3. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a crucial step in the data analysis process, including when dealing with transactional data. EDA helps to gain insights, identify patterns, understand the structure of the data, and uncover potential issues.

3.1. DATASET SIZE AND STRUCTURE

The dataset is the result of comprehensive data cleaning and the integration of the transactions fact table with relevant dimension tables (demographic, address). This dataset is well-prepared for further analysis, insights extraction, and modeling. Several steps were taken to clean and prepare the data:

- Missing values were handled using appropriate methods such as imputation or dropping.
- Spelling mistakes, data inconsistencies, and anomalies were corrected.
- Irrelevant and not interpretable columns dropped out.
- Calculated fields added (profit, age, transaction months etc.)
- Merging data into a single dataset for further analysis

The final dataset consists of 19,319 records and 27 fields and contains data on approved transactions made by 3,411 distinct customers out of 3,494 customers given in the original transaction's dataset, which is 97.6%. Among transactional data the dataset represents a comprehensive collection of customer-related information. The following tables presents short description of dataset`s attributes:

Table 3 – Dataset

Attribute	Description	Comment
transaction_id, product_id, customer_id	Unique ID for each transaction, purchased product and customer associated with the transaction	19,319 transactions, 101 products and 3,411 customers
transaction_date	Date of the transaction	Year 2017
online_order	Whether the order was placed online	'1'/'0'/'unknown'
brand	Brand of the product	6 brands
product_line	Line of the product	'Standard', 'Road', 'Touring', 'Mountain'
product_class	Classification of the product	'medium', 'high', 'low'
product_size	Size category of the product	'medium', 'large', 'small'
profit	Profit generated from the transaction	Calculated as (list_price – standard_cost)
gender	Gender of the customer	'M'/'F'
past_3_years_bike_related_purchases	Number of purchases	Given within the original dataset
job_industry_category	Industry category of the customer's job	10 industry categories
wealth_segment	Customer's wealth segment	'Mass, 'Affluent, 'High Net Worth'
owns_car	Whether the customer owns a car	'Yes', 'No'
age	Age of the customer.	Calculated using given DOB
postcode	Postal code of the customer	Australian postcodes
state	State where the customer resides	'VIC', 'NSW', 'QLD'
property_valuation	Valuation of the customer's property	From 1 to 12

3.2. SUMMARY STATISTICS FOR NUMERIC COLUMNS

Basic descriptive statistics help to understand the distribution of values, identify outliers, and get a sense of the overall characteristics of the dataset. Summary statistics for numerical columns (mean, median, min, max, standard deviation) are given in the following table:

Table 4 – Basic descriptive statistics for numerical columns

descriptive statistics	count	mean	std	min	25%	50%	75%	max
list_price	19,319	1,106.29	582.86	12.01	575.27	1,163.89	1,635.30	2,091.47
standard_cost	19,319	555.76	405.69	7.21	215.14	507.58	795.10	1,759.85
profit	19,319	550.53	492.92	4.80	133.78	445.21	827.16	1,702.55
past_purchases	19,319	48.92	28.64	0.00	24.00	48.00	73.00	99.00
tenure	19,319	10.68	5.67	1.00	6.00	11.00	15.00	22.00
age	19,319	45.49	12.61	21.00	36.00	45.00	55.00	91.00
property_valuation	19,319	7.52	2.83	1.00	6.00	8.00	10.00	12.00

3.3. SALES TREND

This sub-chapter provides the general overview of the sales trend for the year of analysis. The more advanced time-series analysis with activities forecasting is provided in the chapter 'Time-Series Analysis'.

The following chart represents monthly sales and number of orders:

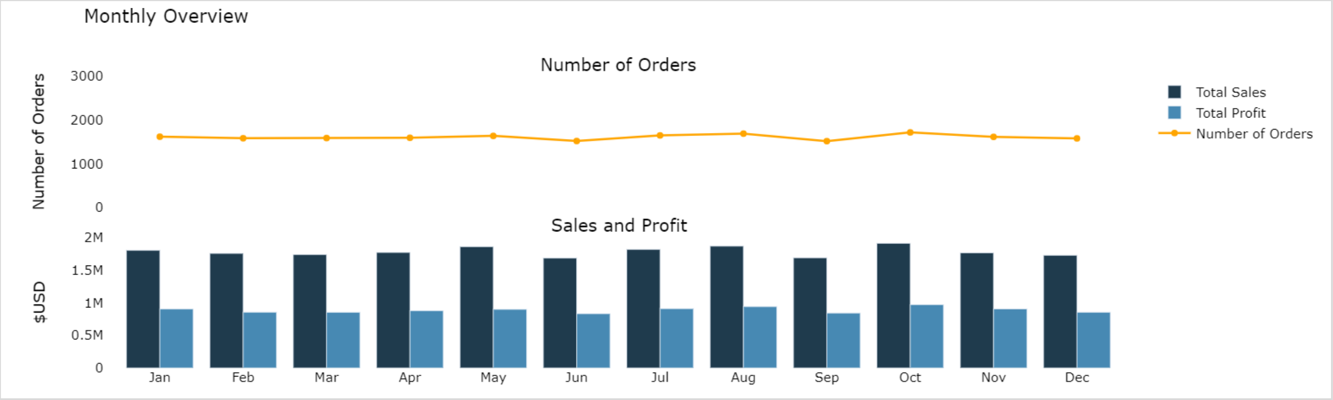

Figure 1 – Sales and Orders: Monthly Trend

On average, there were approximately **1,610 orders per month**, generating a total sales revenue of around $1.78 million and a total profit of approximately $886,301. Despite the fact, that the lowest number of orders is observed in September (1,518 transactions), the minimum values of sales and profit are observed in June. October had the highest number of orders, resulting in the highest total sales revenue and total profit during this period.

Overall, the standard deviation for the sales indicators is relatively low, indicating consistent transaction activity. The analysis reveals that October experienced the most robust transaction activity, leading to the highest revenue and profit, while June witnessed comparatively lower transaction performance.

3.4. PRODUCTS AND BRANDS

Product analysis involves examining the characteristics and performance of the products within the dataset. The main focus maid on the popularity of products, their profitability and customer preferences. The overall product performance can be seen on the following figure:

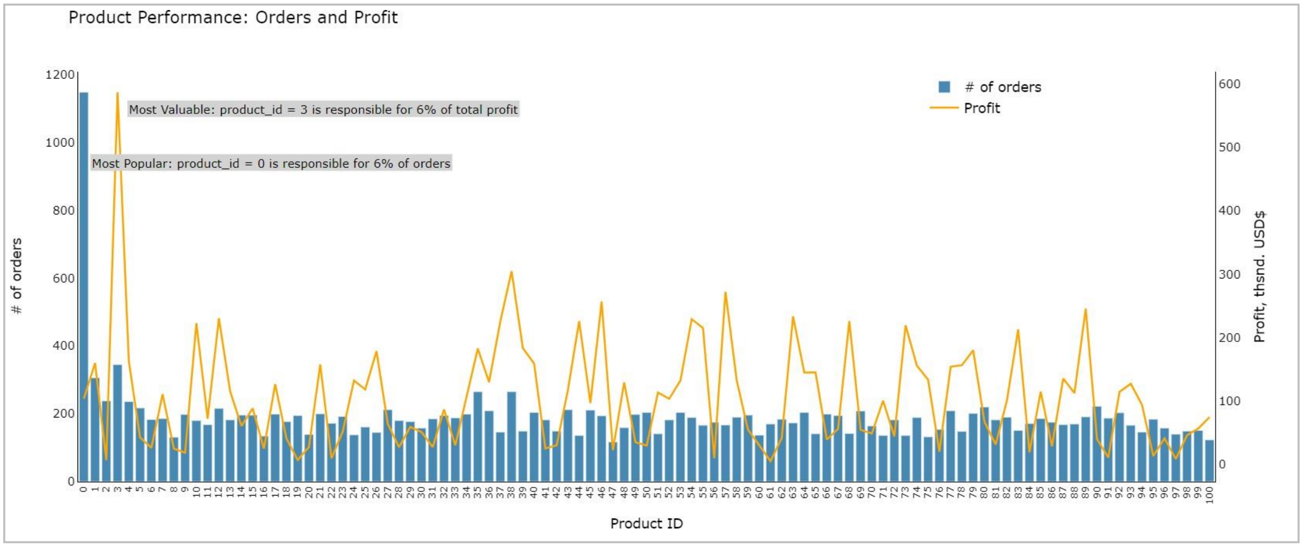

Figure 2 – Product Performance: Popularity and Profit

As can be seen on the figure above, product_id=0 stands out as the most popular item, contributing to a significant 6% of all orders. However, its contribution to the total income remains below 1%, emphasizing its wide customer appeal. On the other hand, product_id=3 emerges as the most valuable asset, accounting for an impressive 6% of the total income, reflecting its significant contribution to the organization's revenue stream.

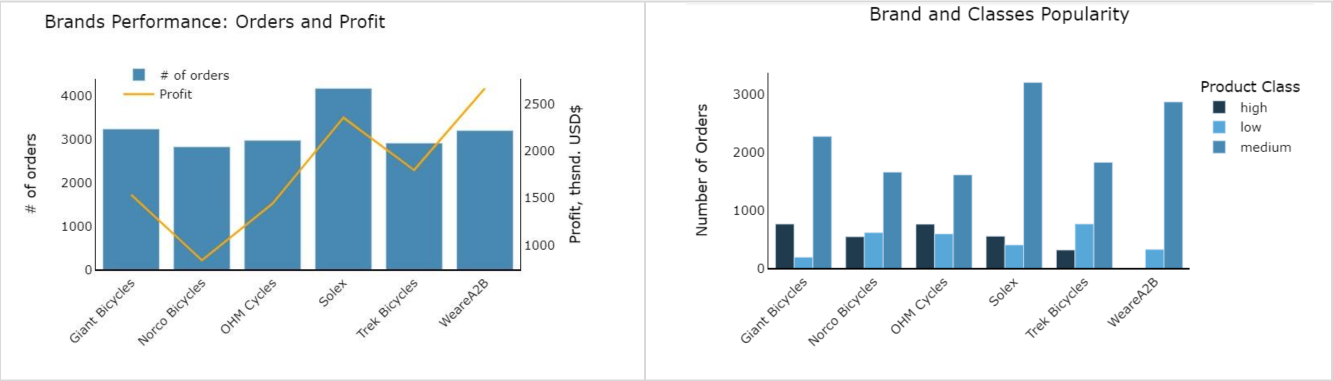

Figure 3 – Brands Performance: Popularity and Profit

Additionally, the data indicates that the "medium" product_class holds both the highest popularity and value across all brands, making it a noteworthy choice for potential marketing strategies.

The analysis reveals that the brand "Silex" is the most popular among customers, capturing the highest number of orders. On the other hand, "WeareA2B" emerges as the most valuable brand in terms of total profit generated. The Norco Bicycles considered to be the less popular brand with the least contribution to the profit.

3.5. ORDERS AND CUSTOMERS

The analysis involves studying customer behavior, preferences, and interactions to gain insights that can drive marketing, sales, and customer service strategies. The chart below describes Distribution of all Orders by number of categories: Customer Gender and Age groups, State, Job Industry:

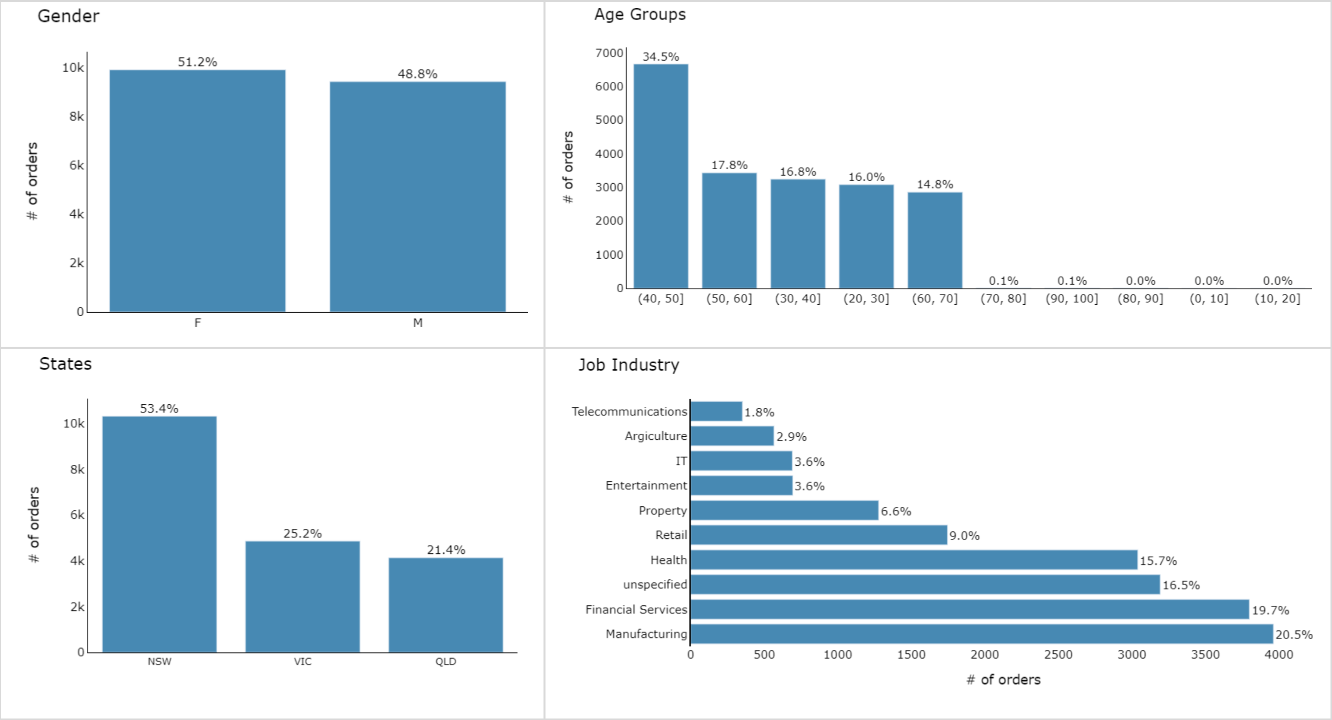

Figure 4 – Orders Distribution: Customers Gender, Age, State and Job Industry

The analysis of gender-based transactions showed that female customers slightly outpaced male customers over the past year, accounting for 51.2% and 48.8% of orders respectively. When considering age groups, customers aged between 40 and 60 years were responsible for the highest percentage of orders at 52%, closely followed by the 20 - 40 age group, contributing to 33% of orders.

In regards with customer`s location the NSQ accounted for the majority of orders at 53.4%, with VIC and QLD following at 25.2% and 21.4% respectively. The job industries distribution shows that although 16% of customers did not specify their associated industry, the most prevalent industries were Manufacturing and Financial Services, each representing 20% of total orders.

The Mass Customers segment emerges as the most valuable across various age groups, contributing significantly to both the number of orders and overall profit. The exception in the 80-100 age group highlights a unique trend in wealth distribution. Moreover, Mass Customers maintain their status as the primary category for both genders and across different states, underscoring their pivotal role in driving success across these dimensions. More Customer and Orders related distributions are shown on the figures below.

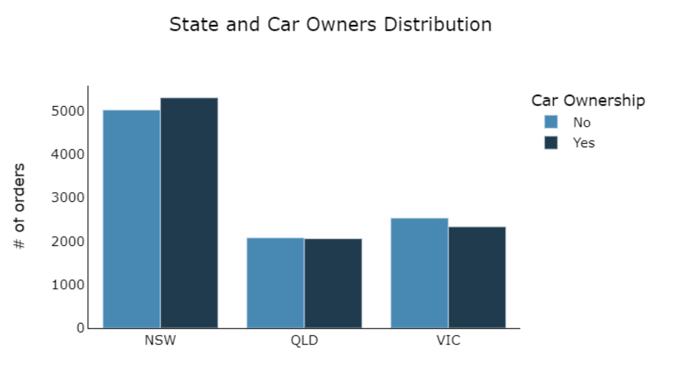

Figure 5 –Car Owners Distribution by States

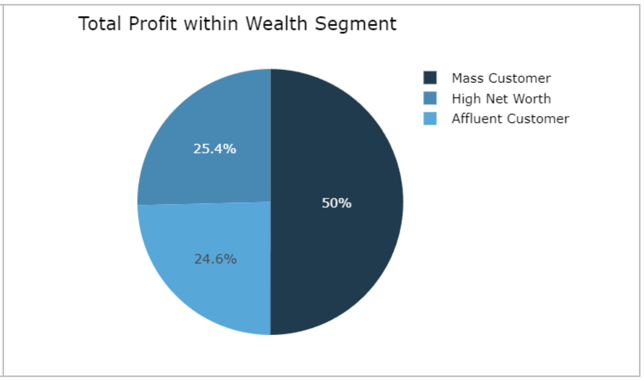

Figure 6 – Wealth Segment by total Profit

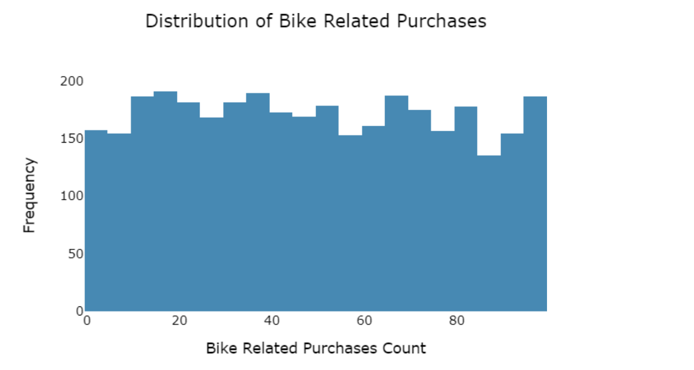

Figure 7 – Past 3 years Bike-related purchases distribution

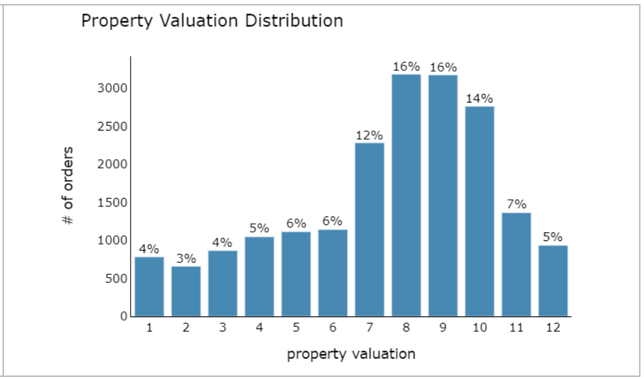

Figure 8– Property Valuation distribution

In conclusion, the ownership of cars across all states is nearly evenly split between car owners and non-car owners. Examining bike-related purchases over the past three years, the frequency distribution of purchases is consistent across all ranges except for the least common range of 85-89 purchases, contrasted with the most common range of 15-19 purchases. Lastly, half of the customers fall within the property valuation range of 7 to 11, signifying a significant middle ground. These insights offer valuable guidance for optimizing marketing strategies, tailoring product offerings, and targeting specific customer segments for enhanced business outcomes.

The analysis reveals several key insights about our customer base and transaction trends:

- The analysis reveals several key insights about our customer base and transaction trends;
- Age groups between 40 and 60 years were the most active in terms of number of transactions;
- Geographically, NSQ dominated in orders;
- Manufacturing and Financial Services emerged as the top industries;
- Mass Customers stood out across age groups as valuable contributors to orders and profits;
- Car ownership is balanced across states;
- Bike-related purchases varied, with 15-19 purchases being the most frequent;
- Customers within property valuation range of 7 to 11 constitute a substantial segment.