澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Faculty of Business Administration

ISOM7015 - FUNDAMENTALS OF BUSINESS ANALYTICS

Group Project Report

(First Semester, 2024/2025)

Theme - Sales Analysis on Vehicle Market

Contents

1 Introduction

This report examines the automotive market's transition toward electric vehicles (EVs) in response to global sustainability goals. Governments worldwide are promoting EV adoption through regulations and incentives, with examples like the EU's 2035 ban on new petrol and diesel cars and Norway's EV dominance, accounting for over 80% of new car sales in 2023. China, the largest EV market, has also advanced rapidly through subsidies and innovations from companies like BYD.

While EVs offer lower operating costs and reduced emissions, challenges such as higher upfront costs, limited charging infrastructure, and battery production concerns persist. Traditional internal combustion engine (ICE) vehicles, despite their affordability and established fueling networks, face growing environmental scrutiny. This study explores these contrasts, market trends, and strategies to support EV adoption and address associated challenges.

The purpose of this analysis is to uncover key factors influencing vehicle sales, including pricing, branding, and regional variations, while comparing the performance of EVs with traditional fuel-powered vehicles. The ultimate goal is to provide actionable insights for manufacturers and policymakers to refine strategies, support EV adoption, and contribute to a more sustainable automotive future.

This analysis utilizes three key datasets, each containing distinct yet complementary information:

'sales data.csv' includes vehicle sales data such as sales volume, regions, time, companies, brands, models, engine displacement, and fuel types. It serves as the foundation for analyzing market trends and regional sales distribution.

'cars to match.csv' contains a list of car brands, models, and technical specifications that correspond to the sales data but lack price information. This dataset acts as a bridge to link sales data with pricing details.

'autohome merged for matching.csv' provides car brands, models, manufacturer IDs, pricing details (e.g., manufacturer suggested retail prices), and technical specifications like engine displacement and transmission type. It is used to supplement the missing price information in the "cars to match.csv" dataset.

By integrating these datasets, the analysis enables the matching of sales data

with pricing information, creating a comprehensive data foundation for exploring market trends and insights.

2 Methodology

Given the large size of the datasets—cars to match.xlsx (87,772 entries), autohome merged for matching.xlsx (67,134 entries), and sales data.xlsx (1,048,575 entries)—they were randomly split into smaller datasets containing 200 entries each for testing the code. Once the code was verified to run correctly, the complete datasets were processed.

2.1 Match The Car Models

The aim here is to identify matching car information between the two datasets and retrieve the manufacturer's suggested retail price (MSRP) from autohome merged for matching.xlsx. The observed matching column headers are as follows:

- 'cars to match.xlsx' : Company, Brand, Model, Version, Vehicle Code, Displacement, Fuel Type, Transmission, Body Type.

- 'autohome merged for matching.xlsx' : Manufacturer Name, Brand Name, Series Name, Model Name, Displacement (L), Energy Type, Transmission, Body Structure.

Figure 2.1.1 The comparison of some labels between 'cars to match.xlsx' and 'autohome merged for matching.xlsx'

Cars to match.xlsx	企业	品牌	车型	款型	车辆型号	排量	燃料	变速器	车身类型
/	北汽新能源	ARCFOX	ARCFOX αT	2021 款 653S	BJ6480BSA-BEV	0.0L	纯电动	1AT	SUV
北汽新能源 ARCFOX ARCFOX αT 2021 款 653S BJ6480BSA-BEV 0.0L 纯电动 1AT SUV									
autohome merged for matching.xlsx	厂商名称	品牌名称	车系名称	车型名称	/	排量（L）	能源类型	变速箱	车身结构
/	北汽新能源	ARCFOX 极狐	极狐 阿尔法 T(ARCFOX αT)	极狐 阿尔法 T(ARCFOX αT) 2021 款 653S 160kW	/		纯电动		5 门 5 座 SUV
北汽新能源 ARCFOX 极狐 极狐 阿尔法 T(ARCFOXαT) 极狐 阿尔法 T(ARCFOX αT) 2021 款 653S 160kW 纯电动 5 门 5 座 SUV									

Steps Taken are as follows:

Data Consistency. All columns in both datasets were converted to string

format to avoid errors during data merging or processing.

Figure 2.1.2

```
from rapidfuzz import process
from joblib import Parallel, delayed
from tqdm import tqdm
import pandas as pd
import re
```

```
cars_to_match = pd.read_excel('C:/Users/huawei/Desktop/cars to match.xlsx')
autohome_merged = pd.read_excel('C:/Users/huawei/Desktop/autohome merged for matching.xlsx')
# 将“排量(L)”和“变速箱”列转换为字符串类型
autohome_merged['排量(L)'] = autohome_merged['排量(L)'].astype(str)
autohome_merged['变速箱'] = autohome_merged['变速箱'].astype(str)
```

(2) *Combined Information Column.* Important content from matching columns was merged into a single combined_info column for both datasets. A clean_text function was used to remove special characters and spaces, ensuring only valid text (Chinese/English characters and numbers) remained.

Figure 2.1.3

```
# Ensure all columns are in string format before concatenation
cars_to_match = cars_to_match.astype(str)
autohome_merged = autohome_merged.astype(str)

# Concatenate relevant columns in 'cars to match' to create a comprehensive text column for matching
cars_to_match['combined_info'] = cars_to_match[['企业', '品牌', '车型', '款型', '排量', '燃料', '变速器', '车身类型']].fill

# Concatenate relevant columns in 'autohome' to create a comprehensive text column for matching
autohome_merged['combined_info'] = autohome_merged[['厂商名称', '品牌名称', '车系名称', '车型名称', '排量(L)', '能源类型', '

# Display the first few rows of the new columns to verify
cars_to_match_combined_head = cars_to_match[['combined_info']].head()
autohome_combined_head = autohome_merged[['combined_info']].head()

cars_to_match_combined_head, autohome_combined_head
```

```
(                                       combined_info
 0  北京汽车 北京 北京40L 北京40L 2018款 2.0T MT 尊贵型 四驱 柴油 2....
 1  北京汽车 北京 北京40L 北京40L 2018款 2.0T MT 尊贵型 四驱 柴油 2....
 2  长城汽车 哈弗 哈弗H9 哈弗H9 2017款 2.0T AT 尊享型 四驱 5座 柴油 2...
 3  江淮汽车 江淮 瑞风M4 瑞风M4 2019款 1.9T MT 商务型 柴油 1.9T 柴油...
 4  江淮汽车 江淮 瑞风M4 瑞风M4 2019款 1.9T MT 舒适型 柴油 1.9T 柴油...,
                                       combined_info
 0  AC Schnitzer AC Schnitzer AC Schnitzer X5 AC S...
 1  AC Schnitzer AC Schnitzer AC Schnitzer M3 AC S...
 2  安凯客车 安凯客车 宝斯通 宝斯通 2010款 2.2L高级版HFC4GA2-1B 2.2 ...
 3  安凯客车 安凯客车 宝斯通 宝斯通 2010款 3.0T高级版NGD3.0-C3HA 3 柴...
 4  安凯客车 安凯客车 宝斯通 宝斯通 2014款 3.0T VIP版NGD3.0-C3HA 3...)
```

(3) *Fuzzy Matching.* Using the rapidfuzz.process.extractOne function, the closest match is searched within the autohome_df['combined_info'] column, with a matching score threshold of ≥60. If a match is found, the corresponding row from autohome_df is extracted, and a new dictionary, matched_record, is created. This dictionary contains the original data of car_row (from "cars to match.xlsx")

along with the manufacturer's suggested retail price (MSRP in yuan) from "autohome merged for matching.xlsx". If no match is found, the result is returned as None.

Figure 2.1.4

```python
# 定义数据清理函数
def clean_text(text):
    # 移除特殊字符和多余空格
    return re.sub(r'\s+|[^A-Za-z0-9\u4e00-\u9fff]', '', text)

# 清理数据中的 combined_info 列
cars_to_match['combined_info'] = cars_to_match['combined_info'].apply(clean_text)
autohome_merged['combined_info'] = autohome_merged['combined_info'].apply(clean_text)

# 定义单条记录的模糊匹配函数
def fuzzy_match_single_record(car_row, autohome_df, threshold=60):
    # 使用 rapidfuzz 找到最佳匹配
    match = process.extractOne(car_row['combined_info'], autohome_df['combined_info'].tolist(), score_cutoff=threshold)

    if match:
        # 获取 autohome 匹配行的数据
        matched_row = autohome_df[autohome_df['combined_info'] == match[0]].iloc[0]

        # 提取相关信息
        matched_record = {
            '企业': car_row['企业'],
            '品牌': car_row['品牌'],
            '车型': car_row['车型'],
            '款型': car_row['款型'],
            '车辆型号': car_row['车辆型号'],
            '排量': car_row['排量'],
            '燃料': car_row['燃料'],
            '变速器': car_row['变速器'],
            '车身类型': car_row['车身类型'],
            '厂商指导价(元)': matched_row['厂商指导价(元)']
        }
        return matched_record
    return None
```

Additionally, to improve the processing speed for large-scale data matching, the Parallel and delayed functions from the joblib library were used to process each car_row record in parallel. The tqdm library was also utilized to display a progress bar, making it easier to monitor the program's execution progress.

Figure 2.1.5

```python
# 定义并行处理的函数
def fuzzy_match_cars_parallel(cars_df, autohome_df, threshold=60, n_jobs=-1):
    # 使用并行处理，每个 car_row 单独进行模糊匹配，并显示进度条
    matched_records = Parallel(n_jobs=n_jobs)(
        delayed(fuzzy_match_single_record)(car_row, autohome_df, threshold)
        for _, car_row in tqdm(cars_df.iterrows(), total=len(cars_df), desc="Processing records")
    )

    # 过滤掉 None 值，并生成 DataFrame
    matched_records = [record for record in matched_records if record is not None]
    matched_df = pd.DataFrame(matched_records)
    return matched_df

# 设置匹配阈值和并行化参数
threshold = 60  # rapidfuzz 的阈值为 0-100
n_jobs = -1  # 使用所有可用 CPU 核心

# 开始并行模糊匹配
matched_cars_data_cleaned = fuzzy_match_cars_parallel(cars_to_match, autohome_merged, threshold, n_jobs)

# 显示前几行结果进行验证
matched_cars_data_cleaned_head = matched_cars_data_cleaned.head()
print(matched_cars_data_cleaned_head)
```

```
Processing records: 100%|██████████████████████████████████████| 87772/87772 [6:23:09<00:00,  3.82it/s]
     企业    品牌      车型                      款型         车辆型号    排量   燃料  \
0  北京汽车    北京   北京40L    北京40L 2018款 2.0T MT 尊贵型 四驱 柴油  BJ2024F7VME  2.0T  柴油
1  北京汽车    北京   北京40L    北京40L 2018款 2.0T MT 尊贵型 四驱 柴油  BJ2024F7VM1E  2.0T  柴油
2  长城汽车    哈弗    哈弗H9  哈弗H9 2017款 2.0T AT 尊享型 四驱 5座 柴油   CC6510WM60  2.0T  柴油
3  江淮汽车    江淮    瑞风M4         瑞风M4 2019款 1.9T MT 商务型 柴油  HFC6521K1C8V  1.9T  柴油
4  江淮汽车    江淮    瑞风M4         瑞风M4 2019款 1.9T MT 舒适型 柴油  HFC6521K1C8V  1.9T  柴油

   变速器 车身类型 厂商指导价(元)
0  6MT  SUV        -
1  6MT  SUV        -
2  8AT  SUV        -
3  6MT  MPV   139800
4  6MT  MPV   129800
```

At the end of the first part, we observed that the "Manufacturer's Suggested Retail Price (MSRP, in yuan)" column in the output .xlsx file had inconsistent formatting. Some entries were written as pure numbers, while others combined numbers with text (e.g., "xx.xx 万"). To facilitate subsequent data analysis, we wrote a script to standardize all entries into numeric format. Additionally, for MSRP values expressed as ranges, we calculated the average value for conversion (e.g., "5.08 万~5.28 万" was converted to its average). After organizing all the formats, the cleaned dataset was saved as matched_cars_data_cleaned3.xlsx.

Upon verification, matched_cars_data_cleaned3.xlsx contained 79,726 vehicle entries, with 8,633 entries unmatched to an MSRP, resulting in an effective match rate of 89.2%.

Figure2.1.6

```
# Define a function to convert all price formats (including text) to numeric values
def convert_to_numeric(price):
    if isinstance(price, str):
        price = price.strip()
        # Handle cases like "XX.XX万" or ranges like "5.08万~5.28万"
        if '万' in price:
            price = price.replace('万', '')
            if '~' in price:  # Handle ranges
                low, high = map(float, price.split('~'))
                return (low + high) / 2 * 10000  # Average and convert to units
            else:
                return float(price) * 10000  # Convert to units
    return price  # Return as is if already numeric or invalid format

# Apply the conversion function to the "厂商指导价(元)" column
matched_cars_data_cleaned['厂商指导价(元)'] = matched_cars_data_cleaned['厂商指导价(元)'].apply(convert_to_numeric)
```

```
output_file_path = r'C:\Users\huawei\Desktop\matched_cars_data_cleaned3.xlsx'
matched_cars_data_cleaned.to_excel(output_file_path, index=False)

output_file_path
```

```
'C:\\Users\\huawei\\Desktop\\matched_cars_data_cleaned3.xlsx'
```

2.2 Adding MSRP to sales data.xlsx

We will add the manufacturer's suggested retail price (MSRP) from matched_cars_data_cleaned3.xlsx to sales data.xlsx, with the goal of achieving the highest possible match rate and accuracy. Upon examination, we observed that sales data.xlsx contains a column for vehicle model codes, which are unique identifiers composed of letters and numbers that correspond to specific vehicle information. Therefore, the matching logic in this part involves performing an exact match based on the vehicle model code column, which is present in both matched_cars_data_cleaned3.xlsx and sales data.xlsx, to retrieve the corresponding

MSRP from matched_cars_data_cleaned3.xlsx.

Figure 2.2.1

```
import pandas as pd
from tqdm import tqdm
```

```
# Load the Excel files containing sales data and matched cars data
sales_data = pd.read_excel('C:/Users/huawei/Desktop/sales data.xlsx')
matched_cars_data_cleaned = pd.read_excel('C:/Users/huawei/Desktop/matched_cars_data_cleaned3.xlsx')
```

We utilized Boolean indexing and array filtering in pandas to check, row by row, whether the vehicle model code in sales_df[key] (from "sales data.xlsx") exists in cars_df[key] (from "matched_cars_data_cleaned3.xlsx"). If a match was found, the corresponding value from the value_column (i.e., the MSRP) was extracted. Finally, the processed dataset was output as merged_sales_data_with_price_progress.xlsx.

After matching the 1,048,575 entries in sales data.xlsx, the resulting dataset, merged_sales_data_with_price_progress.xlsx, was verified. It contained 168,266 vehicle entries without a matched MSRP, resulting in an effective match rate of 84%.

Figure 2.2.2

```
# Function to merge with progress bar
def merge_with_progress(sales_df, cars_df, key, value_column):
    # Initialize tqdm progress bar for the merging process
    tqdm.pandas(desc="Matching rows")
    # Apply row-by-row search with progress monitoring
    merged_column = sales_df[key].progress_apply(
        lambda x: cars_df[cars_df[key] == x][value_column].values[0] if x in cars_df[key].values else None
    )
    return merged_column

# Perform the merge with progress tracking
sales_data['厂商指导价(元)'] = merge_with_progress(sales_data, matched_cars_data_cleaned, '车辆型号', '厂商指导价(元)')

# Display the first few rows of the resulting dataframe for verification
sales_data.head()
```

```
Matching rows: 100%|██████████| 1048575/1048575 [3:39:49<00:00, 79.50it/s]
```

	Unnamed: 0	年月	省	市	企业	品牌	车型	款型	车辆型号	排量	燃料	变速器	车身类型	数量	厂商指导价(元)
0	282682	202108	江苏	镇江	一汽轿车	红旗	红旗H5	2020款 1.5T DCT旗悦版	CA7150HA6T	1.5T	汽油	7DCT	NB	1	265000
1	85260	202107	甘肃	兰州市	一汽丰田	丰田	亚洲狮	2021款 2.0L 尊贵版	TV7206L	2.0L	汽油	CVT	NB	1	179800
2	730379	202107	四川	成都市	上海通用	别克	昂科威	2021款 532T 两驱精英型	SGM6475DBA2	1.5T	汽油	7DCT	SUV	4	239900
3	307609	202107	河南	信阳市	上海大众	斯柯达	柯珞克	2021款 TSI280 豪华版	SVW6449GTD	1.4T	汽油	7DCT	SUV	1	159900
4	141194	202106	广东	深圳市	北京奔驰	奔驰	奔驰E级	2021款 E 300 L 尊贵运动型	BJ7205MEL	2.0T	汽油	9AT	NB	2	-

```
# Save the final merged result to a new Excel file
output_path_progress = 'C:/Users/huawei/Desktop/merged_sales_data_with_price_progress.xlsx'
sales_data.to_excel(output_path_progress, index=False)

output_path_progress
```

```
'C:/Users/huawei/Desktop/merged_sales_data_with_price_progress.xlsx'
```

3 Results and Discussion

3.1 Descriptive Statistics

Because the summary statistics should be for the numerical columns, the first step we need to do is to filtrate the columns in 'merged_sales_data_with_price_progress.xlsx' that meet the requirement. By screening the dataset, we can decide that the columns of 'MSRPs (Manufacturer Suggested Retail Price)' and 'Sales Quantities" are numerical, then we do descriptive statistics.

The descriptive statistics provide valuable insights into the distribution and variability of the 'MSRPs' and 'Sales Quantities'. The average MSRP is 212,422.22 yuan, indicating a higher-end pricing trend in the dataset. However, the median MSRP of 166,900.80 yuan suggests that half of the vehicles are priced below this value, reflecting a broader distribution. The mode, 69,880.08 yuan, represents the most frequently occurring MSRP, which points to a significant number of lower-priced vehicles. The standard deviation of 132,419.03 yuan highlights a wide variation in pricing, likely driven by the inclusion of both economy and luxury vehicles.

Figure 3.1.1

```
count    8.803090e+05
mean     2.124222e+05
std      1.324191e+05
min      2.890000e+04
25%      1.229000e+05
50%      1.669000e+05
75%      2.606500e+05
max      1.043900e+06
Name: 厂商指导价(元), dtype: float64
平均厂商指导价: 212422.22 元
中位数厂商指导价: 166900.00 元
众数厂商指导价: 69800.00 元
厂商指导价的标准差: 132419.03 元
```

For sales quantities, the average value is 2.12, with a median of 1.80, indicating that most vehicles are sold in low volumes. The mode, 1.88, confirms that a small number of units is the most common scenario in the dataset. The standard deviation of 4.76 reflects some variability in sales volumes, but the relatively small average and median values suggest a dataset dominated by vehicles with limited sales numbers.

Figure 3.1.2

```
平均数量: 2.12
中位数数量: 1.00
众数数量: 1.00
数量的标准差: 4.76
```

Overall, the data indicates a diverse pricing strategy within the automotive market and a sales distribution skewed towards lower volumes per vehicle, potentially reflecting niche or high-cost products.

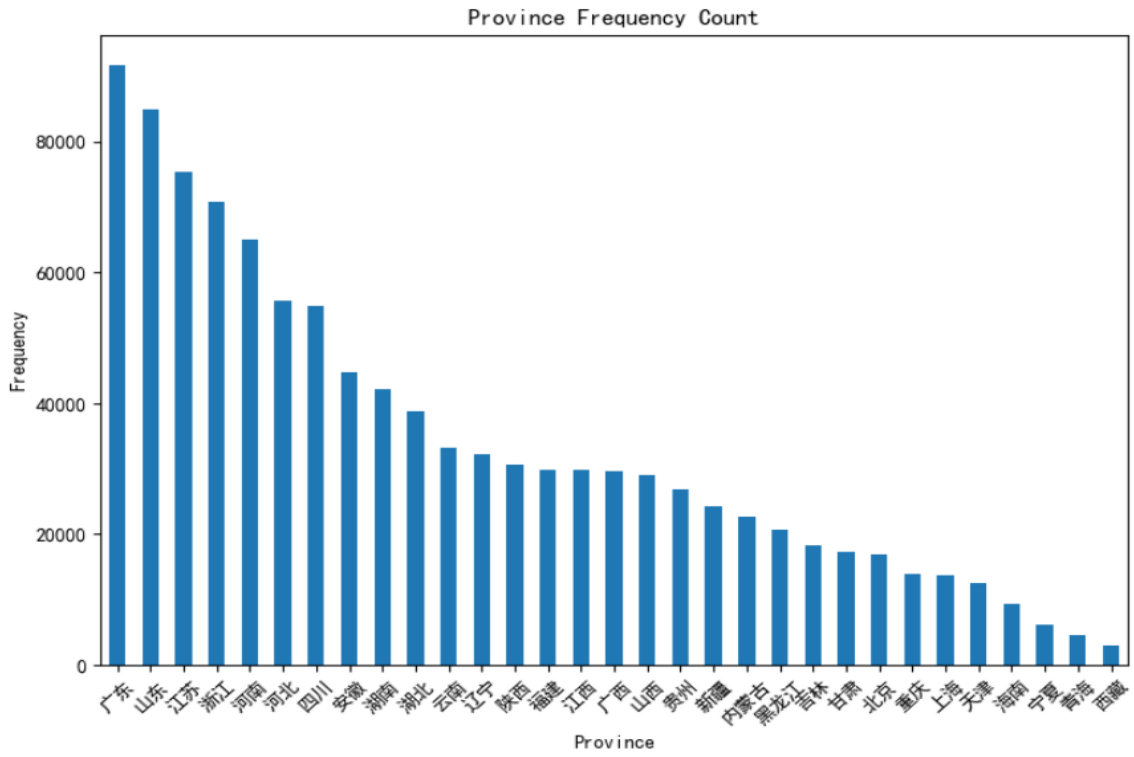
The province frequency count chart illustrates the distribution of vehicle sales across various provinces, highlighting significant regional differences. Guangdong

province leads with the highest frequency, surpassing 90,000 sales, followed closely by Shandong and Jiangsu, each exceeding 80,000. These provinces are major economic hubs, reflecting higher demand for automobiles due to their larger populations and stronger purchasing power.

In the mid-range, provinces such as Hebei, Henan, and Sichuan show moderate sales frequencies, indicating substantial but relatively less intense market activity compared to the leading regions. On the lower end, provinces like Ningxia, Qinghai, and Tibet have minimal sales frequencies, likely due to smaller populations, less developed economies, or challenging geographic conditions.

The data highlights a concentration of vehicle sales in economically developed eastern and coastal provinces, with a gradual decline moving toward less urbanized and economically weaker regions in the west and north. This pattern emphasizes the influence of regional economic disparity and urbanization levels on automotive market demand.

Figure 3.1.3



The city frequency count analysis reveals the distribution of data entries across

different cities, indicating key market trends and regional focuses. Beijing leads with the highest frequency of 16,981 entries, followed closely by Chengdu (15,448) and Guangzhou (14,842). Other major cities such as Chongqing (14,012), Shanghai (13,460), and Hangzhou (13,299) also show significant representation. These cities, typically characterized by their large populations and developed economies, are likely the primary markets for the analyzed data.

The top cities emphasize regions of concentrated activity or interest, reflecting patterns in consumer demand or market presence. Cities such as Tianjin, Suzhou, Shenzhen, and Xi'an all show frequencies above 12,000, indicating robust engagement in these areas. Notably, the frequency distribution decreases steadily after the top 15 cities, suggesting that activity is more centralized in major urban hubs compared to smaller or less economically dominant cities. This trend underscores the importance of these urban centers in shaping broader market dynamics and serves as a valuable insight for targeted analysis or strategic planning.

Figure 3.1.4

```
print("\n市的频率计数:")
city_counts = df['市'].dropna().value_counts().reset_index()
pd.set_option('display.max_rows', 1000)  # 设置最大显示行数为1000
print(city_counts)
```

```
市的频率计数:
    市  count
0   北京  16981
1   成都  15448
2   广州  14842
3   重庆  14012
4   上海  13644
5   杭州  13299
6   郑州  12897
7   天津  12592
8   苏州  12508
9   深圳  12053
10  西安  12007
11  武汉  11855
12  长沙  10484
13  昆明  10092
14  济南   9615
```

```
print("\n市的频率计数:")
city_counts = df['市'].dropna().value_counts().reset_index()
pd.set_option('display.max_rows', 1000)  # 设置最大显示行数为1000
print(city_counts)
```

```
375          阿拉善      30
376           海南      29
377           昌都      28
378         山南地区      25
379           海北      24
380     克孜勒苏柯尔克孜      19
381           阿里      19
382     甘南藏族自治州      18
383           玉树      16
384           山南      12
385     果洛藏族自治州      10
386           甘孜       8
387     黄南藏族自治州       6
388           甘南       6
389           果洛       4
390           琼南       3
391           黄南       3
392           三沙       2
```

The first dataset reveals a significantly larger distribution of records across companies compared to the second one. For instance, in the first image, 一汽大众 (FAW-Volkswagen) leads with a count of 83,624, followed by 吉利汽车 (Geely) with 67,346. Other prominent companies include 上汽通用五菱 (SAIC-GM-Wuling) and 上海大众 (Shanghai Volkswagen), with counts of 63,788 and 63,080, respectively. This indicates that these companies dominate in terms of representation in the dataset, likely due to their strong market presence or larger sales volumes.

In contrast, the second dataset represents a smaller and less evenly distributed dataset. For example, the highest count, 新日汽车 (Xinday), has only 23 occurrences, followed by 众泰汽车 (Zotye) with 21 occurrences. This dataset is more fragmented, possibly due to limited scope or data size.

Therefore, the first dataset gives a broader perspective of the industry with clear leaders, while the second dataset may represent a niche segment or a filtered view.

Figure 3.1.5

```
print("\n企业的频率计数:")
company_counts = df['企业'].dropna().value_counts().reset_index()
print(company_counts)
```

```
企业的频率计数:
          企业  count
0       一汽大众  83624
1       吉利汽车  67346
2     上汽通用五菱  63788
3       上海大众  63080
4       长安汽车  62927
5       上海通用  61798
6       长城汽车  57494
7       东风日产  47754
8       奇瑞汽车  44964
9       广汽丰田  43736
10      一汽丰田  42364
11      东风本田  38503
12      华晨宝马  37321
13     比亚迪汽车  32326
14      广汽本田  31522
```

```
print("\n企业的频率计数:")
company_counts = df['企业'].dropna().value_counts().reset_index()
print(company_counts)
```

```
92      新日汽车      23
93      众泰汽车      21
94      新特汽车      11
95      速达汽车       9
96      恒润汽车       9
97      汉龙汽车       8
98      领途汽车       8
99      江苏九龙       8
100     摩登汽车       7
101     江铃雷诺       5
102     开沃汽车       5
103     御捷汽车       4
104     云雀汽车       2
105     恒天汽车       1
106    比德文汽车       1
107     河北中兴       1
108     国金汽车       1
109     北汽银翔       1
```

The frequency count of the column “品牌” (brands) highlights the distribution of records for various automobile brands. In the first dataset (larger data range), 大众 (Volkswagen) has the highest count of 94,689, indicating its leading presence in the dataset, followed by 丰田 (Toyota) with 85,872, and 本田 (Honda) with 68,401. Other significant players include 长安 (Changan) and 吉利 (Geely), with counts of

61,459 and 54,586, respectively. This shows that the dataset is heavily skewed towards these popular brands, likely reflecting their strong market position and substantial sales volumes. The dominance of global and domestic top players indicates that the data captures a comprehensive picture of the competitive landscape in the automobile market.

In contrast, the second dataset represents a much smaller and fragmented distribution, with the highest counts, such as 恒润 (Hengrui) and BEIJING, each having only 9 records. Other brands like 大马 (Dama) and 摩登 (Moden) have even fewer records. This dataset may represent niche or lesser-known brands, possibly focusing on specific segments of the market.

In summary, the first dataset offers a broad view of the automobile brand landscape, emphasizing top global and domestic players, while the second dataset highlights a limited scope with minor or emerging brands, suggesting a focus on niche or specific market studies.

Figure 3.1.6

```
print("\n品牌的频率计数:")
brand_counts = df['品牌'].dropna().value_counts().reset_index()
print(brand_counts)
```

```
品牌的频率计数:
          品牌  count
0         大众  94689
1         丰田  85872
2         本田  68401
3         长安  61459
4         吉利  54506
5         五菱  46858
6         日产  43803
7         哈弗  42127
8         宝马  37321
9         奥迪  35794
10        别克  33428
11       比亚迪  32326
12        奔驰  30568
13        奇瑞  27145
14        传祺  25275
```

```
print("\n品牌的频率计数:")
brand_counts = df['品牌'].dropna().value_counts().reset_index()
print(brand_counts)
```

```
110       恒润      9
111  BEIJING      9
112       汉龙      8
113       大马      8
114       摩登      7
115    EZoom      5
116       开沃      5
117       御捷      4
118       云雀      2
119       极氪      2
120       中华      2
121       中兴      1
122       国金      1
123       恒天      1
124     北汽威旺      1
125      比德文      1
126       领途      1
127     北汽幻速      1
```

The frequency count of the column “车型” (vehicle models) provides insights into the popularity and representation of various car models in the dataset.

In the first dataset (larger and more detailed), the most frequent model is 宏光 MINI (Wuling Hongguang MINI) with 22,237 records, followed by 哈弗 H6 (Haval H6) with 21,477 records and 轩逸 (Nissan Sylphy) with 19,532 records. Other popular models include 卡罗拉 (Toyota Corolla) and 朗逸 (Volkswagen Lavida), with 15,432 and 12,635 records, respectively. These numbers indicate a strong dominance of economical, compact, and family-oriented vehicles, reflecting consumer preferences in the dataset, possibly indicative of broader market trends.

In contrast, the second dataset is much smaller and highly fragmented, with all models appearing only once. Examples include 哈弗 H5 (Haval H5), 捷达 (Jetta), 马自达 3 星骋 (Mazda3 Xingcheng), and 五菱宏光 S (Wuling Hongguang S). This suggests the dataset may focus on unique or less commonly recorded models, possibly representing outliers or specialized cases.

Thus, the first dataset reflects widespread consumer preferences and trends, highlighting top-selling, mass-market models. The second dataset, with its unique and non-repeated entries, might cater to niche, rare, or individualized aspects of the market. Together, these datasets illustrate the contrast between mass-market trends and specialized subsets.

Figure 3.1.7

```
print("\n车型的频率计数:")
print(df['车型'].dropna().value_counts())
```

```
车型的频率计数:
车型
宏光MINI          22237
哈弗H6            21477
轩逸              19532
卡罗拉             14198
朗逸              12635
宝马3系             11548
RAV4             10924
博越               10755
雷凌               10695
CR-V             10626
A6L              10250
宝马5系              10055
宝来                9878
GL8               9529
凯美瑞               9181
```

```
print("\n车型的频率计数:")
print(df['车型'].dropna().value_counts())
```

```
哈弗H5                    1
捷达                     1
马自达3星骋                   1
五菱宏光S                   1
奔腾70S                   1
翼搏                     1
奔腾B50                   1
江淮iEV7S                 1
奥迪A6L                   1
海马6P                    1
国金GM3                   1
瑞虎8 PLUS                1
海马爱尚                    1
BJ40                   1
九龙EM5                   1
凯翼C3R                   1
五菱Nano                  1
Name: count, dtype: int64
```

The pie chart represents the frequency distribution of various fuel types in the dataset.

Gasoline (汽油) is overwhelmingly dominant, accounting for 84.4% of the total entries. This suggests that traditional internal combustion engine vehicles still make up the vast majority of vehicles in this dataset.

Pure Electric (纯电动) is the second most common fuel type, comprising 10.2%

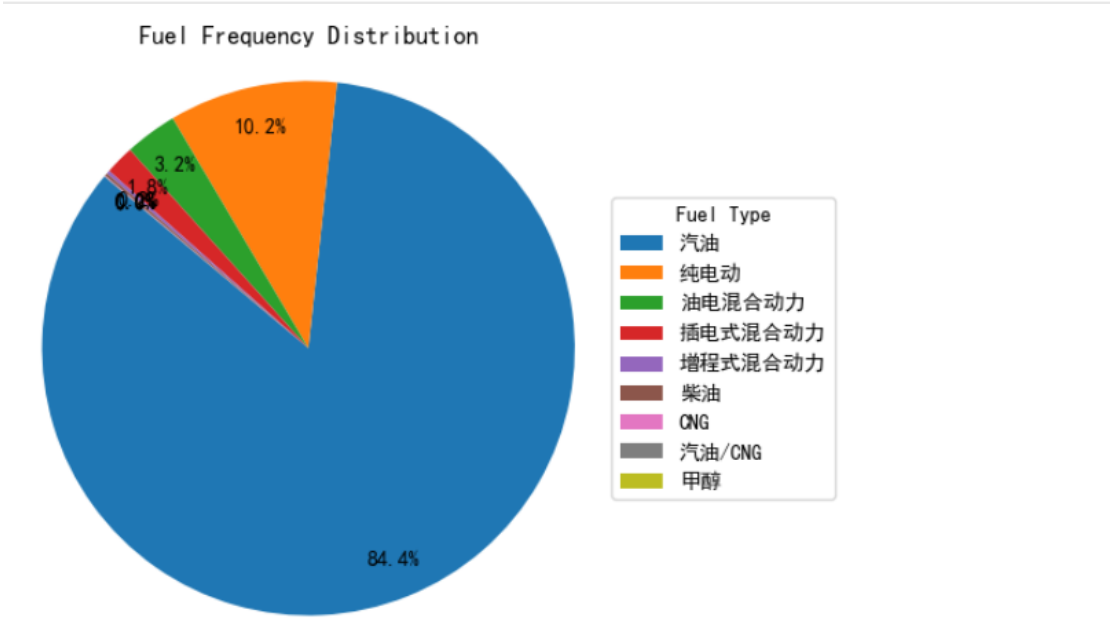
of the data. This indicates a growing interest and adoption of electric vehicles, likely driven by environmental concerns and advancements in electric vehicle technology.

Other fuel types such as Hybrid Electric (油电混合动力) (3.2%) and Plug-in Hybrid Electric (插电式混合动力) (0.8%) make up smaller but notable portions of the dataset. These options reflect the transition towards cleaner and more energy-efficient technologies.

Less common fuel types include Diesel (柴油), CNG (Compressed Natural Gas), Gasoline/CNG (汽油/CNG), and Methanol (甲醇), each accounting for a negligible share. This indicates limited market penetration or specialized use cases for these fuels.

All in all, the dataset highlights a clear preference for gasoline-powered vehicles, with emerging adoption of electric and hybrid technologies signaling a shift toward sustainable alternatives. However, alternative fuels such as CNG and methanol remain niche, possibly reflecting their limited infrastructure or market demand. This distribution provides valuable insight into current trends and potential areas of growth in the automotive fuel market.

Figure 3.1.8



The bar chart illustrates the frequency distribution of different car body types in

the dataset.

SUV dominates the dataset with the highest frequency, surpassing 500,000 entries. This reflects the growing consumer preference for SUVs, likely due to their versatility, spaciousness, and suitability for various driving conditions.

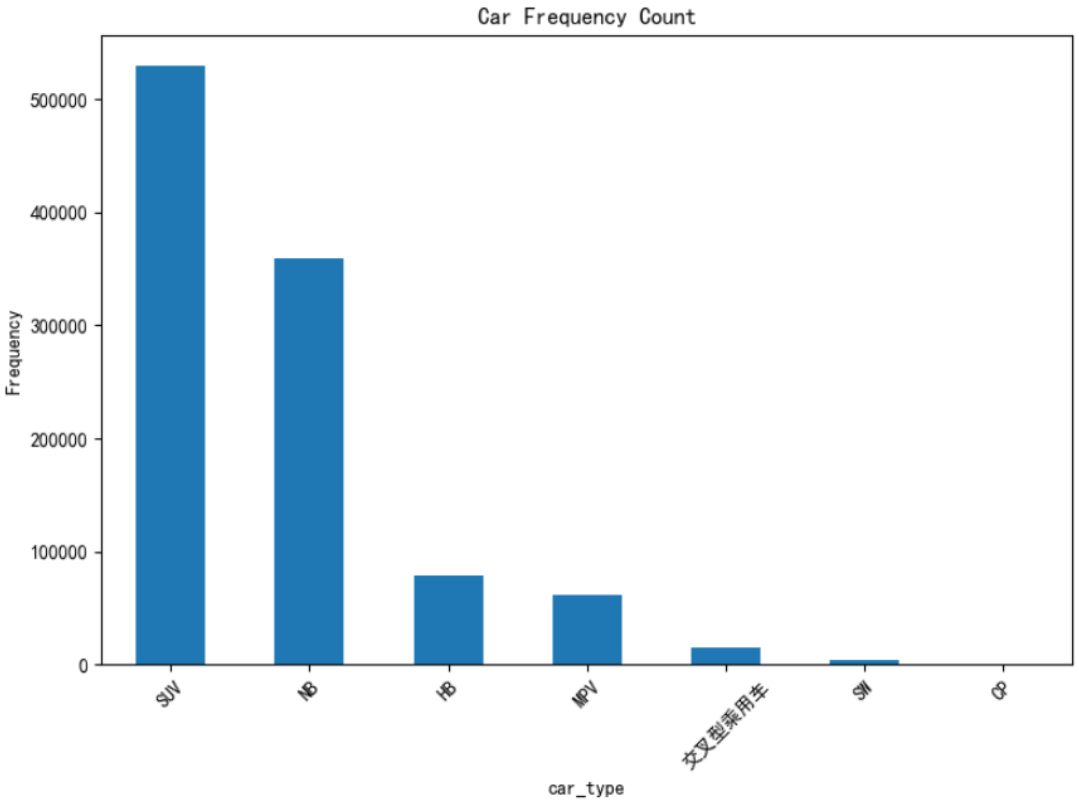
NB (Notchback) is the second most frequent car body type, with over 300,000 entries, suggesting that traditional sedan-style vehicles remain a popular choice for many buyers, possibly due to their affordability and practicality.

Other body types, such as HB (Hatchback) and MPV (Multi-Purpose Vehicle), show moderate representation, indicating their niche but significant appeal to specific market segments, such as families or urban commuters.

The remaining body types, including Passenger Commercial Vehicles (交叉型乘用车), SW (Station Wagon), and QP (Convertible), have minimal frequencies. These represent either specialized uses or less demand in the mainstream market.

In summary, the dataset highlights the dominance of SUVs and sedans (NB) in the automotive market, reflecting their broad consumer appeal. In contrast, body types like hatchbacks and MPVs cater to more specific needs, while station wagons and convertibles remain niche. This distribution underscores the current trends and preferences in vehicle design and functionality, with a clear shift toward versatile and family-oriented vehicles.

Figure 3.1.9



The bar chart represents the frequency distribution of vehicle engine displacements in the dataset.

2.0L and 1.5T displacements dominate the dataset, with frequencies exceeding 200,000 entries each. This indicates a strong preference for mid-sized engines that balance performance and efficiency, particularly turbocharged engines (indicated by "T") such as 1.5T.

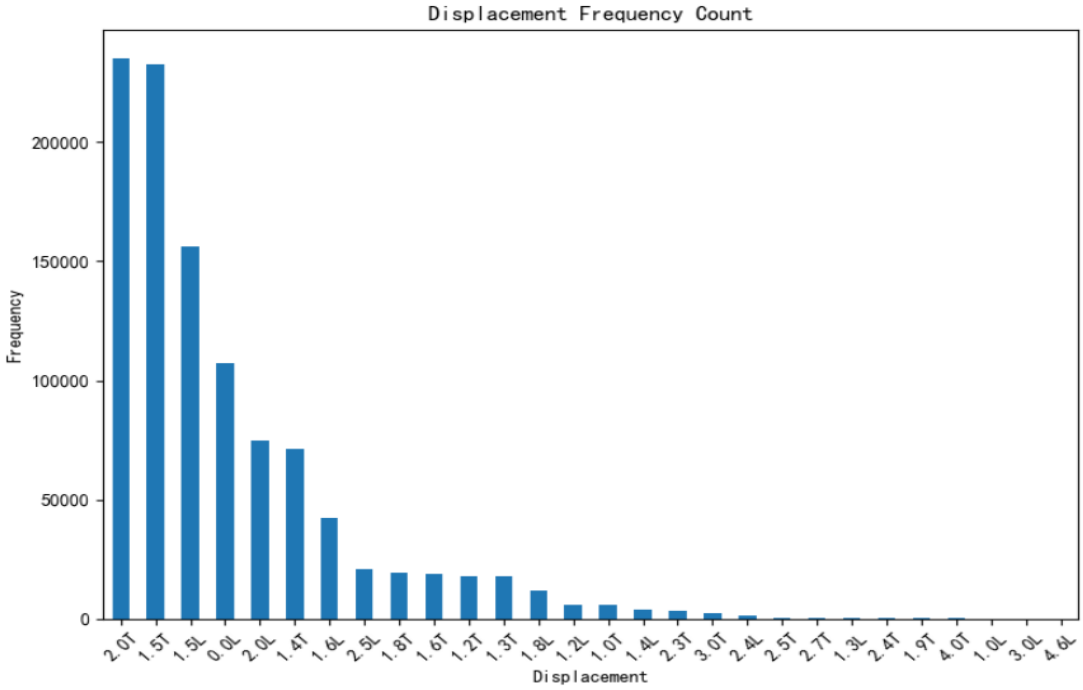
1.5L and 2.0T follow as the next most common displacements, showing a steady interest in both naturally aspirated and turbocharged engines with smaller to medium capacities. Their popularity likely stems from their efficiency and suitability for a wide range of vehicle types, including sedans and SUVs.

Lower displacements such as 1.0L and 1.4T have moderate representation, appealing to consumers seeking fuel-efficient and cost-effective vehicles, particularly in urban areas.

Larger displacements like 2.5L, 3.0L, and above are less frequent, reflecting their niche market appeal. These engines are typically found in premium or performance vehicles, which cater to a smaller segment of the market.

Overall, the dataset underlines a clear trend towards mid-sized engine displacements (2.0L and 1.5T), reflecting a balance between performance and fuel efficiency. Smaller displacements target the cost-sensitive and eco-conscious consumer base, while larger engines cater to a niche market for high-performance or luxury vehicles. This distribution aligns with global trends emphasizing fuel efficiency and environmental sustainability without sacrificing power.

Figure 3.1.10



The bar chart presents the frequency distribution of different transmission types in the dataset. Below is the analysis:

7DCT (7-speed Dual-Clutch Transmission) and CVT (Continuously Variable Transmission) are the most common transmission types, with frequencies exceeding 250,000 records each. This reflects a strong market preference for these transmissions due to their balance of efficiency, smooth driving experience, and performance.

6AT (6-speed Automatic Transmission) and 9AT (9-speed Automatic Transmission) follow as the next most frequent types, with significant representation in the dataset. These automatic transmissions are favored for their ability to enhance driving comfort and improve fuel efficiency.

1AT (Single-speed Automatic Transmission), typically associated with electric

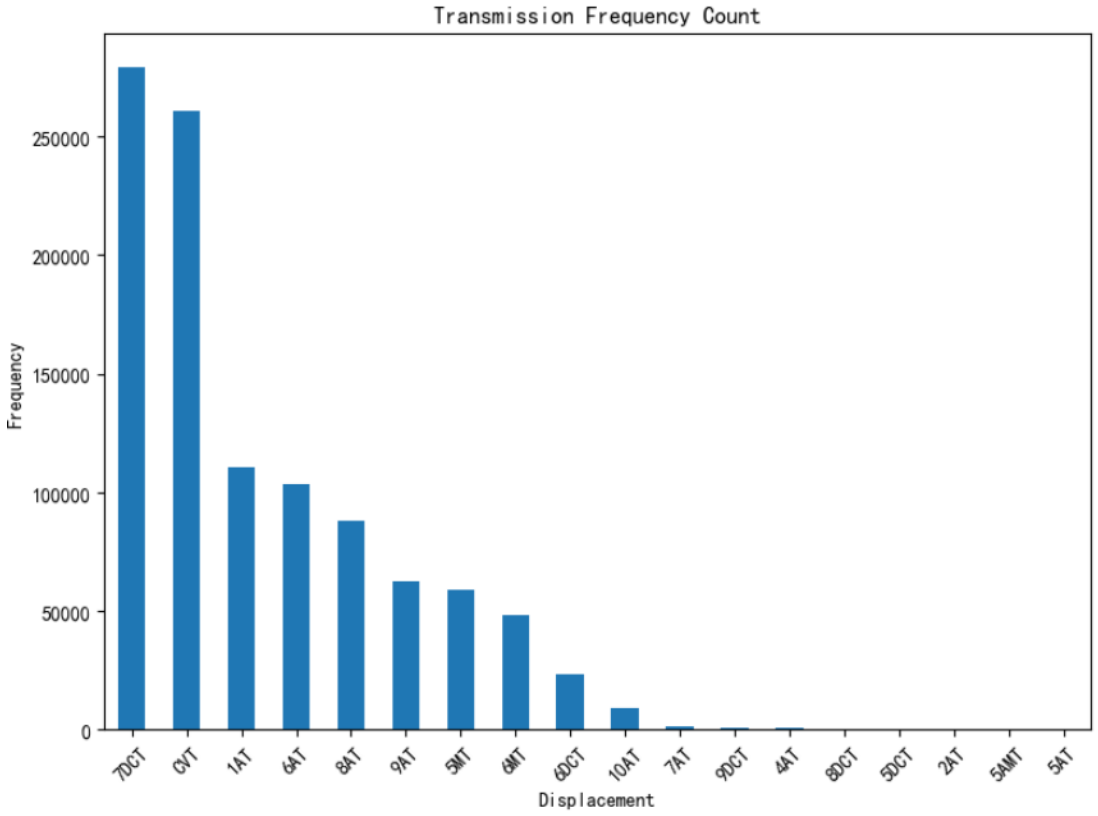
vehicles, also has a notable presence, indicating the growing adoption of EVs and their simplified transmission requirements.

Manual transmissions, such as 5MT (5-speed Manual Transmission) and 6MT (6-speed Manual Transmission), show lower frequencies, reflecting a declining preference for manual cars as automatics become more affordable and widely available.

Advanced and less common options, such as 10AT and 8DCT, have minimal representation, suggesting they are currently limited to premium or high-performance vehicle segments.

Therefore, the dataset stresses the dominance of modern automatic and dual-clutch transmissions, such as 7DCT and CVT, reflecting trends toward enhanced efficiency and ease of use. While manual transmissions maintain a small share, their popularity is waning in favor of automatic alternatives. The significant presence of 1AT suggests the impact of increasing EV adoption on the transmission market. These trends align with the global automotive industry's focus on efficiency, technology, and user convenience.

Figure 3.1.11



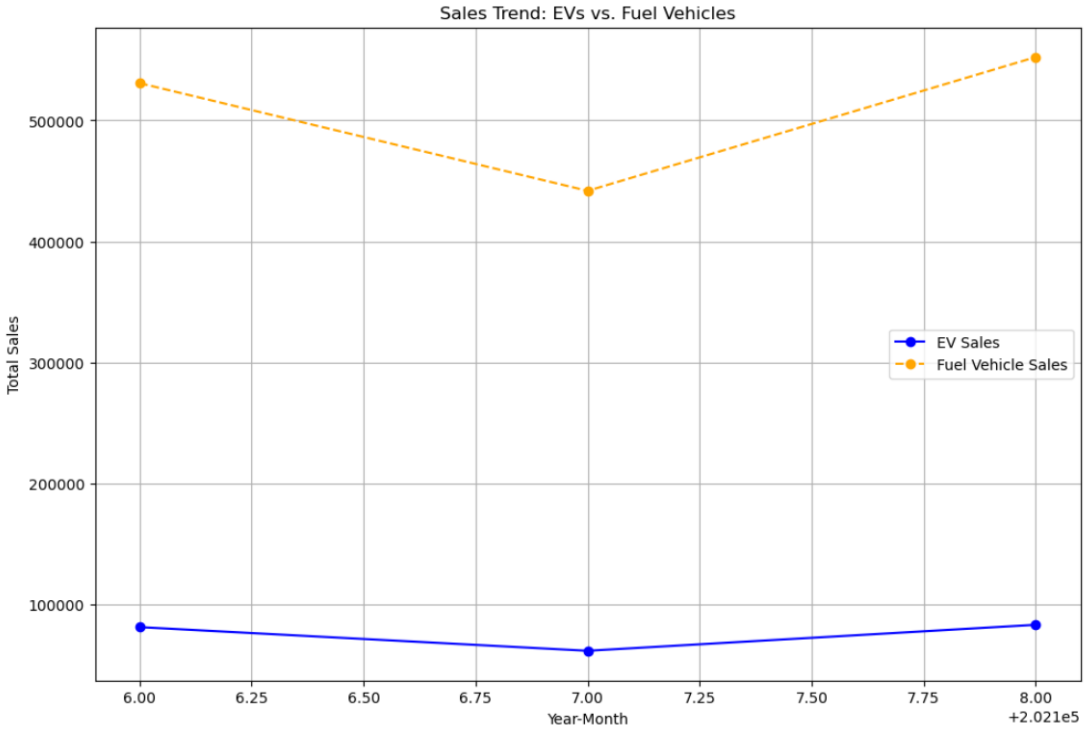
3.2 Sales Analysis

The sales trend chart illustrates the monthly sales performance of Electric Vehicles (EVs) compared to Fuel Vehicles over the observed period. The data reveals a consistent dominance of fuel vehicle sales, which remain significantly higher than those of EVs throughout the timeline. However, a noticeable fluctuation in sales trends for both categories is observed.

Fuel vehicle sales exhibit a decline from the beginning of the period, reaching a low point midway, and then recovering toward the end. This pattern could be influenced by external factors such as seasonal demand, economic conditions, or market-specific events. On the other hand, EV sales follow a relatively stable trajectory, with a slight dip in the middle of the timeline, mirroring the trend in fuel vehicles, and a subsequent recovery towards the end. Despite the lower overall sales figures for EVs, the gradual upward trend in the latter part of the timeline suggests growing consumer interest or improving market conditions for electric vehicles.

The gap between the two categories highlights the significant lead of fuel vehicles in the market, yet the stability and potential growth of EV sales reflect an evolving market dynamic. The consistent performance of EVs, coupled with their recovery towards the end of the period, underscores their increasing relevance in the automotive market, likely driven by advancements in EV technology, government incentives, or shifting consumer preferences. The data points to an ongoing transition, with EVs steadily carving out a more substantial share of the market despite the current dominance of fuel vehicles.

Figure 3.2.1



Based on the analysis of the top-selling car models and brands, the results indicate significant dominance by specific models and manufacturers in the automotive market. Among the car models, "轩逸" takes the lead with 59,496 units sold, followed by "朗逸" with 45,219 units and "宏光 MINI" with 42,443 units. These models are characterized by their affordability and practicality, which make them popular choices among consumers. Additionally, higher-end models such as "卡罗拉" (38,487 units) and "雷凌" (30,422 units) also perform well, reflecting a balanced preference for both budget and premium options in the market. Other notable models include "英朗" (29,819 units), "凯美瑞" (25,454 units), "RAV4" (23,098 units), "长安 CS75PLUS" (22,596 units), and "XR-V" (21,610 units), showcasing a diverse range of sedan and SUV categories among top sellers.

On the brand level, Volkswagen emerges as the market leader with an impressive 233,319 units sold, reflecting its strong presence and reputation in delivering reliable and versatile vehicles. Toyota follows with 193,610 units, further cementing its position as a top global brand. Honda secures third place with 152,792 units, showing robust consumer demand for its models. Domestic brands also perform notably well, with Changan (128,078 units) and Geely (116,635 units) leading the charge, demonstrating the increasing competitiveness of Chinese automakers. Other prominent brands include Nissan (119,729 units), Wuling (96,622 units), BYD (89,887 units), Buick (87,758 units), and BMW (73,800 units). This highlights a blend

of both international and domestic brands achieving significant market share.

Therefore, the results underline the dominance of well-established brands and their ability to offer diverse vehicle models catering to a wide array of consumer preferences. Sedans and SUVs remain the most popular categories, with affordability, brand reputation, and practicality being key drivers of sales performance.

Figure 3.2.2

Top-selling brands:

	品牌	数量
40	大众	233319
12	丰田	193610
77	本田	152792
116	长安	128078
73	日产	119729
30	吉利	116635
16	五菱	90622
85	比亚迪	89087
22	别克	87758
52	宝马	73800
49	奥迪	73409
35	哈弗	71236
48	奔驰	55848
18	传祺	49947
97	现代	47282
46	奇瑞	45101
105	荣威	41033
102	红旗	33406
95	特斯拉	29076
118	雪佛兰	28757
21	凯迪拉克	27113
100	福特	25434
123	领克	25134
53	宝骏	23343
8	东风	22447
66	捷达	20097
126	马自达	19999
110	起亚	19149
91	沃尔沃	17628
4	WEY	16600

	车型	数量	厂商指导价(元)
529	轩逸	59496	144919.061028
351	朗逸	45219	144507.281978
231	宏光MINI	42443	69170.625534
146	卡罗拉	38407	146572.151007
594	雷凌	30042	170568.115942
498	英朗	29819	138369.855506
117	凯美瑞	25454	243109.879098
49	RAV4	23098	232316.477481
571	长安CS75PLUS	22596	137206.139184
57	XR-V	21610	161918.891865

The analysis of sales performance across different regions in China reveals notable differences among the Northwest, Southern, Northern, and Central regions. The Southern region, particularly Guangdong Province, dominates the sales landscape with an impressive 259,091 units sold, significantly outpacing all other provinces. This can be attributed to the region's strong economic development, dense urban population, and robust infrastructure. The Northern region, represented by provinces like Shandong and Hebei, also performs well, with Shandong achieving 171,247 units and Hebei contributing 115,816 units. These provinces benefit from their

industrialized economies and sizable populations, which drive consumer demand for automobiles.

In the Central region, provinces like Henan and Hunan showcase strong performances, with sales of 136,387 and 76,631 units, respectively. This highlights the growing economic strength and rising disposable incomes in central China. Meanwhile, the Northwest region, which includes provinces like Shaanxi and Xinjiang, records relatively lower sales volumes. For example, Shaanxi reports 66,398 units, while Xinjiang achieves 42,364 units. The lower population density and economic development in the Northwest region contribute to its more modest performance compared to the more urbanized and industrialized regions.

The analysis of sales distribution across different cities and provinces reveals significant disparities, driven by regional economic development, population density, and purchasing power. Guangdong Province, the standout performer with 259,091 units sold, is home to several major cities such as Guangzhou and Shenzhen, which serve as economic powerhouses in southern China. These cities not only have a high density of affluent consumers but also benefit from well-established transportation networks and robust urban development, making them key hubs for automobile sales.

Shandong, the second-highest province with 171,247 units sold, showcases the strength of the Northern region. Major cities like Qingdao and Jinan contribute significantly to these figures, supported by their industrial bases and growing middle-class populations. Similarly, Zhejiang and Jiangsu, reporting 154,197 and 149,152 units respectively, demonstrate the influence of the Yangtze River Delta's economic zone. Cities such as Hangzhou, Ningbo, Nanjing, and Suzhou play pivotal roles in these provinces' strong performances due to their high urbanization rates and thriving local economies.

In the Central region, Henan leads with 136,387 units sold, followed by Hunan with 76,631 units. These provinces, while not coastal, benefit from a burgeoning consumer base in cities like Zhengzhou and Changsha, which are rapidly growing as regional economic centers. The increasing disposable income levels in these areas, coupled with improving road infrastructure, have created a fertile market for automobile sales. Provinces like Anhui (75,787 units) and Hubei (71,647 units) also reflect this trend, with cities like Wuhan acting as regional growth engines.

The Western and Northwestern provinces, such as Shaanxi (66,398 units) and

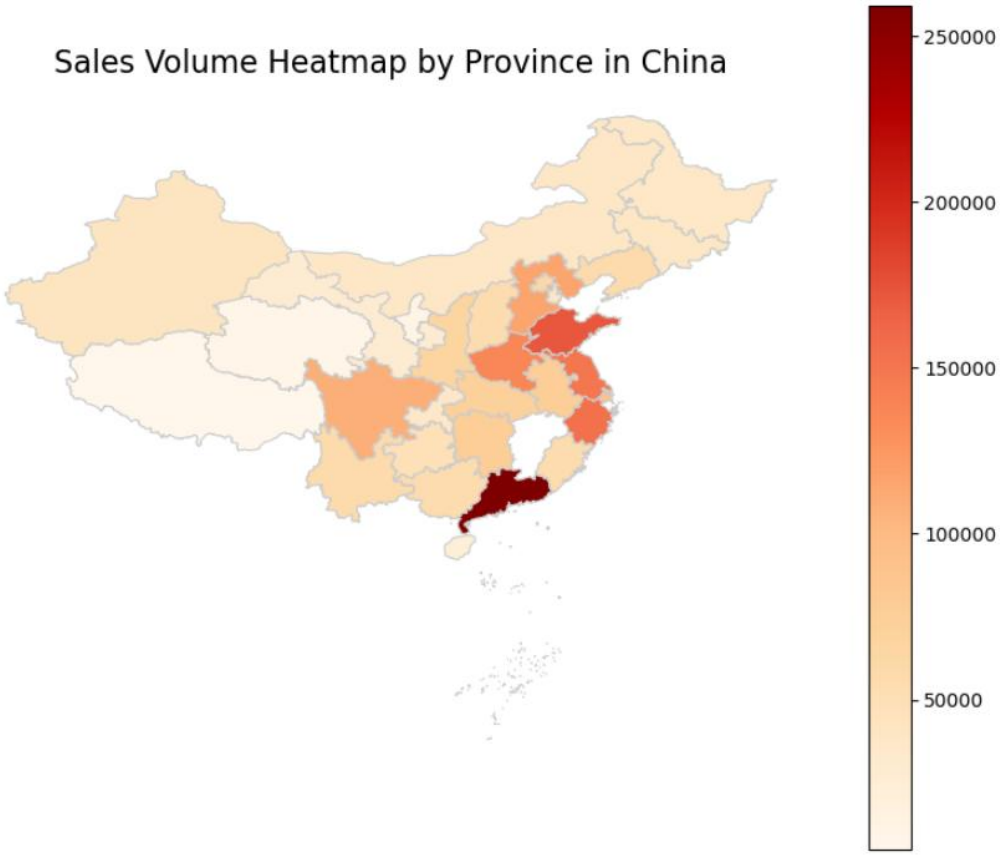
Xinjiang (42,364 units), show comparatively lower sales figures. Xi’an, the capital of Shaanxi, stands out as a regional hub with higher sales than surrounding areas, but the overall contribution from these regions remains modest due to lower population density and economic development. Similarly, provinces like Tibet (4,811 units) and Ningxia (9,639 units) report the lowest sales, highlighting the challenges posed by sparse populations, rugged terrain, and less developed infrastructure.

Municipalities like Shanghai and Beijing also warrant attention. Despite their smaller geographical size compared to provinces, they report sales of 89,918 and 62,836 units respectively, underscoring their importance as highly urbanized and economically affluent markets. In Shanghai, the high concentration of wealth and the city's status as an international business hub drive significant automobile demand. Beijing, as the capital, benefits from similar factors, although its stricter vehicle restrictions and environmental policies may slightly temper its overall sales volume.

Additionally, provinces like Sichuan (188,514 units) and Chongqing (35,939 units) highlight the potential of Southwestern China. Chengdu, the capital of Sichuan, is a key market due to its status as a regional transportation and economic hub, while Chongqing, with its industrial focus, demonstrates strong growth potential despite relatively lower sales compared to top-tier provinces.

In summary, the sales distribution in cities and provinces is heavily influenced by regional economic factors. Coastal provinces and major urban centers dominate the market, reflecting their economic affluence and infrastructure advantages. Inland provinces and cities show emerging potential, driven by urbanization and rising incomes, but they still lag behind due to structural challenges. Future growth strategies could focus on tailoring products to meet the unique demands of less urbanized regions while leveraging infrastructure investments to improve accessibility in these areas.

Figure 3.2.3



The comparison of sales between Electric Vehicles (EVs) and Fuel Vehicles demonstrates a significant disparity, both in terms of total sales volume and price distribution. Fuel Vehicles account for a much larger share of the market, as reflected in their total sales volume of over 1.5 million units, compared to just under 100,000 units for EVs. This substantial difference highlights the current dominance of fuel vehicles in the automotive market, which is likely driven by their long-established infrastructure, greater availability, and consumer familiarity.

From the summary statistics, the average factory price of EVs is ¥185,405, slightly lower than the average price of fuel vehicles at ¥212,289. However, the price range for EVs is broader, with a maximum price of ¥1,812,000, compared to ¥759,800 for fuel vehicles. This indicates that the EV market includes a diverse range of models, from budget-friendly options to premium and luxury models. In contrast, the price distribution of fuel vehicles is more concentrated, reflecting the mature nature of the market.

The sales disparity also suggests that EVs, while growing in popularity, still face challenges in competing with fuel vehicles on a larger scale. These challenges may include limited charging infrastructure, higher upfront costs for some models, and

consumer hesitancy toward new technology. Despite these obstacles, the presence of high-end EV models and a competitive average price indicate a growing shift toward EV adoption, particularly as governments implement incentives and regulations to encourage greener transportation.

In conclusion, the current sales figures underscore the dominance of fuel vehicles, but they also highlight the potential of EVs to expand their market share, particularly as advancements in technology and infrastructure make them more accessible and appealing to a broader consumer base. The automotive market appears to be in a transitional phase, with EVs poised to capture a larger share in the coming years.

Figure 3.2.4

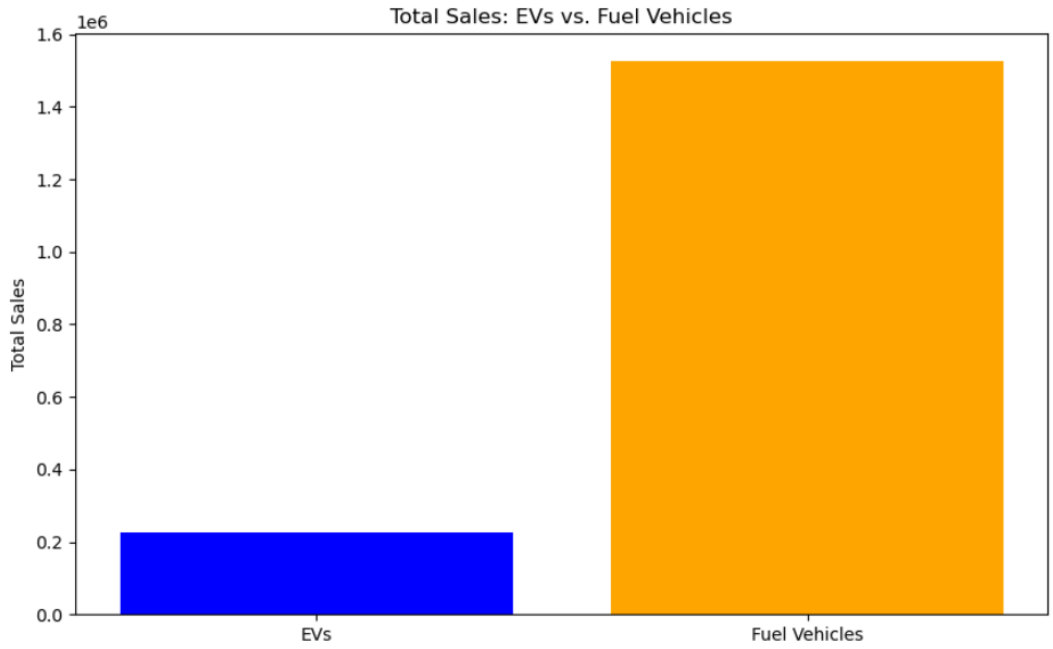
EV Summary Statistics:

	数量	厂商指导价(元)
count	92915.000000	92915.000000
mean	2.435559	185405.019362
std	9.308934	142366.308777
min	1.000000	32800.000000
25%	1.000000	69800.000000
50%	1.000000	146800.000000
75%	2.000000	239800.000000
max	1812.000000	800000.000000

Fuel Vehicle Summary Statistics:

	数量	厂商指导价(元)
count	734701.000000	734701.000000
mean	2.075507	212289.907756
std	3.883209	131965.159570
min	1.000000	28900.000000
25%	1.000000	123900.000000
50%	1.000000	159900.000000
75%	2.000000	254700.000000
max	861.000000	759800.000000

```
# Plot comparison of sales
plt.figure(figsize=(10, 6))
plt.bar(['EVs', 'Fuel Vehicles'], [ev_data['数量'].sum(), fuel_data['数量'].sum()], color=['blue', 'orange'])
plt.title('Total Sales: EVs vs. Fuel Vehicles')
plt.ylabel('Total Sales')
plt.show()
```



3.3 Price Analysis

Due to consider that the price may be decided for the kind of cars, we will divide cars into electric and fuel to discuss their price distribution.

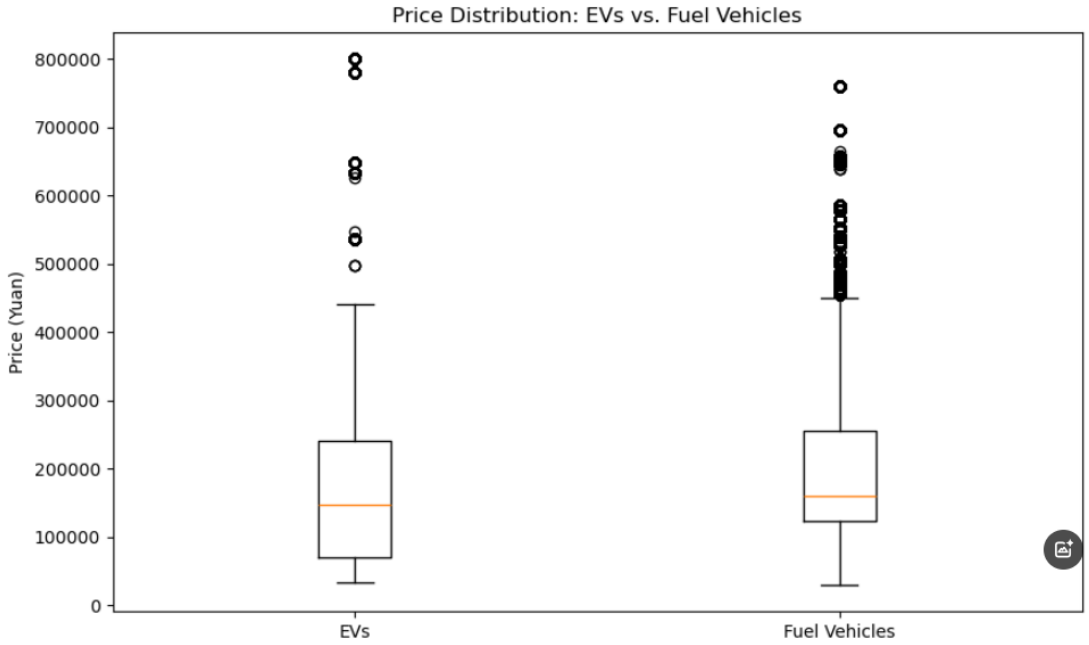
The price distribution of Electric Vehicles (EVs) and Fuel Vehicles reveals significant differences in their market positioning and variability. Overall, the median price of EVs is higher than that of fuel vehicles, indicating that EVs are generally positioned as more premium offerings. The interquartile range (IQR) for EVs is also broader, reflecting greater variability in pricing. This is likely due to the availability of both entry-level models targeting budget-conscious consumers and high-end luxury EVs designed for affluent buyers.

The presence of outliers in both categories highlights the existence of luxury or specialized models; however, EVs exhibit fewer but more extreme outliers, with prices exceeding ¥800,000. These outliers correspond to ultra-premium models, such as high-end electric SUVs or sedans. In contrast, fuel vehicles have a larger number of moderate outliers, representing a broader distribution of luxury fuel-powered cars. Notably, EVs have a higher upper price limit compared to fuel vehicles, emphasizing their ability to compete at the highest end of the market. Despite this, the lower price bounds of both categories are comparable, indicating the availability of affordable

options in both markets.

In summary, the analysis of price distributions reveals that EVs occupy a diverse market space with offerings ranging from budget-friendly to ultra-premium models, while fuel vehicles exhibit a more consistent pricing structure characteristic of a mature and saturated market. The higher median price and greater variability of EVs underscore the growing diversity and competitiveness of the electric vehicle market.

Figure 3.3.1



The analysis reveals several key correlations between car features and their manufacturer suggested retail price (MSRP). From the correlation matrix and feature importance analysis, it is evident that fuel type, car body type, and engine displacement are among the most significant factors influencing price.

The fuel type exhibits the highest importance score, suggesting that vehicles using alternative or premium fuel types generally have higher MSRP values. This aligns with trends where electric or hybrid vehicles, as well as vehicles requiring high-performance fuels, are often positioned as premium products in the market. Similarly, car body type plays a critical role, as larger or more specialized body types, such as SUVs or luxury sedans, tend to command higher prices compared to compact or hatchback models.

Engine displacement (排量) also shows a positive correlation with MSRP, which

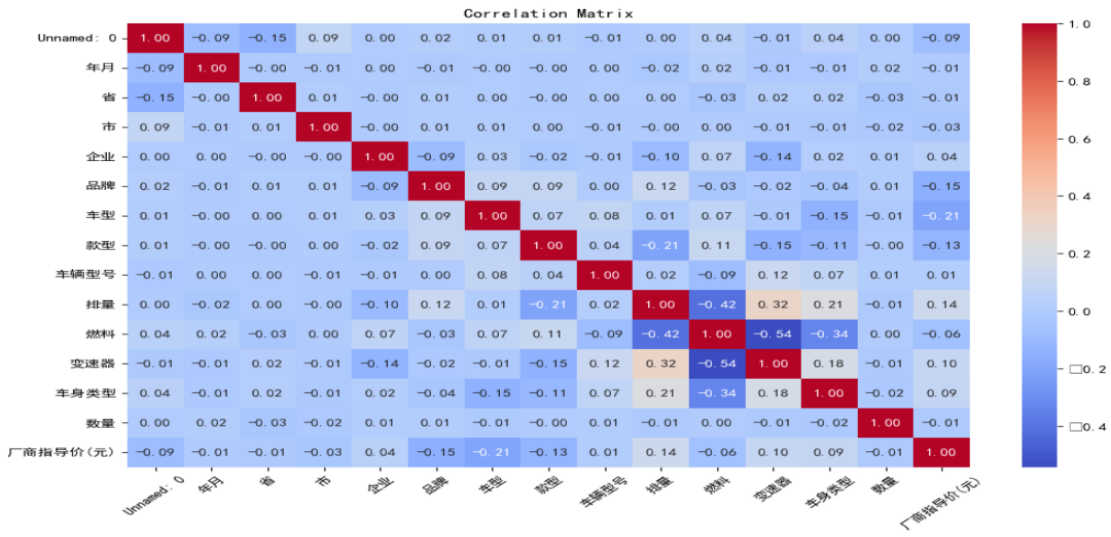
is consistent with the general market positioning of vehicles with larger engines as performance or luxury models. Transmissions (变速器) also display a moderate positive correlation, reflecting the impact of advanced transmission systems (e.g., automatic or dual-clutch) on vehicle pricing.

Conversely, certain features such as brand (品牌) and car model (车型) exhibit negative or weak correlations with MSRP. This could reflect the diverse pricing strategies employed by manufacturers within the same brand or model range, where entry-level variants may dilute the overall correlation. Additionally, variables like city (市) and province (省) have negligible correlations, indicating that geographic factors are not significant direct determinants of MSRP at the macro level.

The R^2 score of the regression model, however, is quite low (0.09), indicating that while certain features like fuel type, engine displacement, and transmission type are important, the overall model does not capture sufficient variance in MSRP. This suggests that other unmeasured factors, such as brand perception, marketing strategies, or additional technical specifications (e.g., safety features, technology packages), may play a crucial role in determining vehicle prices.

In summary, while key factors like fuel type, body type, and engine displacement are positively correlated with MSRP, the relatively low predictive power of the model underscores the complexity of pricing strategies in the automotive market, warranting further investigation into additional variables that influence MSRP.

Figure 3.3.2



```
R^2 score: 0.09943903802583565
Mean Squared Error: 15746856855.188566
  Feature   Importance
7      燃料  7238.794468
9    车身类型  3364.815321
6      排量  2739.345467
8     变速器  1977.356013
2      企业   258.473452
5      款型    -6.121237
1       市   -28.119529
0       省  -132.506975
4      车型  -138.239010
3      品牌  -553.320332

Correlations with 厂商指导价(元):
厂商指导价(元)      1.000000
排量             0.135773
变速器            0.100692
车身类型           0.089869
企业             0.038526
车辆型号           0.006282
年月            -0.005806
省             -0.008349
数量            -0.012023
市             -0.027523
燃料            -0.062661
Unnamed: 0   -0.087867
款型            -0.132690
品牌            -0.154172
车型            -0.208939
Name: 厂商指导价(元), dtype: float64
```

The analysis of the top-selling models and their average prices demonstrates notable differences in pricing strategies and market positioning among these popular vehicles. The Nissan Sylphy ("轩逸") leads in sales volume with over 59,000 units sold, paired with a relatively moderate average price of approximately 144,000 RMB. This balance between affordability and quality explains its strong appeal to a wide range of consumers in the compact sedan segment. Similarly, the Volkswagen Lavida ("朗逸"), another high-volume seller with over 45,000 units sold, maintains an average price close to Sylphy's, reinforcing the dominance of well-priced, reliable sedans in the Chinese automotive market.

In contrast, the Hongguang MINI EV ("宏光 MINI") stands out as a price outlier among the top sellers, with an average price significantly lower at around 69,000 RMB. Its affordability has made it an attractive choice for budget-conscious consumers, particularly in urban areas where compact electric vehicles are gaining popularity for short-distance commuting. The Toyota Corolla ("卡罗拉") and Toyota Levin ("雷凌") both occupy higher price brackets in the range of 145,000–178,000 RMB, reflecting their reputation for reliability and their positioning in the mid-tier

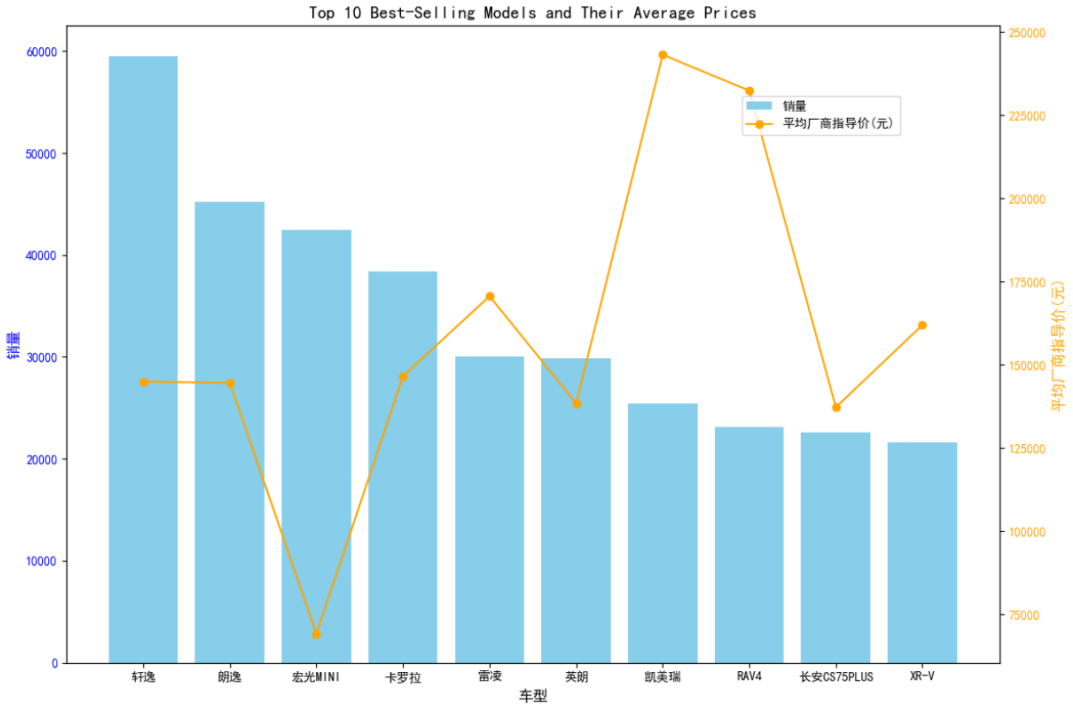
compact sedan market.

Premium models such as the Toyota RAV4 and the Honda CR-V show a distinct shift in pricing trends, with average prices exceeding 200,000 RMB. These higher price points are indicative of the SUV segment's ability to command a premium due to its larger size, better performance, and enhanced features. Interestingly, the Honda XR-V, while still categorized as an SUV, is priced significantly lower at around 161,000 RMB, making it a competitive choice for consumers seeking affordability in this segment.

Other models, such as the Buick Excelle ("英朗") and Toyota Camry ("凯美瑞"), exhibit balanced pricing strategies that cater to middle-income consumers, priced at approximately 138,000 RMB and 154,000 RMB, respectively. These models appeal to buyers who prioritize comfort and brand reliability without venturing into the luxury price segment.

Thus, the top-selling models illustrate a diverse range of price points, from the highly affordable Hongguang MINI EV to the more premium Toyota RAV4, reflecting the varied preferences and economic capabilities of Chinese consumers. While compact sedans dominate in terms of volume, the higher-priced SUVs showcase their growing popularity among affluent buyers seeking space and versatility. The data underscores how price positioning directly correlates with market segmentation and consumer priorities in the automotive landscape.

Figure 3.3.3

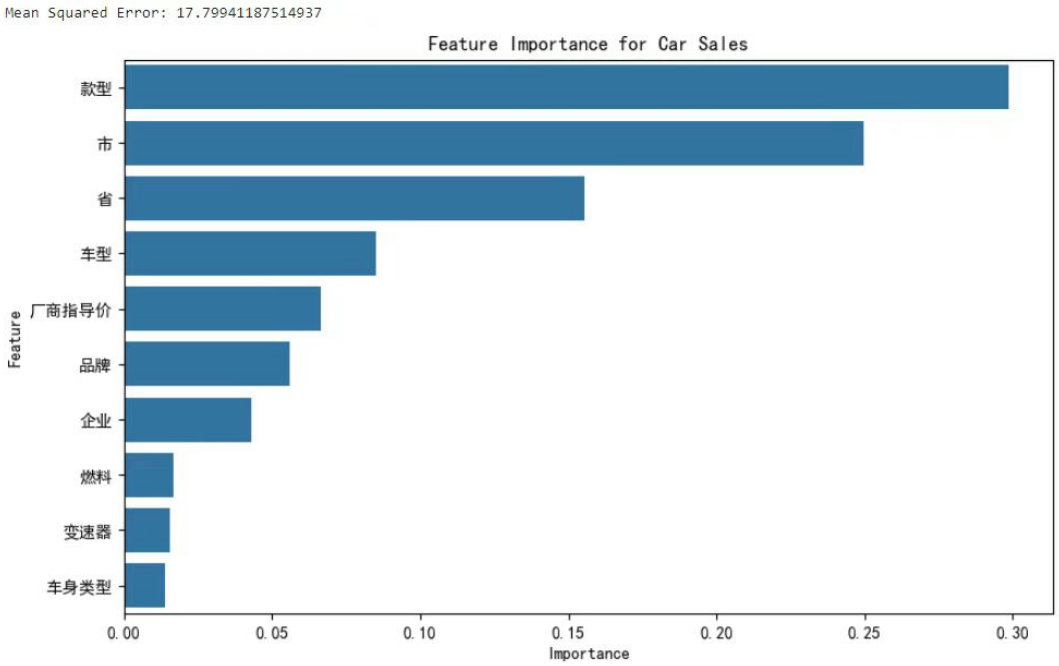


3.4 Business Insights

By analyzing the data, the feature importance chart clearly highlights the most influential factors in the automotive sales domain. Among all the factors, the most significant ones affecting car sales volume are: style, city, model, manufacturer's suggested retail price (MSRP), and brand. Style determines attributes such as aesthetics, comfort, safety, and energy efficiency, which are directly perceived by consumers. The scale, economic level, and urban development of different cities influence consumer demand for automobiles. Cost is a primary consideration for consumers. In highly price-sensitive market segments, competitive pricing can significantly influence consumer choices. Brand reputation and loyalty play a critical role in consumer decision-making. A strong brand commands a market premium, reflecting consumer trust and reliability in branded vehicles.

Other factors such as enterprise, fuel type, transmission, and body type also contribute to consumer purchase decisions but have a relatively smaller impact compared to the aforementioned attributes.

Figure 3.4.1



The bar chart comparing the total sales of electric vehicles (EVs) and fuel vehicles clearly shows that fuel vehicles vastly outperform EVs in total sales units. Fuel vehicles dominate the market with a significantly larger share, consistent with current global market trends. This dominance is supported by well-established infrastructure and consumer familiarity with fuel vehicles. Although the average sales per EV model (2.43) slightly exceed that of fuel vehicles (2.08), the difference is minimal, indicating that individual EV models can compete with fuel vehicles in terms of model-level sales.

The MSRP range for EVs is notably wide, encompassing both highly economical and ultra-high-end models, indicating a broad market targeting various consumer segments. The maximum MSRP for EVs is significantly higher than that of fuel vehicles, reflecting the inclusion of premium, high-cost electric models. Fuel vehicles have a narrower MSRP range, with a slightly higher average MSRP, possibly reflecting the market's maturity and focus on mid-to-high-end models.

The sheer number of fuel vehicles (more than seven times that of EVs) underscores their greater market penetration compared to EVs. This highlights that the automotive industry is undergoing a transitional phase, with EV adoption increasing but still far from the maturity level of the fuel vehicle market.

In summary, fuel vehicles currently hold a significant advantage in terms of total sales and market penetration, supported by a mature market and established infrastructure. However, EVs are rapidly catching up, with competitive model-level sales performance and a wide price range spanning from economical to luxury segments. The data suggests that as technology advances and consumer preferences shift towards more environmentally friendly options, EVs may gain a larger market share. The broader MSRP range and higher maximum price points for EVs also highlight growing investment and consumer interest in high-end electric models.

Figure 3.4.2

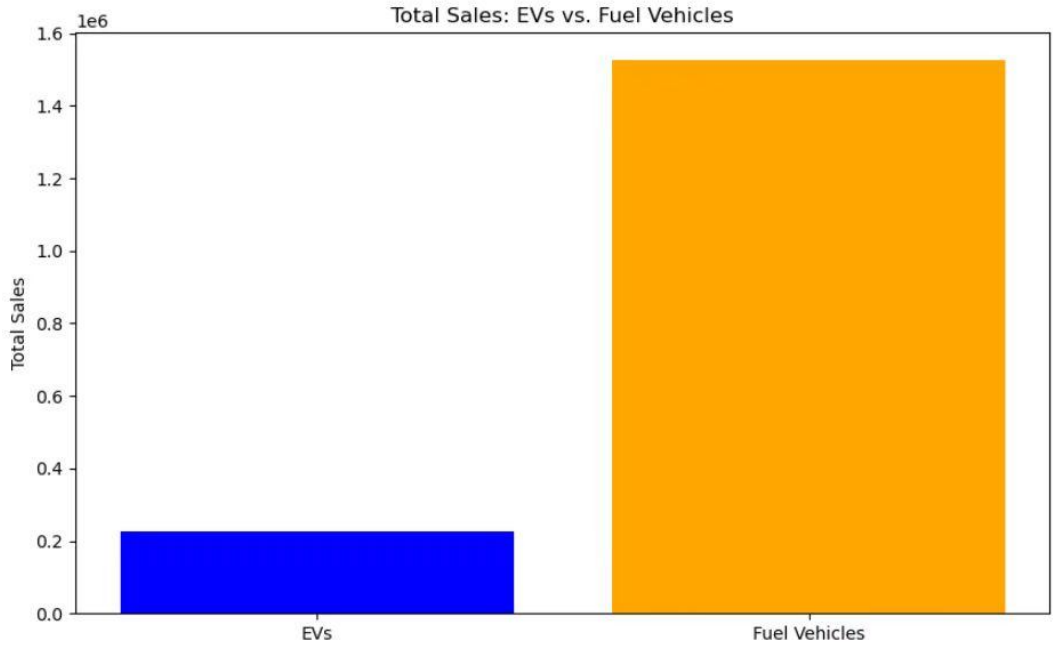
EV Summary Statistics:

	数量	厂商指导价(元)
count	92915.000000	92915.000000
mean	2.435559	185405.019362
std	9.308934	142366.308777
min	1.000000	32800.000000
25%	1.000000	69800.000000
50%	1.000000	146800.000000
75%	2.000000	239800.000000
max	1812.000000	800000.000000

Fuel Vehicle Summary Statistics:

	数量	厂商指导价(元)
count	734701.000000	734701.000000
mean	2.075507	212289.907756
std	3.883209	131965.159570
min	1.000000	28900.000000
25%	1.000000	123900.000000
50%	1.000000	159900.000000
75%	2.000000	254700.000000
max	861.000000	759800.000000

```
# Plot comparison of sales
plt.figure(figsize=(10, 6))
plt.bar(['EVs', 'Fuel Vehicles'], [ev_data['数量'].sum(), fuel_data['数量'].sum()], color=['blue', 'orange'])
plt.title('Total Sales: EVs vs. Fuel Vehicles')
plt.ylabel('Total Sales')
plt.show()
```



4 Conclution

4.1 Summary of key findings

Based on the analysis of the prices, sales, and various factors of both types of vehicles, this report draws the following conclusions:

（1） Internal Combustion Engine (ICE) Vehicles Still Dominate the Domestic Market

ICE vehicles currently dominate the market, with sales exceeding 1.5 million units, while electric vehicles (EVs) account for less than 100,000 units. The advantages of ICE vehicles are mainly due to the well-established infrastructure, such as gas station networks, and consumer familiarity and trust in their use. Additionally, ICE vehicles are generally more attractively priced, especially in the mid- to low-end market, offering a broad range that meets the needs of different consumers. Despite growing environmental awareness, ICE vehicles remain clearly ahead in terms of sales and market penetration due to their historical accumulation and mature market

positioning.

（2）The Electric Vehicle Market Shows an Upward Trend

Although current sales are relatively low, the electric vehicle market is steadily growing, indicating a gradual increase in consumer interest in new energy technologies. This growth is driven by policy support (such as government subsidies and tax incentives), technological improvements (like longer battery life and faster charging technology), and the widespread awareness of environmental issues. At the same time, EVs cover a wide price range, from economy models to high-end luxury vehicles, catering to consumers across different income levels. The potential market for electric vehicles is vast, especially as technology matures and consumer confidence grows.

（3）Price Distribution Differences Between Electric and Internal Combustion Engine Vehicles

The average price of electric vehicles is slightly lower than that of ICE vehicles, but their price range is broader, with high-end models significantly exceeding the price of the most expensive ICE vehicles. In comparison, ICE vehicle prices are more concentrated, reflecting the market's maturity and the higher proportion of mid- to high-end models. EVs include both budget-friendly models and ultra-high-end luxury vehicles, indicating that the market positioning of electric vehicles is gradually diversifying and expanding into higher-value segments. As more brands enter the EV market, price competition is likely to intensify, further enhancing the market appeal of electric vehicles.

（4）Regional and Economic Factors Significantly Affect the Car Market

Sales distribution reflects the direct impact of regional economic development and population density on car demand. Coastal provinces like Guangdong and Shandong, with their developed economies, high urbanization levels, and dense populations, are the primary car sales markets. In contrast, inland and economically underdeveloped regions, such as the northwest, show relatively low sales, suggesting

significant untapped market potential in these areas. By launching more affordable models and strengthening infrastructure, the consumer potential in these regions can be further explored, while balancing market development across different regions.

（5）Consumer Preferences Impact Car Sales

Consumers clearly prefer sedans and SUVs, as these vehicle types are popular for their practicality, versatility, and family-friendliness. Electric vehicles are also gaining favor, particularly in the economy segment, where price and functionality are well-balanced, as well as in the high-end EV market catering to luxury vehicle demands. Overall, consumers focus more on the cost-performance ratio, safety, and long-term operating costs when purchasing a vehicle. The low operational costs and environmental benefits of electric vehicles are allowing them to gradually gain popularity, especially among younger consumers.

（6）Brand Effect and Company Characteristics Impact Car Sales

In terms of brands, internationally known names like Volkswagen, Toyota, and Honda dominate the top sales positions, with their models trusted by consumers for reliability and consistent quality. Budget-friendly electric vehicles, such as the Hongguang MINI EV, stand out in the market due to their exceptional cost-performance ratio, while mid-range sedans like the Nissan Sylphy attract large numbers of consumers with their excellent overall performance. At the same time, domestic brands like Changan and Geely are gradually closing the gap with international brands, showing significant progress in technological innovation and market expansion by Chinese automakers.

4.2 Business insights and recommendations

Based on the conclusions derived from data analysis, this report proposes the following business recommendations:

（1）Strengthen the Electric Vehicle Market Presence

The market acceptance of electric vehicles is improving, but further efforts are needed for widespread adoption. Companies can focus on developing and promoting

mid-range priced EVs, which are affordable yet offer reasonable range and practicality, appealing to more mainstream consumers. Additionally, partnering with charging network operators to expand charging stations and optimize user experience will address consumer concerns about charging convenience. Moreover, businesses can help consumers understand the long-term value of EVs through transparent cost analyses and the promotion of operational advantages, fostering greater market recognition.

（2）Capitalize on Regional Market Opportunities

In terms of market strategy, companies should leverage the demand in coastal and developed urban markets, while increasing penetration in inland and economically underdeveloped regions. These areas may be more interested in affordable and practical models, making it effective to introduce competitively priced vehicles. Tailoring localized marketing strategies to regional needs, such as showcasing how EVs fit the local driving environment and offering regional purchase incentives, can significantly boost sales.

（3）Enhance Brand and Product Differentiation

Brand reputation and consumer trust are key factors in purchase decisions. Companies should increase investment in brand development, strengthen after-sales service, and enhance product quality and innovation to build a positive brand image. Additionally, creating unique product advantages, such as higher safety features, smarter in-car systems, and lower energy consumption, will help companies stand out in the competitive market. Introducing a diversified product line targeting different consumer segments can further expand market reach.

（4）Enhance Consumer Education

For potential consumers, education and guidance are crucial. Hosting test-drive events, information sessions, and online promotions will help consumers better understand the advantages of electric vehicles, such as lower operating costs and environmental benefits. Companies can also collaborate with government agencies to promote green car purchase incentives and discounted charging fees, lowering the initial purchase barriers and driving EV adoption.

（5）Utilize Pricing Dynamics

In price-sensitive markets, competitive pricing strategies will be an important

tool to attract consumers. For example, offering seasonal discounts or bulk purchase incentives can attract price-conscious consumer groups. At the same time, companies should expand the product line for high-end models to meet the luxury car demands of high-income consumers, increasing brand premium capabilities.

（6）Monitor Market Transformation

As environmental regulations tighten and consumer focus on sustainability increases, the car market is undergoing significant transformation. Companies need to closely monitor market policy changes, such as adjustments to new energy vehicle subsidies, and quickly optimize product lines to stay competitive. Moreover, increasing R&D investment, especially in battery technology and intelligent connectivity, will help ensure that companies remain technologically advanced in the future market.

（7）Collaborate with Policymakers

Companies can work with local governments and relevant institutions to promote the construction of charging networks and optimize policies for new energy vehicles. For example, assisting in promoting green purchase incentives, such as reducing purchase taxes or providing exclusive parking spaces, can drive EV adoption. Additionally, collaborating with city management to optimize dedicated EV lanes or improve charging convenience will further enhance consumer acceptance of EVs, expanding market potential.

By implementing these recommendations, companies can not only establish a strong foothold in the current car market but also gain a strategic advantage in the future transition to new energy vehicles.