Sales Analysis of Retail Stores

Evgenii Kozin

Table of Contents

Introduction

Introduction

- The aim of this project is to conduct a comprehensive analysis of the retail chain's sales performance across different regions and periods. The focus is on identifying the main factors influencing sales, including store locations, product categories, and marketing activities. The insights gained will guide strategic decisions and help drive business growth.
- To stay competitive and enhance results, the company is dedicated to leveraging data-driven insights to refine its sales strategies and increase operational efficiency.

Methodology

Data Details

When I downloaded the data first I went back & forth to understand it's format,structure,size of the dataset before jumping into the task

The dataset consist of the following files
***sales.csv*, *product_hierarchy.csv*, *store_cities.csv*, *store_names.csv*, *city_names.csv*, and *product_names.csv*.**

Where sales.csv is the target file to analyse and derive insights.sales.csv consist of the date of purchase, productid information,storeid information , types of promo-code, revenue, stock & price information

product_hierarchy.csv* and *product_names.csv consists of product category, product names, product id which is a unique identity for each product.

store_names.csv & store_cities.csv consists of storeid info and storetypeid info to cityid to establish the relationship between variables.

city_names.csv consists of cityid info and city names

Module 1: Data Cleaning and Preparation Using Excel

The initial and most important step in obtaining meaningful insights is to remove unnecessary or irrelevant data from the dataset. For this task, I used Excel, which is an excellent tool for data analysis.

Step 1: Detecting missing values is critical for any data-driven business, as unaddressed gaps can lead to incorrect conclusions. Using the filter feature, I located columns containing null values and removed those records from the dataset.

Step 2: Applied Excel's *Remove Duplicates* function to identify and eliminate duplicate entries, ensuring only unique records remained for further analysis. This step safeguards data integrity.

Step 3: Standardized the data formatting throughout the worksheet to keep it consistent and presentable, such as aligning date formats and setting numerical values to two decimal places.

Step 4: Removed extra spaces within data fields, converted text to uppercase, and inserted spaces where needed to maintain uniformity. This was achieved using the TRIM() function in Excel.

These steps were applied across all datasets to ensure clean, reliable data for analysis.

Module 2: Data Querying and Analysis Using PostgreSQL

- Moving on to PostgreSQL, several steps were followed to load the data into the server:
- **Step 1:** Created a database and tables, then populated them using the BI-dump.sql file, which contains datasets such as sales.csv and product.csv.
- **Step 2:** In the command-line interface, a new database named *BICapstone* was created and connected by running the command \connect capstone. The tables and data from BI-dump.sql were restored using the command \include BI-dump.sql.
- **Step 3:** Verified that the data loaded correctly using *pgAdmin*, the graphical interface for PostgreSQL. For each table, ran the query SELECT * FROM table_name LIMIT 5; to confirm the presence and accuracy of the data.
- **Step 4:** Performed sales analysis using features such as *Cubes* and *Rollups* to calculate total sales, as well as breakdowns by product, city, and store.

Module 3: Data Visualization and Statistical Analysis

- **Regression Analysis: A Step-by-Step Guide**
- The core purpose of this statistical analysis is to interpret regression results to understand how sales are influenced by various factors such as **date**, **revenue**, **stock**, and **price**. This process provides a detailed report on sales performance across these key variables.
- **The Process**
- **Define Variables**: Begin by identifying **sales** as the dependent variable. Analyze how other variables correlate with sales to find any strong relationships.
- **Run the Analysis**: Use a regression analysis tool, such as the one in Excel, to generate a summary of the output that will explain the relationships between the variables.
- **Use a Sample Dataset**: To save time and computational resources, analyze a **sample** of about 20,000 rows from the dataset instead of the entire dataset.
- **Interpret Coefficients**: Examine the coefficients to determine the expected change in sales for a one-unit increase in each independent variable.
- **Assess Significance**: Evaluate the **p-values** (significance levels) to confirm if the influence of the independent variables is statistically significant.
- **Evaluate Model Fit**: Analyze the **R-squared** value to understand the proportion of the variance in sales that can be explained by the independent variables.

Module 4: Data Visualization and Dashboards Using Tableau

The next task is to develop an **interactive dashboard** that effectively visualizes the entire story through charts. This process involves a structured approach to data handling, visualization, and final presentation.

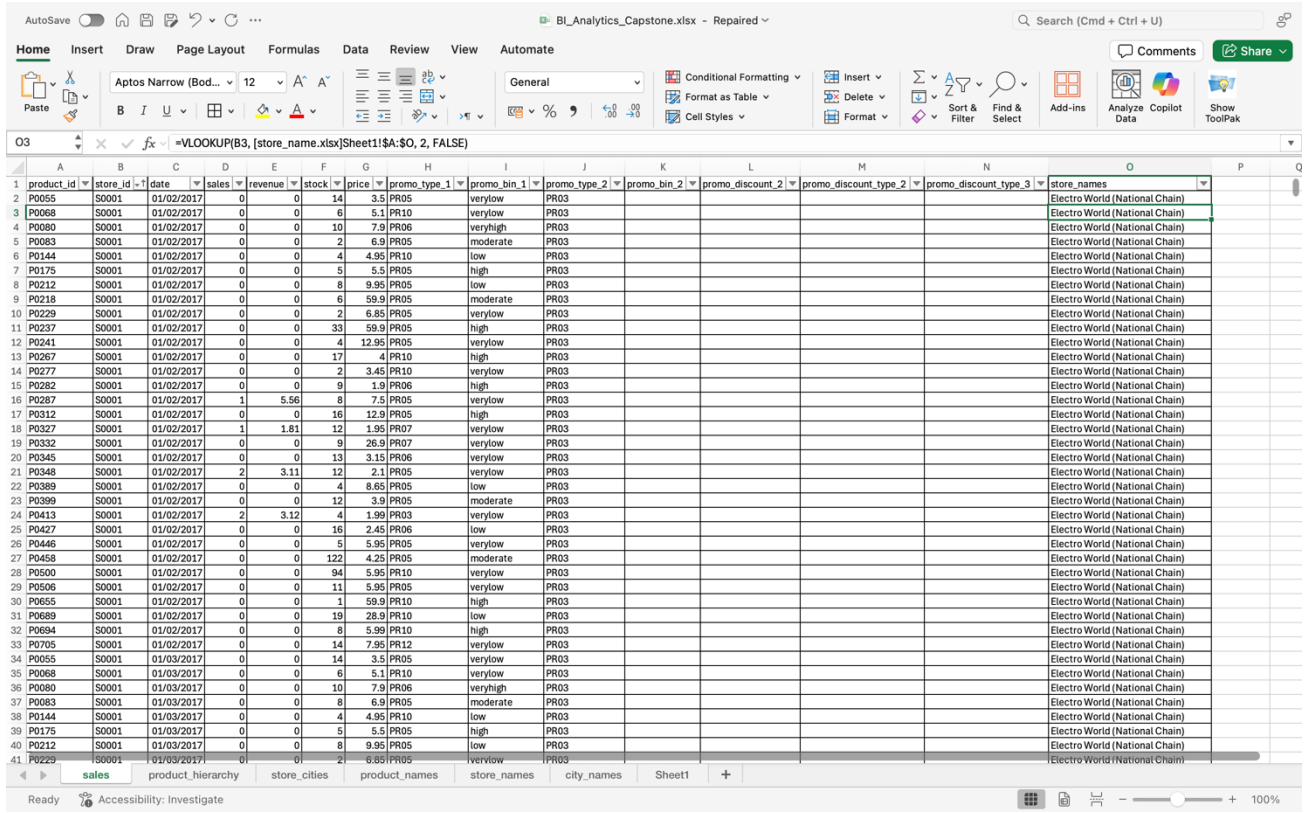
Step-by-Step Process

- **Data Integration and Preparation**: The first step is to **import data** into a business intelligence tool like **Tableau**. Once the data is loaded, you must establish the correct relationships between the variables to ensure the output is accurate. You can also create **calculated fields** at this stage to simplify complex calculations and aggregations for later use in visualizations.
- **Visualization Development**: Next, you will create individual sheets to visualize the data. This involves building various **charts**, such as line charts, bar charts, and bubble charts, to represent the data in a clear and compelling manner. Each chart should be designed to tell a specific part of the story.
- **Dashboard Assembly**: After creating the individual sheets and charts, they are combined into a single **dashboard**. The goal is to arrange these sheets in a logical flow to present a cohesive and insightful narrative for the audience.
- **Enhancing Interactivity**: The final step is to refine the dashboard's user experience. This is achieved by utilizing Tableau's features like **filters** and **highlighters**. These features allow stakeholders to interact with the data directly; for example, clicking on an item in one chart will **highlight** or **filter** the corresponding data points in all other related charts, making the analysis intuitive and easy to understand.

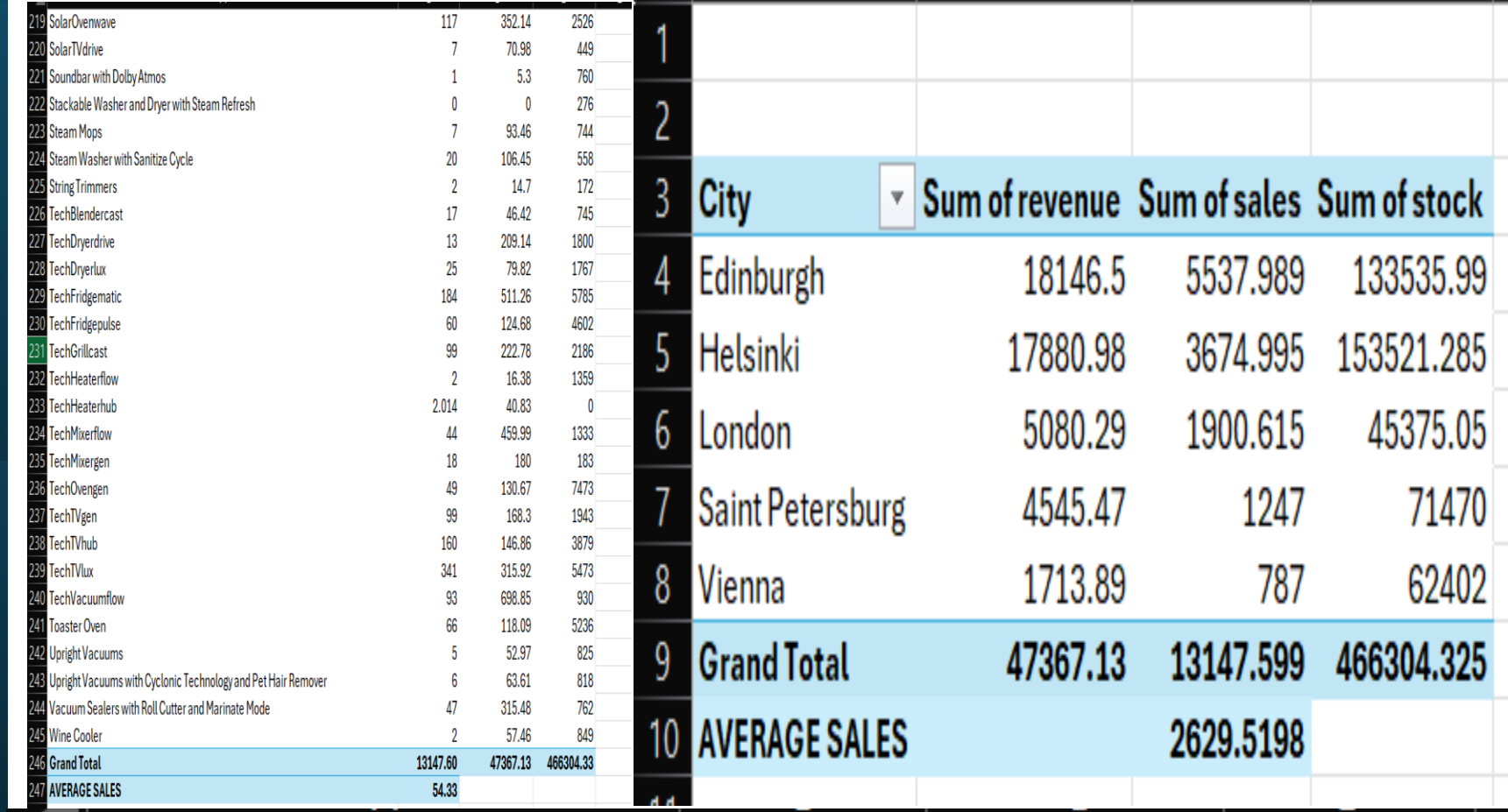
Results

Module 1
Lesson 1: Data Cleaning and Preparation

Data cleaning and preparation are fundamental to any successful data project, whether it's for analysis, machine learning, or generating business insights. These processes are crucial because they ensure the data used is **accurate**, **consistent**, and in a **usable** format.
Properly prepared data serves as a reliable foundation for all subsequent steps. This includes building accurate predictive models and making informed, data-driven decisions. Without this critical phase, the quality and reliability of any conclusions or outputs would be compromised, potentially leading to flawed results.



Module 1, Lesson 2: Data Analysis Using Pivot Tables

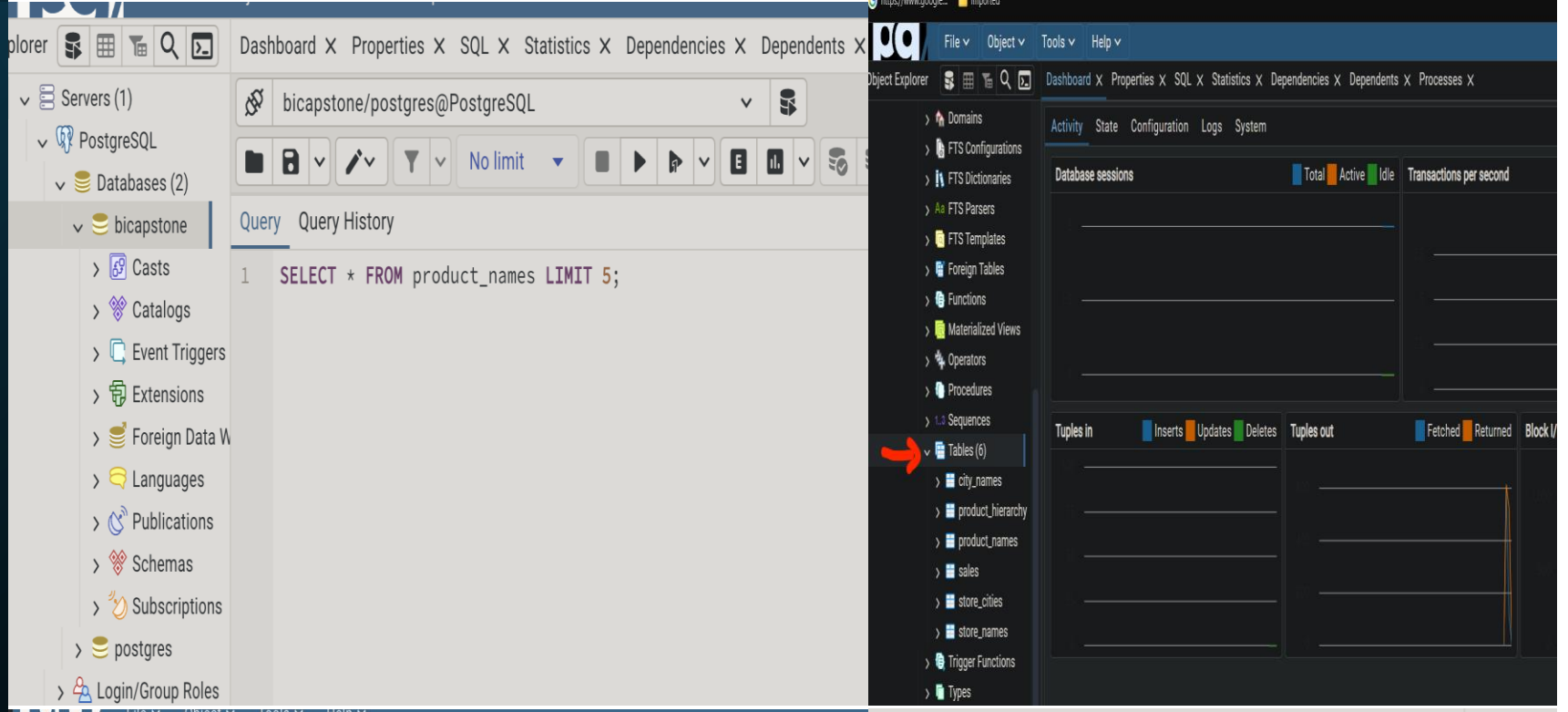


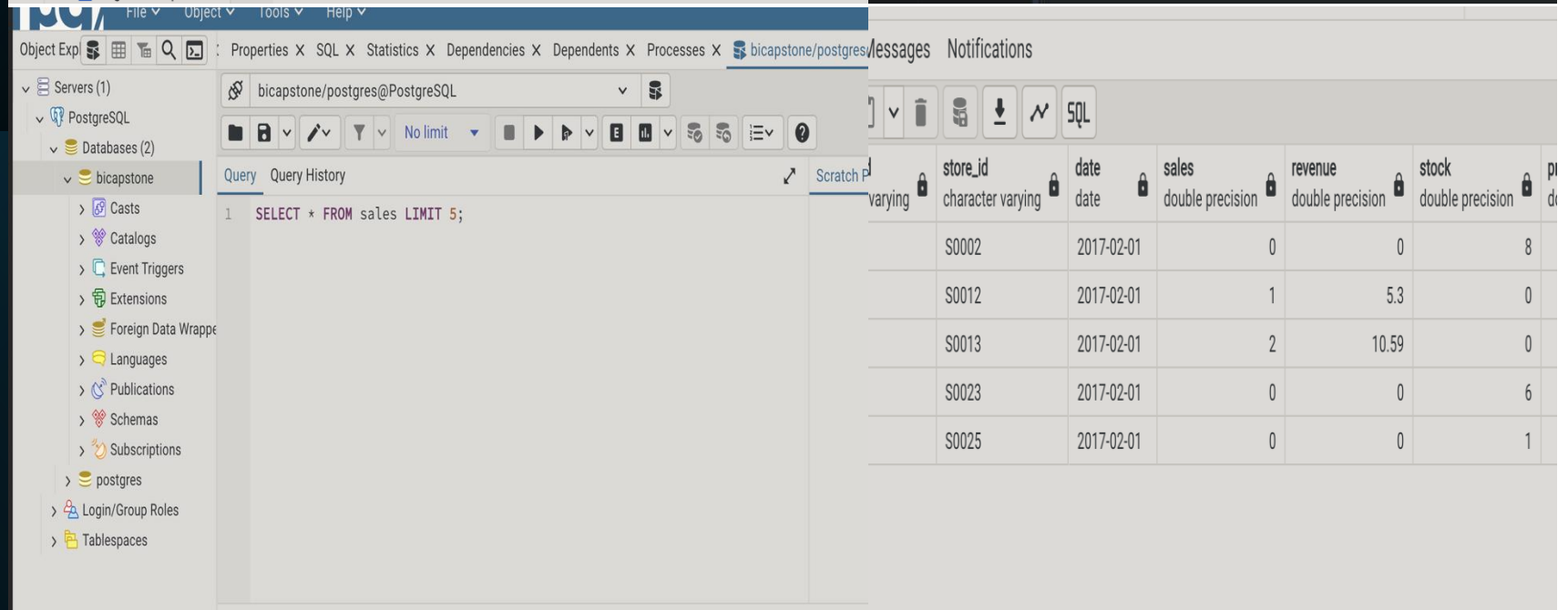
219	SolarOvenwave	117	352.14	2526
220	SolarTVdrive	7	70.98	449
221	Soundbar with Dolby Atmos	1	5.3	760
222	Stackable Washer and Dryer with Steam Refresh	0	0	276
223	Steam Mops	7	93.46	744
224	Steam Washer with Sanitize Cycle	20	106.45	558
225	String Trimmers	2	14.7	172
226	TechBlendercast	17	46.42	745
227	TechDryerdrive	13	209.14	1800
228	TechDryerlux	25	79.82	1767
229	TechFridgematic	184	511.26	5785
230	TechFridgepulse	60	124.68	4602
231	TechGrillcast	99	222.78	2186
232	TechHeaterflow	2	16.38	1359
233	TechHeaterhub	2.014	40.83	0
234	TechMixerflow	44	459.99	1333
235	TechMixergen	18	180	183
236	TechOvengen	49	130.67	7473
237	TechTVgen	99	168.3	1943
238	TechTVhub	160	146.86	3879
239	TechTVlux	341	315.92	5473
240	TechVacuumflow	93	698.85	930
241	Toaster Oven	66	118.09	5236
242	Upright Vacuums	5	52.97	825
243	Upright Vacuums with Cyclonic Technology and Pet Hair Remover	6	63.61	818
244	Vacuum Sealers with Roll Cutter and Marinate Mode	47	315.48	762
245	Wine Cooler	2	57.46	849
246	**Grand Total**	**13147.60**	**47367.13**	**466304.33**
247	**AVERAGE SALES**	**54.33**		

	City	Sum of revenue	Sum of sales	Sum of stock
4	Edinburgh	18146.5	5537.989	133535.99
5	Helsinki	17880.98	3674.995	153521.285
6	London	5080.29	1900.615	45375.05
7	Saint Petersburg	4545.47	1247	71470
8	Vienna	1713.89	787	62402
9	**Grand Total**	**47367.13**	**13147.599**	**466304.325**
10	**AVERAGE SALES**		**2629.5198**	

	Store_name	Sum of sales	Sum of revenue	Sum of stock
4	Currys (National Chain)	389	1055.21	31473
5	Darty	1900.615	5080.29	45375.05
6	DIGI	5537.989	18146.5	133535.99
7	Electro World (National Chain)	3674.995	17880.98	153521.285
8	Elettrodomestici Rossi	1247	4545.47	71470
9	Euronics Lisboa (National Chain)	398	658.68	30929
10	**Grand Total**	**13147.599**	**47367.13**	**466304.325**
11	**AVERAGE_SALES**	**2191.27**		

Module 2 Lesson 1: Data Querying Using PostgreSQL





Module 2, Lesson 2: Data Analysis Using PostgreSQL

bicapstone/postgres@PostgreSQL

No limit

Query | Query History

```sql
SELECT
  p.hierarchy1_id,
  p.hierarchy2_id,
  SUM(sl.sales) AS total_sales
FROM
  sales sl
JOIN
  product_hierarchy p
ON
  sl.product_id = p.product_id
GROUP BY rollup
 (p.hierarchy1_id,
  p.hierarchy2_id);
```

Data Output | Messages | Notifications

	hierarchy1_id character varying	hierarchy2_id character varying	total_sales double precision
1	[null]	[null]	9856.880999999998
2	H03	H0317	2
3	H00	H0001	617
4	H03	H0316	0
5	H03	H0314	190
6	H01	H0107	374

```sql
  TO_CHAR(sl.date, 'YYYY-MM') AS fmt_date,
  SUM(sl.sales) AS total_sales
FROM
  sales sl
GROUP BY rollup
  (sl.store_id,
  fmt_date)
```

Data Output | Messages | Notifications

	store_id character varying	fmt_date text	total_sales double precision
1	[null]	[null]	9856.880999999998
2	S0052	2017-02	23.675
3	S0120	2017-02	19
4	S0135	2017-02	32.305
5	S0024	2017-02	85
6	S0127	2017-03	9
7	S0107	2017-03	27
8	S0072	2017-02	80
9	S0102	2017-03	23
10	S0082	2017-03	42

Tools | Help

Dependencies | Dependents | Processes | bicapstone/postgres@PostgreSQL* | public.c

bicapstone/postgres@PostgreSQL

No limit

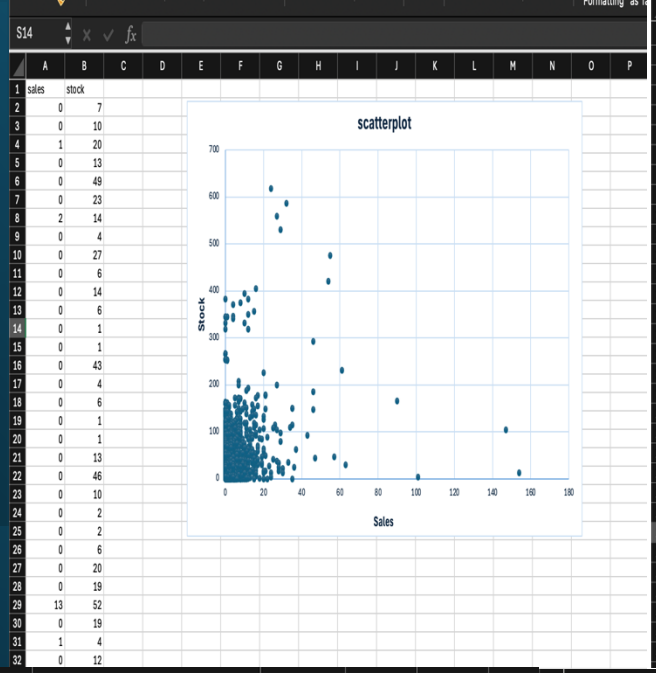
Query | Query History

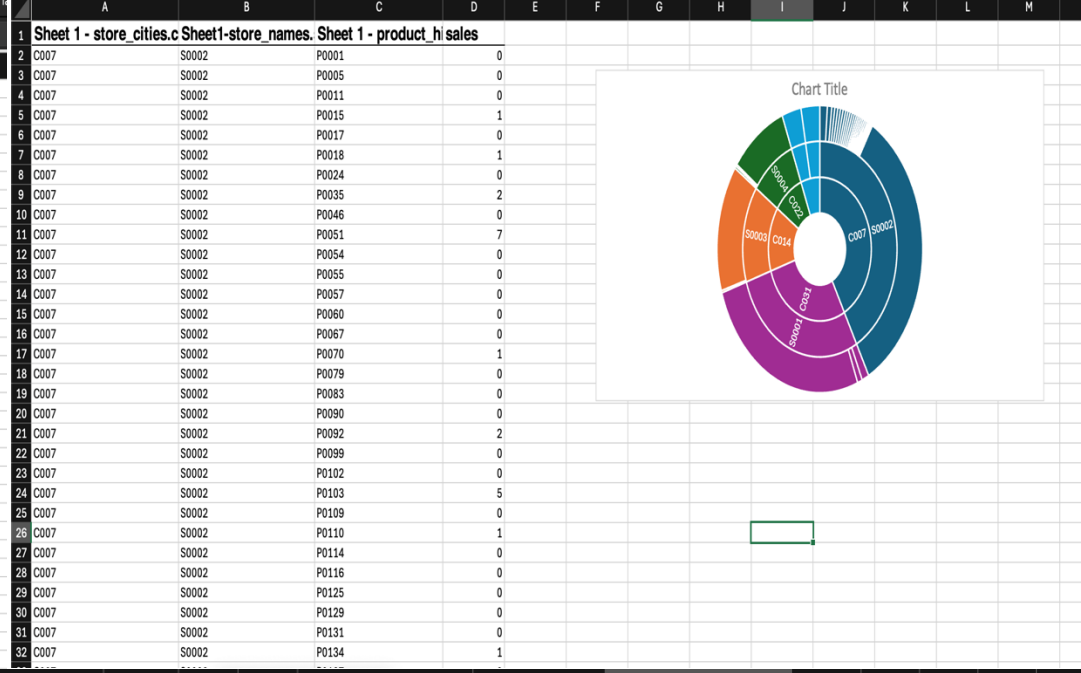
```sql
SELECT
  c.city_id,
  cn.city_name,
  TO_CHAR(s.date, 'YYYY-MM') AS month_sale,
  SUM(s.sales) AS total_sales_sum
FROM
  sales s

JOIN
  store_cities c
ON
  s.store_id = c.store_id
  JOIN
  city_names cn
 ON
  c.city_id=cn.city_id
GROUP BY
  CUBE (c.city_id, cn.city_name,TO_CHAR(s.date, 'YYYY-MM'));
```

Data Output | Messages | Notifications

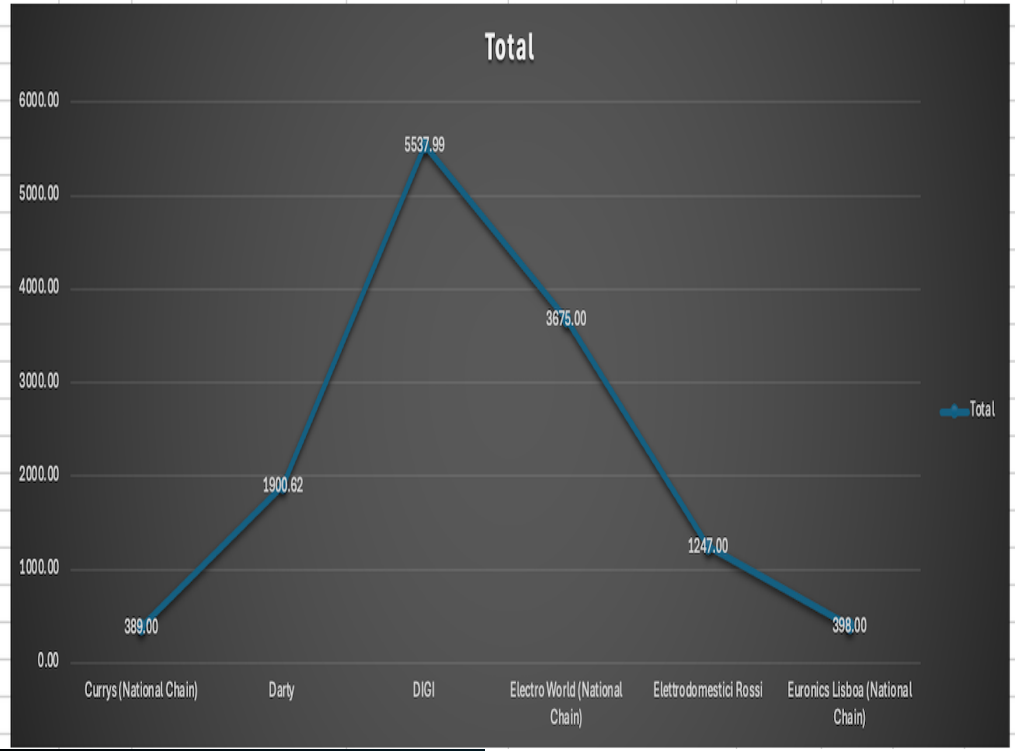
	city_id character varying	city_name character varying (255)	month_sale text	total_sales_sum double precision
1	C002	Berlin	2017-02	100
2	C002	Berlin	2017-03	59
3	C002	Berlin	[null]	159
4	C002	[null]	[null]	159
5	C003	Barcelona	2017-02	32.84
6	C003	Barcelona	2017-03	14
7	C003	Barcelona	[null]	46.84
8	C003	[null]	[null]	46.84
9	C004	Budapest	2017-02	223.49

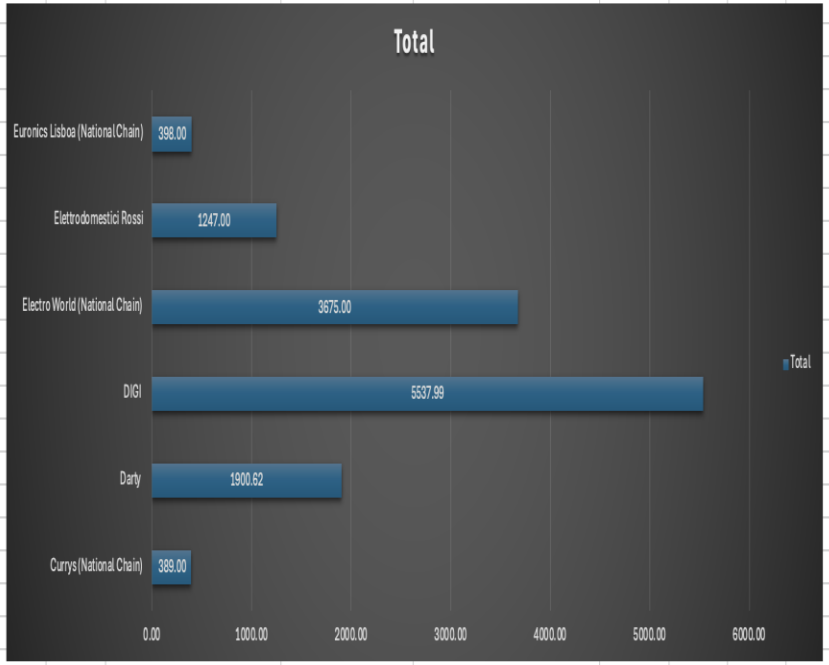
Module 3, Lesson 1: Data Visualization Using Excel





Row Labels	Sum of sales
Currys (National Chain)	389.00
Darty	1900.62
DIGI	5537.99
Electro World (National Chain)	3675.00
Elettrodomestici Rossi	1247.00
Euronics Lisboa (National Chain)	398.00
Grand Total	**13147.60**





Module 3, Lesson 2: Statistical Analysis

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.73269124
R Square	0.53683645
Adjusted R S	0.53674379
Standard Err	1.90960891
Observations	19999

ANOVA

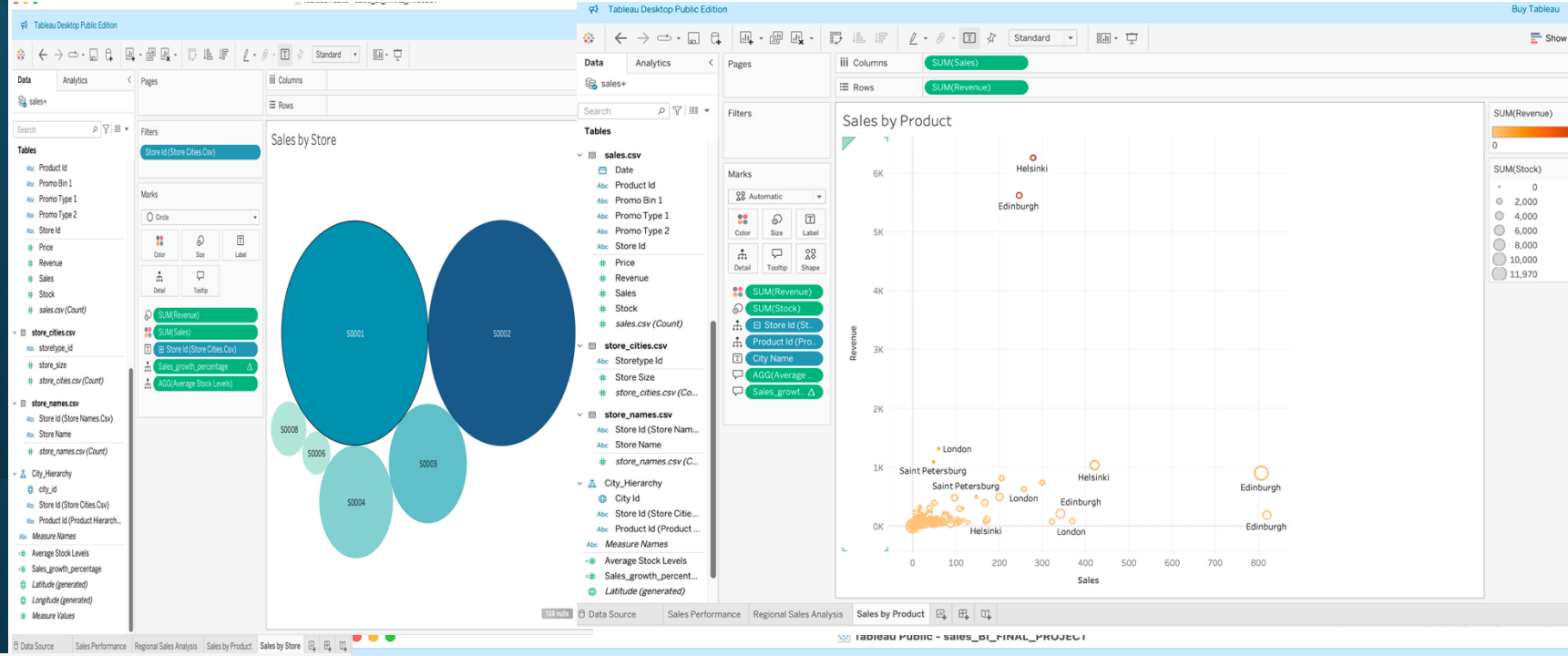
	df	SS	MS	F	Significance F
Regression	4	84507.6798	21126.92	5793.5842	0
Residual	19994	72910.2439	3.64660618		
Total	19998	157417.924			

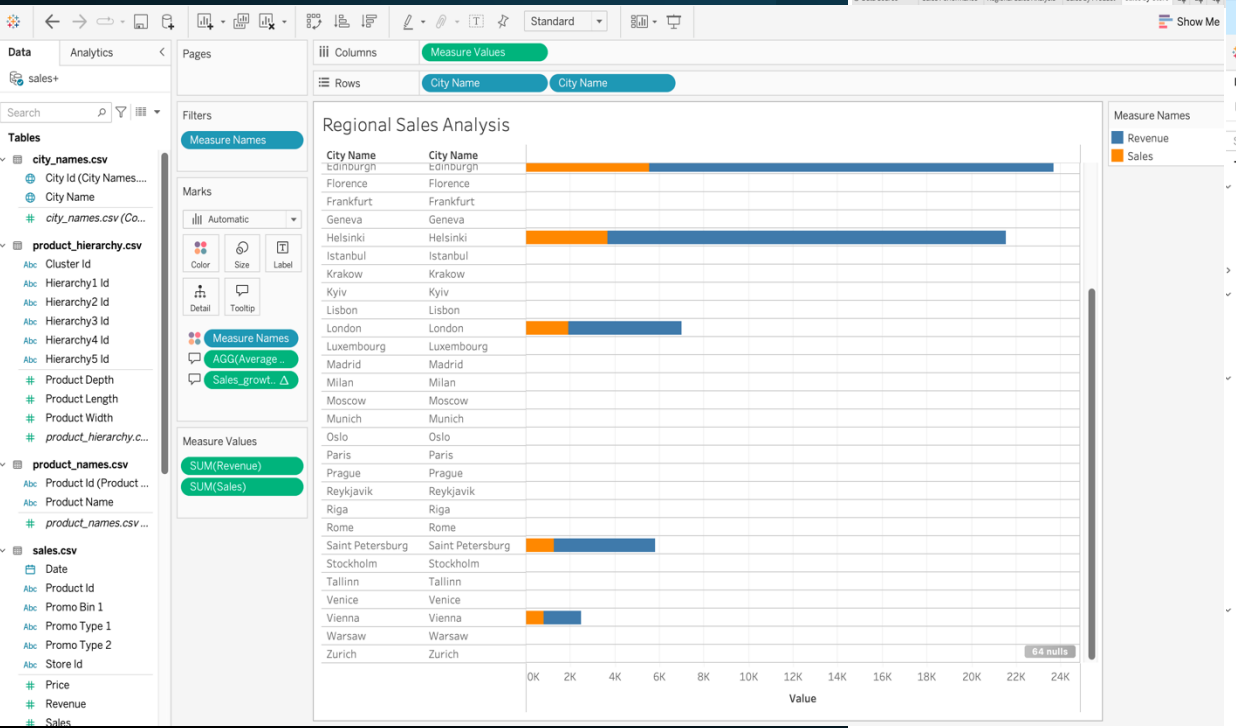
	Coefficients	Standard Err	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-0.5933672	44.846394	-0.0132311	0.98944355	-88.496006	87.3092711	-88.496006	87.3092711
date	1.88E-05	0.00104885	0.0179283	0.98569623	-0.002037	0.00207463	-0.002037	0.00207463
revenue	0.04606812	0.00032757	140.636675	0	0.04542606	0.04671018	0.04542606	0.04671018
stock	0.02416673	0.00049781	48.5463819	0	0.02319098	0.02514247	0.02319098	0.02514247
price	-0.0148461	0.00107736	-13.780087	5.28E-43	-0.0169579	-0.0127344	-0.0169579	-0.0127344

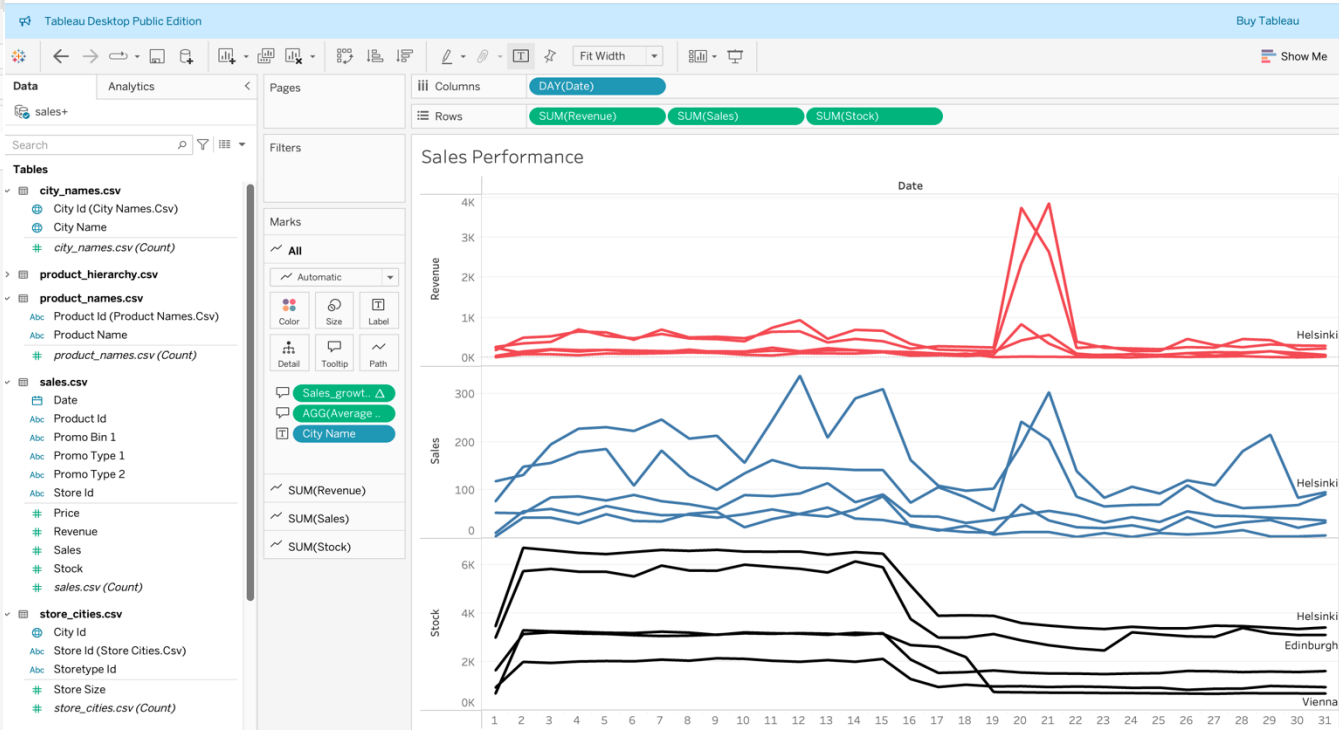
- **R-Square (0.5368):**
  - About 53.68% of the variability in the dependent variable (likely "sales") is explained by the independent variables (date, revenue, stock, price). This suggests a decent model fit but indicates that other factors not included in the model might explain the remaining 46.32%.
  - **Interpretation of Predictors:**

1. **Intercept (-0.5934):**
   - Represents the predicted value of the dependent variable when all independent variables are zero. This has no practical interpretation here due to the high p-value (0.9894), indicating it is not statistically significant.
2. **Date (Coefficient = 1.8804E-05, p-value = 0.9857):**
   - This predictor is not statistically significant (p-value > 0.05). It likely has no meaningful impact on the dependent variable in this model.
3. **Revenue (Coefficient = 0.0461, p-value = 0):**
   - This is a highly significant predictor (p-value < 0.05).
   - For every unit increase in revenue, the dependent variable (e.g., sales) increases by **0.0461** units, holding all other variables constant.
4. **Stock (Coefficient = 0.0242, p-value = 0):**
   - Another highly significant predictor.
   - For every unit increase in stock, the dependent variable increases by **0.0242** units, holding other variables constant.
5. **Price (Coefficient = -0.0148, p-value = 5.28E-43):**
   - This is also highly significant.
   - For every unit increase in price, the dependent variable decreases by **0.0148** units, holding other variables constant. This suggests that as prices increase, sales decrease, which aligns with typical economic behavior.

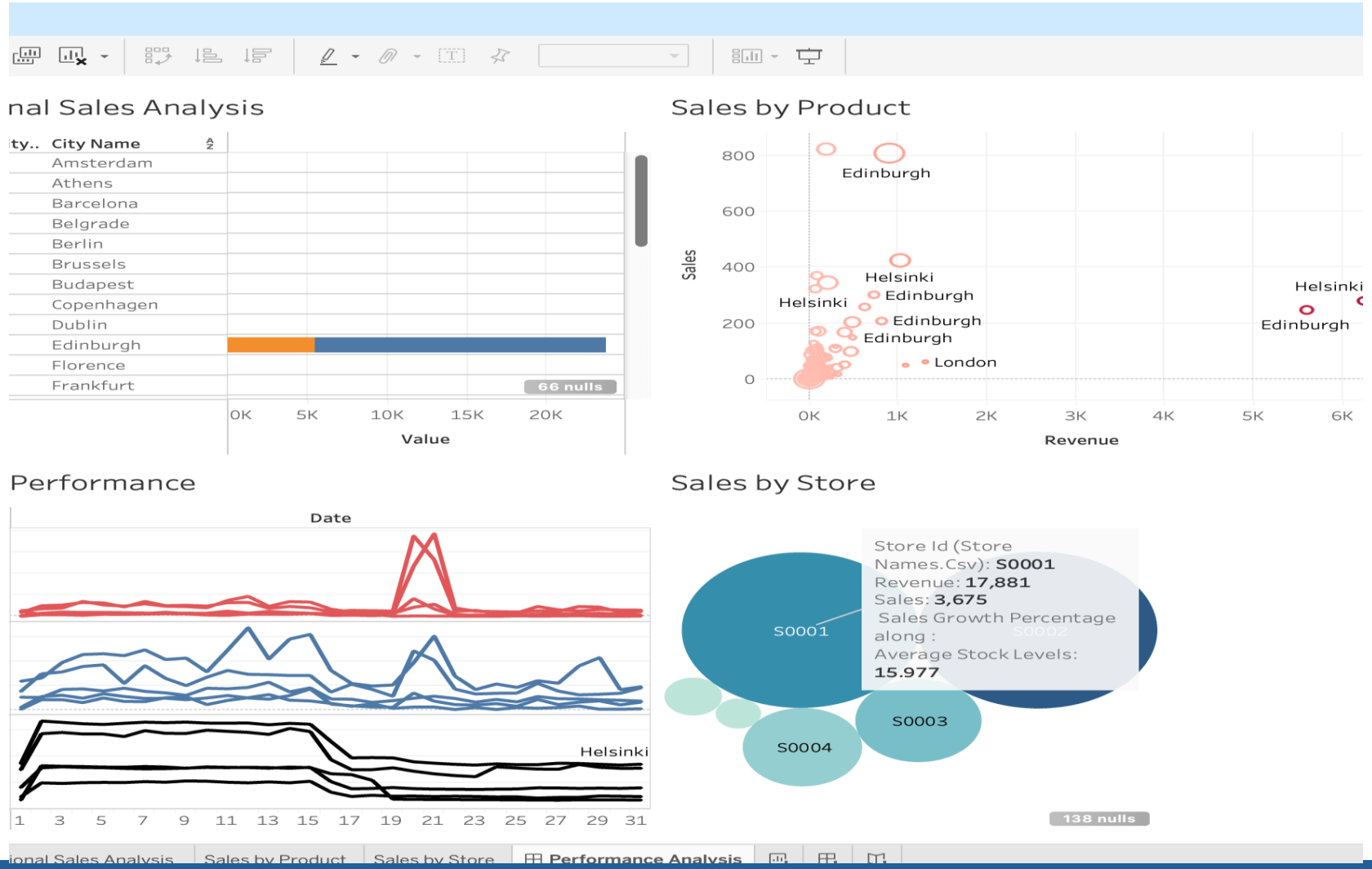
Module 4, Lesson 1: Basic Tableau Visualizations







Module 4, Lesson 2: Advanced Visualizations Using Tableau



Discussion

Insights and Recommendations

Revenue is the Most Important Factor:

The analysis shows that revenue has a strong and direct impact on sales. As revenue increases, sales also increase significantly.

Implication: The retail chain should focus on strategies to boost revenue, such as promotions, targeted marketing, and expanding product offerings.

Having more stock available contributes positively to sales. When stock levels increase, sales also go up.

Implication: Ensuring that popular products are always in stock can help maintain and boost sales.

Higher prices tend to reduce sales. Even small price increases can lead to a noticeable drop in sales.

Implication: The retail chain should consider competitive pricing strategies and discount offers to attract more customers.

Conclusion

Summary

Analysis of Sales Performance Across Cities

Based on the provided key findings, here is a rewritten summary and conclusion, emphasizing clarity and actionable insights.

Key Findings

The analysis reveals distinct performance patterns across different cities:

- **Helsinki and Edinburgh** show strong revenue despite moderate sales volumes. This suggests that these locations are either selling **high-value products** or have adopted successful **premium pricing strategies**.
- In contrast, **London and Saint Petersburg** generate lower revenue relative to their high sales volume. This indicates a potential reliance on **lower-priced products** or smaller transaction values.
- **Stock management** appears to be a key area for improvement. Some cities maintain high stock levels without a corresponding increase in revenue, which points to potential inefficiencies in inventory and logistics.

Conclusion and Recommendations

The company's strategy should focus on two primary areas for optimization:

- **Pricing Strategy**: Review and adjust the pricing model in cities like **London and Saint Petersburg** to increase revenue without compromising sales volume. The success of Helsinki and Edinburgh can serve as a benchmark for this effort.
- **Stock Management**: Improve inventory allocation by linking **stock levels to revenue potential**. This will help prevent overstocking in low-revenue cities and ensure that high-revenue locations have adequate supply to meet demand.