VII Simpósio de Iniciação Científica 2024
Ciência, Tecnologia e Inovação para um Brasil Justo, Sustentável e Desenvolvido

CONSTRUÇÃO DE UMA FERRAMENTA DE ANÁLISE E VISUALIZAÇÃO DE DADOS COM A LINGUAGEM PYTHON

André Faria de Moraes[1] (IC), Alexandre de Ferreira Pinho (PQ)[1]

[1] *Universidade Federal de Itajubá.*

Palavras-chave: Inteligência de negócio. CRISP-DM. Python. Pandas.

Introdução

A pesquisa de Iniciação Científica em questão está inserida no contexto da Ciência de Dados e é fundamentada na construção de uma ferramenta de análise e visualização de dados a partir do Python, uma linguagem de programação orientada a objetos surgida no ano de 1991, a partir do programador holandês Guido van Rossum. O objetivo principal foi avaliar a viabilidade desta linguagem como instrumento de análise, compará-la com outras ferramentas de *business intelligence*, como o Excel e o Power BI e construir o algoritmo de análise e visualização que pode ser utilizado em diferentes bases de dados. É possível observar a relevância do trabalho em função do crescimento da importância dos dados dentro do contexto de organizações, que estão, progressivamente, passando a tomar decisões baseadas nos dados, e não somente na intuição dos gestores. Vercellis (2009) diz que o propósito das ferramentas de Sistemas de Informação é a aplicação de algoritmos matemáticos que gerem informações para otimizar o trabalho daqueles responsáveis pela decisão empresarial. Urge, portanto, a necessidade de ferramentas que transformem os dados brutos em informações úteis para o negócio, que possam otimizar o trabalho daqueles responsáveis pelos rumos do negócio e tornar o processo de tomada de decisão mais assertivo e confiável. A metodologia utilizada como base foi a CRISP-DM e, para o desenvolvimento, foram manipuladas duas bases principais: uma de casos de COVID-19 no estado de São Paulo e outra de falhas elétricas em cidades do interior do mesmo estado. Ambas as bases de dados passaram pelos processos de limpeza, tratamento e análise exploratória. A segunda, referente às ocorrências na rede elétrica, foi apresentada em um dashboard de visualização de negócios desenvolvido com a biblioteca Streamlit, criada pelos programadores Adrien Treuille, Thiago Teixeira e Amanda Kelly. Isso permitiu alcançar resultados consistentes e relevantes.

Metodologia

Como metodologia de pesquisa, foi utilizada a CRISP-DM (*CRoss Industry Standard Process for Data Mining)*, que consiste em um ciclo de relacionamento com dados, a fim de extrair valor deles, criada em 1996 por um consórcio de empresas formado pelas empresas NCR (*National Cash Register*), Daimler-Benz e OHRA(*Onderlinge ziektekostenverzekeringen van Hoogere Rijks Ambtenaren*). Apesar de ter surgido no contexto empresarial, essa metodologia pode ser aplicada a outros meios, como a saúde e a prestação de serviços, como foi comprovado no desenvolvimento do trabalho. As etapas do CRISP-DM são: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, validação e aplicação. Todas estas foram aplicadas em ambas as bases de dados desenvolvidas. Na fase de entendimento, foram compreendidas as necessidades de resultado do trabalho de manipulação de dados. A etapa de entendimento dos dados consistiu na compreensão do significado de cada variável e cada registro, para auxiliar na modelagem. A preparação foi a limpeza e o tratamento dos dados brutos, a exemplo da exclusão de dados faltantes, alteração de tipos de dados e criação de novas variáveis. A modelagem foi realizada a partir da construção das análises e do painel de visualização. A validação foi a análise dos resultados mostrados pelas ferramentas construídas. E, por fim, a etapa de aplicação foram as mudanças propostas para a resolução dos problemas identificados na fase anterior. Além do CRISP-DM, foram utilizadas, como método, consultas nas documentações oficiais do Python e das bibliotecas utilizadas.

Resultados e discussão

A análise exploratória de dados é um processo posterior

à limpeza e tratamento. No seu início, os dados já estão prontos para serem modelados, mas ainda estão na sua forma bruta, isto é, organizados em variáveis e registros, de forma semelhante à que são gerados. Após a aplicação das técnicas de análise, devem ser extraídas informações que resumam e auxiliem na compreensão dos dados em questão, como as medidas estatísticas, a distribuição de frequências, os *boxplots* e outras formas técnicas de exploração. A figura 1 demonstra, de maneira gráfica, o funcionamento de um trabalho de análise exploratória.

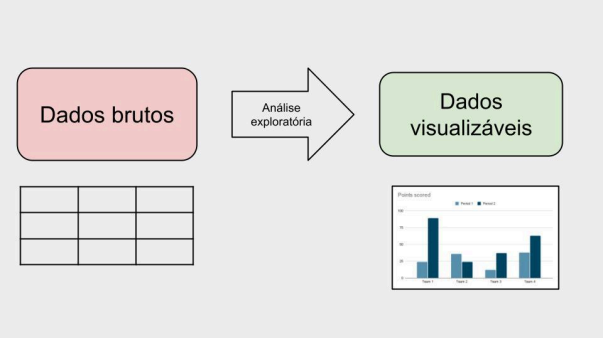

Figura 1: Representação da função da análise exploratória.
Fonte: O autor.

E, inserido nessas definições, é possível observar que a etapa de análises exploratórias foi capaz de expor informações palpáveis acerca das duas bases de dados.

A análise de dados de COVID-19 revelou uma correlação quase perfeita entre o número de casos e o número de óbitos na cidade de Campinas, além de uma forte correlação, de 0.8, entre os casos novos e óbitos novos nessa mesma cidade, como pode ser constatado pela figura 2, que expõe todas as correlações entre variáveis numéricas segundo o método de *Spearman*.

```
df_new = filtro_campinas[['casos', 'casos_novos', 'obitos', 'o

correlacoes = df_new.corr(method='spearman')
print(correlacoes)

import seaborn as sns

plt.figure()
sns.heatmap(correlacoes, annot = True)
```

	casos	casos_novos	obitos	obitos_novos
casos	1.000000	0.400064	0.999914	0.357675
casos_novos	0.400064	1.000000	0.400228	0.795481
obitos	0.999914	0.400228	1.000000	0.358285
obitos_novos	0.357675	0.795481	0.358285	1.000000

Figura 2: Correlações na cidade de Campinas.

Ademais, foi constatado que a variável "óbitos novos" seguia uma distribuição semelhante à normal em Campinas, com poucos outliers. A frequência dos casos foi mais concentrada nos primeiros meses do ano, se assemelhando a uma distribuição logarítmica, em grande parte devido aos avanços científicos, como a vacinação, que contribuíram para conter a propagação da doença.

```
hist1 = px.histogram(filtro_campinas_2021, x = 'casos_novos', nbins=30)
hist1.update_layout(width=400, height=400, title_text= 'Óbitos novos de Campinas em 2021')
hist1.show()

box_camp_casosnovos = px.box(filtro_campinas, y = 'casos_novos')
box_camp_casosnovos.show()
```

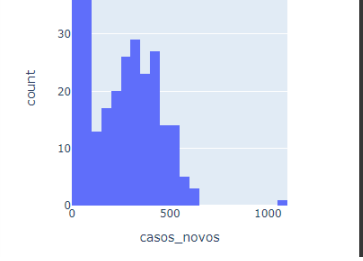

Figura 3: Distribuição dos óbitos em Campinas.
Fonte: O autor.

Na análise de ocorrências elétricas, houve uma correlação negativa entre as variáveis "consumidores afetados" e "duração da ocorrência", e a maioria das ocorrências teve duração próxima a 30 segundos. A Cidade 10 destacou-se com um número elevado de falhas e consumidores afetados, enquanto o ano de 2014 foi notável em termos de falhas, consumidores afetados e duração de ocorrências. As correlações podem ser observadas na figura 4.

```
import seaborn as sns

[59] df_new = filtro_cidade10[['Cons_afetados', 'Duracao[s]']].reset_index()
     df_new.drop(columns='index', inplace=True)

[60] correlacoes = df_new.corr(method='spearman')
     print(correlacoes)
```

	Cons_afetados	Duracao[s]
Cons_afetados	1.000000	-0.406238
Duracao[s]	-0.406238	1.000000

Figura 4: Correlação de dados de ocorrências elétricas.Fonte: O autor.

Já o *dashboard* de visualização apresentou as principais

características dos dados de falhas elétricas. Ele foi dividido entre três seções. A primeira diz respeito aos registros, que são as falhas. Ou seja, como as falhas foram distribuídas entre as variáveis categóricas e no decorrer do tempo. A segunda é uma análise focada individualmente em cada cidade. As cidades podem ser selecionadas a partir do filtro da barra lateral e há diferentes gráficos que expõem a quantidade total de consumidores afetados e média de duração de ocorrência distribuídos segundo classificação de duração, período do dia e causa de ocorrência. A terceira e última parte é semelhante à segunda. Contudo, a análise individual é referente aos anos. Além disso, é exposta uma tabela com as principais informações estatísticas de consumidores afetados e duração de ocorrência. Entre os principais resultados, observou-se um crescimento significativo de 92% nas falhas em 2014 em comparação com o ano anterior, ao passo que, entre 2012 e 2013 o crescimento foi quase insignificante. A maioria das falhas (80%) foi causada por desligamentos de equipamentos, sendo mais frequentes em dezembro nos três anos, possivelmente devido ao aumento do consumo de energia.

A Cidade 10 e a deterioração de equipamentos foram responsáveis pelo maior número de consumidores afetados. Houve uma tendência de que a maior parte das ocorrências tivesse duração inferior a 60 segundos, enquanto o período da noite foi associado às ocorrências de maior duração.

Análises focadas nas cidades destacaram as cidades 10, 15 e 24 como aquelas com mais consumidores afetados. As cidades 10 e 24 apresentaram padrões semelhantes, enquanto a cidade 15 apresentou características distintas. Observou-se a necessidade de melhorar o monitoramento das causas das ocorrências, uma vez que muitas falhas não foram devidamente registradas.

É válido ressaltar que os dados estão criptografados devido ao sigilo das informações que pertencem a uma empresa de energia elétrica.

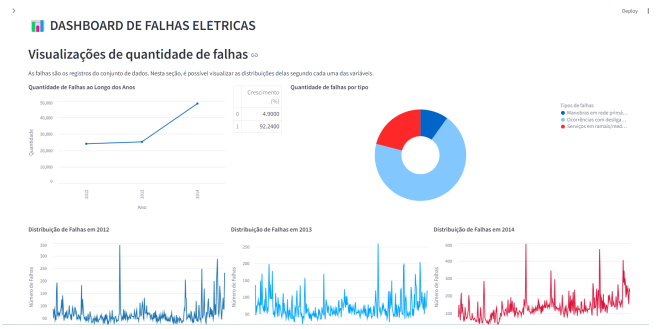

Figura 5: Visão geral do painel de visualização.
Fonte: O autor

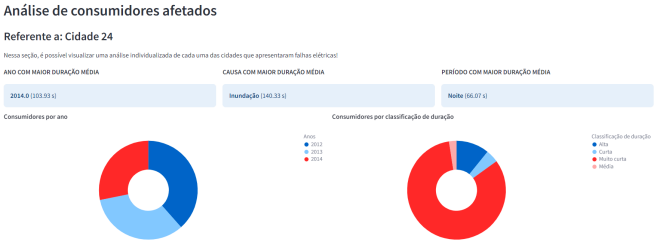

Figura 6: Visão da cidade 24 do painel de visualização.
Fonte: O autor.

Figura 7: Visão do ano 2014 do painel de visualização.
Fonte: O autor.

Conclusões

O trabalho teve como objetivo desenvolver uma ferramenta de análise exploratória e um painel de visualização de dados utilizando Python e suas bibliotecas especializadas. Duas bases de dados foram analisadas: uma sobre casos de COVID-19 no estado de São Paulo e outra sobre ocorrências de falhas elétricas em uma empresa. Através dessas análises, foi possível demonstrar a eficiência do Python na manipulação de grandes volumes de dados, posicionando-o como uma alternativa eficiente a métodos tradicionais, como Excel

e Power BI, especialmente no contexto de "Big Data".

Durante o desenvolvimento, foram aplicadas técnicas de limpeza, segmentação e análise exploratória. A biblioteca Pandas desempenhou um papel crucial na manipulação dos dados, enquanto a Streamlit permitiu a criação de um dashboard dinâmico que facilitou a visualização interativa dos resultados. A metodologia CRISP-DM foi utilizada para estruturar o processo de análise, guiando desde o entendimento do problema até a aplicação prática dos resultados.

Os resultados obtidos foram significativos, revelando correlações entre variáveis, a distribuição de ocorrências ao longo do tempo, e destacando padrões importantes, como os períodos do dia e os anos com mais falhas, além de identificar as principais causas dos problemas analisados.

A ferramenta de análise exploratória automática, utilizando o módulo Ydata Profiling, foi útil para análises rápidas e abrangentes, embora tenha limitações quanto à personalização e controle. Em contrapartida, a criação manual dos códigos ofereceu maior flexibilidade, permitindo um controle mais detalhado sobre as etapas de tratamento e análise dos dados.

Em conclusão, o trabalho atingiu seus objetivos ao criar uma solução prática e eficiente para a análise e visualização de dados, permitindo a transformação de dados brutos em informações acionáveis. O Python provou ser uma ferramenta poderosa, com grande potencial para apoiar a tomada de decisões estratégicas, sendo relevante em diversos contextos, como gestão pública e setor privado.

Agradecimentos

Eu agradeço ao professor Alexandre Pinho, que me auxiliou durante o desenvolvimento do trabalho, fornecendo diretrizes para os rumos da pesquisa, à UNIFEI e à CNPq, que me deram condições de fazer uma pesquisa fundamental para a minha formação profissional e forneceram o auxílio financeiro necessário.

Referências

LAUDON, Kenneth C.; LAUDON, Jane P. **Sistemas de informação gerenciais**. 11. ed. São Paulo: Pearson Education do Brasil, 2014.

VANDERPLAS, Jake. **Python data science handbook: essential tools for working with data.** Sebastopol: O'Reilly Media, 2016.

GRUS, Joel. Da**ta science from scratch: first principles with Python**. 2. ed. Sebastopol: O'Reilly Media, 2019.

MATTOS, Antonio Carlos Marques. **Sistemas de informação**. 2. ed. São Paulo: Saraiva, 2017. Disponível em: https://books.google.com.br/books?hl=pt-BR&lr=&id=SYNnDwAAQBAJ&oi=fnd&pg=PT8&dq=mattos+2017&ots=aBN2bL6v24&sig=JT014-OMz_gsNbGKGgaG_9IZ7VU#v=onepage&q=mattos%202017&f=false. Acesso em: 12 jul. 2024.

RASCHKA, Sebastian; MIRJALILI, Vahid. **Python machine learning**. 3. ed. Birmingham: Packt Publishing, 2019.

PROVOST, Foster; FAWCETT, Tom. **Data science for business: what you need to know about data mining and data-analytic thinking**. 1. ed. Sebastopol: O'Reilly Media, 2013.

TURBAN, Efraim; SHARDA, Ramesh; DELEN, Dursun; KING, David. **Business intelligence: a managerial approach**. Upper Saddle River: Pearson Education, 2011.

WHEELAN, Charles**. Estatística: o que é, para que serve, como funciona**. 1. ed. Rio de Janeiro: Zahar, 2016.