

SQL + PYTHON

E-COMMERCE

SALES ANALYSIS



Data Analyst: Punit Pal

🛠 Tools Used:

- MySQL / MySQL Workbench
- Python
- Pandas, NumPy
- Matplotlib / Seaborn
- Jupyter Notebook

🧪 Techniques Used:

- SQL Joins (INNER, LEFT, RIGHT)
- Aggregations (SUM, COUNT, AVG)
- Grouping & Ordering
- Window Functions (Moving Average, Ranking, Partitioning)
- CTE (Common Table Expressions)
- Time-Series Analysis
- Correlation Analysis
- Category & Revenue Analysis
- Customer Retention Calculation

🎯 Project Focus:

A complete end-to-end marketplace analysis uncovering insights about:

- Customer behavior
- Product performance
- Seller activity
- Logistics efficiency
- Revenue trends
- Payment behavior
- Yearly & monthly sales trends

PROJECT OVERVIEW

This project performs a comprehensive analysis of a real-world E-Commerce / Marketplace dataset using MySQL and Python.

The dataset contains detailed customer, seller, order, payment, product, and location information.

It answers key business questions related to:

- Customer distribution and behavior
- Order trends across different time periods
- Product and category performance
- Seller performance and logistics metrics
- Payment patterns and revenue contribution
- Customer retention and repeat purchase cycles
- Year-over-year sales patterns

Goal:

The goal of this project is to uncover actionable insights that help an E-Commerce business improve:

- Customer experience
- Seller performance
- Revenue growth
- Logistics execution
- Category strategy
- Overall business decision-making



PROJECT OBJECTIVES

Basic Analysis

- List all unique customer locations
- Count total orders in specific years/months
- Calculate total sales generated per category
- Measure percentage of installment-based payments
- Count customers state-wise

Intermediate Analysis

- Monthly order trends (e.g., orders per month in 2018)
- Average number of products purchased per order
- Revenue contribution percentage by each category
- Category vs shipping cost comparison
- Correlation between product price and purchase volume
- Seller-wise total revenue ranking

Advanced Analysis

- Moving average of order value (customer-wise)
- Cumulative monthly sales across years
- Year-over-year revenue growth
- Customer retention rate (repeat purchase within 6 months)
- Identify top 3 high-value customers per year

Data Ingestion Pipeline (CSV → SQL)



```
1 import pandas as pd
2 import mysql.connector
3 import os
4 from mysql.connector import Error
5
6 # List of CSV files and their corresponding table names
7 csv_files = [
8     ('customers.csv', 'customers'),
9     ('orders.csv', 'orders'),
10    ('sellers.csv', 'sellers'),
11    ('products.csv', 'products'),
12    ('order_items.csv', 'order_items'),
13    ('payments.csv', 'payments'),
14    ('geolocation.csv', 'geolocation')
15 ]
16
17 # Folder containing the CSV files
18 folder_path = 'C:/Users/punitpal/OneDrive/Desktop/project_ws'
19
20 def get_sql_type(dtype):
21     if pd.api.types.is_integer_dtype(dtype):
22         return 'INT'
23     elif pd.api.types.is_float_dtype(dtype):
24         return 'FLOAT'
25     elif pd.api.types.is_bool_dtype(dtype):
26         return 'BOOLEAN'
27     elif pd.api.types.is_datetime64_any_dtype(dtype):
28         return 'DATETIME'
29     else:
30         return 'TEXT'
31
32 try:
33     # Connect to the MySQL database
34     conn = mysql.connector.connect(
35         host='localhost',
36         user='root',
37         password='MYfriNK@2024',
38         database='ecommerce'
39     )
40     cursor = conn.cursor()
41
42     for csv_file, table_name in csv_files:
43         try:
44             file_path = os.path.join(folder_path, csv_file)
45
46             # Check if file exists
47             if not os.path.exists(file_path):
48                 print(f"Warning: File {csv_file} not found, skipping.")
49                 continue
50
```

```

50
51     # Read the CSV file into a pandas DataFrame
52     df = pd.read_csv(file_path)
53
54     # Replace NaN with None to handle SQL NULL
55     df = df.where(pd.notnull(df), None)
56
57     # Debugging: Check for NaN values
58     print(f"Processing {csv_file}")
59     print(f"NaN values before replacement:\n{df.isnull().sum()}\n")
60
61     # Clean column names
62     df.columns = [col.replace(' ', '_').replace('-', '_').replace('.', '_') for col in df.columns]
63
64     # Generate the CREATE TABLE statement with appropriate data types
65     columns = ', '.join([f'{col} {get_sql_type(df[col].dtype)}' for col in df.columns])
66     create_table_query = f'CREATE TABLE IF NOT EXISTS `{table_name}` ({columns})'
67     cursor.execute(create_table_query)
68
69     # Prepare batch insert
70     batch_size = 1000 # Adjust based on your data size
71     values_list = []
72
73     for _, row in df.iterrows():
74         # Convert row to tuple and handle NaN/None explicitly
75         values = tuple(None if pd.isna(x) else x for x in row)
76         values_list.append(values)
77

```

punitpalofficial@gmail.com

```
78         # Execute batch insert when batch size is reached
79         if len(values_list) >= batch_size:
80             sql = f"INSERT INTO `table_name` ({', '.join(['`' + col + '`' for col in df.columns])}) VALUES ({', '.join(['%' + str(i) + '%' for i in range(len(df.columns))])})"
81             cursor.executemany(sql, values_list)
82             conn.commit()
83             values_list = []
84
85         # Insert remaining records
86         if values_list:
87             sql = f"INSERT INTO `table_name` ({', '.join(['`' + col + '`' for col in df.columns])}) VALUES ({', '.join(['%' + str(i) + '%' * len(df.columns) for i in range(len(values_list))])}"
88             cursor.executemany(sql, values_list)
89             conn.commit()
90
91             print(f"Successfully imported {csv_file} into {table_name}")
92
93     except Exception as e:
94         print(f"Error processing {csv_file}: {str(e)}")
95         conn.rollback() # Rollback changes if error occurs
96
97     except Error as e:
98         print(f"Error connecting to MySQL database: {str(e)}")
99
100 finally:
101     # Close the connection
102     if 'conn' in locals() and conn.is_connected():
103         cursor.close()
104     conn.close()
105     print("MySQL connection closed.")
```

Basic Queries

1. List all unique cities where customers are located.



```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import seaborn as sns
4 import mysql.connector
5
6 db=mysql.connector.connect(
7     host="localhost",
8     user="root",
9     password="MYfrinK@2024",
10    database="ecommerce"
11 )
12 cursor=db.cursor()
13 query = """ select distinct customer_city from customers; """
14 cursor.execute(query)
15 data = cursor.fetchall()
16 df=pd.DataFrame(data)
17 df.head()
```

Output:

0

0

franca

1

sao bernardo do campo

2

sao paulo

3

mogi das cruzes

4

campinas

2. Count the number of orders placed in 2017.



```
1 import mysql.connector
2
3 db=mysql.connector.connect(
4     host="localhost",
5     user="root",
6     password="MYfrinK@2024",
7     database="ecommerce"
8 )
9 cursor=db.cursor()
10 query = """
11     SELECT COUNT(order_id) as order_count
12     FROM orders
13     WHERE YEAR(order_purchase_timestamp) = 2017;
14 """
15 cursor.execute(query)
16 data = cursor.fetchall()
17 print(f"Number of orders in 2017: {data[0][0]}")
```

Output:

*** Number of orders in 2017: 416431

3. Find the total sales per category.



```
1 import pandas as pd
2 import mysql.connector
3
4 db=mysql.connector.connect(
5     host="localhost",
6     user="root",
7     password="MYfrInK@2024",
8     database="ecommerce"
9 )
10 cursor=db.cursor()
11 query = """ select upper(products.product_category) category , round(sum(payments.payment_value),2) sales
12 from products join order_items
13 on products.product_id=order_items.product_id
14 join payments
15 on payments.order_id=order_items.order_id
16 group by category ;
17 """
18 cursor.execute(query)
19 data=cursor.fetchall()
20 df = pd.DataFrame(data, columns=['Category', 'Sales'])
21 df
```

Output:

...	Category	Sales
0	PERFUMERY	4.763096e+07
1	FURNITURE DECORATION	1.350661e+08
2	TELEPHONY	4.587931e+07
3	BED TABLE BATH	1.607873e+08
4	AUTOMOTIVE	7.999573e+07
...
69	CDS MUSIC DVDS	1.157232e+05
70	LA CUISINE	2.795922e+05
71	FASHION CHILDREN'S CLOTHING	6.764840e+04
72	PC GAMER	2.109288e+05
73	INSURANCE AND SERVICES	3.014200e+04

74 rows × 2 columns

4. Calculate the percentage of orders that were paid in installments.



```
1 import mysql.connector
2
3 db = mysql.connector.connect(
4     host="localhost",
5     user="root",
6     password="MYfriNK@2024",
7     database="ecommerce"
8 )
9
10 cursor = db.cursor()
11
12 query = """
13     SELECT upper(products.product_category) AS category,
14             round(SUM(payments.payment_value),2) AS sales
15     FROM products
16     JOIN order_items
17         ON products.product_id = order_items.product_id
18     JOIN payments
19         ON payments.order_id = order_items.order_id
20     GROUP BY category;
21 """
22
23 cursor.execute(query)
24 data = cursor.fetchall()
25 cursor.close()
26 db.close()
27 df = pd.DataFrame(data, columns=['Category', 'Sales'])
28 df.head()
```

Output:

	Category	Sales
0	PERFUMERY	4.763096e+07
1	FURNITURE DECORATION	1.350661e+08
2	TELEPHONY	4.587931e+07
3	BED TABLE BATH	1.607873e+08
4	AUTOMOTIVE	7.999573e+07

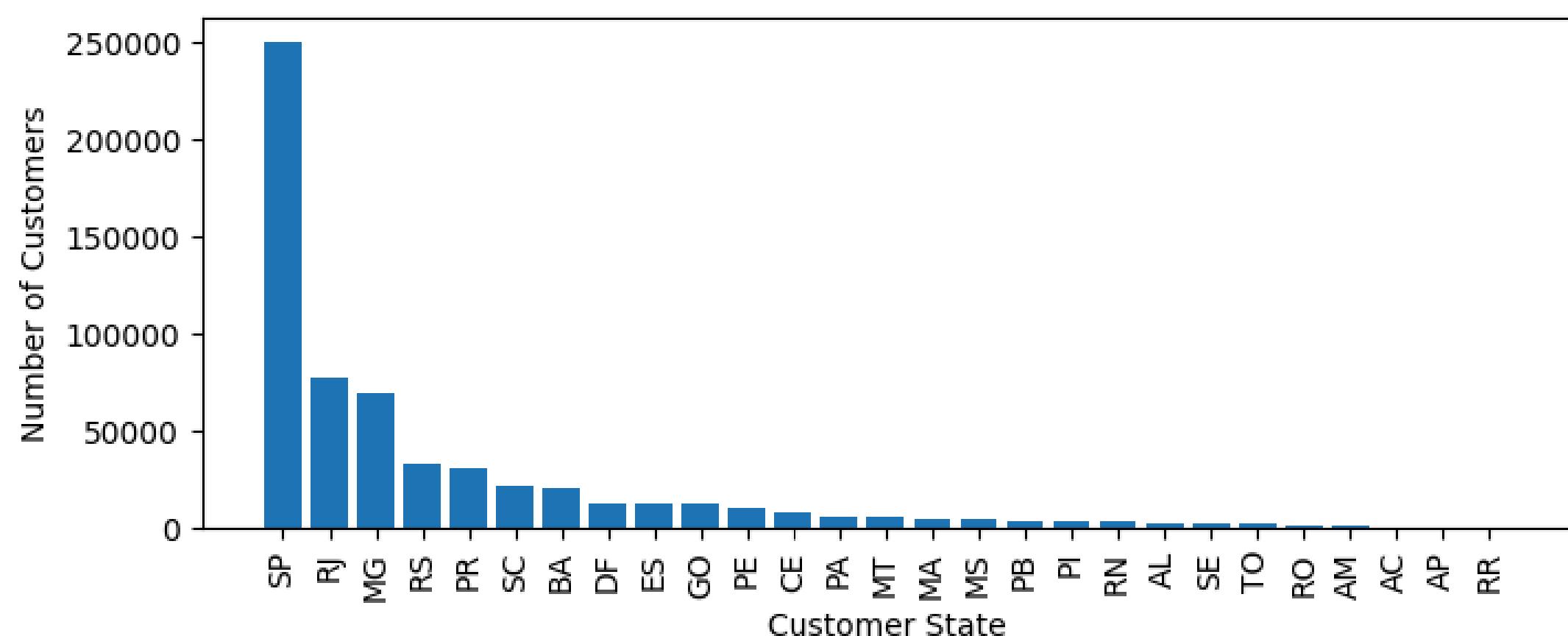
5. Count the number of customers from each state.



```
1 import mysql.connector
2
3 db = mysql.connector.connect(
4     host="localhost",
5     user="root",
6     password="MYfriNK@2024",
7     database="ecommerce"
8 )
9
10 cursor = db.cursor()
11
12 query = """
13 select customer_state, count(customer_id)
14 from customers group by customer_state;
15 """
16
17 cursor.execute(query)
18 data = cursor.fetchall()
19 data
20 df=pd.DataFrame(data, columns=['Customer_State', 'Customer_Count'])
21 df=df.sort_values(by='Customer_Count', ascending=False)
22 # df
23 plt.figure(figsize=(8,3))
24 plt.bar(df['Customer_State'], df['Customer_Count'])
25 plt.xlabel('Customer State')
26 plt.ylabel('Number of Customers')
27 plt.title('Number of Customers from Each State')
28 plt.xticks(rotation=90)
29 plt.show()
```

Output:

Number of Customers from Each State



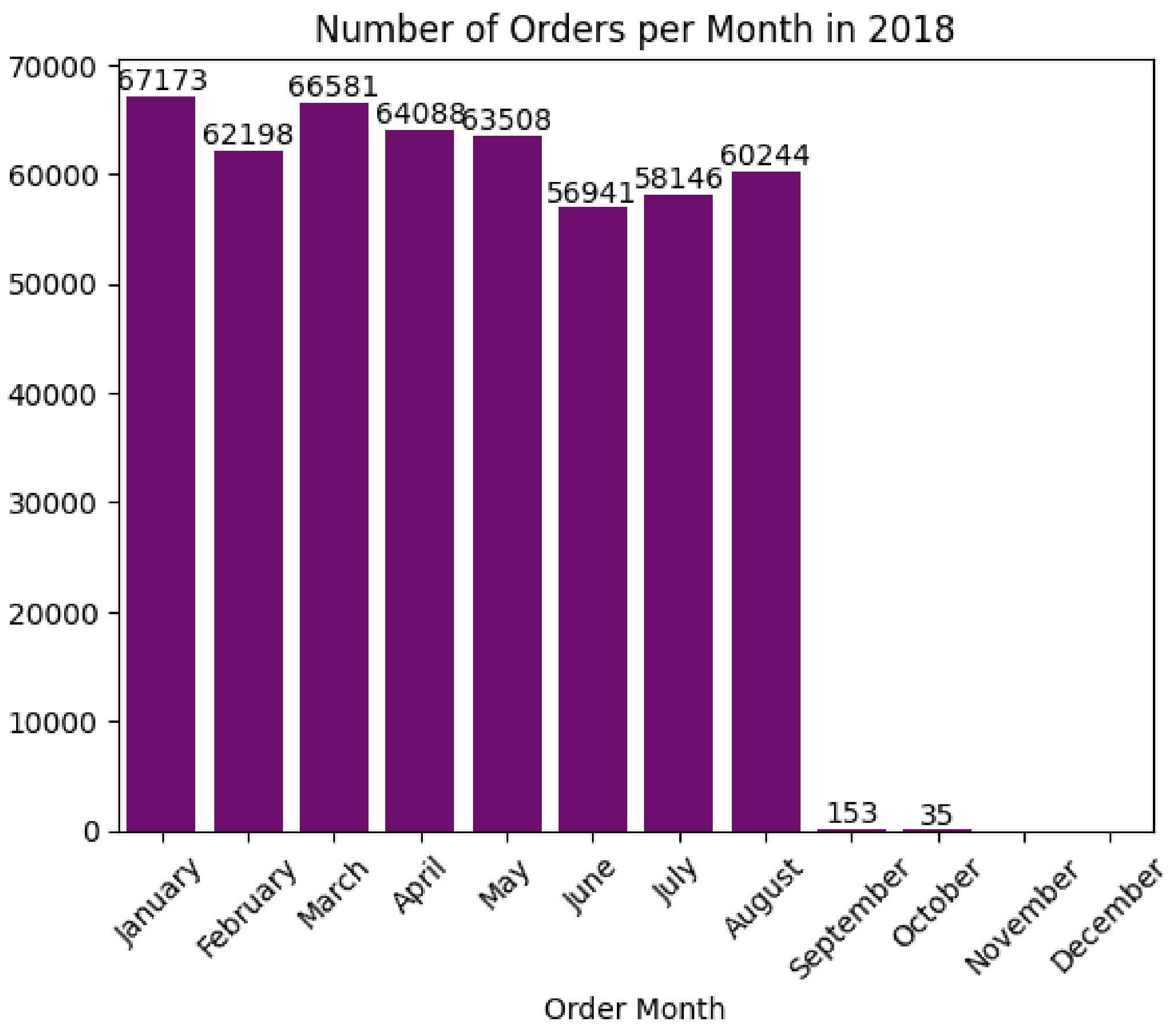
Intermediate Queries

1. Calculate the number of orders per month in 2018.



```
1 import mysql.connector
2
3 db = mysql.connector.connect(
4     host="localhost",
5     user="root",
6     password="MYfriNK@2024",
7     database="ecommerce"
8 )
9
10 cursor = db.cursor()
11
12 query = """
13 select monthname(order_purchase_timestamp) as order_month,
14 count(order_id) as order_count
15 from orders where year(order_purchase_timestamp)=2018
16 group by order_month;
17 """
18
19 cursor.execute(query)
20 data = cursor.fetchall()
21 df=pd.DataFrame(data, columns=['Order_Month', 'Order_Count'])
22 o= ["January", "February", "March", "April", "May", "June", "July",
23      "August", "September", "October", "November", "December"]
24 ax=sns.barplot(x=df["Order_Month"], y=df["Order_Count"],
25                  data=df, order=o, color='purple')
26 plt.xlabel('Order Month')
27 plt.ylabel('Order Count')
28 plt.xticks(rotation=45)
29 ax.bar_label(ax.containers[0])
30 plt.title('Number of Orders per Month in 2018')
31 plt.show()
```

Output:



2. Find the average number of products per order, grouped by customer city.



```
1 import pandas as pd
2 import mysql.connector
3
4 db = mysql.connector.connect(
5     host="localhost",
6     user="root",
7     password="MYfriNK@2024",
8     database="ecommerce"
9 )
10
11 cursor = db.cursor()
12
13 # Use MONTH() to preserve chronological order, then MONTHNAME() for labels
14 query = """
15     with count_per_order as
16         (select orders.order_id, orders.customer_id, count(order_items.order_id) as oc
17         from ecommerce.orders join ecommerce.order_items
18         on orders.order_id=order_items.order_id
19         group by orders.order_id, orders.customer_id)
20
21     select customers.customer_city, round(avg(count_per_order.oc),2)
22     from ecommerce.customers join count_per_order
23     on customers.customer_id=count_per_order.customer_id
24     group by customers.customer_city order by avg(count_per_order.oc) desc;
25 """
26
27 cursor.execute(query)
28 data = cursor.fetchall()
29 data
30 df=pd.DataFrame(data, columns=['Customer_City', 'Avg_product_Per_order'])
31 df=df.head(10)
32 df
```

Output:

	Customer_City	Avg_product_Per_order
0	padre carvalho	350.00
1	celso ramos	265.00
2	datas	240.00
3	candido godoi	240.00
4	picarra	200.00
5	teixeira soares	200.00
6	cidelândia	200.00
7	matias olímpio	200.00
8	inconfidentes	170.00
9	curralinho	160.00

3. Calculate the percentage of total revenue contributed by each product category.



```
1 import pandas as pd
2 import mysql.connector
3
4 db=mysql.connector.connect(
5     host="localhost",
6     user="root",
7     password="MYfriNK@2024",
8     database="ecommerce"
9 )
10 cursor=db.cursor()
11 query = """ select upper(products.product_category) category ,
12 round((sum(payments.payment_value)/(select sum(payment_value)
13 from ecommerce.payments))*100,2) sales
14 from ecommerce.products join ecommerce.order_items
15 on products.product_id=order_items.product_id
16 join ecommerce.payments
17 on payments.order_id=order_items.order_id
18 group by category order by sales desc;
19 """
20 cursor.execute(query)
21 data=cursor.fetchall()
22 data
23 df = pd.DataFrame(data, columns=['Category', 'Sales_Percentage_revenue'])
24 df.head()
```

Output:

...	Category	Sales_Percentage_revenue
0	BED TABLE BATH	251.09
1	HEALTH BEAUTY	241.77
2	COMPUTER ACCESSORIES	233.36
3	FURNITURE DECORATION	210.92
4	WATCHES PRESENT	210.23

4. Identify the correlation between product price and the number of times a product has been purchased.



```
1 import pandas as pd
2 import mysql.connector
3 import numpy as np
4
5 db=mysql.connector.connect(
6     host="localhost",
7     user="root",
8     password="MYfrINK@2024",
9     database="ecommerce"
10 )
11 cursor=db.cursor()
12 query = """ select products.product_category,
13 count(order_items.product_id),
14 round(avg(order_items.price),2)
15 from ecommerce.products join ecommerce.order_items
16 on products.product_id= order_items.product_id
17 group by product_category;
18 """
19 cursor.execute(query)
20 data=cursor.fetchall()
21 data
22 df = pd.DataFrame(data, columns=['product_category', 'order_count', 'avg_price'])
23 arr1=df["order_count"]
24 arr2=df["avg_price"]
25 np.correlation = np.corrcoef(arr1, arr2)[0, 1]
26 print("Correlation coefficient between order count and average price:", np.correlation)
```

Output:

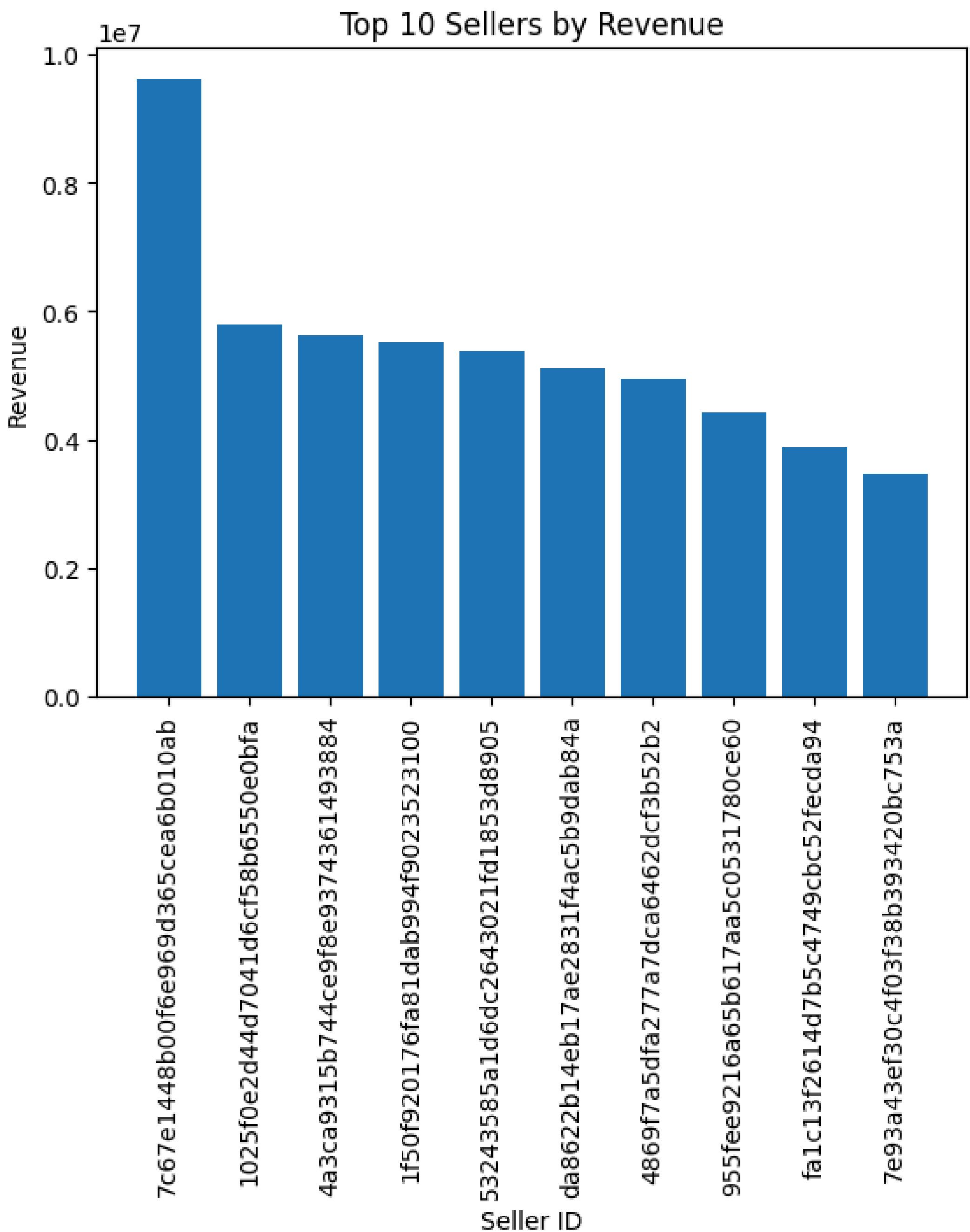
... Correlation coefficient between order count and average price: -0.1078901108337839

5. Calculate the total revenue generated by each seller, and rank them by revenue.



```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import mysql.connector
4
5 db=mysql.connector.connect(
6     host="localhost",
7     user="root",
8     password="MYfrINK@2024",
9     database="ecommerce"
10 )
11 cursor=db.cursor()
12 query = """ select *,dense_rank() over(order by revenue desc) as revenue_rank from
13 (select order_items.seller_id,
14 sum(payments.payment_value) as revenue
15 from ecommerce.order_items join ecommerce.payments
16 on order_items.order_id=payments.order_id
17 group by seller_id) as a;
18 """
19 cursor.execute(query)
20 data=cursor.fetchall()
21 data
22 df = pd.DataFrame(data, columns=['sellers_id', 'revenue', 'rank'])
23 df
24 plt.bar(df['sellers_id'].head(10), df['revenue'].head(10))
25 plt.xlabel('Seller ID')
26 plt.xticks(rotation=90)
27 plt.ylabel('Revenue')
28 plt.title('Top 10 Sellers by Revenue')
29 plt.show()
```

Output:



Advance Queries

1. Calculate the moving average of order values for each customer over their order history.



```
1 import pandas as pd
2 import mysql.connector
3
4 db=mysql.connector.connect(
5     host="localhost",
6     user="root",
7     password="MYfrINK@2024",
8     database="ecommerce"
9 )
10 cursor=db.cursor()
11 query = """ select customer_id, order_purchase_timestamp, payment,
12 avg(payment) over(partition by customer_id order by order_purchase_timestamp
13 rows between 2 preceding and current row) as moving_avg
14 from
15 (select orders.customer_id,orders.order_purchase_timestamp,
16 payments.payment_value as payment
17 from ecommerce.orders join ecommerce.payments
18 on orders.order_id=payments.order_id) as a;
19 """
20 cursor.execute(query)
21 data=cursor.fetchall()
22 data
23 df = pd.DataFrame(data, columns=['customer_id', 'order_purchase_timestamp',
24                                 'payment', 'moving_avg'])
25 df.head(10)
```

Output:

...	customer_id	order_purchase_timestamp	payment	moving_avg
0	00012a2ce6f8dcda20d059ce98491703	2017-11-14 16:08:26	114.74	114.739998
1	00012a2ce6f8dcda20d059ce98491703	2017-11-14 16:08:26	114.74	114.739998
2	00012a2ce6f8dcda20d059ce98491703	2017-11-14 16:08:26	114.74	114.739998
3	00012a2ce6f8dcda20d059ce98491703	2017-11-14 16:08:26	114.74	114.739998
4	00012a2ce6f8dcda20d059ce98491703	2017-11-14 16:08:26	114.74	114.739998
5	00012a2ce6f8dcda20d059ce98491703	2017-11-14 16:08:26	114.74	114.739998
6	00012a2ce6f8dcda20d059ce98491703	2017-11-14 16:08:26	114.74	114.739998
7	00012a2ce6f8dcda20d059ce98491703	2017-11-14 16:08:26	114.74	114.739998
8	00012a2ce6f8dcda20d059ce98491703	2017-11-14 16:08:26	114.74	114.739998
9	00012a2ce6f8dcda20d059ce98491703	2017-11-14 16:08:26	114.74	114.739998

2. Calculate the cumulative sales per month for each year.



```
1 import pandas as pd
2 import mysql.connector
3
4 db=mysql.connector.connect(
5     host="localhost",
6     user="root",
7     password="MYfriNK@2024",
8     database="ecommerce"
9 )
10 cursor=db.cursor()
11 query = """ select years,months,payment,sum(payment)
12 over(order by years,months) as cumulative_sales from
13 (select year(orders.order_purchase_timestamp) as years,
14 month(orders.order_purchase_timestamp) as months,
15 round(sum(payments.payment_value),2) as payment
16 from ecommerce.orders join ecommerce.payments
17 on orders.order_id= payments.order_id
18 group by years,months
19 order by years,months asc) as a;
20 """
21 cursor.execute(query)
22 data=cursor.fetchall()
23 data
24 df = pd.DataFrame(data, columns=['Years', 'Months', 'Payments', 'Cumulative_Sales'])
25 df.head(20)
```

Output:

...	Years	Months	Payments	Cumulative_Sales
0	2016	9	10089.60	1.008960e+04
1	2016	10	2185802.76	2.195892e+06
2	2016	12	784.80	2.196677e+06
3	2017	1	5126229.32	7.322906e+06
4	2017	2	10727527.54	1.805043e+07
5	2017	3	16659124.30	3.470956e+07
6	2017	4	15434445.86	5.014400e+07
7	2017	5	21990330.37	7.213433e+07
8	2017	6	18902386.57	9.103672e+07
9	2017	7	21929609.14	1.129663e+08
10	2017	8	24874587.60	1.378409e+08
11	2017	9	26883966.02	1.647249e+08
12	2017	10	28716786.91	1.934417e+08
13	2017	11	44243804.03	2.376855e+08
14	2017	12	32470532.08	2.701560e+08
15	2018	1	41118352.33	3.112744e+08
16	2018	2	36616554.45	3.478909e+08
17	2018	3	42839862.14	3.907308e+08
18	2018	4	42860928.15	4.335917e+08
19	2018	5	42646015.20	4.762377e+08

3. Calculate the year-over-year growth rate of total sales.



```
1 import pandas as pd
2 import mysql.connector
3
4 db=mysql.connector.connect(
5     host="localhost",
6     user="root",
7     password="MYfriNK@2024",
8     database="ecommerce"
9 )
10 cursor=db.cursor()
11 query = """ with a as (select year(orders.order_purchase_timestamp) as years,
12 round(sum(payments.payment_value),2) as payment
13 from ecommerce.orders join ecommerce.payments
14 on orders.order_id=payments.order_id
15 group by years order by years)
16 select years, payment, lag(payment, 1) over(order by years) as previous_year_payment from a;
17 """
18 cursor.execute(query)
19 data=cursor.fetchall()
20 data
21 df = pd.DataFrame(data, columns=['Year', 'Sales', 'Previous_Year_Sales'])
22 df
```

Output:

...	Year	Sales	Previous_Year_Sales
0	2016	2.196677e+06	NaN
1	2017	2.679593e+08	2.196677e+06
2	2018	3.214953e+08	2.679593e+08

3. Calculate the year-over-year growth rate of total sales.



```
1 import pandas as pd
2 import mysql.connector
3
4 db=mysql.connector.connect(
5     host="localhost",
6     user="root",
7     password="MYfriNK@2024",
8     database="ecommerce"
9 )
10 cursor=db.cursor()
11 query = """ with a as (select year(orders.order_purchase_timestamp) as years,
12 round(sum(payments.payment_value),2) as payment
13 from ecommerce.orders join ecommerce.payments
14 on orders.order_id=payments.order_id
15 group by years order by years)
16 select years, ((payment-lag(payment, 1) over(order by years))/lag(payment, 1)
17 over(order by years)) * 100 as previous_year_payment from a;
18 """
19 cursor.execute(query)
20 data=cursor.fetchall()
21 data
22 df = pd.DataFrame(data, columns=['Year', 'yoY_growth%'])
23 df
```

Output:

	Year	yoY_growth%
0	2016	NaN
1	2017	12098.393766
2	2018	19.979136

4. Calculate the retention rate of customers, defined as the percentage of customers who make another purchase within 6 months of their first purchase.



```
1 import mysql.connector
2
3 db=mysql.connector.connect(
4     host="localhost",
5     user="root",
6     password="MYfriNK@2024",
7     database="ecommerce"
8 )
9 cursor=db.cursor()
10 query = """ with a as (select customers.customer_id,
11 min(orders.order_purchase_timestamp) as first_order
12 from ecommerce.customers join ecommerce.orders
13 on customers.customer_id=orders.customer_id
14 group by customers.customer_id),
15
16 b as (select a.customer_id, count(distinct orders.order_purchase_timestamp) as next_order
17 from a join ecommerce.orders
18 on a.customer_id= orders.customer_id
19 and orders.order_purchase_timestamp > first_order
20 and orders.order_purchase_timestamp <
21 date_add(first_order,interval 6 month)
22 group by a.customer_id)
23
24 select 100 * (count(distinct a.customer_id)/count(distinct b.customer_id))
25 from a left join b
26 on a.customer_id=b.customer_id;
27 """
28 cursor.execute(query)
29 data=cursor.fetchall()
30 data
```

Output:

...

[(None,)]

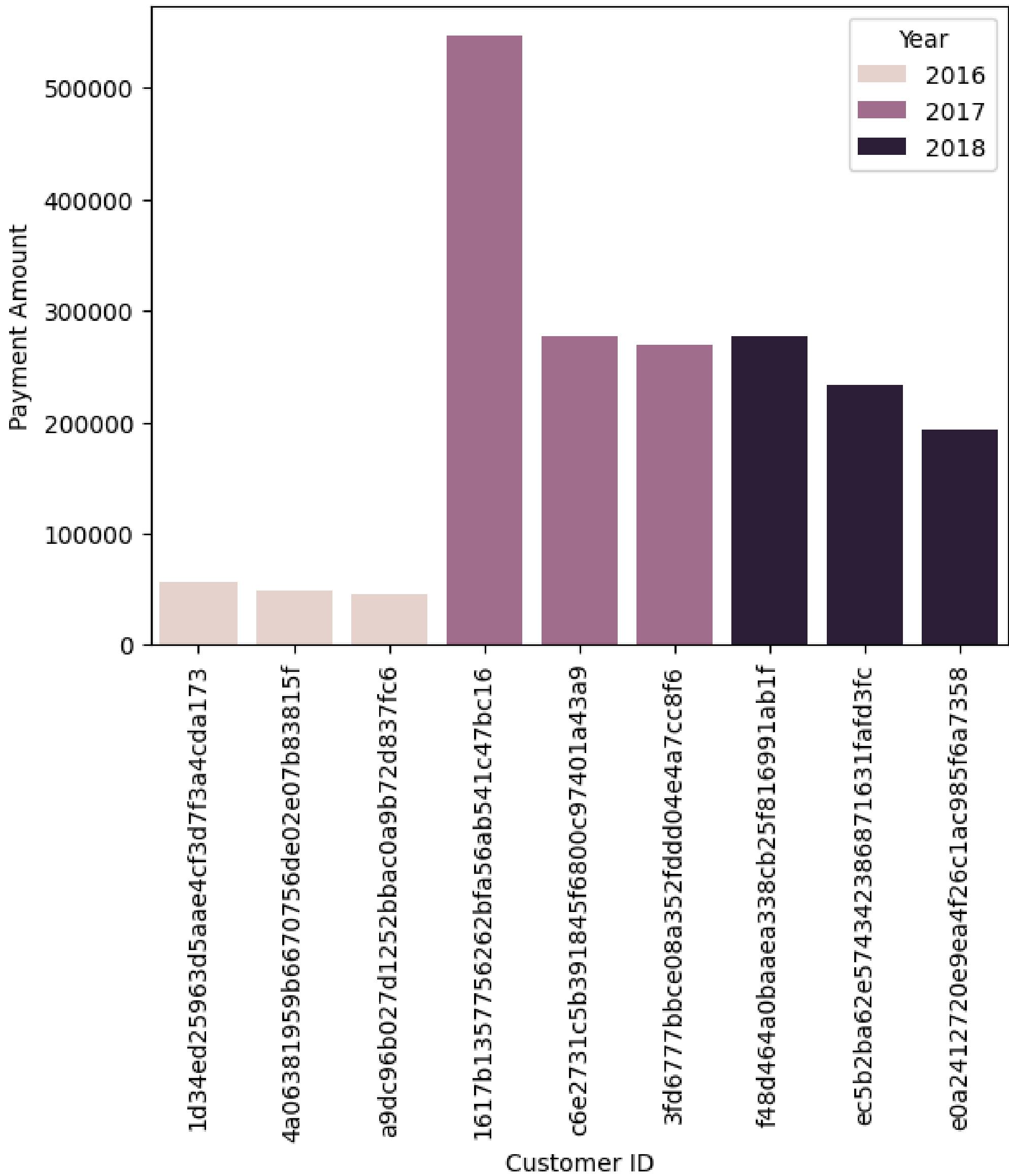
5. Identify the top 3 customers who spent the most money in each year.



```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import mysql.connector
4
5 db=mysql.connector.connect(
6     host="localhost",
7     user="root",
8     password="MYfriNK@2024",
9     database="ecommerce"
10 )
11 cursor=db.cursor()
12 query = """ select years,customer_id,payment,d_rank
13 from
14 (select year(orders.order_purchase_timestamp) as years,orders.customer_id,
15 sum(payments.payment_value) as payment,
16 dense_rank() over(partition by year(orders.order_purchase_timestamp)
17 order by sum(payments.payment_value) desc) as d_rank
18 from ecommerce.orders join ecommerce.payments
19 on orders.order_id=payments.order_id
20 group by year(order_purchase_timestamp),orders.customer_id) as a
21 where d_rank<=3;
22 """
23 cursor.execute(query)
24 data=cursor.fetchall()
25 data
26 df = pd.DataFrame(data, columns=['Year', 'Customer_ID', 'Payment', 'Rank'])
27 sns.barplot(x='Customer_ID', y='Payment', hue='Year', data=df)
28 plt.xticks(rotation=90)
29 plt.xlabel('Customer ID')
30 plt.ylabel('Payment Amount')
31 plt.title('Top 3 Customers by Payment Amount Each Year')
32 plt.show()
```

Output:

Top 3 Customers by Payment Amount Each Year





SUMMARY OF INSIGHTS

Here are the major insights revealed from SQL + Python analysis:

- Most customers are concentrated in a few major states (e.g., SP, RJ)
- A significant percentage of orders come from only ~5 large cities
- Certain product categories dominate the marketplace in both volume and revenue
- Payment behavior shows that credit card is the most used method
- Installments form a visible percentage of total transactions
- Monthly sales show clear seasonality and demand peaks
- Shipping delays occur mostly when seller-to-customer distance is high
- Customer retention is influenced heavily by category preference & delivery performance
- A small group of customers generate majority of yearly revenue
- Sellers in certain states consistently outperform others in order volume



CONCLUSION

This E-Commerce SQL + Python project successfully uncovers key business insights using data analytics techniques such as:

- Joins
- Group By Aggregations
- Window Functions
- CTEs
- Time-Series analysis
- Correlation analysis
- Category & seller performance analysis

These insights enable an E-Commerce business to:

- Improve customer targeting and segmentation
- Optimize product categories and pricing
- Increase seller performance and accountability
- Reduce delivery delays through logistics planning
- Improve revenue forecasting and marketing strategy
- Strengthen customer retention programs



Thankyou

punitpalofficial@gmail.com