BI Analyst Capstone Project

Jody Upright

Table of Contents

Introduction

Introduction

- This Turkish retail company is a leading chain that operates globally.
- This company is committed to customer satisfaction and offers a diverse range of products.
- This company is seeking to leverage data-driven insights to maintain its competitive edge, and enhance sales strategies and operational efficiency.

Methodology

Data Details

- This data coming from a Turkish retail company was originally published on Kaggle
- The original data spans the years 2017-2019
- Data was provided across six spreadsheets in normalized format spanning the following categories:
  - Sales – delineated by date, product, and store
  - Product Hierarchy – details product dimensions and categories
  - Product Names – organized by product id
  - Store names – organized by store id
  - Store cities – organized by store id
  - City Names – organized by city id

Module 1: Data Cleaning and Preparation Using Excel

Part 1: Data Cleaning

- The Sales data was limited to 50000 rows to accommodate limitations of Excel for the Web
- Six spreadsheets were combined into a single .xlsx file
- 710 rows missing price data were filtered out of the data
- Adjusted number formatting for consistency, including date format
- Looked for duplicates on City and Product spreadsheets

Part 2: Pivot Tables

- The Sales data was further limited to 30000 rows
- VLOOKUPs were used to combine the six spreadsheets into a single spreadsheet
- Pivot tables were created to examine different sales scenarios
- Pivot tables were given filters to allow for additional drilling down

Module 2: Data Querying and Analysis Using PostgreSQL

Part 1: Data Querying

- A database with six tables was created
- SELECT statements were used to confirm correct table creation
- Sales performance was grouped by store and city for analysis

Part 2: Data Analysis

- Sales performance by store and month was analyzed with a rollup
- Sales performance by product hierarchy was analyzed with a rollup
- Sales performance by city and month was analyzed with a cube

Module 3: Data Visualization and Statistical Analysis

Part 1: Data Visualization

- Consolidated spreadsheets into a single sheet with VLOOKUPS
- Created a bar chart, line chart, and sunburst chart for various aspects of sales analysis
- Due to Excel limitations for scatter plots, unique points comparing stock and sales data were used to create a scatter plot and trendline

Part 2: Statistical Analysis

- XLMiner Analysis ToolPak was used for regression analysis
  - Due to Excel limitations, 20,000 rows were used
- Analysis performed to determined significance of the linear regression model

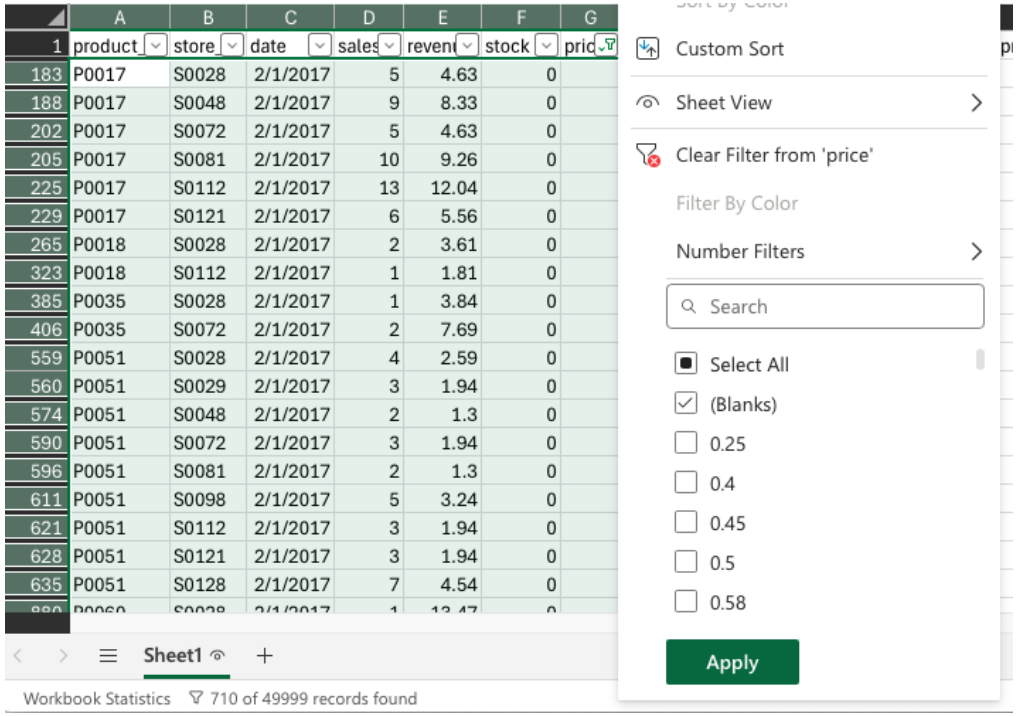
Module 4: Data Visualization and Dashboards Using Tableau

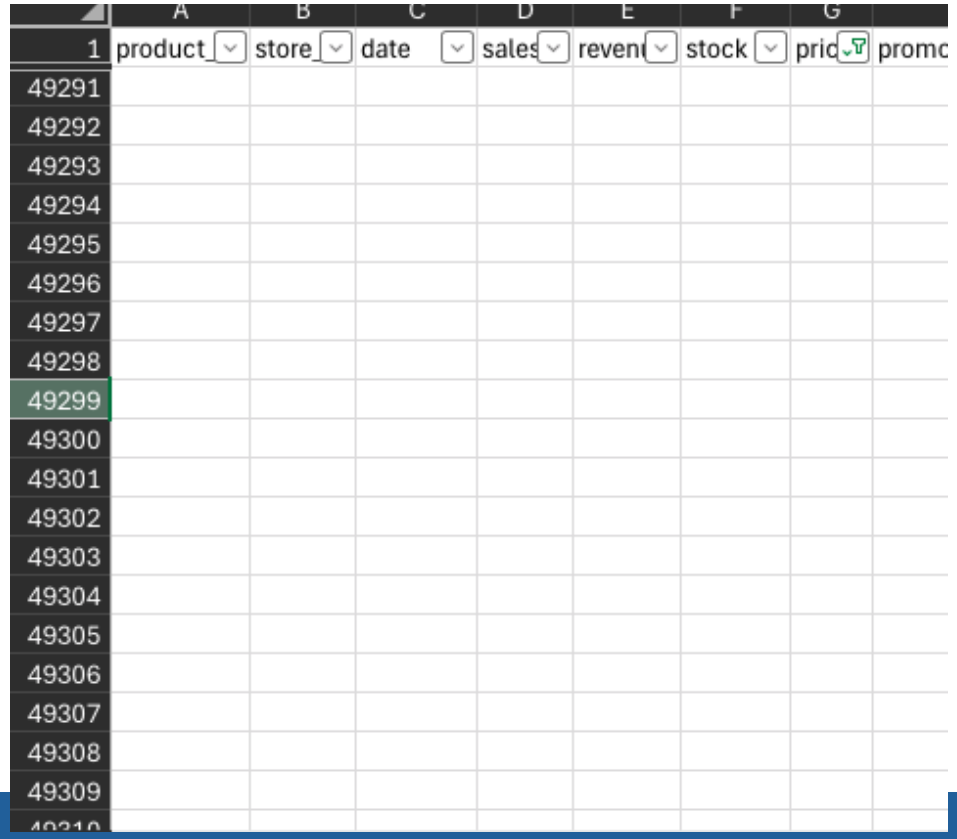
- Relationships across six spreadsheets created within Tableau Public
- Calculated fields created for Sales Growth Percentage and Average Stock Levels
- Converted date from dd/MM/yy to MM/dd/yy format for future time-based analysis
- Created various charts for sales analysis
- Compiled charts onto a single dashboard for side-by-side visual analysis

Results

Module 1, Lesson 1: Data Cleaning and Preparation

- Dropping rows with missing price values ensures missing data does not skew final results.





Module 1, Lesson 1: Data Cleaning and Preparation

- Extra spaces were removed from the data to ensure consistent formatting

	A	B	C	D
1	store_id	store_name		
2	S0005	MediaMarkt(National Chain)		
3	S0036	Saturn		
4	S0104	Euronics		
5	S0068	FNAC		
6	S0086	Darty		
7	S0038	Currys		

	store_id	store_name
2	S0005	MediaMarkt(National Chain)
3	S0036	Saturn
4	S0104	Euronics
5	S0068	FNAC
6	S0086	Darty
7	S0038	Currys
8	S0012	El Corte Inglz

Module 1, Lesson 1: Data Cleaning and Preparation

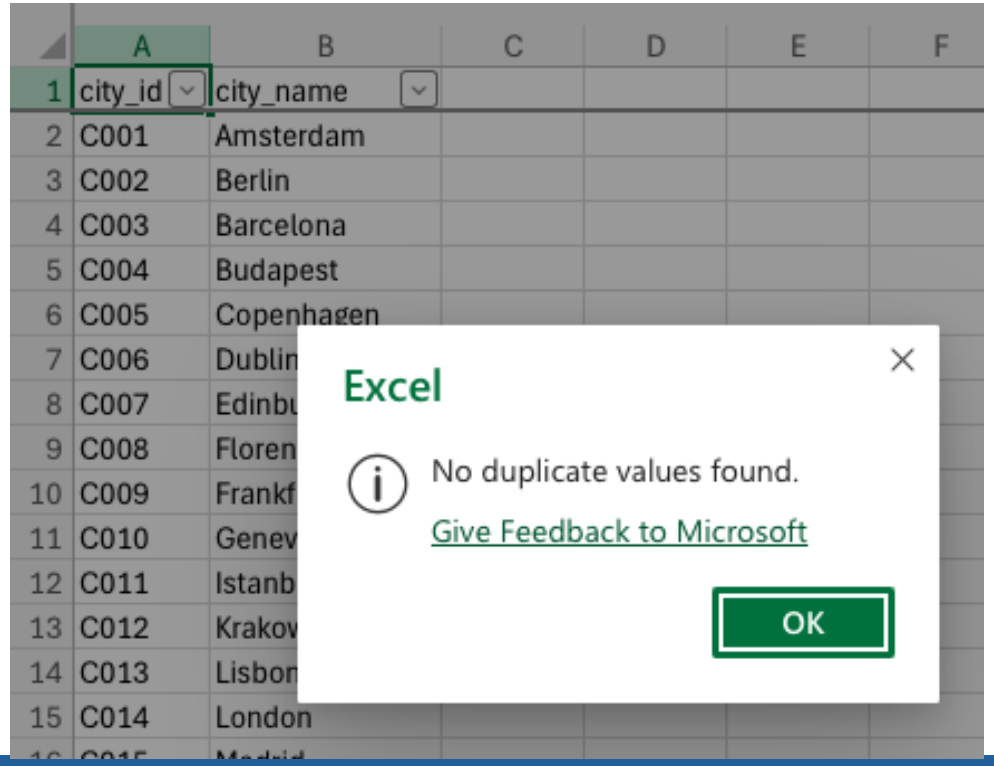
- Dates were converted from dd/MM/yyyy format to MM/dd/yyyy format
- This step was taken because Excel was reading 2/1/2017 as February 1 when the original source had this as January 2
- Dates will be used to analyze sales through the passage of time

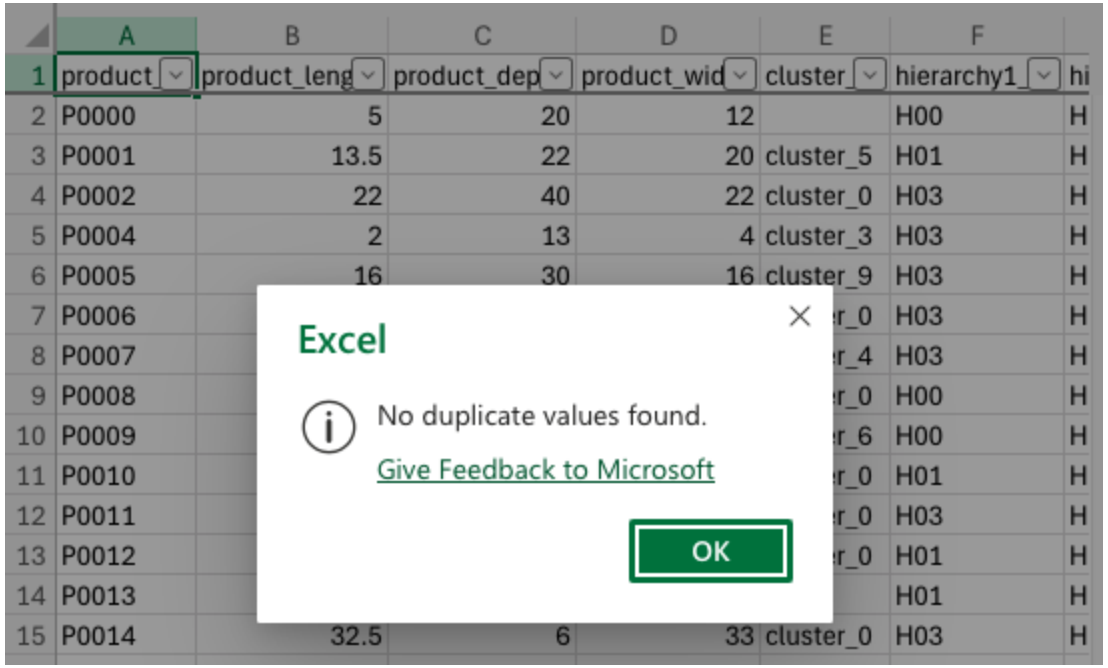
	A	B	C	D	
1	product_id	store_id	date	sales	re
2	P0001	S0002	2/1/2017	0	
3	P0001	S0012	2/1/2017	1	
4	P0001	S0013	2/1/2017	2	
5	P0001	S0023	2/1/2017	0	
6	P0001	S0025	2/1/2017	0	
7	P0001	S0027	2/1/2017	0	
8	P0001	S0040	2/1/2017	0	

	A	B	C	D	
1	product_id	store_id	date	sales	rev
2	P0001	S0002	1/2/2017	0	
3	P0001	S0012	1/2/2017	1	
4	P0001	S0013	1/2/2017	2	
5	P0001	S0023	1/2/2017	0	
6	P0001	S0025	1/2/2017	0	
7	P0001	S0027	1/2/2017	0	
8	P0001	S0040	1/2/2017	0	

Module 1, Lesson 1: Data Cleaning and Preparation

- The data was examined for duplicate values as duplicate values could impact regression analysis
- No duplicate values were found





Module 1, Lesson 2: Data Analysis Using Pivot Tables

- Sales by Store Analysis
- Jan. 2-18, 2017

3	store_name	date	Total sales	Total stock	Total revenu	Average of sales
4	⊞Currys (National Chain)		389	31473	1055.21	0.152429467
5	⊞Darty		1900.615	45375.05	5080.29	0.644494744
6	⊞DIGI		5537.989	133535.99	18146.5	0.675117518
7	⊞Electro World (National Chain)		3674.995	153521.285	17880.98	0.382453429
8	⊞Elettrodomestici Rossi		1247	71470	4545.47	0.245859621
9	⊞Euronics Lisboa (National Chain)		398	30929	658.68	0.246592317
10	**Grand Total**		**13147.599**	**466304.325**	**47367.13**	**0.438267909**

- DIGI and Electro World are the most profitable based on total sales and total revenue
- DIGI and Darty have the highest average of sales
- DIGI, Electro World, and Darty would be good candidates for future store expansion
- Currys and Euronics Lisboa have the lowest values in all categories and would be the candidates for targeted intervention to improve store performance

Module 1, Lesson 2: Data Analysis Using Pivot Tables

- Sales by City Analysis
- Jan. 2-18, 2017

city_name	date	Total sales	Total revenu	Average of sales
⊞Edinburgh		5537.989	18146.5	0.675117518
⊞Helsinki		3674.995	17880.98	0.382453429
⊞London		1900.615	5080.29	0.644494744
⊞Saint Petersburg		1247	4545.47	0.245859621
⊞Vienna		787	1713.89	0.188910226
Grand Total		**13147.599**	**47367.13**	**0.438267909**

- Sales figures for the top four cities align with sales figures for the four highest performing stores.
- Edinburgh and Helsinki have the highest total sales and revenue figure, while Edinburgh and London have the highest average sales.
- Edinburgh, Helsinki, and London would be good candidates for store expansion.
- Vienna has the lowest values in all three categories, combining sales figures from the two lowest-performing stores.

Module 1, Lesson 2: Data Analysis Using Pivot Tables

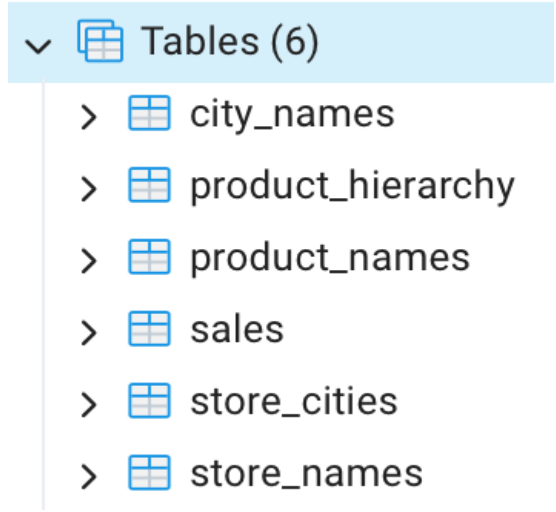
- Product Performance Analysis
- Jan. 2-18, 2017
- The top 15 products by total sales provided are items to continue stocking
- The bottom list of products that had no sales in the 17 day window are items to consider discontinuing sales, particularly if sales performance for these items has not shown improvement across later data

product_name	date	Total sales	Total revenu	Average of sales
NanoGrillflow		1649	381.68	6.842323651
Handheld Milk Frother		1218	2988.52	6.09
PowerDryerflow		867	966.95	6.826771654
Smart Speakers (with voice assistants)		700	440.13	2.834008097
Portable Heater		637	14378.98	3.283505155
CyberHeaterhub		499	231.03	2.070539419
TechTVlux		341	315.92	1.432773109
SmartFridgewave		338	800.06	1.402489627
SolarGrillpulse		276	1123.66	1.15
SolarBlenderlux		273.375	846.66	2.025
SmartFridgedrive		246	170.7	1.016528926
EcoVacuumcast		244	417.94	1.008264463
SmartVacuumcast		197	227.97	0.817427386
TechFridgematic		184	511.26	0.934010152
Pizza Oven		160	147.35	1.032258065

product_name	date	Total sales	Total revenu	Average of sales
MaxGrilltron		0	0	0
Stackable Washer and Dryer with Steam Refresh		0	0	0
Induction Cooktop		0	0	0
SmartMixergen		0	0	0
MaxFridgehub		0	0	0
CyberGrilldrive		0	0	0
AutoFridgetron		0	0	0
CyberWashercast		0	0	0
NanoDryermatic		0	0	0
EcoTVgen		0	0	0
NanoMixertron		0	0	0
Handheld Milk Frother with Adjustable Temperatures		0	0	0
NanoTVflow		0	0	0
CyberMixerhub		0	0	0
Patio Heaters		0	0	0
Electric Grills		0	0	0
Portable Speakers		0	0	0
EcoMixerhub		0	0	0
PowerVacuumgen		0	0	0
Chest Freezer with Lock and Key		0	0	0
Sewing Machines with Computerized Stitching and Automatic Needle Thread		0	0	0

Module 2, Lesson 1: Data Querying Using PostgreSQL

- A database with six tables was created using a dump file
- SELECT statements like the one below were used to verify the data was correctly imported.




```
1  SELECT * FROM sales LIMIT 5;
```

Data Output | Messages | Notifications

	product_id character varying	store_id character varying	date date	sales double precision	revenue double precision	stock double precision	price double precision	promo_type_1 character varying
1	P0001	S0002	2017-02-01	0	0	8	6.25	PR14
2	P0001	S0012	2017-02-01	1	5.3	0	6.25	PR14
3	P0001	S0013	2017-02-01	2	10.59	0	6.25	PR14
4	P0001	S0023	2017-02-01	0	0	6	6.25	PR14
5	P0001	S0025	2017-02-01	0	0	1	6.25	PR14

Module 2, Lesson 1: Data Querying Using PostgreSQL

- Sales performance totals by product, store, and city id were found using the SQL query to the right
- Setting up the database allows for analysis of larger, more complete datasets than what Excel is able to handle

Query | Query History | Scratch Pad

```
1  SELECT p.product_id,
2  c.store_id,
3  c.city_id,
4  SUM (sl.sales) AS total_sales,
5  SUM (sl.revenue) AS total_revenue,
6  AVG (sl.price) AS average_price
7  FROM sales sl
8  INNER JOIN product_hierarchy p ON sl.product_id = p.product_id
9  INNER JOIN store_cities c ON sl.store_id = c.store_id
10 GROUP BY p.product_id, c.store_id, c.city_id;
```

Data Output | Messages | Notifications

SQL

	product_id character varying	store_id character varying	city_id character varying	total_sales double precision	total_revenue double precision	average_price double precision
1	P0001	S0002	C007	0	0	6.25
2	P0001	S0012	C005	1	5.3	6.25
3	P0001	S0013	C026	2	10.59	6.25
4	P0001	S0023	C008	0	0	6.25
5	P0001	S0025	C024	0	0	6.25
6	P0001	S0027	C022	0	0	6.25
7	P0001	S0040	C017	0	0	6.25
8	P0001	S0049	C031	0	0	6.25
9	P0001	S0050	C014	0	0	6.25
10	P0001	S0051	C027	0	0	6.25
11	P0001	S0055	C014	0	0	6.25

Total rows: 1000 of 13220 Query complete 00:00:00.371 Ln 10, Col 46

Module 2, Lesson 2: Data Analysis Using PostgreSQL

- Sales by Store over time
- [null] values indicate grand totals for each store
- Stores generally saw a decrease in sales between February and March, in many cases 50% or more
- Maintaining stock levels equivalent to February sales would be ideal as some stores not show a substantial drop in sales
    - i.e. S0008 and S0029

Query | Query History

```
1  SELECT store_id,
2      TO_CHAR (date, 'YYYY-MM') AS month_of_sale,
3      SUM (sales) AS total_sales
4      FROM SALES sl
5      INNER JOIN product_hierarchy p ON sl.product_id = p.product_id
6      GROUP BY ROLLUP (store_id, month_of_sale)
7      ORDER BY store_id, month_of_sale;
```

Data Output | Messages | Notifications

	store_id character varying	month_of_sale text	total_sales double precision
1	S0001	2017-02	67.695
2	S0001	2017-03	35
3	S0001	[null]	102.695
4	S0002	2017-02	73.965
5	S0002	2017-03	28
6	S0002	[null]	101.965
7	S0003	2017-02	32
8	S0003	2017-03	20.835
9	S0003	[null]	52.835
10	S0004	2017-02	31
11	S0004	2017-03	8
12	S0004	[null]	39
13	S0006	2017-02	8
14	S0006	2017-03	4
15	S0006	[null]	12
16	S0008	2017-02	16
17	S0008	2017-03	15

Total rows: 352 of 352 Query complete 00:00:00.201 Ln 7, C

Module 2, Lesson 2: Data Analysi

- Sales by Product Hierarchy
- The five highest-performing sub-hierarchies are all in hierarchy H00
- Three subhierarchies in H03 had five or fewer sales; these would be candidates for either a sales promotion or for having their sales discontinued

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

	hierarchy1_id character varying	hierarchy2_id character varying	total_sales double precision
1	H00	H0000	1380
2	H00	H0001	617
3	H00	H0002	849
4	H00	H0003	4268
5	H00	H0004	613.366
6	H00	[null]	7727.366000000001
7	H01	H0105	175
8	H01	H0106	166
9	H01	H0107	374
10	H01	H0108	172
11	H01	[null]	887
12	H02	H0209	43.209999999999994
13	H02	H0210	182.30500000000004
14	H02	[null]	225.51500000000001
15	H03	H0311	17
16	H03	H0312	222
17	H03	H0313	581
18	H03	H0314	190
19	H03	H0315	5
20	H03	H0316	0
21	H03	H0317	2
22	H03	[null]	1017
23	[null]	[null]	9856.880999999998

Module 2, Lesson 2: Data Analysis Using PostgreSQL

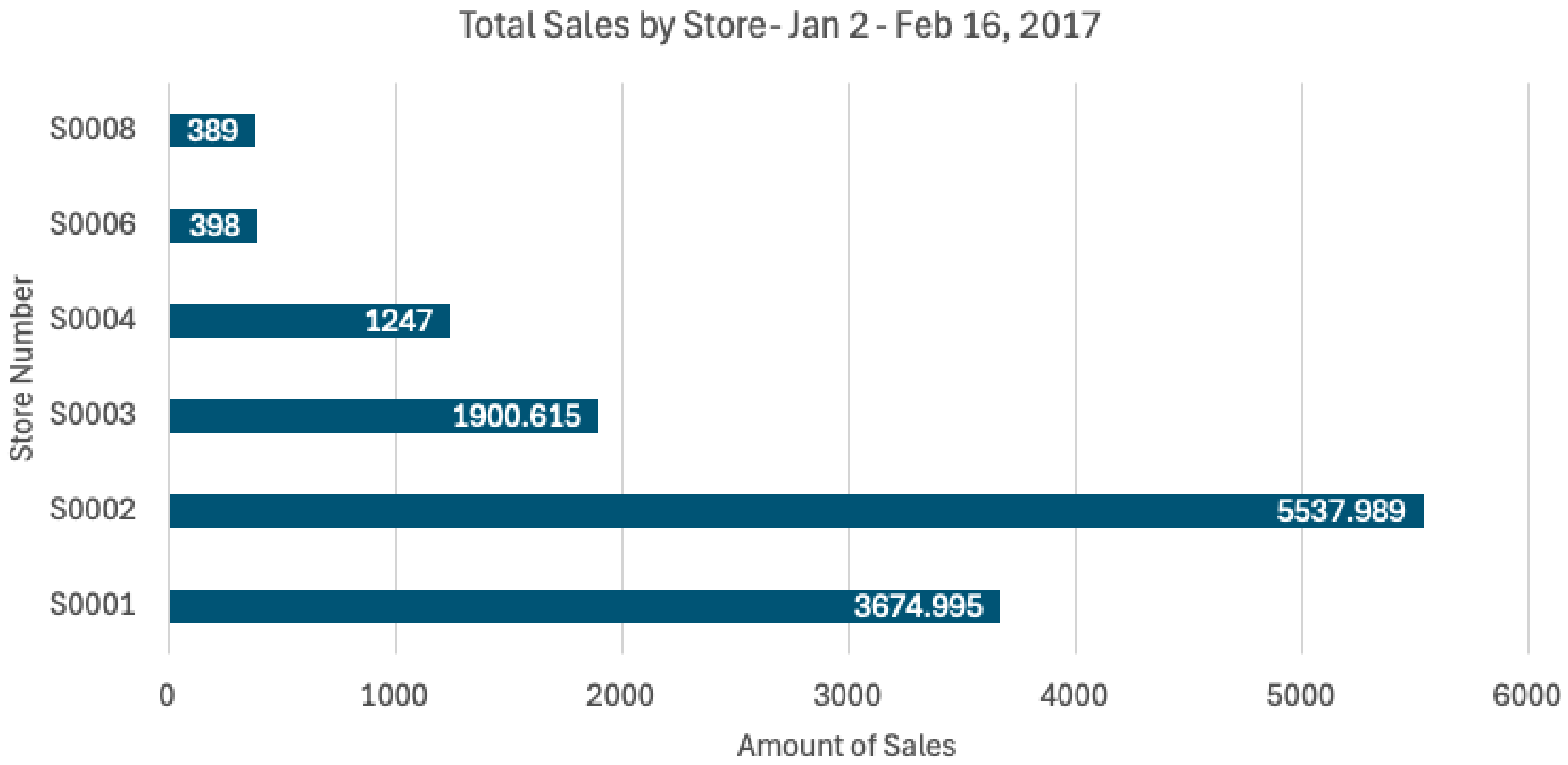
- Sales by City
- Most cities saw a 30-50% decrease in sales from February to March
- City C033 saw a 15% decrease in sales, the smallest percentage change of any city

```sql
SELECT c.city_id,
    TO_CHAR (date, 'YYYY-MM') AS month_of_sale,
    SUM (sales) AS total_sales
    FROM SALES sl
    INNER JOIN store_cities c ON sl.store_id = c.store_id
    GROUP BY CUBE (c.city_id, month_of_sale)
    ORDER BY c.city_id, month_of_sale;
```

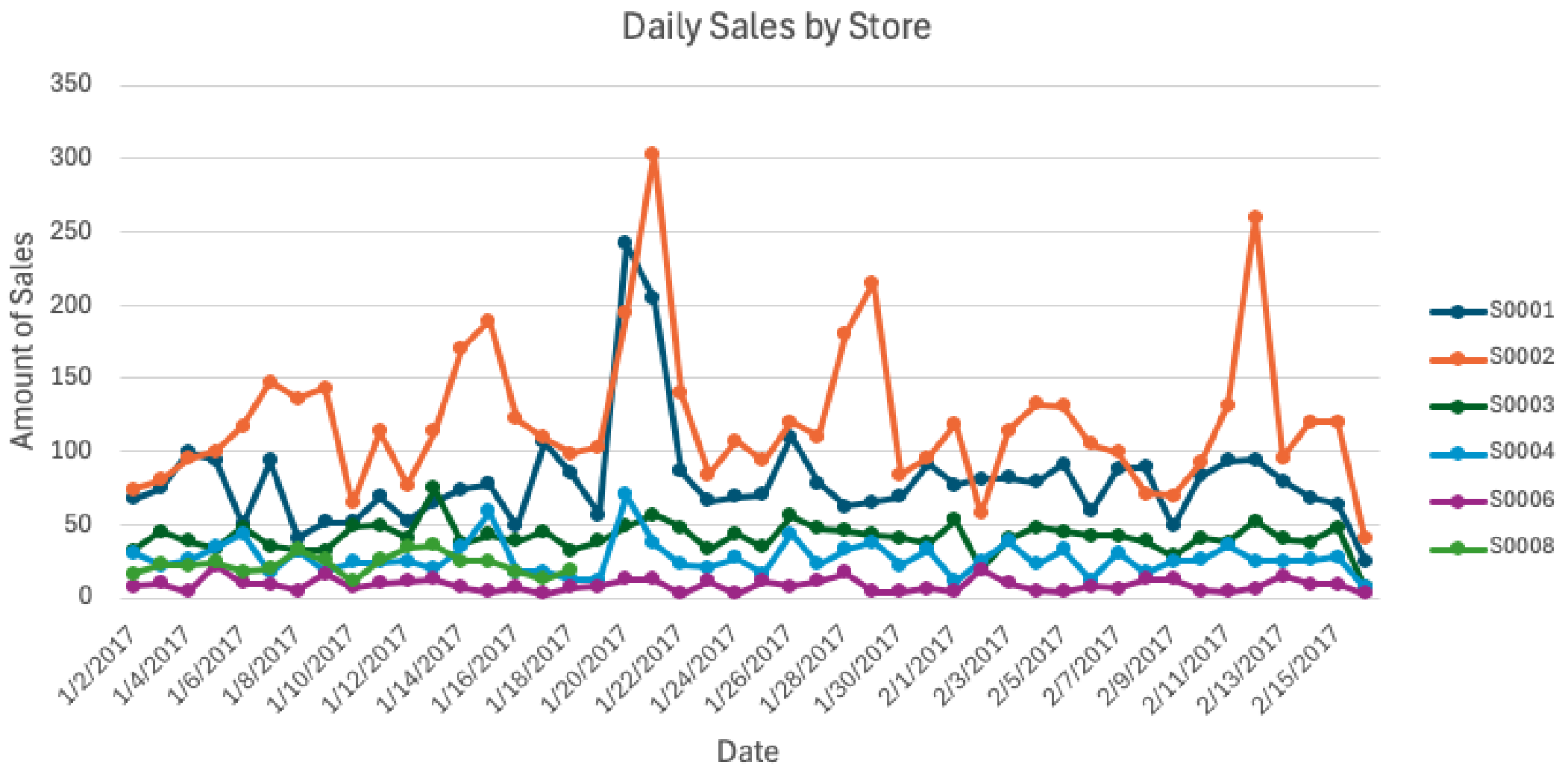
	city_id character varying	month_of_sale text	total_sales double precision
1	C002	2017-02	100
2	C002	2017-03	59
3	C002	[null]	159
4	C003	2017-02	32.84
5	C003	2017-03	14
6	C003	[null]	46.84
7	C004	2017-02	223.49
8	C004	2017-03	110.208
9	C004	[null]	333.69800000000004
10	C005	2017-02	54
11	C005	2017-03	35
12	C005	[null]	89
13	C006	2017-02	124
14	C006	2017-03	29

Total rows: 102 of 102 Query complete 00:00:00.306 Ln 7, C

Module 3, Lesson 1: Data Visualization Using Excel



Module 3, Lesson 1: Data Visualization Using Excel



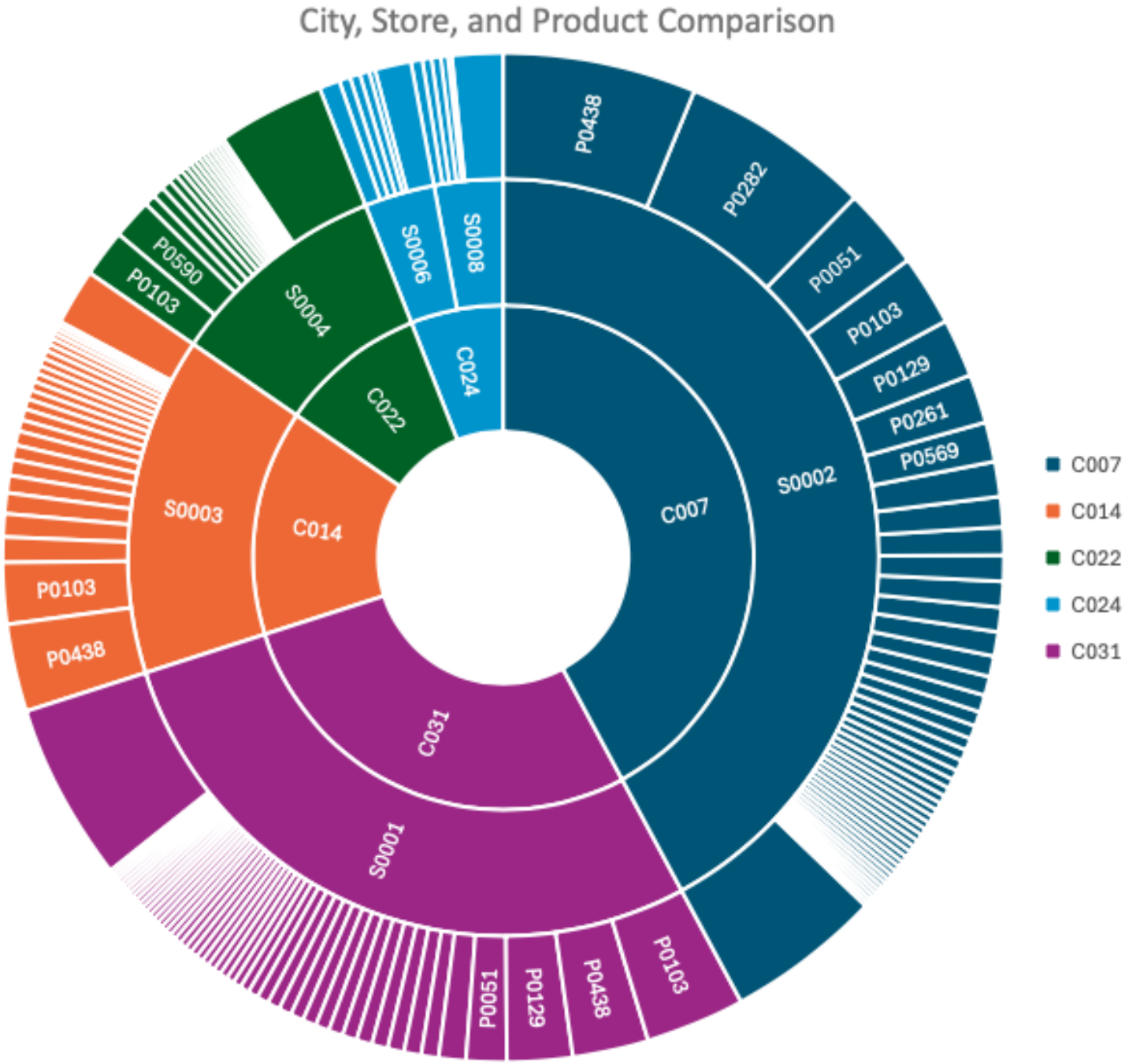
Module 3, Lesson 1: Data Visualization Using Excel

- The bar chart shows S0008 and S0006 having the lowest total sales figures
- The line chart shows S0006 having the lowest daily sales figures each day
- The line chart shows that S0008 has no sales data after January 18, 2017
  - This indicates either sales data got cut off abruptly, or a possible store closure
- S0002 had the highest total sales figure, and held the highest daily sales most days
- S0001 had the did outperform S0002 four days in the time frame

- A dataset with more dates would reveal monthly or seasonal sales trends

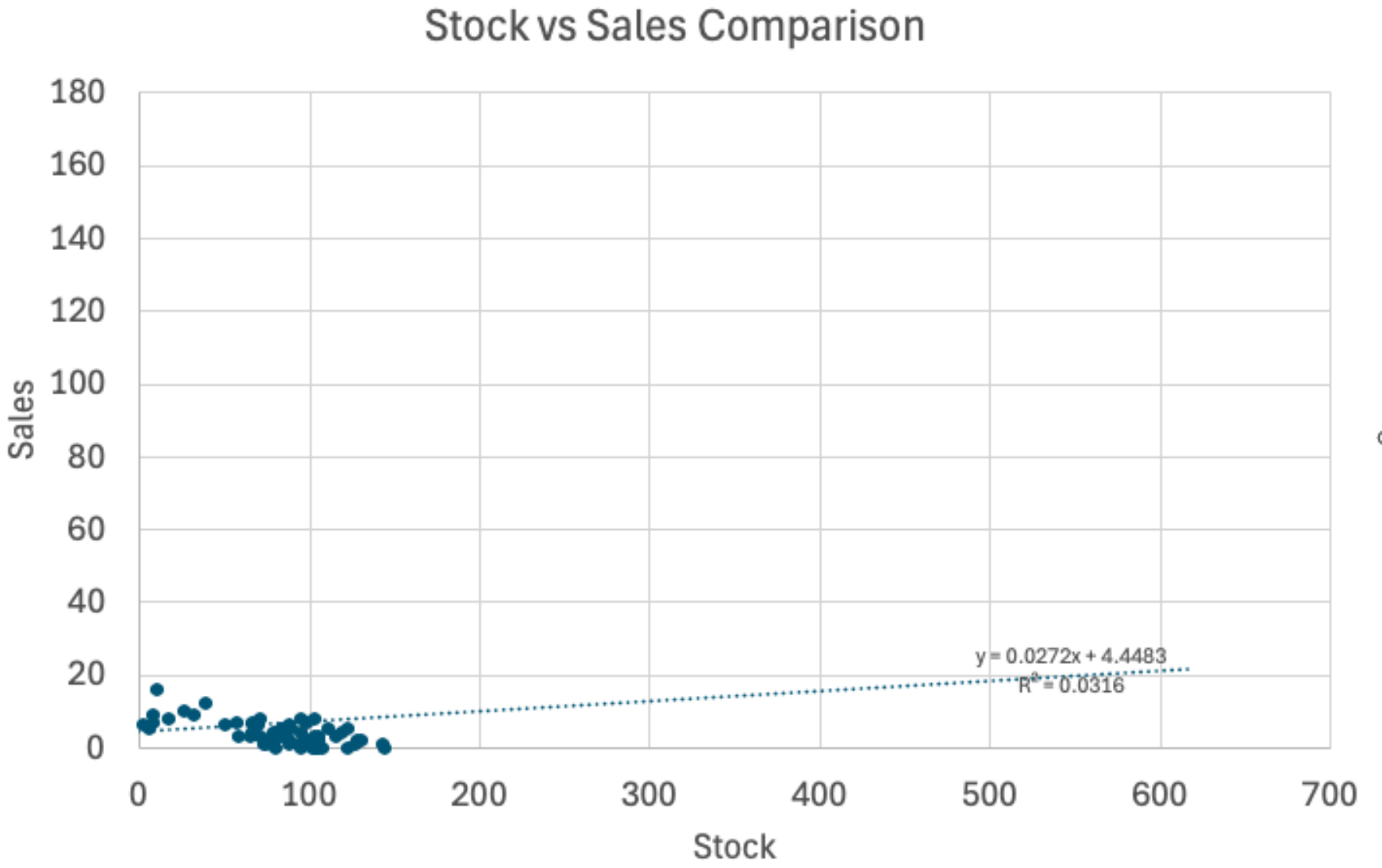
Module 3, Lesson 1: D

- City 024 was the only city with two stores but had the lowest overall sales figures.
- P0438 was one of the top 2 highest performing products in each of the three highest-performing cities
- P0103 was one of the highest performing products in each of the four highest-performing cities



Module 3, Lesson 1: Data Visualization Using Excel

- The scatter plot shows a very weak positive correlation between daily stock levels and sales figures.



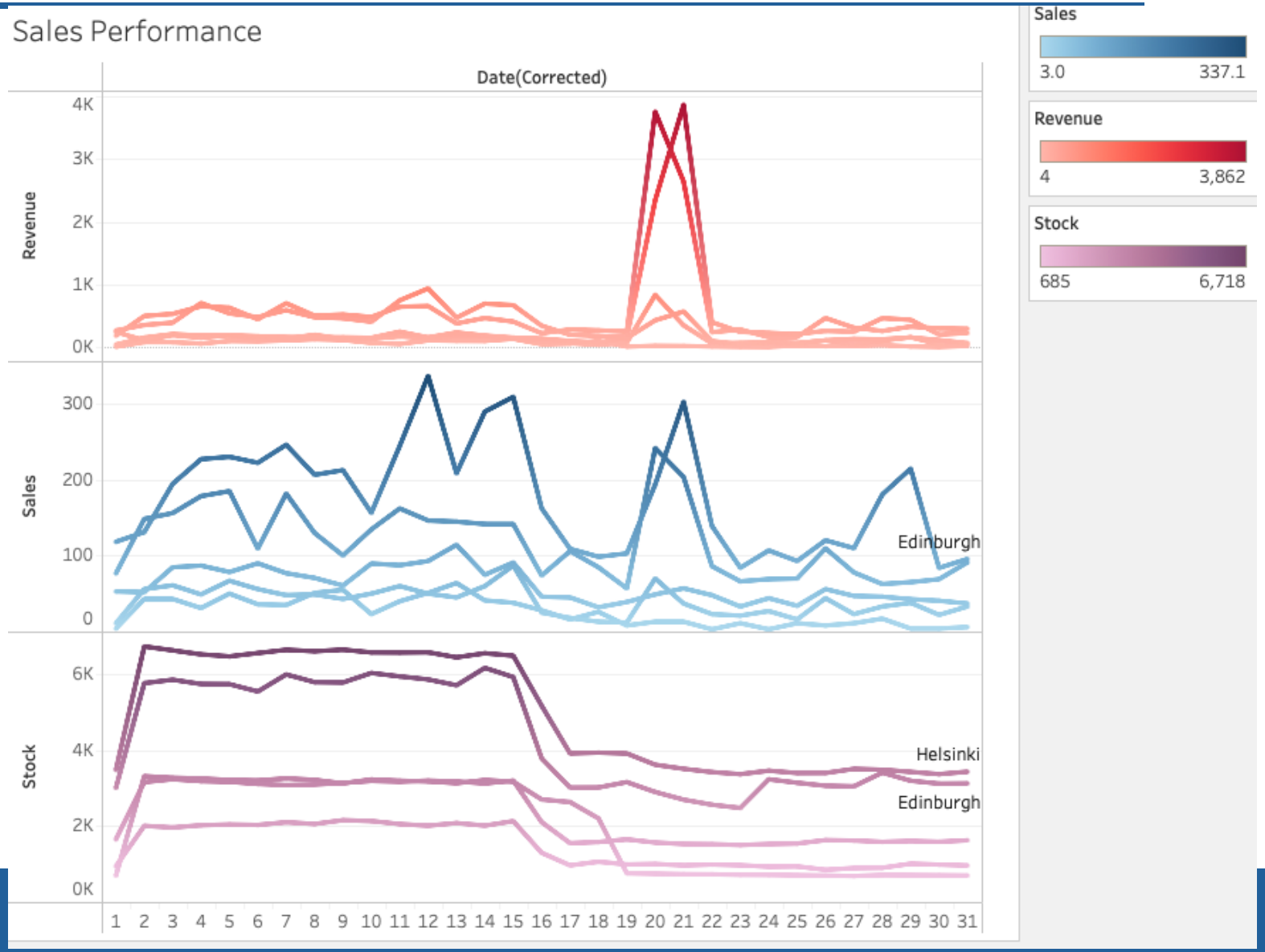
Module 3, Lesson 2: Statistical Analysis

- The R-squared value of 6.49E-06 shows that 0.000649% of variance is explained by the model.
- The Significance F of 0.7186 is the p-value of the model. Since it is greater than 0.05, the model is not statistically significant.

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.002547841					
R Square	6.49149E-06					
Adjusted R Square	-4.35157E-05					
Standard Error	2.805713082					
Observations	19999					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	1.021877363	1.02187736	0.12981123	0.71863165	
Residual	19997	157416.9019	7.8720259			
Total	19998	157417.9238				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-23.18840106	65.8376272	-0.3522059	0.72468749	-152.23559	105.858784
date	0.000554764	0.001539758	0.36029325	0.71863173	-0.0024633	0.00357282

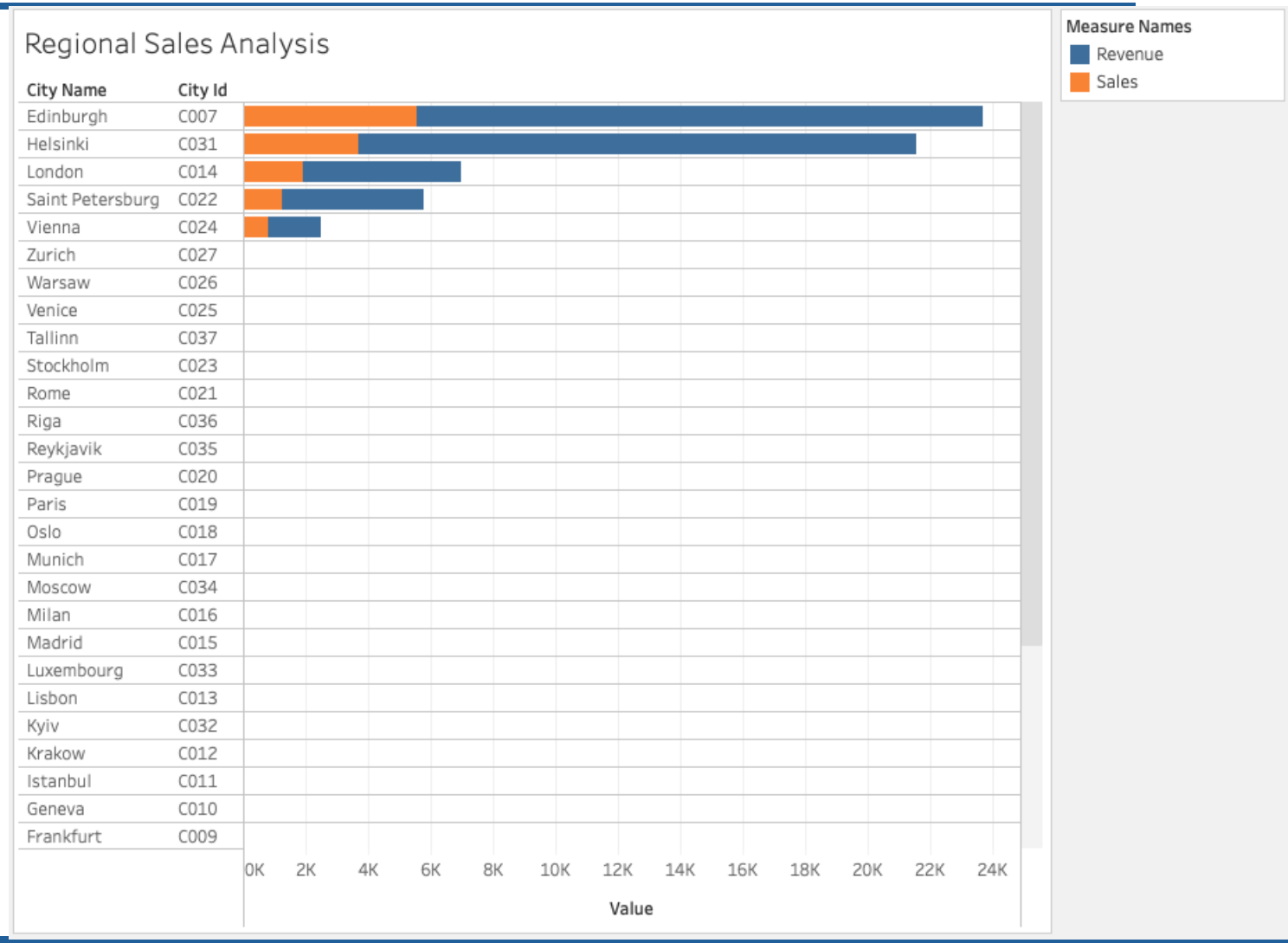
Module 4, Lesson 1: Basic Tableau Visualizations

- The line charts allow for a comparison of sales and revenue performance against stock level based on the day of the month.
- Individual lines represent individual cities.
- Hovering over the lines in Tableau reveal the city, sales growth percentage from the previous day, and applicable totals.



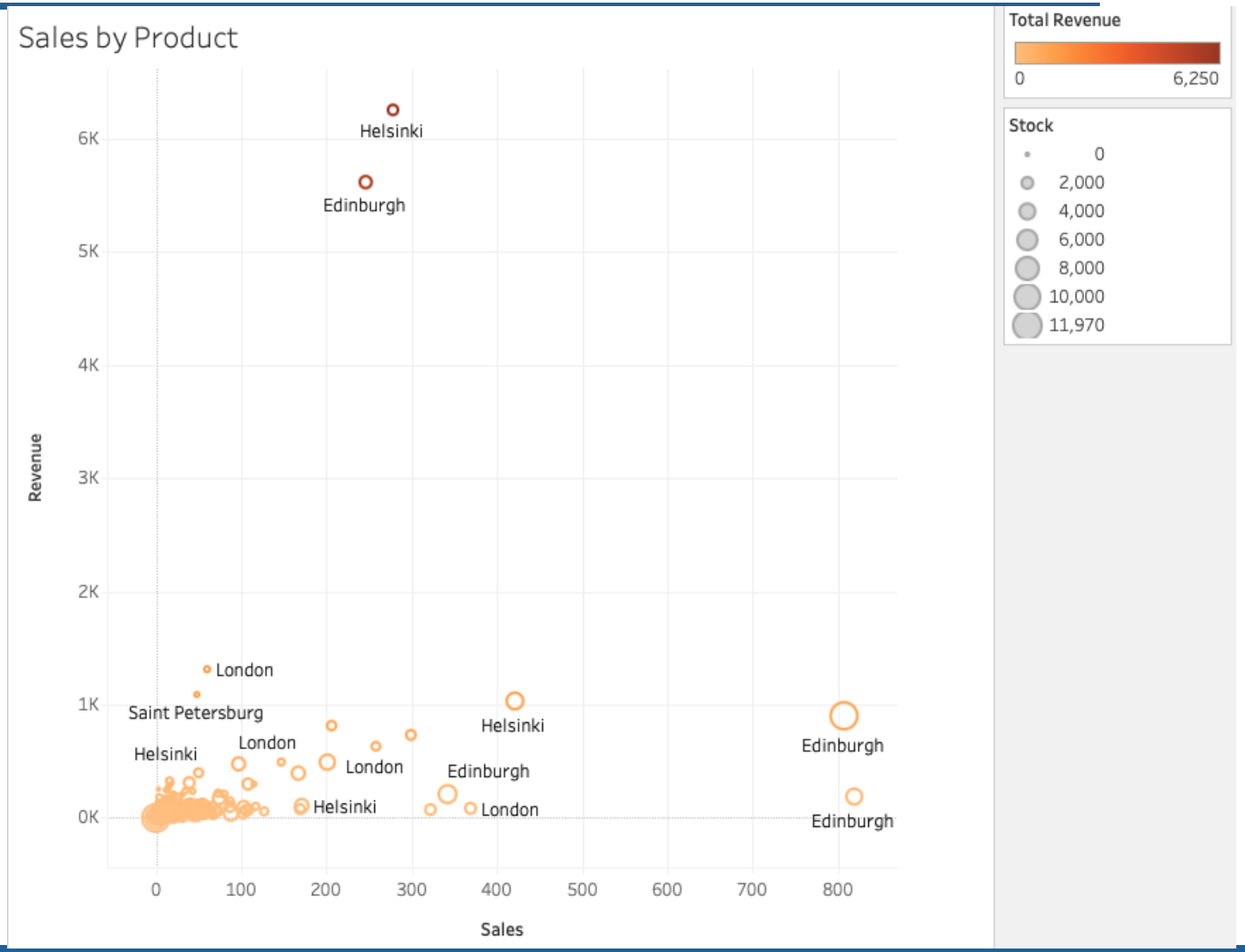
Module 4, Lesson 1: Basic Tableau Visualizations

- The bar chart shows Edinburgh and Helsinki have the highest sales and revenue figures
- Hovering over each bar reveals exact sales and revenue figures, and the change in sales growth as you move through each city



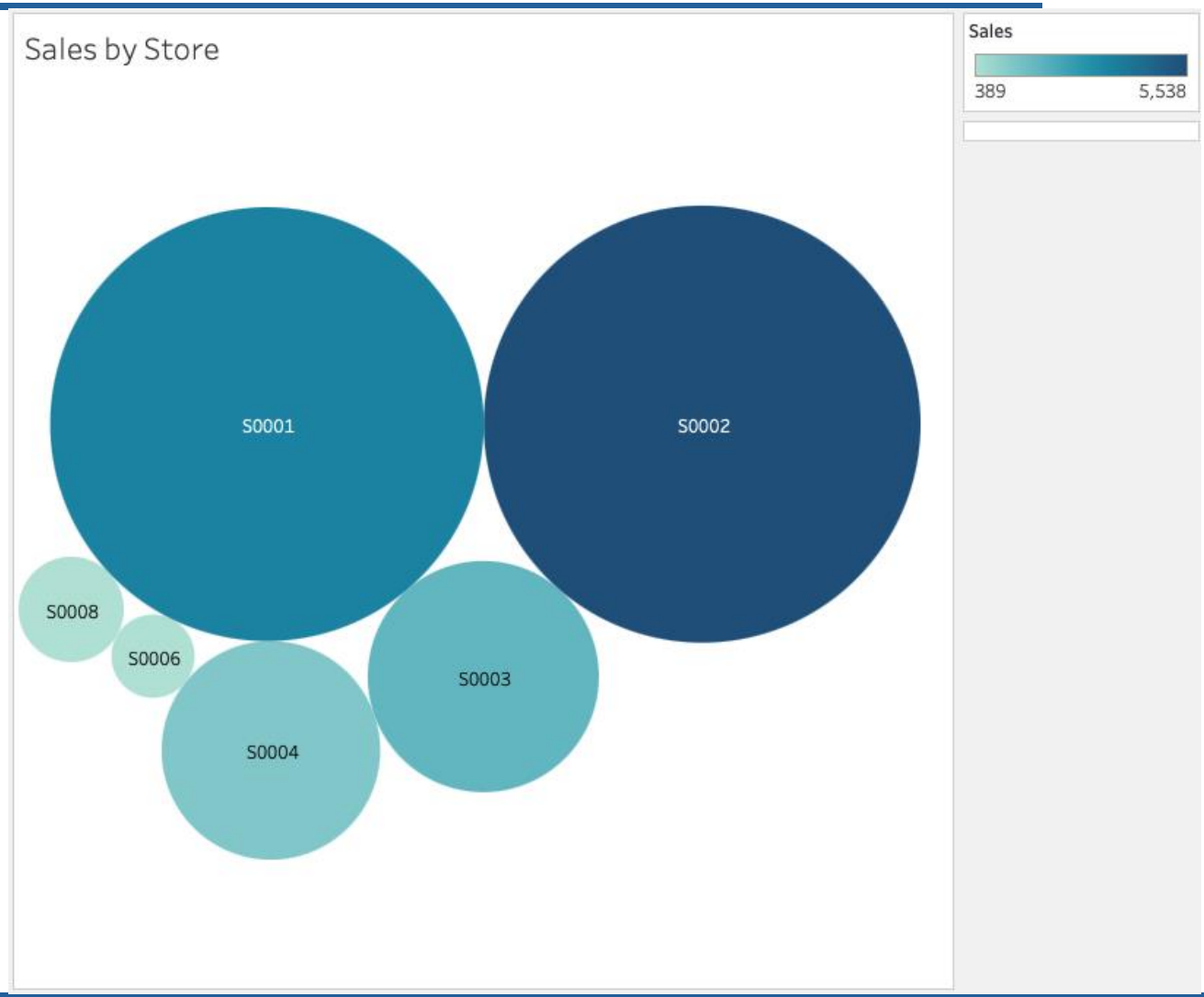
Module 4, Lesson 1: Basic Tableau Visualizations

- The scatter plot shows a comparison of sales and revenue split by product, city, and store.
- Hovering over points on the plot reveals information about the product's performance at each location

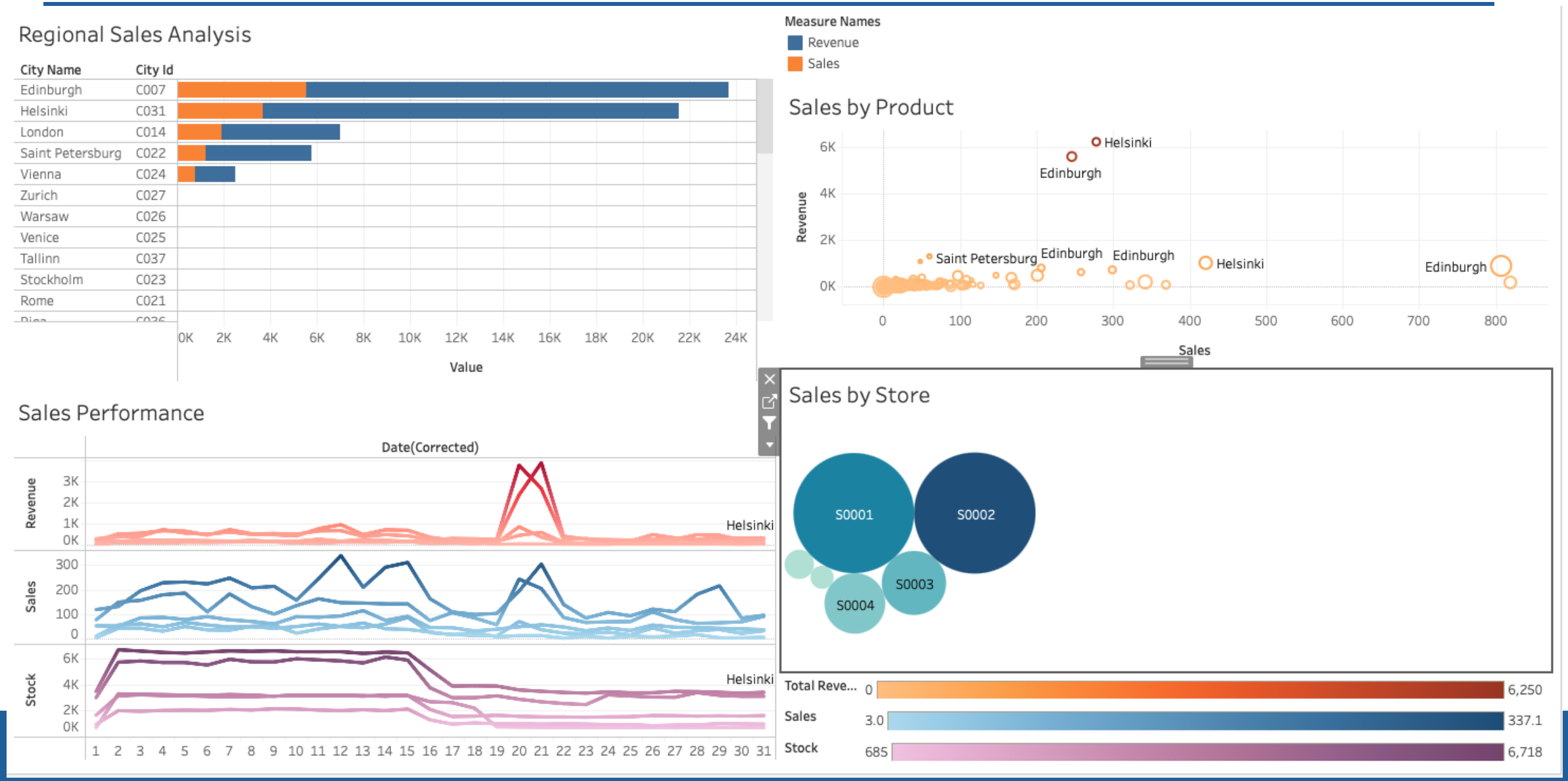


Module 4, Lesson 1: Basic Tableau Visualizations

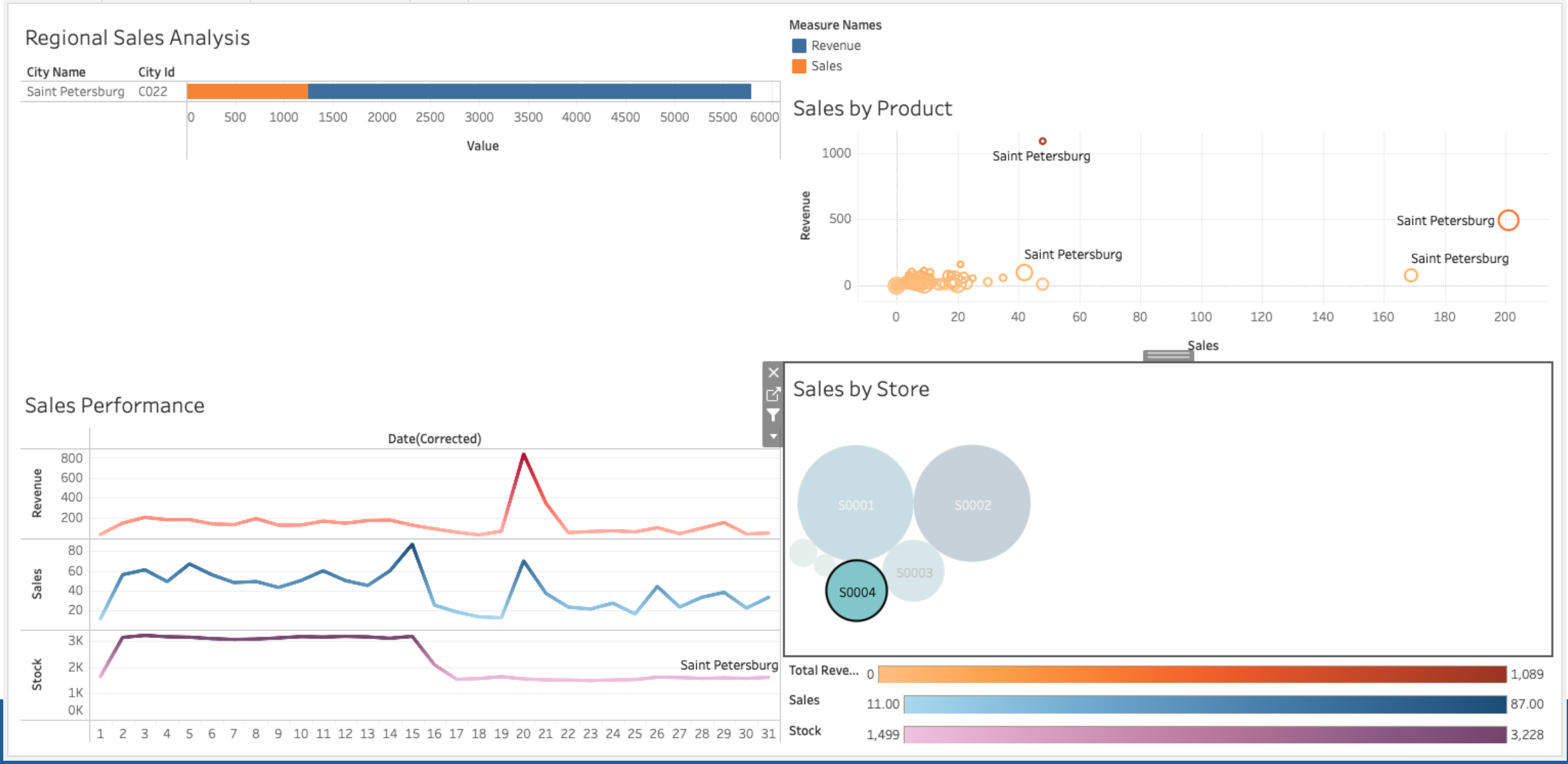
- The bubble chart shows the relative performance each store.
- Hovering over a bubble reveals the store location and performance figures.



Module 4, Lesson 2: Advanced Visualizations Using Tableau



Module 4, Lesson 2: Advanced Visualizations Using Tableau



Module 4, Lesson 2: Advanced Visualizations Using Tableau

- The dashboard allows for viewing multiple visualizations at the same time to allow for comparisons and analysis.
- Filtering by clicking on a bubble allows for drilling down by store to look for trends that would indicate why certain locations are achieving their level of success, and analyse performance trends across the passage of time.

Discussion

Insights and Recommendations

- Edinburgh and Helsinki were shown to have the highest sales and revenue figures
  - These two cities would be good candidates for store expansion
  - Identify strategies these stores are using that other stores could implement
- Vienna was shown to have the lowest sales figures, though two stores were located in Vienna
  - Consider analysis of pricing or offering promotions to help these stores improve their performance
  - If these stores continue to show low performance, closing one of the stores may need to be considered
- Most stores saw a decrease in sales from February to March, 2017
  - Consideration should be made for any promotions or events that may have influenced customer behavior

Insights and Recommendations

- Several products were shown to have no or few sales during the time frame
  - Consider if promotions should be offered to encourage customers to purchase these items, or discontinuing sales of these items to make room for other products
- Several products were shown to have high sales performance
  - Monitor stock levels to make sure these items do not sell out to ensure continued success

Insights and Recommendations

- Pivot charts allow for quickly summarizing data and creating charts.
  - Pivot Charts can be set up with filters to allow for drilling down with data
  - Excel does limit the amount of data that can be used for certain charts, which limited the analysis that could be performed across several stores and dates
- PostgreSQL allows for quickly summarized and sorting larger data sets
  - This analysis can be exported for creating visualizations in other programs
- Tableau allows for the creation of interactive visualizations
  - Tableau Public requires that visualizations be publically available for all to view
  - If using Tableau for company-specific visualizations, consider a paid version of Tableau

Conclusion

Summary

- Edinburgh and Helsinki were shown to have the highest sales and revenue figures
- Vienna's two stores were shown to have the lowest sales performance
- Most stores saw a decrease in sales from February to March, 2017
- Several products were shown to have no or few sales during the time period

Appendix

Appendix

- Tableau public workbook: https://public.tableau.com/views/Upright_Workbook/PerformanceAnalysis?:language=en-US&:sid=&:display_count=n&:origin=viz_share_link

Acknowledgements

- Alan, Berkay. “Retail Sales Data.” *Kaggle*, 12 Apr. 2021, www.kaggle.com/datasets/berkayalan/retail-sales-data/data.