Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Лабораторная работа №1
по дисциплине
«Методы машинного обучения»
на тему

«Разведочный анализ данных. Исследование и визуализация данных»

Выполнил:
студент группы ИУ5-22М
ЧжаоЛян

Москва — 2022 г.

1. Цель лабораторной работы

Изучить различные методы визуализации данных [1].

2. Задание

Требуется выполнить следующие действия [1]:

• Выбрать набор данных (датасет).

• Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного набора данных.
2. Основные характеристики датасета.
3. Визуальное исследование датасета.
4. Информация о корреляции признаков.

• Сформировать отчет и разместить его в своем репозитории на GitHub

3. Ход выполнения работы

3.1. Текстовое описание набора данных

В этой тетради я буду использовать графики для визуализации взаимосвязи между переменными в наборе данных "Успеваемость студентов на экзаменах".

- **The dataset includes the following columns:**
  Gender (object) - The gender of the participant
  - female
  - male
- **Race/Ethnicity** (object) - The race or ethnicity of the participant
  - Group A
  - Group B
  - Group C
  - Group D
  - Group E
- **Parental level of education** (object) - The parental level of education of the participant
  - bachelor's degree
  - some college
  - master's degree
  - associate's degree
  - high school
  - some high school
- **Lunch** (object) - Whether the participant is:

    - standard
    - free/ reduced
- **Test preparation course** (object) - Whether the participant took the test preparation course or not
    - none
    - completed
- **Math score** (int64) - The participants math score
- **Reading score** (int64) - The participants reading score
- **Writing score** (int64) - The participants writing score

3.2. Основные характеристики набора данных

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.impute import SimpleImputer
from sklearn.impute import MissingIndicator
from sklearn.impute import KNNImputer
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.model_selection import GridSearchCV
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer
from IPython.display import Image
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

In [3]:
```
data = pd.read_csv(r'C:\Users\80667\Desktop\文件\ИУ5\研一下\MMO\数据集\印度人口\india_population.csv')
data.head()
```

Out[3]:

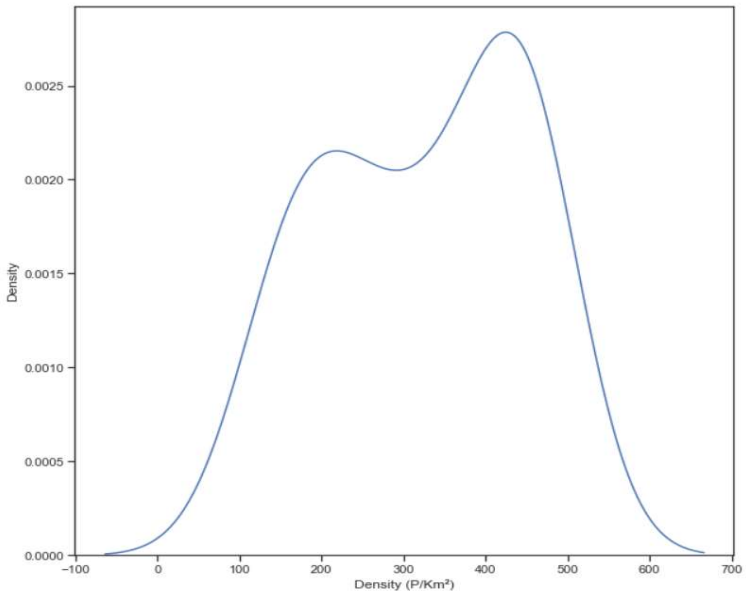
	Year	Population	Yearly % Change	Yearly Change	Migrants (net)	Median Age	Fertility Rate	Density (P/Km²)	Urban Pop %	Urban Population	Country's Share of World Pop	World Population	India Global Rank
0	2020	1380004385	0.99	13586631	-532687	28.4	2.24	464	35.0	483098640	17.70	7794798739	2
1	2019	1366417754	1.02	13775474	-532687	27.1	2.36	460	34.5	471828295	17.71	7713468100	2
2	2018	1352642280	1.04	13965495	-532687	27.1	2.36	455	34.1	460779764	17.73	7631091040	2
3	2017	1338676785	1.07	14159536	-532687	27.1	2.36	450	33.6	449963381	17.74	7547858925	2
4	2016	1324517249	1.10	14364846	-532687	27.1	2.36	445	33.2	439391699	17.75	7464022049	2

In [4]:
```
data.dtypes
```

Out[4]:
```
Year                            int64
Population                      int64
Yearly % Change               float64
Yearly Change                   int64
Migrants (net)                  int64
Median Age                    float64
Fertility Rate                float64
Density (P/Km²)                 int64
Urban Pop %                   float64
Urban Population                int64
Country's Share of World Pop  float64
World Population                int64
India Global Rank               int64
dtype: object
```

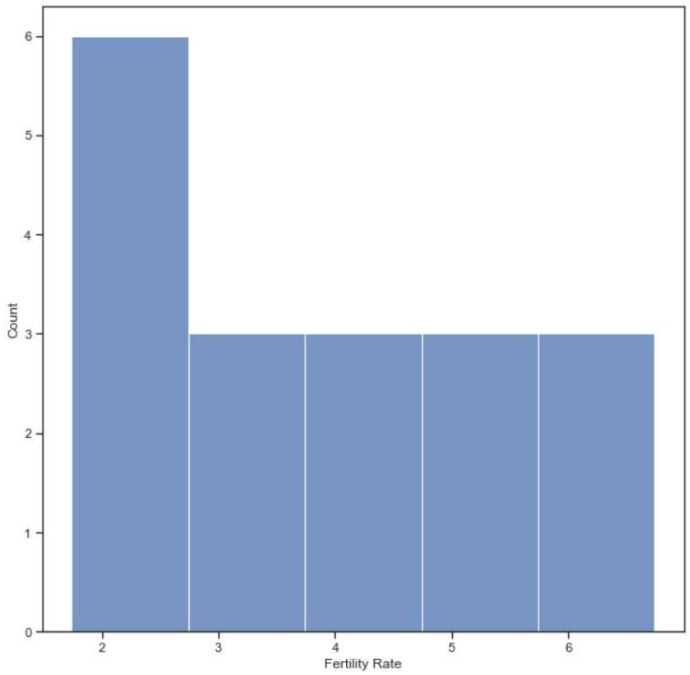
In [7]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.kdeplot(data=data, x="Density (P/Km²)")
```

Out[7]: <AxesSubplot:xlabel='Density (P/Km²)', ylabel='Density'>



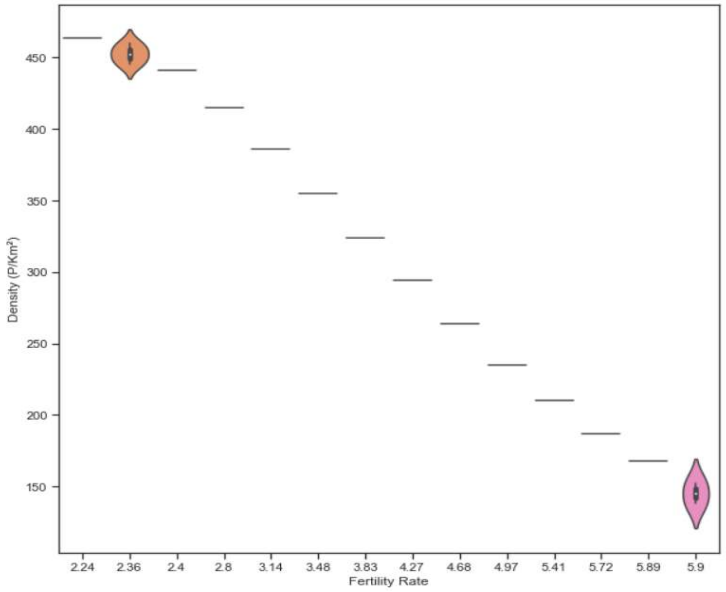
In [8]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.histplot(data['Fertility Rate'], discrete=True)
```

Out[8]: <AxesSubplot:xlabel='Fertility Rate', ylabel='Count'>



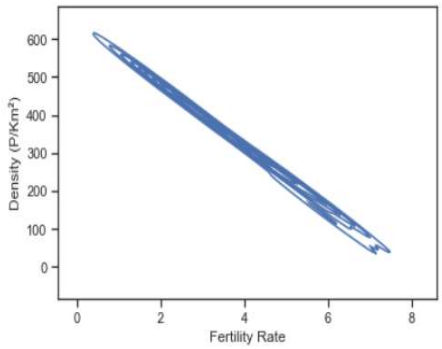
In [9]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.violinplot(x='Fertility Rate', y='Density (P/Km²)', data=data)
```

Out[9]: <AxesSubplot:xlabel='Fertility Rate', ylabel='Density (P/Km²)'>



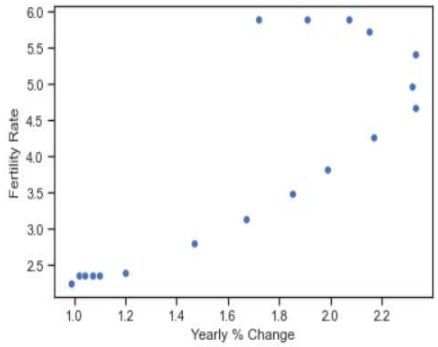
In [11]:
```
sns.kdeplot(data=data, x="Fertility Rate", y="Density (P/Km²)")
```

Out[11]: <AxesSubplot:xlabel='Fertility Rate', ylabel='Density (P/Km²)'>



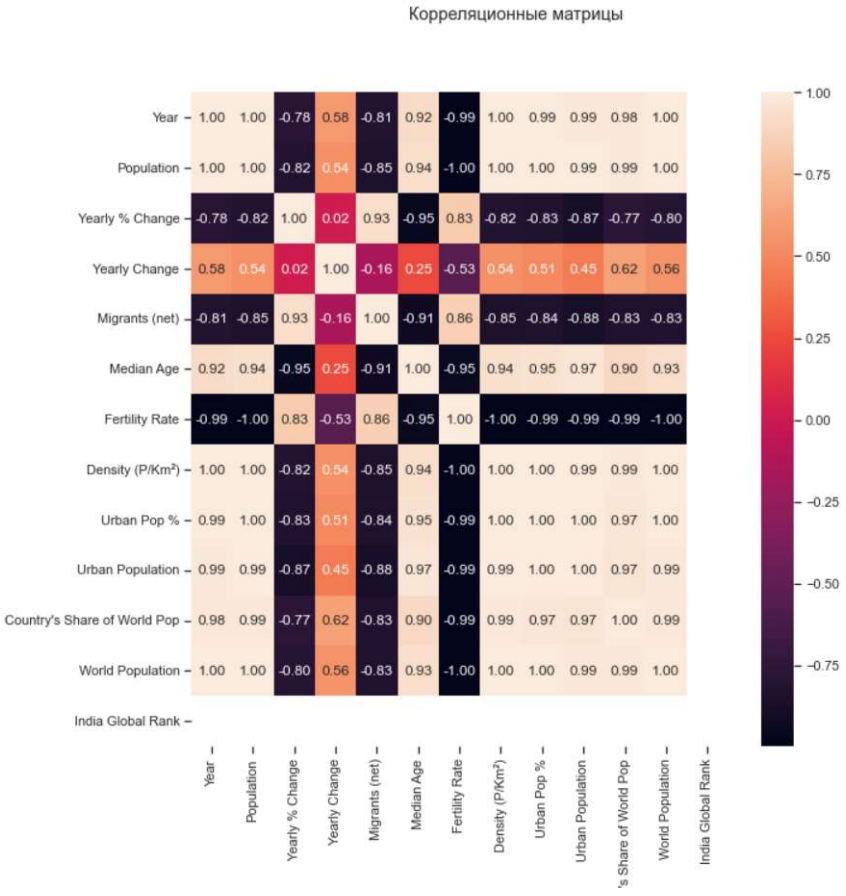
In [12]:
```
sns.scatterplot(x='Yearly % Change', y='Fertility Rate', data=data)
```

Out[12]:
```
<AxesSubplot:xlabel='Yearly % Change', ylabel='Fertility Rate'>
```

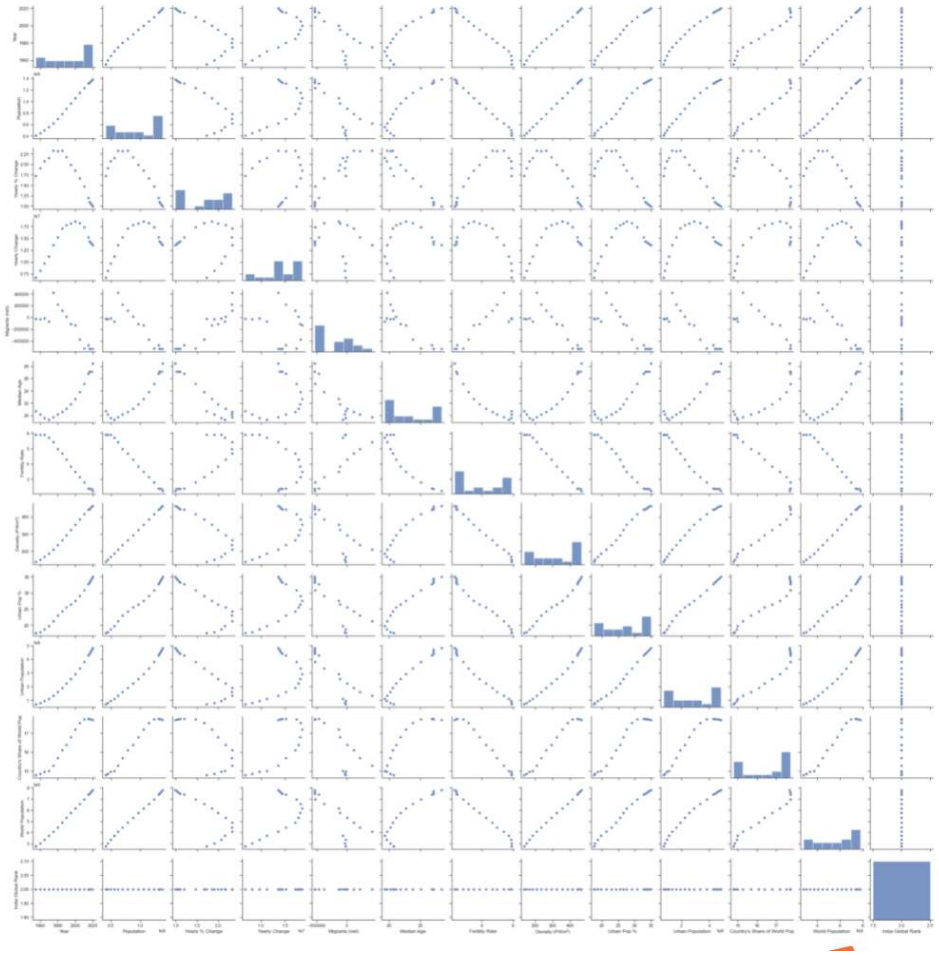


In [13]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.heatmap(data.corr(method='pearson'), annot=True, fmt='.2f')
fig.suptitle('Корреляционные матрицы')
```

Out[13]:
```
Text(0.5, 0.98, 'Корреляционные матрицы')
```



In [14]:
```
sns.pairplot(data)
plt.show()
```



Список литературы

[1] Гапанюк Ю. Е. Лабораторная работа «Разведочный анализ данных. Исследование и визуализация данных» [Электронный ресурс] // GitHub. — 2019. — Режим доступа: https://github.com/ugapanyuk/ml_course/wiki/LAB_EDA_VISUALIZATION (дата обращения: 13.02.2019)

[2] https://www.kaggle.com/datasets