Método Estatístico de Apoio à Decisão
Semana 2

Douglas Rodrigues
Karina Yaginuma

Universidade Federal Fluminense

Video 1

Iniciando um projeto

- Vamos iniciar um novo projeto, onde analisaremos os dados de consumo de água, separado por tipo de imóvel, da cidade de Austin, entre janeiro de 2012 até outubro de 2017.
- Crie um novo projeto, chamado **Austin**, e coloque na pasta do projeto o arquivo dos dados *austin_agua.csv*.
- Carregue os dados para o **R**. Verifique se todos os campos foram classificados corretamente pelo sistema.

- Corrija a coluna da data. Observe que só há informação do ano e do mês.
- Crie duas novas colunas, uma contendo apenas o ano, e outra com o mês do consumo. Isso vai ajudar na comparação do consumo por mês e por ano.

  Sugestão: use os comandos `month()` e `year()` do pacote *lubridate*.

Natureza das Variáveis

- As variáveis possuem naturezas diferentes em relação aos possíveis valores que podem assumir:

Exemplo

No banco de dados *austin_agua.csv*, as variáveis:

- ***Customer_Class*** é uma categoria, o tipo de residência do cliente;
- ***Total_Gallons*** é um valor numérico, neste caso estamos contando o número de galões consumidos.

- Existem dos tipos de variáveis **Qualitativas (Categóricas)** e **Quantitativas**.

Variáveis Categóricas

Variáveis Categóricas: *descrevem* características de elementos de uma população e podem ser medidas em escala nominal ou ordinal.

- **Nominais:** não existe ordenação entre as categorias.
  Exemplos: sexo, cor dos olhos, fumante/não fumante, doente/sadio.

- **Ordinais:** existe uma ordenação natural entre as categorias.
  Exemplos: escolaridade (fundamental, médio e superior), estágio da doença (inicial, intermediário, terminal), mês (janeiro, fevereiro,..., dezembro).

Variáveis Quantitativas

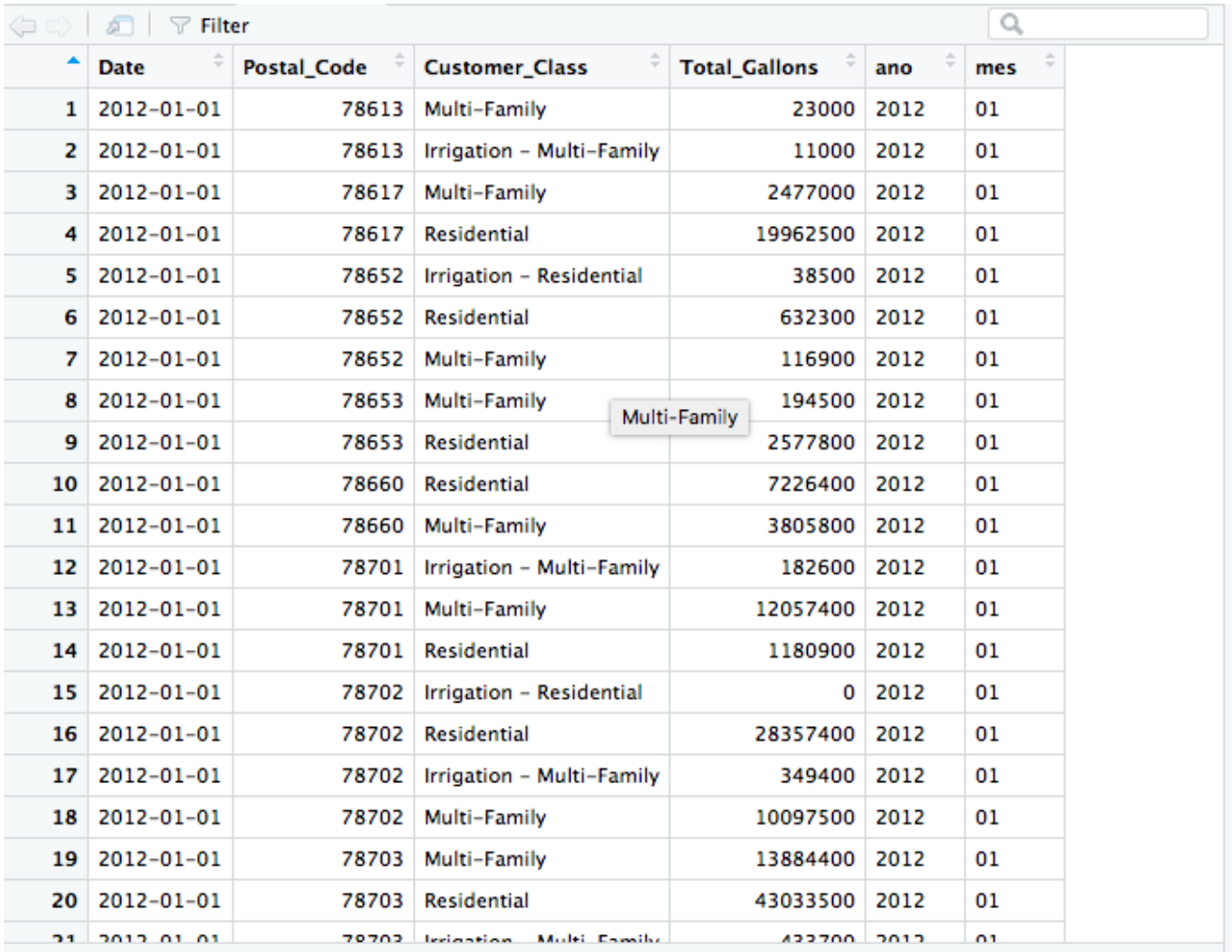
Variáveis Quantitativas: *medem* características de elementos de uma população e são expressar por valores numéricos. As variáveis quantitativas, por sua vez, podem ser discretas ou contínuas.

- **Discreta:** assumem valores pertencentes a um subconjunto dos números inteiros; em geral, resultam de processos de contagem.
  Exemplos: número de filhos, número de bactérias por litro de leite, número de cigarros fumados por dia.
- **Contínuas:** assumem valores pertencentes a um subconjunto dos números reais; em geral resultam de processos de medição.
  Exemplos: peso (balança), altura (régua), tempo (cronômetro).

Comentários

- A natureza da variável deve ser levada em consideração para interpretar os resultados.
- Se o indivíduo A tem 40 anos e o indivíduo B tem 20 anos, é **correto** afirmar que A tem o dobro da idade de B.
- Agora vamos considerar o **grau de especialização**, podemos caracterizá-los da seguinte maneira:
  - não especializado $\Rightarrow$ 1;
  - especializado $\Rightarrow$ 2;
  - muito especializado $\Rightarrow$ 3;
- Podemos dizer que um trabalhador muito especializado tem o dobro da especialização de um não especializado?

E agora? Como obter informação deste banco de dados?



Filter

	Date	Postal_Code	Customer_Class	Total_Gallons	ano	mes
1	2012-01-01	78613	Multi-Family	23000	2012	01
2	2012-01-01	78613	Irrigation - Multi-Family	11000	2012	01
3	2012-01-01	78617	Multi-Family	2477000	2012	01
4	2012-01-01	78617	Residential	19962500	2012	01
5	2012-01-01	78652	Irrigation - Residential	38500	2012	01
6	2012-01-01	78652	Residential	632300	2012	01
7	2012-01-01	78652	Multi-Family	116900	2012	01
8	2012-01-01	78653	Multi-Family	194500	2012	01
9	2012-01-01	78653	Residential	2577800	2012	01
10	2012-01-01	78660	Residential	7226400	2012	01
11	2012-01-01	78660	Multi-Family	3805800	2012	01
12	2012-01-01	78701	Irrigation - Multi-Family	182600	2012	01
13	2012-01-01	78701	Multi-Family	12057400	2012	01
14	2012-01-01	78701	Residential	1180900	2012	01
15	2012-01-01	78702	Irrigation - Residential	0	2012	01
16	2012-01-01	78702	Residential	28357400	2012	01
17	2012-01-01	78702	Irrigation - Multi-Family	349400	2012	01
18	2012-01-01	78702	Multi-Family	10097500	2012	01
19	2012-01-01	78703	Multi-Family	13884400	2012	01
20	2012-01-01	78703	Residential	43033500	2012	01

Multi-Family

Import Dataset

Global Environment

Data

dados 11132 obs. of

Files Plots Packages Help Viewer

Zoom Export

Video 2

Tabela de Frequência

- Uma maneira de tornar as informações presentes nos dados mais evidente é através da tabela de frequência.
- Para cada variável, a tabela de frequência fornece a informação sobre sua distribuição. zesp
- Para um banco de dados com **n** observações, a estrutura básica da tabela de frequência:

Variável	Freq. Absoluta	Porcentagem	Porcentagem Acum.
Total	n	1	100%

Variáveis Categóricas

- Considere a variável *Customer_Class*, podemos utilizar o comando `table` e criar a tabela com as frequências absolutas de cada categoria.

```
> table(dados$Customer_Class)

Irrigation - Multi-Family  Irrigation - Residential
                     2569                       2409
            Multi-Family                Residential
                     2934                       3220
```

- Para construir uma tabela mais completa, com a porcentagem e porcentagem acumulada, podemos construir um *data frame* manualmente, ou utilizar um pacote para nos auxiliar
- Vamos apresentar o comando `tab1()` do pacote *epiDisplay*.

Pacote epiDisplay - Função tab1

- A função `tab1()` gera tabelas de frequência em conjunto com um gráfico de barras.

```
> install.packages("epiDisplay")
> library(epiDisplay)
> tabela1 <- tab1(dados$Customer_Class)
```

- `tabela1` é uma lista: o primeiro componente da lista fornece o nome da variável e o segungo componente fornece a tabela de frequência.

Função `tab1()`

- Observe que o comando `tab1()` nos retorna uma tabela e um gráfico de barras.
- O gráfico de barras é uma das melhores formas de se apresentar dados CATEGÓRICOS.
- Observe que cada barra representa uma categoria da variável.

Função `tab1()`

Alguns argumentos da função `tab1()`

- `cex.names` e `cex.axis`: modifica a magnitude dos nomes das barras e do eixo, são valores entre 0 e 1;
- `las`: estilo dos valores dos eixos (0 - paralelos aos eixos, 1 - horizontais, 2 - perpendicular aos eixos e 3 - verticais);
- `xlim` e `ylim`: define os limites dos eixos o intervalo `[a,b] = c(a,b)`;
- `col`: define as cores das barras, pode ser um vetor de cores;
- `main`: título do gráfico.

Função tab1

- `decimal`: número de decimais para porcentagem na tabela (default = 1).
- `sort.group`: padrão para ordenar as categorias ("decreasing"ou "increasing"), default nenhuma ordenação.
- `cum.percent`: coluna da porcentagem acumulada, default é `TRUE`.
- `graph`: se igual a `FALSE` o gráfico de barras não é construído.
- `missing`: inclui valores ausentes como categoria.
- `bar.values`: define o tipo do valor de frequência em cada barra ("percent"ou "none")

Gráfico de setores (Pizza)

- Gráfico de setores é um diagrama circular onde os valores de cada categoria de uma variável são proporcionais às respectivas frequências.
- Deve ser utilizados quando queremos evidenciar as proporções de cada categoria.
- Adequada para variáveis com um número pequeno de possíveis categorias.
- Para criar um gráfico de pizza, utilizamos a função `pie`.

```
> freq <- table(dados$Customer_Class)
> pie(freq)
```

Video 3

Variáveis Quantitativas

- Quando trabalhamos com dados quantitativos, as vezes há muitos valores distintos. Uma uma tabela de frequência simples pode não nos dar muita informação.

  ```
  > tab1(dados$Total_Gallons)
  ```

- Nesse caso, o melhor é construir tabelas de frequência agrupadas (ou por classes).

Histograma

- Podemos definir manualmente a quantidade e o tamanho das categorias, mas o R faz isso automaticamente.
- É comum usar gráficos de barras em dados quantitativos, no entanto, o correto é usar o HISTOGRAMA.
- O histograma também usa barras para indicar a frequência, mas dividido em intervalos reais, e não em categorias. Utilizamos o comando `hist()`.

Histograma

Como criar tabela de frequência por classes através do histograma:

- Crie um histograma dos dados, com o comando `hist()`.
- Verifique a quantidade e o tamanho dos intervalos de dados, além no maior e menor valor do eixo das abcissas do histograma.
- Com o comando `cut()`, vamos "fatiar" os dados que queremos criar a tabela de frequência;
- Use o comando `tab1()` para construir a tabela de frequência com os dados "fatiados".

Exemplo

- Para fazer uma tabela visualmente mais bonita, vamos alterar a escala dos dados, de galão para unidades de 10.000.000 galões.

```
> dados <- mutate(dados,
dezmil_gallons=Total_Gallons/10000000)
```

- Em seguida, vamos visualizar o histograma de dezmil_gallons.

```
> hist(dados$dezmil_gallons)
```

- Para obter informações mais apuradas, criamos o histograma como um objeto.

```
> histograma<- hist(dados$dezmil_gallons)
> histograma
```

Exemplo

- Observe que temos várias informações, como onde são delimitados os intervalos das classes (*breaks*). frequência de cada classe (*counts*), entre outros.
- Vamos "fatiar" os dados, com os mesmos intervalos (*counts*) do histograma.

```
> consumo <- cut(dados$dezmil_gallons,
breaks=c(0:13))
```

- Criamos então a tabela de frequência.

```
> tab1(consumo,graph=F)
```

Exemplo

- Observe que há vários NA's nos dados. Isso ocorre porque os intervalos estão aberto na esquerda e fechado na direita, ou seja, os valores 0's foram excluídos.
- Para resolver isso, fechamos o intervalo à esquerda.

```
> consumo <- cut(dados$dezmil_gallons,
breaks=c(0:13),right=F)
```

```
> tab1(consumo,graph=F)
```

Observação

- Podemos utilizar outro comando no gráfico de histograma, como colorir, mudar limites dos eixos, renomear gráfico, etc.

```
> hist(dados$dezmil_gallons,col=rainbow(10),
xlim=c(0,14), main="Histograma",xlab="Consumo (em
10mil Galoes)")
```

Observação

- Podemos trocar a frequência para valores relativos, com o comando *prob=T*. Podemos, também, traçar linhas por cima do gráfico, utilizando o comando `lines()` ou `abline()`.

```
> hist(dados$dezmil_gallons,col=rainbow(10),
ylim=c(0,1),xlim=c(0,14),main="Histograma",
xlab="Consumo (em 10mil Galoes)",prob=T)
```

```
> lines(density(dados$dezmil_gallons))
```

- Em breve aprenderemos o conceito de densidade.