

APPLIED GOVERNED DATA MANAGEMENT FRAMEWORK

Table of Contents

Table of Contents	2
Applied Governed Data Management Framework	3
Project A: Sub-domains	3
Project B: Principles	3
Principle 1: Ensuring Data Integrity through Standardized Metadata	3
Principle 2: Documentation	3
Principle 3: Root Cause Analysis	4
Principle 4: Consistency In Data Quality	4
Project C: Policies	4
Policy 1: Metadata Conformity Policy	5
Policy 2: Documenting the Transformation of Raw Data	5
Policy 3: Data Quality Accuracy Policy	5
Policy 4: Ensuring Data Consistency Policy	6
Project D: Procedures	8
Metadata Conformity Policy Procedures	8
Procedure 1: Metadata Verification Procedure	8
Procedure 2: Metadata Training and Compliance Enhancement Procedure	8
Documenting the Transformation of Raw Data Procedures	9
Procedure 1: Documentation of Data Transformations	9
Procedure 2: Archiving and Linking Transformation Documentation with Data	10
Data Quality Accuracy Policy Procedures	11
Procedure 1: Data Entry Verification Procedure	11
Procedure 2: Data Consistency Checks	12
Ensuring Data Consistency Policy Procedures	13
Procedure 1: Data Entry and Validation	13
Procedure 2: Procedure for Data Cleaning and Normalization	14
Operational Report:	15
Executive report	21
Annex	22
Exercise 1	22
Exercise 2	29
Exercise 3	42

Applied Governed Data Management Framework

Project A: Sub-domains

We have selected the **Metadata Management** and **Data Quality** subdomains to highlight

Project B: Principles

Principle 1: Ensuring Data Integrity through Standardized Metadata

Description:

The integrity of our data hinges on the standardized structure of metadata. Consistent definitions and formats for customer, order, product, regional manager, and region data ensure that our database reflects accurate, reliable, and up-to-date information, directly impacting our customer service and business analytics capabilities.

Principle 2: Documentation

Description:

Our [ETL process](#) is critical metadata and this documentation must be kept with aspects of the project as it has every bit the value that our final database has. Since we chose to make certain concessions in the name of expediency (the argument can also be made for performance), without including our ETL process as metadata, there would be a loss of information that could not be recovered

Principle 3: Root Cause Analysis

Description:

Look closely at the root causes of data quality issues in order to prevent recurrence and improve overall data quality. It is not enough to address the visible symptoms of data quality issues alone, as elucidated. Root cause analysis allows us to identify underlying systemic faults or process gaps that cause problems with data quality. By tackling these fundamental problems, we can implement more workable, long-term fixes that will gradually improve data quality.

Principle 4: Consistency in Data Quality

Description:

In evaluating the dataset, which encapsulates a variety of information ranging from order and shipping details to customer demographics and financial metrics, the principle of Consistency emerges as a pivotal aspect of data quality. This principle is fundamental for ensuring the reliability and uniformity of data throughout the dataset, facilitating accurate analysis and decision-making processes. Key areas to focus on are Date Formats, Categorical Data Uniformity, and Numerical Data Integrity.

Project C: Policies

Policy 1: Metadata Conformity Policy

Purpose: To facilitate the integration of various data streams, ensuring that customer data, sales information, product and order details flow seamlessly into the analytics and reporting tools. This uniformity is essential for maintaining the integrity of the data-driven insights that are critical for strategic business.

Key Elements of the Metadata Conformity Policy:

- **Conformity to ERD:** Adherence to the entity-relationship diagrams ensures that all relationships and constraints between entities are respected, which is essential for data consistency and referential integrity.
- **Standardized Domains:** Domains for each attribute, such as the specific numeric identifiers for 'Region ID' or the text values for 'Product Name', must be strictly followed to ensure data uniformity.
- **Data Type Specifications:** Each column must have the correct data type (e.g., VARCHAR, INT, DATE) as outlined in the metadata tables to prevent type mismatch errors.
- **Column Naming Conventions:** Consistent naming conventions must be used across all tables to avoid confusion and ensure that automated processes can correctly reference data.
- **Value Constraints:** Any constraints on the values that can be stored in each column (e.g., 'Returned' must be 'Yes' or 'No') to maintain data validity must be specified.
- **Format Standards:** For fields with a specific format requirement, such as 'Postal Code' or 'Order Date', a standardized format must be defined and enforced.

Alignment with Principle: The Metadata Conformity Policy is instrumental in preserving data integrity. By mandating uniform metadata, the policy ensures that data entry, storage, and retrieval processes are consistent across the entire database.

Policy 2: Documenting the Transformation of Raw Data

- **Purpose:** To retain the full value of our original data in spite of modifications made when parsing during the ETL Process
- **The policy:**
Every modification of Raw Data made during the ETL process including, but not limited to: change of text-case, truncation, or rearrangement must be documented and included with **Current and Archived Raw Data** as well as **The Database when exported for archiving**

Alignment with Principle: For example, we derived our table keys from IDs in our original data which had greater detail than we needed. We truncated the non-numeric part of these IDs since it gave us the integer data type we needed and did not result in any major conflict. If we needed this data in the future, it could be parsed and inserted into the database. It would be crucial to know the location of this data and how to extract it programmatically while keeping referential integrity in the database

Policy 3: Data Quality Accuracy Policy

- **Purpose:** Ensure the accuracy and consistency of sales transaction data. The goal of this policy is to guarantee that all recorded transactions—including product specifications, client information, and financial data—accurately represent the reality of each sale by carefully validating each one. The policy attempts to improve the dataset's reliability for analysis and decision-making processes by reducing mistakes, inconsistencies, and discrepancies. It provides the fundamental structure for preserving data dependability and integrity, which eventually boosts the organization's operational performance and strategic efficiency."

Key Elements of the Data Quality Assurance Policy

- **Verification Processes:** Implement rigorous verification processes to validate the accuracy of each sales transaction recorded in the dataset.
- **Error Detection and Correction:** Define protocols for locating and fixing inconsistencies or mistakes in the dataset.

- **Quality Assurance Checks:** Regular quality assurance tests are necessary to maintain accuracy standards and pinpoint areas in the dataset that need to be improved.
- **Training and Education:** Provide training and education to personnel involved in data entry and management, emphasizing the importance of accurately recording sales transactions.
- **Documentation and Reporting:** To guarantee accountability and transparency in data management operations and to adhere to the concept of correctness, keep track of paperwork about data accuracy methods and report on verification results.

Alignment with Principle: The emphasis on accurate sales transaction records in the Data Quality Accuracy Policy is consistent with the accuracy principle. The strategy guarantees the correctness of the dataset by carefully examining each item and reducing inaccuracies. This dedication strengthens the data's usefulness as a trustworthy source of business insights and improves its dependability for analysis and decision-making.

Policy 4: Ensuring Data Consistency Policy

Purpose: The purpose of this policy is to establish standards and procedures to ensure data consistency across the dataset. It aims to maintain uniform formats, naming conventions, and categorizations to facilitate accurate analysis, reporting, and data integration. Ensuring consistency is vital for reliable decision-making and gaining meaningful insights from the data.

Key Elements of Ensuring Data Consistency Policy

- **Standard Formats:** Define standard formats for all data types within the dataset. For example, dates should follow the ISO 8601 format (YYYY-MM-DD), numerical data should have a consistent number of decimal places, and text data should adhere to a standardized case (e.g., all caps, title case, or lower case) as appropriate.
- **Categorical Data Standardization:** Establish a controlled vocabulary for all categorical data fields such as Ship Mode, Segment, Category, and Sub-Category. This includes a predefined list of acceptable values for each category to prevent variations that essentially mean the same thing (e.g., "Corporate" should not be entered as "corporate" or "Corporation").
- **Data Entry and Validation Procedures:** Implement data entry guidelines and validation rules to ensure that data is entered correctly and consistently. This may include dropdown menus for categorical fields, date pickers for date fields,

and validation checks for numerical fields to ensure they fall within expected ranges.

- **Data Cleaning and Normalization:** Regularly review and clean the dataset to correct inconsistencies, such as mismatched formats, incorrect categorizations, and typographical errors. This process should include normalization techniques to standardize data entries.
- **Training and Awareness:** Provide training for all staff involved in data entry, management, and analysis to ensure they understand the importance of data consistency and how to achieve it. This should include familiarization with the standard formats, categorical data standardization, and data entry/validation procedures.
- **Periodic Audits:** Conduct periodic audits of the dataset to identify and rectify inconsistencies. This may involve random sampling of data entries to ensure adherence to the established standards and procedures.
- **Feedback and Improvement:** Establish a feedback mechanism for data users to report inconsistencies or suggest improvements to the data consistency policy. Regularly review and update the policy and its procedures based on this feedback to continuously improve data quality.

Alignment with Principle: The policy and its associated procedures for ensuring data consistency directly align with the principle of maintaining uniformity in formats, categorizations, and other types of data across a dataset. By establishing standards for data formats and categorizations, implementing data entry and validation procedures, and conducting regular data cleaning and normalization, the policy aims to prevent and correct inconsistencies. This comprehensive approach ensures that data remains reliable and accurate for analysis and decision-making, directly supporting the principle of data consistency by promoting uniformity and preventing discrepancies within the dataset.

Project D: Procedures

Metadata Conformity Policy Procedures

Procedure 1: Metadata Verification Procedure

Purpose:

To verify the conformity of metadata entries with the established standards in Superstore's databases, ensuring consistency and accuracy in data recording.

Steps for Implementation:

1. Development of a Verification Checklist:
 - Create a checklist based on the ERD and metadata standards for data stewards to use during verification.
2. Routine Data Audits:
 - Schedule and perform regular audits on the database entries to ensure conformity to the metadata standards.
3. Discrepancy Resolution:
 - Outline a process for immediate correction of any discrepancies found during audits.
4. Audit Reporting:
 - Compile audit findings and report them to the Data Governance Committee for oversight.

Alignment with Policy:

This procedure directly supports the Metadata Conformity Policy by establishing a systematic approach to verify and correct data entries, ensuring they meet Superstore's standardized metadata requirements.

Procedure 2: Metadata Training and Compliance Enhancement Procedure

Purpose:

To enhance the understanding and application of metadata standards among all relevant Superstore personnel, promoting compliance through education and system design.

Steps for Implementation:

1. Performance Review and Incentivization:
 - Include metadata conformity in performance reviews for data handlers.
 - Introduce incentives for consistent compliance with metadata standards.
2. Feedback Integration:
 - Implement a feedback mechanism for continuous improvement of the training programs and data entry systems.

Alignment with Policy:

This procedure aligns with the Metadata Conformity Policy by fostering an environment that supports and encourages adherence to metadata standards through training and system design.

Documenting the Transformation of Raw Data Procedure

Procedure 1: Documentation of Data Transformations

Purpose:

To create a comprehensive and accessible record of all transformations applied to raw data during the ETL process, ensuring that details of all modifications, such as text-case changes, truncations, and rearrangements, are fully documented.

Steps for Implementation:

1. Define Documentation Standards:
 - a. Establish a clear standard for documenting data transformations, including the level of detail required and the format in which this documentation should be maintained.
2. Create a Transformation Log:
 - a. Implement a system for logging every transformation applied to the data. This log should include the nature of the transformation, the reason for its application, and the original and transformed values, where applicable.
3. Automate Documentation Where Possible:
 - a. Use ETL tools that support automatic logging of data transformations or develop custom scripts to capture this information during the ETL process.
4. Review and Update Documentation Regularly:
 - a. Periodically review the transformation log to ensure completeness and accuracy. Update the documentation to reflect any new transformations or changes to the ETL process.
5. Ensure Accessibility of Documentation:

- a. Store the transformation documentation in a central, accessible location, ensuring that it is available to all relevant stakeholders and can be easily referenced or queried.

Alignment with Policy: This procedure directly supports the policy's objective to retain the full value of the original data despite modifications made during the ETL process. By meticulously documenting every transformation, it ensures that the lineage of the data is preserved, allowing for the future retrieval and understanding of data modifications. This is crucial for maintaining data integrity and ensuring that any transformations do not obscure the original data's meaning or utility.

Procedure 2: Archiving and Linking Transformation Documentation with Data

Purpose:

To ensure that documentation of data transformations is effectively linked with both the current and archived versions of the raw data, as well as with the database, facilitating traceability and referential integrity.

Steps for Implementation:

1. Integrate Documentation with Data Archives:
 - a. When archiving raw data, include the transformation documentation as part of the archive package. This ensures that the context of the data transformations is preserved alongside the data itself.
2. Link Documentation to Database Exports:
 - a. For each database export intended for archiving, ensure that an export of the transformation documentation is included, clearly linking the documentation to the specific dataset export.
3. Maintain Referential Integrity:
 - a. Develop procedures to maintain referential integrity when transformations involve key fields, such as truncating non-numeric parts of IDs. Ensure that the documentation details how these transformations can be reversed or referenced to maintain data integrity.
4. Provide Guidelines for Data Reconstitution:
 - a. Include in the documentation guidelines on how the original data can be reconstituted from the transformed data if necessary. This includes scripts or instructions for parsing and inserting data back into the database while preserving referential integrity.
5. Regular Audits and Updates:

- a. Conduct regular audits of the archiving and linking processes to ensure that the transformation documentation remains accurate and effectively linked to the data it describes. Update procedures as needed to adapt to changes in the ETL process or data structure.

Alignment with Policy: Aligning with the policy, this procedure ensures that detailed records of data transformations are not only documented but also effectively archived and made retrievable. This enables the organization to maintain the utility of transformed data, ensuring that modifications made during the ETL process do not permanently obscure the insights or value of the original raw data. The archival and retrieval protocols support the policy's aim of preserving the full value of the original data, ensuring that the transformations are transparent and reversible, thereby upholding data integrity and referential integrity within the database.

Data Quality Accuracy Policy Procedures

Procedure 1: Data Entry Verification Procedure

Purpose:

To make sure that sales transaction records are accurate, each entry will be methodically checked against predetermined criteria using the information that has been given..

Steps for Implementation:

1. Define Verification Criteria: Establish criteria for verifying sales transaction details, including product information, customer details, dates, and financial data, based on the dataset columns.
2. Conduct Verification Checks: Utilize automated validation tools and manual reviews to verify each transaction against the defined criteria, ensuring that all fields are accurately recorded.
3. Identify Discrepancies: Compare each entry against the verification criteria to identify any discrepancies or errors, such as missing or incorrect information.
4. Correct Errors: Promptly address and correct any inaccuracies found in the dataset, ensuring that the corrected information is accurately reflected to maintain data integrity.
5. Document Verification Results: Maintain detailed documentation of verification results, including discrepancies found and actions taken to rectify them, linking them to specific entries in the dataset for traceability and audit purposes.

Alignment with Policy: This procedure aligns with the Data Quality Accuracy Policy by prioritizing the accuracy of sales transaction records, ensuring that each entry accurately reflects the details of the corresponding sale, as demonstrated by the dataset provided.

Procedure 2: Data Consistency Checks

Purpose:

To maintain consistency in data formatting and standards across all sales transaction records, ensuring uniformity and reliability in the dataset, based on the dataset provided.

Steps for Implementation:

1. **Standardize Data Formats:** Establish standardized formats for key data fields, such as dates, currency, and product codes, based on the specific formatting observed in the dataset columns.
2. **Validate Data Consistency:** Implement checks to validate that data entries adhere to predefined formatting rules and standards, ensuring consistency across all records.
3. **Identify Formatting Inconsistencies:** Utilize automated tools to flag any entries that deviate from the standardized formats or fail to meet predefined standards, +highlighting them for further review.
4. **Address Formatting Issues:** Take corrective action to rectify formatting inconsistencies by updating entries to comply with predefined standards or reformatting data fields as necessary, ensuring data uniformity.
5. **Monitor Consistency Over Time:** Continuously monitor data consistency and adjust formatting rules as needed to maintain uniformity and accuracy in the dataset, considering changes in data patterns over time and ensuring ongoing compliance with established standards

Alignment with Policy:

This procedure aligns with the Data Quality Accuracy Policy by promoting consistency and uniformity in data formatting and standards, thereby enhancing the dataset's reliability and trustworthiness, as evidenced by the dataset provided.

Ensuring Data Consistency Policy Procedures

Procedure 1: Data Entry and Validation

Purpose:

The purpose of this procedure is to ensure accurate and consistent data entry across the dataset. It aims to prevent common data entry errors and enforce standard formats and categorizations, thereby enhancing the overall data quality and reliability.

Steps for Implementation:

1. Define Data Standards:
 - a. Establish clear guidelines for each type of data in the dataset, including formats for dates (YYYY-MM-DD), numerical values (e.g., decimal places), and text (e.g., case sensitivity).
 - b. Create a controlled vocabulary for categorical fields to ensure uniformity in entries.
2. Configure Data Entry Interfaces:
 - a. Set up data entry forms or software with predefined fields that enforce the data standards. Use dropdown menus for categorical fields, date pickers for dates, and validation rules for numerical fields.
3. Implement Validation Rules:
 - a. Incorporate validation rules within the data entry system to automatically check the data against the established standards. This includes format checks, range limits for numerical fields, and mandatory field completion.
4. Error Reporting and Correction Mechanisms:
 - a. Design the system to provide immediate feedback when data entered does not comply with the validation rules, explaining the error and how to correct it.
5. Training and Support:
 - a. Provide comprehensive training for all personnel involved in data entry, focusing on the importance of data quality, understanding the data standards, and using the data entry system effectively.
 - b. Offer ongoing support to address any questions or issues that arise during data entry.
6. Regular Reviews and Updates:
 - a. Periodically review the data entry guidelines and system functionality to ensure they remain effective and adjust them as necessary based on feedback and evolving data needs.

Alignment with Policy: The Data Entry and Validation procedure aligns with the policy by establishing a structured approach to ensure that data entered into the dataset

meets predefined standards of consistency. By configuring data entry interfaces with predefined fields, implementing validation rules, and providing immediate feedback on errors, this procedure directly supports the policy's goal of maintaining uniform data formats and categorizations. Training and support for data entry personnel further ensure adherence to these standards, thereby upholding the policy's emphasis on data consistency from the point of entry.

Procedure 2: Procedure for Data Cleaning and Normalization

Purpose:

This procedure aims to maintain and improve the dataset's quality by regularly identifying and correcting inconsistencies, errors, and duplicates. It ensures that the data remains accurate, consistent, and usable for analysis and decision-making.

Steps for Implementation:

1. Schedule Regular Data Cleaning Cycles:
 - a. Establish a routine schedule for data cleaning activities, such as monthly or quarterly, depending on the data volume and update frequency.
2. Identify Common Data Inconsistencies:
 - a. Use historical data cleaning logs and user feedback to identify common issues within the dataset, such as misformatted dates, inconsistent categorizations, and typographical errors.
3. Utilize Data Cleaning Tools:
 - a. Implement software tools or scripts that can automate the detection and correction of identified inconsistencies. These tools should be capable of handling tasks like format standardization, duplicate removal, and outlier detection.
4. Manual Review and Correction:
 - a. For complex issues or those requiring contextual understanding, conduct a manual review and correction process. Assign this task to team members with in-depth knowledge of the data and its application.
5. Document Changes and Rationale:
 - a. Keep a detailed log of all corrections made during the data cleaning process, including the nature of the issue, how it was resolved, and the date of correction. This documentation is essential for transparency and future reference.
6. Feedback Loop for Continuous Improvement:
 - a. Create a mechanism for users to easily report data quality issues. Regularly review these reports, and incorporate the insights gained into the data cleaning process to prevent similar issues in the future.
7. Update Data Standards as Necessary:

- a. Based on the findings from data cleaning cycles and user feedback, periodically update the data standards and cleaning methodologies to address new types of inconsistencies or changes in data use cases.

Alignment with Policy: The Data Cleaning and Normalization procedure supports the policy by addressing inconsistencies and errors in the dataset post-entry. Through scheduled cleaning cycles, the use of data cleaning tools, and manual review processes, this procedure corrects deviations from established data standards. It ensures that all data, regardless of its entry point, conforms to the uniform formats and categorizations outlined in the policy. Documentation of changes and a feedback loop for continuous improvement ensure that the procedure evolves to meet the policy's objectives, reinforcing the overarching goal of data consistency.

Operational Report:

Superstore Sales Operational Report over the ~4 Year Period (3 years and 362 days)
03/01/2018 – 30/12/2021

Region	State	Category	Total Quantity	Average Discount (%)	Gross Sales (\$)	Profit (\$)
Central	Illinois	Furniture	432	0.463865546	27907.49	-8705.7226
		Office Supplies	1037	0.411111111	19118.124	-8147.9047
		Technology	302	0.205952381	31983.673	4822.5592
	Illinois Total	1771	0.360309679	79009.287	-12031.0681	
	Indiana	Furniture	83	0	11496.71	2181.2753
		Office Supplies	374	0	15654.47	5162.2849
		Technology	106	0	26323.25	11000.8773
	Indiana Total	563	0	53474.43	18344.4375	
	Iowa	Furniture	24	0	2642.31	520.0385
		Office Supplies	66	0	723.16	317.9622
		Technology	13	0	1154.3	318.3682
	Iowa Total	103	0	4519.77	1156.3689	
	Kansas	Furniture	8	0	111.12	36.9696
		Office Supplies	47	0	1954.15	624.4873
		Technology	19	0	849.04	174.9866
	Kansas Total	74	0	2914.31	836.4435	
	Michigan	Furniture	181	0	22260.26	4652.4324
		Office Supplies	589	0.010457516	37521.349	14910.2044
		Technology	148	0.004444444	16209.975	4778.3119
	Michigan Total	918	0.00496732	75991.584	24340.9487	
	Minnesota	Furniture	52	0	7611.35	2023.8871
		Office Supplies	241	0	19406.54	7780.4995
		Technology	38	0	2845.26	1018.8008
	Minnesota Total	331	0	29863.15	10823.1874	
	Missouri	Furniture	29	0	2390.57	537.8761
		Office Supplies	158	0	12080.82	2747.42
		Technology	54	0	7086.52	3014.7268
	Missouri Total	241	0	21557.91	6300.0229	
	Nebraska	Furniture	21	0	1944.7	518.4364
		Office Supplies	86	0	2216.85	548.4992
		Technology	26	0	3285.74	961.515
	Nebraska Total	133	0	7447.29	2028.4506	
	North Dakota	Office Supplies	30	0	919.91	230.1497
	North Dakota Total	30	0	919.91	230.1497	
	Oklahoma	Furniture	57	0	8284.1	2153.8622
		Office Supplies	119	0	4489.6	1113.7811
		Technology	67	0	6368.45	1580.9003
	Oklahoma Total	243	0	20981.97	4848.5436	
	South Dakota	Furniture	5	0	324.9	67.1898
		Office Supplies	31	0	597.72	193.7419
		Technology	6	0	392.94	133.8966
	South Dakota Total	42	0	1315.56	394.8283	
	Texas	Furniture	758	0.4212	60553.0238	-10410.4531
		Office Supplies	2221	0.394871795	43896.384	-18415.9623
		Technology	659	0.21452514	65104.224	3291.429
	Texas Total	3638	0.343532312	169553.6318	-25534.9864	
	Wisconsin	Furniture	144	0	17256.61	3838.9545
		Office Supplies	211	0	6037.12	1955.5522
		Technology	104	0	8798.16	2597.0697
	Wisconsin Total	459	0	32091.89	8391.5764	
Central Grand Total			8516	0.059067443	498349.372	39898.7533

East	Connecticut	Furniture	46	0.046153846	5174.987	1226.2805
		Office Supplies	202	0	5351.78	1473.6601
		Technology	29	0	2791.03	780.9336
	Connecticut Total		277	0.015384615	13317.797	3480.8742
	Delaware	Furniture	62	0.035294118	4547.359	828.3152
		Office Supplies	233	0	7926.59	2769.0582
		Technology	59	0	14562.22	6239.0508
	Delaware Total		354	0.011764706	27036.169	9836.4242
	District of Columbia	Furniture	8	0	1346.58	350.0835
		Office Supplies	24	0	138.52	60.9434
		Technology	8	0	1379.92	648.5624
	District of Columbia Total		40	0	2865.02	1059.5893
	Maine	Furniture	2	0	109.48	33.9388
		Office Supplies	19	0	399.8	169.5146
		Technology	14	0	761.25	251.0328
	Maine Total		35	0	1270.53	454.4862
	Maryland	Furniture	104	0.021428571	9149.253	1905.8274
		Office Supplies	267	0	10345.48	3781.8888
		Technology	44	0	4166.04	1322.8776
	Maryland Total		415	0.007142857	23660.773	7010.5938
	Massachusetts	Furniture	103	0.065625	10873.224	1074.4871
		Office Supplies	306	0	11828.35	3863.5205
		Technology	69	0	5726.63	1755.7501
	Massachusetts Total		478	0.021875	28428.204	6693.7577
	New Hampshire	Furniture	27	0.05	1886.474	153.937
		Office Supplies	80	0	1769.25	649.8585
		Technology	20	0	3636.8	902.7073
	New Hampshire Total		127	0.016666667	7292.524	1706.5028
	New Jersey	Furniture	85	0.023076923	6307.042	932.3293
		Office Supplies	268	0	13792.37	4633.6443
		Technology	86	0	14501.16	4170.198
	New Jersey Total		439	0.007692308	34600.572	9736.1716
	New York	Furniture	860	0.109090909	92661.569	5572.5977
		Office Supplies	2510	0.051212121	89338.534	25761.8278
		Technology	759	0.006763285	127453.53	42172.6997
	New York Total		4129	0.055688772	309453.633	73507.1252
	Ohio	Furniture	328	0.288764045	23141.241	-4366.8568
		Office Supplies	987	0.334469697	17800.412	-54.4197
		Technology	384	0.339215686	35675.992	-12649.9401
	Ohio Total		1699	0.320816476	76617.645	-17071.2166
	Pennsylvania	Furniture	462	0.280672269	38630.939	-7237.4739
		Office Supplies	1158	0.332075472	34136.957	-5009.2946
		Technology	406	0.341525424	42143.341	-3199.6132
	Pennsylvania Total		2026	0.318091055	114911.237	-15446.3817
	Rhode Island	Furniture	53	0.075	5918.756	913.377
		Office Supplies	106	0	6207.59	1761.7968
		Technology	35	0	10474.41	4598.0123
	Rhode Island Total		194	0.025	22600.756	7273.1861
	West Virginia	Furniture	3	0.3	673.344	-76.9536
		Office Supplies	15	0	536.48	262.8752
	West Virginia Total		18	0.15	1209.824	185.9216
	East Grand Total		10231	0.073086343	663264.684	88427.0344

South	Alabama	Furniture	54	0	6332.48	1231.3882
		Office Supplies	149	0	4209.08	1257.6342
		Technology	53	0	8969.08	3297.8029
	Alabama Total		256	0	19510.64	5786.8253
	Arkansas	Furniture	43	0	3187.55	781.4552
		Office Supplies	132	0	4382.39	1879.8117
		Technology	62	0	3925.25	1261.4384
	Arkansas Total		237	0	11495.19	3922.7053
	Florida	Furniture	292	0.23313253	22556.87	-2252.7049
		Office Supplies	805	0.344144144	19363.801	-1659.7791
		Technology	258	0.226470588	46956.212	529.7702
	Florida Total		1355	0.267915754	88876.883	-3382.7138
	Georgia	Furniture	110	0	7726.1	1751.768
		Office Supplies	426	0	26511.65	9782.2192
		Technology	155	0	14058.55	4399.6597
	Georgia Total		691	0	48296.3	15933.6469
	Kentucky	Furniture	106	0	12126.84	3210.9932
		Office Supplies	320	0	11894.27	3832.067
		Technology	97	0	12570.64	4156.6364
	Kentucky Total		523	0	36591.75	11199.6966
	Louisiana	Furniture	36	0	2963.03	685.9946
		Office Supplies	79	0	3423.16	495.0925
		Technology	41	0	2830.84	1015.0152
	Louisiana Total		156	0	9217.03	2196.1023
	Mississippi	Furniture	43	0	4317.85	944.8196
		Office Supplies	135	0	3589.29	1221.4665
		Technology	38	0	2822.33	986.4024
	Mississippi Total		216	0	10729.47	3152.6885
	North Carolina	Furniture	174	0.238095238	15155.484	-3486.4633
		Office Supplies	619	0.311842105	14309.609	-417.1361
		Technology	174	0.223529412	26083.119	-3583.304
	North Carolina Total		967	0.257822252	55548.212	-7486.9034
	South Carolina	Furniture	22	0	3078.25	612.8439
		Office Supplies	117	0	3316.38	465.5674
		Technology	20	0	1591.62	453.5875
	South Carolina Total		159	0	7986.25	1531.9988
	Tennessee	Furniture	165	0.235555556	13506.732	-2208.6291
		Office Supplies	413	0.313636364	12108.755	-3136.7774
		Technology	74	0.23	4807.283	66.0337
	Tennessee Total		652	0.25973064	30422.77	-5279.3728
	Virginia	Furniture	233	0	25321.95	5204.3265
		Office Supplies	492	0	20871.54	5856.9156
		Technology	140	0	24115.6	7399.7732
	Virginia Total		865	0	70309.09	18461.0153
	South Grand Total		6077	0.071406241	388983.585	46035.689

South Grand Total			6077	0.071406241	388983.585	46035.689	
West	Arizona	Furniture	177	0.293333333	12882.627	-2825.517	
		Office Supplies	488	0.342276423	9651.393	-726.3014	
		Technology	173	0.23125	11750.885	112.5012	
	Arizona Total		838	0.288953252	34284.905	-3439.3172	
	California	Furniture	1664	0.110091743	153836.0715	8495.438	
		Office Supplies	4370	0.045778938	137548.35	36733.3991	
		Technology	1394	0.117647059	159183.17	29440.3665	
	California Total		7428	0.09117258	450567.5915	74669.2036	
	Colorado	Furniture	191	0.316326531	12467.981	-2669.8166	
		Office Supplies	339	0.343617021	7850.926	-348.2794	
		Technology	149	0.242857143	10966.329	-3471.5845	
	Colorado Total		679	0.300933565	31285.236	-6489.6805	
	Idaho	Furniture	18	0.033333333	2595.482	533.9665	
		Office Supplies	27	0.111111111	840.506	149.9901	
		Technology	12	0.15	837.498	92.0973	
	Idaho Total		57	0.098148148	4273.486	776.0539	
	Montana	Furniture	7	0	63.98	21.7532	
		Office Supplies	29	0.044444444	1856.342	286.3301	
		Technology	18	0.1	3662.934	1523.0354	
	Montana Total		54	0.048148148	5583.256	1831.1187	
	Nevada	Furniture	37	0.022222222	4635.172	524.5705	
		Office Supplies	93	0.060869565	6870.724	2254.3518	
		Technology	33	0.133333333	5137.006	512.8456	
	Nevada Total		163	0.072141707	16642.902	3291.7679	
	New Mexico	Furniture	19	0.1	1701.412	251.5917	
		Office Supplies	91	0.041666667	1384.182	568.2416	
		Technology	41	0.088888889	1697.928	337.2828	
	New Mexico total		151	0.076851852	4783.522	1157.1161	
	Oregon	Furniture	69	0.342857143	6338.13	-1487.5769	
		Office Supplies	299	0.290277778	5122.416	154.4406	
		Technology	107	0.237037037	5821.556	126.3564	
	Oregon Total		475	0.290057319	17282.102	-1206.7799	
	Utah	Furniture	29	0	4822.35	631.7557	
		Office Supplies	113	0.054545455	3442.782	1141.365	
		Technology	45	0.114285714	2309.904	581.7351	
	Utah Total		187	0.056277056	10575.036	2354.8558	
	Washington	Furniture	417	0.061818182	47503.892	7012.0558	
		Office Supplies	1041	0.058273381	38557.562	10947.9131	
		Technology	351	0.082828283	50528.718	15016.6462	
	Washington Total		1809	0.067639949	136590.172	32976.6151	
	Wyoming	Furniture	4	0.2	1603.136	100.196	
	Wyoming Total		4	0.2	1603.136	100.196	
West Grand Total			11845	0.144574871	713471.3445	106021.1495	
			United States Grand Total				
			Total Quantity	Average Discounts	Gross Sales	Total Profit	
			36669	0.087033725	2264068.986	280382.6262	

Notes:

1. All Gross Sales and profit figures are in USD.
2. **Breakdown by regions –**
 - a. **Central:**
Total Quantity sold – 8516 products.
Gross Sales - \$ 498349.372
Total Profit - \$ 39898.7533
 - b. **East:**
Total Quantity sold – 10231 products.
Gross Sales - \$ 663264.684
Total Profit - \$ 88427.0344
 - c. **South:**
Total Quantity sold – 6077 products.
Gross Sales - \$ 388983.585
Total Profit - \$ 46035.689
 - d. **West:**
Total Quantity sold – 11845 products.
Gross Sales - \$ 713471.3445
Total Profit - \$ 106021.1495
3. **Grand Total for all regions (USA):**
Total Quantity sold – 36669 products.
Gross Sales - \$ 2264068.986
Total Profit - \$ 280382.6262
4. The operational report is adaptable to suit the specific needs of stakeholders.
Depending on their requirements, additional sub-categories or city-level data can be incorporated to enhance the level of detail and comprehensiveness within the report.
5. The report offers comprehensive insights into the superstore's sales performance spanning a four-year period (03/01/2018 – 30/12/2021). It delves into regional and city-specific sales data, shedding light on top-selling product categories.
Furthermore, the report meticulously outlines key financial metrics including profits, sales figures, discounts, and product quantities sold across various regions and states.

Note: Please refer Annex Lab Exercise 3 for detailed information and snapshots of the script used to create the operational report.

Executive report

Superstore Yearly Executive Report Gross Profit (\$)

Version 2.0

Author: Group 4

Region	2018 Profit (\$)	2019 Profit (\$)	2019 Profit vs 2018 Profit (\$)	2020 Profit (\$)	2020 Profit vs 2019 Profit (\$)	2021 Profit (\$)	2021 Profit vs 2020 Profit (\$)
South	11745.8151	7981.6786	-3764.1365	17712.8445	9731.1659	8595.3508	-9117.4937
West	19728.449	20184.8163	456.3673	23035.8991	2851.0828	43071.9851	20036.086
Central	584.693	11691.6566	11106.9636	19734.3279	8042.6713	8118.2255	-11616.1024
East	16985.4692	19857.6112	2872.142	18858.5055	-999.1057	32725.4485	13866.943
Overall Growth (\$)			10671.3364		19625.8143		13169.4329

Notes –

1. Profit Variance between consecutive years:

a. 2019 VS 2018 –

Total Growth in Profit (\$) 10671.33.

b. 2019 VS 2018 –

Total Growth in Profit (\$) 19625.81.

c. 2019 VS 2018 –

Total Growth in Profit (\$) 13169.43.

2. This executive report provides the information about the overall growth in the profit (in \$) for the superstore over the 4-year period (03/01/2018 – 30/12/2021).

Note: Please refer Annex Lab Exercise 3 for detailed information and snapshots of the script used to create the executive report.

Annex

Lab Exercise 1

1. Analysis

1.1. Data Analysis

1.1.1. Completeness This section assesses the presence of missing values in the dataset. Below is a summary of the findings:

1. Postal Code:

- Number of Missing Values: 11
- Impact on Analysis: The 11 missing values in the "Postal Code" column may impact analyses that rely on geographic information. It's essential to understand the nature of the missing values and determine whether imputation or removal is appropriate.

2. Overall Assessment:

- The majority of the dataset is complete, with no missing values in most columns.
- The impact of missing values on analysis depends on the specific goals and the significance of the "Postal Code" column.

3. Recommended Actions:

- Investigate the nature and potential reasons for missing "Postal Code" values.
- Decide whether to impute missing values, if possible, or consider removing the affected rows.
- Document any decisions made regarding missing values for transparency in subsequent analyses. Completeness is a critical aspect of data quality, and addressing missing values appropriately is essential for reliable and accurate analyses. The specific actions taken will depend on the nature of the dataset and the goals of the analysis.

1.1.2. Duplicates The analysis of the dataset reveals that there are no duplicate records present. The absence of duplicate records is a positive indication for data quality and integrity. It suggests that each record in the dataset is unique, reducing the risk of redundancy and ensuring accuracy in subsequent analyses.

Summary:

- Number of Duplicate Records: 0
- Dataset Integrity: The dataset exhibits a high level of uniqueness, contributing to the reliability of the data for analytical purposes.

Recommendation:

- No immediate action is required regarding duplicate records.
- Periodic monitoring and checks for duplicates can be incorporated into ongoing data quality assurance processes to ensure continued data integrity. This Duplicates Report provides a quick overview of the duplicate record status in the dataset, offering confidence in the uniqueness of the data for further analysis and decision-making.

1.1.3. Inconsistency

Analysis:

- Order ID: No duplicate records found. All Order IDs are unique.
- Ship Date: No duplicate records found. Dates vary across the dataset.
- Ship Mode: Four unique values representing different shipping modes.
- Customer ID: No duplicate records found. Each customer has a unique ID.
- Customer Name: Unique names for each customer.

Conclusion: The dataset appears to be consistent with unique Order IDs, Ship Dates, Ship Modes, Customer IDs, and Customer Names. No duplicate records were identified.

1.1.4. Redundancies Report

1. Row ID

- Redundancy: The Row ID correlation with itself is always 1.000000 (perfect correlation), which is expected as it is a self-correlation.
- Recommendation: Consider excluding the Row ID column from the correlation analysis as it doesn't provide meaningful insights.

2. Postal Code

- Redundancy: The Postal Code correlation with itself is always 1.000000 (perfect correlation), which is expected as it is a self-correlation.
- Recommendation: Consider excluding the Postal Code column from the correlation analysis as it doesn't provide meaningful insights.

3. Sales vs. Profit

- Redundancy: The correlation between Sales and Profit is 0.479064, indicating a moderate positive correlation.
- Recommendation: While this correlation is informative, it's crucial to note that Profit is a more direct measure of financial success. Consider evaluating Profit independently for a more direct analysis.

4. Discount Rate vs. Profit

- Redundancy: The correlation between Discount and Profit is -0.219487, indicating a moderate negative correlation.
- Recommendation: While this correlation is informative, it's essential to interpret the relationship carefully. High discounts might contribute to lower profits. Consider a detailed analysis of the impact of discounts on profit.

5. Sales vs. Quantity

- Redundancy: The correlation between Sales and Quantity is 0.200795, indicating a weak positive correlation.

- Recommendation: While this correlation is informative, it might be more insightful to analyze the relationship between Sales and Quantity in the context of the business. Consider exploring how changes in quantity sold affect overall sales.

6. Quantity vs. Profit

- Redundancy: The correlation between Quantity and Profit is 0.066253, indicating a weak positive correlation.

- Recommendation: While this correlation is informative, it's essential to recognize that Quantity alone might not be a strong predictor of profit. Consider exploring other factors that contribute to profit.

7. Sales vs. Discount

- Redundancy: The correlation between Sales and Discount is -0.028190, indicating a weak negative correlation.

- Recommendation: While this correlation is informative, it's essential to interpret the relationship carefully. Consider a more detailed analysis to understand the impact of discounts on sales.

8. Note

- It's recommended to carefully interpret correlations and consider additional factors for a comprehensive analysis.
- Depending on the business context, consider consulting domain experts for a deeper understanding of the relationships observed in the data. Correlation does not imply causation, and additional analyses are often needed for robust insights.

1.2. Statistical Analysis

1. Sales Distribution

- The mean sales value is \$229.86, indicating the average transaction amount.
- The minimum sales value is \$0.44, suggesting there are some transactions with very low sales.
- The maximum sales value is \$22,638.48, indicating a wide range of sales amounts.
- The standard deviation (std) is relatively high at \$623.25, indicating variability in sales.

2. Quantity Ordered

- The average quantity ordered is approximately 3.79 items per transaction.
- The minimum quantity ordered is 1, and the maximum is 14, with a standard deviation of 2.23.

3. Discounts

- On average, a discount of 15.62% is applied to transactions.
- The minimum discount is 0%, and the maximum is 80%.
- 75% of the transactions have a discount of 20% or less.

4. Profit and Loss

- The mean profit per transaction is \$28.66.
- The minimum profit is -\$6,599.98, indicating some transactions incurred significant losses.
- The maximum profit is \$8,399.98.
- 25% of transactions have a profit of \$1.73 or less.

5. Postal Code Analysis

- The postal codes vary widely, with a mean of 55,245.23 and a standard deviation of 32,038.72.
 - This suggests that sales and other metrics may vary across different regions.
6. General Trends • The 25th, 50th (median), and 75th percentiles provide insights into the distribution of data. • For example, 75% of transactions have a sales value of \$209.94 or less, and 50% have a sales value of \$54.49 or less.
7. Outliers • The presence of outliers is suggested by the significant difference between the mean and median values, especially in Sales and Profit. This indicates a potential skewness in the data.
8. Negative Profit • The presence of negative profit values indicates that some transactions resulted in losses. Further investigation may be needed to understand the reasons behind these losses.
- 1.3. Data Quality Summary • The dataset contains missing values in the "Postal Code" column, which may need to be addressed depending on the analysis requirements. • No duplicate records were found, indicating a clean dataset in terms of duplications. • The unique values in categorical columns provide insights into the cardinality of each feature. • Outliers are present in the numerical columns, especially in "Sales," "Quantity," "Discount," and "Profit." These outliers may need further investigation to understand their impact on analysis results.

Detailed quality report:

1. Missing Values • The "Postal Code" column has 11 missing values.
2. Duplicate Records • There are no duplicate records in the dataset.
3. Unique Values in Categorical Columns • "Order ID" has 5009 unique values. • "Order Date" has 1236 unique values. • "Ship Date" has 1334 unique values. • "Ship Mode" has 4 unique values. • "Customer ID" has 793 unique values. • "Customer Name" has 793 unique values. • "Segment" has 3 unique values. • "Country/Region" has 1 unique value. • "City" has 531 unique values. • "State" has 49 unique values. • "Region" has 4 unique values. • "Product ID" has 1862 unique values. • "Category" has 3 unique values. • "Sub-Category" has 17 unique values. • "Product Name" has 1849 unique values.
4. Outliers • "Sales" column has 1167 outliers. • "Quantity" column has 170 outliers. • "Discount" column has 856 outliers. • "Profit" column has 1881 outliers.

2. Target Audience We will assume general characteristics of the business, the data, and organizational structure in order to define an audience for operational and executive reports based on the sales data provided. With the provided data that covers sales, customers, products, and profits, among others, we can customize these reports based on the needs of different stakeholders in the organization

2.1 Operational Reports Target Audience: • Sales Managers • Marketing Teams • Product Managers
 Intended Use: The Operational reports are prepared for management and monitoring trends on monthly basis. They include information on sales performance. These reports help in monitoring and controlling the sales performance in relation to goals. Recognizing those places where sales and marketing tactics can be improved, product categories can be better evaluated and optimized. By using operational reports, Superstore can enhance its sales performance, operational efficiency, manage inventory more effectively, and adjust sales as well as marketing strategies to meet short-term objectives.

2.2 Executive Reports Target Audience: • C-Level Executives (CEO, CFO, COO, CMO) • Functional Heads (VP sales, VP marketing) • Strategy and Planning Teams
 Intended Use: The Executive report comprises: sales performance, key performance indicators (KPIs) from data that have been gathered from the monthly reports

done. They are higher level and broader, though, than operational reports, reporting on trends and strategic opportunities. These reports are used for: Decision-Making: Excelling in providing top-level strategic decisions involving market expansion, product development, and resource allocation. Performance Improvement: Analyzing long-term patterns evident in sales performance, profitability, and market share. Research Analysis: Directing research essence into new markets, customers segments, and product innovation.

Achievables: With executive reports, Superstore can make informed strategic decisions, anticipate market changes, improve long-term financial performance, and identify opportunities for growth and innovation. Conclusion: Operational and executive reports are equally important in defining a company's success. In contrast to operational reports which focus on the day-to-day management of sales performance. Executive reports provide the relevant strategic overview required for long-term planning and decision-making. Using such reports well, Superstore will not only be able to monitor and improve its current performance but also position itself strategically for future threats and opportunities.

3. Context and Additional Assumptions:

3.1 Context: The reports are being created in the context of analyzing the growth or decline of the superstore business. These reports can be utilized by operational teams, managers, and executives to monitor the business's performance, identify the best-performing regions, states, product categories and subcategories, and also to enhance the business's future market share and profitability.

3.2 Creation of Reports: Frequency: Monthly operational reports are generated to ensure timely monitoring of the superstore's performance without causing disruptions or delays in decision-making. Given the dynamic nature of the retail industry, monthly intervals strike a balance between providing relevant insights and avoiding excessive fluctuations in data. Data Analysis: Monthly intervals allow for a granular analysis of sales, profitability, and other key metrics, facilitating a deeper understanding of trends and patterns within shorter time frames. This enables operational teams including sales and marketing managers to promptly identify areas of growth or decline and take corrective actions as needed.

3.3 Maintenance: Yearly Analysis: While monthly reports are crucial for ongoing performance monitoring and decision-making, the annual executive report serves a broader purpose. Year-end analysis enables the superstore to conduct a comprehensive review of its annual performance, identify long-term trends, and assess the effectiveness of strategic initiatives implemented throughout the year. Business Evaluation: Yearly report also facilitates strategic planning for the upcoming year by providing valuable insights into overall business performance, market trends, and customer behavior. It serves as a basis for setting goals, refining strategies, and allocating resources effectively.

3.4 Utilization of the Reports: Decision Making: Previous month's reports are to be utilized throughout the subsequent month by operational teams, managers, and executives for decision-making. These reports serve as a guide for adjusting inventory levels, optimizing pricing, sales and marketing strategies, and allocating resources to maximize sales and profitability. Product Analysis and Promotion: Reports are leveraged to analyze the performance of different product categories and sub-categories. Products with low sales can be identified and targeted for promotional campaigns or markdowns to stimulate demand and clear inventory. Furthermore, insights from reports inform product development and procurement strategies to meet evolving customer preferences and market demands. Strategy Review and Adjustment: Reports play a pivotal role in reviewing current strategies and making necessary adjustments to

align with business objectives. Whether it's refining marketing tactics, enhancing customer service initiatives, or optimizing supply chain operations, data-driven insights from reports inform strategic decision-making and drive continuous improvement across all facets of the superstore's operations.

3.4 Assumptions:

1. We have assumed that the data is mostly accurate and reliable.
2. Since we have the data of 4 years and the superstore is effectively seeing an increase in the number of products sold per year, we can assume that it is an established online business which is trying to improve on profit margins.
3. We have assumed that the superstore produces its own products rather than depending on other stores/manufacturers for their products.
4. We have assumed that the market was stable throughout the period of the data provided.
5. It is assumed that the primary clients/customers of the business includes office spaces and small businesses (it is observed that one customer often purchases multiple products across various categories). The superstore might want to look towards other potential customer segments such as individual consumers and educational institutions.
6. The presence of outliers and negative profit (because of excessive discounts) indicates that the Superstore business might be looking to improve in the aspect of sales strategies.
7. From the varied sales and profits across different regions it can be assumed that the superstore would like to tailor their operational and sales strategies according to specific areas to optimize their sales.

4) Operational and Executive Reports

Operational Report We have selected a monthly frequency for our Operational Report since the volume of sales is relatively low. We will support and enhance day-to-day operations by facilitating monitoring geographic and product category performance by sales and shipping time as well as providing the ability to flag products yielding low or negative profits. Specific information contained in the report will include State and broader regional sales revenues and volumes by product category as well as product categories' sales including average and peak discount percentages and profit earned.

Executive Report We have selected a yearly frequency for our Executive Report to support and enhance strategic decision-making in Product Development, Logistics and Advertising. We are assuming there will be inventory, shipping, taxation, advertising, and other implications for our business based on the data (multi-state / interstate sales, some large items, and product types that may need to be updated regularly like Phones). Specific information contained in the report will include revenue and profit by Region and Product Category. Previous quarters' figures will be displayed alongside the current quarter's to easily see growth at a glance.

Key Performance Indicators (KPIs) will include target revenue and profit and percentage differential where applicable. The business is seeking an overall annual growth in revenue of 20%. Additionally, change in Average Order Value, Customer Retention Rate, and Customer-base growth will be calculated.

Sample formulas: $(\text{avg_total_order_value} / \text{num_customers}) ((\text{num_customers_end} - \text{num_customers_new}) / \text{num_customers_start}) * 100$

Operational and Executive Report Templates

Operational Report - Template									
Monthly Sales Performance Report									
Gross Sales(\$)									
12/01/21 - 12/31/21									
Region	States	Category	Sub-Category	Sales			Profit Margin		
				Actual	Target	Variance(%)	Actual	Target	Variance(%)
West									
	Total								
East									
	Total								
Central									
	Total								
South									
	Total								
TOTAL									

Notes:

1. This report encapsulates a detailed breakdown of monthly sales and profit margins for Superstore, segmented into four regions in the United States: West, East, Central, and South. Each region's data is further dissected by category and sub-category to provide a granular view of the sales performance.
2. The primary objective of this report is to meticulously monitor sales performance and profit margins—denominated in US dollars and percentages, respectively. This data-driven approach is intended to inform our sales and marketing strategies, ultimately aiming to maximize profitability in the forthcoming month.
3. The sales figures represent actual revenue, while the profit margin columns reflect the percentage of sales that has translated into profit. Variances are calculated as the percentage difference between actual figures and predetermined targets, offering a clear indicator of our performance against set goals.
4. In order to maintain a streamlined and efficient decision-making process, this report is prepared monthly and is due by the 3rd of each month. It undergoes a rigorous review process by the regional sales managers and finance managers. The final version of the report is to be shared with respective parties by the 7th of January.
5. This report contains sensitive business information, and it is strictly confidential and should not be disclosed to any third party without prior consent from Superstore management.
6. The report can be transformed based on the management's needs.

Executive Report - Template

Annual Sales Performance Report

Gross Sales(\$)

Updated on: 30-01-2022

Region	Sales						Profit Margin					
	Actual				Target	Variance(%)	Actual				Target	Variance(%)
	2021	2020	2019	2018	2021	2021	2021	2020	2019	2018	2021	2021
West												
East												
Central												
South												
Total												

Note:

This Annual Sales Performance Report provides an overview of the sales and profitability metrics across Superstore's regional divisions for the period of 2018 to 2021. The report is structured to facilitate a detailed analysis of performance trends and variance against set targets for 2021, with the intention to inform strategic decisions and improve profitability.

Key elements of the report include:

- 1.A breakdown by region, enabling a comparison of performance across West, East, Central, and South regions.
- 2.Year-over-year sales and profit figures, highlighting trends and enabling performance tracking over time.
- 3.Profit margins for the years 2018 through 2021, reflecting the effectiveness of Superstore's sales and operational strategies.
- 4.A comparison of actual profit margins against the targets for 2021, providing insight into Superstore's alignment with financial objectives.
- 5.The report is to be prepared quarterly, with a final comprehensive version generated at the fiscal year-end. The next update will be due by 30-04-2022.A preliminary report will be shared with the Regional Managers and the executive team one week prior to the final submission date for feedback and validation.
- 6.This report contains sensitive business information and should be handled in accordance with Superstore's confidentiality policies.

Lab Exercise 2

1. Analysis

Entities

1. Orders:

- Each row represents a specific order made by a customer.
- An order consists of multiple products.
- Contains information such as order ID, order date, shipping date, shipping method, payment method, and total price.

2. Customers:

- Represents individuals or organizations making purchases.
- Each customer may place multiple orders.

- Contains information such as customer ID, name, segment (e.g., Consumer, Corporate), country, city, state, postal code, and region.
3. **Products:**
 - Represents the items available for sale in the superstore.
 - Each product belongs to a specific category and sub-category.
 - Contains information such as product ID, product name, category (e.g., Furniture, Office Supplies), sub-category, and price.
 4. **Shipments:**
 - Contains details about the shipping associated with each order.
 - Each shipment corresponds to an order and contains information such as shipment ID, order ID (foreign key), shipment date, carrier, and tracking number.

Attributes:

1. **Orders:**
 - Order ID: Alphanumeric identifier for each order (Primary Key).
 - Order Date: Date when the order was placed.
 - Ship Date: Date when the order was shipped.
 - Shipping Method: Method used for shipping (e.g., Standard Class, Second Class).
 - Payment Method: Method used for payment (e.g., Cash, Credit Card).
 - Total Price: Total price of the order.
2. **Customers:**
 - Customer ID: Alphanumeric identifier for each customer (Primary Key).
 - Name: Name of the customer.
 - Segment: Segment to which the customer belongs (e.g., Consumer, Corporate).
 - Country, City, State, Postal Code, Region: Location details of the customer.
3. **Products:**
 - Product ID: Alphanumeric identifier for each product (Primary Key).
 - Product Name: Name of the product.
 - Category: Category to which the product belongs (e.g., Furniture, Office Supplies).
 - Subcategory: Subcategory of the product.
 - Price: Price of the product.
4. **Shipments:**
 - Shipment ID: Alphanumeric identifier for each shipment (Primary Key).
 - Order ID: Alphanumeric identifier linking the shipment to the corresponding order (Foreign Key).
 - Shipment Date: Date when the shipment was made.
 - Carrier: Shipping carrier responsible for delivering the shipment.
 - Tracking Number: Unique identifier to track the shipment.

Domains:

1. **Order ID:** Alphanumeric, unique identifier.
2. **Order Date, Ship Date, Shipment Date:** Date format.
3. **Shipping Method, Payment Method:** Text.
4. **Total Price, Price:** Numeric, currency format.

5. **Customer ID:** Alphanumeric, unique identifier.
6. **Name, Country, City, State, Region:** Text.
7. **Postal Code:** Numeric, specific format.
8. **Product Name, Category, Sub-category:** Text.

Referential Integrity:

1. **Customers to Orders:** Customer ID in Orders references Customer ID in Customers.
2. **Orders to Shipments:** Order ID in Shipments references Order ID in Orders.

2. Data Cleansing

1. Handling Missing Values

1. Missing Values Identification

The dataset was examined, revealing that the "Postal Code" column contained approximately 11 missing values.

2. Complete Column Comparison

While other columns were complete, a specific focus was placed on addressing the missing values in the "Postal Code" column.

3. Retaining Rows with Missing Postal Codes

Instead of opting for the removal of the 11 rows with missing postal codes, a strategic decision was made to retain them.

4. Rationale for Retention

The decision to retain rows with missing postal codes was grounded in the assessment that their absence would not significantly impact the ongoing analysis.

5. Potential for Future Enhancement

It was acknowledged that the missing values in the "Postal Code" column might be populated at a later time as additional details are obtained from customers.

6. Preservation of Valuable Data

Choosing to keep these rows preserves potentially valuable data, preventing the loss of information that may contribute to a more comprehensive analysis.

7. Correction Flexibility

The retention strategy allows for correction at a later point in time, aligning with the consideration of balancing the need for dataset completeness with the potential for future data enhancement.

8. Strategic Decision Alignment

This decision is aligned with a strategic approach that values the preservation of existing data while remaining open to future improvements.

Retaining rows with missing postal codes is a proactive decision, acknowledging the potential for future data enrichment and ensuring the preservation of valuable information within the dataset. This approach reflects a thoughtful balance between the immediate needs of completeness and the anticipation of future data enhancements.

2. Dropping Duplicates

1. Duplication Assessment

A thorough examination of the dataset was conducted to identify any duplicate records.

2. Result Summary

The investigation revealed that no duplicate entries were present in the dataset.

3. Ensuring Data Integrity

The absence of duplicate records is a crucial factor in ensuring data integrity.

4. Reliability Boost

The verification process contributes significantly to the overall reliability of the dataset.

5. Commitment to Accuracy

The decision to address and eliminate duplicates underscores a commitment to maintaining a dataset characterized by cleanliness and accuracy.

6. Enhanced Trustworthiness

This commitment enhances the dataset's trustworthiness, laying the foundation for more robust and dependable analytical outcomes.

The absence of duplicate records reinforces the dataset's integrity and sets the stage for reliable and trustworthy analytical endeavors.

3. Field Changes

3.1. Date Column Conversion

- The "Order Date" and "Ship Date" columns underwent a transformation and were converted to the datetime type.
- This conversion enhances the dataset's ability to handle and analyze temporal information more effectively.

3.2. Feature Engineering - Month and Year Extraction

- From the "Order Date" column, two new features were engineered: "Order Month" and "Order Year."
- The extraction of these attributes allows for a more detailed exploration of temporal patterns in the dataset.

3.3. Purpose of Extraction

- Extracting the month and year from the "Order Date" column enables a granular analysis of ordering trends over time.
- The new features provide valuable insights into monthly and yearly variations in the dataset, facilitating a more nuanced understanding of temporal patterns.

3.4. Enhanced Analytical Capabilities

- These modifications not only improve the dataset's usability but also lay the groundwork for more sophisticated temporal analyses.
- Users can now explore and interpret trends and patterns with greater precision, leveraging the added dimensions of month and year.

3.5. Total Sales per Customer

- A new column, "Total Sales per Customer," was created by aggregating the total sales for each customer.
- This additional metric provides a holistic view of customer spending, aiding in the identification of high-value customers and contributing to customer-centric analyses.

3.6. Profit Margin Calculation

- Another new column, "Profit Margin," was introduced by calculating the percentage of profit relative to sales.
- This metric offers insights into the profitability of transactions and can help identify products or orders with the highest and lowest profit margins.

3.7. Reasoning for Additional Columns

- The creation of "Total Sales per Customer" and "Profit Margin" columns enriches the dataset with valuable metrics for customer-centric and profitability analyses.
- These new features enhance the depth of analytical exploration, providing a more comprehensive understanding of customer behavior and transaction profitability.

4. Outlier handling

Before Outlier Handling

1. Understanding the Distribution

- The "Sales" column exhibits a wide range of values, as evident from the descriptive statistics:
- Mean: \$229.86
- Standard Deviation: \$623.25
- Minimum: \$0.44
- 25th Percentile (Q1): \$17.28

- Median (50th Percentile): \$54.49
- 75th Percentile (Q3): \$209.94
- Maximum: \$22,638.48

The distribution has a substantial spread, with a notable difference between the median and the mean, indicating potential skewness.

The maximum value of \$22,638.48 suggests the presence of outliers at the upper end of the distribution.

Outlier Handling using IQR by Sub-Category

1. IQR Calculation by Sub-Category

The Interquartile Range (IQR) was calculated separately for each sub-category within the "Sales" column.

2. Identifying and Handling Outliers by Sub-Category

A threshold of 1.5 times the IQR was applied to identify potential outliers within each sub-category.

Outliers were replaced with None within each sub-category.

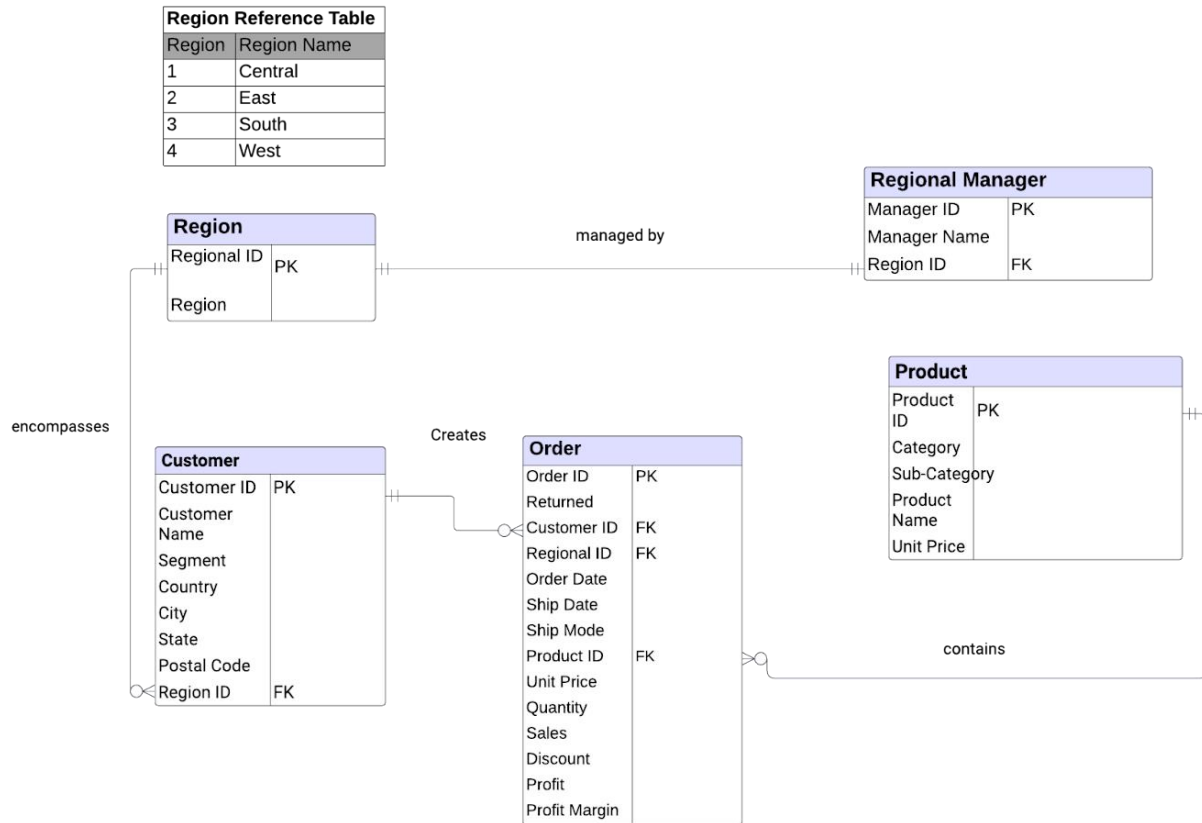
3. Handling Strategy by Sub-Category

Outliers in the "Sales" column were replaced with None within each sub-category as part of the handling strategy.

After Outlier Handling

This modification ensures that outlier handling is performed independently within each sub-category, allowing for a more nuanced and category-specific approach. The report sections provide insights into the handling strategy and its impact on the "Sales" distribution within each sub-category.

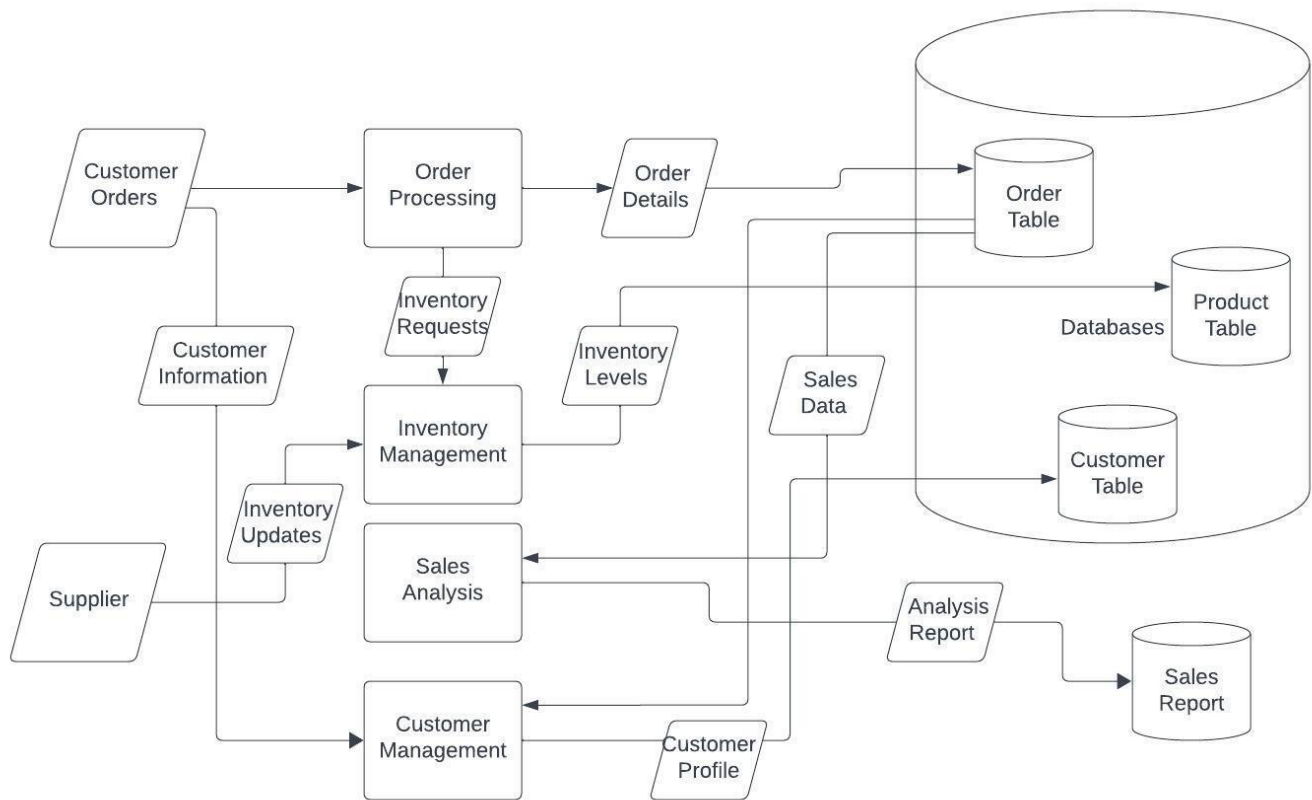
3. Logical-level ERD



Customer Master Data Table							
Customer ID (PK)	Customer	Segment	Country	City	State	Postal	Region ID
(unique value)	(full name)	(segment)	(name)	(name)	(name)	(number)	(unique value)

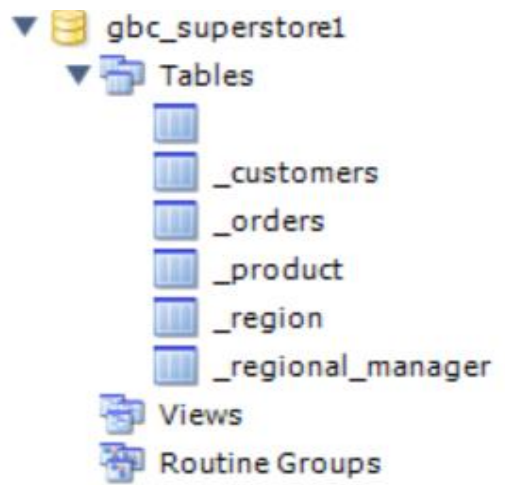
Metadata - Customer Table						
Column	Data Type	Domain	Description	Last	Updated By	
Customer ID	VARCHAR	Alphanumeric	Unique identifier for each customer			
Customer Name	VARCHAR	Text	Full name of the customer			
Segment	VARCHAR	{'Consumer', 'Corporate', 'Home	Classification of customers based on the segment			
Country	VARCHAR	{'United States'}	Country where the customer is located			
Region ID	INT	{'1', '2', '3', '4'}	Unique identifier for each geographical region where			
State	VARCHAR	Text	State where the customer is located			
City	VARCHAR	Text	City where the customer is located			
Postal Code	INT	Numeric	Postal code of the customer's location			

4. Data Flow

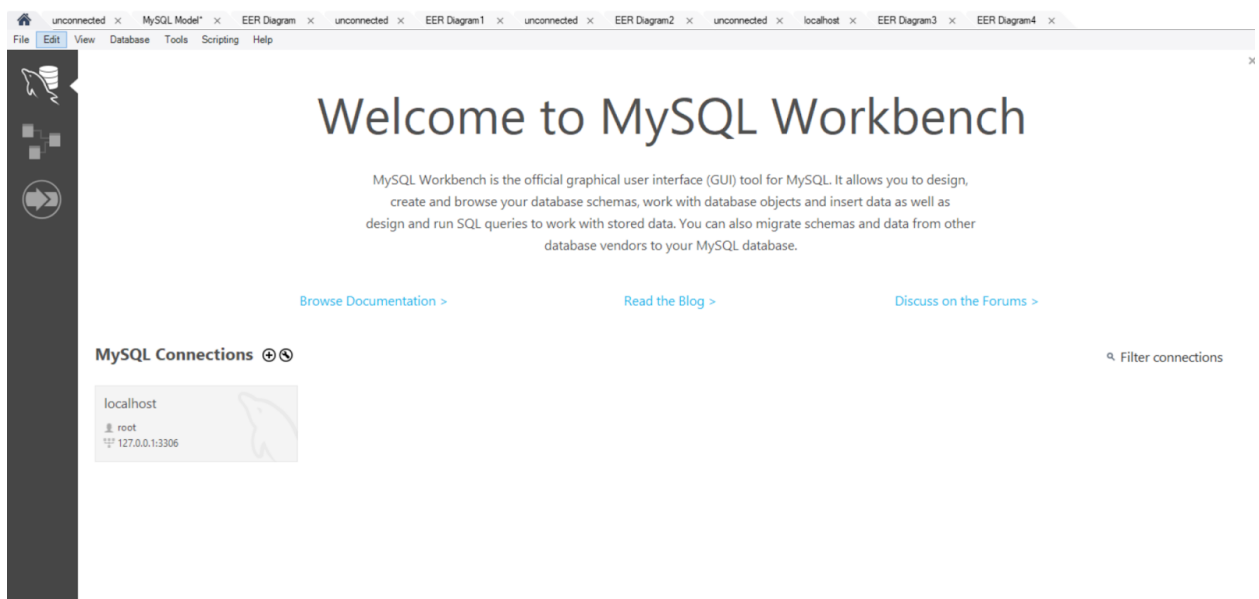


5. MySQLInstallationL Installation

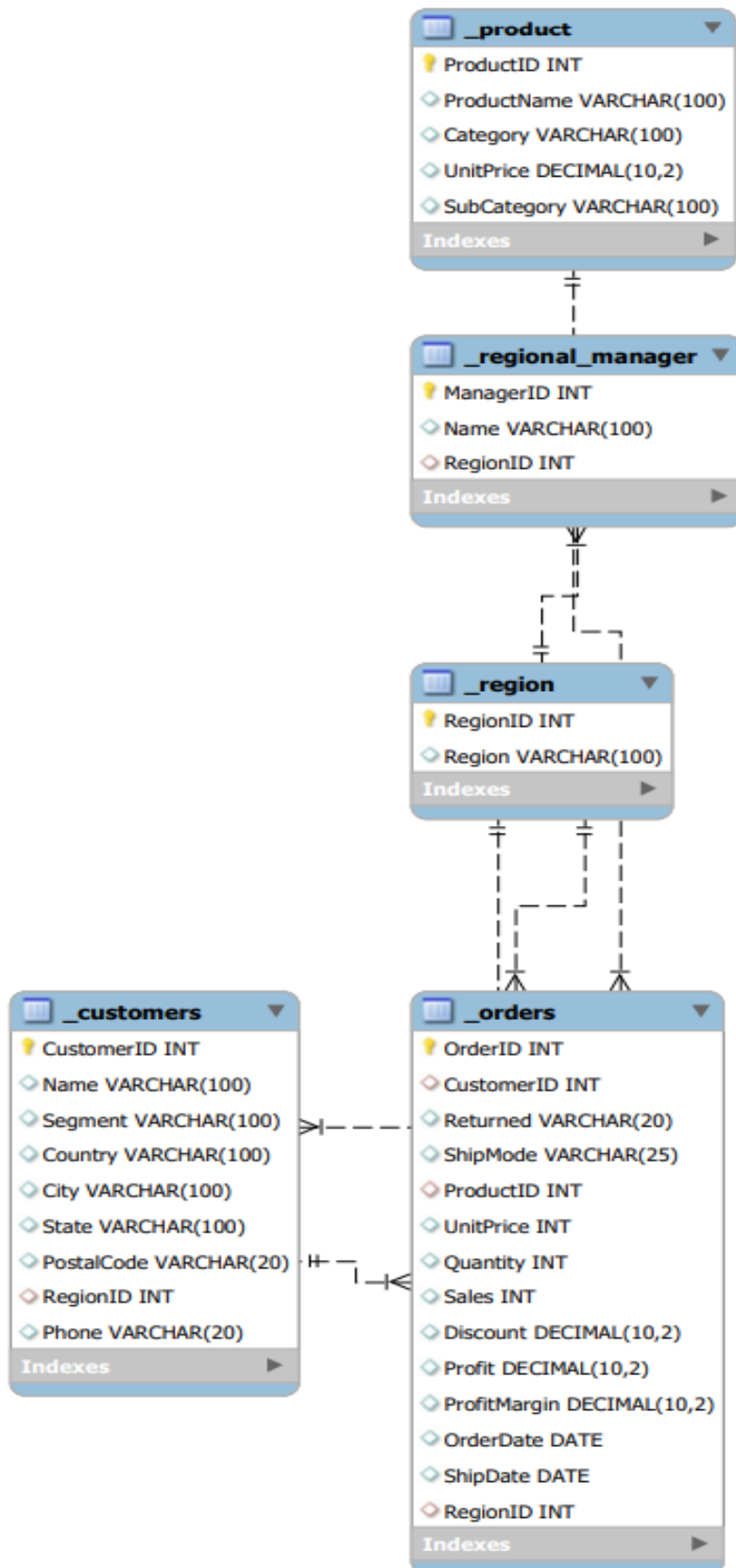
5.1. Database Creation



5.2. Workbench:

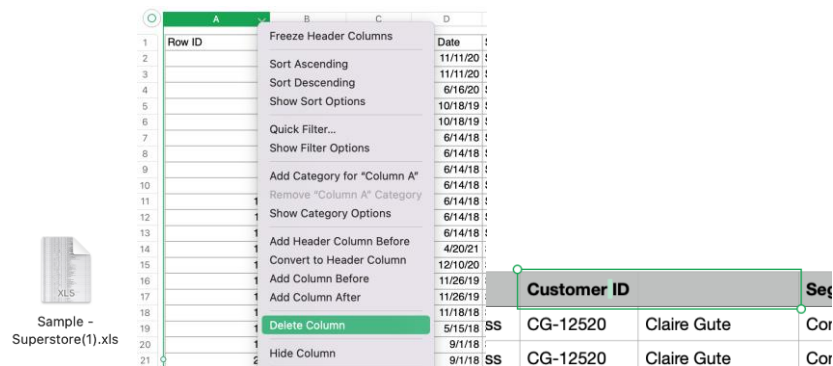


6. Database Schema:



7. ETL

Deleted unneeded columns, renamed columns, and exported XLS as CSV using Numbers



Export Your Spreadsheet

Excel CSV TSV

- Tables: ☐ Create a file for each table
☒ Combine tables into a single file
☐ Include table names

Advanced Options

Text Encoding: Unicode (UTF-8)

Imported CSV into Pandas DataFrame using Python

```
[3]: import pandas as pd
```

```
[4]: superstore_df = pd.read_csv('~/.JUPYTER/DATA/Superstore.csv')  
superstore_df.head()
```

[4]:	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country/Region	City
0	CA-2020-152156	11/8/20	11/11/20	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson
1	CA-2020-152156	11/8/20	11/11/20	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson

Created Foreign Keys based on data as needed

```
superstore_df['RegionID'].value_counts()
```

```
Region
West      3203
East      2848
Central   2323
South     1620
Name: count, dtype: int64
```

```
superstore_df['RegionID'] = superstore_df['Region'].apply(lambda x: 1 if x == 'West' else 2 if x == 'East' else 3 if x == 'Central' else 4)
```

```
superstore_df.head()
```

ier ne	Segment	Country/Region	City	State	...	Region	Product ID	Category	Sub- Category	Product Name	Sales	Quantity	Discount	Profit	RegionID
ire jte	Consumer	United States	Henderson	Kentucky	...	South	FUR-BO- 10001798	Furniture	Bookcases	Bush Somerset Collection Bookcase	261.9600	2.0	0.00	41.9136	4

Note: This could be challenging with many values but in our case there were few

Updated ID column values to be the correct types

```
superstore_df['CustomerID'] = superstore_df['CustomerID'].str.slice(start=3)
```

Note: some 'Customers' table data may have been lost here. We hadn't accounted for using an Integer ID when the data's IDs were meaningful and had string prefixes

```
superstore_df['OrderDate'] = pd.to_datetime(superstore_df['OrderDate'], format='%m/%d/%y', errors='coerce')
superstore_df['ShipDate'] = pd.to_datetime(superstore_df['ShipDate'], format='%m/%d/%y', errors='coerce')
```

Dropped duplicate IDs and rows missing ID

```
customers_df = superstore_df[['CustomerID', 'Name', 'Segment', 'Country', 'City', 'State', 'PostalCode', 'RegionID']].copy()
```

```
customers_df.drop_duplicates(subset='CustomerID', inplace=True)
customers_df.dropna(subset='CustomerID', inplace=True)
customers_df.head()
```

Inserting rows

```
from sqlalchemy import create_engine, text
import pymysql
```

```
engine = create_engine('mysql+pymysql://root:root@localhost:3306/GBC_Superstore1')
```

```
table_name = '_Product'
products_df.to_sql(table_name, con=engine, if_exists='append', index=False)
```

1596

Referential integrity

```
import pandas as pd

superstore_df = pd.read_csv('~/.JUPYTER/DATA/Superstore.csv')
superstore_df.head(1)

superstore_df.columns

superstore_df['ProductID'] = superstore_df['ProductID'].str.slice(start=7)

products_df = superstore_df[['ProductID', 'Category', 'SubCategory']].copy()
products_df['UnitPrice'] = (superstore_df['Sales'] / (superstore_df['Discount'] + 1)) / superstore_df['Quantity']
products_df.drop_duplicates(subset='ProductID', inplace=True)
products_df.head()

from sqlalchemy import create_engine, text
import pymysql

engine = create_engine('mysql+pymysql://root:root@localhost:3306/GBC_Superstore1')

table_name = '_Product'
products_df.to_sql(table_name, con=engine, if_exists='append', index=False)

superstore_df['Region'].value_counts()

superstore_df['RegionID'] = superstore_df['Region'].apply(lambda x: 1 if x == 'West' else 2 if x == 'East' else 3 if x == 'Central' else 4)
#superstore_df.head()

region_df = superstore_df[['RegionID', 'Region']].dropna().drop_duplicates()

table_name = '_Region'
region_df.to_sql(table_name, con=engine, if_exists='append', index=False)

superstore_df['CustomerID'] = superstore_df['CustomerID'].str.slice(start=3)
superstore_df.head(1)

customers_df = superstore_df[['CustomerID', 'Name', 'Segment', 'Country', 'City', 'State', 'PostalCode', 'RegionID']].copy()
customers_df.drop_duplicates(subset='CustomerID', inplace=True)
print(len(customers_df))
customers_df.dropna(subset='CustomerID', inplace=True)
print(len(customers_df))
customers_df.head()

table_name = '_Customers'
customers_df.to_sql(table_name, con=engine, if_exists='append', index=False)

superstore_df['OrderID'] = superstore_df['OrderID'].str.slice(start=3)
superstore_df['OrderID'] = superstore_df['OrderID'].str.replace('-', '')

superstore_df['OrderDate'] = pd.to_datetime(superstore_df['OrderDate'], format='%m/%d/%y', errors='coerce')
superstore_df['ShipDate'] = pd.to_datetime(superstore_df['ShipDate'], format='%m/%d/%y', errors='coerce')
superstore_df.head(1)

orders_df = superstore_df[['OrderID', 'CustomerID', 'ShipMode', 'ProductID', 'Quantity',
                           'Sales', 'Discount', 'Profit', 'OrderDate', 'ShipDate', 'RegionID']].copy()
orders_df.drop_duplicates(subset='OrderID', inplace=True)
orders_df.drop_duplicates(subset='ProductID', inplace=True)
orders_df.head()

table_name = '_Orders'
orders_df.to_sql(table_name, con=engine, if_exists='append', index=False)
```

“Screenshots from Jupyter Notebook 7.0.6, Python 3.11.7, MySQL 5.7.39”

Since we inserted the rows into the tables having their constraints enabled in the following order: Products, Regions, Customers, Orders; since there were not errors, we can be sure there is referential integrity

Lab Exercise 3

Operational Report

Superstore Sales Operational Report over the ~4 Year Period (3 years and 362 days)
03/01/2018 – 30/12/2021

Region	State	Category	Total Quantity	Average Discount (%)	Gross Sales (\$)	Profit (\$)
Central	Illinois	Furniture	432	0.463865546	27907.49	-8705.7226
		Office Supplies	1037	0.411111111	19118.124	-8147.9047
		Technology	302	0.205952381	31983.673	4822.5592
	Illinois Total	1771	0.360309679	79009.287	-12031.0681	
	Indiana	Furniture	83	0	11496.71	2181.2753
		Office Supplies	374	0	15654.47	5162.2849
		Technology	106	0	26323.25	11000.8773
	Indiana Total	563	0	53474.43	18344.4375	
	Iowa	Furniture	24	0	2642.31	520.0385
		Office Supplies	66	0	723.16	317.9622
		Technology	13	0	1154.3	318.3682
	Iowa Total	103	0	4519.77	1156.3689	
	Kansas	Furniture	8	0	111.12	36.9696
		Office Supplies	47	0	1954.15	624.4873
		Technology	19	0	849.04	174.9866
	Kansas Total	74	0	2914.31	836.4435	
	Michigan	Furniture	181	0	22260.26	4652.4324
		Office Supplies	589	0.010457516	37521.349	14910.2044
		Technology	148	0.004444444	16209.975	4778.3119
	Michigan Total	918	0.00496732	75991.584	24340.9487	
	Minnesota	Furniture	52	0	7611.35	2023.8871
		Office Supplies	241	0	19406.54	7780.4995
		Technology	38	0	2845.26	1018.8008
	Minnesota Total	331	0	29863.15	10823.1874	
	Missouri	Furniture	29	0	2390.57	537.8761
		Office Supplies	158	0	12080.82	2747.42
		Technology	54	0	7086.52	3014.7268
	Missouri Total	241	0	21557.91	6300.0229	
	Nebraska	Furniture	21	0	1944.7	518.4364
		Office Supplies	86	0	2216.85	548.4992
		Technology	26	0	3285.74	961.515
	Nebraska Total	133	0	7447.29	2028.4506	
	North Dakota	Office Supplies	30	0	919.91	230.1497
	North Dakota Total	30	0	919.91	230.1497	
	Oklahoma	Furniture	57	0	8284.1	2153.8622
		Office Supplies	119	0	4489.6	1113.7811
		Technology	67	0	6368.45	1580.9003
	Oklahoma Total	243	0	20981.97	4848.5436	
	South Dakota	Furniture	5	0	324.9	67.1898
		Office Supplies	31	0	597.72	193.7419
		Technology	6	0	392.94	133.8966
	South Dakota Total	42	0	1315.56	394.8283	
	Texas	Furniture	758	0.4212	60553.0238	-10410.4531
		Office Supplies	2221	0.394871795	43896.384	-18415.9623
		Technology	659	0.21452514	65104.224	3291.429
	Texas Total	3638	0.343532312	169553.6318	-25534.9864	
	Wisconsin	Furniture	144	0	17256.61	3838.9545
		Office Supplies	211	0	6037.12	1955.5522
		Technology	104	0	8798.16	2597.0697
	Wisconsin Total	459	0	32091.89	8391.5764	
Central Grand Total			8516	0.059067443	498349.372	39898.7533

East	Connecticut	Furniture	46	0.046153846	5174.987	1226.2805
		Office Supplies	202	0	5351.78	1473.6601
		Technology	29	0	2791.03	780.9336
	Connecticut Total		277	0.015384615	13317.797	3480.8742
	Delaware	Furniture	62	0.035294118	4547.359	828.3152
		Office Supplies	233	0	7926.59	2769.0582
		Technology	59	0	14562.22	6239.0508
	Delaware Total		354	0.011764706	27036.169	9836.4242
	District of Columbia	Furniture	8	0	1346.58	350.0835
		Office Supplies	24	0	138.52	60.9434
		Technology	8	0	1379.92	648.5624
	District of Columbia Total		40	0	2865.02	1059.5893
	Maine	Furniture	2	0	109.48	33.9388
		Office Supplies	19	0	399.8	169.5146
		Technology	14	0	761.25	251.0328
	Maine Total		35	0	1270.53	454.4862
	Maryland	Furniture	104	0.021428571	9149.253	1905.8274
		Office Supplies	267	0	10345.48	3781.8888
		Technology	44	0	4166.04	1322.8776
	Maryland Total		415	0.007142857	23660.773	7010.5938
	Massachusetts	Furniture	103	0.065625	10873.224	1074.4871
		Office Supplies	306	0	11828.35	3863.5205
		Technology	69	0	5726.63	1755.7501
	Massachusetts Total		478	0.021875	28428.204	6693.7577
	New Hampshire	Furniture	27	0.05	1886.474	153.937
		Office Supplies	80	0	1769.25	649.8585
		Technology	20	0	3636.8	902.7073
	New Hampshire Total		127	0.016666667	7292.524	1706.5028
	New Jersey	Furniture	85	0.023076923	6307.042	932.3293
		Office Supplies	268	0	13792.37	4633.6443
		Technology	86	0	14501.16	4170.198
	New Jersey Total		439	0.007692308	34600.572	9736.1716
	New York	Furniture	860	0.109090909	92661.569	5572.5977
		Office Supplies	2510	0.051212121	89338.534	25761.8278
		Technology	759	0.006763285	127453.53	42172.6997
	New York Total		4129	0.055688772	309453.633	73507.1252
	Ohio	Furniture	328	0.288764045	23141.241	-4366.8568
		Office Supplies	987	0.334469697	17800.412	-54.4197
		Technology	384	0.339215686	35675.992	-12649.9401
	Ohio Total		1699	0.320816476	76617.645	-17071.2166
	Pennsylvania	Furniture	462	0.280672269	38630.939	-7237.4739
		Office Supplies	1158	0.332075472	34136.957	-5009.2946
		Technology	406	0.341525424	42143.341	-3199.6132
	Pennsylvania Total		2026	0.318091055	114911.237	-15446.3817
	Rhode Island	Furniture	53	0.075	5918.756	913.377
		Office Supplies	106	0	6207.59	1761.7968
		Technology	35	0	10474.41	4598.0123
	Rhode Island Total		194	0.025	22600.756	7273.1861
	West Virginia	Furniture	3	0.3	673.344	-76.9536
		Office Supplies	15	0	536.48	262.8752
	West Virginia Total		18	0.15	1209.824	185.9216
East Grand Total		10231	0.073086343	663264.684	88427.0344	

South	Alabama	Furniture	54	0	6332.48	1231.3882
		Office Supplies	149	0	4209.08	1257.6342
		Technology	53	0	8969.08	3297.8029
	Alabama Total		256	0	19510.64	5786.8253
	Arkansas	Furniture	43	0	3187.55	781.4552
		Office Supplies	132	0	4382.39	1879.8117
		Technology	62	0	3925.25	1261.4384
	Arkansas Total		237	0	11495.19	3922.7053
	Florida	Furniture	292	0.23313253	22556.87	-2252.7049
		Office Supplies	805	0.344144144	19363.801	-1659.7791
		Technology	258	0.226470588	46956.212	529.7702
	Florida Total		1355	0.267915754	88876.883	-3382.7138
	Georgia	Furniture	110	0	7726.1	1751.768
		Office Supplies	426	0	26511.65	9782.2192
		Technology	155	0	14058.55	4399.6597
	Georgia Total		691	0	48296.3	15933.6469
	Kentucky	Furniture	106	0	12126.84	3210.9932
		Office Supplies	320	0	11894.27	3832.067
		Technology	97	0	12570.64	4156.6364
	Kentucky Total		523	0	36591.75	11199.6966
	Louisiana	Furniture	36	0	2963.03	685.9946
		Office Supplies	79	0	3423.16	495.0925
		Technology	41	0	2830.84	1015.0152
	Louisiana Total		156	0	9217.03	2196.1023
	Mississippi	Furniture	43	0	4317.85	944.8196
		Office Supplies	135	0	3589.29	1221.4665
		Technology	38	0	2822.33	986.4024
	Mississippi Total		216	0	10729.47	3152.6885
	North Carolina	Furniture	174	0.238095238	15155.484	-3486.4633
		Office Supplies	619	0.311842105	14309.609	-417.1361
		Technology	174	0.223529412	26083.119	-3583.304
	North Carolina Total		967	0.257822252	55548.212	-7486.9034
	South Carolina	Furniture	22	0	3078.25	612.8439
		Office Supplies	117	0	3316.38	465.5674
		Technology	20	0	1591.62	453.5875
	South Carolina Total		159	0	7986.25	1531.9988
	Tennessee	Furniture	165	0.235555556	13506.732	-2208.6291
		Office Supplies	413	0.313636364	12108.755	-3136.7774
		Technology	74	0.23	4807.283	66.0337
	Tennessee Total		652	0.25973064	30422.77	-5279.3728
	Virginia	Furniture	233	0	25321.95	5204.3265
		Office Supplies	492	0	20871.54	5856.9156
		Technology	140	0	24115.6	7399.7732
	Virginia Total		865	0	70309.09	18461.0153
	South Grand Total		6077	0.071406241	388983.585	46035.689

South Grand Total			6077	0.071406241	388983.585	46035.689	
West	Arizona	Furniture	177	0.293333333	12882.627	-2825.517	
		Office Supplies	488	0.342276423	9651.393	-726.3014	
		Technology	173	0.23125	11750.885	112.5012	
	Arizona Total		838	0.288953252	34284.905	-3439.3172	
	California	Furniture	1664	0.110091743	153836.0715	8495.438	
		Office Supplies	4370	0.045778938	137548.35	36733.3991	
		Technology	1394	0.117647059	159183.17	29440.3665	
	California Total		7428	0.09117258	450567.5915	74669.2036	
	Colorado	Furniture	191	0.316326531	12467.981	-2669.8166	
		Office Supplies	339	0.343617021	7850.926	-348.2794	
		Technology	149	0.242857143	10966.329	-3471.5845	
	Colorado Total		679	0.300933565	31285.236	-6489.6805	
	Idaho	Furniture	18	0.033333333	2595.482	533.9665	
		Office Supplies	27	0.111111111	840.506	149.9901	
		Technology	12	0.15	837.498	92.0973	
	Idaho Total		57	0.098148148	4273.486	776.0539	
	Montana	Furniture	7	0	63.98	21.7532	
		Office Supplies	29	0.044444444	1856.342	286.3301	
		Technology	18	0.1	3662.934	1523.0354	
	Montana Total		54	0.048148148	5583.256	1831.1187	
	Nevada	Furniture	37	0.022222222	4635.172	524.5705	
		Office Supplies	93	0.060869565	6870.724	2254.3518	
		Technology	33	0.133333333	5137.006	512.8456	
	Nevada Total		163	0.072141707	16642.902	3291.7679	
	New Mexico	Furniture	19	0.1	1701.412	251.5917	
		Office Supplies	91	0.041666667	1384.182	568.2416	
		Technology	41	0.088888889	1697.928	337.2828	
	New Mexico total		151	0.076851852	4783.522	1157.1161	
	Oregon	Furniture	69	0.342857143	6338.13	-1487.5769	
		Office Supplies	299	0.290277778	5122.416	154.4406	
		Technology	107	0.237037037	5821.556	126.3564	
	Oregon Total		475	0.290057319	17282.102	-1206.7799	
	Utah	Furniture	29	0	4822.35	631.7557	
		Office Supplies	113	0.054545455	3442.782	1141.365	
		Technology	45	0.114285714	2309.904	581.7351	
	Utah Total		187	0.056277056	10575.036	2354.8558	
	Washington	Furniture	417	0.061818182	47503.892	7012.0558	
		Office Supplies	1041	0.058273381	38557.562	10947.9131	
		Technology	351	0.082828283	50528.718	15016.6462	
	Washington Total		1809	0.067639949	136590.172	32976.6151	
	Wyoming	Furniture	4	0.2	1603.136	100.196	
	Wyoming Total		4	0.2	1603.136	100.196	
West Grand Total			11845	0.144574871	713471.3445	106021.1495	
			United States Grand Total				
			Total Quantity	Average Discounts	Gross Sales	Total Profit	
			36669	0.087033725	2264068.986	280382.6262	

Notes –

1. Breakdown by regions –

a. Central:

Total Quantity sold – 8516 products.

Gross Sales - \$ 498349.372

Total Profit - \$ 39898.7533

b. East:

Total Quantity sold – 10231 products.

Gross Sales - \$ 663264.684

Total Profit - \$ 88427.0344

c. South:

Total Quantity sold – 6077 products.

Gross Sales - \$ 388983.585

Total Profit - \$ 46035.689

d. West:

Total Quantity sold – 11845 products.

Gross Sales - \$ 713471.3445

Total Profit - \$ 106021.1495

2. Grand Total for all regions (USA):

Total Quantity sold – 36669 products.

Gross Sales - \$ 2264068.986

Total Profit - \$ 280382.6262

3. The operational report is adaptable to suit the specific needs of stakeholders. Depending on their requirements, additional sub-categories or city-level data can be incorporated to enhance the level of detail and comprehensiveness within the report.
4. The report offers comprehensive insights into the superstore's sales performance spanning a four-year period (03/01/2018 – 30/12/2021). It delves into regional and city-specific sales data, shedding light on top-selling product categories. Furthermore, the report meticulously outlines key financial metrics including profits, sales figures, discounts, and product quantities sold across various regions and states.

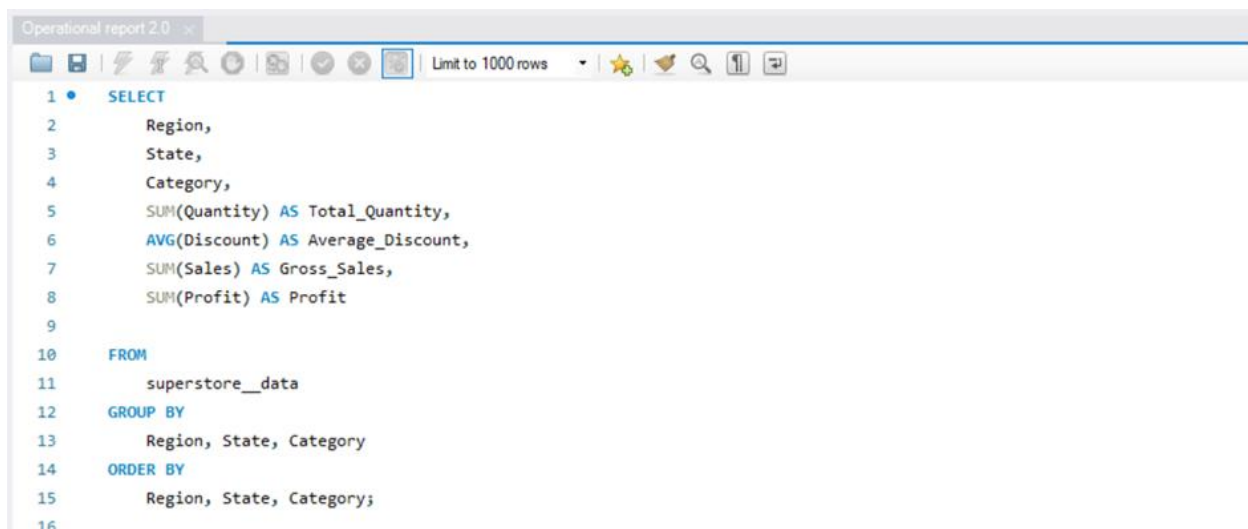
Operational Report Creation Process -

The cleaned superstore data was loaded into the table 'superstore__data' using SQL Workbench.

The operational report was generated using the SQL script titled 'Operational report 2.0' (please refer to the attached SQL script file and the screenshots below for the script). The CSV file containing the operational report was generated using SQL Workbench by executing the script. Subsequently, the report was formatted in Excel to enhance readability.

The CSV and SQL script files for the operational report are submitted along with the final project report.

SQL Script:

A screenshot of the SQL Workbench interface. The title bar shows 'Operational report 2.0'. The toolbar includes icons for file operations, execution, and a 'Limit to 1000 rows' dropdown. The SQL editor contains the following script:

```
1 • SELECT
2     Region,
3     State,
4     Category,
5     SUM(Quantity) AS Total_Quantity,
6     AVG(Discount) AS Average_Discount,
7     SUM(Sales) AS Gross_Sales,
8     SUM(Profit) AS Profit
9
10  FROM
11     superstore__data
12  GROUP BY
13     Region, State, Category
14  ORDER BY
15     Region, State, Category;
16
```

Executive report

Superstore Yearly Executive Report Gross Profit (\$)

Version 2.0

Author: Group 4

Region	2018 Profit (\$)	2019 Profit (\$)	2019 Profit vs 2018 Profit (\$)	2020 Profit (\$)	2020 Profit vs 2019 Profit (\$)	2021 Profit (\$)	2021 Profit vs 2020 Profit (\$)
South	11745.8151	7981.6786	-3764.1365	17712.8445	9731.1659	8595.3508	-9117.4937
West	19728.449	20184.8163	456.3673	23035.8991	2851.0828	43071.9851	20036.086
Central	584.693	11691.6566	11106.9636	19734.3279	8042.6713	8118.2255	-11616.1024
East	16985.4692	19857.6112	2872.142	18858.5055	-999.1057	32725.4485	13866.943
Overall Growth (\$)			10671.3364		19625.8143		13169.4329

Notes –

1. Profit Variance between consecutive years:

a. 2019 VS 2018 –

Total Growth in Profit (\$) 10671.33.

b. 2019 VS 2018 –

Total Growth in Profit (\$) 19625.81.

c. 2019 VS 2018 –

Total Growth in Profit (\$) 13169.43.

2. This executive report provides the information about the overall growth in the profit (in \$) for the superstore over the 4-year period (03/01/2018 – 30/12/2021).

Executive report Creation Process:

The executive report was generated using the SQL script titled 'executivereport1 script' (please refer to the attached SQL script file and the screenshots below for the script). The CSV file of the executive report was also formatted in Excel to improve readability.


```
Operational report 2.0 | executivereport1 script x
Limit to 1000 rows

1
2 • SELECT Region, SUM(Profit) AS '2018 Profit'
3 FROM superstore__data
4 WHERE YEAR(OrderDate) = 2018
5 GROUP BY Region;
6
7 -- 2019 Sales
8 • SELECT Region, SUM(Profit) AS '2019 Profit'
9 FROM superstore__data
10 WHERE YEAR(OrderDate) = 2019
11 GROUP BY Region;
12
13 -- 2020 Sales
14 • SELECT Region, SUM(Profit) AS '2020 Profit'
15 FROM superstore__data
16 WHERE YEAR(OrderDate) = 2020
17 GROUP BY Region;
18
19 -- 2021 Sales
20 • SELECT Region, SUM(Profit) AS '2021 Profit'
21 FROM superstore__data
22 WHERE YEAR(OrderDate) = 2021
23 GROUP BY Region;
24
24 • SELECT s.Region,
25         s18.`2018 Profit`,
26         s19.`2019 Profit`,
27         (s19.`2019 Profit` - s18.`2018 Profit`) AS '2019 Profit vs 2018 Profit',
28         s20.`2020 Profit`,
29         (s20.`2020 Profit` - s19.`2019 Profit`) AS '2020 Profit vs 2019 Profit',
30         s21.`2021 Profit`,
31         (s21.`2021 Profit` - s20.`2020 Profit`) AS '2021 Profit vs 2020 Profit'
32 FROM
33     (SELECT DISTINCT Region FROM superstore__data) s
34 LEFT JOIN
35     (SELECT Region, SUM(Profit) AS '2018 Profit'
36      FROM superstore__data
37      WHERE SUBSTRING_INDEX(OrderDate, '-', -1) = '2018'
38      GROUP BY Region) s18
39 ON s.Region = s18.Region
40 LEFT JOIN
41     (SELECT Region, SUM(Profit) AS '2019 Profit'
42      FROM superstore__data
43      WHERE SUBSTRING_INDEX(OrderDate, '-', -1) = '2019'
44      GROUP BY Region) s19
45 ON s.Region = s19.Region
```

```
46     LEFT JOIN
47     (SELECT Region, SUM(Profit) AS '2020 Profit'
48     FROM superstore__data
49     WHERE SUBSTRING_INDEX(OrderDate, '-', -1) = '2020'
50     GROUP BY Region) s20
51     ON s.Region = s20.Region
52     LEFT JOIN
53     (SELECT Region, SUM(Profit) AS '2021 Profit'
54     FROM superstore__data
55     WHERE SUBSTRING_INDEX(OrderDate, '-', -1) = '2021'
56     GROUP BY Region) s21
57     ON s.Region = s21.Region;
58
```