Университет ИТМО

Практическая работа №3

по дисциплине «Визуализация и моделирование»

Автор: Власов Матвей Иванович
Поток: ВИМ 1.1
Группа: K3240
Факультет: ИКТ
Преподаватель: Чернышева А. В.

Санкт-Петербург, 2021 г.

Датасет

Для дальнейшей работы выбран датасет: "Russian Presidental Elections 2018 Voting Data": https://www.kaggle.com/valenzione/russian-presidental-elections-2018-voting-data

Описание датасета

В нашем датасете содержится информация об итогах выборов 2018 года, полученная с официального сайта ЦИК РФ.

Названия большинства стобцов исходного датасета представим в графе "Описание а сами названия сократим для удобства в дальнейшем:

Столбец	Описание	Тип	Шкала	Предобработка
PS_ID	Идентификатор избирательного участка	INT	Номинальная	Не требуется
REGION	Название региона	STRING	Номинальная	Убрать цифры в начале строки (если есть)
SUBREGION	Название округа	STRING	Номинальная	Убрать цифры в начале строки (если есть)
N_ALL	Число избирателей, включенных в список избирателей	INT	Относительная	Не требуется
N_GIVEN	Число избирательных бюллетеней, полученных участковой избирательной комиссией	INT	Относительная	Удалить после подсчёта N_VOTED
N_EARLY	Число избирательных бюллетеней, выданных избирателям, проголосовавшим досрочно	INT	Относительная	Не требуется
N_IN	Число избирательных бюллетеней, выданных в помещении для голосования в день голосования	INT	Относительная	Удалить (избыточные данные)
N_OUT	Число избирательных бюллетеней, выданных вне помещения для голосования в день голосования	INT	Относительная	Не требуется
N_LEFT	Число погашенных избирательных бюллетеней	INT	Относительная	Удалить после подсчёта N_VOTED
N_PORTABLE	Число избирательных бюллетеней в переносных ящиках для голосования	INT	Относительная	Удалить (избыточные данные)
N_STATIC	Число бюллетеней в стационарных ящиках для голосования	INT	Относительная	Удалить (избыточные данные)
N_INVALID	Число недействительных избирательных бюллетеней	INT	Относительная	Не требуется
N_VALID	Число действительных избирательных бюллетеней	INT	Относительная	Удалить (избыточные данные)
N_LOST	Число утраченных избирательных бюллетеней	INT	Относительная	Удалить (незначительные данные)
N_UNUSED	Число избирательных бюллетеней, не учтенных при получении	INT	Относительная	Удалить (незначительные данные)

Столбец	Описание	Тип	Шкала	Предобработка
BABURIN	Бабурин Сергей Николаевич	INT	Относительная	Не требуется
GRUDININ	Грудинин Павел Николаевич	INT	Относительная	Не требуется
ZHIRINOVSKY	Жириновский Владимир Вольфович	INT	Относительная	Не требуется
PUTIN	Путин Владимир Владимирович	INT	Относительная	Не требуется
SOBCHAK	Собчак Ксения Анатольевна	INT	Относительная	Не требуется
SURAYKIN	Сурайкин Максим Александрович	INT	Относительная	Не требуется
TITOV	Титов Борис Юрьевич	INT	Относительная	Не требуется
YAVLINSKY	Явлинский Григорий Алексеевич	INT	Относительная	Не требуется

Задачи, решаемые при помощи датасета

1. Визуализация результатов выборов.
2. Анализ данных на предмет возможных фальсификаций.
3. Выявление особенностей голосования в различных регионах.

Гипотезы

1. В Москве и Санкт-Петербурге ниже, чем в среднем, и процент за Путина, и явка (в больших городах более образованное население, а также большое количество наблюдателей, что затрудняет фальсификации).
2. В Крыму высокая явка и поддержка президента (из-за присоединения территории).
3. В регионах с большим количеством избирателей, проголосовавших досрочно, процент за Путина выше, чем в среднем (голоса, поданные досрочно, легче сфальсифицировать).
4. Есть регионы, где победил не Путин (у Грудинина в среднем больше 11 процентов - вполне возможно, что где-то он набрал больше Путина).

Предобработка данных

Посмотрим, какие столбцы у нас есть

```
list(df.columns)
```

```
['PS_ID',
 'REGION',
 'SUBREGION',
 'N_ALL',
 'N_GIVEN',
 'N_EARLY',
 'N_IN',
 'N_OUT',
 'N_LEFT',
 'N_PORTABLE',
 'N_STATIC',
 'N_INVALID',
 'N_VALID',
 'N_LOST',
 'N_UNUSED',
 'BABURIN',
 'GRUDININ',
 'ZHIRINOVSKY',
 'PUTIN',
 'SOBCHAK',
 'SURAYKIN',
 'TITOV',
 'YAVLINSKY',
 'N_VOTED']
```

Заметим, что в таблице много столбцов с избыточными/ненужными данными, а именно:
N_GIVEN, N_LEFT - могут использоваться только для подсчёта N_VOTED, что мы уже сделали
N_IN - вряд ли будем использовать, но в случае необходимости посчитаем как N_VOTED - N_OUT - N_EARLY
N_PORTABLE, N_STATIC - почти полностью совпадают с N_OUT и N_IN, нет необходимости их хранения
N_VALID - можно посчитать как N_VOTED - N_INVALID
N_UNUSED, N_LOST - почти всегда равны 0 (см. ниже) и не влияют на общую картину

```
df['N_UNUSED'].sum()
```

```
104
```

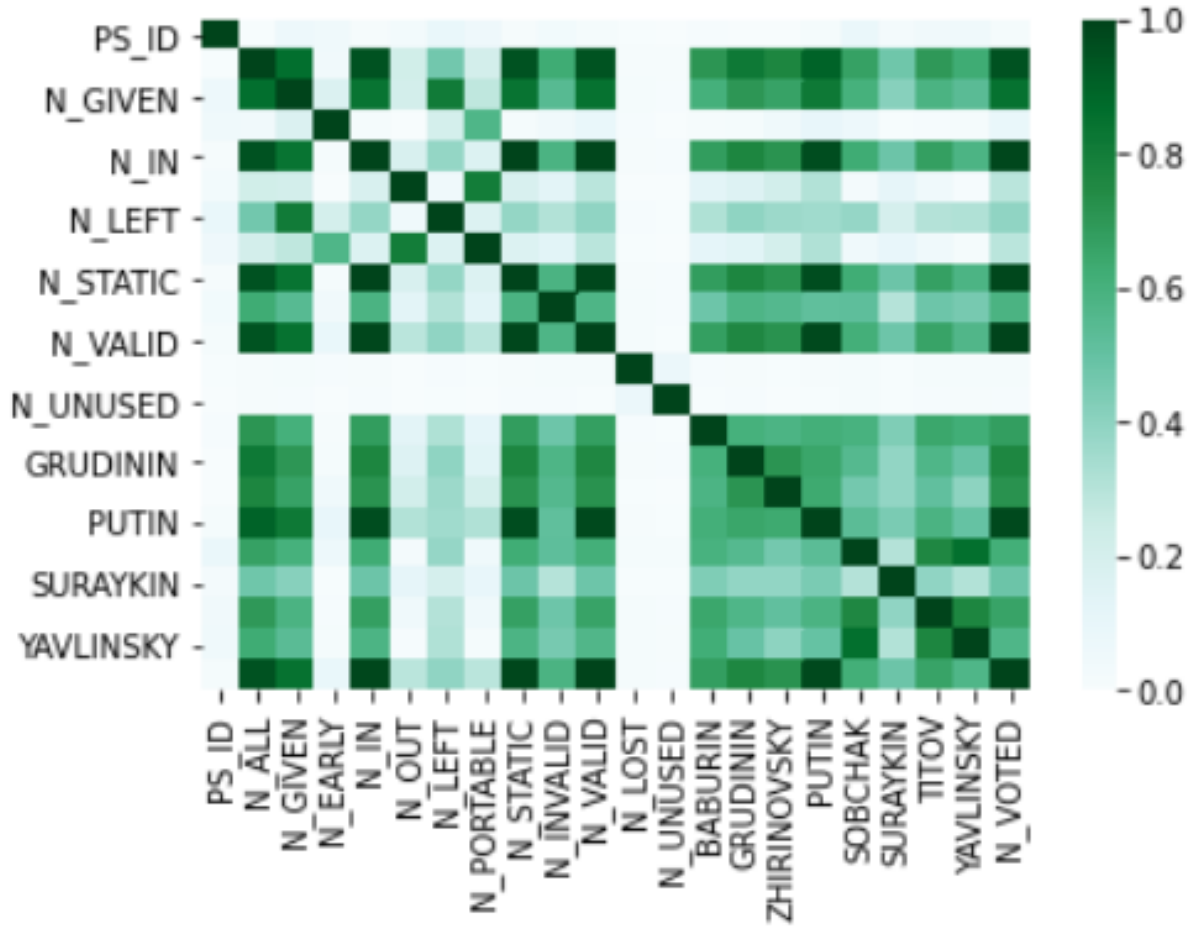
```
df['N_LOST'].sum()
```

```
1047
```

Покажем корреляцию столбцов

```
corr = df.corr()
sns.heatmap(corr, cmap="BuGn", vmin=0)
```

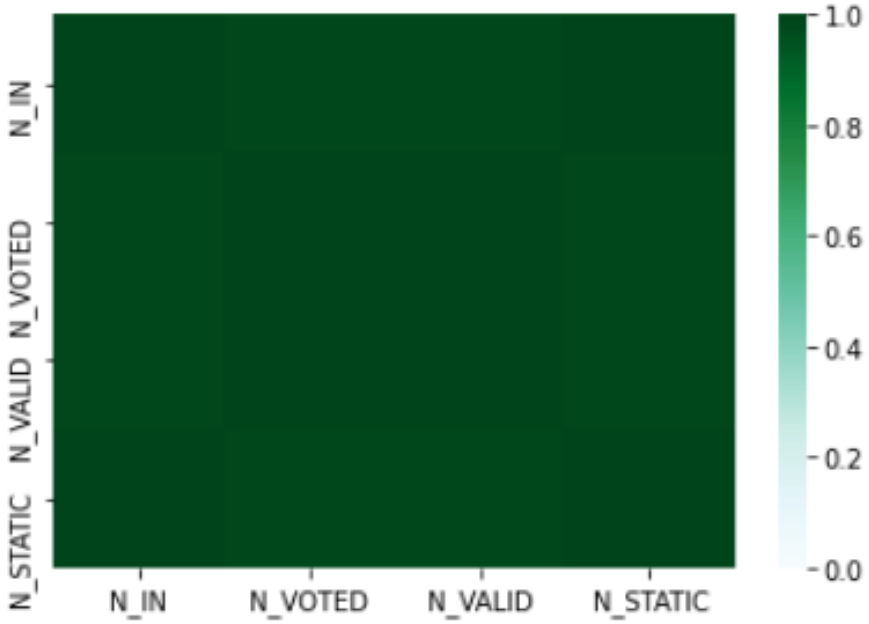
```
<AxesSubplot:>
```



Рассмотрим ближе интересующие нас стобцы

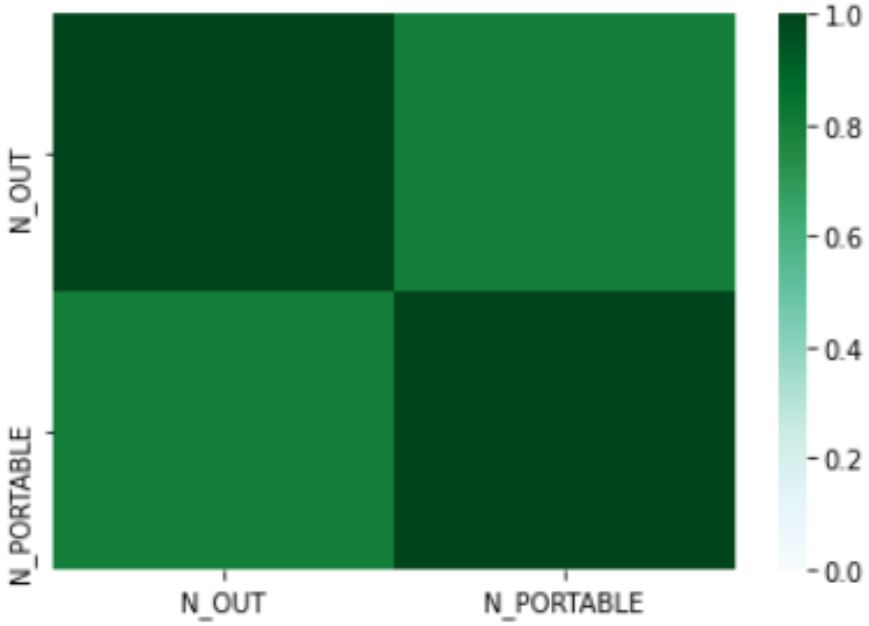
```
corr = df[['N_IN', 'N_VOTED', 'N_VALID', 'N_STATIC']].corr()
sns.heatmap(corr, cmap="BuGn", vmin=0)
```

```
<AxesSubplot:>
```




```
corr = df[['N_OUT', 'N_PORTABLE']].corr()
sns.heatmap(corr, cmap="BuGn", vmin=0)
```

```
<AxesSubplot:>
```



```
df = df.drop(columns=['N_GIVEN', 'N_LEFT', 'N_IN', 'N_PORTABLE', 'N_STATIC', 'N_VALID', 'N_UNUSED', 'N_LOST'], erro
df_regions = df_regions.drop(columns=['N_GIVEN', 'N_LEFT', 'N_IN', 'N_PORTABLE', 'N_STATIC', 'N_VALID', 'N_UNUSED',
DF_REGIONS_LIST = [i for i in range(df_regions.shape[0])]
df_regions
```

	REGION	N_ALL	N_EARLY	N_OUT	N_INVALID	BABURIN	GRUDININ	ZHIRINOVSKY	PUTIN	SOBCHAK	SURAYKIN	TITOV	YAVLINSKY	N_VO
0	Город Байконур (Республика Казахстан)	14575	0	168	104	32	1026	628	7568	130	51	50	70	
1	Территория за пределами РФ	483957	53482	18026	5801	2010	23871	8224	403306	19203	1439	3079	7433	47
2	Республика Адыгея (Адыгея)	336720	0	15183	2647	1165	28711	8923	203095	2060	1317	1088	1197	25
3	Республика Алтай	158891	1186	4630	974	446	21259	5376	72674	929	431	340	483	10
4	Республика Башкортостан	3045698	0	111258	18506	15845	277797	115635	1784626	28983	20429	15891	19390	229
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
82	Еврейская автономная область	128844	106	4123	1331	501	14066	7387	52374	771	404	407	369	7
83	Ненецкий автономный округ	39470	6116	579	267	185	3397	2482	17863	448	135	172	158	2
84	Ханты-Мансийский автономный округ - Югра	1130343	26633	15804	10824	3916	94785	53569	600404	10884	4095	4363	5060	78
85	Чукотский автономный округ	33541	2531	874	293	115	1616	2018	22709	358	142	180	160	2
86	Ямало-Ненецкий автономный округ	370823	41135	4008	2616	1052	19488	19409	291409	2394	1794	1154	1367	34

87 rows × 15 columns