Sales Data Analysis using R Studio

Caleb Kiplangat

2025-06-28

1.Loading Required Packages

```
library(tidyverse)     # For data manipulation and visualization
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.2     ✔ tibble    3.3.0
## ✔ lubridate 1.9.4     ✔ tidyr     1.3.1
## ✔ purrr     1.0.4
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to be
come errors
```

```
library(lubridate)     # For working with dates
library(caret)         # For machine learning modeling
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(ggplot2)       # For plotting
library(cluster)       # For clustering
library(forecast)      # For time series forecasting
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(readr)         # For reading CSV files
```

2.Loading the Dataset

```
# Set working directory or use relative path
sales_data <- read_csv("C:/Users/NEPHIC  840G3/Desktop/FYJ/DATA_ANALYSIS FILES/sales_data.cs
v")
```

```
## Rows: 1000 Columns: 14
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (8): Sale_Date, Sales_Rep, Region, Product_Category, Customer_Type, Paym...
## dbl (6): Product_ID, Sales_Amount, Quantity_Sold, Unit_Cost, Unit_Price, Dis...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Preview the data
glimpse(sales_data)
```

```
## Rows: 1,000
## Columns: 14
## $ Product_ID          <dbl> 1052, 1093, 1015, 1072, 1061, 1021, 1083, 1087, 1…
## $ Sale_Date           <chr> "03/02/2023", "21/04/2023", "21/09/2023", "24/08/…
## $ Sales_Rep           <chr> "Bob", "Bob", "David", "Bob", "Charlie", "Charlie…
## $ Region              <chr> "North", "West", "South", "South", "East", "West"…
## $ Sales_Amount        <dbl> 5053.97, 4384.02, 4631.23, 2167.94, 3750.20, 3761…
## $ Quantity_Sold       <dbl> 18, 17, 30, 39, 13, 32, 29, 46, 30, 18, 13, 43, 2…
## $ Product_Category    <chr> "Furniture", "Furniture", "Food", "Clothing", "El…
## $ Unit_Cost           <dbl> 152.75, 3816.39, 261.56, 4330.03, 637.37, 900.79,…
## $ Unit_Price          <dbl> 267.22, 4209.44, 371.40, 4467.75, 692.71, 1106.51…
## $ Customer_Type       <chr> "Returning", "Returning", "Returning", "New", "Ne…
## $ Discount            <dbl> 0.09, 0.11, 0.20, 0.02, 0.08, 0.21, 0.14, 0.12, 0…
## $ Payment_Method      <chr> "Cash", "Cash", "Bank Transfer", "Credit Card", "…
## $ Sales_Channel       <chr> "Online", "Retail", "Retail", "Retail", "Online",…
## $ Region_and_Sales_Rep <chr> "North-Bob", "West-Bob", "South-David", "South-Bo…
```

```
summary(sales_data)
```

```
##    Product_ID    Sale_Date          Sales_Rep           Region         
##  Min.   :1001   Length:1000        Length:1000        Length:1000       
##  1st Qu.:1024   Class :character   Class :character   Class :character  
##  Median :1051   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :1050                                                           
##  3rd Qu.:1075                                                           
##  Max.   :1100                                                           
##   Sales_Amount    Quantity_Sold   Product_Category     Unit_Cost      
##  Min.   : 100.1   Min.   : 1.00   Length:1000        Min.   :  60.28  
##  1st Qu.:2550.3   1st Qu.:13.00   Class :character   1st Qu.:1238.38  
##  Median :5019.3   Median :25.00   Mode  :character   Median :2467.24  
##  Mean   :5019.3   Mean   :25.36                      Mean   :2475.30  
##  3rd Qu.:7507.4   3rd Qu.:38.00                      3rd Qu.:3702.86  
##  Max.   :9989.0   Max.   :49.00                      Max.   :4995.30  
##    Unit_Price     Customer_Type         Discount      Payment_Method    
##  Min.   : 167.1   Length:1000        Min.   :0.0000   Length:1000       
##  1st Qu.:1509.1   Class :character   1st Qu.:0.0800   Class :character  
##  Median :2696.4   Mode  :character   Median :0.1500   Mode  :character  
##  Mean   :2728.4                      Mean   :0.1524                     
##  3rd Qu.:3958.0                      3rd Qu.:0.2300                     
##  Max.   :5442.1                      Max.   :0.3000                     
##  Sales_Channel      Region_and_Sales_Rep
##  Length:1000        Length:1000         
##  Class :character   Class :character    
##  Mode  :character   Mode  :character    
##                                         
##                                         
## 
```

3.Data Cleaning and Transformation

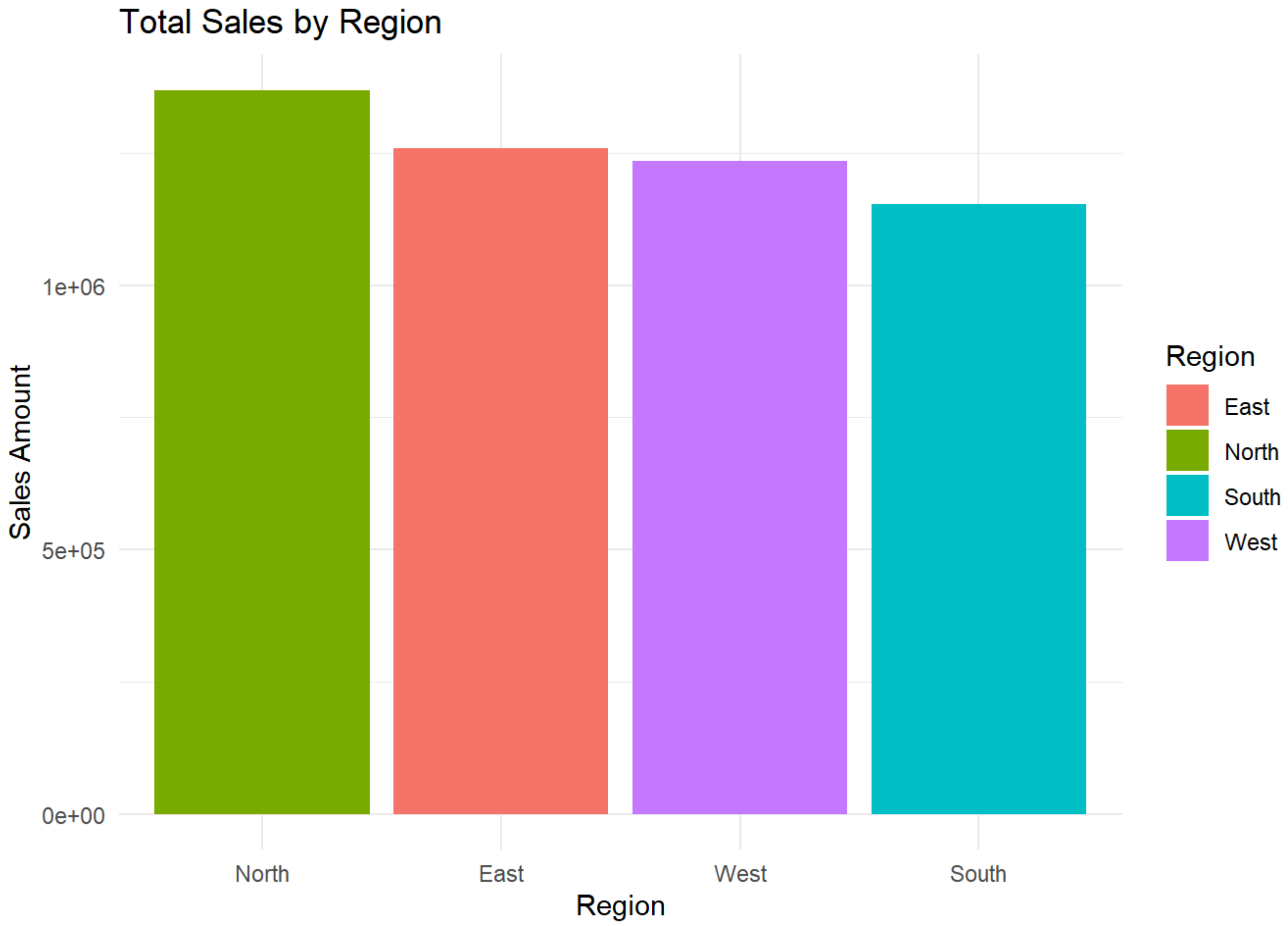
```
# Convert Sale_Date to Date format
sales_data <- sales_data %>%
  mutate(Sale_Date = dmy(Sale_Date))

# Create Profit and Profit_Margin
sales_data <- sales_data %>%
  mutate(
    Profit = (Unit_Price - Unit_Cost) * Quantity_Sold,
    Profit_Margin = Profit / Sales_Amount
  )
```

4. Exploratory Data Analysis (EDA)

Total Sales by Region

```
# Total sales by region
sales_data %>%
  group_by(Region) %>%
  summarise(Total_Sales = sum(Sales_Amount)) %>%
  ggplot(aes(x = reorder(Region, -Total_Sales), y = Total_Sales, fill = Region)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Total Sales by Region", x = "Region", y = "Sales Amount")
```



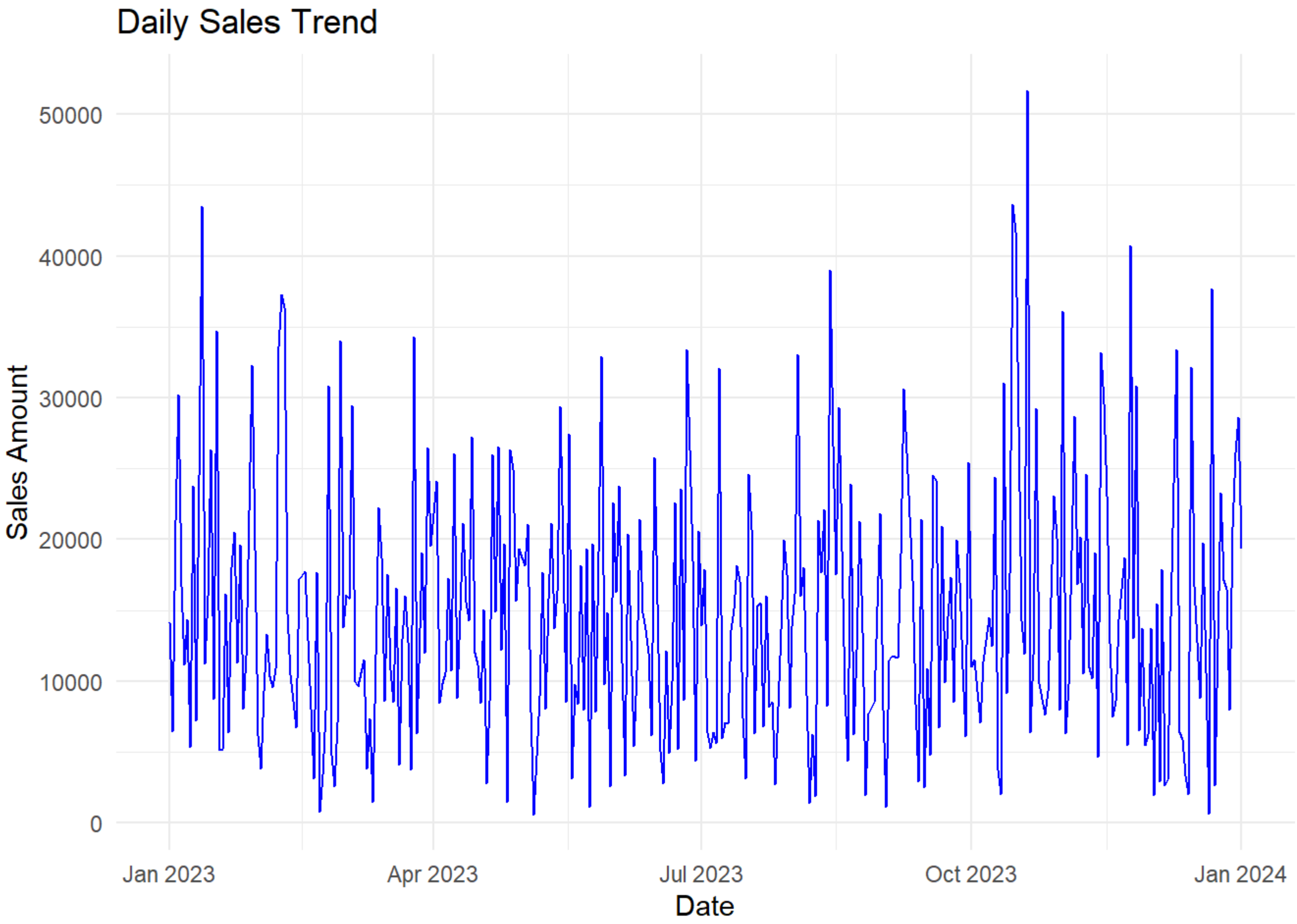
Interpretation: Total Sales by Region The bar chart above summarizes total sales across all regions in the dataset. Each bar represents a region. The height of the bar reflects the cumulative sales amount from that region. Regions are reordered from highest to lowest sales, so you can quickly see which regions are top performers.

Key Takeaways: The region at the far left of the chart is the top-performing region in terms of total sales. The differences in bar heights reveal how sales are distributed — if a few regions have much taller bars, it suggests sales are concentrated in specific areas. Regions with shorter bars may be underperforming or represent smaller markets.

Recommendation: Investigate why top regions are performing better — is it due to stronger reps, more customers, or product mix? For lower-performing regions, consider targeted strategies like discounts, rep training, or more product availability to boost sales. Use this insight to guide resource allocation and regional marketing plans

Daily Sales Trend

```
# Sales trend over time
sales_data %>%
  group_by(Sale_Date) %>%
  summarise(Daily_Sales = sum(Sales_Amount)) %>%
  ggplot(aes(x = Sale_Date, y = Daily_Sales)) +
  geom_line(color = "blue") +
  theme_minimal() +
  labs(title = "Daily Sales Trend", x = "Date", y = "Sales Amount")
```



Interpretation: Daily Sales Trend The line chart above shows how total sales have changed over time, based on the Sale_Date column. The x-axis represents dates (chronological order). The y-axis shows the total sales made on each day. The blue line connects daily sales amounts, making it easier to spot patterns, spikes, or drops over time.

Key Insights: Rising segments of the line indicate days where sales increased. Sharp dips may point to low-activity days — possibly weekends, holidays, or stock shortages. Consistent peaks may suggest strong performance on specific days or promotional cycles. A flat line or downward trend over a long period could be a concern worth investigating.

Recommendations: Identify dates with peak sales and examine what may have caused the spike — promotions, product launches, or high-demand periods. Look into low-sales periods and check for issues like downtime, supply delays, or reduced demand. Use these trends to support sales planning and forecasting — especially to prepare for high-demand days and optimize staffing or inventory.

5.Hypothesis Testing

Hypothesis 1: Discounts significantly increase Sales Amount

Hypothesis Statement:

Null (H_0): There is no significant difference in sales between transactions with and without a discount.

Alternative (H_1): Transactions with a discount have significantly different sales than those without

```
# Create two groups: with and without discount
sales_data <- sales_data %>%
  mutate(Has_Discount = ifelse(Discount > 0, "Yes", "No"))

# Run t-test
t_test_result <- t.test(Sales_Amount ~ Has_Discount, data = sales_data)
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  Sales_Amount by Has_Discount
## t = -0.80562, df = 15.623, p-value = 0.4326
## alternative hypothesis: true difference in means between group No and group Yes is not equ
al to 0
## 95 percent confidence interval:
##  -1876.5606   844.5073
## sample estimates:
##  mean in group No mean in group Yes
##          4511.495          5027.522
```

Results Summary: Group Means:

No Discount: 4,511.50

With Discount: 5,027.52

t-statistic: -0.81

Degrees of Freedom: ~15.6

p-value: 0.4326

95% Confidence Interval for the Difference in Means: [-1876.56, 844.51]

Interpretation:

The p-value (0.4326) is much greater than the typical threshold of 0.05.

This means the observed difference in average sales between discounted and non-discounted transactions is not statistically significant.

The confidence interval includes zero, which further supports that there’s no strong evidence of a difference.

Conclusion:

There is no statistically significant difference in average sales between transactions with and without a discount in your dataset.

Hypothesis 2: Average sales differ significantly across regions

Hypothesis Statement:

Null (H_0): All regions have the same average sales amount

Alternative (H_1): At least one region has a different average sales amount

```
# One-way ANOVA
anova_result <- aov(Sales_Amount ~ Region, data = sales_data)
summary(anova_result)
```

```
##              Df    Sum Sq Mean Sq F value Pr(>F)
## Region        3 1.931e+07 6435087   0.794  0.498
## Residuals   996 8.077e+09 8109242
```

```
# Optional: Post-hoc test to see which regions differ
TukeyHSD(anova_result)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Sales_Amount ~ Region, data = sales_data)
##
## $Region
##                  diff       lwr      upr     p adj
## North-East  339.54768 -297.0974 976.1928 0.5169653
## South-East  317.21750 -347.4666 981.9016 0.6090661
## West-East   273.88384 -377.4808 925.2484 0.7006096
## South-North -22.33018 -684.7092 640.0488 0.9997660
## West-North  -65.66384 -714.6760 583.3484 0.9938153
## West-South  -43.33367 -719.8725 633.2052 0.9984067
```

ANOVA Results Term Df Sum Sq Mean Sq F value Pr(>F) Region 3 19,310,000 6,435,087 0.794 0.498 Residual 996 8,077,000,000 8,109,242

p-value: 0.498 — this is much greater than 0.05

F value: 0.794 — shows weak between-group variance compared to within-group variance

Interpretation:

There is no statistically significant difference in average sales amounts across the regions. The variation in sales within regions is far greater than the variation between them.

Tukey Post-Hoc Results Interpretation

None of the pairwise regional comparisons show a statistically significant difference in mean sales.

All confidence intervals include 0, and adjusted p-values are much higher than 0.05.

This confirms what ANOVA already indicated — regional differences in average sales are not meaningful.

Conclusion:

There’s no strong evidence that any region outperforms others significantly in terms of average sales.

The difference in sales across regions is not statistically meaningful based on this dataset.

6. Predictive Modeling: Regression

```
# Linear regression model to predict Sales_Amount
model_data <- sales_data %>%
  select(Sales_Amount, Quantity_Sold, Unit_Cost, Unit_Price, Discount)

model <- lm(Sales_Amount ~ ., data = model_data)
summary(model)
```

```
## 
## Call:
## lm(formula = Sales_Amount ~ ., data = model_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5178.0 -2504.3    26.4  2493.7  4981.7 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   4942.7647   329.2024  15.014   <2e-16 ***
## Quantity_Sold   -8.6305     6.3785  -1.353    0.176    
## Unit_Cost       -0.4274     0.6403  -0.668    0.505    
## Unit_Price       0.4543     0.6397   0.710    0.478    
## Discount       747.5224  1033.6741   0.723    0.470    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2848 on 995 degrees of freedom
## Multiple R-squared:  0.002911,   Adjusted R-squared:  -0.001097 
## F-statistic: 0.7263 on 4 and 995 DF,  p-value: 0.574
```

Model Statistics:

Multiple R-squared: 0.0029

→ The model explains only 0.29% of the variation in Sales_Amount, which is extremely low.

Adjusted R-squared: -0.0011

→ After adjusting for number of predictors, the model does worse than a horizontal average line.

F-statistic p-value: 0.574

→ The model is not statistically significant overall.

Interpretation:

None of the individual predictors significantly explain changes in Sales_Amount.

The overall model does not provide useful predictive power.

This suggests that other unmeasured factors are likely driving sales — not just quantity, price, cost, or discount.

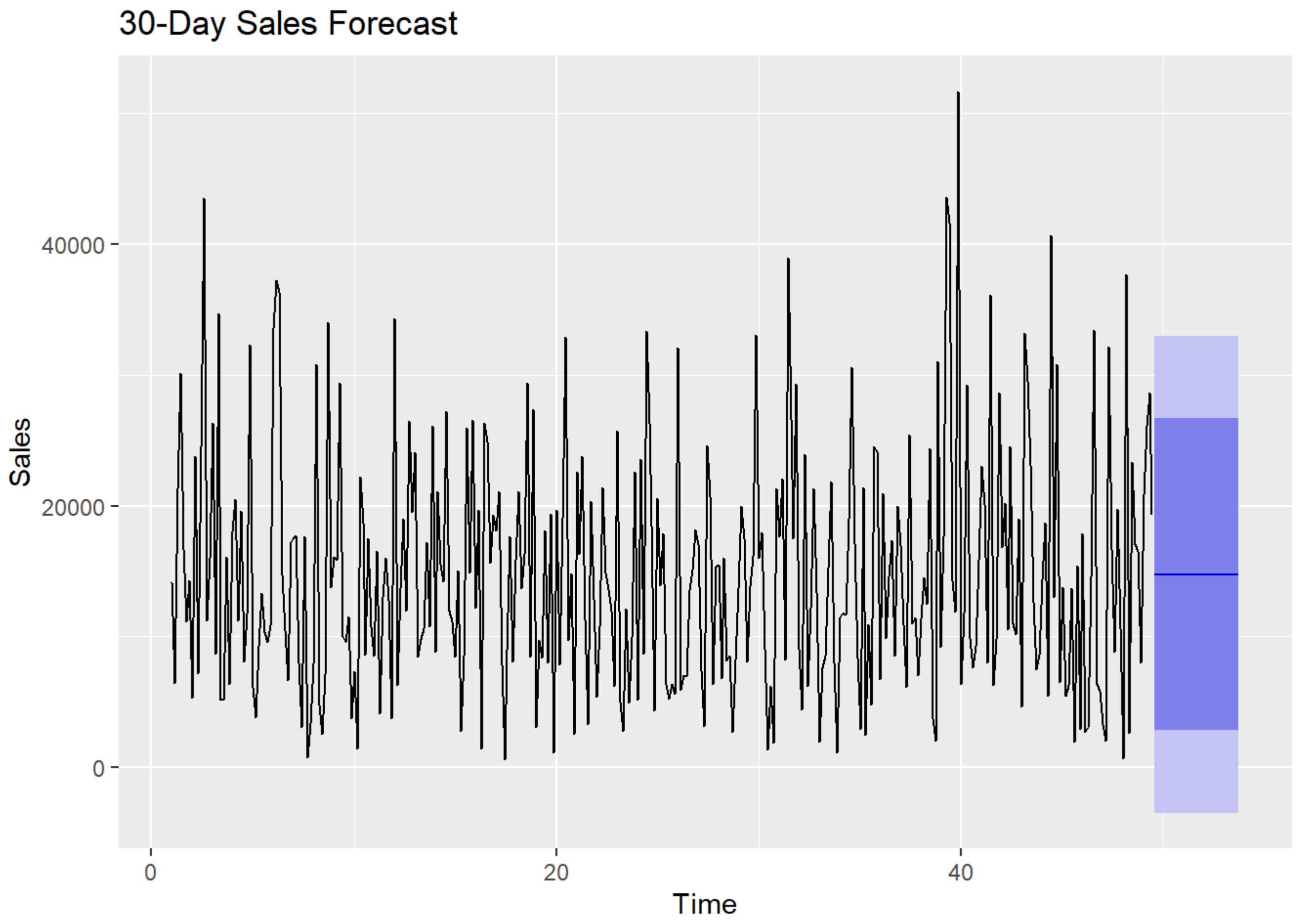
7. Time Series Forecasting

```
# Aggregate by date
ts_data <- sales_data %>%
  group_by(Sale_Date) %>%
  summarise(Total_Sales = sum(Sales_Amount))

# Convert to time series object
sales_ts <- ts(ts_data$Total_Sales, frequency = 7)

# Fit ARIMA model
fit <- auto.arima(sales_ts)
forecasted <- forecast(fit, h = 30)

# Plot forecast
autoplot(forecasted) +
  labs(title = "30-Day Sales Forecast", y = "Sales", x = "Time")
```



How to interpret the forecast plot:

Forecasted line: This shows the predicted total sales for each day in the next 30 days based on past sales trends.

Confidence intervals (usually shaded area): These bands give a range where the true sales values are expected to fall with a certain probability (usually 80% and 95% intervals). Wider bands mean more uncertainty.

Pattern insights:

If the forecasted line is increasing, it suggests sales are expected to rise.

If it’s flat or decreasing, sales are expected to stay steady or drop.

Model fit: The ARIMA model accounts for trends, seasonality (weekly), and noise in your data to make these predictions.