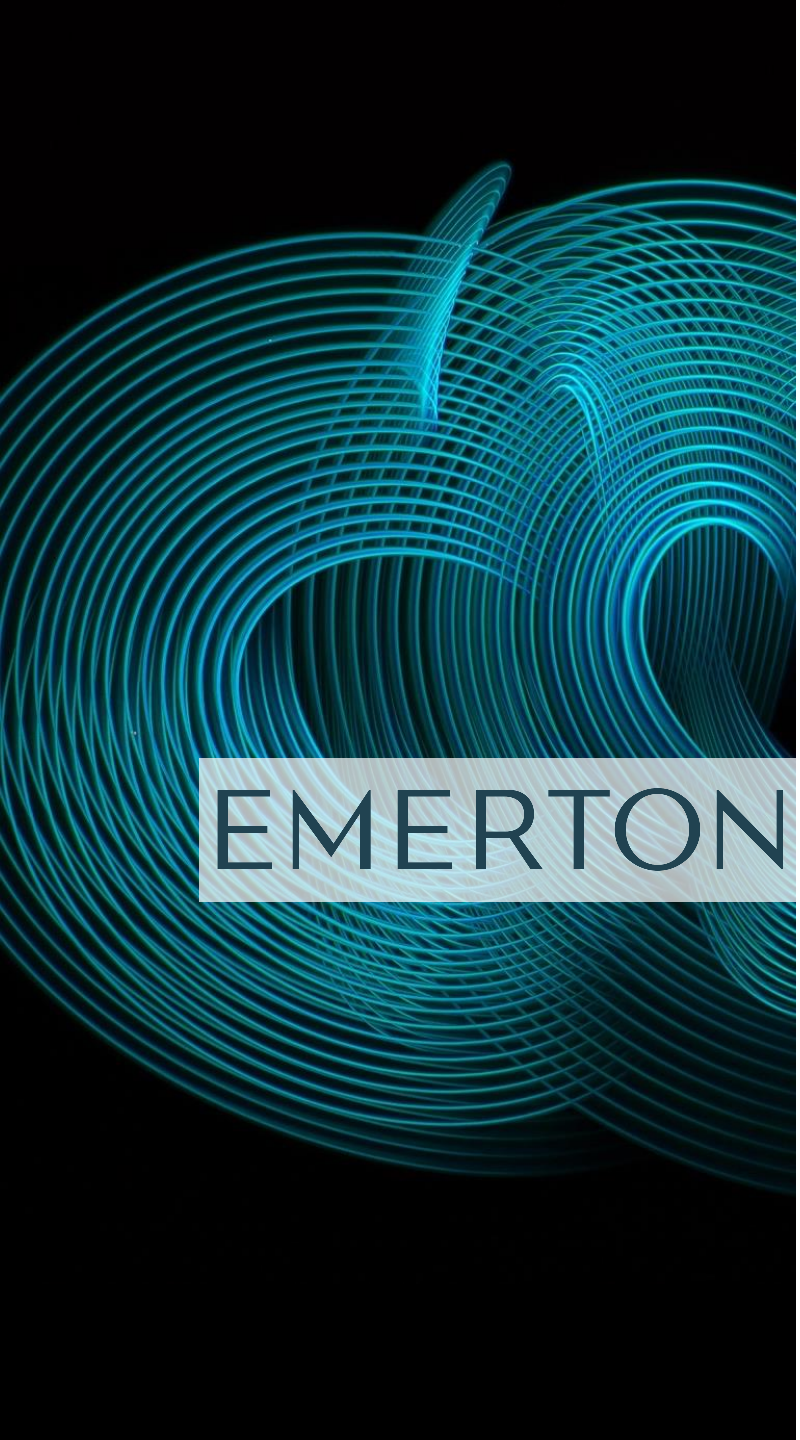


Contact details:

Yannick Léo

Associate Partner & Data Science Director, Emerton Data

yannick.leo@emerton-data.com

EMERTONDATA

Data Science for Business: project exercises and instructions

To deliver before March 12

Data Science for Business: dates and timeline

Dates	Timeline
Tuesday, 3th January, 17:15-20:30	Sharing of the regression exercise
Thursday, 5th January, 17:15-20:30	Sharing of the regression exercise
Thursday, 26th January, 17:15-20:30	Sharing of the shap extension exercise
Sunday, 12th March, 23:59	Projects delivery including 1. regression with 2. shap extension and 3. clustering
Sunday, 19th March, 23:59	Feedback regarding the projects
Thursday, 24th March, 10:00-12:00	Exam

Exercises and instructions

Exercises

1. Build ML model for regression (auto insurance problem)
2. Perform an interpretability study of the built model in the previous exercise
3. Implement the clustering exercises

Delivery instructions

1. Create a google colab with Python environment
2. Access to dataset (see next page)
3. Develop your methodology on a notebook
4. Export colab notebook in .ipynb format
5. The complete restitution is the March 12 and should contain all the exercices

Important to get all points

1. Follow the steps presented in the lessons
2. A cell not ran is to considered
3. Work to get a good result (high mape)
4. Easy to read notebook with few comments and logical steps
5. Follow the delivery instructions

Agenda

1. **Regression exercise**
2. Extension with the interpretability
3. Clustering exercise

Exercise: build full ML methodology to predict Auto Insurance pricing

Problem statement

In the context of price optimization, we need to inject in the demand model a variable that reflects competitiveness. Fortunately we have at disposal a market index price on a set of quotes. We don't know the details on this index (it could be the best price offer, or any kinds of averaged price offer), but we know it is useful to segment elasticity.

Presentation of the dataset and information regarding access

DATA DESCRIPTION

- Id: categorical, quotes identifier number
- Y: response numerical variable, ‘Response_Market_Index’
- X: 38 variables (17 are categorical, 21 are numerical), they are related to:
  - → Behavioral information (BEH prefix): 3 variables
  - → Claims information (CLA prefix): 5 variables
  - → Geocoded information (GEO): 2 variables
  - → Personal information (PER): 13 variables
  - → Product information (PRO): 1 variable
  - → Vehicle information (VEH): 14 variables

FILES DESCRIPTION AND ACCESS

- Files in the folder data_market_prediction:
  - → data_train_insurance.csv: (; separated) dataset to be used, contains Id,Y,X
  - → data_type.csv: contains the list of variables and related
- Access to files:
  - → https://drive.google.com/drive/folders/19g17pqo2DWT1lcOdkwX3lSgj-tyATWNx?usp=sharing

Description of metrics for assessment

CRITERIA

- Mean Absolute Percentage Error – MAPE

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \widehat{y_i}|}{y_i}$$

- Inverse percentile of APE at 20% level

$$IP_{20\%} = card\left\{i \ \in 1, \dots, N : \frac{|y_i - \widehat{y_i}|}{y_i} < 20\%\right\}/N$$

EVALUATION METRICS

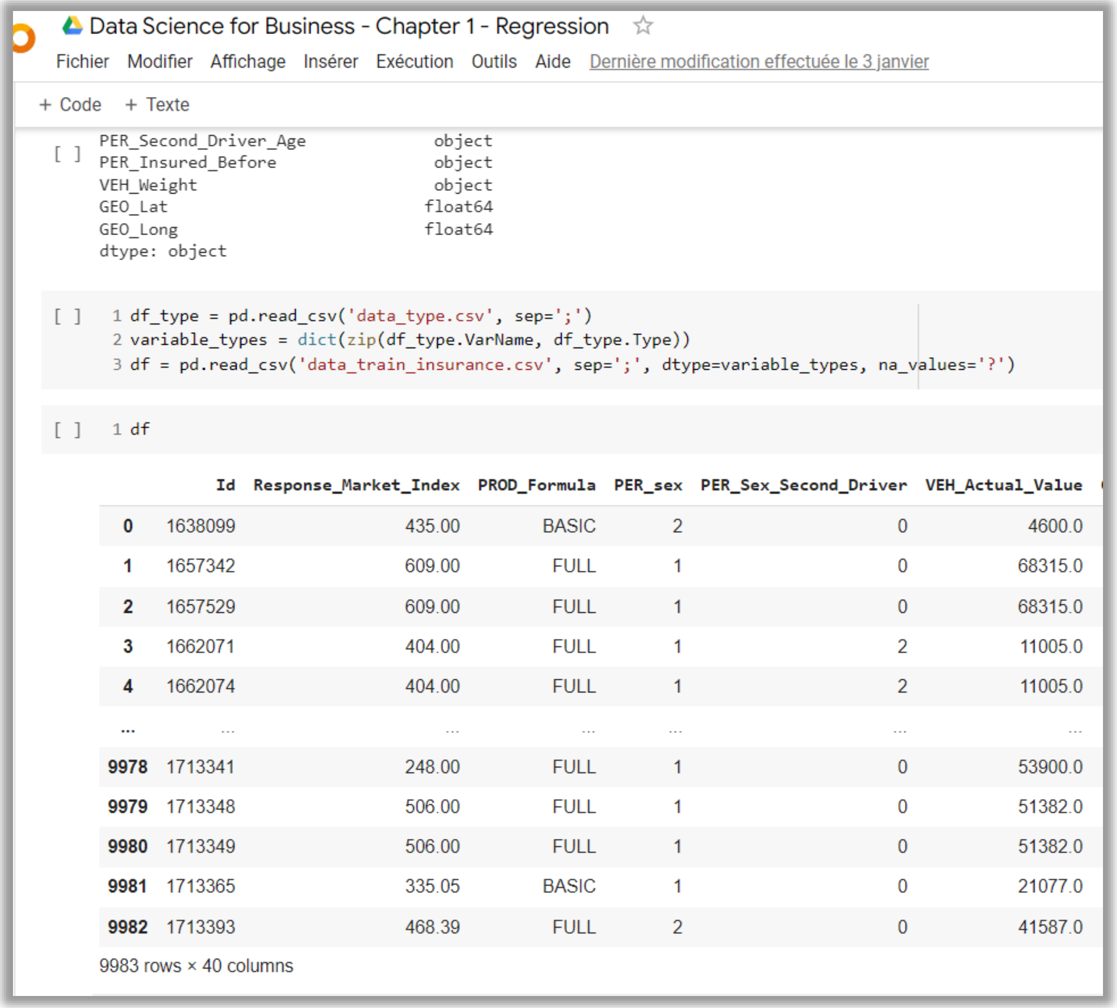
- Mean Absolute Percentage Error – MAPE

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \widehat{y_i}|}{y_i}$$

- Inverse percentile of APE at 20% level

$$IP_{20\%} = card\left\{i \ \in 1, \dots, N : \frac{|y_i - \widehat{y_i}|}{y_i} < 20\%\right\}/N$$

Tips to open the files in the regression notebook


Data Science for Business - Chapter 1 - Regression
Fichier
Modifier
Affichage
Insérer
Exécution
Outils
Aide
Dernière modification effectuée le 3 janvier
+ Code
+ Texte

```
PER_Second_Driver_Age             object
PER_Insured_Before                object
VEH_Weight                        object
GEO_Lat                          float64
GEO_Long                         float64
dtype: object
```

```python
df_type = pd.read_csv('data_type.csv', sep=';')
variable_types = dict(zip(df_type.VarName, df_type.Type))
df = pd.read_csv('data_train_insurance.csv', sep=';', dtype=variable_types, na_values='?')
```

```python
df
```

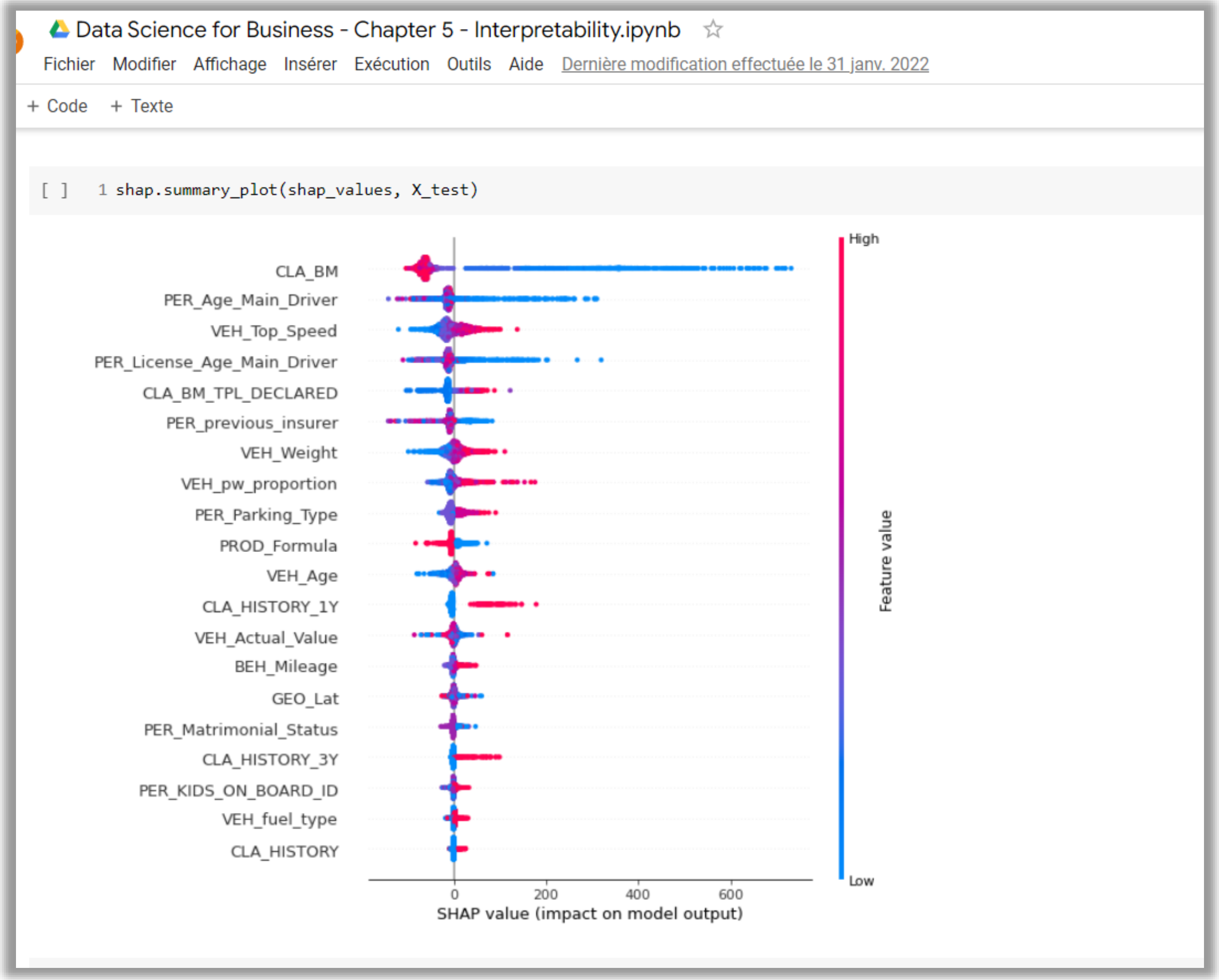
	Id	Response_Market_Index	PROD_Formula	PER_sex	PER_Sex_Second_Driver	VEH_Actual_Value
0	1638099	435.00	BASIC	2	0	4600.0
1	1657342	609.00	FULL	1	0	68315.0
2	1657529	609.00	FULL	1	0	68315.0
3	1662071	404.00	FULL	1	2	11005.0
4	1662074	404.00	FULL	1	2	11005.0
...	...	...	...	...	...	...
9978	1713341	248.00	FULL	1	0	53900.0
9979	1713348	506.00	FULL	1	0	51382.0
9980	1713349	506.00	FULL	1	0	51382.0
9981	1713365	335.05	BASIC	1	0	21077.0
9982	1713393	468.39	FULL	2	0	41587.0

9983 rows × 40 columns

Agenda

1. Regression exercise
2. **Extension with the interpretability**
3. Clustering exercise

Exercise: perform an interpretability study should be implemented during the regression problem as an extension



Data Science for Business - Chapter 5 - Interpretability.ipynb

```
shap.summary_plot(shap_values, X_test)
```

Agenda

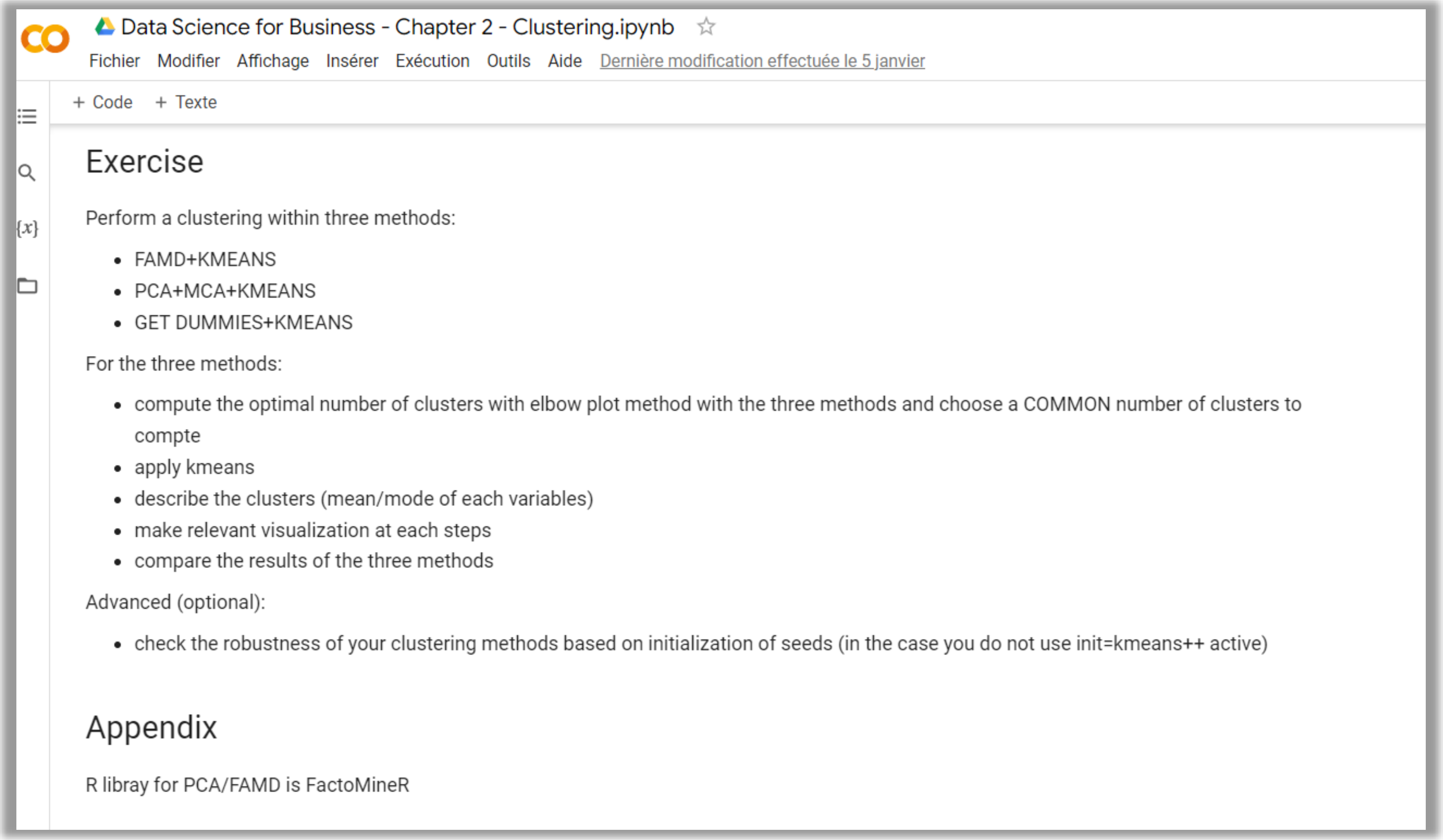
1. Regression exercise
2. Extension with the interpretability
3. **Clustering exercise**

Exercise: build three clustering methodologies on credit bank data

Problem statement:

- Perform a clustering within three methods:
  - Get dummies+KMEANS
  - FAMD+KMEANS
  - PCA+MCA+KMEANS
- For the three methods:
  - Compute the optimal number of clusters with elbow plot method with the three methods and choose a COMMON number of clusters to compute the cluster
  - Apply kmeans other the three methods
  - Describe the clusters (mean/mode of each variables)
  - Make relevant visualization at each steps
  - Compare the results of the three methods
- Advanced (optional):
  - Check the robustness of your clustering methods based on initialization of seeds (in the case you do not use init=kmeans++ active)

Exercise on clustering is positioned at the end of the clustering notebook



CO
Data Science for Business - Chapter 2 - Clustering.ipynb ☆
Fichier Modifier Affichage Insérer Exécution Outils Aide Dernière modification effectuée le 5 janvier
+ Code + Texte

Exercise

Perform a clustering within three methods:

- FAMD+KMEANS
- PCA+MCA+KMEANS
- GET DUMMIES+KMEANS

For the three methods:

- compute the optimal number of clusters with elbow plot method with the three methods and choose a COMMON number of clusters to compte
- apply kmeans
- describe the clusters (mean/mode of each variables)
- make relevant visualization at each steps
- compare the results of the three methods

Advanced (optional):

- check the robustness of your clustering methods based on initialization of seeds (in the case you do not use init=kmeans++ active)

Appendix

R libray for PCA/FAMD is FactoMineR

EMERTONDATA

PARIS | 16 avenue Hoche | 75008 Paris | France

www.emerton-data.com