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

Отчет

Лабораторная работа № 1

По курсу «Технологии машинного обучения»

«Разведочный анализ данных. Исследование и визуализация данных»

ИСПОЛНИТЕЛЬ:

Молева Анастасия
Группа ИУ5-61Б

_____

"__"_____2020 г.

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

_____

"__"_____2020 г.

Москва 2020

Цель лабораторной работы:

Изучение различных методов визуализация данных.

Краткое описание.

Построение основных графиков, входящих в этап разведочного анализа данных.

1) **Текстовое описание набора данных**

Dataset (www.kaggle.com):
2019 Coronavirus dataset (January - February 2020)
Tracking the spread of 2019-nCoV
Context:

The 2019-nCoV is a contagious coronavirus that hailed from Wuhan, China. This new strain of virus has striked fear in many countries as cities are quarantined and hospitals are overcrowded. This dataset will help us understand how 2019-nCoV is spread aroud the world.

Датасет содержит следующие колонки:

- Province/State – провинция, город
- Country – страна
- Date last updated – последнее обновление данных
- Confirmed – подтвержден вирус, сколько заболевших
- Suspected – ожидают лечение
- Recovered – выздоровело людей
- Deaths – смертей от вируса

2) **Загрузка данных**

Загрузим файлы датасета в помощью библиотеки Pandas.

```
In [1]: import pandas

        data = pandas.read_csv('2019_nC0v_20200121_20200126 - SUMMARY.csv')
```

3) Основные характеристики датасета

In [2]:
```
data[:]
```

Out[2]:

	Province/State	Country	Date last updated	Confirmed	Suspected	Recovered	Deaths
0	Shanghai	Mainland China	1/21/2020	9.0	10.0	NaN	NaN
1	Yunnan	Mainland China	1/21/2020	1.0	NaN	NaN	NaN
2	Beijing	Mainland China	1/21/2020	10.0	NaN	NaN	NaN
3	Taiwan	Mainland China	1/21/2020	1.0	NaN	NaN	NaN
4	Jilin	Mainland China	1/21/2020	NaN	1.0	NaN	NaN
...	...	...	...	...	...	...	...
363	NaN	France	1/26/2020 11:00 AM	3.0	NaN	NaN	NaN
364	NaN	Australia	1/26/2020 11:00 AM	4.0	NaN	NaN	NaN
365	NaN	Nepal	1/26/2020 11:00 AM	1.0	NaN	NaN	NaN
366	NaN	Malaysia	1/26/2020 11:00 AM	4.0	NaN	NaN	NaN
367	Ontario	Canada	1/26/2020 11:00 AM	1.0	NaN	NaN	NaN

368 rows × 7 columns

In [3]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 368 entries, 0 to 367
Data columns (total 7 columns):
Province/State       304 non-null object
Country              368 non-null object
Date last updated    368 non-null object
Confirmed            339 non-null float64
Suspected            88 non-null float64
Recovered            36 non-null float64
Deaths               24 non-null float64
dtypes: float64(4), object(3)
memory usage: 20.2+ KB
```

In [68]:
```
data['Recovered'].value_counts()
```

Out[68]:
```
1.0     18
2.0     11
42.0     2
32.0     2
31.0     2
28.0     1
Name: Recovered, dtype: int64
```

In [40]:
```
data.describe()
```

Out[40]:

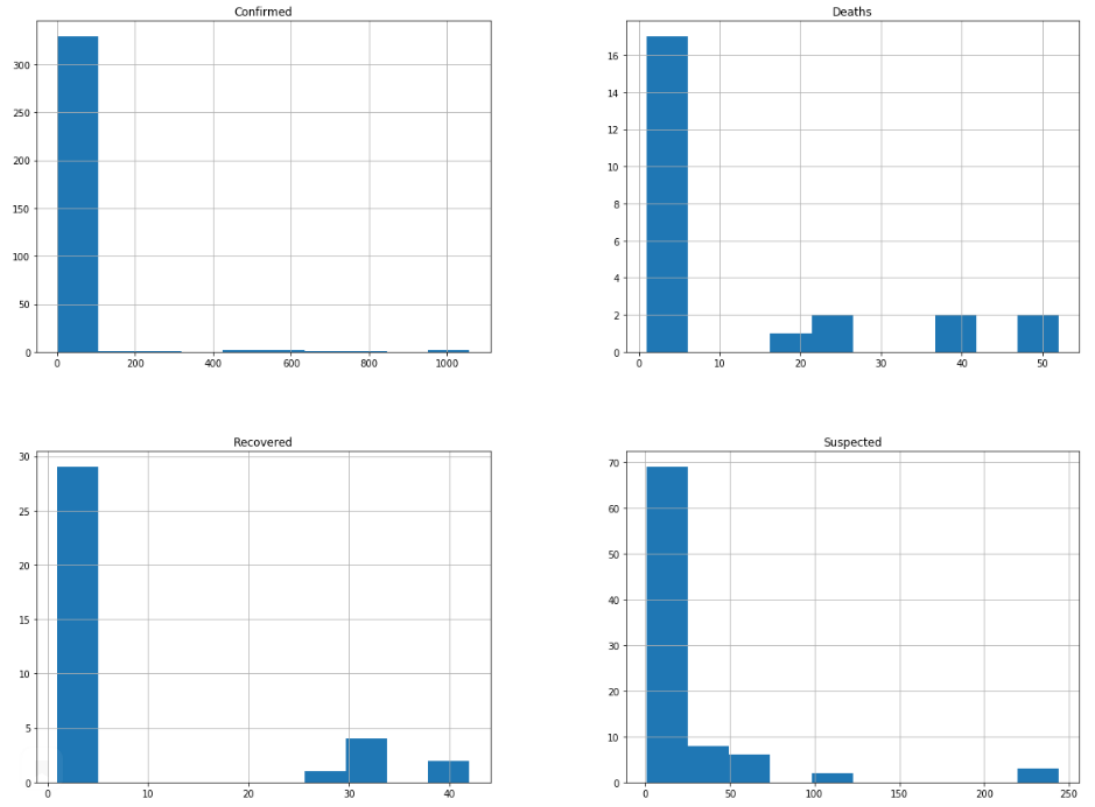
	Confirmed	Suspected	Recovered	Deaths
count	339.000000	88.000000	36.000000	24.000000
mean	30.351032	22.613636	7.722222	11.041667
std	112.556169	48.177696	13.306521	17.521364
min	1.000000	1.000000	1.000000	1.000000
25%	2.000000	1.000000	1.000000	1.000000
50%	5.000000	4.000000	1.500000	1.000000
75%	18.000000	22.000000	2.000000	18.750000
max	1058.000000	244.000000	42.000000	52.000000

In [38]:
```
#Количество смертей от коронавируса
import math
for i in range(len(data['Deaths'])):
    if math.isnan(data['Deaths'][i]) == False:
        print(str(data['Province/State'][i]) + '\t' + str(data['Deaths'][i]))
```

```
Hebei   1.0
Hubei   17.0
Hubei   24.0
Heilongjiang    1.0
Hebei   1.0
Hubei   24.0
Heilongjiang    1.0
Hebei   1.0
Hubei   39.0
Heilongjiang    1.0
Hebei   1.0
Hubei   40.0
Heilongjiang    1.0
Hebei   1.0
Hubei   52.0
Henan   1.0
Shanghai        1.0
Fujian  1.0
Hebei   1.0
Hubei   52.0
Henan   1.0
Shanghai        1.0
Heilongjiang    1.0
Hebei   1.0
```

4) Визуальное исследование датасета

In [58]:
```
import matplotlib.pyplot as plt
%matplotlib inline
data.hist(bins=10, figsize= (20, 15))
plt.show()
```



In [69]:
```
# Создание испытательного набора
from sklearn.model_selection import train_test_split

train_set, test_set = train_test_split(data, test_size=0.2, random_state=42)
print(len(train_set), "train +", len(test_set), "test")
test_set.head()
```

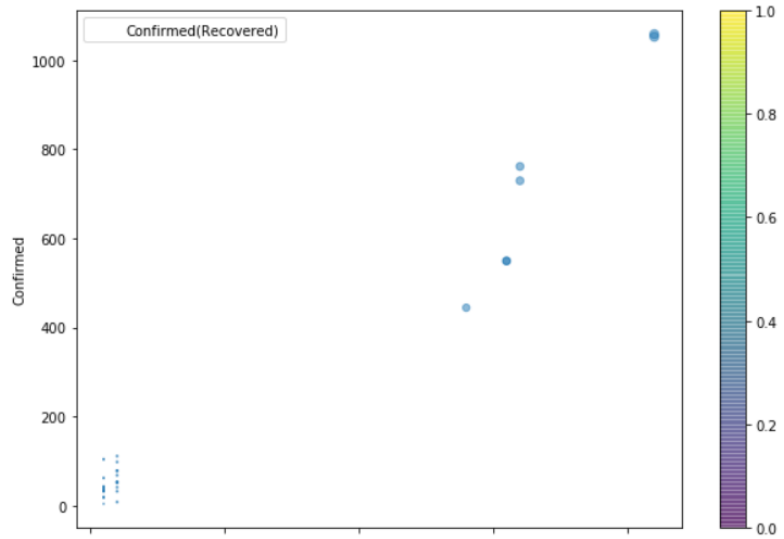
294 train + 74 test

Out[69]:

	Province/State	Country	Date last updated	Confirmed	Suspected	Recovered	Deaths
165	Tianjin	Mainland China	1/24/2020 12:00 PM	8.0	NaN	NaN	NaN
33	Guangxi	Mainland China	1/22/2020 12:00	2.0	1.0	NaN	NaN
15	Hainan	Mainland China	1/21/2020	NaN	1.0	NaN	NaN
312	Illinois	US	1/25/2020 12:00 PM	1.0	NaN	NaN	NaN
57	Tibet	China	1/22/2020 12:00	NaN	NaN	NaN	NaN

In [95]:
```
#s - радиус
# c - цвет
data.plot(kind="scatter", x="Recovered", y="Confirmed", alpha=0.5, s=data["Recovered"],
          figsize=(10,7), cmap=plt.get_cmap("jet"), colorbar=True, label="Confirmed(Recovered)")
```

Out[95]: <matplotlib.axes._subplots.AxesSubplot at 0x25bdbcecfd0>



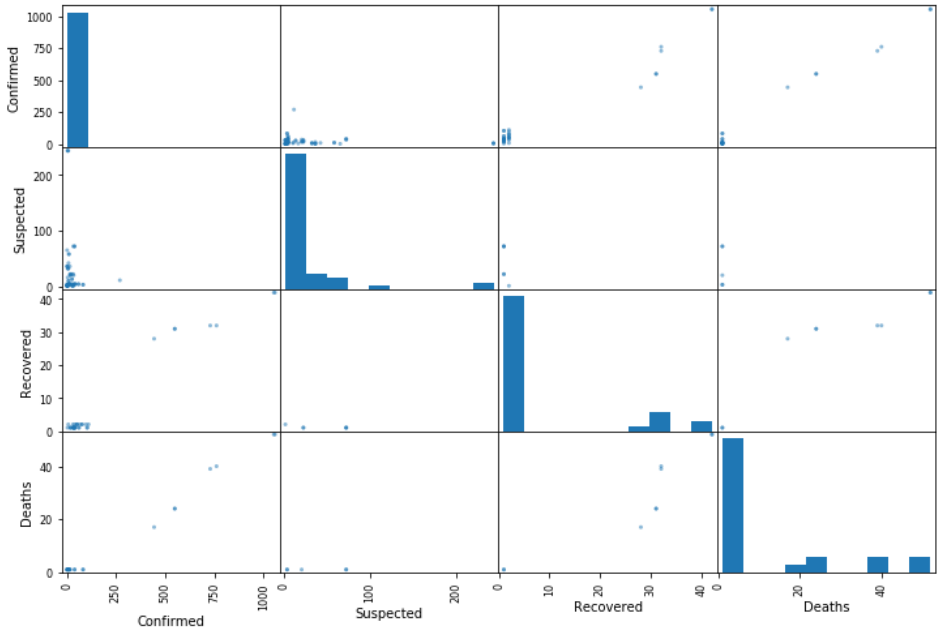
5) Корреляция

In [97]:
```
#Корреляция по коэффициенту корреляции Пирсона
corr_matrix = data.corr()

corr_matrix["Deaths"].sort_values(ascending=False)
```

Out[97]:
```
Deaths       1.000000
Confirmed    0.994090
Recovered    0.919208
Suspected         NaN
Name: Deaths, dtype: float64
```

In [105]:
```
#Тоже корреляция - вычерчивает каждый числовой атрибут по отношению к каждому другому числовому атрибуту
from pandas.plotting import scatter_matrix

scatter_matrix(data, figsize=(12,8))
```

Out[105]:
```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x0000025BE160BCF8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000025BE16200B8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000025BE1342240>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000025BE13747F0>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x0000025BE13A5DA0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000025BE13E3390>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000025BE1411940>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000025BE1444F28>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x0000025BE1444F60>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000025BE14B1A90>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000025BE14F2080>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000025BE1521630>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x0000025BE1554BE0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000025BE15931D0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000025BE15C1780>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000025BE1854D30>]],
      dtype=object)
```



Вывод:

Научился работе в Jupyter notebook. Познакомился с ML.