Used Car Dataset in 2019

Jash Jasani  
eBay Germany Classified *(of Affiliation)*  
kaggle.comSykesville, Maryland USA  
jjasan1@students.towson.edu

*Abstract*—By examining the past data of Used Car database in 2019, valuable insights into sale of the car and prediction may be gained by application of data mining techniques. (*Abstract*)

Keywords—csv, encoding, ansi, gearbox, past data, data mining, data analysis, regression (key words)

# Introduction

The Dataset used is found on Kaggle. This report will be helpful to those working in the automotive industry to get a better understanding of the used cars market. The data available is massive that can be enhanced for data analysis and data prediction. The report will focus on some predictions relevant to car availability in the market and its condition. An exploratory data analysis of attributes like dataCreated, lateen, price, notRepairedDamage will bolster the prediction.

# The Dataset

## Description

The data set includes over 300000 used cars used for data analysis is beautifully structured, there are some free text fields, missing values as well. Additionally, as the name phonetic sounds German there is some of the data in German. I was able to translate most terms using free tool Google Translate. Though the name column was a bust, so I decided not to play with it. I believe the column “nrOfPictures” would be disregarded due to very negligible importance.

Note: - For reading the csv file, you will need to set encoding=’ansi’ because some data have special character, ut-8 won’t work in this case.

The data attributes include dataCrawled, dataCreated, lastSeen are of “Date” datatype, price, yearOfRegistration, powerPS, kilometre, monthOfRegistration, nrOfPictures, postalCode are of ‘Integer” datatype, name, seller, offerType, abtest,vehicleType,gearbox, model, fuelType, brand, notRepairedDamage are of “String” datatype. There are 371528 tuples in this dataset with no missing values.

# Explorative data analysis

## Categorical Variable

1) Gearbox

The categorical variable “Gearbox” will tell us about the transmission type of a specific car. The dataset will have either manual or auto as filed values. Thus, there are 77105 “automatic” and “274214 “manual”.

2) vehicleType

The categorical variable “vehicleType” will tell us about the type of car. This helps in predicting which type of car was sold faster than the others.

3) model

The categorical variable “model” indicates the model of car. Based on the analysis of lastSeen, one can predict what kind of model people look for when they buy used cars.

4) fuelType

The categorical variable “fuelType” indicates the type of fuel the car will need. For the prediction, this variable also plays a dynamic role in determining the chances of how soon the car can be sold.

5) notRepairedDamage

The categorical variable “notRepariedDamage” shows if the damage in the car is repairable or not. If it is affirmative, then the price would be lowest. It might be possible that some data might not be present due to the same cause. This all leads to one thing the seller wants to get rid of the car. Thus, in the dataset the author uses negation. The assumption made by me is when the data field is “nien” means the damage is repairable and “ja” means the damage is not repairable.

## Numerical Variables

1) Price

The numerical variable “price” impacts how soon can a car can be sold quicker. Most buyers will see the value over the price, if price of the car is less than the average market value then it is more likely to sell fast but the opposite might not be true. For maximum and minimum price refer to TABLE 2. For Five summaries of all numeric variables refer to TABLE 1.

You would see some data fields in “Price” is 0 which will not be treated as missing value. It would be true to say with price as 0 that car has a salvage title. This observation is made based on the data from two data fields:- ‘price’ with 0 and ‘notRepairedDamage’ with ja. Refer to TABLE 2 for maximum and minimum price of the dataset.

TABLE 1 FIVE NUMBER SUMMARIES OF ALL NUMERICAL VARIABLES

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | price | PowerPs | kilometre | Time\_diff\_hrs |
| Min | 0.00 | 0.00 | 5000 | 0.00 |
| 1st Q | 1.15 | 70.00 | 125000 | 61.00 |
| Median | 2.95 | 105 | 150000 | 156.00 |
| Mean | 1.72 | 115.55 | 125618.7 | 223.35 |
| 3rd Q | 7.20 | 150 | 150000 | 338.00 |
| Max | 2.147 | 20000 | 150000 | 18220.00 |
| Mode | 0.00 | 0.00 | 150000 | 12.0 |
| SD | 3.58 | 192.14 | 40112.34 | 207.96 |

2) Kilometer

The numerical variable “kilometer” has a conditional hammer on sale of the car. Some buyers like to buy the cars with highway mileage, some like to buy with high mileage as a starting car. There are more to this and that is why “mileage” is anvil for selling a car. Refer to TABLE 3 for maximum and minimum value of the dataset. Histogram for kilometres shows that large number of cars have more than 140km odometer reading.

3) powerPS

This numerical variable provides information on the power of the engine. It is quite often mistaken with “horsepower”. The first use was done by German car makers to reflect the infamous automotive work of their engineers.

4) dateCreated

The numerical variable “dataCreated” is entry point in the dataset for any car. From the dataset, on running some query one can see there is no null values. This time variable basically tells you when the car was available in the market.

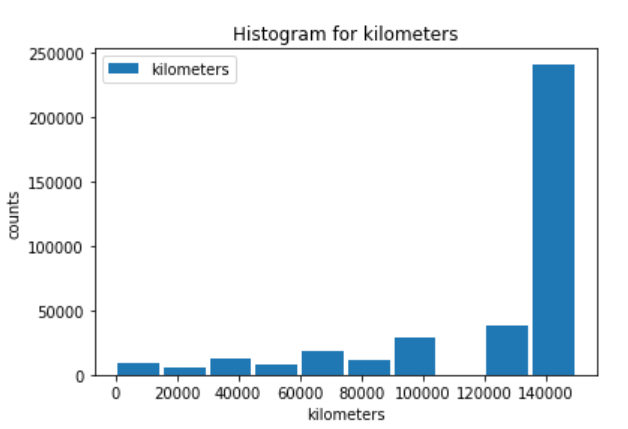
5) lastSeen

The numerical variable “lastSeen” is the last checkpoint before a car is taken off the rails from the market. With pandas null checking methods, it is found to have no empty values. This period attribute will give data analysts insight about the presence of the car on the market.

|  |  |
| --- | --- |
| *TABLE 3* | |
| *Kilometres Margins* | |
| Max | Min |
| 150000 | 5000 |

|  |  |
| --- | --- |
| *TABLE 2* | |
| *Price Margins* | |
| Max | Min |
| 2147483647 | 1 |

FIGURE 1 HISTOGRAM OF KILOMETERS



## Relationship Between Attributes

### dataCreated

This numerical variable not only tells us about the time a car appeared on the market but is an important component in estimating how long a car will be online before it is sold. Refer to TABLE 4 “for time\_on\_market”.

### lastSeen

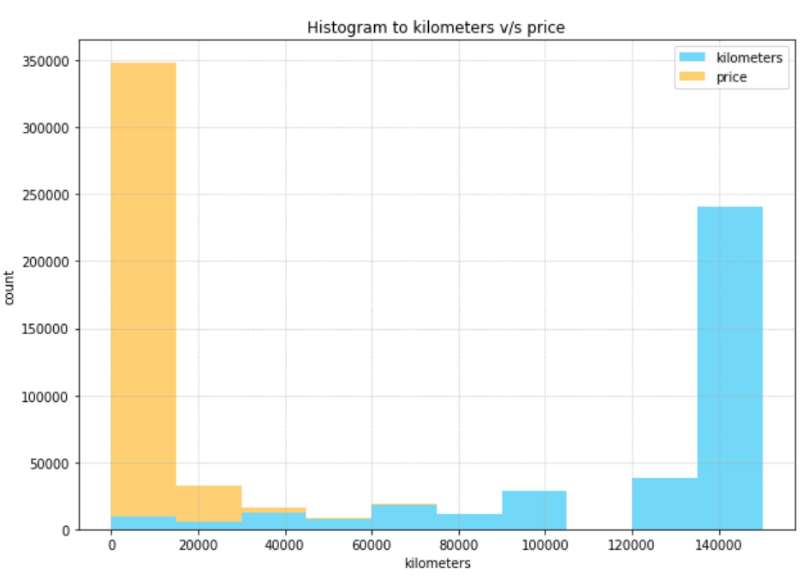
This numerical variable will tell you when the car was last seen online before it get sold and is another pivotal part of the estimation of a car’s online life on ebay. Refer to TABEL 3 “for time\_on\_market”.

### Price

There are two variables taken to elect a salvage title. The first one is ‘Price’. After cleaning the dataset, the values are taken into consideration to determine the value of the car. The prime attention is the zero values. Refer to TABLE 3 that tells us how to decide the title of the car as “salvage”.

To further validate the salvage condition, I have used a histogram of price and kilometre. For most thinkers, an empty space in the histogram will come as a red flag but not for this dataset, the hard to miss insight. In the below histogram, the blank space is proof that there are salvaged cars in the dataset. This bolsters the data analysis of the dataset.

FIGURE 2 HISTOGRAM OF KILOMETERS AND PRICE



### notRepairedDamage

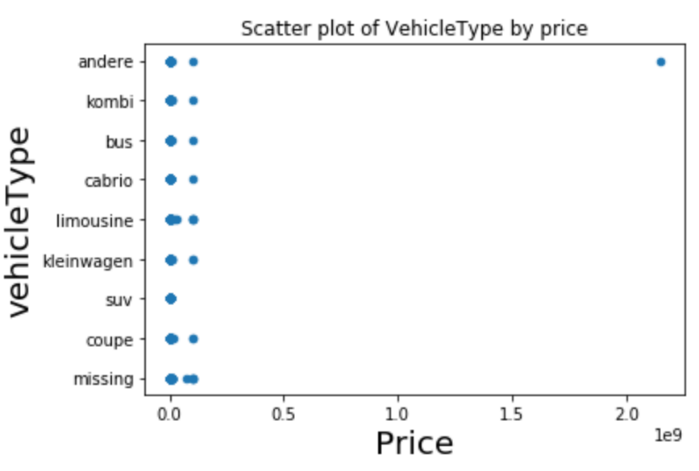
As mentioned, about negation this is the second variable in renouncing a car to be salvage. By the power of negating the “yes” value will sanction a car recovered. Refer to TABLE 3 for salvage title attribute.

|  |  |  |
| --- | --- | --- |
| TABLE 4 | | |
| **Is the car salvaged?** | | |
| *Price* | *notRepairedDamage* | *Salvage Title* |
| *0* | *Yes* | *Yes* |
| *>0* | *No* | *No* |

### vehicleType

This relationship detect outliers in the dataset. It is combined with the price to give a better understanding of the presence of outliers. To achieve this, scatterplot comes to the rescue as we need to define the axes and show the relationship among them. From the scatterplot figure, the other category of vehicle type is found to show diverse data. This points to the assumption that in the dataset there are cars that have been modified to gain price margin. This has an impact on the sale of the car.

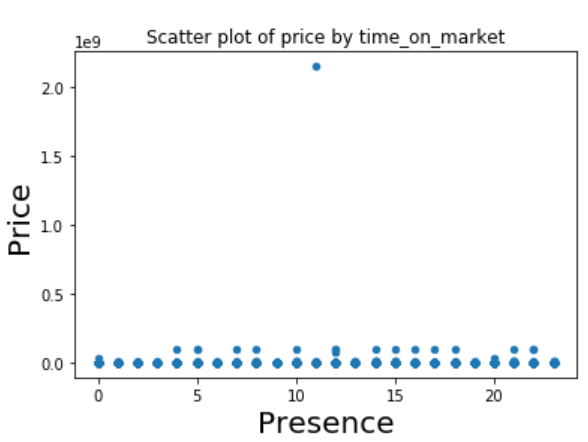
FIGURE 3 SCATTERPLOT OF VEHICLETYPE V/S PRICE



### time\_diff\_hours

This categorical variable is derived by subtracting dateCreated from lastSeen. If the assumption like when the car gets sold its advertisement is generally erased. Thus, we can take difference between lastSeen and dateCreated that will tell us how long it took for the car to be sold. Future insight can be generated by plotting against relevant attributes. For eg:- Scatterplot of price v/s time\_diff\_hours will reflect some light on the relationship. (Refer to Figure 4)

FIGURE 4 SCATTERPLOT OF TIME\_DIFF\_HRS V/S PRICE



# DATA PREPROCESSING

## Normalization

In this dataset, one could consider normalization of one or more attributes depending on the data analysis requirement. The author sightsees normalization of time\_diff\_hours attribute. Decimal Scaling has no effect on skewness, thus it is not taken into consideration.

### Min-Max Normalization

Looking at the data from Table 1, the minimum value was 0, and the maximum was 18220, which gives a skewness of 0.972. When the min-max normalization is applied to the values, the median becomes 0.008 and mean is 0.012, which still directs left-skewed value distribution. By comparing the skewness of non-normalized data with skewness of min-max normalized data, it turns to be equal.

### Z-score Standardization

By applying the z-score standardization, the range of values is -1.07 to 8.653. The skewness of the Z-score values is still 0.972 which proves Z-score standardization has no effect on skewness.

## Transformation

With transformation techniques, we could achieve more normalized behavior of the dataset. To eliminate infinite values, the values with ‘0 hrs’ are converted to ‘1 hrs’. The skewness of non-transformed data is 0.972.

### Natural Log Transformation

With the natural log of time\_diff\_hours variable, we see in Table 5 that Mean and Median are now closer in value 5 respectively. This indicates a more normal distribution of values. The skew for natural log is -0.485, that is almost perfect than original data’s skewness. (closer to 0).

### Square Root Transformation

With this method, we get skewness 0.35 which is better than natural log’s skewness.

### Inverse Sqaure Root Transformation

As the name goes “inverse”, the skewness will also have inversed effect; it will increase to 1.07 from original 0.35.

|  |  |  |  |
| --- | --- | --- | --- |
| TABEL 6 TRANSFORMATION STATISTICAL VALUES OF TIME\_DIFF\_HRS | | | |
|  | *Median* | *Mean* | *Skew* |
| *Natural Log* | 5 | 4.85 | 0.485 |
| *Square Root* | 13.3 | 12.5 | 0.35 |
| *Inverse Square Root* | 0.08 | 0.11 | 1.07 |

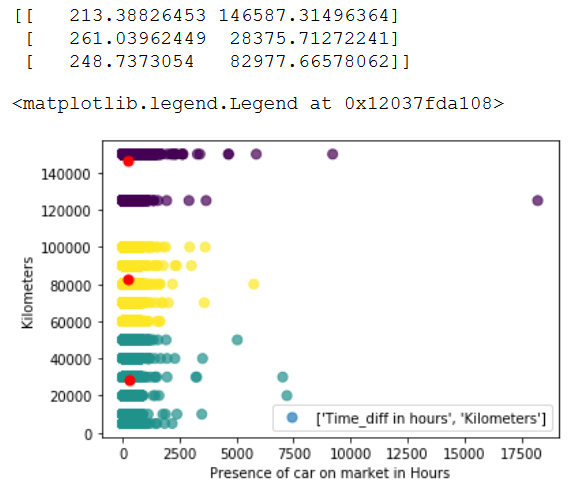
## Binning

As there is a large range of values for time\_diff\_hours in the dataset (Minimum is 0 hrs and Maximum is 18220 hrs), these values could be binned using K-means Clustering. The equal frequency binning technique is omitted here as misleading information is generated due to outliers that is not beneficial for analysis.

### K-Means Clustering

K-means clustering would be the better option when it comes to predict the presence of cat on market based on how many kilometres the car has. From figure 5, the centroids are in list above the scatterplot. Note at the centre of each cluster (in red) represents the mean of all observations that belong to that cluster.

FIGURE 5 SCATTERPLOT OF K-MEANS CLUSTERING OF TIME\_DIFF\_HRS



# Regression Analysis

## Predication Question

Based on data of yearOfregistration of the car we can predict kilometer variable in future. From the dataset, the question rises how accurately we can predict about kilometers gained by a car. Depending on which year the car was registered, the prediction on the kilometers gained can be done using regression analysis.

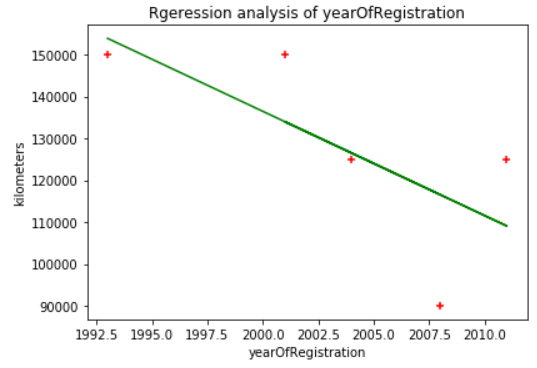
## Regression Model

We will be using yearofRegistration to predict time that any car will take before it gets sold online. The clear assumption made here is kilometer has a direct relationship with yearOfRegistrattion. The equation for the regression line is: -

*kilometer = 2290060.64 + -1080.35 x yearOfRegistration*

For each increase in yearOfRegistration, the kilometer increases by 1080.35. The r2 is 0.492 of this regression that fit linear relationship between kilometer and yearOfRegistration. Therefore, it is not a good predictor (Refer to figure 6). It might be useful to do a multiple regression of yearOfRegistration, kilometer, price or time\_diff\_hours.

FIGURE 6 REGRESSION ANALYSIS OF YEAROFREGISTRATION



# Conclusion and Future Prediction

The time\_diff\_hours provided a good look at the time prediction of cars on the market in 2019. The presence of some outliers was discovered during data distribution. Normalization and Transformation of the time\_diff\_hours has an effect on the skewness of the data, expect inverse square root as the skewness increased under that transformation. The outlier might be needed to be taken care of before doing any further analysis.

The categorical variable “salvage\_title” was added to actually categorized the cars in the dataset using two variables price and notRepairedDamage. The valuable insight derived from this variable was to recognize the car’s title. As on the internet, anyone can be a victim of false information, further analysis can be performed to improve this part of the horizon. This analysis can lead to a better and cleaner dataset for future data analysts.

On the other hand, the numerical variable price might need to makeover due to unexpectedly high and low values which hinders in making accurate predictions. The other most important further investigation might be needed is of vehicle type, brand and vehicle name. The obvious error can occur while entering data into the database by a human or if there is an automatic information gatherer is set up then due to careless verification one can have a faulty dataset. This action will have a tremendous impact on the imminent explanation.

If you look at the yearOfRegistration column, there are some data fields that do not follow normally. To correct this error, it is recommended to filter the rows by a conditional selection of yearOfRegistration from 1885 (the first care ever made) to the present year. This is will eliminate the wrong values so that any prediction made involving yearofRegistration or even any kind of analysis will not deliver false patterns.

Lastly, to detect the scammers one can set up a combination of variables checks depending on the requirement of the locality. The reason it can’t be kept universal is because of the uncertainty of respective geographical domain. The forthcoming prediction would be needing a huge, stable and continuous improvement to keep up the hoodwinks from the sellers.

##### VII References

1. Orges Leka, https://www.kaggle.com/orgesleka/used-cars-database, November 2019. *(reference)*