UNIVERSITY OF DOHA FOR SCIENCE AND TECHNOLOGY

HIGH SPENDER PREDICTION USING MACHINE LEARNING AND SHAP EXPLAINABILITY

MACHINE LEARNING - 3201 PRESENTATION BY

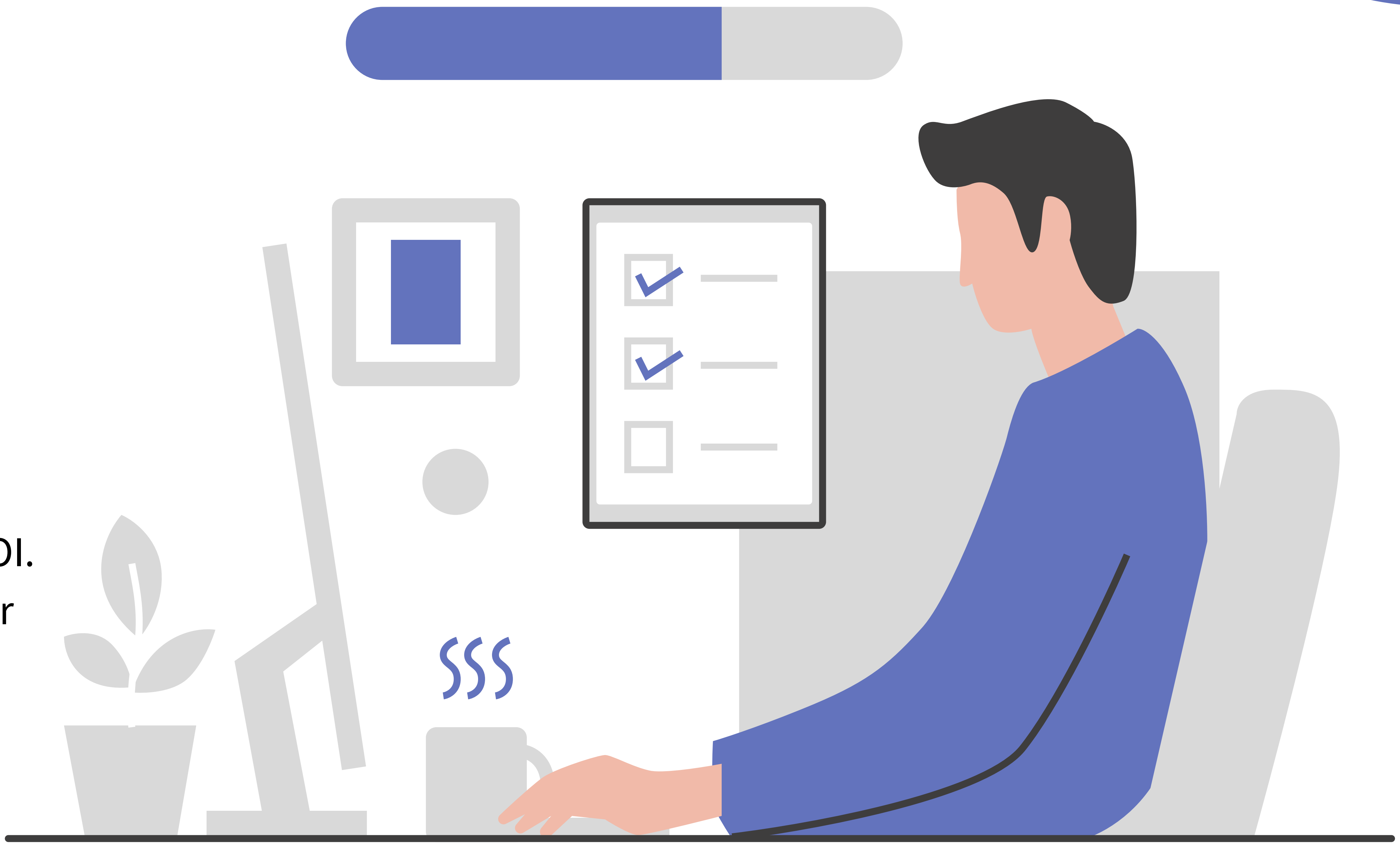
SYED ABDULLAH - 60104641
SHAIMAH MOHAMMED -

OVERVIEW OF THE PROJECT

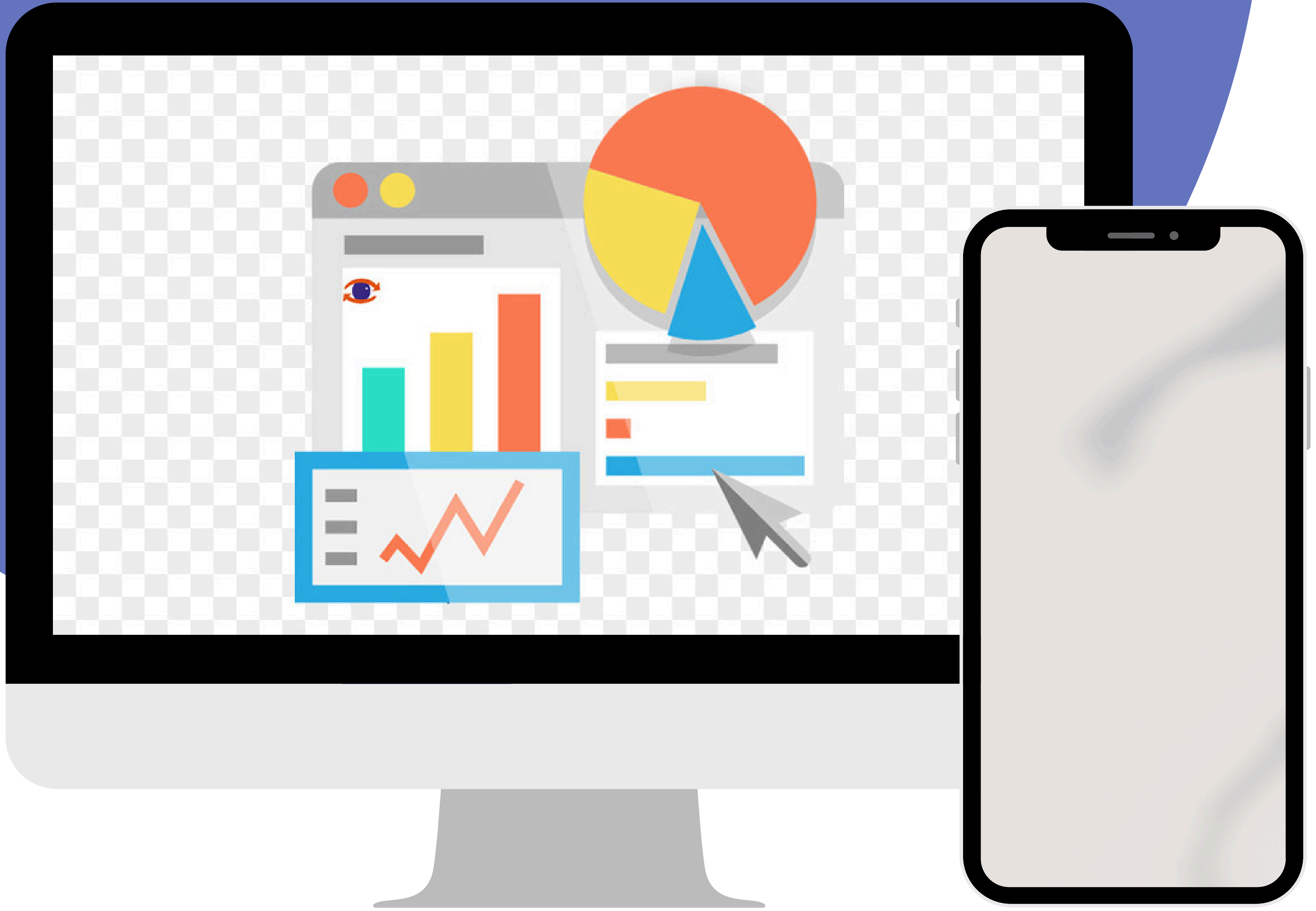
PROJECT OBJECTIVE

In this project, my goal was to analyze customer behavior and predict high-spending customers using customer and purchase data. By identifying these high-value customers, businesses can optimize their marketing efforts, personalize campaigns, and ultimately improve their ROI. Understanding customer spending behavior is crucial for targeted marketing, customer retention, and resource allocation.

LETS GET STARTED



DATASET DESCRIPTION & BACKGROUND CONTEXT



BUSINESS PROBLEM

The goal is to predict high-spending customers based on their purchase behavior and demographics. Identifying high spenders allows businesses to target them with personalized offers, optimize marketing efforts, and improve customer retention strategies.

DATASET OVERVIEW

The dataset is sourced from Kaggle and consists of three files: customer_data.csv, product_data.csv, and purchase_data.csv. It contains a total of 40 columns and approximately 15,000 rows. The dataset includes customer demographics, product details, and transaction history. The main objective of the analysis is to predict high-spending customers based on their purchase behavior. This will help optimize marketing strategies and improve customer retention efforts.

HOW THE PROBLEM IS ADDRESSED

By merging these datasets, I built a comprehensive dataset to analyze customer behavior. After cleaning and engineering relevant features, I used machine learning models to predict high spenders, helping businesses focus resources on high-value customers.

WHY I CHOSE THIS DATASET

I chose this dataset because it provides a comprehensive view of customer behavior, product details, and purchasing patterns, which directly aligns with my objective to predict high-spending customers.

RELEVANCE

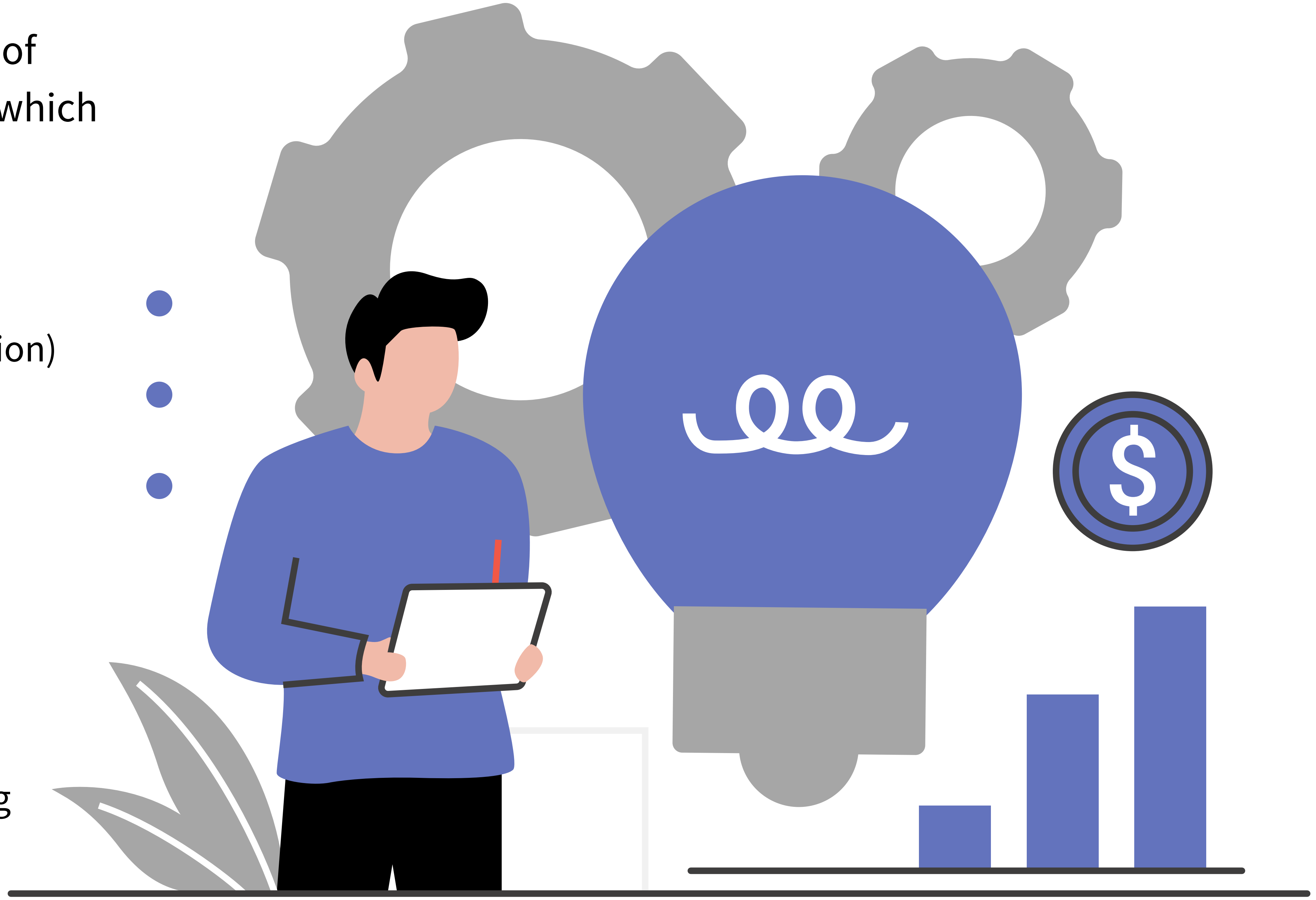
The dataset includes customer demographics (e.g., age, gender, region) and purchase history, essential for identifying spending patterns.

COMPREHENSIVE

It contains both numerical (e.g., total spend) and categorical (e.g., product category, gender) features, providing a solid foundation for analysis.

DATA QUALITY

The data was well-structured and cleaned, with minor preprocessing required, making it suitable for model building

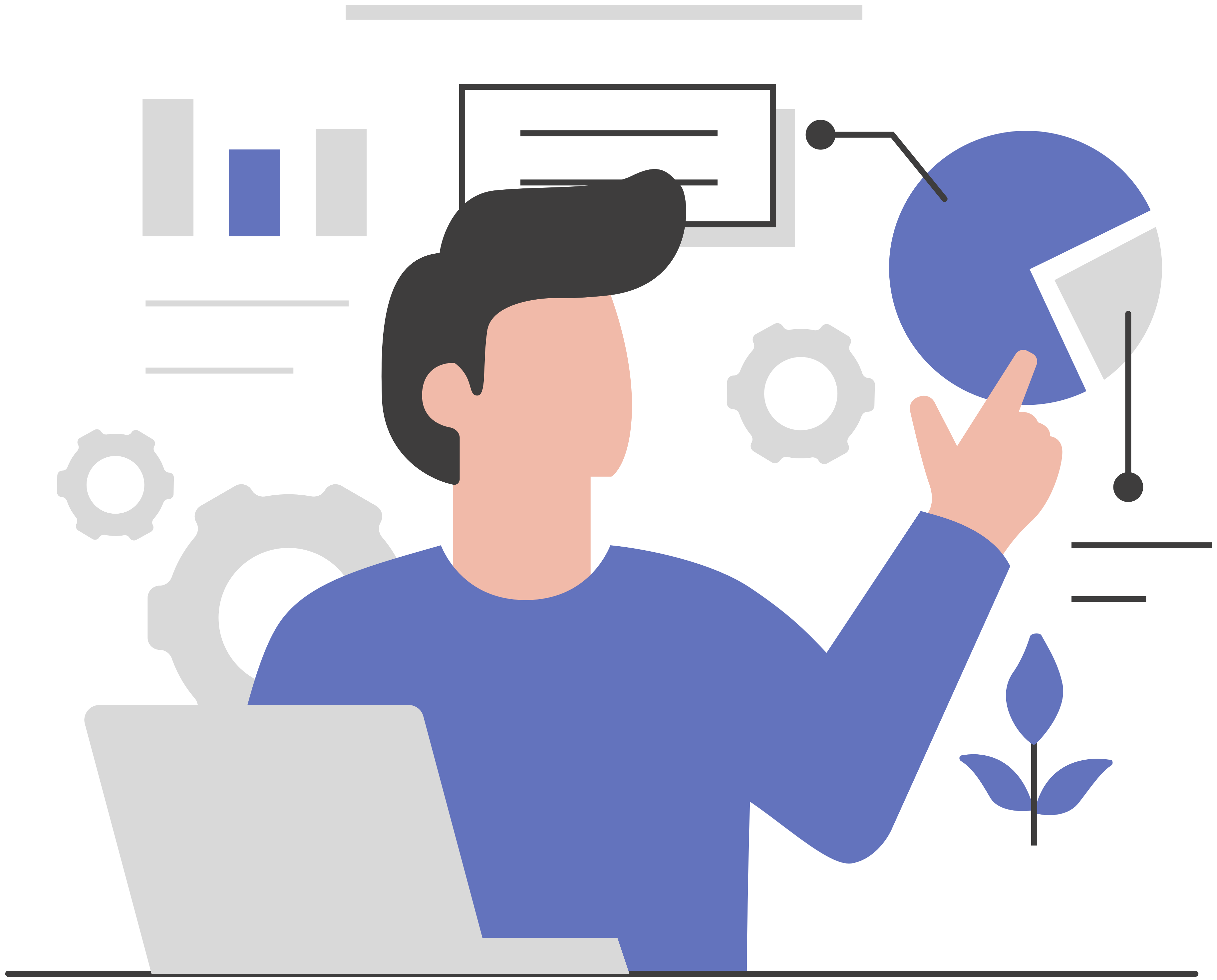


VISUAL DATASET SNAPSHOT

SAMPLE OF DATASET (CUSTOMER, PRODUCT, AND PURCHASE DATA) WITH A FEW ROWS TO PROVIDE CLARITY ON THE STRUCTURE.

	purchase_id	customer_id	product_id	purchase_date	quantity	total_amount	first_name	last_name	gender	date_of_birth	...	signup_date	address	city	state
0	1	1	42	2018-04-15 14:08:01	3	37.642074	Robert	Smith	Female	1994-06-14 21:40:27	...	2016-10-16 17:23:25	8465 Main St	San Antonio	CA
1	2	1	138	2022-07-10 23:33:47	4	70.247106	Robert	Smith	Female	1994-06-14 21:40:27	...	2016-10-16 17:23:25	8465 Main St	San Antonio	CA
2	3	1	403	2021-12-31 03:53:33	3	89.168896	Robert	Smith	Female	1994-06-14 21:40:27	...	2016-10-16 17:23:25	8465 Main St	San Antonio	CA
3	4	1	193	2017-01-14 01:25:11	2	59.705059	Robert	Smith	Female	1994-06-14 21:40:27	...	2016-10-16 17:23:25	8465 Main St	San Antonio	CA
4	5	1	26	2018-04-06 11:01:06	3	101.778864	Robert	Smith	Female	1994-06-14 21:40:27	...	2016-10-16 17:23:25	8465 Main St	San Antonio	CA

5 rows × 22 columns



APPROACH:

DATA PREPROCESSING

I cleaned and transformed the raw data, handling missing values, encoding categorical features, and performing necessary feature engineering

EXPLORATORY DATA ANALYSIS (EDA):

I explored the data visually to identify trends, correlations, and patterns in customer behavior, which helped me select features for the model.

MODEL BUILDING

I used Logistic Regression and Shap to automatically train and tune machine learning models.

DATA CLEANING & PREPROCESSING

Before starting the analysis, I performed essential data cleaning and preprocessing steps to ensure the data was ready for modeling:

Missing Values

I handled missing values by either imputing (filling in) or removing rows/columns with too many missing values

Duplicate Records

I identified and removed any duplicate entries in the dataset to avoid bias.

Data Transformation

I created new features like signup years (calculated from signup_date) and total spending per customer to improve the model's predictive power

Feature Encoding

Categorical features like gender and product category were encoded into numerical values for compatibility with the machine learning model.

VISUAL DATA CLEANING SNAPSHOT

DATA CLEANING: A SCREENSHOT OF THE CODE USED FOR HANDLING MISSING VALUES OR ENCODING FEATURES.

```
print("\nMissing values in Product Data:")
print(product_df.isnull().sum())

print("\nMissing values in Purchase Data:")
print(purchase_df.isnull().sum())

# Drop duplicates
customer_df.drop_duplicates(inplace=True)
product_df.drop_duplicates(inplace=True)
purchase_df.drop_duplicates(inplace=True)
```

```
Missing values in Customer Data:
customer_id      0
first_name       0
last_name        0
gender           0
date_of_birth    0
email            0
phone_number     0
signup_date      0
address          0
```

EDA OVERVIEW

EXPLORATORY DATA ANALYSIS (EDA) WAS AN ESSENTIAL PART OF THIS PROJECT, AS IT HELPED ME UNDERSTAND THE DATASET AND UNCOVER MEANINGFUL PATTERNS BEFORE BUILDING THE PREDICTIVE MODEL.

THE MAIN OBJECTIVES OF THE EDA WERE:

- To understand the distribution of key variables.
- To identify any patterns or correlations between features.
- To visualize how different factors (like age, gender, and region) impact spending behavior.

EXPLORATORY DATA ANALYSIS

In this section, we perform descriptive statistics, correlations, and visualizations to explore trends and patterns. These insights will guide our modeling and business recommendations.

[196]:
```
# General statistical summary
full_df.describe(include='all')
```

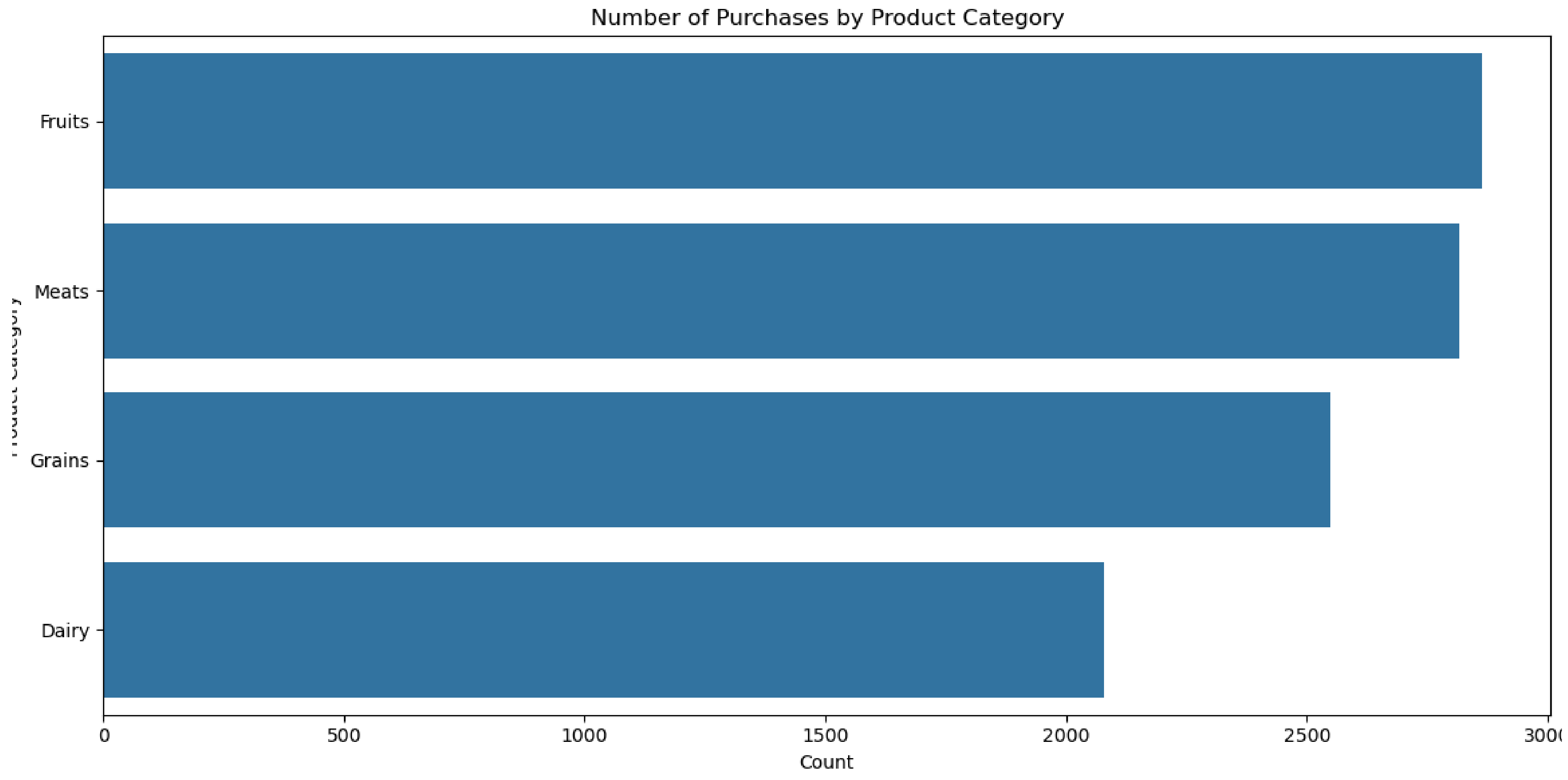
[196]:

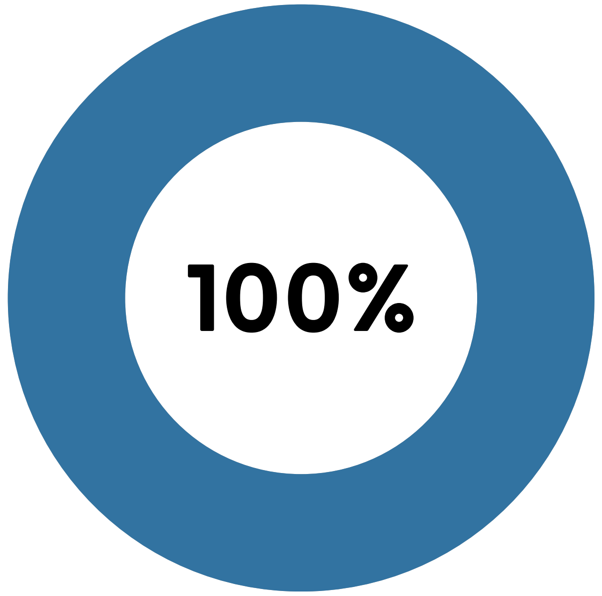
	purchase_id	customer_id	product_id	purchase_date	quantity	total_amount	first_name	last_name	gender	date_of_birth	...	signup_date	address
count	10308.000000	10308.000000	10308.000000	10308	10308.000000	10308.000000	10308	10308	10308	10308	...	10308	10308
unique	NaN	NaN	NaN	10308	NaN	NaN	10	10	2	1000	...	1000	990
top	NaN	NaN	NaN	2018-04-15 14:08:01	NaN	NaN	Alex	Smith	Female	1973-02-01 13:43:23	...	2016-06-18 01:41:15	7346 Main St
freq	NaN	NaN	NaN	1	NaN	NaN	1240	1158	5496	20	...	20	28
mean	5154.500000	504.540648	251.363795	NaN	3.030656	77.423841	NaN	NaN	NaN	NaN	...	NaN	NaN
std	2975.807621	292.026758	143.690280	NaN	1.412852	58.719304	NaN	NaN	NaN	NaN	...	NaN	NaN
min	1.000000	1.000000	1.000000	NaN	1.000000	1.526648	NaN	NaN	NaN	NaN	...	NaN	NaN
25%	2577.750000	245.750000	127.000000	NaN	2.000000	30.143436	NaN	NaN	NaN	NaN	...	NaN	NaN
50%	5154.500000	510.000000	253.000000	NaN	3.000000	62.499946	NaN	NaN	NaN	NaN	...	NaN	NaN
75%	7731.250000	758.000000	375.250000	NaN	4.000000	113.776378	NaN	NaN	NaN	NaN	...	NaN	NaN
max	10308.000000	1000.000000	500.000000	NaN	5.000000	249.963513	NaN	NaN	NaN	NaN	...	NaN	NaN

11 rows × 22 columns

DISTRIBUTION OF PURCHASES BY PRODUCT CATEGORY

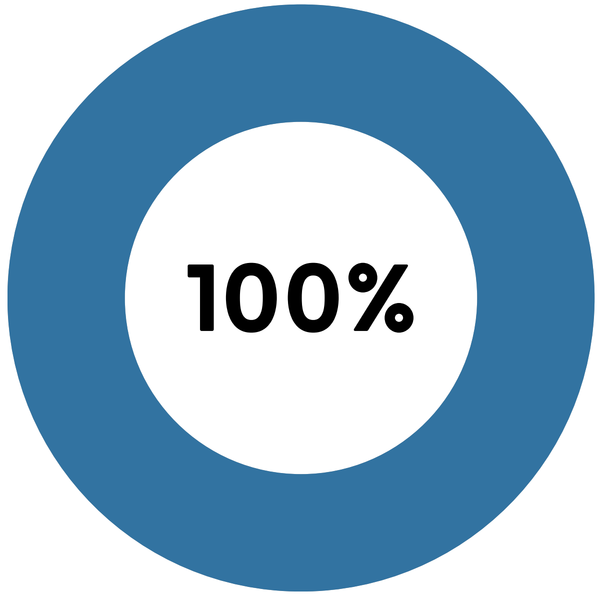
In this step of the Exploratory Data Analysis (EDA), I analyzed the distribution of purchases across different product categories to identify which categories generate the highest number of purchases. This helps in understanding customer preferences and pinpointing high-performing product segments.





ANALYSIS

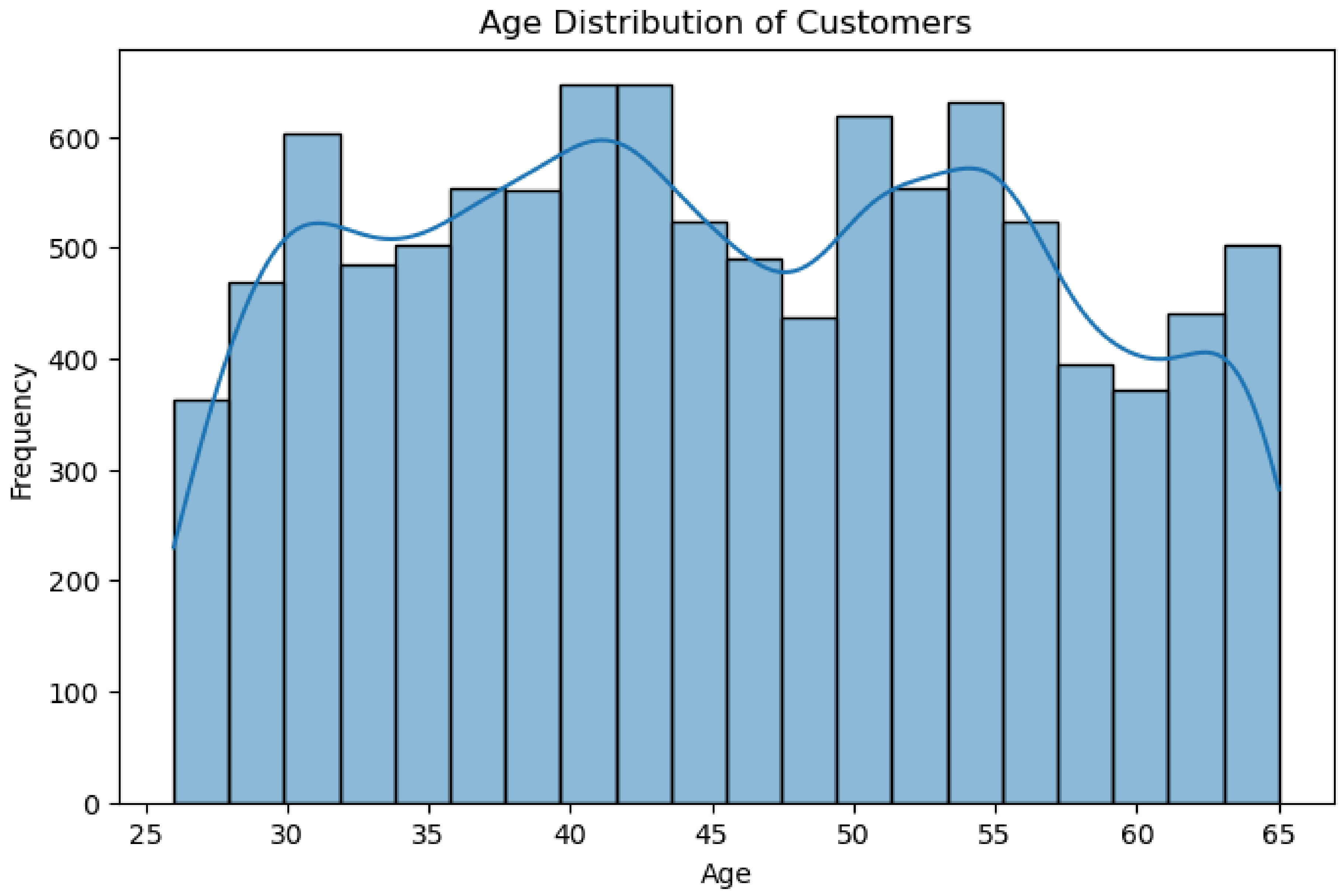
I used a horizontal bar chart to visualize the number of purchases in each product category.



INSIGHT

The analysis revealed that Fruits and Meats are the highest purchased categories, followed by Grains and Dairy. This indicates the potential for businesses to focus on these categories for targeted promotions and marketing efforts.

AGE DISTRIBUTION OF CUSTOMERS



400K

HIGHEST SALES

A histogram with a KDE (Kernel Density Estimate) overlay was used to visualize the distribution of customer ages. The histogram shows the frequency of customers in each age group, while the KDE provides a smooth estimation of the distribution.

70%

INSIGHTS

The age distribution appears relatively uniform with a slight peak in the 30-45 age range. This suggests that customers in their 30s to 40s form a significant portion of the customer base, and businesses can target this group with age-appropriate products and services.

CHART SUMMARY

In this part of the Exploratory Data Analysis (EDA), I examined the age distribution of customers. This helps to understand the demographics of the customer base, which can inform targeted marketing efforts, product recommendations, and customer service

CUSTOMERS BY TOTAL SPEND

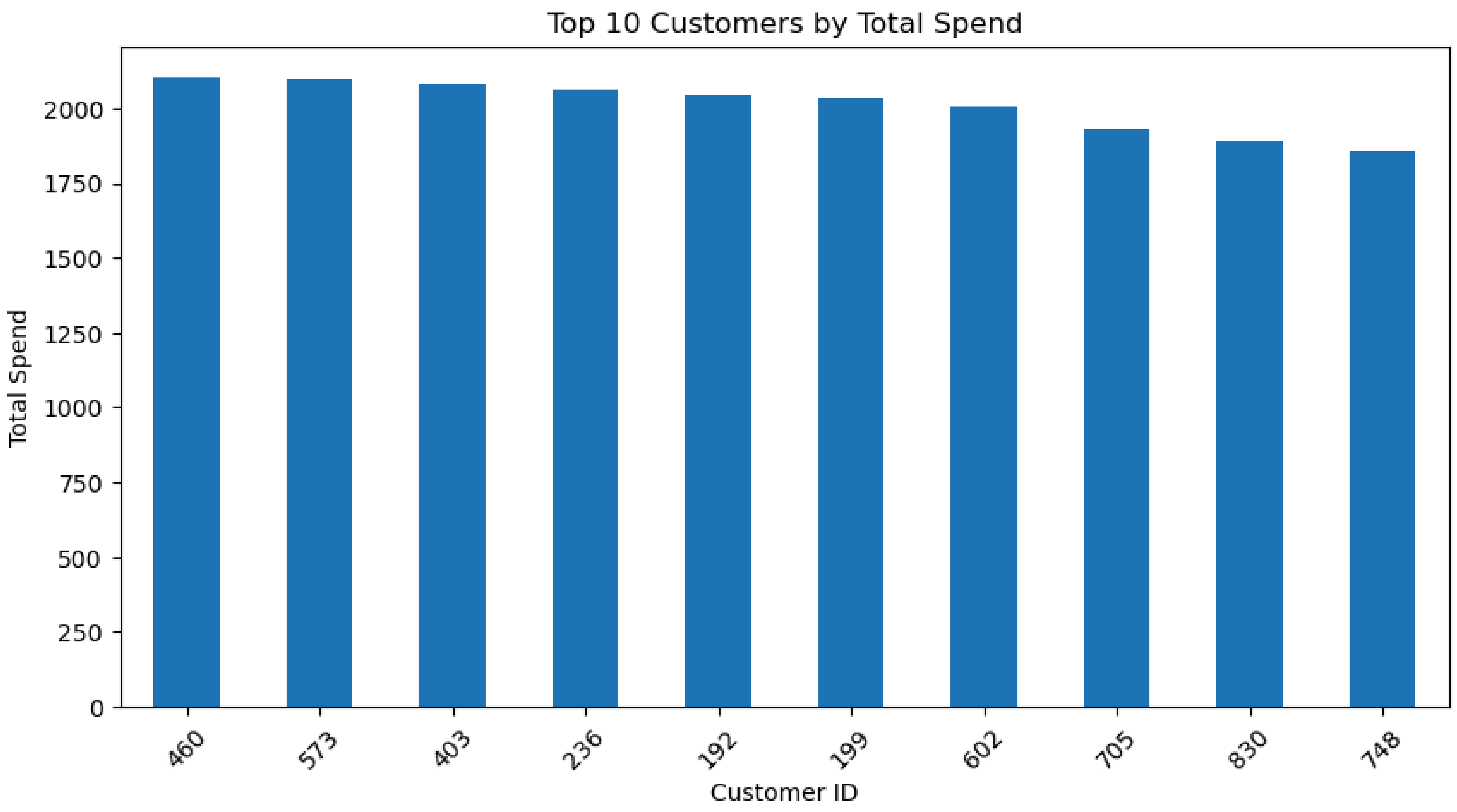
In this analysis, I identified the top 10 customers based on their total spend. This is crucial for understanding high-value customers and can be used to create loyalty programs or targeted marketing campaigns.

ANALYSIS

The bar chart highlights the total amount spent by the top 10 customers. Each customer is represented by their Customer ID and the corresponding total amount spent.

INSIGHT

The customers with the highest total spend are likely the most valuable to the business. Businesses should focus on maintaining and nurturing these high-spending customers, as they contribute a significant portion of revenue.



FEATURE ENGINEERING & TARGET DEFINITION

Feature engineering is the process of transforming raw data into meaningful features to improve machine learning model performance. For this project, we aimed to predict whether a customer is a high spender based on their purchase behavior and profile

IDENTIFYING RELEVANT FEATURES

Focused on critical features impacting high-spender prediction, such as:
Total Purchase Amount, Product Categories, Age, Signup Year, Gender

FEATURE CREATION

Aggregating Data: Created new features like Avg Spend and Total Quantity to summarize customer purchase behavior.

TARGET DEFINITION

HighSpender: Defined as customers with total spend above the 75th percentile.

DATA PREPROCESSING

- Removed irrelevant columns (purchase_id, purchase_date).
- Encoded categorical features like Category and Gender using One-Hot Encoding.

SHAP SUMMARY PLOT

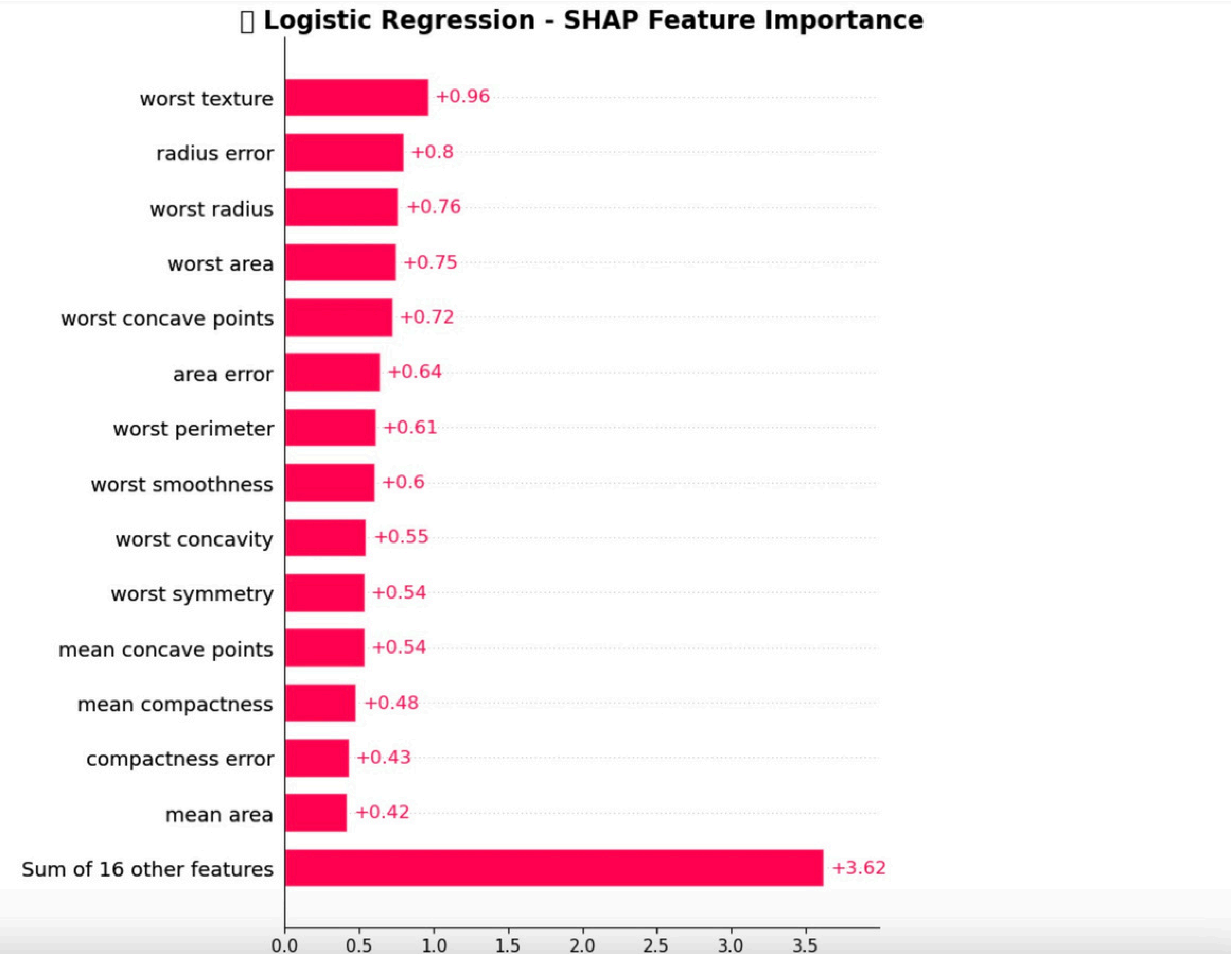
DISTRIBUTION OF IMPACT

EXPLANATION

This plot shows which features had the most impact on the model’s predictions.
Each dot is one prediction. The color shows the feature value (red = high, blue = low).
Position on the x-axis shows how much that feature increased or decreased the prediction.

WHY IT MATTERS

It helps us understand how different features affect predictions, not just how important they are.

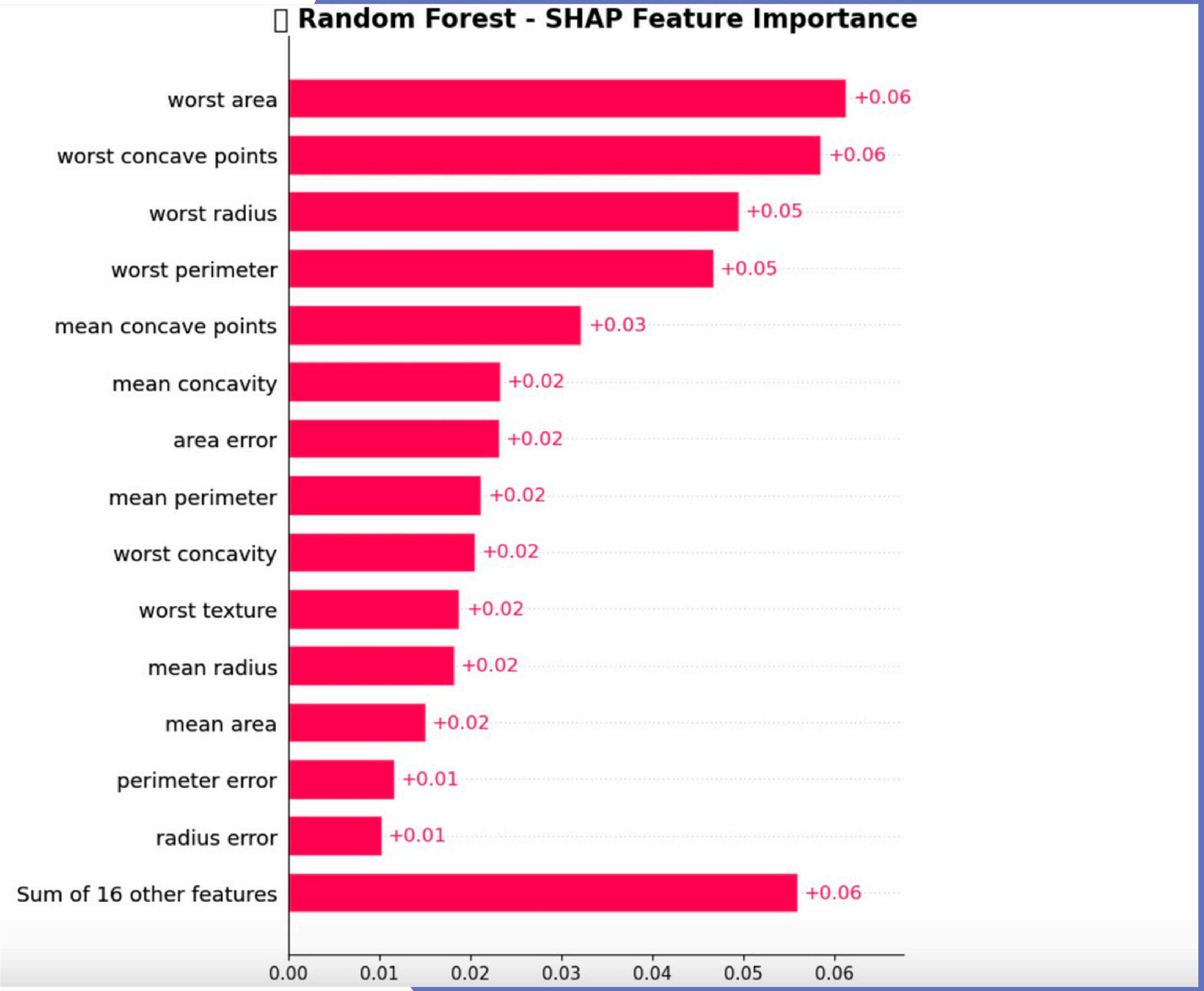


SHAP FEATURE IMPORTANCE

AVERAGE IMPACT ON MODEL PREDICTION

EXPLANATION

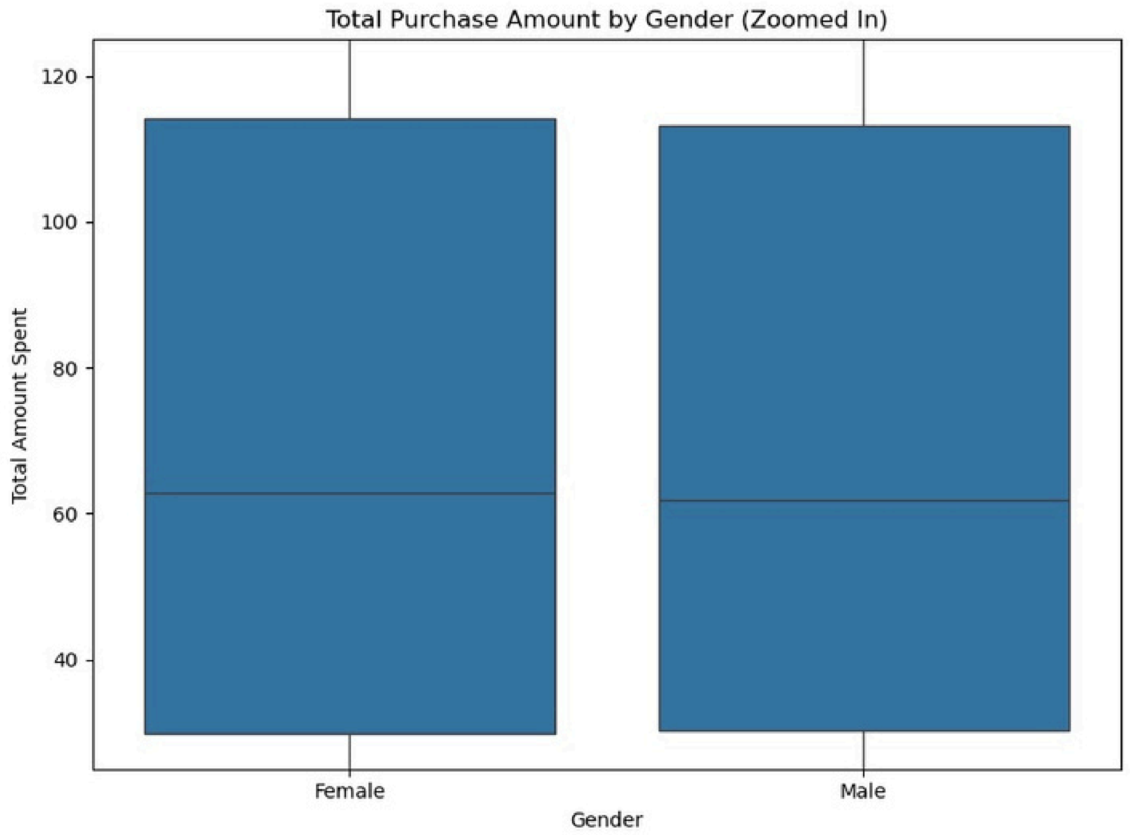
This bar chart shows the average impact of each feature across all predictions.
The higher the bar, the more important the feature.

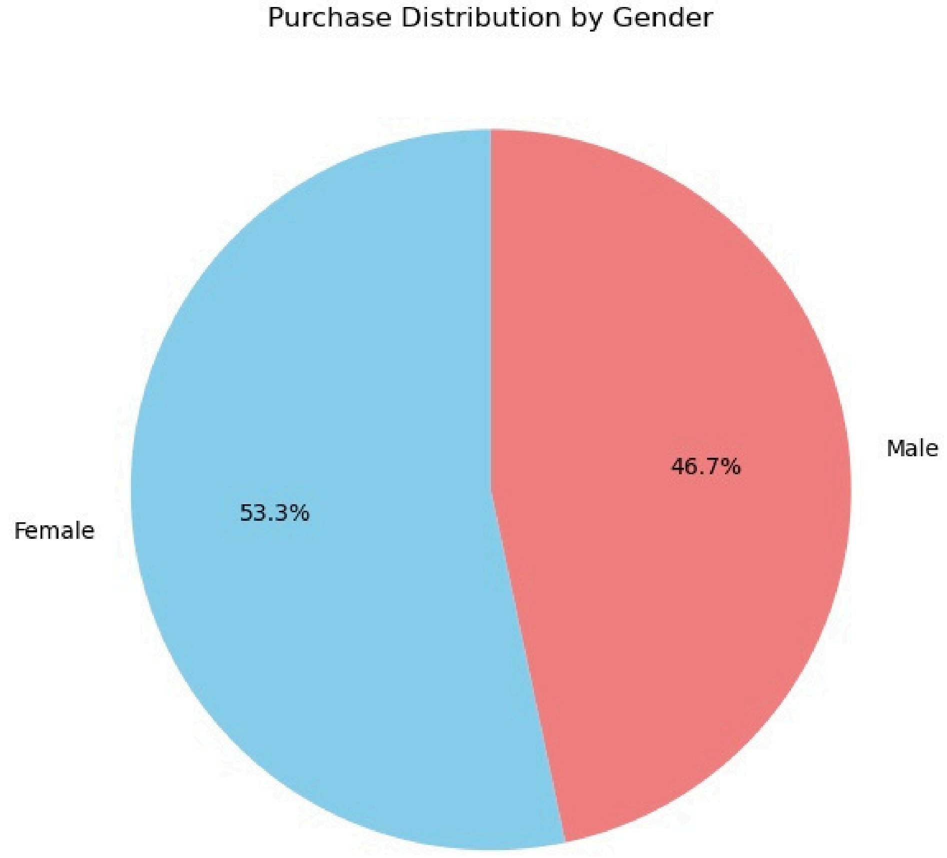


WHY IT MATTERS

It shows which features the model relies on most when predicting high-spending customers.

BIAS ANALYSIS





1

GENDER COMPARISON

- We checked if males or females spent more or were predicted differently.
- Results show both genders have similar total spending and almost equal purchase share, with females slightly higher at 53.3%.

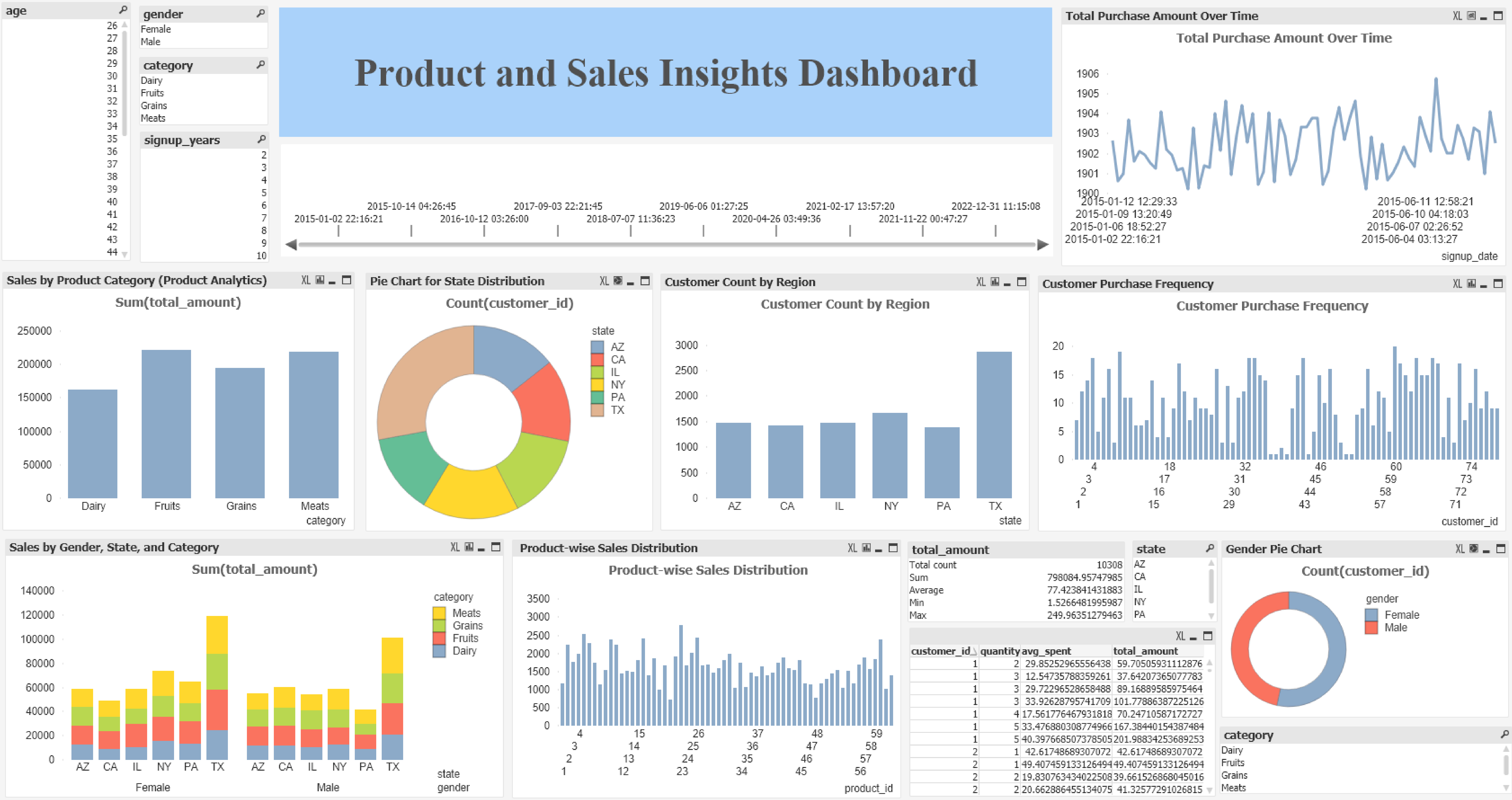
2

WHY IT'S IMPORTANT

- To make sure the model is fair and not favoring one gender.
- Our data shows a balanced pattern in total spending and number of purchases between males and females, so no strong gender bias.

UNIVERSITY OF DOHA FOR SCIENCE AND TECHNOLOGY

DASHBOARD DEMO



Product and Sales Insights Dashboard
age
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
gender
Female
Male
category
Dairy
Fruits
Grains
Meats
signup_years
2
3
4
5
6
7
8
9
10
2015-10-14 04:26:45
2017-09-03 22:21:45
2019-06-06 01:27:25
2021-02-17 13:57:20
2022-12-31 11:15:08
2015-01-02 22:16:21
2016-10-12 03:26:00
2018-07-07 11:36:23
2020-04-26 03:49:36
2021-11-22 00:47:27
Total Purchase Amount Over Time
Total Purchase Amount Over Time
1906
1905
1904
1903
1902
1901
1900
2015-01-12 12:29:33
2015-01-09 13:20:49
2015-01-06 18:52:27
2015-01-02 22:16:21
2015-06-11 12:58:21
2015-06-10 04:18:03
2015-06-07 02:26:52
2015-06-04 03:13:27
signup_date
Sales by Product Category (Product Analytics)
Sum(total_amount)
250000
200000
150000
100000
50000
0
Dairy
Fruits
Grains
Meats
category
Pie Chart for State Distribution
Count(customer_id)
state
AZ
CA
IL
NY
PA
TX
Customer Count by Region
Customer Count by Region
3000
2500
2000
1500
1000
500
0
AZ
CA
IL
NY
PA
TX
state
Customer Purchase Frequency
Customer Purchase Frequency
20
15
10
5
0
4 18 32 46 60 74
3 17 31 45 59 73
2 16 30 44 58 72
1 15 29 43 57 71
customer_id
Sales by Gender, State, and Category
Sum(total_amount)
140000
120000
100000
80000
60000
40000
20000
0
category
Meats
Grains
Fruits
Dairy
AZ CA IL NY PA TX
Female
AZ CA IL NY PA TX
Male
state
gender
Product-wise Sales Distribution
Product-wise Sales Distribution
3500
3000
2500
2000
1500
1000
500
0
4 15 26 37 48 59
3 14 25 36 47 58
2 13 24 35 46 57
1 12 23 34 45 56
product_id
total_amount
Total count 10308
Sum 798084.95747985
Average 77.423841431883
Min 1.5266481995987
Max 249.96351279463
state
AZ
CA
IL
NY
PA
customer_id quantity avg_spent total_amount
1 2 29.85252965556438 59.70505931112876
1 3 12.54735788359261 37.64207365077783
1 3 29.72296528658488 89.16889585975464
1 3 33.92628795741709 101.77886387225126
1 4 17.561776467931818 70.24710587172727
1 5 33.476880308774966 167.38440154387484
1 5 40.39766850737850 201.98834253689253
2 1 42.61748689307072 42.61748689307072
2 1 49.40745913312649 49.40745913312649
2 2 19.830763434022508 39.661526868045016
2 2 20.662886455134075 41.32577291026815
Gender Pie Chart
Count(customer_id)
gender
Female
Male
category
Dairy
Fruits
Grains
Meats