Assignment Report:
Understanding Customer Purchase Behaviour at 2Market

Introduction

2Market is a global supermarket chain operating both online and in-store, aiming to enhance its understanding of customer purchase behaviour.

Problem Statement: The company seeks to identify the demographics of its customers, determine the most effective advertising channels, and understand which products sell best among different demographic groups.

The key objectives of this analysis are to segment customers based on demographics, evaluate the performance of various advertising channels, and analyse product sales trends in relation to customer demographics. This information will support strategic decision-making to improve marketing efforts and product offerings.

Analytical Approach

To address 2Market's objectives, I utilised PostgreSQL to perform data analysis on two datasets: marketing_data.csv and ad_data.csv. The first step involved data cleaning and preparation. I imported the datasets into SQL tables and ensured data types were correctly assigned—for instance, converting income fields to numeric values by stripping currency symbols and commas, and parsing date fields was done while creating the table.

```
--Converting Income to Numeric Values

UPDATE public."Marketing_Data"
SET "Income" = REPLACE(REPLACE("Income", '$', ''), ',', '')::NUMERIC;

--Checking for Missing Values

SELECT *
FROM public."Marketing_Data"
WHERE "Income" IS NULL OR "Year_Birth" IS NULL;
```

I performed data validation checks to handle missing or inconsistent values. For example, I checked for nulls in critical fields like Income and Year_Birth, and assessed the distribution of categorical variables such as Education and Marital_Status. Any anomalies were documented and addressed accordingly.

```
--Handling Categorical Variables

SELECT DISTINCT "Education", COUNT("ID") FROM public."Marketing_Data" GROUP BY "Education";
SELECT DISTINCT "Marital_Status", COUNT("ID") FROM public."Marketing_Data" GROUP BY "Marital_Status";
```

In the demographic segmentation, I calculated customer ages by subtracting the “Year_Birth” from the current year. I then categorised customers into age groups: Young (<30), Middle-Aged (30-50), and Senior (>50). Similarly, I segmented income into brackets: Low Income (<$30,000), Middle Income ($30,000-$60,000), and High Income (>$60,000). These segments allowed me to perform group-wise analyses.

```
--Calculating Age, Creating Age Groups and Segmenting Income into Brackets
SELECT
  "ID",
  EXTRACT(YEAR FROM CURRENT_DATE) - "Year_Birth" AS Age,
  CASE
    WHEN EXTRACT(YEAR FROM CURRENT_DATE) - "Year_Birth" < 30 THEN 'Young'
    WHEN EXTRACT(YEAR FROM CURRENT_DATE) - "Year_Birth" BETWEEN 30 AND 50 THEN 'Middle-Aged'
    ELSE 'Senior'
  END AS "Age_Group",
  "Income",
  CASE
    WHEN CAST("Income" AS DECIMAL) < 30000 THEN 'Low Income'
    WHEN CAST("Income" AS DECIMAL) BETWEEN 30000 AND 60000 THEN 'Middle Income'
    ELSE 'High Income'
  END AS "Income_Bracket"
FROM public."Marketing_Data";
```

For the advertising channel effectiveness, I joined the marketing_data and ad_data tables on the ID field. I calculated conversion rates for each advertising channel by dividing the number of successful conversions by the total number of customers. This involved aggregating the boolean fields (Bulkmail_ad, Twitter_ad, etc.) to determine total conversions per channel.

```
-- Join marketing_data and ad_data on ID
SELECT
  m."ID",
  COALESCE(CASE WHEN a."Bulkmail_ad" THEN 1 ELSE 0 END, 0) AS "Bulkmail_Conversions",
  COALESCE(CASE WHEN a."Twitter_ad" THEN 1 ELSE 0 END, 0, 0) AS "Twitter_Conversions",
  COALESCE(CASE WHEN a."Instagram_ad" THEN 1 ELSE 0 END, 0, 0) AS "Instagram_Conversions",
  COALESCE(CASE WHEN a."Facebook_ad" THEN 1 ELSE 0 END, 0, 0) AS "Facebook_Conversions",
  COALESCE(CASE WHEN a."Brochure_ad" THEN 1 ELSE 0 END, 0, 0) AS "Brochure_Conversions"
FROM public."Marketing_Data" m
LEFT JOIN public."Ad_Data" a
ON m."ID" = a."ID";
```

```sql
SELECT 'Bulkmail' AS "Channel",
    SUM(CASE WHEN "Bulkmail_ad" THEN 1 ELSE 0 END) * 100.0 / COUNT(*) AS "Conversion_Rate"
FROM
    public."Ad_Data"
UNION ALL
SELECT 'Twitter' AS "Channel",
    SUM(CASE WHEN "Twitter_ad" THEN 1 ELSE 0 END) * 100.0 / COUNT(*) AS "Conversion_Rate"
FROM
    public."Ad_Data"
UNION ALL
SELECT 'Instagram' AS "Channel",
    SUM(CASE WHEN "Instagram_ad" THEN 1 ELSE 0 END) * 100.0 / COUNT(*) AS "Conversion_Rate"
FROM
    public."Ad_Data"
UNION ALL
SELECT 'Facebook' AS "Channel",
    SUM(CASE WHEN "Facebook_ad" THEN 1 ELSE 0 END) * 100.0 / COUNT(*) AS "Conversion_Rate"
FROM
    public."Ad_Data"
UNION ALL
SELECT 'Brochure' AS "Channel",
    SUM(CASE WHEN "Brochure_ad" THEN 1 ELSE 0 END) * 100.0 / COUNT(*) AS "Conversion_Rate"
FROM
    public."Ad_Data";

--Total Conversions by Channel

SELECT 'Bulkmail' AS "Channel",
    SUM(CASE WHEN "Bulkmail_ad" THEN 1 ELSE 0 END) AS "Total_Conversions"
FROM
    public."Ad_Data"
UNION ALL
SELECT 'Twitter' AS "Channel",
    SUM(CASE WHEN "Twitter_ad" THEN 1 ELSE 0 END) AS "Total_Conversions"
FROM
    public."Ad_Data"
UNION ALL
SELECT 'Instagram' AS "Channel",
    SUM(CASE WHEN "Instagram_ad" THEN 1 ELSE 0 END) AS "Total_Conversions"
FROM
    public."Ad_Data"
```

In analysing product spending, I summed the amounts spent on each product category (AmtLiq, AmtVege, AmtNonVeg, AmtPes, AmtChocolates, AmtComm) and grouped the results by demographic segments. This allowed me to identify which products were most popular among different age groups, income levels, and family sizes (number of kids and teenagers at home).

```sql
SELECT
  CASE
    WHEN (EXTRACT(YEAR FROM CURRENT_DATE) - "Year_Birth") < 30 THEN 'Young'
    WHEN (EXTRACT(YEAR FROM CURRENT_DATE) - "Year_Birth") BETWEEN 30 AND 50 THEN 'Middle-Aged'
    ELSE 'Senior'
  END AS "Age_Group",
  CASE
    WHEN CAST("Income" AS DECIMAL) < 30000 THEN 'Low Income'
    WHEN CAST("Income" AS DECIMAL) BETWEEN 30000 AND 60000 THEN 'Middle Income'
    ELSE 'High Income'
  END AS "Income_Segment",
  SUM("AmtLiq") AS "Total_Spent_On_Alcohol",
  SUM("AmtVege") AS "Total_Spent_On_Vegetables",
  SUM("AmtNonVeg") AS "Total_Spent_On_Meat",
  SUM("AmtPes") AS "Total_Spent_On_Fish",
  SUM("AmtChocolates") AS "Total_Spent_On_Chocolates",
  SUM("AmtComm") AS "Total_Spent_On_Commodities"
FROM public."Marketing_Data"
GROUP BY "Age_Group", "Income_Segment";
```

```sql
-- Analyze spending based on number of kids and teenagers
SELECT
  "Kidhome",
  "Teenhome",
  SUM("AmtLiq") AS "Total_Spent_On_Alcohol",
  SUM("AmtVege") AS "Total_Spent_On_Vegetables",
  SUM("AmtNonVeg") AS "Total_Spent_On_Meat",
  SUM("AmtPes") AS "Total_Spent_On_Fish",
  SUM("AmtChocolates") AS "Total_Spent_On_Chocolates",
  SUM("AmtComm") AS "Total_Spent_On_Commodities"
FROM public."Marketing_Data"
GROUP BY "Kidhome", "Teenhome";
```

During the analysis, I faced challenges such as ensuring accurate age calculations, considering the current date versus the data's timeframe, etc. I also had to address potential outliers in income and spending amounts, which could skew average values. To mitigate this, I used median values and percentile distributions where appropriate.

By performing these SQL queries and aggregations, I was able to extract meaningful insights from the data, which informed the subsequent dashboard design and recommendations.

Upon further analysis, I tried to efficiently explore the relationship between demographic factors and customer responsiveness to different advertising channels, while identifying the top products across customer segments.

I created a new table called Customer_Ad_Performance, this query combines customer demographic information with advertising performance data using the common ID field:

```
CREATE TABLE public."Customer_Ad_Performance" AS
SELECT
    m."ID",
    m."Year_Birth",
    m."Education",
    m."Marital_Status",
    m."Income",
    m."Country",
    a."Bulkmail_ad",
    a."Twitter_ad",
    a."Instagram_ad",
    a."Facebook_ad",
    a."Brochure_ad"
FROM
    public."Marketing_Data" m
JOIN
    public."Ad_Data" a
ON
    m."ID" = a."ID";
```

Then I tried to get some insights into customer demographics by country, average income, and total ad conversions.

```
SELECT
    "Country",
    COUNT(*) AS "Total_Customers",
    AVG(CAST("Income" AS DECIMAL)) AS "Avg_Income",
    SUM(CASE WHEN "Bulkmail_ad" THEN 1 ELSE 0 END
        + CASE WHEN "Twitter_ad" THEN 1 ELSE 0 END
        + CASE WHEN "Instagram_ad" THEN 1 ELSE 0 END
        + CASE WHEN "Facebook_ad" THEN 1 ELSE 0 END
        + CASE WHEN "Brochure_ad" THEN 1 ELSE 0 END) AS "Total_Conversions"
FROM
    public."Customer_Ad_Performance"
GROUP BY
    "Country";
```

I then identified which product categories are favoured by different demographic groups based on total spending. I also did a similar analysis using “Income” and ended up combining the queries for effective analysis:

```
SELECT
    CASE
        WHEN CAST(REPLACE(REPLACE(REPLACE("Income", '$', ''), ',', ''), '.00', '') AS INTEGER) <= 30000 THEN
        'Low Income (<=30K)'
        WHEN CAST(REPLACE(REPLACE(REPLACE("Income", '$', ''), ',', ''), '.00', '') AS INTEGER) BETWEEN 30001 AND 60000 THEN
        'Mid Income (30K-60K)'
        WHEN CAST(REPLACE(REPLACE(REPLACE("Income", '$', ''), ',', ''), '.00', '') AS INTEGER) BETWEEN 60001 AND 90000 THEN
        'High Income (60K-90K)'
        ELSE 'Very High Income (>90K)'
    END AS "Income_Bracket",
    CASE
        WHEN EXTRACT(YEAR FROM CURRENT_DATE) - "Year_Birth" BETWEEN 18 AND 29 THEN '18-29'
        WHEN EXTRACT(YEAR FROM CURRENT_DATE) - "Year_Birth" BETWEEN 30 AND 39 THEN '30-39'
        WHEN EXTRACT(YEAR FROM CURRENT_DATE) - "Year_Birth" BETWEEN 40 AND 49 THEN '40-49'
        ELSE '50+'
    END AS "Age_Group",
    COALESCE(SUM("AmtLiq"), 0) AS "Total_Liquor_Sales",
    COALESCE(SUM("AmtVege"), 0) AS "Total_Vegetable_Sales",
    COALESCE(SUM("AmtNonVeg"), 0) AS "Total_Meat_Sales",
    COALESCE(SUM("AmtPes"), 0) AS "Total_Fish_Sales",
    COALESCE(SUM("AmtChocolates"), 0) AS "Total_Chocolate_Sales",
    COALESCE(SUM("AmtComm"), 0) AS "Total_Commodity_Sales"
FROM
    public."Marketing_Data"
GROUP BY
    "Income_Bracket",
    "Age_Group"
ORDER BY
    "Income_Bracket",
    "Age_Group";
```

For an even more in-depth analysis, I created a query to provide a summary of conversion rates and total conversions, to measure the effectiveness of each advertising channel (the next 2 screenshots).

```
WITH "Conversion_Rates" AS (
    SELECT 'Bulkmail' AS "Channel",
        SUM(CASE WHEN "Bulkmail_ad" THEN 1 ELSE 0 END) * 100.0 / COUNT(*) AS "Conversion_Rate"
    FROM
        public."Ad_Data"
    UNION ALL
    SELECT 'Twitter' AS "Channel",
        SUM(CASE WHEN "Twitter_ad" THEN 1 ELSE 0 END) * 100.0 / COUNT(*) AS "Conversion_Rate"
    FROM
        public."Ad_Data"
    UNION ALL
    SELECT 'Instagram' AS "Channel",
        SUM(CASE WHEN "Instagram_ad" THEN 1 ELSE 0 END) * 100.0 / COUNT(*) AS "Conversion_Rate"
    FROM
        public."Ad_Data"
    UNION ALL
    SELECT 'Facebook' AS "Channel",
        SUM(CASE WHEN "Facebook_ad" THEN 1 ELSE 0 END) * 100.0 / COUNT(*) AS "Conversion_Rate"
    FROM
        public."Ad_Data"
    UNION ALL
    SELECT 'Brochure' AS "Channel",
        SUM(CASE WHEN "Brochure_ad" THEN 1 ELSE 0 END) * 100.0 / COUNT(*) AS "Conversion_Rate"
    FROM
        public."Ad_Data"
),
"Total_Conversions" AS (
    SELECT 'Bulkmail' AS "Channel",
        SUM(CASE WHEN "Bulkmail_ad" THEN 1 ELSE 0 END) AS "Total_Conversions"
    FROM
        public."Ad_Data"
    UNION ALL
    SELECT 'Twitter' AS "Channel",
        SUM(CASE WHEN "Twitter_ad" THEN 1 ELSE 0 END) AS "Total_Conversions"
    FROM
        public."Ad_Data"
    UNION ALL
    SELECT 'Instagram' AS "Channel",
        SUM(CASE WHEN "Instagram_ad" THEN 1 ELSE 0 END) AS "Total_Conversions"
    FROM
        public."Ad_Data"
```

```
    FROM
        public."Ad_Data"
    UNION ALL
    SELECT 'Facebook' AS "Channel",
        SUM(CASE WHEN "Facebook_ad" THEN 1 ELSE 0 END) AS "Total_Conversions"
    FROM
        public."Ad_Data"
    UNION ALL
    SELECT 'Brochure' AS "Channel",
        SUM(CASE WHEN "Brochure_ad" THEN 1 ELSE 0 END) AS "Total_Conversions"
    FROM
        public."Ad_Data"
)
SELECT
    c."Channel",
    c."Conversion_Rate",
    t."Total_Conversions"
FROM
    "Conversion_Rates" c
JOIN
    "Total_Conversions" t
ON
    c."Channel" = t."Channel";
```

The above queries helped in providing more information to visualise the dashboard and its design.

Dashboard Design and Development

The dashboard design incorporates new visual elements while maintaining the focus on accessibility, interactivity, and clarity to serve 2Market's business stakeholders effectively. Below are the revised elements:

1. **Demographic overview (Average income by education level)**:
   Visual: A vertical bar chart, where education levels are displayed on the x-axis (rows), and the average income is plotted on the y-axis (columns).
   Rationale: A bar chart is ideal for comparing categorical data like education levels. Stakeholders can easily assess how education correlates with income among the customer base. This insight is valuable for refining targeted marketing strategies, understanding which education groups have higher purchasing power, and crafting tailored messaging for each segment.

2. **Customer spend across product categories (Product spending trends)**:
   Visual: An area chart illustrating total customer spending over time, with the x-axis representing customer joining dates and the y-axis displaying total spend across all categories.
   Rationale: The area chart is suitable for displaying spending trends over time, allowing stakeholders to see how customer spending has evolved since joining 2Market. This helps identify spending patterns, seasonal spikes, and long-term growth. It also assists in making decisions about customer retention strategies, predicting future revenue streams, and adjusting product promotions based on trends.

3. **Advertising channel effectiveness** :
   Visual: A bar chart where the y-axis lists various advertising channels (e.g., Bulkmail, Twitter, Instagram, Facebook, Brochure), and the x-axis shows the successful conversions for each channel.
   Rationale: A straightforward bar chart allows stakeholders to quickly grasp the effectiveness of each advertising channel by comparing total conversions. This visual is crucial for optimising ad spend and ensuring resources are allocated to the most impactful channels. It provides clear data for determining which platforms are driving the most conversions and where additional marketing investment may be needed.

Dashboard Integration

The three visualisations are integrated into a single dashboard with a universal filter for "Country." This filter enables stakeholders to adjust the visuals dynamically and focus on the data for specific regions, providing a localised perspective on demographic income, spending behaviour, and advertising performance.

Design Principles:

- **Colour**: Differentiating colours are used consistently across the dashboard to help distinguish between categories. For example, individual education levels, advertising channels, and spending patterns each have their distinct colour schemes. This helps avoid confusion and promotes easy interpretation.
- **Layout**: The dashboard layout remains intuitive, with charts arranged logically from demographic insights to spending behaviour and ad performance. This progression mirrors the flow of business insights, enabling users to understand customer profiles before analysing their behaviour and the effectiveness of advertising efforts.
- **Interactivity**: The universal filter on "Country" ensures that all visuals update in real-time based on the region selected. This interactivity fosters a deeper understanding of customer behaviour across different geographical segments and provides tailored insights that can drive region-specific strategies.

Patterns, Trends, and Insights

Several key patterns emerged from the analysis:

1. Demographic Insights: Customers aged between 30-50, particularly those with higher income levels, contribute significantly to the total sales. Married customers and those with children also tend to spend more, especially on commodities like meat and fish.
2. Advertising Effectiveness: Among the advertising channels, Facebook and Instagram showed the highest conversion rates, particularly for younger customers. Brochure ads, on the other hand, performed well with older demographics.
3. Product Performance: Alcoholic beverages and meat products are the top-performing categories across most customer segments, with these categories showing consistent high spending across age groups. However, younger customers tend to spend more on vegetables and chocolates.

Recommendations:

- Increase investment in Facebook and Instagram ads, as they drive the highest conversions among the most valuable customer segments (age 30-50).
- Consider targeted promotions on alcoholic beverages and meat products to maximise sales, while also expanding offerings in high-growth areas such as chocolates for younger customers.
- Use brochures more effectively by targeting older customers who respond well to this medium.

These findings offer actionable insights that can enhance 2Market’s marketing strategies and improve overall profitability.