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 Системы обработки информации и управления

Отчёт по лабораторной работе №1

По дисциплине:
«Технологии машинного обучения»

Выполнил:

Студент группы ИУ5-62 _____ **Чепкин Д.А.**

(Подпись, дата) (Фамилия И.О.)

Проверил:

_____ **Гапанюк Ю. Е.**

(Подпись, дата) (Фамилия И.О.)

Москва, 2021

Задание:

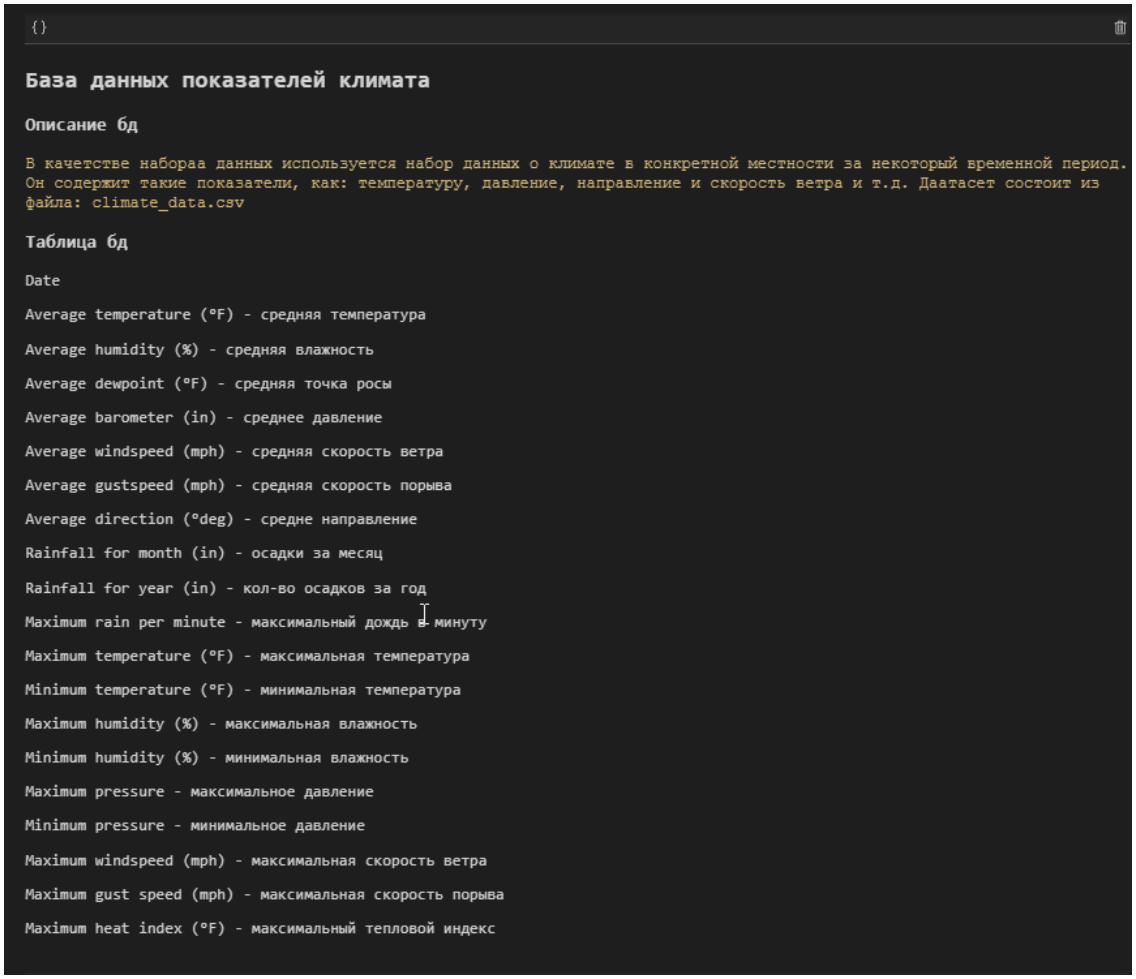
- Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь.
- Для первой лабораторной работы рекомендуется использовать датасет без пропусков в данных, например из Scikit-learn.
- Пример преобразования датасетов Scikit-learn в Pandas Dataframe можно посмотреть здесь.

Для лабораторных работ не рекомендуется выбирать датасеты большого размера.

- 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.



[4]
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

[8]
```
Data = pd.read_csv('climate_data.csv', sep=",")
```

[9]
```
print("Размер таблицы climate_data: ", Data.shape)
```

```
Размер таблицы climate_data:  (3902, 23)
```

[15]
```
total_count = Data.shape[0]
print('Всего строк: {}'.format(total_count))
```

```
Всего строк: 3902
```

[10]
```
print("climate_data:")
Trans.head()
```

```
climate_data:
```

	Date	Average temperature (°F)	Average humidity (%)	Average dewpoint (°F)	Average barometer (in)	Average windspeed (mph)	Average gustspeed (mph)	Average direction (°deg)	Rainfall for month (in)	Rainfall for year (in)	...	Maximum humidity (%)	Minimum humidity (%)	Maximum pressure	Minim pressu
0	2009-01-01	37.8	35.0	12.7	29.7	26.4	36.8	274.0	0.0	0.0	...	4.0	27.0	29.762	29.5
1	2009-01-02	43.2	32.0	14.7	29.5	12.8	18.0	240.0	0.0	0.0	...	4.0	16.0	29.669	29.2
2	2009-01-03	25.7	60.0	12.7	29.7	8.3	12.2	290.0	0.0	0.0	...	8.0	35.0	30.232	29.2
3	2009-01-04	9.3	67.0	0.1	30.4	2.9	4.5	47.0	0.0	0.0	...	7.0	35.0	30.566	30.2
4	2009-01-05	23.5	30.0	-5.3	29.9	16.7	23.1	265.0	0.0	0.0	...	5.0	13.0	30.233	29.5

```
5 rows × 23 columns
```

[17]
```
# Список колонок с типами данных
Data.dtypes
```

```
Date                          object
Average temperature (°F)     float64
Average humidity (%)         float64
Average dewpoint (°F)        float64
Average barometer (in)       float64
Average windspeed (mph)      float64
Average gustspeed (mph)      float64
Average direction (°deg)     float64
Rainfall for month (in)      float64
Rainfall for year (in)       float64
Maximum rain per minute      float64
Maximum temperature (°F)     float64
Minimum temperature (°F)     float64
Maximum humidity (%)         float64
Minimum humidity (%)         float64
Maximum pressure             float64
Minimum pressure             float64
Maximum windspeed (mph)      float64
Maximum gust speed (mph)     float64
Maximum heat index (°F)      float64
Date1                         object
Month                          int64
diff_pressure                float64
dtype: object
```

[11]
```
print('Количество пустых ячеек в таблице TransactionBase:')
for col in Data.columns:
    temp_null_count = Data[Data[col].isnull()].shape[0]
    print('{} - {}'.format(col, temp_null_count))
```

```
Количество пустых ячеек в таблице TransactionBase:
Date - 0
Average temperature (°F) - 0
Average humidity (%) - 0
Average dewpoint (°F) - 0
Average barometer (in) - 0
Average windspeed (mph) - 0
Average gustspeed (mph) - 0
Average direction (°deg) - 0
Rainfall for month (in) - 0
Rainfall for year (in) - 0
Maximum rain per minute - 0
Maximum temperature (°F) - 0
Minimum temperature (°F) - 0
Maximum humidity (%) - 0
Minimum humidity (%) - 0
Maximum pressure - 0
Minimum pressure - 0
Maximum windspeed (mph) - 0
```

[12]

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

	Average temperature (°F)	Average humidity (%)	Average dewpoint (°F)	Average barometer (in)	Average windspeed (mph)	Average gustspeed (mph)	Average direction (°deg)	Rainfall for month (in)	Rainfall for year (in)	Maximum rain per minute	...	Minimu temperatur (°F
count	3902.000000	3902.000000	3902.000000	3902.000000	3902.000000	3902.000000	3902.000000	3902.000000	3902.000000	3902.0	...	3902.00000
mean	44.670733	48.878011	23.127037	29.881420	5.758893	10.011968	216.037417	0.451105	5.486171	0.0	...	31.22752
std	15.326793	17.438153	14.634088	0.250395	4.022485	14.117446	97.677761	0.603462	4.534444	0.0	...	14.12442
min	-12.100000	9.000000	-22.200000	28.200000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	-27.70000
25%	33.700000	36.000000	12.100000	29.700000	2.700000	4.500000	116.000000	0.050000	0.980000	0.0	...	23.00000
50%	45.100000	47.000000	22.500000	29.900000	4.600000	7.100000	253.000000	0.220000	5.080000	0.0	...	32.80000
75%	58.000000	61.000000	35.400000	30.000000	8.000000	12.100000	282.000000	0.670000	9.047500	0.0	...	41.80000
max	76.300000	94.000000	55.100000	31.000000	26.400000	240.400000	360.000000	4.480000	16.410000	0.0	...	65.70000

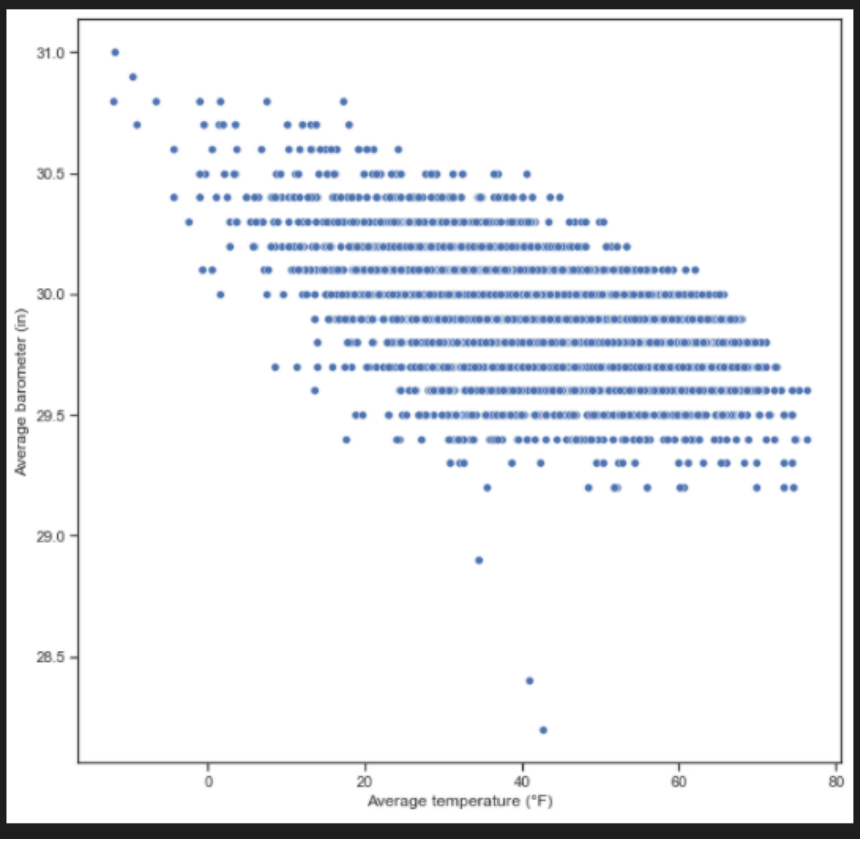
8 rows × 21 columns

Визуальное исследование

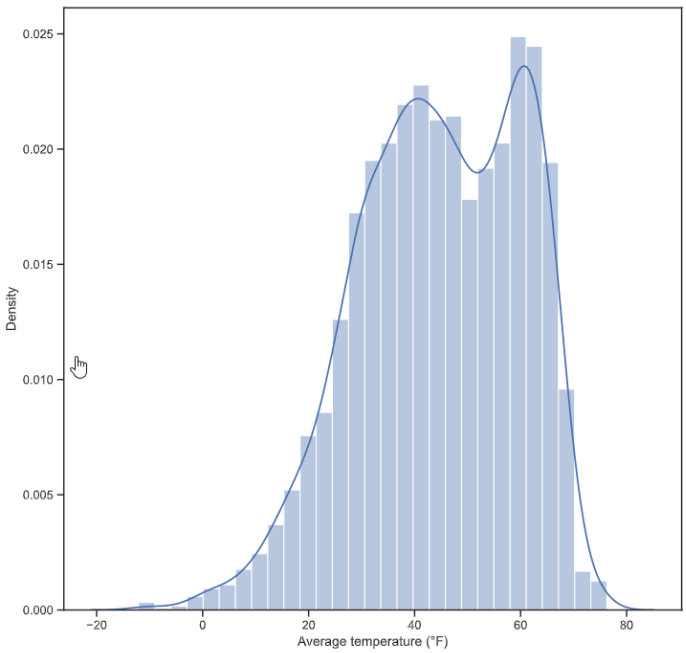
[19]

```
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='Average temperature (°F)', y='Average barometer (in)', data=Data)
```

```
<AxesSubplot:xlabel='Average temperature (°F)', ylabel='Average barometer (in)'>
```



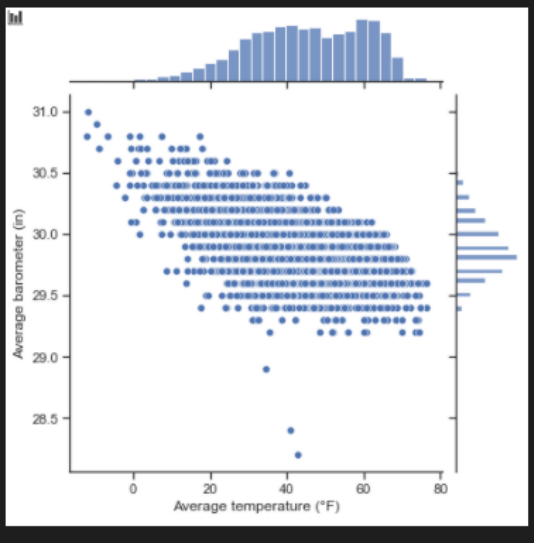

```
fig, ax = plt.subplots(figsize=(10,10))
sns.distplot(Data['Average temperature (°F)'])
```



[21]

```
sns.jointplot(x='Average temperature (°F)', y='Average barometer (in)', data=Data)
```

```
<seaborn.axisgrid.JointGrid at 0x11097e2da60>
```



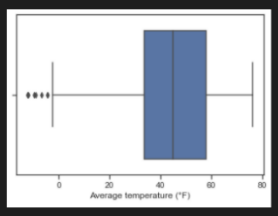
Информация о корреляции признаков.

[5]
```
Data.corr()
```

	Average temperature (°F)	Average humidity (%)	Average dewpoint (°F)	Average barometer (in)	Average windspeed (mph)	Average gustspeed (mph)	Average direction (°deg)	Rainfall for month (in)	Rainfall for year (in)	Maximum rain per minute	...	Minimum temperature (°F)	M hu
Average temperature (°F)	1.000000	-0.258103	0.764830	-0.550897	-0.167162	0.000737	0.034183	0.299229	0.203111	NaN	...	0.919248	0.
Average humidity (%)	-0.258103	1.000000	0.404557	0.178934	-0.516141	-0.197759	-0.315470	0.227414	0.006378	NaN	...	-0.170000	0.
Average dewpoint (°F)	0.764830	0.404557	1.000000	-0.403144	-0.455355	-0.112658	-0.156349	0.429974	0.194450	NaN	...	0.762414	0.
Average barometer (in)	-0.550897	0.178934	-0.403144	1.000000	-0.121518	-0.062563	-0.143962	-0.128451	0.013939	NaN	...	-0.581604	-0.
Average windspeed (mph)	-0.167162	-0.516141	-0.455355	-0.121518	1.000000	0.393666	0.291648	-0.209548	-0.106904	NaN	...	-0.056171	-0.
Average gustspeed (mph)	0.000737	-0.197759	-0.112658	-0.062563	0.393666	1.000000	0.076630	-0.057578	0.007168	NaN	...	0.032247	-0.
Average direction (°deg)	0.034183	-0.315470	-0.156349	-0.143962	0.291648	0.076630	1.000000	-0.045220	-0.049204	NaN	...	0.033013	-0.
Rainfall for month (in)	0.299229	0.227414	0.429974	-0.128451	-0.209548	-0.057578	-0.045220	1.000000	0.116040	NaN	...	0.299055	0.
Rainfall for year (in)	0.203111	0.006378	0.194450	0.013939	-0.106904	0.007168	-0.049204	0.116040	1.000000	NaN	...	0.193721	0.
Maximum rain per minute	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
Maximum temperature (°F)	0.963821	-0.230784	0.739464	-0.462309	-0.277724	-0.051876	0.014190	0.286233	0.204039	NaN	...	0.817127	0.
Minimum temperature (°F)	0.919248	-0.170000	0.762414	-0.581604	-0.056171	0.032247	0.033013	0.299055	0.193721	NaN	...	1.000000	0.
Maximum humidity (%)	0.068721	0.696767	0.515906	-0.029233	-0.542426	-0.189880	-0.201194	0.277425	0.105984	NaN	...	0.007406	1.

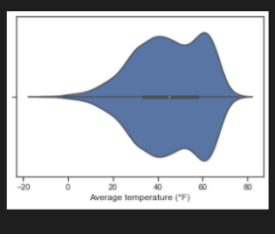
[23]
```
sns.boxplot(x=Data['Average temperature (°F)'])
```

```
<AxesSubplot:xlabel='Average temperature (°F)'>
```



[24]
```
sns.violinplot(x=Data['Average temperature (°F)'])
```

```
<AxesSubplot:xlabel='Average temperature (°F)'>
```




```
sns.heatmap(Data.corr())
```

```
<AxesSubplot:>
```