[7]:
```
# Load required libraries
library(dplyr)      # For data manipulation
library(ggplot2)    # For data visualization
library(corrplot)   # For correlation plots
```

```
Attaching package: ‘dplyr’


The following objects are masked from ‘package:stats’:

    filter, lag


The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union


corrplot 0.92 loaded
```

Data Exploration and Cleaning

[8]:
```
# Load the data (adjust the path to your dataset)
data <- read.csv("/kaggle/input/sustainable-fashion-eco-friendly-trends/sustainable_fashion_trends_2024.csv")
```

[9]:
```
# Overview of the data
str(data)          # Structure of the dataset
```

```
'data.frame':   5000 obs. of  15 variables:
 $ Brand_ID                 : chr  "BRAND-0001" "BRAND-0002" "BRAND-0003" "BRAND-0004" ...
 $ Brand_Name               : chr  "Brand_1" "Brand_2" "Brand_3" "Brand_4" ...
 $ Country                  : chr  "Australia" "Japan" "USA" "Italy" ...
 $ Year                     : int  2018 2015 2024 2023 2016 2017 2015 2022 2018 2011 ...
 $ Sustainability_Rating    : chr  "D" "D" "A" "D" ...
 $ Material_Type            : chr  "Tencel" "Vegan Leather" "Vegan Leather" "Bamboo Fabric" ...
 $ Eco_Friendly_Manufacturing: chr  "No" "Yes" "No" "No" ...
 $ Carbon_Footprint_MT      : num  1.75 124.39 336.66 152.04 415.63 ...
 $ Water_Usage_Liters       : num  4511153 1951566 467455 899577 1809220 ...
 $ Waste_Production_KG      : num  97844 37268 38386 32665 37295 ...
 $ Recycling_Programs       : chr  "No" "No" "No" "No" ...
 $ Product_Lines            : int  2 15 2 13 19 10 17 18 11 9 ...
 $ Average_Price_USD        : num  38.3 250.1 146.2 165.5 211.6 ...
 $ Market_Trend             : chr  "Growing" "Growing" "Growing" "Growing" "Stable" ...
 $ Certifications           : chr  "GOTS" "GOTS" "B Corp" "OEKO-TEX" ...
```

[10]:
```
summary(data)      # Summary statistics for numerical columns
```

```
   Brand_ID          Brand_Name          Country              Year
 Length:5000        Length:5000        Length:5000        Min.   :2010
 Class :character   Class :character   Class :character   1st Qu.:2013
 Mode  :character   Mode  :character   Mode  :character   Median :2017
                                                          Mean   :2017
                                                          3rd Qu.:2021
                                                          Max.   :2024
 Sustainability_Rating Material_Type      Eco_Friendly_Manufacturing
 Length:5000           Length:5000        Length:5000
 Class :character      Class :character   Class :character
 Mode  :character      Mode  :character   Mode  :character


 Carbon_Footprint_MT Water_Usage_Liters Waste_Production_KG Recycling_Programs
 Min.   :  1.04      Min.   :  50106    Min.   : 1026       Length:5000
 1st Qu.:126.61      1st Qu.:1293087    1st Qu.:25341       Class :character
 Median :250.65      Median :2499096    Median :50466       Mode  :character
 Mean   :250.32      Mean   :2517862    Mean   :50107
 3rd Qu.:372.25      3rd Qu.:3763860    3rd Qu.:74985
 Max.   :499.93      Max.   :4999597    Max.   :99948
 Product_Lines   Average_Price_USD Market_Trend       Certifications
 Min.   : 1.00   Min.   : 20.02    Length:5000        Length:5000
 1st Qu.: 5.00   1st Qu.:142.87    Class :character   Class :character
 Median :10.00   Median :258.62    Mode  :character   Mode  :character
 Mean   :10.43   Mean   :259.35
 3rd Qu.:15.00   3rd Qu.:378.60
 Max.   :20.00   Max.   :499.94
```

[11]:
```
head(data)         # View the first few rows
```

A data.frame: 6 × 15

	Brand_ID	Brand_Name	Country	Year	Sustainability_Rating	Material_Type	Eco_Friendly_Manufacturing	Carbon_Footprint_MT	Water_Usage_Liters	Waste_Production_KG	Recycling_Programs	Product_Lines	Average_Price_USD	Market_Trend	Certifications
	<chr>	<chr>	<chr>	<int>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<chr>	<int>	<dbl>	<chr>	<chr>
1	BRAND-0001	Brand_1	Australia	2018	D	Tencel	No	1.75	4511152.8	97844.11	No	2	38.33	Growing	GOTS
2	BRAND-0002	Brand_2	Japan	2015	D	Vegan Leather	Yes	124.39	1951566.3	37267.75	No	15	250.07	Growing	GOTS
3	BRAND-0003	Brand_3	USA	2024	A	Vegan Leather	No	336.66	467454.5	38385.92	No	2	146.16	Growing	B Corp
4	BRAND-0004	Brand_4	Italy	2023	D	Bamboo Fabric	No	152.04	899576.9	32665.45	No	13	165.52	Stable	OEKO-TEX
5	BRAND-0005	Brand_5	USA	2016	D	Bamboo Fabric	Yes	415.63	1809219.9	37295.47	Yes	19	211.63	Stable	Fair Trade
6	BRAND-0006	Brand_6	Italy	2017	B	Recycled Polyester	No	447.65	2244115.4	69017.63	Yes	10	196.45	Stable	B Corp

[12]:
```
# Check for missing values
colSums(is.na(data))
```

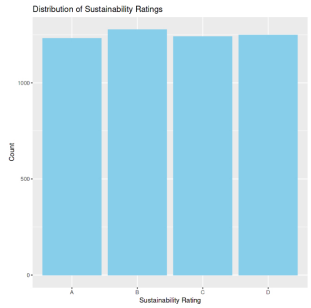
Brand_ID: 0 **Brand_Name:** 0 **Country:** 0 **Year:** 0 **Sustainability_Rating:** 0 **Material_Type:** 0 **Eco_Friendly_Manufacturing:** 0 **Carbon_Footprint_MT:** 0 **Water_Usage_Liters:** 0 **Waste_Production_KG:** 0 **Recycling_Programs:** 0 **Product_Lines:** 0 **Average_Price_USD:** 0 **Market_Trend:** 0 **Certifications:** 0

[13]:
```
# Clean the data if necessary
data <- na.omit(data)  # Remove rows with missing values
```

Exploratory Data Analysis (EDA)

> a) Sustainability Ratings Distribution

[14]:
```r
# Count of sustainability ratings
ggplot(data, aes(x = Sustainability_Rating)) +
  geom_bar(fill = "skyblue") +
  ggtitle("Distribution of Sustainability Ratings") +
  xlab("Sustainability Rating") +
  ylab("Count")
```



We can observe that the distribution is fairly even across all rating categories. There are approximately the same number of items with ratings A, B, C, and D.

> b) Country-wise Analysis of Sustainability Ratings

[15]:
```r
# Sustainability ratings by country
country_ratings <- data %>%
  group_by(Country, Sustainability_Rating) %>%
  summarise(Count = n())

ggplot(country_ratings, aes(x = Country, y = Count, fill = Sustainability_Rating)) +
  geom_bar(stat = "identity", position = "dodge") +
  ggtitle("Sustainability Ratings by Country") +
  xlab("Country") +
  ylab("Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
`summarise()` has grouped output by 'Country'. You can override using the
`.groups` argument.
```

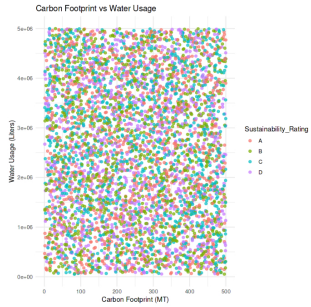
The countries included are Australia, Brazil, China, France, Germany, India, Italy, Japan, UK, and USA. The sustainability ratings are categorized into four groups: A, B, C, and D. The height of each bar represents the number of items in that country that fall into a particular rating category.

We can observe that:

1. Rating A is the most common rating for most countries, with the exception of Germany and Italy.
2. Rating B is also relatively common in most countries.
3. Rating C is less common than A and B for most countries.
4. Rating D is the least common rating for all countries.

> c) Carbon Footprint and Water Usage

[16]:
```r
# Scatter plot of Carbon Footprint vs Water Usage
ggplot(data, aes(x = Carbon_Footprint_MT, y = Water_Usage_Liters, color = Sustainability_Rating)) +
  geom_point(alpha = 0.7) +
  ggtitle("Carbon Footprint vs Water Usage") +
  xlab("Carbon Footprint (MT)") +
  ylab("Water Usage (Liters)") +
  theme_minimal()
```



The plot displays the relationship between carbon footprint and water usage for items with different sustainability ratings (A, B, C, and D). Each dot represents an item, and its position on the graph indicates its carbon footprint (x-axis) and water usage (y-axis). The color of the dot corresponds to the item's sustainability rating.

Here are some observations from the plot:

No clear pattern: There doesn't seem to be a strong correlation between carbon footprint and water usage. The dots are scattered across the graph, indicating that items with high carbon footprints can have both high and low water usage, and vice versa. Overlap in ratings: Items with different sustainability ratings are clustered together in some areas of the graph, suggesting that carbon footprint and water usage alone might not be sufficient to distinguish between the ratings.

d) Average Price and Market Trend

[17]:
```
# Boxplot of Average Price by Market Trend
ggplot(data, aes(x = Market_Trend, y = Average_Price_USD, fill = Market_Trend)) +
  geom_boxplot() +
  ggtitle("Average Price by Market Trend") +
  xlab("Market Trend") +
  ylab("Average Price (USD)")
```

Overall Observations:

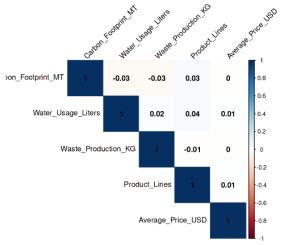
The median prices for all three market trends are relatively close to each other, suggesting that the average price isn't drastically different across these trends. The boxes representing the interquartile range (IQR) are roughly similar in size, indicating that the spread of prices within each trend is comparable. There are outliers present in all three trends, represented by the points beyond the whiskers. Specific Observations:

- Declining Trend: The median price for declining products is slightly higher than the other two trends. There's a wider range of prices, as indicated by the longer box and whiskers.
- Growing Trend: The median price for growing products is slightly lower than the declining trend but similar to the stable trend. The distribution is relatively symmetrical.
- Stable Trend: The median price for stable products is similar to the growing trend. The distribution is also relatively symmetrical, with a slightly narrower range than the declining trend. Key Takeaway:

While the average prices across the three market trends are not significantly different, there are variations in the spread of prices within each trend. Declining products tend to have a wider range of prices compared to growing and stable products.

Correlation Analysis

[18]:
```
# Subset numerical columns
numerical_data <- data %>%
  select(Carbon_Footprint_MT, Water_Usage_Liters, Waste_Production_KG, Product_Lines, Average_Price_USD)

# Calculate correlation matrix
cor_matrix <- cor(numerical_data)

# Plot the correlation matrix
corrplot(cor_matrix, method = "color", type = "upper",
         tl.col = "black", tl.srt = 45, addCoef.col = "black")
```



Logistic Regression for Predictive Modeling

a) Prepare the Data

[19]:
```
# Convert sustainability rating to binary: 1 = High (A/B), 0 = Low (C/D)
data$Sustainability_Binary <- ifelse(data$Sustainability_Rating %in% c("A", "B"), 1, 0)

# Logistic Regression
model <- glm(Sustainability_Binary ~ Carbon_Footprint_MT + Water_Usage_Liters +
               Waste_Production_KG + Product_Lines + Average_Price_USD,
             data = data, family = binomial)

# Summary of the model
summary(model)
```

```
Call:
glm(formula = Sustainability_Binary ~ Carbon_Footprint_MT + Water_Usage_Liters +
    Waste_Production_KG + Product_Lines + Average_Price_USD,
    family = binomial, data = data)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          1.109e-01  1.164e-01   0.952    0.341
Carbon_Footprint_MT  1.046e-04  1.984e-04   0.527    0.598
Water_Usage_Liters  -3.315e-09  1.982e-08  -0.167    0.867
Waste_Production_KG -1.341e-06  9.850e-07  -1.362    0.173
Product_Lines       -2.498e-03  4.964e-03  -0.503    0.615
Average_Price_USD   -1.088e-04  2.055e-04  -0.530    0.596

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6931.4  on 4999  degrees of freedom
Residual deviance: 6928.7  on 4994  degrees of freedom
AIC: 6940.7

Number of Fisher Scoring iterations: 3
```

- None of the predictors (e.g., carbon footprint, water usage, waste production, product lines, and price) significantly influence the likelihood of a high sustainability rating.
- The p-values for all coefficients are greater than 0.05, meaning there is no strong evidence that these variables impact sustainability ratings.
- The model has very little predictive power as indicated by the minimal change in deviance (from 6931.4 to 6928.7) and a high AIC value (6940.7).

b) Model Predictions and Performance

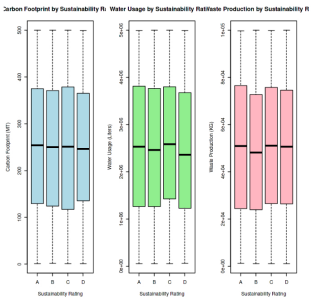
[20]:
```r
# Predict the probability of high sustainability rating
data$Prediction_Prob <- predict(model, type = "response")

# Set threshold for classification
data$Prediction <- ifelse(data$Prediction_Prob > 0.5, 1, 0)

# Confusion Matrix
table(Predicted = data$Prediction, Actual = data$Sustainability_Binary)
```

```
         Actual
Predicted    0    1
        0 1114 1103
        1 1377 1406
```

Analysis of Environmental Impact vs Sustainability Ratings

a) Carbon Footprint, Water Usage, and Waste Production by Rating

[21]:
```r
# Boxplots for environmental metrics by Sustainability Rating
par(mfrow = c(1, 3))  # Arrange plots side by side

# Carbon Footprint
boxplot(Carbon_Footprint_MT ~ Sustainability_Rating, data = data,
        main = "Carbon Footprint by Sustainability Rating",
        xlab = "Sustainability Rating", ylab = "Carbon Footprint (MT)", col = "lightblue")

# Water Usage
boxplot(Water_Usage_Liters ~ Sustainability_Rating, data = data,
        main = "Water Usage by Sustainability Rating",
        xlab = "Sustainability Rating", ylab = "Water Usage (Liters)", col = "lightgreen")

# Waste Production
boxplot(Waste_Production_KG ~ Sustainability_Rating, data = data,
        main = "Waste Production by Sustainability Rating",
        xlab = "Sustainability Rating", ylab = "Waste Production (KG)", col = "lightpink")
```



Looking at the carbon footprint plot, we observe that the median carbon footprint is relatively similar across all four ratings. However, there is a wider spread in the carbon footprint for ratings B and C compared to A and D.

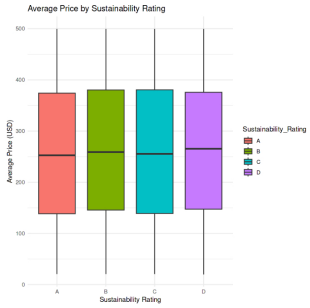
Similarly, in the water usage plot, the median water usage is consistent across all ratings. The spread of water usage is also relatively similar for all ratings.

In the waste production plot, the median waste production is again similar across all ratings. However, the spread of waste production is wider for ratings B and C compared to A and D.

Brand Price Analysis

Relationship Between Price and Sustainability Rating

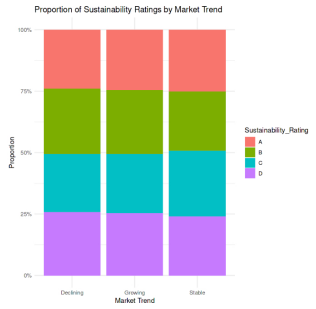
[22]:
```r
# Boxplot for Average Price by Sustainability Rating
ggplot(data, aes(x = Sustainability_Rating, y = Average_Price_USD, fill = Sustainability_Rating)) +
  geom_boxplot() +
  ggtitle("Average Price by Sustainability Rating") +
  xlab("Sustainability Rating") +
  ylab("Average Price (USD)") +
  theme_minimal()
```



Market Trend vs Sustainability Ratings

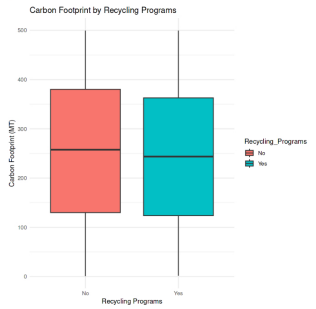
[24]:
```r
# Proportion of Sustainability Ratings within Market Trends
market_trend_ratings <- data %>%
  group_by(Market_Trend, Sustainability_Rating) %>%
  summarise(Count = n(), .groups = "drop")

ggplot(market_trend_ratings, aes(x = Market_Trend, y = Count, fill = Sustainability_Rating)) +
```

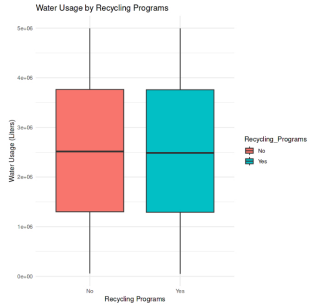
```
  geom_bar(stat = "identity", position = "fill") +
  ggtitle("Proportion of Sustainability Ratings by Market Trend") +
  xlab("Market Trend") +
  ylab("Proportion") +
  scale_y_continuous(labels = scales::percent) +
  theme_minimal()
```



Recycling Programs and Environmental Metrics

[25]:
```
# Compare Carbon Footprint for brands with and without recycling programs
ggplot(data, aes(x = Recycling_Programs, y = Carbon_Footprint_MT, fill = Recycling_Programs)) +
  geom_boxplot() +
  ggtitle("Carbon Footprint by Recycling Programs") +
  xlab("Recycling Programs") +
  ylab("Carbon Footprint (MT)") +
  theme_minimal()

# Water Usage
ggplot(data, aes(x = Recycling_Programs, y = Water_Usage_Liters, fill = Recycling_Programs)) +
  geom_boxplot() +
  ggtitle("Water Usage by Recycling Programs") +
  xlab("Recycling Programs") +
  ylab("Water Usage (Liters)") +
  theme_minimal()
```





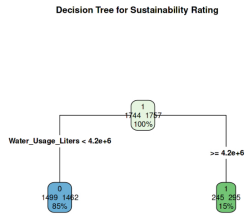
[26]:
```
# Load required libraries
library(rpart)        # For Decision Trees
library(rpart.plot)   # To plot Decision Trees
library(randomForest) # For Random Forests
library(caret)        # For model evaluation

# Convert Sustainability_Binary to a factor
data$Sustainability_Binary <- as.factor(data$Sustainability_Binary)

# Split data into training and testing sets
set.seed(123)  # For reproducibility
trainIndex <- createDataPartition(data$Sustainability_Binary, p = 0.7, list = FALSE)
train_data <- data[trainIndex, ]
test_data <- data[-trainIndex, ]
```

```
randomForest 4.7-1.1

Type rfNews() to see new features/changes/bug fixes.


Attaching package: ‘randomForest’


The following object is masked from ‘package:ggplot2’:

    margin


The following object is masked from ‘package:dplyr’:

    combine


Loading required package: lattice


Attaching package: ‘caret’
```

```
The following object is masked from ‘package:httr’:

    progress
```

Decision Tree Model

[27]:
```
# Train a Decision Tree
decision_tree <- rpart(Sustainability_Binary ~ Carbon_Footprint_MT + Water_Usage_Liters +
                         Waste_Production_KG + Product_Lines + Average_Price_USD,
                       data = train_data, method = "class", cp = 0.01)

# Plot the Decision Tree
rpart.plot(decision_tree, type = 4, extra = 101, main = "Decision Tree for Sustainability Rating")
```



[28]:
```
# Predict on test data
pred_tree <- predict(decision_tree, test_data, type = "class")

# Confusion Matrix
confusionMatrix(pred_tree, test_data$Sustainability_Binary)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 637 642
         1 110 110

               Accuracy : 0.4983
                 95% CI : (0.4727, 0.524)
    No Information Rate : 0.5017
    P-Value [Acc > NIR] : 0.6118

                  Kappa : -0.001

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8527
            Specificity : 0.1463
         Pos Pred Value : 0.4980
         Neg Pred Value : 0.5000
             Prevalence : 0.4983
         Detection Rate : 0.4249
   Detection Prevalence : 0.8532
      Balanced Accuracy : 0.4995

       'Positive' Class : 0
```

Random Forest Model

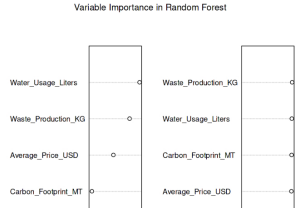
[29]:
```
# Train a Random Forest model
set.seed(123)  # For reproducibility
random_forest <- randomForest(Sustainability_Binary ~ Carbon_Footprint_MT + Water_Usage_Liters +
                                Waste_Production_KG + Product_Lines + Average_Price_USD,
                              data = train_data, importance = TRUE, ntree = 500)

# Print the model summary
print(random_forest)
```

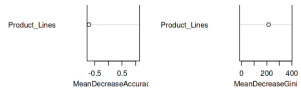
```
Call:
 randomForest(formula = Sustainability_Binary ~ Carbon_Footprint_MT +      Water_Usage_Liters + Waste_Production_KG + Product_Lines +      Average_Price_USD, data = train_data, importance = TRUE,      ntree = 500)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 50.01%
Confusion matrix:
    0   1 class.error
0 859 885   0.5074541
1 866 891   0.4928856
```

[30]:
```
# Variable Importance Plot
importance(random_forest)
varImpPlot(random_forest, main = "Variable Importance in Random Forest")
```

A matrix: 5 × 4 of type dbl

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
Carbon_Footprint_MT	-1.3067924	0.42781950	-0.6777185	382.7996
Water_Usage_Liters	-1.1520508	2.60937855	1.0673933	384.1663
Waste_Production_KG	0.9190482	0.09270344	0.7024959	386.7296
Product_Lines	-1.9028337	0.83848584	-0.7122584	215.2676
Average_Price_USD	0.1524856	0.03245839	0.1095578	380.9675





[31]:
```r
# Predict on test data
pred_rf <- predict(random_forest, test_data)

# Confusion Matrix
confusionMatrix(pred_rf, test_data$Sustainability_Binary)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 350 405
         1 397 347

               Accuracy : 0.465
                 95% CI : (0.4395, 0.4906)
    No Information Rate : 0.5017
    P-Value [Acc > NIR] : 0.9979

                  Kappa : -0.07

 Mcnemar's Test P-Value : 0.8048

            Sensitivity : 0.4685
            Specificity : 0.4614
         Pos Pred Value : 0.4636
         Neg Pred Value : 0.4664
             Prevalence : 0.4983
         Detection Rate : 0.2335
   Detection Prevalence : 0.5037
      Balanced Accuracy : 0.4650

       'Positive' Class : 0
```

Compare Model Performance

```r
# Compare Accuracy of Decision Tree and Random Forest
accuracy_tree <- confusionMatrix(pred_tree, test_data$Sustainability_Binary)$overall['Accuracy']
accuracy_rf <- confusionMatrix(pred_rf, test_data$Sustainability_Binary)$overall['Accuracy']

# Print Accuracy
cat("Decision Tree Accuracy:", accuracy_tree, "\n")
cat("Random Forest Accuracy:", accuracy_rf, "\n")
```

```
Decision Tree Accuracy: 0.4983322
Random Forest Accuracy: 0.4649767
```

+ Code + Markdown

Both models have relatively low accuracy, indicating that they are not very effective in predicting the target variable. The Decision Tree model performs slightly better than the Random Forest model in this case. Further analysis and improvement of the models may be necessary to achieve better performance.