Introduction à la programmation en R
M1 ROAD & M1 IREF
Automne 2024

Devoir terminal

Laurent R. Bergé*

Date limite de rendu. 30 novembre 2024 à 23h59 (*nota : possibilité de décaler deux semaines plus tard si le professeur est prévenu avant le 23 novembre et si la classe est unanime*).

Groupes. Devoir à faire par groupe de 5 maximum et 4 minimum.

Attendus. Vous devrez rendre :

1. un ou plusieurs documents R, voir détails ici
2. un site web hébergé sur Github qui contiendra la présentation des résultats, voir détails ici

Rendu.

- Vous devez absolument rendre les documents R via Moodle (ici), les rendus par mail ne sont pas considérés.
- 1 point de pénalité par 12h de retard (les premières 12h sont gratuites).

Accès aux données.

- Accèdez aux données via la page Moodle du cours.

Notation.

6 points	*Traitement BDD brevets*
6 points	*Traitement BDD offres d'emploi*
2 points	*Appariement des deux BDD*
2 points	*Création de site web*
2 points	*Statistiques descriptives*
3 points	*Analyse des données*
Total :	*21 points*

*B$_x$SE, UMR CNRS 6060, Université de Bordeaux, *email*: laurent.berge@u-bordeaux.fr

1. Objectif

L'objectif de ce projet est de comprendre le lien entre la performance d'innovation des entreprises (mesurée par les dépots de brevets) et leur demande de compétences.

Voici un exemple de questions auxquelles on va essayer de répondre:
Est-ce que ce sont les entreprises les plus innovantes:

- qui paient le mieux ?
- demandent le plus de data scientists ?
- demandent le plus de compétences en machine learning ?
- etc

2. Première partie: Traitement des données

2.1. Format attendu

L'entièreté de cette partie doit être contenue dans un ou plusieurs fichiers R.

Vous pouvez rendre:

- **Soit** un seul fichier principal nommé `xxxx_main.R` avec `xxxx` les initiales de chaque personne personne. Par exemple si Dupont et Gautier travaillent ensemble, ils rendent le fichier R `dg_main.R`.
- **Soit** un fichier pour chaque partie nommés: `xxxx_brevet.R`, `xxxx_emploi.R` et `xxxx_match.R`. De même qu'au-dessus `xxxx` correspond aux initiales des noms des membres du groupe.

Si vous créez des fonctions que vous utilisez dans vos fichiers principaux ci-dessus:

- Toutes les fonctions doivent se localiser dans un fichier nommé `xxxx_src_utilities.R` (avec `xxxx` comme au dessus), pas dans le fichier de code principal. Dans les fichiers qui utilisent ces fonctions, mettez `source("xxxx_src_utilities.R")` au début du code pour charger ces fonctions.

Ces fichiers doivent avoir les caractéristiques suivantes:

- Tout en haut de **chaque** document R, il faudra mettre, en commentaire, les noms et prénoms de chaque personne du groupe.
- Le code est censé fonctionner si je le fais tourner sur mon ordinateur.
- Toutes les sources de données devront se trouver dans le répertoire "`DATA`". La présence d'un chemin de fichier absolu entraînera une pénalité de 5 points.

2.2. Conseils

- si besoin sauvegardez des fichiers intermédiaires
- il y a plusieurs façons de répondre aux questions. Certaines sont plus malines que d'autres. Réfléchissez avant d'agir.

2.3. Brevets

Lisez bien toutes les consignes avant de vous lancer.

6 pts Q1. Créez la base de données nommée `base_brevets` qui contient les variables suivantes:

`firm_name`	nom de l'entreprise
`n_patents`	nombre de brevets
`ipc_main_code`	code de la classe IPC principale de l'entreprise
`ipc_main_desc`	code de la classe IPC principale de l'entreprise
`ipc_second_code`	code de la classe IPC principale de l'entreprise
`ipc_second_desc`	code de la classe IPC principale de l'entreprise
`addr_city_main`	ville principale de l'entreprise
`addr_dept_main`	département principal de l'entreprise

Cette base ne doit contenir qu'une seule ligne par entreprise.

2.3.1. Détails sur le création de la base

Restrictions. Cette base de données concernera seulement:
- les entreprises françaises (dont le pays de l'*applicant* est la France)
- les brevets qui sont déposés entre 2010 et 2020

IPC codes. Les codes IPCs sont une classification hiérarchique des brevets qui permet d'identifier très finement sur quoi porte l'invention du brevet. Ces codes contiennent 14 caractères, un exemple est donné en dessous.

Exemple

	code	signification
IPC-1	G	Physics
IPC-3	G06	Computing; calculating or counting
IPC-4	**G06T**	**Image data processing or generation, in general**
IPC-8	G06T0007	Image analysis
IPC-14	G06T0007000143	... involving probabilistic approaches

Les codes IPCs que vous trouverez sur les brevets sont complets, c'est à dire font 14 caractères: IPC-14. Dans ce projet, on ne sera intéressé que par les codes IPC à 4 caractères (en gras dans la table au dessus). On l'appellera l'IPC-4.

La variable `ipc_main_code` correspond, pour une entreprise donnée, à l'IPC-4 le plus fréquent dans son portefeuille de brevets.

Exemple

(Exemple fictif.) Entre 2010 et 2020 l'entreprise Renault a déposé 1500 brevets. Parmi ceux-ci, et dans l'ordre: 900 brevets contiennent l'IPC `"H01M"`, suivi de 500 qui contiennent `"B60L"`, suivi de 200 qui contiennent `"F16F"`, etc.

Ici on aura:
- `ipc_main_code = "H01M"`
- `ipc_main_desc = "Processes or means, e.g. batteries, for the direct conversion of chemical energy into electrical energy"`
- `ipc_second_code = "B60L"`

- `ipc_second_desc = "Propulsion of electrically-propelled vehicles"`

Attention: un brevet contient en général plusieurs IPC-14 différents, et donc plusieurs IPC-4 par brevet sont possibles.

Addresse. Un même entreprise peut tout à fait avoir des adresses différentes. Par exemple Renault a plusieurs établissements en France, ceux-ci ont naturellement différentes adresses.

Les variables `addr_city_main` et `addr_dept_main` donneront la ville et le département le plus fréquent dans le portefeuille de brevet d'une entreprise.

Identification des entreprises. Dans ce projet on identifiera les entreprises par leur nom. Plusieurs problèmes existent.

Le premier problème est que deux entreprises différentes peuvent avoir le même nom: *vous ignorerez complètement ce problème.*

Le deuxième problème, plus important, est que deux entreprises identiques peuvent avoir des noms différents reportés sur les brevets. Par exemple: "Renault" et "Renault, SA" ont des noms différents mais il s'agit bien de la même entreprise.

Soyez conscients de ces problème mais ne vous cassez pas trop la tête dessus car la vrai identification d'entreprise est très complexe. Donc faites en un peu, mais pas trop.

2.3.2. Sources de données

Les données proviennent de deux sources d'information principales :
- l'OCDE, qui fournit plusieurs bases de données sur les brevets déposés au sein de l'European Patent Office depuis les années 1980
- la liste officielle des dénominations de la classification internationale des brevets (IPC)

Les données que je vous donne sont les données brutes. Ci-dessous je ne liste que les variables importantes.

1. `EPO_APP_REG.txt`
   - **`Appln_id`** identifiant unique des brevets
   - **`App_name`** nom de l'entreprise aui a déposé le brevet
   - **`City`** nom de la ville de l'entreprise qui a déposé le brevet
   - **`Postal_code`** code postal de l'entreprise qui a déposé le brevet
   - **`Ctry_code`** pays de l'entreprise qui a déposé le brevet
2. `EPO_IPC.txt`
   - **`Appln_id`** identifiant unique des brevets
   - **`Prio_year`** année où le brevet a été déposé
   - **`IPC`** code IPC-14, c'est à dire le code IPC complet
3. `EN_ipc_section_A_title_list_20120101.txt to EN_ipc_section_H_title_list_20120101.txt`
   - **`Column_1`** the IPC code, it can be IPC-1 to IPC-14
   - **`Column_2`** definition, in English, of the IPC code

2.4. Offres d'emploi

Lisez bien toutes les consignes avant de vous lancer.

6 pts Q2. Créez la base de données nommée `base_emp` contenant les variables suivantes:

`firm_name`	nom de l'entreprise postant les offres d'emploi
`n_offres`	nombre d'offres d'emploi proposées par l'entreprise
`sector_main`	secteur principal d'activité de l'entreprise
`avg_req_exp`	experience requise moyenne pour les postes offerts
`top_skill_req`	compétence demandée qui revient le plus au sein de toutes les annonces de l'entreprise
`avg_wage`	salaire annuel moyen des offres proposées, doit être une variable numérique
`addr_dept_main`	département principal de l'entreprise

Cette base ne doit contenir qu'une seule ligne par entreprise.

2.4.1. Détails sur la création des variables

Harmonisation du texte. Les données que vous utiliserez ici sont des données brutes issues d'internet. Les variables nécessiteront d'être harmonisées. L'exemple type est le nom d'une entreprise écrit légèrement différemment (ex: `"Renault"` vs `"Renault SA"`), ou de deux compétences qui diffèrent à un `s` près (ex: `"Statistique"` et `"Statistiques"`). Attention: on peut passer un temps infini à harmoniser ce genre de chose. Ce que je demande c'est que vous montriez que vous avez vu le problème et que vous savez faire (un peu). Je ne m'attends pas du tout à ce que ce soit parfait.

Expérience. Au sein de toutes les offres d'une même entreprise, quelle est l'expérience moyenne demandée. Ignorez les valeurs manquantes. Si toutes les valeurs sont manquantes, alors cette variable est manquante.

Compétences. Les compétences demandées sont au format suivant, ex: `"SQL, Spark, Git, Database, équipe, Esprit Critique, Collaboration"`. C'est à dire que chaque compétence est séparée par une virgule.

Les entreprises proposent plusieurs offre d'emploi. La variable `top_skill_req` rapporte les compétences qui apparaissent le plus au sein de toutes les offres.

Exemple

Si une entreprise a deux offres d'emploi qui listent les compétences suivantes:

- `"SQL, Spark, Git, Database, équipe, Esprit Critique, Collaboration"`
- `"SQL, Statistique, Power BI, Collaboration"`

Alors la variable `top_skill_req = "SQL, Collaboration"`. L'ordre des compétences au sein de la chaine de caractères n'a pas d'importance.

Salaire. Le salaire est rempli par les entreprises dans un format libre. Ex: `"Salaire : 55K à 60K€"` ou `"50 000 - 63 000 EUR par an"`. Il faudra convertir au format numérique. Les étapes sont les suivantes:

1. transformer la chaine de caractère pour qu'elle affiche uniquement un nombre
2. convertir la chaine de caractère en numérique

A noter: dans les exemple du dessus il y a une fourchette de salaire. Vous pouvez vous contenter de ne prendre qu'un des deux nombres. Calculer la moyenne est mieux mais vous n'êtes pas obligé de le faire pour avoir le maximum de points.

2.4.2. Sources des données d'offre d'emploi

Ces données ont été collectées lors d'un projet webscrapping mené par Maxime Goutte, Yoann Pull, Louis Quenault et Danny Morgant de la promotion M2-IREF 2023-2024.[1]

Ce sont des offres d'emploi correspondant à la recherche de `"data scientist"` repértoriées dans deux sites bien connus listant des offres d'emploi.

Toutes les données sont contenues dans la base suivante:

4. `base_emp_fmt.tsv`

`intitule_poste`	intitulé du poste dans l'annonce
`entreprise`	nom de l'entreprise
`type_emploi`	type de contrat proposé
`secteur`	secteur donné par l'entreprise
`experience_requise`	années d'expérience requises
`competences_requises`	ensemble de compétences demandées
`poste_desc`	description du poste (cette variable contient beaucoup de texte)
`salaire`	salaire proposé par l'entreprise et affiché publiquement sur l'offre
`departement`	département où se situerait le lieu de travail

2.5. Appariement des deux bases de données

2 pts Q3. Créez la base de données `base_emp_inno` qui est l'appariement des bases `base_brevets` et `base_emp`.

Ne perdez pas d'information ni sur les entreprises qui brevetent, ni sur les entreprises qui proposent des offres d'emploi.

3. Deuxième partie : Résultats

Important. Dans cette partie le code R ne sera pas examiné. Seul le site web et son contenu sera noté.

3.1. Création de site web

2 pts Q4. Créez un site web hébergé sur Github qui contiendra les résultats de votre analyse.

3.1.1. Détails

Dans cette partie, vous devrez créer une page web hébergée sur Github avec les caractéristiques suivantes :

- elle doit contenir au moins trois pages:
  1. une page d'accueil
  2. une page sur les statistiques descriptives
  3. une page sur l'analyse de données
- la page d'accueil contiendra, au moins :
  - un message d'accueil
  - les noms et prénoms de chaque membre du groupe
  - les liens vers les deux autres pages

[1] C'est dans le cours `Big Data Tools`, des M2-IREF parcours ERDS.

- bien sûr c'est mieux si c'est joli !

3.2. Statistiques descriptives

2 pts Q5. Sur votre site web, créez une page qui reportera des statistiques descriptives pour chaque base de donnée.

Pour les variables numériques, reportez:
- min, médiane, max, moyenne, écart-type, nombre de manquants

Pour les variables chaîne de caractère:
- reportez les 5 valeurs qui ont le plus grand nombre de brevets/plus haut salaire
- ex: dans la `base_brevets` pour la variable `addr_city_main`, calculez le nombre total de brevets par ville et reportez les 5 ville avec les plus de brevet

Pour la base `base_brevets`, reportez les statistiques pour les variables suivantes:
- numérique:
  1. `n_patents`
- caractère (5 valeurs ayant le plus de **brevets**):
  1. `firm_name`
  2. `ipc_main_desc`
  3. `addr_city_main`
  4. `addr_dept_main`

Pour la base `base_emp`:
- numérique:
  1. `n_offres`
  2. `avg_wage`
- caractères (5 valeurs ayant le plus haut **salaire moyen**):
  1. `firm_name`
  2. `sector_main`
  3. `addr_dept_main`

Pour la base `base_emp_inno`:
- numérique:
  1. `n_patents`
  2. `n_offres`
  3. `avg_wage`

Ici la notation est 0.2 point par variable bien faite.

3.3. Analyse des données

3 pts Q6. Sur votre site web, créez une page qui reportera des relations entre l'innovation des entreprises et la demande de compétences en data science.

Ici vous pouvez faire ce que vous voulez. Plus c'est créatif et joli (les graphiques) et mieux c'est.

Un exemple de choses possibles:
- graphiques bivariés:
  - innovation vs: 1) salaire moyen, 2) nombre d'offres d'emploi, etc
  - les mêmes graphiques d'au dessus mais au niveau du secteur et pas de l'entreprise
  - idem mais pour au niveau des codes IPCs (est-ce qu'il y a des codes IPCs spécifiques pour lequel la demande est forte?)

    - idem mais au niveau des mots-clefs des compétences demandées (y a-t-il des compétences associées avec plus d'innovation des entreprises?)
- estimations économétriques simples
- autres graphiques:
    - word cloud des compétences, pondérées par le salaire
    - word cloud des compétences, pondérées par les brevets
    - etc !

Ici la notation est : 1 point par élément bien fait.

4. Ressources

Thème	Ressource
manipulation de données	• introduction à data.table, par Laurent Bergé
manipulation de chaînes de caractères	• intro[ion] aux expressions régulières, par Laurent Bergé
création de site web avec RStudio	• créer des sites webs avec Quarto, par Posit
Git (pour Github)	• introduction pas à pas à Git, par Peter Cottle
Publier un site sur Github	• doc officielle • des tonnes de vidéos en ligne
graphiques	• qu'est-ce qu'un joli graphique, par Laurent Bergé • introduction à ggplot2, par Garrick Aden-Buie