Problem Statement

The question we want to answer is how to predict the resale price and find the main factors impacting the resale price of used cars. We consider brands, mileage, number of owners, classes, color, fuel, damage, accidents and some other features as major factors and make rational attempts to build trustworthy models to predict the resale price.
Our project contributes to the pricing of the second hand cars market in three aspects. First, we implement intensive data-driven approaches to generate business insights. The online data of used cars has a large and highly disorganized dataset, containing both structured and unstructured data, which makes it difficult to collect and standardize the data. Our project deals with this problem by scraping the biggest online used car website, carmax.com, to ensure we gather the most comprehensive and up-to-date data. By automating the data collection process, we can systematically extract both structured and unstructured data and eliminate unwanted data. EDA operations are then performed on our dataset to lay the foundation for subsequent stages. Second, we aim to develop reliable and transferable statistical models that can accurately predict the resale price of used cars. We want to provide a tool that helps both buyer and sellers to make reasonable decisions when evaluating the used cars' value and ensure a transparent and fast trade. Third, we hope to make data visualization with dashboards and automatically generated reports that are accessible for users in user-friendly format. These will provide real-time updates about the market that are suitable for individual needs.
Our contribution is crucial because it aims to provide transparency, efficiency and reliability to the used car market. We provide buyers and sellers with authentic data as well as prediction models that help them make rational decisions. We aim to develop a platform for users to assess used car's market value in a way to avoid overpricing or underpricing.

Ask Questions

1. **How does mileage of a car impact its resale value across different brands? (Te Shi)**
   Mileage of a car is commonly understood as an important deciding indicator of a car's conditions and durability and customers value this metric very much. Therefore, understanding how mileage affects resale value for different brands is critical for making the final prediction model.
2. **Does the popularity of different car classes affect a car's resale value? (Te Shi)**
   Car class is a critical factor which influences customer's demands and it can also reflect a general portrait of each potential buyer.
3. **What effect do different colors have on the price of a used car? (Jiabao Yao)**
   The color is the value of the car itself. Choosing a popular color when buying a car will make it easier to resell, and even get a better price. Color will be an important indicator when the model reaches a high accuracy.
4. **How does fuel of a car impact its resale value across different brands? (Jiabao Yao)**

Fuel is an important indicator of a car's value, which determines the durability, maintenance and use costs of the car. Understanding the relationship between fuel standards and car prices is very important for building a predictive model.

5. **Is the brand recognition of different cars an important factor affecting the resale price?(Chao Wu)**
   Studying brand recognition can reveal the influence and competitiveness of certain brands thus supporting the prediction of resale value. Understanding this question is important because it reflects customers' decisions and loyalness, which are both vital for dealers and buyers to make decisions.
6. **Does the number of owners of used cars affect the resale price?(Chao Wu)**
   Typically, cars with fewer previous owners are perceived as better maintained and more reliable, which can lead to higher resale prices and vice versa. The question is significant because it guides both buyers and dealers in the market on pricing. Understanding it can help buyers to make informed decisions.
7. **How do the accidents or damage records of the used cars affect the resale price? (Shijie Zhou)**
   Whether the used car has been through an accident/accidents and how severe the damage on this car is a crucial element to determine the price of the used car as the severity of the damage from the accident can affect the safety and reliability of the vehicle which is an important factor considered by customers.
8. **For used cars with different makes, will the accident record affect the used cars' price differently? (Shijie Zhou)**
   The price of the used car might be more sensitive regarding the accident record for some high-end auto makes.

Data Retrieval

- **Data Source**
  Our dataset of used cars was obtained by scraping car information from carmax.com, a widely-used auto-trader platform with the scrape script we wrote from scratch.
  Reference: https://www.carmax.com/

- **Amount of Data**
  We scraped 10367 pieces of car information from 31 different brands.

  The scrapping script was designed to collect approximately 400 entries per brand. However, due to variations in the availability of cars for each brand on the platform and missing data on some pages (e.g., unavailable prices), certain entries were dropped during the process.

- **Content of Dataset**
  Each row/record represents a car. Each car has the following 29 features before preliminary data cleaning.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10367 entries, 0 to 10366
Data columns (total 29 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   year                                  10367 non-null  int64
 1   make                                  10367 non-null  object
 2   model                                 10367 non-null  object
 3   price                                 10367 non-null  int64
 4   mileage                               10367 non-null  int64
 5   Miles per gallon                      9677 non-null   object
 6   Transmission                          9878 non-null   object
 7   Color                                 10367 non-null  object
 8   owner                                 10367 non-null  object
 9   frame_damage                          10367 non-null  object
 10  Odometer_problem                      10367 non-null  object
 11  VIN                                   10367 non-null  object
 12  class                                 10367 non-null  object
 13  State Title Brand                     10367 non-null  object
 14  Auction Brand / Issues                10367 non-null  object
 15  Accident / Damage                     10367 non-null  object
 16  Open Recall Check                     10367 non-null  object
 17  Insurance Loss / Transfer             10367 non-null  object
 18  Odometer Check                        10367 non-null  object
 19  Certified Pre-Owned                   10367 non-null  object
 20  Service / Repair                      10367 non-null  object
 21  cylinders                             9348 non-null   object
 22  fuel                                  9348 non-null   object
 23  Drive type                            9307 non-null   object
 24  Miles per gallon equivalent (MPGe)    462 non-null    object
 25  Range (when new)                      479 non-null    object
 26  Time to fully charge battery (240V)   450 non-null    object
 27  Motor                                 308 non-null    object
 28  Bed Length                            784 non-null    object
```

After preliminary data cleaning, several single-value features were dropped, changing the total number of features to 25. (see Data Cleaning Section)

- **Process**

  The ***Car_Info.py*** uses Playwright as a headless browser and Selectolax to parse HTML content. Playwright browser was used to bypass 403 errors and handle pagination, while Selectolax was used to extract targeted information from the webpage. The process transferred unstructured data obtained from HTML into final csv files using Pandas library.

  Specifically, the script controls the Playwright browser to navigate each car brand's search page on Carmax. Based on the number of pages to scrape,it collects links to each car's display page. The script then loop through these links, open each of them and extract relevant car information. Then the script navigates to each car's check report page to extract information about the car's condition for further analysis. All the extracted car data is stored in a list of dictionaries, which is then converted into a Pandas DataFrame and saved as CSV files.

  During the process, some data cleaning strategies were applied, such as filtering out cars with unavailable prices, and converting string representation of mileage into float number. (e.g., changing "31K" to 31,000).

  We designed our scrape script to save separate csv files for each car brand. Therefore, after scrapping all the car records, ***Data_Combine.py*** is used to merge all the csv files into a single dataset ***scrapping_data.csv***, which will be manipulated in the next steps.

Data Cleaning/Processing

- **Some data processing/cleaning operations were done in the data retrieval process.**
    1. **drop car information with price unavailable**

```
# Extract price
price_literal = tree.css_first("span#default-price-display").text()
if "unavailable" not in price_literal:
    price = int(price_literal.replace('$', '').replace(',', '').replace('*', ''))
    car["price"] = price
else:
    car["price"]="Price unavailable"
```

    2. **converting string representation of mileage to numerical**

```
# Extract mileage
mileage_literal = tree.css_first("span.car-header-mileage").text()
mileage = int(mileage_literal.split(" ")[0].replace('K', '')) * 1000
car["mileage"] = mileage
```

- ***scrapping_data.csv* was underwent a preliminary cleaning process done by *preliminary_data_processing_cleaning.ipynb,* the result *carinfo_after_pre_clean.csv* is our final dataset**

    3. **Change feature owner and cylinder's values from string to integer type**
    The owner and cylinder features were originally displayed as numbers followed by a string such as "1 Owner",or "4-cyl" which is unnecessary and creates challenges for further manipulation. We simplified it to an integer format for easier handling.

owner	frame_damage	...	Certified Pre-Owned	Service / Repair	cylinders
2.0	No flood or frame damage	...	No CPO Info Available	No Issue	NaN
1.0	No flood or frame damage	...	No CPO Info Available	No Issue	4.0
1.0	No flood or frame damage	...	No CPO Info Available	No Issue	4.0
3.0	No flood or frame damage	...	No CPO Info Available	No Issue	4.0
1.0	No flood or frame damage	...	No CPO Info Available	No Issue	4.0

    4. **Delete erroneous value in feature "owner"**
    Upon inspection, we found that some records in the "owner" feature records 0 owners but they are all old cars with significant mileage. We identified these as erroneous records and removed them from our dataset. 21 pieces of entries were deleted.

```
print("number of records before removal:", df.shape[0])
df=df[df["owner"]!=0]
print("number of records after removal: ", df.shape[0])
```

```
number of records before removal: 10367
number of records after removal:  10346
```

5. **Separate feature "color" to "exterior color" and "interior color"**
   In the original scraped data, the values in the "color" feature combine two color information, such as "red/black" which represent a car's exterior and interior color respectively. We believe these may have separate impacts on a car's resale value so we divided the "color" feature into "exterior color" and "interior color" for convenience of future analysis.

```
df[['exterior_color', 'interior_color']] = df['Color'].str.split('/', expand=True)
df.drop(columns=['Color'],inplace=True) # remove the color column
```

6. **Delete single value features**
   Features containing only 1 single value across all records are considered as non-informative because they provide no insights and are not helpful in the learning process either. Therefore we identified and removed them from the dataset.
   Single value features we found in our data set are 'frame_damage', 'Odometer_problem', 'State Title Brand', 'Insurance Loss / Transfer', 'Service / Repair' and 'Motor'.
   After inspection, we found that the Motor feature contains either "electric" or "N/A," where "N/A" indicates a non-electric car. We decided to keep this feature as it allows us to easily distinguish between electric and gas cars. We removed the rest of them.

```
single_value_features=[col for col in df.columns if df[col].nunique()==1] # identify single-value features
print("single value features: ", single_value_features)
```

```
single value features:  ['frame_damage', 'Odometer_problem', 'State Title Brand', 'Insurance Loss / Transfer', 'Service / Repair', 'Motor']
```

After inspection, we found that the Motor feature contains either "electric" or "N/A," where "N/A" indicates a non-electric car. We decided to keep this feature as it allows us to easily distinguish between electric and gas cars. We removed rest of them

```
single_value_features=single_value_features[:-1] # exclude motor feature
```

```
print("number of features before removal: ", df.shape[1])
df.drop(columns=single_value_features,inplace=True)
print("number of features after removal: ", df.shape[1])
```

```
number of features before removal:  30
number of features after removal:  25
```

7. **Fill N/A value in motor feature**
   As mentioned above, the N/A value in the motor indicates it is a non-electric car, thus we fill records with none value in motor feature to "non-electric".

```
df['Motor']=df['Motor'].fillna('non-electric')
```

- **Some Data operations were completed during our EDA process**

8. **Drop rows with price equals to 0 and rows with NaN values in mileage, price or make**

```
df=df.dropna(subset=['mileage','price','make','class']) # drop rows with NaN values in mileage, price or make
df = df[df['price']!=0] # drop rows with price equals to 0
```

9. **Divide used cars into groups with different damage/accident levels.**

```
clean_car_data = df_dmg.loc[df_dmg['Accident / Damage'].isin(['No Issue'])]
minor_data = df_dmg.loc[df_dmg['Accident / Damage'].isin(['Minor', 'Very Minor'])]
moderate_data = df_dmg.loc[df_dmg['Accident / Damage'].isin(['Moderate', 'Minor-Moderate','Moderate-Severe'])]
damage_reported_data = df_dmg.loc[df_dmg['Accident / Damage'].isin(['Damage Reported'])]
```

10. **Drop rows whose color is NaN.**

```
df_cleaned_1 = df.dropna(subset = ['exterior_color', 'interior_color'])
```

11. **Fill “Electric” for the “fuel” column if “Motor” column is “Electric” and drop rows whose fuel is NaN.**

```
df.loc[(df['fuel'].isna()) & (df['Motor'] == 'Electric'), 'fuel'] = 'Electric'
df_cleaned = df.dropna(subset = ['fuel'])
```

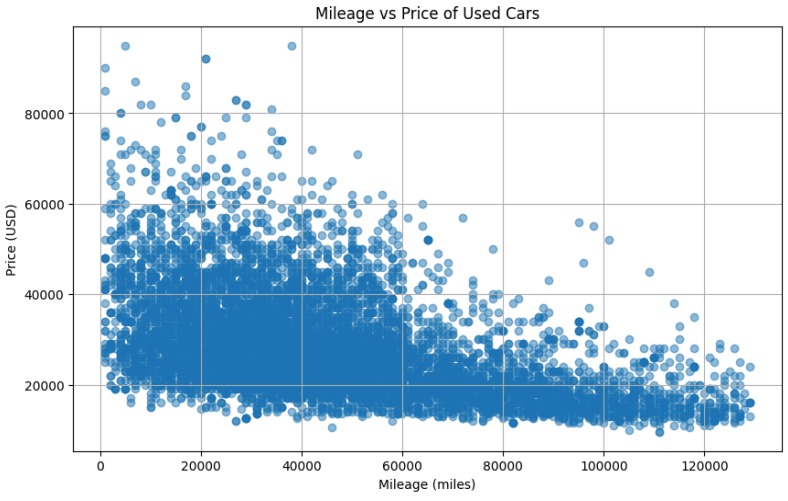
12. **Remove cars with duplicate VIN and empty color**

```
df = pd.DataFrame(clean_data)
df_cleaned_1 = df.dropna(subset = ['exterior_color', 'interior_color'])
df_cleaned = df_cleaned_1[df_cleaned_1['price'] != 0]
df_unique = df_cleaned.drop_duplicates(subset = 'VIN', keep = False)
```

Exploratory Data Analysis

Question 1: How does mileage of a car impact its resale value across different brands? (Te Shi)

Hypothesis 1: Generally higher mileage will correlate with a lower resale car value across all car brands



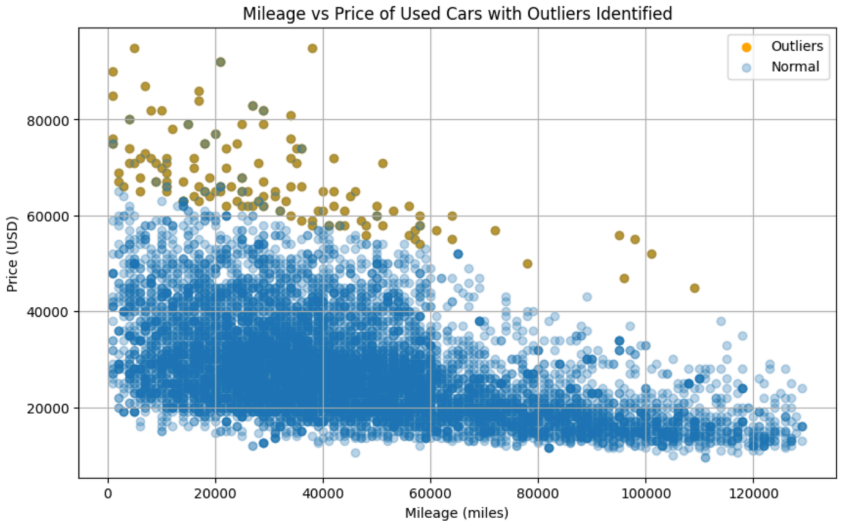
The graph shows that there is a clear negative correlation between mileage and price across all car information we collected, regardless of brands.

```
# calcualte general correlation value between resale price and mileage
general_corr=df['mileage'].corr(df['price'])
print(general_corr)
```

```
-0.49604939623020256
```

The General Pearson correlation -0.49 indicates that the mileage has a moderate to fairly strong impact on a car's resale value.

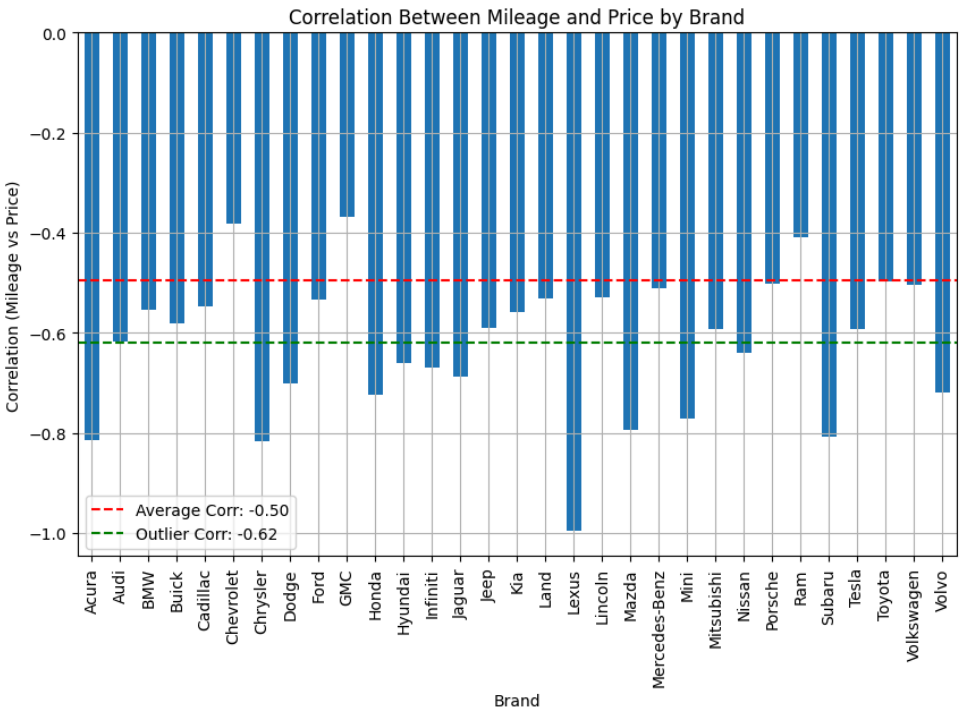
From the graph, although as expected, there is a general trend that with the increase of mileage, the price decreases, some outliers can be identified. Thus linear-regression is used to find cars with abnormal relationship between price and mileage.



The above graph shows the relationship between mileage and resale price with outliers identified. It demonstrates that the outliers generally have a higher resale value compared to other cars with similar mileage. Despite this, the graph still demonstrates a general negative correlation between mileage and resale price. The Pearson correlation of outliers is -0.6201513912533334

In conclusion, the above outputs proved the hypothesis.

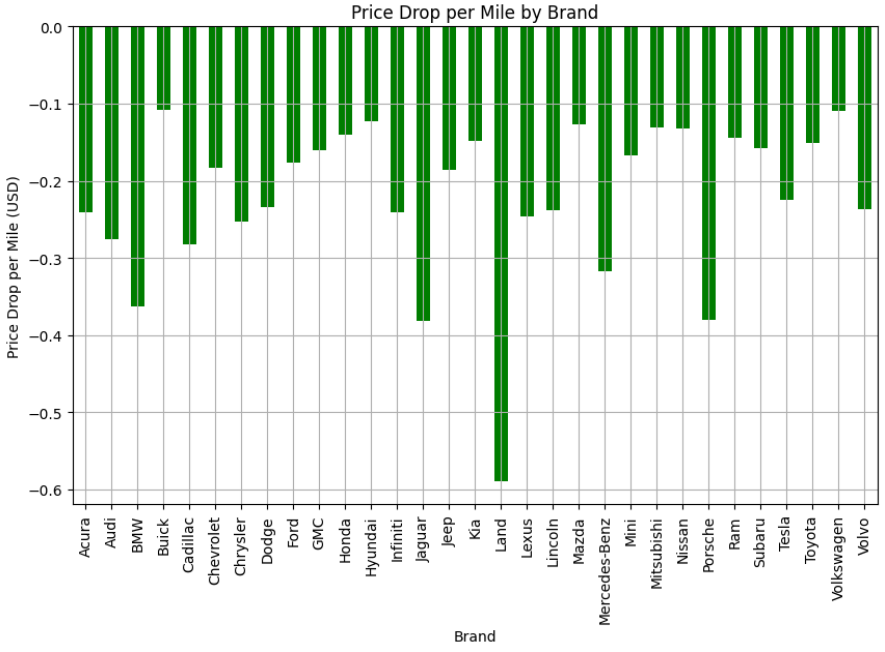
Hypothesis 2: Although there is a general negative correlation between mileage and resale value, different car brands may have significantly different correlation values.



The above graph shows correlations between mileage and resale value for each car brand. The result proves our hypothesis that different cars have different correlation values.

Among them 4 brands' correlation values exceed -0.8, which are Acura, Chrysler, Lexus and Subaru, indicating that mileage plays a very crucial role in determining their resale values.

Except Subaru, the other 3 makes are commonly considered as luxury car manufacturers, suggesting that luxury cars experience more damage when mileage increases. However, further analysis considering more features need to be conducted to validate this hypothesis.



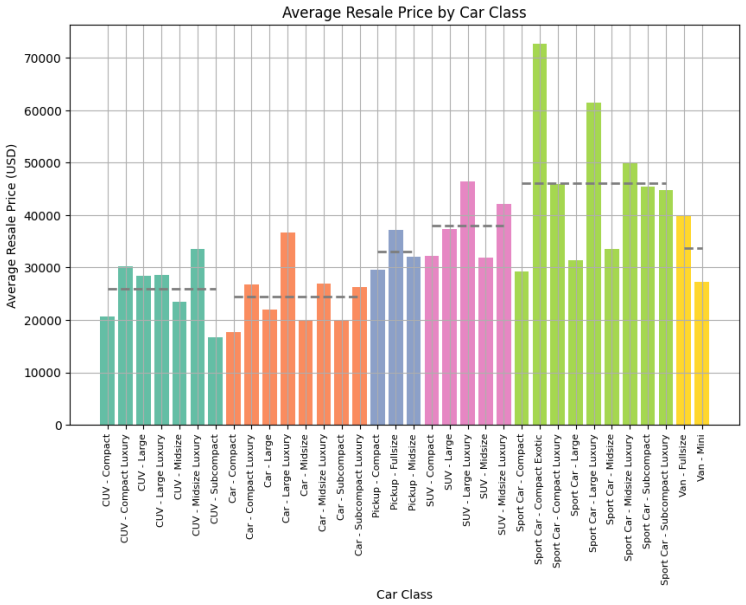
In addition, as shown above, price drops per mile for each brand were calculated. The weight obtained by linear regression model is used here as the measurement of price drop per mile.
Compared to correlation values, price drop per mile provides an actual overview about how mileage affects each car's resale values. Both metrics could be useful in the next modeling stage for estimating a car's resale value.

Question 2: Does the popularity of different car classes affect a car's resale value? (Te Shi)

Hypothesis 1: Car size is an important factor for customers when purchasing a car, and therefore, it significantly influences the car's resale value.

Firstly, average car prices by class is calculated for the 34 different car class

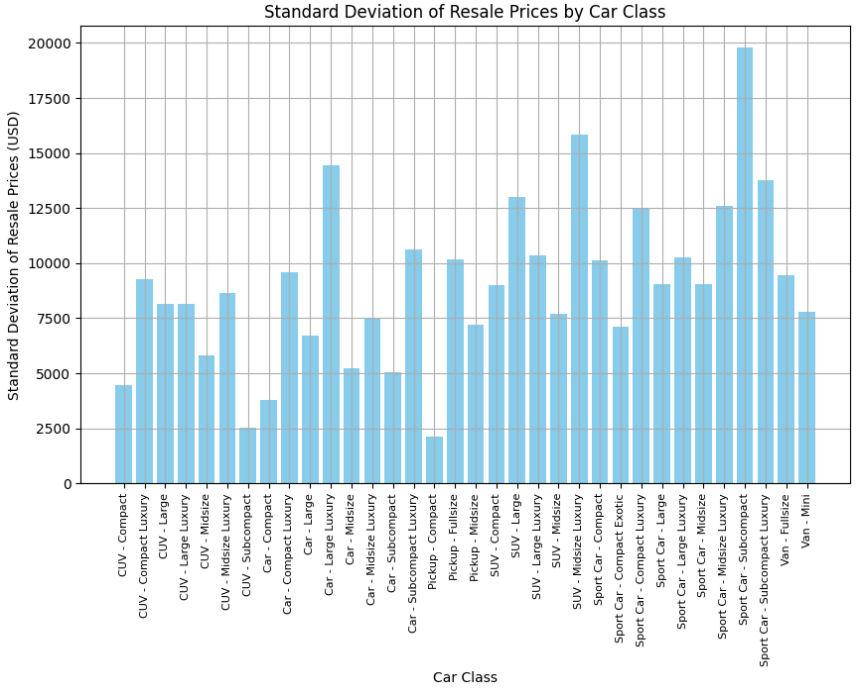
	class	price
0	CUV - Compact	20731.541347
1	CUV - Compact Luxury	30222.759223
2	CUV - Large	28419.798077
3	CUV - Large Luxury	28611.207547
4	CUV - Midsize	23475.808405
5	CUV - Midsize Luxury	33605.538803
6	CUV - Subcompact	16664.666667
7	Car - Compact	17737.345283
8	Car - Compact Luxury	26825.413994
9	Car - Large	22009.075862
10	Car - Large Luxury	36637.344262
11	Car - Midsize	19917.294268
12	Car - Midsize Luxury	26853.599455
13	Car - Subcompact	20019.160000
14	Car - Subcompact Luxury	26257.740260
15	Pickup - Compact	29553.555556
16	Pickup - Fullsize	37254.451613
17	Pickup - Midsize	32094.590909
18	SUV - Compact	32240.424242
19	SUV - Large	37333.365854
20	SUV - Large Luxury	46406.045977
21	SUV - Midsize	31919.980769
22	SUV - Midsize Luxury	42209.604096
23	Sport Car - Compact	29179.818182
24	Sport Car - Compact Exotic	72664.666667
25	Sport Car - Compact Luxury	45998.000000
26	Sport Car - Large	31474.190476
27	Sport Car - Large Luxury	61442.444444
28	Sport Car - Midsize	33464.666667
29	Sport Car - Midsize Luxury	50133.135135
30	Sport Car - Subcompact	45426.571429
31	Sport Car - Subcompact Luxury	44804.451613
32	Van - Fullsize	39998.000000
33	Van - Fullsize Luxury	38998.000000
34	Van - Mini	27309.953353

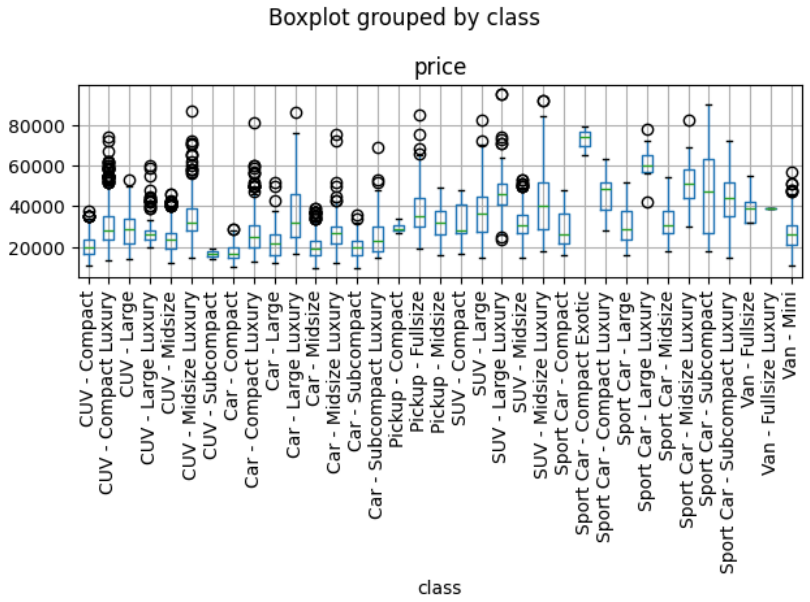


Then the above graph was plotted to represent the average resale price for each car class. The graph grouped car's with the same car class by color and the gray dashed line is the average price for each car class.

The graph indicates that SUV and Sport cars have higher resale values, while CUV and regular cars have relatively lower resale values. Also, among the same class, the larger cars and luxury versions have a understandably higher resale value. The result therefore indicates car class and size are important contributors for determining a car's resale value. In the further modeling stages, this feature can be divided into car class and car size, encoded, and used as input for the model.

Hypothesis 2: Car class may exhibit differences in the standard deviation of resale prices across different car classes



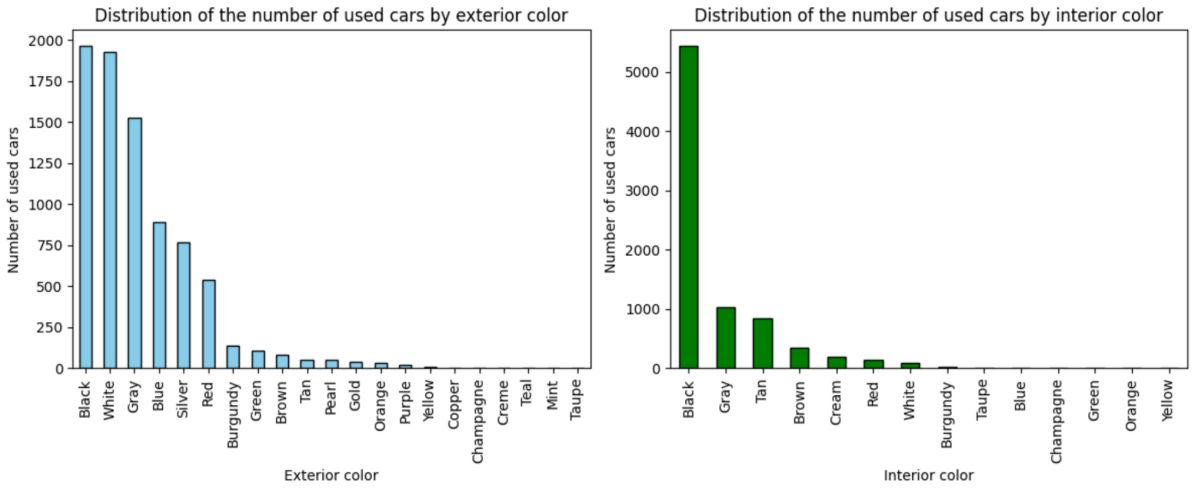


Firstly the standard deviation for each car class was calculated and visualized. Next the box plot was then generated to provide more information related to car class.

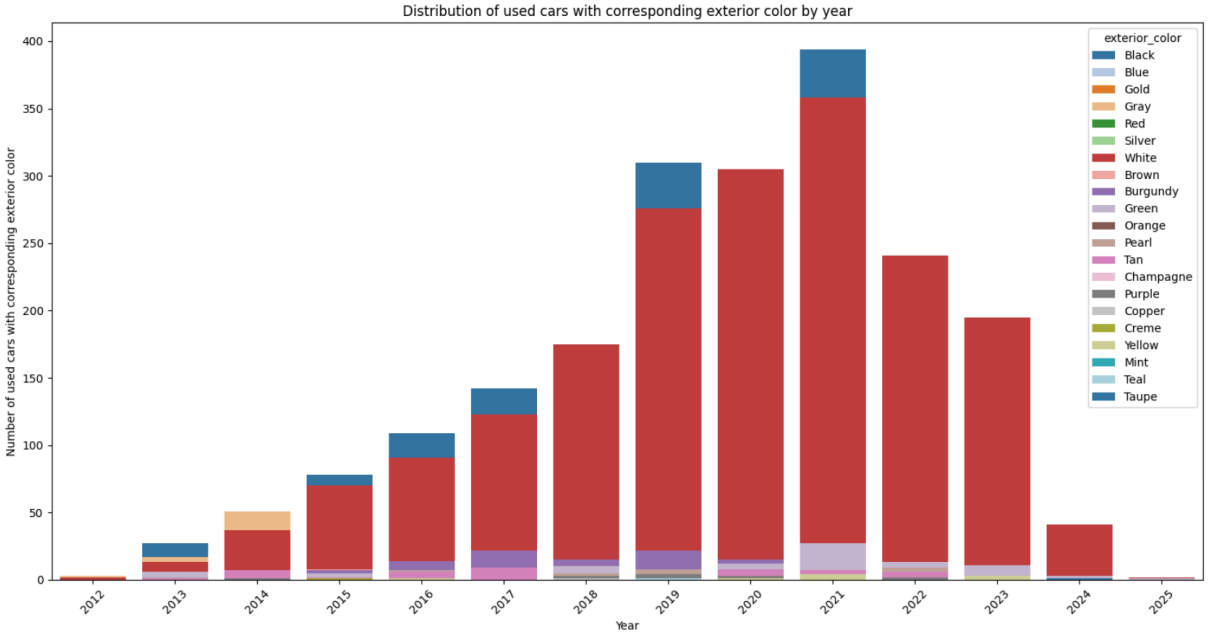
The taller boxes of Luxury and Sports Car suggests that they have more variability in resale values. Furthermore, large luxury car classes tend to have longer whiskers, which indicates a wide range of resale values. This may imply that it is more challenging to estimate the value for those types of car. Therefore, more investigations for them should be conducted in the further steps, also more data related to these types could be added into our dataset to counter the influence of their high variability.

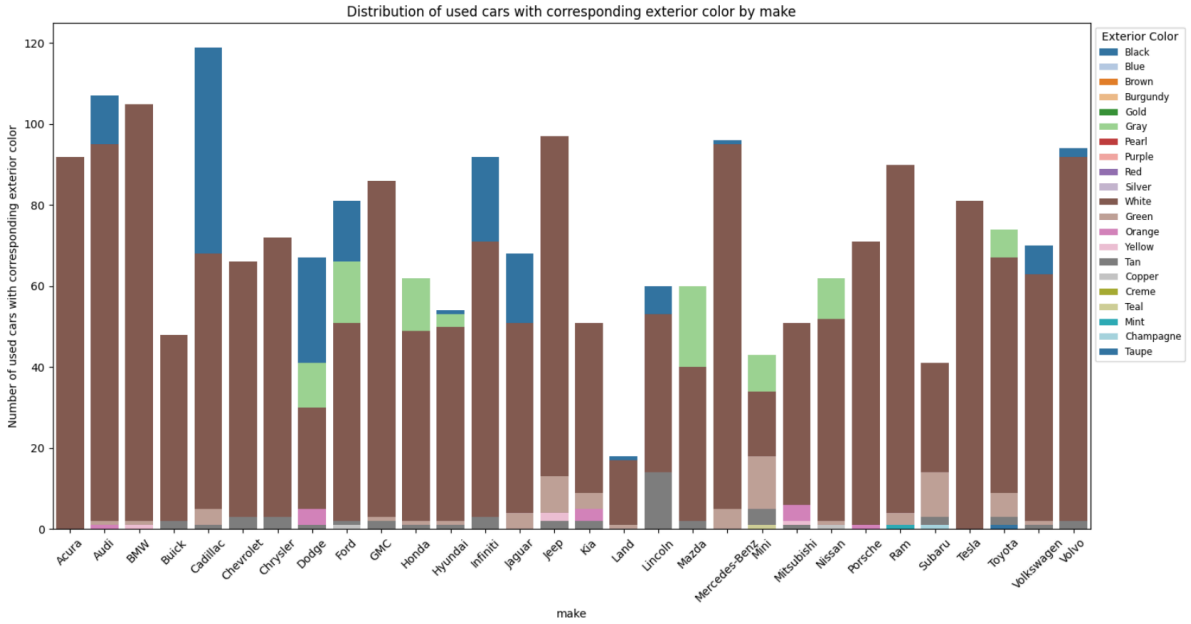
Question 3: What effect do different colors have on the price of a used car? (Jiabao Yao)

Hypothesis 1: The exterior colors of most used cars are relatively concentrated, except for special colors such as red. The used car price will not fluctuate significantly due to color.

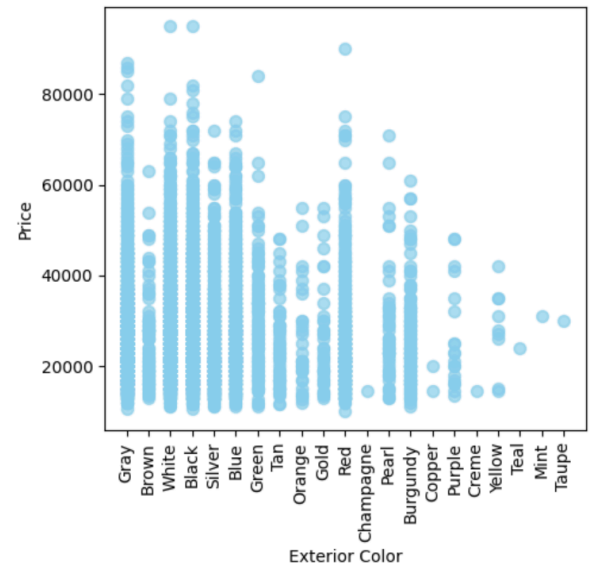


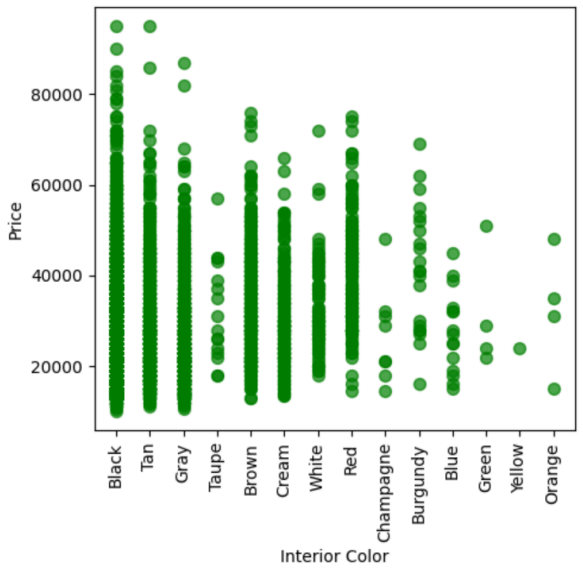
Firstly, based on our dataset, the exterior color of used cars is concentrated on white.

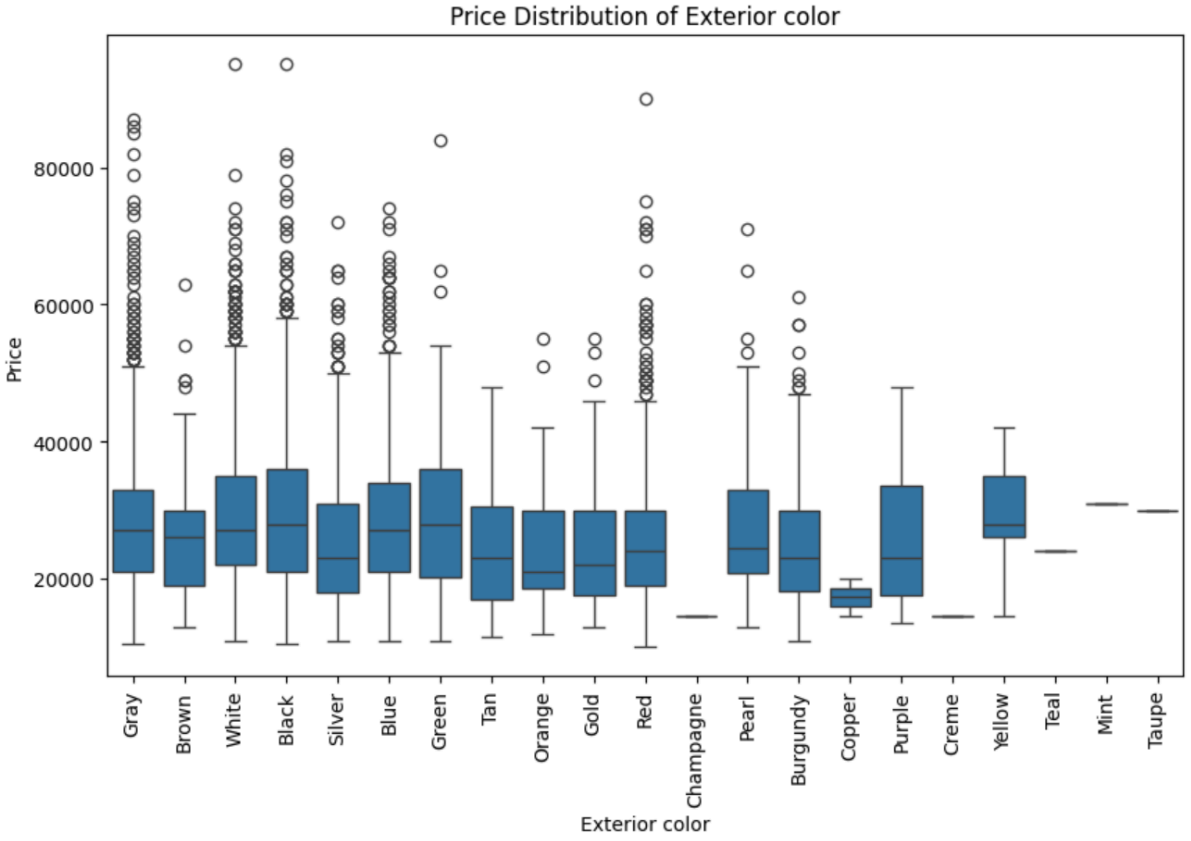




Secondly, as the years increase, most used cars have a white exterior color. This trend is shown in the graph above, which displays the exterior color of used cars by different manufacturers.



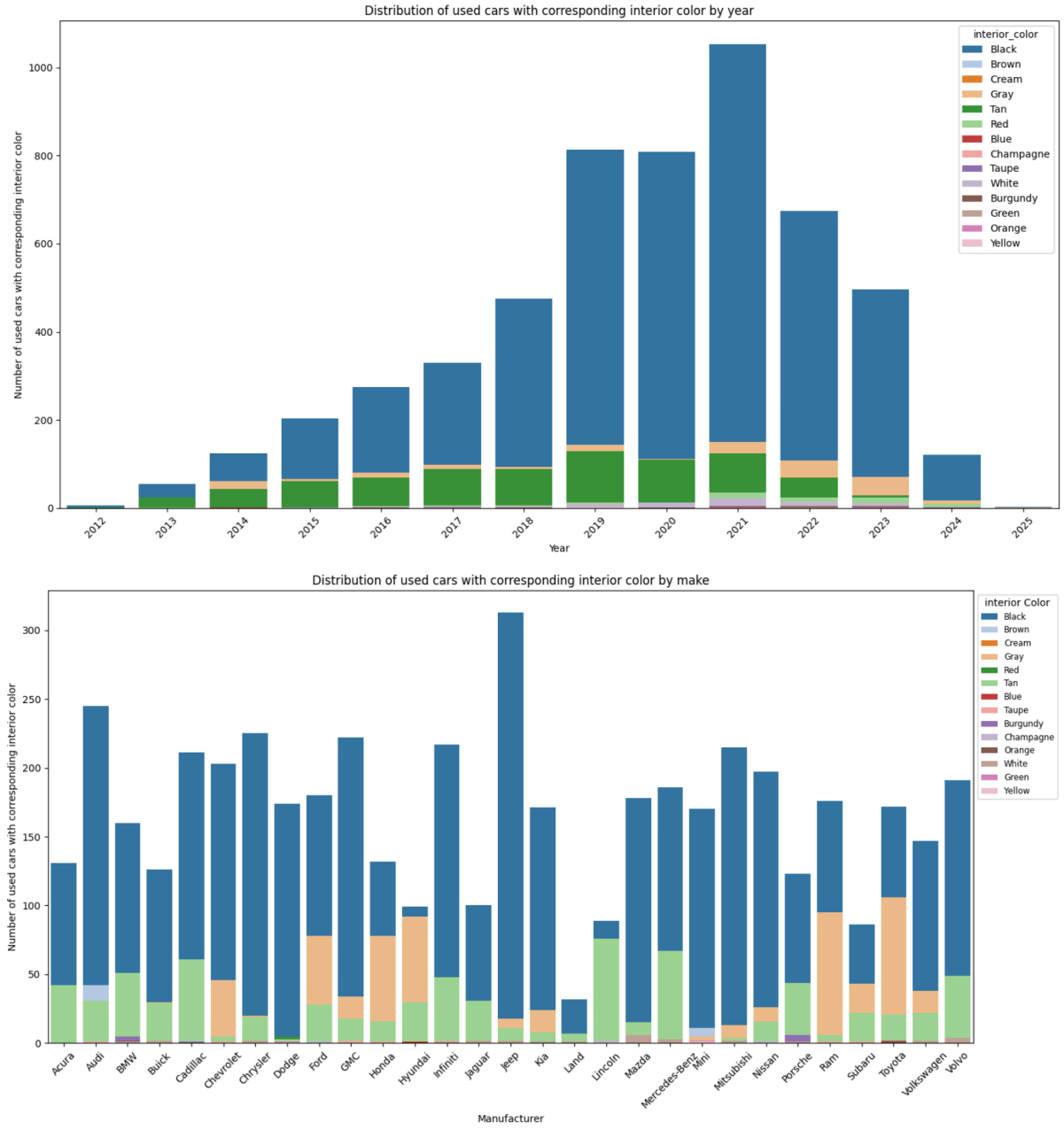




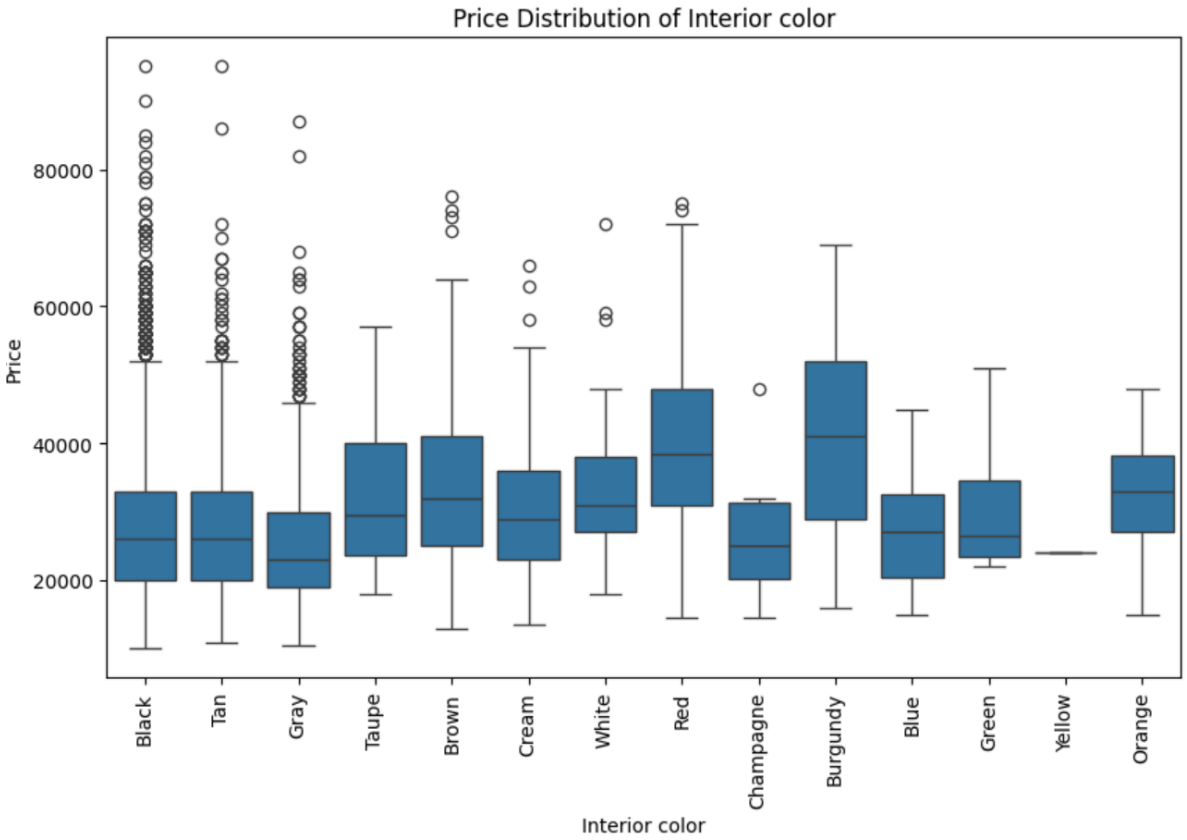
According to the graphs above, there is no clear relationship between exterior color and car price. However, we can observe that popular colors tend to have higher prices compared to other colors.

In conclusion, the exterior color of used cars has little impact on their prices because the color distribution is too concentrated. Popular exterior colors are generally not associated with lower prices.

Hypothesis 2: The interior colors of high-end vehicles are generally red, so the price of used cars with these colors will be relatively higher.



From the first hypothesis, we know that the interior colors of used cars are concentrated on black, gray, and tan. As shown in the graph above, as the years progress, most used cars have black interiors. This trend is evident in the graph, which shows the distribution of interior colors across different manufacturers, although Honda and Hyundai have some different preferences.

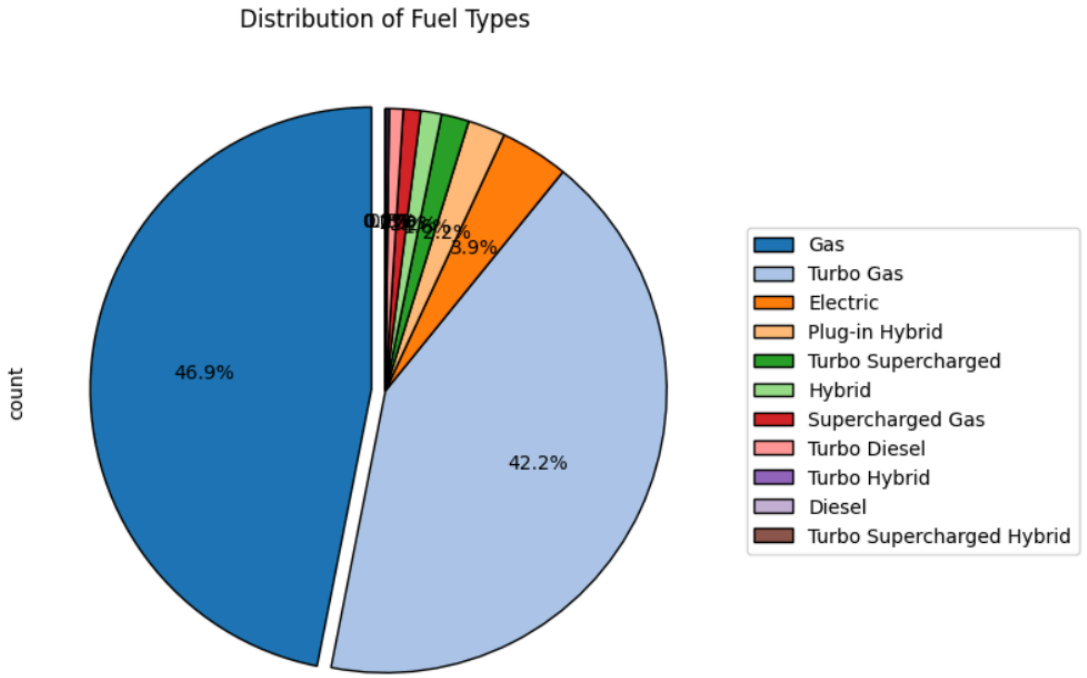


According to the graph above, the price of used cars with red and burgundy interior colors is relatively higher.

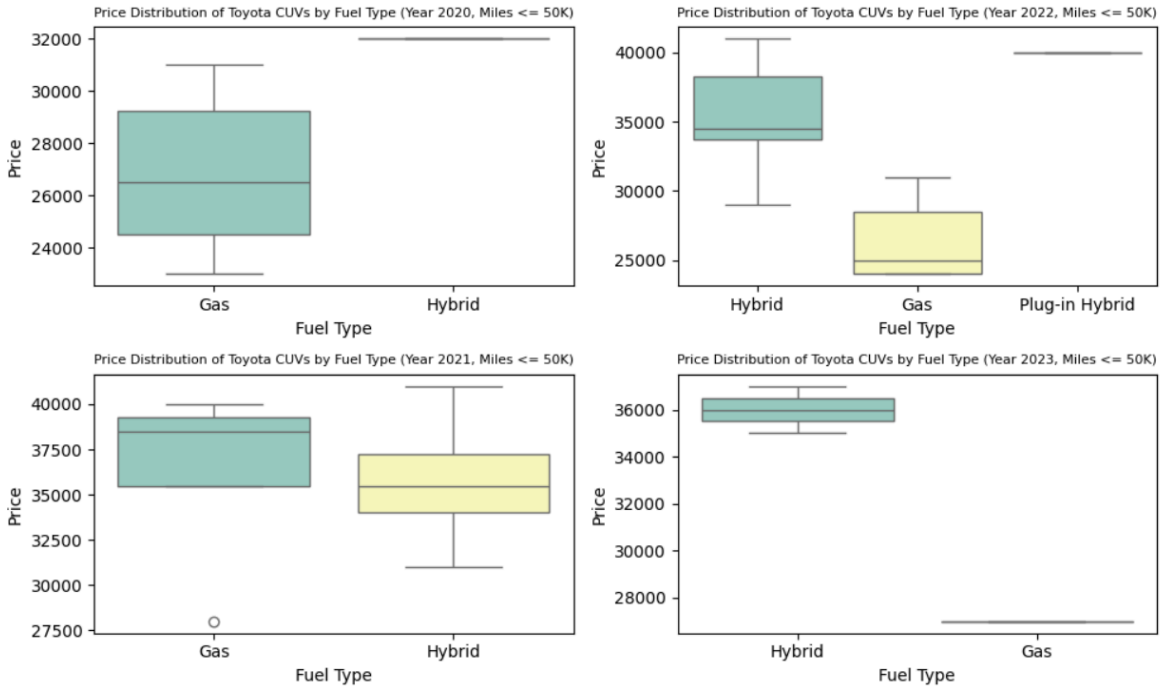
In conclusion, interior color has a slight impact on car prices, particularly for red and burgundy.

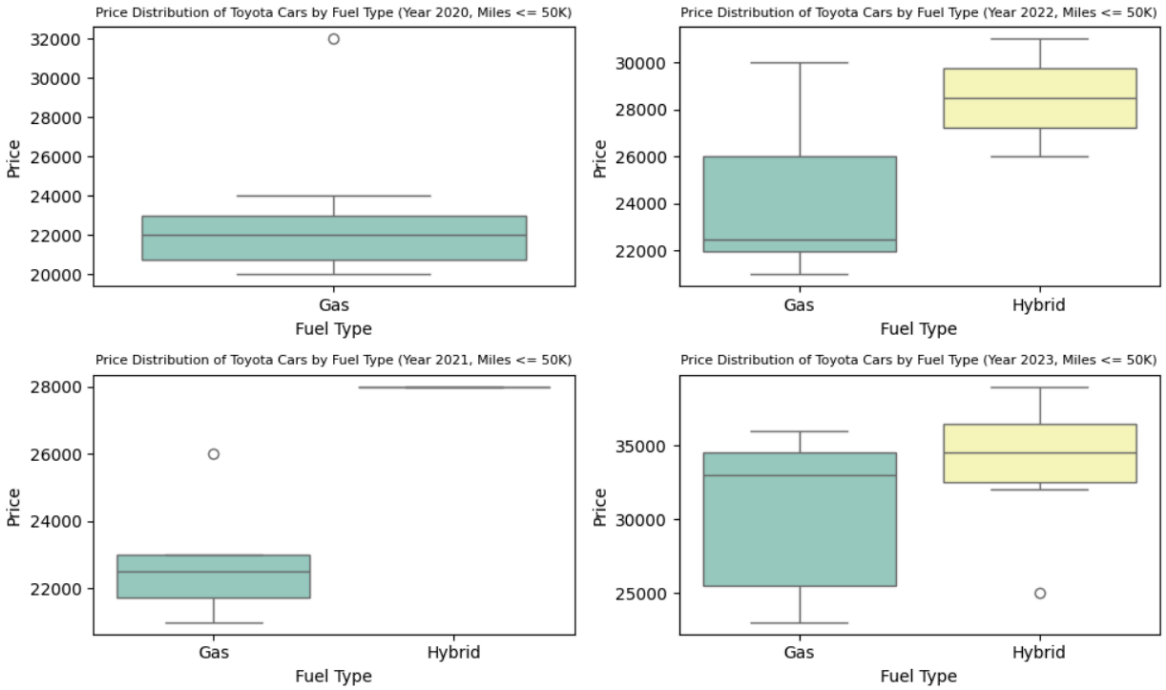
Question 4: How does fuel of a car impact its resale value across different brands? (Jiabao Yao)

Hypothesis 1: Hybrid cars have significant advantages in terms of cost and performance, so the price of these models may be relatively high.

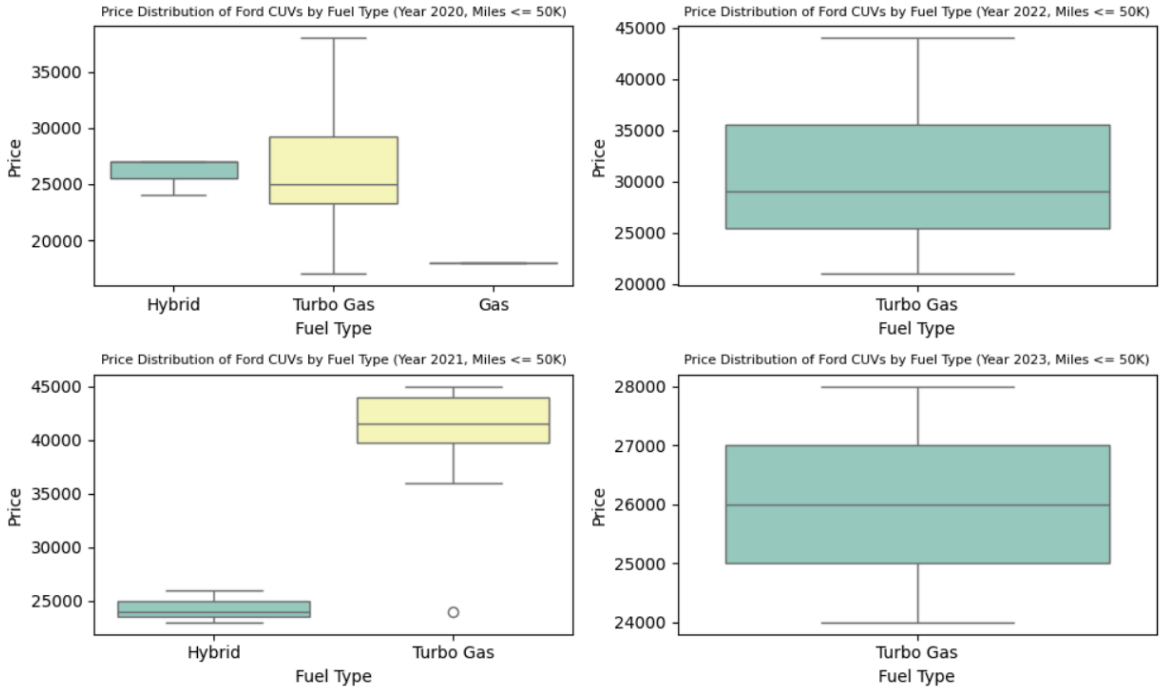


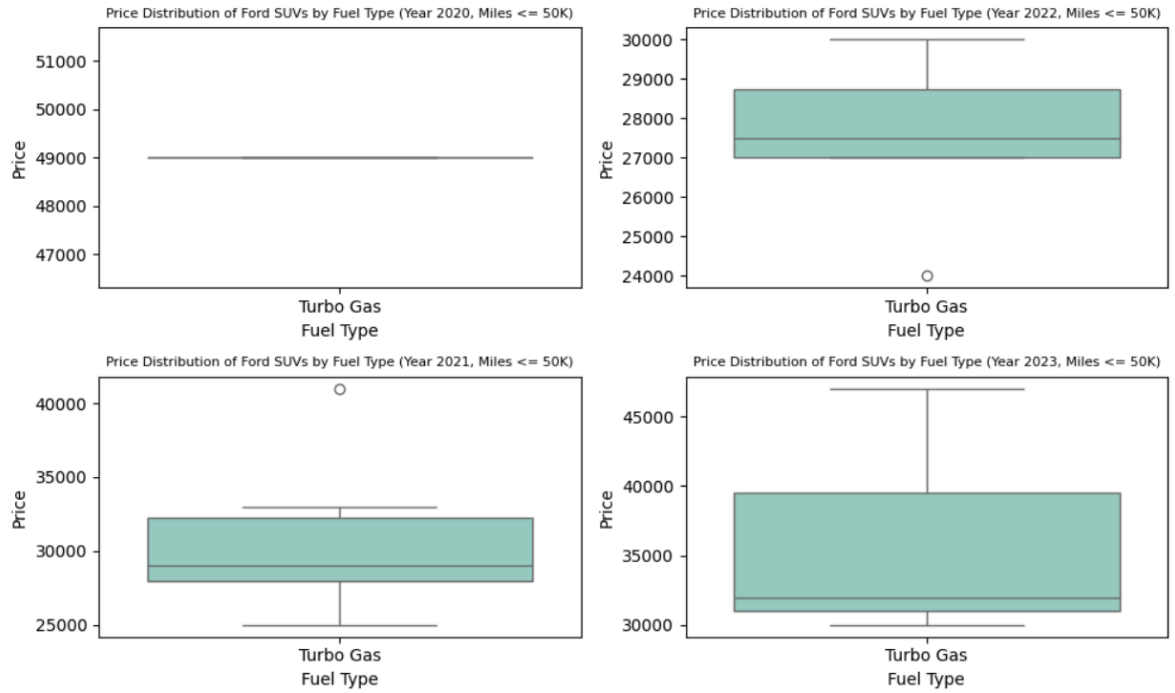
The distribution of fuel types for used cars indicates that most used cars on the market are gas and turbo gas powered, rather than hybrid.



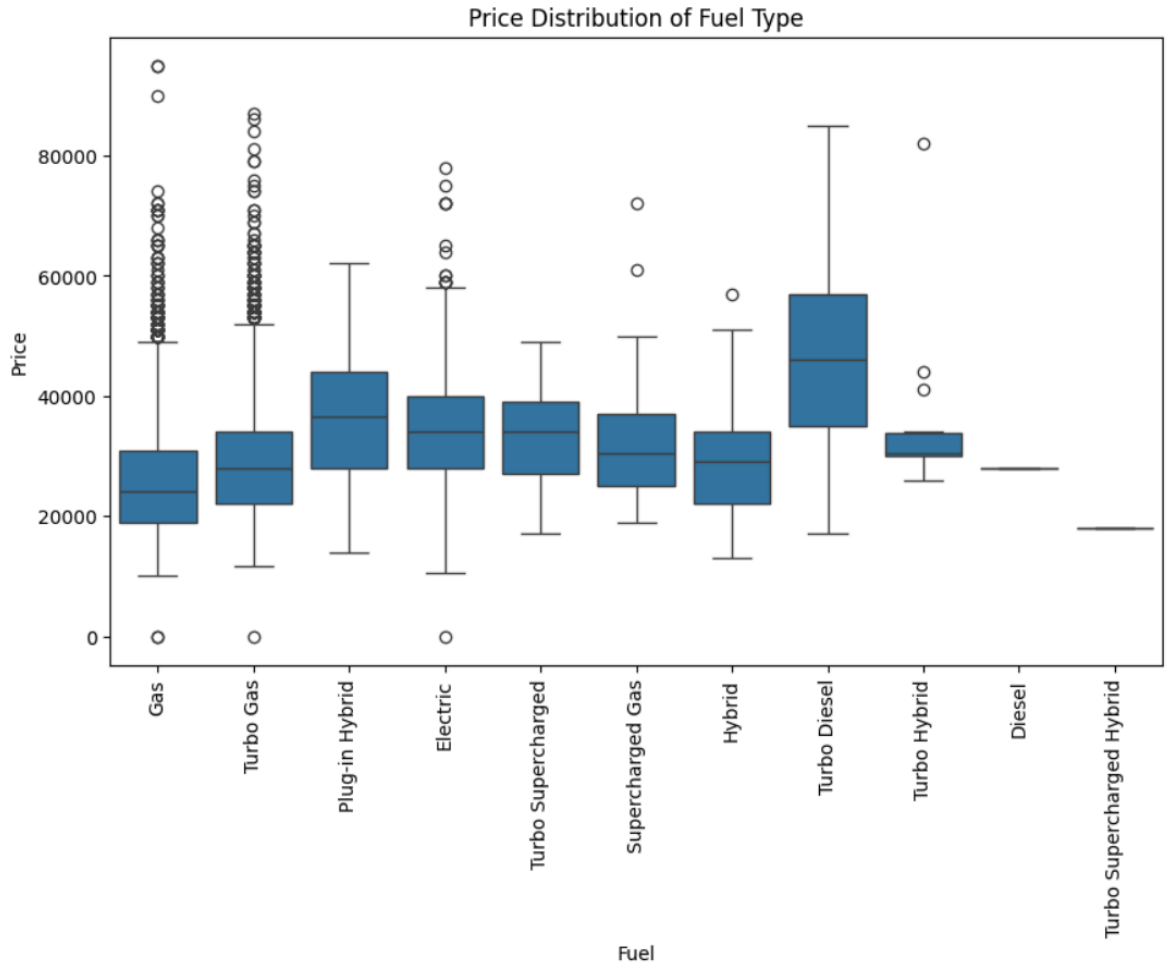


When focusing on Toyota CUVs, we notice that gas-powered models are very popular and tend to have higher prices compared to hybrids overall. However, different trends emerge when focusing on other types of Toyota cars.



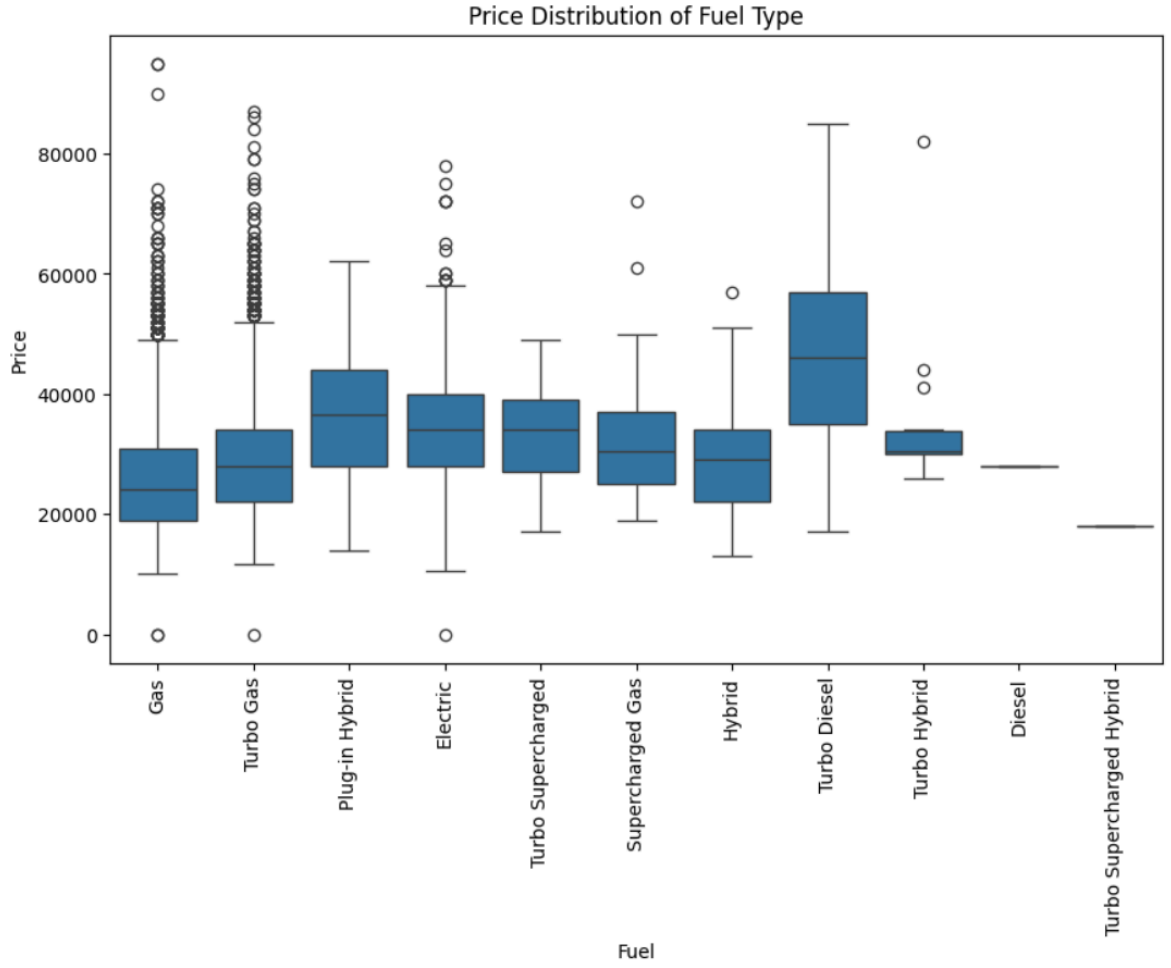


For Ford, turbo gas seems to be more in favor for most people.



In conclusion, most used cars on the market are powered by gas and turbo gas. Based on the median price of different fuel types, plug-in hybrids have the highest prices, excluding turbo diesel, while gas-powered cars have the lowest prices.

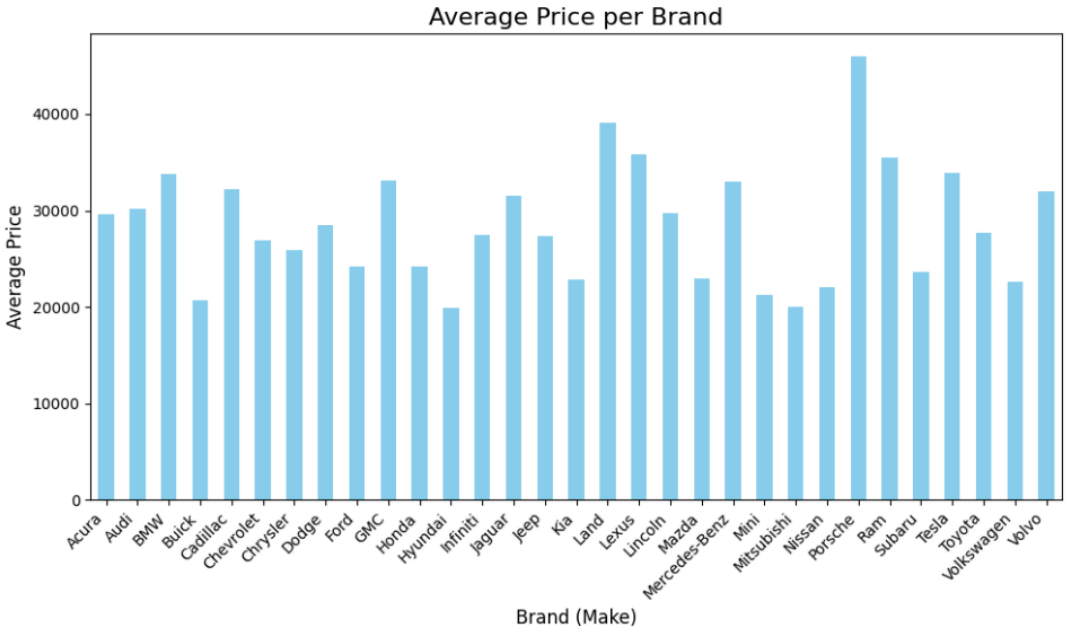
Hypothesis 2: Traditional vehicles, such as gas-powered ones, do not differ significantly in price from newer electric vehicle models due to their durability in the used car market.



Based on the price distribution of fuel types across all manufacturers, the price of gas is still influenced by new forms of energy. According to the median price of used cars, the prices of traditional fuel types, such as gas and turbo gas, are still lower than those of electric cars.

Question 5: Is the brand recognition of different cars an important factor affecting the resale price? (Chao Wu)

Hypothesis 1: Generally speaking, when purchasing used cars, some brands have relevant higher average resale price than other brands.




```
: top_5_brands = brand_avg_price.sort_values(ascending=False).head(5)
  low_5_brands = brand_avg_price.sort_values(ascending=True).head(5)
  top_5_brands, low_5_brands
```

```
: (make
   Porsche    45998.000000
   Land       39102.956268
   Lexus      35798.000000
   Ram        35472.193548
   Tesla      33868.689655
   Name: price, dtype: float64,
   make
   Hyundai       19954.468672
   Mitsubishi    20048.484252
   Buick         20720.965368
   Mini          21246.323276
   Nissan        22001.951100
   Name: price, dtype: float64)
```

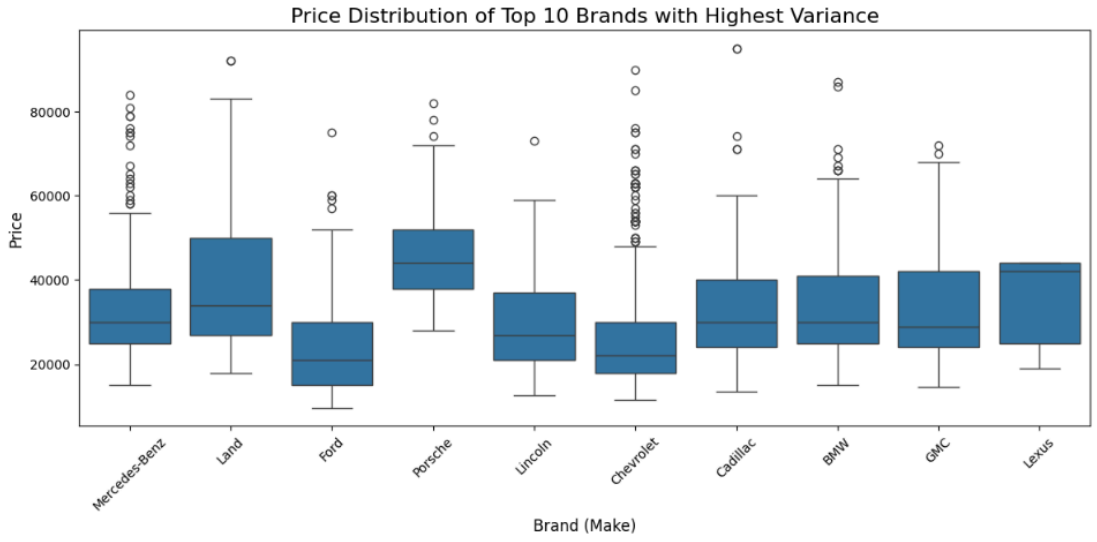
As shown in the bar chart of average price, the brand has certain impact to the reasale price. The Top 5 brand are Porsche, Land, Lexus, Ram, Tesla, which are common luxury or popular brands, with Hyundai, Mitsubishi, Buick, Minil,Nissan as the lowest 5 brands which are more affordable brands.

Hypothesis 2: Though brands contribute to resale price, different brands have varying price variances, and the depreciation rate of cars varies across different brands over years.

```
brand_price_variance = df_q5.groupby('make')['price'].var()
pd.set_option('display.float_format', '{:.2f}'.format)
brand_price_variance.sort_values(ascending=False).head(10)
```

```
make
Land            232462636.18
Chevrolet       187549853.09
GMC             155383789.20
BMW             145614411.35
Mercedes-Benz   139009542.46
Cadillac        137421356.55
Lincoln         121481604.06
Ford            118154131.27
Lexus           114885714.29
Porsche         113021459.23
Name: price, dtype: float64
```

The top 10 brands with the highest variance overall are mostly luxury brands but also include common brands such as Chevrolet, and Ford. Luxury brands have large differences between entry-level and high-end models, while some mass-market brands have a wide range of models (such as sedans, SUVs, and trucks), resulting in significant price differences. There can also be some outliers in certain brands(tailored cars in mass-market brands). To illustrate the assumption, the box plot is used to identify the outliers and distribution.

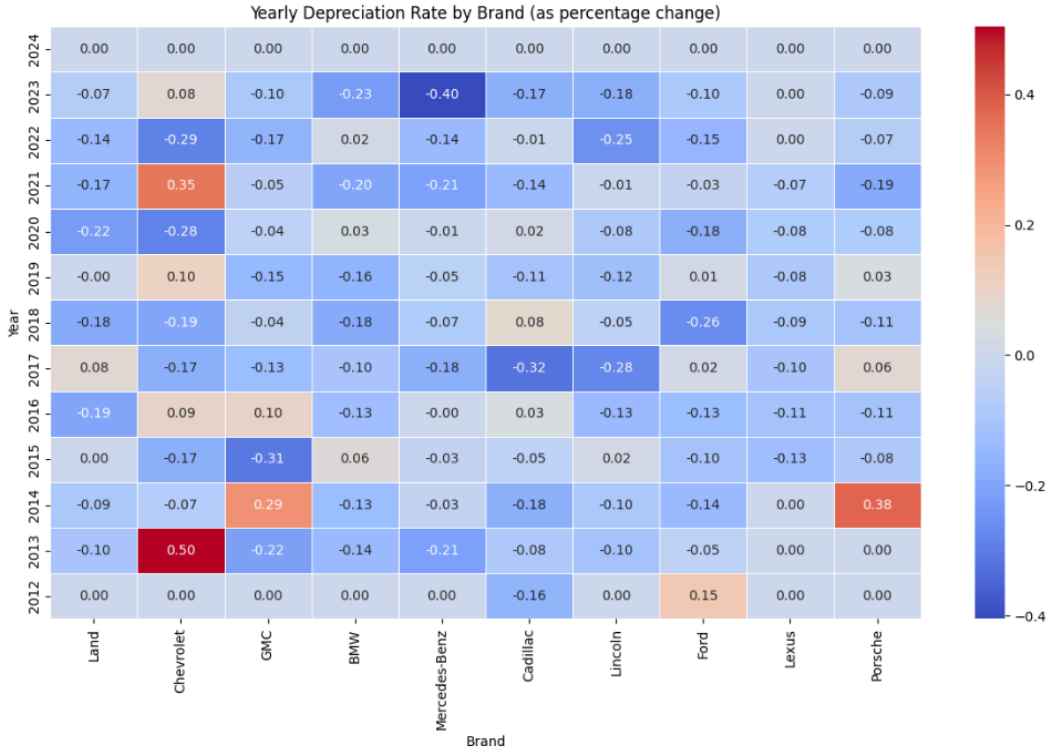


It is notable that Mercedes-Benz and Chevrolet are with the most outliers according to box-plot in the top 10 variance brands. This phenomenon can be analyzed below in two parts:

Luxury brands such as Land Rover, Mercedes-Benz, and Porsche have a wide price range, with prices from tens of thousands to over a million dollars. This large variation is due to these brands offering both entry-level luxury vehicles as well as high-end luxury or limited edition models. That's why Mercedes-Benz, Porsche, Cadillac and BMW have outliers above the upper bound.

For brands like Ford and Chevrolet, the price differences are also substantial, likely because these manufacturers produce a wide variety of vehicles, ranging from

economy sedans to large trucks and SUVs. This diverse lineup leads to a broader price variance. That's why Ford, GMC, Chevrolet have outliers above the upper bound.
We cannot simply remove these outliers in the analysis because they still contribute part of the resale price.

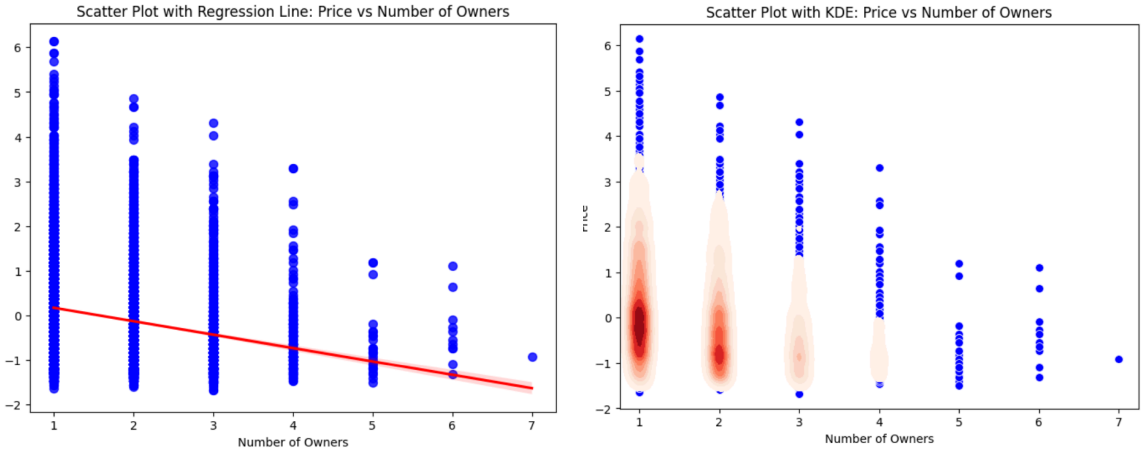


As shown in the heatmap, we calculate the deprecation rate for the top 10 brands from 2012 to 2014, most brands have shown varying degrees of reduction in depreciation rates (price increases) after 2020, especially after 2021, brands such as Land Rover and GMC have very small depreciation rates after 2021, and even tend to recover in price, indicating that these brands may have been driven by market demand during this period. The are unusual cases for Chevrolet, Ford and Porsche, when in 2013, 2014, and 2021, their prices iincreased significantly(50% for Chevrolet in the 2013 to 2014 period) instead of depreciating. This could be due to the severe variation of the supply and demand relationship in the auto market due to the manufacturer's marketing strategy.

Q6: Does the number of owners of used cars affect the resale price? (Chao Wu)

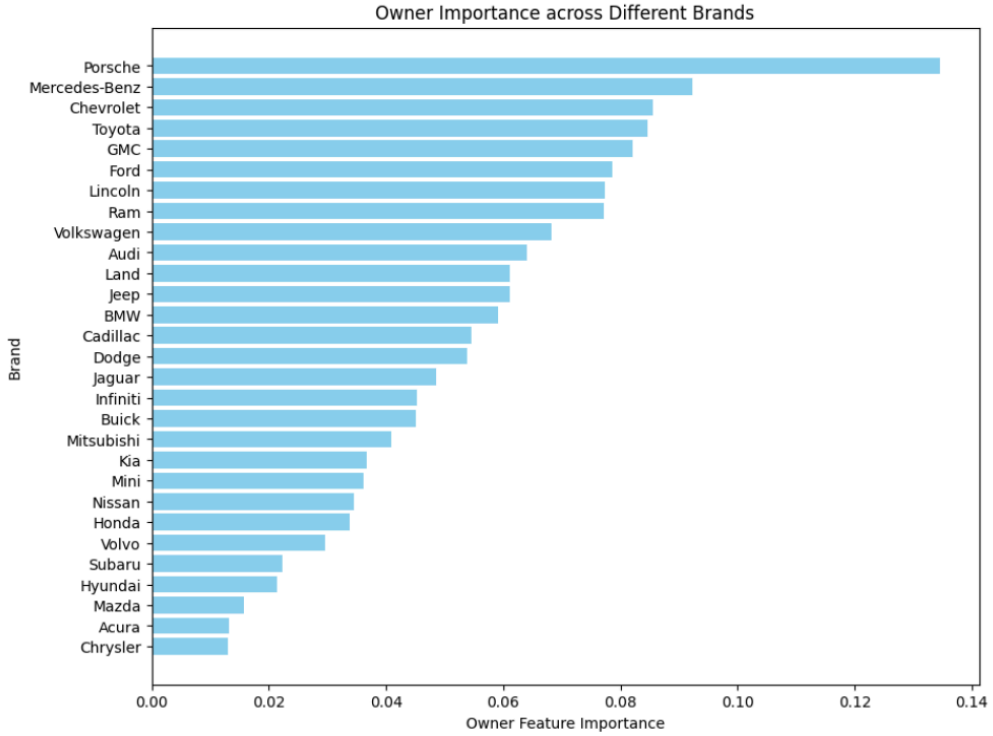
Data cleaning for Q6:
1. Drop NaN values for 'price', 'year', 'owner', 'mileage'
2. Use Z-score to normalize 'price', 'owner', and 'mileage' to contribute equally to the distance computation and avoid screwing the data.
3. Check if there are unknown outliers for price and year, remove the price = 0 rows and year = 2025 rows

Hypothesis 1: Number of owners has a negative impact on resale price.



There is a significant negative correlation between the number of owners and the price. However, vehicles with more owners show more dispersed and lower price distributions, and vehicles with fewer owners (such as 1 or 2) tend to have higher prices concentrated in a higher range.We can see that this trend is not absolute, especially for vehicles with fewer owners, where price variance is greater.

Hypothesis 2: Different brands have different price sensitivity to number of owners.

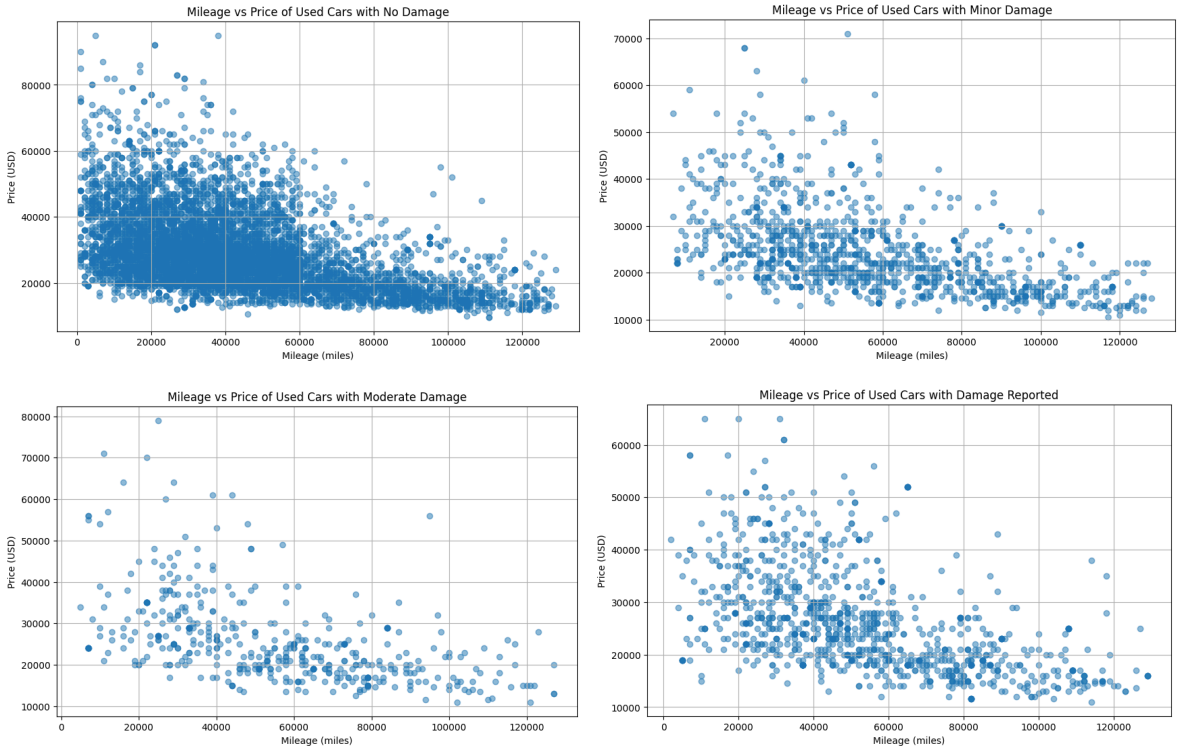


Luxury brands, such as Porsche and Mercedes-Benz, show significantly higher sensitivity to the number of owners compared to other brands.This is because buyers of luxury cars place more importance on the vehicle's ownership history. In contrast, non-luxury brands like Chrysler, and Mazda, buyers of non-luxury vehicles may be less concerned with the number of previous owners and more focused on

other factors like mileage or overall condition. For middle-class brands such as Toyota and Chevrolet, they have moderate sensitivity. These brands often maintain good resale value, indicating that they are moderate in value retention.
However, we can see from the training loop that brands smaller than 100 sam- ples will be excluded from training, this could result in bias. For example, Lexus, one of the luxury brands, is excluded from training. If Lexus is added, we can see it generates the highest feature importance of number of owners, this is the result of bias. We may need more balanced samples for calculating feature importance for different brands.

Question 7: How do the accidents or damage records of the used cars affect the resale price? (Shijie)

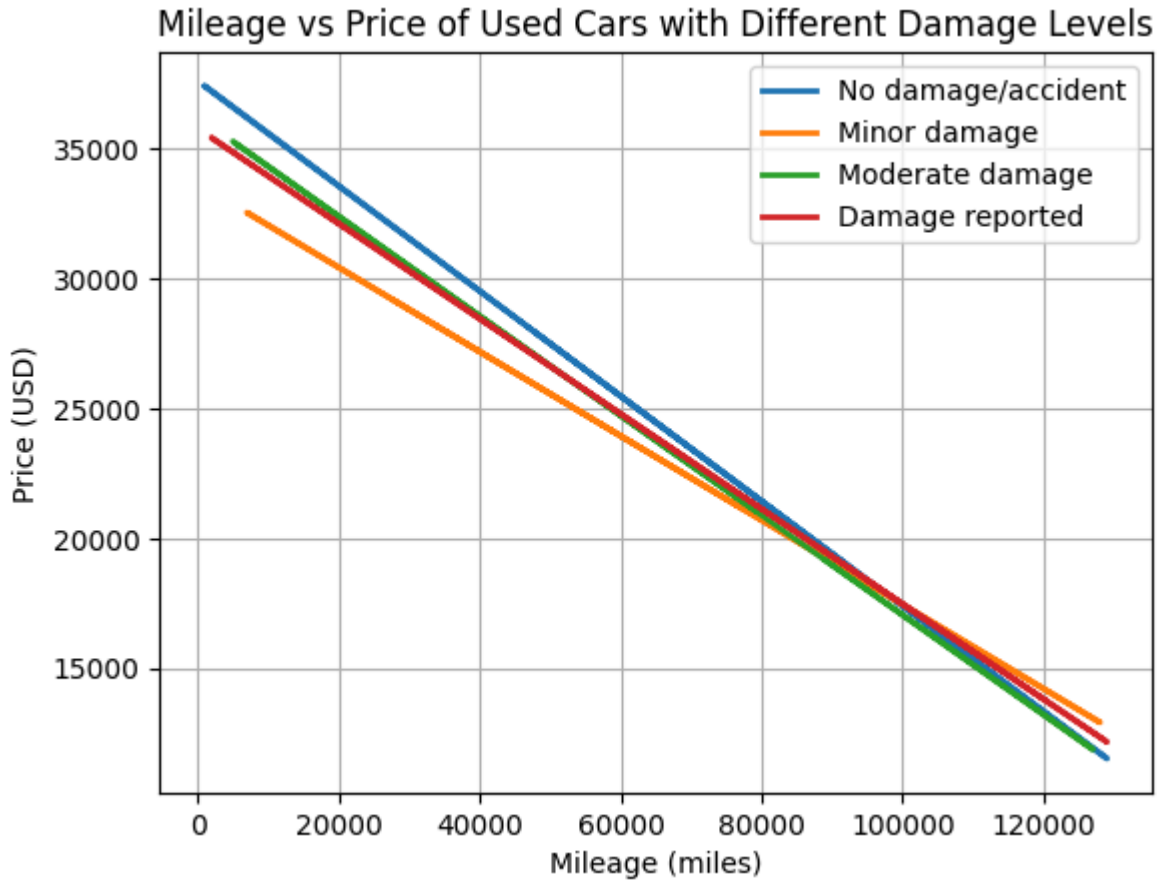
Hypothesis 1: the used cars with different damage levels share the similar pattern between mileage and price.



We can observe that used cars with different damage levels maintain the relationship between mileage and price. We further check it by computing the Pearson correlation scores as following and still get verify our assumptionL

Pearson correlation of used car with no damage: -0.47140781824890604
Pearson correlation of used car with minor damage: -0.5117868757602223
Pearson correlation of used car with moderate damage: -0.5172322127037858
Pearson correlation of used car with damage reported: -0.520546544670029

Hypothesis 2: used cars with higher damage level tend to have lower price.



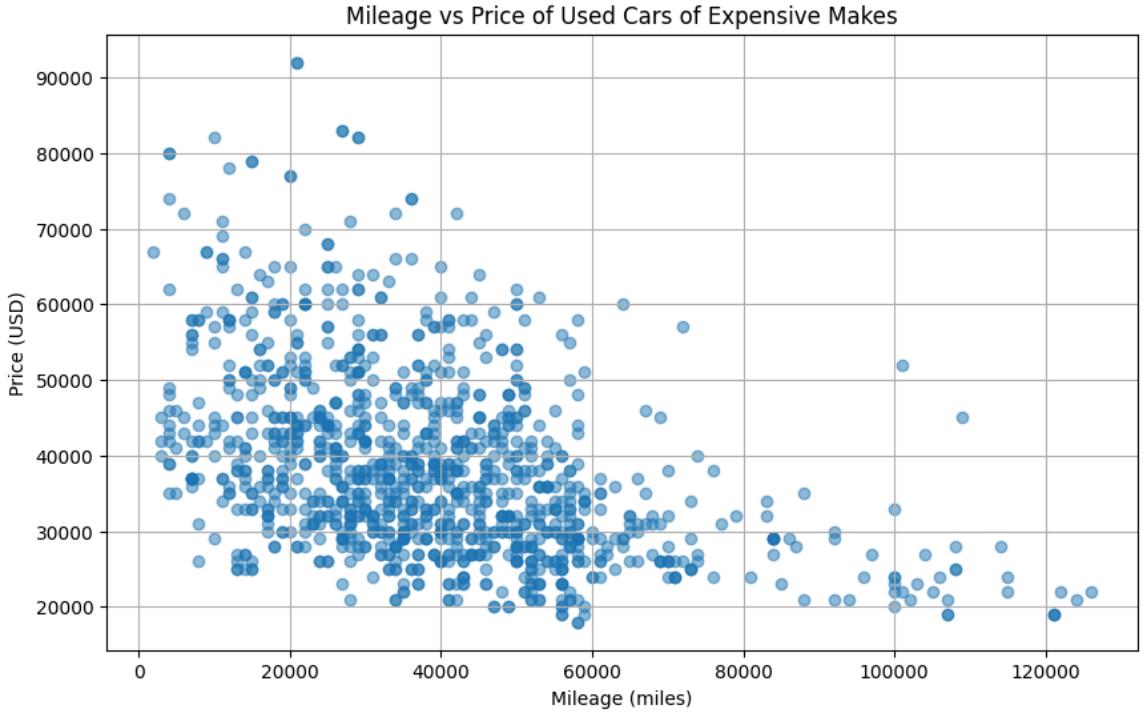
We conduct linear regression on the above data.
We can observe that in most mileage ranges, used cars with damage will have lower resale price than cars with no damage/accident which verify our hypothesis.

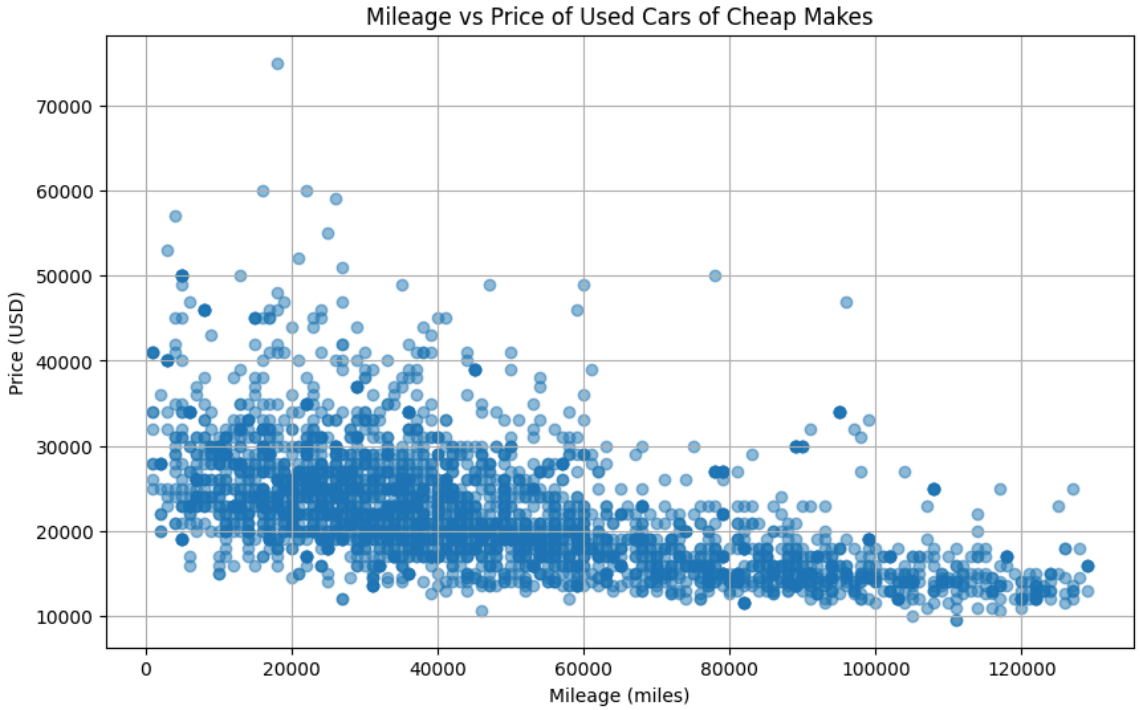
Question 7: For used cars with different makes, will the accident record affect the used cars' price differently? (Shijie)
Hypothesis 1: the used cars with different makes share the similar pattern between mileage and price.
We first split data into two groups: cheap and expensive brand groups. Here's the listed makes and their average prices:

lowest_ten_makes
make
Hyundai 19954.468672
Mitsubishi 20048.484252
Buick 20720.965368
Mini 21246.323276
Nissan 22001.951100
Volkswagen 22651.104938
Mazda 22927.378917
Kia 22931.190981
Subaru 23680.248148
Ford 24181.816742
highest_five_makes
make

Porsche 45998.000000
Land 39102.956268
Lexus 35798.000000
Ram 35472.193548
Tesla 33868.689655

And then:



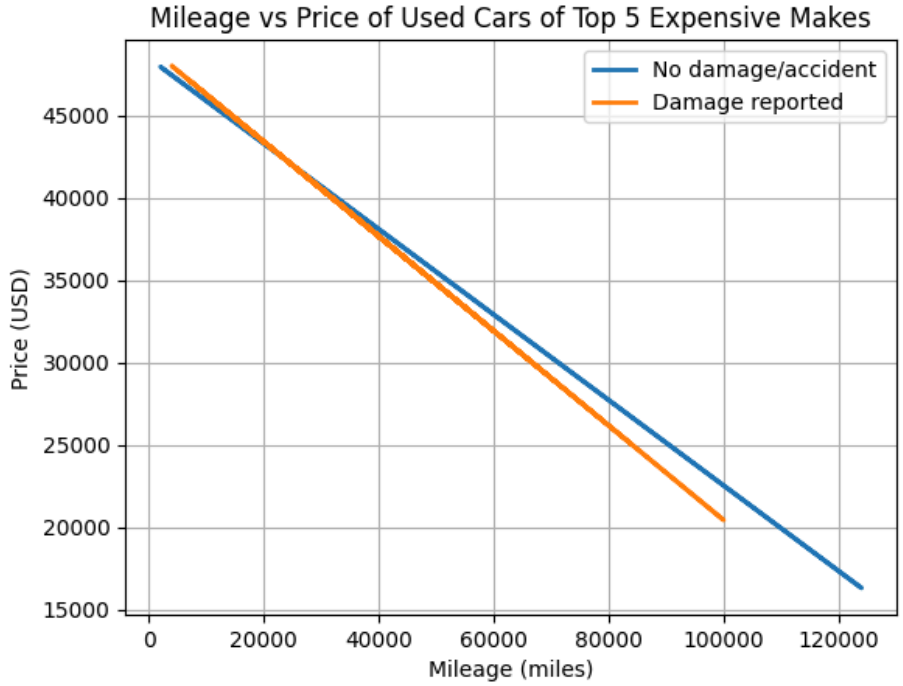


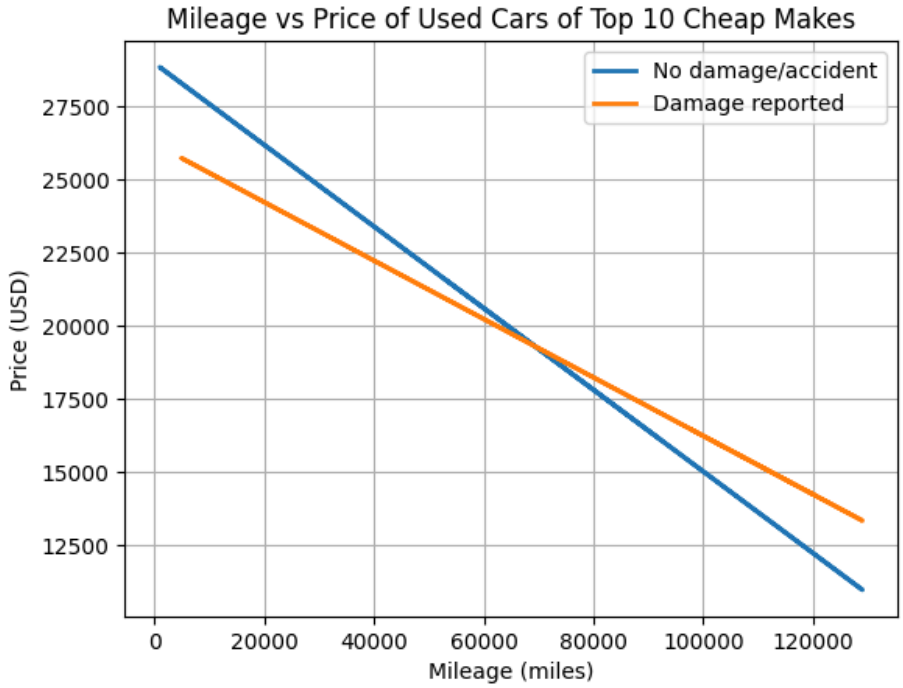
And compute Pearson correlation scores for each groups of used car:

Pearson correlation of used car of expensive makes: -0.4487958243732463
Pearson correlation of used car of cheap makes: -0.5643709389958534

From both plots and Pearson correlation values of used cars with different price levels share the similar pattern between mileage and price.

Hypothesis 2: Used cars with different brands will have different sensitivity on price towards the accident/damage record.





We conduct linear regression on the above two groups of data.

Since most used cars on sale have mileage smaller than 80000 (1114/107 for samples in plot). Thus we focus on this range.

We can observe that used cars with cheap auto makes are more sensitive with whether the used cars have been through accidents (the used car with accident

reported has much less price.) But for used cars of expensive makes, the effect from accidents is much more slight. It verifies our Hypothesis 2.