

Projeto Final - Pandas

Artur Porto – RA:791330

Escolha da base de dados

- Foi escolhida a base de dados "Popularidade de nomes de bebes nos EUA"
- Disponível em:
<https://www.kaggle.com/datasets/robikscube/us-baby-name-popularity>

Exploração inicial

Temos dois arquivos ao baixar essa base de dados. O primeiro, "**names.csv**", contem 4 colunas (name, sex, count, year).

O segundo, "**states.csv**", contem 7 colunas (state, sex, year, name, count, total, count_normalized).

Importando os dados:

```
[ ] names = pd.read_csv('names.csv')  
    states = pd.read_csv('states.csv')
```

Exploração inicial

Descrição geral da base

```
[ ] names.shape
```

```
(2052781, 4)
```

2052781 linhas e 4 colunas

```
[ ] names.columns
```

```
Index(['Name', 'Sex', 'Count', 'Year'], dtype='object')
```

```
[ ] states.shape
```

```
(6337734, 7)
```

6337734 linhas e 7 colunas

```
[ ] states.columns
```

```
Index(['State', 'Sex', 'Year', 'Name', 'Count', 'Total', 'Count_Normalized'], dtype='object')
```

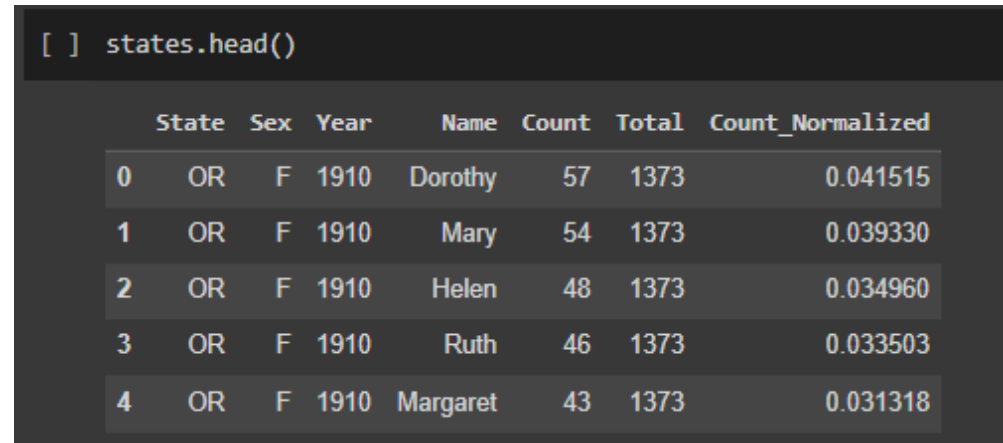
Exploração inicial

Visualização inicial:



A screenshot of a Jupyter Notebook cell. The command `names.head()` is entered in the input area. The output is a table with 5 rows and 5 columns: Name, Sex, Count, and Year. The rows show the top 5 names by count for the year 1997.

	Name	Sex	Count	Year
0	Emily	F	25735	1997
1	Jessica	F	21044	1997
2	Ashley	F	20895	1997
3	Sarah	F	20712	1997
4	Hannah	F	20594	1997



A screenshot of a Jupyter Notebook cell. The command `states.head()` is entered in the input area. The output is a table with 5 rows and 8 columns: State, Sex, Year, Name, Count, Total, and Count_Normalized. The rows show the top 5 names by count for the year 1910, grouped by state (OR).

	State	Sex	Year	Name	Count	Total	Count_Normalized
0	OR	F	1910	Dorothy	57	1373	0.041515
1	OR	F	1910	Mary	54	1373	0.039330
2	OR	F	1910	Helen	48	1373	0.034960
3	OR	F	1910	Ruth	46	1373	0.033503
4	OR	F	1910	Margaret	43	1373	0.031318

Exploração inicial

- Verificando atributos faltantes:

```
[ ] names.isna().mean().round(4).mul(100).sort_values(ascending=False)[:50]
```

```
Name      0.0  
Sex        0.0  
Count      0.0  
Year       0.0  
dtype: float64
```

Sem dados em branco/faltando

```
[ ] states.isna().mean().round(4).mul(100).sort_values(ascending=False)[:50]
```

```
State      0.0  
Sex        0.0  
Year       0.0  
Name       0.0  
Count      0.0  
Total      0.0  
Count_Normalized 0.0  
dtype: float64
```

Sem dados em branco/faltando

Isso significa que temos todos os dados a respeito do registro de nomes nos EUA??

registrados==nascidos?

NÃO

Análises das Variáveis

- Verificando a amplitude dos anos:
- Para o conjunto “names.csv” : 1880 - 2021
- Para o conjunto “states.csv” : 1910 – 2021
- Qual seria o ano com mais nomes registrados??

```
Ano com mais nomes registrados
```

```
[ ] st.mode(list(ano_states))
```

```
2008
```

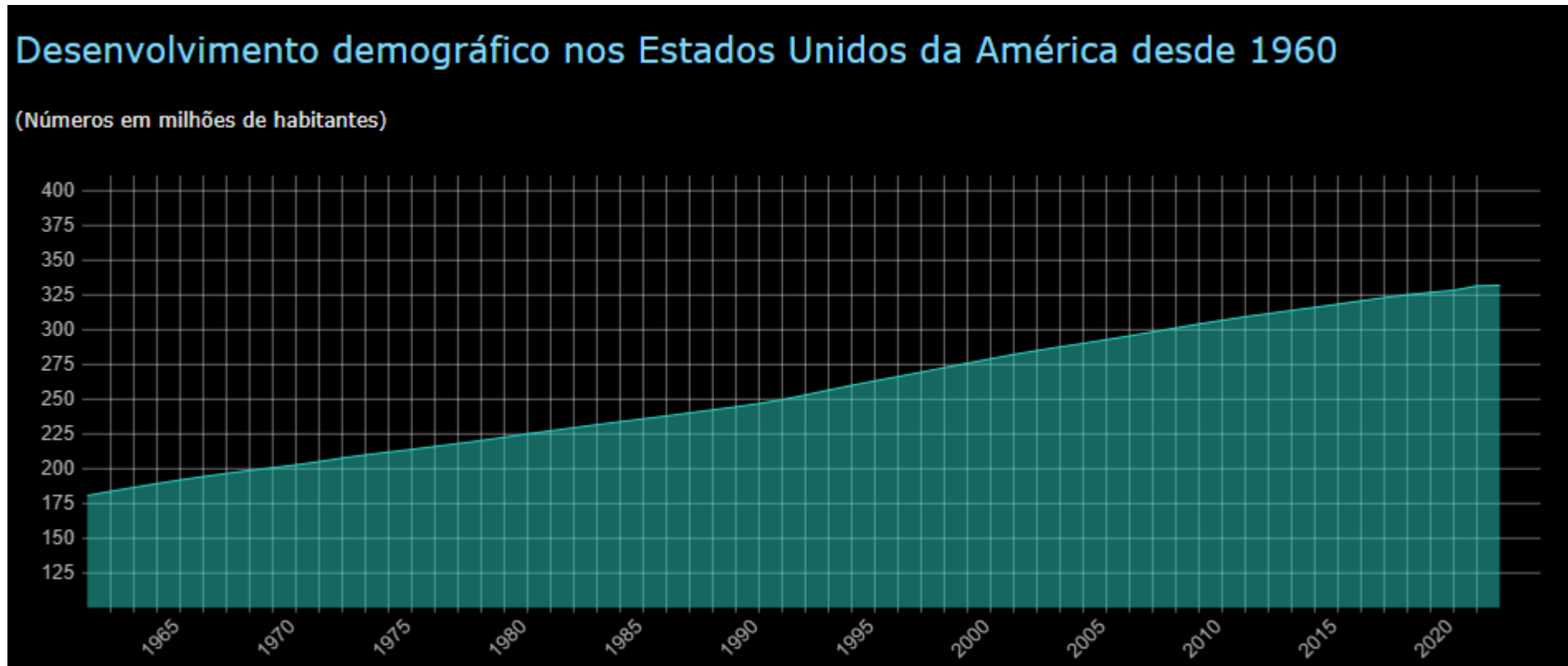
```
[ ] st.mode(list(ano_names))
```

```
2008
```

Esta em concordância
com a realidade??

Análises das Variáveis

Se olharmos o crescimento da população dos estados unidos, vemos que “O maior aumento foi registrado em 1961 com 1,67%”

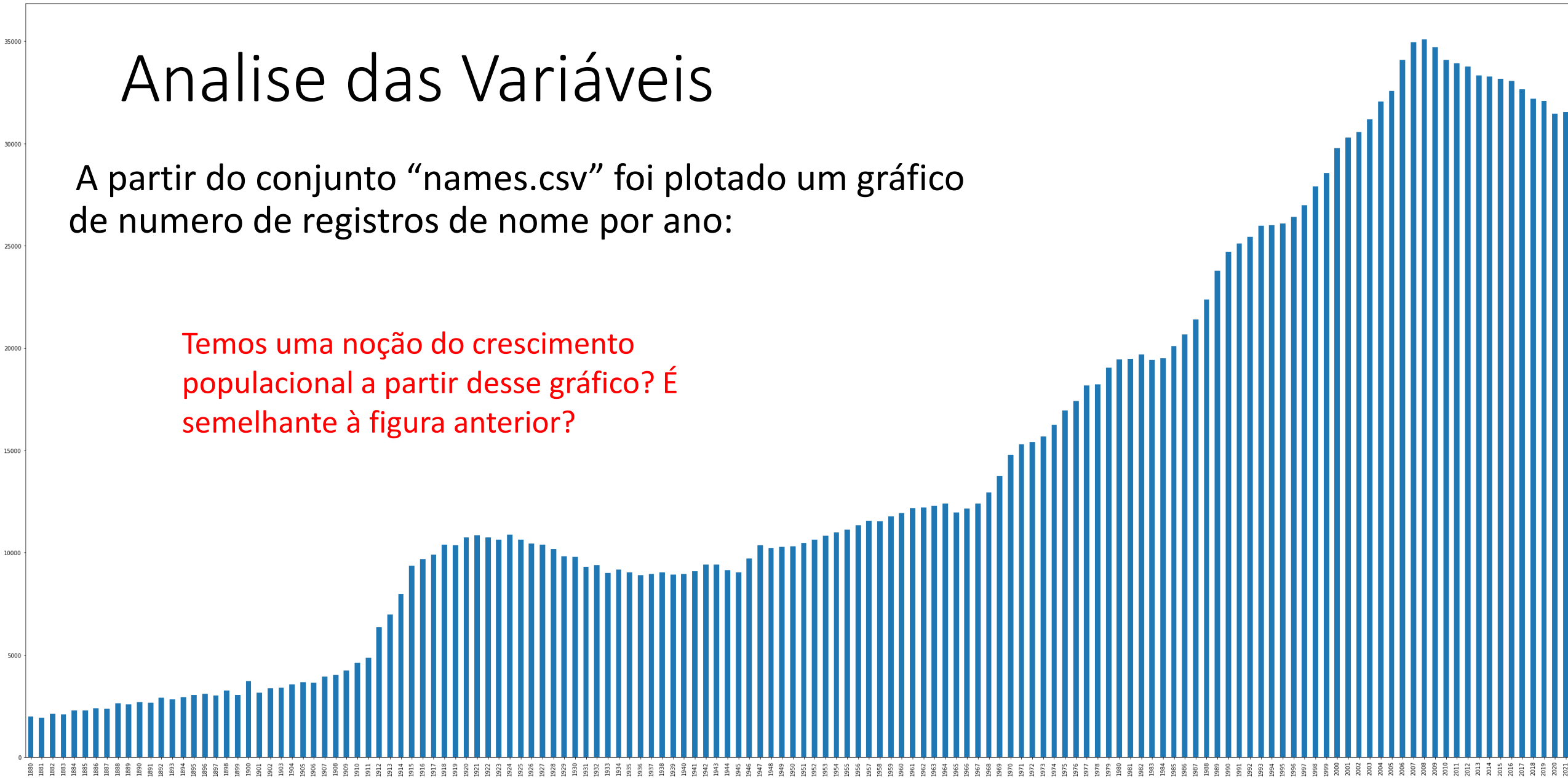


(Disponível: [Crescimento da população nos Estados Unidos da América \(dadosmundiais.com\)](https://dadosmundiais.com))

Analise das Variáveis

A partir do conjunto “names.csv” foi plotado um gráfico de numero de registros de nome por ano:

Temos uma noção do crescimento populacional a partir desse gráfico? É semelhante à figura anterior?



Análises das Variáveis

- Variável “Name” :

```
[ ] names['Name'].value_counts()
```

```
Jean      284
Johnnie   284
Jessie     284
Marion     284
Lee        284
...
Demeatra   1
Getrudes   1
Fordie     1
Flosie     1
Zkye       1
Name: Name, Length: 101338, dtype: int64
```

```
[ ] states['Name'].value_counts()
```

```
James      7453
Leslie     7418
Lee        7318
John       7267
Robert     7213
...
Morio       1
Seiki       1
Tsuyoshi   1
Yoshikatsu 1
Thyago      1
Name: Name, Length: 33492, dtype: int64
```

Análises das Variáveis

Temos nomes únicos?

```
Nomes unicos

[ ] nomes=states['Name']
    nome_unico_states = nomes.unique()

[ ] nomes=names['Name']
    nome_unico_names = nomes.unique()

[ ] nome_unico_names.size

101338

[ ] nome_unico_states.size

33492
```

Quais os nomes que mais aparecem?

```
[ ] nomes = names['Name']
    nomes.mode()

0    Francis
1     James
2      Jean
3      Jesse
4     Jessie
5       John
6    Johnnie
7    Joseph
8        Lee
9     Leslie
10    Marion
11     Ollie
12    Sidney
13    Tommie
14   William
dtype: object

[ ] nomes = states['Name']
    nomes.mode()

0    James
dtype: object
```

Para o conjunto maior
("states.csv") qual
será o nome feminino
mais registrado?

Conjunto com menor número de dados apresentou maior
numero de nomes únicos e é multimodal. (mais pobre?)

Análise das Variáveis

Qual será o nome feminino em “states.csv” que mais aparece?

Nome feminino que mais aparece

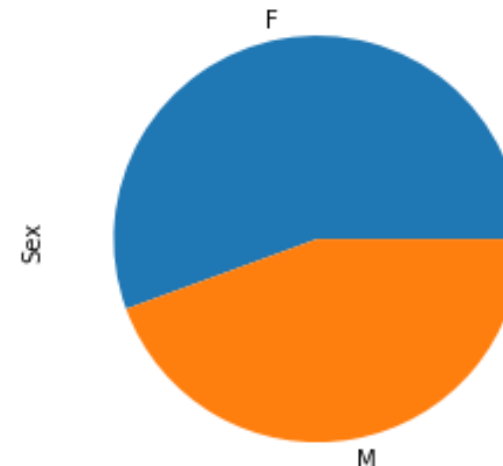
```
[ ] Filtro_sexo=states['Sex']=='F'  
    states_filtrado = states[Filtro_sexo]  
    nome_feminino = states_filtrado['Name']  
    nome_feminino.mode()
```

```
0    Elizabeth  
dtype: object
```

Qual será a divisão de nomes por sexo?

Nomes por sexo:

```
[ ] sexo=states['Sex']  
    z= sexo.value_counts().plot(kind = 'pie')
```



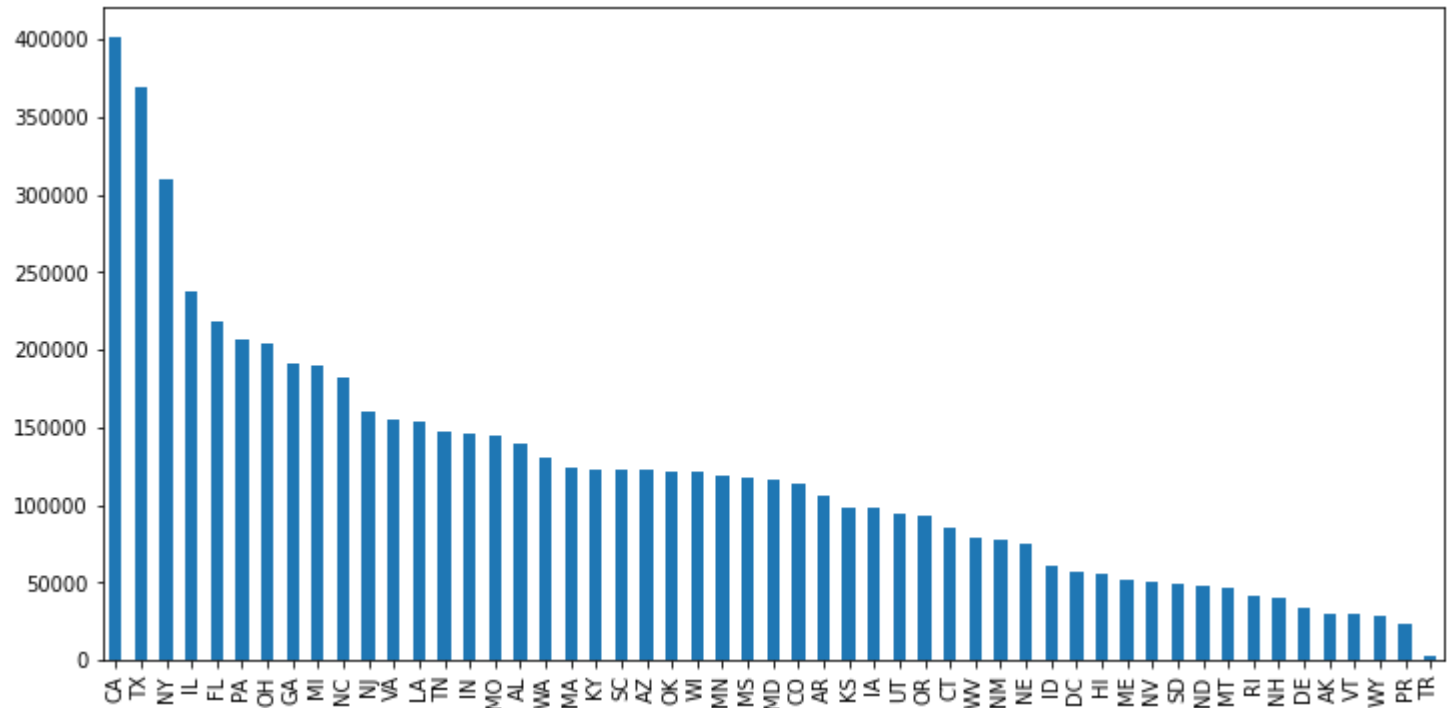
Análise das Variáveis

- E os estados?

Vamos olhar o número de nomes por estado:

```
[ ] estado=states['State']
```

```
[ ] x = estado.value_counts()  
    x.plot.bar(figsize=(12,6))
```



Por que a distribuição assume esse formato?

Análise das Variáveis

Por que a distribuição assume esse formato?

Vemos que os estados mais populosos naturalmente registram mais nomes ([Lista de estados dos Estados Unidos por população – Wikipédia, a enciclopédia livre \(wikipedia.org\)](#)).

Será que o nome mais popular (James) é o mais registrado no Estado mais populoso?

```
Nome que mais aparece na CA

CA - estado com mais nomes registrados

[ ] Filtro_CA=states['State']=='CA'
    states_filtrado_CA = states[Filtro_CA]
    nome_CA = states_filtrado_CA['Name']
    nome_CA.mode()

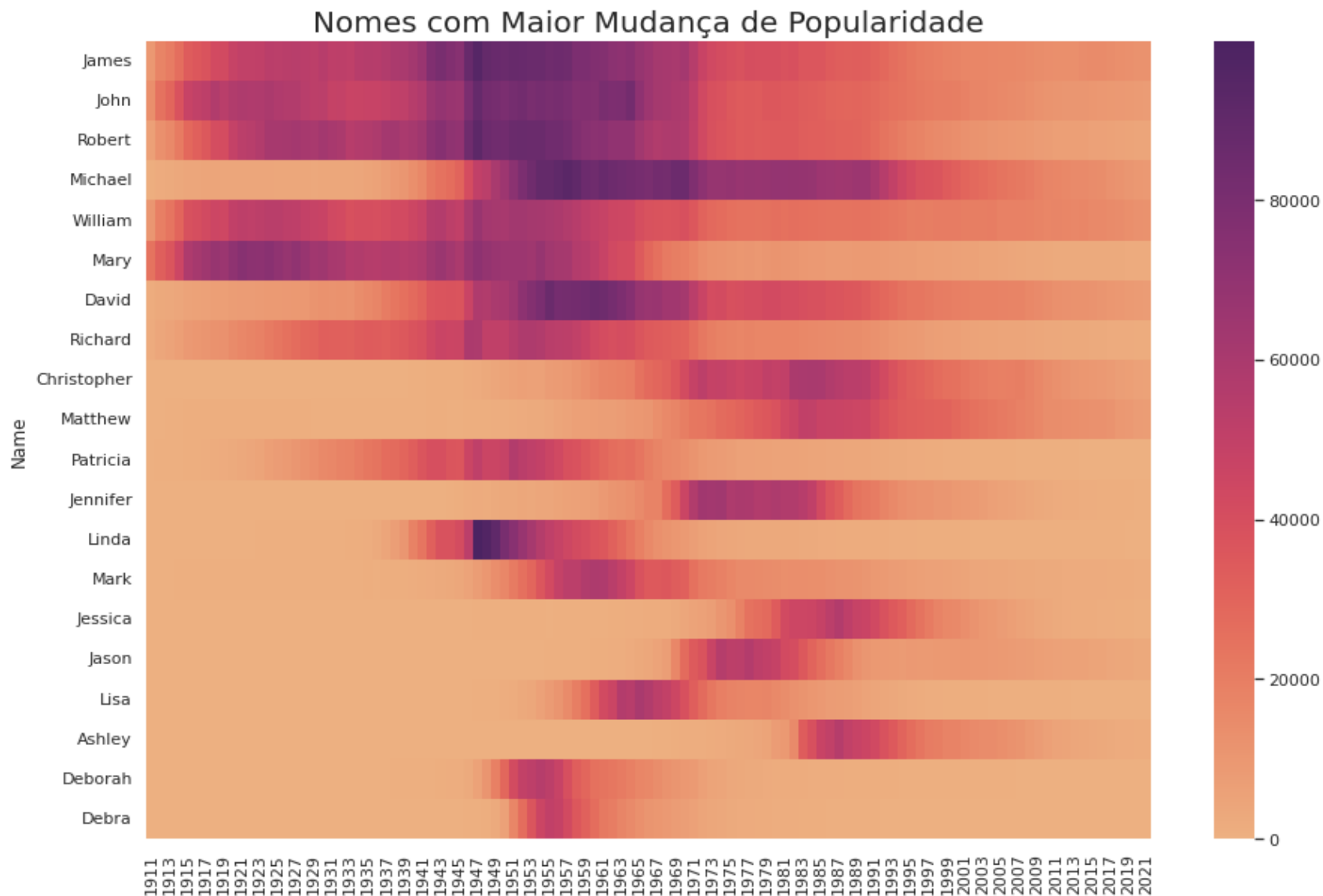
0    Jean
dtype: object
```

Não

Análise das Variáveis

Por fim, podemos fazer algumas análises utilizando mais de uma variável apenas.

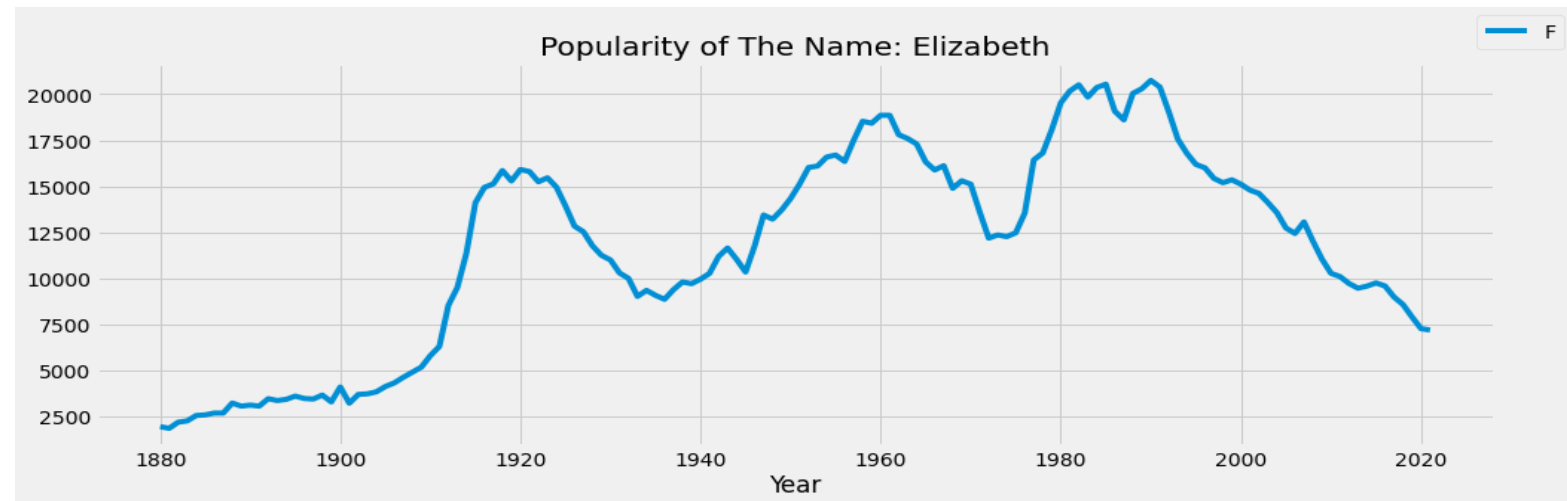
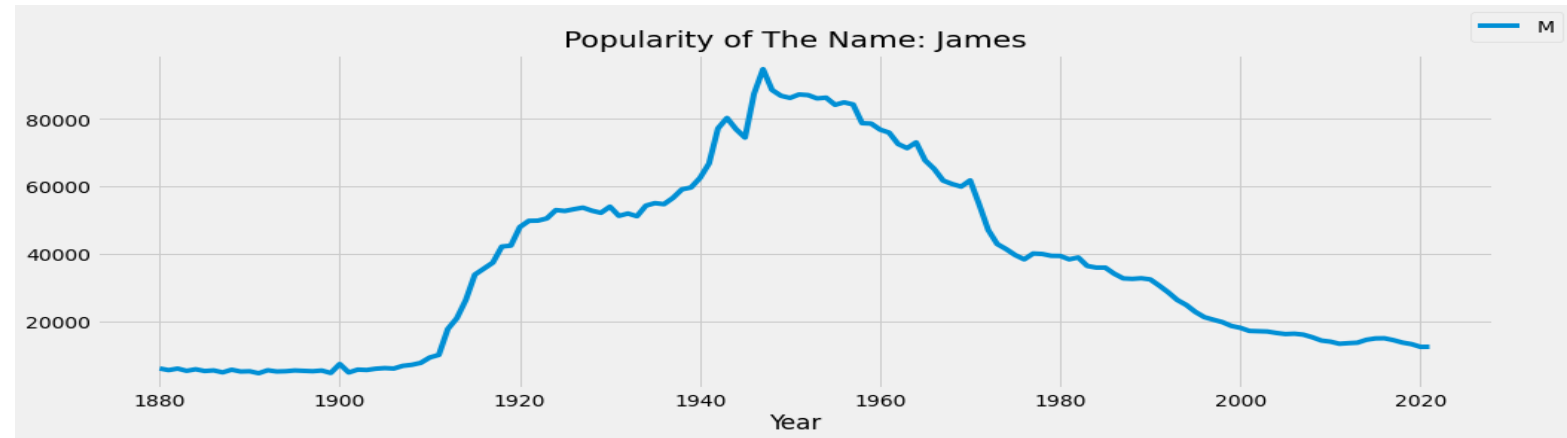
Aqui podemos ver a popularidade de alguns nomes no decorrer dos anos.



Análise das Variáveis

Essa análise pode
ser feita por nome.
Individualmente:

Aqui podemos ver
a popularidade do
nome no decorrer
dos anos.



Análise das Variáveis

Código: [US Baby Name Popularity | Kaggle](#)

Muito obrigado pela sua atenção!