**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра ИС**

**Отчет**

**по лабораторной работе №1**

**по дисциплине «Машинное обучение»**

**Тема: «Исследование набора данных»**

|  |  |
| --- | --- |
| Студентка гр. 2373 | Маркова З.А. |
| Преподаватель | Татчина Я.А. |

Санкт-Петербург 2024

**Цель работы**: исследовать алгоритмы классификации и кластеризации на выбранном наборе данных.

1. **Краткое описание набора данных**

Этот набор данных моделирует набор ключевых экономических, социальных и экологических показателей Всемирного банка для 20 стран за период с 2010 по 2019 год. Датасет был взят с сайта [www.kaggle.com](http://www.kaggle.com).

Характер данных: Смоделированные данные, сгенерированные на основе реальных сценариев на основе типичных показателей в стиле Всемирного банка.

Атрибуты датасета:

* Country - Название страны, для которой записаны данные. Тип данных: string.
* Year - Год, в котором наблюдаются данные. Тип данных: integer.
* GDP(USD) - Валовой внутренний продукт страны в миллиардах долларов США. Тип данных: float.
* Population - Общая численность населения страны в миллионах. Тип данных: float.
* Life Expectancy - Средняя продолжительность жизни (в годах). Тип данных: float.
* Unemployment Rate (%) - Уровень безработицы в процентах от общей численности рабочей силы. Тип данных: float.
* CO2 Emissions (metric tons per capita) - Количество выбросов углекислого газа на человека в стране, измеряемое в метрических тоннах. Тип данных: float.
* Access to Electricity (%) - Процент населения, имеющего доступ к электричеству. Тип данных: float.

1. **Среднее значение и СКО**

Среднее значение и СКО были посчитаны с помощью методов mean() и std() библиотеки Pandas.

В результате чего были получены следующие результаты:

Year: среднее = 2014.5; СКО = 2.8722813232690143

GDP: среднее = 10568674550000.0; СКО= 5533816584665.849

Population: среднее = 738790850.0; СКО = 437896696.8815562

Life Expectancy: среднее = 66.3245; СКО = 9.794281482069014

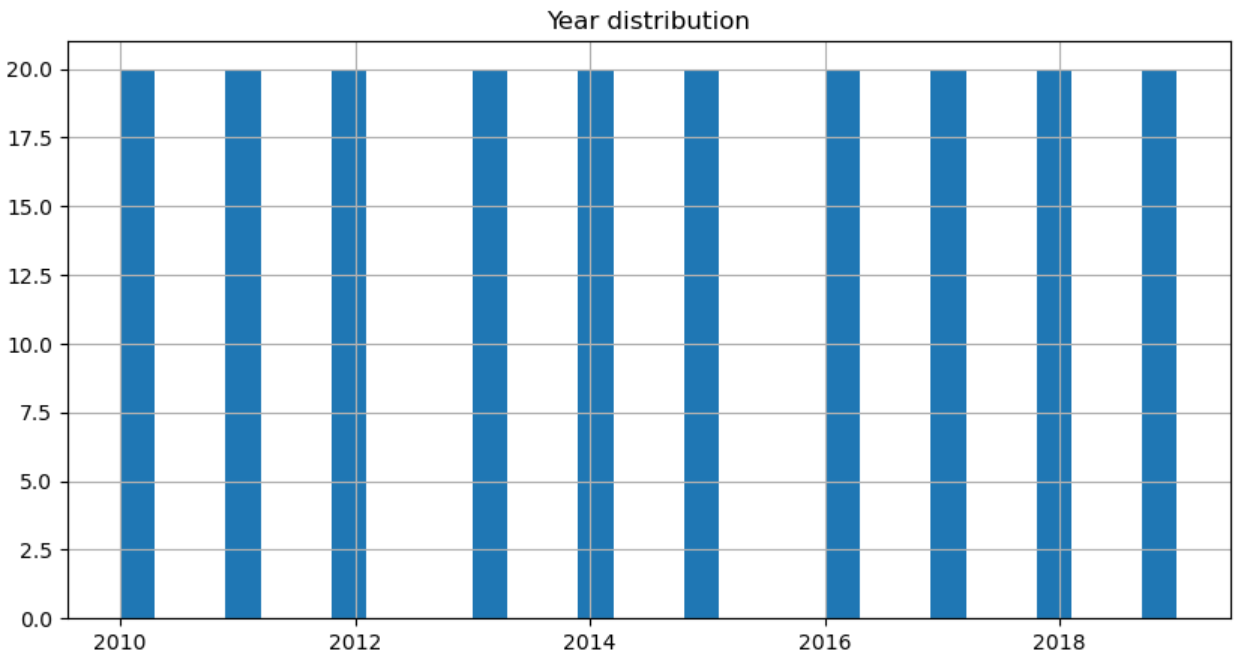
Unemployment Rate: среднее = 13.27165; СКО = 6.787134135811669

CO2 Emissions: среднее = 10.0582; СКО = 5.697826933138633

Access to Electricity: среднее = 72.87674999999999; СКО = 14.75426687224749

Наличие пропусков было определено с помощью методов df['…'].isna().sum(), что позволило установить, что ни у одного атрибута нет пропущенных значений.

1. **Гистограммы и выбросы**

С помощью функций библиотеки Matplotlib были построены следующие гистограммы атрибутов:

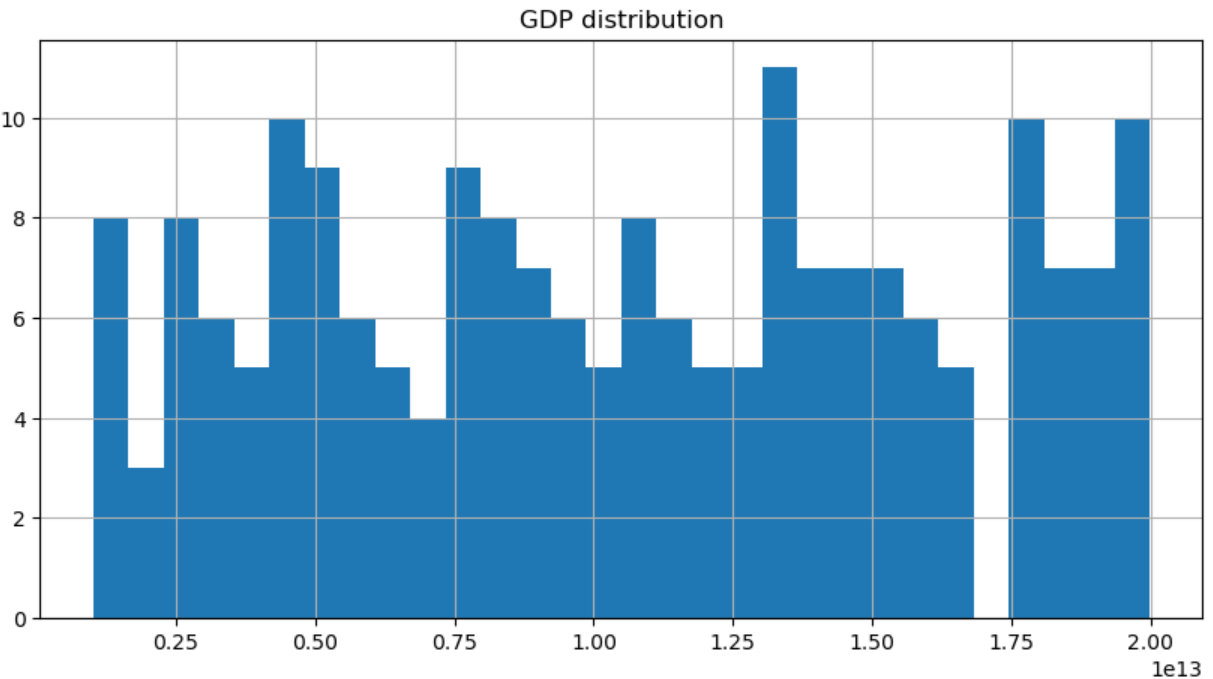
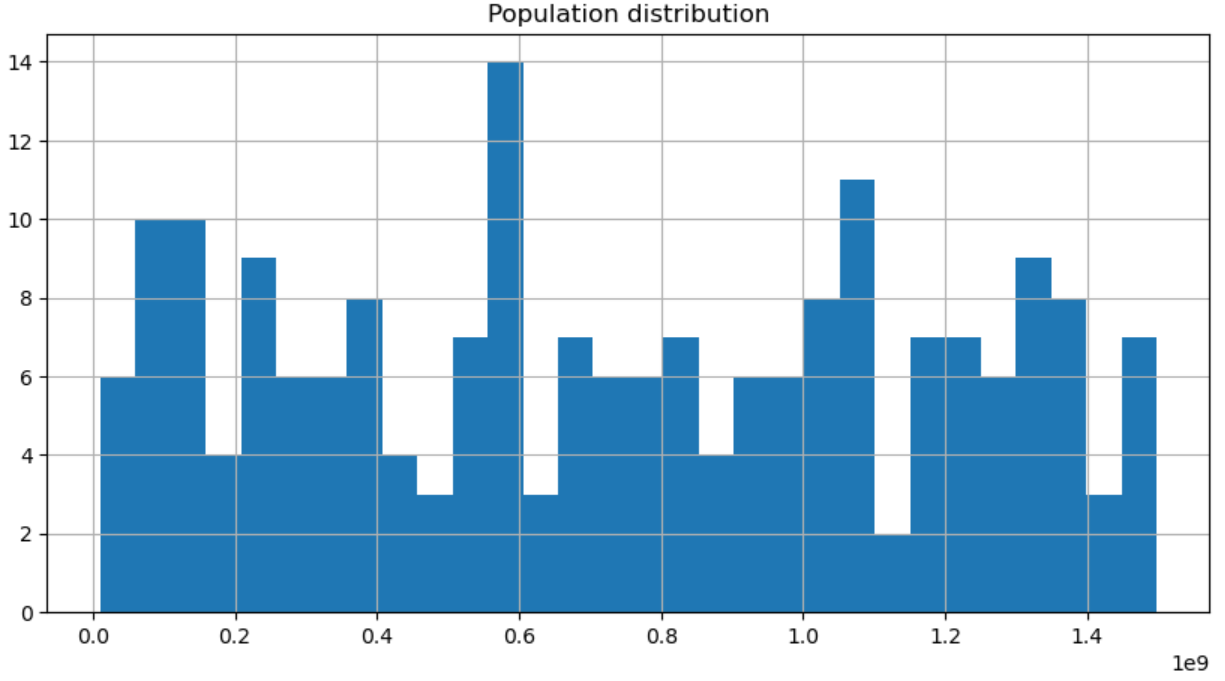
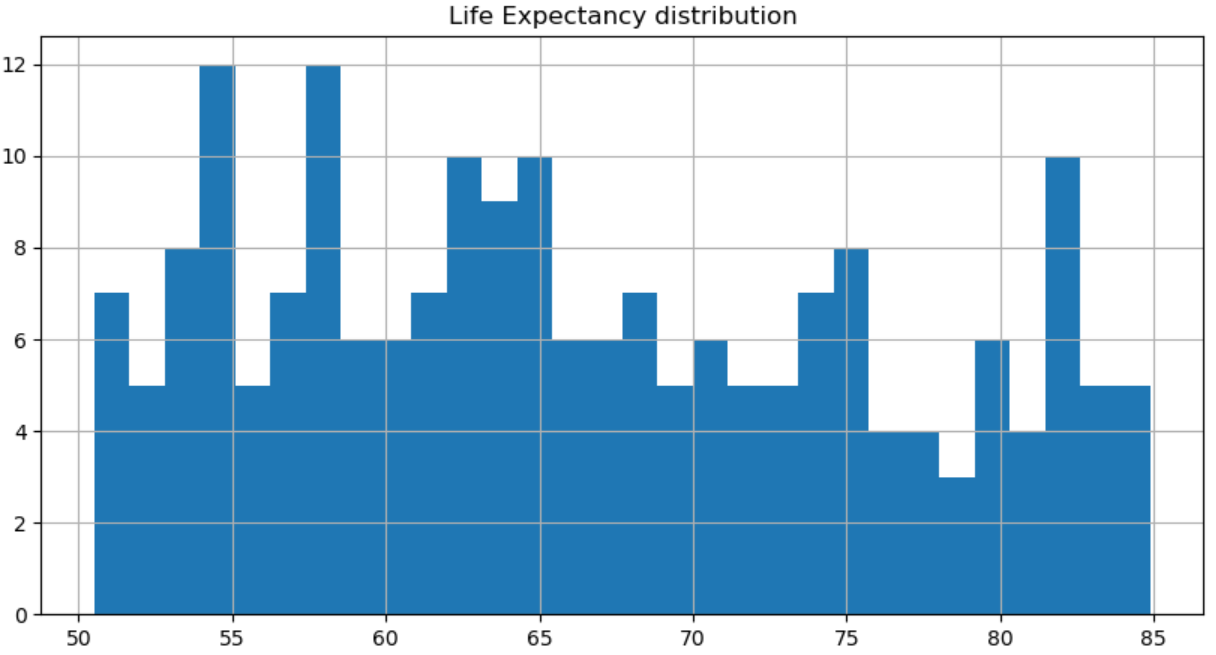
Рис. 1 Гистограмма атрибута «Year»

Рис. 2 Гистограмма атрибута «GDP»

Рис. 3 Гистограмма атрибута «Population»

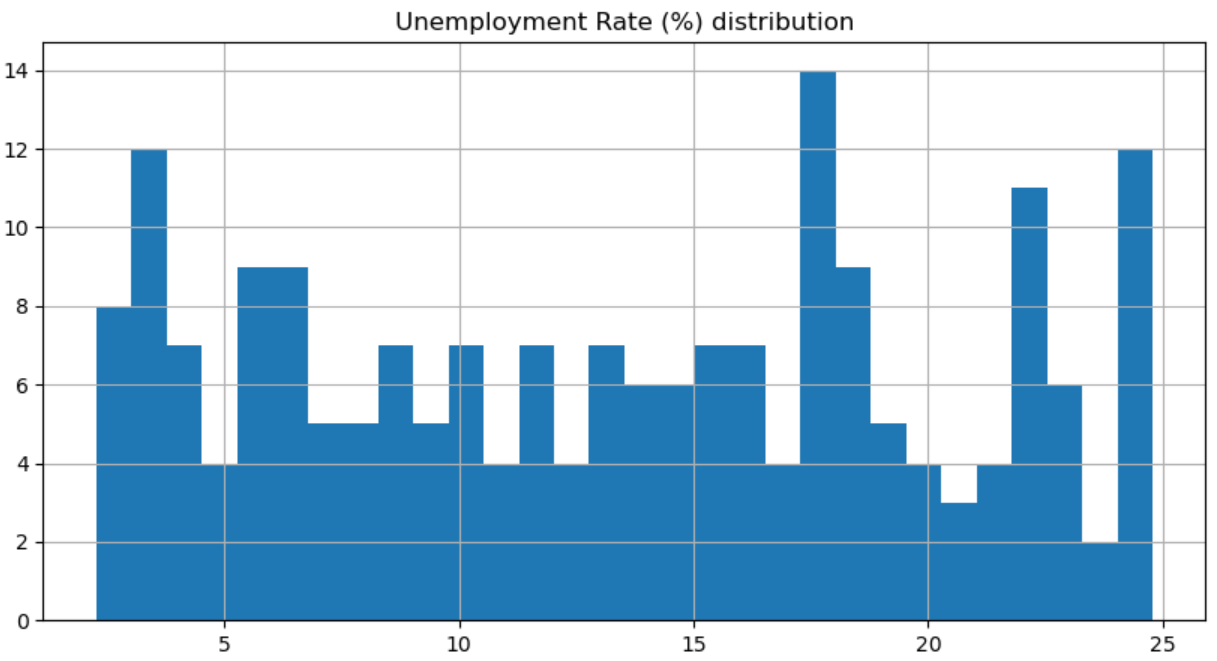
Рис. 4 Гистограмма атрибута «Life Expectancy»

Рис. 5 Гистограмма атрибута «Unemployment Rate»

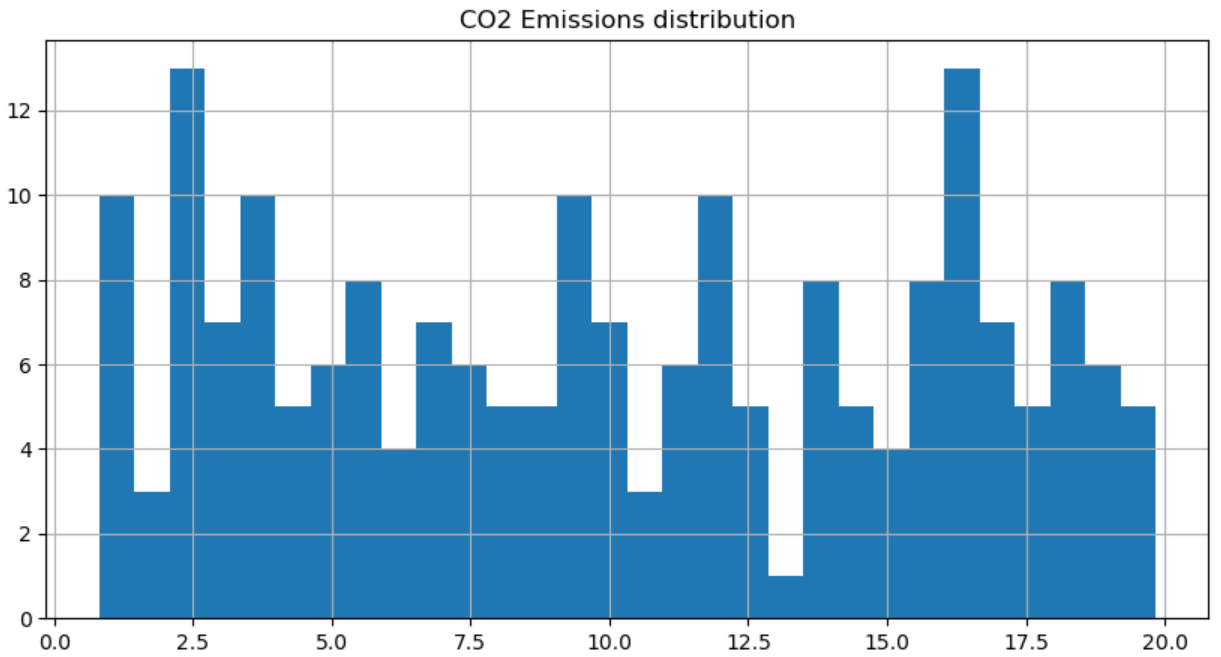
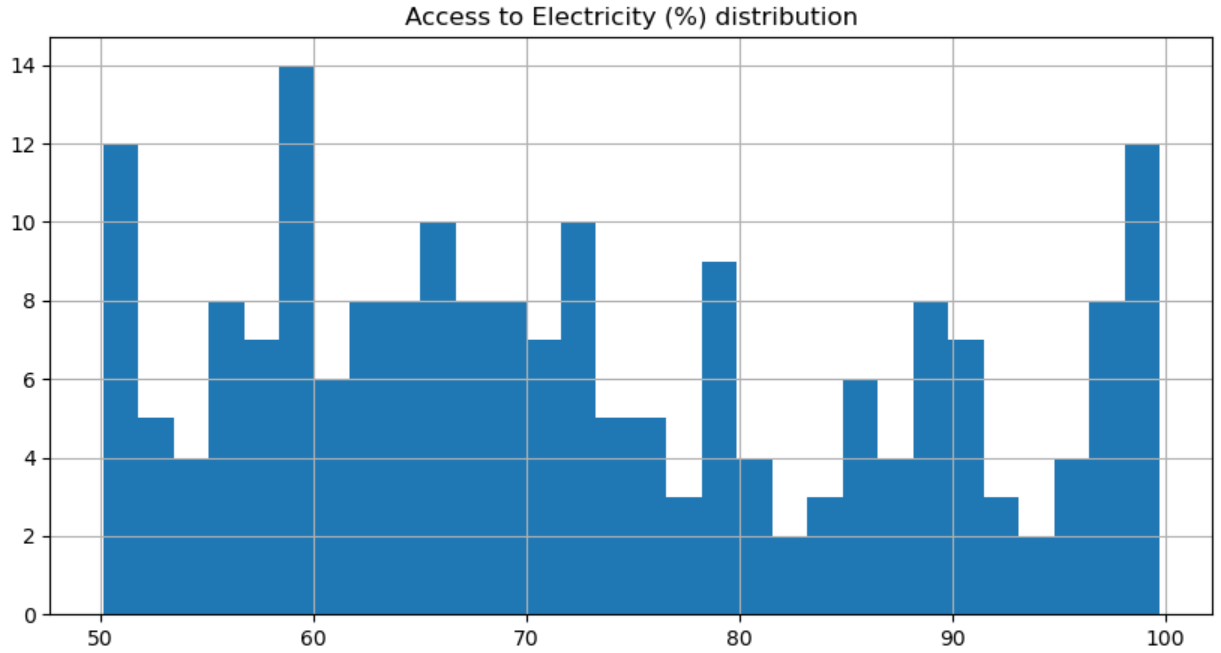
Рис. 6 Гистограмма атрибута «CO2 Emissions»

Рис. 7 Гистограмма атрибута «Access to Electricity»

Наличие выбросов было определено с помощью межквартильного диапазона:

1. Данные разделены на квартили: Q1 — 25-я процентиль данных, Q3 — 75-я процентиль данных. С использованием функции percentile() библиотеки Numpy.
2. IQR(Межквартальный диапазон)=Q3-Q1
3. Значения, которые попадают ниже Q1 — 1,5\* IQR или выше Q3 + 1,5 IQR, считаются выбросами.

***У всех атрибутов выбросы отсутствуют***

1. **Корреляция.**

Коэффициент корреляции был найден с помощью матрицы корреляции, которая была построена функцией corrcoef() библиотеки Numpy.

В данной работе были рассмотрены 7 зависимостей:

* 1. *ВВП от Уровня безработицы*

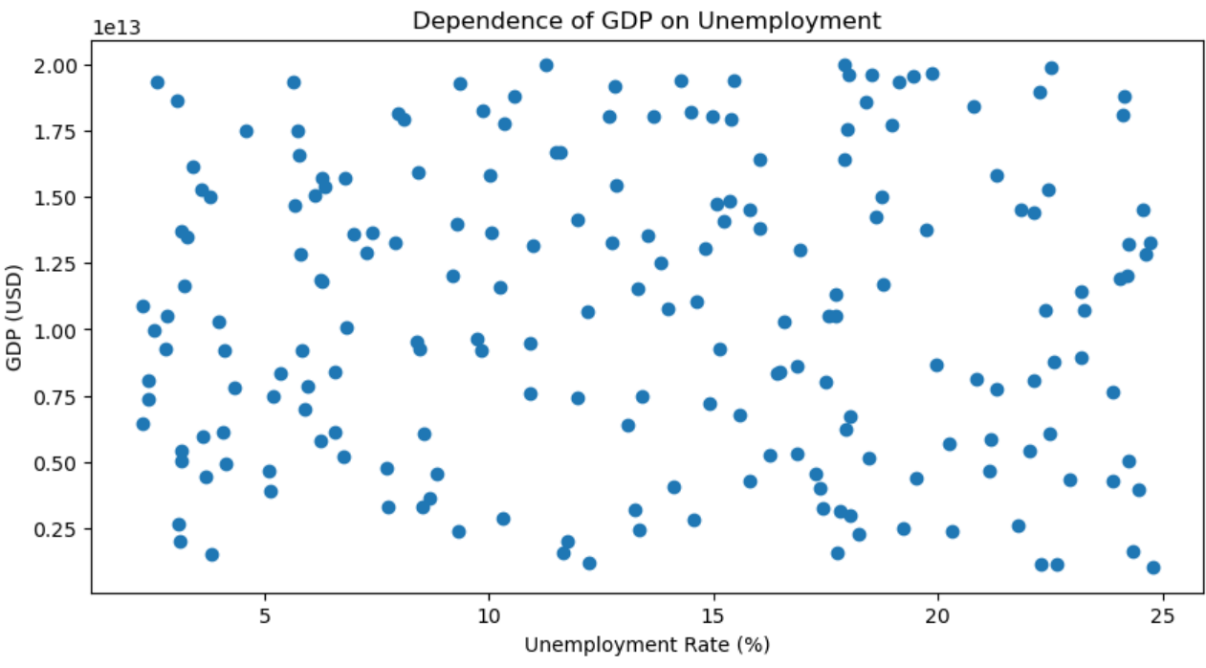
Коэффициент корреляции = -0.01196997. Слабая отрицательная корреляция, значит атрибуты слабо зависимы, а при увеличении одного другой будет уменьшаться.

Рис. 8 График рассеивания «Зависимость ВВП от Безработицы»

* 1. *Доступ к электричеству от ВВП*

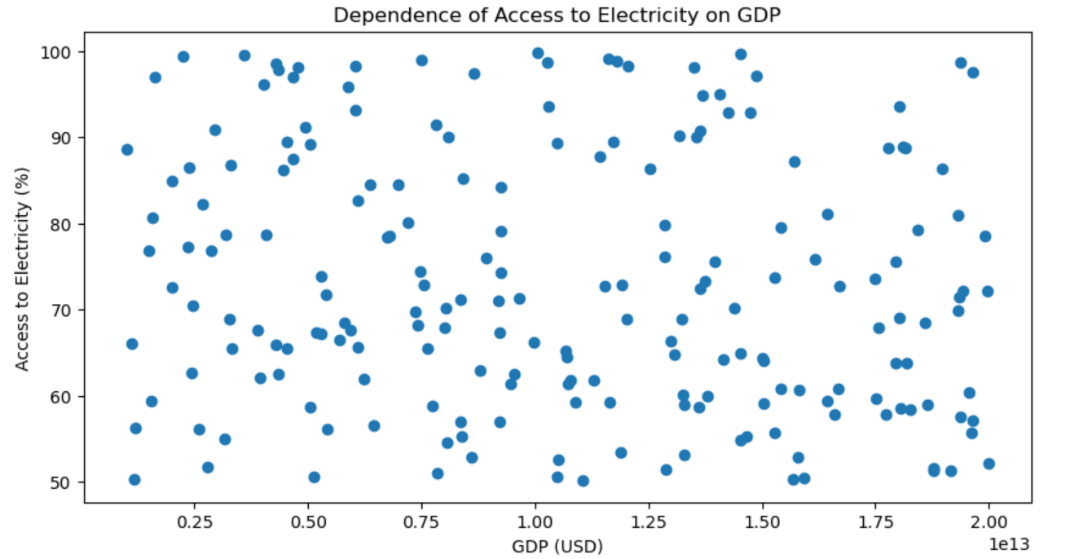
Коэффициент корреляции = -0.14895434. Слабая отрицательная корреляция, значит атрибуты слабо зависимы, а при увеличении одного другой будет уменьшаться.

Рис. 9 График рассеивания «Зависимость Доступа к электричеству от ВВП»

* 1. *Продолжительность жизни от Выбросов CO2*

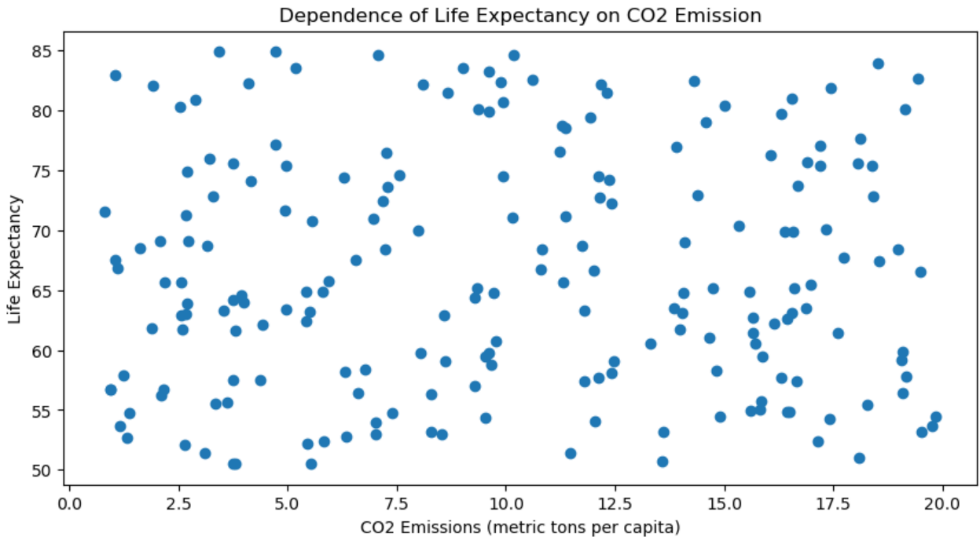
 Коэффициент корреляции = 0.02301529. Слабая положительная корреляция, значит атрибуты слабо зависимы, а при увеличении одного другой будет тоже увеличиваться

Рис. 10 График рассеивания «Зависимость Прод. Жизни от Выбросов CO2»

* 1. *Выбросы CO2 от Численности населения*

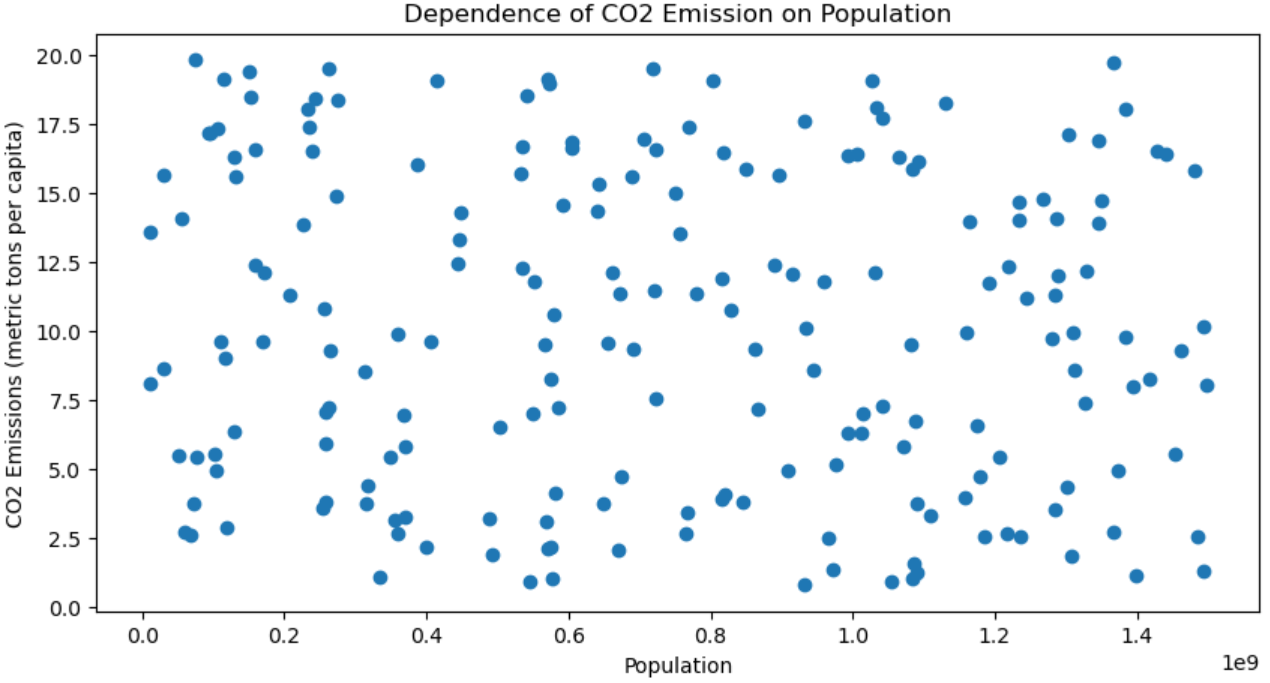
 Коэффициент корреляции = -0.08116944. Слабая отрицательная корреляция, значит атрибуты слабо зависимы, а при увеличении одного другой будет уменьшаться.

Рис. 11 График рассеивания «Зависимость Выбросов CO2 от Численности населения»

* 1. *Уровень безработицы от Численности населения*

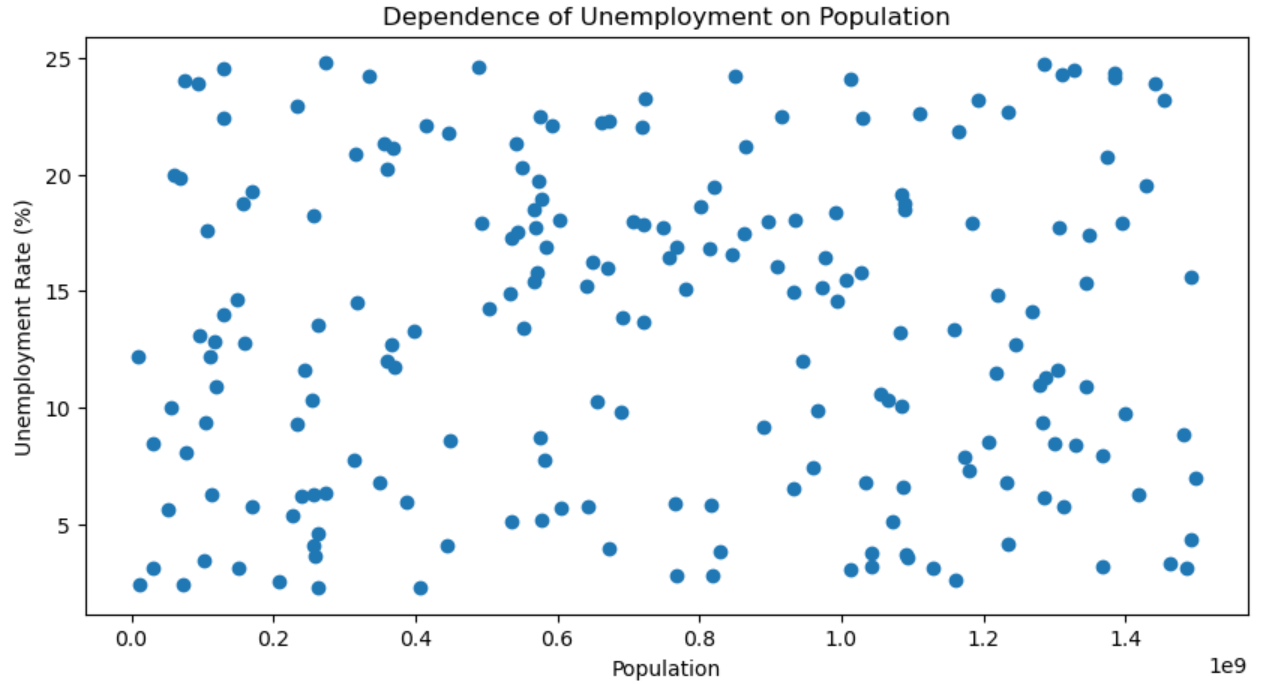
 Коэффициент корреляции = 0.02803804. Слабая положительная корреляция, значит атрибуты слабо зависимы, а при увеличении одного другой будет тоже увеличиваться

Рис. 12 График рассеивания «Зависимость Уровня безработицы от Численности населения»

* 1. *Численность населения от Продолжительности жизни*

Коэффициент корреляции = -0.09969296. Слабая отрицательная корреляция, значит атрибуты слабо зависимы, а при увеличении одного другой будет уменьшаться.

Рис. 13 График рассеивания «Зависимость Численности населения от Прод. жизни»

Из полученных расчётов можно сделать вывод, что атрибуты слабо зависимы, что может быть связано с тем, что эти данные, сгенерированны на основе типичных показателей в стиле Всемирного банка и могут отличаться от реальных.