Final Project – Sales Analysis

Salma Ahmed

15/07/2024

Phase 1: Data Exploration & Understanding (SQL)

Objective: Gain initial insights into the data structure and customer purchase information

* Data Importing was through uploading the CSV file provided on sqliteonline.com then data exploration started
* Insights from data exploration and understanding:
* It’s a sales dataset for more than one country, it combines lots of different cities throughout the years of 2003 till 2005.
* It’s focused on the selling of vehicles like cars, trains and motorcycles. Total of 7 different categories.
* Each customer made more than one order, ordering different stuff. Total of 92 unique customer and 307 invoices.
* The order status is not only shipped but also there is in process, disputed or even cancelled.
* Total sales generated throughout the whole dataset is $ 10,032,629. (assuming that the transactions are made in US dollars.)
* Link for my work: <https://drive.google.com/file/d/11FhomEWtvrKcO4zIeiCZuApl4zkwhdRE/view?usp=sharing>

Phase 2: Data Cleaning & Analysis (Python)

Objective: Clean, analyze, and visualize customer purchase data.

* Started off by opening a new Google Collaboratory notebook on my Google Drive.
* Mounting my drive on to the notebook for importing the CSV file is a crucial step.
* I imported the libraries than I’m going to need in my analysis while working which were [pandas, NumPy and matplotlib].
* Data Cleaning:
* I removed all rows that were blank if any.
* Removed as well the unnecessary columns that I did not want in my analysis like [‘ADDRESSLINE2', 'CITY', 'TERRITORY' and 'POSTALCODE’] and others as well.
* Data Analysis:
* I created a new column called ‘Customer Segment’, that gives the value of “High-Value” to customers with sales more than $ 2000 and “Low-Value” to customers with sales less than $ 2000.
* I calculated the purchase frequency for each customer by counting the ‘ORDERNUMBER’ for each customer and I inserted it as a column in the dataset for further analysis.
* Another column is created which is ‘CLTV’ which stands for Customer Lifetime Value.

I calculated the ‘CV’ which is the Customer Value by multiplying the ‘ORDERLINENUMBER’ and the price for each customer. Then the average frequency was taken from the previous column made which was the ‘frequency’. Lastly, to calculate ‘CLTV’ I multiplied the ‘CV’ to the average frequency.

* Last thing made to analyze the dataset is to see the top 10 products sold. I did it by grouping the data by the ‘PRODUCTCODE’ and getting the top 10 sold products.
* Data Visualization:
* Matplotlib was used as the library for visualizing the findings in this dataset.
* I plotted monthly and quarterly sales using a line graph and made some changes to customize it to my own preference.
* The second graph was a bar chart plotting the top 10 products sold in descending order.
* I plotted the total sales generated each year in a bar chart, which resulted in showing that 2004 had the most sales from all three years.
* I used a pie chart to visualize the product categories as percentage from total sales, which showed us that the most wanted category is the classical cars which was 39.1% of the total sales.
* The next graph was a line graph showing the average quantity ordered for the three years.
* A bar graph was used to see which customer segment ordered the most and the result was that the ‘High-Value’ Customers are way more than the ‘Low-Value’ Customers.
* The order status was visualized is a bar chart as well that showed us that most of the orders were shipped.
* Lastly, I used a pie chart again to plot the deal size as a percentage of sales. It showed that Medium deals took 49% of the total sales.
* Last step was exporting the new prepared dataset as a csv file to move to the next phase.
* Link to my work: <https://colab.research.google.com/drive/1Sk-XpO7py5xO_6TYMzTuacXplMrutstW?usp=sharing>

Phase 3: Data Visualization & Storytelling (Power BI)

Objective: Create compelling visuals and communicate data-driven insights.

* I used 3 slicers in my report [ 1. Quarters for each year, 2. Years as a drop-down list and 3. A checklist for the countries] to customize the report as the user wants.
* A clustered bar chart is used to graph the average frequency and the average quantity ordered for each product category which shows us that as previously seen in SQL and python that the most sold products category is the classical cars.
* A donut chart is used to see the status for orders made.
* A bar chart is used to see the total sales per year.
* The pie chart is like the one in the previous step with the deal size.
* A ribbon chart is used to visualize sales per customer segment.
* Link to my work: <https://drive.google.com/file/d/1qwwOPTUqaUcNZ-RRqbllWRpN_fmXsIcJ/view?usp=sharing>