Fatec
Ipiranga

PROJETO DE ARQUITETURA DE BI E BIG DATA

Componentes do Grupo
Hamilton Alves da Silva,
Juscilene Cecilia S. Varandas,
Nicolas Yudji Kondo
Marcelo Augusto Luvizutto
Augusto Pinho de Freitas

Professor IZAIAS PORFIRIO FARIA

Hamilton Alves da Silva,
Juscilene Cecilia S. Varandas,
Nicolas Yudji Kondo
Marcelo Augusto Luvizutto
Augusto Pinho de Freitas

PROJETO DE ARQUITETURA DE BI E BIG DATA
GCP - Google Cloud Platform

Trabalho para a conclusão da
matéria de Arquitetura de Big Data e Projeto Integrador IV,
feito em conjunto com todas as
disciplinas

Professor Izaias Porfirio faria

São Paulo - SP
Junho/2024

RESUMO

Este trabalho apresenta a implementação de um processo de Extração, Transformação e Carga (ETL) utilizando ferramentas como Google Cloud Storage, Google Colab, Python e Pandas. O objetivo é demonstrar a viabilidade e eficiência dessas ferramentas no tratamento de grandes volumes de dados, focando em conjuntos de dados de vendas de bebidas alcoólicas em diversos países, dados de e-commerce, dados de clientes de cartões de crédito e localizações de lojas. Também exploraremos a utilização do Google Cloud Platform (GCP) para Machine Learning.

Palavras-chave: ETL, Google Cloud Storage, Google Colab, Python, Pandas, BigQuery, Machine Learning, Kaggle.

ABSTRACT

This work presents the implementation of an Extract, Transform, and Load (ETL) process using tools such as Google Cloud Storage, Google Colab, Python, and Pandas. The goal is to demonstrate the feasibility and efficiency of these tools in handling large volumes of data, focusing on datasets of alcoholic beverage sales, e-commerce data, credit card customer data, and store locations. We will also explore the use of Google Cloud Platform (GCP) for Machine Learning.

Keywords: ETL, Google Cloud Storage, Google Colab, Python, Pandas, BigQuery, Machine Learning, Kaggle.

INTRODUÇÃO

O presente trabalho tem como principal objetivo emular um banco de dados de e-commerce a fim de gerar insights e demonstrativos analitcs de como funciona um e-commerce no varejo de nível mundial.

A análise de dados é uma disciplina essencial para organizações que buscam obter insights valiosos e tomar decisões estratégicas fundamentadas. O processo de Extração, Transformação e Carga (ETL) é uma etapa crucial dentro desse contexto, pois permite a integração e preparação dos dados provenientes de diversas fontes, garantindo sua disponibilidade, confiabilidade e eficiência.

Este trabalho apresenta a implementação de um processo ETL robusto e eficaz utilizando as tecnologias Google Cloud Platform (GCP) e Dataproc, combinadas com ferramentas como Google Colab, Pandas e Python. O objetivo principal é transformar dados brutos em informações acionáveis, viabilizando análises abrangentes e a geração de insights valiosos.

O projeto integra e transforma dados de várias fontes em um ambiente de nuvem e em um ambiente distribuído. Utilizando Google Cloud Storage para o armazenamento, Google Colab para a execução de notebooks interativos, e Python com a biblioteca Pandas para a manipulação e transformação dos dados, garantimos um processo ETL eficiente e escalável. Além disso, exploramos o uso de Machine Learning (ML) no Google Cloud Platform (GCP) para enriquecer a análise de dados e gerar insights adicionais.

A escolha das ferramentas e tecnologias mencionadas não é arbitrária; cada uma desempenha um papel fundamental na cadeia de processamento de dados. O Google Cloud Storage oferece uma solução de armazenamento escalável e confiável, enquanto o Google Colab proporciona um ambiente interativo e colaborativo para o desenvolvimento de scripts e análises. O Pandas, com sua poderosa funcionalidade de manipulação de dados, é ideal para a transformação e preparação dos dados.

Este trabalho busca demonstrar a importância de um processo ETL bem estruturado e como a combinação das ferramentas e tecnologias mencionadas pode resultar em um sistema eficiente e eficaz para a análise de dados. A implementação prática será descrita detalhadamente, destacando os desafios enfrentados e as soluções adotadas, com o intuito de servir como referência para futuros projetos de análise de dados. Além disso, a aplicação de Machine Learning no GCP será explorada, evidenciando como essa abordagem pode agregar valor ao processo de tomada de decisões.

ARQUITETURA DE BIG DATA

Para gerenciar grandes volumes de dados de maneira eficiente, a arquitetura de Big Data é fundamental. Essa arquitetura consiste em organizar sistemas e tecnologias que permitem a coleta, armazenamento, processamento e análise de dados em grande escala. Durante o processo de pesquisa foram exploradas algumas ferramentas da AWS Amazon Web Services.

AMAZON WEB SERVICES (AWS)

A Amazon Web Services (AWS) é a plataforma de nuvem mais adotada e mais abrangente do mundo, oferecendo mais de 200 serviços completos de data centers em todo o mundo. Milhões de clientes, incluindo as startups que crescem mais rápido, as maiores empresas e os maiores órgãos governamentais, estão usando a AWS para reduzir custos, ganhar agilidade e inovar mais rapidamente (AWS, 2024).

Abaixo algumas ferramentas apresentadas e suas funcionalidades:

AWS Lambda é um serviço de computação que executa seu código em resposta a eventos e gerencia automaticamente os recursos de computação, tornando-se a maneira mais rápida de transformar uma ideia em aplicações de produção modernas e com tecnologia sem servidor (AWS, 2024).

Kinesis é um serviço de streaming de dados totalmente gerenciado pela Amazon Web Services (AWS). Ele permite a ingestão, processamento e análise em tempo real de grandes volumes de dados de streaming, como logs, métricas, transmissões de mídia e dados de IoT. Com o Kinesis, as empresas podem extrair insights valiosos dos seus dados em tempo real, tomar decisões mais rápidas e tomar ações imediatas (AWS, 2024).

AWS Glacier é um serviço de armazenamento de arquivos em nuvem de baixo custo que fornece armazenamento seguro e durável para arquivamento de dados e backup online. Para manter os custos baixos, o S3 Glacier fornece três classes de armazenamento, de alguns milissegundos a horas. O S3 Glacier Flexible Retrieval e o S3 Glacier Deep Archive oferecem opções adicionais com base na rapidez com que você precisa restaurar os dados (AWS, 2024)

AWS S3 (Amazon Simple Storage Service) é um serviço de armazenamento de objetos que oferece escalabilidade, disponibilidade de dados, segurança e performance líderes do setor (AWS, 2024)

EMR (anteriormente chamado de Amazon Elastic MapReduce) é uma plataforma de cluster gerenciada que simplifica a execução de estruturas de big data, como Apache Hadoop e Apache Spark, para processar e analisar grandes quantidades de dados (AWS, 2024).

Sage Maker é um serviço totalmente gerenciado para criar, treinar e implantar modelos de ML com infraestrutura, ferramentas e fluxos de trabalho também totalmente gerenciados (AWS, 2024).

O **Apache Hadoop** é um projeto de software de código aberto que pode ser usado para processar de modo eficiente grandes conjuntos de dados. Em vez de usar um grande computador para processar e armazenar os dados, o Hadoop permite o agrupamento de hardware padrão em clusters para analisar em paralelo grandes conjuntos de dados (AWS, 2024).

Redshift é um data warehouse em nuvem totalmente gerenciado e escalável que acelera a geração de insights com análises rápidas, fáceis e seguras em grande escala (AWS, 2024).

RDS (Amazon Relational Database Service) é um serviço web que facilita a configuração, a operação e a escalabilidade de um banco de dados relacional na AWS nuvem. Se você for um Amazon RDS usuário, poderá usar Amazon Kendra para indexar sua fonte Amazon RDS (MySQL) de dados

AWS Glue é um serviço de integração de dados sem servidor que facilita descobrir, preparar e combinar dados para análise, machine learning (ML) e desenvolvimento de aplicações.

Amazon DynamoDB é um serviço de banco de dados NoSQL sem servidor que permite desenvolver aplicações modernas em qualquer escala. Como um banco de dados sem servidor, você paga apenas pelo que usa. Além disso, o DynamoDB é escalável até zero, não tem inicialização a frio, upgrades de versão, janelas de manutenção, aplicação de patches nem manutenção com tempo de inatividade (AWS, 2024).

O **Amazon Athena** é um serviço de consultas interativas que facilita a análise de dados diretamente no Amazon Simple Storage Service (Amazon S3) usando SQL padrão (AWS, 2024)

AWS QuickSight é um serviço de BI baseado em ML escalável, sem servidor (AWS, 2024).

TEMA

O comércio eletrônico, ou e-commerce, tem se consolidado como uma força motriz na economia global, transformando a maneira como produtos e serviços são comprados e vendidos. Este fenômeno representa uma mudança significativa nas interações comerciais, trazendo consigo uma série de vantagens e desafios. No contexto deste trabalho, o e-commerce será analisado sob a perspectiva da integração e transformação de dados para otimização de operações e estratégias empresariais.

OBJETIVOS GERAL E ESPECÍFICO

Objetivo Geral:

- Implementar um processo de Extração, Transformação e Carga (ETL) eficiente e escalável utilizando tecnologias de nuvem e ferramentas avançadas, com o objetivo de integrar e transformar dados de várias fontes no contexto de um sistema de e-commerce.

Objetivos específicos:

- Aplicar Machine Learning no GCP
- Garantir a Escalabilidade e Eficiência
- Gerar Insights para a Tomada de Decisões

JUSTIFICATIVA

A escolha do tema e-commerce para este trabalho é motivada por vários fatores que destacam a relevância e a importância desta área no cenário atual. O comércio eletrônico tem se estabelecido como uma das principais vertentes do comércio global, impulsionado pelo avanço das tecnologias digitais e pela mudança nos hábitos de consumo dos consumidores. Esta seção apresenta os principais motivos que justificam a seleção do tema e-commerce, com foco na integração e transformação de dados para otimização de operações e estratégias empresariais.

METODOLOGIA

Nossa metodologia está dividida em várias etapas, abrangendo desde a ingestão dos dados até a visualização e análise:

- Ingestão dos Dados: Extração dos dados a partir do Google Cloud Storage e carregamento no Google Colab.
- Tratamento dos Dados: Remoção de valores nulos, filtragem, renomeação de colunas e ajuste de tipos de dados utilizando Python e Pandas.
- Armazenamento dos Dados: Armazenamento dos dados tratados no Google Cloud Storage e no BigQuery.
- Visualização e Análise: Utilização de ferramentas de visualização para gerar insights a partir dos dados tratados.
- Machine Learning: Aplicação de técnicas de Machine Learning utilizando os serviços do GCP para obter previsões e análises avançadas.

FUNDAMENTAÇÃO TEÓRICA

E-commerce de varejo

O comércio eletrônico refere-se à compra e venda de bens e serviços através de plataformas digitais. Nos últimos anos, a popularidade do e-commerce cresceu exponencialmente, impulsionada pela penetração da internet e pela mudança nos comportamentos dos consumidores. De acordo com estudos recentes, o e-commerce representa uma parcela significativa do comércio global, com expectativas de crescimento contínuo nos próximos anos (Smith & Anderson, 2021).

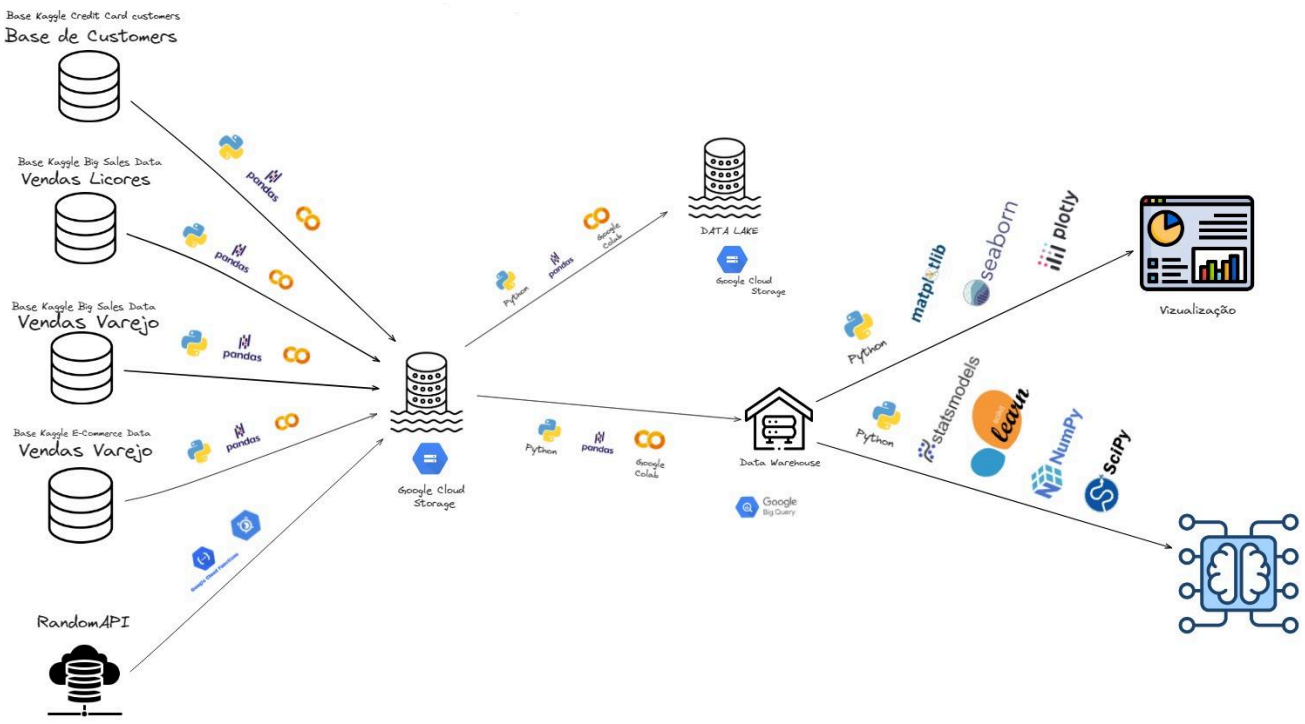
O e-commerce começou a ganhar tração na década de 1990 com o advento da internet pública. Inicialmente, as transações eram limitadas a simples vendas de produtos em sites estáticos. Com o desenvolvimento de tecnologias como SSL (Secure Sockets Layer) para segurança nas transações e a popularização de métodos de pagamento online, o e-commerce evoluiu rapidamente. Empresas pioneiras como Amazon e eBay estabeleceram modelos de negócios que se tornaram referência na indústria.

O futuro do e-commerce promete ser excitante, com tendências emergentes como o comércio social, onde as transações ocorrem diretamente em plataformas de mídia social, e o uso de IA para personalização e automação de processos. Além disso, o aumento do uso de tecnologias de voz e o crescimento do comércio móvel (m-commerce) são áreas que prometem transformar ainda mais o cenário do e-commerce.

PLANEJAMENTO DO PROJETO

A ideia inicial para fazer esse projeto seria usar o Amazon AWS (Amazon Web Services) para a implementação de um processo de Extração, Transformação e Carga (ETL), porém na realização acabamos mudando de ideia e usando o Google Cloud Storage no lugar. Primeiro escolhemos a base de dados sobre e-commerce, separamos as vendas de licores e usamos uma random API para gerar mais dados fictícios. Tratamos os dados para fazer as análises e vamos utilizar o machine learning com o objetivo de realizar uma previsão. Por último, vamos utilizar o BigQuery como data warehouse.

Fluxo de dados



Fonte: elaborada pelo autor.

FERRAMENTAS UTILIZADAS

As ferramentas utilizadas para a composição desse projeto são programas e linguagens.

Pandas é uma biblioteca de software de código aberto escrita para a linguagem de programação Python, amplamente utilizada em análise e manipulação de dados. Ela oferece estruturas de dados de alto desempenho, como DataFrames, que facilitam a manipulação e análise de dados tabulares. Com Pandas, é possível realizar operações como limpeza de dados, transformação, agregação e visualização, bem como importar e exportar dados de diferentes formatos, como CSV, Excel e SQL. (GOOGLE a).

Python é uma linguagem de programação de alto nível, versátil e de fácil leitura, possuindo uma vasta coleção de bibliotecas e frameworks, como Pandas, NumPy, Django e Tensor Flow, que expandem suas capacidades e permitem a resolução eficiente de problemas complexos. (GOOGLE a, 2024).

Google Cloud Storage é um serviço de armazenamento de objetos na nuvem oferecido pelo Google Cloud Platform, projetado para armazenar e acessar grandes volumes de dados de maneira escalável e segura. O serviço oferece alta durabilidade, disponibilidade e segurança, com opções de armazenamento de várias classes, otimizadas para diferentes necessidades de desempenho e custo. (GOOGLE a, 2024).

Ele permite que os usuários realizem consultas SQL em grandes conjuntos de dados com rapidez e eficiência, sem a necessidade de gerenciar infraestrutura. BigQuery é projetado para analisar terabytes e até petabytes de dados, facilitando a obtenção de insights acionáveis em tempo real.

Google Colab é um ambiente de notebooks Jupyter baseado na nuvem, oferecido gratuitamente pelo Google, que facilita a execução e compartilhamento de código Python. Ele é amplamente utilizado para análise de dados, aprendizado de máquina e desenvolvimento de projetos de inteligência artificial, permitindo que os

usuários executem código em GPUs e TPUs sem necessidade de configuração complexa.

INGESTÃO DE DADOS

A ingestão de dados é o processo de coletar e importar dados de várias fontes para um sistema de armazenamento ou banco de dados. Esse processo pode ser realizado em tempo real (*streaming*) ou em lotes (*batch*). Existem duas abordagens principais para a ingestão de dados:

Ingestão em Tempo Real (*Real-time Ingestion*):

a) Os dados são coletados e processados continuamente, quase ao mesmo tempo em que são gerados.

b) É útil para aplicações que requerem dados atualizados instantaneamente, como monitoramento de sensores IoT ou análises em tempo real.

EXTRACT, TRANSFORM, LOAD - ETL

É um processo crítico em pipelines de dados, responsável por mover e transformar dados de fontes diversas para um destino centralizado. O processo é dividido em três etapas principais:

a) Extração (*Extract*):

A primeira etapa do processo ETL consiste em extrair dados de várias fontes de dados, que podem ser bancos de dados relacionais, arquivos CSV, sistemas ERP, APIs, entre outros. Durante a extração, é importante garantir que os dados sejam recuperados de maneira eficiente e sem perda de informações.

b) Transformação *(Transform)*:

Após a extração, os dados são transformados para adequá-los ao formato e à estrutura desejada no destino. A transformação pode incluir limpeza de dados (remoção de duplicatas, correção de erros), agregação (soma, média), filtragem, enriquecimento (adição de dados adicionais), e transformação de formatos de dados.

Esta etapa é crucial para garantir que os dados estejam em um estado consistente e utilizável para análise.

c) Carregamento (*Load*):

A última etapa do processo ETL é carregar os dados transformados no sistema de destino, como um *data warehouse, data lake* ou banco de dados analítico. O carregamento pode ser feito de maneira incremental (apenas os dados novos ou alterados são carregados) ou em modo completo (todos os dados são recarregados).

É importante garantir que o processo de carregamento seja eficiente e cause o mínimo de interrupção possível ao sistema de destino.

ARMAZENAMENTO

O processo de armazenamento de dados é crucial para qualquer organização, pois permite a retenção, organização e proteção de informações essenciais.

DATA LAKE

As bases tratadas foram armazenadas no Google Cloud Storage novamente. Uma vez que esta é a ferramenta Data Lake escolhida para o desenvolvimento do projeto. Em um novo *bucket* foram salvos os arquivos tratados. Essa ação é realizada pela própria ferramenta.

COMPUTAÇÃO E PREPARAÇÃO

Nesta fase, o Machine Learning foi configurado para processar uma base de dados em um ambiente de programação Python. Para acessar essas bases de dados por meio do ambiente de programação, foi essencial mover as informações armazenadas na nuvem para o BigQuery. O BigQuery, que é uma plataforma de armazenamento de dados corporativa do Google Cloud, foi criado para facilitar a aquisição, armazenamento, análise e visualização de dados. A transferência de dados para o BigQuery pode ser feita por meio de carregamento em lotes ou streaming direto, permitindo a obtenção de insights em tempo real. Como o Google Cloud cuida da infraestrutura como um todo, você pode se concentrar em analisar seus dados em uma escala específica, incluindo.

IMPLEMENTAÇÃO

A análise de dados é crucial para organizações que buscam insights valiosos e tomadas de decisões estratégicas fundamentadas. Nesse contexto, o processo de Extração, Transformação e Carga (ETL) desempenha um papel essencial ao integrar e preparar dados de diversas fontes para análises abrangentes. Este projeto utiliza tecnologias como Google Cloud Platform (GCP) e Dataproc, aliadas a ferramentas como Google Colab, Pandas e Python, para criar um processo ETL robusto e eficiente.

O ambiente de desenvolvimento escolhido foi o Google Colab, proporcionando um ambiente interativo e colaborativo para o desenvolvimento de scripts em Python. Os dados foram extraídos de arquivos CSV armazenados no Google Cloud Storage utilizando a biblioteca Google Cloud Storage Client para Python. Em seguida, foram aplicadas várias etapas de transformação de dados utilizando Pandas, incluindo a remoção de valores nulos, filtragem de linhas relevantes, renomeação de colunas e ajuste de tipos de dados.

Após o tratamento dos dados, eles foram armazenados novamente no Google Cloud Storage e carregados no BigQuery para análises futuras. Durante esse processo, foram enfrentados desafios como garantir a integridade dos dados e

otimizar a eficiência na carga para o BigQuery. Para resolver esses desafios, foram utilizadas técnicas de limpeza e validação dos dados em cada etapa do processo, além de otimizações e configurações adequadas para maximizar a velocidade de carga no BigQuery.

Além do processo ETL tradicional, exploramos também a aplicação de Machine Learning (ML) no GCP para enriquecer a análise de dados e gerar insights adicionais. Isso incluiu o uso de técnicas de ML para análises preditivas e descoberta de padrões nos dados, agregando uma camada adicional de inteligência ao processo de análise de dados.

Em resumo, a implementação prática do processo ETL utilizando GCP, Dataproc, Google Colab, Pandas e Python demonstrou a eficiência e escalabilidade desse sistema para análise de dados. A combinação dessas tecnologias permitiu a integração e preparação de dados de forma robusta, preparando-os para análises profundas e a geração de insights valiosos para a tomada de decisões estratégicas.

VISUALIZAÇÃO DE DADOS

A etapa de visualização foi fundamental para a interpretação e análise dos dados tratados no projeto de ETL. Utilizando ferramentas como Matplotlib e Seaborn em Python, conseguimos criar gráficos e dashboards que proporcionam insights valiosos sobre os dados de vendas de licores, dados de e-commerce, dados de clientes de cartões de crédito e localizações de lojas.

Durante a análise exploratória, identificamos padrões significativos nos dados, como os produtos mais vendidos em diferentes regiões, a distribuição das vendas ao longo do tempo e o comportamento dos clientes. Essas análises foram cruciais para entender a dinâmica do mercado e identificar oportunidades de melhoria e otimização nas estratégias de vendas e marketing.

Além disso, a aplicação de técnicas de Machine Learning na visualização nos permitiu prever tendências futuras e identificar padrões ocultos nos dados, proporcionando uma visão mais profunda e abrangente do cenário analisado.

Em resumo, a visualização dos dados foi essencial para transformar números em informações acionáveis e insights valiosos. A combinação de ferramentas de visualização avançadas com análises detalhadas nos ajudou a extrair o máximo valor dos dados tratados, contribuindo para uma tomada de decisões mais fundamentada e estratégica para as organizações envolvidas.

MACHINE LEARNING

Os notebooks do Colab foram base para desenvolver o Machine Learning (ML). O ML foi a fundação das principais informações que contribuíram para tomada de decisão e análises futuras do e-commerce no varejo.

Usamos as bibliotecas Scikit-learn, Scipy, e Statsmodels, para o desenvolvimento do ML. Utilizamos as ferramentas, pandas para lidar com a base em um dataframe no python, matplotlib e seaborn para parte de gráficos e outras visualizações. Na parte de Machine Learning, usamos o ARIMA (AutoRegressive Integrated Moving Average), da StatsModels, é um modelo utilizado para a análise e previsão de séries temporais, fazer previsões para um período de teste e visualizar os futuros resultados. Usamos o algoritmo K-Means para agrupar os clientes em 5 clusters distintos e, em seguida, visualizar os resultados usando um gráfico de dispersão onde a cor indica a qual cluster cada cliente pertence. Isso pode ser útil para identificar diferentes segmentos de clientes e entender seus comportamentos de compra.

RESULTADOS E DISCUSSÕES

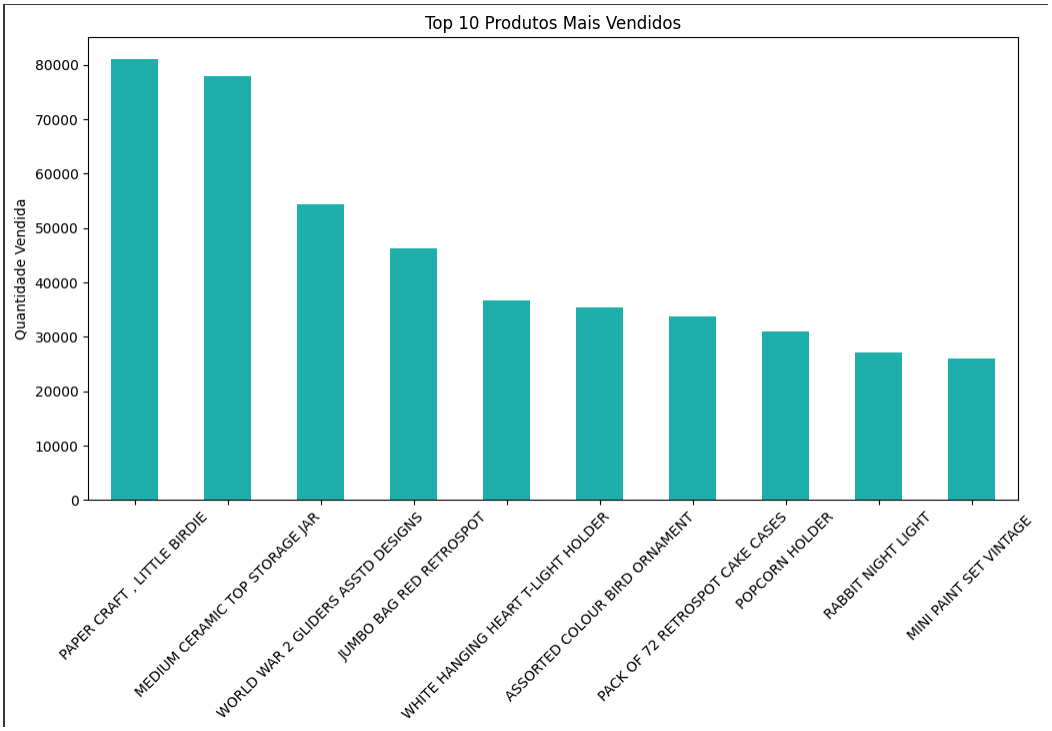
Os resultados obtidos ao longo deste projeto demonstram a eficiência e a capacidade das ferramentas e tecnologias utilizadas no tratamento de grandes volumes de dados. O processo de Extração, Transformação e Carga (ETL)

implementado mostrou-se robusto e eficaz ao lidar com diversos conjuntos de dados, provenientes de diferentes fontes, garantindo a disponibilidade, confiabilidade e eficiência necessárias para análises abrangentes. Reforçando a importância de um processo ETL bem estruturado e a eficácia das ferramentas e tecnologias utilizadas na análise de dados com a combinação de um processo ETL robusto, ferramentas avançadas de manipulação e visualização de dados, juntamente com a aplicação de Machine Learning, proporciona insights valiosos e contribuiu para uma tomada de decisões mais embasada e estratégica para as organizações em questão.

GRÁFICOS ESTATÍSTICOS

TOP 10 Produtos mais vendidos:

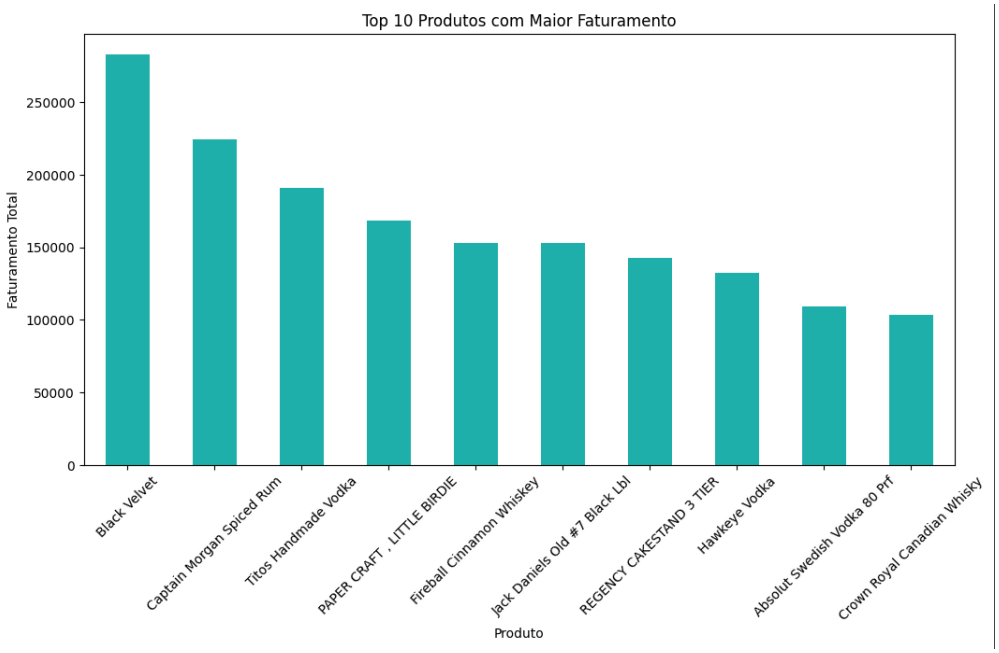
Identificação dos produtos com maior volume de vendas



Fonte: elaborada pelo autor.

TOP 10 Produtos com maior faturamento:

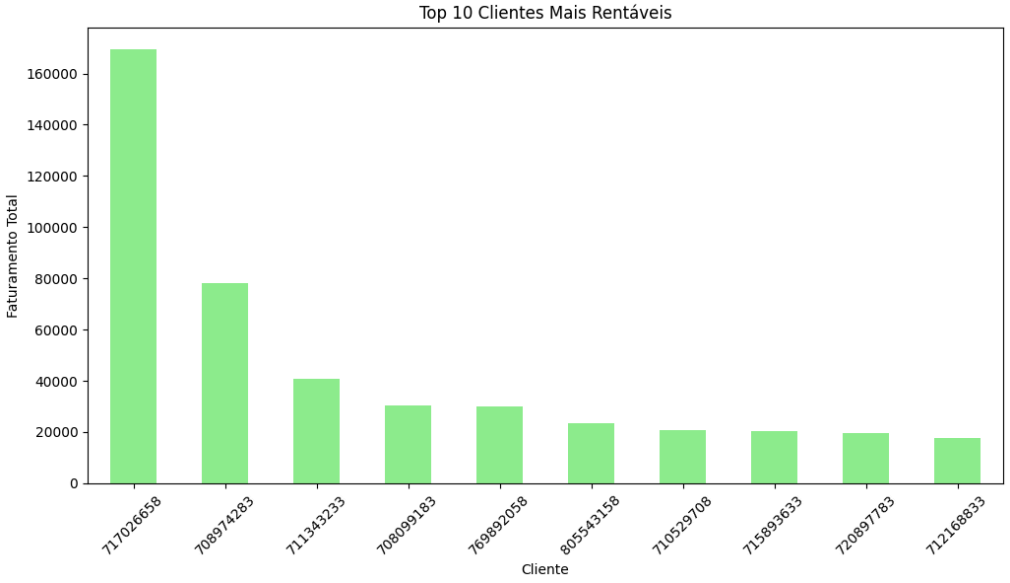
Destaque para os produtos que geram maior receita.



Fonte: elaborada pelo autor.

Clientes mais rentáveis:

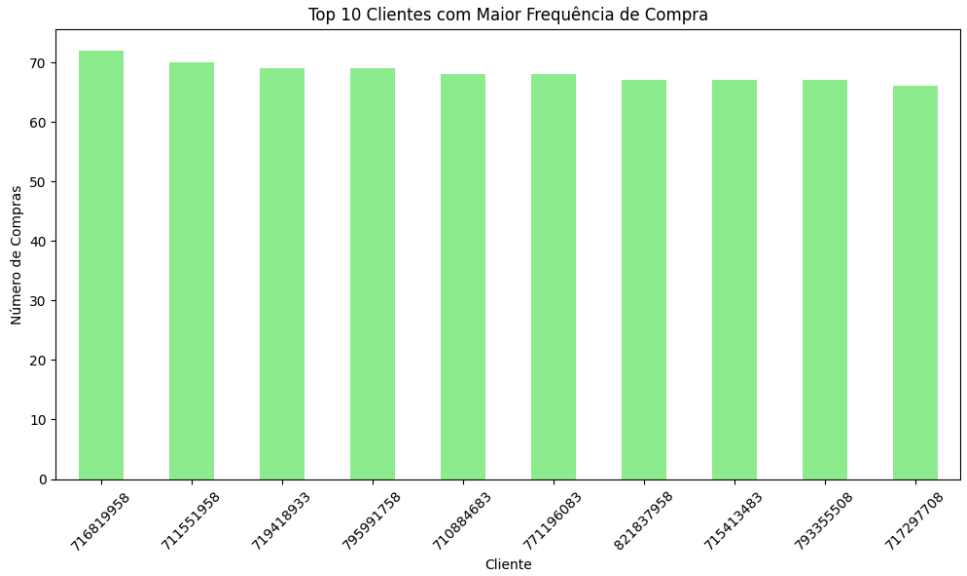
Segmentação dos clientes que proporcionam maior lucro.



Fonte: elaborada pelo autor.

Clientes com mais frequência de compra:

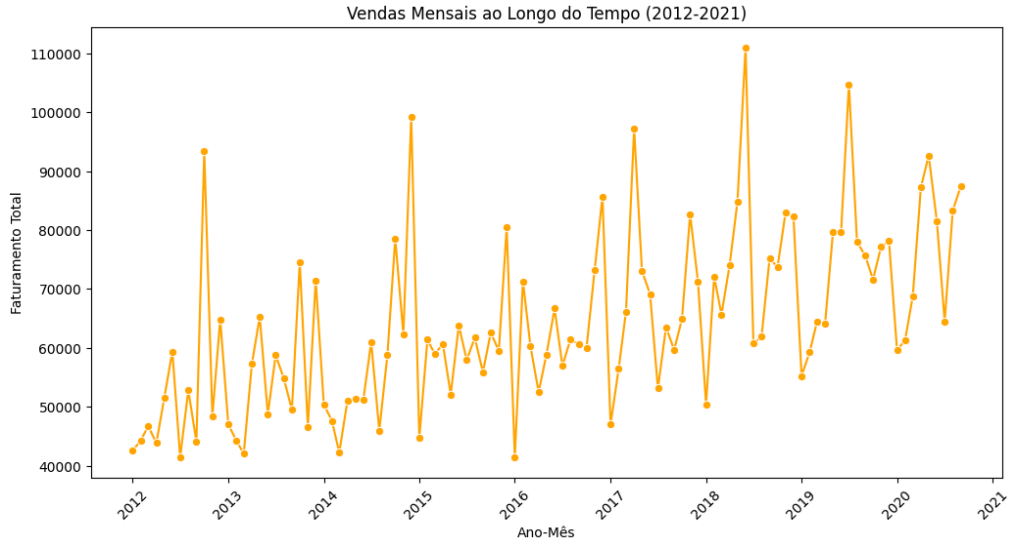
Análise dos clientes com maior recorrência de compras.



Fonte: elaborada pelo autor.

Vendas Mensais x Faturamento:

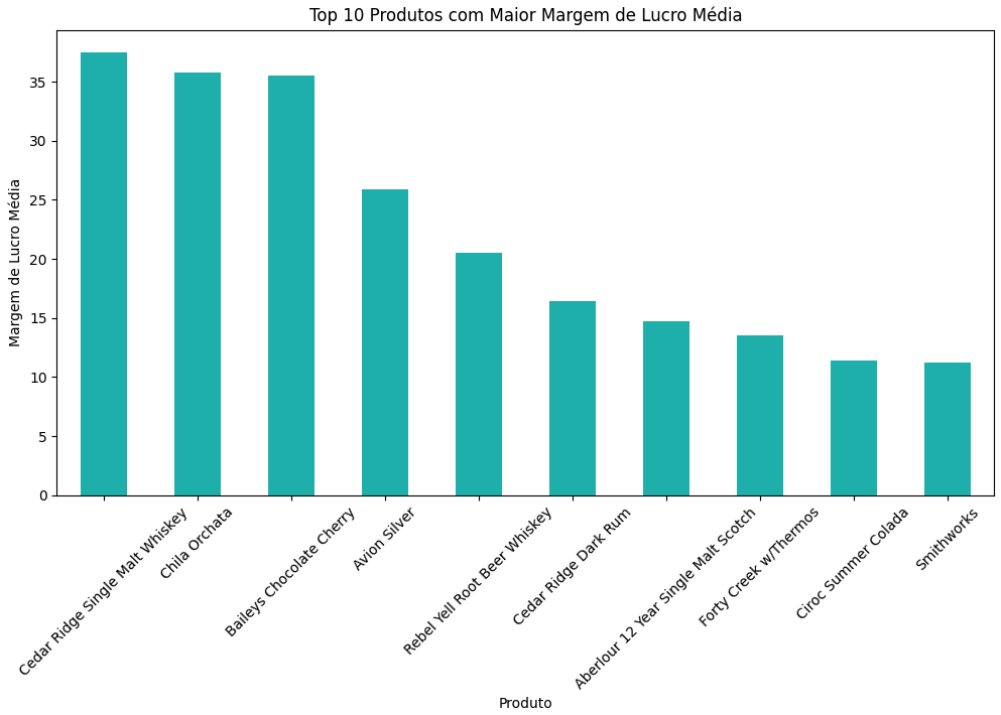
Correlação entre o volume de vendas e o faturamento mensal.



Fonte: elaborada pelo autor.

Produtos com maior margem de lucro:

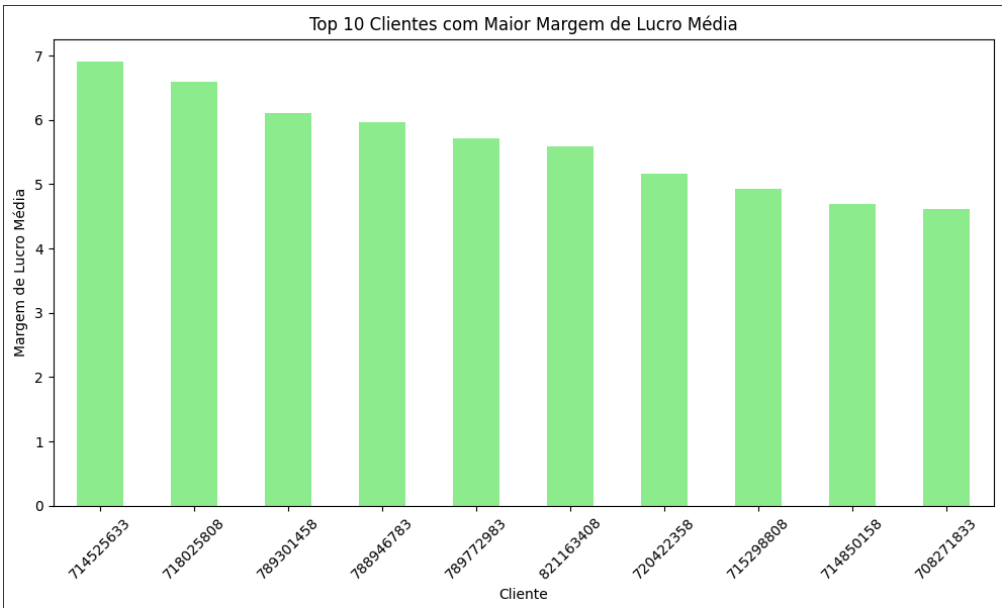
Produtos que apresentaram a maior margem de lucro.



Fonte: elaborada pelo autor.

Clientes com maior margem de lucro:

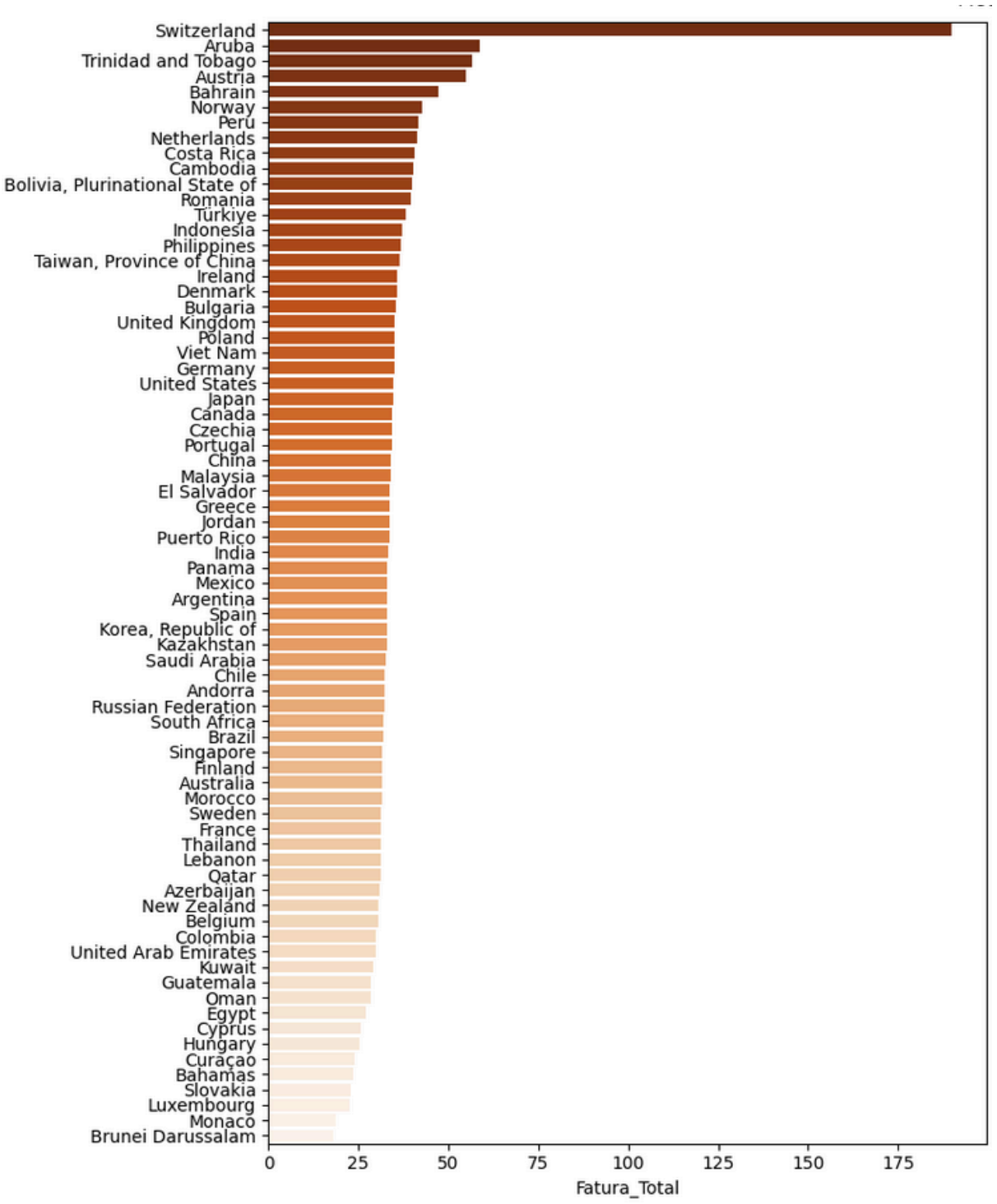
Identificação dos clientes que proporcionaram maior margem de lucro.



Fonte: elaborada pelo autor.

Fatura total por país:

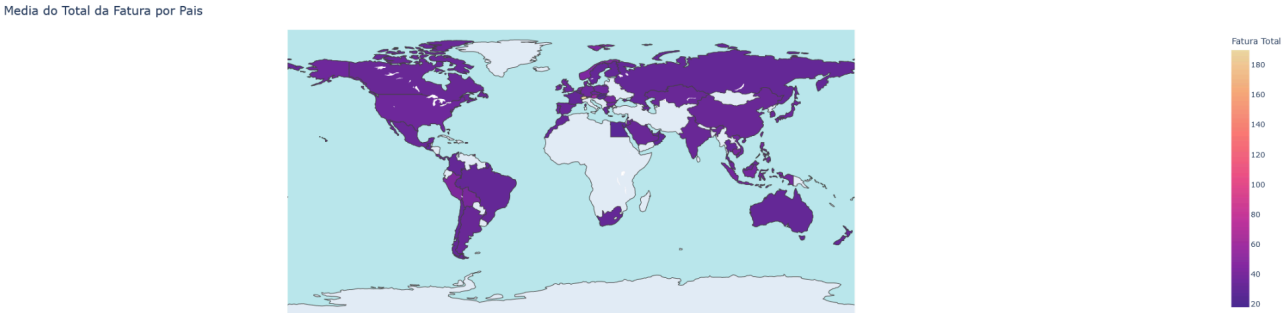
Comparativo do faturamento total por país.



Fonte: elaborada pelo autor.

Média de Fatura por mais:

Análise da média de faturamento mensal.



Fonte: elaborada pelo autor.

Melhores países por produto e % de vendas por países:

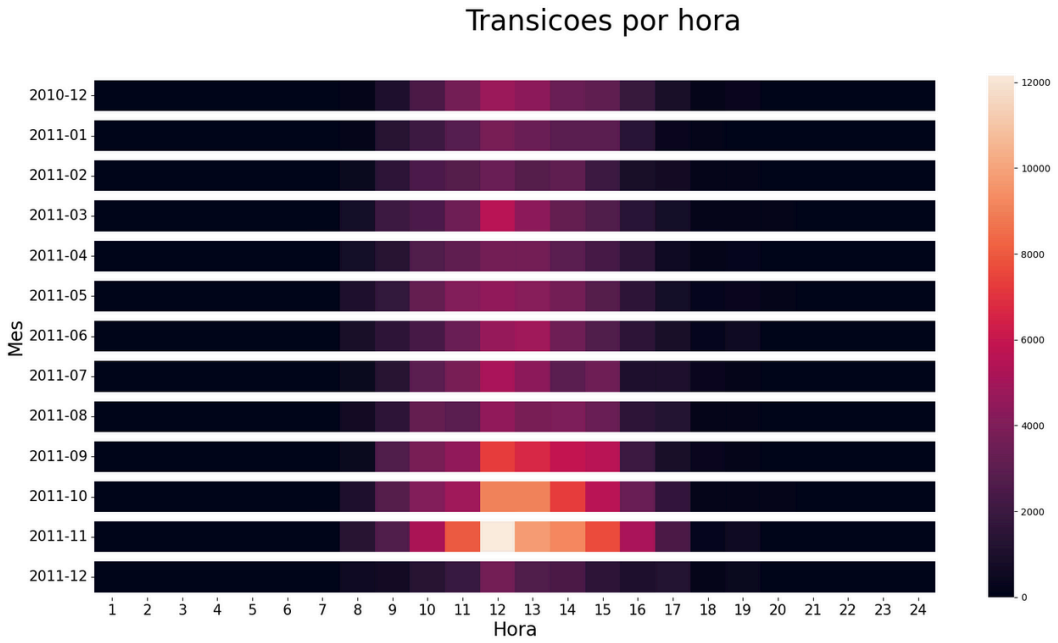
Identificação dos países com melhor desempenho por produto e a porcentagem de vendas por país.

	Pais	Melhor Produto	Vendas	Vendas total do pais	% Vendas do pais
0	Andorra	Bailey's Original Irish Cream	£311	£612	50.76%
1	Argentina	Fireball Cinnamon Whiskey	£1,320	£61,580	2.14%
2	Aruba	SPOTTY BUNTING	£813	£2,943	27.64%
3	Australia	Bacardi Superior	£1,080	£11,991	9.01%
4	Austria	Fireball Cinnamon Whiskey	£7,193	£17,161	41.91%
5	Azerbaijan	DOORMAT RED RETROSPOT	£270	£2,218	12.17%
6	Bahamas	Grey Goose Vodka	£333	£4,597	7.24%
7	Bahrain	Captain Morgan Spiced Rum	£4,860	£19,021	25.55%
8	Belgium	3 HOOK PHOTO SHELF ANTIQUE WHITE	£635	£11,236	5.65%
9	Bolivia, Plurinational State of	LUNCH BOX I LOVE LONDON	£532	£3,164	16.81%
10	Brazil	CREAM HEART CARD HOLDER	£2,297	£55,989	4.10%
11	Brunei Darussalam	Herradura Gold Reposado 6pak	£212	£1,425	14.89%
12	Bulgaria	VINTAGE CREAM DOG FOOD CONTAINER	£459	£3,291	13.95%
13	Cambodia	Patron Tequila Silver	£486	£2,212	21.97%
14	Canada	Titos Handmade Vodka	£43,278	£896,462	4.83%
15	Chile	Hawkeye Vodka	£2,975	£54,462	5.46%
16	China	Black Velvet	£29,354	£1,635,828	1.79%
17	Colombia	Smirnoff Vodka Traveller	£446	£5,953	7.49%
18	Costa Rica	Fireball Cinnamon Whiskey	£2,178	£7,415	29.37%
19	Curaçao	Jeffersons Ocean Aged at Sea	£382	£1,147	33.34%
20	Cyprus	Smirnoff PET 80prf	£266	£4,890	5.43%
21	Czechia	BLUE VINTAGE SPOT BEAKER	£2,028	£18,406	11.02%
22	Denmark	Absolut Swedish Vodka 80 Prf	£1,988	£13,037	15.25%
23	Egypt	LUXURY BATH BOMB SET	£1,007	£14,726	6.84%
24	El Salvador	MISELTOE HEART WREATH CREAM	£996	£7,658	13.01%
25	Finland	FELTCRAFT CUSHION BUTTERFLY	£651	£4,624	14.08%
26	France	PANTRY CHOPPING BOARD	£3,855	£73,949	5.21%
27	Germany	Titos Handmade Vodka	£4,746	£97,663	4.86%
28	Greece	Templeton Rye	£3,093	£16,772	18.44%
29	Guatemala	MINI PAINT SET VINTAGE	£317	£3,342	9.48%

Fonte: elaborada pelo autor

Mais transições por hora:

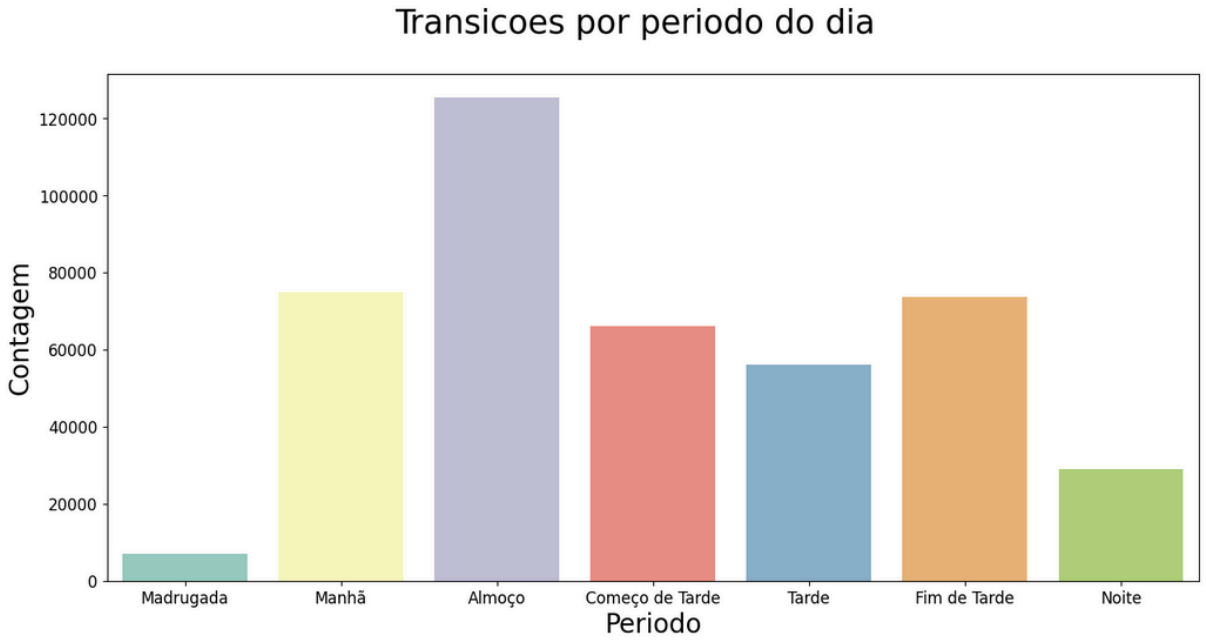
Análise das horas com maior número de transações.



Fonte: elaborada pelo autor.

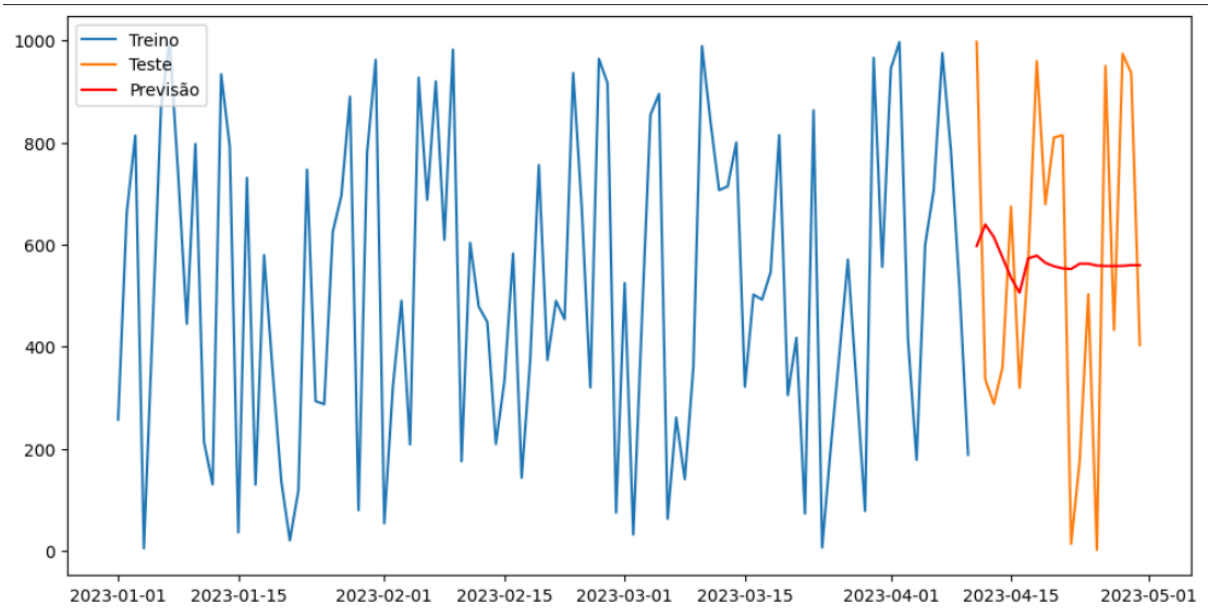
Melhores períodos:

Identificação dos períodos com melhor desempenho de vendas.



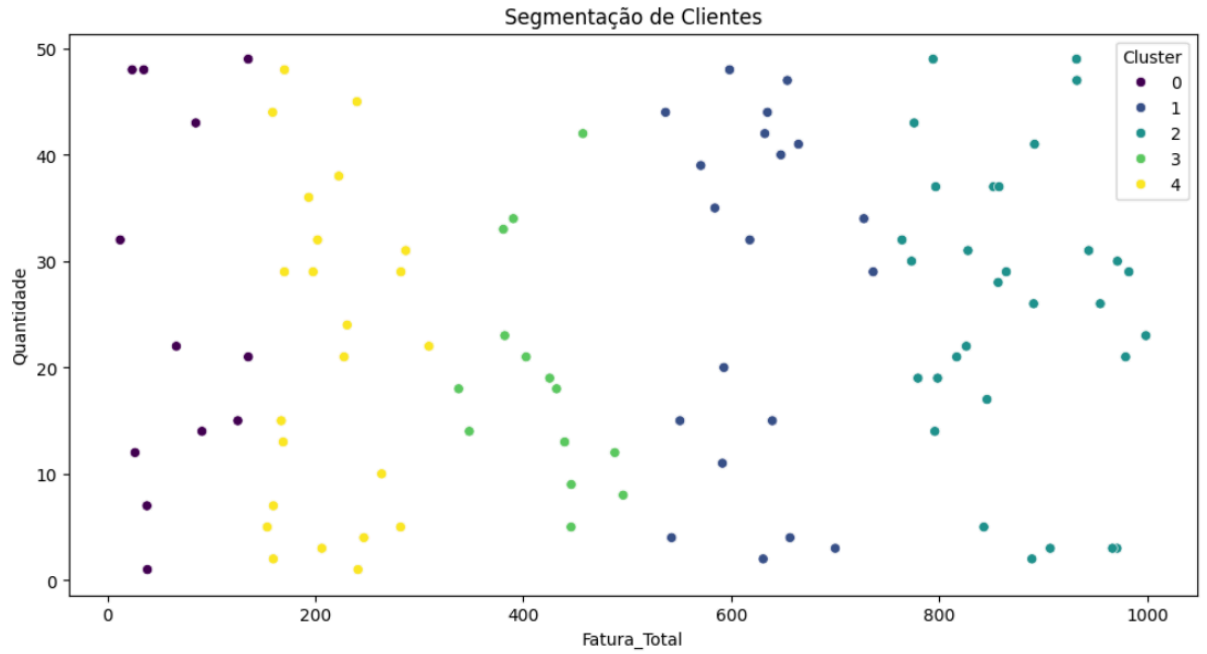
Fonte: elaborada pelo autor.

Machine Learning - Teste ARIMA, previsão de faturamento para os próximos anos: Projeções de faturamento com base em modelos preditivos.



Fonte: elaborada pelo autor.

Machine Learning - Segmentação de clientes:
Classificação de clientes com base em clusterização.



Fonte: elaborada pelo autor.

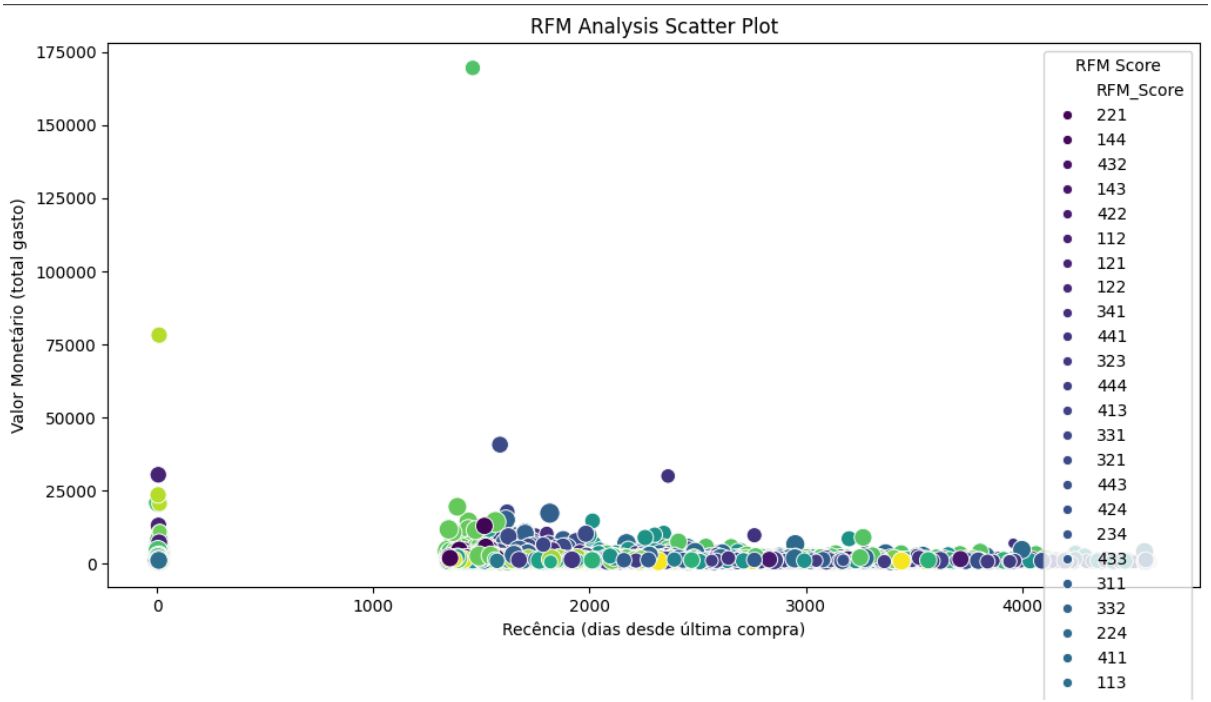
RFM Clientes

A Análise RFM é uma técnica de segmentação de clientes baseada em três métricas:

Recência: Quanto tempo passou desde a última compra.

Frequência: Quantas compras o cliente fez.

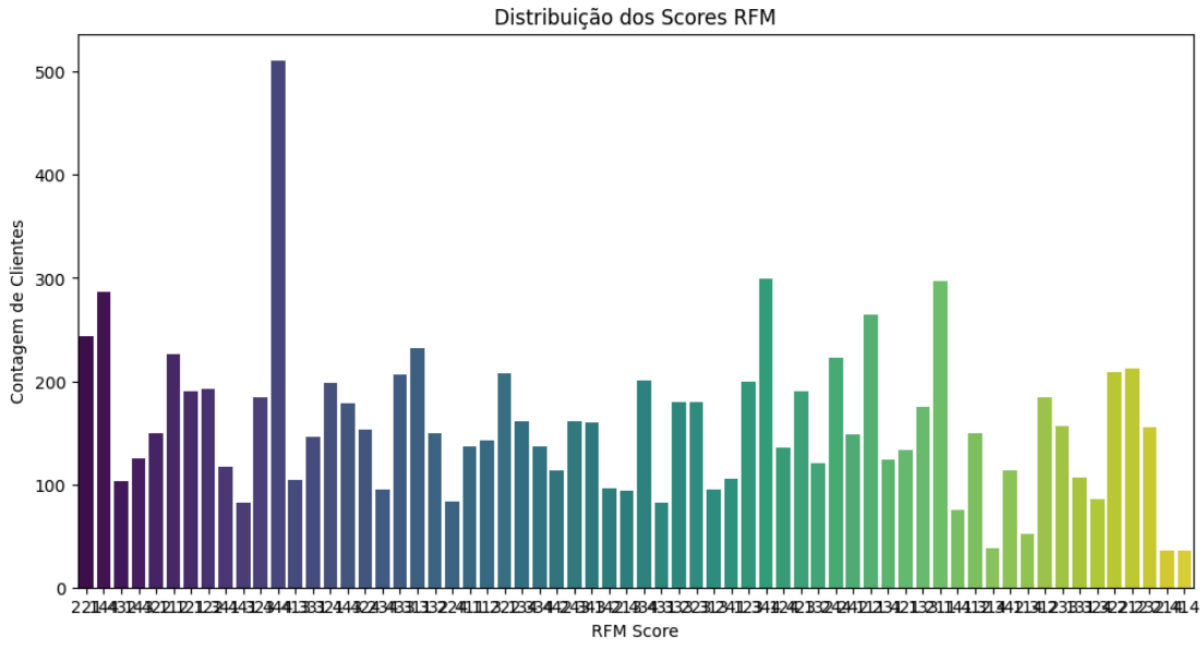
Valor Monetário: Quanto dinheiro o cliente gastou no total



Fonte: elaborada pelo autor.

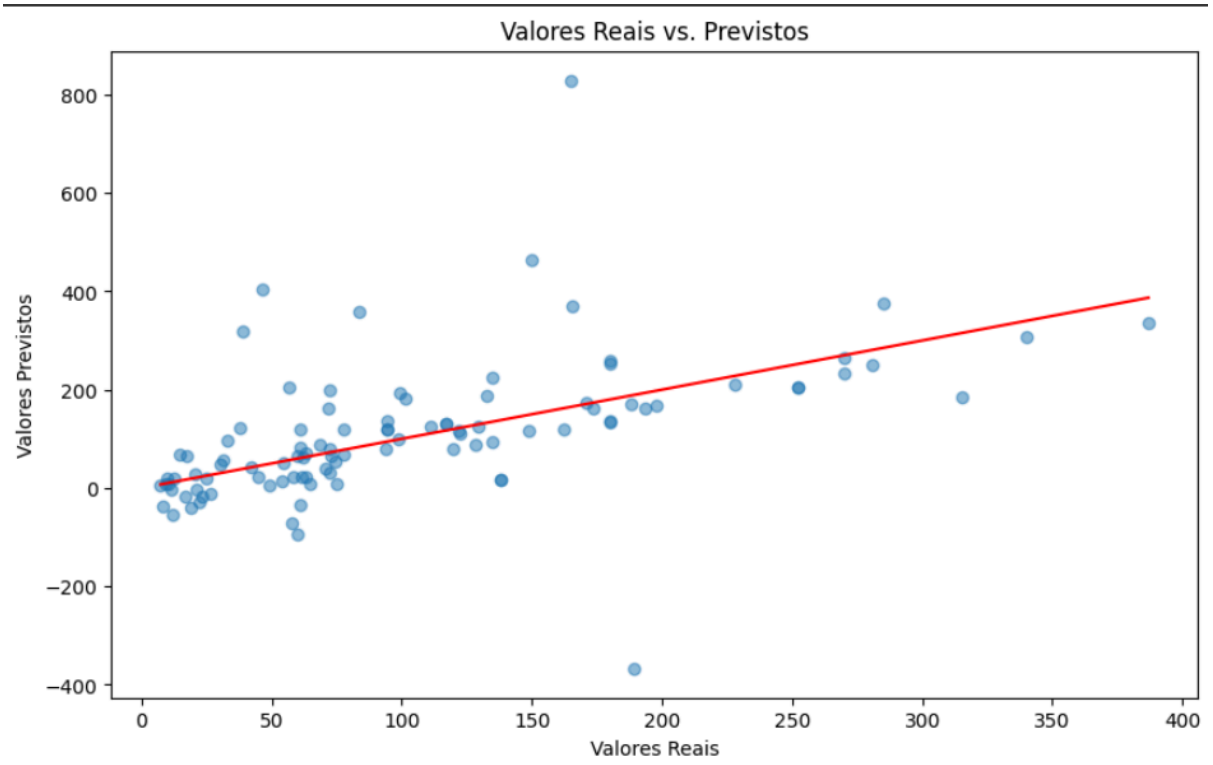
```
[10127 rows x 7 columns]
Clientes com os maiores RFM Scores:
             Recência  Frequência  Valor_Monetário R_Score F_Score M_Score  \
Cliente_Cod
771534408        3979          30           311.83       4       4       4
808273533        2789          28           462.19       4       4       4
720371733        3251          37          1002.24       4       4       4
716510658        3176          31           584.72       4       4       4
709888383        2333          40           830.75       4       4       4

            RFM_Score
Cliente_Cod
771534408         444
808273533         444
720371733         444
716510658         444
709888383         444
```

Fonte: elaborada pelo autor.



Fonte: elaborada pelo autor.

Valores reais vs Previsto



Fonte: elaborada pelo autor.

Conclusão

Este trabalho demonstrou a viabilidade e a eficácia da implementação de um processo ETL utilizando tecnologias de nuvem e ferramentas avançadas. A combinação do Google Cloud Platform, Google Colab, Python e Pandas mostrou-se eficiente para a integração e transformação de dados, proporcionando uma base sólida para análises abrangentes e a geração de insights valiosos.

A aplicação de Machine Learning no GCP agregou valor ao processo, permitindo a obtenção de previsões e análises avançadas. A metodologia adotada pode servir como referência para futuros projetos de análise de dados, evidenciando a importância de um processo ETL bem estruturado e a utilização de tecnologias adequadas para alcançar resultados satisfatórios.

REFERÊNCIAS

https://aws.amazon.com/pt/s3/?nc2=h_ql_prod_fs_s3 . Acesso em: 09 abr. 2024.

BARON, Joe; BAZ, Hisham; BIXLER, Tim; GAUT, Biff; KELLY, Kevin E.; SENIOR, Sean. AWS Certified Solutions Architect Official Study Guide: Associate Exam. 2. ed. New Jersey: Wiley, 2018.

ALTEEN, Nick; STAMPER, John; KHASHA, Kunal. AWS Certified Developer Official Study Guide: Associate Exam. 1. ed. New Jersey: Wiley, 2019.

WITTIG, Andreas; WITTIG, Michael. Amazon Web Services in Action. 2. ed. Shelter Island: Manning Publications, 2018.

PIPER, Ben; CLINTON, David. AWS Certified Solutions Architect Study Guide: Associate SAA-C01 Exam. 2. ed. New Jersey: Wiley, 2019.

ATCHISON, Lee. Architecting for Scale: High Availability for Your Growing Applications. 2. ed. Sebastopol: O'Reilly Media, 2020.

POCCIA, Danilo. AWS Lambda in Action: Event-Driven Serverless Applications. 1. ed. Shelter Island: Manning Publications, 2016.

WILKINS, Mark. Learning Amazon Web Services (AWS): A Hands-On Guide to the Fundamentals of AWS Cloud. 1. ed. Indianapolis: Addison-Wesley Professional, 2019.

ANTHONY, Albert. AWS Security Best Practices. 1. ed. Birmingham: Packt Publishing, 2018.

PALANISAMY, Praveen. Cloud Computing with AWS: Learn and implement cloud computing using AWS. 1. ed. Birmingham: Packt Publishing, 2018.

GOUNDER, Premkumar. DevOps on the AWS Cloud: Building Secure, Reliable, and Scalable Applications. 1. ed. Birmingham: Packt Publishing, 2018.

MCLAUGHLIN, Brett; SBARSKI, Peter et al. AWS Certified Solutions Architect – Professional Study Guide: SAP-C01 Exam. 1. ed. New Jersey: Wiley, 2020.

SCHWARTZ, Mark. AWS Well-Architected Framework. Publicado pela Amazon Web Services (AWS), atualizado regularmente. Disponível em: https://d1.awsstatic.com/whitepapers/architecture/AWS_Well-Architected_Framework.pdf. Acesso em: 19 maio. 2024.

SBARSKI, Peter. Serverless Architectures on AWS: With Examples Using AWS Lambda. 1. ed. Shelter Island: Manning Publications, 2017.

WADIA, Yohan. AWS Administration – The Definitive Guide. 2. ed. Birmingham: Packt Publishing, 2018.

GOPINATH, Ramesh. Building Data Lakes on AWS: A Comprehensive Guide to Architecting and Managing Data Lakes. 1. ed. Birmingham: Packt Publishing, 2021.

ARTASANCHEZ, Alberto. AWS for Solutions Architects: Design your cloud infrastructure by implementing DevOps, containers, and Amazon Web Services. 1. ed. Birmingham: Packt Publishing, 2021.

WADIA, Yohan; KUMARASWAMY, Narayan. Mastering AWS Lambda: Learn how to build and deploy serverless applications on AWS. 2. ed. Birmingham: Packt Publishing, 2019.

DAS, Satyajit. AWS Networking Cookbook. 1. ed. Birmingham: Packt Publishing, 2017.

VORA, Zeal. AWS Certified Security – Specialty Exam Guide: Master AWS security fundamentals. 1. ed. Birmingham: Packt Publishing, 2020.

Referências Bibliográficas (ABNT) para Google Cloud Platform

SPICELAND, James; COUSINS, Dan; MCINTYRE, Randal; RUTLEDGE, Jackie. Official Google Cloud Certified Professional Cloud Architect Study Guide. 1. ed. New Jersey: Wiley, 2020.

JILDHARD, Priyanka Vergadia; WALZ, Mark. Visualizing Google Cloud: 101 Illustrated References for Cloud Engineers and Architects. 1. ed. Sebastopol: O'Reilly Media, 2021.

PITHER, Seth. Google Cloud Certified Associate Cloud Engineer All-in-One Exam Guide. 1. ed. New York: McGraw-Hill Education, 2021.

SADA, J.C. Google Cloud for DevOps Engineers: A Practical Guide to SRE and Achieving Google-Grade Reliability. 1. ed. Sebastopol: O'Reilly Media, 2021.

NUDELMAN, Brian. Google Cloud Platform Cookbook: Practical Solutions for Building, Deploying, and Managing GCP Applications. 1. ed. Birmingham: Packt Publishing, 2021.

GRINBERG, Miguel. Flask Web Development: Developing Web Applications with Python. 2. ed. Sebastopol: O'Reilly Media, 2018.

PAGALDAY, Sandeep. Google Cloud Platform for Architects: Design and Manage Powerful Cloud Solutions. 1. ed. Birmingham: Packt Publishing, 2018.

MARELLI, Giuseppe; KROGH, Micheal. Data Science on the Google Cloud Platform: Implementing End-to-End Real-Time Data Pipelines: From Ingest to Machine Learning. 1. ed. Sebastopol: O'Reilly Media, 2018.

ROMERO, Micheal; JAMES, James. Cloud Data Management and Machine Learning on Google Cloud Platform. 1. ed. Birmingham: Packt Publishing, 2020.

SONG, Zhe. Python and Jupyter Notebook for Beginners: The Ultimate Beginners Guide to Data Science and Machine Learning on Google Cloud Platform. 1. ed. San Bernardino: Independently Published, 2020.