The Data Analysis Report

Fatma Beyza Çepni
Computer Engineering of 3rd Year
University of Galatasaray
Turkey,Istanbul
Email: f.beyza0216@gmail.com

Abstract: This document investigates the one-year visitor data of the Coulombia.com.tr website, aiming to explore the factors influencing the conversion of site visitors into shoppers. The analysis is conducted to comprehend the site's performance and assess customer behaviors.

I. Introduction

A. Overview

In today's world, our shopping habits have significantly evolved with the opportunities provided by the internet. Online shopping has become central to our lives due to its advantages such as a diverse range of products, ease of use, and accessibility from anywhere at any time. We turn to online shopping platforms to research products, discover new trends, or simply pass the time. The time spent on online shopping applications is now competing with social media usage in our screen time. Particularly under global pandemic conditions, the secure and fast solutions offered by online shopping have further popularized this habit. People are turning to online platforms to meet their needs and experience safe shopping without leaving their homes. In this context, online shopping sites have become a crucial factor shaping the shopping experience of today's consumers. Each of us, by occasionally visiting these sites, define our shopping preferences, expectations and demands.

B. Motivation

At times, we all become visitors to e-commerce websites. This research aims to examine the factors that influence the completion of an online shopping experience and the elements that play a role in determining the intention of a visitor to an e-commerce site. Understanding customer behaviors on online shopping platforms plays a critical role in businesses gaining a competitive advantage. This analysis aims to delve deeper into understanding user shopping intentions and guide businesses in making strategic decisions. Understanding customer behaviors can provide valuable insights for businesses to optimize marketing strategies and improve service quality.

C. Research Questions

1) High-Level Research Question: In this research, we aim to identify the elements that businesses providing online shopping services should focus on in their marketing strategies to be more successful in the industry. Increasing the likelihood of website visits resulting in purchases is a crucial factor influencing the positions of businesses within the sector. In this context, our research aims to answer the following question: "What are the key factors that influence user intention to result in a purchase during a visit to an online shopping site?".

2) Low-Level Technical Questions: In the dataset used during the research, there are many parameters that could influence the intentions of visitors. Therefore, it is important to ask the following questions:

- What is the impact of the date of the visit to the website on customer behavior?
- Which types of pages are more influential in positively affecting the shopping intentions of visitors who spend more time on them?
- Do new visitors or returning visitors have a higher potential to convert their website visits into purchases?

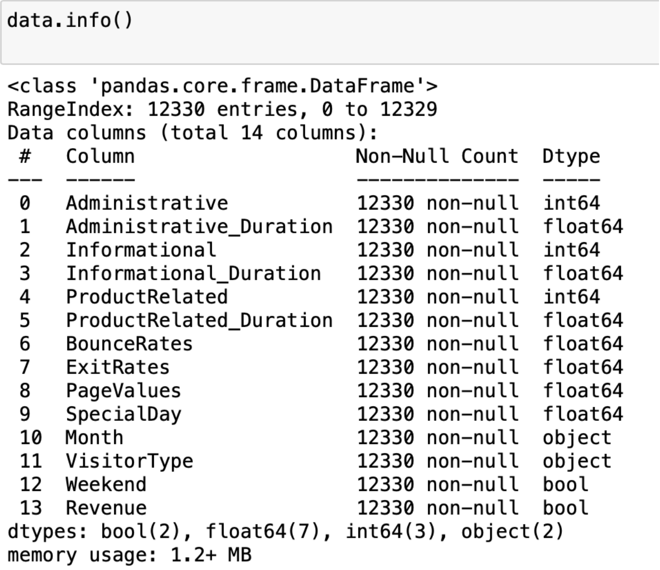
These questions aim to provide important information for understanding the online shopping experience in detail and assisting businesses in optimizing their marketing strategies.

II. Method

A. Dataset

1) Story/Overview: The dataset used in the research is obtained from "coulombia.com.tr" and consists of feature vectors from 12,330 sessions. The dataset is created so that each session corresponds to a unique user, avoiding trends specific to a certain campaign, special day, user profile, or period. This approach is implemented to avoid biases related to specific campaigns, special days, user profiles, or periods.

2) Attributes: First of all, what we need for analysis is a detailed introduction of data. It is a data set with a sample size of approximately 12330 and has 14 main columns. The dataset consists of 10 numerical and 4 categorical attributes. The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for each page in the e-commerce site.

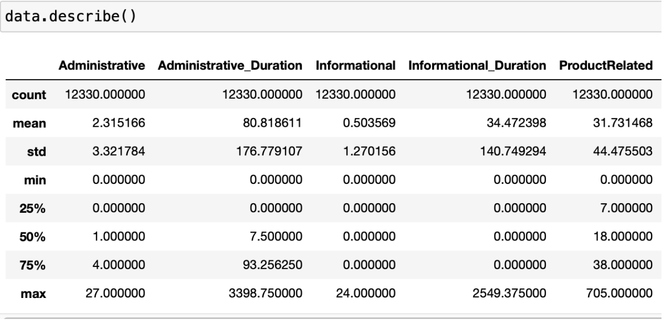


```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12330 entries, 0 to 12329
Data columns (total 14 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  VisitorType              12330 non-null  object 
 12  Weekend                  12330 non-null  bool   
 13  Revenue                  12330 non-null  bool   
dtypes: bool(2), float64(7), int64(3), object(2)
memory usage: 1.2+ MB
```

Fig. 1. Information about data

- *Administrative:* Represents the number of pages visited by the visitor in the ”Administrative” category (e.g. sign in page, my account page).Numerical.
- *Administrative Duration:* Total time spent by the visitor on pages in the ”Administrative” category (second).Numerical.
- *Informational:* Indicates the number of pages visited by the visitor in the ”Informational” category.Numerical.
- *Informational Duration:* Total time spent by the visitor on pages in the ”Informational” category (second).Numerical.
- *Product Related:* Represents the number of pages visited by the visitor in the ”Product Related” category. Numerical.
- *Product Related Duration:* Total time spent by the visitor on pages in the ”Product Related” category (second).Numerical.
- *Bounce Rate:* Percentage of visitors who enter the site from a page and then leave without further action.Numerical.
- *Exit Rate:* Percentage of all page views for a specific page that were the last in the session.Numerical.
- *Page Value:* Average value of a page that a user visited before completing an e-commerce transaction.Numerical.
- *Special Day:* Indicates the proximity of the site visit time to a specific special day.Numerical.
- *Visitor Type:* Indicates whether the visitor is a returning or new visitor.Categorical.
- *Weekend:* Boolean value indicating whether the date of the visit is a weekend.Categorical.
- *Month:*The month of the year in which the visit occurred.Categorical.



```
data.describe()
```

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated
count	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000
mean	2.315166	80.818611	0.503569	34.472398	31.731468
std	3.321784	176.779107	1.270156	140.749294	44.475503
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	7.000000
50%	1.000000	7.500000	0.000000	0.000000	18.000000
75%	4.000000	93.256250	0.000000	0.000000	38.000000
max	27.000000	3398.750000	24.000000	2549.375000	705.000000

Fig. 2. Data Descriptions (part 1)

- *Revenue:*The class label indicating whether the session resulted in a purchase or not.Categorical.

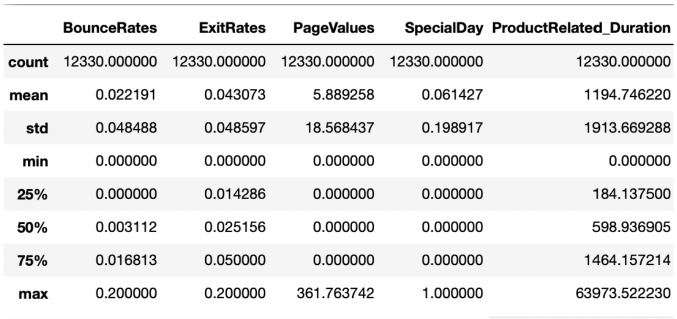
B. Method for Answering Research Questions

- H0: There is no difference in shopping rates between weekdays and weekends.
  HA: There is a significant difference in shopping rates between weekdays and weekends.
  Method: We will use an independent two-sample t-test to the difference between the two groups.We will also examine the percentage and numerical results of visitors' shopping behavior on weekdays and weekends.
- H0: There is no difference in shopping rates between months.
  HA: Specific months exhibit significantly different shopping rates.
  Method: We will employ an ANOVA test to assess differences in shopping rates between months.
- H0: There is no significant correlation between the time spent on different site types.
  HA: The time spent on specific site types correlates with user behavior.
  Method: The relationship between them will be examined by using correlation statistics.
- H0: There is no relationship between visitor type and the likelihood of revenue.
  HA: Visitor types influences the likelihood of revenue.
  Method: Evaluate the relationship between Visitor Type and Purchase using the Kendall Tau-b correlation coefficient.

III. Descriptive Analysis

A. Administrative

The analysis results indicate that the majority of visitors did not visit pages in the ”Administrative” category. However, the median value of 1 suggests that at least one visitor has visited a page in this category. With a mean value of 2.32, it can be inferred that, on average, a visitor visited around 2 pages in the ”Administrative” category.



	BounceRates	ExitRates	PageValues	SpecialDay	ProductRelated_Duration
count	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000
mean	0.022191	0.043073	5.889258	0.061427	1194.746220
std	0.048488	0.048597	18.568437	0.198917	1913.669288
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.014286	0.000000	0.000000	184.137500
50%	0.003112	0.025156	0.000000	0.000000	598.936905
75%	0.016813	0.050000	0.000000	0.000000	1464.157214
max	0.200000	0.200000	361.763742	1.000000	63973.522230

Fig. 3. Data Descriptions (part 2)

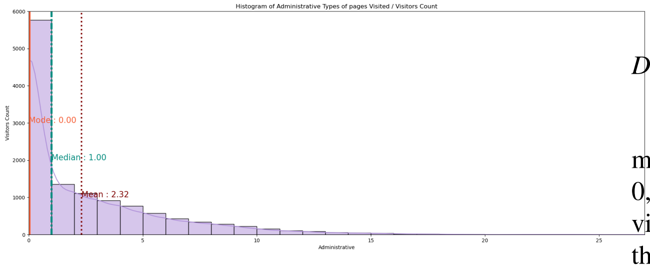


Fig. 4. The Histogram of Administrative Types of pages Visited / Visitors Count

B. Administrative Duration

Upon examining the data, it is observed that the majority of visitors spend either no time or very little time on pages in the ”Administrative” category. The median value of 7.5 seconds indicates that half of the visitors spend less than 7.5 seconds, while the other half spends more time. The average duration for all visitors is 80.82 seconds. Additionally, the data exhibits a positively skewed distribution, implying that, in general, most visitors spend short durations on the pages, but a few visitors spend longer durations.

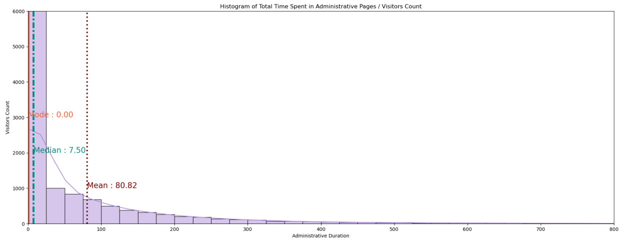


Fig. 5. The Histogram of Total Time Spent in Administrative Pages

C. Informational

Based on the assessments, it is evident that the majority of visitors either do not visit or spend very little time on pages in the ”Informational” category. The median value of 0 indicates that half of the visitors do not visit ”Informational” pages.This information highlights that pages in the ”Informational” category are generally not preferred by visitors, and visits to pages in this category are low.

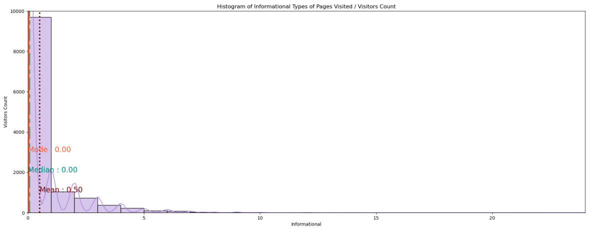


Fig. 6. The Histogram of Informational Types of Pages Visited / Visitors Count

D. Informational Duration

The total time spent by visitors on pages in the ”Informational” category is analyzed as follows; mode: 0, median: 0, mean: 34.47s. This analysis indicates that the majority of visitors either spend no time or very little time on pages in the ”Informational” category. The mode and median values of 0 suggest that a significant portion of visitors did not spend any time on these pages. However, the mean value of 34.47 indicates that there are instances where visitors spent a relatively longer time on ”Informational” pages, contributing to the higher average.

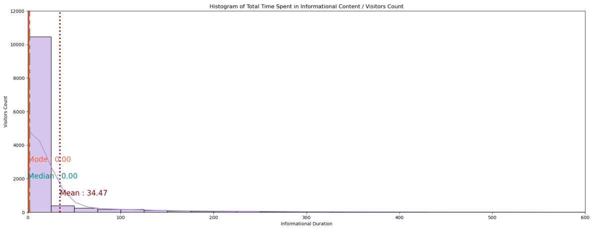


Fig. 7. The Histogram of Total Time Spent in Informational Pages

E. Product Related

The statistics derived from the analysis of visitor page visits in the ”Product Related” category are as follows; mode: 1, median: 18, mean: 31.73. These assessments reveal varying levels of interaction among visitors with pages in this category. The mode value of 1 represents the most frequently observed specific visit count in the dataset. A median value of 18 indicates that half of the visitors explored fewer than 18 pages in the ”Product Related” category, while the other half visited more pages. The mean value of 31.73 signifies the average number of page visits in this category, indicating a moderate level of engagement. These statistics demonstrate that pages in the ”Product Related” category trigger diverse visitor behaviors

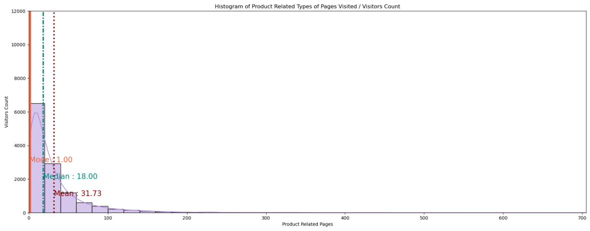


Fig. 8. The Histogram of Product Related Types of Pages Visited / Visitors Count

F. Product Related Duration

According to data, the outcomes of this analysis are as follows; mode: 0, median: 598.94 mean: 1194.75. A mode value of 0 indicates the most frequently observed specific duration in the dataset. The calculated median of 598.94 seconds represents the median value of the time visitors spend on ”Product Related” pages. Additionally, the mean of 1194.75 seconds suggests that outliers or significantly longer durations might impact the overall mean.The dataset shows a positively skewed distribution, indicating that it is mostly concentrated on lower values but extends to the right with a few higher values. This implies that the time spent on product related pages is generally short, but some visitors spend significantly longer periods.

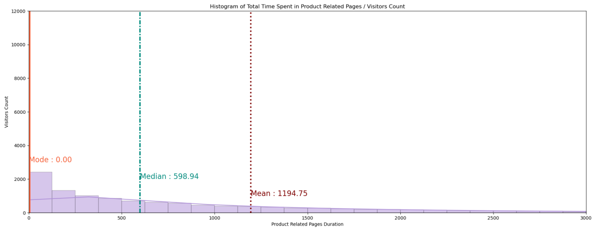


Fig. 9. The Histogram of Total Time Spent in Product Related Pages / Visitors Count

G. Bounce Rates

The value of ”Bounce Rate” feature for a web page refers to the percentage of visitors who enter the site from that page and then leave (”bounce”) without triggering any other requests to the analytics server during that session(mean:0.00, median: 0.003, mean: 0.02). The mode value of 0 suggests that a significant portion of the web pages has a ”Bounce Rate” of 0, indicating that many visitors do not leave the site immediately after entering from those pages. The median value of 0.003 provides the middle point of the data, suggesting that half of the pages have a ”Bounce Rate” less than or equal to 0.003. The mean value of 0.02 represents the average ”Bounce Rate” across all pages.This analysis indicates that, on average, the web pages in the dataset have a relatively low ”Bounce Rate,” with the majority having a ”Bounce Rate” close to zero.

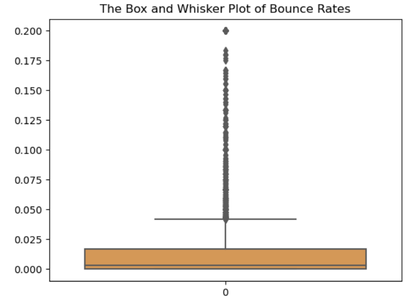


Fig. 10. The Box and Whisker Plot of Bounce Rates

H. Exit Rates

The value of ”Exit Rate” feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session.The mode value of 0.2 indicates that there is a significant frequency of pages with an ”Exit Rate” of 0.2, signifying that a considerable portion of sessions concludes with the last view on those pages. The median value of 0.025 represents the middle point of the data, suggesting that half of the pages have an ”Exit Rate” less than or equal to 0.025. The mean value of 0.04 denotes the average ”Exit Rate” across all pages.This analysis implies that, on average, the web pages in the dataset have a moderate ”Exit Rate,” with a substantial number of pages having an ”Exit Rate” around 0.2.

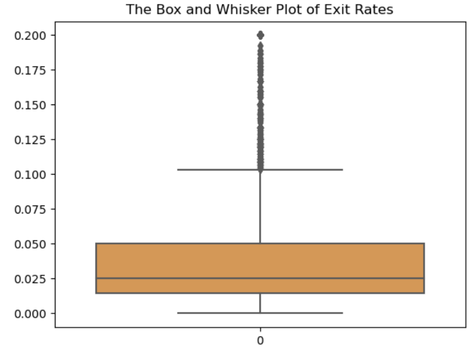


Fig. 11. The Box and Whisker Plot of Exit Rates

I. Page Value

The ”Page Value” feature signifies the average value assigned to a web page that a user visited before finalizing an e-commerce transaction.(mode: 0, median: 0, Mean: 5.89)The mode value of 0 suggests that there is a prevalent frequency of pages with a ”Page Value” of 0, indicating that a considerable portion of web pages might not directly contribute to the e-commerce transaction’s value. The median value of 0 implies that half of the pages have a page value less than or equal to 0. The mean value of 5.89 represents the average page value across all pages.

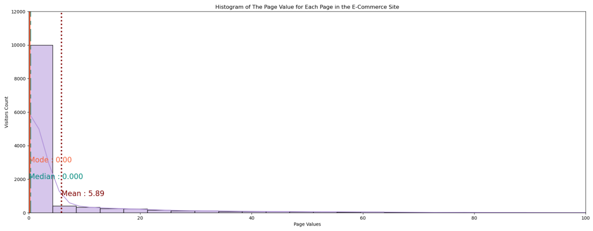


Fig. 12. The Histogram of The Page Value for Each Page in the E-Commerce Site

J. Special Day

The ”Special Day” feature reflects the closeness of a site visit to specific occasions (e.g., Mother’s Day, Valentine’s Day) when transactions are more likely to take place. The attribute value is determined by considering e-commerce dynamics such as the duration between order and delivery dates. For example, on Valentine’s Day, the value is nonzero from February 2 to February 12, zero before and after unless close to another special day, reaching its maximum value of 1 on February 8.(Mode: 0, Median: 0, Mean: 0.06) The mode value of 0 indicates that there is a prevalent frequency of instances where the special day feature has a value of 0, suggesting that a substantial portion of site visits might not be closely associated with specific occasions. The median value of 0 implies that half of the instances have a special day value less than or equal to 0. The mean value of 0.06 represents the average ”Special Day” value across all instances.This analysis suggests that, on average, site visits in the dataset are not strongly correlated with special occasions, with a substantial number having a special day value of 0.

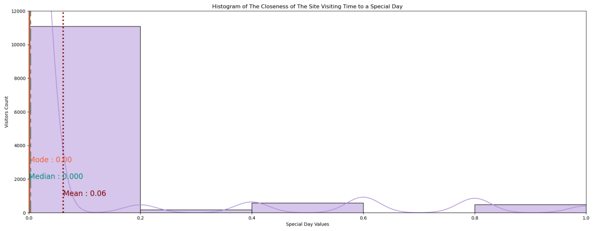


Fig. 13. The Histogram of The Closeness of The Site Visiting Time to a Special Day

K. Visitor Type

Among the 12,330 visitors in the dataset, it is observed that more than 10,000 fall into the category of returning Visitor. This suggests that the site tends to attract users who have visited before, forming a loyal customer base. The remaining visitors are categorized as new visitor, indicating those who are exploring the site for the first time.

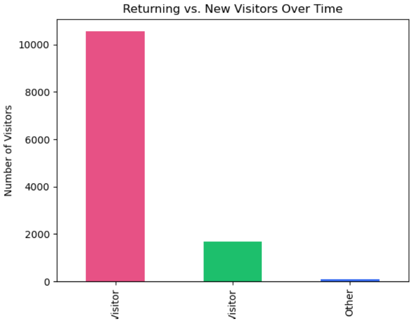


Fig. 14. Returning Visitors vs. New Visitors

L. Weekend

The attribute indicating whether the date of the visit falls on a weekend is a binary value. Out of the total visits, approximately %23.3 are on weekends, while the majority, accounting for %76.7, occur on weekdays.This insight provides an understanding of the distribution of site visits across different days of the week. The fact that a significant portion of visits occurs on weekdays suggests that users frequently engage with the site as part of their regular weekday activities.

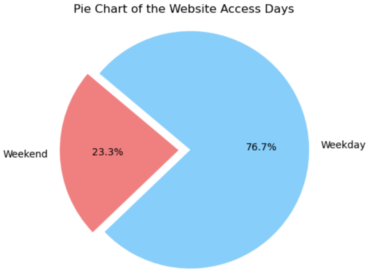


Fig. 15. The Pie Chart of The Website Access Days

M. Month

This attribute indicates the month in which the visit occurred. The site experiences its peak visits in May and November.

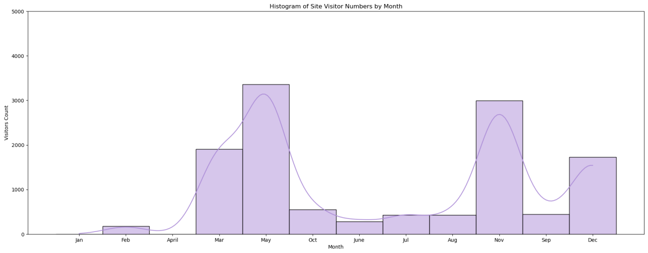


Fig. 16. The Histogram of Site Visitor Numbers by Month

N. Revenue

The class label represents whether the session resulted in a purchase. Approximately &15.5 of sessions are labeled as ”False”, indicating no purchase, while the majority, around %84.5, are labeled as ”True”, denoting a successful purchase.This information is crucial for understanding the conversion rate and the success of the site in turning visits into actual transactions.The high percentage of ”True” labels suggests that a significant portion of the sessions ends with a successful purchase

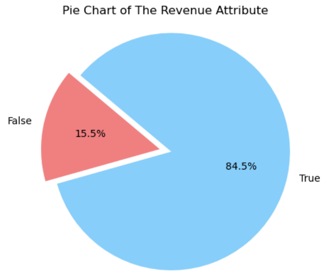


Fig. 17. The Pie Chart of The Revenue Attribute

IV. ANALYSIS FOR ANSWERING RESEARCH QUESTIONS

A. First Hypothesis

H0: There is no difference in shopping rates between weekdays and weekends.
HA: There is a significant difference in shopping rates between weekdays and weekends.

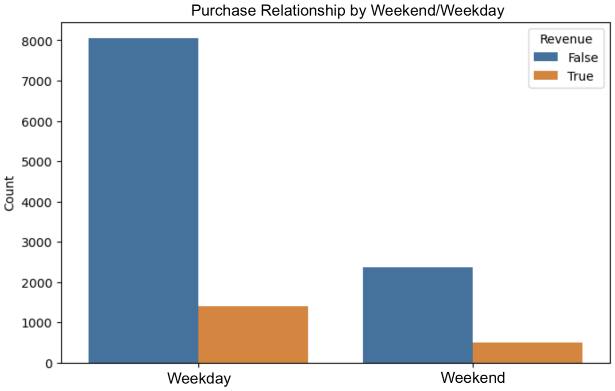
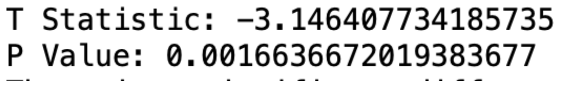


Fig. 18. The Bar Chart of The Revenue&Day Relationship

For weekday shopping data, we observed that around 8000 individuals did not engage in shopping, while approximately 1600 did. On weekends, we noted that about 2500 did not shop, whereas 500 did.



```
T Statistic: -3.146407734185735
P Value: 0.0016636672019383677
```

Fig. 19. The Values of The T-Test

In this analysis, we examined the shopping behaviors on weekdays and weekends, utilizing an independent two-sample t-test to investigate whether a significant difference exists between the two groups.
According to the t-test results, there is a significant difference between weekday and weekend shopping behaviors (T-Statistic: -3.1464, P-Value: 0.00166). This finding indicates a statistically significant distinction in shopping behaviors between weekdays and weekends. However, the similarities observed in the graphs might suggest otherwise.

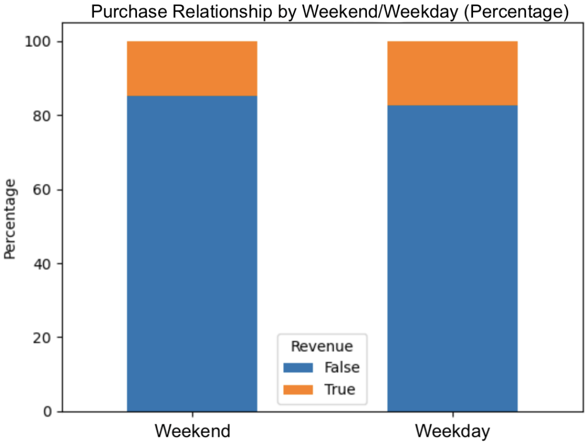


Fig. 20. The Bar Chart of Percentage Relationship Between Revenue&Day

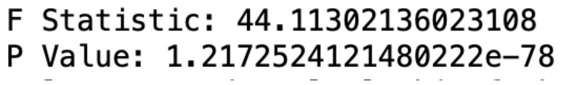
The fundamental reason for this contradiction could lie in the differences in the distributions within the dataset. The t-test can identify significant differences in distributions, even when graphical representations appear similar.

Additionally, when examined as percentages, we found that the rates of non-shopping individuals on weekdays and weekends are quite similar. For instance, the percentage of individuals not shopping on weekdays might be around 80%, while on weekends, it could be approximately 83%.

In conclusion, evaluating numerical and percentage values together can help elucidate the contradiction between analysis results.According to the result we obtained by examining the income percentage bar chart, we do not reject the H0 hypothesis.

B. Second Hypothesis

H0: There is no difference in shopping rates between months.
HA: Specific months exhibit significantly different shopping rates.



```
F Statistic: 44.11302136023108
P Value: 1.2172524121480222e-78
```

Fig. 21. The Values of The Anova Test

In this analysis, we focused on the shopping behaviors across months, evaluating the shopping conversion rates for each month. The F-statistic, indicating a significant difference between months in terms of shopping rates, is notably high (F-statistic: 44.11). Additionally, the p-value is very small (P-value: 1.21725e-78), suggesting that there is a significant impact of months on shopping.

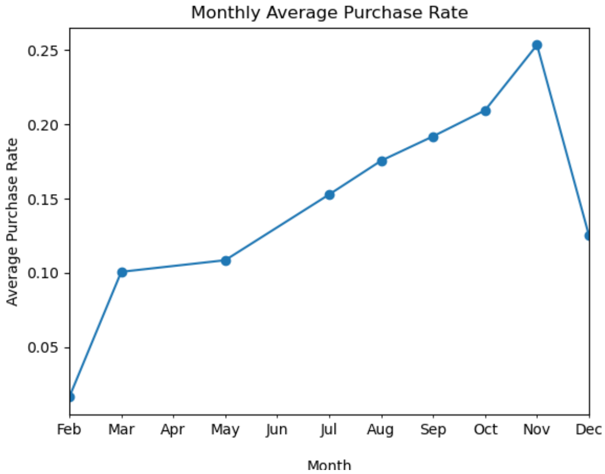
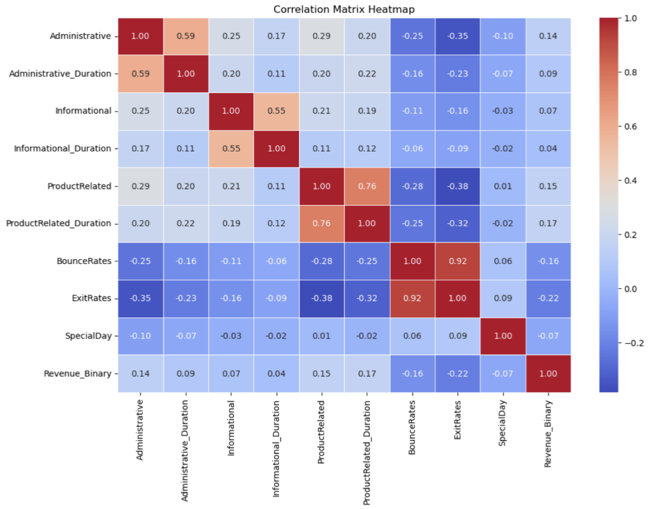


Fig. 22. The Line Chart of Percentage Relationship Between Revenue&Month

Furthermore, when examining the line chart, the elevated shopping conversion rate in November highlights the significance of this month in terms of shopping. The probability of online shopping in November is noticeably higher compared to other months.
We reject the H0 hypothesis according to the p-value and the f statistic value obtained from the ANOVA test.

C. Third Hypothesis

H0: There is no significant correlation between the time spent on different site types.
HA: The time spent on specific site types correlates with user behavior.
The correlation table is a crucial tool that visually presents the relationships between variables in the dataset. The table expresses the correlation between variables with values ranging from -1 to 1. Positive correlation values indicate that an increase in one variable is associated with an increase in another, while negative correlation values signify that an increase in one variable is associated with a decrease in another.



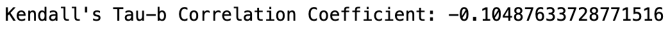
	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	SpecialDay	Revenue_Binary
Administrative	1.00	0.59	0.25	0.17	0.29	0.20	-0.25	-0.35	-0.10	0.14
Administrative_Duration	0.59	1.00	0.20	0.11	0.20	0.22	-0.16	-0.23	-0.07	0.09
Informational	0.25	0.20	1.00	0.55	0.21	0.19	-0.11	-0.16	-0.03	0.07
Informational_Duration	0.17	0.11	0.55	1.00	0.11	0.12	-0.06	-0.09	-0.02	0.04
ProductRelated	0.29	0.20	0.21	0.11	1.00	0.76	-0.28	-0.38	0.01	0.15
ProductRelated_Duration	0.20	0.22	0.19	0.12	0.76	1.00	-0.25	-0.32	-0.02	0.17
BounceRates	-0.25	-0.16	-0.11	-0.06	-0.28	-0.25	1.00	0.92	0.06	-0.16
ExitRates	-0.35	-0.23	-0.16	-0.09	-0.38	-0.32	0.92	1.00	0.09	-0.22
SpecialDay	-0.10	-0.07	-0.03	-0.02	0.01	-0.02	0.06	0.09	1.00	-0.07
Revenue_Binary	0.14	0.09	0.07	0.04	0.15	0.17	-0.16	-0.22	-0.07	1.00

Fig. 23. The Correlation Heatmap

When examining the correlation table, it can be interpreted that there is no significant relationship between the durations spent on different page types (administrative duration, informational duration, product-related duration) and the revenue variable. This interpretation is drawn from the fact that the correlation values are very close to 0. The values approaching 0 suggest a weak or negligible linear relationship between these variables. Therefore, it can be inferred that the time spent on various page types is not significantly correlated with the revenue, as indicated by the correlation values near 0.
According to the correlation values we obtained, we do not reject the H0 hypothesis.

D. Fourth Hypothesis

H0: There is no relationship between visitor type and the likelihood of revenue.
HA: Visitor types influences the likelihood of revenue.
The Kendall's Tau-b correlation coefficient is a statistical measure specifically used to assess the ordinal relationship between binary data. This coefficient serves the purpose of determining the ordered relationship between two variables, particularly effective when categorical data is ordinal. Kendall's Tau-b expresses the strength and direction of the relationship between two variables. A positive value indicates that an increase in one variable is associated with an increase in the other, while a negative value signifies an association where an increase in one variable is associated with a decrease in the other.



```
Kendall's Tau-b Correlation Coefficient: -0.10487633728771516
```

Fig. 24. The Value of The Kendall's Tau-b Correlation Coefficient

The Kendall's Tau-b correlation coefficient for the visitor type and revenue values is approximately -0.1049. This negative correlation suggests a weak, negative association between the visitor type and revenue, indicating that as one variable increases, the other tends to decrease slightly.

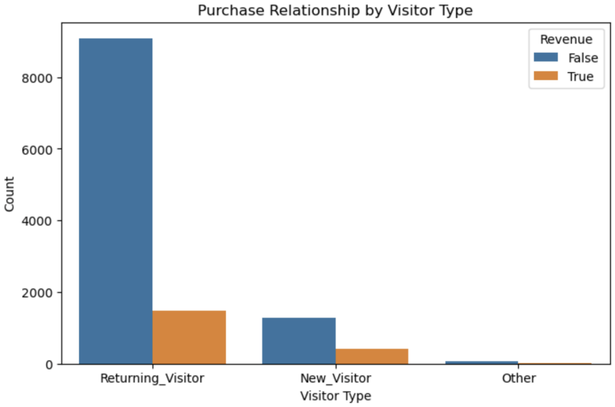


Fig. 25. The Bar Chart of The Visitor Type&Revenue

Upon closer examination of the bar charts representing visitor types, a notable trend emerges, indicating that new visitors exhibit higher shopping engagement rates than returning visitors.The data showcases a meaningful distinction in shopping behaviors between these two visitor segments, emphasizing the significance of visitor type in influencing the likelihood of making a purchase.

We reject the H0 hypothesis according to the value of the correlation coefficient obtained the Kendall's Tau-b Correlation.

V. CONCLUSION

In conclusion, our data analysis project aimed to predict user behaviors on the e-commerce platform. Despite achieving accurate predictions, there is inherent unpredictability in online shopping dynamics. Numerous variables influence user interactions, and there are nuances in customer decisions that cannot be explained through statistical inferences. However, our analysis, encompassing various parameters and their relationships, provides valuable insights into potential trends and patterns.

This approach contributes to a better understanding of user interactions on the platform and serves as a resource for e-commerce sites to inform their strategies. It should be noted that continual monitoring and adaptation are crucial to keeping up with the constant changes in the e-commerce landscape.