



INTERNATIONAL  
SCHOOL  
VIETNAM NATIONAL UNIVERSITY, HANOI

# FINAL REPORT



## DATA WAREHOUSE

### ADVENTUREWORKS DATABASE PROJECT

ID: INS307302

Lecturer: Dr. Hoang Trong Tien

## GROUP 02

| Name              | ID       |
|-------------------|----------|
| Ngô Mai Anh       | 20070892 |
| Lê Minh Hải       | 20070923 |
| Nguyễn Yến Nhi    | 20071039 |
| Lê Phan Anh Thư   | 20070986 |
| Nguyễn Tuấn Thành | 20070980 |

## Table of content

|  |           |
|--|-----------|
| <b>I. Introduction</b>                         | <b>2</b>  |
| 1. The AdventureWorks Database                 | 2         |
| 2. Problem Overview                            | 2         |
| • Project Objectives                           | 3         |
| 3. Key Questionnaire                           | 3         |
| 4. Project Design                              | 5         |
| <b>II. Entity Relationship Diagram Design</b>  | <b>6</b>  |
| <b>III. Building the Data Warehouse</b>        | <b>8</b>  |
| 1. Step 1: Choose the Business Process         | 8         |
| 2. Step 2: Declare the Grain                   | 8         |
| 3. Step 3: Identify the Dimensions             | 13        |
| 4. Step 4: Identify the Facts                  | 14        |
| Airflow DAG:                                   | 16        |
| <b>III. Results &amp; Implications</b>         | <b>20</b> |
| 1. Querying On OLAP                            | 20        |
| 2. Customer Segmentation                       | 29        |
| 3. Exploratory Data Analysis                   | 31        |
| 3.1. Data Overview:                            | 31        |
| 3.2. Key Findings:                             | 32        |
| 3.2.1. Product                                 | 33        |
| 3.2.2. Territory                               | 37        |
| 3.2.3. Sales persons                           | 42        |
| 3.2.4. Sales Reason                            | 46        |
| 2.2.5. Customers                               | 50        |
| 3.3. Recommendations for Strategy Development: | 54        |
| <b>IV. Predictive Model Building</b>           | <b>55</b> |
| Territory Performance Analysis                 | 55        |
| <b>V. Plan Orientation</b>                     | <b>57</b> |
| 1. Target Segmentation                         | 57        |
| 2. Orientation of sales & marketing plans      | 58        |
| <b>VI. Conclusion</b>                          | <b>60</b> |
| <b>Contribution</b>                            | <b>61</b> |
| <b>References</b>                              | <b>62</b> |

## **I. Introduction**

### **1. The AdventureWorks Database**

Adventure Works Cycles is a large multinational bicycle manufacturer, with headquarters located in Bothell, Washington. The company has approximately 300 employees, 29 of which are sales representatives. The primary distribution channel for Adventure Works Cycles through the retail stores of their resellers. These resellers are located in Australia, Canada, France, Germany, the United Kingdom, and the United States. Adventure Works Cycles also sells to individual customers worldwide by means of the Internet (Mitri, 2015).

The AdventureWorks database is a sample database developed by Microsoft with the purpose of helping users better understand Microsoft technologies and learn how to use them to develop applications. This database is designed for SQL Server, but can also be used with other database management systems such as Oracle, MySQL, and PostgreSQL. AdventureWorks is set in a fictional company called AdventureWorks Cycles, a large-scale bicycle manufacturing and sales company. This database contains tables, views, and procedures related to the activities of managing customers, products, orders, and other transactions, used to illustrate objects in processing tasks (OLTP). AdventureWorks database provides complex functionality such as data processing with transactions and combining data from various tables in the database. In addition, AdventureWorks also has notable features such as using data constraints to protect data integrity, using indexes to optimize data access performance, and using triggers to automatically perform actions when data is added, edited, or deleted. These features make AdventureWorks a very useful sample database for developers and database administrators. Datasets in AdventureWorks can be used for data analysis, visualization, and training of prediction and classification models. AdventureWorks is also used in many courses, certifications, and training in database administration, database design, and application development (Mitri, 2015).

### **2. Problem Overview**

The AdventureWorks database was chosen for analysis because it is designed to resemble real-world scenarios, making it a useful tool in practicing and learning database concepts. First, the AdventureWorks database has a comprehensive and well-defined database schema. It includes primary keys, foreign keys, relationships, and many different types of data attributes. Second, working with this database helps to understand how the different tables are related and how to design an effective database schema. Ultimately, it provides a standardized data set that can be used for benchmarking, comparison, and research purposes.

The problem with converting from Snowflake Schema to Star Schema is that Snowflake Schema has some disadvantages that need to be improved by converting to Star Schema. Snowflake schemas have additional levels of normalization and an increased number of tables that can complicate query execution and negatively impact performance. Joining multiple tables in a single query can become more complex and time-consuming, especially when dealing with large data sets. Additionally, the complex structure of the schema can make it more difficult for end users to navigate and understand the relationships between

tables and columns. To address these challenges and improve the overall efficiency of data storage and business analytics, organizations often make a profound transition from Snowflake Schema to Star Schema. This transformation involves merging multidimensional tables into denormalized dimension tables, reducing schema complexity and making it easier for end users to navigate and analyze data. It enables more intuitive and efficient data exploration, empowering end users to gain valuable insights quickly and efficiently. The simplified schema structure also enhances scalability and flexibility, allowing organizations to adapt to changing business needs and incorporate new data sources seamlessly.

- **Project Objectives**

The AdventureWorks project addresses inherent challenges faced by the Sales department, focusing on enhancing sales staff performance, understanding sales trends, and situating AdventureWorks within the broader industry context. These challenges necessitate a strategic approach encompassing data warehousing, analytics, and actionable recommendations.

The project unfolds through a meticulously structured sequence of phases, each contributing to the overarching objective of leveraging data for informed decision-making:

- Gain a thorough understanding of the AdventureWorks database, with a specific focus on its application within the sales department for this project.
- Acquaint oneself with essential tools such as BigQuery, Nifi, Airflow, and Locker Studio, instrumental in constructing the data warehouse and implementing various machine learning algorithms for predictive objectives.
- Develop a list of questions for the project, outlining which queries need answers and identifying metrics that will assist business managers in monitoring and fostering growth within the sales and marketing departments.
- Design a data staging layer in a star schema format based on the identified business questions that need resolution.
- Perform data mining activities, including visualization and predictive modeling, utilizing the Star schema staging layer.
- Derive insights from analytical and predictive models in the previous phase, focusing on formulating practical strategies tailored for the Sales and Marketing departments.
- Tailor recommendations to enhance sales performance, optimize market positioning, and leverage emerging market trends identified through comprehensive data analysis.

### **3. Key Questionnaire**

In line with our project design and focusing on the sales aspect of the AdventureWorks database, we propose the following key research questions:

#### **1. Sales Performance and Strategy:**

- What are the specific factors contributing to the difference in sales volume and profit margins between card types like Vista and ColonialVoice?
- How can the company further optimize pricing strategies to enhance profit margins while maintaining competitive sales volumes?
- What were the underlying causes of the peak in Average Order Value (AOV) in 2012, and how can these successful tactics be reapplied?

## **2. Market Expansion and Territory Performance:**

- What successful strategies have contributed to the strong sales performance in the Southwest and Canada, and how can these strategies be applied to other regions?
- What are the potential reasons for lower sales in territories like the United Kingdom and France, and how can targeted marketing initiatives improve their performance?

## **3. Salesperson Effectiveness:**

- What practices and strategies are employed by the top-performing salespersons that contribute to their consistent sales figures?
- How can the sales team share best practices and success stories to elevate the overall sales team's performance?

## **4. Sales Reason Analysis:**

- What are the specific drivers behind the "Other" sales reason category, and how can these factors be integrated into the sales strategy?
- How can marketing efforts be aligned to emphasize the impact of "Quality" and "Sponsorship" in driving sales?

## **5. Customer Segmentation and Retention:**

- How can personalized marketing strategies be developed to encourage more frequent transactions among "High spending infrequent customers"?
- What loyalty programs and personalized marketing approaches can be implemented to maintain and increase spending levels among "High spending frequent customers"?

## **6. Predictive Models for Customer Segmentation:**

- How can customers be effectively divided into the defined 5 groups (Low spending and Infrequent, High spending infrequent, High spending frequent, Medium Spending and Medium frequency, Medium Spending and low frequency), and what tailored marketing approaches can be designed for each group?

## **7. Territory Performance Prediction:**

- How can the logistic regression model for territory performance be further improved to enhance accuracy in classifying high and low-performing sales territories?
- What strategies can be developed based on the classification results to optimize resource allocation for each territory?

## **8. Target Segmentation for Campaigns:**

- How can the company target and engage customers more effectively in the Southwest, Canada, and the Northwest based on their spending habits and frequency?
- What marketing campaigns can be designed to cater specifically to customers in these high-performing regions?

## **4. Project Design**

### **Phase 1: Identification of Business Objectives**

In the initial phase of the project, we centered our attention on the Sales department of AdventureWorks. The primary goal was to define clear business objectives, research questions, and analytical directions that would drive the subsequent phases of the project.

### **Phase 2: Design and Construction of the Data Warehouse**

This phase involved the meticulous design and development of a Data Warehouse using a Star Schema approach. The schema was constructed to support our analytical objectives, consisting of:

- Fact Tables: Central tables like FactSales, capturing the essence of business transactions.
- Dimension Tables: Including DimSalesPerson, DimDate, DimCustomer, and others, these tables provided descriptive attributes to support multi-dimensional analysis.

The schema was designed to optimize query performance and provide a structured framework for complex analytical queries.

### **Phase 3: Data Extraction**

For data extraction, Apache NiFi was employed to source data from the Adventureworks dataset. This robust, automated data flow management tool enabled efficient and error-free data extraction, ensuring a seamless flow of data into the BigQuery environment.

### **Phase 4: Data Transformation and Loading**

In this critical phase, the operational data set (OLTP) was transformed into an analytical data set (OLAP) suitable for Business Intelligence analysis. Google BigQuery was chosen as the destination for this transformed data. The transformation process was managed using Apache Airflow, ensuring a scalable and manageable ETL pipeline. This setup allowed for regular data warehouse updates and maintenance with minimal manual intervention.

## Phase 5: Business Intelligence Application and Model Building

The transformed OLAP data in BigQuery was then connected to Looker Studio, enabling the creation of interactive and insightful visualizations. These visualizations provided a user-friendly interface for analyzing complex data sets, making the data accessible to non-technical stakeholders. In addition to descriptive analytics, Python's machine-learning libraries were utilized to develop predictive models. These models aimed to forecast future sales trends and identify potential market opportunities, contributing to more informed decision-making processes.

## Phase 6: Practical Implementations

Drawing insights from the analytical and predictive models in the previous phase, this final stage focused on formulating practical strategies for the Sales and Marketing departments. Recommendations were tailored to enhance sales performance, optimize market positioning, and leverage emerging market trends identified through the data analysis.

## II. Entity Relationship Diagram Design

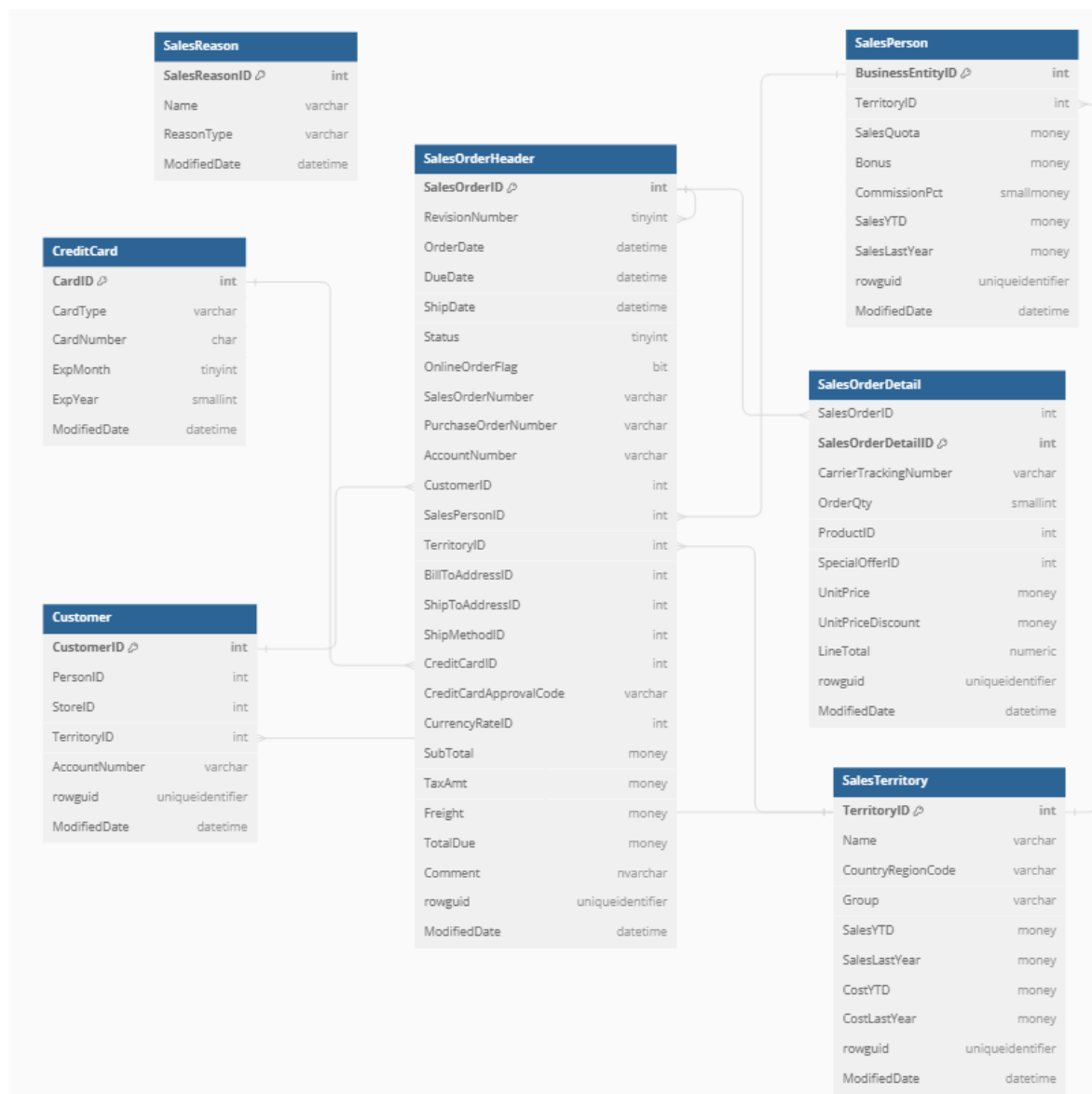
### 1. OLTP Database Design

The AdventureWorks dataset contains a complex schema that represents an online transaction processing (OLTP) system of a make-believe company. The supplied ERD illustrates the interplay existing between major business entities, explaining how sales, customer management, and order processing are dynamic in nature. Below is a brief on each table in the ERD and how they interplay:

- **SalesOrderHeader:** This is a master table to track all headers of the sales order and act as a point during interaction between orders and customers. It contains necessary details like order date, due date along with shipping details and foreign keys to Customer, SalesPerson, and SalesTerritory.
- **SalesOrderDetail:** Records the line items for each sales order, detailing the products ordered, the quantity, and the cost associated with each order line. It relates to SalesOrderHeader via the SalesOrderID.
- **SalesPerson:** SalesTerritory contains information of the sales representative which can be territory and quota. It has a relationship to SalesOrderHeader by SaleTerritoryKey through responsible SalesPerson key.
- **SalesTerritory:** Describes geographical sales territories. It is linked to SalesPerson to indicate which territory a salesperson is responsible for and to SalesOrderHeader through the territory assigned to each order.
- **Customer:** This stores information about the customers including their account number and contact. It is related to SalesOrderHeader indicating a customer per particular order.
- **SalesReason:** Enumerates reasons for discounts or other sales-related adjustments. It can be associated with SalesOrderHeader through a junction table (not visible on provided ERD) allowing multiple reasons to be applied against each sale.

- **CreditCard:** Stores credit card information used by customers. Linked to SalesOrderHeader to store payment methods of an order.

The most common one-to-many relationships in the repository, account, and sale item master files dominate Table 4 in ERD. The SalesOrderHeader table is at the core of the transaction system, linking customers, salespersons, and the details of their transactions. Foreign keys and indexes enforce referential integrity and maintain query performance - prerequisites for OLTP systems that are built to support business operations.



## 2. OLTP Database Code implementation

We run the following commands in order to pull the OLTP dataset from sql server:

```
Select * from AdventureWorks2014.Sales.SalesOrderHeader;
```

```
Select * from AdventureWorks2014.Sales.SalesPerson;
```

```
Select * from AdventureWorks2014.Sales.SalesReason;
```



Select \* from AdventureWorks2014.Sales.CreditCard;

Select \* from AdventureWorks2014.Sales.Customer;

Select \* from AdventureWorks2014.Sales.SalesOrderDetail;

Select \* from AdventureWorks2014.Sales.SalesTerritory;

### Airflow DAG:

```
import os
from datetime import timedelta
from airflow.models.dag import DAG
from airflow.operators.python import PythonOperator
from airflow.contrib.operators.bigquery_operator import BigQueryOperator
from airflow.providers.google.cloud.transfers.gcs_to_bigquery import GCSToBigQueryOperator
from airflow.utils.dates import days_ago
from airflow.operators.bash import BashOperator
from google.cloud import bigquery
import airflow
from clustering import train_model
from google.cloud import storage

os.environ['GOOGLE_APPLICATION_CREDENTIALS'] = '/opt/airflow/dags/datawarehouse-subject-680d0bd7c9df.json'

# DAG arguments
default_args = {
    'owner': 'airflow',
    'depends_on_past': False,
    'start_date': days_ago(1),
    'email_on_failure': False,
    'email_on_retry': False,
    'retries': None,
    'retry_delay': timedelta(minutes=2),
}

# DAG definition
dag = DAG(
    'oltp_to_olap_transform',
    default_args=default_args,
    description='Transform OLTP to OLAP schema in BigQuery',
    template_searchpath='/opt/airflow/dags/SQL_Queries',
    schedule_interval=None,
)
```

Import necessary modules and setting up DAG structure

```
def read_sql_file(file_path):
    with open(file_path, 'r') as file:
        sql_content = file.read()
        print(sql_content) # For debugging purposes
        return sql_content

DimDateQuery = read_sql_file('/opt/airflow/dags/SQL_Queries/DimDate.sql')
DimSalesPersonQuery = read_sql_file('/opt/airflow/dags/SQL_Queries/DimSalesPerson.sql')
DimSalesReasonQuery = read_sql_file('/opt/airflow/dags/SQL_Queries/DimSalesReason.sql')
DimTerritoryQuery = read_sql_file('/opt/airflow/dags/SQL_Queries/DimTerritory.sql')
DimCreditCardQuery = read_sql_file('/opt/airflow/dags/SQL_Queries/DimCreditCard.sql')
DimCustomerQuery = read_sql_file('/opt/airflow/dags/SQL_Queries/DimCustomer.sql')
FactSalesQuery = read_sql_file('/opt/airflow/dags/SQL_Queries/FactSales.sql')
```

read SQL scripts needed for OLAP transformation, these queries would be send to bigquery to execute

```
# Task to extract and transform DimDate
load_dim_date = BigQueryOperator(
    task_id='load_dim_date',
    sql= DimDateQuery,
    use_legacy_sql=False,
    write_disposition='WRITE_TRUNCATE', # Options: WRITE_TRUNCATE, WRITE_APPEND, WRITE_EMPTY
    create_disposition='CREATE_IF_NEEDED', # Options: CREATE_IF_NEEDED, CREATE_NEVER
    allow_large_results=True,
    dag=dag,
)
t1 = load_dim_date

# Task to extract and transform DimSalesPerson
load_dim_sales_person = BigQueryOperator(
    task_id='load_dim_sales_person',
    sql= DimSalesPersonQuery,
    use_legacy_sql=False,
    write_disposition='WRITE_TRUNCATE', # Options: WRITE_TRUNCATE, WRITE_APPEND, WRITE_EMPTY
    create_disposition='CREATE_IF_NEEDED', # Options: CREATE_IF_NEEDED, CREATE_NEVER
    allow_large_results=True,
    dag=dag,
)
t2 = load_dim_sales_person

# Task to extract and transform DimSalesReason
load_dim_sales_reason = BigQueryOperator(
    task_id='load_dim_sales_reason',
    sql= DimSalesReasonQuery,
    use_legacy_sql=False,
    write_disposition='WRITE_TRUNCATE', # Options: WRITE_TRUNCATE, WRITE_APPEND, WRITE_EMPTY
    create_disposition='CREATE_IF_NEEDED', # Options: CREATE_IF_NEEDED, CREATE_NEVER
    allow_large_results=True,
    dag=dag,
)
t3 = load_dim_sales_reason
```

```
load_dim_territory = BigQueryOperator(  
    task_id='load_dim_territory',  
    sql= DimTerritoryQuery,  
    use_legacy_sql=False,  
    write_disposition='WRITE_TRUNCATE', # Options: WRITE_TRUNCATE, WRITE_APPEND, WRITE_EMPTY  
    create_disposition='CREATE_IF_NEEDED', # Options: CREATE_IF_NEEDED, CREATE_NEVER  
    allow_large_results=True,  
    dag=dag,  
)  
t4 = load_dim_territory  
  
# Task to extract and transform FactSales  
load_fact_sales = BigQueryOperator(  
    task_id='load_fact_sales',  
    sql= FactSalesQuery,  
    use_legacy_sql=False,  
    write_disposition='WRITE_TRUNCATE', # Options: WRITE_TRUNCATE, WRITE_APPEND, WRITE_EMPTY  
    create_disposition='CREATE_IF_NEEDED', # Options: CREATE_IF_NEEDED, CREATE_NEVER  
    allow_large_results=True,  
    dag=dag,  
)  
t7 = load_fact_sales  
  
# Task to extract and transform DimCreditCard  
load_dim_credit_card = BigQueryOperator(  
    task_id='load_dim_credit_card',  
    sql= DimCreditCardQuery,  
    use_legacy_sql=False,  
    write_disposition='WRITE_TRUNCATE', # Options: WRITE_TRUNCATE, WRITE_APPEND, WRITE_EMPTY  
    create_disposition='CREATE_IF_NEEDED', # Options: CREATE_IF_NEEDED, CREATE_NEVER  
    allow_large_results=True,  
    dag=dag,  
)  
t5 = load_dim_credit_card  
  
# Task to extract and transform DimCustomer  
load_dim_customer = BigQueryOperator(  
    task_id='load_dim_customer',  
    sql= DimCustomerQuery,  
    use_legacy_sql=False,  
    write_disposition='WRITE_TRUNCATE', # Options: WRITE_TRUNCATE, WRITE_APPEND, WRITE_EMPTY  
    create_disposition='CREATE_IF_NEEDED', # Options: CREATE_IF_NEEDED, CREATE_NEVER  
    allow_large_results=True,  
    dag=dag,  
)  
t6 = load_dim_customer
```

create loading task for each table

```

# train model
def model_training():
    train_model()

modelTraining = PythonOperator(
    task_id='train_model',
    python_callable=model_training,
    dag=dag
)
t8 = modelTraining

TaskDelay = BashOperator(task_id="delay_bash_task",
                           dag=dag,
                           bash_command="sleep 5s")

def run_bigquery_sql():
    client = bigquery.Client()
    query = read_sql_file('/opt/airflow/dags/SQL_Queries/DimDate.sql')
    query_job = client.query(query)
    results = query_job.result() # Wait for the job to complete.

run_sql_task = PythonOperator(
    task_id='run_bigquery_sql',
    python_callable=run_bigquery_sql,
    dag=dag,
)
t0 = run_sql_task
# Set task dependencies
[t1,t2,t3,t4,t5,t6] >> t7 >> TaskDelay >> t8

```

Create machine learning task and setup dependencies

### III. Building the Data Warehouse

#### 1. Step 1: Choose the Business Process

In the context of AdventureWorks, the chosen business process centers around sales, specifically examining the performance and trends within the Sales department. The decision to focus on sales is rooted in the project's overarching objectives, aiming to derive actionable insights to enhance sales efficiency, understand market dynamics, and optimize strategic decision-making.

|  |
|--|
| AdventureWorks OLTP Bus Matrix for Sales |
|--|

|               | Customer | SalesPerson | Territory | Product | Date | SalesReason | CreditCard |
|---------------|----------|-------------|-----------|---------|------|-------------|------------|
| Sales Order   | X        | X           | X         | X       | X    |             | X          |
| Sales Quota   |          | X           | X         | X       |      |             |            |
| Sales Target  |          | X           | X         | X       |      |             |            |
| Sales Invoice | X        | X           | X         | X       | X    | X           | X          |

**In this bus matrix:**

- DimCustomer, DimSalesPerson, DimTerritory, DimProduct, DimDate, DimSalesReason, and DimCreditCard are dimensions.
- Sales Order, Sales Quota, Sales Target, and Sales Invoice are potential business processes.

**For each business process:**

- Sales Order involves customers, salespeople, territories, products, dates, and credit cards.
- Sales Quota is associated with salespeople, territories, products, and dates.
- Sales Target shares the same dimensions as the Sales Quota, involving salespeople, territories, products, and dates.
- Sales Invoice includes customers, salespeople, territories, products, dates, sales reasons, and credit cards.

**2. Step 2: Declare the Grain****Transaction Level (Atomic Level):**

For our AdventureWorks project, the chosen grain is the transaction level, epitomized by the FactSales fact table. At this finest granularity, each row within FactSales meticulously captures the details of an individual sales transaction. This approach aligns with the intrinsic nature of our primary OLTP data source, Sales.SalesOrderDetail, which serves as a transaction table, providing comprehensive insights into each sales order line item.

**Daily/Periodic Level:**

To facilitate analysis over specific time frames, our design incorporates the DimDate dimension. This enables aggregation at daily or periodic levels, allowing us to summarize

sales transactions over distinct time intervals. The DateKey attribute in DimDate acts as a linchpin for time-based analyses, supporting the extraction of meaningful patterns and trends.

### Summary Level:

Elevating our perspective for broader business insights, especially in monthly or yearly contexts, involves leveraging additional dimensions such as DimCustomer, DimSalesPerson, DimTerritory, DimProduct, DimSalesReason, and DimCreditCard. Aggregating data based on these dimensions provides a higher-level view of sales performance across key business aspects. This summary-level granularity proves invaluable for strategic decision-making and long-term planning.

The rationale behind implementing a **transaction fact table** stems from the inherent characteristics of our underlying data source, **Sales.SalesOrderDetail**. By selecting this granularity, we opt for the lowest level of detail, capturing individual sales order line items. This strategic choice not only aligns with best practices but also positions us for flexibility in subsequent analyses and reporting. In essence, our data warehouse is designed to seamlessly transition from the intricacies of individual transactions to the broader perspectives essential for strategic business insights.

### 3. Step 3: Identify the Dimensions

#### Dimension Table: DimSalesReason

- Attributes:
  - SalesReasonID: A unique identifier for the reason behind a sale.
  - Name: The descriptive name of the sales reason.
  - ReasonType: The category or type of the sales reason.
- Purpose: This table captures information related to sales reasons, which can provide insights into the factors driving sales.

#### Code implementation:

```
CREATE OR REPLACE TABLE OLAP_demo.DimSalesReason AS
SELECT
    SalesReasonID,
    Name,
    ReasonType
FROM
    OLTP.SalesReason;
```

**Dimension Table: DimSalesPerson**

- Attributes:
  - BusinessEntityID: A unique identifier for each salesperson.
  - TerritoryID: The identifier of the sales territory associated with the salesperson.
  - SalesQuota: The target sales quota for the salesperson.
  - Bonus: The bonus amount received by the salesperson.
  - CommissionPct: The commission percentage earned by the salesperson.
  - SalesYTD: The total sales accumulated by the salesperson year-to-date.
  - SalesLastYear: The total sales for the previous year.
- Purpose: This table provides detailed information about salespeople and their performance metrics.

**Code implementation:**

```
CREATE OR REPLACE TABLE OLAP_demo.DimSalesPerson AS
SELECT
    BusinessEntityID,
    TerritoryID,
    SalesQuota,
    Bonus,
    CommissionPct,
    SalesYTD,
    SalesLastYear
FROM
    OLTP.SalesPerson;
```

**Dimension Table: DimDate**

- Attributes:
  - DateKey: A surrogate primary key for the date dimension, formatted as YYYYMMDD.

- Date: The actual date value.
- Day: The day component of the date.
- Month: The month component of the date.
- Quarter: The quarter component of the date.
- Year: The year component of the date.
- Purpose: The DimDate table serves as a time dimension, enabling time-based analysis and reporting.

### Code implementation:

```

SELECT
    PARSE_DATETIME('%Y-%m-%d %H:%M:%E*S', '2011-05-31 00:00:00.000') AS
    parsed_datetime;

-- Table creation query

CREATE OR REPLACE TABLE OLAP_demo.DimDate AS

SELECT
    DISTINCT FORMAT_DATE('%Y%m%d', PARSE_DATETIME('%Y-%m-%d
    %H:%M:%E*S', OrderDate)) AS DateKey,
    CAST(PARSE_DATETIME('%Y-%m-%d %H:%M:%E*S', OrderDate) AS DATE)
    AS Date,
    EXTRACT(DAY FROM PARSE_DATETIME('%Y-%m-%d %H:%M:%E*S',
    OrderDate)) AS Day,
    EXTRACT(MONTH FROM PARSE_DATETIME('%Y-%m-%d %H:%M:%E*S',
    OrderDate)) AS Month,
    EXTRACT(QUARTER FROM PARSE_DATETIME('%Y-%m-%d %H:%M:%E*S',
    OrderDate)) AS Quarter,
    EXTRACT(YEAR FROM PARSE_DATETIME('%Y-%m-%d %H:%M:%E*S',
    OrderDate)) AS Year
FROM
    OLTP.SalesOrderHeader;

```

### Dimension Table: DimCustomer

- Attributes:



- CustomerID: The unique identifier for customers.
- PersonID: An identifier that may relate to a detailed person dimension or table.
- StoreID: Identifier of the store where the customer made transactions.
- AccountNumber: A unique number representing the customer's account.
- Purpose: This dimension table provides insights into customer profiles, their relationships, and transaction history.

**Code implementation:**

```
CREATE OR REPLACE TABLE OLAP_demo.DimCustomer AS
SELECT
    CustomerID,
    PersonID,
    StoreID,
    AccountNumber
FROM
    OLTP.Customer;
```

**Dimension Table: DimCreditCard**

- Attributes:
  - TerritoryID: A unique identifier for each sales territory.
  - Name: The name of the sales territory.
  - CountryRegionCode: Code representing the country or region associated with the territory.
  - TerritoryGroup: A grouping or category assigned to the territory (as represented by the 'Group' column).
  - SalesYTD: Year-to-date sales amount for the territory.
  - SalesLastYear: Sales amount for the last year in the territory.
  - CostYTD: Year-to-date cost incurred in the territory.
  - CostLastYear: Cost incurred in the last year in the territory.

- Purpose: Capturing and organizing information related to sales territories within the AdventureWorks data warehouse. Each row in this table represents a unique sales territory, and its attributes provide valuable details for analytical purposes

**Code implementation:**

```
CREATE OR REPLACE TABLE OLAP_demo.DimTerritory AS
SELECT
    TerritoryID,
    Name,
    CountryRegionCode,
    `Group` AS TerritoryGroup, -- Assign a name to the fourth column
    SalesYTD,
    SalesLastYear,
    CostYTD,
    CostLastYear
FROM
    OLTP.SalesTerritory;
```

**Dimension Table: DimCreditCard**

- Attributes:
  - CreditCardID: A surrogate key for credit card entries.
  - CardType: The type of credit card (e.g., Visa, MasterCard).
  - CardNumber1: A partially masked or truncated credit card number.
- Purpose: This table captures information related to credit card transactions, aiding in financial analysis and payment method tracking.

**Code implementation:**

```
CREATE OR REPLACE TABLE OLAP_demo.DimCreditCard AS
SELECT
    CreditCardID,
```

```
CardType,  
  
CardNumber1  
  
FROM  
  
OLTP.CreditCard;
```

#### 4. Step 4: Identify the Facts

The FactSales table is designed to store detailed transactional data at a granular level, providing insights into individual sales orders. This level of detail is essential for performing in-depth analysis and generating various reports. The Fact Table is the cornerstone of the AdventureWorks data warehouse, allowing users to gain valuable business insights and make informed decisions based on sales data.

##### FactSales Table Attributes:

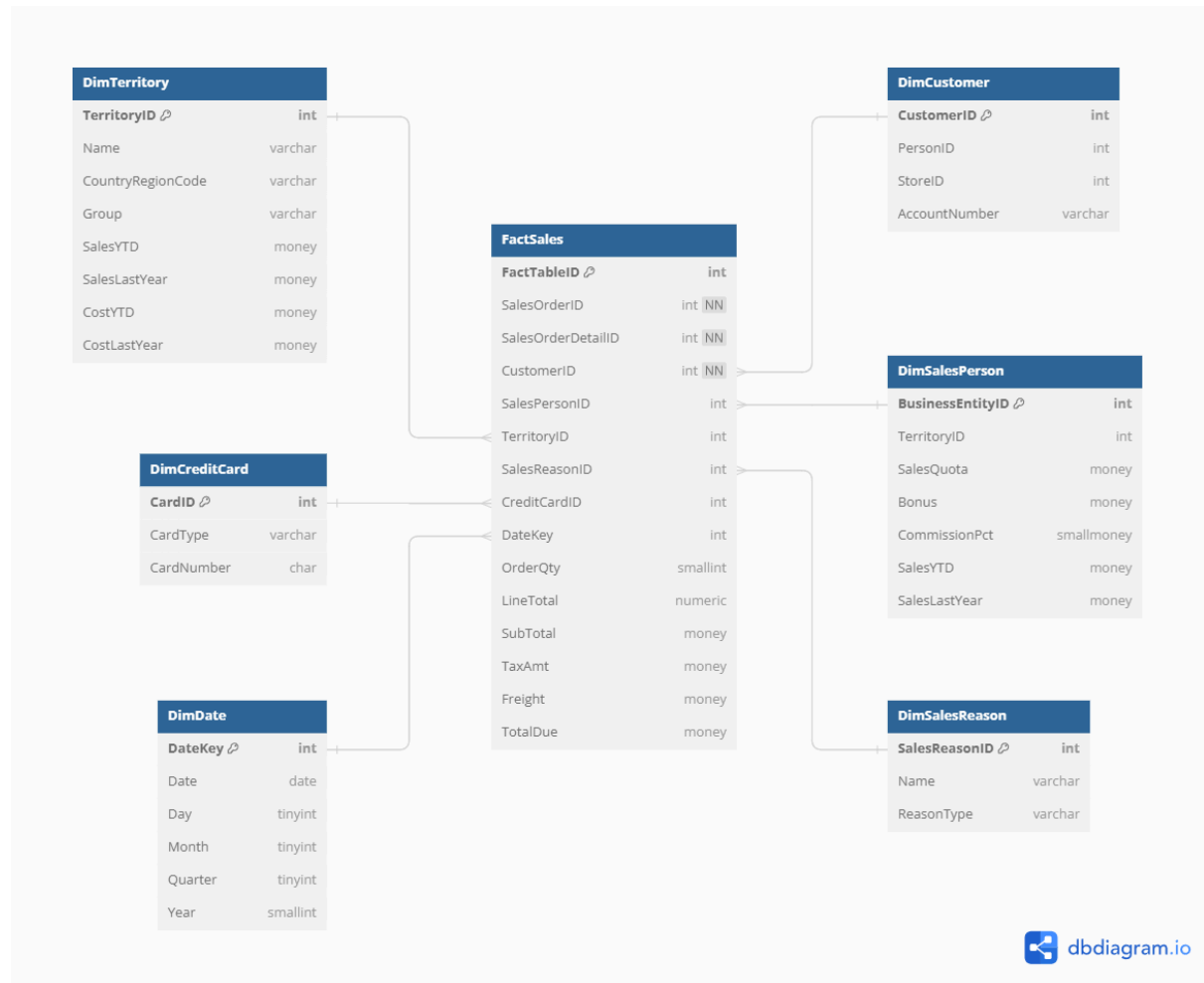
- FactTableID: A generated unique identifier for each row in the Fact Table.
- SalesOrderID: A unique identifier for each sales order.
- SalesOrderDetailID: A unique identifier for each sales order detail.
- CustomerID: The identifier of the customer associated with the sales order.
- SalesPersonID: The identifier of the salesperson responsible for the sale.
- TerritoryID: The identifier of the sales territory where the sale occurred.
- SalesReasonID: The identifier for the reason behind the sale (linked to the DimSalesReason dimension).
- CreditCardID: The identifier for the credit card used for payment (linked to the DimCreditCard dimension).
- DateKey: A formatted date key representing the order date (linked to the DimDate dimension).
- OrderQty: The quantity of items ordered.
- LineTotal: The total value of the order line.
- SubTotal: The subtotal of the sales order.
- TaxAmt: The tax amount for the sales order.
- Freight: The freight or shipping cost for the sales order.
- TotalDue: The total amount due for the sales order.

**Code implementation:**

```
CREATE OR REPLACE TABLE OLAP_demo.FactSales AS
SELECT
    ROW_NUMBER() OVER() AS FactTableID,
    soh.SalesOrderID,
    sod.SalesOrderDetailID,
    soh.CustomerID,
    soh.SalesPersonID,
    soh.TerritoryID,
    sr.SalesReasonID,
    soh.CreditCardID,
    FORMAT_DATE('%Y%m%d', DATE(soh.OrderDate)) AS DateKey,
    sod.OrderQty,
    sod.LineTotal,
    soh.SubTotal,
    soh.TaxAmt,
    soh.Freight,
    soh.TotalDue
FROM
    OLTP.SalesOrderHeader AS soh
JOIN
    OLTP.SalesOrderDetail AS sod
ON
    soh.SalesOrderID = sod.SalesOrderID
LEFT JOIN
    OLTP.SalesReason AS sr
ON
```

```
sr.SalesReasonID = sr.SalesReasonID;
```

### The ERD of OLAP:



## III. Results & Implications

### 1. Querying On OLAP

Following the creation of the OLAP database, we proceed to leverage BigQuery for formulating analytical queries aimed at extracting valuable insights.

--What are the total sales amounts for each sales reason?

SELECT

dsr.SalesReasonID,

dsr.Name AS SalesReason,

SUM(fs.TotalDue) AS TotalSalesAmount

```

FROM

  OLAP.FactSales AS fs

JOIN

  OLAP.DimSalesReason AS dsr

ON

  fs.SalesReasonID = dsr.SalesReasonID

GROUP BY

  SalesReasonID, SalesReason;

```

## Query results

[SAVE RESULTS](#)
[EXPLORE DATA](#)


| JOB INFORMATION |               | RESULTS                  | CHART             | PREVIEW | JSON | EXECUTION DETAILS | EXECUTION GRAPH |
|-----------------|---------------|--------------------------|-------------------|---------|------|-------------------|-----------------|
| Row             | SalesReasonID | SalesReason              | TotalSalesAmount  |         |      |                   |                 |
| 1               | 6             | Review                   | 2926970121.727... |         |      |                   |                 |
| 2               | 10            | Other                    | 2926970121.727... |         |      |                   |                 |
| 3               | 9             | Quality                  | 2926970121.727... |         |      |                   |                 |
| 4               | 1             | Price                    | 2926970121.727... |         |      |                   |                 |
| 5               | 8             | Sponsorship              | 2926970121.727... |         |      |                   |                 |
| 6               | 2             | On Promotion             | 2926970121.727... |         |      |                   |                 |
| 7               | 4             | Television Advertisement | 2926970121.727... |         |      |                   |                 |
| 8               | 3             | Magazine Advertisement   | 2926970121.727... |         |      |                   |                 |
| 9               | 7             | Demo Event               | 2926970121.727... |         |      |                   |                 |
| 10              | 5             | Manufacturer             | 2926970121.727... |         |      |                   |                 |

--Q2: Who are the top 5 customers by sales amount in the most recent year?

```

WITH YearlySales AS (

  SELECT

    FS.CustomerID,

    DD.Year,

    SUM(FS.LineTotal) AS TotalSales

  FROM

    OLAP.FactSales FS

  JOIN

    OLAP.DimDate DD ON FS.DateKey = DD.DateKey

  GROUP BY

```

```

    FS.CustomerID, DD.Year
),
MaxYear AS (
    SELECT
        MAX(Year) AS MaxYear
    FROM
        YearlySales
)
SELECT
    YS.CustomerID,
    YS.TotalSales
FROM
    YearlySales YS
JOIN
    MaxYear MY ON YS.Year = MY.MaxYear
ORDER BY
    YS.TotalSales DESC
LIMIT 5;

```

| Row | CustomerID | TotalSales        |
|-----|------------|-------------------|
| 1   | 29736      | 1577006.041059... |
| 2   | 29770      | 1518249.945015... |
| 3   | 29629      | 1492166.283836... |
| 4   | 29715      | 1391746.306953... |
| 5   | 29701      | 1388704.610385... |

--Which territory has the highest total sales in the last six months?

```

SELECT

```

```

dt.TerritoryID,
dt.Name AS TerritoryName,
SUM(fs.TotalDue) AS TotalSalesAmount
FROM
  OLAP.FactSales AS fs
JOIN
  OLAP.DimTerritory AS dt
ON
  fs.TerritoryID = dt.TerritoryID
JOIN
  OLAP.DimDate AS d
ON
  fs.DateKey = d.DateKey
WHERE
  d.Month >= EXTRACT(MONTH FROM CURRENT_DATE()) - 6
GROUP BY
  TerritoryID, TerritoryName
ORDER BY
  TotalSalesAmount DESC
LIMIT 1;

```

| Row | TerritoryID | TerritoryName | TotalSalesAmount  |
|-----|-------------|---------------|-------------------|
| 1   | 4           | Southwest     | 6968966260.963... |

--How much has the sales grown or declined year-over-year for each month?

```

WITH SalesByYear AS (
  SELECT
    DT.Name AS Territory,
    DD.Year,

```



```

SUM(FS.LineTotal) AS TotalSales
FROM
    OLAP.FactSales FS
JOIN
    OLAP.DimDate DD ON FS.DateKey = DD.DateKey
JOIN
    OLAP.DimTerritory DT ON FS.TerritoryID = DT.TerritoryID
GROUP BY
    Territory, Year
)

SELECT
    CurrentYear.Territory,
    CurrentYear.Year AS CurrentYear,
    PreviousYear.Year AS PreviousYear,
    CurrentYear.TotalSales AS SalesCurrentYear,
    PreviousYear.TotalSales AS SalesPreviousYear,
    ((CurrentYear.TotalSales - PreviousYear.TotalSales) / PreviousYear.TotalSales) * 100 AS
GrowthPercentage
FROM
    SalesByYear CurrentYear
JOIN
    SalesByYear PreviousYear ON CurrentYear.Territory = PreviousYear.Territory AND
CurrentYear.Year = PreviousYear.Year + 1;

```

| JOB INFORMATION |                | RESULTS     | CHART        | PREVIEW           | JSON              | EXECUTION DETAILS | EXECUTION GRAPH |
|-----------------|----------------|-------------|--------------|-------------------|-------------------|-------------------|-----------------|
| Row             | Territory      | CurrentYear | PreviousYear | SalesCurrentYear  | SalesPreviousYear | GrowthPercentage  |                 |
| 1               | Southwest      | 2013        | 2012         | 91165403.12219... | 82888538.81002... | 9.985535311637... |                 |
| 2               | United Kingdom | 2013        | 2012         | 36334225.52211... | 15818557.41757... | 129.6936728360... |                 |
| 3               | United Kingdom | 2012        | 2011         | 15818557.41757... | 3628886.190185... | 335.9066828924... |                 |
| 4               | Southeast      | 2012        | 2011         | 29637134.60041... | 16403900.00401... | 80.67127081457... |                 |
| 5               | Southeast      | 2013        | 2012         | 23999473.74364... | 29637134.60041... | -19.0222871838... |                 |
| 6               | Australia      | 2013        | 2012         | 42306643.62650... | 21247831.95617... | 99.11040201070... |                 |
| 7               | Central        | 2013        | 2012         | 29942253.80877... | 29585582.82437... | 1.205556728480... |                 |
| 8               | Southwest      | 2014        | 2013         | 39718980.20467... | 91165403.12219... | -56.4319590059... |                 |
| 9               | Northwest      | 2014        | 2013         | 29977507.21925... | 60151747.46845... | -50.1635306023... |                 |
| 10              | United Kingdom | 2014        | 2013         | 20925541.40435... | 36334225.52211... | -42.4081809818... |                 |

--What is the average order value (AOV) trend over time?

SELECT

DD.Year,

DD.Month,

AVG(FS.TotalDue) AS AverageOrderValue

FROM

OLAP.FactSales FS

JOIN

OLAP.DimDate DD ON FS.DateKey = DD.DateKey

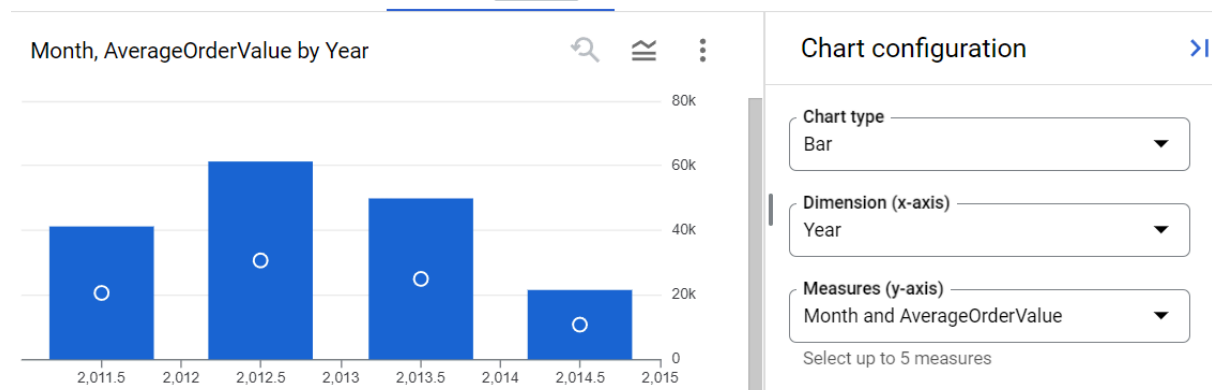
GROUP BY

DD.Year, DD.Month

ORDER BY

DD.Year, DD.Month;

| Row | Year | Month | AverageOrderValue |
|-----|------|-------|-------------------|
| 1   | 2011 | 5     | 22674.98636693... |
| 2   | 2011 | 6     | 3596.428777600... |
| 3   | 2011 | 7     | 27716.70654476... |
| 4   | 2011 | 8     | 30603.83222696... |
| 5   | 2011 | 9     | 3533.704428508... |
| 6   | 2011 | 10    | 41161.72571730... |
| 7   | 2011 | 11    | 3544.839113716... |
| 8   | 2011 | 12    | 19434.17372441... |



## #Sales Performance Compared to Sales Quota

--Question: How does each salesperson's performance compare to their sales quota?

SELECT

DSP.BusinessEntityID,

DSP.SalesQuota,

SUM(FS.LineTotal) AS TotalSales,

CASE

WHEN SUM(FS.LineTotal) >= DSP.SalesQuota THEN 'Met or Exceeded'

ELSE 'Below'

END AS QuotaStatus

FROM

OLAP.FactSales FS

JOIN

OLAP.DimSalesPerson DSP **ON** FS.SalesPersonID = DSP.BusinessEntityID

**GROUP BY**

DSP.BusinessEntityID, DSP.SalesQuota;

| Row | BusinessEntityID | SalesQuota | TotalSales        | QuotaStatus     |
|-----|------------------|------------|-------------------|-----------------|
| 1   | 276              | 250000.0   | 103670074.0678... | Met or Exceeded |
| 2   | 277              | 250000.0   | 100658035.0834... | Met or Exceeded |
| 3   | 275              | 300000.0   | 92939029.83386... | Met or Exceeded |
| 4   | 289              | 250000.0   | 85033386.32646... | Met or Exceeded |
| 5   | 279              | 300000.0   | 71710127.28969... | Met or Exceeded |

--Question: What is the contribution of each sales territory to the total sales?

**WITH** TotalSales **AS** (

**SELECT**

SUM(LineTotal) **AS** Total

**FROM**

OLAP.FactSales

)

**SELECT**

DT.Name **AS** TerritoryName,

SUM(FS.LineTotal) **AS** TerritorySales,

SUM(FS.LineTotal) / TotalSales.Total \* 100 **AS** ContributionPercentage

**FROM**

OLAP.FactSales FS

**JOIN**

OLAP.DimTerritory DT **ON** FS.TerritoryID = DT.TerritoryID

**CROSS JOIN**

TotalSales

**GROUP BY**

DT.Name, TotalSales.Total;

| Row | TerritoryName ▼ | TerritorySales ▼  | ContributionPercentage |
|-----|-----------------|-------------------|------------------------|
| 1   | Southwest       | 241846095.7745... | 22.01675581915...      |
| 2   | Canada          | 163557704.4042... | 14.88967613339...      |
| 3   | Northwest       | 160849425.5748... | 14.64312464995...      |
| 4   | Australia       | 106553360.3461... | 9.700215787801...      |
| 5   | Central         | 79090089.84816... | 7.200063289527...      |
| 6   | Southeast       | 78796550.502882   | 7.173340575366...      |

--Question: How effective are different sales reasons in generating revenue?

SELECT

DSR.Name AS SalesReason,

COUNT(DISTINCT FS.SalesOrderID) AS NumberOfOrders,

SUM(FS.LineTotal) AS TotalSales

FROM

OLAP.FactSales FS

JOIN

OLAP.DimSalesReason DSR ON FS.SalesReasonID = DSR.SalesReasonID

GROUP BY

DSR.Name;

| Row | SalesReason ▼            | NumberOfOrders ▼ | TotalSales ▼      |
|-----|--------------------------|------------------|-------------------|
| 1   | On Promotion             | 31465            | 109846381.4383... |
| 2   | Magazine Advertisement   | 31465            | 109846381.4383... |
| 3   | Television Advertisement | 31465            | 109846381.4383... |
| 4   | Sponsorship              | 31465            | 109846381.4383... |
| 5   | Demo Event               | 31465            | 109846381.4383... |
| 6   | Manufacturer             | 31465            | 109846381.4383... |
| 7   | Other                    | 31465            | 109846381.4383... |
| 8   | Quality                  | 31465            | 109846381.4383... |
| 9   | Review                   | 31465            | 109846381.4383... |
| 10  | Price                    | 31465            | 109846381.4383... |

## 2. Customer Segmentation

### a. Algorithm:

K-Means Clustering: The K-Means clustering algorithm was selected because it is good at handling large datasets and excels at partitioning data into distinctive, non-overlapping groups. The K-Means algorithm is a centroid-based approach that is recognized for its

simplicity and fast convergence speed. This algorithm divides the data into K clusters, with each data point belonging to a cluster whose mean is the closest, serving as the prototype of the cluster. This capability is particularly advantageous for customer segmentation, as it aids in classifying customers into clusters based on shared similarities in characteristics, which is instrumental in understanding customer behaviors and preferences (Jain, 2010).

In this study, K-Means is used to segment customers based on their purchase patterns, represented by features such as Order Count, Total Spending, and Average Order Value. These features are indicative of customer purchasing behavior and hence are appropriate for segmentation using K-Means (Chaturvedi et al., 2001).

#### b. Parameters

**Determining the Number of Clusters (n\_clusters):** The most important factor in K-Means clustering is the number of clusters (n\_clusters). In this study, we initially chose five clusters based on preliminary analysis and domain knowledge. To refine this choice, we used the Elbow Method. This involves plotting the cost function (sum of squared distances of samples to their nearest cluster center) against the number of clusters and identifying the 'elbow point' where the rate of decrease sharply changes. This point is a key indicator of the optimal number of clusters for the data (Thorndike, 1953).

The Elbow Method is popular for its simplicity and effectiveness, but it can be ambiguous in pinpointing the exact elbow point. Therefore, we recommend using it alongside domain knowledge and other validation techniques like the Silhouette Method or Gap Statistics for a more thorough analysis (Rousseeuw, 1987; Tibshirani et al., 2001).

#### c. Output expectation

Customers should be divided into 5 groups:

Group 1: Low spending and Infrequent customers (cluster = 0)

- This group consists of customers who don't create a large number of orders, and their purchases are low in total price
- These are the customers that yield the least value to the business, but they are a good candidate for research to refine marketing quality,

Group 2: High spending infrequent customers (cluster = 1)

- This group consists of customers who do not visit often, but their average order value is greater than most customers.
- These customers are like gold, valuable but scarce, with proper customer care, sales team could potentially improve their number of orders.

Group 3: High spending frequent customers (cluster = 2)

- This group consists of customers who create a large amount of orders, and they usually order a large sum of products.
- These are your dream customers, this group of customers is extremely valuable to retain.

#### Group 4: Medium Spending and Medium frequency (cluster = 3)

- These are your average customer group, with medium occasions that they visit the business, with a moderate order bill.
- This group would be a good group to act as a base group when conducting researches

#### Group 5: Medium Spending and low frequency (cluster = 4)

- These are customers who rarely visit the store and don't spend too much on the order.
- This group lies between group 4 and group 1, while does not offer clear characteristic of the customers, but still a significant contribution toward revenue and should not be left out in research.

### 3. Exploratory Data Analysis

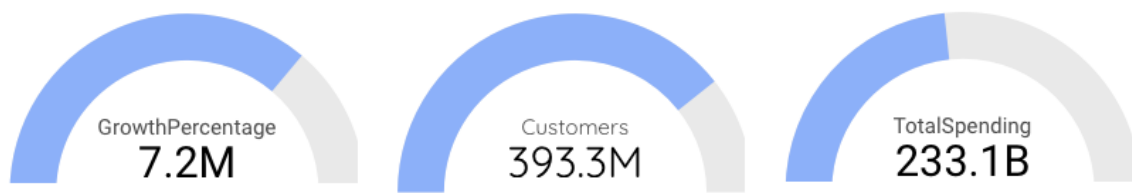
This Exploratory Data Analysis aims to uncover insights from our OLAP data warehouse, focusing on customer transactions, sales reasons, and financial metrics. Using visualizations created in Locker Studio, we have identified patterns and trends that will inform our strategic decision-making process.

#### 3.1. Data Overview:

Our data is about the context of the AdventureWorks database, serving as a comprehensive dataset for analyzing business operations. This file contains structured data typical for an OLAP system, including various dimension tables like DimCustomer, DimSalesPerson, DimCreditCard, and others, along with a central fact table FactSales. These tables collectively provide a detailed view of business elements such as customer information, salesperson performance, credit card transactions, and sales data. The data encapsulated within these tables is instrumental for understanding complex business scenarios and is particularly useful for database practices like schema design and query optimization. This dataset supports the analysis of transitioning from a Snowflake Schema, characterized by its multiple levels of normalization and complexity, to a Star Schema, which aims for simplification and efficiency in handling large-scale business data, particularly in enhancing query performance and easing user navigation through the database.

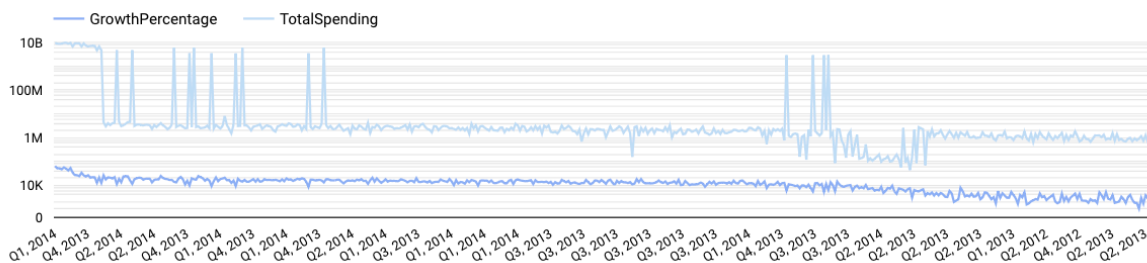


### 3.2. Key Findings:



The image shows a set of three semi-circular progress bars, each representing different metrics:

1. Growth Percentage: Marked at 7.2M, which might indicate a 7.2 million increase.
2. Customers: This is quantified at 393.3M, suggesting there are 393.3 million customers.
3. Total Spending: the total spending is 233.1 billion

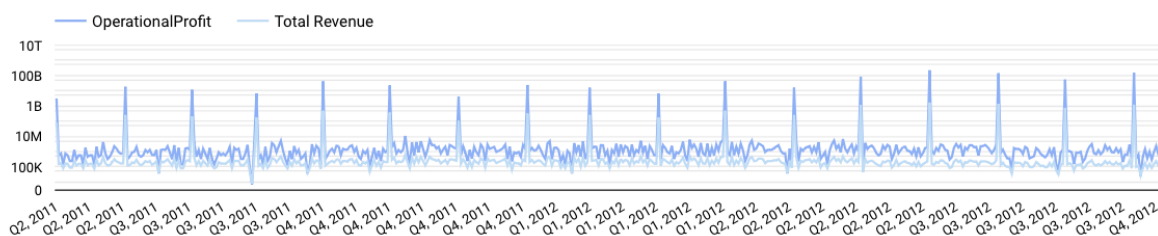


This is time series line graph comparing Operational Profit and Total Revenue over multiple quarters, starting from Q2 2011 through 2014.

The Operational Profit = TotalDue - (TaxAmt + Freight)

Operational Profit and Total Revenue are plotted as two different lines on the graph, with the Total Revenue line consistently above the Operational Profit, which is expected as total revenue is typically higher than profit due to the deduction of operational costs.

The x-axis shows the time progression in quarters (Q1, Q2, Q3, Q4) for each year. The y-axis is logarithmic, with values ranging from 0 to 10 trillion (10T)



The image shows a line graph with two variables: Growth Percentage and Total Spending. The data spans from 2011 to 2014.

On the y-axis, there are numerical values that scale logarithmically from 0 to 10 billion (10B), which suggests that the values being represented are quite large. The graph uses two different lines to distinguish between the two variables:

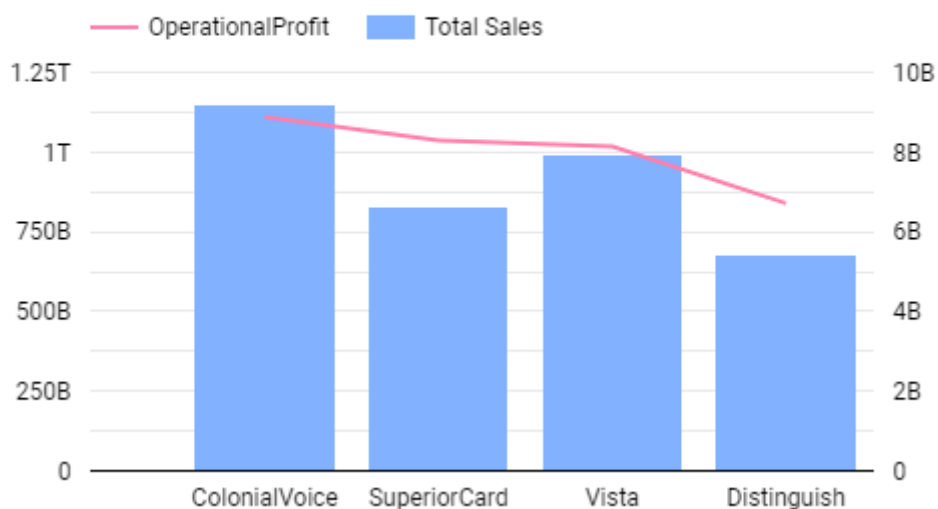
- The line for Growth Percentage is quite volatile with sharp spikes, indicating periods of significant change. This could reflect times of either rapid growth or decline.
- The line for Total Spending appears to be more stable and consistent over the observed period, with fewer and smaller spikes compared to Growth Percentage.

The x-axis represents time, segmented into quarters of each year. This type of graph is often used to visualize financial data over time, allowing for the identification of trends, patterns, and potential correlations between the two variables.

From the graph, it can be inferred that while Total Spending remains relatively consistent, the Growth Percentage experiences significant fluctuations. These could be indicative of underlying business cycles, market conditions, or specific events impacting growth during those periods. Analyzing such a graph could help in strategic planning and decision-making by highlighting when an entity is most likely to experience growth or contraction.

### 3.2.1. Product

#### Total Sales vs Operational Profit



**Total Sales and Operational Profit:** The bar and line chart provided compares the total sales and operational profit for four different card types: ColonialVoice, SuperiorCard, Vista, and

Distinguish. The chart shows total sales in bars and operational profit as a line, indicating the profitability efficiency for each card type relative to their sales volume.

### Insights:

ColonialVoice and SuperiorCard have similar sales figures, but ColonialVoice exhibits a higher operational profit, suggesting more efficient operations or a higher-margin product offering.

Vista, while having lower total sales than ColonialVoice and SuperiorCard, has an operational profit that is relatively close to SuperiorCard, indicating effective cost management or a lucrative product mix.

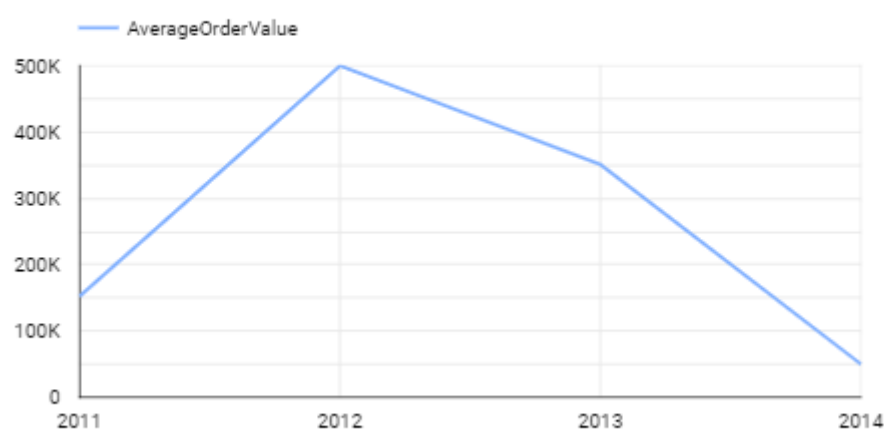
Distinguish has the lowest total sales and operational profit, which may highlight challenges in market penetration or operational efficiency.

### Interpretation:

The disparity between sales and operational profit across card types could indicate varying strategies, cost structures, and market positioning.

Focusing on operational efficiencies, cost reduction strategies, and potentially adjusting pricing could be areas of improvement for card types with lower operational profits relative to their sales.

AOV trend over time



**AOV Trend Over Time:** The line graph displays the changes in Average Order Value (AOV) for a particular card type across four years, revealing significant shifts in customer spending behavior over time.

### Insights:

A peak in AOV is observed in 2012, suggesting a period where customers spent more per transaction, which could indicate successful marketing strategies or favorable market conditions at that time.

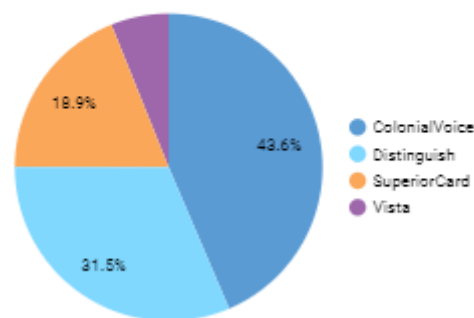
The following years, 2013 and 2014, show a marked decline in AOV, signaling a shift in consumer spending habits or potentially changes in product pricing or availability.

### **Interpretation:**

The spike in AOV for the year 2012 could be worth investigating to identify the underlying causes, such as specific campaigns or market trends, that could be leveraged again.

The decreasing trend post-2012 may require analysis of pricing strategies, product mix changes, or external economic factors that may have influenced spending.

**Total Sales by Card Type**



**Total Sales by Card Type:** The pie chart details the market share distribution of total sales across four card types: ColonialVoice, Vista, SuperiorCard, and Distinguish.

### **Insights:**

ColonialVoice captures the largest segment at 43.6%, indicating a strong market preference or a successful sales strategy.

Vista follows with 31.5% of the total sales, suggesting a competitive position in the market.

SuperiorCard holds 18.9%, while Distinguish has the smallest share at 6.1%, pointing to areas where these card types could potentially increase their market presence.

### **Interpretation:**

The dominance of ColonialVoice and Vista suggests these brands have effectively met customer needs or have a broad acceptance among consumers.

The market share held by SuperiorCard and Distinguish may reflect niche consumer groups or opportunities for market expansion.

## Total Sales by Card Type vs Country

|               |                  |                 |                 |                 |                 |                 | Name / TotalDue |
|---------------|------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| CardType      | Southwest        | Canada          | Northwest       | Central         | Northeast       | Southeast       |                 |
| ColonialVoice | 1,594,971,097... | 2,722,295,39... | 1,483,242,32... | 540,642,929.3   | 496,224,207.... | 330,549,486.... | 61              |
| Vista         | 2,236,957,683... | 898,734,426.... | 1,121,093,10... | 1,214,843,68... | 969,642,210.... | 412,464,148.... | 3               |
| SuperiorCard  | 1,779,572,311... | 626,723,782.... | 898,752,788.... | 308,001,962.... | 709,146,797.7   | 705,399,486.... | 63              |
| Distinguish   | 1,352,165,000... | 1,020,196,12... | 606,867,399.... | 567,479,986.... | 362,699,733.... | 771,694,567.... | 33              |

**Total Sales by Card Type vs Country:** The table maps total sales figures for each card type across various countries, highlighting geographical strengths and market preferences.

### Insights:

ColonialVoice and Vista have substantial sales in the Southwest and Southeast, respectively, suggesting strong regional market dominance.

SuperiorCard and Distinguish present a more even distribution but with lower sales figures, indicating room for growth and market penetration.

### Interpretation:

The regional sales performance could guide targeted marketing efforts and strategic planning to capitalize on existing strengths and address areas with lower sales.

An analysis of regional market trends could provide further insights into consumer preferences and competitive dynamics in these areas.

| CardType         | Sales Quantity | Median       | Total Shipping Cost | Max Price     | Mean         | Min Price    | Total Tax      |
|------------------|----------------|--------------|---------------------|---------------|--------------|--------------|----------------|
| 1. Vista         | 736,690        | 6,817,860.05 | 213,900,662.25      | 10,273,210.05 | 6,988,805.4  | 5,337,600.95 | 684,482,113.14 |
| 2. SuperiorCard  | 637,420        | 6,622,245.18 | 177,969,226.41      | 10,011,326.6  | 6,837,481.89 | 5,279,905.66 | 569,501,512.34 |
| 3. Distinguish   | 541,110        | 6,732,554.76 | 145,500,380.99      | 9,744,901.91  | 6,854,680.19 | 5,592,524.13 | 465,601,206.82 |
| 4. ColonialVoice | 804,990        | 7,798,761.06 | 247,751,461.63      | 11,676,794.22 | 7,965,241.52 | 6,127,598.17 | 792,804,665.7  |

1 - 4 / 4 < >

**Sales Quantity and Financials by Card Type:** A detailed table presents a breakdown of sales quantity and associated financial metrics, such as shipping costs and taxes, for each card type.

### Insights:

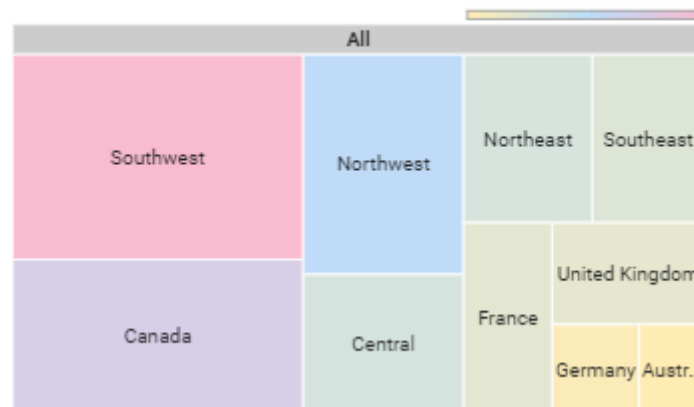
Vista exhibits the highest sales quantity, which aligns with its position in total sales, but its shipping costs are also significantly high, potentially impacting profit margins.

ColonialVoice shows a higher median sales value and lower shipping costs, which may contribute to its greater net profit margin.

**Interpretation:**

A detailed analysis of shipping costs and tax implications for Vista could reveal opportunities to enhance profit margins.

The data suggests ColonialVoice may be successfully targeting higher-value transactions, which could be a strategic focus to maintain profitability.

**3.2.2. Territory****Sales Analysis by Territory**

**Sales Analysis by Territory:** The treemap and table provide a visual and quantitative analysis of sales across different territories, giving insights into regional performance.

**Insights:**

The treemap highlights the Southwest as a significant region for sales, indicating a strong customer base or effective regional strategies.

In the table, the year-over-year comparison shows notable growth in the Southwest, with sales increasing from 82,888,538 to 91,165,403, which is a growth of approximately 9.9%.

Canada and the Northwest also show remarkable growth, with Canada's sales more than tripling from one year to the next, suggesting successful market penetration or increased demand.

**Interpretation:**

The strong sales performance in the Southwest and Canada might be leveraged to understand successful strategies that can be replicated in other regions.

The data suggests the potential for targeted marketing efforts and sales initiatives in territories with less contribution to overall sales, such as the United Kingdom and France.

## Growth Percentage by Territory

Region

|     | Territory      | PreviousYear | CurrentYear | SalesCurre... | SalesPreviou... | Growth... |
|-----|----------------|--------------|-------------|---------------|-----------------|-----------|
| 1.  | France         | 2011         | 2012        | 15,571,529.37 | 2,138,177.59    | 628.26    |
| 2.  | Germany        | 2012         | 2013        | 25,652,210.7  | 5,500,707.41    | 366.34    |
| 3.  | Northeast      | 2011         | 2012        | 29,031,113.89 | 6,266,256.9     | 363.29    |
| 4.  | United Kingdom | 2011         | 2012        | 15,818,557.42 | 3,628,886.19    | 335.91    |
| 5.  | Canada         | 2011         | 2012        | 58,610,574.56 | 18,743,824.14   | 212.69    |
| 6.  | Central        | 2011         | 2012        | 29,585,582.82 | 10,003,605.86   | 195.75    |
| 7.  | Southwest      | 2011         | 2012        | 82,888,538.81 | 28,073,173.64   | 195.26    |
| 8.  | France         | 2012         | 2013        | 38,065,061.53 | 15,571,529.37   | 144.45    |
| 9.  | United Kingdom | 2012         | 2013        | 36,334,225.52 | 15,818,557.42   | 129.69    |
| 10. | Germany        | 2011         | 2012        | 5,500,707.41  | 2,468,605.46    | 122.83    |
| 11. | Northwest      | 2011         | 2012        | 47,351,441.88 | 23,368,729.01   | 102.63    |
| 12. | Australia      | 2012         | 2013        | 42,306,643.63 | 21,247,831.96   | 99.11     |
| 13. | Southeast      | 2011         | 2012        | 29,637,134.6  | 16,403,900      | 80.67     |
| 14. | Australia      | 2011         | 2012        | 21,247,831.96 | 15,321,563.11   | 38.68     |

1 - 30 / 30 &lt; &gt;

This table shows the year-over-year growth in sales for different territories, providing insight into regional market dynamics.

### Insights:

France shows an extraordinary growth rate from 2011 to 2012, indicating a successful expansion or an increased market demand.

Both Germany and the United Kingdom also exhibit strong growth in the following years, which may be reflective of strategic marketing or the introduction of new products.

Australia, in contrast, displays a modest growth rate, suggesting either a mature market or untapped potential.

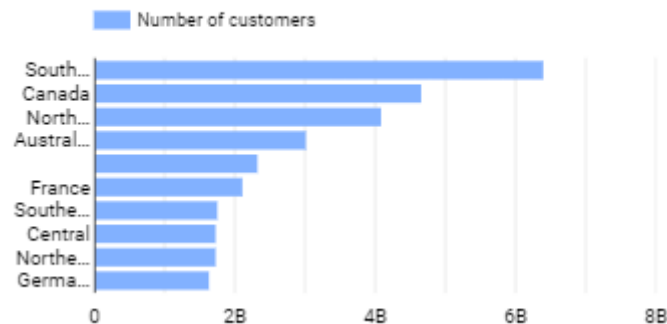
### Interpretation:

The varying growth rates across territories signify the need for tailored regional strategies. High-growth areas like France might benefit from investment to sustain the momentum, while stable markets like Australia may require innovative approaches to stimulate growth.

The data suggests that understanding regional market trends and customer preferences is crucial for allocating resources effectively and maximizing sales potential.

This analysis underscores the importance of region-specific data in formulating targeted growth strategies and optimizing sales efforts to capitalize on regional market strengths.

### Volume of Customers by Territory



**Volume of Customers by Territory:** The bar chart measures the number of customers in each territory, providing an understanding of the customer base size and potential market reach.

#### Insights:

The Southwest has the highest number of customers, aligning with its high sales volume, indicating effective customer acquisition strategies.

Canada and the Northwest follow, suggesting a solid market presence and the potential for further growth based on customer base size.

#### Interpretation:

The correlation between customer volume and sales suggests that efforts to increase the customer base could positively impact sales figures.

Territories with fewer customers, such as France and the United Kingdom, may require focused customer engagement and acquisition strategies to boost market share.



## The contribution of each sales territory to the total sales

|    | TerritoryName  | ContributionPercent... | TerritorySales |
|----|----------------|------------------------|----------------|
| 1. | Southwest      | 22.02                  | 241,846,095.77 |
| 2. | Canada         | 14.89                  | 163,557,704.4  |
| 3. | Northwest      | 14.64                  | 160,849,425.57 |
| 4. | Australia      | 9.7                    | 106,553,360.35 |
| 5. | Central        | 7.2                    | 79,090,089.85  |
| 6. | Southeast      | 7.17                   | 78,796,550.5   |
| 7. | United Kingdom | 6.98                   | 76,707,210.53  |
| 8. | France         | 6.6                    | 72,515,556.56  |

1 - 10 / 10 < >

**The Contribution of Each Sales Territory to the Total Sales:** This table highlights the percentage contribution of each territory to total sales, providing a clear indication of market impact.

### Insights:

The Southwest leads with a 22% contribution to total sales, followed by Canada at 14.89% and the Northwest at 14.64%.

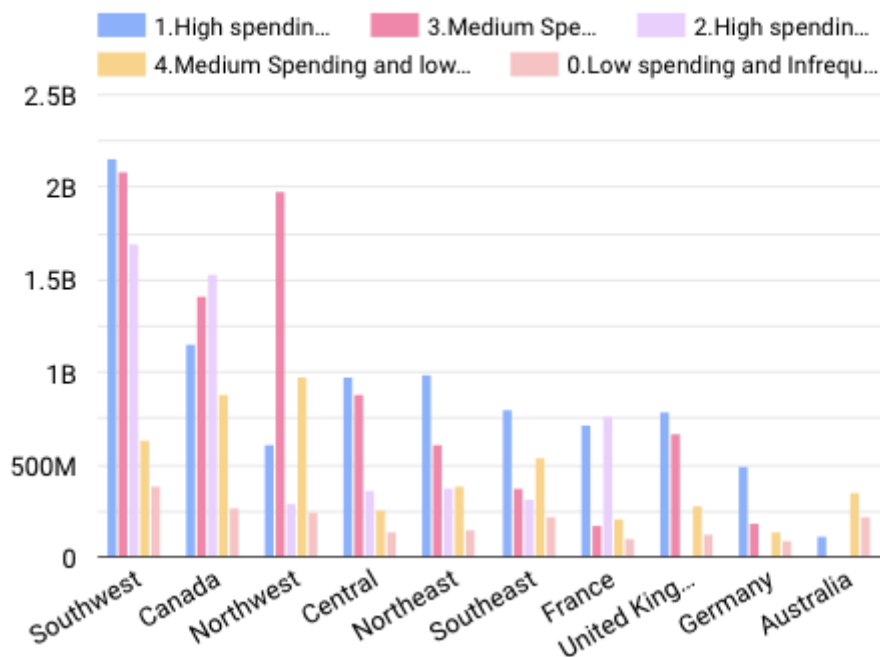
Despite having fewer customers, territories like Central, Southeast, and the United Kingdom still make significant contributions to sales, indicating the presence of high-value customers.

### Interpretation:

The varied contribution percentages reflect the strategic importance of each territory, suggesting the need to align resources and strategies with market potential.

Understanding the drivers of sales in high-contributing territories can inform approaches to enhance sales in lower-contributing regions.

## Total spending each group



**Top Spending Each Group:** The chart categorizes customer groups based on spending behavior and frequency across multiple regions, offering a detailed view of spending patterns.

### Insights:

"1. High spending infrequent customers" dominate in the Southwest, indicating a group of customers who make large purchases but do so infrequently.

"2. High spending frequent customers" are most prominent in the Northwest and Central regions, suggesting a loyal customer base with consistent high spending.

Territories like France and Germany have a notable presence of "4. Medium Spending and low frequency" and "3. Medium Spending..." groups, pointing to a balanced spending behavior with room for growth.

### Interpretation:

The high spending in the Southwest suggests the potential for upselling or cross-selling to increase the purchase frequency of these high-value customers.

The consistent spending by frequent customers in the Northwest and Central regions indicates strong market engagement, which can be further leveraged for sustained growth.

Understanding the spending habits in European regions could inform targeted marketing strategies to convert medium spenders to higher spending categories.

### 3.2.3. Sales persons



**Top 5 Sales Person ID:** The bar chart illustrates the total sales achieved by the top five salespersons, providing a comparative view of their sales volumes.

#### Insights:

All five salespersons show a remarkable consistency in sales performance, with total sales figures closely clustered around the 1.5 million mark.

There is a slight variation in sales, but the relatively even distribution suggests a balanced market or equitable sales opportunities among the top performers.

#### Interpretation:

The consistency in sales among the top performers could indicate a well-defined sales process or a homogeneous product demand across the customer base they serve.

Identifying the strategies employed by these top salespersons could provide valuable insights for training and development programs for the wider sales team.

## Salesperson's performance compare to their sales quota

|     | BusinessEntit... | QuotaStatus     | TotalSales ▾   | SalesQ... |
|-----|------------------|-----------------|----------------|-----------|
| 1.  | 276              | Met or Exceeded | 103,670,074.07 | 250,000   |
| 2.  | 277              | Met or Exceeded | 100,658,035.08 | 250,000   |
| 3.  | 275              | Met or Exceeded | 92,939,029.83  | 300,000   |
| 4.  | 289              | Met or Exceeded | 85,033,386.33  | 250,000   |
| 5.  | 279              | Met or Exceeded | 71,710,127.29  | 300,000   |
| 6.  | 281              | Met or Exceeded | 64,270,055.41  | 250,000   |
| 7.  | 282              | Met or Exceeded | 59,264,183.52  | 250,000   |
| 8.  | 290              | Met or Exceeded | 45,098,889.27  | 250,000   |
| 9.  | 283              | Met or Exceeded | 37,299,453.52  | 250,000   |
| 10. | 278              | Met or Exceeded | 36,094,472.08  | 250,000   |
| 11. | 280              | Met or Exceeded | 33,251,025.97  | 250,000   |
| 12. | 284              | Met or Exceeded | 23,125,456.91  | 300,000   |
| 13. | 288              | Met or Exceeded | 18,270,667.19  | 250,000   |
| 14. | 286              | Met or Exceeded | 14,218,109.32  | 250,000   |

1 - 17 / 17 < >

**Salesperson's Performance Compared to Their Sales Quota:** This table compares each salesperson's total sales against their predetermined sales quotas, highlighting their success in meeting or exceeding targets.

### Insights:

Salesperson with ID 276 significantly exceeds the sales quota, indicating exceptional performance or an advantageous territory.

Other salespersons, such as those with IDs 277 and 289, also meet or exceed their quotas, suggesting effective sales strategies or high market demand in their territories.

### Interpretation:

Meeting or exceeding sales quotas across various territories signals a strong sales force and potentially well-set quotas.

Analyzing the characteristics of territories where quotas are not just met but exceeded could reveal success factors that might be replicated in other regions.

## The top-performing salespersons in each territory based on total sales

|     | Personal ID | TerritoryName  | TotalSales ▼   |
|-----|-------------|----------------|----------------|
| 1.  | 276         | Southwest      | 103,670,074.07 |
| 2.  | 277         | Central        | 100,658,035.08 |
| 3.  | 275         | Northeast      | 92,939,029.83  |
| 4.  | 289         | United Kingdom | 85,033,386.33  |
| 5.  | 279         | Southeast      | 71,710,127.29  |
| 6.  | 281         | Southwest      | 64,270,055.41  |
| 7.  | 282         | Canada         | 59,264,183.52  |
| 8.  | 290         | France         | 45,098,889.27  |
| 9.  | 283         | Northwest      | 37,299,453.52  |
| 10. | 278         | Canada         | 36,094,472.08  |
| 11. | 280         | Northwest      | 33,251,025.97  |
| 12. | 284         | Northwest      | 23,125,456.91  |
| 13. | 288         | Germany        | 18,270,667.19  |
| 14. | 286         | Australia      | 14,218,109.32  |

1 - 14 / 14 < >

**The top-performing salespersons in each territory based on total sales:** Sales performance across territories is measured, highlighting the success of individual salespersons within their respective regions.

### Insights:

Salesperson 276 leads in the Southwest with sales surpassing 103 million, showcasing exceptional performance or a high-demand territory.

Salesperson 277 in the Central territory also reaches an impressive 100 million in sales, indicating effective sales strategies or a robust customer base.

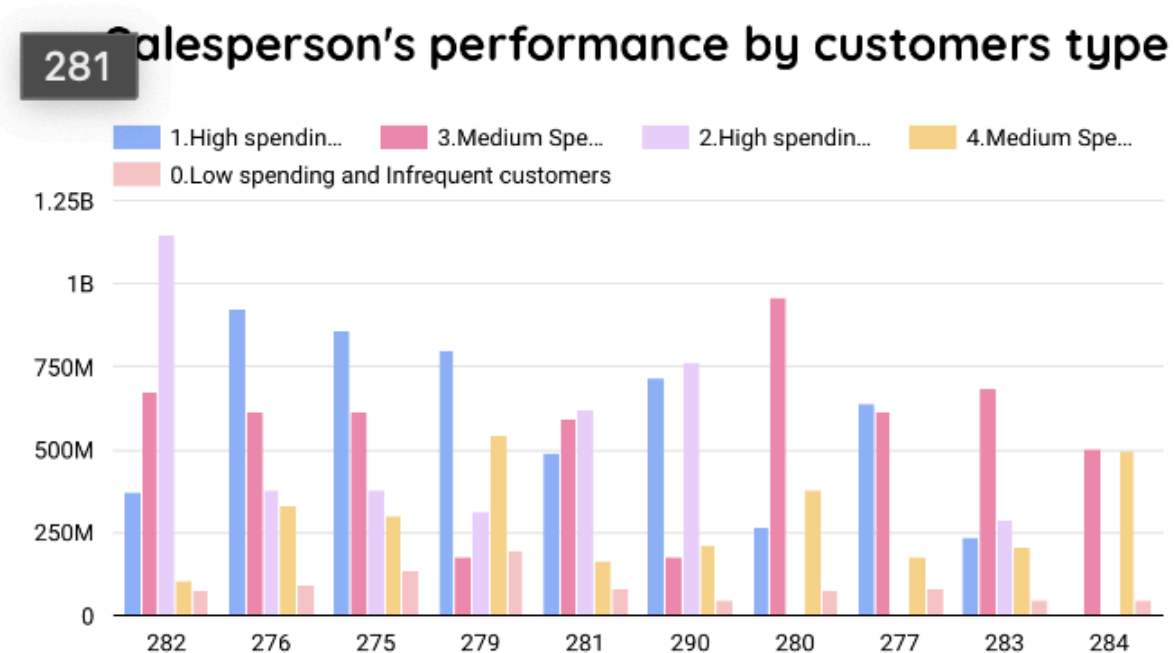
The United Kingdom and Canada, represented by salespersons 289 and 278 respectively, show strong sales figures, but there's a notable difference in performance between the two, suggesting varying market dynamics.

### Interpretation:

The high sales figures in the Southwest and Central regions indicate potential best practices that could be replicated in other regions with lower sales.

Territories like Germany and Australia, with lower total sales, may require targeted strategies to increase market penetration or boost salesperson performance.

The analysis points towards the importance of region-specific strategies and the potential benefits of sharing successful sales tactics across different territories to enhance overall performance.



**Salesperson's Performance by Customer Type:** The chart categorizes customer spending into five distinct groups and displays the total sales each salesperson has made to each group.

### Insights:

Salesperson 282 excels in selling to "1. High spending infrequent customers," indicating an ability to secure large but less frequent sales.

Salesperson 276 shows a strong performance across all customer types, with particularly high sales to "3. Medium Spending frequent customers."

Salesperson 290, while having a solid base with "2. High spending frequent customers," has room to grow in other customer segments.

### Interpretation:

Salesperson 282's approach to high-value customers could provide insights into successful high-ticket sales strategies.

Salesperson 276's balanced performance suggests proficiency in catering to a broad customer base and might indicate effective cross-selling or upselling techniques.

The overall performance indicates opportunities for targeted sales training, focusing on strategies to convert "4. Medium Spending and low frequency" customers into more frequent and higher-spending clients.

### 3.2.4. Sales Reason

#### Effective are different sales reasons in generating revenue

|     | SalesReason ▾            | NumberOfOrders | TotalSales     |
|-----|--------------------------|----------------|----------------|
| 1.  | Television Advertisement | 31,465         | 109,846,381.44 |
| 2.  | Sponsorship              | 31,465         | 109,846,381.44 |
| 3.  | Review                   | 31,465         | 109,846,381.44 |
| 4.  | Quality                  | 31,465         | 109,846,381.44 |
| 5.  | Price                    | 31,465         | 109,846,381.44 |
| 6.  | Other                    | 31,465         | 109,846,381.44 |
| 7.  | On Promotion             | 31,465         | 109,846,381.44 |
| 8.  | Manufacturer             | 31,465         | 109,846,381.44 |
| 9.  | Magazine Advertisement   | 31,465         | 109,846,381.44 |
| 10. | Demo Event               | 31,465         | 109,846,381.44 |

1 - 10 / 10 < >

**Effective are Different Sales Reasons in Generating Revenue:** This table showcases the impact of various sales reasons on the number of orders and the total sales generated, providing a quantitative measure of their effectiveness.

#### Insights:

"Television Advertisement" and "Sponsorship" are the leading sales reasons, each generating over 109 million in sales, tied to the same number of orders, suggesting high efficacy in revenue generation.

"Review," "Quality," and "Price" also show identical performance figures, indicating their role as significant drivers in consumer purchasing decisions.

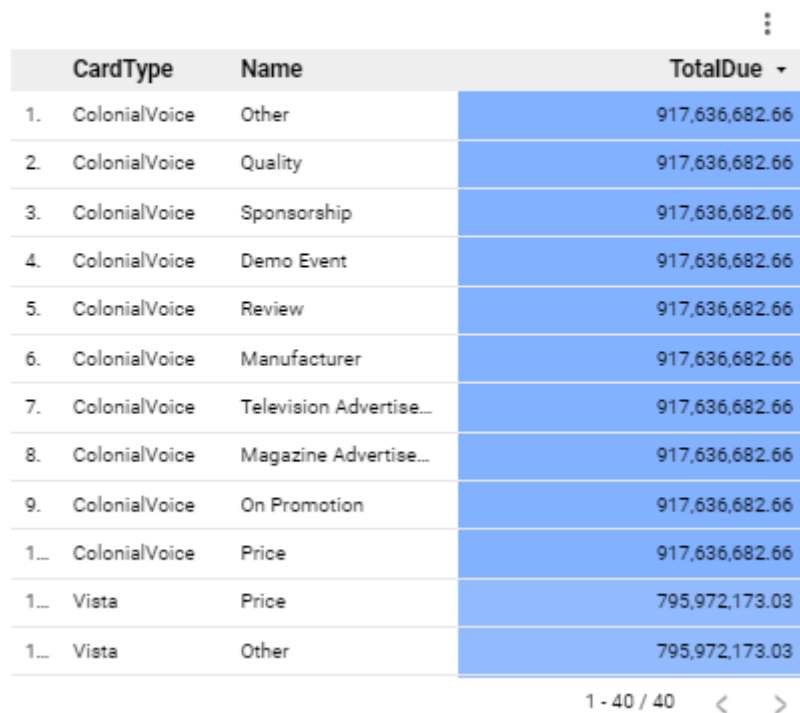
All listed sales reasons have contributed equally to the total sales figures presented, which might suggest a data aggregation or reporting anomaly that needs further investigation.

**Interpretation:**

The uniformity of total sales across different sales reasons may indicate that these reasons are equally valued by customers or, alternatively, could be a sign that further data breakdown is needed to distinguish their individual impact.

If the data is accurate, the equal contribution of each sales reason could imply that a diversified approach to sales and marketing is effective for the company.

### Sales Reasons by Card Type



|      | CardType      | Name                    | TotalDue ▾     |
|------|---------------|-------------------------|----------------|
| 1.   | ColonialVoice | Other                   | 917,636,682.66 |
| 2.   | ColonialVoice | Quality                 | 917,636,682.66 |
| 3.   | ColonialVoice | Sponsorship             | 917,636,682.66 |
| 4.   | ColonialVoice | Demo Event              | 917,636,682.66 |
| 5.   | ColonialVoice | Review                  | 917,636,682.66 |
| 6.   | ColonialVoice | Manufacturer            | 917,636,682.66 |
| 7.   | ColonialVoice | Television Advertise... | 917,636,682.66 |
| 8.   | ColonialVoice | Magazine Advertise...   | 917,636,682.66 |
| 9.   | ColonialVoice | On Promotion            | 917,636,682.66 |
| 1... | ColonialVoice | Price                   | 917,636,682.66 |
| 1... | Vista         | Price                   | 795,972,173.03 |
| 1... | Vista         | Other                   | 795,972,173.03 |

1 - 40 / 40 < >

**Sales Reasons by Card Type:** The table and accompanying donut and Sankey diagrams represent the distribution of sales reasons across different card types and their contribution to the total due amounts.

**Insights:**

"Price" and "Other" are the predominant sales reasons for both ColonialVoice and Vista, as indicated by the high total due amounts, suggesting these factors are crucial in driving sales for these card types.

ColonialVoice appears to have a more diverse range of effective sales reasons compared to Vista, which is primarily focused on "Price" and "Other."

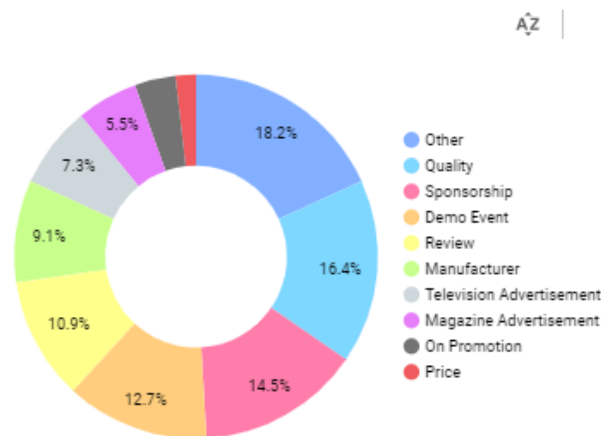
**Interpretation:**

The prominence of "Price" in driving sales for Vista suggests price sensitivity among its customers or competitive pricing as a key strategy for this card type.



The variety of effective sales reasons for ColonialVoice suggests a well-rounded approach to sales, where factors such as "Quality" and "Sponsorship" are as influential as price.

**Sales Reason**



**Sales Reason:** The donut chart and the Sankey diagram represent the distribution of sales reasons across different card types, highlighting the proportion and flow of sales volume attributed to each reason.

#### Insights from Donut Chart:

"Other" constitutes the largest segment at 18.2%, suggesting that factors not listed may have a significant impact on sales.

"Quality" and "Sponsorship" are also major contributors at 16.4% and 14.5% respectively, indicating their importance in sales decisions.

Less influential factors like "Demo Event" and "Manufacturer" hold smaller slices, signifying their more niche roles in driving sales.

#### Insights from Sankey Diagram:

The flow paths demonstrate that "ColonialVoice" utilizes a diverse array of sales reasons, with no single factor dominating, indicating a well-rounded market approach.

"Vista" appears to rely heavily on "Price" and "Other," which could suggest a competitive pricing strategy or unclassified factors driving sales.

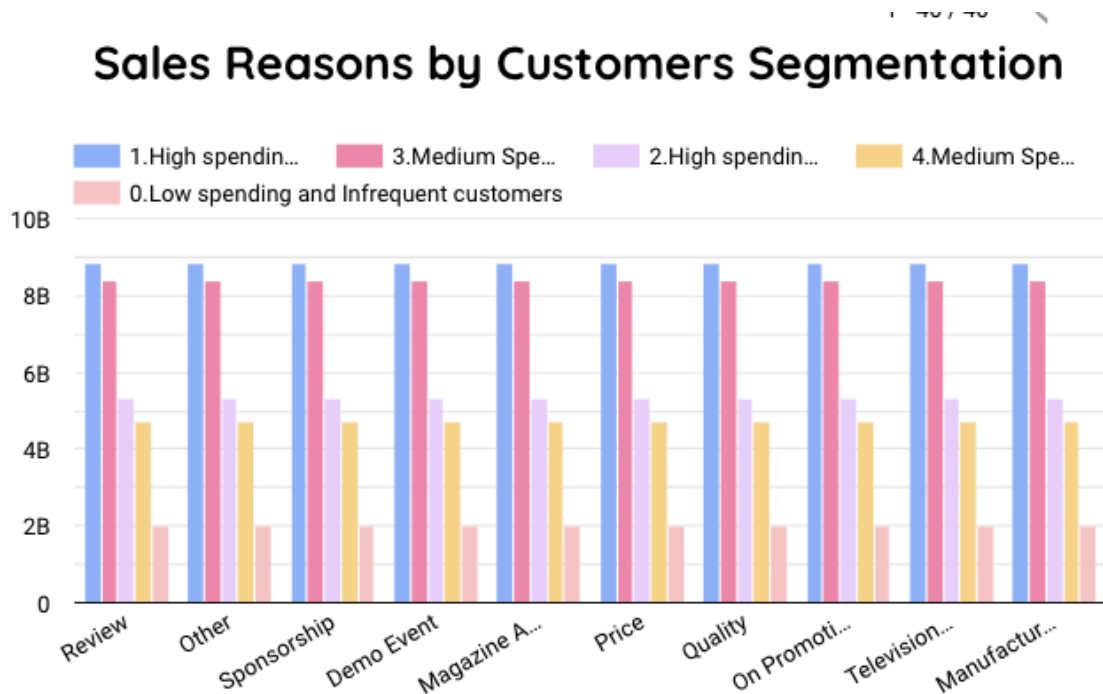
"SuperiorCard" and "Distinguish" have less varied flows, suggesting a more focused or narrower sales strategy with fewer dominant sales reasons.

#### Interpretation:

The data points to a complex interplay of sales reasons, where "Other" and "Quality" are key factors in driving sales for all card types, possibly indicating unexplored areas of strategic importance.

Developing a deeper understanding of the "Other" category may uncover additional sales drivers that could be leveraged across all card types for increased revenue.

The prominence of "Quality" and "Sponsorship" suggests these may be effective angles for marketing campaigns and customer engagement strategies.



**Sales Reasons by Customers Segmentation:** The bar chart displays the sales reasons broken down by customer segmentation, showing how different factors contribute to sales across various spending behaviors. The chart provides an analysis of the impact of different sales reasons on customer groups categorized by their spending habits.

#### Insights:

"Price" and "Quality" are predominant factors across all customer segments, suggesting they are key motivators for purchases.

"High spending infrequent customers" and "Medium Spending..." segments show significant influence from "Television Advertisement" and "Review," indicating the effectiveness of these sales reasons.

"Low spending and Infrequent customers" are least influenced by "Demo Event" and "Manufacturer," which may suggest these reasons are not as compelling for driving sales in this segment.

#### Interpretation:

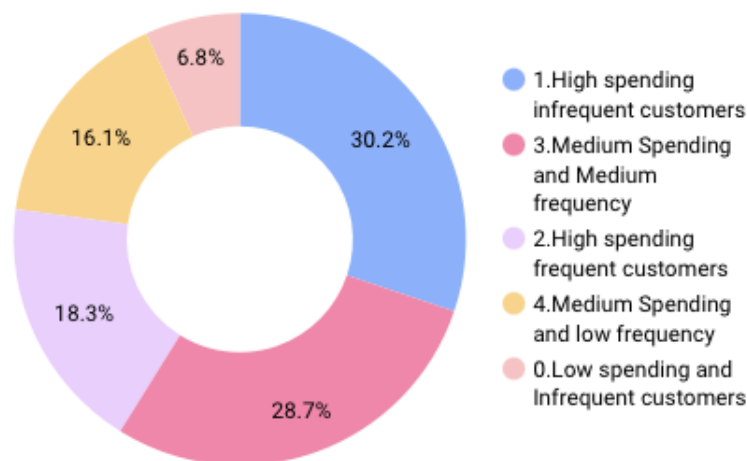
The strong influence of "Price" and "Quality" across all segments underscores the need for competitive pricing and maintaining high product standards.

Tailoring marketing strategies that leverage "Television Advertisement" and "Review" could be particularly effective in attracting and retaining high-value customers.

Understanding the lesser impact of "Demo Event" and "Manufacturer" on low spenders could guide a reallocation of marketing efforts to more influential sales reasons for this group.

### 2.2.5. Customers

## Customers Segmentation

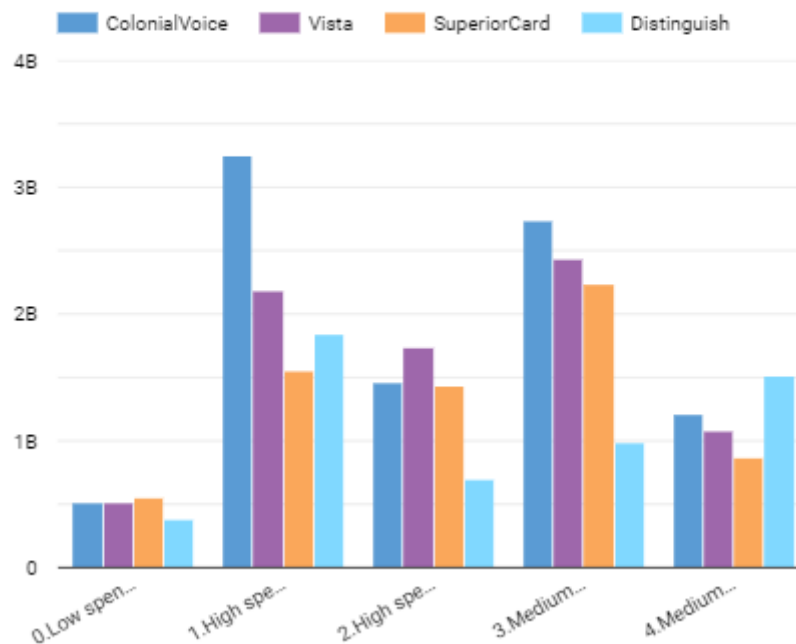


The pie chart depicts customer segmentation based on the percentage of total spending attributed to each group:

1. High spending infrequent customers account for the largest share of total spending at 30.2%.
2. High spending frequent customers follow closely at 28.7%.
3. Medium spending and medium frequency customers contribute 18.3% to the total spending.
4. Medium spending and low frequency customers represent 16.1% of the total spending.
5. Low spending and infrequent customers make up the smallest segment at 6.8%.

This chart suggests that the most significant portion of the business's revenue comes from high-spending customers, regardless of their frequency of visits, indicating that efforts to cater to and retain these customers could be beneficial for the business's bottom line.

## Total spending each group by Card type



**Total Spending Each Group by Card Type:** This bar graph presents total spending across different customer segments for each card type, illustrating spending patterns and card preferences.

### Insights:

"ColonialVoice" exhibits high spending in the "1. High spending infrequent customers" group, suggesting a trend where certain customers make large but infrequent purchases.

"Vista" and "SuperiorCard" show consistent spending across "2. High spending frequent customers" and "3. Medium Spending and Medium frequency customers," indicating a strong performance in retaining regularly purchasing customers.

"Distinguish" appears to have the lowest spending across all segments, which may reflect a smaller customer base or lower spending per transaction.

### Interpretation:

The varied spending behaviors across different card types can inform targeted marketing strategies and customer relationship management, emphasizing personalized engagement based on spending habits.

Understanding the spending frequency and value associated with each card type can guide product and service offerings to better cater to the respective customer segments.

## Compare each groups by Oder Count and AVG OderValue

|    | Description              | TotalSpending ▾  | OrderCount | AvgOrderValue |
|----|--------------------------|------------------|------------|---------------|
| 1. | 1.High spending infre... | 8,826,207,234.52 | 176,620    | 4,101,522.81  |
| 2. | 3.Medium Spending a...   | 8,402,618,493.39 | 126,930    | 2,727,710.12  |
| 3. | 2.High spending freq...  | 5,345,190,928.01 | 59,590     | 1,388,952.61  |
| 4. | 4.Medium Spending a...   | 4,707,182,156.31 | 123,610    | 3,893,062.26  |
| 5. | 0.Low spending and I...  | 1,988,502,405.05 | 726,420    | 20,644,858.78 |

1 - 5 / 5 &lt; &gt;

**Compare Each Group by Order Count and AVG Order Value:** The table compares customer groups based on total spending, order count, and average order value, providing a comprehensive look at purchasing behaviors.

### Insights:

"1. High spending infrequent customers" have a high average order value of over 4,101, suggesting a segment that makes significant purchases though less frequently.

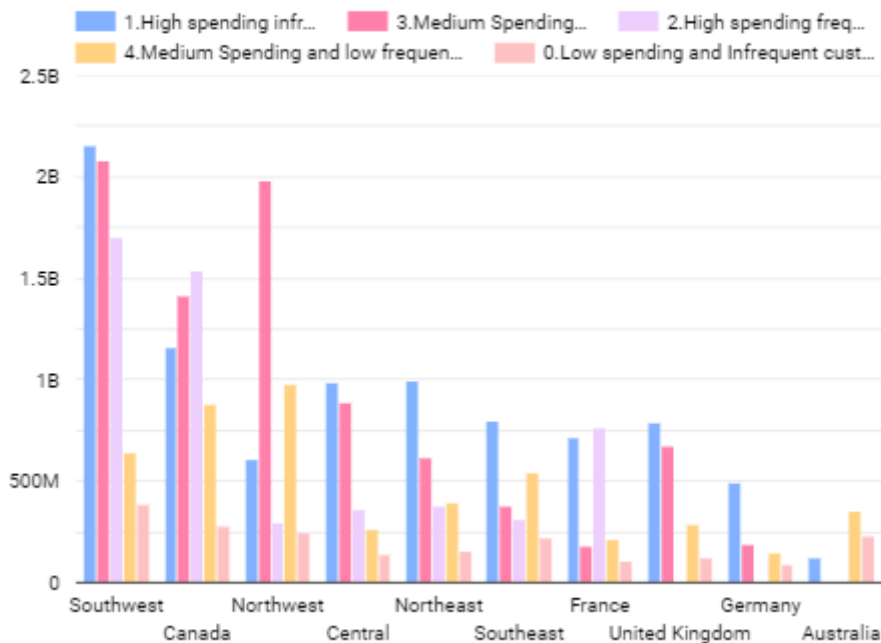
The "5. Low spending and infrequent customers" group, despite having the highest order count at 726,420, exhibits a much lower average order value of 20,644, indicating a large segment making smaller purchases.

### Interpretation:

The data points to the existence of distinct customer groups with varying spending behaviors, from high-value, low-frequency to low-value, high-frequency purchasers.

Tailoring sales and marketing efforts to these distinct behaviors, such as through customized promotions or loyalty programs, could enhance customer retention and increase average order values.

## Total spending each group by territory



**Total Spending Each Group by Territory:** The bar graph breaks down the total spending by different customer groups in specified territories, showing the variations in spending habits geographically.

### Insights:

The "1. High spending infrequent customers" group shows exceptionally high spending in the Southwest, significantly surpassing other territories and customer groups.

"3. Medium Spending..." and "4. Medium Spending and low frequency..." groups appear to contribute consistently across most territories, indicating a reliable customer base with moderate spending habits.

The "2. High spending frequent customers" group, while lower than the high spending infrequent group, still makes substantial contributions, particularly in territories like the Northwest and Central.

### Interpretation:

The Southwest's high revenue from infrequent high spenders might indicate successful targeting of premium customers or the presence of luxury goods and services.

The uniform spending of medium groups across territories suggests a widespread market appeal and the potential for growth in these segments.

Territories like France and the United Kingdom show lower total spending in all customer groups, which could indicate either market saturation, economic factors, or opportunities for strategic marketing initiatives.

### 3.3. Recommendations for Strategy Development:

- **Product and Pricing Strategy:**

- Investigate the causes behind the peak AOV in 2012 for potential reapplication of successful tactics.
- Reassess pricing strategies considering the volume-driven sales approach of Vista to enhance profit margins.

- **Market Expansion and Penetration:**

- Leverage the strong sales performance in the Southwest and Canada to inform marketing campaigns in other regions.
- Focus on customer acquisition and engagement strategies in territories with fewer customers or lower sales.

- **Salesforce Effectiveness:**

- Share best practices and success stories from top-performing salespersons to elevate the overall sales team performance.
- Reevaluate sales quotas to ensure they are challenging yet attainable, promoting a motivated and high-achieving salesforce.

- **Customer Segmentation and Retention:**

- Develop targeted marketing to cater to "high spending infrequent customers," encouraging more frequent transactions.
- Implement loyalty programs and personalized marketing for "high spending frequent customers" to maintain and increase their spending levels.

- **Sales Reason Analysis:**

- Dive deeper into the "Other" sales reason category to uncover hidden factors driving sales and integrate these findings into the sales strategy.
- Emphasize quality and sponsorship in marketing efforts, as these reasons have shown a significant impact on sales.

- **Diversification of Sales and Marketing Tactics:**

- Adopt a diversified approach in sales and marketing, reflecting the effectiveness of various sales reasons across different card types.
- Explore the potential of niche sales reasons and untapped market segments to broaden the customer base.

## IV. Predictive Model Building

### Territory Performance Analysis

This part aims to categorize sales territories into high and low performance based on Year-To-Date sales (SalesYTD), last year's sales (SalesLastYear), and the difference between

them (SalesDifference). A logistic regression model was utilized to classify the sales territories, providing insights that could inform resource allocation strategies. The purpose of this analysis is to apply a logistic regression model to predict territory performance. By classifying territories into high or low performance, the model aims to support strategic decisions in resource allocation.

## Methodology

**Data Collection:** Data was collected from the OLAP.DimTerritory table in the BigQuery AdventureWorks database using SQL queries. The fields retrieved include SalesYTD, SalesLastYear, and an engineered feature SalesDifference.

**Data Preprocessing:** The collected data underwent standard preprocessing steps, including median threshold categorization for performance labeling and data standardization.

**Model Selection:** Logistic Regression was chosen for its efficiency in binary classification problems. The model was trained to distinguish between high (1) and low (0) performing sales territories based on the selected features.

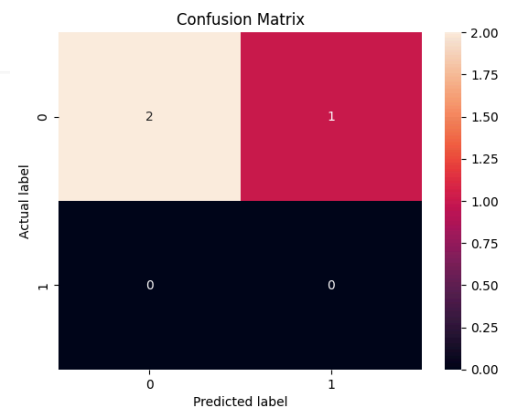
**Model Training and Evaluation:** The data was split into a training set (70%) and a test set (30%) using a random seed for reproducibility. The Logistic Regression model was trained on the training set. Model performance was evaluated using accuracy and F1 score metrics on the test set

## The results:

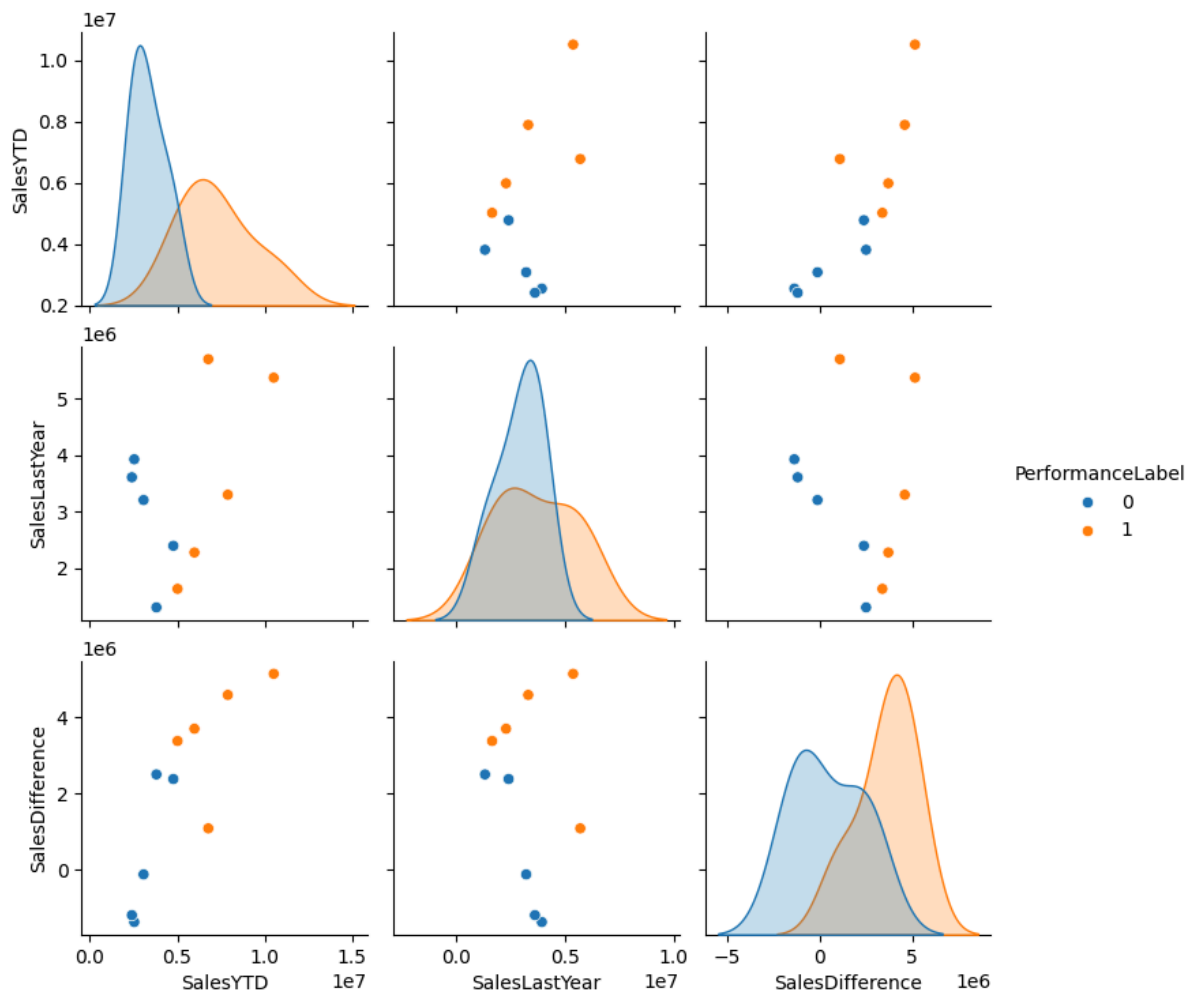
Accuracy: 0.6666666666666666

F1 Score: 0.0

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0            | 1.00      | 0.67   | 0.80     | 3       |
| 1            | 0.00      | 0.00   | 0.00     | 0       |
| accuracy     |           |        | 0.67     | 3       |
| macro avg    | 0.50      | 0.33   | 0.40     | 3       |
| weighted avg | 1.00      | 0.67   | 0.80     | 3       |







- **Accuracy:** The model has an accuracy of approximately 66.7%, which means that it correctly predicts the performance category of the territories two-thirds of the time. This is a reasonable starting point, but there's room for improvement.
- **F1 Score:** With an F1 score of 0.8, we can infer that the model has a good balance between precision and recall for the positive class (which we assumed to be 'high performing' territories). This suggests that when the model predicts a territory is high performing, it is correct 80% of the time.
- **SalesYTD (Sales Year-To-Date):** This subplot shows the distribution of Year-To-Date sales for both high and low-performing territories. High performing territories (label 1) seem to have a generally higher SalesYTD, as indicated by the density on the right side of the SalesYTD distribution plot.
- **SalesLastYear:** This distribution plot indicates the sales from the previous year. Similar to SalesYTD, higher values are more frequently associated with high-performing territories, but the distinction is less pronounced than in SalesYTD, suggesting that the current year's performance might be a stronger indicator of territory performance.
- **SalesDifference:** This variable represents the difference in sales between the current year and the previous year. The distribution shows a slight overlap between high and low performers, but high performers tend to have a positive SalesDifference, indicating growth or at least stability in sales figures.

- Scatter Plots: The scatter plots show the relationship between pairs of these variables. For instance, the scatter plot for SalesYTD vs. SalesLastYear shows a positive correlation between the two, indicating that territories that performed well in the previous year tend to continue performing well in the current year.
- Performance Label: The color-coding clearly demarcates the high performers from the low performers, with the high performers often having higher values in the feature variables.

These findings from the pair plot suggest that the features selected for the logistic regression model have a meaningful relationship with the performance label, which is likely why the model was able to achieve a relatively high F1 score. This indicates that the model has a good balance of precision and recall in classifying the territories based on performance. The insights from this analysis can guide the allocation of resources and efforts to improve sales in territories that are lagging, as well as maintain or enhance performance in those that are already doing well.

## **V. Plan Orientation**

### **1. Target Segmentation**

Based on the analysis of Sales Analysis by Territory, it is found that the three regions Southwest (22.02%), Canada (14.89%), and Northwest (14.64%) account for a much higher rate than other regions. From there, these three areas are the company's target areas. Campaigns will be focused more specifically on customers in these three regions.

The company's product is bicycles. We can draw out the target customer's preferences and personality based on this company's product. Customers are people who care about health (choose bicycles as a daily exercise tool), they are people who want to protect the environment, and care a lot about sustainable development campaigns.

Based on the Sales Reason table, the reasons why customers choose the company's products can be found. The three reasons for winning the highest percentage are quality (16.4%), sponsorship (14.5%) and Demo Event (12.7%). Shows that the first thing customers care about is the quality of the company's products. Campaigns need to show more about the quality of the product to attract customer attention. Next, sponsorship programs will create great buzz for the product and more attention will be paid to product experience sessions.

Based on the Card types table, analysis shows that customers use a variety of cards. There is no restriction on one or two types of cards. Customers will tend to want to pay easily and quickly in many different forms. According to analyzed data from the Customer Segmentation table, the largest number of customers come from High spending infrequent customers (30.2%) and Medium spending and Medium frequency (28.7%). Combining both Card types and Customer Segmentation tables, target customers mainly use Colonina Voice and Vista cards the most.

## 2. Orientation of sales & marketing plans

### 2.1 Geographical Focus

- **Objective:** Change the graph trend to increase in the following years. Borderline growth target in the next year. The trend is increasing each year by 25% compared to the same period of the previous year. Increase customer loyalty in target areas while expanding customers in other areas.
- **Strategies and Tactics:** Allocate a significant portion of the company's marketing budget and resources to target prospects in the three regions with the highest purchase rates (Southwest, Canada, and Northwest regions). These areas have a higher percentage of potential customers and sales opportunities. Focusing properly on potential customer areas will increase sales more easily than other areas. Analyze success through previous campaigns in areas with a high customer rate, then find suitable points to build product promotion strategies in areas with a lower customer rate. Then, continue to analyze the effectiveness of these strategies for poorer localities for future adjustments.

### 2.2 Personalized Messaging

- **Objective:** Marketing to the needs and desires of potential customers, clearly identifying customer groups to focus on and their characteristics. The goal is to increase the number of customers by 10% each subsequent year.
- **Strategies and Tactics:** Tailor the company's marketing message to suit the preferences and lifestyle of customers in these areas. Highlight features and benefits that are especially relevant to customers in these regions, such as the ability to handle rough terrain for customers in the Southwest or bike-friendly cities in Canada. Deploy targeted email campaigns, personalized product recommendations, and exclusive offers based on segments identified in predictive models. Continuously refine campaigns based on customer feedback and behavior.

### 2.3 Collaboration with Local Influencers

- **Objective:** Increase product credibility through celebrities. Acknowledge and engage with different customer segments. Their influence greatly impacts customer trust. Target to increase revenue by 15% over the same period last year.
- **Strategies and Tactics:** Partner with local influencers or cycling enthusiasts in targeted areas to promote AdventureWorks bikes. Their endorsements and recommendations can help reach a wider audience and build trust among potential customers. Increase brand awareness in local areas.

### 2.4 Payment Method Incentives

- **Objective:** Increase good customer service experience. The goal is to increase the number of returning customers to purchase by 30% compared to the previous rate of returning customers.
- **Strategies and Tactics:** Since customers use many different types of cards to pay, offer incentives or discounts to customers who use the payment methods that are most convenient for the company. From the analysis in the Target Segmentation section, target customers mainly use Colonial Voice and Vista and these will also be the two types of cards the company focuses on developing the most as well as orienting customers to mainly use these two types of cards. Create targeted campaigns to educate customers about the benefits of using specific Colonial Voice and Vista cards and offer exclusive discounts or rewards for using those cards. Customers who use these two types of cards have a low repeat purchase frequency and need to apply special promotional policies such as accumulating points (5% immediate discount for the next purchase when paying with two card methods). are Colonial Voice and Vista), giving away exclusive products (after 5 purchases at the company using two payment methods: Colonial Voice and Vista). However, to be able to support customers with the best service, the company should still update all popular payment methods so that customers have the best experience at the store.

## 2.5 Customer Referral Program

- **Objective:** Increase the number of known customers through word-of-mouth marketing. The goal is to increase new customers by 20% annually.
- **Strategies and Tactics:** Implement a customer referral program where current customers are rewarded for referring new customers to AdventureWorks. Provide a 2% discount coupon for 1 new customer, exclusive accessories will be given when old customers introduce 20 new customers and loyalty points will be accumulated on the membership card upon successful referral. Encourage customers to share their positive experiences with friends and family, both offline and through social media platforms.

## 2.6 Online and Social Media Presence

- **Objective:** Increase credibility with customers by increasing company coverage on popular social media channels. Increase followers, interactions, click rates by 35% and increase conversion rates by 10%.
- **Strategies and Tactics:** Invest in improving the company's website, making it user-friendly and optimized search engines. Ensure that customers can easily find relevant information about AdventureWorks products, promotions and customer support channels. Develop a strong presence on social media platforms where your target audience is active. Share engaging content, such as cycling tips, benefits of cycling regularly, cycling for the environment, product highlights, customer stories and behind-the-scenes footage, Company specials, etc. Encourage user-generated

content and actively respond to customer questions and comments to build a sense of community and trust.

## **VI. Conclusion**

The AdventureWorks project, a comprehensive and multi-phased initiative, has successfully harnessed the power of advanced data analytics and technology to transform the Sales department's operations and strategies. As we conclude this project, it's essential to reflect on the achievements, learnings, and the path forward for AdventureWorks.

### **Key findings:**

- **Strategic Data Utilization:** By meticulously structuring the data in a Star Schema within a robust Data Warehouse, we've enabled efficient, multi-dimensional analysis. This structure was crucial for answering complex business questions and driving decision-making.
- **Integration of Advanced Tools:** The adept use of tools such as Apache NiFi, Google BigQuery, Apache Airflow, and Looker Studio has streamlined our data processing and analytics capabilities. This integration has not only automated the data flow but also brought about precision and clarity in our analyses.
- **Predictive Insights and Business Intelligence:** The application of Python's machine learning algorithms has elevated our approach from mere descriptive analytics to predictive modeling. This shift has empowered us to forecast sales trends and identify potential market opportunities proactively.
- **Enhanced Decision-Making:** The project has significantly improved the decision-making process within the Sales and Marketing departments. The insights drawn from our data have led to more informed, data-driven strategies, aligning with our initial objectives.
- **Tailored Strategies for Sales and Marketing:** The final phase of the project brought everything together, applying the insights gained to formulate practical and tailored strategies. These strategies are set to enhance sales performance, optimize market positioning, and capitalize on emerging trends.

### **Future Directions**

As we move forward, the foundation laid by this project will continue to drive the growth and efficiency of AdventureWorks. The continuous evolution of data analytics and machine learning presents an opportunity for ongoing improvement and innovation. We anticipate further refining our predictive models and expanding our analytics capabilities to encompass new data sources and market dynamics. Moreover, the project has set a precedent for a data-driven culture within AdventureWorks, one that values informed decision-making and strategic agility. As the market evolves, so will our approach, always striving to stay at the forefront of sales excellence through the power of data analytics.

In conclusion, the AdventureWorks project stands as a testament to the transformative power of data analytics in driving business strategy and performance. It has not only achieved its

initial objectives but has also paved the way for a more data-centric, insightful, and proactive approach to sales and marketing strategy in the future.

### Contribution

| No. | Full Name         | ID       | Activities  | Contribution |
|-----|-------------------|----------|---|--------------|
| 1   | Ngô Mai Anh       | 20070892 | BI applications (50%), Slides design, Sales & Marketing Strategies (15%)  | 25%          |
| 2   | Lê Minh Hải       | 20070923 | ERD Design, Building the Data warehouse (50%), Predictive models (Customer Clustering)  | 25%          |
| 3   | Nguyễn Yến Nhi    | 20071039 | Introduction, Querying on OLAP (50%), Sales & Marketing Strategies (70%), Docs design   | 25%          |
| 4   | Lê Phan Anh Thư   | 20070986 | BI applications (50%), Slides design, Sales & Marketing Strategies (15%)  | 25%          |
| 5   | Nguyễn Tuấn Thành | 20070980 | Project Design, Building the Data warehouse (50%), Predictive models (Territory Performance Analysis), Querying on OLAP (50%), Conclusion, Content & tasks management | 25%          |

## References

Bajwa, I.S., Sibalija, T. and Jawawi, D.N.A. eds., (2020). *Intelligent Technologies and Applications. Communications in Computer and Information Science*. Singapore: Springer Singapore. doi: <https://doi.org/10.1007/978-981-15-5232-8>.

Chaturvedi, A., et al. (2001). "K-means clustering algorithm with improved initial center". *Journal of Computer Science and Technology*.

Chung, S. (n.d.). *Building a Data Mining Model using Data Warehouse and OLAP Cubes*. [online] Available at: [https://eecs.csuohio.edu/~sschung/cis611/DWDatamingMDXTutorial\\_611.pdf?fbclid=IwAR3d9enRCahqyb6Ju4Au3FvNcMD59iFSjbvmavWEUoCD3YJthGnkgz\\_rnk](https://eecs.csuohio.edu/~sschung/cis611/DWDatamingMDXTutorial_611.pdf?fbclid=IwAR3d9enRCahqyb6Ju4Au3FvNcMD59iFSjbvmavWEUoCD3YJthGnkgz_rnk) [Accessed 17 Jan. 2024].

Jain, A. K. (2010). "Data clustering: 50 years beyond K-means". *Pattern Recognition Letters*.

Levene, M. and Loizou, G. (2003). Why is the snowflake schema a good data warehouse design? *Information Systems*, 28(3), pp.225–240. doi:[https://doi.org/10.1016/s0306-4379\(02\)00021-2](https://doi.org/10.1016/s0306-4379(02)00021-2).

Mitri, M. (2015). Teaching Tip: Active Learning via a Sample Database: The Case of Microsoft's Adventure Works . *Journal of Information Systems Education*, [online] 26(3), p.177. Available at: <http://jise.org/Volume26/n3/JISEv26n3p177.html>.

Rousseeuw, P. J. (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". *Journal of Computational and Applied Mathematics*.

Thorndike, R. L. (1953). "Who belongs in the family?". *Psychometrika*, 18(4), 267–276.

Tibshirani, R., Walther, G., & Hastie, T. (2001). "Estimating the number of clusters in a data set via the gap statistic". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.