# OneDot Remote Data Task

Amit Kumar Gangwar

# Table of contents


1. Task Overview
2. Data Profiling (Before Pre-processing)
3. Data Preprocessing
4. Data Profiling (After Pre-processing)
5. Data Normalisation
6. Data Integration
7. Take-away message for Customers

# Task Overview

An e-commerce shop would like to onboard new suppliers efficiently. To enable the onboarding process, the customer needs to integrate product data from suppliers in various formats and styles into the pre-defined data structure of their e-commerce shop application.

Our goal is to transform the supplier data so that it could be directly loaded into the target dataset without any other changes. For each step, you should first profile the data to understand what you can do for the customer, then implement a few selected functions (you can keep it lightweight) and tell the customer what you can also potentially do for them, using examples to illustrate.

# Data Profiling (Before Preprocessing)

- Dataframe derived from the JSON file has 21906 rows and 9 Attributes.
- There were no duplicate rows found in the dataframe.
- Only one attribute had missing values i.e. ModelText. It has only 4.3% of missing values which is very less and acceptable.
- Attributes of supplier data are : 'ID', 'MakeText', 'TypeName', 'TypeNameFull', 'ModelText',  'ModelTypeText', 'Attribute Names', 'Attribute Values', 'entity_id'

# Data Preprocessing

- On analysing the data it was concluded that ‘ID’ uniquely identify a vehicle.
- The column ‘Attribute Names’ actually consists of more attributes like 'Doors', 'Seats', ‘city’ etc.
- Data wrangling was conducted in order to get all these attributes as individual columns using pivot table methodology which is described in the code file in comments.
- Now we have expanded dataframe with less rows and more attributes.
- There are many numeric attributes like ‘Ccm’, ‘FirstRegYear’, ‘Hp’ etc whose datataype was object, they were converted to int or float accordingly.
- There were two attributes ‘Co2EmissionText’ and ‘ConsumptionTotalText’ who are suppose to be numeric but because of units attached to them, they were strings. They were converted to numeric datatype after splitting the unit parts.

# Data Profiling (After preprocessing)

## Structure Discovery

- Now there are 1153 rows and 25 attributes.
- Still there are no duplicate rows.
- The attributes now are : 'ID', 'MakeText', 'TypeName', 'TypeNameFull', 'ModelText', 'ModelTypeText', 'BodyColorText', 'BodyTypeText', 'Ccm', 'City', 'ConditionTypeText', 'ConsumptionRatingText', 'Doors', 'DriveTypeText', 'FirstRegMonth', 'FirstRegYear', 'FuelTypeText', 'Hp', 'InteriorColorText', 'Km', 'Properties', 'Seats', 'TransmissionTypeText', 'Co2Emission(g/km)', 'Consumption(l/100km)'
- There are 5 attributes which have zero values and 4 attributes which have missing values.
- Co2Emission(g/km), Consumption(l/100km) have highest percentage of total Zero and Missing Values i.e. 29.6% and 26.3% respectively.

Below are some more attributes with some zero values which does not make any sense and could be termed as noise possibly:

- ‘FirstRegMonth’, it cannot be zero since maximum is 12.
- ‘Seats’, number of seats can't be zero.
- ‘Hp’, which is horse-power cannot be zero.
- ‘Ccm’, which Cubic centimeter and which denotes engine capacity, can't be zero.
- ‘Doors’, number of doors also can't be zero.

## Content Discovery

- The registration year of the vehicles range from 1927 to 2016.
- The vehicles are from 70 different manufacturers.
- There are total 1153 vehicles.
- AUDI RS6 Avant quattro, MERCEDES-BENZ SL 500 and PORSCHE 911 Turbo models have highest numbers, i.e. 31, 30, 30 respectively.
- Most vehicles are of color Black met. and Silver met., 232 and 208 in numbers respectively.
- Coupé, Limousine, convertible  are kind of vehicles with highest numbers, i.e. 369, 325 and 248 respectively.
- All cities are from Switzerland, one with highest numbers of vehicles is Zuzwil, it has 1118 number of vehicles out of total 1153, which is almost 97% of vehicles just from one city.
- Most number of cars are double door cars, which is 607, more than half.
- Almost 40% are 5 seater vehicles.

# Data Normalisation

- There were special characters in the suppliers data like "Ã¼", 'Ã©' , 'Ã¨' and 'Ã‰' which were decoded and replaced by their equivalent according to the UTF-8 Encoding Debugging Chart (https://www.i18nqa.com/debug/utf8-debug.html )
- I'll choose 'BodyTypeText', 'BodyColorText' and ‘ConditionTypeText’ attributes for normalisation because they are present in the target data as well in the form of 'carType', 'color'  and ‘condition’ respectively. They are in english in target data but in supplier data they are in German , partially or entirely, so they need to be translated to english.
- Other attributes which need normalisation are 'ConditionTypeText', 'DriveTypeText', 'FuelTypeText', 'InteriorColorText', 'Properties', 'TransmissionTypeText' because all of them are either completely or partially in German language. So they need to be converted to english.
- Not sure if ‘fuel_consumption_unit’ in target data is same as the unit of Consumption(l/100km) attribute in supplier data, So it wasn’t touched.

# Data Integration

Target data attributes which are also present in Supplier data with same or different names. They were merged in the same columns. They are:

| Target Data Attribute | Supplier Data attributes |
|---|---|
| city | City |
| make | MakeText |
| manufacture_year | FirstRegYear |
| mileage | Km |
| model | ModelText |
| model_variant | ModelTypeText |
| manufacture_month | FirstRegMonth |
| carType | BodyTypeText |
| color | BodyColorText |
| condition | ConditionTypeText |

Attributes present in Target data but missing in Supplier data, they will be marked as NA while integration. They are :

| **Attribute** |
| --- |
| currency |
| drive |
| price_on_request |
| zip |
| fuel_consumption_unit |

The ‘mileage_unit’ and ‘country’ attributes needs to be hard coded. ‘CH’ for ‘country’ because all are swiss cities in supplier data and ‘km’ for ‘mileage_unit’ because the unit of mileage, which is denoted by attribute ‘km’ in supplier data is km.

A new column, ‘type’ will be added in supplier data and values will be assigned on the basis of 'BodyTypeText' values. A table research was conducted to decide what categories these 'BodyTypeText' vehicles belong to. Find values in table below:

| **'BodyTypeText' values in supplier data** | **'type' value** |
|---|---|
| Limousine, combi, Coupé, SUV / cross-country, small car | car |
| Demountable Camper | Camper |
| compactvan / Minivan | Van |
| articulated lorry, Pick-up | Truck |

# Take-away message for Customers

1. New supplier has new variety of  vehicles like Trucks, Vans and Campers unlike previous suppliers.
2. Supplier operates in Switzerland because all cities in supplier data are from Switzerland.
3. City of Zuzwil has 1118 number of vehicles out of total 1153, which is almost 97% of vehicles just from one city.
4. The registration year of the vehicles ranges from 1927 to 2016.
5. Most number of cars are double door cars, which is 607, more than half.
6. Coupé, Limousine, convertible  are kind of vehicles with highest numbers, i.e. 369, 325 and 248 respectively.

# THANK YOU