Conteúdo

Capítulo 1

O banco de dados

1.1 Informações pessoais

1.1.1 Nome

Os nomes são gerados aleatoriamente de acordo com listas de frequência, sendo que o primeiro nome depende do sexo do indivíduo. As listas foram retiradas de...

1.1.2 Idade

A idade é representada por um número inteiro de anos, variando de 0 a 99. Para obter idades mais precisas (como meses, úteis para bebês), pode-se usar a data de aniversário.

1.1.3 Data de nascimento

A data de nascimento é representada no formato dd/mm/aaaa. No momento, não existe nascimento no dia 29 de fevereiro.

1.1.4 Signo do zodíaco

O signo do zodíaco é representado por uma palavra, dependente da data de nascimento.

1.1.5 Signo chinês

O signo chinês é representado por uma palavra, dependente do ano de nascimento. Embora o signo chinês correto não depende apenas do ano, mas também do dia do nascimento, esse detalhe não está presente nesse banco de dados.

1.1.6 Gênero

O gênero é representado por uma letra: "M" para masculino e "F" para feminino.

1.1.7 Altura

A altura é representada em centímetros, com precisão de 1 casa decimal. Para obter a altura em metros, basta dividir a coluna (ou a célula) por 100.

1.1.8 Peso

O peso é representado em kg, com precisão de 2 casas decimais.

1.1.9 IMC

O IMC é representado em kg/m^2, com precisão de 2 casas decimais.

1.1.10 Percentual de gordura corporal (BF%)

O percentual de gordura corporal é representado em porcentagem, com 1 casa decimal.

1.1.11 Raça

A raça é representada por uma palavra. São consideradas 5 raças:

- Branco
- Preto
- Amarelo
- Pardo
- Indígena

1.1.12 Cor do cabelo

A cor do cabelo é representada por uma ou mais palavras. São consideradas 9 cores:

- Loiro comum
- Castanho claro
- Castanho médio
- Castanho escuro
- Castanho-ruivo

- Ruivo
- Preto
- Grisalho
- Branco

1.1.13 Cor dos olhos

A cor dos olhos é representada por uma palavra. São consideradas 4 cores:

- Castanho
- Azul
- Âmbar
- Verde

1.1.14 Cidade e estado

Cidade e estados são representados em duas colunas, por palavras. Cuidado, pois existem cidades de mesmo nome em mais de um estado. Para selecionar uma cidade específica, filtre por cidade e estado, ou crie uma coluna que agregue as duas informações. Uma sugestão é criar uma coluna no modelo "Cidade (UF)".

1.1.15 Estado civil

Existem 4 estados civis: solteiro, casado, divorciado e viúvo. Indivíduos com menos de 16 anos não podem se casar legalmente, então recebem NA. Em algumas abordagens, pode ser interessante substituir NA por "Solteiro(a)".

1.2 Personalidade e preferências

1.2.1 Personalidade

A personalidade é descrita por 3 palavras, separadas por vírgula e espaço. A maneira mais fácil de fazer análises nessa coluna é separar os 3 traços.

1.2.2 Time de futebol

O time de futebol preferido é dado por um termo. São considerados 26 times, outros times, ou não torcer.

1.2.3 Destino de viagem

Existem 3 tipos de destino de viagem preferidos: praia, campo e cidade.

1.2.4 Período do dia

Existem 3 períodos do dia preferidos: manhã, tarde e noite.

1.2.5 Refeição

Existem 3 refeições preferidas: café da manhã, almoço e jantar.

1.2.6 Prato

Existem vários pratos preferidos. Indivíduos com NA não têm prato preferido.

1.3 Dados socioeconômicos

1.3.1 Alfabetização

A alfabetização é representada por uma letra: "S" representa indivíduos alfabetizados, e "N" representa indivíduos analfabetos.

1.3.2 Escolaridade

A escolaridade, ou nível de instrução, é representada por um termo ou por NA. Indivíduos com NA têm menos de 6 anos, e por isso não são considerados "Sem instrução". Dependendo da análise, recomenda-se substituir NA por "Sem instrução". Os níveis de escolaridade são:

- Sem instrução
- Ensino fundamental incompleto
- Ensino fundamental completo
- Ensino médio incompleto
- Ensino médio completo
- Ensino superior incompleto
- Ensino superior completo

1.3.3 Acesso à educação privada

O acesso à educação privada se refere a escolas públicas ou privadas. É representado por uma palavra; indivíduos que não estão em idade escolar recebem NA.

1.3.4 Emprego

O emprego é representado por um código numérico no **formato de palavra**. Esse código numérico é advindo da Classificação Brasileira de Ocupações (CBO), do Ministério do Trabalho. Uma chave com todos os códigos e seus nomes principais (pois alguns códigos são usados para um grupo de ocupações semelhantes) está disponível. "Do lar" representa indivíduos sem ocupação oficial, mas que cuidam da própria casa. Indivíduos desempregados, que não se ocupam com a própria residência são representados por NA.

1.3.5 Renda

A renda está representada por um número no formato R$reais,centavos. Indivíduos que não têm renda recebem NA.

1.3.6 Velocidade de acesso à internet

A velocidade da internet à qual o indivíduo tem acesso é representada por em Mbps, com precisão de 1 casa decimal. Indivíduos sem acesso à internet recebem NA.

1.3.7 Plano de saúde privado

O acesso a um plano de saúde privado é representado por uma letra: "S" representa indivíduos com um plano de saúde privado, e "N" representa indivíduos sem acesso a plano de saúde privado.

1.4 Saúde

1.4.1 Atividade esportiva

A prática de atividade esportiva é representada em duas colunas. A primeira indica o esporte práticado pelo indivíduo, sendo que indivíduos que não praticam qualquer esporte recebem o valor "Sedentário". A segunda indica o tempo médio em horas, por semana, de prática do esporte (com precisão de uma casa decimal); indivíduos sedentários recebem o valor NA.

1.4.2 Tipo sanguíneo

O tipo sanguíneo é representado por um termo: o primeiro se refere ao tipo no sistema ABO, e o segundo no sistema Rh . Dependendo da análise desejada, pode ser conveniente separar os dois sistemas em colunas diferentes.

Capítulo 2

Técnicas úteis

2.1 Tratar NAs

Algumas colunas do banco de dados têm "dados faltantes". No R, dados faltantes são representados por "NA" (Not Available). Em uma análise de dados, é importante definir como lidar com dados indisponíveis, já que nem sempre se pode simplesmente ignorar os dados indisponíveis. No mínimo, em alguns casos, eles podem diminuir o tamanho amostral de parte da análise; em outros, podem afetar a validade da análise se não forem considerados.

Muitas funções não ignoram NAs automaticamente, como uma forma de forçar o usuário a considerar como tratá-los, além de deixar evidente caso um conjunto de dados tenha algum NA perdido no meio, que poderia passar despercebido. Um exemplo é a função `sum()`. Por padrão, essa função retorna NA se encontrar qualquer NA dentro dos dados entregues a ela. Para contornar isso, basta passar o argumento `na.rm = TRUE`. Muitas funções têm um parâmetro semelhante.

A permanência de NAs na sua análise depende da sua seleção de dados. Por exemplo: se você decidir analisar pessoas com renda maior que 1 salário mínimo mensal, naturalmente vai retirar os NAs da coluna. Porém, se você fizer uma amostragem de mulheres pretas da cidade de São Paulo, e decidir analisar as rendas, provavelmente vai encontrar alguns NAs no meio dos seus dados. Dependendo da sua análise, considere de que forma a falta de alguns dados afeta a interpretação. De forma alguma ignore os NAs sem ter certeza que faz sentido ignorá-los.

2.1.1 Códigos para lidar com NAs

Verificando a existência de NAs

Você pode procurar se existe ao menos um NA em um conjunto de dados com o operador `%in%`, o que retorna TRUE ou FALSE:

```
1 NA %in% data
```

Você também pode encontrar quais valores são NA com `is.na()` (retorna TRUE ou FALSE, ou uma sequência deles). Essa função pode servir de filtro, também (no primeiro caso, mostra apenas os dados que são NA; no segundo, apenas os que não são):

```
1 is.na(data)
2 data[is.na(data)]
3 data[!is.na(data)]
```

Lidando com NAs nativamente

Ao invés da opção `na.rm`, algumas funções usam `na.action`. Nesses casos, elas utilizam a ação padrão atribuída nas opções do R, mas permitem uma mudança local nessas opções (como a função `lm()`):

`na.omit/na.exclude` Remove as observações que contém NA e retorna o objeto

`na.pass` Retorna o objeto sem modificar

`na.fail` Retorna o objeto apenas se não tiver NA

Lidando com NAs usando o Tidyr (do Tidyverse)

O pacote Tidyr oferece 3 formas de lidar com NAs:

`drop_na()` Remove linhas contenho NAs nas colunas apontadas

`fill()` Completa com o valor que veio antes ou que vem depois (dependendo da escolha)

`replace_na()` Substitui os NAs pelo valor escolhido.

Note que você pode utilizar `replace_na()` para colocar um valor calculado, como a mediana ou a média dos valores disponíveis.

2.2 Manipulação de datas

2.3 Manipulação de texto

2.4 Criação de colunas