#### Introduction à la Data Science avec R

Fousseynou Bah

12-Jun-2020

- Définir la data science
- 2 Pourquoi développer du contenu pour un textbook?
- R
- RStudio

#### Section 1

#### Définir la data science

### Commençons par quelques exemples

#### Fait de la data science:

- l'économiste qui examine le niveau du PIB sur 30 ans et cherche à dégager des scénarii pour des futures évolutions
- le sociologue qui s'appuie le taux de natalité et le taux de participation des femmes au marché du travail pour comprendre l'évolution de la place de la femme dans la société
- le météorologue qui cherche à prédire la pluviométrie de la semaine à venir en modélisant les données historiques
- l'épidémiologue qui cartographie le taux de prévalence du paludisme pour appuiyer un programme stratégique
- ..

Tout ça, pour dire que la *data science* est une discipline transversale. Elle se pratique dans plusieurs domaines.

#### Une discipline carrefour

Selon Wikipédia, la data science est un champ interdisciplinaire qui utilise les méthode, processus, algorithmes et systèmes scientifiques pour extraire des données - tant structurées que non structurées - des informations utiles à la compréhension et à la prise de décision. De ce fait, elle s'appuie sur diverses méthodes (mathématiques, statistiques, informatiques, etc.) pour tirer des données une compréhension meilleure de phénomènes d'intérêt.

#### Le data scientist, le métier du 21eme siècle

Face à la génération massive des données, dans tous les secteurs, le besoin de *data scientist* se fait pressant.

Qu'est-ce qu'un data scientist? Quelqu'un qui pratique la data science, évidemment! Mais plus que ça, c'est un détective des données, quelqu'un qui investigue les données en vue de découvrir des récurrences, de reveler des éléments surprenants ou tout simplement de prendre la mesure des faits déjà connus…et tout ça dans le dessein de guider la prise de décision.

Comme tout bon détective, il se doit de posseder ou de cultiver un certains nombres de compétences. Quelles sont celles-ci?

### Ce qu'il faut pour être data scientist

- Pas nécéssairement un diplôme avancé en mathématiques ou en statistiques...quoiqu'il est utile de maîtriser des concepts de bases (vecteur, matrice, moyenne, écart-type, etc.)
- Pas forcément un diplôme en informatique ou en programmation...quoiqu'il est utile de connaître les notions de bases (qu'est-ce qu'un objet, un environnement? quels types d'objets peut-on manipuler dans un environnement donné...)
- Une connaissance avérée dans un domaine spécifique dans lequel l'on peut soulever des questions, mobiliser des outils théoriques auxquels on confronte les résultats de l'analyse conduite sur les données
- Un esprit curieux, quelle que soit l'avenue que l'on emprunte.

Vous pourrez avoir une meilleure idée en surfant sur le net. Pour commencer, regardez ici.

#### Section 2

Pourquoi développer du contenu pour un textbook?

#### Un abondance étourdissante

L'Internet compte un nombre indénombrable de ressources pour apprendre la data science. Pour s'en rendre compte, il faut de rechercher le terme dans Google (ou tout autre moteur de recherche...dont il faut chercher le nom dans Google).

Les résultats sont aussi nombreux que:

- les programmes et langages se prêtant à la data science;
- les questions de recherche sollicitant ses outils.

De ce fait, on peut se demander: "pourquoi faire un autre textbook ou page sur Internet?"

#### Réponse 1 : le contexte personnel

Pour s'assurer qu'on a compris une chose, il n'y a rien de mieux comme test que de tenter de l'expliquer à quelqu'un autre. Une idée assez datée maintenant. Il suffit de se référer à la désormais très célèbre technique de Feynman pour se rendre compte de l'importance de l'enseignement dans le processus d'apprentissage.

Ce exercice répondait dont à un impératif personnel. Il me semblait utile de me lancer dans l'élaboration d'outils pédagogiques pour tester ma propre maîtrise des notions apprises dans le cadre de mes cours en data science.

J'espère que les lecteurs partageront cet avis et accepteront de se livrer à un exercice similaire.

#### Réponse 2: le contexte géographique

Les données qui servent à illustrer les concepts ou techniques présentés sont souvent celles d'un monde étranger l'étudiant africain: les pays industrialisés. Les données des banques centrales, les données de recensement ou d'enquête, ou encore les données des marchés financiers...de ces pays. Il me paraissait d'utile de mobiliser autant que possible des données locales pour contruire les illustration de cet nième \*textbook\* en data science.

#### Section 3

R

#### Qu'est-ce que c'est que R?

R est un langage de programmation et un logiciel gratuit et libre. Il est surtout utilisé pour le développement de programmes statistiques et des analyses de données. Il gagne en popularité depuis quelques années avec l'émergence de la *data science* et du fait qu'il est gratuit et ouvert (*open-source*).

R est née d'un projet de recherche mené par deux chercheurs, Ross Ihaka et Robert Gentleman à l'université d'Auckland (Nouvelle-Zélande) en 1993. En 1997 est mis en place le *Comprehension R Archive Network (CRAN)* qui centralise les contributions au projet. Depuis le projet connaît une croissance soutenue, grâce à des contributions de la part de milliers de personnes à travers le monde.

#### Pourquoi R?

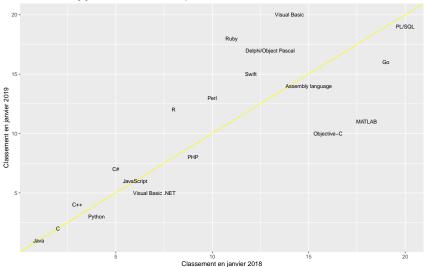
Pour un apprenti data scientist, le choix du langage et/ou du programme est une décision critique. Considérant le temps qu'il investira en apprentissage et le retour qu'il espéra à travers l'utilisation de ses nouvelles connaissances dans sa profession, il est utile de considerer divers critères dont:

- l'accessibilité de l'outil en termes de coûts (tout n'est pas gratuit comme R!);
- l'accessibilité du langage en termes de syntaxe;
- la popularité du langage parmi ses paires (en vue de faciliter des collaborations);
- la dynamique de développement du langage.

Et sur tous ces points, R présente des avantages.

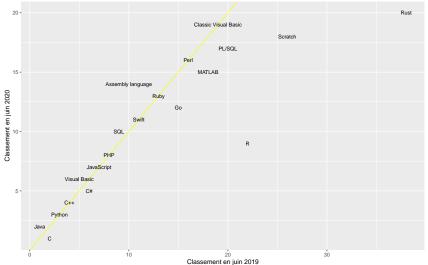
## R dans l'écosystème des langages: TIOBE Index (1)

TIORE Index Classement des langages sur la base de divers moteurs, janvier 2019



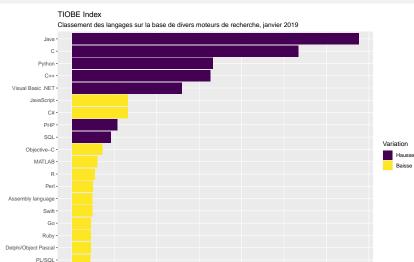
## R dans l'écosystème des langages: TIOBE Index (2)





Introduction à la Data Science avec R

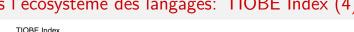
# R dans l'écosystème des langages: TIOBE Index (3)

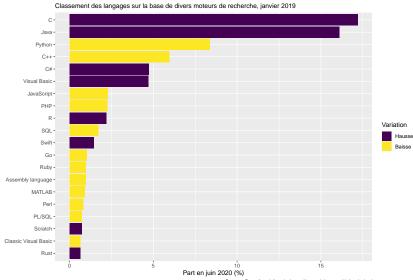


Visual Basic -

Part en janvier 2019 (%)

## R dans l'écosystème des langages: TIOBE Index (4)

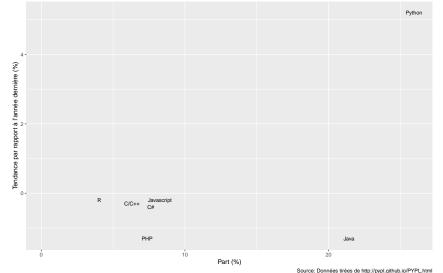




## R dans l'écosystème des langages: PYPL Index (1)

PopularitY of Programming Language Index

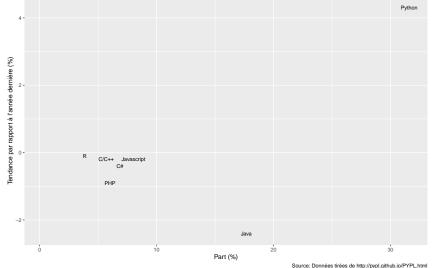
Classement des langages sur la base des recherches dans Google, janvier 2019



## R dans l'écosystème des langages: PYPL Index (2)

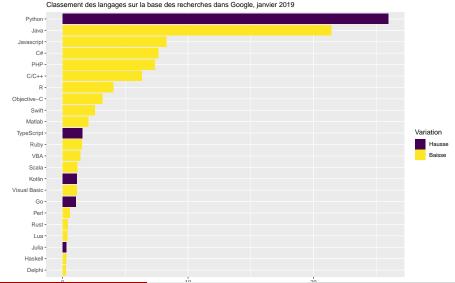
PopularitY of Programming Language Index

Classement des langages sur la base des recherches dans Google, juin 2020

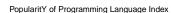


## R dans l'écosystème des langages: PYPL Index (3)

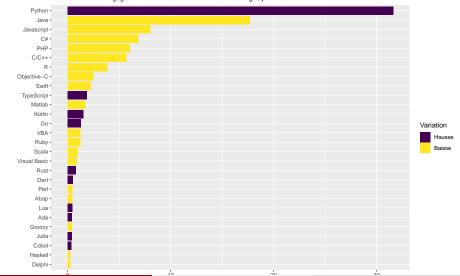




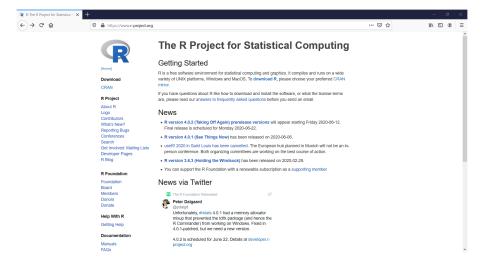
## R dans l'écosystème des langages: PYPL Index (4)



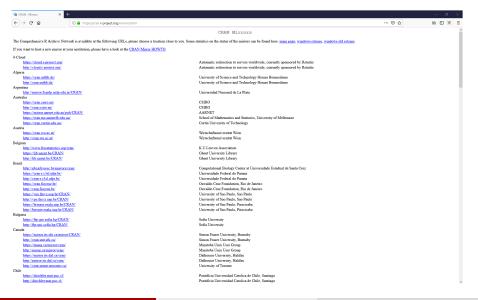
Classement des langages sur la base des recherches dans Google, juin 2020



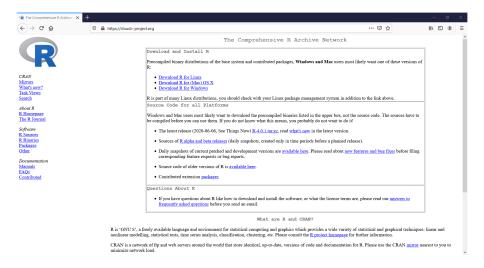
## Télécharger R (1)



## Télécharger R (2)



## Télécharger R (3)



## Télécharger R (4)

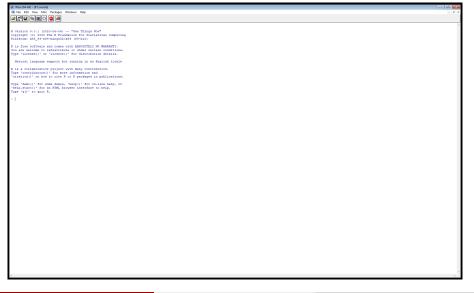


Manuals FAQs Contributed

## Télécharger R (5)



### Un aperçu de R: GUI



### Un aperçu de R: Terminal

```
R is free software and comes with ABSOLUTELY NO MARRAWITY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
 'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
```

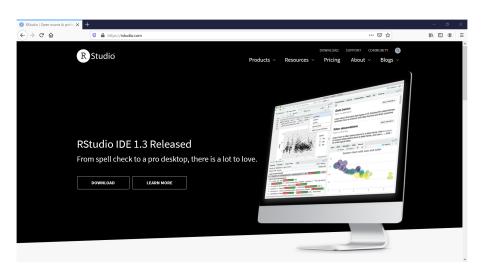
#### Section 4

#### **RStudio**

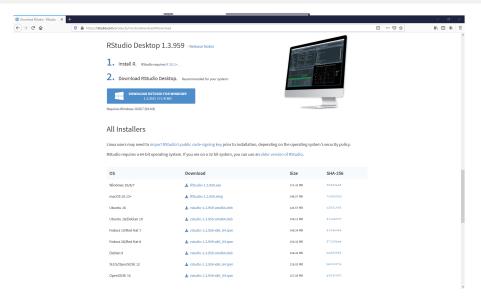
#### Qu'est-ce que c'est que RStudio

- C'est une IDE (*Integrated Development Environment*) ou Environnement Intégré de Développement
- Il sert d'interface entre R et l'utilisateur, offre à celui diverses commodités d'utilisation

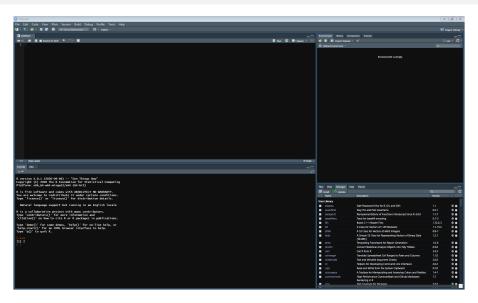
### Télécharger RStudio (1)



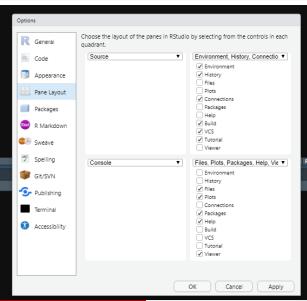
# Télécharger RStudio (2)



## Un aperçu de RStudio (1)



## Un aperçu de RStudio (2)



#### Et on démarre

Maintenant, vous avez les outils nécéssaires pour commencer la formidable aventuRe!