Статистика и емпирични методи

Домашна работа №1

Предаване до 30.Април.2023 23:00:00

Данните са във файла *tips.csv*. Записаните колони са:

- *total_bill*: стойността на сметката в долари;
- *tip*: стойността на бакшиша в долари;
- *sex*: Пол на платеца на сметката в заведението;
- *smoker*: Дали е имало пушачи на масата;
- *day*: Ден от седмицата, в който е посетено заведението;
- *time*: Част от деня, в която е било посетено заведението;
- *size*: Брой лица на масата.

Задачи:

1. (1 т.) Прочетете данните и ги запишете в data frame *tips* в *R*. Покажете, че сте заредили правилно данните, като ги покажете на екрана във формата на таблица.
2. (0,5 т.) Отговорете в коментар на въпроса: Структурата data frame в *R* съдържа еднакъв тип по ... и еднаква дължина по ...?
3. (0,5 т.) Изведете колко реда и колко колони притежава data frame-а *tips*.
4. (0,5 т.) Изведете имената на колоните на data frame-а *tips*.
5. (6 т.) Добавете нова колона *bill.without.tip* в data frame-а *tips* съдържаща стойността на сметката минус стойността на бакшиша. Преместете колоната на 3то място между колоните *tip* и *sex*. Проверете дали стойността на бакшиша може да е включена в стойността на сметката (стойността на сметката ще е по-голяма от стойността на бакшиша). Преименувайте първата колона на data frame-а от *total_bill* на *bill*. Изтрийте колоната *bill.without.tip* от data frame-а *tips*. Преименувайте редовете на $Order1$, $Order2$, $Order3$, ...
6. (1 т.) Изкарайте на екрана първите 7 реда и отделно последните 3 реда използвайки функции в *R*.
7. (1,5 т.) Напишете в коментар в *R*: Какъв вид данни качествени/количествени непрекъснати/дискретни са записани във всяка от колоните? Трансформирайте колоните с качествени данни до колони с факторни променливи.
8. (4 т.) Заредете колоните в паметта. Изведете дескриптивни статистики за всяка една от променливите. Напишете няколко изречения, като коментари в *R* за да опишете резултатите, като отговорите например на въпросите:

(a) Колко е минималната, максималната и средната заплатена сметка?

(b) 75% от сметките са били на стойност по-малка от колко?

(c) Как се нарича още '0.75-тия квантил'?

(d) На колко от масите в заведението е имало пушачи?

(e) Кой е най-често срещания ден от седмицата за посещаване на заведението?

(f) Колко от посещенията на заведението са направени за вечеря?

и т.н.

9. (9 т.) Напишете скрип в R, който да изведе:

   (a) Вариационния ред на стойността на сметките.

   (b) Ранга на 15-тото наблюдение от стойността на сметките.

   (c) Data frame-а *tips* подреден по стойностите в колоната *size*. Без да използвате допълнителни библиотеки.

   (d) Инсталирайте и заредете библиотеката *dplyr*. Използвайте функцията *arrange*() и оператора % > % за да сортирате data frame-а *tips* първо по стойностите в колоната *size* и при наличие на повтарящи се стойности в колоната, сортирайте по колоната *time*.

   Бележка: Пример за приложение на Chain оператора (Pipe) % > %:

   ```
   tips %>% subset(smoker == 'Yes') %>% summary
   ```

   (e) 30% от сметките са били на стойност по-малка от колко?

   (f) 1% от бакшишите са били на стойност по-голяма от колко?

   (g) Само стойностите на долния и горния hinge за бакшишите.

   (h) Стандартното отклонение на стойността на сметките.

   (i) Дисперсията на стойността на бакшишите.

   (j) Най-често колко души са присъствали на една маса? (Сортирайте таблицата за да получите резултата)

10. (9,5 т.) Оценете вероятностите на следните събития в R. Напишете в коментари кои вероятности оценявате и как ги пресмятате с формули.

    (a) 'На случайно избрана маса в заведението да има пушач';

    (b) 'На случайно избрана маса в заведението да има поне 4ма души';

    (c) 'На случайно избрана маса от 3ма души да има пушач';

    (d) 'Случайно избрана поръчка да е от събота вечер';

    (e) 'Случайно избрана поръчка в събота да е за вечеря';

    (f) 'На случайно избрана маса за вечеря да плати мъж';

    (g) 'На случайно избрана маса, на която е платил мъж, да са вечеряли';

    (h) 'На случайно избрана маса да плати мъж и клиентите да са вечеряли';

    За четните факултетни номера (ф.н.) подточки (c), (d) и (e). За нечетните ф.н. подточки (f), (g) и (h).

11. (2 т.) Изведете:

(a) редовете на масите, на които е имало най-много лица и са дошли за обяд (не използвайте константа за максимума);

(b) броя на масите, на които сметката заедно с бакшиша е била по-голяма от $20 и на която са били пушачи, за обяд.

12. (1 т.) Направете подходяща графика за да представите натоварването на заведението в различните дни от седмицата. Добавете заглавие.

13. (2 т.) Направете подходяща графика за да представите пола на платеца в различните части от деня. Добавете цветове и легенда към графиката.

14. (3,5 т.) Инсталирайте и заредете библиотеката *ggplot*2. Напишете скрипт в *R*, с който да пресметнете честотната таблица за пола на платеца в различните дни и части от деня. Напишете коментар, какъв е броят на мъжете, платили сметката в петък на обяд? Направете подходяща графика, с която да представите разпределението на платците по пола в различните части от деня.

Бележка: Проверете как работят функциите *ggplot*(), *aes*(), $facet\_grid()$ и $geom\_bar()$:

```
ggplot(data = <data>, mapping = aes(<mappings>)) + <geom_function>()
```

aes() - дефинира 'aesthetic mapping', като казва кои променливи и по какъв начин (например: size, shape, color и т.н.) да бъдат представени на графиката. Параметърът $fill$ променя цвета на барчетата, когато му се подаде качествена променлива се използва дискретна скала на цветовете, съответстваща на всяка от категориите.

Специална техника наречена 'faceting' позволява да се разбие една графика на много графики по редове или колони спрямо някаква качествена променлива.

```
+   facet_grid(rows = vars(...), cols = ...)
```

<geom_function>() (например $geom\_bar()$) - казва как да бъдат представени данните на графиката (например: bars, points, lines и т.н.). За да фиксирате големината на барчетата може да използвате:

```
+   geom_bar(position = position_dodge2(preserve = "single"))
```

15. (2,5 т.) Направете хистограма на стойността на сметката за вечеря и добавете в червено графика на оценка на плътността върху нея. Симетрична ли е графиката?

16. (5,5 т.) Направете boxplot на стойността на бакшиша за вечеря. Напишете скрипт в *R*, с който да пресметнете интерквартилния размах. Опишете в коментари и пресметнете къде точно се намират линийте на този boxplot. Симетрична ли е графиката? Наблюдавате ли outliers? Ако има такива, колко е стойността на първите два най-малки outlier-а на графиката?

17. (1,5 т.) Направете boxplot на големината на бакшиша в зависимост от пола на платеца (без да използвате допълнителни библиотеки). Анализирайте графиката?

18. (2,5 т.) Инсталирайте и заредете библиотеката *lattice*. Направете подходяща графика за да представите стойността на бакшиша в зависимост от това, в коя част на деня е посетено заведението и пола на платеца. Можете ли да направите някакъв извод?

19. (6 т.) Изследвайте съвместно стойността на сметката *bill* и стойността на бакшиша *tip* с подходящ графичен метод. Пресметнете корелацията на Пирсън и корелацията на Спирман между величините. Напишете коментар в R за да отговорите на въпросите: Бихте ли казали, че съществува линейна връзка между тях?; Бихте ли казали, че съществува някаква зависимост между тях? Начертайте регресионната права. Изведете оценките на коефициентите на линейната функция. Напишете как ще изглежда уравнението на линейния модел. Ако сметката е била $25, каква е очакваната стойност на бакшиша, на базата на линейния модел?

Инструкции за предаване на домашната работа:

- Предаването на домашната работа ще бъде през страницата на курса в moodle.
- Домашната работа трябва да е наименувана с факултетния ви номер и да бъде в .R формат.
- На първите 5 реда в коментари трябва да пише:

```
# ф.н.
# името ви
# специалността
# административната група и дали сте от минали години
# името на асистента, при когото сте посещавали практикум
```

- Всяка нова задача трябва да е отделена с 3 нови реда от следващата. Всяка задача започва с коментар '### Задача ...', който показва номера ѝ. Всяка подточка започва с коментар '## (...)', който показва номера ѝ. Като спазвате номерацията от условията:

```
### Задача 1.
...



### Задача 9.
## (a)
...
## (b)
...
```

- В коментари, където е поискано трябва да бъдат описани анализите и необходимите отговори.
- R скриптът трябва да съдържа коментарите и кода. Не е нужно да копирате резултата от изпълнението на кода.
- Точките на задачите са орентировъчни.
- Общия брой точки е 60. Като:

```
30 точки - Среден 3
40 точки - Добър 4
50 точки - Мн. Добър 5
60 точки - Отличен 6
```