***Sitchenko Anastasiya, Marchenko Marina, IT2-2116***

**Mental Health**

1. Какая тенденция развития каждого психического расстройства в будущем? (снижение/рост)

2. Какое заболевание наиболее распространено?

3. Когда были самые высокие показатели? (с какими событиями в мире/странах может быть связано?)

4. Когда были самые низкие показатели? (с какими событиями в мире/странах может быть связано?)

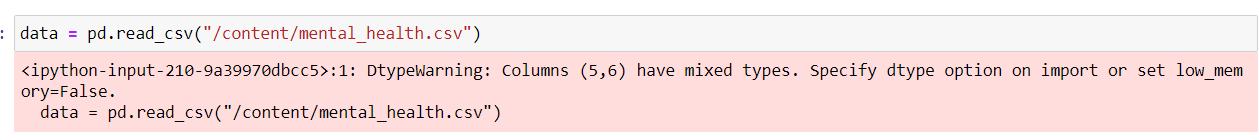
**Dataset collection.**

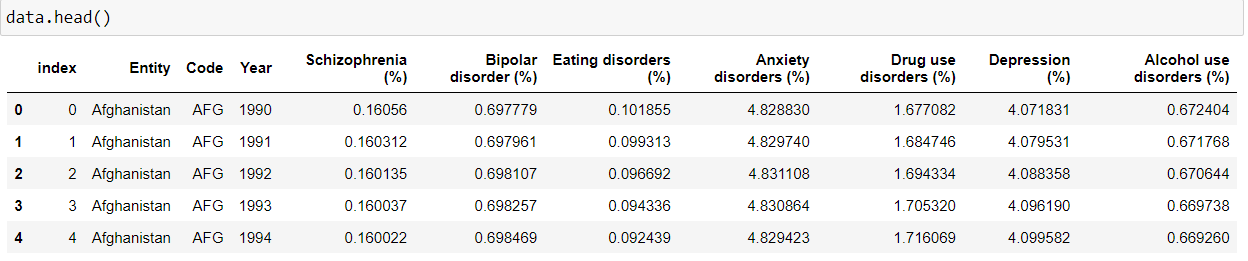
Мы нашли и скачали датасет с сайта Kaggle.com на тему ментального здоровья в мире.

**First look at data.**

Для начала, просматриваем как выглядит наша база данных (первые и последние 5 строчек, название колон, типы данных, количество NAN и т.д.):





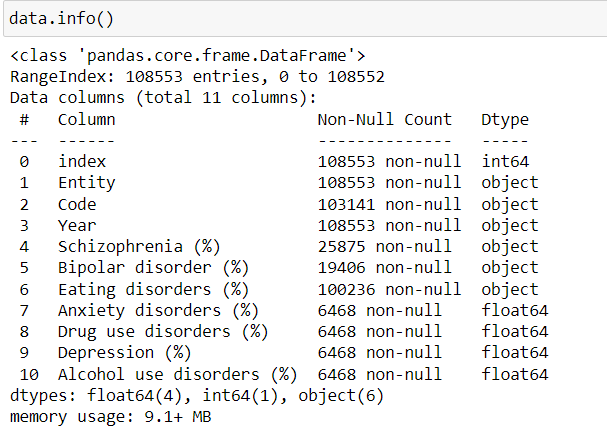


Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как текст

Автоматически созданное описание

Изображение выглядит как текст

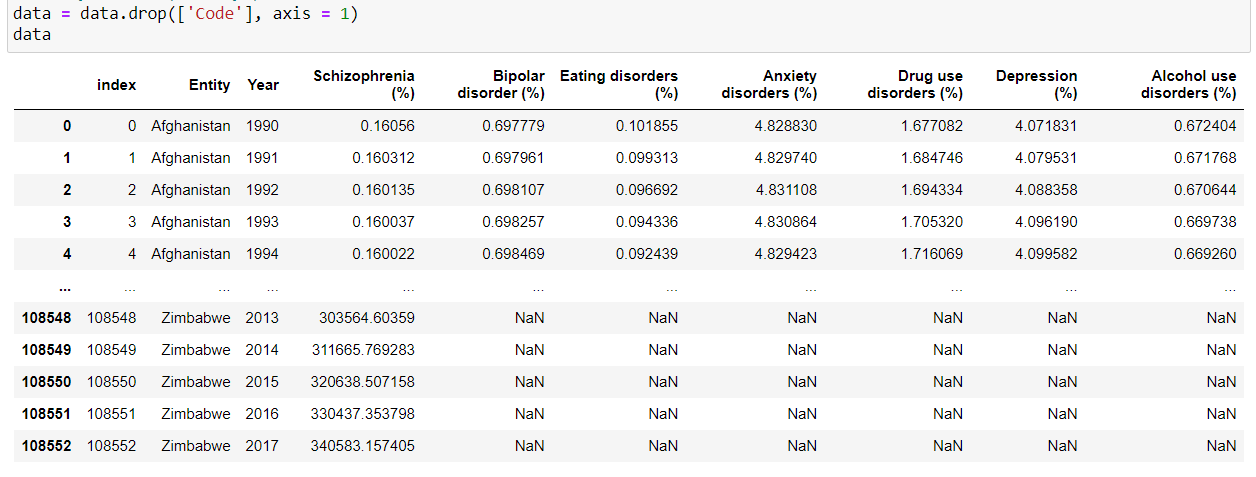
Автоматически созданное описание

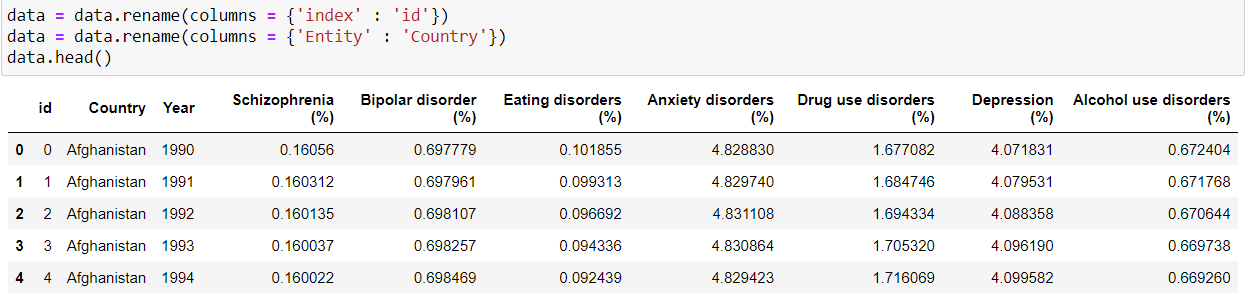
Изображение выглядит как текст

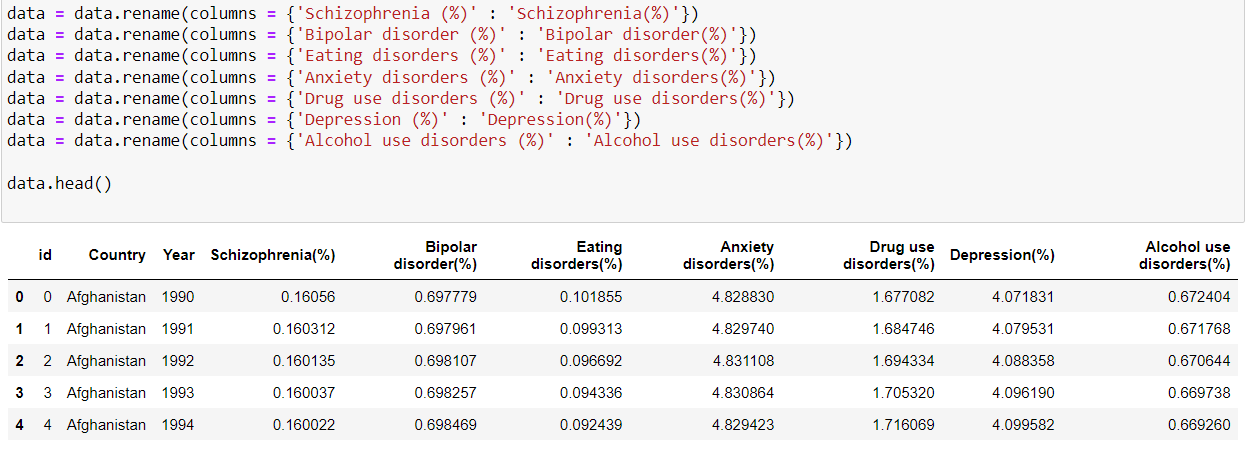
Автоматически созданное описание

**Data Cleaning**

Удаляем строку CODE, тк имеем колону ENTITY. Переименовываем колоны INDEX в ID, а ENTITY в COUNTRY. Убираем пробелы в названии колон для удобства.







Выяснилось, что скачанный датасет изначально заполнен некорректно, поэтому сейчас мы будем искать и исправлять данные, которые должны находиться в отдельных колонках.

Это индексы тех рядов, где начинается заполнение новых данных:

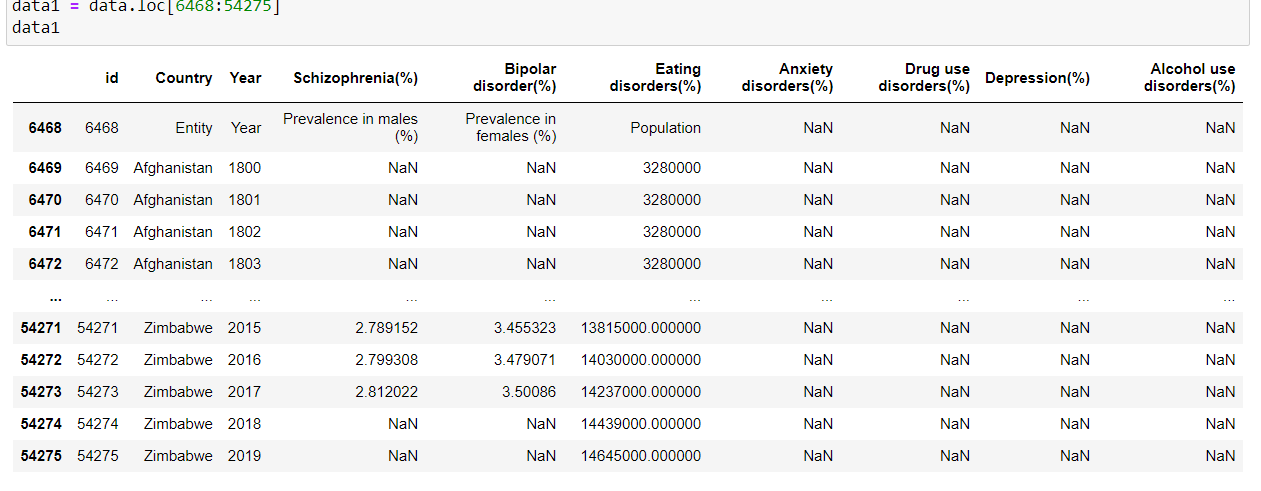
Изображение выглядит как текст

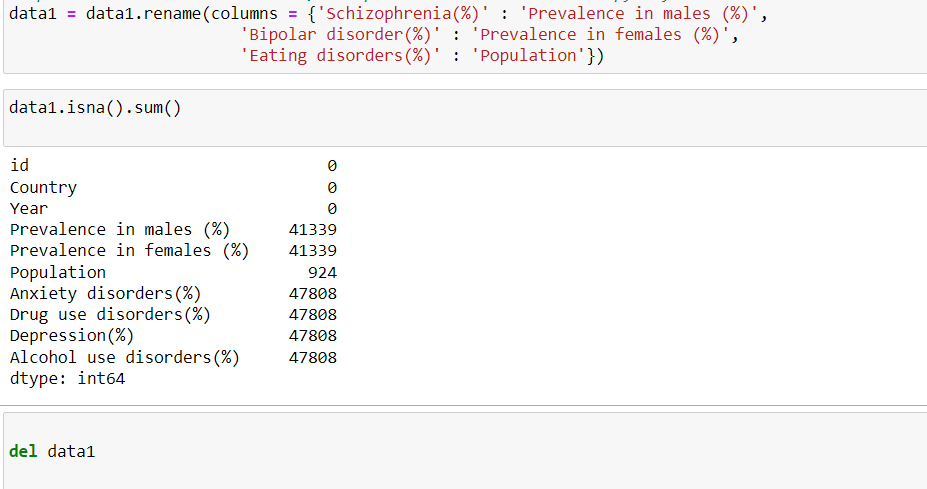
Автоматически созданное описание

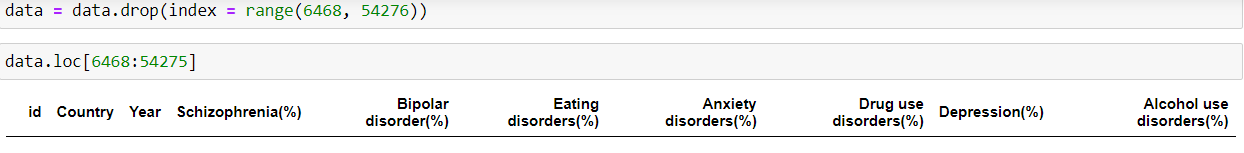
Изображение выглядит как текст

Автоматически созданное описание

Записываем данные в новый дата сет, переименовываем колоны, просматриваем количество NAN и удаляем новый датасет (т.к. большое количество NAN):

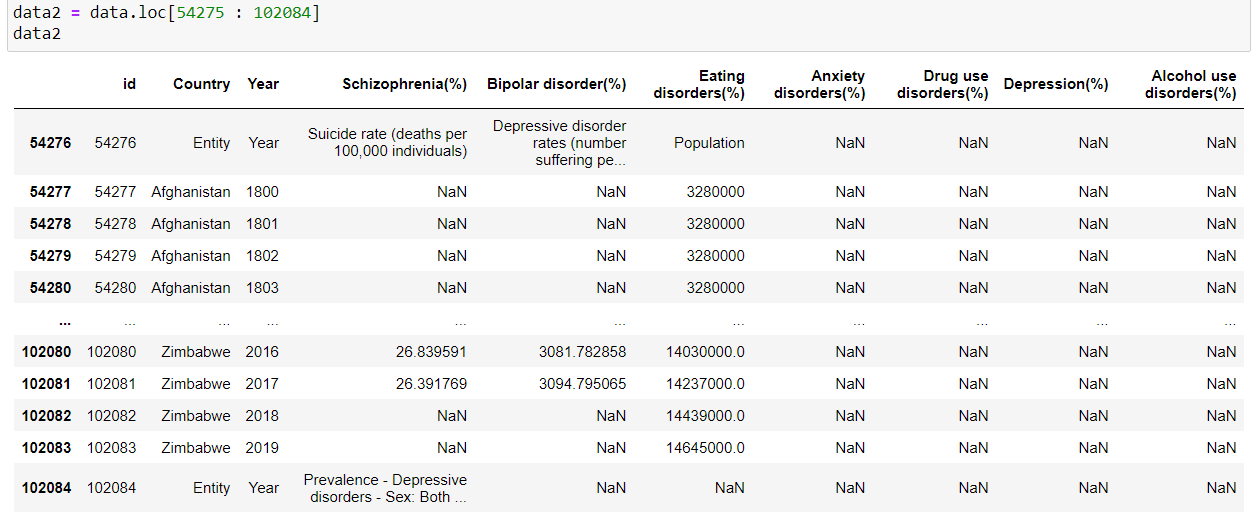


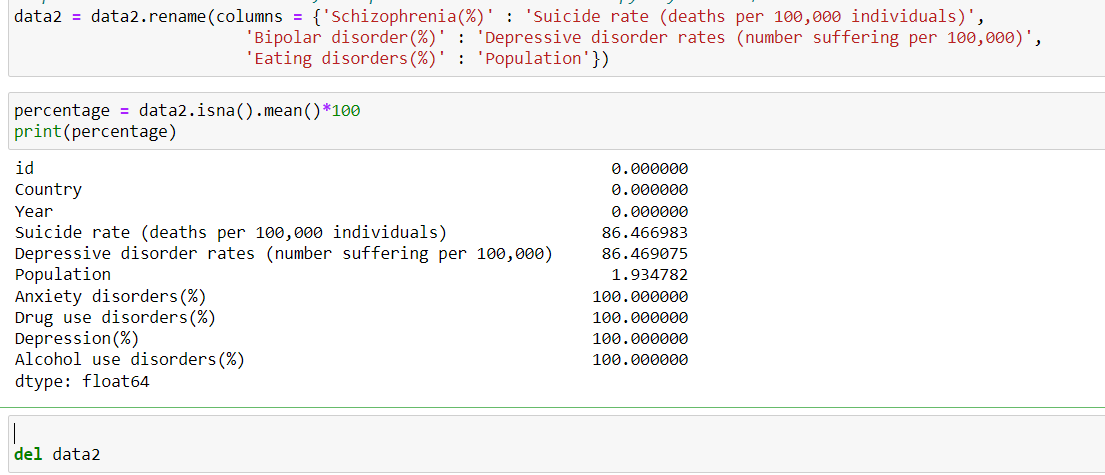


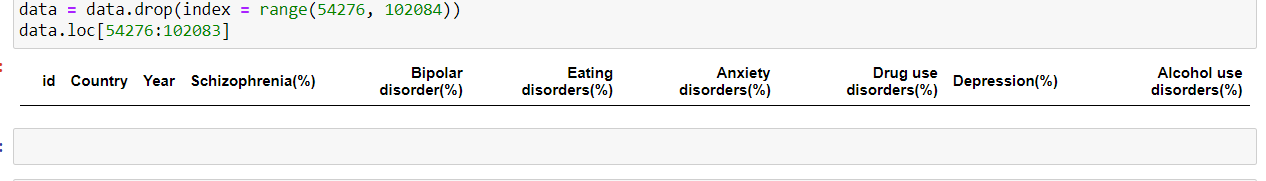


Записываем все данные с 54275 по 102084 в новый датасет, переименовали названия колонок, которые должны называться по-другому. Мы также удалим весь этот датасет, поскольку в колонках 'Prevalence in males(%) and females(%)'

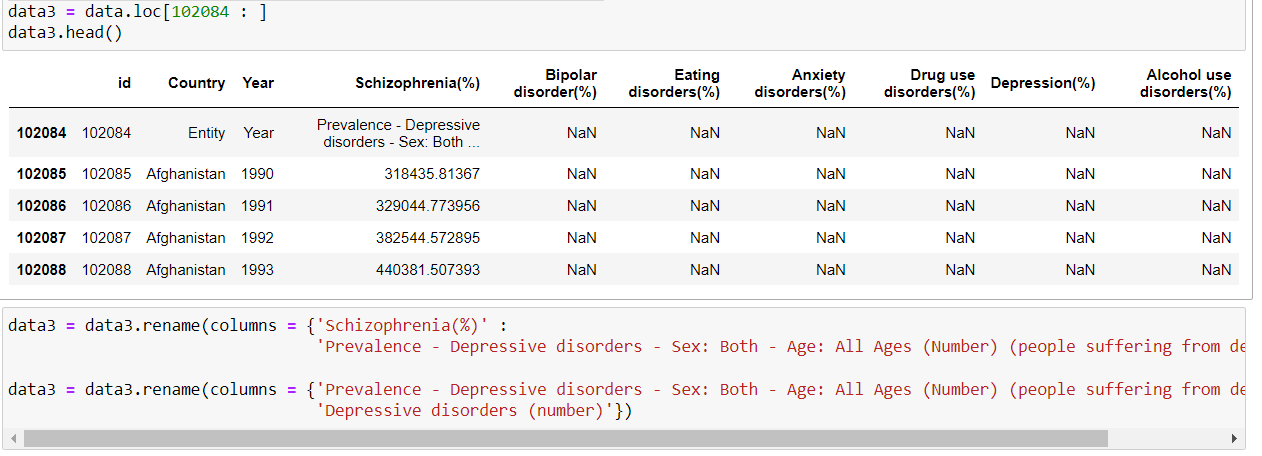
слишком много значений NaN (почти 100%), а в остальных нужных колонках, за исключением id, страны и года, их вовсе нет.

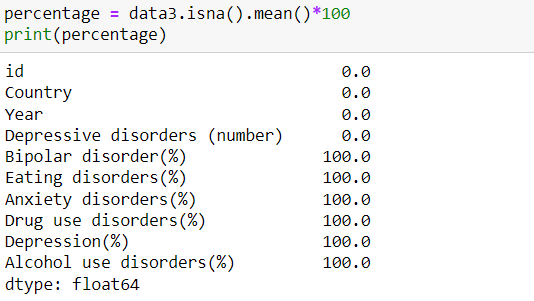




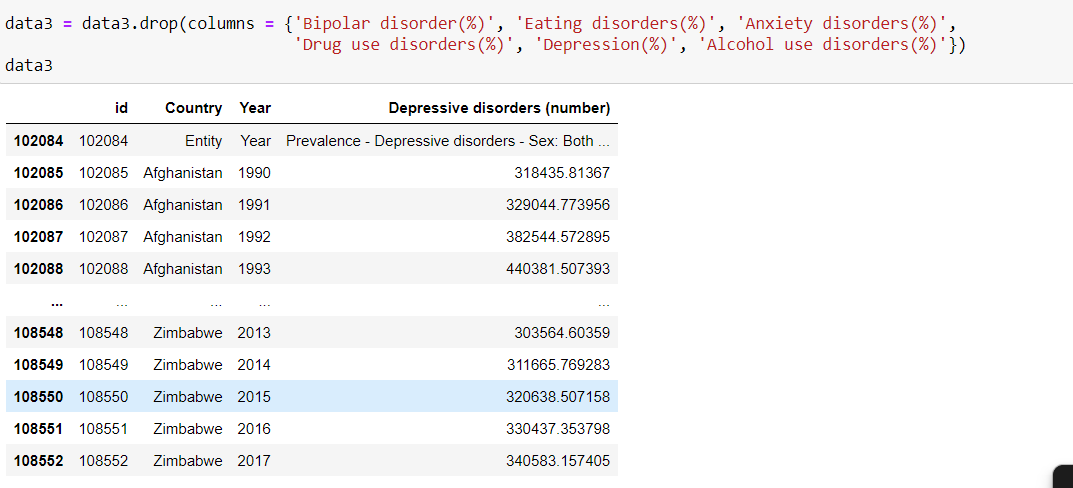
Теперь удалим все строки с 6468 по 54276, так как там совсем нет нужных нам данных, которые можно было бы перенести в отдельные колонки:

Записали все данные с 102084 и до конца списка в новый датасет:





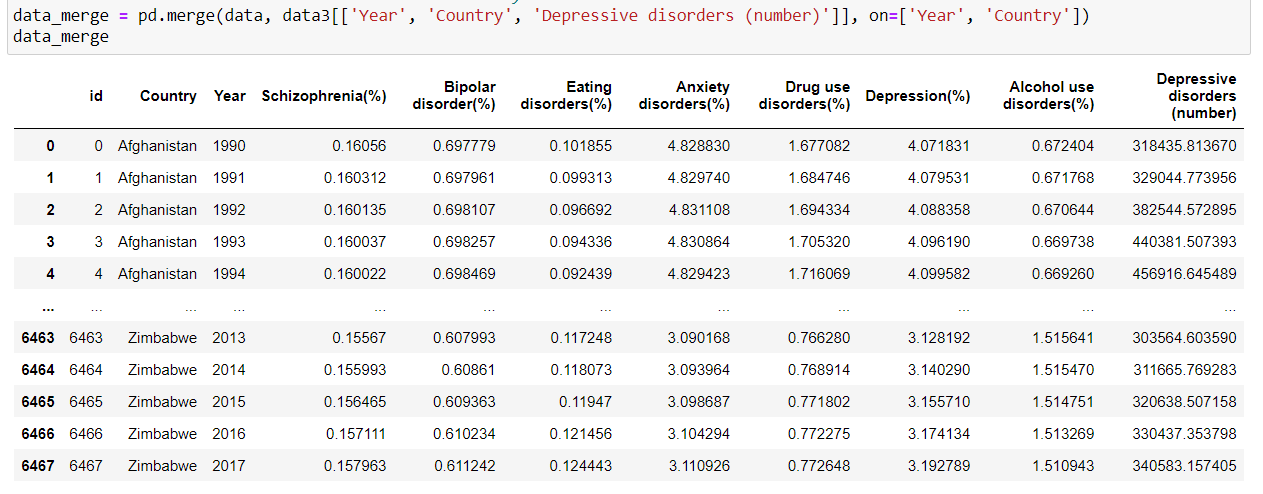
Поскольку здесь у нас нет пустых значений в важной информации, оставим это в новом датасете, а пустые колонки удаляем.



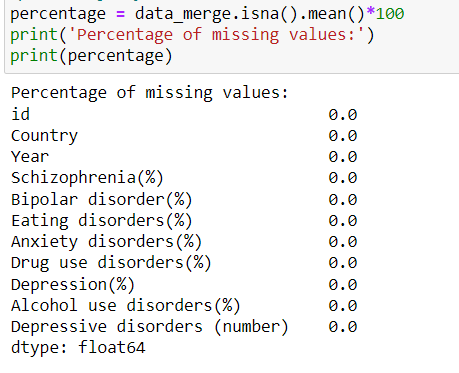
Также удаляем ряды с этими значениями из сновного датасета, поскольку затем мы будем присоединять

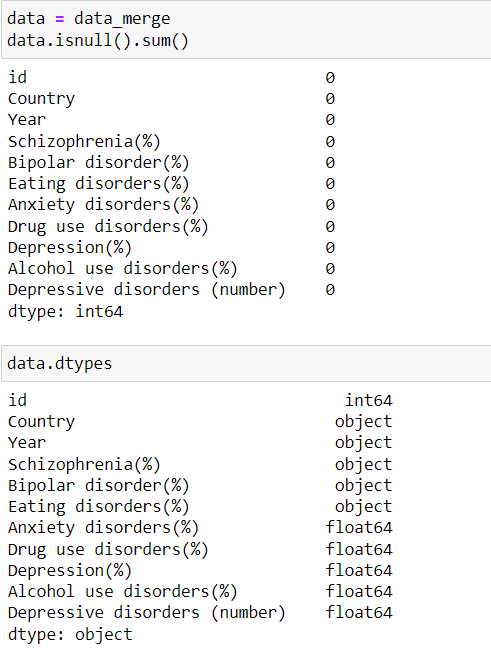


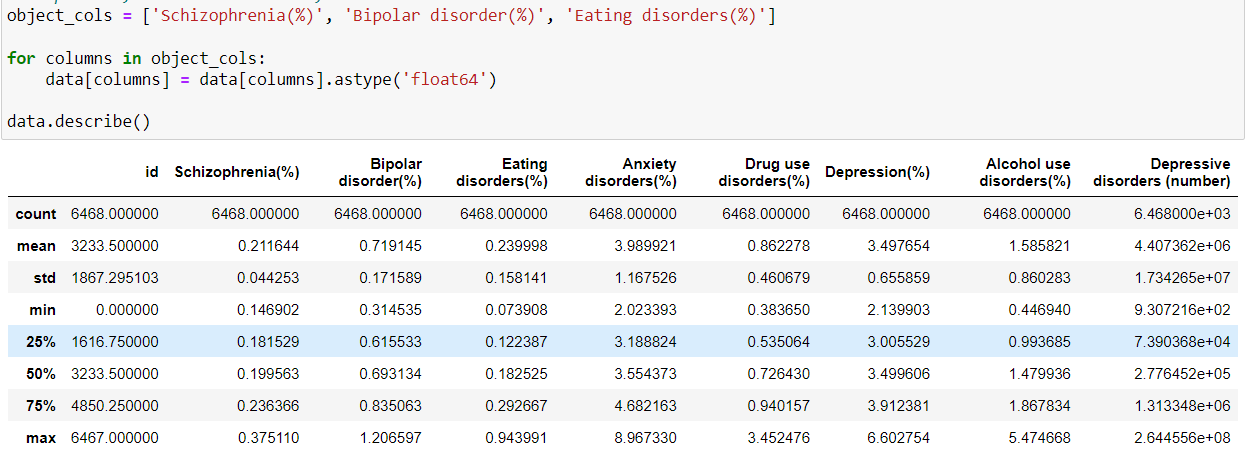
Соединили два датасета по колонкам 'Year' и 'Country':



Считаем процент значений NaN:

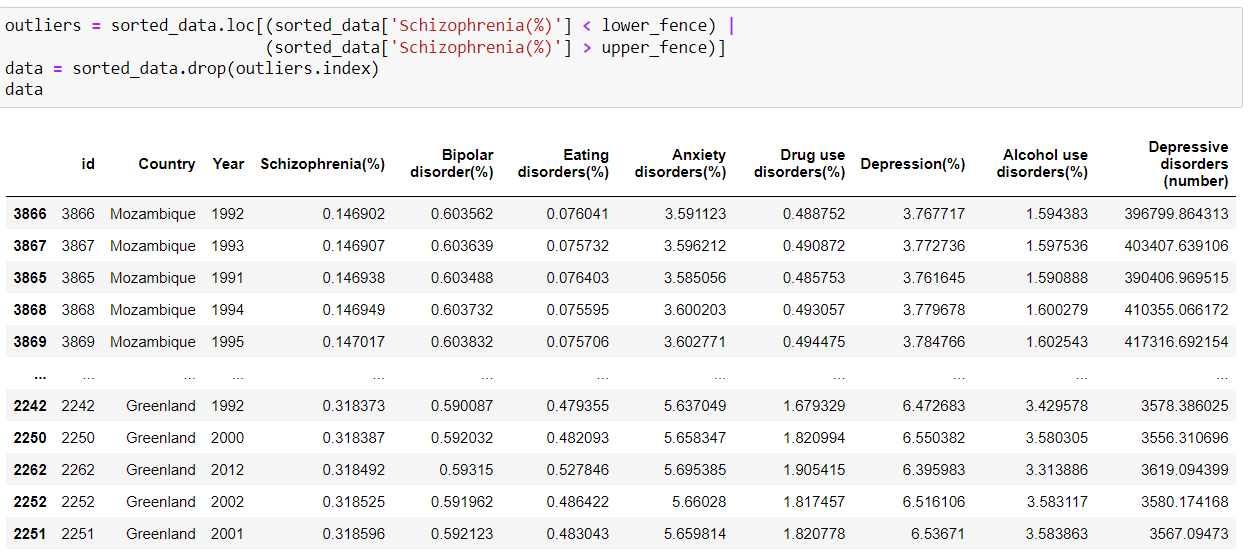
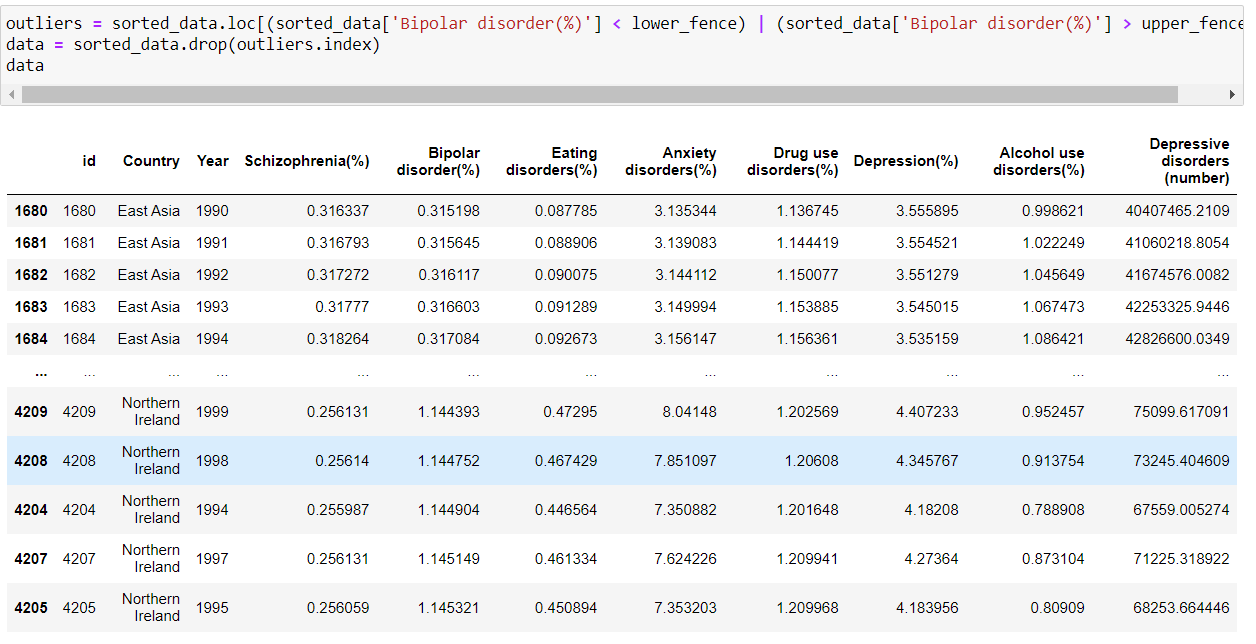
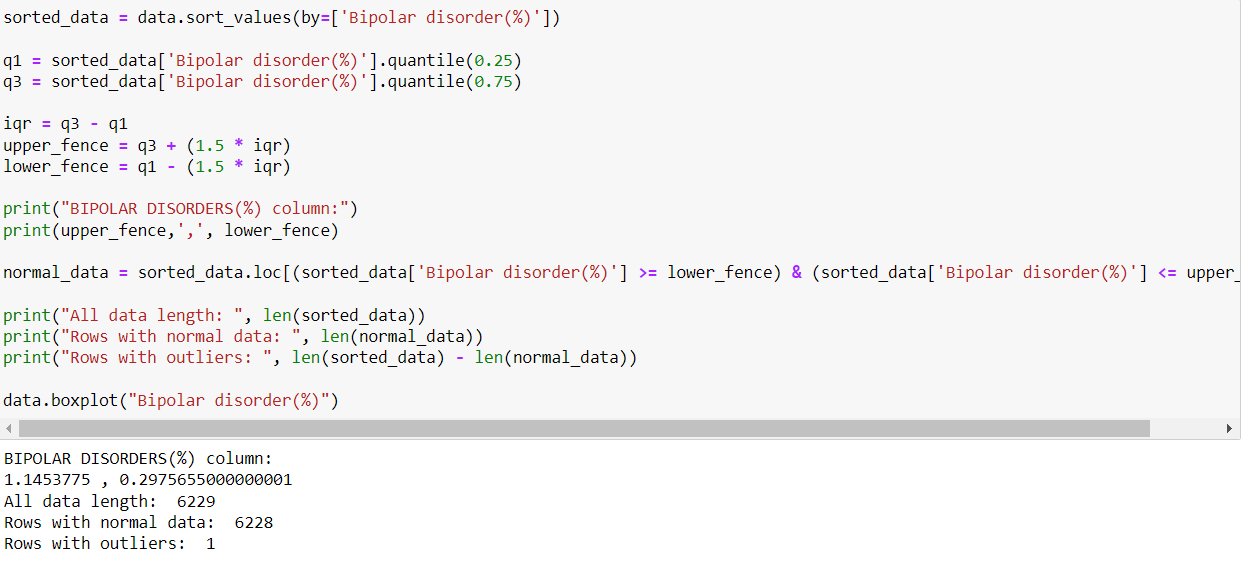




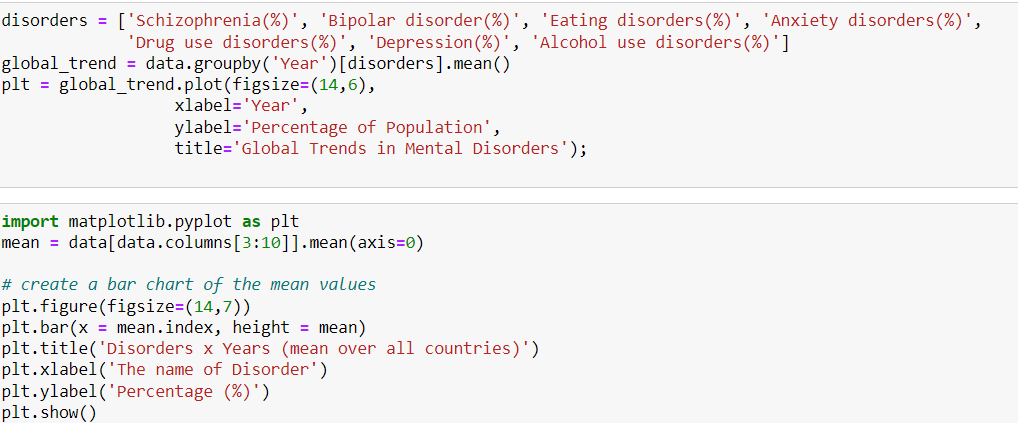
Descriptive statistics включает в себя о central tendency, dispersion, min and max values, и т.д.:

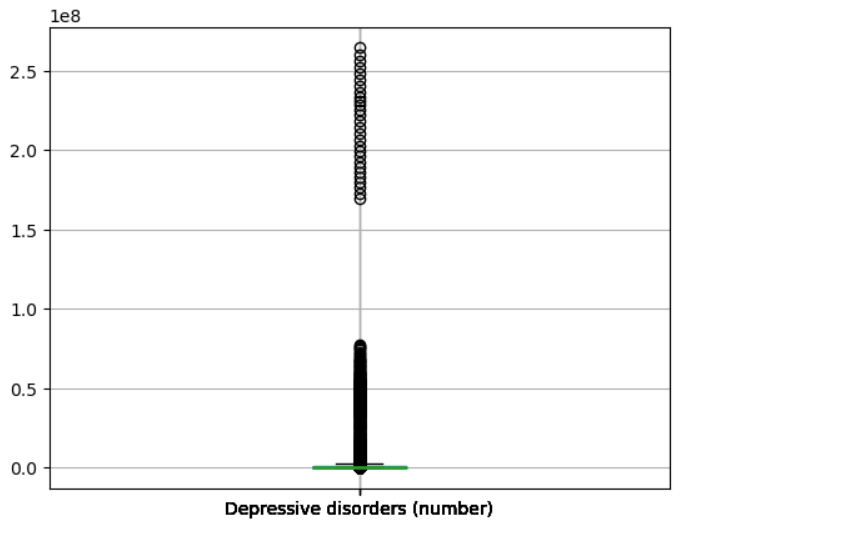
Мы просматриваем выбросы в каждой колоне, после чего удаляем их (показано для 2 столбцов, с остальными - аналогично):

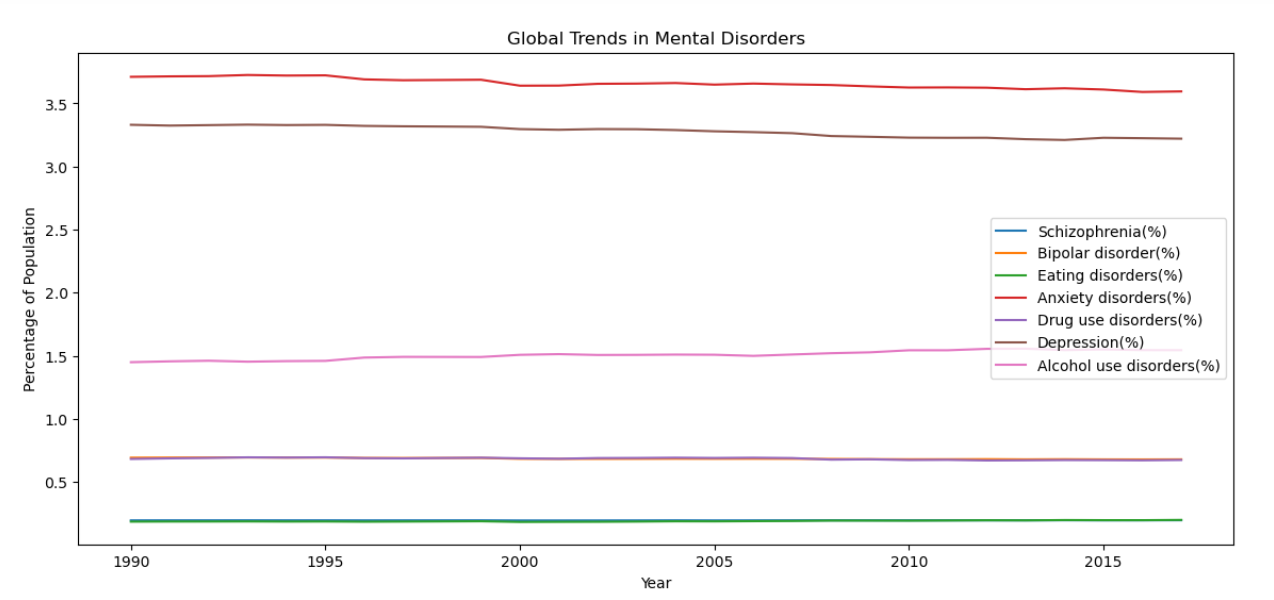


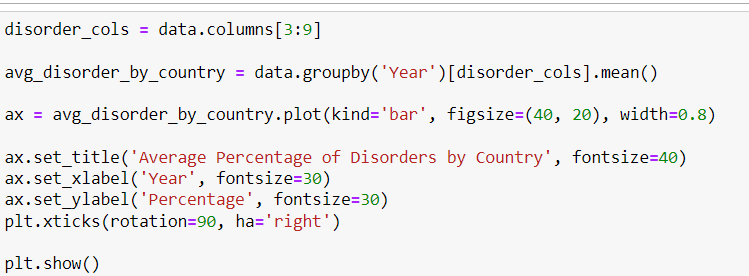


# ****Data exploration (visualisation)****

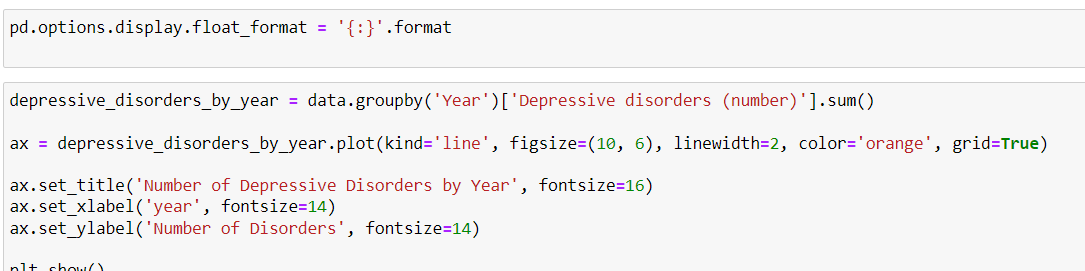


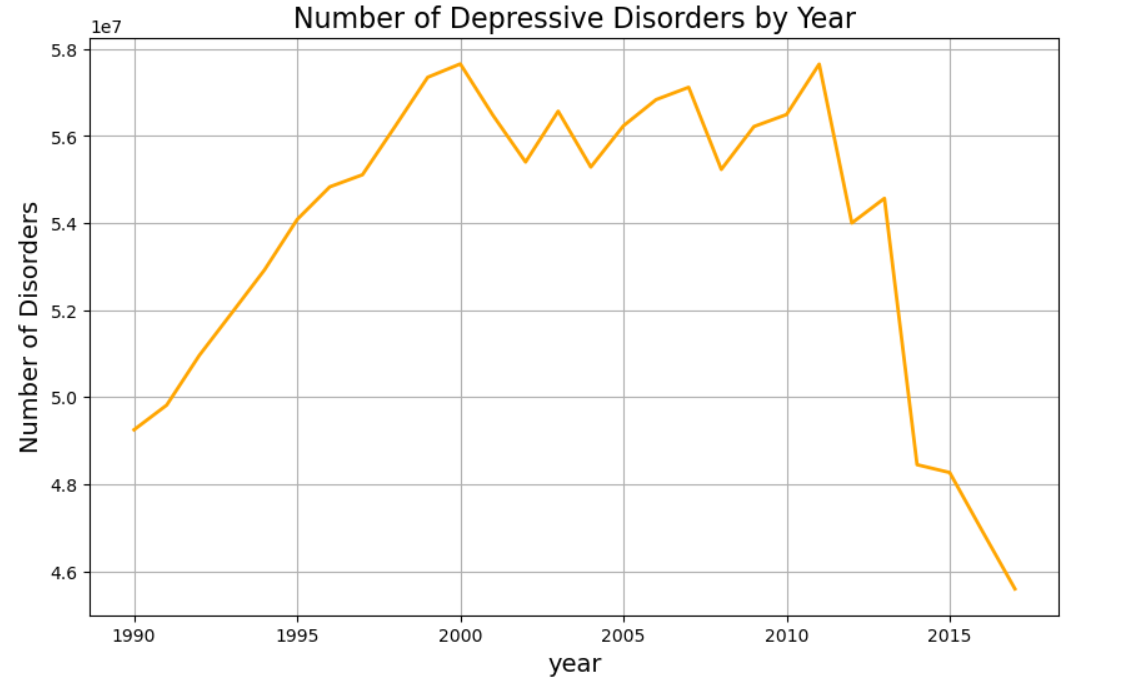


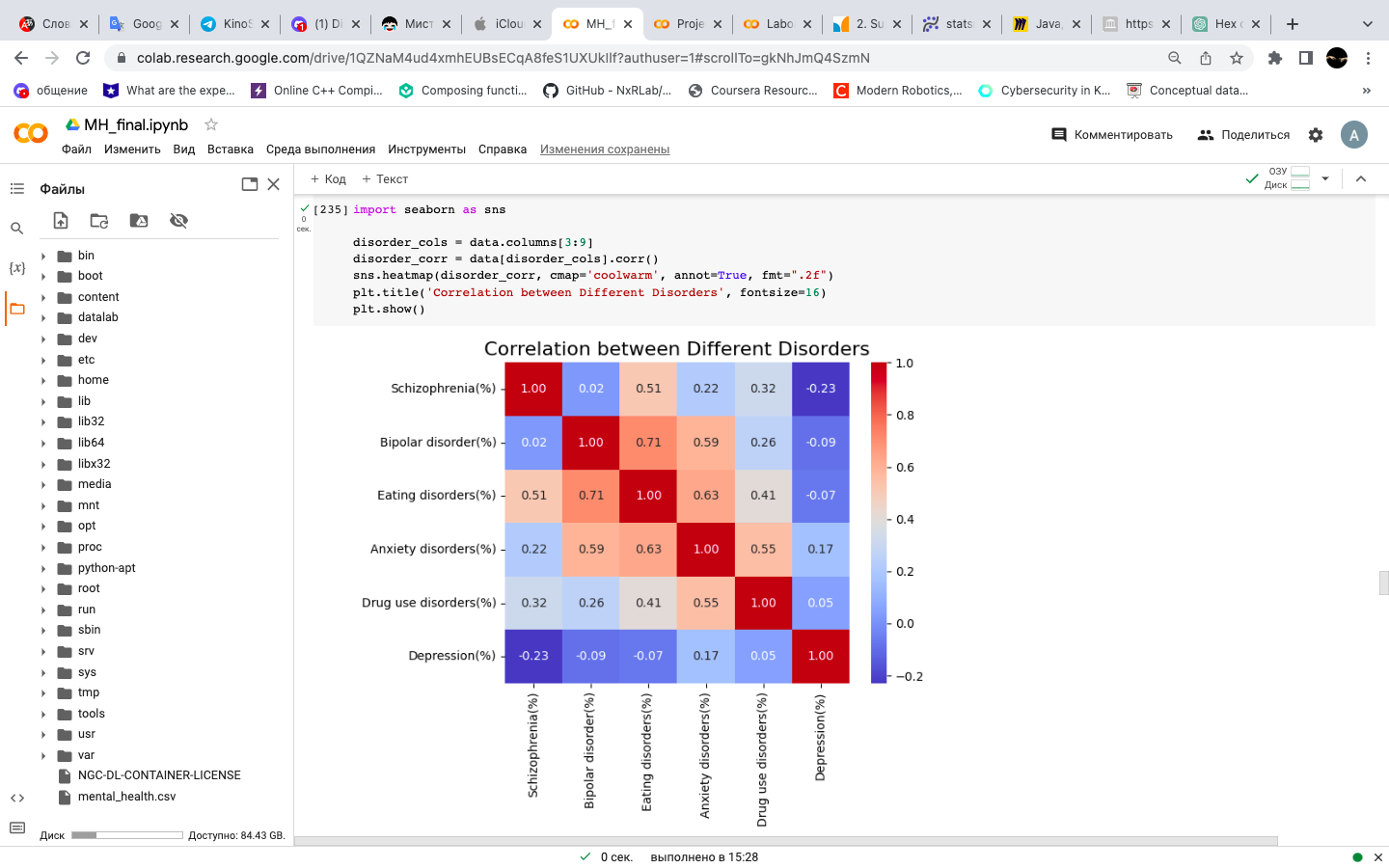


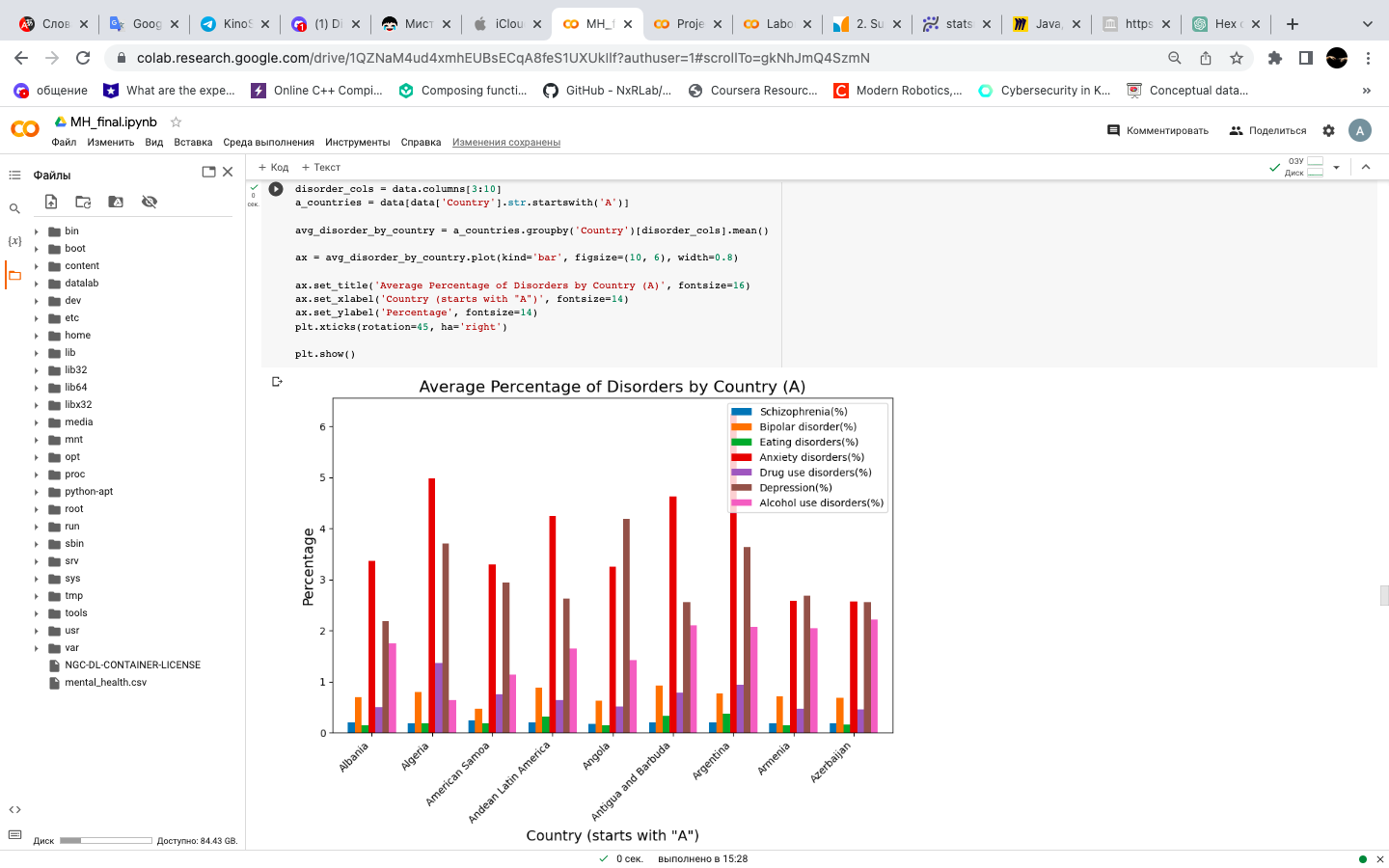


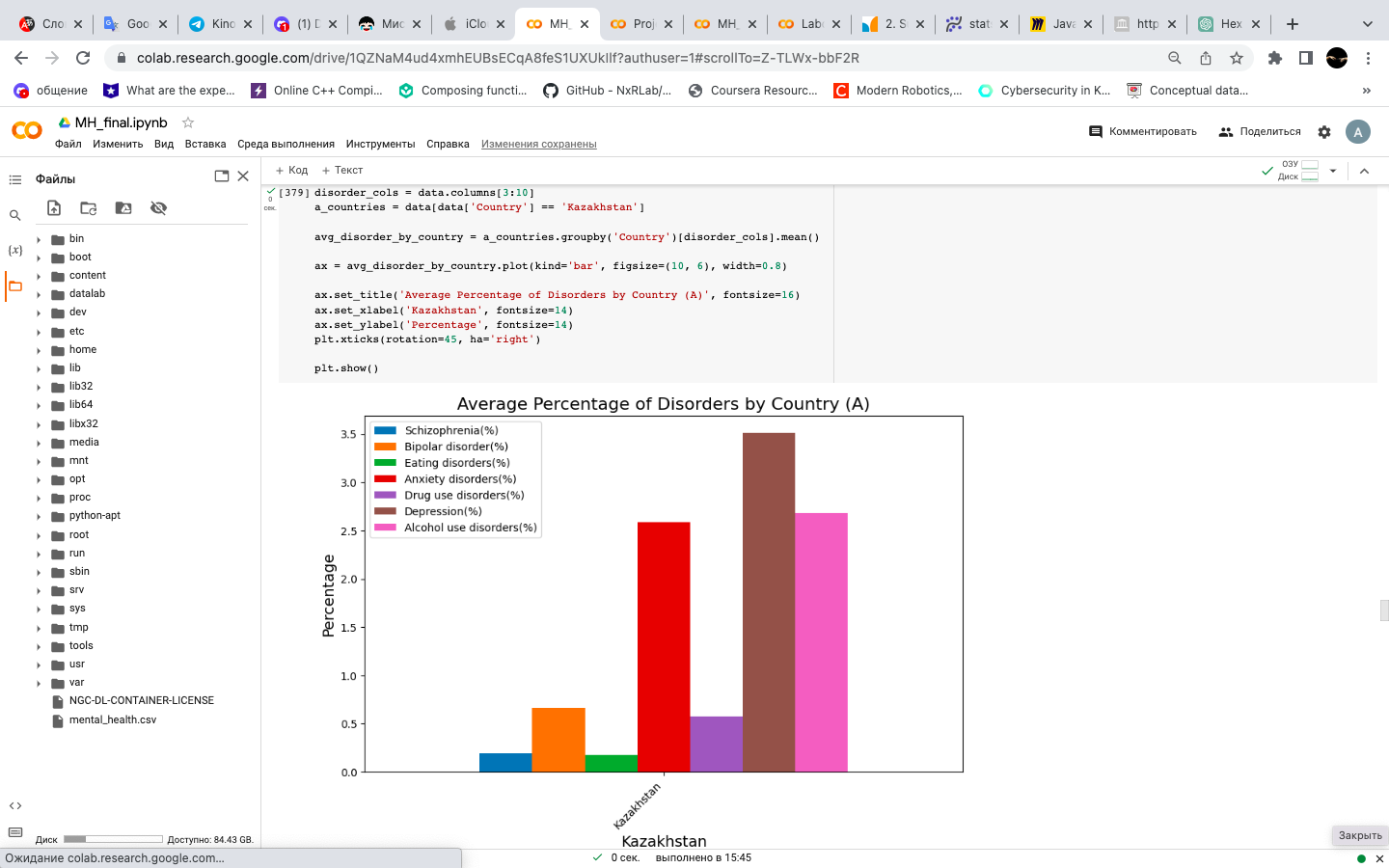


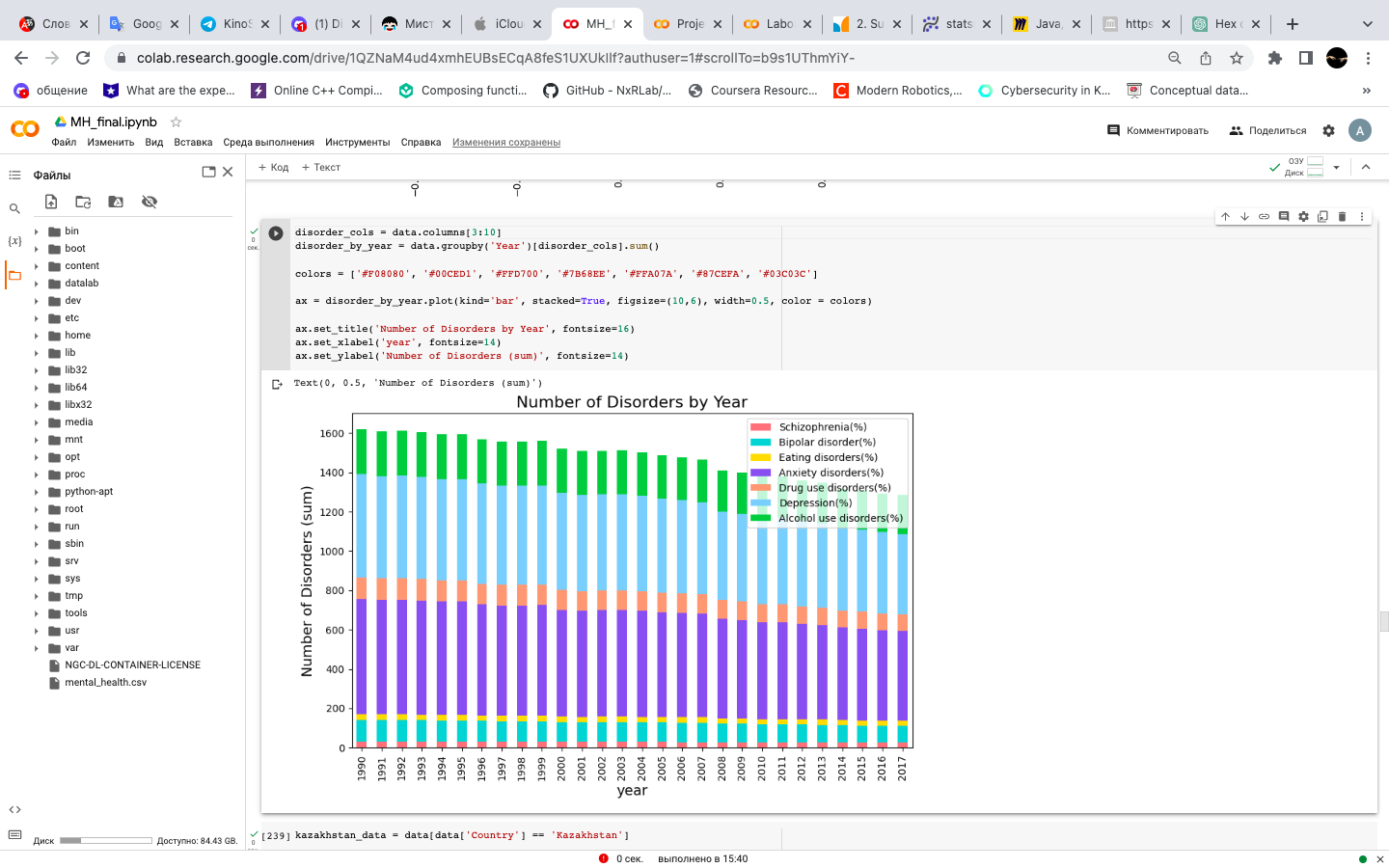


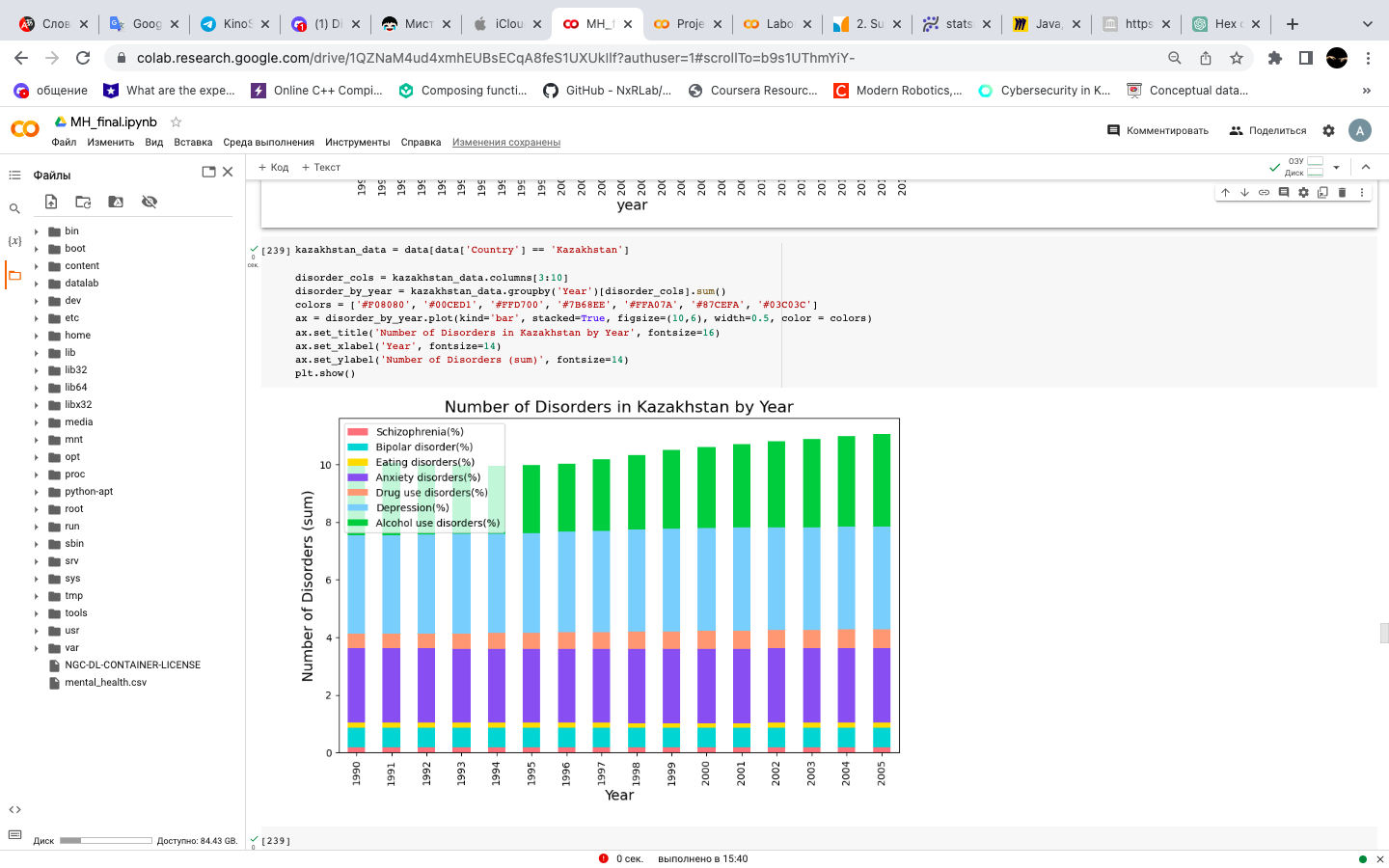




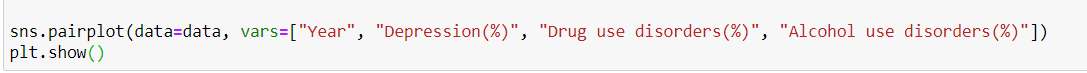


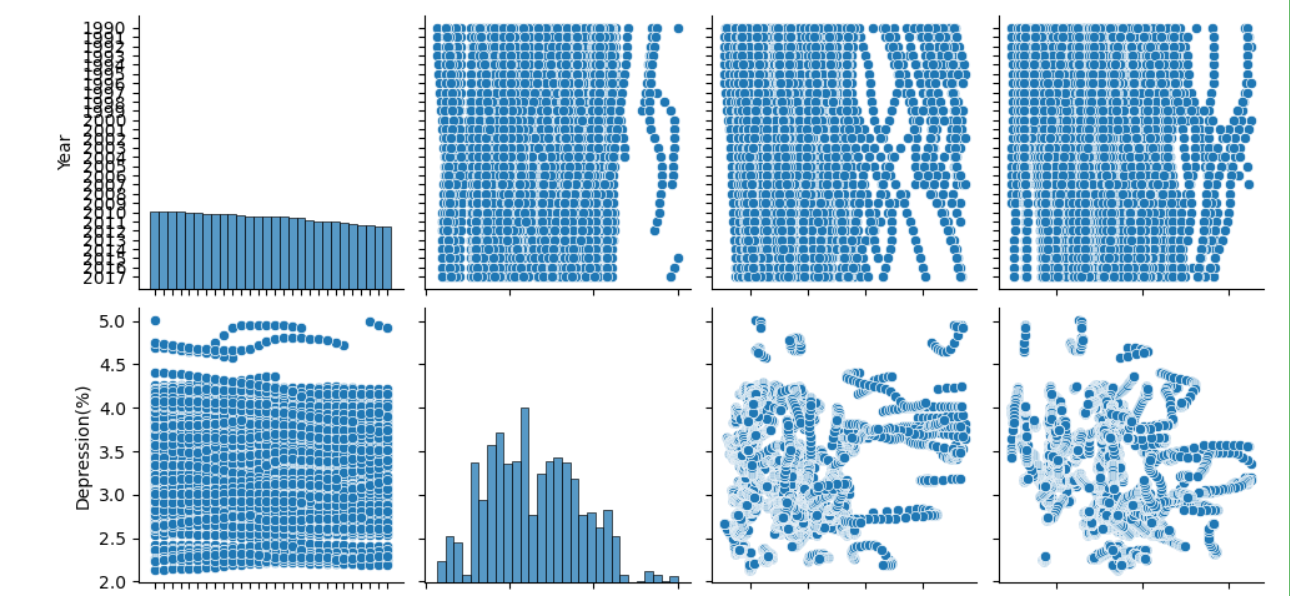


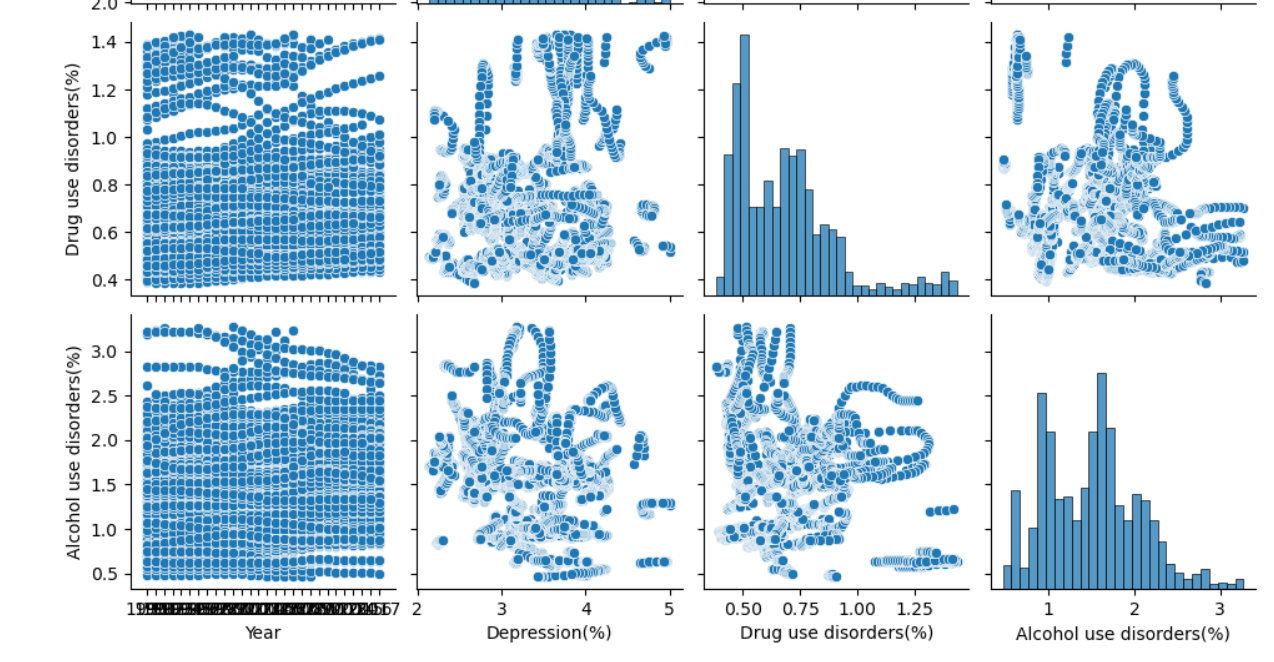




Посмотрим на распределение и связь между количественными переменными например, "Year", "Depression(%)", "Drug use disorders(%)", "Alcohol use disorders(%)" с помощью диаграмм рассеяния.



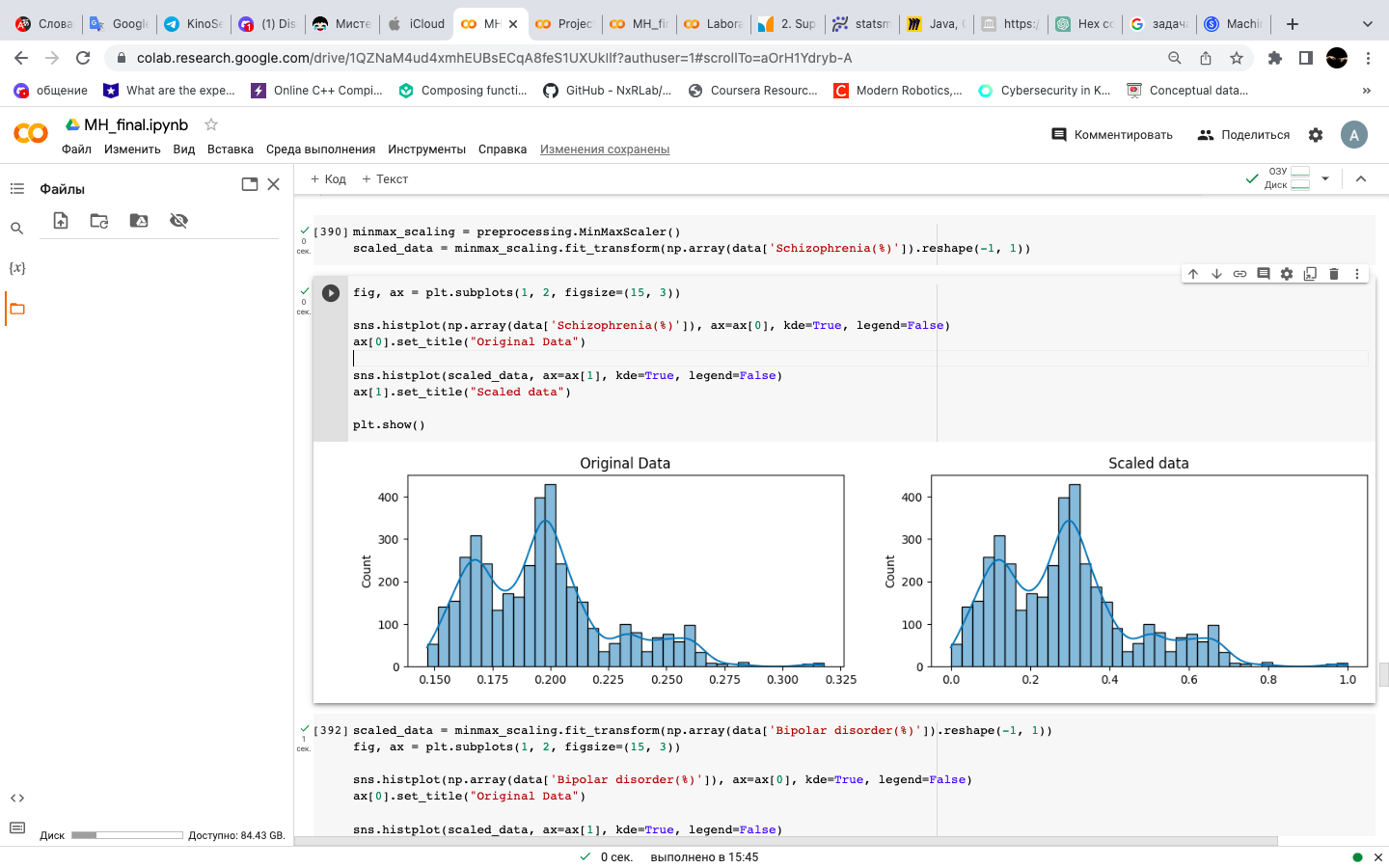




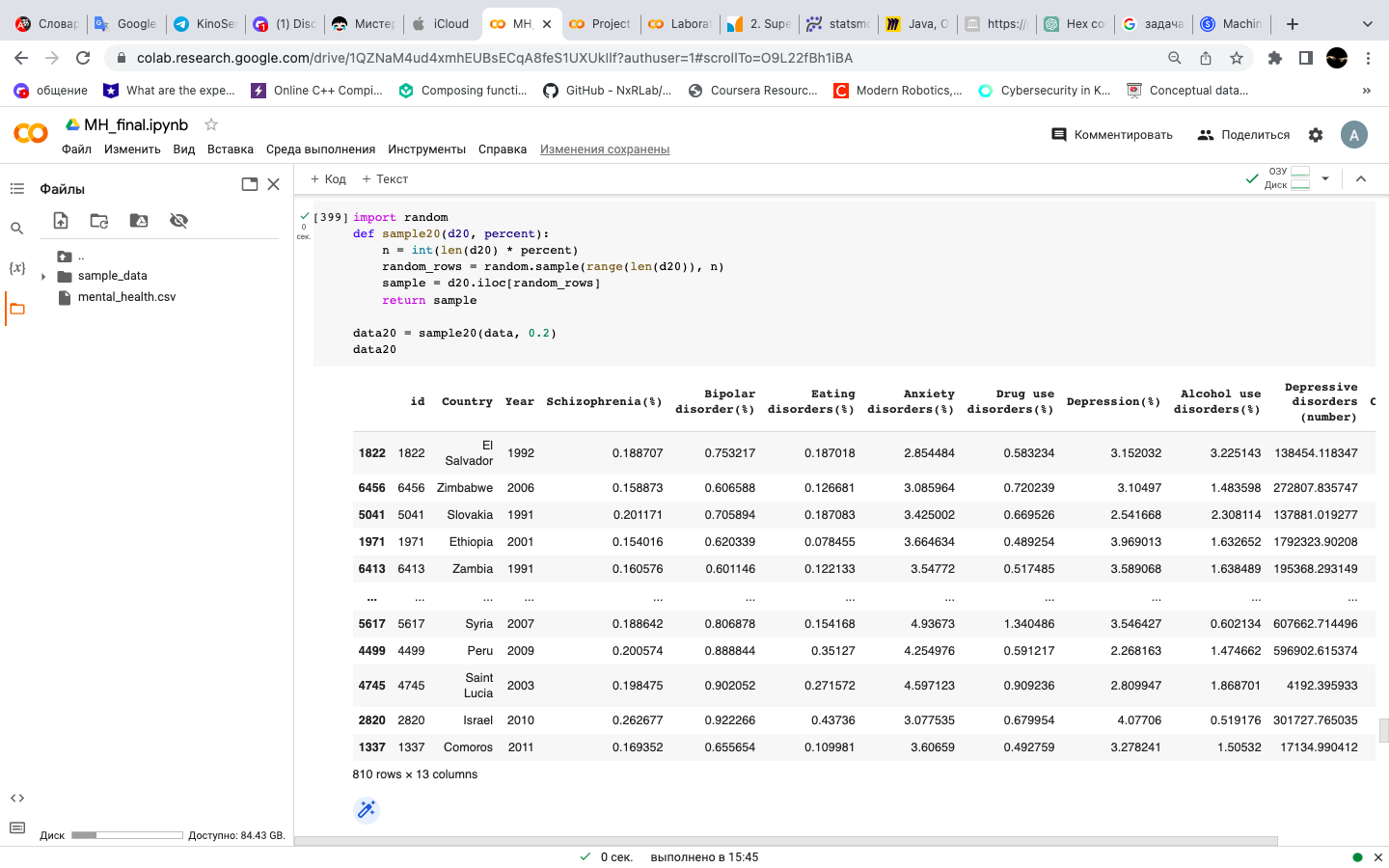
# ****Data Encoding****

# **У нас есть только одна колонка с категориальными значениями – это «Country», поэтому мы закодируем только эту колонку в новую колонку «Country\_enc». (Пока что не буде удалять основную колонку для удобства, удалим её позже):**

**Data Normalization and Data Scaling**

**Мы можем видеть, что scaling уже применён для нашего датасета:**

**Далее мы прописали несколько разных функций для samplings, чтобы можно было извлечь при необходимости рандомные данные в разных объёмах и с определённым «шагом»:**



**Model training and testing**

**Decision tree:**

