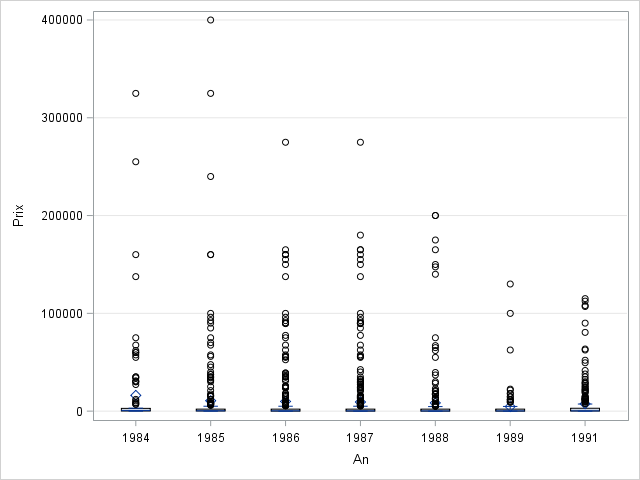
Projet SAS : Prédiction du prix de la drogue

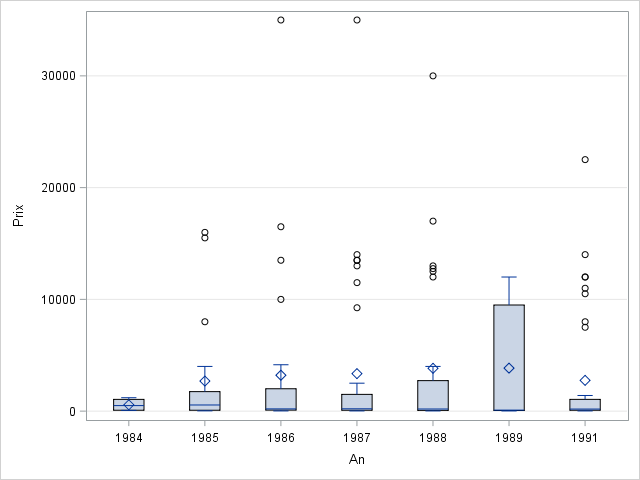
Le but de ce projet est de produire une régression linéaire prenant les bons paramètres en compte, afin de prédire le prix de la drogue suivant différents critères.

Nous avons eu pour cela à disposition une base de données contenant le prix de différentes drogues, en fonction des critères suivants : endroit de saisie, poids, type de drogue, pureté, année de saisie

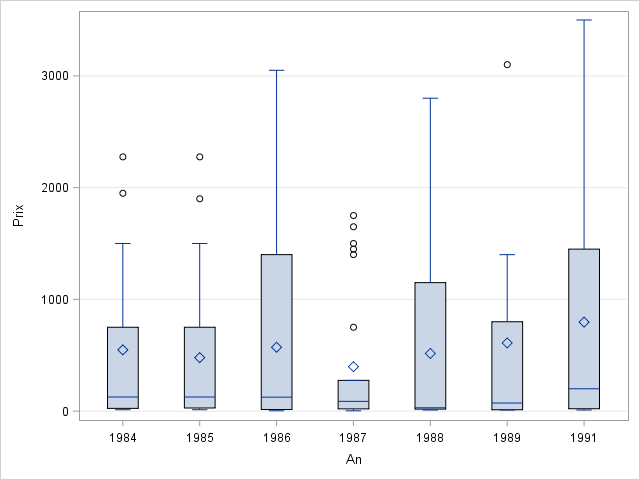
Concernant le critère pour l’année de saisie, entre 1984 et 1991, il est choisi de le laisser discret pour chaque année, sans regrouper certaines années, car d’un point de vue global, l’année a peu d’influence sur les ventes de drogue, comme on peut le voir ci-dessous :



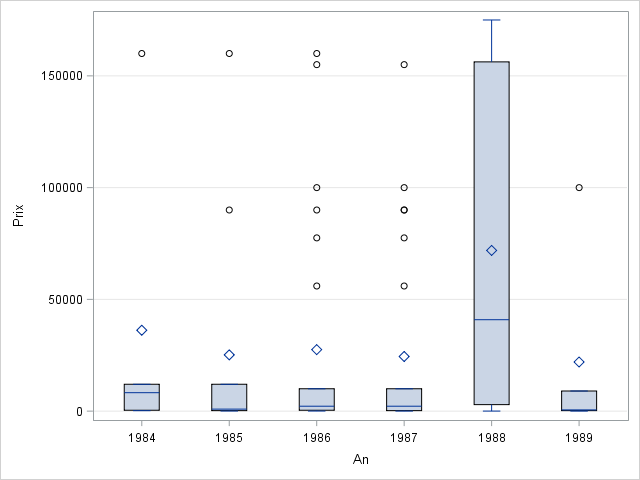
On remarque notamment en termes de médiane, que chaque année est à peu près similaire, malgré des étendues assez différentes. Pour montrer cela, on prend l’exemple de la meth :



Un regroupement aurait été néanmoins possible pour les années allant de 1985 à 1987, avec des prix pour chaque drogue à peu près similaire pour ces années (notamment en termes de médiane), avec un léger écart, le plus marqué étant pour le haschisch, comme présenté ci-dessous :

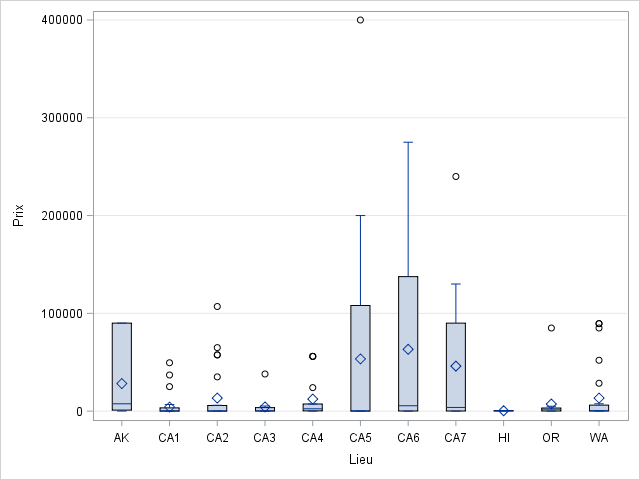


Toutefois, les autres années ne pouvaient être regroupés en raison d’écart important deuprix de certaines drogues, notamment entre 1988 et le reste, où certaines drogues ont vu leur prix très élevés ou au contraire très faibles, comme par exemple pour l’héroine W :

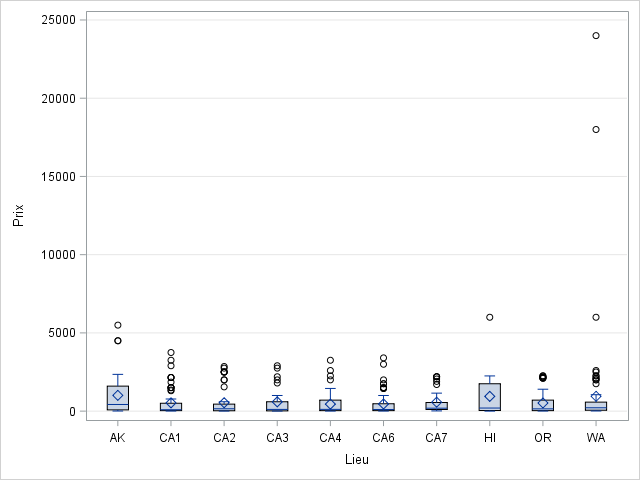


Ensuite, on se doute que le poids de la drogue a une influence sur le prix de chaque drogue, et que cette influence est à peu près linéaire. Toutefois il a fallu discrétiser le poids pour plus de visibilité, et il a été choisi de le classer en 3 catégories pour plus de lisibilité : Low, Medium, et High.

Concernant l’influence de l’endroit de saisie sur le prix de la drogue, nous avons étudié la répartition des prix en fonction des lieux pour chaque drogue, tout comme pour les années avant. Après analyse, nous avons choisi de regrouper les lieux géographiquement, car malgré certaines différences pour quelques drogues, globalement le prix était soit peu influencé par le lieu, soit au contraire quelques groupes se dégageaient clairement comme le montre l’histogramme suivant concernant l’héroine T :



Concernant l’absence d’influence du lieu sur certaines drogues, on peut prendre l’exemple de la marijuana :



Ainsi, 6 catégories ont été créées pour les lieux :

* Californie1 regroupant CA1, CA2, et CA3
* Californie2 regroupant CA4, CA5, CA6 et CA7
* AK pour AK
* Autre1 pour HI (Hawai)
* Autre2 pour OR (Oregon)
* Autre3 pour WA (Washington)

Enfin, pour la pureté, on remarque que tous les types de marijuana et le haschish sont purs à 100%. On décide ainsi de faire deux régressions linéaires : une sur les drogues pures (tous types de marijuana et haschish), où on ne prendra pas en compte la pureté, et les drogues non-pures (cocaine, crack, héroine et meth), où la pureté sera prise en compte.

Pour la seconde régression linéaire, une discrétisation de la pureté aurait peut-être été intéressante pour avoir un meilleur modèle, mais cela n’a pas été fait faute de délimitation facile à trouver pour créer différentes catégories, contrairement au poids.

Ainsi, deux types de régressions linéaires ont été faites, regroupant les drogues contre décrit au-dessus, la première (sur les drogues non-pures) prenant en compte l’année, la pureté, le lieu et le poids, ces deux derniers étant classés par catégorie, et l’autre regroupent les drogues pures, sur les mêmes critères minus la pureté.

De plus, pour chaque type de régression, on a voulu exprimer soit le prix en fonction de ces critères, soit le log du prix, soit la racine du prix. Au final, pour le premier type de régression, la racine du prix en fonction des critères obtient le milleur R-carré ajusté, avec près de 0.70, tandis que pour le second type, c’est le log du prix en fonction des critères qui permet d’obtenir le meilleur modèle, avec un R-carré ajusté d’environ 0.65 .

Ces modèles sont satisfaisants, mais peuvent être améliorés, par exemple en prenant en compte un recoupage des années différent, en trois catégorie par exemple, avec les années 1984 à 1987, puis 1988, puis 1989 à 1991, ou encore en regroupant différemments les drogues ou les lieux, à partir d’observations plus précises. De plus, on pourrait chercher à s’avoir si exprimer le prix en fonction du log ou de la racine carrée de chaque critère ne permettrait pas d’obtenir un modèle plus exact.