# Processo Seletivo 2024: IEEE Computational Intelligence Society, UnB

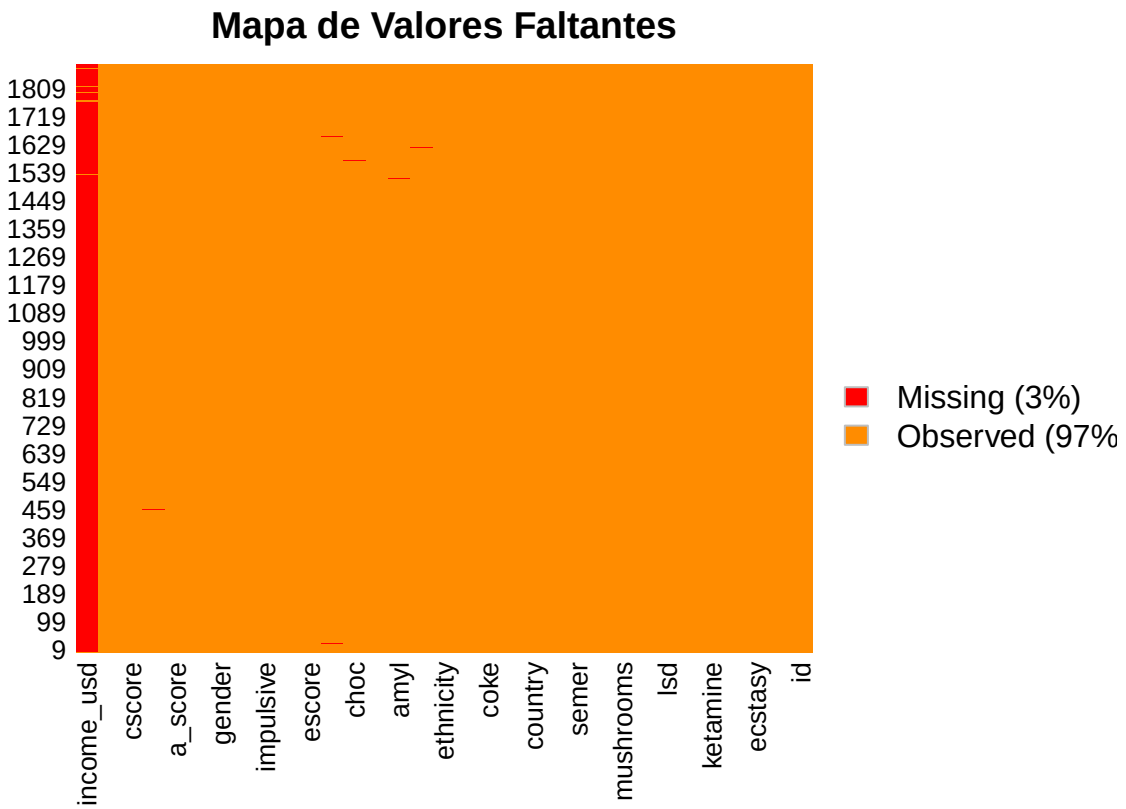
Luiz Paulo Tavares Gonçalves

2024-07-04

## Import dataset, cleaning & Validation

Após a importação do dataset os nomes das colunas são padronizados, assim, retirando caracteres especiais e espaços e, em segundo, é validado a não existência de duplicação de pacientes (no presente caso, duplicação de ID's).

## QUESTÃO 01: No dataset existem alguns valores faltantes. Antes de começar a manipular os dados, trate essas informações e descreva sucintamente as alterações feitas



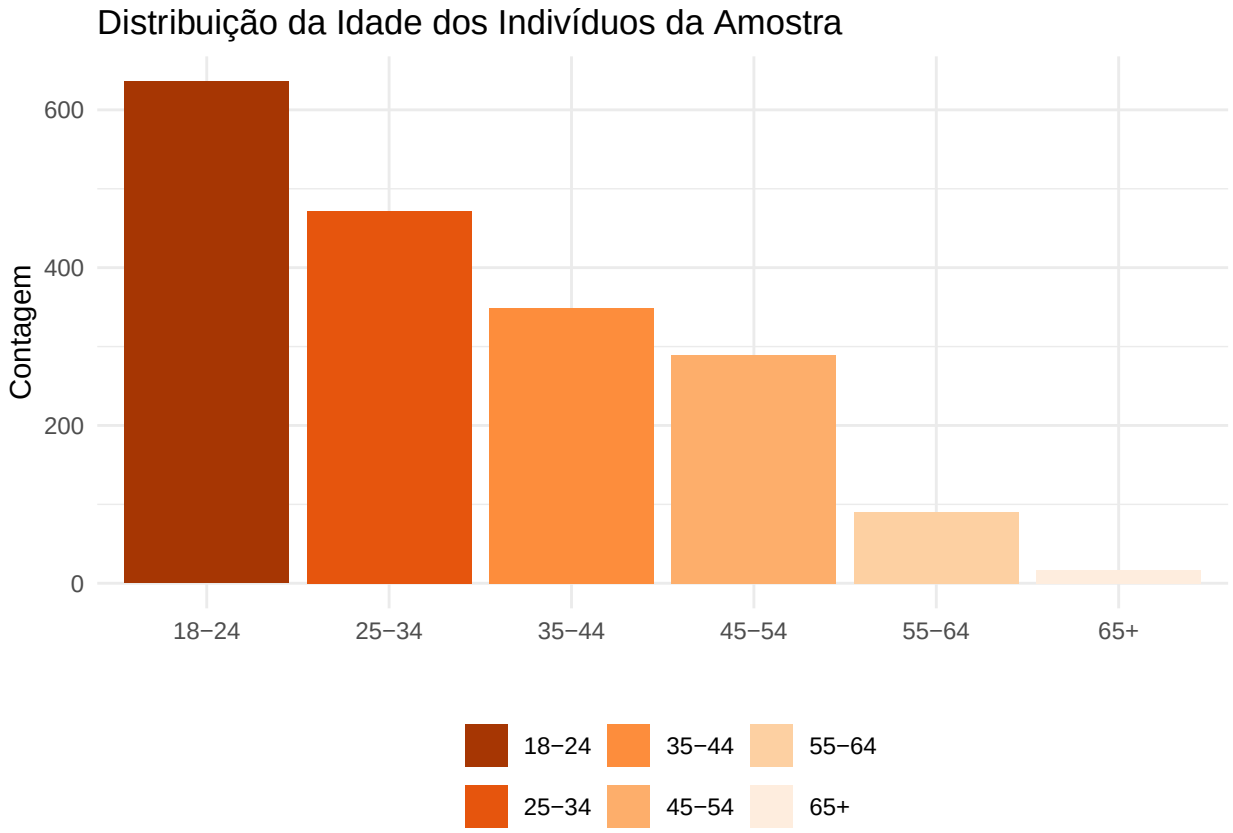
Como pode ser verificado no gráfico anterior, há aproximadamente 3% de missings no dataset.Com destaque para a variável **income**. Assim, a seguir é calculado e apresentado a porcentagem de NA's em cada variável:

Table 1: Porcentagem de NA's nas variáveis do Dataset

| Variáveis do Dataset | Porcentagem |
|---|---|
| income_usd | 98.7791932 |
| ss | 0.3715499 |
| cscore | 0.2653928 |
| gender | 0.2123142 |
| nscore | 0.2123142 |
| a_score | 0.2123142 |
| caff | 0.2123142 |
| education | 0.1592357 |
| escore | 0.1592357 |
| oscore | 0.1592357 |
| impulsive | 0.1592357 |
| alcohol | 0.1592357 |
| ethnicity | 0.1061571 |
| amphet | 0.1061571 |
| amyl | 0.1061571 |
| benzos | 0.1061571 |
| choc | 0.1061571 |
| country | 0.0530786 |
| cannabis | 0.0530786 |
| coke | 0.0530786 |
| crack | 0.0530786 |
| id | 0.0000000 |
| age | 0.0000000 |
| ecstasy | 0.0000000 |
| heroin | 0.0000000 |
| ketamine | 0.0000000 |
| legalh | 0.0000000 |
| lsd | 0.0000000 |
| meth | 0.0000000 |
| mushrooms | 0.0000000 |
| nicotine | 0.0000000 |
| semer | 0.0000000 |
| vsa | 0.0000000 |

A variável renda (income) tem os incríveis 98,78% de observações como NA's. O restante de variáveis não ultrapassa 0.40% quando o assunto é a presença de NA's. **Por simplificidade e tempo de análise, assume-se a partir de agora a remoção da variável renda e das linhas que tenha NA, isto é, do ID que tenha em alguma variável dados faltantes. Assim, restando 1853 linhas, ou seja, 1853 id's diferentes na base de dados para análise.**

## QUESTÃO 02: Qual é a distribuição da idade dos indivíduos na amostra? Existem diferenças significativas nas faixas etárias predominantes de consumo entre os grupos de usuários de diferentes substâncias?

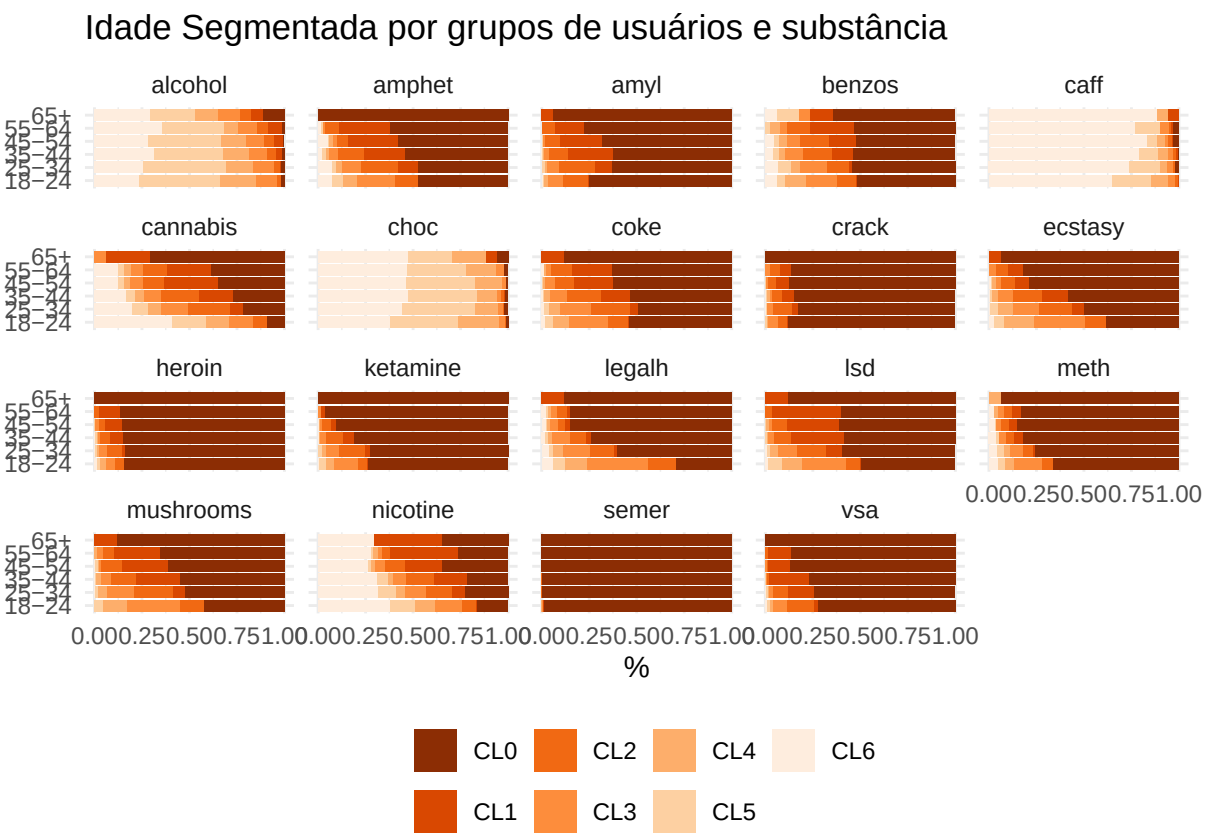



**Na tabela a seguir pode-se visualizar a frequência absoluta e relativa para a idade dos indivíduos da amostra:**

Table 2: Estatísticas Descritivas da variável Idade

| Idade | Frequência Absoluta | Frequência Relativa |
|---|---|---|
| 18-24 | 636 | 34.32 |
| 25-34 | 472 | 25.47 |
| 35-44 | 349 | 18.83 |
| 45-54 | 289 | 15.60 |
| 55-64 | 90 | 4.86 |
| 65+ | 17 | 0.92 |

Como pode ser observado, a população jovem, entre 18 a 34, ocupa a maior parcela da amostra; somando-se um total de 59,79% da amostra. Agora vamos segmentar o mesmo cálculo por grupos de usuários de diferentes substâncias.




Há 19 substâncias diferentes com 7 classificações possíveis:

CL0: Nunca Usou

CL1: Usou Mais de Uma Década Atrás

CL2: Usou nos Últimos Dez Anos

CL3: Usou no Último Ano (59 vezes)

CL4: Usou nos Últimos Meses

CL5: Usou na Última Semana

CL6: Usou Hoje

**Aparentemente há uma relação, menor a idade, maior a proporção de uso de substâncias, isto é, menor a chancer para a classificação "nunca usou".**

Pois bem, buscando visualizar a relação entre idade e consumo de substâncias, pode-se modelar uma correlação de Spearman ($\rho$) e Kendall ($\tau$). Na qual a idade é ordenada da menor para a maior e, por sua vez, o uso de substâncias de CL0 a CL6 (de nunca usou para uso no dia da pesquisa). As correlações podem ser calculadas como segue, respetivamente:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

$$\tau = \frac{2(C-D)}{n(n-1)}$$

Table 3: Relação entre Idade e uso de Substâncias

| | Kendall | P-valor Kendall | ***Kendall | Spearman | P-valor Spearman | ***Spearman |
|---|---|---|---|---|---|---|
| alcohol | 0.0155 | 0.4217 | Insignificante | 0.0182 | 0.4327 | Insignificante |
| amphet | -0.1822 | 0.0000 | Significativo | -0.2130 | 0.0000 | Significativo |
| amyl | -0.0132 | 0.5108 | Insignificante | -0.0100 | 0.6674 | Insignificante |
| benzos | -0.0764 | 0.0001 | Significativo | -0.0923 | 0.0001 | Significativo |
| caff | 0.1177 | 0.0000 | Significativo | 0.1345 | 0.0000 | Significativo |
| cannabis | -0.3672 | 0.0000 | Significativo | -0.4474 | 0.0000 | Significativo |
| choc | 0.0727 | 0.0002 | Significativo | 0.0851 | 0.0002 | Significativo |
| coke | -0.1454 | 0.0000 | Significativo | -0.1663 | 0.0000 | Significativo |
| crack | -0.0015 | 0.9405 | Insignificante | -0.0007 | 0.9749 | Insignificante |
| ecstasy | -0.3192 | 0.0000 | Significativo | -0.3733 | 0.0000 | Significativo |
| heroin | -0.0359 | 0.0798 | Insignificante | -0.0396 | 0.0884 | Insignificante |
| ketamine | -0.1664 | 0.0000 | Significativo | -0.1859 | 0.0000 | Significativo |
| legalh | -0.3833 | 0.0000 | Significativo | -0.4520 | 0.0000 | Significativo |
| lsd | -0.1969 | 0.0000 | Significativo | -0.2314 | 0.0000 | Significativo |
| meth | -0.1613 | 0.0000 | Significativo | -0.1874 | 0.0000 | Significativo |
| mushrooms | -0.2474 | 0.0000 | Significativo | -0.2902 | 0.0000 | Significativo |
| nicotine | -0.1856 | 0.0000 | Significativo | -0.2265 | 0.0000 | Significativo |
| semer | -0.0484 | 0.0216 | Significativo | -0.0534 | 0.0215 | Significativo |
| vsa | -0.1447 | 0.0000 | Significativo | -0.1619 | 0.0000 | Significativo |

**Conclusão:** Nota-se que nas correlações significativas, há uma negativa correlação fraca entre idade e consumo de substância, ou seja, o aumento da idade está correlacionada negativamente com o uso de substâncias. Com exceção de caff e choc, café e chocolate, respectivamente. **Observe que a resposta da questão 12 já está respondida: correlação negativa**.

## QUESTÃO 03: Há uma relação entre o nível educacional e o consumo de substâncias?

Para analisar a relação entre as variáveis, primeiro, vamos ordenar o nível de escolaridade em ordem crescente, isto é, da menor escolaridade (Left school before 16 years) a maior (Doctorate degree). Em segundo, vamos transformar em fator (ordinal) o uso de substâncias: iniciando em zero (CLO) até CL6 – isto é, nunca usou qualquer substância até o uso no dia da pesquisa. A seguir pode-ser osbervar a ordenação do nível de escolaridade:

Table 4: Nível Educacional - do menor para o maior

| Nível |
|---|
| Left school before 16 years |
| Left school at 16 years |
| Left school at 17 years |
| Left school at 18 years |
| Some college or university, no certificate or degree |
| Professional certificate/ diploma |
| University degree |
| Masters degree |
| Master degree |
| Doctorate degree |

**Para verificar a relação entre as variáveis pode-se aplicar a correlação de Spearman ($\rho$) e a correlação de Kendall ($\tau$) como segue, respectivamente**:

Table 5: Relação entre Educação e uso de Substâncias

| | Kendall | P-valor Kendall | ***Kendall | Spearman | P-valor Spearman | ***Spearman |
|---|---|---|---|---|---|---|
| alcohol | 0.1197 | 0.0000 | Significativo | 0.1480 | 0.0000 | Significativo |
| amphet | -0.1406 | 0.0000 | Significativo | -0.1758 | 0.0000 | Significativo |
| amyl | 0.0229 | 0.2419 | Insignificante | 0.0272 | 0.2413 | Insignificante |
| benzos | -0.1112 | 0.0000 | Significativo | -0.1376 | 0.0000 | Significativo |
| caff | 0.0417 | 0.0348 | Significativo | 0.0490 | 0.0348 | Significativo |
| cannabis | -0.2309 | 0.0000 | Significativo | -0.2943 | 0.0000 | Significativo |
| choc | 0.0600 | 0.0019 | Significativo | 0.0722 | 0.0019 | Significativo |
| coke | -0.0971 | 0.0000 | Significativo | -0.1192 | 0.0000 | Significativo |
| crack | -0.1212 | 0.0000 | Significativo | -0.1396 | 0.0000 | Significativo |
| ecstasy | -0.1555 | 0.0000 | Significativo | -0.1916 | 0.0000 | Significativo |
| heroin | -0.1200 | 0.0000 | Significativo | -0.1394 | 0.0000 | Significativo |
| ketamine | -0.0719 | 0.0003 | Significativo | -0.0845 | 0.0003 | Significativo |
| legalh | -0.2011 | 0.0000 | Significativo | -0.2443 | 0.0000 | Significativo |
| lsd | -0.1591 | 0.0000 | Significativo | -0.1939 | 0.0000 | Significativo |
| meth | -0.1610 | 0.0000 | Significativo | -0.1905 | 0.0000 | Significativo |
| mushrooms | -0.1546 | 0.0000 | Significativo | -0.1896 | 0.0000 | Significativo |
| nicotine | -0.2054 | 0.0000 | Significativo | -0.2604 | 0.0000 | Significativo |
| semer | -0.0416 | 0.0433 | Significativo | -0.0470 | 0.0432 | Significativo |
| vsa | -0.1057 | 0.0000 | Significativo | -0.1245 | 0.0000 | Significativo |

**Conclusão:** Como pode ser observado na tabela 4 considerando p-valor 0.05, as correlações significativas no geral são fracas e negativas, isto é, com aumento do nível de escolaridade, menor o uso de substância

(**mais próximo de nunca usou: CL0**), com exceção das substâncias caff e choc (cafeína e chocolate, respetivamente) que têm correlação positiva.

**Observe que a questão 10 já foi respondida aqui**

# QUESTÃO 04: Como o gênero influencia no consumo de drogas alucinógenas (LSD, Ecstasy, Ketamine, Cannabis e Mushrooms)? Explique.

Para esse problema podemos modelar um modelo multinominal, isto é, um modelo logit (binário) expandido para inferir sobre a existência de influência do gênero no consumo de drogas alucionógenas. Assim, pode-se tomar as ordens de consumo(CLO...CL6) como variável depedente e, por sua vez, a idade como variável explicativa. Porém, observe que a variável dependente no presente caso não é binária, então, precisamos estender o modelo logit para estimar os parâmetros tornando-se um modelo multinominal. O qual assume a forma de logito para a k-ésima categoria em relação a uma categoria de referência (**no presente caso, CL0, isto é, a categoria "nunca usou"**), expresso como:

$$\log\left(\frac{P(Y=k)}{P(Y=1)}\right) = \beta_{k0} + \beta_{k1}x_1 + \ldots + \beta_{kp}x_p$$

Assim, as probabilidades para todas as k categorias são obtidas através da função softmax, que normaliza as probabilidades:

$$P(Y=k) = \frac{\exp(\beta_{k0} + \beta_{k1}x_1 + \ldots + \beta_{kp}x_p)}{\sum_{l=1}^{K}\exp(\beta_{l0} + \beta_{l1}x_1 + \ldots + \beta_{lp}x_p)}$$

**Estimando o modelo temos:**

Table 6: Coeficientes do Modelo Multinominal

| | (Intercept) | genderM |
|---|---|---|
| CL1 | 0.1410304 | 1.447603 |
| CL2 | 0.1629182 | 1.845671 |
| CL3 | 0.1229562 | 3.030887 |
| CL4 | 0.0549895 | 3.527430 |
| CL5 | 0.0368921 | 3.611629 |
| CL6 | 0.0556780 | 3.080820 |

## Interpretação de cada coeficiente:

**CL1 - Usou Mais de Uma Década Atrás:**

Intercepto: A chance de uma mulher estar na categoria CL1 em vez de CL0 é aproximadamente 0.141.

Gênero Masculino: A chance de um homem estar na categoria CL1 em vez de CL0 é 1.448 vezes a chance de uma mulher.

**CL2 - Usou nos Últimos Dez Anos:**

Intercepto: A chance de uma mulher estar na categoria CL2 em vez de CL0 é aproximadamente 0.163.

Gênero Masculino: A chance de um homem estar na categoria CL2 em vez de CL0 é 1.846 vezes a chance de uma mulher.

**CL3 - Usou no Último Ano:**

Intercepto: A chance de uma mulher estar na categoria CL3 em vez de CL0 é aproximadamente 0.123.

Gênero Masculino: A chance de um homem estar na categoria CL3 em vez de CL0 é 3.031 vezes a chance de uma mulher.

**CL4 - Usou nos Últimos Meses:**

Intercepto: A chance de uma mulher estar na categoria CL4 em vez de CL0 é aproximadamente 0.055.

Gênero Masculino: A chance de um homem estar na categoria CL4 em vez de CL0 é 3.527 vezes a chance de uma mulher.

**CL5 - Usou na Última Semana:**

Intercepto: A chance de uma mulher estar na categoria CL5 em vez de CL0 é aproximadamente 0.037.

Gênero Masculino: A chance de um homem estar na categoria CL5 em vez de CL0 é 3.612 vezes a chance de uma mulher.

**CL6 - Usou Hoje:**

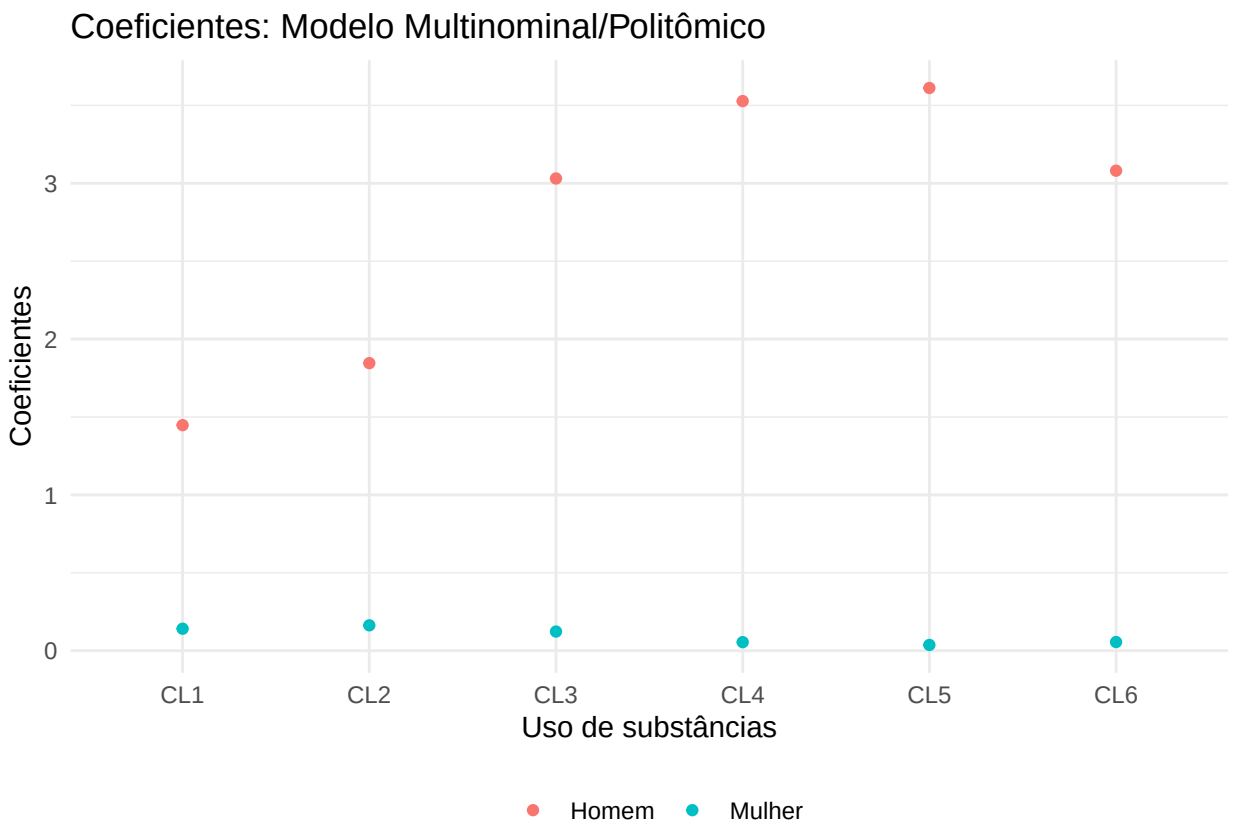
Intercepto: A chance de uma mulher estar na categoria CL6 em vez de CL0 é aproximadamente 0.056.

Gênero Masculino: A chance de um homem estar na categoria CL6 em vez de CL0 é 3.081 vezes a chance de uma mulher.

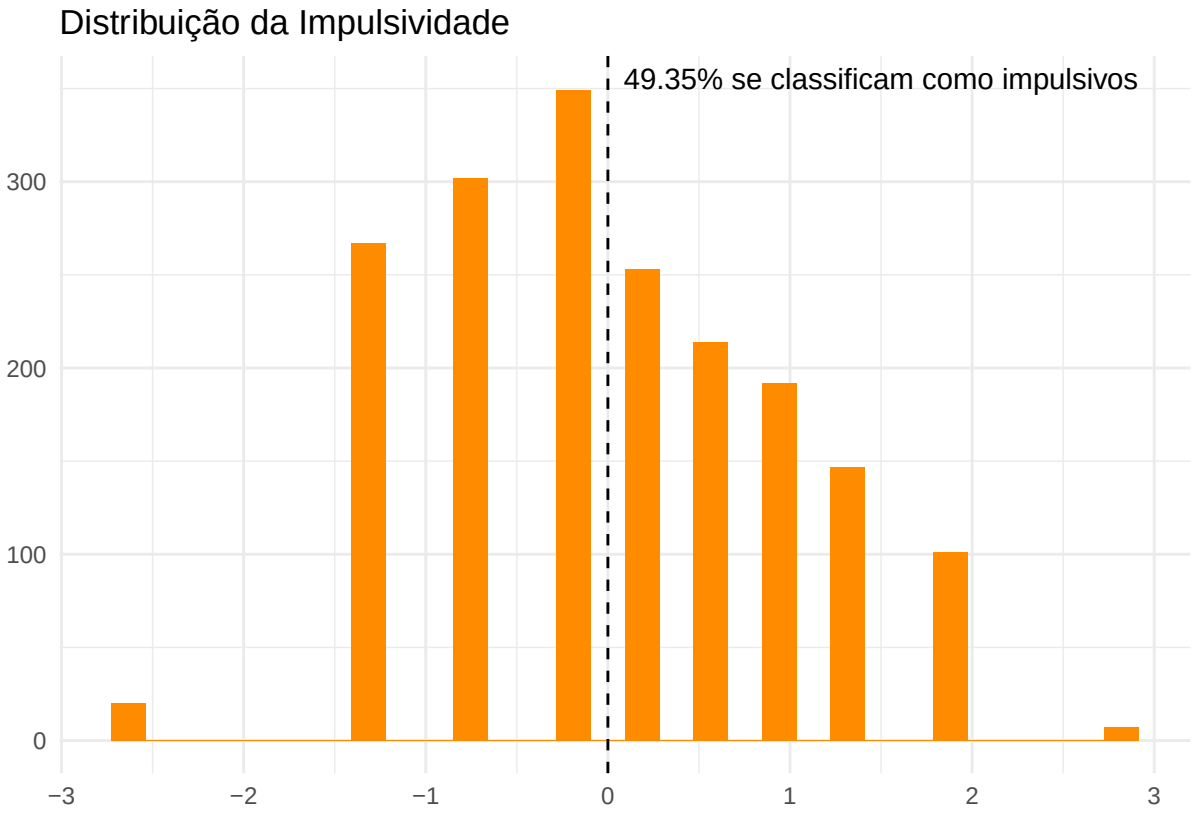
## Conclusão geral:

**Nota-se que os homens são mais vuneráveis na possibilidade de consumir alguma substância. A chance de um homem consumir "hoje" alguma droga é 3.081 vezes a chance de uma mulher (enquanto de uma mulher consumir "hoje" é de apenas 0.056).**

**Os coeficientes podem ser resumidos no seguinte plote para uma melhor visualização:**



## QUESTÃO 05: Qual é a proporção de participantes que se auto-classificam como impulsivos (score superior a zero)? Existe uma correlação entre a impulsividade e o consumo de substâncias?




Sabendo que quase 50% dos entrevistados consideram-se impulsivos, vamos agora calcular a correlação de Spearman (considerando não normalidade entre a correlação da impulsividade com as substâncias - variável ordinal de 1 a 7). A correlação de Spearman pode ser calculada como segue:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Table 7: Correlação de Spearman

| Substâncias | Correlação |
|---|---|
| alcohol | 0.0407541 |
| amphet | 0.2944064 |
| amyl | 0.1313166 |
| benzos | 0.2325970 |
| caff | 0.0160944 |
| cannabis | 0.3151627 |
| choc | -0.0214221 |
| coke | 0.2626520 |
| crack | 0.1937967 |
| ecstasy | 0.2704343 |
| heroin | 0.1978586 |

| Substâncias | Correlação |
|---|---|
| ketamine | 0.1859128 |
| legalh | 0.2849797 |
| lsd | 0.2541038 |
| meth | 0.1991995 |
| mushrooms | 0.2825223 |
| nicotine | 0.2547454 |
| semer | 0.0360670 |
| vsa | 0.1956084 |

**Conclusão: como pode ser observado, todas as variáveis têm correlação fraca e positiva com a impulsividade, com exceção da variável choc (chocolate). Com destaque para a maconha (cannabis) com maior correlação.**

## QUESTÃO 06: Classifique as variáveis entre qualitativas (ordinal ou nominal), ou quantitativas (discreta, contínuas).

**A classificação das variáveis pode ser visualizado na tabela a seguir:**

Table 8: Estrutura dos dados

| var | qualitativa | classe |
|---|---|---|
| id | sim | nominal |
| age | sim | ordinal |
| gender | sim | nominal |
| education | sim | ordinal |
| country | sim | nominal |
| ethnicity | sim | nominal |
| income_usd | não | contínua |
| nscore | não | contínua |
| escore | não | contínua |
| oscore | não | contínua |
| a_score | não | contínua |
| cscore | não | contínua |
| impulsive | não | contínua |
| ss | não | contínua |
| alcohol | sim | ordinal |
| amphet | sim | ordinal |
| amyl | sim | ordinal |
| benzos | sim | ordinal |
| caff | sim | ordinal |
| cannabis | sim | ordinal |
| choc | sim | ordinal |
| coke | sim | ordinal |
| crack | sim | ordinal |
| ecstasy | sim | ordinal |
| heroin | sim | ordinal |
| ketamine | sim | ordinal |
| legalh | sim | ordinal |
| lsd | sim | ordinal |
| meth | sim | ordinal |
| mushrooms | sim | ordinal |
| nicotine | sim | ordinal |
| semer | sim | ordinal |
| vsa | sim | ordinal |

**QUESTÃO 07: Qual é a proporção de consumo de substâncias legais versus ilícitas na amostra (considere a definição de legalidade segundo a legislação brasileira)?**

| droga | status |
|---|---|
| alcohol | lícita |
| amphet | ilícita |
| amyl | ilícita |
| benzos | ilícita |
| caff | lícita |
| cannabis | ilícita |
| choc | lícita |
| coke | ilícita |
| crack | ilícita |
| ecstasy | ilícita |
| heroin | ilícita |
| ketamine | ilícita |
| legalh | ilícita |
| lsd | ilícita |
| meth | ilícita |
| mushrooms | ilícita |
| nicotine | lícita |
| semer | ilícita |
| vsa | ilícita |

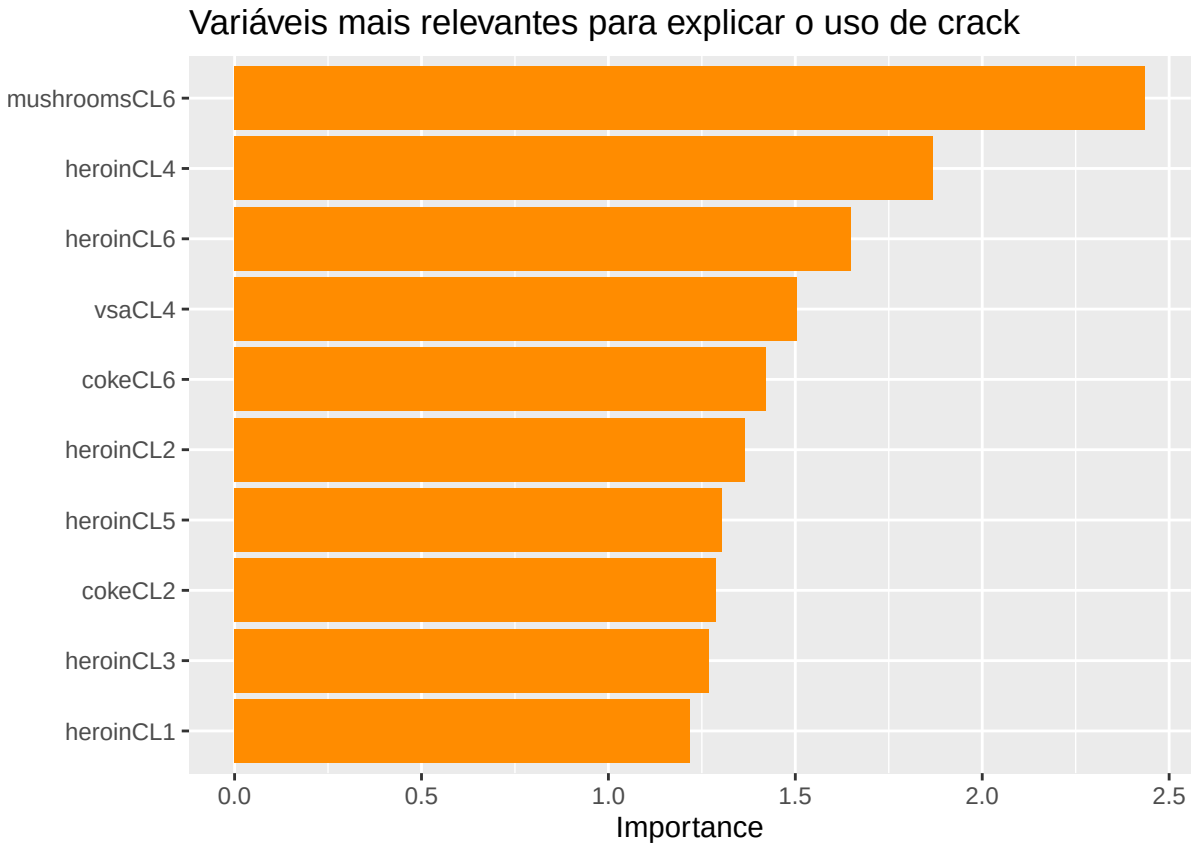
## QUESTÃO 08:Quais fatores predizem a probabilidade de um indivíduo consumir crack (Crack)?

Para verificar quais variáveis são mais relevantes para predizer o consumo de crack, pode-se modelar uma regressão via LASSO. Na qual temos a seguinte função de custo com parâmetro $\lambda$ de regularização:

$$\min_{\beta_0,\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Observe que com $\lambda$ igual a zero temos a minimazação clássica de mínimos quadrados. Porém, à medida que $\lambda$ aumenta, a penalização faz com que os coeficientes sejam encolhidos em direção a zero, ou seja, removendo alguns coeficientes que são exatamente iguais a zero (menos relevantes).

Assim, no presente caso, pode-se modelar uma regressão na qual a variável dependente é o uso ou não de crack (0 = caso nunca usou, 1 = caso tenha usado alguma vez na vida) contra todas as variáveis do dataset. Por fim, via LASSO, filtra-se as 10 mais relevantes para explicar o uso de crack:




**Conclusão: De acordo com o método adotado, o uso de cogumelos mágicos "hoje", heroína no "último mês" e "hoje", respectivamente, são as três variáveis mais relevantes para explicar o uso de crack. O restante das variáveis relevantes são: VSA (classe de consumo de abuso de substâncias voláteis), uso de heroína e cocaína. Drogas que, de fato, costumeiramente estão correlacionadas com o uso de drogas mais pesadas como o crack.**

## QUESTÃO 09: Qual é a média das pontuações Nscore, Escore, Oscore, AScore, Cscore? Calcule a correlação entre elas.

A segue pode-se visualizar a média e o teste de Shapiro-Wilk para verificar a normalidade das respectivas variáveis, pois o teste de correlação de Pearson pressupõe distribuição Gaussiana. Nota-se que escore, oscore, a_score não tem distribuição Gaussiana, logo, é necessário tomar os resultados com cautela.

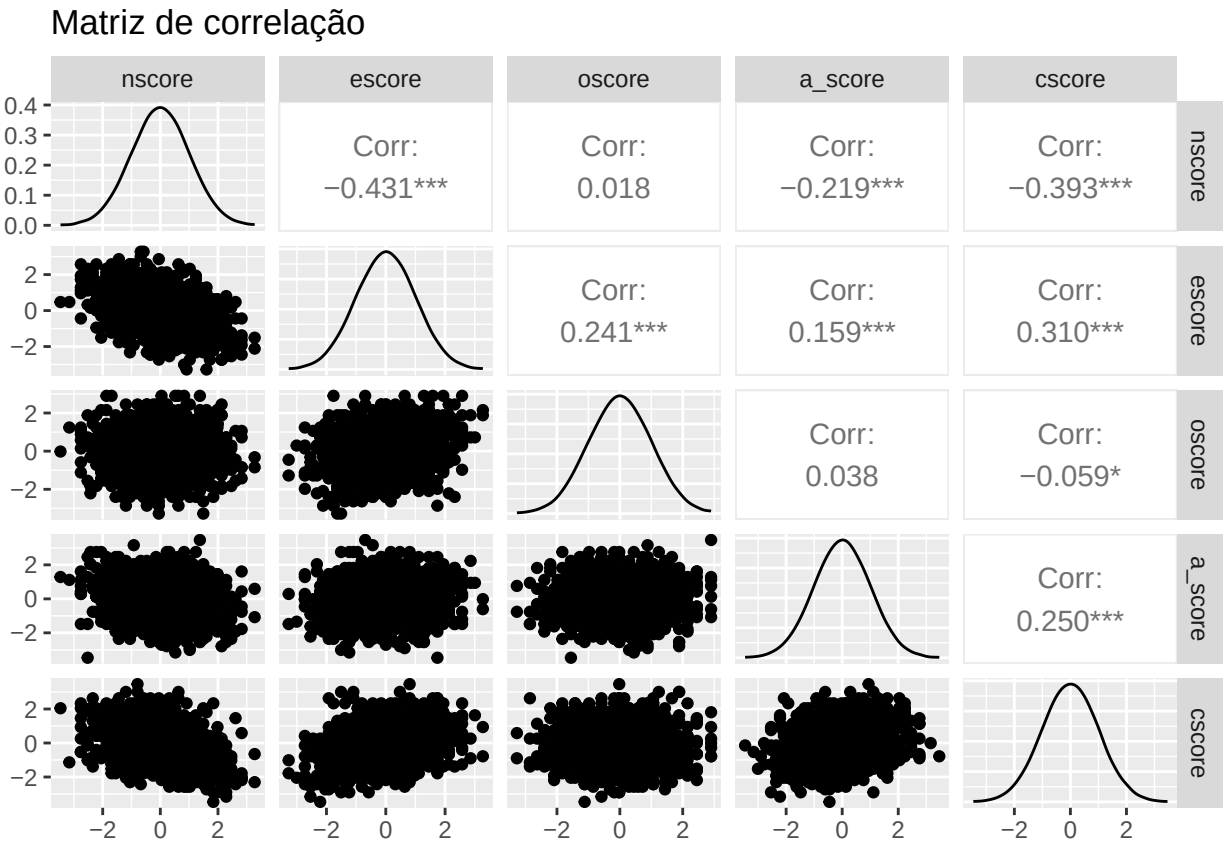
A correlação pode ser calculada como segue:

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}}$$

```
## [1] "**********************************"
## Teste de Shapiro-Wilk para a variável:  nscore
## Variável com média:  0.001072148
##
##  Shapiro-Wilk normality test
##
## data:  scores[[i]]
## W = 0.99869, p-value = 0.1739
##
##
## Variável com distribuição normal
## [1] "**********************************"
## Teste de Shapiro-Wilk para a variável:  escore
## Variável com média:  0.0040685
##
##  Shapiro-Wilk normality test
##
## data:  scores[[i]]
## W = 0.99785, p-value = 0.01396
##
##
## Variável não tem normalidade
## [1] "**********************************"
## Teste de Shapiro-Wilk para a variável:  oscore
## Variável com média:  0.004893411
##
##  Shapiro-Wilk normality test
##
## data:  scores[[i]]
## W = 0.99735, p-value = 0.003206
##
##
## Variável não tem normalidade
## [1] "**********************************"
## Teste de Shapiro-Wilk para a variável:  a_score
## Variável com média:  -0.0003985969
##
##  Shapiro-Wilk normality test
##
## data:  scores[[i]]
```

```
## W = 0.99779, p-value = 0.01178
##
##
## Variável não tem normalidade
```

**Assim, pode-se visualizar a correlação de Pearson:**



**QUESTÃO 10: Analise a relação entre o nível de educação (Education) e o consumo de diferentes substâncias ilícitas (como LSD, Amphet, Cannabis, etc.). Identifique se há uma correlação significativa entre essas variáveis e, em caso afirmativo, explore a natureza dessa correlação (positiva/negativa).**

**Questão respondida na questão de número 03**

## QUESTÃO 11: Treine uma árvore de decisão para prever se um indivíduo consome uma determinada substância (por exemplo, álcool, anfetaminas, cannabis) com base em suas características demográficas e pontuações de personalidade. Utilize a acurácia para avaliar os seus resultados.

Vamos modelar com a substância álcool, mas poderia ser com qualquer outra, pois as funções ficaram flexíveis e automatizadas. A seguir desenvolve-se as funções (considerando 75% para treino da árvore) considerando uma modelagem binária (já usou a substância alguma vez na vida ou nunca usou a substância). A tabela de estatísticas e acurácia pode ser verificada ao final:

*já usou = 1*

*nunca usou = 0*

```
get_pre_processing <- function(db, transformation, pct_train, var_target){

  base::set.seed(123)

  # Divisao entre base de Treino e Teste

  data_split <- rsample::initial_split(data = db,
                                       prop = pct_train)

  bases <- list(base_train = rsample::training(data_split),
                base_test  = rsample::testing(data_split))

# Pré processando Treino e Teste
# Loop para aplicar as transformações em treino e teste

  for(i in seq_along(bases)) {
    bases[[i]] <- bases[[i]] %>%
                  dplyr::select(-dplyr::all_of(var_target)) %>%
                  # Imputação pré-definida: mediana *\\
                  # Na presença de dados faltantes imputa de forma automática
                  dplyr::mutate_if(is.numeric,
                                   zoo::na.aggregate,
                                   FUN = median,
                                   na.rm = TRUE) %>%
                   dplyr::mutate_if(is.numeric,
                                    ~dlookr::transform(., method = transformation)) %>%
                  dplyr::bind_cols(bases[[i]][dplyr::all_of(var_target)])

    # Salvando bases pré-processadas
    switch(names(bases)[i],
           base_train = base_train <- base::data.frame(bases[[i]]),
           base_test = base_test <- base::data.frame(bases[[i]]))

  }

  return(bases)

}
```

```r
db_model = data_clean %>%
           dplyr::select(age,
                         gender,
                         education,
                         country,
                         ethnicity,
                         alcohol) %>%
           mutate(age = factor(age),
                  gender = factor(gender),
                  education = factor(education),
                  country = factor(country),
                  ethnicity = factor(ethnicity),
                  alcohol = ifelse(alcohol == "CL0", 0, 1),
                  alcohol = as.factor(alcohol)) %>%
           relocate(alcohol, .after = NULL)
```

```r
# Chamando a função de pré-processamento

data_pre_processing = get_pre_processing(db = db_model,
                                         transformation = "zscore",
                                         pct_train = 0.75,
                                         var_target = "alcohol")
```

```r
# Definindo o controle de treino com validação cruzada de 10 vezes

train_control <- caret::trainControl(method = "cv", number = 10)

# Treinando o modelo

model <- caret::train(alcohol ~ .,
                      data = data_pre_processing[["base_train"]],
                      method = "rpart",
                      trControl = train_control)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0   0   0
##          1   7 457
##
##                Accuracy : 0.9849
##                  95% CI : (0.9692, 0.9939)
##     No Information Rate : 0.9849
##     P-Value [Acc > NIR] : 0.59872
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : 0.02334
##
##             Sensitivity : 0.00000
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
```

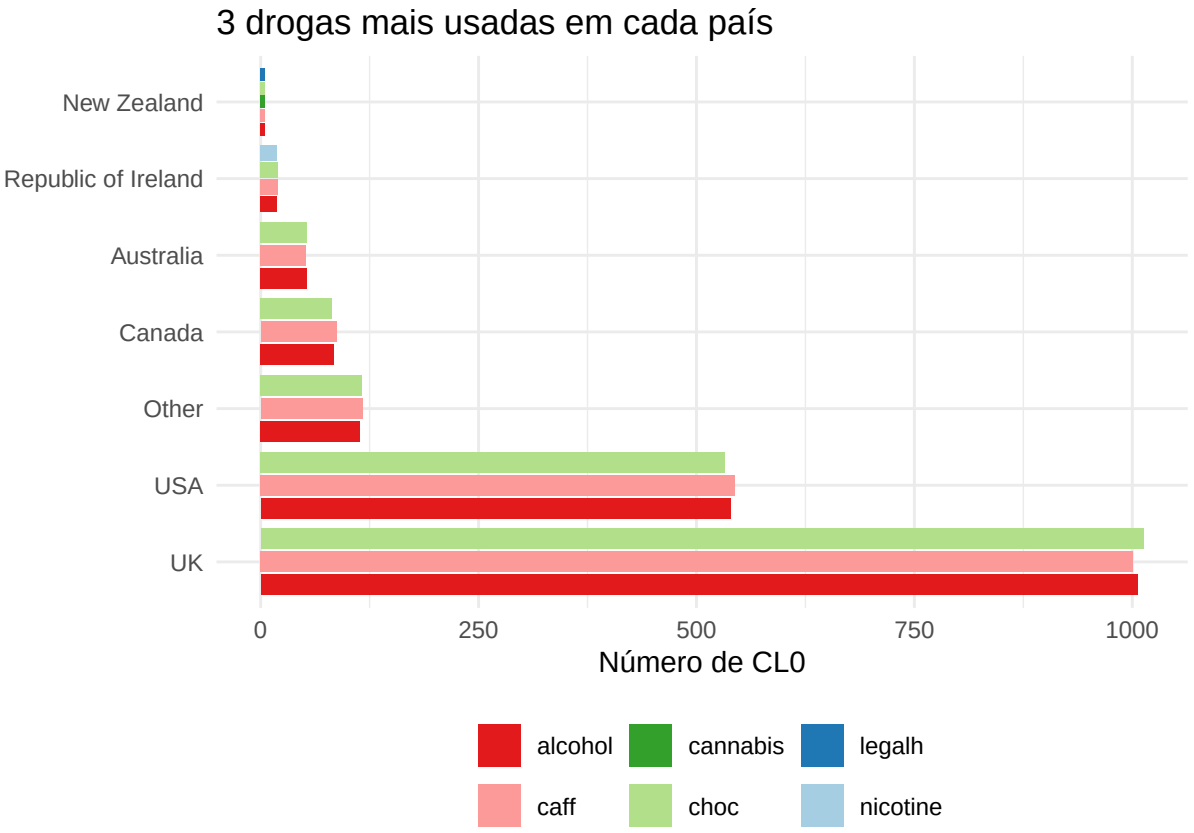
```
##          Neg Pred Value : 0.98491
##              Prevalence : 0.01509
##          Detection Rate : 0.00000
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
##
##        'Positive' Class : 0
##
```

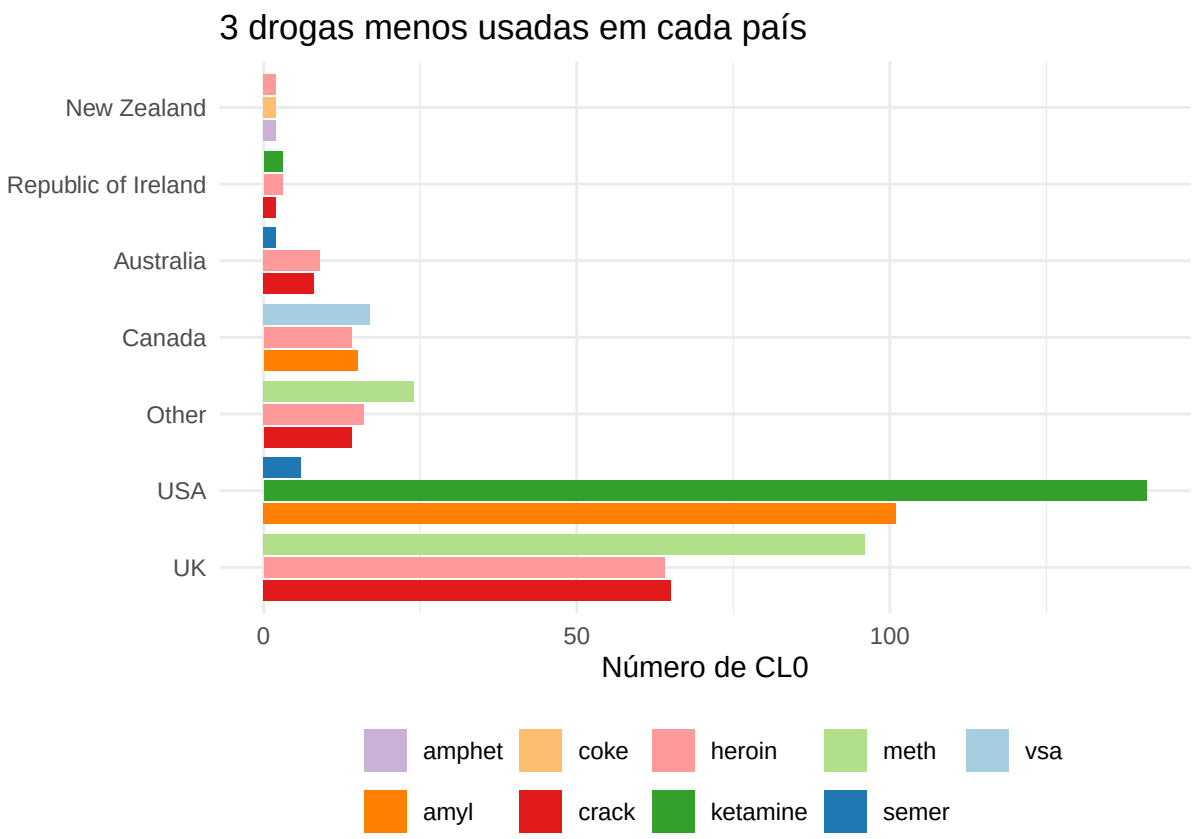
**QUESTÃO 12:Explore a correlação entre a idade (variável Age) e a experimentação de diferentes substâncias ilícitas. Verifique se há uma tendência de aumento ou diminuição do consumo conforme a idade avança.**

**Questão respondida na questão de número 02**

## QUESTÃO 13: Quais são as 3 drogas mais utilizadas para cada país presente na amostra? E quais são as 3 menos utilizadas?

Como queremos encontrar a substância com a maior frequência de uso, pode-se excluir CL0 (isto é, nunca usou). Com o restante podemos fazer uma cotagem, eis a resposta para as 3 substâncias mais consumidas em cada país (como esperado, as substâncias lícitas lideram):




3 drogas menos usadas em cada país
New Zealand
Republic of Ireland
Australia
Canada
Other
USA
UK
0
50
100
Número de CL0
amphet
coke
heroin
meth
vsa
amyl
crack
ketamine
semer