Рогович Татьяна

Анализ данных в Python (политология)

Дедлайн: 1 декабря 23.59

Домашнее задание 1

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 пять файлов csv (те, которые вы создаете или изменяете во время работы).

Оригинальные файлы, пожалуйста, загружайте по ссылкам с git.

Данные:
https://github.com/rogovich/Data/tree/master/docs/Plotly_HW

1 балл
Общее оформление кода, графиков, написание промежуточных выводов.
Штрафы по 0.25 балла за каждый случай отсутствия заголовков, неработающий код, ошибки в коде (например, ненужные преобразования переменных), отсутствие выводов в ячейках markdown и т.д.

1.
    a. Пройти по ссылке
    https://nbviewer.jupyter.org/github/rogovich/Data/blob/master/docs/Plotly_HW/FairVote%20-%20Press%20Room.html
    В статье есть три таблицы, нужно написать код, который сможет обратиться к странице, достать информацию для каждой таблицы, преобразует ее в pandas dataframe и сохраните каждую в csv, положите эти три файла в архив с выполненным заданием.

    Названия колонок в таблице можно задать вручную (с помощью парсинга тоже можно, но иногда слишком много трудозатрат, поэтому, если не понимаете, как сделать легко, то просто делайте вручную :)! В таблице три достаточно одного уровня заголовка — Year, Landslide, Comfortable, Competitive

2,4 балла (каждая корректно сохраненная таблица + работающий код для ее парсинга оценивается в 0.8 балла)

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

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

Table 3. Increasing partisanship of states over time

	States within each two-party partisanship range		
	(60%+)	(55%+)	(under 55%)
Year	Landslide	Comfortable	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

b. Загрузите файлы csv, которые вы сохранили в предыдущем пункте для каждой таблицы. В plotly построить интерактивную визуализацию для таблицы 3 из статьи, которая будет отражать, как на протяжении лет менялось соотношение landslide, comfortable и competitive штатов. Написать вывод.
1 балл

c. 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.
1 балл (0.5 балла за каждую визуализацию)

Постройте для одного из графиков выше древовидную карту.
https://plot.ly/python/treemaps/
https://python-graph-gallery.com/200-basic-treemap-with-python/
Дополнительные 0.5 балла

2. По ссылке загрузите данные о ходе кампании в 2008 году https://github.com/rogovich/Data/blob/master/docs/Plotly_HW/2008_campaign_data.xlsx?raw=true
Создайте новые переменные - общее количество визитов депутатов (переменные Visit), общее количество расходов на рекламу по радио и телевидению (переменные air), расходы на рекламу на душу населения (переменные Spend и StatePop).
0.6 балла
3. По ссылке загрузите данные опроса National Annenberg Election Survey. https://github.com/rogovich/Data/blob/master/docs/Plotly_HW/naes_phone.xlsx?raw=true
В нем мы создаем переменную political knowledge.
   a. С помощью codebook https://github.com/rogovich/Data/blob/master/docs/Plotly_HW/NAES08-Phone-Codebook.pdf
   Найдите в наборе данных naes_phone переменные измеряющие political knowledge и штат проживания респондента и выделите их из датасета. 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:
      i. Supreme Court
      ii. Two-thirds

iii. Democratic
iv. Nominated by the president and confirmed by Senate

2 балла

b. С помощью текстового файла state_of_residency.txt перекодируйте колонку с номерами штатов в текстовый формат (двухбуквенный код).
1 балл

c. Сохраните получившийся датафрейм в csv и положите в архив. Перед сохранением в датафрейме должно быть пять колонок для каждого респондента: штат (двухбуквенный код), ответы на четыре вопроса (1/0), средний уровень political knowledge.

d. В датафрейме, в котором находится информация о ходе кампаний (файл 2008_campaign_data.xlsx) создайте новую переменную на основе датафрейма, полученного в предыдущем пункте.
Найдите значение переменной political knowledge для каждого штата как среднее значение для всех респондентов из этого штата. Например, его можно получить, если сгруппировать датафрейм из предыдущего пункта по штатам и взять среднее значение PK. А дальше соединить таблицы с помощью pandas.merge(), или подставить нужные значения через apply, или...
1.5 балла

4. В plotly создать две тепловые карты (географические) США для трат на рекламу по штатам и визитов кандидатов в штаты для 2008 года (новые переменные, созданные в п. 2). Написать вывод о трендах.
1 балл
5. 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 и знаниями о политике жителей этих штатов (три переменных на одном графике — y, x и цвет или размер). Написать общий вывод.
0.5 балла