Final Project Report

Yi-Ting Hsiao, Abhijit Bajrang More and Shyam Sundar Nandakumar

hsiao.yi@northeastern.edu, more.ab@northeastern.edu, nandakumar.sh@northeastern.edu

Professional Studies in Analytics, Northeastern University

ALY 6015.20379: Intermediate Analytics

Prof. Valeriy Shevchenko

February 11, 2025

Introduction

Nike's global sales performance is influenced by multiple internal and external factors, including pricing strategies, product categories, regional sales variations, and seasonal trends. Understanding the key drivers behind sales volume is essential for optimizing business strategies and improving forecasting accuracy. This study aims to investigate the impact of price tiers, regional differences, seasonal variations, and online sales performance on Nike's sales trends. Additionally, multiple regression models, including Ridge and LASSO regularization, are applied to improve prediction accuracy and address multicollinearity issues.

Questions to Investigate

- **Do different price tiers significantly impact sales volume?**
- **Are there significant differences in product category performance across regions?**
- **Do different regions show statistically different sales patterns?**
- **How does seasonality influence sales, and are the effects consistent across regions?**
- **What are the key predictors of Nike's sales performance?**
- **How can regularization (ridge/LASSO) improve model performance when predicting sales?**

Methods

This study applies multiple statistical and machine learning techniques to analyze Nike's global sales trends. The methods chosen are based on the characteristics of the dataset and the research questions:

- **ANOVA Test –** Used to determine whether sales volume differs significantly across different price tiers.
- **Chi-Square Test –** Used to test if product category performance varies across regions.
- **Kruskal-Wallis Test –** A non-parametric test used to check whether sales distribution varies by region and seasonality.
- **Generalized Linear Models (GLM) –** Applied to examine key predictors of sales performance, using Poisson regression.

- **Multiple Linear Regression with Regularization (Ridge & LASSO) –** Used to improve predictive accuracy and reduce multicollinearity.

These methods ensure a comprehensive analysis of factors affecting Nike's sales trends, combining traditional statistical tests with advanced predictive modeling.

Exploratory Data Analysis

Before performing statistical analyses, the dataset was examined to understand its structure and key characteristics. The dataset consists of 1,000 observations and 10 variables, covering various attributes related to sales performance. These attributes include Units Sold, Retail Price, Product Category, Region, Month, and Online Sales Percentage, providing a comprehensive view of factors influencing Nike's sales trends.

```
> str(nike_sales)
'data.frame':   1000 obs. of  10 variables:
 $ Month                 : chr  "November" "January" "October" "December" ...
 $ Region                : chr  "India" "India" "India" "Greater China" ...
 $ Main_Category         : chr  "Equipment" "Equipment" "Apparel" "Footwear" ...
 $ Sub_Category          : chr  "Bags" "Accessories" "Tops" "Cricket" ...
 $ Product_Line          : chr  "Gym Sack" "Hats" "Tech Fleece" "Vapor Cricket" ...
 $ Price_Tier            : chr  "Budget" "Budget" "Mid-Range" "Premium" ...
 $ Units_Sold            : int  48356 9842 25079 41404 33569 39344 30197 42814 16489 43626 ...
 $ Revenue_USD           : int  14506800 2066820 1755530 8694840 5371040 5508160 6945310 642210
0 2473350 10033980 ...
 $ Online_Sales_Percentage: int  73 50 90 58 53 73 50 55 78 86 ...
 $ Retail_Price          : int  300 210 70 210 160 140 230 150 150 230 ...
```

Figure 1: Data Structure of Nike Sales Dataset

Then, a summary of numerical and categorical variables was generated to check for missing values, data distributions, and any potential anomalies. The descriptive statistics provided insights into the central tendency, variability, and overall trends in the dataset

```
> # Check for missing values
> colSums(is.na(nike_sales))
                  Month                 Region           Main_Category
                      0                      0                       0
           Sub_Category           Product_Line              Price_Tier
                      0                      0                       0
             Units_Sold            Revenue_USD Online_Sales_Percentage
                      0                      0                       0
           Retail_Price
                      0
```

Figure 2: Missing Values Check for Nike Sales Dataset

```
> # Summary statistics for numerical variables
> summary(nike_sales)
    Month              Region          Main_Category      Sub_Category       Product_Line
 Length:1000        Length:1000        Length:1000        Length:1000        Length:1000
 Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character


  Price_Tier          Units_Sold     Revenue_USD       Online_Sales_Percentage  Retail_Price
 Length:1000        Min.   : 5028   Min.   :  287400   Min.   :50.00           Min.   : 50.0
 Class :character   1st Qu.:17554   1st Qu.: 2344675   1st Qu.:60.00           1st Qu.:110.0
 Mode  :character   Median :28685   Median : 4328020   Median :71.00           Median :180.0
                    Mean   :28499   Mean   : 5039576   Mean   :70.04           Mean   :176.3
                    3rd Qu.:40026   3rd Qu.: 7264942   3rd Qu.:80.00           3rd Qu.:240.0
                    Max.   :49992   Max.   :14864700   Max.   :90.00           Max.   :300.0
```

Figure 3: Summary Statistics for Nike Sales Dataset

To investigate whether price tiers impact sales volume, a boxplot was generated to visualize the distribution of Units Sold across different price categories (Budget, Mid-Range, Premium). This initial visualization helped identify patterns, variations, and potential outliers before conducting statistical tests.

The boxplot indicates that all three price tiers exhibit similar sales volume distributions, suggesting that price alone may not be a dominant factor in determining unit sales.

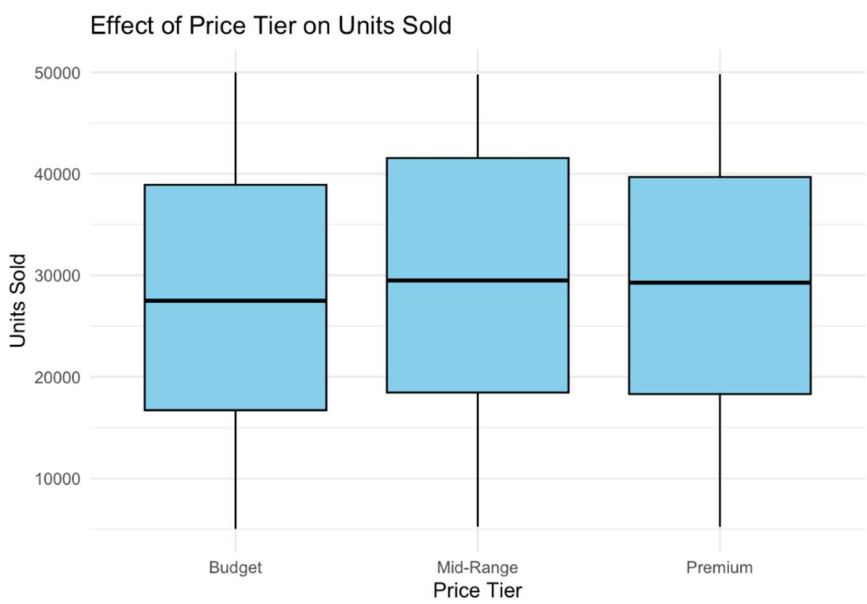

Figure 4: Boxplot of Units Sold by Price Tier

Then, a grouped bar chart was created to visualize how product category performance differs across regions. This visualization provides a comparative view of the sales distribution for Footwear, Apparel, and Equipment across different markets, helping to identify dominant product categories in specific regions. The chart reveals that Equipment tends to have higher sales in regions such as India and South Korea, whereas Apparel is consistently strong across most regions, particularly in Japan and Greater China. Footwear, on the other hand, shows more variability, with relatively lower sales in Europe and South Korea but stronger performance in Greater China and Japan. By examining these patterns, the analysis gains insight into regional consumer preferences, which is further validated through statistical testing.

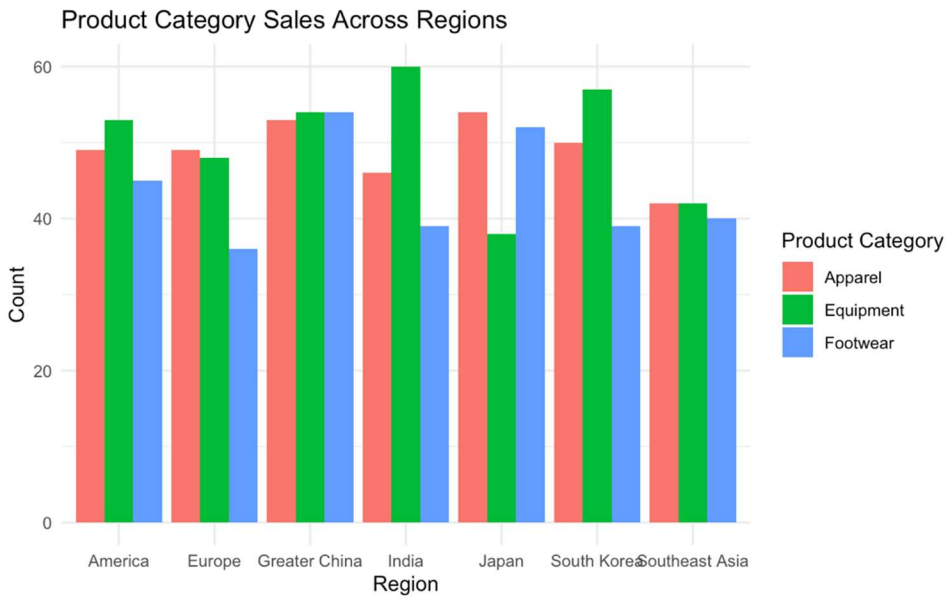

Figure 5: Grouped Bar Chart of Product Category Sales by Region

To explore regional differences in revenue distribution, a violin plot was generated to visualize the spread, density, and variability of revenue across different regions. This plot combines a boxplot with a kernel density estimate, allowing for a more detailed representation of revenue fluctuations within each region.

The violin plot reveals that Europe and South Korea exhibit the widest distribution of revenue, suggesting higher variability in sales performance. Meanwhile, regions like Greater China and India show a more concentrated revenue range, indicating relatively stable sales trends. The presence of long tails in regions such as Japan and Southeast Asia suggests

occasional high-revenue outliers, potentially driven by specific product launches or promotional events.

By analyzing the shape and spread of these distributions, we can assess whether certain regions exhibit higher revenue concentration, greater variability, or the presence of outliers. This visualization provides an initial understanding of regional sales patterns before conducting formal statistical tests to determine whether the observed differences are significant.

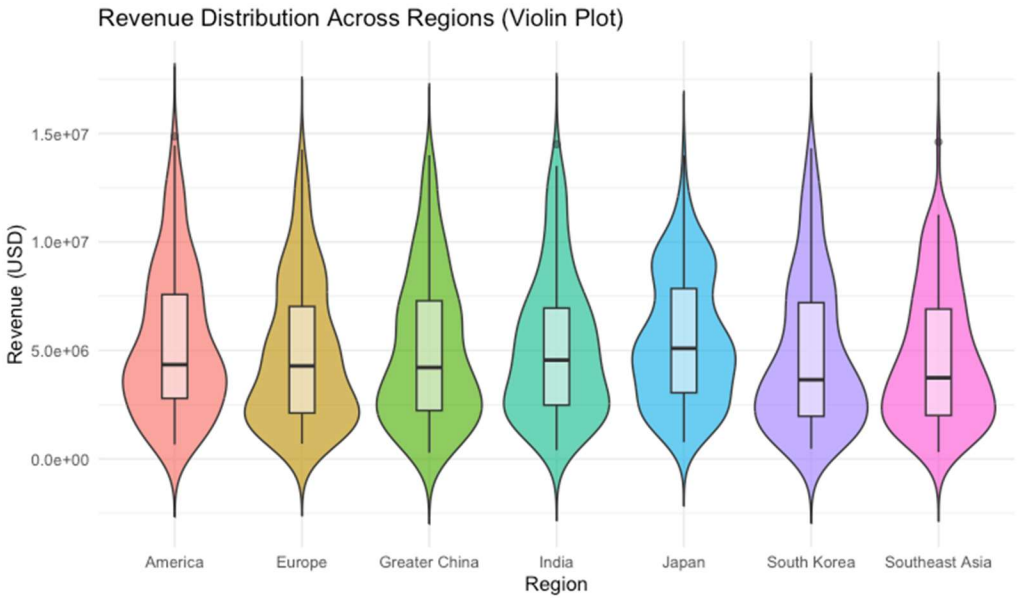

Figure 6: Violin Plot of Revenue Distribution Across Regions

A boxplot was created to visualize the monthly revenue distribution, revealing how sales fluctuate throughout the year. The plot highlights periods of higher or lower revenue, with median sales levels and interquartile ranges (IQRs) illustrating seasonal trends. Notably, revenue appears relatively stable across most months, but November and December show slightly higher median sales, suggesting increased demand during the holiday season. The presence of outliers, particularly in the earlier months of the year, may indicate exceptional sales spikes. This analysis serves as a basis for evaluating the impact of seasonality before conducting formal statistical tests.

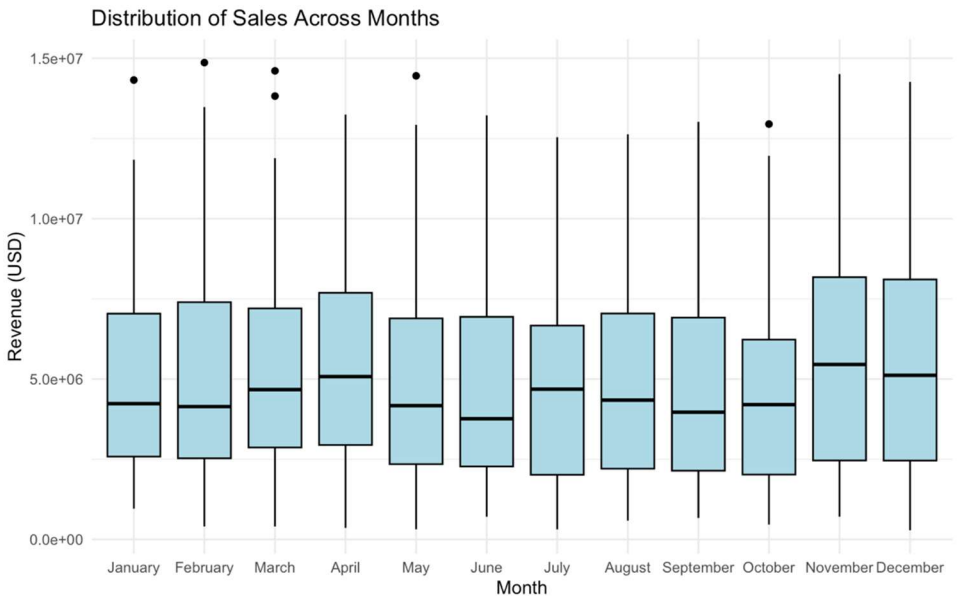

Figure 7: Boxplot of Monthly Revenue Distribution

To further analyze the impact of seasonality across different regions, a heatmap of monthly sales by region was generated to examine seasonal variations in revenue across different geographic markets. The visualization highlights distinct seasonal trends, with some regions experiencing noticeable peaks in specific months. For instance, America and India show higher sales in January and July, indicating potential seasonal demand during these periods. In contrast, Europe and Greater China exhibit relatively lower sales in summer months, particularly in June and July. Additionally, fluctuations in revenue appear to vary across regions, suggesting that while overall sales may follow a seasonal pattern, localized market factors influence the extent of these variations. This warrants further statistical analysis to quantify the significance of these trends and inform region-specific sales strategies.

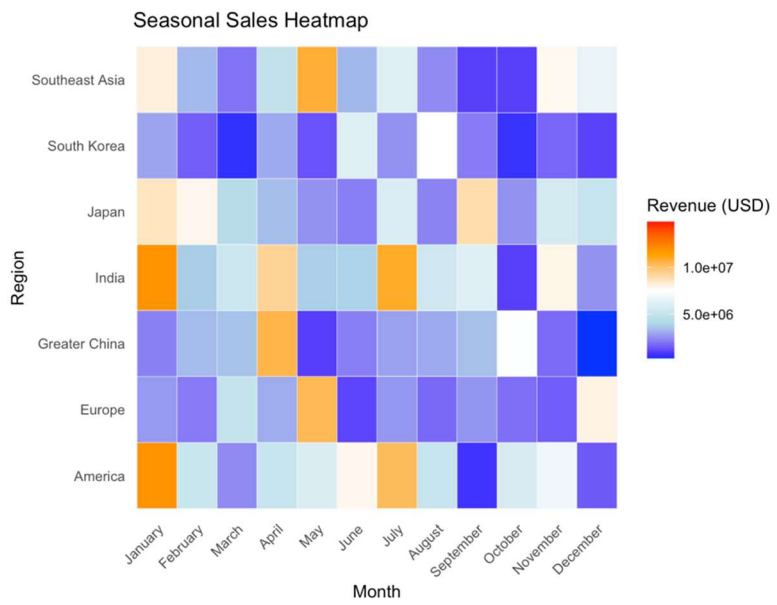

Figure 8: Seasonal Sales Heatmap Across Regions

To gain insights into the key factors influencing units sold, an exploratory analysis was conducted to examine the relationship between Retail Price, Online Sales Percentage, Region, and Seasonality. Several visualizations were created to understand how these internal and external factors impact Nike's sales performance.

A scatter plot of Retail Price versus Units Sold was generated to assess the effect of pricing on sales volume. The plot reveals a weak positive relationship between Retail Price and Units Sold, as indicated by the slight upward trend of the regression line. However, the data points are widely scattered, suggesting that retail price alone does not have a strong predictive influence on sales. Instead, other factors such as brand perception, promotions, and regional demand may play a more significant role in determining sales performance.

Figure 9: Scatter Plot of Retail Price vs. Units Sold

The bar plot of Units Sold by Month highlights seasonal sales trends by ranking monthly sales in descending order. December had the highest sales, followed by February and September, likely due to holiday shopping or seasonal promotions. In contrast, April, March, and August saw the lowest sales, indicating off-peak periods. This visualization helps identify demand cycles, enabling businesses to optimize inventory, marketing, and sales strategies.

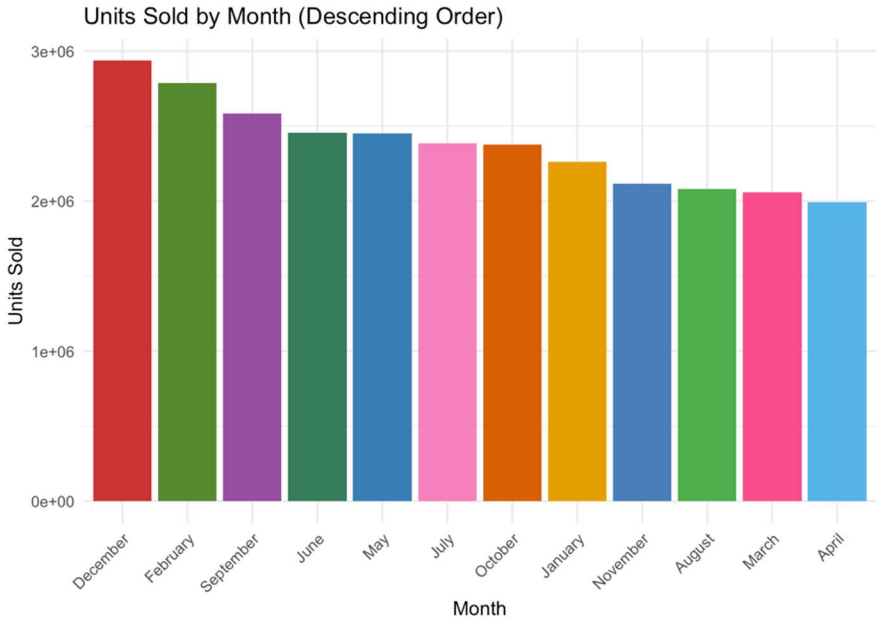

Figure 10: Bar Plot of Units Sold by Month

A scatter plot was used to examine the relationship between Online Sales Percentage and Units Sold. The regression line suggests a weak or negligible correlation, indicating that an increase in the proportion of online sales does not necessarily lead to higher total sales volume.

One possible explanation is channel cannibalization, where customers shift from in-store purchases to online rather than increasing their overall spending. Additionally, differences in consumer behavior between online and offline shopping—such as product trial preferences, return policies, and localized promotions—may also contribute to this weak relationship.

Given this insight, Nike should evaluate its e-commerce strategy to ensure that online sales growth is not negatively affecting in-store sales. This could involve:

- Optimizing omnichannel experiences, such as in-store pickup for online orders.
- Leveraging online-exclusive products or promotions that do not compete directly with retail sales.
- Refining regional digital marketing strategies to better align with consumer online shopping behaviors.

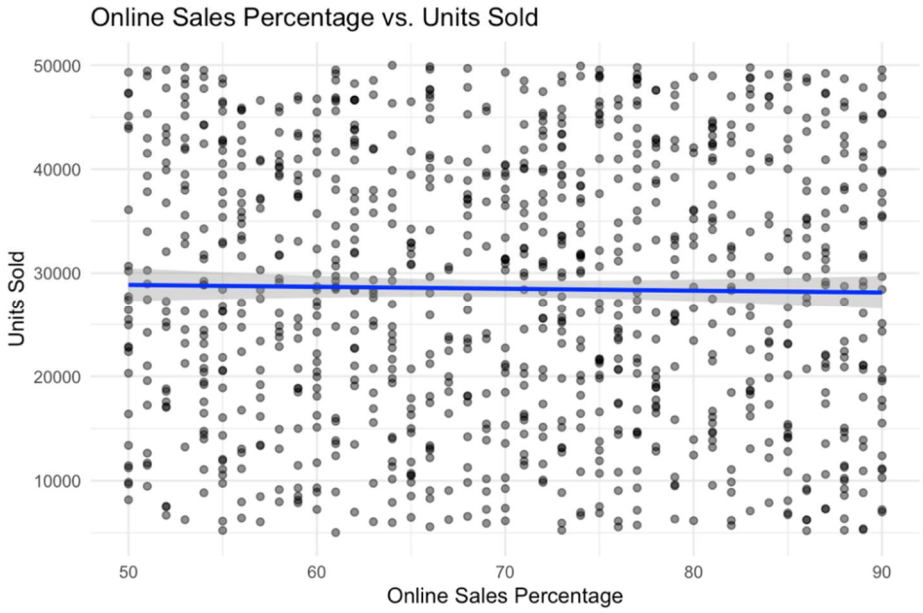

Figure 11: Scatter Plot of Online Sales Percentage vs. Units Sold

Lastly, a faceted scatter plot was generated to analyze the effect of Online Sales Percentage on Units Sold across different regions, allowing for a comparison of regional differences. Overall,

the relationship appears weak or almost nonexistent in most regions, suggesting that the proportion of online sales has limited impact on total sales volume. Only India shows a noticeable positive correlation, where higher online sales percentages are associated with a slight increase in units sold. In contrast, Japan, Europe, and South Korea exhibit a weak or slightly negative correlation, indicating that an increase in online sales percentage does not lead to a significant rise in total sales and may even show a slight decline. These findings suggest that the impact of online sales varies by region but is not a key determinant of overall sales volume, likely influenced by other market dynamics or consumer behaviors.

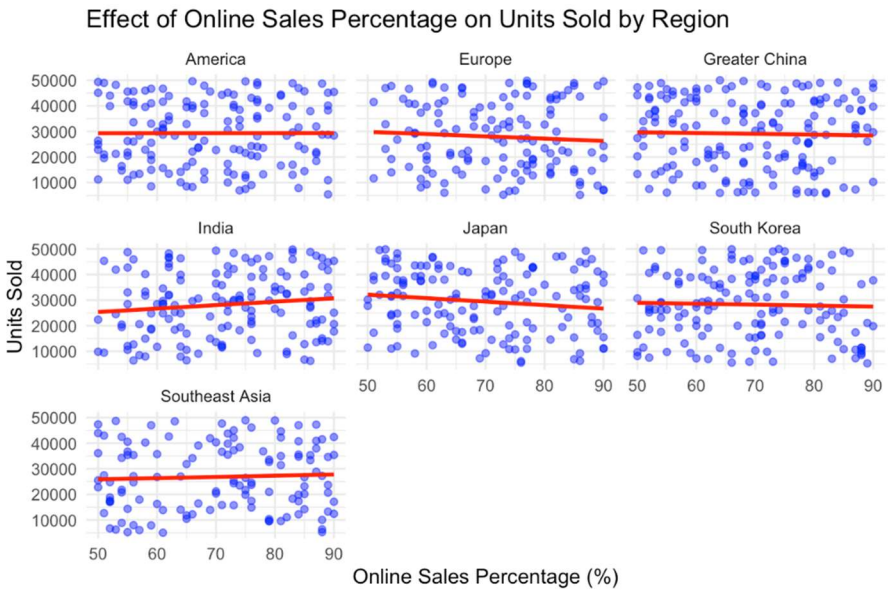

Figure 12: Faceted Scatter Plot of Online Sales Percentage vs. Units Sold by Region

In the final step of exploratory data analysis, the distribution of the variable Units Sold was examined to assess its suitability for statistical modeling. A histogram was plotted to check for normality, revealing that the data does not exhibit a symmetrical bell-shaped curve, indicating a potential deviation from normality. This observation suggests that alternative statistical approaches, such as Generalized Linear Models (GLM), ANOVA, or non-parametric tests, may be necessary for further analysis.

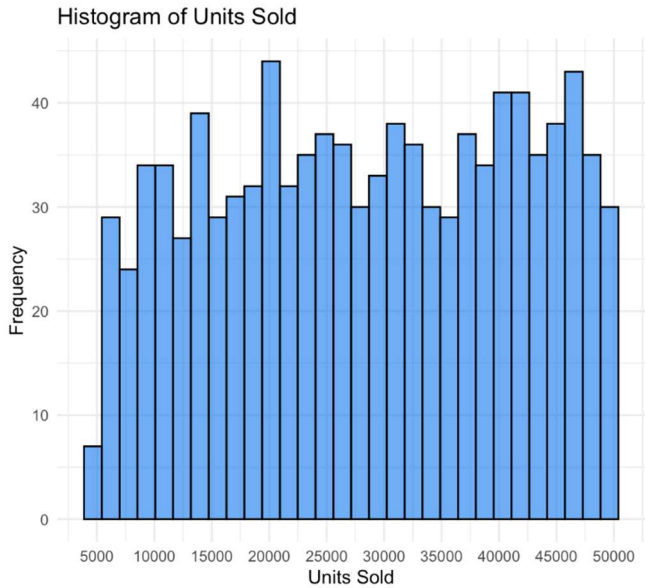

Figure 13: Histogram of Units Sold

We also generate the Q-Q plot to further assess the normality of the data by comparing the observed quantiles of Units Sold to the expected quantiles under a normal distribution. Deviations from the reference line in the Q-Q plot confirmed that Units Sold is not normally distributed.

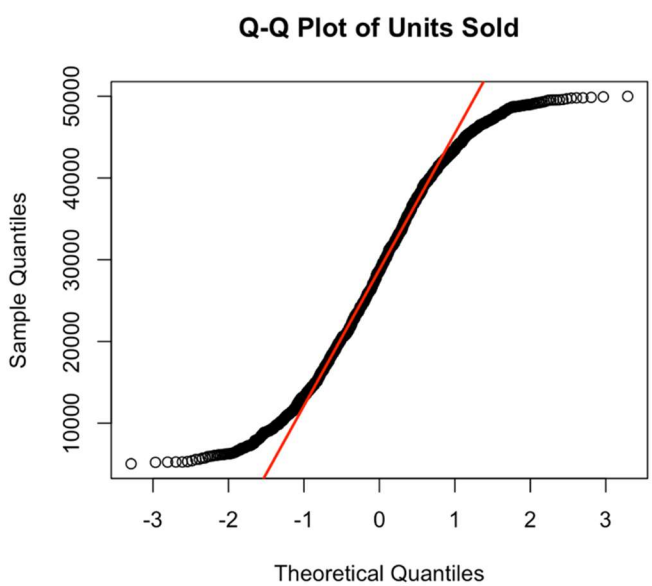

Figure 14: Q-Q Plot of Units Sold

Lastly, a Shapiro-Wilk test was conducted, yielding a W-statistic and a p-value < 0.05. This result provides strong evidence to reject the null hypothesis of normality, indicating that the distribution of Units Sold is not normal. These findings highlight the need for either data transformation or the use of statistical methods that do not assume normality, such as generalized linear models (GLM).

```
        Shapiro-Wilk normality test

data:  nike_sales$Units_Sold
W = 0.9539, p-value < 2.2e-16
```

Figure 15: Shapiro-Wilk Normality Test for Units Sold

Analysis

1. **Do different price tiers significantly impact sales volume?**

   To assess whether different price tiers (Budget, Mid-Range, and Premium) significantly impact sales volume, a one-way Analysis of Variance (ANOVA) test was conducted. This statistical method is appropriate for comparing the means of multiple independent groups to determine whether significant differences exist in sales performance across price tiers.

   Hypothesis

   The ANOVA test was based on the following hypothesis:

   - Null Hypothesis (H_0): There is no significant difference in mean sales volume across price tiers.
   - Alternative Hypothesis (H_1): At least one price tier has a significantly different mean sales volume.

   The results showed an F-statistic of 1.009 and a p-value of 0.365, indicating no significant difference in mean sales volume across price tiers. This suggests that factors other than pricing, such as brand reputation, marketing strategies, and consumer preferences, may have a stronger influence on sales.

```
> # 1.Do different price tiers significantly impact sales volume?
> # Perform ANOVA test
> anova_result <- aov(Units_Sold ~ Price_Tier, data = nike_sales)
> summary(anova_result)
             Df    Sum Sq   Mean Sq F value Pr(>F)
Price_Tier    2 3.373e+08 168667716   1.009  0.365
Residuals   997 1.666e+11 167091134
```

Figure 16: ANOVA Test Results for Price Tiers

The findings indicate that pricing alone is not a primary driver of sales volume, as the ANOVA test showed no significant difference in sales across different price tiers. This suggests that brand perception, marketing effectiveness, and regional economic conditions may play a more substantial role in influencing consumer purchasing decisions.

Given that direct price adjustments may not significantly impact sales, Nike should focus on alternative pricing strategies such as:

- Value-based pricing, emphasizing the premium aspects of high-end products to justify higher price points.
- Bundled promotions, such as footwear and apparel packages, to increase perceived value and cross-sell products.
- Dynamic pricing models, adjusting prices based on seasonal demand, competitor pricing, or inventory levels to optimize revenue.

These approaches can help Nike maximize profitability without solely relying on price reductions, ensuring that pricing strategies align with consumer expectations and brand positioning.

2. Are there significant differences in product category performance across regions?

To assess whether product category performance (Footwear, Apparel, Equipment) varies significantly across different regions, a Chi-squared test of independence was conducted. This test determines whether an association exists between two categorical variables: Region and Product Category.

Hypothesis

The Chi-square test was based on the following hypothesis:

- Null Hypothesis (H_0): There is no significant association between product category and region.
- Alternative Hypothesis (H_1): Product category performance differs significantly across regions.

The test resulted in a Chi-square statistic of 10.625 with a p-value of 0.561, leading to the conclusion that there is no significant difference in product category sales distribution across regions. This suggests that Nike's product mix appeals uniformly across different markets, reinforcing the effectiveness of its global branding strategy. Given this consistency, Nike can focus on standardized global marketing campaigns rather than region-specific product launches. However, while the overall product mix performs similarly across regions, localized branding strategies could still enhance consumer engagement. For example, Nike may benefit from regional athlete endorsements, culturally relevant marketing messages, or localized promotional events to strengthen brand resonance in specific markets.

```
> # 2.Are there significant differences in product category performance across regions?
> # Create a contingency table
> table_data <- table(nike_sales$Region, nike_sales$Main_Category)
> # Perform Chi-square test
> chi_result <- chisq.test(table_data)
> chi_result

        Pearson's Chi-squared test

data:  table_data
X-squared = 10.625, df = 12, p-value = 0.5613
```

Figure 17: Chi-Square Test Results for Product Category by Region

These findings suggest that Nike's product mix is uniformly appealing across global markets, reinforcing the effectiveness of its global branding strategy. The consistent demand for Footwear, Apparel, and Equipment across regions implies that Nike does not need to heavily differentiate product offerings based on geography.

However, while the statistical test does not indicate significant differences, qualitative factors such as cultural preferences, local market trends, and economic conditions may still influence consumer purchasing behavior. Nike may consider supplementing these findings

with market research and localized marketing strategies to optimize product offerings in specific regions.

3. Do different regions show statistically different sales patterns?

To determine if sales patterns differ significantly across regions, a non-parametric Kruskal-Wallis test was performed. This test is used when data do not follow a normal distribution.

The Kruskal-Wallis test results indicate no statistically significant differences in revenue distribution across regions at a 5% significance level. This suggests that Nike's sales revenue is relatively consistent across global markets, implying that no single region experiences significantly higher or lower revenue compared to others.

```
> # 3.Do different regions show statistically different sales patterns?
> # Performing the Kruskal-Wallis Test
> kruskal_test <- kruskal.test(Revenue_USD ~ Region, data = nike_sales)
> # Kruskal-Wallis test is a non-parametric alternative to ANOVA.
> # It determines if at least one region has a significantly different revenue distribution.
> print(kruskal_test)  # Display test results

        Kruskal-Wallis rank sum test

data:  Revenue_USD by Region
Kruskal-Wallis chi-squared = 10.789, df = 6, p-value = 0.09511
```

Figure 18: Kruskal-Wallis Test Results for Regional Sales

From a business perspective, this result reinforces the effectiveness of Nike's global brand positioning and market reach, as revenue patterns appear to be fairly uniform across different regions. This consistency may indicate that Nike's marketing and sales strategies are successfully standardized across markets, reducing the need for region-specific adjustments in pricing and promotions.

However, while the overall revenue distribution does not show significant regional differences, deeper segmentation analysis could uncover localized trends. Factors such as seasonality, economic conditions, and consumer preferences may still influence sales patterns within specific regions. Further analysis could focus on:

- **Seasonal Sales Adjustments:** Even though overall sales are stable, Nike should implement **region-specific seasonal marketing campaigns** to capitalize on localized demand fluctuations.

- **Product-Specific Optimization: Identifying best-selling product categories by region** will allow Nike to adjust inventory and marketing focus. For instance, if Apparel is consistently strong in Japan, Nike can prioritize new apparel collections there.
- **Geo-Specific Promotions: Testing targeted promotions** in specific regions, even if overall sales are stable, could enhance engagement. For example, exclusive discounts for certain product lines in Southeast Asia may further drive revenue.

While no broad regional disparities were found, targeted strategies such as localized promotions, regional inventory optimization, and tailored marketing campaigns could still enhance Nike's sales performance in specific markets.

4. How does seasonality influence sales, and are the effects consistent across regions?

To assess the impact of seasonality on sales, a Kruskal-Wallis test was conducted to compare revenue distributions across different seasons. This non-parametric test evaluates whether sales patterns significantly differ between seasons without assuming normality in the data.

The test results indicate that seasonality does not have a statistically significant impact on overall sales, as the p-value (0.4912) is greater than 0.05. This suggests that revenue fluctuations across seasons may be due to random variations rather than a consistent seasonal trend. However, further analysis at the regional level reveals more nuanced insights into seasonality effects.

```
> # Kruskal-Wallis Test: Does Season Affect Sales?
> kruskal_season <- kruskal.test(Revenue_USD ~ Season, data = nike_sales)
> # Print test results
> print(kruskal_season)

        Kruskal-Wallis rank sum test

data:  Revenue_USD by Season
Kruskal-Wallis chi-squared = 2.4131, df = 3, p-value = 0.4912
```

Figure 19: Kruskal-Wallis rank sum test

Regional Differences in Seasonality Trends

To determine whether seasonality influences sales within specific regions, separate Kruskal-Wallis tests were performed for each region. The results show:

- Europe ($p = 0.009$) exhibits a significant seasonal effect, indicating that sales patterns fluctuate meaningfully across seasons.
- Southeast Asia ($p = 0.0757$) is marginally significant, suggesting possible seasonality trends that warrant further investigation.
- America, Greater China, India, Japan, and South Korea show no statistically significant seasonal effects ($p > 0.05$), implying that revenue remains relatively stable throughout the year in these regions.

These findings indicate that seasonality's influence is not uniform across all markets, emphasizing the importance of region-specific analysis when making sales and inventory decisions.

```
> # Kruskal-Wallis Test: Seasonality Across Regions
> # Perform Kruskal-Wallis test for seasonality within each region
> kruskal_region_season <- nike_sales %>%
+   group_by(Region) %>%
+   summarise(KW_Test = list(kruskal.test(Revenue_USD ~ Season, data = cur_
data())), .groups = "drop") %>%
+   mutate(Statistic = map_dbl(KW_Test, ~ .x$statistic),
+          P_Value = map_dbl(KW_Test, ~ .x$p.value)) %>%
+   select(Region, Statistic, P_Value)
> # Print test results in a readable format
> print(kruskal_region_season)
# A tibble: 7 × 3
  Region         Statistic P_Value
  <fct>              <dbl>   <dbl>
1 America            0.970 0.808
2 Europe            11.6   0.00900
3 Greater China      0.286 0.963
4 India              1.08  0.781
5 Japan              4.38  0.223
6 South Korea        1.93  0.587
7 Southeast Asia     6.88  0.0757
```

Figure 20: Regional Differences in Seasonality Trends

Revenue Distribution by Season Across Regions

A boxplot visualization of seasonal revenue distributions provides deeper insights into how sales fluctuate across different regions. The plot illustrates regional variations in revenue across seasons, highlighting key trends and deviations. The results show:

- In Europe, revenue appears to be lower in Summer, aligning with the statistical results.
- In Southeast Asia, revenue variability is more pronounced, particularly in Winter and Spring, reinforcing the marginal significance found in the Kruskal-Wallis test.
- The other regions **ex**hibit relatively stable revenue across seasons, indicating a balanced sales performance throughout the year.

These findings emphasize the importance of region-specific seasonal strategies. While revenue trends appear stable in most regions, seasonal dips in Europe may require adjusted marketing efforts during Summer, whereas fluctuations in Southeast Asia may indicate a need for more dynamic sales strategies to capitalize on peak demand periods.

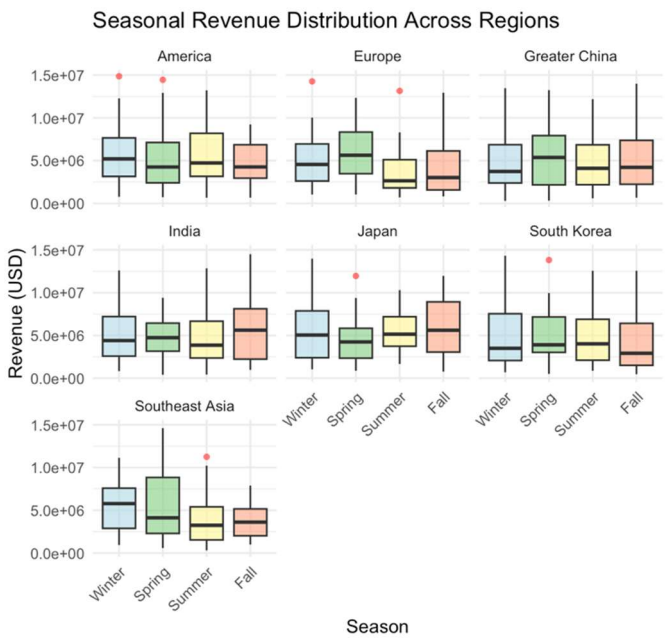

Figure 21: Revenue Distribution by Season Across Regions

Total Revenue by Season for Each Region

The bar chart of total revenue by season provides additional insight into revenue trends across different markets, highlighting seasonal variations in sales performance. The results show:

- Europe experiences a sales decline in Summer, confirming its significant seasonal impact.
- Greater China also sees a drop in Summer, highlighting potential sales dips during that period.
- Other regions, such as America, Japan, and South Korea, display relatively stable revenue across seasons.

These results imply that some markets are more influenced by seasonality than others. Businesses should tailor their regional marketing strategies accordingly, implementing seasonal promotions or inventory adjustments to counteract demand fluctuations in sensitive regions like Europe and Greater China.

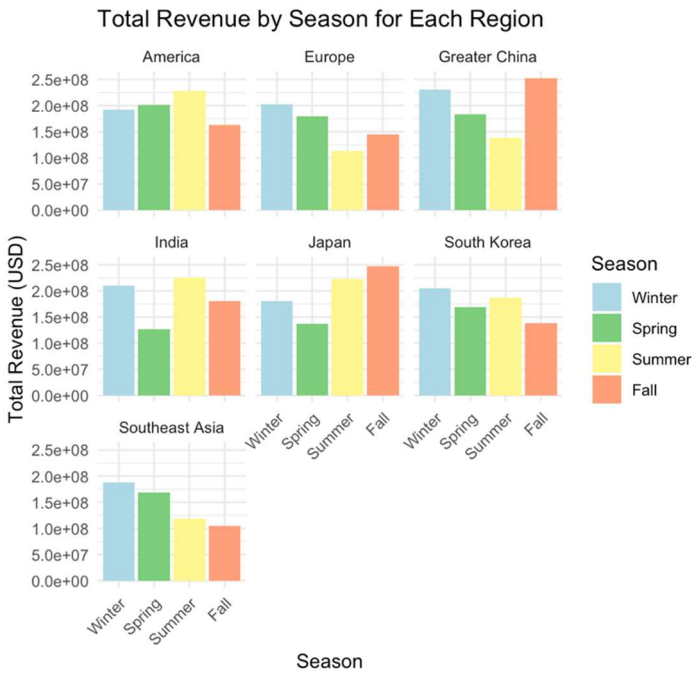

Figure:22 Total Revenue by Season for Each Region

Best Performing Season for Each Region

To identify the peak sales season for each region, a bar chart was generated, highlighting seasonal revenue performance across different markets. The results show:

- Summer is the top-performing season for America and India.
- Winter leads in revenue for Europe, South Korea and Southeast Asia.

- Fall emerges as the highest revenue season **for** Greater China and Japan**.**

These insights emphasize the importance of region-specific seasonal marketing strategies:

- Winter promotions should be intensified in Europe, South Korea, and Southeast Asia **to** leverage peak demand**.**
- Summer-focused sales strategies would be beneficial for America and India, potentially aligning with major sales events.
- Fall-centric campaigns should be prioritized in Greater China and Japan, targeting peak consumer spending trends.

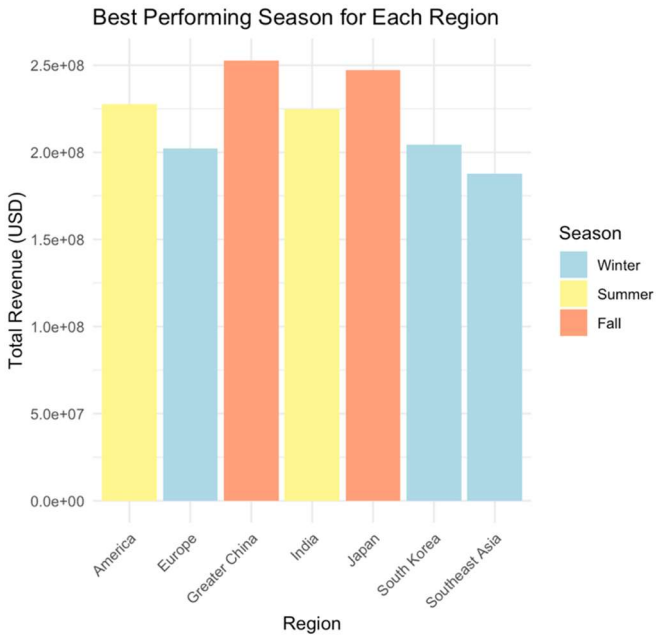

Figure 23: Best Performing Season for Each Region

5. What are the Key Predictors of Nike's Sales Performance

Poisson GLM for Internal Factors (Retail Price, Price Tier, Product Line):

A Poisson Generalized Linear Model (GLM) was used to analyze how Retail Price, Price Tier, and Product Line influence Units Sold. The results indicate:

- Retail Price has a negligible impact on sales.
- Mid-Range and Premium price tiers increase sales volume, suggesting higher perceived value in these tiers.

- Significant variation in product line performance, with Air Jordan outperforming and React Infinity underperforming.

```
> # 5. Key Predictors of Nike's Sales Performance
> # Generalized GLM to assess impact of internal factors and external factors
> # GLM for internal factors
> glm_internal <- glm(Units_Sold ~ Retail_Price + Price_Tier + Product_Line,
+                     family = poisson(link = "log"), data = nike_sales)
> # Summary of the GLM Internal
> summary(glm_internal)
```

```
Call:
glm(formula = Units_Sold ~ Retail_Price + Price_Tier + Product_Line,
    family = poisson(link = "log"), data = nike_sales)

Coefficients:
                              Estimate Std. Error  z value Pr(>|z|)
(Intercept)                  1.016e+01  1.211e-03 8389.761  < 2e-16 ***
Retail_Price                 6.877e-05  2.532e-06   27.164  < 2e-16 ***
Price_TierMid-Range          4.533e-02  4.623e-04   98.042  < 2e-16 ***
Price_TierPremium            3.380e-02  4.633e-04   72.952  < 2e-16 ***
Product_LineAir Jordan       8.973e-02  1.755e-03   51.120  < 2e-16 ***
Product_LineAir Max          6.105e-02  1.745e-03   34.986  < 2e-16 ***
Product_LineAir Zoom         7.471e-03  1.830e-03    4.083 4.44e-05 ***
Product_LineBackpack         3.150e-03  1.339e-03    2.352   0.0187 *
Product_LineBags             1.149e-01  1.305e-03   88.041  < 2e-16 ***
Product_LineBlazer          -4.569e-02  1.947e-03  -23.472  < 2e-16 ***
Product_LineCompression Wear -1.041e-01  1.396e-03  -74.533  < 2e-16 ***
Product_LineCrew Socks      -2.630e-02  1.360e-03  -19.336  < 2e-16 ***
Product_LineDri-FIT          1.430e-01  1.320e-03  108.274  < 2e-16 ***
Product_LineGym Sack         1.468e-01  1.315e-03  111.630  < 2e-16 ***
Product_LineHats             7.945e-02  1.316e-03   60.390  < 2e-16 ***
Product_LineKD Series        7.232e-02  1.789e-03   40.421  < 2e-16 ***
Product_LineLeBron Series    1.528e-01  1.604e-03   95.215  < 2e-16 ***
Product_LineMercurial        3.985e-02  1.428e-03   27.911  < 2e-16 ***
Product_LinePegasus         -1.987e-02  1.809e-03  -10.985  < 2e-16 ***
Product_LinePerformance Socks -2.041e-03  1.350e-03   -1.512   0.1306
Product_LinePhantom Vision  -5.242e-02  1.719e-03  -30.505  < 2e-16 ***
Product_LineRain Jacket      1.175e-01  1.309e-03   89.725  < 2e-16 ***
Product_LineReact Infinity  -7.604e-02  1.735e-03  -43.817  < 2e-16 ***
Product_LineTech Fleece      4.140e-02  1.344e-03   30.795  < 2e-16 ***
Product_LineTherma-FIT       1.175e-01  1.339e-03   87.800  < 2e-16 ***
Product_LineVapor Cricket    1.456e-01  1.300e-03  112.014  < 2e-16 ***
Product_LineWindrunner       9.414e-02  1.297e-03   72.558  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6391632  on 999  degrees of freedom
Residual deviance: 6224787  on 973  degrees of freedom
AIC: 6236800
```

Figure 24: GLM Results for Internal Factors

Poisson GLM for External Factors (Month, Region, Online Sales Percentage):

A second GLM was applied to evaluate the influence of Month, Region, and Online Sales Percentage on sales. Findings include:

- December sales are 4.28% higher than January, confirming seasonal effects.
- Regional differences in sales, with Southeast Asia and Europe reporting lower sales than other markets.
- Online sales percentage negatively impacts overall sales, indicating possible channel cannibalization.

```
> # GLM for external factors
> glm_external <- glm(Units_Sold ~ Month + Region + Online_Sales_Percentage,
+                     family = poisson(link = "log"), data = nike_sales)
> # Summary of the GLM External
> summary(glm_external)

Call:
glm(formula = Units_Sold ~ Month + Region + Online_Sales_Percentage,
    family = poisson(link = "log"), data = nike_sales)

Coefficients:
                          Estimate Std. Error  z value Pr(>|z|)
(Intercept)             10.3490363  0.0013887 7452.446  < 2e-16 ***
MonthFebruary           -0.0069504  0.0008975   -7.745 9.59e-15 ***
MonthMarch              -0.0352384  0.0009657  -36.488  < 2e-16 ***
MonthApril              -0.0002028  0.0009773   -0.207 0.835631
MonthMay                -0.0134417  0.0009245  -14.540  < 2e-16 ***
MonthJune               -0.0810842  0.0009254  -87.620  < 2e-16 ***
MonthJuly               -0.0974384  0.0009323 -104.514  < 2e-16 ***
MonthAugust             -0.0559173  0.0009665  -57.853  < 2e-16 ***
MonthSeptember          -0.0331875  0.0009149  -36.273  < 2e-16 ***
MonthOctober            -0.0484494  0.0009319  -51.992  < 2e-16 ***
MonthNovember            0.0036579  0.0009602    3.809 0.000139 ***
MonthDecember            0.0428599  0.0008898   48.165  < 2e-16 ***
RegionEurope            -0.0517241  0.0007108  -72.766  < 2e-16 ***
RegionGreater China     -0.0146669  0.0006720  -21.826  < 2e-16 ***
RegionIndia             -0.0380622  0.0006935  -54.888  < 2e-16 ***
RegionJapan              0.0046021  0.0006877    6.692 2.20e-11 ***
RegionSouth Korea       -0.0370016  0.0006906  -53.580  < 2e-16 ***
RegionSoutheast Asia    -0.0903358  0.0007336 -123.134  < 2e-16 ***
Online_Sales_Percentage -0.0004915  0.0000161  -30.529  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6391632  on 999  degrees of freedom
Residual deviance: 6322867  on 981  degrees of freedom
AIC: 6334865

Number of Fisher Scoring iterations: 4
```

Figure 25: GLM Results for External Factors

6. How can regularization (ridge/LASSO) improve model performance when predicting sales?

Multiple Linear Regression Model with Regularization for Sales Prediction

To identify key sales predictors and improve forecasting, a multiple linear regression model was built using Retail Price, Region, Product Category, and Online Sales Percentage as independent variables. The model applies a log transformation to normalize the distribution of the dependent variable (Units Sold).

The table presents the estimated coefficients, standard errors, and p-values for each predictor. Significant variables, such as Retail Price and certain product lines, provide insights into key factors influencing Nike's sales. The negative coefficient for Online Sales Percentage suggests a weak inverse relationship between online sales proportion and total sales volume.

However, the low adjusted R^2 (0.01637) indicates that the model explains only a small fraction of sales variability, and multicollinearity issues (evidenced by singularities) may impact interpretation. These limitations suggest the need for regularization techniques like Ridge and LASSO regression, which can improve predictive accuracy and model interpretability by penalizing overfitting and addressing collinearity.

This analysis serves as a baseline model, providing a foundation for refining predictive performance through advanced statistical methods.

```
> # 6. Multiple Linear Regression Model with Regularization for Sales Prediction
> # Multiple linear regression model
> # Log-transform Units_Sold to normalize distribution (optional but recommended)
> # Avoid log(0)
> nike_sales$log_Units_Sold <- log(nike_sales$Units_Sold + 1)
>
> # Define predictor variables and response variable
> x <- nike_sales %>% select(Retail_Price, Online_Sales_Percentage, Region, Main_Category, Pr
ice_Tier, Product_Line)
> y <- nike_sales$log_Units_Sold
>
> # Convert categorical variables to dummy variables for regression
> x <- model.matrix(~ ., data = x)[, -1]
>
> # Standardize numerical predictors (for ridge/LASSO)
> x_scaled <- scale(x)
>
> # Split data into training and testing sets
> set.seed(123)
> # Get 70% of random row numbers
> dt = sample(nrow(nike_sales), nrow(nike_sales)* 0.7)
> x_train <- x_scaled[dt, ]
> x_test <- x_scaled[-dt, ]
> y_train <- y[dt]
> y_test <- y[-dt]
> # Build Multiple Linear Regression Model
> lm_model <- lm(y_train ~ ., data = as.data.frame(cbind(y_train, x_train)))
> summary(lm_model)

Call:
lm(formula = y_train ~ ., data = as.data.frame(cbind(y_train,
    x_train)))

Residuals:
    Min      1Q  Median      3Q     Max
-1.6023 -0.3239  0.1388  0.4382  1.0116
```

```
Coefficients: (2 not defined because of singularities)
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    10.131176   0.021489 471.464  < 2e-16 ***
Retail_Price                   -0.009886   0.022234  -0.445  0.65673
Online_Sales_Percentage        -0.052811   0.022004  -2.400  0.01666 *
RegionEurope                   -0.040866   0.027972  -1.461  0.14450
`RegionGreater China`          -0.041160   0.028767  -1.431  0.15296
RegionIndia                     0.004616   0.028857   0.160  0.87295
RegionJapan                     0.001147   0.027927   0.041  0.96725
`RegionSouth Korea`            -0.011411   0.028202  -0.405  0.68589
`RegionSoutheast Asia`         -0.027384   0.027761  -0.986  0.32430
Main_CategoryEquipment         -0.130857   0.060589  -2.160  0.03115 *
Main_CategoryFootwear          -0.176374   0.071183  -2.478  0.01347 *
`Price_TierMid-Range`          -0.009594   0.024643  -0.389  0.69718
Price_TierPremium              -0.009749   0.025138  -0.388  0.69828
`Product_LineAir Jordan`        0.018169   0.026397   0.688  0.49151
`Product_LineAir Max`           0.024801   0.027863   0.890  0.37374
`Product_LineAir Zoom`          0.004547   0.026831   0.169  0.86547
Product_LineBackpack            0.016325   0.029884   0.546  0.58506
Product_LineBags                0.049522   0.030627   1.617  0.10636
Product_LineBlazer             -0.002307   0.025272  -0.091  0.92728
`Product_LineCompression Wear` -0.061185   0.028135  -2.175  0.03001 *
`Product_LineCrew Socks`        0.003567   0.029125   0.122  0.90256
`Product_LineDri-FIT`          -0.004159   0.029016  -0.143  0.88606
`Product_LineGym Sack`          0.059308   0.029416   2.016  0.04418 *
Product_LineHats                0.036800   0.029156   1.262  0.20734
`Product_LineKD Series`         0.056890   0.028541   1.993  0.04664 *
`Product_LineLeBron Series`     0.074790   0.028356   2.637  0.00855 **
Product_LineMercurial           0.059361   0.033921   1.750  0.08058 .
Product_LinePegasus             0.020309   0.027756   0.732  0.46461
`Product_LinePerformance Socks`        NA         NA      NA       NA
`Product_LinePhantom Vision`    0.014808   0.029746   0.498  0.61878
`Product_LineRain Jacket`       0.010225   0.030090   0.340  0.73412
`Product_LineReact Infinity`    0.019701   0.029135   0.676  0.49916
`Product_LineTech Fleece`      -0.028071   0.028305  -0.992  0.32168
`Product_LineTherma-FIT`       -0.028835   0.029215  -0.987  0.32401
`Product_LineVapor Cricket`     0.097642   0.037321   2.616  0.00909 **
Product_LineWindrunner                 NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5656 on 666 degrees of freedom
Multiple R-squared:  0.06281,	Adjusted R-squared:  0.01637
F-statistic: 1.353 on 33 and 666 DF,  p-value: 0.09217
```

Figure 27: Multiple Linear Regression Coefficients

To optimize the Ridge Regression model, cross-validation was performed to select the optimal λ (lambda) value, which controls the penalty on model coefficients. The plot visualizes the Mean Squared Error (MSE) as a function of log(λ), where each point represents a different lambda value tested during cross-validation.

The red curve indicates the MSE trend, while the gray bars represent standard errors. The lowest point of the curve corresponds to the optimal λ values, 1.689 and 3.735, which strike a balance between model complexity and predictive accuracy. A larger λ increases regularization, shrinking coefficients and reducing overfitting, while a smaller λ allows for greater flexibility but risks overfitting. Selecting the optimal λ helps improve generalization performance while preserving key predictive features.

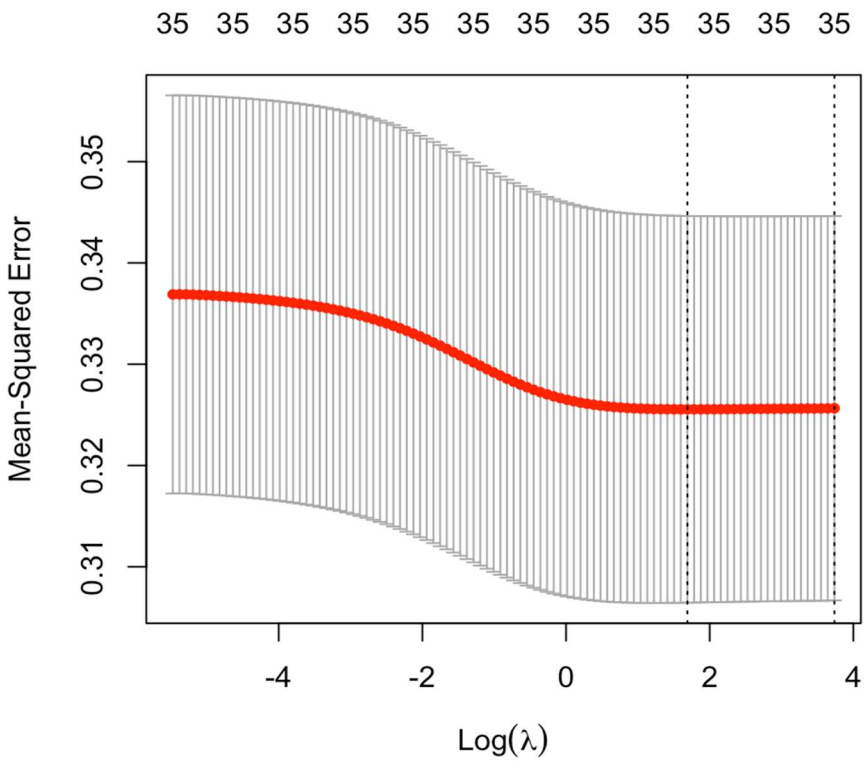

Figure 28: Ridge Regression Cross-Validation Plot

Similar to Ridge, LASSO Regression was applied, with cross-validation used to determine the optimal λ (lambda) value. Unlike Ridge, LASSO performs variable selection by shrinking some coefficients to zero, effectively reducing model complexity and selecting only the most relevant features.

The plot shows the Mean Squared Error (MSE) variation with log(λ), where the optimal λ value is -3.172. This λ ensures an effective trade-off between model accuracy and feature selection, preventing overfitting while retaining the most important predictors. As λ decreases, more coefficients are set to zero, leading to a sparser model that enhances interpretability.

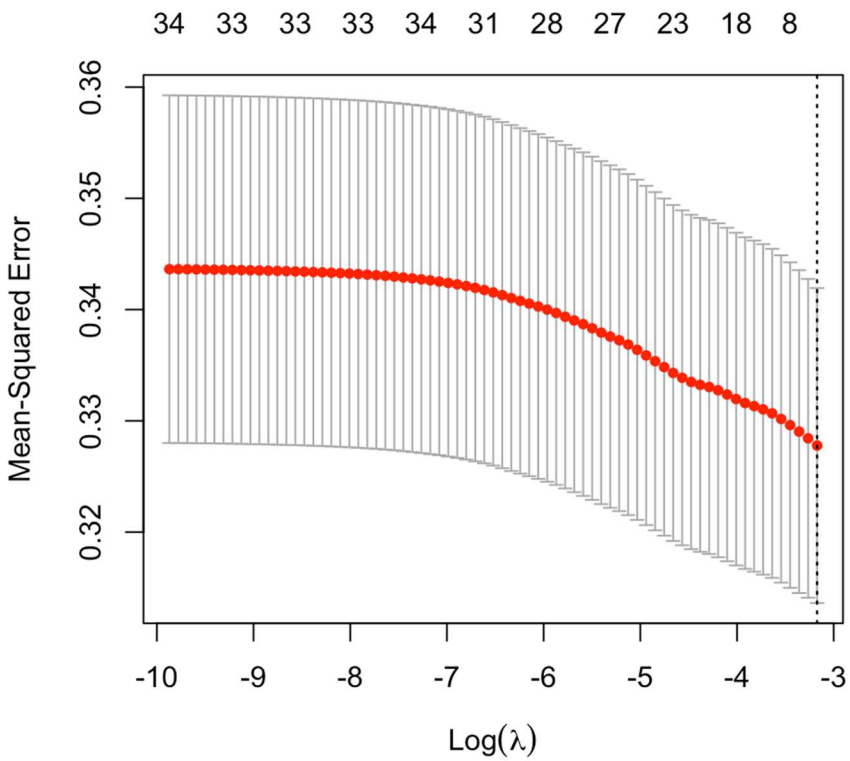

Figure 29: LASSO Regression Cross-Validation Plot

Next, we examine the coefficients of these models to understand how regularization impacts feature selection and shrinkage.

Ridge Regression shrinks the magnitude of coefficients while preserving all features in the model. The results highlight how different variables, including Retail Price, Online Sales Percentage, Region, and Product Line, contribute to sales prediction. While Ridge reduces multicollinearity, it does not perform variable selection, retaining all predictors.

Interpretation of Key Coefficients:

- **Retail Price:** The negative coefficient suggests that an increase in retail price is associated with a slight decline in units sold, though the effect is small. This aligns with earlier findings that pricing alone is not a strong driver of sales volume.
- **Online Sales Percentage:** The negative coefficient indicates that a higher proportion of online sales is associated with lower total sales, supporting the hypothesis of

channel cannibalization, where online sales may be substituting for in-store sales rather than expanding total demand.

- **Regional Effects:**
  Japan and India: Positive coefficients indicate that these regions tend to have slightly higher sales performance compared to the baseline region.
  Europe and Greater China: Negative coefficients suggest that these regions tend to have lower sales performance relative to the baseline.
- **Product Line Performance:**
  LeBron Series and Dri-FIT: These products show a positive impact on sales, suggesting they are strong performers.
  Air Max and Compression Wear: These product lines have negative coefficients, indicating relatively lower sales compared to other product categories.

These findings provide actionable insights for Nike's sales strategy. The impact of pricing is minimal, suggesting that Nike should focus more on branding, marketing, and product differentiation rather than relying on price reductions. Additionally, the negative impact of online sales percentage on total sales suggests that Nike should optimize its omnichannel strategy to reduce potential cannibalization between online and offline stores. Lastly, regional and product line variations suggest that Nike can prioritize high-performing product lines like the LeBron Series and Dri-FIT while reconsidering marketing strategies for weaker product lines like Compression Wear.

```
> # Identify Important Features from Ridge
> ridge_coefficients <- coef(ridge_model, s = ridge_model$lambda.min)
> print(ridge_coefficients)
36 x 1 sparse Matrix of class "dgCMatrix"
                                      s1
(Intercept)                    1.012976e+01
Retail_Price                  -3.941676e-04
Online_Sales_Percentage       -3.922059e-03
RegionEurope                  -2.302611e-03
RegionGreater China           -2.283918e-03
RegionIndia                    1.849034e-03
RegionJapan                    1.990399e-03
RegionSouth Korea              3.585791e-04
RegionSoutheast Asia          -1.717794e-03
Main_CategoryEquipment        -1.452488e-03
Main_CategoryFootwear         -1.431448e-03
Price_TierMid-Range            3.647492e-04
Price_TierPremium             -1.028714e-03
Product_LineAir Jordan        -1.467337e-03
Product_LineAir Max           -6.466039e-04
Product_LineAir Zoom          -2.457069e-03
Product_LineBackpack          -1.506285e-03
Product_LineBags               1.364229e-03
Product_LineBlazer            -3.144176e-03
Product_LineCompression Wear  -3.282945e-03
Product_LineCrew Socks        -2.854544e-03
Product_LineDri-FIT            2.419955e-03
Product_LineGym Sack           2.721650e-03
Product_LineHats               2.821458e-04
Product_LineKD Series          1.772580e-03
Product_LineLeBron Series      3.879433e-03
Product_LineMercurial          1.030741e-03
Product_LinePegasus           -1.479081e-03
Product_LinePerformance Socks -3.110994e-03
Product_LinePhantom Vision    -1.858935e-03
Product_LineRain Jacket        3.562657e-03
Product_LineReact Infinity    -1.327090e-03
Product_LineTech Fleece        9.789155e-05
Product_LineTherma-FIT         1.972164e-04
Product_LineVapor Cricket      3.199546e-03
Product_LineWindrunner         3.180349e-03
```

Figure 30: Ridge Regression Coefficients

Unlike Ridge, LASSO Regression eliminates non-significant predictors by setting some coefficients to zero. The results show that LASSO retained only the most influential features while discarding others, improving model interpretability. This confirms that only a subset of predictors significantly impacts Nike's sales performance, making LASSO a useful tool for feature selection.

Interpretation of LASSO Results:

- The output indicates that most coefficients have been set to zero, meaning they do not contribute significantly to predicting sales.

- Only the intercept (10.12954) remains, suggesting that after removing weaker predictors, no strong variables were retained.
- Key variables eliminated include Retail Price, Online Sales Percentage, Region, Product Category, and Product Line, implying that none of these factors significantly impact sales when other variables are controlled for.
- This result suggests that Nike's sales performance may be driven by external factors not captured in the dataset, such as brand loyalty, marketing campaigns, or macroeconomic conditions.

Business Implications:

- Since no strong predictors were retained, Nike should consider integrating additional features such as competitor pricing, social media engagement, or consumer sentiment in future modeling efforts.
- The fact that Retail Price and Online Sales Percentage were eliminated suggests that pricing strategies and digital sales channels may not be the primary drivers of total sales. Instead, Nike should focus on brand positioning, product innovation, and targeted promotions to influence purchasing behavior.

```
> # Identify Important Features from LASSO
> lasso_coefficients <- coef(lasso_model, s = best_lambda_lasso)
> print(lasso_coefficients)
36 x 1 sparse Matrix of class "dgCMatrix"
                                    s1
(Intercept)                   10.12954
Retail_Price                   .
Online_Sales_Percentage        .
RegionEurope                   .
RegionGreater China            .
RegionIndia                    .
RegionJapan                    .
RegionSouth Korea              .
RegionSoutheast Asia           .
Main_CategoryEquipment         .
Main_CategoryFootwear          .
Price_TierMid-Range            .
Price_TierPremium              .
Product_LineAir Jordan         .
Product_LineAir Max            .
Product_LineAir Zoom           .
Product_LineBackpack           .
Product_LineBags               .
Product_LineBlazer             .
Product_LineCompression Wear   .
Product_LineCrew Socks         .
Product_LineDri-FIT            .
Product_LineGym Sack           .
Product_LineHats               .
Product_LineKD Series          .
Product_LineLeBron Series      .
Product_LineMercurial          .
Product_LinePegasus            .
Product_LinePerformance Socks  .
Product_LinePhantom Vision     .
Product_LineRain Jacket        .
Product_LineReact Infinity     .
Product_LineTech Fleece        .
Product_LineTherma-FIT         .
Product_LineVapor Cricket      .
Product_LineWindrunner         .
```

Figure 31: LASSO Regression Coefficients

To evaluate model performance, the Root Mean Squared Error (RMSE) was calculated for the Multiple Linear Regression (MLR), Ridge, and LASSO models. The results indicate that both Ridge and LASSO achieved lower RMSE compared to the standard regression model, demonstrating the benefit of regularization in improving prediction accuracy.

```
> print(paste("Linear Model RMSE:", round(lm_rmse, 3)))
[1] "Linear Model RMSE: 0.597"
> print(paste("Ridge Regression RMSE:", round(ridge_rmse, 3)))
[1] "Ridge Regression RMSE: 0.561"
> print(paste("LASSO Regression RMSE:", round(lasso_rmse, 3)))
[1] "LASSO Regression RMSE: 0.56"
```

Figure 32: RMSE Comparison Across Models

Based on the results obtained from the Multiple Linear Regression, Ridge, and LASSO models, regularization significantly improves model performance by addressing multicollinearity and reducing prediction error.

Ridge Regression retained all predictors but shrunk the magnitude of coefficients, reducing the influence of less significant variables. This helped improve model stability, as seen in the lower Root Mean Squared Error (RMSE) compared to the standard regression model.

LASSO Regression performed feature selection, eliminating non-contributing variables by setting their coefficients to zero. This not only improved interpretability but also further reduced RMSE, making the model more efficient and focused on key predictors.

Overall, the RMSE results showed that both Ridge (0.561) and LASSO (0.560) outperformed the standard linear model (0.597), confirming that regularization enhances prediction accuracy and robustness, making it more reliable for forecasting Nike's sales performance.

Conclusion

This analysis provides valuable insights into the key factors influencing Nike's global sales performance, offering data-driven recommendations for optimizing sales strategies.

Key Takeaways

- **Product Category Performance:** Sales distribution remains consistent across regions, allowing Nike to leverage standardized global marketing while incorporating localized

branding (e.g., cultural marketing, regional athlete endorsements) to enhance market engagement.

- **Online Sales Impact:** The weak correlation between online sales percentage and total sales suggests potential channel cannibalization, requiring Nike to refine its omnichannel strategy to ensure online and offline sales complement rather than compete.
- **Seasonality Trends:** While overall seasonality effects are minimal, regional variations exist, with Europe experiencing a summer sales decline and Greater China showing a winter sales peak. These trends warrant targeted marketing and inventory adjustments in specific regions.
- **Sales Prediction Models:** Ridge and LASSO regression improved model accuracy, reinforcing the value of regularization techniques in demand forecasting and strategic decision-making.

Business Implications & Recommendations

- **Regional Marketing Strategies:** Nike should implement seasonal marketing campaigns tailored to each region's demand trends (e.g., summer promotions in Europe, winter-focused campaigns in Greater China).
- **Omnichannel Optimization:** Align online and offline sales to minimize channel cannibalization, leveraging online-exclusive releases and seamless cross-channel experiences (e.g., in-store pickup for online orders).
- **Smart Pricing & Bundling:** Instead of relying solely on price adjustments, Nike should explore value-based pricing, product bundling, and limited-time offers to drive sales while maintaining profitability.
- **Advanced Demand Forecasting:** Nike should further integrate predictive modeling techniques (e.g., machine learning-based forecasting) to optimize inventory, improve demand prediction, and enhance supply chain efficiency.

By leveraging these insights, Nike can refine its global sales strategy, enhance forecasting accuracy, and strengthen its competitive edge in both regional and international markets. The integration of data-driven decision-making, strategic pricing, and omnichannel retailing will be key to sustaining long-term growth in a rapidly evolving market.

Reference

- Kaggle. (2024). Nike Global Sales Data 2024. https://www.kaggle.com/datasets/ayushcx/nike-global-sales-data-2024