

Customer Sentiment Analysis

Women's E-Commerce Clothing Reviews

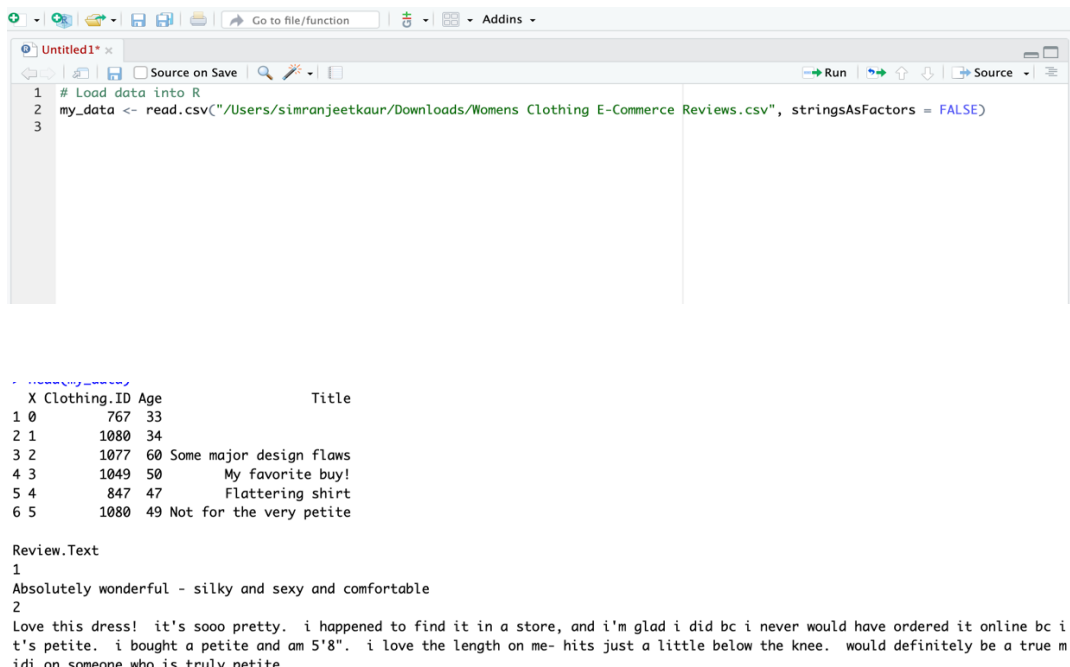
Objective:

To conduct sentiment analysis on customer reviews for a specific product or service category using R. This analysis will help us understand overall customer sentiment (positive, negative, neutral) towards the product or service and identify key themes or issues that frequently arise in customer feedback.

Goals:

- **Sentiment Classification:** Classify each review into positive, negative, or neutral categories.
- **Trend Analysis:** Identify any trends in sentiment over time. For instance, does product sentiment improve following updates or changes?
- **Key Themes Identification:** Extract common themes or topics mentioned in reviews, such as product features, customer service experience, pricing, etc.
- **Insight Generation:** Provide actionable insights based on the analysis that can help improve the product or service.

Load the Data



```
1 # Load data into R
2 my_data <- read.csv("/Users/simranjeetkaur/Downloads/Womens Clothing E-Commerce Reviews.csv", stringsAsFactors = FALSE)
3
```


	X	Clothing.ID	Age	Title
1	0	767	33	
2	1	1080	34	
3	2	1077	60	Some major design flaws
4	3	1049	50	My favorite buy!
5	4	847	47	Flattering shirt
6	5	1080	49	Not for the very petite


```
Review.Text
1
Absolutely wonderful - silky and sexy and comfortable
2
Love this dress! it's sooo pretty. i happened to find it in a store, and i'm glad i did bc i never would have ordered it online bc i
t's petite. i bought a petite and am 5'8". i love the length on me- hits just a little below the knee. would definitely be a true m
idi on someone who is truly petite.
```

Exploratory Data Analysis (EDA):

- Summarize the data:

```
> summary(my_data_frame)
      X      Clothing.ID      Age      Title      Review.Text      Rating      Recommended.IND
Min.   : 0      Min.   : 0.0      Min.   :18.0      Length:23486      Length:23486      Min.   :1.000      Min.   :0.0000
1st Qu.:5871    1st Qu.: 861.0    1st Qu.:34.0    Class :character      Class :character      1st Qu.:4.000      1st Qu.:1.0000
Median :11742   Median : 936.0    Median :41.0    Mode  :character      Mode  :character      Median :5.000      Median :1.0000
Mean   :11742   Mean   : 918.1    Mean   :43.2                                Mean   :4.196      Mean   :0.8224
3rd Qu.:17614   3rd Qu.:1078.0    3rd Qu.:52.0                                3rd Qu.:5.000      3rd Qu.:1.0000
Max.   :23485   Max.   :1205.0    Max.   :99.0                                Max.   :5.000      Max.   :1.0000

Positive.Feedback.Count Division.Name Department.Name Class.Name
Min.   : 0.000      Length:23486      Length:23486      Length:23486
1st Qu.: 0.000      Class :character      Class :character      Class :character
Median : 1.000      Mode  :character      Mode  :character      Mode  :character
Mean   : 2.536
3rd Qu.: 3.000
Max.   :122.000
```

- Check for NA values.

1. The count of NA values in each column.

```
> sapply(my_data_frame, function(x) sum(is.na(x)))
      X      Clothing.ID      Age      Title      Review.Text
      0              0          0          0              0
Rating Recommended.IND Positive.Feedback.Count Division.Name Department.Name
      0              0          0              0              0
Class.Name
      0
```

2. Total count of NA values in the entire data frame

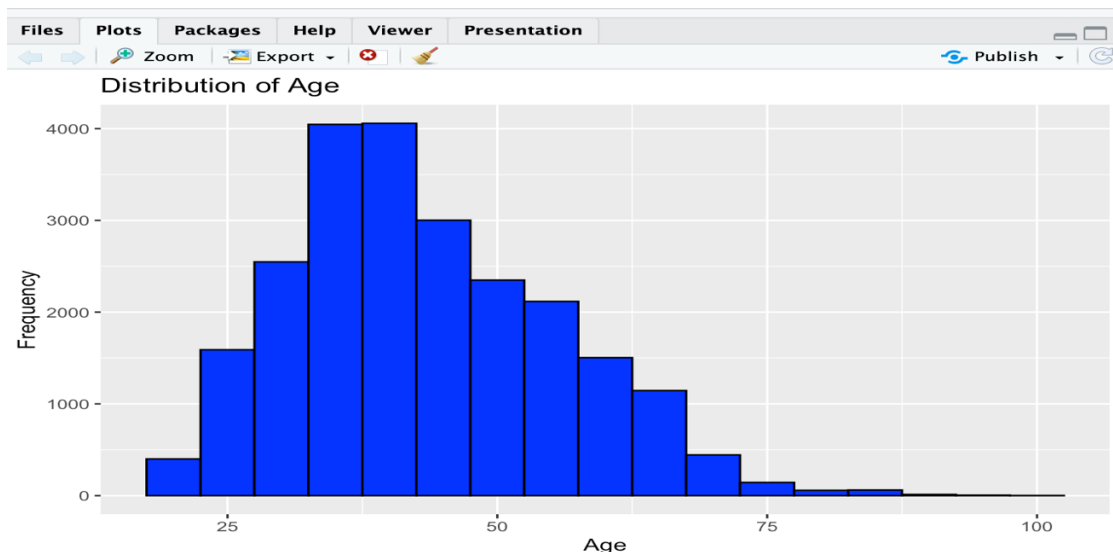
```
> sum(is.na(my_data_frame))
```

```
[1] 0
```

- Visualize the data.

1. For Age distribution

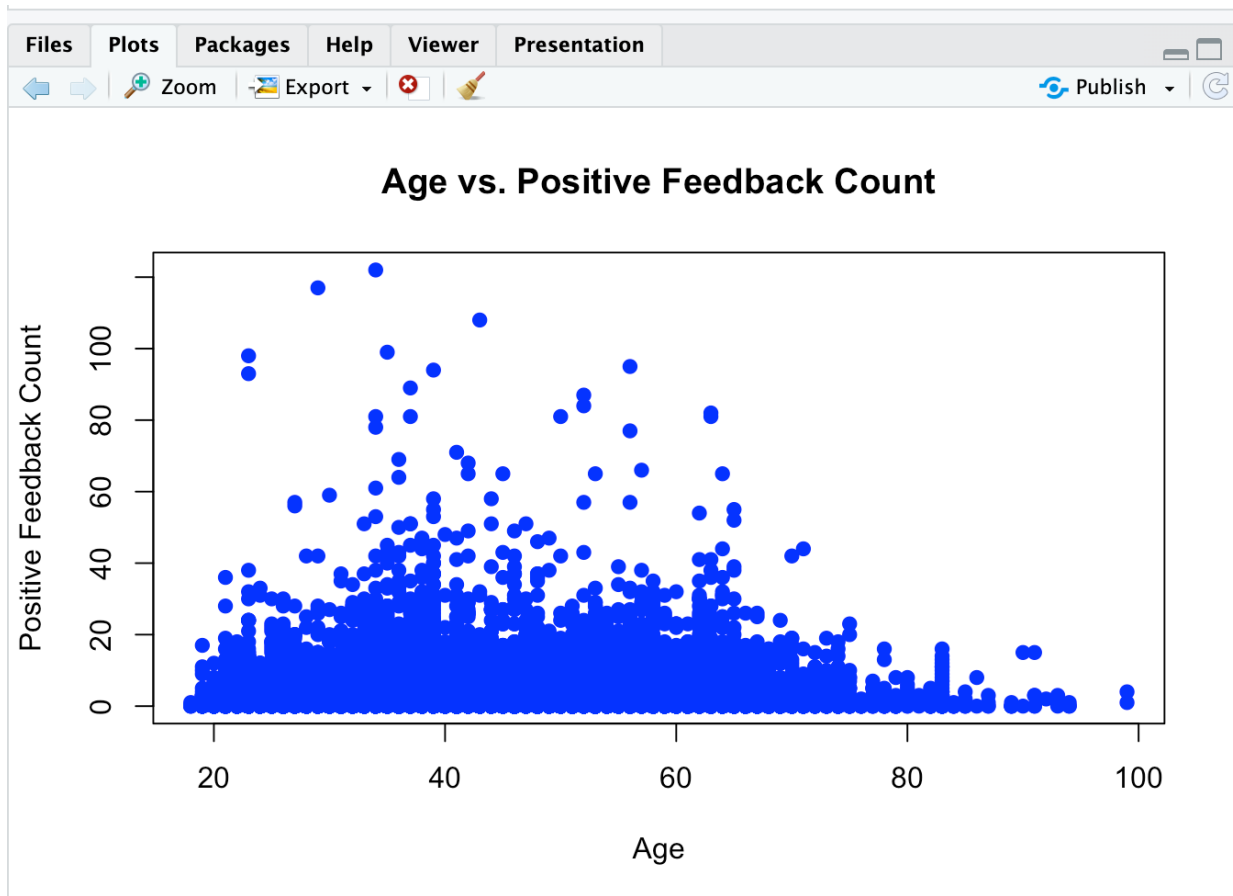
```
# For Age distribution
ggplot(my_data_frame, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  labs(title = "Distribution of Age", x = "Age", y = "Frequency")
```



2. Scatter Plot for Age vs. Positive Feedback Count

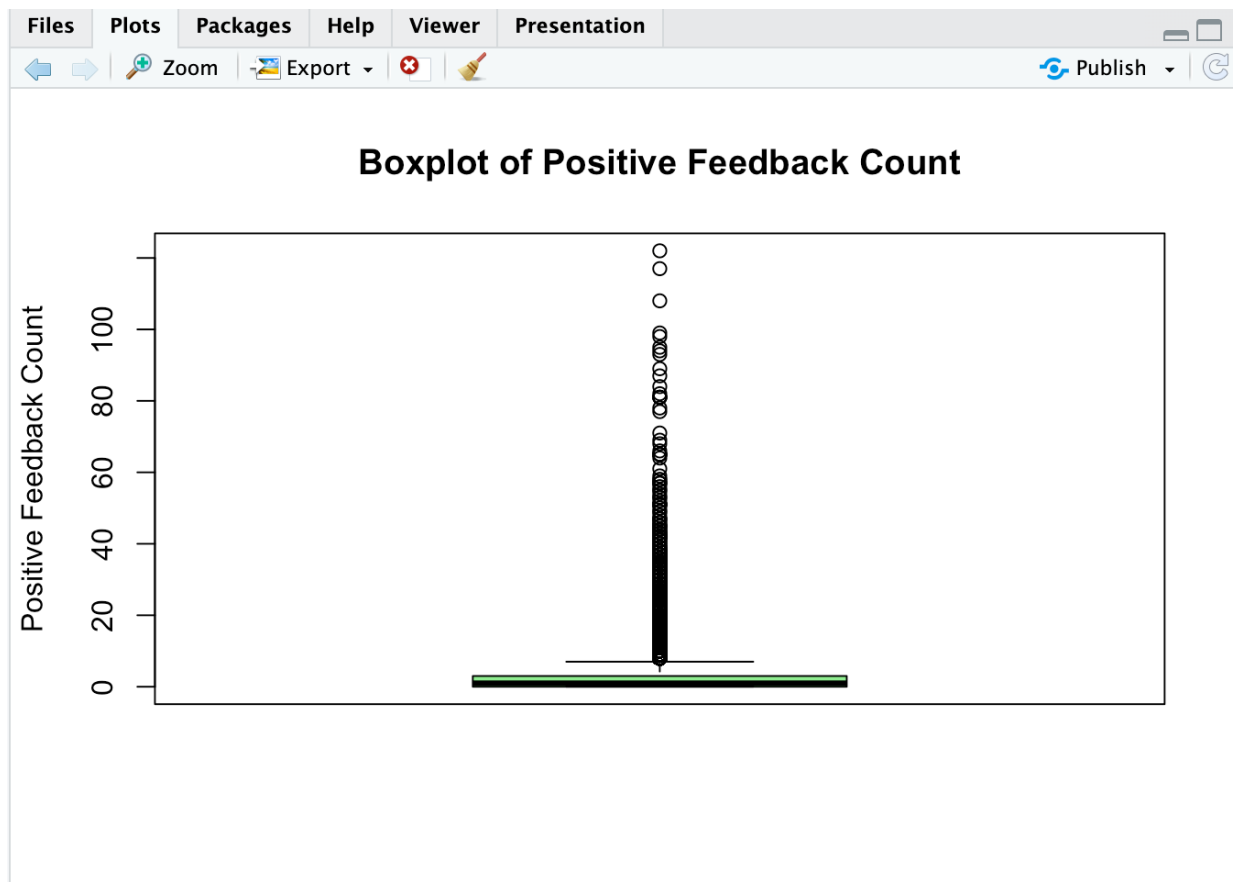
```
# Scatter Plot for Age vs. Positive Feedback Count
```

```
plot(my_data_frame$Age, my_data_frame$Positive.Feedback.Count,  
     main = "Age vs. Positive Feedback Count",  
     xlab = "Age",  
     ylab = "Positive Feedback Count",  
     pch = 19,  
     col = "blue")
```



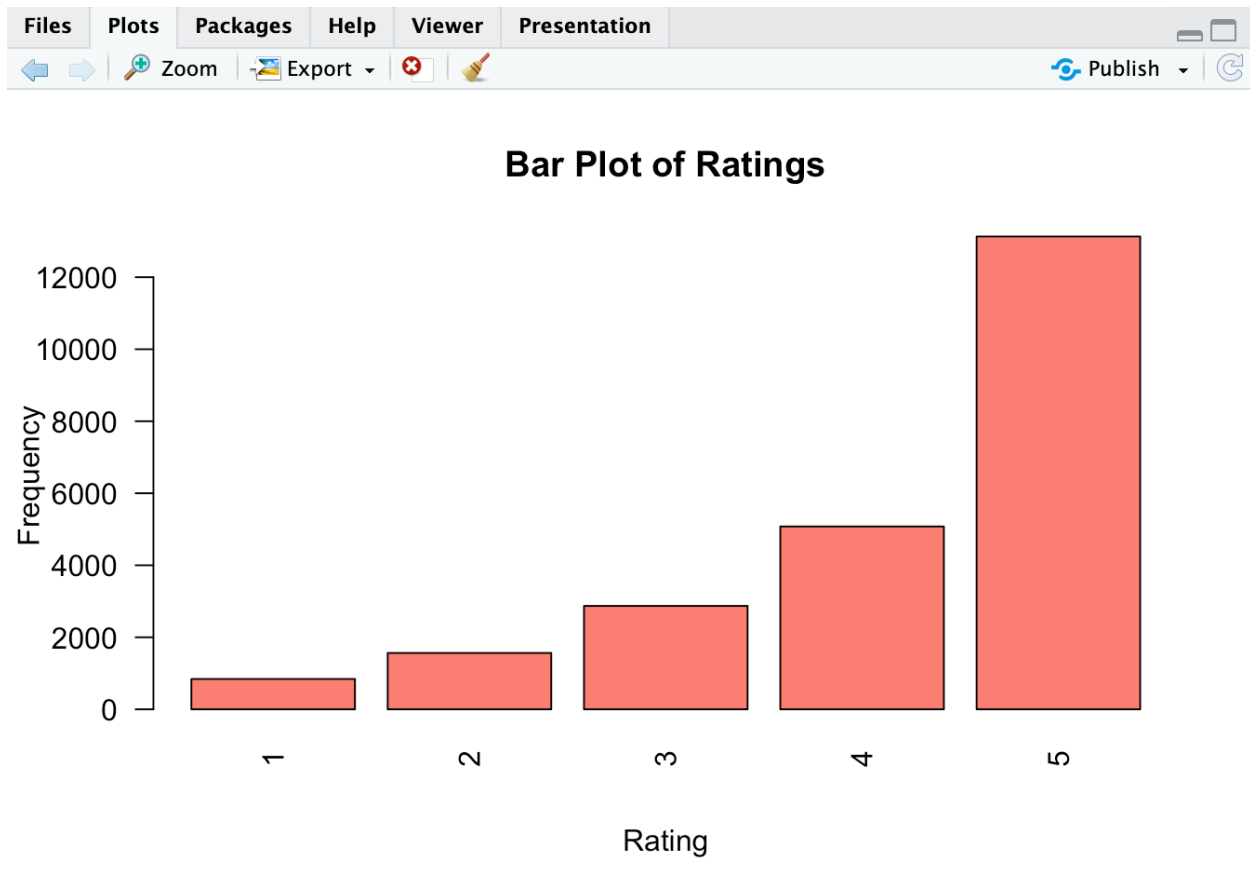
3. Outliers in the positive feedback count

```
col = "blue")  
boxplot(my_data_frame$Positive.Feedback.Count,  
        main = "Boxplot of Positive Feedback Count",  
        ylab = "Positive Feedback Count",  
        col = "lightgreen")
```



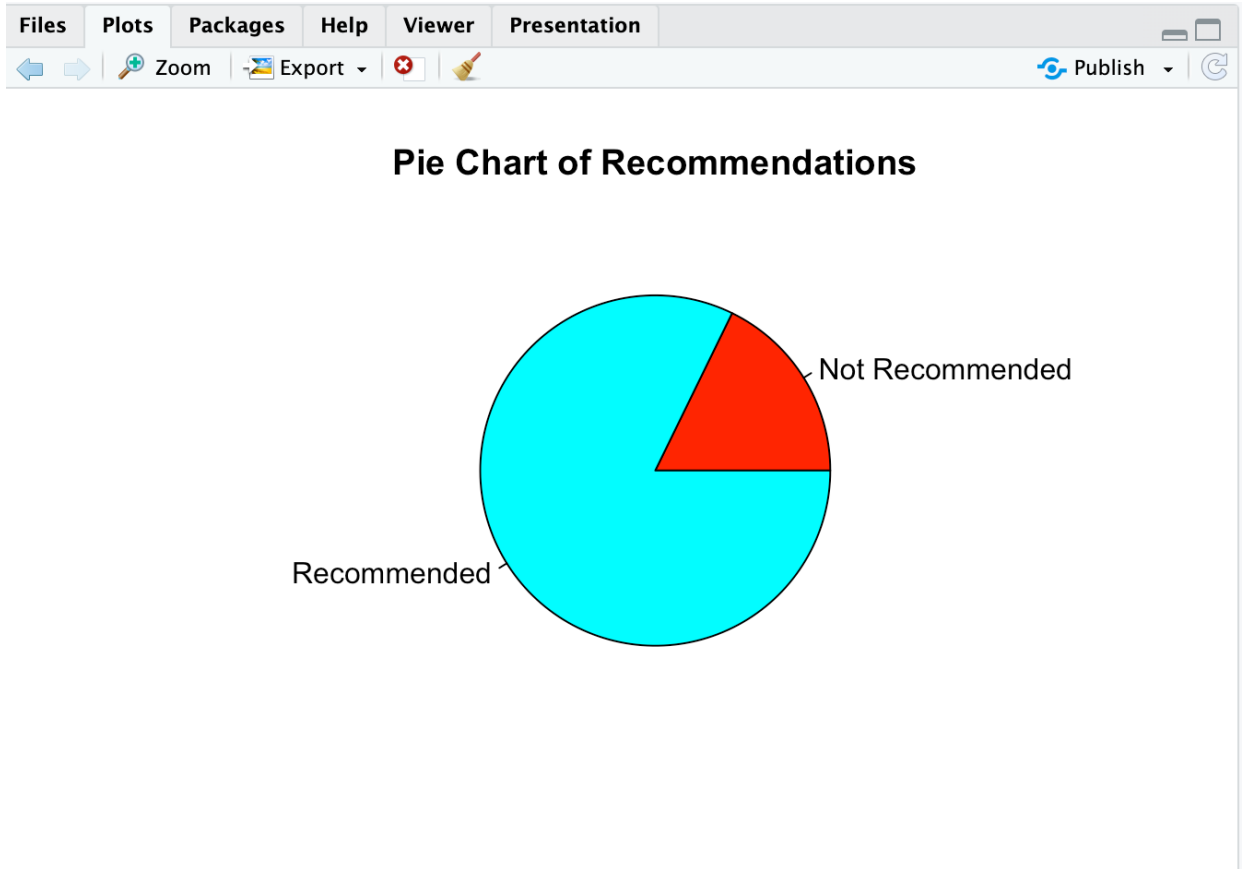
4. Frequency of Ratings

```
barplot(tableRating,  
  main = "Bar Plot of Ratings",  
  xlab = "Rating",  
  ylab = "Frequency",  
  col = "salmon",  
  las = 2)
```



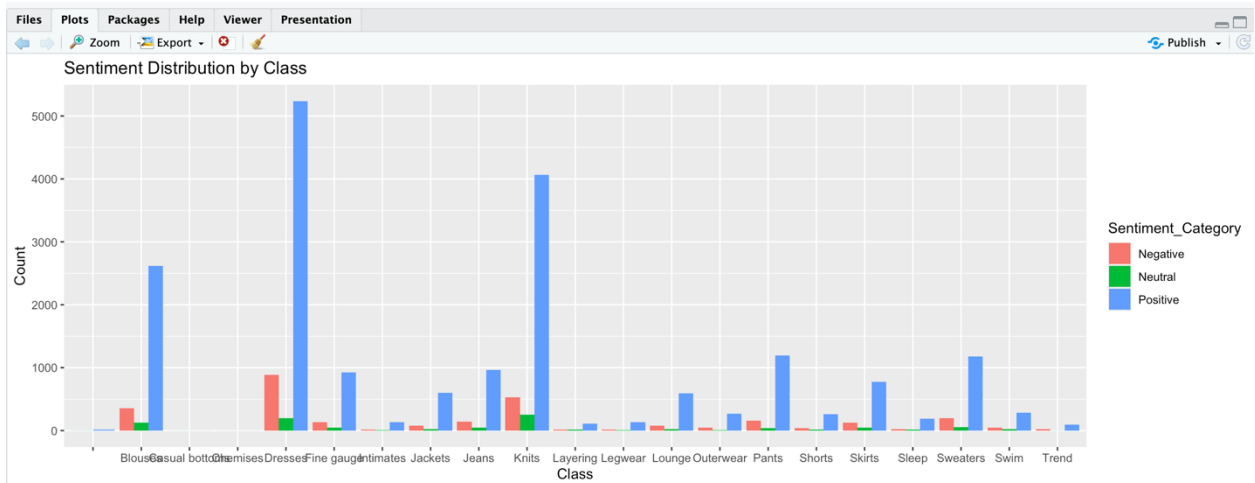
5. proportion of recommendations

```
pie(tableRecommended,  
    main = "Pie Chart of Recommendations",  
    col = rainbow(length(tableRecommended)),  
    labels = c("Not Recommended", "Recommended"))
```



- Reviews

```
ggplot(reviews, aes(x = Class.Name, fill = Sentiment_Category)) +  
  geom_bar(position = "dodge") +  
  labs(title = "Sentiment Distribution by Class", x = "Class", y = "Count")
```



Data Cleaning

- Handling Missing Values

```
# Count missing values in each column
```

```
col_missing <- colSums(is.na(my_data_frame))
```

- Removing Duplicates

```
# Identify duplicate rows
```

```
duplicate_rows <- my_data_frame[duplicated(my_data_frame), ]
```

```
# Remove duplicates
```

```
unique_data <- unique(my_data_frame)
```

duplicate_rows	0 obs. of 11 variables
\$ X	: int
\$ Clothing.ID	: int
\$ Age	: int
\$ Title	: chr
\$ Review.Text	: chr
\$ Rating	: int
\$ Recommended.IND	: int
\$ Positive.Feedback.Count	: int
\$ Division.Name	: chr
\$ Department.Name	: chr
\$ Class.Name	: chr

unique_data	23486 obs. of 11 variables												
\$ X	:	int	0	1	2	3	4	5	6	7	8	9	...
\$ Clothing.ID	:	int	767	1080	1077	1049	847	1080	858	858	1077	1077...	
\$ Age	:	int	33	34	60	50	47	49	39	39	24	34	...
\$ Title	:	chr	"	"	"Some major design flaws"	"My favorite	...						
\$ Review.Text	:	chr	"Absolutely wonderful - silky and sexy and co...										
\$ Rating	:	int	4	5	3	5	5	2	5	4	5	5	...
\$ Recommended.IND	:	int	1	1	0	1	1	0	1	1	1	1	...
\$ Positive.Feedb	\$ Recommended.IND	:	int	1	1	0	1	1	0	1	1	1	...
\$ Division.Name	:	chr	"Initmates"	"General"	"General"	"General Peti...							
\$ Department.Name	:	chr	"Intimate"	"Dresses"	"Dresses"	"Bottoms"	...						
\$ Class.Name	:	chr	"Intimates"	"Dresses"	"Dresses"	"Pants"	...						

Text Normalization

- Lowercasing

```
reviews$title <- tolower(reviews$title)
reviews$Review.Text <- tolower(reviews$Review.Text)
```

- Remove Punctuation and Numbers

```
> reviews$Review.Text <- str_remove_all(reviews$Review.Text, "[[:punct:]]")
> reviews$Review.Text <- str_remove_all(reviews$Review.Text, "[[:digit:]]")
```

	X	Clothing.ID	Age	Title	Review.Text	Rating	Recommended.IND
1	0	767	33		absolutely wonderful silky and sexy and comfortable	4	1
2	1	1080	34		love this dress its sooo pretty i happened to find it i...	5	1
3	2	1077	60	some major design flaws	i had such high hopes for this dress and really wante...	3	0
4	3	1049	50	my favorite buy	i love love love this jumpsuit its fun flirty and fabulou...	5	1
5	4	847	47	flattering shirt	this shirt is very flattering to all due to the adjustable ...	5	1
6	5	1080	49	not for the very petite	i love tracy reese dresses but this one is not for the v...	2	0
7	6	858	39	cagrcoal shimmer fun	i aded this in my basket at hte last mintue to see wha...	5	1
8	7	858	39	shimmer surprisingly goes with lots	i ordered this in carbon for store pick up and had a to...	4	1
9	8	1077	24	flattering	i love this dress i usually get an xs but it runs a little ...	5	1
10	9	1077	34	such a fun dress	im and lbs i ordered the s petite to make sure the le...	5	1
11	10	1077	53	dress looks like its made of cheap material	dress runs small esp where the zipper area runs i ord...	3	0
12	11	1095	39		this dress is perfection so pretty and flattering	5	1
13	12	1095	53	perfect	more and more i find myself reliant on the reviews wr...	5	1
14	13	767	44	runs big	bought the black xs to go under the larkspur midi dre...	5	1
15	14	1077	50	pretty party dress with some issues	this is a nice choice for holiday gatherings i like that t...	3	1
16	15	1065	47	nice but not for my body	i took these out of the package and wanted them to fi...	4	1

- Tokenization


```
library(tokenizers)
reviews$Title.Tokens <- tokenize_words(reviews$Title)
reviews$Review.Text.Tokens <- tokenize_words(reviews$Review.Text)
```

ime	Title.Tokens	Review.Text.Tokens
	character(0)	c("beautiful", "top", "but", "delicate", "i", "wor [...]
	c("super", "cute", "and", "unique", "top")	c("just", "received", "this", "in", "the", "mail", [...]
	c("beautiful", "top")	c("love", "this", "top", "i", "kept", "eyeing", "i [...]
	c("i", "wanted", "to", "love", "this", "top")	c("i", "really", "loved", "this", "top", "online", [...]
	c("great", "summer", "fabric")	c("i", "really", "wanted", "this", "to", "work", " [...]
	c("cute", "but", "massive", "sweep")	c("this", "top", "is", "so", "cute", "but", "it", [...]
	c("short", "and", "boxy")	c("why", "do", "designers", "keep", "making", "cro [...]
	c("beautiful", "design")	c("i", "have", "a", "short", "torso", "and", "this [...]
	character(0)	c("i", "love", "this", "top", "it", "is", "loose", [...]
	character(0)	c("i", "passed", "up", "this", "dress", "so", "man [...]
	c("very", "very", "cute", "but", "a", "lot", "of", [...]	c("i", "am", "so", "drawn", "to", "baby", "doll", [...]
	c("runs", "large", "in", "top", "and", "waist")	c("i", "would", "have", "loved", "this", "dress", [...]
	c("zipper", "broke")	c("the", "zipper", "broke", "on", "this", "piece", [...]
	obsessed	c("i", "usually", "size", "up", "with", "this", "b [...]
	c("so", "cutebut", "dry", "clean")	c("this", "dress", "is", "adorable", "its", "a", " [...]
	c("elegant", "classic", "dress")	c("i", "am", "usually", "a", "petite", "but", "sin [...]

Sentiment Analysis

```
# Assuming you've already performed sentiment analysis
reviews$Sentiment_Category <- ifelse(reviews$Sentiment_Score > 0, "Positive",
                                     ifelse(reviews$Sentiment_Score < 0, "Negative", "Neutral"))

# Summarize the sentiment categories
table(reviews$Sentiment_Category)
```

```
Negative  Neutral  Positive
    2903      949    19634
>
```