Estudo de caso: transformação e análise de dados para ONG

Vitor L. Guimarães, Daniela Marques

[1]Instituto Federal de São Paulo – Campus Hortolândia (IFSP)
CEP 13183-250 –Hortolândia –SP –Brazil

vitor.luciano@aluno.ifsp.edu.br

Abstract. *One of the most used tools today in the business environment is the electronic spreadsheet, due to its practicality and organization, which make the user experience more pleasant. Even though the spreadsheet facilitates the work of several people, it remains a manual way of management, and limits and hinders a descriptive analysis of the data and possible decision making. This work presents an application that deals with an eletronic spreadsheet that is then used to create dashboards in Power BI.*

Resumo. *Uma das ferramentas mais utilizadas hoje no ambiente empresarial é a planilha eletrônica, por conta de sua praticidade e organização, que fazem a experiência do usuário ser mais agradável. Mesmo que a planilha facilite o trabalho de diversas pessoas, ela continua sendo uma maneira manual de gerir, e limita e dificulta uma análise descritiva dos dados e uma possível tomada de decisão. Este trabalho apresenta uma aplicação que trata uma planilha eletrônica que em seguida é utilizada para criação de dashboards em Power BI para facilitar a tomada de decisão pelos seus usuários.*

1. Introdução

Segundo [Araújo 2014], uma planilha eletrônica é um tipo de programa computacional que utiliza tabelas. Cada tabela é formada por uma grade, composta por linhas e colunas. O nome eletrônica deve-se ao fato de sua implementação ser por meio de programas de computador.

Dentro de pequenas organizações, muitos dados ainda são armazenados e controlados a partir de planilhas do Excel. Uma pesquisa realizada em 2019 pela empresa Capterra mostra que entre 10 empresas do ramo contábil, ao menos 6 usam planilhas eletrônicas para gestão de seus negócios [Rossi 2019]. Isso se dá por conta do baixo investimento dessas organizações com ferramentas de banco de dados e criação de *dashboard*[1] para analisar esses dados.

Um cliente conhecido do autor deste projeto trabalha em uma dessas organizações sem fins lucrativos, que controla presenças de participantes de reuniões através de planilhas. Com a grande quantidade de participantes e sendo somente uma pessoa a realizar este controle, o uso da planilha se torna um fator problemático, por dificultar a análise e organização dos dados, além de tomar muito tempo por se tratar de um processo manual. Pensando em como melhorar esses processos, este trabalho se propõe a automatizar o tratamento e transformação da planilha utilizada pelo cliente, e usá-la como base para criação de *dashboards* em Power BI.

[1]Um painel visual onde é possível analisar dados importantes para qualquer operação de uma empresa.

O objetivo deste projeto é realizar o estudo de caso utilizando dados reais de uma ONG, atuando no tratamento de dados utilizando Python e gerando representações gráficas na plataforma Microsoft Power BI, para facilitar as tomadas de decisões pelo cliente final.

Este artigo está organizado da seguinda forma: Seção 2 descreve a fundamentação teórica utilizada, Seção 3 apresenta os trabalhos correlatos, a Seção 4 descreve a metodologia utilizada no projeto, o desenvolvimento do trabalho é detalhado na Seção 5, a conclusão e trabalhos futuros são apresentados nas Seções 6 e 7.

2. Fundamentação Teórica

Nesta seção aborda-se os principais conceitos que fundamentam os aspectos técnicos e o embasamento utilizado para o processo de tratamento e análise de dados proposto.

2.1. Dados e seu valor

Atualmente ouve-se muito falar sobre dados, mas qual o real valor dos dados? Os dados são gerados no mundo computacional a todo momento, quando se faz buscas no Google, procuramos um vídeo engraçado no Youtube ou acesso a um anúncio daquela calça que despertou interesse.

Isoladamente os dados não apresentam nenhuma relevância. Segundo [Menegat et al. 2020], dados são construções humanas para tentar abstrair fenômenos complexos da realidade através de elementos quantificáveis, que por sua vez são passíveis de serem analisados.

A partir daí que o papel de um cientista e analista de dados se torna importante, para gerar um valor, uma riqueza, a partir das informações que os dados podem fornecer. Informações é um conjunto de dados, que associados, fazem sentido e geram um valor. As informações quando bem utilizadas, processadas e transformadas, oferecem conhecimentos importantes para tomadas de decisões, favorecem novas atitudes, comportamentos, resultados e novas perspectivas [Elias 2017].

Segundo [Fergunson 2012], em uma pesquisa realizada pela empresa McKinsey, somente 65% das organizações utilizam de maneira eficiente os dados obtidos, sendo que 4% usam todos os dados que coletam.

2.2. CRISP-DM

CRISP-DM (*Cross Industry Standard Process for Data Mining*), "Processo padrão inter-industriais para mineração de dados" em português, é uma metodologia de projeto utilizada com objetivo de desenvolver modelos a partir de uma análise de dados e informações para prever falhas e soluções [Roberto 2022].

Na documentação oficial de uso do CRISP-DM, desenvolvido pelas empresas criadoras da metodologia, *NCR Systems Engineering Copenhagen, DaimlerChrysler AG, SPSS Inc e OHRA Verzekeringen en Bank Groep B.V*, destaca que os passos da metodologia são o entendimento do negócio, entendimento dos dados, seguido da preparação dos mesmos, a criação de um modelo, avaliação deste modelo criado, e possível implementação deste. A Figura 1 mostra as etapas do processo e como essas etapas se relacionam.

Figura 1. Etapas do CRISP-DM. Fonte:[Chapman et al. 2000]

2.2.1. Entendimento do negócio

É a etapa inicial do projeto, cujo o foco é entender a necessidade do cliente e definir um objetivo para que alcance a solução com sucesso. Apesar de ser no início do projeto, essa fase pode se repetir de acordo com as avaliações feitas durante o processo de desenvolvimento.

2.2.2. Entendimento dos dados

Com os problemas e objetivos definidos, é necessário identificar como os dados serão analisados, coletados e tratados. Nessa etapa, o cliente é questionado se há algum banco de dados, quantas e quais as fontes de dados, qual a qualidade destes dados, etc. É nessa etapa que é gerado as primeiras ideias e hipóteses para alcançar o objetivo proposto da etapa anterior.

2.2.3. Preparação dos dados

Com os dados coletados, é hora de organizá-los de modo que mostrem o que devem mostrar, ou seja, de modo que formam-se informações. Nessa etapa, é realizada a seleção dos dados, limpeza e transformação. Quando se trata de tratamento dos dados, envolve-se maneiras de lidar com dados nulos, dados com diferentes formatos, fusão de dados, etc.

2.2.4. Modelagem

Nesta etapa, um modelo do projeto é criado para ser avaliado posteriormente. O tipo de modelagem a ser utilizada varia de acordo com a necessidade do negócio e os tipos de dados. Durante a criação do modelo, é comum voltar à primeira etapa para conferir os objetivos e encontrar novas possibilidades.

2.2.5. Avaliação

O objetivo dessa etapa é verificar se o modelo criado atende as necessidades e objetivos definidos na primeira etapa. A equipe de avaliação precisa olhar com uma visão mais real possível para o modelo. Para isso é preciso cuidado na seleção dos dados que serão utilizados nos testes. Caso a avaliação seja negativa e haja espaço para mudanças e melhorias, a equipe deve trabalhar em cima dessas mudanças de maneira que não fuja do escopo inicial.

2.2.6. Implementação

Esta etapa é iniciada caso o *feedback* da etapa anterior seja positivo. Caso o processo tenha sido feito corretamente, o modelo é colocado em produção, de modo a agregar valor para o negócio. Nesta etapa, a exposição do produto desenvolvido não é necessariamente realizada por um cientista de dados.

2.3. Transformação e Análise de dados

A transformação de dados é um processo de conversão de dados brutos para dados utilizáveis pelo sistema ou aplicativo de destino. Inclui várias atividades como 'transformar' os dados, filtrar com base nas regras de negócio da organização e unir diferentes campos para obter uma visão consolidada [Rehan 2020]. A análise de dados é esse processo de formação de sentido além dos dados. É um processo complexo que envolve dados pouco concretos e conceitos abstratos, e relações deles com raciocínio indutivo e dedutivo.

A análise tem como objetivo organizar e sumarizar os dados de tal forma que possibilitem o fornecimento de respostas ao problema proposto para investigação [Gil 1999].

2.4. Python

Python é uma linguagem de programação de alto nível popularmente usada nas áreas de análise de dados, *Machine Learning* e Inteligência Artificial. Segundo [Kriger 2022], Python foi criado com o objetivo de otimizar a leitura de códigos e estimular a produtividade de quem os cria, razão pela qual é considerada uma linguagem de fácil compreensão. Além da economia de tempo e melhora na eficiência em um projeto, Python também tem a vantagem de conter um grande número de bibliotecas, sendo elas nativas ou de terceiros, que auxilia no desenvolvimento do projeto.

2.5. Pandas

Pandas é uma biblioteca construída sobre a linguagem Python que providencia uma abordagem rápida e flexível em projetos das áreas de banco de dados, *webscraping*[2], *Machine Learning*, visualização e mineração de dados, etc. [Alura 2023] aponta que Pandas é bastante comparado à planilhas eletrônicas pela maneira de trabalharem e estruturar os dados, mas que a maior diferença é que enquanto as planilhas eletrônicas suporta até 1.048.576 linhas por 16.384 colunas, o Pandas contém uma limitação baseada na quantidade de memória disponível, então é possível ter uma grande variedade de linhas e colunas desde que a memória alocada não ultrapasse a quantidade disponível na máquina.

[2]Uma forma de mineração que permite a extração de dados de sites da web convertendo-os em informação estruturada para posterior análise.

2.6. Microsoft Power BI

O Microsoft Power BI se trata de um conjunto de serviços de *softwares* que, juntos, oferecem funções de transformações de suas fontes de dados não relacionadas em informações valiosas, com visual agradável e envolvente [Microsoft 2023].

Ao se trabalhar com dados dentro do Power BI e com o esquema estrela (Figura 2), que será usado no projeto, há a necessidade de uma separação entre dados FATOS e dados DIMENSÃO.

2.6.1. Esquema Estrela

O esquema estrela utilizado no Power BI é um modo de separar as tabelas de forma que as ligações entre elas se pareça com uma estrela. Para entender a ideia do esquema, é necessário entender como os dados são divididos.

- Dados Dimensões: são dados que servem como legendas de alguma informação, o significado bruto, por exemplo: Cliente de codigo 5 refere ao nome Vitor Luciano.
- Dados Fatos: são dados que reúnem as dimensões e criam um fato, um evento e uma relação entre essas dimensões, por exemplo: Cliente de codigo 5 participou do encontro de codigo 10 na data 05/10/2022, no horário 20:30, com 1 hora de duração.

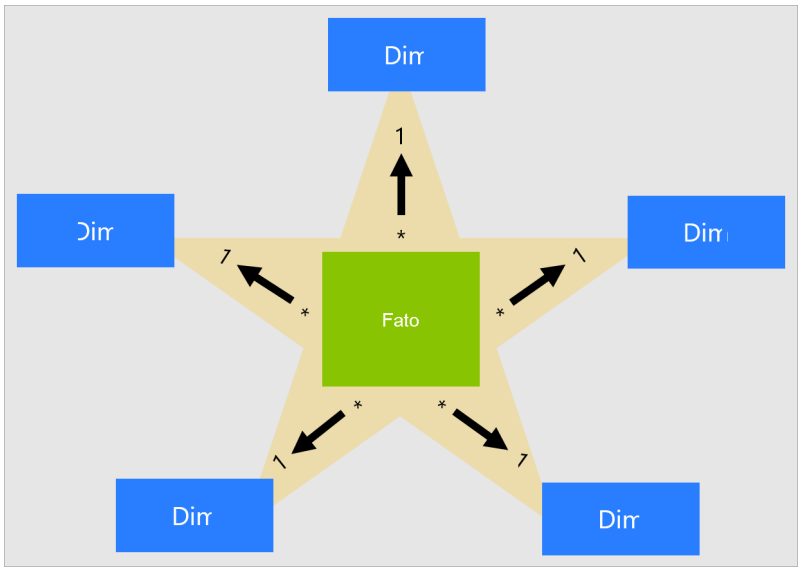


Figura 2. Design do esquema estrela. Fonte:[Microsoft 2023]

3. Trabalhos correlatos

Esta seção visa apresentar resumidamente os trabalhos correlatos a este projeto, encontrados em pesquisa através do Google Acadêmico e portal de Trabalhos de Conclusão de Curso do IFSP câmpus Hortolândia, no ano de 2023, utilizando as palavras-chave: Análise de dados, dados, Python e *dashboards*.

Como base do conhecimento e da ideia do projeto, o trabalho de conclusão de curso entitulado "Desenvolvimento de um processo de ETL[3] e uso de *dashboard* para visualização de casos de Covid", utiliza o processo de ETL para coletar, transformar e carregar uma grande massa de dados, e traz conceitos de ciência de dados e Big Data em seu projeto [Santos and Noda 2022].

Outro trabalho de conclusão de curso similar é "BigPy: um Sistema WEB para captura, tratamento e visualização de dados de defesa do consumidor utilizando Python". Neste trabalho, há uso da linguagem de programação Python como principal ferramenta de tratamento e visualização de dados, assim como criação de gráficos e *dashboards* para uma demonstração mais atraente das informações ao seu receptor. Tais ferramentas do projeto também serão utilizados neste trabalho, e por conta desta relação, será usado como auxílio para escolhas de gráficos em determinados tipos de dados [Ribas et al. 2022].

O projeto entitulado "Python na Análise de dados: Estudo de caso com dados de acidentes aéreos no Brasil", desenvolvido pelo aluno Philipe de Araújo Fernandes Formigoni junto ao orientador José Kimio Ando [Formigoni and Ando 2021], serviu como base para que a ideia do projeto se consolidasse e que o uso da metodologia CRISP-DM fosse inserida no escopo, além de também utilizar a linguagem de programação Python para lidar com os dados.

4. Metodologia

O desenvolvimento do projeto iniciou-se com uma reunião com usuários finais do projeto, para entender quais as informações valiosas a serem analisadas e quais possíveis decisões a serem tomadas após análise dos dados. Após isso, seguiu-se com a importação e tratamento dos dados contidos em uma planilha, sendo possível determinar quais dados serão desconsiderados no tratamento e quais dados serão relevantes o suficiente para serem mantidos.

Para esse projeto, estudou-se a linguagem de programação Python, com objetivo de conhecer as principais funções e comandos voltados para tratamento de dados e planilhas em Excel. Após os estudos, os conhecimentos foram aplicados de forma que a aplicação atuasse como responsável por tratar esses dados e transformá-los de acordo com a necessidade do cliente.

Com os dados tratados devidamente, houve a necessidade de uma validação desses dados por parte do cliente final. A validação tem o objetivo de confirmar se os dados finais estão realmente corretos e se são de fato relevantes para uma futura análise. Com os dados revisados e sua importância confirmada, a aplicação gerou uma nova planilha, dessa vez somente com os dados valiosos para análise, que foram utilizados para criar representações gráficas ao usuário. Nessa etapa, a ferramenta Microsoft Power BI foi estudada para aprender como realizar os vínculos entre os dados tratados e como representá-los graficamente.

No Microsoft Power BI foi configurado a fonte dos dados a serem importados e a definição dos dados FATO e DIMENSÃO. Com os dados DIMENSÃO e FATO já organizados, iniciou-se a última parte do projeto, que visava representar os dados em um

[3]Sigla para o processo de extrair, transformar e carregar dados. É uma forma tradicionalmente aceita para que as organizações combinem dados de vários sistemas em um único banco de dados.

dashboard de gráficos. Neste momento, foi analisado qual o gráfico mais adequado para determinado tipo de dado, para que a representação fosse feita de maneira adequada. Por exemplo: em casos de dados que envolvam data, é mais recomendado usar gráficos de série temporal; em casos de dados que envolvam características quantitativas, os gráficos de barras ou pirâmide são mais adequados.

Com o *dashboard* de gráficos já criado, o projeto estava pronto para entrega ao cliente final. O cliente terá acesso a esse *dashboards* por meio do Microsoft Power BI versão WEB e Mobile, sendo possível acessar de qualquer lugar e a qualquer momento. Após o projeto ser entregue, um formulário para o cliente final será enviado, com objetivo de coletar um *feedback* do produto, possíveis melhorias ou irregularidades.

5. Desenvolvimento

5.1. Entendimento do negócio

A primeira etapa do processo CRISP-DM é o entendimento do negócio. Para isso, uma reunião inicial com o cliente foi realizada para que o objetivo do projeto fosse determinado, os dados relevantes fossem definidos e para que uma planilha preenchida fosse disponibilizada para mapeamento e tratamento dos dados.

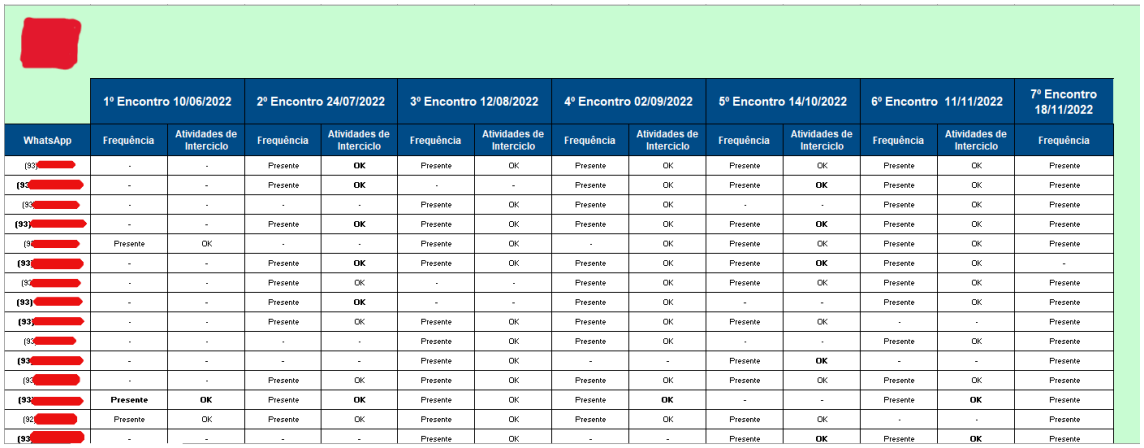
A ONG atua na capacitação de professores de escolas da rede pública, e promove encontros durante o ano para acompanhamento desses professores e realização de atividades. A planilha coletada se refere à uma lista de presenças do ano de 2022, contendo dados referentes à participantes (Figura 3), suas frequências e conclusão das atividades nos encontros realizados (Figura 4).

MONITORAMENTO DE FREQUÊNCIA DA FRENTE DE AVALIAÇÃO

Qtde	Frente de Formação	Município	Nome Completo	Nome da Escola	Cargo/Função Atual	WhatsApp
1	Avaliação das Aprendizagens	Juruti	Adson	EMEI		(93)
2	Avaliação das Aprendizagens	Juruti	**Ana**	**EMEI**	**Professora de 9º ano**	**(93) 9**
3	Avaliação das Aprendizagens	Juruti	Antonio	EMEI	Professor de 4º ano	(93) 9
4	Avaliação das Aprendizagens	Juruti	**Carlos**	**EMEI**		**(93) 9**
5	Avaliação das Aprendizagens	Juruti	Carlos	EMEIF	Professor de 5º ano	(93) 9
6	Avaliação das Aprendizagens	Juruti	**Climério**	**EMEI**		**(93) 9**
7	Avaliação das Aprendizagens	Juruti	Darlen	EMEI	Professora de 5º ano	(93) 9
8	Avaliação das Aprendizagens	Juruti	**Denilce**	**EMEI**	**Professor de 9º ano de Matemáti**	**(93) 99**
9	Avaliação das Aprendizagens	Juruti	Deuziane		Professora	**(93)**
10	Avaliação das Aprendizagens	Juruti	Erlon	EMEI	Professor de 8º ano de Ciências	(93) 9
11	Avaliação das Aprendizagens	Juruti	**Francisco**	**EMEI**	**Professor de 8º ano**	**(93) 9**
12	Avaliação das Aprendizagens	Juruti	Francisco	EMEI	Professor de 5º ano	(93) 9

Figura 3. Planilha original - primeira parte

Com a obtenção da planilha, foi iniciado a capacitação sobre a linguagem de programação Python e a biblioteca Pandas.



	1º Encontro 10/06/2022		2º Encontro 24/07/2022		3º Encontro 12/08/2022		4º Encontro 02/09/2022		5º Encontro 14/10/2022		6º Encontro 11/11/2022		7º Encontro 18/11/2022
WhatsApp	Frequência	Atividades de Intercíclo	Frequência	Atividades de Intercíclo	Frequência	Atividades de Intercíclo	Frequência	Atividades de Intercíclo	Frequência	Atividades de Intercíclo	Frequência	Atividades de Intercíclo	Frequência
(93)	-	-	Presente	**OK**	Presente	OK	Presente	OK	Presente	OK	Presente	OK	Presente
(93	-	-	Presente	**OK**	-	-	Presente	OK	Presente	**OK**	Presente	OK	Presente
(93	-	-	-	-	Presente	OK	Presente	OK	-	-	Presente	OK	Presente
(93)	-	-	Presente	**OK**	Presente	OK	Presente	OK	Presente	**OK**	Presente	OK	Presente
(9	Presente	OK	-	-	Presente	OK	-	OK	Presente	OK	Presente	OK	Presente
(93	-	-	Presente	**OK**	Presente	OK	Presente	OK	Presente	**OK**	Presente	OK	-
(93	-	-	Presente	OK	-	-	Presente	OK	Presente	OK	Presente	OK	Presente
(93)	-	-	Presente	**OK**	-	-	Presente	OK	-	-	Presente	OK	Presente
(93)	-	-	Presente	OK	Presente	OK	Presente	OK	Presente	OK	-	-	Presente
(93	-	-	-	-	Presente	OK	Presente	OK	-	-	Presente	OK	Presente
(93	-	-	-	-	Presente	OK	-	-	Presente	**OK**	-	-	Presente
(93	-	-	Presente	OK	Presente	OK	Presente	OK	Presente	OK	Presente	OK	Presente
(93	**Presente**	**OK**	Presente	**OK**	Presente	OK	Presente	**OK**	-	-	Presente	**OK**	Presente
(92	Presente	OK	Presente	OK	Presente	OK	Presente	OK	Presente	OK	-	-	Presente
(93	-	-	-	-	Presente	OK	-	-	Presente	**OK**	Presente	**OK**	Presente

Figura 4. Planilha original - segunda parte

5.2. Entendimento dos dados

Dando continuidade à metodologia adotada, o segundo passo do CRISP-DM é o entendimento dos dados. Nessa etapa, foram identificados os dados cadastrais de cada participante e encontro, e a maneira que se relacionam. A planilha contém uma estrutura que dificulta a análise posterior no Power BI: As linhas condizem com cada participante, e as colunas contém informações sobre os participantes e sobre os encontros, e em cada encontro contém duas colunas que se referem à atividade e à frequência. Todas as informações sobre o participante e sua relação com os encontros se encontram em uma única linha, o que não é recomendado quando se trata de organização dos dados.

5.3. Preparação dos dados

Inicia-se então a terceira etapa do CRISP-DM, a preparação dos dados. Essa etapa é o momento em que colocamos em prática o conhecimento aprendido da linguagem Python e da biblioteca Pandas. O processo de tratamento foi dividido em passos, para facilitar a manutenção do código e entendimento.

5.3.1. Importação da planilha

Inicialmente foi necessário importar a planilha no Jupyter Notebook, para trabalhar com os dados. A importação foi realizada já desconsiderando as 6 primeiras linhas da planilha, pois se tratava do cabeçalho e logos das empresas (Figura 5). Para cada passo realizado é gerada uma visualização da planilha Figura (6).

5.3.2. Separação das tabelas dimensões

Para separar os dados dimensão de participantes e encontros, foi necessário criar uma nova planilha para cada dimensão. Foi criado a tabela de participantes inicialmente inserindo as mesmas colunas que se encontram na planilha original, sendo "Nome completo", "Nome da Escola", "Cargo"e "Whatsapp", e para indexar cada participante foi criada uma coluna "ID"(Figura 7).

```python
import pandas as pd


planilha = pd.read_excel("Inscritos_Avaliação_Juruti_Formadores 3.xlsx")
planilha = planilha.drop([0, 1, 2, 3, 4, 5])
planilha = planilha.drop(columns=["Unnamed: 0","Unnamed: 1", "Unnamed: 2"])

display(planilha)
```

Figura 5. Importação da planilha no Python

	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unnamed: 12	Unnamed: 13	Unnamed: 14
6	NaN	NaN	NaN	NaN	NaN	1º Encontro 10/06/2022	NaN	2º Encontro 24/07/2022	NaN	3º Encontro 12/08/2022	NaN	4º Encontro 02/09/2022
7	Município	Nome Completo	Nome da Escola	Cargo/Função Atual	WhatsApp	Frequência	Atividades de Interciclo	Frequência	Atividades de Interciclo	Frequência	Atividades de Interciclo	Frequência
8	Juruti	Adson Luiz [redacted]	EMEI [redacted]	NaN	(93) [redacted]	-	-	Presente	OK	Presente	OK	Presente
9	Juruti	Ana Cintia [redacted]	EMEI [redacted]	Professora de 9º ano	(93) [redacted]	-	-	Presente	OK	-	-	Presente
10	Juruti	Antonio [redacted]	EMEI [redacted]	Professor de 4º ano	(93) [redacted]	-	-	-	-	Presente	OK	Presente

Figura 6. Visualização da planilha importada

	A	B	C	D	E
1	ID	Nome completo	Nome da Escola	Cargo	Whatsapp
2	1	Adson [redacted]	EMEI [redacted]		(93) [redacted]
3	2	Ana [redacted]	EMEI [redacted]	Professora de 9º ano	(93) [redacted]
4	3	Antonio [redacted]	EMEI [redacted]	Professor de 4º ano	(93) [redacted]
5	4	Carlos [redacted]	EMEI [redacted]		(93) [redacted]
6	5	Carlos [redacted]	EMEIF [redacted]	Professor de 5º ano	(93) [redacted]
7	6	Climério [redacted]	EMEI [redacted]		(93) [redacted]
8	7	Darlen [redacted]	EMEI [redacted]	Professora de 5º ano	(93) [redacted]
9	8	Denilce [redacted]	EMEI [redacted]	Professor de 9º ano de Matemática	(93) [redacted]
10	9	Deuziane [redacted]		Professora	(93) [redacted]
11	10	Erlon [redacted]	EMEI [redacted]	Professor de 8º ano de Ciências	(93) [redacted]
12	11	Francisco [redacted]	EMEI [redacted]	Professor de 8º ano	(93) [redacted]

Figura 7. Tabela dimensão de participantes.

Partindo para os dados dimensão de encontros, foram utilizados somente os valores da linha 1 da planilha original, que condizia com os dados de cada encontro. Após passar esses valores para uma nova planilha, foi verificado que a data do encontro estava junto ao nome do encontro, portanto foi preciso separar o campo de texto de cada célula, sendo uma parte movida para "Encontro"e outra para "Data". Além deste tratamento, assim como feito com participantes, também foi necessário indexar cada encontro criando um "ID"para eles. A planilha resultante é demonstrada na Figura 8.

5.3.3. Substituição de textos para valores inteiros

Para documentar se o participante esteve frequente no encontro, a coluna correspondente é preenchida com a palavra "Presente". O mesmo caso vale para a conclusão da atividade,

	A	B	C
1	ID	Encontro	Data
2	1	1º Encontro	10/06/2022
3	2	2º Encontro	24/07/2022
4	3	3º Encontro	12/08/2022
5	4	4º Encontro	02/09/2022
6	5	5º Encontro	14/10/2022
7	6	6º Encontro	11/11/2022
8	7	7º Encontro	18/11/2022

Figura 8. Imagem da tabela dimensão de encontros.

cujo a coluna é preenchida com "OK". Caso o participante não contemple nenhuma das duas condições, as colunas são preenchidas com hífen.

Para facilitar as futuras análises e possibilitar a criação de visões no *dashboard*, esses dados foram substituídos por 1 ou 0, sendo 1 para valores de "Presente"ou "OK", e 0 para hífens, como observado na Figura 9, nas linhas 3, 4 e 5.

```
fato = planilha
fato = fato.replace('Presente', '1')
fato = fato.replace('OK', '1')
fato = fato.replace('-', '0')

fato = fato.drop(columns=["Unnamed: 3","Unnamed: 5","Unnamed: 6","Unnamed: 7"])
fato = fato.drop([46,47,48])

display(fato)
```

Figura 9. Tratamento dos dados incompatíveis no Python.

Ao aplicar as alterações mostradas na Figura 9, gera-se uma planilha resultante com valores adequados para uma futura análise (Figura 10).

	Unnamed: 4	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unnamed: 12	Unnamed: 13	Unnamed: 14	Unnamed: 15	Unnamed: 16	Unnamed: 17	Unnamed: 18	Unnamed: 19	Unn
6	NaN	1° Encontro 10/06/2022	NaN	2° Encontro 24/07/2022	NaN	3° Encontro 12/08/2022	NaN	4° Encontro 02/09/2022	NaN	5° Encontro 14/10/2022	NaN	6° Encontro 11/11/2022	NaN	En 18/1
7	Nome Completo	Frequência	Atividades de Interciclo	Frequência	Atividades de Interciclo	Frequência	Atividades de Interciclo	Frequência	Atividades de Interciclo	Frequência	Atividades de Interciclo	Frequência	Atividades de Interciclo	Freq
8	Adson Luiz [redacted]	0	0	1	1	1	1	1	1	1	1	1	1	
9	Ana Cintia [redacted]	0	0	1	1	0	0	1	1	1	1	1	1	
10	Antonio [redacted]	0	0	0	0	1	1	1	1	0	0	1	1	

Figura 10. Planilha resultante com dados substituídos.

5.3.4. Criação da tabela fato

Com as novas planilhas de dados dimensões de participantes e encontros, é necessário criar a terceira e última planilha, que relaciona as duas já criadas. A nova planilha necessita ter a mesma estrutura de um banco de dados relacional, nesse caso sendo formada por

somente 4 colunas, sendo chaves estrangeiras de encontro e participante, "Frequência"e "Atividade".

Para iniciar a criação da tabela fato, foi utilizado a planilha original já com os tratamentos citados acima. O primeiro passo foi inserir o ID dos participantes e dos encontros, portanto foram criadas as colunas "ID.Encontro"e "ID.Participante"e utilizado um laço de repetição FOR para preencher as células de acordo com a quantidade de encontros e participantes (Figura 11).

```
novas_linhas = []
colunas_alvo= ['ID_Participante', 'Unnamed: 4','Unnamed: 8','Unnamed: 9', 'Unnamed: 10', 'Unnamed: 11','Unnamed: 12', 'Unnamed: 1



for indice, linha in fato.iloc[2:].iterrows():
    valores_combinados = [linha[coluna] for coluna in colunas_alvo]
    for _ in range(7):
        nova_linha = {coluna: valor for coluna, valor in zip(colunas_alvo, valores_combinados)}
        novas_linhas.append(nova_linha)


novo_fato = pd.DataFrame(novas_linhas)
#apagando a coluna com nomes
novo_fato = novo_fato.drop(columns=['Unnamed: 4'])
#inserindo coluna para os encontros
novo_fato['ID_Encontro'] = [(num % 7) + 1 for num in range(266)]
novo_fato = novo_fato[['ID_Encontro'] + [col for col in novo_fato.columns if col != 'ID_Encontro']]

display(novo_fato)
```

	ID_Encontro	ID_Participante	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unnamed: 12	Unnamed: 13	Unnamed: 14	Unnamed: 15	Unnamed: 16	Unnamed: 17	Unname
0	1	1	0	0	1	1	1	1	1	1	1	1	
1	2	1	0	0	1	1	1	1	1	1	1	1	
2	3	1	0	0	1	1	1	1	1	1	1	1	
3	4	1	0	0	1	1	1	1	1	1	1	1	
4	5	1	0	0	1	1	1	1	1	1	1	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	
261	3	38	1	1	1	1	1	1	1	1	1	1	
262	4	38	1	1	1	1	1	1	1	1	1	1	
263	5	38	1	1	1	1	1	1	1	1	1	1	
264	6	38	1	1	1	1	1	1	1	1	1	1	

Figura 11. Tabela fato com IDs dos participantes e encontros.

Com os IDs inseridos, foram criadas as duas colunas faltantes de "Frequência"e "Atividade". Para o preenchimento das duas novas colunas, foram utilizados laços de repetições FOR aninhados para percorrer as colunas e linhas da planilha original e mover os dados de cada iteração para as colunas adequadas da planilha fato (Figura 12).

```
num_colunas_fato = len(fato.columns)
linha_novo_fato = 0


for indice_linha, linha in fato.iterrows():
    # Dentro do loop de linhas, itere sobre as colunas
    for indice_coluna, (nome_coluna, valor_coluna) in enumerate(linha.iteritems()):

        if indice_coluna % 2 == 0:
            novo_fato.at[linha_novo_fato, 'Frequencia'] = valor_coluna
        else:
            novo_fato.at[linha_novo_fato, 'Atividade'] = valor_coluna
            linha_novo_fato = linha_novo_fato + 1



novo_fato.to_excel(writer, sheet_name='Fato', index=False)
writer.close()
```

Figura 12. Código para ajuste das frequências e atividades para os respectivos participantes e encontros.

Com isso, a planilha fato é gerada corretamente com dados binários relacionando encontros e participantes, com seu resultado mostrado na Figura 13.

	A	B	C	D	E
1	ID_Encontro	ID_Participante	Frequencia	Atividade	
2	1	1	0	0	
3	2	1	1	1	
4	3	1	1	1	
5	4	1	1	1	
6	5	1	1	1	
7	6	1	1	1	
8	7	1	1	0	
9	1	2	0	0	
10	2	2	1	1	
11	3	2	0	0	
12	4	2	1	1	
13	5	2	1	1	
14	6	2	1	1	
15	7	2	1	0	
16	1	3	0	0	
17	2	3	0	0	
18	3	3	1	1	

Figura 13. Tabela fato relacionando participantes e encontros.

Ao fim deste passo, as 3 planilhas se tornaram aptas para serem importadas no Microsoft Power BI e realizar a criação da modelagem do *dashboard*.

5.4. Modelagem

Assim como no tratamento dos dados em Python, o primeiro passo da modelagem é realizar a importação dos dados. A plataforma Microsoft Power BI oferece diferentes tipos de fontes de dados para importar, e para o projeto é selecionado "Pasta de trabalho do Excel". Ao importar as planilhas, o relacionamento automático entre elas não é feita de forma adequada pela ferramenta, portanto foi necessário gerenciar as relações de forma manual. A configuração foi realizada de maneira que as tabelas ENCONTROS e PARTICIPANTES se relacionem com a tabela FATO através de seus IDs. Além disso, foi necessário especificar que os registros de Encontros e Participantes ocorrem somente 1 vez nas tabelas de origem, mas múltiplas vezes na tabela FATO, formando um relacionamento de 1 para N (múltiplos), como mostrado na Figura 14.

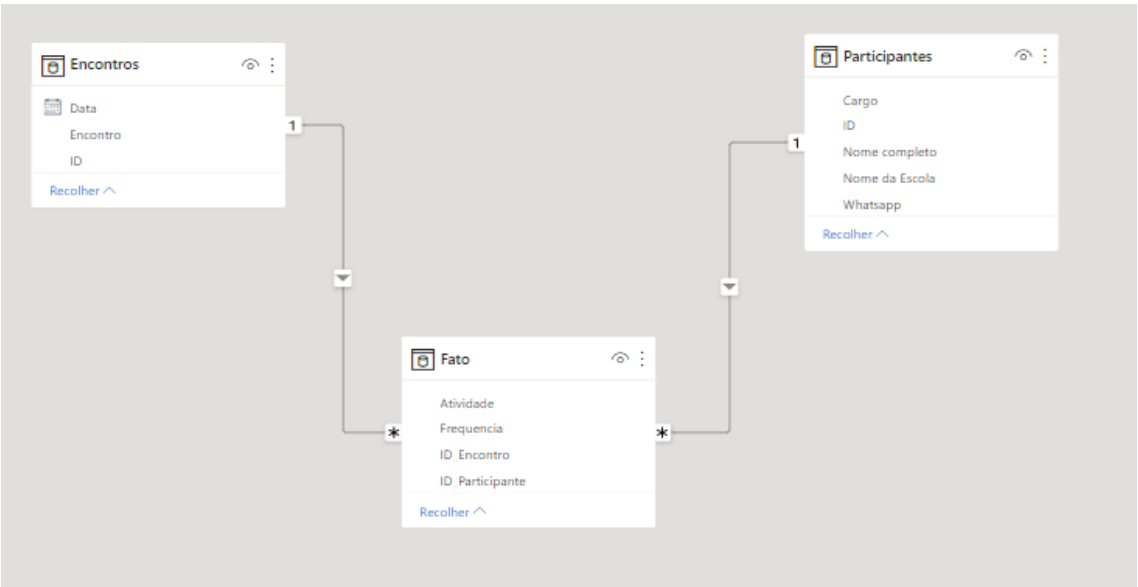


Figura 14. Relacionamento entre as tabelas dimensões e fato.

Com o relacionamento adequado entre as tabelas, a criação dos gráficos se tornou possível, assim foram escolhidos 4 tipos de gráficos para formar o *dashboard*: gráfico de setores (pizza), gráfico de linhas, Treemap[4] e tabela de dados. É importante ressaltar que os gráficos foram escolhidos de acordo com os tipos de dados a serem trabalhados, para que a análise seja mais precisa e eficaz.

O gráfico de setores apresentado na Figura 15 foi utilizado para representar o total de frequências para cada encontro, portanto é utilizado Encontros como a legenda do gráfico, e a soma das frequências como valores de cada setor.

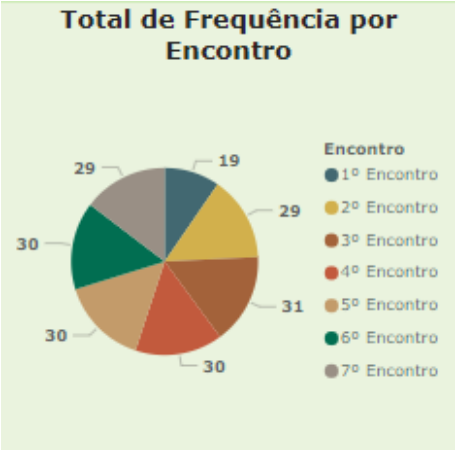


Figura 15. Gráfico de setores.

O gráfico de linhas foi utilizado para representar a quantidade de atividades realizadas a cada encontro. Com ele é possível analisar se houve comprometimento dos participantes em realizar as atividades dos encontros de acordo com o tempo. O gráfico completo é mostrado na Figura 16.

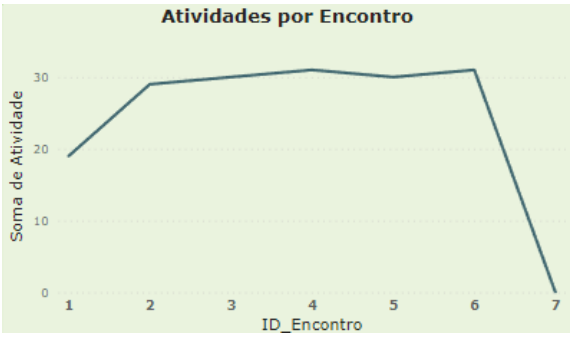


Figura 16. Gráfico de linhas.

O gráfico Treemap (Figura 17) contém uma técnica de visualização que consiste em retângulos aninhados que seu tamanho varia de acordo com o dado representado. No nosso projeto, o gráfico foi utilizado para representar o total de frequência dos encontros por dia da semana, portanto quanto maior o total de frequência do encontro, maior o retângulo. Com esse gráfico é possível análisar os dias da semana que tiveram menos frequência nos encontros, assim podendo tomar um plano de ação para tal fato.

[4]Uma técnica de visualização para representar dados hierárquicos usando retângulos aninhados.

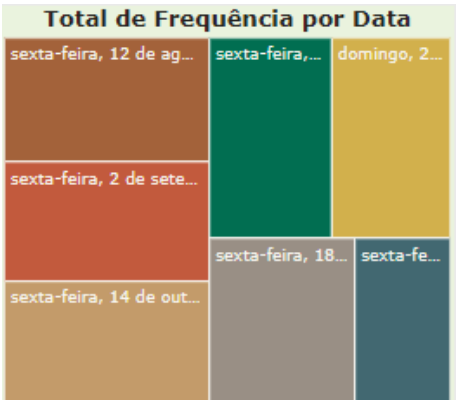


Figura 17. Treemap.

Para finalizar a criação dos gráficos, foram utilizados 2 com o modelo de tabela de dados. A primeira (Figura 18) mostra os dados de cada participante e a soma total de frequências e atividades realizadas nos encontros, sendo possível constatar qual participante teve menos frequência durante o ano, o que teve mais frequência, e o mesmo caso para as atividades feitas. A segunda tabela (Figura 19) monta uma relação entre os encontros e cada participante. Ela nos oferece uma visão de qual encontro o participante especificado frequentou e realizou as atividades ou não.

Nome completo	Whatsapp	Soma de Frequencia	Soma de Atividade
Wilderns [redacted]	(93) [redacted]	5	6
Vanderlan [redacted]	(93) [redacted]	5	4
Tânia [redacted]	(93) [redacted]	5	4
Sandrey [redacted]	(93) [redacted]	5	5
Rubia [redacted]	(93) [redacted]	6	5
Rozivaldo [redacted]	(93) [redacted]	5	4
[redacted]	[redacted]	7	6
Total		198	170

Figura 18. Soma de frequência e atividade por participante.

Encontro	Nome completo	Frequencia	Atividade
1º Encontro	Wilderns [redacted]	1	1
2º Encontro	Wilderns [redacted]	1	1
3º Encontro	Wilderns [redacted]	1	1
4º Encontro	Wilderns [redacted]	1	1
5º Encontro	Wilderns [redacted]	1	1
6º Encontro	Wilderns [redacted]	0	1
7º Encontro	Wilderns [redacted]	0	0
1º Encontro	Vanderlan [redacted]	0	0
2º Encontro	Vanderlan [redacted]	1	1

Figura 19. Relação de encontros com participante.

Após a criação e configuração dos gráficos, foi realizado uma estilização do painel de visualização, adicionando cor de fundo de acordo com a cor tradicional da ONG e

logo da ONG e parceiros. Com isso, o *dashboard* (Figura 20) foi finalizado e postado no ambiente de produção disponibilizado pela plataforma do Power BI.

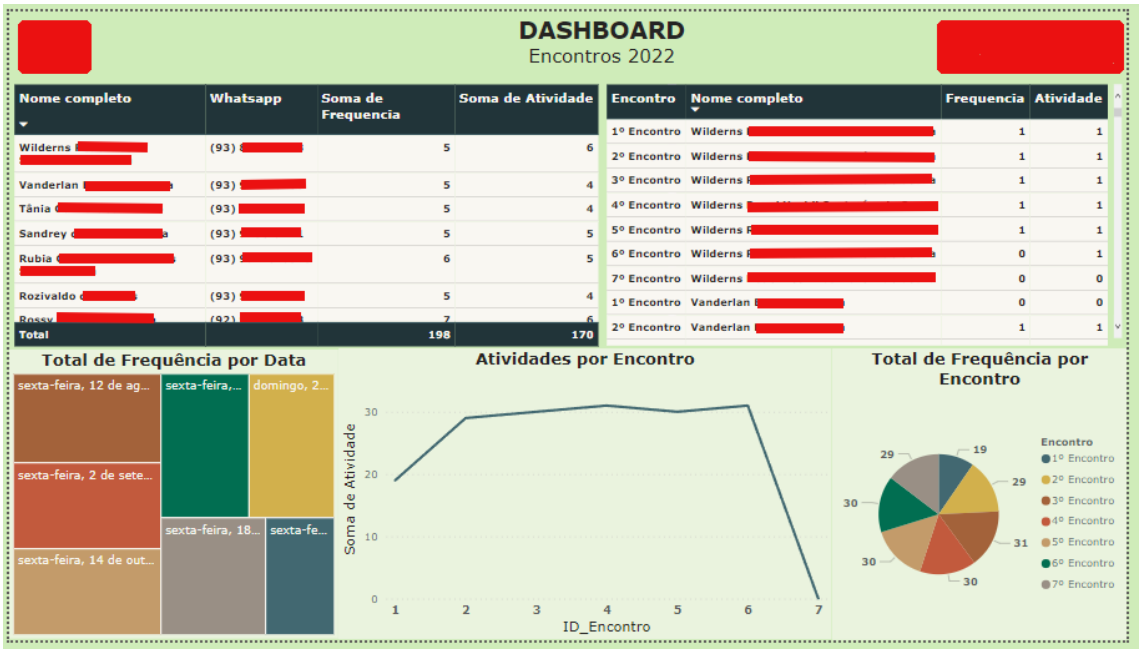


Figura 20. ***Dashboard***

6. Conclusão

Conclui-se com o projeto que obteve-se êxito em executar o tratamento de dados com Python e gerar as representações gráficas no estudo de caso. O processo de transformação e tratamento dos dados utilizando Python se mostrou eficaz no projeto para gerar dados possíveis de serem estruturados e analisados. Com a transformação dos dados em Python, foi possível utilizar a ferramenta Power BI e seus recursos para a criação do *dashboard*, e o mesmo gerou análises para o cliente, como número total de frequências e atividades de cada encontro e participante, dias da semana que tiveram maior número de frequência e até mesmo engajamento dos participantes ao passar do tempo, possibilitando tomadas de decisões como realizar encontros em dias mais convenientes, premiação para participantes com maiores frequências, etc.

Apesar de uma conclusão positiva, houveram desafios a serem enfrentados durante o desenvolvimento do projeto. A planilha original se encontrava em uma estrutura que dificultava a manipulação e reestruturação para um modelo de banco de dados relacional, e isso demandou mais tempo no passo de tratamento dos dados do que previsto inicialmente.

O passo 5 do CRISP-DM não pôde ser realizado pois durante o processo de desenvolvimento, o contato com o cliente final foi perdido e isso impossibilitou de receber um *feedback* final. Além dessas dificuldades, houve-se também desafios técnicos, e para superá-los foram de grande ajuda os conhecimentos absorvidos do curso de Análise e Desenvolvimento de Sistemas do Instituto, como Linguagem de Programação, Banco de Dados, Estatística e Metodologia Científica, além da ferramenta ChatGPT.

7. Trabalhos futuros

No tratamento da planilha utilizada no projeto, a quantidade de encontros e de participantes foram fixadas no código, assim, se outra planilha com quantidades diferentes de encontros ou participantes for importada no projeto, não será possível realizar o tratamento e transformação dos dados. Existe essa condições de melhoria que pode ser implementada como trabalho futuro, realizando a manutenção do tratamento de dados em Python de forma que permita aos usuários adicionar novos participantes e encontros, e dessa forma o *dashboard* permitirá analisar os dados do ano corrente.

Referências

Alura (2023). Pandas: o que é, para que serve e como instalar. https://www.alura.com.br/artigos/pandas-o-que-e-para-que-serve-como-instalar. [Online; Acessado em 02 de maio de 2023].

Araújo, I. S. (2014). *Planilhas eletrônicas*. NT Editora.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). *CRISP-DM 1.0 - Step-by-step data mining guide*. SPSS.

Elias, D. (2017). Dados vs informação: Qual a diferença? https://www.binapratica.com.br/dados-x-informacao. [Online; Acessado em 05 de abril de 2023].

Fergunson, R. (2012). Data analytics and the information transfer gap. https://sloanreview.mit.edu/article/data-analytics-and-the-information-transfer-gap/. [Online; Acessado em 04 de março de 2023].

Formigoni, P. and Ando, J. K. (2021). Python na anÁlise de dados: Estudo de caso com dados de acidentes aÉreos no brasil. https://app.uff.br/riuff/handle/1/23160. [Online; Acessado em 16 de novembro de 2023].

Gil, A. C. (1999). *Métodos e técnicas de pesquisa social.* Atlas.

Kriger, D. (2022). O que é python, para que serve e por que aprender? https://kenzie.com.br/blog/o-que-e-python/. [Online; Acessado em 02 de maio de 2023].

Menegat, R., Gehrke, M., da Costa, A. B. F., and Cubas, M. (2020). *Fluxo de trabalho com dados - Do zero à prática*. Open Knowledge Brasil.

Microsoft (2023). Entenda o esquema em estrela e a importância para o power bi. https://learn.microsoft.com/pt-br/power-bi/guidance/star-schema. [Online; Acessado em 06 de abril de 2023].

Rehan, A. (2020). O que é transformação de dados e como otimiza processos de negócios. https://www.astera.com/pt/type/blog/data-transformation-tools/. [Online; Acessado em 01 de março de 2023].

Ribas, W. A., Noda, E., and Marques, D. (2022). Bigpy: um sistema web para captura, tratamento e visualização de dados de defesa do consumidor utilizando python.

Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP) – Campus Hortolândia - São Paulo – SP – Brasil.

Roberto, C. (2022). Crisp-dm: as 6 etapas da metodologia do futuro. https://blog.mbauspesalq.com/2022/04/12/crisp-dm-as-6-etapas-da-metodologia-do-futuro/. [Online; Acessado em 15 de março de 2023].

Rossi, L. (2019). Contabilidade online: 61% das pmes brasileiras ainda usam excel. https://www.capterra.com.br/blog/1108/contabilidade-online. [Online; Acessado em 05 de abril 2023].

Santos, E. and Noda, E. (2022). Desenvolvimento de um processo de etl e uso de dashboard para visualização de casos de covid. *Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP) – Campus Hortolândia - São Paulo – SP – Brasil*.

Documento Digitalizado Público

Anexo I - Entrega do TCC final

Assunto: Anexo I - Entrega do TCC final
Assinado por: Daniela Marques
Tipo do Documento: Projeto
Situação: Finalizado
Nível de Acesso: Público
Tipo do Conferência: Documento Digital

Documento assinado eletronicamente por:

- **Daniela Marques**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 30/11/2023 15:20:57.

Este documento foi armazenado no SUAP em 30/11/2023. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifsp.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 1506369
Código de Autenticação: 53f2aa9345