



DESAFIO DE INFERÊNCIA



FERNANDA PERES
PROFESSORA E
CONSULTORA NA ÁREA DE
ANÁLISE DE DADOS



AGENDA

- **Entendendo o problema**
- **Sugestão de roteiro**
- **Resolução do desafio**
- **Resumo da aula**

T



ENTENDENDO O PROBLEMA

A vertical bar on the left side of the slide with a gradient from light green at the top to light blue at the bottom.

Transtorno depressivo

Problema de saúde pública (prevalência de $\approx 10\%$)

Multifatorial (fatores sociais, psicológicos, biológicos)

Estilo de vida saudável (OMS)

Prática regular de atividade física

Alimentação saudável

Os dados

Resultados da pesquisa NHNES (*National Health and Nutrition Examination Survey*), realizada anualmente nos EUA para avaliar a **saúde e nutrição** de adultos e crianças. Inclui dados demográficos, socioeconômicos, dietéticos e relacionados à saúde.
Período de 2005-2006

Perguntas a serem respondidas

- 1) Qual o **perfil** de indivíduos (adultos maiores de 18 anos) com sintomas depressivos nos EUA no período de 2005-2006?
- 2) Hábitos saudáveis de alimentação e atividade física estão **associados** a menores índices de depressão nesta população?

Duas bases de dados

DEMO_PHQ.csv

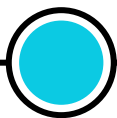
Dados **demográficos** e resultados do *Patient Health Questionnaire-9* (PHQ-9), usado para avaliar o grau de **sintomas depressivos**. Inclui apenas adultos.

PAG_HEI.csv

Dados referentes à **atividade física** e ao Healthy Eating Index (HEI), um questionário que avalia a **qualidade da dieta**. Inclui adultos e crianças.



SUGESTÃO DE ROTEIRO




**TRATAMENTO
DOS DADOS**

**ANÁLISE
EXPLORATÓRIA
(EDA)
UNIVARIADA**

**ANÁLISE
EXPLORATÓRIA
(EDA)
BIVARIADA**

**TESTES DE
HIPÓTESES**

**ANÁLISE
CRÍTICA DOS
RESULTADOS**


A vertical bar on the left side of the slide, transitioning from green at the top to blue at the bottom.

Faça a **leitura** dos bancos e os **tratamentos** que julgar necessários nas variáveis.

Por exemplo, você pode realizar o tratamento das categorias 7 = “Se recusou a responder” e 9 = “Não sabe”.

Combine os dois bancos fornecidos, utilizando a variável SEQN como chave única.

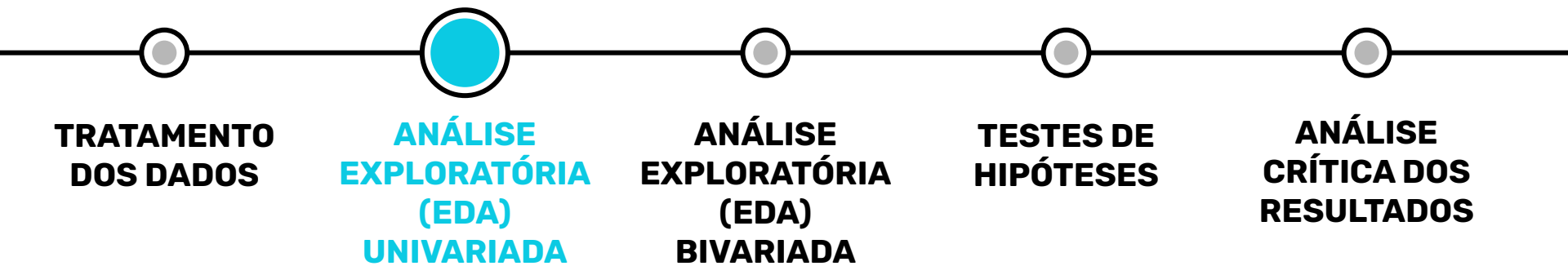
O banco de dados final deve conter 5334 observações dos adultos respondentes do NHANES.




Crie a variável **phq9**, correspondente ao escore do questionário PHQ-9, somando as variáveis DPQ010 a DPQ090.

Crie a variável **phq_grp** com a classificação do escore **phq9**, assumindo os valores:

- 0 (“sem sintomas”) se $\text{phq9} < 5$
- 1 (“sintomas leves”) se $5 \leq \text{phq9} < 10$
- 2 (“sintomas moderados”) se $10 \leq \text{phq9} < 15$
- 3 (“sintomas moderadamente severos”) se $15 \leq \text{phq9} < 19$
- 4 (“sintomas severos”) se $\text{phq9} \geq 20$.



A vertical bar on the left side of the slide, transitioning from green at the top to blue at the bottom.

Faça uma análise descritiva **univariada**
de **todas** as variáveis da análise.

Para isso, use gráficos e tabelas.



Variáveis quantitativas (numéricas)

Exemplos:

Idade (anos)

Altura (m)

Valor gasto em compras (R\$)

Variáveis qualitativas (categóricas)

Exemplos:

Gênero

Nível de escolaridade

Faixa etária



Variáveis quantitativas (numéricas)

Medidas de posição e de dispersão:

Média

Mediana

Moda

Desvio-padrão

Percentis

Variáveis qualitativas (categóricas)

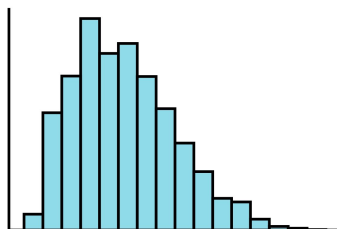
Frequências:

Absoluta (n)

Relativa (%)

Qual gráfico utilizar?

Distribuição de
variável numérica



Histograma

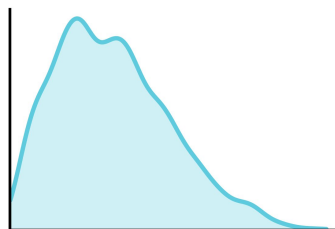


Gráfico de
densidade

Distribuição de
variável categórica

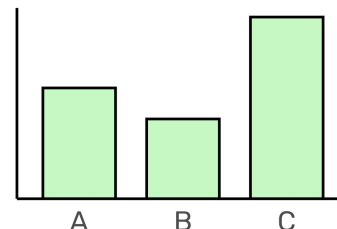



Gráfico de barras

A vertical bar on the left side of the slide with a gradient from light green at the top to light blue at the bottom.


Faça uma análise descritiva **univariada** de **todas** as variáveis da análise.

Para isso, use gráficos e tabelas.

Avalie a quantidade de **casos faltantes** nas variáveis e decida como proceder em relação a eles.

Reagrupe ou **recodifique** as variáveis que julgar necessário.



Faça uma análise bivariada de **sintomas de depressão** (phq9 ou phq_grp) com as **características demográficas**. Descreva o perfil com maiores prevalências de sintomas de depressão.

Faça uma análise bivariada de **características demográficas** x **hábitos saudáveis**. Qual perfil possui hábitos mais saudáveis?

Faça uma análise bivariada de **sintomas de depressão** (phq9 ou phq_grp) com os **hábitos saudáveis** (variáveis de atividade física e *Healthy Eating Index*). O que podemos observar?

Qual gráfico utilizar?

Relação entre **duas**
variáveis **numéricas**

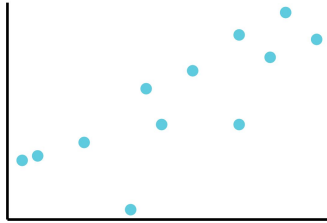


Gráfico de
dispersão

Relação entre **duas**
variáveis **categóricas**

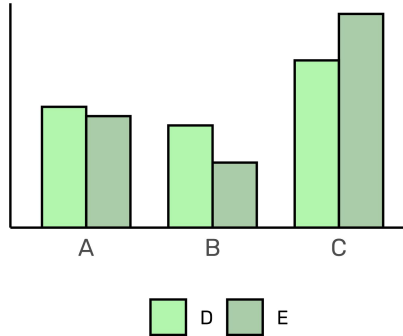


Gráfico de barras
agrupadas

Relação entre uma
variável **numérica** e
uma **categórica**

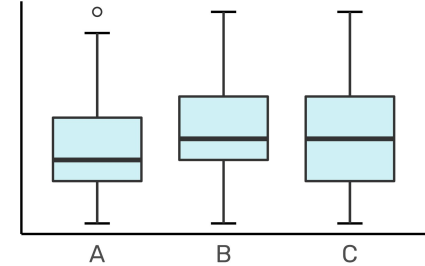
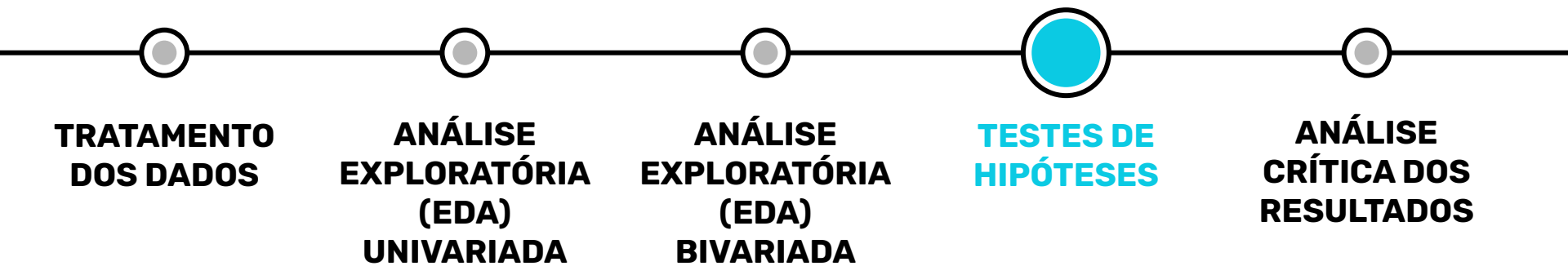



Gráfico boxplot



A vertical bar with a gradient from light green at the top to light blue at the bottom.

Faça o(s) teste(s) de hipóteses adequado(s) para avaliar a **significância estatística** das diferenças nas **características demográficas** apontadas na análise bivariada.

Quais são as características que apresentam diferenças estatisticamente significativas para a frequência de sintomas de depressão?

Faça o(s) teste(s) de hipóteses adequado(s) para avaliar se existe **associação** entre **hábitos saudáveis** e sintomas de depressão.

Algumas opções de testes de hipóteses:

Avaliar a relação entre
duas variáveis **numéricas**



Teste de correlação de Pearson

Avaliar a associação entre
duas variáveis **categóricas**



Teste qui-quadrado de independência

Comparar as **médias** de **dois**
grupos independentes



Teste-t independente

Comparar as **médias** de **mais**
de dois grupos independentes



Teste F (ANOVA)



**TRATAMENTO
DOS DADOS**



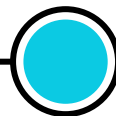
**ANÁLISE
EXPLORATÓRIA
(EDA)
UNIVARIADA**




**ANÁLISE
EXPLORATÓRIA
(EDA)
BIVARIADA**



**TESTES DE
HIPÓTESES**



**ANÁLISE
CRÍTICA DOS
RESULTADOS**

A vertical bar with a gradient from green at the top to blue at the bottom.

1. Qual **tipo de estudo** está sendo empregado pelo NHANES?
Experimental ou observacional?

2. Avalie as possíveis fontes de **viés** presentes na análise.

3. A partir da análise realizada, podemos afirmar que hábitos saudáveis possuem um **efeito causal** na prevenção de depressão?

4. Quais são as **limitações** das análises realizadas?
O que pode ser feito para **melhorar**?

5. Quais **outras variáveis/informações** poderiam ter sido coletadas para esta análise?

T



RESOLUÇÃO DO DESAFIO

A vertical bar on the left side of the slide, transitioning from green at the top to blue at the bottom.

LEITURA DO BANCO E TRATAMENTO DOS DADOS

Importação das bibliotecas

```
[ ] import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.api as sm
```

```
[ ] from google.colab import drive
drive.mount('/content/drive', force_remount = True)
```

Mounted at /content/drive

Carregamento dos bancos de dados

```
[ ] demo_phq = pd.read_csv('/content/drive/My Drive/Colab Notebooks/Desafio de inferência - Tera/DEMO_PHQ.csv')

pag_hei = pd.read_csv('/content/drive/My Drive/Colab Notebooks/Desafio de inferência - Tera/PAG_HEI.csv')
```

```
[5] demo_phq.info()  
# Informações de 5334 adultos
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 5334 entries, 0 to 5333  
Data columns (total 15 columns):
```

#	Column	Non-Null Count	Dtype
0	SEQN	5334 non-null	int64
1	DPQ010	4836 non-null	float64
2	DPQ020	4836 non-null	float64
3	DPQ030	4836 non-null	float64
4	DPQ040	4836 non-null	float64
5	DPQ050	4836 non-null	float64
6	DPQ060	4836 non-null	float64
7	DPQ070	4836 non-null	float64
8	DPQ080	4836 non-null	float64
9	DPQ090	4836 non-null	float64
10	RIAGENDR	5334 non-null	int64
11	RIDAGEYR	5334 non-null	int64
12	RIDRETH1	5334 non-null	int64
13	DMDEDUC	5334 non-null	int64
14	INDFMINC	5287 non-null	float64

```
dtypes: float64(10), int64(5)  
memory usage: 625.2 KB
```

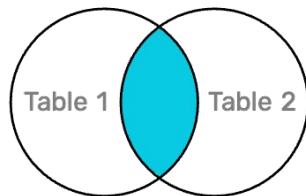
```
▶ pag_hei.info()  
# Informações de 9424 crianças e adultos
```

```
↳ <class 'pandas.core.frame.DataFrame'>  
RangeIndex: 9424 entries, 0 to 9423  
Data columns (total 17 columns):
```

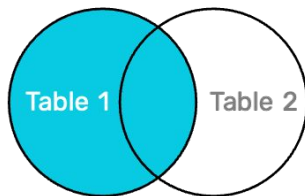
#	Column	Non-Null Count	Dtype
0	SEQN	9424 non-null	int64
1	PAG_MINW	7150 non-null	float64
2	ADHERENCE	7150 non-null	float64
3	HEI2015C1_TOTALVEG	8549 non-null	float64
4	HEI2015C2_GREEN_AND_BEAN	8549 non-null	float64
5	HEI2015C3_TOTALFRUIT	8549 non-null	float64
6	HEI2015C4_WHOLEFRUIT	8549 non-null	float64
7	HEI2015C5_WHOLEGRAIN	8549 non-null	float64
8	HEI2015C6_TOTALDAIRY	8549 non-null	float64
9	HEI2015C7_TOTPROT	8549 non-null	float64
10	HEI2015C8_SEAPLANT_PROT	8549 non-null	float64
11	HEI2015C9_FATTYACID	8549 non-null	float64
12	HEI2015C10_SODIUM	8549 non-null	float64
13	HEI2015C11_REFINEDGRAIN	8549 non-null	float64
14	HEI2015C12_SFAT	8549 non-null	float64
15	HEI2015C13_ADDSUG	8549 non-null	float64
16	HEI2015_TOTAL_SCORE	8549 non-null	float64

```
dtypes: float64(16), int64(1)  
memory usage: 1.2 MB
```

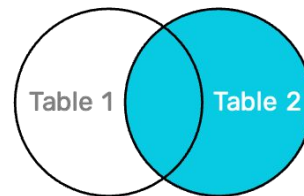
Inner Join



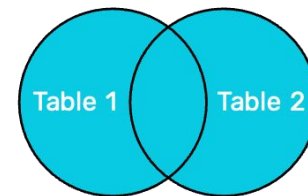
Left Join



Right Join



Full Outer Join



```
▶ df = demo_phq.merge(pag_hei, on = 'SEQN', how = 'left')
df.info()
# Resulta em um banco com 5334 pessoas, só adultos
```

```
☐ <class 'pandas.core.frame.DataFrame'>
Int64Index: 5334 entries, 0 to 5333
Data columns (total 31 columns):
#   Column          Non-Null Count  Dtype
---  -
0   SEQN            5334 non-null   int64
1   DPQ010          4836 non-null   float64
```

Tratamento das variáveis

Vamos substituir por valores ausentes os códigos que, de acordo com o dicionário, correspondiam a categorias ausentes.

Para a variável RIDRETH1, vamos agrupar as categorias 5 e 2, que correspondem a "outros".

Para a variável renda ("INDFMINC"), como há muitas categorias, vamos tratá-la como numérica, usando a média dos valores da categoria.

```
▶ replace_map = {  
    "DPQ010": {7: np.nan, 9: np.nan},  
    "DPQ020": {7: np.nan, 9: np.nan},  
    "DPQ030": {7: np.nan, 9: np.nan},  
    "DPQ040": {7: np.nan, 9: np.nan},  
    "DPQ050": {7: np.nan, 9: np.nan},  
    "DPQ060": {7: np.nan, 9: np.nan},  
    "DPQ070": {7: np.nan, 9: np.nan},  
    "DPQ080": {7: np.nan, 9: np.nan},  
    "DPQ090": {7: np.nan, 9: np.nan},  
    "RIDRETH1": {5: 2}, # Other  
    "DMDEDUC": {7: np.nan, 9: np.nan},  
    "INDFMINC": {1: np.mean([0, 4999]), 2: np.mean([5000, 9999]),  
                 3: np.mean([10000, 14999]), 4: np.mean([15000, 19999]),  
                 5: np.mean([20000, 24999]), 6: np.mean([25000, 34999]),  
                 7: np.mean([35000, 44999]), 8: np.mean([45000, 54999]),  
                 9: np.mean([55000, 64999]), 10: np.mean([65000, 74999]),  
                 11: 75000, 12: np.mean([20000, 90000]), 13: np.mean([0, 19999]),  
                 77: np.nan, 99: np.nan}  
}  
  
df2 = df.replace(replace_map)
```

Análise dos valores ausentes

```
df2.isna().sum().sort_values(ascending=False)
```

phq9	535
phq_grp2	535
phq_grp	535
DPQ080	507
DPQ010	507
DPQ030	505
DPQ020	503
DPQ040	503
DPQ060	503
DPQ050	502
DPQ070	502
DPQ090	500
HEI2015C2_GREEN_AND_BEAN	274
HEI2015C7_TOTPROT	274
HEI2015C8_SEAPLANT_PROT	274
HEI2015C11_REFINEDGRAIN	274
HEI2015C9_FATTYACID	274
HEI2015C10_SODIUM	274
HEI2015C5_WHOLEGRAIN	274
HEI2015C6_TOTALDAIRY	274
HEI2015C12_SFAT	274
HEI2015C4_WHOLEFRUIT	274
HEI2015C3_TOTALFRUIT	274
HEI2015C1_TOTALVEG	274
HEI2015_TOTAL_SCORE	274
HEI2015C13_ADDSUG	274
INDFMINC	160
DMDEDUC	8

```
100*df2.isna().sum().sort_values(ascending=False)/df2.shape[0]
```

phq9	10.029996
phq_grp2	10.029996
phq_grp	10.029996
DPQ080	9.505062
DPQ010	9.505062
DPQ030	9.467567
DPQ020	9.430071
DPQ040	9.430071
DPQ060	9.430071
DPQ050	9.411324
DPQ070	9.411324
DPQ090	9.373828
HEI2015C2_GREEN_AND_BEAN	5.136858
HEI2015C7_TOTPROT	5.136858
HEI2015C8_SEAPLANT_PROT	5.136858
HEI2015C11_REFINEDGRAIN	5.136858
HEI2015C9_FATTYACID	5.136858
HEI2015C10_SODIUM	5.136858
HEI2015C5_WHOLEGRAIN	5.136858
HEI2015C6_TOTALDAIRY	5.136858
HEI2015C12_SFAT	5.136858
HEI2015C4_WHOLEFRUIT	5.136858
HEI2015C3_TOTALFRUIT	5.136858
HEI2015C1_TOTALVEG	5.136858
HEI2015_TOTAL_SCORE	5.136858
HEI2015C13_ADDSUG	5.136858
INDFMINC	2.999625
DMDEDUC	0.149981
-----	- - - - -

Criação da variável escore PHQ

```
[13] df2["phq9"] = df2[["DPQ010",  
                      "DPQ020",  
                      "DPQ030",  
                      "DPQ040",  
                      "DPQ050",  
                      "DPQ060",  
                      "DPQ070",  
                      "DPQ080",  
                      "DPQ090"]].sum(axis = 'columns', skipna = False)
```

```
[14] df2['phq9'].describe()
```

```
count    4799.000000  
mean      2.732236  
std       3.727676  
min       0.000000  
25%      0.000000  
50%      1.000000  
75%      4.000000  
max      27.000000  
Name: phq9, dtype: float64
```

Criação da variável phq_grp

Vamos classificar os escores de acordo com a literatura

```
▶ conditions = [  
    (df2['phq9'].isna()),  
    (df2['phq9'] <= 5),  
    (df2['phq9'] > 5) & (df2['phq9'] <= 9),  
    (df2['phq9'] > 9) & (df2['phq9'] <= 14),  
    (df2['phq9'] > 14) & (df2['phq9'] <= 19),  
    (df2['phq9'] > 19)  
]  
values = [np.nan, 0, 1, 2, 3, 4]  
  
df2["phq_grp"] = np.select(conditions, values) # Construindo a variável  
  
df2[["phq_grp"]].value_counts(sort = False) # Avaliando as frequências
```

```
☞ phq_grp  
0.0      4013  
1.0       489  
2.0       201  
3.0        73  
4.0         23  
dtype: int64
```

Como há poucas pessoas nas categorias 2, 3 e 4, vamos agrupá-las:

```
[16] df2["phq_grp2"] = df2["phq_grp"].replace([3, 4], 2)
```

```
df2[["phq_grp2"]].value_counts(sort = False)
```

```
phq_grp2
0.0      4013
1.0       489
2.0       297
dtype: int64
```

- 1) Para as etapas de análise exploratória e teste de hipótese, utilizaremos a variável `phq_grp2`, com 3 níveis de sintomas de depressão.
- 2) Como o percentual de missing está abaixo de 10% para todas as variáveis, não faremos nenhum tratamento para os casos faltantes.

A vertical bar on the left side of the slide, transitioning from green at the top to blue at the bottom.

ANÁLISE EXPLORATÓRIA (EDA) UNIVARIADA

100

1

}

Variáveis numéricas

```
df2[var_quant].describe(percentiles = [.01, .25, .5, .75, .99]).round(2)
```



	RIDAGEYR	INDFMINC	PAG_MINW	HEI2015C1_TOTALVEG	HEI2015C2_GREEN_AND_BEAN
count	5334.00	5174.00	5334.00	5060.00	5060.00
mean	45.09	40787.68	471.77	3.20	1.81
std	20.15	24247.24	780.35	1.47	2.14
min	18.00	2499.50	0.00	0.00	0.00
1%	18.00	2499.50	0.00	0.00	0.00
25%	27.00	17499.50	35.00	2.06	0.00
50%	43.00	39999.50	210.00	3.22	0.06
75%	62.00	69999.50	568.93	4.83	4.50
99%	85.00	75000.00	3672.06	5.00	5.00
max	85.00	75000.00	10777.83	5.00	5.00

Dado que uma semana tem, no máximo, 160 horas (10.080 min) não é possível que alguém tenha feito 10777 min de exercícios semanais. Nesse caso, iremos truncar a variável PAG_MINW em 3600 min (60 horas semanais).

```
[93] df2['PAG_MINW_trunc'] = np.where(df2['PAG_MINW'] > 3600, 3600, df2['PAG_MINW'])
```

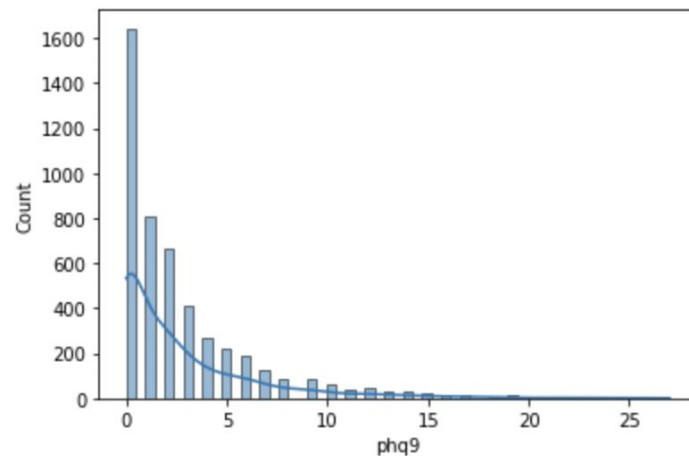
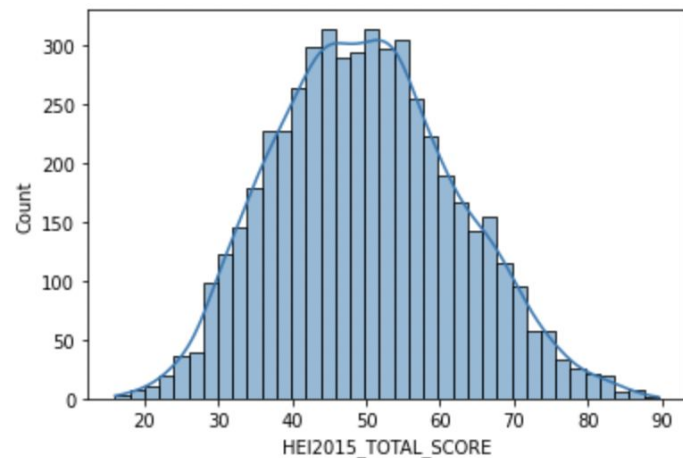
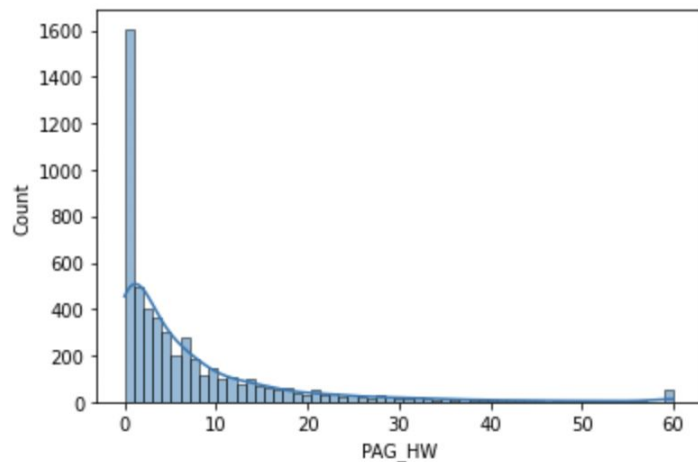
```
[94] df2[['PAG_MINW', 'PAG_MINW_trunc']].describe(percentiles = [.01, .25, .5, .75, .99]).round(2)
```

	PAG_MINW	PAG_MINW_trunc
count	5334.00	5334.00
mean	471.77	454.86
std	780.35	663.05
min	0.00	0.00
1%	0.00	0.00
25%	35.00	35.00
50%	210.00	210.00
75%	568.93	568.93
99%	3672.06	3600.00
max	10777.83	3600.00

```
[95] # Criando a variável PAG_MIN em horas  
df2['PAG_HW'] = df2['PAG_MINW_trunc']/60
```

Gráficos histograma

```
[97] for var in var_quant:  
      sns.histplot(df2[var], kde=True)  
      plt.show()
```

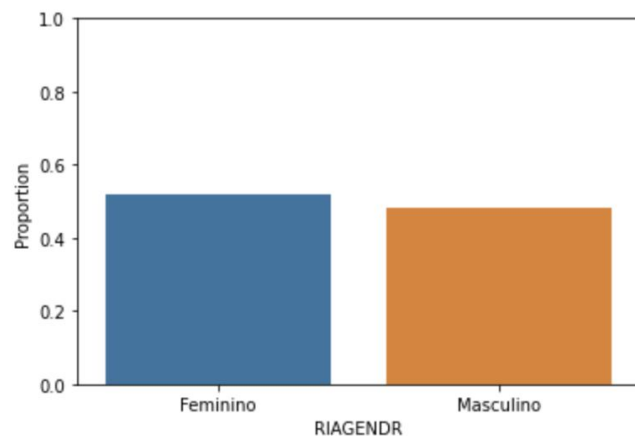


Gráficos de barras para as variáveis categóricas

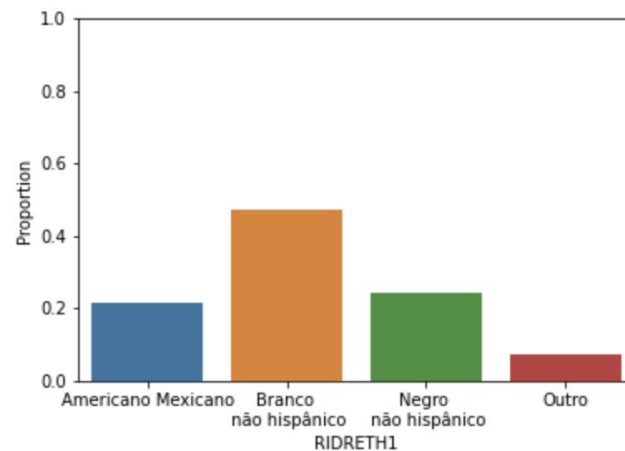
Definição de uma função para criar esses gráficos

```
[ ] def grafico_barras_prop(data, variable):  
    (data[[variable]]  
     .value_counts(normalize=True, sort = False)  
     .rename("Proportion")  
     .reset_index()  
     .pipe((sns.barplot, "data"), x=variable, y="Proportion"))  
    plt.ylim(0,1)  
    plt.show()
```

```
grafico_barras_prop(df2.replace(label_quali),  
                    variable = "RIAGENDR")
```



```
grafico_barras_prop(df2.replace(label_quali),  
                    variable = "RIDRETH1")
```



A vertical bar on the left side of the slide, transitioning from green at the top to blue at the bottom.

ANÁLISE EXPLORATÓRIA (EDA) BIVARIADA

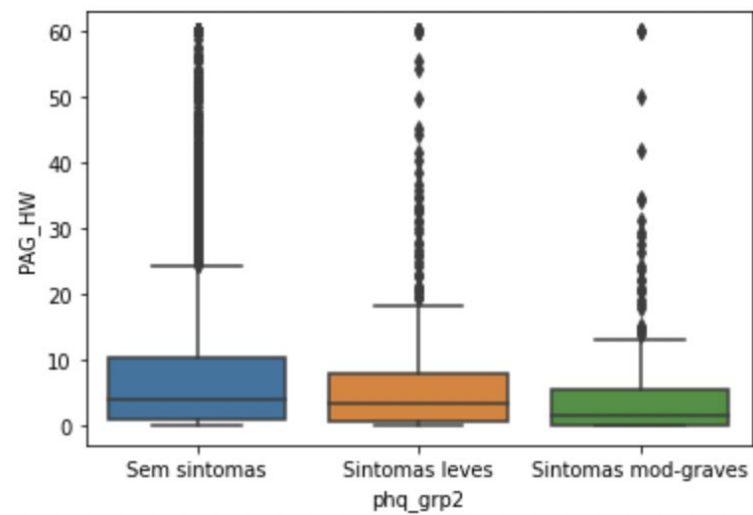
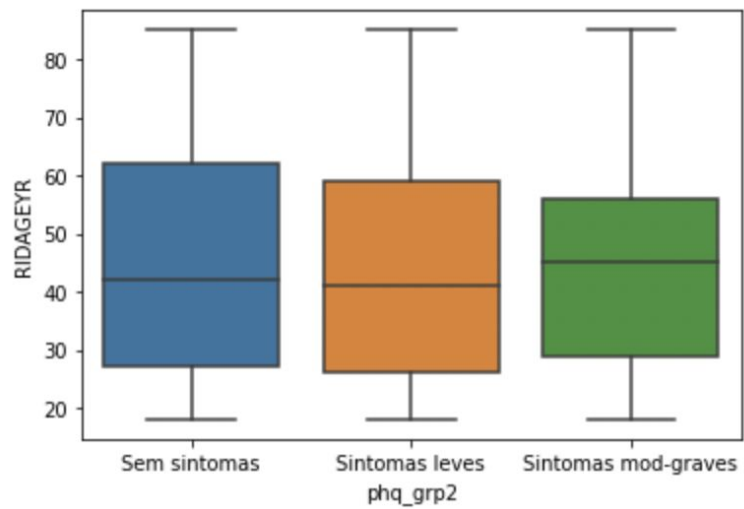
Gráficos boxplot das variáveis numéricas por grupo

Definição de uma função para criar esses gráficos

```
[109] def grafico_boxplot_grp(data, variable, label):  
  
    if label == "": label = variable  
    sns.boxplot(x="phq_grp2", y=variable, data=data.replace({'phq_grp2': {0: "Sem sintomas",  
                                                                           1: "Sintomas leves",  
                                                                           2: "Sintomas mod-graves"}}}))  
  
    plt.ylabel(label)  
    plt.show()
```

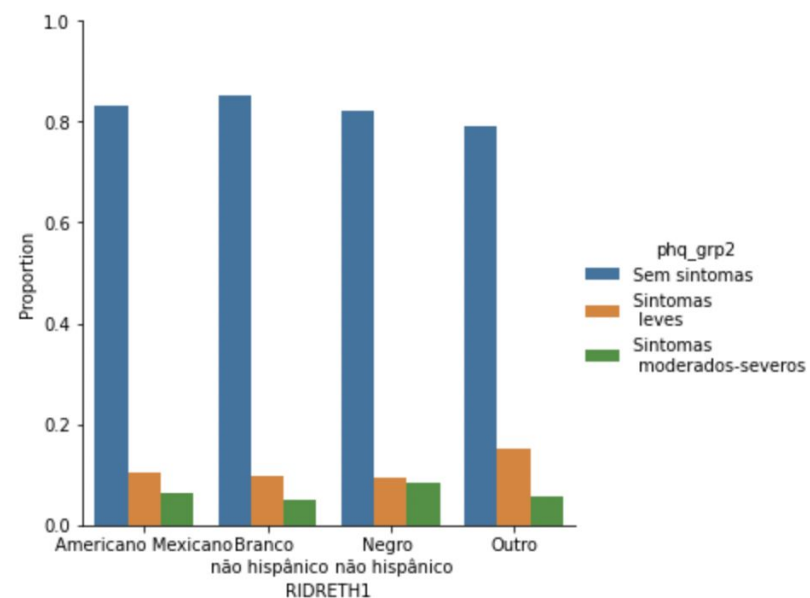
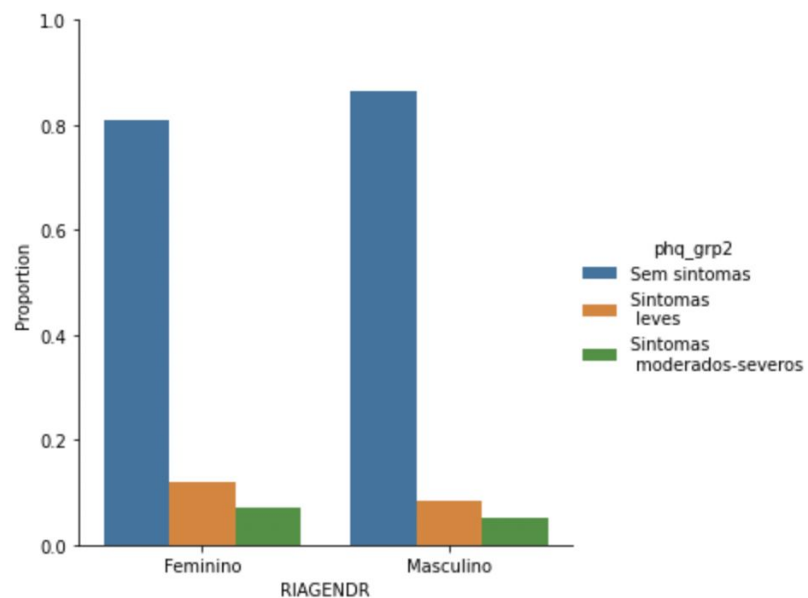


```
for var in var_quant:  
    grafico_boxplot_grp(df2, var, "")
```



```
def grafico_barras_prop_grp(data, variable):  
    (data  
     .groupby(variable)['phq_grp2']  
     .value_counts(normalize=True, sort = False)  
     .rename("Proportion")  
     .reset_index()  
     .pipe((sns.catplot, "data"), x=variable, y="Proportion", hue = 'phq_grp2', kind="bar"))  
    plt.ylim(0,1)  
    plt.show()
```

```
for var in var_quali:  
    grafico_barras_prop_grp(df2.replace(label_quali), var)
```



A vertical bar on the left side of the slide, transitioning from green at the top to blue at the bottom.

TESTES DE HIPÓTESES

A vertical bar on the left side of the slide, transitioning from green at the top to blue at the bottom.

IDADE E SINTOMAS DEPRESSIVOS

Os três **grupos** de depressão diferem quanto à **idade**?

Teste **ANOVA** (F): compara as médias de mais de dois grupos

Hipóteses do teste ANOVA:

- H0: **não há** diferença entre as médias dos grupos
- H1: **há** pelo menos uma diferença entre as médias dos grupos

Os três **grupos** de depressão diferem quanto à **idade**?

Teste **ANOVA** (F): compara as médias de mais de dois grupos

Hipóteses do teste ANOVA:

- H_0 : os três grupos de depressão **não diferem** quanto à média de idade
- H_1 : **há pelo menos uma diferença** na média de idade entre os três grupos de depressão

Os três **grupos** de depressão diferem quanto à **idade**?

```
from scipy.stats import f_oneway

df_aux = df2[["phq_grp2", "RIDAGEYR"]].dropna()

stat, p = f_oneway(df_aux[(df_aux.phq_grp2 == 0)][ "RIDAGEYR" ],
                  df_aux[(df_aux.phq_grp2 == 1)][ "RIDAGEYR" ],
                  df_aux[(df_aux.phq_grp2 == 2)][ "RIDAGEYR" ])

print('stat=%.3f, p=%.3f' % (stat, p))

stat=1.002, p=0.367
```

Como $p > 0,05 \rightarrow$ Não rejeitamos H_0

Conclusão: os três grupos de depressão não diferem estatisticamente quanto à média de idade

A vertical bar on the left side of the slide, transitioning from green at the top to blue at the bottom.

Exercícios físicos e sintomas depressivos

Os três **grupos** de depressão diferem quanto ao tempo dedicado a **exercício físico**, por semana?

Teste **ANOVA** (F): compara as médias de mais de dois grupos

Hipóteses do teste ANOVA:

- H0: os três grupos de depressão **não diferem** quanto à média de tempo gasto com exercício físico
- H1: **há pelo menos uma diferença** na média de tempo gasto com exercício físico entre os três grupos de depressão

Os três **grupos** de depressão diferem quanto ao tempo dedicado a **exercício físico**, por semana?

```
from scipy.stats import f_oneway

df_aux = df2[["phq_grp2", "PAG_HW"]].dropna()

stat, p = f_oneway(df_aux[(df_aux.phq_grp2 == 0)][ "PAG_HW" ],
                  df_aux[(df_aux.phq_grp2 == 1)][ "PAG_HW" ],
                  df_aux[(df_aux.phq_grp2 == 2)][ "PAG_HW" ])

print('stat=%.3f, p=%.3f' % (stat, p))

stat=12.652, p=0.000
```

Como $p < 0,05 \rightarrow$ Rejeitamos H_0

Conclusão: há pelo menos uma diferença na média de tempo gasto com exercício físico entre os três grupos de depressão

Quais grupos diferem entre si?

Para responder a essa pergunta, faremos o teste post-hoc de Tukey

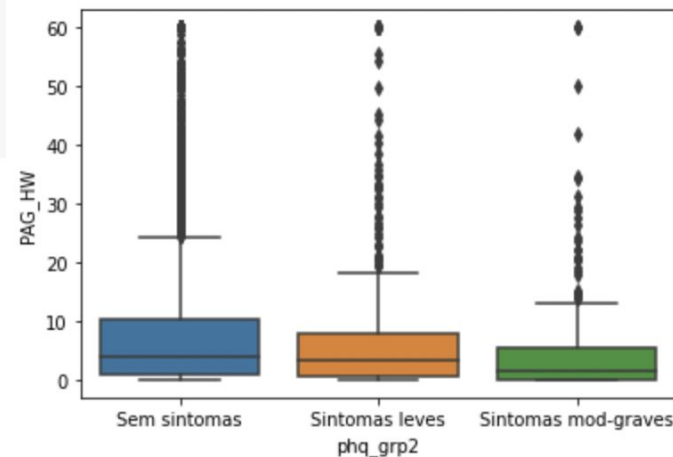
```
from statsmodels.stats.multicomp import pairwise_tukeyhsd

tukey = pairwise_tukeyhsd(df_aux["PAG_HW"],
                           df_aux['phq_grp2'],
                           alpha = 0.05)

print(tukey)
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0.0	1.0	-0.6978	0.3919	-1.948	0.5524	False
0.0	2.0	-3.3156	0.001	-4.8852	-1.7459	True
1.0	2.0	-2.6178	0.004	-4.538	-0.6975	True



A vertical bar on the left side of the slide, transitioning from green at the top to blue at the bottom.

GÊNERO E SINTOMAS DEPRESSIVOS

A proporção de pessoas em cada **grupo** de depressão varia de acordo com o **gênero**?

Teste **qui-quadrado**: avalia a associação entre duas variáveis categóricas

Hipóteses do teste qui-quadrado:

- H0: **não há** associação entre as duas variáveis
- H1: **há** associação entre as duas variáveis

A proporção de pessoas em cada **grupo** de depressão varia de acordo com o **gênero**?

Teste **qui-quadrado**: avalia a associação entre duas variáveis categóricas

Hipóteses do teste qui-quadrado:

- H0: **não há** associação entre gênero e a presença de sintomas depressivos
- H1: **há** associação entre gênero e a presença de sintomas depressivos

A proporção de pessoas em cada **grupo** de depressão varia de acordo com o **gênero**?

```
from scipy.stats import chi2_contingency

crosstab = pd.crosstab(df2.replace(label_quali)['RIAGENDR'],
                       df2.replace(label_quali)['phq_grp2'])

stat, p, df, expected = chi2_contingency(crosstab)
print('stat = %.3f, p = %.3f' % (stat, p))

stat = 28.811, p = 0.000
```

Como $p < 0,05 \rightarrow$ Rejeitamos H_0

Conclusão: há associação entre gênero e a presença de sintomas depressivos

Como se dá essa diferença?

```
100*pd.crosstab(df2.replace(label_quali)[ 'RIAGENDR' ],  
                df2.replace(label_quali)[ 'phq_grp2' ],  
                normalize='index')
```

phq_grp2 Sem sintomas Sintomas \n leves Sintomas \n moderados-severos

RIAGENDR


Feminino	80.860474	11.982308	7.157218
Masculino	86.591696	8.261246	5.147059

T

A vertical bar with a gradient from light green at the top to light blue at the bottom.


RESUMO DA AULA

TAKEAWAY #1

A vertical bar with a gradient from light green at the top to light blue at the bottom.

**Não há uma única
solução. Há várias
respostas e caminhos
possíveis**

TAKEAWAY #2



Há associação entre os sintomas depressivos, gênero e realização de exercícios físicos

Não podemos fazer inferências causais - estudo observacional.

TAKEAWAY #3

A vertical bar with a gradient from green at the top to blue at the bottom.

**A ideia é que tenha sido
desafiador!**

TAKEAWAY #4

A vertical bar with a gradient from light blue at the top to dark blue at the bottom.

**O desafio é diferente
para cada um.**

Estamos em níveis diferentes e tudo bem.

