Automobile Sales On The Go

Dataset consists of 370000+ used cars which is on sale on eBay.

**SUMMARY**

The content of the data is in german, so one has to translate it first if one can not speak german. Those fields are included:

**dateCrawled** : when this ad was first crawled, all field-values are taken from this date

**name** : "name" of the car

**seller** : private or dealer

**offerType** : types of offer

**price** : the price on the ad to sell the car

**abtest** : type of testing

**vehicleType** : types of vehicle

**yearOfRegistration** : at which year the car was first registered

**gearbox** : types of gear available

**powerPS** : power of the car in PS

**model** : car's model

**kilometer** : how many kilometers the car has driven

**monthOfRegistration** : at which month the car was first registered

**fuelType** : types of fuel

**brand** : available car's brand

**notRepairedDamage** : if the car has a damage which is not repaired yet

**dateCreated** : the date for which the ad at ebay was created

**nrOfPictures** : number of pictures in the ad

**postalCode** : car sold in area

**lastSeenOnline** : when the crawler saw this ad last online

The fields lastSeen and dateCrawled could be used to estimate how long a car will be at least online before it is sold.

Dataset is having to be exact 371528 observation & 20 Features to be explored.

**Missing Values Handling**

Among all of them 5 features [vehicleType, gearbox, model, fuelType, notRepairedDamage] having missing values.

As all of them are categorical features, sorted the value by taking the count of them individually and imputed the missing value by highest value of Categorical Features respectively.

**Univariate Analysis**

Found new feature “Sold\_In\_Days” by taking difference of dateCrawled & lastseenOnline which gives us result for how many days the car was online before it was sold.

**Sold\_In\_Days** : Most of the car was sold in <= 2 days and this is the high trending time for selling.

No. Of cars sold on the day the ad was published : 48418

No. Of cars sold on the 1st day the ad was published : 23370

No. Of cars sold on the 2nd day the ad was published : 23466

**Sellers** : There were 2 types of seller was present.

Privat : 99.99% of car was from this seller

Gewerblich : 0.01% only was from this seller

**Offer Type** : There were 2 types of offer was going on.

Angelbot : 99.9% of the offer was this type

Gesuch : 0.1% of the offer was this type

**Price** : Number of observations where price is 0 : 10258, indicates missing price or we could consider this, the car was sold free of cost.

2.7 % of the data includes 0.

Some of the car costs more than 200000, count is : 163

Again car costs less than 200 is : 16671

**AB Test** : There were 2 types of testing available

Test : 52% of testing was this type

Control : 48% of the testing was this type

**Vehicle Type** : All Total 8 types of vehicles were present. Most sold to less sold type of vehicles are : 'limousine', 'kleinwagen', 'kombi', 'bus', 'cabrio', 'coupe', 'SUV', 'andere' respectively.

Top most sold type of car was : Limousine

Less sold type car was : Andere

**Year of Registration** : Tried to visualize top 10 years when car was registered for sale.

2000, 1999, 2005, 2006, 2001, 2003, 2004, 2002, 1998, 2007 - in these consecutive year car was registered more than any year. Among which 1999 & 2006 was highest no of car was registered.

Taking the observation of year in between 1989 & 2019 only

**Gearbox** : 2 Types of car was available.

Manuell : 80% car was this type

Automatic : 20% car was this type

**Power PS** : There were several types of car available powered by different horse power.

But before that it has been observed that No of observation having value of 0 was 40820.

Ignoring this observation took the top 10 cars sold by top most liked horse powered car which was - 75, 60, 150, 140, 101, 90, 116, 170, 105, 125

Among those most sold car was : 75 & 150 powered by HorsePower.

**Car’s Model** : AMong several model found top 10 car was sold in that sell which was - 'golf', 'andere', '3er', 'polo', 'corsa', 'astra', 'passat', 'a4', 'c\_klasse', '5er'

Golf & Andere was the top most cars sold respectively.

**No. Of Cars Sold in Month** :

Less amount of car was sold in the month of Jan & Mar

Whereas max no. of car was sold in the month followed by Apr & Jul.

**Types of Fuel** :

There are 7 types of car available with different fuel types. Among which benzin & diesel type of engine were the most.

**Types of Car’s Brand** : Among various types of brands of Car took top 10 most available car.

They are : 'volkswagen', 'bmw', 'opel', 'mercedes\_benz', 'audi', 'ford', 'renault', 'peugeot', 'fiat', 'seat'

Above list is arranged in chronological order from highest availability to lowest.

**No Of Car Not Repaired Damaged:** Among all the cars present 95% of them was not repaired and 5% of them was repaired.

**Postal Code** : By analyzing the postal code it was found that 99% of the postal code belongs to Germany only.

Not a single Picture was needed to sell those cars as per observation.

**Bi-Variate Analysis**

Filtered out the observation where car was sold within 0-4 days pf time interval.

Observation :

* Vehicle Type - Limousin, Kleinwagen, Kombi was sold among any vehicle within 0-4 days.
* Car with Automatic gear was most preferable among customer
* Benzin & Diesel was top sold fuel type cars

**Outliers Handling**

**monthOfRegistration** : Unique values were present from 0-12, whereas in calendar only 12 months available. Removed observation equal to 12 having observation of : ~12k

**Price** : There is no make sense of having car’s price of 0, less than $200 or $200000 as it is out of range and suspicious as well. Removed all the observation which is < $200 & > $200000

**PowerPS (Horsepower)** : Car without horsepower can’t be operational. In the same place the highest value of powered HP in the world as of now in 662. So, removed all the observation which is having horsepower of < 0 & > 662

**yearOfRegistration** : Taking into consideration where the year is in between 1989 & 2019 only to handle and understand the data more easily.

**Sold\_In\_Days :** Found this column by subtracting values from lastSeen to dateCrwaled having days in int format. Removed all the observation having negative as it is meaningless.

**Correlation** :

There are almost no correlation among price, yearofRegistration, powerPS, kilometer, monthOfRegistration, nrOfPictures, postalCode

**Overall Observation**

* Privat car was sold more than any seller present on the observation
* Angelbot was most accepted offer than others
* Distribution of car price is normal
* Both types (test & control) of testing was accepted
* Limousine, Kleinwagen & Kombi were the most sold cars
* In 1999 & 2006 car was mostly registered for sale
* Manuell car was most preferable & liked by all the buyers
* Car powered by 75 & 150 HP was most likely to be bought
* Golf & Andere was the top model car sold
* Most of the car was registered for sale in the month of March & June
* Benzin & Diesel was most preferred car engine to be brought
* Volkswagen, BMW & Opel was top most brand of car was sold
* Interestingly most of the cars which was available for sell was not repaired
* Not a single picture was needed to sell the car