Description du projet

Stéphane Caron

2018-01-01

Abstract

Ce document a pour but d'expliquer le projet qui sera travaillé par des joueurs de l'équipe de hockey des Dynamiques du Cégep de Sainte-Foy (2018). Le projet proprement dit est une brève introduction aux statistiques, à la programmation et aux techniques d'apprentissage automatique permettant de résoudre des problèmes concrets.

Contents

Mise en contexte

Ce projet sera réalisé dans le cadre du programme de "tutorat" mis en place par Christian Larue, entraîneur de l'équipe. Le programme a pour but de présenter aux joueurs actuels de l'équipe différents domaines dans lesquels certains anciens joueurs oeuvrent actuellement.

Ce projet spécifique permet de donner une brève introduction aux domaines des mathématiques et des statistiques, en plus de toucher à plusieurs concepts en lien avec la programmation et l'analyse de données. Ces concepts peuvent évidemment s'appliquer à plusieurs autres domaines, notamment l'informatique et l'actuariat. Pour plus d'informations sur ces domaines en particulier, voici quelques liens pertinents:

Mathématiques et statistiques:

- *Département de mathématiques et statistique de l'Université Laval*
- *Data science and statistics jobs*

Informatique et programmation:

- *Département d'informatique et génie logiciel de l'Université Laval*
- *McGill School of Computer Science*
- *Data science and analytics in sports*

Actuariat:

- *École d'actuariat de l'Université Laval*
- *Society of Actuaries*
- *Casualty Actuarial Society*

Description du projet

Dans ce projet, nous utiliserons les statistiques individuelles de la saison 2016-2017 de la LNH pour tenter de comprendre quels joueurs de hockey partagent des styles de jeu similaires. Ainsi, nous pourrions possiblement être en mesure de mieux comprendre des exemples comme:

- Pourquoi Max Paciorrety et Philippe Danault se complète bien alors que Galchenyuk semble avoir moins de chimie avec le capitaine du CH?
- Est-ce que le CH possède vraiment un attaquant comparable aux gros canons de la ligue?
- Est-ce qu'un défenseur comme Jeff Petry se compare à des défenseurs plus défensifs ou offensifs?

Pour tenter de répondre à ces questions, nous ferons une brève introduction de certaines méthodes mathématiques et statistiques. Le concept principal mis de l'avant dans ce projet s'appelle le "clustering". Le *clustering* est une méthode statistique permettant de regrouper des données dans différents groupes partageant des caractéristiques similaires à l'intérieur du groupe, mais différentes de celles des autres groupes. Il existe plusieurs méthodes de clustering, basées sur différents algorithmes, qui permettent d'obtenir des résultats distincts dépendemment du contexte. Ces méthodes ont d'ailleurs plusieurs applications dans d'autres domaines:

- Marketing:
  - Pour faire la segmentation d'un marché et l'analyse de prospects potentiels.
  - Pour définir certaines caractéristiques des clients "perdus" et ainsi améliorer la rétention d'une clientèle.
  - Pour l'analyse géographique de marchés potentiels.
- Finance:
  - Pour faire le regroupement d'actions présentant des caractéristiques similaires et ainsi améliorer la diversité d'un portefeuille.
  - Pour établir certaines caractéristiques communes de clients qui ne remboursent par leurs dettes (enquête de crédit).
- Médecine:
  - Pour la recherche de caractéristiques présents chez les patients contractant un certain virus.
  - Pour la gestion des dépenses médicales (personnels, équipements, investissements, etc) en lien avec le type de demande de certains établissements médicaux.

Dans notre cas, nous utiliserons une méthode de clustering précise, la méthode k-means. Cette méthode est décrite plus en détails dans les prochaines sections. Comme introduit un peu plus haut, nous utiliserons cette méthode pour établir des styles de joueurs de hockey. Finalement, nous pourrons utiliser les résultats de l'analyse et tenter de voir si certaines questions peuvent être clarifiées.

Description des données

Jeu de données

Le jeu de données correspond aux statistiques individuelles des joueurs de la LNH pour la saison 2016-2017. Le jeu de données a été extrait sur ce *site*.

Nettoyage des données

Une étape inévitable dans l'analyse de données et dans l'application de la grande majorité des méthodes statistiques consiste à nettoyer et préparer les données. Dans notre situation, nous devrons réaliser certaines de ces étapes:

- Sélectionner les données pertinentes.
- Modifier certaines données pour les rendre adaptées à notre analyse.
- Faire la gestion des données manquantes.
- Etc

Méthodologie

Une fois que les données sont nettoyées et prêtes pour l'analyse, il faut maintenant passer à l'étape d'appliquer notre algorithme et de procéder à l'analyse. Avant tout, il est important de comprendre le fonctionnement de l'algorithme.

Méthode k-means

La méthode de clustering introduite dans ce projet se nomme la méthode k-means. Cette méthode se base sur le fait que le nombre de groupes (k) à determiner est connu (ou supposé) d'avance. Cela peut paraître contre-intuitif de déterminer un nombre de groupes à l'avanc, mais nous verrons un peu plus loin qu'il est possible après coup de trouver un nombre de groupes "optimal".

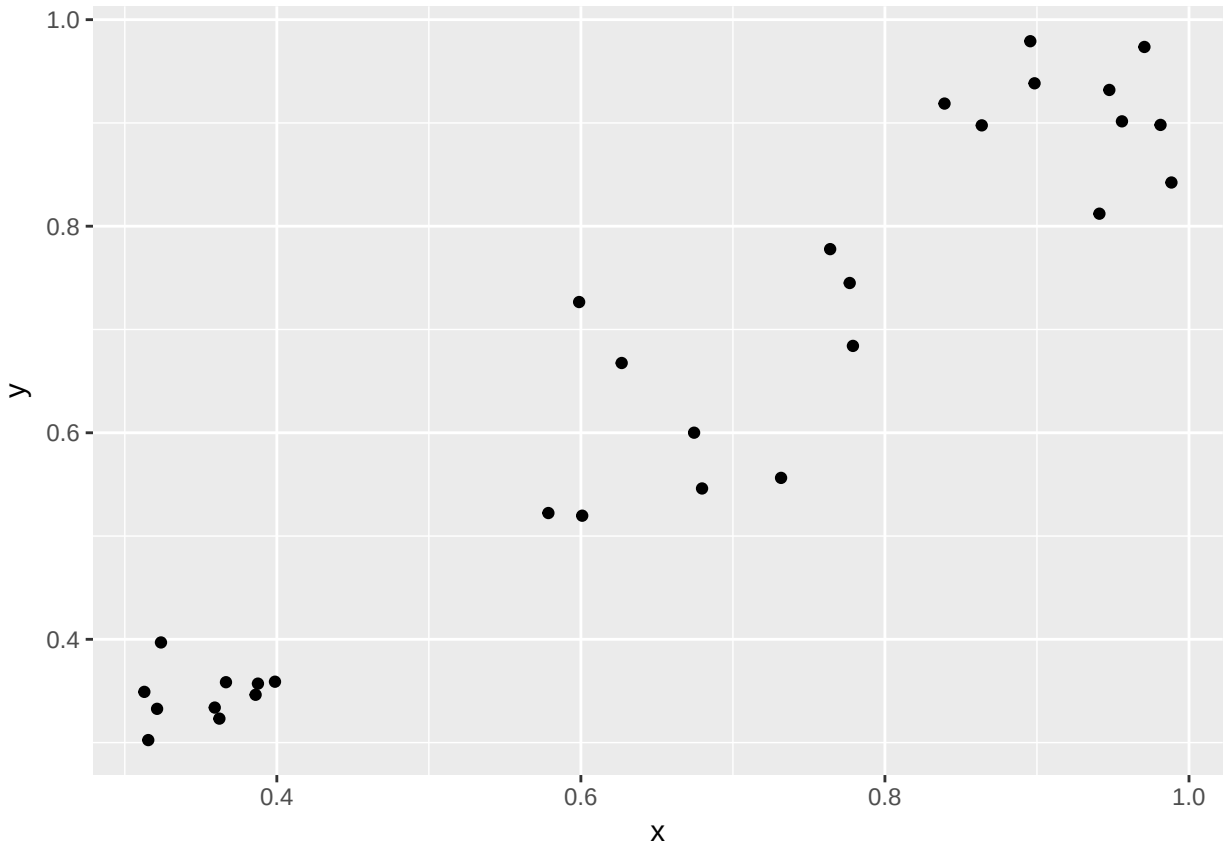
En résumé, l'algorithme fonctionne comme suit:

1. On commence par initialiser aléatoirement les groupes.
2. On détermine le centroïde de chacun des groupes. Le centroïde correspond à la valeur centrale de chacun des groupes.
3. Pour chacune des données, on attribue celle-ci au groupe correspondant au centroïde le plus près. Pour déterminer quel centroïde est le plus "près", il faut se définir une mesure de distance. Dans le cas le plus fréqeuent, on utilise la distance euclidienne, soit la longueur du trajet entre deux points.
4. On réitère les étapes 2 et 3 jusqu'à temps que les groupes ne changent pratiquement plus.

Voici un exemple en deux dimensions qui illustre les étapes décrites un peu plus haut.

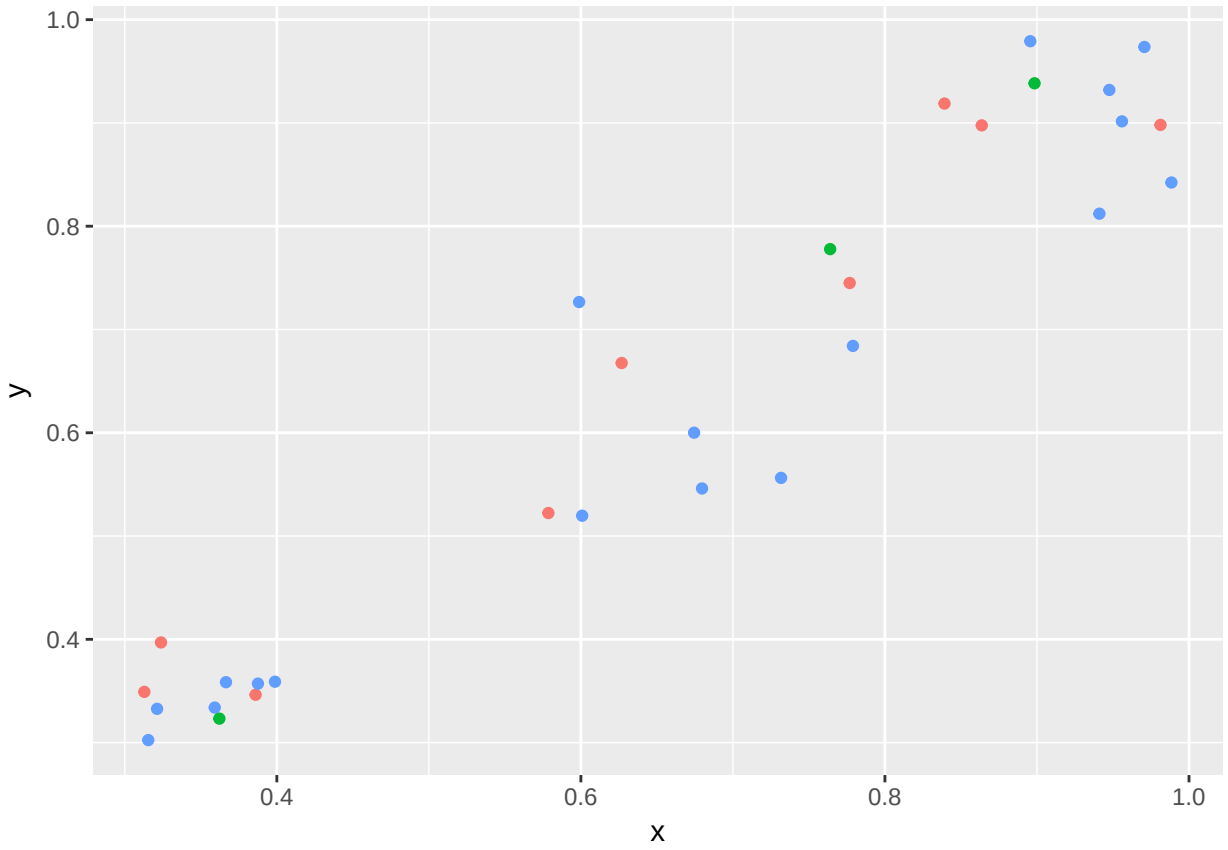
Étape 0

Voici le jeu de données (exemple) dont nous voulons appliquer notre méthode de clustering. On remarque rapidement qu'il semble avoir 3 groupes apparents.



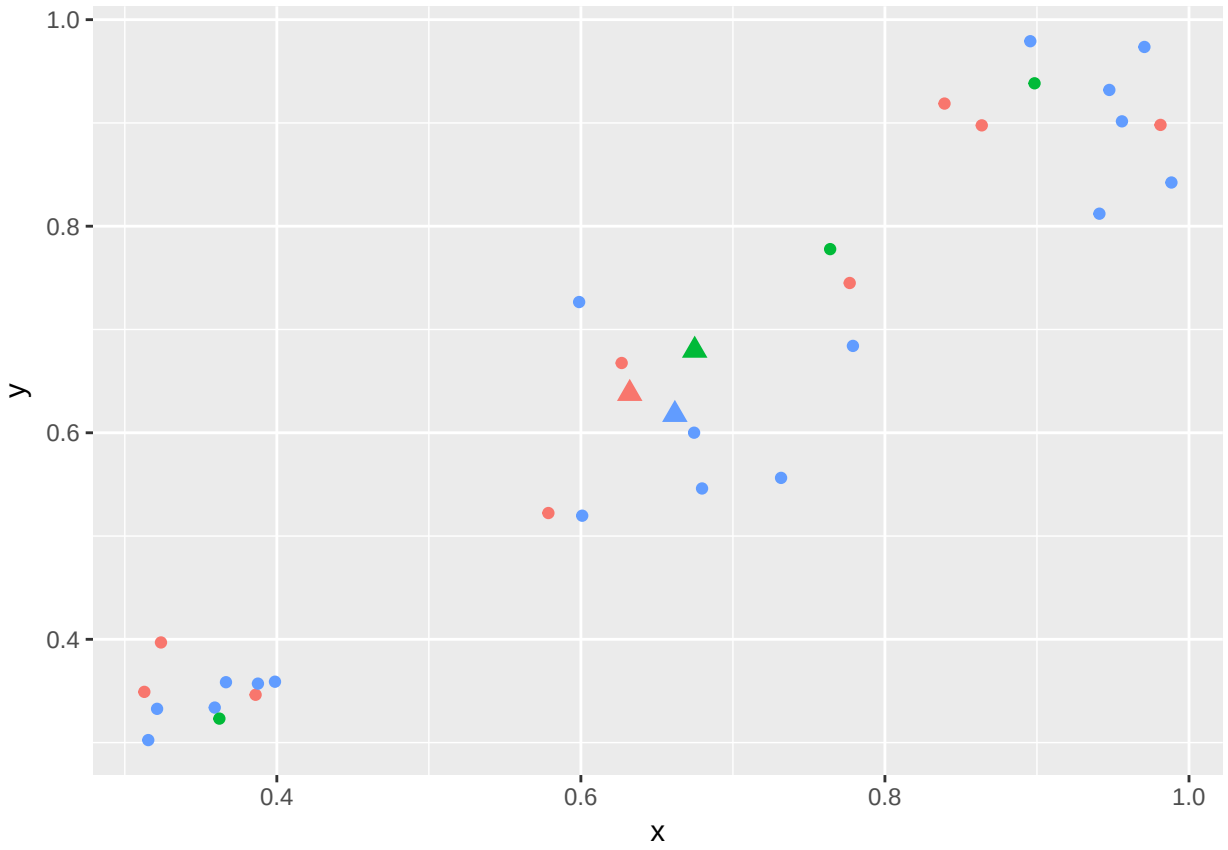
Étape 1

On initialise l'algorithme. Ici, on attribue aléatoirement chacune des données à un groupe. Dans le cas ci-dessus on commence en définissant 3 groupes (k = 3).



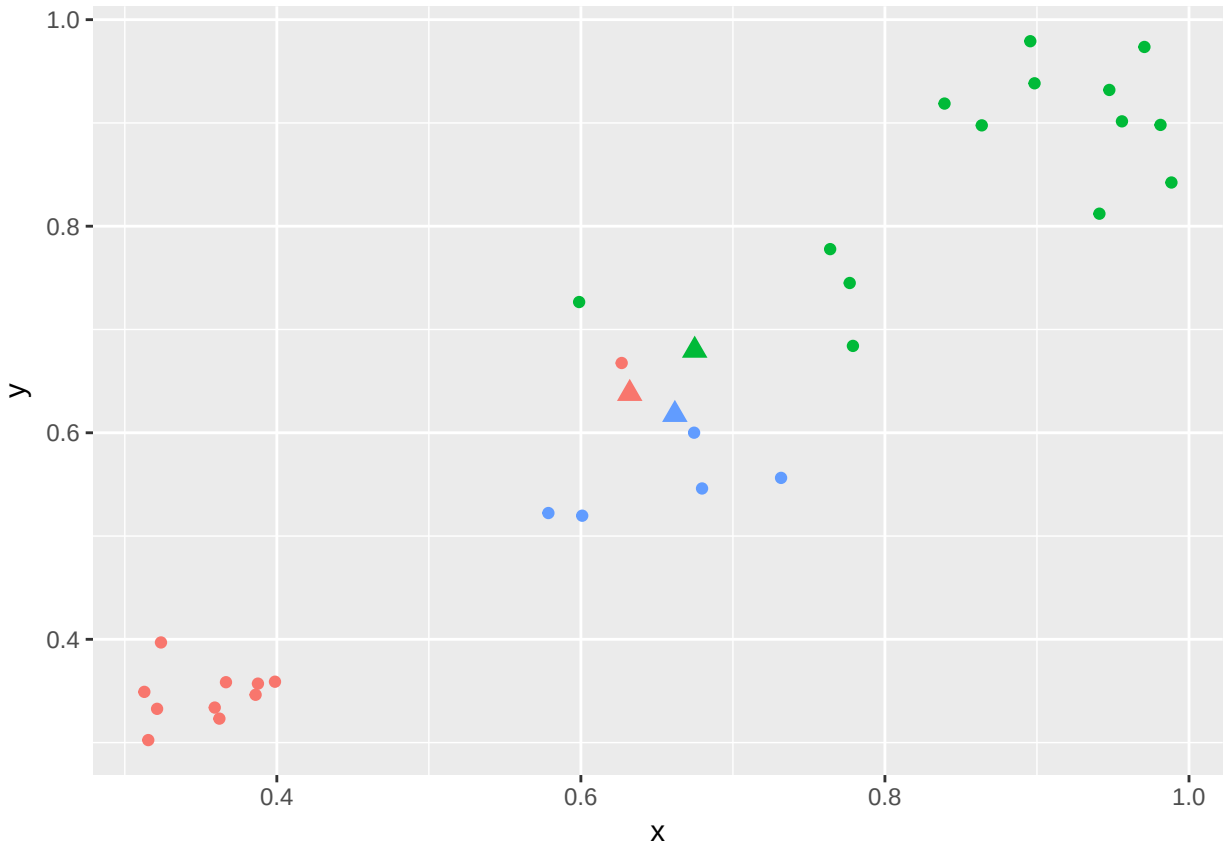
Étape 2

On trouve le centroïde, ou la valeur centrale, de chacun des groupes. Ces valeurs sont illustrées par des triangles. Les points ronds sont les données que nous voulons grouper bien sûr.



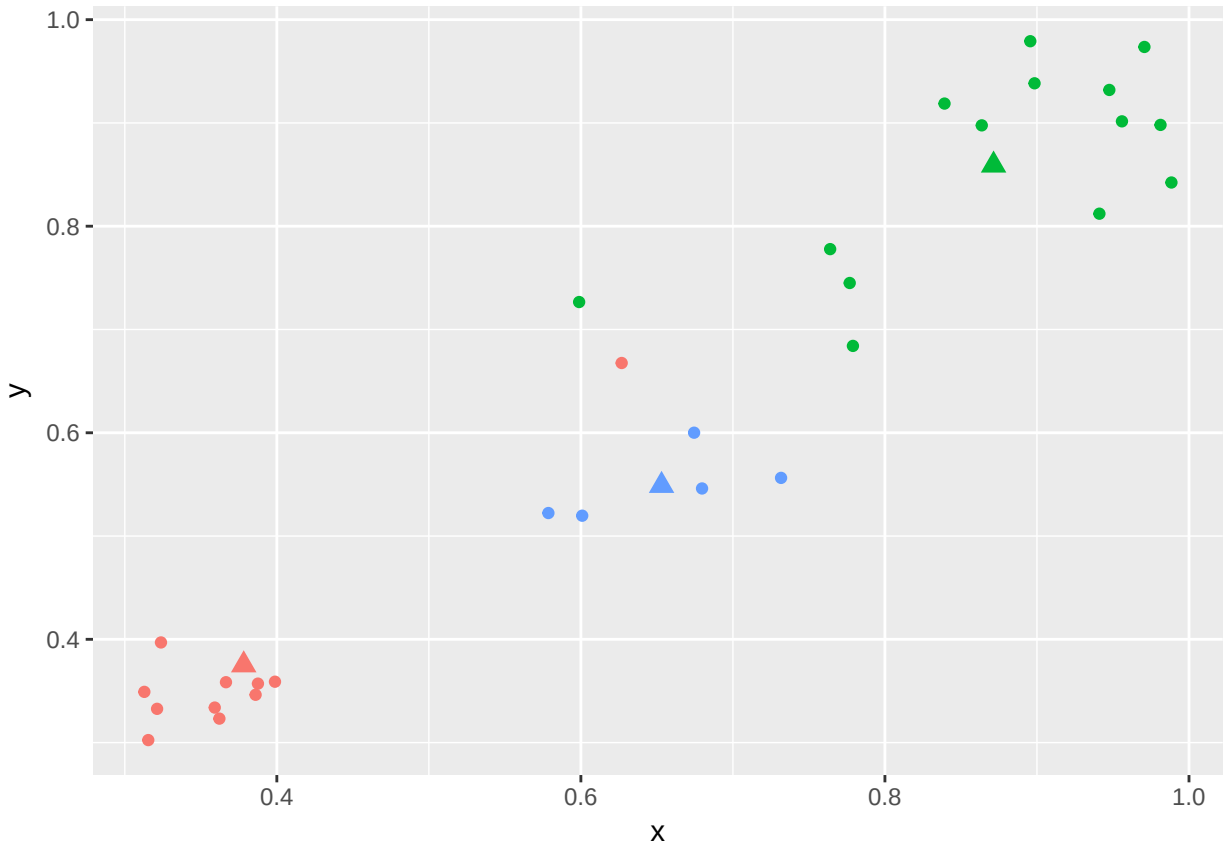
Étape 3

Pour chaque donnée, on attribue le groupe correspondant au groupe du centroïde le plus près de celle-ci.



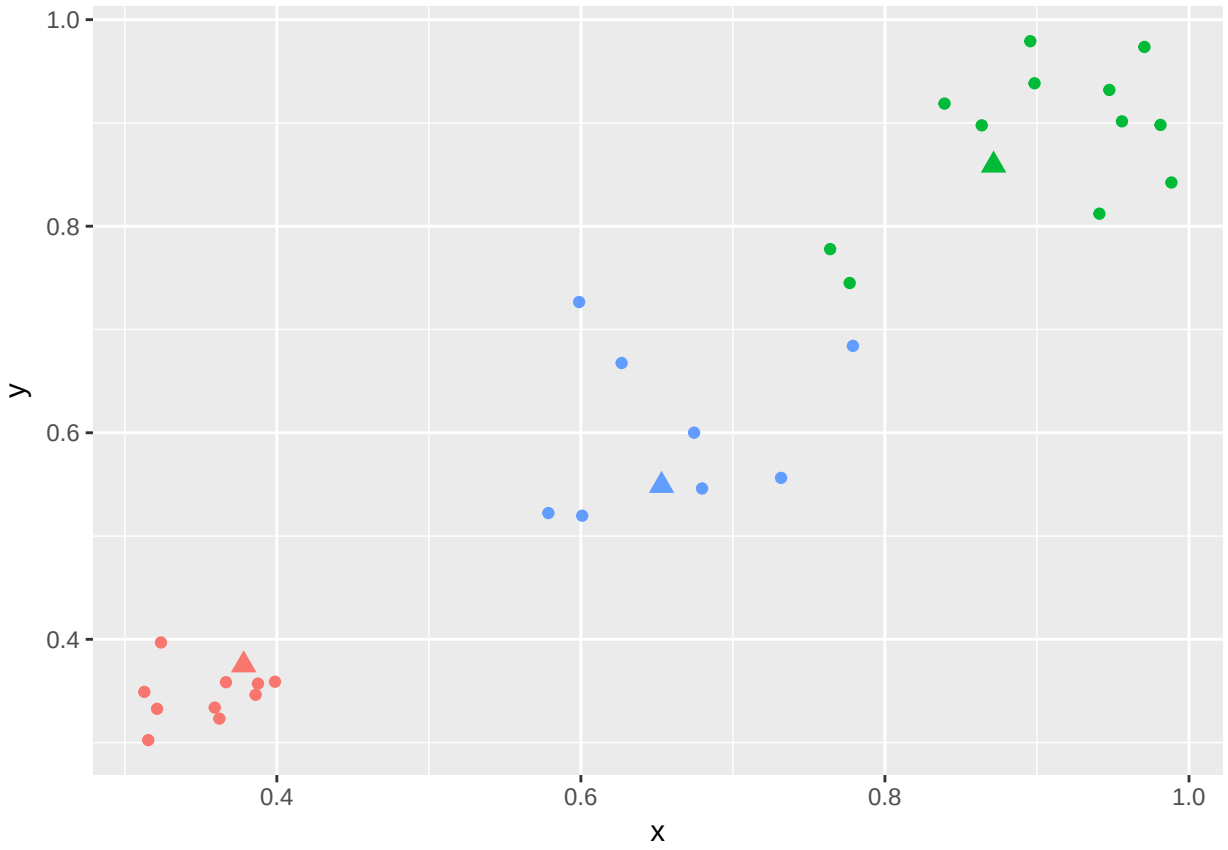
Étape 4

On recalcule les valeurs de centroïde pour chacun des groupes nouvellement définis.



Étape 5

Comme à l'étape 3, on attribue le groupe correspondant au groupe du centroïde le plus près de chaque donnée.



Étape 6

On recommence les deux dernières étapes jusqu'à temps que les groupes soient stables. Pour déterminer quand les groupes sont stables et quand il est souhaitable d'arrêter l'algorithme, on peut décider, par exemple, que si les groupes demeurent inchangés après 5 ou 10 itérations, on arrête.

Ce *site* constitut un bel example supplémentaire de comment l'algorithme fonctionne à chaque itération.

Choix du nombre de groupes

Comme mentionné un peu plus tôt, il peut paraître contre-intuitif de définir le nombre de groupes avant même de commencer l'analyse. Dans l'exemple décrit un peu plus haut, il était relativement facile de voir les 3 groupes apparents. Toutefois, il n'est pas toujours aussi aisé de voir de tels groupes ou il peut être difficile de les visualiser dans des situations ayant plus de 2 ou 3 dimensions.

Il existe donc une méthode permettant de définir le nombre de groupes "optimal". Au départ, nous avons introduit le clustering comme étant un technique permettant de regrouper certaines données ayant des caractéristiques similaires. Il est donc possible d'obtenir une mesure nous permettant de valider à quel point les données à l'intérieur d'un même groupe sont similaires. Pour ce faire, nous calculerons la somme totale des carrés à l'intérieur des groupes. En gros, on somme les distances qui séparent chacune des données à l'intérieur d'un groupe avec le centroïde de ce même groupe. Pour avoir des groupes homogèenes, on souhaite évidemment que cette mesure soit petite puisque cela nous indique que les valeurs à l'intérieur d'un groupe sont relativement similaires. Cependant, il est important de comprendre que plus le nombre de groupes est grand, plus cette mesure devient petite. À l'ultime, si on définit un nombre de groupes égale aux nombre de données, chaque donnée deviendra son propre groupe et la mesure sera nulle. Pour trouver un nombre de groupe optimal, on calcule cette mesure pour différentes valeurs de k (nombre de groupes) et on fait un graphique de la mesure en fonction du nombre de groupes. À partir de là, on cherche ce qu'on appelle le

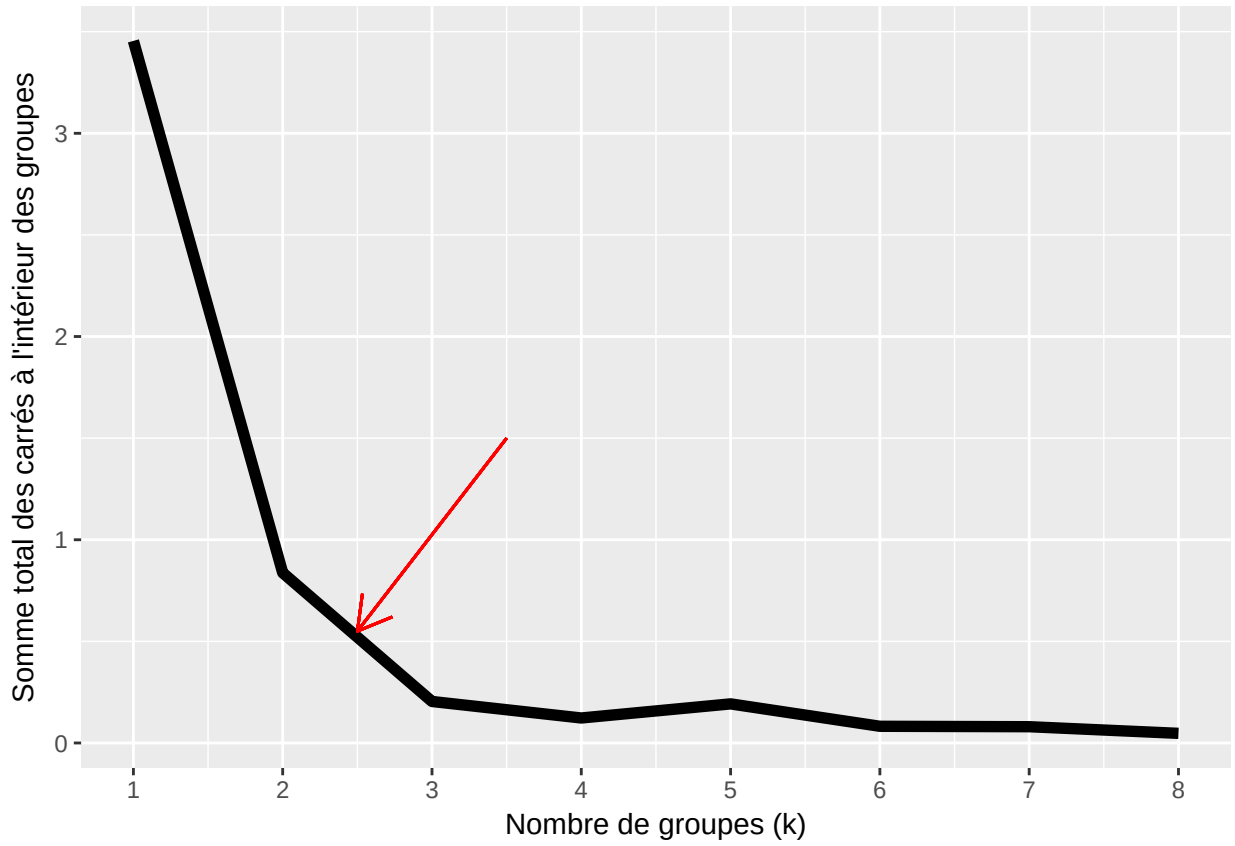


Figure 1: Graphique en coude permettant de trouver le nombre de groupes optimale

"coude" de la droite, soit le point où on apperçoit une flexion. La figure 1, construite à partir de l'exemple illustré plus tôt, nous montre que le coude à lieu environ entre 2 et 3 groupes. Il serait donc approprié de garder 3 groupes, mais il pourrait également être correct de garder seulement 2 groupes, alors que les deux groupes en haut à droite pourraient possiblement former un seul groupe.

Application aux données de hockey

Dans ce projet, nous appliquerons cet algorithme a des statistiques individuelles de joueurs de hockey. C'est donc dire que dépendemment du nombre de statistiques inclut dans l'analyse, nous effectuerons le même genre de procédure d'illustré un peu plus haut. Cependant, nous ne travaillerons pas nécéssairement dans un espace à deux dimensions commme illustré dans l'exemple. En effet, si nous décidons d'inclure comme statistiques: les buts, les passes, les mises en échec et les minutes de pénalités pour construire nos groupes de joueurs, nous travaillerons sur un espace à 4 dimensions (car 4 statistiques différentes). Le principe de l'algorithme reste le même, il est juste plus difficile de visualiser les étapes une à une. Toutefois, nous pourrons quand même utiliser la méthode de validation décrite plus haut pour définir le bon nombre de groupes et nous pourrons analyser les joueurs placés dans chacun des groupes et tirer des conclusions.

Analyse des résultats

Une fois que les données ont été nettoyées et que nous aurons appliqué notre algorithme de clusering à celles-ci, il ne reste plus qu'à analyser les résultats. Dans ce projet, nous tenterons de produire un outil flexible qui nous permettra de modifier certains paramètres et d'analyser les différents résultats.

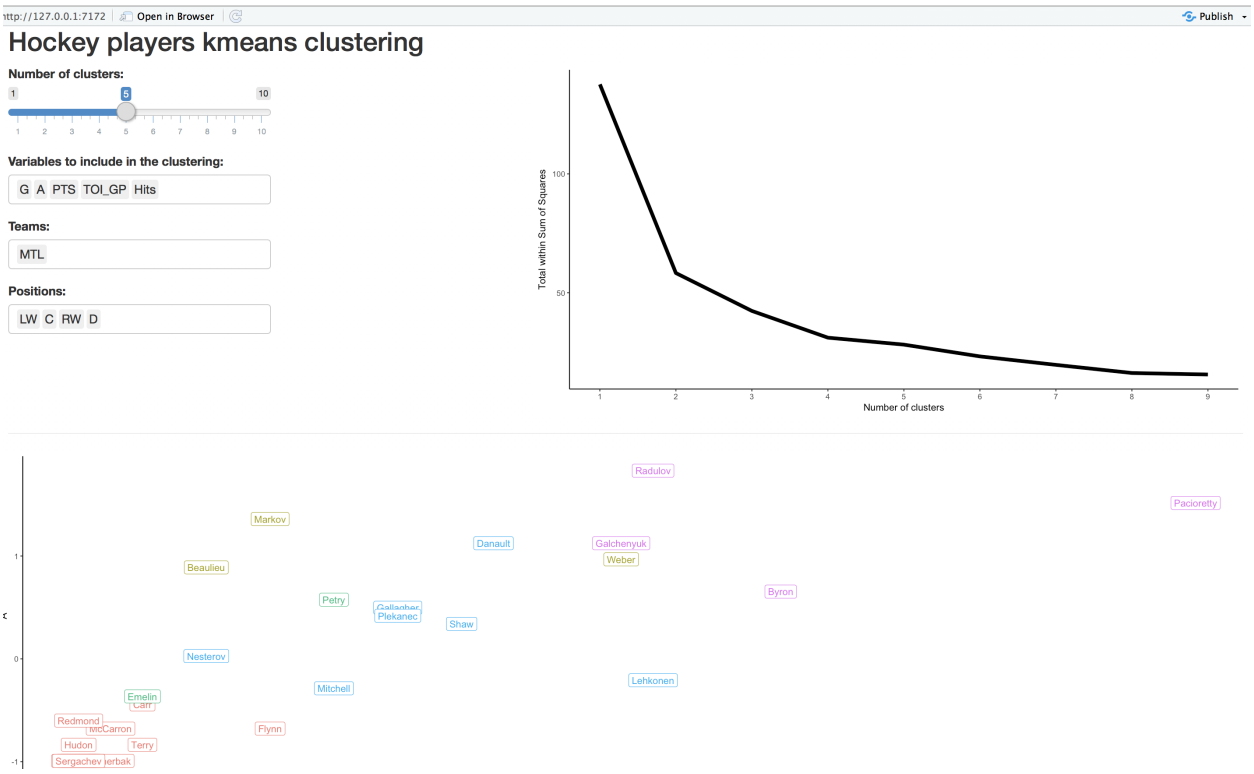


Figure 2: Example d'application permettant d'analyser les résultats de nos analyses

Nous bâtierons donc une application flexible dans laquelle nous implanterons notre algorithme et laisserons le choix à l'utilisateur de modifier certains paramètres comme ceux-ci:

- Le nombre de groupes.
- Les statistiques inclus dans l'analyse (buts, passes, points, mises en échec, etc).
- Les joueurs considérés dans l'analyse (choix de l'équipe).
- Les positions des joueurs analysés à l'intérieur des équipes sélectionnées.

La figure 2 illustre un aperçu du genre d'application que nous bâtierons.