Relatório de teste de estágio – análise de dados – Hilab
Candidato: Rodrigo P. Di Napoli

Para uma melhor abordagem do teste, separei os scripts em duas etapas: i) Dataset; ii) Intepretações dos dados

PARTE 1 – Dataset (Napoli_Estagio_HiLab.R)

Dataset

O dataset foi compilado a partir de suas 3 partes, provenientes de fontes diferentes (JavaScript, CSV, SQLite) perfazendo um dataset com 1.924.665 observações com 5 variáveis.

PARTE 2 – Interpretação dos resultados (Napoli_Estagio_HiLab_2.R)

Pergunta 1 - Quantos nomes diferentes existem por ano a partir de 2000? Forneça uma tabela e demonstre os resultados graficamente.

Na tabela 1 e figura 1 é possível observar o número de nomes únicos separados por ano (a partir de 2000), sendo possível concluir que, apesar de haver certa variação entre os anos, este valor é pequeno (amplitude de 4.998 e coeficiente de variação de 4,99%)

Tabela 1 – valores de nomes únicos separados por ano

year	count
2000	27512
2001	27980
2002	28279
2003	28886
2004	29501
2005	30153
2006	31624
2007	32416
2008	32510
2009	32242
2010	31623
2011	31449
2012	31279
2013	30834
2014	30731
2015	30583
2016	30386
2017	29910

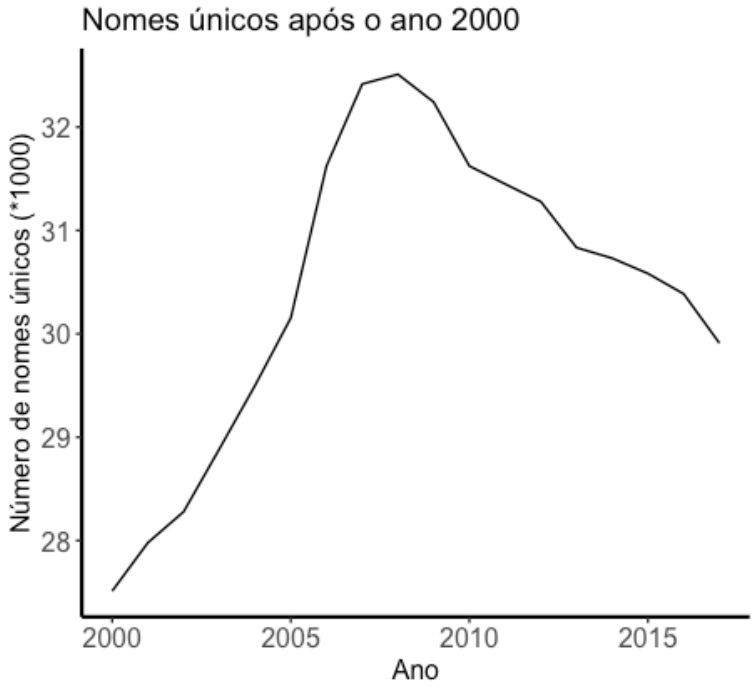

Figura 1 – Número de nomes únicos ao longo dos anos

Pergunta 2 - Qual a média e a mediana da contagem de bebês no dataset. Qual dessas medidas de tendência central você escolheria para descrever esse dado, justifique sua opção.

A média anual da contagem de bebês entre os anos de 1880 e 2017 é 2.522.612, já a mediana é de 3.037.679. Uma vez que há grande variação de dados (coeficiente de variação = 54,51%), a medida mais apropriada é a mediana, uma vez que esta não é afetada por valores extremos - caso da média.
Desta forma a interpretação seria que em 50% dos anos estudados houveram 3.037.679 nascimentos ou menos e nos outros 50% houveram 3.037.679 ou mais nascimentos.

Insight sobre a pergunta 2:
O gráfico de linhas (figura 2) evidencia de forma clara o aumento no número total de nascimentos ao longo dos anos. Este crescimento é esperado conforme as condições econômicas, sanitárias e médicas melhoram desde 1900. Após 1960 observa-se uma diminuição (possivelmente posterior aos "*baby boomers*" do momento pós guerra - 1952), e um certo grau de estabilização sugerindo um maior grau de desenvolvimento do país.

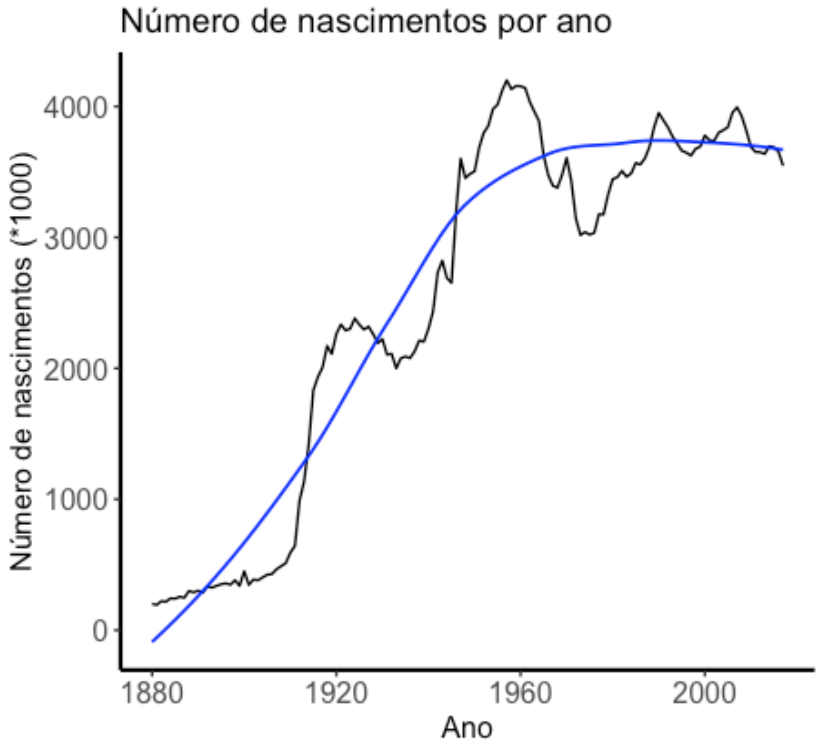

Figura 2 – Número total de nascimentos entre 1880 e 2017, linha de tendência em azul.

Abaixo (figura 3) é possível observar com mais detalhes o processo entre 1950 e 1980.

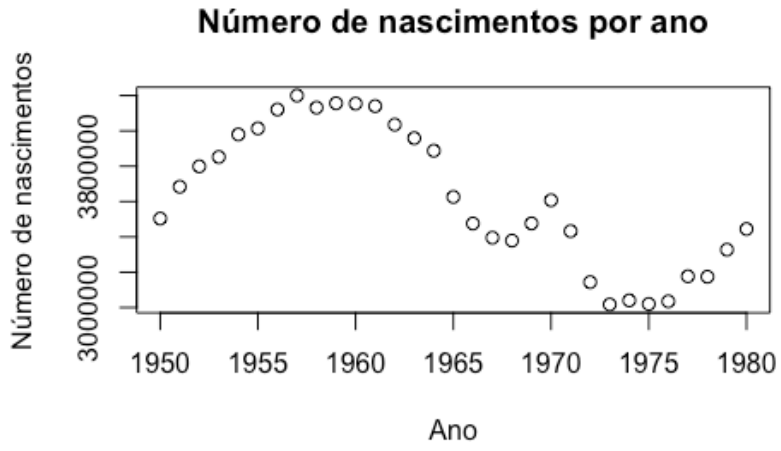

Figura 3 - Número total de nascimentos entre 1950 e 1980

Interessante observar que o ano de maior nascimento foi 1957 (4.200.007) e o ano com menor nascimentos foi 1881 (192.696).

Pergunta 3 - Qual a média e desvio padrão da contagem de bebês no ano de 1997?

Os valores calculados para média (134,40) e desvio padrão (1024,32) para bebês nascidos em 1997 não são cálculos que trazem grandes informações, uma vez que o que estes valores estão dizendo é que há em torno de 134 crianças nascidas em 1997 para cada nome distinto (com um desvio padrão de 1024, coeficiente de variação de 762%!!!).
Aqui o valor que me parece mais interessante é o valor total do número de nascimentos para este ano = 3.624.799

Pergunta 4 - Levando em conta que o dataset engloba o nascimento de todos os bebês do país imaginário Hilablândia. Qual o total de nascimentos no ano de 2002? Desses, quantos são do sexo feminino e quantos do sexo masculino?

Em 2002 nasceram 3.736.042 crianças, sendo 1.940.301 do sexo masculino (51,93%) e 1.795.741 do sexo feminino (48,07%) (tabela 2, figura 4)

Tabela 2 – numero de nascimentos no ano de 2002 separado por sexo

sexo	n	percentual
F	1795741	48,07
M	1940301	51,93

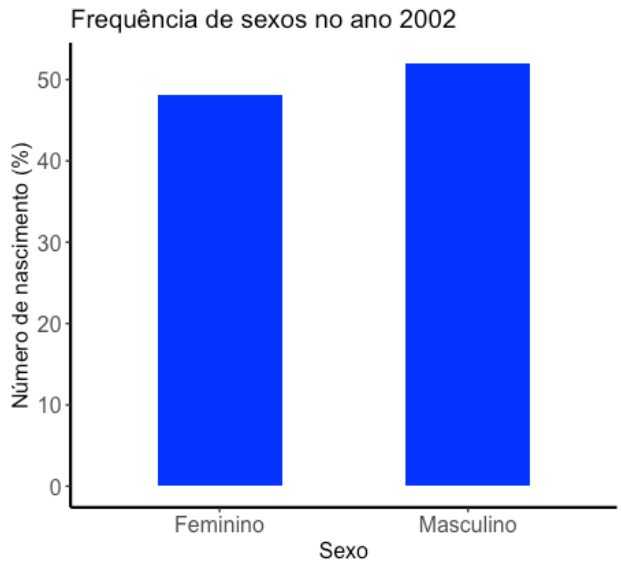

Figura 4 – Proporção de sexos nos nascimentos no ano de 2002.

Pergunta 5 - (Opcional)- Use sua criatividade para explorar a base e tentar retirar algum insight como por exemplo variação dos nomes Luke e Leia de acordo com lançamento dos filmes da série Star Wars.

Nomes mais populares de todos os tempos

O nome mais comum de todos os tempos em um único ano foi "Linda", com 99.686 meninas recebendo este nome no ano de 1947.

Nomes mais populares ao longo de todos os tempos

Obeservamos na tabela abaixo que os nomes mais populares em numeros absolutos foram (em ordem): James, John, Robert (Tabela 3). Em uma observação mais apurada desta tabela 3 é possível evidenciar que 9 dos 10 nomes são masculinos o que sugere que os meninos recebem menos nomes únicos do que meninas. Este fato é corroborado pelo fato de haverem 67.046 nomes únicos do sexo feminino (entre1880 a 2017) e 40.927 nomes únicos do sexo masculino (entre 1880 e 2017)

Tabela 3 – Nomes mais populares em termos absolutos ao entre os anos de 1880 e 2017

name	n
James	5173828
John	5137142
Robert	4834915
Michael	4372536
Mary	4138360
William	4118553
David	3624225
Joseph	2614083
Richard	2572613
Charles	2398453

Nomes mais populares nos últimos 20 anos (1998 – 2017)

No gráfico abaixo (figura 5) é possível observar quais os nomes foram mais comuns para meninos e meninas nos últimos 20 anos. Entre os anos 1999 e 2012 o nome Jacob foi o mais popular entre o sexo masculino e entre 1998 e 2007 o nome Emily foi o mais popular entre o sexo feminino. O que chama a atenção neste gráfico é que, uma vez que um nome deixa de ser popular, ele não retorna mais (pelo menos na escala de tempo observada)

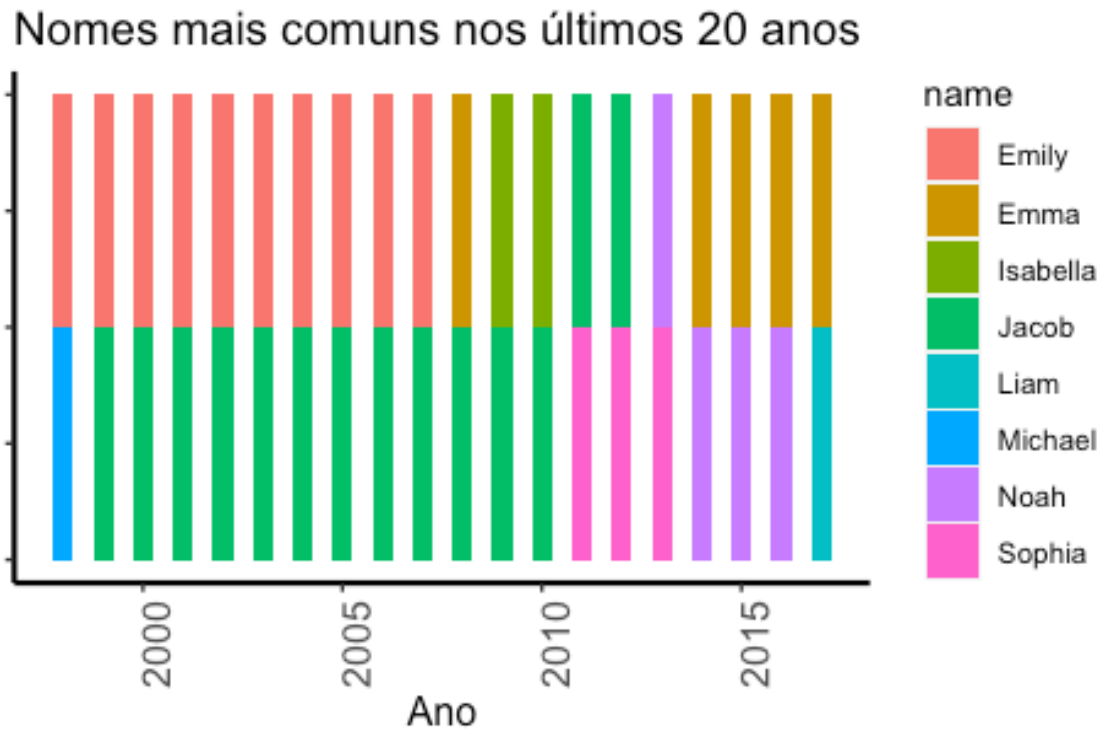

Figura 5 – Nomes de meninos e meninas mais comuns ao longo de 20 anos (1998-2017)

Além disso, podemos ver (tabela 4) que nos últimos 20 anos os nomes mais comuns na maioria dos anos foram:
Emily (para meninas) em 11 dos 20 anos estudados
Jacob (para meninos) em 14 dos 20 anos estudados

Tabela 4 – Nomes mais comuns (por ano) nos últimos 20 anos

Nome	Freq
Jacob	14
Emily	10
Emma	5
Noah	4
Sophia	3
Isabella	2
Liam	1
Michael	1

Sugestão de análise feita no roteiro: Influência dos filmes da franquia Star Wars ao nomear os filhos e filhas de Luke e Leia

Os dados mostram que os nomes Luke e Leia se tornaram mais populares em momentos diferentes, Luke já vinha aumentando sua proporção ao longo dos anos, e pareceu ter um aumento ainda maior após o primeiro filme da franquia (1977) (figura 6). Já o nome Leia parece ter sofrido algum impacto com a primeira trilogia, mas posteriormente ao 6º filme (2005) houve um incremento considerável (figura 6). É possível propor que os indevidos que eram crianças ao assistiram a primeira trilogia estejam em idade reprodutiva por volta dos anos 2000. Este efeito pode não ter sido observado em relação ao nome Luke, pois já era um nome muito mais comum, independente dos filmes, o que pode ser observado ao comparar os dois nomes na mesma escala (figura 7).

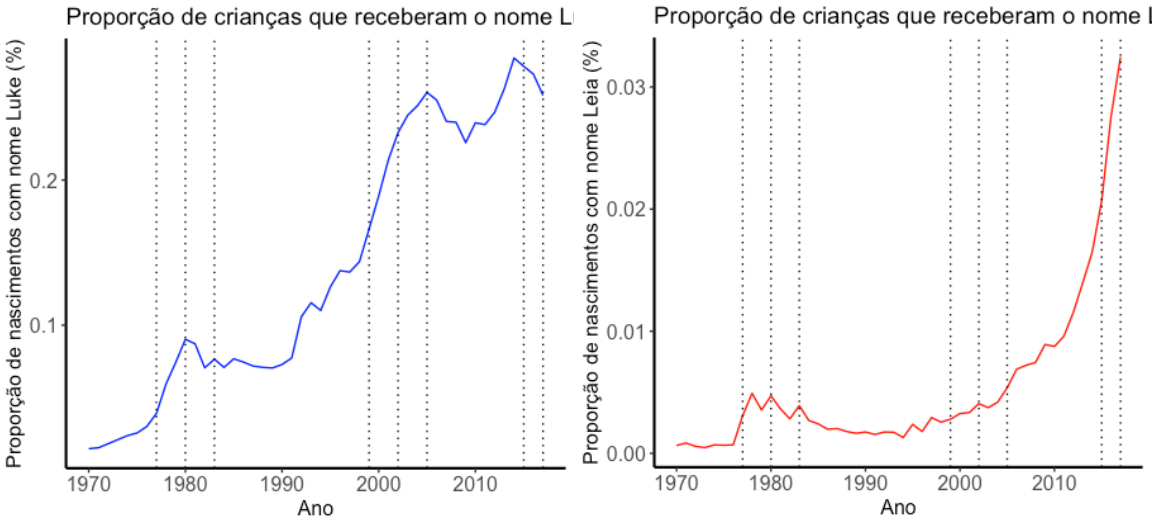

Figura 6 – Proporção de crianças que receberam o nome de “Luke” (gráfico à esquerda, linha azul) e “Leia” (gráfico à direita, linha vermelha) entre os anos de 1970 e 2017. Linhas pontilhadas representam os anos de lançamento de filmes da franquia Star Wars

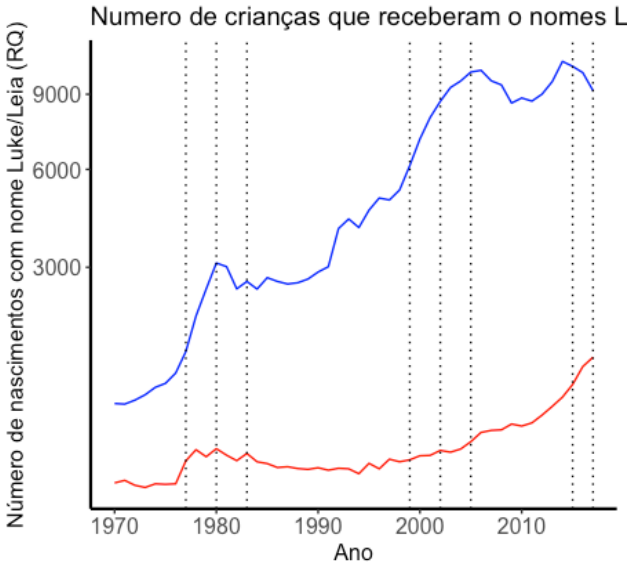

Figura 7 – Número de crianças que receberam os nomes “Luke” (em azul) e “Leia” (em vermelho) entre os anos de 1970 e 2017. Linhas pontilhadas representam os anos de lançamento de filmes da franquia Star Wars. Uma vez que os valores são muito distantes foi utilizada uma transformação de raiz quadrada do eixo Y.