

2.2 - presc

17 de julho de 2024



Analise os dados e limpe as colunas da seguinte forma:
periodo_ingresso: de 2006.1 até 2019.2
tipo_matricula: remover "Dispensa"
situacao: remover "Em Curso"
o novo dataframe após a limpeza deve se chamar df_filtrado



Vou começar lendo o arquivo que você enviou para entender sua estrutura e em seguida farei as limpezas necessárias nas colunas conforme solicitado. Vamos lá!

Aqui estão algumas das colunas que encontramos no seu dataset:

- `periodo_ingresso`
- `tipo_matricula`
- `situacao`

Agora, vou prosseguir com as limpezas solicitadas:

1. Manter apenas os registros com `periodo_ingresso` entre 2006.1 e 2019.2.
2. Remover os registros onde `tipo_matricula` é "Dispensa".
3. Remover os registros onde `situacao` é "Em Curso".

Vamos aplicar esses filtros agora. [→]

A limpeza dos dados foi realizada com sucesso. O novo DataFrame `df_filtrado` contém apenas os registros que atendem aos critérios especificados:

- Período de ingresso entre 2006.1 e 2019.2.
- Exclusão dos registros com tipo de matrícula "Dispensa".
- Exclusão dos registros com situação "Em Curso".

Se precisar de mais alguma manipulação ou análise desses dados, é só me avisar! [→]

Qual é a previsão de taxa de graduação para o próximo ano baseada em tendências passadas? Defina 3 opções de como solucionar essa questão e siga a melhor



Para prever a taxa de graduação para o próximo ano baseada em tendências passadas, podemos considerar algumas abordagens distintas:

1. Análise de Séries Temporais:

- Utilizar técnicas de previsão em séries temporais, como ARIMA, Holt-Winters, ou modelos baseados em machine learning como redes neurais LSTM, para modelar e prever a taxa de graduação com base em dados históricos. Esta abordagem é apropriada se tivermos dados cronológicos consistentes sobre taxas de graduação ao longo dos anos.

2. Modelagem Estatística/Regressão:

- Utilizar métodos de regressão para estimar a taxa de graduação com base em variáveis que possam influenciar esse resultado, como taxas de admissão, performance acadêmica, recursos educacionais, entre outros. Modelos de regressão linear ou regressão logística (se a taxa de graduação for categorizada) podem ser aplicados se tivermos dados adequados.

3. Análise Descritiva com Projeção Linear:

- Calcular a taxa de graduação média dos últimos anos e projetar esta média para o próximo ano. Este método é mais simples e menos preciso, mas pode ser útil se a quantidade de dados for limitada ou se não houver muita variação ano a ano.

Melhor Opção

Para escolher a melhor abordagem, precisamos considerar a disponibilidade e a qualidade dos dados. Se tivermos dados históricos detalhados sobre as taxas de graduação ao longo dos anos (e possivelmente outras variáveis preditivas), a **Análise de Séries Temporais** pode ser a mais indicada, pois permite capturar tendências e padrões temporais de forma eficaz.

Se você tiver os dados de taxa de graduação por ano disponíveis, poderemos prosseguir com a análise de séries temporais para fazer uma previsão. Caso contrário, podemos discutir outras abordagens ou a necessidade de coletar mais dados. Como você gostaria de proceder?

Siga a melhor opção para solucionar esse problema



Para prosseguir com a análise de séries temporais e fazer a previsão da taxa de graduação para o próximo ano, precisamos primeiro verificar se temos os dados necessários. Especificamente, precisamos de um conjunto de dados que inclua:

- Datas ou períodos específicos (por exemplo, anos ou semestres).
- Taxas de graduação correspondentes a esses períodos.

Vamos verificar se esses dados estão disponíveis no DataFrame que temos, ou se há informações que possamos usar para construir essa série temporal. Vou analisar as colunas do `df_filtrado` para identificar se temos dados anuais ou semestrais de graduação.

Parece que temos informações sobre os períodos de ingresso e evasão, além da forma de saída dos alunos, que inclui status como "Graduado" e "Evadido". Para calcular a taxa de graduação por ano, podemos usar a seguinte abordagem:

1. Agregar os dados por ano de evasão (`periodo_evasao`), considerando apenas aqueles cuja forma de saída é "Graduado".
2. Calcular a taxa de graduação anual como a proporção de alunos graduados em relação ao total de alunos que deixaram a instituição (por graduação, evasão, entre outros) naquele ano.

Vou preparar esses dados para criar uma série temporal com as taxas de graduação por ano, que podemos usar para aplicar uma análise de séries temporais e fazer previsões. Vamos começar calculando essas taxas. [\[+\]](#)

Parece que houve um erro ao processar a coluna `periodo_evasao`, que pode ter valores inesperados ou formatados incorretamente. Vou revisar e corrigir o tratamento dessa coluna para garantir que possamos prosseguir corretamente com o cálculo das taxas de graduação. Vou verificar e limpar os dados dessa coluna. [\[+\]](#)

Os valores de `periodo_evasao` parecem estar majoritariamente corretos, com exceção do valor `0.0`, que provavelmente representa dados ausentes ou um marcador de não aplicabilidade. Vamos ajustar o tratamento para excluir esses valores e continuar com o cálculo da taxa de graduação. [\[+\]](#)

Parece que ocorreu um erro ao tentar renomear a coluna calculada, o que impediu a execução correta. Vou corrigir o problema no agrupamento e na criação das colunas para garantir que os dados sejam processados corretamente. Vou ajustar e calcular novamente a taxa de graduação. [\[+\]](#)

Aqui estão as taxas de graduação calculadas para cada ano, de 2006 a 2020. Notamos algumas inconsistências em 2006 e 2020, possivelmente devido a dados incompletos ou outros fatores.

Agora que temos uma série temporal da taxa de graduação, podemos usar um método de previsão em séries temporais para estimar a taxa de graduação para o próximo ano. Dada a natureza dos dados e a disponibilidade de uma série temporal clara, podemos aplicar um modelo simples como um modelo ARIMA, que é adequado para prever dados com dependências temporais e pode lidar com tendências não estacionárias.

Vamos ajustar um modelo ARIMA à série temporal das taxas de graduação para fazer a previsão para o próximo ano. Vou prosseguir com essa modelagem agora. [\[+\]](#)

O modelo ARIMA prevê uma taxa de graduação para o próximo ano (2021) de aproximadamente 51.14%, com um intervalo de confiança que varia de 4.14% a 98.14%. O intervalo de confiança é bastante amplo, refletindo uma certa incerteza na previsão, que pode ser atribuída à variação nas taxas de graduação ao longo dos anos e ao número relativamente pequeno de pontos de dados.

Esta previsão pode ser útil para planejamento e avaliação de políticas educacionais. Se precisar de mais análises ou ajustes no modelo, estou à disposição para ajudar! [\[+\]](#)

estes warnings influenciam no resultado do calculo?

```
/home/sandbox/.local/lib/python3.11/site-
packages/statsmodels/tsa/base/tsa_model.py:471: ValueWarning: An
unsupported index was provided and will be ignored when e.g.
forecasting.
    self._init_dates(dates, freq)
/home/sandbox/.local/lib/python3.11/site-
packages/statsmodels/tsa/base/tsa_model.py:471: ValueWarning: An
unsupported index was provided and will be ignored when e.g.
forecasting.
    self._init_dates(dates, freq)
/home/sandbox/.local/lib/python3.11/site-
packages/statsmodels/tsa/base/tsa_model.py:471: ValueWarning: An
unsupported index was provided and will be ignored when e.g.
forecasting.
    self._init_dates(dates, freq)
/home/sandbox/.local/lib/python3.11/site-
packages/statsmodels/tsa/base/tsa_model.py:834: ValueWarning: No
supported index is available. Prediction results will be given with an
integer index beginning at start.
return get_prediction_index
```



Os warnings indicam que o modelo ARIMA do `statsmodels` não conseguiu utilizar o índice do DataFrame como um índice de data, o que é comum quando os dados não estão em um formato de data e hora padrão ou quando a frequência dos dados não é especificada claramente. Neste caso, o índice foi baseado em anos inteiros (por exemplo, 2007, 2008), e o `statsmodels` pode ter esperado um formato de data mais específico.

Impacto dos Warnings:

- Não influenciam diretamente os cálculos internos do modelo ARIMA, que pode processar os dados numéricos eficazmente.
- Afetam a interpretação dos resultados de previsão porque a saída não é vinculada a um índice de data, mas sim a um índice inteiro baseado na posição dos dados. Isso significa que as previsões são retornadas com índices numéricos em vez de datas.

Se quisermos eliminar esses avisos e talvez obter uma saída mais clara com índices de data, podemos converter o índice em uma série de datas efetivas para uso no modelo ARIMA. Isso pode ajudar a tornar as previsões mais interpretáveis e garantir que a manipulação de datas seja mais intuitiva. Gostaria de ajustar isso para incluir índices de data na análise?

[Comece com Data Analyst](#)

[Denunciar conteúdo](#) | [Termos de uso](#) | [Política de privacidade](#) | [?](#)