DATA CONSULTING

Data Transformation and Integration

TABLE OF CONTENTS

01

Insights to input data

02

Process of data integration

03

Key take-aways from the integrated data

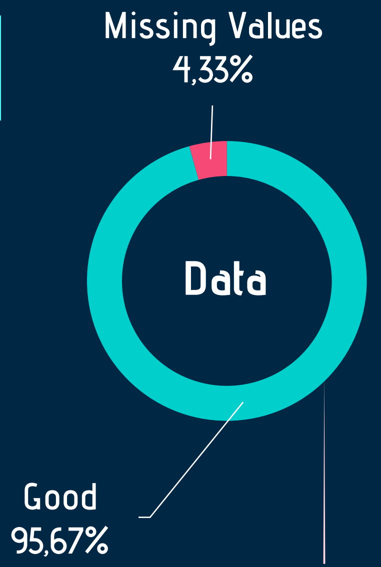
01. Insight to input data

Orginal data

Number of records :
21 ,906
Features / Columns :
9

General information

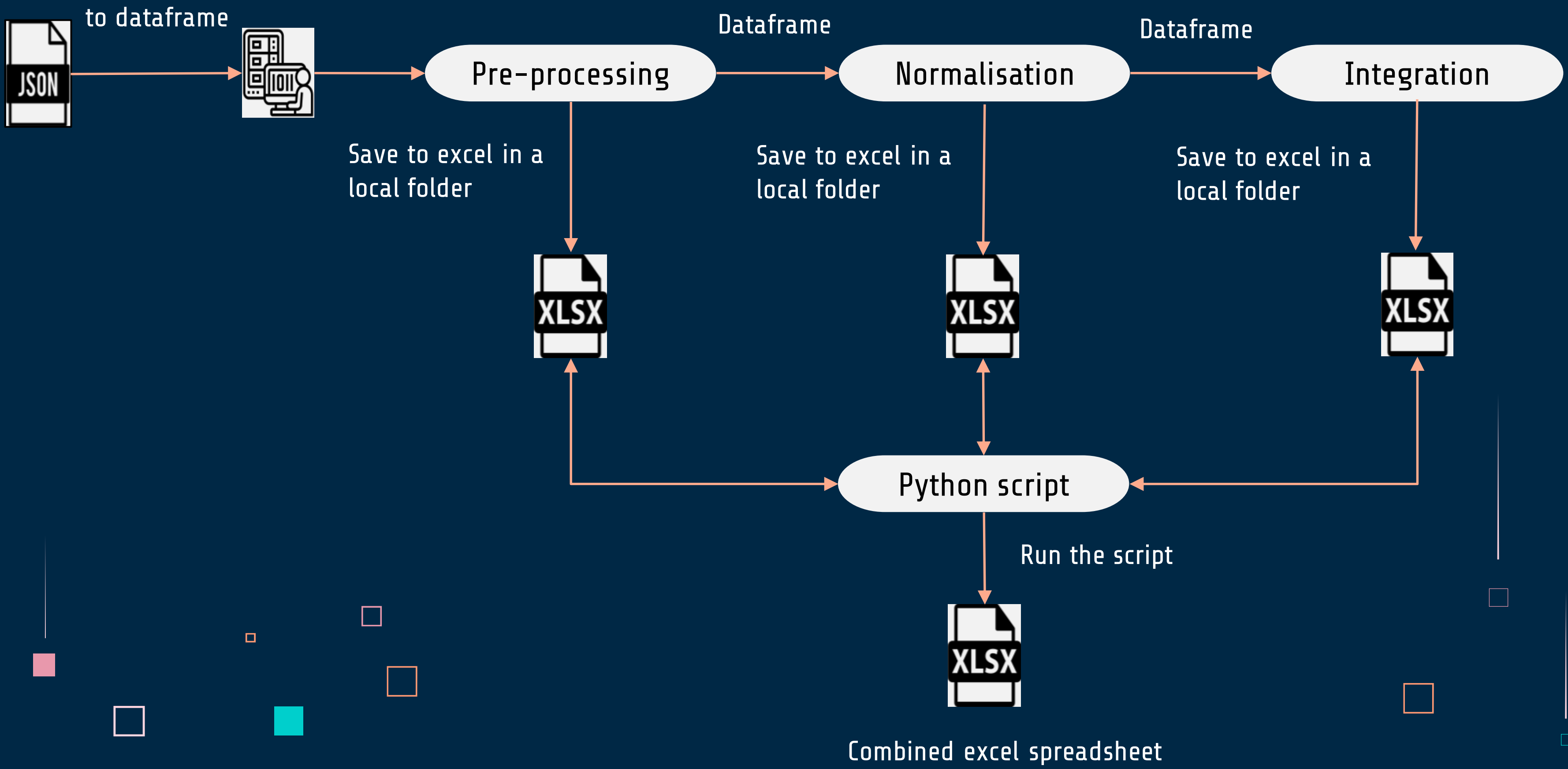
1. Each model has 19 attributes as records
2. Missing values in model type



What we know about the input data

1. Input data file is json file type
2. Input file needs to be converted to excel
3. Inconsistent naming convention
4. Missing values
5. Feature names as measuring unit.
6. Inconsistent letter cases.
7. Redundant data

02. Process of data integration


JSON
to dataframe
Pre-processing
Dataframe
Normalisation
Dataframe
Integration
Save to excel in a local folder
XLSX
Save to excel in a local folder
XLSX
Save to excel in a local folder
XLSX
Python script
Run the script
XLSX
Combined excel spreadsheet

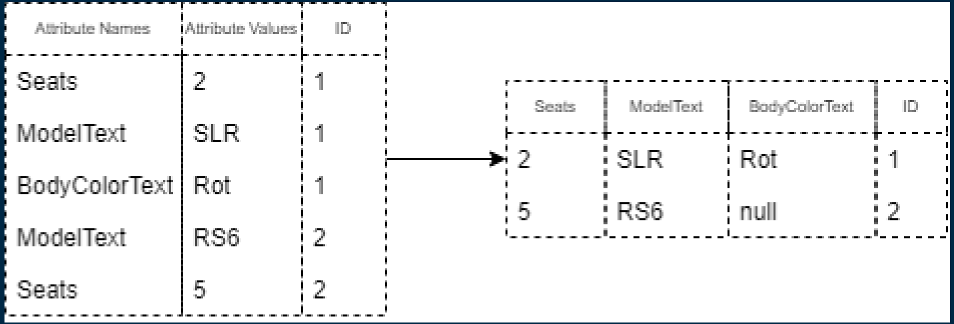
Pre-processing

After reading the json file into dataframe attributes are transformed to features, this reduces the number of records to around 1000.

Dataframe and the excel files are saved as output

PRE

POST



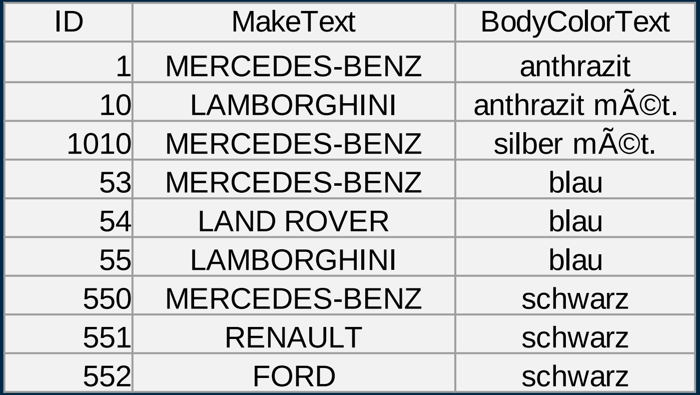
Attribute Names	Attribute Values	ID
Seats	2	1
ModelText	SLR	1
BodyColorText	Rot	1
ModelText	RS6	2
Seats	5	2

Seats	ModelText	BodyColorText	ID
2	SLR	Rot	1
5	RS6	null	2

Normalisation

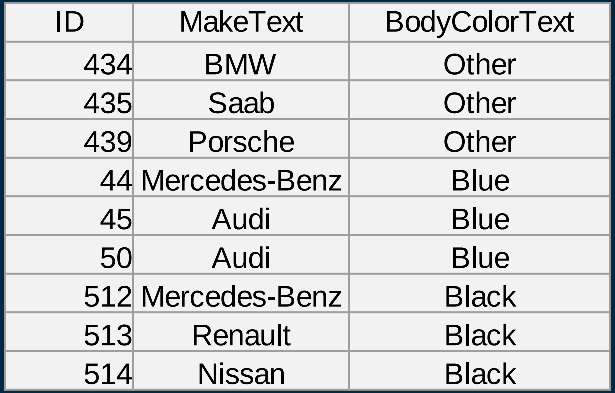
The ‘BodyColorText’ and ‘MakeText’ features of the Input Pre-processed dataframe are normalized and saved as excel file and dataframe as output

PRE



ID	MakeText	BodyColorText
1	MERCEDES-BENZ	anthrazit
10	LAMBORGHINI	anthrazit mÃ©t.
1010	MERCEDES-BENZ	silber mÃ©t.
53	MERCEDES-BENZ	blau
54	LAND ROVER	blau
55	LAMBORGHINI	blau
550	MERCEDES-BENZ	schwarz
551	RENAULT	schwarz
552	FORD	schwarz

POST



ID	MakeText	BodyColorText
434	BMW	Other
435	Saab	Other
439	Porsche	Other
44	Mercedes-Benz	Blue
45	Audi	Blue
50	Audi	Blue
512	Mercedes-Benz	Black
513	Renault	Black
514	Nissan	Black

Normalisation required

Some other attributes that require normalisation

"ConditionTypeText"
Change German to English
Example : VorfÃ¼hrmodell to Used, Neu to New

"InteriorColorText"
Change German to english
Example : schwarz to Black

"DriveTypeText"
Change German to English
Example : Hinterradantrieb to Rear-wheel drive

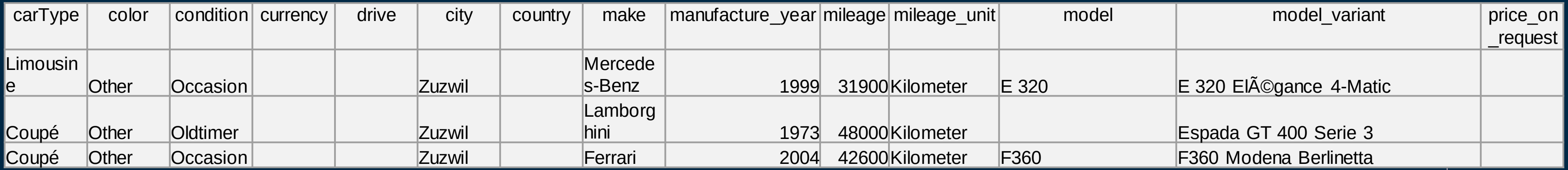
"km"
Change the km column from data type object to float
Example :48000 to 48000.0

Integration

Taking normalized data-frame as input, data is integrated to the copied schema of the target dataset.

unmapped features are removed,

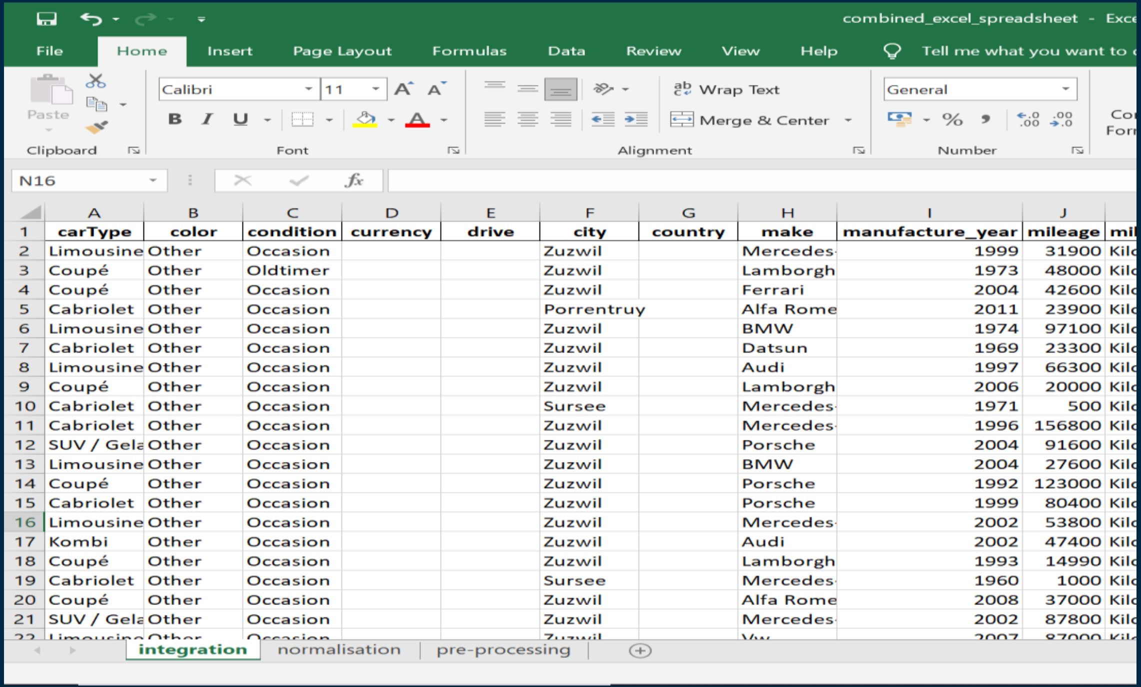
Finally the data-frame is saved as excel file as output.



carType	color	condition	currency	drive	city	country	make	manufacture_year	mileage	mileage_unit	model	model_variant	price_on_request
Limousine	Other	Occasion			Zuzwil		Mercedes-Benz	1999	31900	Kilometer	E 320	E 320 ElÃ©gance 4-Matic	
Coupé	Other	Oldtimer			Zuzwil		Lamborghini	1973	48000	Kilometer		Espada GT 400 Serie 3	
Coupé	Other	Occasion			Zuzwil		Ferrari	2004	42600	Kilometer	F360	F360 Modena Berlinetta	

Combined excel file extraction

Run the python script which searches for all the excel files in the folder and combines them into one excel spreadsheet with different tabs.



	A	B	C	D	E	F	G	H	I	J	K
1	carType	color	condition	currency	drive	city	country	make	manufacture_year	mileage	mil
2	Limousine	Other	Occasion			Zuzwil		Mercedes-	1999	31900	Kil
3	Coupé	Other	Oldtimer			Zuzwil		Lamborgh	1973	48000	Kil
4	Coupé	Other	Occasion			Zuzwil		Ferrari	2004	42600	Kil
5	Cabriolet	Other	Occasion			Porrentruy		Alfa Rome	2011	23900	Kil
6	Limousine	Other	Occasion			Zuzwil		BMW	1974	97100	Kil
7	Cabriolet	Other	Occasion			Zuzwil		Datsun	1969	23300	Kil
8	Limousine	Other	Occasion			Zuzwil		Audi	1997	66300	Kil
9	Coupé	Other	Occasion			Zuzwil		Lamborgh	2006	20000	Kil
10	Cabriolet	Other	Occasion			Sursee		Mercedes-	1971	500	Kil
11	Cabriolet	Other	Occasion			Zuzwil		Mercedes-	1996	156800	Kil
12	SUV / Gela	Other	Occasion			Zuzwil		Porsche	2004	91600	Kil
13	Limousine	Other	Occasion			Zuzwil		BMW	2004	27600	Kil
14	Coupé	Other	Occasion			Zuzwil		Porsche	1992	123000	Kil
15	Cabriolet	Other	Occasion			Zuzwil		Porsche	1999	80400	Kil
16	Limousine	Other	Occasion			Zuzwil		Mercedes-	2002	53800	Kil
17	Kombi	Other	Occasion			Zuzwil		Audi	2002	47400	Kil
18	Coupé	Other	Occasion			Zuzwil		Lamborgh	1993	14990	Kil
19	Cabriolet	Other	Occasion			Sursee		Mercedes-	1960	1000	Kil
20	Coupé	Other	Occasion			Zuzwil		Alfa Rome	2008	37000	Kil
21	SUV / Gela	Other	Occasion			Zuzwil		Mercedes-	2002	87800	Kil
22	Limousine	Other	Occasion			Zuzwil		Vw	2007	87000	Kil

Key take-aways from the integrated data

Key facts

1. Missing values
2. Output file is more clear and small
3. Consistent naming
4. Consistent letter case
5. Redundant datas are removed.
6. "km" feature in supplier data column is assumed to be mileage in target dataset
7. "mileage_unit" feature in target data filled with unit "kilometer" target dataset
8. "FirstRegYear" feature is assumed as "manufacture_year" in target dataset
9. "FirstRegMonth" feature is assumed as "manufacture_month" in target dataset

Recomendation

1. Avoid redundant data
2. Keep naming convention consistent
3. Consider important features like price and feul consumption
4. Try to avoid measuring units as features
5. Country and Zip can be useful to add

Do you have any questions?

Chandrakantha HA

ckanth_ha@yahoo.com
+49 15901286283

THANKS