Министерство науки и высшего образования Российской Федерации Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)» (МГТУ им. Н.Э. Баумана)

Факультет «Информатика и системы управления»

Кафедра ИУ5 «Системы обработки информации и управления»

Курс «Технологии машинного обучения»

Отчёт по рубежному контролю №1

«Технологии разведочного анализа и обработки данных»

Вариант №1

Выполнил:

студент группы ИУ5-63Б

Абрамов В. Г.

Преподаватель:

Гапанюк Ю. Е.

2023 г.

Выполнение работы

Для выполнения задачи проведения корреляционного анализа данных был представлен набор данных sklearn iris dataset

```
[21]
import numpy as np
import pandas as pd
import itertools
import matplotlib.pyplot as plt
import seaborn as sns
import sklearn
✓ 0.1s

[22]
from sklearn.datasets import load_iris
iris = load_iris()
✓ 0.0s

[23]
iris.data.shape
✓ 0.0s

(150, 4)

[24]
iris.feature_names
✓ 0.0s

['sepal length (cm)',
 'sepal width (cm)',
 'petal length (cm)',
 'petal width (cm)']
```

Создадим датафрейм и узнаем, что тип каждого поля является числовым (float 64). Также мы можем увидеть в нашем наборе данных отсутствуют пропуски

```
[25]
irisFrame = pd.DataFrame(iris.data, columns=iris.feature_names)
irisFrame.head()
✓ 0.1s
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2

```
[26]
irisFrame.info()
✓ 0.0s

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 4 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   sepal length (cm)  150 non-null    float64
 1   sepal width (cm)   150 non-null    float64
 2   petal length (cm)  150 non-null    float64
 3   petal width (cm)   150 non-null    float64
dtypes: float64(4)
memory usage: 4.8 KB

[27]
irisFrame.isna().sum()
✓ 0.0s

sepal length (cm)    0
sepal width (cm)     0
petal length (cm)    0
petal width (cm)     0
dtype: int64
```

Визуализируем корелляционную матрицу

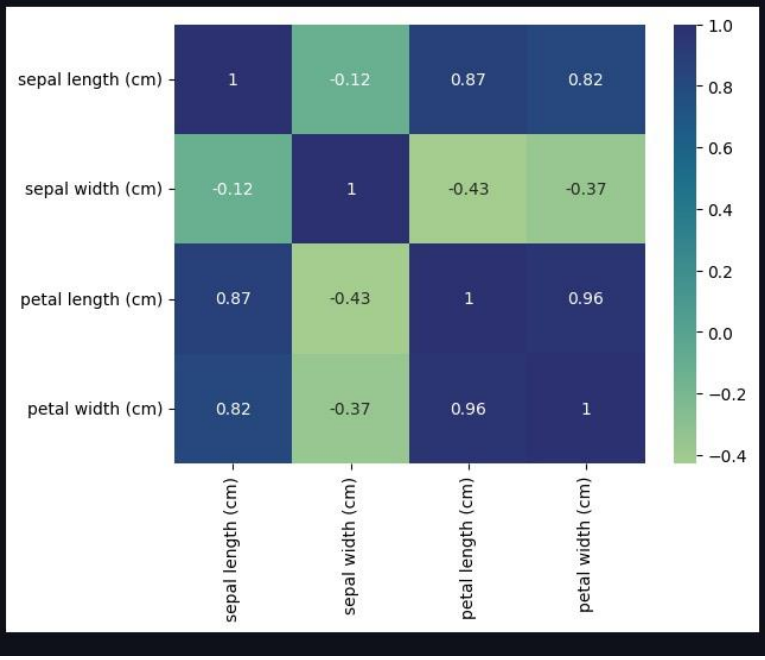
```
irisFrame.duplicated().sum()
```
[28] ✓ 0.0s

```
1
```

Пропуски отсутствуют

Корреляционный анализ

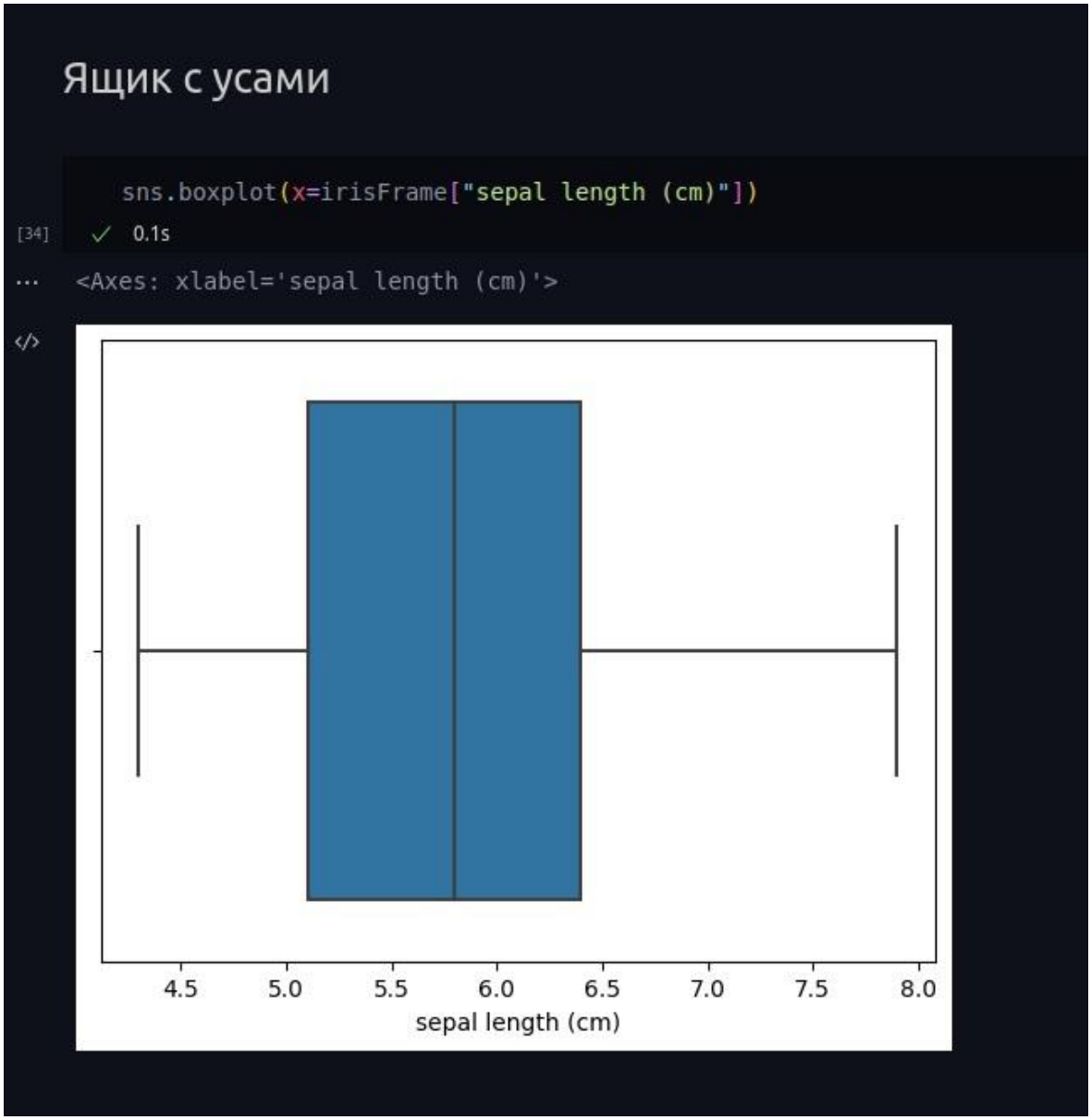
```
dataplot = sns.heatmap(irisFrame.corr(), cmap="crest", annot=True)
```
[39] ✓ 0.2s



Вывод по анализу:

Наиболее сильную зависимость можно заметить между переменными *petal length (cm)* и *petal width (cm)*, также между *sepal length (cm)* и *petal length (cm)* и *sepal length (cm)* и *petal width (cm)*. Следовательно, эти признаки будут наиболее информативными при построении моделей машинного обучения.

Таким образом, на основе признаков *petal length (cm), petal width (cm), sepal length (cm)* могут быть построены модели машинного обучения.



Данный график позволяет увидеть минимальные/максимальные значения, медиану, верхний и нижний квартили.

Так как в представленном датасете не были обнаружены пропуски, возьмемдополнительный датасет, содержащий пропуски.

Для этого был взят набор данных, содержащий данные об отзывах критиков на сайте Rotten Tomatoes

Дополнительный датасет

В выбранном датасете по варианту отсутствовали пропуски, поэтому возьмем датасет с пропусками, для дальнейшей обработки

```
df = pd.read_csv("rotten_tomatoes_movie_reviews.csv")
```
[47] ✓ 4.4s Python

```
df.head()
```
[48] ✓ 0.0s Python

	id	reviewId	creationDate	criticName	isTopCritic	originalScore	reviewState	publicatioName	reviewText	scoreSentiment	reviewUrl
0	beavers	1145982	2003-05-23	Ivan M. Lincoln	False	3.5/4	fresh	Deseret News (Salt Lake City)	Timed to be just long enough for most youngste...	POSITIVE	http://www.deseretnews.com/article/700003233/B...
1	blood_mask	1636744	2007-06-02	The Foywonder	False	1/5	rotten	Dread Central	It doesn't matter if a movie costs 300 million...	NEGATIVE	http://www.dreadcentral.com/index.php?name=Rev...
2	city_hunter_shinjuku_private_eyes	2590987	2019-05-28	Reuben Baron	False	NaN	fresh	CBR	The choreography is so precise and lifelike at...	POSITIVE	https://www.cbr.com/city-hunter-shinjuku-priva...
3	city_hunter_shinjuku_private_eyes	2558908	2019-02-14	Matt Schley	False	2.5/5	rotten	Japan Times	The film's out-of-touch attempts at humor may ...	NEGATIVE	https://www.japantimes.co.jp/culture/2019/02/0...
4	dangerous_men_2015	2504681	2018-08-29	Pat Padua	False	NaN	fresh	DCist	Its clumsy determination is endearing and some...	POSITIVE	http://dcist.com/2015/11/out_of_frame_dangerou...

Посчитаем количество пропусков и удалим пропущенные значения

```
df.isna().sum()
```
[49] ✓ 0.3s

```
id                   0
reviewId             0
creationDate         0
criticName           0
isTopCritic          0
originalScore     435218
reviewState          0
publicatioName       0
reviewText       69225
scoreSentiment       0
reviewUrl        210925
dtype: int64
```

```
df.describe().T
```
[50] ✓ 0.0s

	count	mean	std	min	25%	50%	75%	max
reviewId	1444963.0	9.035203e+06	2.575716e+07	1.0	1610366.5	2200337.0	2587023.5	102796154.0

```
df = df.dropna()
```
[51] ✓ 0.4s

```
df.isna().sum()
```
[52] ✓ 0.2s

```
id               0
reviewId         0
creationDate     0
criticName       0
isTopCritic      0
originalScore    0
reviewState      0
publicatioName   0
reviewText       0
scoreSentiment   0
reviewUrl        0
dtype: int64
```