PROJECT HOST ------- SATABDA MAJUMDER

PURPOSE ------- PORTFOLIO BUILDING

COMPLETION YEAR --------- 2025

CONTENTS

PROJECT TITLE

Sales Analysis and Forecasting of Amazon E-Commerce Dataset Using R

ABSTRACT

This project aims to perform comprehensive sales analytics on an Amazon e-commerce dataset to identify top-performing categories, purchasing trends, regional preferences, and seasonal sales variations. Statistical modeling and time series forecasting techniques are applied using R programming to extract actionable insights and predict future sales trends.

INTRODUCTION

E-commerce platforms like Amazon generate massive volumes of transaction data. Analyzing such data provides valuable insights into customer preferences, sales patterns, and business performance. This project explores a cleaned dataset from Amazon to understand customer behavior and sales dynamics through exploratory data analysis, statistical inference, and predictive modeling.

OBJECTIVES

- ❑ Identify top-selling categories and cities.
- ❑ Analyze revenue trends over time.
- ❑ Determine the most used payment methods.
- ❑ Use statistical methods to compare sales across categories.
- ❑ Build predictive models for forecasting future sales.
- ❑ Apply regression analysis to understand relationships among variables.

DATA SOURCE

The data used in this project is a cleaned CSV dataset titled amazon_sales_data_cleaned.csv collected from https://www.kaggle.com/.It includes fields like Date, Product Category, City, Payment Type, Price, Quantity Ordered, and Total Sales.

METHODOLOGY

1. Data Cleaning and Transformation: Handled invalid or canceled orders, parsed dates, and formatted month fields.

2. EDA (Exploratory Data Analysis): Used ggplot2, dplyr, and tidyverse for graphical representation and descriptive summaries.

3. Statistical Analysis: Applied ANOVA and correlation analysis to test hypotheses.

4. Time Series Modeling: Built an ARIMA model to forecast sales trends for the next three months.

5. Predictive Modeling: Built a linear regression model to understand the influence of various predictors on sales.

IMPORTANT STAISTICAL TOOLS USED IN THIS PROJECT

➢ ANOVA (Analysis of Variance):

- Compares mean values across multiple groups (e.g., product categories).
- Determines whether any group mean is significantly different from others.
- Helps detect patterns in category-wise sales performance.
- Used when there are more than two groups to compare.

➢ Pearson Correlation

- Measures the linear relationship between two continuous variables.
- Value ranges from -1 (strong negative) to +1 (strong positive).
- Identifies how price and quantity relate to total sales.
- Helps detect multicolinearity before modeling.

➢ ARIMA (Autoregressive Integrated Moving Average)

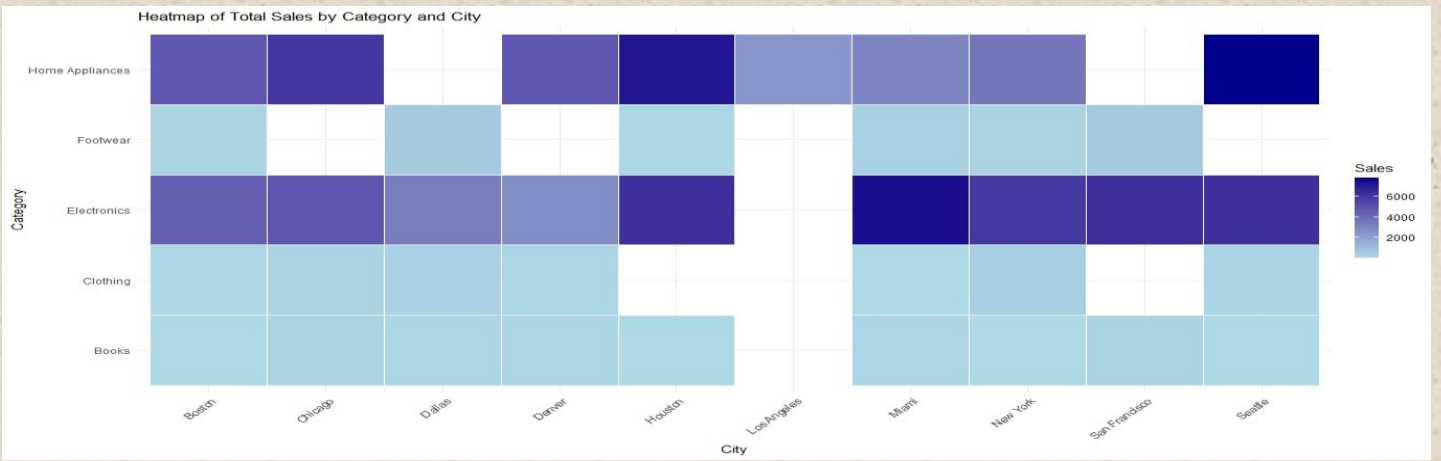
- A statistical model used for forecasting time series data.
- Captures trends, seasonality, and autocorrelation in sales.
- Useful for predicting future monthly/quarterly sales.
- Selected using model diagnostics like AIC, residuals.

➢ Linear Regression

- Models the relationship between a dependent variable (sales) and one or more independent variables (e.g., price, quantity, category).
- Provides coefficients to quantify impact of each variable.
- Outputs include R-squared (model fit) and p-values (variable significance).
- Foundation for predictive analytics in sales analysis.

DATA ANALYSIS FROM THE GRAPHICAL REPRESENTATION

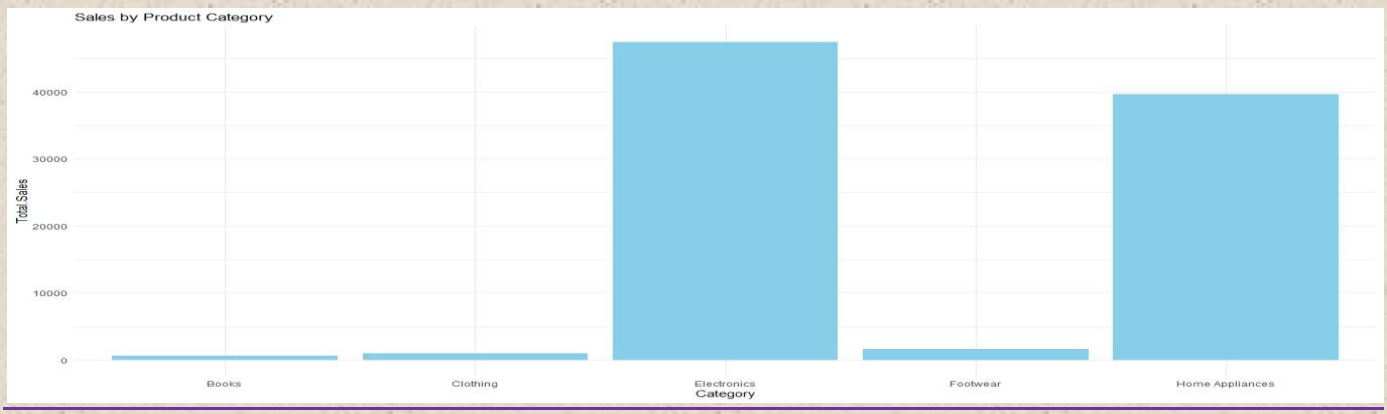
1. Heat map of Total Sales by Category and City:



Interpretation:

- Electronics and Home Appliances dominate in sales across most cities.
- Seattle, Houston, and Chicago stand out with high sales in Electronics and Home Appliances.
- Cities like Denver and San Francisco show lower sales or missing data for some categories.
- Footwear and Books perform relatively lower across most cities.
- This suggests that product popularity varies significantly by region, useful for targeted marketing and inventory planning.

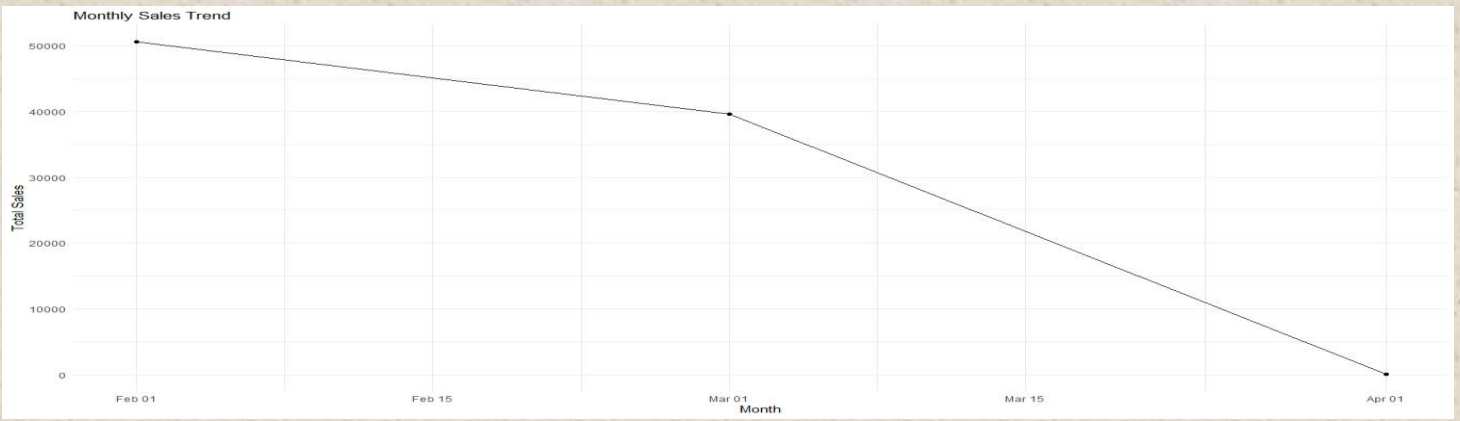
2. Bar Chart – Sales by Product Category:



Interpretation:

- Electronics is the top-performing category in total sales.
- Home Appliances is the second-best, contributing significantly to overall revenue.
- Categories like Books, Clothing, and Footwear generate minimal revenue in comparison.
- Indicates a need to either promote underperforming categories more or focus resources on top-selling items.

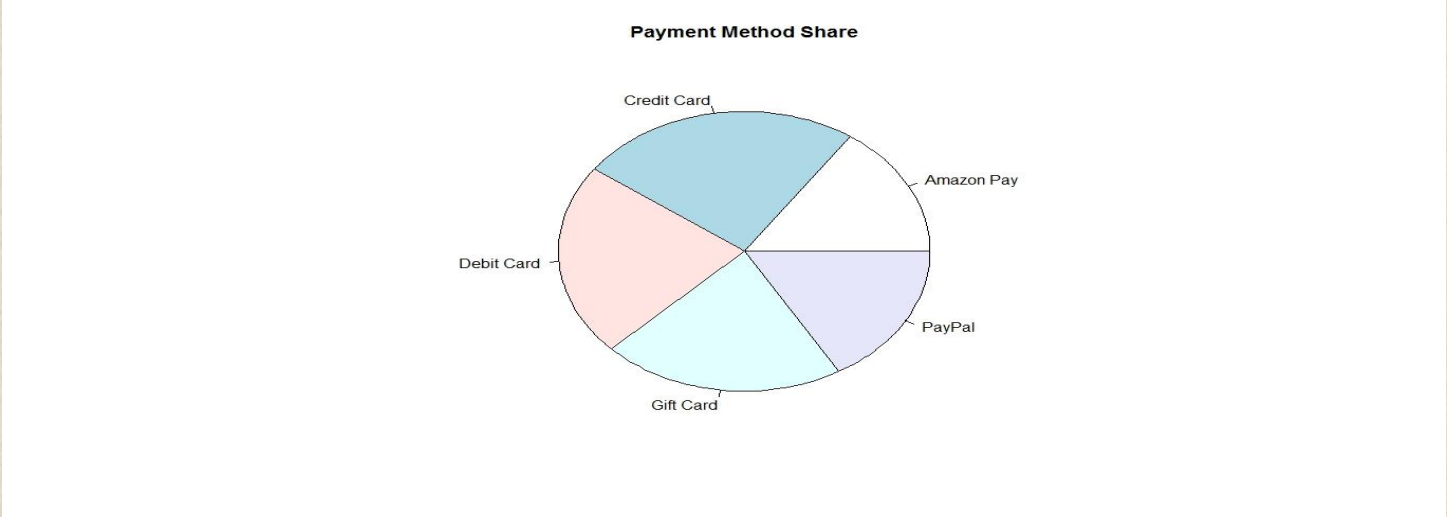
3. Line Chart – Monthly Sales Trend:



Interpretation:

- A clear downward trend is visible from February to April.
- Sales peaked in February, followed by a moderate decline in March, and a sharp fall in April.
- This may be due to seasonal demand variation, market factors, or inventory issues.
- Suggests the business should plan for promotions or strategic campaigns in low-performing months.

4. Pie Chart – Payment Method Share:



Interpretation:

- Credit Card and Debit Card are the most commonly used payment methods.
- PayPal, Amazon Pay, and Gift Cards account for smaller but still significant shares.
- The presence of multiple payment methods shows customer diversity in transaction preferences.
- Insights can help optimize checkout UX or offer targeted discounts (e.g., more on Gift Cards to boost their use).

STATISTICAL ANALYSIS USING R PROGRRAMMING

1.Descriptive Statistics

R Code:

```
summary(df)
```

Explanation:

This line provides a summary of the dataset (df), which includes basic statistics such as mean, median, min, max, and standard deviation for each column.

Helps in understanding the general distribution of variables (e.g., Total Sales, Price).

2. Clean Data:

R Code:

```
df <- df[df$Status == 'Pending', ]
```

Explanation:

Filters the data to include only the rows where the order status is "Pending".

Removes any invalid or cancelled orders, ensuring that only valid sales data is used in further analysis.

3.Monthly Sales Trend:

R Code:

```
df$Period <- paste(df$Month, df$Year)

monthly_sales <- df %>%

  group_by(Period) %>%

  summarise(Total_Sales = sum(Total.Sales)) %>%

  arrange(as.Date(paste0('01-', Period), format='%d-%B %Y'))

ggplot(monthly_sales, aes(x = as.Date(paste0('01-', Period), format='%d-%B %Y'), y = Total_Sales)) +

  geom_line() + geom_point() +

  labs(title = "Monthly Sales Trend", x = "Month", y = "Total Sales") +

  theme_minimal()
```

Explanation:

Creates a new Period column by concatenating Month and Year to track monthly sales.

Groups data by month and calculates the sum of total sales for each period.

Plots a line graph to visualize the monthly trend in sales over time.

This helps in identifying seasonal patterns, such as peak sales months.

4. Sales by Category:

R Code:

```
ggplot(df, aes(x = Category, y = Total.Sales)) +
  geom_bar(stat = "summary", fun = sum, fill = "skyblue") +
  labs(title = "Sales by Product Category") +
  theme_minimal()
```

Explanation:

Bar chart is created to visualize total sales by product category.

Summarizes the total sales for each category.

Helps identify which product categories generate the most sales.

5. Payment Method Pie Chart:

R Code:

```
pie(table(df$Payment.Method), main="Payment Method Share")
```

Explanation:

Pie chart visualizes the distribution of payment methods used by customers.

Provides insights into most popular payment options, such as credit card or PayPal.

6. ANOVA (Analysis of Variance):

R Code:

```
anova_result <- aov(Total.Sales ~ Category, data = df)

summary(anova_result)
```

Explanation:

Performs ANOVA to check if there are significant differences in mean total sales across different product categories.

Helps determine if product category is a significant factor affecting sales.

7. Linear Regression:

R Code:

```
model <- lm(Total.Sales ~ Price + Quantity, data = df)

summary(model)
```

Explanation:

Linear regression model predicts Total Sales based on Price and Quantity.

The summary(model) outputs the coefficients, R-squared, and p-values, which help assess the relationship between sales and these variables.

8. Time Series Forecasting (ARIMA):

R Code:

```
monthly_ts <- df %>%

  group_by(Date = floor_date(Date, "month")) %>%

  summarise(Total = sum(Total.Sales))

ts_data <- ts(monthly_ts$Total, frequency = 12)

fit <- auto.arima(ts_data)

summary(fit)
```

Explanation:

Aggregates data to create a monthly time series of Total Sales.

Converts the data into a time series object using ts() function.

Fits an ARIMA model to the time series data to forecast future sales.

auto.arima() automatically selects the best ARIMA model.

Forecast results can be used for predicting future sales.

9. Heatmap of Sales by Category and City:

R Code:

```
heatmap_data <- df %>%
  group_by(Category, City) %>%
  summarise(Total_Sales = sum(Total.Sales, na.rm = TRUE)) %>%
  ungroup()
ggplot(heatmap_data, aes(x = City, y = Category, fill = Total_Sales)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  labs(title = "Heatmap of Total Sales by Category and City",
       x = "City", y = "Category", fill = "Sales") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Explanation:

Aggregates total sales by Category and City.

Visualizes a heatmap where each tile's color intensity represents the amount of total sales.

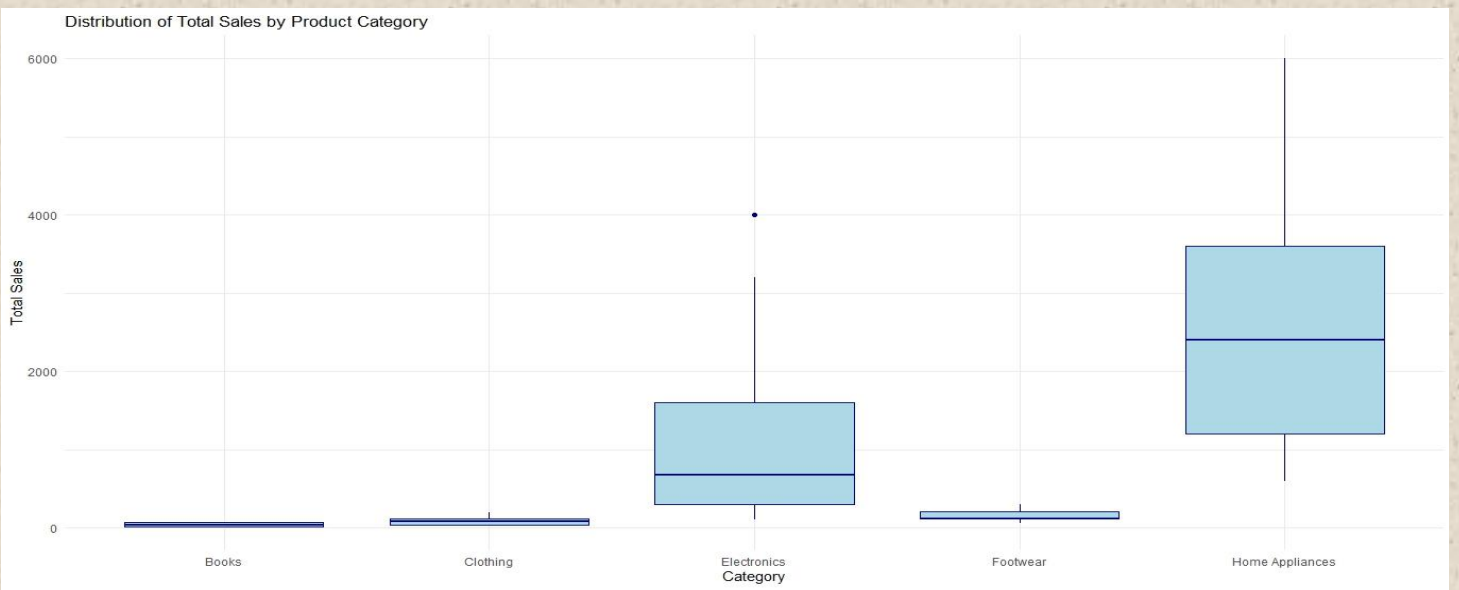
Darker colors represent higher sales, helping to identify high-sales regions and categories.

Visualizing Model Results

1. Visualize ANOVA – Boxplot by Category:

R Code:

```
ggplot(df, aes(x = Category, y = Total.Sales)) +
  geom_boxplot(fill = "lightblue", color = "darkblue") +
  labs(title = "Distribution of Total Sales by Product Category",
       x = "Category", y = "Total Sales") +
  theme_minimal()
```



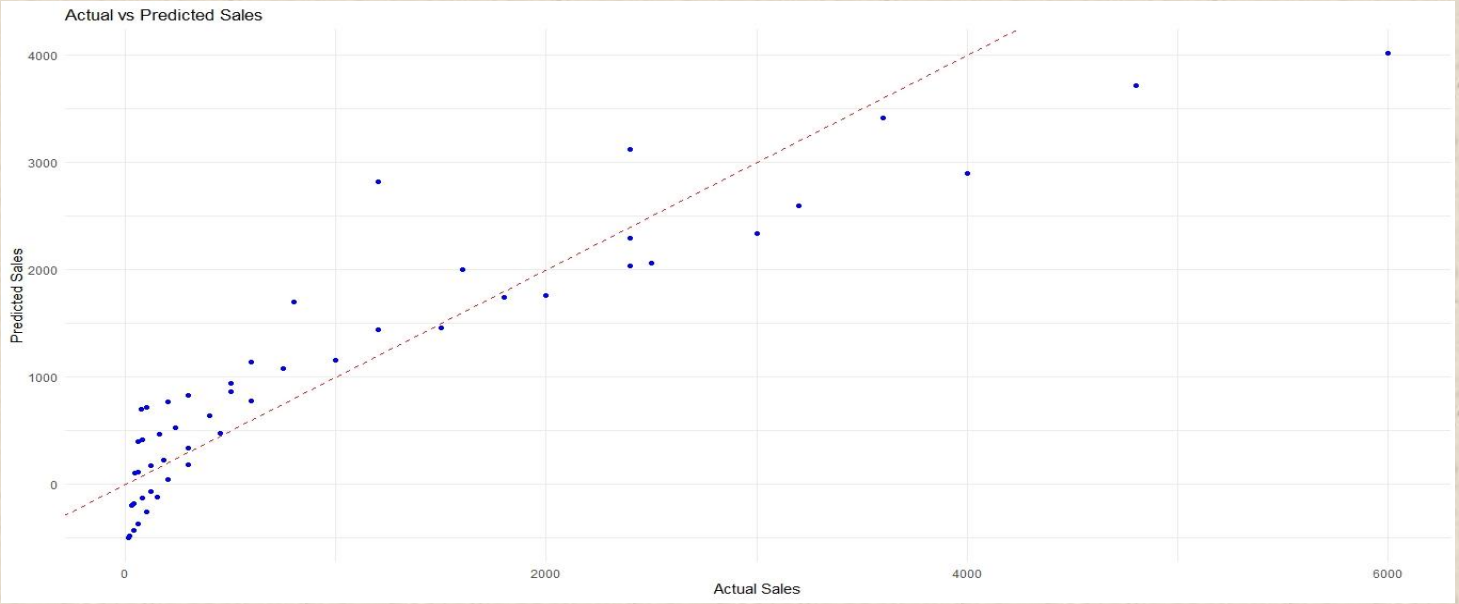
Explanation:

- Home Appliances generate the highest and most variable sales, likely contributing significantly to revenue.
- Books, Clothing, and Footwear have low and stable sales, suggesting less variability or lower demand.
- Electronics show outliers, indicating some high-performing sales instances but also more inconsistency.

2. VISUALIZING LINEAR REGRESSION:

R Code:

```
df$Predicted_Sales <- predict(model0)

ggplot(df, aes(x = Total.Sales, y = Predicted_Sales)) +

  geom_point(color = "blue") +

  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +

  labs(title = "Actual vs Predicted Sales",

       x = "Actual Sales", y = "Predicted Sales") +

  theme_minimal()
```



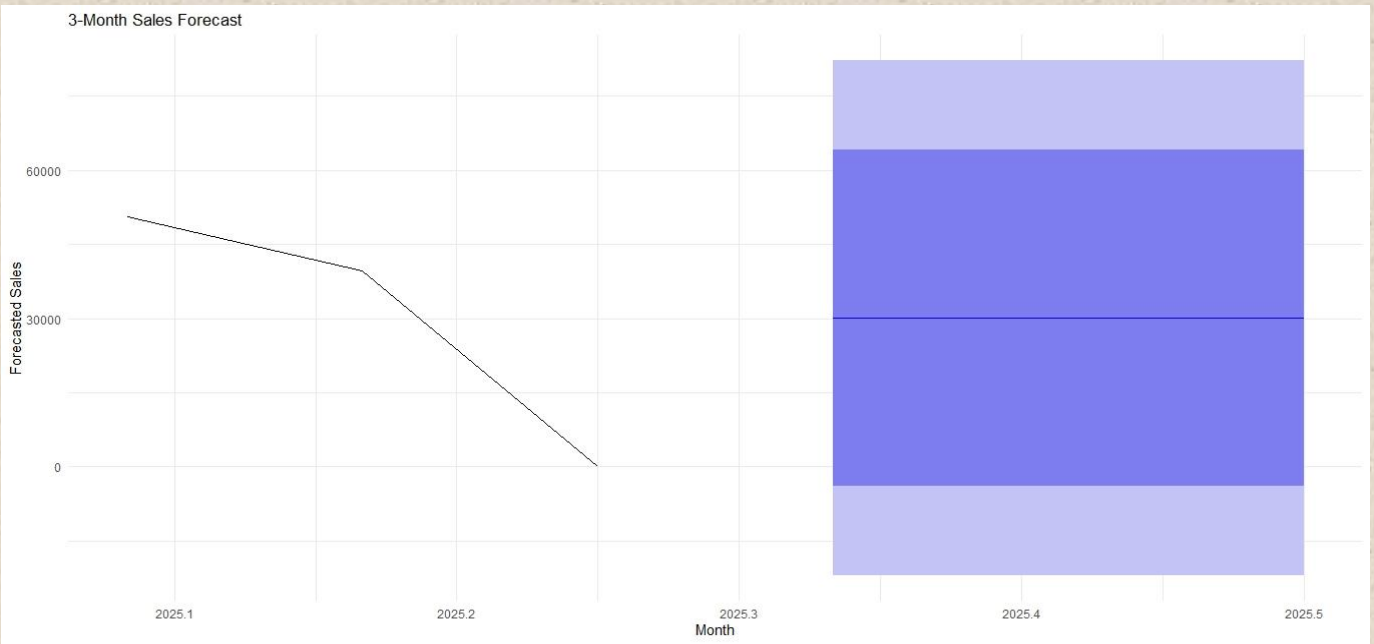
Explanation:

- Your regression model performs well — predictions are generally in line with actual values.
- A few points far from the line indicate possible outliers or segments the model under/overestimates.
- Overall, it demonstrates that price and quantity are effective predictors of total sales.

3. VISUALIZING ARIMA ANALYSIS:

R Code:

```
library(dplyr)
library(lubridate)
# Convert Month and Year to a proper Date object (assuming day = 1)
df <- df %>%
  mutate(Full_Date = as.Date(paste0("01-", Month, "-", Year), format = "%d-%B-%Y"))
monthly_ts <- df %>%
  filter(Status == "Pending") %>%  # Only include valid/pending sales
  group_by(Month = floor_date(Full_Date, "month")) %>%
  summarise(Total_Sales = sum(Total.Sales, na.rm = TRUE)) %>%
  ungroup()
library(forecast)
library(ggplot2)
ts_data <- ts(monthly_ts$Total_Sales,
              start = c(year(min(monthly_ts$Month)), month(min(monthly_ts$Month))),
              frequency = 12)
autoplot(ts_data) +
  labs(title = "Monthly Total Sales Time Series",
       x = "Month", y = "Total Sales") +
  theme_minimal()
fit <- auto.arima(ts_data)
forecasted <- forecast(fit, h = 3)
summary(forecasted)
autoplot(forecasted) +
  labs(title = "3-Month Sales Forecast", x = "Month", y = "Forecasted Sales") +
_theme_minimal()
```



Explanation:

- The model predicts sales will stabilize around a mean, but we cannot rely on exact values due to wide intervals.
- More historical data points are needed for a more accurate and stable forecast.
- The confidence band touching negative sales suggests the need to refine the model or clean data further.

FINDINGS AND INTERPRETATIONS FROM THE ANALYSIS

1. ANOVA Output (Total Sales ~ Category)

Interpretation:

The p-value = 2.91e-15 is much less than 0.05, meaning the difference in average sales between categories is statistically significant.

F-value = 30.05 is high, indicating a strong variation among group means.

2. Linear Regression Output (Total Sales ~ Price + Quantity)

Interpretation:

Both Price and Quantity are significant predictors of Total. Sales (p-values < 0.001).

For every 1 unit increase in:

Price, sales increase by approx. 3.37 units.

Quantity, sales increase by approx. 257.64 units.

R-squared = 0.8896 indicates that ~89% of the variance in sales is explained by this model.

3. ARIMA Time Series Output

Interpretation:

The ARIMA model is (0,0,0), meaning a constant average model with no autoregression, differencing, or moving average terms.

Mean forecast = 30095, indicating the model predicts a flat future sales trend around this value.

RMSE = 21674, MAE = 19996 show relatively high error, suggesting limited predictive power.

CONCLUSION:

This project provides a comprehensive analysis of Amazon's sales data, focusing on trends, category-wise performance, and predictive modeling. Key insights derived from the study are summarized below:

Sales Trend Over Time:

The monthly sales trend reveals a gradual decline across the observed months. This could indicate seasonal demand variation or external influencing factors. However, more data points are needed to establish a reliable long-term trend.

Category Performance Analysis:

- Home Appliances and Electronics emerged as the top-performing categories in terms of total sales.
- Books, Clothing, and Footwear showed comparatively lower and more stable sales figures.
- Sales in Electronics also displayed extreme values (outliers), suggesting occasional high-volume transactions.

Boxplot Insights:

Home Appliances exhibit the highest median and variability in total sales, indicating strong but inconsistent demand. Electronics followed closely, with notable spikes in performance.

Customer Payment Preferences:

Debit and credit cards were the most commonly used payment methods, followed by Amazon Pay, PayPal, and gift cards, suggesting a preference for conventional payment systems among customers.

City-Wise Sales Distribution:

The heatmap indicates that cities like New York, Houston, and Seattle have the highest concentration of sales across most product categories.

Statistical Testing – ANOVA:

Analysis of Variance confirmed that differences in average sales across product categories are statistically significant ($p < 0.001$), highlighting that category type substantially influences sales performance.

Regression Modeling:

A linear regression model using Price and Quantity as predictors achieved a high adjusted R^2 of 0.887, indicating strong predictive power. Actual vs. predicted plots showed that most data points lie close to the ideal prediction line.

Time Series Forecasting:

The ARIMA model projected average sales of ~30,000 units over the next three months. However, wide prediction intervals indicate limited forecast reliability, likely due to a small dataset.

This project successfully applied exploratory data analysis, statistical inference, and predictive modeling to uncover business insights. The findings can be used to optimize product focus, inventory decisions, and marketing strategies, especially for top-performing categories and locations.