Computational Mathematics CW

Nisitha Nimsara, IIT-ID:20240281, RGU-ID:2506755

2025-08-03

```
suppressPackageStartupMessages({
  #load the libraries
  suppressWarnings(library(tidyverse))
})
```

```
#import the csv file
df <- read.csv("customer_purchases.csv")
```

3.1 Exploratory Data Analysis(EDA)

3.1.1 Analysis

3.1.1.a summary statistics for all numerical variables

```
summary(df)
```

```
##   customer_id     purchase_amount  time_spent_on_site    region
##  Min.   :   1.0   Min.   : 10.00   Min.   : 1.000     Length:2000
##  1st Qu.: 500.8   1st Qu.: 61.62   1st Qu.: 7.995     Class :character
##  Median :1000.5   Median : 74.74   Median : 9.950     Mode  :character
##  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
```

Purchase amount: Minimun= $10,
Maximum= $146.69, Mean $74.69, Median $74.74

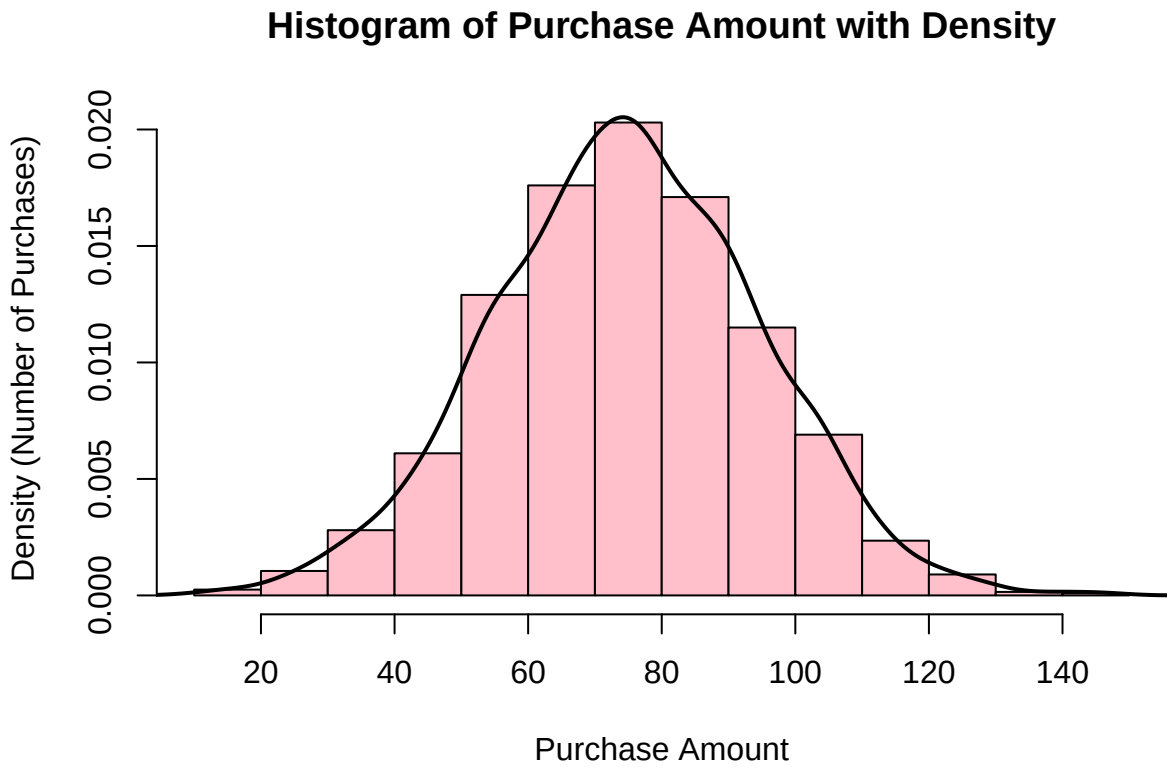
Time spent on site: Min= 1 min, Max= 20.41 mins, Median 9.950, Mean 9.963, Most customers spend between 8 and 12 minutes.(1st Qu - 3rd Qu)

Number of previous purchases: Mean 2.02 purchases, Range= 0 to 9: some customers are repeat buyers, some are first-time.

Used discount: 55.7% of transactions used a discount code. (mean)

3.1.1.b.1 Histogram of purchase_amount with density overlay

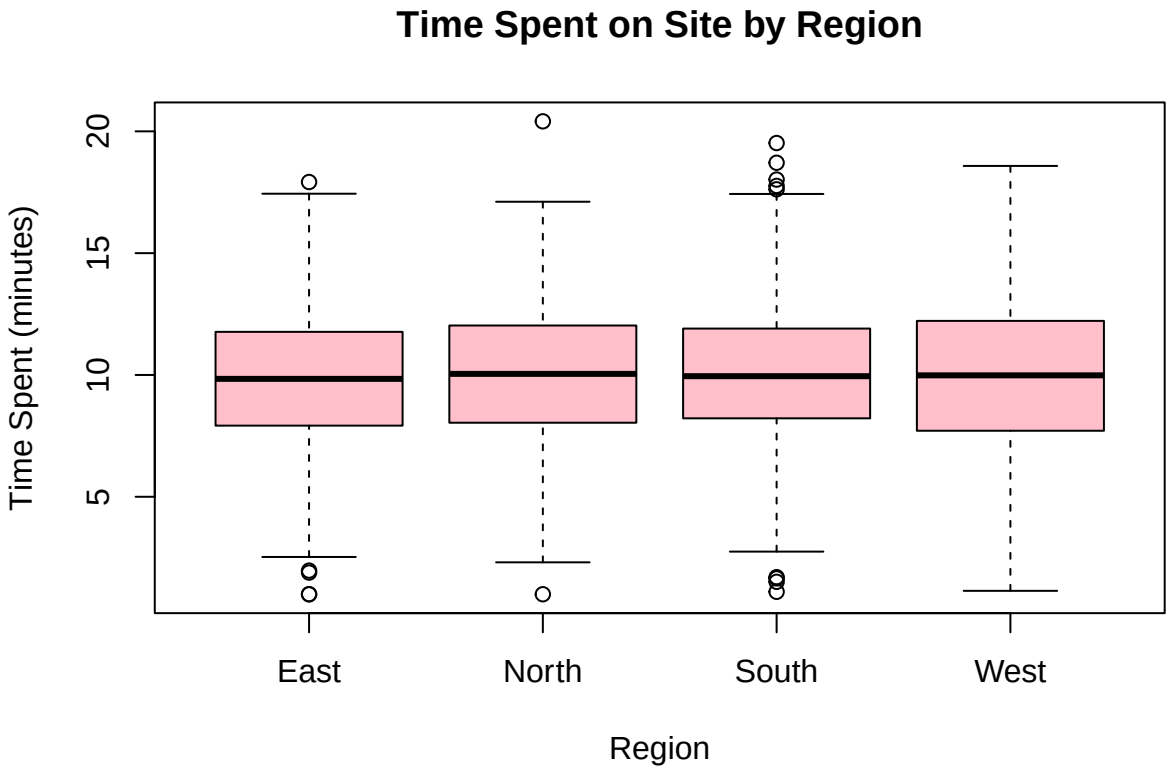
```
hist(df$purchase_amount, probability = TRUE,
     main = "Histogram of Purchase Amount with Density",
     xlab = "Purchase Amount",
     ylab = "Density (Number of Purchases)",
     col = "pink",
     border = "black",
     breaks = 10)
lines(density(df$purchase_amount),col="black",lwd=2)
```



This shows that both low and high spenders are fairly balanced around the average, with most purchases near the center.

3.1.1.b.2 Boxplot of time_spent_on_site by region

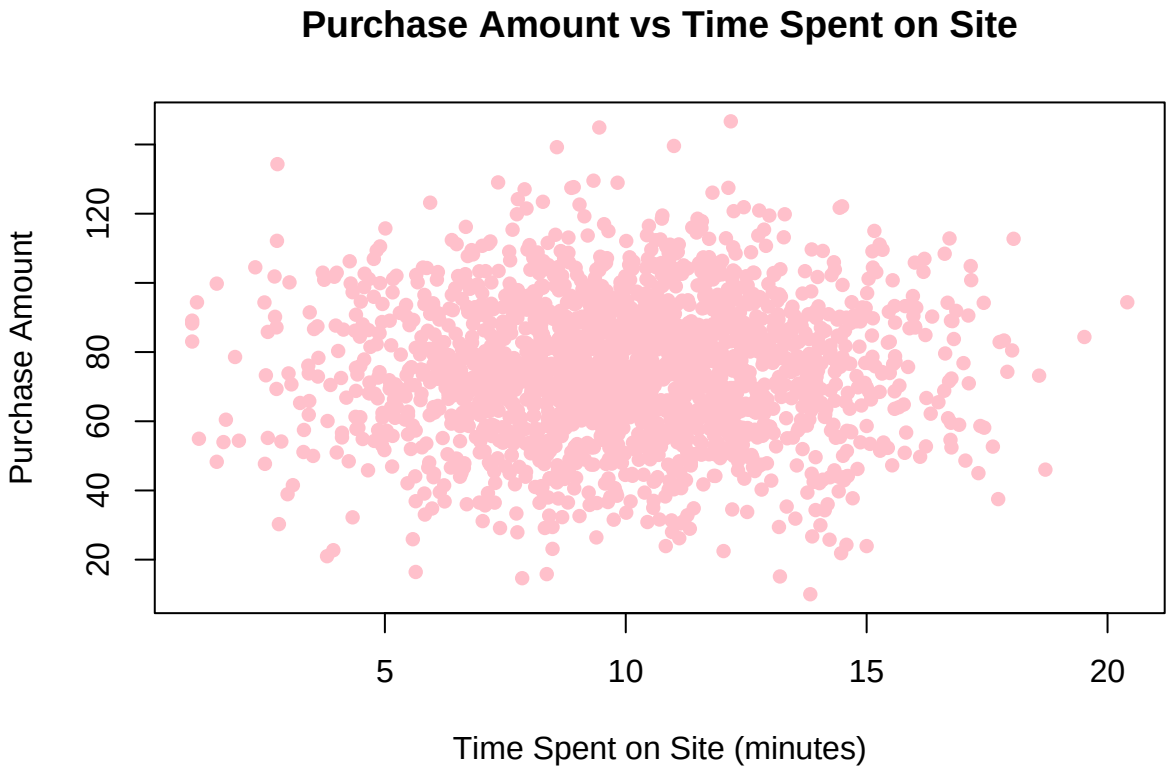
```
boxplot(time_spent_on_site ~ region, data = df,
        main = "Time Spent on Site by Region",
        xlab = "Region",
        ylab = "Time Spent (minutes)",
        col="pink")
```



Some regions are engage longer; but most are around 10 minutes. In south region showers more outliers than the rest.

3.1.1.b.3 Scatterplot of purchase_amount vs time_spent_on_site

```
plot(df$time_spent_on_site, df$purchase_amount,
     main = "Purchase Amount vs Time Spent on Site",
     xlab = "Time Spent on Site (minutes)",
     ylab = "Purchase Amount",
     pch= 16, col="pink")
```



Most customers spend between 6–12 minutes and purchase $60–$90 worth of products.

3.1.1.c Identify and handle any missing values

```
# Count missing per column
# Count missing values in each column
colSums(is.na(df))
```

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

No missing values in any column, Data completeness is perfect meaning no imputation or deletion is needed.

3.1.1.d Outliers

```
flag_outliers_iqr <- function(x) {
  Q1 <- quantile(x, 0.25, na.rm = TRUE)
  Q3 <- quantile(x, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower <- Q1 - 1.5 * IQR
  upper <- Q3 + 1.5 * IQR
```

```
  (x < lower) | (x > upper)
}

summary_outliers <- df %>% #start with my data freme
  summarise(across(where(is.numeric), #selects only numeric columns for each.
                   ~ sum(flag_outliers_iqr(.)), # count of all returns a logical vector of TRUE
                   .names = "{.col}")) %>% #keeps the result columns named the same as the original
  pivot_longer(everything(),
               names_to = "variable",
               values_to = "outlier_count") #Converts that one-row wide summary into long format

summary_outliers
```

```
## # A tibble: 5 x 2
##   variable                     outlier_count
##   <chr>                                <int>
## 1 customer_id                              0
## 2 purchase_amount                         14
## 3 time_spent_on_site                      18
## 4 number_of_previous_purchases            15
## 5 used_discount                            0
```

Detected 14 outliers for purchase_amount, 18 for time_spent_on_site and 15 for number_of_previous_purchases

3.1.2 Interpretation

a. Distribution of key variables

Purchase Amount: Most purchase amounts are between $60 and $90. Based on mean and median values are nearly equal, which might represent a fairly symmetric distribution.

Time Spent on Site: Most customers spend 8–12 minutes browsing. A few outliers spend much longer, which might represent highly engaged customers or those struggling to find what they need.

Number of Previous Purchases: The median is 2 purchases with a maximum of 9,indicating graph slightly skewed to the right, showing a mix of first-time buyers and repeat customers.

Used Discount: Over 55% of customers used a discount code, indicating that promotional offers are influential in driving sales(customer base may be price-sensitive).

b. Unusual patterns

No missing values: the dataset is complete.

Regional Differences: some regions spend more or less time on the site.

3.2 Probability Analysis

3.2.1 Calculations

3.2.1.a.1 P(Purchase > $75)

```
p_purchase_gt_75 <- mean(df$purchase_amount > 75)
p_purchase_gt_75
```

```
## [1] 0.4905
```

49% of customers make purchases above $75, indicating mid and high value buying tendency.

3.2.1.a.2 P(Used Discount | Purchase > $100)

```
p_discount_given_gt_100 <- mean(df$used_discount[df$purchase_amount > 100] == 1)
p_discount_given_gt_100
```

```
## [1] 0.7067308
```

70% spenders use discounts frequently, meaning price sensitivity.

3.2.1.b.1 Region vs Used Discount

```
table(df$region, df$used_discount)
```

```
##
##           0   1
##   East  213 284
##   North 224 294
##   South 233 242
##   West  216 294
```

In all regions, more customers used a discount than didn't, except in the South where usage is nearly equal. The North and West regions show the highest number of discount users. This suggests that discount offers are popular across regions, but slightly less so in the South.

3.2.1.b.2 Previous Purchases vs Discount Usage

```
table(df$number_of_previous_purchases, df$used_discount)
```

```
##
##       0   1
##   0 123 170
##   1 223 280
##   2 248 292
##   3 157 206
##   4  87 110
##   5  28  33
##   6  14  14
##   7   4   6
##   8   2   2
##   9   0   1
```

Most customers have 0–3 previous purchases, and in all these groups, discount usage is higher than non-usage. Discount usage is common across all customer loyalty levels, but the number of customers drops sharply after 4 previous purchases.

3.2.1.c conditional probabilities by region

```
region_discount <- matrix(c(213,284,
                            224,294,
                            233,242,
                            216,294),
                          nrow = 4, byrow = TRUE,
                          dimnames = list(
                            Region = c("East", "North", "South", "West"),
                            Discount = c("No", "Yes")
                          ))

# Conditional probabilities P(Discount | Region)
prop_region <- prop.table(region_discount, margin = 1)
print(prop_region)
```

```
##        Discount
## Region         No       Yes
##   East  0.4285714 0.5714286
##   North 0.4324324 0.5675676
##   South 0.4905263 0.5094737
##   West  0.4235294 0.5764706
```

East: ~57.2% of customers used discounts.

North: ~56.8% used discounts.

South: Only ~50.9% used discounts, almost equal split.

West: ~57.6% used discounts.

3.2.2 Interpretation

a. What probabilities reveal about customer behavior

Most customers use discounts across all regions, with around 57% usage in East, 56%in North, and 57% in West. The South shows a lower discount usage near 51%, suggesting customers there are less likely to use

discounts. In summary, this shows that discounts are generally effective, but marketing efforts in the South might need improvement.

b. Probabilities across different segments

Highest: West (57.6%) - slightly more likely than other regions to use discounts.

Lowest: South (50.9%) - only about half of customers use discounts.

East and North: Both around 57%, showing similar behavior.

Gap: There's about a 6-7 percentage point difference between the highest (West) and lowest (South), indicating a modest but notable variation in discount usage across regions.

c. how marketing could target specific groups

South region - Run targeted ads, local promotions, and short-term deals to raise discount usage.

High purchase history customers - Offer VIP discounts, tiered rewards, and loyalties.

First-time buyers - Give welcome discounts, bundle offers.

High usage regions (East, North, West) - Keep regular promotions, cross-sell with discounts, and link offers to seasonal events.

d. limitations of the probability analysis

It only looks at one factor at a time and can't explain why differences happen. The results might not be accurate if the data is incomplete, if some groups are too small, or if customer behavior changes in the future. Also, we didn't test if small differences are actually important.

3.3 Distribution Fitting

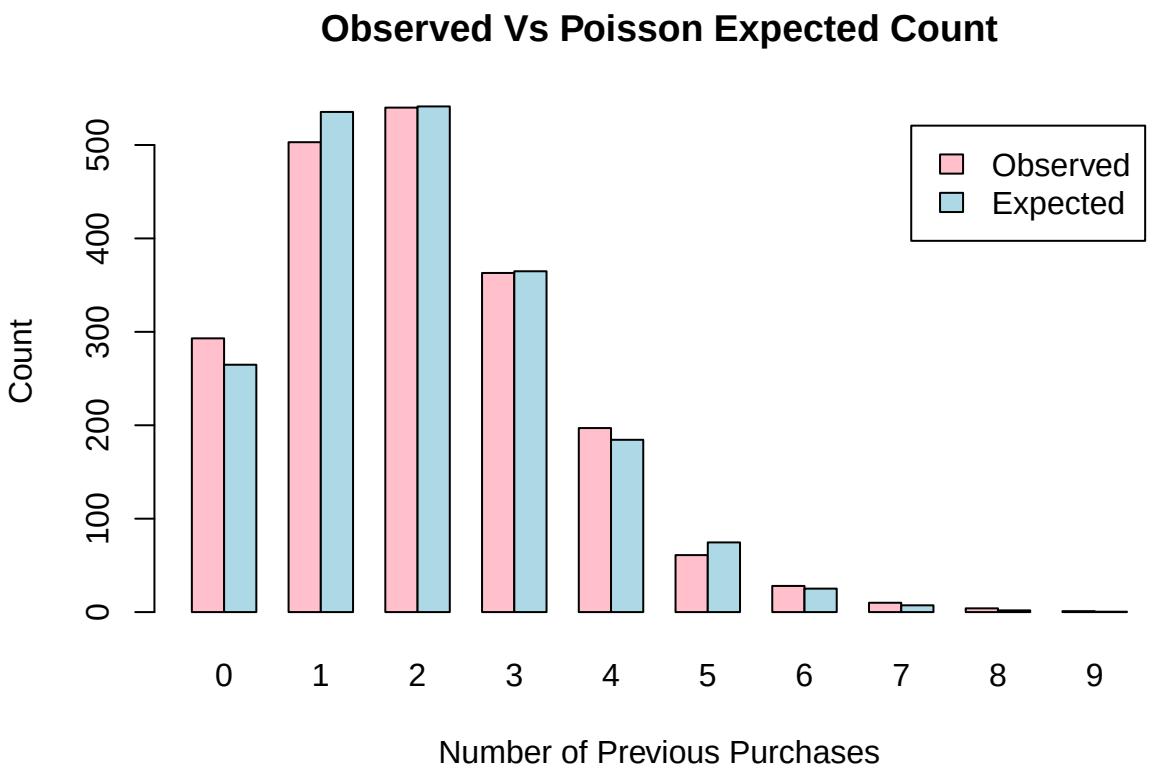
3.3.1 Analysis

3.3.1.a.1 Poisson for number_of_previous_purchases

```
lambda <- mean(df$number_of_previous_purchases)

observed <- table(df$number_of_previous_purchases)
expected <- dpois(as.numeric(names(observed)), lambda) * nrow(df)

barplot(rbind(observed, expected), beside = TRUE,
        col = c("pink", "lightblue"), legend.text = c("Observed", "Expected"),
        main = "Observed Vs Poisson Expected Count",
        xlab = "Number of Previous Purchases",
        ylab = "Count")
```

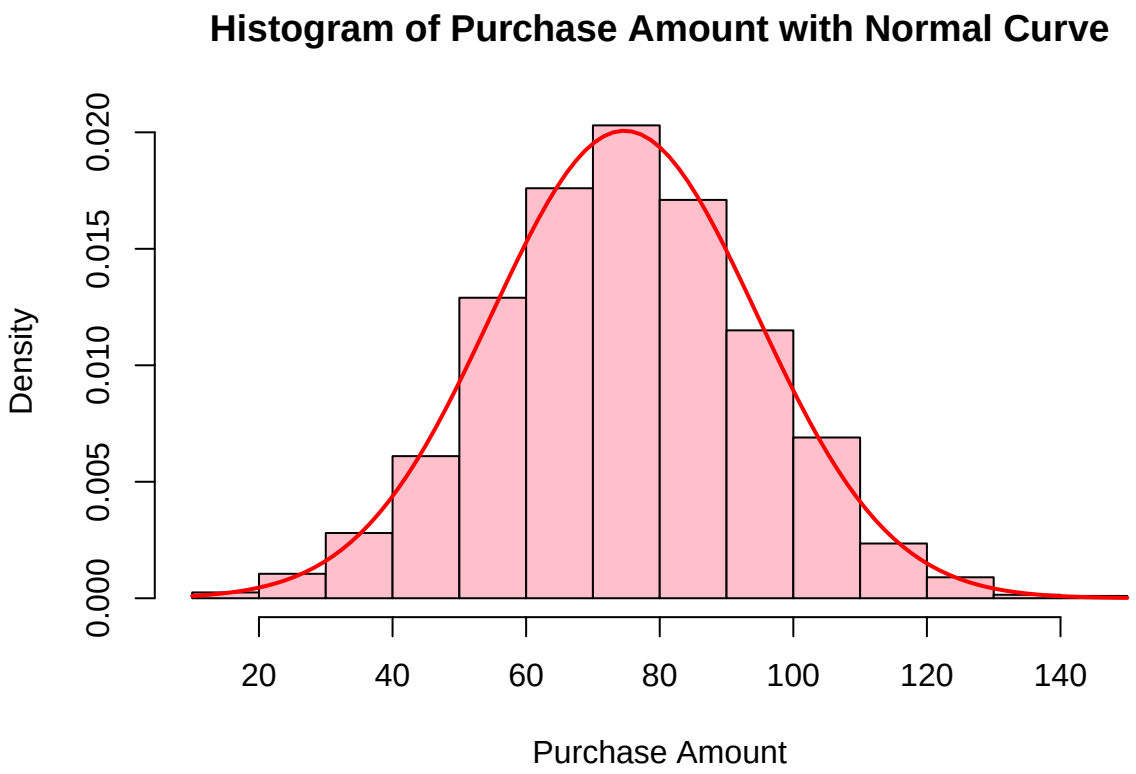


The histogram shows that how well a Poisson distribution with the observed average matches the actual number of previous purchases. Most customers had 1 to 3 past purchases, and very few had more than that.This suggests that repeat purchases are not very high, and most customers are still early in their buying journey.

3.3.1.a.2 Normal for purchase_amount

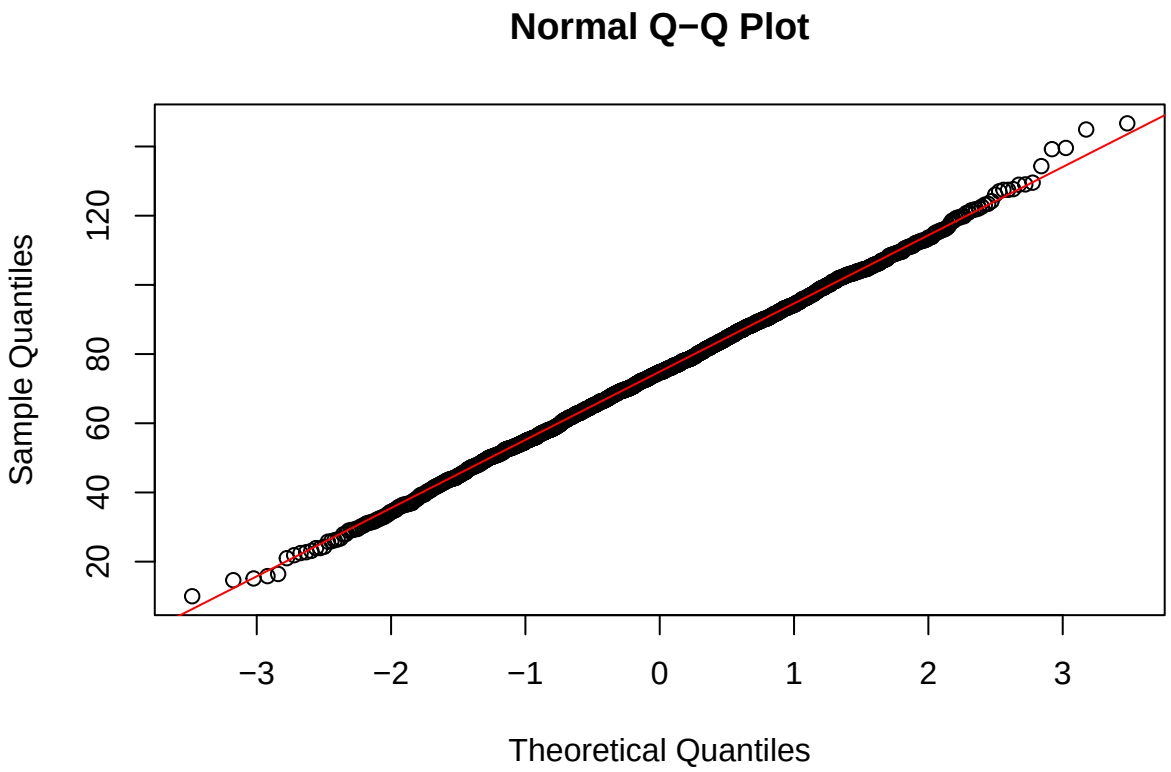
```
hist(df$purchase_amount ,probability = TRUE,
     main = "Histogram of Purchase Amount with Normal Curve",
     col = "pink", xlab = "Purchase Amount")

# Add the normal curve
curve(dnorm(x, mean = mean(df$purchase_amount),
            sd = sd(df$purchase_amount)),
      col = "red", lwd = 2, add = TRUE)
```



The histogram shows that the data follows a fairly symmetrical bell-shaped curve, with most values centered around the mean. This suggests that, customer spending amounts are normally distributed, with most people spending close to the average, and fewer making very small or very large purchases.

```
qqnorm(df$purchase_amount)
qqline(df$purchase_amount, col = "red")
```



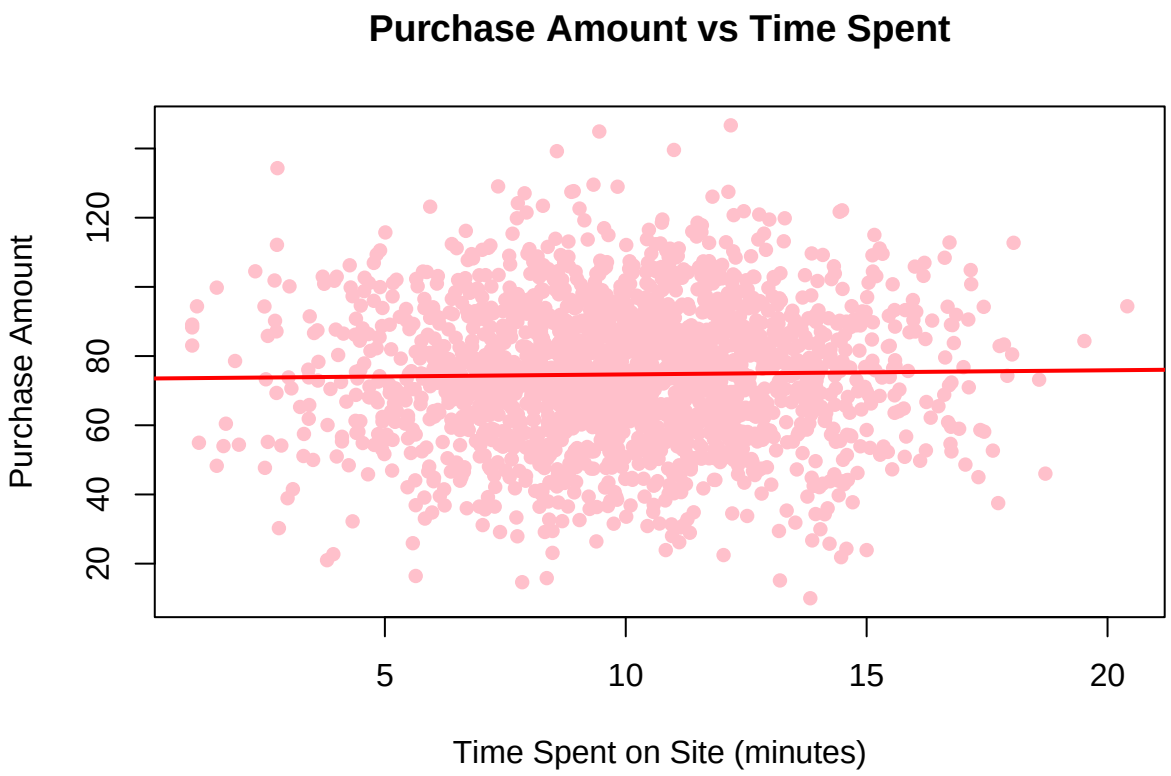
The Q-Q plot also shows that the data are mostly follows the straight line, which means the data is fairly normally distributed, but a few values are different (possibly outliers). Overall, the normal distribution is a suitable model for customer spending behavior in this dataset.

3.4 Predictive Modeling

3.4.a Purchase Amount vs Time Spent

```
plot(df$time_spent_on_site, df$purchase_amount,
     main = "Purchase Amount vs Time Spent",
     xlab = "Time Spent on Site (minutes)",
     ylab = "Purchase Amount",
     col = "pink", pch = 16)

# Add regression line
model <- lm(purchase_amount ~ time_spent_on_site, data = df)
abline(model, col = "red", lwd = 2)
```



There is a positive linear relationship between time spent on the site and purchase amount. This means customers who spend more time tend to make higher purchases.

3.4.b Data about fitted values and regression line over the data

```
model <- lm(purchase_amount ~ time_spent_on_site, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = purchase_amount ~ time_spent_on_site, data = df)
##
## 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
```

```r
coef(model)
```

```
##        (Intercept) time_spent_on_site
##         73.4981130          0.1195997
```

3.4.c Fitted equation

purchase_amount = 73.4981 + 0.1196 * time_spent_on_site

3.4.d Prediction of purchase_amount for a time_spent of 12 minutes

```r
purchase_amount = 73.4981 + 0.1196 * 12
purchase_amount
```

```
## [1] 74.9333
```

This means purchase amount is $74.93 for someone who spends 12 minutes on the site. Which means: for every extra minute on the site, the average purchase goes up by about $0.12; In summery, encouraging users to stay on the site longer could lead to slightly higher purchases.