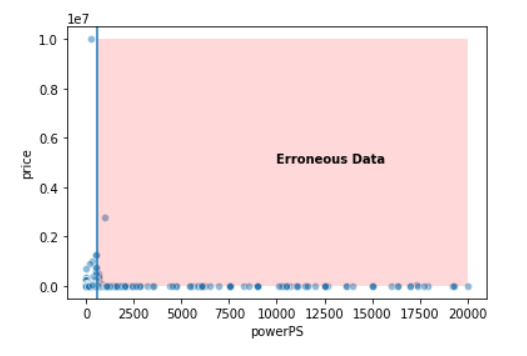
**CAPSTONE DATA STORYTELLING**

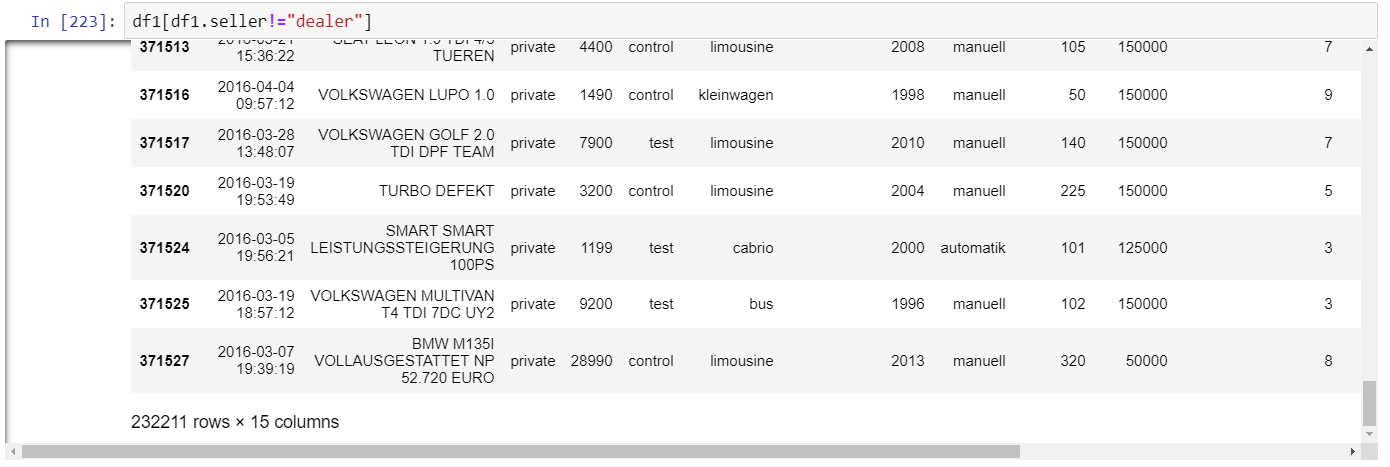
A large data frame consisting of 377 thousand arbitrarily entered car sale advertisements have been processed and cleaned to be used in a value prediction model. Now the dataset looks thinner with 230 thousand data points, but it is sanitized and ready to be used to provide valuable information through data storytelling, EDA and machine learning algorithms.

Having strong confidence of the cleanliness of the data, I started writing code for drawing graphs immediately. The first few graphs drawn immediately revealed I had missed some outliers, as seen on the right, which were removed immediately. 

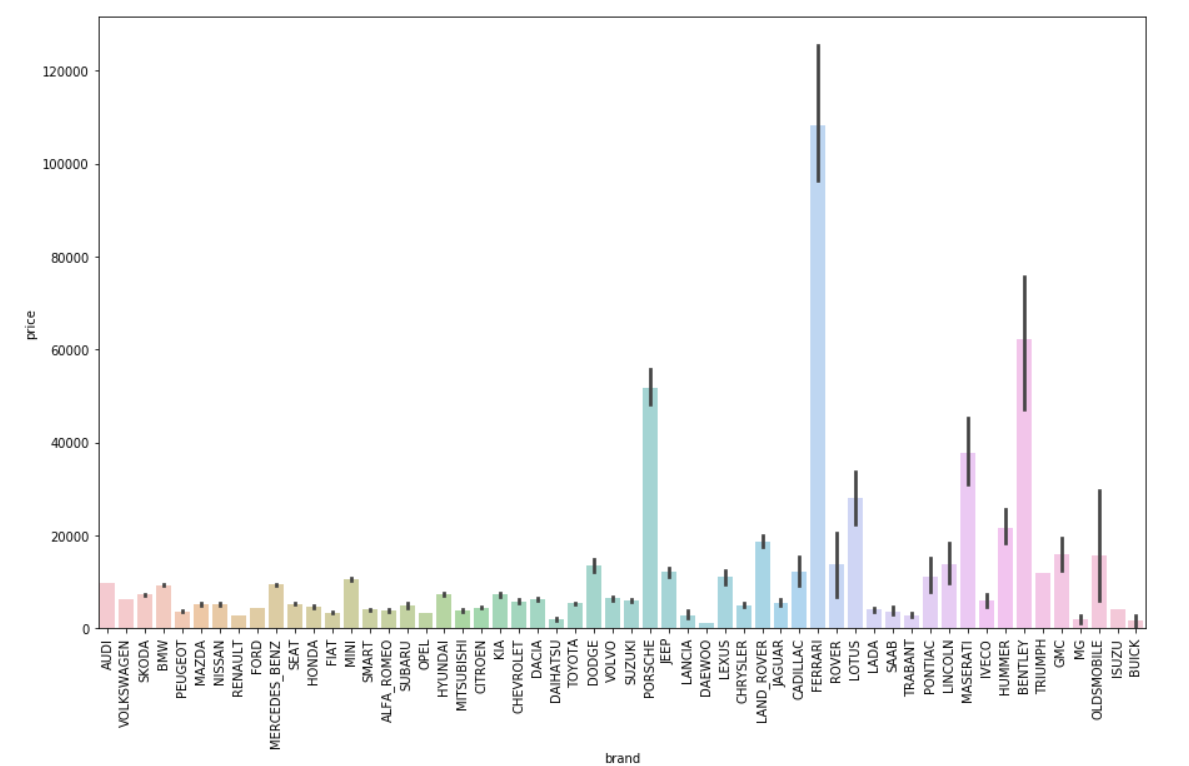
Outliers, once ignored, can cause great problems for basic calculations such as mean, standard deviation and they would have great impact on the correlation coefficients. A prediction model would never work precisely, once error is introduced as input to its mechanism.

After revisiting and cleaning the dataset, I immediately found something interesting to count: of all the ads placed, only 1 out of 233.000 belonged to dealers! It appears like dealers like to advertise as private party sellers. This tendency is very common worldwide, simply because buyers prefer private sellers.

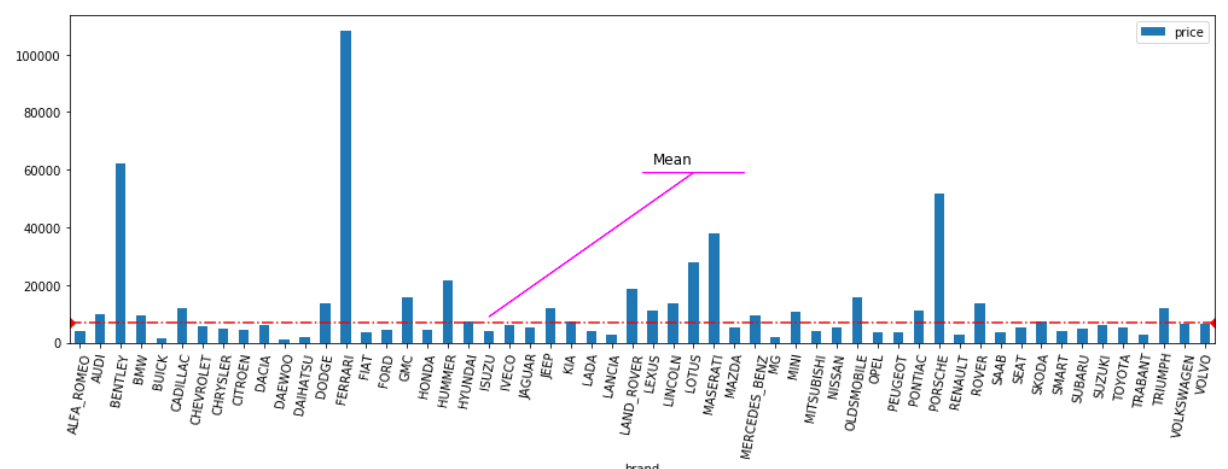




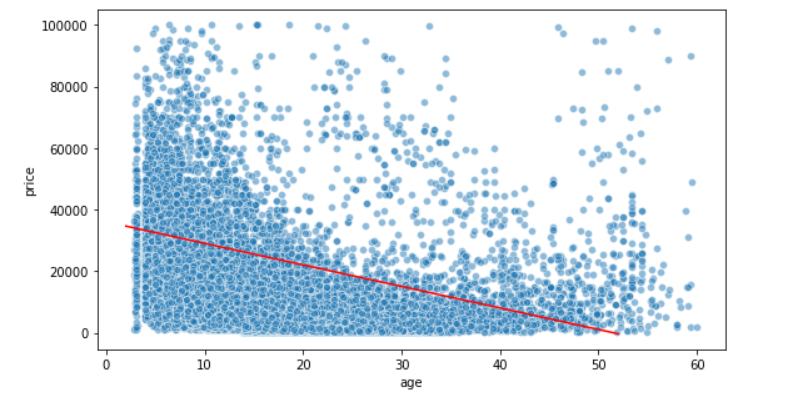
The graph below shows all brands and their mean prices as well as interquartile ranges. Ferrari, Porsche, Bentley and Maserati are the most expensive brands.

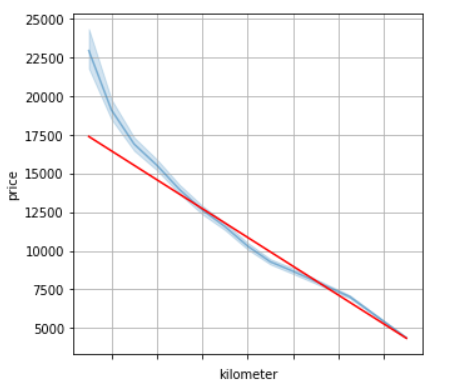


Compared to the database mean, the top three brands are clearly seen to be dwarfing others.

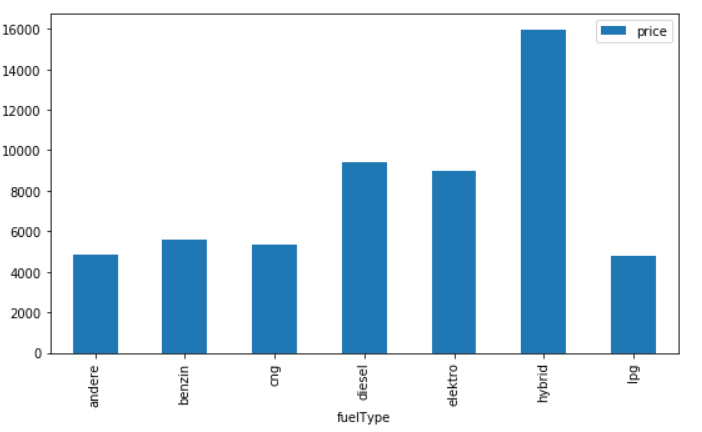
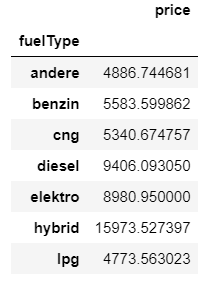


The red line represents the regression, in this (age,price) graph below. The regression line was shifted upwards in order to overlay. The trend is clearly visible, the age and price are negatively correlated.

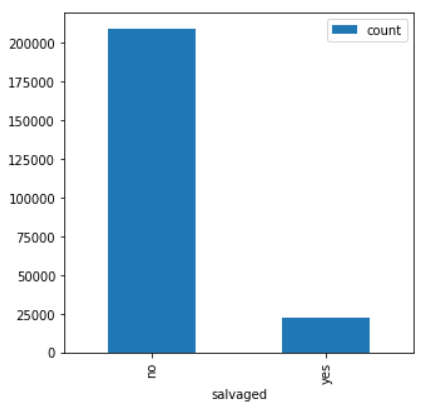
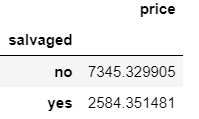
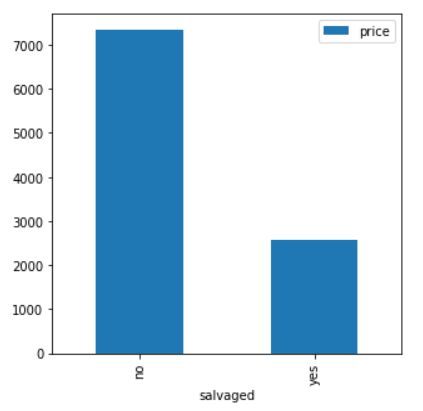




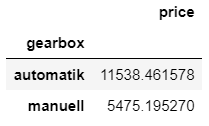
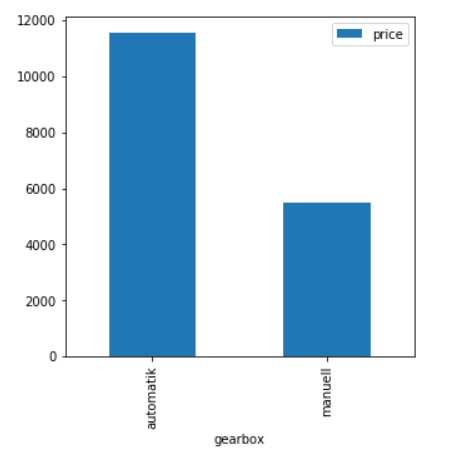
The (kilometer,price) graph on the left reveals that the price is inversely proportional to the mileage. This information is also backed by the regression line, depicted in red, and the negative sign of the Pearson coefficient, which has been calculated as -0.37.

The price according to fuel types have been depicted below. The fact that hybrid vehicles are the most expensive proves to be because of this reason: the only electric cars available at second hand market -thus this database- were “Smart fourtwo” ultra compact city cars that has front row of seats only. Those vehicles are very small, therefore not so expensive. Many brands have hybrid models, that have cutting age technology to curb down fuel consumption, making them considerably higher priced. Diesel vehicles burn less amount of fuel and diesel is considerably cheaper than gasoline in Europe, therefore these cars are preferred over gasoline ones and they are more expensive.

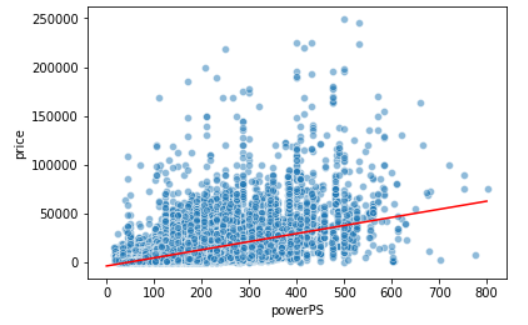


The average price of a salvaged vehicle is almost three times lower than an accident free vehicle, which is quite obvious, since salvaged vehicles are less preferred. The actual wording in the original data frame is not provided, however, I believe it is “nicht unfall frei” which means not accident free.It was originally translated as not repaired damage which I believe is wrong. Some information may be lost in translation, because people tend to sell damaged vehicles, even totaled ones.

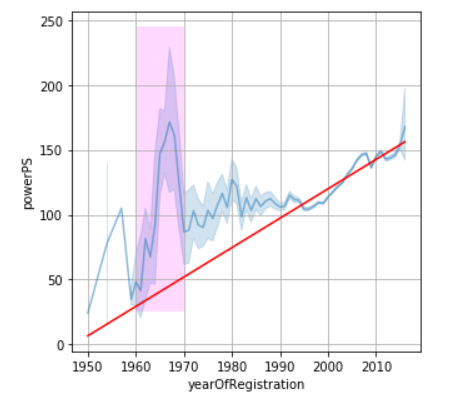


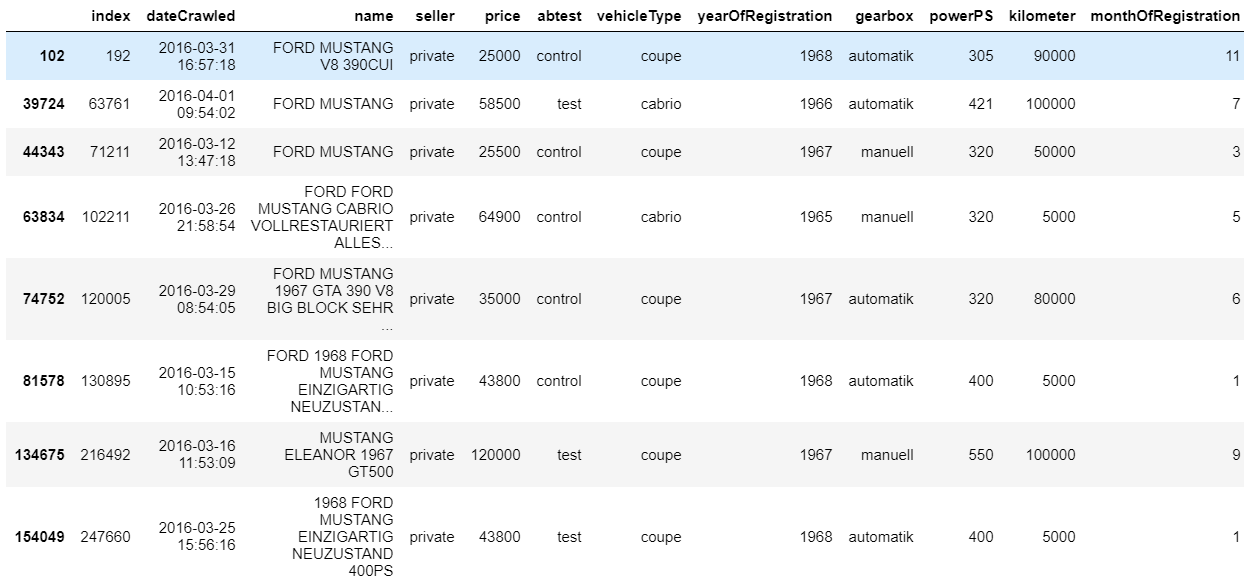
Automatic transmission vehicles are twice as expensive as manual transmission vehicles. Automatic transmission is more sophisticated, therefore it is more expensive. Moreover, some luxurious models only come in automatic transmission.

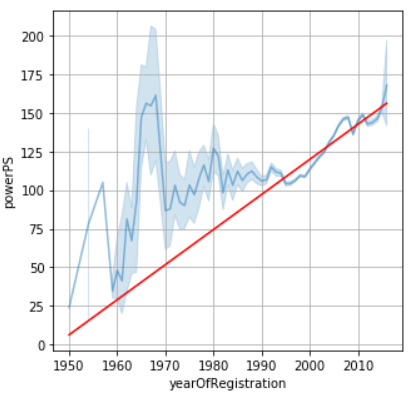
**Correlation Calculations**



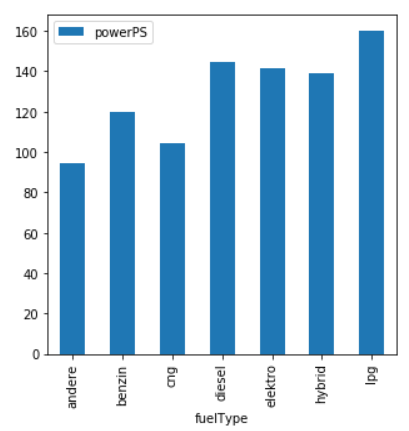
In this project, correlation was measured by pearson coefficient r, which shows strong correlation between continuous numeric variables if r = 1, no correlation if r=0, and strong negative correlation if r=-1.

Initial r for the horsepower - price correlation was as low as 0.2. This number seemed to be too small for me, and then after performing another check, I noticed that I failed to realize PS outliers with extremely high values. These were removed and then another pearson correlation value calculation revealed the actual r value is 0.5. The removal of only 130 rows out of 232 thousand changed the correlation coefficient dramatically.

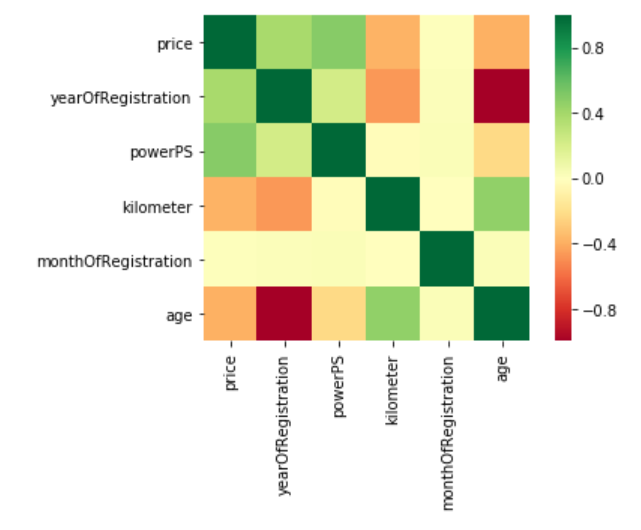
The year vs power chart drew my attention immediately. The power of the vehicles registered in 1960-1970 era create a very noticeable bump in the graph. The data seemed to be wrong, and needed an examination, which revealed only a limited number of data points that were off. The Fiat 500, which is a compact car, has engine volume of 598cc, and this number was entered in the HP column. However, after removing these data points, the shape of the graph did not change. This was due to the increasing role of the American muscle cars in European market! The mean horsepower in 50’s was 50.8, and with the introduction of the American vehicles, the mean increased to 132 in just a decade. In fact, there are only two vehicles in the database in the 50’s that are more powerful than 100 PS. The American invasion of the German car market with hot rods can be clearly seen 

below. A decade ago, 2 vehicles in this DB exceeded two digit horsepower values, but just in a few years, export vehicles as powerful as 550 hp were introduced. The graph on the left was created after removal of erroneous Fiat 500 values. 

It is really interesting to see that the most powerful vehicles are LPG powered. This data seems erroneous at first glance, but a brief look at the vehicle brands and models reveals that the data is right. LPG conversion kits are very common in Europe, and when people buy vehicles that have low gas mileage, they tend to have the vehicles converted to LPG, which cuts down the fuel costs. It is quite obvious that this conversion would be more prevalently visible in bigger engine cars: small engine burns less fuel and kit price is comparable to the savings that would have been



made after the conversion. However, it is a must to convert powerful vehicles to LPG, since conversion cost would be negligible in the long run. Actually some people prefer to have their vehicles converted as soon as they buy them, even if it is brand new.



The heat map generated to display the correlation coefficients can be seen on the left. An internal algorithm coded into Pandas automatically blocks out columns that are not continuous numeric variables, leaving out transmission type, fuel type etc. These will be handled as a part of the classification problem in the machine learning phase.