Sales Performance Analysis and Customer Segmentation for a Bicycle Retail Client

Scenario

I am acting as a junior data analyst at a business intelligence consulting firm. Our new client is a bicycle retail company with multiple branches that wants to leverage their data to grow their business. I have been assigned to lead this project, starting from defining the business problem to presenting strategic recommendations based on the data.

Step 1: ASK

This stage focuses on understanding and defining the business problem and the goals the client wants to achieve.

Business Task

Analyze historical sales data to identify key trends, top-performing products and stores, as well as customer behavior. The goal is to provide high-level strategic recommendations to the client's executive and marketing teams in order to increase revenue and marketing effectiveness.

Guiding Questions

This analysis will be guided by the following questions:

- What are the key trends in product usage and sales across all stores?
- How can these trends be applied to the existing customer base to boost sales?
- How can insights from this data help inform and influence the client's future marketing strategy?

Stakeholders

- **Client's Executive Team**: Requires a high-level summary to support strategic decision-making regarding business growth.
- **Client's Marketing Team**: Needs in-depth insights into customer behavior to design more effective and targeted marketing campaigns.

Deliverable

A well-defined business statement summarizing the objective of this project

Step 2: PREPARE

This phase involves gathering, assessing, and preparing the data to be used for analysis.

Data Source

This analysis uses the public dataset titled **"Bike Store Sample Database"**, available on Kaggle. The dataset consists of **9 separate .csv files** containing data related to:

- **Customers**
- **Orders (orders, order_items)**
- **Products (products, categories, brands)**
- **Operations (stores, staffs, stocks)**

Data Storage and Organization

- **Storage**: The dataset downloaded from Kaggle was first stored locally. Then, the data was uploaded to **Google BigQuery**, a cloud-based data warehouse selected for its capability to efficiently handle complex SQL queries across multiple tables.
- **Organization**: Each CSV file was uploaded as a separate table within a single **BigQuery dataset named** `Bikesales`. This structure preserves the integrity of the raw data before any merging or transformation is performed.

Data Credibility and Integrity (ROCCC)

- **Reliable**: The dataset is fairly reliable as it is well-structured and complete. However, it has limitations as it is a static sample dataset, not derived from a live production environment.
- **Original**: The data source is clearly identified as coming from the Kaggle platform.
- **Comprehensive**: The dataset covers multiple business areas (sales, products, customers, operations), making it highly relevant for answering various business questions.
- **Current**: The data is historical and not real-time. This is a limitation because the trends discovered may not fully reflect the current market conditions.
- **Cited**: The source of the dataset has been clearly referenced.

Final Deliverable

A complete description of all data sources used in the analysis.

Step 3: PROCESS

This stage focuses on cleaning and transforming the raw data into a tidy, analysis-ready format.

Tool Used

Google BigQuery and **SQL** were used for processing. SQL is particularly effective for cleaning, transforming, and joining data from multiple relational tables.

Cleaning & Transformation Documentation

The following SQL-based steps were taken:

- **Null Value Checks**: Ensured key columns (e.g., `order_id`, `customer_id`) did not contain missing values.
- **Duplicate Checks**: Verified primary key uniqueness (e.g., `order_id`) across tables to prevent double counting.

- **Data Type Validation**: Confirmed appropriate data formats:
  - Date columns in `DATE` format.
  - Price and quantity columns in numeric formats.
- **Data Transformation (JOINs)**:

  All nine tables were joined into a single master table named `sales_transactions`. This step was essential to simplify the analytical process and enable insights across multiple business dimensions.

```
select * from `Bikesales.customers`;
select * from `Bikesales.order_items`;
select * from `Bikesales.orders`;
select * from `Bikesales.products`;
select * from `Bikesales.stocks`;

SELECT
  COUNT(*) AS total_rows,
  COUNTIF(product_name IS NULL) AS product_name_nulls,
  COUNTIF(list_price IS NULL) AS list_price_nulls
FROM `Bikesales.products`;

SELECT
  COUNT(*) AS total_rows,
  COUNTIF(first_name IS NULL) AS nama_depan,
  COUNTIF(last_name IS NULL) AS nama_belakang,
  COUNTIF(city IS NULL) AS kota,
  COUNTIF(state IS NULL) AS bagian,
FROM `Bikesales.customers`;

SELECT
  COUNT(*) AS total_rows,
  COUNTIF(quantity IS NULL) AS jumlah,
  COUNTIF(list_price IS NULL) AS harga,
FROM `Bikesales.order_items`;

SELECT
```

```
  COUNT(*) AS total_rows,
  COUNTIF(order_status IS NULL) AS status,
  COUNTIF(order_date IS NULL) AS tglpesan,
  COUNTIF(shipped_date IS NULL) AS tgl_pengiriman,
  COUNTIF(required_date IS NULL) AS tgl,
FROM `Bikesales.orders`;

SELECT
  COUNT(*) AS total_rows,
  COUNTIF(quantity IS NULL) AS jmlh,
FROM `Bikesales.stocks`;

-- Memeriksa duplikasi pada customer_id
SELECT
    customer_id,
    COUNT(customer_id) AS jumlah_duplikat
FROM
    `Bikesales.customers`
GROUP BY
    customer_id
HAVING
    COUNT(customer_id) > 1;

--Penggabungan untuk membuat table Sales Transaction

CREATE OR REPLACE TABLE `Bikesales.sales_transactions` AS
SELECT
    ord.order_id,
    cus.customer_id,
    ord.order_date,
    ord.required_date,
    ord.shipped_date,
    ROUND(IFNULL((ite.quantity * ite.list_price * (1 - ite.discount)), 0), 2) AS
    ite.quantity,
    cus.first_name,
    cus.last_name,
    cus.city AS customer_city,
    cus.state AS customer_state,
```

```
    pro.product_name,
    cat.category_name,
    bra.brand_name,
    sto.store_name,
    sta.first_name AS staff_first_name,
    sta.last_name AS staff_last_name
FROM
    `Bikesales.orders` AS ord
LEFT JOIN
    `Bikesales.order_items` AS ite ON ord.order_id = ite.order_id
LEFT JOIN
    `Bikesales.customers` AS cus ON ord.customer_id = cus.customer_id
LEFT JOIN
    `Bikesales.products` AS pro ON ite.product_id = pro.product_id
LEFT JOIN
    `Bikesales.categories` AS cat ON pro.category_id = cat.category_id
LEFT JOIN
    `Bikesales.brands` AS bra ON pro.brand_id = bra.brand_id
LEFT JOIN
    `Bikesales.stores` AS sto ON ord.store_id = sto.store_id
LEFT JOIN
    `Bikesales.staffs` AS sta ON ord.staff_id = sta.staff_id;

select * from `Bikesales.sales_transactions`;


-- Membuat tabel RFM
-- Membuat tabel RFM dengan segmentasi pelanggan
CREATE OR REPLACE TABLE `Bikesales.rfm_scores` AS
WITH rfm_base AS (
    SELECT
        CONCAT(first_name, ' ', last_name, ' - ', customer_city) AS customer_
        DATE_DIFF((SELECT MAX(order_date) FROM `Bikesales.sales_transac
        COUNT(DISTINCT order_id) AS frequency,
        SUM(total_price) AS monetary
    FROM
        `Bikesales.sales_transactions`
    GROUP BY
```

```
        customer_id
),
rfm_scores AS (
    SELECT
        customer_id,
        recency,
        frequency,
        monetary,
        -- Skor Recency: makin kecil recency makin bagus (DESC)
        NTILE(5) OVER (ORDER BY recency DESC) AS R_score,
        -- Skor Frequency: makin besar frequency makin bagus (ASC)
        NTILE(5) OVER (ORDER BY frequency ASC) AS F_score,
        -- Skor Monetary: makin besar monetary makin bagus (ASC)
        NTILE(5) OVER (ORDER BY monetary ASC) AS M_score
    FROM
        rfm_base
),
rfm_labeled AS (
    SELECT
        *,
        -- Gabungan skor jadi satu string, misal: 555, 431, dst.
        CAST(R_score AS STRING) || CAST(F_score AS STRING) || CAST(M_s

        -- Segmentasi pelanggan berdasarkan kombinasi skor RFM
        CASE
            WHEN R_score >= 4 AND F_score >= 4 AND M_score >= 4 THEN '
            WHEN R_score >= 4 AND F_score >= 3 THEN 'Loyal Customers'
            WHEN R_score >= 4 AND F_score <= 2 THEN 'Potential Loyalist'
            WHEN R_score = 5 AND F_score = 1 THEN 'Recent Customers'
            WHEN R_score BETWEEN 3 AND 4 AND F_score <= 2 THEN 'Prom
            WHEN R_score = 3 AND F_score = 3 THEN 'Needs Attention'
            WHEN R_score <= 2 AND F_score >= 4 THEN 'At Risk'
            WHEN R_score = 1 AND F_score = 1 THEN 'Lost'
            WHEN R_score <= 2 AND F_score <= 2 AND M_score <= 2 THEN '
            ELSE 'Others'
        END AS segment
    FROM
        rfm_scores
```

```
)

SELECT * FROM rfm_labeled;
```

Deliverable

A step-by-step record of the cleaning and transformation process.

Step 4: ANALYZE

At this stage, the cleaned data is analyzed to identify trends, patterns, and insights that address key business questions.

Summary of Analysis

Sales Performance Analysis

- **Total Revenue & Customers:**

  The total recorded revenue is **$7,689,110**, generated from **1,445 unique customers**.

- **Monthly Revenue Trends:**

  Data from **January 2016 to November 2018** shows fluctuations in monthly revenue. A significant revenue peak occurred in **May 2018**, exceeding **$800,000**, which stood out from the rest of the timeline.

- **Top Product Categories:**

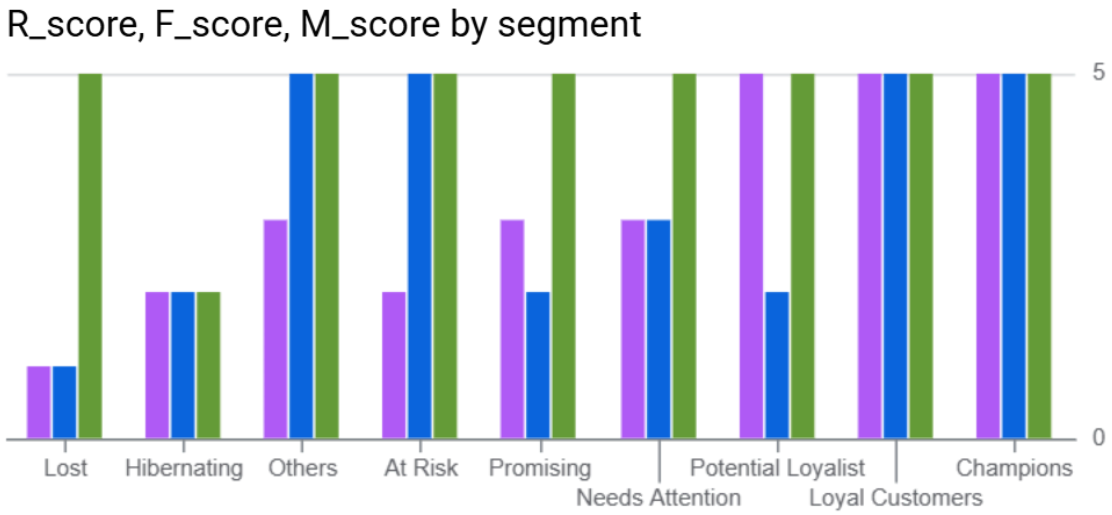
  Based on total revenue, the **Mountain Bikes** category leads the sales performance, followed by **Road Bikes**. **Cruiser Bicycles** also contributed significantly to overall revenue.

- **Customer Distribution:**

  Customers are geographically distributed across multiple states, with the highest concentrations in **California** and **New York**, accounting for a combined **1,019 customers**.

Customer Segmentation Analysis (RFM)

Customers were segmented using the **Recency (R)**, **Frequency (F)**, and **Monetary (M)** metrics. The RFM segmentation identified key customer groups, including:



- **Champions:** Customers scoring the highest (5) in Recency, Frequency, and Monetary.
- **At Risk:** Customers with high Frequency and Monetary scores but low Recency, indicating they haven't engaged recently.
- Other segments identified include **Loyal Customers**, **Potential Loyalists**, **Needs Attention**, **Promising**, **Hibernating**, and **Lost** customers.

Surprising Finding

One of the most notable insights from the analysis was the **sharp revenue spike in May 2018**, which appeared highly anomalous.

- **Anomalous Trend:**

  This spike was extremely distinct from regular monthly fluctuations and was followed by a rapid drop in revenue in the subsequent months, nearing zero.

- **Single-Event Pattern:**

The pattern did not repeat in the same months across other years, suggesting it was **a one-time event** rather than a seasonal trend.

- **Implications:**

  This phenomenon warrants further investigation. Potential causes include:

  - A single **large-volume bulk order**
  - A **highly successful, short-lived marketing campaign**
  - Or even a **data entry anomaly**

  Understanding the root cause is crucial to determine whether this spike represents a **repeatable success** or a **data quality issue**.

Final Deliverable

A concise summary of key analytical insights that address the primary business questions, providing data-driven recommendations for decision-making.

Step 5: SHARE

This stage focuses on presenting analytical findings to stakeholders through effective visualizations and compelling narratives.

Supporting Visualizations

To communicate the findings, an interactive *dashboard* was created using Looker Studio. This *dashboard* allows stakeholders to explore the data independently.

Dashboard Development in Looker Studio

Looker Studio was chosen due to its seamless connectivity with BigQuery and its capability to create interactive, easily shareable reports.

Final Deliverable

An interactive *dashboard* in Looker Studio presenting key insights and supporting visualizations.

Bike_Sales_Dashboard.pdf

https://lookerstudio.google.com/reporting/3d7e1875-7573-4a7e-b43f-4985251096c4

Step 6: ACT

The final step involves translating insights into concrete, actionable business recommendations to drive sales growth and inform future marketing strategies.

Business Recommendations

1. What are the key trends in product usage and sales?

Based on the data, several key trends were identified:

- **Dominance of Specific Product Categories:** Sales are consistently dominated by *Mountain Bikes*, the top revenue contributor, followed by *Road Bikes*. This indicates that the current business focus and market demand are centered around these two types of bikes.
- **Revenue Volatility with a Major Anomaly:** Monthly revenue trends show fluctuations, with a significant spike in May 2018. This spike is not part of a normal seasonal pattern, but rather an anomaly suggesting a large-scale sales event or a highly successful short-term campaign.
- **Geographic Concentration:** Customer distribution is uneven, with very high concentrations in key states like California and New York, where the number of customers reaches up to 1,019. This indicates that the current market is heavily geographically concentrated.

2. How can these trends be applied to the existing customer base to boost sales?

These trends can be leveraged to increase sales from the existing customer base through the following strategies:

- **Apply RFM Segmentation for Personalization:**

- For the *"Champions"* and *"Loyal Customers"* segments (high R, F, and M scores): Offer early access to new products (especially new Mountain and Road Bike models), exclusive loyalty programs, and special promotions to retain them.
- For the *"At Risk"* segment (low R score, high F & M scores): These customers previously made large purchases but haven't returned recently. Launch targeted *re-engagement* campaigns with personalized discounts for categories they previously purchased.
- For *"Promising"* or *"Potential Loyalists"*: These are newer or frequent buyers with lower monetary value. Encourage them to increase purchase value through product bundles (e.g., bike + accessories).

- **Focus on Cross-selling and Upselling:**

  Since most customers are interested in Mountain and Road Bikes, offer complementary products such as helmets, cycling apparel, or spare parts. Promote upgraded models to customers who already own basic versions.

3. How can these insights inform and influence future marketing strategies?

Insights from the data provide a solid foundation for smarter, more efficient marketing strategies:

- **Investigate the Sales Anomaly as a Top Priority:**

  The marketing and sales teams should investigate the cause of the revenue spike in May 2018. Was it due to a viral campaign, a large B2B deal, or a data error? If it was a successful strategy, it should be analyzed for replication. If it was a B2B sale, it may indicate a new market opportunity worth developing.

- **Focused Marketing Budget Allocation:**

  Rather than spreading the budget evenly, concentrate marketing resources where ROI is highest:

  - **Product Focus:** Create campaigns specifically highlighting the strengths of Mountain Bikes and Road Bikes.
  - **Geographic Focus:** Intensify digital and local marketing efforts in states with the highest customer concentration.

- **Develop Segment-Based Marketing Strategies:**

Shift from a one-size-fits-all approach to a personalized strategy. Use RFM data to create tailored messages for each segment, such as a *“We Miss You”* campaign for the *At Risk* group and *“VIP Offers”* for the *Champions* segment.

Final Deliverable

A list of high-level strategic recommendations based on data analysis.