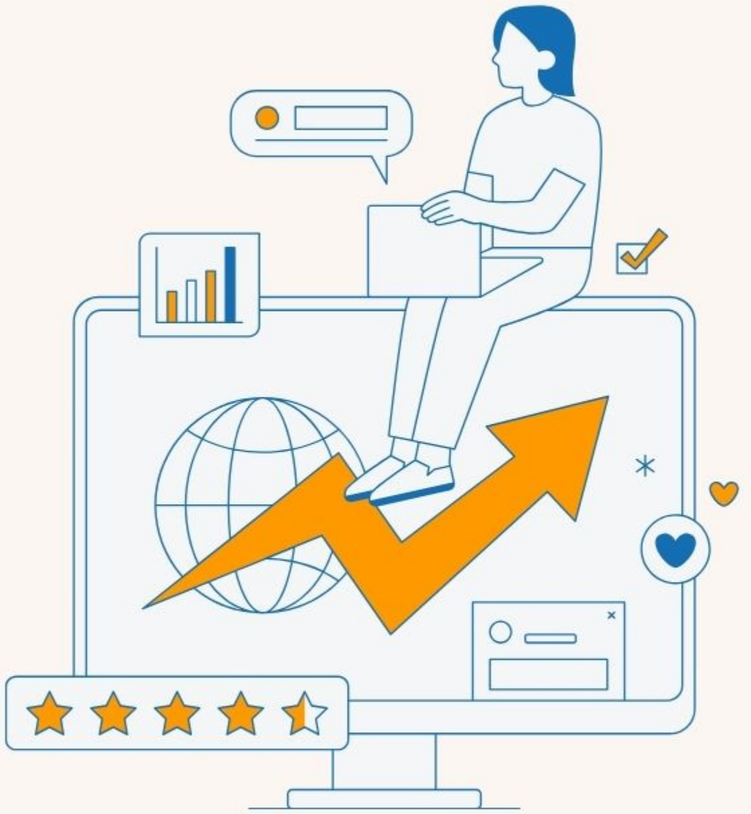
IE 6200: Engineering Probability and Statistics

Prof. Rehab Ali

Statistical Analysis of Amazon Prime day Sales

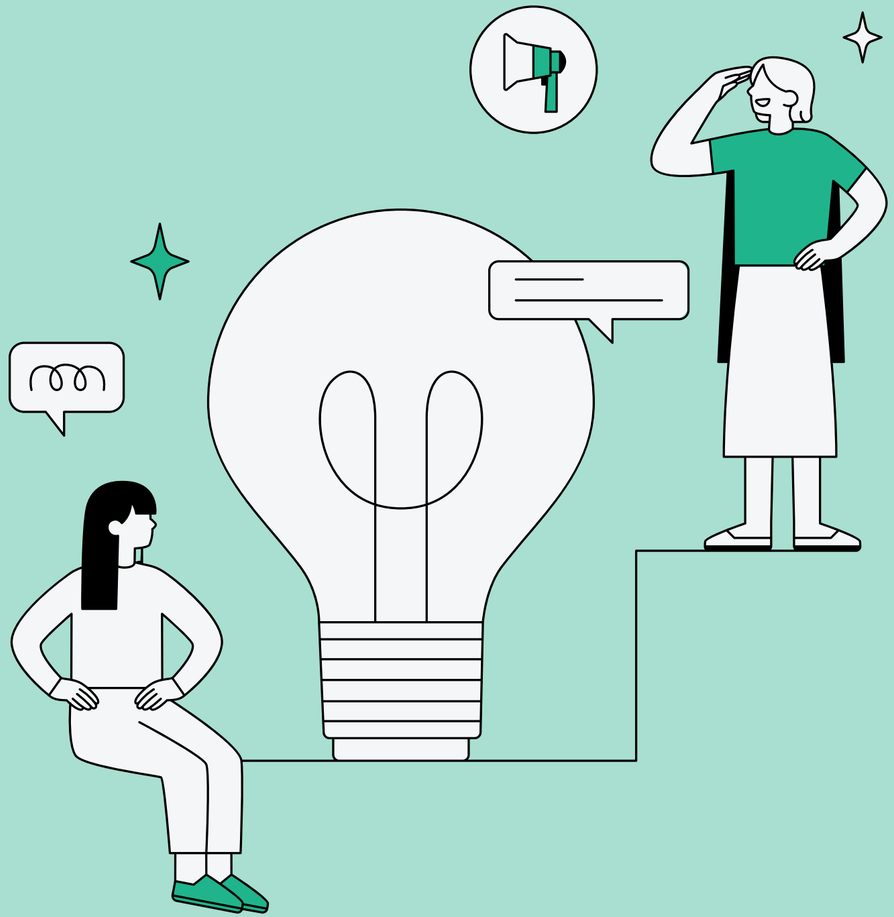
Group 7

Vishwanath Basavaraj Badiger
Vaishnavi Poluru
Shweta Shinde
Prateek Narayan Shetty
Harshitha Chandrashekar
Aryan C



PROBLEM STATEMENT

This annual online shopping extravaganza provides a unique lens to dissect e-commerce dynamics. By scrutinizing product performance across categories, our goal is to extract actionable insights for optimizing sales strategies and enhancing customer satisfaction. Our class project centers on a rigorous statistical analysis of Amazon Prime Day 2023 sales data using R. This project offers a hands-on exploration of real-world e-commerce data, bridging theory and practice in the context of one of the largest online retail events.



Scope Of The Project

01.

Included in the Project:

- Analyze sales data with descriptive statistics.
- Assess impact of discounts on sales and revenue.
- Identify key factors influencing sales.
- Compare product category performance.
- Use ANOVA to analyze data statistically.

02.

Excluded from the Project:

- No deep dive into customer reviews (qualitative analysis).
- No recommendations for optimizing future Prime Days.

Project Objectives

- Quantify sales trends: Identify top-performing product categories for resource allocation.
- Uncover customer behavior: Analyze how ratings, discount percentages, and sales correlations impact customer decisions.
- Visualize key findings: Create clear and impactful data representations to communicate insights.

Expected benefits:

- Strengthen statistical skills: Apply probability and statistics concepts to real-world data.
- Develop analytical thinking: Solve a practical business problem using statistical methods.

DATA COLLECTION

- Data Sources
  - Primary Source: Kaggle
  - Datasets: Amazon Prime Day sales datasets

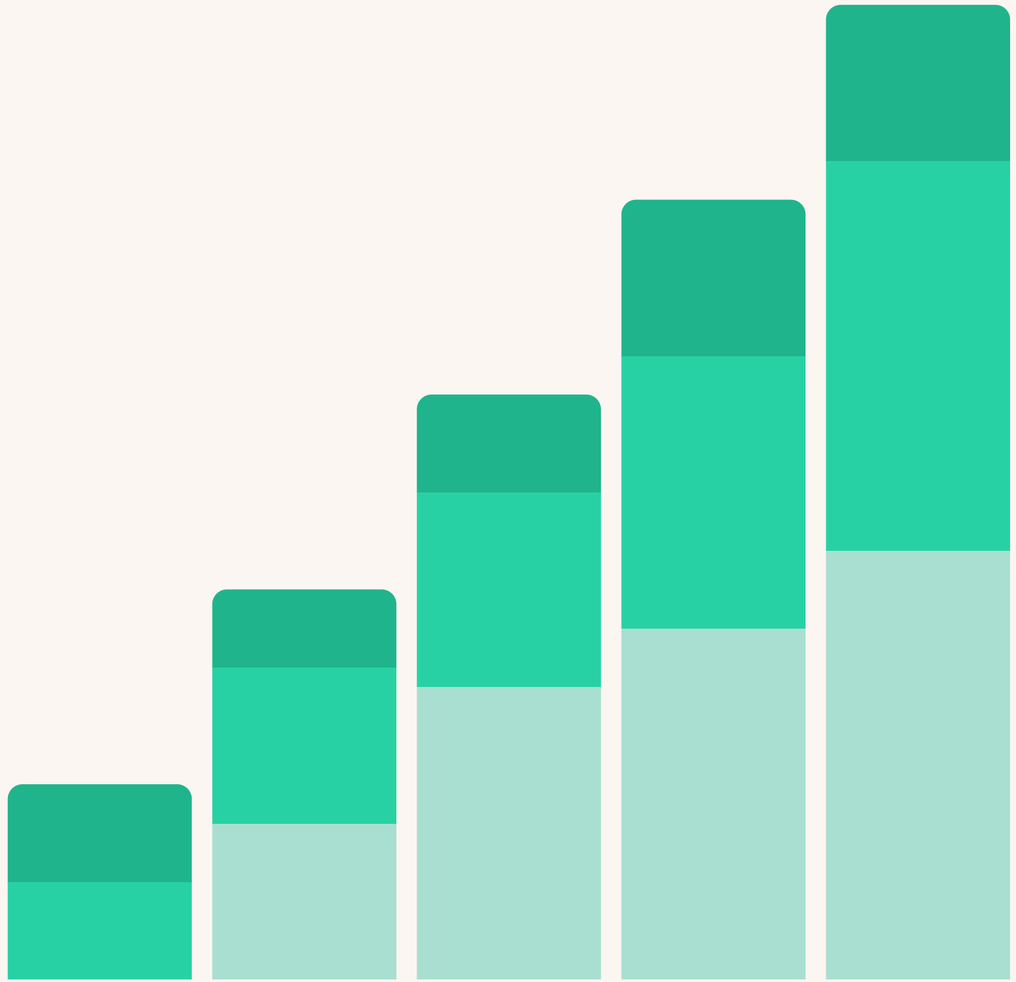
- Variable Explanation
  - Variables: 13 : 2 Categorical and 11 Numerical
    - Categorical: Product ID, Category
    - Numerical: Discounted Prices on a regular day, Discounted Prices on a prime day, MRP, Discount Percentage on a prime day, Discount percentage on a regular day, Ratings, Rating Count, Units Sold on a prime day, Units sold on a regular day, Revenue on a prime day, Revenue on a regular day.

		AMAZON PRIME DAY SALES 2023										
Product ID	Category	Discounted_price on prime day	Discounted_ price on regular day	MRP	Discount_percentage on prime day	Discount percentage on a regular day	Rating	Rating_ Count	No of Units Sold on Prime Day	No of units sold on a regular day	Revenue on a Prime day	Revenue on a regular day

DATA PREPARATION

Model Assumptions

- **Accuracy: Assumes Kaggle datasets are accurate.**
- **Ethics: Adheres to rigorous ethical standards and data privacy regulations.**
- **Independence: Sales data of product categories is independent.**
- **Time Frame: Data represents Amazon Prime Day Sales 2023.**
- **Independence: Sales data across product categories is not dependent.**
- **Data Quality: Assumes accuracy and quality of sales data from Kaggle.**
- **Normality: Acknowledges normal distribution of data points.**

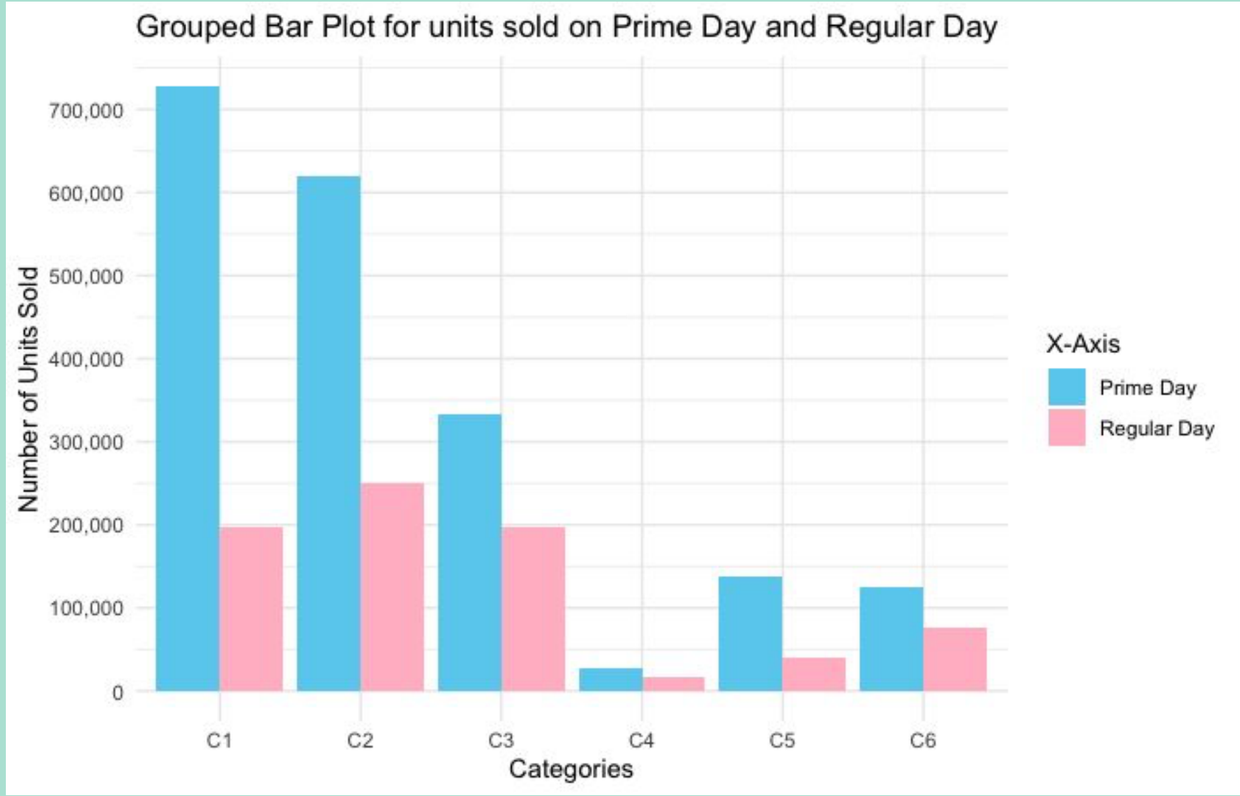


DATA VISUALIZATION

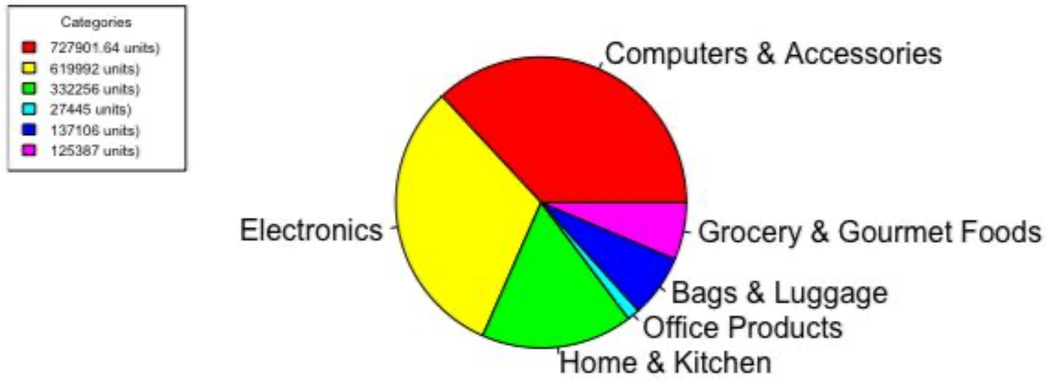
Grouped Bar Plot: Units Sold on Prime Day vs. Regular Day

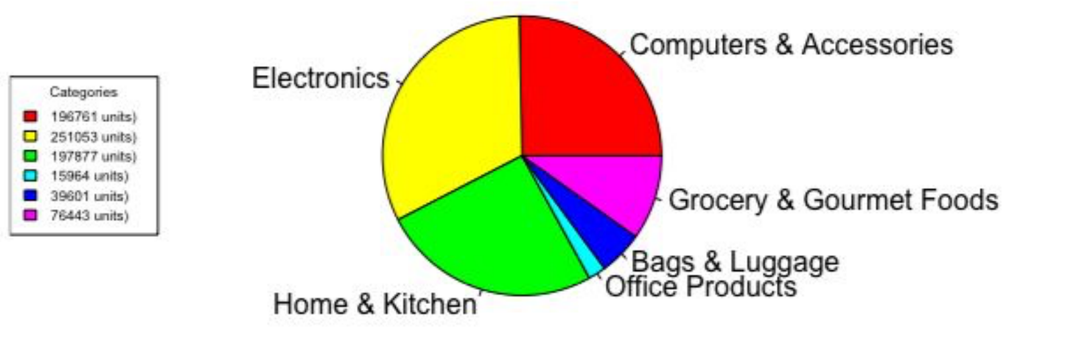
Categories:
C1 - Computer and Accessories,
C2 – Electronics,
C3 – Home & Kitchen,
C4 – Office Products,
C5 – Bags & Luggage,
C6 – Grocery & Gourmet foods.

Computer & Accessories dominate Prime Day sales, while Office Products show the least sales figure.



Pie Charts


Category Distribution For Units Sold on Prime Day
Categories
727901.64 units)
619992 units)
332256 units)
27445 units)
137106 units)
125387 units)
Computers & Accessories
Electronics
Grocery & Gourmet Foods
Bags & Luggage
Office Products
Home & Kitchen

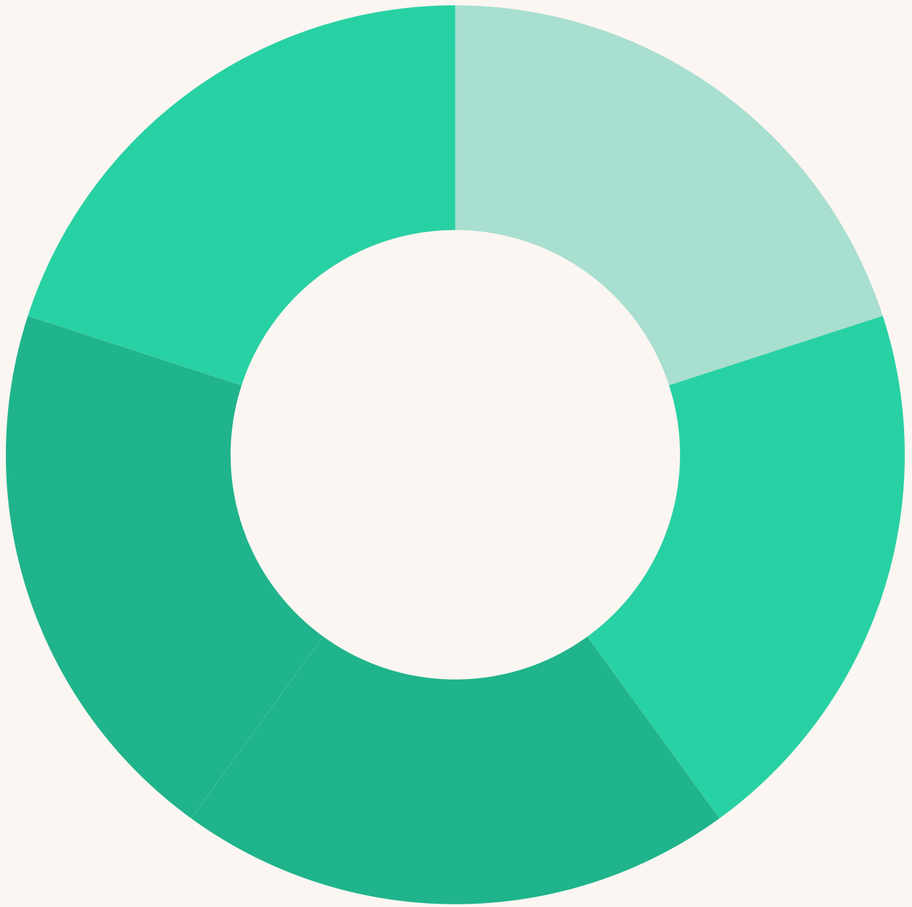

Category Distribution For Units Sold on Regular Day
Categories
196761 units)
251053 units)
197877 units)
15964 units)
39601 units)
76443 units)
Computers & Accessories
Electronics
Grocery & Gourmet Foods
Bags & Luggage
Office Products
Home & Kitchen

Scatter Plots

STATISTICAL ANALYSIS

Confidence Interval of Difference in Means (Revenue)

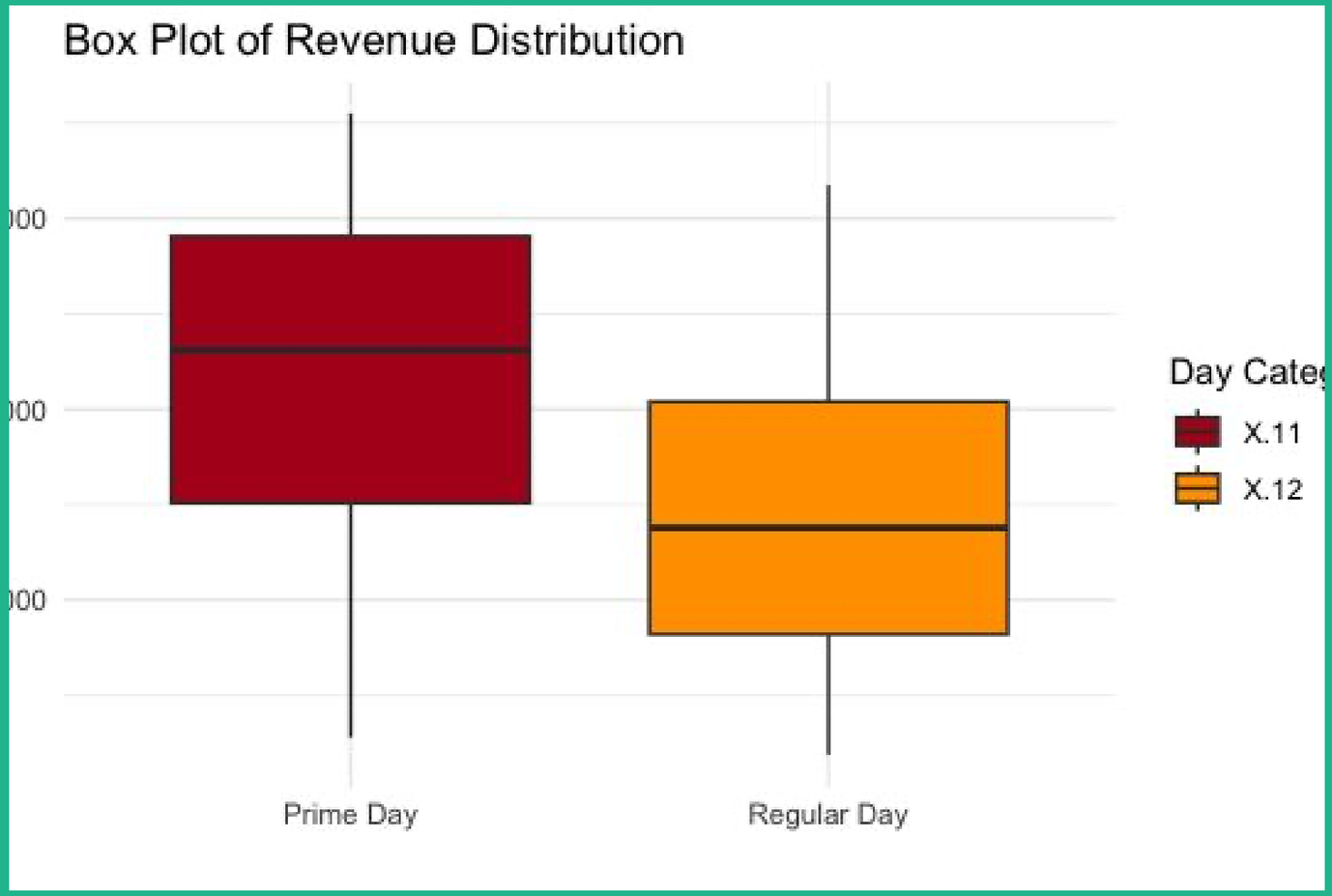
Categories	Sample size (n)	CI of difference in means
Computer and accessories	45	(7421.843, 26369.128)
Electronics	50	(-189187, 738391)
Home and kitchen	25	(-940017.74, -54417.22)
Office products	27	(-14035.71, 185751.64)
Bags and Luggage	33	(-32621.09, 482116.18)
Grocery and Gourmet foods	25	(-61158.39, 143048.95)



Paired Sample T Test

- **Null hypothesis: The revenue on a Prime day is equal to that of on a regular day. Category under analysis: Computer & Accessories. (Ho: PrimeDay = RegularDay)**

- **Alternative hypothesis: The revenue on a Prime day is higher than that of on a regular day for the same category. (Ha: PrimeDay > RegularDay)**



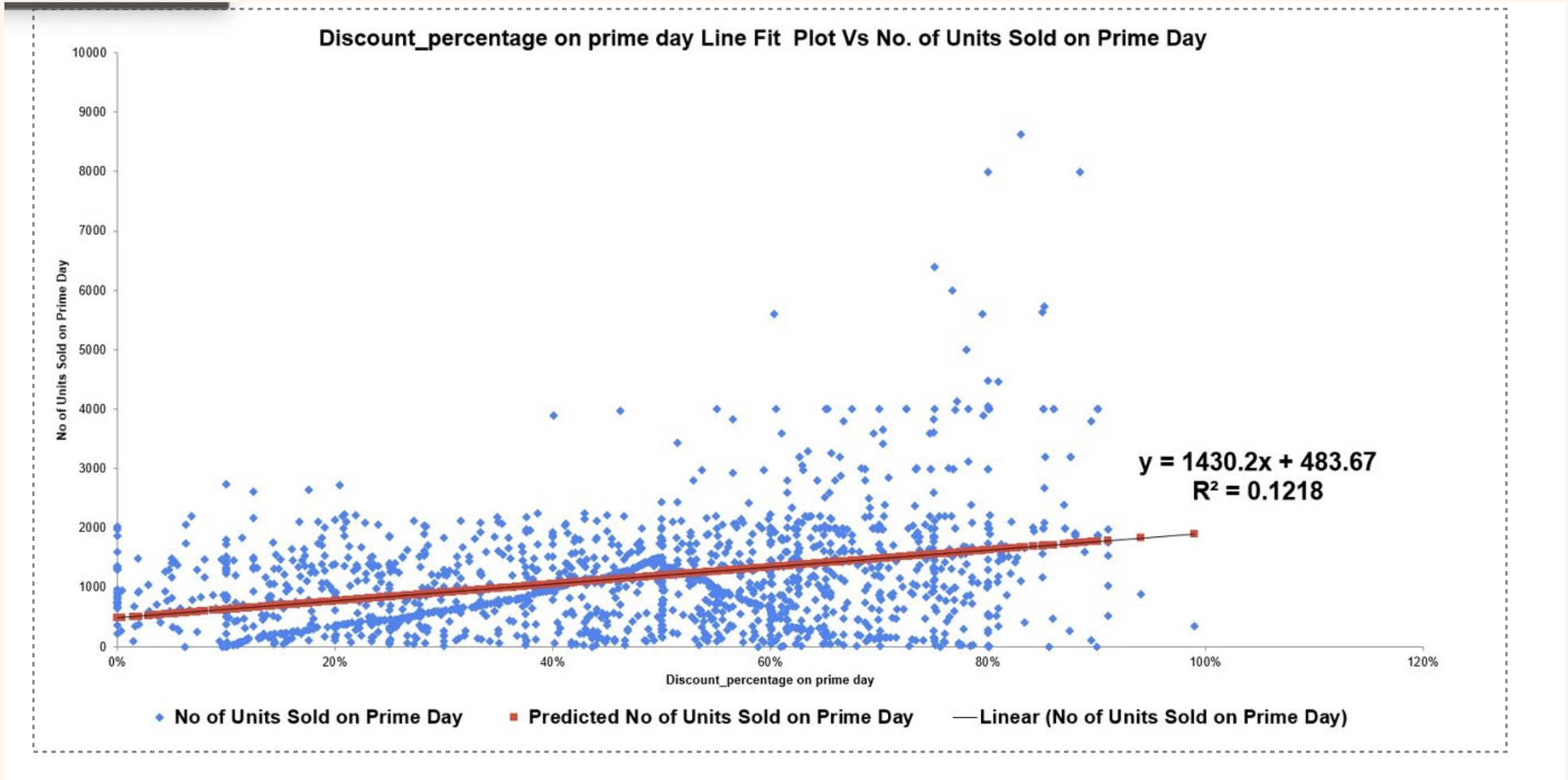
- The analysis compares Prime Day and regular day revenue for "Computer and Accessories" with a sample size of 45.
- A t-value of 8.7964, degrees of freedom 43, and p-value 1.821e-11 (<0.05) strongly reject the null hypothesis.
- The one-sided test suggests a positive mean difference, supported by a 95% confidence interval (13666.6, +∞).
- The mean difference of 16895.49 is significantly greater than zero, providing strong evidence for rejecting the null hypothesis.

LINEAR REGRESSION ANALYSIS

- **Null Hypothesis(Ho): There is no significant relationship between the discount percentage discount and the number of units sold on Prime Day.**
- **Alternative hypothesis(Ha): The number of units sold increases with an increase in the discount percentage on Prime Day.**

- The scatter plot illustrates the relationship between percentage discount and units sold.
- Mean discount: 46.268 (SD = 0.220); mean units sold: 1145.399 (SD = 903.158).
- The regression model (intercept = 1430.195, slope = 483.669) suggests a weak association (R-squared = 0.1218).
- With a p-value < 0.05, we don't reject the null hypothesis, supporting that percentage discount doesn't significantly impact units sold.

SCATTER PLOT FOR REGRESSION ANALYSIS

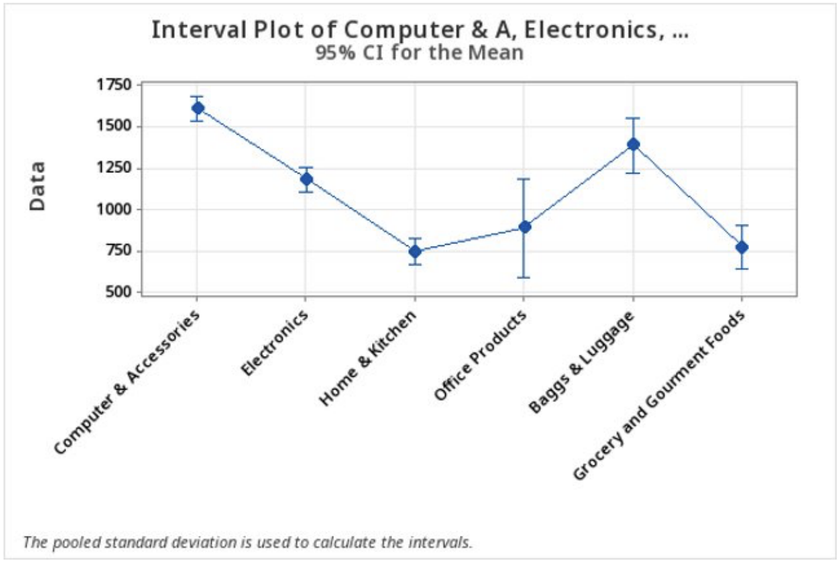


ANOVA TESTING

- **Null Hypothesis (Ho): The mean values of the number of units sold are the same across all product categories.**
- **Alternative Hypothesis (Ha): At least one of the product categories has a different mean value.**

- Based on the ANOVA test results for No of units Sold on Prime Day , we accept the alternative hypothesis and reject the null hypothesis because the p-value is very small ($p<0.05$) and the F-value is high.
- A high F-value, coupled with a low p-value indicates that there is more variation between group means than there is within groups. It is reasonable to assume that there are statistically significant variations in the average of No of Units Sold on Prime day between a minimum of two distinct categories.

ANOVA Test for No of Units Sold on Prime Day



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Factor	5	200913820	40182764	57.33	0.000
Error	1714	1201265317	700855		
Total	1719	1402179137			

Means

Factor	N	Mean	StDev	95% CI
Computer & Accessories	453	1606.8	1254.8	(1529.7, 1684.0)
Electronics	526	1178.7	652.5	(1107.1, 1250.3)
Home & Kitchen	448	741.6	431.3	(664.1, 819.2)
Office Products	31	885.3	435.1	(590.4, 1180.2)
Baggs & Luggage	99	1385	1219	(1220, 1550)
Grocery and Gourment Foods	163	769.2	440.8	(640.6, 897.9)

Pooled StDev = 837.171

Result and Conclusion

- Prime Day Revenue Boost: The confidence interval for the difference in revenue between Prime Day and regular days for computers and accessories (7421.843, 26369.128) excludes zero, providing strong evidence that Prime Day significantly boosts revenue in this category
- Discount Impact on Units Sold: The scatter plot and regression analysis suggest a weak relationship between discount percentage and units sold (R-squared = 0.1218). Further, the p-value for the slope coefficient indicates no significant increase in units sold with a higher discount.
- Category-wise Sales Variation: ANOVA tests for both units sold and discounted prices on Prime Day reveal significant differences across categories (p-value < 0.05). This suggests at least two categories differ in average performance.

Limitations

- We could not find partially/ readily available variables information for a smoother data collection and preparation process.
- Figuring out the tests that would be compatible with our data was a hassle but also a good learning process.
- Understanding and applying the R code to visualize and hypothesize the data sets was a long and tedious process

PROPOSED NEXT STEPS AND FUTURE WORK

- Can extend the project to delve deeper into the trends over the past few years with respect to the Prime Day and the impact of various factors on the revenue and the sales.
- Can look further into the trends revolving around the purchase preferences of customers from various demographics.
- Develop more sophisticated predictive models to forecast Prime Day sales and the preferred categories of customers.

Thank you very much!