Sales Data Analysis

September 16, 2025

[1]:
```
# 1. Install required packages
# Run this once in your environment
# pip install kaggle pandas sqlite3

import os
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import sqlite3
# %pip install seaborn
```

[2]:
```
print(os.getcwd())
```

```
/Users/mac2025/Desktop/Continuous
Education/Scaler/Sessions/Portfolio_projects/BeginnerSQLProject
```

[5]:
```
# 3. Load CSV into pandas
df = pd.read_csv('/Users/mac2025/Desktop/Continuous Education/Scaler/Sessions/
 ↪Portfolio_projects/BeginnerSQLProject/Sales Data.csv')
```

[6]:
```
conn = sqlite3.connect('SalesData.db')
df.to_sql('SalesData', conn, index=False, if_exists='replace')
```

[6]:
```
185950
```

[18]:
```
def query_and_print(sql):
    result = pd.read_sql(sql, conn)
    print(result, '\n')

analysis_queries = {
    'Segment customers by total spending': '''
        SELECT
          "Purchase Address" AS customer_identifier,
          SUM("Sales") AS total_spent,
          CASE
            WHEN SUM("Sales") > 1000 THEN 'High-value'
            WHEN SUM("Sales") BETWEEN 500 AND 1000 THEN 'Medium-value'
            ELSE 'Low-value'
```

```
        END AS spending_segment
      FROM SalesData
      GROUP BY "Purchase Address"
      ORDER BY total_spent DESC;
  ''',
  'Segment customers by purchase frequency': '''
      SELECT
        "Purchase Address" AS customer_identifier,
        COUNT("Order ID") AS purchase_count,
        CASE
          WHEN COUNT("Order ID") > 20 THEN 'Frequent Buyer'
          WHEN COUNT("Order ID") BETWEEN 10 AND 20 THEN 'Moderate Buyer'
          ELSE 'Occasional Buyer'
        END AS frequency_segment
      FROM SalesData
      GROUP BY "Purchase Address"
      ORDER BY purchase_count DESC;
  ''',
  'Segment customers top product category': '''
      SELECT
        "Purchase Address" AS customer_identifier,
        Product,
        COUNT("Order ID") AS purchases
      FROM SalesData
      GROUP BY "Purchase Address", Product
      ORDER BY "Purchase Address", purchases DESC;
  ''',
  'Segment customers by last purchase date': '''
      SELECT
        "Purchase Address",
        MAX("Order Date") AS last_purchase_date,
        CASE
          WHEN MAX("Order Date") > DATE('now', '-30 day') THEN 'Recent'
          WHEN MAX("Order Date") BETWEEN DATE('now', '-90 day') AND␣
↪DATE('now', '-31 day') THEN 'Dormant'
          ELSE 'Inactive'
        END AS recency_segment
      FROM SalesData
      GROUP BY "Purchase Address"
      ORDER BY last_purchase_date DESC;
  ''',
  'Segment customers RFM': '''
      WITH rfm AS (
        SELECT
          "Purchase Address",
          MAX("Order Date") AS last_purchase_date,
          COUNT("Order ID") AS frequency,
```

```
      SUM("Sales") AS monetary
    FROM SalesData
    GROUP BY "Purchase Address"
  )
  SELECT
    "Purchase Address",
    CASE
      WHEN last_purchase_date > DATE('now', '-30 day') THEN 'Recent'
      ELSE 'Lapsed'
    END AS recency,
    CASE
      WHEN frequency > 20 THEN 'Frequent'
      WHEN frequency BETWEEN 10 AND 20 THEN 'Moderate'
      ELSE 'Rare'
    END AS frequency_segment,
    CASE
      WHEN monetary > 1000 THEN 'High Spender'
      WHEN monetary BETWEEN 500 AND 1000 THEN 'Medium Spender'
      ELSE 'Low Spender'
    END AS monetary_segment
  FROM rfm
  ORDER BY monetary DESC;
''',
'Top 10 Customers by revenue and purchase frequency': '''
  WITH customer_metrics AS (
    SELECT
      "Purchase Address",
      SUM("Sales") AS total_revenue,
      COUNT("Order ID") AS purchase_count
    FROM SalesData
    GROUP BY "Purchase Address"
  ),
  ranked_customers AS (
    SELECT
      "Purchase Address",
      total_revenue,
      purchase_count,
      NTILE(10) OVER (ORDER BY total_revenue DESC) AS revenue_decile,
      NTILE(10) OVER (ORDER BY purchase_count DESC) AS frequency_decile
    FROM customer_metrics
  )
  SELECT
    "Purchase Address",
    total_revenue,
    purchase_count
  FROM ranked_customers
  WHERE revenue_decile = 1
```

```
        AND frequency_decile = 1
      ORDER BY total_revenue DESC, purchase_count DESC
      LIMIT 10;
  ''',
  'Cluster Customers': '''
      WITH customer_category_summary AS (
        SELECT
          "Purchase Address",
          Product,
          SUM("Sales") AS category_spent,
          COUNT("Order ID") AS category_purchases
        FROM SalesData
        GROUP BY "Purchase Address", Product
      ),
      customer_ltv AS (
        SELECT
          "Purchase Address",
          SUM("Sales") AS lifetime_value
        FROM SalesData
        GROUP BY "Purchase Address"
      ),
      category_counts AS (
        SELECT
          "Purchase Address",
          COUNT(DISTINCT Product) AS distinct_categories
        FROM SalesData
        GROUP BY "Purchase Address"
      ),
      customer_clusters AS (
        SELECT
          c."Purchase Address",
          l.lifetime_value,
          cc.distinct_categories,
          CASE
            WHEN l.lifetime_value > 1000 THEN 'High LTV'
            WHEN l.lifetime_value BETWEEN 500 AND 1000 THEN 'Medium LTV'
            ELSE 'Low LTV'
          END AS ltv_segment,
          CASE
            WHEN cc.distinct_categories > 5 THEN 'Diverse Buyer'
            WHEN cc.distinct_categories BETWEEN 2 AND 5 THEN 'Moderate Buyer'
            ELSE 'Niche Buyer'
          END AS category_segment
        FROM customer_category_summary c
        JOIN customer_ltv l ON c."Purchase Address" = l."Purchase Address"
        JOIN category_counts cc ON c."Purchase Address" = cc."Purchase␣
↪Address"
```

```
        GROUP BY c."Purchase Address", l.lifetime_value, cc.
↪distinct_categories
      )
      SELECT *
      FROM customer_clusters;
    '''
}
```

[19]:
```
for desc, sql in analysis_queries.items():
    print(f'-- {desc} --')
    query_and_print(sql)

conn.close()
```

```
-- Segment customers by total spending --
                        customer_identifier  total_spent spending_segment
0       668 Park St, San Francisco, CA 94016      4379.99       High-value
1             795 1st St, Atlanta, GA 30301      4100.00       High-value
2             391 1st St, Seattle, WA 98101      4100.00       High-value
3         10 1st St, San Francisco, CA 94016      4000.00       High-value
4       731 11th St, New York City, NY 10001      3919.88       High-value
…                                       …           …              …
140782      1 Willow St, Portland, OR 97035         2.99        Low-value
140783   1 Lincoln St, Los Angeles, CA 90001         2.99        Low-value
140784    1 Lake St, New York City, NY 10001         2.99        Low-value
140785  1 Church St, New York City, NY 10001         2.99        Low-value
140786           1 4th St, Seattle, WA 98101         2.99        Low-value

[140787 rows x 3 columns]

-- Segment customers by purchase frequency --
                          customer_identifier  purchase_count  \
0       193 Forest St, San Francisco, CA 94016               9
1       279 Sunset St, San Francisco, CA 94016               8
2            223 Elm St, Los Angeles, CA 90001               8
3          727 9th St, San Francisco, CA 94016               7
4          716 5th St, San Francisco, CA 94016               7
…                                        …              …
140782      1 12th St, New York City, NY 10001               1
140783        1 12th St, Los Angeles, CA 90001               1
140784      1 11th St, San Francisco, CA 94016               1
140785        1 11th St, Los Angeles, CA 90001               1
140786            1 11th St, Atlanta, GA 30301               1

       frequency_segment
0       Occasional Buyer
1       Occasional Buyer
2       Occasional Buyer
```

```
3       Occasional Buyer
4       Occasional Buyer
…                  …
140782  Occasional Buyer
140783  Occasional Buyer
140784  Occasional Buyer
140785  Occasional Buyer
140786  Occasional Buyer

[140787 rows x 3 columns]

-- Segment customers top product category --
                          customer_identifier                     Product  \
0                1 11th St, Atlanta, GA 30301        USB-C Charging Cable
1            1 11th St, Los Angeles, CA 90001          Macbook Pro Laptop
2          1 11th St, San Francisco, CA 94016                      iPhone
3            1 12th St, Los Angeles, CA 90001    Apple Airpods Headphones
4          1 12th St, New York City, NY 10001            Wired Headphones
…                                        …                         …
181494       999 Wilson St, Atlanta, GA 30301  Bose SoundSport Headphones
181495   999 Wilson St, Los Angeles, CA 90001             ThinkPad Laptop
181496  999 Wilson St, New York City, NY 10001    Apple Airpods Headphones
181497      999 Wilson St, Portland, OR 97035      AAA Batteries (4-pack)
181498  999 Wilson St, San Francisco, CA 94016    Apple Airpods Headphones

        purchases
0               1
1               1
2               1
3               1
4               1
…             …
181494          1
181495          1
181496          1
181497          1
181498          1

[181499 rows x 3 columns]

-- Segment customers by last purchase date --
                             Purchase Address   last_purchase_date  \
0        657 Spruce St, New York City, NY 10001  2020-01-01 05:13:00
1         784 River St, San Francisco, CA 94016  2020-01-01 04:54:00
2       754 Hickory St, New York City, NY 10001  2020-01-01 04:21:00
3             825 Adams St, Portland, OR 97035  2020-01-01 04:13:00
4         202 Maple St, San Francisco, CA 94016  2020-01-01 04:06:00
…                                          …                  …
```

```
140782            232 12th St, Boston, MA 02215  2019-01-01 06:41:00
140783            943 2nd St, Atlanta, GA 30301  2019-01-01 06:03:00
140784      735 5th St, New York City, NY 10001  2019-01-01 04:56:00
140785   760 Church St, San Francisco, CA 94016  2019-01-01 03:40:00
140786       9 Lake St, New York City, NY 10001  2019-01-01 03:07:00

       recency_segment
0             Inactive
1             Inactive
2             Inactive
3             Inactive
4             Inactive
…                ⋯
140782        Inactive
140783        Inactive
140784        Inactive
140785        Inactive
140786        Inactive

[140787 rows x 3 columns]

-- Segment customers RFM --
                           Purchase Address recency frequency_segment  \
0       668 Park St, San Francisco, CA 94016  Lapsed              Rare
1              391 1st St, Seattle, WA 98101  Lapsed              Rare
2              795 1st St, Atlanta, GA 30301  Lapsed              Rare
3        10 1st St, San Francisco, CA 94016  Lapsed              Rare
4       731 11th St, New York City, NY 10001  Lapsed              Rare
…                                        …     …                 …
140782  999 Center St, Los Angeles, CA 90001  Lapsed              Rare
140783      999 Church St, Atlanta, GA 30301  Lapsed              Rare
140784        999 Main St, Atlanta, GA 30301  Lapsed              Rare
140785         999 West St, Dallas, TX 75001  Lapsed              Rare
140786     999 Wilson St, Portland, OR 97035  Lapsed              Rare

       monetary_segment
0          High Spender
1          High Spender
2          High Spender
3          High Spender
4          High Spender
…                 ⋯
140782      Low Spender
140783      Low Spender
140784      Low Spender
140785      Low Spender
140786      Low Spender
```

```
[140787 rows x 4 columns]

-- Top 10 Customers by revenue and purchase frequency --
                         Purchase Address  total_revenue  purchase_count
0      668 Park St, San Francisco, CA 94016        4379.99               3
1            391 1st St, Seattle, WA 98101        4100.00               3
2            795 1st St, Atlanta, GA 30301        4100.00               3
3       10 1st St, San Francisco, CA 94016        4000.00               3
4      731 11th St, New York City, NY 10001        3919.88               5
5  208 Chestnut St, San Francisco, CA 94016        3789.99               3
6   949 Hickory St, New York City, NY 10001        3779.99               3
7    611 Wilson St, San Francisco, CA 94016        3718.78               6
8        610 14th St, Los Angeles, CA 90001        3699.98               3
9      256 Hill St, San Francisco, CA 94016        3561.95               4

-- Cluster Customers --
                            Purchase Address  lifetime_value  \
0                1 11th St, Atlanta, GA 30301           11.95
1             1 11th St, Los Angeles, CA 90001         1700.00
2           1 11th St, San Francisco, CA 94016          700.00
3             1 12th St, Los Angeles, CA 90001          150.00
4           1 12th St, New York City, NY 10001           11.99
…                                         …             …
140782        999 Wilson St, Atlanta, GA 30301           99.99
140783    999 Wilson St, Los Angeles, CA 90001          999.99
140784  999 Wilson St, New York City, NY 10001          150.00
140785       999 Wilson St, Portland, OR 97035            2.99
140786  999 Wilson St, San Francisco, CA 94016          150.00

        distinct_categories ltv_segment category_segment
0                         1     Low LTV      Niche Buyer
1                         1    High LTV      Niche Buyer
2                         1  Medium LTV      Niche Buyer
3                         1     Low LTV      Niche Buyer
4                         1     Low LTV      Niche Buyer
…                       …         …              …
140782                    1     Low LTV      Niche Buyer
140783                    1  Medium LTV      Niche Buyer
140784                    1     Low LTV      Niche Buyer
140785                    1     Low LTV      Niche Buyer
140786                    1     Low LTV      Niche Buyer

[140787 rows x 5 columns]
```

[ ]: