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

Рубежный контроль №1
по дисциплине
«Методы машинного обучения»
на тему

«Методы обработки данных.»

Выполнил:
студент группы ИУ5и-22М
Лун Сыхань

Москва — 2024 г.

Варианты заданий

Номер варианта	Номер задачи №1	Номер задачи №2
18	18	38

Задача №18.

Для набора данных проведите масштабирование данных для одного (произвольного) числового признака на основе Z-оценки.

Задача №38.

Для набора данных проведите процедуру отбора признаков (feature selection). Используйте класс SelectKBest для 10 лучших признаков, и метод, основанный на взаимной информации.

Для студентов групп ИУ5-22М и ИУ5И-22М - для произвольной колонки данных построить гистограмму.

Задача №18

```
import pandas as pd

from sklearn.preprocessing import StandardScaler

import matplotlib.pyplot as plt


data = pd.read_csv(r'C:\Users\Loong\Desktop\MCU Movies.csv')

numeric_feature = "Runtime (min)"

scaler = StandardScaler()

data[numeric_feature + "_scaled"] = scaler.fit_transform(data[[numeric_feature]])

print(data[[numeric_feature, numeric_feature + "_scaled"]].head())

plt.hist(data[numeric_feature + "_scaled"], bins=20, color='skyblue',
edgecolor='black')

plt.xlabel(numeric_feature + " (Scaled)")

plt.ylabel('Frequency')

plt.title('Histogram of Scaled ' + numeric_feature)

plt.show()
```

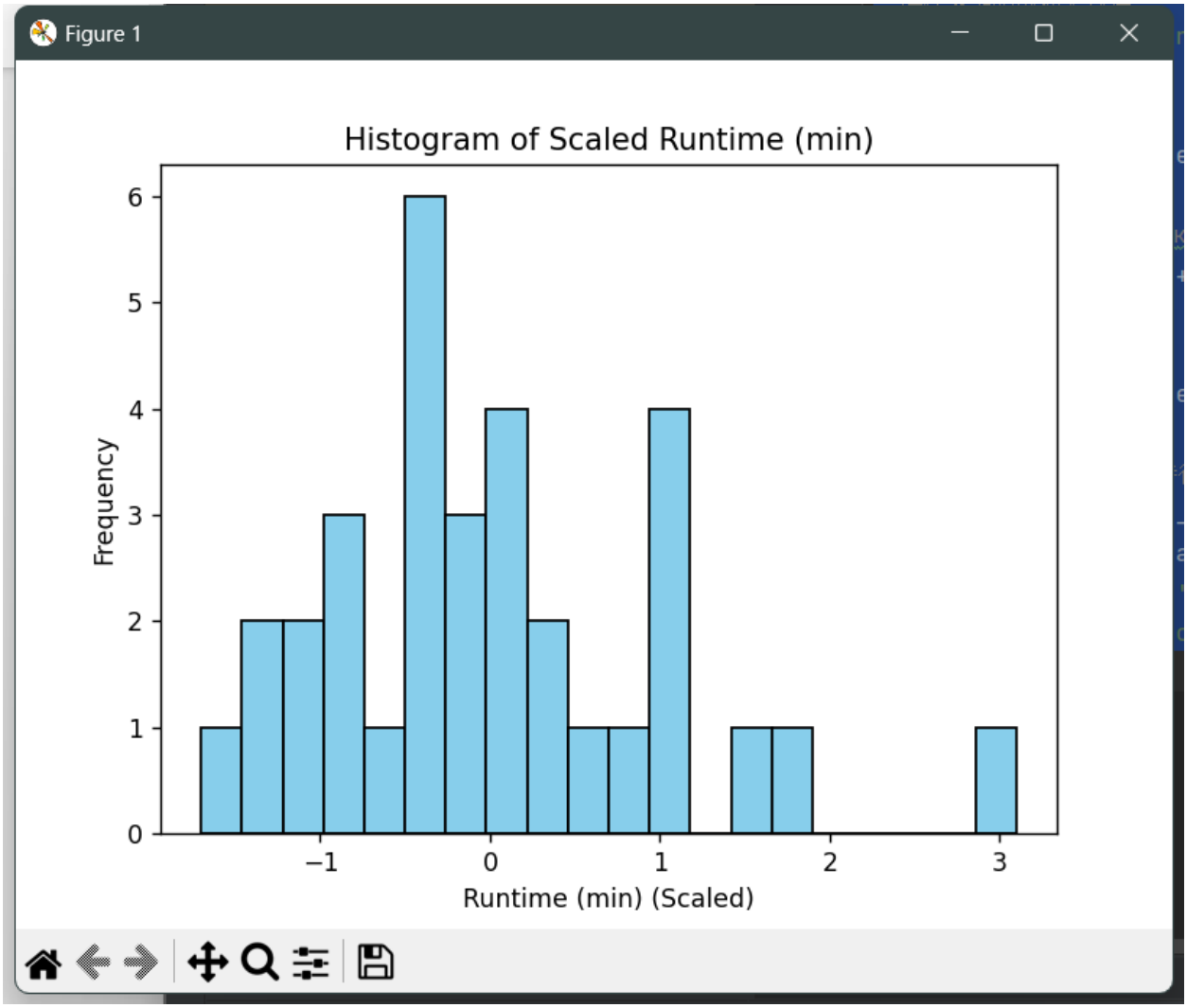


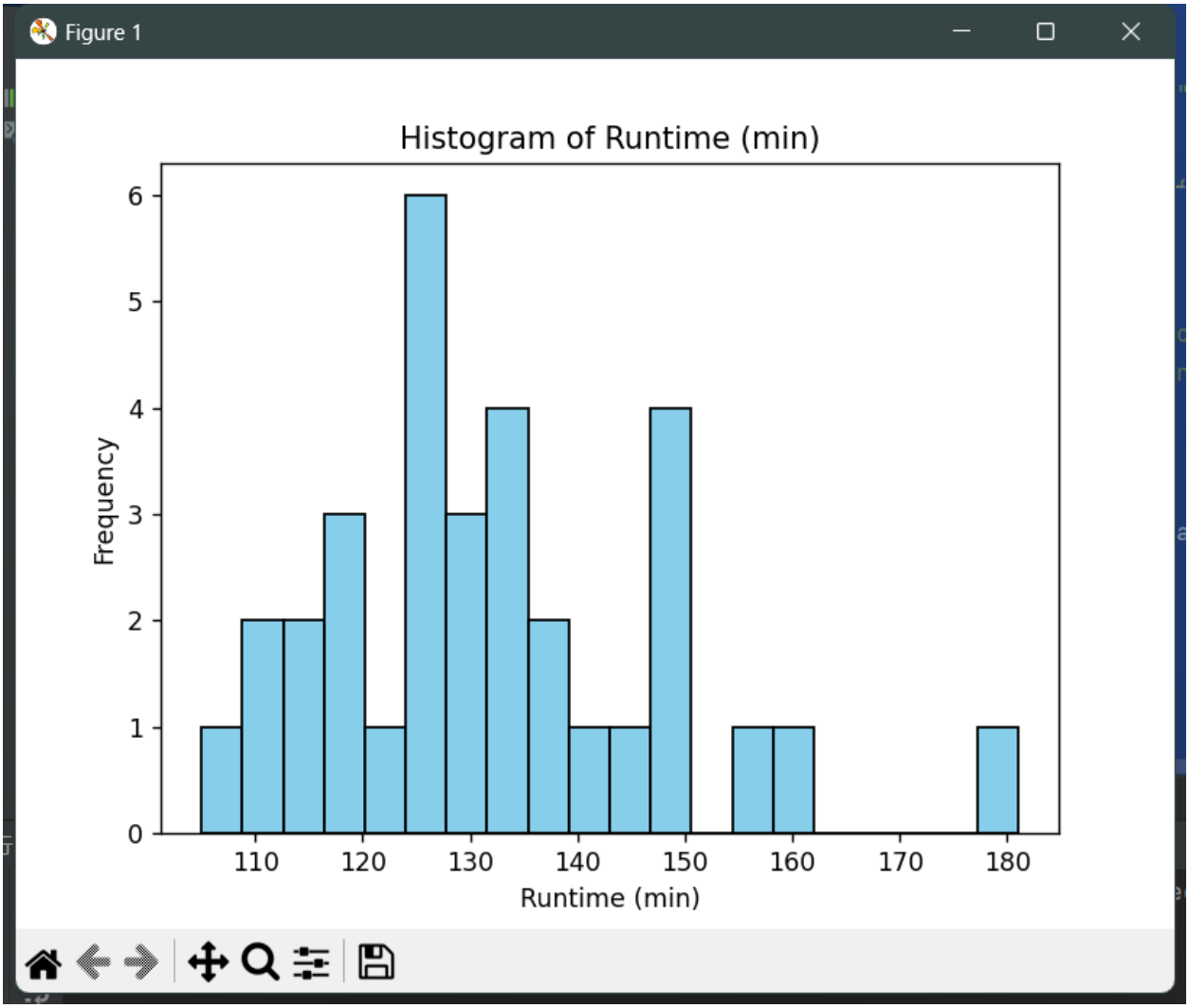
Figure 1
Histogram of Scaled Runtime (min)
Frequency
Runtime (min) (Scaled)
0
1
2
3
4
5
6
−1
0
1
2
3

Задача №38

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.feature_selection import SelectKBest, mutual_info_regression
import matplotlib.pyplot as plt


data = pd.read_csv(r'C:\Users\Loong\Desktop\MCU Movies.csv')
currency_columns = ["Production Budget", "Box Office (Local)", "Box Office (International)", "Total Box Office Earnings"]
for col in currency_columns:
    data[col] = data[col].replace('[\$,]', '', regex=True).astype(float)
X = data.drop(["Movie Title", "Release Date (USA)", "Phase", "Genre", "Movie Rating", "Lead Role",
               "Production Budget", "Box Office (Local)", "Box Office (International)",
               "Total Box Office Earnings", "Rotten Tomatoes Ratings (%)", "IMDb Ratings"], axis=1)
y = data["Total Box Office Earnings"]
selector = SelectKBest(score_func=mutual_info_regression, k=min(10, X.shape[1]))
X_selected = selector.fit_transform(X, y)
selected_features = X.columns[selector.get_support(indices=True)]
print(selected_features)
for feature in selected_features:
    plt.hist(data[feature], bins=20, color='skyblue', edgecolor='black')
    plt.xlabel(feature)
    plt.ylabel('Frequency')
```

```
plt.title('Histogram of ' + feature)
plt.show()
```



Список литературы

[1] Гапанюк Ю. Е. LAB_MMO__DATA_STORYЛабораторная работа №1Создание "истории о данных" (Data Storytelling)// GitHub. — 2024. — Режим доступа:https://github.com/ugapanyuk/courses_current/wiki/MMO_RK_1