**IMPORTÂNCIA DO TRATAMENTO DE INTERFERÊNCIAS NAS**

**BASES DE DADOS DO COVID E VACINAÇÃO.**

*IMPORTANCE OF TREATMENT OF INTERFERENCES IN COVID AND*

*VACCINATION DATABASE.*

**Helvio Rezende**

Graduação em sistemas de informação pelo Centro Universitário Una de Uberlândia-MG, servidor público municipal lotado na Secretaria Municipal de Saúde, com experiência nos sistemas, SIA, SIH, CNES.

email: [helviorezende.ti@gmail.com](mailto:helviorezende.ti@gmail.com),

**Orientador (a) Me: Josiane Araujo da Cunha**

e-mail: josiane.araujo@lais.huol.ufrn.br

**RESUMO**

A área da saúde detém grandes volumes de informações em seus bancos de dados, que contém além de informações pessoais dos pacientes, dados sobre os respectivos atendimentos, minerar esses dados para descobrir conhecimento não é tarefa trivial. É preciso conhecer os dados, o processo de análise e descoberta, as tarefas e técnicas de mineração e as ferramentas matemáticas e computacionais que se aplicam nesse contexto, esse texto pretende trazer um relato de experiência no estudo das bases de dados do Covid19 e Vacinação contra covid19, realizando um pré-processamento, identificando ruídos, sujeiras e valores ausentes contidos nessas bases de dados que podem atrapalhar a respectiva análise, a forma que esses ruídos e valores ausentes foram tratados, conforme embasamento teórico da revisão literária. Assim, por exemplo, na base do covid os nomes dos municípios foram digitados com acento, há presença de valores ausentes, por exemplo na coluna de em acompanhamentos novos, assim como a base de dados da vacina contém data de aplicação da vacina que indica erro na digitação.

PALAVRAS-CHAVE: Pré-processamento, filtro, valores ausentes,concatenação.

**ABSTRACT**

The health area holds large volumes of information in its databases, which contain, in addition to patients’personal information, data on the respective visits, mining this data to discover knowledge is not a trivial task. It is necessary to know the data, the analysis and discovery process, the mining tasks and techniques and the mathematical and computational tools tha apply in this context, this text intends to bring an experience report in study os the Covid19 and vaccination databases. Covid19, carrying out a pre-processing, identifying noises, dirt and missing values contained in these databases tha can interfere with the respective analysis, the way these noises and missing values were treated and the information extracted from the databases that helps so much to take of decidision, for example, in the covid database, the names of the municipalities were typed with an accent, there are missing values for example in the column of in new follow-ups, as well the vaccine database constains the date of vaccine application that indicates typing error.

**KEYWORDS:** Preprocessing, Filtering, Missing Values, Concatenation.

**1 INTRODUÇÃO**

Na área da saúde são encontrados diversos programas que alimentam extensas bases de dados, contendo as mais variadas informações, por exemplo, Sistema de Informação sobre Mortalidade (SIM), Sistemas de Informações sobre Nascidos Vivos (SINASC), Sistema de Informação Sobre Agravos Notificados (SINAN), dentre outras.

O Governo do estado de Minas Gerais, considerando a legislação e normas de utilização, dos usuários da saúde pública e aplicações desenvolvidas para registro e contabilidade dos casos notificados, padronizou esse registro através da construção de fichas de atendimento, contendo informações dos pacientes, informações de sua localização, dos sintomas e exames realizados. Tais fichas de notificação, normativas são encontradas no endereço eletrônico: http://vigilancia.saude.mg.gov.br/index.php/sistemas-de-informacao/agravos-de-notificacao-sinan/

Esse trabalho foi realizado utilizando as bases de dados, do covid19 extraida do site covid.saude.gov.br, e a base da vacina pelo site opendatasus.saude.gov.br, que respeitaram a Lei Geral de Proteção de Dados, por ser dado público, impossibilitando identificação de qualquer paciente registrado nessas bases.

O objetivo principal deste artigo científico é analisar base de dados da covid19 e da campanha de vacinação de enfrentamento à doença, verificando valores ausentes, ruídos, dados digitados incorretamente que possam dificultar a análise das informações e objetivo específico, após identificar os ruídos, valores ausentes é conseguir extrair informação dessas bases de dados.

2 REVISÃO LITERÁRIA

Diante de uma pesquisa bibliográfica em artigos e livros sobre o tema de Análise de dados, Mineração de dados, tem-se nesta seção, a definição das variáveis que serão estudadas, esclarecendo o significado de cada uma delas e como acontecem as relações dentro do tema.

McKinney, explica, que decorrer da análise e modelagem dos dados , 80% do tempo é gasto em sua preparação como carga, limpeza, transformação e reorganização, algumas bases de dados não constituem o formato correto para determinada tarefa, porém existem ferramentas de alto nível, rápido e flexível, para permitir manipulação dos dados e deixá-los no formato correto, explica McKinney(2018).

Silva, Peres e Boscarioli (2016), definem Mineração de Dados como uma área que existe devido a uma grande quantidade de dados disponível para ser explorada. Além da evolução tecnológica, evidenciada pelo aumento da capacidade de armazenamento decorrente da popularização dos gadgets, massificação do uso da internet e influência das redes sociais, contribuíram para o aumento na capacidade de gerar dados, não só pelos dados estruturados como os não estruturados.

No cenário de crescimento exponencial na quantidade de dados coletados e armazenados não há exclusividade apenas para a internet. Por causa do desenvolvimento tecnológico, tanto na qualidade quanto na quantidade de sensores geradores e monitores de dados, as empresas melhoraram sua capacidade de armazenamento e disponibilidade de dados. Além do tempo e esforço despendidos nas organizações para construção e manutenção de bases de dados, gerando especialidades como administradores e indexadores de bancos de dados, que não se tem ideia do que pode ser extraído. Cabe ressaltar que, frequentemente, os dados não podem ser analisados manualmente, em virtude de fatores como grande quantidade de registros, elevado número de atributos, valores ausentes, presença de dados qualitativos e não quantitativos, entre outros. Dessa forma, utiliza-se a mineração de dados como um processo sistemático, interativo e iterativo, de preparação e extração de conhecimentos a partir de grandes bases de dados (Ferrari, 2016).

A mineração de dados faz parte de um processo mais amplo, conhecido como descoberta de conhecimento em bases de dados (Knowledge Discovery in Databases, ou KDD)(AMARAL, 2016).

Segundo De Castro e Ferrari (2016), o processo de KDD consiste em quatro etapas, sendo:

* a primeira, a base de dados. Uma coleção organizada de dados, compostos por valores quantitativos ou qualitativos, referentes a um conjunto de itens que propiciam, eficientemente, a recuperação de dados.
* a segunda etapa é a preparação ou pré-processamento de dados, antecede a mineração, pois prepara os dados, constitui-se na remoção de ruídos e dados inconsistentes, com combinação de dados obtidos de várias fontes, ou seja, integração, seleção ou redução, na escolha dos dados relevantes à análise e à transformação ou consolidação dos dados em formatos apropriados à mineração;
* a terceira etapa, mineração de dados, corresponde à aplicação de algoritmos capazes de extrair conhecimentos, a partir dos dados pré-processados, através de técnicas de análise descritiva, agrupamento, predição, associação e detecção de anomalias; e,
* finalmente, avaliação ou validação do conhecimento, responsável por identificar conhecimentos verdadeiramente úteis e não triviais, através da avaliação dos resultados da mineração.

Contudo, tais etapas são correlacionadas e interdependentes, e consiste em considerar suas inter-relações, de forma ideal à abordagem para extrair informações relevantes em bancos de dados e sua influência no resultado, como ilustrado na Figura 2.

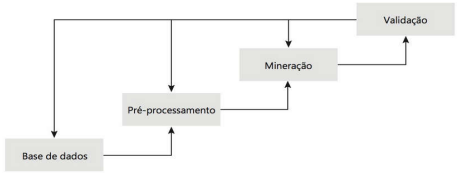


Figura 1 - Processo de Descoberta do conhecimento em bases de dados

Fonte: DE CASTRO e FERRARI, 2016

2.1 Entendimento e Preparação dos Dados

Para melhor entendimento dos dados e de sua análise é preciso compreender a forma de armazenamento desses.

Amaral (2016), explica que uma tarefa de banco de dados é composta por linhas e colunas. No aprendizado de máquina, as colunas são atributos e as linhas uma instância. Por outro lado, Silva e Peres e Boscarioli (2016), afirmam que a estatística descritiva é uma ferramenta capaz de descrever e resumir dados, mostrando aspectos importantes do conjunto de dados, por exemplo, tipo de distribuição associada e os valores mais representativos do conjunto. Além disso, faz parte das fases de descoberta do conhecimento a coleta de dados, o pré e o pós-processamento dos resultados, provenientes da mineração, o que resulta numa prévia preparação dos mesmos para que sejam numerados.

* Incompletude: podem ocorrer devido à falta de valores de um dado atributo. Entretanto, essa falta pode passar despercebida;
* inconsistência: são dados diferentes e conflitantes versões do mesmo dado. Aparecem em locais variados;
* Ruído: dependendo do contexto, em vídeo pode ser chuvisco na imagem; em rádio, interferência do sinal de áudio, mas em mineração de dados, aproxima-se do conceito de ruído em estatística e do processamento de sinais.

Diante disso é preciso fazer o pré-processamento dos dados. Para isso Castro e Ferrari (2016), definem as principais tarefas de pré-processamento como:

* Limpeza: para imputação de valores ausentes, correção de inconsistências e remoção de ruídos;
* Integração: unir várias fontes em único local, como data warehouse;
* Redução: reduzir a dimensão da base de dados, agrupando ou eliminando atributos redundantes, ou reduzir quantidade de objetivos da base;
* Transformação: para padronizar e deixar os dados em formato possível, mas as técnicas de mineração.

Segundo Castro e Ferrari (2016), limpeza dos dados é um problema que afeta a maior parte das bases de dados reais. As respectivas ferramentas atuam no sentido de imputar valores ausentes, suavizar ruídos, identificar valores discrepantes (outliers) e corrigir inconsistências. Nesse sentido, são apresentados os métodos tradicionais de imputação de valores ausentes:

* Ignorar objeto: remover todos os objetos que possuem um ou mais valores ausentes;
* Imputar manualmente os valores ausentes, ou seja, substituir os valores ausentes por outro valor;
* Utilizar uma constante global para imputar os valores ausentes: imputar um valor constante ao valor ausente;
* Imputação do tipo hot-ceck: imputar um valor de objeto similar a um objeto ausente;
* Imputar, de acordo com a última observação: ordenar a base e imputar valor ausente com valor da célula anterior;
* Usar a média pela média ou moda dos valores do atributo;
* Usar modelos preditivos para imputar valor ausente.

Ainda na fase de pré-processamento dos dados, Castro e Ferrari (2016), explicam que a concatenação de todos os dados se torna um dos passos essenciais antes de se realizar a mineração de dados. Entretanto, a integração desses dados pode resultar em vários problemas, como por exemplo, nas formas de armazenagem, convenções dos dados, datas, chaves de acesso, padronizações e outras características. Para os autores, o processo de integração abrange três aspectos:

* Redundância: quando um mesmo dado aparece em dois locais diferentes da base;
* Duplicidade;
* Conflitos: para uma mesma entidade, diferentes valores aparecem em diferentes fontes de dados.

Quando a base de dados para análise é imensa, a mineração consome um esforço computacional representado pelo espaço e tempo de processamento. Nesse momento, necessita-se da aplicação de técnicas de redução de dados, no intuito de reduzir a quantidade de objetos da base e da quantidade de atributos (CASTRO e FERRARI, 2016). A seguir, observa-se os métodos empregados nos casos de dimensionalidade:

Seleção de atributos (ou características): atributos irrelevantes, pouco relevantes ou redundantes detectados e removidos na redução de dimensionalidade;

Compressão de atributos: realizado através de algoritmos de codificação ou transformação de dados para a redução de dimensionalidade, em vez de seleção;

Redução no número de dados: dados são removidos, substituídos ou estimados por representações menores, modelos paramétricos e métodos não paramétricos, como agrupamento, amostragem e histogramas;

Discretização: os valores de atributos são substituídos por intervalos ou níveis conceituais mais elevados, reduzindo a quantidade final de atributos.

2.2 Tratando valores ausentes

Segundo McKinney(2018), dados ausentes são representados em objetos do pandas, para dados numéricos. Determinadas aplicações utilizam o valor de ponto flutuante NaN(Not a Number), chamado também de valor de sentinela, convenção usada na linguagem de programação R, referenciando os dados ausentes como NA, que significa Not Available(indisponível). Em aplicações estatísticas, dados NA podem ser dados inexistentes ou dados que existem, porém não observados.

Abaixo (Quadro 1), pode-se observar uma lista contendo algumas funções relacionadas ao tratamento de dados ausentes (MCKINNEY, 2018).

|  |  |
| --- | --- |
| Argumento | Descrição |
| dropna | Filtra rótulos de eixos, baseado no fato de os valores para cada rótulo terem dados ausentes, com limites variados para a quantidade de dados ausentes a ser tolerada. |
| fillna | Preenche os dados ausentes com algum valor ou utilizando um método de interpolação como ‘ffill' ou ‘bfill’. |
| isnull | Desenvolve valores booleanos informando quais valores estão ausentes/são NA. |
| notnull | Negação de isnull |

Quadro 1 - Tratamento para valores ausentes

Sob uma perspectiva de armazenagem de dados (data warehouse), o processo de mineração pode ser visto como um estágio avançado do processamento analítico on-line (On-line Analytical Processing – OLAP), conforme indica De Castro e Ferrari (2016).

Portanto, nessa etapa do trabalho, a revisão bibliográfica identificou as etapas de se trabalhar com análise de dados, iniciando-se com a obtenção da base de dados. As bases de dados relacionadas ao Covid 19 e à vacinação estão disponibilizadas em sites do Ministério da Saúde (covid.saude.gov.br - forneceu a base do covid e opendatasus.saude.gov.br - forneceu a base de vacinação contra covid19). O próprio site do covid apresenta também a descrição das informações fornecidas na base de dados e conceitos básicos para entendimento da mesma. também foi realizada uma análise das bases de dados, para identificar ruídos, sujeitas e valores ausentes, cumprindo assim as etapas, descritas na referida revisão bibliográfica, que se dispõe à análise de dados.

**3 METODOLOGIA**

A Base de dados do covid19 e a base de vacinação contra covid19, utilizadas nesse relato de experiência, públicas, de forma que qualquer pessoa consiga obtê-la, e obedecendo a Lei Geral de Proteção de Dados, as quais não disponibilizam informações dos pacientes, impossibilitando, respetiva identificação e na base de dados do covid avaliou a quantidade de óbitos, por alguns municípios, inclusive o Município de residência do autor desse trabalho, Uberlândia, e outros Municípios próximos,e fazer um comparativo dos casos de óbitos nesses municípios.

Primeira etapa foi conhecer as informações a respeito das bases de dados, as informações contidas no site covid.saude.gov.br e da vacinação contra covid, após isso obteve os arquivos, no formato csv, para análise, no caso do covid são 5 arquivos separados por semestre, no arquivo das vacinas, tem o arquivo completo contendo todos os estados, porém são arquivos muito grandes, de forma que se optou a trabalhar com arquivo de minas gerais.

Na Análise da base de dados, foi necessário concatenar os arquivos, pois no covid eram 5 arquivos e da vacina 3 e após concatenar, verificou-se quais dados estavam disponibilizados, tipo desses dados, a quantidade de colunas e linhas existentes nessas bases.

Identificou-se campo de data, o qual foi configurado, no formato de data e facilitar seleção das informações por ano ou mês, assim como a presença de valores ausentes, que poderiam ser substituídos ou excluídas linhas ou colunas e caracteres não reconhecidos.

Os Caracteres não reconhecidos, municípios digitados com acento, foram corrigidos, pois a localização dos mesmos ocorreu através da condição de identificar o código do IBGE, na base de dados, localizado na coluna codmun, do respetivo município e assim substituindo o nome com carácter não reconhecido pelo valor correto.

Os valores ausentes poderiam ser substituídos por valor de medias de outros valores existentes, somas, etc, porém ao selecionar determinada região, estado ou município, esses valores não apareceram na seleção.

Na extração dos resultados, após realização dos filtros selecionando região, ou estado, período e agrupou tais informações extraindo o total de óbitos por ano, por região e estado.

Finalmente na base de vacinação contra covid19, base muito grande, foi necessário escolher algumas colunas antes de trabalhar com a base de dados, houve a configuração do campo data de aplicação da vacina, para ficar no formato de data, verificou-se que os caracteres estavam em letra maiúscula e sem acento, não havia valores ausentes. Como a base de dados utilizada era do estado de minas gerais houve a seleção do município de Uberlândia, para obter os resultados.

**4 RESULTADOS E ANÁLISE DOS DADOS**

Na etapa do Pré-processamento, foi possível identificar os valores ausentes, sua localização e quantidade dos mesmos na base de dados do covid, conforme mostra quadro 3.

|  |  |
| --- | --- |
| NOME DAS COLUNAS | QUANTIDADE DE VALORES AUSENTES |
| Região | 0 |
| Estado | 1.094 |
| Município | 38.411 |
| Coduf | 0 |
| Codmun | 22.451 |
| Codregiaosaude | 38.411 |
| Nomeregiaosaude | 38.411 |
| Data | 0 |
| Semanaepi | 0 |
| Populaçãotcu2019 | 159.600 |
| Casosacumulados | 0 |
| Casosnovos | 0 |
| Obitosacumulado | 0 |
| Obitosnovos | 0 |
| Recuperadosnovos | 4.270.517 |
| Emacompanhamentonovos | 4.270.517 |
| interior/metropolitana | 38.411 |

Quadro 2 – quantidade de valores ausentes na base de dados covid.

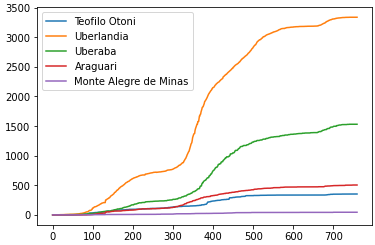
A configuração do formato das datas e correção dos caracteres não reconhecidos, gerados por nome dos municípios digitados com acento, permitiram obter a evolução dos óbitos por Município, conforme quadro 4.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ANO | MES | QUANTIDADE DE ÓBITOS POR MUNICÍPIO | | | | |
| Uberlandia | Araguari | Uberaba | Monte Alegre | Teofilo Otoni |
| 2020 | Abril | 8 | 1 | 3 | 0 | 0 |
| Maio | 17 | 1 | 6 | 0 | 6 |
| Junho | 78 | 5 | 26 | 3 | 31 |
| Julho | 204 | 30 | 55 | 5 | 49 |
| Agosto | 389 | 65 | 89 | 8 | 76 |
| Setembro | 561 | 78 | 141 | 8 | 90 |
| Outubro | 666 | 98 | 205 | 9 | 98 |
| Novembro | 712 | 106 | 227 | 9 | 109 |
| Dezembro | 737 | 114 | 239 | 13 | 117 |
| 2021 | Janeiro | 809 | 140 | 270 | 19 | 139 |
| Fevereiro | 1.048 | 211 | 345 | 19 | 148 |
| Março | 1.701 | 284 | 484 | 25 | 162 |
| Abril | 2.147 | 324 | 736 | 27 | 234 |
| Maio | 2.367 | 362 | 957 | 31 | 260 |
| Junho | 2.547 | 391 | 1.090 | 38 | 303 |
| Julho | 2.750 | 417 | 1.202 | 39 | 326 |
| Agosto | 2.983 | 443 | 1.265 | 41 | 331 |
| Setembro | 3.090 | 460 | 1.319 | 41 | 334 |
| Outubro | 3.161 | 470 | 1.349 | 42 | 336 |
| Novembro | 3.180 | 473 | 1.374 | 42 | 336 |
| Dezembro | 3.187 | 474 | 1.388 | 43 | 336 |
| 2022 | Janeiro | 3.233 | 476 | 1.429 | 43 | 336 |
| Fevereiro | 3.312 | 495 | 1.497 | 45 | 350 |
| Marco | 3.337 | 502 | 1.528 | 45 | 352 |
| abril | 3.338 | 505 | 1.531 | 45 | 353 |

Quadro 3 - Evolução dos óbitos por municípios.

Fonte Própria

Conforme o quadro 3, a figura 2, mostra a evolução dos óbitos nos Municípios de Uberlândia, Uberaba, Araguari, Teófilo Otoni e Monte Alegre de Minas.

Figura 2 – Evolução dos óbitos por município.

O Quadro 3 mostra a evolução dos óbitos por município, no caso, Araguari, Uberaba, Uberlândia, Monte Alegre de Minas e Teófilo Otoni, em todos eles aumento dos óbitos durante 2020, porém novembro e dezembro, apresenta uma queda e entre janeiro a maio uma queda, aumenta novamente, provavelmente período de festas, como natal, ano novo e carnaval, momento em que a população se aglomerou e provavelmente se descuidou quanto as medidas de proteção, em Uberlândia houve queda a partir de maio, acompanhada pelos outros municípios, devido o começo da vacinação, demonstrado também no quadro 4 que mostra a evolução em percentual, dos óbitos por mês.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ANO | MES | UBERLANDIA | ARAGUARI | UBERABA | MONTE  ALEGRE | TEÓFILO  OTONI |
| 2020 | Maio | 113 | 0 | 100 |  |  |
| Junho | 359 | 400 | 333 |  | 417 |
| Julho | 162 | 500 | 112 | 67 | 58 |
| Agosto | 91 | 117 | 62 | 60 | 55 |
| Setembro | 44 | 20 | 58 | 0 | 18 |
| Outubro | 19 | 26 | 45 | 13 | 9 |
| Novembro | 7 | 8 | 11 | 0 | 11 |
| Dezembro | 4 | 8 | 5 | 44 | 7 |
| 2021 | Janeiro | 10 | 23 | 13 | 46 | 19 |
| Fevereiro | 30 | 51 | 28 | 0 | 6 |
| Março | 62 | 35 | 40 | 32 | 9 |
| Abril | 26 | 14 | 52 | 8 | 44 |
| Maio | 10 | 12 | 30 | 15 | 11 |
| Junho | 8 | 8 | 14 | 23 | 17 |
| Julho | 8 | 7 | 10 | 3 | 8 |
| Agosto | 8 | 6 | 5 | 5 | 2 |
| Setembro | 4 | 4 | 4 | 0 | 1 |
| Outubro | 2 | 2 | 2 | 2 | 1 |
| Novembro | 1 | 1 | 2 | 0 | 0 |
| Dezembro | 0 | 0 | 1 | 2 | 0 |
| 2022 | Janeiro | 1 | 0 | 3 | 0 | 0 |
| Fevereiro | 2 | 4 | 5 | 5 | 4 |
| Marco | 1 | 1 | 2 | 0 | 1 |
| abril | 0 | 1 | 0 | 0 | 0 |

Quadro 4 Percentual dos óbitos por mês.

Fonte própria

No quadro 5 é possível visualizar aumento das doses de vacinação nos municípios principalmente durante 2021 que provavelmente contribuiu para diminuição dos casos de óbitos por covid.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ANO | MÊS | QUANTIDADE DE DOSES APLICADAS | | | | |
| UBERLÂNDIA | ARAGUARI | UBERABA | MONTE  ALEGRE | TEÓFILO  OTONI |
| 2021 | JANEIRO | 5.069 | 893 | 2.360 | 107 | 1.393 |
| FEVEREIRO | 21.473 | 3.790 | 11.106 | 466 | 5.016 |
| MARÇO | 39.397 | 12.704 | 28.727 | 2.560 | 12.793 |
| ABRIL | 88.423 | 13.248 | 32.287 | 2.484 | 16.613 |
| MAIO | 77.000 | 9.204 | 31.673 | 947 | 12.734 |
| JUNHO | 86.794 | 12.849 | 42.384 | 3.312 | 16.870 |
| JULHO | 158.260 | 26.899 | 78.804 | 4.551 | 29.217 |
| AGOSTO | 137.477 | 25.983 | 64.998 | 3.685 | 27.911 |
| SETEMBRO | 193.626 | 20.096 | 54.200 | 3.292 | 17.773 |
| OUTUBRO | 184.326 | 26.833 | 58.934 | 4.107 | 31.153 |
| NOVEMBRO | 112.601 | 12.601 | 57.972 | 3.810 | 18.022 |
| DEZEMBRO | 88.264 | 7.508 | 35.041 | 2.097 | 13.980 |
| 2022 | JANEIRO | 106.881 | 15.417 | 41.477 | 1.948 | 9.354 |
| FEVEREIRO | 95.987 | 11.715 | 33.594 | 2.384 | 11.194 |
| MARÇO | 65.819 | 9.723 | 31.806 | 2.571 | 6.168 |
| ABRIL | 44.718 | 6.247 | 23.536 | 1.819 | 4.605 |
| MAIO | 43.825 | 6.273 | 21.625 | 1.330 | 5.109 |
| JUNHO | 32.176 | 6.100 | 5.848 | 495 | 1.697 |
| TOTAL | | 1.582.116 | 228.083 | 656.372 | 41.965 | 241.602 |

Quadro 5 – distribuição de vacinas nos municípios

Fonte Própria

Portanto através desses resultados foi possível verificar aumento dos óbitos nos municípios, durante 2020, ainda por poucos casos mas aumentando em percentual alto até 2021 quando começou a vacinação, que se percebe uma queda no numero de óbitos, durante o período de junho a dezembro, porém com relaxamento das medidas prevenção, quanto sensação de proteção devido a vacina e aglomeração os casos aumentam no começo de 2022.

CONSIDERAÇÕES FINAIS

Esta pesquisa foi extremamente gratificante, no sentido de extrair informação de uma base de dados extensa, trabalho que não se realizaria por fórmulas avançadas do Excel na tentativa de quantificar alguns valores. informações obtidas rapidamente em detrimento da manipulação manual.

As referências teóricas, que nortearam esse relato de experiência, analisando as bases de dados do covid19 e da vacinação contra covid19, auxiliaram no entendimento da base de dados, os registros encontrados, identificação de ruídos, valores ausentes, caracteres não reconhecidos, datas incorretas.

Mas além da otimização de tempo, foi possível analisar a necessidade do tratamento e da padronização dos dados, ora digitados de forma incorreta, seja por excesso de informação ou pela falta dela. Talvez fosse pensar em uma padronização da digitação das informações, desta forma, evitar dados não reconhecidos pela tabela ASCII, ou dados digitados em locais errados.

A aplicação utilizada para manipulação das bases de dados nesse trabalho correspondeu bem, no tratamento dos dados, agilizando os cálculos solicitados e os resultados obitidos auxiliariam na tomada de decisão, num momento de apreensão diante da pandemia do Corona Vírus.

Esse relato também nos faz pensar, quantas vidas poderiam ser evitadas isolamento fosse tratado com mais seriedade pelos governantes, protegendo as pessoas mais carentes, o pequenos comerciantes, etc.

**REFERENCIA**

AMARAL, Fernando. **Aprenda Mineração de Dados: teoria e prática**. Alta Books, 2016

CARVALHO, Luís A. V. de. **DATAMINING: A Mineração de Dados no Marketing,Medicina, Economia, Engenharia e Administração.** 2a Edição. São Paulo-SP: Érica, 2002.

CANAL DO SANDECO. Analisando Covid19 no Brasil com Python e dados oficiais, disponível em

CANAL DO SANDECO. Analisando Covid19 no Brasil com Python e dados oficiais, disponível em [https://www.youtube.com/watch?v=30FijUpQ1po&t=1586s.](https://www.youtube.com/watch?v=30FijUpQ1po&t=1586s,) . Acesso em 23/12/2021.PROGRAME PYTHON. Python PANDAS - como filtrar dados com LOC e ILOC. Disponível em:<https://www.youtube.com/watch?v=30FijUpQ1po&t=1586s>. Acessado em 12/01/2022.

CÔRTES, Sergio da Costa; PORCARO, Rosa Maria; LIFSCHITZ, Sergio. Mineração de Dados - Funcionalidade, Técnicas e Abordagens. PUC-INFO, 2002. Disponível em ftp://obaluae.inf.puc-rio.br/pub/docs/techreports/02\_10\_cortes.pdf. Acesso em 26/05/2018.

De CASTRO, Leandro Nunes; FERRARI, Daniel Gomes. Introdução à Mineração de dados: conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

GREGORY, Guilherme; PRETTO, Fabrício. Mineração de dados para descoberta de conhecimento em dados de promoção à saúde. Revista Destaques Acadêmicos, Lajeado, v. 8, n. 4, 2016. ISSN 2176-3070. Disponível em http://univates.br/revistas/index.php/destaques/article/viewFile/1234/1091. Acesso em 19/04/2018.

MCKINNEY, WES. **Python para Análise de Dados; Tratamento de dados com pandas, numpy e python**. São Paulo: Novatec Editora Ltda. 2018.

OPENDATASUS. Campanha Nacional de Vacinação Contra Covid-19. Contexto O Ministério da Sáude, por meio do Sistema de informação ro Programa Nacional de Imunização. Disponível em:

PROGRAME PYTHON. Python PANDAS - como filtrar dados com LOC e ILOC. Disponível em:

PROGRAME PYTHON.Python PANDAS - Como excluir linhas e colunas com Nan. Disponível em:

SILVA, Leandro Augusta da; PERES, Sarajane Marques; BOSCARIOLLI, Clodis. **Introdução à Mineração de Dados. Com aplicações em r**. Rio de Janeiro, Elsevier Editora ltda, 2016.

SWEIGART, AL. A**utomatize tarefas maçantes com python: Programação prática para verdadeiros iniciantes**. São Paulo: Novatec Editora Ltda. 2015.

STEINBACH, Michael; KUMAR, Vipin. Introdução ao Datamining:Mineração de Dados. Rio de Janeiro: Ciência Moderna, 2009.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. Introdução ao Datamining: Mineração de Dados. Rio de Janeiro: Ciência Moderna, 2009

WITTEN, Ian H.; FRANK, Eibe. Data Mining: Practical Machine Learning Tools andTechiniques,. San Francisco, 2a Edição. Elsevier, 2005.SWEIGART, AL. A**utomatize tarefas maçantes com python: Programação prática para verdadeiros iniciantes**. São Paulo: Novatec Editora Ltda. 2015.

ANEXO 1

Procedimentos realizados no python, para analise da base de dados do covid e vacinação contra covid, realização do pré-processamento, e obtenção dos resultados agrupando e filtrando as informações.

Os arquivos destacados nesse anexo referente a base de dados do covid19 processada no python, podem ser acessados pelo link: https://github.com/helviorezende/arquivos-python/blob/0a921fe0880b9d0547ba9bf6efd6fe4d166330d4/covid.ipynb, a base de dados da vacinação contra covid19 também pode ser acesada pelo link: https://github.com/helviorezende/COVID-19/blob/114800d8ce1340792d7d4de4fa136c37e5f3b26b/vacina.ipynb.

Importando as bibliotecas no python

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from datetime import date

import csv

from google.colab import drive

drive.mount('/content/drive')

importando as base de dados do covid

dfcovid20\_1 = pd.read\_csv("/content/drive/MyDrive/Colab Notebooks/HIST\_PAINEL\_COVIDBR\_25abr2022/HIST\_PAINEL\_COVIDBR\_2020\_Parte1\_25abr2022.csv", sep=';', parse\_dates=['data'], encoding='ISO-8859-1')

dfcovid20\_2 = pd.read\_csv("/content/drive/MyDrive/Colab Notebooks/HIST\_PAINEL\_COVIDBR\_25abr2022/HIST\_PAINEL\_COVIDBR\_2020\_Parte2\_25abr2022.csv", sep=';', parse\_dates=['data'], encoding='ISO-8859-1')

dfcovid21\_1 = pd.read\_csv("/content/drive/MyDrive/Colab Notebooks/HIST\_PAINEL\_COVIDBR\_25abr2022/HIST\_PAINEL\_COVIDBR\_2021\_Parte1\_25abr2022.csv", sep=';', parse\_dates=['data'], encoding='ISO-8859-1')

dfcovid21\_2 = pd.read\_csv("/content/drive/MyDrive/Colab Notebooks/HIST\_PAINEL\_COVIDBR\_25abr2022/HIST\_PAINEL\_COVIDBR\_2021\_Parte2\_25abr2022.csv", sep=';', parse\_dates=['data'], encoding='ISO-8859-1')

dfcovid22\_1 = pd.read\_csv("/content/drive/MyDrive/Colab Notebooks/HIST\_PAINEL\_COVIDBR\_25abr2022/HIST\_PAINEL\_COVIDBR\_2022\_Parte1\_25abr2022.csv", sep=';', parse\_dates=['data'],encoding='ISO-8859-1')

realizando concatenação das bases

covid21=pd.concat([dfcovid21\_1,dfcovid21\_2])

covid20=pd.concat([dfcovid20\_1,dfcovid20\_2])

covid22 = dfcovid22\_1

geral = pd.concat([covid20,covid21,covid22])

verificando campo data

geral.data.describe()

verificando colunas existentes

geral.columns

verificando e calculando as quantidade de valores ausentes

geral.isnull().sum()

configurando campo data, no formato ano, mês e dia

geral['ano'] = geral['data'].dt.year

geral['mes'] = geral['data'].dt.month

geral['dia'] = geral['data'].dt.day

convertendo o numero dos meses pelo nome

geral['mes'].replace(1,'janeiro', inplace=True)

geral['mes'].replace(2,'fevereiro', inplace=True)

geral['mes'].replace(3,'marco', inplace=True)

geral['mes'].replace(4,'abril', inplace=True)

geral['mes'].replace(5,'maio', inplace=True)

geral['mes'].replace(6,'junho', inplace=True)

geral['mes'].replace(7,'julho', inplace=True)

geral['mes'].replace(8,'agosto', inplace=True)

geral['mes'].replace(9,'setembro', inplace=True)

geral['mes'].replace(10,'outubro', inplace=True)

geral['mes'].replace(11,'novembro', inplace=True)

geral['mes'].replace(12,'dezembro', inplace=True)

convertendo o nome dos municpios conforme codigo do ibge no campo codmun:

geral.loc[geral.codmun==314810.0, 'municipio']="Patrocinio"

geral.loc[geral.codmun==317020.3, 'municipio']="Abaete"

geral.loc[geral.codmun==317020.0, 'municipio']="Uberlandia"

geral.loc[geral.codmun==310350.4, 'municipio']="Araguari"

geral.loc[geral.codmun==530010.0, 'municipio']="Brasilia"

#preenchendo missing values, primeira linha pela palavra ausente e segunda linha pelo numero 0

#geral[['estado', 'municipio', 'codmun', 'codRegiaoSaude', 'nomeRegiaoSaude']] = geral[['estado', 'municipio', 'codmun', 'codRegiaoSaude', 'nomeRegiaoSaude']].fillna('ausente')

#geral[['Recuperadosnovos', 'emAcompanhamentoNovos', 'interior/metropolitana']] = geral[['Recuperadosnovos', 'emAcompanhamentoNovos', 'interior/metropolitana']].fillna(0)

geral['estado'].value\_counts()

formando novo datasetchamado covid2

covid2 = pd.DataFrame()

função que mostra a evolução dos óbitos por municipio e data

municipio = "Uberaba"

situacao = "obitosAcumulado"

uberlandia = geral.loc[geral.municipio==municipio,:]

uberlandia = uberlandia.groupby('data').sum()[situacao].reset\_index()

uberlandia = uberlandia.drop('data', axis=1)

uberlandia.columns = [municipio]

covid2 = pd.concat([covid2,uberlandia], ignore\_index=False, axis=1)

covid2

imprimindo o grafico

covid2.plot()

excluindo colunas e gravando no dataset regiaoTotal

regiaoTotal = geral.drop(['municipio','coduf', 'codmun', 'codRegiaoSaude','nomeRegiaoSaude', 'data', 'semanaEpi', 'populacaoTCU2019','casosAcumulado', 'casosNovos','obitosNovos','Recuperadosnovos', 'emAcompanhamentoNovos', 'interior/metropolitana', 'mes','dia'],axis=1)

selecionando as regiões diferentes de brasil pois nas colunas que tem Brasil como regiao, são os dados totalizado de todo país.

obitoRegiao = regiaoTotal.loc[regiaoTotal.regiao!="Brasil",:]

agrupando as informações por ano, regiao, estado e obitos acumulados.

Regiao.groupby(['ano','regiao','estado','obitosAcumulado']).sum().reset\_index()

selecionando regiao centro oeste

centroOeste = obitoRegiao.loc[obitoRegiao.regiao=="Centro-Oeste",:]

verificando estados existente na regiao do Centro Oeste

centroOeste['estado'].unique()

centrooeste = geral.loc[geral.regiao=="Centro-Oeste",:]

centrooeste['estado'].unique()

obitosMS = centroOeste.loc[centroOeste.estado=="MS",['ano','obitosAcumulado']]

após selecionar a regiao, selecionou o estado e agrupou por ano e obitos acumulados para realizar calculo da quantidade de obito por estado e regiao.

obitosMS.groupby(['ano','obitosAcumulado']).sum()