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

In [4]:
```
list(df.columns)
```

Out[4]:
```
['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

In [5]:
```
df['N_UNUSED'].sum()
```

Out[5]: 104

In [6]:
```
df['N_LOST'].sum()
```

Out[6]: 1047

Удалим ненужные столбцы

In [7]:
```
df = df.drop(columns=['N_GIVEN', 'N_LEFT', 'N_IN', 'N_PORTABLE', 'N_STATIC', 'N_VALID', 'N_UNUSED', 'N_LOST'], erro
df
```

Out[7]:

	PS_ID	REGION	SUBREGION	N_ALL	N_EARLY	N_OUT	N_INVALID	BABURIN	GRUDININ	ZHIRINOVSKY	PUTIN	SOBCHAK	SURAYKIN	TITOV
0	8140	98 Город Байконур (Республика Казахстан)	98 Город Байконур (Республика Казахстан)	2132	0	11	9	4	176	79	1136	30	9	5
1	8141	98 Город Байконур (Республика Казахстан)	98 Город Байконур (Республика Казахстан)	2207	0	14	14	2	128	87	1214	19	4	7
2	8142	98 Город Байконур (Республика Казахстан)	98 Город Байконур (Республика Казахстан)	2249	0	7	27	5	171	94	1162	17	3	12
3	8143	98 Город Байконур (Республика Казахстан)	98 Город Байконур (Республика Казахстан)	1769	0	48	20	5	98	72	882	17	8	5
4	8144	98 Город Байконур (Республика Казахстан)	98 Город Байконур (Республика Казахстан)	1880	0	13	10	7	124	105	902	7	9	10

У нас есть два региона, названия которых в датасете начинаются с цифры. Для удобства эти цифры удалим

In [8]:
```
def remove_digits(text):
    if text[0].isdigit():
        temp = list(map(lambda x: '' if x.isdigit() else x, text))
        text = ''.join(temp).strip()
    return text
```

In [9]:
```
df['REGION'] = df['REGION'].apply(remove_digits)
df['SUBREGION'] = df['SUBREGION'].apply(remove_digits)
df['REGION'].unique()[:2]
```

Out[9]:
```
array(['Город Байконур (Республика Казахстан)',
       'Территория за пределами РФ'], dtype=object)
```