EDA:ONLINE SHOPPERS PURCHASING INTENTION

FINAL PROJECT | DS5010

BOLAI(BRYAN) YIN

INTRODUCTION

DATASET:

- ~12,330 SESSIONS ON AN E-COMMERCE PLATFORM
- 15.5% (1,908) OF WHICH WERE POSITIVE CLASS SAMPLES ENDING WITH SHOPPING.

CONCERNED VARIABLE: 'REVENUE'

- (WHETHER USER MADE A PURCHASE)

GOAL:

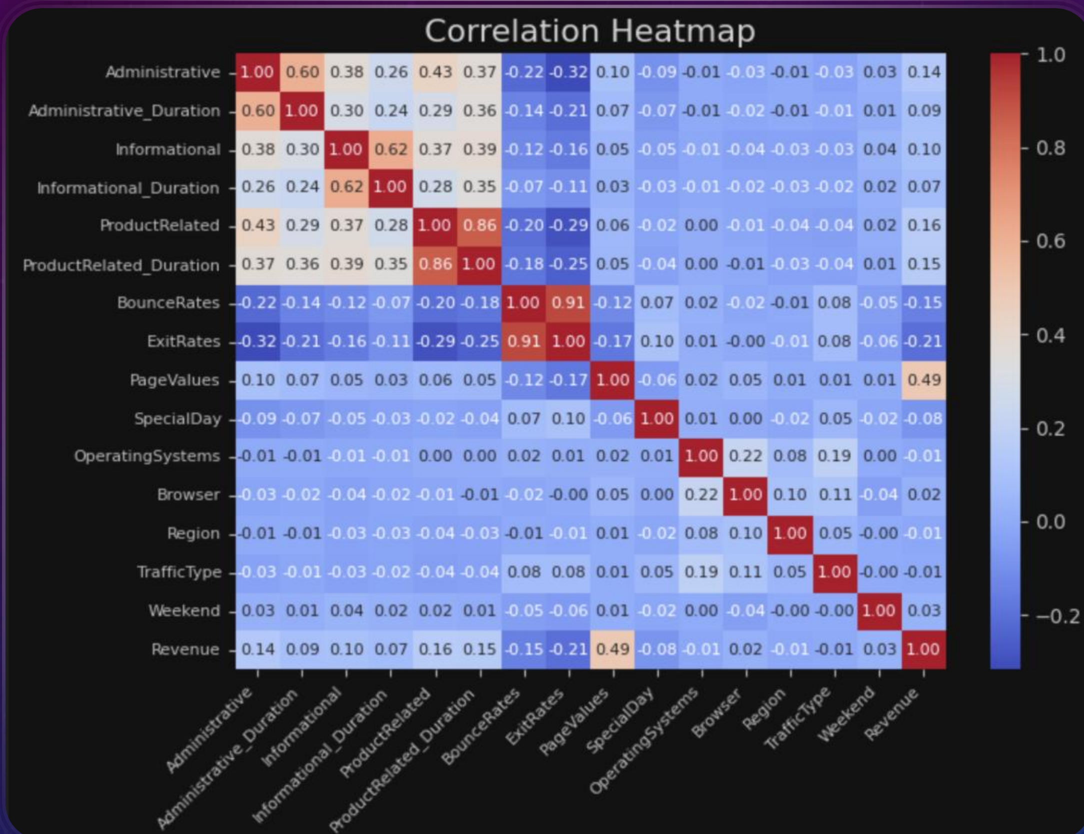
- IDENTIFY KEY PATTERNS ASSOCIATED WITH PURCHASING INTENTION

RAW DATA COLUMNS

- 18 features with two types:
  - ❖ Numerical(Time spent on pages, Bounce & Exit Rates, SpecialDay, etc.)
  - ❖ Categorical(Revenue, Weekend, Region, Month, etc)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12330 entries, 0 to 12329
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Administrative           12330 non-null  int64
 1   Administrative_Duration  12330 non-null  float64
 2   Informational            12330 non-null  int64
 3   Informational_Duration   12330 non-null  float64
 4   ProductRelated           12330 non-null  int64
 5   ProductRelated_Duration  12330 non-null  float64
 6   BounceRates              12330 non-null  float64
 7   ExitRates                12330 non-null  float64
 8   PageValues               12330 non-null  float64
 9   SpecialDay               12330 non-null  float64
 10  Month                    12330 non-null  object
 11  OperatingSystems         12330 non-null  int64
 12  Browser                  12330 non-null  int64
 13  Region                   12330 non-null  int64
 14  TrafficType              12330 non-null  int64
 15  VisitorType              12330 non-null  object
 16  Weekend                  12330 non-null  bool
 17  Revenue                  12330 non-null  bool
dtypes: bool(2), float64(7), int64(7), object(2)
```

CORRELATED FEATURES



Strong Correlations:
- ProductRelated & Duration

Moderate Correlations:
- PageValues
- Administrative & Duration
- Cross Pages: Informational, Administrative & Product

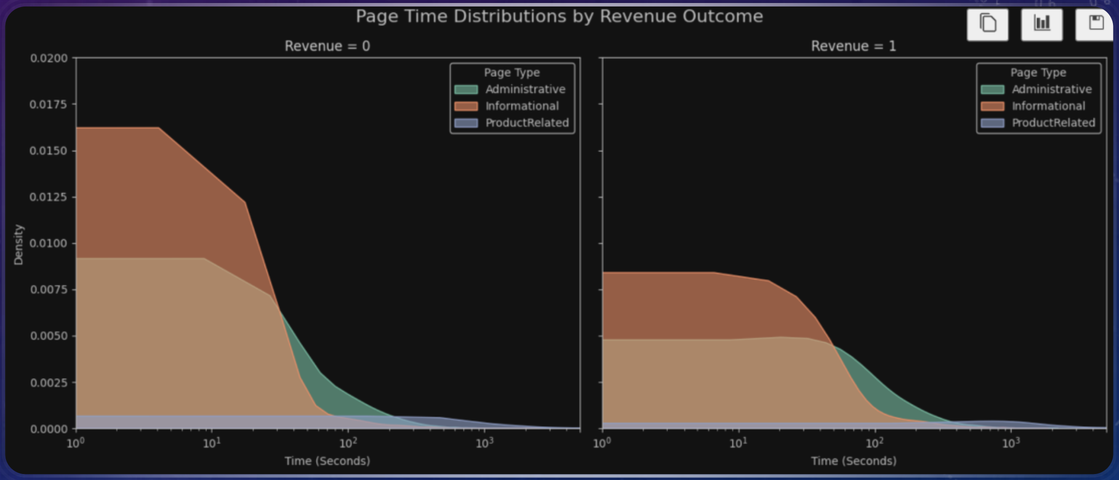
Weak Correlation:
- ExitRates
- BounceRates

AEVERAGE TIME SPENT VS. PURCHASE

- 'Avg. Time’ spent across pages
- Compare time distribution - Insight: Purchasers generally spend more time

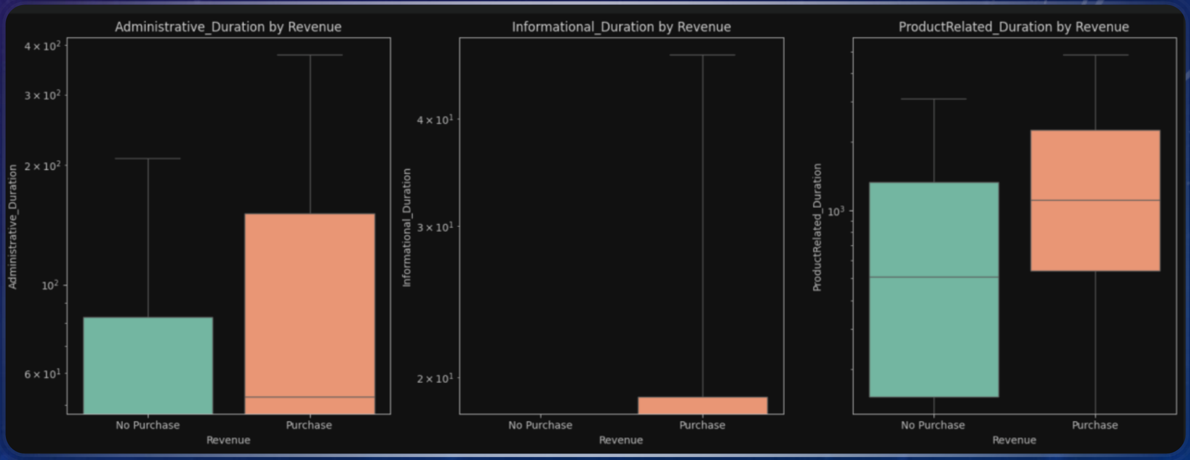
TYPES OF DURATION VS. REVENUE

- Regardless of purchase outcome, spend little time on each page type.
- ProductRelated pages show similar duration distributions across both scenarios
- Administrative and Informational pages have slightly higher densities for non-purchasers.
- The log scale highlights a strong right skew across all page types.



TYPES OF DURATION VS. REVENUE(CONTD.)

- The median line is barely visible or merged with the lower quartile, especially for Informational_Duration, likely due to strong left skewness (many small values)
- ProductRelated_Duration shows a clear difference: median and upper spread are higher for Purchase.
- Whiskers are long due to extreme outliers, which dominate the scale

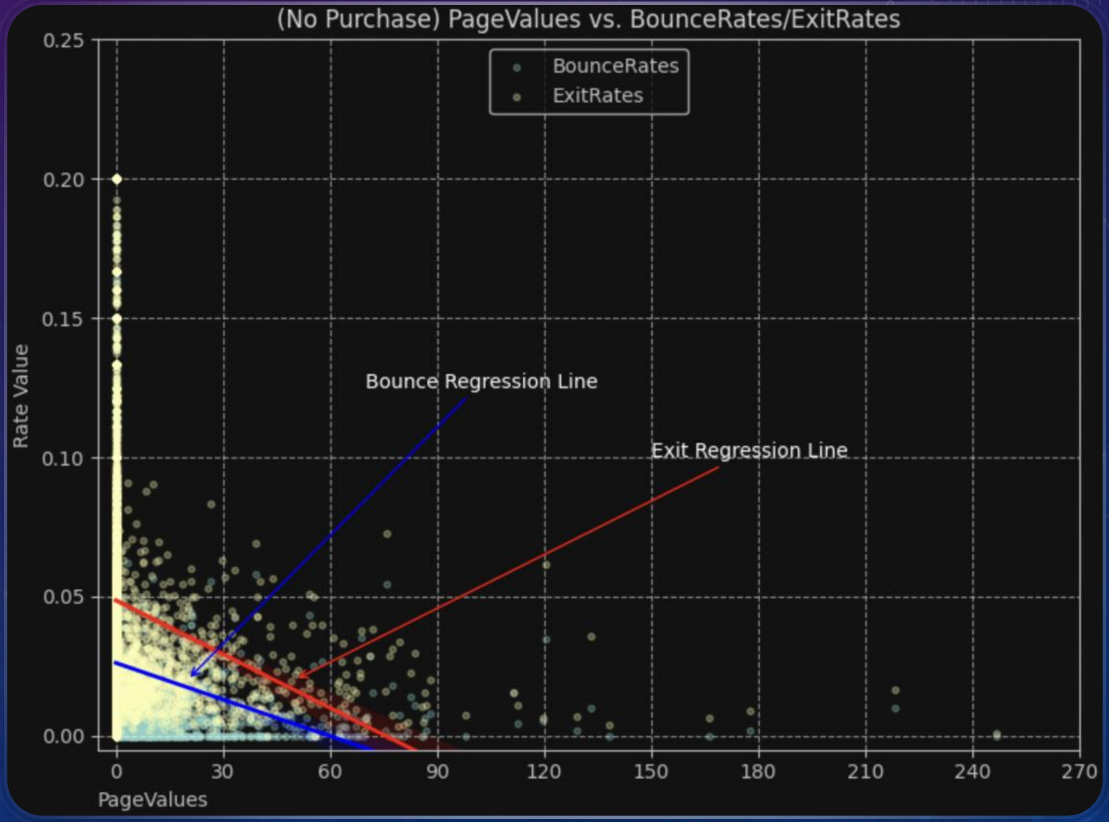


BOUNCE & EXIT RATES VS. REVENUE

- “True” dots are clustered in the bottom-left corner
- Insight: Lower rates associated with higher likelihood of purchasing

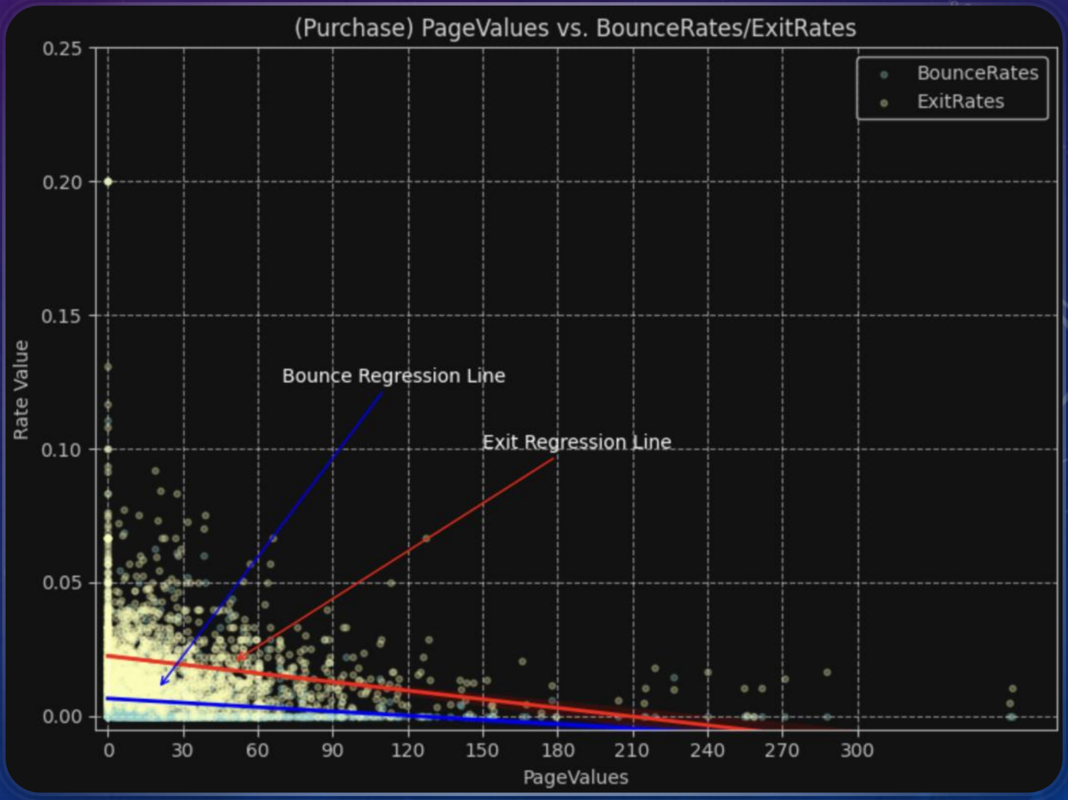
NO REVENUE SAMPLES (BOUNCE & EXIT RATES)

- No Purchase samples are densely clustered in those low PageValues(<30)
- Steeper regression line on exit rate
- Different Threshold:
    - Bounce reduces to zero at lower PageValues, which implies more effectiveness gained even via moderate



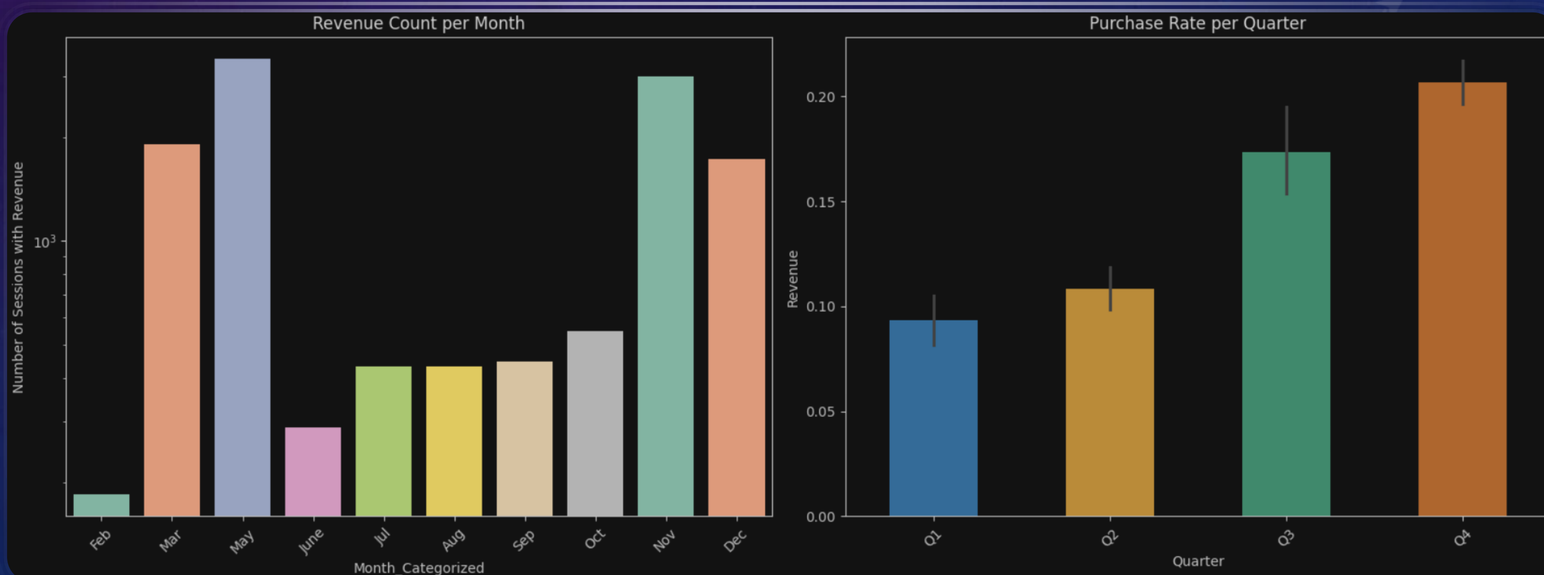
REVENUE SAMPLES (BOUNCE & EXIT RATES VS. PAGEVALUES)

- Flatten two lines compared to the previous scenario.
- Relative steeper for Exit Line when it declines
- High returns for reducing exit rates through page optimization



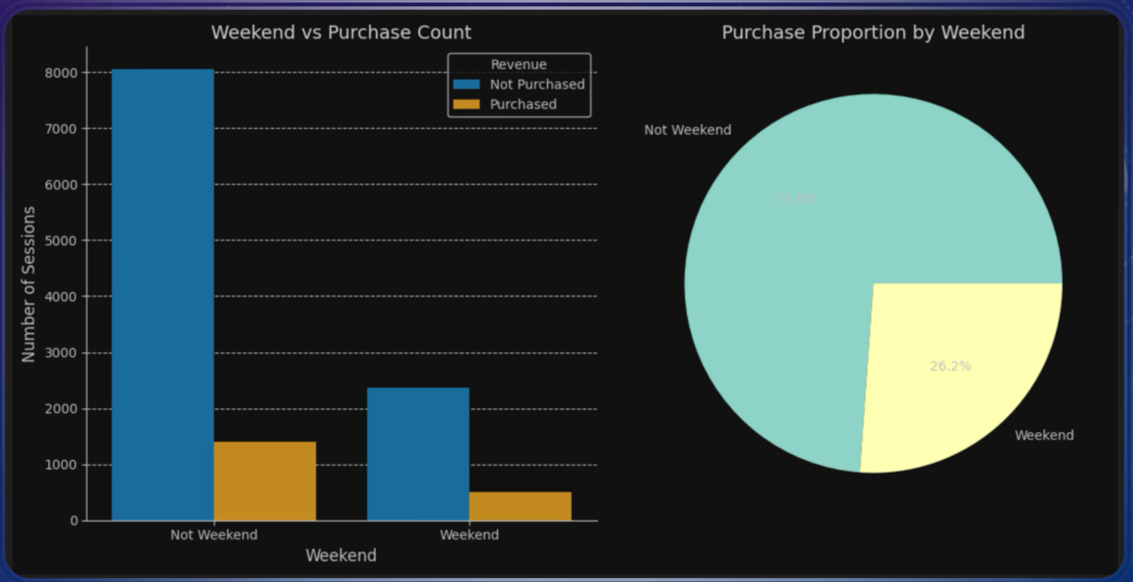
WHAT'S REVENUE BEHAVIOR LOOK LIKE THROUGH A YEAR?

- Revenue ratio increase quarterly
- The spikes in March & May is likely attributed to Spring Break, Mother's Day and Memorial Day
- Hallowean, Thanks-giving, Black Friday and X'mas holiday boost engagement in Q4 as always

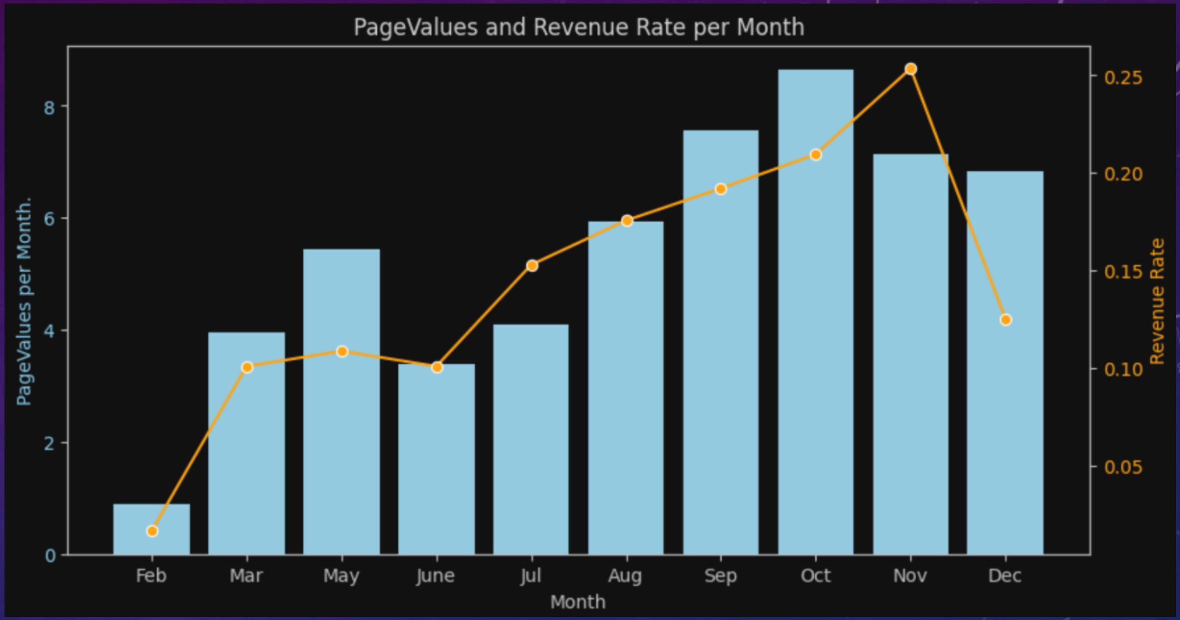


WEEKEND & WEEKDAY?

- People in this dataset are willing to purchase in weekdays rather than weekend.



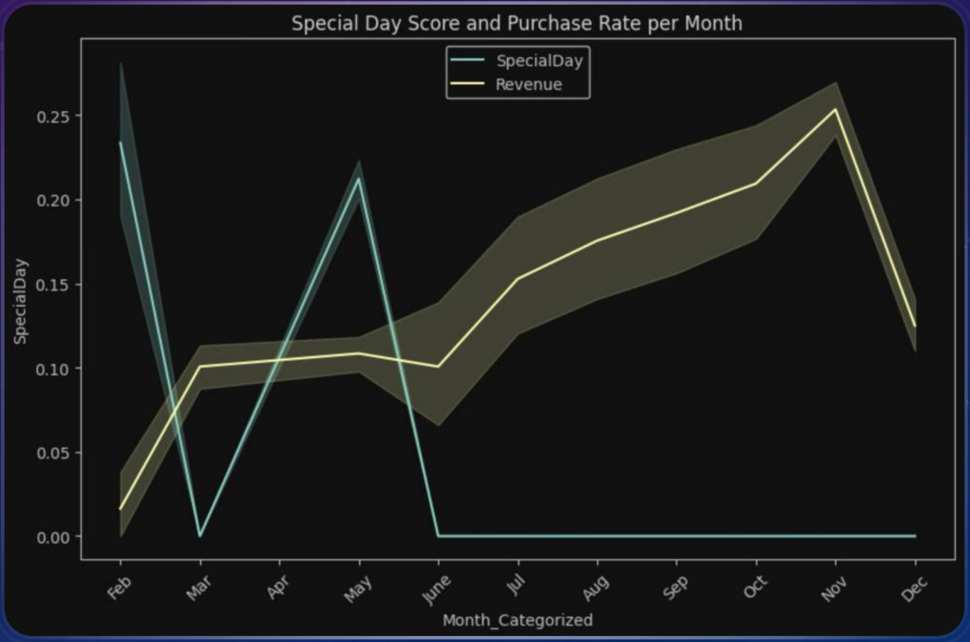
PURCHASE RATE CORRELATES TO PAGEVALUES



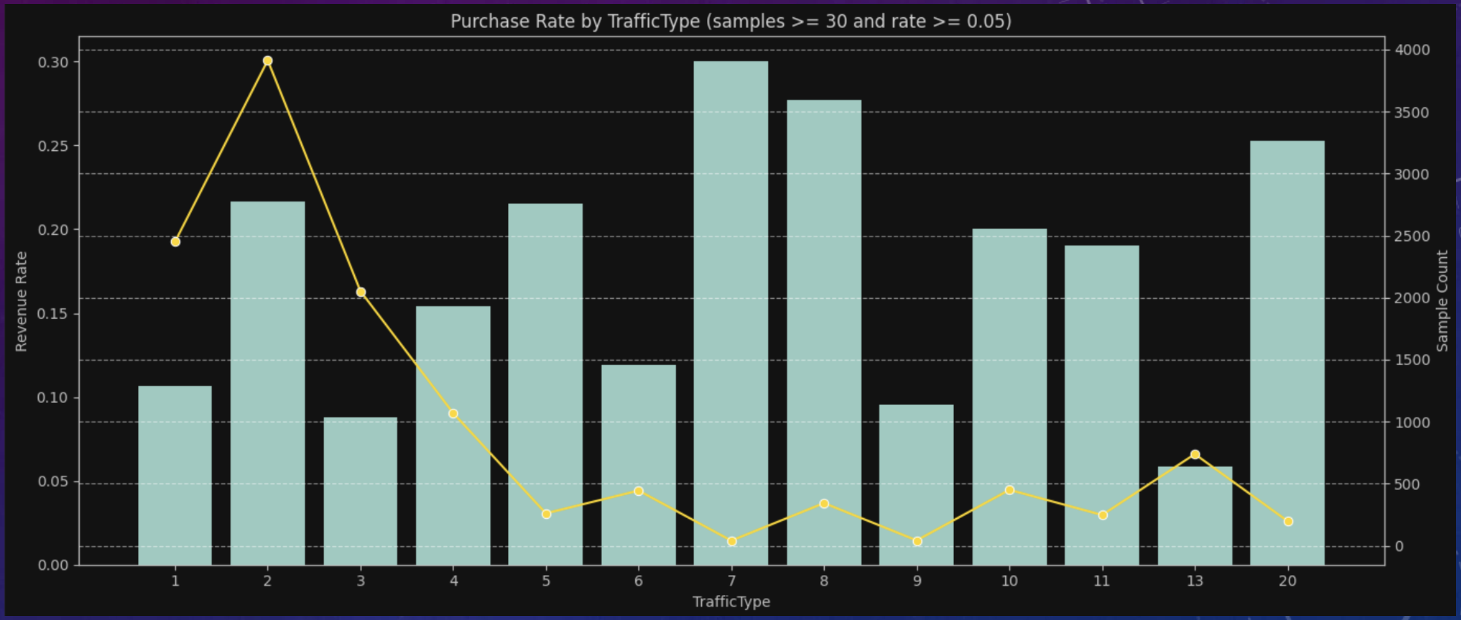
- Revenue rate goes up through months
- Clear positive relationship between PageValues(avg. per month) and Purchase Rate
- Probably another influence factor as revenue ratio drops sharply in Dec

SPECIAL DAY VS REVENUE RATE

- ‘Special Day’ variable doesn’t seems to have strong correlation with revenue ratio

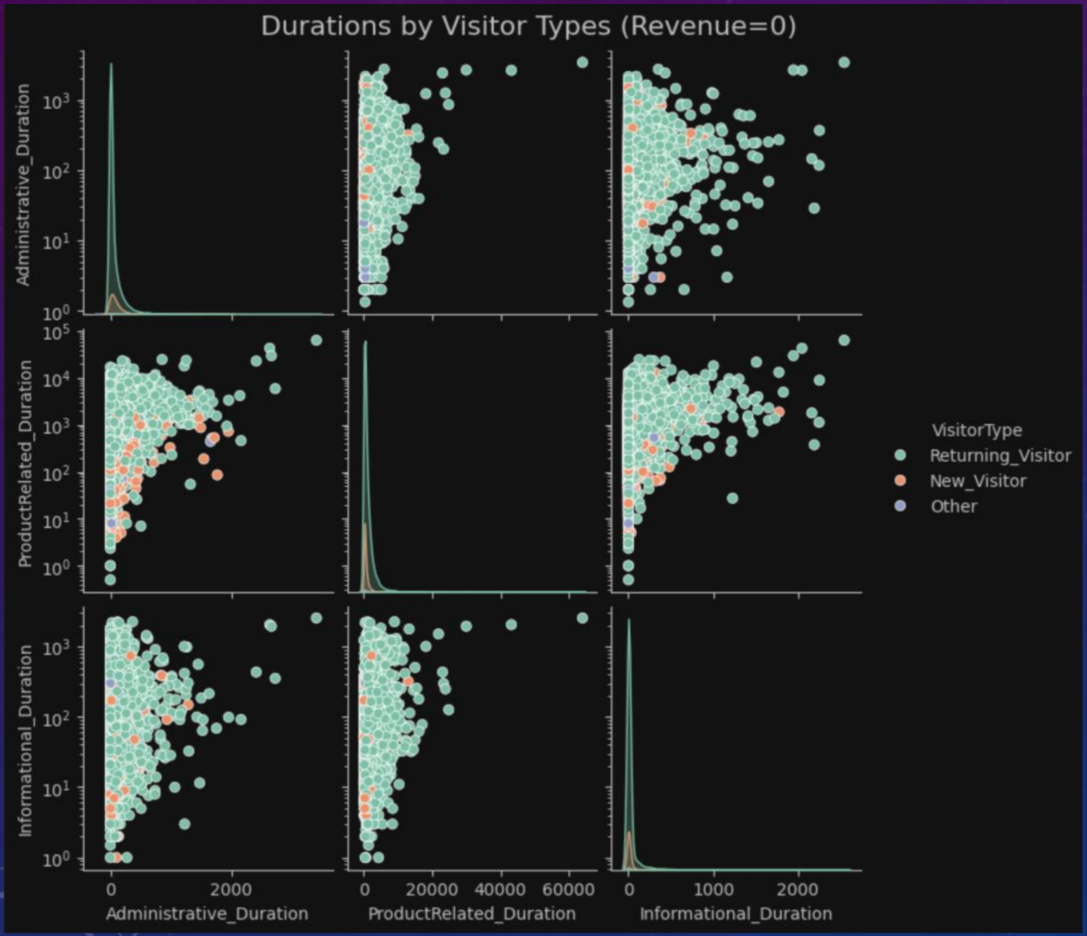


TRAFFIC TYPE CONTRIBUTION



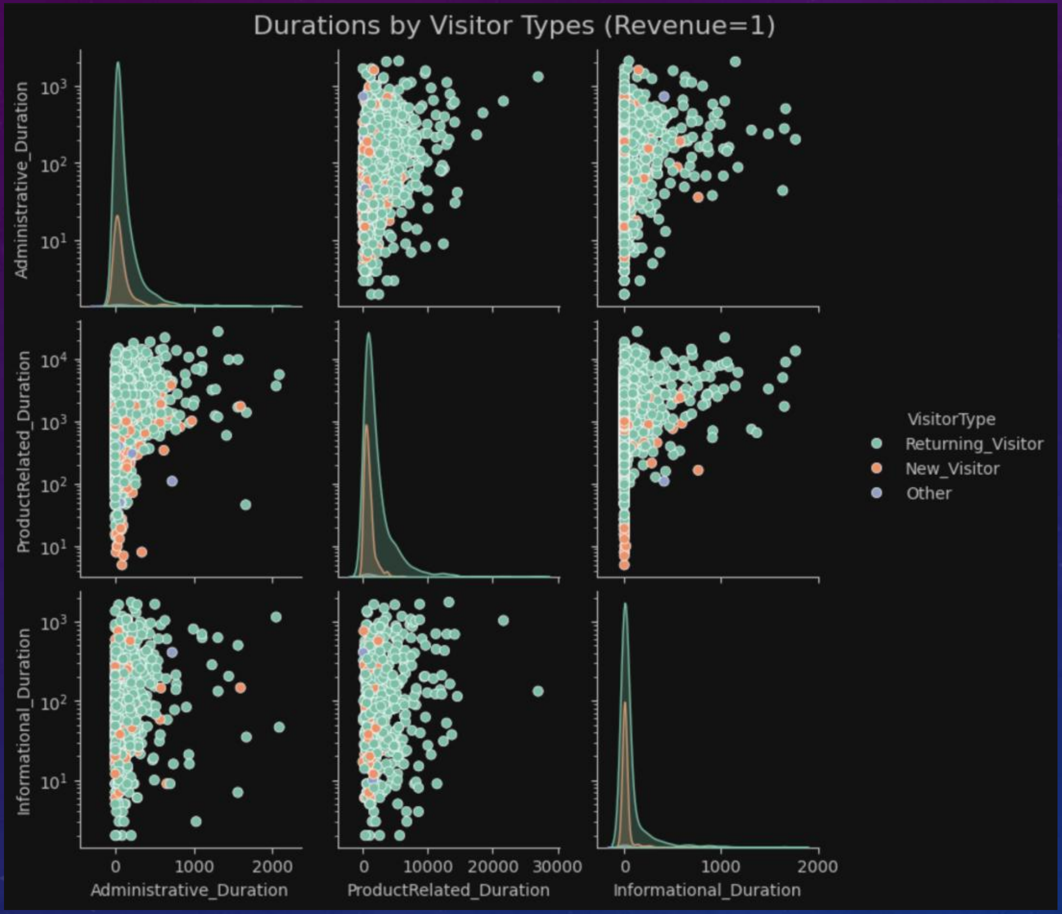
- Traffict Type: 7 has the highest revenue ratio with less number of session(high conversion ratio)
- Second highest is Trafic Type: 8
- Outlier : Extremely small samples (n=1, n=10) are likely noise—prioritize high-volume channels for actionable insights.

DURATION VS. VISTOR TYPES



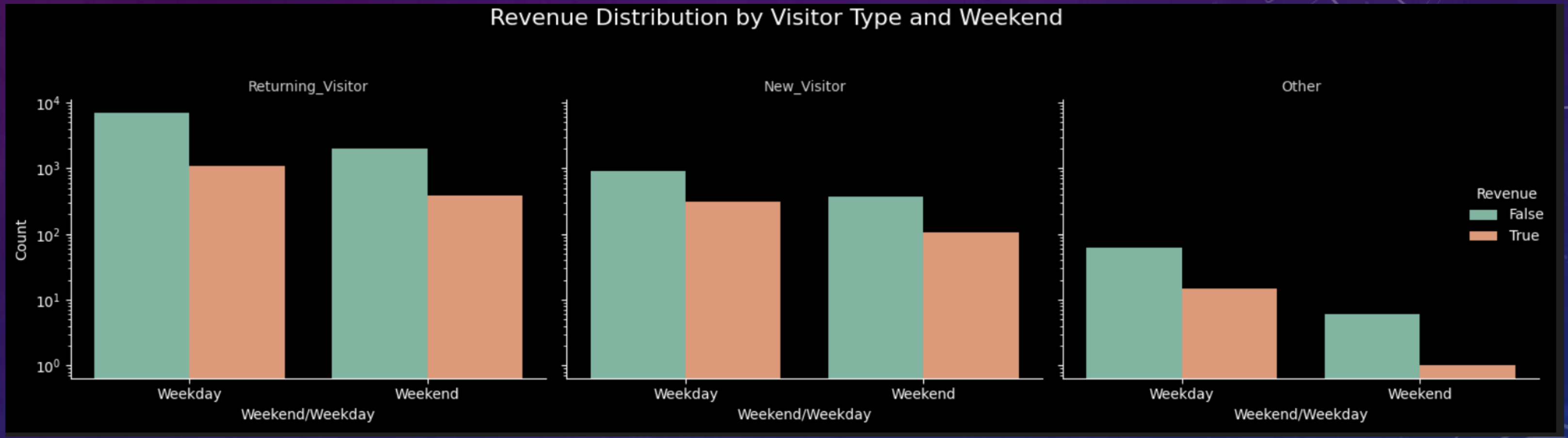
- Majority of samples are returning vistors
- Some new visitors spend more time viewing administrative pages than informational pages when they're viewing product.

DURATION VS. VISTOR TYPES (CON.)



- Majority of samples are returning vistors
- New visitors spent more time on ProductRelated pages

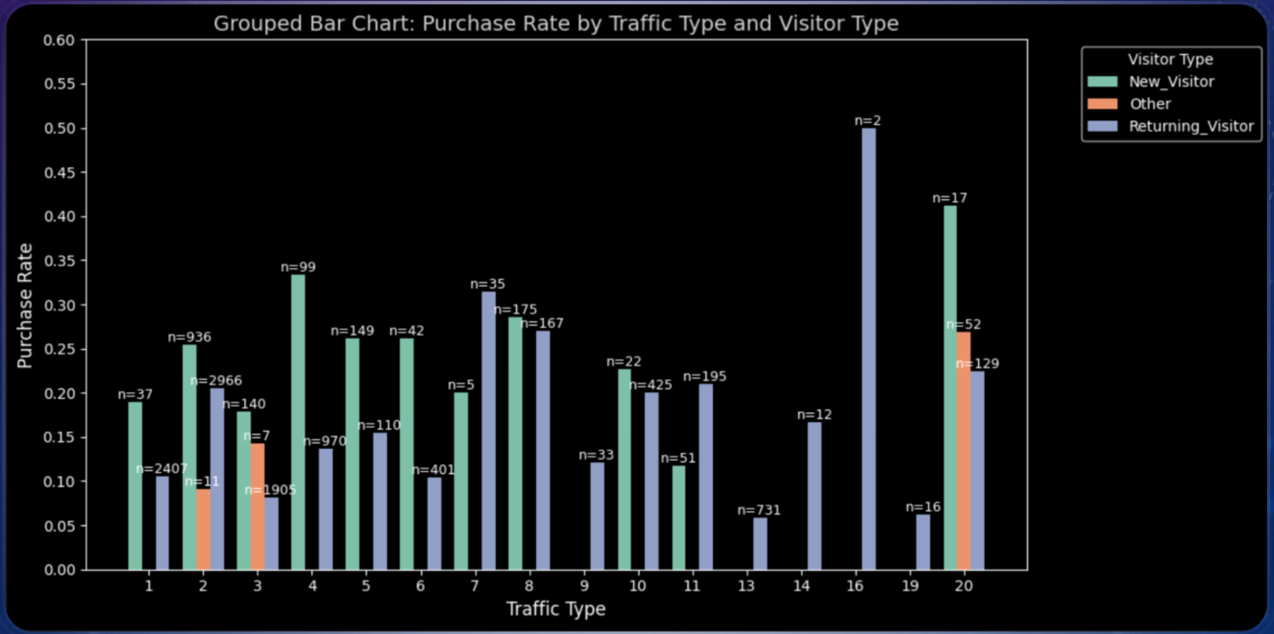
VISITOR TYPES VS. WEEKEND



- No too much behavior difference on this dataset
- Returning visitor has higher number of purchase in weekday; same to the new visitors

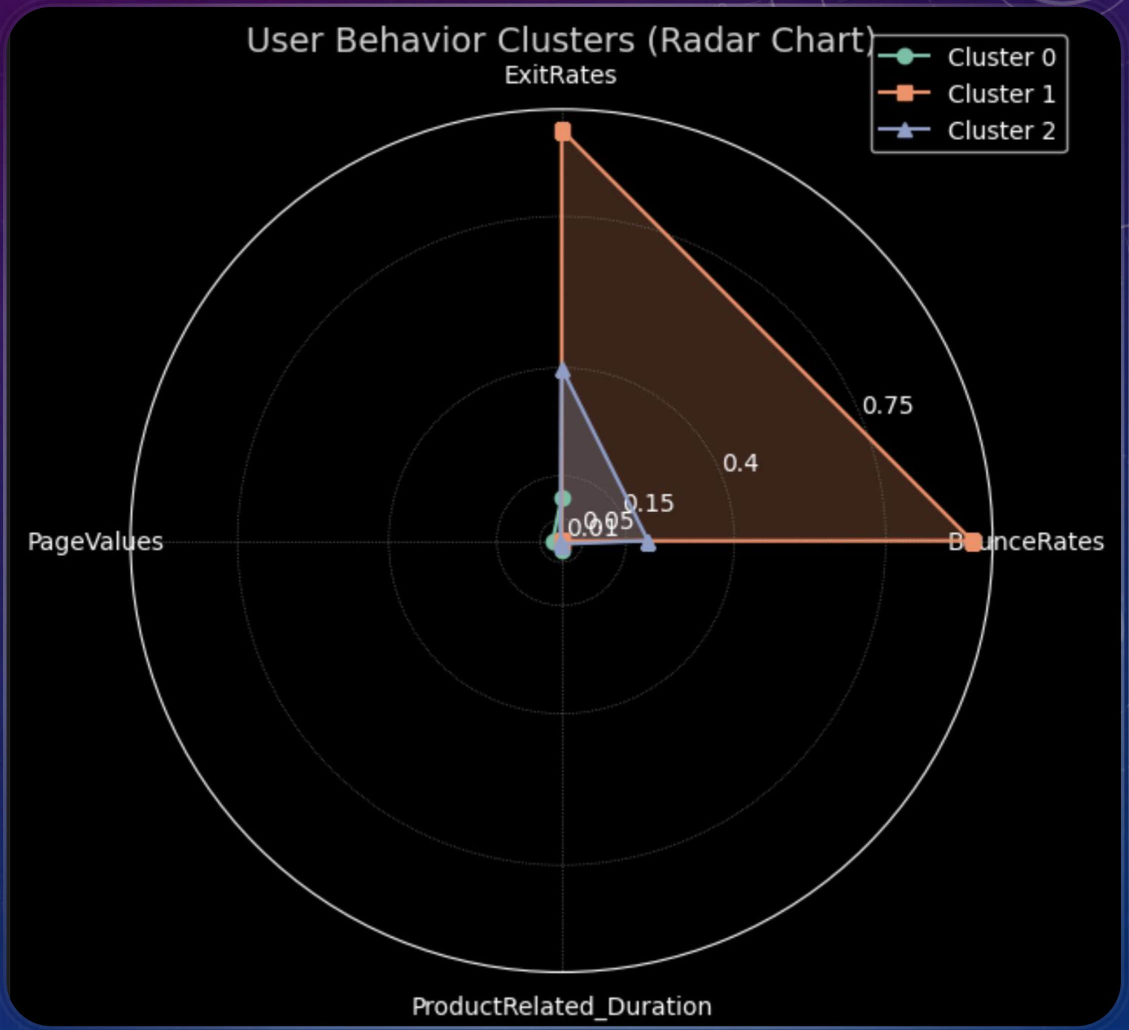
VISITORS VS. TRAFFIC TYPES

- Majority of returning visitors like to use type:2, which also contribute high revene ratio for new visitors.
- Other types likes to use traffic type:20, and some of them uses traffic type 2 and 3.



RADAR CHART SHOWS DISTINCT BEHAVIOR PATTERNS ACROSS 3 CLUSTERS

- Cluster 0:
  - relatively high engaged visitor
  - Spend relatively more time on pages
- Clustor 1:
  - Extremely High Bounce Rates and elevated Exit Rates.
  - PageValues and ProductRelated_Duration near zero.
- Cluster 2:
  - Exit Rates close to Cluster 1, but nuch lower Bounce Rates.
  - Viewing less pages



KEY TAKEAWAYS

- High PageValues and Product Page engagement → Higher likelihood of purchase, which can be identified through visual exploration of feature relationships using Matplotlib and Seaborn. These tools help us identify patterns in user behavior and make data-driven marketing decisions.
- Clustering reveals clear user behavior segments: Using KMeans for clustering, I leveraged unsupervised learning techniques to segment users based on key features like BounceRates, ExitRates, and PageValues. Visualization through Seaborn's pair plots and Matplotlib's radar charts helped visualize these groups and their distinct behaviors.
- Bounce Rates and Exit Rates are critical predictors of early exits, emphasized through bar plots and heatmaps. These visualizations help us understand the correlation between high BounceRates and ExitRates with the likelihood of churn, offering valuable insights for improving user engagement.
- Temporal trend analysis: Line plots from Matplotlib helped identify seasonal patterns in user behavior, aiding in adjusting marketing strategies based on the time of day, week, or promotional cycles.

FUTURE

- In this project, KMeans clustering was used to segment users based on their behaviors, and further applied Matplotlib and Seaborn to visualize these segments and understand their characteristics, which are crucial for predictive modeling.
- For future steps, integrating supervised models (such as logistic regression or random forests) can predict conversion likelihoods, while leveraging time-series analysis techniques (e.g., ARIMA) could forecast future user behaviors. Using Seaborn's advanced visualizations and Matplotlib's flexibility, I aim to build detailed, informative graphics to aid in the optimization of marketing and product strategies.

THANK YOU

Questions