Библиотека pandas: часть 2

Алла Тамбовцева

Домашнее задание

Описание домашнего задания и формат сдачи

В домашнем задании необходимо решить предложенные задачи по программированию – вписать свой код в ячейки после условий задач вместо комментария `### YOUR CODE HERE ###` в файле *homework-pandas2.ipynb* и сохранить изменения, используя опцию *Save and Checkpoint* из вкладки меню *File* или кнопку *Save and Checkpoint* на панели инструментов. Итоговый файл в формате `.ipynb` (файл Jupyter Notebook) необходимо загрузить в личный кабинет обучающей онлайн платформы Skillbox (https://go.skillbox.ru/ (https://go.skillbox.ru/)) и отправить на проверку.

Файл `Fishing.csv` содержит результаты опроса о рыбалке: респонденты, заполняя опросник, подробно описывали свою недавнюю рыбалку.

Описание переменных в датафрейме:

- `mode` : выбранный тип рыбалки: на берегу (`beach`), на пирсе (`pier`), в своей лодке (`boat`) и в арендованной лодке (`charter`);
- `price` : стоимость выбранного типа рыбалки;
- `catch` : коэффициент улова при выбранном типе рыбалки;
- `pbeach` : стоимость рыбалки на берегу;
- `ppier` : стоимость рыбалки на пирсе;
- `pboat` : стоимость рыбалки на своей лодке;
- `pcharter` : стоимость рыбалки на арендованной лодке;
- `cbeach` : коэффициент улова на рыбалке на берегу;
- `cpier` : коэффициент улова на рыбалке на пирсе;
- `cboat` : коэффициент улова на рыбалке на своей лодке;
- `ccharter` : коэффициент улова на рыбалке на арендованной лодке;
- `income` : доход в месяц.

Подробнее об опросе и исследовании можно почитать в статье (https://core.ac.uk/download/pdf/38934845.pdf) J.Herriges, C.Kling *"Nonlinear Income Effects in Random Utility Models"* (1999).

Задание 1

Загрузить таблицу из файла `Fishing.csv` и сохранить её в датафрейм `dat` . Вывести на экран первые 8 строк загруженного датафрейма.

In [0]:
```
### YOUR CODE HERE ###
```

Задание 2

Добавить, используя метод `.apply()` , столбец `log_income` , содержащий натуральный логарифм доходов респондентов.

In [0]:
```
### YOUR CODE HERE ###
```

Задание 3

Посчитать для каждого респондента абсолютное значение отклонения `price` от `pbeach` и сохранить результат в столбец `pdiff`.

Подсказка 1: для нахождения абсолютного значения числа используется функция `abs()`. Пример:

```
abs(-8)
8
```

Подсказка 2: пример с lamda-функцией в первом уроке этого модуля.

In [0]:
```
### YOUR CODE HERE ###
```

Задание 4

Сгруппировать наблюдения в таблице по признаку тип рыбалки (`mode`) и вывести для каждого типа среднюю цену (`price`), которую респонденты заплатили за рыбалку.

In [0]:
```
### YOUR CODE HERE ###
```

Задание 5

Сгруппировать наблюдения в таблице по признаку тип рыбалки (`mode`) и вывести для каждого типа разницу между медианным и средним значением цены (`price`), которую респонденты заплатили за рыбалку.

Посказка: можно написать свою lambda-функцию для подсчёта разницы между медианой и средним и применить её внутри метода для агрегирования. Внимание: название самостоятельно написанной функции будет уже вводиться без кавычек.

In [0]:
```
### YOUR CODE HERE ###
```

Задание 6

Сгруппировать наблюдения в таблице по признаку тип рыбалки (`mode`) и сохранить полученные датафреймы (один для каждого типа рыбалки) в отдельные csv-файлы. В итоге должно получиться четыре разных csv-файла.

Подсказка 1: можно запустить следующий код и посмотреть, что получится:

In [0]:
```
for name, data in dat.groupby("mode"):
    print(name, data)
```

Подсказка 2: для сохранения датафрейма в файл используется метод `.to_csv()`. Например, такой код сохранит датафрейм `dat` в файл `"Fish.csv"`:

In [0]:
```
dat.to_csv("Fish.csv")
```

Подсказка 3: для склеивания строк можно использовать оператор `+`, например:

In [0]:
```
"my_file" + ".xlsx"
```

Out[1]:
```
'my_file.xlsx'
```

In [0]:
```
### YOUR CODE HERE ###
```

Задание 7

Отсортировать строки в датафрейме в соответствии со значениями `income` в порядке убывания таким образом, чтобы результаты сортировки сохранились в исходном датафрейме.

In [0]:
```
### YOUR CODE HERE ###
```

Задание 8

Отсортировать строки в датафрейме в соответствии со значениями `price` и `income` в порядке возрастания. Можно ли сказать, что люди с более низким доходом и выбравшие более дешёвый тип рыбалки, в целом, предпочитают один тип рыбалки, а люди с более высоким доходом и более дорогой рыбалкой – другой? Ответ записать в виде текстовой ячейки или в виде комментария.

In [0]:
```
### YOUR CODE HERE ###
```

Задание 9

Любым известным способом проверить, есть ли в датафрейме пропущенные значения. Если есть, удалить строки с пропущенными значениями. Если нет, написать комментарий, что таких нет.

In [0]:
```
### YOUR CODE HERE ###
```

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

Дополнительное (необязательное) задание.

Задание 1

Загрузить датафрейм из файла `wgi_fh.csv`, учитывая, что в качестве разделителя столбцов используется точка с запятой, а в качестве десятичного разделителя – запятая (опции `sep=` и `decimal=` в функции `read_csv()` соответственно).

Файл содержит данные за 2016 по различным политологическим индексам. Датафрейм содержит следующие столбцы:

- `country` : страна;
- `cnt_code` : код страны (аббревиатура);
- `year` : год;
- `va` : индекс подотчётности *Voice & Accountability (WGI)*;
- `ps` : индекс политической стабильности *Political Stability and Lack of Violence (WGI)*;
- `ge` : индекс эффективности правительства *Government Effectiveness (WGI)*;
- `rq` : индекс качества управления *Regulatory Quality (WGI)*;
- `rl` : индекс верховенства закона *Rule of Law (WGI)*;
- `cc` : индекс контроля коррупции *Control of Corruption (WGI)*;
- `fh` : индекс свободы *Freedom House (Freedom Rating)*.

Подробнее про индексы можно почитать на этой странице (https://www.hse.ru/org/hse/4432173/mathbase/databases/db_18).

In [0]:
```
### YOUR CODE HERE ###
```

Задание 2

Вывести общую информацию по датафрейму: число строк и столбцов, типы данных в таблице. Есть ли в таблице пропущенные значения? Привести код и дать ответ в виде комментария.

In [0]:
```
### YOUR CODE HERE ###
```

Задание 3

Если в датафрейме есть строки с пропущенными значениями, удалить их. Сохранить изменения в исходном датафрейме.

In [0]:
```
### YOUR CODE HERE ###
```

Задание 4

Назвать строки в датафрейме в соответствии со столбцом `cnt_code`. Удалить данный столбец из датафрейма.

In [0]:
```
### YOUR CODE HERE ###
```

Задание 5

Отсортировать строки в таблице в соответствии со значениями столбцов с индексами *Control of Corruption* и *Voice & Accountability* таким образом, чтобы результаты сортировки были сохранены сразу в исходном датафрейме.

In [0]:
```
### YOUR CODE HERE ###
```

Задание 6

Используя метод `.apply()`, создать столбец `cc_round` со значениями индекса *Control of Corruption*, округлёнными до первого знака после запятой.

In [0]:
```
### YOUR CODE HERE ###
```

Задание 7

Добавить в датафрейм столбец `fh_status`, в котором будут храниться типы стран в зависимости от значения индекса *Freedom House* (значения типов стран: `"free"`, `"partly free"`, `"not free"`). Соответствие значений `fh` типам стран см. в Table 3 в конце этой (https://freedomhouse.org/sites/default/files/2020-02/Methodology_FIW_2016.pdf) страницы.

Подсказка: здесь понадобится функция, которая возвращает разные значения в зависимости от выполнения условий. Её можно написать через `def` или `lambda`. Больше про функции можно почитать на pythontutor.ru (https://pythontutor.ru/lessons/functions/).

In [0]:
```
### YOUR CODE HERE ###
```

Задание 8

Сгруппировать строки в датафрейме в соответствии со значениями столбца `fh_status`, полученного в предыдущем задании и вывести минимальное, среднее и максимальное значение показателя *Political Stability and Lack of Violence* по каждой группе.

In [0]:
```
### YOUR CODE HERE ###
```

Задание 9

Сгруппировать строки в датафрейме в соответствии со значениями столбца `fh_status` и записать строки, относящиеся к разным группам, в отдельные csv-файлы.

Подсказка 1: цикл `for`.

Подсказка 2: метод (https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.to_csv.html) `.to_csv()` для выгрузки датафреймов `pandas` в csv-файлы.

In [0]:
```
### YOUR CODE HERE ###
```

Задание 10

Создайте (любым способом) маленький датафрейм, состоящий из двух столбцов:

- `fh_type` : тип страны;
- `count` : число стран данного типа.

Постройте, используя полученный датафрейм, столбиковую диаграмму (*barplot*), опираясь на эту (https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.plot.bar.html) документацию. Чтобы увидеть график явно, прямо в текущем ноутбуке, допишите в начале ячейки с кодом для графика следующую строку:

```
%matplotlib inline
```

Подсказка: число наблюдений — это функция `count` , её можно использовать наравне с `min` , `mean` и прочими.

In [0]:
```
### YOUR CODE HERE ###
```