car_analysis

June 3, 2024

1 1. Problem Statement

- Dataset "Vehicle Sales and Market Trends" menyediakan informasi komprehensif tentang transaksi penjualan kendaraan. Data ini mencakup tahun pembuatan kendaraan, merek atau produsen kendaraan, model spesifik kendaraan, versi atau paket opsi tertentu dari model, jenis bodi kendaraan, jenis transmisi kendaraan, Nomor Identifikasi Kendaraan, negara bagian tempat kendaraan terdaftar, penilaian kondisi kendaraan, jarak tempuh kendaraan, warna eksterior kendaraan, warna interior kendaraan, entitas atau perusahaan yang menjual kendaraan, nilai Manheim Market Report, harga jual kendaraan, serta tanggal dan waktu penjualan kendaraan.
- Proyek ini dibuat untuk menganalisis data penjualan kendaraan guna memahami pola dan faktor-faktor yang mempengaruhi harga jual serta tren pasar. Dengan meningkatnya persaingan di pasar otomotif, memahami dinamika ini menjadi sangat penting untuk Mengidentifikasi tren penjualan dari tahun ke tahun,Menentukan faktor-faktor yang mempengaruhi harga jual kendaraan,Mengenali merek dan jenis bodi kendaraan yang paling populer dan paling mahal,Mengembangkan strategi pemasaran dan penjualan yang lebih efektif.
- Specific questions to be addressed include: Mengidentifikasi tren penjualan dari tahun ke tahun, Menentukan faktor-faktor yang mempengaruhi harga jual kendaraan, Mengenali merek dan jenis bodi kendaraan yang paling populer dan paling mahal, Mengembangkan strategi pemasaran dan penjualan yang lebih efektif.
- Dengan menjawab pertanyaan-pertanyaan di atas, analisis ini akan memberikan kontribusi signifikan pada strategi penjualan dan pemasaran kendaraan. Pertama, mengidentifikasi tren pasar akan membantu perusahaan memahami bagaimana penjualan kendaraan berubah dari tahun ke tahun, yang sangat penting untuk perencanaan jangka panjang dan penyesuaian strategi pemasaran agar sesuai dengan tren pasar. Kedua, menentukan faktor-faktor utama yang mempengaruhi harga jual kendaraan memungkinkan perusahaan untuk mengoptimalkan harga jual berdasarkan kondisi pasar dan karakteristik kendaraan, sehingga dapat meningkatkan profitabilitas. Ketiga, mengetahui 10 merek dan tipe bodi kendaraan yang memiliki harga jual tertinggi membantu perusahaan menyesuaikan inventaris dan strategi penjualan untuk memaksimalkan keuntungan. Terakhir, mengidentifikasi merek dan tipe bodi kendaraan yang paling populer membantu dalam pengelolaan stok dan penyesuaian penawaran produk untuk memenuhi permintaan pasar. Analisis ini akan memberikan wawasan yang mendalam untuk mendukung pengambilan keputusan yang lebih baik dalam strategi penjualan dan pemasaran kendaraan.

2 2. Assumptions

- Completeness of data:Assuming that the dataset does not have many missing or empty values for crucial columns such as year, brand, model, selling price, and sales date.
- Accuracy of data:Assuming that the provided data is accurate and consistent, especially for numeric variables such as selling price, odometer, and MMR value.
- Independence of observations:Assuming that each row of data (vehicle sales) is independent and not related to other observations in the dataset.
- No multicollinearity:Assuming that there is no strong correlation between predictor variables such as year, brand, model, body type, and vehicle condition.
- Distribution of numeric data:Assumptions about the distribution of numeric data such as selling price, odometer, and MMR value. For example, whether the data follows a normal distribution or other distributions.

3 3. Research Question

- Bagaimana pola evolusi tren penjualan kendaraan dari tahun ke tahun? Apakah ada lonjakan atau penurunan dramatis yang bisa diidentifikasi, serta faktor-faktor utama yang mungkin mempengaruhinya?
- Apa yang menjadi pemicu fluktuasi dalam harga jual kendaraan? Selain faktor-faktor umum seperti kondisi kendaraan dan model, apakah ada variabel-variabel tidak terduga yang secara signifikan memengaruhi harga jual?
- Bagaimana merek dan jenis bodi kendaraan tertentu mempengaruhi dinamika pasar dengan mengeksplorasi 10 merek mobil dan jenis bodi yang secara konsisten mempertahankan harga jual tertinggi?
- Dengan memperhatikan preferensi konsumen dan tren pasar, merek dan jenis bodi kendaraan apa yang paling diminati dan paling berhasil dalam distribusi, dan bagaimana hal ini mempengaruhi strategi pemasaran dan stok perusahaan?

4 4. Execution

Pada bagian ini, kita akan mengimpor dataset dan melakukan eksplorasi data, pembersihan data, dan analisis data.

[2]:
```
# Import necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
df = pd.read_csv('car_prices.csv')

# Display the first few rows of the dataset
df.head()
```

[2]:
```
   year   make                model        trim   body transmission  \
0  2015    Kia              Sorento          LX    SUV    automatic
1  2015    Kia              Sorento          LX    SUV    automatic
2  2014    BMW            3 Series  328i SULEV  Sedan    automatic
3  2015  Volvo                 S60          T5  Sedan    automatic
4  2014    BMW  6 Series Gran Coupe        650i  Sedan    automatic

                 vin state  condition  odometer  color interior  \
0  5xyktca69fg566472    ca        5.0   16639.0  white    black
1  5xyktca69fg561319    ca        5.0    9393.0  white    beige
2  wba3c1c51ek116351    ca       45.0    1331.0   gray    black
3  yv1612tb4f1310987    ca       41.0   14282.0  white    black
4  wba6b2c57ed129731    ca       43.0    2641.0   gray    black

                                  seller      mmr  sellingprice  \
0                kia motors america  inc  20500.0       21500.0
1                kia motors america  inc  20800.0       21500.0
2  financial services remarketing (lease)  31900.0       30000.0
3                volvo na rep/world omni  27500.0       27750.0
4  financial services remarketing (lease)  66000.0       67000.0

                                 saledate
0  Tue Dec 16 2014 12:30:00 GMT-0800 (PST)
1  Tue Dec 16 2014 12:30:00 GMT-0800 (PST)
2  Thu Jan 15 2015 04:30:00 GMT-0800 (PST)
3  Thu Jan 29 2015 04:30:00 GMT-0800 (PST)
4  Thu Dec 18 2014 12:30:00 GMT-0800 (PST)
```

[5]:
```
df.shape
```

[5]:
```
(558837, 16)
```

[6]:
```
df.columns
```

[6]:
```
Index(['year', 'make', 'model', 'trim', 'body', 'transmission', 'vin', 'state',
       'condition', 'odometer', 'color', 'interior', 'seller', 'mmr',
       'sellingprice', 'saledate'],
      dtype='object')
```

[7]:
```
df.duplicated().any()
```

[7]:
```
False
```

Handling Missing Values In Categorical Columns

When dealing with a categorical column like 'make' with a significant number of null values, filling them requires careful consideration. Since 'make' represents the brand or manufacturer of the vehicle, blindly filling null values with the most common value may introduce bias.

1. Fill with a Placeholder Category

One approach is to fill the missing values with a placeholder category, such as 'Unknown' or 'Other'. This preserves the fact that the value was missing and does not introduce bias by favoring the most frequent category.

2. Use Mode, Median, Mean (most frequent category)

Another approach is to fill the missing values with the mode, median, or mean of the column. However, this method may introduce bias if the most frequent category dominates the dataset.

3. Remove Null Values

If the missing values are too numerous or cannot be imputed accurately, another option is to remove the rows with missing values. This ensures that the analysis is based only on complete data but may result in a loss of information.

We are going to use all of the above imputation techniques to handle missing values in our categorical columns.

[9]:
```
df.isnull().sum()
```

[9]:
```
year               0
make           10301
model          10399
trim           10651
body           13195
transmission   65352
vin                4
state              0
condition      11820
odometer          94
color            749
interior         749
seller             0
mmr               38
sellingprice      12
saledate          12
dtype: int64
```

[10]:
```
df['make'] = df['make'].fillna('Other')
df['model'] = df['model'].fillna('Other')
df['trim'] = df['trim'].fillna('Other')
df['color'] = df['color'].fillna('Other')
```

[11]:
```
df['body'] = df['body'].fillna(df['body'].mode()[0])
df['transmission'] = df['transmission'].fillna(df['transmission'].mode()[0])
df['interior'] = df['interior'].fillna(df['interior'].mode()[0])
```

[12]:
```
df.dropna(subset=['vin'], inplace=True)
df.dropna(subset=['saledate'], inplace=True)
```

[13]:
```
df.isnull().sum()
```

[13]:
```
year                0
make                0
model               0
trim                0
body                0
transmission        0
vin                 0
state               0
condition       11816
odometer           94
color               0
interior            0
seller              0
mmr                22
sellingprice        0
saledate            0
dtype: int64
```

Handling Missing Values in Numerical Columns

[14]:
```
df['condition'].fillna(df['condition'].median(), inplace=True)
df['odometer'].fillna(df['odometer'].mean(), inplace=True)
df['mmr'].fillna(df['mmr'].mean(), inplace=True)
```

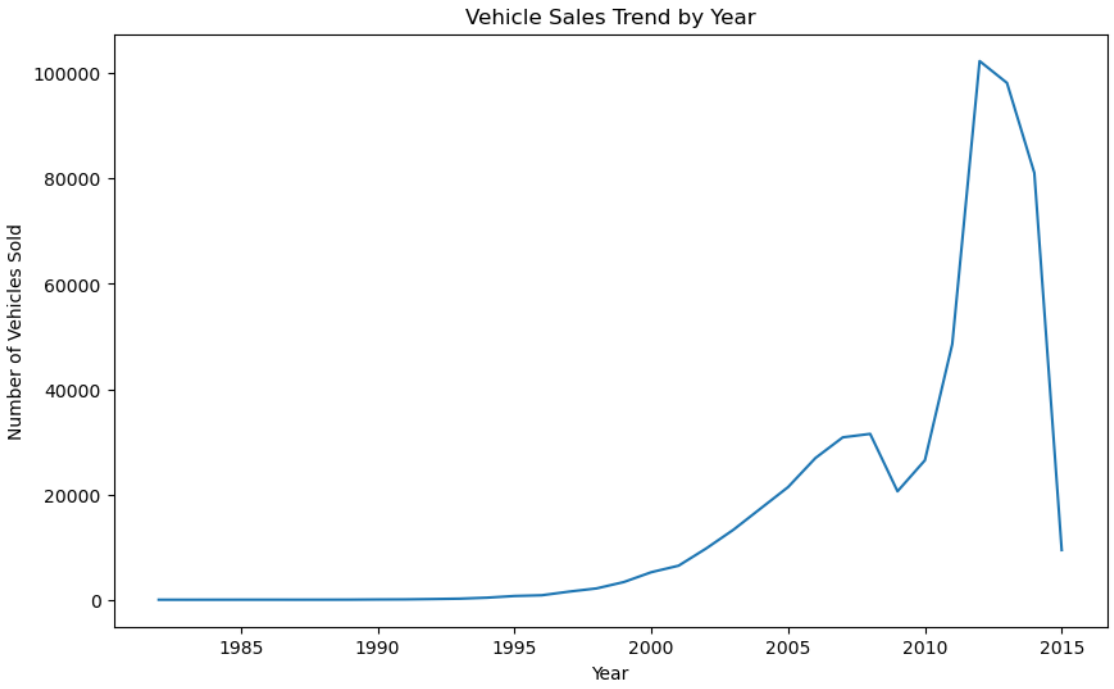
[15]:
```
df.isnull().sum()
```

[15]:
```
year            0
make            0
model           0
trim            0
body            0
transmission    0
vin             0
state           0
condition       0
odometer        0
color           0
interior        0
seller          0
mmr             0
sellingprice    0
saledate        0
dtype: int64
```

4.0.1 Research Answer

1. Bagaimana pola evolusi tren penjualan kendaraan dari tahun ke tahun? Apakah ada lonjakan atau penurunan dramatis yang bisa diidentifikasi, serta faktor-faktor utama yang mungkin mempengaruhinya?

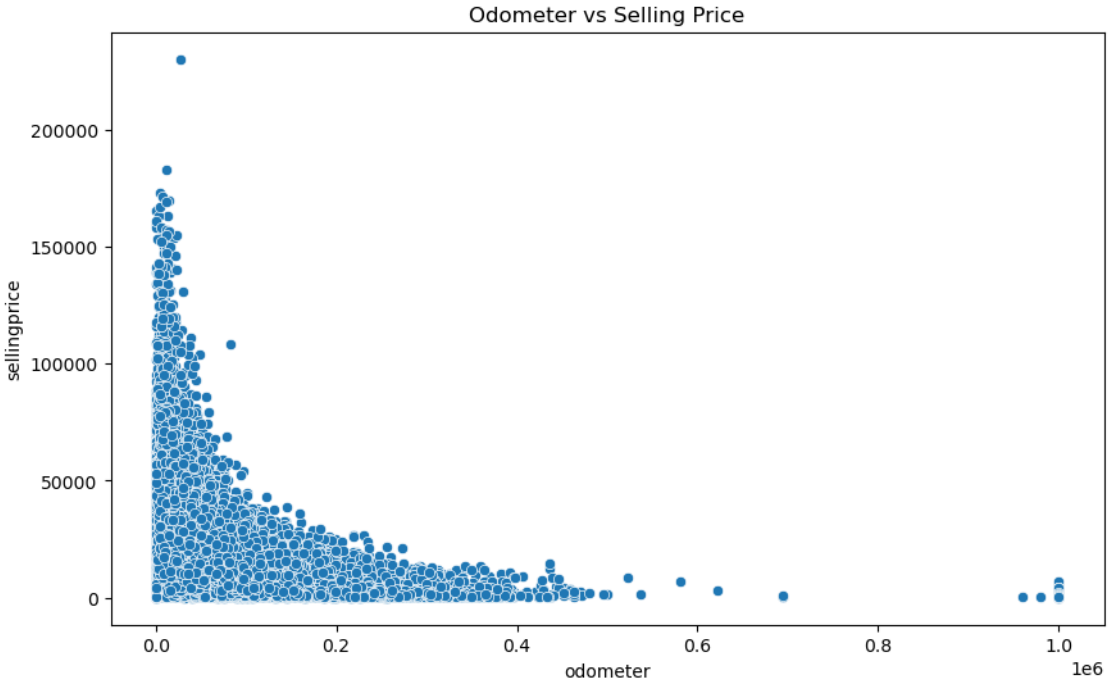
[16]:
```
sales_per_year = df.groupby('year').size()

plt.figure(figsize=(10, 6))
sales_per_year.plot(kind='line')
plt.title('Vehicle Sales Trend by Year')
plt.xlabel('Year')
plt.ylabel('Number of Vehicles Sold')
plt.show()
```

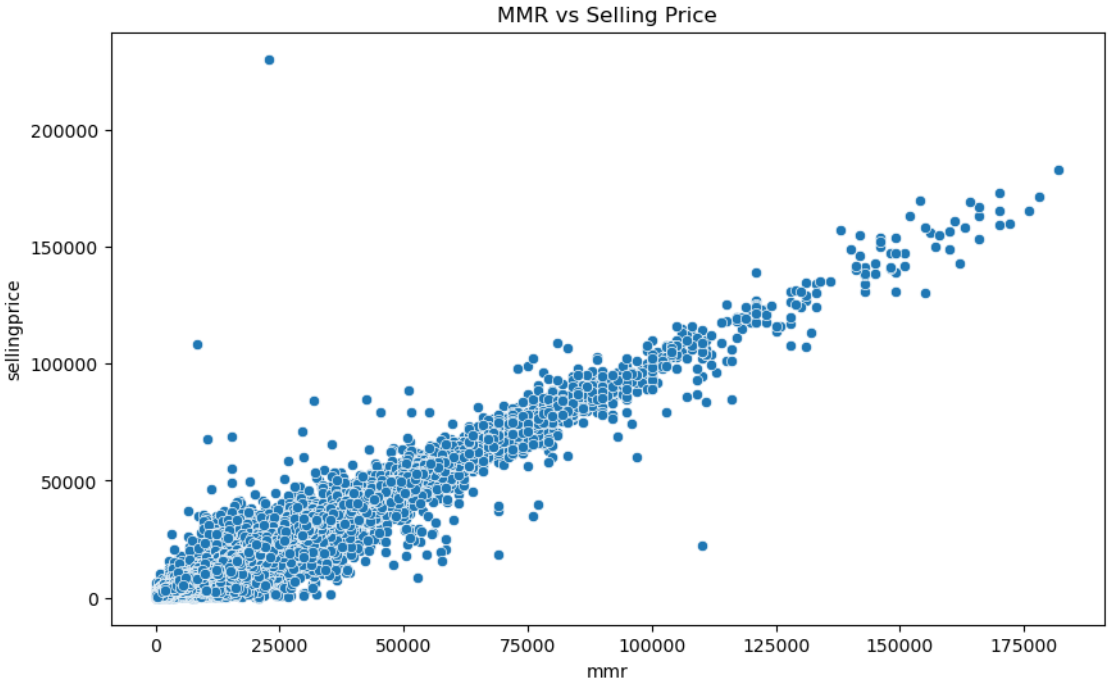


2. Apa yang menjadi pemicu fluktuasi dalam harga jual kendaraan? Selain faktor-faktor umum seperti kondisi kendaraan dan model, apakah ada variabel-variabel tidak terduga yang secara signifikan memengaruhi harga jual?

[52]:
```
# Condition vs Selling Price
plt.figure(figsize=(10, 6))
sns.scatterplot(x='condition', y='sellingprice', data=df)
plt.title('Condition vs Selling Price')
plt.show()

# Odometer vs Selling Price
```

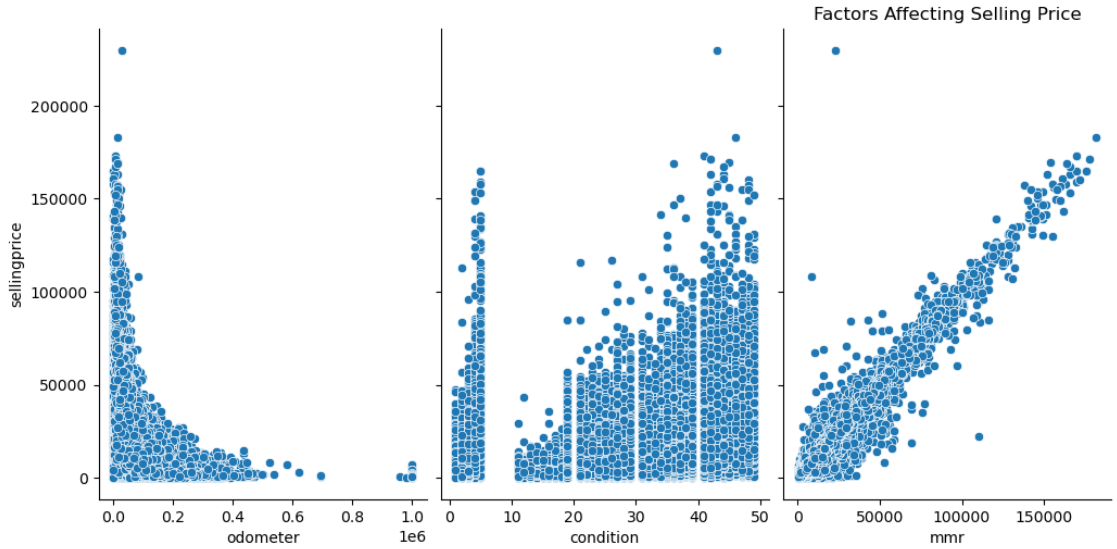
```
plt.figure(figsize=(10, 6))
sns.scatterplot(x='odometer', y='sellingprice', data=df)
plt.title('Odometer vs Selling Price')
plt.show()

# mmr vs selling price
plt.figure(figsize=(10, 6))
sns.scatterplot(x='mmr', y='sellingprice', data=df)
plt.title('MMR vs Selling Price')
plt.show()
```





[20]:
```
# Pairplot to see relationships
```

```
sns.pairplot(df, x_vars=['odometer', 'condition', 'mmr'],␣
 ↪y_vars='sellingprice', height=5, aspect=0.7, kind='scatter')
plt.title('Factors Affecting Selling Price')
plt.show()
```



3. Bagaimana merek dan jenis bodi kendaraan tertentu mempengaruhi dinamika pasar dengan mengeksplorasi 10 merek mobil dan jenis bodi yang secara konsisten mempertahankan harga jual tertinggi?

[36]:
```
# Hitung harga rata-rata untuk setiap merek
average_price_per_make = df.groupby('make')['sellingprice'].mean()

# Urutkan merek berdasarkan harga rata-rata secara menurun
sorted_average_price = average_price_per_make.sort_values(ascending=False)

# Ambil 10 merek dengan harga rata-rata tertinggi
top_10_makes = sorted_average_price.head(10)

print(top_10_makes)
```

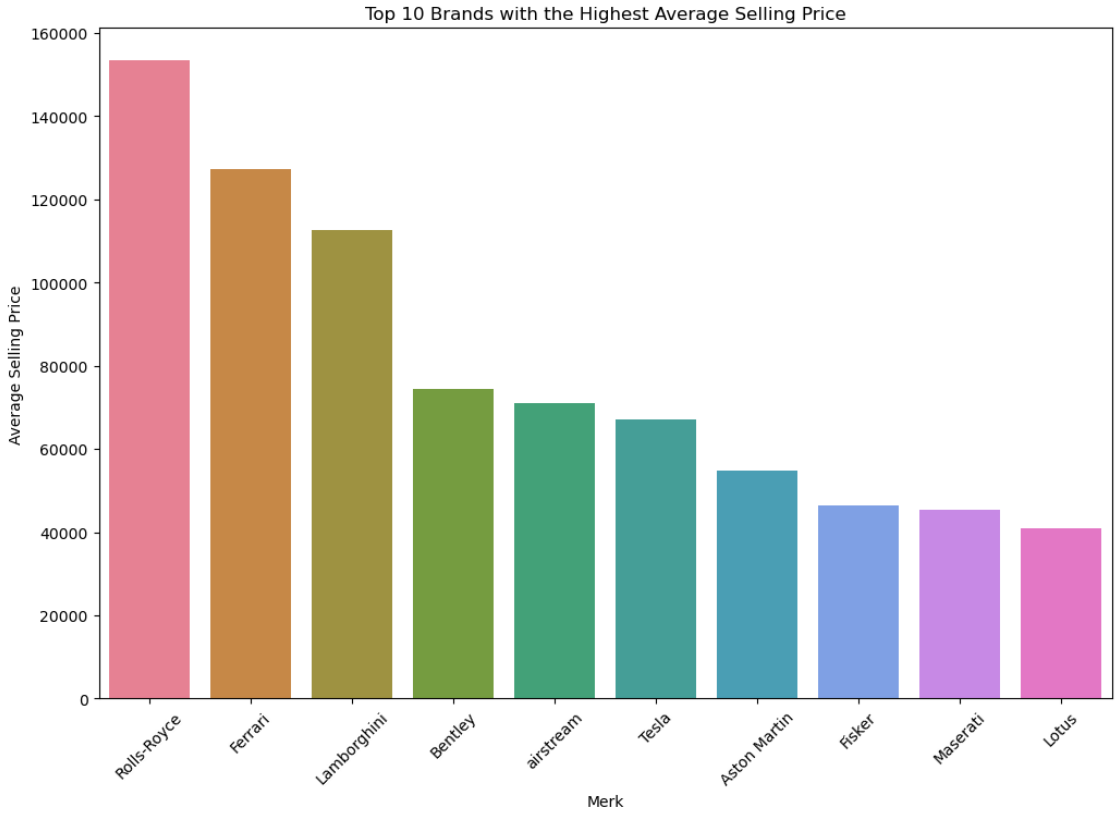
```
make
Rolls-Royce      153488.235294
Ferrari          127210.526316
Lamborghini      112625.000000
Bentley           74367.672414
airstream         71000.000000
Tesla             67054.347826
Aston Martin      54812.000000
Fisker            46461.111111
```

```
Maserati          45320.300752
Lotus             40800.000000
Name: sellingprice, dtype: float64
```

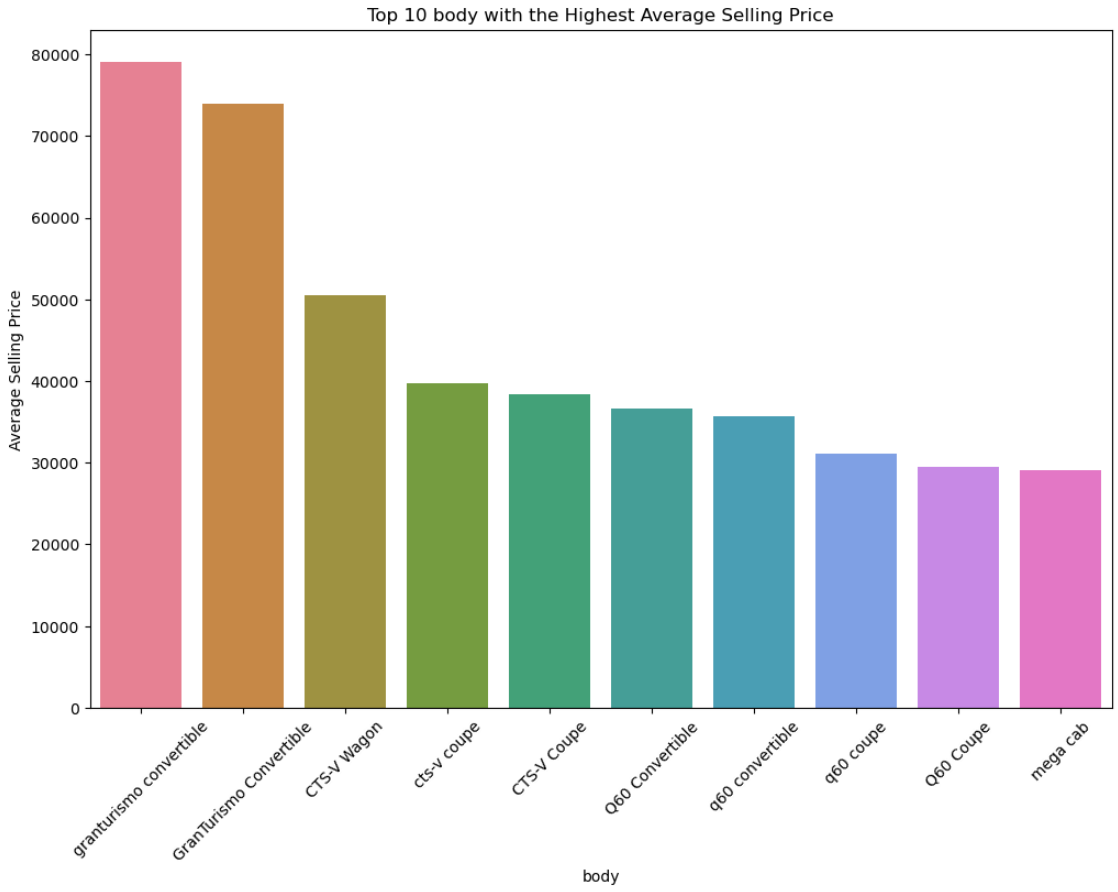
[49]:
```python
# Hitung harga rata-rata untuk setiap merek
average_price_per_make = df.groupby('body')['sellingprice'].mean()

# Urutkan merek berdasarkan harga rata-rata secara menurun
sorted_average_price = average_price_per_make.sort_values(ascending=False)

# Ambil 10 merek dengan harga rata-rata tertinggi
top_10_body = sorted_average_price.head(10)

print(top_10_body)
```

```
body
granturismo convertible    79041.666667
GranTurismo Convertible    74000.000000
CTS-V Wagon                50500.000000
cts-v coupe                39707.142857
CTS-V Coupe                38425.750000
Q60 Convertible            36667.105263
q60 convertible            35725.000000
q60 coupe                  31112.500000
Q60 Coupe                  29479.687500
mega cab                   29055.769231
Name: sellingprice, dtype: float64
```

[40]:
```python
plt.figure(figsize=(12, 8))
sns.barplot(x=top_10_makes.index, y=top_10_makes.values, palette='husl')
plt.title('Top 10 Brands with the Highest Average Selling Price')
plt.xlabel('Merk')
plt.ylabel('Average Selling Price')
plt.xticks(rotation=45)
plt.show()
```



[50]:
```
plt.figure(figsize=(12, 8))
sns.barplot(x=top_10_body.index, y=top_10_body.values, palette='husl')
plt.title('Top 10 body with the Highest Average Selling Price')
plt.xlabel('body')
plt.ylabel('Average Selling Price')
plt.xticks(rotation=45)
plt.show()
```



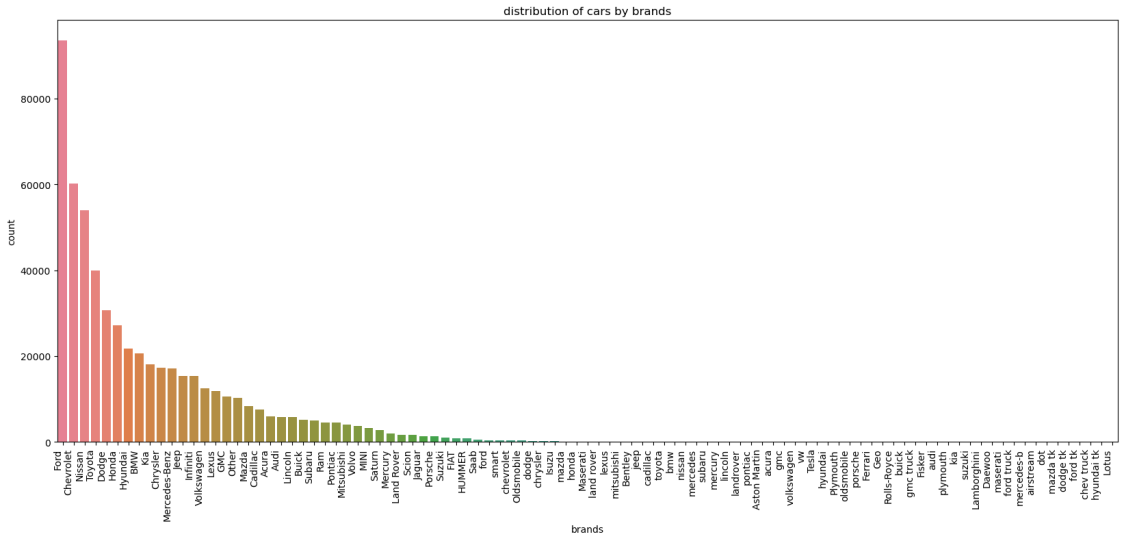
4. Dengan memperhatikan preferensi konsumen dan tren pasar, merek dan jenis bodi kendaraan apa yang paling diminati dan paling berhasil dalam distribusi, dan bagaimana hal ini mempengaruhi strategi pemasaran dan stok perusahaan?

[58]:
```
# Menghitung distribusi merek mobil
brand_distribution = df['make'].value_counts()

# Menghitung distribusi jenis body mobil
body_distribution = df['body'].value_counts()

print("merek yang paling sering muncul",brand_distribution)
print("tipe yang paling sering muncul",body_distribution)
```

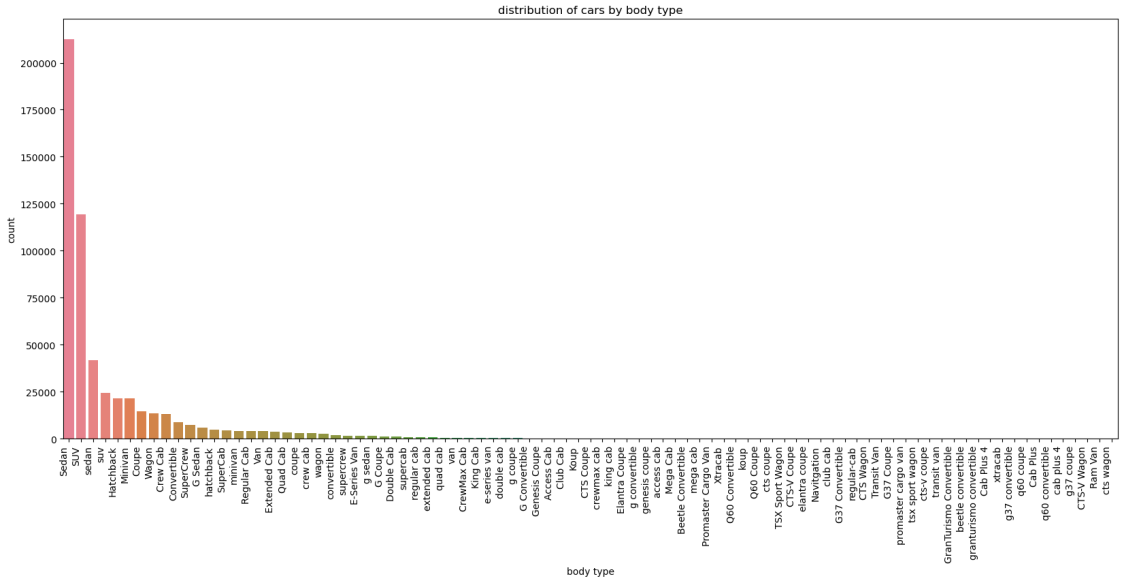
```
merek yang paling sering muncul make
Ford          93553
Chevrolet     60197
Nissan        53946
Toyota        39871
Dodge         30708
              ...
```

```
dodge tk           1
ford tk            1
chev truck         1
hyundai tk         1
Lotus              1
Name: count, Length: 97, dtype: int64
tipe yang paling sering muncul body
Sedan          212624
SUV            119292
sedan           41903
suv             24552
Hatchback       21380
                  …
cab plus 4            1
g37 coupe             1
CTS-V Wagon           1
Ram Van               1
cts wagon             1
Name: count, Length: 87, dtype: int64
```

[67]:
```
plt.figure(figsize=(20, 8))
sns.barplot(x=brand_distribution.index, y=brand_distribution.values,␣
 ↪palette='husl')
plt.title('distribution of cars by brands')
plt.xlabel('brands')
plt.ylabel('count')
plt.xticks(rotation=90, ha='right')
plt.show()
```



[68]:
```
plt.figure(figsize=(20, 8))
sns.barplot(x=body_distribution.index, y=body_distribution.values,␣
 ↪palette='husl')
plt.title('distribution of cars by body type')
plt.xlabel('body type')
plt.ylabel('count')
plt.xticks(rotation=90, ha='right')
plt.show()
```



[ ]:
```

```