Exploring Factors Affecting Clothes Sales

Final Project–Group 8 XinyuXie TianyuDuan

Online shopping has become a major channel for selling clothing items. It is becoming increasingly important to help merchants as well as designers to understand what factors influence the main factors of affordable clothing.The purpose of this project is to analyze the consumption trends of clothing and identify the main factors that influence customer purchasing behavior or purchasing preferences. By utilizing all the explored data sets, this study will investigate the relationship between customer age, merchant discounts, and the amount of money spent by the purchasing customer, providing businesses and beneficiaries with better pricing and customer attraction strategies.

Data Description

Primary Dataset: Kaggle – Customer Shopping Trends

URL:***https://www.kaggle.com/datasets/bhadramohit/customer-shopping-latest-trends-dataset***

Content: The dataset offers a comprehensive view of consumer shopping trends, aiming to uncover patterns and behaviors in retail purchasing. It contains detailed transactional data across various product categories, customer demographics, and purchase channels.Case is a ***Customer***. The following data are included:

Transaction Details: Purchase date, transaction value, product category, and payment method.

Customer Information: Age group, gender, location, and loyalty status.

Shopping Behavior: Frequency of purchases, average spend per transaction, and seasonal trends.

Secondary Dataset: library(Stat2Data) Data(Clothing)

Provides additional data to backup the findings of the primary dataset.

Content: This dataset represents a random sample of 60 customers from a large clothing retailer. Data on 60 customers at a clothing retailer. Case is a ***Customer***.The following data are included:

ID Case ID
Amount Net dollar amount spent by customers in their latest purchase from this retailer
Recency Number of months since the last purchase
Freq12 Number of purchases in the last 12 months
Dollar12 Dollar amount of purchases in the last 12 months
Freq24 Number of purchases in the last 24 months
Dollar24 Dollar amount of purchases in the last 24 months
Card 1 for customers who have a private-label credit card with the retailer, 0 if not

Data Cleaning

First we check the our main dataset, and we are lucky that it is a tidy data. Our plan for data processing is to first filter out the useful variables, and then to analyze the consumption for the type of clothes.The customer number has no real meaning, in fact we want to analyze the consumption behavior, not the product or the consumption process. So we can drop the following variables "CustomerID, Size, Color, Payment Method, Shipping Type, Preferred Payment Method".Let's look at the remaining variables in table 1.

Table 1: Retained Variables in clothing_data

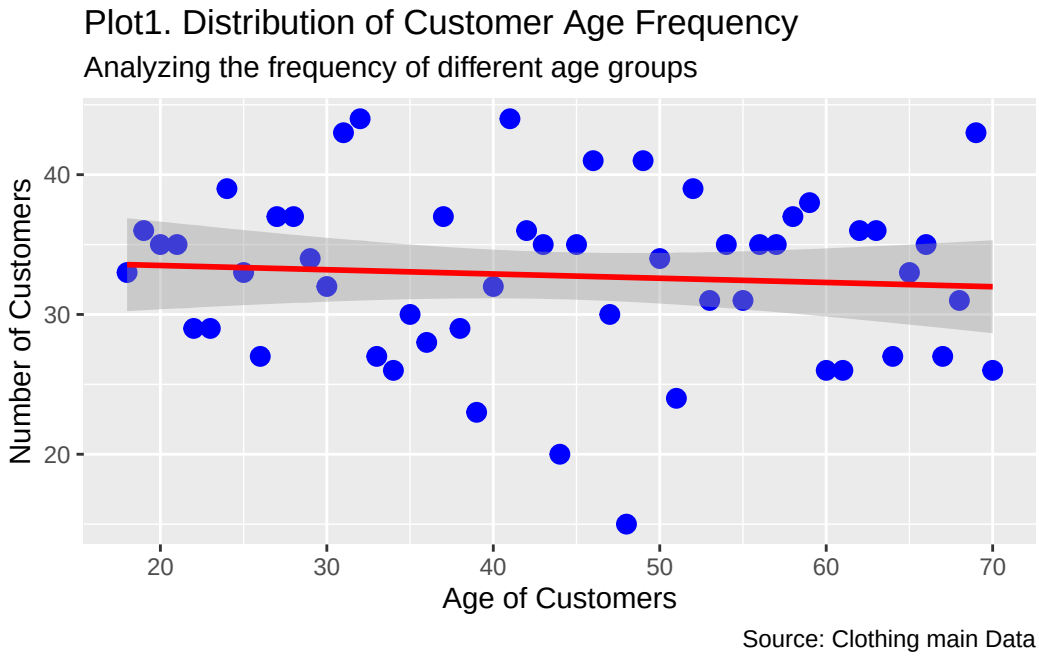
Variable Name
Age
Gender
Category
Purchase.Amount..USD.
Season
Review.Rating
Subscription.Status
Discount.Applied
Promo.Code.Used
Previous.Purchases
Frequency.of.Purchases

Descriptive Statistics

We need to look at our sample, our case is every customer who has ever spent money, they are important for us to analyze.We first want our sample to be representative of the majority of the consumer population, which we can see well using the variable age, so we will analyze age first. The average spend is used to look at the average unit price and the average rating and standard deviation are used to see if the product has a consistent level of satisfaction among the population.

Table 2: Summary Statistics for Clothing Data

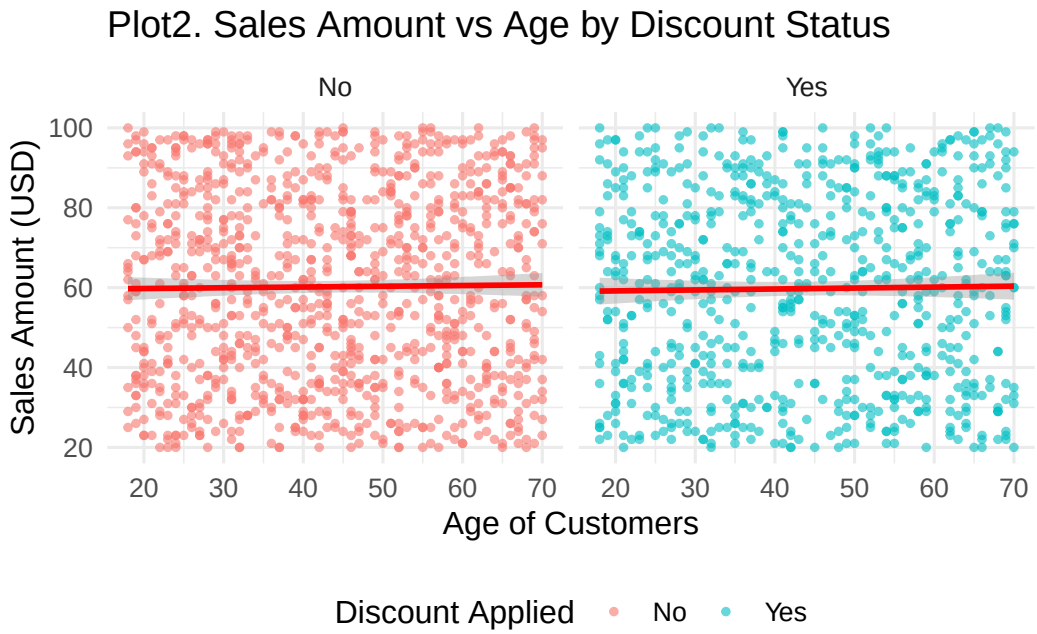
Total Customers	Average Age	Median Age	Average Purchase ($)	Average Rating	Rating Std Dev
1737	43.78296	43	60.02533	3.723143	0.717671



According to Table 2, the data contains 1737 customers, the sample size is large enough for statistical analysis and the results are representative.The average age is 43.78 years old, and the median is 43 years old, which means the age distribution is more symmetrical. The average purchase amount is $60.03, which indicates a stable level of single purchase. According to Tables 2 and Plot 1, we can conclude that the number of customers is evenly distributed across all age groups, with no upward or downward trend with age.The correlation between the independent variable and the dependent variable is very weak and may be close to 0. This represents that customers regardless of age groups show similar interest in purchasing clothes online, which is in line with what we know about the context that e-commerce is an important

clothing sales channel nowadays. And our sample is representative. We are confident to proceed to the next step of the analysis.

Next we argue that discounts are an important variable that affects sales Because in the past there was the idea that discounts would encourage people to spend more, and we can take whether or not we use discounts and analyze the impact on sales.

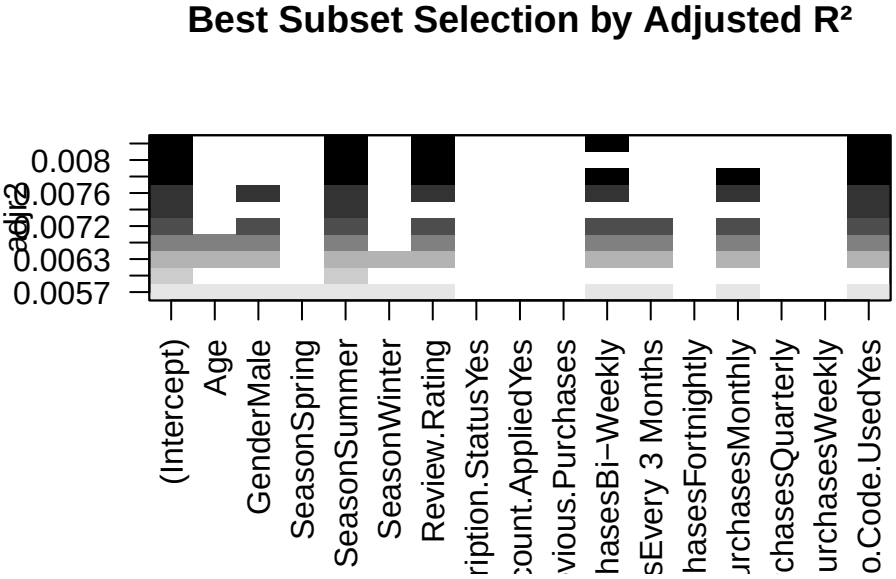


Based on plot 2 the lack of a significant linear relationship between sales and age, with or without the use of discounts, has left us in a quandary, and we should conduct further exploratory analyses.

Multiple Linear Regression Modeling Analysis

Exploratory analysis can bring us more information that we should explore further, we can use multiple linear regression to find out what are the important factors that affect sales. The use of multiple linear regression not only clarifies the independent contributions of the data we want to explore, but also reveals the interactions between them, thus providing data to support the optimization of discounting strategies. With this approach we are able to make more accurate business decisions based on actual data and improve sales results.

`Reordering variables and trying again:`



Best subset method Successfully screened out key variables affecting sales: 'summer season', 'rating scores', 'bi-weekly purchases' use of code' These variables have good explanatory power in the model for thestrong support for further modeling. Although we can tell from Table 3 that none of the variables are significant except for summer, we can nevertheless compare the best variables except for summer. Our regression equation is as follows:

$$\hat{PurchaseAmount} = 57.995 - 4.774 \times Summer + 1.0943 \times Rating + 1.0943 \times BuyBiWeekly - 0.2955 \times Code$$

Table 3: Regression Coefficients and P-values

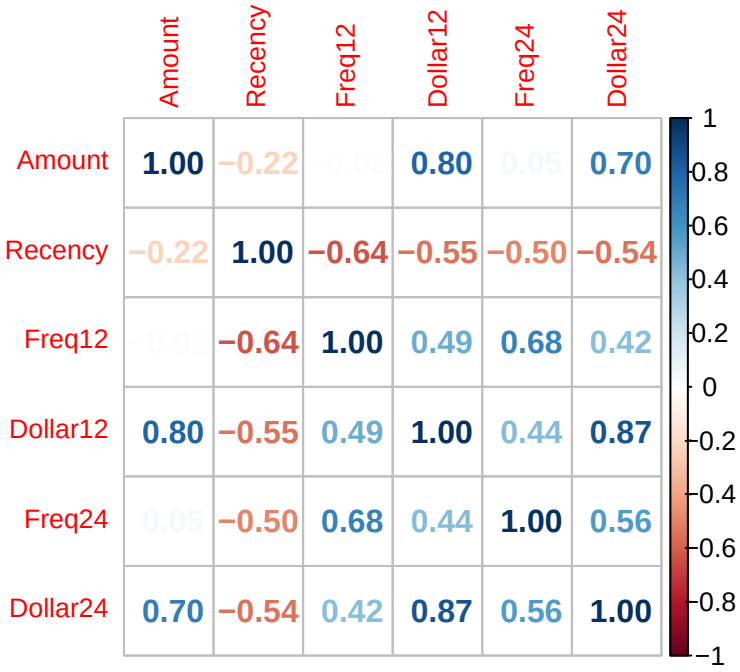
Variable	Beta Coefficient	P-value
(Intercept)	57.9951	0.0000
Review.Rating	1.0943	0.1694
SeasonSpring	-0.3530	0.8261
SeasonSummer	-4.7748	0.0038
SeasonWinter	-0.5057	0.7537
Frequency.of.PurchasesBi-Weekly	-0.6294	0.7663
Frequency.of.PurchasesEvery 3 Months	0.1427	0.9453
Frequency.of.PurchasesFortnightly	-1.2156	0.5758
Frequency.of.PurchasesMonthly	0.2611	0.9002
Frequency.of.PurchasesQuarterly	0.4982	0.8153
Frequency.of.PurchasesWeekly	-3.2983	0.1210
Promo.Code.UsedYes	-0.2955	0.7988

Supporting data validation

To further investigate the impact of repeat purchases and discount usage on sales, we will utilize a second dataset for validation and comparison. This dataset provides an opportunity to analyze whether the patterns observed in the first dataset are consistent across environments or customer segments. By examining the relationship between repeat purchase frequency and discount usage and sales performance in the second dataset.This approach not only enhances the robustness of our conclusions, but also provides deeper insight into the drivers of sales and ensures that the trends observed are not specific to the dataset, but are indicators of broader consumer behavior.

Data Cleaning

First of all that we have variables that we don't need to ID as well as we should remove the extreme values before we proceed with the analysis we need to introduce a correlation matrix to help us understand whether the variables are independent from each other or not. Afterwards making a box plot to analyze if we card is an important variable.



Scatterplot Matrix of Numerical Variables

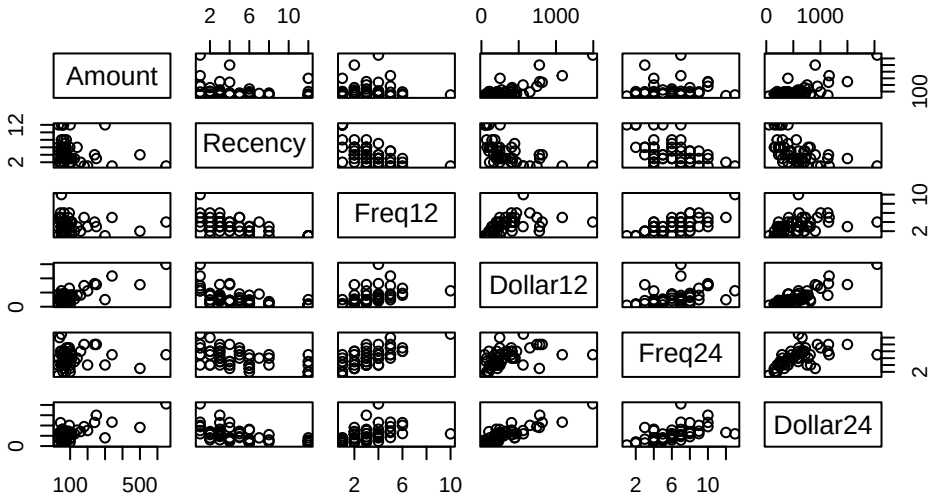
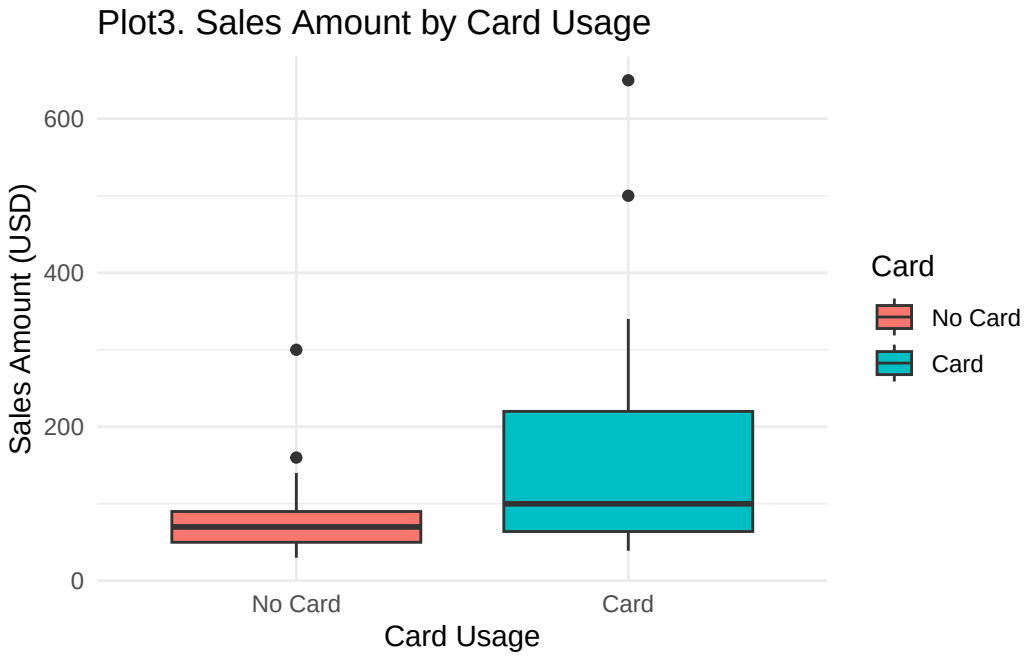


Table 4: ANOVA Results for Sales Amount by Card Usage

Source	Df	F_Value	P_Value
Card	1	8.8282	0.0046
Residuals	50	NA	NA



Based on Plot 3 we find that the variance of the use card is significantly and slightly larger than the No card, but we cannot tell if the means are the same.We decided to use ANOVA to determine this.
Based on Table 5, we can conclude that there is a difference in the mean value of sales with or without cards, albeit slightly violating the basic assumptions of the ANOVA, although this is not in our main considerations.

The correlation coefficients of Amount with Dollar12 and Dollar24 are close to 0.9, indicating that the amount of money spent in the past 12 and 24 months has a significant effect on the amount of money spent currently. The positive correlation between Freq12 and Freq24 is high, reflecting the consistency in the number of purchases made in the past 12 and 24 months. From the scatter plot of Amount with other variables, it can be seen that Amount shows a strong positive correlation distribution with Dollar12 and Dollar24, which is consistent with the results of the correlation matrix in the first graph. In addition, the scatter plots of Freq12 and Freq24 are densely distributed, further indicating a strong correlation between the two.

We can find the VIF values of **Dollar12** and **Dollar24** in Table 6, and we usually think that

Table 5: Variance Inflation Factor (VIF) Table

Variable	VIF
Recency	2.0569
Freq12	2.9538
Dollar12	5.1902
Freq24	2.7677
Dollar24	6.1348
Card	1.6724

Table 6: Regression Coefficients and P-values for Best Subset Model

Variable	Beta Coefficient	P-value
(Intercept)	79.3224	0.0029
CardCard	19.8729	0.4338
Dollar24	0.2902	0.0000
Freq24	-23.0330	0.0000

there is some covariance if the VIF is greater than 5, so we should be careful to use these two variables for prediction.

We try to process it to avoid multicollinearity by turning it into a variable.

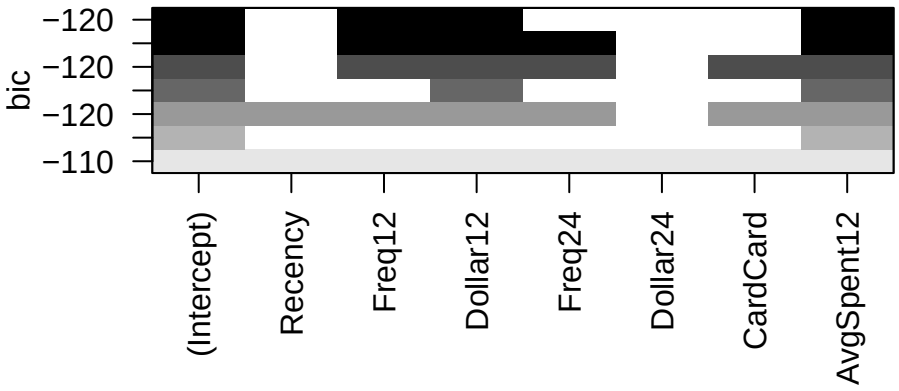
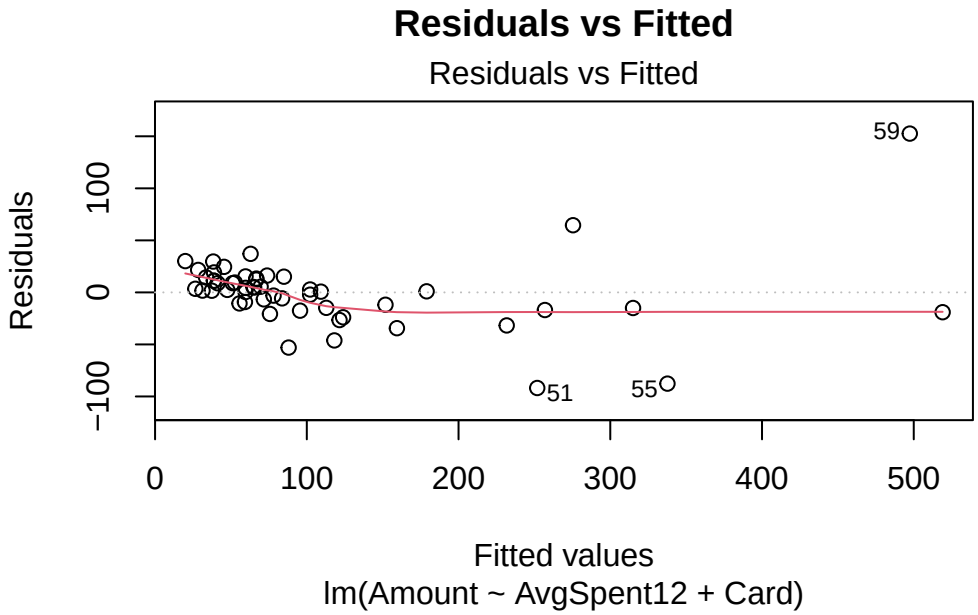
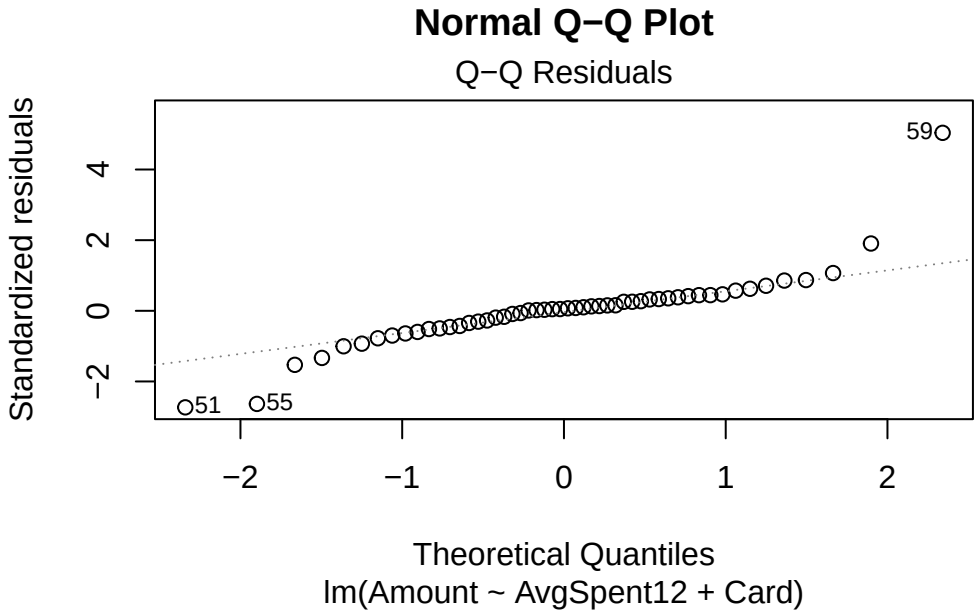


Table 7: Regression Coefficients and P-values for Best Subset Model

Variable	Beta Coefficient	P-value
(Intercept)	-39.6651	0.0000
AvgSpent12	1.4187	0.0000
CardCard	7.5408	0.4943

We understand that the average spend comes from the fact that we decided not to introduce frequency and money to avoid multicollinearity, but we decided to introduce card because he is the highest influencing variable besides the two mentioned above.





Let us evaluate the underlying assumptions of the mode:

Linearity: The fact that the reference line of the Residual vs fitted plot is approximate linear relationships distributed at both ends can be said not to violate the linear relationship.

Independence: Assuming what we have is a random sample of people it will be reasonable to assume errors are independent.

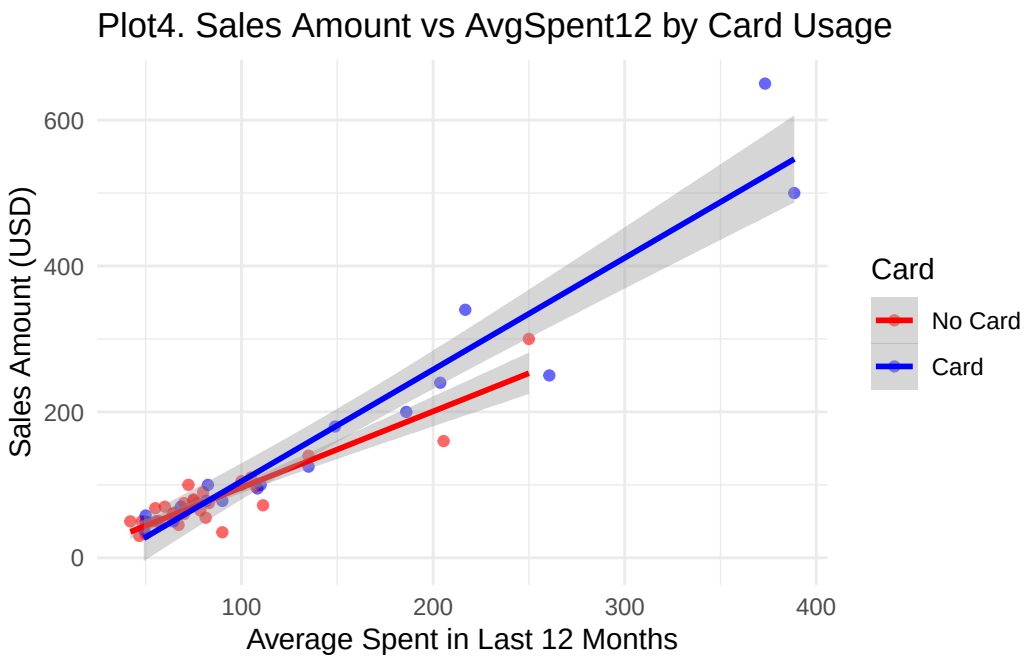
Normality: The Q-Q plot shows that most points lie on the reference line, except for some deviations in the tails particularly at the upper right. This suggests that while the majority of the residuals follow a normal distribution, there may be some outliers or heavy-tailed behavior. The normality assumption is mostly satisfied.

Equal Variance: It appears that the variance is a little bit fanning out and there seems to be an increase in variance, it's possible having a heteroskedasticity problem.

This is one of the better predictive models although the card is not significant, but the adjusted R-squared is an impressive 0.9064.

We can learn from Table 8 that Our model_2 is:

$$\hat{Amount} = -39.6651 + 1.4187 \times Avgspent12 + 1.0943 \times Card$$



According to the fitting diagram Plot 4. Customers who use cards contribute more to sales and have higher average spending.Customers who do not use their cards have a smaller increase in sales, even if they have higher average expenditures.But average spend is clearly an excellent predictor variable anyway.

Discussions

Possible measures

The results of this study show that there is a positive correlation between "Repeat Purchase Behavior" and "Average Spending" and sales, which not only reflects the value of loyal customers, but also the "Loyalty Incentive" behind the data " mechanism. Similar to credit cards that incentivize spending through points and cashback, Code's negative coefficient reveals a negative correlation between promo code users and spending power. We hypothesize that this contrasts with the high spending power exhibited by credit card users (Code) becoming a mapping of the lower spending power customer segment, while (Card) reflects the characteristics of the high spending power customer segment. The effect of this loyalty incentive comes from customers' continued spending habits and trust in the brand. Enterprises can identify these customers as high-value customers and further increase their consumption frequency and single-use spending through precision marketing and personalized offer strategies. In addition, by combining payment methods such as credit cards issued by the company itself with loyalty

incentives, a multi-level customer retention system can be formed to promote sustained sales growth.

FAIR and CARE

Fair

Findable: In the data preparation phase, we make the data well findable through clear variable definitions and appropriate naming conventions.The clarity of variables makes the analysis process highly reproducible and transparent.

Accessible: During the analysis, We chose to use all publicly available datasets, steps that ensured that the data could be easily accessed and utilized by subsequent studies.

Interoperable: by standardizing the data format, we ensured that the data could be passed seamlessly between analysts.

Reusable: The final results clearly documented and summarized, providing high-value information for further marketing optimization.This reuse of data supports companies in implementing similar analyses at different times or regions.

Care

Collective Benefit: The core objective of the analysis is to explore key factors that influence clothing sales, such as seasonal variations and customer loyalty.The results are not only important for the optimization of companies' marketing strategies, but also provide academics with lessons to be learned in the field of consumer behavior.

Authority to Control: In the analysis, variables such as Card and Promo Code reflect customers' autonomy of choice.These variables reflect how customers participate in purchasing decisions through payment methods and promotional tools, emphasizing the importance of firms respecting customer choice in data collection and use.

Responsibility: Outliers (The extreme value of 1506000 in sales) were excluded and the data was processed in a reasonable manner to avoid misleading the results of the analysis with unreasonable data and to ensure the fairness of the results.

Ethics: The analysis process emphasizes respect for client privacy and transparency of data sources.For both datasets, their behavior was analyzed only by aggregating data rather than tracking personal information, reflecting the spirit of data ethics in open science.

Conclusion

By analyzing the two datasets, we can draw the following conclusions:

In the first dataset, season is an important factor affecting sales, there is a significant difference between different seasons on sales performance, in the summer people are more reluctant to buy which is the off-season, you can reduce investment at this time.Repeat purchase has a significant positive effect on sales, indicating that the consumption behavior of loyal customers contributes more to sales.

The second dataset, there is an interaction between repeat purchase and amount spent, further confirming the importance of repeat purchase behavior.Repeat purchase behavior, as a core variable, has a stable and positive impact on sales increase.

Enterprises can combine seasonal promotion strategies, loyal customer maintenance, and credit card incentives to accurately increase the frequency of customer purchases and average expenditures to achieve sales growth.

Reference

Bhadra, Mohit. *Customer Shopping (Latest Trends) Dataset.* Kaggle, 23 Nov. 2024, https://www.kaggle.com/datasets/bhadramohit/customer-shopping-latest-trends-dataset.

Erickson, Timothy, and Kari Lock Morgan (2022). *Stat2Data: Data Sets for Stat2, Second Edition.* R package version 2.0.0. https://CRAN.R-project.org/package=Stat2Data

```r
#Checking Availability of packages
if (!requireNamespace("dplyr", quietly = TRUE)) install.packages("dplyr")
if (!requireNamespace("ggplot2", quietly = TRUE)) install.packages("ggplot2")
if (!requireNamespace("knitr", quietly = TRUE)) install.packages("knitr")
if (!requireNamespace("kableExtra", quietly = TRUE)) install.packages("kableExtra")
if (!requireNamespace("corrplot", quietly = TRUE)) install.packages("corrplot")
if (!requireNamespace("Stat2Data", quietly = TRUE)) install.packages("Stat2Data")
if (!requireNamespace("GGally", quietly = TRUE)) install.packages("GGally")
if (!requireNamespace("broom", quietly = TRUE)) install.packages("broom")
if (!requireNamespace("leaps", quietly = TRUE)) install.packages("leaps")
if (!requireNamespace("car", quietly = TRUE)) install.packages("car")

#Loading packages
library(dplyr)
library(ggplot2)
library(knitr)
library(kableExtra)
library(corrplot)
library(Stat2Data)
library(GGally)
library(broom)
library(leaps)
library(car)

#Loading data
Shopping_trends <- read.csv("Shopping_trends.csv")


#Delete these columns
clothing_data <- Shopping_trends %>%
  select(
    -`Customer.ID`,
    -Size,
    -Color,
    -`Payment.Method`,
    -`Shipping.Type`,
    -`Preferred.Payment.Method`,
    -`Item.Purchased`,
    -Location
  ) %>%
  filter(Category == "Clothing")
```

```r
#Cheak variable name
#View(clothing_data)

variable_names1 <- data.frame(Variable = colnames(clothing_data))

#Displaying a table using kable
kable(variable_names1,
      col.names = c("Variable Name"),
      caption = "Retained Variables in clothing_data"
)




#Give descriptive statistics
summary_stats <- clothing_data %>%
  summarise(
    total_customers = n(),                          #Total customers
    avg_age = mean(Age, na.rm = TRUE),              #Average age
    median_age = median(Age, na.rm = TRUE),         #Median age
    avg_purchase = mean(Purchase.Amount..USD., na.rm = TRUE),   #Average purchase amount
    avg_rating = mean(Review.Rating, na.rm = TRUE), #Average rating
    sd_rating = sd(Review.Rating, na.rm = TRUE)     #standard deviation of ratings
)

#Provide new readable listings
colnames(summary_stats) <- c(
  "Total Customers",
  "Average Age",
  "Median Age",
  "Average Purchase ($)",
  "Average Rating",
  "Rating Std Dev"
)

#Displaying tables with kable
kable(summary_stats,
      caption = "Summary Statistics for Clothing Data"
) %>%
  kable_styling(
    bootstrap_options = c(
      "striped",
```

```
      "hover",
      "condensed",
      "responsive"
    ),
    full_width = FALSE,
    position = "center",
    font_size = 8
  ) %>%
  row_spec(0, bold = TRUE)

#Grouping of data for graphing purposes
age_counts <- clothing_data %>%
  group_by(Age) %>%
  summarise(count = n()) %>%
  ungroup()

#Plotting scatterplots by age
ggplot(
  age_counts,
  aes(
    x = Age,
    y = count
 )
) +
  geom_point(
    color = "blue",
    size = 3
) +
   geom_smooth(   #Add linear fit line with confidence intervals
     method = "lm",
     color = "red",
     se = TRUE
) +
   labs(          #Add title and tags
    title = "Plot1. Distribution of Customer Age Frequency",
    subtitle = "Analyzing the frequency of different age groups",
    x = "Age of Customers",
    y = "Number of Customers",
    caption = "Source: Clothing main Data"
)
```

```r
#Relationship between the use of discounts and sales
ggplot(data = clothing_data,
       aes(
         x = Age,
         y = Purchase.Amount..USD.,
         color = Discount.Applied
  )
) +
  geom_point(
    alpha = 0.6,
    size = 1
) +
  geom_smooth(
    method = "lm",
    color = "red"
) +      #Add a linear regression fit line
  labs(
    title = "Plot2. Sales Amount vs Age by Discount Status",  #Name it
    x = "Age of Customers",
    y = "Sales Amount (USD)",
    color = "Discount Applied"
) +
  facet_grid(~ Discount.Applied) +                #Grouped by discount
  theme_minimal(base_size = 12) +
  theme(
    legend.position = "bottom"
)

#Elimination of the single level variable 'Category'
clothing_data_clean <- clothing_data[, !(names(clothing_data) %in% "Category")]

#Implementation of the optimal subset method
best_subset_cloth1 <- regsubsets(Purchase.Amount..USD. ~ ., data = clothing_data_clean, nvma

#View Summary of Results
best_summary <- summary(best_subset_cloth1)


#Visualization results

plot(best_subset_cloth1,
```

```
     scale = "adjr2",
     main = "Best Subset Selection by Adjusted R²"
     )
#Fit the final model
model1_1 <- lm(Purchase.Amount..USD. ~ Review.Rating + Season + Review.Rating + Frequency.of

#Fitting model
model1_1 <- lm(Purchase.Amount..USD. ~ Review.Rating + Season + Frequency.of.Purchases + Pro

#Extract model results (Beta coefficients and P-values)
model_results <- tidy(
  model1_1
)  #Use tidy() from the broom package to extract model results

#Select required columns: variable name, Beta coefficient, and P-value
beta_pvalue_table <- model_results[, c("term", "estimate", "p.value")]

#Rename columns for readability
colnames(beta_pvalue_table) <- c(
  "Variable",
  "Beta Coefficient",
  "P-value"
)

#Create table using knitr::kable
kable(beta_pvalue_table,
      format = "latex",
      caption = "Regression Coefficients and P-values",
      digits = 4,            #Retain 4 decimal places
      booktabs = TRUE)
library(Stat2Data)
data("Clothing")
first_rows_table <- kable(
  x = head(Clothing, 5),
  format = "latex",      #Suitable for PDF output in LaTeX format
  caption = "First 5 Rows of Clothing Data",
  booktabs = TRUE,       #Use aesthetically pleasing table lines in LaTeX
  align = "c"
)           #Center-align columns

#Data preprocessing
Clothing <- subset(
```

```
  Clothing,
  select = -ID
) #Remove ID column
Clothing <- subset(
  Clothing,
  Amount != 1506000
) #Remove outliers
Clothing <- subset(
  Clothing,
  Freq12 > 0
) #Remove records with Freq12 of 0
Clothing$Card <- factor(
  Clothing$Card,
  levels = c(0, 1),
  labels = c("No Card", "Card")
) #Convert to factor

#Separate numeric and factor variables
numeric_clothing <- Clothing[, sapply(Clothing, is.numeric)]   #Numeric variables
factor_clothing <- Clothing[, sapply(Clothing, is.factor)]     #Factor variables

#Compute correlation matrix for numeric variables
cor_matrix <- cor(
  x = numeric_clothing,
  method = "pearson"
)

#Visualize correlation matrix
corrplot(cor_matrix, method = "number", tl.cex = 0.8)

#Scatterplot matrix for numeric variables
pairs(
  x = numeric_clothing,
  main = "Scatterplot Matrix of Numerical Variables"
)

#Box plot (Card's effect on Amount)
ggplot(Clothing, aes(
  x = Card,
  y = Amount,
  fill = Card
  )
```

```
) +
  geom_boxplot() +
  labs(
    title = "Plot3. Sales Amount by Card Usage",    #Name it
    x = "Card Usage",
    y = "Sales Amount (USD)"
  ) +
  theme_minimal() +
  scale_fill_manual(
    values = c(
      "No Card" = "#F8766D",
      "Card" = "#00BFC4"
 )
)

#Group statistics: Effect of Card on numeric variables
group_stats <- Clothing %>%
  group_by(Card) %>%
  summarise(
    Mean_Amount = mean(Amount, na.rm = TRUE),
    SD_Amount = sd(Amount, na.rm = TRUE),
    Mean_Freq12 = mean(Freq12, na.rm = TRUE),
    Mean_Freq24 = mean(Freq24, na.rm = TRUE)
  )



#Perform ANOVA: Test significance of Card on Amount
anova_result <- aov(
  formula = Amount ~ Card,
  data = Clothing
)
anova_summary <- summary(
  anova_result
)

anova_table <- data.frame(
  Source = rownames(anova_summary[[1]]),
  Df = anova_summary[[1]]$Df,
  F_Value = anova_summary[[1]]$`F value`,
  P_Value = anova_summary[[1]]$`Pr(>F)`
)
```

```r
#Tabulation of ANOVA using kable
kable(anova_table,
      format = "latex",
      caption = "ANOVA Results for Sales Amount by Card Usage",
      digits = 4,            #Retain 4 decimal places
      booktabs = TRUE)


#full-model
model_all <- lm(Amount ~ ., data = Clothing)
vif_values <- vif(model_all)


#Convert VIF values to a data frame
vif_table <- data.frame(
  Variable = names(vif_values),
  VIF = as.numeric(vif_values)
)

#Display VIF table using knitr::kable
kable(vif_table,
      format = "latex",
      caption = "Variance Inflation Factor (VIF) Table",
      digits = 4,                       #Retain 4 decimal places
      booktabs = TRUE
)


Clothing$AvgSpent12 <- Clothing$Dollar12 / Clothing$Freq12

best_subset_Cloth <- regsubsets(Amount ~ ., data = Clothing)
plot(best_subset_Cloth)

best_subset_Cloth_lm <- lm(Amount ~ Card + Dollar24 + Freq24, data = Clothing)

#Extract regression model results
model_summary <- tidy(best_subset_Cloth_lm)
#Select Beta values and P-values
beta_p_table <- model_summary[, c(
  "term",
  "estimate",
  "p.value"
```

```
  )
]

#Rename columns for readability
colnames(beta_p_table) <- c(
  "Variable",
  "Beta Coefficient",
  "P-value"
)

#Display regression results using kable
kable(beta_p_table,
      format = "latex",
      caption = "Regression Coefficients and P-values for Best Subset Model",
      digits = 4,         #Retain 4 decimal places
      booktabs = TRUE
)
model_Cloth_k <- lm(Amount ~ AvgSpent12 + Card, data = Clothing)

#Extract regression model results
model_summary2 <- tidy(model_Cloth_k)

#Select Beta values and P-values
beta_p_table2 <- model_summary2[, c("term", "estimate", "p.value")]

#Rename columns for readability
colnames(beta_p_table2) <- c(
  "Variable",
  "Beta Coefficient",
  "P-value"
)

#Display regression results using kable
kable(beta_p_table2,
      format = "latex",
      caption = "Regression Coefficients and P-values for Best Subset Model",
      digits = 4,          #Retain 4 decimal places
      booktabs = TRUE)

cooks_model_k <- cooks.distance(model_Cloth_k)

#Generate residual graph
```

```r
plot(
  x = model_Cloth_k,
  which = 1,
  main = "Residuals vs Fitted"
)

#Generate Normal Q-Q plot
plot(
  x = model_Cloth_k,
  which = 2,
  main = "Normal Q-Q Plot"
)
#Plot two regression lines categorized according to the cards
ggplot(Clothing, aes(x = AvgSpent12, y = Amount, color = Card)) +
  geom_point(
    alpha = 0.6
  ) +
  geom_smooth(
    method = "lm",
    formula = y ~ x,
    se = TRUE
  ) +
  labs(
    title = "Plot4. Sales Amount vs AvgSpent12 by Card Usage",
    x = "Average Spent in Last 12 Months",
    y = "Sales Amount (USD)"
  ) +
  theme_minimal() +
  scale_color_manual(
    values = c(
      "No Card" = "red",
      "Card" = "blue"
      )
    )
```