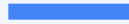




Consumer Spending Habits During Holiday Seasons



Natalie Raver-Goldsby, Drew Johnson, Ellen Grant

Agenda



- Overview
 - Data Cleaning and Preparation
 - Project Scope
 - Exploratory Analysis
- Key Findings and Results
 - Models and Optimization
- Conclusions and Insights

Objective:

Analyze consumer spending patterns to uncover relationships between holidays, seasonal trends, and purchasing behavior.

Goals:

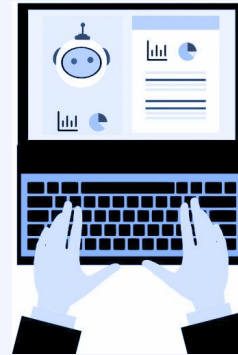
- Uncover holiday and seasonal trends in consumer spending.
- Identify demographic factors (e.g., gender, location) that influence spending.
- Measure the effectiveness of promotional strategies, such as discounts and coupons.



Overview

Data Kaggle Online Shopping Dataset

- Original format 52,955 rows, 23 features over 12/31/2018 - 12/30/2019)
- Features of interest: Product Categories, Total Cost, Total Sales, Quantity, Delivery Charges, Coupon Usage,



Scope of Analysis

- Investigating holiday-related spending, demographic influences, and promotional effectiveness
 - Demographic and promotional breakdowns to analyze patterns across different customer segments.
- Data cleaning included identifying duplicates and NA values, transformation and standardization, calculating additional features and normalization.
- Analysis included spending and discount breakdowns by demographic, traditional time series, a time series regression, a random forest decision tree, and a neural network.

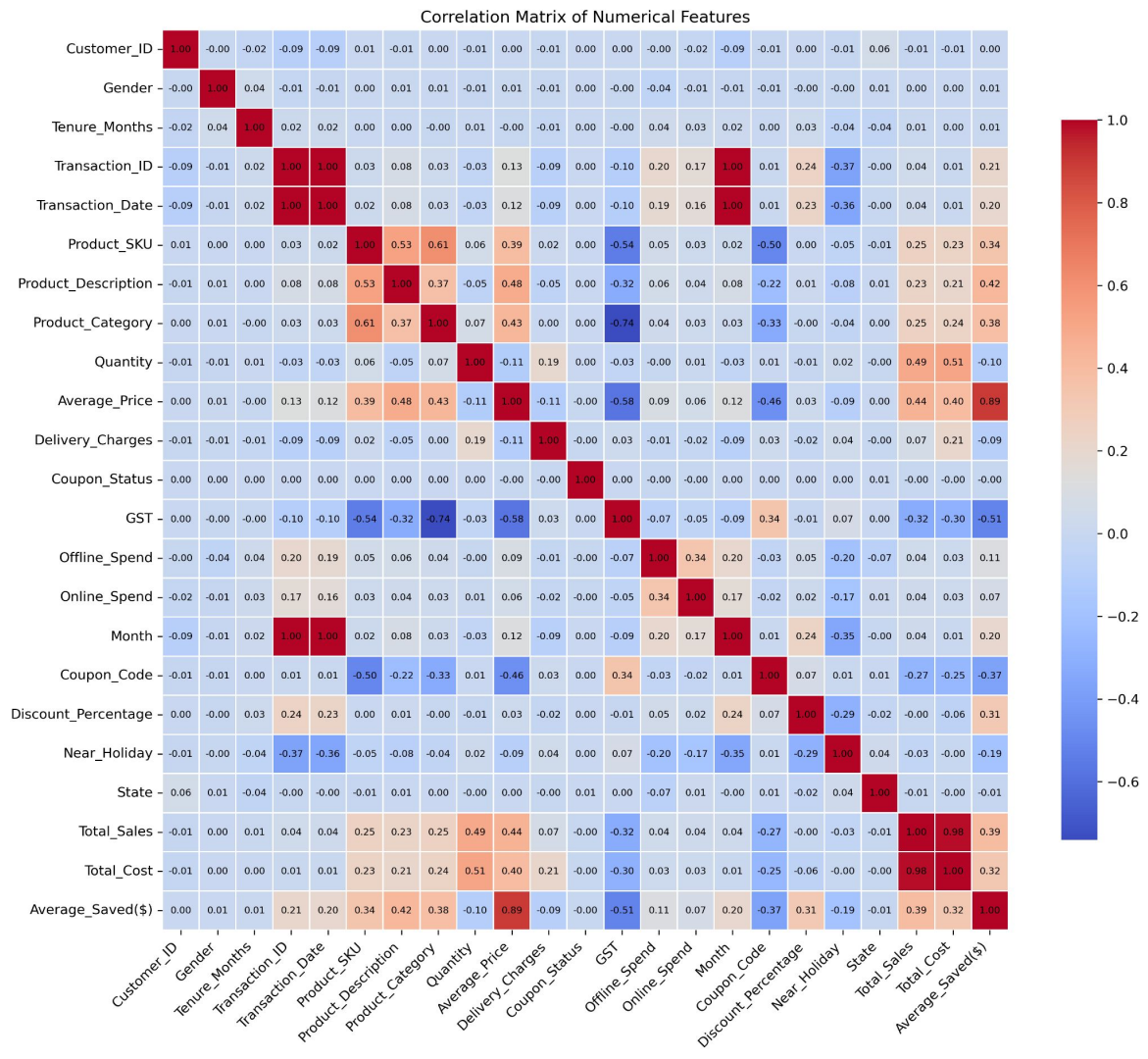


- Key Positive Correlations
- Key Negative Correlations
- Insights

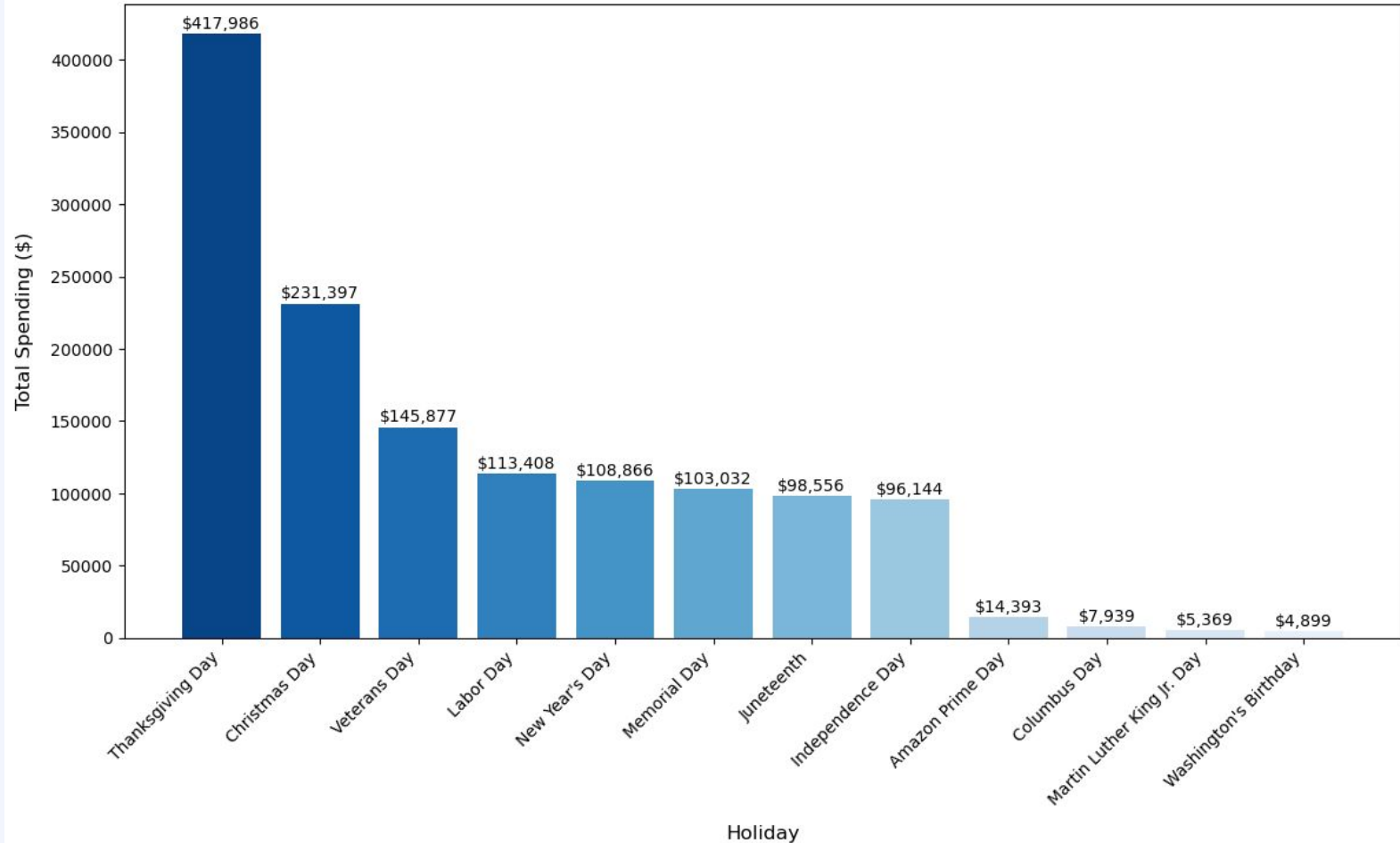
- Offline vs. Online Spend: Independent behavior
- Discount Strategy: Bigger savings on high-cost items

- Actions

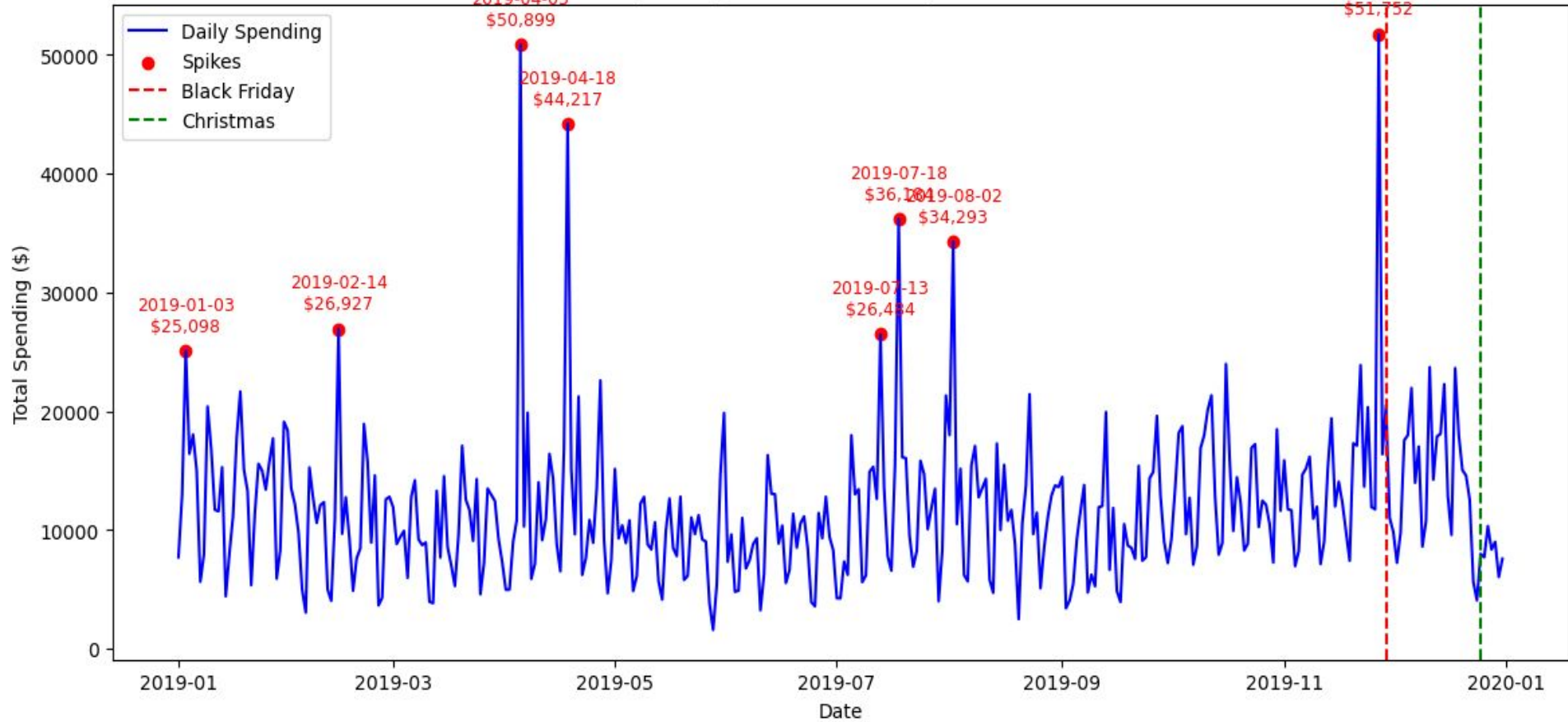
- Promotions
- Holiday Planning
- Discounting



Total Spending by Holiday



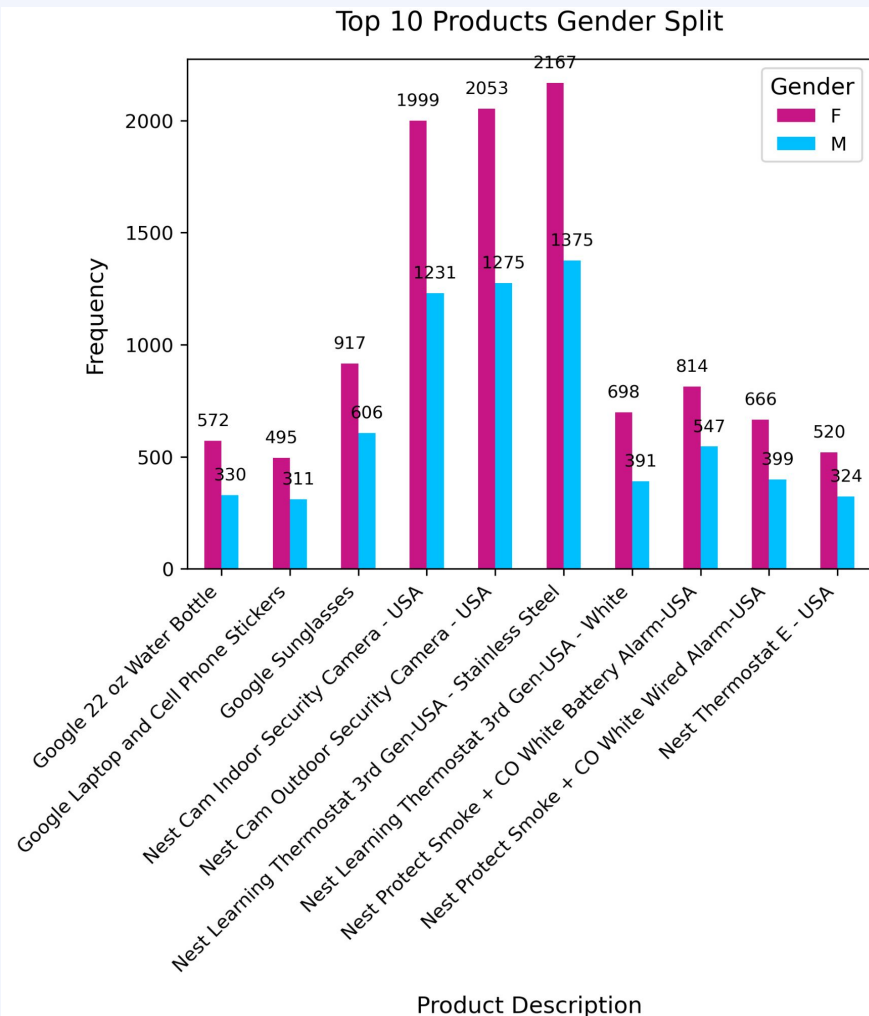
Total Spending Over Time with Spike Annotations



Spending Breakdown

- Women and men on average spend the same amount despite the gender skew
- Applied discounts does not explain the difference and this average spend is reflected in average savings.
- Of the ten most popular products, women are outpurchasing men as well →

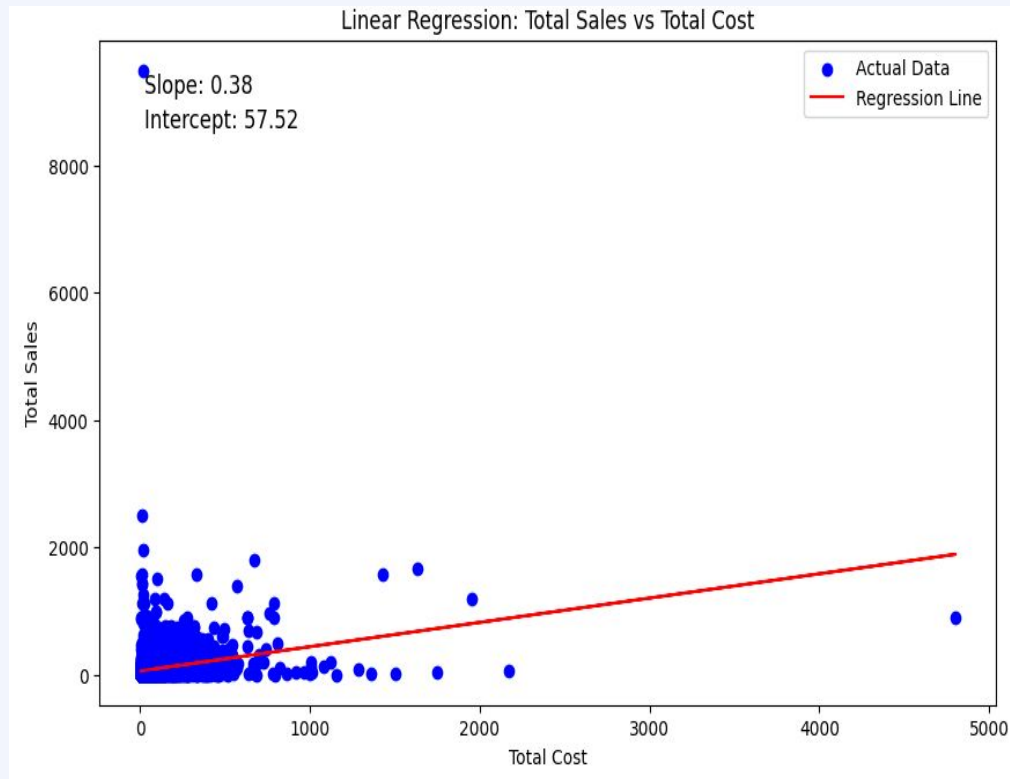
	F	M
Sample	33,038	19,917
Spend(Sum)	2,681,119	1,621,060
Spend(Mean)	81.15	81.39
Applied Discounts	19.98%	19.90%
Savings(Mean)	10.49	10.69



Key Findings and Results

Regression

- The model predicts Total Sales based on Total Cost using a Linear Regression approach
 - By using Total Cost to predict Total Sales, a business can understand if and how their spending (on products, marketing, etc.) influences sales.
- **R-squared (R^2): 0.119**
 - Interpretation: The model explains only 11.9% of the variance in Total Sales, which suggests that the linear relationship between Total Cost and Total Sales is weak.
- **Total_Sales = $0.382 \times \text{Total_Cost} + 57.52$**
 - Interpretation: For each 1 unit increase in Total Cost, Total Sales increase by 0.382 units. This means that higher costs are associated with higher sales, but the relationship is relatively weak (a small increase in sales for every increase in cost).



Random Forest

- Our model performed best with a max depth of 100 branches and 600 trees
- R^2 remained high across each iteration showing that random forest captured 98% of deviation in the data.
- MSE of 297 is still high considering Total_Cost std is 138, however the data ranges from 3.7 to 8,552

Parameters		Results		
Max Depth	Trees	MSE	MAE	R^2
None	100	364.57	0.95	0.98
10 ★	100	326.99	1.91	0.98
60 ★	200	336.49	0.94	0.98
100 ★	600	297.05	0.92	0.98
200 ★	1500	327.18	0.94	0.98
None ★	3000	314.83	0.94	0.98

Random Forest with Outlier Control

```
# Outlier controls
def remove_outliers_iqr(df, column):
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    df_out = df[(df[column] >= lower_bound) & (df[column] <= upper_bound)]
    return df_out







# Apply outlier removal to y_train
y_train = pd.DataFrame(y_train, columns=['Total_Cost'])
y_train = remove_outliers_iqr(y_train, 'Total_Cost')
y_train = y_train['Total_Cost']

# Now filter X_train based on the final indices of y_train
X_train = X_train.loc[y_train.index]
```

- This is an ineffective model, only explaining 45% of the variance in the data
- Outliers do not account for the large error in the previous model

Parameters		Results		
Max Depth	Trees	MSE	MAE	R ²
None	100	10,803.24	9.94	0.45
None	250	10,803.05	9.94	0.45
None	800	10,802.79	9.94	0.45
10	100	10,803.78	10.24	0.45
100	250	10,803.05	9.94	0.45
500	800	10,802.79	9.94	0.45

Neural Network

Parameters				Results	
Activation Func.	N Layers	Nodes	Epochs	MSE	MAE
relu / linear	3	6 6 z	50	32.17	1.60
relu / linear	3	8 6 z 	50	33.47	2.66
relu/linear	3	6 6 z	100 	71.04	2.08
relu/ linear	4 	6 6 6 z	50	82.83	4.27
relu/ linear	4	8 6 6 z 	50	27.36	1.98
relu/ linear	4	8 8 8 z 	50	58.26	2.75
relu/ linear	4	8 6 6 z	100 	63.91	5.16

- The model performed best with 4 layers with varying nodes numbers and 50 epochs.
- Otherwise the model performed best with only three layers in the first two iterations.

Conclusions and Insights

- We evaluated linear regression, random forest, and neural network models.
 - We did a grid search for optimized parameters; this could have been performed more efficiently with a loop
- Random forest without controlling for outliers minimized error and was the most effective in modeling our sales data.
- Are the products in our data skewed towards certain holidays?
- Future Considerations
 - Introduce other methods of analysis like ARIMA
 - Utilize feature engineering to reduce number of features
 - A more organized feature optimization process
 - Further dive into gender gap in spending

Appendix

Data Cleaning and Preparation



- **Quality Assessment:**
 - **Handling Missing Values:** Imputed or removed incomplete entries.
 - **Removing Duplicates:** Ensured each transaction is unique.
- **Data Transformation**
 - **Standardizing Categories:** Unified product categories and descriptions.
 - **Feature Engineering:**
 - Added a **Near_Holiday** indicator for transactions close to holidays.
 - Extracted **Month** and **Day_of_Week** from transaction dates.
 - Calculated **Total_Cost** and **Avg_Dollars_Saved**.
- **Encoding and Scaling**
 - **Categorical Variables:**
 - **Numerical Features:** Scaled using StandardScaler.

Data Cleaning and Preparation



Data Summary

- **Final Dataset Size:** 52,955 rows, 23 columns after cleaning.
- **Key Columns:**
 - **Transaction-related:** Transaction_Date, Quantity, Total_Cost
 - **Demographic:** Gender, Location
 - **Promotional:** Discount_Percentage, Coupon_Status
 - **Seasonal:** Month, Near_Holiday

Research Questions



Holiday and Seasonal Trends

- How do spending behaviors change around major holidays? Are there notable spikes in spending associated with specific holidays?
- Do certain times of the year consistently show higher or lower spending trends?

Demographic Influences

- How does gender affect spending behavior? Are there notable differences in average spending between male and female customers?
- Does location impact spending trends? Do customers from specific states or regions spend more or less, especially during holiday seasons?

Research Questions



Promotional Effectiveness

- How effective are discounts and coupon codes in driving sales? Is there a direct correlation between coupon usage and increased spending?
- What percentage of total spending is attributed to discounted purchases, and do discounts result in higher purchase quantities?

Product Insights

- Which product categories are most purchased during holidays, and what is their average price point?

Trend Prediction

- Can historical data be used to predict spending behavior for upcoming holidays or seasons?

Product Breakdown

