**ANALYSIS OF CAR SALES DATA**

**TEAM MEMBERS:**

* Adnan Shareef
* Praveen Reddy Kadasani
* Aseef Basha Shaik
* Yog Chaudhary
* Nishith Reddy Mannuru

**About the dataset:**

The dataset was gathered from Kaggle**.** [LINK](https://www.kaggle.com/datasets/suraj520/car-sales-data)

The dataset contains information about the sale of different cars. To make the dataset more precise we are considering the data for the car year from 2022 – 2023. This information can help study car sales patterns over time. It can also help assess how well each car salesperson in the dealership is doing. The data can be used to predict future sales and understand how different things impact car sales, like the type of car, model, year, and salesperson. Additionally, analyzing metrics like average sale price and commission rates by vehicle provides insights into profitability differences across car types to guide inventory investments. Identifying repeat versus one-time customers informs loyalty program decisions. Examining monthly and yearly sales trends enables better alignment of operations to seasonal volume swings. Together this detailed dataset allows both long-term strategy planning around optimal inventory mix, pricing, and sales team staffing as well as short-term tactical forecasts of demand by car segment. The granular sales data gives a comprehensive 360-degree view of the dealership’s true sales productivity, customer value, and market competitiveness.

|  |  |
| --- | --- |
| **Attributes** | **Description** |
| Date | The date of the car sale |
| Salesperson | The name of the salesperson who made the sale |
| Customer Name | The name of the customer who purchased the car |
| Car Make | The make of the car that was purchased |
| Car Model | The model of the car that was purchased |
| Car Year | The year of the car that was purchased |
| Sale Price | The sale price of the car in USD |
| Commission Rate | The commission rate paid to the salesperson on the sale |
| Commission Earned | The amount of commission earned by the salesperson on the sale |

**Problem statement:**

The sales manager aims to conduct a thorough analysis of sales data across multiple dimensions in order to derive insights that can inform better business strategies and incentive programs. Having detailed information on salesperson performance, customer preferences and purchase patterns, and pricing trends by vehicle model will enable more targeted, data-driven decision-making to maximize dealership sales and earnings.

For example, identifying which salespeople generate the highest volume and commission earnings can guide recruiting and training investments in stellar performers. Determining which customers purchase the most frequently or spend the most money could shape customer loyalty and incentive programs to cultivate these valuable patrons. Tracking sales volumes and profit margins over time by vehicle make and model would allow dynamic adjustment of purchasing decisions, pricing policies, and sales team commissions in response to demand trends and profit potential. Diving into metrics across key categories provides focused intelligence to continually optimize operations, sales contests, vehicle stocking plans, and staff rewards relative to current realities. By continuously monitoring this data and response metrics, the dealership will gain the opportunity to experiment, assess results, and evolve their strategies using analysis rather than guesswork.

In summary, comprehensive, and ongoing analysis of granular sales data will enable proactive, informed decisions to boost dealership performance.

**Data Preprocessing**

**STEP 1:** We loaded the dataset into OpenRefine.

A screenshot of a computer

Description automatically generated

We can then see that a total of 241205 records are loaded.

A screenshot of a computer

Description automatically generated

**STEP** **2**: Analyzing the data

* Checking for unique values in Car make using Text Facet.

A screenshot of a computer

Description automatically generated

* You can see there are 5 different car makes in the dataset

A screenshot of a computer

Description automatically generated

**STEP 3:** Checking for missing values

* We’ve used the Facet Panel from the drop-down arrow next to the column header to check the missing values.
* In the facet menu, I’ve selected the "Customized facets" option and selected “Facet by Blank”. I’ve checked all the columns.
* On the left side, you can see that the facet will now display the number of blank or missing values in that column. You can see the number of missing values for each facet as shown below for all columns. As we can see there are no missing values.

A screenshot of a computer

Description automatically generated

A screenshot of a chat

Description automatically generated A screenshot of a computer

Description automatically generated

**STEP 4**: Checking for Duplicate values

* We’ve checked for any spelling mistakes and unique words, that may be spelled wrong through the Facet and Text Facet option.

A screenshot of a computer

Description automatically generated

**Type of Storage Used**

Google Cloud Platform's (GCP) BigQuery was used to analyze the car sales data. BigQuery would be an optimal data warehouse solution for analyzing this automotive sales dataset to uncover actionable insights. Specifically, BigQuery is a serverless, highly scalable cloud data warehouse that enables running SQL queries over enormous datasets, in the petabyte range and beyond (Bhaumik, 2022). This makes it ideally suited for performing fast, complex analysis on large volumes of granular sales transaction information aggregated over time to guide data-driven business strategy decisions through a user-friendly SQL interface (Ali et al., 2021). It enables running SQL queries on huge amounts of data quickly and cost-effectively.

Reasons BigQuery is well-suited for the dealership's needs:

**Performance** - Advanced capabilities ensure fast SQL queries, even on billions of rows of granular sales transactions over time(Cheng, 2022). This allows rapid analysis.

**Cost** - As a serverless platform, BigQuery is very cost efficient compared to traditional data warehouses because you only pay for what you use (Masmoudi, 2021). This saves money for the analysis.

**Security** - GCP meets major compliance standards like HIPAA for data privacy assurance when dealing with sensitive customer sales data (Ignyte Team, 2020).

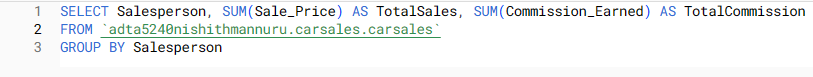
**Accessibility** - The SQL interface facilitates complex sales analytics without advanced data science expertise. Integrations with data visualization tools also enable interactive dashboards.

In summary, BigQuery grants the dealership a flexible, affordable way to generate data-driven insights on inventory needs, customer trends, and sales team effectiveness to enhance strategic planning in a secure cloud platform. The savings and performance make it simpler than traditional warehouse approaches. Therefore GCP’s BigQuery would be an optimal data warehouse solution for analyzing this automotive sales dataset to uncover actionable insights. This makes it ideally suited for performing fast, complex analysis on large volumes of granular sales data to guide data-driven strategy decisions to improve inventory management, pricing structure, sales compensation, and promotional initiatives in a secure and compliant cloud environment.

**QUERIES**

**Query 1**: **Total Sales and Commissions by Salesperson**

In order to evaluate how effectively each individual sales representative converts opportunities into sales and commissions earned, we will write a query to aggregate performance data by representative over a set time period. By selecting the salesperson's name, summing the total dollar amounts of their closed sales, and summing the total commissions they have earned across all transactions, we can view both total sales revenue generated and total commissions made for each person. Comparing these totals relative to their peers allows sales management to quickly identify top performers to replicate success as well as lower performers who may require additional training, mentoring, or transition to other roles to optimize the overall output of the sales organization. This data is pivotal for informing critical management decisions.



A screenshot of a data report

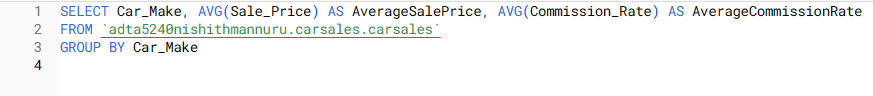
Description automatically generated

A graph of a number of people

Description automatically generated with medium confidence

**Query 2: Average Sale Price and Commission Rate by Car Make**

The goal of this query is to analyze differences in sale prices and commission rates by vehicle make. Grouping sales data by car make allows calculating the average sale price and average commission rate per make. This reveals pricing and profit trends across different vehicle types sold by the dealership. Some makes likely allow for higher sale prices and commissions. Understanding which makes currently perform better allows making data-driven decisions to improve future profitability. This includes optimizing inventory purchases to focus on high-selling makes and models, adjusting pricing and incentives on certain vehicle types, and negotiating improved manufacturer commission rates on top-selling makes. The analysis helps maximize profits by aligning sales, inventory, and manufacturer relationships with customer vehicle preferences and purchasing trends.

****

**A screenshot of a chart

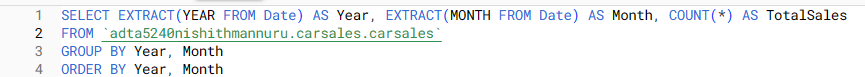
Description automatically generated**

A screenshot of a computer

Description automatically generated

**Query 3: Monthly Sales Trends**

This query will break down the last 12 months of sales data by month to reveal trends over the 1-year period. Extracting the month attribute from the date field enables segmenting records into monthly buckets. Counting total monthly sales shows if there are cyclical demand swings - certain months may see recurring spikes or dips. For a single year, this specifically highlights seasonality around summer driving, holidays, or year-end clearances. The granular month-by-month analysis tracks volumes to uncover why sales rise or fall at certain times annually. Visualizing these demand cycles and understanding unique monthly factors allows better inventory planning, sales staffing, promotions, and manufacturer ordering to align operations with predicted volumes. Even a year of historical data reveals seasonality potentials to optimize the business for the upcoming summer surge or year-end wind down based on last year's trends.



A screenshot of a table

Description automatically generated

A graph with a line

Description automatically generated

**Query 4: Top-Selling Car Models**

Analyzing which specific car models sell the most units can optimize future inventory investments. Grouping sales records by model and counting total sales per model reveals model-level demand. Ordering models descending by sales volume ranks models from best to worst sellers. Taking the top 5 highest-volume models shows the most popular purchases. Increased access to hot sellers boosts both customer satisfaction from model availability and business revenue from inventory aligned to proven sales trends. Tracking which model variants are fastest-moving improves cash flow by ensuring displayed inventory expediently converts to sales.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A pie chart with numbers

Description automatically generated

**Query 5: Sales Breakdown by Car Year**

Analyzing sales performance by model year reveals trends in the relative demand for new versus used vehicles. Grouping sales records by model year and summing total sale price per group shows the sales contribution of each vehicle age segment. The comparison could reveal that brand new model year vehicles command the highest total sales revenue due to higher pricing. Or alternatively used vehicles contribute greater overall profit if the pricing difference exceeds volume gains from new cars. Understanding preferences for vehicle age guides trade-in offers, pre-owned vehicle acquisition at auction, and new model year inventory purchasing to optimize cash investments against validated sales demand across the age spectrum.

A screenshot of a computer

Description automatically generated

A screenshot of a calendar

Description automatically generated

A graph with blue lines

Description automatically generated with medium confidence

**Query 6: Customer Loyalty**

Analyzing sales data to identify instances where the same customer has purchased multiple vehicles over time reveals loyal repeat buyers. These customers likely enjoy their ownership experience, have ongoing vehicle needs, or remain brand devotees. Understanding who our best customers are who continuously return allows for creating targeted loyalty programs with special perks for continued patronage. Additional retention tactics like service discounts and trade-in bonuses could incentivize further repeat purchases. Idealizing this viral cycle of repeat sales & rewards boosts customer lifetime value. Retaining loyal buyers is more profitable long term than constantly acquiring new customers.

A screenshot of a computer

Description automatically generated

A screenshot of a table

Description automatically generated

A graph of a number of people

Description automatically generated with medium confidence

**Query 7: Car Make Popularity Over Time**

Tracking vehicle preference shifts by customers over the years guides business strategy adaptations and future decision-making. The query groups total sales records by both car make and year extracted from the sale date. This structures data to analyze make-level demand annually. Identifying consistent high sales validates further development of dealer relationships and incentives with their manufacturers. Adapting operations based on popularity changes driven by consumer preference shifts ensures showroom inventory, technician training, manufacturer ties, and sales efforts align with ever-evolving market dynamics for sales growth.

A close-up of a website

Description automatically generated

A screenshot of a computer

Description automatically generated

A graph of blue and orange bars

Description automatically generated

**REFERENCES:**

Ali, M. H., Hosain, M. S., & Hossain, M. A. (2021). Big Data analysis using BigQuery on cloud computing platform. *Australian JofEng Inno Tech*, *3*(1), 1-9.

Bhaumik, S. (2022, November 1). BigQuery: Basics & Internals: Part # 1 [Medium]. *Google Cloud - Community*. https://medium.com/google-cloud/bigquery-basics-internals-part-1-9eda52b65960

Cheng, C. (2022). *14 Best Practices to Tune BigQuery SQL Performance*. Medium. https://towardsdatascience.com/14-ways-to-optimize-bigquery-sql-for-ferrari-speed-at-honda-cost-632ec705979

Ignyte Team. (2020, January 5). *Platform Compliance & Risk Management*. https://www.ignyteplatform.com/platform-compliance-risk-management/

Masmoudi, A. (2021, May 27). BigQuery Cost Efficiency. *Medium*. https://datadice.medium.com/bigquery-cost-efficiency-a6223b46376c