Міністерство освіти і науки України

Національний університет «Львівська політехніка»

Інститут комп’ютерних наук та інформаційних технологій

Кафедра автоматизованих систем управління



**Звіт**

до розрахунково-графічної роботи

з дисципліни

Основи смарттехнологій і систем

Виконав:

студент групи ОІ-31

Шпак Артем

Прийняла:

Нарушинська Ольга

Львів – 2024

**Етап розуміння бізнесу**

Для своєї роботи я обрав датасет який описує якість сну людини відносно його стилю життя.

**Основна бізнес-мета**:

Навчитись передбачати якість сну, відповідно до кількості годин витрачених на навчання, спорт, кількість кофеїну, часу пробудження та часу, коли людина лягає.

**Основна мета моделювання**:

Розроблення моделі, яка буде передбачати якість сну людини використовуючи її дані того, як вона провела день(кількість годин витрачених на навчання, спорт, кількість кофеїну, час пробудження та час, коли людина лягає.)

**Критерії успіху бізнесу:**

1. Покращення якості сну людини.
2. Розуміння того, який тип активності найбільше впливає на якість сну людини.
3. Використання моделі для надання індивідуальних рекомендацій для покращення якості сну, що сприяє підвищенню загального рівня задоволеності користувачів сервісу на 15-20%.

**Критерії успіху моделювання:**

1. Точність передбачень: Модель повинна досягати RMSE < 1.5 або MSE зниження на 20% у порівнянні з базовою лінією.
2. Інтерпретованість моделі: Модель має дозволяти аналіз впливу ключових факторів (кофеїн, фізична активність, екранний час) на якість сну.
3. Збалансованість результатів: Модель повинна показувати стабільні результати для різних груп студентів за віком, статтю та курсом навчання.

**Етап розуміння даних**

Мій датасет показує залежність якості сну від кількості часу витраченого за телефоном, навчанням, заняттям спортом, загалом за типом життя.

**Опис ознак:**

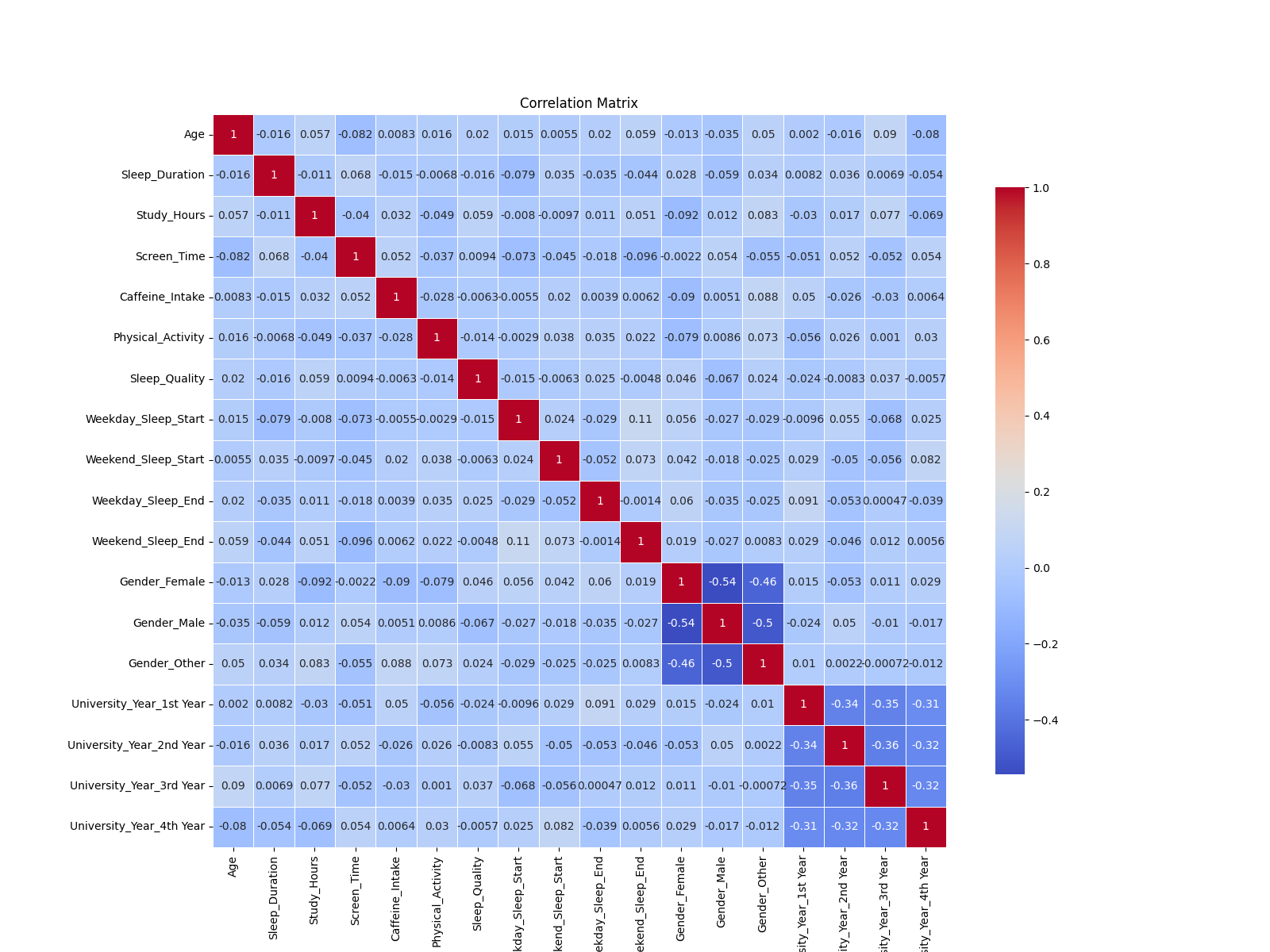
Student\_ID: Unique identifier for each student (integer).  
Age: Age of the student in years (integer).  
Gender: Gender of the student (categorical: 'Male', 'Female', 'Other').  
University\_Year: The student's current year of university (categorical: '1st Year', '2nd Year', '3rd Year', '4th Year').  
Sleep\_Duration: Total hours of sleep per night (float).  
Study\_Hours: Average number of hours spent studying per day (float).  
Screen\_Time: Average number of hours spent on screens (excluding studying) per day (float).  
Caffeine\_Intake: Average number of caffeinated beverages consumed per day (integer).  
Physical\_Activity: Average minutes spent on physical activity per day (integer).  
Sleep\_Quality: Subjective rating of sleep quality on a scale of 1 to 10 (1 being the worst, 10 being the best) (integer).  
Weekday\_Sleep\_Start: Time the student typically goes to sleep on weekdays (float, in 24-hour format, e.g., 23.00 is 11 PM).  
Weekend\_Sleep\_Start: Time the student typically goes to sleep on weekends (float, in 24-hour format).  
Weekday\_Sleep\_End: Time the student typically wakes up on weekdays (float, in 24-hour format).  
Weekend\_Sleep\_End: Time the student typically wakes up on weekends (float, in 24-hour format).

RangeIndex: 500 entries, 0 to 499

Data columns (total 14 columns)

Мій датасет містить 14 стовпців з характеристиками та 500 рядків з об’єктами. Як можна бачити по результатами інформації програми, пропущених значень ми не маємо, отже можна використати всі дані.

Цільовою змінною є Sleep\_Quality. Мої характеристики мають такі типи даних: цілі числа(Вік, ID, якість сну, тривалість сну, тривалість занять спортом та інші), маю категоріальні дані(рік навчання та гендер).



Кореляційна матриця показує:

1. Поведінкові фактори сну:

Sleep\_Duration, Sleep\_Quality, Weekday\_Sleep\_Start, Weekend\_Sleep\_Start, Weekday\_Sleep\_End, Weekend\_Sleep\_End

Ці змінні показують слабкі кореляції між собою, що вказує на зв'язок часу початку і закінчення сну з його тривалістю і якістю. Наприклад, фізична активність та ранній час сну трохи покращують якість і тривалість сну.

2. Стать:

Gender\_Female, Gender\_Male, Gender\_Other

Статеві змінні мають слабкі кореляції з деякими поведінковими факторами, як-от "Screen\_Time" та "Caffeine\_Intake". Це може відображати різні звички між статями щодо використання екранного часу та вживання кофеїну.

3. Навчальні фактори та рік навчання:

Study\_Hours, University\_Year\_1st Year, University\_Year\_2nd Year, University\_Year\_3rd Year, University\_Year\_4th Year

Роки навчання мають слабкі негативні кореляції між собою та зі змінною "Study\_Hours". Студенти на різних курсах демонструють відмінності в навчальних звичках: старші курси можуть мати менше навчальних годин.

4. Здоров'я та активність:

Physical\_Activity, Caffeine\_Intake

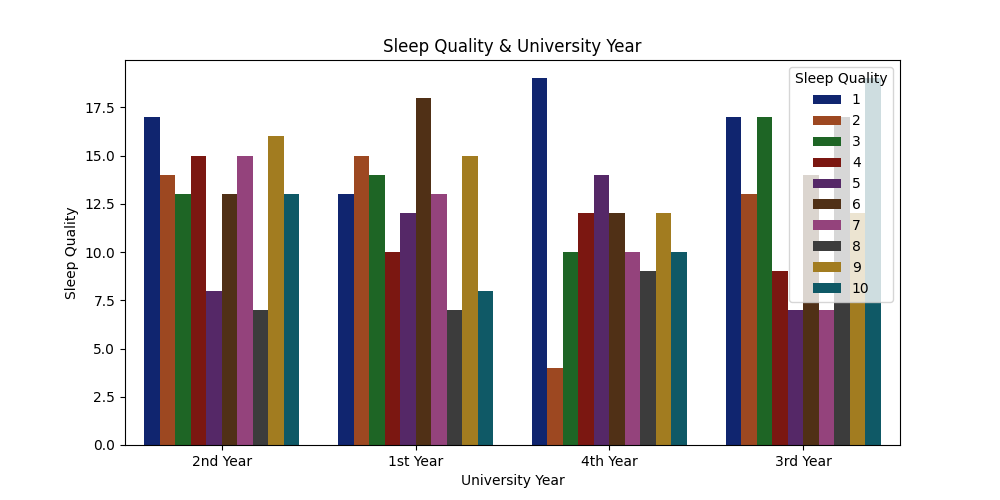
Фізична активність і вживання кофеїну мають слабкі кореляції з якістю сну і тривалістю навчання. Наприклад, активніші люди можуть споживати менше кофеїну і мати кращу якість сну.

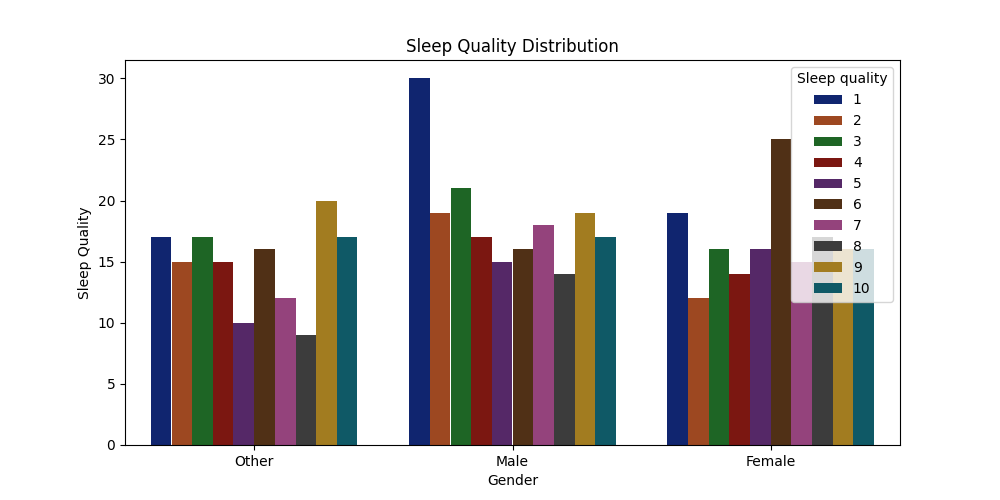
5. Екранний час:

Screen\_Time

Екранний час має слабкі негативні кореляції з "Study\_Hours" і "Physical\_Activity", що може означати, що більше часу перед екраном зменшує час на навчання та фізичну активність.Загалом, графік демонструє, що концентрація PM2.5 значною мірою залежить від сезону, з високими значеннями взимку і відносно низькими влітку, що є важливим фактором для врахування при прогнозуванні та аналізі якості повітря.

Переглянемо деякі залежності та їх графіки:

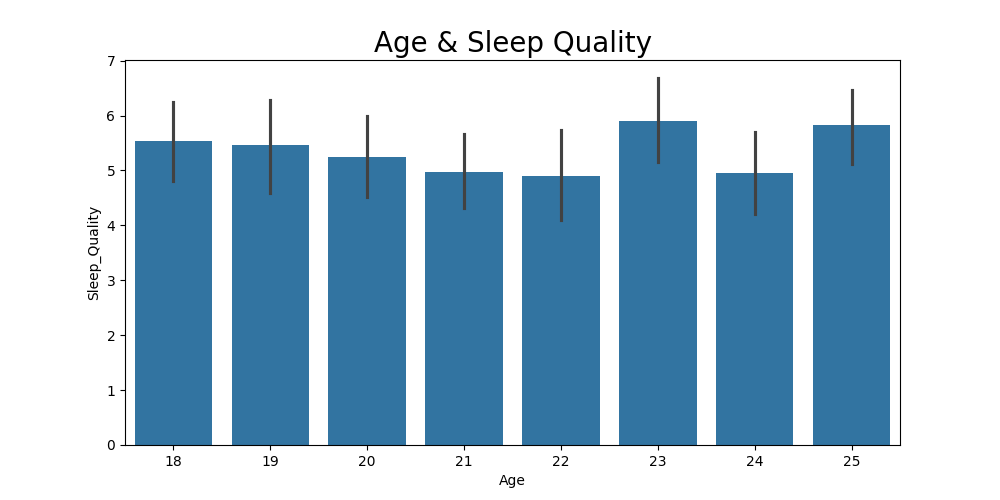
Переглянемо графік між якістю сну та курсом навчання.

 Як бачимо всі курси студентів мають рівну якість сну. Так само як і на графіку зі статтю людини та якості сну.

Так само переглядаючи графік якості сну відносно віку людини, можемо зробити такий самий висновок. Скоріш за все, такий баланс даних є через те, що це є синтетичні дані і саме через це, результати навчання моделі може мати дивні результати.

**Етап підготовки даних (Feature engineering)**

Перевірка датасету на пропущені комірки.



**Етап підготовки даних (Feature engineering)**

Перевірка датасету на пропущені комірки.

RangeIndex: 500 entries, 0 to 499

Data columns (total 14 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Student\_ID 500 non-null int64

1 Age 500 non-null int64

2 Gender 500 non-null object

3 University\_Year 500 non-null object

4 Sleep\_Duration 500 non-null float64

5 Study\_Hours 500 non-null float64

6 Screen\_Time 500 non-null float64

7 Caffeine\_Intake 500 non-null int64

8 Physical\_Activity 500 non-null int64

9 Sleep\_Quality 500 non-null int64

10 Weekday\_Sleep\_Start 500 non-null float64

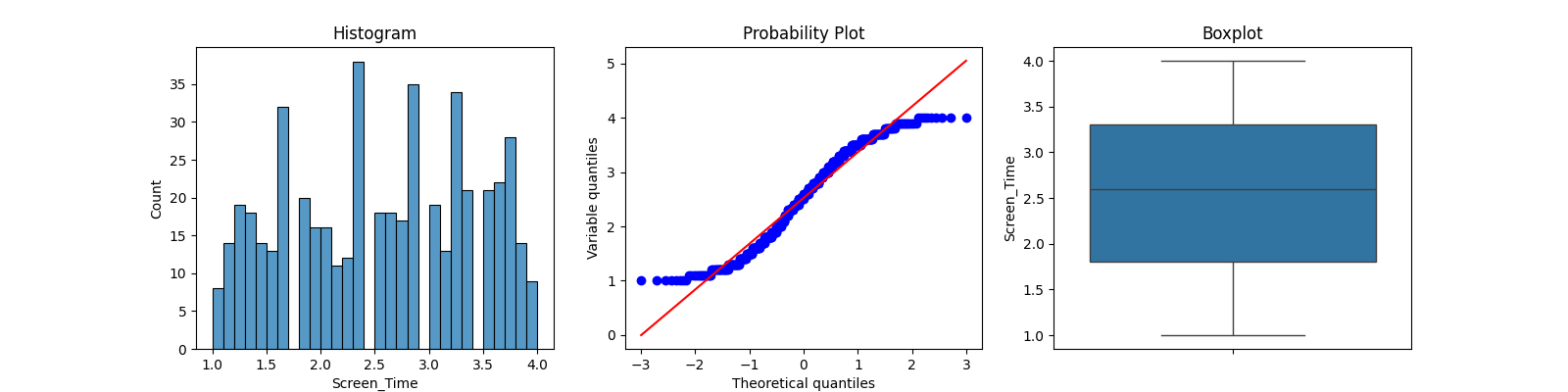
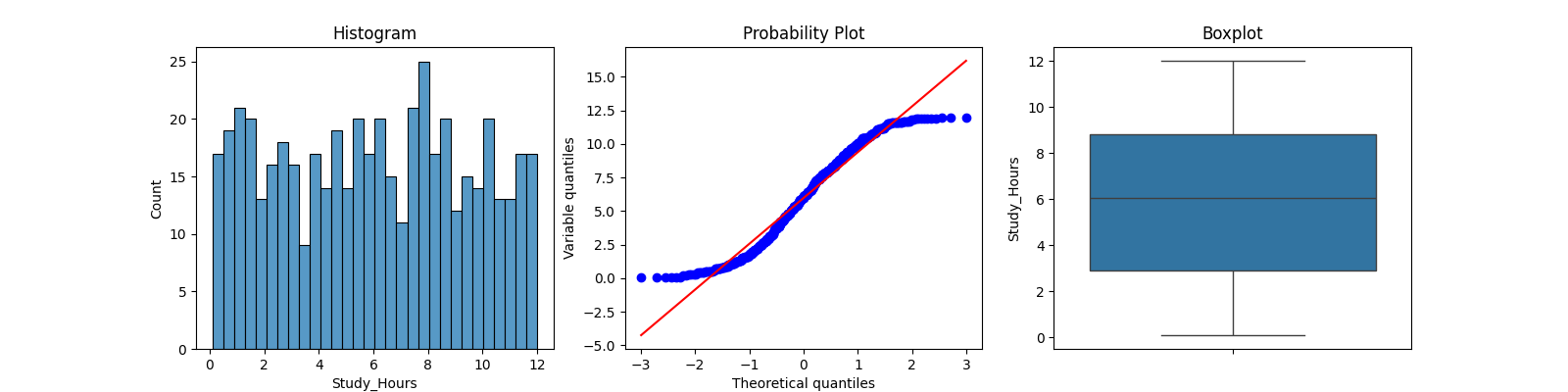
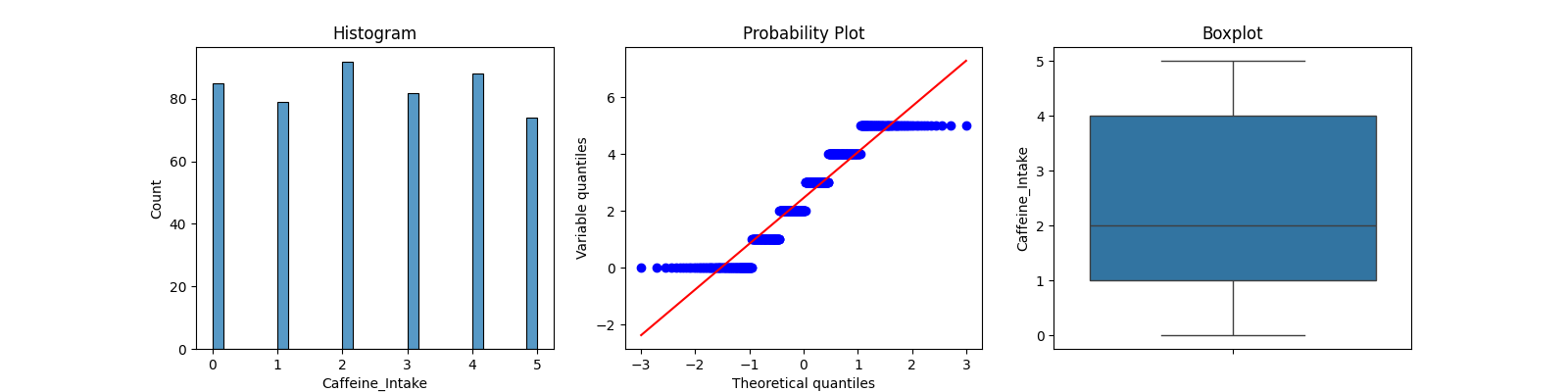
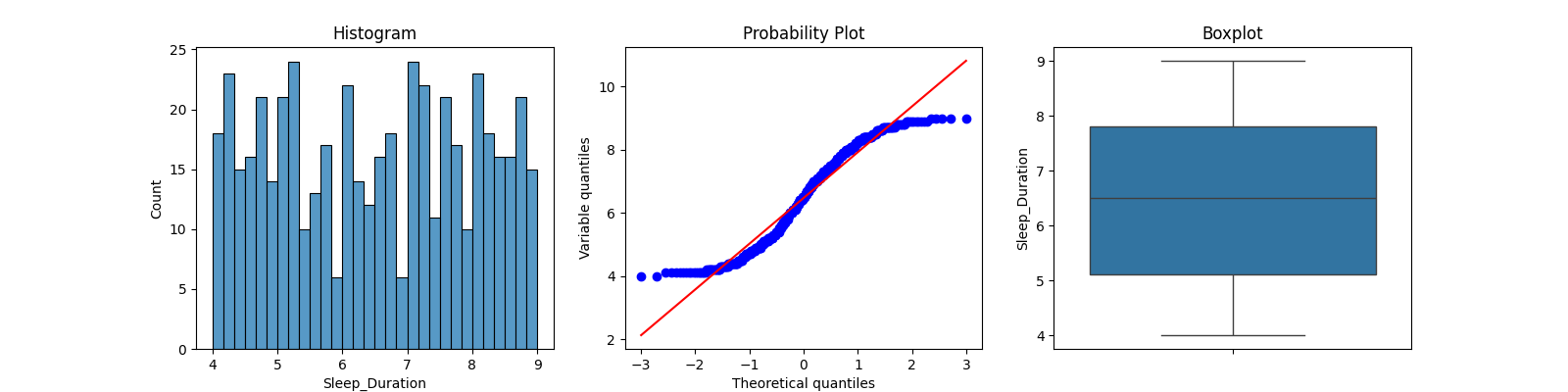
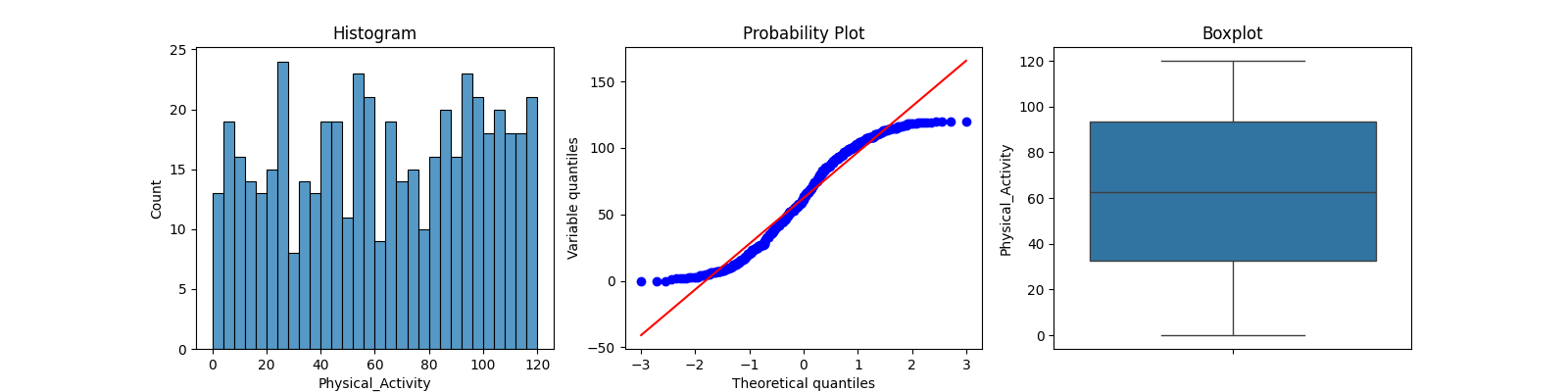
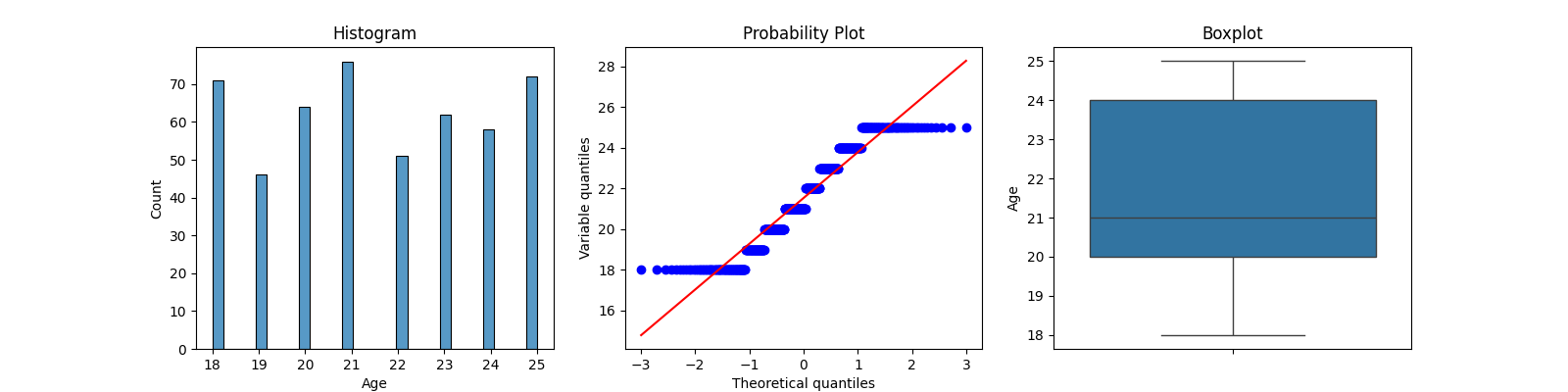
11 Weekend\_Sleep\_Start 500 non-null float64

12 Weekday\_Sleep\_End 500 non-null float64

13 Weekend\_Sleep\_End 500 non-null float64

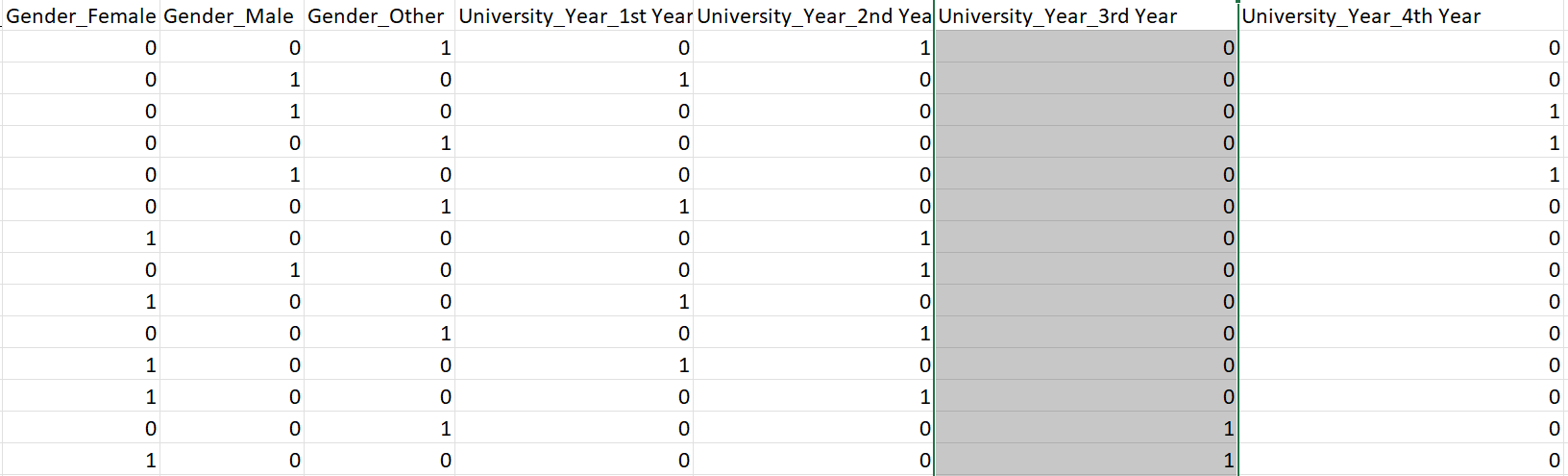
dtypes: float64(7), int64(5), object(2)

Пропущених даних не виявлено, тому перейдемо до перевірки на аномалії в колонках:

****

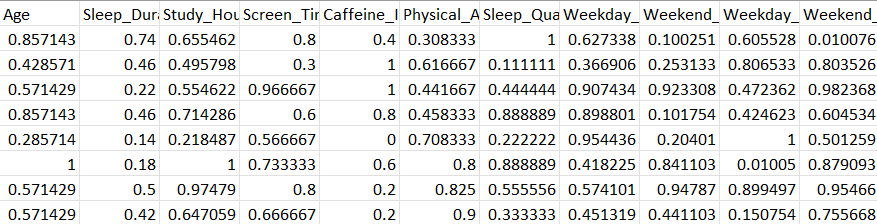
Переглянувши основні колонки, в яких можуть бути аномалії, прийнято рішення що аномалії відсутні і ми можемо переходити до кодування наших категоріальних даних.

Я використовую метод One Hot Encoding для кодування категоріальних даних(використання бінарних значень для колонок).

Після кодування перейдемо до маштабування.

Маштабування

Я використовую метод MinMaxScaller для стандартизації даних (приведення даних в діапазон від 0 до 1). Це допоможе зробити ознаки однаково важливими для моделей, які чутливі до різних масштабів ознак.

**Етап моделювання**

Для вибору найкращого методу моделювання спочатку розділяю датасет тренувальний та тестувальний (80 на 20). Для завдання регресії з прогнозування рівня забруднення повітря (концентрації PM2.5) можна розглянути кілька моделей: LinearRegression, Lasso, Ridge, XGBRegressor,

DecisionTreeRegressor, SVR, KNeighborsRegressor, RandomForestRegressor.

Результати тренування моделей:

training LinearRegression

MAE: 2.776976697258746

MAPE: 1.2230961905239337

MSE: 9.963897104933114

R2 -0.04513479744620241

------------------------------------------------------------------------

training Lasso

MAE: 2.745771346028695

MAPE: 1.2231728462264089

MSE: 9.660959433135451

R2 -0.01335900741959506

------------------------------------------------------------------------

training Ridge

MAE: 2.776908711598151

MAPE: 1.2230742203080522

MSE: 9.963024982089012

R2 -0.04504331858783783

------------------------------------------------------------------------

training XGBRegressor

MAE: 3.104977397918701

MAPE: 1.4471627482033909

MSE: 14.004942768406961

R2 -0.46900880336761475

------------------------------------------------------------------------

training DecisionTreeRegressor

MAE: 3.59

MAPE: 1.4027936507936507

MSE: 18.63

R2 -0.9541411429050934

------------------------------------------------------------------------

training SVR

MAE: 2.77297795621531

MAPE: 1.2522319563567088

MSE: 9.892609399808194

R2 -0.037657275300851056

------------------------------------------------------------------------

training KNeighborsRegressor

MAE: 2.94

MAPE: 1.3103690476190477

MSE: 12.098400000000002

R2 -0.26902743979189414

------------------------------------------------------------------------

training RandomForestRegressor

MAE: 2.8608

MAPE: 1.2907848412698413

MSE: 11.170624

R2 -0.17171100109087867

На основі отриманих результатів можна зробити такі висновки щодо ефективності різних моделей для прогнозування якості сну:

**LinearRegression** та **Ridge** мають схожі результати з MAE (2.7769), MAPE (1.223), MSE (9.96), та низьким значенням R² (-0.045), що свідчить про їхню обмежену здатність описувати дані. Ці моделі не дають задовільного результату, оскільки значення R² близьке до нуля або навіть негативне, що вказує на те, що вони не можуть пояснити значну частину варіації в даних.

**Lasso** демонструє трохи кращий результат у порівнянні з LinearRegression та Ridge, з меншою MAE (2.7458) та MSE (9.66) і R² (-0.013). Хоча різниця незначна, це свідчить про невелику перевагу Lasso у цьому конкретному завданні. Однак, значення R² все одно негативне, що означає недостатню якість моделі.

**XGBRegressor** демонструє гірші показники, з MAE (3.105), MAPE (1.447) та R² (-0.469). Це вказує на те, що модель значно переоцінює чи недооцінює прогнози, і вона не є ефективною для цього набору даних.

**DecisionTreeRegressor** має найгірші результати серед усіх моделей, з MAE (3.59), MSE (18.63), і дуже низьким R² (-0.954), що свідчить про низьку стабільність та точність у прогнозах. Це може бути результатом перенавчання або нечутливості дерева до особливостей даних.

**SVR** показує схожий результат з LinearRegression та Ridge, з MAE (2.773), MAPE (1.252), MSE (9.89), і трохи кращим R² (-0.037), що однак незначно покращує загальну картину.

**KNeighborsRegressor** також не демонструє хороших результатів, з MAE (2.94), MAPE (1.310), і негативним R² (-0.269). Ця модель не може точно моделювати залежність, ймовірно, через складність або нерівномірність даних.

**RandomForestRegressor** показує дещо кращий результат порівняно з іншими моделями з MAE (2.861), MAPE (1.291), MSE (11.17), та R² (-0.171). Хоча RandomForest має дещо кращі показники точності, R² все одно залишається негативним.

Я проаналізував результати і обрав метод Lasso для навчання.

Для **валідації моделі** використано метод **KSplit** з 5 ітераціями та 20% даних для тестування на кожному спліті, а також ввмікну тасування.

Результати оцінки моделі на п'яти сплітах з використанням метрик **Lasso** демонструють хорошу стабільність і точність моделі для прогнозування значень.

Lasso Regression:

MSE: 9.660959433135451

MAE: 2.745771346028695

R² Score: -0.01335900741959506

RMSE: 3.108208396027437

----------------------------------------------------------------------

k-fold set metrics:

Mean Squared Error (MSE): 8.728769823350241

Mean Absolute Error (MAE): 2.6077402728045227

R² Score: -0.020610327196754152

k-fold set metrics:

Mean Squared Error (MSE): 7.678930702747491

Mean Absolute Error (MAE): 2.356794931158394

R² Score: -0.023942755595965037

k-fold set metrics:

Mean Squared Error (MSE): 8.887744140625

Mean Absolute Error (MAE): 2.60453125

R² Score: -0.0015064881329669433

k-fold set metrics:

Mean Squared Error (MSE): 9.108393936558054

Mean Absolute Error (MAE): 2.6605255895775146

R² Score: -0.0009224106107750263

k-fold set metrics:

Mean Squared Error (MSE): 8.72136103993636

Mean Absolute Error (MAE): 2.5504727443147344

R² Score: -0.008997101458679602

Test set metrics:

Mean Squared Error (MSE): 9.657856250000002

Mean Absolute Error (MAE): 2.7432499999999997

R² Score: -0.013033507803977606

**Основна оцінка моделі:**

* MSE на тестовому наборі становить 9.66, MAE — 2.74, R² — -0.013. Це означає, що модель неадекватно пояснює варіативність даних (R² близький до нуля або негативний), і, відповідно, її точність залишається низькою.

**Оцінка моделі з використанням k-fold cross-validation:**

* Показники MSE варіюються в діапазоні від 7.68 до 9.11, а MAE — від 2.36 до 2.66. R² також залишався негативним у всіх фолдах, що свідчить про нестабільність та обмежену ефективність моделі на різних розбиттях даних.

**Загальний висновок:**

* Результати свідчать про те, що Lasso Regression не дає якісного прогнозування для цього набору даних. Негативні та близькі до нуля значення R² у тестовому наборі та k-fold cross-validation вказують на те, що модель не може належним чином описати залежність у даних.

**Оптимізація моделі**

Для налаштування гіперпараметрів я обирала **GridSearch** (Випадковий

вибір комбінацій гіперпараметрів)

### 

### Пояснення параметрів:

**tol = 0.0001:** Точність для зупинки оптимізації. Чим менше значення, тим більше ітерацій для збіжності.

**max\_iter = 1000:** Максимальна кількість ітерацій. Дає більше часу для збіжності.

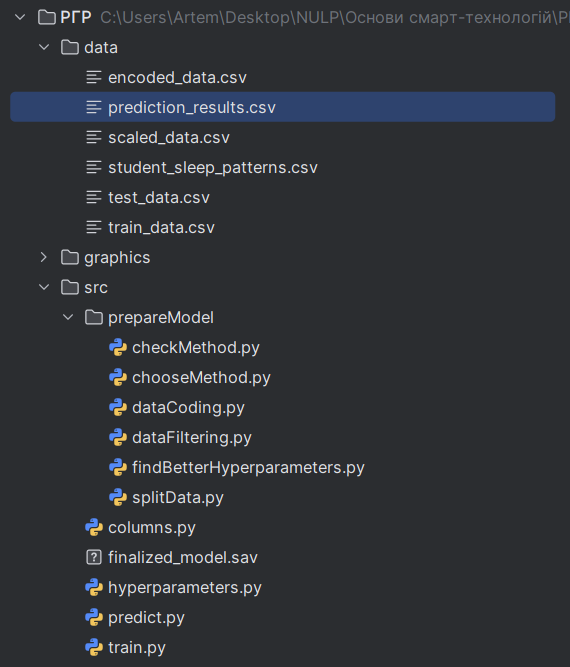
**alpha = 3:** Коефіцієнт регуляризації. Зменшує перенавчання, але може знизити точність при завеликому значенні.

Результат:



**Розгортання моделі**

**Структура проекту**

****

Для запуску цього проєкту знадобляться наступні пакети:

python 3.12  
  
numpy==2.1.1  
pandas==2.2.2  
sklearn==1.5.2

Matplotlib

Інструкція з використання

1. Копіюйте або завантажте репозиторій проєкту.

2. Переконайтесь, що всі необхідні бібліотеки встановлені.

3. Для навчання моделі виконайте файл train.py.

3. Далі для передбачення результатів на нових даних:

   predict.py

4. Результати передбачення будуть збережені у папці data/prediction\_results.csv.

**Висновок:**

Модель **Lasso Regression** продемонструвала обмежену ефективність у прогнозуванні якості сну на основі наданих даних. Основні метрики, такі як середня квадратична похибка (MSE), абсолютна похибка (MAE) та коефіцієнт детермінації (R²), вказують на те, що модель не здатна точно відображати залежність між змінними. Зокрема, середня квадратична похибка коливається від 7.68 до 9.66, що є досить високим показником для регресійної моделі, яка мала б демонструвати вищу точність прогнозів. Такий рівень похибки свідчить про те, що модель не може належним чином врахувати складність зв’язків між ознаками.

Абсолютна похибка (MAE) на рівні близько 2.7 також показує, що модель відхиляється від фактичних значень, що може призводити до неточних прогнозів якості сну. Особливу увагу слід звернути на коефіцієнт R², який залишається негативним навіть після застосування крос-валідації (k-fold cross-validation), що вказує на нездатність моделі пояснити варіативність у даних. Це означає, що модель Lasso Regression не тільки не описує змінну якість сну, а й може давати менш точні прогнози, ніж проста середня, оскільки негативний R² свідчить про те, що модель не змогла знайти адекватну залежність у вибірці.

Такі результати можуть бути наслідком невідповідності лінійної моделі з L1-регуляризацією для цього конкретного завдання або можливих проблем з масштабом та структурою даних, які Lasso Regression не може захопити. Можливо, набір даних містить складні або нелінійні взаємозв’язки між змінними, що може вимагати застосування більш гнучких алгоритмів або ретельної підготовки даних, наприклад, створення додаткових ознак чи відбору інформативних змінних.