Student Name:

[Email address]

DATA ANALYSIS PROJECT

[Document subtitle]

Table of Contents

[Introduction 2](#_Toc102329156)

[Data Engineering 2](#_Toc102329157)

[Exploratory Data Analysis 3](#_Toc102329158)

[MODELLING 7](#_Toc102329159)

[Recommendations 8](#_Toc102329160)

## Introduction

The project investigated data similar to that which Zappos' Web Analytics team would use to gauge the company's overall performance. The project's purpose, which was done as part of this firm's Web Analytics team, was to understand the demographics of visitors to the site, the customers' actions on the site, and how well the company is doing.

The dataset had the following attributes:

|  |  |
| --- | --- |
| DAY | The calendar day |
| VISITS | The sum of visits for a particular day |
| ORDERS | The sum of orders for a particular day |
| HAS\_PURCHASED\_PRIOR | Denotes whether or not metrics apply to a website visitor that has previously placed an order or not |
| DEVICE | Metrics by the device used |
| BOUNCES | The sum of bounces for a particular day. A bounce represents a visitor who enters the site and then leaves ("bounces") rather than continuing to view other pages within the same site. |
| ADD\_TO\_CART | The sum of add to cart events for a particular day |
| PRODUCT\_PAGE\_VIEWS | The sum of product details page views for a particular day |
| SEARCH\_PAGE\_VIEWS | The sum of search results pages viewed for a particular day |
| GENDER | Metrics by customer gender |
| AGE | Metrics by customer age |
| INCOME | Metrics by customer income |
|  |  |

The Data Analysis Process;

Recommendations

Modelling

Exploratory Data Analysis

Data Engineerng

## Data Engineering

The project's first step was to calculate the following metrics to determine how the company was performing. This stage was referred to as the data engineering stage.

|  |  |
| --- | --- |
| CONVERSION\_RATE | ='ORDERS'/'VISITS' |
| BOUNCE\_RATE | ='BOUNCES'/'VISITS' |
| ADD\_TO\_CART\_RATE | ='ADD\_TO\_CART'/'VISITS' |

## Exploratory Data Analysis

Data scientists utilize exploratory data analysis (EDA) to study and investigate data sets and describe their primary properties, frequently using data visualization techniques. It assists data scientists in determining how to manipulate data sources best to obtain the answers they require, making it easier for them to find patterns, test hypotheses, and verify assumptions.

EDA is primarily used to examine what data can disclose outside of formal modeling or hypothesis testing tasks and better understand data set variables and their interactions. It might also assist you in determining whether the statistical techniques you're contemplating for data analysis are suitable.

The primary objective of EDA is to assist in the analysis of data before making any assumptions. It can aid in detecting evident errors, as well as a better understanding of data patterns, the detection of outliers or unusual events, and the discovery of interesting relationships between variables.

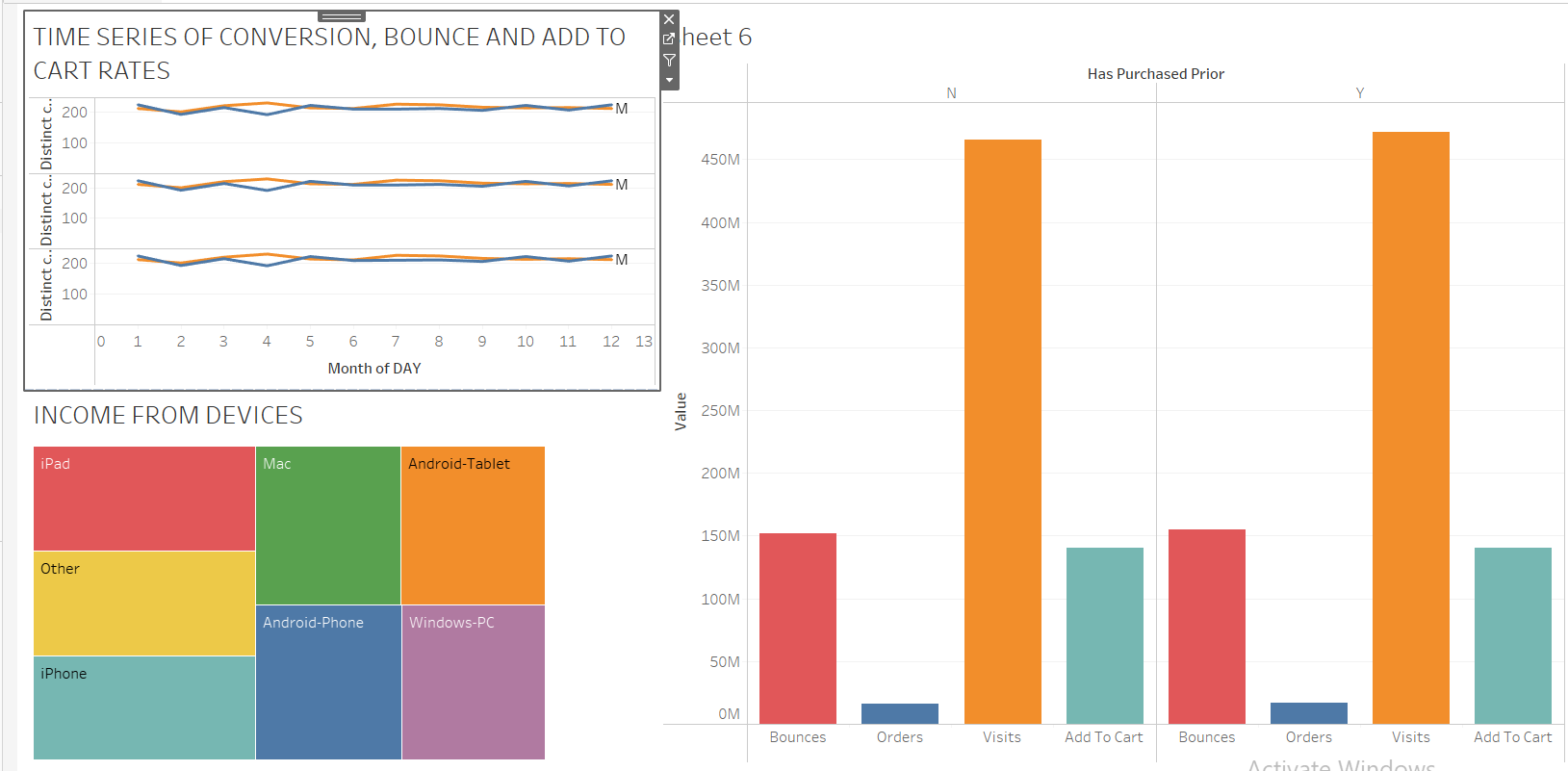
The tools used for exploratory data analysis in this project were Tableau and Python using the Matplotlib and Seaborn libraries.

The exploratory had the following benefit to the stakeholders; it was to identify key performance indicators. This stage helped the stakeholders know the site's performance at a glance through the use of plots and dashboards.

The key performance indicators for this project were:

1. Income
2. Bounce Rate
3. Add to cart rate
4. Conversion Rate
5. Search Pageviews
6. Product page views
7. If the Clients had purchased from the site before

Here is the first dashboard:



From the first plot, we note the following:

The add to cart rate - The proportion of the visitors to the site that added items to their carts to the total number of visitors).

The Bounce Rate - the proportion of the clients who enter the site and then leave ("bounces") rather than continuing to view other pages within the same site to the total number of visits to the site.

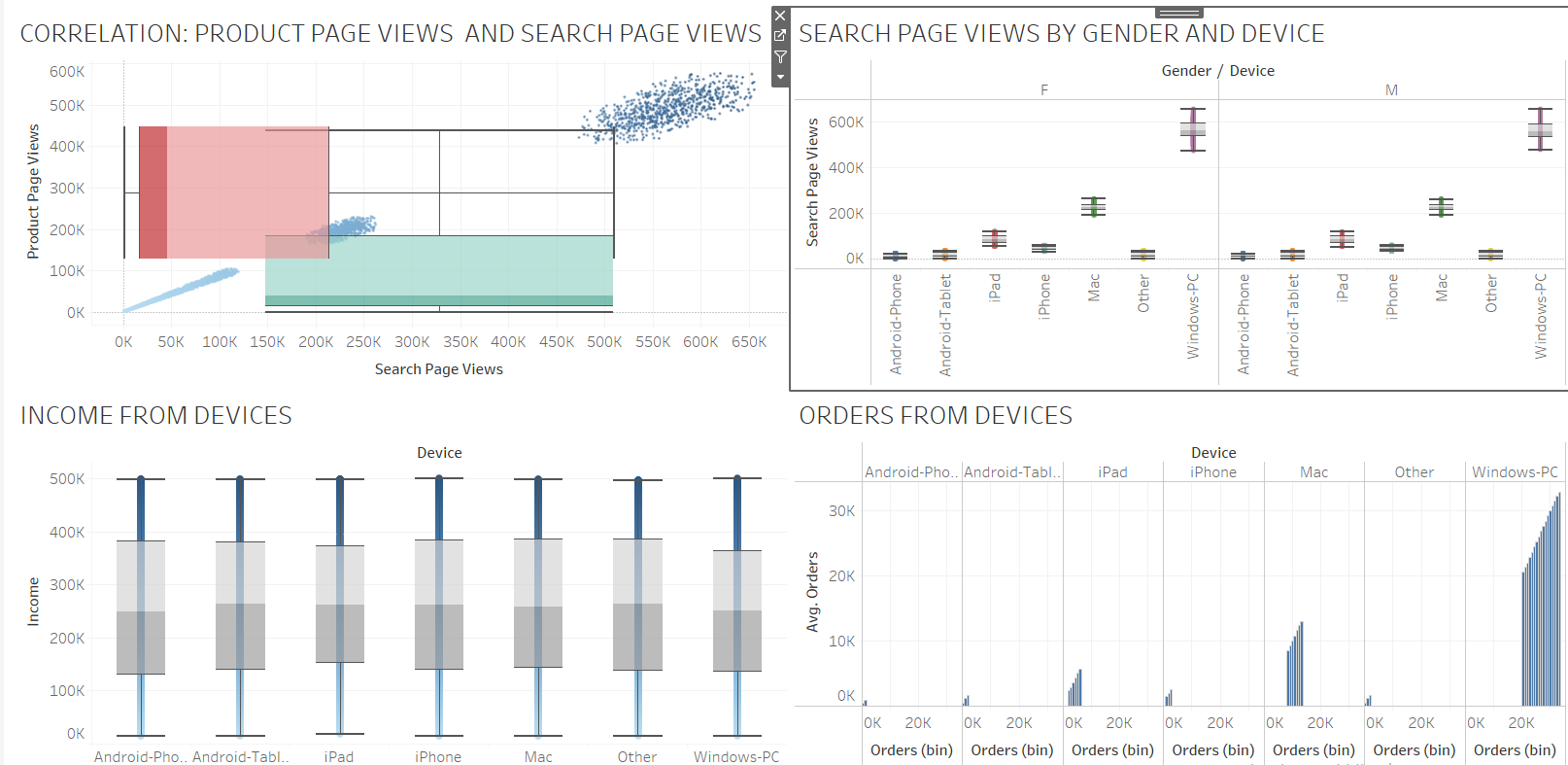
Add to the Cart Rate – The proportion of the add-to-cart events for a particular day to the total number of visits to the site.

We note that the three key performance indicators all followed the same trend the whole year for both Gender; there was not much difference in trend for both genders except for April.

The performance indicators prompted a look into the proportions of the visits, orders, and Bounces: The number of bounces, add to carts, and orders were only at most a third of the visits.

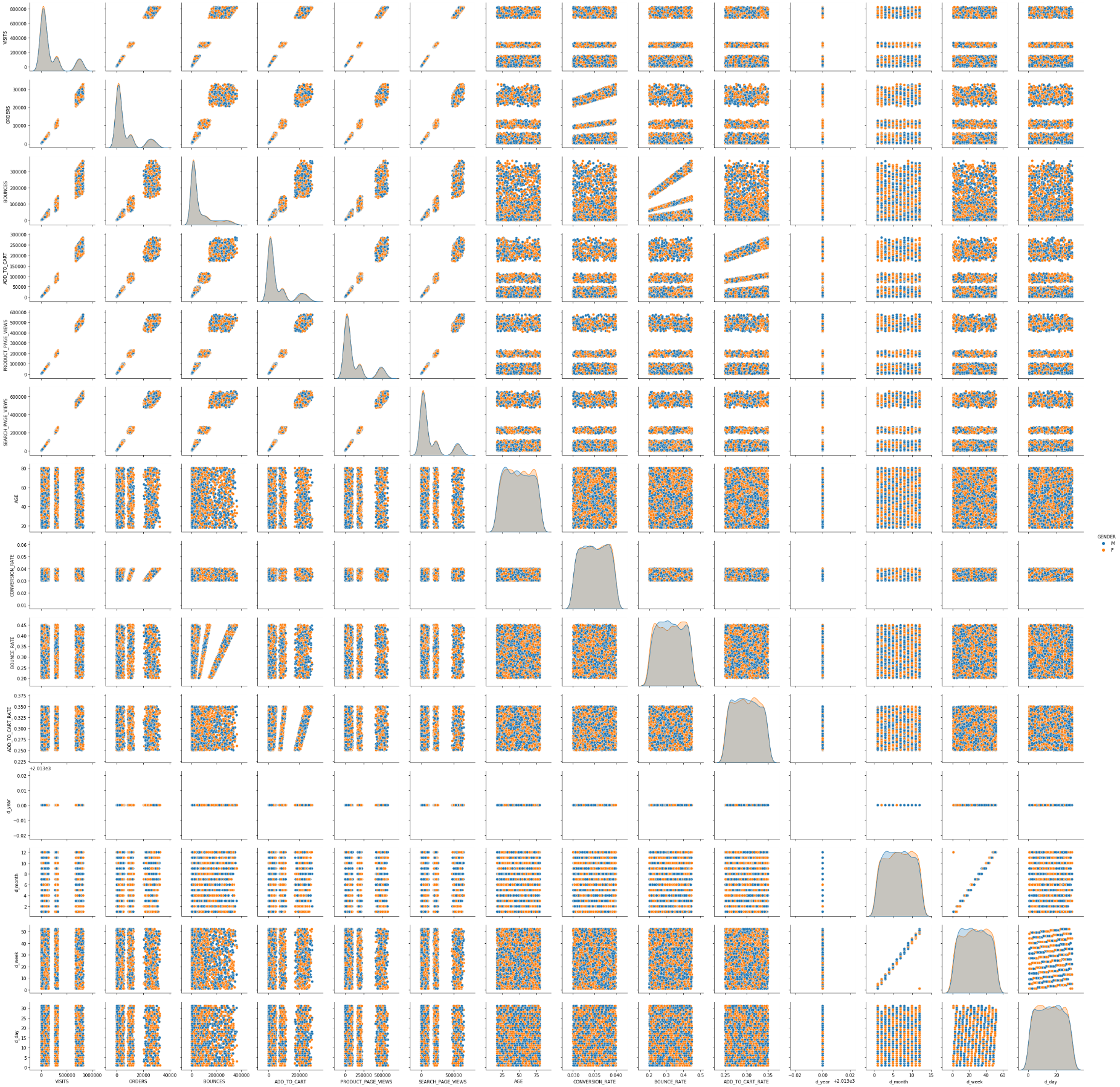
Another explored aspect was the income from the devices; the devices had almost the same amount of income. The box plots from the second dashboard (below) emphasize the same thing. This could be due to clients having connected accounts on their devices. The project went a step further to determine where the orders were made from.

Here is the second dashboard:



We note that most orders were made from windows computers from the dashboard, followed by Macs. This meant that first, most clients own windows laptops, and customers prefer to view the site from large screens hence windows laptops and MACs.

Both genders have no significant differences in where to view the site; they prefer windows and MACs, which could be due to the large screens. Clients first visit the search page and then proceed to view the product page, which means an increase in search page views also means an increase in product page views. This suggests a positive correlation between search page views and product page views.

We note from the pair plot that the only noticeable distributions/correlations are the visits, bounces, add to cart, orders, product page search views, and search page views, which are positive correlations. Age and Gender have no noticeable correlation. 

## Modeling

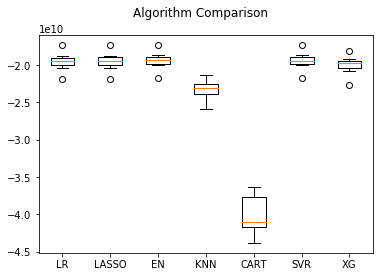
The model was prepared with the main objective of determining which features were most important in increasing the income from the site.

The machine learning approach was taken:

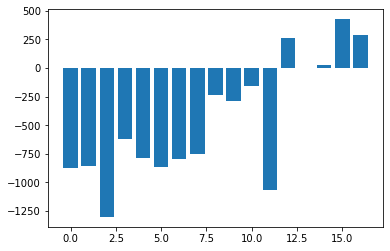
The following models were trained :

1. Linear Regression
2. Lasso Regression
3. Elastic Net Regression
4. K-Nearest Neighbors
5. XG Boost Model
6. Decision Trees Model
7. Support Vector Machine Model

The best performing model was the Elastic Net Regression Model; here is a plot of their performance:



The model was further trained independently, and the features were plotted according to their importance.



We note that the model emphasized the number of events added to the cart, the day, week, and month to be the most important features determining how much income will be earned.

Here is the order of importance:

1. Week
2. Day
3. Add to Cart
4. Month

## Recommendations

1. The company should focus on attracting mobile device users, making the site mobile-friendly since most users prefer to make orders and browse the site from laptops. This recommendation will profit both site owners and customers. By making creating a better experience for visitors to the site, means more purchases can be made at the site
2. The site should make offers at specific times of the month, especially the end months and beginning of the month when clients are willing and able to make purchases.