

Case Study: Data Analyst Task Summary

As a Data Analyst, I was tasked with providing insights for the product team using four datasets. These datasets were provided to help the product team identify trends and patterns to guide their project strategy. Below, I'll outline the steps I took to complete the tasks, as well as the SQL and Python techniques used to analyze the data.

Step 1: Data Preparation

I started by loading the four datasets into a database from the **Salla platform**. Salla is an e-commerce platform that allows merchants to create an online store or expand their business online.

I used (**Microsoft SQL Server**) for this case, as it is a robust database system for handling large datasets and complex queries. After importing the data, I explored the dataset schema to understand the relationships between the tables. This step was crucial as it enabled me to effectively join the tables later in SQL.

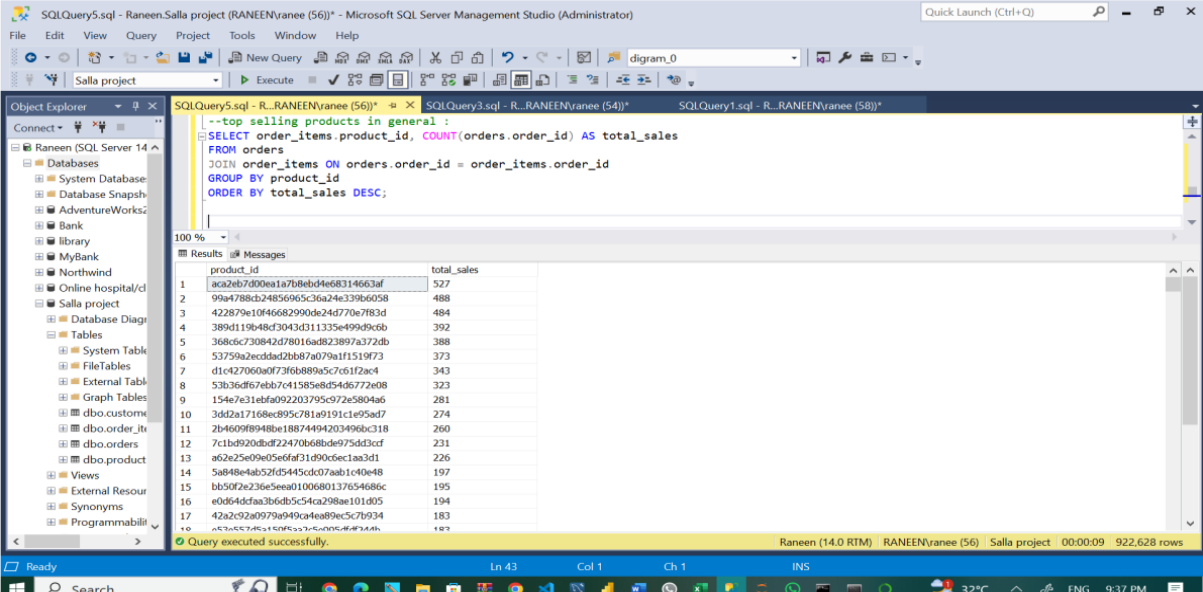
Step 2: SQL Tasks

I completed the first four tasks using SQL to extract insights from the data.

Task 1: Top Selling Products (General and by Region)

- **Top selling products in general:**

To identify the top-selling products, both overall, I used SQL queries that aggregated sales data from relevant tables. For the overall top-selling products, I performed a ****GROUP BY**** on product identifiers and summed the total sales.



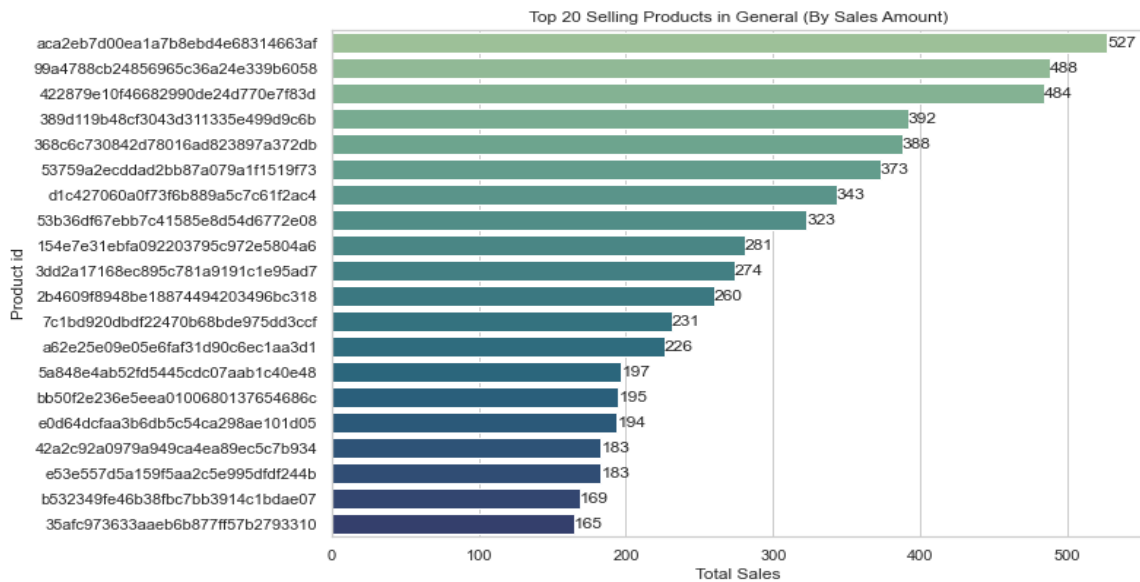
The screenshot displays the Microsoft SQL Server Management Studio (SSMS) interface. The main window shows a SQL query in the 'Query Editor' pane, which is titled 'SQLQuery5.sql - R...RANEEN\vanee (56)*'. The query is as follows:

```
--top selling products in general :
SELECT order_items.product_id, COUNT(orders.order_id) AS total_sales
FROM orders
JOIN order_items ON orders.order_id = order_items.order_id
GROUP BY product_id
ORDER BY total_sales DESC;
```

Below the query, the 'Results' pane shows the output of the query. It contains a table with two columns: 'product_id' and 'total_sales'. The table lists the top-selling products based on their total sales.

product_id	total_sales
aca2eb7000ea1a7b8ebd4e68314663af	527
99a778bd324856965c36a24e33986058	488
422879610f466829096a24d770e783d	484
389d119b48cf3043d311335e499d9c6b	392
368c6c730842d78016ad823897a372db	388
53759a2ecddad2bb87a079a1f151973	373
d1c427060a0f736b889a5c7c61f2ac4	343
53b36df67ebb7c41585e8d54d6772e08	323
154e7e31ebfa092203795c972e5804a6	281
3dd2a17168ec895c781a9191c1e95ad7	274
2b4609f8948be18874494203496bc318	260
7c1b920dbdf22470b68bd975dd3cdf	231
a62c25e09d056df31890d6e1a3d1	226
5a98be4ub52f5445cd07aab1c40e48	197
bb50f2e236e5ead100680137654686c	195
e0d54dca3b6db5c54ca298ae101d05	194
42a2c92a0979a949ca4e89ec5c7b934	183
...	...

The status bar at the bottom indicates that the query was executed successfully, returning 922,628 rows.

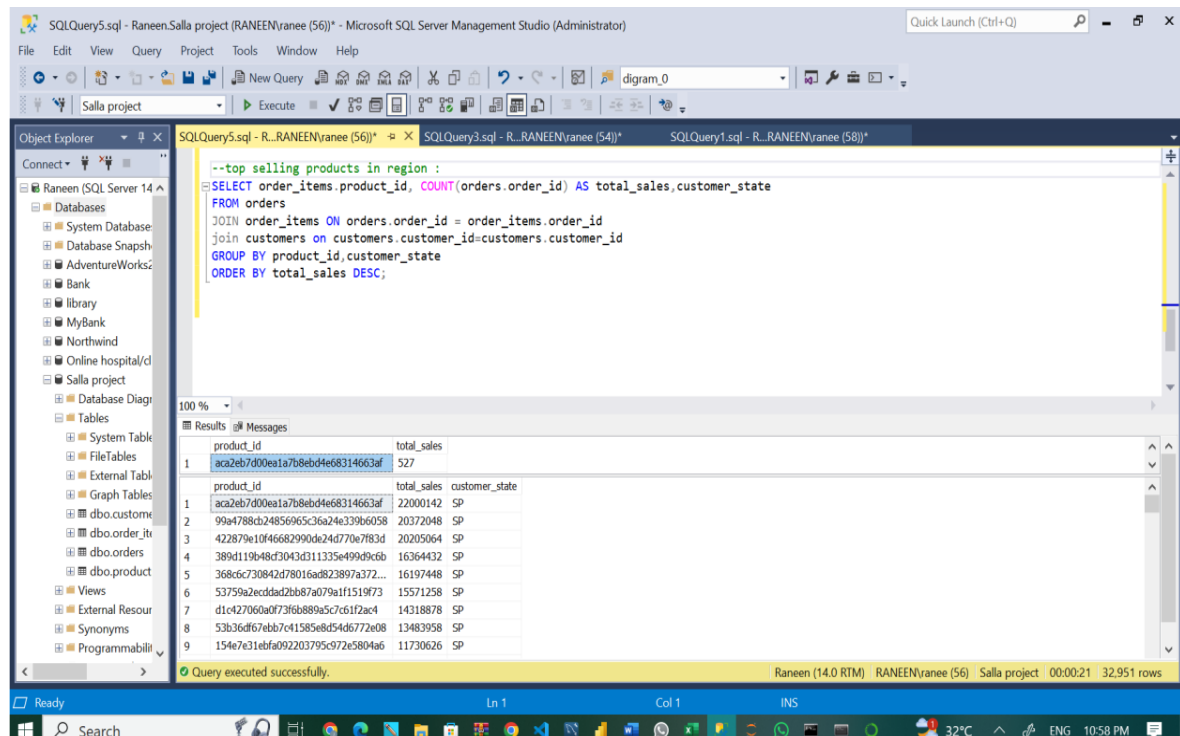


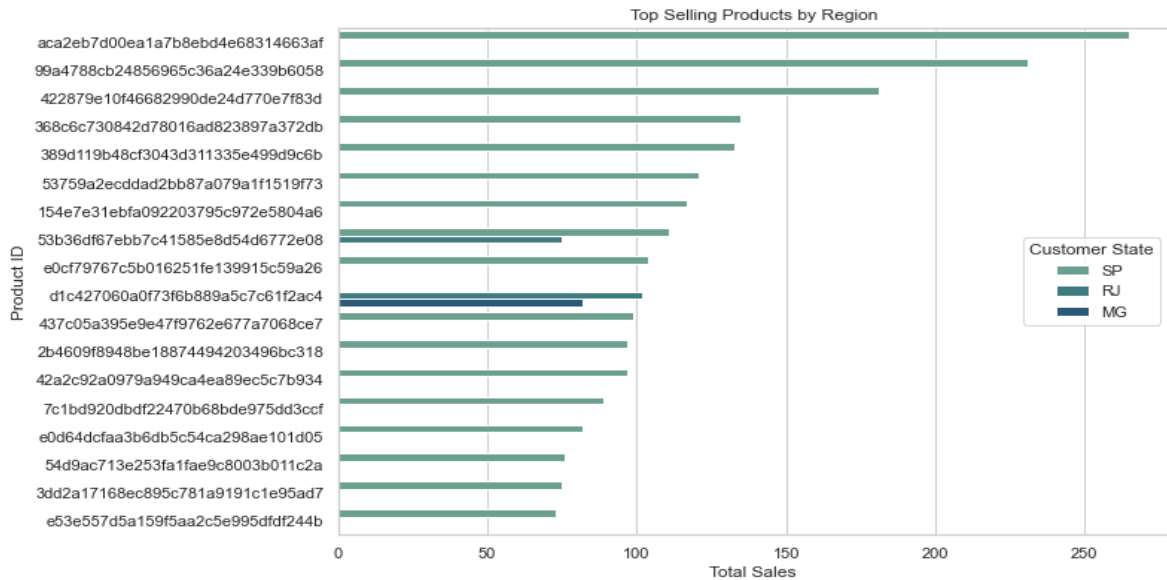
The product with ID aca2eb7d00ea1a7b8ebd4e68314663af has the highest sales overall

Other top-selling products follow, with IDs like 99a4788cb24856965c36a24e339b6058 and 422879e10f46682990de24d770e7f83d also showing high sales volumes.

- **Top selling products by region:**

For top-selling products by region, I modified the query to include the region and grouped by both the product and the region.



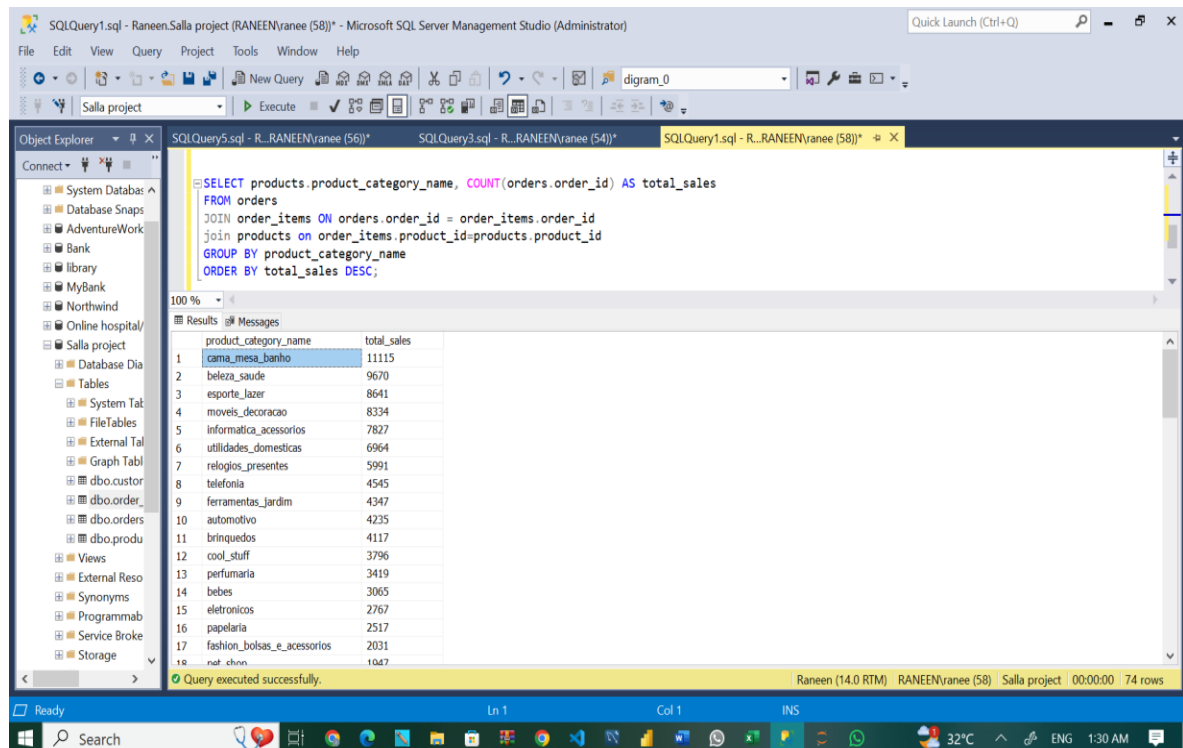


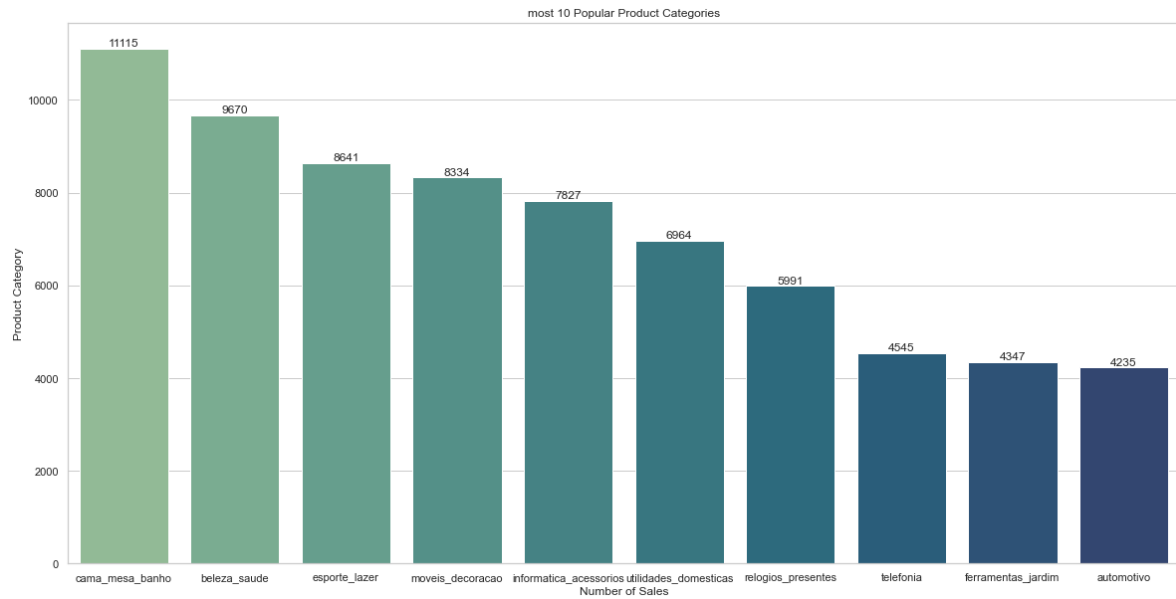
The top-selling product aca2eb7d00ea1a7b8ebd4e68314663af also leads in the state of SP (São Paulo) with 265 units sold there.

This is followed closely by product 99a4788cb24856965c36a24e339b6058 with 231 units in SP.

Task 2: Most Popular Product Categories

To determine the most popular product categories, I used a ****JOIN**** to link the product data with the category table and then grouped by category to calculate the sales.



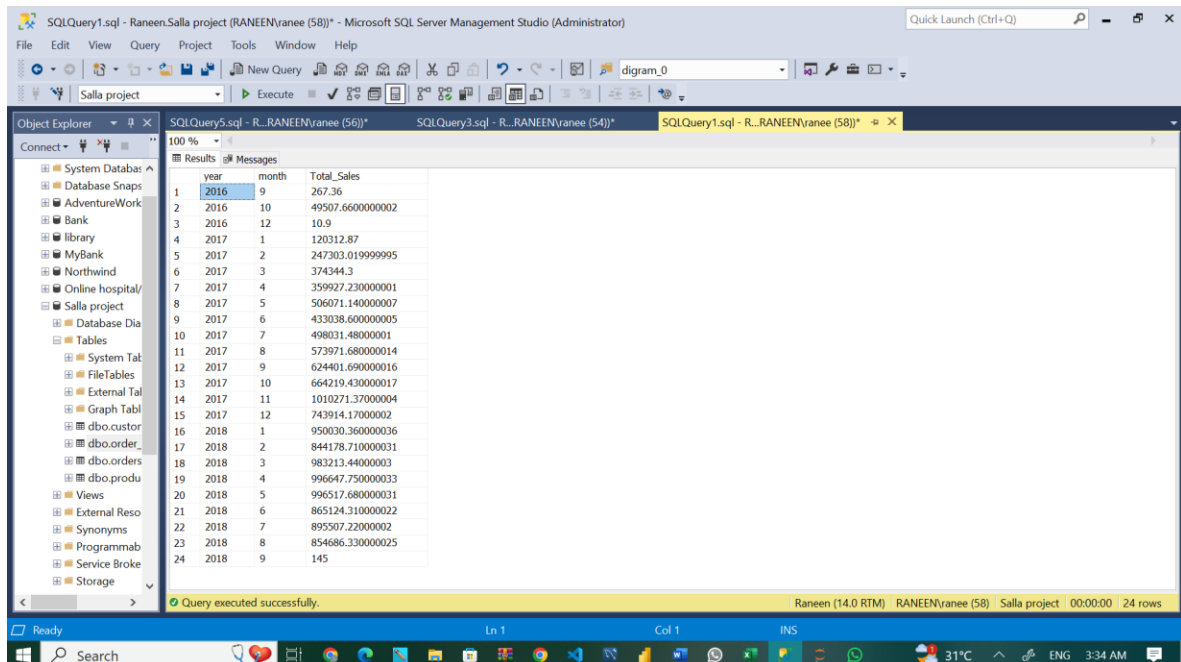


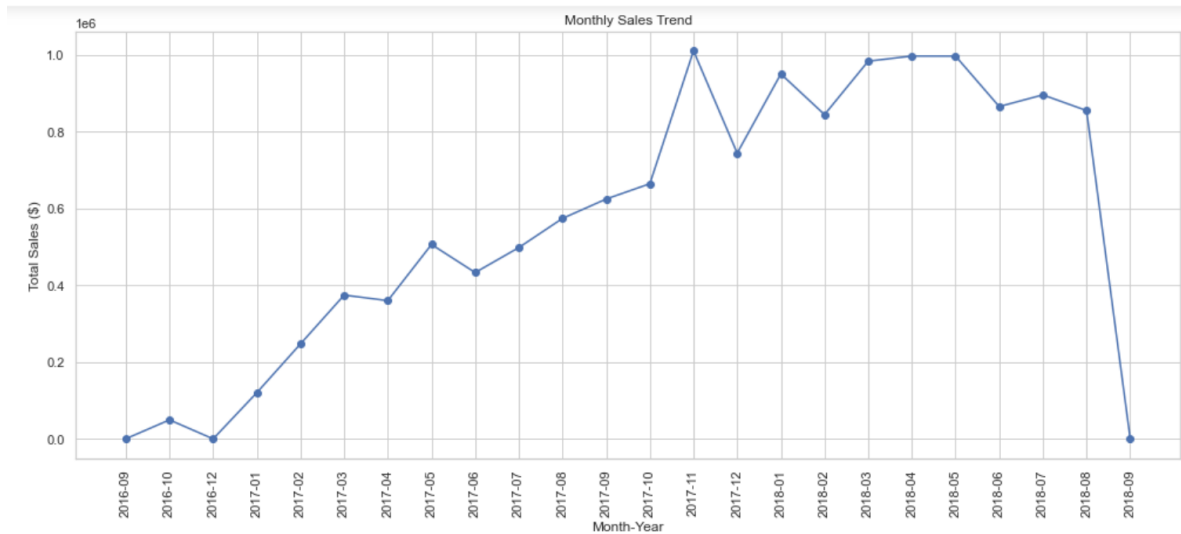
Cama_mesa_banho (Bed, Table, Bath) is the most popular category with 11,115 sales.

Other popular categories include Beleza_saude (Beauty and Health) and Esporte_lazer (Sports and Leisure).

Task 3: Monthly, Quarterly, and Yearly Sales

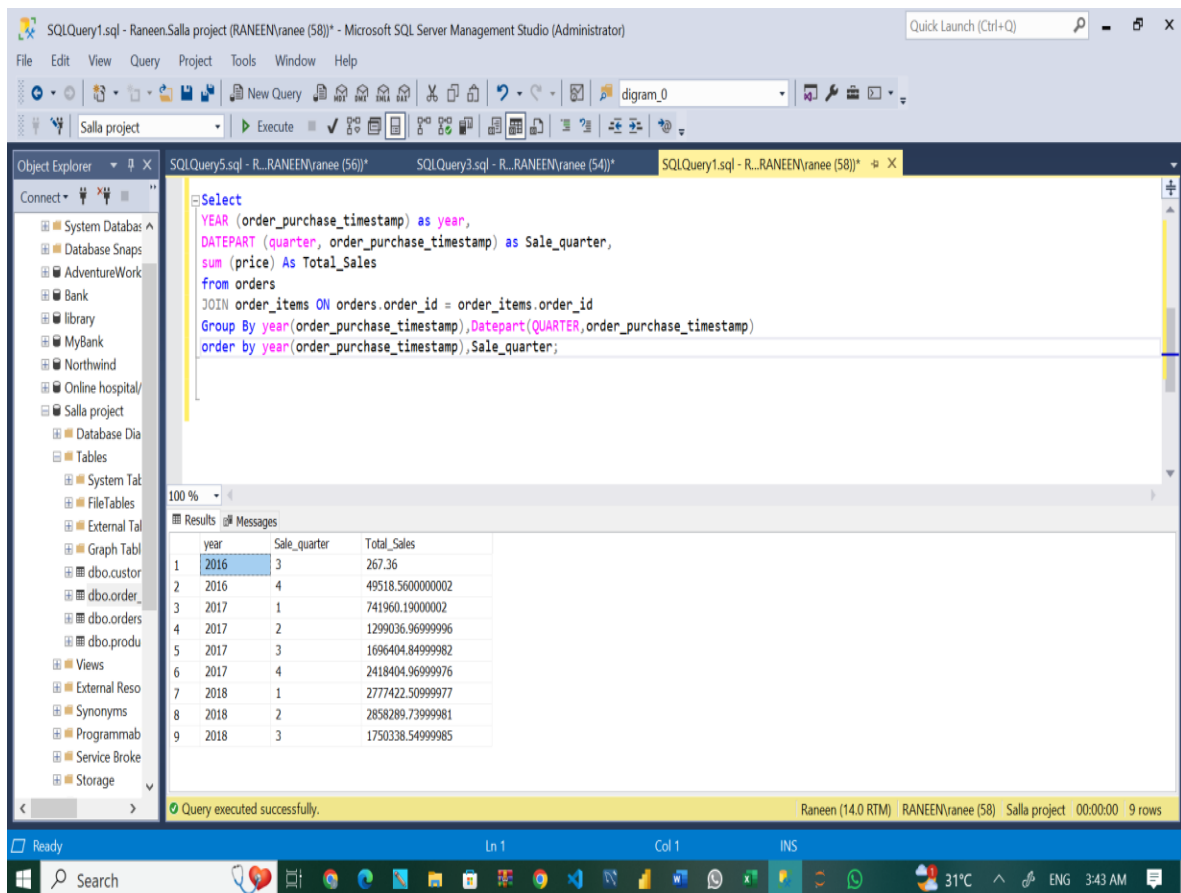
- Monthly Sales

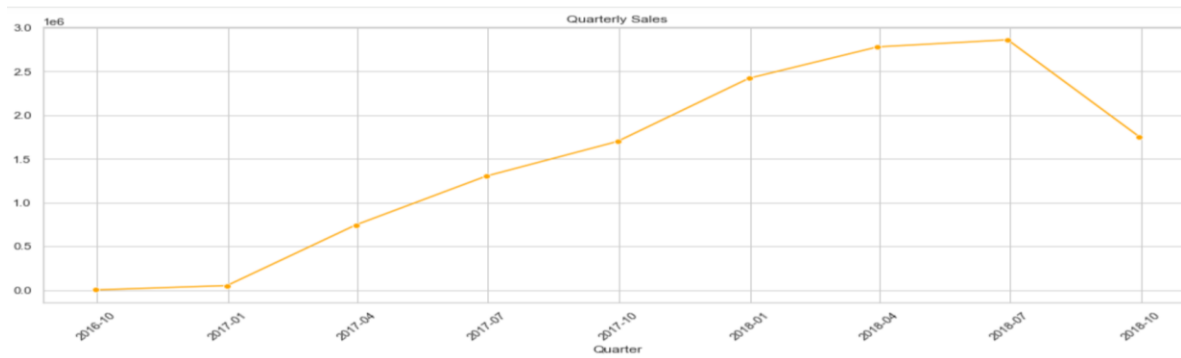




Sales show variation across months, with significant sales noted in November 2017 amounting to \$120,312.87.

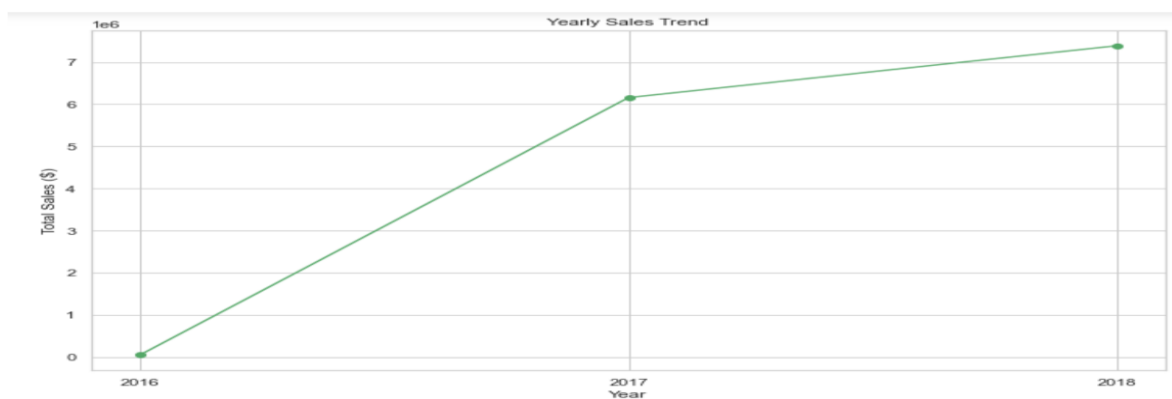
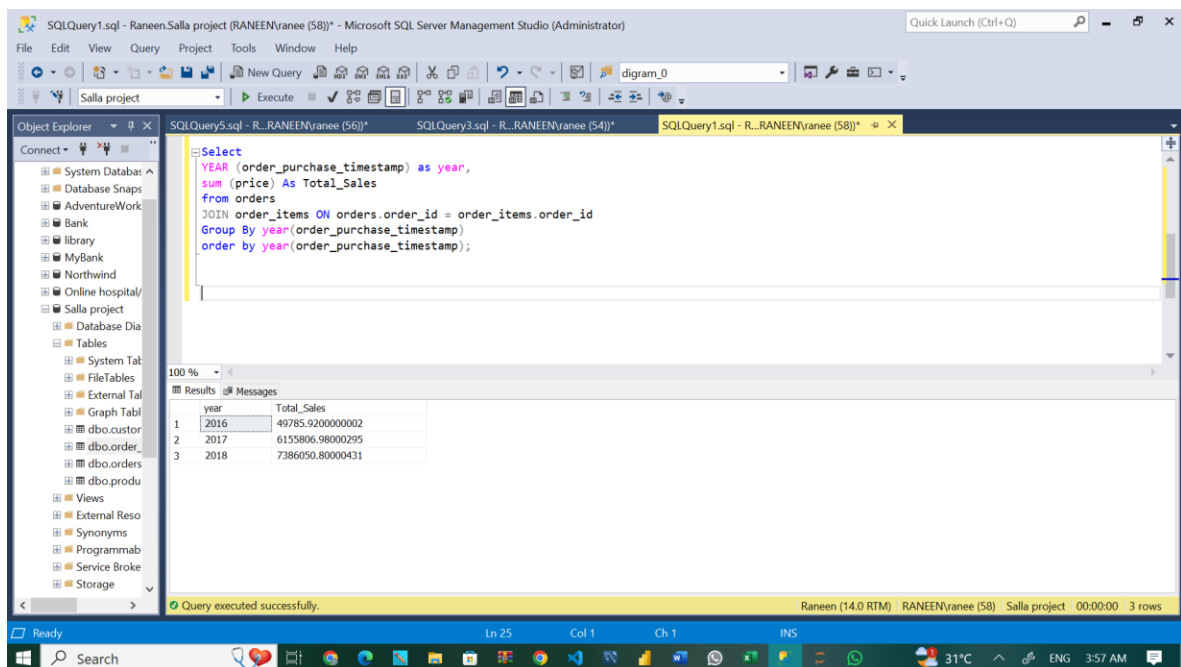
- Quarterly Sales





The second quarter of 2018 saw the highest sales volume, accumulating to \$2858289.74.

- Yearly Sales

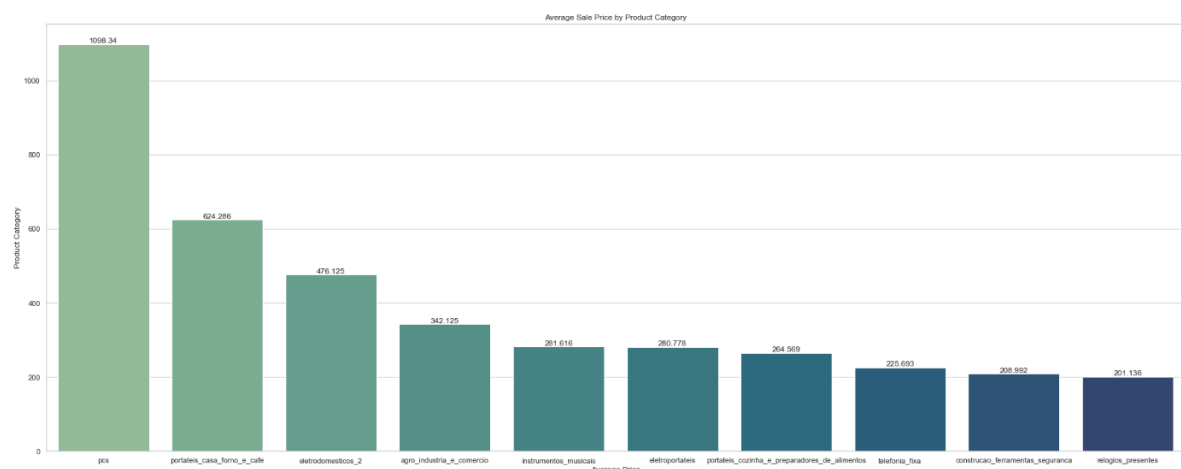
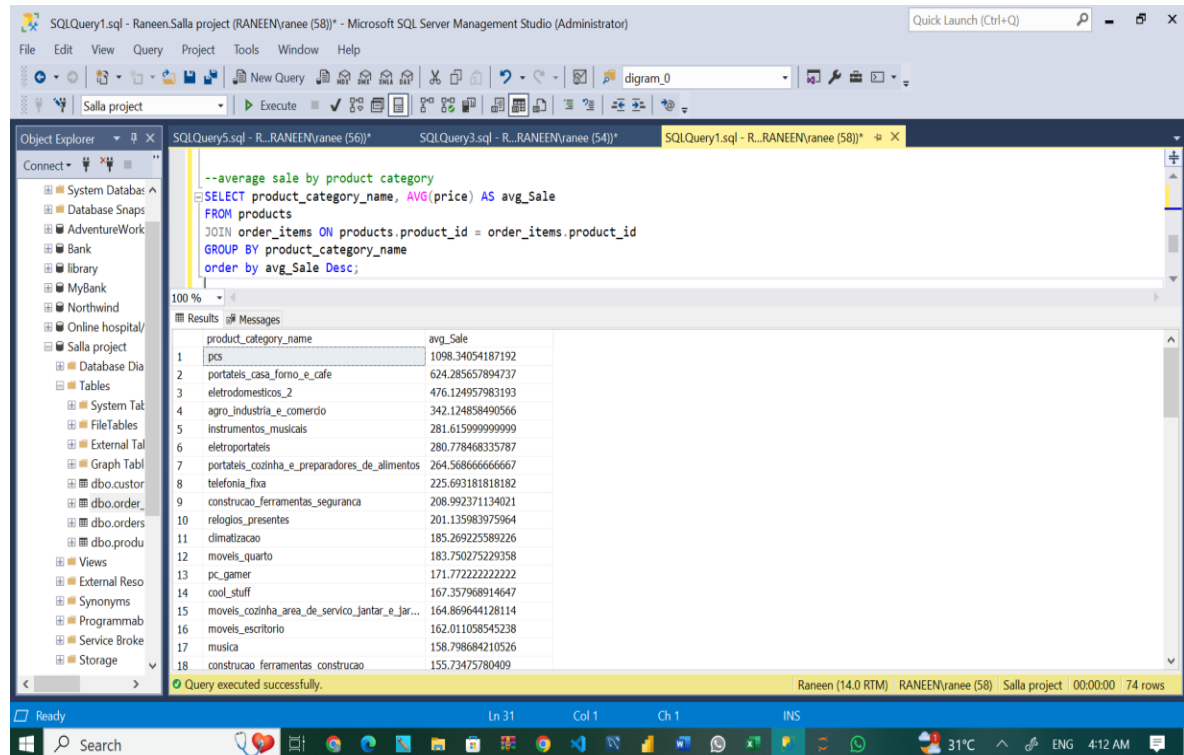


The year 2018 had the highest sales with a total of \$7386050.80.

Task 4: Average Sale by Product Category and Top Categories by Location

- **Average Sale by Product Category**

To calculate the average sale per category, I used ****AVG**** and joined the necessary tables.

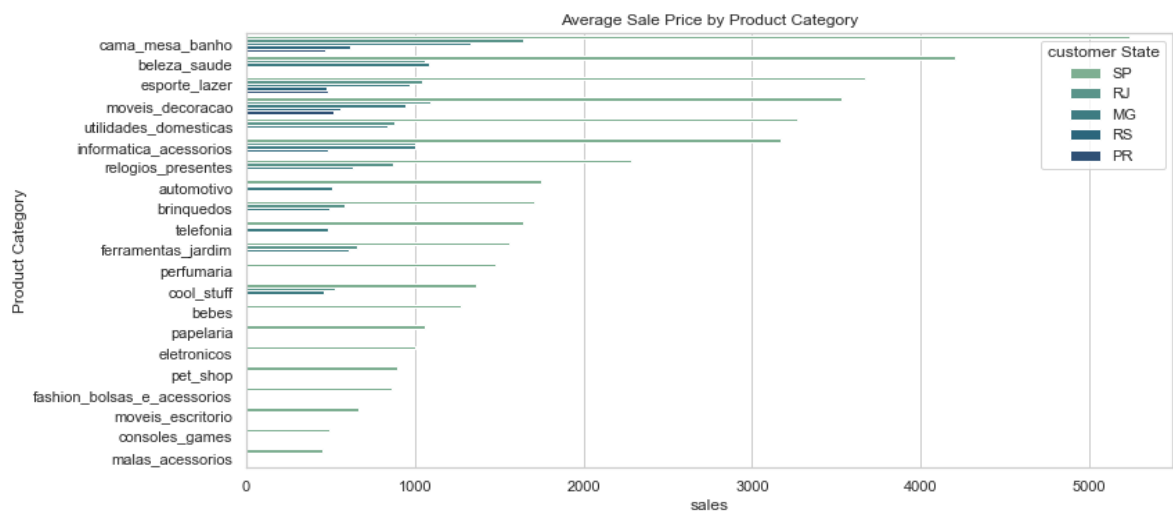
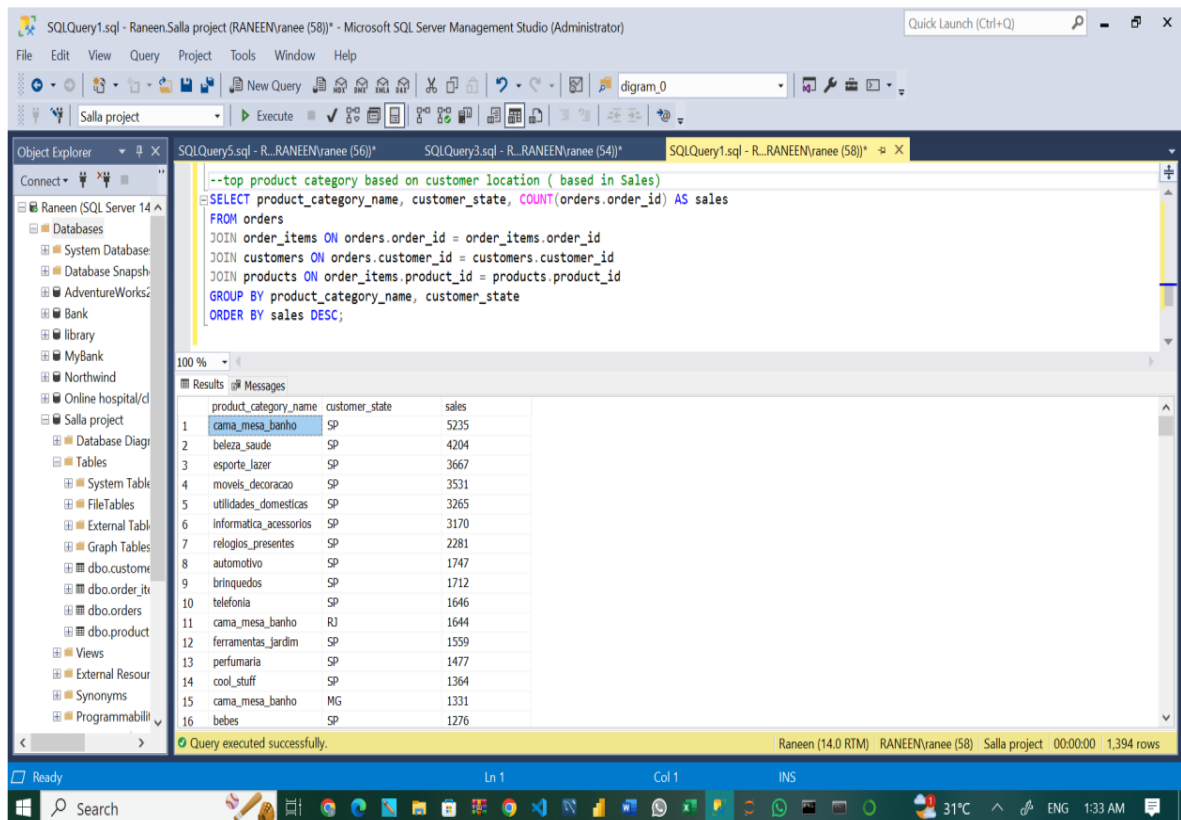


The category pcs (computers) has the highest average sale price at approximately \$109842.34.

Other notable categories by average price include portateis_casa_forno_e_cafe (small appliances, home oven and coffee) and eletrodomesticos_2 (home appliances 2).

- **Top Categories by Customer Location**

For the top product categories based on customer location, I grouped by location and category.



in São Paulo (SP), the cama_mesa_banho (Bed, Table, Bath) category ranks first with 5,235 sales, representing the highest orders.

- **Top Product Categories by order count for each Customer State**

customer_state	product_category_name	order_count
AC	moveis_decoracao	12
AL	beleza_saude	63
AM	beleza_saude	20
AP	beleza_saude	10
BA	beleza_saude	350
CE	beleza_saude	167
DF	beleza_saude	246
ES	cama_mesa_banho	225
GO	cama_mesa_banho	235
MA	beleza_saude	89
MG	cama_mesa_banho	1331
MS	esporte_lazer	75
MT	beleza_saude	90
PA	beleza_saude	107
PB	beleza_saude	78
PE	beleza_saude	240
PI	beleza_saude	54
PR	moveis_decoracao	520
RJ	cama_mesa_banho	1644
RN	beleza_saude	58
RO	beleza_saude	25
RR	esporte_lazer	8
RS	cama_mesa_banho	614
SC	esporte_lazer	363
SE	beleza_saude	40
SP	cama_mesa_banho	5235
TO	beleza_saude	1



in São Paulo (SP), the cama_mesa_banho (Bed, Table, Bath) category ranks first with 5,235 sales, representing the highest orders among all states.

In Paraná (RP), the esporte_lazer (sports leisure) category has 8 sales, representing the lowest orders among all states.

Step 3: Python Tasks

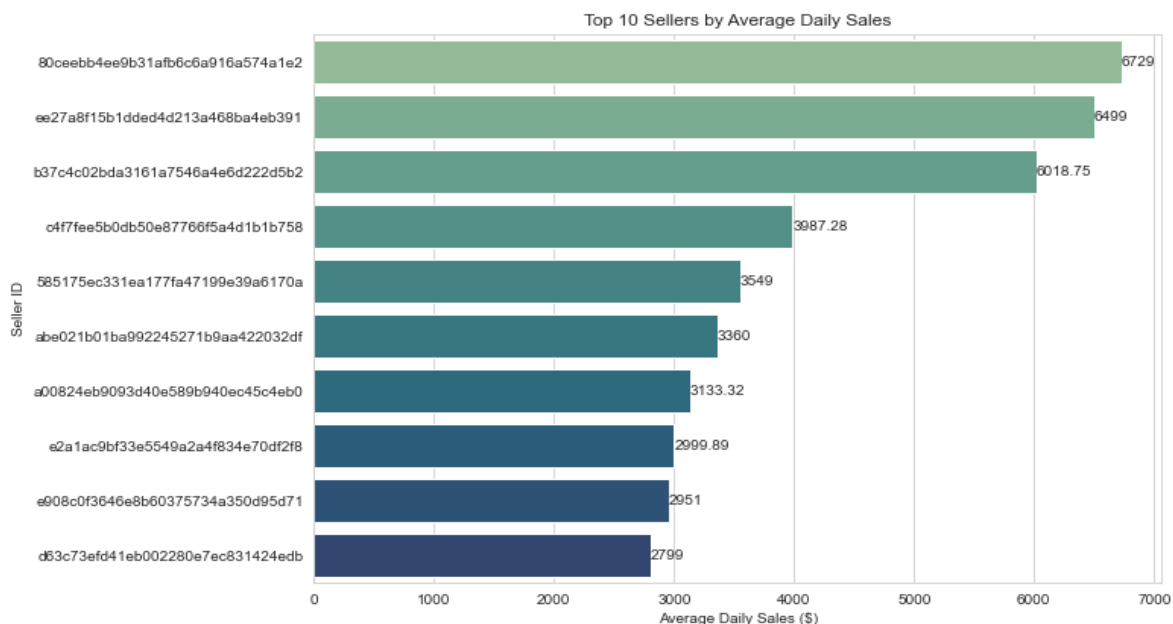
For the next three tasks, I used **Python** along with libraries like Pandas for data manipulation, Matplotlib and Seaborn for visualization.

Task 1: Top 10 Stores with the Highest Average Daily Sales

To calculate the top 10 stores with the highest average daily sales, I first grouped the sales data by store and date, calculated the daily sales, and then took the average.

```
1 # Sort and get the top 10 sellers with the highest average daily sales
2 top_sellers_average_daily_sales = average_daily_sales.sort_values(by='average_sales', ascending=False).head
3 top_sellers_average_daily_sales
```

	seller_id	average_sales
1583	80ceebb4ee9b31afb6c6a916a574a1e2	6729.000000
2881	ee27a8f15b1dded4d213a468ba4eb391	6499.000000
2157	b37c4c02bda3161a7546a4e6d222d5b2	6018.750000
2378	c4f7fee5b0db50e8776f5a4d1b1b758	3987.276667
1078	585175ec331ea177fa47199e39a6170a	3549.000000
2073	abe021b01ba992245271b9aa422032df	3360.000000
1943	a00824eb9093d40e589b940ec45c4eb0	3133.323333
2730	e2a1ac9bf33e5549a2a4f834e70df2f8	2999.890000
2813	e908c0f3646e8b60375734a350d95d71	2951.000000
2589	d63c73efd41eb002280e7ec831424edb	2799.000000



The product with ID 80ceebb4ee9b31afb6c6a916a574a1e2 has the highest average daily sales overall with 6729 average daily sales.

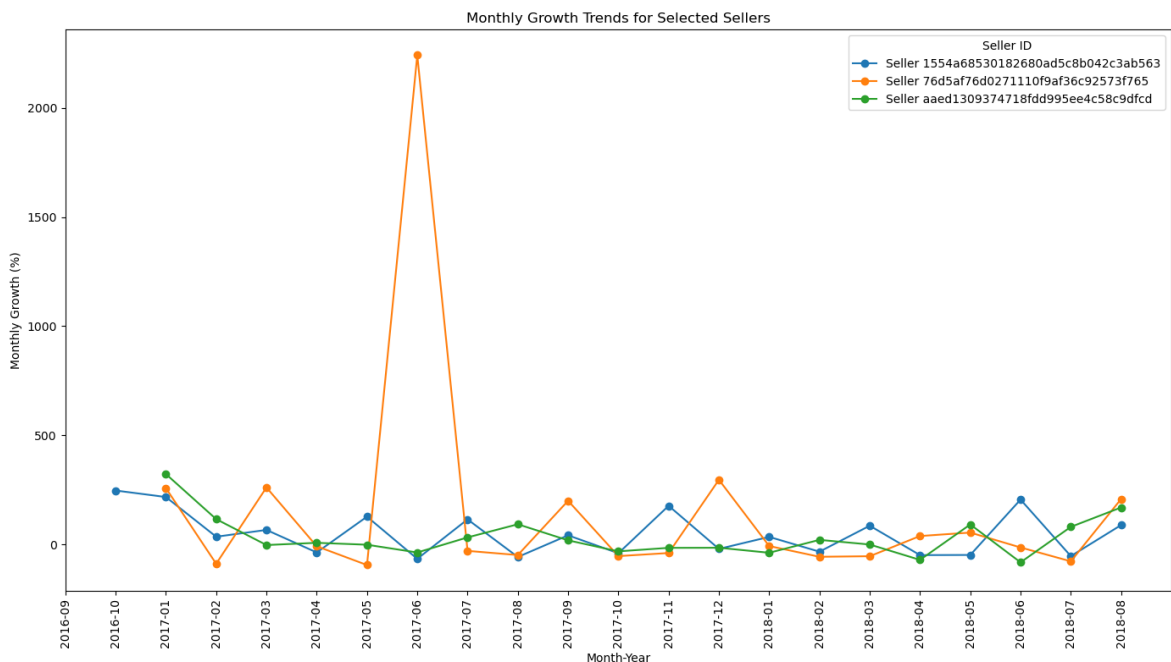
Other top-selling products follow, with IDs like ee27a8f15b1dded4d213a468ba4eb391 and b37c4c02bda3161a7546a4e6d222d5b2 also showing average daily sales.

Task 2: Monthly Growth Percentage per Store

To calculate the percentage of monthly growth, I used the formula:

$$\text{Growth Percentage} = \frac{\text{Current Month Sales} - \text{Previous Month Sales}}{\text{Previous Month Sales}} \times 100$$

seller_id	1554a68530182680ad5c8b042c3ab563	76d5af76d0271110f9af36c92573f765	aaed1309374718fdd995ee4c58c9dfcd
order_purchase_timestamp			
2016-09	NaN	NaN	NaN
2016-10	247.029771	NaN	NaN
2017-01	217.161494	257.967519	322.416534
2017-02	35.350136	-88.617299	115.863756
2017-03	66.439504	261.109227	-2.780926
2017-04	-36.805179	-8.454276	7.406743
2017-05	127.258081	-94.371761	-1.369177
2017-06	-66.018123	2241.659070	-36.905367
2017-07	114.728691	-29.516506	32.197478
2017-08	-57.515246	-48.657336	92.368581
2017-09	41.734192	199.924699	18.759232
2017-10	-40.356877	-53.491984	-31.725302
2017-11	176.972949	-39.360988	-15.627846
2017-12	-20.473256	296.178303	-15.222085
2018-01	34.499353	-7.399644	-38.166273
2018-02	-33.246657	-56.641037	20.917917
2018-03	85.697825	-54.087547	-0.364964
2018-04	-48.954951	38.511101	-70.818071
2018-05	-48.271102	54.268293	91.464435
2018-06	205.488258	-13.907339	-82.124126
2018-07	-53.542733	-77.413858	80.440098
2018-08	88.660400	208.046940	169.376694



The data begins in October 2016. **Seller 1** (1554a68530182680ad5c8b042c3ab563) shows early growth, while the other sellers have no recorded sales, suggesting different entry times. Seller 1 also faces negative growth in several months, like -66.02 in June 2017, indicating possible market saturation or competition issues.

Seller 2 (76d5af76d0271110f9af36c92573f765) has fluctuating sales, peaking at 2241.66 in June 2017 due to a likely successful promotion, but then experiences negative growth, showing inconsistent performance. In contrast, **Seller 3** (aaed1309374718fdd995ee4c58c9dfcd) demonstrates steady positive growth, suggesting a stronger customer base or better product fit.

Task 3: Cohort Analysis

For cohort analysis, I first identified the month of a customer's first purchase and then tracked their behavior over time.

:	cohort_index	0
	cohort	
	2016-09	4.0
	2016-10	324.0
	2016-12	1.0
	2017-01	800.0
	2017-02	1780.0
	2017-03	2682.0
	2017-04	2404.0
	2017-05	3700.0
	2017-06	3245.0
	2017-07	4026.0
	2017-08	4331.0
	2017-09	4285.0
	2017-10	4631.0
	2017-11	7544.0
	2017-12	5673.0
	2018-01	7269.0
	2018-02	6728.0
	2018-03	7211.0
	2018-04	6939.0
	2018-05	6873.0
	2018-06	6167.0
	2018-07	6292.0
	2018-08	6512.0
	2018-09	16.0
	2018-10	4.0

Cohort Growth:

Customers who first purchased in October 2016 grew from 4 in September to 324 in October, indicating effective marketing efforts.

Strong Retention:

Retention rates were high for late 2016 to 2017 cohorts, with the January 2017 group (800 customers) staying engaged through mid-2018.

Engagement Decline:

A notable drop in customer counts in late 2018 suggests potential issues with engagement, likely due to increased competition.

Conclusion

Through the analysis, I was able to generate insights that would guide the product team's decision-making. I used SQL to answer product-related questions, and Python to conduct more complex analyses, like cohort analysis and store performance evaluation. Visualizing the results made the insights easier to interpret and helped the team strategize for future product development.