Devi Sree Pendyala

FINAL REPORT

An Analysis on Vehicle Sales Data

Executive Summary:

This project's primary objective was to thoroughly examine the Vehicle sales data that was accessible by utilizing the power of Exploratory Data Analysis (EDA) and reliable tools like NumPy and Pandas. Finding important insights and forecasting variables that affect vehicle sales turnover is our goal. Our team is set out to find significant patterns and trends in the information to make more informed strategic decisions about vehicles sales bases on its different specifications.

Data Pre-Processing:

Data preprocessing is an important initial step in converting raw data into a format compatible with machine learning analysis. The relevant material consists of 16 columns and 1379 rows. The main characteristics identified for analysis are year, make, model, trim, body, transmission, vin, state, condition, odometer, color, interior, seller, mmr, selling price, sale date.

Preprocessing Tasks Include:

Data Cleaning: When preparing data, data cleansing is an essential step, especially for machine learning and analytics applications. To guarantee that the dataset is reliable, consistent, and suitable for modeling, several strategies are used. Handling missing numbers, handling duplicates, fixing mistakes, and dealing with outliers are common tasks in detailed data cleaning procedures.

Analyzing exploratory data (EDA):

The most important stage in the data analysis process is visually summarizing and assessing the key features of a dataset to find trends, detect patterns, identify deviations, and test concepts. Understanding the data's underlying structure, directing the methods for more analysis, and assisting in choosing of modeling strategies and data transformations are all made feasible with the use of EDA.

Project Motivation and Background

The automobile sector is distinguished by its dynamic market conditions, evolving quickly, and adapting to changing consumer demands and technology. To improve market share and profits, stakeholders, and business owners must have a thorough understanding of pricing models.

This project's main goal is to examine this extensive dataset to glean insightful knowledge about the automotive industry. We have two objectives:

Determine Important Factors:

Our goal is to identify the crucial elements influencing car sales. This entails examining the relationships between many characteristics, including the manufacturer, model, year, and condition of the car, and sales results.

Improve Sales Strategies:

We hope to find efficient ways to increase sales of various brands by comprehending these dynamics. This could entail identifying prospective areas for feature enhancements in vehicles, modifying price plans, or refining marketing techniques considering the traits that consumers find most appealing.

Data Description

Dataset Name: Vehicle Sales Data

"Vehicle Sales and Market Trends Dataset" offers an extensive compilation of data concerning the sales transactions of different types of cars. The year, make, model, trim, body type, transmission type, VIN (Vehicle Identification Number), condition rating, state of registration, odometer reading, exterior and interior colors, seller details, Manheim Market Report (MMR) values, selling prices, and sale dates are all included in this dataset.

The Columns include:

Year- The manufacturing year of the vehicle.

Make- The brand or manufacturer of the vehicle.

Model- The specific model of the vehicle.

Trim- additional designation for the vehicle model.

Body- The body type of the vehicle (e.g., SUV, Sedan).

Transmission- The type of transmission in the vehicle (e.g., automatic).

Vin- Vehicle Identification Number, a unique code for each vehicle.

State- The state where the vehicle is registered.

Condition- Condition of the vehicle, possibly rated on a scale.

Odometer: The vehicle's distance traveled.

Color- Exterior color of the vehicle.

Interior- Interior color of the vehicle.

Seller- The entity selling the vehicle.

MMR- Manheim Market Report, possibly indicating the estimated market value of the vehicle.

Selling Price- The price at which the vehicle was sold.

Sale Date- The date and time when the vehicle was sold.

Data Transformation/Exploratory Data Analysis

- The dataset contains 1379 makes and 1378 models.
- There are 213 distinct models, 17 different colors, and 38 various make types present in this dataset.
- The most prevalent color among the entries is white.
- Model G Sedan ranks as the most common model type.
- The model with the highest frequency occurs 247 times.

```
df[['make', 'model','color']].describe(include='all')
```

	make	model	color
count	1379	1378	1378
unique	38	213	17
top	Infiniti	G Sedan	white
freq	250	247	323

The make type "***Acura***" has the highest selling price in the dataset.

```
df.groupby('make').agg({'sellingprice': 'max'}).reset_index()
```

	make	sellingprice
0	Acura	41500
1	Audi	91000
2	BMW	77250
3	Bentley	96000
4	Buick	24500
5	Cadillac	44000
6	Chevrolet	36000

In this dataset, we have 213 unique models and 17 different colors for the models.

```
df.nunique()
```

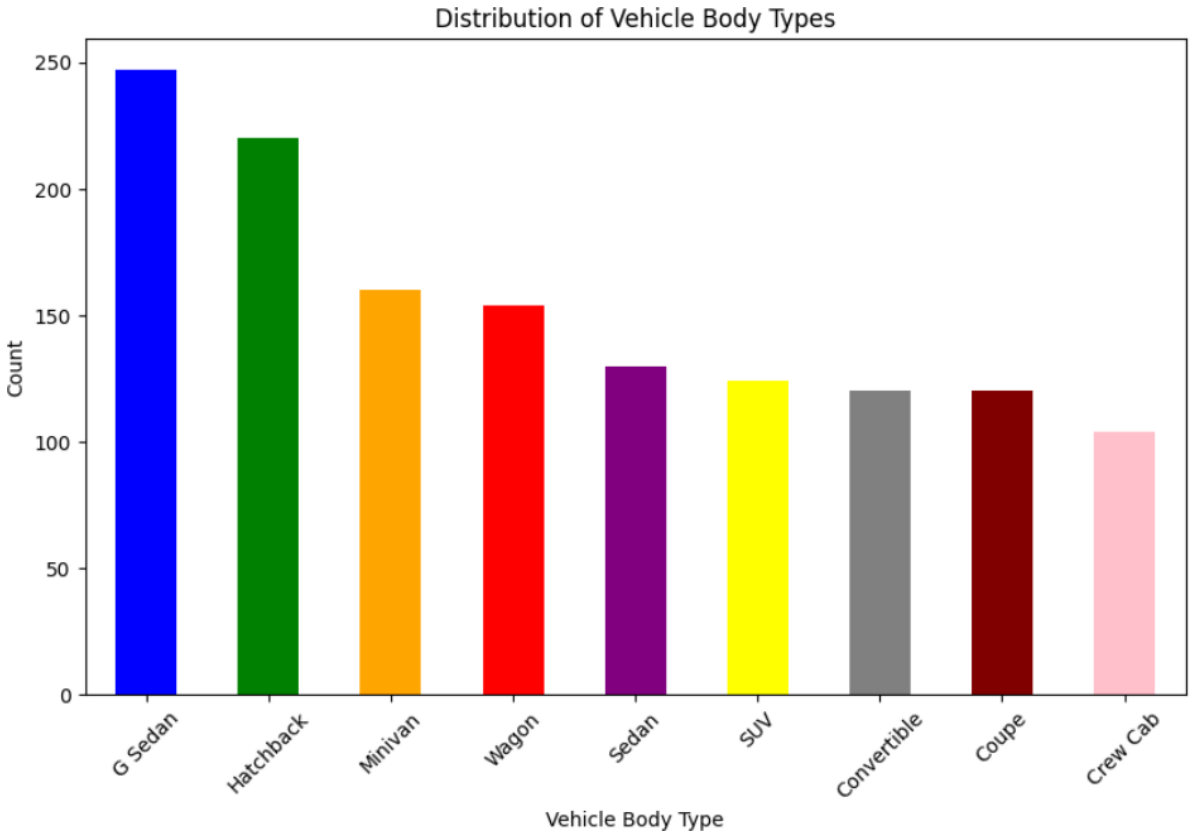
```
year               15
make               38
model             213
trim              207
body                9
transmission        2
vin              1379
state               1
condition          36
odometer         1359
color              17
interior           12
seller            278
mmr               584
sellingprice      373
saledate           30
dtype: int64
```

Analysis

Distribution of Vehicle Body Types:

What is the distribution of vehicle body (e.g., sedan, SUV, truck) in the dataset?

```
import pandas as pd
import matplotlib.pyplot as plt


data = pd.read_csv("/content/sample_data/Car_Prices_cleaned.csv")


body_type_counts = data["body"].value_counts()


colors = ['blue', 'green', 'orange', 'red', 'purple', 'yellow','grey','maroon','pink']

body_type_counts.plot(kind="bar", figsize=(10, 6), color=colors)
plt.xlabel("Vehicle Body Type")
plt.ylabel("Count")
plt.title("Distribution of Vehicle Body Types")
plt.xticks(rotation=45)
plt.show()
```

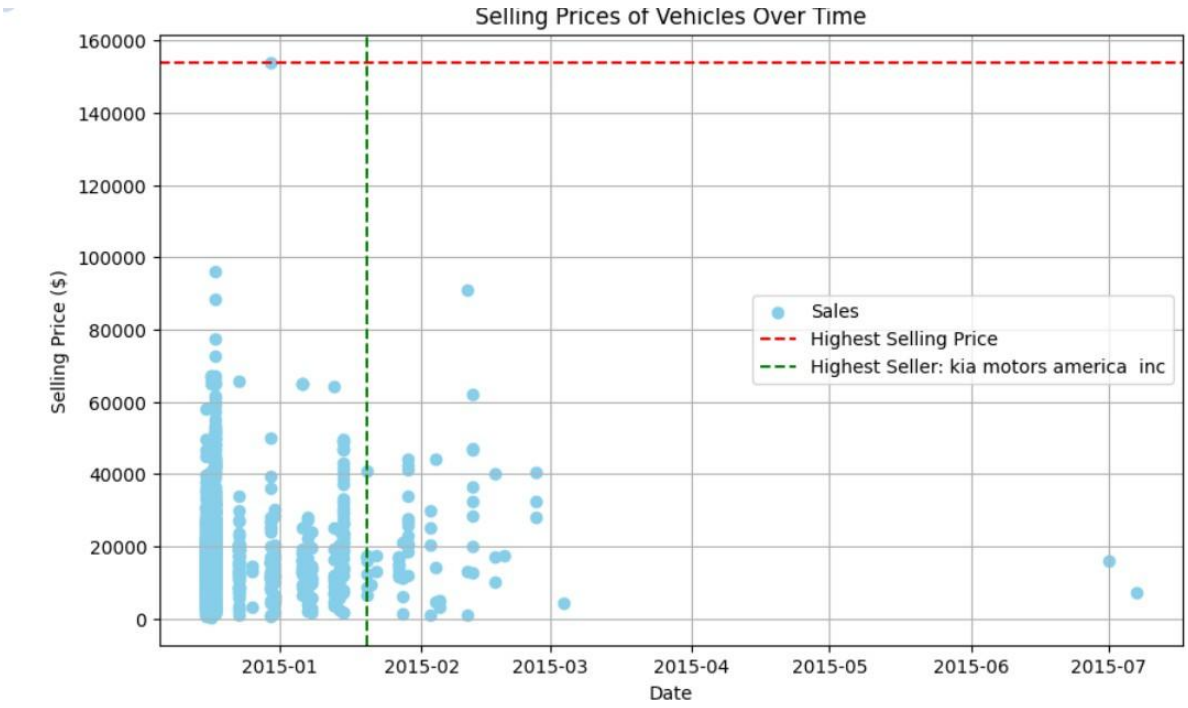


The distribution of various car body shapes within a dataset is depicted in this vibrant bar chart. With the labels for the various car body types positioned along the x-axis and their corresponding counts on the y-axis, each bar indicates the number of vehicles for a specific body type. The bars are color-coded to visually distinguish different body types; the colors range from blue to pink. The headline "Distribution of Vehicle Body Types" makes it evident what the chart is showing, and the labels on the x-axis are made easier to read by rotating them by 45 degrees. The frequency of each vehicle body type in the dataset is well communicated by the chart.

Vehicle Selling Price:

Which car seller has the highest selling price recorded on a specific date?

```
import pandas as pd
import matplotlib.pyplot as plt


df = pd.read_csv('/content/sample_data/Car_Prices_cleaned.csv')


df['saledate'] = pd.to_datetime(df['saledate'])

def highest_selling_seller_on_date(df, date):

    df_date = df[df['saledate'] == date]

    max_price_row = df_date[df_date['sellingprice'] == df_date['sellingprice'].max()]


    seller = max_price_row['seller'].values[0]

    return seller


date_of_interest = pd.to_datetime('2015-01-20')


seller = highest_selling_seller_on_date(df, date_of_interest)


plt.figure(figsize=(10, 6))
plt.scatter(df['saledate'], df['sellingprice'], color='skyblue', label='Sales')
plt.title('Selling Prices of Vehicles Over Time')
plt.xlabel('Date')
plt.ylabel('Selling Price ($)')
plt.axhline(y=df['sellingprice'].max(), color='red', linestyle='--', label='Highest Selling Price')
plt.axvline(x=date_of_interest, color='green', linestyle='--', label=f'Highest Seller: {seller}')
plt.legend()
plt.grid(True)
plt.show()
```

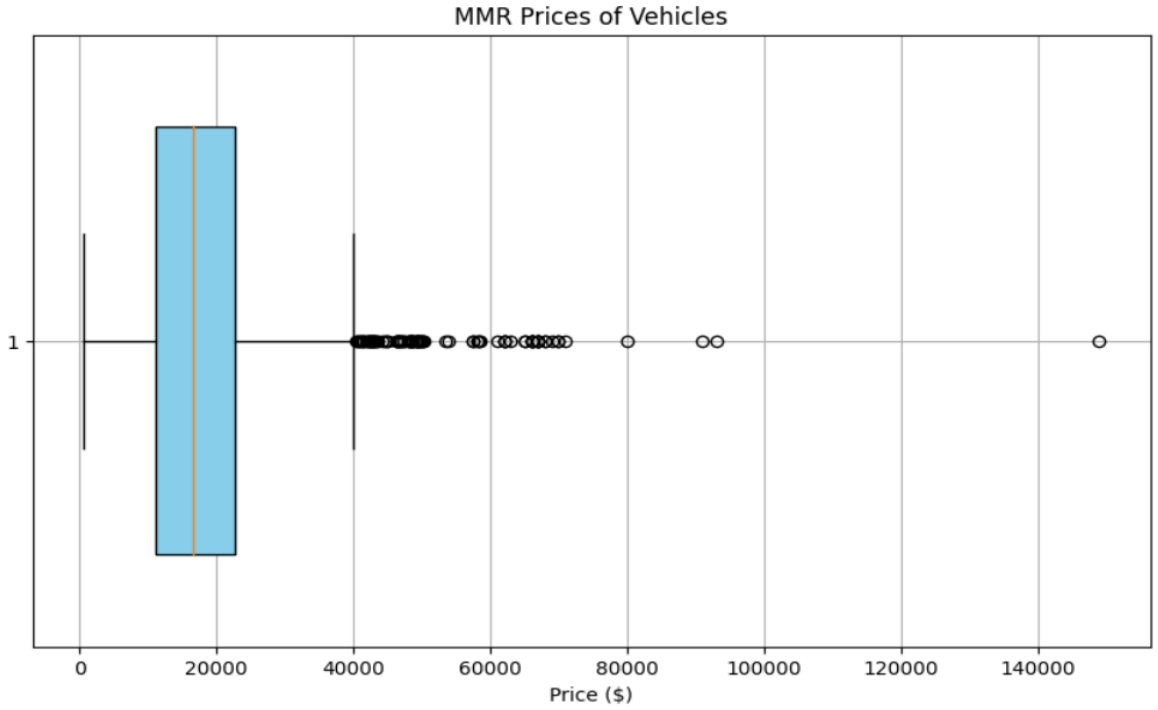


A scatter plot is used in the visualization to display the selling prices of cars over time. To enable datetime format for plotting, the sale date' column in the dataset is processed. With the selling price on the y-axis and the sale date on the x-axis, each point on the scatter plot represents a vehicle sale. The highest selling price over all dates is shown as a red dashed line in a horizontal orientation. A vertical dashed green line indicates the best-selling seller on a given date of interest, as determined by the function offered. The plot, "Selling Prices of Vehicles Over Time," has grid lines for clarity and a legend that explains these components.

Average MMR Prices of Vehicle:

What is the average mmr price of vehicles in the dataset?

```
import pandas as pd
import matplotlib.pyplot as plt


df = pd.read_csv('/content/sample_data/Car_Prices_cleaned.csv')

plt.figure(figsize=(10, 6))
plt.boxplot(df['mmr'], vert=False, widths=0.7, patch_artist=True, boxprops=dict(facecolor='skyblue'))
plt.title('MMR Prices of Vehicles')
plt.xlabel('Price ($)')
plt.grid(True)
plt.show()
```

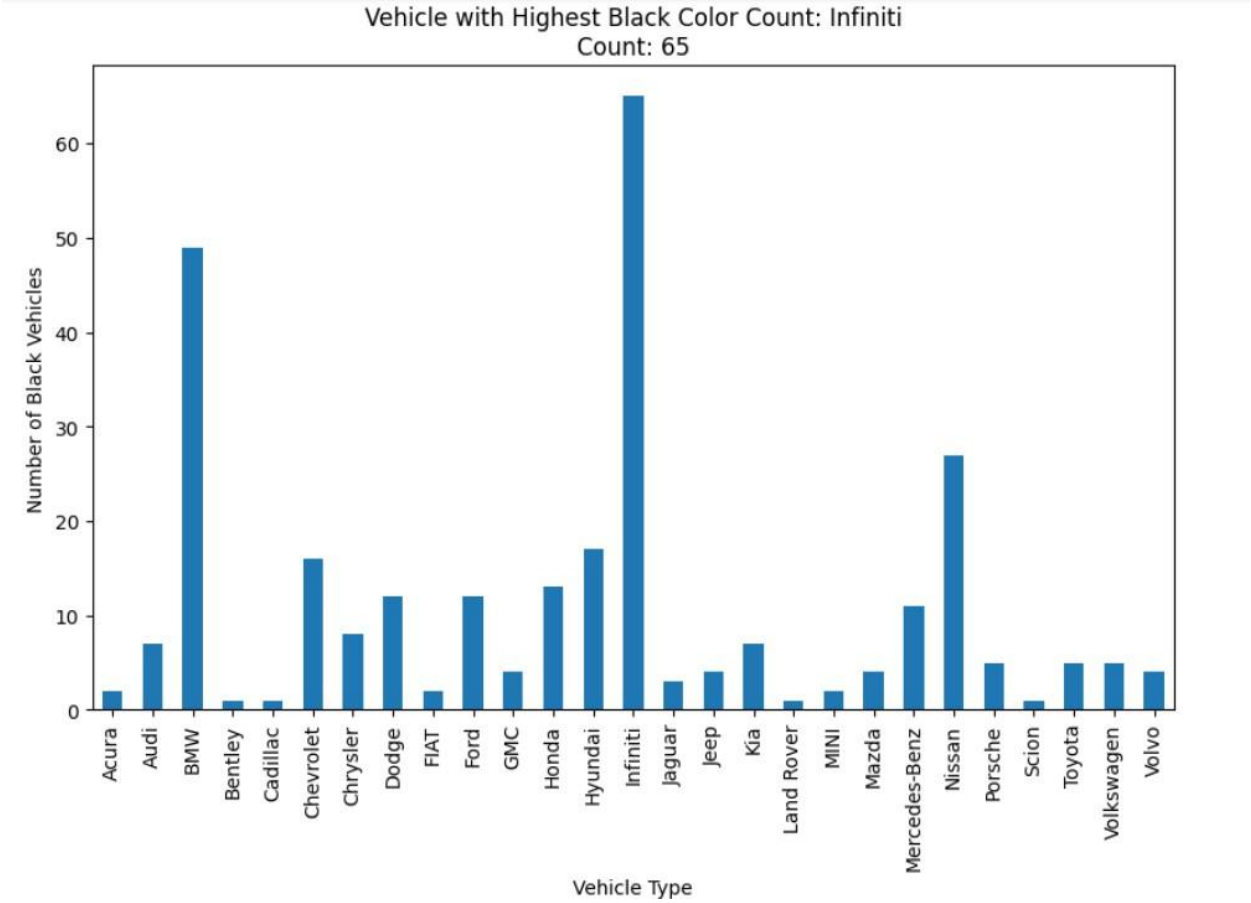


The Manufacturer's Suggested Retail Price (MMR) distribution for cars in a dataset is shown in the boxplot visualization that is supplied. The y-axis displays the number of cars, which is probably incorrectly labeled as a boxplot's y-axis should normally reflect the range of the data set. The MMR pricing of cars are shown on the x-axis in US dollars. The MMR prices' median, interquartile range (IQR), and outliers are displayed in the boxplot. Individual points outside of the whiskers are called outliers, and they represent MMR prices that are abnormally high or low in relation to the rest of the data.

Vehicle Brand with Highest Number of Black Cars:

Which vehicle has the highest black color recorded in the dataset?

```
import pandas as pd
import matplotlib.pyplot as plt


df = pd.read_csv('/content/sample_data/Car_Prices_cleaned.csv')

black_counts = df[df['color'] == 'black'].groupby('make')['color'].count()

highest_black_count = black_counts.max()
highest_black_vehicle = black_counts.idxmax()


plt.figure(figsize=(10, 6))
black_counts.plot(kind='bar')


plt.xlabel('Vehicle Type')
plt.ylabel('Number of Black Vehicles')
plt.title(f"Vehicle with Highest Black Color Count: {highest_black_vehicle}\nCount: {highest_black_count}")


plt.show()
```

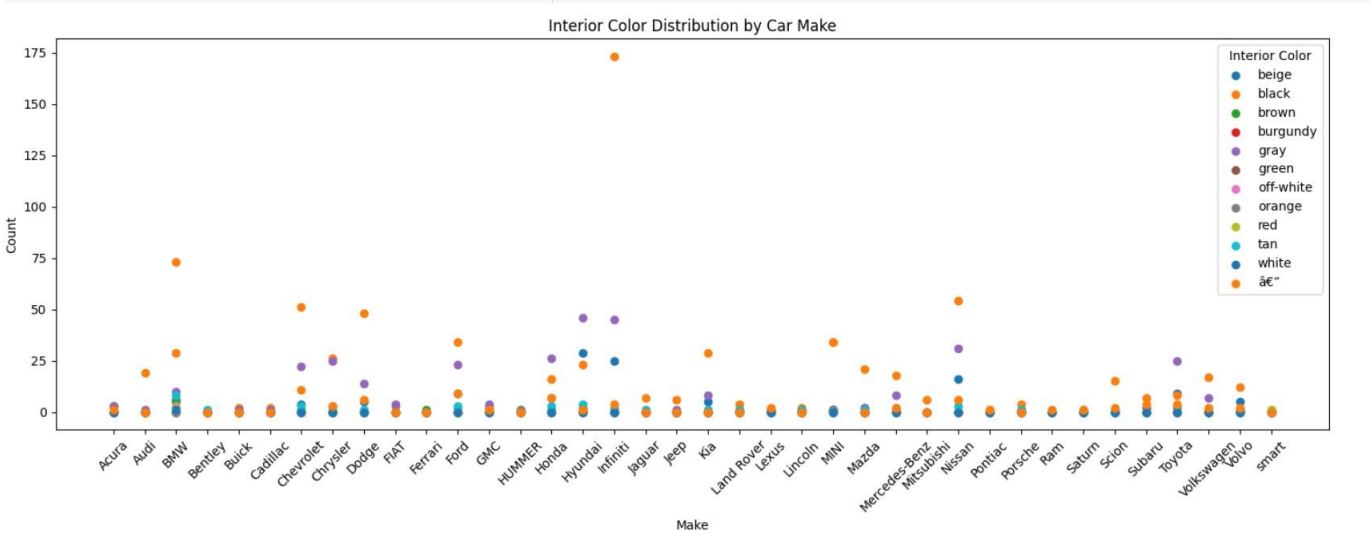


Based on a dataset, the bar chart shows the proportion of black vehicles by kind of vehicle. The information is first filtered to only include black-colored cars, after which it is categorized by make—referred to as "vehicle type" in this context. Each manufacture is displayed on the x-axis of the bar chart, while the y-axis displays the matching number of black cars. The title highlights the car type with the highest number of black automobiles, in this case the Infiniti with a total of 65. Within this particular dataset, this visualization facilitates the easy identification of the vehicle type with the highest proportion of black automobiles.

Interior Car Color Analysis by Make:

What is the interior color of a particular car based on make and model?

```
import pandas as pd
import matplotlib.pyplot as plt

def load_data(file_path):
    """Load data from a CSV file into a pandas DataFrame."""
    return pd.read_csv(file_path)

def process_data(df):
    """Group the data by make and interior color, and count the occurrences."""
    interior_counts = df.groupby(['make', 'interior']).size().unstack(fill_value=0)
    return interior_counts

def plot_data(interior_counts):
    """Create a scatter plot of interior color distribution by car make."""
    plt.figure(figsize=(15, 6))
    for column in interior_counts.columns:
        plt.scatter(interior_counts.index, interior_counts[column], label=column)
    plt.xlabel('Make')
    plt.ylabel('Count')
    plt.title('Interior Color Distribution by Car Make')
    plt.legend(title='Interior Color')
    plt.xticks(rotation=45)
    plt.xticks(range(len(interior_counts.index)), interior_counts.index)
    plt.tight_layout()
    plt.show()

def main():

    file_path = '/content/sample_data/Car_Prices_cleaned.csv'
    df = load_data(file_path)


    interior_counts = process_data(df)


    plot_data(interior_counts)

if __name__ == "__main__":
    main()
```



A visual comparison of interior color frequencies across different automobiles is possible with the scatter plot. 'Make' and 'interior' color groupings are applied to the data from a CSV file, and the resulting plot displays the number of interior color for each automobile make. According to the legend, various colors on the plot correspond to various inside color. Car makes are listed on the x-axis, and the number of vehicles for each interior color within those makes is displayed on the y-axis. The x-axis labels are rotated to make them readable. The goal of the visualization is concisely described in the plot title, "Interior Color Distribution by Car Make". This graphic makes it possible to quickly compare the popularity of interior colors among various automakers.

Findings and Managerial Implications

Kinds of Transmission (Manual and Automatic):

Observation:

The dataset notifies two transmission types as manual and automatic.

Implications:

Commercial organizations must bear in mind the need to stock both hand and machine-based vehicles, considering their customers' choices in their market. Although automatic transmission is the most favorite by many in different parts of the as manual transmission are still preferred in certain segments, especially in sports cars and small vehicles. Communicate with customers and play to the strength of different channels to narrow down the gap between the audience.

Styles of Vehicles Users:

Observation:

The catalog acknowledges nine styles of car bodies, each ranging from sedan to convertible to SUV.

Implications:

The variability of car body types helps businesses to establish a large customer base on the ground of the different consumers' needs in terms of family-friendly transportation and fashionable and luxury vehicles. Companies must indicate that they have this variation in marketing materials so that consumers will see that their demands on the market are being taken care of. Bearing this in mind, as well as flooring the market with the most appealing body types, such as sedans and SUVs, inventory management and sales can be upgraded.

Color and Interior Selection:

Observation:

The data set offers 17 exterior colors and 11 interior styles that are customizable implying that buyers have boundless options.

Implications:

The fact that there are a lot of varieties in colors and choices on interior furnishing provides an opportunity for putting up individual personalization choices. Currently, businesses can utilize this diversity by enabling customers to include specific features to the vehicle they have selected before purchasing the vehicle. Marketing efforts can display the current trends in colors and décor in a variety of scenarios to appeal to whoever sees them. Also, businesses can use this information to make their stocking decisions where they can buy vehicles faced with the most expected or desired colors and interior configurations.

Note Worthy Sale Dates:

Observation:

The dataset includes 30 dates of sale possibly signifying sale dates or the data collecting pattern.

Implications:

As soon as the data reflect a specific sales period, it becomes possible to detect seasonal trends, and unearth patterns in customer purchases. This data can be utilized in order to come up with campaigns which will highlight trendy periods with a purpose to make the biggest possible earnings. Furthermore, the companies need to discover the reasons for sales on those special days to find out the successful promotion or marketing approaches.

Extremely Spread between Mileage:

Observation:

The miles traveled vary from 1 mile to 355,898 miles, with the lowest being for vehicles that are used frequently, and the highest for infrequently used ones.

Implications:

Such a substantial difference in mileage could be an opportunity of certain companies to become the dealers of new cars as well as used ones. Businesses need to make a mix of selections so that everyone can afford or buy the one they like. For used vehicles, certifying them, having warranties, and proper inspections are measures that will develop the trust of the customers. A marketing effort can address the low maintenance and reasonable cost for used cars with different mileages.

Extreme Price Variances:

Observation:

They are going from about $300 to $154,000, with a mean close to $18,001, which shows that these cars have different values.

Implications:

The fact that the price of the car varies greatly indicates that the cars may have a group of customers ranging from those who have a budget to the ones that look for cars as luxury vehicles. For better income streams, each business should approach marketing at different price brackets. Moreover, pricing strategy should cover the environmental span of vehicle assets with different price levels in the whole segment range.

Size of this Market Includes Vast Amount of Fashion Products with Various Makes and Models:

Observation:

According to scenario, the data set contains 213 different vehicle models, and 38 vehicle makes, thereby providing a broad picture on the market.

Implications:

The fact that there is a massive range of models and makes from different companies allows businesses to stock a wide variety of products and cars that are to their customers' satisfaction. As the means to raise positive customers experience, the number of models and brands companies should be expanded to give customers a wide range of options. Marketing campaigns can be dedicated to the fact that the market of vehicles is broad and as a result the shoppers who would like to be particular in selection of brand or model will be pleased. Besides, companies can estimate which models and makes order was the most and update the warehouse in accordance with their findings.

Conclusion:

The cars distribution dataset analysis resulted in an in-depth understanding of the dynamics of the automotive industry and a basis for strategic and tactical decision making, marketing and operations. Taking components for example such as transmission types, the number and range of car body options, color design, interior, sales volumes, mileage, price variability, and the diversity of automobile models and makes, a holistic picture of the business' success and customers satisfaction was drawn.

It is apparent that transmission preferences are critical, and there is a combination of manual and automatic vehicles in the data set. Automatic transmissions have the largest share of the market, but the market retains a considerable base of manual transmission lovers, such as sports car enthusiasts, budget car buyers and so on. This argument implies that firms need to have a balanced stock of the various transmission types and also the marketing should emphasize the particular benefits of every transmission type.

The data set has nine different types of car bodies, like sedans, SUVs, and convertibles, making it suitable for all car types of buyers' tastes. This enhances the diversification requirement for the business to supply products according to the different needs of the consumers. Having different types of cars in marketing campaigns is one of those things that can attract a bigger audience and increase sales, because it shows how flexible the car company is when it comes to the different transportation needs.

Customization is an essential feature too, and according to the data, the model will come in 17 exterior colors and 11 interior options. This customization makes for individual types of cars that are a unique and specific feature that attracts the consumers who value uniqueness. If a company has a good number of individualizing choices, customers may perceive the shopping process as more engaging. When marketing campaigns put an emphasis on personalization, interested customers are attracted and satisfied.

Interestingly, based on the 30 different sale dates, a probable pattern or seasonal peaks in customer purchase behavior could be traced. Businesses can take advantage of it for execution of promotions and for tuning inventory according to the expected demand level. The analysis of these trends can help to discover perfect marketing strategies, which can then be used to optimize the number of sales at busiest hours.

The varied milage of 1 mile to over 355,000 shows the extent to which vehicles have been used, and this range can help you to enter both the new and used vehicle market. Businesses can split up their customers into segments by offering different models of vehicles with varying mileage. The confidence of a client should be earned through nationally recognized certificates and guaranteed warranties which indicate quality and assured dependability.

The price dependency makes it difficult to decide the selling prices, which range from $300 to $154,000 and with a mean value of almost $18,001. This shows that businesses bookmaking a flexible price to attract customers of different income levels. The manufacturing of vehicles in price ranges from lower to higher can catch at once both demand of budget-conscious prospects as well as those looking for luxurious models. When manufacturers differentiate their product based on different price points, marketing campaigns should be tailored to ensure optimum revenue and customer engagement.

In conclusion, the data analysis of car distribution produces actionable ideas both related to the customers' preference and the market trend. Through data-driven integrations of inventory, marketing and sales strategies with the collected data, automotive businesses will be empowered to increase their competitiveness, customer approval, and income.

Appendix: Python Codes with Proper Documentations

```
import numpy as np
import pandas as pd
```

```
df = pd.read_csv('/content/sample_data/Car_Prices_cleaned.csv')
```

```
df.columns
```

```
df.dtypes
```

```
df["full_model"] = df[('make')] + df[('make')]
```

```
df.columns
```

```
df.head()
```

What is the distribution of vehicle body (e.g., sedan, SUV, truck) in the dataset?

```
import pandas as pd
import matplotlib.pyplot as plt


data = pd.read_csv("/content/sample_data/Car_Prices_cleaned.csv")


body_type_counts = data["body"].value_counts()


colors = ['blue', 'green', 'orange', 'red', 'purple', 'yellow','grey','maroon','pink']

body_type_counts.plot(kind="bar", figsize=(10, 6), color=colors)
plt.xlabel("Vehicle Body Type")
plt.ylabel("Count")
plt.title("Distribution of Vehicle Body Types")
plt.xticks(rotation=45)
plt.show()
```

Which car seller has the highest selling price recorded on a specific date?

```
import pandas as pd
import matplotlib.pyplot as plt


df = pd.read_csv('/content/sample_data/Car_Prices_cleaned.csv')


df['saledate'] = pd.to_datetime(df['saledate'])

def highest_selling_seller_on_date(df, date):

    df_date = df[df['saledate'] == date]

    max_price_row = df_date[df_date['sellingprice'] == df_date['sellingprice'].max()]


    seller = max_price_row['seller'].values[0]

    return seller


date_of_interest = pd.to_datetime('2015-01-20')


seller = highest_selling_seller_on_date(df, date_of_interest)


plt.figure(figsize=(10, 6))
plt.scatter(df['saledate'], df['sellingprice'], color='skyblue', label='Sales')
plt.title('Selling Prices of Vehicles Over Time')
plt.xlabel('Date')
plt.ylabel('Selling Price ($)')
plt.axhline(y=df['sellingprice'].max(), color='red', linestyle='--', label='Highest Selling Price')
plt.axvline(x=date_of_interest, color='green', linestyle='--', label=f'Highest Seller: {seller}')
plt.legend()
plt.grid(True)
plt.show()
```

What is the average mmr price of vehicles in the dataset?

```
import pandas as pd
import matplotlib.pyplot as plt


df = pd.read_csv('/content/sample_data/Car_Prices_cleaned.csv')

plt.figure(figsize=(10, 6))
plt.boxplot(df['mmr'], vert=False, widths=0.7, patch_artist=True, boxprops=dict(facecolor='skyblue'))
plt.title('MMR Prices of Vehicles')
plt.xlabel('Price ($)')
plt.grid(True)
plt.show()
```

Which vehicle has the highest black color recorded in the dataset?

```
import pandas as pd
import matplotlib.pyplot as plt


df = pd.read_csv('/content/sample_data/Car_Prices_cleaned.csv')

black_counts = df[df['color'] == 'black'].groupby('make')['color'].count()

highest_black_count = black_counts.max()
highest_black_vehicle = black_counts.idxmax()


plt.figure(figsize=(10, 6))
black_counts.plot(kind='bar')


plt.xlabel('Vehicle Type')
plt.ylabel('Number of Black Vehicles')
plt.title(f"Vehicle with Highest Black Color Count: {highest_black_vehicle}\nCount: {highest_black_count}")


plt.show()
```

What is the interior color of a particular car based on make and model?

```python
import pandas as pd
import matplotlib.pyplot as plt

def load_data(file_path):
    """Load data from a CSV file into a pandas DataFrame."""
    return pd.read_csv(file_path)

def process_data(df):
    """Group the data by make and interior color, and count the occurrences."""
    interior_counts = df.groupby(['make', 'interior']).size().unstack(fill_value=0)
    return interior_counts

def plot_data(interior_counts):
    """Create a scatter plot of interior color distribution by car make."""
    plt.figure(figsize=(15, 6))
    for column in interior_counts.columns:
        plt.scatter(interior_counts.index, interior_counts[column], label=column)
    plt.xlabel('Make')
    plt.ylabel('Count')
    plt.title('Interior Color Distribution by Car Make')
    plt.legend(title='Interior Color')
    plt.xticks(rotation=45)
    plt.xticks(range(len(interior_counts.index)), interior_counts.index)
    plt.tight_layout()
    plt.show()

def main():

    file_path = '/content/sample_data/Car_Prices_cleaned.csv'
    df = load_data(file_path)


    interior_counts = process_data(df)


    plot_data(interior_counts)

if __name__ == "__main__":
    main()
```

References

- The dataset is taken from Kaggle:

  Anwar, S. (2024, February 21). *Vehicle sales data*. Kaggle. https://www.kaggle.com/datasets/syedanwarafridi/vehicle-sales-data

- Dr. Melody White, Department of Information Technology and Decision Sciences (2024). *Programming Languages in Business Analytics* [PowerPoint slides]. University of North Texas, Denton.

  Data analysis and visualization.

  https://unt.instructure.com/courses/98627/files/25554039/download?download_frd=1

  Introduction to NumPy, pandas and matplotlib.

  https://unt.instructure.com/courses/98627/files/25268437/download?download_frd=1