Лабораторна робота 2

Завдання 1: Виявлення проблем.

(A) Представити атрибути (поля) вашого dataset в якому неприйнятні типи;
(B) Представити атрибути з пропущеними даними або показати, що пропущених даних не має в dataset;
(C) Зробити припущення стосовно групування атрибутів.

1. В даному датасеті більш змінних - неперервні, типу double, тому вони нас цілком влаштовують.Проблеми можуть виникнути з Country, Age group and Sex при спробі візуалізації. Тому зробимо їх факторними.

2. Відсутніх значень майже немає(0.05%), тому маємо право просто позбутись від них методом delete. Хоча в такому випадку ми втратимо майже всю інформацію про дві країни Argentina and Barabados. Заповнити нулі середніми значеннями в даному кейсі також не вийде

3. Групування по країнам, по рокам, по віковим групам. Можливо ще варто спробувати вкладене групування, наприклад по рокам та віковим групам

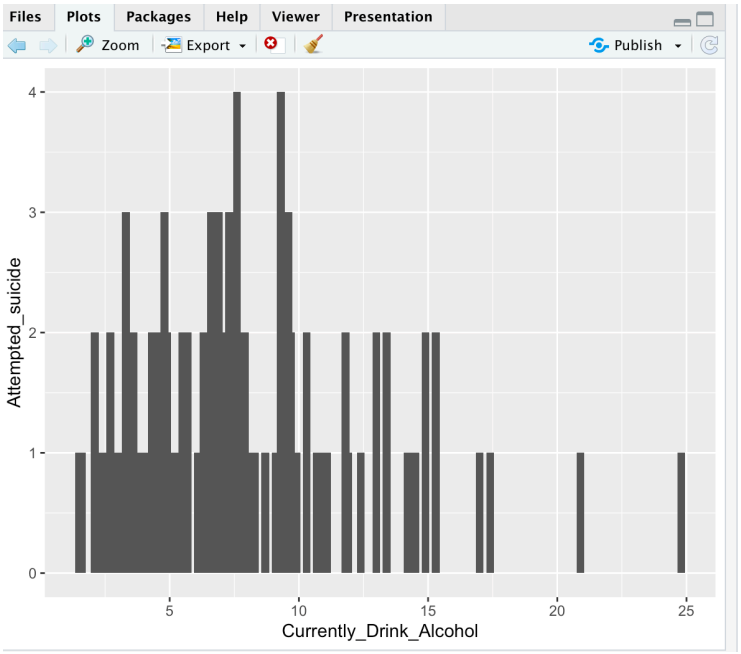
Завдання 2: Усунення проблем.

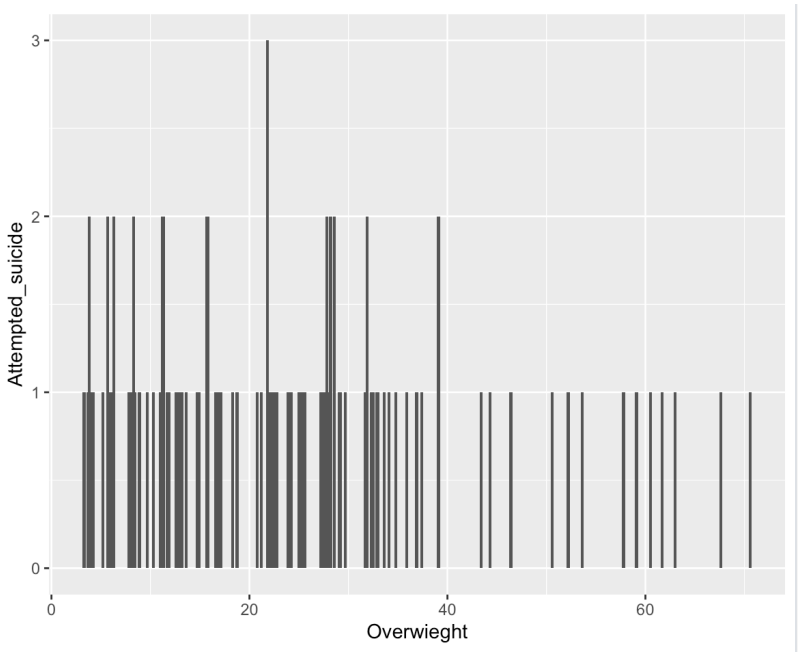
(A) Виконати заміну типу в атрибутах з неприйнятним типом;

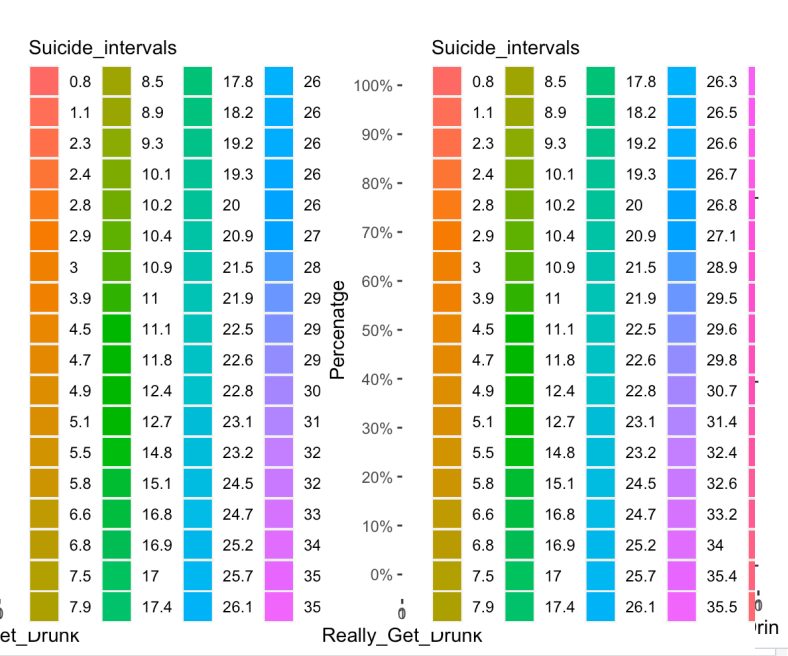
```
# 1. Виконати заміну типу в атрибутах з неприйнятним типом;

suicidal_behaviours$Country <- factor(suicidal_behaviours$Country)
suicidal_behaviours$Sex <- factor(suicidal_behaviours$Sex)
suicidal_behaviours$`Age Group` <- factor(suicidal_behaviours$`Age Group`)

is.factor(suicidal_behaviours$Country)
is.factor(suicidal_behaviours$Sex)
is.factor(suicidal_behaviours$`Age Group`)

str(suicidal_behaviours)
```

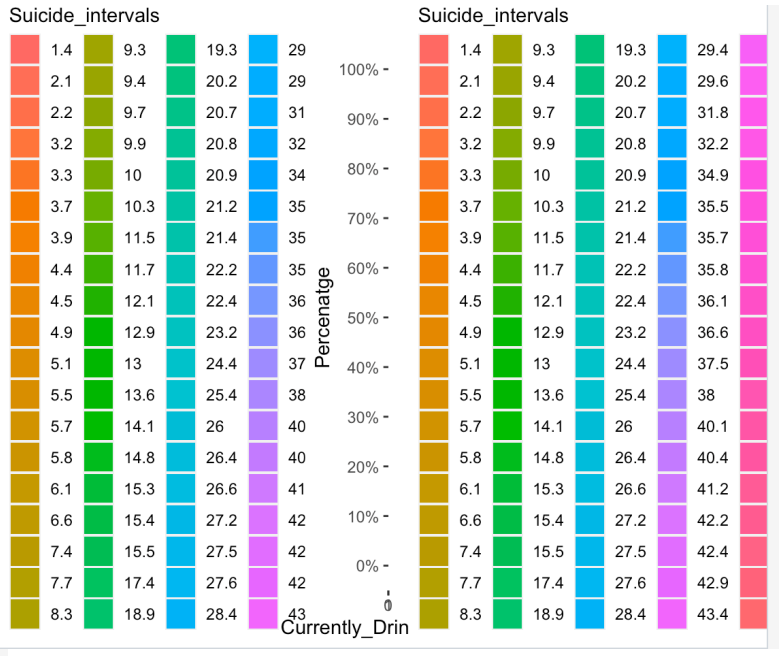
B) Представте графічно зв'язок залежності кожного атрибута від залежної змінної в кількісному та відсотковому значенні;

Справа в тому, що більша кількість незалежних змінних в нас також неперервного типу, тому без додаткових групувань та перетворень складно щось візуалізувати.









(C) Для пропущених значень запропонуйте та виконайте дозаповнення даних (якщо їх багато, то не менше ніж для 3-х);

```
> missing_vars(suicidal_behaviours)
                              var missing missing_prop
1                         Bullied       4   0.03773585
2             Smoke_cig_currently       2   0.01886792
3                         Country       0   0.00000000
4                            Year       0   0.00000000
5                       Age Group       0   0.00000000
6                             Sex       0   0.00000000
7       Currently_Drink_Alcohol       0   0.00000000
8              Really_Get_Drunk       0   0.00000000
9                     Overwieght       0   0.00000000
10                   Use_Marijuana       0   0.00000000
11      Have_Understanding_Parents       0   0.00000000
12 Missed_classes_without_permssion       0   0.00000000
13            Had_sexual_relation       0   0.00000000
14                    Had_fights       0   0.00000000
15          Got_Seriously_injured       0   0.00000000
16               No_close_friends       0   0.00000000
17              Attempted_suicide       0   0.00000000
18              Suicide_intervals       0   0.00000000
>
```

Спробуємо заповнити пропущені значення відповідною медіаною, взятою з усього датасету таким чином:

suicidal_behaviours$Bullied <- ifelse((is.na(suicidal_behaviours$Bullied) == TRUE), 27.55,suicidal_behaviours$Bullied)

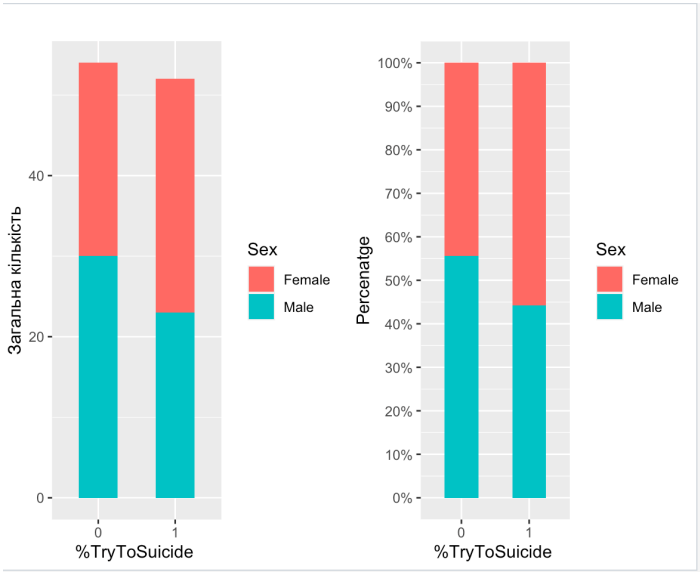
suicidal_behaviours$Smoke_cig_currently <- ifelse((is.na(suicidal_behaviours$Smoke_cig_currently) == TRUE), 12.60,suicidal_behaviours$Smoke_cig_currently)

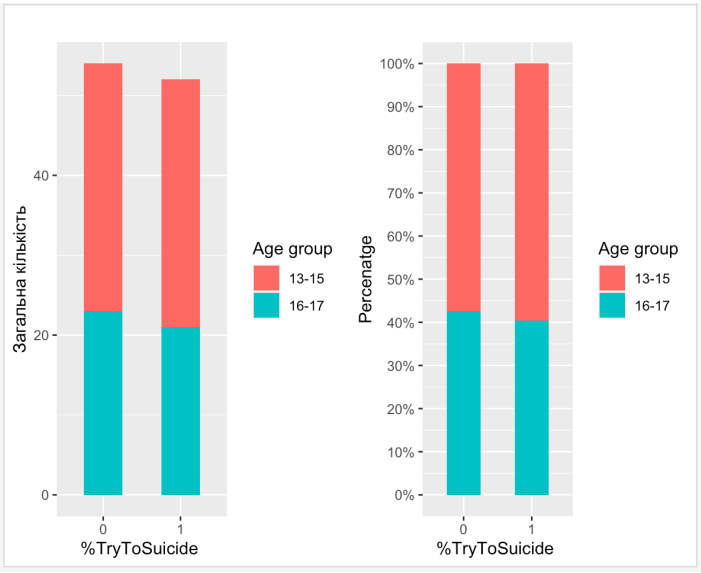
(D) Намалюйте діаграми зв'язку в кількісному та відсотковому значенні для пункту (C) та порівняйте з графіками до перетворення. Опишіть що змінилось.

Так як пропущений значень в нас майже не було(приблизно 0,05%), було прийнято рішення заповнити їх медіаною. Від таких маніпуляцій графіки істотна не змінять свою форму, тому вирішили не візуалізовувати дані ще раз.

(E) Виконайте для даних які потребують групування або перетворення відповідні заміни (якщо їх багато, то не менше ніж для 3-х). Представте діаграми зв'язку.

Групуємо по Sex, Age group та рахуємо середня по залежній змінній, дивимось на результат.
Також так як залежна змінна неперервна, створили додаткову бінарну змінну (0 - < медіани, 1 - > медіани)


Загальна кількість
40
20
0
0
1
%TryToSuicide
Sex
Female
Male
Percenatge
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0
1
%TryToSuicide
Sex
Female
Male


Загальна кількість
40
20
0
0
1
%TryToSuicide
Age group
13-15
16-17
Percenatge
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0
1
%TryToSuicide
Age group
13-15
16-17