Business Insights from Car Sales Data

Introduction

In today's competitive automotive market, understanding customer behavior and regional performance is crucial for sustaining growth and optimizing sales strategies. This project aims to provide actionable business insights through a comprehensive data analysis of car sales records across multiple regions and customer demographics.

The dataset used in this project was obtained from **Kaggle**, a popular online platform for data science and machine learning datasets. It includes detailed information such as customer income, gender, dealer region, car brand, body style, and price, offering a robust foundation for business analysis.

Using Python within a Google Colab environment, this analysis applies data cleaning, visualization, statistical exploration, and predictive modeling techniques to explore:

- Revenue distribution across companies and regions
- Patterns in car model preferences based on income and gender
- Key factors influencing customer purchase decisions
- Feature importance in predicting brand category choice
- Strategic opportunities for dealership optimization

This end-to-end analysis simulates the type of value-driven insights expected from a business analyst in a consulting setting, such as those at Big 4 firms. The findings are supported by visualizations and practical recommendations aimed at enhancing data-driven decision-making.

Data Overview

The dataset used for this analysis contains structured information on customer car purchases. It includes **20 columns** and **several thousand rows** representing individual transactions. The features span across demographic data, dealership details, and vehicle specifications.

Key features in the dataset

Column Name	Description
Car_id	Unique identifier for each car transaction
Date	Purchase date
Customer_Name	Name of the buyer
Gender	Gender of the customer
Annual_Income	Exact yearly income of the customer (USD)
Dealer_Name	Name of the dealership
Company	Car manufacturer (e.g., Toyota, BMW)
Model	Car model name
Engine	Engine specifications (e.g., 2.0L Turbo)
Transmission	Transmission type (Automatic/Manual)
Color	Color of the car
Price_$	Price of the car in USD
Dealer_No	Unique dealership ID
Body_Style	Type of vehicle (e.g., SUV, Sedan)
Phone	Contact number of the customer
Dealer_Region	Region where the dealership is located
Income_Segment	Categorized income group (Low/Mid/High)
Segment	Customer class/segment
Month	Month of the purchase (extracted from Date)
Brand_Category	Simplified brand classification (e.g., Economy, Premium)

Data Preparation and cleaning

Before diving into the analysis, the dataset underwent essential data preparation and cleaning steps to ensure accuracy and consistency. After importing the data into a Google Colab environment using Python, I conducted an initial inspection to understand the structure, detect missing values, and review data types. Minimal missing values were addressed appropriately. To enhance business relevance, I performed feature engineering — such as extracting the purchase month from the date, categorizing annual income into segments (Low, Mid, High), and creating a simplified brand category (Economy, Mid-range, Luxury). I also renamed key columns to improve clarity, like changing Annual_Income to Exact Income (USD) and Income_Segment to Income Category (Low/Mid/High). These steps ensured the dataset was clean, interpretable, and ready for meaningful business analysis and modeling.

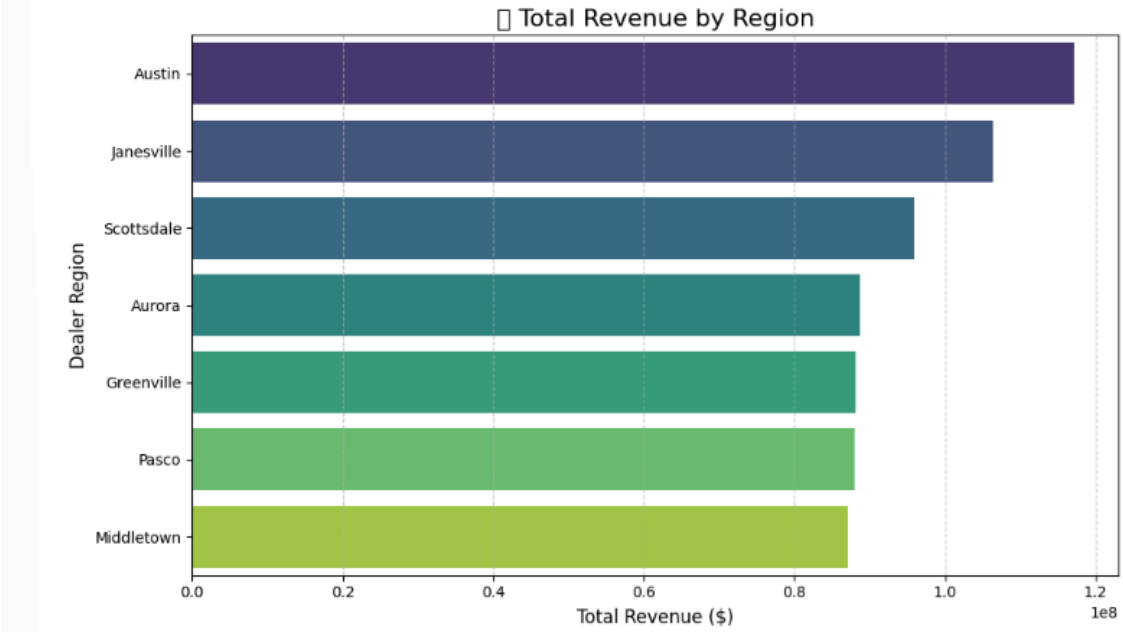
Company analysis

- **Total revenue by region**

To begin the business analysis, we evaluated revenue performance across different dealership regions. Using the `groupby()` function on the `Dealer_Region` column and summing the corresponding `Price_$`, we ranked each region by total revenue.

Dealer_Region	Price_$
Austin	117192531
Janesville	106351234
Scottsdale	95969374
Aurora	88687382
Greenville	88149602
Pasco	88040714
Middletown	87134628

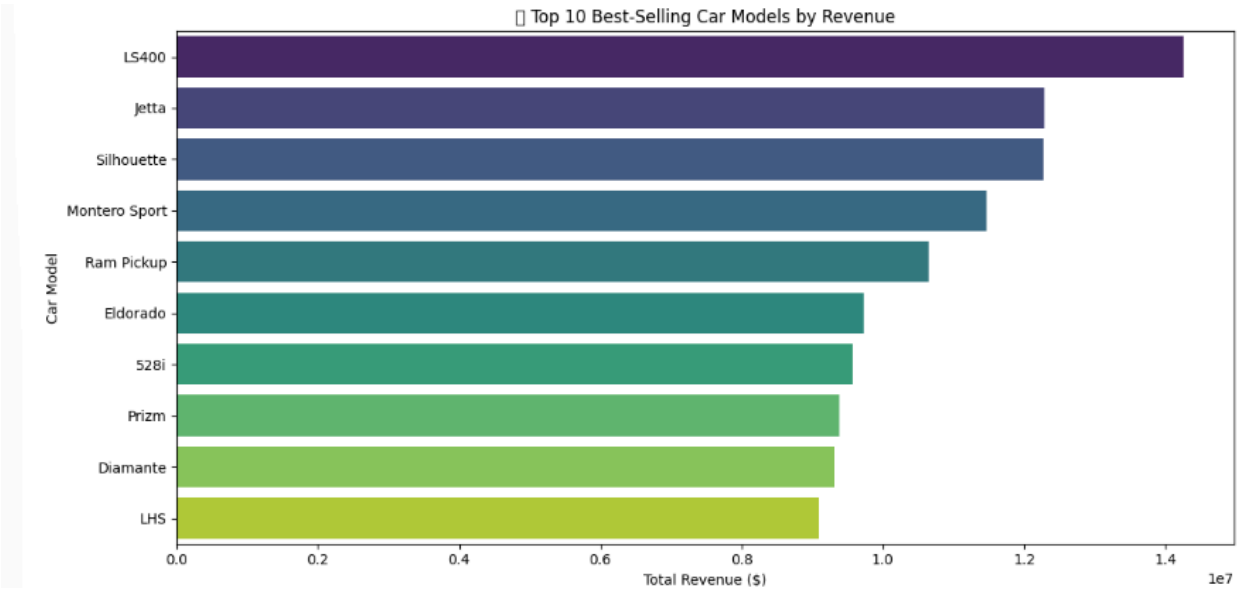
dtype: int64



The analysis revealed that Austin generated the highest revenue among all dealer regions, reaching over $117 million in sales. This suggests that Austin is a high-performing location, likely due to a combination of strong customer demand, an effective dealer network, and potentially higher car prices or volume.

Regions such as Janesville and Scottsdale also demonstrated robust performance, making them key contributors to overall revenue. On the other hand, although Middletown has a slightly lower figure, its performance is still relatively strong and could be boosted further with targeted promotions or operational improvements.

- **Top 10 Best-Selling car models by Revenue**



After analyzing the performance by region, I wanted to explore which specific car models were generating the highest revenue across the dataset. Understanding which models contribute the most to total revenue can guide strategic decisions in marketing, inventory, and dealership focus.

To do this, I grouped the dataset by the `Model` column and calculated the total revenue (`Price_$`) for each. I then visualized the top 10 models using a horizontal bar chart, which clearly displays how each model performs relative to the others in terms of revenue generation.

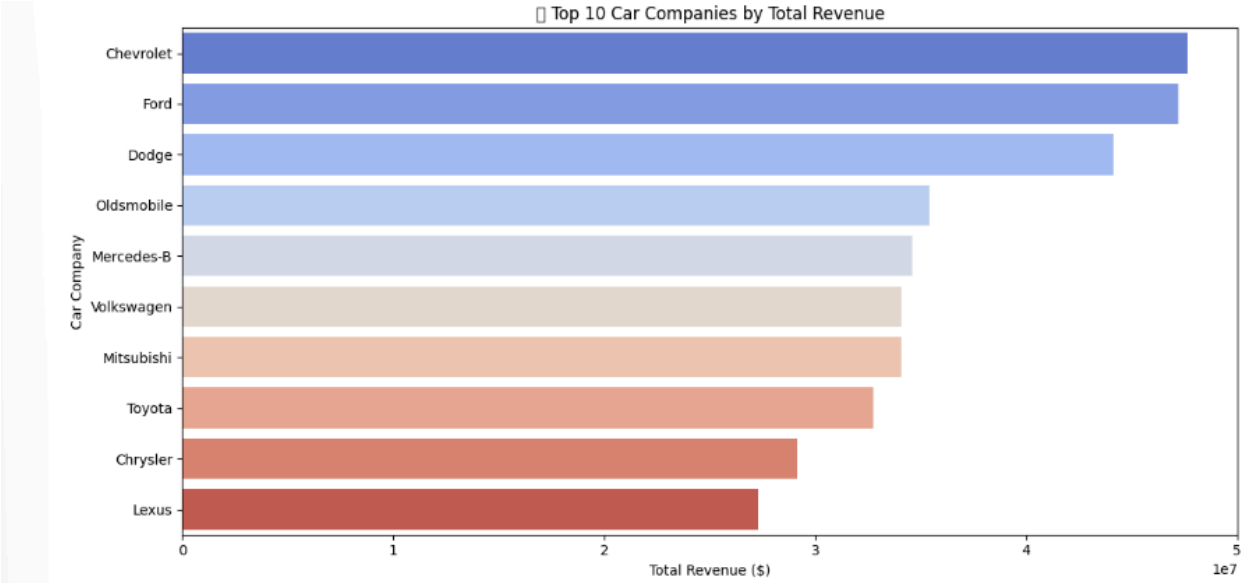
LS400 stands out significantly as the highest revenue-generating car model, outperforming all others by a notable margin. This indicates that the LS400 not only sells well but likely also has a high unit price.
Jetta and Silhouette follow closely behind, generating comparable revenues. Their performance highlights strong customer demand and suggests these models are consistent top-sellers.
Models such as Montero Sport, Ram Pickup, and Eldorado form a solid middle tier. These vehicles likely appeal to a broad segment of customers, balancing popularity with reasonable pricing.
The bottom three of the top 10528i, Prizm, and LHS while still among the best performers, generate slightly lower revenue compared to the others. Their inclusion in the top 10, however, confirms they are still vital contributors to overall sales.

This analysis reveals which models dealers should prioritize in terms of marketing and stock availability. The LS400, being the leader, may benefit from premium placement and promotional campaigns to further capitalize on its popularity. At the same time, strong performers like Jetta and Silhouette could be leveraged in competitive pricing or bundling strategies.

Additionally, this insight can help refine future procurement decisions and adjust regional focus if specific models align with high-performing areas. A deeper analysis could involve cross-referencing model popularity with region or customer demographics to uncover more granular insights.

- **Top 10 Car companies by Total Revenue**

To gain a high-level understanding of brand performance, I analyzed total revenue generated by each car company in the dataset. By grouping the data by the `Company` column and summing the `Price_$`, I was able to identify which companies were the most profitable across all dealerships.



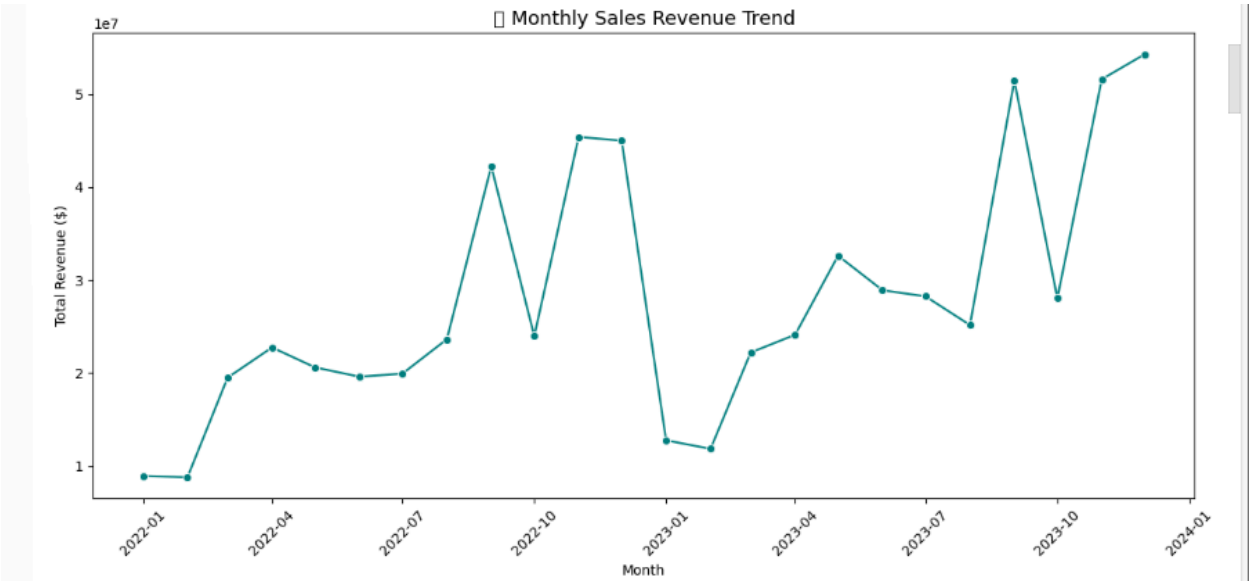
The visualization above highlights the top 10 car manufacturers based on cumulative sales revenue. Chevrolet stands out as the top-performing brand, closely followed by Ford and Dodge. These three companies clearly dominate the market in terms of revenue generation, each earning well over 40 million dollars.

Oldsmobile and Mercedes-Benz also show strong performance, occupying the fourth and fifth positions respectively. Interestingly, the companies ranked from sixth to tenth Volkswagen, Mitsubishi, Toyota, Chrysler, and Lexus are relatively close in revenue, suggesting a competitive mid-tier segment.

This ranking provides valuable insights for dealerships and business strategists. High-performing companies might be prioritized in future sales strategies, inventory planning, and marketing investments, while lower-performing ones may need promotional support or reevaluation in certain regions.

- **Monthly sales Revenue Trend**

To better understand sales dynamics over time, I analyzed the monthly sales revenue trend. I converted the transaction dates into a standardized month format and aggregated the total revenue by month. The resulting line chart illustrates the fluctuations in monthly revenue across the dataset's time range.



From the plot, several key observations emerge:

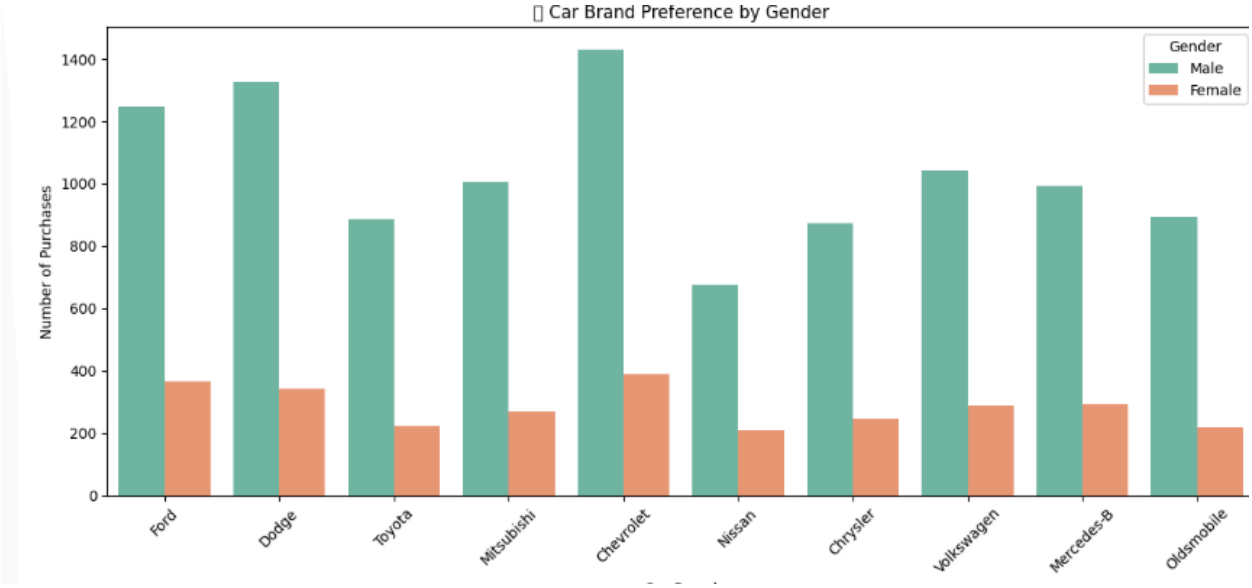
- **Seasonal Peaks:** Notable revenue spikes occur around October to December 2022 and October to December 2023. This suggests a seasonal sales pattern, potentially linked to end-of-year promotions, holiday shopping behavior, or new model releases.

- **Growth Trend:** While fluctuations exist, there's a noticeable upward trend in revenue as we move from early 2022 to late 2023. The final months of 2023 particularly November and December record the highest sales figures of the entire period, each surpassing $50 million in revenue.

- **Dips in Sales:** January and February of both years show significantly lower sales, indicating a potential post-holiday sales slump or reduced inventory cycles at the beginning of the year.

- **Car Brand Preference by Gender**

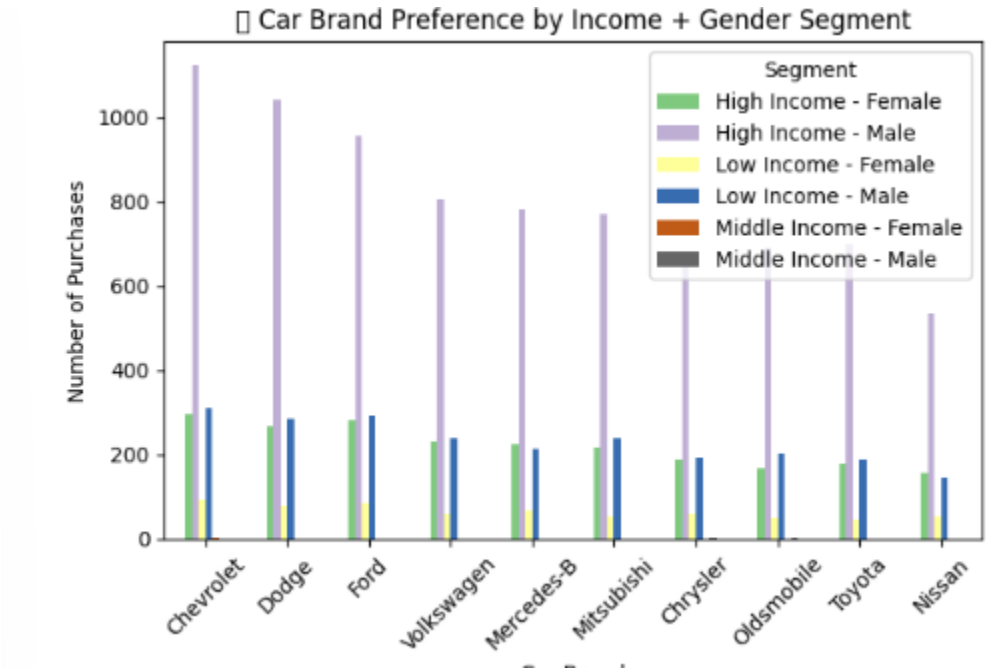
To uncover gender-based trends in car brand selection, I created a grouped bar chart showing the number of purchases made by males and females across the top-selling automotive brands. This comparison provides a lens through which marketing strategies can be tailored to better engage each demographic group.



Male Buyers Dominate:
Across all brands, male customers significantly outnumber female customers in terms of car purchases. Chevrolet, Dodge, and Ford have the highest male buyer counts, each exceeding 1,200 purchases. This dominance may reflect broader market trends, dealership engagement strategies, or decision-making dynamics in households.
Female Preferences:
While female purchase volumes are generally lower, brands such as Chevrolet, Ford, and Volkswagen see relatively higher engagement from female buyers compared to others.
The gap is notably smaller for Volkswagen and Mercedes-B, suggesting these brands may already be resonating better with female customers.
Balanced Appeal:
Although no brand exhibits parity, Chevrolet and Ford show the most balanced gender reach, with females purchasing over 350 units, indicating their broad brand appeal and trustworthiness across demographics.

- **Car Brand Preference by Income + Gender Segment**

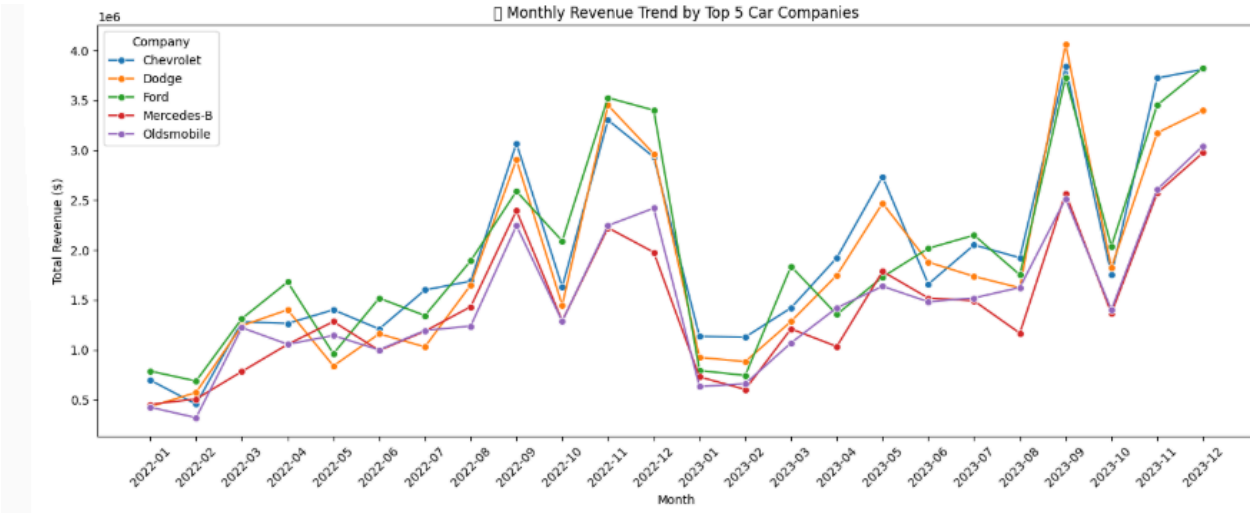
To gain deeper insights into customer behavior, I analyzed the intersection of income level and gender in relation to car brand preferences. This multidimensional segmentation helps in understanding how purchasing decisions vary across distinct socio-demographic groups.



The results reveal that low-income male customers make up the largest portion of buyers across all brands, particularly for Chevrolet, Dodge, and Ford. This suggests that these brands resonate strongly with this demographic, potentially due to affordability or perceived reliability. In contrast, female buyers in all income segments show noticeably fewer purchases, though high-income females demonstrate relatively stronger interest in brands like Chevrolet and Volkswagen. Interestingly, middle-income buyers, both male and female, are significantly underrepresented, which may suggest a gap in product-market fit or marketing outreach. These insights indicate an opportunity for brands to expand their appeal by tailoring marketing strategies toward underserved segments especially females and the middle-income group while continuing to strengthen their positioning among their core low-income male customer base.

- **Monthly Revenue Trend by Top 5 Car companies**

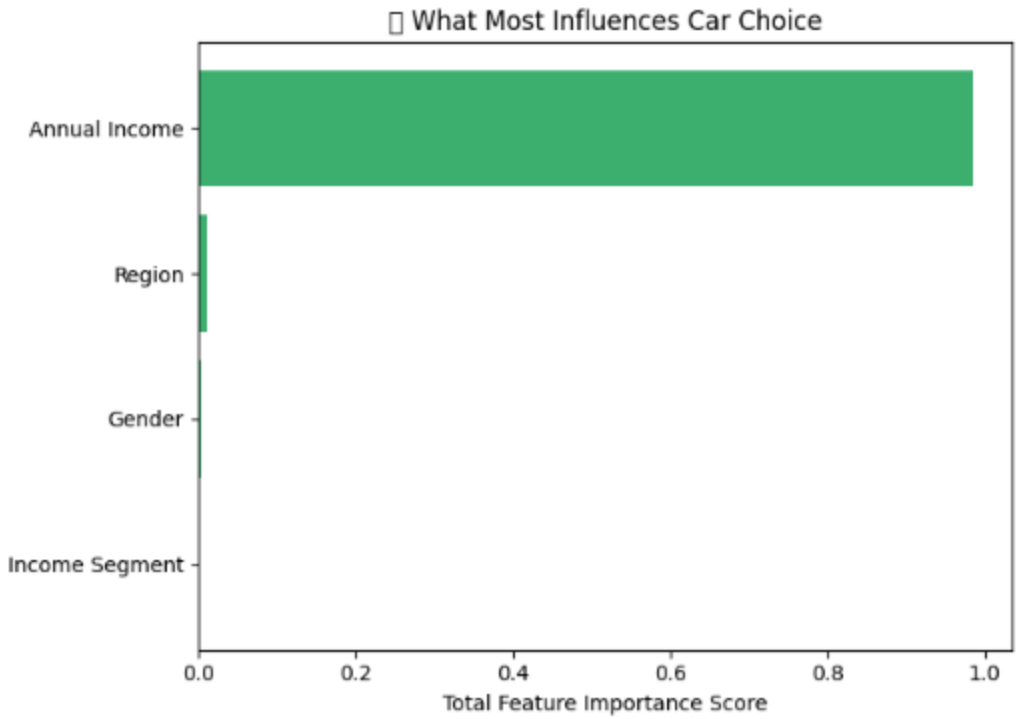
To better understand performance over time, I examined the monthly revenue trend of the top 5 car companies: Chevrolet, Dodge, Ford, Mercedes-Benz, and Oldsmobile.



The analysis reveals a consistent seasonal fluctuation, with noticeable peaks around September to November each year. These spikes may correlate with promotional periods, new model releases, or seasonal buying behavior. Among the five, Ford and Chevrolet consistently lead in total revenue, often alternating in top position. Notably, all companies experienced a sharp dip in early 2023, followed by a strong recovery in the latter half of the year—suggesting a temporary market disruption, possibly due to economic or supply chain factors. Despite some volatility, the general trend is upward, reflecting either increased sales volume or higher-value transactions. These insights can help inform future inventory planning, marketing campaigns, and strategic launches around high-performing months.

- **What most influences car choices**

To conclude the analysis, I wanted to understand what factors most significantly influence a customer's car choice. Using feature importance scores from a Random Forest model, I grouped individual dummy variables into broader categories such as Annual Income, Region, Gender, and Income Segment.



The results clearly show that Annual Income is the dominant factor driving car selection decisions, far outweighing all other variables. This highlights that customers' purchasing power is the most influential component when choosing a vehicle. In contrast, demographic attributes like Gender, Region, and Income Segment had a comparatively minimal impact.

- **Recommendations**

After analyzing the car sales dataset, I derived several strategic recommendations to support business growth. First, it is clear that annual income is a major factor influencing a customer's vehicle choice. Therefore, marketing efforts should be segmented by income levels, with luxury models being targeted specifically at high-income consumers through exclusive campaigns and premium financing offers. Additionally, regions like Austin, Janesville, and Scottsdale generated the highest revenues, suggesting that inventory allocation, marketing resources, and dealership support should be prioritized there. I also found that specific models such as the LS400, Jetta, and Silhouette were top performers in terms of revenue; these models should be consistently featured in sales promotions and value-added service bundles. From a brand perspective, companies like Chevrolet, Ford, and Dodge stood out as strong performers, and deeper collaboration with these brands could yield higher profitability. Moreover, the data shows a significant skew toward male buyers; to expand market reach, companies should consider gender-specific marketing strategies to better engage female customers. Finally, there is an opportunity to align sales campaigns with seasonal peaks and even introduce income-based recommendation systems on digital platforms to enhance personalization and drive conversions.

- **Conclusion**

This project was a complete end-to-end business analysis of a car sales dataset sourced from Kaggle. By combining the power of AI-assisted coding with my own analytical thinking, I was able to explore the dataset, extract meaningful insights, and build a predictive model from scratch using Python. I carried out thorough data cleaning, segmentation, and visual exploration, uncovering trends such as the influence of annual income on brand preference, regional revenue disparities, and seasonal buying behaviors. With AI guiding the technical implementation and my strategic perspective driving the business context, I translated raw data into actionable insights that could help optimize dealership performance, marketing strategies, and customer targeting. This experience not only enhanced my data analysis and visualization skills but also demonstrated how AI can be effectively leveraged to support impactful decision-making positioning me strongly for a role in business analytics or consulting.

References

MissionJEE. (n.d.). *Car Sales Report* [Data set]. Kaggle. https://www.kaggle.com/datasets/missionjee/car-sales-report