ECOMMERCE ANALYSIS

25 OCTOBER, 2024

OVERVIEW

1 Question

2 Dataset Summary

3 Hypothesis and testing

4 Conclusion

CUSTOMER BEHAVIOR

QUSTIONS

1. Do the customer likely to buy cheap or expensive product

2. what time and what day have the highest traffic of people viewing products

3. what date have the most sales and least sales.

4. How loyal are people to a brand.

5. When the person purchased the item, what other item did the person also purchase at the same time? (using the association rules)

DATA SUMMARY

	event_time	event_type	product_id	category_id	category_code	brand	price	user_id	user_session
0	2019-10-01 00:00:00 UTC	view	44600062	2103807459595387724	NaN	shiseido	35.79	541312140	72d76fde-8bb3-4e00-8c23-a032dfed738c
1	2019-10-01 00:00:00 UTC	view	3900821	2053013552326770905	appliances.environment.water_heater	aqua	33.20	554748717	9333dfbd-b87a-4708-9857-6336556b0fcc

# of rows: 42mil

# of Cols: 9

Property	Description	Type	Note
event_time	Time when event happened at (in UTC).	datetime64	total of 42448764 time stamp
event_type	Only one kind of event: purchase.	object	3 event type in total view = 96%, cart = 2.2% and purchase = 1.7%
product_id	ID of a product	int64	166794 unique product
category_id	Product's category ID	int64	624 unique category ID
category_code	Product's category taxonomy (code name) if it was possible to make it. Usually present for meaningful categories and skipped for different kinds of accessories	object	126 unique category code
brand	Downcased string of brand name. Can be missed.	object	3444 unique brand
price	Float price of a product. Present.	float64	min spend = $0.00, spend = $2547 and average spending = $290
user_id	Permanent user ID.	int64	3022290 user active this month
user_session	Temporary user's session ID.	object	

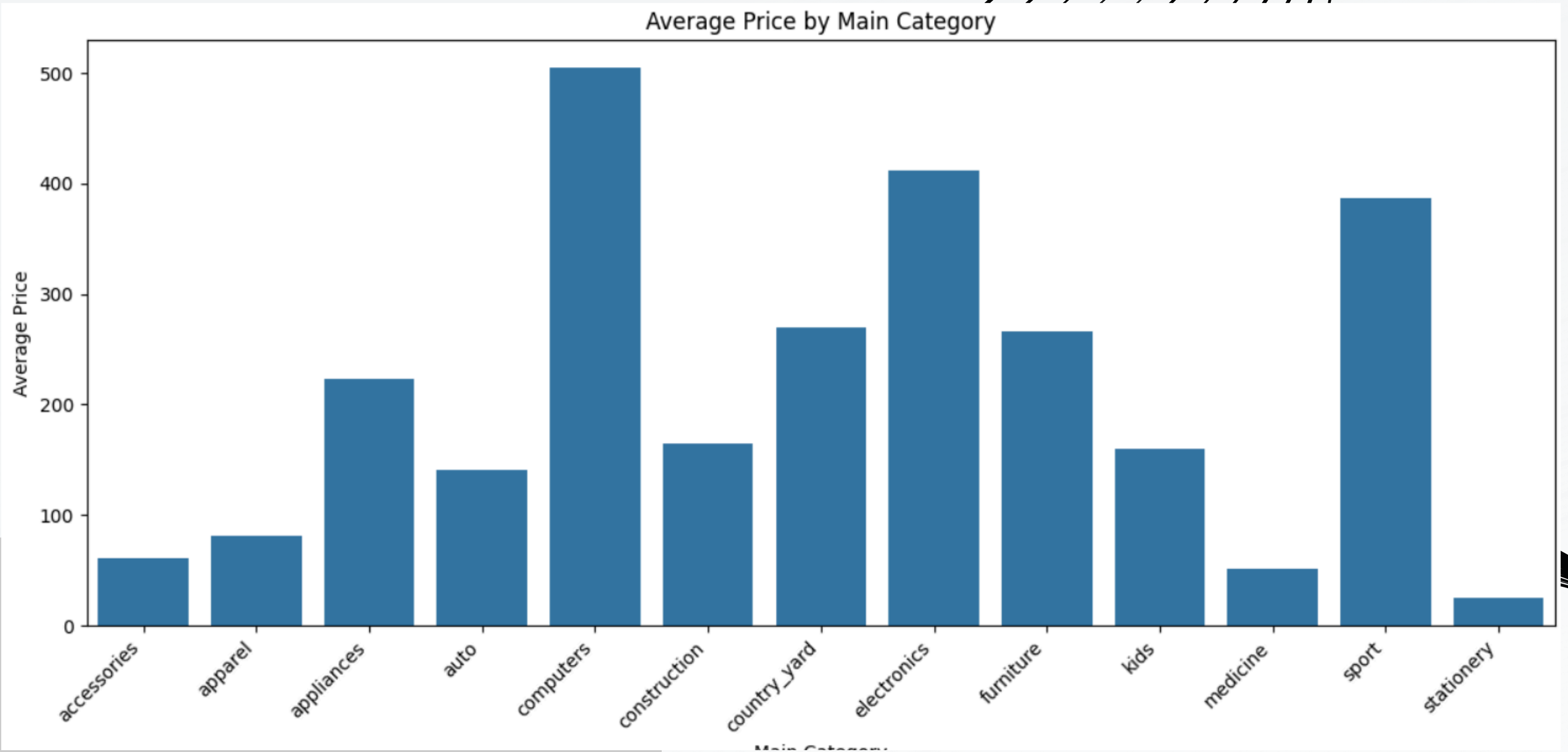
EDA 1

DO PEOPLE LIKELY TO BUY CHEAP OR EXPENSIVE PRODUCT/S

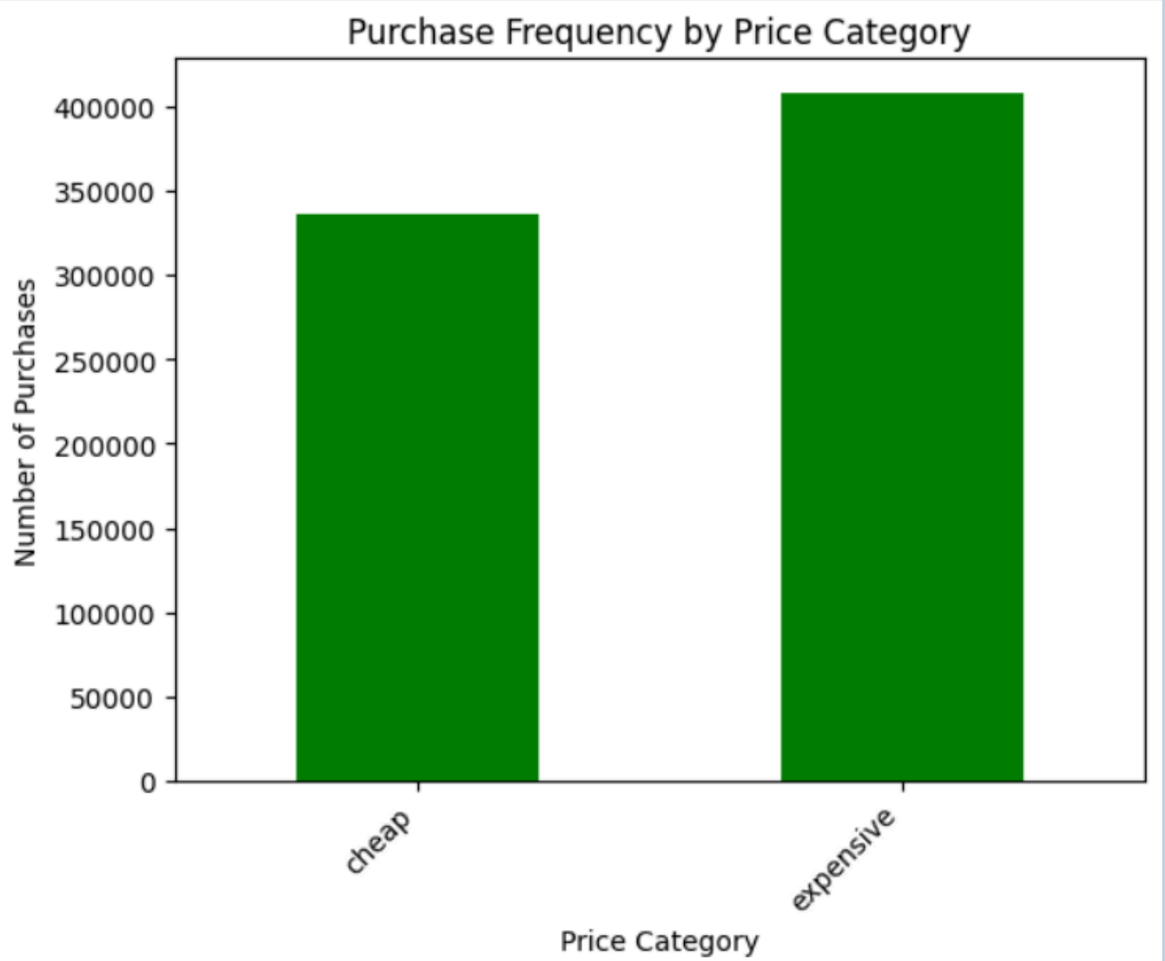
Hypothesis: There is a practical difference in purchase preference between cheap and expensive products

EDA 1

VISUALIZATION



EDA 1




```
# Group by 'price_category' and count purchases
purchase_frequency = new_df[new_df['event_type'] == 'purchase'].groupby('price_category')['event_type'].count()

   # Print the result
print(purchase_frequency)

# Visualize the result (optional)
purchase_frequency.plot(kind='bar', color='green')
plt.title('Purchase Frequency by Price Category')
plt.xlabel('Price Category')
plt.ylabel('Number of Purchases')
```

HYPOTHESIS TESTING

Null Hypothesis: There is no difference in purchase preference between cheap and expensive products.

Alternative Hypothesis: There is a difference in purchase preference. People prefer one price category over the other.

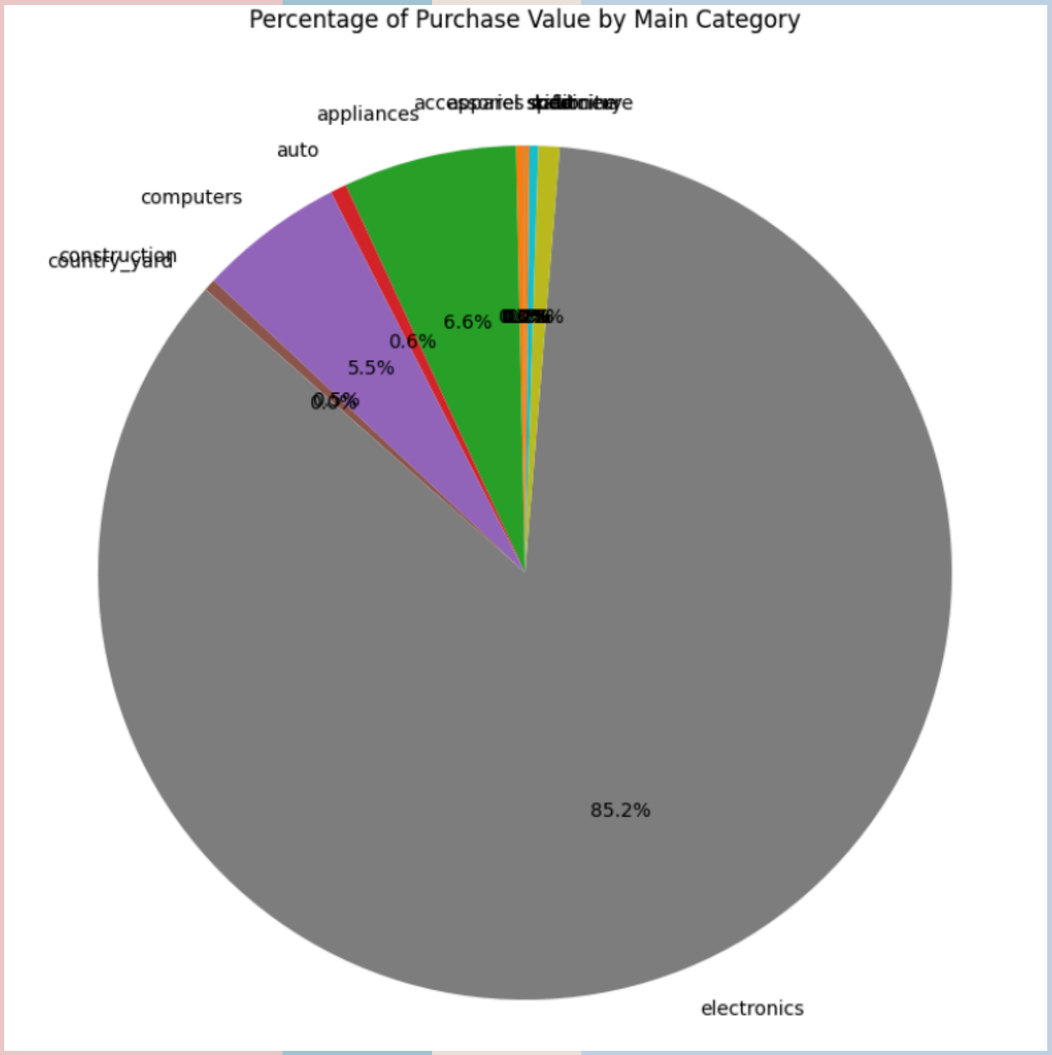
HYPOTHESIS TESTING

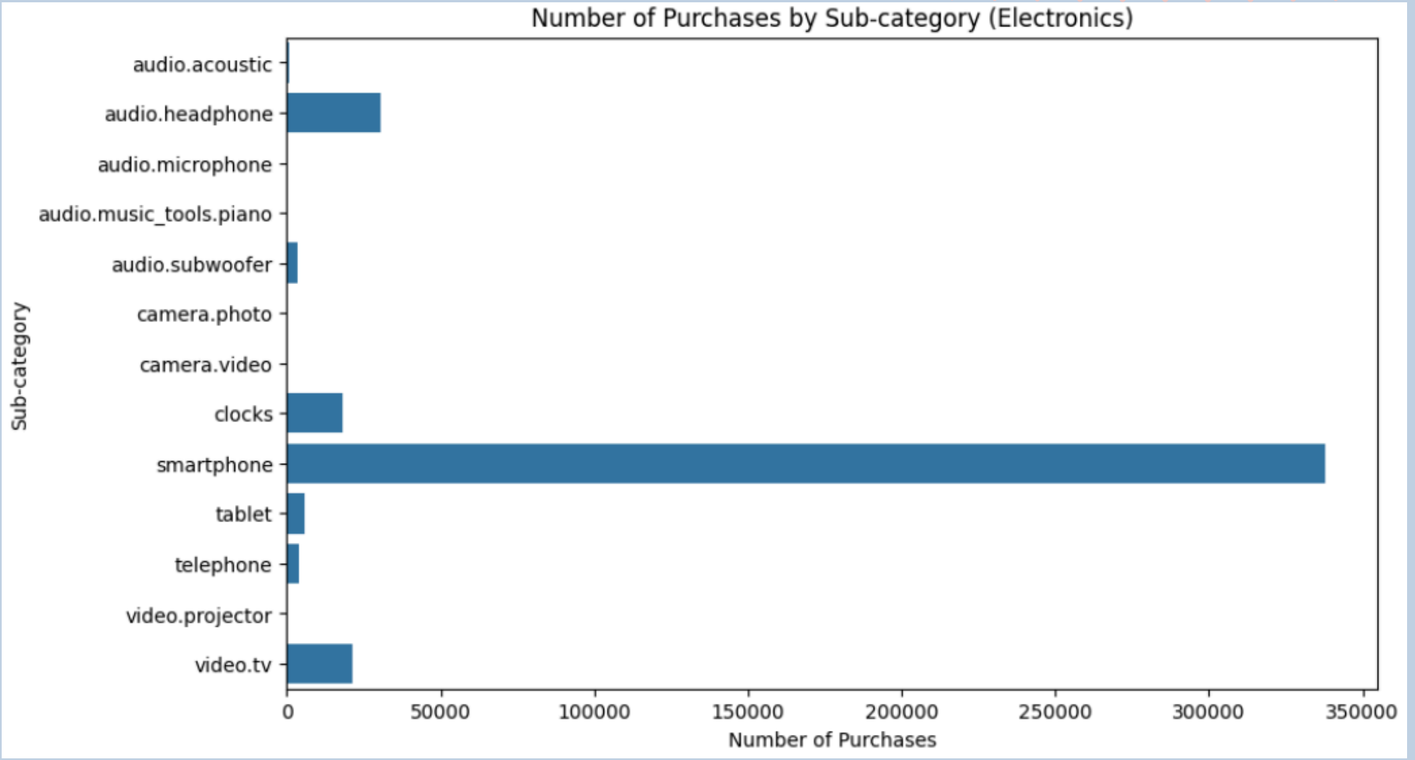
```python
import pandas as pd
from scipy.stats import chi2_contingency

# Create price bins
new_df['price_bin'] = pd.qcut(new_df['price'], q=3, labels=['low', 'medium', 'high'])

# Create contingency table
contingency_table = pd.crosstab(new_df['main_category'], new_df['price_bin'], values=new_df['event_type']=='purchase', aggfunc=sum)

# Perform Chi-Square test
chi2, p, dof, expected = chi2_contingency(contingency_table)

print(f"Chi-Square Statistic: {chi2}")
print(f"P-value: {p}")
print(f"Degrees of Freedom: {dof}")
```

Chi-Square Statistic: 118878.53748678904
P-value: 0.0

Conclusion: There is a significant difference in purchase preference between cheap and expensive products. Meaning that people have “preference”.

EDA 1





85% of people purchased electronics and most of them are buying smartphone

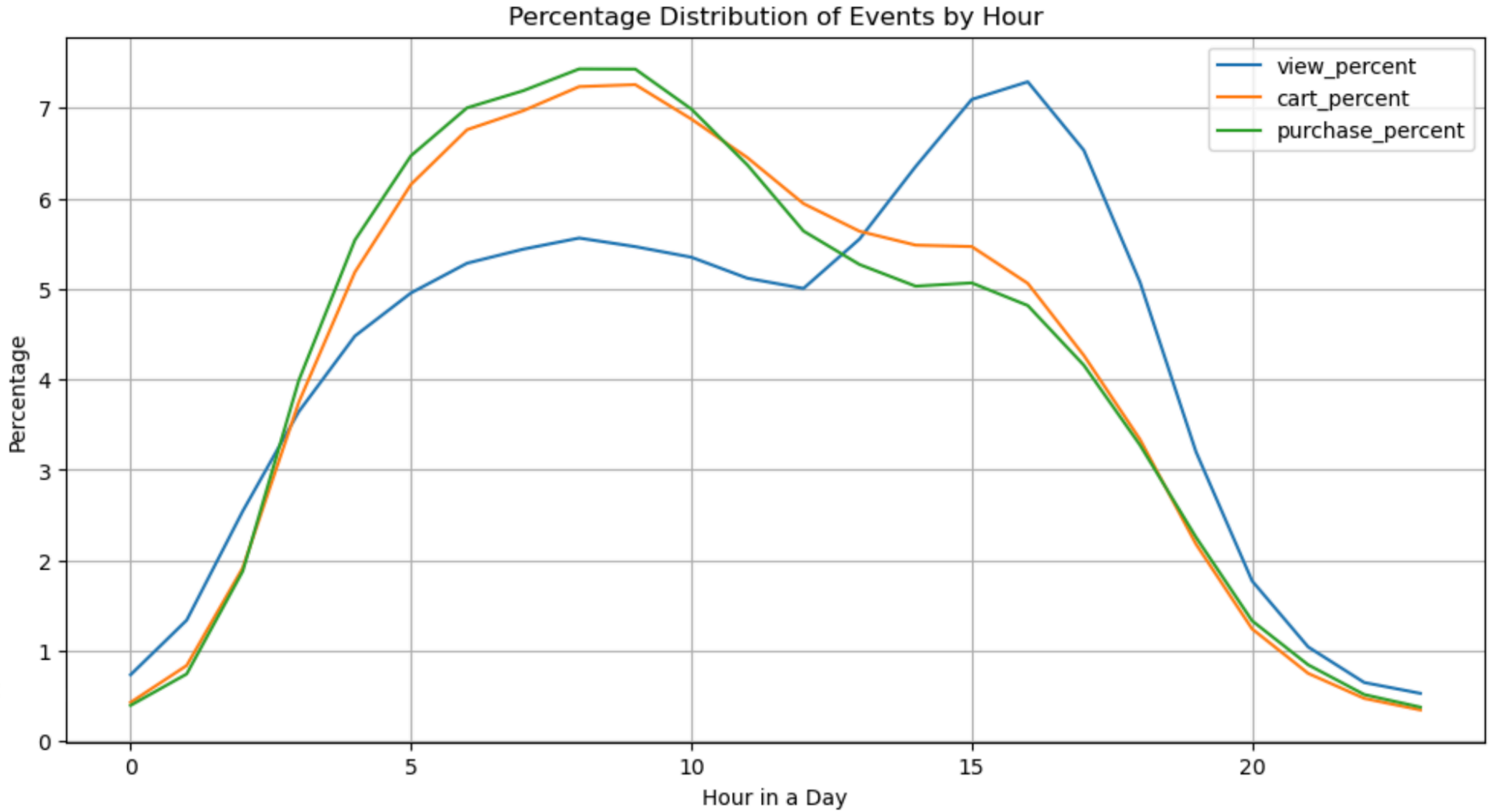
EDA 2

WHAT TIME AND WHAT DAY HAVE THE HIGHEST TRAFFIC OF PEOPLE VIEWING PRODUCTS

Hypothesis: most product views occur during weekends or evenings when users have more free time compared to weekdays or working hours

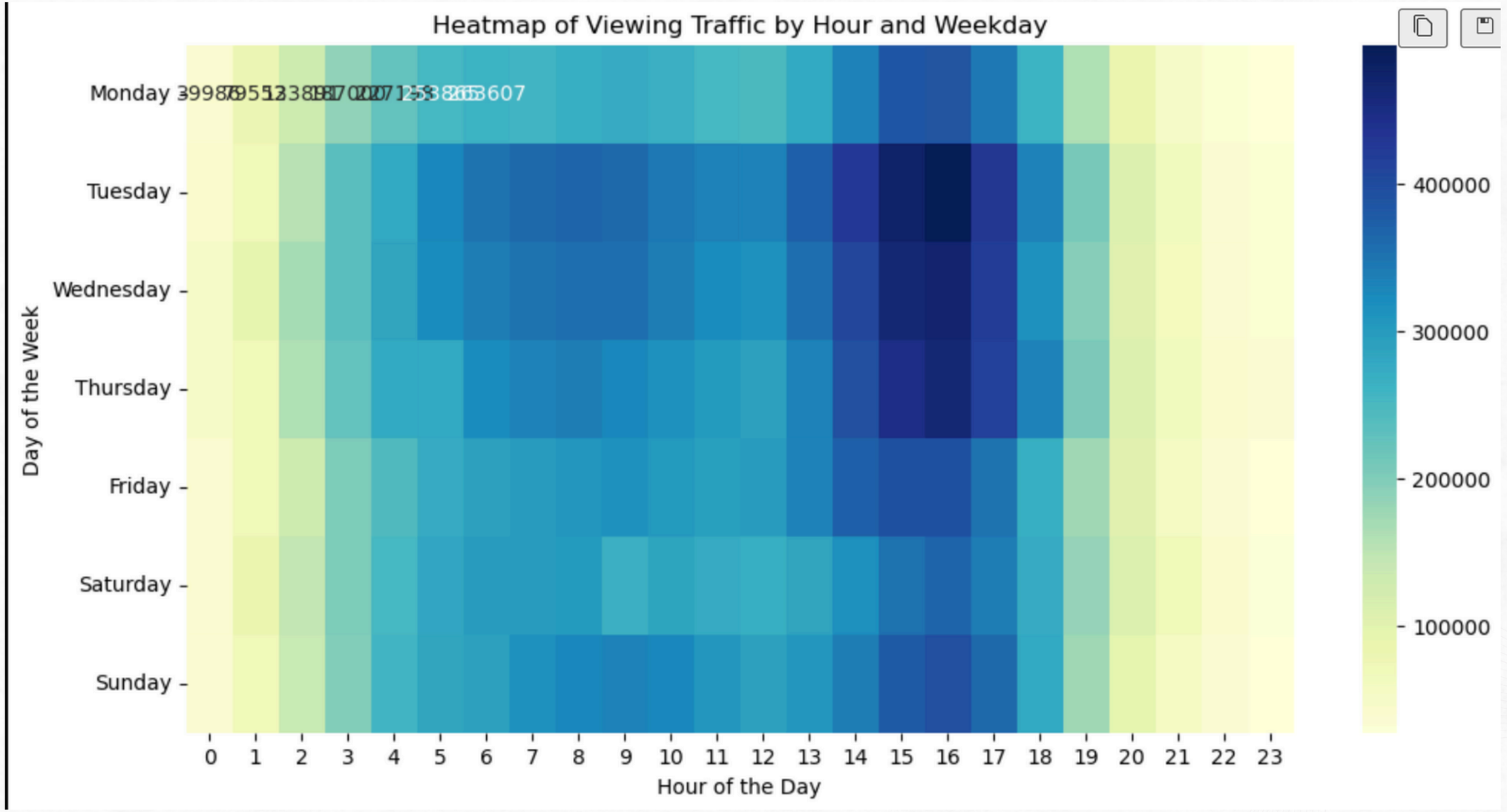
EDA 2

VISUALIZATION



EDA 2

VISUALIZATION



EDA 2

HYPOTHESIS TESTING

Null Hypothesis: Product views are independent of the day of the week and the time of day (i.e., there is no difference in viewing patterns

Alternative Hypothesis: Product views are dependent on the day of the week and time of day (i.e., there are more views during weekends/evenings.

```
df['event_time'] = pd.to_datetime(df['event_time'])

df['hour'] = df['event_time'].dt.hour
df['day_of_week'] = df['event_time'].dt.dayofweek    "dayofweek": Unknown word.

df['is_weekend'] = df['day_of_week'] >= 5
df['is_evening'] = (df['hour'] >= 12) & (df['hour'] < 17)
view_data = df[df['event_type'] == 'view']
contingency_table = pd.crosstab(view_data['is_weekend'], view_data['is_evening'])

chi2_stat, p_val, dof, expected = chi2_contingency(contingency_table)
```

with Chi2 Stat: 11202.846276248023,
P-value: 0.0
Reject the null hypothesis (H_0):

Thus Product views are dependent on the day of the week and time of day. Which mean more views occur during the afternoon. When they have more free time.

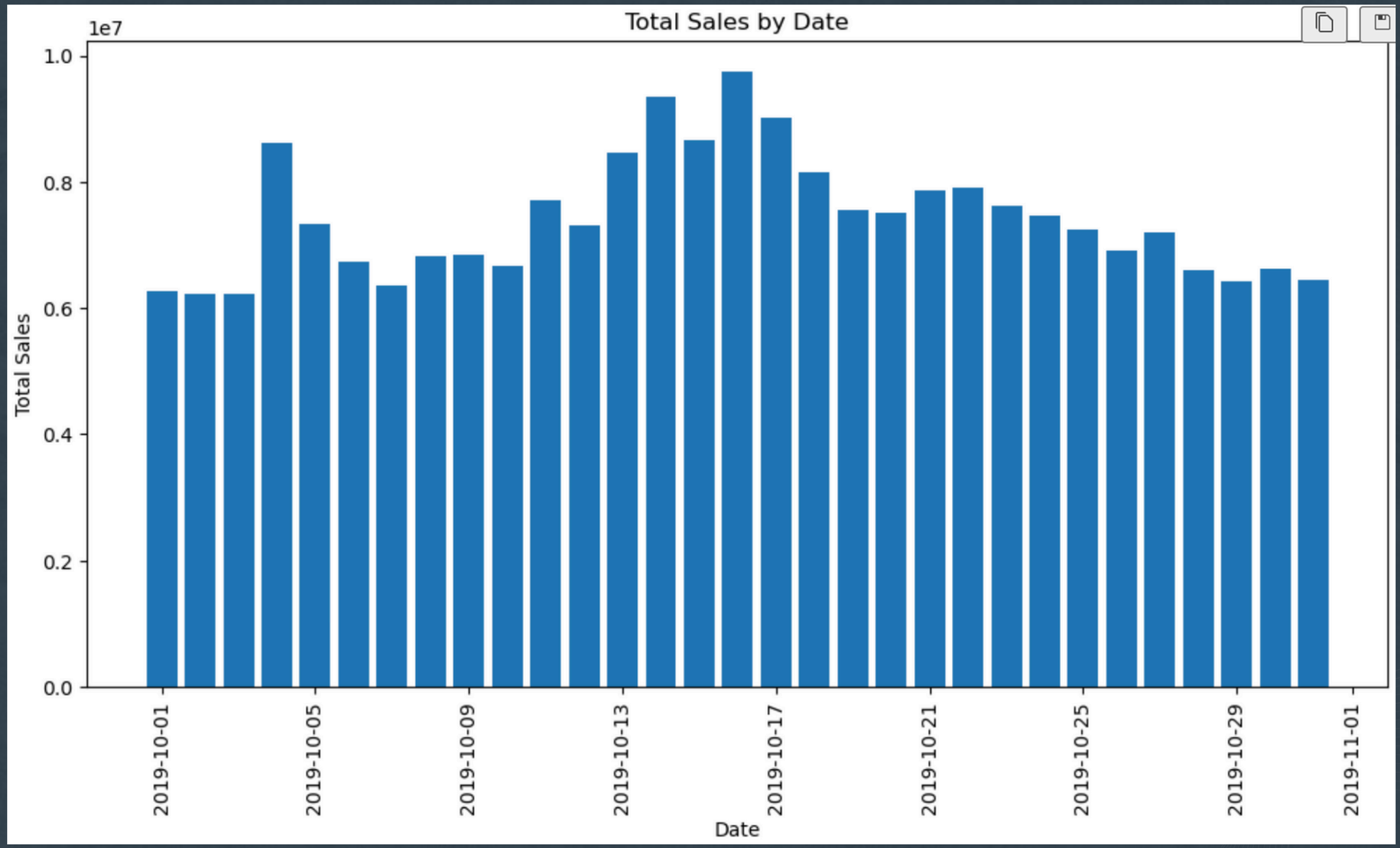
EDA 3

WHAT DATE HAVE THE MOST SALES AND LEAST SALES.

Hypothesis: Sales are higher on certain dates (e.g., weekends) compared to regular weekdays.

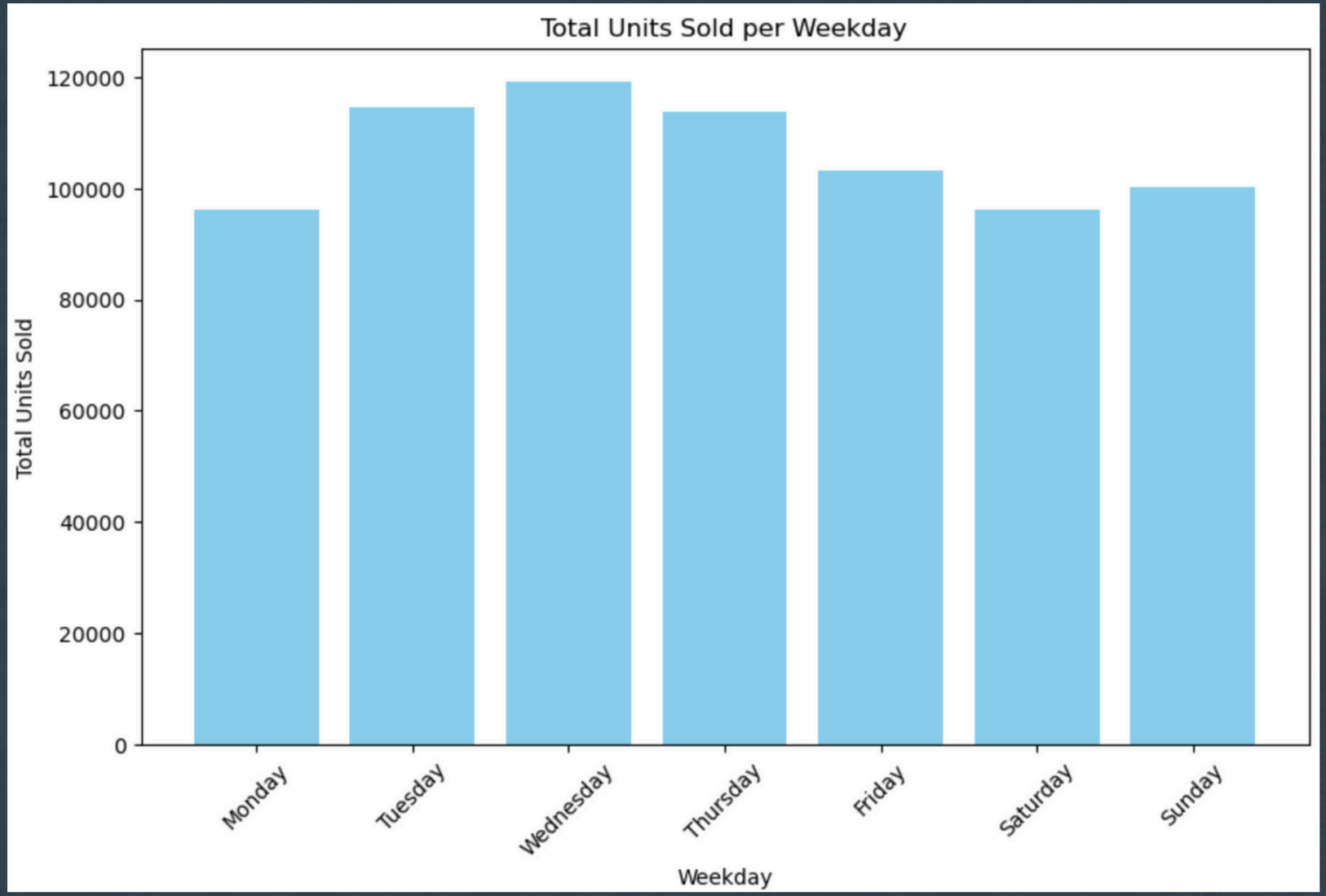
EDA 3

VISUALIZATION



EDA 3

VISUALIZATION



EDA 3

HYPOTHESIS TESTING

Null Hypothesis: There is no difference in sales between weekends and weekdays.

Alternative Hypothesis: There is a Sales on weekends are significantly different from weekdays.

with T-statistic: 0.6201632574068122, P-value: 0.539993425770068

Fail to Reject the null hypothesis (H_0):

```
df['purchase_time'] = pd.to_datetime(df['event_time'])

sales_by_date = df[df['event_type'] == 'purchase'].groupby(df['purchase_time'].dt.date).size().reset_index(name='sales_count')

sales_by_date['date_ordinal'] = pd.to_datetime(sales_by_date['purchase_time']).apply(lambda x: x.toordinal())

correlation, p_value = pearsonr(sales_by_date['date_ordinal'], sales_by_date['sales_count'])    "pearsonr": Unknown word.
```

After performing an t-test we got a P-value of 0.53 which is higher than 0.05 thus we fail to reject the null hypothesis. Which show that there are no high differences between the sales in weekdays and weekends.

with Chi2 Stat: 11202.846276248023,

P-value: 0.0

Reject the null hypothesis (H_0):

EDA 4

BRAND LOYALTY

Hypothesis

People are more likely to be loyal to a brand and purchase multiple items from the same brand.

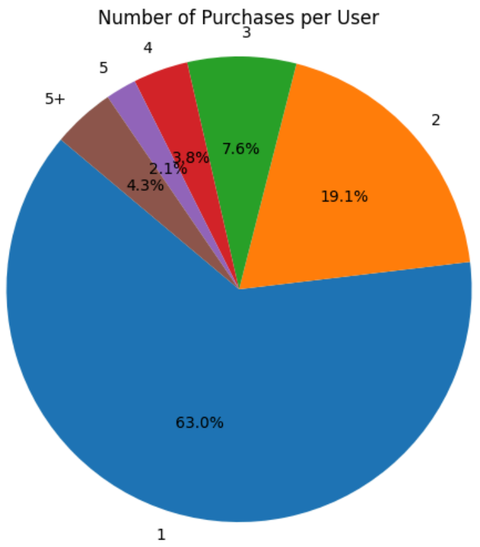
40

Average Purchases

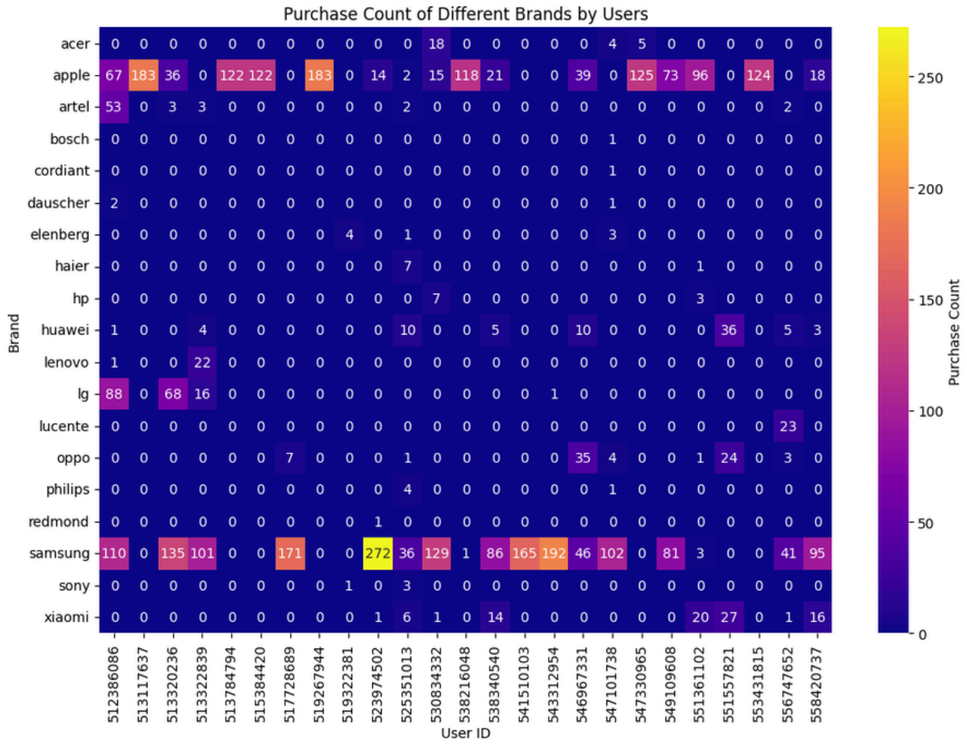
The average number of item purchased from users who had purchased a product is 40.

DATA VISULISATION

NUMBER OF PERCHASES PER USER



PURCHASES OF DIFFRENT BRANDS BY TOP BUYERS



EDA 4

NUMBER OF LOYAL CUSTOMERS

24.7%

LOYAL CUSTOMERS

24.7% of customers are loyal to a brand and continue to purchase product from the same brand

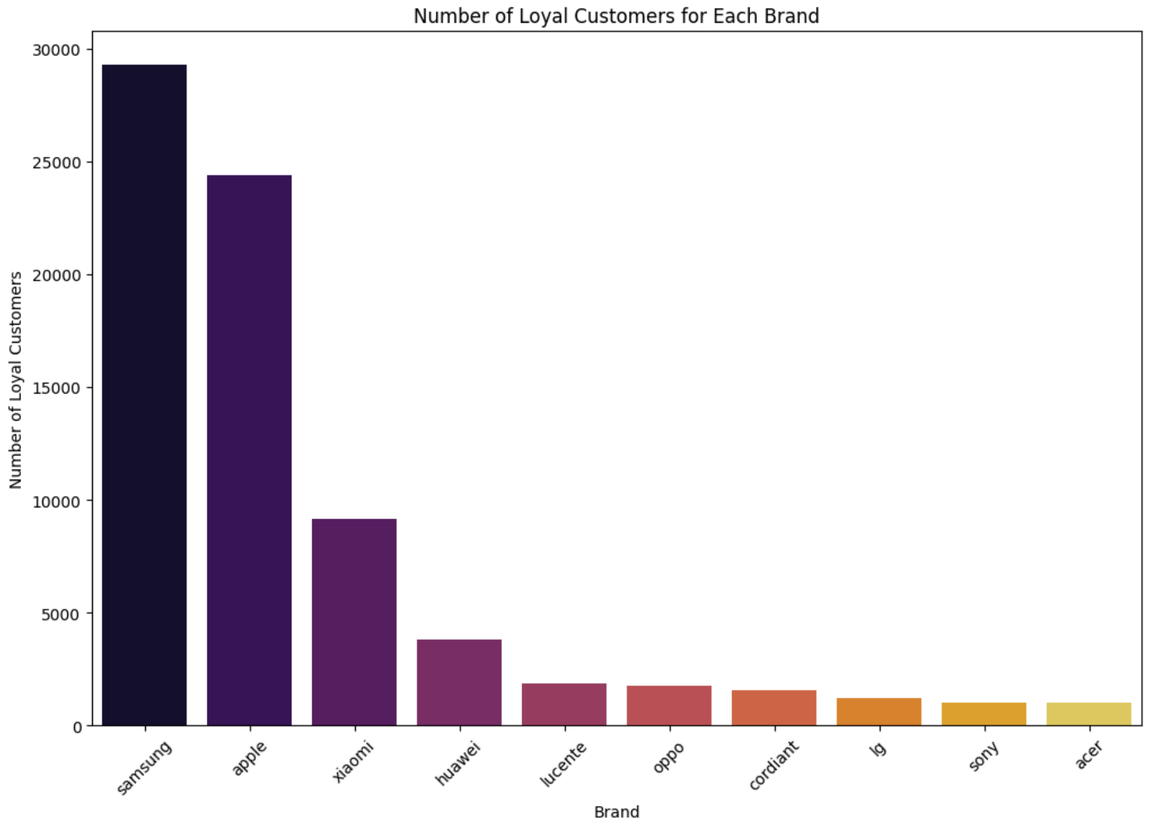
```python
# Calculate the number of loyal customers for each brand
purchase_df = df[df['event_type'] == 'purchase']
purchase_df = purchase_df.sort_values(by=['user_id', 'event_time'])
purchase_counts = purchase_df.groupby(['user_id', 'brand']).size().reset_index(name='purchase_count')
purchase_counts['loyal'] = purchase_counts['purchase_count'] > 1
loyal_customers = purchase_counts[purchase_counts['loyal'] == True]
percent_loyal_customers = len(loyal_customers) / len(purchase_counts) * 100

same_brand = len(loyal_customers)
different_brand = len(purchase_counts) - same_brand
contingency_table = [[same_brand, different_brand]]

# Chi-squared test
chi2, p, _, _ = chi2_contingency(contingency_table)

print("Number of Loyal Customers for Each Brand:")
print(same_brand)

print("Number of Customers Purchasing from Different Brands:")
print(different_brand)

print("Percentage of Loyal Customers:")
print(percent_loyal_customers)

print("Chi-squared Test Statistic:", chi2)
print("p-value:", p)
```

75.3%

NON-LOYAL CUSTOMERS

Over 75% of customer did not stick to the same brand when purchasing products.

CONCLUSION



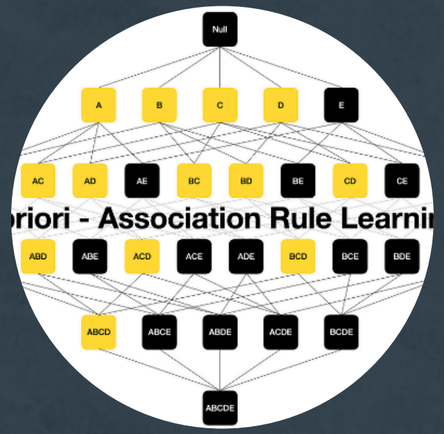

```
Number of Loyal Customers for Each Brand:
105047
Number of Customers Purchasing from Different Brands:
319752
Percentage of Loyal Customers:
24.728636366846438
Chi-squared Test Statistic: 0.0
p-value: 1.0
```

Result

Through over calculation and test we found that most customer are not loyal to a single brand and chose to purchase product from a variety of brands proving our hypothesis wrong

EDA 5

WHEN AN ITEM IS PURCHASED, WHAT OTHER ITEM IS ALSO PURCHASE AT THE SAME TIME?



Apriori

We used Apriori to generate association rules for items which were purchased in one user session to find which product were usually purchased together

24

Rules

We found 24 rules using a minimum support of 0.001 and minimum confidence of 0.1 we had to use low support and confidence as the data set was very large and the chances of the same group of item being purchased is very low

EDA 4

DATA VISULISATION

```
purchase_df = df[df['event_type'] == 'purchase']
purchase_df = purchase_df.sort_values(by=['user_session', 'event_time'])
purchase_sessions = purchase_df.groupby('user_session')['product_id'].apply(list).reset_index(name='products')
purchase_sessions = purchase_sessions[purchase_sessions['products'].apply(len) > 1]

dataset = purchase_sessions['products'].tolist()
dataset = [session for session in dataset if len(session) > 0]

F, support_data = apriori(dataset, min_support=0.001, verbose=True)
rules = generate_rules(F, support_data, min_confidence=0.1, verbose=True)

for rule in rules:
    antecedent, consequent, confidence = rule
    print(f"Rule: {set(antecedent)} -> {set(consequent)}, confidence: {confidence}")
```

NUMBER OF PRODUCTS PUCHASED IN EACH SESSION

	user_session	products
2	00004ada-8f93-49a6-956d-4ed71ae94791	[1005031, 1005031]
4	00005b76-13ba-4afe-b80d-2f2b337d3e92	[1004806, 1005066]
5	00007b3f-ec7a-4d2e-ac44-64dfec829806	[12719135, 12719135]
9	0000de39-dc74-414d-8da6-83ad56135bf5	[1004246, 1002544, 1004767, 1004833]
11	0000fa47-9577-480a-9fa4-be5c25e8dd59	[1004767, 1004836]
...	...	...
629511	fffabe59-a6ad-4159-ac40-c7f091c944c4	[6600082, 6600082]
629523	fffc0956-c9fb-402a-a8b6-80c833487834	[1002544, 1004767]
629526	fffc6f01-ba1f-4f96-af05-399cb2d6b39f	[1004833, 1004833]
629531	fffcc946-4682-4328-8d68-92d155112fd7	[26400301, 5300097]
629555	ffff15c0-016a-4ad9-8a32-982fab93ae39	[1005105, 1005105]

80261 rows × 2 columns

ASSOCIATION RULES

```
Rule: {1003306} -> {1002544}, confidence: 0.13360323886639677
Rule: {1005100} -> {1004856}, confidence: 0.1287793952967525
Rule: {4804055} -> {4804056}, confidence: 0.18552875695732837
Rule: {1004836} -> {1004856}, confidence: 0.10950819672131148
Rule: {1004750} -> {1004856}, confidence: 0.1335282651072125
Rule: {1004870} -> {1004767}, confidence: 0.1415617128463476
Rule: {1004750} -> {1004833}, confidence: 0.1286549707602339
Rule: {1002524} -> {1002544}, confidence: 0.11172796668979873
Rule: {1003317} -> {1005115}, confidence: 0.12437810945273631
Rule: {1004249} -> {1005115}, confidence: 0.11431623931623931
Rule: {1004750} -> {1004767}, confidence: 0.14132553606237816
Rule: {1004873} -> {1004767}, confidence: 0.13194959229058562
Rule: {1004209} -> {1004856}, confidence: 0.23716012084592145
Rule: {1004873} -> {1004870}, confidence: 0.10971089696071164
Rule: {1004858} -> {1004856}, confidence: 0.19969512195121952
Rule: {1004836} -> {1004833}, confidence: 0.11606557377049181
Rule: {1004767} -> {1004856}, confidence: 0.131246967491509
Rule: {1004856} -> {1004767}, confidence: 0.10551979715233079
Rule: {1002633} -> {1002544}, confidence: 0.15055467511885895
Rule: {1004833} -> {1004856}, confidence: 0.1941747572815534
Rule: {1004856} -> {1004833}, confidence: 0.10142383460113126
Rule: {1005105} -> {1005115}, confidence: 0.16027874564459932
Rule: {1004836} -> {1004767}, confidence: 0.1318032786885246
Rule: {1004833} -> {1004767}, confidence: 0.12957430918595966
```

We sort the data for only user session which had purchased more than 1 product.

Then turn the list of products into a set which we used to generate the association rules and found 24 rules with over 0.1 confidence.

The highest confidence is 0.23 where product 1004209 is usually purchased with product 1004856.

Both product being Samsung smart phone

ECONOMIC ANALYSIS

CONCLUSIONS

We found that customers tends to purchase cheaper/more expensive product. Most customers actively view products at around noon on week days. Customers are not particularly loyal to one brand when purchasing a product. Lastly, only around 24 products are bought together in one session and this is rarely the case as most customer prefer to only purchase one products in a session.

IMPROVEMENTS

The data have no specific demographic data therefore the times in which users are most active could be distorted as the customers could be in a different time zone as the data is from an ecommerce site.

Our data is only from a time frame of one month which may not show a clear and truthful trend of the customer's purchase history and only give us a monthly trend.

Most of our data is on customer viewing a product which cause our runtime of a lot of our question to take very long and eat a lot of ram. Therefore, we had to change up our question to focus more on the purchase side of the data which is not the majority of the data we had access to.

WE WANT TO SAY

THANK YOU

FOR YOUR ATTENTION