

Анализ данных в Python

Домашнее задание 3

Дедлайн:

16 мая 23.59. Архив с файлами отправляется через online.hse.ru

Поздний дедлайн:

23 мая 23.59 (штраф 1 балл)

Тестируем гипотезу, влияет ли снижение количества конкурентных штатов во время президентских выборах в США на уровень знаний граждан о политике.

Обратите внимание, что задание принимается архивом, в котором лежит блокнот и два или пять файлов csv (те, которые вы создаете или изменяете во время работы).

Оригинальные файлы, пожалуйста, загружайте по ссылкам с git, а не с компьютера.

Всего можно набрать 13 баллов (бонусные баллы идут в ведомость в счет этой домашней работы)

Общее оформление (1 балл)

Общее оформление кода, графиков, написание промежуточных выводов.

Штрафы по 0.25 балла за каждый случай отсутствия заголовков графиков, шкалы, неработающий код, отсутствие выводов в ячейках markdown и т.д.

Сбор данных (3 балла) (бонусные баллы)

3 балла (каждая корректно сохраненная таблица + работающий код для ее парсинга оценивается в 1 балл)

Пройти по [ссылке](#). В статье есть три таблицы. Нужно написать код, который сможет обратиться к странице, достать информацию для каждой таблицы, преобразовать ее в pandas dataframe и сохранить каждую в csv. Приложите эти три файла в архив с выполненным заданием.

Подсказка:

1. Текст страницы доставайте с помощью [requests.get\(\)](#). Вам нужен именно атрибут **text**.

2. BeautifulSoup тут не поможет. С помощью `.find()` для каждой таблицы найдите ее начало и конец и сделайте срез.
3. В таблице колонки разделены пробельными символами (' '), которые повторяются много раз подряд. Замените их таким образом, чтобы вместо серии разделителей везде остался один. Например, подумайте про `while` и `.replace()`
4. Далее посмотрите как организованы строки и найдите разделители рядов. Запустите цикл, почистите каждую строку от лишних символов и разбейте по разделителю. Создайте список списков.
5. Из списка списков сделайте датафрейм – вы восхитительны!

Table 1. Campaign Events between September 5 and November 4, 2008*

Rank	State	Events	% of total	Cumulative %
1	OH	62	20.7	20.7
2	FL	46	15.3	36.0
3	PA	40	13.3	49.3
4	VA	23	7.7	57.0
5	MO	21	7.0	64.0
6	CO	20	6.7	70.7
7	NC	15	5.0	75.7
8	NV	12	4.0	79.7
9	NH	12	4.0	83.7
10	MI	10	3.3	87.0
11	IN	9	3.0	90.0
12	NM	8	2.7	92.7
13	WI	8	2.7	95.3
14	IA	7	2.3	97.7
15	MN	2	0.7	98.3
16	ME	2	0.7	99.0
17	DC	1	0.3	99.3
18	TN	1	0.3	99.7
19	WV	1	0.3	100.0

Заголовок таблицы (Table 1 и т.д.) сохранять не нужно.

Как таблица должна выглядеть:

	Rank	State	Events	% of total	Cumulative %
0	1	OH	62	20.7	20.7
1	2	FL	46	15.3	36.0
2	3	PA	40	13.3	49.3
3	4	VA	23	7.7	57.0
4	5	MO	21	7.0	64.0
5	6	CO	20	6.7	70.7
6	7	NC	15	5.0	75.7
7	8	NV	12	4.0	79.7
8	9	NH	12	4.0	83.7
9	10	MI	10	3.3	87.0
10	11	IN	9	3.0	90.0
11	12	NM	8	2.7	92.7
12	13	WI	8	2.7	95.3
13	14	IA	7	2.3	97.7
14	15	MN	2	0.7	98.3
15	16	ME	2	0.7	99.0
16	17	DC	1	0.3	99.3
17	18	TN	1	0.3	99.7
18	19	WV	1	0.3	100.0

Table 2. Ad spending by candidates from September 24 to November 4, 2008*

Rank	State	Ad \$ spent	% of total	Cumulative %
1	FL	\$29,249,985	18.2	18.2
2	PA	\$24,903,675	15.5	33.7
3	OH	\$16,845,415	10.5	44.1
4	VA	\$16,634,262	10.3	54.5
5	NC	\$9,556,598	5.9	60.4
6	IN	\$8,964,817	5.6	66.0
7	WI	\$8,936,200	5.6	71.5
8	MO	\$7,970,313	5.0	76.5
9	CO	\$7,944,875	4.9	81.4
10	NV	\$7,108,542	4.4	85.9
11	MI	\$5,780,198	3.6	89.5
12	MN	\$4,262,784	2.6	92.1
13	IA	\$3,713,223	2.3	94.4
14	NM	\$3,134,146	1.9	96.4
15	NH	\$2,924,839	1.8	98.2
16	MT	\$971,040	0.6	98.8
17	ME	\$832,204	0.5	99.3

Заголовок таблицы (Table 1 и т.д.) сохранять не нужно.

Как таблица должна выглядеть:

	Rank	State	Ad \$ spent	% of total	Cumulative %
0	1	FL	\$29,249,985	18.2	18.2
1	2	PA	\$24,903,675	15.5	33.7
2	3	OH	\$16,845,415	10.5	44.1
3	4	VA	\$16,634,262	10.3	54.5
4	5	NC	\$9,556,598	5.9	60.4
5	6	IN	\$8,964,817	5.6	66.0
6	7	WI	\$8,936,200	5.6	71.5
7	8	MO	\$7,970,313	5.0	76.5
8	9	CO	\$7,944,875	4.9	81.4
9	10	NV	\$7,108,542	4.4	85.9
10	11	MI	\$5,780,198	3.6	89.5
11	12	MN	\$4,262,784	2.6	92.1
12	13	IA	\$3,713,223	2.3	94.4
13	14	NM	\$3,134,146	1.9	96.4
14	15	NH	\$2,924,839	1.8	98.2
15	16	MT	\$971,040	0.6	98.8
16	17	ME	\$832,204	0.5	99.3

Table 3. Increasing partisanship of states over time

Year	States within each two-party partisanship range		
	(60%+) Landslide	(55%+) Comfortable	(under 55%) Competitive
1960	6	11	33
1964	12	15	24
1968	6	14	31
1972	7	16	28
1976	8	10	33
1980	10	11	30
1984	5	16	30
1988	2	23	26
1992	3	16	32
1996	10	14	27
2000	15	14	22
2004	15	18	18
2008	16	19	16

Заголовок таблицы (Table 1 и т.д.) сохранять не нужно, сохраняем только один уровень заголовков (Year, Landslide, Comfortable, Competitive).

Как таблица должна выглядеть:

	Year	Landslide	Comfortable	Competitive
0	1960	6	11	33
1	1964	12	15	24
2	1968	6	14	31
3	1972	7	16	28
4	1976	8	10	33
5	1980	10	11	30
6	1984	5	16	30
7	1988	2	23	26
8	1992	3	16	32
9	1996	10	14	27
10	2000	15	14	22
11	2004	15	18	18

Визуализация собранных данных (2 балла)

Если в прошлом шаге вы не собрали данные, загрузите их по [ссылке](#).

а. **1 балл**

Загрузите файлы csv, которые вы сохранили в предыдущем пункте, как pandas dataframes.

В plotly построить интерактивную визуализацию для таблицы 3 из статьи, которая будет отражать, как на протяжении лет менялось соотношение landslide, comfortable и competitive штатов (составной барчарт). Коротко описать тренды.

б. **1 балл (0.5 балла за каждую визуализацию)**

Постройте в plotly визуализации для таблиц 1 и 2, которые отражают долю приведенных штатов в рекламном бюджете/политических событиях. Для таблицы 1 используйте столбчатый график. Для таблицы 2 – круговую диаграмму. Обратите внимание, что в таблице 2 в сумме нет 100% и перед тем как строить круговую диаграмму, нужно решить эту проблему. Напишите вывод к каждой визуализации.

Работа с таблицами (7 баллов)

а. **0.6 балла (0.2 за каждую переменную)**

По ссылке загрузите [данные о ходе кампании в 2008 году](#).

Создайте новые переменные: общее количество визитов депутатов (используйте переменные _Visits), общее количество расходов на рекламу по радио и

телевидению (переменные `_Air`), расходы на рекламу на душу населения (переменные `_Spend` и `StatePop`).

b. 2 балла

По ссылке загрузите [данные опроса National Annenberg Election Survey](#). В нем мы создадим переменную `political knowledge`. С помощью [codebook](#) найдите, какие переменные измеряют `political knowledge` и штат проживания респондента и выделите их из данных. Разберитесь как закодированы переменные `political knowledge` и приведите их к виду 1 – правильный ответ, 0 – все остальное (нет ответа/неправильный ответ и т.д.). Создайте новую переменную, которая измеряет средний уровень `political knowledge` для респондента (сумма правильных ответов разделенная на количество вопросов).

Правильные ответы на вопросы:

1. Supreme Court
2. Two-thirds
3. Democratic
4. Nominated by the president and confirmed by Senate

Затем создайте новую переменную, которая для каждого респондента будет измерять средний уровень `political knowledge` для штата, в котором он проживает. Это можно сделать разными способами, но рекомендую начать с метода объекта `groupby transform` (пример есть в самом конце [этого блокнота](#)).

c. 2 балла

С помощью текстового файла [state_of_residency.txt](#) перекодируйте колонку с номерами штатов в текстовый формат (двухбуквенный код).

Итоговый датафрейм должен выглядеть так: колонка, отвечающая за штат (двухбуквенный код); ответы на четыре вопроса (1/0); колонка со средним РК для респондента; колонка со средним РК для штата респондента. Сохраните получившийся датафрейм как `csv` и приложите к архиву с вашим блокнотом.

d. 0.4 балла

В датафрейм, в котором находится информация о ходе кампаний (файл [2008_campaign_data.xlsx](#)), добавьте данные о среднем `political knowledge` для каждого штата. Соединить таблицы можно с помощью [pandas.merge\(\)](#). Получившийся датафрейм сохраните в `csv` и тоже приложите к архиву.

e. 1 балл

В `plotly` создать две тепловые карты (географические – хороплеты) США для трат на рекламу по штатам и визитов кандидатов в штаты для 2008 года (новые переменные, созданные в п. а). Написать вывод о трендах.

[Ссылка](#) на блокнот `Plotly` с примером.

[Ссылка](#) на документацию.

f. **1 балл**

В plotly построить пузырьковый график, который будет отображать наличие или отсутствие связи между расходами на рекламу и визитами кандидата в штат во время предвыборной кампании 2008 года (новые переменные, созданные в п. а) и средним уровне знаний о политике жителей этих штатов (три переменных на одном графике – y, x и цвет или размер). Написать общий вывод.