Statistical Analysis of Zara Sales Data

Helen Zhou, Amy Tang, Emily Xu

I. Abstract

The fast fashion clothing industry operates under short product life cycles and rapid selling cycles. As one of the most well-known brands in the fast fashion clothing industry, Zara has been widely used as an example to investigate its operational efficiency and relatively stable profit margins. We used Zara sales data in 2024, exploring how section, price, and promotion influence sales volume, a key measure of consumer choice. We employed a multiple regression with interaction terms to capture the joint effects among these factors and applied a log-transformation to better satisfy model assumptions. Results show that increasing price can negatively affect the sales volume, but promotion can effectively improve the sales volume. In addition, products in the women 's section are associated with higher sales volume than the men's section. To ensure that the model balances fit and complexity well, we did hypothesis testing and used other model comparison metrics, and concluded that only the interaction effect of price and promotion contributes to explaining more variation, suggesting promotional activities reduce consumer price sensitivity.

II. Introduction

The fast fashion clothing industry has grown rapidly in size in recent years due to its ability to translate changing fashion trends into affordable products for a massive consumers. However, one common challenge for these brands is how to maximize profits while increasing sales, under high product turnover and short time margins. Among all these brands, Zara has been successful, known for its diverse design styles and seasonally driven production. In terms of improving profits, Zara has its advantages of lower product prices, high production efficiency, and low production costs. (Duoyan (2021)). According to Yahoo Finance (Vimal (2025)), Zara had a Gross profit of 6.2% in Q3 2025, which beat its competitors like Gap, which proves the success of its sales strategy.

Being a benchmark in the fashion industry, Zara invested time and resources into understanding consumer behaviors by analyzing real time data and pushing corresponding new inventory in response to build an efficient supply chain and become dominant in the industry, as reported by the research from Harvard Business School (Uberoi (2017)). While existing literature emphasizes Zara's excellent performance and operational strategies, there is comparatively less empirical analysis on consumer purchasing behaviors and choices in Zara. Through exploring consumers' responses to different sales-related factors, Zara and other brands in the industry can better understand the mechanisms that drive sales volume.

Motivated by this gap in the literature, we aim to investigate consumer behavior in Zara by analyzing **how price, promotion status, and gender section affect sales volume**, which is an important factor to represent consumer demand. Specifically, we chose a **multiple regression model with interactive categorical variables** to fit the data set of Zara sales. This model allows us to capture not only the individual effects of price, promotion, and section (gender) on sales volume, but also the interactive effects of multiple variables. This is particularly important in a retail context, as the effectiveness of promotions or pricing strategies may differ across product sections. Then, based on **exploratory data analysis** and diagnostic checks, we **log-transformed** the response variable 'Sales Volume' and the numerical explanatory variable 'Price' and fit the multiple regression again. We then did **hypothesis testing** and computed **model selection metric** to determine whether the terms in the model contribute significantly to explain more variation. These steps lead to a final model that best captures the relationships of interest between the sales-related factors and sales volume.

III. Data Description

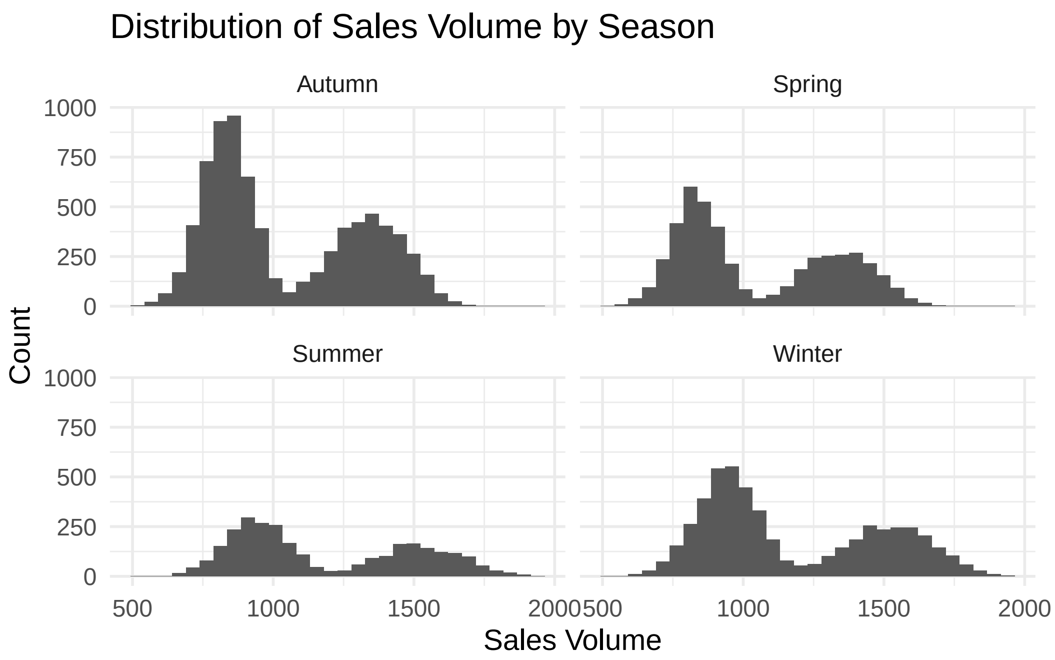
1. Data Selection

```r
library(readr)
library(tidyverse)
Zara_sales_EDA <- read_delim("Zara_sales_EDA.csv",
    delim = ";", escape_double = FALSE, trim_ws = TRUE)
Zara_sales_EDA <- Zara_sales_EDA |> drop_na() # drop missing data
```

The data set we selected is called Zara Sales for EDA, and it was created by combining several public fashion datasets from GitHub and Kaggle. The specific methods of data collection is not mentioned in the data description provided with the dataset. It comprises rich information on price, sales volume, season, gender categories, and promotional status on 20252 Zara products after dropping missing values. To improve the feasibility of data analysis, we filter the data set to get a small subset. Since we're not investigating the effect of variable 'Season' on sales volume, we decide to limit our statistical analysis to one season.

```r
library(ggplot2)

ggplot(Zara_sales_EDA, aes(x = `Sales Volume`)) +
  geom_histogram() +
  facet_wrap(~ season, ncol = 2) +
  labs(title = "Distribution of Sales Volume by Season", x = "Sales Volume", y =
   ↪  "Count") +
  theme_minimal()
```



By the histogram of sales volume for four seasons, from observation, summer has the fewest number of observations, so we filter the data to include only products from summer season with 2,906 observations. Each observation represents a product-level sales record from Zara.

2. Variable Selection

```r
Zara_sales_EDA_summer <- Zara_sales_EDA |>
  filter(season == "Summer")
glimpse(Zara_sales_EDA_summer)
```

```
Rows: 2,906
Columns: 17
$ `Product ID`       <dbl> 182157, 183064, 182306, 131298, 168571, 154224, 154~
$ `Product Position` <chr> "Aisle", "Front of Store", "Front of Store", "End-c~
$ Promotion          <chr> "Yes", "Yes", "No", "Yes", "Yes", "Yes", "Yes", "Ye~
$ `Product Category` <chr> "clothing", "clothing", "clothing", "clothing", "cl~
$ Seasonal           <chr> "Yes", "No", "Yes", "Yes", "No", "Yes", "Yes", "Yes~
$ `Sales Volume`     <dbl> 1711, 1289, 980, 1590, 1529, 1525, 1404, 1501, 1300~
$ brand              <chr> "Zara", "Zara", "Zara", "Zara", "Zara", "Zara", "Za~
$ url                <chr> "https://www.zara.com/us/en/suit-jacket-in-100-line~
$ name               <chr> "SUIT JACKET IN 100% LINEN", "POCKET OVERSHIRT", "C~
$ description        <chr> "Straight fit blazer made of linen. Notched lapel c~
$ price              <dbl> 24.00, 69.95, 33.95, 25.95, 23.95, 28.99, 79.99, 32~
$ currency           <chr> "USD", "USD", "USD", "USD", "USD", "USD", "USD", "U~
$ terms              <chr> "jackets", "jackets", "jackets", "jackets", "jacket~
$ section            <chr> "WOMAN", "WOMAN", "WOMAN", "WOMAN", "MAN", "WOMAN",~
$ season             <chr> "Summer", "Summer", "Summer", "Summer", "Summer", "~
$ material           <chr> "Cotton", "Linen", "Cotton", "Linen", "Linen", "Lin~
$ origin             <chr> "Portugal", "Spain", "Brazil", "Spain", "Turkey", "~
```

From the 2,906 observations, the response variable we chose for this project is **Sales Volume**, representing the approximate number of units sold. The explanatory variable includes **price**, a numerical variable representing the unit price a product is sold, measured in dollars; **Promotion**, a categorical variable representing whether the product was part of a promotion or discount campaign with value of 1 indicating there is a promotion, otherwise 0; **Section**, a categorical variable representing whether the consumer of the product is a man or a women, with value of 1 representing a women, otherwise 0. The choice of variables allow us to investigate the influence of price, promotion and section (gender) on sales volume, satisfying our research aim.

To justify our choices of variables, because of our focus in how Zara uses multiple strategies to tackle the difficulty of maximizing profits under high product turnover and short time margins, we choose sales volume as the response variable since it directly reflects consumer demand and largely decides the earned interest. Among the factors related to sales volume, pricing is the one of the most influential factor of consumer purchasing behaviors. Lower prices reduce the perceived risk of dissatisfaction, increasing consumers' willingness to buy. A second key factor is promotion status. Prior research suggests that promotional activities tend to stimulate buying behaviors of consumers (Familmaleki, Aghighi, and Hamidi (2015)). Finally, we decide to include section (gender) as a variable, as women tend to engage in shopping more frequently than men (Pradhana and Sastiono (2019)). Therefore, price, promotion and section are selected as three representative predictors that are expected to influence sales volume.

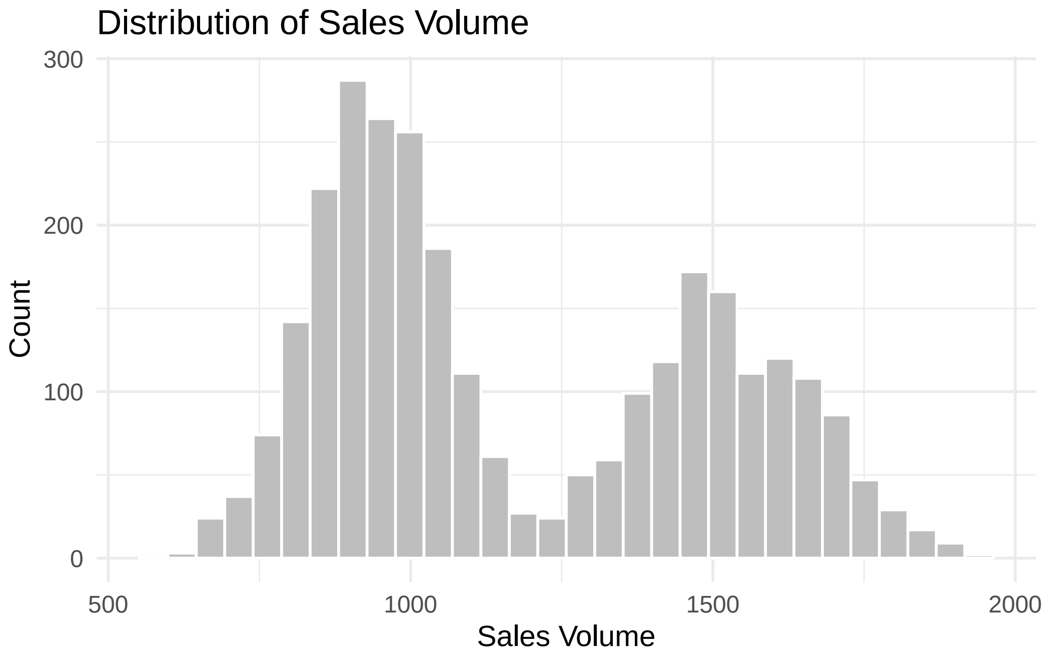
3. Glimpse of response variable

```r
summary_sales_volume_stats <- Zara_sales_EDA_summer|>
  summarise(
    mean_sales_volume = mean(`Sales Volume`),
    median_sales_volume = median(`Sales Volume`),
    min_sales_volume = min(`Sales Volume`),
    max_sales_volume = max(`Sales Volume`),
  )
print(summary_sales_volume_stats)
```

```
# A tibble: 1 x 4
  mean_sales_volume median_sales_volume min_sales_volume max_sales_volume
              <dbl>               <dbl>            <dbl>            <dbl>
1             1185.                1059              575             1940
```

Providing a glimpse to the response variable **sales volume**, the average sales volume in summer season is 1185 units, with the max of 1940 units, min of 575 units and a median of 1059 units.

```
ggplot(Zara_sales_EDA_summer, aes(x = `Sales Volume`)) +
  geom_histogram(fill = "gray", color = "white") +
  labs(title = "Distribution of Sales Volume",
       x = "Sales Volume", y = "Count") +
  theme_minimal()
```



Placing the sales volume on a histogram, there are clearly two peaks in count around sales volume of 800 and around 1500. The remaining values cluster around the two peaks, suggesting that the bimodal distribution may be caused by an underlying categorical variable such as promotion status or product section. that lead to distinct product sales volume.

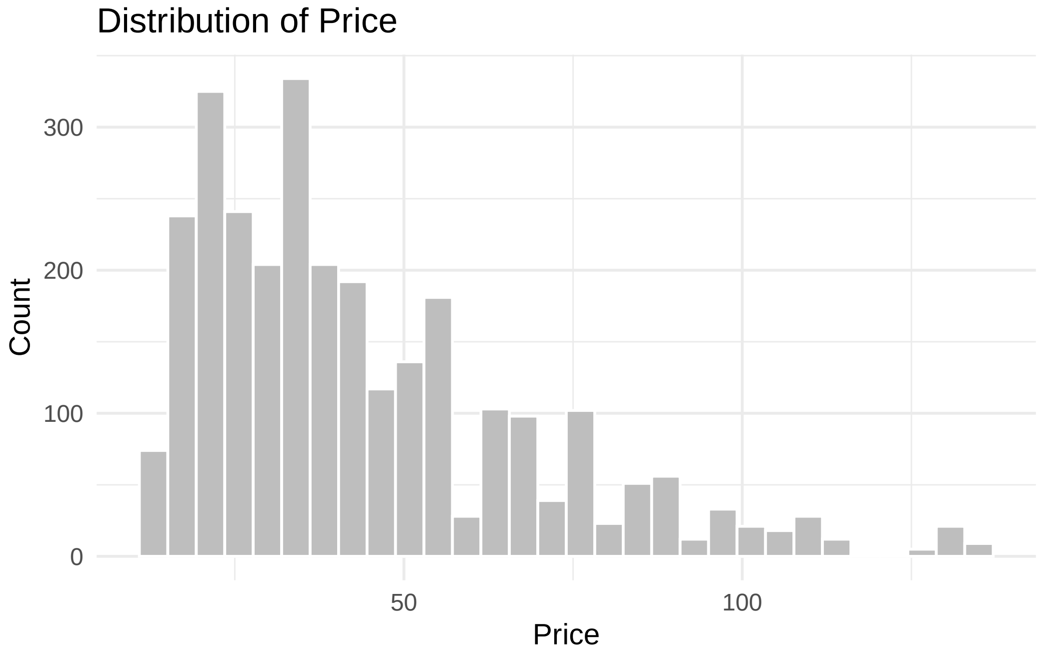
4. Glimpse of explanatory variable

```
summary_price_stats <- Zara_sales_EDA_summer|>
  summarise(
    mean_price = mean(`price`),
    median_price = median(`price`),
    min_price = min(`price`),
    max_price = max(`price`),
  )
print(summary_price_stats)
```

```
# A tibble: 1 x 4
  mean_price median_price min_price max_price
       <dbl>        <dbl>     <dbl>     <dbl>
1       44.1         38.0        13      135.
```

Providing a glimpse for the numerical explanatory variable **price**, it has a mean of 44.1 dollars with the max price at 135 dollars, min price at 13 dollars, and a median of 38 dollars.

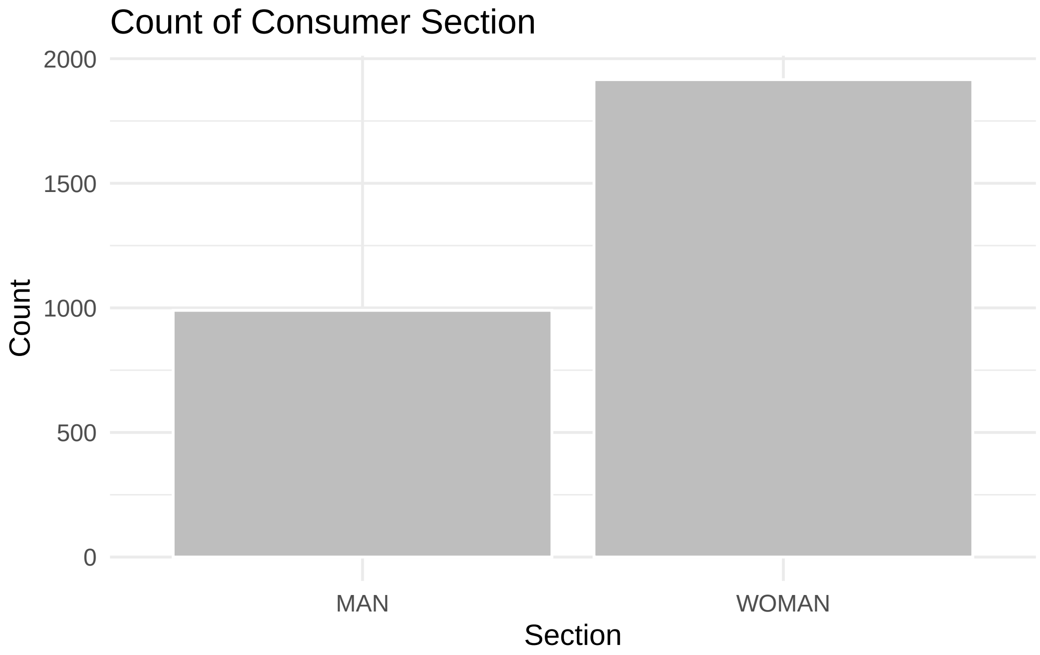
```r
ggplot(Zara_sales_EDA_summer, aes(x = price)) +
  geom_histogram(fill = "gray", color = "white") +
  labs(title = "Distribution of Price",
       x = "Price", y = "Count") +
  theme_minimal()
```



Indicated by the large mean and small median and the histogram above, the distribution for price is right-skewed. The values of price cluster around the peak at around 25 dollars.

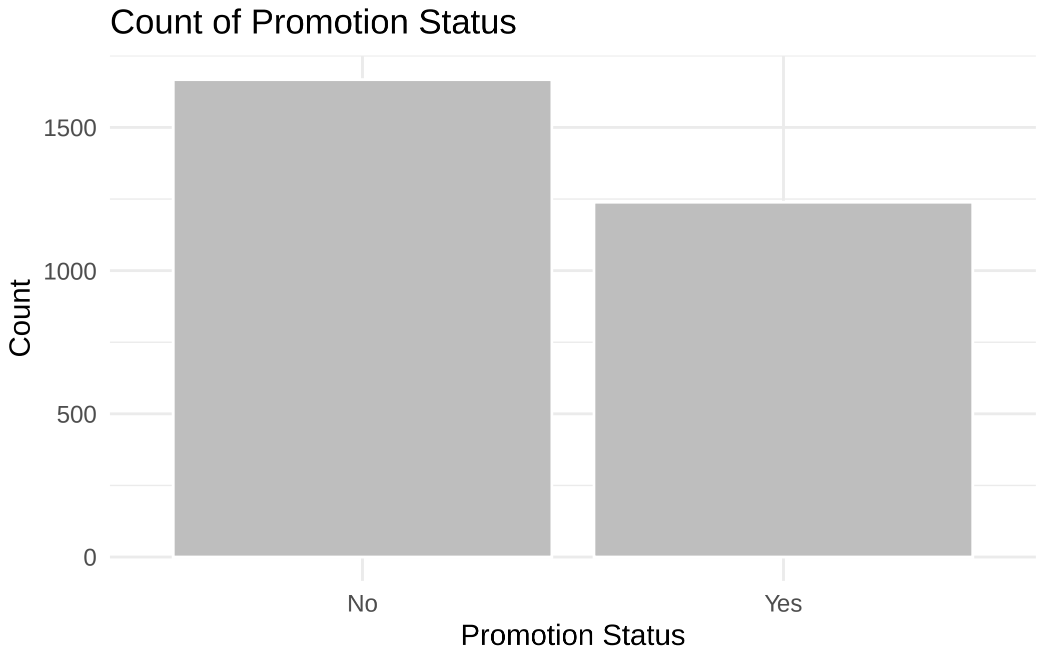
Providing an overview for the explanatory variable **Section**, within the Zara sales in summer data, more women are buying the products compared to men based on the bar plot. The difference is relatively big as women consumers are twice as many as men consumers.

```r
ggplot(Zara_sales_EDA_summer, aes(x = section)) +
  geom_bar(fill = "gray", color = "white") +
  labs(title = "Count of Consumer Section",
       x = "Section",
       y = "Count") +
  theme_minimal()
```



Providing an overview for the explanatory variable **Promotion Status**, within the Zara sales in summer data, more products are not in promotion than products in promotion. The difference in number of products in promotion is around 250.

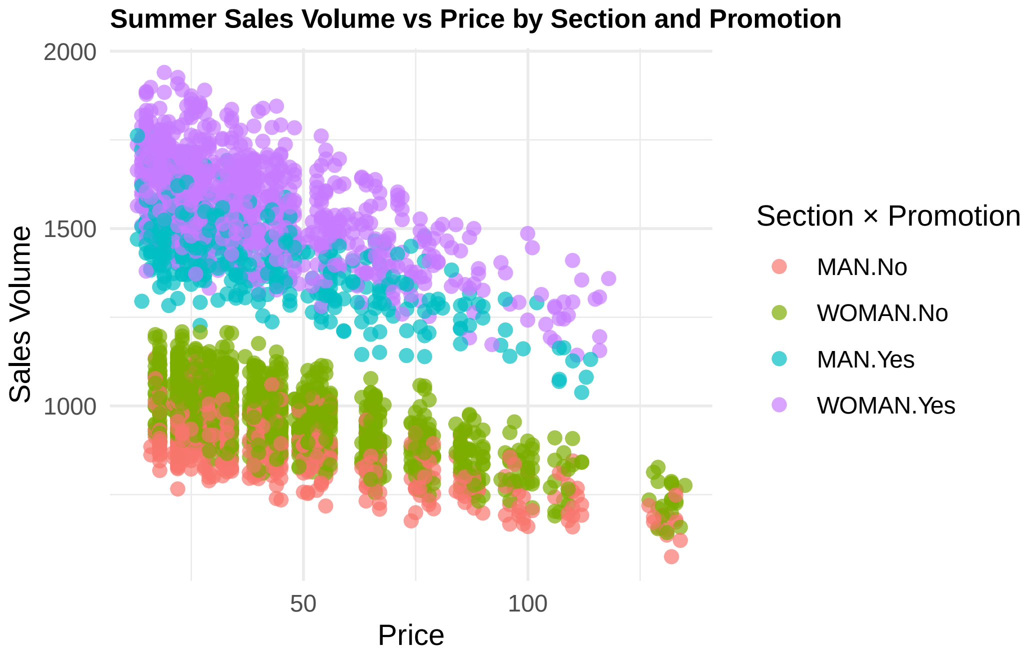
```
ggplot(Zara_sales_EDA_summer, aes(x = Promotion)) +
  geom_bar(fill = "gray", color = "white") +
  labs(title = "Count of Promotion Status",
       x = "Promotion Status",
       y = "Count") +
  theme_minimal()
```



5. Visualization

Visualizing the data set on a scatter plot with different colors representing different categories of the categorical variable in the following plot,

```r
Zara_sales_EDA_summer |>
  ggplot(aes(x = price,
             y = `Sales Volume`,
             color = interaction(section, Promotion))) +
  geom_point(size = 2, alpha = 0.7) +
  labs(
    title = "Summer Sales Volume vs Price by Section and Promotion",
    x = "Price",
    y = "Sales Volume",
    color = "Section × Promotion",
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 10, face = "bold"),
    legend.position = "right")
```



In the scatter plot, looking at the general trend, the purple and blue points together represents the product sales with promotion, which means the categorical variable Promotion has value of one; and the green and red points together represents product sales without promotion, which means the categorical variable Promotion is set to zero. Specifically, for products with promotion, the purple points represent products bought by female consumers and blue points represent products bought by male consumers. For products without promotion, the green points represent products bought by female consumers while the red points represent products bought by male consumers.

From observing the graph, overall speaking, products with promotion usually have higher sales volume compared to those that are not in promotion. Also, for both products with or without promotion, products in the women's section tend to exhibit higher sales volume than those in the men's section.

IV. Method

Since there are one response variable sales volume and several explanatory variables including both numerical variables and categorical variables, a multiple regression with categorical variables will be used. Additionally, as shown by the scatter plot in data description, the effect of price on sales volume may differ by promotion status and product section, motivating the inclusion of interaction terms.Therefore, we choose to include interactive terms in the model. The population form of the multiple regression model is

$$E[SalesVolume_i \mid Price_i, Promotion_i, SectionWomen_i] = \beta_0 + \beta_1 \cdot Price_i + \beta_2 \cdot Promotion_i$$

$$+\beta_3 \cdot SectionWomen_i + \beta_4 \cdot Price_i \cdot Promotion_i$$

$$+\beta_5 \cdot Price_i \cdot SectionWomen_i + \beta_6 \cdot Promotion_i \cdot SectionWomen_i$$

After fitting the data to the multiple regression mode, we will assess whether the linearity, equal variance, normality, and independence assumptions are met for this model in order to ensure that interpretations and inferences are reliable.

Then, based on the conditions assessed, we can decide if we need to transform our model to satisfy the assumptions in order to make valid inferences. Furthermore, to further test the reliability of the statistical model, we plan to use hypothesis testing including T test for single coefficients, F test for nested model, and the model selection metric AIC, BIC, and $Adjusted\ R^2$ to determine possible redundant terms that do not contribute significantly in the regression model. In this way, we will be able to construct a model to answer our research question of how does price, promotion and gender influence the sales volume for Zara.

V. Initial Multiple Regression With Interaction Terms

1. Fitting Multiple Regression Model

Fitting the data set to the multiple regression model to construct a statistical model using the following R output table.

```
zara_sales_summer_lm <- lm(`Sales Volume` ~ price * (Promotion + section) +
 ↪  Promotion*section, data = Zara_sales_EDA_summer)
summary(zara_sales_summer_lm)
```

```

Call:
lm(formula = `Sales Volume` ~ price * (Promotion + section) +
    Promotion * section, data = Zara_sales_EDA_summer)

Residuals:
     Min       1Q   Median       3Q      Max
-290.724  -58.047   -1.975   53.355  294.179

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                992.4118     6.5371 151.813  < 2e-16 ***
price                       -2.4502     0.1154 -21.235  < 2e-16 ***
PromotionYes               582.9427     7.9162  73.639  < 2e-16 ***
sectionWOMAN               103.6421     7.6675  13.517  < 2e-16 ***
price:PromotionYes          -1.3925     0.1344 -10.364  < 2e-16 ***
price:sectionWOMAN          -0.2863     0.1330  -2.153   0.0314 *
PromotionYes:sectionWOMAN   53.4547     6.6971   7.982 2.06e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 83.6 on 2899 degrees of freedom
Multiple R-squared:  0.9301,    Adjusted R-squared:  0.9299
```

```
F-statistic:  6427 on 6 and 2899 DF,  p-value: < 2.2e-16
```

$$E[SalesVolume_i \mid Price_i, Promotion_i, SectionWomen_i] = 992.4118 - 2.4502 \cdot Price_i$$
$$+582.9427 \cdot Promotion_i + 103.6421 \cdot SectionWomen_i - 1.3925 \cdot Price_i \cdot Promotion_i$$
$$-0.2863 \cdot Price_i \cdot SectionWomen_i + 53.4547 \cdot Promotion_i \cdot SectionWomen_i$$

where the reference level for this model is the average estimated sales volume for male without promotion;

$SalesVolume_i$ is the predicted average number of units sold for a product;

$Price_i$ is the sales price, measured in dollar;

$Promotion_i$ is a dummy variable, if it is equals to 1 then there is a promotion for the product promoted, 0 otherwise;

$SectionWomen_i$ is a dummy variable, if it is equals to 1 the product belongs to the women's section, 0 belongs to men's section;

$\beta_0 = 992.4118$ represents the predicted average sales volume for non-promoted men's products when the sales price is zero (reference level) is 992.4118 units;

$\beta_1 = -2.4502$ measures the predicted change in average sales volume for a dollar increase in price non promoted men‘s product, and the sales volume is expected to decrease by 2.4502 units for every dollar increase in price, holding other variables constant;

$\beta_2 = 582.9427$ measures the difference in predicted average sales volume between promoted and non-promoted products, holding other variables constant, and the product with promotion are expected to sell on average 582.9427 units more than the product without promotion, holding section constant;

$\beta_3 = 103.6421$ measures the difference in average predicted sales volume between women's and men's product, holding other variables constant, and women's product are expected sell 103.6421 more units on average than men's products, holding promotion status constant;

$\beta_4 = -1.3925$ measures the difference in average change in sales volume for promoted compared to non-promoted products for each dollar increase in price, holding section constant, and for promoted product, each dollar increase in price would lead to decrease in sales volume by on average 1.3925 units less than the decrease in sales volume for the product without promotion, holding section constant;

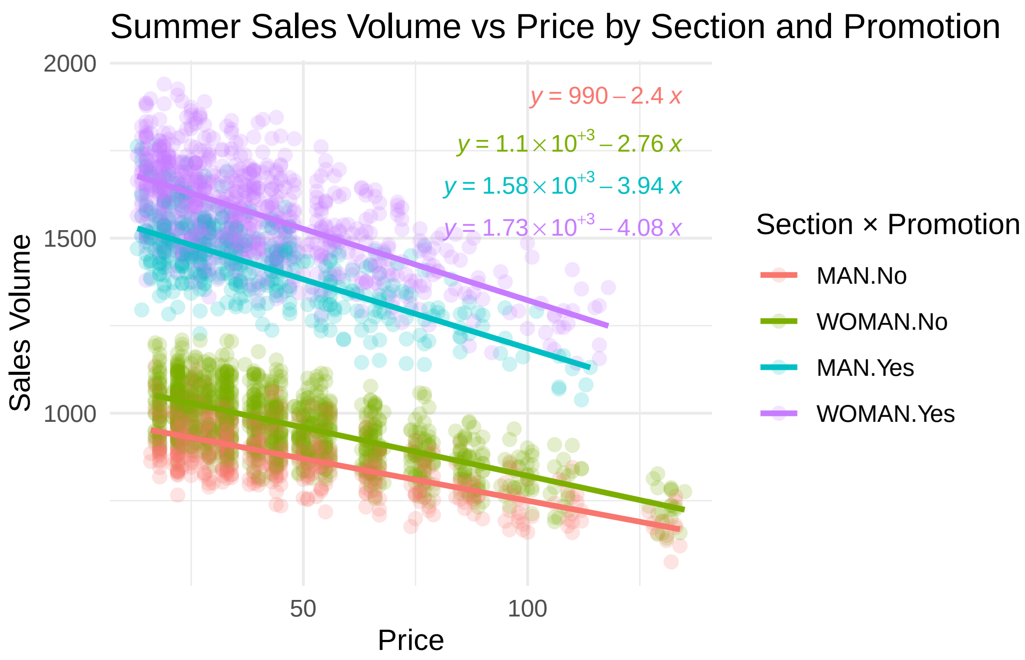
$\beta_5 = -0.2863$ measures the difference in average change in sales volume for products for women compared to products for men for each dollar increase in price, holding other variables constant, and one additional dollar increase in sales price will lead to decrease in sales volume for women consumer of 0.2863 units less than if the buyer is men, holding promotion status constant;

$\beta_6 = 53.4547$ measures the additional change in sales volume is the buyer is women and there is promotion by an average of 53.4547 units.

2. Visualizing

Visualizing the fitted line, the regression line under four possible conditions with different levels of categorical variable is shown below.

```r
library(ggpmisc)
Zara_sales_EDA_summer |>
  ggplot(aes(
    x = price,
    y = `Sales Volume`,
    color = interaction(section, Promotion)
  )) +
  geom_point(size = 2, alpha = 0.2) +
  geom_smooth(method = "lm", se = FALSE, linewidth = 1) +
  stat_poly_eq(
    aes(
      label = paste(..eq.label.., sep = "~~~"),
      group = interaction(section, Promotion)
    ),
    formula = y ~ x,
    parse = TRUE,
    label.x = "right",
    label.y = "top",
    vstep = 0.08,
    size = 3
  ) +
  labs(
    title = "Summer Sales Volume vs Price by Section and Promotion",
    x = "Price",
    y = "Sales Volume",
    color = "Section × Promotion"
  ) +
  theme_minimal()
```

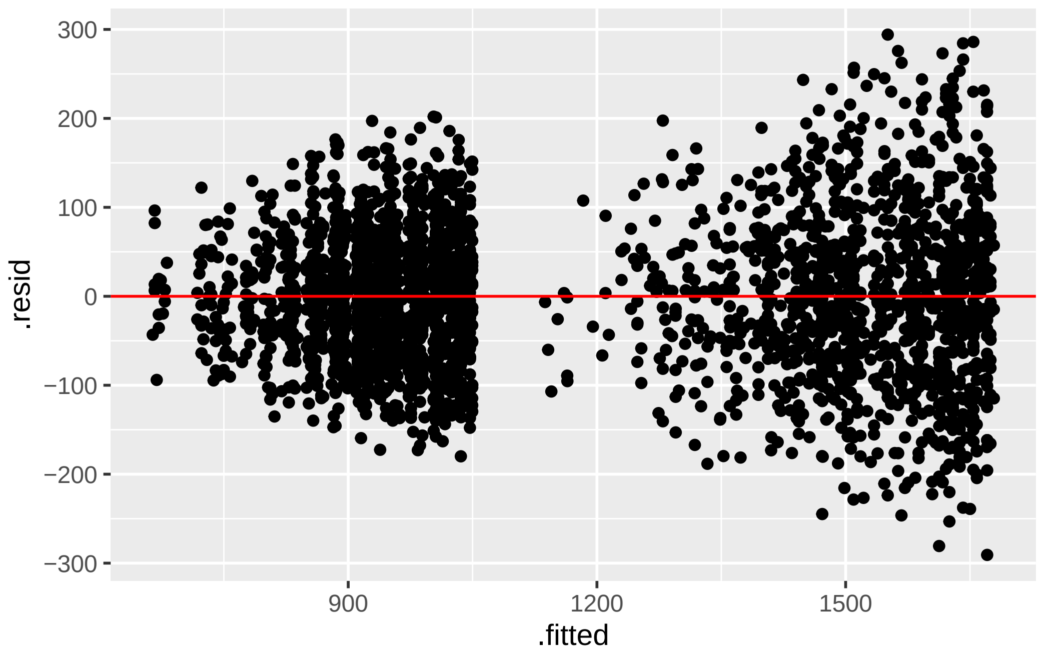


3. Assessing Condition with Multiple Regression Model

Checking Independence

The independence condition is generally satisfied because each observation in the dataset represents a distinct product-level sales record. Specifically, the measurements are collected from separate individual consumers, ensuring that the observations do not interact with each other, satisfying the independence condition.

```
library(broom)
zara_sales_summer_lm |> augment() |> ggplot(aes(x = .fitted, y = .resid)) +
↪  geom_point()+geom_hline(yintercept = 0, col = "red")
```



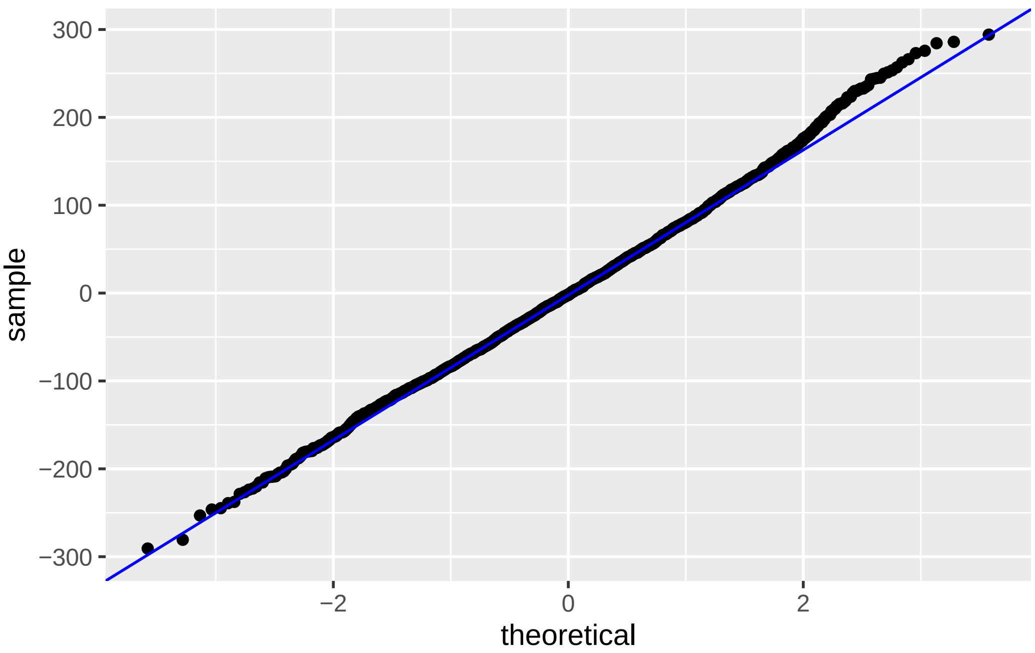
Checking Linearity

The above residual versus fitted plot shows two distinct clusters. The pattern is because both section and promotion are binary variables, resulting in group-based mean structure in the fitted values rather than a continuous spread. In other words, the large positive effects of promotion and section shift the fitted values upward substantial, causing fitted value shift into two groups. From the residual-fitted plot, there are no systematic relationship between fitted values and their residuals and centers around zero, which means the linearity condition is satisfied.

Checking Equal Variance

Also from the residual-fitted graph, the spread of residual is noticeably higher for higher fitted value, which means variance of the residuals may increase with fitted values, showing the violation to equal variance condition.

```
zara_sales_summer_lm |> augment() |> ggplot(aes(sample = .resid)) +
↪  geom_qq()+geom_qq_line(col = "blue") + xlab("theoretical") + ylab("sample")
```



Checking Normality

As we can see from the QQ plot, the points mostly cluster tightly along the reference line, with only minor deviations appear at the tail. Therefore, the residuals are approximately normally distributed with mean equals to zero, satisfying the normality condition.

VI. Transforming the Multiple Regression Model

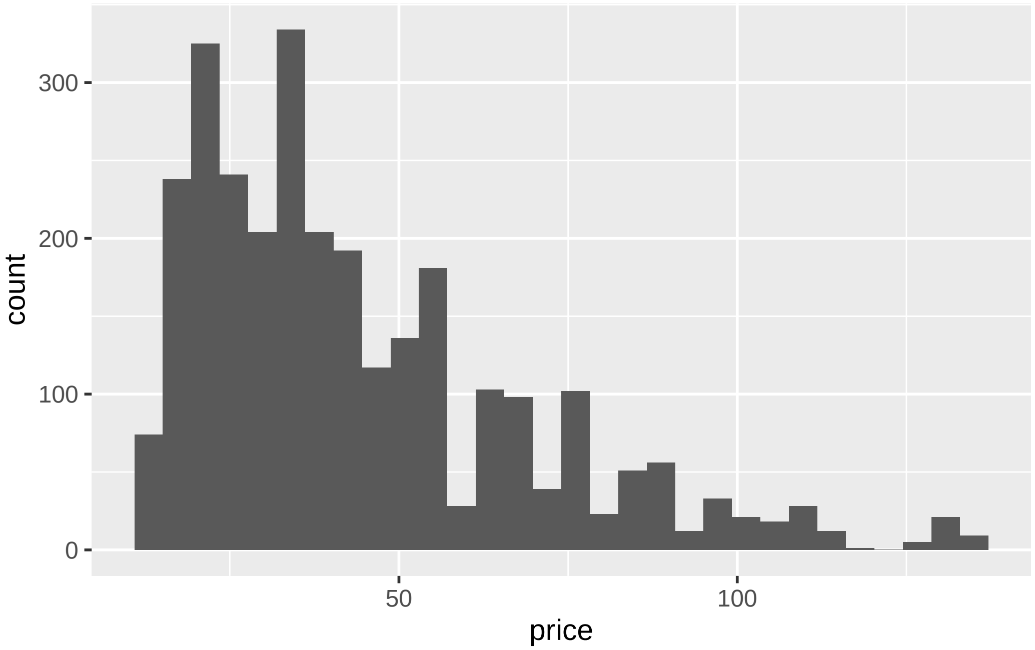
By assessing the conditions of the multiple regression model, we found violation of equal variance condition. Therefore, we will transform to original model to satisfy the conditions in order to make valid interpretation and inferences.

1. Exploratory variable analysis

Exploratory variable analysis was conducted to investigate which variable we should transform. Specifically, we plot the histograms for numerical variables price and sales volume to determine which variables needs transformation.

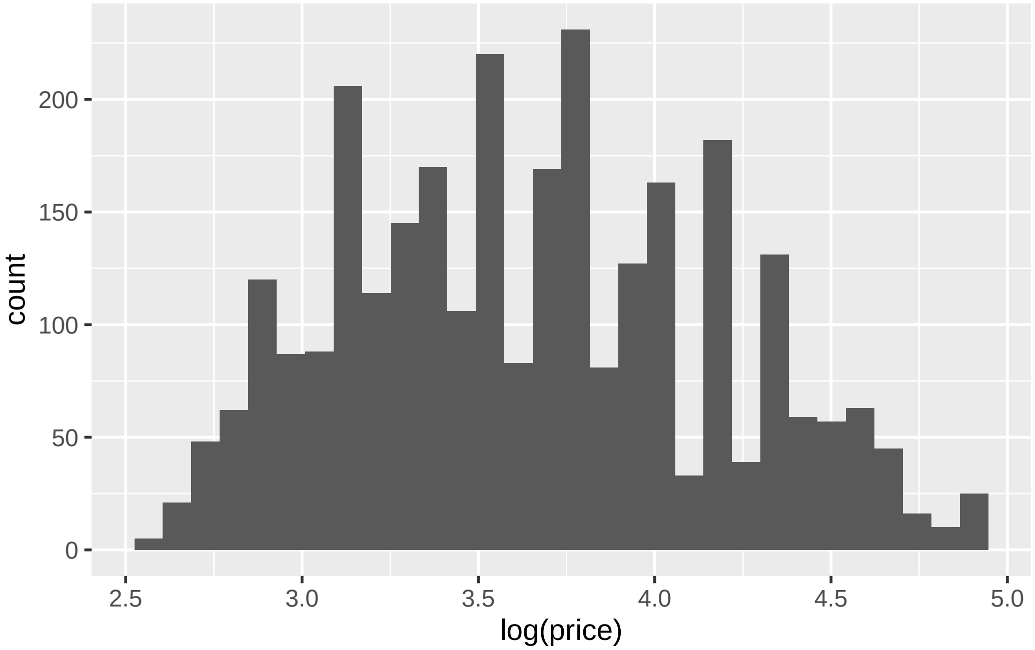
The following histogram shows the distribution for price variable. As shown, the distribution is highly right skewed, so we have to do a log-transformation to address this problem.

```
Zara_sales_EDA_summer |> ggplot(aes(x = price)) + geom_histogram()
```



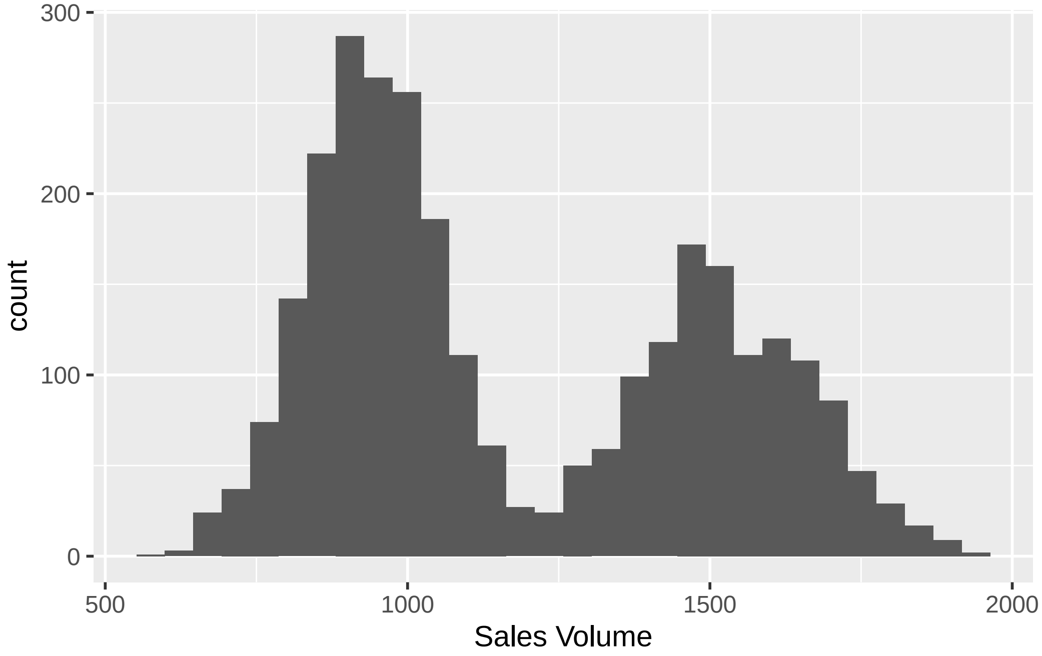
The following histogram shows the log-transformed price distribution, which is much more symmetrical and may meet the assumptions for further analysis.

```r
Zara_sales_EDA_summer |> ggplot(aes(x = log(price))) + geom_histogram()
```



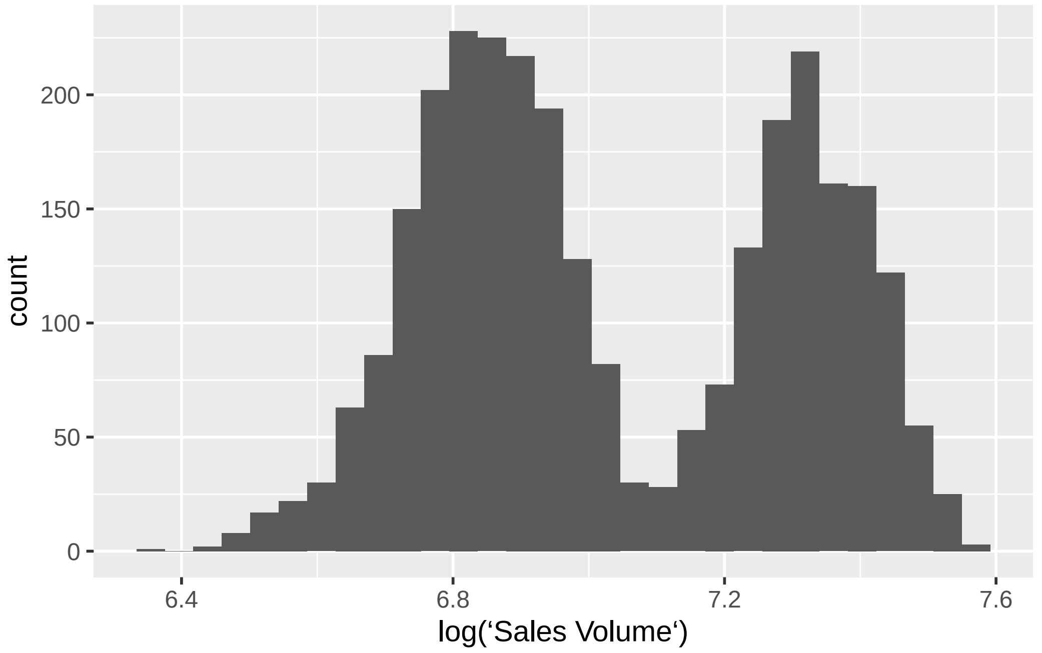
The following histogram shows the distribution for sales volume. As show, the distribution is right-skewed as well, so we need to log-transform to address this problem that may lead to violation of assumptions.

```
Zara_sales_EDA_summer |> ggplot(aes(x = `Sales Volume`)) + geom_histogram()
```



The log-transformed sales volume distribution is much more symmetrical and suitable for constructing linear regression model.

```
Zara_sales_EDA_summer |> ggplot(aes(x = log(`Sales Volume`))) + geom_histogram()
```



2. Log-Transforming the Model

Based on above analysis, we transformed both sales volume and price variable and fit the regression model again.

```r
log_zara_sales_summer_lm <- lm(log(`Sales Volume`) ~ log(price)* (Promotion +
 ↪  section) + Promotion*section, data = Zara_sales_EDA_summer)
summary(log_zara_sales_summer_lm)
```

```

Call:
lm(formula = log(`Sales Volume`) ~ log(price) * (Promotion +
    section) + Promotion * section, data = Zara_sales_EDA_summer)

Residuals:
     Min       1Q   Median       3Q      Max
-0.24982 -0.05125  0.00203  0.05128  0.20644

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                7.298940   0.018908 386.031  < 2e-16 ***
log(price)                -0.142288   0.005008 -28.411  < 2e-16 ***
PromotionYes               0.354860   0.019986  17.756  < 2e-16 ***
sectionWOMAN               0.090695   0.021007   4.317 1.63e-05 ***
log(price):PromotionYes    0.030201   0.005345   5.650 1.76e-08 ***
log(price):sectionWOMAN    0.002335   0.005546   0.421    0.674
PromotionYes:sectionWOMAN -0.001767   0.005919  -0.299    0.765
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07375 on 2899 degrees of freedom
Multiple R-squared:  0.9226,    Adjusted R-squared:  0.9224
F-statistic:  5756 on 6 and 2899 DF,  p-value: < 2.2e-16
```

The new fitted model can be expressed as:

$$E[log(SalesVolume_i) \mid log(Price_i), Promotion_i, SectionWomen_i] = 7.298940 - 0.142288 \cdot log(Price_i)$$

$$+0.354860 \cdot Promotion_i + 0.090695 \cdot SectionWomen_i + 0.030201 \cdot log(Price_i) \cdot Promotion_i$$

$$0.002335 \cdot log(Price_i) \cdot SectionWomen_i - 0.001767 \cdot Promotion_i \cdot SectionWomen_i$$

where the reference level for this model is the average estimated sales volume for male without promotion;

$log(SalesVolume_i)$ is the log-transformed predicted average number of units sold for a product;

$log(Price_i)$ is the log-transformed sales price, measured in log-transformed dollars;

$Promotion_i$ is a dummy variable, if it is equals to 1 then there is a promotion for the product promoted, 0 otherwise;

$SectionWomen_i$ is a dummy variable, if it is equals to 1 the product belongs to the women's section, 0 belongs to men's section;

$\beta_0 = 7.298940$ measures the log-transformed value of sales volume when price is zero at the reference level. The median sales volume when price is equal to zero is $e^{7.298940} = 1478.732$ units.

$\beta_1 = -0.142288$ measures the log-transformed slope. A two-fold multiplicative change in price is associated with a multiplicative change of $2^{-0.142288} = 0.906081$ in median sales volume, holding other variables constant.

$\beta_2 = 0.354860$ means that when price is set equal to zero, the median sales volume for promoted items is $e^{0.354860} = 1.425981$ times the median sales volume for non-promoted items, holding section constant.

$\beta_3 = 0.090695$ means that when price is set to zero, the median sales volume by women is $e^{0.090695} = 1.094935$ times the median sales volume for men, holding promotion status constant.

$\beta_4 = 0.030201$ means that a two-fold multiplicative change in price is associated with a increase in median of sales volume in promoted products of $2^{0.030201} = 1.021154$ times higher than the increase in median sales volume in non-promoted products, holding section constant.

$\beta_5 = 0.002335$ means that a two-fold multiplicative change in price is associated with a increase in median of sales volume by women of $2^{0.002335} = 1.00162$ times higher than the increase in median sales volume by men, holding promotion status constant.

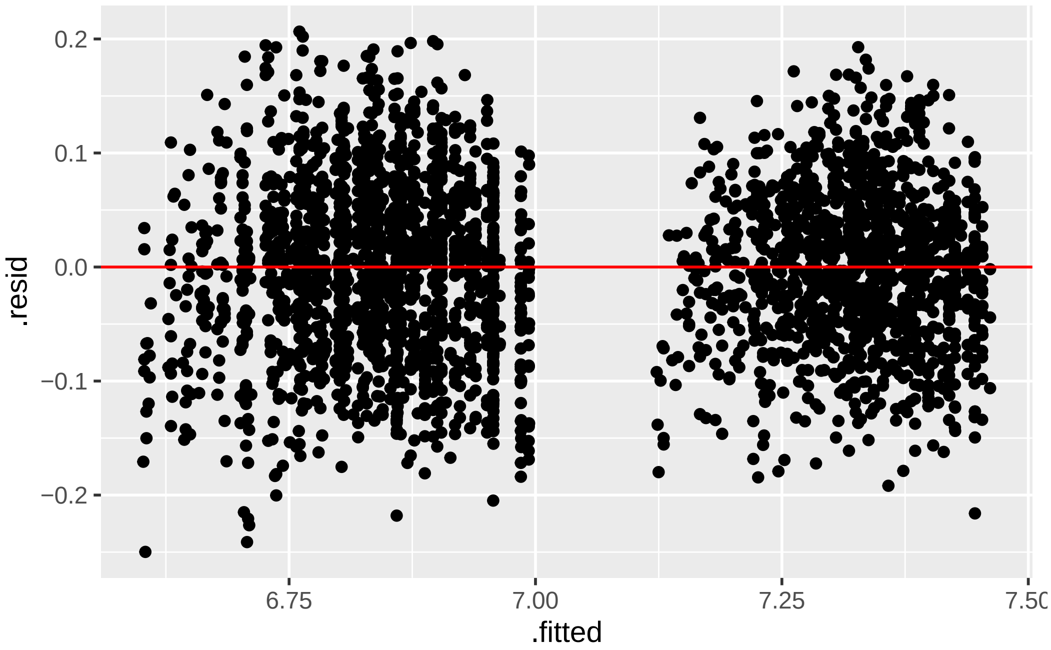
$\beta_6 = -0.001767$ measures the additional multiplicative change in median sales volume of $e^{-0.001767} = 0.9982346$ if there is promotion and the purchase is by women, setting price equal to zero.

3. Assessing Conditions of the Transformed Model

Checking Independence

Independence is still met after the log transform. Each row of data represents a distinct product-level sales record in the summer section, so the random errors from each observation are independent and identically distributed. Therefore, independence condition is met.

```r
library(broom)
log_zara_sales_summer_lm |> augment() |> ggplot(aes(x = .fitted, y = .resid)) +
↪   geom_point()+geom_hline(yintercept = 0, col = "red")
```



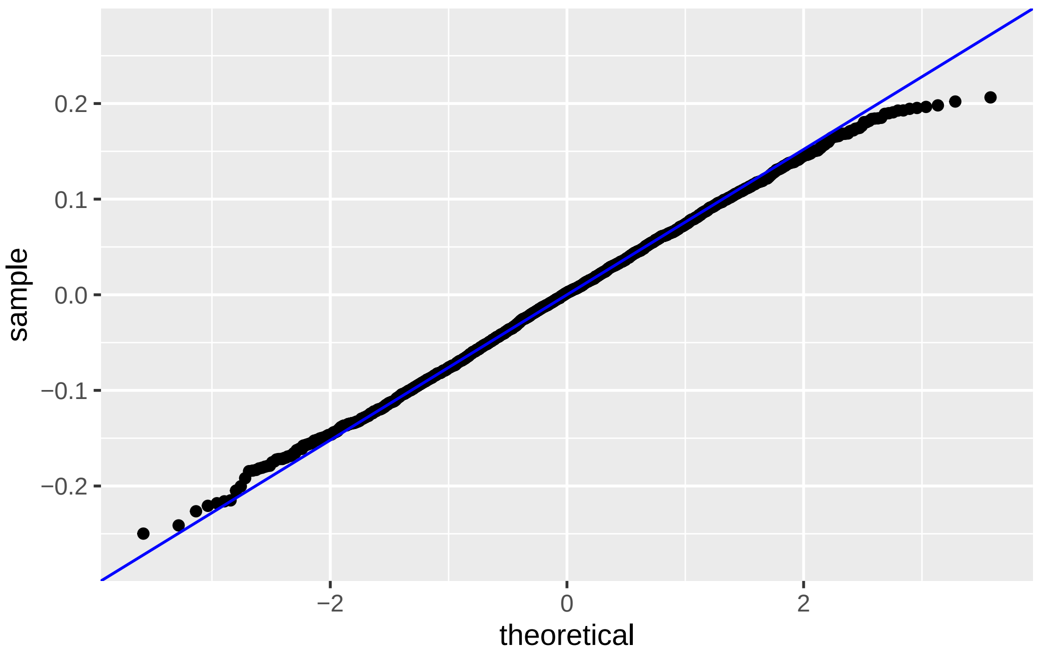
Checking Linearity

The residuals-versus-fitted plot for the log-transformed model still shows two distinct clusters of fitted values. The data forms two clusters because the binary variables promotion status and section shifts the general sales volume greatly. In each cluster, the residuals are concentrated around zero with no curved or systematic pattern as the value increases. This suggests that linearity condition is reasonably met.

Checking Equal Variance

From the residual versus fitted plot, equal variance can be verified by checking the vertical spread of residuals across the fitted values. Both clusters show similar vertical disparity around zero without a funnel shape of systematic increase. As a result, after the log transform, the variability of the residuals is constant in both cluster without increasing patter, showing that equal variance is now satisfied.

```
log_zara_sales_summer_lm |> augment() |> ggplot(aes(sample = .resid)) +
↪  geom_qq()+geom_qq_line(col = "blue") + xlab("theoretical") + ylab("sample")
```



Checking Normality

Q–Q plot of the residuals helps to assess the normality condition. In the plot, the points are closely aligned with the reference line, suggesting that residuals are approximately normally distributed around mean zero. Slight deviation at the end of the tail would not influence normality substantially because the large sample size in data set satisfies central limit theorem. As a result, the normality condition is still satisfied for the log-transformed model.

Since the transformed model satisfies the assumptions of linearity, independence, normality and equal variance for multiple linear regression, we are capable of doing further valid statistical analysis and inferences.

4. Hypothesis Testing

The log transformed fitted multiple regression model shows how $log(SalesVolume_i)$ is related to $log(Price_i)$, $Promotion_i$, $SectionWomen_i$. Through hypothesis testing, we aim to test:

- whether there is statistically significant relationship between sales volume and price
- whether there is statistically significant relationship between sales volume and promotion
- whether there is statistically significant relationship between sales volume and gender section
- whether the interactive terms contribute meaningfully to the model

Specifically, we will first conduct t-tests on coefficients of non-interactive terms of the log-transformed model. Then, we will apply nested F-test to determine whether the interactive terms contribute meaningfully to the model. Conditional on a significant F-test, we further assess the statistical significance of individual interaction coefficients.

Additionally, since the log transformation is monotonic, inference conducted on the transformed model remains valid for assessing relationships in the original scale. Thus, we can do hypothesis testing with the transformed model that satisfies the assumptions and then directly apply the results of hypothesis testing in the context of the original model.

Testing A: The effect of price on sales volume

Controlling for promotion and gender section, the equation can be simplify into

$$E[log(SalesVolume_i) \mid log(Price_i), Promotion_i = 0, SectionWomen_i = 0]$$

$$= 7.298940 - 0.142288 \cdot log(Price_i)$$

Null hypothesis: Holding promotion and section constant, $log(Price_i)$ is no statistically significant associate with $log(SalesVolume_i)$.

$$H_0 : \beta_1 = 0$$

Alternative Hypothesis: Holding promotion and section constant, there is a statistically significant relationship between $log(Price_i)$ and $log(SalesVolume_i)$

$$H_A : \beta_1 \neq 0$$

```r
coef_table <- summary(log_zara_sales_summer_lm)$coefficients
coef_table["log(price)", ]
```

```
     Estimate     Std. Error       t value       Pr(>|t|)
-1.422883e-01   5.008157e-03  -2.841132e+01  7.364593e-157
```

The output R table shows that the test statistic t value is $t = -28.41$, meaning that the observed statistics is -28.41 standard errors away from zero. The resulting p-value is 7.36×10^{-157}, meaning that the probability of observing a t-value that is as extreme as the test statistics is 7.36×10^{-157}.

Using the two sided t-test with $\alpha = 0.05$ significance level the critical values are approximately ± 1.96. With t value larger than the critical value and p value smaller than the significant level, we have enough statistical evidence to reject the null hypothesis. Therefore, we conclude that holding other variable constant, there is a statistically significant relationship between $log(Price)$ and $log(SalesVolume)$ among non-promoted men's products. Transferring the conclusion back to the original scale, we conclude that price is significantly associated with sales volume.

From the log-log model coefficient,$\beta_1 = -0.142288$ shows a price elasticity that 1% increase in price is associated with an estimated 0.142% decrease in expected sales volume on average while holding other factors constant.

Testing B: The effect of promotion on sales volume

This hypothesis aims to determine the whether promotion is associated with statistically significant change in $log(SalesVolume_i)$, controlling for gender section at a representative price.

In the model, there is a interaction term between $log(price)$ and promotion. This means that promotion effect on sales volume depend on price. To make this hypothesis effective in determine the individual effect of promotion on sales volume, we center $log(Price)$ using this equation.

$$X_i = log(Price_i) - log(Price)$$

After defining this, $X_i = 0$ would be presenting a product's typical price. The equation can be simplify into:

$$E[log(SalesVolume_i) \mid X_i = 0, SectionWomen_i = 0]$$

$$= 7.298940 + 0.354860 \cdot Promotion_i$$

Null Hypothesis: Holding price at centered value and section constant, there is no statistically significant association between $Promotion_i$ and $log(SalesVolume_i)$

$$H_0 : \beta_2 = 0$$

Alternative Hypothesis: Holding price at centered value and section constant, there is statistically significant association between $Promotion_i$ and $log(SalesVolume_i)$

$$H_A : \beta_2 \neq 0$$

```
Zara_sales_EDA_summer <- Zara_sales_EDA_summer |>
  mutate(log_price = log(price))

xbar <- mean(Zara_sales_EDA_summer$log_price, na.rm = TRUE)

Zara_sales_EDA_summer <- Zara_sales_EDA_summer |>
  mutate(log_price_c = log_price - xbar)

log_zara_sales_summer_lm_c <- lm(
  log(`Sales Volume`) ~ log_price_c * (Promotion + section) + Promotion *
   ↪  section,
  data = Zara_sales_EDA_summer
)

coef_table <- summary(log_zara_sales_summer_lm_c)$coefficients
coef_table["PromotionYes", ]
```

```
    Estimate   Std. Error      t value     Pr(>|t|)
 0.464958063  0.004794954 96.968196866  0.000000000
```

After centralizing the $log(Price)$ in R, the output table shows that the test statistics t value for $Promotion_i$ is $t = 96.96$, meaning that the observed statistics is 96.96 standard errors away from zero. The resulting p-value is extremely small, very close to zero.

Using the two sided t-test with $\alpha = 0.05$ significance level, the critical values are approximately ±1.96. With t value larger than the critical value and p value smaller than the significant level, we have enough statistical evidence to reject the null hypothesis. Therefore, we conclude that at the centered price level and setting $SectionWomen_i = 0$, $Promotion_i$ has a statistically significant effect on $log(SalesVolume_i)$. Transferring the conclusion back to the original scale, there is a statistically significant relationship between promotion and sales volume.

The estimated coefficient $\hat{\beta}_2 = 0.4649$ shows that at a centered price level for men's products, promoted product are expected to have sales volume multiply by $e^{0.4649} = 1.59$.

Testing C: The effect of gender section on sales volume

This hypothesis aims to determine the whether gender section (women vs men) is statistically significant associated with the change in $log(SalesVolume_i)$, controlling for promotion and price. To make this hypothesis effective in determine section's effect on sales volume, we center $log(Price_i)$ using this equation:

$$X_i = log(Price_i) - log(Price)$$

After defining this, $X_i = 0$ would be presenting a product's typical price. The equation can be simplify into:

$$E[log(SalesVolume_i) \mid X_i = 0, Promotion_i = 0] = 7.298940 + 0.090695 \cdot SectionWomen_i$$

Null Hypothesis: Holding price at centered value and promotion constant, there is no statistically significant association between $SectionWomen_i$ and $log(SalesVolume_i)$.

$$H_0 : \beta_3 = 0$$

Alternative Hypothesis: Holding price at centered value and promotion constant, there is statistically significant association between $SectionWomen_i$ and $log(SalesVolume_i)$.

$$H_A : \beta_3 \neq 0$$

```r
coef_table <- summary(log_zara_sales_summer_lm_c)$coefficients
coef_table["sectionWOMAN",  ]
```

```
     Estimate    Std. Error       t value      Pr(>|t|)
 9.920864e-02  3.838040e-03  2.584878e+01 9.357653e-133
```

After centralizing the $log(Price)$ in R, the output table shows that the test statistics t value for $SectionWomen_i$ is $t = 25.85$, meaning that the observed statistics is 25.85 standard errors away from zero. The resulting p-value is 9.36×10^{-133}, meaning that the probability of observing a t-value that is as extreme as the test statistics is 9.36×10^{-133}.

Using the two sided t-test with $\alpha = 0.05$ significance level, the critical values are approximately ± 1.96. With t value larger than the critical value and p value smaller than the significant level, we have enough statistical evidence to reject the null hypothesis. Therefore, we conclude that at the centered price level and setting $Promotion_i = 0$, $SectionWomen_i$ has a statistically significant effect on $log(SalesVolume_i)$. Transferring the conclusion back to the original scale, there is a statistically significant relationship between section and sales volume.

The estimated coefficient $\hat{\beta}_3 = 0.09921$ shows that at a centered price level for non- promoted product, women's section are expected to have sales volume by $e^{0.09921} = 1.1$, which is about 10% more sales volume.

Testing D: The effect of interactive terms on sales volume (Nested-F test)

Then, we will use a nested F-test to test the combined effect of all interactive terms in explaining variation in the model.

Null Hypothesis: the interactive terms in the model does not contribute significantly in explaining variation.

$$H_0 : \beta_{Price \times Promotion} = \beta_{Price \times SectionWomen} = \beta_{Promotion \times SectionWomen}$$

$$= \beta_4 = \beta_5 = \beta_6 = 0$$

Alternative hypothesis: at least one of the interactive terms contribute statistically significant to the model.

$$H_A : at\ least\ one\ of\ \beta_4, \beta_5, \beta_6 \neq 0$$

```r
#Reduced log transformed model
log_zara_sales_reduced <- lm(
  log(`Sales Volume`) ~ log(price) + Promotion + section,
  data = Zara_sales_EDA_summer
)
```

The reduced model we will use for testing is

$$E[log(SalesVolume_i) \mid log(Price_i), Promotion_i, SectionWomen_i]$$

$$= \beta_0 + \beta_1 \cdot log(Price_i) + \beta_2 \cdot Promotion_i + \beta_3 \cdot SectionWomen_i$$

```r
anova(log_zara_sales_reduced, log_zara_sales_summer_lm)
```

```
Analysis of Variance Table

Model 1: log(`Sales Volume`) ~ log(price) + Promotion + section
Model 2: log(`Sales Volume`) ~ log(price) * (Promotion + section) + Promotion *
    section
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1   2902 15.942
2   2899 15.766  3   0.17611 10.794 4.73e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the R output table for the nested F-test, the test statistics F-value has a value of 10.794. A large F indicates that the inclusion of interactive terms improves the model fit compared to the reduced model. The resulting p-value is 4.73×10^{-7}, meaning that the probability of observing a F value that is as extreme as the observed value is 4.73×10^{-7}. Since the p-value is smaller than the critical value of $\alpha = 0.05$, we can have sufficient statistical evidence to reject the null hypothesis. Therefore, we can conclude that the interactive terms contribute to the model meaningfully. Transferring the conclusion back to the original scale, the interactive terms contribute significantly to the model.

To further evaluate two models, we calculate the corresponding AIC, BIC, and $Adjusted\ R^2$.

```r
data.frame(
  Model = c("Reduced", "Full"),
  Adj_R2 = c(summary(log_zara_sales_reduced)$adj.r.squared,
             summary(log_zara_sales_summer_lm)$adj.r.squared),
  AIC = c(AIC(log_zara_sales_reduced), AIC(log_zara_sales_summer_lm)),
  BIC = c(BIC(log_zara_sales_reduced), BIC(log_zara_sales_summer_lm))
)
```

```
    Model    Adj_R2       AIC       BIC
1 Reduced 0.9216177 -6870.499 -6840.627
2    Full 0.9224034 -6896.780 -6848.983
```

Comparing $Adjusted\ R^2$ for the two model, the full model has a value of 0.9224034 and the reduced model has the value of 0.9216177, so the full model explains more variation than the reduced one. Comparing AIC, the full model has a value of -6896.780 and the reduced model has a value of -6870.499. Since better model has smaller value of AIC, the full model better balances the trade-off between model complexity and model fit. Comparing BIC, the full model has a value of -6848.983 and the reduced model has a value of -6840.627. Since better model has smaller value of BIC, the full model is better even the calculation of BIC penalizes models with more terms. Combining the above factors altogether, full model has larger $Adjusted\ R^2$ and smaller AIC and BIC, the full model including the interactive terms is preferred.

Because of the conclusions made from nested F-test and model selection metric, the interactive terms contribute significantly to improve the model. Therefore, we will preserve these terms and perform hypothesis testing on each of them to test their individual effects to the model.

Testing E: The effect of Price × Promotion interaction on sales volume

This hypothesis aims to determine whether the interactive term for price and promotion contribute significantly to the statistical model for sales volume.

Null Hypothesis: Controlling variable section constant, the interaction between log-price and promotion does not contribute significantly to the model.

$$H_0 : \beta_4 = 0$$

Alternative Hypothesis: Controlling variable section constant, the interaction between log-price and promotion contribute significantly to the model.

$$H_A : \beta_4 \neq 0$$

```r
coef_table <- summary(log_zara_sales_summer_lm_c)$coefficients
coef_table["log_price_c:PromotionYes", ]
```

```
    Estimate   Std. Error      t value     Pr(>|t|)
3.020136e-02 5.345079e-03 5.650310e+00 1.757091e-08
```

After controlling for the variable $SectionWomen_i$, the output table shows that the test statistics t value for term $log(Price_i) \times Promotion_i$ is $t = 5.650310$, meaning that the observed statistics is 5.650310 standard errors away from zero. The resulting p-value is 1.757091×10^{-8}, meaning that the probability of observing a t-value that is as extreme as the test statistics is 1.757091×10^{-8}.

Using the two sided t-test with $\alpha = 0.05$ significance level, the critical values are approximately ± 1.96. With t value larger than the critical value and p value smaller than the significant level, we have enough statistical evidence to reject the null hypothesis. Therefore, we conclude that at the constant $SectionWomen_i$ level, the interactive term $log(Price_i) \times Promotion_i$ contribute significantly to the model. Transferring the conclusion back to the original scale, the term $Price_i \times Promotion_i$ contributes significantly as well.

The slope coefficients of the log-log model represent elasticity. The positive $\hat{\beta}_4 = 0.030201$ shows that the price elasticity for promoted products is 0.030201. This means that the sales column of the promoted item is less negatively influenced by this price increase. This suggests that sales volume for promoted products is less sensitive to price increases compared to non-promoted products.

Testing F: The effect of Price × SectionWomen interaction on sales volume

This hypothesis aims to determine whether the interactive term for price and section contribute significantly to the statistical model for sales volume.

Null Hypothesis: Controlling for promotion, the interaction between log-price and gender section does not contribute significantly to the model.

$$H_0 : \beta_5 = 0$$

Alternative Hypothesis: Controlling for promotion, the interaction between log-price and gender section does contribute significantly to the model.

$$H_A : \beta_5 \neq 0$$

```
coef_table <- summary(log_zara_sales_summer_lm_c)$coefficients
coef_table["log_price_c:sectionWOMAN", ]
```

```
   Estimate  Std. Error     t value    Pr(>|t|)
0.002335356 0.005545821 0.421101849 0.673711928
```

After controlling for the variable $Promotion_i$, the output table shows that the test statistics t value for term $log(Price_i) \times SectionWomen_i$ is $t = 0.421101849$, meaning that the observed statistics is 0.421101849 standard errors away from zero. The resulting p-value is 0.673711928, meaning that the probability of observing a t-value that is as extreme as the test statistics is 0.673711928.

Using the two sided t-test with $\alpha = 0.05$ significance level, the critical values are approximately ± 1.96. With small t-value and p-value larger than the critical value, we failed to reject the null hypothesis. Therefore, we conclude that at the constant $Promotion_i$ level, the interactive term $log(Price_i) \times SectionWomen_i$ does not contribute significantly to the model. Transferring the conclusion back to the original scale, the term $Price_i \times SectionWomen_i$ does not contributes significantly as well.

Then, to determine whether we should delete the term from the model, we compute the model selection metrics.

```
#Reduced log transformed model
log_zara_sales_reduced_section <- lm(
  log(`Sales Volume`) ~ log(price)*Promotion + section + Promotion*section,
  data = Zara_sales_EDA_summer
)
```

The reduced model we will use for testing is

$$E[log(SalesVolume_i) \mid log(Price_i), Promotion_i, SectionWomen_i] = \beta_0 + \beta_1 \cdot log(Price_i)$$
$$+\beta_2 \cdot Promotion_i + \beta_3 \cdot SectionWomen_i + \beta_4 \cdot log(Price_i) \cdot Promotion_i$$
$$+\beta_5 \cdot Promotion_i \cdot SectionWomen_i$$

```
# reduced model
log_zara_sales_reduced_section |> glance()
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared  sigma statistic p.value    df logLik    AIC    BIC
      <dbl>         <dbl>  <dbl>     <dbl>   <dbl> <dbl>  <dbl>  <dbl>  <dbl>
1     0.923         0.922 0.0737     6910.       0     5  3456. -6899. -6857.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
# full model
log_zara_sales_summer_lm |> glance()
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared  sigma statistic p.value    df logLik    AIC    BIC
      <dbl>         <dbl>  <dbl>     <dbl>   <dbl> <dbl>  <dbl>  <dbl>  <dbl>
1     0.923         0.922 0.0737     5756.       0     6  3456. -6897. -6849.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Comparing $Adjusted\ R^2$ for the two model, both model have similar $Adjusted\ R^2$ of approximately 0.923.

Comparing AIC, the full model has a value of -6897 and the reduced model has a value of -6899. Since better model has smaller value of AIC, the reduced model better balances the trade-off between model complexity and model fit.

Comparing BIC, the full model has a value of -6849 and the reduced model has a value of -6857. Since better model has smaller value of BIC, the reduced model is better even the calculation of BIC penalizes models with more terms.

Combining the above factors altogether, two model has similar $Adjusted\ R^2$, whereas the reduced model has smaller AIC and BIC, so the reduced model excluding the interactive term between section and log-price is preferred. Therefore, we will delete the term $log(Price_i) \times SectionWomen_i$ from the model, which also applies to the model with original scale.

Testing G: The effect of Promotion × SectionWomen interaction on sales volume

This hypothesis aims to determine whether the interactive term between promotion and section contribute significantly to the statistical model for sales volume.

Null Hypothesis: Controlling price at constant level, the interaction between promotion and section does not contribute significantly to the model.

$$H_0 : \beta_6 = 0$$

Alternative Hypothesis: Controlling price at constant level, the interaction between promotion and section contribute significantly to the model.

$$H_A : \beta_6 \neq 0$$

```
coef_table <- summary(log_zara_sales_summer_lm_c)$coefficients
coef_table["PromotionYes:sectionWOMAN", ]
```

```
    Estimate    Std. Error       t value      Pr(>|t|)
-0.001767356  0.005919094 -0.298585475  0.765277749
```

After controlling for the variable $log(Price_i)$, the output table shows that the test statistics t value for term $Promotion_i \times SectionWomen_i$ is $t = -0.298585475$, meaning that the observed statistics is 0.298585475 standard errors away from zero. The resulting p-value is 0.765277749, meaning that the probability of observing a t-value that is as extreme as the test statistics is 0.765277749.

Using the two sided t-test with $\alpha = 0.05$ significance level, the critical values are approximately ± 1.96. With small t-value and p-value larger than the critical value, we failed to reject the null hypothesis. Therefore, we conclude that at the constant $log(Price_i)$ level, the interactive term $Promotion_i \times SectionWomen_i$ does not contribute significantly to the model. Transferring the conclusion back to the original scale, the term $Promotion_i \times SectionWomen_i$ does not contributes significantly as well.

Then, to determine whether we should delete the term from the model, we compute the model selection metrics.

```r
#Reduced log transformed model
log_zara_sales_reduced_section_promotion <- lm(
  log(`Sales Volume`) ~ log(price)*Promotion + section,
  data = Zara_sales_EDA_summer
)
```

The reduced model we will use for testing is

$$E[log(SalesVolume_i) \mid log(Price_i), Promotion_i, SectionWomen_i] = \beta_0 + \beta_1 \cdot log(Price_i)$$

$$+\beta_2 \cdot Promotion_i + \beta_3 \cdot SectionWomen_i + \beta_4 \cdot log(Price_i) \cdot Promotion_i$$

```r
# reduced model
log_zara_sales_reduced_section_promotion |> glance()
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared  sigma statistic p.value    df logLik    AIC    BIC
      <dbl>         <dbl>  <dbl>     <dbl>   <dbl> <dbl>  <dbl>  <dbl>  <dbl>
1     0.923         0.922 0.0737     8639.       0     4  3456. -6900. -6865.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```r
# full model
log_zara_sales_summer_lm |> glance()
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared  sigma statistic p.value    df logLik    AIC    BIC
      <dbl>         <dbl>  <dbl>     <dbl>   <dbl> <dbl>  <dbl>  <dbl>  <dbl>
1     0.923         0.922 0.0737     5756.       0     6  3456. -6897. -6849.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Comparing $Adjusted\ R^2$ for the two model, both model have similar $Adjusted\ R^2$ of approximately 0.923.

Comparing AIC, the full model has a value of -6897 and the reduced model has a value of -6900. Since better model has smaller value of AIC, the reduced model better balances the trade-off between model complexity and model fit.

Comparing BIC, the full model has a value of -6849 and the reduced model has a value of -6865. Since better model has smaller value of BIC, the reduced model is better even the calculation of BIC penalizes models with more terms.

Combining the above factors altogether, two model has similar $Adjusted\ R^2$, whereas the reduced model has smaller AIC and BIC, so the reduced model excluding the interactive term between section and log-price is preferred. Therefore, we will delete the term $Promotion_i \times SectionWomen_i$ from the model which also applies to the model with original scale.

VII. Results

Concluding from the results of hypothesis testing and model selection metric, we decide to exclude the terms $Promotion_i \times SectionWomen_i$ and $log(Price_i) \times SectionWomen_i$ from the model as they do not contribute significantly to the model.

```r
log_zara_sales_reduced_section_promotion <- lm(
  log(`Sales Volume`) ~ log(price)*Promotion + section,
  data = Zara_sales_EDA_summer
)
summary(log_zara_sales_reduced_section_promotion)
```

```
Call:
lm(formula = log(`Sales Volume`) ~ log(price) * Promotion + section,
    data = Zara_sales_EDA_summer)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.25217 -0.05113  0.00206  0.05149  0.20574 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)              7.293942   0.013326 547.367  < 2e-16 ***
log(price)              -0.140782   0.003507 -40.139  < 2e-16 ***
PromotionYes             0.353459   0.019535  18.093  < 2e-16 ***
sectionWOMAN             0.098465   0.002886  34.115  < 2e-16 ***
log(price):PromotionYes  0.030261   0.005342   5.664 1.62e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.07372 on 2901 degrees of freedom
Multiple R-squared:  0.9226,    Adjusted R-squared:  0.9224
F-statistic:  8639 on 4 and 2901 DF,  p-value: < 2.2e-16
```

From the R output, the new fitted model is:

$$E[log(SalesVolume_i) \mid log(Price_i), Promotion_i, SectionWomen_i] = 7.293942 - 0.140782 \cdot log(Price_i)$$

$$+0.353459 \cdot Promotion_i + 0.098465 \cdot SectionWomen_i + 0.030261 \cdot log(Price_i) \cdot Promotion_i$$

where the reference level for this model is the average estimated sales volume for male without promotion;

$log(SalesVolume_i)$ is the log-transformed predicted average number of units sold for a product;

$log(Price_i)$ is the log-transformed sales price, measured in log-transformed dollars;

$Promotion_i$ is a dummy variable, if it is equals to 1 then there is a promotion for the product promoted, 0 otherwise;

$SectionWomen_i$ is a dummy variable, if it is equals to 1 the product belongs to the women's section, 0 belongs to men's section;

$\beta_0 = 7.293942$ measures the log-transformed value of sales volume when price is zero. The median sales volume when price is equal to zero is $e^{7.293942} = 1471.359$ units.

$\beta_1 = -0.140782$ measures the log-transformed slope. A two-fold multiplicative change in price is associated with a multiplicative change of $2^{-0.140782} = 0.9070274$ in median sales volume, holding other variables constant.

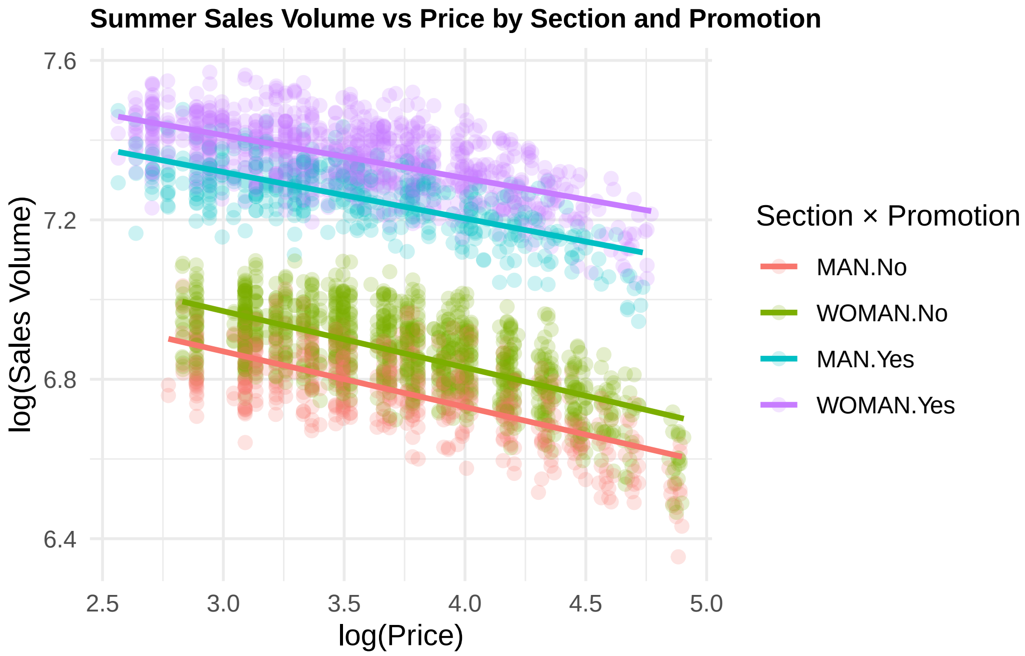
$\beta_2 = 0.353459$ means that when price is set equal to zero, the median sales volume for promoted items is $e^{0.353459} = 1.423985$ times the median sales volume for non-promoted items, holding section constant.

$\beta_3 = 0.098465$ means that when price is set to zero, the median sales volume by women is $e^{0.098465} = 1.103476$ times the median sales volume for men, holding promotion status constant.

$\beta_4 = 0.030261$ means that a two-fold multiplicative change in price is associated with a increase in median of sales volume in promoted products of $2^{0.030261} = 1.021197$ times higher than the increase in median sales volume in non-promoted products, holding section constant.

Discussing the fitted multiple regression model under different conditions referring to the visualization. The fitted regression line is shown in the graph below

```r
library(ggplot2)
Zara_sales_EDA_summer |>
  ggplot(aes(x = log(price),
             y = log(`Sales Volume`),
             color = interaction(section, Promotion))) +
  geom_point(size = 2, alpha = 0.2) +
  geom_smooth(method = "lm", se = FALSE, linewidth = 1) +
  labs(
    title = "Summer Sales Volume vs Price by Section and Promotion",
    x = "log(Price)",
    y = "log(Sales Volume)",
    color = "Section × Promotion"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 10, face = "bold"),
    legend.position = "right"
  )
```



For products for men without promotion, the relationship between log-transformed sales volume and log-transformed price is shown by the red line.

$$E[log(SalesVolume_i) \mid log(Price_i), Promotion_i = 0, SectionWomen_i = 0]$$

$$= 7.293942 - 0.140782 \cdot log(Price_i)$$

For products for women without promotion, the relationship between log-transformed sales volume and log-transformed price is shown by the green line.

$$E[log(SalesVolume_i) \mid log(Price_i), Promotion_i = 0, SectionWomen_i = 1]$$

$$= 7.392407 - 0.140782 \cdot log(Price_i)$$

For products for men with promotion, the relationship between log-transformed sales volume and log-transformed price is shown by the blue line.

$$E[log(SalesVolume_i) \mid log(Price_i), Promotion_i = 1, SectionWomen_i = 0]$$

$$= 7.647401 - 0.110521 \cdot log(Price_i)$$

For products for women with promotion, the relationship between log-transformed sales volume and log-transformed price is shown by the purple line.

$$E[log(SalesVolume_i) \mid log(Price_i), Promotion_i = 1, SectionWomen_i = 1]$$

$$= 7.745866 - 0.110521 \cdot log(Price_i)$$

To summarize, within the groups with $Promotion_i = 1$ and $Promotion_i = 0$, the slope of regression line for men and women is the same with different intercept, whereas between the groups with $Promotion_i = 1$ and $Promotion_i = 0$, the slope and intercept both differs. Therefore, only the promotion status changes consumers' sensitivity in price measured by the sales volume.

We also calculate the confident interval to measure the multiplicative effects on sales volume.

```
conf <- confint(log_zara_sales_reduced_section_promotion)

conf_no_intercept <- conf[rownames(conf) != "(Intercept)", ]

pct_conf <- (exp(conf_no_intercept) - 1) * 100
pct_conf
```

```
                              2.5 %     97.5 %
log(price)               -13.727470 -12.532661
PromotionYes              37.047157  47.958821
sectionWOMAN               9.724819  10.973818
log(price):PromotionYes    1.998262   4.157724
```

We are 95% confident that a 1% increase in price is associated with an estimated 12.5–13.7% decrease in sales volume, holding promotion and section constant.This result suggests strong price sensitivity among consumers, which is consistent with the competitive nature of the fast fashion industry where substitutes are readily available. Promotion also has a significant effect in improving the sales volume and stimulating consumers' demand, which we are 95% confident that applying promotion will increase 37.04% - 47.9% of sales volume. The coefficient for section women is also statistically important. From the 95% confidence interval test, we can that the there is an 9.7% to 10.9% improvement in sales volume compared to men's section. For the interaction term of log(price) and Promotion, we are 95% confident to say that negative effect of increasing price per percent was mitigated for 2.0% and 4.2% when promotions are applied. It shows that applying promotion can reduce the price sensitivity for consumers, matching the conclusions from the visualization, which is also a strategy Zara could apply in future sales.

VIII. Conclusion

To answer our research question – how the sales-related factors, including price, gender section, and promotion status in Zara, influence consumer choices and preferences as measured by sales volume

– we choose to fit a multiple regression model including interactive terms. This choice allows us to investigate the complex relationship how pricing relates to sales volume under different promotional contexts, providing insights to consumer behaviors.

Additionally, after constructing the initial model, we assessed conditions and found violations of the conditions, so we transformed the model by using log-transformation in both response variables and the explanatory variable, ensuring the validity and reliability of statistical analysis and inferences. Then, the hypothesis testing and model selection processes ensure that the terms in the constructed transformed multiple regression models all contribute significantly to reducing model variation. Our final model only includes the interaction effects of price and promotion. Therefore, the final model we generated balances model fit and model complexity well, strengthening the credibility of conclusions drawn from the analysis.

There are still several limitations in our research and model as well, even though the model helps explain part of the effects of factors improving sales volume. The Zara data does not include product-level characteristics such as style, consumers' perceived satisfaction with products, which are all very important factors that influence consumer preferences that are represented by sales volume in our analysis. Such an omission may lead to bias in coefficient estimation. To improve this limitation, we can find other Zara data sets that include the above-mentioned factors and join them with the current sales product information dataset to produce a more comprehensive model that helps us understand the data in a different context.

Moreover, according to the original dataset description, there were seven thousand observations, so they used oversampling to increase the number of records and balance categories for better statistical analysis. Although it improves model stability and performance, it may also introduce duplicated information and affect the variance of estimation, thereby reducing its ability to be representative of the population. To improve this, we may find the original dataset with seven thousand observations and refit a statistical model to the dataset.

In addition, considering that we only select data from the "summer" season for analysis, seasonality may also be a factor influencing consumer decisions. For instance, in summer, consumers may place more emphasis on current fashion trends over the functionality of clothes. In contrast, in winter, the functionality and quality are more valued than the fashion design. Therefore, the relationship identified in this study only reflects consumer behaviors in summer and may not be generalized to other seasons for the entire population. To construct a more applicable model, we will, in the future, construct models for the rest three seasons separately, providing deeper insights into consumer behaviors for Zara and other brands, allowing them to better understand the mechanisms that drive sales volume.

IX. References

Duoyan, Hu. 2021. "Research on ZARA Strategy from the Perspective of SWOT Analysis Method." *Proceedings of the 2021 6th International Conference on Social Sciences and Economic Development (ICSSED 2021)* 543. https://www.researchgate.net/publication/350916569_Research_on_ZARA_Strategy_from_the_Perspective_of_SWOT_Analysis_Method.

Familmaleki, Mahsa, Alireza Aghighi, and Kambiz Hamidi. 2015. "Analyzing the Influence of Sales Promotion on Customer Purchasing Behavior." *International Journal of Economics and Management Sciences* 04: 1–6. https://doi.org/10.4172/2162-6359.1000243.

Pradhana, Faizal, and Prani Sastiono. 2019. "Gender Differences in Online Shopping: Are Men More Shopaholic Online?" *Proceedings of the 12th International Conference on Business and Management Research (ICBMR 2018)* 72. https://doi.org/10.2991/icbmr-18.2019.21.

Uberoi, Ravneet. 2017. "ZARA: Achieving the 'Fast' in Fast Fashion Through Analytics." Digital Innovation; Transformation. https://d3.harvard.edu/platform-digit/submission/zara-achieving-the-fast-in-fast-fashion-through-analytics/.

Vimal, Shubhendu. 2025. "Zara Owner Inditex Reports Q3 2025 Sales up with Positive Start to Q4." Yahoo Finance. https://finance.yahoo.com/news/zara-owner-inditex-reports-q3-152609858.html.