Data Analytics Final Project

Used Car Price Analysis in Germany

Ci Chen, Fu-Chi Shih, Naixin Li

Project Overview

Idea and Motivation

Our project focuses on the used car market in Germany. We act as a used car dealer in Germany and analyze the most recent data on ebay to do a basic analysis on used car price and factors that have significant influence on it. Overall, we hope this analysis can help us to make a better assessment on used car prices when we buy and sell used cars.

Over the past few decades, the used car market experienced significant growth worldwide. Households are given incentives to buy used cars over new cars as market regulation improves. As a used car dealer, we seek to maximize the profit for each used car. We understand that mileage, registration date and some other factors have impact on car prices. But in this project, we seek to have several detailed data-driven strategies in order to improve our current performance.

Data Description

We selected the used cars database from Kaggle. This dataset contains over 370,000 German used cars information on Ebay.[link: https://www.kaggle.com/orgesleka/used-cars-database]

To clean the data, we deleted observations by following three guidelines: unreliable data, insufficient data amount, and invalid data format. Unreliable data consists of observations that the team does not think would reflect the real market. An example of this would be when an observation for a used car price is zero. Insufficient data amount includes some categories with a small dataset. For instance, the “electric” category under feature “fuelType” only contains 80 observations, which is 0.022% of the entire dataset. Because of this, we do not believe this small category could provide enough observations to build the accurate model. Lastly, we delete the observations that contain empty or NaN information. (See the python notebook for more details on data cleaning.)

After the data cleaning process, there were a total of 126,869 useable observations to build the model. The descriptions for each feature are listed below:

- vehicleType: kleinwagen (compact car) or limousine (mid-size family car)
- gearbox: automatik (automatic transmission) or manuell (manual transmission)
- powerPS: horsepower
- kilometer: distance travelled by the vehicle

- fuelType: benzin(petrol), diesel, lpg(liquefied petroleum gas)
- Brand: the brand for the vehicle (there are 25 different brands in our dataset)
- notRepairedDamage: whether the vehicle has unrepaired damage. (nein:no, ja:yes)
- age: the age of the vehicle (2016 - year of registration)

Model Description

Model Fitting and Evaluation

Our team first divided the entire dataset into three categories - 80% of the shuffled data to the train dataset, 10% goes to the validation set, and the remaining 10% goes to test set. We then fit a linear regression model on the train data and selected the final model which best estimated the parameters and minimized error on the validation set. **Exhibit 1** presents the summary of the model. The model has a decent R-square value of 0.698, which means the data generally fit the regression line. Each chosen feature has a relatively low "p-value", with a few brands as an exception. These low "p-values" show that it is unlikely the model predicts by chance alone. The train data has a 3190 RMS error while the validation set has a 3335 RMS error. The validation set has a 0.83 correlation between the predicted and real price. Both RMS error and correlation data indicated that there were no occurrences of overfitting in the model.. With this positive summary of the model, our team was confident to move forward.

Model Assessment

Finally, our team computed the test error of our model to approximate the generalization error. The RMS error is 3240, meaning our prediction deviates from the actual price $3,240 on average. The correlation of predicted price and actual price is 0.84, which indicates our model has an exceptional prediction.

Model Conclusion

Figure 1 through **Figure 7** shows the relationship between each significant factor and price. Age and kilometer are negatively correlated with price, while horsepower has a positive coefficient with price. In conclusion, cars with automatic transmission, diesel fuel type, and zero unrepaired damage can be sold at higher price than others.

Strategic Suggestion

Based on our team's analysis of the dataset and conclusion, we found that multiple factors have significant influence on used car prices. Factors such as mileage and age proved play a large part in our current assessment strategy. However, to better improve the strategy according to our data analysis, we made the following two recommendations.

1. **Repair the damage of the used car or not**

One option to increase the selling price of the used car is to consider repairing any existing damages. However prior to doing so, it is necessary to take into consideration the repair price required, approximate the future sale price and assess whether the repairs will increase or reduce profit. Our model presents a clear representation and assists in the decision making of whether or not to repair damages on a used car. The coefficient of variable C(notRepairedDamage) [T.nein] is 1221, which indicates that having the damage repaired before selling will increase the value by €1221 on average. Thus, we can choose to repair the damage before selling if the cost is less than €1221. Otherwise, we will choose not to repair prior to a detailed assessment of the damage.

2. **Brand Analysis**

After analyzing the coefficients for vehicle types, we summarize five "most profitable" brands in used cars. We interpreted profitable brand as holding everything else as constants, the brand has the biggest influence on used car price. These five brands are Mini (with coefficient 3214), Audi (with coefficient 3139), BMW (with coefficient 2688), Volkswagen (with coefficient 2557) and Mercedes_benz (with coefficient 2331). To improve our performance, our team plans to consider individual's budgets into the model and ask employees to close more deals within these brands.

Exhibit 1. Model Summary

Dep. Variable:	price	**R-squared:**	0.698
Model:	OLS	**Adj. R-squared:**	0.698
Method:	Least Squares	**F-statistic:**	7337.
Date:	Fri, 13 Jan 2017	**Prob (F-statistic):**	0.00
Time:	21:16:46	**Log-Likelihood:**	-9.6285e+05
No. Observations:	101495	**AIC:**	1.926e+06
Df Residuals:	101462	**BIC:**	1.926e+06
Df Model:	32		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	7691.1845	138.105	55.691	0.000	7420.500 7961.869
C(vehicleType)[T.limousine]	-163.6415	27.417	-5.969	0.000	-217.378 -109.905
C(gearbox)[T.manuell]	-889.7086	31.332	-28.396	0.000	-951.119 -828.298
C(brand)[T.audi]	3138.7249	124.043	25.304	0.000	2895.602 3381.848
C(brand)[T.bmw]	2687.8781	122.639	21.917	0.000	2447.507 2928.249
C(brand)[T.chevrolet]	-950.9002	187.415	-5.074	0.000	-1318.232 -583.569
C(brand)[T.citroen]	842.7063	147.101	5.729	0.000	554.389 1131.023
C(brand)[T.fiat]	1031.0137	131.651	7.831	0.000	772.979 1289.048
C(brand)[T.ford]	939.3870	125.764	7.469	0.000	692.891 1185.883
C(brand)[T.honda]	1599.5255	158.680	10.080	0.000	1288.514 1910.537
C(brand)[T.hyundai]	-188.3102	153.537	-1.226	0.220	-489.241 112.620
C(brand)[T.kia]	107.3039	171.595	0.625	0.532	-229.019 443.627
C(brand)[T.mazda]	919.7889	149.508	6.152	0.000	626.755 1212.823
C(brand)[T.mercedes_benz]	2331.2363	124.311	18.753	0.000	2087.588 2574.885
C(brand)[T.mini]	3213.5175	146.658	21.912	0.000	2926.071 3500.965
C(brand)[T.mitsubishi]	871.8095	161.445	5.400	0.000	555.380 1188.239
C(brand)[T.nissan]	1589.3875	143.709	11.060	0.000	1307.721 1871.054
C(brand)[T.opel]	1467.0998	123.201	11.908	0.000	1225.627 1708.572
C(brand)[T.peugeot]	839.7699	132.628	6.332	0.000	579.821 1099.719
C(brand)[T.renault]	1089.5544	128.290	8.493	0.000	838.107 1341.002
C(brand)[T.seat]	1433.5488	132.737	10.800	0.000	1173.387 1693.711
C(brand)[T.skoda]	1047.2151	148.613	7.047	0.000	755.936 1338.494
C(brand)[T.smart]	-120.5080	145.517	-0.828	0.408	-405.719 164.704
C(brand)[T.suzuki]	523.0790	169.579	3.085	0.002	190.707 855.451
C(brand)[T.toyota]	920.9734	143.933	6.399	0.000	638.867 1203.079
C(brand)[T.volkswagen]	2557.2414	121.017	21.131	0.000	2320.051 2794.432
C(brand)[T.volvo]	1723.8983	193.160	8.925	0.000	1345.307 2102.490
C(fuelType)[T.diesel]	1904.5421	27.654	68.871	0.000	1850.341 1958.743
C(fuelType)[T.lpg]	-926.0823	91.453	-10.126	0.000	-1105.328 -746.836
C(notRepairedDamage)[T.nein]	1221.6171	33.238	36.753	0.000	1156.470 1286.764
powerPS	43.1884	0.272	158.645	0.000	42.655 43.722
age	-268.6588	2.056	-130.668	0.000	-272.689 -264.629
kilometer	-0.0534	0.000	-180.124	0.000	-0.054 -0.053

Exhibit 2. Visualization

Figure 1 and Figure 2 show that: everything else being equal, if the age of a car is older, the price is lower. However, some antique cars may have high value due to its unique value.

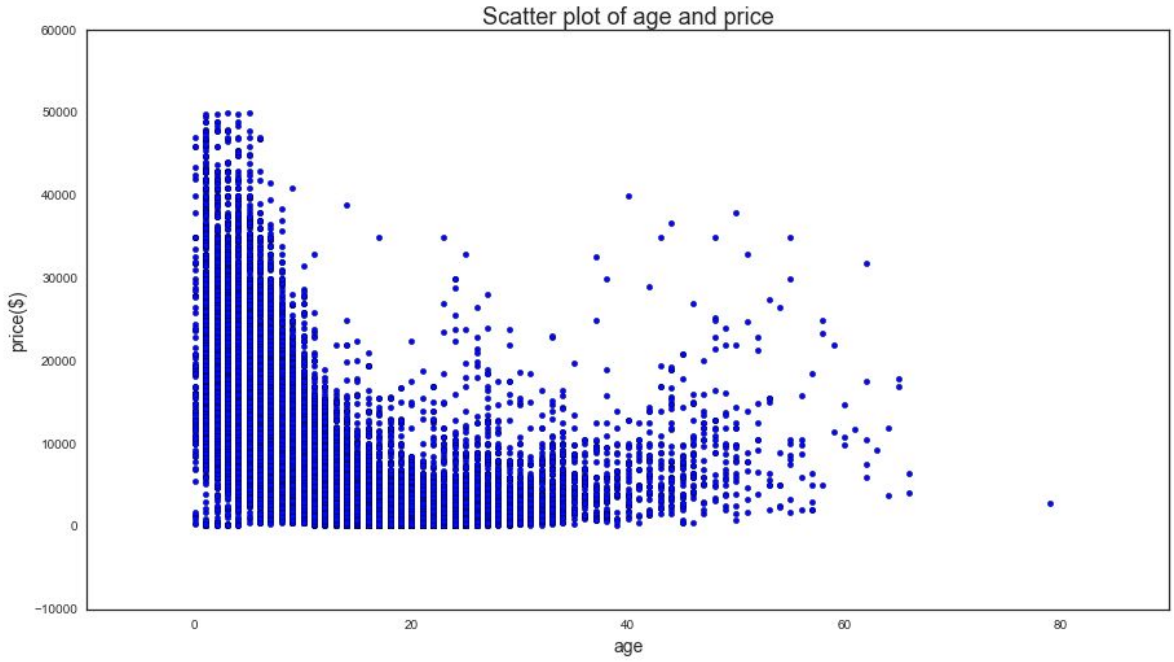

Figure 1. Scatter plot between age and price.

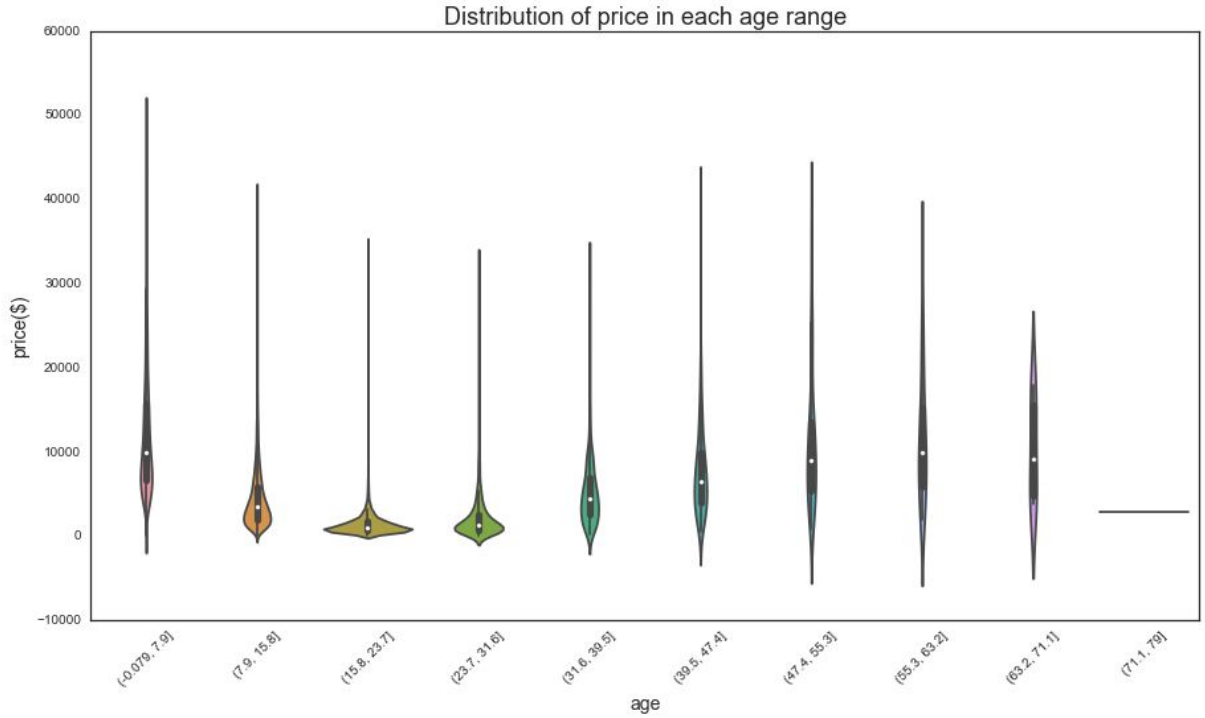

Figure 2. Violin plot between age and price.

Generally speaking, if the vehicle milage is higher, the price is generally lower.

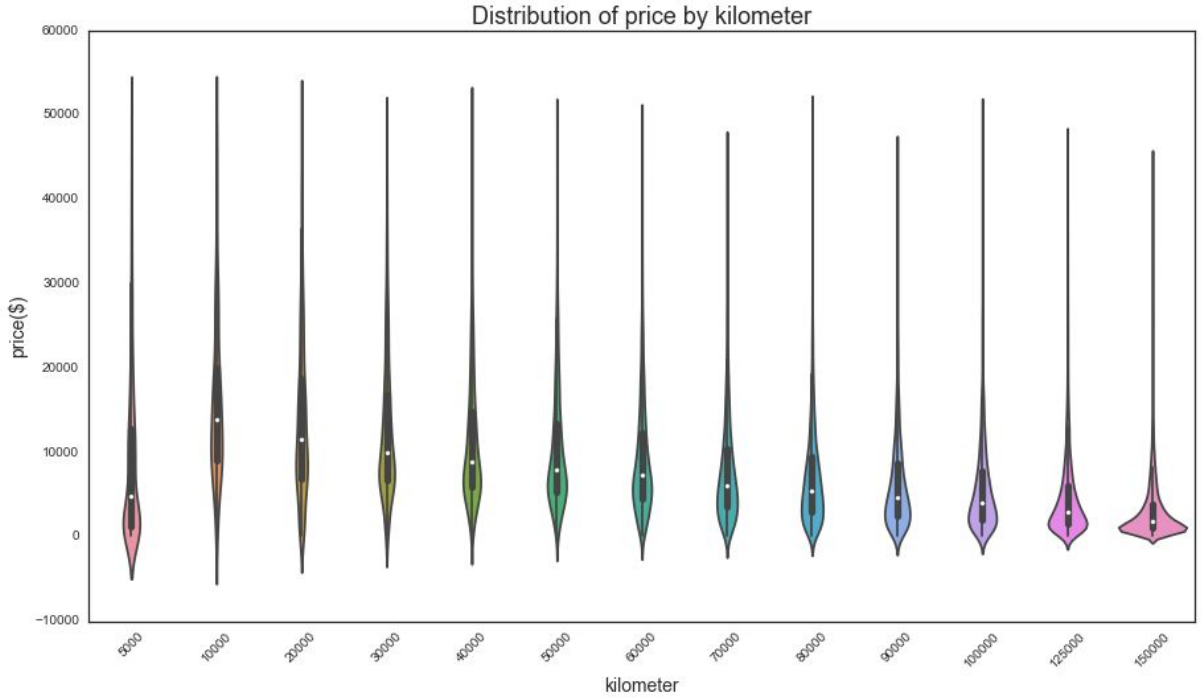

Figure 3. Violin plot between kilometer and price.

Generally speaking, a vehicle with automatic transmission can be sold at higher price.

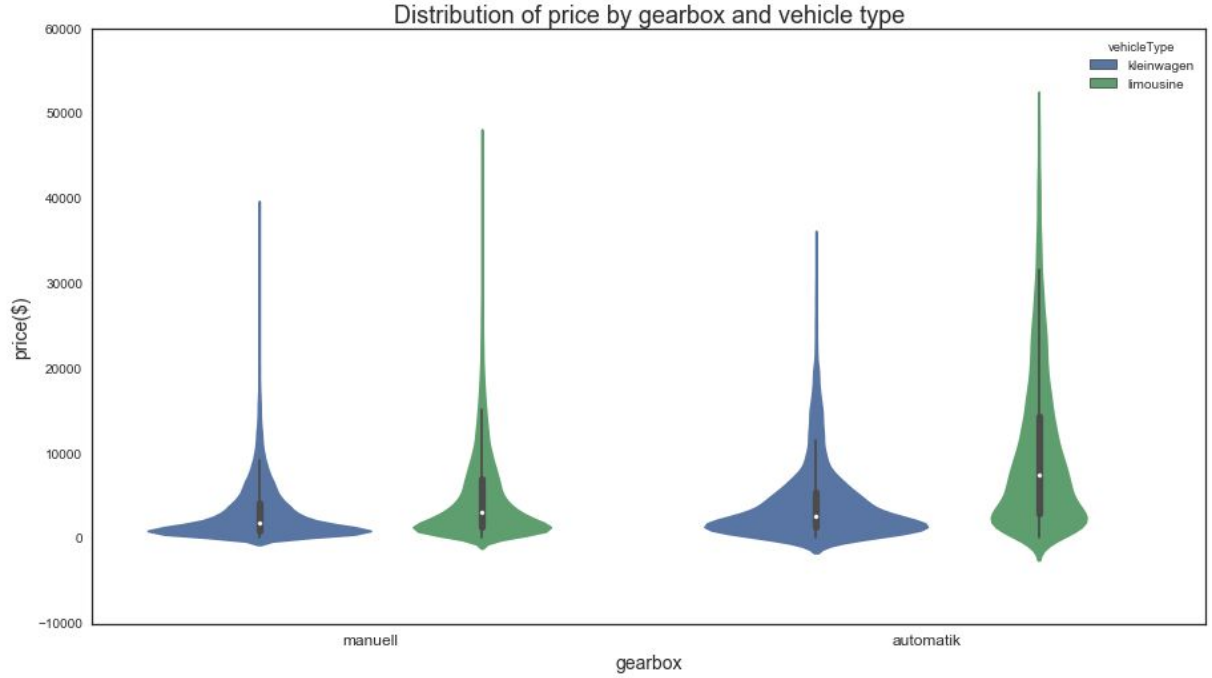

Figure 4. Violin plot between gearbox and price.

Generally speaking, a vehicle with fuel type diesel can be sold at higher price.

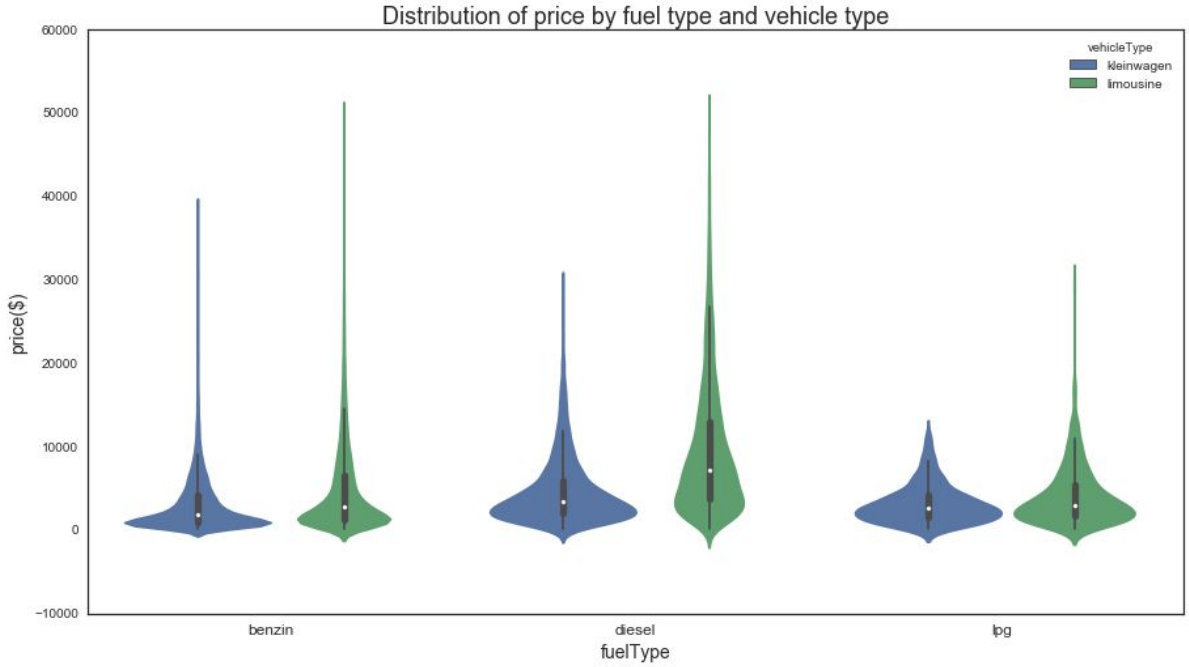

Figure 5. Violet plot between fuelType and price.

Generally speaking, a vehicle with higher horsepower can be sold at higher price.

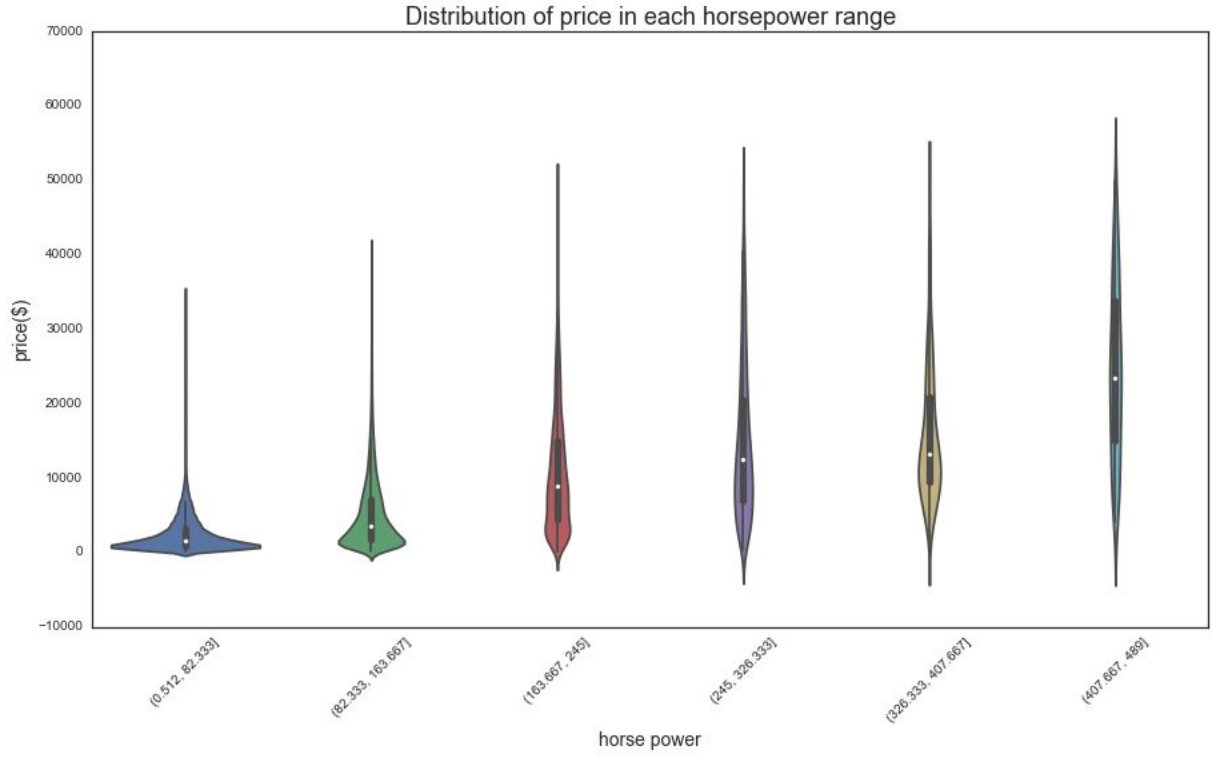

Figure 6.Violet plot between horsepower and price.

Generally speaking, a vehicle without “notRepairedDamage” can be sold at higher price.

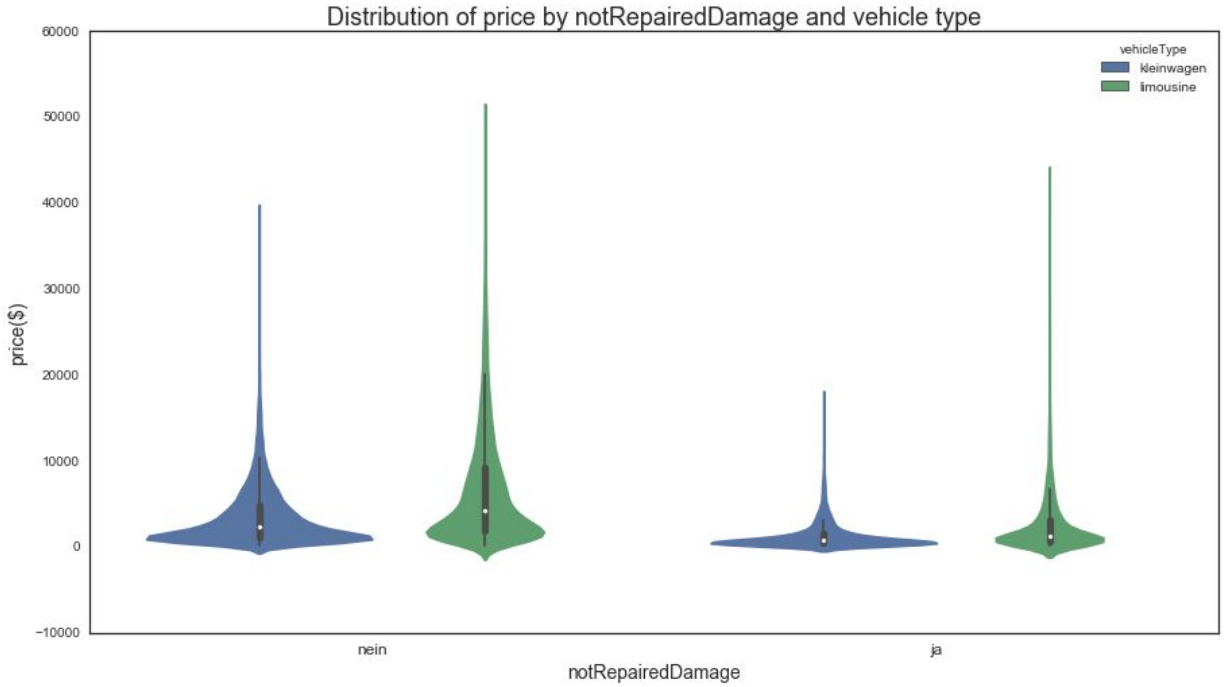

Figure 7.Violet plot between notRepairedDamage and price.