Projet

Le projet est à réaliser individuellement. Les étudiants devront envoyer au plus tard le **10 avril 2022** l'ensemble des fichiers (code et word/pdf) sur le mail suivant : quentin_lajaunie@hotmail.fr.

Dans le cadre de ce projet, il est demandé aux étudiants de respecter les règles suivantes :

- Le code doit être commenté ;
- Le nom des variables doit être adapté ;
- Le nom des variables ne doit pas contenir d'accent ;
- L'indentation doit être respectée.

Ces critères compteront significativement dans la note finale.

Le projet se décompose en trois parties. Dans la première, une série de fonctions est à réaliser. **Ces fonctions ne pourront pas faire appel à des packages et fonctions déjà existantes, sauf si cela est explicitement précisé.** Dans la deuxième, les étudiants retraiteront les données à partir de packages, à partir des fonctions développées dans la première partie, ou encore en utilisant des requêtes vues en cours. Enfin, dans la troisième partie, les étudiants devront créer un indice composite. Cet indice s'inspirera des étapes présentées en cours. Certaines des questions pourront éventuellement s'appuyer sur des packages et des fonctions à chercher en ligne (pour la partie 2 et pour la partie 3). Notez également que l'ensemble des fonctions développées ne seront pas forcément utilisées. Un document word/pdf sera à rédiger pour présenter les résultats.

Première partie - Fonctions

1. Développez quatre fonctions permettant de calculer la moyenne, l'écart-type, la Skewness et le Kurtosis. Chacune de ces fonctions aura comme input un vecteur de données. Pour rappel :

   - **moyenne** : la moyenne $\overline{x}$ se calcule comme suit :

   $$\overline{x} = \frac{1}{n}\sum_{i=1}^{n}(x_i).$$

   - **écart-type** : c'est un indicateur de dispersion, noté σ, qui se calcule comme suit :

   $$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

   - **Skewness** : c'est un indicateur qui mesure l'asymétrie d'une distribution, notée S, qui se calcule comme suit :

   $$S = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)^3}.$$

   - **Kurtosis** : c'est un indicateur qui mesure l'applatissement d'une distribution, noté κ, qui se calcule comme suit :

   $$\kappa = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)^4}.$$

2. Développez une fonction pour calculer la statistique du test de Jarque-Bera. Il permet de tester la normalité d'une distribution, et revient à tester l'hypothèse suivante :

   $$\begin{cases} H_0\text{: les données suivent une loi normale.} \\ H_1\text{: les données ne suivent pas une loi normale.} \end{cases}$$

   La statistique du test de Jarque-Bera se calcule comme suit :

   $$JB = \frac{n}{6}\left(S^2 + \frac{(\kappa - 3)^2}{4}\right)$$

3. Développez une fonction permettant de calculer la covariance entre X et Y. La covariance se calcule comme suit :

   $$cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

4. Développez une fonction permettant de calculer la corrélation de Pearson entre deux variables X et Y. Notée $\rho_{X,Y}$, la corrélation se calcule comme suit :

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y},$$

où $cov(X,Y)$ désigne la covariance entre X et Y, et σ_X et σ_Y désignent l'écart-type de X et Y respectivement.

5. Développez une fonction permettant de calculer la corrélation de Spearman entre deux variables X et Y. Notée $\rho_{rg(X),rg(Y)}$, la corrélation se calcule comme suit:

$$\rho_{rg(X),rg(Y)} = \frac{cov(rg(X),rg(Y))}{\sigma_{rg(X)}\sigma_{rg(Y}},$$

où $rg(X)$ et $rg(Y)$ désignent les rangs associés aux variables X et Y, $cov(rg(X),rg(Y))$ désigne la covariance entre $rg(X)$ et $rg(Y)$, et $\sigma_{rg(X)}$ et $\sigma_{rg(Y)}$ désignent l'écart-type de $rg(X)$ et $rg(Y)$ respectivement.

6. L'interpolation linéaire est une méthode qui consiste à déterminer une valeur manquante à partir d'une fonction affine. L'estimation de la valeur manquante est réalisée par cette fonction en prenant en compte deux points déterminés.

   Développez une fonction d'interpolation linéaire qui prend en paramètres x, deux points x_a et x_b, ainsi que les deux valeurs associées y_a et y_b. La fonction renverra y : la valeur associée à x, et sera calculée à partir de la fonction affine dépendant des couples (x_a, y_a) et (x_b, y_b).

7. Développez une fonction permettant d'estimer les coefficients d'une régression linéaire.

   Soit une équation de la forme $Y = X\beta + \varepsilon$, où Y est un vecteur de dimension $[n \times 1]$ contenant les données de la variable que l'on cherche à expliquer, X une matrice de dimension $[n \times k]$ contenant les données des k variables explicatives, β un vecteur de dimension $[k \times 1]$ contenant les coefficients associées aux k variables explicatives, et ε le vecteur des erreurs, de dimension $[n \times 1]$.

   La fonction doit permettre d'estimer les coefficients β en minimisant la somme des carrés des erreurs. Cet estimateur est noté $\hat{\beta}$ et se calcule comme suit :

$$\hat{\beta} = (X'X)^{-1}X'Y$$

8. Développez une fonction maximum et une fonction minimum. Ces deux fonctions doivent prendre en paramètre un vecteur de données, et renvoyer la valeur la plus grande pour la fonction maximum, et la valeur la plus petite pour la fonction minimum (les fonctions max et min ne doivent pas être utilisées).

9. Développez une fonction qui renvoie dans un dataframe le minimum, le premier quartile, la médiane, la moyenne, le troisième quartile et le maximum d'un vecteur entré en paramètre. Vous pouvez utiliser la fonction `quantile` uniquement pour cette fonction.

10. Développez une fonction de tri qui renvoie une matrice de deux colonnes dans laquelle :

    - la première colonne renvoie le classement sous forme de nombres allant de 1 à n (1 pour la plus grande valeur et n pour la plus petite), et où n représente la taille du vecteur entré en paramètre ;

- la deuxième colonne renvoie le vecteur des nombres entrés en paramètre, triés du plus grand au plus petit.

11. La standardisation, ou Z-Score, est une approche qui permet de normaliser un vecteur de données. Après normalisation, le nouveau vecteur à une moyenne nulle et un écart-type à 1. La standardisation se calcule comme suit :
$$X_{stand} = \frac{(X_i - \overline{X})}{\sigma_X}$$
Développez une fonction de normalisation basée sur la standardisation.

12. La méthode min-max est une autre approche qui permet de normaliser un vecteur de données. Elle se calcule comme suit :
$$X_{mm} = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$
Développez une fonction de normalisation basée sur la méthode min-max.

13. L'échelle catégorielle (Categorical scale) attribue un score pour chaque indicateur. Le résultat peut-être numérique (valeur allant de 1 à 5 par exemple) ou qualitatif (objectif atteint vs objectif non atteint). Cette méthode s'appuie généralement sur les percentiles de la distributions, et masque ainsi une partie de la variance des indicateurs.
Développez une fonction d'échelle catégorielle qui prend en paramètres un vecteur de données, les 5 quantiles, et qui renvoie les valeurs de 0/20/40/60/80/100 conditionnellement à la valeur de chaque de données et de sa position par rapport aux quantiles.

14. De nombreux indicateurs composites reposent sur une pondération "Equal Weighting" (EW). Cela consiste à avoir un indicateur composite construit à partir de n variables, où chaque variable aura un poids de $\frac{1}{n}$.
Développez une fonction qui prenne en argument une matrice de n variables, et qui renvoie un indicateur qui agrège chaque variable avec un poids égal.

15. La méthode d'agrégation additive avec pondération linéaire se fait comme suit :
$$CI_i = \sum_{k=1}^{K} w_k I_{qi}$$
Développez une fonction d'agrégation additive qui prend en paramètre un vecteur de poids et une matrice de variables, et qui calcule l'indice composite à partir de ces deux paramètres. La fonction doit vérifier que les dimensions des paramètres soient égales (gestion des erreurs).

16. Sur le même principe, réalisez une fonction d'agrégation géométrique.

17. L'agrégation avec l'Analyse en Composante Principale se réalise en 3 étapes. Dans la troisième étape, une rotation des facteurs est faite pour que chaque variable ne contribue qu'à une composante principale.
En vous appuyant sur les éléments du cours, développez une fonction qui permette de réaliser cette étape. Vous pouvez vous appuyez sur le package FactoMineR et factoextra.

Le projet ne limite d'aucune façon les étudiants à ne développer que ces fonctions. Aussi, vous avez la liberté de vous appuyer sur une fonction développée par vos soins pour répondre à une ou plusieurs des questions des parties suivantes.

Deuxième partie - Récupération et retraitement des données

Le fichier excel `data_def` contient les définitions des données. Ces données sont les composantes des indicateurs et sous-indicateurs de développement financier étendus par Svirydzenka en 2016. Pour la partie Institutions, il y a 4 variables pour l'indice FID, 2 variables pour l'indice FIA et 5 variables pour l'indice FIE. Pour la partie marché, il y a 5 variables pour l'indice FMD, 2 indices pour l'indice FMA et 1 variable pour l'indice FME.

Le second fichier `data_base` est un fichier au format .csv qui contient les valeurs de chaque variable. Ces valeurs sont celles de pays développés, à fréquence annuelle.

L'indicateur de développement financier étendu par Svirydzenka en 2016 se présente comme suit:

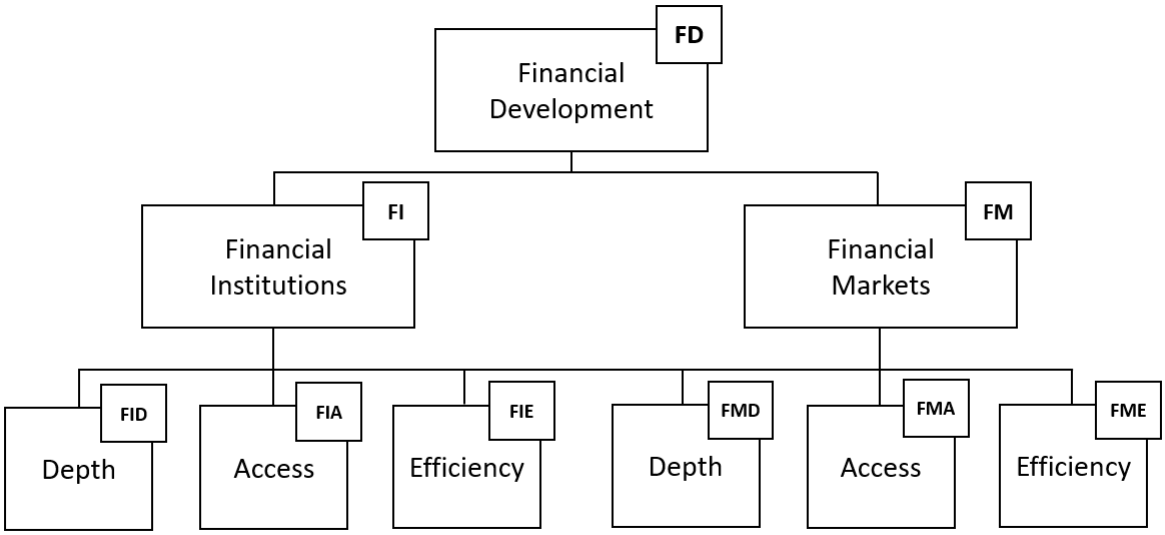

Figure 0.1: Indicateurs du développement financier - Svirydzenka (2016)

Dans cette partie, il vous est demandé de retraiter ces données en plusieurs étapes. Ce retraitement vous permettra ensuite de réaliser la troisième partie du projet :

1. Importez les données dans le logiciel R Studio, et stockez les dans une variable.
2. En utilisant les fonctions et commandes de votre choix, déterminez le nombre de pays présents dans la base. Déterminez également le nombre d'années pour chaque pays.
3. Calculez les statistiques descriptives des différentes variables en utilisant les fonctions de votre choix codées dans la première partie.
4. Deux données manquantes apparaissent pour le pays *Chile* avec le code ISO *CHL* pour la variable `ID1`. Déterminez ces valeurs manquantes avec les fonctions les plus appropriées (fonctions développées dans la première partie).
5. Stockez dans 6 matrices différentes les variables *Institutions* (respectivement les variables ID, IA, IE) et les variables *Marché* (respectivement les variables MD, MA, ME).
6. En utilisant la méthode min-max, normalisez chaque variable dans chaque matrice créée précédemment, et stockez les résultats dans six nouvelles matrices.
7. En utilisant la méthode de standardisation, normalisez chaque variable dans chaque matrice créée dans l'étape 5, et stockez les résultats dans six nouvelles matrices.

Troisième partie - Analyse et Interprétation

Le fichier pdf fourni avec ces deux fichiers de définitions et de données correspond au papier de Svirydzenka de 2016. Il présente la construction de l'indice de développement financier, les données utilisées, les retraitements effectués, et l'interprétation des résultats.

Dans cette partie, il vous sera demandé de vous appuyer sur ce papier pour motiver certaines réponses, de coder et/ou de rédiger des réponses dans un document word à part. Les réponses devront être précises, détaillées et proprement rédigées.
Objectif : créer deux indicateurs Institutions et Marchés

1. En vous appuyant sur le papier de Svirydzenka (2016), expliquez l'intérêt d'avoir des indicateurs de développement financier, et présentez rapidement les 6 sous indicateurs que vous allez créer, et qui sont présentés dans la Figure de la partie 2. (équivalent à *Etape 1 - Cadre théorique* vue en cours)

2. Présentez rapidement les variables pour chaque sous-indicateur. (équivalent à *Etape 2 - Sélection des données* vue en cours)

3. Les étapes données manquantes, retraitement, et normalisation ont été réalisées dans la Partie 2. Toutefois, pour connaître la structure des données, calculez pour les 6 sous-indicateurs les corrélations entre les variables qui les composent (utilisez les 6 matrices normalisées créées dans l'étape 6 ou 7). *Remarque : si une matrice n'est composée que d'une variable, il n'est pas nécessaire de calculer de corrélation.*

4. Pour chacun des 6 sous indicateurs, en suivant les étapes du cours :

   - Utilisez la fonction `PCA`, avec l'option `print=TRUE` ;
   - Utilisez ensuite la fonction `get_eigenvalue` et la fonction `fviz_eig`. Récupérez les graphiques et commentez vos résultats ;
   - Utilisez la fonction `get_pca_var` puis `corrplot`, puis `fviz_pca_var`. Commentez également vos résultats ;
   - Utilisez ensuite la méthode en 3 étapes de pondération avec l'ACP pour créez les 6 indicateurs agrégés.

5. Une fois les 6 sous-indicateurs obtenus, agrégez les indicateurs institutions ensemble et les indicateurs marchés ensemble. Cette agrégation doit être faite de façon équipondérée, avec deux méthodes : additive, et géométrique. Discutez de la différence entre ces deux méthodes.

6. Calculez la moyenne annuelle des deux indicateurs Institutions et Marchés obtenus avec la méthode additive et avec la méthode géométrique, et représentez les résultats dans deux graphiques séparés. Commentez les résultats.