EDA mini project - Online Shopping in the USA

Objectives

- Diagnose essential factors contributing to Revenue
- Analyze transaction frequency (time-series)
- Examine product sales performance
- Study customer demographics
- Investigate customer behaviors in different diversities, main focus on tenure and coupon.
- Optimize business revenue (online shopping)

Dataset Description

We used a dataset from Kaggle, in which the dataset was collected and consolidated. It consists of 52,955 transaction records of a business in the USA covering the year 2019. Variables are identified as below:

Type	Variable
Categorical	Coupon status, Gender, Product Category
Numeric	Average Price, Customer ID, Delivery Charges, Discount Percentage, GST (Goods & Services Tax), Offline Spends, Online Spends, Transaction ID, Month
Date	Date, Transaction Date
Text	Coupon Code, Location, Product SKU, Product Description

Methodology

Data Cleaning

We have created a new column for effective analysis, which is called Gross Revenue, and is calculated by the following formula: [[(Quantity * Average Price)*(1-Discount Rate in % {only "Used" is adopted})]] * (1+ GST in %) + Delivery Charge. We have assumed that the delivery charge was paid by the customer and the delivery charge was equal to its cost.

We have also deleted the missing data and the noisy data, like the columns - offline and online spend.

```
present_columns = ['Coupon_Code','Discount_pct','CustomerID','Gender','Location','Tenure_Months','Transaction_ID','Transaction_Date','Product_SKU','Product_Descri
sales_data = sales_data[sales_data.isnull().any(axis=1)]
sales_data[present_columns]
```

0.0s

	Coupon_Code	Discount_pct	CustomerID	Gender	Location	Tenure_Months
52927	NE10	10.0	Missing value	Missing value	Missing value	Missing value
52947	HOU10	10.0	Missing value	Missing value	Missing value	Missing value
52937	NOTES10	10.0	Missing value	Missing value	Missing value	Missing value
52952	AND10	10.0	Missing value	Missing value	Missing value	Missing value
52934	NE10	10.0	Missing value	Missing value	Missing value	Missing value
52924	NOTES10	10.0	Missing value	Missing value	Missing value	Missing value
3895	Missing value	Missing value	17850.0	M	Chicago	12.0
3896	Missing value	Missing value	17850.0	M	Chicago	12.0
3897	Missing value	Missing value	14060.0	F	Chicago	22.0
3898	Missing value	Missing value	15061.0	M	California	7.0
3899	Missing value	Missing value	18074.0	F	California	10.0
3900	Missing value	Missing value	16029.0	F	Washington DC	40.0

Transaction_Date	CustomerID	Offline_Spend	Online_Spend	Gender	Location
1/1/2019	17850	4500	2424.5	M	Chicago
1/1/2019	17850	4500	2424.5	M	Chicago
1/1/2019	17850	4500	2424.5	M	Chicago
1/1/2019	17850	4500	2424.5	M	Chicago
1/1/2019	17850	4500	2424.5	M	Chicago
1/1/2019	13047	4500	2424.5	M	California
1/1/2019	13047	4500	2424.5	M	California
1/1/2019	13047	4500	2424.5	M	California
1/1/2019	13047	4500	2424.5	M	California
1/1/2019	12583	4500	2424.5	M	Chicago
1/1/2019	15100	4500	2424.5	M	California
1/1/2019	14688	4500	2424.5	F	New York
1/1/2019	14688	4500	2424.5	F	New York
1/1/2019	14688	4500	2424.5	F	New York

Data Transformation

Customers are divided into clusters for further analysis.
The cluster 1,2,3 (non-red color) will be defined as the normal-scale customer.
The cluster 3 (red color) will be defined as the large-scale customer.

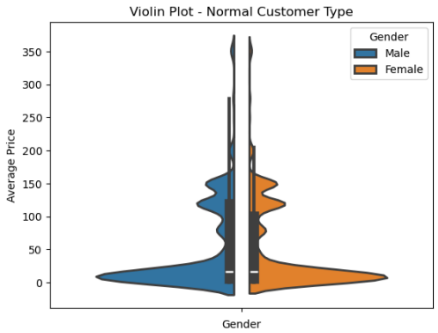
Analysis and Findings

There are several directions in our study, which are listed in the following:

	Quantity	Average_Price	Total_Price	Customer Tenure Months
count	52,924.00	52,924.00	52,924.00	52,924.00
mean	4.50	52.24	103.45	26.13
std	20.10	64.01	172.76	13.48
min	1.00	0.39	4.60	2.00
25%	1.00	5.70	21.24	15.00
50%	1.00	16.99	47.18	27.00
75%	2.00	102.13	138.05	37.00
max	900.00	355.74	8,979.60	50.00

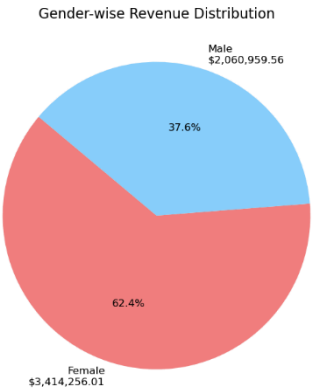
The dataset contains 52,924 records for each variable.
The average quantity purchased is approximately 4.50, with a standard deviation of 20.10, indicating high variability.
The average price is $52.24, with a wide range from $0.39 to $355.74.
The total price ranges from $4.60 to $8,979.60, with an average of $103.45.
Customer tenure spans from 2 to 50 months, with an average of 26.13 months.

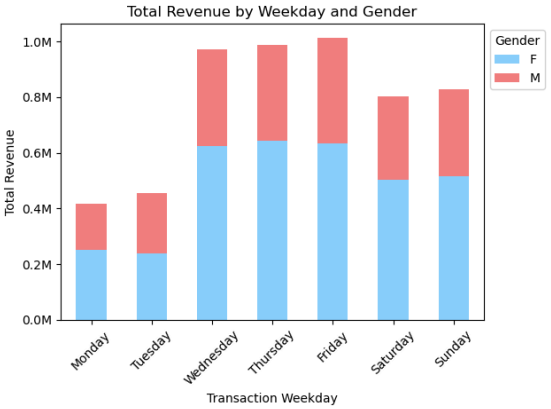
Average Price



The average purchase price is fairly close between male and females. The average prices are concentrated on US $6 and US $50. However, male group shows a slightly higher upper limit and this suggests that male customers occasionally splurge more. Sales strategies could possibly be augmented by providing high-end offers to male customers in a limited period.

Revenue by Gender / Weekday

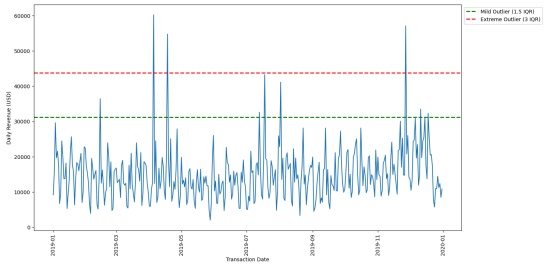


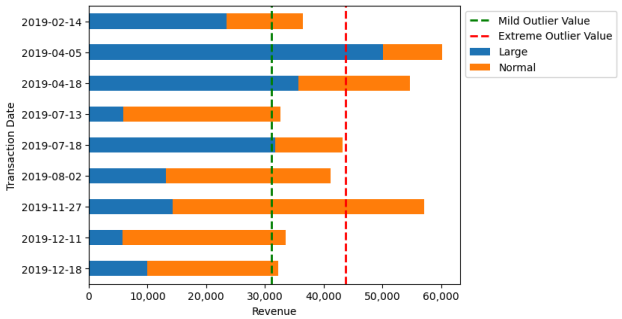


From a gender perspective, the Gender-wise Revenue Distribution highlights a significant difference in revenue contributions between genders. Female and Male customers generated $3,414,256.01 (62.36%) and $2,060,959.56 (37.64%) respectively to the total revenue. This disparity underscores the importance of understanding gender-specific preferences and behaviors to optimize marketing strategies.

From days of the week perspective, it reveals that females shopped heavily from Wednesday to Friday, with daily revenue doubling Monday's figures. Monday and Tuesday total revenues are half of other days. This presents an opportunity—targeted promotions for female shoppers on Mondays and Tuesdays to boost sales.

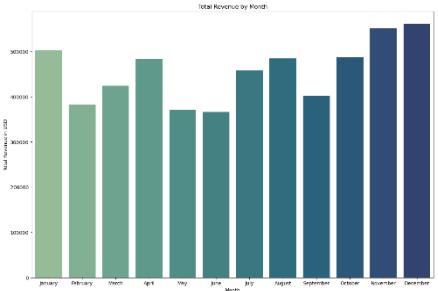
Revenue by day





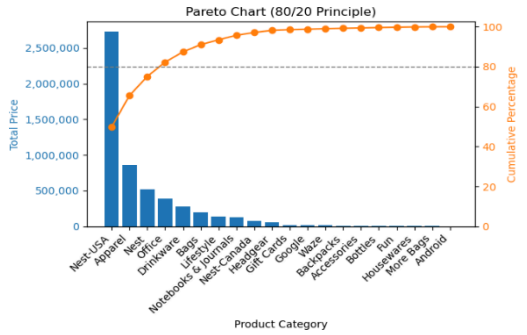
The extreme online sales outliers include April 5th at 60,239 dollars, which was nearly 5 times the IQR with 50,000 dollars coming from just three customers, April 18th at 54,761 dollars, up 4.3 times likely due to Easter, and November 27th at 57,081 dollars, up 4.58 times likely from Black Friday. These peaks all exceeded 54,000 dollars, marking major sales events. To prevent stock issues, reach out to key customers like those behind the April 5th spike to better plan inventory.

Gross revenue

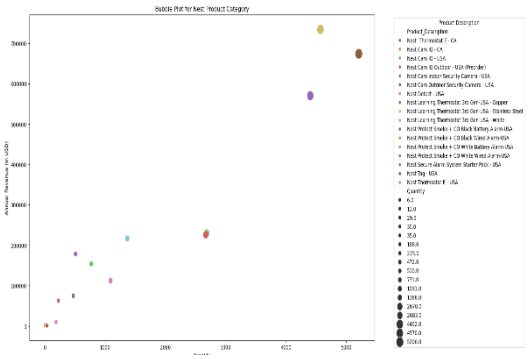


It is observed that higher revenue was recorded at the Start and End of the Year. Possible reason is seasonal demand as there were a total of 10 holidays in 2019 in the US, with 3 holidays in Q1 and 4 holidays in Q4.

Product Categories



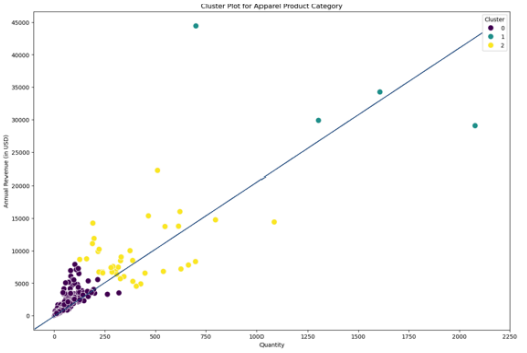
The top 3 and worst 3 product categories can be recognized easily from this chart. For top 3: Nest-USA, Nest, Apparel categories contributed the top 3 revenue amounts among all categories and are worth emphasis. For the worst 3: Fun, More Bags, Android categories contribute the least and the business might have to strictly control inventory planning of these products.



By breaking it down into Nest Product Category (Top 1, Top 3), key popular products are Security Camera, Learning Thermostat, Cam IQ.

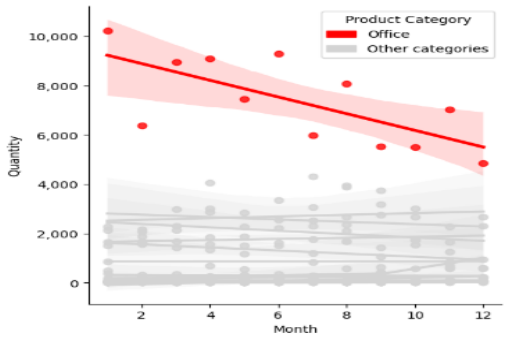
- Security Camera: Both Outdoor and Indoor had high demand, but Outdoors' demand was higher
- Thermostat: Stainless Steel is better than all the other learning thermostat and thermostat.
- Camera: USA / CA version are roughly the same.

Understanding this best-selling information could boost our sales via selecting them in our product marketing or recommendation in a face-to-face scenario.



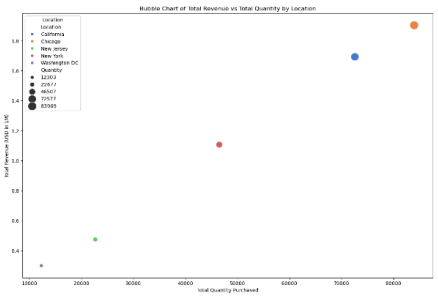
The Apparel Category has 200 sub-categories and thus cluster plot is used to find the appropriate cluster. Three groups are clearly classified to purple (quantity <250 , revenue < 6000), yellow (250 < quantity < 1250 , 6000 < Annual Revenue < 25000) and green (bulk quantity and revenue from 30000 – 45000).
To maximize revenue of the business:

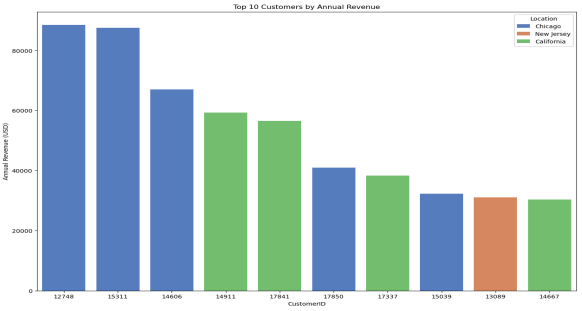
- Business focus (i.e. Marketing & Inventory) on group 1 & 2 should be equally important because group 2 has more number of categories and its aggregated outcome is similar to group 1.
- In all groups, those points above the diagonal (with higher annual revenue in the same x-axis) should be emphasized because they generate higher revenue with the same quantity sold.



Furthermore, it is observed that Office Category is undergoing a downward trend given that coupons are provided. This shift suggests that the change in customer preferences or market dynamics will have to be explored further from the business side. Are customers moving to alternatives, or is demand softening?

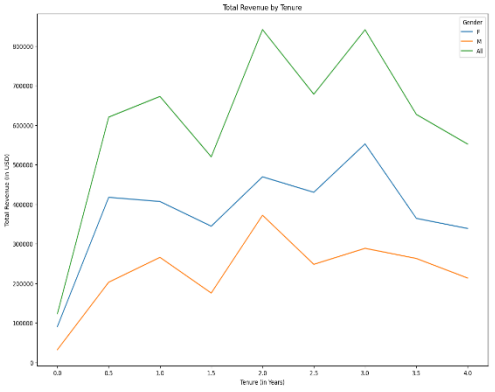
Customer Segments

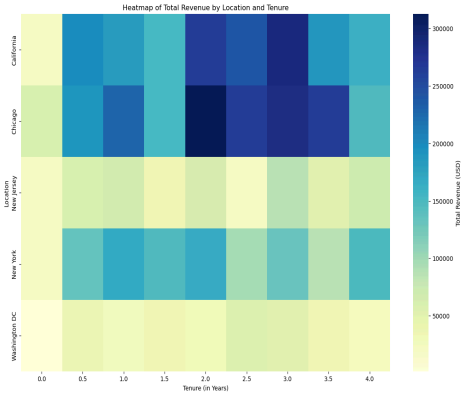




1. Revenue by Customer Location: Chicago > California > New York > New Jersey > Washington DC.
   A rough positive correlation between total quantity purchased and total revenue could be observed but the small number of observations might not be confident enough to form a comprehension prediction.
2. Majority of Top 10 customers come from Chicago and California. Another spotlight is customer 13089 from New Jersey. These customers should be treated with priority.

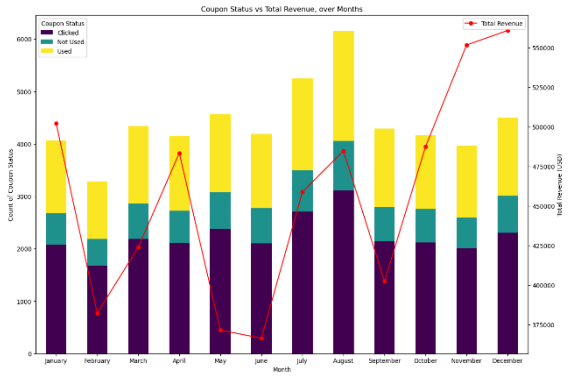
Tenure Study (Customer Loyalty)

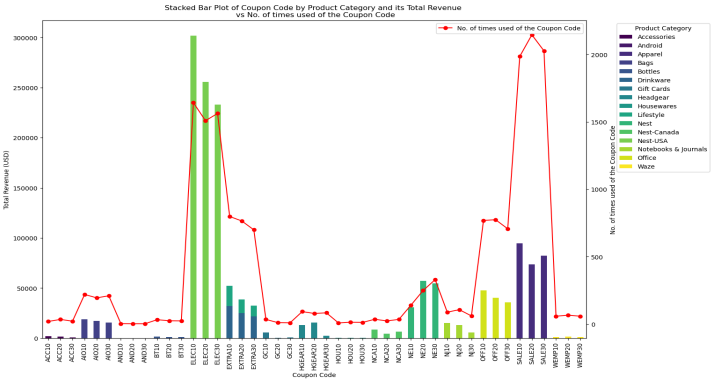




Generally, it is confident to claim “Higher tenure, Higher total revenue “.
The effect of "higher tenure, higher total revenue" is less obvious in the location New Jersey and Washington DC. The business has to find ways to improve in these two locations.

Coupon Study (Customer Incentive)

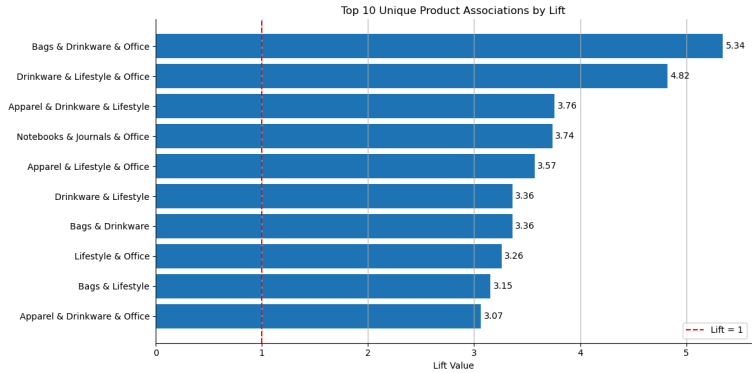


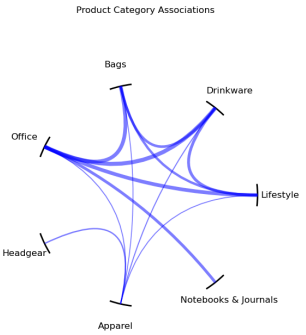


1. In LHS, Q2 and Q3 generally stay at a relatively lower annual revenue level compared to Q1 & Q4. The increase in use of coupons in July and August could actually boost revenue and vice versa in February. It is confident that coupon incentive plays a vital role in boosting revenue and there is room for improvement in issuing more appropriate coupon incentives in Q1 & Q4.
2. The most effective Coupon Codes (most used and notable revenue) are as followings:
   - ELEC 10 to ELEC 30: mainly used for Nest Electrical Products
   - SALE 10 to SALE 30: mainly used for Accessories, Android, Apparel, which is general usage purpose

It is believed that issuing more similar coupon code and terms could help to boost revenue.

Product category associations





Drinkware frequently co-occurs with Bags and Office products (high lift >4).
Office is a common antecedent, suggesting bundling opportunities (e.g., promote Drinkware with Office supplies).
Lifestyle products show moderate ties to Drinkware + Office (lift ~4.8).
Cross-Selling: Target customers buying Office or Bags with Drinkware recommendations.

Challenges and Limitations

Challenges	● Lack of information Operating costs & original price to calculate the net profit. ● Determinations in spending The online spends & Offline spends columns are undefined in the dataset. ● Idiopathic reasons for not using coupons The majority of customers clicked the coupons without using them.
Limitations of the approach or dataset	● Missing information, such as product cost and who paid the delivery fee. ● Offline & Online spends in determination ● Lack of cost information to calculate the net revenue column
Potential areas for future improvement	● Find out the definition of offline & online spending ● Assessment of pre- & post-order customer service satisfaction to stimulate sales. ● Questionnaires about further improvement on delivery service, product durability & website design to maintain customer loyalty.

Future Work

- Description of any additional ideas or approaches that were not implemented
  - Deep investigation of Individual Products
  - Because of the time constraint, there could be a hundred individuals in one product category.
- Recommendations
  - Assessment of pre- & post-order customer service satisfaction to stimulate sales.
  - Questionnaires about further improvement on delivery service, product durability & website design to maintain customer loyalty.

Conclusion

Summary of the project's objectives and achievements	Objectives Diagnose essential factors contributing to Revenue. Afterwards, provide business owners with data-driven recommendations for Revenue Maximization (thus Net Profit). Achievements: Identified best business approaches in total revenue, product categories, customer segments, customer loyalty, and customer incentives.
Key takeaways and insights gained	Revenue: Seasonal Demand Product Categories: Best Selling Products (Nest & Apparel) Customer Segments: Top Locations, Top Customers Customer Loyalty: Tenure matters Customer Incentives: Coupon matters and the best Coupon Code were ELEC and SALE.

Work Distribution

Even distribution among team members