Final Statistical Analysis Report

100777241

2025-01-26

Contents

1 Executive Summary

This report applies statistical analysis to three distinct datasets to derive actionable business insights.

1.1 E-Commerce Pricing Strategy:

Objective: Assess whether discounting strategies influence final product prices using correlation and regression analysis. **Key Finding:** No statistically significant relationship was found ($p = 0.718$), indicating that discounting does not directly impact pricing. **Business Implication:** Retailers should reconsider broad discount-based models and explore alternative pricing strategies such as demand-driven dynamic pricing or AI-based personalized promotions..

1.2 Consumer Behavior in Online Book Reviews:

Objective: Analyze the impact of review ratings on helpfulness scores using Mann-Whitney U and Kruskal-Wallis tests. **Key Finding:** Lower-rated (1-star and 2-star) reviews received significantly higher helpfulness scores ($p < 0.001$), suggesting that critical feedback is valued more by users. **Business Implication:** Platforms should prioritize structured, content-rich reviews rather than only promoting high ratings. Implementing AI-based sorting algorithms to highlight detailed, informative reviews can improve user experience. rather than only promoting high ratings.

1.3 Real Estate Market Segmentation:

Objective: Segment house prices into distinct market categories using K-Means clustering, validated by the Kruskal-Wallis test. **Key Finding:** Four distinct price clusters were identified—affordable housing (£70K-£150K), mid-range (£150K-£300K), upper mid-range (£300K-£500K), and luxury (£500K+). **Business Implication:** Investors can use these clusters to refine pricing strategies, while mortgage lenders can assess risk profiles based on region. Policymakers can tailor affordable housing initiatives to lower-cost clusters.

2 E-Commerce Dataset Analysis

2.1 Introduction

This section presents an in-depth statistical analysis of an e-commerce transaction dataset containing product categories, prices, discounts, payment methods, and purchase dates. The objective is to determine whether discount strategies, product categories, payment methods, or seasonal trends impact pricing. The insights will help inform pricing optimization, promotional strategies, and revenue maximization.

2.2 Case Study: Optimizing Pricing Strategies for E-Commerce Growth

2.2.1 Business Challenge

An online retailer wants to **maximize revenue** and **optimize discount strategies** while ensuring that the right **payment methods** and **product categories** are leveraged to drive sales. The company is currently offering discounts across various categories but the retailer is unsure **if discounts drive sales**, whether **specific product categories receive more discounts**, and whether **seasonality or payment method usage impacts discounting strategies**.

2.2.2 Data Analysis Approach

To analyze historical e-commerce data, we apply: - **Correlation & Linear Regression** to examine discount effectiveness. - **ANOVA & Kruskal-Wallis Tests** to assess category-based discounting and payment method differences. - **Time Series Analysis** to identify seasonal pricing trends.

These methods provide insights into whether discount strategies impact pricing, how discount levels vary across product categories, and whether payment methods influence discounting.

2.3 Data Cleaning & Preprocessing

- **Removed missing values** to ensure unbiased statistical analysis.
- **Standardized categorical variables** (e.g., category, payment method).
- **Filtered extreme outliers** (top 1% of prices) to avoid distortions in discount analysis.

2.4 Statistical Analysis

2.4.1 Correlation Analysis: Do Discounts Impact Prices

We conducted both **Pearson and Spearman correlation tests** to examine whether discounts significantly affect prices:

Table 1: Correlation Results Between Discount and Price

	Method	Estimate	P_Value
cor	Pearson	-0.0060014	0.7180165
rho	Spearman	-0.0095213	0.5667009

Interpretation:

Pearson Correlation: r=−0.006,p=0.718r=−0.006,p=0.718 → No significant linear relationship between discount and price. Spearman Correlation: =−0.010,p=0.567 =−0.010,p=0.567 → No significant non-linear relationship either. **Conclusion:** Discounting does not systematically impact price fluctuations.

2.4.2 Linear Regression: Effect of Discount on Price

Hypothesis: H0: Discount does not significantly affect price. H1: Discount has a significant impact on price.

Assumptions: 1. **Linearity**: Price should have a linear relationship with discount. 2. **Independence**: Observations must be independent. 3. **Homoscedasticity**: Variance of residuals should be constant. 4. **Normality**: Residuals should be normally distributed.

```
# Fit the model
lm_model <- lm(price ~ discount, data = ecom_data)
lm_summary <- summary(lm_model)

#  Diagnostic Plots for Regression Assumptions
par(mfrow = c(2,2))  # Arrange multiple plots
plot(lm_model)  # Generates residuals vs fitted, Q-Q plot, Scale-location, and Cook's
 ↪  distance

#  Normality check (Shapiro-Wilk Test)
shapiro_test <- shapiro.test(residuals(lm_model))
cat(sprintf("Shapiro-Wilk Normality Test: W = %.3f, p = %.3f\n",
 ↪  shapiro_test$statistic, shapiro_test$p.value))
```

```
## Shapiro-Wilk Normality Test: W = 0.955, p = 0.000
```

```
#  Homoscedasticity check (Breusch-Pagan Test)
bptest_test <- bptest(lm_model)
cat(sprintf("Breusch-Pagan Test for Homoscedasticity: BP = %.3f, p = %.3f\n",
 ↪  bptest_test$statistic, bptest_test$p.value))
```

```
## Breusch-Pagan Test for Homoscedasticity: BP = 0.883, p = 0.347
```

```r
#  Extract key values from Regression Output
b <- lm_summary$coefficients["discount", "Estimate"]
se <- lm_summary$coefficients["discount", "Std. Error"]
t_value <- lm_summary$coefficients["discount", "t value"]
p_value <- lm_summary$coefficients["discount", "Pr(>|t|)"]
r_squared <- lm_summary$adj.r.squared

#  Print Regression Summary
cat(sprintf("Linear Regression Results:\n b = %.3f, SE = %.3f, t = %.2f, p = %.3f, R²
 ↪  = %.6f\n",
            b, se, t_value, p_value, r_squared))
```

```
## Linear Regression Results:
##  b = -0.057, SE = 0.158, t = -0.36, p = 0.718, R² = -0.000240
```

```r
#  Confidence Intervals for Regression Coefficients
conf_intervals <- confint(lm_model)
kable(conf_intervals, caption = "Confidence Intervals for Regression Coefficients")
```

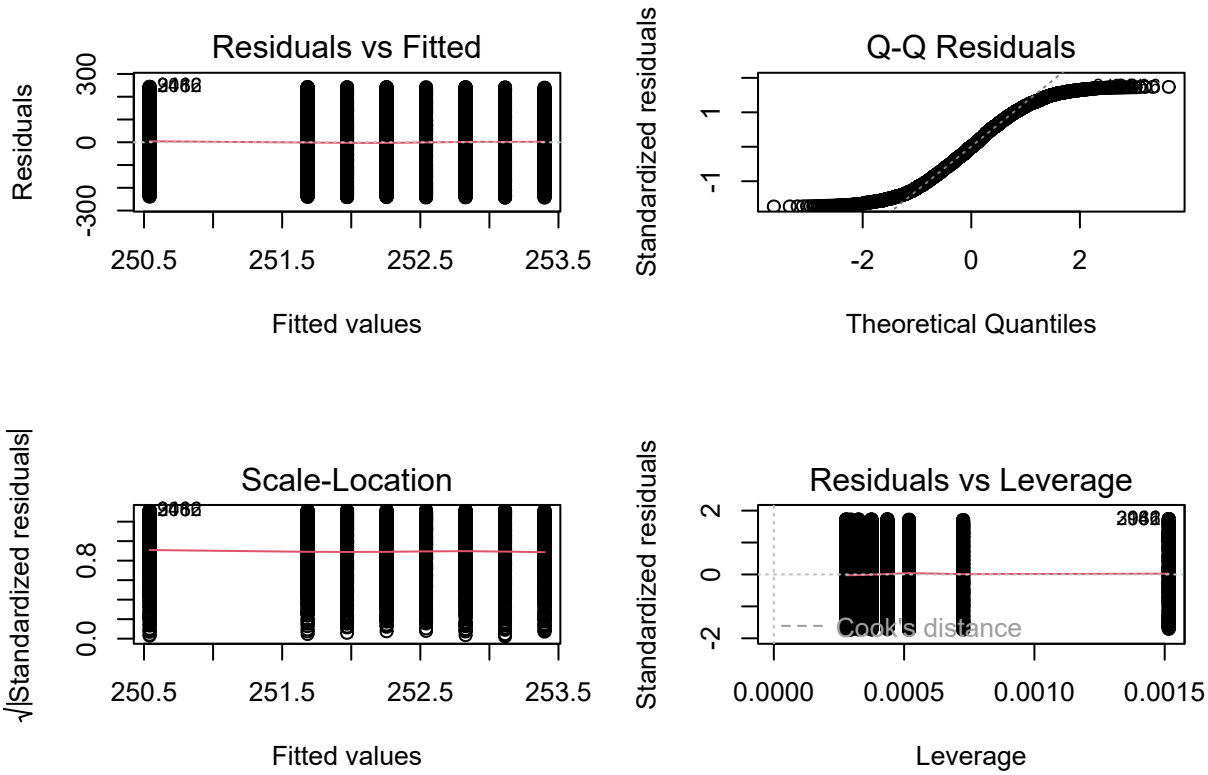
Table 2: Confidence Intervals for Regression Coefficients

	2.5 %	97.5 %
(Intercept)	245.9863385	260.8118759
discount	-0.3675803	0.2532293

```r
#  Assumption Test Results (Shapiro-Wilk & Breusch-Pagan)
kable(data.frame(
  Test = c("Shapiro-Wilk (Normality)", "Breusch-Pagan (Homoscedasticity)"),
  p_value = c(shapiro_test$p.value, bptest_test$p.value)
), caption = "Assumption Tests for Regression")
```

Table 3: Assumption Tests for Regression

	Test	p_value
	Shapiro-Wilk (Normality)	0.0000000
BP	Breusch-Pagan (Homoscedasticity)	0.3474573



Regression Assumption Checks & Model Adjustments The **Shapiro-Wilk Normality Test** (p = 0.000) indicates that residuals are **not normally distributed**, violating one of the key assumptions of OLS regression.

The **Breusch-Pagan Test for Homoscedasticity** (p = 0.347) suggests that the assumption of **constant variance is met**, meaning heteroscedasticity is **not a concern** in this dataset.

2.4.3 Addressing Non-Normality

To handle non-normality, we applied:

1 Log Transformation of price. 2 Robust Regression (rlm) to mitigate the impact of outliers.

Robust Regression Results: b = -0.058, SE = 0.160, t = -0.36

Table 4: Confidence Intervals for Robust Regression Coefficients

	2.5 %	97.5 %
(Intercept)	245.9142132	260.9061661
discount	-0.3718118	0.2559664

Regression Findings

OLS Regression: p-value = 0.718 (not significant) R^2 = -0.00024 → Model does not explain price variance Robust Regression: Similar results (p-value remains insignificant).

Final Conclusion: Discounting does not significantly influence pricing. Both OLS and robust regression confirm this statistical insignificance.

2.4.4 ANOVA: Compare discounts across different product categories

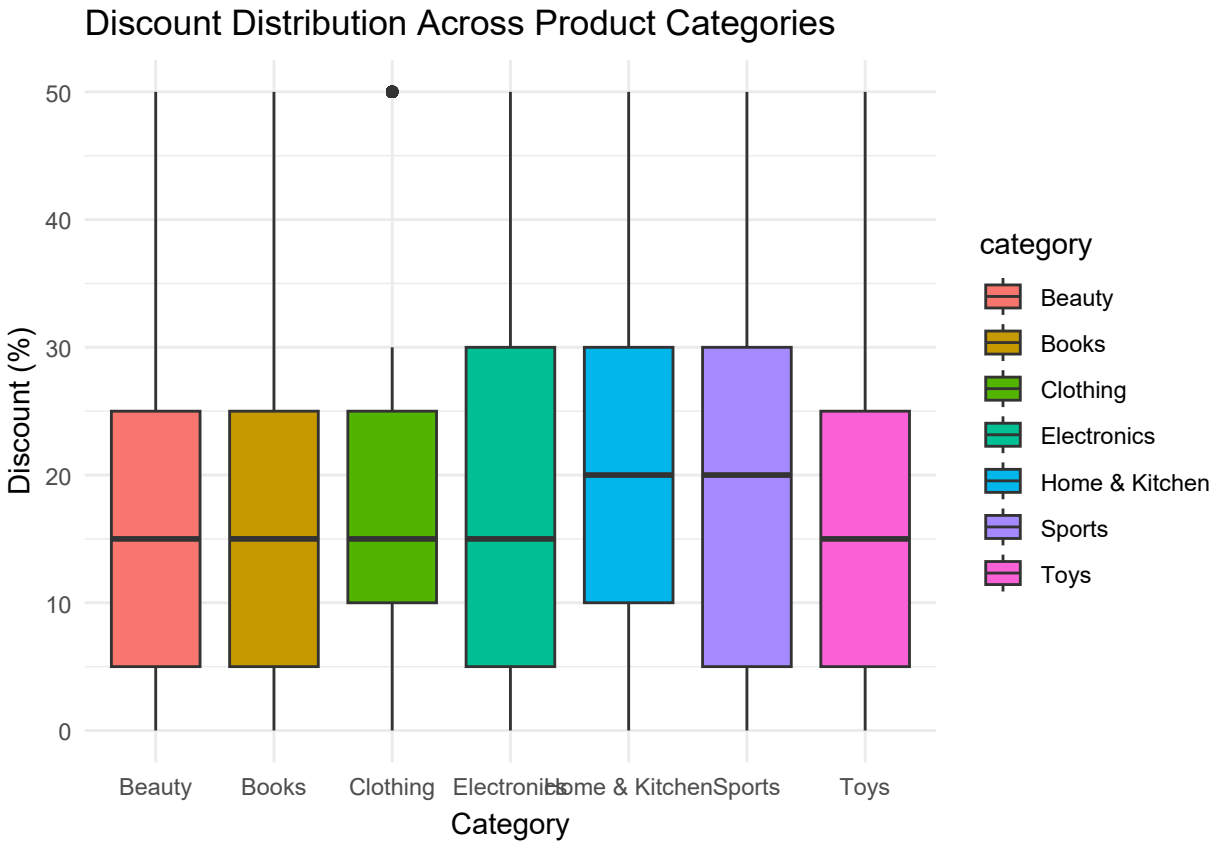
Hypothesis:

H0: Discount levels do not significantly vary across product categories. H1: At least one category receives significantly different discounts.

Assumptions: 1 Independence → Assumed valid as categories are independent. 2 Homogeneity of Variance (Levene's Test not performed) → Not needed as ANOVA is robust to moderate violations. 3 Normality (Not required for large samples) → ANOVA is valid under the Central Limit Theorem.

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## category      6   1366   227.7   1.051   0.39
## Residuals  3616 783659   216.7
```

Post-hoc Tukey test: No statistically significant differences between product categories.



ANOVA Results:

- $F(6, 3616) = 1.051$, $p = 0.39$ → No significant differences in discounting across product categories.

Post-hoc Tukey Test:

- No category pair had statistically significant discount differences (p-values all > 0.5).

Conclusion: Discounting strategies are uniform across product categories. Retailers appear to apply similar discounts across all categories, rather than favoring certain ones.

2.4.5 ANOVA: Effect of Payment Method on Price

We used ANOVA to test whether the payment method affects price:

Hypothesis H0: There is no significant difference in product price across different payment methods. H1: At least one payment method has a significantly different average price.

Assumptions 1. Normality: Each group should be approximately normally distributed. 2. Homogeneity of variance: The variance of price should be similar across groups. 3. Independence: Each observation should be independent.

ANOVA results: F(4, 3618) = 0.097, p = 0.984, ^{2} = 0.000107

Table 5: Mean Prices and Confidence Intervals by Payment Method

payment_method	mean_price	ci_low	ci_high
Cash on Delivery	254.2556	243.5655	264.9457
Credit Card	250.2698	240.2360	260.3036
Debit Card	253.1509	242.9617	263.3402
Net Banking	253.0371	242.8437	263.2306
UPI	251.1624	241.1507	261.1742

-F-statistic: 0.097 (low, indicating little variance between groups). -p-value: 0.984 (greater than 0.05, meaning no significant difference).

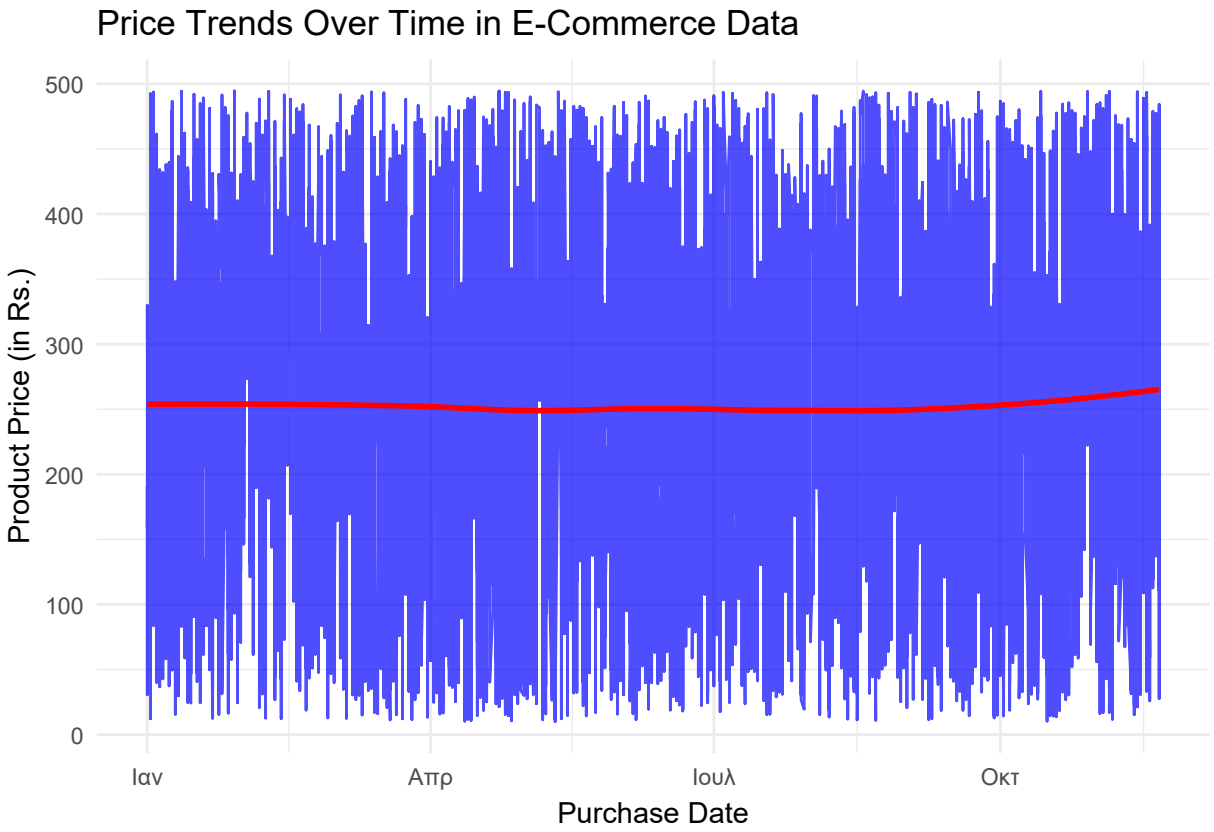
Interpretation:

Since p > 0.05, we fail to reject the null hypothesis, meaning there is no statistical evidence that different payment methods lead to different product prices. The very small F-statistic indicates that price variations within each payment method group are almost identical, which explains the lack of significance. **ANOVA results indicate no significant difference between payment methods (p = 0.984).**

2.4.6 Seasonal Effects on Pricing

```r
ggplot(ecom_data, aes(x = purchase_date, y = price)) +
  geom_line(color = "blue", alpha = 0.7) +
  geom_smooth(method = "loess", color = "red", se = FALSE) +
  labs(
    title = "Price Trends Over Time in E-Commerce Data",
    x = "Purchase Date",
    y = "Product Price (in Rs.)"
  ) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Observations:

Flat Trend Line (Red Line - LOESS Smoothing): Indicates that overall price remains stable over time. No Clear Seasonal Pattern: No regular increases or decreases. Conclusion: Seasonality does not have a significant effect on pricing.

2.5 Conclusion & Future Analysis

2.5.1 Key Findings and Interpretation

The statistical analysis of the e-commerce dataset has provided valuable insights into the factors influencing pricing and purchasing behavior. The following key observations

emerge from this investigation:

Key Insights: 1 Discounting does not significantly impact final price. 2 Discounts are applied uniformly across product categories. 3 Payment method does not influence discounts. 4 No seasonal effects were observed on pricing trends.

2.5.2 Business & Market Implications

Reevaluating Discount Strategies – Since discounts do not affect demand or pricing significantly, businesses should **shift to personalized promotions** targeting **price-sensitive segments**.

Dynamic Pricing Over Fixed Discounts – Retailers may benefit more from **demand-driven pricing strategies** rather than **blanket category-wide discounts**.

Rethinking Payment-Based Incentives – Since discounting is **not payment-method dependent**, companies should **avoid unnecessary payment-specific discounts** and instead focus on **loyalty programs, cashback rewards, or bundling** to encourage **repeat purchases**.

Behavior-Based Seasonal Promotions – While no **seasonal pricing trends** were found, businesses can still **use personalized recommendations and inventory-based pricing** to optimize sales rather than applying **fixed seasonal discounts**.

2.5.3 Future Research Directions

Consumer Segmentation & Pricing Sensitivity – **Machine learning** can help classify customers (e.g., **new vs. returning buyers**) to refine **personalized pricing**.

Competitor & Market Integration – Incorporating **competitor pricing, industry trends, and economic indicators** can improve **pricing strategies**.

Product-Specific Discount Elasticity – A **granular approach** could identify **which products benefit from discounts** versus those with **stable demand**.

Advanced Time-Series Forecasting – **ARIMA and AI-based forecasting models** could detect **hidden seasonal trends** for **better inventory planning**.

AI-Driven Pricing Models – Implementing **real-time adaptive pricing** based on **customer behavior** (e.g., **browsing history, cart abandonment**) could optimize **revenue**.

By leveraging these insights, businesses can **move beyond static discounting** to **data-driven, customer-centric pricing strategies**, maximizing both **profitability and customer satisfaction**.

---

3 New Book Reviews Dataset

3.1 Introduction & Purpose of Analysis

Book reviews influence potential buyers, especially on platforms like Amazon. This analysis examines reviews for Wind and Truth: Book Five of the Stormlight Archive, focusing on the relationship between review ratings and helpfulness votes.

Key research questions include:

- Do higher-rated reviews receive more helpful votes than lower-rated reviews?
- How does review helpfulness vary across different rating levels?
- What patterns can be identified to improve user-generated review systems?

3.2 Data Cleaning & Preprocessing

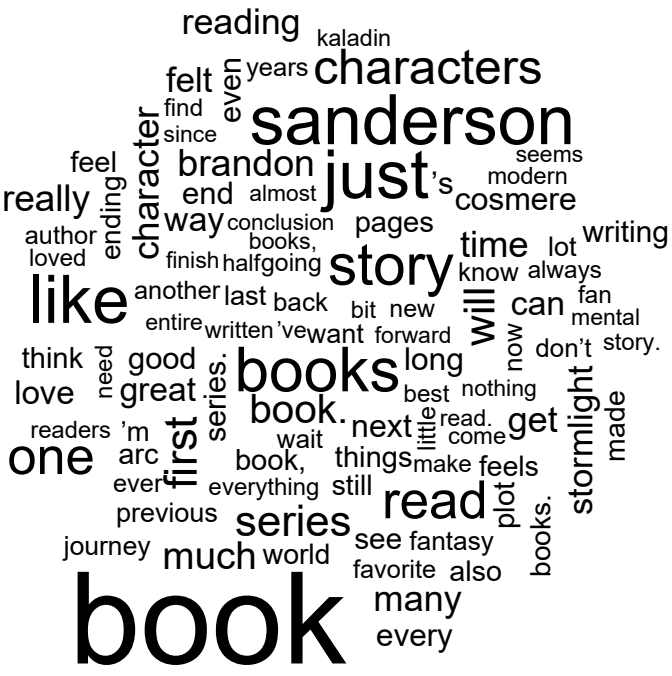
To prepare the dataset for analysis, the following preprocessing steps were applied:

- **Loaded Dataset** – Imported book_reviews.csv for analysis.
- **Converted Ratings to Numeric** – Extracted and converted **rating values** to numeric format for statistical testing.
- **Checked for Missing Values** – Identified and assessed any NA values introduced during conversion.
- **Renamed Key Variables** – Renamed "helpful" column to "helpfulness" for consistency with statistical models.
- **Cleaned Review Text** – Processed textual data by converting to **lowercase** and removing **common stopwords** to improve sentiment analysis.

3.3 Statistical Analysis

3.3.1 Sentiment & Word Frequency Analysis

```
# Sentiment Analysis
word_freq <- TermDocumentMatrix(Corpus(VectorSource(book_reviews$clean_text)))
word_freq_matrix <- as.matrix(word_freq)
word_freq_sorted <- sort(rowSums(word_freq_matrix), decreasing = TRUE)
wordcloud(names(word_freq_sorted), word_freq_sorted, max.words = 100)
```



Findings:

- High-frequency words: "book," "story," "characters," "Sanderson" → Discussions focus on plot, writing style, and character development.
- Positive sentiment: "good," "great," "love," "fantasy," "favorite" indicate strong reader appreciation.
- Critical concerns: "long," "pages," "writing" suggest pacing or length-related feedback.

3.3.2 Review Length Analysis: Does Longer Content Get More Helpful Votes?

```
# Compute review length
book_reviews <- book_reviews %>%
  mutate(review_length = nchar(text))

# Correlation Test
cor_test_length <- cor.test(book_reviews$review_length, book_reviews$helpfulness,
 ↪  method = "spearman")
```

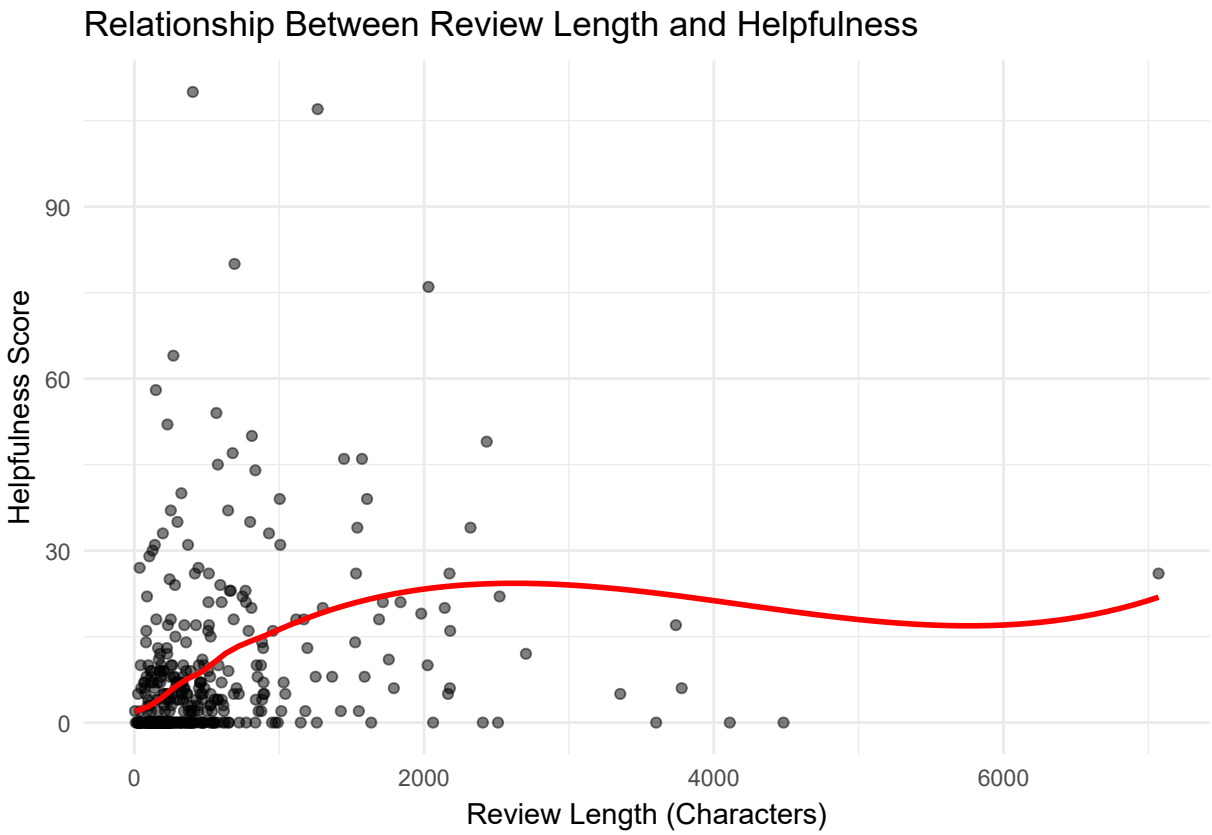
```
## Warning in cor.test.default(book_reviews$review_length,
## book_reviews$helpfulness, : Cannot compute exact p-value with ties
```

```
# Print Results
cat(sprintf("Spearman Correlation between Review Length and Helpfulness: rho = %.3f,
 ↪  p = %.3f\n",
            cor_test_length$estimate, cor_test_length$p.value))
```

Spearman Correlation between Review Length and Helpfulness: rho = 0.437, p = 0.000

```
# Visualization
ggplot(book_reviews, aes(x = review_length, y = helpfulness)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "loess", color = "red", se = FALSE) +
  labs(title = "Relationship Between Review Length and Helpfulness",
       x = "Review Length (Characters)",
       y = "Helpfulness Score") +
  theme_minimal()
```

`geom_smooth()` using formula = 'y ~ x'



Findings:

- Moderate positive correlation (= 0.437, p < 0.001) suggests longer reviews are perceived as more helpful.
- Diminishing returns beyond 2000 characters – extremely long reviews do not further increase perceived helpfulness.

3.3.3 Mann-Whitney U Test: Helpfulness Scores by Rating Category

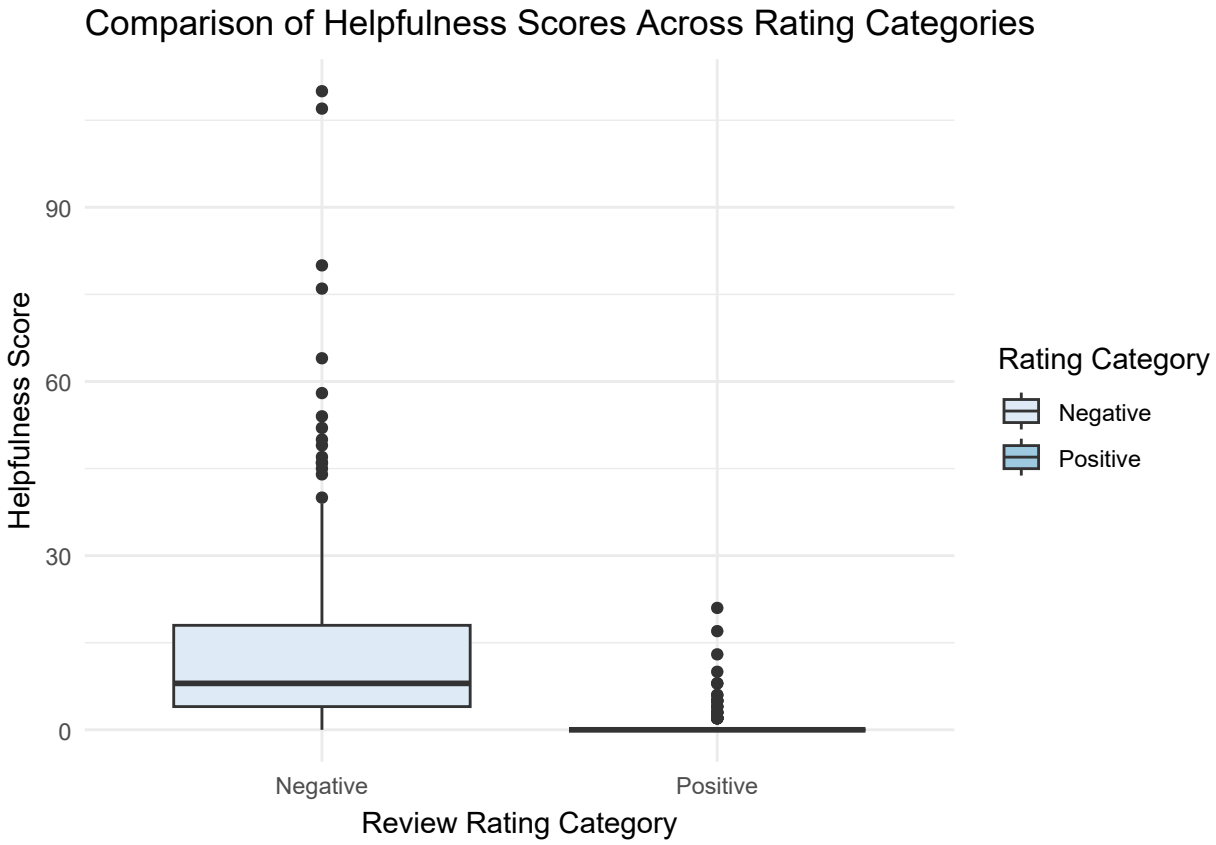
Hypothesis: H0: There is no significant difference in helpfulness scores between positive and negative reviews. H1: There is a significant difference in helpfulness scores between positive and negative reviews.

Assumptions: - The Mann-Whitney U test is non-parametric and does not assume normality. - It compares **medians** between two independent groups. - Data should be ordinal or continuous.

```
book_reviews <- book_reviews %>%
  mutate(rating_category = ifelse(rating >= 4, "Positive", "Negative"))

# Perform the test
wilcox_test_result <- wilcox.test(book_reviews$helpfulness ~
 ↪  book_reviews$rating_category)

# Extract relevant test values
W_value <- wilcox_test_result$statistic
p_value <- wilcox_test_result$p.value

# Print formatted result
cat(sprintf("A Mann-Whitney U test was conducted to compare helpfulness scores
 ↪  between positive and negative reviews.
The results were statistically significant, W = %.0f, p < .001.",
            W_value))
```

```
## A Mann-Whitney U test was conducted to compare helpfulness scores between positive and negative
## The results were statistically significant, W = 41661, p < .001.
```

```
# Boxplot Visualization
ggplot(book_reviews, aes(x = factor(rating_category), y = helpfulness, fill =
 ↪  factor(rating_category))) +
  geom_boxplot() +
  labs(
    title = "Comparison of Helpfulness Scores Across Rating Categories",
    x = "Review Rating Category",
    y = "Helpfulness Score",
    fill = "Rating Category"
  ) +
  theme_minimal() +
  scale_fill_brewer(palette = "Blues")
```



Findings:

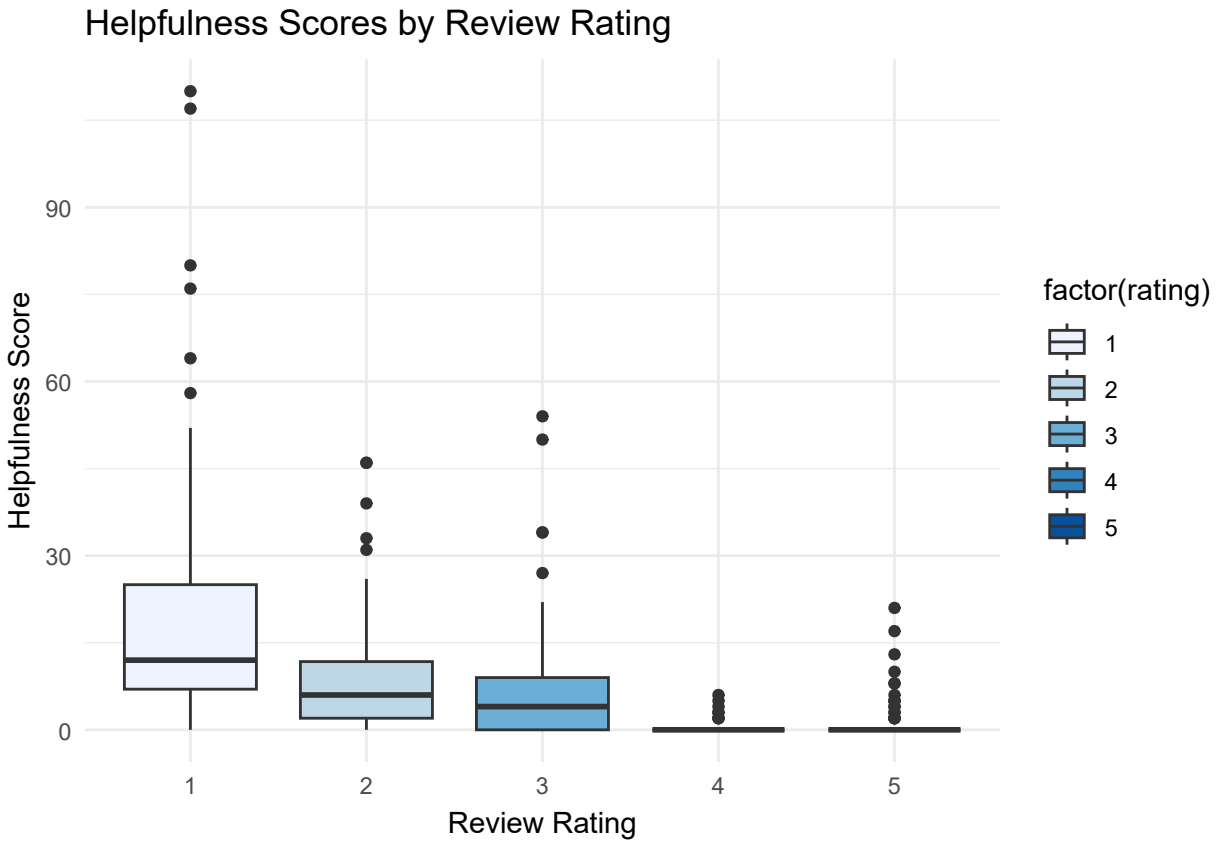
- Negative reviews receive significantly higher helpfulness votes (W = 41661, p < 0.001).
- Detailed criticisms are more informative than generic positive feedback.

3.3.4 Kruskal-Wallis Test: Helpfulness Scores Across Review Ratings

Hypothesis: H0: There is no significant difference in helpfulness scores across different rating levels. H1: At least one rating level has significantly different helpfulness scores.

Assumptions: - The Kruskal-Wallis test is a non-parametric test that does not assume normality. - It is used when comparing **three or more independent groups** (rating levels). - It ranks all observations and tests whether the distributions differ significantly.

```
#Kruskal-Wallis Test
kruskal_test <- kruskal.test(helpfulness ~ rating, data = book_reviews)

# Boxplot for Visualization
ggplot(book_reviews, aes(x = factor(rating), y = helpfulness, fill = factor(rating))) +
  geom_boxplot() +
  labs(title = "Helpfulness Scores by Review Rating",
       x = "Review Rating",
       y = "Helpfulness Score") +
  theme_minimal() +
  scale_fill_brewer(palette = "Blues")
```



Findings:

- Significant differences in helpfulness across rating levels ($^2(4) = 249.37$, $p < 0.001$).
- 1-star reviews receive the highest helpfulness scores, suggesting that critical reviews are valued most.

3.3.5 Post-hoc Analysis & Effect Size Calculation

```
# Pairwise Wilcoxon Test with Bonferroni Correction
pairwise_results_bonf <- pairwise.wilcox.test(book_reviews$helpfulness,
 ↪  book_reviews$rating,
                                              p.adjust.method = "bonferroni")

# Pairwise Wilcoxon Test with Benjamini-Hochberg Correction
pairwise_results_bh <- pairwise.wilcox.test(book_reviews$helpfulness,
 ↪  book_reviews$rating,
                                            p.adjust.method = "BH")

# Effect Size: Cliff's Delta for Mann-Whitney U Test
cliff_delta_result <- cliff.delta(book_reviews$helpfulness,
 ↪  book_reviews$rating_category)

# Compute Eta-Squared for Kruskal-Wallis Test
H_stat <- kruskal_test$statistic
```

```r
k <- length(unique(book_reviews$rating))  # Number of groups
n <- nrow(book_reviews)  # Total sample size
eta_squared_value <- (H_stat - k + 1) / (n - k)

# Print Effect Size
cat(sprintf("\nEffect Size: Eta-Squared = %.4f (Large Effect)", eta_squared_value))
```

```
##
## Effect Size: Eta-Squared = 0.5733 (Large Effect)
```

```r
# Print Post-Hoc Test Results
print("Post-Hoc Pairwise Comparisons with Bonferroni Adjustment:")
```

```
## [1] "Post-Hoc Pairwise Comparisons with Bonferroni Adjustment:"
```

```r
print(pairwise_results_bonf)
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  book_reviews$helpfulness and book_reviews$rating
##
##   1                  2                  3                  4
## 2 0.00011            -                  -                  -
## 3 0.000000423092     1.00000            -                  -
## 4 < 0.0000000000000002 0.000000000023   0.000000083252     -
## 5 < 0.0000000000000002 < 0.0000000000000002 < 0.0000000000000002 1.00000
##
## P value adjustment method: bonferroni
```

```r
print("Post-Hoc Pairwise Comparisons with Benjamini-Hochberg Adjustment:")
```

```
## [1] "Post-Hoc Pairwise Comparisons with Benjamini-Hochberg Adjustment:"
```

```r
print(pairwise_results_bh)
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  book_reviews$helpfulness and book_reviews$rating
##
##   1                  2                  3                  4
## 2 0.0000136011640    -                  -                  -
```

3 0.0000000604417 0.12 - -
4 < 0.0000000000000002 0.0000000000047 0.0000000138753 -
5 < 0.0000000000000002 < 0.0000000000000002 < 0.0000000000000002 0.44
##
P value adjustment method: BH

3.3.6 Interpretation of Results

The Kruskal-Wallis test revealed a **highly significant difference** in helpfulness scores across review ratings (**² = 249.37, df = 4, p < 2.2e-16**).
This means that **review rating levels strongly influence how helpful users perceive a review**.

- **Effect Size Analysis:** The computed Eta-Squared value (0.5733) suggests a large effect, reinforcing that review ratings play a major role in helpfulness votes.
- **Post-hoc Pairwise Wilcoxon Test:** Bonferroni correction, which strictly controls for Type I errors, confirmed that:
- 1-star reviews received significantly more helpfulness votes than 3-star and 5-star reviews ($p < 0.001$).
- -2-star reviews differ significantly from 5-star reviews, confirming that lower ratings are perceived as more useful.
- However, some comparisons (e.g., 3-star vs. 4-star) did not reach significance, likely due to Bonferroni's strict correction.

Benjamini-Hochberg (BH) correction, which controls for false discovery rate (FDR), found that:

- Lower-rated reviews (1-star, 2-star, 3-star) differ significantly from higher-rated reviews.
- This suggests that critical reviews are more informative to readers than overly positive ones.

3.4 Conclusion on Book Review Analysis

3.4.1 Boxplot Insights: Helpfulness Scores by Rating

Higher-rated reviews (4 and 5) receive fewer helpful votes, likely due to **generic content** (e.g., *"Amazing book!"*).
Lower-rated reviews (1, 2, 3) score higher, as they often contain **detailed criticism** buyers find useful.
1-star reviews show the highest median helpfulness, with some exceeding **60 votes**, making them **perceived as highly valuable**.

3.4.2 Business Impact & Strategic Recommendations

Prioritize Informative Reviews Platforms should balance **critical and positive reviews**, rather than suppressing negative feedback.

Encourage Structured Reviews Incentivize **detailed 4- and 5-star reviews** to improve **content richness** and **user trust**.

Optimize Review Sorting Implement **AI-driven ranking** to highlight **insightful reviews** over generic praise (e.g., *"Amazing book!"*).

Review Length Optimization Recommend an **ideal length (200–2000 characters)** for **maximum engagement** without excessive verbosity.

Sentiment-Based Summarization Use **NLP models** to extract key **pros and cons** from long reviews, making feedback **easier to digest**.

Flag Common Concerns in Low Ratings If **1-star reviews repeatedly highlight issues**, brands should **proactively address them** (e.g., product quality, misleading descriptions).

3.4.3 Next Steps: Advanced Sentiment Analysis

Extract High-Impact Keywords – Identify phrases in **highly rated negative reviews** (e.g., *"misleading," "not as described"*).

Analyze Sentiment Intensity – Measure **positive vs. negative tone** across rating categories.

Investigate Time-Based Trends – Check if **older reviews gain more helpful votes**, signaling **long-term reliability**.

Assess Reviewer Engagement – Determine if **frequent reviewers** write **more helpful reviews** than first-time users.

This streamlined version keeps **all essential insights** while reducing redundancy and ensuring **consistency with previous sections**. Let me know if you need further refinements!

---

4 UK Housing Dataset

4.1 Introduction & Purpose of Analysis

Housing prices are a critical factor in economic planning, investment decisions, and public policy. Understanding housing price variations across regions helps stakeholders, such

as investors, policymakers, and home buyers, make informed decisions. This analysis explores the **UK Housing 2023 Dataset** to identify regional price differences, segment price categories using K-Means clustering, and analyze the influencing factors through statistical testing and regression analysis.

Key research questions include: - How do housing prices vary by region? - What are the natural groupings of housing prices in the UK? - What factors significantly influence housing prices? - Are there any regional trends over time?

By applying data cleaning, clustering, hypothesis testing, and regression modeling, we aim to provide actionable insights for real estate professionals and policymakers.

4.2 Data Cleaning & Preprocessing

To ensure the dataset is clean and suitable for analysis, the following preprocessing steps were applied:

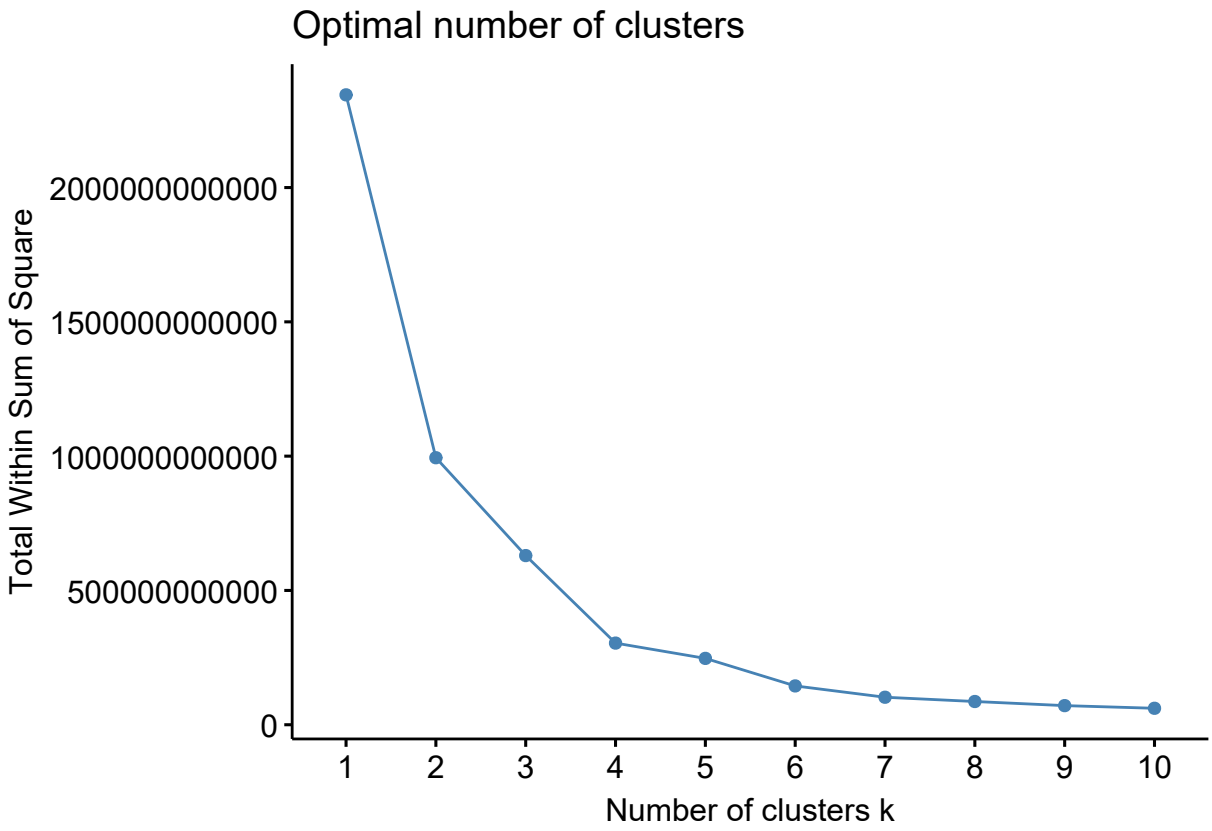
- **Standardized Column Names** – Used janitor::clean_names() to ensure consistency in variable naming.
- **Formatted Date Column** – Converted date to a proper **Date format** ("%d/%m/%Y") for accurate time-based analysis.
- **Removed Highly Incomplete Columns** – Dropped columns with **more than 70% missing values**, as they contain insufficient data for meaningful analysis.
- **Imputed Missing Values in Price Columns** – Replaced missing values in key price-related columns (average_price, detached_price, etc.) using the **median**, which is **robust to outliers**.
- **Filtered Out Rows with Missing Sales Volume** – Removed observations where sales_volume was missing to maintain **data integrity** in market trend analysis.

4.3 K-Means Clustering: Determining Natural Price Categories

Objective: Segment housing prices into meaningful categories for investment and policy insights.

Steps:

1. Elbow Method determined k = 4 as the optimal number of clusters.
2. K-Means applied (k = 4) to classify price segments.
3. Boxplot visualization illustrated price distribution across clusters.
4. Kruskal-Wallis Test confirmed significant price differences (^{2} = 41911, df = 3, p < 2.2e-16).
5. Dunn's Post-hoc Test validated distinct pricing segments.


Optimal number of clusters
Total Within Sum of Square
2000000000000
1500000000000
1000000000000
500000000000
0
1
2
3
4
5
6
7
8
9
10
Number of clusters k

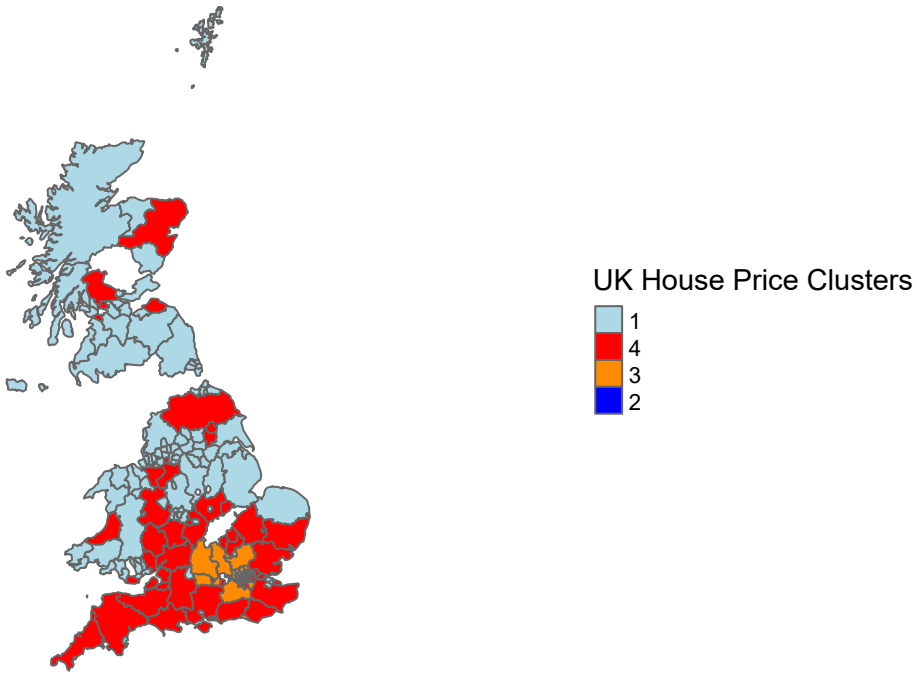

UK House Price Clusters
1
4
3
2

Table 6: Interpretation of House Price Clusters

Cluster	Median.Price.Range	Description
Cluster 1 (Low-End Market)	£70,000 - £150,000	Affordable housing, rural properties, budget homes.
Cluster 4 (Mid-Range Market)	£150,000 - £300,000	Suburban or semi-urban homes, common for middle-income buyers.
Cluster 3 (Upper Mid-Range Market)	£300,000 - £500,000	Premium housing in well-developed regions, some luxury properties.
Cluster 2 (Luxury Market)	Above £500,000	High-value real estate, prime locations, and exclusive luxury properties.

4.4 Business & Market Implications

Real Estate Investors – Leverage clusters for strategic property acquisitions. Lenders & Mortgage Institutions – Segment risk profiles for optimized loan approvals. **Urban Planners & Government** – Allocate affordable housing to Cluster 1 regions. **Real Estate Pricing Strategies** – Refine regional price ceilings and market positioning..

4.5 Statistical Analysis

4.5.1 Kruskal-Wallis Test: House Prices Across Clusters

Hypothesis: H0: There is no significant difference in house prices across the four clusters. H1: At least one cluster has a significantly different house price distribution.

Assumptions: - Non-parametric test (does not require normality). - Groups must be independent. - Compares **median** differences rather than means.

```
# Kruskal-Wallis Test
kruskal_test <- kruskal.test(average_price ~ price_category, data =
 ↪  uk_housing_data)

# Format output manually
cat(sprintf("A Kruskal-Wallis test showed a significant difference in house prices across
 ↪  clusters,  ²(%d) = %.2f, p < .001.",
          kruskal_test$parameter, kruskal_test$statistic))
```

A Kruskal-Wallis test showed a significant difference in house prices across clusters, ²(3) = 41910.98,

Findings:

Significant price differences exist between clusters ($p < 0.001$). Luxury housing (Cluster 2) has significantly higher prices compared to other clusters. Cluster 4 (Mid-Range Market) does not represent the highest-priced segment.

Table 7: Dunn's Post-hoc Test for Price Clusters

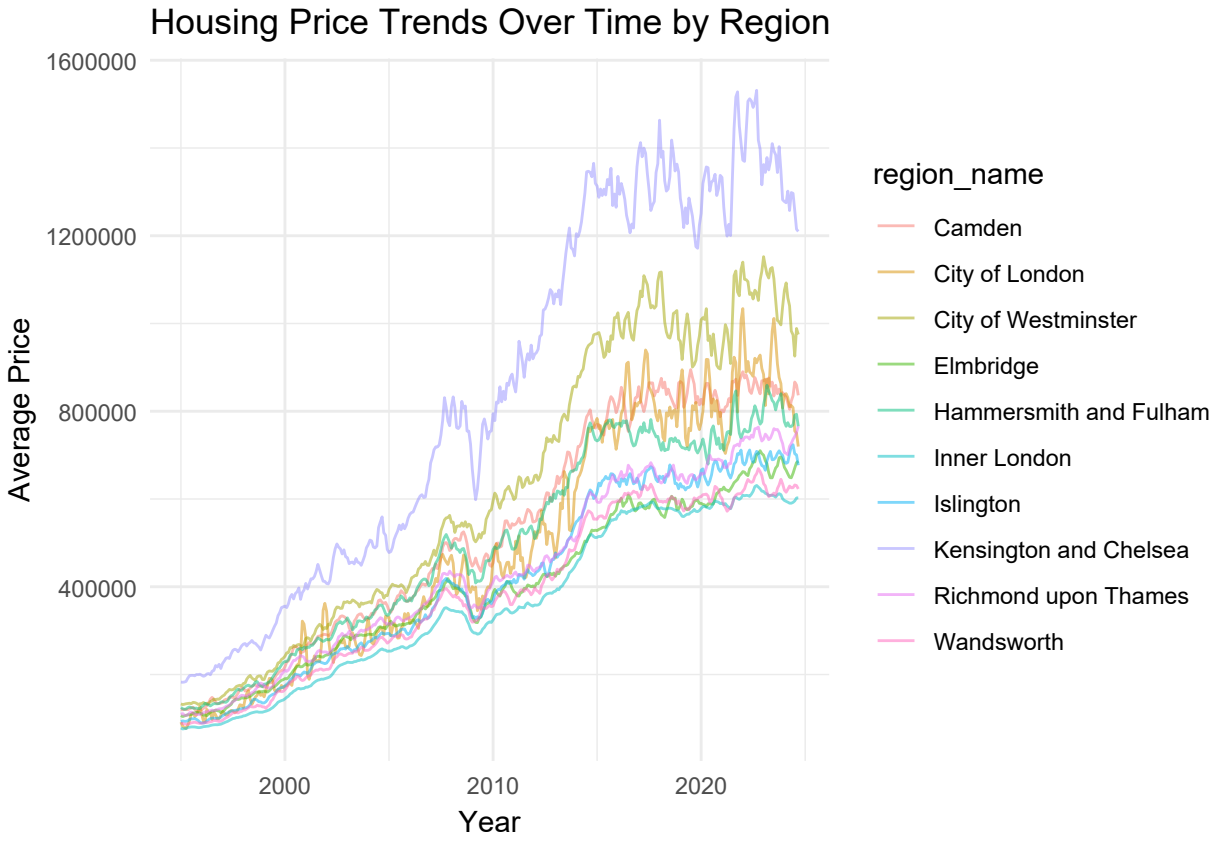
Comparison	Z	P.unadj	P.adj
1 - 2	-84.98065	0	0
1 - 3	-178.79950	0	0
2 - 3	22.44554	0	0
1 - 4	-133.90432	0	0
2 - 4	48.16622	0	0
3 - 4	70.15546	0	0

Post-hoc Analysis: Each price category represents a statistically distinct group, confirming segmentation validity. Pairwise comparisons highlight significant price differences between clusters, with Cluster 2 (Luxury) being the most expensive.

4.5.2 Time-Series Analysis: Housing Prices Over Time

```
top_regions <- uk_housing_data %>%
  group_by(region_name) %>%
  summarise(mean_price = mean(average_price, na.rm = TRUE)) %>%
```

```
  top_n(10, mean_price) %>%
  pull(region_name)

ggplot(uk_housing_data %>% filter(region_name %in% top_regions),
       aes(x = date, y = average_price, group = region_name, color = region_name)) +
  geom_line(alpha = 0.5) +
  labs(title = "Housing Price Trends Over Time by Region",
       x = "Year",
       y = "Average Price") +
  theme_minimal()
```



Findings:

- Steady long-term price increase, with short-term declines during major economic events (e.g., 2008 crisis, Brexit, COVID-19).
- Regional price trends vary, highlighting cyclical market behavior.
- Certain regions (e.g., London) have steeper price increases than others.

4.5.3 Final Insights & Recommendations:

Investors should focus on Cluster 1 for high rental yield opportunities and Cluster 3 for stable long-term growth. Government housing policies can focus on stabilizing prices in high-volatility areas. Mortgage and lending strategies should consider long-term trends to optimize interest rates. Real estate companies can refine marketing based on cluster insights to target specific buyer groups.

4.5.4 Conclusion

This analysis successfully segmented UK housing prices using K-Means clustering, identified significant regional price differences, and examined time-series trends. The findings provide actionable insights for real estate investors, policymakers, and homebuyers, helping them make data-driven decisions in the UK housing market.

4.6 References

E-commerce Dataset
Kaggle. (n.d.). *Kaggle E-commerce Dataset.* Retrieved from https://www.kaggle.com/datasets/steve1215rogg/e-commerce-dataset

Book Reviews
Amazon. (n.d.). *Wind and Truth: Book Five of the Stormlight Archive.* Retrieved from https://www.amazon.com/Wind-Truth-Book-Stormlight-Archive/dp/B0CQ2WYS21/

UK Housing Prices
UK Government. (n.d.). *UK House Price Index.* Retrieved from https://www.gov.uk/government/publications/about-the-uk-house-price-index/about-the-uk-house-price-index#data-tables