Statistical analysis.R

2024-04-16

# Load required libraries  
library(dplyr)

##   
## 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

library(ggplot2)  
library(caret)

## Loading required package: lattice

library(randomForest)

## randomForest 4.7-1.1

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

##   
## Attaching package: 'randomForest'

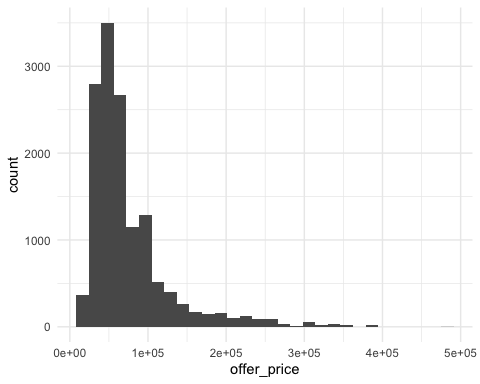
## The following object is masked from 'package:ggplot2':  
##   
## margin

## The following object is masked from 'package:dplyr':  
##   
## combine

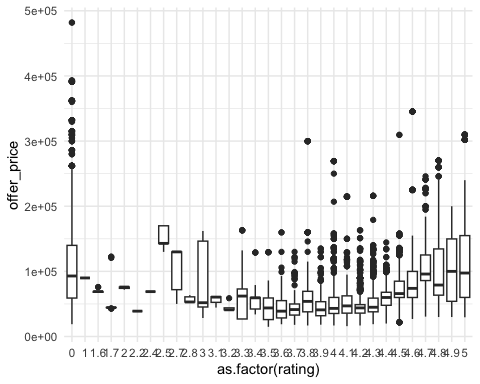
# Read the data  
laptops\_data <- read.csv("Laptop\_Merged(cleaned).csv")  
  
# Summary statistics  
summary(laptops\_data)

## u\_id name offer\_price original\_price   
## Length:14040 Length:14040 Min. : 14890 Min. : 18890   
## Class :character Class :character 1st Qu.: 41900 1st Qu.: 59054   
## Mode :character Mode :character Median : 58990 Median : 76990   
## Mean : 74545 Mean : 95740   
## 3rd Qu.: 89890 3rd Qu.:112608   
## Max. :481990 Max. :481990   
## off\_now total\_ratings total\_reviews rating   
## Length:14040 Min. : 0.0 Min. : 0.00 Min. :0.000   
## Class :character 1st Qu.: 2.0 1st Qu.: 0.00 1st Qu.:3.000   
## Mode :character Median : 43.0 Median : 5.00 Median :4.200   
## Mean : 450.6 Mean : 56.78 Mean :3.215   
## 3rd Qu.: 261.0 3rd Qu.: 30.00 3rd Qu.:4.400   
## Max. :30936.0 Max. :3710.00 Max. :5.000   
## description item\_link created\_at   
## Length:14040 Length:14040 Length:14040   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##

# Histograms for numerical variables  
ggplot(laptops\_data, aes(x = offer\_price)) + geom\_histogram(bins = 30) + theme\_minimal()



# Box plots for comparing price distributions  
ggplot(laptops\_data, aes(x = as.factor(rating), y = offer\_price)) + geom\_boxplot() + theme\_minimal()



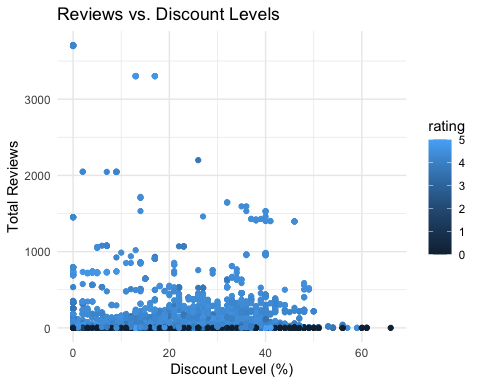
# Correlation matrix  
correlations <- cor(laptops\_data %>% select(offer\_price, original\_price, total\_ratings, total\_reviews, rating), use = "complete.obs")  
  
# Heatmap  
library(corrplot)

## corrplot 0.92 loaded

corrplot(correlations, method = "circle")



# Convert discount levels from percentage to numerical  
laptops\_data$discount\_numeric <- as.numeric(sub("% off", "", laptops\_data$off\_now))  
  
  
# Create scatter plot for Number of Reviews vs. Discount with color based on Rating  
ggplot(laptops\_data, aes(x = discount\_numeric, y = total\_reviews, color = rating)) +  
 geom\_point() +  
 labs(x = "Discount Level (%)", y = "Total Reviews", title = "Reviews vs. Discount Levels") +  
 theme\_minimal()



# Split the data into training and testing sets  
set.seed(123)  
train\_index <- createDataPartition(laptops\_data$offer\_price, p = 0.8, list = FALSE)  
train\_data <- laptops\_data[train\_index, ]  
test\_data <- laptops\_data[-train\_index, ]  
  
# Build the random forest model  
rf\_model <- randomForest(offer\_price ~ original\_price + off\_now + total\_ratings + total\_reviews + rating,  
 data = train\_data, importance = TRUE)  
  
# Get the column names of the importance data frame  
importance\_cols <- names(rf\_model$importance)  
  
# Determine the column name for importance values  
importance\_col\_name <- ifelse("IncNodePurity" %in% importance\_cols, "IncNodePurity", "%IncMSE")  
  
# Create a data frame with variable importance  
importance\_df <- data.frame(Variable = rownames(rf\_model$importance),  
 Importance = rf\_model$importance[, importance\_col\_name])  
  
# Plot variable importance››  
ggplot(importance\_df, aes(x = Importance, y = reorder(Variable, Importance))) +  
 geom\_col(fill = "steelblue") +  
 labs(x = paste0("Importance (", importance\_col\_name, ")"), y = "Variable") +  
 ggtitle("Variable Importance Plot") +  
 theme\_minimal()

