Visualizing Data with ggplot2

Visualizações são grandes aliadas ao investigar bases de dados, enxergar padrões, extrair insights e descobrir tendências. No entanto não basta criar um gráfico qualquer, existem tipos diferentes, princípios a serem seguidos para evitar má interpretações, vieses e uma análise mal embasada.

Dados para o exercício

Para o exercício vamos utilizar uma base do National Center for Health Statistics sobre tendências de mortalidade nos Estados Unidos ao longo dos anos que pode ser encontrada clicando aqui.

Bibliotecas

```
library(readr)
library(dplyr)
library(purrr)
library(ggplot2)
```

Importando o dataset

```
base <-
  read.csv("https://data.cdc.gov/api/views/w9j2-ggv5/rows.csv?accessType=DOWNLOAD")

glimpse(base)
```

```
## Rows: 1,071
## Columns: 5
## $ Year                              <int> 1900, 1901, 1902, 1903, 1904, 1905, 19~
## $ Race                              <chr> "All Races", "All Races", "All Races",~
## $ Sex                               <chr> "Both Sexes", "Both Sexes", "Both Sexe~
## $ Average.Life.Expectancy..Years. <dbl> 47.3, 49.1, 51.5, 50.5, 47.6, 48.7, 48~
## $ Age.adjusted.Death.Rate           <dbl> 2518.0, 2473.1, 2301.3, 2379.0, 2502.5~
```

Campo	Significado
Year	Ano de Nascimento
Race	A raça da população avaliada
Sex	Sexo da população avaliada
Average.Life.Expectancy..Years.	A expectativa de vida em anos dado o ano do nascimento
Age.adjusted.Death.Rate	Taxa de mortalidade ajustada por idade de pessoas nascidas em um determinado ano

Padronizando nome dos campos

```
names(base)[4] <- "Avg_Life_Expect"
names(base)[5] <- "Age_Adj_Death_Rate"

names(base)
```

```
## [1] "Year"                "Race"                "Sex"
## [4] "Avg_Life_Expect"     "Age_Adj_Death_Rate"
```

Investigando os dados da base

Temos dados de 1900 a 2018, para diferente sexos e raças. O interessante será analisar a mudança da expectativa de vida ao longo do tempo, bem como a mudança na taxa de mortalidade.

```
base %>% select(Year,Avg_Life_Expect,Age_Adj_Death_Rate) %>% summary()
```

```
##       Year      Avg_Life_Expect Age_Adj_Death_Rate
##  Min.   :1900   Min.   :29.1    Min.   : 611.3
##  1st Qu.:1929   1st Qu.:57.1    1st Qu.:1013.0
##  Median :1959   Median :66.8    Median :1513.7
##  Mean   :1959   Mean   :64.5    Mean   :1593.1
##  3rd Qu.:1989   3rd Qu.:73.9    3rd Qu.:2057.2
##  Max.   :2018   Max.   :81.4    Max.   :3845.7
##                 NA's   :6
```

```
base %>% select(Race, Sex) %>% map(table)
```

```
## $Race
##
## All Races     Black     White
##       357       357       357
##
## $Sex
##
## Both Sexes     Female       Male
##        357        357        357
```

ggplot2

As letras gg da biblioteca ggplot2 significa *grammar of graphics*, é um sistema de visualização de dados em que o ggplot2 se baseia para criar um gráfico em camadas.

- Precisamos de dados x e y que vão alimentar o gráfico
- Precisamos de um eixo x e um eixo y
- Precisamos de uma forma geométrica para desenhar o gráfico, nesse exemplo uma linha
- Precisamos de rótulos para explicar do que se trata os dados apresentados

E assim combinando todos os elementos chegamos numa figura que explica a relação entre os dados.

Essa relação pode ser positiva quando conforme x avança y aumenta, ou negativa quando conforme x avança o y diminui.

Camada 1: Os dados

Começando o gráfico apenas com os dados, temos o seguinte resultado.

```
base %>% ggplot()
```

Camada 2: Os Eixos

Só colocando os dados sem informar qual representa o eixo vertical y e o eixo horizontal x, não é possível enxergar nada além de um quadrado branco. Com o código abaixo informamos os eixos, e assim o gráfico começa a ter forma.

Pode ser difícil decidir qual informação usar em qual eixo. Então pense da seguinte forma:

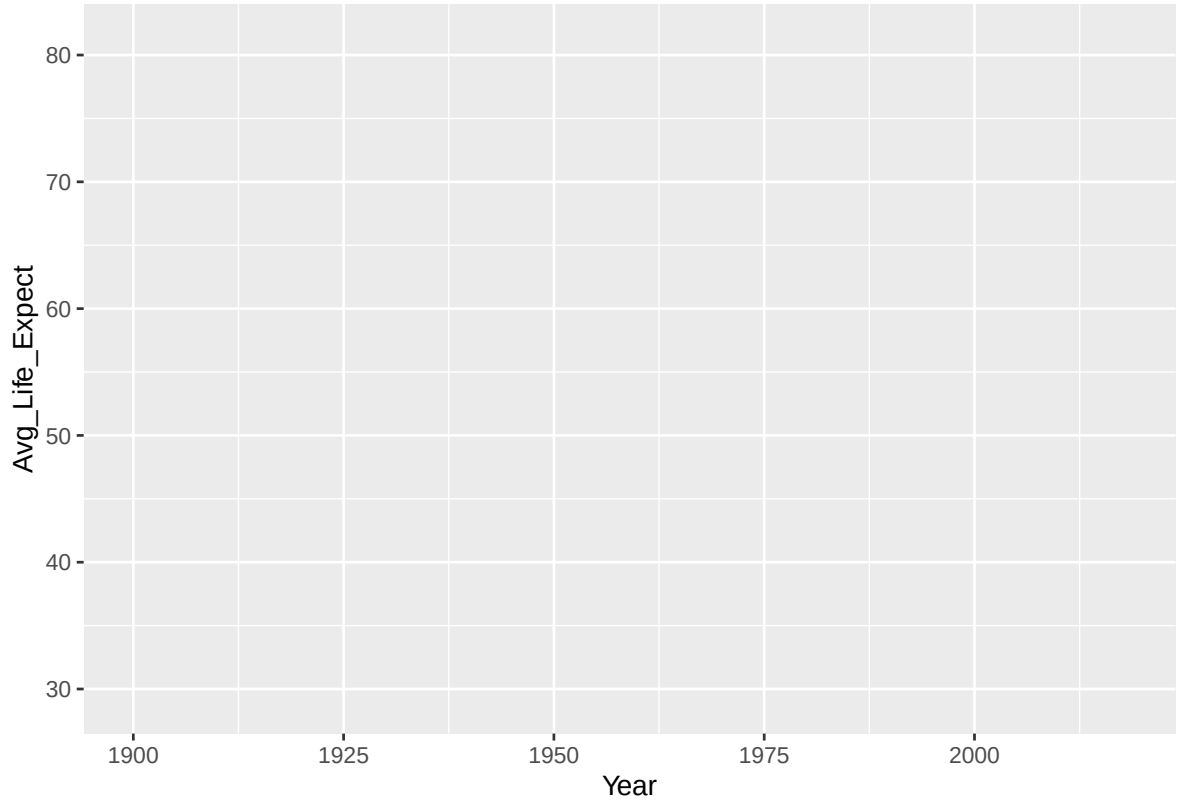
- Conforme x muda o que acontece com y?
- Qual a relação entre x e y?

A ideia é que x seja a variável que conseguimos controlar ou medir, já y é algo que muda em resposta a mudança de x, acompanha a mudança de x, depende de x.

- Eixo horizontal x: a variável que nós temos o controle, que muda, é a variável independente.
- Eixo Vertical y: a variável que muda de acordo com a variável independente, portanto é a variável dependente.

No nosso caso, queremos ver como a Expectativa de Vida muda ao longo dos Anos, portanto a Expectativa de vida fica no eixo y por depender da variável tempo em anos que fica no eixo x.

```
base %>% ggplot(aes(x=Year,y=Avg_Life_Expect))
```



Camada 3: Forma geométrica

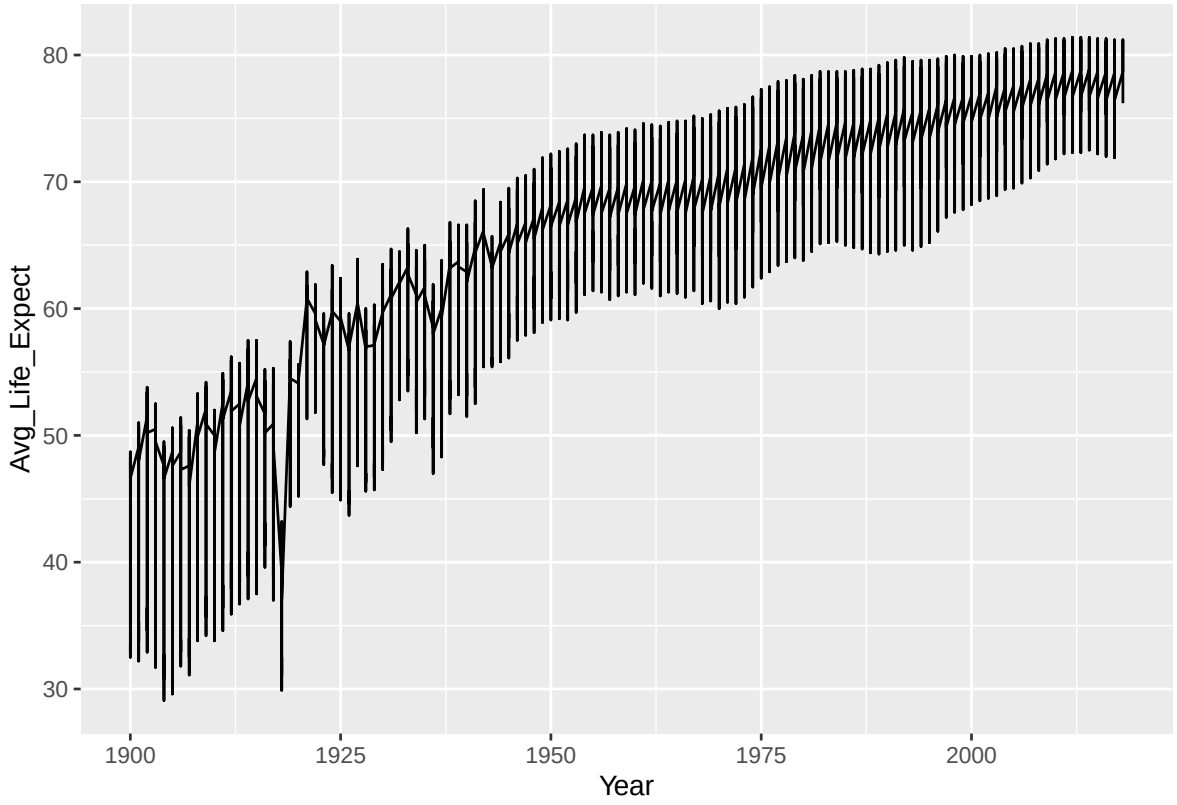
Agora que temos os eixos, temos que desenhar os pontos x e y nesse nosso plano cartesiano, e para isso precisamos de uma forma geomátrica, nesse caso uma linha.

A adição das camadas ao gráfico aqui será representadas pelo símbolo de adição + e nas próximas camadas continuaremos a utilizar esse símbolo sempre a partir da função que plota o gráfico.

As formas geométricas a serem adicionadas no gráfico tem como convenção começar com o prefixo *geom* tornado padronizado e intuitivo, facilitando nossa vida ao criar o gráfico.

```
base %>% ggplot(aes(x=Year,y=Avg_Life_Expect)) +
          geom_line()
```

```
## Warning: Removed 6 row(s) containing missing values (geom_path).
```



No entanto o gráfico está mais complexo do que o esperado, esperávamos uma linha simples, e vemos muitos altos e baixos a cada ponto desenhado.

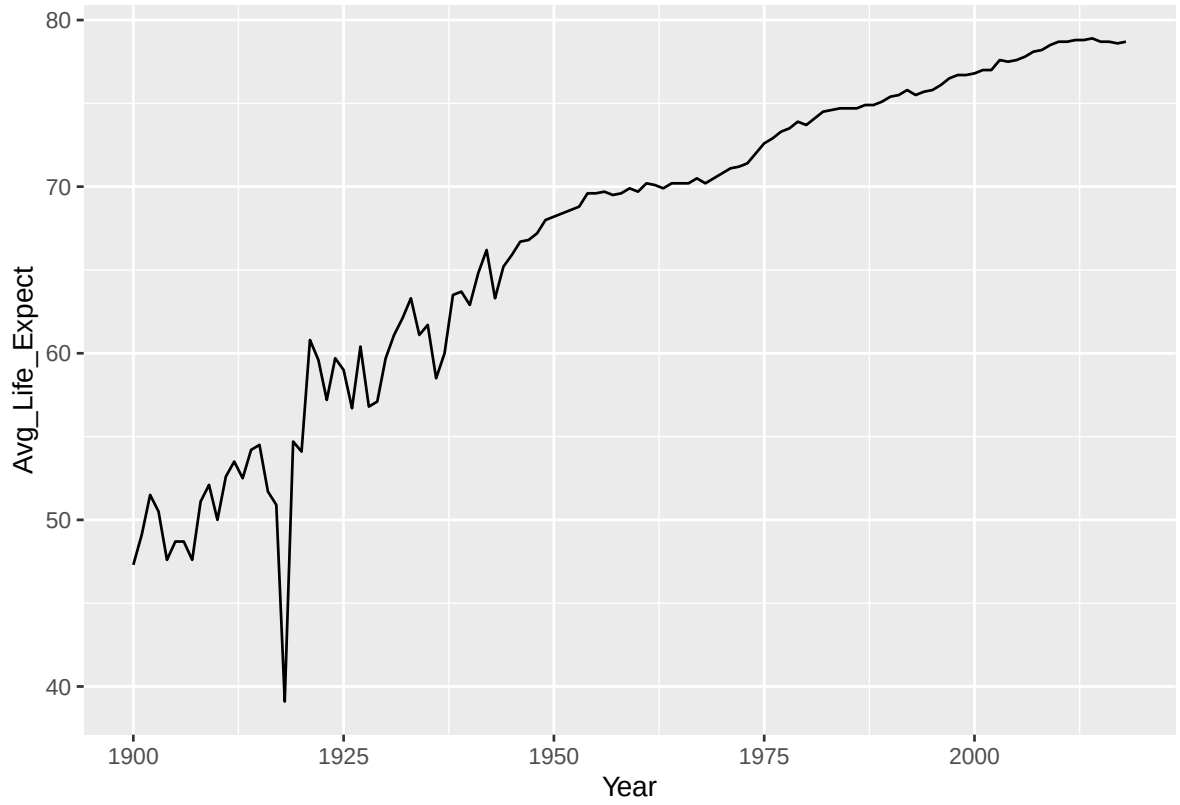
Ao investigar os dados notamos que para um mesmo ano temos muitas informações e tudo está entrando no gráfico.

```
base %>% filter(Year == 2000) %>% head()
```

```
##   Year      Race        Sex Avg_Life_Expect Age_Adj_Death_Rate
## 1 2000 All Races Both Sexes            76.8               869.0
## 2 2000 All Races     Female            79.7               731.4
## 3 2000 All Races       Male            74.3              1053.8
## 4 2000     Black Both Sexes            71.8              1121.4
## 5 2000     Black     Female            75.1               927.6
## 6 2000     Black       Male            68.2              1403.5
```

Em nossa análise vamos partir de um ponto mais genérico e portanto filtrar os registros que representam ambos sexos e raças.

```
base %>% filter(Sex == "Both Sexes",
                Race == "All Races") %>%
         ggplot(aes(x=Year,y=Avg_Life_Expect)) +
          geom_line()
```



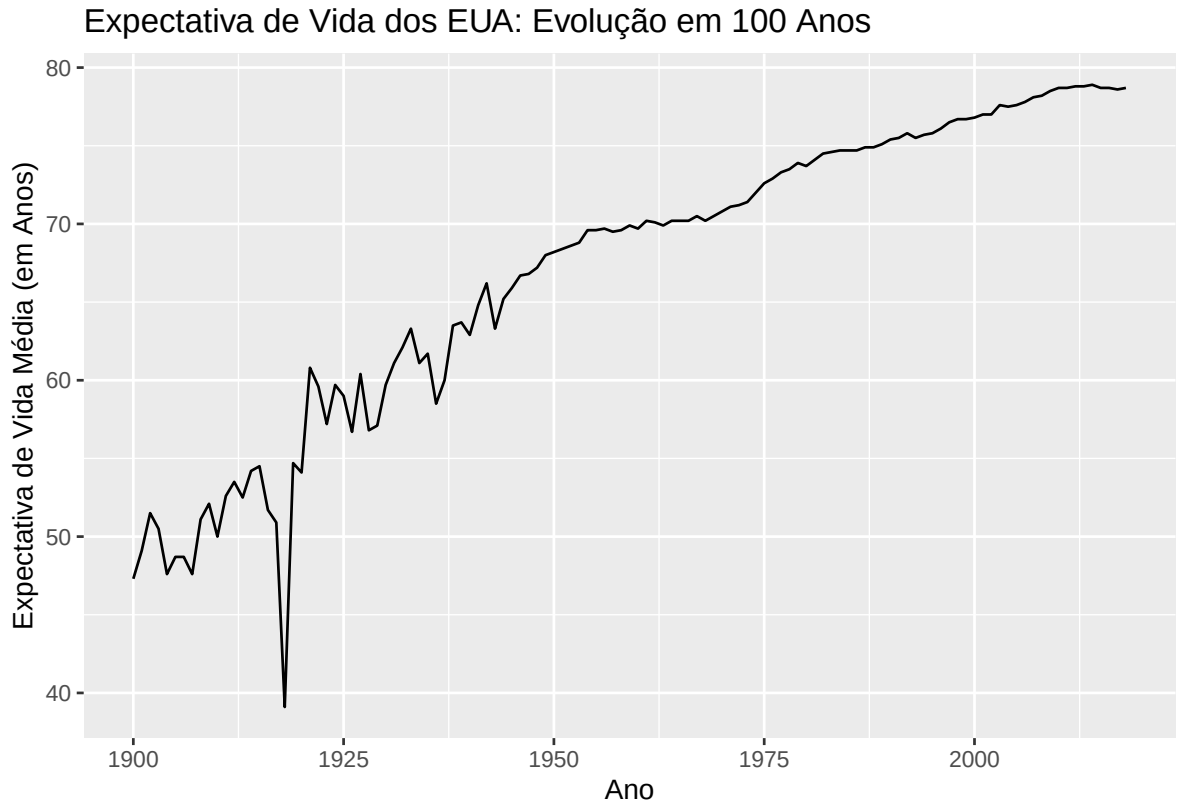
Camada 4: Rótulos

O gráfico parece completo, para nós que estamos invesstigando a base desde o início. Mas se formos apresentar esse gráfico apenas para alguém que desconhece a base, e precisa entender só de olhar a figura e quem sabe tomar decisões, as informações precisam estar claras.

Um exemplo é o eixo y se chamar Avg_Life_Expect, pode não ser óbvio o que significa ou até cada pessoa interpretar de uma forma diferente.

Para evitar esse tipo de situação utilizamos rótulos para melhorar a interpretação dos dados.

```
base %>% filter(Sex == "Both Sexes",
                Race == "All Races") %>%
         ggplot(aes(x=Year,y=Avg_Life_Expect)) +
          geom_line() +
          labs(
            title = "Expectativa de Vida dos EUA: Evolução em 100 Anos",
            x     = "Ano",
            y     = "Expectativa de Vida Média (em Anos)")
```

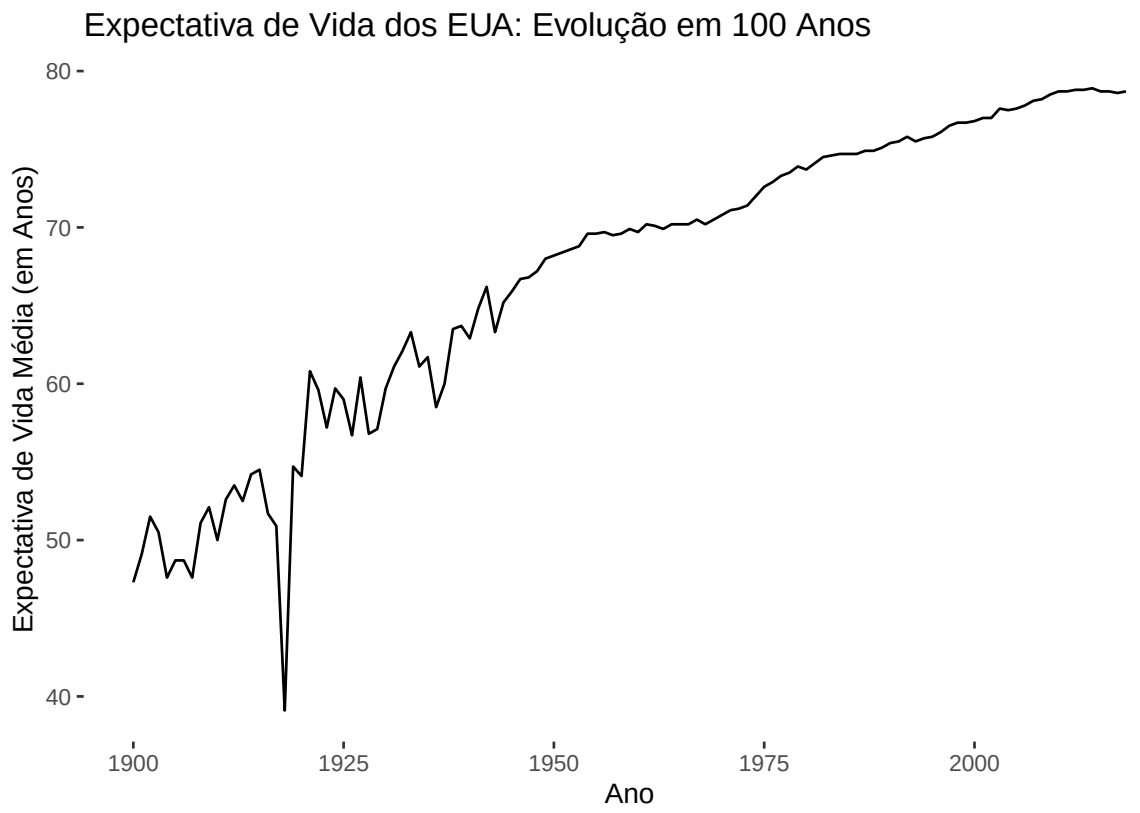


Personalização

Finalmente concluímos o gráfico, possui todas as camadas, possui informações claras e conseguimos fazer observações. Mas é claro que dependendo do trabalho, criar um gráfico mais sofisticado por fazer a diferença. E a biblioteca ggplot2 possui a função theme que permite estilizar nosso gráfico.

```
base %>% filter(Sex == "Both Sexes",
                Race == "All Races") %>%
         ggplot(aes(x=Year,y=Avg_Life_Expect)) +
```

```
        geom_line() +
        labs(
            title = "Expectativa de Vida dos EUA: Evolução em 100 Anos",
            x     = "Ano",
            y     = "Expectativa de Vida Média (em Anos)") +
        theme(
            panel.background = element_rect(fill = "white"))
```



Pronto! Agora o gráfico ficou mais limpo, sem excesso de elementos visuais e podemos focar no que interessa, nos dados!

Vemos que num geral a expectativa de vida aumentou, provavelmente graças a tecnologia e ciência. Um ponto que chama atenção é a queda brusca que torno de 1920 que provavelmente se deu pela pandemia de gripe que a população viveu naquela época.