File name: VehicleToOrganize.xlsx

progr	*freq*	*Excel label*	*English label*	*Value*
1	100%	**CodVeicolo**	cod_vehicle	3
2	99%	**anno**	year	1934
3	100%	**Marca**	brand_name	Alfa Romeo
4	100%	**Modello_A_Serie**	model_part_a	6C 2300
5	50%	**Modello_B_Versisone**	model_part_b	Gran Sport
6		**Mod_Verificato**	verified	
7	100%	**Mod_Famiglia**	family	6C
8	50%	**Mod_Versione**	version	
9	100%	**Mod_Serie**	series	I
10	50%	**Mod_Allestimento**	dressing	
11	100%	**Mod_Stage**	stage	1
12		**Mod_Potenza**	power	
13	100%	**Modello_Descriz_Originale**	description	1930 Alfa Romeo 6C 1750 Gran Sport

1) CodVeicolo: this value is the key_value. You have to copy this value in new elaborated file without any change.

2) anno: [extract from (13)**]** (\d{4}) - attention, it's not always first element on the left side; sometimes you can find values like 99 (= 1999), 60' (1960-1969 = we consider 1965) - definition order = 1

3) Marca (= brand name): [extract from (13)**] -** usually is the first text value on the right of [year]. verify existing value into file Brand.xlsx - attention: you can alias values like: Alfa Romeo, Alfa, Alfa Romeo. In any case brand name is omitted and you'll find only family or model name . Verify that [brand_name] is inside year range [year_begin ….year_end inside file brand.xlsx] . If description contains [clone, replica, handmade, etc…] than, brand_name = brand-name_replica. Copy "Record_ID" as "ID_Record_brand"value into new sheet definition order = 2

4) Modello_A_Serie: Usually is the same value of [Mod_Famiglia - Mod_Serie] (7 - 9). To define [Mod_Serie] you need to verify inside Model.xlsx file if columns "Mod_FamigliaA" = Mod_Famiglia (7). If so, count repetition. If repetition = 1 than Modello_A_Serie = [Mod_FamigliaA] , else find into range [year_begin ….year_end inside file Model.xlsx] the value in column Mod_FamigliaB . Copy "ID_Record" as "ID_Record_model" value into new sheet. definition order = 4

5) Modello_B_Versisone: all text of point 13, less text of point 2,3,7,4. definition order = 5

6) Mod_Verificato:

7) Mod_Famiglia: usually is the first text value on the right of [Brand_name] verify existing value into file Family.xlsx. - Verify that [family] is inside year range [year_begin ….year_end inside file Family.xlsx] . Copy "ID_Record" as "ID_Record_family" value into new sheet . definition order = 3

8) Mod_Versione:
9) Mod_Serie: same than point 4
10) Mod_Allestimento:
11) Mod_Stage: column Stage in Model.xlsx of corresponding model value
12) Mod_Potenza: not to elaborate
13) **Modello_Descriz_Originale**: text to elaborate

File name: Brand.xlsx

progr	*freq*	*Excel label*	*English label*	*Value*
1	100%	**Record_ID**	id	1
2	100%	**Marca**	brand_name	Lotus
3	10%	**MarcaCompleto**	full_name	
4	10%	**MarcaCorto**	short_name	
5	10%	**NomeStoricoCostante**	recursive_name	
6	99%	**DataInizio**	year_begin	1952
7	50%	**DataFine**	year_end	

1) Record_ID: this value is the key_value. You have to copy this value in new elaborated file without any change.
2) Marca: always present (=> create a hard code dictionary to add values like: “General Motors”, “GM”, “General motor corporation”, etc……..)
3) MarcaCompleto: possible alias
4) MarcaCorto: possible alias
5) NomeStoricoCostante: possible alias
6) DataInizio: if not present, do not evaluete
7) DataFine: if not present consider still existing

File name: Family.xlsx

progr	*freq*	*Excel label*	*English label*	*Value*
1	100%	**Cod_Famiglia**	family_code	39393
2	100%	**ID_Record**	Id	39396
3	100%	**Marca**	brand_name	Abarth
4	100%	**Famiglia**	family	1500
5	50%	**AnnoFine**	year_begin	
6	50%	**AnnoInizio**	year_end	
7	10%	**EsemplariProdotti**	production	
8	10%	**Fonte_web**	source_web	
9	100%	**Marca::Record_ID**	id_brand	29
10	10%	**Conta_VeicoliPresenti**	DB_vehicles	1
11	10%	**Conta_Modelli**	count	1
12	100%	**AnnoInizioCalcolato**	year_begin_calc	1952
13	100%	**AnnoFineCalcolato**	year_end_calc	1952

1) Cod_Famiglia: this value is the key_value. You have to copy this value in new elaborated file without any change.
2) ID_Record: this value is the key_value. You have to copy this value in new elaborated file without any change.
3) Marca: brand_name
4) Famiglia: family name
5) AnnoFine: year end, if not existing assume value 13 “year_end_calculated”
6) AnnoInizio: year begin, if not existing assume value 12 “year_begin_calculated”
7) EsemplariProdotti: do not consider
8) Fonte_web: do not consider

9) Marca::Record_ID: do not consider
10) Conta_VeicoliPresenti: do not consider
11) Conta_Modelli: do not consider
12) AnnoInizioCalcolato: calculated year begin
13) AnnoFineCalcolato: calculated year begin

File name: Model.xlsx

progr	*freq*	*Excel label*	*English label*	*Value*
1	100%	**ID_Record**	id	33791
2	100%	**Marca**	brand_name	Abarth
3	100%	**Mod_FamigliaA**	model_name_a	204
4	100%	**Mod_FamigliaB**	model_name_b	204
5	100%	**Serie**	seires	I
6	100%	**Stage**	stage	1
7	100%	**AnnoInizioCalcolato**	year_begin	1948
8	100%	**AnnoFineCalcolato**	year_end	1948

1) ID_Record: this value is the key_value. You have to copy this value in new elaborated file without any change.
2) Marca: brand_name
3) Mod_FamigliaA: = “family”
4) Mod_FamigliaB: = “Mod_FamigliaA” - “Serie”
5) Serie: if count(Mod_FamigliaA)=1 than “I”, else verify “year-begin” and “year_end” to define “series”.
6) Stage:
7) AnnoInizioCalcolato:
8) AnnoFineCalcolato:

File name: VehcilesOrganized.xlsx

progr	*freq*	*Excel label*	*English label*	*Value*
1	100%	**CodVeicolo**	id	1
2	100%	**anno**	year	1982
3	100%	**Marca**	brand_name	Alfa Romeo
4	100%	**Modello_A_Serie**	model_name_a	F1
5	50%	**Modello_B_Versisone**	model_name_b	182
6		**Mod_Verificato**	verified	no
7	100%	**Mod_Famiglia**	family	F1
8	50%	**Mod_Versione**	version	
9	100%	**Mod_Serie**	series	I
10	50%	**Mod_Allestimento**	dressing	
11	100%	**Mod_Stage**	stage	1
12		**Mod_Potenza**	power	
13	100%	**Modello_Descriz_Originale**	description	1982 Alfa Romeo 182

Same structure of VehicleToOrganize.xlsx

use this file as a reference to the previous conversions of the "description" and the relative distribution of the data in the relevant columns

1) CodVeicolo
2) anno
3) Marca
4) Modello_A_Serie
5) Modello_B_Versisone
6) Mod_Verificato
7) Mod_Famiglia
8) Mod_Versione
9) Mod_Serie
10) Mod_Allestimento
11) Mod_Stage
12) Mod_Potenza
13) Modello_Descriz_Originale

File name: AI_VehcilesOrganized.xlsx

progr	*freq*	*Excel label*	*English label*	*Value*
1	100%	**CodVeicolo**	id	1
2	100%	**anno**	year	1982
3	100%	**Marca**	brand_name	Alfa Romeo
4	100%	**Modello_A_Serie**	model_name_a	F1
5	50%	**Modello_B_Versisone**	model_name_b	182
6		**Mod_Verificato**	verified	no
7	100%	**Mod_Famiglia**	family	F1
8	50%	**Mod_Versione**	version	
9	100%	**Mod_Serie**	series	I
10	50%	**Mod_Allestimento**	dressing	
11	100%	**Mod_Stage**	stage	1
12		**Mod_Potenza**	power	
13	100%	**Modello_Descriz_Originale**	description	1982 Alfa Romeo 182
14	100%	**Cod_brand**	Cod_brand	
15	100%	**Cod_family**	Cod_family	
16	100%	**Cod_model**	Cod_model	
17	100%	**Accuracy**	Accuracy	

Accuracy:
A = 100% (Brand, family, model is inside year-range and in description)
B = 99% (Brand, family, model is inside year-range but not in description)
C = 80% (Brand, family, model is inside year-range and in description)
D = 50% (Brand, family, in description but year range is not inside range +/- 2 year)
E = 40% (Brand, family, in description but year range is not inside range MORE +/- 2 year)

F = 25% (Only Year and Brand in description, and inside year range))
G = 20% (Only Year and Brand in description, but not inside year range)
H = 5% (Family is not listed in family)
J = 0% (Brand is not listed in brand.xlsx)