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 вычислительная техника»
Кафедра ИУ5 «Системы обработки информации и управления»

Курс «Технологии машинного обучения»

Отчет по лабораторной работе №1
«Разведочный анализ данных. Исследование и визуализация данных»

Выполнил:
студент группы ИУ5-62Б

Веревкина Диана В.
Подпись и дата:

Проверил:
преподаватель каф. ИУ5
Гапанюк Ю.Е.
Подпись и дата:

Москва, 2022 г.

Цель лабораторной работы

Изучение различных методов визуализация данных.

Описание задания

- Выбрать набор данных (датасет).
- Для первой лабораторной работы рекомендуется использовать датасет без пропусков в данных.

- Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного Вами 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.

Текст программы

1. Текстовое описание набора данных

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")

df = pd.read_csv('penguins_size.csv', sep=",")
#поиск пропусков
#df.isna()
#заполнение пропусков
#df.fillna(value=df.mean())
df.dropna()
#поиск дубликатов
df.duplicated()
#df.drop_duplicates()#удаление дубликатов
```

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	MALE
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	FEMALE
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	FEMALE
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	FEMALE
5	Adelie	Torgersen	39.3	20.6	190.0	3650.0	MALE
...	...	...	...	...	...	...	...
338	Gentoo	Biscoe	47.2	13.7	214.0	4925.0	FEMALE
340	Gentoo	Biscoe	46.8	14.3	215.0	4850.0	FEMALE
341	Gentoo	Biscoe	50.4	15.7	222.0	5750.0	MALE
342	Gentoo	Biscoe	45.2	14.8	212.0	5200.0	FEMALE

2. Основные характеристики датасета

```
# Первые 5 строк датасета
df.head()
```

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	MALE
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	FEMALE
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	FEMALE
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	FEMALE
5	Adelie	Torgersen	39.3	20.6	190.0	3650.0	MALE

```
# Размер датасета - 344 строки, 7 колонок
data.shape
```

```
(344, 7)
```

```
# Список колонок
df.columns
```

```
Index(['species', 'island', 'culmen_length_mm', 'culmen_depth_mm',
       'flipper_length_mm', 'body_mass_g', 'sex'],
      dtype='object')
```

```
# Список колонок с типами данных
df.dtypes
```

```
species               object
island                object
culmen_length_mm     float64
culmen_depth_mm      float64
flipper_length_mm    float64
body_mass_g          float64
sex                   object
dtype: object
```

```
# Основные статистические характеристки набора данных
df.describe()
```

	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g
count	334.000000	334.000000	334.000000	334.000000
mean	43.994311	17.160479	201.014970	4209.056886
std	5.460521	1.967909	14.022175	804.836129
min	32.100000	13.100000	172.000000	2700.000000
25%	39.500000	15.600000	190.000000	3550.000000
50%	44.500000	17.300000	197.000000	4050.000000
75%	48.575000	18.700000	213.000000	4793.750000
max	59.600000	21.500000	231.000000	6300.000000

```
# Определим уникальные значения для видов
data['species'].unique()
```

```
array(['Adelie', 'Chinstrap', 'Gentoo'], dtype=object)
```

```
# Определим уникальные значения для островов
data['island'].unique()
```

```
array(['Torgersen', 'Biscoe', 'Dream'], dtype=object)
```

3. Визуальное исследование датасета

```
#визуализация
#зависимость длины крыла от массы
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='body_mass_g', y='flipper_length_mm', data=df)
```

```
<AxesSubplot:xlabel='body_mass_g', ylabel='flipper_length_mm'>
```

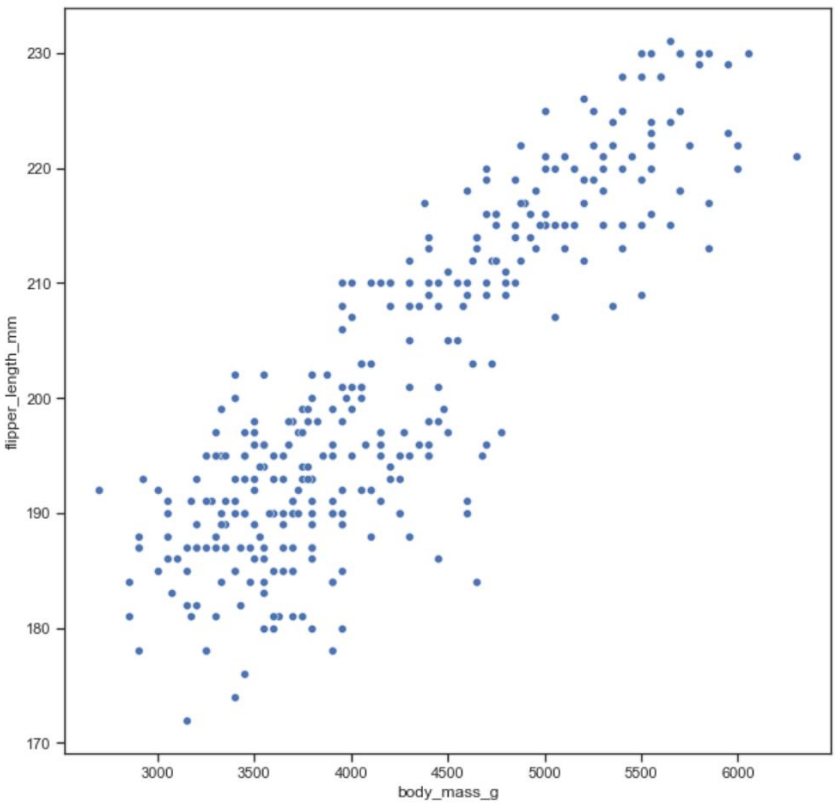
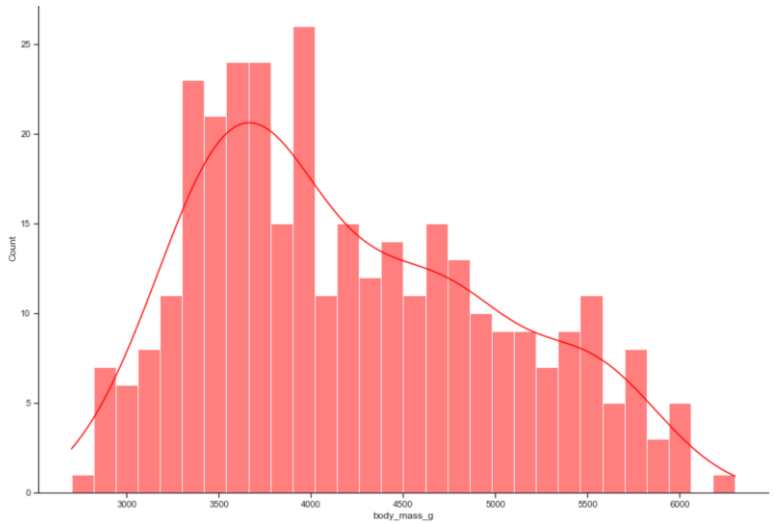


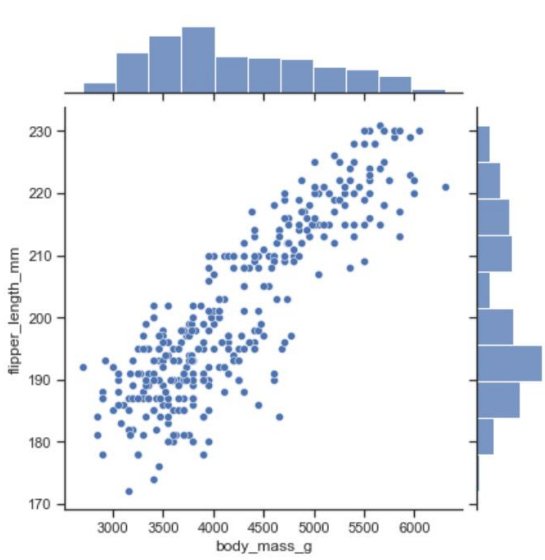
Диаграмма рассеяния показывает зависимость длины крыла пингвина от его массы. Из результатов проглядывается линейная зависимость с отклонениями. Отклонения могут быть связаны со значением веса выше или ниже нормы (ожирение или дефицит массы).

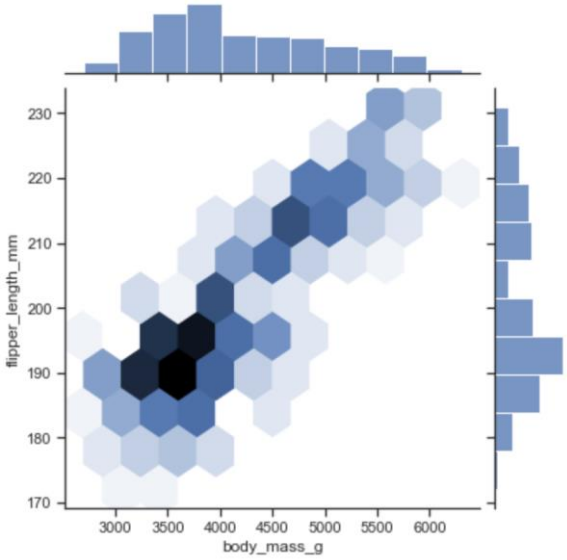
```
#гистограмма - распределение массы
ax = sns.displot(data=df,x='body_mass_g',color='red',bins=30,kde=True);
ax.fig.set_figheight(10)
ax.fig.set_figwidth(15)
```



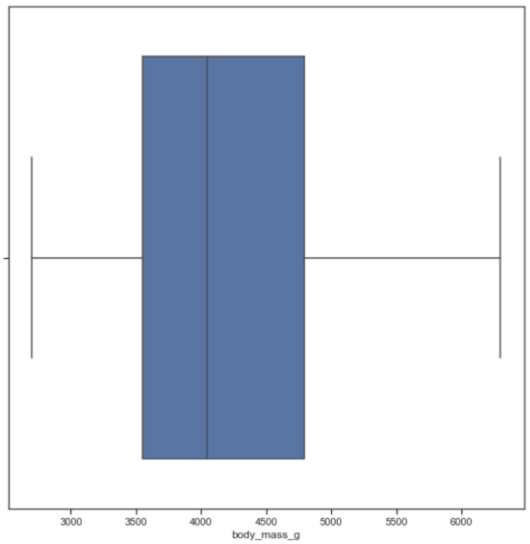
```
#Комбинация гистограмм и диаграмм рассеивания
sns.jointplot(x='body_mass_g', y='flipper_length_mm', data=df)
```

```
#Комбинация гистограмм и диаграмм рассеивания
sns.jointplot(x='body_mass_g', y='flipper_length_mm', data=df, kind="hex")
```

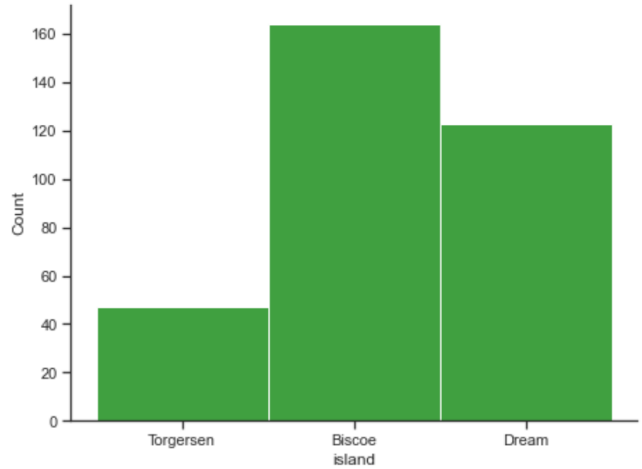





```
#одномерное распределение вероятности
sns.boxplot(x=df['body_mass_g'])
```

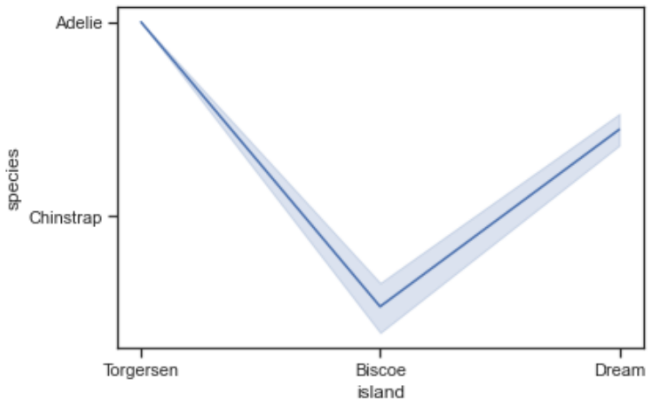



```
#распределение островов обитания
ax = sns.displot(data=df,x='island',color='green',bins=10,kde=False);
ax.fig.set_figheight(5)
ax.fig.set_figwidth(7)
```



```
#зависимость проживания видов от островов
sns.lineplot(x='island', y='species', data=df)
```

```
<AxesSubplot:xlabel='island', ylabel='species'>
```



4. Информация о корреляции признаков

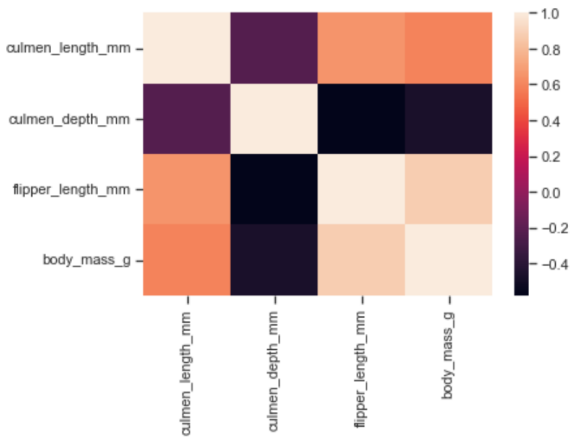
```
#Информация о корреляции признаков
df.corr()
```

	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g
culmen_length_mm	1.000000	-0.228640	0.652126	0.589066
culmen_depth_mm	-0.228640	1.000000	-0.578730	-0.472987
flipper_length_mm	0.652126	-0.578730	1.000000	0.873211
body_mass_g	0.589066	-0.472987	0.873211	1.000000

Довольно высокий процент корреляции можно наблюдать между длиной крыла и клюва и между длиной крыла и массой тела.

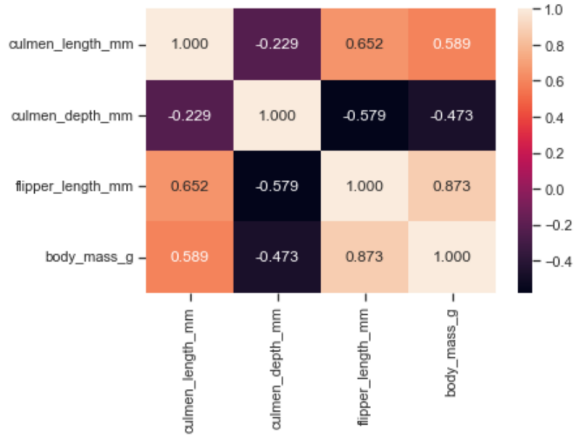
```
#тепловая карта для корреляции
sns.heatmap(df.corr())
```

```
<AxesSubplot:>
```




```
sns.heatmap(df.corr(), annot=True, fmt='.3f')
```

```
<AxesSubplot:>
```



Вывод

В ходе выполнения данной лабораторной работы я повторила язык программирования Python и работу с юпитер тетрадками. Также вспомнила функции для визуализации данных.