---  
title: "Sales Analysis and Forecasting"  
author: "Mark Abong'o"  
date: "2025-04-09"  
output: html\_document  
---  
  
  
  
## Installing and Loading Packages  
  
  
``` r  
# install.packages("forecast") # Uncomment if not already installed  
  
library(ggplot2)  
library(tidyverse)  
library(readr)  
library(forecast)

## Loading Sales Data

sales\_data <- read\_csv("D:/Documents/train.csv") # Make sure path is correct

## Rows: 9800 Columns: 18  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (15): Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Customer ...  
## dbl (3): Row ID, Postal Code, Sales  
##   
## 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.

## Data Exploration

dim(sales\_data)

## [1] 9800 18

head(sales\_data)

## # A tibble: 6 × 18  
## `Row ID` `Order ID` `Order Date` `Ship Date` `Ship Mode` `Customer ID`  
## <dbl> <chr> <chr> <chr> <chr> <chr>   
## 1 1 CA-2017-152156 08/11/2017 11/11/2017 Second Class CG-12520   
## 2 2 CA-2017-152156 08/11/2017 11/11/2017 Second Class CG-12520   
## 3 3 CA-2017-138688 12/06/2017 16/06/2017 Second Class DV-13045   
## 4 4 US-2016-108966 11/10/2016 18/10/2016 Standard Class SO-20335   
## 5 5 US-2016-108966 11/10/2016 18/10/2016 Standard Class SO-20335   
## 6 6 CA-2015-115812 09/06/2015 14/06/2015 Standard Class BH-11710   
## # ℹ 12 more variables: `Customer Name` <chr>, Segment <chr>, Country <chr>,  
## # City <chr>, State <chr>, `Postal Code` <dbl>, Region <chr>,  
## # `Product ID` <chr>, Category <chr>, `Sub-Category` <chr>,  
## # `Product Name` <chr>, Sales <dbl>

names(sales\_data)

## [1] "Row ID" "Order ID" "Order Date" "Ship Date"   
## [5] "Ship Mode" "Customer ID" "Customer Name" "Segment"   
## [9] "Country" "City" "State" "Postal Code"   
## [13] "Region" "Product ID" "Category" "Sub-Category"   
## [17] "Product Name" "Sales"

glimpse(sales\_data)

## Rows: 9,800  
## Columns: 18  
## $ `Row ID` <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,…  
## $ `Order ID` <chr> "CA-2017-152156", "CA-2017-152156", "CA-2017-138688", …  
## $ `Order Date` <chr> "08/11/2017", "08/11/2017", "12/06/2017", "11/10/2016"…  
## $ `Ship Date` <chr> "11/11/2017", "11/11/2017", "16/06/2017", "18/10/2016"…  
## $ `Ship Mode` <chr> "Second Class", "Second Class", "Second Class", "Stand…  
## $ `Customer ID` <chr> "CG-12520", "CG-12520", "DV-13045", "SO-20335", "SO-20…  
## $ `Customer Name` <chr> "Claire Gute", "Claire Gute", "Darrin Van Huff", "Sean…  
## $ Segment <chr> "Consumer", "Consumer", "Corporate", "Consumer", "Cons…  
## $ Country <chr> "United States", "United States", "United States", "Un…  
## $ City <chr> "Henderson", "Henderson", "Los Angeles", "Fort Lauderd…  
## $ State <chr> "Kentucky", "Kentucky", "California", "Florida", "Flor…  
## $ `Postal Code` <dbl> 42420, 42420, 90036, 33311, 33311, 90032, 90032, 90032…  
## $ Region <chr> "South", "South", "West", "South", "South", "West", "W…  
## $ `Product ID` <chr> "FUR-BO-10001798", "FUR-CH-10000454", "OFF-LA-10000240…  
## $ Category <chr> "Furniture", "Furniture", "Office Supplies", "Furnitur…  
## $ `Sub-Category` <chr> "Bookcases", "Chairs", "Labels", "Tables", "Storage", …  
## $ `Product Name` <chr> "Bush Somerset Collection Bookcase", "Hon Deluxe Fabri…  
## $ Sales <dbl> 261.9600, 731.9400, 14.6200, 957.5775, 22.3680, 48.860…

## 🧹 Data Cleaning

missingValues <- !complete.cases(sales\_data)  
sales\_data[missingValues,]

## # A tibble: 11 × 18  
## `Row ID` `Order ID` `Order Date` `Ship Date` `Ship Mode` `Customer ID`  
## <dbl> <chr> <chr> <chr> <chr> <chr>   
## 1 2235 CA-2018-104066 05/12/2018 10/12/2018 Standard Class QJ-19255   
## 2 5275 CA-2016-162887 07/11/2016 09/11/2016 Second Class SV-20785   
## 3 8799 US-2017-150140 06/04/2017 10/04/2017 Standard Class VM-21685   
## 4 9147 US-2017-165505 23/01/2017 27/01/2017 Standard Class CB-12535   
## 5 9148 US-2017-165505 23/01/2017 27/01/2017 Standard Class CB-12535   
## 6 9149 US-2017-165505 23/01/2017 27/01/2017 Standard Class CB-12535   
## 7 9387 US-2018-127292 19/01/2018 23/01/2018 Standard Class RM-19375   
## 8 9388 US-2018-127292 19/01/2018 23/01/2018 Standard Class RM-19375   
## 9 9389 US-2018-127292 19/01/2018 23/01/2018 Standard Class RM-19375   
## 10 9390 US-2018-127292 19/01/2018 23/01/2018 Standard Class RM-19375   
## 11 9742 CA-2016-117086 08/11/2016 12/11/2016 Standard Class QJ-19255   
## # ℹ 12 more variables: `Customer Name` <chr>, Segment <chr>, Country <chr>,  
## # City <chr>, State <chr>, `Postal Code` <dbl>, Region <chr>,  
## # `Product ID` <chr>, Category <chr>, `Sub-Category` <chr>,  
## # `Product Name` <chr>, Sales <dbl>

sales\_data <- na.omit(sales\_data)  
  
# Confirm missing row (2235) is gone  
sales\_data[2235, 1]

## # A tibble: 1 × 1  
## `Row ID`  
## <dbl>  
## 1 2236

## Date Conversion

date\_columns <- grep("Date", names(sales\_data), value = TRUE, ignore.case = TRUE)  
date\_columns

## [1] "Order Date" "Ship Date"

sales\_data$`Order Date` <- as.Date(sales\_data$`Order Date`, format = "%d/%m/%Y")  
sales\_data$`Ship Date` <- as.Date(sales\_data$`Ship Date`, format = "%d/%m/%Y")  
  
glimpse(sales\_data)

## Rows: 9,789  
## Columns: 18  
## $ `Row ID` <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,…  
## $ `Order ID` <chr> "CA-2017-152156", "CA-2017-152156", "CA-2017-138688", …  
## $ `Order Date` <date> 2017-11-08, 2017-11-08, 2017-06-12, 2016-10-11, 2016-…  
## $ `Ship Date` <date> 2017-11-11, 2017-11-11, 2017-06-16, 2016-10-18, 2016-…  
## $ `Ship Mode` <chr> "Second Class", "Second Class", "Second Class", "Stand…  
## $ `Customer ID` <chr> "CG-12520", "CG-12520", "DV-13045", "SO-20335", "SO-20…  
## $ `Customer Name` <chr> "Claire Gute", "Claire Gute", "Darrin Van Huff", "Sean…  
## $ Segment <chr> "Consumer", "Consumer", "Corporate", "Consumer", "Cons…  
## $ Country <chr> "United States", "United States", "United States", "Un…  
## $ City <chr> "Henderson", "Henderson", "Los Angeles", "Fort Lauderd…  
## $ State <chr> "Kentucky", "Kentucky", "California", "Florida", "Flor…  
## $ `Postal Code` <dbl> 42420, 42420, 90036, 33311, 33311, 90032, 90032, 90032…  
## $ Region <chr> "South", "South", "West", "South", "South", "West", "W…  
## $ `Product ID` <chr> "FUR-BO-10001798", "FUR-CH-10000454", "OFF-LA-10000240…  
## $ Category <chr> "Furniture", "Furniture", "Office Supplies", "Furnitur…  
## $ `Sub-Category` <chr> "Bookcases", "Chairs", "Labels", "Tables", "Storage", …  
## $ `Product Name` <chr> "Bush Somerset Collection Bookcase", "Hon Deluxe Fabri…  
## $ Sales <dbl> 261.9600, 731.9400, 14.6200, 957.5775, 22.3680, 48.860…

## Exploratory Data Analysis

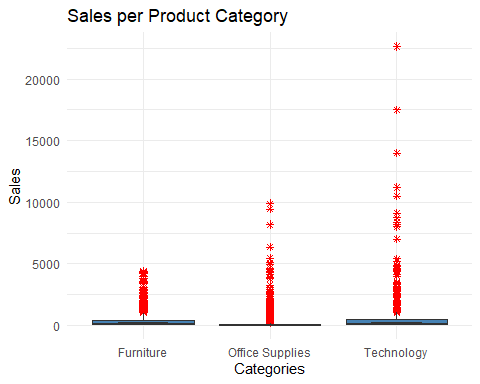
### Summary of Sales

summary(sales\_data$Sales)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.444 17.248 54.384 230.116 210.392 22638.480

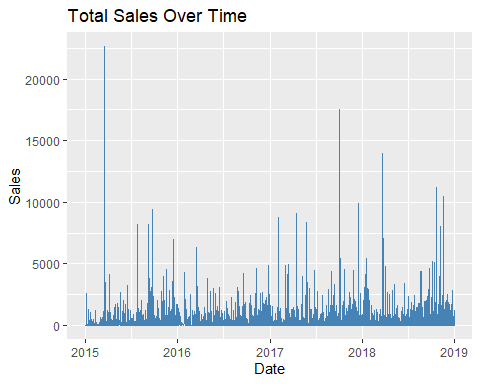
### Boxplot: Sales per Product Category

ggplot(sales\_data, aes(x = Category, y = Sales)) +  
 geom\_boxplot(fill = "steelblue", width = 0.8, notch = TRUE, outlier.colour = "red", outlier.shape = 8) +  
 labs(title = "Sales per Product Category", x = "Categories", y = "Sales") +  
 theme\_minimal()



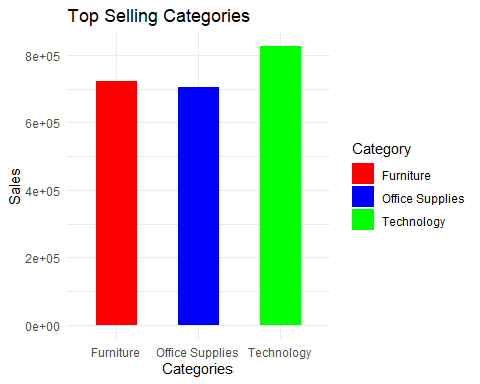
### Line Chart: Total Sales Over Time

ggplot(sales\_data, aes(x = `Order Date`, y = Sales)) +  
 geom\_line(color = "steelblue") +  
 labs(title = "Total Sales Over Time", x = "Date", y = "Sales")



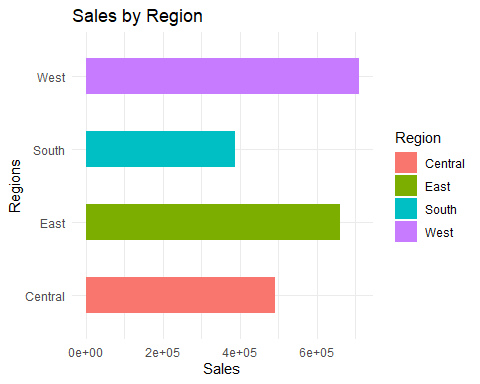
### Bar Chart: Top Selling Categories

ggplot(sales\_data, aes(x = Category, y = Sales, fill = Category)) +  
 geom\_bar(stat = "identity", width = 0.5) +  
 scale\_fill\_manual(values = c("red", "blue", "green")) +  
 labs(title = "Top Selling Categories", x = "Categories", y = "Sales") +  
 theme\_minimal()



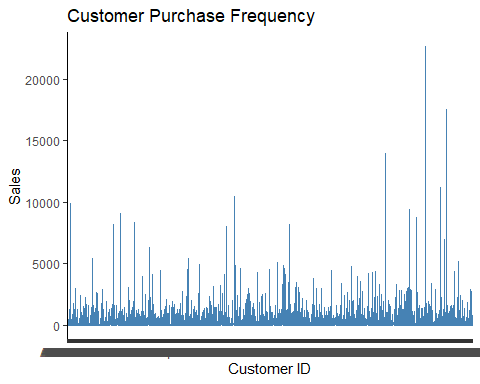
### Bar Chart: Sales by Region

ggplot(sales\_data, aes(x = Region, y = Sales, fill = Region)) +  
 geom\_bar(stat = "identity", width = 0.5) +  
 labs(title = "Sales by Region", x = "Regions", y = "Sales") +  
 coord\_flip() +  
 theme\_minimal()



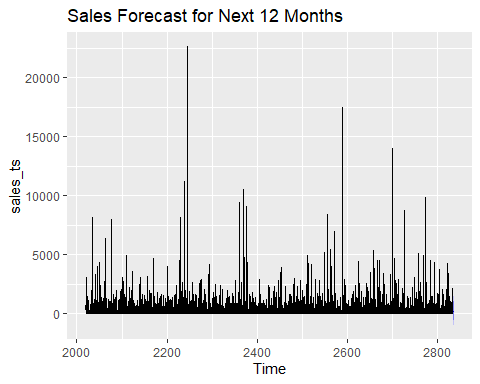
### Line Chart: Customer Purchase Frequency

ggplot(sales\_data, aes(x = `Customer ID`, y = Sales)) +  
 geom\_line(color = "steelblue") +  
 labs(title = "Customer Purchase Frequency", x = "Customer ID", y = "Sales") +  
 theme\_classic()



## Forecasting with ARIMA

sales\_ts <- ts(sales\_data$Sales, start = c(2020, 1), frequency = 12)  
arima\_model <- auto.arima(sales\_ts)  
forecasted\_sales <- forecast(arima\_model, h = 12)  
  
autoplot(forecasted\_sales) + labs(title = "Sales Forecast for Next 12 Months")



## Export Cleaned Data

write\_csv(sales\_data, "cleaned\_customer\_sales.csv")

```