ABOUZAID Mehdi
LEE Nathan

Projet M8

Recherche de liens entre le comportement et la consommation de drogues

A l’attention de M. Canu

Table des matières

Introduction **1**

La variable Drogue **2**

L'Age et le Genre des individus **5**

Les pays d'origine et ethnicité **6**

La variable Comportement **7**

L'art est graisse ion multiple **7**

Tableau des R^2 8

Régression multiple avec différentes catégories de drogues **9**

Diagnostic de la régression multiple **11**

Elimination des points aberrants 13

L'As Epais **14**

Régression en utilisant les données de l'ACP 17

Test du Q'Hideux **18**

Corrélation entre les variables **20**

Conclusion **21**

Introduction

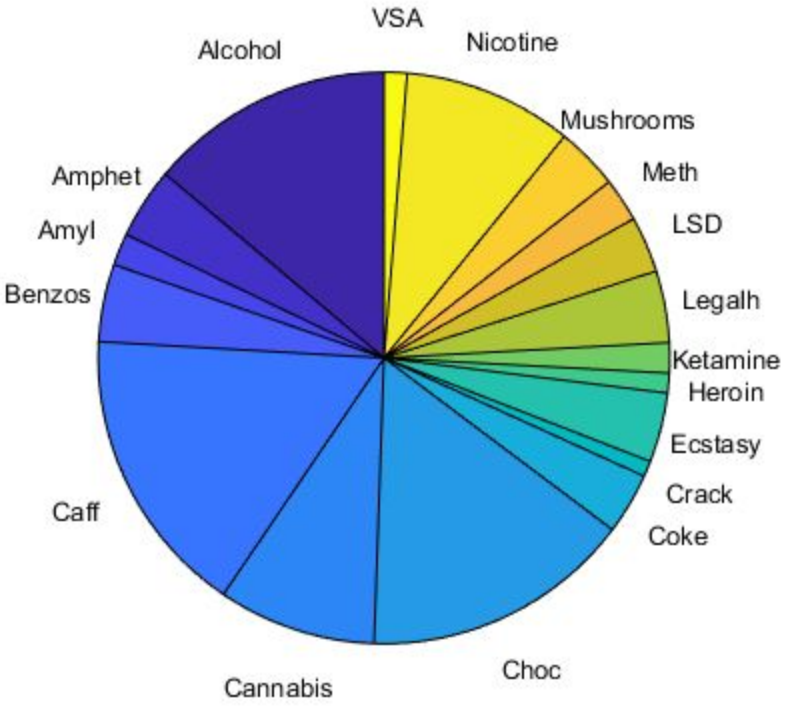
Pour notre EC de M8 d'introduction à la science des données, un projet fut mis en place pour analyser et exploiter les données de notre choix. Nous avons mis la main sur des données quantifiant la consommation de différentes drogues par des individus classés selon différents critères. La diversité et l'importance des variables (1885 au total) nous permettent d'avoir des données fiables et variés. C'est ainsi que les questions de liaison entre mode de vie et consommation de drogues et d'influence de ces drogues sur les individus se posent.

Nous disposons donc d'une base de données de 1885 individus et 31 observations : âge, genre, niveau d'éducation, pays, ethnie, traits de caractère (neuroticisme, extraversion, ouverture à l'expérience, agréabilité, conscienciosité, impulsivité, adepte de sensations) et leur consommation de différentes drogues (alcool, amphétamine, pentylamine, benzodiazépine, caféine, cannabis, chocolat, cocaïne, crack, ecstasy, héroïne, kétamine, euphorisant légal (legal high), LSD, méthadone, champignons hallucinogènes, nicotine, semeron, substances volatiles). Toutes ces données ont été quantifiés. Les participants ont été questionnés sur l'usage de 18 drogues légales ou illégales (incluant une drogue fictive Semeron pour détecter les menteurs). Pour chaque drogue, il fallait choisir parmi l'une des réponses suivantes : n'a jamais consommé cette drogue, l'a consommé il y a plus de 10 ans, moins de 10 ans, moins d'un an, moins d'un mois, moins d'une semaine ou dans la journée.

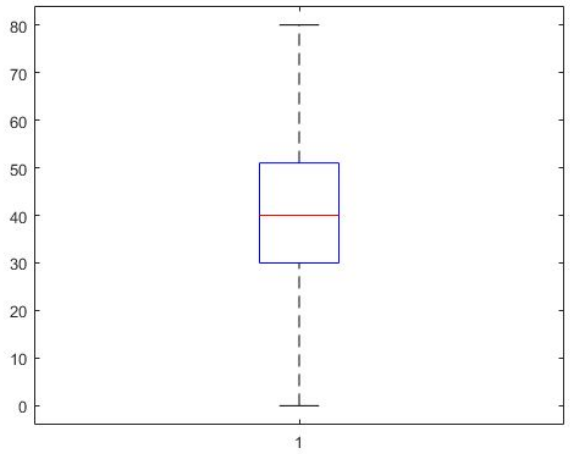
Nous avons commencé par organiser ces données en ordonnant de 0 à 6 la consommation de chaque drogue dans le temps et en éliminant les individus ayant répondus favorablement à la consommation de la drogue fictive Semer puisque ce sont probablement des menteurs et auraient donc fourni des valeurs aberrantes. Nous sommes donc passé de 1885 à 1877 personnes, soit 8 retraits.

La variable Drogue

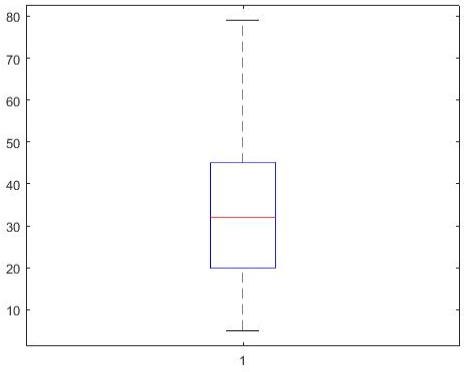
Nous avons d'abord décidé de nous intéresser aux drogues et à leur usage. Le plus simple a été de créer un camembert pour observer les drogues les plus consommées. Ainsi 5 de ces drogues en sont ressorties clairement : Nicotine, Chocolat, Caféine, Cannabis et Alcool. En effet ces drogues là sont légales et facilement disponibles, à part le cannabis dans certains pays ou états. Il y a une drogue qui n'apparaît pas sur le schéma, car celle-ci est la drogue fictive Semer : puisque celle-ci a très peu de poids par rapport aux autres, nous ne l'avons pas mise sur ce graphe.

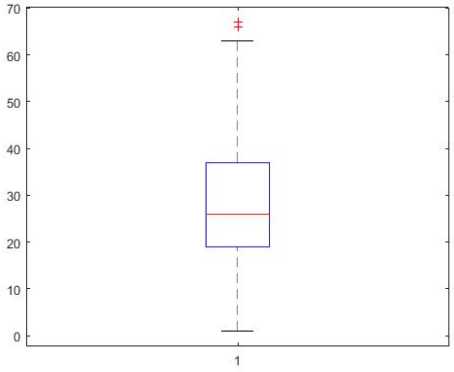


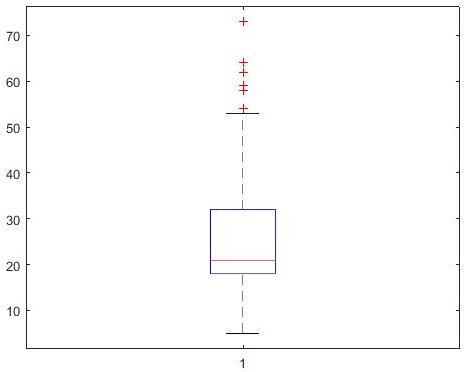
Nous nous sommes ensuite intéressés à l'étendu de l'usage de ces drogues : en faisant un cumul pour chaque personne, nous avons créé une boîte à moustache, permettant de mieux comprendre comment se découpait les valeurs.

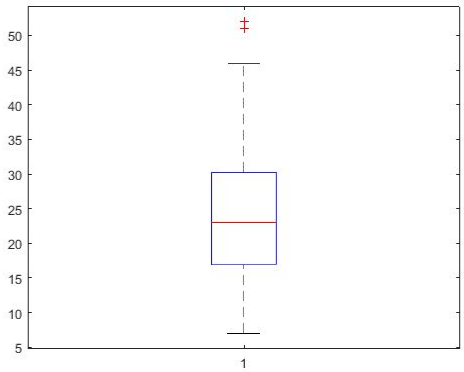


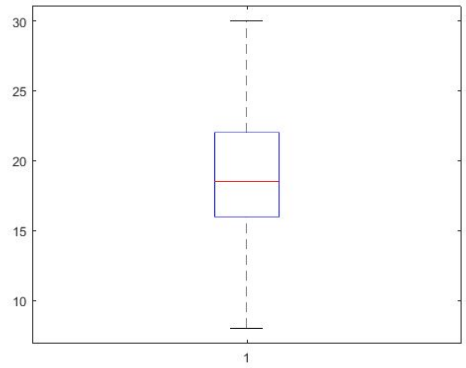
Ainsi la valeur maximale ne dépasse pas 80, alors que techniquement elle aurait pu aller jusqu'à 112. La médiane se trouvant aux alentours de 40, avec un premier quartile à 30 et un troisième quartile à 50, on n'en apprend pas grand-chose, puisque d'un premier coup d'oeil les valeurs sont assez étendues. Ainsi nous avons décidé de refaire des boîtes à moustache pour chaque tranche d'âge.







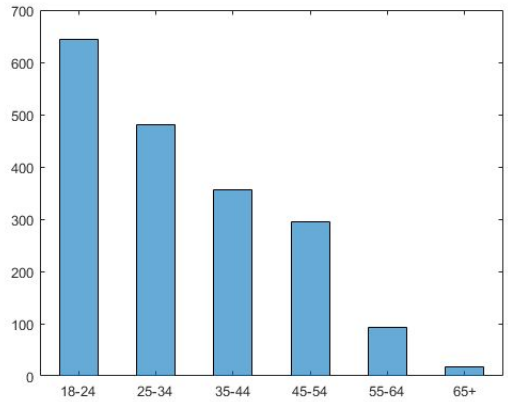


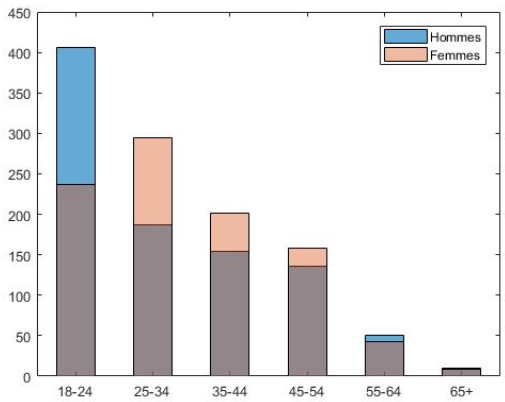


Nous pouvons alors observer le fait que la répartition des mesures (drogues) est assez homogène malgré quelques points hors épure, sauf pour la tranche d'âge 35-44. En effet, celle-ci possède plusieurs points hors épure et sa médiane est proche du premier quartile. On peut donc en déduire que certaines individus âgés entre 35 et 44 ans consomment plus de drogues que les autres du même âge, qui eux en consomment globalement peu par rapport aux autres tranches d'âge.

L'Age et le Genre des individus

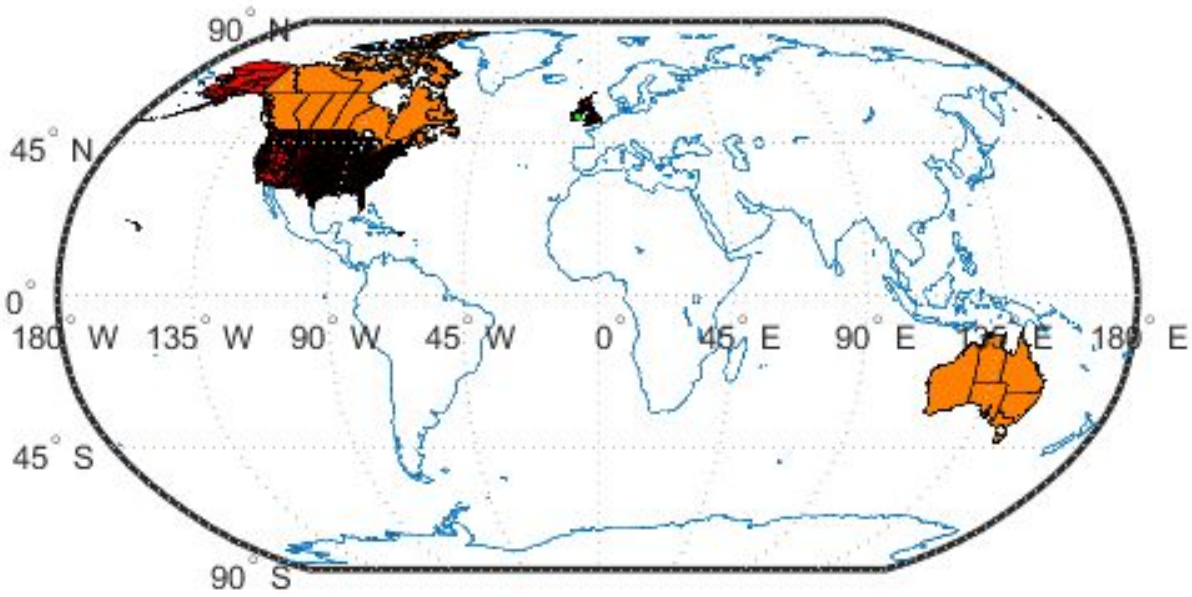
L'âge des participants n'est pas donné exactement mais par intervalle. A l'aide d'un histogramme, nous avons pu observer que plus de 50% des participants ont entre 18 et 34 ans. En effet, 643 individus sont âgés de 18 à 24 ans, 481 de 25 à 34 ans, 356 de 35 à 44 ans, 294 de 45 à 54 ans, 93 de 55 à 64 ans et 18 âgés de plus de 65 ans.
Le faible nombre d'individus âgés de plus de 55 ans nous indique qu'il sera plus compliqué de savoir si les drogues ont un effet à long terme sur le comportement.





Nous avons ensuite essayé de voir si le sexe était un paramètre à prendre en compte dans notre étude étant donné qu'il y a 50% de réponses féminines et masculines. Après avoir superposé l'histogramme des hommes et celui des femmes, on remarque que les hommes interrogés sont en moyenne plus jeunes par rapport aux femmes qui elles sont assez dispersées sur les tranches d'âge. Or nos boîtes à moustaches ont montrés que les personnes de 18-24 ont tendance à boir beaucoup : ceci est probablement dû aux hommes étant en majorité par rapport aux femmes. Donc nos données sont contrebalancés par le fait que les hommes soient en général plus jeunes que les femmes parmi les participants.

Les pays d'origine et ethnicité



Nous avons, à l'aide de trois couleurs (rouge pour une forte concentration, orange pour une concentration moyenne et vert pour une faible concentration), déterminé la proportion d'individus selon leur pays d'origine. Nous trouvons que les individus interrogés viennent principalement d'Angleterre (1044) et des Etats-Unis (557). On retrouve derrière 87 personnes du Canada, 54 d'Australie, 20 d'Irlande et en dernière place la Nouvelle-Zélande avec seulement 5 individus. Nos données pourraient déjà avoir des préjugés puisque ces échantillons ne sont peut-être pas représentatifs de l'échelle mondiale. De plus si on regarde la variable Ethnicité, on se rend compte que plus de 90% des participants sont blancs : ce qui implique encore qu'on risque de ne pas avoir la réalité représentée correctement.

La variable Comportement

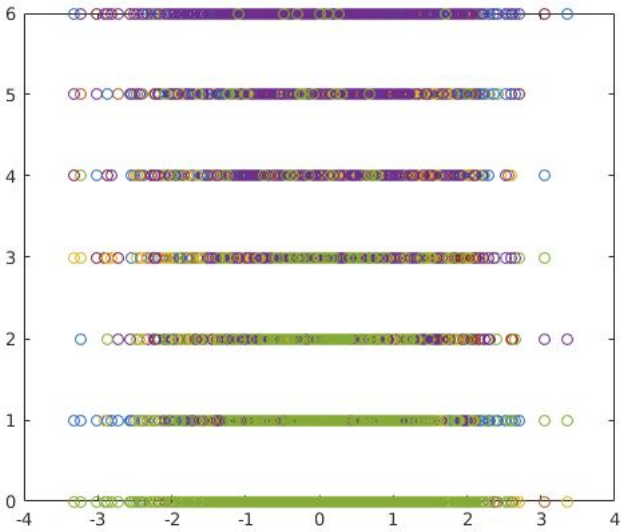
Il faut comprendre que la variable comportement comporte 7 traits, qui sont censés ensemble définir pour neuroticisme, extraversion, ouverture à l'expérience, agréabilité, conscienciosité, les 5 traits supposant définir la personnalité d'une personne. Donc dire que les drogues définissent la personalité d'une personne est une hypothèse tout de même assez poussée, même si logiquement quelqu'un qui consomme des drogues sera ouvert à d'autres expériences. Les deux derniers, l'impulsivité et adepte des sensations sont des traits à part, mais intéressants dans le sens que l'on pourrait les croire liés à l'usage de drogues.

L'art est graisse ion multiple

Nous avons d'abord effectué une régression multiple de chaque comportement en fonction des drogues : nous avons posé le modèle y=a0+a1x1+a2x2+a3x3+etc.. avec x1,x2,x3… les différentes drogues pour chacun des différents comportements, et de là nous avons calculé les coefficients de corrélations. Nous avons posé ce modèle puisque logiquement plus on utilise de drogues , plus le comportement d'une personne deviendra excessif, de façon linéaire.

```
n=length(Drogue);
X=[ones(n,1) Drogue];
for i=1:7
        y=Comportement(:,i)%on l'a fait pour chaque comportement
        a = (X'*X)\(X'*y); % vecteur des coefficients a et b du modèle y=ax+b
        e = y-X*a; % résidus
        R2 = 1 - e'*e/sum((y-mean(y)).^2); %coefficient de corrélation
end;
```

Nous avons voulu observer les résidus en fonction de y puis en fonction de X pour déterminer si notre modèle était correctement posé.



Pour le graphe des données X en fonction des résidus e, nous avons obtenu quelque chose d'assez anormal que nous n'avons pas tout de suite compris.

Tableau des R²

Nous avons décidé de garder notre modèle linéaire et de continuer à tenter de calculer les R².

Comportement	**R2**
Nscore	0.0925
Escore	0.0487
Oscore	0.2205
Ascore	0.0617
Cscore	0.1171
Impulsive	0.1477
SS	0.2807

Etant donné que les R² sont assez faibles, nous avons décidé de séparer les drogues en catégorie pour essayer de trouver une dépendance entre un groupe de drogues et un comportement.

Régression multiple avec différentes catégories de drogues

Après quelques recherches sur Internet, nous avons pu distinguer 3 grandes catégories : perturbateurs, dépresseurs et stimulants.

Catégorie	**Drogue**
Perturbateur	Ketamine, Cannabis, Mushrooms, LSD, VSA
Dépresseur	Heroin, Alcool, Benzos
Stimulant	Coke, Ecstasy, Amphet, Amyl, Choc, Crack, Caff, Legalh, Meth, Nicotine

Nous avons donc fait une régression avec chaque catégorie de drogues. Nous nous apercevons que les R^2 sont toujours faibles. Les différentes catégories de drogues n'ont apparemment pas de lien direct avec les différents comportements, ou pas suffisant.

Comportement	**R2 Perturbateur**	**R2 Dépresseur**	**R2 Stimulant**
Nscore	0.0202	0.078	0.0503
Escore	0.0032	0.0206	0.0384
Oscore	0.2091	0.0445	0.1334
Ascore	0.0325	0.0399	0.0534
Cscore	0.089	0.0489	0.1063
Impulsive	0.1172	0.0656	0.1333
SS	0.2471	0.0893	0.2331

Nous avons aussi tenté de cumuler l'usage des drogues en une seule variable *X=sum(Drogue)* pour ensuite refaire un système du type y=ax+b, mais celui-ci a donné des R^2 encore plus petits, ce qui implique qu'il existe quand même une faible linéarité entre différentes drogues et comportements. Nous avons aussi tenté d'inverser les valeurs dans drogue : privilégier ainsi les effets à long terme sur le comportement plutôt que les effets à court terme, mais encore une fois, cela n'a pas non plus aider au niveau des R^2 qui sont au contraire devenus plus petits encore une fois. Et enfin, nous avons décidé d'inverser la tendance, pour voir si au contraire, un comportement singulier pouvait pousser une personne à consommer de la drogue.

Drogues	R2
Alcool	0.0202
Amphet	0.1620
Amyl	0.0472
Benzos	0.1523
Caff	0.0066
Cannabis	0.3145
Choc	0.0043
Coke	0.1535
Crack	0.0568

Drogues	R2
Ecstasy	0.1928
Heroin	0.0909
Ketamine	0.0818
Legalh	0.2333
LSD	0.2073
Meth	0.1111
Mushrooms	0.2201
Nicotine	0.1302
VSA	0.0844

Nous remarquons que l'on monte plus avec ce modèle au niveau du coefficient de R2 grâce au Cannabis, mais 0.3145 n'est tout de même pas suffisant pour établir une bonne corrélation entre Comportement et Drogues.

Puisque Oscore et SS ont pour leur part le R^2 le plus élevé, nous avons voulu sélectionner seulement ces deux-ci pour refaire ce que nous venons de faire. Nous obtenons le tableau suivant:

Drogues	R2
Alcool	0.0124
Amphet	0.1195
Amyl	0.0396
Benzos	0.0729
Caff	0.0027
Cannabis	0.2661
Choc	0.0017
Coke	0.1206
Crack	0.0363

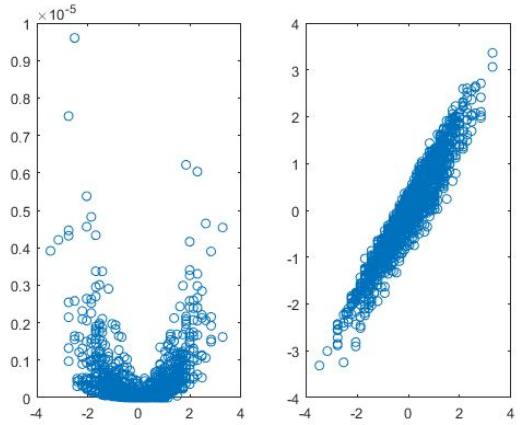
Drogues	R2
Ecstasy	0.1710
Heroin	0.0482
Ketamine	0.0669
Legalh	0.1912
LSD	0.1885
Meth	0.0565
Mushrooms	0.1969
Nicotine	0.0977
VSA	0.0649

Nous remarquons que même si les R^2 ont baissés, ils restent tout de même assez élevés par rapport aux autres, ce qui soutiendrait l'idée que Oscore et SS ont le plus d'influence sur les drogues, et vice-versa. Mais nous n'avons toujours pas de coefficients de corrélation satisfaisants, c'est pourquoi nous devons effectuer un diagnostic de cette régression multiple.

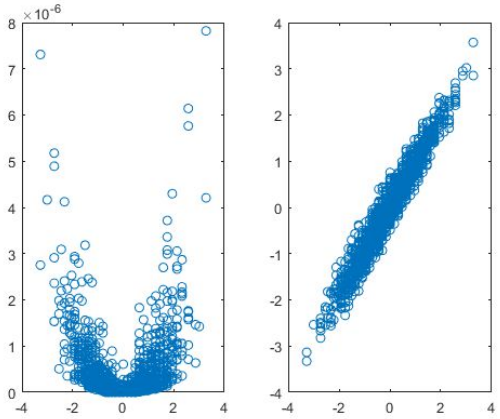
Diagnostic de la régression multiple

On fait donc un diagnostic pour trouver et éliminer les valeurs aberrantes. Voici le code Matlab permettant de calculer la contribution de chaque point selon son erreur et résidu.

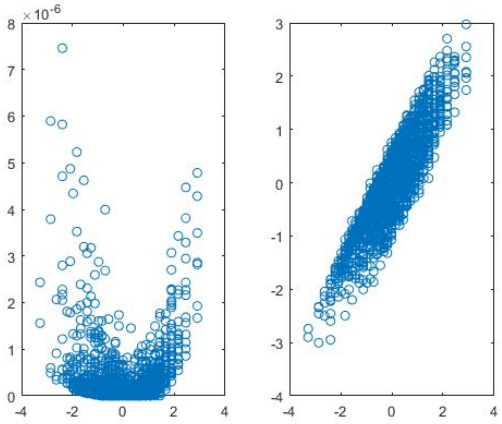
```
n=length(Drogue);
X=[ones(n,1) Drogue];
for i=1:7
    y=Comportement(:,i);%ici y prend les comportements de 1 à 7
    [n,p] =size(X);
    a = (X'*X)\(X'*y);
    e = y-X*a;
    s2 = e'*e/(n-p);  % calculs des résidus standardisés
    R2 = 1 - e'*e/sum((y-mean(y)).^2);
    h =diag(X*((X'*X)\(X')));
    c = h./(1-h).^2/p.*e.^2/(e'*e); % contribution
    figure(i+9)
    subplot(1,2,1)
    plot(y,c,'o') % pour obtenir le graphe des contributions de chaque comportement
    hold on
    subplot(1,2,2)
    plot(y,e,'o') %graphe des erreurs pour chaque comportement
end;
```



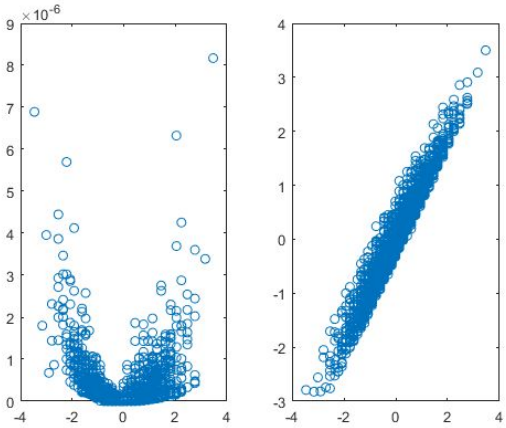
Contributions et erreurs pour SS



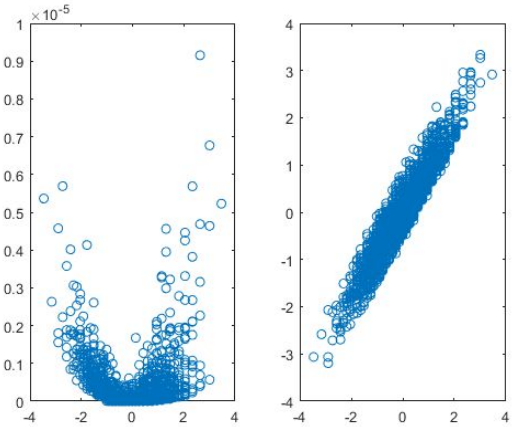
Contributions et erreurs pour Impulsive



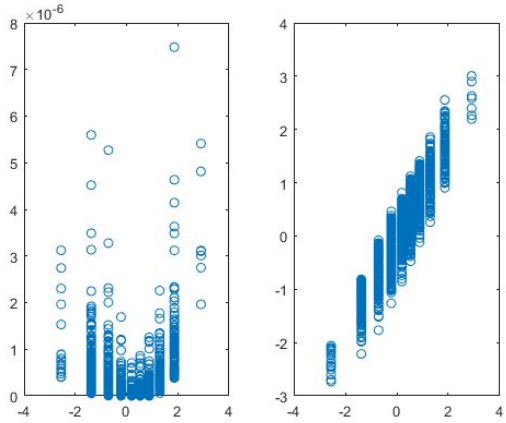
Contributions et erreurs pour Cscore



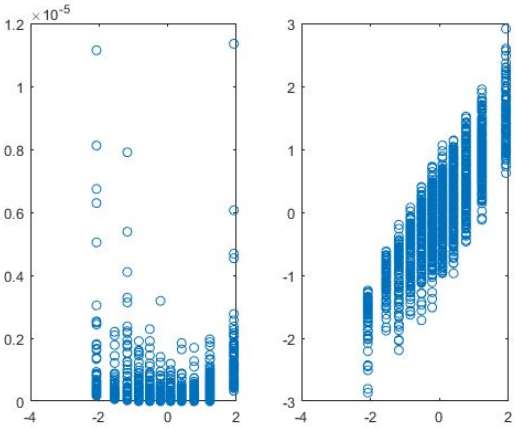
Contributions et erreurs pour Ascore



Contributions et erreurs pour Oscore



Contributions et erreurs pour Escore



Contributions et erreurs pour Nscore

En observant le nuage de points des contributions et des erreurs, on remarque que le nuage est **structuré** et qu'il y a des points aberrants. Nous avons décidé d'éliminer les points aberrants qui seraient trop éloignés par rapport aux autres et avons cherché un autre modèle à appliquer.

Elimination des points aberrants

Nous avons considérés à chaque fois que des contributions supérieures à 1.5e-06 correspondent à des points aberrants comme les graphes se ressemblent tous, même si au final nous n'avons pas forcément le même nombre de valeurs restantes à chaque fois :

```
j=1;
for i=1:length(X)
        if c(i)<1.5e-06 % le même pour tous
           xs(j,:)=X(i,:);
           ys(j,:)=y(i);
           j=j+1;
        end
end
```

Nous avons fait ceci pour chaque Comportement de manière individuel. Nous avons calculés deux R2 pour chaque Comportement en variant la condition sur la contribution c. L'un des R^2 est plus grand que l'autre car la condition sur la contribution c est plus forte mais nous perdons beaucoup de valeurs.
Puis nous avons réalisé que nous pourrions regrouper tout cela en enlevant les valeurs aberrantes pour chaque comportement et calculer ensuite le R2 avec toutes ces valeurs en moins. Le tableau ci-dessous montre les différents résultats obtenus.

Comportement	R2 avec c<1.5e-06	R2 avec c<1.0e-07	R2 avec 1179 valeurs
Nscore	0.1301 avec 1785 valeurs	0.3091 avec 990 valeurs	0.2986
Escore	0.504 avec 1794 valeurs	0.1287 avec 961 valeurs	0.1472
Oscore	0.2548 avec 1799 valeurs	0.5589 avec 993 valeurs	0.1417
Ascore	0.766 avec 1798 valeurs	0.1794 avec 959 valeurs	0.0585
Cscore	0.1597 avec 1792 valeurs	0.3916 avec 1017 valeurs	0.2393
Impulsive	0.1724 avec 1821 valeurs	0.4141 avec 983 valeurs	0.0506
SS	0.3468 avec 1805 valeurs	0.6263 avec 994 valeurs	0.1184

Nous constatons qu'il y a beaucoup de points aberrants, et que même en forçant la main, c'est à dire en éliminant la moitié des valeurs, les R^2 sont pour la plupart bien trop petits pour qu'il y ait une réelle corrélation, mis à part Oscore et SS. On remarque aussi que les valeurs aberrantes pour chaque comportement ne sont pas les mêmes, d'où le fait que les R^2 n'augmentent pas pour la plupart lorsque les valeurs aberrantes sont enlevées. Ainsi les comportements seraient plus ou moins indépendants mais ceci ne nous aide en rien pour notre hypothèse.
Nous avons aussi essayé de faire la régression suivant d'autres modèles tels que les modèles $y=ax^2+bx+c$ ou encore $y=ax^3+bx^2+c$ mais aucun changement significatif n'apparait.

Nous sommes donc obligés de passer par l'ACP pour voir si notre hypothèse s'avère correcte.

L'As Epais

L'analyse en composantes principales (ACP) est une méthode qui consiste à transformer des variables liées entre elles (corrélées) en nouvelles variables décorrélées les unes des autres. Cela permet au statisticien de réduire le nombre de variables et de rendre l'information moins redondante. *(Wikipédia)*

Nous commençons donc par centrer et réduire les données :

```
Xc = Drogue - ones(n,1)*mean(Drogue); % centrer = ramener l'origine au centre du nuage de points
Xn = Xc./ (ones(n,1)*std(Drogue,1));       % réduire =  atténuer l'effet d'échelle (si unités différentes)
```

Nous calculons ensuite les plus grande valeurs propres :

```
C = (Xn'*Xn)/n;
[V L] =eig(C);
V         % vecteurs propres
diag(L) % valeurs propres
```

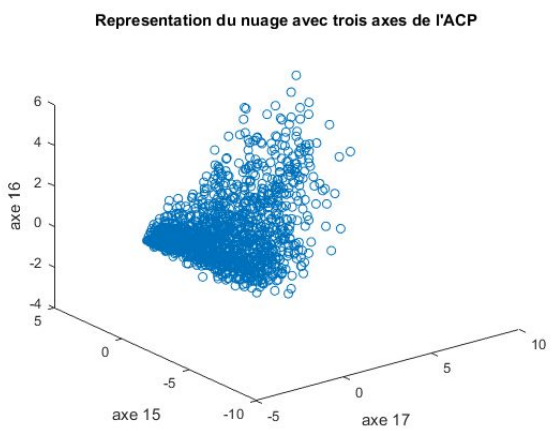
Nous calculons ensuite l'influence de chaque valeur propre puis dressons un tableau des influences cumulées grâce au script Matlab suivant :

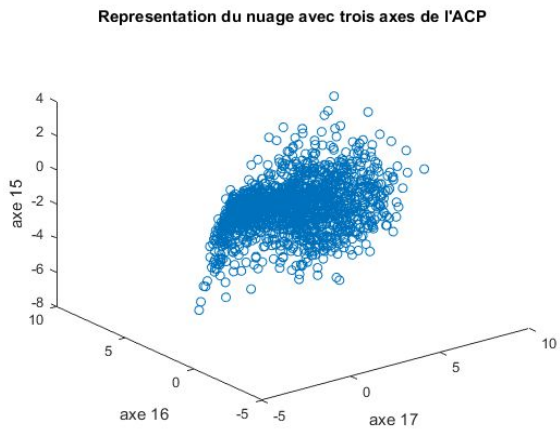
```
vp = sort(diag(L),'descend');
vpc = 100*cumsum(vp)/sum(vp);
[vp vpc]
```

Valeurs propres	Influence cumulée (%)
5.6890	33.4647
1.5403	42.5253
1.2569	49.9187
1.0041	55.8249
0.9455	61.3865
0.8725	66.5189
0.8067	71.2642
0.7594	75.7310
0.7559	80.1778
0.5407	83.3582
0.5101	86.3590
0.4768	89.1636
0.4437	91.7736
0.4163	94.2227
0.3691	96.3940
0.3370	98.3761
0.2761	100

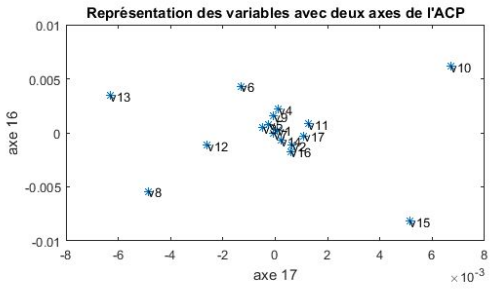
Nous remarquons que pour représenter plus de 50% des données il nous faut 4 valeurs propres. Comme il est compliqué de faire une projection suivant 4 axes, nous allons donc nous limiter aux 3 premières principales à savoir ceux représentant 49,9% de l'information.

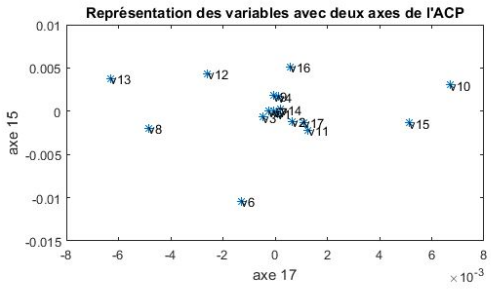
Nous projetons les observations suivant différents axes principaux avec le script Matlab $U = Xn*V$ et nous obtenons les graphes suivants :

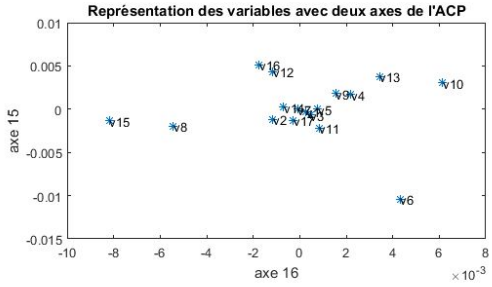




Comme il est compliqué d'en tirer des informations, nous projetons ensuite les variables en fonction de nos axes principaux pour discerner plus facilement les différents groupes et ainsi voir si il y a une similitude des variables







Plus les variables sont proches entre elles et plus elles sont corrélées. Ici l'échelle étant en 10^−3, nos variables sont donc très corrélées car elles sont très proches.

Nous allons essayer de faire une régression linéaire sur ces variables pour voir si notre R2 augmente.

Régression en utilisant les données de l'ACP

On crée une nouvelle matrice C qui calcule la composante principale des individus :
$C = Xn*V$ où Xn est notre matrice de données centrées et réduites et V la matrice des vecteurs propres

On approxime ensuite nos données X :
$X = C*Vn'$ où et Vn est la matrice des vecteurs propres centrés et réduits.

Comportement	**R2 de base**	**R2 de base avec ACP**	**R2 sans valeurs aberrantes avec ACP**
Nscore	0.0925	0.1181	0.1301
Escore	0.0487	0.0530	0.0504
Oscore	0.2205	0.2619	0.2548
Ascore	0.0617	0.1075	0.0766
Cscore	0.1171	0.1728	0.1597
Impulsive	0.1477	0.2432	0.1724
SS	0.2807	0.3150	0.3468

Après avoir fait l'ACP sur les données "brutes" et les données sans valeurs aberrantes, nous nous rendons compte que les coefficients de corrélation sont sensiblement les mêmes. L'ACP n'a pas été utile, ce qui est normal car nos variables sont fortement corrélés au départ. Il semble donc que nos variables ne soient pas vraiment linéaires comme observé précédemment.

Nous effectuons désormais des tests, pour voir si au final on ne s'est pas trompé quelque part.

Test du Q'Hideux

Nous avons effectué un test du Chi2 sous la forme d'un test d'adéquation, permettant de savoir si des données sont indépendantes ou non. On pose H0 : elles sont indépendantes et H1 : elles ne le sont pas.
Dans notre cas nous allons poser ce test sur nos différentes drogues en comptant combien de fois 0,1,2,3,4,5,6 (la consommation selon le temps) apparaît pour chaque variable. Voici notre tableau de contingence:

Apparition / Drogue	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
0	33	973	1299	999	27	413	32	1036	1622	1020	1600	1488	1092	1069	1424	982	428	1452
1	34	230	210	116	10	207	2	160	67	112	68	43	29	257	39	208	193	199
2	68	241	236	230	24	266	10	267	109	232	91	140	195	175	95	259	203	133
3	197	196	91	234	59	210	53	257	59	275	65	129	321	213	148	272	184	59
4	284	75	24	119	106	138	295	98	9	154	24	40	109	96	50	114	106	13
5	758	61	14	84	271	185	680	40	9	63	16	33	64	55	48	39	156	14
6	503	101	3	95	1380	458	805	19	2	21	13	4	67	12	73	3	607	7

On remarque que le test du chi2 ne peut pas s'effectuer sur ce tableau car il y a des effectifs inférieurs à 5. Il faut donc fusionner des colonnes afin d'obtenir un nouveau tableau valide.

Apparition\Drogue	1	2 et 3	4	5	6 et 7	8 et 9	10	11 et 12	13	14	15 et 16	17	18
0	33	2272	999	27	445	2658	1020	3088	1092	1069	2424	428	1452
1	34	440	116	10	209	227	112	111	29	257	247	193	199
2	68	477	230	24	276	376	232	231	195	175	354	203	133
3	197	287	234	59	263	316	275	194	321	213	420	184	59
4	284	99	119	106	433	107	154	64	109	96	164	106	13
5	758	75	84	271	865	49	63	49	64	55	87	156	14
6	503	104	95	1380	1263	21	21	17	67	12	76	607	7

On effectue le test sur le nouveau tableau avec le script Matlab suivant :

```
m =sum(O);
n =sum(O,2);
T = n*m/sum(n)
D =sum(sum((O-T).^2./T))
ddl = (length(n)-1)*(length(m)-1)
pval = 1 - cdf('chi2',D,ddl)
```

On trouve le tableau théorique suivant :

T	1	2 et 3	4	5	6 et 7	8 et 9	10	11 et 12	13	14	15 et 16	17	18
0	944.3	1888.7	944.3	944.3	1888.7	1888.7	944.3	1897.7	944.3	944.3	1897.7	944.3	944.3
1	121.3	242.5	121.3	121.3	242.5	242.5	121.3	243.7	121.3	121.3	243.7	121.3	121.3
2	165.1	330.3	165.1	165.1	330.3	330.3	165.1	331.9	165.1	165.1	331.9	165.1	165.1
3	167.8	335.6	167.8	167.8	335.6	335.6	167.8	337.2	167.8	167.8	337.2	167.8	167.8
4	102.9	205.9	102.9	102.9	205.9	205.9	102.9	206.9	102.9	102.9	206.9	102.9	102.9
5	143.8	287.6	143.8	143.8	287.6	287.6	143.8	289	143.8	143.8	289	143.8	143.8
6	231.7	463.4	231.7	231.7	463.4	463.4	231.7	465.6	231.7	231.7	465.6	231.7	231.7

On a une distance du chi2 égale à 22387, 72 degrés de liberté et une p-valeur de 0.
Comme la distance du chi2 est très grande, que les tableaux sont différents et que la p-valeur est inférieur à 0.05 (alpha), on rejette H0 et donc accepte H1. On en déduit donc que les variables sont fortement liées.

Corrélation entre les variables

Comme les variables sont fortement liés, nous testons maintenant si le comportement de chaque variable peut être fonction des autres à l'aide d'une nouvelle régression suivant le modèle y=ax+b

Nous faisons d'abord le cas de chaque drogue en fonction des autres drogues puis établissons le tableau suivant :

Drogue	R2
Alcool	0.0455
Amphet	0.4653
Amyl	0.2353
Benzos	0.4223
Caff	0.0539
Cannabis	0.5396
Choc	0.0330
Coke	0.5379
Crack	0.3378

Drogue	R2
Ecstasy	0.5977
Heroin	0.4193
Ketamine	0.3530
Legalh	0.4751
LSD	0.5363
Meth	0.3953
Mushrooms	0.5584
Nicotine	0.3274
VSA	0.1854

Nous remarquons que le R^2 des drogues usuels est très faible comme le chocolat, la caféine ou encore la nicotine car plusieurs personnes ne consomment que celles-ci et pas les autres, ainsi ces drogues usuels ne peuvent pas représenter les autres drogues. Au contraire les drogues avec un R^2 élevé sont consommés par des personnes prenant les drogues dites usuels ainsi que la drogue en question.

Conclusion

Ce projet nous aura été très utile puisqu'il nous aura permis d'utiliser les connaissances que nous avons utilisés en cours pour les appliquer sur des données que nous avons pris au hasard, données qui se sont avérées assez uniques.
Ce projet, étant malgré tout un échec de notre part dans le sens que nous n'avons pas réussi à prouver notre hypothèse, nous aura tout de même permis de progresser, et de comprendre comment à l'avenir nous pourrons mieux utiliser les connaissances à notre disposition pour réussir un projet.

Les régressions que nous aurons effectués n'auront pas abouti, et même en utilisant différents modèles, nous n'avons pas réussi à augmenter le R2 de façon satisfaisante. Mais en enlevant les valeurs aberrantes et en passant par une ACP, nous avons tout de même réalisé que 2 comportements ressortent : **l'ouverture à l'expérience, et l'adepte des sensations**. Le fait que ceux-ci ressortent s'avère plutôt logique puisque quelqu'un qui consomme des drogues aurait en effet tendance à être ouvert puisqu'il effectue des actions que la plupart de gens n'oserait jamais faire.

Le test du chi2 aura permis de nous montrer que les variables sont en effets liées puisque la pvaleur nous donne 0. Ce qui est logique puisque l'usage d'une drogue pousserait l'usage d'autres drogues chez les personnes concernées, sous une certaine limite bien sur.

Au final il s'avère que la consommation des drogues n'a pas de réelle influence sur le comportement d'un individu, mais qu'au contraire, un comportement dangereux pousserait les gens à consommer des drogues.