CentraleSupélec

OPENCLASSROOMS

Parcours Data Scientist

Projet 2 :

Analyse des données nutritionnelles

04/02/2018

MINE Nicolas

Sommaire

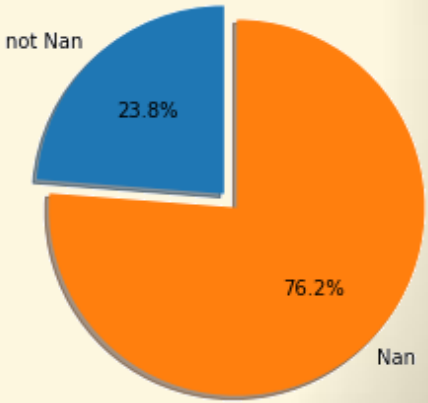
- Présentation et Objectifs
- Nettoyage
  - Dataset original
  - Choix des features
  - Simplification de features
  - Nettoyage des données aberrantes
- Exploration
  - Analyse Univariée
  - Analyse Multivariée
- Exploitation
  - Résultats actuels
  - Modèles possibles
- Conclusion

Présentation

- Entrée
  - Données nutritionnelles d'aliments
  - Dataset peu « fourni »
- Objectifs
  - Paramètres influant la qualité nutritionnelle
  - Avantages et désavantages nutritionnels des aliments
- Sortie
  - Etre capable d'évaluer si une recette a une bonne qualité nutritionnelle

Nettoyage

- 320772 lignes
- 162 features
- 75 % are Nan
  - 51,965 M values
  - 39,6 M Nan



Nettoyage

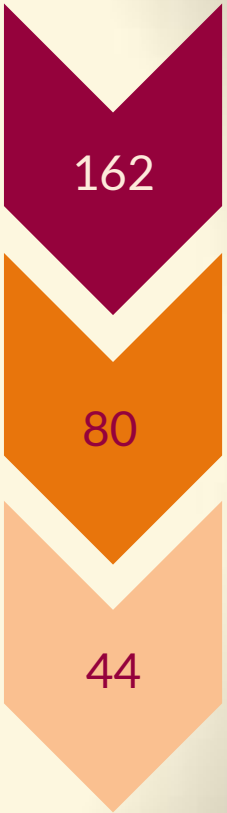
Nettoyage
- ✓ Dataset original
- ➢ Choix des features
- ➢ Simplification de features
- ❑ Nettoyage des données abérantes

Exploration
- ❑ Analyse Univariée
- ❑ Analyse Multivariée

Exploitation
- ❑ Résultats actuels
- ❑ Modèles possibles

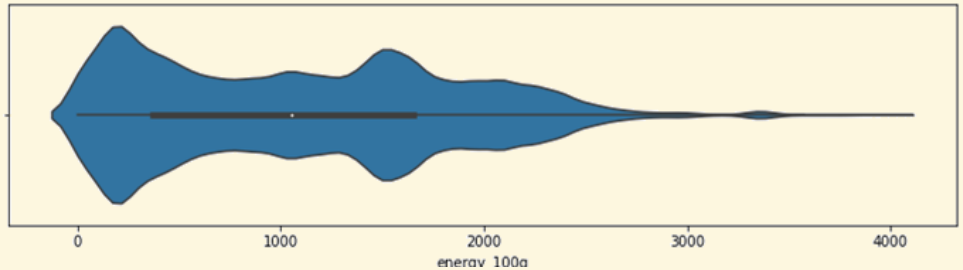
Conclusion

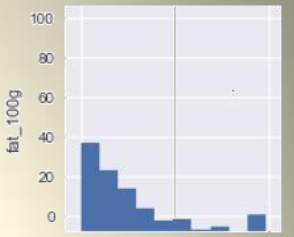
- Suppression Feature:
  - Trop de données manquantes
    - < 1000 pts post-nettoyage
  - Données inutiles
    - Temporelles
    - url, nom, origine, stores, CO2, …
- Agrégation sous un count :
  - Labels, Huile de Palme, Additifs, Allergènes
- Correction certaines features
  - Pnns_groups

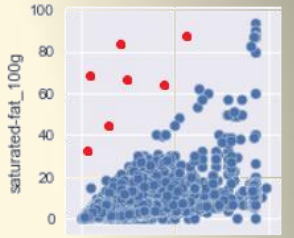


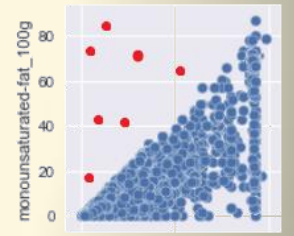
Nettoyage

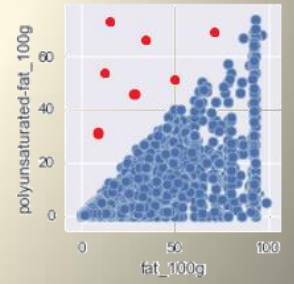
- En parallèle avec l'exploration
  - Suppression des outliers
    - +/- Médiane + 5 x Standard Déviation
  - Suppression de « faux points »
    - Graisses spécifiques > Graisse totale
    - Valeurs négatives











Nettoyage
- ✓ Dataset original
- ✓ Choix des features
- ✓ Simplification de features
- ➢ Nettoyage des données abérantes

Exploration
- ❑ Analyse Univariée
- ❑ Analyse Multivariée

Exploitation
- ❑ Résultats actuels
- ❑ Modèles possibles

Conclusion

Nettoyage

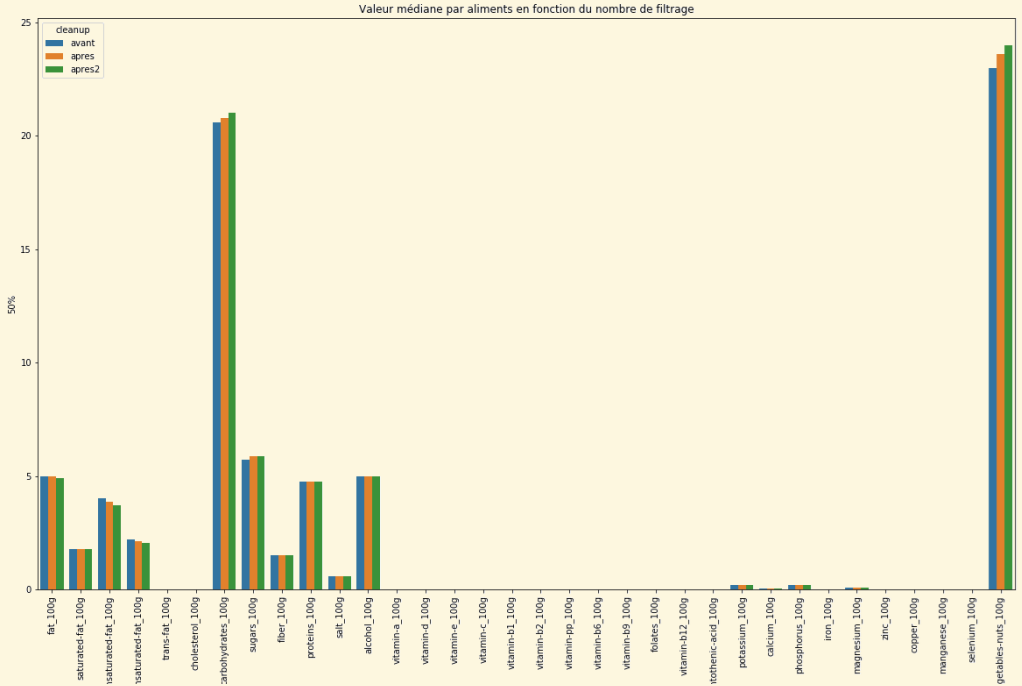
- Intérêt du filtrage
  - Suppression des outliers
  - Peu d'impacte sur la médiane
  - Stabilisation de la Standard Déviation
  - Correction de la moyenne
- Inconvénients
  - Nb d'itérations ?

	index	count	mean	std	min	25%	50%	75%	max	cleanup
10	salt_100g	255510.0	2.028624	128.269454	0.0	0.0635	0.58166	1.37414	64312.8000	avant
44	salt_100g	251684.0	1.655391	7.142057	0.0	0.0635	0.58928	1.36906	604.7613	apres
78	salt_100g	243786.0	1.122654	2.404598	0.0	0.0635	0.57404	1.34366	37.4650	apres2

	index	count	mean	std	min	25%	50%	75%	max	cleanup
33	fruits-vege...	3036.0	31.458587	31.967918	0.0	0.0	23.0	51.0	100.0	avant
67	fruits-vege...	3015.0	31.594783	31.935997	0.0	0.0	23.6	51.0	100.0	apres
101	fruits-vege...	2999.0	31.650640	31.913197	0.0	0.0	24.0	51.0	100.0	apres2

Nettoyage

- Intérêt du filtrage



Nettoyage
- ✓ Dataset original
- ✓ Choix des features
- ✓ Simplification de features
- ➢ Nettoyage des données abérantes

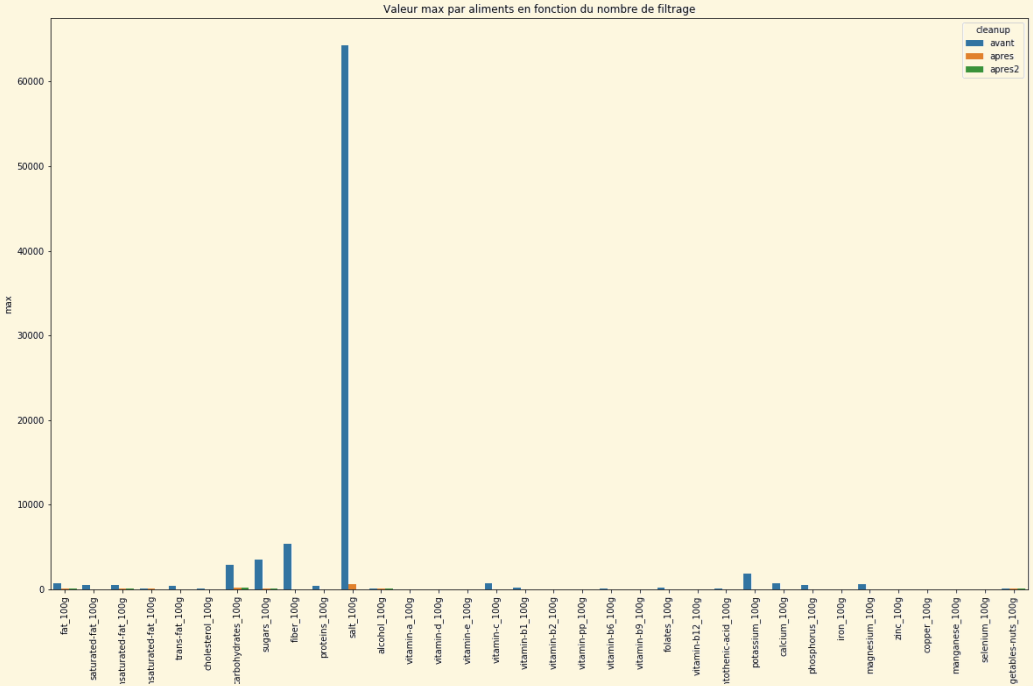
Exploration
- ❑ Analyse Univariée
- ❑ Analyse Multivariée

Exploitation
- ❑ Résultats actuels
- ❑ Modèles possibles

Conclusion

Nettoyage

- Intérêt du filtrage



Nettoyage
- ✓ Dataset original
- ✓ Choix des features
- ✓ Simplification de features
- ➢ Nettoyage des données abérantes

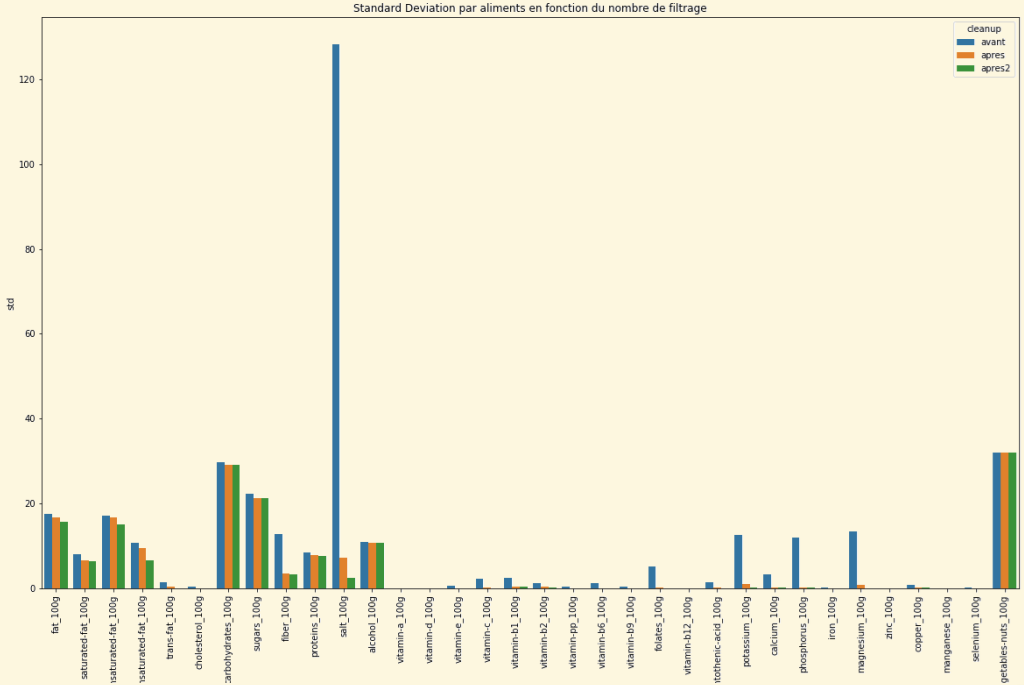
Exploration
- ❑ Analyse Univariée
- ❑ Analyse Multivariée

Exploitation
- ❑ Résultats actuels
- ❑ Modèles possibles

Conclusion

Nettoyage

- Intérêt du filtrage



Nettoyage
- ✓ Dataset original
- ✓ Choix des features
- ✓ Simplification de features
- ➢ Nettoyage des données abérantes

Exploration
- ❑ Analyse Univariée
- ❑ Analyse Multivariée

Exploitation
- ❑ Résultats actuels
- ❑ Modèles possibles

Conclusion

Nettoyage

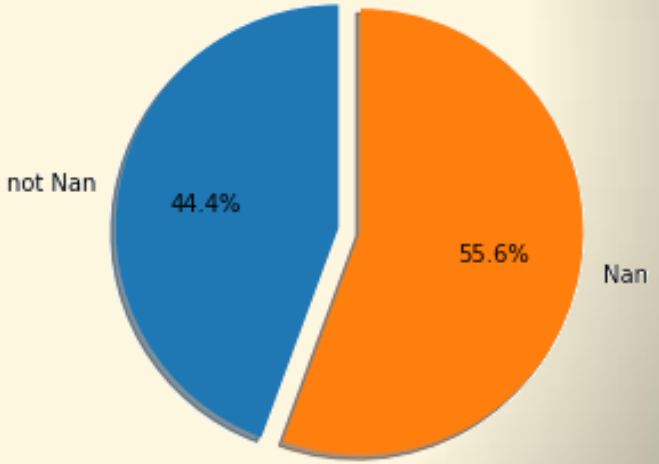
Nettoyage
- ✓ Dataset original
- ✓ Choix des features
- ✓ Simplification de features
- ✓ Nettoyage des données abérantes

Exploration
- ❑ Analyse Univariée
- ❑ Analyse Multivariée

Exploitation
- ❑ Résultats actuels
- ❑ Modèles possibles

Conclusion

- 248463 lignes (-23%)
- 44 features (-73%)
- 56 % are Nan (-19%)
  - 10,93 M values
  - 6,22 M Nan



Exploration

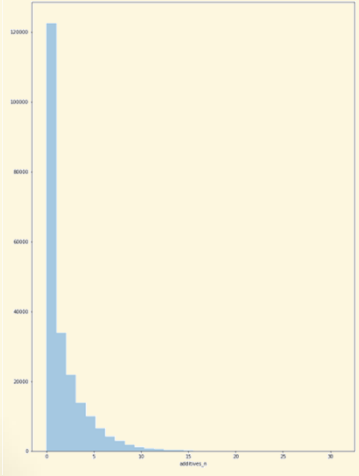
- Histogramme : données catégoriques
- Violinplot : données continues
- Pie : données presque binaires

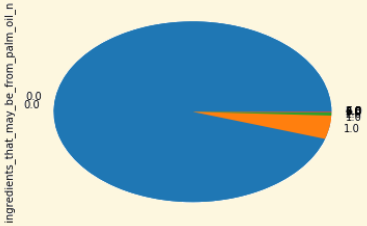
Nettoyage
- ✓ Dataset original
- ✓ Choix des features
- ✓ Simplification de features
- ✓ Nettoyage des données aberrantes

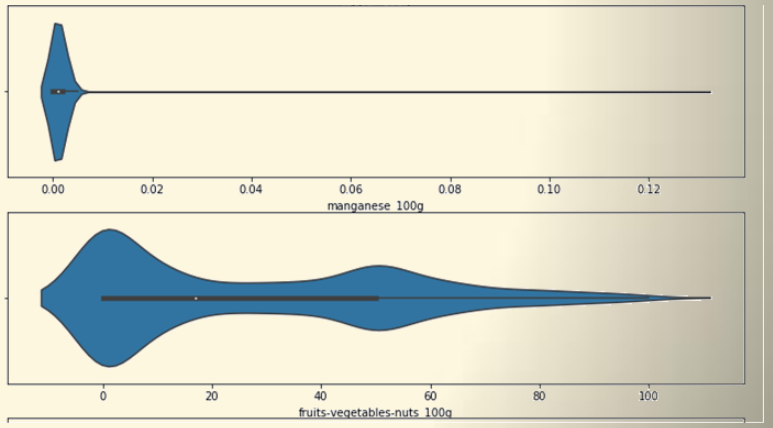
Exploration
- ➢ Analyse Univariée
- ❑ Analyse Multivariée

Exploitation
- ❑ Résultats actuels
- ❑ Modèles possibles

Conclusion







Exploration

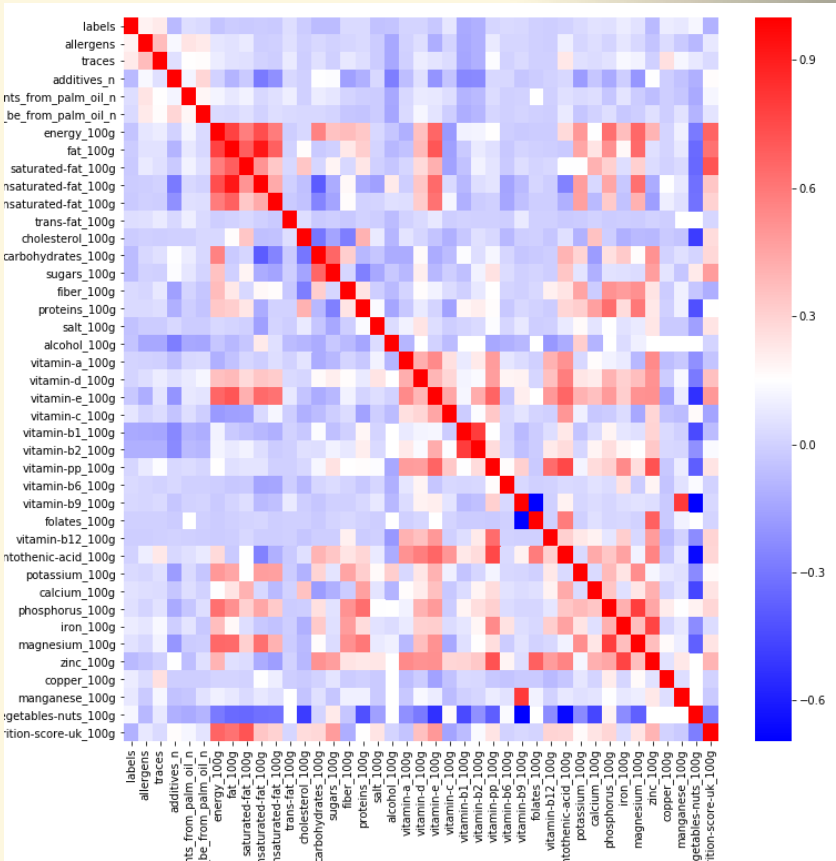
Nettoyage
- ✓ Dataset original
- ✓ Choix des features
- ✓ Simplification de features
- ✓ Nettoyage des données abérantes

Exploration
- ✓ Analyse Univariée
- ➢ Analyse Multivariée

Exploitation
- ❑ Résultats actuels
- ❑ Modèles possibles

Conclusion

- Analyse des Corrélations
  - Vitamine b9 & folates
  - Ligne fruit, veget. & nuts
  - Graisses
  - Vitamine E et graisses
  - Energie et magnésium, potassium, phosphore
  - Sucres et carbohydrates
- Autres analyses



MINE Nicolas

Exploration

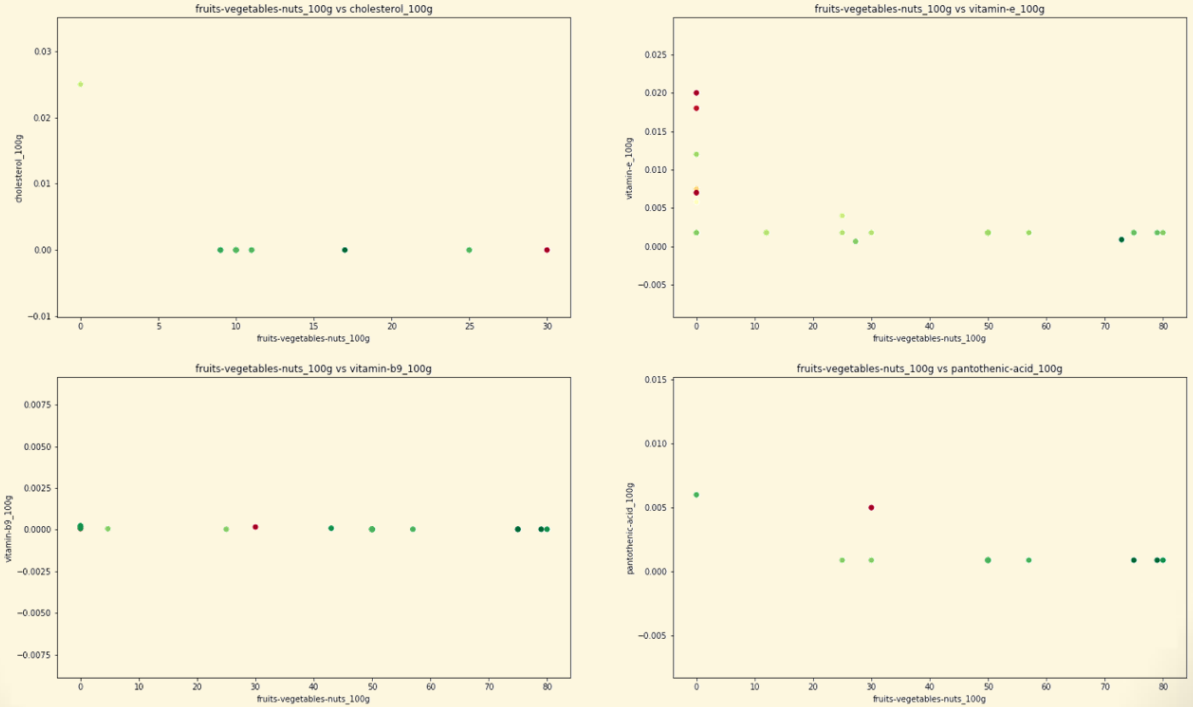
- Ligne fruit, veget. & nuts

Nettoyage
- ✓ Dataset original
- ✓ Choix des features
- ✓ Simplification de features
- ✓ Nettoyage des données abérantes

Exploration
- ✓ Analyse Univariée
- ➢ Analyse Multivariée

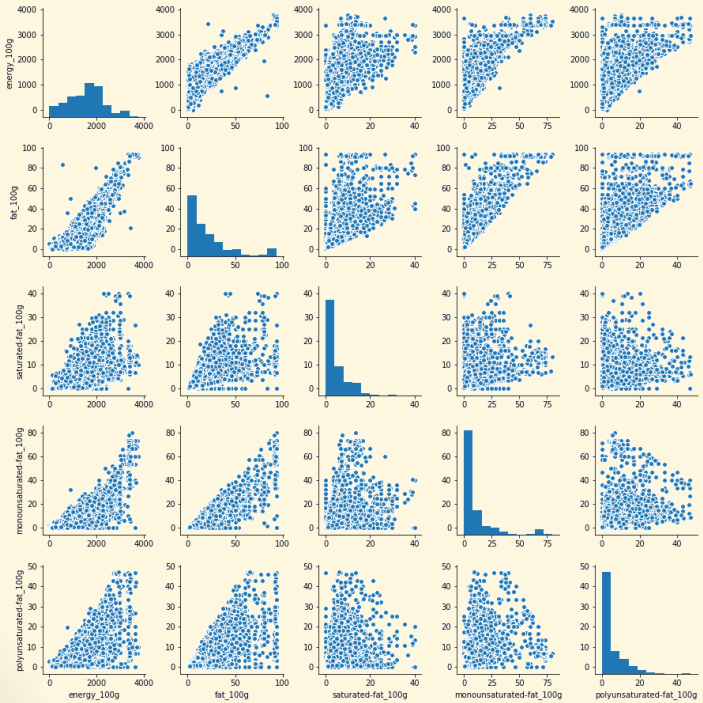
Exploitation
- ❑ Résultats actuels
- ❑ Modèles possibles

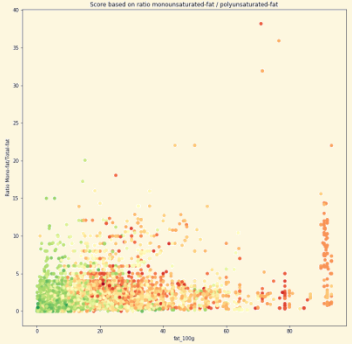
Conclusion



Exploration

- Graisses





Nettoyage
- ✓ Dataset original
- ✓ Choix des features
- ✓ Simplification de features
- ✓ Nettoyage des données abérantes

Exploration
- ✓ Analyse Univariée
- ➢ Analyse Multivariée

Exploitation
- ❑ Résultats actuels
- ❑ Modèles possibles

Conclusion

Exploration

- Vitamine E et graisses

Nettoyage
- ✓ Dataset original
- ✓ Choix des features
- ✓ Simplification de features
- ✓ Nettoyage des données abérantes

Exploration
- ✓ Analyse Univariée
- ➢ Analyse Multivariée

Exploitation
- ❑ Résultats actuels
- ❑ Modèles possibles

Conclusion

La vitamine E ou tocophérol est une vitamine soluble dans les graisses

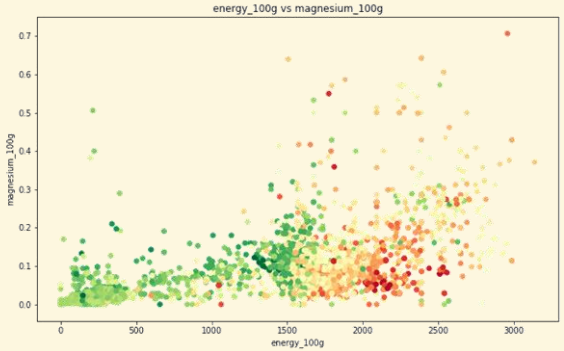
caducee.net

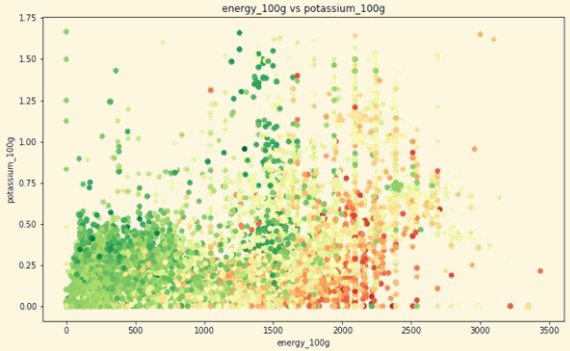
Besoin : 15mg/j

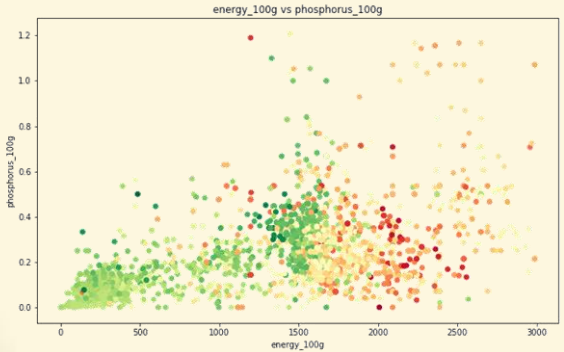
wikipedia.org

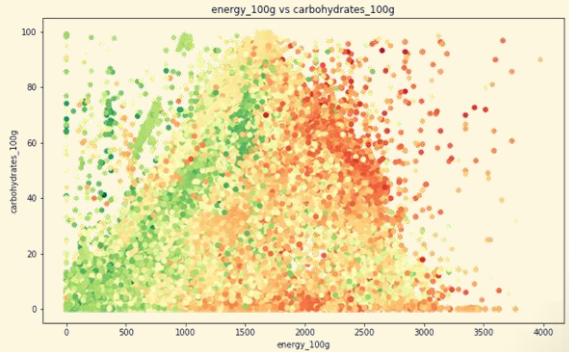
Exploration

- Energie et magnésium, potassium, phosphore, carbohydrates









Nettoyage
- ✓ Dataset original
- ✓ Choix des features
- ✓ Simplification de features
- ✓ Nettoyage des données abérantes

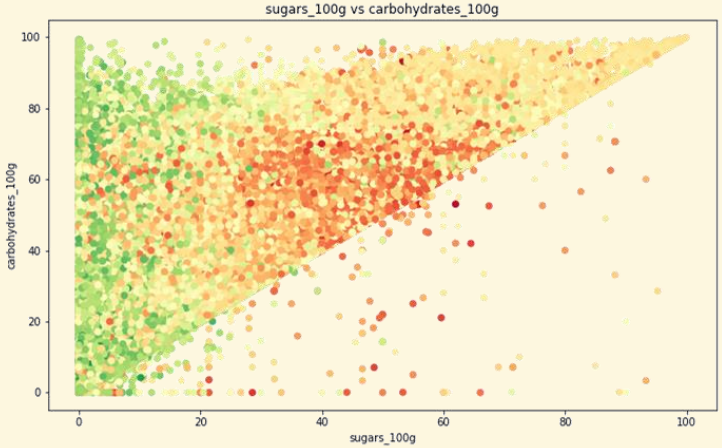
Exploration
- ✓ Analyse Univariée
- ➢ Analyse Multivariée

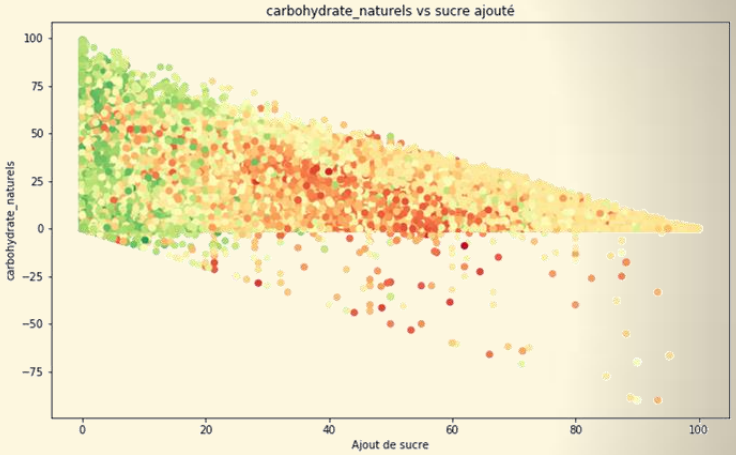
Exploitation
- ❑ Résultats actuels
- ❑ Modèles possibles

Conclusion

Exploration

- Sucres et carbohydrates





Les Carbohydrates proviennent en partie du sucre

Nettoyage
- ✓ Dataset original
- ✓ Choix des features
- ✓ Simplification de features
- ✓ Nettoyage des données abérantes

Exploration
- ✓ Analyse Univariée
- ➢ Analyse Multivariée

Exploitation
- ❑ Résultats actuels
- ❑ Modèles possibles

Conclusion

Exploration

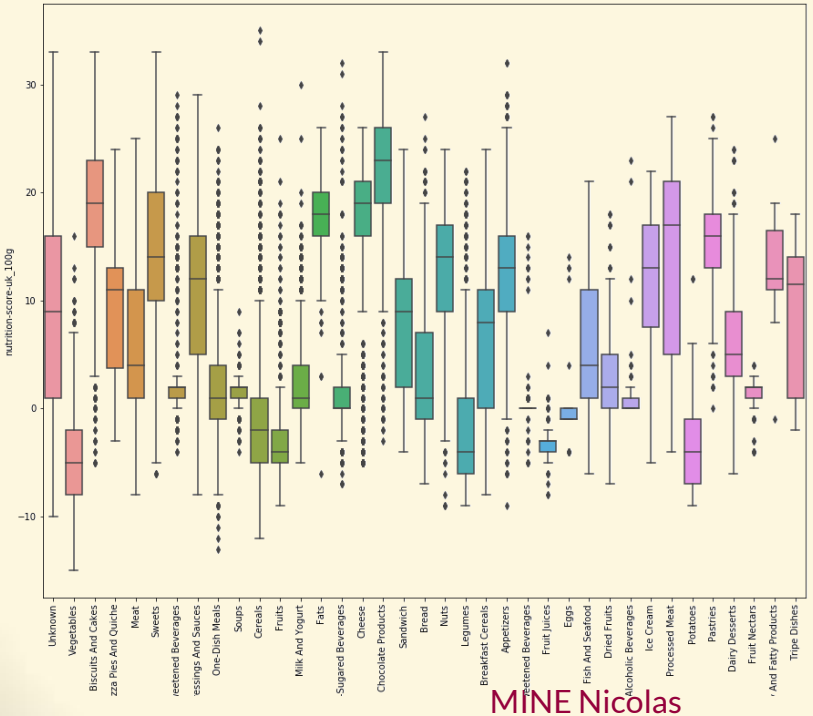
- Autres analyses

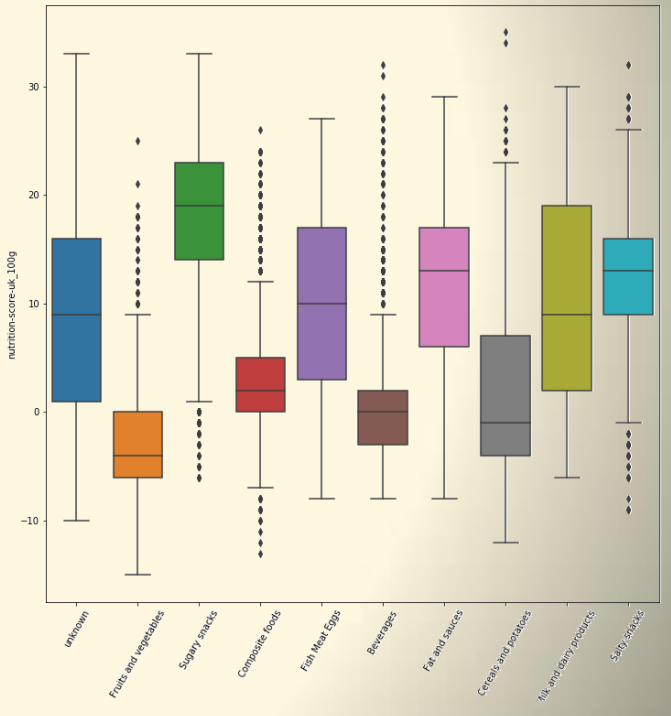
Nettoyage
- ✓ Dataset original
- ✓ Choix des features
- ✓ Simplification de features
- ✓ Nettoyage des données abérantes

Exploration
- ✓ Analyse Univariée
- ➢ Analyse Multivariée

Exploitation
- ❑ Résultats actuels
- ❑ Modèles possibles

Conclusion





Exploration

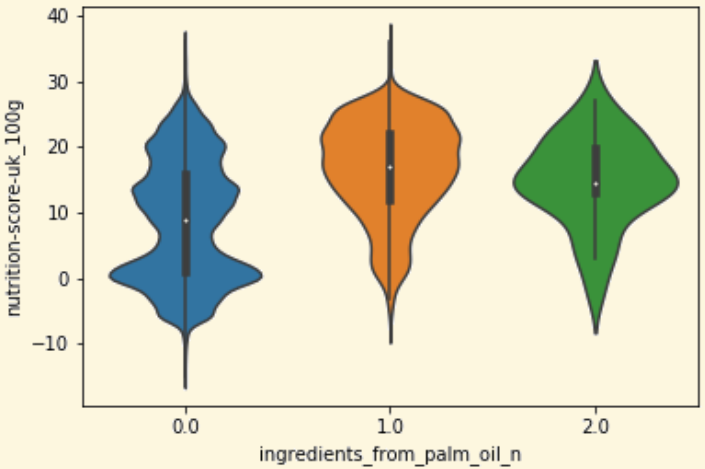
- Autres analyses

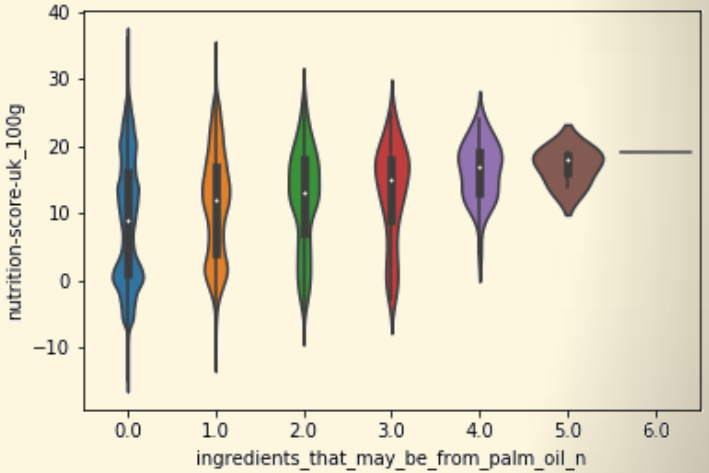
Nettoyage
- ✓ Dataset original
- ✓ Choix des features
- ✓ Simplification de features
- ✓ Nettoyage des données abérantes

Exploration
- ✓ Analyse Univariée
- ➢ Analyse Multivariée

Exploitation
- ❑ Résultats actuels
- ❑ Modèles possibles

Conclusion





Exploration

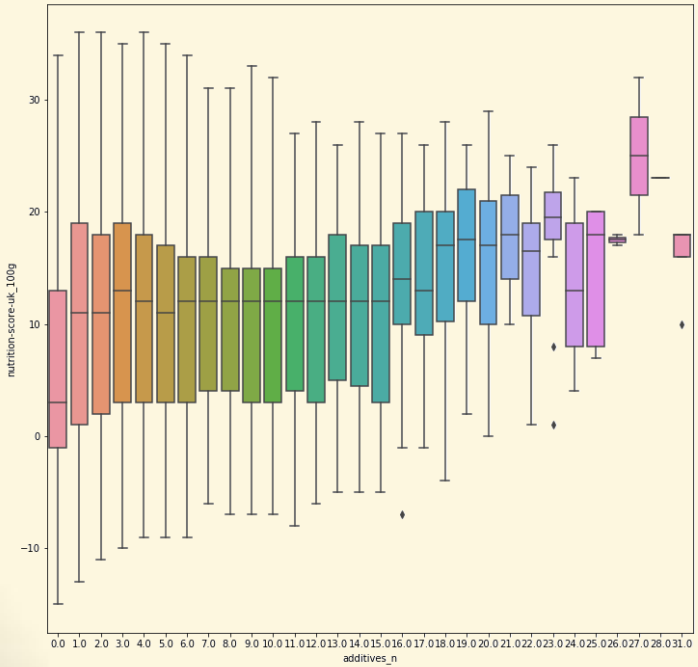
- Autres analyses

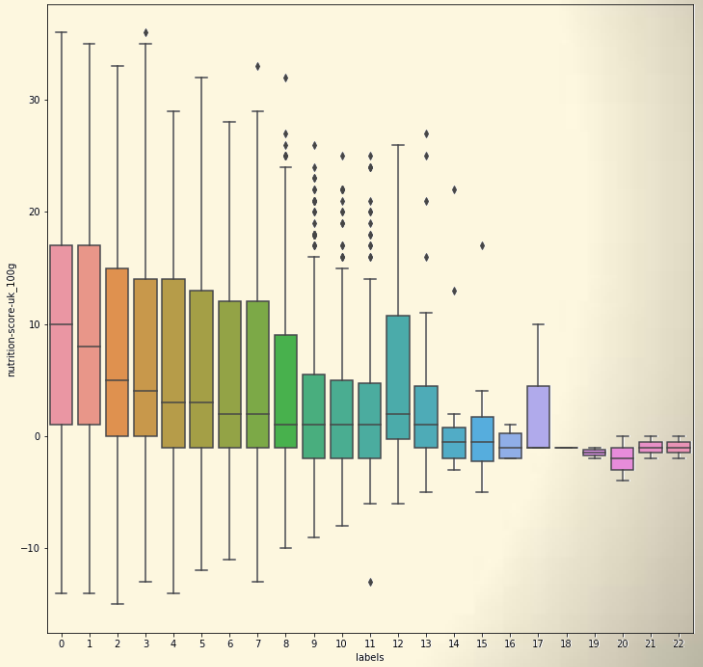
Nettoyage
- ✓ Dataset original
- ✓ Choix des features
- ✓ Simplification de features
- ✓ Nettoyage des données abérantes

Exploration
- ✓ Analyse Univariée
- ➢ Analyse Multivariée

Exploitation
- ❑ Résultats actuels
- ❑ Modèles possibles

Conclusion





Exploration

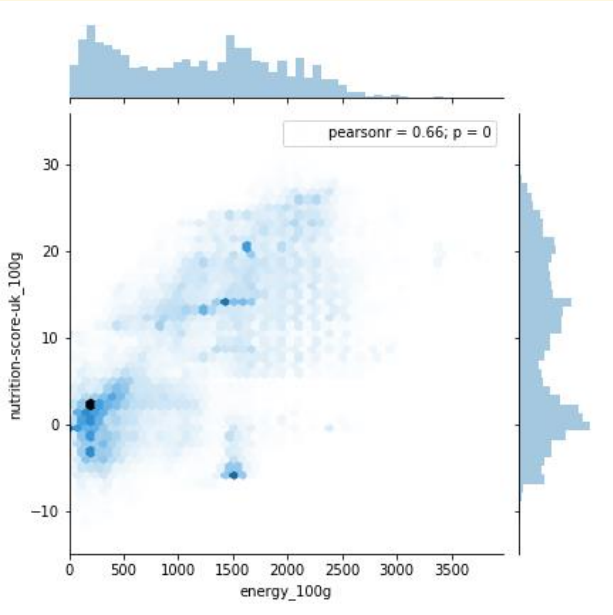
- Autres analyses

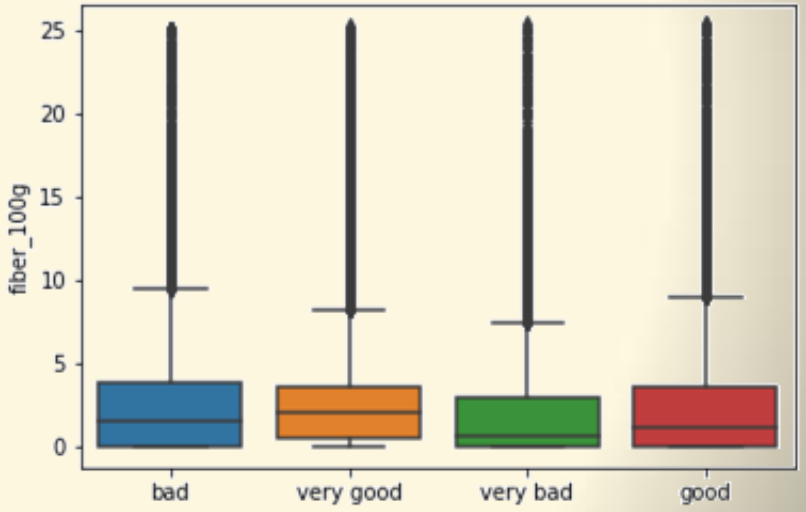
Nettoyage
- ✓ Dataset original
- ✓ Choix des features
- ✓ Simplification de features
- ✓ Nettoyage des données abérantes

Exploration
- ✓ Analyse Univariée
- ➢ Analyse Multivariée

Exploitation
- ❑ Résultats actuels
- ❑ Modèles possibles

Conclusion





Exploitation

- Nettoyage correct
  - Possibilité d'agréger plusieurs features
    - Vitamines et graisses
  - Essayer de récupérer de la donnée
  - Ajouter les apports journaliers

- Synthèse
  - Energie lié à la graisse (privilégier graisses non saturées)
  - Produit naturels (peu d'additifs, allergènes, avec labels)
  - Privilégier fruits, légumes, œufs, céréales
  - Éviter sucres ajoutés, gâteaux, sodas, fromages
  - Peu d'infos sur les vitamines

Exploitation

Nettoyage
- ✓ Dataset original
- ✓ Choix des features
- ✓ Simplification de features
- ✓ Nettoyage des données abérantes

Exploration
- ✓ Analyse Univariée
- ✓ Analyse Multivariée

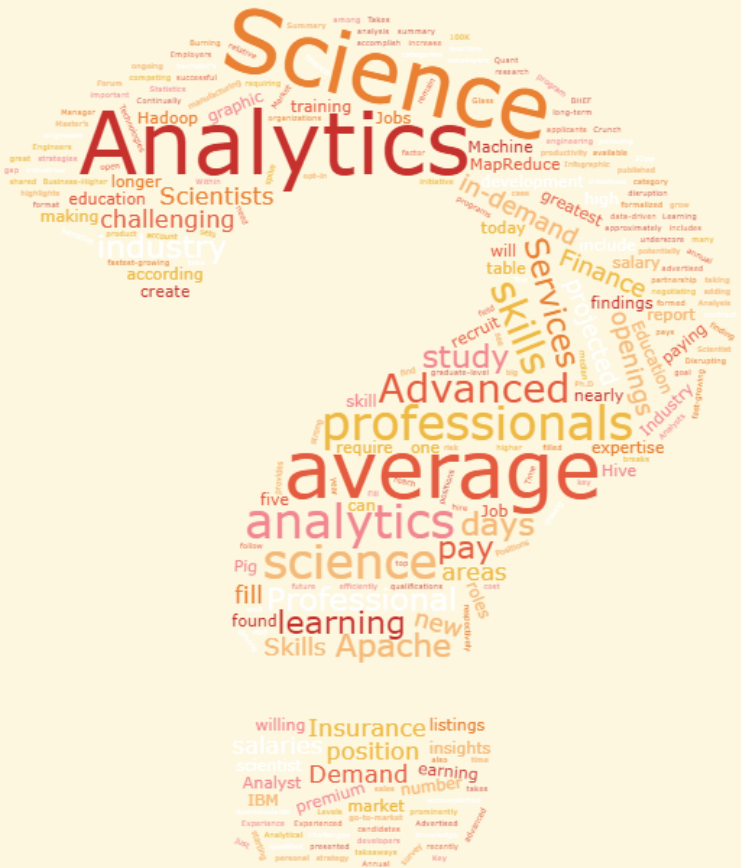
Exploitation
- ✓ Résultats actuels
- ➢ Modèles possibles

Conclusion

- Problèmes :
  - Points manquants (imputer KNN, mean, 0, …) ?
- Régression (Score = f(Features))
  - Modèles linéaire (SGD Regressor)
  - Régression par proximité (KNN Regressor)
  - Modèle non linéaire (Linearisation, SVM)
- Classification (bon/pas bon)
  - Possibilité de multi-classes (OVA, OVO)
  - Modèle linéaire (Reg. Logistique)
  - Régression par proximité (KNN Classifier)
  - Modèle non linéaire (Linearisation, SVC)

Conclusion

- Nettoyage correct mais améliorable
- Certains critères ont été extraits
  - Produit labélisés
  - Peu d'additifs (naturels)
  - Peu gras et calorifique (fruits, légumes, pommes de terres, …)
- Bémol :
  - reversed engineering plus que feature engineering


Science
Analytics
Hadoop
graphic
training
Jobs
Machine
MapReduce
Scientists
longer
education
making
challenging
industry
according
create
in-demand
today
will
table
Services
skills
Finance
salary
findings
report
paying
openings
Education
Industry
study
Advanced
skill
nearly
professionals
require
one
expertise
Hive
average
five
can
Job
analytics
days
pay
Pig
science
areas
fill
Professional
roles
found
learning
new
Skills
Apache
willing
Insurance
listings
salaries
position
insights
Demand
earning
Analyst
number
premium
market
IBM