SQL Sakila Dataset

Basic Queries

1. List all actors' first and last names in alphabetical order.
   Expected output:
   First name, Last name
   Adam, Grant
   Amy, Adams
   Angela, Basset

2. Find all films released in the year 2006.
   Expected output:
   Title, Release year
   Academy Dinosaur, 2006
   Adaptation Holes, 2006

3. Retrieve all unique film categories.
   Expected output:
   Name
   Action
   Comedy
   Drama

4. Get the total number of customers.
   Expected output:
   Total customers
   599

5. Find all rental records where the return date is null.
   Expected output:
   Rental id, Customer id, Inventory id, Rental date
   1523, 45, 1023, 2024-03-10 14:30:00
   1678, 98, 1289, 2024-03-12 10:15:00

6. Display the names and emails of customers from the city of London.
   Expected output:
   First name, Last name, Email
   John, Doe, john.doe@email.com
   Jane, Smith, jane.smith@email.com

7. List the staff members along with the store they work at.
   Expected output:
   Staff id, First name, Last name, Store id
   1, Mike, Johnson, 1
   2, Sarah, Adams, 2

8. Retrieve a list of countries that have customers.
   Expected output:
   Country
   USA
   Canada
   UK

9. Find all films with a rental duration greater than five days.
   Expected output:
   Film id, Title, Rental duration
   12, The Grand Escape, 7
   25, Mystery Island, 6

10. Get the list of movies that belong to the Action category.
    Expected output:
    Film id, Title
    45, Action Hero Returns
    78, Speed Chase

Intermediate Queries

11. Find the number of films available in each category.
    Expected output:
    Category, Film count
    Action, 55
    Comedy, 47
    Drama, 62

12. List the top five most rented movies along with their rental count.
    Expected output:
    Title, Rental count
    Jurassic Hunt, 85
    Mystery Island, 78

13. Retrieve the total number of rentals for each customer sorted by most rentals.
    Expected output:
    Customer id, First name, Last name, Rental count
    102, John, Smith, 15
    98, Alice, Johnson, 13

14. Find the highest rental rate of movies in each category.
    Expected output:
    Category, Max rental rate
    Action, 4.99
    Comedy, 3.99

15. Get the names of customers who have rented the movie Academy Dinosaur.
    Expected output:
    Customer id, First name, Last name
    205, Peter, Griffin
    310, Lois, Lane

16. Display a report of total payments collected by each staff member.
    Expected output:
    Staff id, First name, Last name, Total collected
    1, John, Doe, 5000.00
    2, Sarah, Smith, 4800.00

17. Find the total revenue generated by each store.
    Expected output:
    Store id, Total revenue

1, 10000.00
2, 9500.00

Advanced Queries

18. Find the most popular movie category based on rental count.
    Expected output:
    Category, Rental count
    Action, 500

19. Identify customers who have spent more than 100 dollars in total rentals.
    Expected output:
    Customer id, First name, Last name, Total spent
    305, Alice, Smith, 120.00

20. Get the second highest rental count for any movie.
    Expected output:
    Rental count
    78

21. List the customers who have rented all movies from the Horror category.
    Expected output:
    Customer id, First name, Last name
    201, Bob, Marley

22. Find the city that has the highest number of customers.
    Expected output:
    City, Customer count
    New York, 58

23. Retrieve the details of films that have never been rented.
    Expected output:
    Film id, Title
    145, Forgotten Island

24. Find customers who have rented more than five movies but have never rented a movie from the Comedy category.
    Expected output:
    Customer id, First name, Last name, Rental count
    308, Emma, Watson, 7

25. Get a ranked list of customers based on total money spent on rentals.
    Expected output:
    Rank, Customer id, First name, Last name, Total spent
    1, 102, John, Smith, 250.00
    2, 305, Alice, Johnson, 220.00

26. Find the month with the highest revenue across all years.
    Expected output:
    Month, Year, Total revenue
    June, 2024, 5500.00

27. List movies with a rental rate higher than the average rental rate of all movies.
    Expected output:
    Film id, Title, Rental rate

102, Action Blast, 4.99
208, Super Spy, 5.99

28. Retrieve the average rental duration for each film category.
    Expected output:
    Category, Average rental duration
    Action, 5.4
    Comedy, 4.9

29. Find the number of times each customer has rented a movie from the Action category.
    Expected output:
    Customer id, First name, Last name, Action rentals
    102, John, Smith, 5
    205, Peter, Griffin, 7

30. Find the films that have been rented by more than 20 different customers.
    Expected output:
    Film id, Title, Unique customers
    305, The Chase, 25
    409, The Mystery, 22

EDA Global-Superstore

Data Cleaning & Preprocessing

1. How many missing values are present in each column?
2. Are there any duplicate rows in the dataset? If yes, remove them.
3. Convert Order Date and Ship Date columns to datetime format.
4. Extract the year, month, and day from the Order Date column.
5. Identify and remove outliers in the Sales column using the IQR method.
6. Calculate the shipping time for each order (difference between Ship Date and Order Date).
7. Check if there are invalid postal codes (non-numeric values).
8. Identify the top 5 most frequent and least frequent Customer ID values.
9. Standardize column names by replacing spaces with underscores.
10. Find and replace any inconsistent values in the Category and Sub-Category columns.
11. How many missing values are present in each column?
12. If missing values exist, should we fill them with the mean, median, or mode? Test different approaches.
13. If missing values exist in categorical columns, should we use the most frequent value or create a separate category?
14. Check if missing values in Ship Date correspond to orders that were canceled.
15. Identify cases where Postal Code is missing and check if they belong to a particular region.
16. Are there any duplicate rows in the dataset? If yes, remove them.
17. Check if Order ID has duplicate entries. Should multiple rows per Order ID exist (e.g., multiple items in one order)?
18. Ensure that Customer Name is consistently formatted (e.g., remove extra spaces, standardize capitalization).
19. Detect and correct misspelled city names in the City column (e.g., "Newyork" vs. "New York").
20. Identify and replace any inconsistent values in the Category and Sub-Category columns.
21. Convert Order Date and Ship Date columns to datetime format.
22. Convert the Postal Code column to an appropriate data type (integer or string).
23. Convert the Sales column from string to float if needed.
24. Ensure that Customer ID and Product ID are treated as categorical variables rather than numerical values.
25. Extract the year, month, and day from the Order Date column.
26. Check if any orders have a Ship Date earlier than the Order Date.

27. Identify orders that took more than 30 days to ship. Are they concentrated in specific regions or categories?
28. Calculate the shipping time for each order (difference between Ship Date and Order Date).
29. Create a new column indicating whether an order was shipped on the same day it was placed.
30. Find the earliest and latest order dates in the dataset.
31. Identify and remove outliers in the Sales column using the IQR method.
32. Create a column that categorizes sales into buckets (e.g., Low, Medium, High) based on quartiles.
33. Identify if any Order ID has different Customer Names associated with it (potential data entry error).
34. Calculate the total number of orders placed per Customer ID.
35. Create a new column Order Month-Year by combining the extracted month and year from Order Date.
36. Calculate the average number of days taken to ship products per Region.
37. Standardize column names by replacing spaces with underscores.
38. Check if Product ID and Product Name have a one-to-one mapping (i.e., ensure there are no mismatches).
39. Identify the top 5 most frequent and least frequent Customer ID values.
40. Check if there are invalid postal codes (non-numeric values).

Exploratory Data Analysis (EDA)

1. What is the total number of unique Order IDs?
2. What is the range of sales values (min, max, mean, median)?
3. How many unique products are there in the dataset?
4. What is the most frequently purchased product?
5. What is the most profitable product category?
6. What are the top 10 cities with the highest sales?
7. What percentage of total sales comes from each region?
8. How does the number of orders vary by Ship Mode?
9. Find the top 5 customers who made the highest purchases.
10. What is the sales trend over the years?
11. What is the total number of unique Order IDs?
12. How many unique customers are there in the dataset?
13. How many unique products are available?
14. What are the most and least frequently purchased products?

15. What are the most and least profitable products?
16. What is the total revenue generated?
17. What is the range of sales values (min, max, mean, median)?
18. What is the total profit generated?
19. What is the overall profit margin (profit-to-sales ratio)?
20. What are the top 10 cities with the highest sales?
21. What percentage of total sales comes from each region?
22. Which region generates the highest and lowest profits?
23. What is the distribution of sales values (histogram)?
24. What is the distribution of profit values? Are there outliers?
25. What is the distribution of order quantities?
26. How does the number of orders vary by Ship Mode?
27. Which Ship Mode has the fastest and slowest average shipping time?
28. What is the average shipping time for each region?
29. Find the top 5 customers who made the highest purchases.
30. What percentage of total revenue comes from the top 10% of customers?
31. What is the sales trend over the years?
32. Are there any seasonal patterns in sales?
33. Which months have the highest and lowest sales?
34. How do sales vary across different days of the week?
35. How does sales distribution vary by customer segment?
36. What is the total revenue and profit contribution of each product category?
37. Which product categories have the highest and lowest profit margins?
38. Which Sub-Category generates the most and least revenue?
39. What is the most common order size (number of items per order)?
40. How does discounting impact total sales?
41. What is the relationship between discount percentage and profit?
42. Are there products that are frequently sold at high discounts?
43. Identify the top 5 states with the highest and lowest total profits.
44. Which states have the longest and shortest average shipping times?
45. What percentage of orders were shipped late?
46. How does customer purchase behavior change over time?

47. Which cities have the highest number of returning customers?
48. Identify patterns in large-value orders (orders above the 90th percentile).
49. Which customer segment contributes the most to total revenue?
50. What is the correlation between order quantity and total sales?
51. How does average order value differ between new and returning customers?
52. What percentage of total sales come from repeat customers?
53. Which products have the highest number of repeat purchases?
54. Are there specific cities where certain categories are more popular?
55. How do average sales differ between different regions?
56. What is the proportion of different Ship Modes used?
57. What is the average discount applied per category?
58. How do sales differ before and after applying a discount?
59. Which months have the highest number of high-value orders?
60. Are there correlations between order size, sales, and profit?

Sales Performance Analysis

1. What is the total revenue generated from each Category?
2. What is the average sales amount per order?
3. Which Sub-Category contributes the most to total sales?
4. Find the sales distribution by Region and State.
5. What is the yearly sales growth percentage?
6. Identify the peak sales month for each year.
7. Which state has the highest average order value?
8. How many orders were placed in each quarter of the year?
9. Identify the top 3 months with the highest and lowest sales.
10. Which Customer Segment generates the most revenue?

Shipping & Delivery Analysis

1. What is the most commonly used Ship Mode?
2. Find the average shipping time per Ship Mode.
3. Identify the states with the fastest and slowest average shipping times.
4. What is the correlation between Sales and Shipping Time?
5. Compare the shipping times for different regions.
6. Find the number of delayed shipments (Ship Date > Order Date + 3 days).

7. What percentage of orders were delivered within 2 days?
8. Identify the states with the highest proportion of delayed shipments.
9. Find the most common shipping time per Product Category.
10. Analyze the relationship between Ship Mode and Order Quantity.

Visualizations & Business Insights

1. Plot a time series graph for monthly sales trends.
2. Create a bar chart for sales across different Regions.
3. Generate a pie chart showing sales contribution by Category.
4. Create a heatmap to visualize the correlation between numerical columns.
5. Plot a box plot to analyze sales distribution across different Customer Segments.
6. Generate a histogram to analyze the distribution of Sales.
7. Create a violin plot to compare Sales across different Ship Modes.
8. Visualize sales trends per Category using a line plot.
9. Create a scatter plot of Sales vs. Order Date to observe sales fluctuations.
10. Build a bar chart showing the top 10 most sold products.