МГТУ имени Баумана
Факультет «Информатика и Системы управления»
Кафедра «Системы обработки информации и управления»
Дисциплина «Теория машинного обучения»

Отчет по лабораторной работе №1

«Разведочный анализ данных. Исследование и визуализация данных»

Выполнил:
Студент группы ИУ5-61Б
Гапчук Л.Д.

Преподаватель:
Гапанюк Ю.Е.

Москва, 2020г.

Цель лабораторной работы: изучение различных методов визуализация данных.

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.

Импорт библиотек

In [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
```

Загрузка данных

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

2)Основные характеристики датасета

In [3]:
```
# Первые пять строк датасета
happy_data.head()
```

Out[3]:

	title	score	id	url	comms_num	created	body	timestamp
0	Health Canada approves AstraZeneca COVID-19 va...	7	lt74vw	https://www.canadaforums.ca/2021/02/health-can...	0	1.614400e+09	NaN	2021-02-27 06:33:45
1	COVID-19 in Canada: 'Vaccination passports' a ...	2	lsh0ij	https://www.canadaforums.ca/2021/02/covid-19-i...	1	1.614316e+09	NaN	2021-02-26 07:11:07
2	Coronavirus variants could fuel Canada's third...	6	lohlle	https://www.canadaforums.ca/2021/02/coronaviru...	0	1.613887e+09	NaN	2021-02-21 07:50:08
3	Canadian government to extend COVID-19 emergen...	1	lnptv8	https://www.canadaforums.ca/2021/02/canadian-g...	0	1.613796e+09	NaN	2021-02-20 06:35:13
4	Canada: Pfizer is 'extremely committed' to mee...	6	lkslm6	https://www.canadaforums.ca/2021/02/canada-pfi...	0	1.613468e+09	NaN	2021-02-16 11:36:28

In [4]:
```
# Размер датасета
happy_data.shape
```

Out[4]:
```
(1429, 8)
```

In [5]:
```
# Количество нулевых элементов
happy_data.isnull().sum()
```

Out[5]:
```
title          0
score          0
id             0
url          984
comms_num      0
created        0
body         365
timestamp      0
dtype: int64
```

In [6]:
```
 # Колонки и их типы данных
happy_data.dtypes
```

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

In [7]:
```
# Описание датасета
happy_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1429 entries, 0 to 1428
Data columns (total 8 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   title      1429 non-null   object 
 1   score      1429 non-null   int64  
 2   id         1429 non-null   object 
 3   url        445 non-null    object 
 4   comms_num  1429 non-null   int64  
 5   created    1429 non-null   float64
 6   body       1064 non-null   object 
 7   timestamp  1429 non-null   object 
dtypes: float64(1), int64(2), object(5)
memory usage: 89.4+ KB
```

In [8]:
```
# Статистические данные
happy_data.describe()
```

Out[8]:

	score	comms_num	created
count	1429.000000	1429.000000	1.429000e+03
mean	3.909727	1.977607	1.538045e+09
std	31.576009	17.054766	6.975050e+07
min	-13.000000	0.000000	1.389624e+09
25%	1.000000	0.000000	1.553489e+09
50%	1.000000	0.000000	1.565014e+09
75%	4.000000	1.000000	1.578769e+09
max	1184.000000	596.000000	1.616487e+09

In [9]:
```
# Удаляем столбец url
happy_data = happy_data.drop('url', axis = 1)
```

In [10]:
```
# Первые пять строк датасета
happy_data.head()
```

Out[10]:

	title	score	id	comms_num	created	body	timestamp
0	Health Canada approves AstraZeneca COVID-19 va...	7	lt74vw	0	1.614400e+09	NaN	2021-02-27 06:33:45
1	COVID-19 in Canada: 'Vaccination passports' a ...	2	lsh0ij	1	1.614316e+09	NaN	2021-02-26 07:11:07
2	Coronavirus variants could fuel Canada's third...	6	lohlle	0	1.613887e+09	NaN	2021-02-21 07:50:08
3	Canadian government to extend COVID-19 emergen...	1	lnptv8	0	1.613796e+09	NaN	2021-02-20 06:35:13
4	Canada: Pfizer is 'extremely committed' to mee...	6	lkslm6	0	1.613468e+09	NaN	2021-02-16 11:36:28

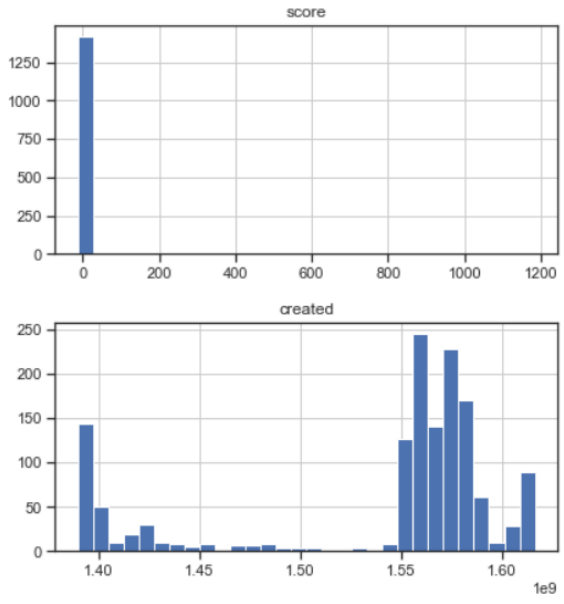
In [11]:
```
# Определим уникальные значения для целевого признака
happy_data['score'].unique()
```

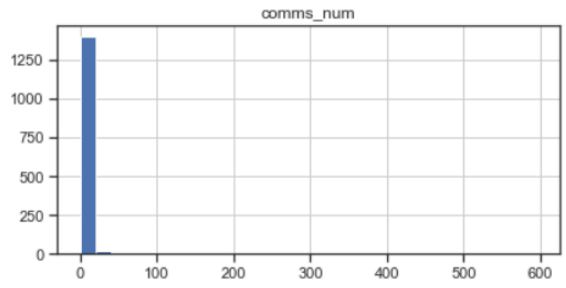
Out[11]:
```
array([   7,    2,    6,    1,    5,   10,    0,    3,   15,   11,   14,
          4,    9,   36,   20,   16,   18,   13,   29,   21,   24,   23,
          8,   25,   12,   31,   37,   22,   27,   26,   17,   19,   43,
       1184,   -1,   -3,   -2,  -13,   -4,   -5,   -6,  -10], dtype=int64)
```

3) Визуальное исследование датасета

In [12]:
```
# Гистограммы для всех признаков
happy_data.hist(bins=30, figsize = (15,7))
```

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

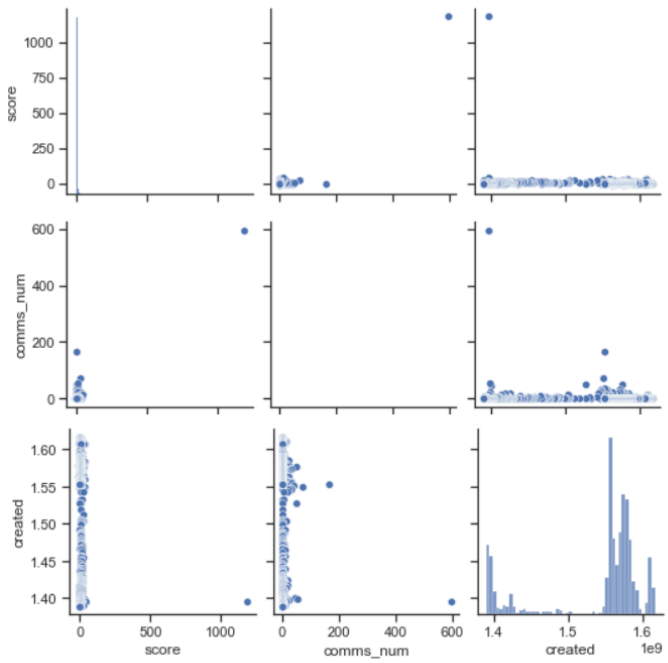




In [13]:
```
# Диаграммы рассеяние для всех признаков
plt.figure(figsize=(12,6))
sns.pairplot(happy_data)
```

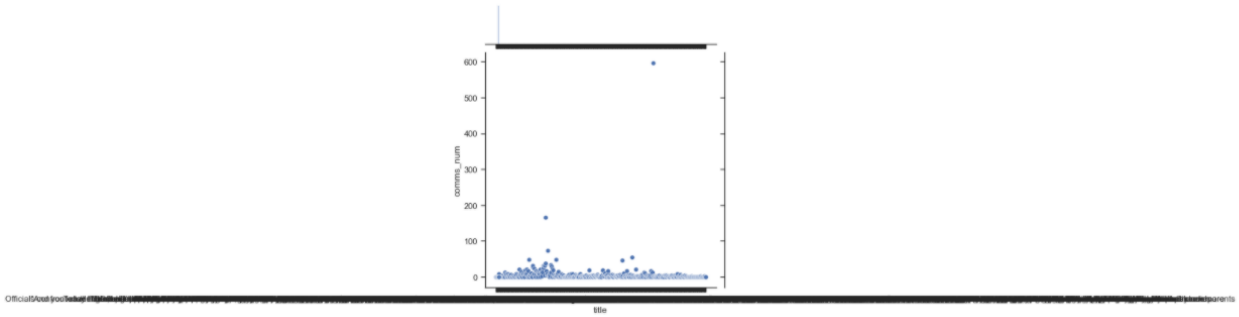
Out[13]:
```
<seaborn.axisgrid.PairGrid at 0x1ec21252dc0>

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



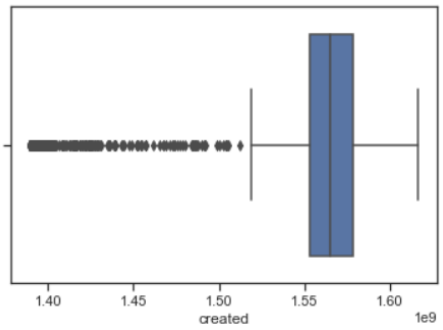
In [14]:
```
# Увеличенные диаграммы рассеяния для признаков, которые имеют зависимость с уровнем счастья
sns.jointplot(x = "title", y = "comms_num", kind="scatter", data = happy_data)
```

Out[14]:
```
<seaborn.axisgrid.JointGrid at 0x1ec2195e4f0>
```



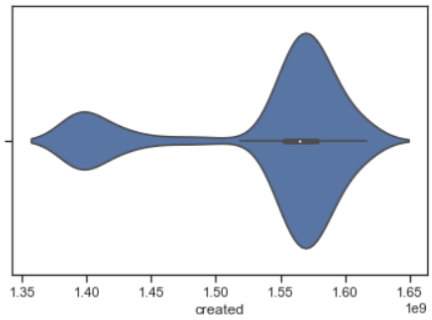
In [20]:
```
# Одномерное распределение вероятности
sns.boxplot(x=happy_data['created'])
```

Out[20]:
```
<AxesSubplot:xlabel='created'>
```



In [16]:
```
sns.violinplot(x=happy_data['created'])
```

Out[16]:
```
<AxesSubplot:xlabel='created'>
```



4) Корреляции признаков

In [17]:
```
corr_matrix = happy_data.corr()
```

In [18]:
```
corr_matrix['comms_num']
```

Out[18]:
```
score        0.924180
comms_num    1.000000
created     -0.101303
Name: comms_num, dtype: float64
```

In [19]:
```
sns.heatmap(happy_data.corr(), annot=True, fmt='.3f')
```

Out[19]:
```
<AxesSubplot:>
```