

**Московский государственный технический  
университет им. Н.Э. Баумана.**

Факультет «Информатика и управление»  
Кафедра ИУ5. Курс «Технологии машинного обучения»

**Отчет по лабораторной работе № 1**

Выполнил:  
студент группы ИУ5-61Б  
**Егор Утробин**  
Подпись и дата:

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

Москва, 2021 г.

March 22, 2021

# 1 №1

## 1.0.1

```
[1]: import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from pandas.plotting import scatter_matrix
import warnings
warnings.filterwarnings('ignore')
sns.set(style="ticks")
%matplotlib inline
```

## 1.0.2

```
[2]: happy_data = pd.read_csv('reddit_wsb.csv', sep = ',', )
```

### 1.1 2)

```
[3]: #
happy_data.head()
```

```
[3]:          title  score      id \
0  It's not about the money, it's about sending a...    55  16ulcx
1  Math Professor Scott Steiner says the numbers ...   110  16uibd
2                           Exit the system      0  16uhhn
3  NEW SEC FILING FOR GME! CAN SOMEONE LESS RETAR...    29  16ugk6
4  Not to distract from GME, just thought our AMC...    71  16ufgy

          url  comms_num      created \
0  https://v.redd.it/6j75regs72e61                  6  1.611863e+09
1  https://v.redd.it/ah50lyny62e61                23  1.611862e+09
2  https://www.reddit.com/r/wallstreetbets/comment...    47  1.611862e+09
3  https://sec.report/Document/0001193125-21-019848/   74  1.611862e+09
4  https://i.redd.it/4h2sukb662e61.jpg              156  1.611862e+09
```

```
          body      timestamp
0           NaN 2021-01-28 21:37:41
1           NaN 2021-01-28 21:32:10
2 The CEO of NASDAQ pushed to halt trading "to g... 2021-01-28 21:30:35
3                           NaN 2021-01-28 21:28:57
4                           NaN 2021-01-28 21:26:56
```

```
[4]: #  
happy_data.shape
```

```
[4]: (36520, 8)
```

```
[5]: #  
happy_data.isnull().sum()
```

```
[5]: title      0  
score       0  
id         0  
url         0  
comms_num   0  
created      0  
body        18048  
timestamp     0  
dtype: int64
```

```
[6]: #  
happy_data.dtypes
```

```
[6]: title      object  
score      int64  
id        object  
url        object  
comms_num  int64  
created    float64  
body       object  
timestamp  object  
dtype: object
```

```
[7]: #  
happy_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 36520 entries, 0 to 36519  
Data columns (total 8 columns):  
 #   Column      Non-Null Count  Dtype     
---  --  -----  -----  -----  
 0   title       36520 non-null  object
```

```
1   score      36520 non-null  int64
2   id         36520 non-null  object
3   url        36520 non-null  object
4   comms_num  36520 non-null  int64
5   created     36520 non-null  float64
6   body       18472 non-null  object
7   timestamp  36520 non-null  object
dtypes: float64(1), int64(2), object(5)
memory usage: 2.2+ MB
```

```
[8]: #  
happy_data.describe()
```

```
[8]:          score      comms_num      created
count    36520.000000  36520.000000  3.652000e+04
mean     1224.997700   164.866046  1.612316e+09
std      8272.922526   2166.834418  6.328753e+05
min      0.000000     0.000000  1.601340e+09
25%     1.000000     1.000000  1.611879e+09
50%     15.000000    9.000000  1.612035e+09
75%    138.000000   38.000000  1.612528e+09
max    348241.000000  89050.000000 1.614447e+09
```

```
[9]: #  
id  
happy_data = happy_data.drop('id', axis = 1)
```

```
[10]: #  
happy_data.head()
```

```
[10]:              title  score  \
0  It's not about the money, it's about sending a...      55
1  Math Professor Scott Steiner says the numbers ...    110
2                           Exit the system                  0
3  NEW SEC FILING FOR GME! CAN SOMEONE LESS RETAR...     29
4  Not to distract from GME, just thought our AMC...    71

                               url  comms_num      created  \
0  https://v.redd.it/6j75regs72e61                   6  1.611863e+09
1  https://v.redd.it/ah50lyny62e61                 23  1.611862e+09
2  https://www.reddit.com/r/wallstreetbets/comment...    47  1.611862e+09
3  https://sec.report/Document/0001193125-21-019848/    74  1.611862e+09
4  https://i.redd.it/4h2sukb662e61.jpg                156 1.611862e+09

                           body      timestamp
0                      NaN  2021-01-28 21:37:41
1                      NaN  2021-01-28 21:32:10
2  The CEO of NASDAQ pushed to halt trading "to g... 2021-01-28 21:30:35
```

```
3  
4
```

```
NaN 2021-01-28 21:28:57  
NaN 2021-01-28 21:26:56
```

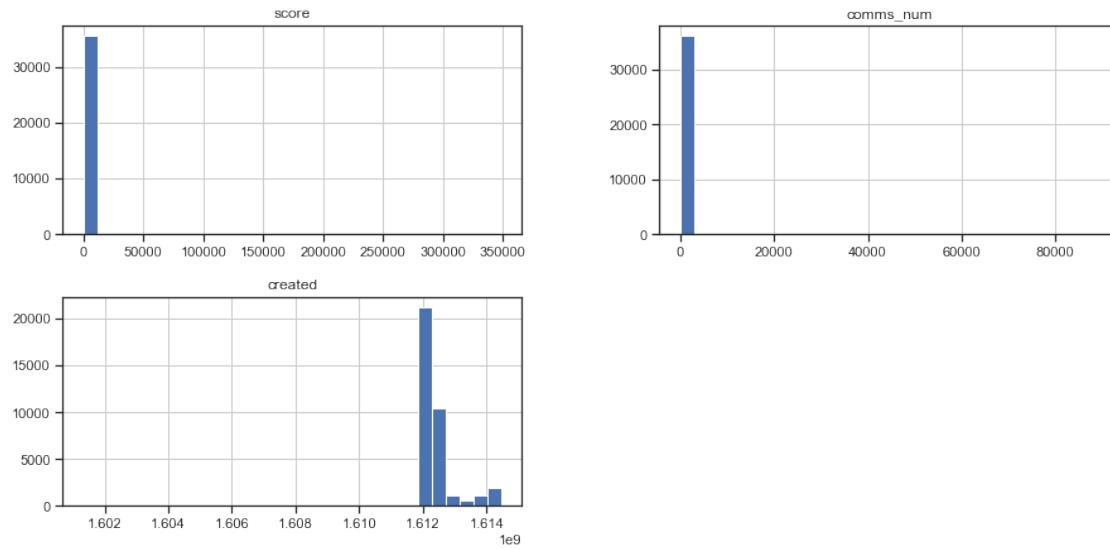
```
[11]: #  
happy_data['score'].unique()
```

```
[11]: array([ 55, 110, 0, ..., 83582, 1289, 37242])
```

## 1.2 3)

```
[12]: #  
happy_data.hist(bins=30, figsize = (15,7))
```

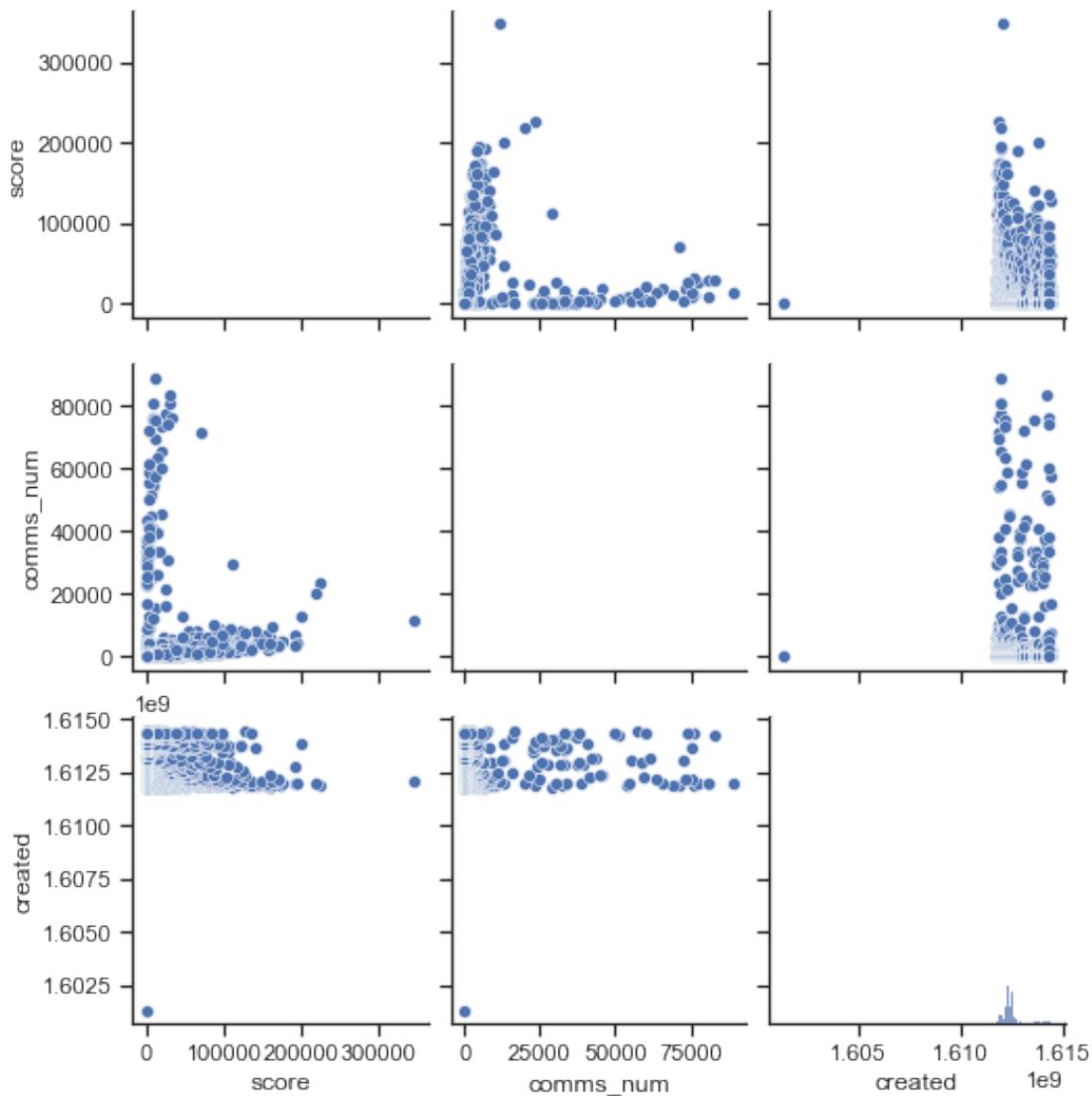
```
[12]: array([[<AxesSubplot:title={'center':'score'}>,  
           <AxesSubplot:title={'center':'comms_num'}>],  
           [<AxesSubplot:title={'center':'created'}>, <AxesSubplot:>]],  
          dtype=object)
```



```
[13]: #  
plt.figure(figsize=(12,6))  
sns.pairplot(happy_data)
```

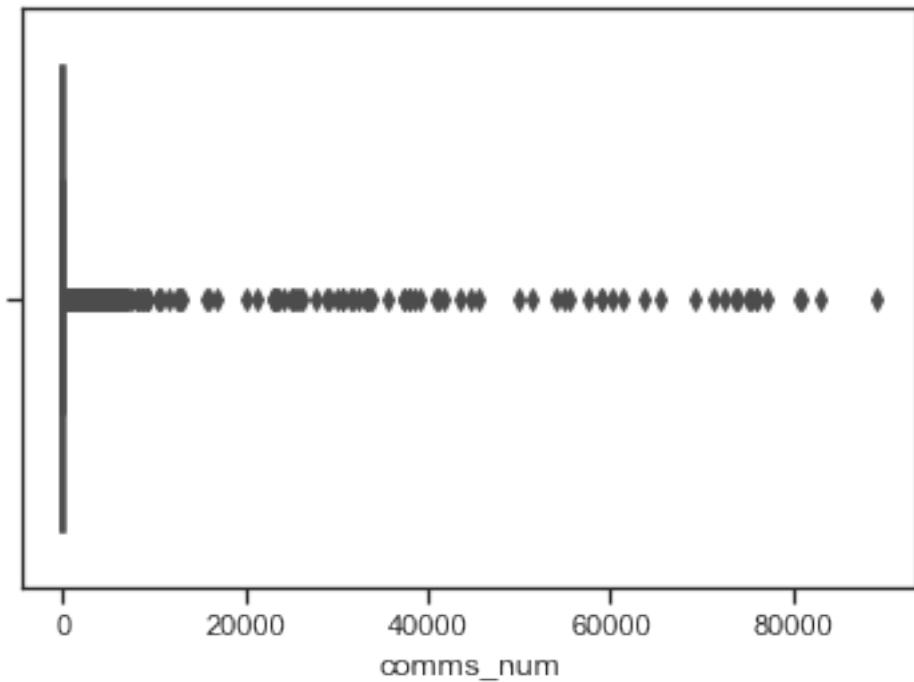
```
[13]: <seaborn.axisgrid.PairGrid at 0x7fc4ffbc4d60>
```

```
<Figure size 864x432 with 0 Axes>
```



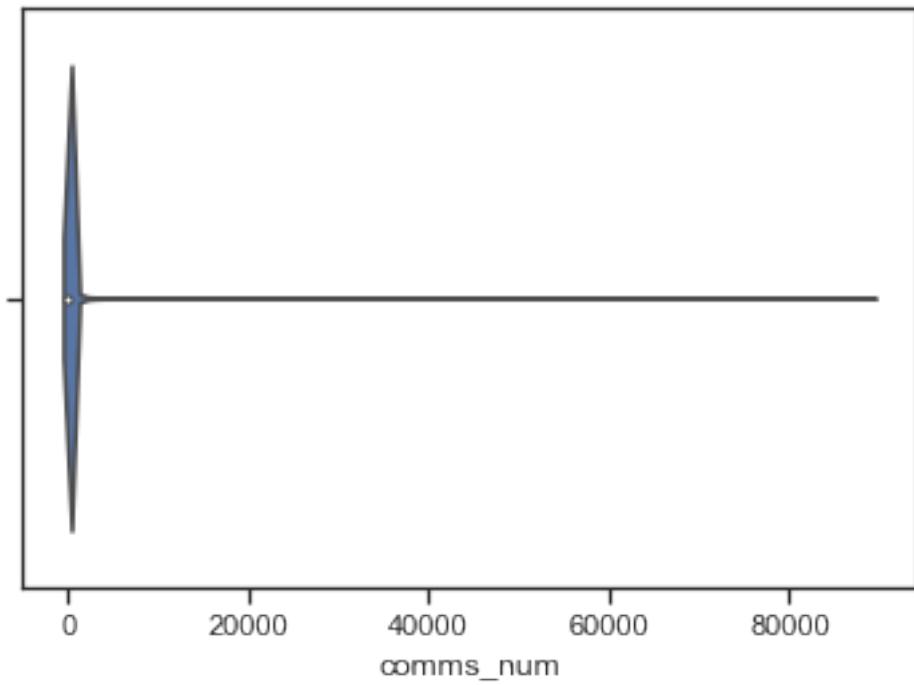
```
[14]: #  
sns.boxplot(x=happy_data['comms_num'])
```

```
[14]: <AxesSubplot:xlabel='comms_num'>
```



```
[15]: sns.violinplot(x=happy_data['comms_num'])
```

```
[15]: <AxesSubplot:xlabel='comms_num'>
```



### 1.3 4)

```
[16]: corr_matrix = happy_data.corr()
```

```
[17]: corr_matrix['created']
```

```
[17]: score      0.072981  
comms_num    0.057357  
created      1.000000  
Name: created, dtype: float64
```

```
[18]: sns.heatmap(happy_data.corr(), annot=True, fmt='.3f')
```

```
[18]: <AxesSubplot:>
```

