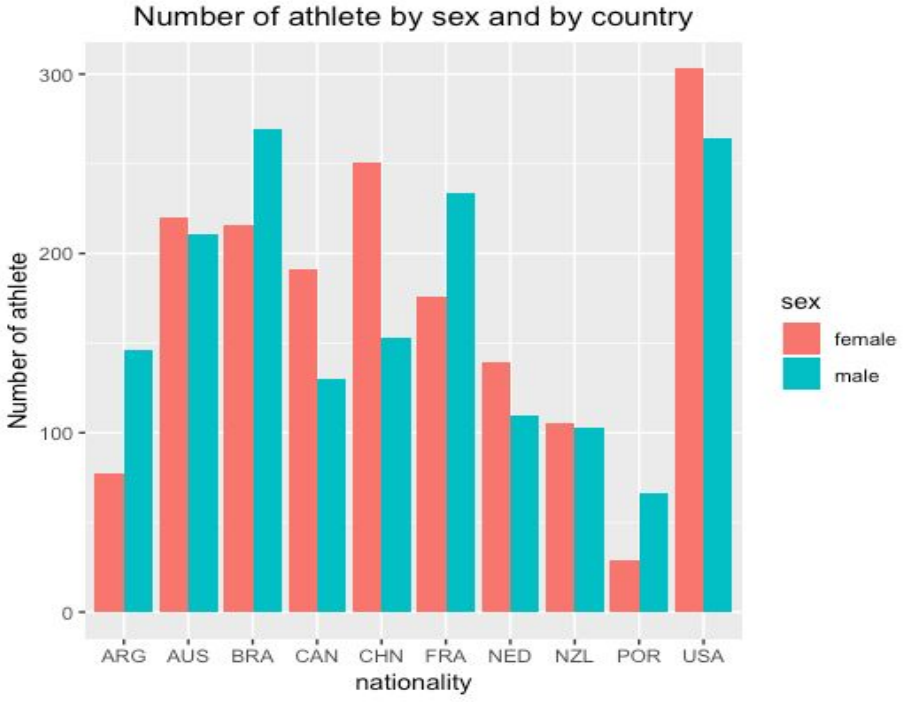
Introduction :

Dans le cadre du projet de traitement de masse de données, nous avons choisi d'analyser deux datasets différents. Le premier recense les athlètes ayant participé aux jeux olympiques d'été de Rio en 2016, pour 10 pays. Le but étant d'analyser la part des athlètes féminines ainsi que leur impact dans les résultats. Le second dataset regroupe tous les crimes ayant eu lieu dans la ville de Boston from June 2015 to October 2018. Nous voulons mettre en relation l'occurrence des crimes avec des paramètres tels que le temps, ou le lieu, puis nous allons essayer de définir quels sont les districts les plus dangereux.
Ces deux datasets proviennent du site : https://www.kaggle.com/datasets. Nous avons inclu les deux datasets dans le dossier du projet.

Dataset 1 : Etude de la part des femmes et de leur influence aux JO de Rio 2016.

Pour cette étude nous utilisons le dataset *"Rio_Olympic_Games.csv"* qui regroupe les athlètes qui ont participé aux JO de Rio pour dix pays : Argentine, Australie, Brésil, Canada, Chine, Pays-Bas, Nouvelle-Zélande, Portugal et Etats-Unis.

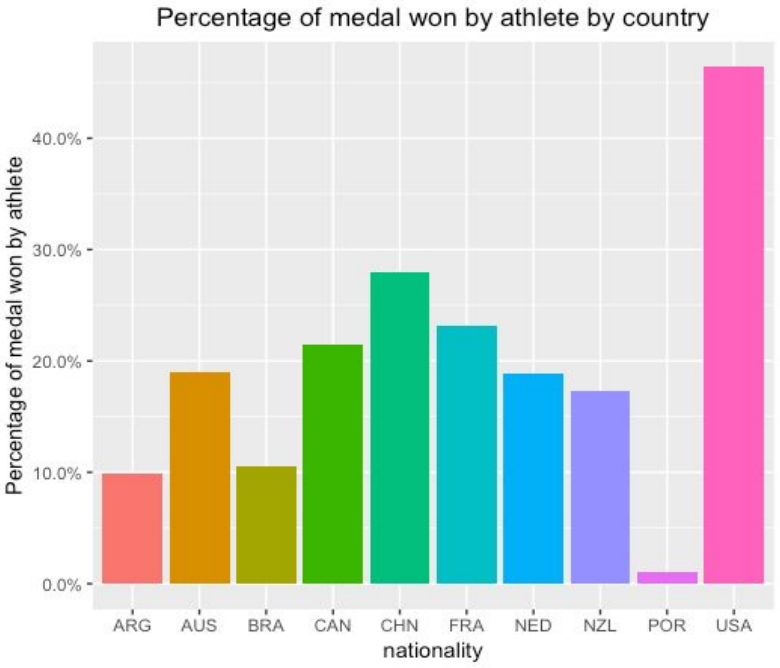
Dans un premier temps, nous allons étudier la parité homme-femme suivant les pays. Pour cela, nous avons décidé de créer un graphique représentant, non pas le nombre d'homme et de femme sélectionné par pays, mais le pourcentage d'homme et de femme sélectionné par pays, afin de se rendre mieux compte de la parité pour chaque pays.



On peut donc voir ici que 6 pays (AUS, CAN, CHN, NED, NZL, USA) sur 10 ont envoyé plus d'athlètes féminin que masculin, avec notamment la Chine et le Canada qui ont envoyé environ 20% plus de femmes que d'hommes. En revanche, l'Argentine et le Portugal ont respectivement envoyé environ 30% et 40% plus d'homme que de femme.

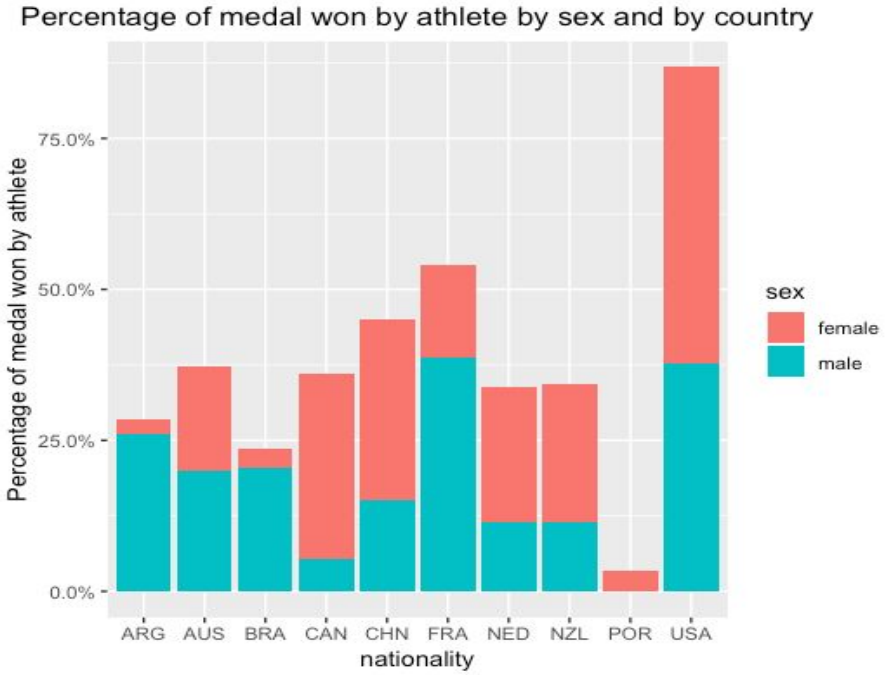
Ensuite, nous avons décidé de regarder les résultats par pays pour ces JO. Pour cela, nous considérons l'ensemble des médailles or, argent et bronze gagnées par l'ensemble des

athlètes d'un pays. Pour ne pas être influencé par la différence du nombre d'athlète sélectionnés par pays, nous avons décidé de représenter le pourcentage de médaille gagnées par athlète. Ce dernier est calculé en divisant le nombre de médailles gagnées par un pays par le nombre total d'athlètes concourant pour ce pays.



Ce graphique montre la très bonne performance des USA avec presque un athlète sur deux qui a gagné une médaille. De plus, mis à part la France, les 6 pays qui ont le plus fort taux de médailles gagnées par athlètes sont les mêmes qui ont sélectionné plus d'athlètes féminines que d'athlètes masculin.

A présent, nous allons étudier le pourcentage d'athlètes qui ont gagné une médaille par pays en faisant le distinguo entre homme et femme. Le but étant de voir si le nombre femmes sélectionné a une influence sur les résultats.
Pour cela, nous avons divisé le nombre de médailles gagnés par pays et par sexe par le nombre total d'athlète par pays et par sexe.



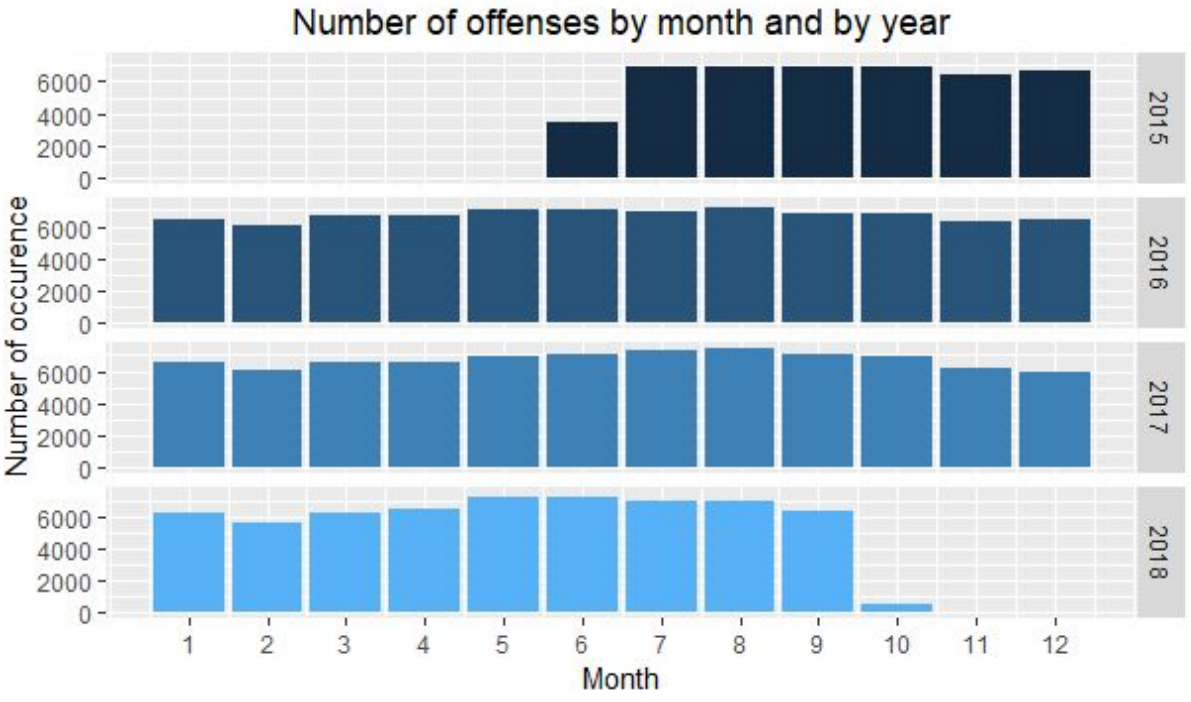
Ainsi, ce graph nous montre que pour les pays (sauf la France) ayant le plus fort taux de médailles gagnées par athlète, les femmes participent majoritairement à ce succès. En effet, on peut voir que, pour ces pays, le pourcentage de femme ayant gagné une médaille est bien supérieur à celui des hommes.

Cependant, cette étude ne porte que sur un échantillon de 10 pays et pour une seule édition des JO. Ainsi, pour que les déductions faites sur cette étude soit généralisées, il faudrait étendre l'analyse sur plus de pays et regarder l'évolution sur plusieurs éditions des JO.

Dataset 2 : Etude de l'occurrence des crimes en fonction du lieu et du temps, puis élection des districts tendus

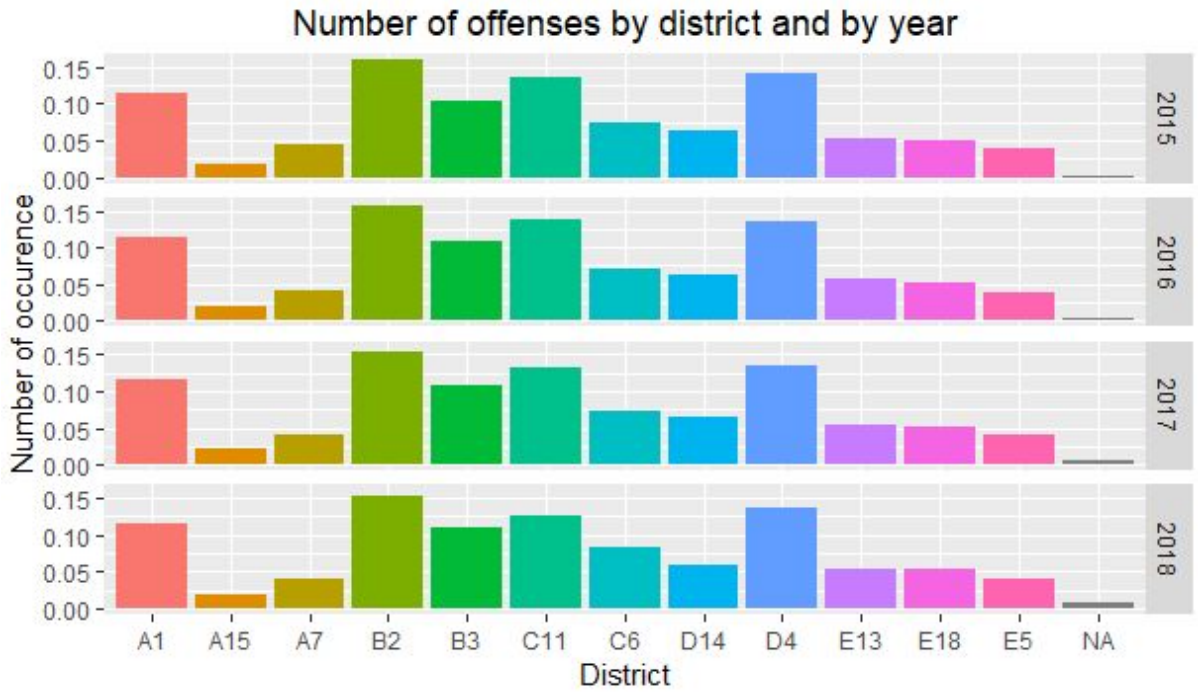
Nous avons utilisé le dataset *"Boston_Offenses.csv"* dans lequel nous pouvons avoir des informations sur le crime, comme son identifiant, la catégorie du crime commis, le type, la date et l'heure précise, ainsi que la localisation et le district précis. Les crimes peuvent être de plusieurs types, par exemple consommation d'alcool sur la voie publique, jusqu'au trafic d'humain.

Après avoir préalablement visualisé le dataset afin de voir l'étendue des crimes commis, ainsi que leur fréquence d'apparition, nous avons tracé le nombre total de crimes réalisés par mois sur les 4 années consécutives.



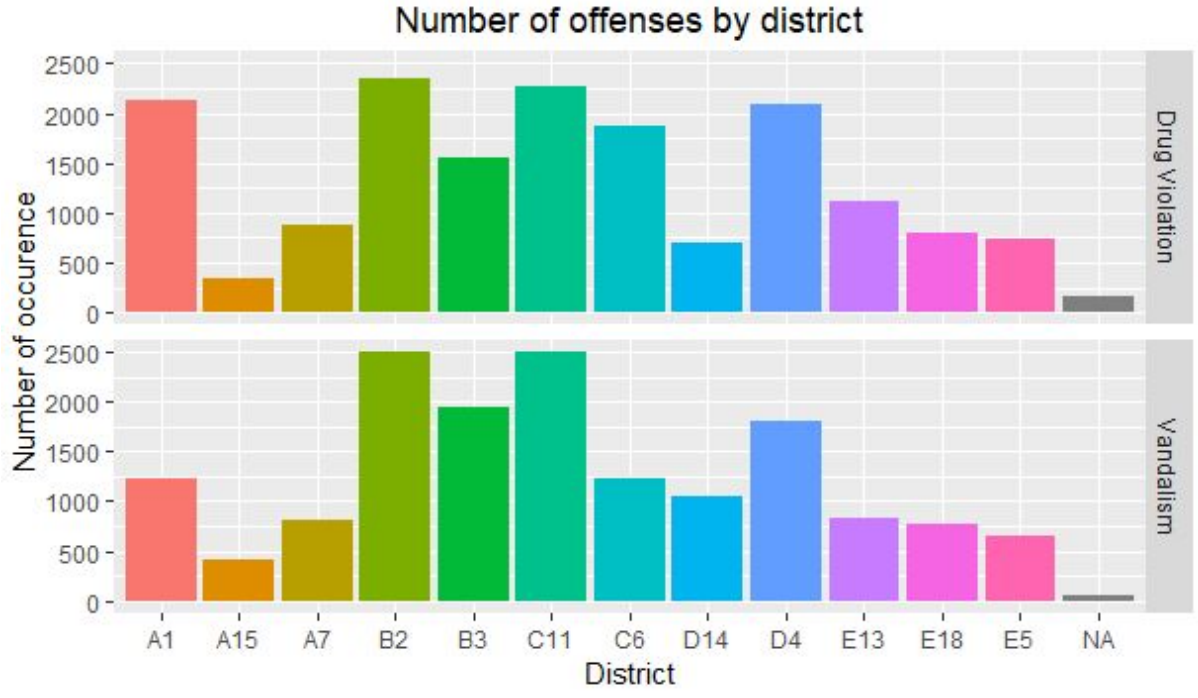
Sur les années 2016, 2017 et 2018, nous pouvons voir un même tendance sur les premiers mois de l'année, c'est-à-dire de janvier à mai. En effet, nous pouvons voir un recrudescence des crimes sur le mois de février, puis une reprise à la hausse des crimes jusqu'en mai. Nous avons bien une relation entre la période de l'année et les crimes commis.

Pour aller plus loin, nous pouvons essayer de corréler les crimes avec leur lieu. Pour cela nous avons tracé sur les 4 années, le pourcentage de crimes commis par district.



Comme les années 2015 et 2018 ne sont pas complètes, nous voulons regarder la proportion de crimes commis par district et non leur nombre total, nous devons donc diviser le nombre de crimes par le total des crimes sur l'année en question. Ainsi, nous pouvons voir que certains districts sont enclins à un taux de crimes plus élevés.

Cependant, nous présentons l'ensemble des crimes comme les fraudes, les insultes verbales, il n'y a donc normalement aucun lien entre le nombre de crimes élevé et la dangerosité potentielle des districts. Nous pouvons nous intéresser maintenant à deux cas de crimes qui dénotent un climat sous tension des districts : la vente de drogue et le vandalisme.



Nous pouvons remarquer une ressemblance flagrante entre la répartition des crimes totals par district avec le nombre de vandalismes et de trafic de stupéfiants par district. En effet, nous pouvons voir la même tendance entre les deux graphes précédents. Nous retrouvons en tête les districts B2, C11 et D4, alors que les districts A5, A7 et E5 répertorient le moins de crimes.

Ainsi comme nous pouvions le prédire, certains districts rassemblent un nombre important de crimes, nous pouvons donc mettre en relation la répartition totale des crimes avec le lieu des crimes. Certains districts semblent plus dangereux que d'autres. Un autre point à souligner, et qui nous permet d'affirmer cela, est que certains crimes sont révélateurs d'un climat de tension du district, comme la vente de drogue ou le vandalisme, mais aussi les homicides, et les fusillades. Ce climat sous tension entraîne une augmentation d'autres crimes.

Fait surprenant, il y a un lien entre la période de l'année et l'apparition des crimes, en effet la répartition des crimes n'est pas homogène sur l'ensemble de l'année, et nous pouvons retrouver la même tendance d'une année sur l'autre. Nous aurions pu aussi regarder la répartition des crimes sur les jours de la semaine.

Malheureusement, nous n'avons pas pu faire une étude approfondie des crimes car nous ne disposions pas d'information sur le criminel en question, comme ses revenus, sa situation familiale, ni même son casier judiciaire. Nous pouvons supposer sans le confirmer qu'il y a un lien entre le lieu des crimes, la gravité avec l'individu en question.