Used Cars Data Exploratory Analysis and Prediction in Belarus

Taohe Zhan, Nianqing Chen, Jie Huang and Liuyi Pan
Group 2

August 17, 2020

1. Introduction

1.1 data introduction

We get the data from the Kaggle website[1].This data was collected on the Internet and represents the car market. Dataset was collected at the dawn of 2019.

1.2 Purpose of this research

In this project, we aim to understand the data structure and the specific connotation of each variable and analyze the relationship between variables. Then we use different models to predict used car prices and give practical advice to people who needed used cars and to the firms of used cars. In addition, we also consider which prediction model is more appropriate when different features are selected. In this project, the four of us not only have division of labor, but also cooperation and communication. Only with our concerted efforts can the final project be completed.

1.3 some statements

In our project, the part Chen is responsible for in the outline is classification, which divides used cars into low-end, mid-end and high-end according to price. In the first attempt, Chen used the random forest classification method and got the result. However, after communicating, we chose another idea that is more in line with our theme, which is to use two different regression models to predict prices, compare the accuracy of the two prediction results. After changing the thinking, we find that it can give this project more practical meaning.

2. Data cleaning and preprocessing

2.1 Data overview

I use the code ‘shape’, ‘info’ and ‘describe’ codes to know more about the data. The ‘used cars’ data has 12 columns and 56244 rows, including key features such as the production year, mileage of the cars, price, brands, etc.

2.2 Deal with missing values

I use ‘isnull.sum’ code to check the number of missing values in every column. According to the outcome, I find that ‘volume (cm3)’ has 47 missing values, ‘drive_unit’ has 1905 missing values, and ‘segment’ has 5291 missing values. Because the ‘volume (cm3)’ is numerical data, I decide to use its mean value to fill the missing values. As for the ‘drive_unit’ and ‘segment’, which is categorical data, I use the previous value to impute the missing values. Finally, I use the code to check whether the missing values have been imputed successfully.

2.3 Handle outliers

I use the ‘describe’ function to check the data. And I find that the maximum value and minimum value for ‘mileage’ is 9999999 and 0 respectively. So I reckon they are outliers in the used cars data and delete them.

2.4 Deal with duplicate values

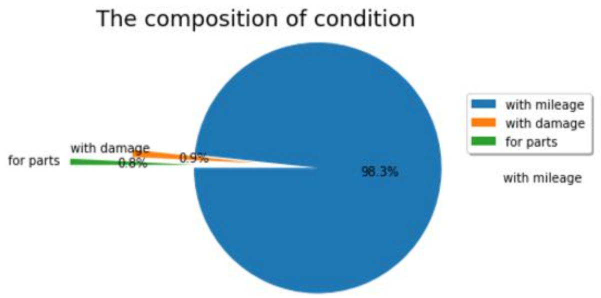
I use ‘duplicated’ code to check the duplicated values in this data and use ‘drop_duplicates’ code to delete them.

3. Fundamental features description

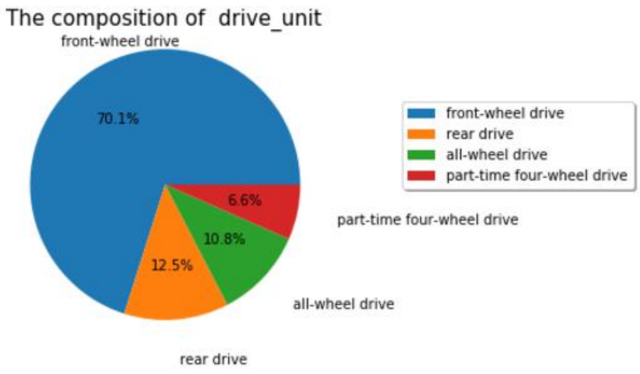
3.1 Initial understanding of the features

In order to better understand these data and present them more intuitively, I made several pie charts to describe these data. From these pie charts, we can clearly know the proportion of second-hand cars under different indicators.

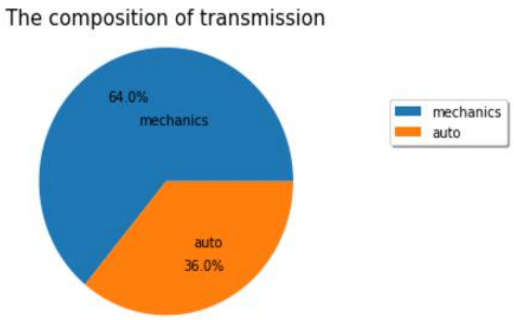
First of all, I download the data from Kaggle and I used pandas to read the data.



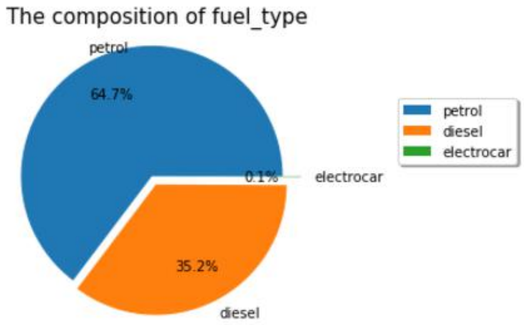
The first figure is the condition of a second-hand car. There are 98.3% of cars with mileage, 0.9 % of cars with damage, and 0.8% of car for parts. From this pie chart we can intuitively see that most of cars are with mileage.



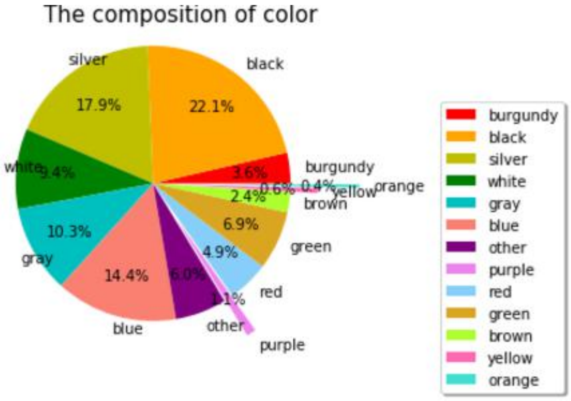
The second graph is about the drive unit. From this chart we can see that most of used car are front-wheel drive, and it account for as much as 70.1%



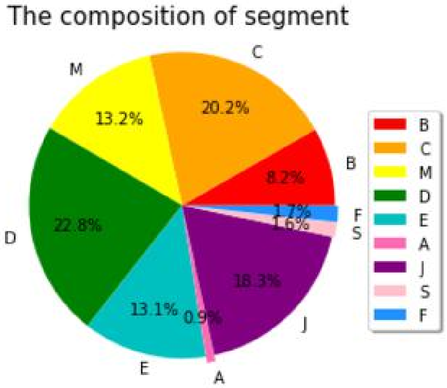
The third chart is about the composition of transmission. This chart indicate that most of used cars are mechanics, which means for the used car industry so that we can know the mechanics transmission is more popular than other types



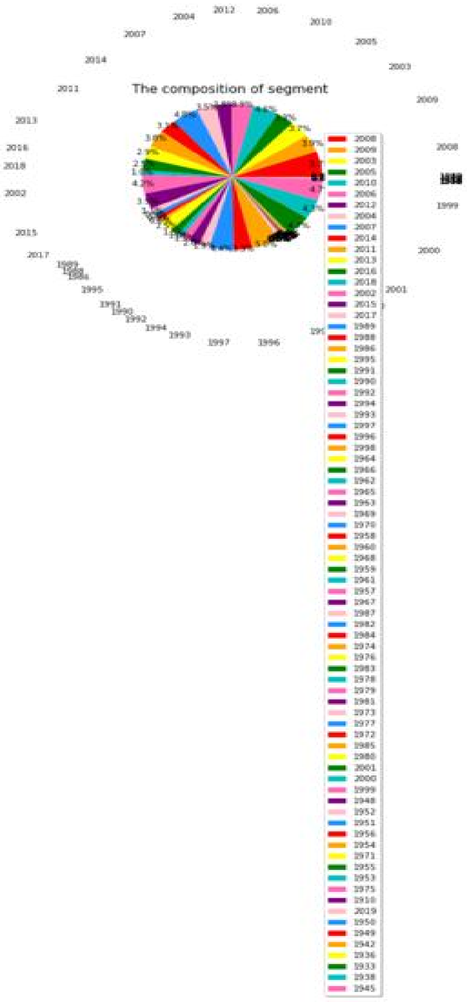
The 4th chart is about the fuel type. We can see there are 64.7%of cars use petrol ,so we can know that in the used cars industry, most of cars are use petrol, and there are only 0.1% of cars use the electricity so that we can know the electronic cars is not as popular as other types of cars



The 5th graph is about the composition of color. From this chart we can easily see which color people like most. For example,we can see that black is the most popular color, and the orange is the fewest, and it's account for as little as 0.4%



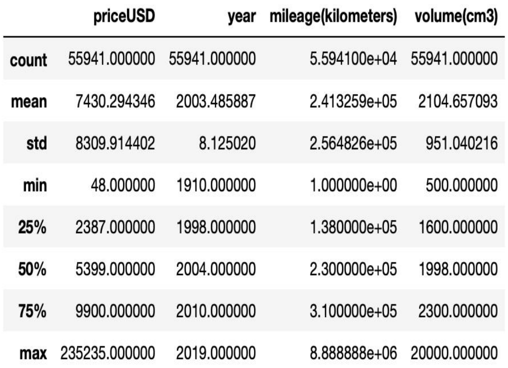
This chart is about the composition of segment. This chart shows which class of cars is the most popular in used cars. This chart shows which class of cars is the most popular in used cars. For example, form the picture we can easily see D is the most popular segment of cars.



The last chart is about the sell of the year, and we can see which year sells the most cars.

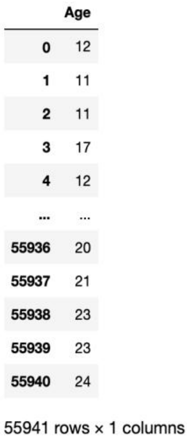
3.2 further understanding of the features

I visualized the data in each column, which is the most appropriate, and linked some of the variables together for analysis. Finally draw the conclusion after analysis.



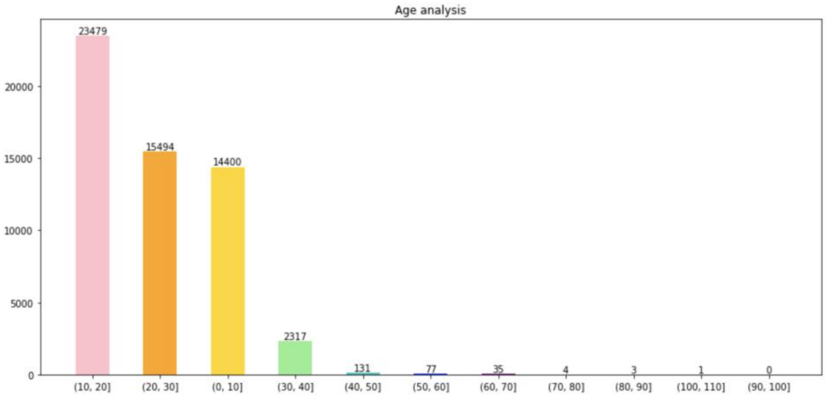
	priceUSD	year	mileage(kilometers)	volume(cm3)
count	55941.000000	55941.000000	5.594100e+04	55941.000000
mean	7430.294346	2003.485887	2.413259e+05	2104.657093
std	8309.914402	8.125020	2.564826e+05	951.040216
min	48.000000	1910.000000	1.000000e+00	500.000000
25%	2387.000000	1998.000000	1.380000e+05	1600.000000
50%	5399.000000	2004.000000	2.300000e+05	1998.000000
75%	9900.000000	2010.000000	3.100000e+05	2300.000000
max	235235.000000	2019.000000	8.888888e+06	20000.000000

3.2.1 Descriptive analysis of data



	Age
0	12
1	11
2	11
3	17
4	12
...	...
55936	20
55937	21
55938	23
55939	23
55940	24

55941 rows × 1 columns

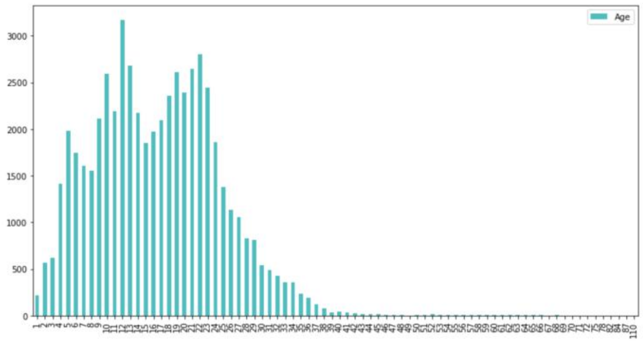


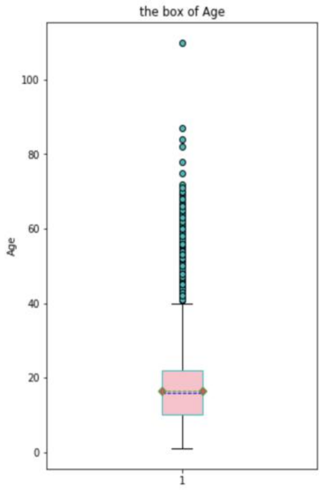
After data preprocessing, I used python to perform a descriptive analysis of the overall data.

Conclusion:

① Understand and be familiar with the basic statistics of the digital columns (priceUSD,year,mileage,volumn) in the used car data, including sum, average, and maximum. Minimum, dichotomous, third, quartile, and std.

3.2.2 Age analysis of second-hand cars





First, I process the data in the year column, and subtract the year column data from 2020 to get the age of each used car. I divide the car ages into several groups and count the total number of cars in each group. Secondly, I list how many used cars there are for each age and plot a figure to show. Finally, I make a box plot of the Age data .

Conclusion:

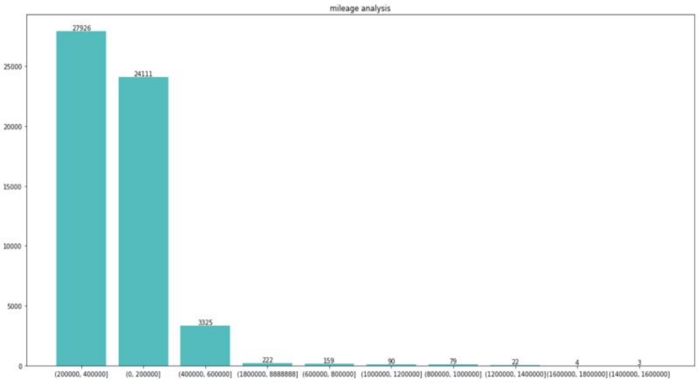
①The number of cars in different sections and the ranking of the sections.

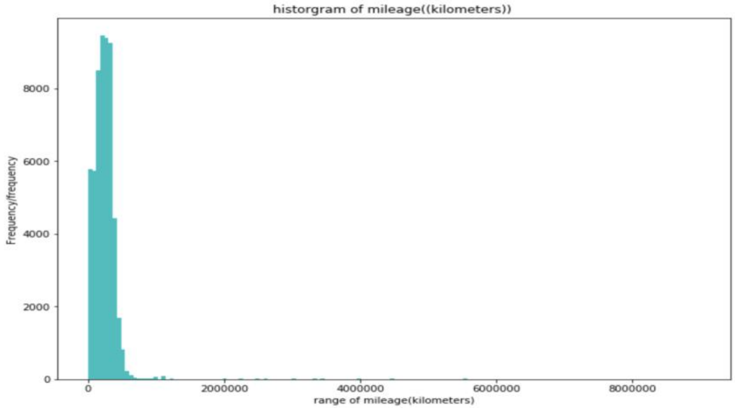
②Generally speaking, the age of used cars is concentrated in 3-30 years. After 30 years old, the number of used cars will gradually decrease with the increase of age.

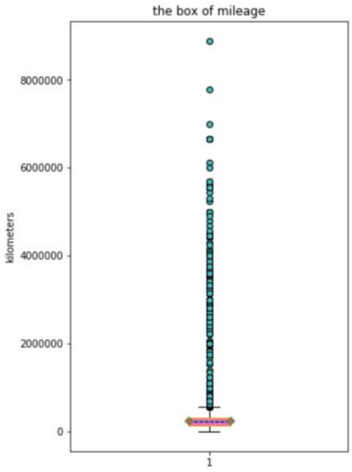
③The total number of second-hand cars with the age of 12 years is the largest.

④There are many outliers in age data, the upper limit is about 40, the lower limit is about 2, the median and mean are very close to about 15, the upper quartile is in the early 20s, and the lower quartile is about 10.

3.2.3 Driving mileage analysis







First of all, I divide the miles into several sections (uneven), and show the number of cars in each section. Then， I plot a histogram about mileage data. Last but not least, I plot a box of mileage data.

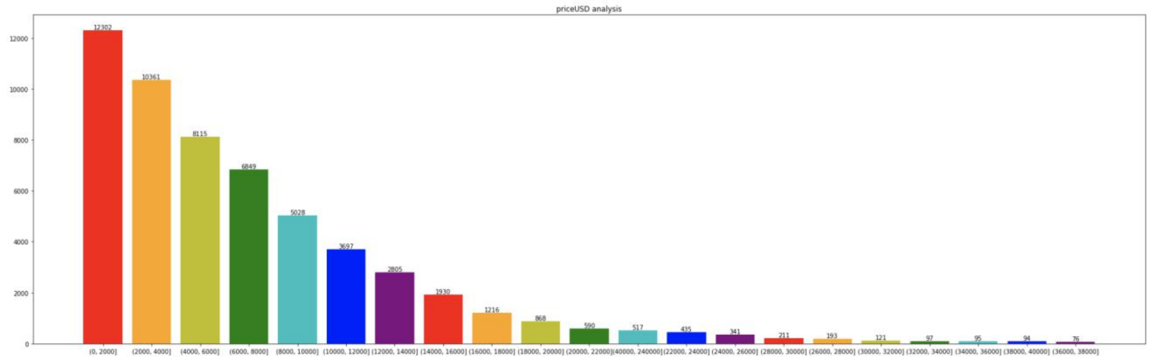
Conclusion:

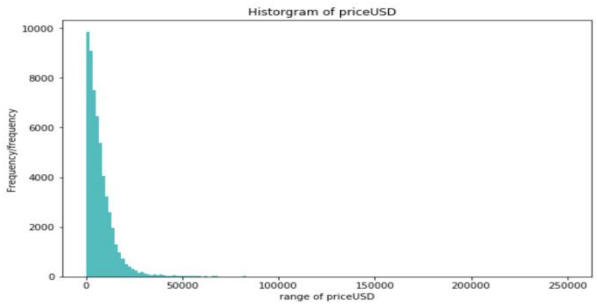
① How many used cars are in different mileage intervals and the order of the intervals. the interval with the most vehicles is 200000-400000 kilometers, and there are 27926 vehicles in total.

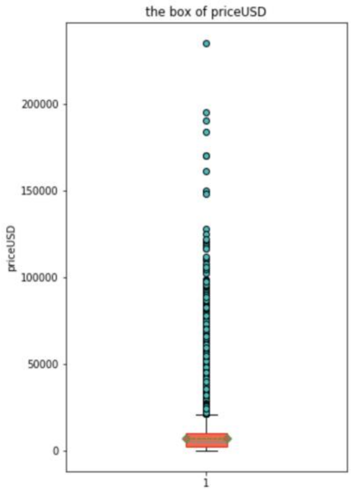
② It can be clearly seen that the mileage of used cars is mainly concentrated within 1,000,000 kilometers.

③the box in the figure is lower than 500,000, with more outliers, the median and the mean are closer, and the difference between the upper quartile and the lower quartile is not very large.

3.2.4 PriceUSD analysis







I divide the prices into different price ranges and count the number of vehicles in each range and plot a bar chart to show. I plot a histogram about priceUSD data. Similarly, I plot a box of PriceUSD.

Conclusion:

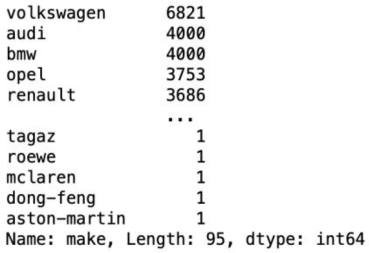
①How many used cars are in different price ranges, and the order of the ranges is in descending order of the number of used cars. there are 12,302 second-hand cars whose prices are between 0 and 2000 dollars, 10,361 are between 2000 and 4000 dollars, etc.

However, only 76 used cars are between 36,000 and 38,000 US dollars.

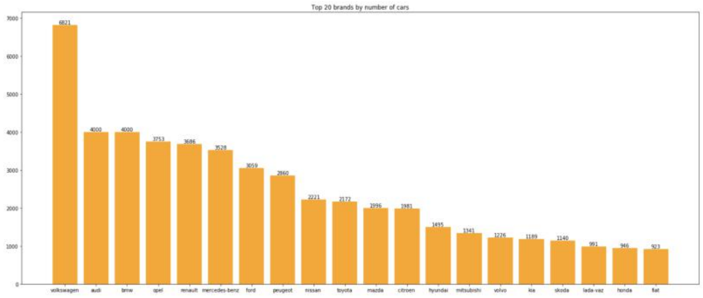
② It can be seen from the histogram that the price of used cars is mainly concentrated in 0-25000 US dollars, which means that the price of used cars is mostly in this price range.

③The relationship between the median and the mean. It can also be found that the position of the upper limit is close to the most concentrated position of the data in the histogram. there are still many outliers in the column price.

3.2.5 Make analysis



```
volkswagen       6821
audi             4000
bmw              4000
opel             3753
renault          3686
                 ...
tagaz               1
roewe               1
mclaren             1
dong-feng           1
aston-martin        1
Name: make, Length: 95, dtype: int64
```

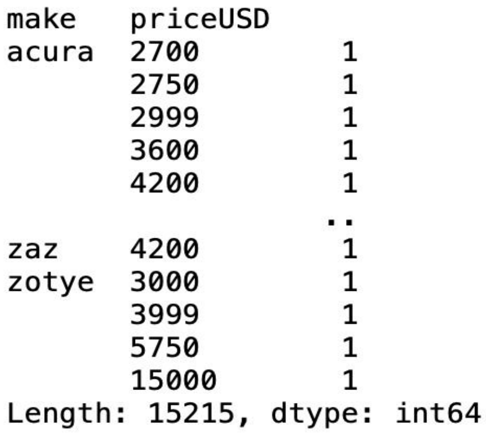


There are 95 categories in this column, so I select the top 20 brands and display the number of cars of each brand.

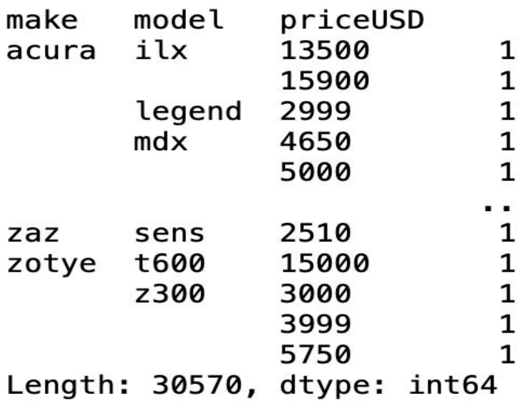
Conclusion:

①the number one brand is volkswagen with 6,821 vehicles, and the 20th is fiat, which has 923 car.

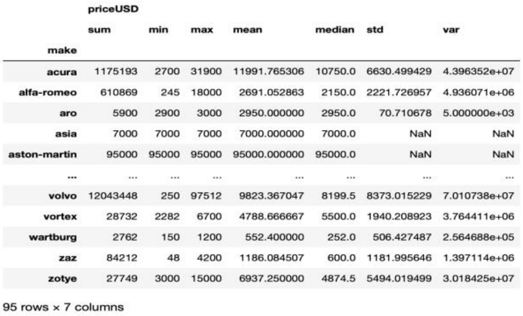
3.2.6 Make &Model analysis



```
make   priceUSD
acura  2700        1
       2750        1
       2999        1
       3600        1
       4200        1
                  ..
zaz    4200        1
zotye  3000        1
       3999        1
       5750        1
       15000       1
Length: 15215, dtype: int64
```

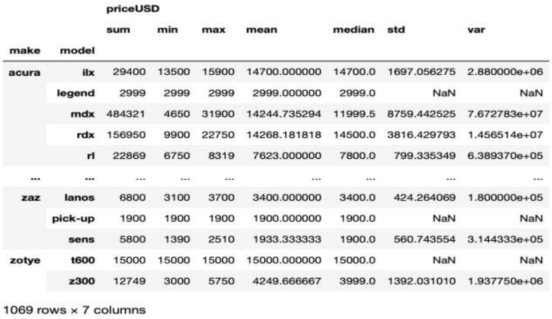


```
make   model   priceUSD
acura  ilx     13500       1
               15900       1
       legend  2999        1
       mdx     4650        1
               5000        1
                          ..
zaz    sens    2510        1
zotye  t600    15000       1
       z300    3000        1
               3999        1
               5750        1
Length: 30570, dtype: int64
```



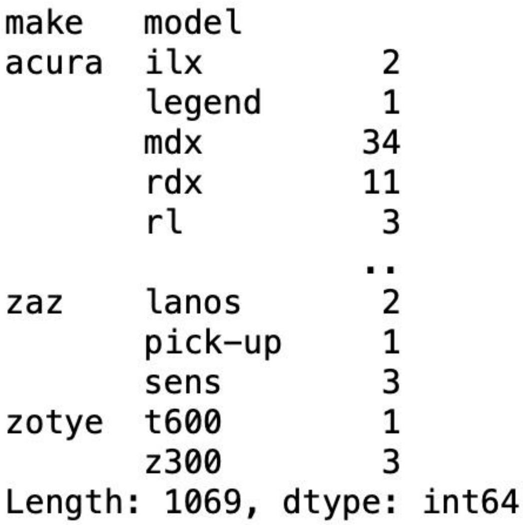
	priceUSD						
	sum	min	max	mean	median	std	var
make							
acura	1175193	2700	31900	11991.765306	10750.0	6630.499429	4.396352e+07
alfa-romeo	610869	245	18000	2691.052863	2150.0	2221.726957	4.936071e+06
aro	5900	2900	3000	2950.000000	2950.0	70.710678	5.000000e+03
asia	7000	7000	7000	7000.000000	7000.0	NaN	NaN
aston-martin	95000	95000	95000	95000.000000	95000.0	NaN	NaN
...	...	...	...	...	...	...	...
volvo	12043448	250	97512	9823.367047	8199.5	8373.015229	7.010738e+07
vortex	28732	2282	6700	4788.666667	5500.0	1940.208923	3.764411e+06
wartburg	2762	150	1200	552.400000	252.0	506.427487	2.564688e+05
zaz	84212	48	4200	1186.084507	600.0	1181.995646	1.397114e+06
zotye	27749	3000	15000	6937.250000	4874.5	5494.019499	3.018425e+07

95 rows × 7 columns

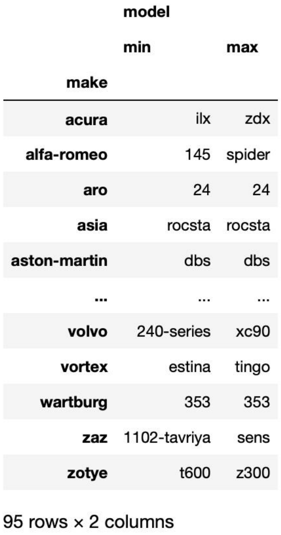


		priceUSD						
		sum	min	max	mean	median	std	var
make	model							
acura	ilx	29400	13500	15900	14700.000000	14700.0	1697.056275	2.880000e+06
	legend	2999	2999	2999	2999.000000	2999.0	NaN	NaN
	mdx	484321	4650	31900	14244.735294	11999.5	8759.442525	7.672783e+07
	rdx	156950	9900	22750	14268.181818	14500.0	3816.429793	1.456514e+07
	rl	22869	6750	8319	7623.000000	7800.0	799.335349	6.389370e+05
...	...	...	...	...	...	...	...	...
zaz	lanos	6800	3100	3700	3400.000000	3400.0	424.264069	1.800000e+05
	pick-up	1900	1900	1900	1900.000000	1900.0	NaN	NaN
	sens	5800	1390	2510	1933.333333	1900.0	560.743554	3.144333e+05
zotye	t600	15000	15000	15000	15000.000000	15000.0	NaN	NaN
	z300	12749	3000	5750	4249.666667	3999.0	1392.031010	1.937750e+06

1069 rows × 7 columns



```
make   model
acura  ilx          2
       legend       1
       mdx         34
       rdx         11
       rl           3
                   ..
zaz    lanos        2
       pick-up      1
       sens         3
zotye  t600         1
       z300         3
Length: 1069, dtype: int64
```



	model	
	min	max
make		
acura	ilx	zdx
alfa-romeo	145	spider
aro	24	24
asia	rocsta	rocsta
aston-martin	dbs	dbs
...	...	...
volvo	240-series	xc90
vortex	estina	tingo
wartburg	353	353
zaz	1102-tavriya	sens
zotye	t600	z300

95 rows × 2 columns

First, I group priceUSD into different brands, and then calculate the sum, mean, maximum, minimum, variance, median, and std of each group.

Secondly, I divide the priceUSD into groups according to different models, and divide the models into groups according to different brands, and also calculate the sum, mean, median, maximum, minimum, variance, std of each group.

Conclusion:

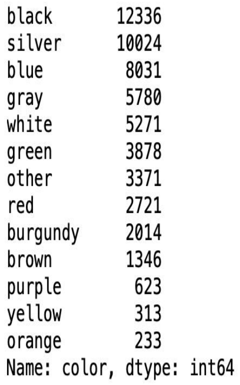
①Learn about the different brands of this group of used cars, the models of each brand, and the number of cars in each model.

②Among used cars, the brand with the largest number of cars and the model with the largest number of cars, as well as the least, as well as the total price of all cars of each brand and model, the average price of the car, the most expensive car, the cheapest car.

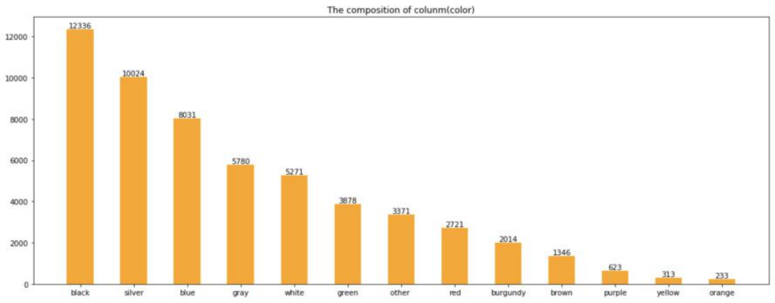
③For example, zotya, this brand has two models, T600 and Z300, and the number of these two models is 1 and 3 respectively.

The model with the largest number of cars in Zotya is Z300, and the smallest is T600.

3.2.7 color analysis



```
black       12336
silver      10024
blue         8031
gray         5780
white        5271
green        3878
other        3371
red          2721
burgundy     2014
brown        1346
purple        623
yellow        313
orange        233
Name: color, dtype: int64
```

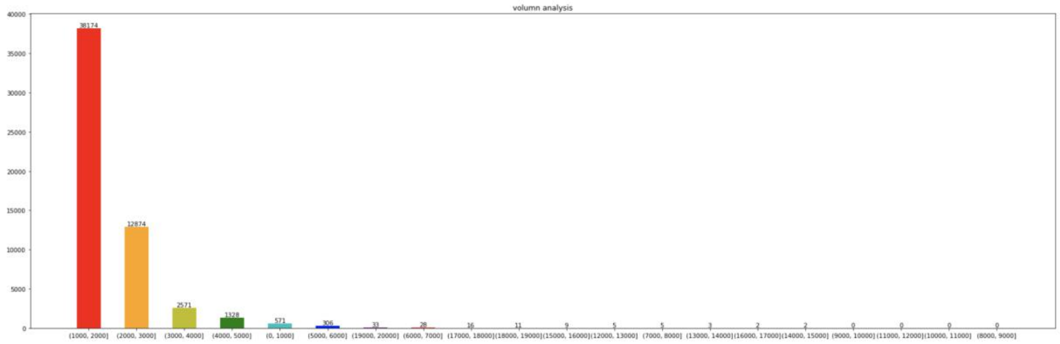


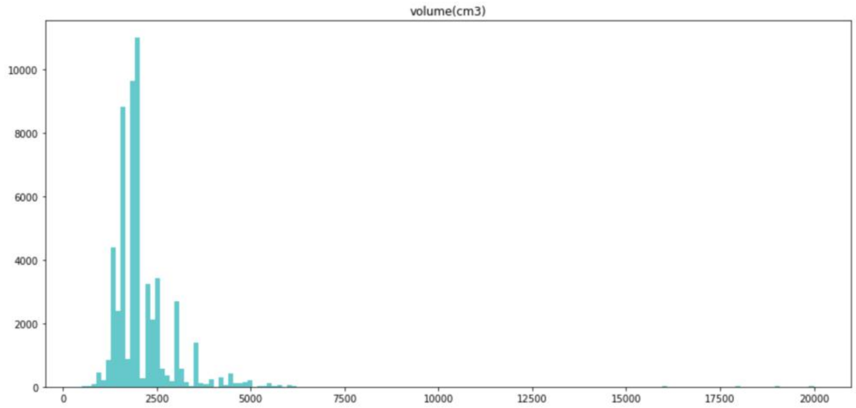
I counte how many colors there are in used cars and how many cars each color has and plot a bar chart to show.

Conclusion:

①The bar chart clearly shows the type of color and the number of cars in each color. The number one is black used cars with 12,334, the least is orange cars with only 233.

3.2.8 volumn analysis





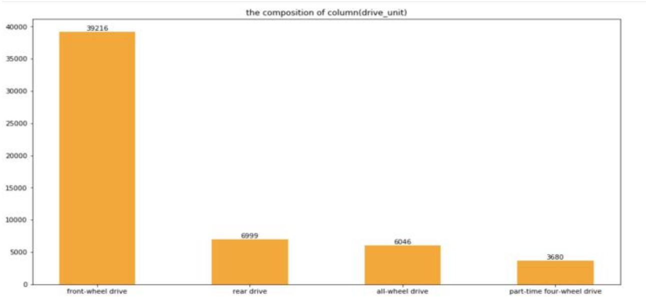
I divide Volumn into several groups, and then calculate the number of cars in each group respectively, as shown in the bar diagram. Plot a histogram of volumn data.

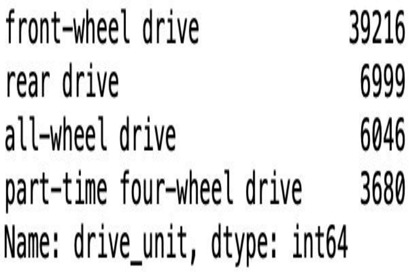
Conclusion

①In the histogram, it can be seen that most of the data are concentrated in the range of 0-5000cm3, which can indicate that the volume of most second-hand cars is within this range, and only a small part of the volume is not in this range.

②For example, the interval with the largest number of vehicles is 1000-2000 cubic centimeters, with 38174 vehicles.

3.2.9 Drive-unit analysis





```
front-wheel drive             39216
rear drive                     6999
all-wheel drive                6046
part-time four-wheel drive     3680
Name: drive_unit, dtype: int64
```

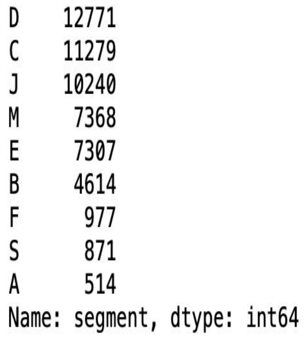
Plot a bar chart to show the composition of Drive-unit .

Conclusion:

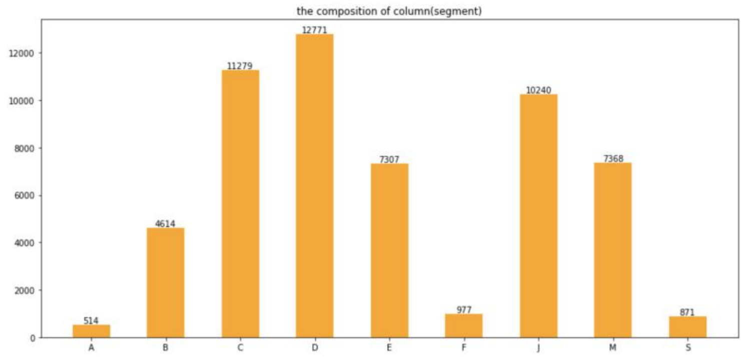
① There are four types of drive_unit, namely front-wheel drive, rear drive, all-wheel drive, and part-time four-wheel drive. The number of used cars in these four categories are 39216, 6,999, 6,046, and 3680, respectively.

②The largest number of used cars in this batch is front-wheel dive(39216), and the least is part-time four-wheel drive(3680).

3.2.10 Segment analysis



```
D    12771
C    11279
J    10240
M     7368
E     7307
B     4614
F      977
S      871
A      514
Name: segment, dtype: int64
```



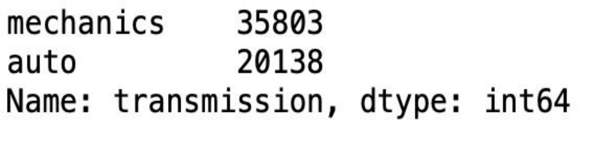
I count that there are 9 types of segments, and get the number of vehicles in each type. Plot a bar chart to show the details.

Conclusion:

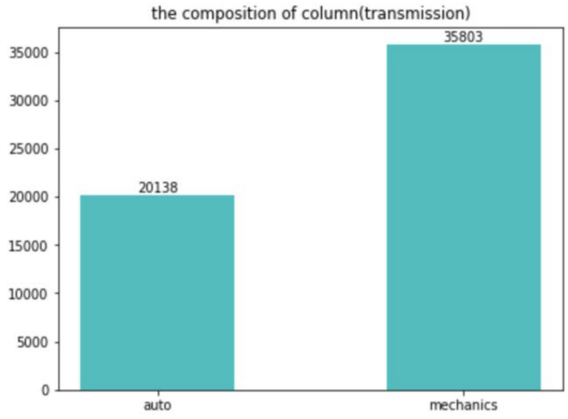
① There are 12,771 second-hand cars with segment D, which is the most among all categories.

②The least-numbered category is A, with only 514 used cars.

3.2.11 Transmission analysis



```
mechanics     35803
auto          20138
Name: transmission, dtype: int64
```

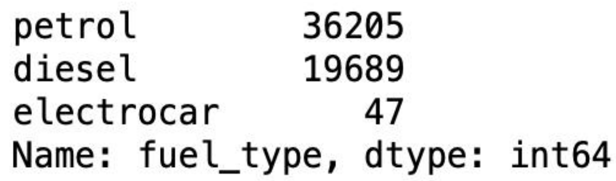


Visualizing transmission data by bar chart.

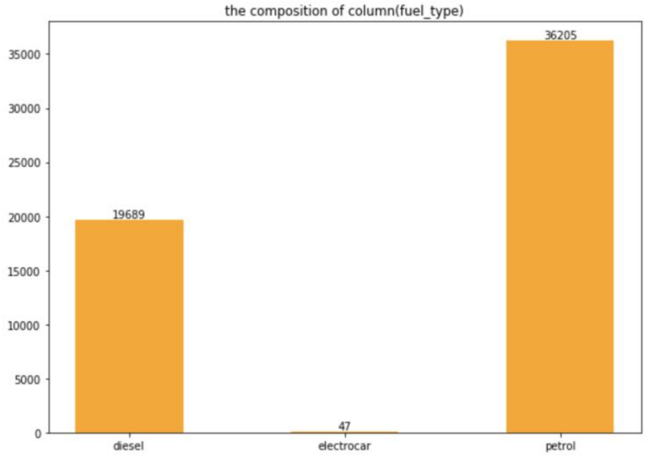
Conclusion:

① There are only two types of data in the transmission column, namely auto and mechanical, with 25138 and 35803 used cars respectively.

3.2.12 Fuel_type analysis



```
petrol        36205
diesel        19689
electrocar       47
Name: fuel_type, dtype: int64
```



Showing the composition of fuel_type by bar chart.

Conclusion:

① Similarly, fuel_unit has only three types of data, petrol, diesel, and electrocar. Among the used cars, there are 36,205 petrol, 19,689 diesel and 47 electrocar.

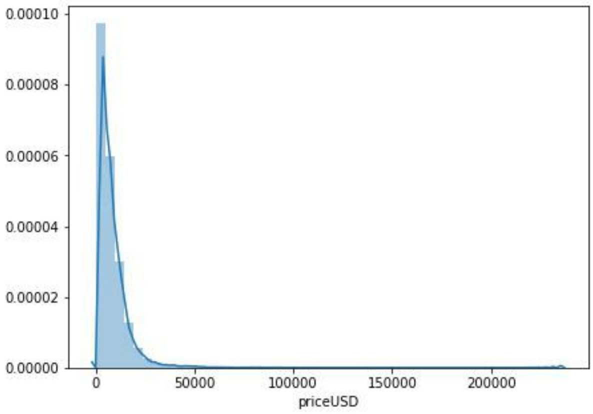
4. Multiple nonlinear regression

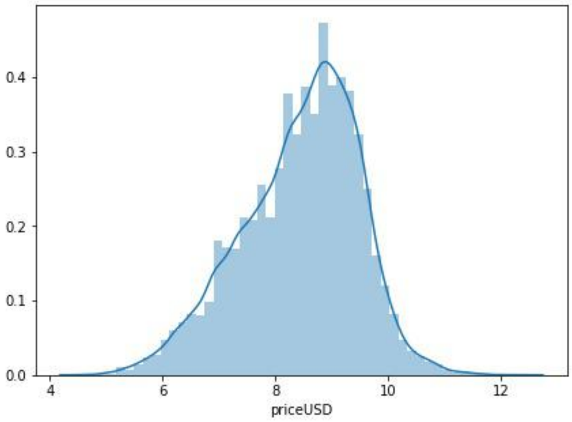
4.1 Preprocess the variables

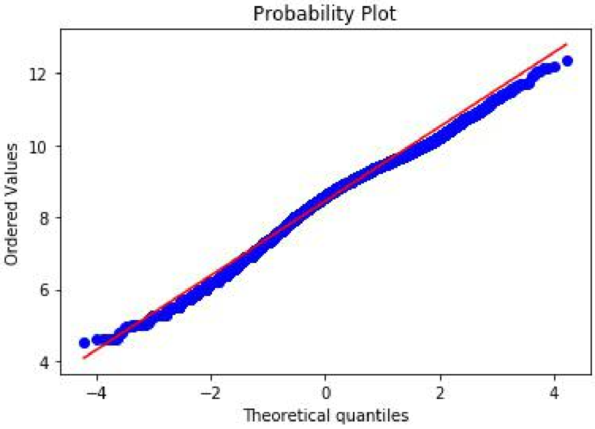
In this data set, we have 12 variables, in which ‘price (USD)’ is the dependent variable y, and ‘year’, ‘mileage (kilometers)’, ‘make’ , etc. are independent variables. I notice that the ‘make’, ‘model’, ‘color’ have a lot of categories, which will make the model become complex and difficult. Therefore, in the consideration of the simplicity of the model, I decide to delete these three variables.

I use the histogram to show the distribution of the price. It is obvious that it is not a normal distribution. However, one of the assumptions of regression model is that dependent variable Y must follow normal distribution. So I need to transform the y into lny. After transformation，the new variable follow the normal distribution roughly.

And I also use the probability plot to observe it. In the probability plot, the points form a straight line roughly, which can also help me get the same conclusion.







4.2 Figure out the correlation between the variables

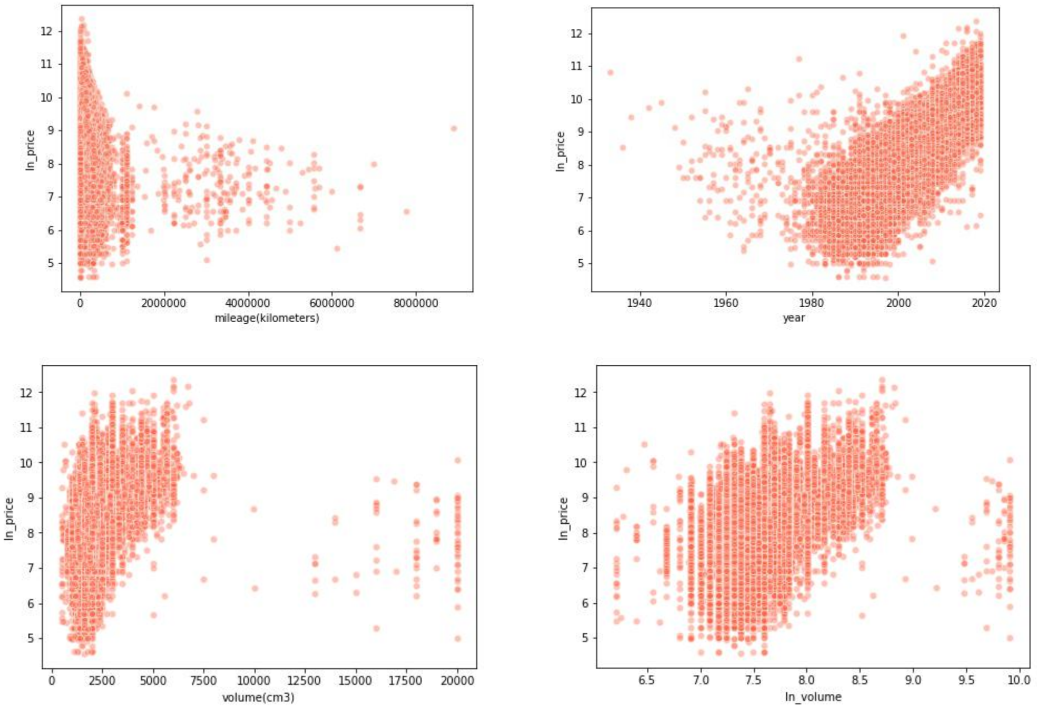
4.2.1 Use graphs to visualize the data

I need to select the independent variables by means of their correlation with y. At first, I use the graphs to visualize the data, so that I can have a sketchy understanding of the correlations. As for the numerical variables, I use scatter plot to show the relationship. Following are the conclusion I made from the figures:

① There is no apparent relationship between mileage and ln(price).

② The year and ln(price) are positively correlated. Besides, I find a outlier, whose production year is less than 1920. Considering the fact that this point may affect the performance of the model, I decide to delete it.

③ There is no apparent relationship between volume and ln(price). But I try to transform the variable, that is, change ‘volume’ into ‘ln(volume)’. Therefore, now there is a positive correlation between them, and the correlation coefficient is higher, rising from 0.25 to 0.33.



As for the categorical variables, I use box plots to visualize them. Following are the conclusions:

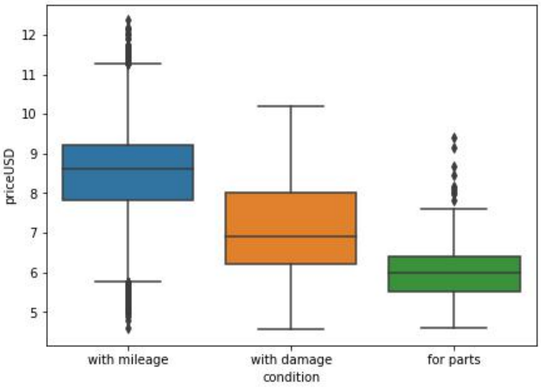
① In general, The price of the used car with mileage is higher than the price of the cars with damage and for parts.

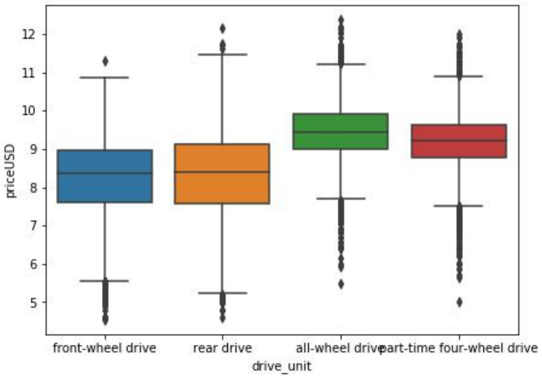
② The price of electrocar is higher than the cars with petrol and diesel

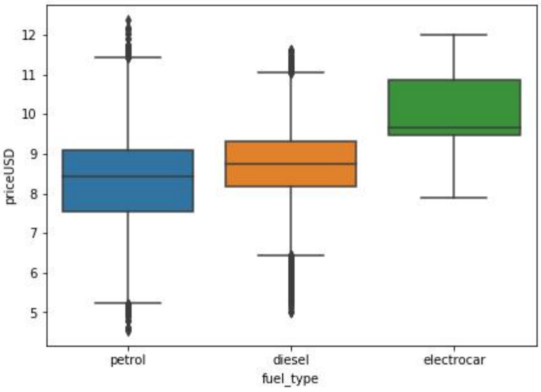
③ Automatic cars are more expensive than mechanical cars

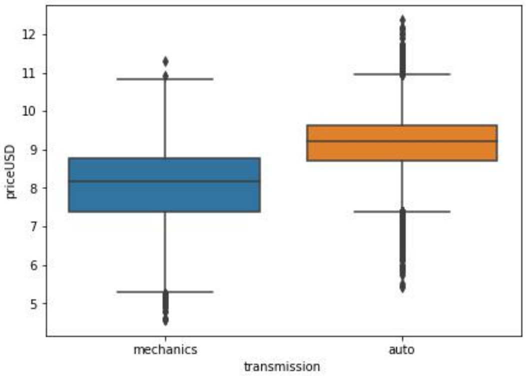
④ Among the various drive unit cars, the used cars with all wheel drive are the most expensive one.

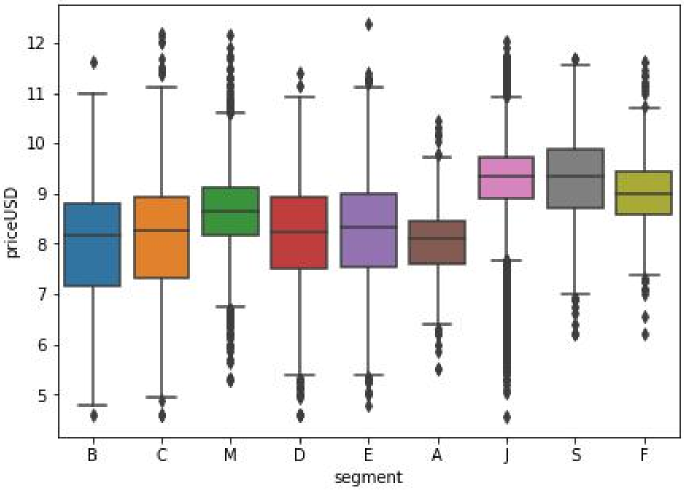
⑤ As for the segment and price, there is no apparent relationship among them. Hence, later when I select the variables, I need to consider whether to delete it or not.





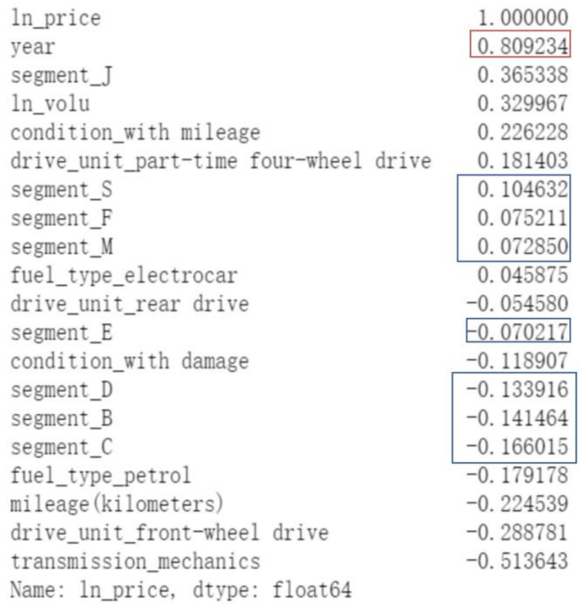




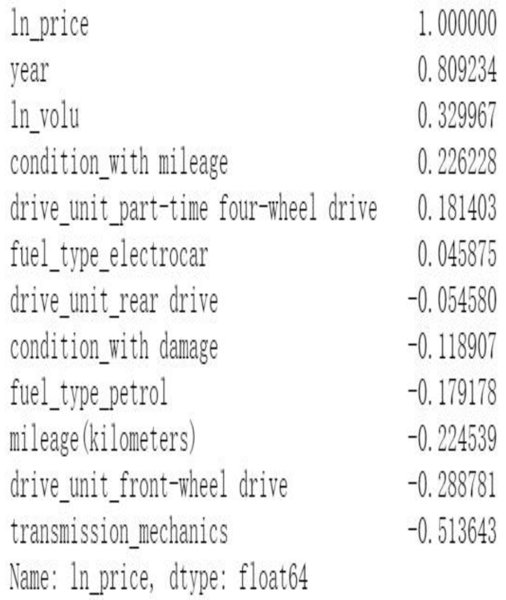


4.2.2 calculate the correlation coefficients.

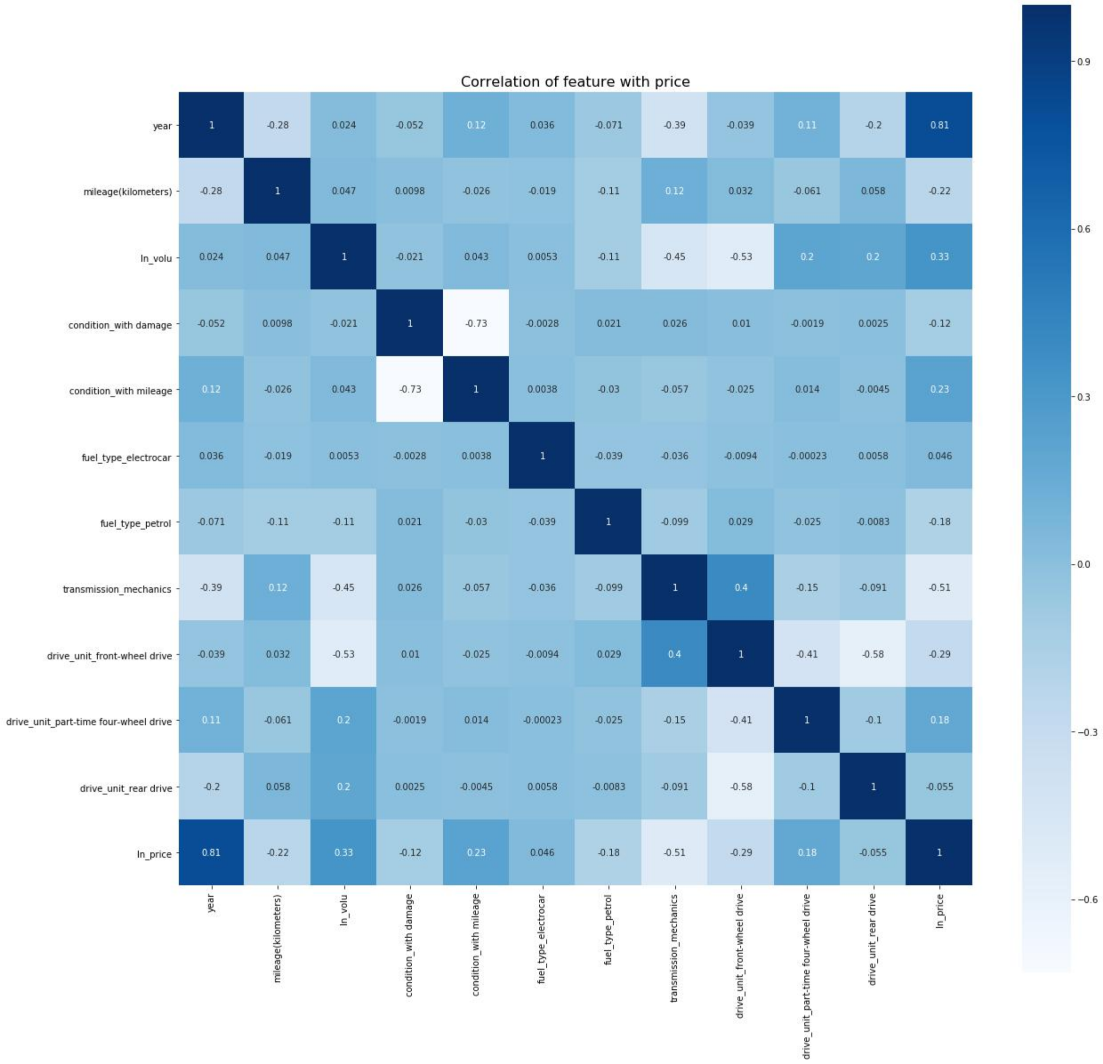
After getting a general idea of their relation, I calculate the correlation coefficients. Before doing so, I need to transform the categorical variables into dummy variables. I use 0 and 1 to represent the different categories. If each variable has n categories, n-1 dummy variables will be used. Take ‘condition’ as an example, this variable has 3 kind of categories, so I will introduce 2 dummy variables, ‘condition_with damage’ and ‘condition_with mileage’. And ‘condition_for parts’ is taken as the benchmark. I use the 'get_dummies' function to transform the variables. Finally, I have 8 dummy variables. Then I use the ‘corr’ function to calculate the correlation coefficients, so that I can have a exact understanding of the correlation between the variables. Through the values, I find that the price and year has strong correlation. However, the correlation between the price and the segment is week. Considering the fact that segment has many categories, which may complicated the model, I decide to delete it. Besides, I use the heat map to get more information. Through the heat map, I find that the correlation between the independent variables are not so strong, meaning that the colinearity problem of the model is not so serious.



```
ln_price                                1.000000
year                                    0.809234
segment_J                               0.365338
ln_volu                                 0.329967
condition_with mileage                  0.226228
drive_unit_part-time four-wheel drive   0.181403
segment_S                               0.104632
segment_F                               0.075211
segment_M                               0.072850
fuel_type_electrocar                    0.045875
drive_unit_rear drive                  -0.054580
segment_E                              -0.070217
condition_with damage                  -0.118907
segment_D                              -0.133916
segment_B                              -0.141464
segment_C                              -0.166015
fuel_type_petrol                       -0.179178
mileage(kilometers)                    -0.224539
drive_unit_front-wheel drive           -0.288781
transmission_mechanics                 -0.513643
Name: ln_price, dtype: float64
```



```
ln_price                                1.000000
year                                    0.809234
ln_volu                                 0.329967
condition_with mileage                  0.226228
drive_unit_part-time four-wheel drive   0.181403
fuel_type_electrocar                    0.045875
drive_unit_rear drive                  -0.054580
condition_with damage                  -0.118907
fuel_type_petrol                       -0.179178
mileage(kilometers)                    -0.224539
drive_unit_front-wheel drive           -0.288781
transmission_mechanics                 -0.513643
Name: ln_price, dtype: float64
```



4.3 Do regression analysis

4.3.1 variables selection

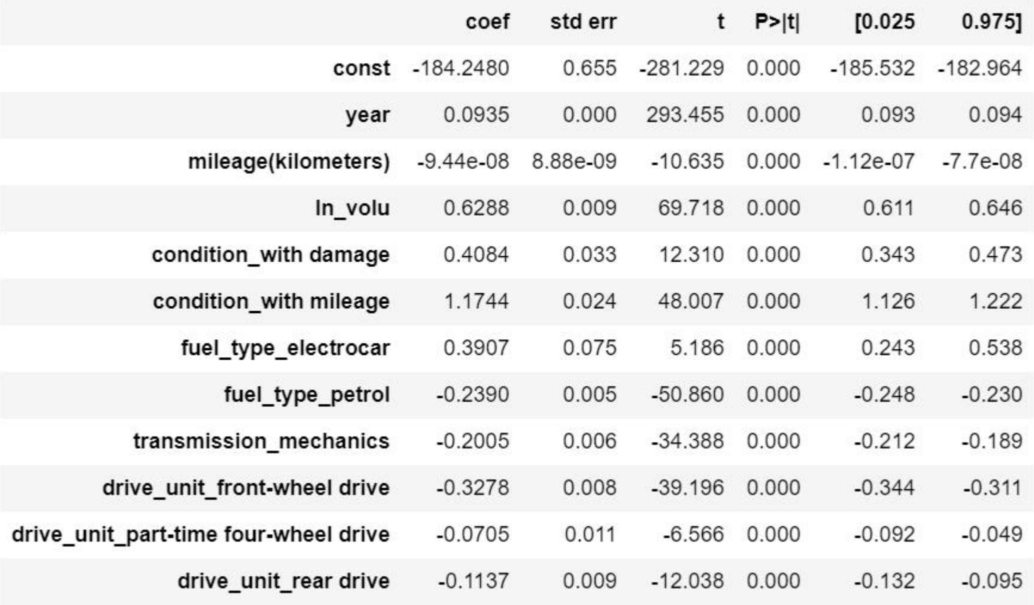
Independent variables:
year, mileage(kilometers), ln(volume), condition_with damage, condition_with mileage, fuel type_electrocar, fuel type_petrol, transmission_mechanics, drive unit_front-wheel drive, drive unit_part-time four-wheel drive, drive unit_rear drive
Dependent variable: ln(price)
(Note: Considering the complexity and difficulty of the model, here I don't take the interaction and drift of the variables into account**)**

4.3.2 split data set into train data set and test data set

Before doing the regression, I divide the data set into train data set and test data set. I use train data to train the model, and use test data to evaluate the performance of the model, which can help improve the model. I use 85% of the data as train data, and 15% of the data for testing. And I use 'train_test_split' function from the sklearn package to divide the data into X_train, X_test, y_train and y_test randomly.

4.3.3 do regression analysis

I use the 'LinearRegression' package and ‘statsmodels.api’ package to do the regression. And the outcome is:
lny=-184.2480+0.0935X1-9.44*10^-8X2+0.6288X3+0.4084D1+1.1744D2+0.3907D3-0.239D4-0.2005D5-0.3278D6-0.0705D7-0.1137D8
(Note: In order to express conveniently, I use the symbols to represent the variables: X1--year, X2--mileage, X3--ln(volume), D1--condition_with damage, D2--condition_with mileage, D3--fuel type_electrocar, D4--fuel type_petrol, D5--transmission_mechanics, D6--drive unit_front-wheel drive, D7--drive unit_part-time four-wheel drive, D8--drive unit_rear drive.)



	coef	std err	t	P>|t|	[0.025	0.975]
const	-184.2480	0.655	-281.229	0.000	-185.532	-182.964
year	0.0935	0.000	293.455	0.000	0.093	0.094
mileage(kilometers)	-9.44e-08	8.88e-09	-10.635	0.000	-1.12e-07	-7.7e-08
ln_volu	0.6288	0.009	69.718	0.000	0.611	0.646
condition_with damage	0.4084	0.033	12.310	0.000	0.343	0.473
condition_with mileage	1.1744	0.024	48.007	0.000	1.126	1.222
fuel_type_electrocar	0.3907	0.075	5.186	0.000	0.243	0.538
fuel_type_petrol	-0.2390	0.005	-50.860	0.000	-0.248	-0.230
transmission_mechanics	-0.2005	0.006	-34.388	0.000	-0.212	-0.189
drive_unit_front-wheel drive	-0.3278	0.008	-39.196	0.000	-0.344	-0.311
drive_unit_part-time four-wheel drive	-0.0705	0.011	-6.566	0.000	-0.092	-0.049
drive_unit_rear drive	-0.1137	0.009	-12.038	0.000	-0.132	-0.095

4.3.4 Use various metrics to evaluate the performance of the model

R^2: 0.7961

Adjusted R^2: 0.796

MAE: 0.3278

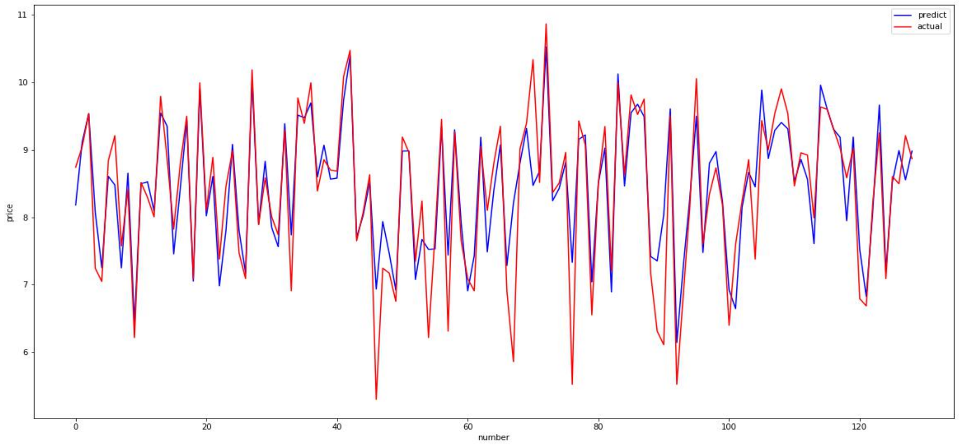
MSE: 0.2413

RMSE: 0.4912

The R^2 is about 80%, which means that the independent variables explain 80% of the variation in Y. And the adjusted R^2 is also about 80%. R^2 will increase when the model introduce new variables, while adjusted R^2 will take the number of the variables into account, which can be a better metrics. And I notice that the P values of the F-statistic and t-statistic are approximately 0, which means that all of the independent variables has great influence on the price. Besides, I use the ‘metrics’ function to calculate the mean absolute error, mean square error and root mean square error , which can give me an idea of the error between predicted values and actual values. For instance, MAE means that the difference between the predicted lny and actual lny is about 0.33 on average. Thus, an conclusion can be made that the fitting effect of the model is good.

4.3.5 Use graph to visualize the predicted value and Actual value.

I also try to use the line chart to show the difference between the predicted values and actual values. Because of the large amount of the data, I just select a small part of it, so the graph can be clear. The red line denotes the actual values and the blue one denotes the predicted values. I find that this two lines are very close.



4.3.6 Make a conclusion

According to the multiple nonlinear model, I draw some conclusions:

① Without considering other factors, as the volume increase 1%, the price of the used cars increase 0.63% .

② As the production year of the used cars increase 1 unit, the price increase 9.35%.

③ The difference of the price between a used car with mechanics transmission and one with auto transmission is $0.20.

④ A petrol car is $0.24 cheaper than a diesel car, while a electrocar is $0.39 more expensive than a diesel car.

5. Random Forest Regression

5.1 Pre-processing data

Above all, I randomly divided data into 85% training and 15% testing data using ‘train_test_split’ function in Scikit-Learn. Then, except for the price that needs to be predicted, I had to transform some categorical data such as color, model name, fuel type, etc. into numerical data. Thus, I decided to use ‘LabelEncoder’ transformer on training sets. Specifically, what I was trying to do is if I have two different kinds of color, for instance, red and green. Red can be transformed into 0 while green will be transformed into 1.

5.2 Random Forest Regressor

With the data parsed and some initial insights to guide us, I applied our random forest regressor. After evaluating, the results indicated that root mean squared error was about 1073 dollars and r square score was really high that up to 98%. At this point, I guessed our model was overfitting since I just tried the default parameter.

5.3 Improvement

Algorithm	MSE	Standard Deviation
Random Forest	2938.1941	198.0764
Linear Regression	5955.4046	183.5885

Using k-fold cross validation to verify my conjecture, it was found that contrary to linear regressor, random forest had a lower mean squared error with approximately 3000 gap, which was successful. However, it had a higher standard deviation(198.1), namely, higher volatility. So I supposed our model may have overfitting.

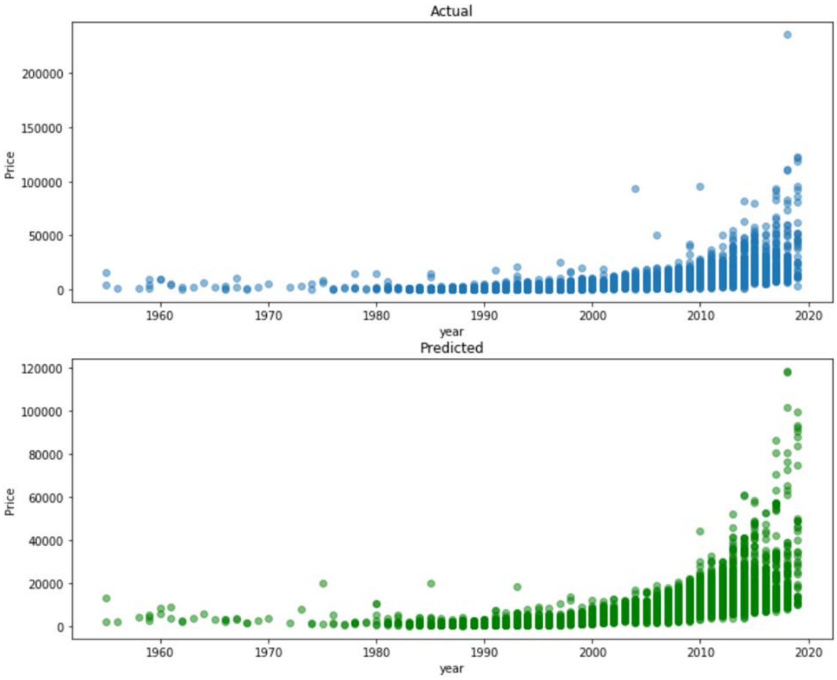
5.4 Adjust parameter

max_features	n_estimators	MSE
2	3	3474.73
2	10	3017.97
2	30	2869.28
4	3	3287.87
4	10	2968.54
4	30	2818.35
6	3	3264.66
6	10	3002.91
6	30	2836.17
8	3	3429.53
8	10	3032.67
8	30	2841.67

The high variance means that I may enable to improve on the algorithm if I change different parameters applied to compare mean squared error, so grid search is a good method to adjust main parameters: n_estimators and max_features. After computing the best hyperparameter, I get 4 was our max_features and n_estimators should be set as thirty. Technically, look at the score of each combination above: 2818.35 was the smallest error in this combination. Recall what I evaluated earlier, new MSE was better than 2938.19.

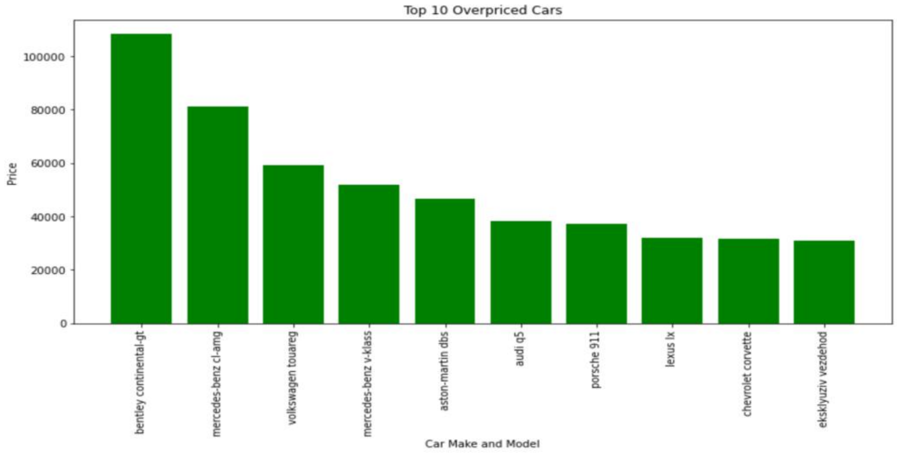
5.5 Results

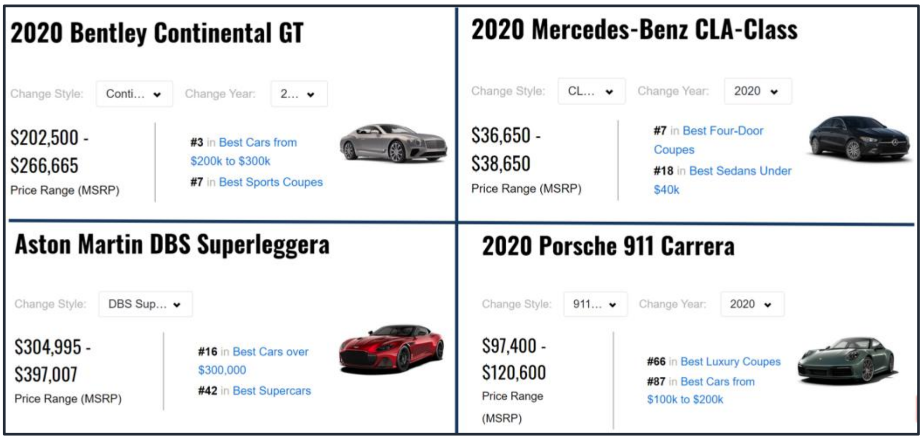
Finally, I tested our testing sets with random forest regression. It produced a prediction with a high r square score of 86%. Although it was not as good as I did before. To some degree, our model was therefore avoiding the overfitting problem.



While visualizing the actual price and predicted price on scatter plot, it also displayed a high similarity.

5.6 Application





2020 Bentley Continental GT

Change Style: Conti... Change Year: 2...

$202,500 - $266,665
Price Range (MSRP)

#3 in Best Cars from $200k to $300k
#7 in Best Sports Coupes

2020 Mercedes-Benz CLA-Class

Change Style: CL... Change Year: 2020

$36,650 - $38,650
Price Range (MSRP)

#7 in Best Four-Door Coupes
#18 in Best Sedans Under $40k

Aston Martin DBS Superleggera

Change Style: DBS Sup...

$304,995 - $397,007
Price Range (MSRP)

#16 in Best Cars over $300,000
#42 in Best Supercars

2020 Porsche 911 Carrera

Change Style: 911... Change Year: 2020

$97,400 - $120,600
Price Range (MSRP)

#66 in Best Luxury Coupes
#87 in Best Cars from $100k to $200k

Next, if I sorted the difference between the actual price and predicted price, I got top 10 overpriced cars. The results did not disappointed. Indeed, the highest-priced cars are those that are popular or ridiculously expensive since everyone wants to buy but cannot afford.

5.7 Suggestions

In general, it is highly recommended that second-hand car dealers are supposed to take advantage of predicted market prices to develop different marketing strategies. Furthermore, the price of overpriced cars should be controlled. As for buyers, be a rational consumer, making sure the car is good value for money.

6. Conclusion

To sum up, our project completed the fundamental features description, using pie charts, bar plots, histograms, etc. Thereby, made a careful analysis of the different features and their relationships.

Algorithm	RMSE	R2
Multiple Nonlinear	4237.96	0.7961
Random Forest	3165.92	0.8662

Moreover, comparing with multiple nonlinear and random forest regressor, I get 4237.96 and 3165.92 root mean square error respectively. The r square metric also reveals random forest model appears better. It is not only because they have different modeling method but also RandomForest applied all independent variables while our multiple nonlinear regressor applied only important features.

As a result, management is able to understand how exactly the prices vary with various features of used cars. They can manipulate the marketing strategy accordingly so as to establish the best price. In addition, this project will be useful for managers to know the pricing dynamics of a new market.

References

[1] Pasedko, S. (2020, February 04). Belarus Used Cars Prices. Retrieved August 15, 2020, from https://www.kaggle.com/slavapasedko/belarus-used-cars-prices

[2] 2020 Bentley Continental GT: Review, Trims, Specs, Price, New Interior Features, Exterior Design, and Specifications. (n.d.). Retrieved August 15, 2020, from https://carbuzz.com/cars/bentley/continental-gt

[3] 2020 Mercedes-Benz CLA-Class: Review, Trims, Specs, Price, New Interior Features, Exterior Design, and Specifications. (n.d.). Retrieved August 15, 2020, from https://carbuzz.com/cars/mercedes-benz/cla

[4] Aston Martin DBS Superleggera: Review, Trims, Specs, Price, New Interior Features, Exterior Design, and Specifications. CarBuzz. https://carbuzz.com/cars/aston-martin/dbs-superleggera.

[5] 2020 Porsche 911 Carrera: Review, Trims, Specs, Price, New Interior Features, Exterior Design, and Specifications. CarBuzz. https://carbuzz.com/cars/porsche/911-carrera/2020.

Note: The codes are attached in the package. In ‘data cleaning and preprocessing’ part, we import ‘cars.csv’, while in other parts, we import ‘cars.data cleaning.csv’.

Work Allocation

Data cleaning and preprocessing: Jie Huang

Fundamental features description: Taohe Zhan and Nianqing Chen

Multiple nonlinear regression: Jie Huang

Random Forest Regression: Liuyi Pan and Nianqing Chen

Conclusion: Liuyi Pan

Coordinating and integrating: Jie Huang