Московский государственный технический университет им. Н. Э. Баумана

Курс «Технологии машинного обучения»

Отчёт по рубежному контролю №1

«Технологии разведочного анализа и обработки данных.»

Вариант № 5

Выполнил:
Каятский П. Е.
группа ИУ5-64Б

Проверил:
Гапанюк Ю.Е.

Дата: 06.04.25

Дата:

Подпись:

Подпись:

Москва, 2025 г.

Задача №1.

Для заданного набора данных проведите корреляционный анализ. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Сделайте выводы о возможности построения моделей машинного обучения и о возможном вкладе признаков в модель.

Набор данных:

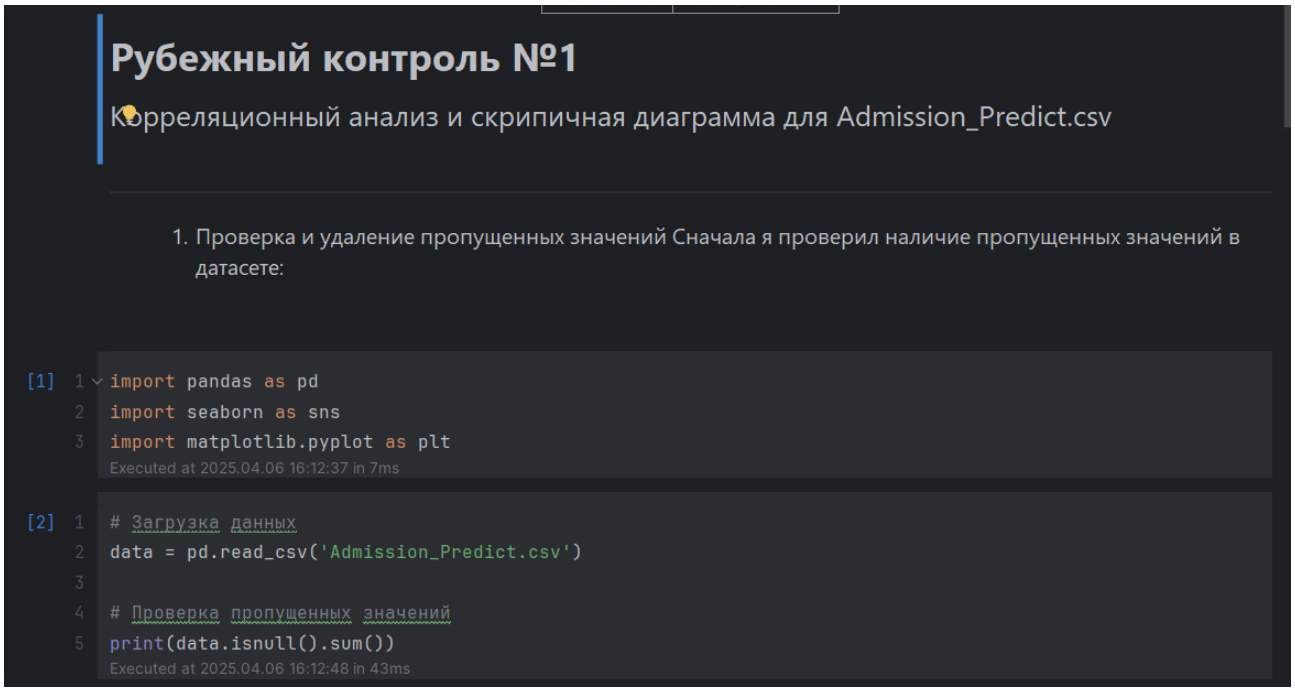
https://www.kaggle.com/mohansacharya/graduate-admissions

(файл Admission_Predict.csv)

Дополнительные требования:

Для студентов группы ИУ5-64Б, ИУ5Ц-84Б - для произвольной колонки данных построить график "Скрипичная диаграмма (violin plot)".

Ход работы:



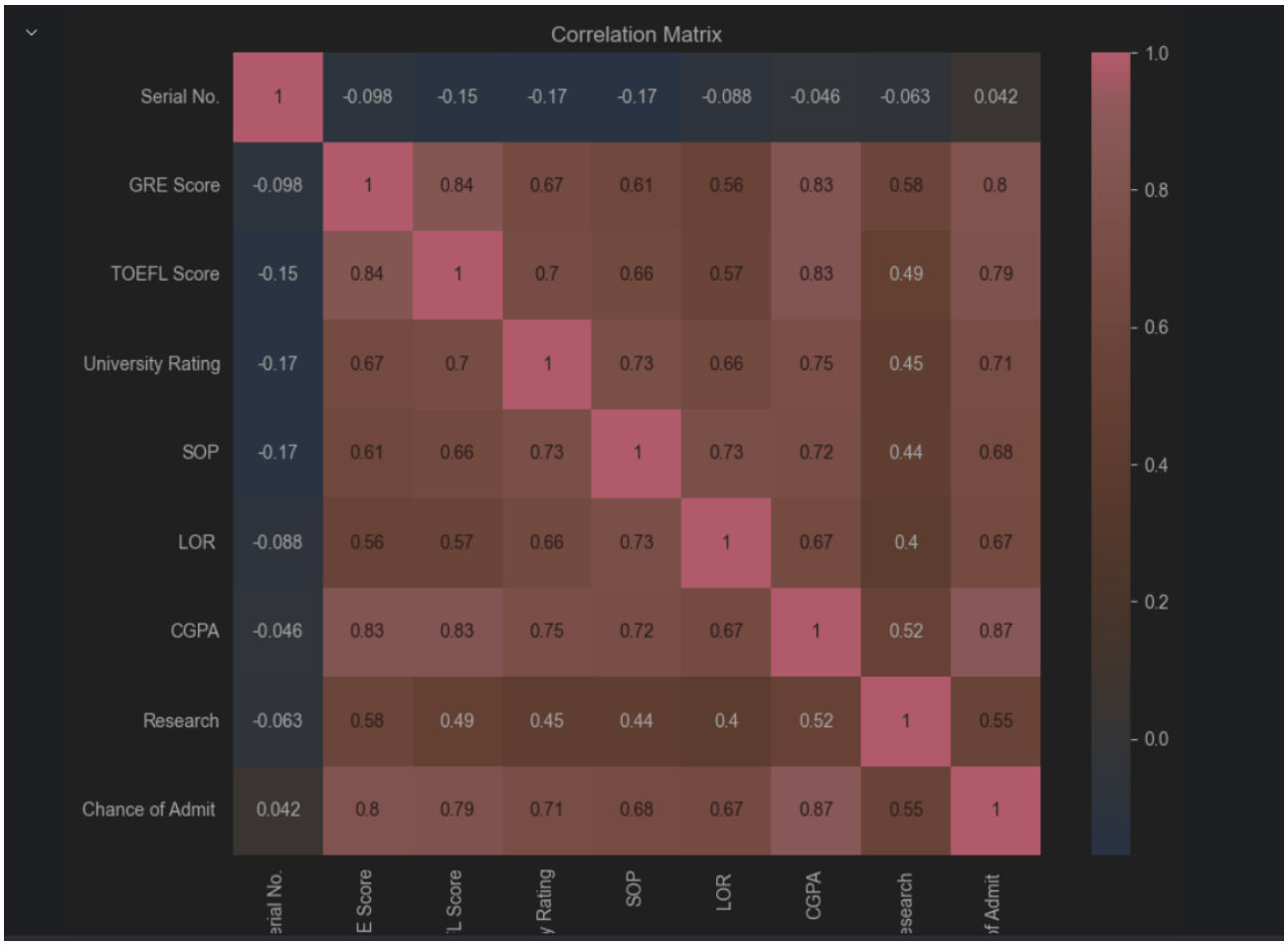
Результат показал, что в данном датасете нет пропущенных значений:

2. Корреляционный анализ Я провел корреляционный анализ, чтобы понять взаимосвязи между признаками:

```
[3] # Вычисление корреляционной матрицы
    corr_matrix = data.corr(numeric_only=True)

    # Визуализация корреляционной матрицы
    plt.figure(figsize=(10, 8))
    sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', center=0)
    plt.title('Correlation Matrix')
    plt.show()
```

Executed at 2025.04.06 16:13:16 in 949ms



Выводы из корреляционного анализа:

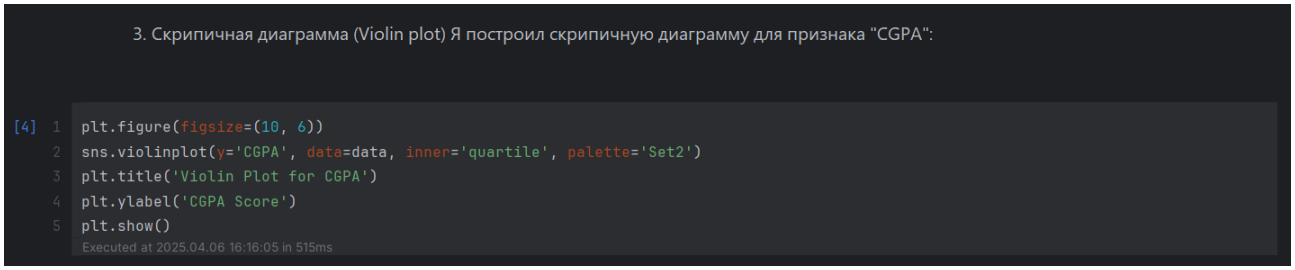
1. Наибольшую корреляцию с целевой переменной "Chance of Admit" имеют:
    - CGPA (0.87) - очень сильная положительная корреляция
    - GRE Score (0.81) - сильная положительная корреляция
    - TOEFL Score (0.79) - сильная положительная корреляция
2. Умеренная корреляция наблюдается с:

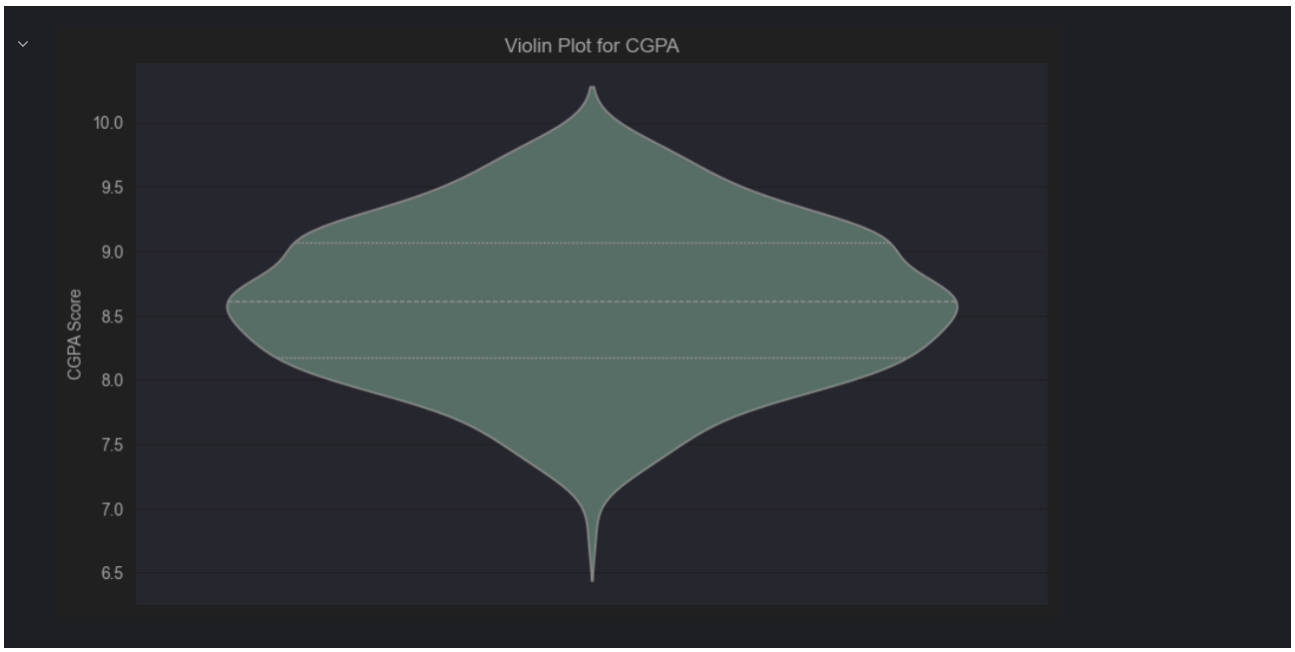
- University Rating (0.66)
- SOP (0.68)
- LOR (0.64)

3. Слабая корреляция с:
   - Research (0.55)
4. Serial No. практически не коррелирует с целевой переменной (-0.001), что логично, так как это просто порядковый номер.

Возможности построения моделей:

1. Данные хорошо подходят для построения моделей регрессии, так как целевая переменная числовая и имеет сильные корреляции с несколькими признаками.
2. Наибольший вклад в модель будут вносить CGPA, GRE Score и TOEFL Score.
3. Serial No. можно исключить из модели, так как это не информативный признак.
4. Имеющиеся корреляции позволяют ожидать хорошее качество прогнозирования.





Анализ скрипичной диаграммы для CGPA:

1. Распределение близко к нормальному с небольшим правосторонним смещением.
2. Основная масса данных сосредоточена между 7.8 и 9.2.
3. Медиана находится около 8.6.
4. Есть несколько выбросов в нижней части распределения.
5. Плотность распределения максимальна около 8.5-9.0.

Этот анализ подтверждает, что CGPA - важный признак с хорошим разбросом значений, что делает его полезным для прогнозирования.