UNIVERSIDADE EVANGÉLICA DE GOIÁS - UNIEVANGÉLICA
ENGENHARIA DE COMPUTAÇÃO/ENGENHARIA DE SOFTWARE

Bianca Abreu Félix de Lima
João Vitor Sponchiado
Natal Junio Barbosa de Souza
Patrick Alves Freitas

TRATATIVA DE DADOS DO E-COMMERCE PARA ANÁLISE DE DEMANDA

Anápolis - GO
Junho, 2022

UNIVERSIDADE EVANGÉLICA DE GOIÁS - UNIEVANGÉLICA
ENGENHARIA DE COMPUTAÇÃO/ENGENHARIA DE SOFTWARE

Bianca Abreu Félix de Lima
João Vitor Sponchiado
Natal Junio Barbosa de Souza
Patrick Alves Freitas

TRATATIVA DE DADOS DO E-COMMERCE PARA ANÁLISE DE DEMANDA

Trabalho apresentado ao Curso de Engenharia de Software da Universidade Evangélica de Goiás – UniEVANGÉLICA, da cidade de Anápolis-GO como requisito parcial para obtenção do Grau de Bacharel em Engenharia de Software.

Orientador (a): Prof. Alexandre Moraes Tannus

Anápolis
Junho, 2022

UNIVERSIDADE EVANGÉLICA DE GOIÁS - UNIEVANGÉLICA
ENGENHARIA DE COMPUTAÇÃO/ENGENHARIA DE SOFTWARE

Bianca Abreu Félix de Lima
João Vitor Sponchiado
Natal Junio Barbosa de Souza
Patrick Alves Freitas

TRATATIVA DE DADOS DO E-COMMERCE PARA ANÁLISE DE DEMANDA

Monografia apresentada para Trabalho de Conclusão de Curso de Engenharia de Software da Universidade Evangélica de Goiás - UniEVANGÉLICA, da cidade de Anápolis-GO como requisito parcial para obtenção do grau de Engenheiro(a) de Software.

Aprovado por:

_____
(ORIENTADOR)

_____
(AVALIADOR)

Anápolis, Junho de 2022.

AGRADECIMENTOS

Agradeço aos meus pais, Leida Maria e Airton Félix, por todo amparo durante esses anos cursando Engenharia de Software e que sem eles, eu não teria chegado até aqui. Eles me guiaram e me orientaram para que eu pudesse progredir. Agradeço a todos que fizeram e que ainda vão continuar fazendo parte da minha vida, durante esses 4 anos estudando juntos. Conheci pessoas incríveis e com toda certeza, contribuíram para o meu aprendizado. Agradeço ao meu irmão, Diego Félix, por sempre acreditar em mim quando eu mesma duvidava e por todo apoio.

Bianca Abreu Félix de Lima.

Venho agradecer primeiramente a minha família que sempre me apoiou e me motivou em tudo o que sempre sonhei, meus pais por nunca deixarem me faltar oportunidades e principalmente meus avós os quais sempre me deram o suporte necessário para que eu realizasse esse desejo de me formar. Aos professores do curso pelos aprendizados tanto nas matérias quanto para a vida e o mercado de trabalho e também aos meus colegas de classe os quais sempre tiveram paciência e me ajudaram no que foi preciso, tanto nos trabalhos como na convivência.

João Vitor Sponchiado.

Agradeço a meus amigos e minha família pelo apoio e forças que me deram nesta jornada que serviram bastante na continuação e realização deste projeto. Sou bastante grato aos meus pais por sempre me incentivarem e acreditarem que eu seria capaz de superar os obstáculos pela frente. Agradeço ao meu orientador, Alexandre Moraes Tannus por estar presente para indicar a direção correta que o trabalho deveria tomar, Também agradeço a meu professor/Amigo Ricardo Dias que sempre me ajudou com sua vasta experiência desde o início deste projeto de pesquisa. Também agradeço à Universidade UniEVANGÉLICA porque sempre trouxe algo diferente para o aprendizado diário.

Natal Junio Barbosa De Sousa.

Agradeço primeiramente a minha mãe que sempre esteve ao meu lado e sempre me motivou a finalizar o curso, aos meus avós que me influenciaram a começar e terminar o curso de engenharia de Software e me auxiliaram bastante tanto em motivação quanto financeiramente. Agradeço também aos meus amigos que sempre estiveram ao meu lado para realização das atividades e também para auxílio de realização de trabalhos e estudo em conjunto proporcionando melhor proveito de cada matéria.

Patrick Alves Freitas.

RESUMO

Neste trabalho iremos realizar uma pesquisa sobre a tratativa de dados, a qual está presente em todos os lugares atualmente. Quando realizamos uma pesquisa de preço ou até mesmo o acesso em algum *site,* implicitamente já estamos participando da análise e tratativa de dados e contribuindo para a realização da coleta de informações e consequentemente já nos trazendo promoções e propagandas referentes a nossa pesquisa requisitada. Essa tratativa envolve vários fatores como a IA (Inteligência Artificial), que envolve o próprio raciocínio da máquina com o mínimo de intervenção humana possível. Um dos tipos de IA que serão utilizados no decorrer desse trabalho será a Inteligência Artificial Limitada (ANI), pois é uma classe de IA considerada fraca, no entanto, possui uma memória limitada mais avançada, capaz de armazenar dados nas escolhas anteriores do usuário e logo é utilizada para tomar decisões. A *Big Data* será apresentado um pouco sobre estatística onde é tratado sobre a melhoria de consulta de tabelas e análise de gráficos, utilizaremos a *Big Data* a favor da grande consulta de dados para otimizar e organizar os resultados obtidos, trabalharemos bastante com a ferramenta dos 5V's. Além destes requisitos, este trabalho apresentará algumas ferramentas mais utilizadas para a análise e tratativa dos dados como a linguagem *Python*, que se trata de uma linguagem tipada e para executar o código será utilizado a ferramenta *Google Colab*, por ser um serviço de armazenamento em nuvem. Os objetivos desse trabalho se resumem em trabalhar com a análise de dados obtendo conhecimento sobre a linguagem *Python* e suas bibliotecas como *Numpy*, *Pandas*, *Matplotlib* e entre outras que serão utilizadas, agregando no conhecimento da utilização de cada uma e também por possuir uma grande base de conhecimento que podemos consultar a qualquer momento e fazer uma tratativa um pouco mais avançada com os resultados obtidos, e com o conhecimento adquirido poder agregar na melhoria das empresas nas quais atuamos.

Palavras-chave: *Big data*, Tratativa de Dados, Inteligência artificial.

ABSTRACT

In this work we will deal with data processing, which is present everywhere today. When we carry out a price survey or even access a website, we are implicitly already participating in the analysis and processing of data and contributing to the collection of information and consequently already bringing us promotions and advertisements regarding our requested research. It is done so quickly today that at the same time we search, the information that is linked to the search is already shown to us. This deal involves several factors such as AI (Artificial Intelligence), which involves the machine's own reasoning with as little human intervention as possible. One of the types of AI that will be used in the course of this work will be Limited Artificial Intelligence (ANI), as it is a class of AI considered weak, but it has a more advanced limited memory, capable of storing data in the user's previous choices and is therefore used to make decisions. Big Data will be presented a little about statistics where it is treated about the improvement of table consultation and graph analysis, we will use Big Data in favor of big data query to optimize and organize the results obtained, we will work a lot with the 5V's tool . In addition to these requirements, this work will present some of the most used tools for analyzing and processing data such as the Python language, which is a typed language in the world of programming in which variables are defined at runtime, and to execute the code, Google Colab tool will be used, as it is a cloud storage service. The objectives of this work are summarized in working with data analysis obtaining knowledge about the Python language and its libraries such as Numpy, Pandas, Matplotlib and among others that will be used, adding to the knowledge of the use of each one and also for having a large base of knowledge that we can consult at any time and make a slightly more advanced treatment with the results obtained, and with the acquired knowledge, we can add to the improvement of the companies in which we operate.

Keywords: *Big data, Data processing, Artificial intelligence.*

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS E SIGLAS

KDD	*Knowledge Discovery in Databases*
IA	*Inteligência Artificial*

SUMÁRIO

1. INTRODUÇÃO

O *E-commerce* nasceu em meados de 1970 nos Estados Unidos, como uma troca de arquivos e solicitações de pedidos. No Brasil o primeiro registro de *E-commerce* foi diretamente de uma grande livraria em 1996, porém, muitas pessoas acreditam que o mesmo teve início em 1999 com o site submarino, que de acordo com uma pesquisa do G1 em 2014, consta dentre os 50 maiores sites de *E-commerce* no *ranking* mundial com um faturamento de cerca US$ 2,477 bilhões.

É importante ressaltar que o *E-commerce* vem sendo agregado com um crescimento contínuo desde 2011. E ao chegar no ano de 2019, mais especificamente a partir de novembro, as vendas por aplicativos *mobile* ultrapassaram as de *desktop*, pois, os *sites* de busca e redes sociais são os principais caminhos para as lojas e também foi comprovado por uma análise feita da E-bit | Nielsen em 2020 de que a *Black Friday* disparou na frente do natal que se consolidava a data mais importante do *E-commerce*.

Juntamente com esse grande crescimento do *E-commerce* encontramos também a criação de ferramentas para estudar e analisar os dados gerados através deste novo modelo de vendas, com isso surgiu vários métodos durante anos até chegarmos às mais conhecidas, como mineração de dados ou KDD. Com estas ferramentas podemos realizar a tratativa dos dados para verificar os parâmetros das vendas, entre outras demandas.

Com os impactos da *COVID-19*, as transações online têm tido um grande aumento devido às modificações feitas no horário comercial e também às métricas aplicadas para quem vai fazer uma compra presencial. Podemos observar que ao longo da pandemia que teve início com grande impacto no Brasil a partir de março de 2020, teve como resultado a aceleração de uma tendência que estava sendo trabalhada a pouco mais de uma década (Alessandro Silveira,2021).

Este trabalho tem como objetivo disponibilizar conhecimento na área de mineração e tratativa de dados com a linguagem *Python*, aplicando a análise da mesma em uma tabela com dados do *E-commerce*, sendo como vendas de produtos em geral, entregas e avaliações. Visando com o objetivo de aplicar o conhecimento adquirido na empresa em que trabalhamos para aumento das vendas e verificações do que precisa ser melhorado.

1.1 PROBLEMA DA PESQUISA

A busca da competência e vantagem competitiva é premissa básica para as empresas que pretendem alcançar o sucesso. Sendo assim, acabar descobrindo formas como as empresas se desenvolvem e mantêm suas vantagens competitivas é um item central na teoria administrativa (Oliveira JR, 1999). Ao longo da pandemia, que iniciou com força no Brasil a partir de março de 2020, e sem dúvidas impactou em um grande aumento nos números de transações *online*, porém, o isolamento não foi a única explicação para o impacto que temos hoje, pois, na verdade apenas acelerou uma tendência que vinha sendo desenhada há pelo menos uma década (Alessandro Silveira, 2021).

Baseando-se em dados divulgados pela empresa Ebit Nilsen, as datas comemorativas tiveram grande impacto no *E-commerce* no ano de 2018, foram os dias da famosa *Black Friday* e em seguida do Natal. Apesar de não ser nenhuma grande surpresa, porém, são épocas que de forma tradicional, acabam agregando maior movimentação de compras e vendas. De acordo com uma pesquisa realizada pela Ebit Nilsen – plataforma de opinião de consumidores do Brasil, foi revelado que o faturamento da *Black Friday* de 2018 ultrapassou o esperado. Devido ao relatório gerado, os pedidos feitos até as 17h da sexta-feira tinham somado um valor de cerca de R$2,1 bilhões, o que representa um aumento de até 27% nas transações do ano anterior (Ebit;Nilsen, 2018).

Conforme os dados disponibilizados pela plataforma *Kaggle* referente a transações do *E-commerce* essa pesquisa tende a realizar a tratativa de dados de uma tabela contendo conteúdos de transações de *marketplaces* em geral. E de acordo com o conhecimento obtido, como podemos aplicar o mesmo dentro da atual empresa que trabalhamos?

1.2 OBJETIVOS

1.2.1 Objetivo Geral:

Uso da tratativa de dados para realizar a verificação das demandas de empresas do *E-commerce*.

1.2.2 Objetivos Específicos:

- Analisar as ferramentas necessárias para a mineração e tratativa dos dados.
- Identificar as bases de dados que serão utilizadas para o estudo.
- Realizar a comparação dos dados entre os gráficos obtidos.
- Aplicar o conhecimento adquirido no mercado de trabalho.

1.3 JUSTIFICATIVA

Com a modernização das tecnologias e as suas evoluções cada vez mais avançadas, verificamos que é fundamental a importância que as organizações disponham do conhecimento de técnicas e ferramentas para análise de dados e de informações, criadas para suportar as decisões estratégicas, táticas e operacionais. Nesse aspecto, a mineração de dados contribui com as descobertas de conhecimentos, pois através de técnicas e ferramentas, ajudam a buscar correlações importantes entre os dados (FAYYAD et al., 1996).

A análise de grandes quantidades de dados pelo homem é inviável sem o auxílio de ferramentas computacionais apropriadas. Portanto, torna-se imprescindível o desenvolvimento de ferramentas que auxiliem o homem, de forma automática e inteligente, na tarefa de analisar, interpretar e relacionar esses dados para que se possa desenvolver e selecionar estratégias de ação em cada contexto de aplicação.(GOLDSCHMIDT R. e PASSOS E.; 2005).

As técnicas de mineração de dados não podem substituir o papel significativo dos especialistas em domínio e seu conhecimento comercial. Porém, pode-se obter resultados úteis combinando com essas técnicas. Como por exemplo, combinar experiência pessoal no campo ou informações de negócios com um modelo de mineração de dados para gerar resultados mais bem-sucedidos. Além disso, esses resultados devem ser sempre avaliados por especialistas. Assim, os conhecimentos do negócio podem ajudar e enriquecer os resultados da mineração de dados (ZIAFAT; SHAKERI, 2014).

A partir destes fatos, a aplicação deste estudo pretende contribuir para o conhecimento pessoal envolvendo análise de dados que lidam com desafios com a operação do *E-commerce*, portanto teremos como base, tabelas de transações *online* envolvendo a venda de produtos, sejam eletrônicos, roupas, objetos em geral. E com os resultados obtidos, uma boa maneira de realizar a tratativa desses dados e a análise entre os gráficos obtidos, tendo em vista crescimentos ou diferenças durante os meses dentro das empresas as quais analisamos os dados.

2. FUNDAMENTAÇÃO TEÓRICA

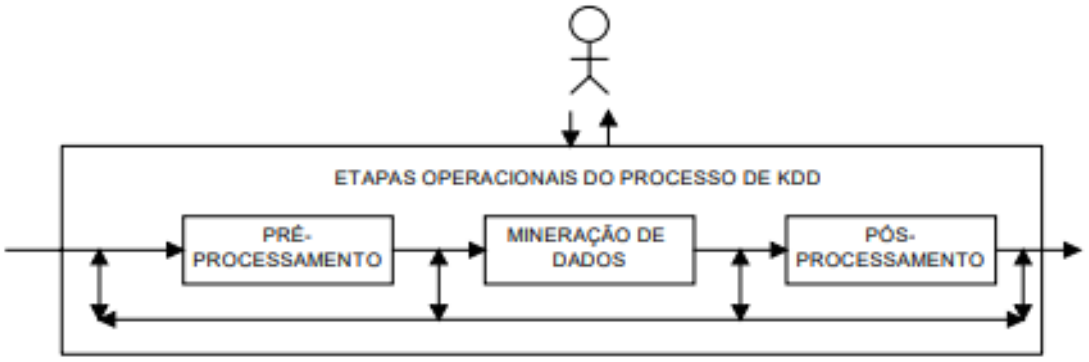
2.1 Knowledge Discovery in Databases - KDD

O termo *Knowledge Discovery in Databases - KDD* foi formalizado em 1989 em referência ao amplo conceito de procurar conhecimento a partir de bases de dados. Uma das definições mais populares foi proposta em 1996, por um grupo de pesquisadores (FAYYAD et al., 1996). É um processo não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados.

A Descoberta de Conhecimento em Bases de Dados é caracterizada como um processo composto por várias etapas operacionais. A Figura 1 apresenta um resumo pragmático das etapas operacionais executadas em processos de KDD. Neste resumo, a etapa de pré-processamento compreende as funções relacionadas à captação, à organização e ao tratamento dos dados. A etapa de pré-processamento tem como objetivo a preparação dos dados para os algoritmos da etapa seguinte, a Mineração de Dados. Durante a etapa de Mineração de Dados, é realizada a busca efetiva por conhecimentos úteis no contexto da aplicação de KDD. A etapa de pós-processamento abrange o tratamento do conhecimento obtido na Mineração de Dados. Tal tratamento, nem sempre necessário, tem como objetivo viabilizar a avaliação da utilidade do conhecimento descoberto (Fayyad et al., 1996a).

O KDD é constituído de três passos básicos que são o pré-processamento, a mineração propriamente dita e o pós-processamento ou interpretação dos resultados, A figura 1 ilustra uma configuração resumida de como funciona esse processo.

Figura 1: Etapas Operacionais do Processo KDD.



Fonte: (BOENTE, A. N. P.; GOLDSCHMIDT, R. R.; ESTRELA, V. V.)

O uso da informação de maneira eficaz e eficiente se torna um elemento essencial para o sucesso das organizações, sendo até incorporado a seu patrimônio. Saber que a

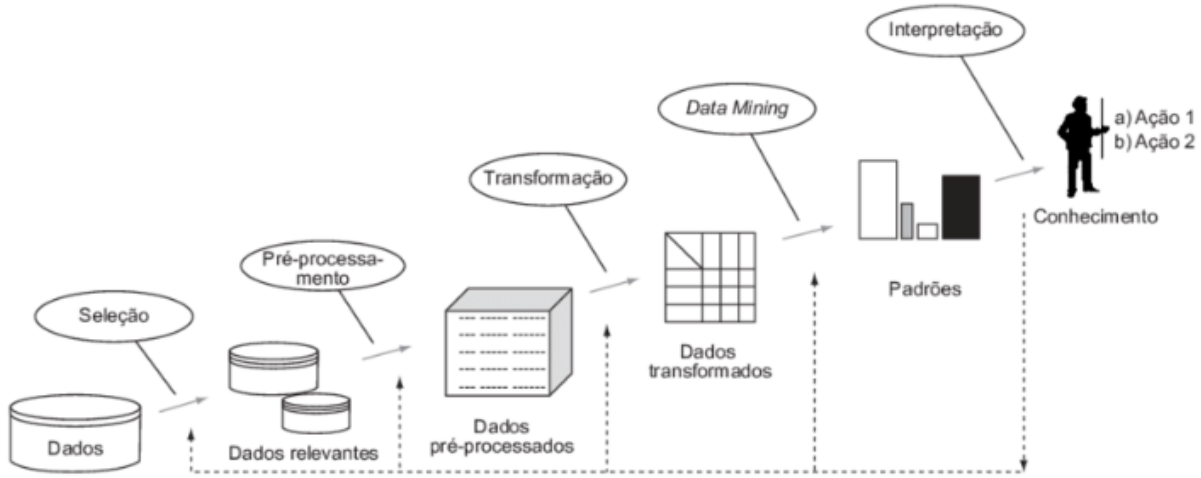
informação é um dos principais recursos estratégicos que a organização dispõe, requer que estas informações estejam estruturadas, disponíveis e sejam íntegras, condições que apenas se fazem possíveis com o uso de tecnologias computacionais, frequentemente conhecidas como Tecnologia da Informação e Comunicação, ou Sistemas de Informação (DALFOVO, 2007).

Para uma melhor compreensão do processo de KDD é necessária uma apresentação dos principais elementos envolvidos em aplicações nesta área. Basicamente, uma aplicação de KDD é composta por três tipos de componentes: o problema em que será aplicado o processo de KDD, os recursos disponíveis para a solução do problema e os resultados obtidos a partir da aplicação dos recursos disponíveis em busca da solução do problema. (GOLDSCHMIDT R. e PASSOS E.; 2005).

Com o avanço da tecnologia um dos efeitos é a quantidade e variedade de informações que são produzidas em um excesso de escala. São várias as aplicações. No momento atual, todas as pesquisas efetuadas em sites da *web* geram dados, especialmente *sites* de comércio eletrônico. Considerando essa concentração de dados, foi desenvolvido o processo do KDD para que conseguisse ser extraído e retirado informações as quais as empresas conseguissem se favorecer e sobrevir os principais objetivos que a empresa tem a aprimorar e a destacar para seu desenvolvimento.

Na figura 2 temos os processos do KDD separados de forma mais abrangente e mais categórica para verificarmos como e todo o processo até chegarmos nas análises dos gráficos obtidos.

Figura 2: Processos do KDD



Fonte: (Fayyad et al. (1996)).

2.1.1 Pré-Processamento

A etapa de pré-processamento inclui todas as funções relacionadas à coleta,

Organização e tratamento de dados, esta etapa é responsável pela preparação

Dados para o algoritmo que será usado na etapa de mineração de dados. As funções principais da etapa do pré-processamento são:

- Seleção de Dados: Nesta etapa focamos na busca pelos dados aos quais queremos tratar dentro do processo de KDD.
- Limpeza dos Dados: Corresponde no tratamento dos dados obtidos de maneira a garantir a qualidade das informações contidas. As informações que estiverem erradas, inconsistentes ou até mesmo ausentes devem ser corrigidas para não prejudicar o banco de dados durante o processo do KDD.
- Codificação dos Dados: Os dados devem ser codificados como Numéricos - forma categórica, converte valores verdadeiros em categorias ou intervalos; Ou Categórico - Numérico, representado numericamente valores de atributos categóricos, para que possam ser usadas como entrada para algoritmos de mineração de dados.
- Enriquecimento dos Dados: Como propriamente dito, é a etapa em que se foca no enriquecimento dos dados das tabelas para que seja o melhor possível e o resultado seja o mais eficiente.

2.1.2 Mineração de Dados

A Mineração de Dados é a principal etapa do processo de KDD. Nessa etapa ocorre a busca efetiva por conhecimentos novos e úteis a partir dos dados. Por este motivo, diversos autores referem-se à Mineração de Dados e ao Processo de KDD de forma indistinta, como se fossem sinônimos. A execução da etapa de Mineração de Dados compreende a aplicação de algoritmos sobre os dados procurando abstrair conhecimento, estes algoritmos são fundamentados em técnicas que procuram, segundo determinados paradigmas, explorar os dados de forma a produzir modelos de conhecimento. (GOLDSCHMIDT R. e PASSOS E.; 2005).

2.1.3 Pós-Processamento

Essa etapa envolve o tratamento do conhecimento obtido na Mineração de Dados. Podem ser citados como exemplos de funções na etapa de Pós-processamento: Elaboração e organização do conhecimento obtido, simplificação do Modelo de

Conhecimento, gráficos, diagramas ou relatórios demonstrativos. O objetivo é basicamente facilitar a visualização do conhecimento adquirido para demonstração, sendo algumas vezes desnecessário esse tratamento (FAYYAD et al., 1996).

2.2 Análise de Dados

Iniciativas de estudo para sistematizar a descoberta de padrões a partir de bases de dados é uma tarefa bastante antiga. Já nos anos 1960, estatísticos se esforçaram na área de Análise de Dados a partir do uso de procedimentos indutivos. Com o passar dos anos e com o aumento no volume de dados, heterogeneidade de formato e necessidade de alta disponibilidade, a área de análise evolui para metodologias exploratórias chamadas, então, de "*Data Fishing*" ou "*Data Dredging*", até que, nos anos 1990, surgiu o termo "*Data Mining*". No mesmo período, Piatetsky Shapiro cunhou o termo *Knowledge Discovery in Databases – KDD*, ou em português "Descoberta de Conhecimento em Bases de Dados ", para o primeiro *workshop* na temática de análise de dados. (Silva, L.A. D., Peres, S. M., & Boscarioli, C.,2016).

Hoje cada vez mais temos acesso a dados dos mais variados tipos e formatos. Além disso, a quantidade de dados com acesso à informação vem aumentando. Para um projeto de análise de dados, é necessário tratar e analisar os dados brutos até se transformarem em informação. (Ferreira, R.G. C.; Miranda, L.B.A. D.; & Pinto,2021). A análise de dados é a arte de transformar dados em conhecimentos e *insights* relevantes. Ou seja, comparar e agregar as informações brutas para entender o que os dados nos dizem e a transformação de números em informação.

As aplicações da análise de dados em vários ramos de atuação deram origem a muitas áreas relacionadas. A categoria geral de análise textual, voltada a agregar valor a partir de texto ou da análise da *Web* que analisa fluxos de dados na *Internet*, muitos fluxos/profissões de análise de dados específicos de um ramo de atuação ou de um problema foram desenvolvidos. Dentre os novos terrenos de aplicação da análise de dados estão *marketing*, varejo, prevenção de fraudes, transportes, saúde, esportes, recrutamento de talentos, ciência comportamental e entre outros. (Sharda, R., Delen, D., & Turban, E. (2019).

Muito embora a análise de dados não seja novidade, a explosão em sua popularidade é bastante nova. Graças à recente explosão em *Big Data*, nos modos de coletar e armazenar esses dados e nas ferramentas de *software* intuitivas, diagnósticos embasados por dados. (Sharda, R., Delen, D., & Turban, E. (2019). O objetivo da análise é examinar os dados para

qualquer aplicabilidade estatística. Com isso, alcançamos conhecimento sobre os dados coletados e, sobretudo, as relações atuais entre as variáveis analisadas. A análise de dados envolve quatro principais tipos, as quais são: Análise Preditiva, Análise Prescritiva, Análise Descritiva e a Análise Diagnóstica.

A análise de dados preditiva visa determinar o que é mais provável de acontecer no futuro. Essa análise se baseia em técnicas estatísticas, bem como em outras técnicas desenvolvidas mais recentemente que recaem na categoria geral de mineração de dados. A meta da análise de dados prescritiva é reconhecer o que está acontecendo, bem como o que deve vir a acontecer, e tomar decisões para garantir o melhor desempenho possível. Seu objetivo geral é otimizar o desempenho de um sistema. A meta aqui é chegar a uma decisão ou a uma recomendação para uma ação específica. A análise de dados descritiva diz respeito a conhecer o que está acontecendo na organização e entender tendências e causas subjacentes de tais ocorrências. Em primeiro lugar, isso envolve a consolidação de fontes de dados e a disponibilidade de todos os dados relevantes de um modo que permita a extração e a análise apropriadas de relatórios. (Sharda, R., Delen, D., & Turban, E.,2019). Análise Diagnóstica tem a finalidade dessa prática é compreender as causas de um evento, ou seja, responder às perguntas: Quem? Quando? Onde? Como? Por quê? A partir disso, pode-se traçar estratégias para aprimorar os resultados.

A coleta desses dados pode ser construída considerando-se a forma na qual esses dados são representados. Dependendo da tecnologia, podemos armazenar dados em um formato considerado *array* dimensional, ou seja, em uma série de dados ou seu cruzamento entre linhas e colunas. (Ferreira, R.G. C.; Miranda, L.B.A. D.; & Pinto,2021).

2.3 Big Data

A definição de *Big Data*, são dados com maior variedade que chegam em volumes crescentes e com velocidade cada vez maior. Simplificando, *Big Data* é um conjunto de dados maior e mais complexo, especialmente de novas fontes de dados. Esses conjuntos de dados são tão volumosos que o *software* tradicional de processamento de dados simplesmente não consegue gerenciá-los. (Oracle, 2020).

Embora o conceito de *Big Data* em si seja realmente novo, as origens de grandes conjuntos de dados remontam às décadas de 1960 e 1970, quando o mundo dos dados estava apenas começando, com os primeiros *data centers* e o desenvolvimento do banco de dados relacional. Por volta de 2005, as pessoas começaram a perceber a quantidade de usuários de dados gerados pelo *Facebook*, *Youtube* e outros serviços *online*. O *Hadoop* (uma estrutura de

código aberto criada especificamente para armazenar e analisar grandes conjuntos de dados) foi desenvolvido no mesmo ano. (Oracle, 2020).

A *Big Data* é estruturada seguindo os conceitos dos 5 Vs, apresentados com base em Barbieri (2011).

O primeiro V refere-se ao termo volume que se refere ao conceito principal da *Big Data*, em que a quantidade de dados a ser coletada e tratada representa um grande volume. Esses dados são advindos de diversas fontes, como redes sociais, motores de busca da internet, *E-commerce*, entre outras. Este talvez tenha sido o grande diferencial em relação aos outros bancos de dados: o tratamento de grandes volumes de informações de dados, com formas diferentes (estruturados, não estruturados e semiestruturados).

Já o segundo V, diz respeito a velocidade, como os dados são gerados em grandes volumes, o seu processamento deve efetuar o tratamento da mesma forma, caso contrário, ocorrem filas e a consequente degradação do serviço. Para tanto, o servidor (*hardware*) deve possuir uma capacidade de processamento adequada às necessidades. Em seguida, o terceiro V se refere à variedade, os bancos de dados que trabalham dentro do conceito de *Big Data* devem possuir técnicas que possibilitem o tratamento dos mais diferentes tipos de dados, como números, tags, imagens e textos.

O quarto V, é voltado para a veracidade, onde acontece a necessidade de constante análise em tempo real, isso significa de dados que condizem com a realidade daquele momento, pois dados passados não podem ser considerados dados verídicos para o momento em que é analisado.

E por fim temos o valor, que mostra a significância dos dados coletados e tratados para a organização, de forma que possam trazer informações relevantes, que possam proporcionar um diferencial de mercado ou, ainda, auxiliar os gestores na tomada de decisão. Esse último ponto das características da *Big Data* é exatamente o que as empresas buscam e precisam, pois os dados de valor podem ser convertidos em informações importantes para a companhia.

Mas, de que forma que os 5 Vs, de fato, são utilizados na aplicação da *Big Data*? Inicialmente, deve-se analisar o cenário no qual será aplicado e utilizado a *Big Data*. Com isso, será possível compreender o volume e a velocidade com que esses dados são gerados, sendo necessário observar a variedade dos dados e utilizar filtros para se garantir a veracidade desses dados coletados. Por fim, deve-se garantir que esses dados sejam de grande significância e valor para a empresa. (Barbieri, 2011).

A *Big Data* surgiu como uma tecnologia inovadora e com alto potencial para auxílio no crescimento das empresas para que possam garantir seu espaço no mercado que vem com a tendência de ficar cada vez mais competitivo. Existem alguns segmentos nos quais o cenário da *Big Data* pode ser aplicado para gerar mais lucro e assertividade no ramo empresarial, onde temos os comportamentos e tendências, a *Big Data* permite compreender o comportamento dos consumidores e a tendência de consumo, apontando quais produtos ou serviços foram mais relevantes em um período de tempo. (Dumbill,2013).

Ainda no ramo empresarial contamos com a estratégia de *marketing*, onde grande parte da aplicação comercial da *Big Data* está ligada às estratégias de *marketing*, pois ele permite realizar análises de dados e direcionar as ofertas de produtos e serviços de forma mais assertiva. Melhoria de produtos e serviços com a estratégia de *marketing*, os *feedbacks* dos consumidores tendem a serem mais fornecidos em redes sociais, melhorando assim o retorno dos produtos oferecidos pelas empresas.

2.4 Python

O *Python* foi criado por Guido Van Rossum em 1991, tendo como origem de seu nome a série humorística britânica *Monty Python's Flying Circus*, e atualmente tem conquistado bastante espaço entre as outras ferramentas de programação, por ter uma interface "amigável" de fácil aprendizagem e pela sua grande aplicabilidade. Atualmente o *Python* é utilizado por grandes empresas de tecnologia, como a *Google*, *Microsoft*, *Instagram*, *Spotify* e várias outras. (SANTANA; GALESI, 2010).

Python é uma linguagem de uso geral, projetada especificamente para tornar os programas bastante legíveis. *Python* também possui uma rica biblioteca, tornando possível criar aplicações sofisticadas usando código de aparência relativamente simples. Por esses motivos, *Python* tornou-se uma linguagem de desenvolvimento de aplicações populares e também uma preferência como "primeira" linguagem de programação. (PERKOVIC, LJUBOMIR, 2016).

A linguagem de programação *Python* é de altíssimo nível, de tipagem dinâmica e forte, iterativa e interpretada e orientada a objeto. Ela possui uma sintaxe clara e concisa, favorecendo a legibilidade do código fonte e com isso faz com que a linguagem seja mais produtiva (BORGES, 2010). Existem muitas ferramentas de desenvolvimento para *Python*, como IDEs, editores e *shells* (que aproveitam da capacidade interativa do *Python*).

A linguagem inclui diversas estruturas de alto nível (listas, tuplas, dicionários, data / hora, complexos e outras) e uma vasta coleção de módulos prontos para uso, além de

frameworks de terceiros que podem ser adicionados. Também possui recursos encontrados em outras linguagens modernas, tais como: geradores, introspecção, persistência, meta classes e unidades de teste. Multiparadigma, a linguagem suporta programação modular e funcional, além da orientação a objetos. Mesmo os tipos básicos no *Python* são objetos. A linguagem é interpretada através de *bytecode* pela máquina virtual *Python*, tornando o código portável. Com isso é possível compilar aplicações em uma plataforma e rodar em outras ou executar direto do código fonte (LUIZ E. BORGES, 2009).

Essa linguagem é muito utilizada como programação de *script*, como *Perl* e *Scheme*. Por outro lado, conta com recursos que a equiparam a linguagens como C, C++ e Java, permitindo o desenvolvimento de grandes projetos que podem ser constituídos por diversos módulos, que acessem bancos de dados, que enviem e recebam dados por meio de redes, trabalhem com recursos multimídia, entre outros. *Python* também dispõe de mecanismos que permitem a integração com *softwares* escritos em outras linguagens, como C. Pode ser utilizado em um grande número de áreas do desenvolvimento de *software*, das quais se destacam: ferramentas para administração e interface com sistemas operacionais; aplicações que trabalhem com grandes volumes de dados armazenados em sistemas gerenciadores de bancos de dados. (Banin, Sérgio L, 2018).

A linguagem tem muitas outras funções e classes definidas na Biblioteca Padrão *Python*. A Biblioteca Padrão *Python* (*Python Standard Library*) consiste em milhares de funções e classes organizadas em componentes chamados módulos. Um módulo é simplesmente um arquivo contendo código *Python*, cujo nome de arquivo termina com *.py* é um módulo *Python*. Cada módulo contém um conjunto de funções e/ou classes relacionadas a determinado domínio de aplicação. Mais de 200 módulos embutidos formam juntos a Biblioteca Padrão *Python*. (PERKOVIC, LJUBOMIR, 2016). Por ser muito extensa, oferece um vasto agrupamento de módulos e funções fundamentais para reduzir o uso de código. Os módulos oferecem acesso a finalidade do sistema, assim, solucionando padrões para diversas complicações que decorrem na programação.

3. METODOLOGIA DA PESQUISA

O fundamento deste tema foi baseado em Tratativa de Dados, que consistirá pela linguagem *Python* para estruturar e ajustar, de forma ordenada todos os dados que será feito a verificação de análise da demanda do *E-Commerce*. Foi utilizada a análise exploratória dos dados, com isso, obtemos entendimento sobre os dados coletados, para tornar o nosso tema mais compreensível, com a pesquisa descritiva será descrito o que será observado, ou seja, os dados que serão analisados no processo de organização e informações em um formato adequado à visualização, conforme a disponibilidade das tabelas oferecidas.

É proposto um estudo categorizado como um estudo de caso, cujo objeto é a Tratativa de Dados do *E-Commerce*. Segundo Gilberto de Andrade, o estudo de caso deve apresentar indicadores de confiabilidade dos instrumentos de coleta de dados utilizados. Uma grande utilidade dos estudos de caso é verificada nas pesquisas exploratórias. Por sua flexibilidade, é recomendável nas fases iniciais de uma investigação sobre temas complexos, para a construção de hipóteses ou reformulação do problema. (GIL 2008)

Para dar início ao trabalho, utilizaremos a plataforma *Kaggle* que disponibiliza uma ampla gama de atividades envolvendo *Data Science*. Nessa plataforma, vários conjuntos de dados estão disponíveis e que pode ser feito a extração dos dados, sendo facilmente importados para o ambiente de análise. O *Kaggle* possibilita que os usuários acessem uma série de dados, analisem e produzam, envolvendo a ciência de dados.

Tendo isso, utilizaremos a ferramenta *Colab*, que é uma ferramenta gratuita do *Google* para criar e executar códigos na linguagem *Python* e com ele, podemos executar programas de forma simples e rápida diretamente do navegador. É uma ferramenta que não há necessidade de instalar o *Python* e qualquer outro *software* na máquina para ser utilizado e, até mesmo, vem pré-instaladas diversas bibliotecas como *Numpy*, *Pandas* e *Matplotlib*, que facilitam a importação e o aproveitamento de seus recursos.

Com o *Colab*, é viável importar agrupamentos de dados de imagem, treinar um classificador de imagens dentro da ferramenta e avaliar o modelo, tudo com apenas algumas linhas de código. É uma ferramenta que possibilita misturar código fonte (geralmente em *Python*) e texto rico (geralmente em *markdown*) com imagens e o resultado desse código. É um método conhecido como *notebook* (“caderno”). Esse termo foi influenciado pelos *notebooks* do *Jupyter*. *Colab* trabalha em especial com a linguagem *Python*, mas com algumas adaptações para compartilhamento de código.

Existem inúmeras bibliotecas *Python* com diversas finalidades. Dentre elas que citamos e as quais iremos utilizar neste trabalho, estão: A biblioteca *NumPy* que oferece o código aglutinador para as estruturas de dados, os algoritmos e a biblioteca necessária à maioria das aplicações científicas que envolvem dados numéricos em *Python*. A *NumPy* acrescenta ao *Python*, um de seus principais usos em análise de dados é como um contêiner para que dados sejam passados entre algoritmos e bibliotecas. A biblioteca *Pandas* oferece estruturas de dados de alto nível e funções, projetadas para fazer com que trabalhar com dados estruturados seja rápido, fácil e expressivo. Desde o surgimento em 2010, o *Pandas* tem ajudado a viabilizar o *Python* como um ambiente eficaz e produtivo para análise de dados. A biblioteca *Matplotlib* é a mais popular para fazer plotagens e gerar outras visualizações de dados bidimensionais. A biblioteca foi projetada para criar plotagens apropriadas para publicação. A *Matplotlib* é a mais amplamente utilizada e, desse modo, tem uma boa integração em geral com o restante do ecossistema. (W. McKinney, 2018). A *Seaborn* é uma biblioteca de visualização de dados da linguagem de programação *Python* baseada na biblioteca de plotagem *Matplotlib*. *Seaborn* oferece uma interface de alto nível para desenhar gráficos estatísticos elegantes e informativos, melhorando a visualização dos dados. (Ferreira, R.G. C., Miranda, L.B.A. D., & Pinto,2021).

Tendo em vista as bibliotecas que serão utilizadas para o desenvolvimento e a plataforma a qual utilizaremos para trabalhar, iremos criar o código para a verificação dos dados e geração das tabelas para que assim nós possamos realizar as análises previstas e necessárias das tabelas.

4. DESENVOLVIMENTO

4.1 Tabela 1 - Vendas comércio eletrônico 2021/2022

Ao analisar a primeira tabela para tratativa nos deparamos com 12 colunas e com alguns campos vazios e também com vários caracteres especiais. A base disponibilizada possuía em sua maioria produtos eletrônicos as informações dos dados eram o número de ordem de compra, data da ordem de compra, comprador, cidade e estado, código do produto, descrição, quantidade, preço total do item e preço do frete, se foi pago na entrega e se a entrega foi realizada ou teve retorno para o vendedor.

Para a primeira tratativa, foi feita a separação de duas colunas: a data e o total de itens vendidos, fizemos a redefinição delas para gerar um gráfico de linha para uma visualização mais ampla do resultado.

O trecho do código utilizado para agrupamento das colunas foi:

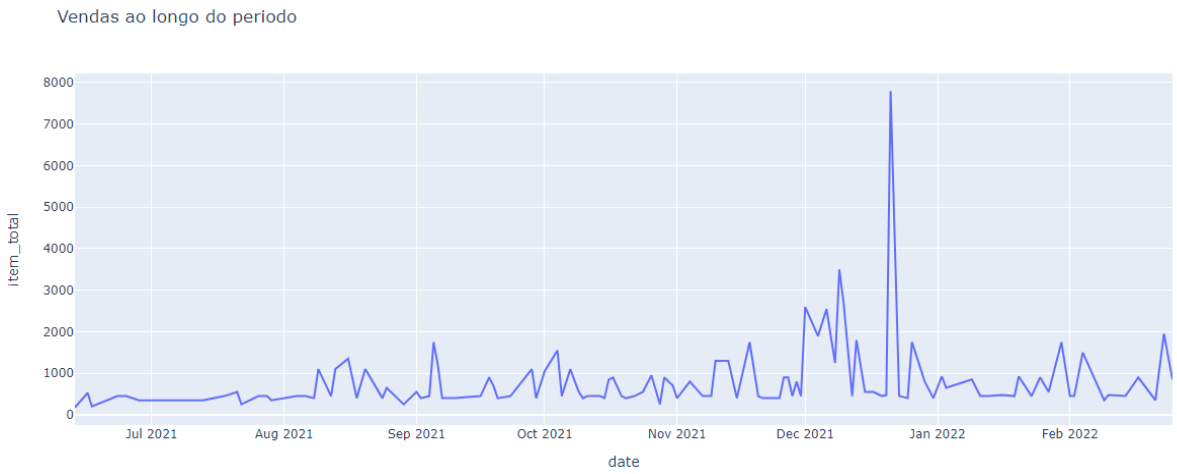
Figura 3: Agrupamento de colunas

```
df_sales = df[['date', 'item_total']].groupby('date').sum().reset_index()
```

Fonte: Os Autores.

Em seguida utilizamos a função *px.line* da biblioteca *Plotly* para gerar o seguinte gráfico:

Gráfico 1: Vendas ao longo do período



Fonte: Os Autores.

Após o resultado do gráfico 1 referente às vendas ao longo do segundo semestre de 2021 para o primeiro semestre de 2022 podemos observar que tivemos um aumento relevante no mês de dezembro, onde comemoramos o natal e ano novo. O gráfico 1, é uma representação básica da análise feita da planilha obtida para visualização de maior aumento de demandas e transações no e-commerce.

Em seguida, tivemos a classificação em barras no gráfico 2 para saber os dias de cada mês que teve maior quantidade de vendas de produtos, então agrupamos os dados por ano, mês e nome dos dias da semana. Segue o trecho do código:

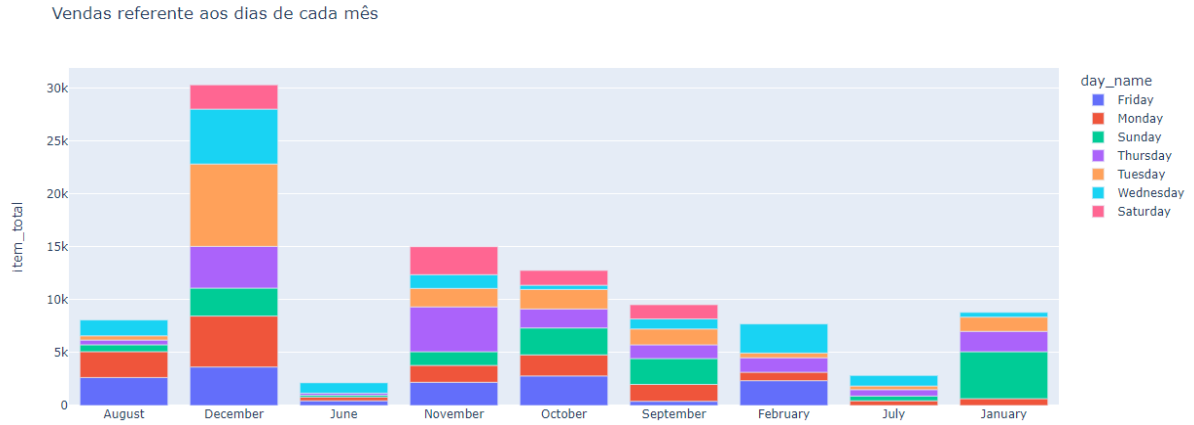
Figura 4: Agrupamento de ano, mês, dia e total de itens.

```
df_days = df[['year', 'month_name', 'day_name', 'item_total']].groupby (['year', 'month_name', 'day_name']).sum().reset_index()
```

Fonte: Os Autores.

Depois utilizamos a função *px.bar* para gerar o gráfico baseado em barras para a devida análise dos valores.

Gráfico 2: Vendas referente aos dias de cada mês.

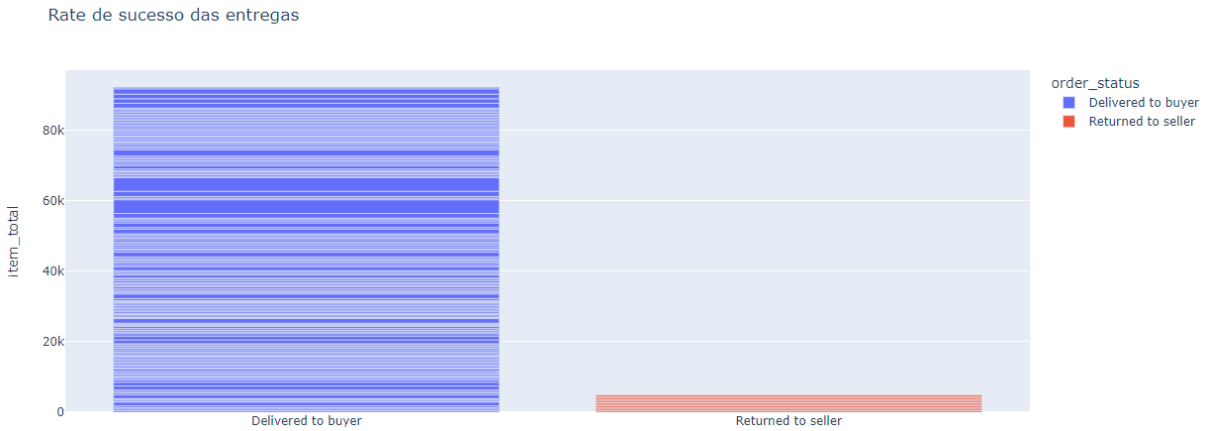


Fonte: Os Autores.

Ao gerar o gráfico 2, e como já era de se esperar conforme o gráfico de linhas anterior (gráfico 1) a maior quantidade foi no mês de dezembro, só que através deste podemos obter uma informação de suma importância que o maior dia de vendas foi em uma terça feira, e podemos ter como base que era um dia de pagamento por ser o quinto dia útil do mês e na semana do natal, era o dia mais viável para compras. Visto que buscamos extrair o máximo de dados disponibilizados na tabela para aplicar o formato de gráficos.

Para finalização da análise, no gráfico 3 utilizamos as últimas colunas que faltaram, que era saber qual a quantidade de entregas deram certo e quantas tiveram retorno. Na primeira análise da tabela, apenas olhando superficialmente os dados já podíamos imaginar que haveria uma diferença exorbitante entre essas colunas e claramente tivemos, entretanto agrupamos as colunas de status da ordem e total dos itens para gerar uma visualização com a função *px.bar* e trazer a *rate* de sucesso das entregas e a *rate* dos retornos ao vendedor.

Gráfico 3: Rate de sucesso das entregas em formato de barras.



Fonte: Os Autores.

4.2 Tabela 2 - Vendas eletrônicos Junho 2021

Ao iniciar a tratativa da segunda tabela, tivemos a visualização de algumas colunas, como código do produto, categoria, link do produto, número de vendas, ranking, rating de avaliações, quantidade de reviews e preço. Já antes de aplicar o código nesta planilha, decidimos ignorar duas colunas que seria a do código do produto e também a do link do produto, pois seriam duas colunas que não iriam apresentar mudanças no resultado final.

Logo em seguida apliquei a consulta da tabela no mesmo código utilizado para análise da tabela 1, porém com formatações um pouco diferentes. Utilizando a função *.replace* modificamos alguns valores em colunas específicas como: *reviews count*, *price*, *rank*, *No of Sellers*, tanto para remover caracteres especiais, quanto para remover vírgulas.

Figura 5: Atribuição da formatação das colunas de visualização, preço, ranking e número de vendas.

```
df['Reviews Count'] = df['Reviews Count'].replace(',','', regex=True)
df['Price'] = df['Price'].str.replace('$','', regex=True)
df['Rank'] = df['Rank'].str.replace('#','', regex=True)
df['No of Sellers'] = df['No of Sellers'].str.replace(' Sellers','', regex=True)
```

Fonte: Os Autores.

Depois geramos uma visualização geral da tabela para analisar os valores e como estavam, e já pudemos identificar que a formatação estava conforme iremos precisar para gerar os gráficos e facilitar a análise.

Usando a função *df* (variável que estava vinculada a leitura da tabela) podemos imprimir os dados formatados.

Figura 6: Dados após tratamento de formatação.

	ASIN	Category	Product Link	No of Sellers	Rank	Rating	Reviews Count	Price
0	B079QHML21	Electronics	https://www.amazon.com/gp/offer-listing/B079QH...	1	1	2022-07-04 00:00:00	640.721	39.99
1	B07FZ8S74R	Electronics	https://www.amazon.com/gp/offer-listing/B07FZ8...	1	2	2022-07-04 00:00:00	854.114	34.99
2	B07XJ8C8F5	Electronics	https://www.amazon.com/gp/offer-listing/B07XJ8...	1	3	2022-07-04 00:00:00	267.821	44.99
3	B07WVFCVJN	Electronics	https://www.amazon.com/gp/offer-listing/B07WVF...	27	4	2022-08-04 00:00:00	114.267	28.48
4	B08YT2N5SX	Electronics	https://www.amazon.com/gp/offer-listing/B08YT2...	1	5	2022-07-04 00:00:00	267.821	49.99
...	...	...	...	...	...	...	...	...
702	B007DW6F34	Toys & Games	https://www.amazon.com/gp/offer-listing/B007DW...	10	95	2022-06-04 00:00:00	8.795	6.99
703	B01N16VX79	Toys & Games	https://www.amazon.com/gp/offer-listing/B01N16...	5	96	2022-07-04 00:00:00	649.000	16.99
704	B09197N995	Toys & Games	https://www.amazon.com/gp/offer-listing/B09197...	1	97	2022-08-04 00:00:00	9.121	8.99
705	B015CCR1FW	Toys & Games	https://www.amazon.com/gp/offer-listing/B015CC...	26	98	2022-07-04 00:00:00	18.449	19.99
706	B07TS96J7Q	Toys & Games	https://www.amazon.com/gp/offer-listing/B07TS9...	2	99	2022-07-04 00:00:00	12.902	24.99

Fonte: Os Autores.

Depois formatamos os valores devidos para inteiro com a função *astype(int)* e para *float* com a função *.astype(float)*, segue o trecho de código usado:

Figura 7: Atribuição de formato de variáveis.

```
df['No of Sellers'] = df['No of Sellers'].astype(int)
df['Rank'] = df['Rank'].astype(float)
df['Reviews Count'] = df['Reviews Count'].astype(int)
df['Price'] = df['Price'].astype(float)
```

Fonte: Os Autores.

Com a transformação do tipo dos valores imprimimos a tabela novamente para verificar se a transformação ocorreu conforme estávamos esperando e tivemos um resultado plausível conforme esperado.

Figura 8: Tabela após a atribuição dos tipos de variáveis.

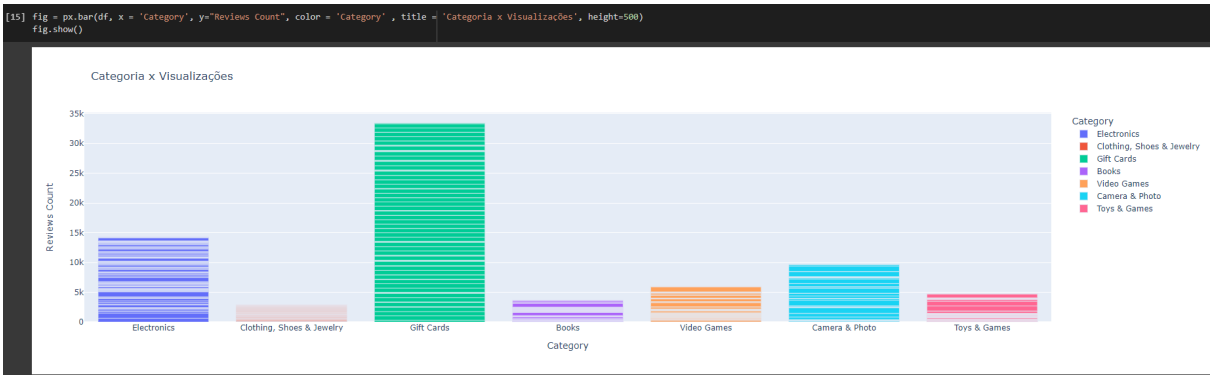
```
[14] df
```

	ASIN	Category	Product Link	No of Sellers	Rank	Rating	Reviews Count	Price
0	B079QHML21	Electronics	https://www.amazon.com/gp/offer-listing/B079QH...	1	1.0	2022-07-04 00:00:00	640	39.99
1	B07FZ8S74R	Electronics	https://www.amazon.com/gp/offer-listing/B07FZ8...	1	2.0	2022-07-04 00:00:00	854	34.99
2	B07XJ8C8F5	Electronics	https://www.amazon.com/gp/offer-listing/B07XJ8...	1	3.0	2022-07-04 00:00:00	267	44.99
3	B07WVFCVJN	Electronics	https://www.amazon.com/gp/offer-listing/B07WVF...	27	4.0	2022-08-04 00:00:00	114	28.48
4	B08YT2N5SX	Electronics	https://www.amazon.com/gp/offer-listing/B08YT2...	1	5.0	2022-07-04 00:00:00	267	49.99
...	...	...	...	...	...	...	...	...
702	B007DW6F34	Toys & Games	https://www.amazon.com/gp/offer-listing/B007DW...	10	95.0	2022-06-04 00:00:00	8	6.99
703	B01N16VX79	Toys & Games	https://www.amazon.com/gp/offer-listing/B01N16...	5	96.0	2022-07-04 00:00:00	649	16.99
704	B09197N995	Toys & Games	https://www.amazon.com/gp/offer-listing/B09197...	1	97.0	2022-08-04 00:00:00	9	8.99
705	B015CCR1FW	Toys & Games	https://www.amazon.com/gp/offer-listing/B015CC...	26	98.0	2022-07-04 00:00:00	18	19.99
706	B07TS96J7Q	Toys & Games	https://www.amazon.com/gp/offer-listing/B07TS9...	2	99.0	2022-07-04 00:00:00	12	24.99

Fonte: Os Autores.

E após essa finalização das transformações de valores, começamos o processo de gerar os gráficos para analisar as demandas, a princípio utilizamos a coluna de categoria como principal, pois basicamente estava interligada com todas as outras. Então utilizando a função *px.bar* geramos o gráfico 4 com agrupamento de categoria e número de visualizações.

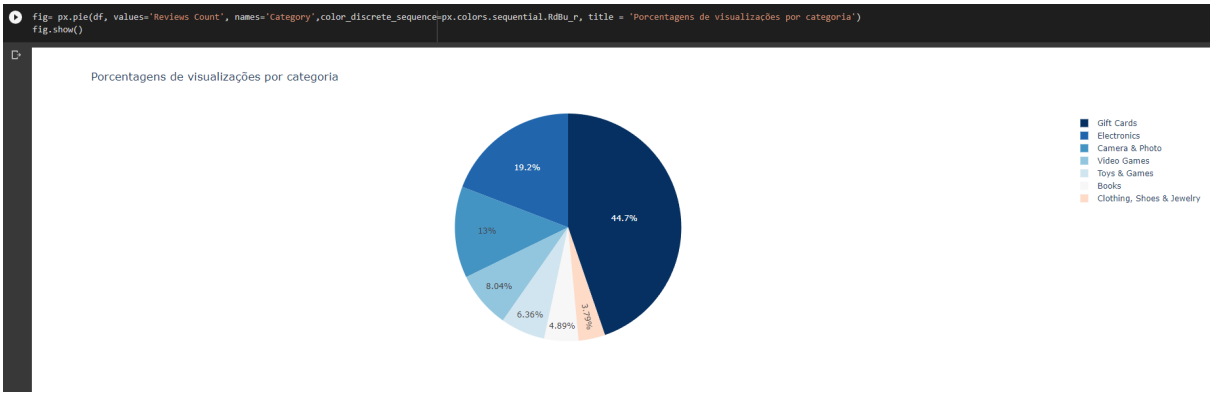
Gráfico 4: Visualizações por categorias em formato de barras.



Fonte: Os Autores.

Tivemos uma breve noção de como foram em mais questão de quantidade, então para uma análise mais aprofundada usamos mais uma função disponibilizada pela biblioteca *plotly* que seria a *px.pie* que nos gerou o gráfico 5 que trouxe os dados em porcentagens.

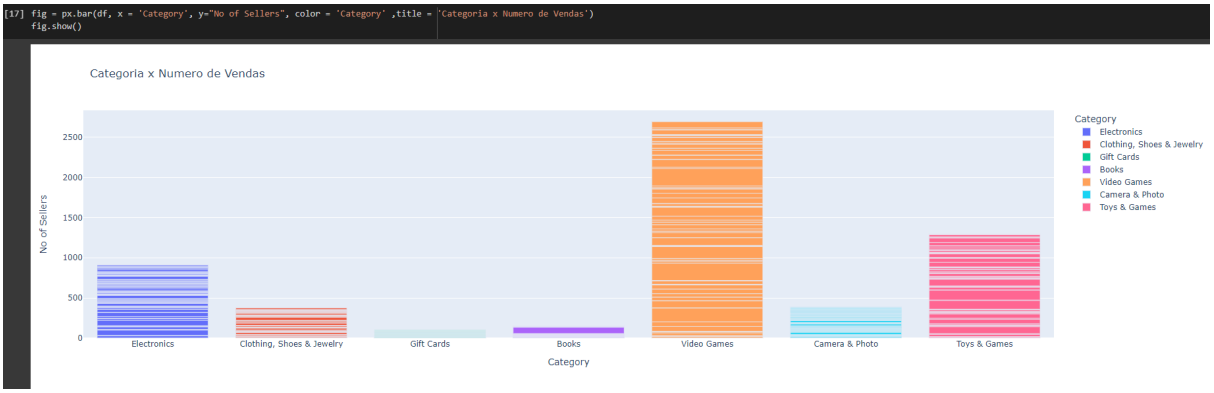
Gráfico 5: Porcentagem de visualização por categoria em formato pizza.



Fonte: Os Autores.

Em seguida, usando as mesmas funções dos gráficos acima, fizemos a análise agrupando categorias e número de vendas. Função *px.bar* gerando o gráfico 6 de barras.

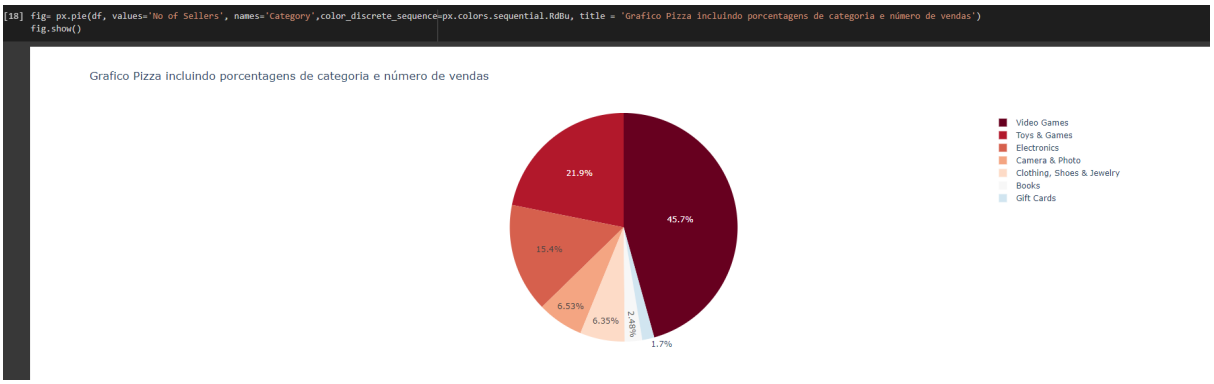
Gráfico 6: Número de vendas por categorias em formato de barra.



Fonte: Os Autores.

E depois a função *px.pie* para as porcentagens no gráfico 7.

Gráfico 7: Porcentagem de categoria e número de vendas em formato pizza.

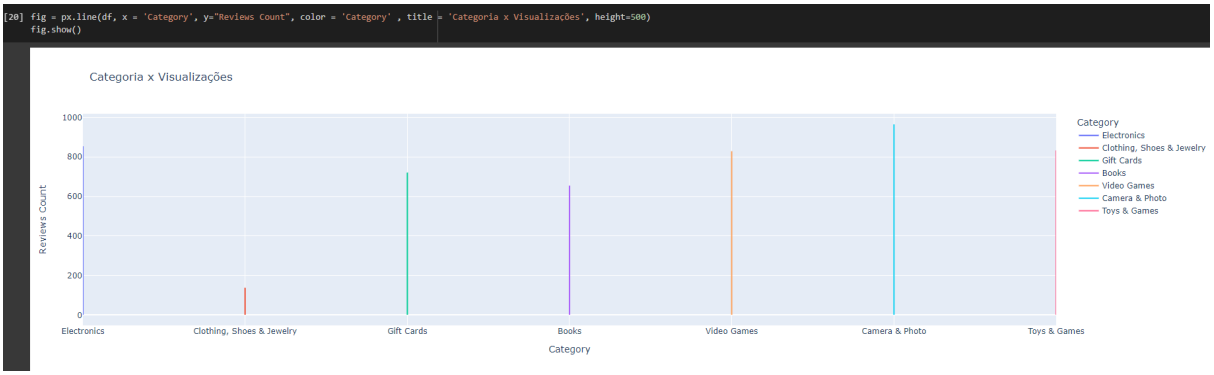


Fonte: Os Autores.

Os gráficos acima foram como os principais usados, e depois começamos a fazer alguns testes com agrupamento de outras colunas para teste de resultados, e tivemos alguns resultados um tanto quanto interessantes, porém, com gráficos meio complicados de se analisar.

Vale ressaltar que foram apenas testes para contribuição do conhecimento, e o porquê de não usá-las dependendo dos valores que possui. Com a função *px.line* e agrupando categoria e visualização tivemos o gráfico 8.

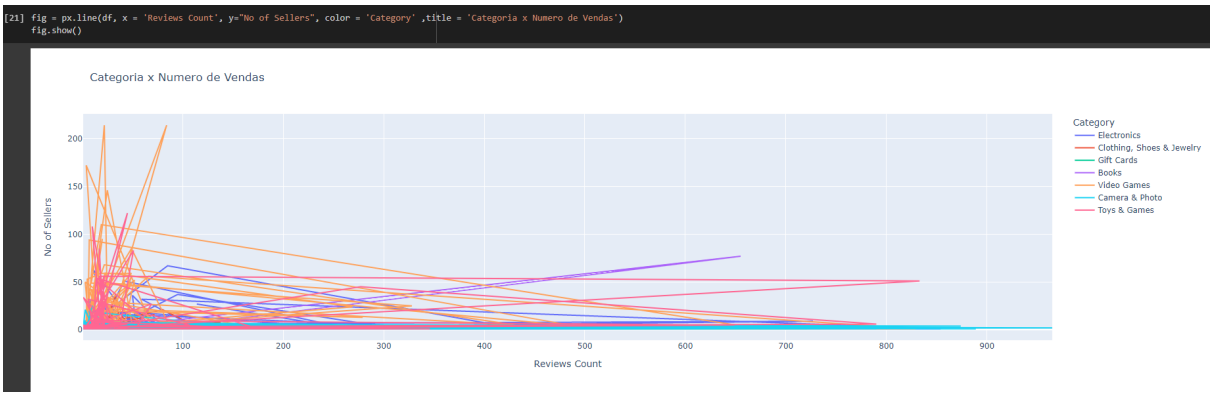
Gráfico 8: Visualizações por categorias em formato de linha



Fonte: Os Autores.

Podemos observar que pelas linhas, a análise fica um pouco complexa, pois as linhas não apresentam uma visualização um tanto quanto ampla sobre o resultado. Em seguida, agrupamos o número de visualizações e o número de vendas, só que obtivemos o gráfico 9 que apresentou uma difícil análise, cuja função utilizada foi a *px.line.*

Gráfico 9: Agrupamento de categoria por número de vendas



Fonte: Os Autores.

4.3 Tabela 3 - Base de dados vendas das Lojas Americanas

Para a análise dessa tabela deparamos com 15 colunas referente aos gastos do primeiro semestre de 2019 ao primeiro semestre de 2022, nessa tabela está apresentando custos das mercadorias vendidas e serviços prestados, lucro bruto, despesas operacionais e compras físicas e digitais. Para essa tratativa, a analisar para demostrar-la usamos algoritmos diferentes da tabela 1 e 2, para gerar a tabela da americanas usamos o seguinte algoritmo simples:

Figura 9: dataframe para pandas.

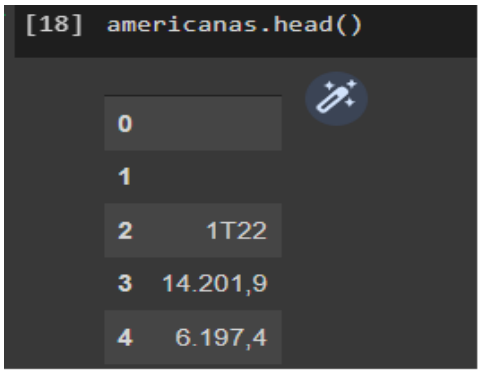
```
[14] americanas = pd.DataFrame(page.get_all_records())

[18] americanas.head()
```

Fonte: Os Autores.

Para não ter que verificar células por células, linha por coluna, sem fazer alteração na tabela gerada pela Americanas SA, usamos um comando onde americanas que no caso seria o nome da tabela, trabalhar em um dataframe do *Google* sem precisar fazer as verificações das colunas por colunas ou linhas por linhas, depois de criar esse dataframe usando a função *.head()*, para usar esse código temos que fazer a importação pandas, seguindo isso visualizamos a seguinte a tabela:

Figura 10: tabela gerada.



Fonte: Os Autores.

Depois de ter gerado a tabela para manipulação o resultado deu de uma tabela completamente bagunçada, sendo assim teríamos que reorganizar a tabela gerada pela Americanas SA, mas ao em vez disso, conseguimos manipular essa tabela de forma manual sem alterações na tabela utilizada nessa análise.

Para conseguir os dados necessários dessa tabelas usamos da seguinte forma:

Figura 11: variável e fileiras usadas.

```
pl = gc.open('PLA')

[23] page = pl.sheet1

[12] page.row_values(8)

[21] page.row_values(9)

[22] page.row_values(12)
```

Fonte: Os Autores.

pl seria o nome da variável que foi atribuída para abrir a planilha no *Google Sheet* que está salvo como "PLA", depois de acessar essa planilha peguei as fileiras que precisava para a análise que está localizada em linha 8,9 e 12, e para isso utilizei o comando *page.row_values* onde pegamos os valores necessários para a criação dos gráficos em *matplotlib*.

Depois dessa manipulação e verificação dos dados na tabela seguimos em pegar os valores do primeiro semestre ao quarto semestre de 2021, pegando o valores dos quatro semestre e atribuindo em um gráfico, importado do *matplotlib*, usando o seguinte código:

Figura 12: importação e criação do gráfico.

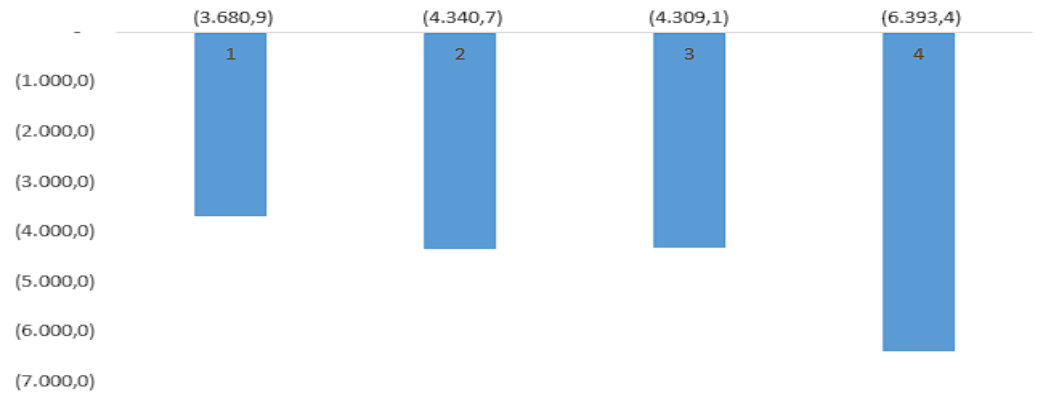
```
import matplotlib.pyplot as plt

data = {'1': 3.680_9, '2': 4.340_7, '3': 4.309_1, '4': 6.393_4}
names = list(data.keys())
values = list(data.values())
```

Fonte: Os Autores.

Que gerou o resultado das colunas das mercadorias vendidas e serviços que gerou o gráfico 10:

Gráfico 10: mercadorias vendidas.

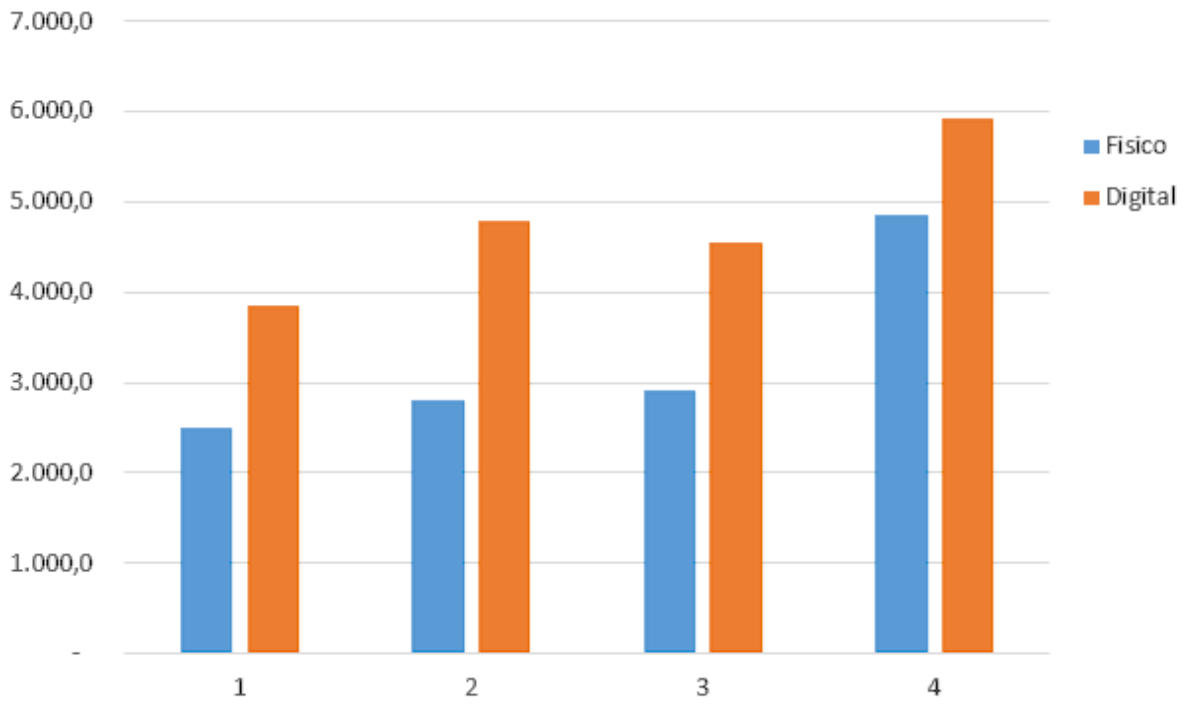


Fonte: Os Autores.

no gráfico 10 mostra a numeração nas barra que estão de 1 a 4 que se refere aos semestres de 2021, reparamos que no primeiro semestre as vendas de mercadorias e serviços prestados chegou aos 3.680,9 e teve um aumento maior em 6.393,4, que se refere ao quarto semestre de 2021 que é referente ao mês de natal e ano novo.

A análise do gráfico 11 a seguir e o comparativo de mídia digital e física nos semestres de 2021. Até o momento presente, com a importação do *matplotlib* da análise da tabela da Americanas SA.

Gráfico 11: físico x digital.



Fonte: Os Autores.

no gráfico 11 mostra o comparativo do primeiro semestre ao quarto, onde a mídia física sempre está atrás comparado a mídia digital que sempre esteve acima desde o primeiro semestre, mas note que entre o primeiro ao terceiro eles sempre esteve em constante diferença, mostrando que o *E-commerce* das Americanas S.A sempre foi o foco de vendas.

O código para criação da tabela ficou da seguinte forma:

Figura 13: Código de criação do gráfico físico e digital.

```
import matplotlib.pyplot as plt
import numpy as np


labels = ['1', '2', '3', '4']
fisico = [2.510, 2.796, 2.920, 4.853]
digital = [3.860, 4.782, 4.542, 5.923]

x = np.arange(len(labels))
width = 0.28
```

Fonte: Os Autores.

Portanto a análise foi feita retirando apenas colunas importantes referente ao *E-commerce*, de acordo com a Americanas S.A 2021 foi um ano histórico para a trajetória de mais de 90 anos.

5. RESULTADOS

Os resultados obtidos através da análise da primeira tabela, referente ao gráfico 1, foi que o aumento das vendas no período de dezembro foi um tanto quanto exorbitante e podemos ter como base que é o mês com duas datas comemorativas como o natal e também ano novo. Logo as vendas naturalmente seriam maiores, em seguida no gráfico 2, fizemos a análise de quantidade de vendas referente ao dia do mês e como no gráfico 1 apresenta um aumento no mês de dezembro, os dias que tiveram mais realizações de pedidos foram na terça-feira e quarta-feira que são dias no meio da semana e tiveram de que como datas de pagamento do mês de dezembro. Já no gráfico 3, extraímos a quantidade de vendas por região em um gráfico no formato pizza, para obtermos uma visualização ampla das cidades, sendo a principal delas localizada na Índia. Por fim no gráfico 4, finalizamos com a consulta de sucesso das entregas, como já era de se esperar tivemos bem mais registros de entregas realizadas com sucesso do que mercadorias que retornaram ao seus vendedores.

Começando a tratativa da segunda tabela, tivemos os dados da plataforma *Amazon*, porém apenas de um determinado mês que era junho de 2021, onde podemos levar em conta que era época de pandemia (Covid-19) e também tínhamos como dados bem importantes as categorias dos produtos vendidos, sendo classificados por: eletrônicos, roupas, sapatos, acessórios, cartões de presentes, livros, vídeo-games, câmeras, fotos, brinquedos e jogos. Logo após a formatação da planilha, resolvemos usar a coluna de categoria para separar uma análise mais concreta sobre os dados. Sendo assim, conforme no gráfico 5 começamos a agrupar as categorias por visualizações e analisar pelo gráfico de barras, a quantidade das mesmas, em seguida geramos o gráfico pizza com os mesmos valores para obter a porcentagem de visualização de cada categoria. Depois usamos o agrupamento de categorias por número de vendas e tivemos uma visualização bem melhor no gráfico pizza do que no de barras, pois como se trata de vendas a análise pela porcentagem ficaria mais clara, além de determinar melhor precisão sobre uma determinada categoria e como era de se esperar, tivemos a categoria de vídeo games como líder na demanda de vendas, e claramente pelo fato de ser um produto alvo de um público geral.

E por fim, nesta segunda tabela decidimos realizar alguns testes para saber se o gráfico seria preciso como os outros, tanto alterando a função quanto alterando os valores obtidos, e infelizmente não tivemos um resultado interessante, pois, conforme no gráfico 9, onde agrupamos categoria e número de visualizações na função *px.line* foi gerado um gráfico de forma complexa para leitura, e no gráfico 10, tentamos realizar o agrupamento de categoria

por número de vendas porém na função *px.line* nos retorna de forma bagunçada a quantidade de vendas, justamente pela planilha não disponibilizar a data das vendas para um gráfico mais organizado.

A análise da terceira tabela foi feita da planilha de vendas das lojas Americanas, o resultado dessa planilha teve o total de todas as vendas feitas a partir do primeiro semestre de 2021 ao último semestre, que seria o quarto semestre de 2021, todos os totais de vendas foram em resultados financeiros de bilhões em vendas, sendo que o semestre que mais teve aumento de vendas foi o quarto, resultando em total de quase ate 7 bilhões em vendas. Isso porque como mostra no décimo segundo gráfico que as vendas digitais sempre esteve dominando as vendas em época de pandemia (Covid-19) com muita diferença entre as vendas físicas, e como mostra no décimo segundo gráfico, a diferença quase sempre foi de até 2 bilhões em vendas, sem as vendas de mídia digital sair do topo. Com os resultados, percebemos que nesses semestres de 2021 fez uma grande diferença no *E-commerce* que vem se tornando líderes do mercado, a necessidade e o aumento na confiança de compras digitais levaram a experimentarem ainda mais o mundo *online* sem precisar sair de sua casa e fazer compras.

6. CONSIDERAÇÕES FINAIS

O objetivo geral deste trabalho, foi obter conhecimento sobre a tratativa de dados e também analisar alguns conceitos sobre análise de dados, além de extrair conhecimentos com bases em arquivos textuais. Nesse modelo, foi feito o estudo de vários conceitos incluindo o KDD que basicamente serve para auxiliar na busca de dados, definir e padronizar elementos. E também o conceito dos 5 V 's da Big Data, pois se trata de uma área que realiza o estudo de como tratar, analisar e adquirir informações em grande escala para serem analisados.

Com o uso da linguagem *python* e suas bibliotecas como *pandas*, *numpy*, *matplotlib*, *plotly*, conseguimos fazer a tratativa de alguns dados referentes ao comércio eletrônico, pois, há pretensão de aplicar o conhecimento obtido nas devidas empresas que atuamos hoje. Aplicamos a análise em três tabelas diferentes porém utilizando basicamente o mesmo algoritmo, alterando apenas alguns parâmetros.

As tabelas foram disponibilizadas pela plataforma *Kaggle*, de todas as tabelas tinha o segmento de dados de vendas sob o *e-commerce*, a primeira tabela era voltada para vendas em geral do comércio eletrônico, porém, na tratativa da mesma apenas fornecia o nome dos produtos e não a categoria dos mesmos, no entanto, ao pesquisar manualmente descobrimos que boa parte se tratava de produtos eletrônicos então tivemos a idéia de fazer a análise voltada para este tipo de segmento, e durante o processo da análise dos gráficos gerados na primeira tab, encontramos uma forte ligação com as vendas do mês de dezembro e também coincidentemente era o mês de datas festivas como Natal e Ano Novo.

Antes do início da tratativa da segunda tabela, vale ressaltar que tentamos buscar uma tabela que possua o mesmo segmento da primeira, porém infelizmente como não tivemos sucesso na busca, resolvemos utilizar uma base de vendas da amazon referente ao mês de Junho em 2021, não era uma base completa, entretanto possuía o segmento de dados que estávamos buscando, logo ao iniciar a análise da tabela fora da aplicação do código vimos que a categoria dos produtos era do mesmo grupo que os produtos da tabela 1, então após a aplicação do código utilizado na tabela 1, porém com alterações nos parâmetros, conseguimos gerar os principais gráficos com o principal agrupamento voltado para categoria, e assim analisando quantas vendas, visualizações e avaliações tivemos por categoria podemos tirar como resultado que nem sempre os produtos que são mais visualizados ou tem mais avaliações são os de maior quantidade de vendas, segue como exemplo no gráfico 7 e no gráfico 8, onde podemos ver que os produtos da categoria Vídeo Games estão no top 1 das vendas do mês. Nos dando a informação de que é um produto que agrada a basicamente quase todos os públicos.

E por fim na análise da segunda tabela, decidimos aplicar alguns testes usando o formato de uma tabela diferente e também com categorias diferentes como no gráfico 9 e gráfico 10, e vimos que é gerado de uma forma que dificulta a análise dos mesmos.

Na terceira tabela, para podermos fazer uma análise um pouco diferente da primeira e da segunda, porém com o mesmo segmento. Resolvemos utilizar um algoritmo com alguns parâmetros diferentes, a princípio a idéia era tentar gerar os gráficos sem a intenção de aplicar nenhum *dataframe* para edição dos valores, porém, a biblioteca *matplotlib*, não consegue encaixar caracteres especiais no processo de gerar o gráfico. Tendo em vista a dificuldade de gerar os gráficos de forma "manual", decidimos formatar fileira por fileira e aplicar os resultados no algoritmo para gerar o gráfico mostrando o total das vendas de mercadorias nas mídias físicas e digitais.

No estudo do processo da tratativa de dados, podemos adquirir não só o conhecimento de quantos produtos ou até mesmo em qual período teve maior quantidade de vendas, mas também a ideia de aplicação de estratégias empresariais para melhoria dos fluxos de venda e recorrências.

Realizada essa pesquisa, podemos obter como conhecimento como funciona a tratativa dos dados e também como aplicar a mesma para melhorias ou soluções empresariais e conseguimos compreender como funciona o processo da tratativa de dados em tabelas que são geradas diretamente pelo banco de dados da empresa.

REFERÊNCIAS BIBLIOGRÁFICAS

ASSUNÇÃO, W. Comércio Eletrônico. Grupo A, 2018.

AFFONSO, A. Como a Mineração de Dados Pode te Ajudar a Alcançar Melhores Resultados. Voitto, 2021. Disponível em: < https://www.voitto.com.br/blog/artigo/mineracao-de-dados >. Acesso em: 20/08/2021.

BOENTE, A. N. P.; GOLDSCHMIDT, R. R.; ESTRELA, V. V. Uma metodologia de suporte ao processo de descoberta de conhecimento em bases de dados. Disponível em: < http://boente.eti.br/publica/seget2008kdd.pdf > Acesso em: 10/12/2021.

DALFOVO, O. Modelo de integração de um sistema de inteligência competitiva com um sistema de gestão da informação e de conhecimento. UFSC; 2007.

Ebit Nielsen. Plataforma de opinião de consumidores do Brasil. Disponível em: < https://www.ebit.com.br/ >.

E-commerce Brasil. E-commerce Brasileiro cresce 73,88% em 2020, revela índice MCC-ENET. E-commerce Brasil, 2021. Disponível em: <https://www.ecommercebrasil.com.br/artigos/comercio-eletronico-antes-e-depois-da-pandemia-do-coronavirus/ >

FERNANDO, A. Aprenda Mineração de dados: Teoria e Prática. Rio de Janeiro: Alta Books Editora, 2016.

GIBBS, G. Análise de dados Qualitativos. Porto Alegre: Artmed Editora S.A., 2009.

GOLDSCHMID R.; PASSOS E. Data mining Um guia prático. São Paulo: Elsevier Editora Ltda.; 2005.

KUNIYOSHI, M. S. Comércio Eletrônico: A Revolução em Tempos Digitais. Revistas PUCSP, 2000. < https://revistas.pucsp.br/index.php/rad/article/view/1689/1083 > . Acesso em: 01/10/2021.

LARRY, U. E-commerce com PHP e MySQL,Novatec Editora, 2014.

LUIZ, E. B. Introdução à Programação com Python: Algoritmos e Lógica de Programação Para Iniciantes, Novatec Editora, 2009.

MCC-ENET. Referência em métricas e indicadores do consumo online no Brasil. Disponível em < https://www.mccenet.com.br/ >.

MOREIRA, P. Comércio Eletrônico: Antes e Depois da Pandemia do Coronavírus. E-commerce Brasil,2020.<https://www.ecommercebrasil.com.br/artigos/comercio-eletronico-antes-e-depois-da-pandemia-do-coronavirus/ >. Acesso em: 25/09/2021.

NOLETO, C. Mineração de dados: O que é e como funciona o Data Mining. Blog da Trybe, 2021. < https://blog.betrybe.com/tecnologia/mineracao-de-dados/ > Acesso em: 28/09/2021.

SAS Insights. Mineração de dados, o que é e qual sua importância? Disponível em: <https://www.sas.com/pt_br/insights/analytics/mineracao-de-dados.html#dmhistory%E2%80%AF>

TURCHI, S. R. Estratégia de Marketing Digital e E-Commerce, 2ª edição. Grupo GEN, 2018.

VIVEROS, M.S. et al. Applying data mining techniques to a health insurance information system. In: VLDB CONFERENCE, 22., 1996, Bombay. Proceedings… Bombay: IIT Bombay, 1996. p. 286-295.

WES MCKINNEY, Python para Análise de Dados: Tratamento de Dados com Pandas, NumPy e Python, 2018.

ZAKI, M.;MEIRA JR.W. Data mining and analysis: fundamental concepts and algorithms. New York: Cambridge University Press, 2014.

ZIAFAT, H; SHAKERI, M. Using Data Mining Techniques in Customer Segmentation. Journal of Engineering Research and Applications, Volume 4, Fascículo 9, Página 70-79. Setembro 2014.
< https://www.ijera.com/papers/Vol4_issue9/Version%203/K49037079.pdf > .

Livro com até três autores:

BANIN, SÉRGIO L. *Python 3 - Conceitos e Aplicações - Uma abordagem didática*. Editora Saraiva, 2018.

OTAVIANO, A. M. H. VANIN R. V. Título: Tratamento Estatístico de Dados: Em Física Experimental. 2° Edição. São Paulo, 1991.

PERKOVIC, LJUBOMIR. Introdução à Computação Usando Python - um Foco no Desenvolvimento de Aplicações, 2016.

YASMINA SANTOS MARIBEL, ISABEL RAMOS. Business Intelligence. Tecnologias Da Informação Na Gestão De Conhecimento,2009.

Livro com quatro autores ou mais:

BARBOZA, FABRÍCIO FELIPE, M. e PEDRO HENRIQUE CHAGAS FREITAS. *Modelagem e desenvolvimento de banco de dados*, 2018.

FERREIRA, Rafael G C.; MIRANDA, Leandro B. A D.; PINTO, Rafael A.; Preparação e Análise Exploratória de Dados, 2021.

RAMESH SHARDA, DURSUN DELEN, EFRAIM TURBAN e ÂNGELA BRODBECK Business Intelligence e Análise de Dados para Gestão do Negócio,2019.

SILVA, L.A. D., PERES, S. M., & BOSCARIOLI, C. (2016). *Introdução à Mineração de Dados - Com Aplicações em R*.

VETORAZZO, Adriana de Souza; revisão técnica: MACHADO, Jeferson F. L. de Souza. Estrutura de Dados. Porto Alegre: SAGAH, 2018.