Лабораторная работа 0
Студент: Введенская П.К.
Группа: 80-304Б

Постановка задачи
Требуется сформировать/получить два набора данных соответствующие следующим критериям:

1. Один из датасетов должен представлять собой корпус документов. Язык, источник и тематика произвольна
2. Второй датасет должен содержать категориальные, количественные признаки. Для данного датасета определить предсказываемые признаки (для задачи регрессии и классфикации). Если такого признака нет, спроектировать

Данные датасеты будут в дальнейшем использованы в оставшихся лабораторных работах. По каждому датасету построить распределения признаков (в случае корпуса документов – построить распределение слов) и объяснить имеющуюся картину. Вычислить статистические характеристики признаков. Обнаружить и решить возможные проблемы с данными. Если решить данную проблему невозможно, объяснить почему.

Задание 1. Работа с корпусом текстовых доументов

Датасет: отзывы на игры в Steam

Нахождение 20 наиболее часто встречающихся слов

```
import numpy as np

import matplotlib

import matplotlib.pyplot as plt

import pandas as pd

import seaborn as sns

import nltk

import re

import string


from collections import Counter

sr = pd.read_csv('steam_reviews.csv', usecols=['review'])


rem = string.punctuation

pattern = r"[{}]".format(rem)

sr['review'] = sr['review'].str.replace(pattern, '')
```

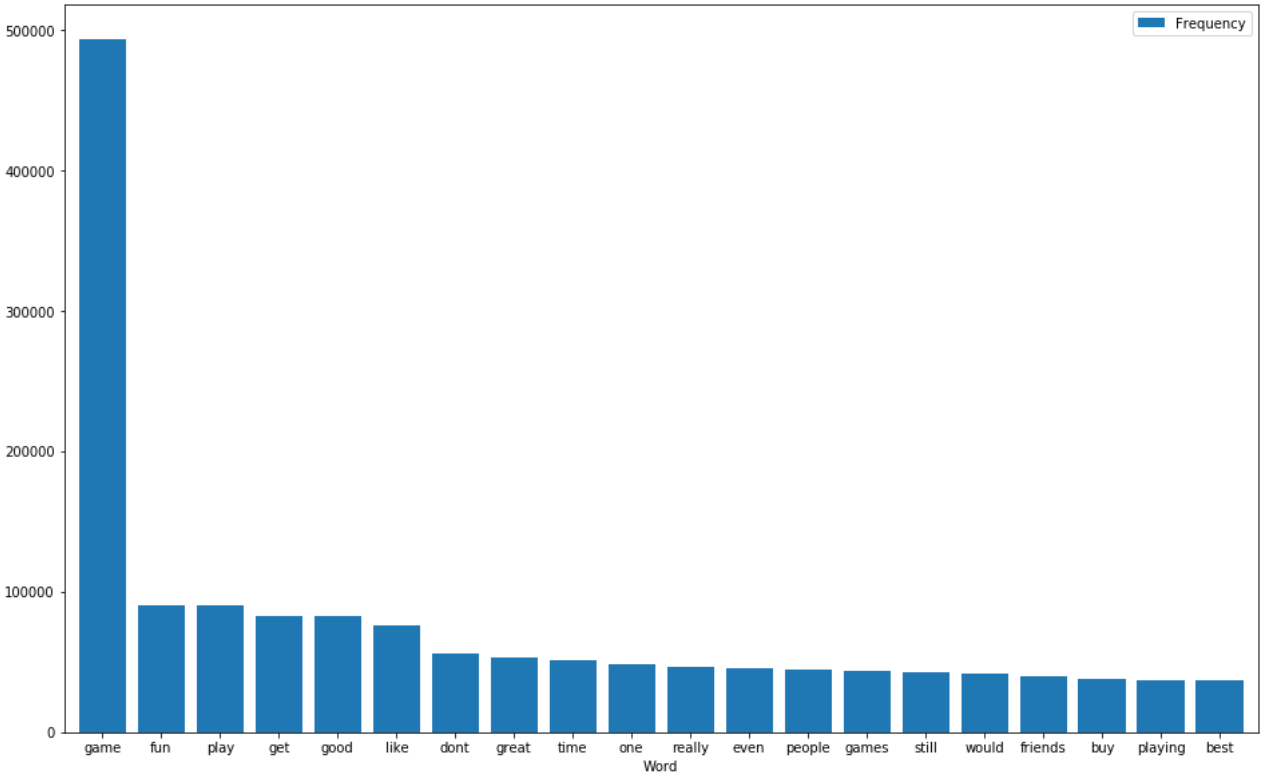
```
nltk.download('stopwords')
stopwords = nltk.corpus.stopwords.words('english')
RE_stopwords = r'\b(?:{})\b'.format('|'.join(stopwords))

words = (sr.review
        .str.lower()
        .replace([r'\|', RE_stopwords], [' ', ''], regex=True)
        .str.cat(sep=' ')
        .split()
)

top_N = 20
rslt = pd.DataFrame(Counter(words).most_common(top_N),
                    columns=['Word', 'Frequency']).set_index('Word')
print(rslt)
```

Word	Frequency
game	493483
fun	90462
play	90210
get	82884
good	82357
like	75874
dont	55946
great	53593
time	51047
one	48811
really	46550
even	46002
people	44396
games	43618
still	43032
would	41839
friends	40060
buy	37633
playing	37432
best	37165

Построение гистограммы распределения наиболее часто встречающихся слов

```
rslt.plot.bar(rot=0, figsize=(16,10), width=0.8)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6c3cc09ed0>
```



Pandas дает пользователю возможность легко удалить из текста знаки препинания и артикли. Но даже отфильтрованый список наиболее часто встречающихся слов не обязательно говорит что-либо примечательное о тональности отзывов, помимо того, что большинство из них - положительные.

Задание 2. Работа с датасетом с категориальными и количественными признаками

Датасет: The Human Freedom Index

Вывод первых пяти строк датасета

hfi = pd.read_csv('hfi_cc_2018.csv')

hfi.head()

Output View

	year	ISO_code	countries	region	pf_rol_procedural	pf_rol_civil	pf_rol_criminal	pf_rol	pf_ss_homicide	pf_ss_disappearances_disap	...	ef_regulation_business_bribes	ef_regulation_business_licensing
0	2016	ALB	Albania	Eastern Europe	6.661503	4.547244	4.666508	5.291752	8.920429	10.0	...	4.050196	7.324582
1	2016	DZA	Algeria	Middle East & North Africa	NaN	NaN	NaN	3.819566	9.456254	10.0	...	3.765515	8.523503
2	2016	AGO	Angola	Sub-Saharan Africa	NaN	NaN	NaN	3.451814	8.060260	5.0	...	1.945540	8.096776
3	2016	ARG	Argentina	Latin America & the Caribbean	7.098483	5.791960	4.343930	5.744791	7.622974	10.0	...	3.260044	5.253411
4	2016	ARM	Armenia	Caucasus & Central Asia	NaN	NaN	NaN	5.003205	8.808750	10.0	...	4.575152	9.319612

5 rows × 123 columns

Названия столбцов,их типы и краткая информация

print(hfi.columns)

```
Index([u'year', u'ISO_code', u'countries', u'region', u'pf_rol_procedural',
       u'pf_rol_civil', u'pf_rol_criminal', u'pf_rol', u'pf_ss_homicide',
       u'pf_ss_disappearances_disap',
       ...
       u'ef_regulation_business_bribes', u'ef_regulation_business_licensing',
       u'ef_regulation_business_compliance', u'ef_regulation_business',
       u'ef_regulation', u'ef_score', u'ef_rank', u'hf_score', u'hf_rank',
       u'hf_quartile'],
      dtype='object', length=123)
```

print(hfi.info())

<class 'pandas.core.frame.DataFrame'>

```
RangeIndex: 1458 entries, 0 to 1457
Columns: 123 entries, year to hf_quartile
dtypes: float64(119), int64(1), object(3)
memory usage: 1.4+ MB
None
```

Количество данных для каждого параметра, медиана, стандартное отклонение, минимум, кватили (0.25, 0.5, 0.75), максимум

hfi.describe()

	year	pf_rol_procedural	pf_rol_civil	pf_rol_criminal	pf_rol	pf_ss_homicide	pf_ss_disappearances_disap	pf_ss_disappearances_violent
count	1458.000000	880.000000	880.000000	880.000000	1378.000000	1378.000000	1369.000000	1378.000000
mean	2012.000000	5.589355	5.474770	5.044070	5.309641	7.412980	8.341855	9.519458
std	2.582875	2.080957	1.428494	1.724886	1.529310	2.832947	3.225902	1.744673
min	2008.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2010.000000	4.133333	4.549550	3.789724	4.131746	6.386978	10.000000	10.000000
50%	2012.000000	5.300000	5.300000	4.575189	4.910797	8.638278	10.000000	10.000000
75%	2014.000000	7.389499	6.410975	6.400000	6.513178	9.454402	10.000000	10.000000
max	2016.000000	9.700000	8.773533	8.719848	8.723094	9.926568	10.000000	10.000000

8 rows × 120 columns

Средние характеристики для страны (Россия)

hfi[hfi['countries'] == 'Russia'].mean()

```
pf_rol_procedural                 3.770599
pf_rol_civil                      4.948302
pf_rol_criminal                   3.696139
pf_rol                            4.147344
pf_ss_homicide                    4.977812
pf_ss_disappearances_disap        0.555556
pf_ss_disappearances_violent      9.214397
pf_ss_disappearances_organized    5.000000
pf_ss_disappearances_fatalities   9.719299
```

```
pf_ss_disappearances_injuries            9.657984
pf_ss_disappearances                     6.829447
pf_ss_women_fgm                         10.000000
pf_ss_women_missing                     10.000000
pf_ss_women_inheritance_widows          10.000000
pf_ss_women_inheritance_daughters       10.000000
pf_ss_women_inheritance                 10.000000
pf_ss_women                             10.000000
pf_ss                                    7.269086
pf_movement_domestic                     4.444444
pf_movement_foreign                      2.222222
pf_movement_women                       10.000000
pf_movement                              5.555556
pf_religion_estop_establish              3.750000
pf_religion_estop_operate                7.500000
pf_religion_estop                        5.416667
pf_religion_harassment                   8.594815
pf_religion_restrictions                 4.071605
pf_religion                              6.027695
pf_association_association               4.166667
                                             ...
ef_trade_black                          10.000000
ef_trade_movement_foreign                4.188675
ef_trade_movement_capital                5.641026
ef_trade_movement_visit                  2.409181
ef_trade_movement                        4.079627
ef_trade                                 6.074130
ef_regulation_credit_ownership           5.000000
ef_regulation_credit_private             9.339980
ef_regulation_credit_interest            9.222222
ef_regulation_credit                     7.854068
ef_regulation_labor_minwage              4.883333
ef_regulation_labor_firing               4.997522
ef_regulation_labor_bargain              6.935498
ef_regulation_labor_hours                6.888889
ef_regulation_labor_dismissal            8.570993
ef_regulation_labor_conscription         3.000000
ef_regulation_labor                      5.879373
ef_regulation_business_adm               2.944651
ef_regulation_business_bureaucracy       5.563153
ef_regulation_business_start             9.392901
ef_regulation_business_bribes            3.619552
ef_regulation_business_licensing         4.939130
ef_regulation_business_compliance        7.563840
ef_regulation_business                   5.670538
ef_regulation                            6.467993
ef_score                                 6.598889
ef_rank                                 96.000000
hf_score                                 6.400402
hf_rank                                112.555556
hf_quartile                              3.111111
Length: 120, dtype: float64
```

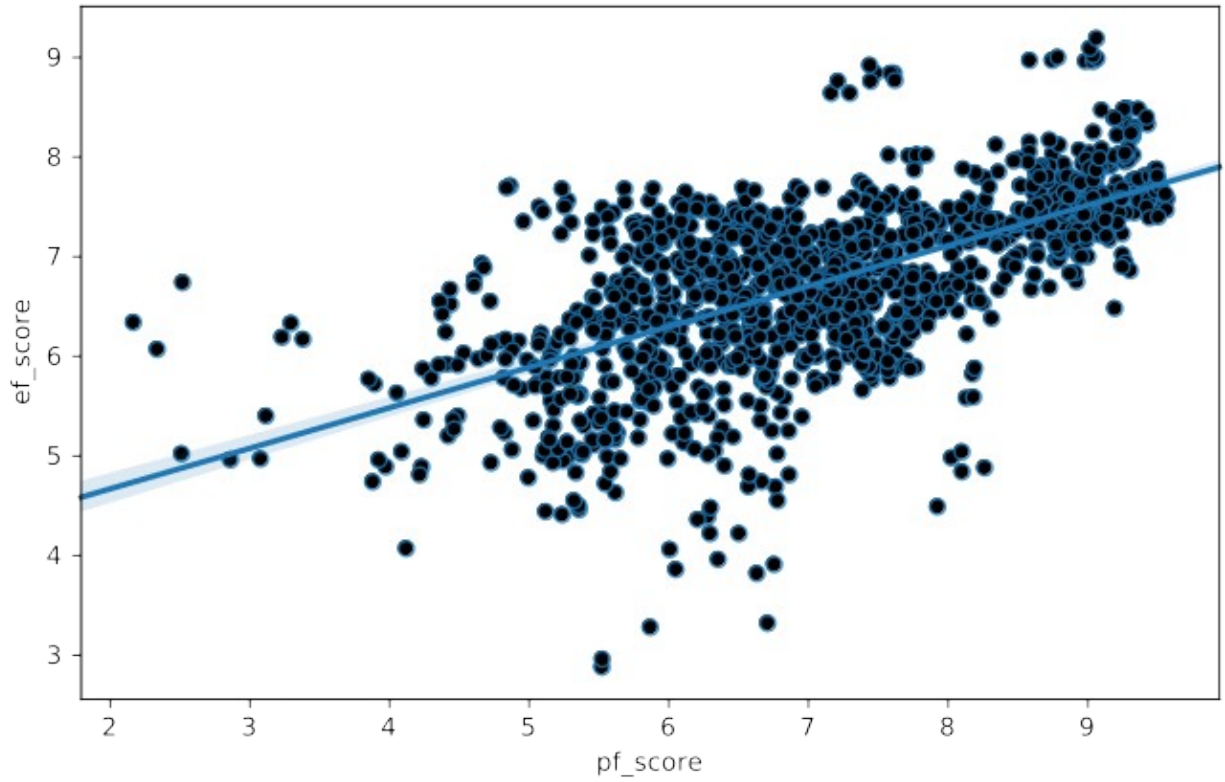
Значение, которое индекс человеческой свободы не превысит с вероятностью 0.5

np.nanpercentile(hfi['hf_score'], 50)

```
6.923840432999999
```

Диаграмма зависимости индекса экономической свободы от индекса личной свободы

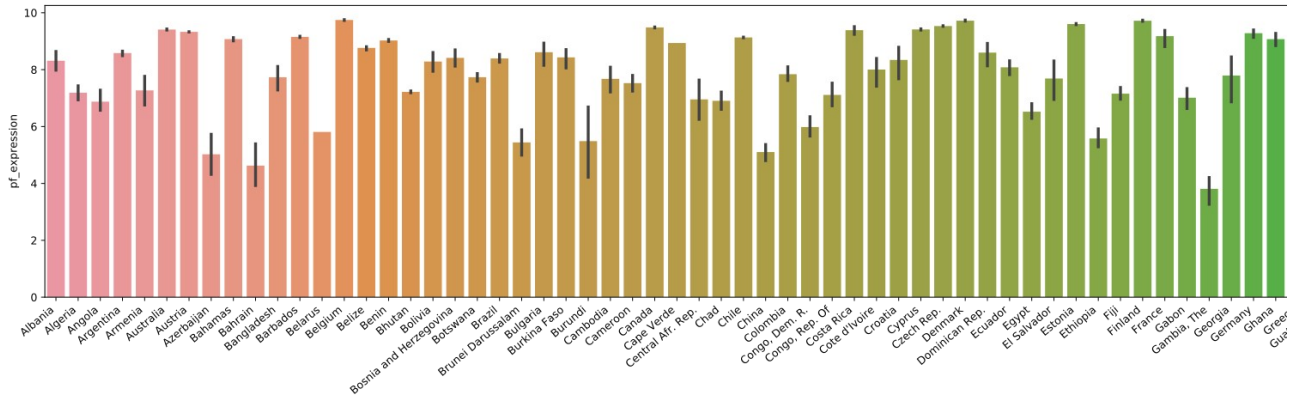
```
%matplotlib inline
%config InlineBackend.figure_format = 'svg'
from pylab import rcParams
rcParams['figure.figsize'] = 8, 5
sns.regplot(x='pf_score', y='ef_score', data=hfi)
```

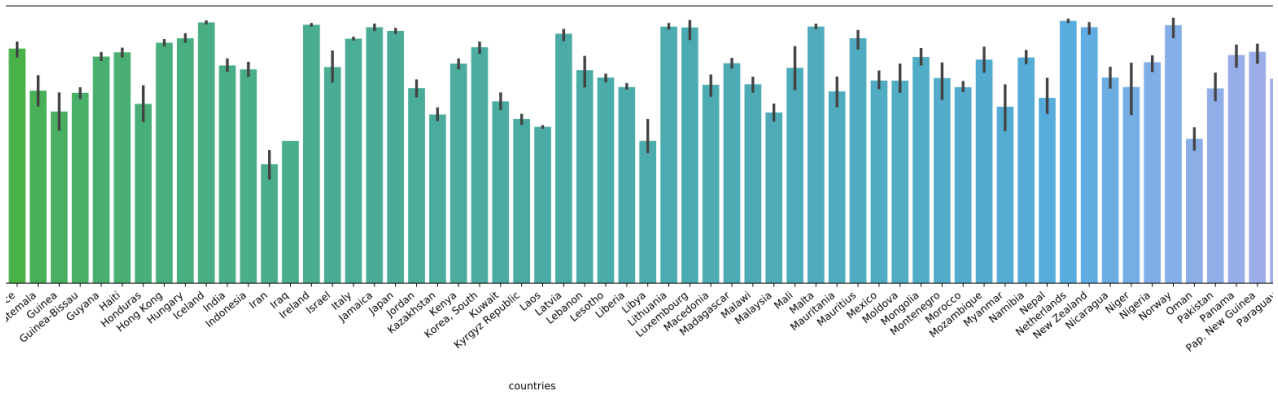


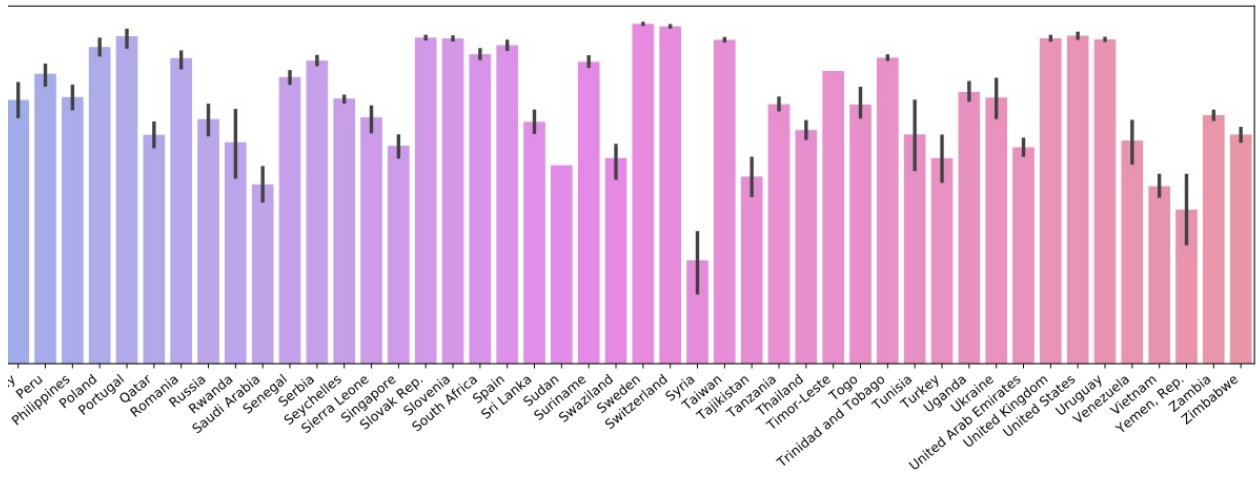
Из графика можно увидеть сильную зависимость.

Гистограмма зависимости свободы слова от страны

```
from pylab import rcParams
rcParams['figure.figsize'] = 60, 5
ax = sns.barplot(data = hfi, x = 'countries', y = 'pf_expression')
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right")
```







Из графика видно распределение индекса свободы слова. Он не обязательно идеально соответствует реальности, но дает общее представление об очевидных проблемах.

Вывод

Базовые инструменты JupyterLab позволяют определить очевидные общие тенденции датасета, но этого недостаточно для полноценного анализа данных. Также стоит отметить, как легко с их помощью отредактировать и подготовить датасет к использованию.