Mestrado em Estatística Aplicada

Ficha de exerciícios Manipulaçao de dados

2023-03-18

Operações de manipulação de dados

Questão I

Para os exercícios a seguir iremos usar base de dados `cereal` disponível na livraria `liver`. Instale a livraria e acesse os dados usando o código `data(cereal)`.

1. Quantas observações e variáveis tem a base de dados `cereal`?
2. Adicione uma nova variável ao conjunto de dados chamada 'totalcarb', que é a soma de `carb` e `sugars`.
3. Quantos cereais na base dados são `hot`?
4. Seleccione o conjunto de cereais cujo frabricante é a Kellogs (K)
5. Seleccione os cereais que tem menos que 80 calorias e mais de 20 unidade de vitaminas.
6. Seleccione os cereais que pelo menos uma unidade de açucar e visualize apenas `Cereal.name`, `calories` and `vitamins`.
7. Exporte cada um dos subconjuntos de dados que seleccionou nos exercícios 4,5 e 6, em um formato `.csv`
8. Renomei a variável `Manufacturer` para `Fabricante`.
9. Quantos fabricantes existem na base de dados?

Questão II

Para a resolução dos exercícios a seguir, primeiro importe a base de dados de biomassa de árvores de Eucalyptus saligna que foram abatidas e medidas. A base tem o nome `esaligna`. **Metadados da base de dados**

```
arvore: número que identifica a árvore em cada talhão
classe: classificação da árvore.
talhao: unidade de produção, definida por uma área na propriedade.
dap: diâmetro a altura do peito (1,3 m altura do tronco)
ht: altura da árvore
tronco: biomassa do tronco
folha: biomassa das folhas
sobra: biomassa restante (outras partes que não folha e tronco)
total: biomassa total
```

1. Verifique se as variáveis `classe` e `talhao` são fator, caso não sejam, faça a coerção para a classe factor mantendo o mesmo nome.
2. Verifique o conteúdo do objeto resultante, com as funções: str, head e summary. Guarde os resultados de head e summary em objetos com os nomes: `headEsal` e `sumEsal`
3. Faça a contagem de quantas vezes cada valor da variável `arvore` aparece, ordenada pelo valor crescente da variável arvore. Guarde esse resultado no objeto `arvTab`.

4. Crie uma variável composta pela junção dos characteres arv somado aos valores de arvore, classe e talhao, sem espaço entre os caracteres e nessa ordem. Essa nova variável deve ser incluia como uma variável de esaligna chamada `arvID`.
5. Faça a contagem de quantas vezes cada código da variável arvID aparece. Guarde esse resultado no objeto `idTab`. Garanta que os valores de contagens tem como nome o código arvID correspondente.
6. Adicione uma nova coluna no objeto esaligna chamado `biomTrFo` com a soma das biomassas de folhas e do tronco de cada árvore, essa nova variável deve vir na última posição das colunas do dataframe.
7. Em uma outra coluna, denominada areabasal calcule o valor da área basal de cada árvore. Considere que o tronco apresenta a secção transversal circular e que a área basal é dada pela área desta secção na altura do peito. Lembre que a coluna dap no seu data frame se refere ao diâmetro da árvore na altura do peito.
8. Crie um novo objeto chamado esaligna15cm, selecionando apenas os dados relativos às árvores com mais de 15 cm de diâmetro na altura do peito do objeto esaligna.
9. Salve os objetos modificados e criados em arquivos texto, com campos separados por tabulação, com os respectivos nomes: `esaligna.txt` e `esaligna15cm.txt`

Questão II

Para os exercícios a seguir vamos usar a base de dados `flights` da livraria `nycflights13`. Primeiro, terão de instalar a livraria para ter acesso aos dados. Podem ver a documentação dos dados no help do R. Chamem os dados usando o código `data(flights)`.

Encontre:

1. Todos os aviões que tiveram um atraso de duas horas.
2. Que partiram para Huoston (IAH ou HOU).
3. Que foram operados pela United, American, ou Delta.
4. Partiram no verão
5. Chegaram com duas horas de atraso tem saido cedo.
6. Atrasaram-se em pelo menos uma hora, mas compensaram mais de 30 minutos de voo.
7. Partiram entre a meia noite e as seis da manhã.
8. Quais são os destinos que receberam mais aviões m Junho?
9. A transportadora que teve a maior distância média por voo?
10. O dia que teve o maior atraso médio de chegada para todos os voos?
11. A distancia total de todos aviões em Janeiro.
12. Quantas companhias aéreas não têm a palavra “air” em seu nome? (Tente procurar uma função na internet que lhe ajude a fazer correspondência de strings, por exemplo `grepl`, veja a documentação)

Junção de base de dados - Merge

Primeiro, execute o código a seguir para construir três dataframes que irá usar para fazer a operção de merge

```
dataset1 <- data.frame(unit=letters[1:9], treatment=rep(LETTERS[1:3],each=3),
Damage=runif(9,50,100))

unitweight <- data.frame(unit=letters[c(1,2,4,6,8,9)], Weight = rnorm(6,100,0.3))
```

```r
treatlocation <- data.frame(treatment=LETTERS[1:3], Glasshouse=c("G1","G2","G3"))
```

Faça o merge das três data frame de dados, criando uma única data frame para que tem as colunas 'unit', 'treatment', 'Glasshouse', 'Damage' e 'Weight'. Algumas observações não possuem informação para a veriável weight. Faça o merge dos dados de duas maneiras, para incluir ou excluir as observaçõs sem informação na variável `Weight`.

Rachid Muleia, PhD in Statistics