Bootcamp: Ciência de Dados
Aluno(a): NILSON ANTÔNIO DE ASSUNÇÃO JÚNIOR

Relatório de Entrega do Desafio Final

Desenvolvimento da Solução do Desafio Final

1. Importação de bibliotecas e carregamento dos dados

Iniciei a análise importando as bibliotecas fundamentais como `pandas`, `matplotlib`, `seaborn`, `scikit-learn`, `statsmodels` e outras. Carreguei os dados a partir dos arquivos `consumo_energia_eletrica.csv` e `estado_regiao.csv`, e fiz uma verificação inicial com `.head()` e `.info()` para entender a estrutura e integridade dos dados.

2. Limpeza e padronização dos dados

Padronizei os nomes das colunas e integrei os dois conjuntos de dados utilizando a sigla do estado como chave de junção. Removi os dados dos tipos “Total” e “Cativo”, pois percebi que causavam sobreposição e poderiam distorcer a análise comparativa. Tratei inicialmente valores ausentes com interpolação temporal, mas optei por excluir esses registros quando percebi que não impactariam negativamente as análises por região e tipo de consumo.

3. Análise Exploratória Inicial

Realizei análises descritivas com agrupamentos por tipo de consumo, estado e região. Tentei utilizar a Lei de Benford para avaliar a distribuição natural dos dados, com a intenção de investigar possíveis padrões de normalidade. No entanto, percebi que ainda não domino totalmente a aplicação dessa técnica, e reconheço essa limitação como um aprendizado importante.

4. Visualizações e análise descritiva

Utilizei gráficos de barras, boxplots e séries temporais para explorar padrões e comportamentos:

- Identifiquei que as regiões Sudeste e Sul concentram os maiores volumes de consumo;

- Notei que o consumo residencial apresenta maior sazonalidade ao longo dos meses;

- Observei comportamentos distintos no consumo industrial e comercial, especialmente em momentos de instabilidade econômica.

5. Análise Temporal: Sazonalidade e Tendência

Agrupei os dados por ano e mês, e apliquei técnicas de decomposição de séries temporais usando PCA. Isso me ajudou a visualizar tendências regionais e padrões sazonais, como o crescimento gradual do consumo industrial no Nordeste.

6. Sessão Extra

Acrescentei análises complementares como gráficos de dispersão, correlação entre tipos de consumo, e comparações por UF e região. Essas análises foram úteis para identificar padrões de consumo semelhantes entre estados com características econômicas parecidas.

Conclusão

i. Aplicação dos Conhecimentos

Utilizei os conhecimentos adquiridos durante o bootcamp em todas as etapas do projeto, da preparação à análise crítica. Apliquei conceitos de manipulação e limpeza de dados, visualizações, estatísticas descritivas, normalização e análise temporal. Consegui montar um fluxo analítico completo com autonomia e clareza.

ii. Principais Dificuldades e Superações

Tive dificuldades para entender a sobreposição dos tipos de consumo “Cativo” e “Total”, mas consegui superar isso ao analisar a estrutura lógica dos dados. Também encontrei obstáculos na tentativa de aplicar técnicas estatísticas mais avançadas, como a Lei de Benford, e percebi que preciso aprofundar meus estudos nesse tema. Por fim, a junção dos dados geográficos exigiu cuidado com o tratamento das chaves de ligação e dos formatos.

iii. Resultados Obtidos

Consegui identificar padrões relevantes de consumo:

- O consumo residencial apresenta forte sazonalidade;
- O consumo industrial tem crescido em algumas regiões como o Nordeste;
- As regiões Sudeste e Sul concentram a maior parte da demanda energética;
- Foi possível observar correlações entre diferentes tipos de consumo em estados com perfis semelhantes.

Busquei também prever o consumo futuro com base nos dados históricos, utilizando modelos de regressão. Inicialmente testei uma regressão linear simples, mas percebi que os resíduos apresentavam padrões que comprometiam a precisão do modelo. Para melhorar os resultados, apliquei a transformação Box-Cox nos dados, o que ajudou a estabilizar a variância e aproximar a normalidade, melhorando o ajuste.

Além disso, testei o modelo de Random Forest Regressor, que permitiu capturar não linearidades e interações entre variáveis de forma mais flexível. Comparei os desempenhos entre os modelos com base em métricas como RMSE e R^2. A Random Forest apresentou desempenho superior na maioria dos testes, especialmente em cenários com múltiplos tipos de consumo e regiões.

Essa etapa reforçou meu entendimento sobre a importância de testar diferentes modelos e transformações estatísticas para obter previsões mais realistas e robustas.

Além das previsões, realizei uma tentativa de classificar perfis de consumo com base em agrupamento (clustering), utilizando o algoritmo KMeans. Para isso, normalizei os dados por tipo de consumo e região e explorei diferentes quantidades de clusters (`k`). A fim de definir o número ideal de grupos, utilizei a técnica do cotovelo (`elbow method`), que consiste em observar o ponto de inflexão da curva de inércia — onde o ganho de separação entre clusters deixa de ser significativo.

A análise permitiu distinguir grupos com perfis distintos de consumo, como estados com predominância residencial e outros com consumo majoritariamente industrial. Essa abordagem foi útil para entender a estrutura no comportamento energético em diferentes regiões.

iv. Lições Aprendidas

Aprendi a importância de validar hipóteses com métodos estatísticos adequados e compreendidos. Também entendi melhor o fluxo completo de uma análise de dados, desde a organização até a extração de conclusões. Me senti mais preparado para realizar análises estruturadas e comunicar os resultados com clareza.

v. Melhorias Futuras

Para aprimorar futuras versões desse trabalho, pretendo:

- Criar visualizações mais elaboradas, como gráficos geoespaciais interativos;
- Aprofundar minha análise de sazonalidade com o uso de técnicas mais robustas;
- Explorar ferramentas de previsão para o consumo futuro;
- Levar em consideração eventos externos como políticas públicas, crises econômicas e clima, que influenciam diretamente o consumo de energia.