Introdução à Estatística com Linguagem R

Ana Grasielle Dionísio Corrêa

Faculdade de Computação e Informática

Programa de Pós-Graduação em Distúrbios do Desenvolvimento

Universidade Presbiteriana Mackenzie

ana.correa@mackenzie.br

SUMÁRIO

CAPÍTULO 3. PREPARAÇÃO E ORGANIZAÇÃO DOS DADOS

O objetivo deste capítulo é apresentar e formatar os dados que iremos trabalhar ao longo deste livro. Usaremos como exemplo os dados do ENADE 2019 para fazer análises estatísticas com R. O Exame Nacional de Desempenho do Estudante (ENADE) integra o ciclo de avaliações do Sistema Nacional de Avaliação do Ensino Superior (SINAES) promovido pelo Ministério da Educação, ao qual todas as Instituições de Ensino Superior estão submetidas (ENADE 202). Seu objetivo é aferir o desempenho dos estudantes em relação aos conteúdos programáticos previstos nas diretrizes curriculares do respectivo curso de graduação, bem como em relação às habilidades e competências em sua formação. O resultado do exame é utilizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep/MEC) como indicador oficial que atribui visibilidade e credibilidade aos cursos universitários (INEP 2021).

3.1. OBTENDO MICRODADOS

Os microdados do Inep se constituem no menor nível de desagregação de dados recolhidos por suas pesquisas estatísticas, avaliações e exames. As informações podem ser obtidas via download, em formato ASCII, e contêm inputs (canais de entrada) para leitura por meio dos softwares de análises estatísticas como R, SAS e SPSS. Para abrir os arquivos, que estão em formato de compressão específico (.zip), é necessário o uso de algum programa descompactador.

Para ter acesso aos dados do ENADE é necessário acessar o site do INEP, na seção de Dados Abertos, disponível no link: https://www.gov.br/inep/pt-br/ace sso-a-informacao/dados-abertos/microdados/enade e baixar o arquivo "*Microdados do Enade 2019 | zip*". Ao descompactar o arquivo, será possível encontrar três diretórios:

1. LEIA-ME
   Contém os seguintes arquivos:
   - Manual do usuário 2019.pdf;
   - Questionário do Estudante_Enade_Edição 2019.pdf;
   - Dicionário de Variáveis dos Microdados do Enade 2019.ods;
   - Dicionário de Variáveis dos Microdados do Enade 2019.xlsx;
2. INPUTS
   Contém os seguintes INPUTS para leitura dos microdados:
   - Sas_input_enade_2019.sas;

- R_input_enade_2019.r;
- Spss_input_enade_2019.spss;

3. DADOS
   Contém o arquivo dos microdados do Enade 2019:
   - microdados_enade_2019.txt.

Nosso arquivo de referência para trabalhar com os dados do ENADE encontra-se dentro da pasta DADOS "microdados_enade_2019.txt" que contém todos os registros com mais de 430 mil linhas e 137 colunas.

Dentro da pasta LEIA-ME encontra-se o arquivo "*Dicionário de Variáveis dos Microdados do Enade 2019*" que iremos utilizar para saber o significado dos rótulos (siglas) das colunas da tabela que iremos analisar, bem como os valores em relação aos registros de cada coluna. Ao abrir o arquivo no software Excel, por exemplo, podemos ver (em uma das colunas) o nome das variáveis (ou siglas), por exemplo NU_ANO e sua respectiva descrição "Ano de realização do exame"; CO_IES "Código da IES (e-MEC)"; CO_CATEGAD "Código da categoria administrativa da IES"; CO_ORGACAD "Código da organização acadêmica da IES" e assim por diante.

Os valores das variáveis, ou seja, das colunas, podem variar. Por exemplo, os valores da coluna CO_ORGACAD "Código da organização acadêmica da IES" podem ter os seguintes valores:

- 10019 = Centro Federal de Educação Tecnológica
- 10020 = Centro Universitário
- 10022 = Faculdade
- 10026 = Instituto Federal de Educação, Ciência e Tecnologia
- 10028 = Universidade

Ao analisar o arquivo de dicionário do ENADE 2019, é possível notar que existem muitas informações. Precisaremos fazer uma filtragem nesses dados, visto que existem muitas informações que não são relevantes para nosso estudo inicial. Além disso, o arquivo é pesado, cada vez que tentamos rodar o arquivo original iremos levar alguns minutos de processamento tornando inviável (dependendo do processador do computador que estivermos utilizando) para o nosso objetivo neste estudo que é o de aprender a usar a ferramenta R para análises estatísticas. Logo, iremos filtrar somente o Estado de São Paulo que irá nos fornecer muitas informações interessantes.

Variáveis e Registros

Por convenção, neste livro iremos adotar o termo "variável" ao nos referirmos aos campos da tabela, ou seja, suas colunas (ex. NU_ANO, CO_ORGACAD, CO_GRUPO etc). Já o termo "registro" se refere aos valores atribuídos em cada campo, ou seja, as linhas da tabela (Figura 1).

NU_ANO	CO_ORGACAD	CO_GRUPO	CO_MODALIDADE	CO_MUNIC_CURSO
2019	10028	5710	1	3548906
2019	10028	5710	1	3548906
2019	10028	5710	1	3548906
2019	10028	5710	1	3548906
2019	10028	5710	1	3548906
2019	10028	5710	1	3548906

Figura 1. Tabela ENADE 2019

3.2. IMPORTANDO E FILTRANDO O BANCO DE DADOS

Antes de abrir o arquivo de microdados do Inep vamos preparar um script para trabalharmos com nossos dados. Primeiro será necessário instalar os pacotes "data.table" e "dyplr". Usaremos o comando *!require* para que a instalação seja feita somente se o pacote ainda não tiver sido instalado. Em seguida, é preciso carregar os pacotes usando o comando "library". Em alguns casos, o Inep solicita que seja feita uma alocação de memória para tratamento dos dados. Neste caso, podemos usar o comando *memory.limit* com o valor (em megabytes) sugerido pelo Inep. Na sequência precisamos fornecer o diretório onde o arquivo de dados será encontrado. Por exemplo, um diretório na raiz do computador (drive C:).

```
################################
#####  ABERTURA DO ARQUIVO  #####
################################

# Instalação dos pacotes (caso não estejam instalados)
#-------------------------------------------------------------
if(!require(data.table)){install.packages('data.table')}
if(!require(dplyr)){install.packages("dplyr")}

# Carregamento do pacote
library (data.table)
```

```
library(dplyr)

# Alocação de memória
#------------------------
memory.limit(24576)
#------------------------

# FORNECENDO DIRETÓRIO DO ARQUIVO
setwd("C:/CursoR")

#CARREGANDO TABLE INEP
ENADE_2019 <- data.table::fread(input='microdados_enade_2019.txt',
              integer64='character', #ler em caracteres
              skip=0,  #Ler do inicio
              nrow=-1, #Ler todos os registros
              na.strings = " ",
              showProgress = TRUE)
```

Em seguida, precisamos puxar os dados da tabela disponibilizada pelo Inep. Vamos criar a variável ENADE_2019 que será nossa tabela de dados (podemos chama-la de dataframe) que vai receber a tabela de dados do Inep com os seguintes parâmetros: input irá receber o arquivo de dados do Inep; integer64 para indicar se é caractere ou numérico; skip=0 indica que a leitura deve ser feita do início da tabela; nrow=-1 indica que será feita a leitura de todos os registros; na.strings indica os valores ausentes; showProgress para indicar se é necessário mostrar o progresso de carregamento do arquivo. Ao executar o trecho de comandos do quadro "CARREGANDO O ARQUIVO" será possível visualizar, na aba Console do RStudio, a barra de progresso indicando o carregamento do arquivo de dados através do sinal ===.

```
> # CARREGANDO O ARQUIVO
> setwd("C:/CursoR")
> ENADE_2019 <-
data.table::fread(input='microdados_enade_2019.txt',
+                 integer64='character', #ler em characteres
+                 skip=0,  #Ler do inicio
+                 nrow=-1, #Ler todos os registros
+                 na.strings = " ",
```

```
+                  showProgress = TRUE)
|----------------------------------------------------------------------|
|======================
```

Filtrando os dados

Neste estudo iremos realizar uma filtragem por Estado. Iremos trabalhar somente com os dados do Estado de São Paulo (SP). Logo, será preciso criar uma nova variável chamada, por exemplo, enade_SP_2019 para armazenar o subgrupo de dados de SP. É preciso indicar os dados carregados do ENADE_2019 seguido do comando "%>%" requerido para uso da função filter(). Também é necessário indicar a variável (coluna) referente ao Estado que queremos filtrar. Podemos consultar o arquivo de dicionário de dados para saber qual coluna se refere ao Estado brasileiro. No dicionário de dados está especificado que a coluna CO_UF_CURSO que significa "Código da UF de funcionamento do curso" possui o valor 35 apenas para os dados relacionados ao estado de SP.

```
# Filtrando apenas os dados do Estado de São Paulo
enade_sp_2019 <- ENADE_2019 %>% filter(CO_UF_CURSO =="35")
```

Após o carregamento completo, torna-se possível visualizar a tabela de dados ENADE_2019 carregada na memória do computador na aba "Environment" do Rstudio (Figura 2). Essa aba mostra que foram carregados 433.930 mil registros (linhas) e 137 variáveis (colunas). Após aplicar a filtragem, a tabela enade_sp_2019 é reduzida à uma quantidade menor de registros, no caso 105.763 registros (linhas).

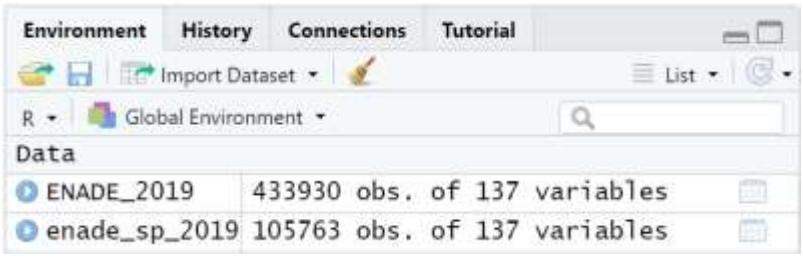

Figura 2. Carregamento da tabela de dados

Podemos também alterar a quantidade de variáveis (colunas) que queremos trabalhar. Nesse caso, precisamos alterar a tabela enade_sp_2019 e selecionar somente as variáveis (colunas) que desejarmos trabalhar neste estudo. A Tabela 2 apresenta algumas variáveis encontradas no dicionáio de microdados do Enade 2019 e que são interessantes para trabalharmos em nossa análise.

Tabela 1. Tabela de microdados do Enade 2019

PARTE 1 - INFORMAÇÕES DA INSTITUIÇÃO DE ENSINO SUPERIOR E DO CURSO	
NU_ANO	Ano de realização do exame
CO_ORGACAD	Código da organização acadêmica da IES
CO_GRUPO	Código da Área de enquadramento do curso no Enade
CO_MODALIDADE	Código da Modalidade de Ensino
CO_MUNIC_CURSO	Código do município de funcionamento do curso
CO_UF_CURSO	Código da UF de funcionamento do curso
CO_REGIAO_CURSO	Código da região de funcionamento do curso
TP_INSCRICAO	Tipo de inscrição
PARTE 2 - INFORMAÇÕES DO ESTUDANTE	
NU_IDADE	Idade do inscrito em 24/11/2019
TP_SEXO	Sexo
ANO_FIM_EM	Ano de conclusão do Ensino Médio
ANO_IN_GRAD	Ano de início da graduação
CO_TURNO_GRADUACAO	Código do turno de graduação
PARTE 5 - TIPOS DE PRESENÇA	
TP_PRES	Tipo de presença no Enade
TP_PR_GER	Tipo de presença na prova
PARTE 6 - TIPOS DE SITUAÇÃO DAS QUESTÕES DA PARTE DISCURSIVA	
TP_SFG_D1	Tipo de situação da questão 1 da parte discursiva da formação geral
TP_SFG_D2	Tipo de situação da questão 2 da parte discursiva da formação geral
TP_SCE_D1	Tipo de situação da questão 1 da parte discursiva do componente específico
TP_SCE_D2	Tipo de situação da questão 2 da parte discursiva do componente específico
TP_SCE_D3	Tipo de situação da questão 3 da parte discursiva do componente específico
PARTE 7 - NOTAS NA FORMAÇÃO GERAL E COMPONENTE ESPECÍFICO	
NT_GER	Nota bruta da prova - Média ponderada da formação geral (25%) e componente específico (75%). (valor

	de 0 a 100)
NT_FG	Nota bruta na formação geral - Média ponderada da parte objetiva (60%) e discursiva (40%) na formação geral. (valor de 0 a 100)
NT_OBJ_FG	Nota bruta na parte objetiva da formação geral. (valor de 0 a 100)
NT_DIS_FG	Nota bruta na parte discursiva da formação geral. (valor de 0 a 100)
NT_FG_D1	Nota da questão 1 da parte discursiva da formação geral - Média ponderada da parte de Língua Portuguesa (20%) e Conteúdo (80%) da Questão 1 da parte discursiva (valor de 0 a 100)
NT_FG_D1_PT	Nota de Língua Portuguesa da questão 1 da parte discursiva da formação geral. (valor de 0 a 100)
NT_FG_D1_CT	Nota de Conteúdo da questão 1 da parte discursiva da formação geral. (valor de 0 a 100)
NT_FG_D2	Nota da questão 2 da parte discursiva na formação geral - Média ponderada da parte de Língua Portuguesa (20%) e Conteúdo (80%) da Questão 2 da parte discursiva. (valor de 0 a 100)
NT_FG_D2_PT	Nota de Língua Portuguesa da questão 2 da parte discursiva da formação geral. (valor de 0 a 100)
NT_FG_D2_CT	Nota de Conteúdo da questão 2 da parte discursiva da formação geral. (valor de 0 a 100)
NT_CE	Nota bruta no componente específico - Média ponderada da parte objetiva (85%) e discursiva (15%) no componente específico. (valor de 0 a 100)
NT_OBJ_CE	Nota bruta na parte objetiva do componente específico. (valor de 0 a 100)
NT_DIS_CE	Nota bruta na parte discursiva do componente específico. (valor de 0 a 100)
NT_CE_D1	Nota da questão 1 da parte discursiva do componente específico. (valor de 0 a 100)
NT_CE_D2	Nota da questão 2 da parte discursiva do componente específico. (valor de 0 a 100)
NT_CE_D3	Nota da questão 3 da parte discursiva do componente específico. (valor de 0 a 100)

O código a seguir mostra como selecionar as variáveis de interesse. Os códigos podem ser consultados na planilha "Dicionário de variáveis dos Microdados do Enade 2019" (Tabela 1).

```
# Selecionando as colunas de interesse ENADE
enade_sp_2019 <- select(enade_sp_2019, NU_ANO, CO_ORGACAD,
CO_GRUPO, CO_MODALIDADE, CO_MUNIC_CURSO, CO_UF_CURSO,
CO_REGIAO_CURSO, NU_IDADE, TP_SEXO, ANO_FIM_EM, ANO_IN_GRAD,
CO_TURNO_GRADUACAO, TP_INSCRICAO, TP_PRES, TP_PR_GER, TP_SFG_D1,
TP_SFG_D2, TP_SCE_D1, TP_SCE_D2, TP_SCE_D3, NT_GER, NT_FG,
NT_OBJ_FG, NT_DIS_FG, NT_FG_D1, NT_FG_D1_PT, NT_FG_D1_CT, NT_FG_D2,
NT_FG_D2_PT, NT_FG_D2_CT, NT_CE, NT_OBJ_CE, NT_DIS_CE, NT_CE_D1,
NT_CE_D2, NT_CE_D3)

#Visualizando a tabela após a filtragem
View(enade_sp_2019)

#Exportando o arquivo .csv
write.table(enade_sp_2019, file ="enade_sp_2019.csv", sep = ",")
```

A Figura 3 mostra a quantidade de variáveis da tabela "enade_sp_2019" após aplicar a filtragem das variáveis (colunas). O número de variáveis diminuiu de 137 para 36 das quais que iremos trabalhar.

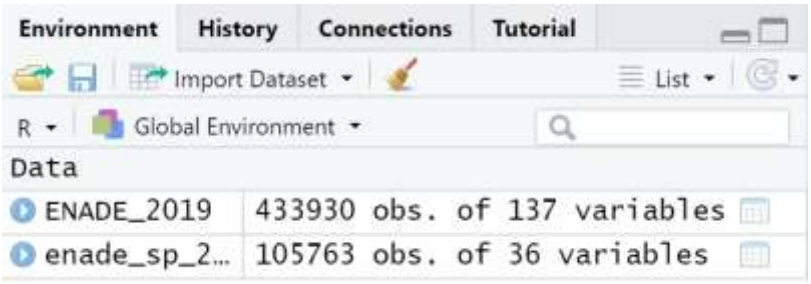

Figura 3. Carregamento da tabela de dados após filtragem

Em seguida, podemos usar o comando "view(enade_sp_2019)" para visualizar todas as variáveis que foram selecionadas (Figura 4).

	NU_ANO	CO_ORGACAD	CO_GRUPO	CO_MODALIDADE	CO_MUNIC_CU
1	2019	10028	5710	1	
2	2019	10028	5710	1	
3	2019	10028	5710	1	
4	2019	10028	5710	1	
5	2019	10028	5710	1	
6	2019	10028	5710	1	
7	2019	10028	5710	1	

Figura 4. Visualizando os dados selecionados

Por último, após aplicação do filtro, é preciso exportar a nova tabela de dados usando o comando "write.table" passando como parâmetro a tabela que queremos exportar, no caso, "enade_2019_sp" e o nome do arquivo .csv que iremos exportar (pode ser o mesmo nome atribuído à tabela de dados). Após executar o comando, o arquivo .scv (Arquivo de Valores Separados por Vírgulas) do Microsoft Excel aparecerá no diretório onde a tabela foi carregada, neste caso, em "C:/CursoR".

Criamos, portanto, nossa tabela de dados final "enade_sp_2019" contendo 105.763 registros (linhas) e 35 variáveis (colunas). Iremos utilizar essa tabela em nossas análises futuras.

3.3. ORGANIZANDO E ESTRUTURANDO OS DADOS

A fase de organização e estruturação dos dados é também conhecida como pré-processamento. O primeiro passo é criar um novo script e, na sequência, verificar se há necessidade de instalar e carregar algum pacote. Como iremos trabalhar com manipulação de dados, o pacote indicado é o "dplyr". Em seguida, é preciso abrir a pasta onde está o arquivo enade_sp_2019.csv e carregar o arquivo usando o comando "read.csv". O script a seguir mostra o carregamento da tabela enade_sp_2019.csv para a variável (tabela) enade_2019. É com essa tabela (somente com dados de SP) que iremos trabalhar. Por último, podemos visualizar a tabela enade_2019 usando o comando View.

```
################################################
# ORGANIZAÇÃO E ESTRUTURAÇÃO DOS DADOS (PRÉ-PROCESSAMENTO)   #
################################################

# BAIXAR PACOTES
install.packages("dplyr") # Manipulação de Dados
# OU
# BAIXAR PACOTES, CASO ELES AINDA NÃO ESTEJAM BAIXADOS
if(!require(dplyr)) install.packages("dplyr")

# CARREGAR PACOTES
library(dplyr)

# BUSCAR DIRETÓRIO (PASTA COM OS ARQUIVOS)
setwd("C:/CursoR")

#ABRIR ARQUIVO
enade_2019 <- read.csv('enade_sp_2019.csv')

#VISUALIZAR TABELA
View(enade_2019)
```

Excluindo campos (colunas)

Propositalmente, a tabela enade_sp_2019.csv (agora enade_2019), foi criada com alguns campos (colunas) desnecessários para que possamos aprender a excluir colunas usando R. Por exemplo, NU_ANO não é necessário, uma vez que já sabemos previamente que a tabela do Inpe contém somente dados do Enade de 2019. Não tem sentido, portanto, manter a tabela com a coluna que contém somente valores iguais a 2019. Idem para a variável CO_UF_CURSO que contém somente valor igual a 35, ou seja, Estado de SP. Em R, podemos excluir apenas uma coluna ou várias colunas ao mesmo tempo, como mostra o script a seguir.

```
# EXCLUIR UMA COLUNA
enade_2019$NU_ANO <- NULL
#OU
# EXCLUIR VÁRIAS COLUNAS
excluir <- c("NU_ANO", "CO_UF_CURSO")
```

```
View(excluir)
enade_2019 <- enade_2019[  , !(names(enade_2019) %in% excluir)]

#VISUALIZAR TABELA
View (enade_2019)
```

Para excluir várias colunas precisamos criar um vetor contendo as colunas que desejamos excluir. No exemplo foi criado um vetor chamado "excluir". O comando View(excluir) foi usado apenas para visualizar o vetor com as duas colunas que serão excluídas. Para efetivamente excluir as duas colunas da tabela de dados precisamos sobrescrever a tabela de dados "enade_2019" usando o operador "%in%" que verifica a interseção entre linhas ou vetores. O comando "enade_2019[, !(names(enade_2019) %in% excluir)]" significa que queremos todas as linhas "[," e todas as colunas "(names(enade_2019)" da tabela enade_2019, exceto "!" a interceção "%in%" com as colunas do vetor excluir, no caso as colunas NU_ANO e CO_UF_CURSO. Perceba que, ao rodar as linhas do script " *EXCLUIR VÁRIAS COLUNAS*", o número de colunas diminui de 35 para 33.

Renomeando campos (colunas)

Muitas vezes é preciso renomear alguns campos da tabela com vista a facilitar o entendimento. Por exemplo, vamos renomear os campos NU_IDADE e TP_SEXO para IDADE e SEXO respectivamente. Em R, é possível renomear os campos individualmente ou vários campos simultaneamente como apresentado no script a seguir. Por último, podemos visualizar a tabela com os campos renomeados.

```
#RENOMEAR UMA COLUNA
enade_2019 <- rename(enade_2019, IDADE = NU_IDADE)

#RENOMEAR VÁRIAS COLUNAS
enade_2019 <- rename(enade_2019, NU_IDADE = IDADE, TP_SEXO = SEXO,
ORGACAD = CO_ORGACAD, GRUPO = CO_GRUPO, CO_MODALIDADE = MODALIDADE,
CO_MUNIC_CURSO = MUNIC_CURSO, CO_REGIAO_CURSO = REGIAO_CURSO,
CO_TURNO_GRADUACAO = TURNO_GRADUACAO)

#VISUALIZAR TABELA
View (enade_2019)
```

Analisando a tipagem dos Atributos (Variáveis)

Existem seis tipos de variáveis em R: *character* (caracteres); *numeric* (números reais); *integer* (números inteiros); *logical* (falso ou verdadeiro); *complex* (números complexos); *factor* (fator: ordenar strings).

Precisamos conhecer como o R reconheceu cada uma das variáveis que foram carregadas da tabela do Inep. Podemos fazer isso de duas formas: usando os comandos str ou glimpse, conforme mostra o script a seguir.

```
# VISUALIZANDO TIPAGENS DAS VARIÁVEIS
str(enade_2019)
# OU
glimpse(enade_2019)
```

Ao rodar um dos comandos, todas as variáveis e seus respectivos tipos são apresentados na janela de Console do RStudio. Na tabela enade_2019 temos dois tipos: <int> que significa inteiro; <chr> que significa <character>.

```
Columns: 33
$ ORGACAD         <int> 10028, 10028, 10028, 10028, 10028,...
$ GRUPO           <int> 5710, 5710, 5710, 5710, 5710, 5710...
$ MODALIDADE      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
$ MUNIC_CURSO     <int> 3548906, 3548906, 3548906, 3548906...
$ REGIAO_CURSO    <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3...
$ IDADE           <int> 23, 26, 24, 27, 24, 28, 22, 23, 26...
$ SEXO            <chr> "F", "M", "F", "M", "M", "M", "M",...
$ ANO_FIM_EM      <int> 2013, 2011, 2013, 2010, 2013, 2008...
$ ANO_IN_GRAD     <int> 2015, 2013, 2014, 2013, 2014, 2013...
$ TURNO_GRADUACAO <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3...
$ TP_PRES         <int> 555, 555, 555, 555, 555, 555, 555,...
$ TP_PR_GER       <int> 555, 555, 555, 555, 555, 555, 555,...
$ TP_SFG_D1       <int> 555, 555, 555, 555, 555, 555, 555,...
$ TP_SFG_D2       <int> 555, 555, 555, 555, 555, 555, 555,...
$ TP_SCE_D1       <int> 555, 555, 555, 555, 555, 555, 555,...
$ TP_SCE_D2       <int> 333, 333, 555, 555, 555, 555, 555,...
$ TP_SCE_D3       <int> 333, 555, 555, 333, 555, 555, 555,...
$ NT_GER          <chr> "33,8", "54", "51,7", "70,2", "38,...
$ NT_FG           <chr> "55,8", "33,7", "76,4", "64,6", "4...
```

```
$ NT_OBJ_FG       <chr> "71,4", "42,9", "85,7", "100", "57...
$ NT_DIS_FG       <chr> "32,5", "20", "62,5", "11,5", "30,...
$ NT_FG_D1        <int> 16, 10, 91, 9, 7, 46, 8, 0, 0, 49,...
$ NT_FG_D1_PT     <int> 80, 50, 75, 45, 35, 50, 40, 0, 0, ...
$ NT_FG_D1_CT     <int> 0, 0, 95, 0, 0, 45, 0, 0, 0, 45, 4...
$ NT_FG_D2        <int> 49, 30, 34, 14, 54, 12, 18, 0, 29,...
$ NT_FG_D2_PT     <int> 45, 50, 70, 70, 70, 60, 90, 0, 45,...
$ NT_FG_D2_CT     <int> 50, 25, 25, 0, 50, 0, 0, 0, 25, 0,...
$ NT_CE           <chr> "26,5", "60,8", "43,5", "72", "36,...
$ NT_OBJ_CE       <chr> "30", "65", "50", "80", "35", "70"...
$ NT_DIS_CE       <chr> "6,7", "36,7", "6,7", "26,7", "43,...
$ NT_CE_D1        <chr> "20", "10", "20", "20", "30", "60"...
$ NT_CE_D2        <int> 0, 0, 0, 60, 0, 0, 30, 100, 0, 0, ...
$ NT_CE_D3        <chr> "0", "100", "0", "0", "100", "100"...
```

Se for preciso, podemos alterar o tipo da variável, por exemplo, o código do município de funcionamento do curso (MUNIC_CURSO) foi carregado como inteiro <int>, mas não iremos fazer operações matemáticas com esse código. Logo, podemos transforma-lo em fator (factor <fct>). Para alterar o tipo da variável precisamos usar o operador "$" para indicar a coluna e o operador "as." seguido do tipo conforme mostra o script a seguir. Iremos fazer mais dessas alterações mais adiante quando estivermos aplicando a análise nos dados.

```
#ALTERANDO O TIPO DAS VARIÁVEIS
enade_2019$MUNIC_CURSO <- as.factor(enade_2019$MUNIC_CURSO)
```

Outro problema que podemos enfrentar é a presença de valores ausentes, conhecidos em R como "*missings*". Por exemplo, nos campos referentes às notas, temos vários valores "NA" ou "NAN", ou seja, valores ausentes indefinidos (erro numérico) respectivamente. Os valores "NA" ou "NAN" são diferentes do valor zero (valor zero no campo nota significa que a nota de um determinado aluno foi zero). Em R nós podemos obter a quantidade de valores ausentes NA e NAN usando o comando *sapply* como mostra o script a seguir.

```
# VERIFICANDO VALORES MISSING (AUSENTES OU INDEFINIDOS)
# NA = valores ausentes
# NAN = not a number(valor indefinido)
sapply(enade_019, function(x) sum(is.na(x)))
```

```
sapply(enade_019, function(x) sum(is.nan(x)))
```

Após rodar o comando sapply, é possível visualizar a quantidade de valores ausentes ou indeterminados na janela de console, por exemplo existem 96.453 valores ausentes para a variável Nota bruta da Prova (NT_GER). De acordo com o dicionário de variáveis dos microdados, valores ausentes podem ser decorrentes de ausências ou desclassificações nas provas. Neste caso, as notas zero não são contabilizadas. Na nossa tabela "enade_2019" não foram encontrados valores indeterminados (NAN), o que significa que teremos um tratamento a menos a se preocupar. Quando formos de fato realizar o estudo estatístico, não podemos considerar os valores ausentes, neste caso teremos que retirar esses valores ou substituir o valor ausente por um outro valor que não irá interferir na análise.

Calculando estudantes concluintes

Foram inscritos no exame ENADE 2019 os estudantes ingressantes e concluintes dos cursos que fazem parte das áreas avaliadas em 2019. Os ingressantes ficam dispensados de participar do exame, mas precisam ser inscritos. Já os concluintes precisam participar para terem o direito a colar grau. Vamos então analisar os estudantes concluintes. No dicionário de microdados do Enade está indicado que os estudantes concluintes possuem valores igual a zero. Vamos então excluir do nosso estudo os estudantes que não eram concluintes. Para isso, é preciso criar uma variável "concluintes" para receber a tabela enade_2019 com o filtro da coluna TP_INSCRICAO (tipo de inscrição) apenas os estudantes concluintes, ou seja, os que estiverem representados com valor zero. Por coincidência, todos os estudantes de SP foram inscritos como concluintes, totalizando 105.763 estudantes.

```
#VERIFICANDO QUANTIDADE DE CONCLUINTES
concluintes <- enade_2019 %>% filter(TP_INSCRICAO==0)
```

Caso tivéssemos um número menor de concluintes seria então necessário excluir (filtrar) os estudantes ingressantes. Então seria necessário prosseguir com o script de exclusão apresentado a seguir. Note que, mesmo excluindo os registros de ingressantes, é interessante armazená-lo em um arquivo .csv. Dessa forma, caso seja necessário usar essa informação futuramente, ela poderá ser recuperada.

```
#CALCULAR TOTAL DE INGRESSANTES
ingressantes <- enade_2019 %>% filter(TP_INSCRICAO==1)

#RETIRAR INGRESSANTES COM FILTRO
concluintes <- enade_2019 %>% filter(TP_INSCRICAO==0)

# EXCLUIR UMA COLUNA
concluintes$TP_INSCRICAO <- NULL

#EXPORTANDO ARQUIVO DE INGRESSANTES
write.table(ingressantes, file ="ingressantes.csv", sep = ",")
```

Removendo estudantes desclassificados

Consideram-se válidos para os procedimentos de cálculo do Enade 2019 apenas os resultados dos concluintes inscritos regularmente pelas IES e com presença atestada no Exame. Estes estudantes possuem a variável "Tipo de Presença" no Enade (TP_PRES=555) na base de Microdados do Enade. Estudantes ausentes (TP_PRES=222), com inscrição indevida (TP_PRES=333); eliminado por participação indevida (TP_PRES=334), ausente devido a dupla graduação (TP_PRES=444) ou com resultados desconsiderados pela empresa aplicadora (TP_PRES=556) não são considerados para o cálculo do Conceito Enade. Portanto, precisamos remover da nossa tabela enade_2019 os estudantes que foram desclassificados da prova. Após rodar o script a seguir, os registros diminuirão de 105.763 para 93.225 (número de inscritos válidos).

```
#FILTRANDO POR TIPO DE PRESENÇA
presenca_prova <- enade_2019 %>% filter(TP_PRES==555)
View (presenca_prova)
```

Eliminando notas ausentes

Depois de todo tratamento realizado nos dados da nossa tabela enade_2019, ainda é preciso verificar se existem valores ausentes (NA ou NAN) no campo presenca_prova. Ao rodar o script a seguir iremos constatar que existem valores ausentes *(missings)* encontrados nos campos respectivos às notas dos estudantes. Todos os campos (NT_DIS_FG, NT_FG_D1, NT_FG_D1_PT, NT_FG_D1_CT, NT_FG_D2, NT_FG_D2_PT, NT_FG_D2_CT, NT_CE, NT_OBJ_CE, NT_DIS_CE, NT_CE_D1, NT_CE_D2, NT_CE_D3) estão com 30 valores ausentes (ou seja, NA).

Considerando que o valor 30 corresponde apenas a 0,3% de 93.225 registros (dados filtrados), devemos eliminar esses 30 registros, pois poderão afetar futuramente nossa análise estatística.

```
# VERIFICANDO VALORES MISSING (AUSENTES OU INDEFINIDOS)
# NA = valores ausentes
# NAN = not a number(valor indefinido)
sapply(presenca_prova, function(x) sum(is.na(x)))
sapply(presenca_prova, function(x) sum(is.nan(x)))
```

Para excluir os valores NA dos campos das notas, devemos usar o comando "dorp_na" em todos os campos das notas conforme mostra o script a seguir. Eliminando os valores ausentes do campo NT_GER irá fazer com que todos os outros campos referentes às notas também sejam eliminados. Após rodar o comando sampply, é possível notar que os valores do nosso dataframe presenca_prova caiu de 93.225 registros para 93.125 registros.

```
#EXCLUINDO VALORES AUSENTES DOS CAMPOS DE NOTAS
presenca_prova <- drop_na(presenca_prova, NT_GER)
sapply(presenca_prova, function(x) sum(is.na(x)))
```

Verificando quantidade de notas zeradas

Notas zeradas podem interferir nos nossos estudos futuros. É recomendado, portanto, verificar a quantidade de notas zero nos campos das notas. Podemos apenas criar uma variável, por exemplo, nota_zero, e ir verificando campo a campo a quantidade de notas zero atribuídas e adicionando os valores como comentários ao lado de cada linha.

```
#OBTENDO QUANTIDADE DE NOTAS ZERO (opcional)
nota_zero <- presenca_prova %>% filter(NT_GER==0) # 26 notas zeros
nota_zero <- presenca_prova %>% filter(NT_DIS_FG==0) # 7099
nota_zero <- presenca_prova %>% filter(NT_FG_D1==0) # 9847
nota_zero <- presenca_prova %>% filter(NT_FG_D1_PT==0) # 55722
nota_zero <- presenca_prova %>% filter(NT_FG_D1_CT==0) # 55722
nota_zero <- presenca_prova %>% filter(NT_FG_D2==0)# 14239
nota_zero <- presenca_prova %>% filter(NT_FG_D2_PT==0) # 13868
nota_zero <- presenca_prova %>% filter(NT_FG_D2_CT==0) # 59217
```

```
nota_zero <- presenca_prova %>% filter(NT_CE==0) # 57
nota_zero <- presenca_prova %>% filter(NT_OBJ_CE==0) # 98
nota_zero <- presenca_prova %>% filter(NT_DIS_CE==0) # 8339
nota_zero <- presenca_prova %>% filter(NT_CE_D1==0) # 18612
nota_zero <- presenca_prova %>% filter(NT_CE_D2==0) # 29247
nota_zero <- presenca_prova %>% filter(NT_CE_D3==0) # 31250
```

É possível observar que as questões discursivas (NT_FG_D1_PT e NT_FG_D1_CT) de língua portuguesa e conteúdo geral foram as que tiveram maior quantidade de notas zero (>55mil), seguido das questões discursivas de componentes específicos (NT_DIS_D1, NT_DIS_D2, NT_DIS_D3). Por enquanto, não podemos excluir essas notas zeradas, faremos essa análise futuramente.

Exportando dados tratados

Os dados já estão filtrados e perfeitos para estudo estatístico. O próximo passo é exportá-lo como arquivo .csv, por exemplo, como o nome "enade_2019_tratado", conforme mostra o script a seguir. Após exportá-lo, procure pelo arquivo salvo na pasta do computador.

```
#EXPORTAR ARQUIVO TRATADO
write.table(presenca_prova, file ="enade_2019_tratado.csv", sep = ",")
```

Agora que nossos dados estão organizados e tratados, podemos dar início aos estudos estatísticos.

3.4. TAREFA → CRIAÇÃO DE DATAFRAMES

Crie outros dois data frames (tabela de dados) por exemplo, um com outro estado e outro com mais de um estado, utilizando como referência os scripts apresentados nessa seção.

CRONOGRAMA DE EXECUÇÃO

As atividades previstas neste projeto serão desenvolvidas durante um período de 12 meses, a contar de janeiro de 2021 (Tabela 1).

Observações:

- Os Capítulos 1, e 2 já estão finalizados;
- O Capítulo 3 está em faz de finalização.
- Pretende-se finalizar todo o conteúdo até 31 de abril de 2021
- Oferecimento do minicurso aos alunos de pós-graduação em maio de 2021 seguindo o conteúdo do curso para testes e ajustes.
- Oferecimento do minicurso aos alunos de pós-graduação em setembro/outubro de 2021 seguindo o conteúdo do curso para novos testes e ajustes.
- Finalização do livro em novembro e dezembro de 2021.

Tabela 1. Cronograma de Atividades

Atividades	**Meses**											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Capítulo 1 Linguagem R e RStudio: 1.1. Instalação; 1.1. Primeiros passos	**x**											
Capítulo 2 Preparação e Organização dos Dados: 2.1. Importando e Filtrando Banco de Dados; 2.2. Preparando e Organizando Dados em R; Tarefa 1: Criação de Novo Banco de Dados	**x**											
Capítulo 3: Introdução à Estatística Conceitos gerais: 3.1. Amostragem Aleatória Simples em R; 3.2. Amostragem estratificada em R; 3.3. Amostragem por Agrupamento em R; 3.4. Tarefa 2: Tipos de Amostragem.		**x**										
Capítulo 4: Estatística Descritiva: 4.1. Distribuição de Frequências (Teoria); 4.2. Distribuição de Frequências em R; 4.3. Gráficos de Frequência em R; 4.4. Medidas de Tendência Central (Teoria); 4.5 Medidas de Tendência Central em R; 4.6. Medidas de Dispersão e Posição (Teoria); 4.7. Medidas de Dispersão e Posição em R; 4.8. Tarefa 2: Estatística Descritiva.			**x**									
Capítulo 5 Estatística Inferencial: 5.1. Intervalo de confiança (Teoria); 5.1. Distribuição t e Qui-quadrado (Teoria); 5.2. Teste Z para uma amostra (Teoria); 5.3. Teste Z para uma amostra em R; 5.3. Teste t para uma amostra (Teoria); 5.4. Teste t para uma amostra em R; 5.5. Teste de Hipótese para duas Amostras Independentes (Teoria); 5.6. Teste t de Student em R; 5.7. Teste Z para duas Amostras Independentes em R;				**x**								

5.8. Teste t pareado; 5.9. Teste t pareado em R; 5.10. Regressão Linear (Teoria); 5.11. Regressão Linear em R; 5.12. Teste Qui-quadrado (Teoria); 5.13. Teste Qui-quadrado em R; 5.14. Teste ANOVA (teoria); 5.15. Teste ANOVA 1 via em R; 5.16. Teste ANOVA 2 vias em R; 5.17. Testes Não Paramétricos (Teoria). 5.18. Teste Mann_Whitney (não paramétrico) em R; 5.19. Teste de Wilcoxon em R; 5.20. Teste de Kruskal_Wallis no R; 5.21. Correlação de Spearman e Kendall no R;												
Oferecimento do curso aos alunos de pós-graduação do PPG-DD (1º Sem 2021).					**x**							
Ajustes de conteúdo seguindo o conteúdo do curso para testes e ajustes						**x**	**x**	**x**				
Oferecimento do curso aos alunos de pós-graduação do PPG-DD (2º Sem 2021)									**x**	**x**		
Ajustes de conteúdo seguindo o conteúdo do curso para testes e ajustes											**x**	
Finalização e submissão do livro												**x**

GLOSSÁRIO

ASCII	American Standard Code for Information Interchange (ASCII), código binário que codifica um conjunto de 128 sinais: 95 sinais gráficos e 33 sinais de controle, utilizando 7 bits para representar todos os seus símbolos.
CSV	formato de arquivo que significa “comma-separated-values” (valores separados por vírgulas)
Dataframe	estrutura de dados rotulada bidimensional (tabela) com colunas de tipos potencialmente diferentes.
ENADE	Exame Nacional de Desempenho do Estudante
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
Microdados	Tabela de dados de menor nível de desagregação de dados recolhidos por suas pesquisas estatísticas, avaliações e exames
R	Ferramenta livre e de código aberto usada para facilitar análises estatísticas, cálculos e manipulações gráficas em muitos campos do conhecimento, seja para fins acadêmicos ou para o mercado
RStudio	Ambiente de desenvolvimento integrado (Integrated Development Environment - IDE) para análise estatística em R.

SAS	conjunto de software estatístico desenvolvido pelo SAS Institute para gerenciamento de dados, análise avançada, análise multivariada, inteligência de negócios, investigação criminal e análise preditiva.
SPSS	Statistical Package for the Social Sciences, plataforma de software estatístico da IBM

REFERÊNCIAS

ALCOFORADO, Luciane Ferreira. Utilizando A Linguagem R: Conceitos, manipulação, visualização, modelagem e elaboração de relatórios. Alta Books, 2021.

BONAFINI, F. C. Estatística. São Paulo: Pearson Education do Brasil, 2012.

ENADE - Exame Nacional de Desempenho do Estudante. Disponível em http://enade.inep.gov.br/enade/. Acesso em 10 de março de 2021.

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Disponível em https://enem.inep.gov.br/. Acesso em 10 de março de 2021.

MORETTIN, L. G. Estatística básica: probabilidade e inferência: volume único. São Paulo: Pearson Prentice Hall, 2010

REISEN, V. A.; SILVA, A. N. O uso da linguagem R para cálculos de estatística básica. Vitória, ES: EDUFES, 2011.

The Comprehensive R Archive Network. Disponível em: https://cran.r-project.org/. Acesso em 10 de Março de 2021.

VENABLES, W. N. Venables, SMITH, D. M. Smith (the R Core Team). An Introduction to R. Version 4.0.4 (2021-02-15). Disponível em https://cran.r-project.org/. Acesso em 10 de Março de 2021.