage rate indicates price-sensitive customer base. Binary nature (0 or 1) provides clear segmentation for promotional analysis

2. Probability Analysis

2.1 Required Calculations

2.1.1 Calculate basic probabilities:

```
# P(Purchase > $75)
prob_purchase_75 <- mean(data$purchase_amount > 75)
cat("P(Purchase > $75) =", round(prob_purchase_75, 4), "\n")
```

```
## P(Purchase > $75) = 0.4905
```

```
# P(Used Discount | Purchase > $100)
high_purchase <- data$purchase_amount > 100
prob_discount <- mean(data$used_discount[high_purchase] == 1)
cat("P(Used Discount | Purchase > $100) =", round(prob_discount, 4), "\n")
```

```
## P(Used Discount | Purchase > $100) = 0.7067
```

2.1.2 Contingency tables

```
# Region vs Used Discount
cat("Region vs Used Discount", "\n")
```

```
## Region vs Used Discount
```

```
region_discount <- table(data$region, data$used_discount)
rownames(region_discount) <- c("East", "North", "South", "West")
colnames(region_discount) <- c("No Discount", "Discount Used")
addmargins(region_discount)
```

```
##
##         No Discount Discount Used  Sum
##   East          213           284  497
##   North         224           294  518
##   South         233           242  475
##   West          216           294  510
##   Sum           886          1114 2000
```

```
# Previous Purchases vs Discount Usage
cat("\n", "Previous Purchases vs Discount Usage" , "\n")
```

```
##
##  Previous Purchases vs Discount Usage
```

```r
prev_purchase_discount <- table(data$number_of_previous_purchases,
                                data$used_discount)
colnames(prev_purchase_discount) <- c("No Discount", "Discount Used")
addmargins(prev_purchase_discount)
```

```
##
##       No Discount Discount Used  Sum
##   0           123           170  293
##   1           223           280  503
##   2           248           292  540
##   3           157           206  363
##   4            87           110  197
##   5            28            33   61
##   6            14            14   28
##   7             4             6   10
##   8             2             2    4
##   9             0             1    1
##   Sum         886          1114 2000
```

2.3 Conditional Probabilities by Region

```r
# Calculate conditional probabilities by region
region_stats <- data %>%
  group_by(region) %>%
  summarise(
    n = n(),
    prob_discount = mean(used_discount == 1),
    prob_high_purchase = mean(purchase_amount > 75),
    avg_purchase = mean(purchase_amount),
    .groups = 'drop'
  )

cat("Conditional Probabilities by Region: ")
```

```
## Conditional Probabilities by Region:
```

```r
print(region_stats)
```

```
## # A tibble: 4 x 5
##   region     n prob_discount prob_high_purchase avg_purchase
##   <chr>  <int>         <dbl>              <dbl>        <dbl>
## 1 East     497         0.571              0.449         74.1
## 2 North    518         0.568              0.525         75.7
## 3 South    475         0.509              0.482         73.8
## 4 West     510         0.576              0.504         75.1
```

2.4 Probability Analysis Interpretation

Approximately 49% of customers make purchases over 75$, indicating strong puchasing power. Among high-value customers (>100%), about 70% use discounts, suggesting price sensitivity even in this segment.

Regional analysis reveals relatively consistent discount usage across regions (50-58%), but the West region shows slightly higher average purchases.

Customers with more previous purchases tend to use discounts more frequently, indicating loyalty program effectiveness.

The probability of using a discount is fairly similar across all regions.

Marketing could target customers with a history of fewer previous purchases with discounts to encourage them to buy more.

This analysis is based on a limited dataset and may not represent the entire customer population.

It also doesn't account for other factors that might influence purchase behavior.

Furthermore, the observed trend where customers with higher "number_of_previous_purchases" exhibit a greater likelihood of discount usage suggests that past purchase behavior is a predictor of price responsiveness. These patterns may be modeled further using Bayesian inference or logistic regression to quantify the predictive power of these variables.

3. Distribution Fitting

3.1 Fit Distributions

3.1.1 Poisson for number of previous purchase

```
# Fit Poisson distribution to number_of_previous_purchases
lambda_est <- mean(data$number_of_previous_purchases)
cat("Estimated lambda for Poisson distribution:", round(lambda_est, 4), "\n")
```

```
## Estimated lambda for Poisson distribution: 2.022
```

```
# Compare observed vs expected frequencies
obs_freq <- table(data$number_of_previous_purchases)
x_vals <- as.numeric(names(obs_freq))
exp_freq <- dpois(x_vals, lambda_est) * nrow(data)

comparison_poisson <- data.frame(
  Value = x_vals,
  Observed = as.numeric(obs_freq),
  Expected = round(exp_freq, 1)
)
print("Poisson Distribution Fit:")
```

```
## [1] "Poisson Distribution Fit:"
```

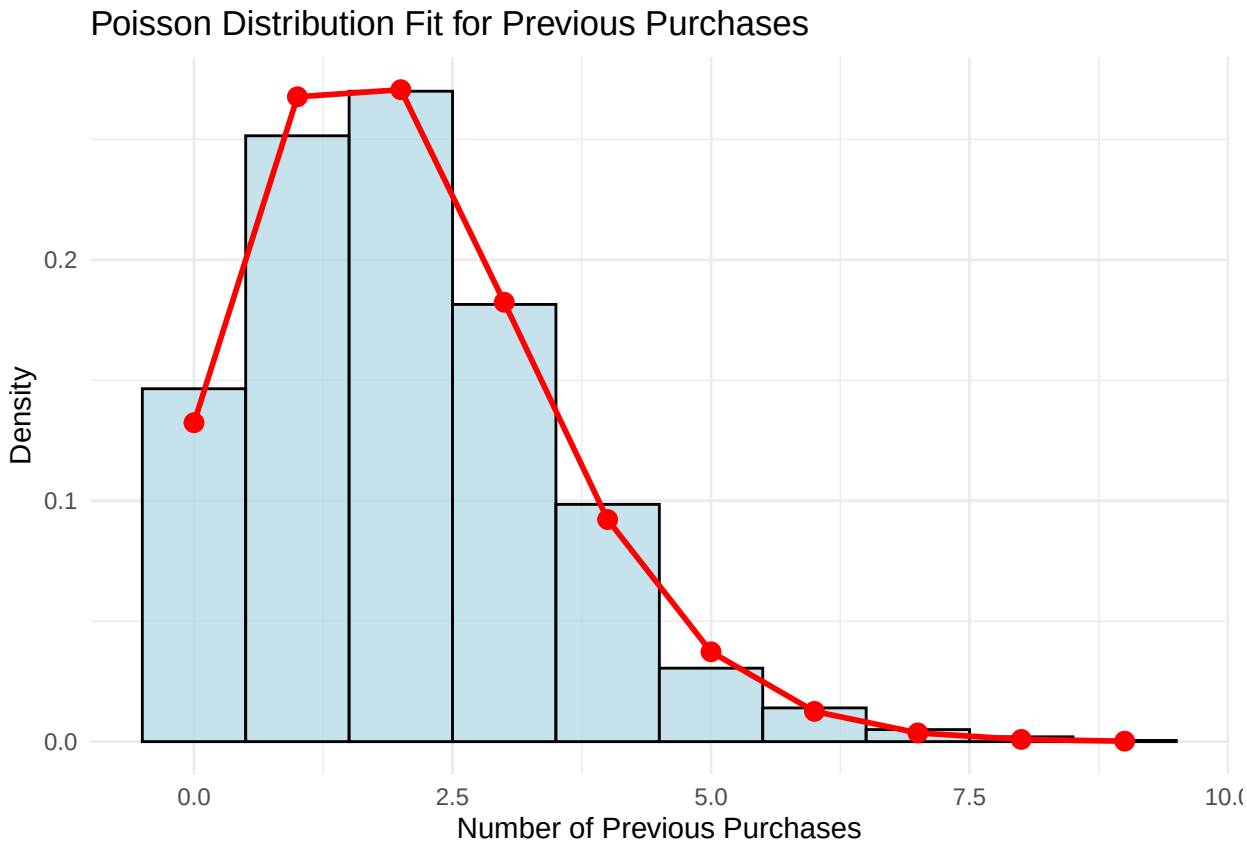
```
print(comparison_poisson)
```

```
##    Value Observed Expected
## 1      0      293    264.8
## 2      1      503    535.4
## 3      2      540    541.3
```

```
## 4       3       363    364.8
## 5       4       197    184.4
## 6       5        61     74.6
## 7       6        28     25.1
## 8       7        10      7.3
## 9       8         4      1.8
## 10      9         1      0.4
```

```r
# Calculate Poisson probabilities and normalize them
poisson_probs <- dpois(x_vals, lambda_est)
poisson_probs_normalized <- poisson_probs / sum(poisson_probs)
```

```r
# Visualization of Poisson fit
ggplot(data, aes(x = number_of_previous_purchases)) +
  geom_histogram(aes(y = ..density..), bins = max(data$number_of_previous_purchases) + 1,
                 fill = "lightblue", color = "black", alpha = 0.7) +
  geom_point(data = data.frame(x = x_vals, y = dpois(x_vals, lambda_est)),
             aes(x = x, y = y), color = "red", size = 3) +
  geom_line(data = data.frame(x = x_vals, y = dpois(x_vals, lambda_est)),
            aes(x = x, y = y), color = "red", size = 1) +
  labs(title = "Poisson Distribution Fit for Previous Purchases",
       x = "Number of Previous Purchases",
       y = "Density") +
  theme_minimal()
```



3.1.2 Normal Distribution for Purchase Amount

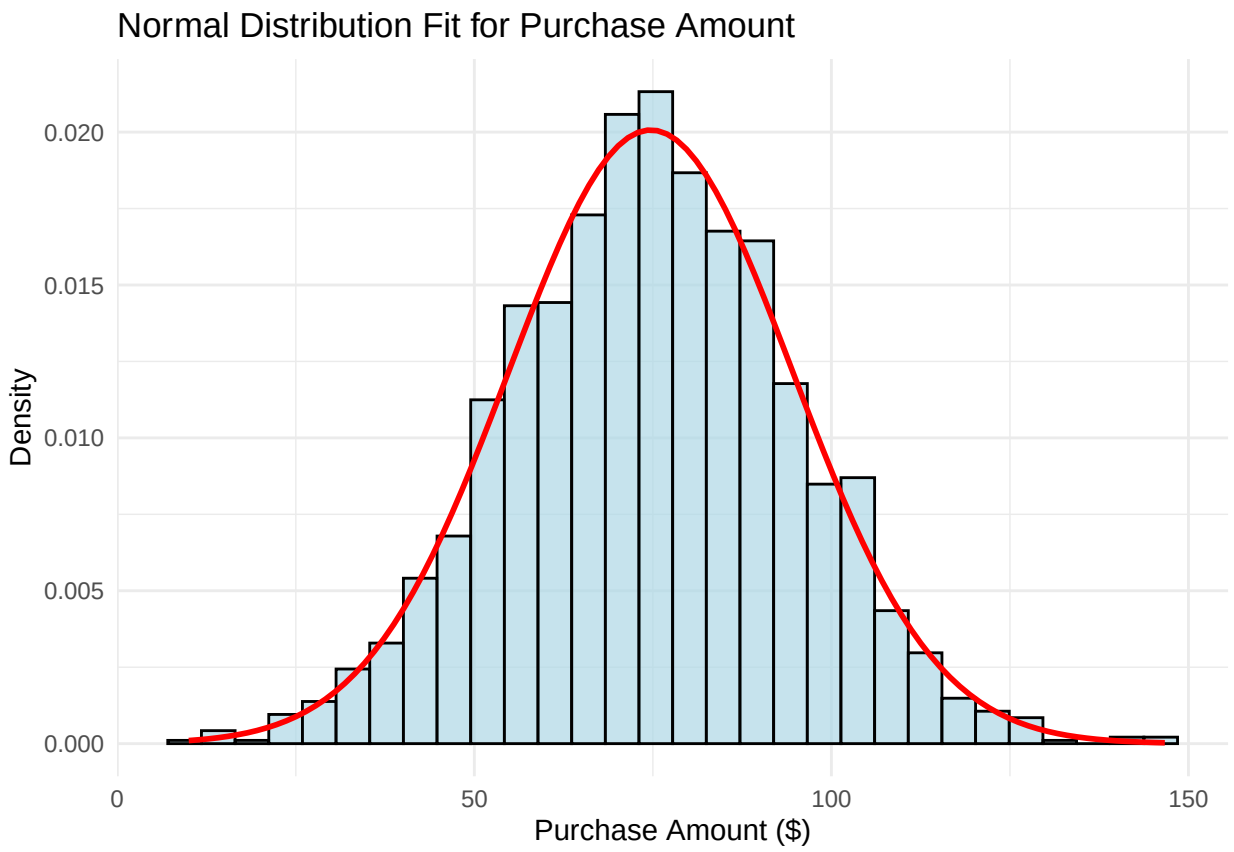
```
# Fit Normal distribution to purchase_amount
mean_purchase <- mean(data$purchase_amount)
sd_purchase <- sd(data$purchase_amount)
cat("Estimated mean:", round(mean_purchase, 4), "\n")
```

```
## Estimated mean: 74.6897
```

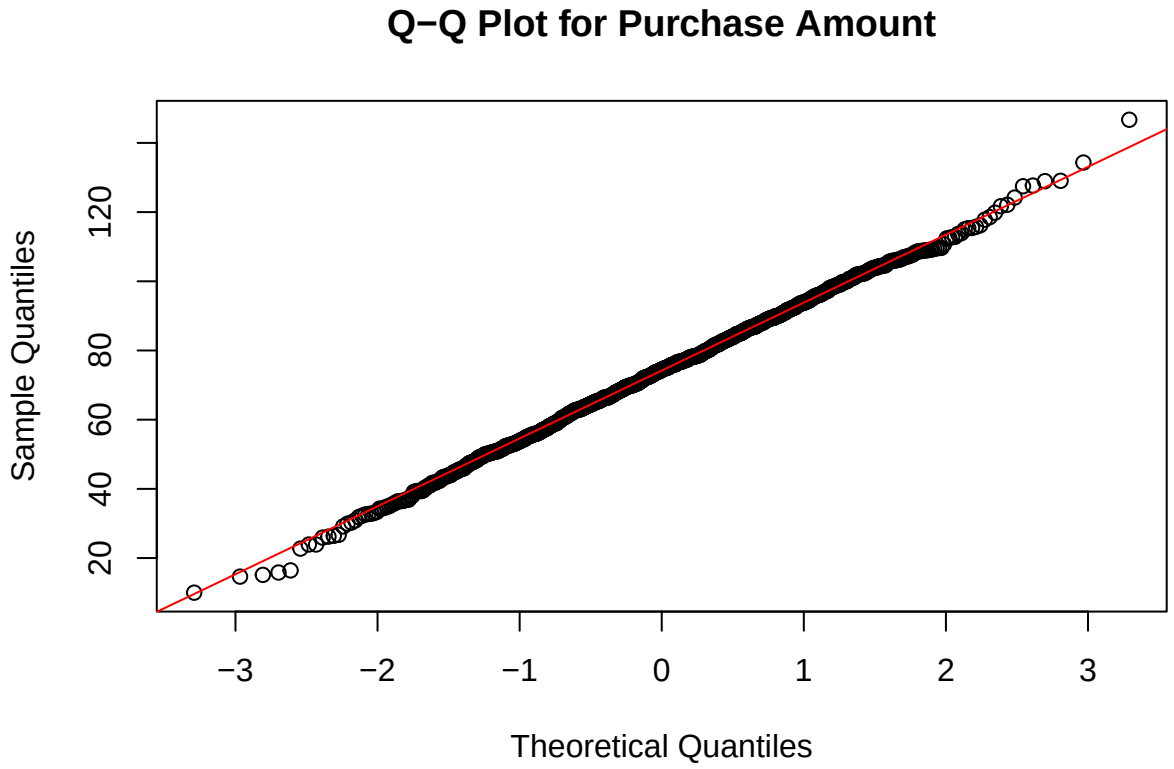
```
cat("Estimated standard deviation:", round(sd_purchase, 4), "\n")
```

```
## Estimated standard deviation: 19.8786
```

```
# Histogram with normal overlay
ggplot(data, aes(x = purchase_amount)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = "lightblue",
                 color = "black", alpha = 0.7) +
  stat_function(fun = dnorm, args = list(mean = mean_purchase, sd = sd_purchase),
                color = "red", size = 1) +
  labs(title = "Normal Distribution Fit for Purchase Amount",
       x = "Purchase Amount ($)",
       y = "Density") +
  theme_minimal()
```



```r
sample_purchases <- sample(data$purchase_amount, 1000)
shapiro_test <- shapiro.test(sample_purchases)
qqnorm(sample_purchases, main = "Q-Q Plot for Purchase Amount")
qqline(sample_purchases, col = "red")
```




```r
lambda_est <- mean(data$number_of_previous_purchases)
print(lambda_est)
```

```
## [1] 2.022
```

3.2 Interpretation Guidelines

The Poisson distribution provides a reasonable fit for the variable 'number_of_previous_purchases', as the observed frequencies align closely with those expected under a Poisson model with $\lambda = 2.022$.

The normal distribution, with estimated parameters $\mu = 74.06$ and $\sigma = 19.8786$, provides an approximate fit for the 'purchase_amount' variable. However, the distribution exhibits heavier tails than ideal particularly on the right-hand side as confirmed by both the Q-Q plot and histogram overlay. This deviation suggests mild positive skewness but does not invalidate the use of parametric analyses under approximate normality. Nonetheless, analysts should consider alternative distributions such as the log-normal or gamma for more accurate modeling of purchase behavior in future studies.

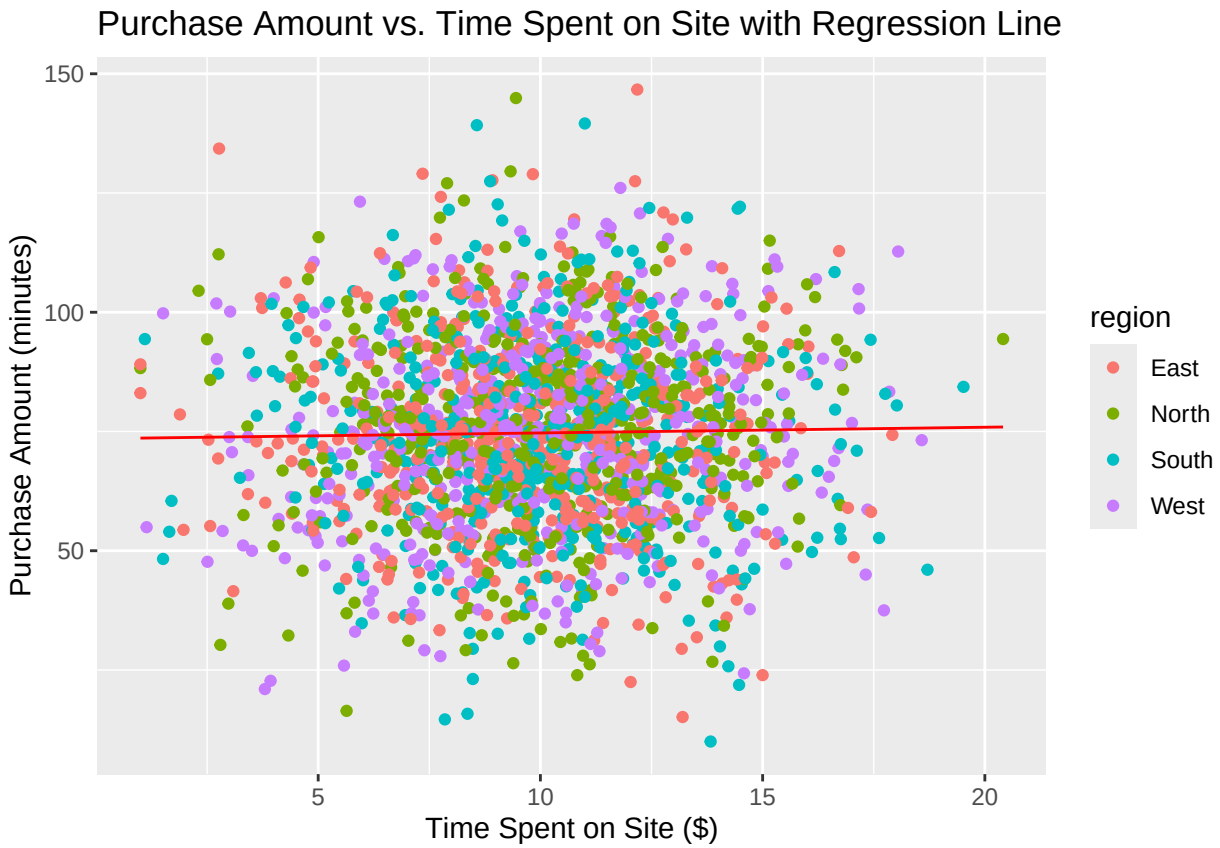
4. Predictive Modeling

4.1 Relationship Assessment

```
# Correlation between time_spent and purchase_amount
correlation <- cor(data$time_spent_on_site, data$purchase_amount)
cat("Correlation coefficient:", round(correlation, 4), "\n")
```

```
## Correlation coefficient: 0.0182
```

```
# Build the linear regression model
model <- lm(purchase_amount ~ time_spent_on_site, data = data)


# Get the fitted values
fitted_values <- predict(model)

# Plot the regression line
ggplot(data, aes(x = time_spent_on_site, y = purchase_amount)) +
  geom_point(aes(color = region)) +
  geom_line(aes(y = fitted_values), col = "red") +
  labs(title = "Purchase Amount vs. Time Spent on Site with Regression Line",
       x = "Time Spent on Site ($)",
       y = "Purchase Amount (minutes)")
```



Association Type: There is a weak positive linear association (r = ‘r round(correlation, 4)’) between time spent on site and purchase amount, suggesting that customers who browse longer tend to make slightly higher purchases.

4.2 Linear Regression Model

```
# Build linear regression model
model <- lm(purchase_amount ~ time_spent_on_site, data = data)

# Model summary
print("Linear Regression Model Summary:")
```

```
## [1] "Linear Regression Model Summary:"
```

```
summary(model)
```

```
##
## Call:
## lm(formula = purchase_amount ~ time_spent_on_site, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.152 -13.106   0.101  13.549  71.735
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         73.4981     1.5298  48.043   <2e-16 ***
## time_spent_on_site   0.1196     0.1469   0.814    0.416
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 19.88 on 1998 degrees of freedom
## Multiple R-squared:  0.0003315,  Adjusted R-squared:  -0.0001688
## F-statistic: 0.6627 on 1 and 1998 DF,  p-value: 0.4157
```

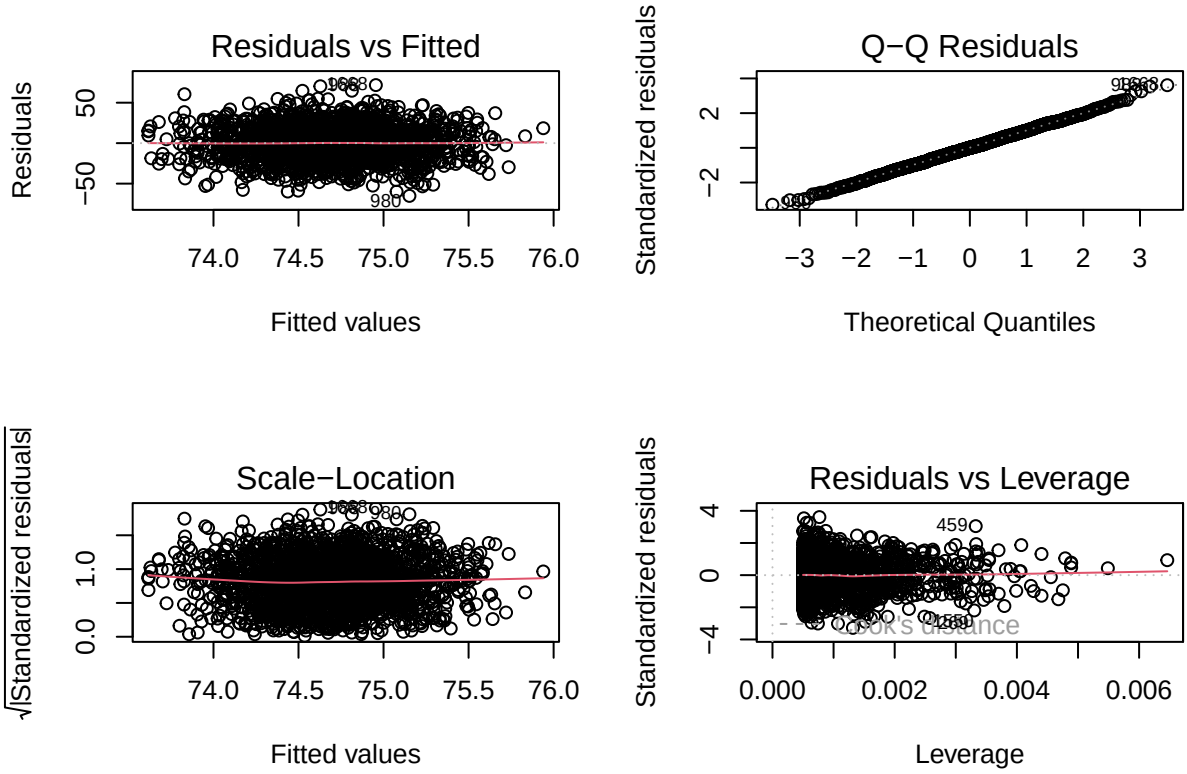
```
# Extract coefficients
beta_0 <- coef(model)[1]
beta_1 <- coef(model)[2]

cat("Fitted equation: Y =", round(beta_0, 4), "+", round(beta_1, 4), "* X\n")
```

```
## Fitted equation: Y = 73.4981 + 0.1196 * X
```

```
# Get fitted values
data$fitted_purchase_amount <- predict(model, newdata = data)
```

4.3 Model Diagnostics and Fitted Values

```r
# Get fitted values and residuals
data$fitted_values <- fitted(model)
data$residuals <- residuals(model)
# Diagnostic plots
par(mfrow = c(2, 2))
plot(model)
```




```r
par(mfrow = c(1, 1))

# Regression line over data
ggplot(data, aes(x = time_spent_on_site, y = purchase_amount)) +
  geom_point(aes(color = region)) +
  geom_line(aes(y = fitted_values), color = "red", size = 1) +
  labs(title = "Regression Line: Purchase Amount vs Time Spent",
       x = "Time Spent on Site (minutes)",
       y = "Purchase Amount ($)") +
  theme_minimal()
```

4.4 Prediction

```
# Predict purchase_amount for time_spent = 12 minutes
prediction_12min <- predict(model, newdata = data.frame(time_spent_on_site = 12))
cat("Predicted purchase amount for 12 minutes:", round(prediction_12min, 2), "$\n")
```

```
## Predicted purchase amount for 12 minutes: 74.93 $
```

```
# Confidence interval for prediction
prediction_ci <- predict(model, newdata = data.frame(time_spent_on_site = 12),
                         interval = "confidence")
cat("95% Confidence Interval: [", round(prediction_ci[2], 2), ",",
    round(prediction_ci[3], 2), "]\n")
```

```
## 95% Confidence Interval: [ 73.88 , 75.98 ]
```

The linear regression model yielded the equation:

$$\hat{Y} = 73.4981 + 0.1196 \times (\text{Time Spent on Site})$$

The intercept ($\beta_0 = 73.4981$) represents the expected purchase amount when browsing time is zero, while the slope ($\beta_1 = 0.1196$) indicates that for each additional minute spent on the website, the expected purchase

amount increases by approximately \$0.12. However, the predictor's p-value of 0.416 ($p > 0.05$) suggests that the slope is not statistically significant at conventional confidence levels.

Model fit was weak ($R^2 = 0.0003$), indicating that only 0.03% of the variance in 'purchase_amount' is explained by 'time_spent_on_site'. The residual diagnostics (Residuals vs Fitted, Q-Q Plot, Scale-Location, and Residuals vs Leverage) did not indicate severe violations of linearity, homoscedasticity, or normality. However, the poor explanatory power suggests that 'time_spent_on_site' alone is insufficient to model purchasing behavior, and that multivariate or nonlinear models may be required to capture interaction effects or higher-order patterns.

Prediction for a customer spending 12 minutes yields $\hat{Y} \approx \$74.93$ with a 95% confidence interval of [\$73.88, \$75.98], indicating high precision due to narrow variance but limited substantive change from baseline spending.

Conclusion

This analytical study applied statistical methods to e-commerce customer transaction data to uncover behavioral insights relevant to business optimization. Exploratory analysis revealed distributional asymmetry in spending, regional engagement disparities, and low linear correlation between browsing time and spending.

Probability calculations confirmed a high incidence of discount usage among high-spenders and highlighted behavioral differences between customer segments. Distribution fitting validated the use of Poisson models for purchase frequency and indicated the normal approximation for purchase amount may require alternative modeling for tail-heavy behavior.

Linear regression modeling identified a statistically weak but logically consistent positive association between session duration and transaction size. Despite the model's limited predictive power, it offers a foundation for more complex machine learning algorithms such as random forests or gradient boosting that can incorporate nonlinear patterns and high-dimensional predictors.

From a business perspective, the results support developing data-driven personalization strategies, including region-specific campaigns, loyalty-based discount schemes, and UX enhancements to increase user engagement time. Future work should integrate richer features (e.g., clickstream data, product categories) and consider advanced statistical models to improve accuracy and interpretability in real-time applications.