UFRJ - ANALYTICA
RELATÓRIO A3

Cauê Caviglioni Daniel Silva

1) MOTIVAÇÃO

Para a realização desta etapa do processo seletivo, fazia-se necessário escolher um banco de dados, de tal forma a realizar previsões de relevância no contexto social brasileiro. Dessa maneira, o banco de dados escolhido dizia respeito à população brasileira ao longo das últimas décadas. O banco de dados pertencia à plataforma Banco de Dados.

2) OBJETIVO

A partir do comportamento da população brasileira, o objetivo é estimar a quantidade de pessoas que habitam o Brasil. Dessa maneira, será possível cruzar com outros bancos de dados (tais como quantidade de alimentos, quantidade de desempregados), de modo a se ter uma noção clara e ampla dos possíveis problemas futuros no nosso país. Algumas perguntas a serem respondidas, com base nas predições, são:

- Qual a importância de fazer predições da população brasileira?
- A análise exploratória foi feita corretamente?
- Os modelos usados se ajustam adequadamente aos dados?
- As predições fazem sentido?
- O que poderia ser feito para melhorar as predições?

3) ANÁLISE EXPLORATÓRIA

Com o CSV importado e carregado pelo Python, foi necessário analisar a estrutura do banco de dados, de forma a identificar problemas, lacunas ou desinformações presentes, que podem prejudicar a eficiência e o desenvolvimento dos modelos de previsão.

Como mostrado no notebook, a estrutura dos dados era bastante simples, composta por duas colunas (“ano” e “população”) e 30 linhas, representando a população de 1991 até 2020.

Figura 1 - Estrutura dos dados

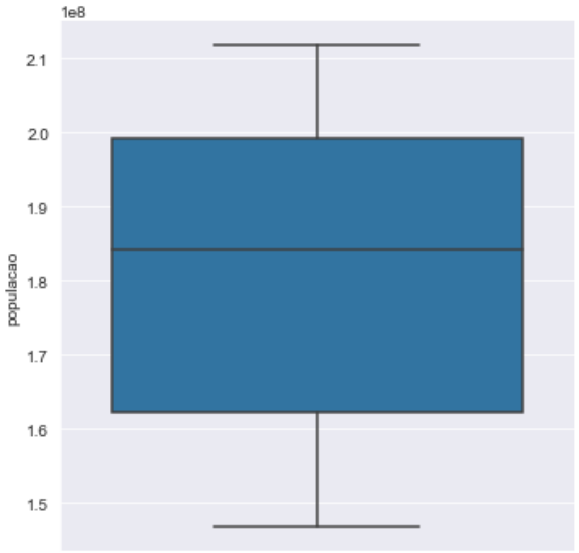
	ano	populacao
0	1991	146815815
1	1992	149236984
2	1993	151571727
3	1994	153725670
4	1995	155822440

Uma vez que o formato dos dados foi entendido, a próxima etapa foi buscar possíveis falhas nas linhas dos dados, de modo a evitar que tais erros influenciassem negativamente as previsões.

Assim, em primeiro lugar, foi checado se havia linhas em que não havia informação disponível (ex.: Nan). Como mostrado no notebook, pode-se verificar que, para cada ano, havia uma população registrada.

Em segundo lugar, foi checado a existência de outliers ou de quaisquer inconsistências nos dados. Para isso, foi plotado um boxplot.

Figura 2 - Boxplot da população

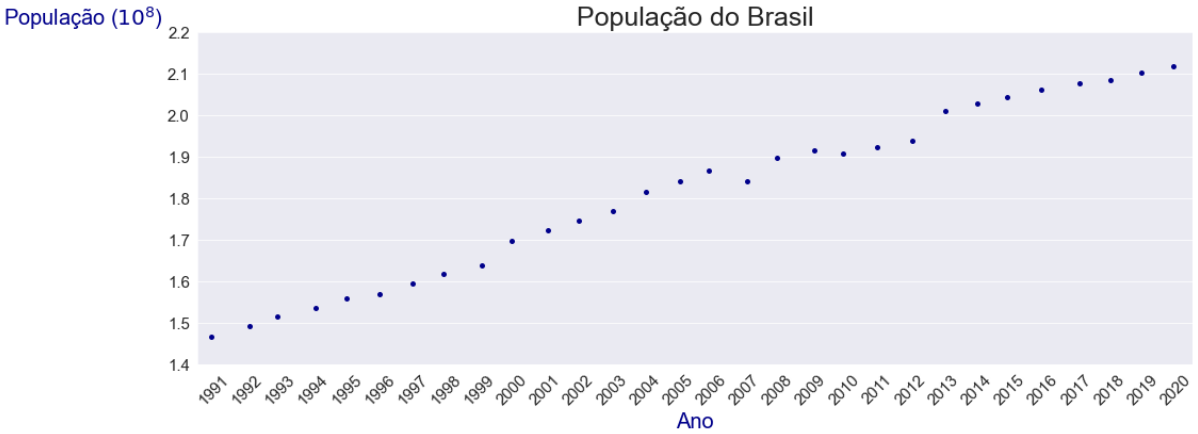


A partir da figura acima, pode-se afirmar que não havia outliers, tampouco dados cujo comportamento era incomum ou até mesmo incompatível.

Portanto, devido à análise prévia e à relativa simplicidade dos dados, pode-se dizer que eles podem ser trabalhados sem mudanças significativas.

Por último, uma vez que o objetivo é realizar previsões, foi preciso entender o comportamento da população brasileira ao longo dos anos. Assim, foi plotado um gráfico de dispersão.

Figura 3 - Gráfico de dispersão da população brasileira



A partir da figura acima, é nítido que o comportamento da população assemelha-se muito ao de uma reta. No entanto, sabe-se que muitas populações se comportam de uma forma exponencial. Dessa forma, não era possível saber se este intervalo temporal disponível de dados representava uma reta ou um trecho da exponencial. Dessa forma, foram feitos dois ajustes a esse conjunto de dados, de modo a compará-los depois.

4) AJUSTE DAS CURVAS AOS DADOS

Como falado no tópico anterior, as curvas escolhidas para representar a população brasileira ao longo dos anos (e fazer previsões) foram a reta e a exponencial, cujas formas estão mostradas abaixo.

Equação 1 - Reta

$$y = ax + b$$

Equação 2 - Exponencial

$$y = be^{ax}$$

Para encontrar tais funções, foi necessário determinar suas constantes. O método escolhido foi o ajuste por mínimos quadrados. As funções inerentes a este ajuste foram desenvolvidas em notebook separado (funcoes_de_apoio.ipynb).

As constantes calculadas pelos algoritmos implementados no notebook secundário foram:

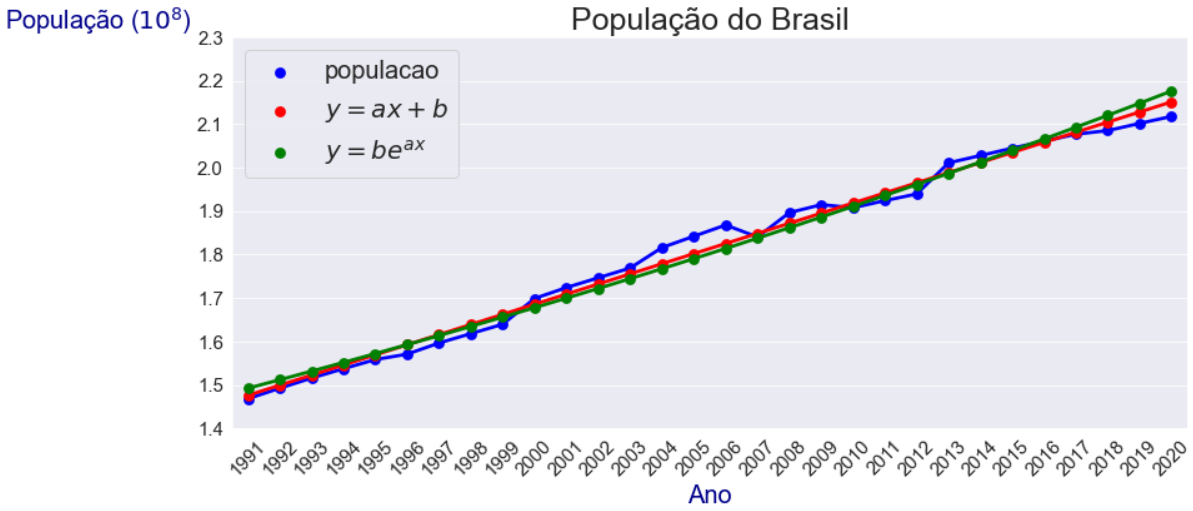
- Para a reta:
a = 2327616.9255
b = -4486687991.4655

- Para a exponencial:
a = 0.0130
b = 0.0008

5) COMPARAÇÃO DAS APROXIMAÇÕES COM OS DADOS REAIS

Uma vez que os coeficientes das curvas foram calculados, tornou-se possível plotar, em um gráfico só, o comportamento real da população, a aproximação pela reta e a aproximação pela exponencial.

Figura 4 - Comparação entre as curvas

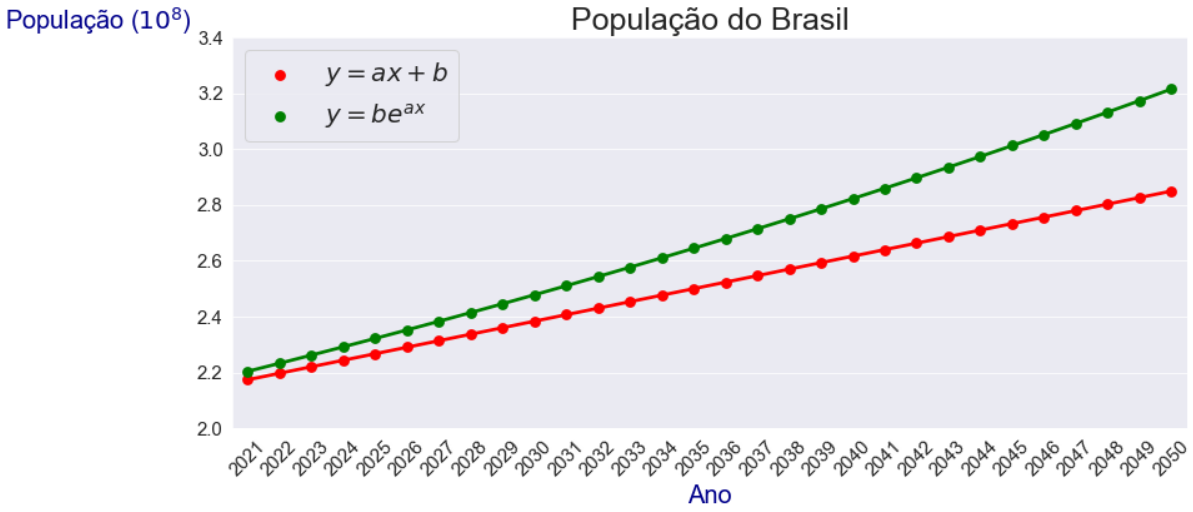


Como observado na imagem acima, as curvas escolhidas, dada à simplicidade dos dados, ajustam-se devidamente à população brasileira real. Apesar de bem próximas, pode-se dizer que o ajuste da reta é mais fiel à população real. Todavia, ao olhar para os extremos do intervalo temporal, nota-se que a exponencial distancia-se consideravelmente da população real. Além disso, em trechos em que a população apresentou um comportamento mais instável, ambas as curvas de ajuste se distanciam da real. De qualquer forma, posto que na maioria das vezes a população do país assume um crescimento constante, façamos algumas estimativas.

6) PREVISÕES DA POPULAÇÃO BRASILEIRA ENTRE 2021 E 2050

Como dito anteriormente, usando as funções de reta e exponencial ajustadas aos dados, foram feitas previsões da população brasileira de 2021 a 2050.

Figura 5 - Previsões



A diferença mais notável entre as previsões da reta e da exponencial é a divergência ao longo do tempo. Outros comentários acerca de tais previsões serão feitas no próximo tópico de conclusões.

7) CONCLUSÕES

Antes de falar sobre as conclusões de fato, deve-se comentar sobre as limitações dos modelos empregados para realizar as predições e do banco de dados em si. A primeira limitação consiste no fato de que, tanto a exponencial quanto a reta, não representam possíveis quedas da população brasileira de um ano para o outro. Em outras palavras, os modelos assumem um comportamento estritamente crescente, enquanto os dados da população, às vezes, apresentam uma diminuição da população. Em relação aos dados usados para a construção dos modelos, vale ressaltar que o fato de apresentar somente duas colunas impede uma análise mais profunda e, consequentemente, o desenvolvimento de modelos mais rigorosos. Isso ocorre, pois é sabido que o comportamento da população de um país depende de vários fatores, tais como taxa de natalidade, taxa de mortalidade, aspectos culturais e economia, por exemplo. Se houvesse informações acerca desses fatores, seria possível implementar um algoritmo genético ou até mesmo uma rede neural artificial, de modo a chegar mais próximo do comportamento real da população brasileira. Além disso, o intervalo temporal do dado é pequeno para se estabelecer relações mais precisas, algo que também restringe a eficiência das predições.

Ademais, tentemos responder as perguntas estabelecidas no tópico 2. Mesmo que já dito, a importância de predizer a população brasileira no futuro não se reduz a simplesmente ter um dado numérico. Uma vez que se tem em mãos a população, torna-se possível cruzar tal informação com outras, de modo a tomar atitudes, traçar planos e evitar futuros problemas. Um exemplo disso (com um dado presente na plataforma de banco de dados recomendada) são os dados da quantidade de alimentos produzidos em cada estado do Brasil. Ao cruzar esses dois dados, seria possível saber se faltaria comida ou, no caso de excesso, como seria a melhor maneira de distribuição.

No tocante à análise exploratória, pode-se afirmar que foi feita de forma correta. O dado era relativamente simples e não apresentava nenhuma incoerência significativa que pudesse atrapalhar o desenvolvimento dos modelos propostos.

Em relação aos modelos utilizados, pode-se dizer que, para os anos em que a população brasileira era conhecida, ambas as curvas escolhidas se adequam aos dados, salvo algumas diferenças já discutidas no tópico 5. No entanto, aqui volta o problema das limitações apontadas no primeiro parágrafo deste tópico. Uma simples mudança no comportamento das pessoas, como casamentos e formação de família tardios, poderia causar quedas ou um crescimento mais achatado na população. Isso faria com que os modelos da reta e exponencial realizassem predições imprecisas e significativamente incorretas. Dessa maneira, vale ressaltar que os modelos implementados não servem para longos prazos, necessitando de constantes atualizações, com base nas novas populações.

Todavia, não se deve descartar as predições calculadas. O ajuste da reta, por exemplo, aproxima-se muito bem da população real, podendo ser usada para predizê-la nos próximos 5-10 anos. O mesmo vale para o ajuste exponencial. Tudo

depende de um estudo mais detalhado do comportamento da população. Se houver uma noção mais profunda, será possível aprimorar tais modelos.

No que tange às melhorias das predições, é interessante fazer alguns comentários de como fazê-las. Melhores previsões estão acompanhadas de melhores informações. Quanto mais informações houver acerca da população brasileira, mais fácil será de criar modelos avançados (redes neurais artificiais, por exemplo) que farão previsões mais precisas e completas. Com os dados que estavam disponíveis na plataforma, talvez se as quedas nas populações fossem corrigidas para médias móveis, o comportamento - mesmo que irreal - seria sempre crescente, o que seria compatível com os modelos implementados. Além disso, outros modelos mais complexos de serem trabalhados, como o modelo logístico, também poderiam ser usados como forma de prever a população brasileira nos anos seguintes.

Por último, gostaria de dizer que o objetivo do trabalho foi cumprido e acredito que teci as críticas relevantes à melhoria dos modelos por mim feitos, bem como à importância de se estudar e estimar a população brasileira.