

# **Avaliação de microdados do Enem 2021: as contribuições dos dados para compreensão do desempenho de estudantes**

**Grupo 1:**  
João Estevam, Leandro Lima, Thiago Brito,  
Vivian Gomes e Willian Wallace


# Motivações deste estudo

- Contexto educacional brasileiro, tendo em vista acesso às universidades públicas mediante o Exame Nacional do Ensino Médio (Enem).
- Ausência de políticas públicas que assegurem um ensino igualitário entre escolas públicas e privadas;
- Diferente construção social de estudantes que constituíram o Exame Nacional do Ensino Médio (Enem) de 2021 em comparação aos anos anteriores com menores índices de inclusão (PALHARES, 2021).



# Objetivo geral

Compreender quais fatores podem colaborar para o desempenho dos participantes a partir da extração, tratamento e análise dos microdados do Enem 2021.



# Hipóteses

**Desempenho  
de participantes**

**1<sup>a</sup>**

Participantes provenientes de escolas particulares possuem melhores desempenhos no Enem do que participantes provenientes de escolas públicas.

**2<sup>a</sup>**

Os participantes cujos pais possuem maior nível de escolaridade têm melhores desempenhos no Enem.

**3<sup>a</sup>**

Participantes cuja a raça declarada é branca possuem melhores desempenhos no Enem do que participantes que se autodeclararam de outras raças.

# Descrição dos dados e abordagem

## Notas

Ciências Humanas;  
Linguagens, Códigos e  
Tecnologias;  
Matemática e  
Tecnologias; Redação;  
Ciências da Natureza.

## Escolas

Públicas;  
Privadas.

## Raça

Não  
declarado;  
branca; preta;  
parda, etc.

## Procedimentos

Limpeza; Contagem;  
Média; Desvio padrão,  
entre outros.

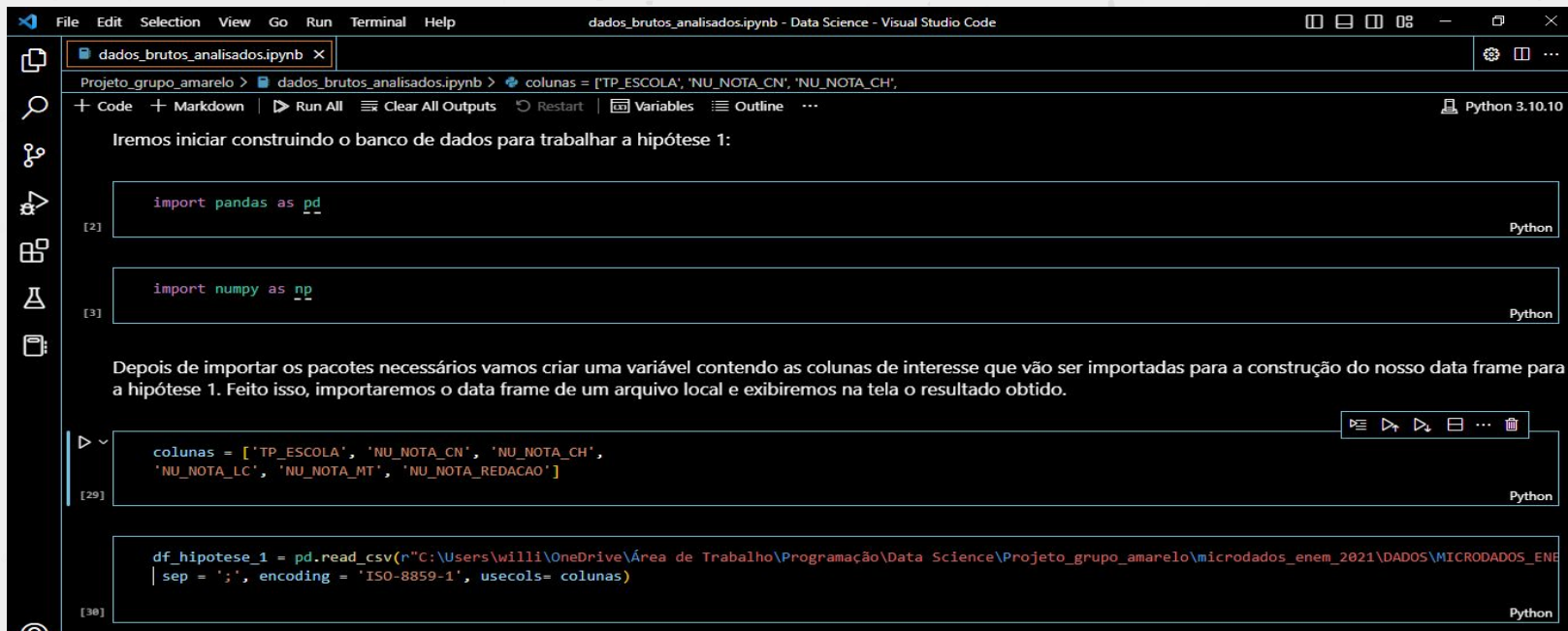
## Dados

Instituto Nacional de  
Estudos e Pesquisas  
Educacionais (Inep)



## Análises

# Resultados (hipótese 1)



The screenshot shows a Jupyter Notebook titled 'dados\_brutos\_analisados.ipynb' in the Visual Studio Code editor. The notebook is open to a cell containing Python code. The code imports 'pandas' as 'pd' and 'numpy' as 'np'. It then defines a list of columns: 'TP\_ESCOLA', 'NU\_NOTA\_CN', 'NU\_NOTA\_CH', 'NU\_NOTA\_LC', 'NU\_NOTA\_MT', and 'NU\_NOTA\_REDACAO'. Finally, it reads a CSV file from a local path and creates a DataFrame named 'df\_hipotese\_1'.

```
import pandas as pd

import numpy as np

colunas = ['TP_ESCOLA', 'NU_NOTA_CN', 'NU_NOTA_CH',
           'NU_NOTA_LC', 'NU_NOTA_MT', 'NU_NOTA_REDACAO']

df_hipotese_1 = pd.read_csv(r"C:\Users\willi\OneDrive\Área de Trabalho\Programação\Data Science\Projeto_grupo_amarelo\microdados_enem_2021\DADOS\MICRODADOS_ENE",
                           sep = ';', encoding = 'ISO-8859-1', usecols= colunas)
```

The notebook interface includes a menu bar (File, Edit, Selection, View, Go, Run, Terminal, Help), a toolbar with icons for file operations, and a status bar at the bottom indicating the Python version (3.10.10). The left sidebar shows the Explorer, Search, and Run and Debug views.

# Resultados (hipótese 1)



Visual Studio Code interface showing a Jupyter Notebook with the following code and output:

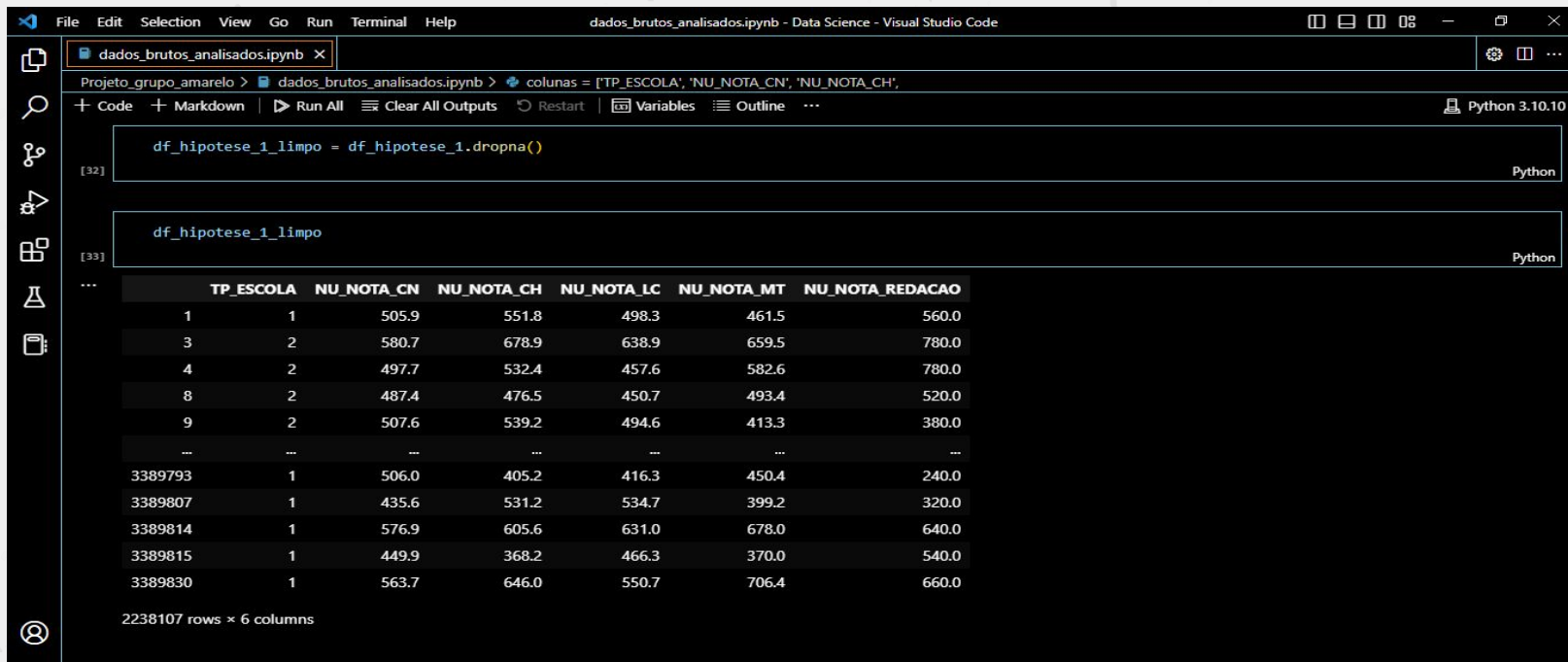
```
dados_brutos_analisados.ipynb > colunas = ['TP_ESCOLA', 'NU_NOTA_CN', 'NU_NOTA_CH',  
[31] df_hipotese_1
```

	TP_ESCOLA	NU_NOTA_CN	NU_NOTA_CH	NU_NOTA_LC	NU_NOTA_MT	NU_NOTA_REDACAO
0	1	NaN	574.6	472.6	NaN	760.0
1	1	505.9	551.8	498.3	461.5	560.0
2	1	NaN	NaN	NaN	NaN	NaN
3	2	580.7	678.9	638.9	659.5	780.0
4	2	497.7	532.4	457.6	582.6	780.0
...	...	...	...	...	...	...
3389827	1	NaN	NaN	NaN	NaN	NaN
3389828	1	NaN	NaN	NaN	NaN	NaN
3389829	1	NaN	NaN	NaN	NaN	NaN
3389830	1	563.7	646.0	550.7	706.4	660.0
3389831	1	NaN	NaN	NaN	NaN	NaN

3389832 rows x 6 columns

Feito isto, é possível notar que existem mais de 3 milhões de linhas em cada uma das colunas que foram importadas. Também é possível notar que muitos valores estão preenchidos com NaN de modo que dificulta a nossa análise. Para lidar com este problema resolvemos eliminar todas as linhas que possuem os valores NaN.

# Resultados (hipótese 1)



The screenshot shows a Jupyter Notebook interface in Visual Studio Code. The notebook is titled "dados\_brutos\_analisados.ipynb" and is running on Python 3.10.10. The code cell [32] shows the command `df_hipotese_1_limpo = df_hipotese_1.dropna()`. The output cell [33] displays the resulting data frame, which has 2238107 rows and 6 columns. The columns are labeled `TP_ESCOLA`, `NU_NOTA_CN`, `NU_NOTA_CH`, `NU_NOTA_LC`, `NU_NOTA_MT`, and `NU_NOTA_REDACAO`. The data is presented in a table format with a header row and several data rows.

```
[32] df_hipotese_1_limpo = df_hipotese_1.dropna()
```

```
[33] df_hipotese_1_limpo
```

	TP_ESCOLA	NU_NOTA_CN	NU_NOTA_CH	NU_NOTA_LC	NU_NOTA_MT	NU_NOTA_REDACAO
1	1	505.9	551.8	498.3	461.5	560.0
3	2	580.7	678.9	638.9	659.5	780.0
4	2	497.7	532.4	457.6	582.6	780.0
8	2	487.4	476.5	450.7	493.4	520.0
9	2	507.6	539.2	494.6	413.3	380.0
...	...	...	...	...	...	...
3389793	1	506.0	405.2	416.3	450.4	240.0
3389807	1	435.6	531.2	534.7	399.2	320.0
3389814	1	576.9	605.6	631.0	678.0	640.0
3389815	1	449.9	368.2	466.3	370.0	540.0
3389830	1	563.7	646.0	550.7	706.4	660.0

2238107 rows x 6 columns



# Resultados (hipótese 1)



The screenshot shows a Jupyter Notebook interface with the following content:

```
dados_brutos_analisados.ipynb x
```

Projeto\_grupo\_amarelo > dados\_brutos\_analisados.ipynb > `colunas = ['TP_ESCOLA', 'NU_NOTA_CN', 'NU_NOTA_CH',`

+ Code + Markdown | ▶ Run All | Clear All Outputs | Restart | Variables | Outline | ... Python 3.10.10

```
[14] dicionario_tipo_escola = {1: 'Não respondeu', 2: 'Pública', 3: 'Privada'}
```

Python

Com a limpeza o número de linhas foi reduzido de 3 milhões e 400 mil, aproximadamente, para 2 milhões e 200 mil, aproximadamente. Em seguida iremos calcular quantos participantes existem em cada tipo de escola.

```
[40] print(df_hipoteses_1_limpo.groupby(df_hipoteses_1_limpo['TP_ESCOLA']).map(dicionario_tipo_escola).count().sort_values(by='NU_NOTA_MT', ascending=False))
```

Python

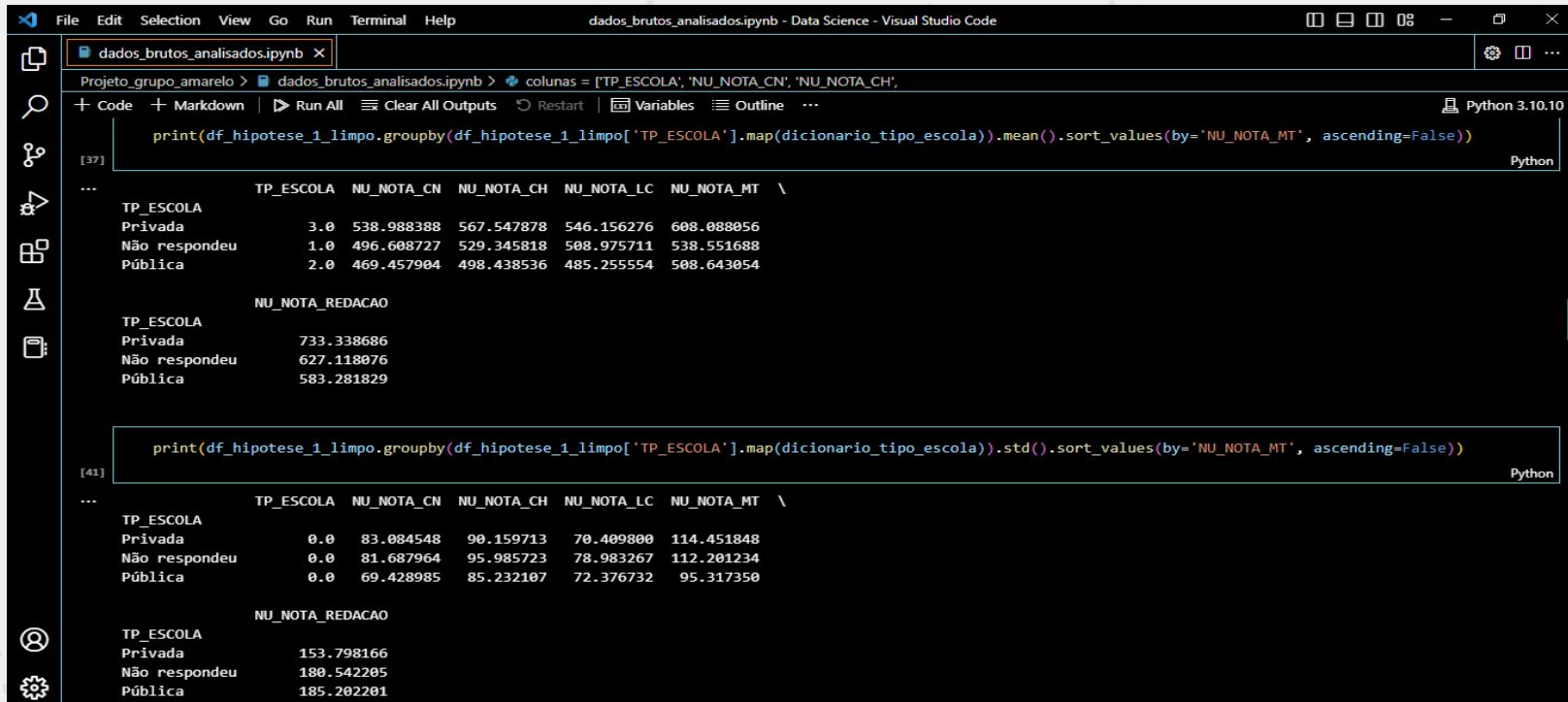
```
...
TP_ESCOLA
Não respondeu 1390710
Pública      668036
Privada      179361

NU_NOTA_REDACAO
TP_ESCOLA
Não respondeu 1390710
Pública      668036
Privada      179361
```

TP_ESCOLA	NU_NOTA_CN	NU_NOTA_CH	NU_NOTA_LC	NU_NOTA_MT
Não respondeu	1390710	1390710	1390710	1390710
Pública	668036	668036	668036	668036
Privada	179361	179361	179361	179361

TP_ESCOLA	NU_NOTA_REDACAO
Não respondeu	1390710
Pública	668036
Privada	179361

# Resultados (hipótese 1)



The screenshot shows a Jupyter Notebook interface with two code cells. The first cell (index 37) calculates the mean of 'NU\_NOTA\_MT' for each school type. The second cell (index 41) calculates the standard deviation of 'NU\_NOTA\_MT' for each school type. Both outputs are displayed as tables.

**Code Cell 37:**

```
print(df_hipotese_1_limpo.groupby(df_hipotese_1_limpo['TP_ESCOLA']).map(dicionario_tipo_escola)).mean().sort_values(by='NU_NOTA_MT', ascending=False))
```

**Output 37:**

TP_ESCOLA	NU_NOTA_CN	NU_NOTA_CH	NU_NOTA_LC	NU_NOTA_MT
Privada	538.988388	567.547878	546.156276	608.088056
Não respondeu	496.608727	529.345818	508.975711	538.551688
Pública	469.457904	498.438536	485.255554	508.643054

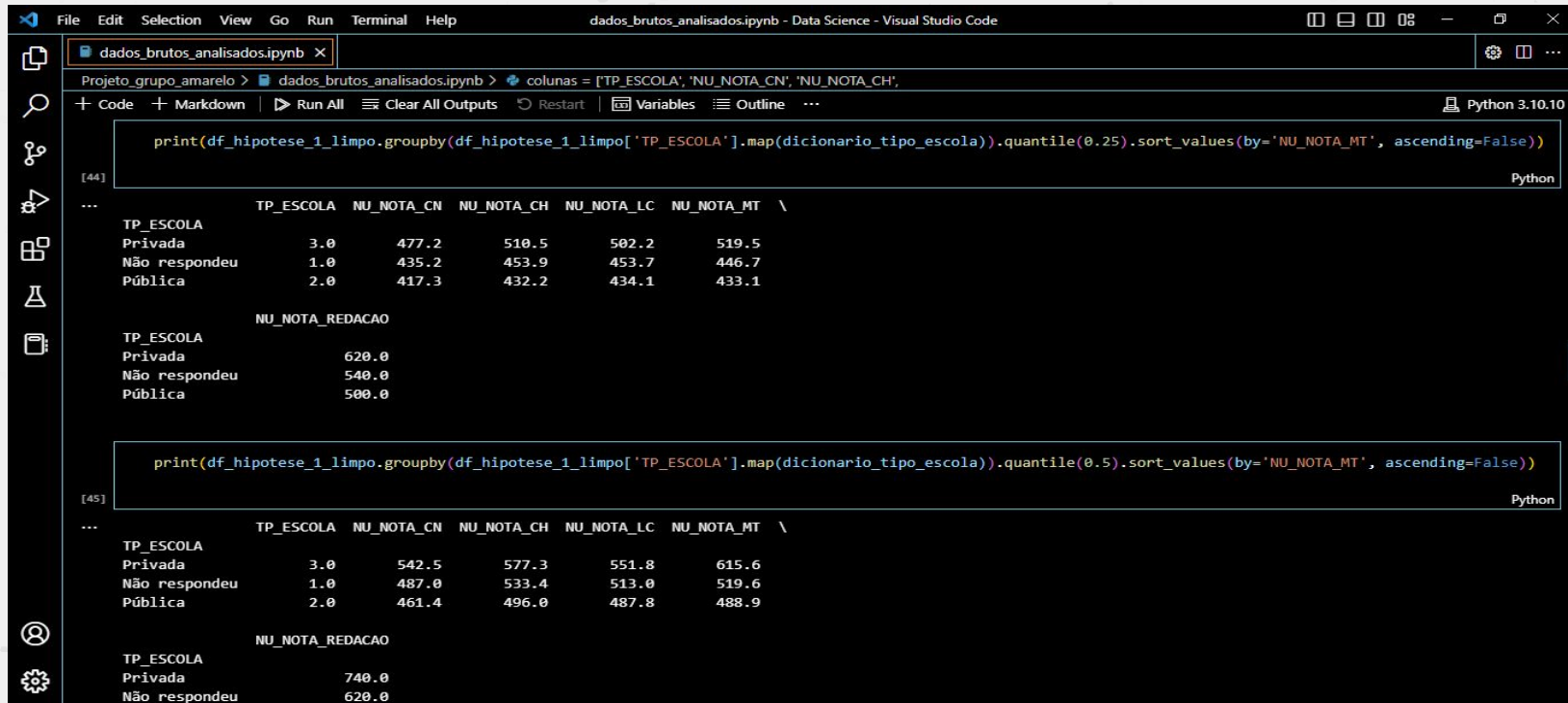
**Code Cell 41:**

```
print(df_hipotese_1_limpo.groupby(df_hipotese_1_limpo['TP_ESCOLA']).map(dicionario_tipo_escola)).std().sort_values(by='NU_NOTA_MT', ascending=False))
```

**Output 41:**

TP_ESCOLA	NU_NOTA_CN	NU_NOTA_CH	NU_NOTA_LC	NU_NOTA_MT
Privada	83.084548	90.150713	70.409800	114.451848
Não respondeu	81.687964	95.985723	78.983267	112.201234
Pública	69.428985	85.232107	72.376732	95.317350

# Resultados (hipótese 1)



Visual Studio Code interface showing a Jupyter Notebook with two code cells and their outputs.

File: dados\_brutos\_analisados.ipynb - Data Science - Visual Studio Code

Code cell [44]:

```
print(df_hipotese_1_limpo.groupby(df_hipotese_1_limpo['TP_ESCOLA'].map(dicionario_tipo_escola)).quantile(0.25).sort_values(by='NU_NOTA_MT', ascending=False))
```

Output [44]:

TP_ESCOLA	TP_ESCOLA	NU_NOTA_CN	NU_NOTA_CH	NU_NOTA_LC	NU_NOTA_MT
Privada	3.0	477.2	510.5	502.2	519.5
Não respondeu	1.0	435.2	453.9	453.7	446.7
Pública	2.0	417.3	432.2	434.1	433.1

NU\_NOTA\_REDACAO

TP_ESCOLA	NU_NOTA_REDACAO
Privada	620.0
Não respondeu	540.0
Pública	500.0

Code cell [45]:

```
print(df_hipotese_1_limpo.groupby(df_hipotese_1_limpo['TP_ESCOLA'].map(dicionario_tipo_escola)).quantile(0.5).sort_values(by='NU_NOTA_MT', ascending=False))
```

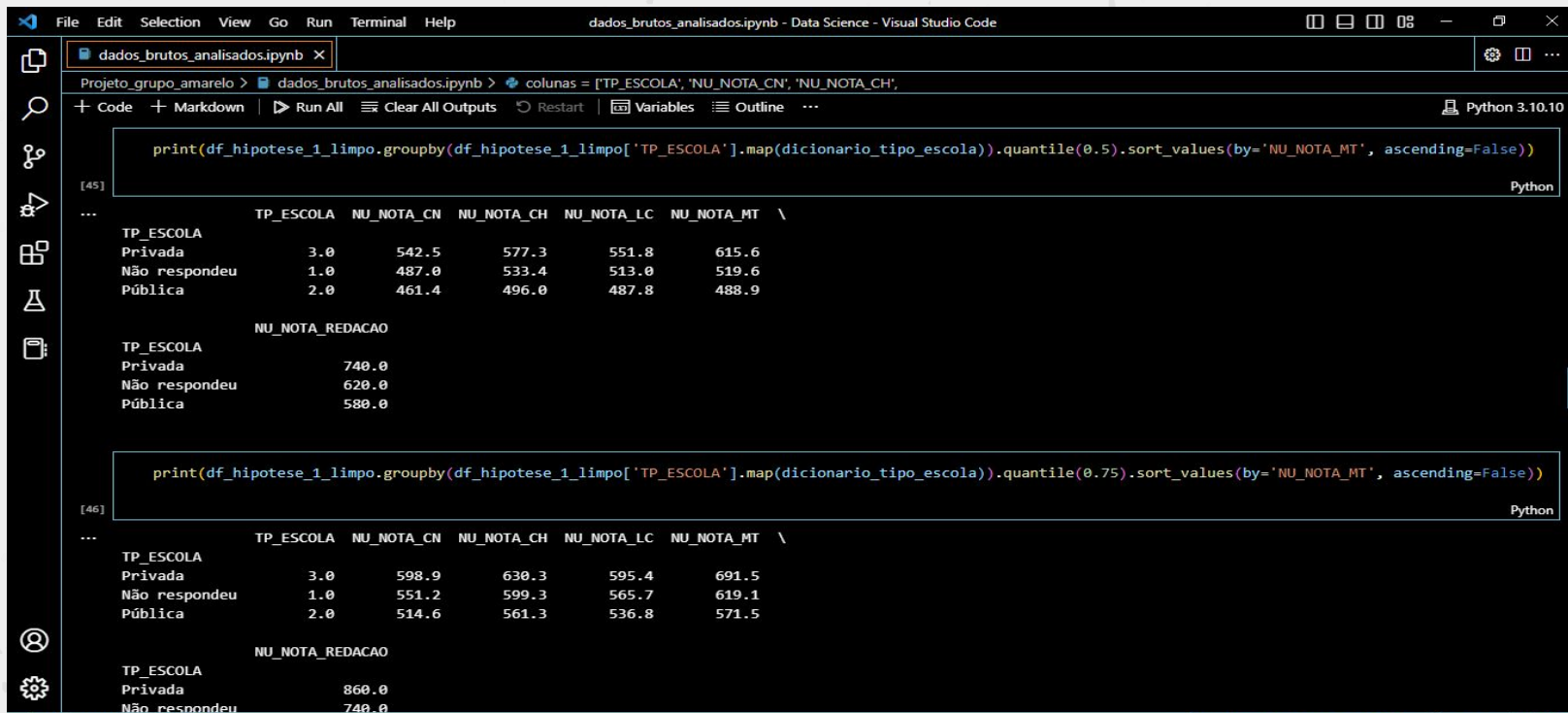
Output [45]:

TP_ESCOLA	TP_ESCOLA	NU_NOTA_CN	NU_NOTA_CH	NU_NOTA_LC	NU_NOTA_MT
Privada	3.0	542.5	577.3	551.8	615.6
Não respondeu	1.0	487.0	533.4	513.0	519.6
Pública	2.0	461.4	496.0	487.8	488.9

NU\_NOTA\_REDACAO

TP_ESCOLA	NU_NOTA_REDACAO
Privada	740.0
Não respondeu	620.0

# Resultados (hipótese 1)



The screenshot shows a Jupyter Notebook titled 'dados\_brutos\_analisados.ipynb' in Visual Studio Code. The notebook contains two code cells, each followed by its output.

**Code Cell [45]:**

```
print(df_hipotese_1_limpo.groupby(df_hipotese_1_limpo['TP_ESCOLA']).map(dicionario_tipo_escola).quantile(0.5).sort_values(by='NU_NOTA_MT', ascending=False))
```

**Output [45]:**

TP_ESCOLA	NU_NOTA_CN	NU_NOTA_CH	NU_NOTA_LC	NU_NOTA_MT
Privada	3.0	542.5	577.3	551.8
Não respondeu	1.0	487.0	533.4	513.0
Pública	2.0	461.4	496.0	487.8

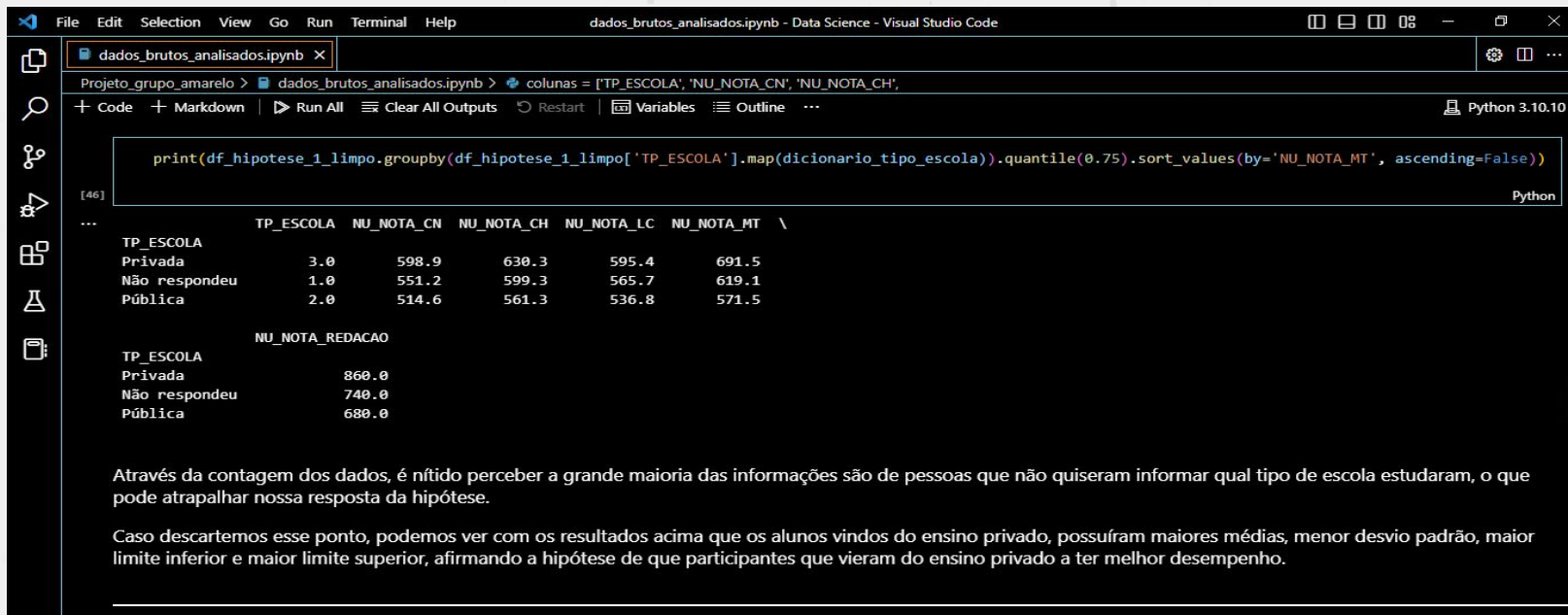
**Code Cell [46]:**

```
print(df_hipotese_1_limpo.groupby(df_hipotese_1_limpo['TP_ESCOLA']).map(dicionario_tipo_escola).quantile(0.75).sort_values(by='NU_NOTA_MT', ascending=False))
```

**Output [46]:**

TP_ESCOLA	NU_NOTA_CN	NU_NOTA_CH	NU_NOTA_LC	NU_NOTA_MT
Privada	3.0	598.9	630.3	595.4
Não respondeu	1.0	551.2	599.3	565.7
Pública	2.0	514.6	561.3	536.8

# Resultados (hipótese 1)



The screenshot shows a Jupyter Notebook interface in Visual Studio Code. The notebook file is named `dados_brutos_analisados.ipynb`. The code cell [46] contains the following Python code:

```
print(df_hipotese_1_limpo.groupby(df_hipotese_1_limpo['TP_ESCOLA']).map(dicionario_tipo_escola).quantile(0.75).sort_values(by='NU_NOTA_MT', ascending=False))
```

The output of the code is a table with 6 columns: `TP_ESCOLA`, `NU_NOTA_CN`, `NU_NOTA_CH`, `NU_NOTA_LC`, `NU_NOTA_MT`, and `\`. The rows represent different school types: `Privada`, `Não respondeu`, and `Pública`.

TP_ESCOLA	NU_NOTA_CN	NU_NOTA_CH	NU_NOTA_LC	NU_NOTA_MT	\
Privada	3.0	598.9	630.3	595.4	691.5
Não respondeu	1.0	551.2	599.3	565.7	619.1
Pública	2.0	514.6	561.3	536.8	571.5

Below the table, there is a section titled `NU_NOTA_REDACAO` with the following data:

TP_ESCOLA	NU_NOTA_REDACAO
Privada	860.0
Não respondeu	740.0
Pública	680.0

Através da contagem dos dados, é nítido perceber a grande maioria das informações são de pessoas que não quiseram informar qual tipo de escola estudaram, o que pode atrapalhar nossa resposta da hipótese.

Caso descartemos esse ponto, podemos ver com os resultados acima que os alunos vindos do ensino privado, possuíram maiores médias, menor desvio padrão, maior limite inferior e maior limite superior, afirmando a hipótese de que participantes que vieram do ensino privado a ter melhor desempenho.

# Resultados (hipótese 1)

- Daqueles participantes válidos, a maioria é de escola pública e a minoria de escola privada;
- Procedimentos utilizados: contagem de participantes; agrupamento de estudantes por escolas; média de notas; desvio padrão; quantis (25%, 50% e 75%);
- Conclui-se que participantes vindos de escola privada possuem notas médias maiores que participantes vindos de escolas públicas.

# Resultados (hipótese 2)

Visual Studio Code interface showing a Jupyter Notebook with a Python script and its output.

**Code Cell [48]:**

```
df_hipotese_2 = pd.read_csv(r"C:\Users\willi\OneDrive\Área de Trabalho\Programação\Data Science\Projeto_grupo_amarelo\microdados_enem_2021\Dados\MICRODADOS_ENE  
| sep = ';', encoding = 'ISO-8859-1', usecols= colunas_hipotese_2)
```

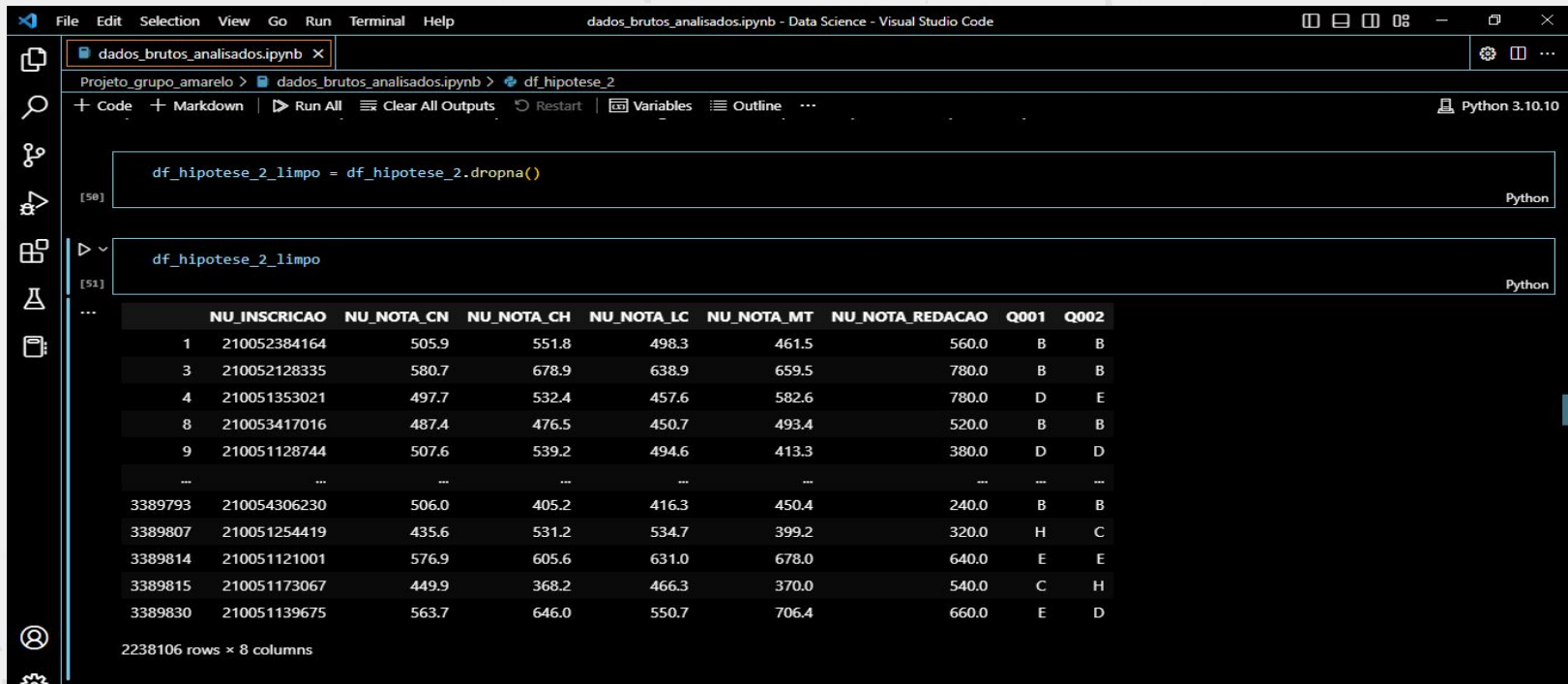
**Output Cell [49]:**

df\_hipotese\_2

	NU_INSCRICAO	NU_NOTA_CN	NU_NOTA_CH	NU_NOTA_LC	NU_NOTA_MT	NU_NOTA_REDACAO	Q001	Q002
0	210053865474	NaN	574.6	472.6	NaN	760.0	F	F
1	210052384164	505.9	551.8	498.3	461.5	560.0	B	B
2	210052589243	NaN	NaN	NaN	NaN	NaN	B	C
3	210052128335	580.7	678.9	638.9	659.5	780.0	B	B
4	210051353021	497.7	532.4	457.6	582.6	780.0	D	E
...	...	...	...	...	...	...	...	...
3389827	210053249138	NaN	NaN	NaN	NaN	NaN	B	B
3389828	210053776013	NaN	NaN	NaN	NaN	NaN	E	E
3389829	210052441508	NaN	NaN	NaN	NaN	NaN	B	C
3389830	210051139675	563.7	646.0	550.7	706.4	660.0	E	D
3389831	210052410399	NaN	NaN	NaN	NaN	NaN	NaN	NaN

3389832 rows x 8 columns

# Resultados (hipótese 2)



The screenshot shows a Jupyter Notebook interface in Visual Studio Code. The notebook is titled "dados\_brutos\_analisados.ipynb" and is open to the "df\_hipotese\_2" cell. The code in the cell is:

```
df_hipotese_2_limpo = df_hipotese_2.dropna()
```

The output of the cell is a preview of the DataFrame "df\_hipotese\_2\_limpo". The preview shows the first 10 rows of the DataFrame, with columns: NU\_INSCRICAO, NU\_NOTA\_CN, NU\_NOTA\_CH, NU\_NOTA\_LC, NU\_NOTA\_MT, NU\_NOTA\_REDACAO, Q001, and Q002. The DataFrame has 2238106 rows and 8 columns.

	NU_INSCRICAO	NU_NOTA_CN	NU_NOTA_CH	NU_NOTA_LC	NU_NOTA_MT	NU_NOTA_REDACAO	Q001	Q002
1	210052384164	505.9	551.8	498.3	461.5	560.0	B	B
3	210052128335	580.7	678.9	638.9	659.5	780.0	B	B
4	210051353021	497.7	532.4	457.6	582.6	780.0	D	E
8	210053417016	487.4	476.5	450.7	493.4	520.0	B	B
9	210051128744	507.6	539.2	494.6	413.3	380.0	D	D
...	...	...	...	...	...	...	...	...
3389793	210054306230	506.0	405.2	416.3	450.4	240.0	B	B
3389807	210051254419	435.6	531.2	534.7	399.2	320.0	H	C
3389814	210051121001	576.9	605.6	631.0	678.0	640.0	E	E
3389815	210051173067	449.9	368.2	466.3	370.0	540.0	C	H
3389830	210051139675	563.7	646.0	550.7	706.4	660.0	E	D

2238106 rows × 8 columns



# Resultados (hipótese 2)

dados_brutos_analisados.ipynb - Data Science - Visual Studio Code											
dados_brutos_analisados.ipynb											
Projeto_grupo_amarelo > dados_brutos_analisados.ipynb > df_hipotese_2_limpo["DESCRICAO_Q001"] = [dicionario_resposta_Q001_Q002[item] for item in df_hipotese_2_limpo.Q001]											
	NU_INSCRICAO	NU_NOTA_CN	NU_NOTA_CH	NU_NOTA_LC	NU_NOTA_MT	NU_NOTA_REDACAO	Q001	Q002	DESCRICAO_Q001	DESCRICAO_Q002	
1	210052384164	505.9	551.8	498.3	461.5	560.0	B	B	Não completou a 4ª série/5º ano do Ensino Fund...	Não completou a 4ª série/5º ano do Ensino Fund...	
3	210052128335	580.7	678.9	638.9	659.5	780.0	B	B	Não completou a 4ª série/5º ano do Ensino Fund...	Não completou a 4ª série/5º ano do Ensino Fund...	
4	210051353021	497.7	532.4	457.6	582.6	780.0	D	E	Completou a 8ª série/9º ano do Ensino Fundamen...	Completou o Ensino Médio, mas não completou a ...	
8	210053417016	487.4	476.5	450.7	493.4	520.0	B	B	Não completou a 4ª série/5º ano do Ensino Fund...	Não completou a 4ª série/5º ano do Ensino Fund...	
9	210051128744	507.6	539.2	494.6	413.3	380.0	D	D	Completou a 8ª série/9º ano do Ensino Fundamen...	Completou a 8ª série/9º ano do Ensino Fundamen...	
...	...	...	...	...	...	...	...	...	...	...	
3389793	210054306230	506.0	405.2	416.3	450.4	240.0	B	B	Não completou a 4ª série/5º ano do Ensino Fund...	Não completou a 4ª série/5º ano do Ensino Fund...	
3389807	210051254419	435.6	531.2	534.7	399.2	320.0	H	C	Não sei	Completou a 4ª série/5º ano, mas não completou...	
3389814	210051121001	576.9	605.6	631.0	678.0	640.0	E	E	Completou o Ensino Médio, mas não completou a ...	Completou o Ensino Médio, mas não completou a ...	
3389815	210051173067	449.9	368.2	466.3	370.0	540.0	C	H	Completou a 4ª série/5º ano, mas não completou...	Não sei	
3389830	210051139675	563.7	646.0	550.7	706.4	660.0	E	D	Completou o Ensino Médio, mas não completou a ...	Completou a 8ª série/9º ano do Ensino Fundamen...	

# Resultados (hipótese 2)

dados\_brutos\_analisados.ipynb - Data Science - Visual Studio Code

Projeto\_grupo\_amarelo > dados\_brutos\_analisados.ipynb > Iniciaremos fazendo um análise para o Pai ou homem responsável pelo candidato:

+ Code + Markdown | ▶ Run All | Clear All Outputs | Restart | Variables | Outline | Python 3.10.10

Iniciaremos fazendo um análise para o Pai ou homem responsável pelo candidato:

```
print(df_hipotese_2_limpo.filter(items=['NU_INSCRICAO', 'DESCRICAO_Q001']).groupby('DESCRICAO_Q001').count().sort_values(by='NU_INSCRICAO', ascending=False))
```

[71]

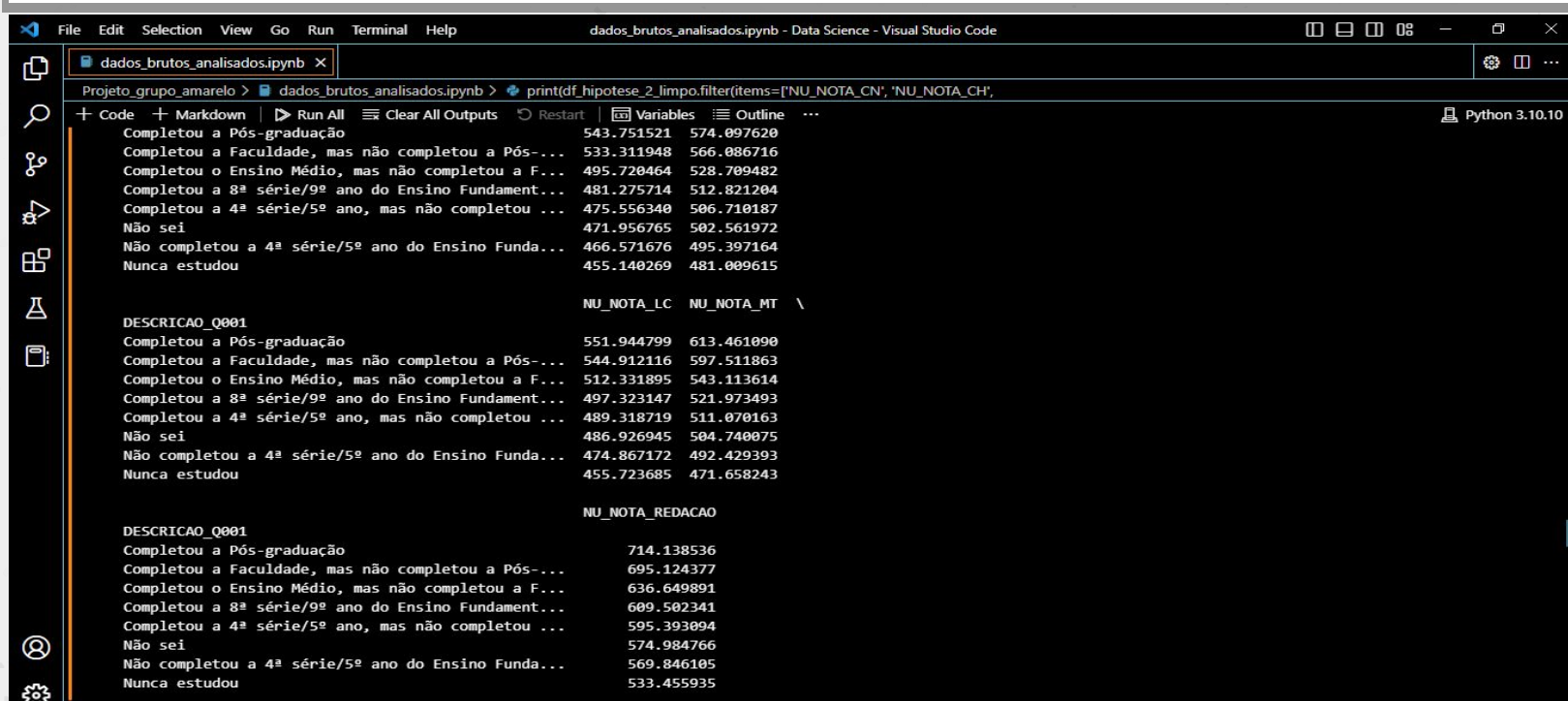
DESCRICAO_Q001	NU_INSCRICAO
Completou o Ensino Médio, mas não completou a F...	681530
Não completou a 4ª série/5º ano do Ensino Funda...	349980
Completou a 4ª série/5º ano, mas não completou ...	281447
Completou a Faculdade, mas não completou a Pós-...	246549
Completou a 8ª série/9º ano do Ensino Fundament...	245630
Completou a Pós-graduação	183606
Não sei	172965
Nunca estudou	76399

```
print(df_hipotese_2_limpo.filter(items=['NU_NOTA_CN', 'NU_NOTA_CH', 'NU_NOTA_LC', 'NU_NOTA_MT', 'NU_NOTA_REDACAO', 'DESCRICAO_Q001']).groupby('DESCRICAO_Q001').mean().sort_values(by='NU_NOTA_MT', ascending=False))
```

[72]

DESCRICAO_Q001	NU_NOTA_CN	NU_NOTA_CH	\
Completou a Pós-graduação	543.751521	574.097620	
Completou a Faculdade, mas não completou a Pós-...	533.311948	566.086716	
Completou o Ensino Médio, mas não completou a F...	495.720464	528.709482	
Completou a 8ª série/9º ano do Ensino Fundament...	481.275714	512.821204	
Completou a 4ª série/5º ano, mas não completou ...	475.556340	506.710187	

# Resultados (hipótese 2)



The screenshot shows a Jupyter Notebook titled 'dados\_brutos\_analisados.ipynb' in the Visual Studio Code editor. The notebook is running on Python 3.10.10. The code cell contains a print statement that filters data based on 'NU\_NOTA\_CN' and 'NU\_NOTA\_CH'. The output displays three tables of results, each with two columns of numerical values.

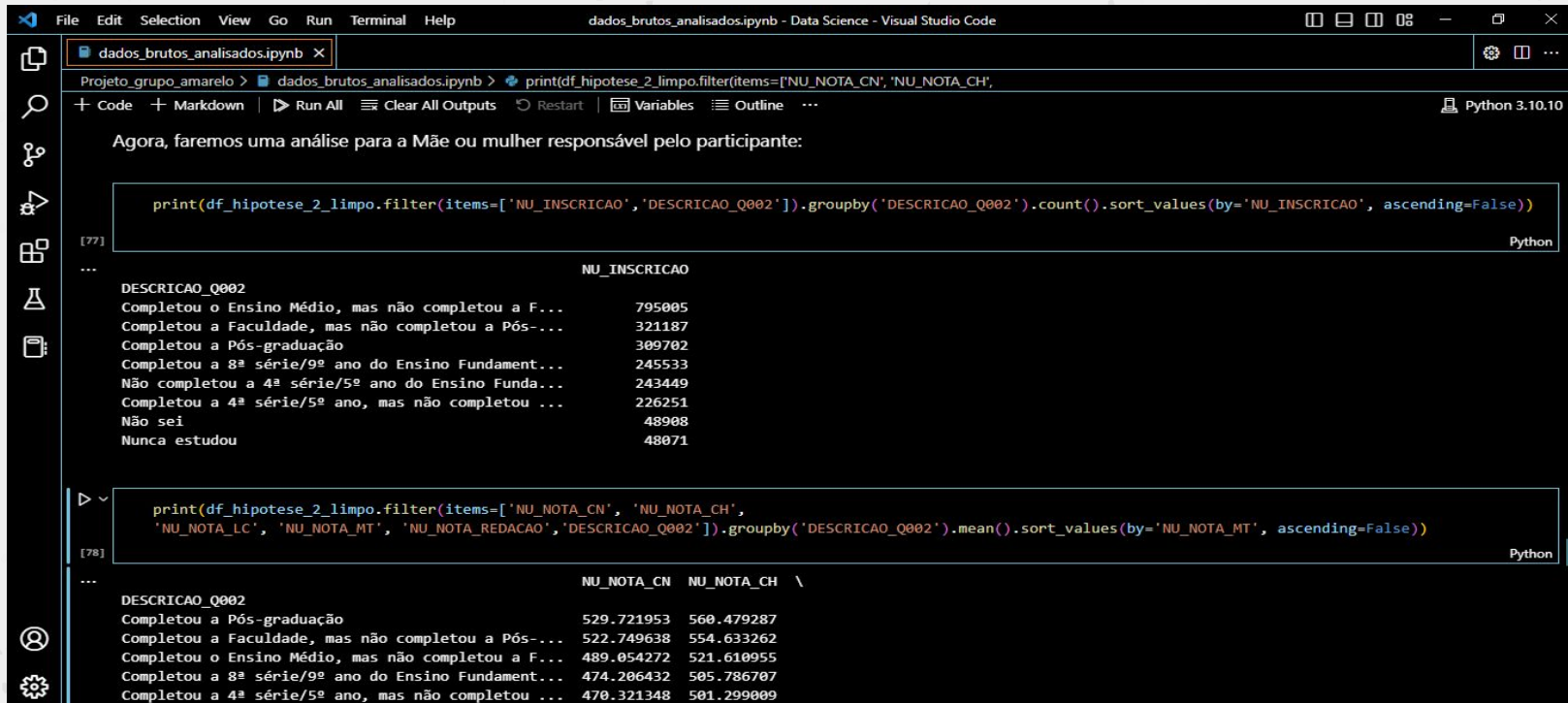
```
Projeto_grupo_amarelo > dados_brutos_analisados.ipynb > print(df_hipoteses_2_limpo.filter(items=['NU_NOTA_CN', 'NU_NOTA_CH'],
```

Completou a Pós-graduação	543.751521	574.097620
Completou a Faculdade, mas não completou a Pós...	533.311948	566.086716
Completou o Ensino Médio, mas não completou a F...	495.720464	528.709482
Completou a 8ª série/9º ano do Ensino Fundament...	481.275714	512.821204
Completou a 4ª série/5º ano, mas não completou ...	475.556340	506.710187
Não sei	471.956765	502.561972
Não completou a 4ª série/5º ano do Ensino Funda...	466.571676	495.397164
Nunca estudou	455.140269	481.009615

	NU_NOTA_LC	NU_NOTA_MT \
DESCRICAO_Q001		
Completou a Pós-graduação	551.944799	613.461090
Completou a Faculdade, mas não completou a Pós...	544.912116	597.511863
Completou o Ensino Médio, mas não completou a F...	512.331895	543.113614
Completou a 8ª série/9º ano do Ensino Fundament...	497.323147	521.973493
Completou a 4ª série/5º ano, mas não completou ...	489.318719	511.070163
Não sei	486.926945	504.740075
Não completou a 4ª série/5º ano do Ensino Funda...	474.867172	492.429393
Nunca estudou	455.723685	471.658243

	NU_NOTA_REDACAO
DESCRICAO_Q001	
Completou a Pós-graduação	714.138536
Completou a Faculdade, mas não completou a Pós...	695.124377
Completou o Ensino Médio, mas não completou a F...	636.649891
Completou a 8ª série/9º ano do Ensino Fundament...	609.502341
Completou a 4ª série/5º ano, mas não completou ...	595.393094
Não sei	574.984766
Não completou a 4ª série/5º ano do Ensino Funda...	569.846105
Nunca estudou	533.455935

# Resultados (hipótese 2)



dados\_brutos\_analisados.ipynb - Data Science - Visual Studio Code

Projeto\_grupo\_amarelo > dados\_brutos\_analisados.ipynb > print(df\_hipotese\_2\_limpo.filter(items=['NU\_NOTA\_CN', 'NU\_NOTA\_CH',

+ Code + Markdown | ▶ Run All | Clear All Outputs | Restart | Variables | Outline | Python 3.10.10

Agora, faremos uma análise para a Mãe ou mulher responsável pelo participante:

```
print(df_hipotese_2_limpo.filter(items=['NU_INSCRICAO', 'DESCRICAO_Q002']).groupby('DESCRICAO_Q002').count().sort_values(by='NU_INSCRICAO', ascending=False))
```

[77] Python

DESCRICAO_Q002	NU_INSCRICAO
Completou o Ensino Médio, mas não completou a F...	795005
Completou a Faculdade, mas não completou a Pós-...	321187
Completou a Pós-graduação	309702
Completou a 8ª série/9º ano do Ensino Fundament...	245533
Não completou a 4ª série/5º ano do Ensino Funda...	243449
Completou a 4ª série/5º ano, mas não completou ...	226251
Não sei	48908
Nunca estudou	48071

```
print(df_hipotese_2_limpo.filter(items=['NU_NOTA_CN', 'NU_NOTA_CH', 'NU_NOTA_LC', 'NU_NOTA_MT', 'NU_NOTA_REDACAO', 'DESCRICAO_Q002']).groupby('DESCRICAO_Q002').mean().sort_values(by='NU_NOTA_MT', ascending=False))
```

[78] Python

DESCRICAO_Q002	NU_NOTA_CN	NU_NOTA_CH	\
Completou a Pós-graduação	529.721953	560.479287	
Completou a Faculdade, mas não completou a Pós-...	522.749638	554.633262	
Completou o Ensino Médio, mas não completou a F...	489.054272	521.610955	
Completou a 8ª série/9º ano do Ensino Fundament...	474.206432	505.786707	
Completou a 4ª série/5º ano, mas não completou ...	470.321348	501.299009	

# Resultados (hipótese 2)

```
File Edit Selection View Go Run Terminal Help dados_brutos_analisados.ipynb - Data Science - Visual Studio Code
dados_brutos_analisados.ipynb X
Projeto_grupo_amarelo > dados_brutos_analisados.ipynb > print(df_hipoteses_2_limpo.filter(items=['NU_NOTA_CN', 'NU_NOTA_CH',
DESCRICAO_Q002
Completou a Pós-graduação 529.721953 560.479287
Completou a Faculdade, mas não completou a Pós-... 522.749638 554.633262
Completou o Ensino Médio, mas não completou a F... 489.054272 521.610955
Completou a 8ª série/9º ano do Ensino Fundament... 474.206432 505.786707
Completou a 4ª série/5º ano, mas não completou ... 470.321348 501.299009
Não sei 463.024280 487.489998
Não completou a 4ª série/5º ano do Ensino Funda... 463.683530 492.952720
Nunca estudou 453.429288 479.985226

NU_NOTA_LC NU_NOTA_MT \
DESCRICAO_Q002
Completou a Pós-graduação 540.180564 592.406291
Completou a Faculdade, mas não completou a Pós-... 535.110313 582.600755
Completou o Ensino Médio, mas não completou a F... 505.873281 533.379838
Completou a 8ª série/9º ano do Ensino Fundament... 489.907826 510.363869
Completou a 4ª série/5º ano, mas não completou ... 482.673122 501.083591
Não sei 470.752268 493.232768
Não completou a 4ª série/5º ano do Ensino Funda... 470.063631 484.995819
Nunca estudou 450.859231 464.279863

NU_NOTA_REDACAO
DESCRICAO_Q002
Completou a Pós-graduação 694.966322
Completou a Faculdade, mas não completou a Pós-... 679.834240
Completou o Ensino Médio, mas não completou a F... 624.292778
Completou a 8ª série/9º ano do Ensino Fundament... 593.212725
Completou a 4ª série/5º ano, mas não completou ... 579.444599
Não sei 538.759712
Não completou a 4ª série/5º ano do Ensino Funda... 556.906785
Nunca estudou 514.674960
```

## Resultados (hipótese 2)

- Procedimentos utilizados: eliminação de dados nulos; construção de dicionários acerca da escolaridade de cada responsável; agrupamento de média de notas de participantes por escolaridade do responsável;
- Conclui-se que a maioria dos responsáveis não possui ensino superior completo, ao passo que quanto maior o nível de escolaridade dos responsáveis, maior o desempenho do candidato.

# Resultados (hipótese 3)

Visual Studio Code interface showing a Jupyter Notebook with a Python script and its output.

The script reads a CSV file and displays the resulting DataFrame, `df_hipotese_3`.

```
df_hipotese_3 = pd.read_csv(r"C:\Users\willi\OneDrive\Área de Trabalho\Programação\Data Science\Projeto_grupo_amarelo\microdados_enem_2021\DAADOS\MICRODADOS_ENEM_2021\microdados_enem_2021\dados_brutos_analisados.csv", sep = ';', encoding = 'ISO-8859-1', usecols= colunas_hipotese_3)
```

The output shows the first 10 rows of the DataFrame:

	TP_COR_RACA	NU_NOTA_CN	NU_NOTA_CH	NU_NOTA_LC	NU_NOTA_MT	NU_NOTA_REDACAO
0	1	NaN	574.6	472.6	NaN	760.0
1	1	505.9	551.8	498.3	461.5	560.0
2	1	NaN	NaN	NaN	NaN	NaN
3	3	580.7	678.9	638.9	659.5	780.0
4	3	497.7	532.4	457.6	582.6	780.0
...	...	...	...	...	...	...
3389827	3	NaN	NaN	NaN	NaN	NaN
3389828	1	NaN	NaN	NaN	NaN	NaN
3389829	3	NaN	NaN	NaN	NaN	NaN
3389830	1	563.7	646.0	550.7	706.4	660.0
3389831	3	NaN	NaN	NaN	NaN	NaN

3389832 rows x 6 columns



# Resultados (hipótese 3)

Visual Studio Code interface showing a Jupyter Notebook with a data frame and a summary of results.

File Edit Selection View Go Run Terminal Help dados\_brutos\_analisados.ipynb - Data Science - Visual Studio Code

dados\_brutos\_analisados.ipynb x

Projeto\_grupo\_amarelo > dados\_brutos\_analisados.ipynb > Com o data frame criado, iremos seguir os mesmos procedimentos das hipóteses anteriores.

+ Code + Markdown | ▶ Run All | Clear All Outputs | Restart | Variables | Outline ... Python 3.10.10

	TP_COR_RACA	NU_NOTA_CN	NU_NOTA_CH	NU_NOTA_LC	NU_NOTA_MT	NU_NOTA_REDACAO	DESCRICAO_COR_RACA
1	1	505.9	551.8	498.3	461.5	560.0	Branca
3	3	580.7	678.9	638.9	659.5	780.0	Parda
4	3	497.7	532.4	457.6	582.6	780.0	Parda
8	3	487.4	476.5	450.7	493.4	520.0	Parda
9	3	507.6	539.2	494.6	413.3	380.0	Parda
...	...	...	...	...	...	...	...
3389793	1	506.0	405.2	416.3	450.4	240.0	Branca
3389807	3	435.6	531.2	534.7	399.2	320.0	Parda
3389814	1	576.9	605.6	631.0	678.0	640.0	Branca
3389815	3	449.9	368.2	466.3	370.0	540.0	Parda
3389830	1	563.7	646.0	550.7	706.4	660.0	Branca

2238107 rows x 7 columns

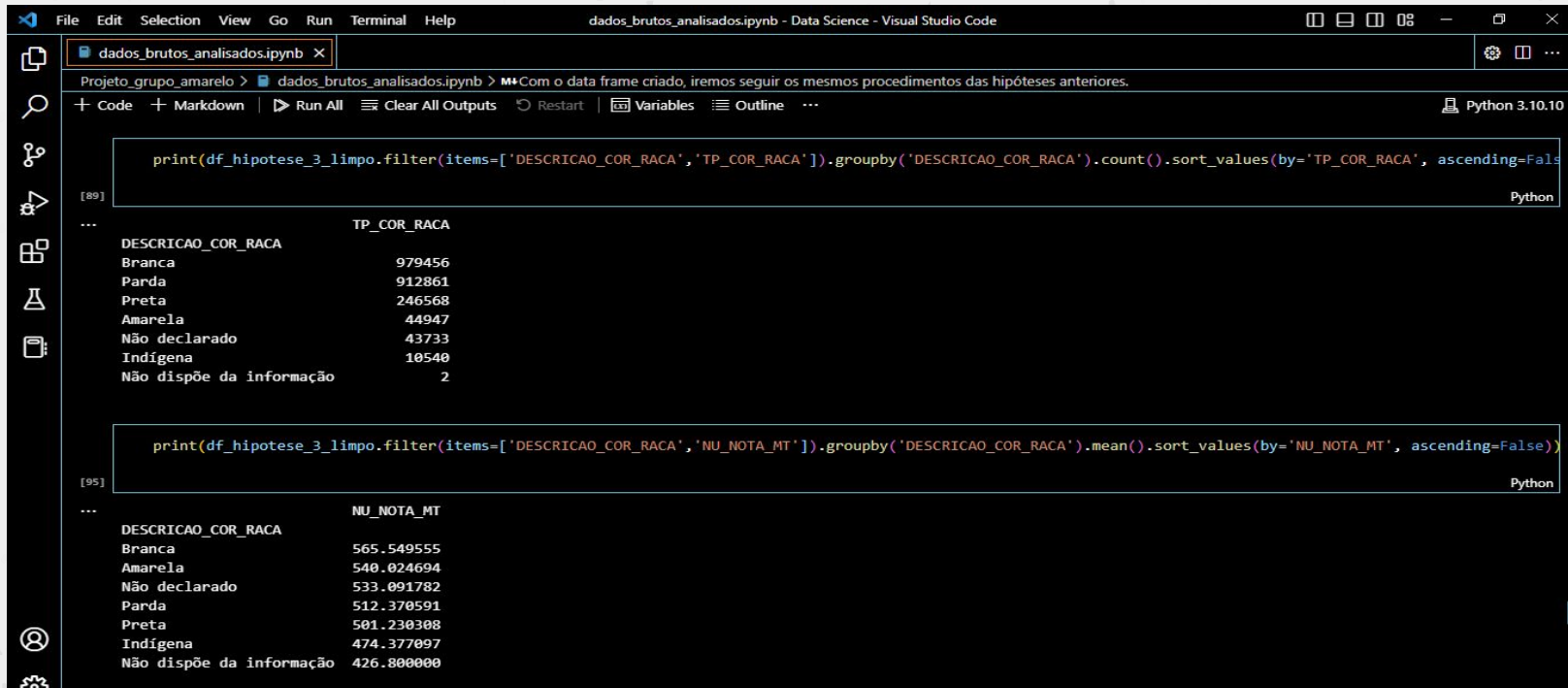
```
print(df_hipotese_3_limpo.filter(items=['DESCRICAO_COR_RACA', 'TP_COR_RACA']).groupby('DESCRICAO_COR_RACA').count().sort_values(by='TP_COR_RACA', ascending=False))
```

[89] Python

DESCRICAO_COR_RACA	TP_COR_RACA
Branca	979456
Parda	912861
Preta	246568



# Resultados (hipótese 3)



The screenshot shows a Jupyter Notebook titled 'dados\_brutos\_analisados.ipynb' in Visual Studio Code. The notebook contains two code cells. The first cell, labeled [89], executes a Python command to group data by 'DESCRICAO\_COR\_RACA' and sort by 'TP\_COR\_RACA'. The output is a table with 7 rows. The second cell, labeled [95], executes a Python command to group data by 'DESCRICAO\_COR\_RACA' and sort by 'NU\_NOTA\_MT'. The output is a table with 7 rows.

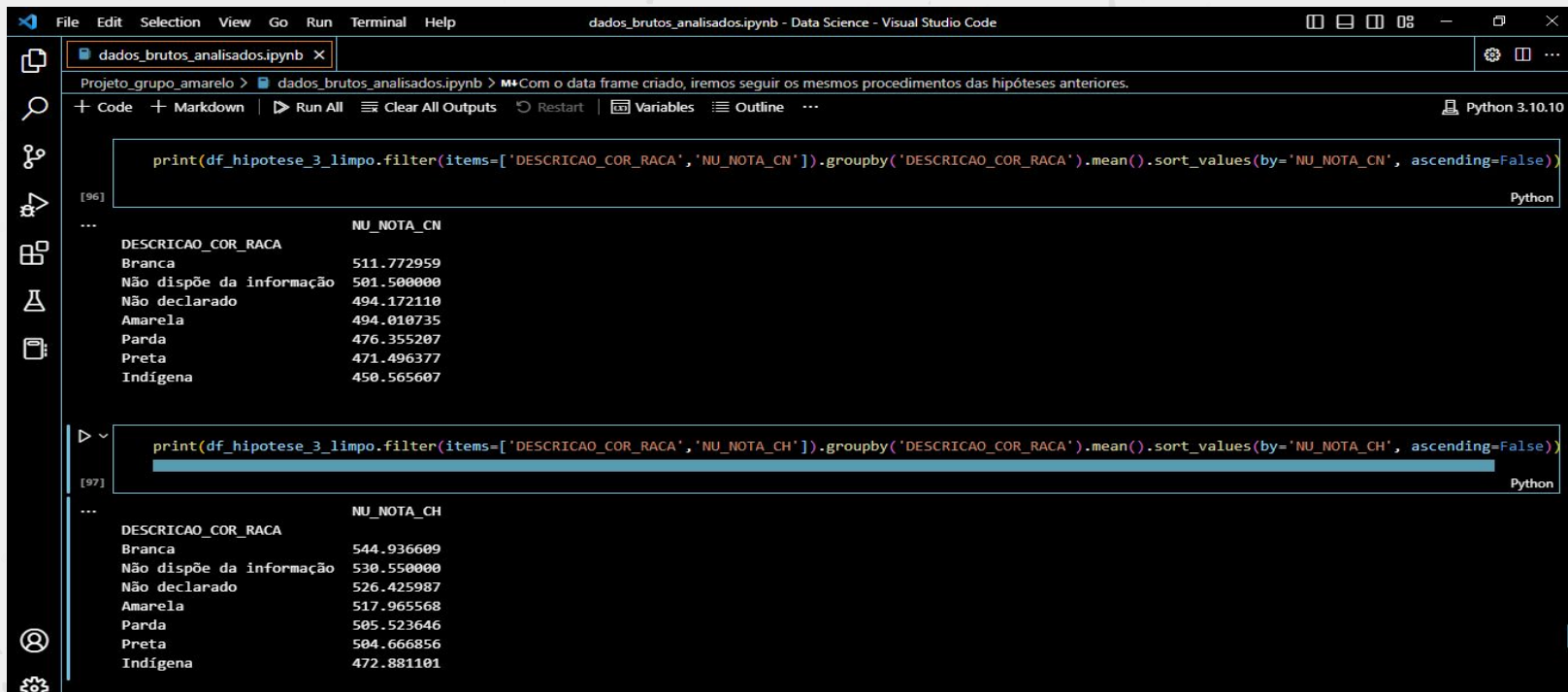
```
print(df_hipotese_3_limpo.filter(items=['DESCRICAO_COR_RACA','TP_COR_RACA']).groupby('DESCRICAO_COR_RACA').count().sort_values(by='TP_COR_RACA', ascending=False))
```

DESCRICAO_COR_RACA	TP_COR_RACA
Branca	979456
Parda	912861
Preta	246568
Amarela	44947
Não declarado	43733
Indígena	10540
Não dispõe da informação	2

```
print(df_hipotese_3_limpo.filter(items=['DESCRICAO_COR_RACA','NU_NOTA_MT']).groupby('DESCRICAO_COR_RACA').mean().sort_values(by='NU_NOTA_MT', ascending=False))
```

DESCRICAO_COR_RACA	NU_NOTA_MT
Branca	565.549555
Amarela	540.024694
Não declarado	533.091782
Parda	512.370591
Preta	501.230308
Indígena	474.377097
Não dispõe da informação	426.800000

# Resultados (hipótese 3)



dados\_brutos\_analisados.ipynb - Data Science - Visual Studio Code

Projeto\_grupo\_amarelo > dados\_brutos\_analisados.ipynb > Com o data frame criado, iremos seguir os mesmos procedimentos das hipóteses anteriores.

+ Code + Markdown | Run All | Clear All Outputs | Restart | Variables | Outline ... Python 3.10.10

```
[96] print(df_hipotese_3_limpo.filter(items=['DESCRICAO_COR_RACA', 'NU_NOTA_CN']).groupby('DESCRICAO_COR_RACA').mean().sort_values(by='NU_NOTA_CN', ascending=False))
```

Python

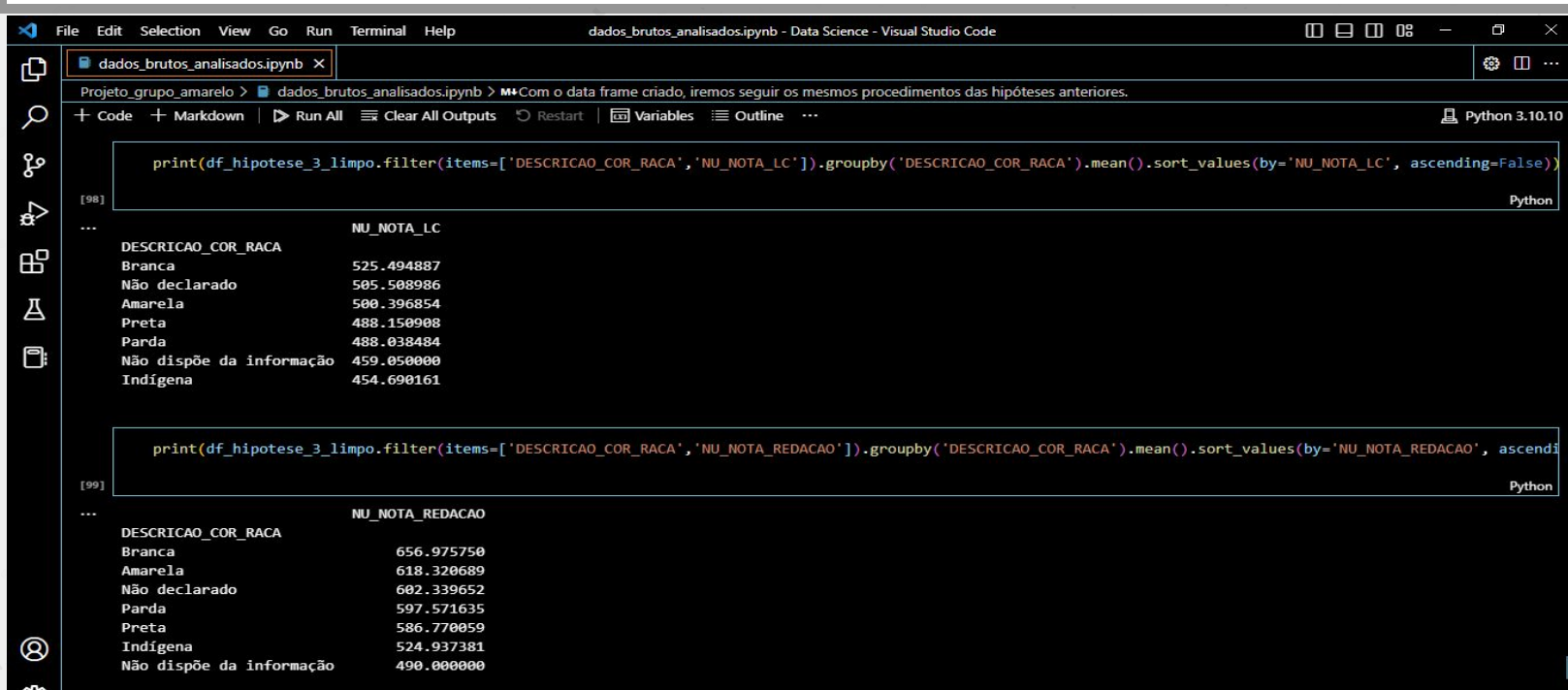
DESCRICAO_COR_RACA	NU_NOTA_CN
Branca	511.772959
Não dispõe da informação	501.500000
Não declarado	494.172110
Amarela	494.010735
Parda	476.355207
Preta	471.496377
Indígena	450.565607

```
[97] print(df_hipotese_3_limpo.filter(items=['DESCRICAO_COR_RACA', 'NU_NOTA_CH']).groupby('DESCRICAO_COR_RACA').mean().sort_values(by='NU_NOTA_CH', ascending=False))
```

Python

DESCRICAO_COR_RACA	NU_NOTA_CH
Branca	544.936609
Não dispõe da informação	530.550000
Não declarado	526.425987
Amarela	517.965568
Parda	505.523646
Preta	504.666856
Indígena	472.881101

# Resultados (hipótese 3)



dados\_brutos\_analisados.ipynb - Data Science - Visual Studio Code

Projeto\_grupo\_amarelo > dados\_brutos\_analisados.ipynb > M Com o data frame criado, iremos seguir os mesmos procedimentos das hipóteses anteriores.

+ Code + Markdown | ▶ Run All | ⌵ Clear All Outputs | ⌵ Restart | 📄 Variables | 📄 Outline ... Python 3.10.10

```
print(df_hipotese_3_limpo.filter(items=['DESCRICAO_COR_RACA','NU_NOTA_LC']).groupby('DESCRICAO_COR_RACA').mean().sort_values(by='NU_NOTA_LC', ascending=False))
```

[98] Python

DESCRICAO_COR_RACA	NU_NOTA_LC
Branca	525.494887
Não declarado	505.508986
Amarela	500.396854
Preta	488.150908
Parda	488.038484
Não dispõe da informação	459.050000
Indígena	454.690161

```
print(df_hipotese_3_limpo.filter(items=['DESCRICAO_COR_RACA','NU_NOTA_REDACAO']).groupby('DESCRICAO_COR_RACA').mean().sort_values(by='NU_NOTA_REDACAO', ascending=False))
```

[99] Python

DESCRICAO_COR_RACA	NU_NOTA_REDACAO
Branca	656.975750
Amarela	618.320689
Não declarado	602.339652
Parda	597.571635
Preta	586.770059
Indígena	524.937381
Não dispõe da informação	490.000000

## Resultados (hipótese 3)

- Procedimentos utilizados: contagem de participantes; agrupamento de estudantes por autodeclaração racial; média de notas considerando cada área do conhecimento.
- Conclui-se que os participantes que se autodeclararam brancos possuem desempenho mais satisfatório que os demais participantes.

# Considerações finais

- O desempenho de participantes no Enem está relacionado a fatores sociais, tais como: pertencimento a escola pública ou privada, raça e nível de escolaridade de responsáveis;
- Os dados indicam a necessidade de implementação de políticas públicas que garantam o acesso ao ensino superior de estudantes de escolas públicas, assim como melhorias do ensino fundamental público;
- Entende-se que outros fatores sociais e procedimentos podem ser selecionados e realizados para compreensão do contexto em estudo, tais como renda familiar e análise de desempenho por campos do conhecimento (Humanas e Exatas, por exemplo).



# Referências

PALHARES, I. Enem 2021 é o mais branco e elitista em mais de uma década. Folha, São Paulo, 2021. Disponível em: <https://www1.folha.uol.com.br/educacao/2021/09/enem-2021-e-o-mais-branco-e-elitista-da-decada.shtml>.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS (INEP). Microdados Enem 2021. Brasília, 2021. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>.