



BIG DATA

INTEGRAÇÃO, GERENCIAMENTO E ANÁLISE

INSTITUTO DE TECNOLOGIA E LIDERANÇA – INTELI

INTEGRAÇÃO, GERENCIAMENTO E ANÁLISE DE BIG DATA

INTEGRATION

Autores: Camila Fernanda de Lima Anacleto

Giovanna Furlan Torres

Izabella Almeida de Faria

João Moreira Tourinho Marques

Kathlyn Diwan

Maria Luísa Vilaronga Maia

Data de criação: 16 de Outubro de 2023

SÃO PAULO – SP

2023

Controle de Documento

Histórico de Revisões

Table 1: Controle de documento

Data	Autor	Versão	Resumo da Atividade
27/10/2023	Giovanna Furlan	0.0.1	Adição da arquitetura da solução e sua descrição.
28/10/2023	Izabella Faria	0.0.2	Adição da matriz de risco e sua descrição.
28/10/2023	Maria Luísa	0.0.3	Adição da análise exploratória
28/10/2023	Maria Luísa	0.0.4	Adição do TAM SAM SOM e sua descrição
29/10/2023	Giovanna Furlan	0.0.5	Adição User Story e Lean Inception
12/11/2023	João Marques	0.0.6	Adição do Resumo das Sprints + sprint1
20/12/2023	Camila Anacleto	0.0.7	Adição ETL, Armazenamento, Organização e Acesso aos Dados.

Sumário

Controle de Documento	3
Sumário	4
1. Introdução	6
1.1 Parceiro de Negócio	6
1.2 Definição do Problema	6
2. Objetivos Gerais	7
2.2 Objetivos Específicos	7
2.3 Justificativa	7
3. Lean Inception	8
3.1 O Produto (É – Não É – Faz – Não Faz)	8
3.2 Funcionalidades	8
3.3 Modelo de dados	9
4. Compreensão do Problema	9
4.1 Proposta de Valor	9
4.2 Matriz de Risco	10
4.3 TAM SAM SOM	18
5. Análise de Experiência do Usuário	19
5.1 Personas	20
5.2 Jornada do Usuário	21
5.3 User Stories	24
6. Descrição dos dados	35
6.1 Identificação dos tipos de dados e suas características.	35
6.2. Dados CSV	35
6.3 Dados CNPJ	37
6.4 API	40
7. Análise exploratória	41
7.1. CNPJs	41
7.2. Dados do Governo	41
8. Arquitetura Macro	42
8.1. Requisitos do pipeline de dados	42
8.2. Identificação dos dados de entrada e saída	43
8.3. Análise das necessidades e objetivos do pipeline	44
8.4. Escolha de serviços adequados para cada etapa do pipeline	45
8.6. Representação visual do pipeline	48
8.7. Consideração de boas práticas para garantir resiliência e escalabilidade	50
8.8. Uso de serviços ou recursos da AWS que suportem resiliência e escalabilidade	50
8.9. Calculadora financeira	51
8.10. Arquitetura e a Integration	51
9. Interface	52
9.1 WireFrame	52
9.2 Prototipação Final	52
9.3 Menu de Navegação	60
9.4 Gráficos	60
9.5 Técnicas avançadas utilizadas no design	60
9.6 Feedbacks e Iterações	60
10. Análise dos dados	61
10.1. IBGE	61

10.2. POF	61
10.3. RAIS e CAGED	61
10.4. Receita Federal Dados Abertos	61
10.5. MEC	61
10.6. INEP	61
10.7. Open Data SUS	61
10.8. Códigos postais mundiais	61
10.9. Estudo de Correlações	61
10.10. Views	61
11. Armazenamento, Organização e Acesso aos Dados	61
11.1. Data Lake	61
11.2. OLAP	61
11.3. Amazon RedShift e Escalabilidade dos dados	62
11.4. Data Warehouse	62
11.5. Views	62
12. ETL	62
12.1 Mapeamento do Fluxo	62
12.2 Serviços utilizados	62
12.3 Processo de ETL	62
13. Ensemble	62
13.1 Modelo RandomForest com CRISP-DM	63
13.2 Spark	63
13.3 Correlação de Dados por Ensemble	63
14. Análise de Custo	63
14.1 Primeiros passos	63
14.2 Configurando serviços	63
14.3 Comparação - AWS X Azure	63
14.4 Custos - Time de desenvolvimento	63
14.5 Cálculo do Custo Total	63
15. Plano de Comunicação	63
15.1 Objetivo	64
15.2 Stakeholders	64
15.3 Mensagens chaves	64
15.4 Canais de comunicação	64
15.5 Plano de implementação	64
15.6 Medidas de sucesso	64
15.7 Feedback e Ajustes	64
16. Análise de Impacto Ético	63
16.1 Introdução	64
16.2 Segurança e Proteção de Dados	64
16.3 Equidade e justiça	64
16.4 Transparência e Consentimento Informado em Projetos de Big Data	64
16.5 Responsabilidade Social	64
15.6 Viés e discriminação	64
15.7 Conclusão	64
17. Referências	65

1. Introdução

Pautada na parceria estabelecida com a Integration, uma consultoria de estratégia e gestão com sede no Brasil e presença global, incluindo escritórios na Argentina, Chile, México, EUA, Reino Unido e Alemanha. A Integration é especializada em fornecer análises estratégicas para empresas alimentícias, entre outras áreas de atuação. O desafio central identificado é a necessidade de oferecer ao cliente uma ferramenta que permita compreender o potencial de consumo de suas categorias de produtos em um nível altamente granular, incluindo informações geográficas e detalhes dos canais de atendimento. A falta dessas informações impacta diretamente a capacidade do cliente em direcionar estrategicamente suas análises e desenvolver categorias ou canais específicos. Para abordar essa questão, o projeto visa criar um pipeline de *Big Data* baseado na AWS para realizar análises estatísticas em dados armazenados em um *data lake* e *data warehouse*. Além disso, busca-se a criação de um infográfico que permitirá ao cliente tomar decisões mais informadas e inteligentes em sua operação diária.

1.1 Parceiro de Negócio

O parceiro de negócio em questão é a Integration, uma consultoria especializada em análises estratégicas para empresas alimentícias e outros setores. Ao longo dos anos, trabalharam com clientes em vários setores, mas em particular Bens de Consumo (Alimentos, Bebidas, Beleza e Cuidados Pessoais), Varejo, *Private Equity & Investimentos*, Financeiro e Pagamentos, Industrial, Agronegócio e Farmacêutico /Assistência médica.

1.2 Definição do Problema

O problema essencial consiste na falta de uma ferramenta que permita avaliar com precisão o potencial de consumo em nível granular nas categorias de produtos alimentícios. Isso prejudica a capacidade do parceiro em fornecer análises estratégicas informadas aos clientes, incluindo o direcionamento de equipe de vendas e ações táticas para o desenvolvimento de categorias ou canais específicos. A ausência de um cubo de dados, que concentrasse os dados em um único local, além disso, uma ausência de representação visual dos dados torna o parceiro incapaz de oferecer análises estratégicas com eficiência para o *Go-To-Market*.

2. Objetivos Gerais

O objetivo principal deste projeto é estabelecer um pipeline de *Big Data*, utilizando recursos da AWS, para realizar análises estatísticas em dados armazenados em um *data lake* e *data warehouse*. Além disso, busca-se a criação de um infográfico que apresente os resultados das análises estatísticas de maneira acessível e valiosa para o cliente.

2.2 Objetivos Específicos

- Coletar, armazenar e processar dados de diversas fontes, incluindo governamentais, parceiros e informações de CNPJ.;
- Realizar análises estatísticas detalhadas para determinar o potencial de consumo em nível granular (região e canal de atendimento) para cada categoria de produtos;
- Desenvolver um infográfico interativo que apresente os insights derivados das análises, auxiliando o cliente na tomada de decisões estratégicas;
- Estabelecer uma arquitetura escalável, segura e portátil para que a solução seja replicável em outros casos semelhantes.

2.3 Justificativa

A solução proposta, baseada em *Big Data* e análises estatísticas, não apenas resolverá o problema imediato, mas também fornecerá um método replicável para resolver desafios semelhantes no futuro. Além disso, a escolha de recursos da AWS e a portabilidade da solução refletem a abordagem pró-ativa do projeto em relação à escalabilidade e ao uso de recursos de nuvem. Este projeto tem o potencial de aprimorar a capacidade do cliente de tomar decisões informadas e direcionar suas operações.

3. Lean Inception

Nesta seção, apresenta-se o Lean Inception, uma técnica baseada na metodologia ágil que visa definir o escopo e os requisitos do produto de forma colaborativa e eficiente, de todo o time e das partes interessadas na solução.

3.1 O Produto (É – Não É – Faz – Não Faz)

Definição das características principais do produto, especificando o que ele “É” e o que “NÃO É”, e o que ele “FAZ” e o que “NÃO FAZ”. Garantindo que todas as partes interessadas tenham uma compreensão comum do produto e evitem mal-entendidos.

- É:
 - Projeto de Go-to-Market;
 - Atuação em lojas físicas e digitais;
 - Mapeamento de público (canal, categoria e região).
- NÃO É:
 - Plataforma que define porque o público compra aquele produto;
 - Sistema de ciência de dados;
 - Sistema de simulação de dados.
- FAZ:
 - Identificação de produtos de alto fluxo;
 - Entendimento de como e onde o consumidor compra;
 - Fornece filtros de dados;
 - Auxilia no input de simulações fora desse ambiente.
- NÃO FAZ;
 - Fornece informações de como a empresa vende;
 - Identifica clientes fora do lar e dentro do lar;
 - Predição de dados;
 - Recebe dados de países estrangeiros.

3.2 Funcionalidades

- Definir como e onde vender produtos / serviços;
- Calibrar investimento com estratégia;
- Mapeamento de concorrentes (Quais são os produtos e as estratégias?);
- Ser intuitivo para pessoas que não tem tanta familiaridade com tech;
- Deve-se transformar os códigos (como por exemplo estados) dos dados em palavras factíveis;
- Flexível o input do dado do cliente, conseguir customizar as colunas (Utilizar em vários setores);
- Mapeamento de público (Canal, categoria e região):

- Brasil : Estados, Cidades e Bairro;
- Onde o consumidor compra? (Canais)

3.3 Modelo de dados

- Formato dos dados são padrão e atualizados em 2 anos;
- Cubo de dados:
 - Canal;
 - Categoria;
 - Região.
- Arquitetura:
 - Migração de nuvem e cloud.

4. Compreensão do Problema

Esta seção apresenta as descrições das análises voltadas ao desenvolvimento de resultados do projeto, para a empresa Integration, a respeito da construção de um MVP (Produto mínimo viável) de um pipeline de *Big Data* baseado em recursos da AWS (Amazon Web Services), para realizar análises estatísticas em dados. Sendo exibido as identificações do mercado e produtos em comparação a solução prevista.

4.1 Proposta de Valor

O Canvas Proposta de Valor é elaborado para a empresa Integration, parceira neste módulo do projeto. Através deste, busca-se entender e mapear de maneira estruturada as soluções oferecidas pela Integration, especialmente no que tange ao mercado consumidor de alimentos brasileiro.

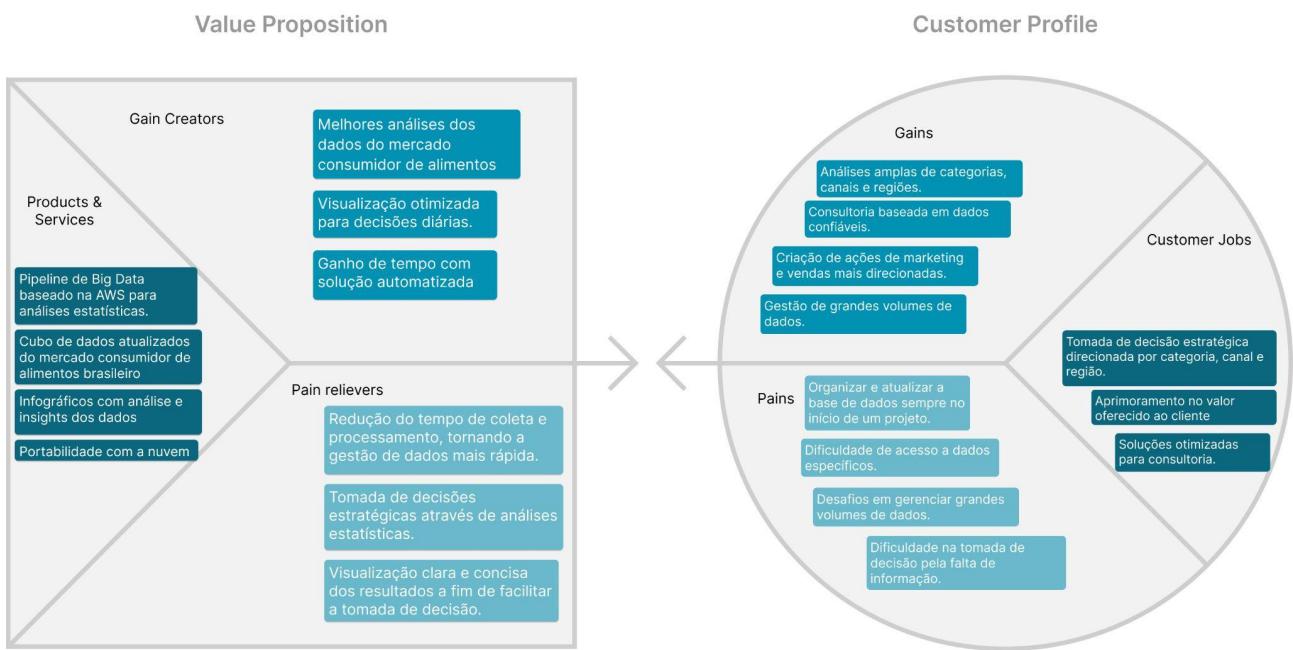
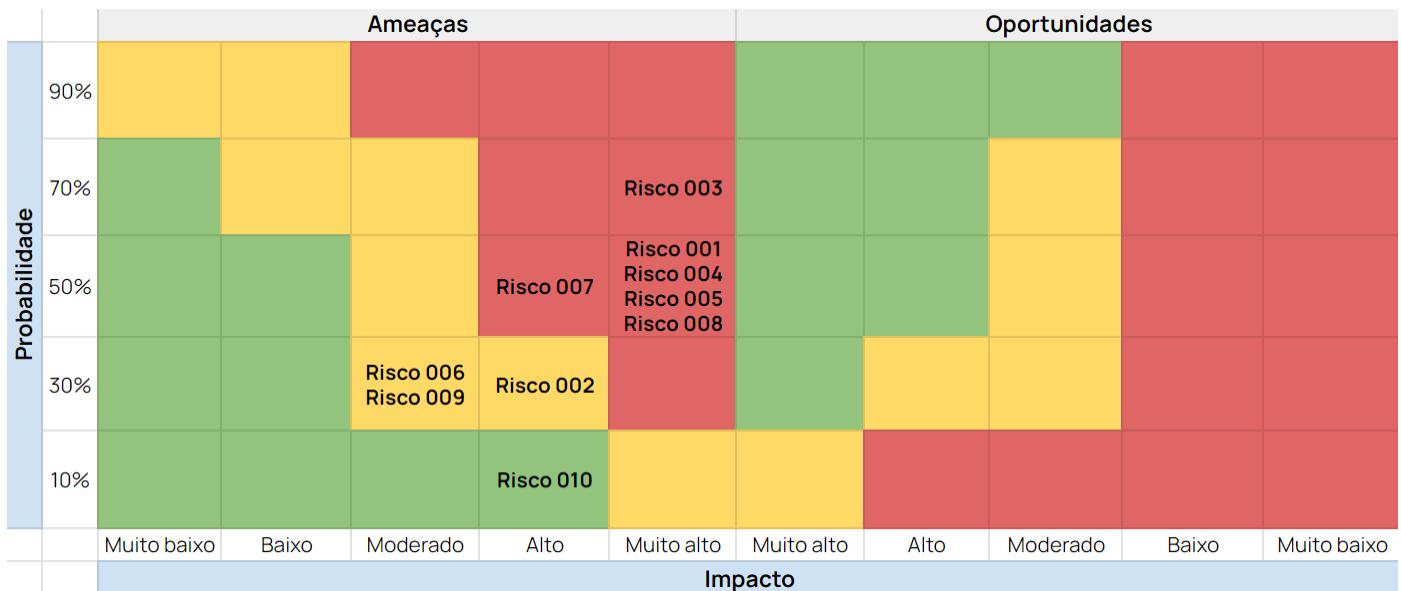


Figura 01: Canvas Proposta de Valor

Fonte: Criação própria.

4.2 Matriz de Risco



A Matriz de Risco é uma das principais ferramentas na análise de negócios, utilizada para o gerenciamento de riscos operacionais existentes na empresa. A Figura 2, ilustra a construção da matriz de risco para o projeto.

Figura 02: Matriz de Risco

Fonte: Criação própria (pode ser acessada em: [Matriz de risco](#))

Risco 001: Problemas com a privacidade e a segurança dos dados.

Problemas de privacidade e segurança de dados são uma preocupação. Em geral, os dados fornecidos pelo parceiro são públicos, o que minimiza o risco de vazamentos. No entanto, à medida que o projeto avança, se houver a necessidade de lidar com informações sensíveis, o risco de vazamento de dados aumenta, especialmente por meio de publicações não intencionais em sistemas hospedados na nuvem, como o GitHub.

Probabilidade: 50%

Impacto: Muito alto

Justificativa: A probabilidade deste risco é moderada, pois os dados públicos minimizam o risco, mas o impacto é muito alto devido à sensibilidade dos dados e às consequências de um vazamento.

Plano de ação: Em caso de vazamento de dados, a pessoa designada para lidar com esse risco terá a responsabilidade de notificar o Professor Afonso e o Orientador Renato, visando à resolução da situação com a colaboração das partes envolvidas.

Responsável: Product owner designado na sprint. Nessa primeira, a pessoa responsável é a Kathlyn Diwan.

Risco 002: Baixo entendimento do escopo.

Uma falta de clareza no escopo do projeto pode prejudicar a construção da solução, seja devido a pontos indefinidos que não estão claros para a equipe ou a uma constante alteração no escopo original devido à adição contínua de novas tarefas. Tais pontos podem impactar o desempenho do projeto.

Probabilidade: 30%

Impacto: Alto

Justificativa: A probabilidade deste risco é moderada, pois a falta de clareza no escopo pode ocorrer, mas o impacto é alto devido aos possíveis atrasos no projeto.

Plano de ação: A pessoa designada como Product Owner na sprint deverá convocar uma reunião com os demais Product Owners da turma e, ao identificar problemas relacionados às mudanças frequentes no escopo, informar o Orientador da turma. O propósito dessa ação é buscar um consenso entre as partes envolvidas e solicitar uma revisão do escopo do projeto.

Responsável: Product owner designado na sprint. Nessa primeira, a pessoa responsável é a Kathlyn Diwan.

Risco 003: Dificuldade na execução de tarefas devido à alta complexidade do projeto.

De forma geral, a equipe enfrenta desafios com tarefas técnicas, possuindo um conhecimento limitado sobre o assunto e as matérias abordadas neste módulo. Essa falta de experiência pode resultar em dificuldades ao lidar com a alta complexidade do projeto, o que pode impactar negativamente a busca por um resultado final satisfatório. Tarefas como a definição da arquitetura da solução, a escolha da infraestrutura e a implementação de computação em nuvem podem se tornar particularmente desafiadoras.

Probabilidade: 70%

Impacto: Muito alto.

Justificativa: A probabilidade desse risco é elevada, pois a equipe possui conhecimentos limitados nas áreas técnicas relevantes para o projeto. O impacto é moderado, pois as dificuldades técnicas podem atrasar o projeto, mas com treinamento e apoio adequado, os desafios podem ser superados.

Plano de ação: Se o time enfrentar uma dificuldade insuperável, o responsável deve notificar os professores de que não será possível concluir todos os incrementos planejados e continuar com as tarefas exigidas devido à incapacidade de lidar com elas. Além disso, os membros do grupo que enfrentarem maiores desafios se comprometem a informar o restante da equipe e buscar ajuda sempre que necessário.

Responsável: Giovanna Furlan, Maria Luisa Maia

Risco 004: Desvio do orçamento previsto para o projeto mediante ao uso irrestrito de soluções em nuvem.

Se ocorrer um uso descontrolado de recursos na nuvem, como armazenamento e computação, o time enfrentará desafios significativos relacionados a custos mais altos do que o inicialmente previsto, afetando negativamente a situação financeira do projeto. Isso pode ocorrer devido à falta de controle adequado do consumo de recursos na nuvem, dimensionamento inefficiente e a ausência de limites orçamentários claros. Essa situação pode impactar o financiamento do projeto, causar atrasos nas entregas e aumentar a complexidade da gestão de custos.

Probabilidade: 50%

Impacto: Muito alto.

Justificativa: A probabilidade deste risco é moderada, pois o dimensionamento inadequado da arquitetura pode ocorrer. Além disso, o impacto é alto, uma vez que a falta de escalabilidade compromete a eficácia do projeto.

Plano de ação: Para mitigar esse risco, é necessário estabelecer uma constante revisão da arquitetura escolhida para ser seguida no projeto. Diante disso, é preciso que

os integrantes da equipe estejam atentos ao consumo da solução, com objetivo de estipular um limite quando necessário ou mudar a tecnologia escolhida inicialmente. Para isso, um sistema de monitoramento constante do consumo de recursos em soluções em nuvem deve ser estabelecido. Devem ser utilizadas ferramentas e métricas apropriadas para rastrear o uso de recursos, identificar tendências de aumento de custos e antecipar possíveis problemas.

Responsável: João Tourinho e Camila Anacleto

Risco 005: Dificuldade na escalabilidade devido à arquitetura inadequada.

Se a arquitetura da solução proposta não for cuidadosamente planejada ou não atender às necessidades previamente definidas, existe a possibilidade de que o projeto não alcance uma escalabilidade satisfatória. Isso, por sua vez, compromete o resultado final e prejudica a viabilidade do uso futuro da solução por parte dos parceiros do projeto. À medida que a quantidade de dados a ser processada aumenta, é provável que enfrentemos desafios relacionados ao desempenho, armazenamento, processamento e distribuição dos dados.

Probabilidade: 50%

Impacto: Muito alto

Justificativa: A probabilidade desse risco é moderada, pois a arquitetura é uma parte crucial do projeto e, se não planejada corretamente, pode levar a problemas de escalabilidade. O impacto é alto, uma vez que uma arquitetura inadequada pode afetar profundamente a viabilidade do projeto.

Plano de ação: Para mitigar esse risco, iremos revisar a arquitetura do projeto com foco em garantir que seja flexível para acomodar o crescimento de dados. Vamos realizar testes de desempenho com volumes moderados de dados e consultar professores e mentores da universidade para orientação. Nossa modelagem de crescimento será simplificada, e implementaremos um monitoramento básico para acompanhar a capacidade de armazenamento e desempenho da arquitetura.

Responsável: Giovanna Furlan e Izabella Faria

Risco 006: Dificuldade na integração e análise de dados provenientes de várias fontes.

Se a integração de dados provenientes de diversas fontes se revelar uma tarefa complexa, o time enfrentará desafios que incluem a possibilidade de atrasos significativos e custos adicionais para o projeto. A complexidade da integração deriva da necessidade

de alinhar diferentes formatos, estruturas e protocolos de dados, o que pode demandar mais tempo e recursos do que o inicialmente previsto. Essa situação pode impactar negativamente o cronograma e o orçamento do projeto.

Probabilidade: 30%

Impacto: Moderado

Justificativa: A probabilidade desse risco é considerada moderada, uma vez que a integração de dados de múltiplas fontes é uma parte comum de projetos de tecnologia, e problemas nessa área são razoavelmente comuns. O impacto é alto, já que a complexidade de integração pode levar a atrasos no projeto e a custos adicionais, afetando tanto o cronograma quanto o orçamento previsto.

Plano de ação: Caso não seja possível dar andamento ao projeto devido à dificuldade de integração entre as fontes de dados, a equipe deverá imediatamente notificar o professor orientador e buscar assistência para avaliar alternativas de integração ou considerar uma abordagem de projeto alternativa que minimize os impactos nos prazos e custos.

Responsável: Maria Luisa Maia

Risco 007: Falha na implementação de políticas de segurança de dados.

Caso a implementação adequada das políticas de segurança de dados não seja realizada, o time se deparará com riscos de violação de privacidade e segurança. Isso inclui a falta de criptografia, autenticação inadequada ou configurações incorretas de acesso aos dados, o que pode expor o projeto a sérias vulnerabilidades.

Probabilidade: 50%

Impacto: Alto

Justificativa: A probabilidade desse risco é moderada, uma vez que a equipe está ciente da importância da segurança, mas as implementações podem conter erros. O impacto é alto devido ao potencial comprometimento da segurança dos dados.

Plano de ação: Para mitigar esse risco, a equipe deve realizar auditorias regulares de segurança, implementar criptografia adequada e seguir as melhores práticas de segurança de dados. A educação contínua da equipe em relação à segurança também é fundamental.

Responsável: Camila Anacleto

Risco 08: Erros na interpretação dos resultados devido à complexidade dos dados.

Se a arquitetura da solução proposta não for cuidadosamente planejada ou não atender às necessidades previamente definidas, existe a possibilidade de que o projeto não alcance uma escalabilidade satisfatória. Isso, por sua vez, compromete o resultado final e prejudica a viabilidade do uso futuro da solução por parte dos parceiros do projeto. À medida que a quantidade de dados a ser processada aumenta, é provável que enfrentemos desafios relacionados ao desempenho, armazenamento, processamento e distribuição dos dados.

Probabilidade: 30%

Impacto: Muito alto

Justificativa: A probabilidade deste risco é moderada, dada a complexidade dos dados. O impacto é muito alto, pois erros na interpretação podem prejudicar o valor do projeto.

Plano de ação: Para mitigar esse risco, é necessário que a equipe busque capacitação em análise de dados, a fim de aprimorar a compreensão dos dados e das técnicas de análise. Além disso, é importante realizar uma validação cruzada de interpretações, envolvendo membros da equipe com diferentes perspectivas e conhecimentos. É fundamental manter uma documentação detalhada dos métodos de análise e resultados, tornando-a uma prática constante para garantir a transparência e facilitar a revisão por pares. Solicitar feedback dos stakeholders também é uma ação importante para assegurar que as descobertas estejam alinhadas com as necessidades e expectativas.

Responsável: Izabella Faria.

Risco 009: Aumento do tempo de processamento dos dados coletados.

Se houver um aumento no tempo necessário para extrair, processar e analisar os dados coletados devido a problemas de desempenho nos sistemas, aumento da carga de trabalho ou problemas técnicos não previstos, o time enfrentará desafios relacionados à capacidade de entregar insights e análises no prazo estabelecido.

Probabilidade: 30%

Impacto: Moderado

Justificativa: A probabilidade desse risco é moderada, pois problemas de desempenho e atrasos no processamento de dados são comuns em projetos de Big Data. O impacto é moderado, uma vez que atrasos podem afetar o cronograma, mas medidas de mitigação podem minimizar o impacto no resultado final.

Plano de ação: Para mitigar esse risco, a equipe deve implementar medidas de otimização de desempenho nos sistemas de processamento de dados. Isso inclui a alocação de recursos adequados, o uso de tecnologias de processamento paralelo e a implementação de estratégias de escalabilidade. Além disso, um monitoramento constante do desempenho do sistema e a identificação antecipada de gargalos podem ajudar a evitar atrasos significativos.

Responsável: João Tourinho

Risco 010: Falta de controle no gerenciamento das atividades individuais e do projeto.

Na hipótese de o time enfrentar dificuldades na organização do cronograma e na mensuração do tempo necessário para desenvolver as tarefas estabelecidas devido ao grande volume de atividades que precisam ser realizadas, tanto no que diz respeito às tarefas individuais quanto às do time, isso pode resultar em atrasos nas entregas e no planejamento geral do projeto, prejudicando o seu progresso.

Probabilidade: 10%

Impacto: Muito alto.

Justificativa: A probabilidade desse risco é moderada, uma vez que a organização do cronograma pode ser desafiadora em projetos complexos com muitas tarefas. O impacto é alto, pois atrasos nas entregas das sprints podem impactar negativamente o progresso do projeto e a satisfação dos parceiros.

Plano de ação: Para mitigar esse risco, a equipe manterá uma gestão rigorosa do cronograma e priorização de tarefas. Será estabelecida uma comunicação eficaz para garantir que todos os membros estejam cientes de suas responsabilidades e prazos. Além disso, a equipe considerará a possibilidade de redistribuir tarefas entre os membros, se necessário.

Responsável: Product owner designado na sprint. Nessa primeira, a pessoa responsável é a Kathlyn Diwan.

Categorização dos riscos

Abaixo, é possível visualizar a classificação dos riscos em três categorias primordiais: segurança, equipe e arquitetura. Além disso, disposto na tabela é possível visualizar os respectivos impactos e a probabilidade de ocorrência.

Risco	Categoria	Impacto	Probabilidade
-------	-----------	---------	---------------

001	Segurança	Muito alto	50%
002	Equipe	Alto	30%
003	Equipe	Muito alto	70%
004	Arquitetura	Muito alto	50%
005	Arquitetura	Muito alto	50%
006	Equipe	Moderado	30%
007	Segurança	Alto	50%
008	Equipe	Muito alto	30%
009	Arquitetura	Moderado	30%
010	Equipe	Alto	10%

Com base nestes dados, é elaborado um gráfico que ilustra a distribuição dos riscos mapeados em cada categoria selecionada. Os resultados revelam que, entre os riscos identificados na primeira sprint, aqueles relacionados à equipe e suas potenciais dificuldades são os mais frequentes. Portanto, é crucial direcionar uma atenção especial às necessidades e desafios da equipe, a fim de prevenir que esses riscos se transformem em problemas reais no futuro.

Distribuição dos riscos em categorias:

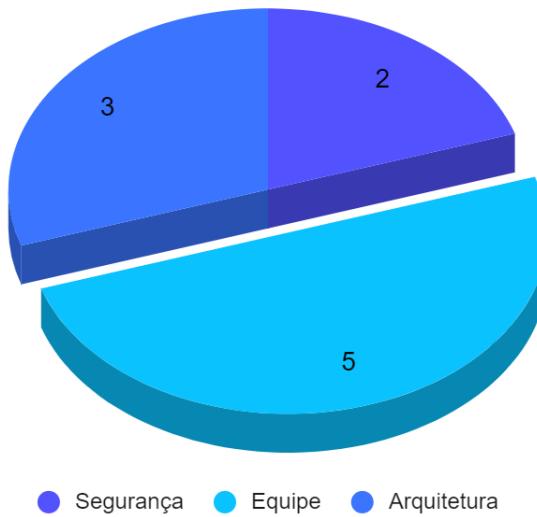


Figura 03: Distribuição dos riscos

Fonte: Criação própria

Além disso, é possível também a classificação dos riscos com base em seu impacto no projeto. Nessa análise, é constatado que a maioria dos riscos possui um impacto de grande magnitude, o que evidencia a necessidade de adotar medidas substanciais para prevenir a materialização desses cenários.

Distribuição do impacto dos riscos mapeados:

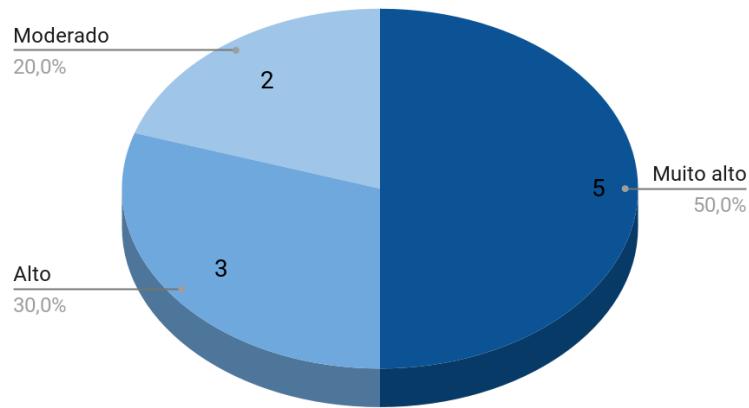


Figura 04: Distribuição do impacto

Fonte: Criação própria

4.3 TAM SAM SOM

O Total Addressable Market, Service Addressable Market e Service Obtainable Market (TAM, SAM e SOM) é utilizado no contexto de análise de mercado e estratégias de negócios, para definir e dimensionar quais são os mercados-alvo. O TAM demonstra a estimativa total do quanto a empresa pode atingir se todos os clientes fossem alcançados, ou seja o mercado total disponível para um produto / serviço, independentemente das limitações. Já o SAM, é uma porção do anterior, na qual a empresa pode efetivamente atingir com seus recursos e estratégias, ou seja, o mercado em que a empresa pode competir de maneira realista. Por último, o SOM é uma fatia do mercado que a empresa pretende atingir a curto prazo, o seu objetivo principal com o seu serviço / produto. Esta análise ajuda as empresas a tomar decisões informadas sobre onde concentrar seus esforços e recursos para alcançar o seu objetivo. A imagem abaixo demonstra essa análise feita para o projeto, figura 05.



Figura 05: TAM SAM SOM

Fonte: Criação própria

O TAM (Total Addressable Market) representa o universo completo de empresas que atuam no setor varejista no Brasil, que possui 6.852.928 CNPJs de empresas ativas em 2023, de acordo com o “Empresômetro”. Por outro lado, o SAM (Serviceable Addressable Market) é uma parcela do TAM que a empresa está focada nesse momento em alcançar com seus recursos e estratégias, que neste caso são 2.022.225 CNPJs de empresas. Esses dados foram retirados de todos os arquivos csv disponibilizados para a realização do projeto. Note que, são empresas com CNAEs específicos, descritos na Análise Descritiva. Já o SOM (Share of Market) representa a fatia específica do mercado que a empresa pretende atingir a curto prazo, que foi definido como 15% do SAM, o que equivale a 303.333 CNPJs de empresas varejistas.

É importante ressaltar que o setor de varejo tem investido cada vez mais em Business Intelligence (BI) nos últimos anos devido à crescente competição e à necessidade de tomar decisões baseadas em dados para melhorar a eficiência operacional e a experiência do cliente. Para isso, alguns processos que podem ser melhorados com a implementação do BI são: coleta, análise e interpretação de dados.

5. Análise de Experiência do Usuário

Nesta seção, apresenta-se a análise de experiência do usuário, a qual através da aplicação de estratégias, visa compreender como os usuários interagem com sistemas, produtos e serviços. O objetivo é melhorar a satisfação e a eficiência dessas interações, levando em conta aspectos subjetivos como emoções, percepções e expectativas dos usuários.

5.1 Personas

As personas desempenham um papel essencial na compreensão e no direcionamento de qualquer projeto ou solução. Elas são representações fictícias, mas altamente detalhadas, dos tipos ideais de clientes que a solução visa atender, neste caso, duas personas distintas: a Consultora de Marketing e Vendas e o Analista de Dados.

Essas personas são baseadas nos setores principais que são fundamentais para a eficácia da solução. Cada uma delas incorpora características, comportamentos e preferências que se alinham com o contexto em que a *Integration* se encontra. Estas, ajudam a equipe a visualizar os usuários finais e definir estratégias e recursos que atendam às necessidades e expectativas de cada um desses perfis.

AMANDA BRAZ

CONSULTORA MARKETING E VENDAS



● BACKGROUND

Amanda é uma mulher de 27 anos que trabalha com a área de Marketing cuidando das empresas alimentícias que a consultoria presta serviço. Nascida em São Paulo, formada na ESPM em Publicidade e Propaganda.

DORES

- Ter que rodar os dados toda vez que um projeto começa;
- Análise dos dados manual demorada;
- Armazenamento dos dados ineficiente.

NECESSIDADES

- Coleta e análise dos dados de maneira eficiente segura e organizada;
- Escalabilidade e automatização;
- Acesso a dados atualizados.

DESEJOS

- Capacidade de previsão de tendência;
- Automatização;
- Visualização personalizada dos dados.

“

"Tenho tantos dados para analisar, e às vezes me sinto afogada neles."

"Perder informações valiosas é meu maior pesadelo"

KPI

- Taxa de conversão de campanhas de marketing.
- Retorno sobre o investimento (ROI)

CENÁRIO DE INTERAÇÃO

- No escritório da consultoria;
- Em reuniões com clientes do segmento alimentício;

INTERESSES

- Acompanhamento de tendências de marketing e tecnologia no segmento alimentício.

LETRAMENTO DIGITAL

- Possui um nível avançado em ferramentas de análise e visualização de dados;
- Não tem tanto conhecimento de AWS.

Figura 06: Persona 1 (Amanda)

Fonte: Criação própria

LUIZ MACHADO

ANALISTA DE DADOS



● BACKGROUND

Luiz é um homem de 25 anos que trabalha na área de Data dentro de uma empresa de Consultoria. Ele é nascido no Rio de Janeiro, estudou Ciência da Computação e está fazendo uma pós-graduação em Cloud Computing.

DORES

- Demora na preparação e limpeza de dados antes da análise;
- Erros na interpretação de dados;
- Pouca segurança no armazenamento dos dados;

”

NECESSIDADES

- Automatização de tarefas de rotina para economizar tempo;
- Uma plataforma que facilite a importação e integração de dados;
- Acesso a dados atualizados em tempo real.

DESEJOS

- Capacidade de previsão;
- Automatização completa;
- Maior personalização e escalabilidade.

KPI

- Métricas relacionadas à precisão e armazenamento de dados;
- Eficiência no processamento dos dados.

CENÁRIO DE INTERAÇÃO

- No escritório do trabalho;
- Em reuniões com outras áreas da empresa;

INTERESSES

- Acompanhamento de tendências de tecnologia;
- Participação em workshops de datascience;
- Datathons;

LETRAMENTO DIGITAL

- Possui conhecimento avançado em linguagens de programação e ferramentas de análise de dados.

Figura 07: Persona 2 (Luiz)

Fonte: Criação própria

5.2 Jornada do Usuário

A jornada do usuário construída consiste na representação das etapas principais que envolvem os consultores de marketing e vendas e o analista de dados. Ambos começam com o acesso à plataforma por meio de autenticação. A consultora se envolve com a análise de dados, usando um painel interativo no qual pode explorar informações sobre o potencial de consumo em categorias específicas, filtrar dados, criar relatórios personalizados e tomar decisões estratégicas. Por outro lado, o analista de dados utiliza o pipeline de dados para acessar e analisar dados brutos, conduzindo análises estatísticas detalhadas e preparando o infográfico final. Essa jornada do usuário proporciona uma experiência que permite que ambos os profissionais extraiam valor dos dados e forneçam

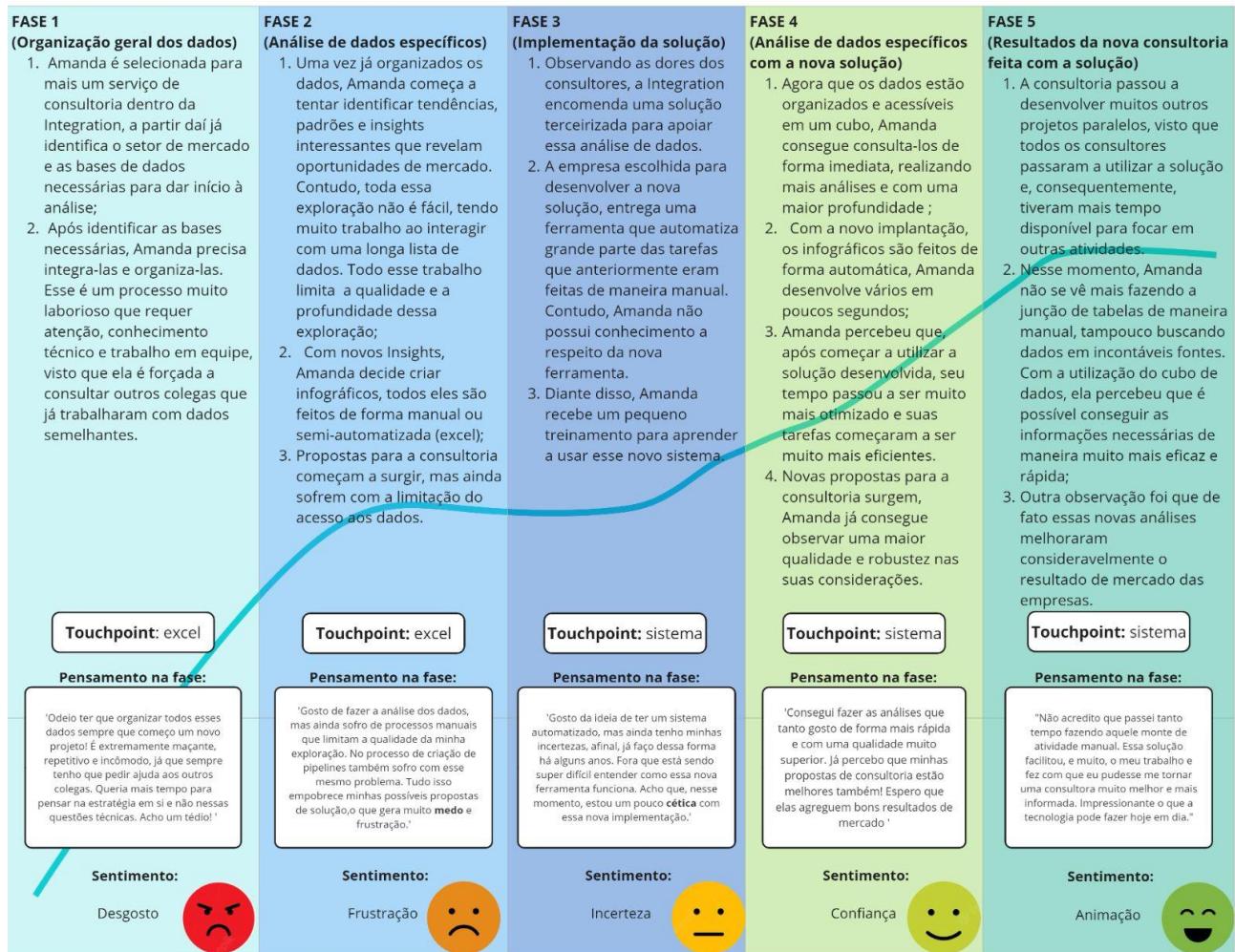
informações valiosas para o cliente.



Consultora de marketing e vendas, Amanda Braz

Cenário: Amanda quer fazer uma análise do mercado consumidor de alimentos de determinada região do Brasil.

Expectativas: Garantir um serviço de consultoria eficiente, com boas soluções de mercado. Visualizar gráficos relevantes com base nos dados utilizados.



Oportunidades

- Criar um manual com linguagem clara e intuitiva que sirva de fato como um treinamento para esse consultor.
- Aprimorar o sistema implementado para realizar algumas tarefas automaticamente, como, por exemplo, conectar à conta da nuvem utilizada.

Momentos da verdade:

Momento Zero da Verdade (ZMOT): O Momento Zero da Verdade ocorre quando Amanda, após ser selecionada para o serviço de consultoria, começa a identificar as bases de dados necessárias para sua análise de mercado. Neste momento, ela toma decisões críticas sobre quais dados usar, as fontes apropriadas e como integrá-los. O ZMOT é essencial para a qualidade e eficácia de sua análise, pois representa o estágio de pesquisa e preparação antes de qualquer ação significativa.

Primeiro Momento da Verdade: O Primeiro Momento da Verdade ocorre quando Amanda adquire a solução terceirizada para analisar dados. Neste momento, a decisão de compra da solução e a experiência de concluir a transação são cruciais. Ela avalia a página de transação, a clareza das informações e a facilidade de uso do sistema. A satisfação nessa etapa depende da eficácia da compra e da experiência de transação.

Segundo Momento da Verdade: O Segundo Momento da Verdade está relacionado à experiência de Amanda ao usar a nova solução para analisar os dados. Neste ponto, Amanda aprende a utilizar a ferramenta, avalia a eficiência da automação na análise de dados e mede a otimização de seu tempo de trabalho. A qualidade de suas análises e sua eficiência dependem da adoção bem-sucedida da nova solução.

Figura 08: Jornada do Usuário 1 (Amanda)

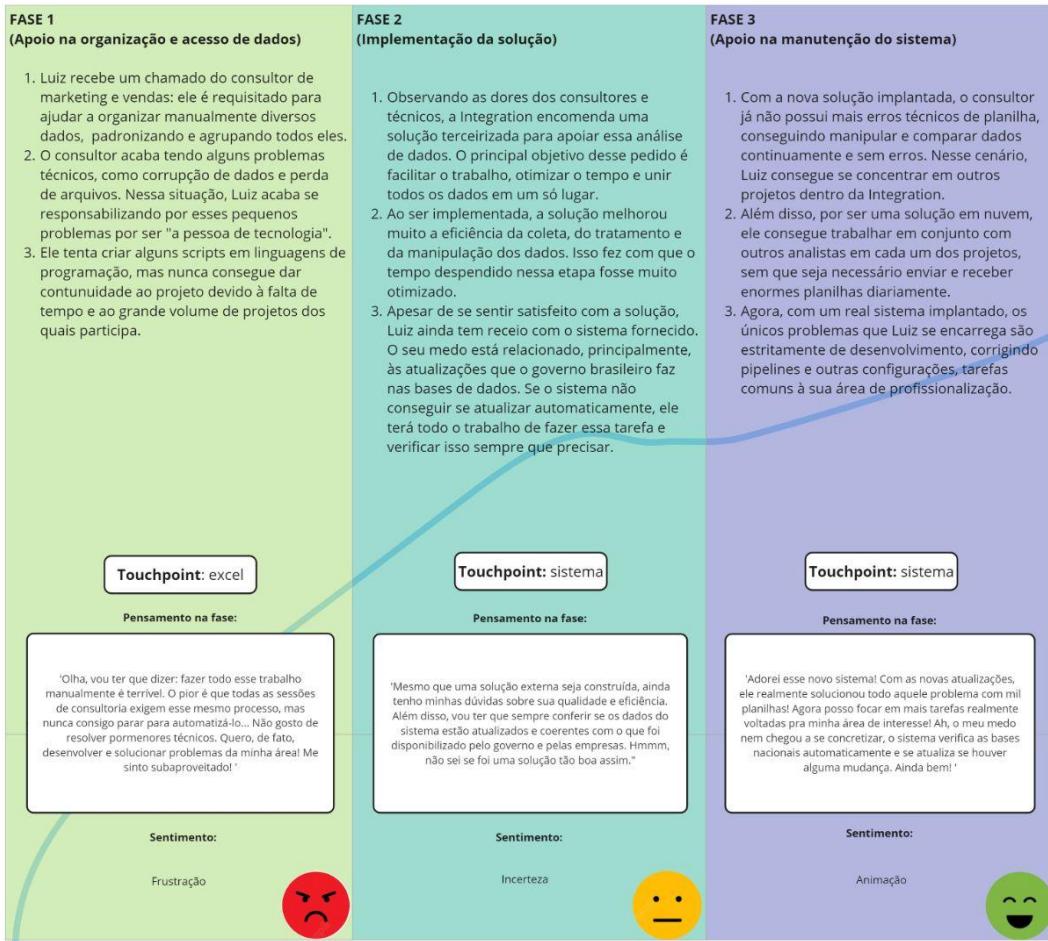
Fonte: Criação própria



Analista de dados, Luiz Machado

Cenário: Luiz deseja garantir, com o apoio das tecnologias, que a análise de dados feita pelos consultores seja a mais eficiente possível.

Expectativas: Garantir um serviço de consultoria eficiente, com boas soluções de mercado. Otimizar o tempo necessário para extrair e analisar os dados.



Oportunidades

- Fazer um manual personalizado para o profissional de tecnologia.
- Implementar a mesma estrutura utilizada nesse processo de automatização em outros projetos da empresa.
- Trabalhar em conjunto com outros profissionais para gerar uma integração entre essa solução e outras ferramentas do mercado de tecnologia.

Momentos da verdade

Momento Zero da Verdade (ZMOT): No início da jornada de Luiz, o Momento Zero da Verdade ocorre quando ele é requisitado para ajudar a organizar manualmente diversos dados e padronizá-los. Ele também enfrenta problemas técnicos, como corrupção de dados e perda de arquivos, que o levam a assumir a responsabilidade de lidar com esses desafios de tecnologia. O ZMOT é evidenciado aqui, pois Luiz está em um estágio de avaliação e decisão sobre como abordar a organização de dados de maneira mais eficiente.

Primeiro Momento da Verdade: O Primeiro Momento da Verdade se manifesta quando a Integration encomenda uma solução terceirizada para otimizar a análise de dados. Luiz é parte fundamental na implementação dessa solução, e o momento da verdade ocorre quando ele percebe a eficácia da nova ferramenta. A experiência de implementação, a otimização do tempo e a facilidade de unir dados em um só lugar são aspectos críticos desta fase.

Segundo Momento da Verdade: O Segundo Momento da Verdade é evidenciado após a implantação da solução. Luiz nota a eficiência da solução na coleta, tratamento e manipulação de dados. No entanto, ele ainda tem preocupações relacionadas às atualizações do governo nas bases de dados. Isso se encaixa no Segundo Momento da Verdade, pois é o momento em que as expectativas se encontram com a realidade e as preocupações com a manutenção do sistema surgem.

Terceiro Momento da Verdade: O Terceiro Momento da Verdade acontece quando a nova solução está totalmente em uso. Luiz relata que o consultor já não enfrenta mais erros técnicos, e ele mesmo consegue se concentrar em outros projetos. O fato de o sistema ser em nuvem e permitir uma colaboração mais eficiente entre os analistas também é um aspecto do Terceiro Momento da Verdade. Neste ponto, Luiz lida principalmente com problemas de desenvolvimento e configurações, o que reflete uma fase de experiência após a adoção bem-sucedida da solução.

Figura 09: Jornada do Usuário 2 (Luiz)

Fonte: Criação própria

5.3 User Stories

Pode-se definir *User Stories* como descrições simplificadas das funcionalidades possíveis que o usuário possui e deseja dentro da aplicação, escrita com a visão dele. Além de transparecer como o sistema espera alcançar tais objetivos. Apresenta-se abaixo as *user stories* referente à aplicação da *Integration*.

5.3.1 US00 - Configuração do Ambiente AWS

Persona: Analista de Dados

História: Como um analista de dados, quero configurar o ambiente AWS para armazenamento, preparação e análise de dados, a fim de estabelecer uma infraestrutura funcional.

Critério de avaliação:

- **Critério 1:** Ambiente AWS configurado corretamente
 - **Condição:** { Verificar se o ambiente AWS foi configurado de acordo com as especificações. Isso inclui a presença e configuração correta do Amazon S3 para armazenamento, do AWS Glue para preparação de dados e do Apache Spark para análise estatística. }
- **Critério 2:** Serviços AWS estão operacionais e interconectados
 - **Condição:** { Verificar a operacionalidade de todos os serviços AWS listados na arquitetura. Além disso, é importante verificar se esses serviços estão conectados e podem se comunicar entre si conforme necessário. }
- **Critério 3:** Documentação da infraestrutura disponível
 - **Condição:** { Existir uma documentação detalhada que descreve a configuração da infraestrutura, incluindo os serviços AWS utilizados, as configurações específicas de cada serviço e qualquer personalização realizada. A documentação deve ser abrangente para permitir referência futura e manutenção. }

Teste de aceitação:

- **Teste 1:** Serviços da AWS operacionais
 - **Aprovado:** { Os serviços da AWS estão operacionais e todos os componentes da aplicação estão funcionando sem problemas. }
 - **Recusado:** { Os serviços da AWS estão inoperantes, resultando em uma interrupção dos serviços da aplicação. Os componentes da aplicação estão inacessíveis e não funcionam. }

- **Teste 2:** Os dados devem ser armazenados com sucesso no S3.
 - Aprovado: { Os dados estão sendo armazenados com sucesso no Amazon S3. Os dados são acessíveis e podem ser recuperados sem problemas. }
 - Recusado: { Não é possível armazenar dados no Amazon S3, ou os dados armazenados estão corrompidos e não podem ser recuperados. }
- **Teste 3:** Preparação de dados com Glue ou Lambda está funcionando corretamente.
 - **Aprovado:** { A preparação de dados com AWS Glue ou AWS Lambda está funcionando corretamente. Os dados são transformados e preparados para análise sem erros ou interrupções. }
 - **Recusado:** { A preparação de dados com Glue ou Lambda está com falhas, causando erros na transformação dos dados ou impedindo que os dados estejam prontos para análise. }

Notas: Esta história é fundamental para a configuração inicial do ambiente AWS necessário para o pipeline de Big Data.

- Prioridade : Alta
- Estimativa : 4 dias
- Relação : N/A

5.3.2 US01 - Ingestão de Dados

Persona: Analista de Dados

História: Como um analista de dados, quero implementar a ingestão de um conjunto de dados para fins de análise estatística, a fim de realizar testes iniciais.

Critério de avaliação:

- **Critério 1:** Conjunto de dados disponibilizado e carregado com sucesso no ambiente AWS.
 - Condição : { Os dados de origem foram carregados com sucesso em um local de armazenamento, como o Amazon S3, e estão disponíveis para processamento. }
- **Critério 2:** Tolerância a falhas, visando segurança (redundância).
 - Condição : { Verificar se a implementação ocorreu com mecanismos de tolerância a falhas, seleção de backup, que garantam a segurança dos serviços e dados. }
- **Critério 3:** Os dados devem ser de origem não estruturada.

- Condição: { Analisar os tipos de dados, como documentos de texto, para confirmar que eles não seguem um formato estruturado, como tabelas de banco de dados. A aprovação deste requisito está relacionada à confirmação de que os dados são, de fato, não estruturados. }

Teste de aceitação:

- **Teste 1:** O conjunto de dados é carregado com sucesso no ambiente AWS.
 - Aprovado: { O conjunto de dados é carregado com sucesso no ambiente AWS sem erros ou problemas. Isso inclui a verificação de que os dados foram transferidos com precisão, que não houve perda de informações durante o processo de carga e que os dados estão acessíveis e disponíveis conforme o esperado. }
 - Recusado: { Se ocorrerem erros durante o carregamento dos dados, se houver perda de informações críticas ou se os dados não estiverem disponíveis conforme o esperado no ambiente AWS. }
- **Teste 2:** Tolerância a falhas
 - Aprovado: { Demonstração de tolerância a falhas e segurança através do uso de redundância adequada. Isso significa que o sistema é capaz de continuar operando mesmo em face de falhas em componentes individuais, garantindo que a disponibilidade e a integridade dos dados sejam mantidas. }
 - Recusado: { Não apresentar tolerância a falhas ou não demonstrar a segurança necessária por meio de redundância. Resultando em vulnerabilidades que comprometem a continuidade das operações e a segurança dos dados. }
- **Teste 3:** Os dados são verificados como não estruturados.
 - Aprovado: { Verificação dos dados for bem-sucedida, classificados como não estruturados conforme as expectativas. Os dados não seguem um formato ou padrão rígido e podem conter informações variadas. }
 - Recusado: { Verificação dos dados identificar que eles são estruturados ou se houver dificuldades em determinar a natureza não estruturada dos dados. Podendo indicar que a classificação dos dados como não estruturados não foi realizada com sucesso. }

Notas: Esta história se concentra na ingestão de dados iniciais para testes.

Prioridade: Alta

Estimativa: 14 dias

Relação: US00

5.3.3 US02 - Análise Estatística Inicial

Persona: Analista de Dados

História: Como um analista de dados, quero realizar uma análise estatística inicial dos dados carregados no ambiente AWS, a fim de identificar tendências e padrões.

Critério de avaliação:

- **Critério 1:** Os dados carregados são processados com sucesso.
 - Condição : { Os dados carregados passaram com sucesso por todo o processo de preparação, transformação e carga (ETL). Ou seja, os dados estão limpos, transformados adequadamente e prontos para análise. }
- **Critério 2:** Análises estatísticas descritivas
 - Condição : { As análises estatísticas foram executadas com sucesso nos dados preparados. As métricas estatísticas, como média, desvio padrão, histogramas ou outras métricas relevantes, foram geradas com precisão. }
- **Critério 3:** Os resultados da análise são armazenados para validação posterior.
 - Condição : { Verificar se os resultados das análises estatísticas foram armazenados de forma adequada e estão disponíveis para validação posterior. Deve-se conter registros com os resultados das análises. }

Teste de aceitação:

- **Teste 1:** Os dados são processados com sucesso
 - Aprovado: { Verificar se os dados passaram por todas as etapas de transformação e limpeza, se aplicável, sem erros críticos. Os dados processados devem estar prontos para análises. }
 - Recusado: { Se ocorrerem erros significativos durante o processamento dos dados, perda de informações importantes ou se os dados processados não estiverem prontos para uso, devido a problemas de qualidade ou integridade. }
- **Teste 2:** As análises estatísticas são geradas de forma precisa.
 - Aprovado: { As análises estatísticas foram geradas de forma precisa e confiável, refletindo com precisão os dados processados. Isso inclui a verificação de que as métricas estatísticas estão corretamente calculadas e representam os dados subjacentes. }

- Recusado: { As análises estatísticas geradas contém erros ou imprecisões significativas. Isso pode ocorrer devido a problemas nos cálculos ou nos dados de entrada. }
- **Teste 3:** Os resultados são armazenados e prontos para validação.
 - Aprovado: { Os resultados das análises são armazenados de forma adequada e estão prontos para validação. Ou seja, os resultados estão acessíveis, bem documentados e podem ser facilmente verificados por partes interessadas. }
 - Recusado: { Os resultados das análises não foram armazenados corretamente, estão incompletos, ou não estão disponíveis para validação. }

Notas: Esta história se concentra na análise estatística inicial dos dados para identificação de tendências.

Prioridade: Média

Estimativa: 7 dias

Relação: US01 - US03

5.3.4 US03 - Load de Dados

Persona: Analista de Dados

História: Como um analista de dados, desejo otimizar minhas atividades diárias através de um script Python personalizado, que permitirá uma padronização de dados rápida usando bases de dados existentes, reduzindo o tempo gasto na construção de novas bases de dados.

Critério de avaliação:

- **Critério 1:** Script Python que seja capaz de acessar e utilizar bases de dados prontas.
 - Condição : { Para aprovar este requisito, é necessário verificar se o script Python foi desenvolvido e é capaz de acessar as bases de dados prontas. A capacidade de se conectar a essas bases de dados, consultar dados e realizar operações relevantes deve ser verificada. }
- **Critério 2:** O script deve permitir uma padronização de dados.
 - Condição : { O script Python deve realizar uma padronização de dados. Isso pode ser avaliado através de métricas de desempenho, como o tempo necessário para executar a padronização em comparação com processos anteriores. }

Teste de aceitação:

- **Teste 1:** O script é capaz de acessar e utilizar bases de dados prontas
 - Aprovado: { O script acessa e usa bases de dados prontas padronizando os dados necessários de maneira precisa e sem erros. Deve ser capaz de manipular os dados conforme necessário. }
 - Recusado: { O script Python não consegue acessar ou utilizar as bases de dados prontas, resultando em erros, tempos de execução excessivos ou problemas na recuperação e manipulação dos dados. }
- **Teste 2:** A padronização de dados usando o script é mais rápida do que o processo manual.
 - Aprovado: { O script Python torna o processo de padronização de dados mais rápido do que o processo manual, demonstrando melhorias no desempenho através de métricas de tempo de execução ou comparações diretas com o processo anterior. }
 - Recusado: { O teste é considerado recusado se a análise de dados usando o script Python não demonstrar melhorias no desempenho em comparação com o processo manual. Isso indica que o script não atendeu às expectativas de otimização. }

Notas: Essa User Story tem como objetivo aprimorar a eficiência do trabalho dos consultores, fornecendo-lhes um script Python personalizado para otimizar a análise de dados com bases de dados existentes.

Prioridade: Alta

Estimativa: 9 dias

Relação: US00 - US01 - US03

5.3.5 US04 - Configuração da estrutura dos dados

Persona: Analista de Dados

História: Como um analista de dados, desejo uma estrutura de banco de dados que me permita estruturar e analisar os dados provenientes de fontes governamentais, parceiros e CNPJs, a fim de formar um cubo de dados e facilitar a análise e manipulação dessas informações.

Critério de avaliação:

- Critério 1: Implementar uma estrutura de banco de dados que seja capaz de acomodar dados de várias fontes.

- Condição : { Verificar se a estrutura de banco de dados foi implementada de forma a permitir a acomodação de dados provenientes de fontes governamentais, parceiros e CNPJs. A estrutura deve ser capaz de receber e exibir esses dados de forma organizada. }
- Critério 2: Os dados recebidos devem ser automaticamente estruturados de maneira consistente, considerando a formação do cubo de dados.
 - Condição : { Confirmação de que os dados recebidos são estruturados de acordo com os requisitos de formação do cubo de dados. Considerando a transformação e organização dos dados de entrada para garantir que eles se encaixem nas dimensões e métricas do cubo de dados. }
- Critério 3: A visualização deve permitir a fácil identificação das métricas e dimensões do cubo de dados.
 - Condição : { Verificar se a visualização permite aos usuários identificar facilmente as métricas e dimensões do cubo de dados. Envolvendo a apresentação clara de rótulos, filtros ou funcionalidades de exploração de dados que tornam as métricas e dimensões visíveis e acessíveis. }

Teste de aceitação:

- Teste 1: Os dados provenientes de fontes governamentais, parceiros e CNPJs podem ser importados para a estrutura de banco de dados.
 - Aprovado: { Os dados provenientes das fontes especificadas podem ser importados com sucesso para a estrutura de dados, sem erros ou problemas significativos. A integração de dados foi realizada. }
 - Recusado: { Se ocorrerem erros durante a importação dos dados, se houver perda de informações críticas ou se os dados não estiverem disponíveis na visualização conforme o esperado. }
- Teste 2: A estruturação dos dados é realizada de forma automática e correta, considerando a formação do cubo de dados.
 - Aprovado: { A transformação e agregação dos dados foram executadas sem erros e que o cubo de dados foi construído conforme as especificações. }
 - Recusado: { Se a estruturação dos dados não for automática ou se for executada de forma incorreta, resultando em erros ou em uma formação inadequada do cubo de dados. }
- Teste 3: A visualização permite uma análise inicial dos dados, identificando as métricas e dimensões necessárias para análises futuras.

- Aprovado: { A estrutura dos dados permite uma análise inicial, identificando as métricas e dimensões necessárias para análises futuras. Os usuários podem explorar os dados com facilidade. }
- Recusado: { Se a estrutura não permitir uma análise inicial dos dados ou se os usuários não conseguirem identificar as métricas e dimensões necessárias. }

Notas: Esta User Story tem como objetivo simplificar o processo de importação e estruturação de dados, possibilitando a criação de um cubo de dados a partir das informações recebidas de diferentes fontes, facilitando assim as análises posteriores.

Prioridade: Média

Estimativa: 14 dias

Relação: US00 - US01 - US03

5.3.6 US05 - Análise de Consumo

Persona: Consultor de Marketing e Vendas

História: Como consultor, desejo ter acesso a uma plataforma de análise de potencial de consumo em várias macrorregiões, apresentada em formato de infográfico.

Critério de avaliação:

- Critério 1: A plataforma deve estar acessível para o consultor, com login e acesso seguros.
 - Condição : { Verificar se a plataforma permite que o consultor acesse com um sistema de login seguro. O acesso deve ser restrito ao consultor autorizado, e as medidas de segurança adequadas, como autenticação e autorização, devem ser implementadas. }
- Critério 2: A análise de potencial de consumo deve ser apresentada de forma clara e concisa em formato de infográfico, facilitando a interpretação dos dados.
 - Condição : { A plataforma deve gerar análises de potencial de consumo em formato de infográfico. Os infográficos devem ser claros, concisos e de fácil interpretação. }
- Critério 3: A infraestrutura da plataforma deve ser baseada na tecnologia AWS e/ou Open Source para garantir escalabilidade, alta disponibilidade e segurança dos dados.

- Condição : { A plataforma deve demonstrar escalabilidade, alta disponibilidade e segurança adequada dos dados, aproveitando os recursos da AWS para garantir essas características. }

Teste de aceitação:

- Teste 1: O consultor realiza login na plataforma de análise, com autenticação segura.
 - Aprovado: { O consultor pode realizar o login com sucesso na plataforma, utilizando um processo de autenticação segura, sem problemas de acesso não autorizado ou falhas de segurança. }
 - Recusado: { O consultor não consegue realizar o login com sucesso devido a problemas de autenticação, ou se a plataforma apresentar falhas de segurança que possam comprometer o acesso não autorizado. }
- Teste 2: A análise do potencial de consumo é apresentada em formato de infográfico.
 - Aprovado: { A análise do potencial de consumo é apresentada em formato de infográfico possibilitando uma compreensão rápida e clara dos dados. As informações estão bem organizadas e visualmente representadas. }
 - Recusado: { Se a apresentação da análise não estiver em formato de infográfico, se for confusa, desorganizada ou se não permitir uma compreensão rápida dos dados. }
- Teste 3: A plataforma é construída na infraestrutura da AWS e/ou Open Source, demonstrando escalabilidade, alta disponibilidade e segurança.
 - Aprovado: { A plataforma pode lidar com cargas de trabalho variáveis, está disponível de forma consistente e é protegida contra ameaças de segurança. }
 - Recusado: { Se a plataforma não for construída na infraestrutura da AWS e/ou Open Source, ou se não demonstrar escalabilidade, alta disponibilidade ou segurança adequadas. Implicando em vulnerabilidades de desempenho ou segurança na plataforma. }

Notas: Esta User Story tem como objetivo fornecer ao consultor da Integration uma plataforma de análise de potencial de consumo com visualização em infográficos, garantindo a escalabilidade e segurança da infraestrutura por meio da tecnologia de nuvem AWS.

Prioridade: Alta

Estimativa: 7 dias

Relação: US03 - US04 - US06

5.3.7 US06 - Filtros para Visualização da Distribuição de Consumo

Persona: Consultor de Marketing e Vendas

História: Como consultor, desejo ter acesso a filtros que permitam uma análise detalhada da distribuição de consumo, com foco nos atributos de produto, região e data, para tomar decisões mais informadas e estratégicas.

Critério de avaliação:

- Critério 1: A plataforma deve disponibilizar filtros interativos que permitam a análise da distribuição de consumo com base em produto, região e data.
 - Condição : { Verificar se a plataforma oferece filtros interativos que permitem aos usuários analisar a distribuição de consumo. Os filtros devem ser capazes de segmentar os dados de acordo com os critérios e exibir os resultados de forma clara. }
- Critério 2: Os filtros devem ser de fácil utilização, permitindo ao usuário ajustar os parâmetros rapidamente.
 - Condição : { Os filtros devem ser intuitivos e de fácil acesso, permitindo que os usuários ajustem os parâmetros de forma rápida e sem dificuldades. A interface do usuário deve permitir uma interação suave com os filtros. }

Teste de aceitação:

- Teste 1: A plataforma apresenta filtros de seleção para os atributos de produto, região e data.
 - Aprovado: { A Plataforma apresenta filtros de seleção para os atributos especificados de forma funcional. Ou seja, os filtros são visíveis, interativos e permitem que os usuários selecionem os atributos desejados. }
 - Recusado: { Se a plataforma não apresentar os filtros de seleção, se eles não estiverem disponíveis, não forem interativos ou se houver problemas na interface de seleção. }
- Teste 2 : Ao aplicar os filtros, a visualização da distribuição de consumo se ajusta de acordo com as seleções feitas.
 - Aprovado: { Quando aplicar os filtros, a visualização da distribuição de consumo se ajusta de acordo com as seleções feitas de forma rápida. A

- visualização é dinâmica e reflete as seleções de atributos, proporcionando uma análise personalizada. }
- Recusado: { Se ao aplicar os filtros, a visualização não se ajustar corretamente de acordo com as seleções feitas, se a atualização da visualização for lenta ou se houver erros na apresentação dos dados após a aplicação dos filtros. }

Notas: Essa User Story visa melhorar a capacidade do consultor de analisar a distribuição de consumo por meio de filtros que consideram atributos críticos, como produto, região e data, permitindo a tomada de decisões mais embasadas e estratégicas.

Prioridade: Média

Estimativa: N+T (Quantidade de trabalho + Tempo) - Necessário ver com a Integration

Relação: US05 - US04

6. Descrição dos dados

6.1 Identificação dos tipos de dados e suas características

O cenário atual do mercado exige uma análise criteriosa e abrangente dos dados para embasar decisões estratégicas bem informadas. Como parte do entregável da primeira *sprint* do módulo 8 (*Big Data*), que corresponde ao artefato “*Arquitetura de Ingestão de Dados do Parceiro*” o presente documento apresenta a identificação detalhada das bases de dados fornecidas pelo parceiro, explorando os diversos tipos de dados e suas características.

6.2. Dados CSV

O conjunto de dados disponível na pasta “Dados CSV” apresenta os micrdados da Pesquisa de Orçamentos Familiares (POF) referentes aos anos de 2017 a 2018. Os arquivos de dados presentes nesta compilação são um reflexo das variáveis exploradas em diversas publicações divulgadas, proporcionando uma visão ampla e detalhada das dinâmicas socioeconômicas e de consumo. Estes dados abrangem uma variedade de temas, desde despesas e rendimentos familiares até indicadores de qualidade de vida no Brasil. Além disso, os micrdados incluem informações demográficas e socioeconômicas, como número de cômodos, idade e educação dos moradores, bem como detalhes sobre a

posse de bens duráveis e características do trabalho. A seguir, é apresentada uma análise aprofundada destes microdados, esclarecendo sua relevância e características.

6.2.1 Tipos de Dados e suas Características

6.2.1.1 Informações Geográficas e Estratificação Social

UF (int64): Identifica a unidade federativa, ajudando a localizar geograficamente os dados;

ESTRATO_POF (int64): Indica a estratificação social, o que pode ser crucial para entender diferentes comportamentos de consumo e necessidades.

6.2.1.2 Identificação e Situação Residencial

As colunas como TIPO_SITUACAO_REG (int64), COD_UPA (int64), NUM_DOM (int64), e NUM_UC (int64) proporcionam uma visão detalhada da situação residencial e identificação dos domicílios, que podem ser essenciais para analisar o mercado de aluguel.

6.2.1.3 Dados Econômicos

Colunas como V9001 (int64), V9002 (int64), V8000 (float64), e RENDA_TOTAL (float64) fornecem informações sobre o poder aquisitivo e condições econômicas, que são fundamentais para entender a demanda e a capacidade de pagamento dos grupos pesquisados.

6.2.1.4 Fatores de Correção e Peso

DEFLATOR (float64), FATOR_ANUALIZACAO (int64), PESO (float64), e PESO_FINAL (float64) ajudam na normalização ou ajuste dos dados para garantir precisão e relevância.

6.3 Dados CNPJ

A base de dados do Cadastro Nacional de Pessoa Jurídica (CNPJ) é um recurso que fornece informações detalhadas sobre entidades empresariais no Brasil. As colunas desta base são estruturadas para oferecer *insights* sobre diferentes aspectos

corporativos, como identificação, localização, atividade econômica e canais de contato. Ao explorar esses dados, pode-se entender a estrutura e operações dessas entidades, mas também estabelecer uma fundação sólida para análises futuras que podem impulsionar decisões de negócios.

6.3.1 Tipos de Dados e suas Características

6.3.1.1 Identificação da Empresa

cnpj (int64): Número completo do CNPJ que identifica de maneira única cada empresa.

cnpj_basico (int64): Parte básica do CNPJ.

cnpj_ordem (int64): Número de ordem do CNPJ.

cnpj_dv (int64): Dígito verificador do CNPJ.

identificador_matriz_filial (int64): Indica se o CNPJ pertence a uma matriz ou filial.

6.3.1.2 Status Cadastral

situacao_cadastral (int64): Status atual do cadastro da empresa.

motivo_situacao_cadastral (int64): Motivo pelo qual a empresa se encontra na situação cadastral informada.

6.3.1.3 Informações de Atividade e Localização:

id_pais (float64): Identificação do país.

cnae_fiscal_principal (int64): Classificação Nacional de Atividades Econômicas principal da empresa.

id_municipio (float64): Identificação do município.

id_municipio_rf (int64): Identificação do município na Receita Federal.

6.3.1.4 Contato

`ddd_1, ddd_2` (float64): Códigos de área para telefonia.

`telefone_2` (float64): Número de telefone secundário.

`ddd_fax` (float64): Código de área para fax.

6.3.1.5 Número de linhas e colunas

O conjunto de dados é dividido em cinco bases distintas, cada uma com 33 colunas. Abaixo estão detalhadas as quantidades de linhas em cada base:

- `cnpj_1`: 409,357 linhas
- `cnpj_2`: 318,897 linhas
- `cnpj_3`: 44,974 linhas
- `cnpj_4`: 78,748 linhas
- `cnpj_5`: 64,565 linhas

Total acumulado entre as cinco bases: 916,541 linhas.

6.4 API

6.4.1 Endpoint

<https://intelfunctiongetdata.azurewebsites.net/api/InteliFunctionGetData>

Método Aceito: GET

Token de uso: pZh3gmJW_87epswrWDuB7CvQle-KqjsVh2ZZJUaifiXd4AzFuOEy98w==

6.4.2 Parâmetros de Consulta (Query Parameters)

code: Este é o *token* de autenticação necessário para fazer a requisição. É um tipo de dado *string*.

table: Este parâmetro informa à API para qual tabela da base de dados a chamada HTTP GET será feita. É um tipo de dado *string* e os valores possíveis são: "Category", "Establishment" e "Sale".

saleDate (opcional): Parâmetro de filtro por data de venda. É um tipo de dado *string* no formato *yyyy-mm-dd* (ex.: 2023-09-23).

saleCnpj (opcional): Parâmetro de filtro por CNPJ. É um tipo de dado *string* que deve ser um CNPJ válido.

saleCategory (opcional): Parâmetro de filtro por nome de categoria. É um tipo de dado *string* que deve ser um nome válido de categoria.

6.4.3 Resposta da API

A resposta da API será em formato JSON. Cada entrada no JSON corresponderá a uma coluna na tabela de dados correspondente e o tipo de dado de cada entrada será determinado pelo tipo de dado da coluna na base de dados.

6.4.4 Tratamento de Exceções

A API está preparada para lidar com quatro tipos de exceções relacionadas a parâmetros faltando ou inválidos, fornecendo mensagens de erro específicas para ajudar a diagnosticar e corrigir problemas com as requisições.

6.4.5 Exemplo de Código

O exemplo de código fornecido mostra como fazer uma requisição GET para a API usando a biblioteca *requests* em Python, passando os parâmetros de consulta como um dicionário e tratando a resposta.

7. Análise exploratória

7.1 CNPJs

A primeira análise exploratória realizada consiste na avaliação dos dados provenientes do Cadastro Nacional de Pessoas Jurídicas, conhecido como CNPJ. Esse processo é fundamental para compreender a estrutura dos arquivos csv disponibilizados pelo parceiro. O CNPJ é um registro obrigatório para todas as empresas ativas, tornando-se uma fonte valiosa de informações sobre a economia e o mercado de trabalho de uma nação.

Para este projeto, foram fornecidos cinco arquivos csv, cada um contendo informações de empresas com uma Classificação Nacional de Atividades Econômicas (CNAE) diferente. O primeiro passo consiste na configuração do ambiente, o chamado *setup*, que engloba a preparação e organização do ambiente de trabalho. Isso inclui a conexão com o Drive, o *download* das bibliotecas necessárias e o acesso aos arquivos.

Posteriormente, realiza-se a análise que coleta informações relevantes a partir dos dados disponibilizados. A seguir, apresentamos uma breve descrição de cada arquivo.

7.1.1. Arquivo “CNPJ 1”

Este dataset contém as empresas cadastradas somente com o CNAE: "5611201", referente à "Restaurantes e similares", de acordo com o Contabilizei. Além disso, há 515.874 CNPJs neste arquivo. O primeiro gráfico foi feito de acordo com a coluna "sigla_uf", com o código abaixo.



```
cnpjjs_1['sigla_uf'].value_counts().plot(kind='bar')
plt.title("Distribuição por UF")
plt.xlabel("UF")
plt.ylabel("Contagem")
plt.show()
```

Figura 10: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

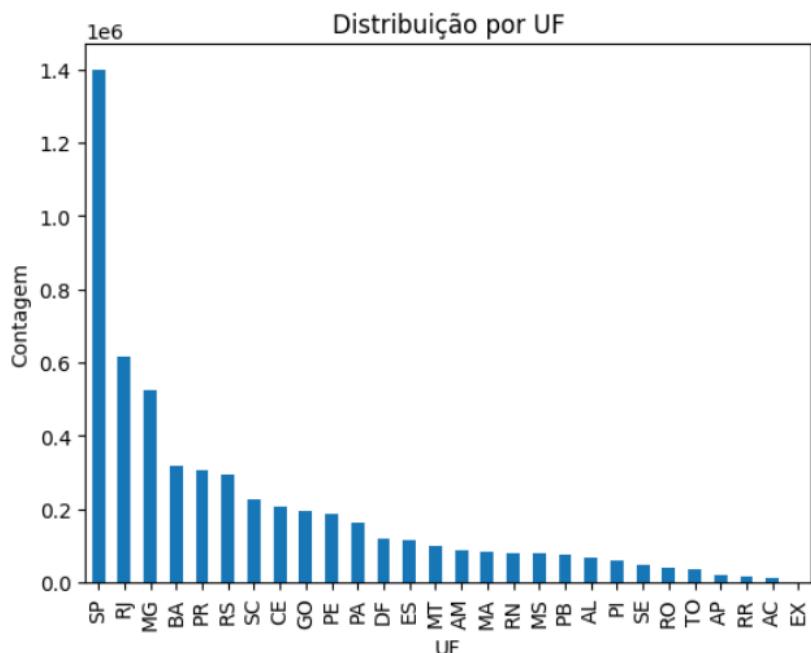


Figura 11: Gráfico “Distribuição por UF” - CNPJ1

Fonte: Autoria Própria

Com esse gráfico é possível observar que os três estados que mais tem "Restaurantes e similares" cadastrados são: São Paulo, Rio de Janeiro e Minas Gerais, respectivamente. Além

disso, há uma inconsistência nos dados, já que o último UF mostrado no gráfico é "EX", que significa "Exterior", este dado não deveria aparecer, já que o CNPJ é um documento brasileiro. O mesmo pode-se observar no gráfico a seguir.



```
cnpj1['id_pais'].value_counts().plot(kind='bar')
plt.title("Distribuição por País")
plt.xlabel("ID do País")
plt.ylabel("Contagem")
plt.show()
```

Figura 12: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

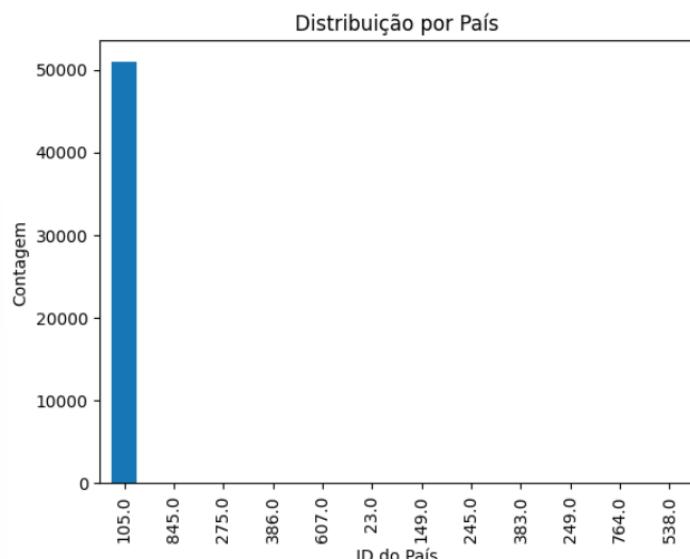


Figura 13: Gráfico “Distribuição por País” - CNPJ1

Fonte: Autoria Própria

O gráfico "Distribuição por País" comprova que há, pelo menos, 11 CNPJs que estão cadastrados em outro país. A primeira coluna, com o ID de "105", é do Brasil. O gráfico a seguir mostra quais empresas são matrizes e quais são filiais.



```
cnpj_1['identificador_matriz_filial'].value_counts().plot(kind='bar')
plt.title("Matriz e Filial")
plt.xlabel("ID")
plt.ylabel("Contagem")
plt.show()
```

Figura 14: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

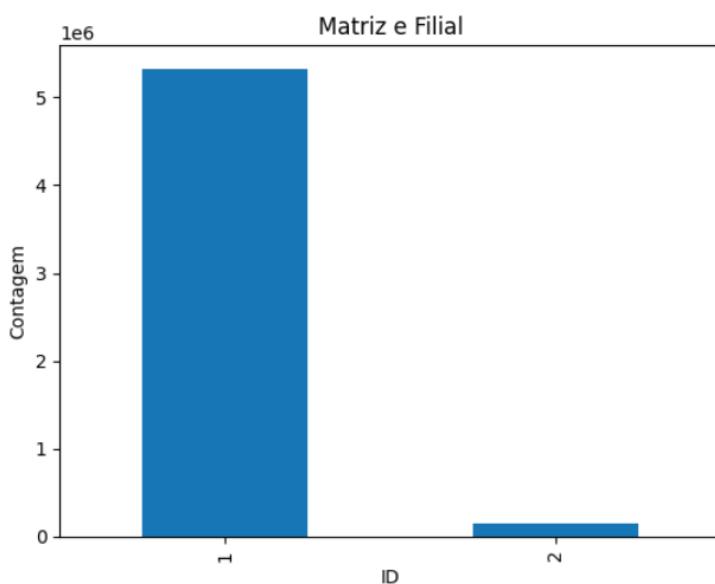


Figura 15: Gráfico “Matriz e Filial” - CNPJ1

Fonte: Autoria Própria

É importante ressaltar neste gráfico que o ID "1" é Matriz, e o ID "2" é Filial. Com isso, pode-se observar que a grande parte dos restaurantes são matriz do CNPJ.

7.1.2. Arquivo “CNPJ 2”

Este dataset contém as empresas cadastradas somente com o CNAE: "5611203", referente à "Lanchonetes, casas de chá, de sucos e similares", de acordo com o Contabilizei. Além disso, há 577.735 CNPJs neste arquivo. O primeiro gráfico foi feito de acordo com a coluna sigla_uf, com o código abaixo.



```
cnpj_1['sigla_uf'].value_counts().plot(kind='bar')
plt.title("Distribuição por UF")
plt.xlabel("UF")
plt.ylabel("Contagem")
plt.show()
```

Figura 16: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

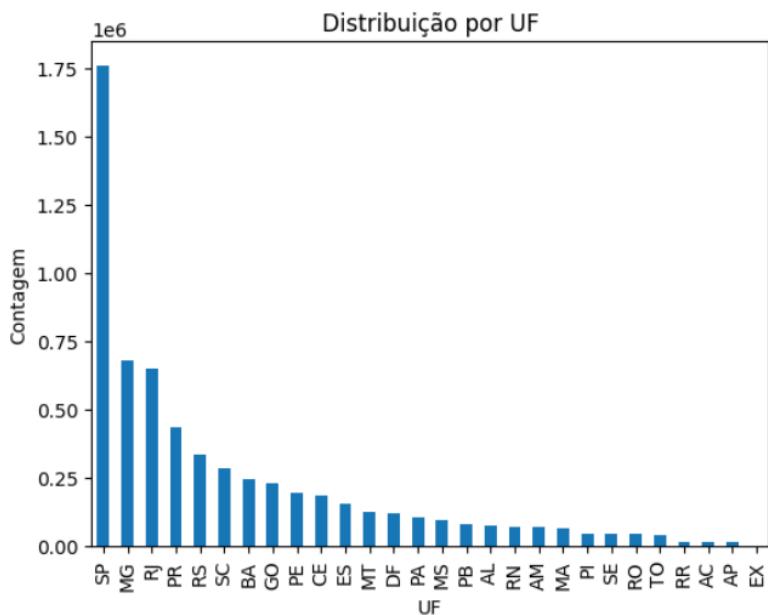


Figura 17: Gráfico “Distribuição por UF” - CNPJ2

Fonte: Autoria Própria

Com esse gráfico é possível observar que os três estados que mais tem "Lanchonetes, casas de chá, de sucos e similares" cadastrados são: São Paulo, Minas Gerais e Rio de Janeiro, respectivamente. Além disso, há uma inconsistência nos dados, já que o último UF mostrado no gráfico é "EX", que significa "Exterior", este dado não deveria aparecer, já que o CNPJ é um documento brasileiro. O mesmo pode-se observar no gráfico a seguir.



```
cnpj_1['id_pais'].value_counts().plot(kind='bar')
plt.title("Distribuição por País")
plt.xlabel("ID do País")
plt.ylabel("Contagem")
plt.show()
```

Figura 18: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

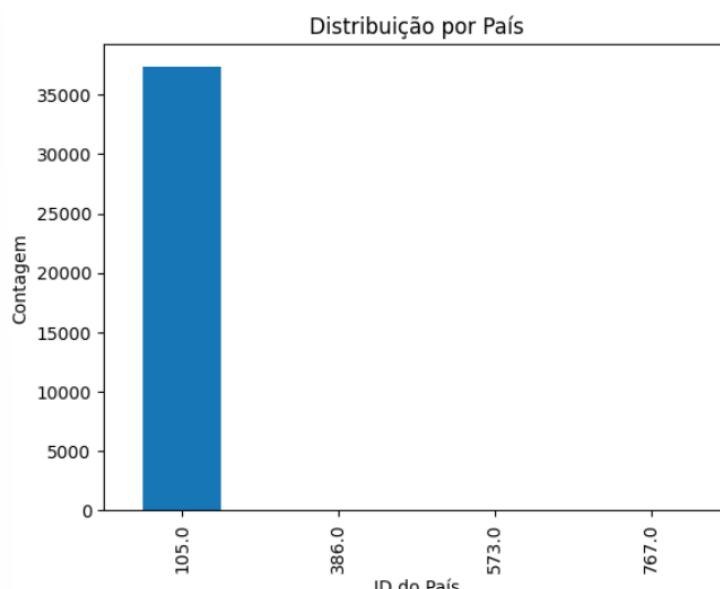


Figura 19: Gráfico “Distribuição por País” - CNPJ2

Fonte: Autoria Própria

O gráfico "Distribuição por País" comprova que há, pelo menos, 3 CNPJs que estão cadastrados em outro país. A primeira coluna, com o ID de "105", é o Brasil. O gráfico a seguir mostra quais empresas são matriz e quais são filiais.



```
cnpj_1['identificador_matriz_filial'].value_counts().plot(kin
d='bar')
plt.title("Matriz e Filial")
plt.xlabel("ID")
plt.ylabel("Contagem")
plt.show()
```

Figura 20: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

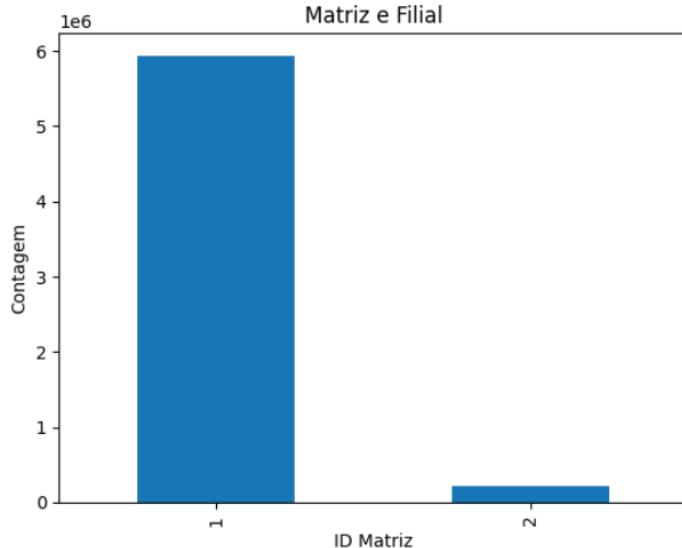


Figura 21: Gráfico “Matriz e Filial” - CNPJ2

Fonte: Autoria Própria

É importante ressaltar neste gráfico que o ID "1" é Matriz, e o ID "2" é Filial. Com isso, pode-se observar que a grande parte das lanchonetes são matriz do CNPJ.

7.1.3. Arquivo “CNPJ 3”

Este dataset contém as empresas cadastradas somente com o CNAE: "5611204", referente à "Bares e outros estabelecimentos especializados em servir bebidas, sem entretenimento", de acordo com o Contabilizei. Além disso, há 184.274 CNPJs neste arquivo. O primeiro gráfico foi feito de acordo com a coluna sigla_uf, com o código abaixo.



```
cnpjjs_1['sigla_uf'].value_counts().plot(kind='bar')
plt.title("Distribuição por UF")
plt.xlabel("UF")
plt.ylabel("Contagem")
plt.show()
```

Figura 22: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

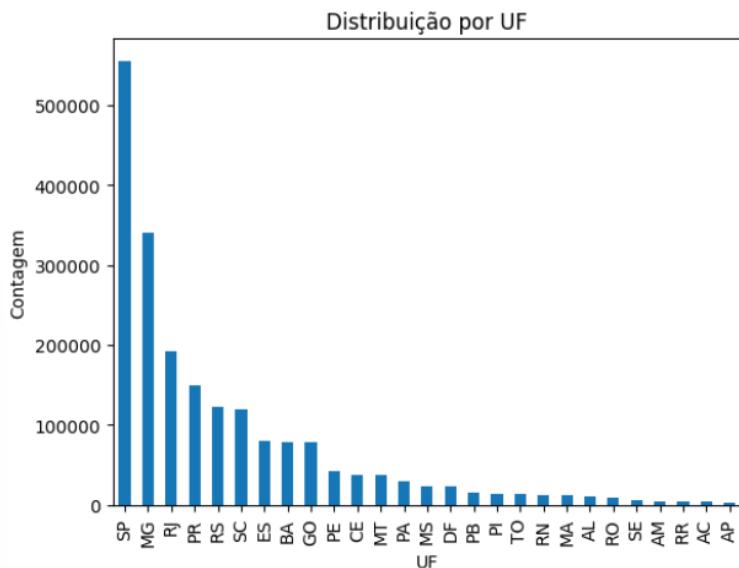


Figura 23: Gráfico “Distribuição por UF” - CNPJ3

Fonte: Autoria Própria

Com esse gráfico é possível observar que os três estados que mais tem "Bares e outros estabelecimentos especializados em servir bebidas, sem entretenimento" cadastrados são: São Paulo, Minas Gerais e Rio de Janeiro, respectivamente. O gráfico a seguir demonstra que a inconsistência encontrada nos últimos arquivos não foi encontrada neste.



```
cnpjjs_1['id_pais'].value_counts().plot(kind='bar')
plt.title("Distribuição por País")
plt.xlabel("ID do País")
plt.ylabel("Contagem")
plt.show()
```

Figura 24: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

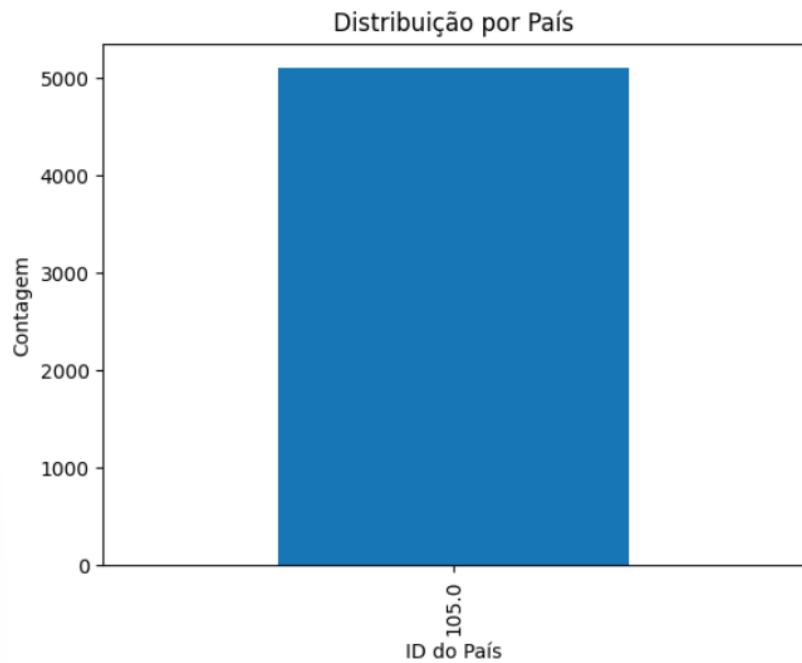


Figura 24: Gráfico “Distribuição por País” - CNPJ3

Fonte: Autoria Própria

A única coluna, com o ID de "105", é o Brasil. O gráfico a seguir mostra quais empresas são matriz e quais são filiais.



```
cnpjjs_1['identificador_matriz_filial'].value_counts().plot(kind='bar')
plt.title("Matriz e Filial")
plt.xlabel("ID")
plt.ylabel("Contagem")
plt.show()
```

Figura 26: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

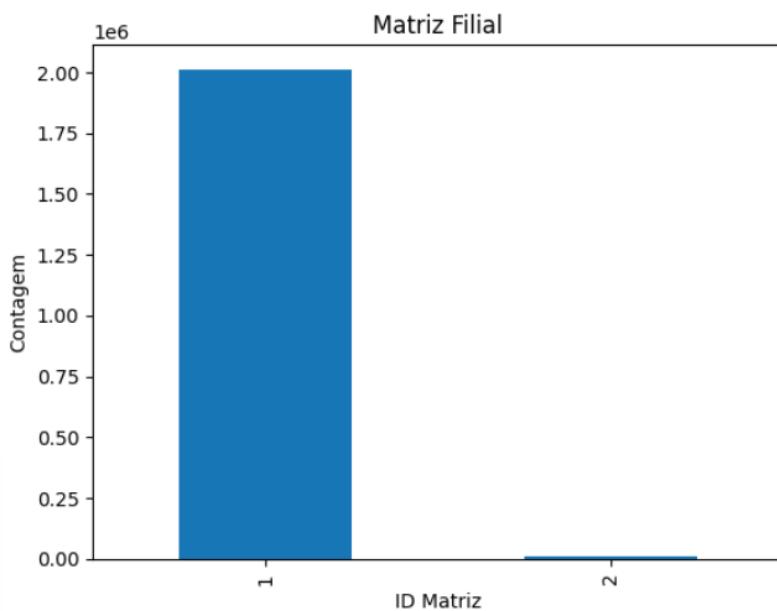


Figura 23: Gráfico “Matriz e Filial” - CNPJ3

Fonte: Autoria Própria

É importante ressaltar neste gráfico que o ID "1" é Matriz, e o ID "2" é Filial. Com isso, pode-se observar que praticamente todos os bares são matrizes do CNPJ.

7.1.3. Arquivo “CNPJ 4”

Este dataset contém as empresas cadastradas somente com o CNAE: "5611205", referente à "Bares e outros estabelecimentos especializados em servir bebidas, com entretenimento", de acordo com o Contabilizei. Além disso, há 92.612 CNPJs neste arquivo. O primeiro gráfico foi feito de acordo com a coluna `sigla_uf`, com o código abaixo.



```
cnpj3['sigla_uf'].value_counts().plot(kind='bar')
plt.title("Distribuição por UF")
plt.xlabel("UF")
plt.ylabel("Contagem")
plt.show()
```

Figura 28: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

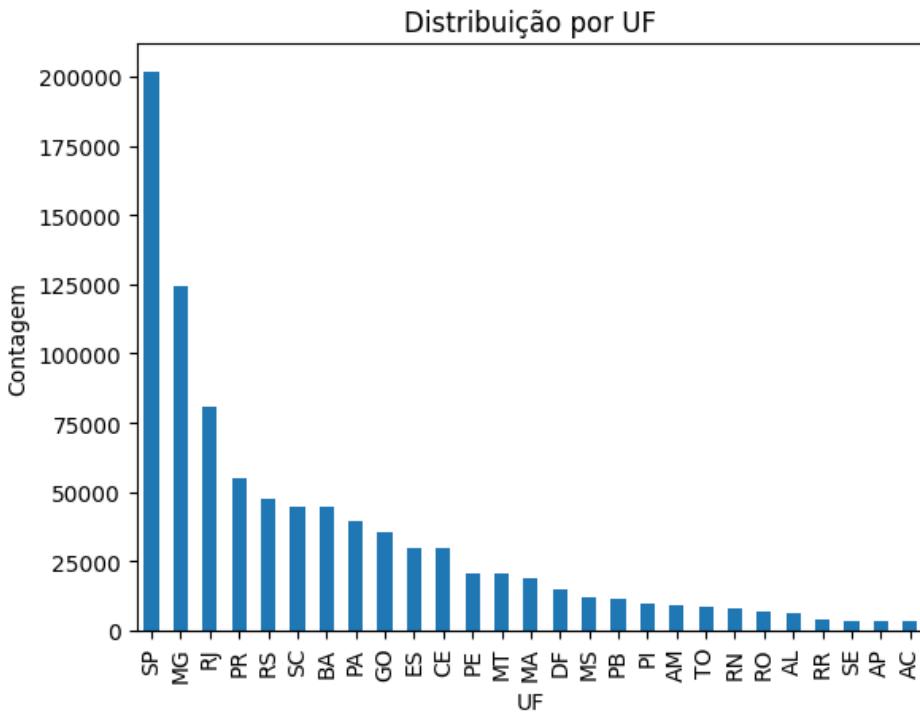


Figura 29: Gráfico “Distribuição por UF” - CNPJ4

Fonte: Autoria Própria

Com esse gráfico é possível observar que os três estados que mais tem "Bares e outros estabelecimentos especializados em servir bebidas, com entretenimento" cadastrados são: São Paulo, Minas Gerais e Rio de Janeiro, respectivamente. O gráfico a seguir demonstra que a inconsistência encontrada nos últimos arquivos não foi encontrada neste.



```
cnpjjs_1['id_pais'].value_counts().plot(kind='bar')
plt.title("Distribuição por País")
plt.xlabel("ID do País")
plt.ylabel("Contagem")
plt.show()
```

Figura 30: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

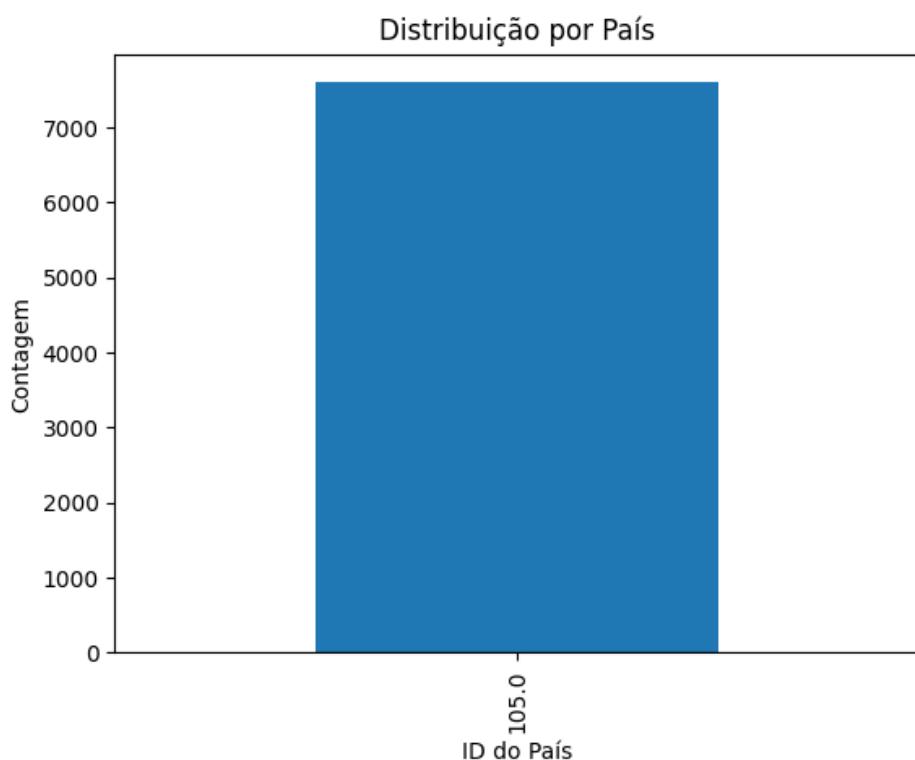


Figura 31: Gráfico “Distribuição por País” - CNPJ4

Fonte: Autoria Própria

A única coluna, com o ID de "105", é o Brasil. O gráfico a seguir mostra quais empresas são matrizes e quais são filiais.

```
cnpj_1['identificador_matriz_filial'].value_counts().plot(kind='bar')
plt.title("Matriz e Filial")
plt.xlabel("ID")
plt.ylabel("Contagem")
plt.show()
```

Figura 32: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

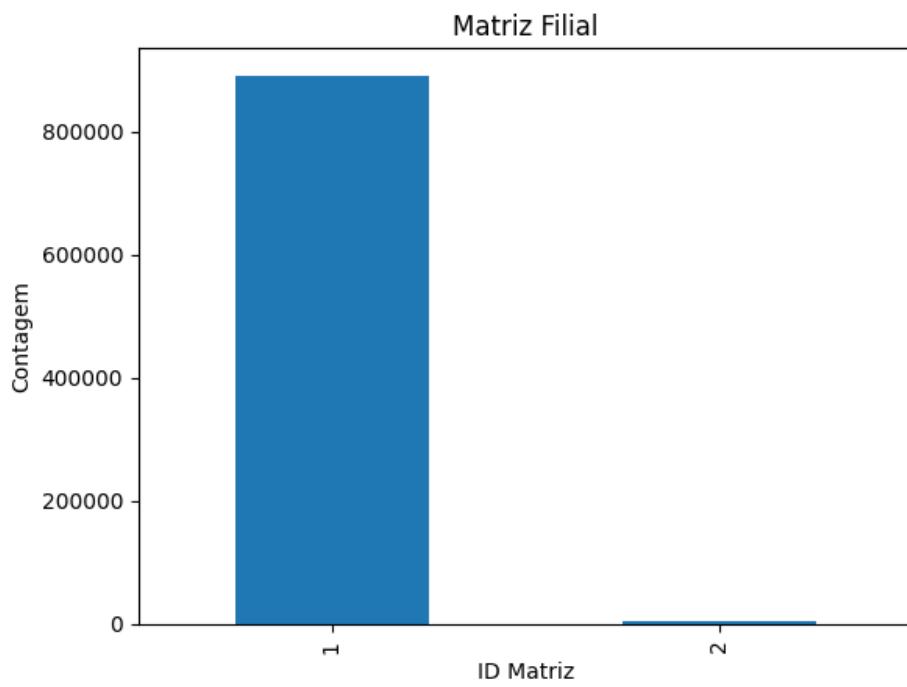


Figura 33: Gráfico “Matriz e Filial” - CNPJ4

Fonte: Autoria Própria

É importante ressaltar neste gráfico que o ID "1" é Matriz, e o ID "2" é Filial. Com isso, pode-se observar que praticamente todos os bares são matrizes do CNPJ.

7.1.3. Arquivo “CNPJ 5”

Este dataset contém as empresas cadastradas com os CNAEs são:

- "4712100" - "Comércio varejista de mercadorias em geral, com predominância de produtos alimentícios - minimercados, mercearias e armazéns";
- "4711302" - "Comércio varejista de mercadorias em geral, com predominância de produtos alimentícios - supermercados";
- "4711301" - "Comércio varejista de mercadorias no geral, com predominância de produtos alimentícios - hipermercados";
- "4691500" - "Comércio atacadista de mercadorias em geral, com predominância de produtos alimentícios";
- "4637107" - "Comércio atacadista de chocolates, confeitos, balas, bombons e semelhantes";

Essas atividades foram buscadas no site do Contabilizei. Além disso, há 651.730 CNPJs neste arquivo. O primeiro gráfico foi feito de acordo com a coluna `sigla_uf`, com o código abaixo.



```
cnpj_1['sigla_uf'].value_counts().plot(kind='bar')
plt.title("Distribuição por UF")
plt.xlabel("UF")
plt.ylabel("Contagem")
plt.show()
```

Figura 34: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

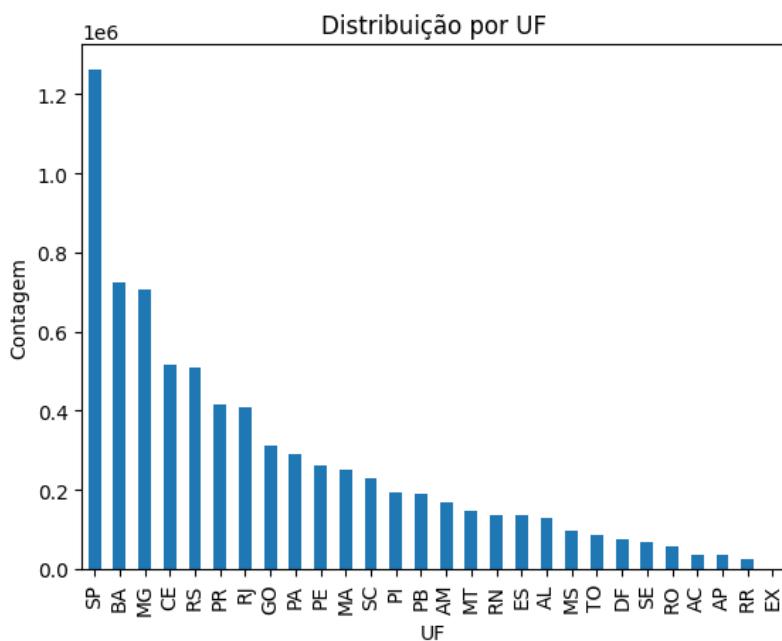


Figura 35: Gráfico “Distribuição por UF” - CNPJ5

Fonte: Autoria Própria

Com esse gráfico é possível observar que os três estados que mais tem essas atividades cadastradas são: São Paulo, Bahia e Minas Gerais, respectivamente. Além disso, há uma inconsistência nos dados, já que o último UF mostrado no gráfico é "EX", que significa "Exterior", este dado não deveria aparecer, já que o CNPJ é um documento brasileiro. O mesmo pode-se observar no gráfico a seguir.



```
cnpj_1['id_pais'].value_counts().plot(kind='bar')
plt.title("Distribuição por País")
plt.xlabel("ID do País")
plt.ylabel("Contagem")
plt.show()
```

Figura 36: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

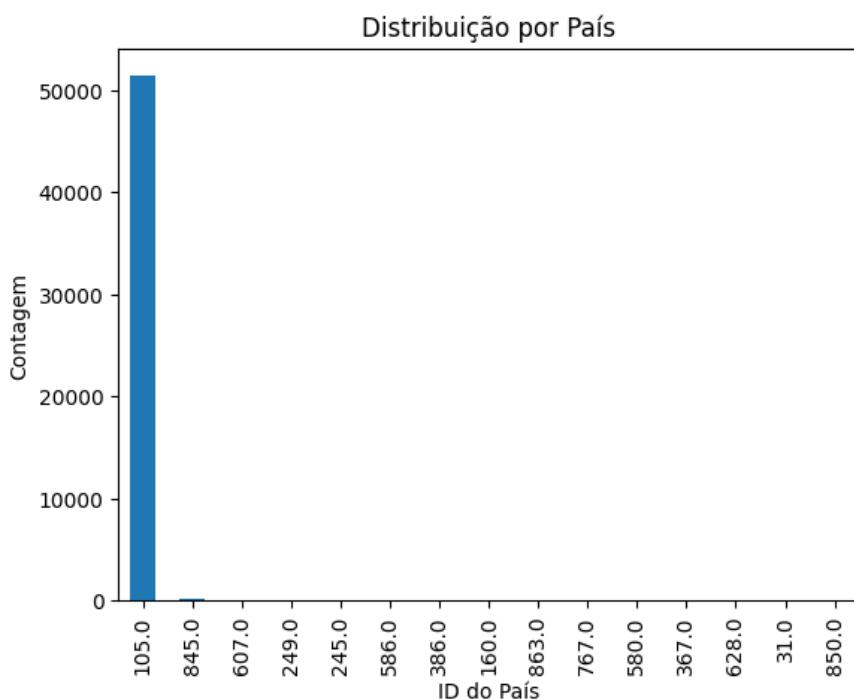


Figura 37: Gráfico “Distribuição por País” - CNPJ5

Fonte: Autoria Própria

O gráfico "Distribuição por País" comprova que há, pelo menos, 14 CNPJs que estão cadastrados em outro país. A primeira coluna, com o ID de "105", é do Brasil. O gráfico a seguir mostra quais empresas são matriz e quais são filiais.



```
cnpj_1['identificador_matriz_filial'].value_counts().plot(kind='bar')
plt.title("Matriz e Filial")
plt.xlabel("ID")
plt.ylabel("Contagem")
plt.show()
```

Figura 38: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

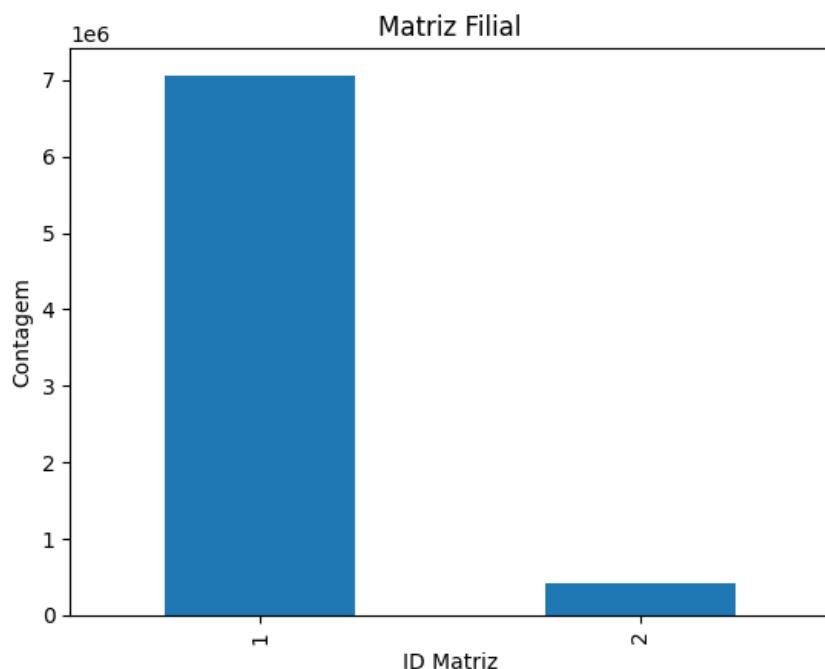


Figura 39: Gráfico “Matriz e Filial” - CNPJ5
Fonte: Autoria Própria

É importante ressaltar neste gráfico que o ID "1" é Matriz, e o ID "2" é Filial. Com isso, pode-se observar que há uma predominância de matrizes do CNPJ.

7.2. Dados do Governo

A segunda análise exploratória feita foi a dos dados disponibilizados pelo governo sobre o POF, Pesquisa de Orçamentos Familiares, os arquivos que serão explicados a seguir estão separados e serão juntados em um próximo momento. Além disso, foi disponibilizado um arquivo Excel que contém o dicionário de variáveis e colunas, que será utilizado para a substituição de valores int para object.

O primeiro passo é realizar a configuração do Setup, que inclui a preparação e organização do ambiente, ou seja, realizar a conexão com o Drive, baixar as bibliotecas e acessar o arquivo. A seguir é realizada a análise que coleta informações sobre os dados disponibilizados. Abaixo há uma descrição sobre cada arquivo.

7.2.1. Aluguel estimado

O primeiro passo é realizar a configuração do Setup, que inclui a preparação e organização do ambiente, ou seja, realizar a conexão com o Drive, baixar as bibliotecas e acessar o arquivo. A seguir é realizada a análise que coleta informações sobre os dados disponibilizados. Abaixo há uma descrição sobre cada arquivo.

```
● ● ●  
aluguel_estimado['UF'] = aluguel_estimado['UF'].replace({11: 'Rondônia', 12: 'Acre', 13: 'Amazonas',  
14: 'Roraima', 15: 'Pará', 16: 'Amapá', 17:'Tocantins', 21:'Maranhão', 22:'Piauí', 23:'Ceará',  
24:'Rio Grande do Norte', 25:'Paraíba', 26:'Pernambuco', 27:'Alagoas', 28:'Sergipe', 29:'Bahia',  
31:'Minas Gerais', 32:'Espírito Santo', 33:"Rio de Janeiro", 35:'São Paulo', 41:'Paraná', 42:'Santa  
Catarina', 43:'Rio Grande do Sul', 50:'Mato Grosso do Sul', 51:'Mato Grosso', 52: 'Goiás',  
53:'Distrito Federal'})
```

Figura 40: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

Um detalhe que foi reparado é que os números dos estados estão divididos nas regiões, por exemplo: 31:'Minas Gerais', 32:'Espírito Santo', 33:"Rio de Janeiro", 35:'São Paulo', todos começam com "3" pois são da região Sudeste. Com o código acima, foi realizado a substituição dos valores inteiros para object. Abaixo, segue o código e o gráfico desta coluna.



```
aluguel_estimado['UF'].value_counts().plot(kind='bar')
plt.title("Distribuição por UF")
plt.xlabel("UF")
plt.ylabel("Contagem")
plt.show()
```

Figura 41: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

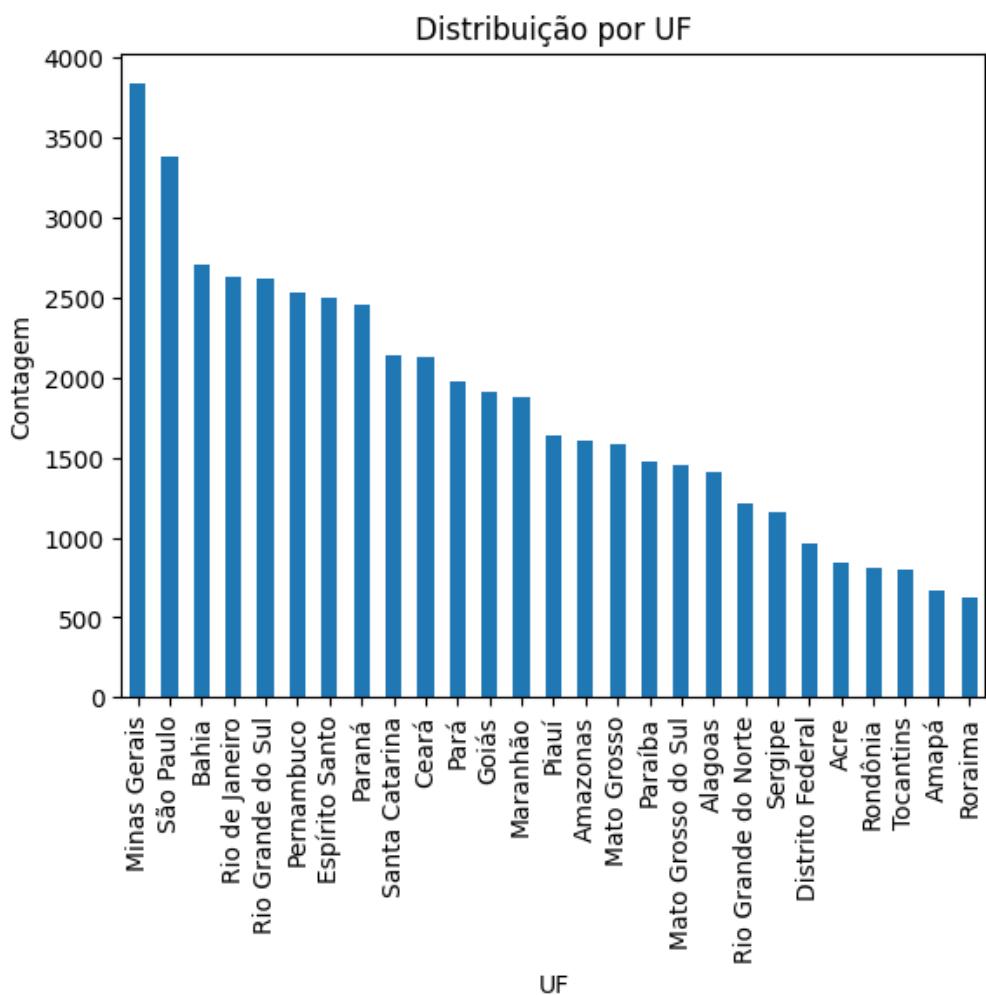


Figura 42: Gráfico “Distribuição por UF” - Aluguel Estimado

Fonte: Autoria Própria

Com esse gráfico, pode-se classificar que os estados que mais têm imóveis alugados são, respectivamente: Minas Gerais, São Paulo e Bahia. O código a seguir foi feito o mesmo processo, só que na coluna "TIPO_SITUACAO_REG", onde mostra se o domicílio se localiza em uma cidade, Urbano, ou em uma área Rural, a seguir é feito o gráfico mostrando essa diferença.



```
aluguel_estimado['TIPO_SITUACAO_REG'] = aluguel_estimado['TIPO_SITUACAO_REG'].replace({1: 'Urbano',  
2: 'Rural'})
```

Figura 43: Código para plotagem de um gráfico em python

Fonte: Autoria Própria



```
aluguel_estimado['TIPO_SITUACAO_REG'].value_counts().plot(kind='pie', autopct='%1.1f%%')  
plt.title("Tipo de situação regional")  
plt.gca().set_ylabel('')  
plt.show()
```

Figura 44: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

Tipo de situação regional

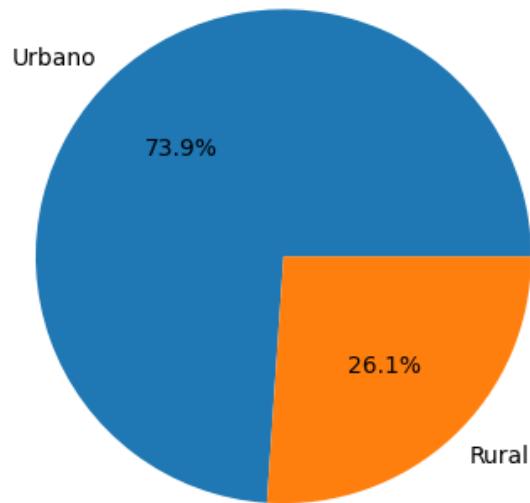


Figura 45: Gráfico “Tipo de situação regional” - Aluguel Estimado

Fonte: Autoria Própria

O gráfico acima demonstra que mais de 73% das casas alugadas estão na cidade, enquanto 26% delas estão na zona rural. Este dado reflete que mesmo que a maioria das pessoas do Brasil moram na cidade, ainda há um número significativo na zona rural. A seguir, o código que também segue o mesmo processo, mas com a coluna "COD_IMPUT_VALOR", mostrando se o valor do aluguel foi ou não imputado, além do gráfico do mesmo.



```
aluguel_estimado['COD_IMPUT_VALOR'] = aluguel_estimado['COD_IMPUT_VALOR'].replace({0: 'Valor não foi  
imputado', 1: 'Valor foi imputado'})  
  
aluguel_estimado['COD_IMPUT_VALOR'].value_counts().plot(kind='pie', autopct='%1.1f%%')  
plt.title("O valor do aluguel estimado foi imputado")  
plt.gca().set_ylabel('')  
plt.show()
```

Figura 46: Código para plotagem de um gráfico em python

Fonte: Autoria Própria



Figura 47: Gráfico “O valor do aluguel estimado foi imputado” - Aluguel Estimado

Fonte: Autoria Própria

Com esse gráfico é possível observar que a maior parte do Brasil não teve o seu valor imputado, isso demonstra a realidade do país.

7.2.2. Domicílio

O dataset domicílio contém informações sobre os domicílios do Brasil. Este é composto por 38 colunas e mais de 57 mil linhas de dados. Acessando o dicionário de variáveis, foi realizado um .replace em 5 colunas que foram consideradas mais importantes para uma primeira análise, o código a seguir demonstra essa mudança.



```
aluguel_estimado['UF'] = aluguel_estimado['UF'].replace({11: 'Rondônia', 12: 'Acre', 13: 'Amazonas',
14: 'Roraima', 15: 'Pará', 16: 'Amapá', 17:'Tocantins', 21:'Maranhão', 22:'Piauí', 23:'Ceará',
24:'Rio Grande do Norte', 25:'Paraíba', 26:'Pernambuco', 27:'Alagoas', 28:'Sergipe', 29:'Bahia',
31:'Minas Gerais', 32:'Espírito Santo', 33:"Rio de Janeiro", 35:'São Paulo', 41:'Paraná', 42:'Santa
Catarina', 43:'Rio Grande do Sul', 50:'Mato Grosso do Sul', 51:'Mato Grosso', 52: 'Goiás',
53:'Distrito Federal'})
```

Figura 48: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

Com o código acima, foi realizado a substituição dos valores inteiros para object. Abaixo, segue o código e o gráfico desta coluna.



```
aluguel_estimado['UF'].value_counts().plot(kind='bar')
plt.title("Distribuição por UF")
plt.xlabel("UF")
plt.ylabel("Contagem")
plt.show()
```

Figura 49: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

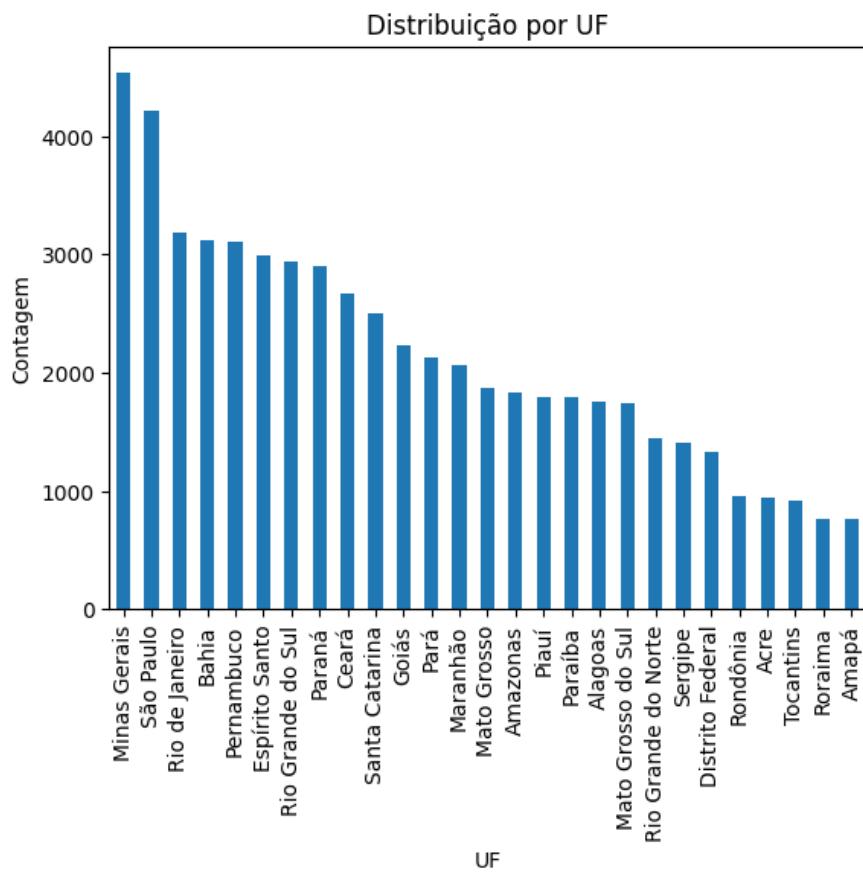


Figura 50: Gráfico “Distribuição por UF” - Domicílio

Fonte: Autoria Própria

Com esse gráfico, pode-se classificar que os estados com mais domicílios são, respectivamente: Minas Gerais, São Paulo e Rio de Janeiro, lembrando que isso não

significa que há mais habitantes. O código a seguir foi feito o mesmo processo, só que na coluna "TIPO_SITUACAO_REG", onde mostra se o domicílio se localiza em uma cidade, Urbano, ou em uma área Rural, a seguir é feito o gráfico mostrando essa diferença.



```
aluguel_estimado['TIPO_SITUACAO_REG'] = aluguel_estimado['TIPO_SITUACAO_REG'].replace({1: 'Urbano',  
2: 'Rural'})
```

Figura 51: Código para plotagem de um gráfico em python

Fonte: Autoria Própria



```
aluguel_estimado['TIPO_SITUACAO_REG'].value_counts().plot(kind='pie', autopct='%1.1f%%')  
plt.title("Tipo de situação regional")  
plt.show()
```

Figura 52: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

Distribuição por Situação Regional

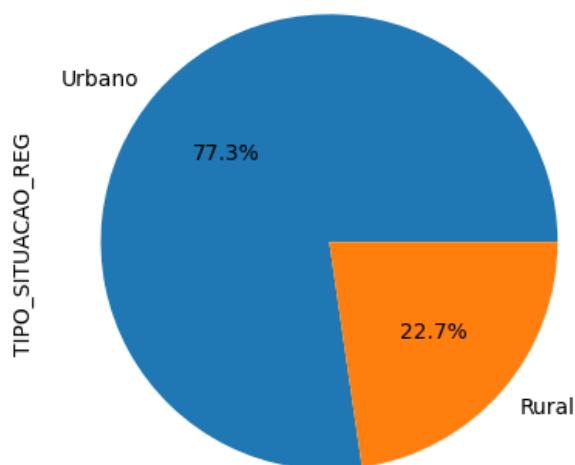


Figura 53: Gráfico “Tipo de situação regional” - Domicílio

Fonte: Autoria Própria

O gráfico acima demonstra que mais de 77% das casas estão na cidade, enquanto 22% delas estão na zona rural. Este dado reflete que mesmo que a maioria das pessoas do Brasil moram na cidade, ainda há um número significativo na zona rural. A seguir, o código que também segue o mesmo processo, mas com a coluna "V0201", indicando o tipo de domicílio.



```
domicilio['V0201'] = domicilio['V0201'].replace({1: 'Casa', 2: 'Apartamento', 3:'Habitação em casa de cômodos, cortiço ou cabeça de porco'})
```

Figura 54: Código para plotagem de um gráfico em python

Fonte: Autoria Própria



```
domicilio['V0201'].value_counts().plot(kind='pie', autopct='%1.1f%%')  
plt.title("Distribuição por tipo de domicílio")  
plt.show()
```

Figura 55: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

Distribuição por tipo de domicílio

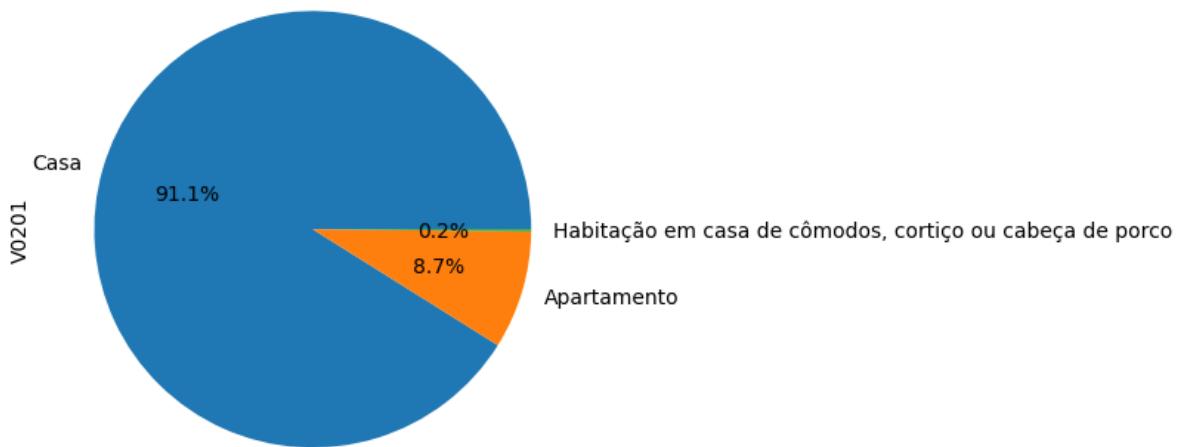


Figura 56: Gráfico “Distribuição por tipo de domicílio” - Domicílio

Fonte: Autoria Própria

Com esse gráfico é possível observar que a maior parte da população brasileira, cerca de 90%, habita em Casa, e mais de 8% mora em Apartamento. A seguir a coluna "V0217", que indica a propriedade do domicílio, passa pelo mesmo processo.



```
domicilio['V0217'] = domicilio['V0217'].replace({1: 'Próprio de algum morador – já pago', 2: 'Próprio de algum morador – ainda pagando', 3:'Alugado', 4:'Cedido por empregador', 5:'Cedido por familiar', 6:'Cedido de outra forma', 7:'Outra condição'})
```

Figura 57: Código para plotagem de um gráfico em python

Fonte: Autoria Própria



```
domicilio['V0217'].value_counts().plot(kind='bar')
plt.title("Este domicílio é:")
plt.show()
```

Figura 58: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

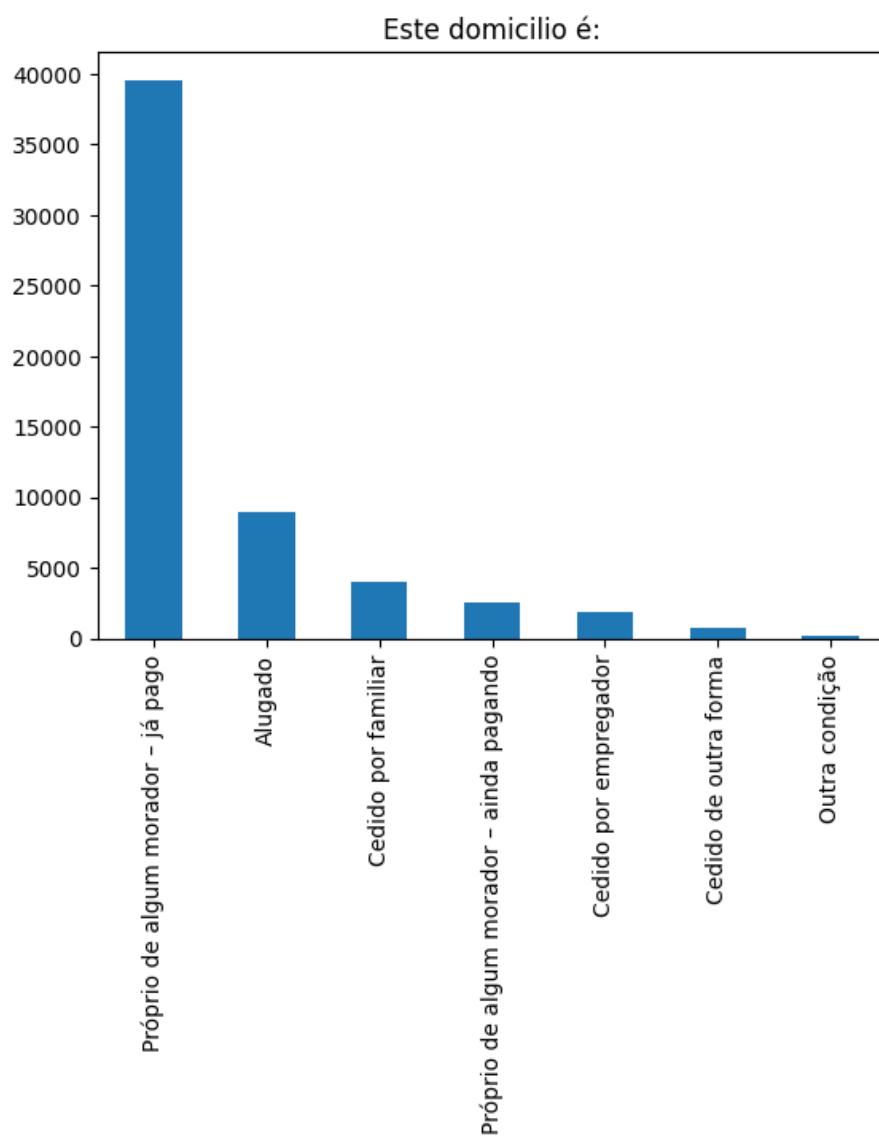


Figura 59: Gráfico “Propriedade do Domicílio” - Domicílio

Fonte: Autoria Própria

Com esse gráfico pode-se concluir que grande parte da população brasileira tem casa própria, e que já está paga. A seguir a coluna "V6199", que indica o nível de segurança alimentar dentro de casa, passa pelo mesmo processo.



```
domicilio['V6199'] = domicilio['V6199'].replace({1: 'Segurança', 2: 'Insegurança leve',
3:'Insegurança moderada', 4:'Insegurança grave'})

domicilio['V6199'].value_counts().plot(kind='pie', autopct='%1.1f%%')
plt.title("Situação de segurança alimentar")
plt.show()
```

Figura 60: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

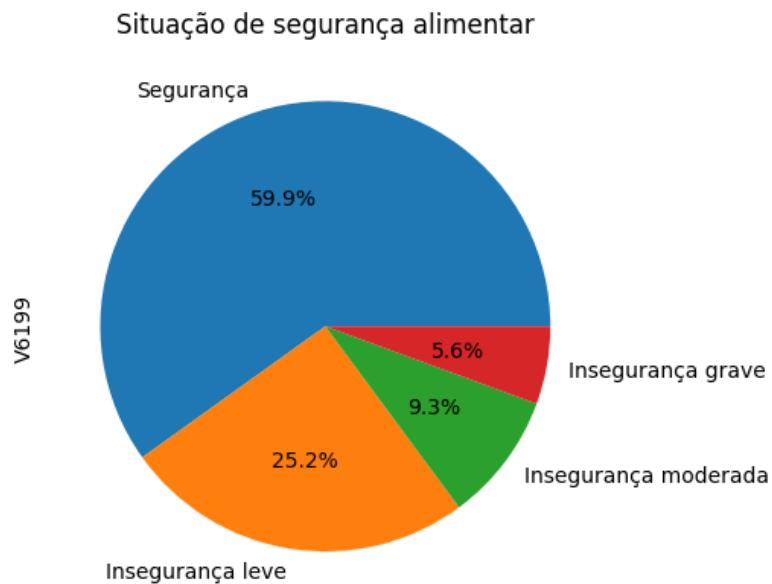


Figura 61: Gráfico “Situação de segurança alimentar” - Domicílio

Fonte: Autoria Própria

Esse gráfico demonstra como as pessoas se sentem seguras em trazer comida dentro de casa, e pode-se observar que mais da metade se sente confortável, esse valor apesar de ser expressivo, ainda não é o ideal, porque os outros 40% ainda sofrem de algum tipo de insegurança. O último processo é aplicar o código `.fillna` em colunas que têm valores nulos, mas estes significam alguma coisa, que pode ser utilizado em um segundo momento da análise. É importante ressaltar que foi feita uma avaliação para entender quais valores poderiam ser substituídos.



```
domicilio['V02101'].fillna(0, inplace=True)
domicilio['V02102'].fillna(0, inplace=True)
domicilio['V02103'].fillna(0, inplace=True)
domicilio['V02104'].fillna(0, inplace=True)
domicilio['V02105'].fillna(0, inplace=True)
domicilio['V02113'].fillna(0, inplace=True)
domicilio['V0212'].fillna(0, inplace=True)
domicilio['V0215'].fillna(0, inplace=True)
domicilio['V0219'].fillna(0, inplace=True)
```

Figura 62: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

7.2.3. Inventário

O dataset do inventário contém informações sobre bens duráveis nos domicílios do Brasil. Este é composto por 16 colunas e mais de 870 mil linhas de dados. Acessando o dicionário de variáveis, foi realizado um .replace em 3 colunas que foram consideradas mais importantes para uma primeira análise, o código a seguir demonstra essa mudança.



```
aluguel_estimado['UF'] = aluguel_estimado['UF'].replace({11: 'Rondônia', 12: 'Acre', 13: 'Amazonas',
14: 'Roraima', 15: 'Pará', 16: 'Amapá', 17:'Tocantins', 21:'Maranhão', 22:'Piauí', 23:'Ceará',
24:'Rio Grande do Norte', 25:'Paraíba', 26:'Pernambuco', 27:'Alagoas', 28:'Sergipe', 29:'Bahia',
31:'Minas Gerais', 32:'Espírito Santo', 33:"Rio de Janeiro", 35:'São Paulo', 41:'Paraná', 42:'Santa
Catarina', 43:'Rio Grande do Sul', 50:'Mato Grosso do Sul', 51:'Mato Grosso', 52: 'Goiás',
53:'Distrito Federal'})
```

Figura 63: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

Com o código acima, foi realizado a substituição dos valores inteiros para object. Abaixo, segue o código e o gráfico desta coluna.



```
aluguel_estimado['UF'].value_counts().plot(kind='bar')
plt.title("Distribuição por UF")
plt.xlabel("UF")
plt.ylabel("Contagem")
plt.show()
```

Figura 64: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

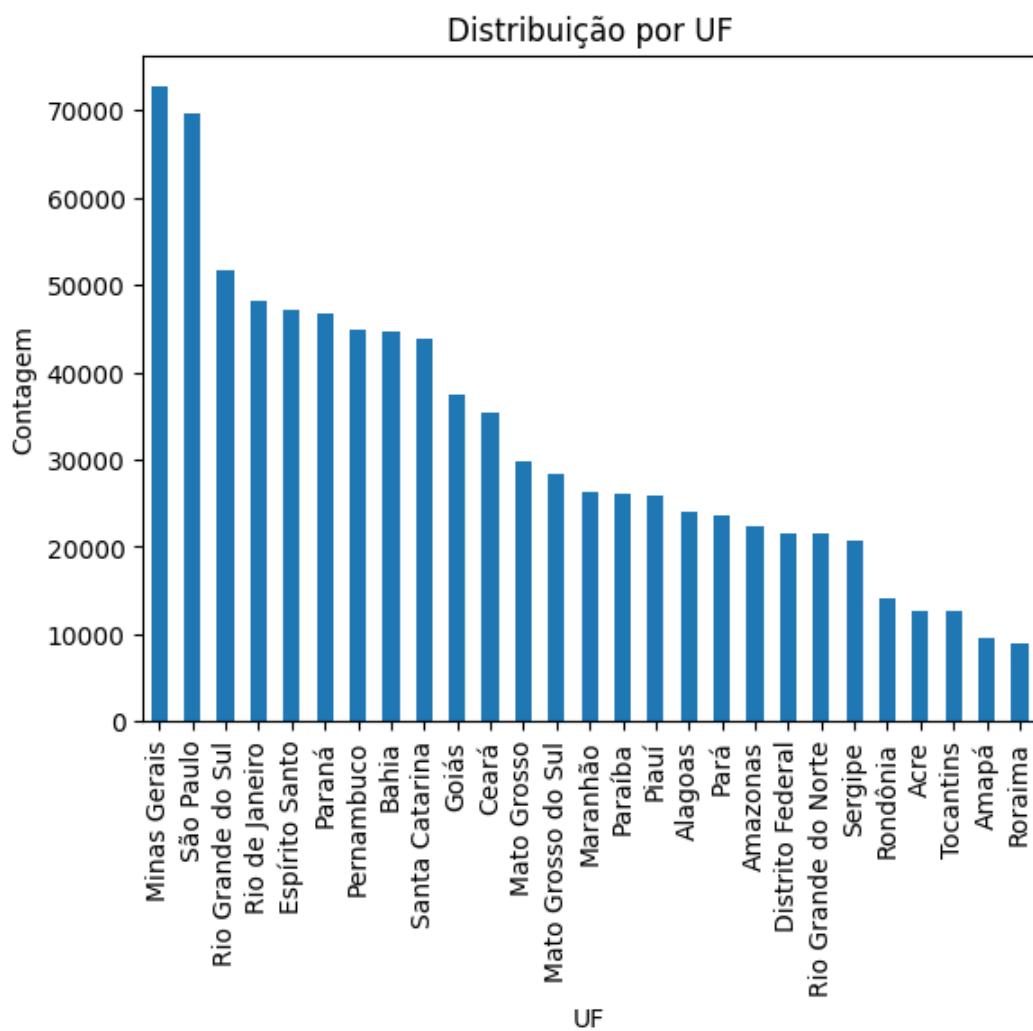


Figura 65: Gráfico “Distribuição por UF” - Inventário

Fonte: Autoria Própria

Com esse gráfico, pode-se classificar que os estados com a maior quantidade de produtos em domicílios são, respectivamente: Minas Gerais, São Paulo e Rio Grande do Sul. O código a seguir foi feito o mesmo processo, só que na coluna "TIPO_SITUACAO_REG", onde mostra se o domicílio se localiza em uma cidade, Urbano, ou em uma área Rural, a seguir é feito o gráfico mostrando essa diferença.



```
aluguel_estimado['TIPO_SITUACAO_REG'] = aluguel_estimado['TIPO_SITUACAO_REG'].replace({1: 'Urbano',  
2: 'Rural'})  
aluguel_estimado['TIPO_SITUACAO_REG'].value_counts().plot(kind='pie', autopct='%1.1f%%')  
plt.title("Tipo de situação regional")  
plt.show()
```

Figura 66: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

Distribuição por Situação Regional

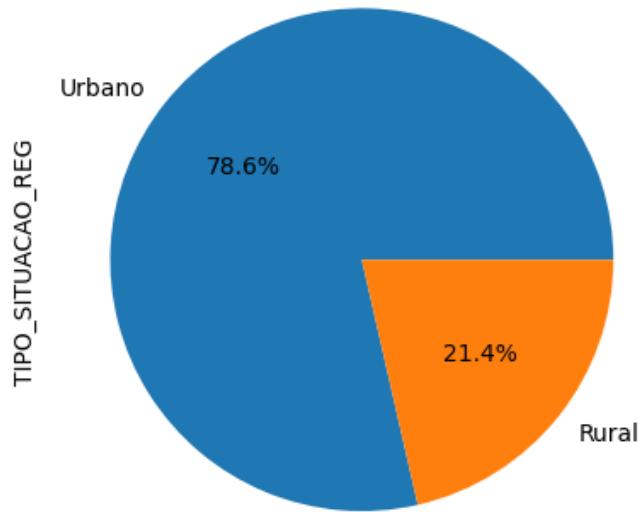


Figura 67: Gráfico “Tipo de situação regional” - Inventário

Fonte: Autoria Própria

A seguir, o código que também segue o mesmo processo, mas com a coluna "V9002", indicando o tipo de domicílio.



```
inventario['V9002'] = inventario['V9002'].replace({1: 'Monetária à vista para a Unidade de Consumo',  
2: 'Monetária à vista para outra Unidade de Consumo', 3:'Monetária a prazo para a Unidade de  
Consumo', 4:'Monetária a prazo para outra Unidade de Consumo', 5:'Cartão de crédito à vista para a  
Unidade de Consumo', 6:'Cartão de crédito à vista para outra Unidade de Consumo', 7:'Doação',  
8:'Retirada do Negócio', 9:'Troca', 10:'Produção Própria', 11:'Outra'})  
domicilio['V0201'].value_counts().plot(kind='pie', autopct='%1.1f%%')  
plt.title("Distribuição por tipo de domicílio")  
plt.show()
```

Figura 68: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

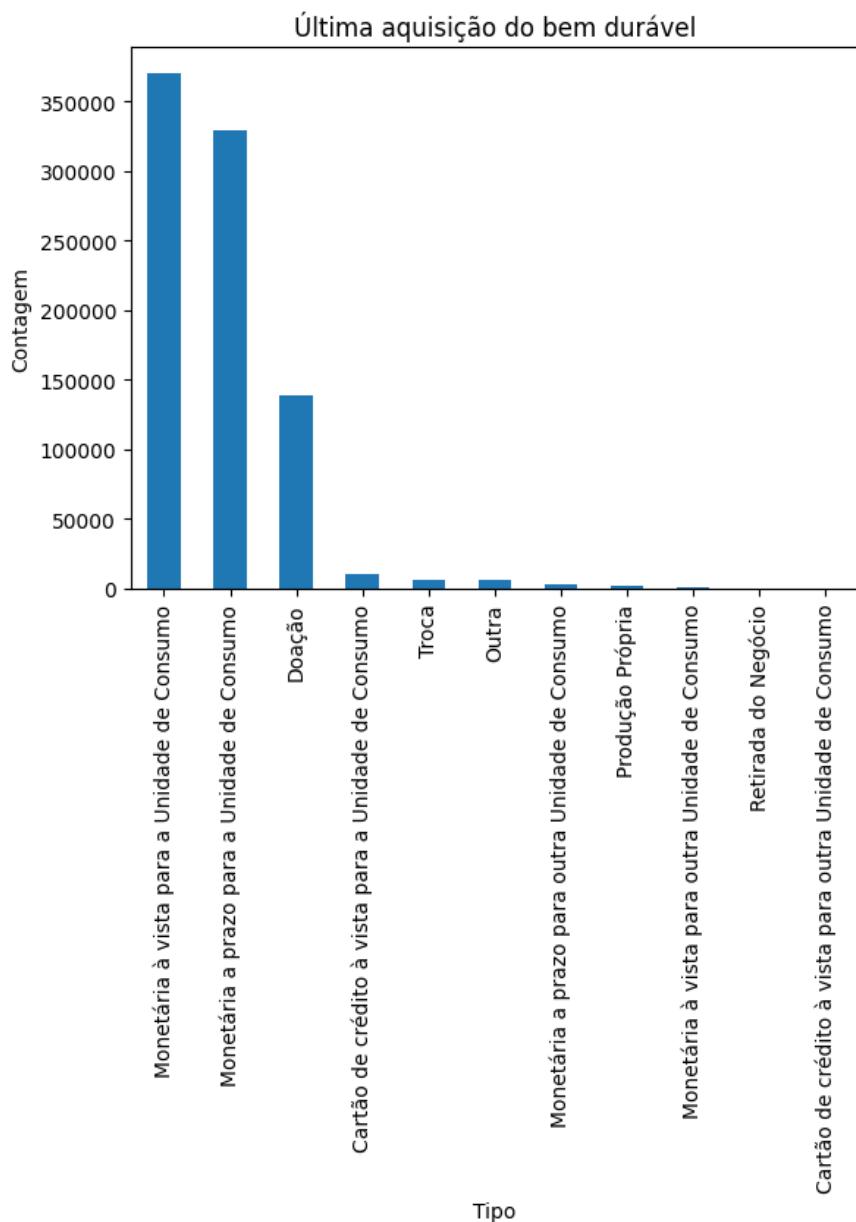


Figura 69: Gráfico “Distribuição por tipo de domicílio” - Domicílio

Fonte: Autoria Própria

7.2.4. Serviço Não Monetário

O dataset Serviço Não Monetário contém informações sobre a renda brasileira, destacando detalhes como a região e distribuição mensal. Acessando o dicionário de variáveis, foi realizado um .replace em 2 colunas que foram consideradas mais importantes para uma primeira análise, o código a seguir demonstra essa mudança:



```
dados['UF'] = dados['UF'].replace({11: 'Rondônia', 12: 'Acre', 13: 'Amazonas', 14: 'Roraima', 15: 'Pará', 16: 'Amapá', 17:'Tocantins', 21:'Maranhão', 22:'Piauí', 23:'Ceará', 24:'Rio Grande do Norte', 25:'Paraíba', 26:'Pernambuco', 27:'Alagoas', 28:'Sergipe', 29:'Bahia', 31:'Minas Gerais', 32:'Espírito Santo', 33:"Rio de Janeiro", 35:'São Paulo', 41:'Paraná', 42:'Santa Catarina', 43:'Rio Grande do Sul', 50:'Mato Grosso do Sul', 51:'Mato Grosso', 52: 'Goiás', 53:'Distrito Federal'})  
dados[ 'TIPO_SITUACAO_REG' ] = dados[ 'TIPO_SITUACAO_REG' ].replace({1: 'Urbano', 2: 'Rural'})
```

Figura 70: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

Aplicação do dicionário da região campo/urbano



```
import pandas as pd
import matplotlib.pyplot as plt
import calendar

# Renomeia as colunas para facilitar o acesso
dados.rename(columns={"V9010": "mes", "RENDAS_TOTAL": "renda"}, inplace=True)

# Agrupa os dados pelo mês e calcula a soma dos gastos em cada mês
gastos_por_mes = dados.groupby("mes")["renda"].sum()

# Converte os números do mês para nomes de mês usando o módulo calendar
meses = [calendar.month_name[int(mes)] for mes in gastos_por_mes.index]

# Lista de valores de gastos para o eixo y do gráfico
gastos = gastos_por_mes.values

# Cria o gráfico de barras
plt.figure(figsize=(10, 6))
plt.bar(meses, gastos, color='forestgreen')
plt.xlabel('Mês')
plt.ylabel('Renda')
plt.title('Renda por Mês')
plt.xticks(rotation=45) # Rotaciona os rótulos do eixo x para melhor visualização
plt.tight_layout() # Ajusta o layout para evitar cortar rótulos
plt.show()
```

Figura 71: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

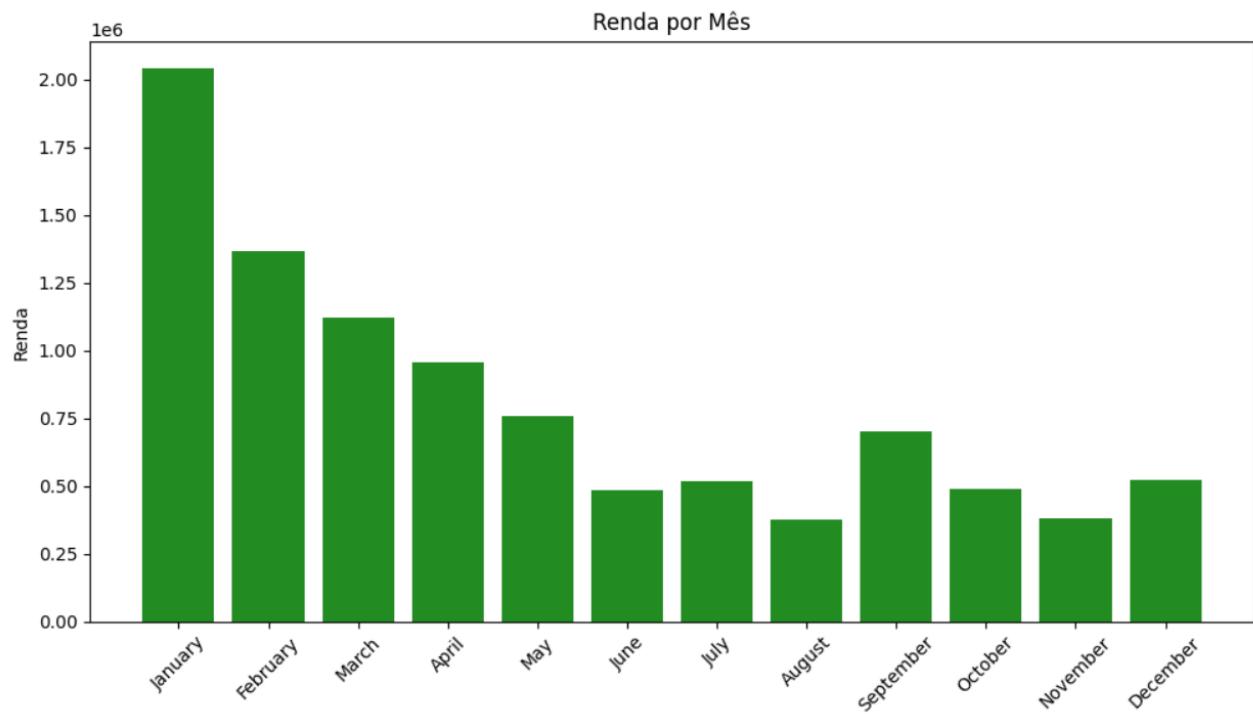


Figura 71: Gráfico “Renda por mês” - Serviço Não Monetário

Fonte: Autoria Própria

O gráfico acima compara a renda total dos brasileiros versus o respectivo mês do ano. O tipo de gráfico de colunas foi escolhido por conta da fácil comparação lado a lado de dados, além de não existir tantas variáveis para essa análise. Fazendo uma análise, é perceptível que a renda brasileira sofre um grande aumento no início do ano, provavelmente por conta do décimo terceiro, depois cai aos poucos e estabiliza.



```
import pandas as pd
import matplotlib.pyplot as plt

# Renomeia as colunas para facilitar o acesso
dados.rename(columns={"UF": "Estado", "RENDAS_TOTAL": "renda"}, inplace=True)

# Agrupa os dados pelo estado e calcula a soma dos gastos em cada estado
gastos_por_estado = dados.groupby("Estado")["renda"].sum()

# Lista de estados para o eixo x do gráfico
estados = gastos_por_estado.index.tolist()

# Lista de valores de gastos para o eixo y do gráfico
gastos = gastos_por_estado.values

# Cria o gráfico de barras
plt.figure(figsize=(10, 6))
plt.bar(estados, gastos, color='darkcyan')
plt.xlabel('Estado')
plt.ylabel('Renda')
plt.title('Renda por Estado')
plt.xticks(rotation=45) # Rotaciona os rótulos do eixo x para melhor visualização
plt.tight_layout() # Ajusta o layout para evitar cortar rótulos
plt.show()
```

Figura 72: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

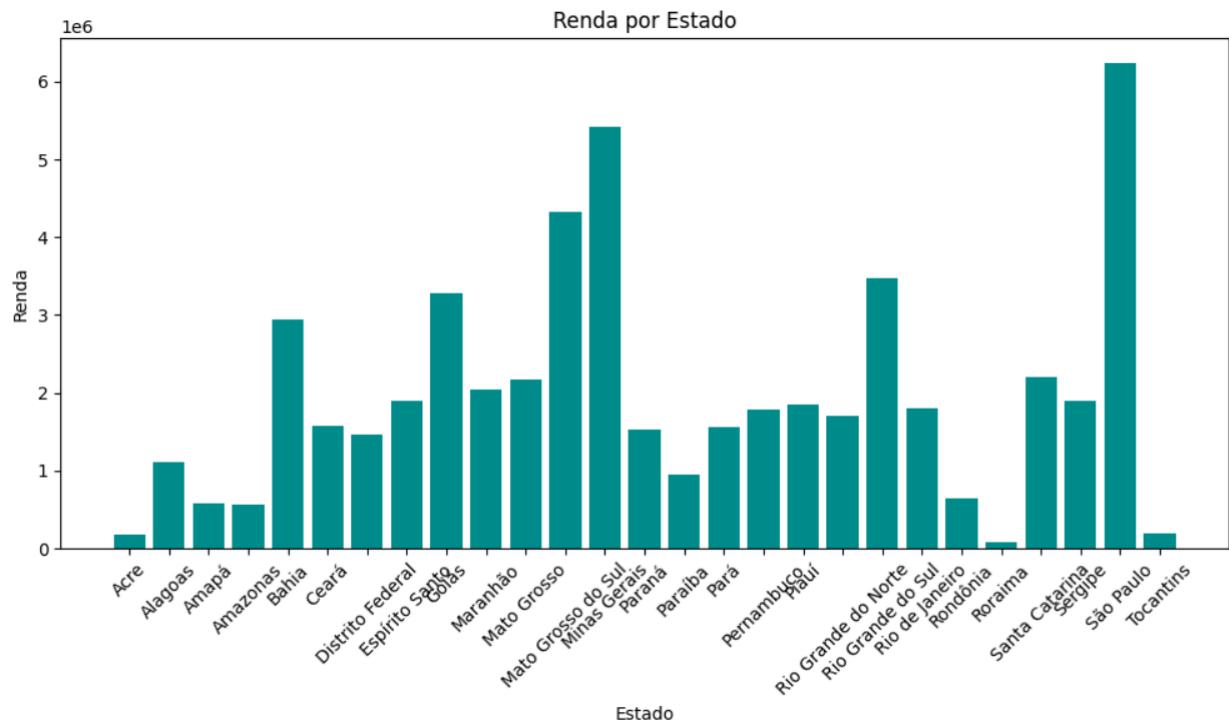


Figura 73: Gráfico “Renda por estado” - Serviço Não Monetário

Fonte: Autoria Própria

O gráfico acima compara a renda total de cada estado brasileiro. O tipo de gráfico de colunas foi escolhido novamente por conta da fácil comparação lado a lado de dados, além de não existirem tantas variáveis para essa análise. Fazendo uma análise, é perceptível que o estado de São Paulo é o mais rico (indústria), acompanhado logo em seguida pelo Mato Grosso do Sul (agropecuária).



```
import pandas as pd
import matplotlib.pyplot as plt
import calendar

# Renomeia as colunas para facilitar o acesso
dados.rename(columns={"V9010": "Mes", "UF": "Estado", "RENTA_TOTAL": "renda"}, inplace=True)

# Remove linhas com valores NaN na coluna de mês
dados = dados.dropna(subset=["mes"])

# Converte os números do mês para nomes de mês usando o módulo calendar
dados["mes"] = dados["mes"].apply(lambda x: calendar.month_name[int(x)])

# Cria o gráfico de barras empilhadas para mostrar o gasto por mês em cada estado
plt.figure(figsize=(12, 6))
dados.pivot_table(index='mes', columns='Estado', values='renda', aggfunc='sum').plot(kind='bar',
stacked=True)
plt.xlabel('Mês')
plt.ylabel('Renda')
plt.title('Renda por Mês e Estado')
plt.xticks(rotation=45)
plt.tight_layout()
plt.legend(title='Estado', bbox_to_anchor=(1.05, 1), loc='upper left') # Move a legenda para fora do
gráfico
plt.show()
```

Figura 74: Código para plotagem de um gráfico em python

Fonte: Autoria Própria

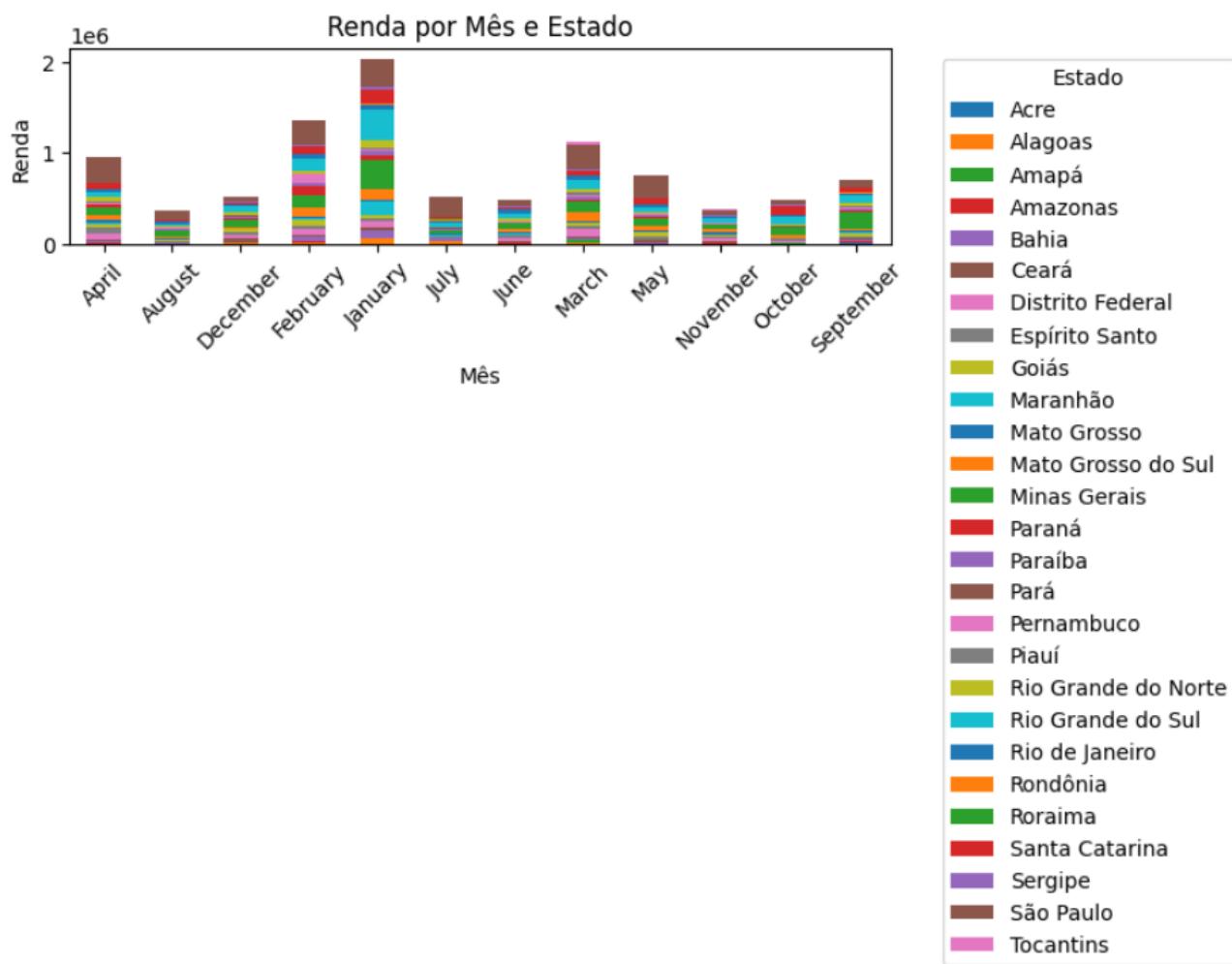


Figura 75: Gráfico “Renda por mês e estado” - Serviço Não Monetário

Fonte: Autoria Própria

O gráfico acima compara a renda total de cada estado brasileiro versus o mês. O tipo de gráfico de colunas empilhadas foi escolhido por conta da fácil comparação lado a lado de dados agregados à sub comparação em cada coluna. Fazendo uma análise, é perceptível que alguns estados mantêm um bom nível de renda durante todo ano (SP, RJ) enquanto outros não (AM).

7.2.5. Serviço Não Monetário

O dataset Restrição Produtos Serviços Saúde contém informações sobre as razões da população brasileira não conseguir comprar determinados produtos. Acessando o dicionário de variáveis, foi realizado um .replace em 3 colunas que foram consideradas mais importantes para uma primeira análise

Distribuição de Proporção para dificuldade em Compra

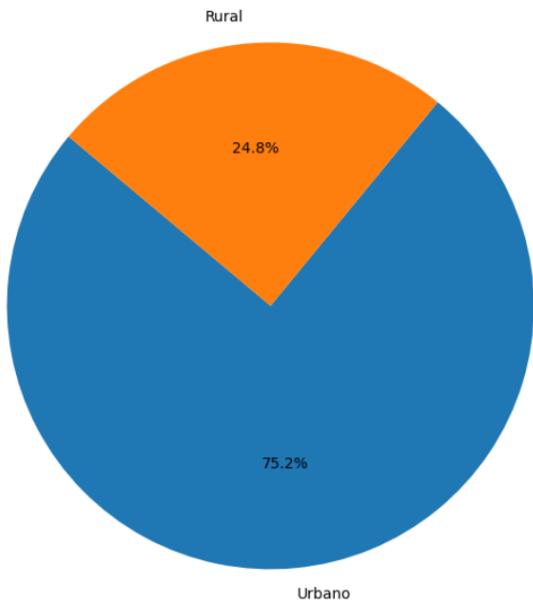


Figura 76: Gráfico “Proporção para a dificuldade de compra” - Serviço Não Monetário

Fonte: Autoria Própria

O gráfico compara a porcentagem da origem local brasileira versus os casos de dificuldade de compra. O tipo de gráfico utilizado é o de pizza por conta do baixo número de conjuntos sendo comparado (apenas 2), além disso, uma distribuição em porcentagem também complementa o gráfico. Já em uma análise, tendo em vista a baixa população rural do Brasil (16% total do país), há uma maior ocorrência proporcional de dificuldades nessa região (25%).

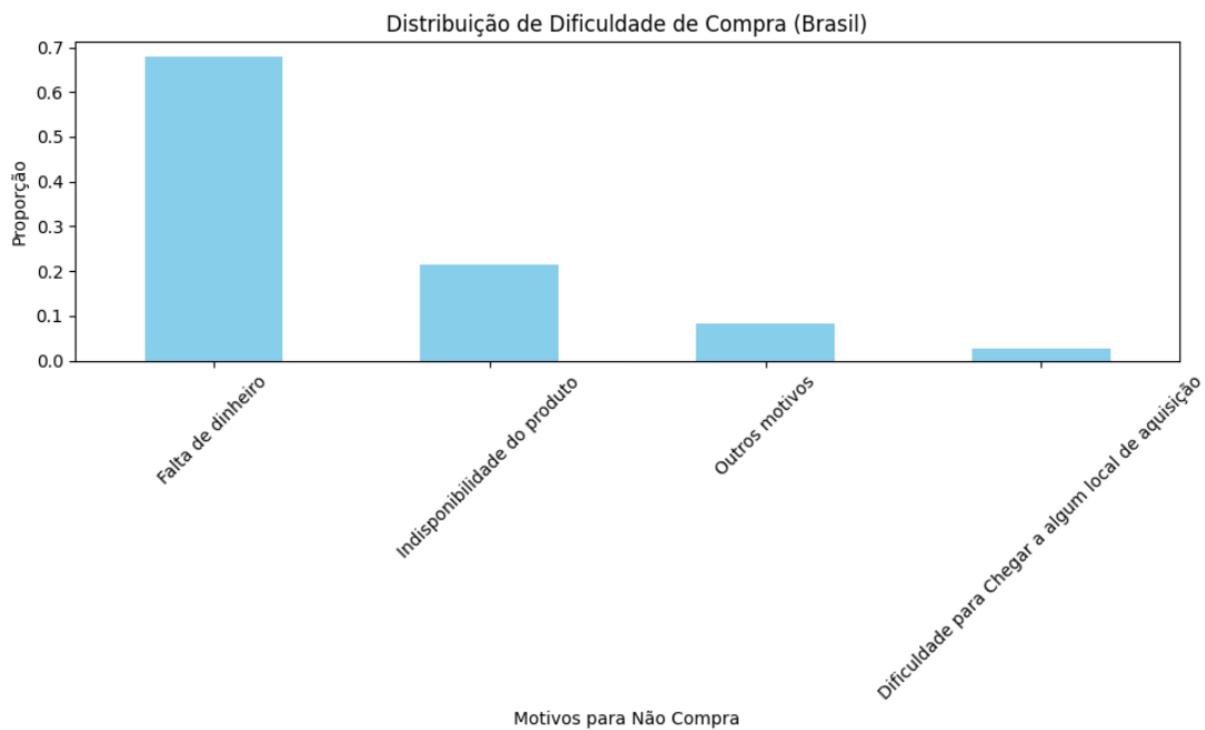


Figura 77: Gráfico “Dificuldade de Compra” - Serviço Não Monetário

Fonte: Autoria Própria

Nesse gráfico é observável que a falta de dinheiro é o principal motivo para o brasileiro não adquirir um item de consumo. O gráfico em barras tem a melhor capacidade de fazer essa comparação lado a lado.

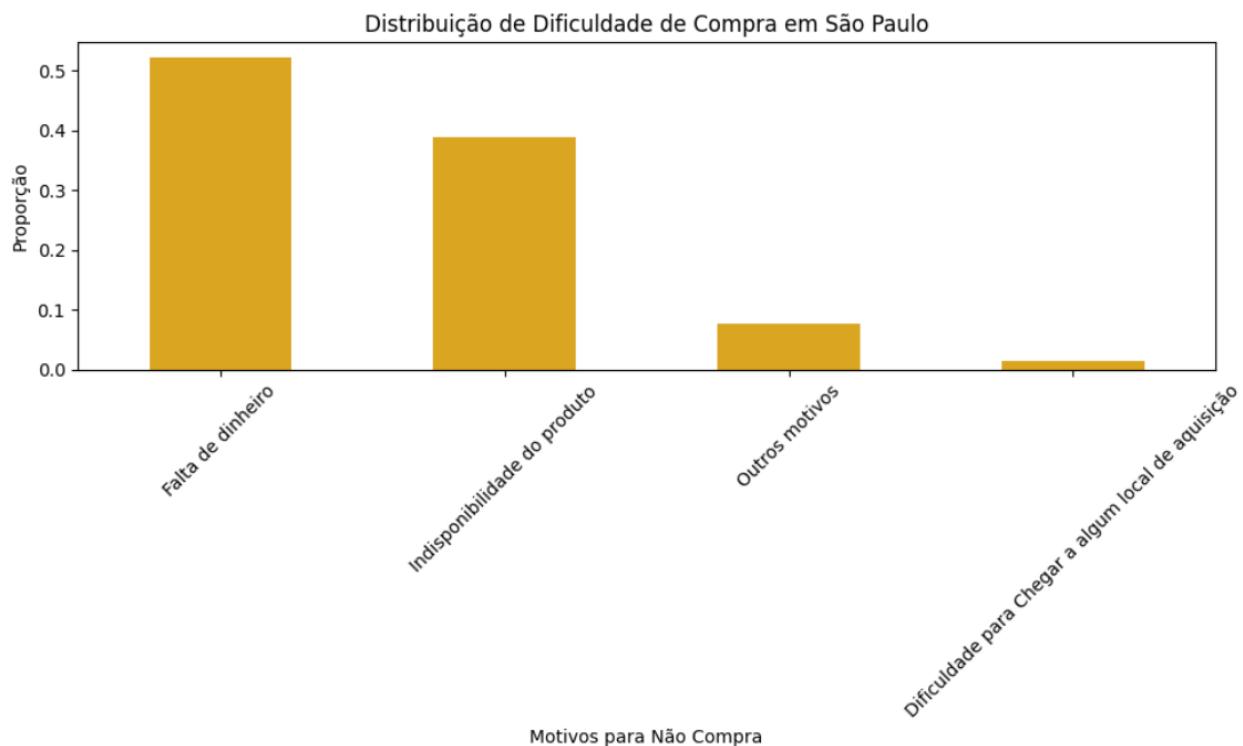


Figura 78: Gráfico “Dificuldade de Compra” - Serviço Não Monetário

Fonte: Autoria Própria

Em São Paulo, o motivo para a não compra de um produto já se altera. Ele também está muito relacionado à indisponibilidade do produto. Novamente, o gráfico de barras é o melhor para essa comparação.

7.2.6. Despesa coletiva

O dataset despesa coletiva contém informações sobre a despesa coletiva referente às famílias brasileiras. Acessando o dicionário de variáveis, foi realizado um .replace em 3 colunas que foram consideradas mais importantes para uma primeira análise.

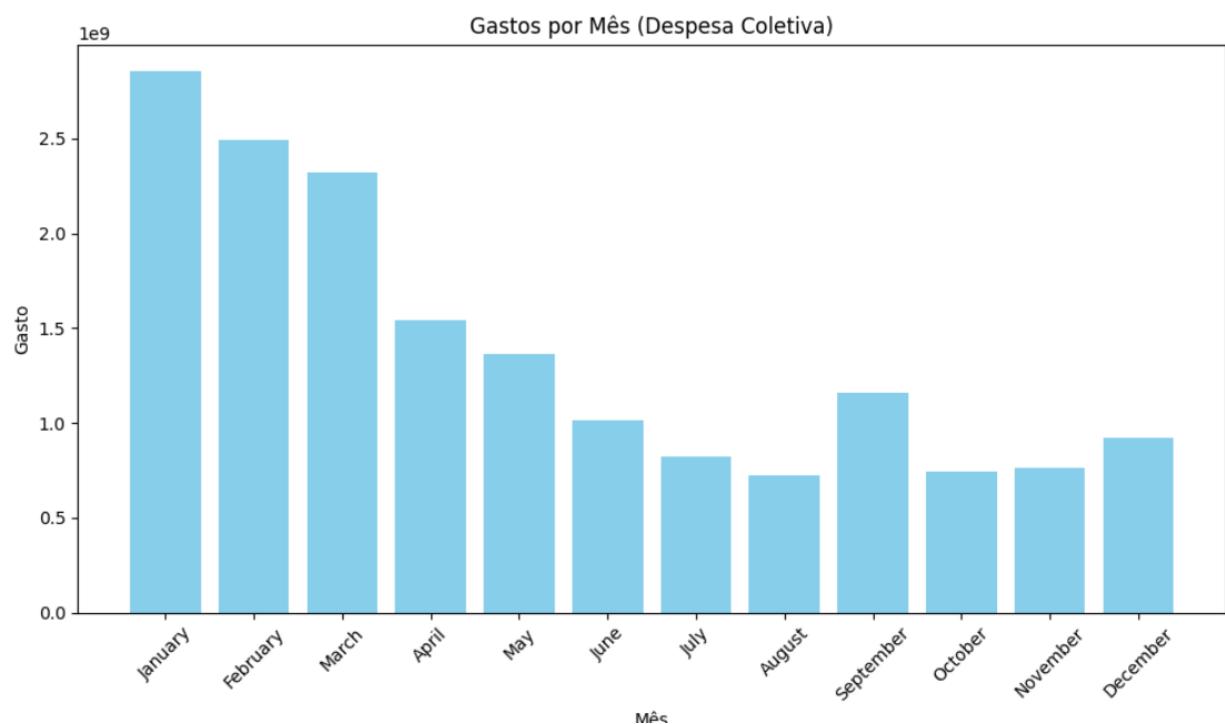


Figura 79: Gráfico “Gastos por mês” - Despesa Coletiva

Fonte: Autoria Própria

Distribuição em gráficos da renda por mês. Comparando com gráficos anteriores de renda, é possível perceber que o gasto é proporcional à renda, reflexo do tipo de consumo brasileiro. O gráfico de barras é o melhor para essa comparação lado a lado

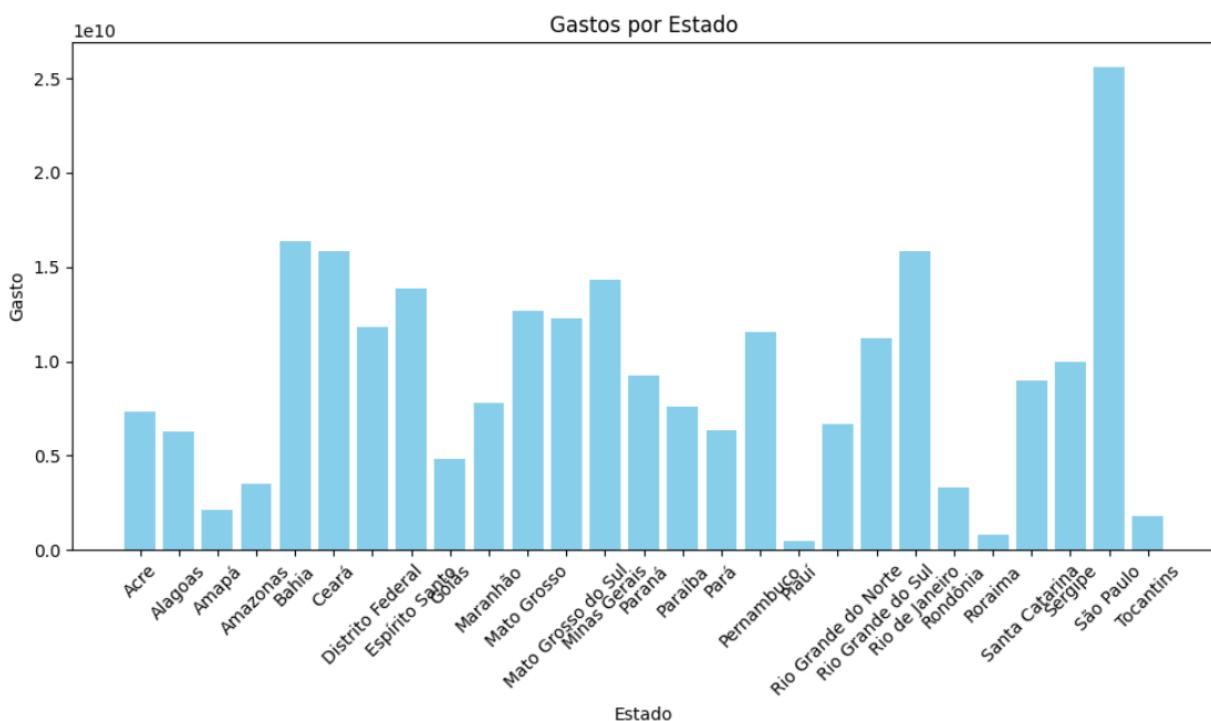


Figura 80: Gráfico “Gastos por mês” - Despesa Coletiva

Fonte: Autoria Própria

Distribuição em gráficos da renda por estado. Comparando com o gráficos, é possível perceber que o estado de São Paulo possui a maior renda mas, olhando agora, o maior custo também. Essa lógica se aplica a todos os outros estados. O gráfico de barras continua sendo o melhor para comparações minuciosas assim.

8. Arquitetura Macro

A arquitetura do sistema se refere às decisões que definem a estrutura e organização dos componentes que constituem a aplicação. Responsável por garantir que a aplicação seja escalável e segura.

8.1 Requisitos do pipeline de dados

Um pipeline de dados é um conjunto de processos e ferramentas que permitem a coleta, processamento, armazenamento e análise de grandes volumes de dados. Apresenta-se abaixo os requisitos estabelecidos para este projeto:

Fontes de Dados: Os dados provêm de três fontes distintas - dados de pesquisas do governo, informações de CNPJs e dados fornecidos pelo parceiro, a *Integration*.

Volume de Dados: O volume de dados varia dependendo das contribuições do parceiro e do crescimento contínuo ao longo dos anos nos dados governamentais e de CNPJs. Em média, se planeja suportar um volume de dados que varia de 6 a 10 gigabytes nesta aplicação.

Velocidade de Ingestão: Os dados são transmitidos via streaming, onde os fluxos são processados durante a visualização das informações do infográfico. A arquitetura foi projetada com serviços que garantem esse processamento ágil.

Transformação e Processamento: Os dados chegam em um formato não estruturado e, durante o processamento, eles são transformados em tabelas estruturadas. Além disso, aplica-se procedimentos de limpeza, verificação de integridade e remoção de dados indesejáveis, incluindo aqueles de origem estrangeira.

Armazenamento: O armazenamento dos dados ocorre em um banco de dados relacional, hospedado na AWS Cloud. No entanto, a arquitetura foi concebida para permitir a portabilidade para outras plataformas de nuvem, fazendo amplo uso de serviços de código aberto.

Segurança: A segurança é mantida por meio de autenticação, com dois níveis distintos. O primeiro nível é para acesso às informações do infográfico, enquanto o segundo abrange toda a parte técnica dos dados, incluindo ingestão, armazenamento e análise estatística. Para implementação utiliza-se o AWS IAM, o que proporciona a flexibilidade de migrar esse serviço para outras plataformas de nuvem, caso necessário.

Escalabilidade: Com foco na escalabilidade e na gestão da demanda, após o processamento, os dados são armazenados em um banco de dados OLAP. Mesmo com grandes volumes de dados, se consegue gerenciar as requisições, pela arquitetura ser modular.

8.2. Identificação dos dados de entrada e saída

A identificação dos dados de entrada e saída é utilizada para esclarecer como os dados fluem através do sistema.

8.2.1 Dados de Entrada

Fontes de Dados:

- Dados de pesquisas do governo;
- Dados de CNPJs;
- Dados fornecidos pelo parceiro.

Formato dos Dados de Entrada: Os dados de entrada são, em sua maioria, não estruturados e podem incluir textos, números e informações variadas.

Método de Ingestão: Os dados de entrada são transmitidos em tempo real por meio de um sistema de ingestão de dados via batch.

8.2.2 Dados de Saída

Infográfico: A principal saída da aplicação é a apresentação de um infográfico, que oferece insights visuais com base nos dados processados.

Formato dos Dados de Saída: Os dados de saída serão apresentados em um formato visual, como gráficos.

Destino dos Dados de Saída: Os dados processados e transformados são exibidos ao usuário final por meio de uma interface.

8.3. Análise das necessidades e objetivos do pipeline

A análise das necessidades e objetivos do pipeline auxilia na definição das diretrizes e no planejamento do sistema. Esta seção detalha as necessidades e metas do pipeline de dados:

8.3.1 Necessidades

Coleta de Dados: O pipeline deve ser capaz de coletar dados de fontes diversas, como dados governamentais, CNPJs e dados do parceiro, de forma confiável, garantindo a integridade e qualidade dos dados.

Processamento: Dada a geração do infográfico, é essencial processar dados em tempo real para fornecer insights atualizados aos usuários.

Transformação e Limpeza de Dados: É necessário aplicar transformações e limpeza aos dados não estruturados, incluindo a estruturação em tabelas e a remoção de dados indesejáveis.

Armazenamento: Os dados processados devem ser armazenados de forma segura em um banco de dados relacional hospedado na AWS Cloud, garantindo que estejam disponíveis quando necessário.

Portabilidade e Flexibilidade: A arquitetura deve ser projetada para permitir a portabilidade para outras nuvens, fazendo uso de serviços de código aberto, caso haja necessidade de migração futura.

Segurança em Duas Camadas: Implementação de dois níveis de segurança, com autenticação para acesso às informações do infográfico e autenticação separada para as operações técnicas do pipeline, garantindo a segurança contra acessos não autorizados.

Escalabilidade: A arquitetura deve ser escalável para acomodar volumes de dados crescentes e manter o desempenho, mesmo com um grande volume de informações.

8.3.2 Objetivos

Infográfico: O principal objetivo é fornecer um infográfico interativo que apresenta informações de forma visual e acessível aos usuários.

Tomada de Decisão: Auxiliar na tomada de decisões com base nas informações apresentadas no infográfico.

Migração Simples: Permitir a migração dos serviços para outras plataformas de nuvem, a fim de garantir a continuidade dos negócios em outros ambientes.

Gerenciamento de Dados: Gerenciar os dados de forma a armazenar em um banco de dados OLAP, além de usar containers Docker para manutenção modularizada.

8.4. Escolha de serviços adequados para cada etapa do pipeline

8.4.1 Fonte de Dados

Dados Governamentais: Dados em formato csv de fontes governamentais e sites para pegar novos dados quando tiver atualizações ou para consultas futuras.

Dados do CNPJ: Dados de empresas em formato csv, incluindo informações sobre CNPJ, setor e localização.

Dados do Parceiro: Informações de parceiros externos através da API, requisições GET.

8.4.2 Automação de Ingestão

Scripts python: *scripts python* que recebem os dados em formato csv, realizam o pré-processamento desses dados e automatizam o seu envio para *buckets* do Amazon S3.

API com Lambda: *script python* que se conecta com a API do parceiro com um intervalo de uma semana. Este, coleta os dados fornecidos por meio de um lambda e realiza o envio para um bucket do Amazon S3, que será excluído mediante interesse das partes envolvidas.

8.4.3 Preparação e Armazenamento:

AWS S3 (Data Lake): Serviço de armazenamento escalável da AWS para dados brutos antes do processamento.

Redshift (Data warehouse): Serviço de armazenamento escalável da AWS para dados previamente tratados e prontos para a utilização.

8.4.4 Análise estatística

Modelo ensemble: modelo ensemble utilizando *python* para verificação da existência de padrões entre os dados utilizados.

8.4.5 Infográfico

Metabase: ferramenta de *Business Intelligence* que simplifica a análise de dados e a tomada de decisões por meio da criação de infográficos e relatórios personalizados.

8.4.6 Segurança

AWS IAM (Identity and Access Management): Gerencia a autenticação de usuários, controlando o acesso aos recursos da AWS, garantindo a segurança dos dados e recursos.

8.5. Escolha dos serviços

A escolha dos serviços nesta arquitetura de Big Data foi planejada considerando principalmente a portabilidade, a ênfase na AWS como principal nuvem, a flexibilidade de custos e a otimização de recursos. Abaixo apresenta separado em tópicos a justificativa para essa seleção:

- **Portabilidade:** A maior parte dos serviços foi escolhida com a portabilidade em mente, evitando depender de soluções proprietárias. Isso permite que a arquitetura seja facilmente migrada para outras plataformas de nuvem, se necessário.
- **Uso de Open Source:** A preferência por tecnologias de código aberto proporciona flexibilidade e custos potencialmente mais baixos.
- **AWS Cloud:** A AWS foi escolhida como a principal nuvem para atender ao escopo inicial do projeto. No entanto, a arquitetura foi planejada de forma a ser portável, o que significa que, se o cliente decidir migrar para outra nuvem no futuro, a transição será suave e eficiente, minimizando a interrupção dos serviços.
- **Flexibilidade de Custos:** Dado que o cliente não estabeleceu um orçamento específico para a arquitetura, a escolha de serviços também considerou a

otimização de custos. A seleção de ferramentas de código aberto e a capacidade de dimensionar recursos conforme necessário permitem que o cliente controle e otimize os custos à medida que o projeto evolui.

8.6. Representação visual do pipeline

A arquitetura proposta atende às necessidades da Integration, fornecendo um sistema que lida com a aquisição e análise de dados de consumo de produtos alimentícios, resultando na criação de infográficos informativos. Esta arquitetura é projetada para abranger todo o processo, desde a coleta de dados a partir de diversas fontes até a entrega de infográficos prontos para análise.

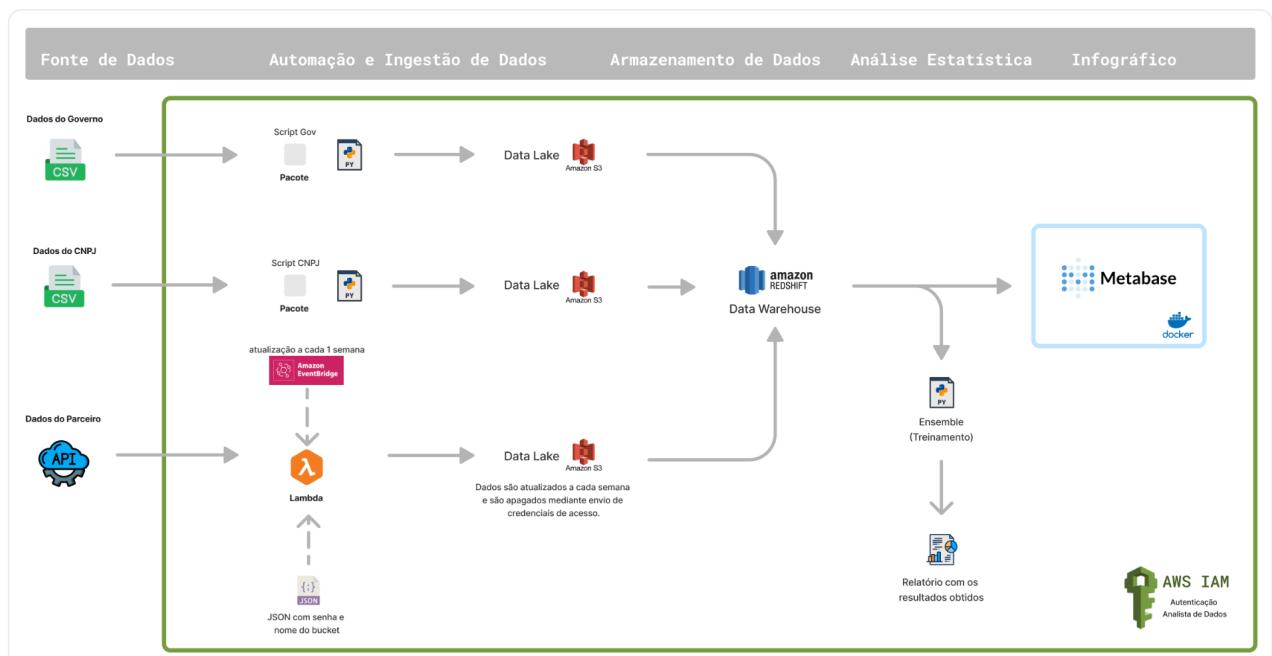


Figura 81: Arquitetura
Fonte: Criação própria ([Arquitetura](#))

A seguir, apresenta-se uma comparação dos serviços utilizados na arquitetura que atualmente fazem uso da infraestrutura na nuvem da AWS, juntamente com suas correspondentes alternativas na nuvem da Azure. Essa comparação assume que o cliente planeja realizar a migração para a plataforma Azure em um momento futuro, exigindo uma compreensão das alternativas disponíveis para uma transição suave.



Figura 82: Comparação AWS x Azure

Fonte: Criação própria

8.7. Boas práticas - resiliência e escalabilidade

Garantir a resiliência e escalabilidade é fundamental em qualquer arquitetura de *Big Data* para lidar com o crescimento de dados e as demandas variáveis. Apresenta-se abaixo as medidas tomadas nesse projeto:

- **Arquitetura Modularizada:** A arquitetura foi projetada de forma modular, com componentes independentes que podem ser escalados e mantidos separadamente. Isso permite que recursos sejam alocados onde mais são necessários, sem impactar o funcionamento de todo o sistema.
- **Serviços de Nuvem:** O uso de serviços em nuvem, como AWS S3 e AWS EC2, proporciona uma escalabilidade sob demanda, permitindo expandir ou reduzir recursos conforme necessário. Além disso, essas plataformas oferecem alta disponibilidade e redundância, contribuindo para a resiliência.
- **Recuperação de falhas:** Planejar estratégias de recuperação de falhas, como redundância de dados e *backups* regulares, para garantir a recuperação eficaz em caso de problemas.

8.8. Serviços ou recursos da AWS - resiliência e escalabilidade

Ao projetar esta arquitetura de *Big Data*, foram incorporados serviços e recursos da Amazon Web Services (AWS) que contribuem para garantir a resiliência e escalabilidade do pipeline de dados:

- **Amazon S3:** Serviço de armazenamento escalável, fornece redundância de dados e replicação entre várias zonas de disponibilidade.
- **AWS Lambda:** Permite a execução de código em resposta a eventos. Ele pode ser usado para lidar com tarefas de processamento de dados, com base na demanda.
- **AWS Identity and Access Management (IAM):** Utilizado para a segurança e a resiliência da arquitetura, permitindo o gerenciamento de permissões e a autenticação dos usuários, garantindo o acesso controlado aos recursos da AWS.

8.9. Calculadora financeira

Para garantir a transparência e o controle dos custos na utilização de serviços em nuvem, é recomendável o uso de calculadoras financeiras. Essas ferramentas ajudam a estimar e gerenciar os gastos com a infraestrutura em nuvem. Aqui estão algumas considerações sobre o uso dessas calculadoras financeiras:

- **AWS Simple Monthly Calculator:** A AWS oferece o "Simple Monthly Calculator", uma ferramenta online que permite estimar os custos mensais com base nos serviços e recursos selecionados.
- **Azure Pricing Calculator:** A Microsoft Azure disponibiliza a "Azure Pricing Calculator", que permite estimar os custos mensais na plataforma Azure. A calculadora oferece uma visão geral dos preços dos serviços, permitindo configurar cenários específicos e avaliar os custos associados a máquinas virtuais, armazenamento, bancos de dados e outros recursos.
- **Comparação de Custos:** Além de calcular os custos em cada plataforma individualmente, é recomendável usar ferramentas de comparação de custos, como o "AWS Total Cost of Ownership (TCO) Calculator" e o "Azure TCO Calculator".

9. Interface

9.1 WireFrame

9.1.1 Desktop

A tela de login foi projetada para proporcionar uma experiência intuitiva, facilitando a conexão com contas já existentes no sistema. Isso garante uma entrada fácil para os usuários, permitindo que insiram suas credenciais nos espaços apropriados para acessar o sistema.

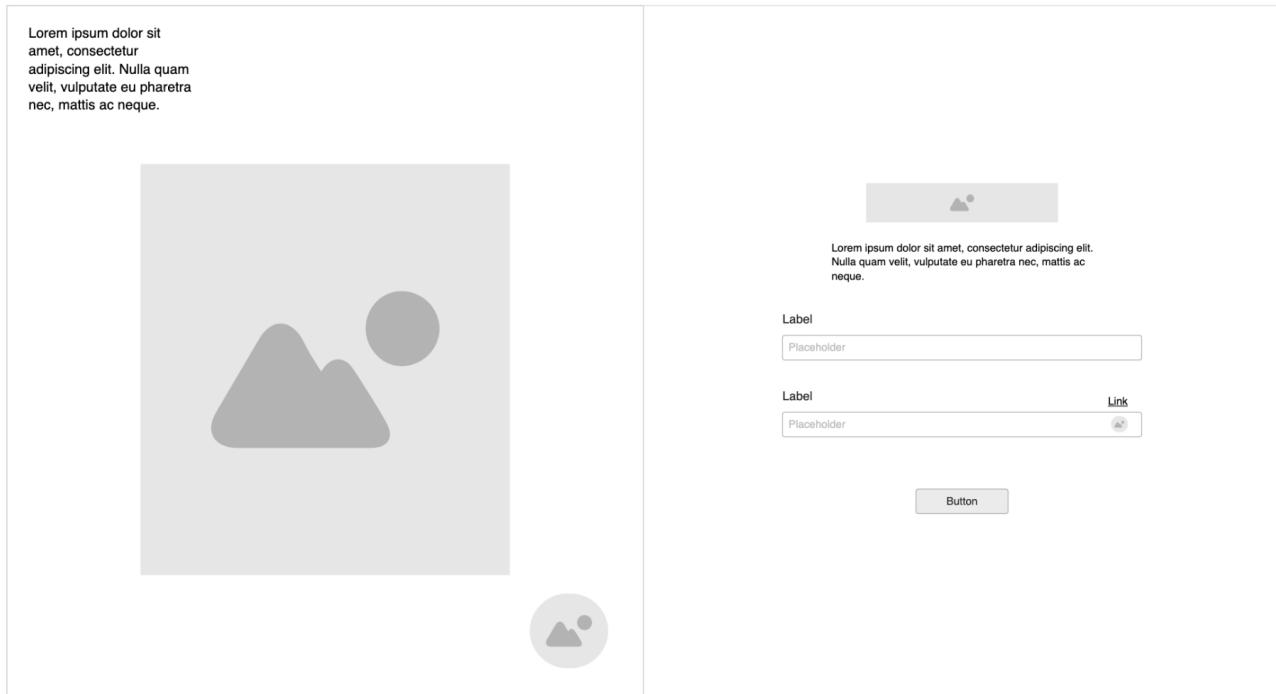


Figura 83: Login Desktop - WireFrame

Fonte: Criação própria

Na tela inicial (*home*), um *sidebar* foi incorporado para simplificar a navegação na página. Esse recurso permitirá a filtragem de informações e servirá como um *hub* para acessar o menu de opções de visualização, filtros e a opção de logout. A tela principal também apresentará cinco gráficos distintos, cada um destacando as tendências de consumo para avaliar o desempenho dos produtos no mercado. A disposição da tela foi estrategicamente segmentada, proporcionando diferentes áreas para títulos específicos de gráficos, seguidos pelos gráficos em si, acompanhados de legendas. As ideias de gráficos incluem o acompanhamento das tendências de vendas de diversos produtos no setor alimentício ao longo do tempo, a quantidade de consumo por regiões, a distribuição de domicílios por região e a tendência de consumo de produtos ao longo do tempo em diferentes categorias.

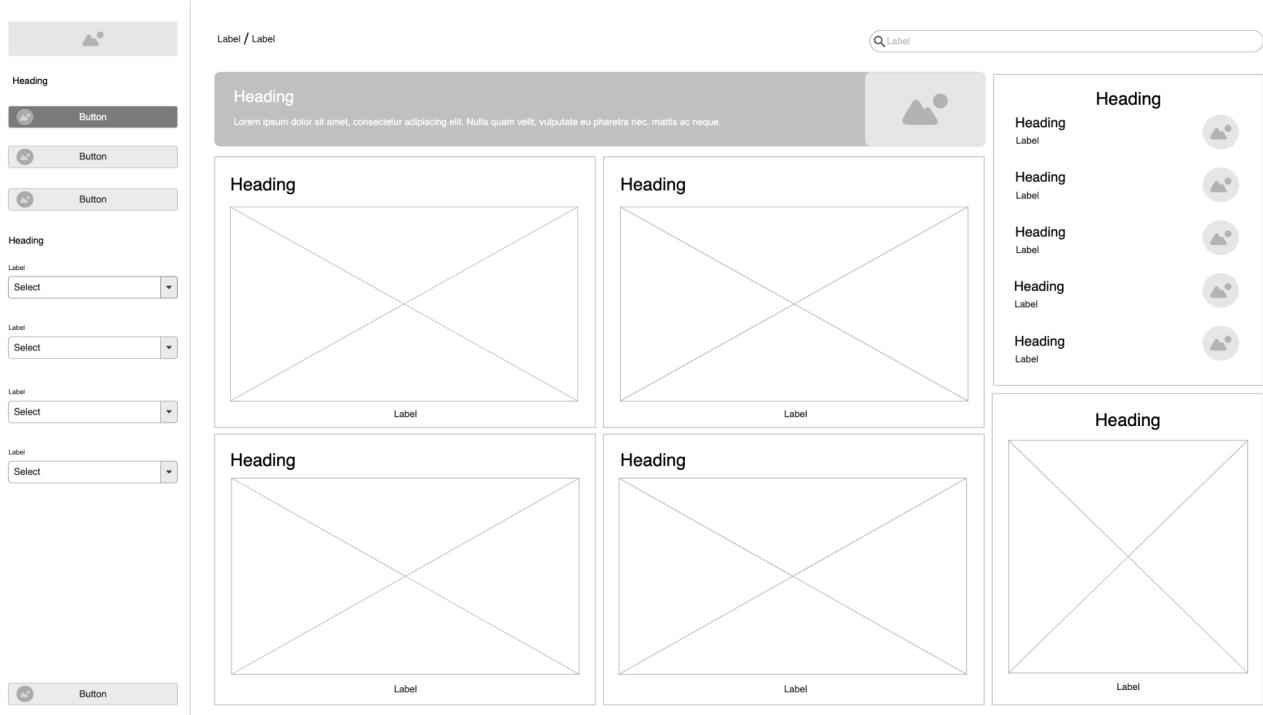


Figura 84: Home Desktop - WireFrame

Fonte: Criação própria

A tela de análise será semelhante à anterior, mantendo a funcionalidade do sidebar. Ela exibirá informações por meio de gráficos e listas, utilizando o modelo mais adequado para apresentar os dados de maneira clara e eficaz.

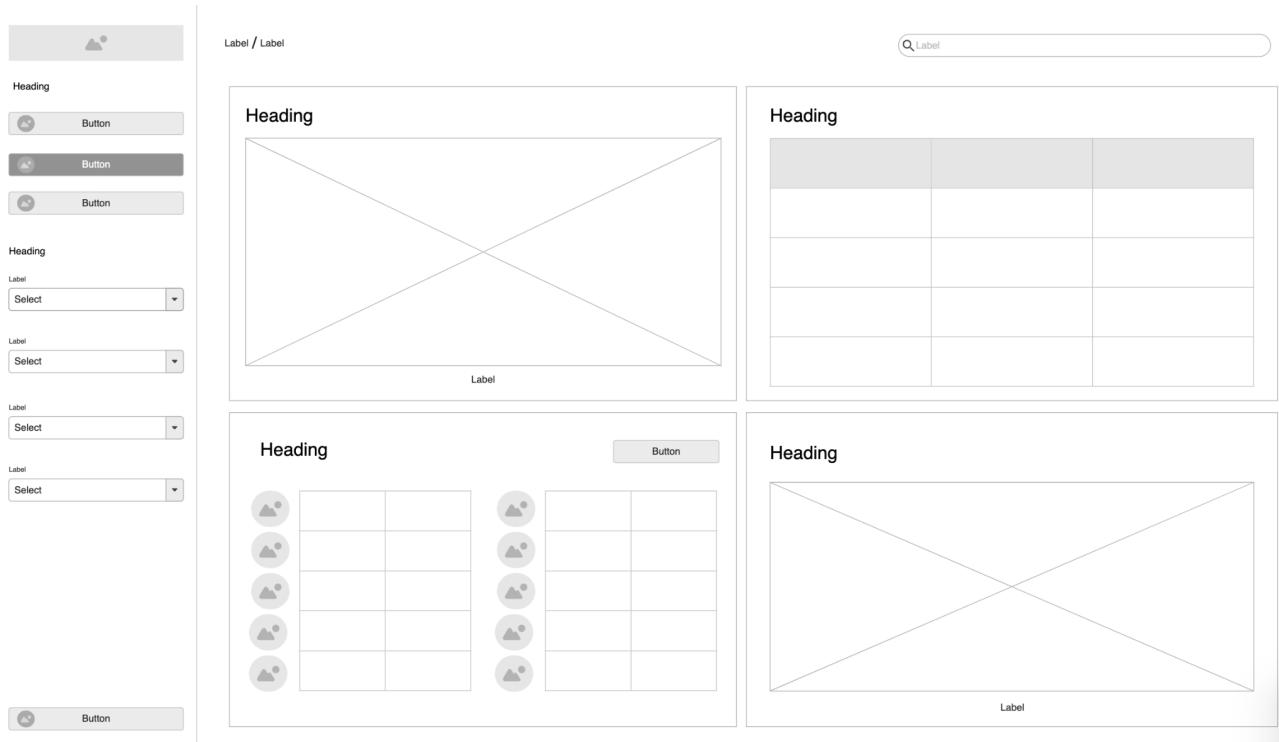


Figura 85: Análise Desktop - WireFrame

Fonte: Criação própria

Na aba de infográfico, será incorporada a opção de busca para facilitar a localização de informações específicas. Além disso, o display de informações incluirá títulos, subtítulos e representações visuais, proporcionando uma experiência de análise detalhada e eficiente. Essa abordagem visa atender às necessidades específicas de análise, garantindo uma interface de usuário intuitiva e eficaz.

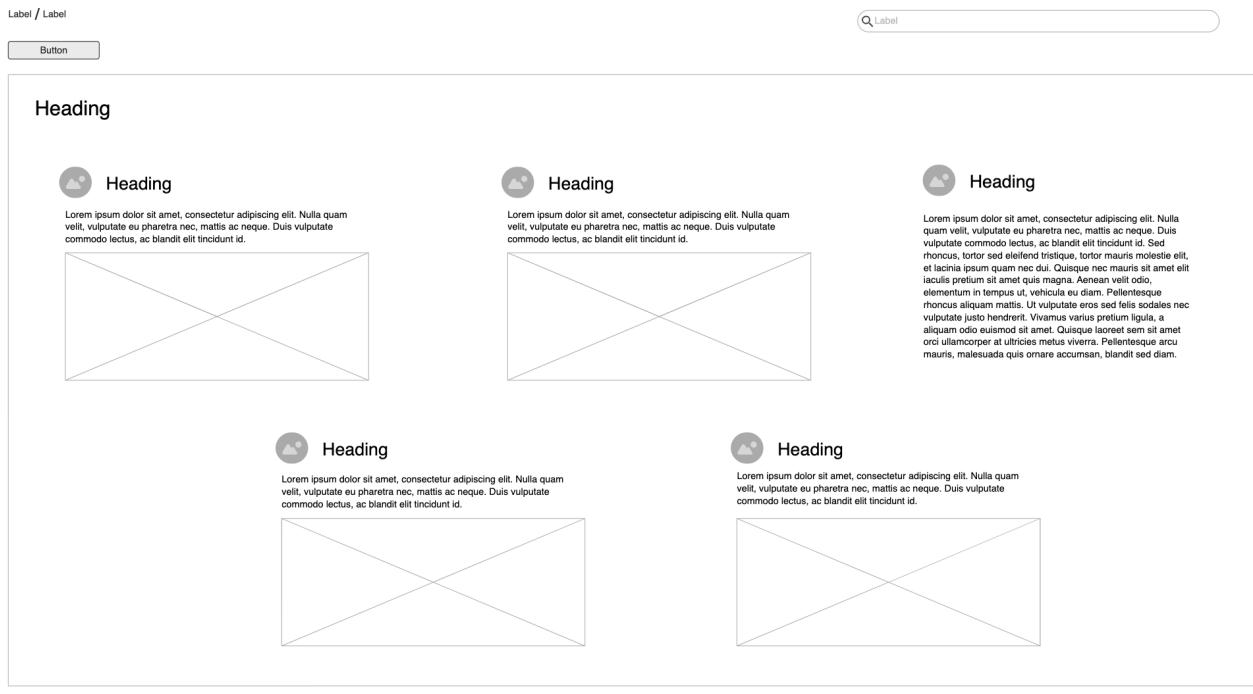


Figura 86: Infográfico Desktop - WireFrame

Fonte: Criação própria

9.1.2 Mobile

Na versão *mobile*, a tela inicial foi projetada com foco na intuição, apresentando uma interface simplificada para proporcionar uma experiência de usuário eficiente. A tela de *login*, semelhante à versão *desktop*, foi otimizada para uma interação intuitiva, facilitando a conexão com contas existentes. Isso assegura uma entrada fácil, permitindo que os usuários insiram suas credenciais nos campos apropriados para acessar o sistema.

Lorem ipsum dolor sit amet,
consectetur adipiscing elit.
Nulla quam velit, vulputate
eu pharetra nec, mattis ac
neque.

Figura 87: Login Mobile 1 - WireFrame

Fonte: Criação própria

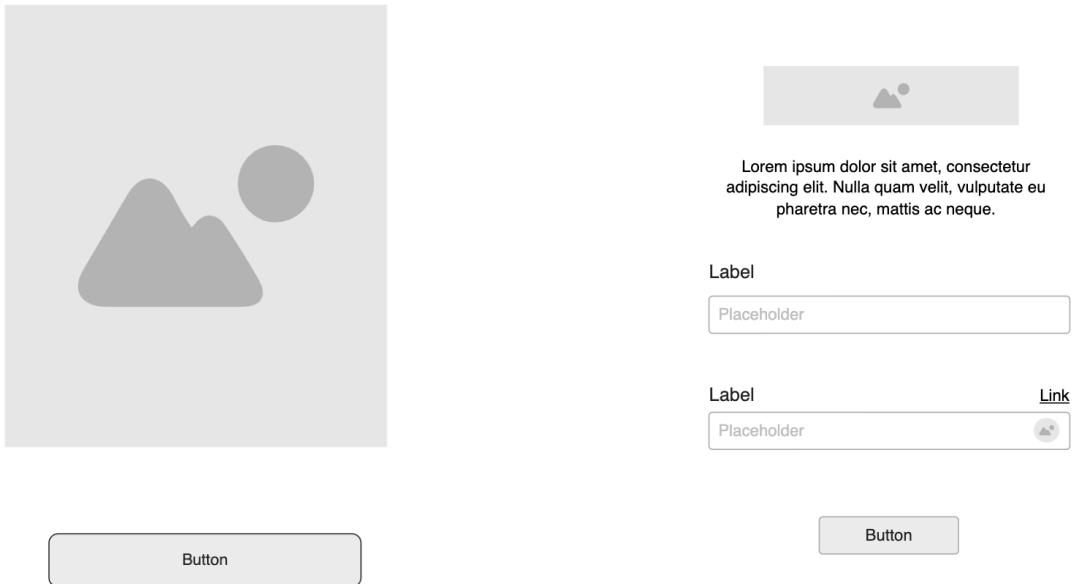


Figura 88: Login Mobile 2 - WireFrame

Fonte: Criação própria



A tela de menu *home*, inspirada no *sidebar* das versões *desktop*, facilita a exploração da interface, garantindo maior acessibilidade e compreensão. Com recursos de busca, seleção e filtro, a disposição da tela adota o método de rolagem (*scroll*) para garantir uma visualização completa.

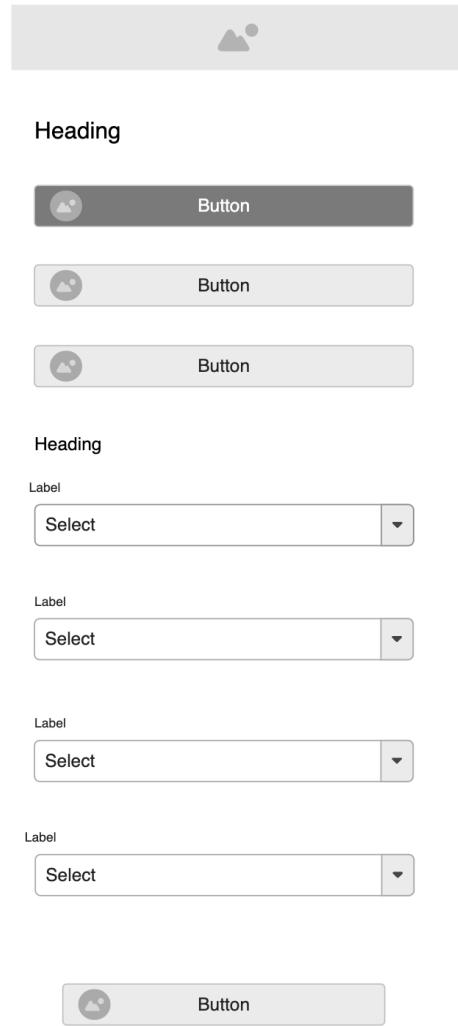


Figura 89: Sidebar Mobile - WireFrame

Fonte: Criação própria

Na aba *home*, os elementos são estrategicamente segmentados, oferecendo áreas distintas para títulos específicos de gráficos, seguidos pelos próprios gráficos e legendas. A variedade de formatos e cores otimiza a visualização, tornando a interpretação dos dados mais eficiente. As ideias de gráficos, como o acompanhamento das tendências de vendas de produtos alimentícios ao longo do tempo e a quantidade de consumo por regiões, são apresentadas de forma amigável ao usuário, com a possibilidade de rolagem para visualização completa das informações.

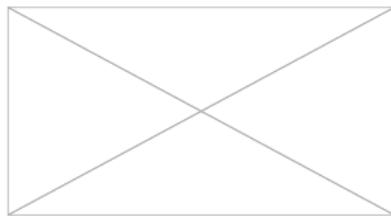
Label / Label Label

Heading

 Lorem ipsum dolor sit amet, consectetur adipiscing elit.
 Nulla quam velit, vulputate eu pharetra nec, mattis ac

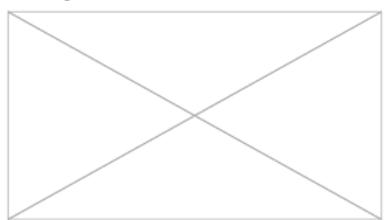


Heading



Label

Heading



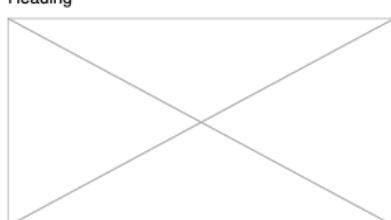
Label

Heading



Label

Heading



Label

Heading

Heading
Label



Heading
Label



Heading
Label



Heading
Label



Heading
Label



Figura 90: Dashboard Mobile - WireFrame

Fonte: Criação própria

A tela de análise, semelhante ao *dashboard*, exibirá informações através de gráficos e listas, adaptando-se ao modelo mais adequado para apresentar dados de maneira clara e eficaz.



Figura 91: Análise Mobile - WireFrame

Fonte: Criação própria

Na aba de infográficos, a opção de busca facilita a localização de informações específicas. O *display* de informações incluirá títulos, subtítulos e representações visuais, proporcionando uma experiência de análise detalhada e eficiente. Essa abordagem visa atender às necessidades específicas de análise, garantindo uma interface de usuário intuitiva e eficaz em dispositivos móveis.

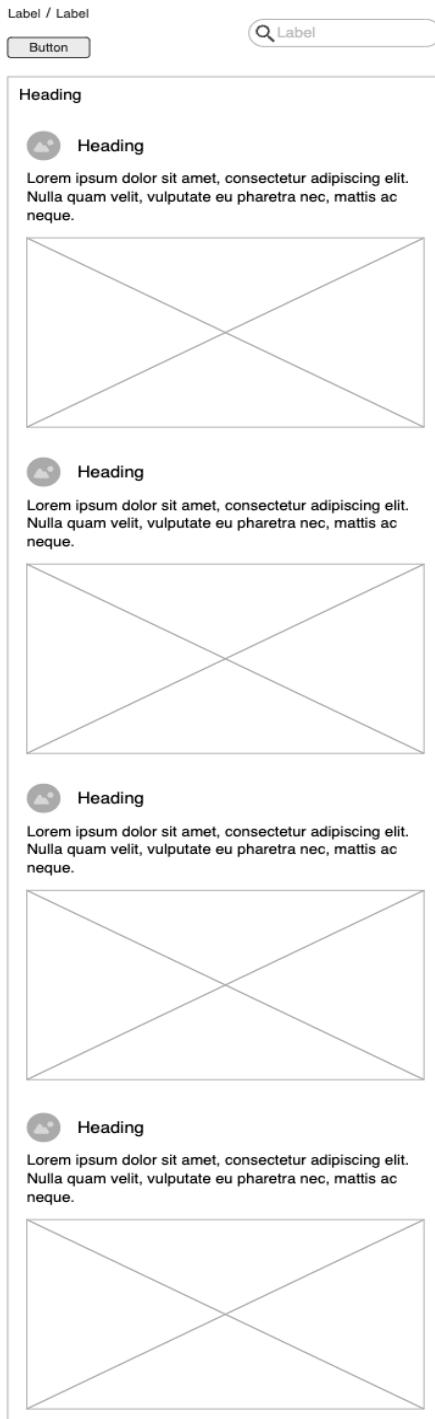


Figura 92: Infográfico Mobile - WireFrame

Fonte: Criação própria

9.2 Prototipação Final

9.2.1 Tela de Login

Esta tela de *login* é o portal de entrada para os consultores de marketing e vendas, garantindo a autenticação segura antes de acessarem os dados. Os usuários deverão inserir suas credenciais de autenticação, como nome de usuário e senha, para acessar a

plataforma. A interface foi projetada para ser amigável e intuitiva, proporcionando uma experiência de login confiável.

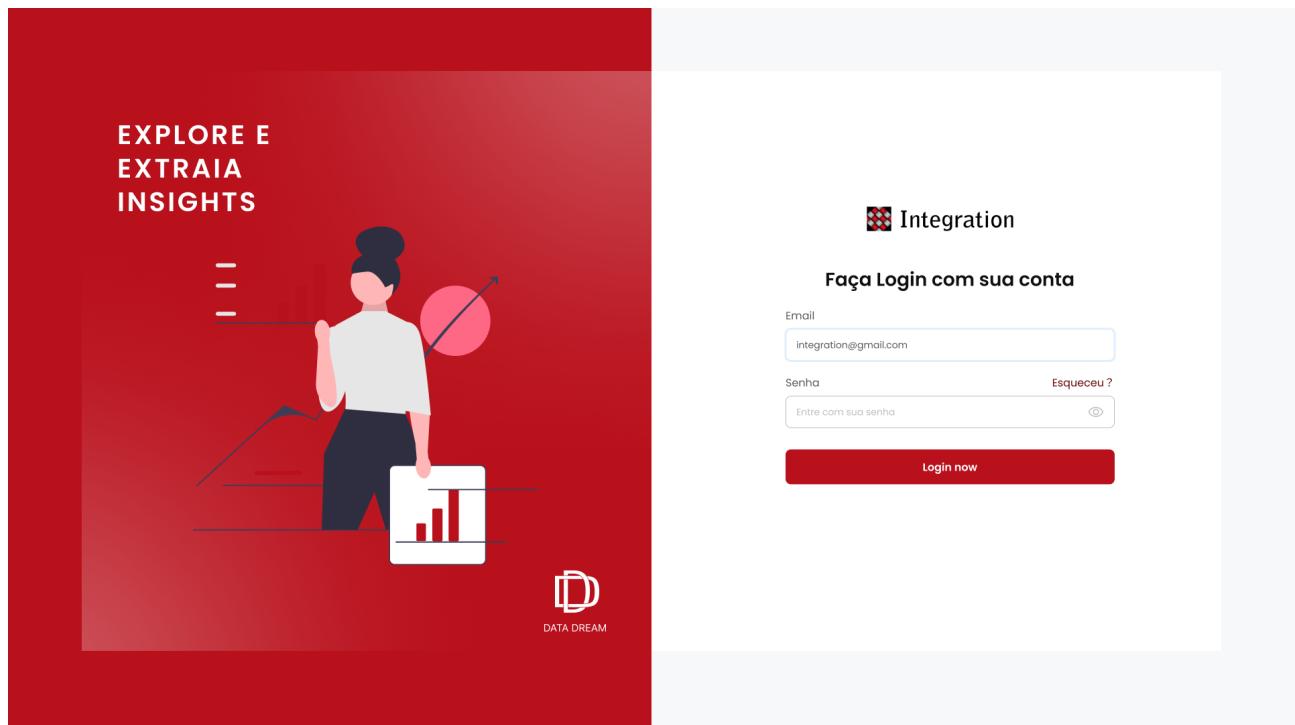


Figura 93: Login Desktop - Prototipação

Fonte: Criação própria

Para dispositivos móveis, a tela anteriormente explicada foi dividida em duas partes: a primeira serve como tela inicial, na qual o usuário pode acessar a tela de login.



Figura 94: Login Mobile 1 - Prototipação

Fonte: Criação própria



Figura 95: Login Mobile 2 - Prototipação

Fonte: Criação própria

9.2.2 Tela de Dashboard

O *Dashboard* é o ponto focal da análise de dados de consumo de empresas e produtos. Esta tela apresenta um painel interativo que permite aos usuários explorar e analisar os dados. Os filtros e opções de personalização permitem que os consultores escolham regiões ou produtos específicos, períodos de tempo e outras variáveis relevantes. Os gráficos e métricas são exibidos de forma clara, fornecendo *insights* vitais para a tomada de decisões estratégicas.

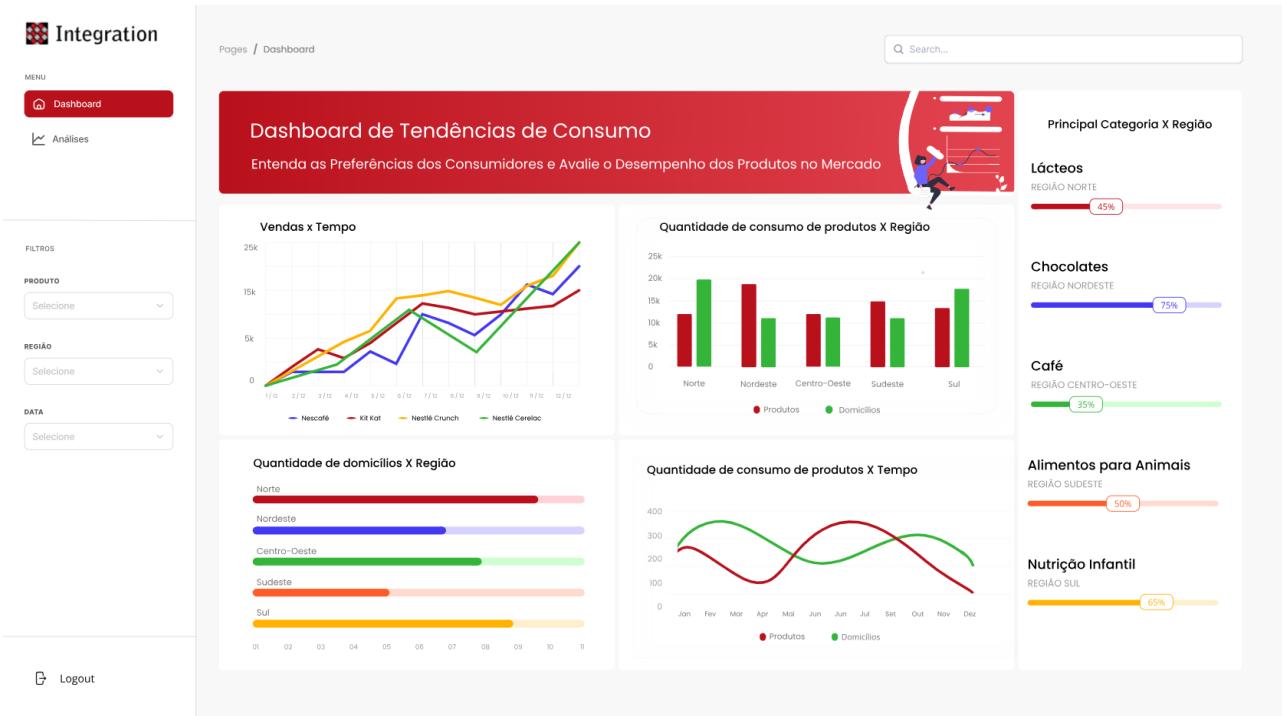


Figura 96: Dashboard Desktop - Prototipação

Fonte: Criação própria

A tela de *Dashboard* para *mobile* é um grande desafio, uma vez que contém diversos gráficos. Portanto, foi necessário criar uma tela para os gráficos (Tela de *Dashboard*) e outra para o menu (Tela de *Menu* - *Dashboard*), que pode ser acessado ao clicar no ícone no canto superior esquerdo.



Figura 97: Dashboard Mobile - Prototipação
Fonte: Elaboração própria

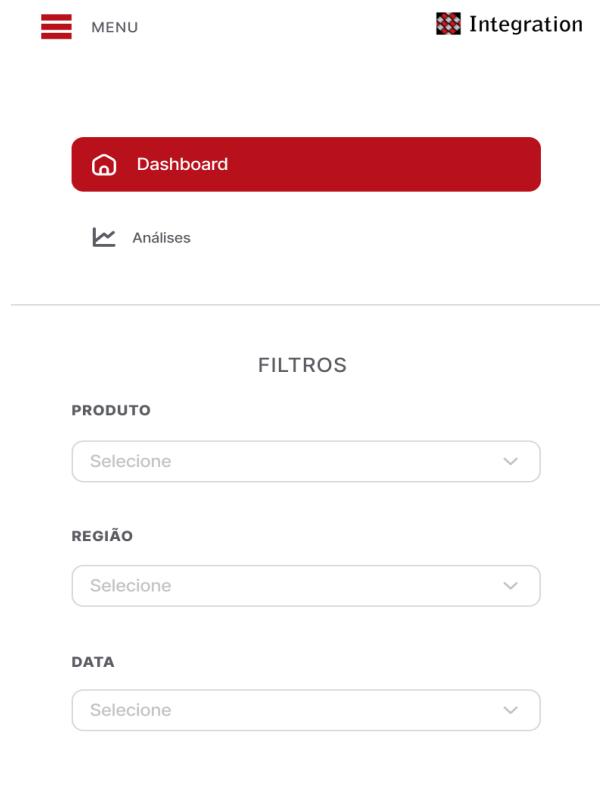


Figura 98: Menu Dashboard Mobile - Prototipação
Fonte: Elaboração própria

9.2.3 Tela de Análises

Nesta tela, os dados recebidos de fontes diversas, como o governo, parceiros e registros de CNPJ, são organizados e apresentados de forma comprehensível. Os consultores podem visualizar as empresas cadastradas, os últimos uploads de dados e explorar informações detalhadas.

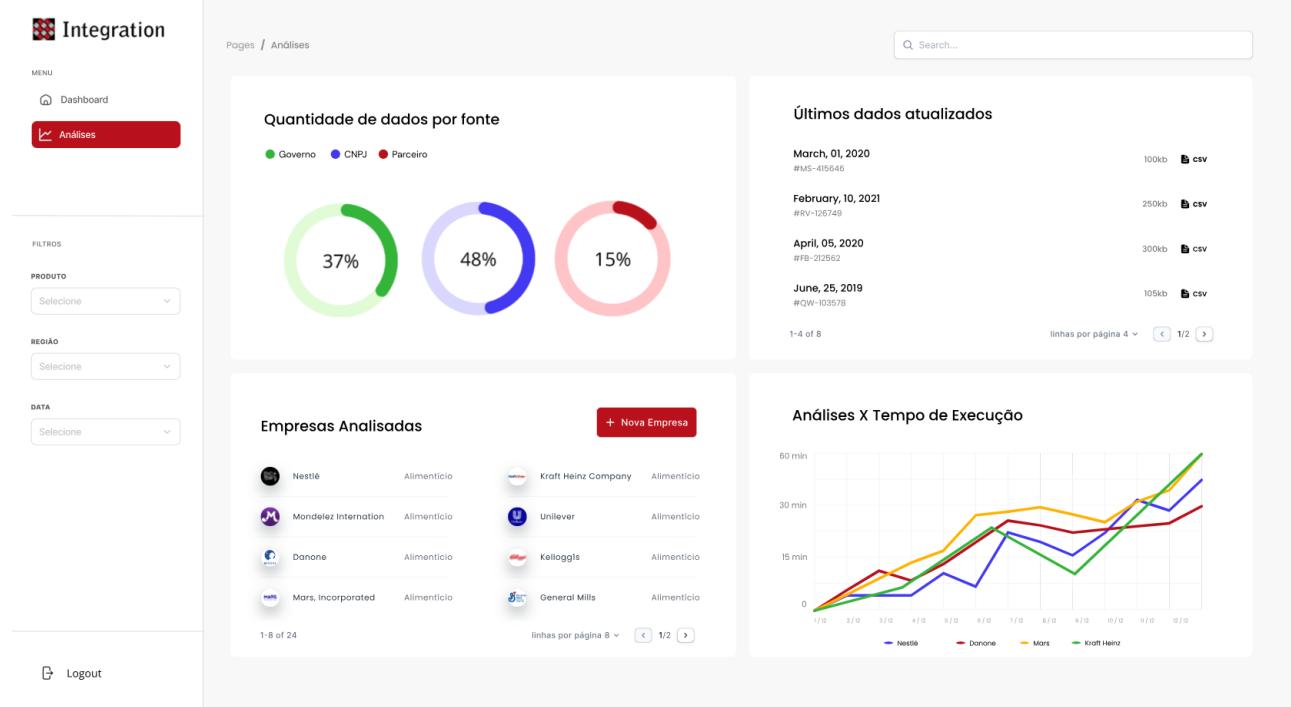


Figura 99: Análises Desktop - Prototipação

Fonte: Criação própria

A tela para *mobile* segue a mesma lógica que a anterior, portanto, foi necessário criar duas telas, uma para análise e outra para o menu.

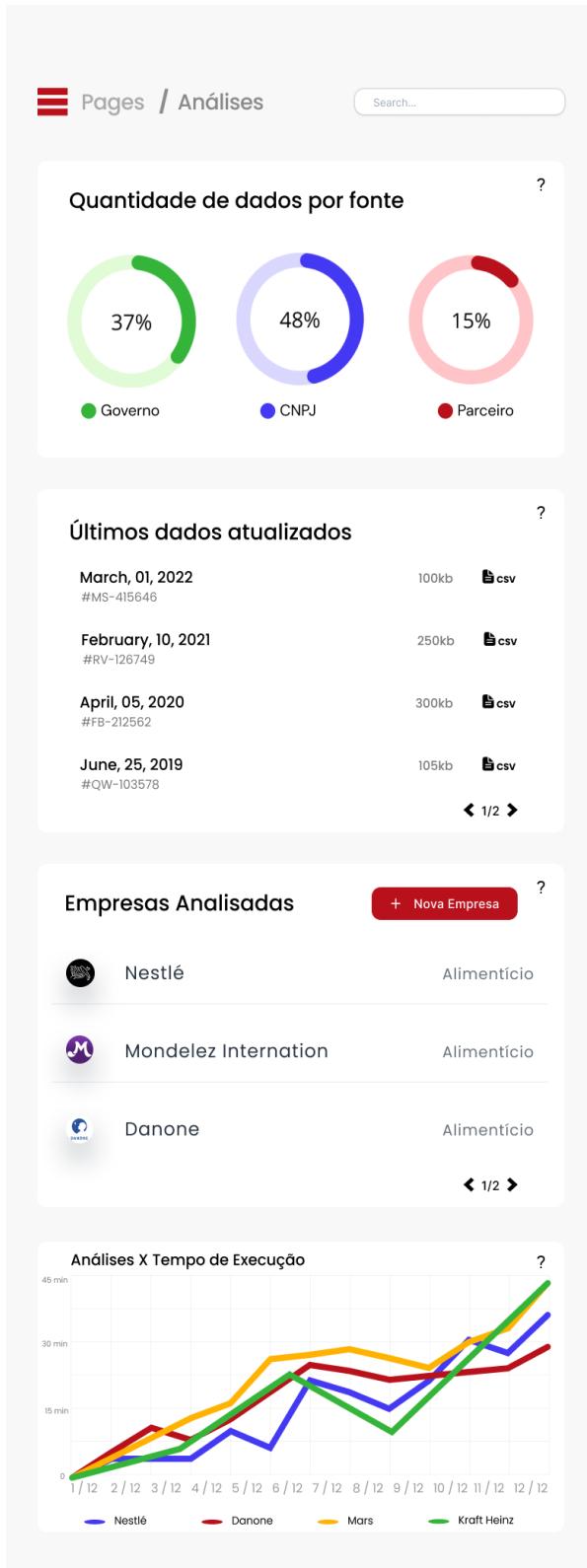


Figura 100: Análises Mobile - Prototipação
Fonte: Elaboração própria

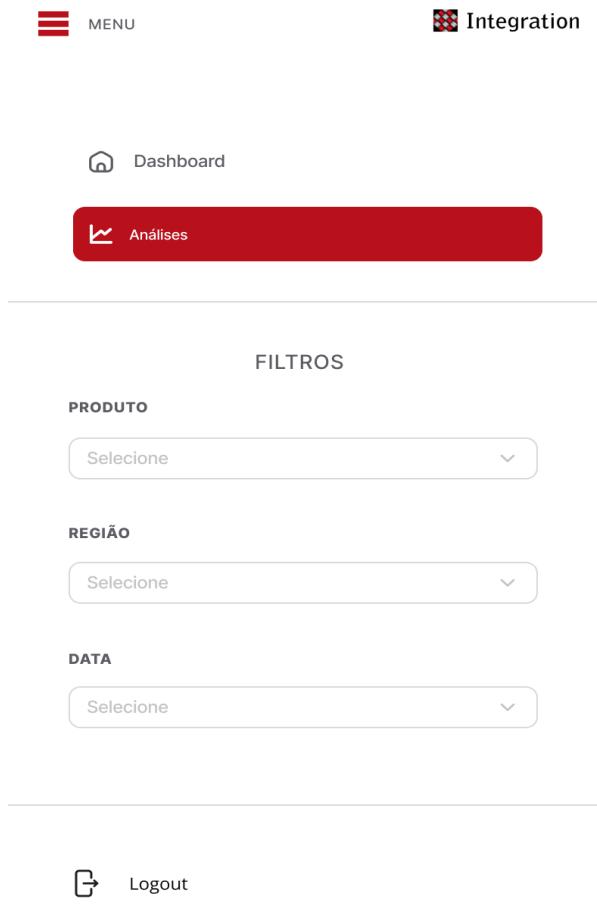


Figura 101: Análises Menu Mobile - Prototipação
Fonte: Elaboração própria

9.2.4 Tela de Infográfico

Um infográfico é uma representação visual de informações ou dados complexos, projetada para tornar a compreensão e a assimilação dessas informações mais acessíveis e eficientes. Combinando elementos gráficos, como gráficos, ícones e ilustrações, com texto conciso, os infográficos transformam dados densos em representações visuais atraentes e de fácil interpretação. A ideia é que seja uma representação para contar algum tipo de história, como neste caso, “Como vender chocolate?”.

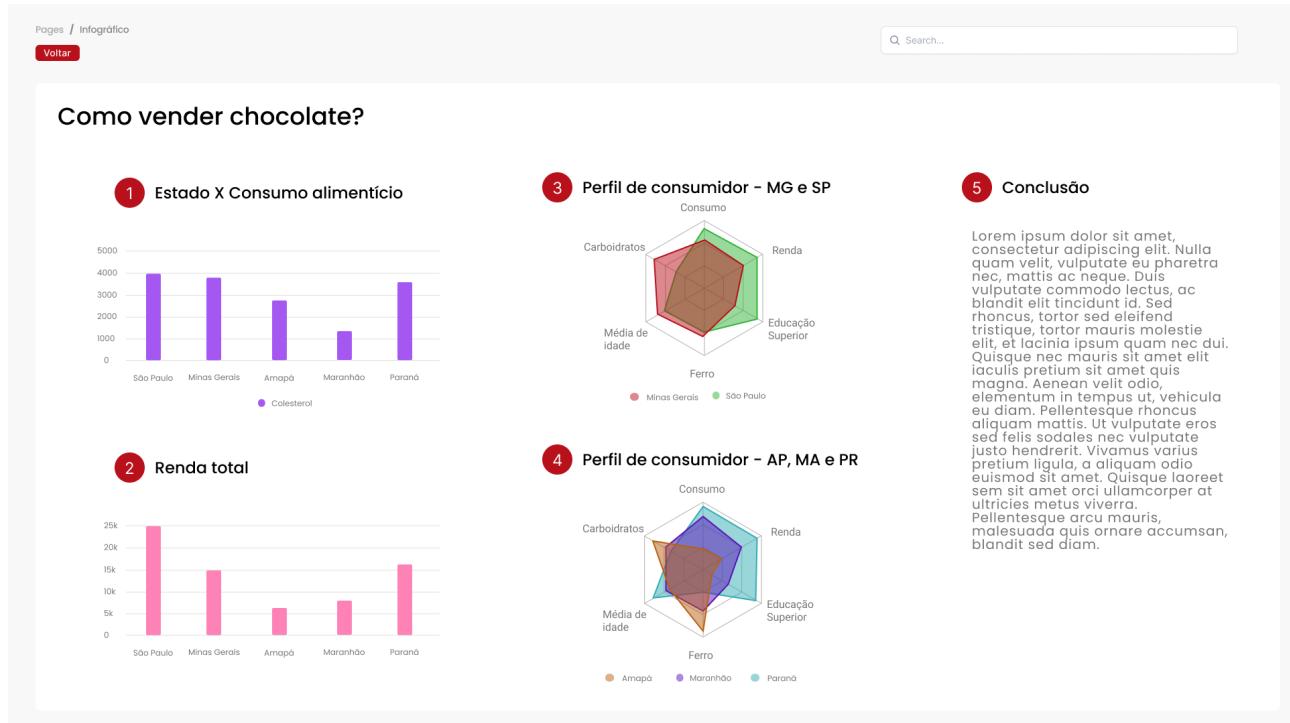


Figura 101: Infográfico de alta fidelidade

Fonte: Autoria Própria

9.3 Menu de Navegação

O menu de navegação é uma parte fundamental do protótipo, fornecendo acesso rápido e fácil às diferentes funcionalidades da plataforma. Os menus disponíveis são os seguintes:

9.3.1 Dashboard

O *Dashboard* é o ponto focal da análise de dados de consumo de empresas e produtos. Esta tela apresenta um painel interativo que permite aos usuários explorar e analisar os dados.

Funcionalidades:

- **Exploração de Dados:** Os usuários podem explorar dados de consumo de empresas e produtos;
- **Filtros e Personalização:** Oferece opções para personalizar os dados, escolhendo regiões, produtos, períodos de tempo e canais;
- **Exibição de Gráficos e Métricas:** Apresenta gráficos e métricas de forma clara para fornecer insights as tomadas de decisões estratégicas.

9.3.2 Análise

Na tela de análise, os dados recebidos de fontes diversas, como o governo, parceiros e registros de CNPJ, são organizados e apresentados de forma comprehensível.

Funcionalidades:

- **Visualização de Empresas:** Permite visualizar informações sobre as empresas cadastradas;
- **Últimos Uploads de Dados:** Mostra os registros dos últimos uploads de dados;
- **Visualização de Tempo por Análise:** Permite visualizar quanto tempo demorou o carregamento de dados para análise de cada empresa.

9.4 Gráficos

9.4.1 Dashboard

Gráficos em linha: A utilização do gráfico em linha foi escolhido para representar a sequência de dados ao longo do tempo, já que ele é ideal para entender as tendências de vendas de produtos ao longo do tempo. Foi utilizado cores diferentes para representar produtos distintos, tornando a análise mais clara e informativa.

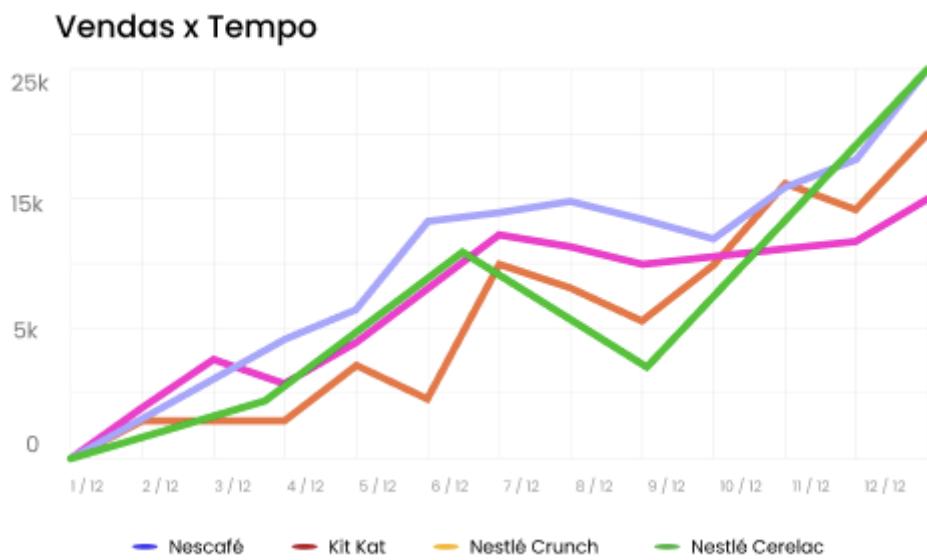


Figura 102: Gráfico de Linha

Fonte: Criação própria

Gráfico de barras duplas em comparação: Este gráfico mostra a quantidade de consumo por regiões, o mesmo foi dividido pelas 5 regiões do Brasil e *color coding* foi utilizado por produto e domicílio para tornar a análise mais clara e concisa. As cores, alturas e tamanhos das colunas facilitam a interpretação. Gráficos de barras são excelentes para comparações de categorias, onde cada barra representa uma categoria distinta, tornando fáceis as comparações entre elas.

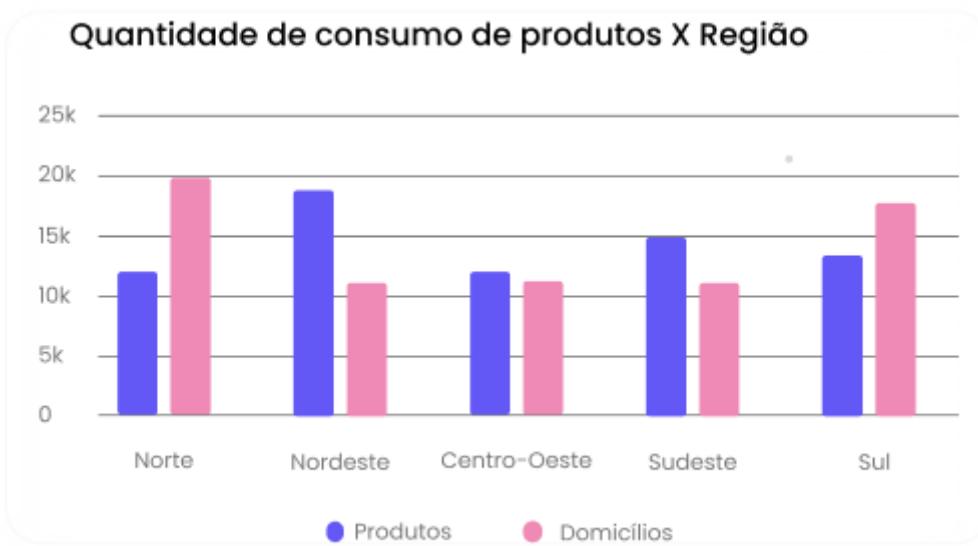


Figura 103: Gráfico de barras duplas

Fonte: Criação própria

Gráfico de barras horizontal: O gráfico de barras horizontal é uma opção que pode ser usada para representar informações de uma forma mais compacta, mas de fácil compreensão. Este gráfico é uma escolha estilística e pode ser útil para apresentar a quantidade de domicílio por regiões de forma visualmente atraente. Este gráfico é utilizado para representar a quantidade de domicílios por região. Também é útil para facilitar comparações quando os números são muito discrepantes, pois a diferença em comprimento pode ser mais fácil de visualizar do que em altura.

Quantidade de domicílios X Região

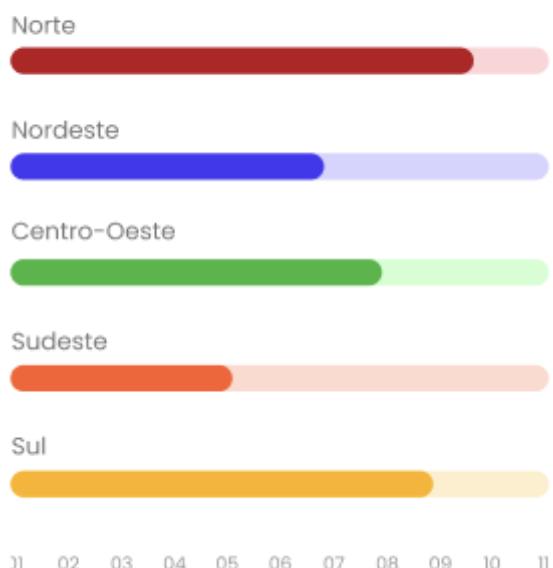


Figura 104: Gráfico de barras horizontais

Fonte: Criação própria

Gráficos em linha: O gráfico de linha é utilizado novamente para representar tendências na quantidade de consumo de produtos ao longo do tempo. Isso é útil para mostrar o crescimento ou queda de produtos e domicílios em um período de 12 meses (1 ano). A linha suave é útil para representar grandes conjuntos de dados onde se quer destacar tendências gerais ao invés de variações pontuais específicas.

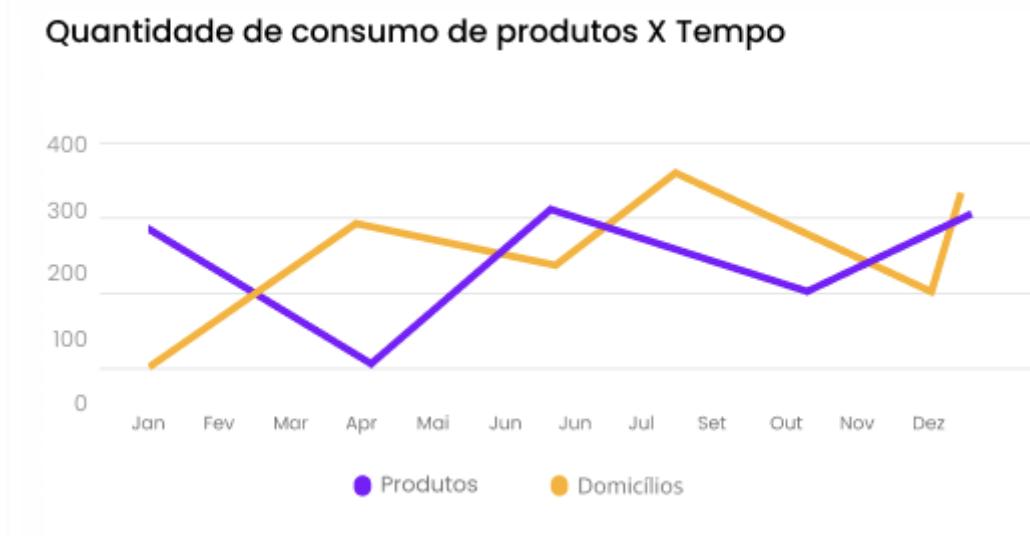


Figura 105: Gráfico de linhas

Fonte: Criação própria

Gráfico de porcentagem: O gráfico de porcentagem é utilizado para destacar a principal categoria de produtos alimentícios consumidos por região.

Principal Categoria X Região



Figura 106: Gráfico de porcentagem

Fonte: Criação própria

Gráfico de barras triplas: O gráfico de colunas triplas em comparação foi desenvolvido com o objetivo de ilustrar as variações entre canal, região e tipo de venda.

Ao segmentar o gráfico em cinco regiões do Brasil e aplicar uma codificação de cores para distinguir franquia, varejo e atacado, juntamente com as diferentes alturas e tamanhos das colunas, possibilitamos uma análise detalhada desses aspectos em cada região.

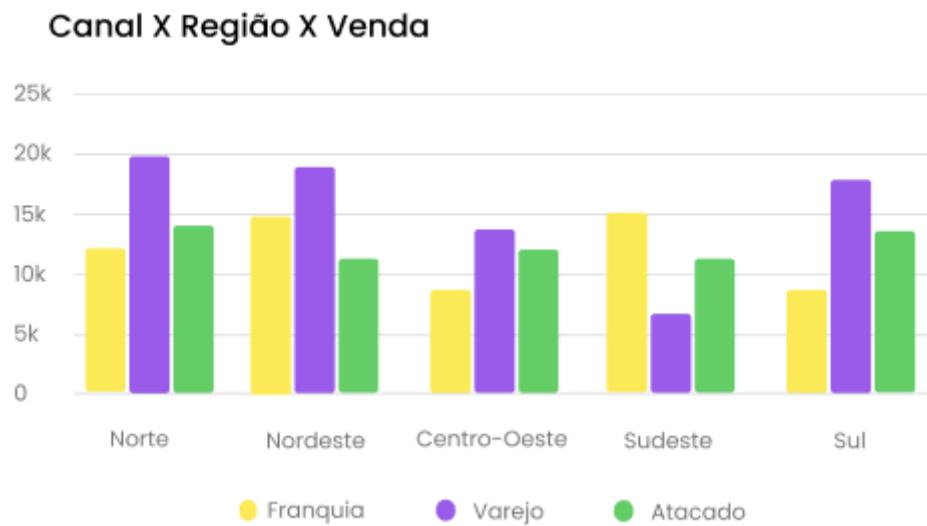


Figura 107: Gráfico de barras triplas

Fonte: Criação própria

9.4.2 Análise

Gráfico de rosquinha: Os gráficos de rosquinha (*Donut Chart*) são empregados para visualizar proporções em diversos componentes, representando assim as relações proporcionais dos dados por fonte. Essa escolha torna as análises mais visuais e de fácil compreensão.

Quantidade de dados por fonte

● Governo ● CNPJ ● Parceiro

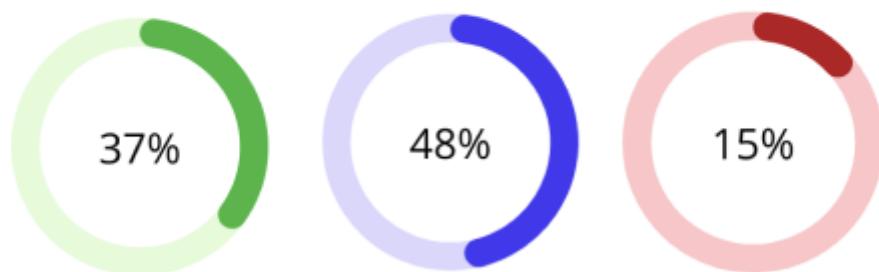


Figura 108: Gráfico de donut

Fonte: Criação própria

A lista abaixo apresenta de maneira clara os dados mais recentes, incluindo informações essenciais como a data de atualização, tamanho do arquivo e *links* para os próprios arquivos *csv* ou *txt*. Uma lista com datas de atualização e tamanhos de arquivos é uma forma direta de apresentar informações sequenciais. A organização em forma de tabela facilita a visualização dos dados e a paginação permite que o usuário navegue por grandes conjuntos de dados sem sobrecarregar a visão.

Últimos dados atualizados

March, 01, 2020 #MS-415646	100kb	 csv
February, 10, 2021 #RV-126749	250kb	 csv
April, 05, 2020 #FB-212562	300kb	 csv
June, 25, 2019 #QW-103578	105kb	 csv

1-4 of 8

1/2

Figura 109: Lista de dados

Fonte: Criação própria

Mantendo consistência, é utilizado novamente o gráficos em linha para representar análises de diferentes produtos e seus tempos de execução ao longo do aplicativo. Este é utilizado para mostrar a relação entre o tempo de execução de análises para várias empresas ao longo do tempo. Geralmente o gráfico é escolhido para esse tipo de dado porque é excelente para visualizar tendências, movimentos e mudanças ao longo de um período contínuo, facilitando a comparação do desempenho entre diferentes entidades em relação ao tempo.

Análises X Tempo de Execução

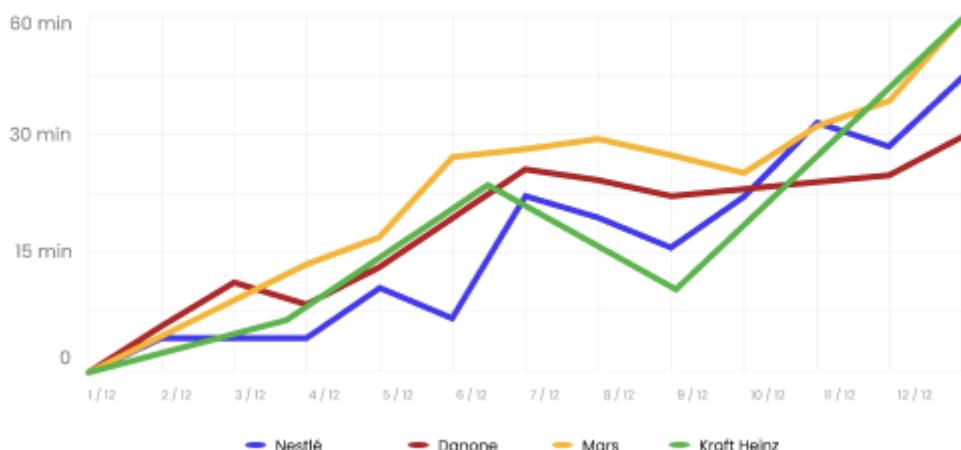


Figura 110: Gráfico de linha

Fonte: Criação própria

Por fim, é incluído mais uma lista, mas que agora exibe as empresas analisadas, fornecendo informações importantes como nome, setor. A apresentação em lista com logos e categorias das empresas permite uma rápida identificação e comparação. A funcionalidade de adicionar uma "Nova Empresa" é colocada de forma destacada, o que sugere a possibilidade de customização ou inserção de novos dados por parte do usuário.

Empresas Analisadas			+ Nova Empresa		
	Nestlé	Alimentício		Kraft Heinz Company	Alimentício
	Mondelez International	Alimentício		Unilever	Alimentício
	Danone	Alimentício		Kellogg's	Alimentício
	Mars, Incorporated	Alimentício		General Mills	Alimentício

1-8 of 24

1/2

Figura 111: Lista de empresas

Fonte: Criação própria

9.5 Técnicas avançadas utilizadas no design

Este tópico visa fornecer uma análise aprofundada e embasada das técnicas avançadas de design, destacando estratégias específicas para otimizar a eficácia do projeto, com uma atenção especial à inclusão de pessoas com deficiências visuais (mais especificamente, diferentes graus de daltonismo). As práticas mencionadas são fundamentadas em diretrizes reconhecidas e apoiadas por especialistas em design e usabilidade.

9.5.1 Cores acessíveis

A abordagem de cores acessíveis se baseia nas diretrizes do Web Content Accessibility Guidelines (WCAG) 3.0, que fornecem critérios rigorosos para garantir uma experiência inclusiva. O projeto adota uma paleta de cores cuidadosamente selecionada, considerando não apenas o contraste, mas também a combinação de tonalidades e a semântica associada a diferentes cores. Esta prática, corroborada por estudos da Awwwards e Smashing Magazine, assegura não apenas a acessibilidade, mas também a estética visual. Diante disso, o projeto conta com três diferentes escolhas de paletas de cores, como é possível observar nas imagens abaixo:

9.5.1.1 Visualização para pessoas com daltonismo

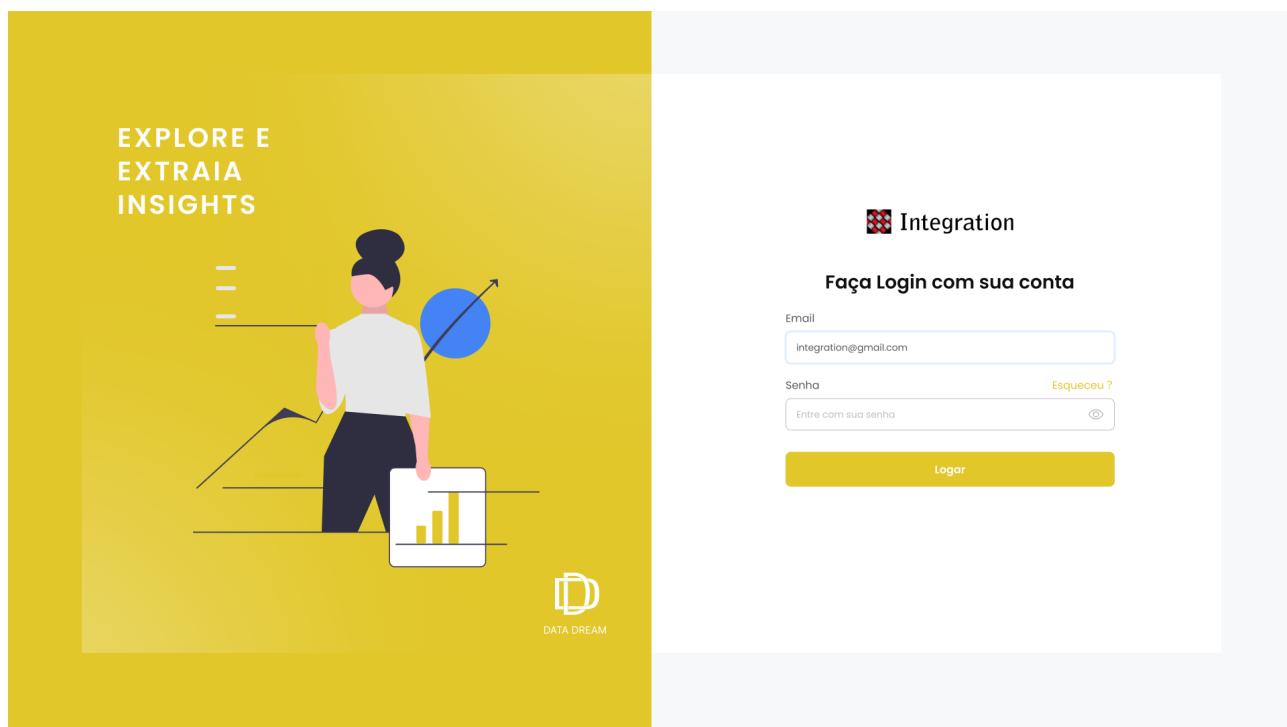


Figura 112: Login - Daltonismo

Fonte: Criação própria

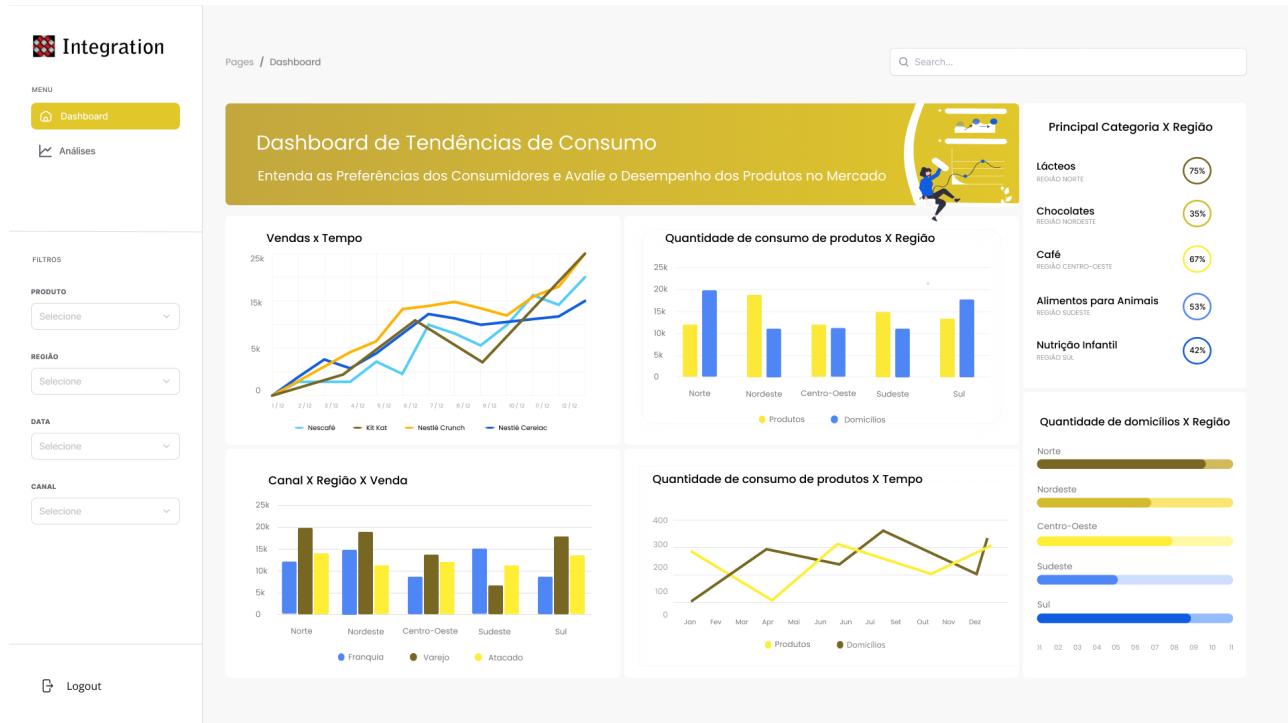


Figura 112: Dashboard - Daltonismo

Fonte: Criação própria

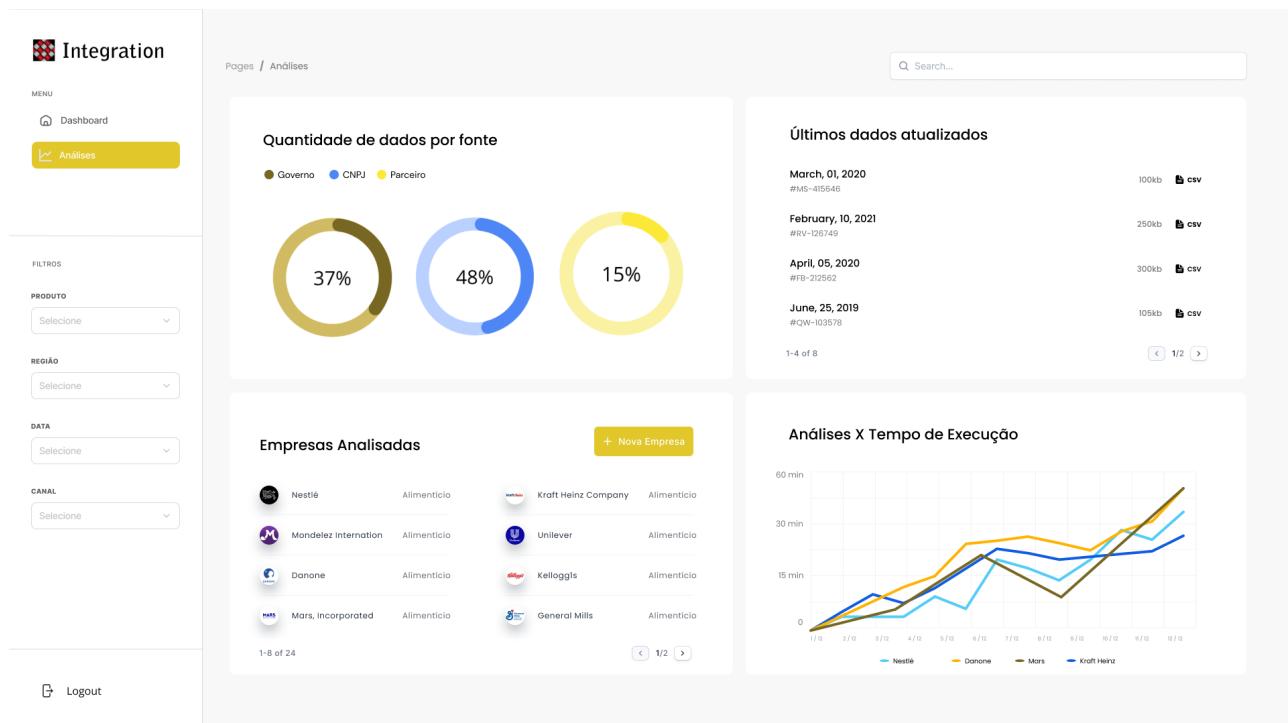


Figura 113: Análises - Daltonismo

Fonte: Criação própria

9.5.1.1 Visualização para pessoas com dark

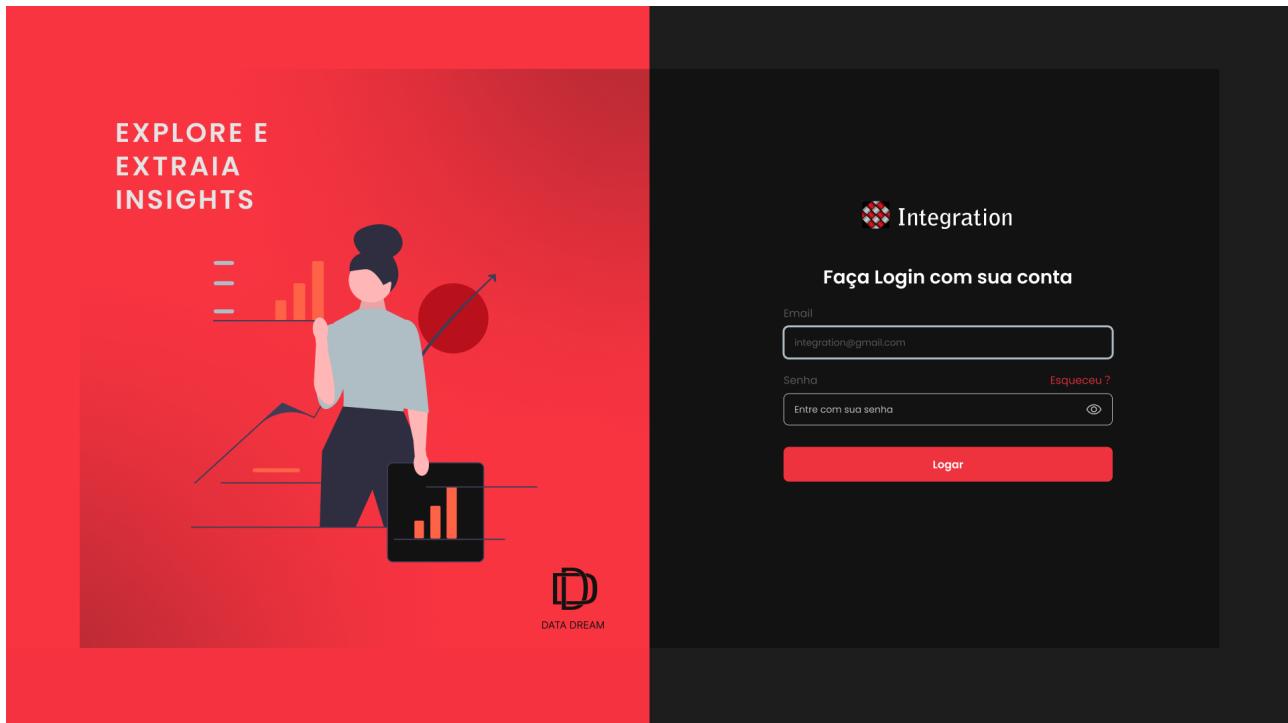


Figura 114: Login - Dark

Fonte: Criação própria

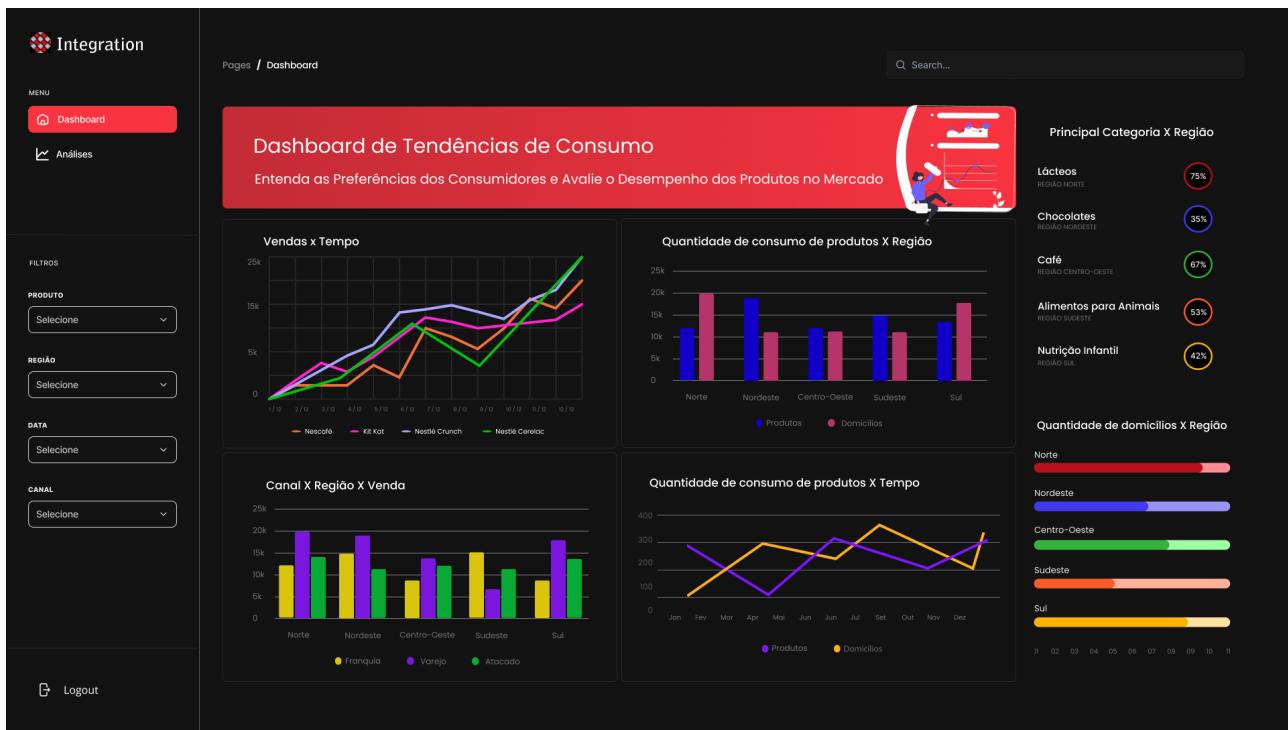


Figura 115: Dashboard- Dark

Fonte: Criação própria

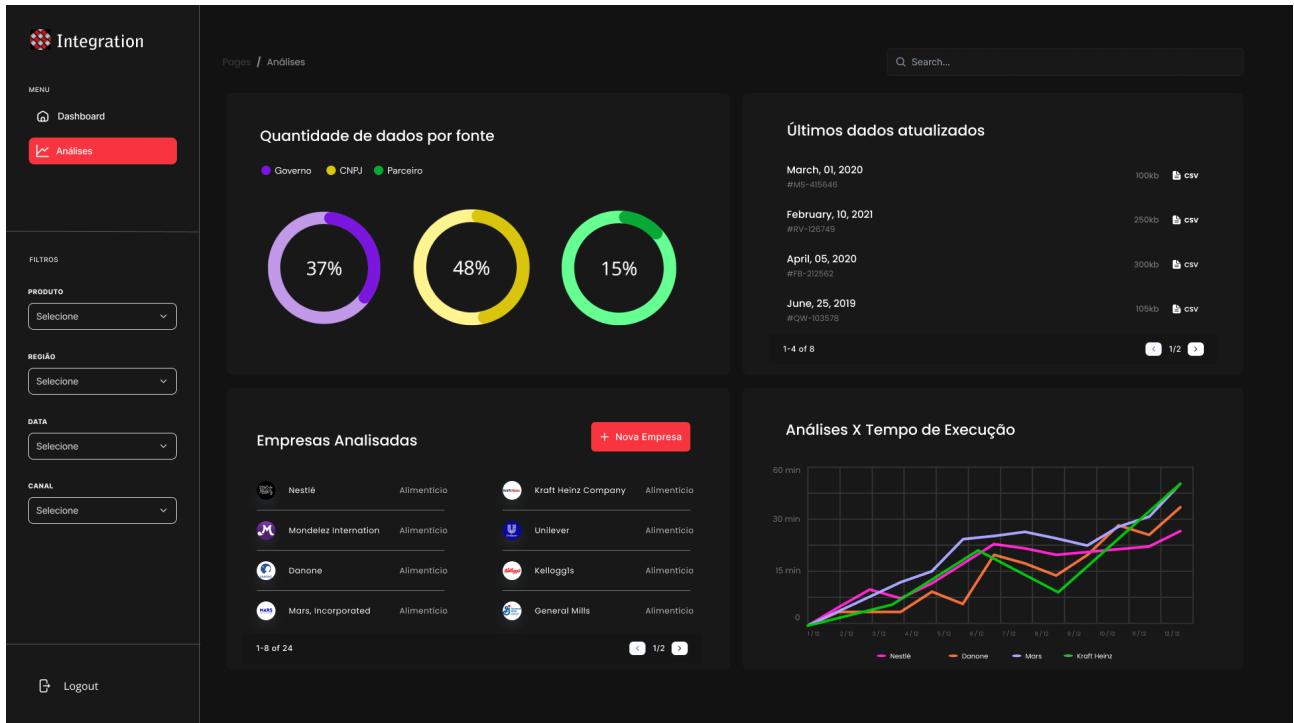


Figura 116: Análises - Dark

Fonte: Criação própria

9.5.2 Sistema de Grid

A implementação do sistema de *grid* é fundamentada em princípios de design visual, como proporção áurea e teoria da *gestalt*, conforme sugerido por especialistas como Ellen Lupton e Timothy Samara. Além disso, é inspirada nas diretrizes de design responsivo do Google Material Design. O projeto utiliza um sistema de *grid* hierárquico, incorporando flexibilidade estrutural e alinhamento consistente para promover uma experiência visual coesa. Essa abordagem é respaldada por estudos da Nielsen Norman Group, que destacam a importância do *layout* na experiência do usuário.

O *Grid* Hierárquico representa uma abordagem inovadora adotada no design do nosso projeto, proporcionando uma estrutura visual que vai além da organização tradicional. Esse sistema de *grid* se destaca pela introdução de múltiplos níveis de organização, criando uma hierarquia visual que guia os usuários de maneira intuitiva e eficaz.

A principal função desse sistema é estabelecer uma hierarquia visual clara, facilitando a compreensão do conteúdo. Elementos posicionados em camadas superiores ganham maior destaque, enquanto elementos em camadas inferiores são percebidos como secundários, proporcionando uma experiência de usuário mais intuitiva. A seguir, encontra-se uma exemplificação de um sistema composto por grid hierárquico.

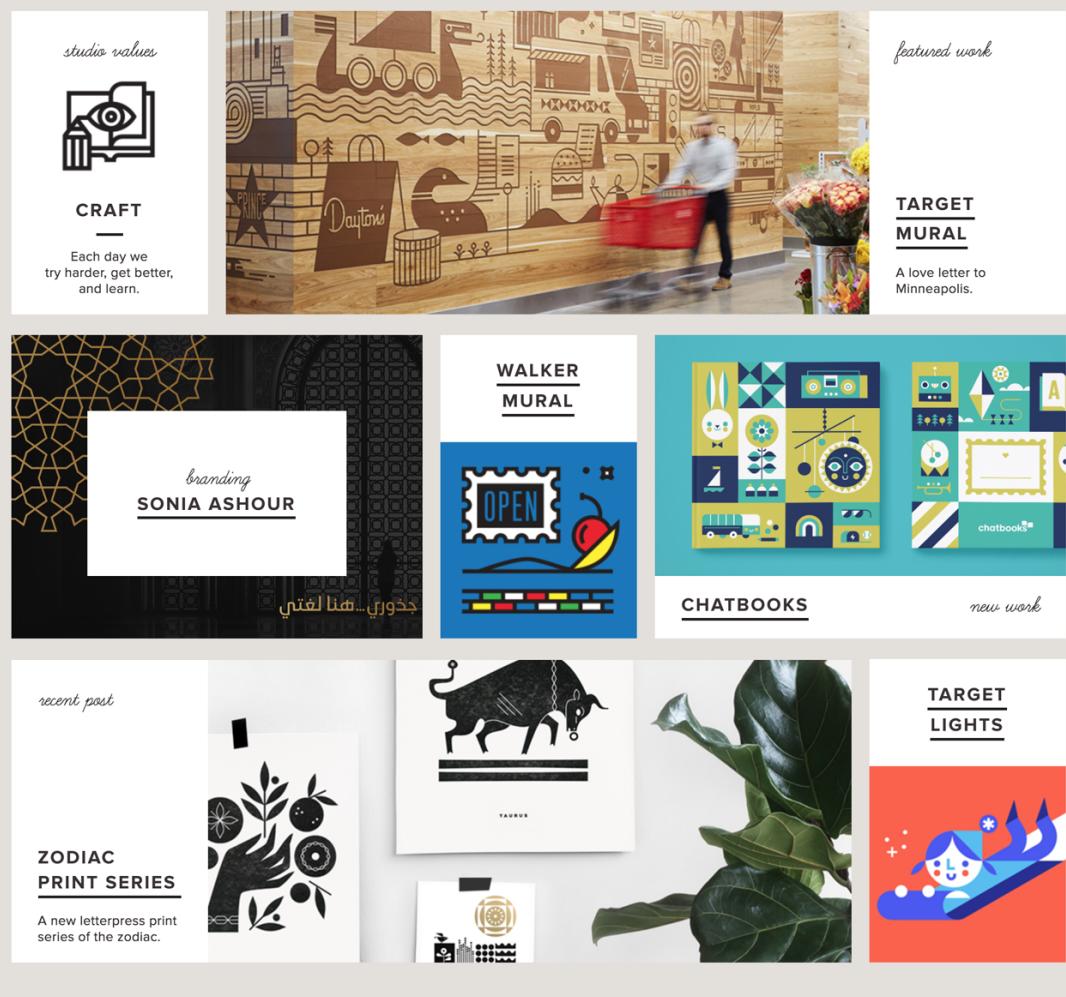


Figura 117: Exemplo de uma aplicação com grid hierárquico

Fonte: [Grids](#)

Abaixo, a nossa tela de dashboard de alta fidelidade com destaque para os diferentes componentes do grid hierárquico que implementamos.

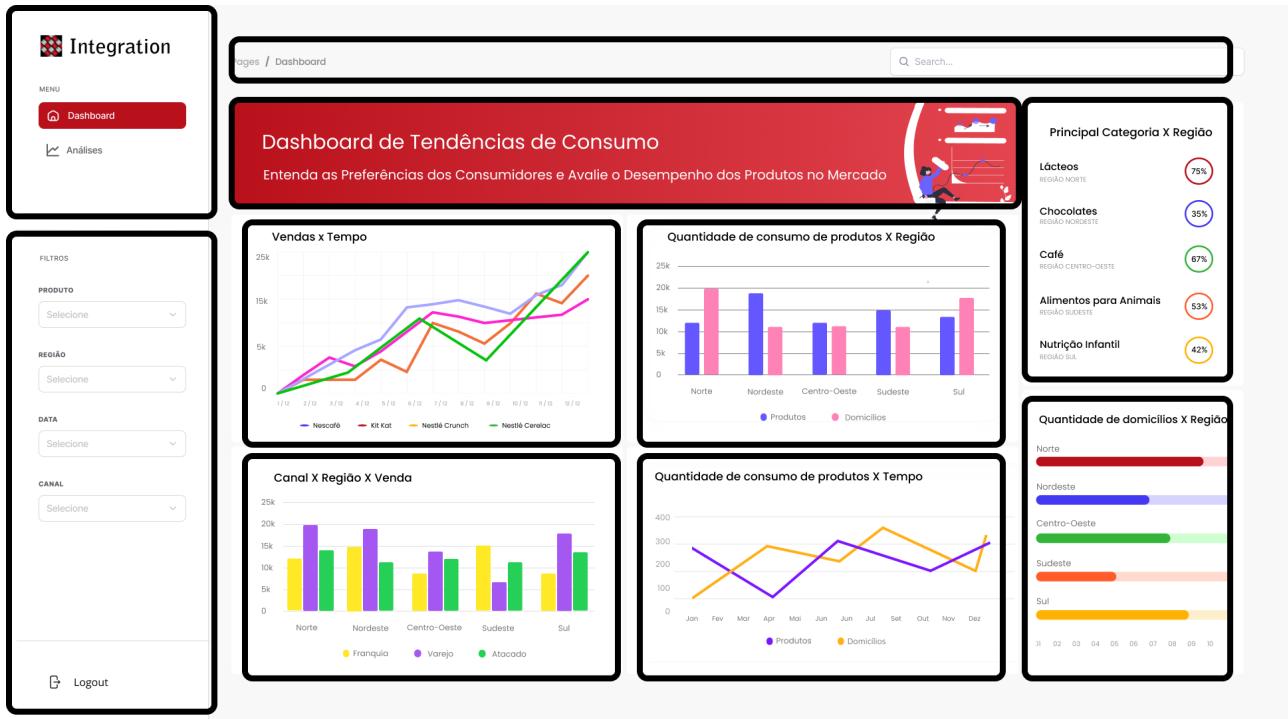


Figura 118: Dashboard com Grid - Prototipação

Fonte: Criação própria

Por fim, a nossa tela de dashboard de baixa fidelidade com destaque para os diferentes componentes do grid hierárquico.

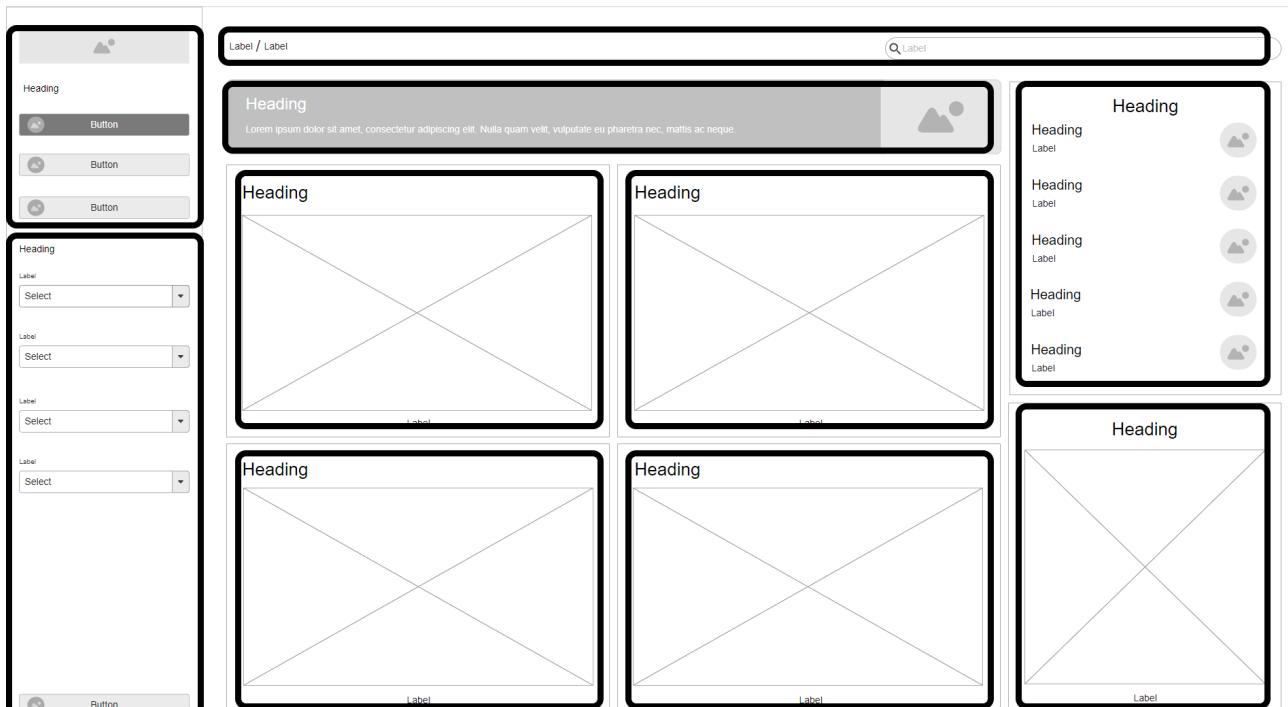


Figura 119: Dashboard com Grid - WireFrame

Fonte: Criação própria

9.5.3 Design responsivo:

A estratégia de design responsivo será implementada considerando as diretrizes do *World Wide Web Consortium* (W3C) e as práticas recomendadas do *Google Developers*. A abordagem inclui não apenas a reorganização de elementos, mas também a otimização de imagens, priorização de conteúdo e considerações de desempenho. Estudos de caso, como os apresentados pela *Smashing Magazine* e *A List Apart*, influenciam diretamente as decisões de design responsivo, garantindo uma experiência consistente e eficaz em todos os dispositivos. Pensando nisso, desenvolvemos uma prototipação inicial, tanto em baixa, quanto em alta fidelidade, que conta com a presença de responsividade.

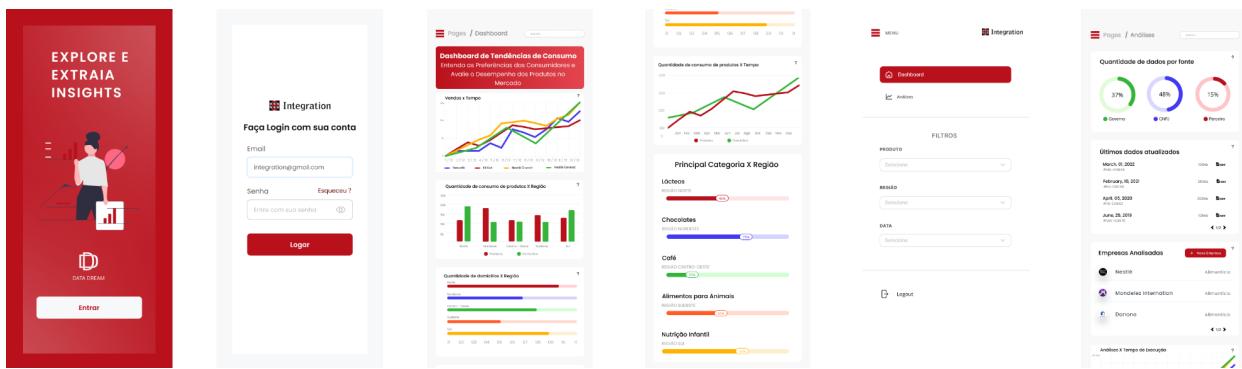


Figura 120: Prototipação do Mobile

Fonte: Criação própria

9.5.4 Conclusão

Ao integrar essas técnicas avançadas de design, o projeto busca não apenas atender, mas superar as expectativas dos usuários. Cada decisão de design é respaldada por princípios fundamentados em pesquisas, estudos de caso e diretrizes reconhecidas, garantindo não apenas uma estética atraente, mas uma experiência de usuário fundamentada em práticas sólidas de design e usabilidade.

9.6 Feedbacks e Iterações

Este documento aborda o impacto significativo do feedback recebido de nosso parceiro, a Integration, e do professor de UX, Francisco Escobar, no processo de desenvolvimento de nosso projeto. Vamos detalhar como essas orientações moldaram as iterações em nosso design de UX, particularmente nos wireframes.

9.6.1 Feedback do parceiro

9.6.1.1 Detalhamento na Documentação do Wireframe

O feedback inicial do nosso parceiro destacou a necessidade de detalhar a documentação, dada a essencial simplicidade do wireframe. Reconhecendo que o wireframe deve focar em layout e funcionalidade, sem detalhes excessivos, a resposta a essa orientação envolveu um aprofundamento na documentação. Realizamos uma análise minuciosa de cada seção do wireframe, descrevendo em detalhes o conteúdo, a interação com outras partes e a finalidade de cada elemento. Essa abordagem permitiu manter o wireframe limpo e focado, enquanto a documentação complementar proporcionou uma compreensão mais clara do design e demonstrou como cada parte contribui para a experiência geral do usuário.

9.6.1.2 Seleção de Gráficos

O parceiro também sugeriu detalhar a escolha de gráficos. Respondendo a esse feedback, justificamos detalhadamente essas escolhas, demonstrando como cada elemento foi selecionado com base nos auto estudos e nas aulas de UX ministradas sobre esse tema.

9.6.1.3 Implementação de Gráficos Geoespaciais

A inclusão de gráficos geoespaciais para visualização de dados por região foi outra melhoria significativa sugerida pelo parceiro. Esses gráficos, que se atualizam com a escolha de localização, mostram áreas de alta demanda ou vendas, melhorando a interatividade e a relevância dos dados apresentados.

9.6.2 Feedback do professor

9.6.2.1 Simplificação do Wireframe

O professor aconselhou a remoção de elementos específicos, como nomes e títulos, do wireframe. Esta orientação nos levou a focar na finalidade principal do wireframe, que é visualizar a disposição e a funcionalidade do layout. Removendo esses detalhes e substituindo-os por marcadores genéricos, conseguimos manter o wireframe focado e eficiente na comunicação da estrutura do design.

10. Análise dos dados

10.1. IBGE

O Plano de Dados Abertos para o período de 2020-2022 é um guia que orienta a disponibilização, atualização e disseminação de informações abertas. Seu objetivo é promover a transparência, melhorar os serviços públicos e atender às necessidades da sociedade civil, permitindo o acesso aos dados publicados pela instituição. Para esse projeto, é utilizado um conjunto de dados, Produto Interno Bruto dos Municípios, disponibilizado que será descrito abaixo.

10.1.1 Tabela 5938 - Produto interno bruto a preços correntes, impostos, líquidos de subsídios, sobre produtos a preços correntes e valor adicionado bruto a preços correntes total e por atividade econômica, e respectivas participações

A estrutura desta tabela consiste em 3 valores: 1. Variáveis, que são as pesquisas e seus valores, por exemplo: Produto interno bruto a preços correntes (47 pesquisas); 2. Ano, quando essa pesquisa rodou no país (2002 - 2020); 3. Unidade Territorial, de onde você deseja ver a pesquisa, exemplo unidades federativas (6 opções). Abaixo, pode-se ver uma figura que demonstra a estrutura dos dados da Tabela 5938.

Produto interno bruto a preços correntes, impostos, líquidos de subsídios, sobre produtos a preços correntes e valor adicionado bruto a preços correntes total e por atividade econômica, e respectivas participações - Referência 2010	
Variável (1)	
	⌚ Ano (1)
Unidade Territorial (27)	

Figura 121: Layout Tabela 5938

Fonte: Criação própria

Diante de 47 pesquisas, o grupo acredita que somente a "PIB a preços correntes (Mil Reais)" é relevante para o projeto, já que esta demonstra o valor total de bens e serviços produzidos em um país durante um determinado período, podendo avaliar o tamanho da economia desse país. Para isso, foi escolhido em cada valor: 1. *Variável - PIB*

a preços correntes (Mil Reais); 2. Ano - 2022; 3. Unidade Territorial - Unidade da Federação.

Além disso, foi realizado um tratamento dos dados para o PIB de todos os municípios em todos os anos (2002-2020). Assim, os valores são os seguintes: 1. Variável - *PIB a preços correntes (Mil Reais)*; 2. Ano - *Todos os anos disponíveis (2002-2020)*; 3. *Unidade Territorial - Município*. Para o tratamento, foi necessária a criação de um Notebook, onde os dados foram lidos e os valores nulos foram removidos, como mostram os códigos abaixo. Além disso, no momento de download é necessário entrar no arquivo e excluir o título.

Fonte para maiores esclarecimentos: [Tabela 5938](#)

10.1.2 Tabela 5939 - Índice de Gini da distribuição do produto interno bruto a preços correntes

A segunda tabela escolhida foi a do índice de Gini, que avalia a desigualdade na distribuição de renda ou riqueza. Este índice varia de 0 a 1, sendo que valores mais próximos de 0 indicam menor desigualdade. Esta tabela segue o mesmo layout anteriormente explicado.

Diante de 5 pesquisas, o grupo acreditou que 2 delas são relevantes para o projeto: 1."Índice de Gini da distribuição do produto interno bruto a preços correntes" e 2."Índice de Gini da distribuição do valor adicionado bruto a preços correntes da indústria". O primeiro avalia esta desigualdade pela distribuição da renda ou riqueza no contexto de toda a economia de um país, já o segundo se concentra especificamente na desigualdade dentro do setor industrial do país. Já que o projeto foca em um setor específico, estes dados podem ajudar. Para isso, foi escolhido em cada valor: 1. Variável - Índice de Gini da distribuição do produto interno bruto a preços correntes ou Índice de Gini da distribuição do valor adicionado bruto a preços correntes da indústria; 2. Ano - 2022; 3. Unidade Territorial - Unidade da Federação.

Fonte para maiores esclarecimentos: [Tabela 5939](#)

10.2. Pesquisa de orçamento familiar (POF)

Os dados que serão analisados nesse documento são de caráter experimental, tendo em vista que tratam-se de novas estatísticas que ainda estão em fase de teste e sob avaliação.

Fonte dos dados: IBGE, Diretoria de Pesquisas, Coordenação de Pesquisas por Amostra de Domicílios, Pesquisa de Orçamentos Familiares.

Formato: xls e ods

Tamanho: o tamanho médio das tabelas de 2008-2009 é de 176 KB.

Frequência de atualização de cada fonte: nos sites oficiais não foi possível encontrar uma data específica para a ocorrência de cada uma das pesquisas, principalmente porque se tratam de dados experimentais. Contudo, foi possível encontrar o tempo de duração da pesquisa (12 meses) e o tempo previsto entre o início da coleta e a divulgação dos dados (18 meses).

10.2.1 Tabelas 2017 - 2018

Abaixo encontram-se as descrições detalhadas e exemplificações com imagens de cada uma das tabelas encontradas na fonte de dados enviada.

10.2.1.1 Tabela 1b

A tabela em questão apresenta dados resumidos acerca da proporção de pessoas das famílias residentes, da proporção de pessoas com algum grau de vulnerabilidade, do índice de vulnerabilidade multidimensional não monetário e contribuições para o índice de vulnerabilidade multidimensional não monetário do Brasil segundo os condicionantes e subgrupos selecionados.

Tabela 1b - Proporção de pessoas das famílias residentes, proporção de pessoas com algum grau de vulnerabilidade, IVM-NM e contribuições para o IVM-NM do Brasil, segundo os condicionantes e subgrupos selecionados-período 2017-2018

Condicionantes e subgrupos selecionados	Proporção de pessoas das famílias residentes (%)	Proporção de pessoas com algum grau de vulnerabilidade (%)	IVM-NM	Contribuição para o IVM-NM do Brasil	Contribuição para o IVM-NM do Brasil (%)
Localização geográfica do domicílio					
Brasil	100,0	63,8	7,7	7,7	100,0
Urbano	85,3	58,8	6,3	5,4	69,8
Rural	14,7	92,9	15,8	2,3	30,2
Grandes Regiões					
Norte	8,6	86,2	13,7	1,2	15,2
Nordeste	27,3	82,3	12,2	3,3	43,1
Sudeste	42,2	52,1	5,0	2,1	27,2
Sul	14,3	47,3	3,9	0,6	7,2
Centro-Oeste	7,7	68,2	7,3	0,6	7,3
Pessoa de referência					
Composição demográfica					
Até 24 anos	3,1	69,8	8,3	0,3	3,4
25 a 49 anos	52,5	63,8	7,7	4,1	52,7
50 a 64 anos	28,9	63,2	7,7	2,2	29,1
65 anos ou mais	15,5	63,8	7,4	1,1	14,9
Cor ou raça					
Brancos	41,4	49,4	4,8	2,0	25,7
Pretos e pardos	57,2	74,5	9,9	5,6	73,2
Sexo					
Homem	59,7	60,8	7,1	4,2	55,0
Mulher	40,3	68,3	8,6	3,5	45,0
Nível de instrução					
Sem instrução	7,0	94,5	16,7	1,2	15,2
Ensino fundamental incompleto	36,8	81,9	11,0	4,0	52,4
Ensino fundamental completo	8,8	64,5	7,0	0,6	8,0
Ensino médio incompleto	5,0	70,1	7,3	0,4	4,8

Figura 122: Tabela 1B

Fonte: POF 2017 - 2018

10.2.1.2 Tabela 2b

A tabela em questão apresenta dados resumidos acerca da proporção de pessoas das famílias residentes, da proporção de pessoas com algum grau de pobreza, do Índice de Pobreza Multidimensional não Monetário do Brasil segundo os condicionantes e subgrupos selecionados.

Tabela 2b - Proporção de pessoas das famílias residentes, proporção de pessoas com algum grau de pobreza, IPM-NM e contribuições para o IPM-NM do Brasil, segundo os condicionantes e subgrupos selecionados-período 2017-2018

Condicionantes e subgrupos selecionados	Proporção de pessoas das famílias residentes (%)	Proporção de pessoas com algum grau de pobreza (%)	IPM-NM	Contribuição para o IPM-NM do Brasil	Contribuição para o IPM-NM do Brasil (%)
Localização geográfica do domicílio					
Brasil	100,0	22,3	2,3	2,3	100,0
Urbano	85,3	17,3	1,6	1,4	59,5
Rural	14,7	51,1	6,4	0,9	40,5
Grandes Regiões					
Norte	8,6	43,8	5,2	0,4	19,4
Nordeste	27,3	38,2	4,3	1,2	51,1
Sudeste	42,2	12,6	1,1	0,5	19,8
Sul	14,3	8,9	0,6	0,1	3,9
Centro-Oeste	7,7	20,1	1,7	0,1	5,8
Pessoa de referência					
Composição demográfica					
Até 24 anos	3,1	24,1	2,5	0,1	3,4
25 a 49 anos	52,5	22,3	2,3	1,2	53,2
50 a 64 anos	28,9	22,6	2,4	0,7	30,2
65 anos ou mais	15,5	21,5	2,0	0,3	13,2
Cor ou raça					
Brancos	41,4	12,1	1,1	0,5	19,7
Pretos e pardos	57,2	29,8	3,2	1,8	79,1
Sexo					
Homem	59,7	20,2	2,1	1,2	53,8
Mulher	40,3	25,4	2,7	1,1	46,2
Nível de instrução					
Sem instrução	7,0	54,6	7,0	0,5	21,2
Ensino fundamental incompleto	36,8	33,4	3,5	1,3	55,7
Ensino fundamental completo	8,8	19,2	1,7	0,2	6,6

Figura 123: Tabela 2B

Fonte: POF 2017 - 2018

10.2.1.3 Tabela 3b

A tabela em questão apresenta dados resumidos acerca do índice de vulnerabilidade multidimensional não monetário, do efeito marginal e da contribuição para a soma dos efeitos marginais, segundo as dimensões selecionadas.

Tabela 3b - Índice de Vulnerabilidade Multidimensional não Monetário, efeito marginal e contribuição para a soma dos efeitos marginais, segundo as dimensões selecionadas - Brasil - período 2017-2018

Dimensões selecionadas	IVM-NM excluindo as perdas da dimensão	Efeito marginal (1)	Contribuição para a soma dos efeitos marginais (%)
Moradia	5,3	2,438	15,0
Serviços de utilidade pública	5,3	2,419	14,9
Saúde e alimentação	5,3	2,446	15,1
Educação	4,7	3,019	18,6
Acesso aos serviços financeiros e padrão de vida	4,5	3,167	19,5
Transporte e Lazer	5,0	2,719	16,8

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Pesquisas por Amostra de Domicílios, Pesquisa de Orçamentos Familiares 2017-2018.

(1) O efeito marginal da dimensão é dado pela diferença entre o valor do IVM-NM e o IVM-NM recalculado com a exclusão das perdas da dimensão.

Figura 124: Tabela 3B

Fonte: POF 2017 - 2018

10.2.1.4 Tabela 4b

A tabela em questão apresenta o índice de pobreza multidimensional não monetário, efeito marginal e contribuição para a soma dos efeitos marginais, segundo as dimensões selecionadas.

Tabela 4b - Índice de Pobreza Multidimensional não Monetário, efeito marginal e contribuição para a soma dos efeitos marginais, segundo as dimensões selecionadas - Brasil - período 2017-2018

Dimensões selecionadas	IPM-NM excluindo as perdas da dimensão	Efeito marginal (1)	Contribuição para a soma dos efeitos marginais (%)
Moradia	1,1	1,214	14,7
Serviços de utilidade pública	1,0	1,297	15,8
Saúde e alimentação	0,9	1,400	17,0
Educação	0,9	1,446	17,6
Acesso aos serviços financeiros e padrão de vida	0,7	1,578	19,2
Transporte e Lazer	1,0	1,294	15,7

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Pesquisas por Amostra de Domicílios, Pesquisa de Orçamentos Familiares 2017-2018.

(1) O efeito marginal da dimensão é dado pela diferença entre o valor do IPM-NM e o IPM-NM recalculado com exclusão das perdas da dimensão.

Figura 125: Tabela 4B

Fonte: POF 2017 - 2018

10.2.1.5 Tabela 5b

A tabela em questão apresenta dados acerca da proporção de pessoas das famílias residentes, da proporção de pessoas com algum grau de vulnerabilidade, do índice de Vulnerabilidade Multidimensional não Monetário e contribuição para o total dos efeitos marginais no Índice de Vulnerabilidade Multidimensional não Monetário, por tipo de dimensão, segundo as unidades de federação.

Tabela 5b - Proporção de pessoas das famílias residentes, proporção de pessoas com algum grau de Vulnerabilidade, IVM-NM e contribuição para o total dos efeitos marginais no IVM-NM, por tipo de dimensão, segundo as Unidades da Federação - período 2017-2018

Unidades da Federação	Proporção de pessoas das famílias residentes (%)	Proporção de pessoas com algum grau de vulnerabilidade (%)	IVM-NM	Contribuição para o total dos efeitos marginais no IVM-NM, por tipo de dimensão (%)					
				Moradia	Acesso aos serviços de utilidade pública	Saúde e alimentação	Educação	Acesso a serviços financeiros e pedágio de vida	Transporte e lazer
Brasil	100,0	63,8	7,7	15,0	14,9	15,1	18,6	19,5	16,8
Rondônia	0,8	84,8	10,2	13,5	24,5	9,5	18,9	17,2	16,4
Acre	0,4	89,0	15,1	15,2	19,5	14,6	17,9	17,2	15,7
Amazonas	1,9	83,6	12,8	16,5	17,5	14,4	15,9	19,8	16,0
Roraima	0,2	72,3	8,4	17,8	19,4	11,2	17,7	19,3	14,6
Pará	4,1	89,1	15,7	15,0	19,5	14,9	16,9	19,4	14,3
Amapá	0,4	88,5	13,5	16,1	21,3	14,9	16,0	17,4	14,5
Tocantins	0,7	80,7	9,7	15,3	16,1	12,6	17,9	20,2	17,9
Maranhão	3,4	93,3	17,4	15,1	18,0	15,7	17,6	19,2	14,4
Piauí	1,6	85,2	12,4	14,2	19,2	10,5	19,5	19,9	16,6
Ceará	4,4	78,9	10,1	15,3	16,3	13,7	19,7	21,5	13,5
Rio Grande do Norte	1,7	81,9	11,7	14,6	16,4	16,6	19,4	19,2	13,8
Paraíba	1,9	81,2	12,1	14,6	17,0	13,6	19,4	20,0	15,4
Pernambuco	4,5	81,4	11,9	15,1	17,4	15,5	17,9	19,6	14,5
Alagoas	1,6	85,3	13,1	15,1	17,9	15,0	19,7	20,3	12,0
Sergipe	1,1	76,4	10,0	14,7	14,1	16,2	19,8	20,4	14,8
Bahia	7,1	79,9	11,3	13,5	15,9	15,2	18,8	20,0	16,6
Minas Gerais	10,1	58,1	5,6	13,9	11,1	15,0	20,4	21,0	18,6
Espírito Santo	1,9	58,8	5,9	15,7	13,2	14,4	19,1	20,4	17,1
Rio de Janeiro	8,3	59,9	6,8	16,8	12,7	14,7	17,2	19,2	19,5
São Paulo	21,9	45,7	3,9	16,5	7,4	17,3	19,0	19,2	20,7
Paraná	5,5	45,7	3,8	14,6	13,3	13,0	21,0	20,2	17,9
Santa Catarina	3,4	40,0	2,6	13,6	16,5	14,0	21,0	19,3	15,6
Rio Grande do Sul	5,5	53,6	4,7	15,3	14,5	18,5	19,2	17,1	15,4
Mato Grosso do Sul	1,3	69,0	6,7	14,9	13,9	13,7	20,2	20,4	16,9
Mato Grosso	1,6	74,7	7,9	14,2	18,2	11,5	18,5	18,0	19,5
Goiás	3,3	69,2	8,1	14,1	15,3	15,0	17,8	17,4	20,4

Figura 126: Tabela 5B

Fonte: POF 2017 - 2018

10.2.1.6 Tabela 6b

A tabela em questão apresenta dados acerca da proporção de pessoas das famílias residentes, da proporção de pessoas com algum grau de pobreza, do Índice de Pobreza Multidimensional não Monetário e contribuição para o total dos efeitos marginais no Índice de Pobreza Multidimensional não Monetário, por tipo de dimensão, segundo as unidades da federação.

Tabela 6b - Proporção de pessoas das famílias residentes, proporção de pessoas com algum grau de pobreza, IPM-NM e contribuição para o total dos efeitos marginais no IPM-NM, por tipo de dimensão, segundo as Unidades da Federação - período 2017-2018

Unidades da Federação	Proporção de pessoas das famílias residentes (%)	Proporção de pessoas com algum grau de pobreza (%)	IPM-NM	Contribuição para o total dos efeitos marginais no IPM-NM, por tipo de dimensão (%)					
				Moradia	Acesso aos serviços de utilidade pública	Saúde e alimentação	Educação	Acesso a serviços financeiros e padrão de vida	Transporte e lazer
Brasil	100,0	22,3	2,3	14,7	15,8	17,0	17,6	19,2	15,7
Rondônia	0,8	28,7	2,6	13,9	19,5	14,7	17,7	17,1	17,1
Acre	0,4	48,4	6,4	14,7	19,1	15,7	18,2	17,3	14,9
Amazonas	1,9	40,7	4,9	15,9	17,4	15,8	16,4	18,9	15,5
Roraima	0,2	23,8	2,3	17,8	18,6	13,3	17,5	16,8	16,0
Pará	4,1	52,0	6,5	14,7	18,6	16,5	17,1	19,2	13,9
Amapá	0,4	43,7	4,9	14,7	19,7	16,8	16,6	17,8	14,4
Tocantins	0,7	28,0	2,7	14,6	15,4	15,3	17,3	19,8	17,6
Maranhão	3,4	58,1	7,7	14,7	17,4	17,2	17,2	18,7	14,8
Piauí	1,6	39,1	4,3	13,8	18,6	13,7	18,3	19,9	15,6
Ceará	4,4	30,9	3,0	14,7	16,6	15,6	18,2	20,5	14,5
Rio Grande do Norte	1,7	36,6	4,0	14,2	14,9	18,9	18,7	18,9	14,5
Paraíba	1,9	38,4	4,2	14,8	16,7	15,8	18,2	19,1	15,6
Pernambuco	4,5	36,6	4,1	14,7	16,6	17,7	17,0	19,4	14,8
Alagoas	1,6	43,6	4,7	15,2	17,3	17,0	17,9	19,7	12,9
Sergipe	1,1	29,9	3,1	14,2	14,3	18,2	18,7	19,5	15,1
Bahia	7,1	34,6	3,9	13,2	16,9	17,1	18,0	19,5	15,3
Minas Gerais	10,1	14,3	1,1	14,3	13,2	17,0	18,4	20,0	17,1
Espírito Santo	1,9	15,6	1,3	15,3	12,9	17,5	18,1	19,6	16,5
Rio de Janeiro	8,3	19,0	2,1	16,2	13,3	16,7	16,6	19,2	18,0
São Paulo	21,9	9,2	0,7	16,9	9,0	18,8	18,4	18,5	18,4
Paraná	5,5	8,7	0,6	14,4	13,1	16,5	18,8	19,9	17,3
Santa Catarina	3,4	5,1	0,3	14,2	14,6	18,1	17,3	20,0	15,8
Rio Grande do Sul	5,5	11,4	0,9	14,8	14,2	20,3	16,6	18,6	15,5
Mato Grosso do Sul	1,3	17,5	1,2	15,6	13,4	17,1	17,7	19,2	17,0
Mato Grosso	1,6	21,0	1,8	14,2	16,3	16,4	17,1	17,8	18,2
Goiás	3,3	23,0	2,2	13,7	13,9	18,2	16,9	18,3	18,9

Figura 127: Tabela 6B

Fonte: POF 2017 - 2018

10.2.1.7 Tabela 7b

A tabela em questão apresenta a proporção de pessoas das famílias residentes, do Índice de Pobreza Multidimensional com Componente Relativo e contribuições para o Índice de Pobreza Multidimensional com Componente Relativo do Brasil, segundo os condicionantes e subgrupos selecionados.

Tabela 7b - Proporção de pessoas das famílias residentes, IPM-CR e contribuições para o IPM-CR do Brasil, segundo os condicionantes e subgrupos selecionados - período 2017-2018

Condicionantes e subgrupos selecionados	Proporção de pessoas das famílias residentes (%)	IPM-CR	Contribuição para o IPM-CR do Brasil	Contribuição para o IPM-CR do Brasil (%)
Localização geográfica do domicílio				
Brasil	100,0	12,0	12,0	100,0
Urbano	85,3	10,6	9,0	75,1
Rural	14,7	20,4	3,0	24,9
Grandes Regiões				
Norte	8,6	18,3	1,6	13,0
Nordeste	27,3	16,7	4,6	37,9
Sudeste	42,2	9,1	3,9	32,0
Sul	14,3	8,0	1,1	9,5
Centro-Oeste	7,7	11,9	0,9	7,6
Pessoa de referência				
Composição demográfica				
Até 24 anos	3,1	12,9	0,4	3,3
25 a 49 anos	52,5	12,1	6,3	52,5
50 a 64 anos	28,9	12,1	3,5	29,0
65 anos ou mais	15,5	11,8	1,8	15,1
Cor ou raça				
Brancos	41,4	8,9	3,7	30,4
Pretos e pardos	57,2	14,4	8,2	68,4
Sexo				
Homem	59,7	11,4	6,8	56,5
Mulher	40,3	13,0	5,2	43,5
Nível de instrução				
Sem instrução	7,0	21,3	1,5	12,4
Ensino fundamental incompleto	36,8	15,7	5,8	48,0
Ensino fundamental completo	8,8	11,6	1,0	8,4
Ensino médio incompleto	5,0	12,1	0,6	5,0

Figura 128: Tabela 7B

Fonte: POF 2017 - 2018

10.2.1.8 Tabela 8b

A tabela em questão apresenta o índice de pobreza multidimensional, com componente relativo, efeito marginal e contribuição para a soma dos efeitos marginais, segundo as dimensões selecionadas.

Tabela 8b - Índice de Pobreza Multidimensional com Componente Relativo, efeito marginal e contribuição para a soma dos efeitos marginais, segundo as dimensões selecionadas - Brasil - período 2017-2018

Dimensões selecionadas	IPM-CR excluindo as perdas da dimensão	Efeito marginal (1)	Contribuição para a soma dos efeitos marginais (%)
Moradia	10,9	1,143	15,4
Serviços de utilidade pública	11,0	1,069	14,4
Saúde e alimentação	11,0	1,065	14,4
Educação	10,6	1,416	19,1
Acesso aos serviços financeiros e padrão de vida	10,6	1,431	19,3
Transporte e Lazer	10,8	1,281	17,3

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Pesquisas por Amostra de Domicílios, Pesquisa de Orçamentos Familiares 2017-2018.

(1) O efeito marginal da dimensão é dado pela diferença entre o valor do IPM-CR e o IPM-CR recalculado com a exclusão das perdas da dimensão.

Figura 129: Tabela 8B

Fonte: POF 2017 - 2018

10.2.1.9 Tabela 9b

A tabela em questão apresenta a proporção de pessoas das famílias residentes, o Índice de Pobreza Multidimensional com Componente Relativo e contribuição para o total dos efeitos marginais do Índice de Pobreza Multidimensional com Componente Relativo, por tipo de dimensão, segundo as unidades da federação.

Tabela 9b - Proporção de pessoas das famílias residentes, IPM-CR e contribuição para o total dos efeitos marginais no IPM-CR, por tipo de dimensão, segundo as Unidades da Federação - período 2017-2018

Unidades da Federação	Proporção de pessoas das famílias residentes (%)	IPM-CR	Contribuição para o total dos efeitos marginais no IPM-CR, por tipo de dimensão (%)					
			Moradia	Acesso aos serviços de utilidade pública	Saúde e alimentação	Educação	Acesso a serviços financeiros e padrão de vida	Transporte e lazer
Brasil	100,0	12,0	15,4	14,4	14,4	19,1	19,3	17,3
Rondônia	0,8	15,0	13,3	25,9	9,2	18,6	16,7	16,3
Acre	0,4	19,8	15,0	20,0	14,5	17,8	16,9	15,7
Amazonas	1,9	17,5	16,4	17,7	14,0	15,8	19,9	16,1
Roraima	0,2	13,0	18,3	19,7	10,4	17,8	19,6	14,1
Pará	4,1	20,2	14,8	19,8	15,1	16,8	19,4	14,1
Amapá	0,4	18,2	16,1	22,5	14,6	15,4	17,1	14,3
Tocantins	0,7	14,5	15,5	16,1	12,4	17,9	20,2	18,0
Maranhão	3,4	22,0	15,0	18,1	15,7	17,6	19,3	14,4
Piauí	1,6	17,1	14,3	19,4	10,3	19,5	20,0	16,5
Ceará	4,4	14,7	15,5	16,1	13,3	20,0	21,5	13,4
Rio Grande do Norte	1,7	16,3	14,6	16,6	16,5	19,5	19,0	13,8
Paraíba	1,9	16,6	14,6	16,9	13,5	19,5	20,1	15,5
Pernambuco	4,5	16,4	15,2	17,6	15,3	17,7	19,5	14,6
Alagoas	1,6	17,6	15,1	18,3	14,9	19,5	20,3	11,9
Sergipe	1,1	14,5	14,9	13,9	16,0	20,0	20,3	15,0
Bahia	7,1	15,9	13,4	15,8	15,1	18,7	20,0	16,9
Minas Gerais	10,1	10,0	14,1	10,2	13,7	21,9	20,9	19,2
Espírito Santo	1,9	10,2	16,6	12,4	13,1	20,0	20,5	17,4
Rio de Janeiro	8,3	11,2	18,0	11,9	13,5	17,2	19,1	20,2
São Paulo	21,9	7,9	17,1	6,3	15,4	20,4	18,7	22,1
Paraná	5,5	7,8	15,7	12,8	11,2	22,6	19,9	17,6
Santa Catarina	3,4	6,7	15,1	17,3	11,5	23,8	17,6	14,7
Rio Grande do Sul	5,4	9,0	16,6	14,2	17,7	20,2	16,2	15,1
Mato Grosso do Sul	1,3	11,2	15,1	13,8	12,9	20,4	20,5	17,3
Mato Grosso	1,6	12,6	14,3	18,6	10,6	18,5	17,7	20,3
Goiás	3,3	12,6	14,5	15,4	14,1	17,6	16,8	21,7

Figura 130: Tabela 9B

Fonte: POF 2017 - 2018

10.2.2 Tabelas 2008 - 2009

10.2.2.1 Tabela 1b

A tabela em questão apresenta dados resumidos acerca da proporção de pessoas das famílias residentes, da proporção de pessoas com algum grau de vulnerabilidade, do índice de vulnerabilidade multidimensional não monetário e contribuições para o índice de vulnerabilidade multidimensional não monetário do Brasil segundo os condicionantes e subgrupos selecionados.

10.2.2.2 Tabela 2b

A tabela em questão apresenta dados resumidos acerca da proporção de pessoas das famílias residentes, da proporção de pessoas com algum grau de pobreza, do Índice de Pobreza Multidimensional não Monetário do Brasil segundo os condicionantes e subgrupos selecionados.

10.2.2.3 Tabela 3b

A tabela em questão apresenta dados resumidos acerca do índice de vulnerabilidade multidimensional não monetário, do efeito marginal e da contribuição para a soma dos efeitos marginais, segundo as dimensões selecionadas.

10.2.2.4 Tabela 4b

A tabela em questão apresenta o índice de pobreza multidimensional não monetário, efeito marginal e contribuição para a soma dos efeitos marginais, segundo as dimensões selecionadas.

10.2.2.5 Tabela 5b

A tabela em questão apresenta dados acerca da proporção de pessoas das famílias residentes, da proporção de pessoas com algum grau de vulnerabilidade, do índice de Vulnerabilidade Multidimensional não Monetário e contribuição para o total dos efeitos marginais no Índice de Vulnerabilidade Multidimensional não Monetário, por tipo de dimensão, segundo as unidades de federação.

10.2.2.6 Tabela 6b

A tabela em questão apresenta dados acerca da proporção de pessoas das famílias residentes, da proporção de pessoas com algum grau de pobreza, do Índice de Pobreza Multidimensional não Monetário e contribuição para o total dos efeitos marginais no Índice de Pobreza Multidimensional não Monetário, por tipo de dimensão, segundo as unidades da federação.

10.2.2.7 Tabela 7b

A tabela em questão apresenta a proporção de pessoas das famílias residentes, do Índice de Pobreza Multidimensional com Componente Relativo e contribuições para o

Índice de Pobreza Multidimensional com Componente Relativo do Brasil, segundo os condicionantes e subgrupos selecionados.

10.2.2.8 Tabela 8b

A tabela em questão apresenta o índice de pobreza multidimensional, com componente relativo, efeito marginal e contribuição para a soma dos efeitos marginais, segundo as dimensões selecionadas.

10.2.2.9 Tabela 9b

A tabela em questão apresenta a proporção de pessoas das famílias residentes, o Índice de Pobreza Multidimensional com Componente Relativo e contribuição para o total dos efeitos marginais do Índice de Pobreza Multidimensional com Componente Relativo, por tipo de dimensão, segundo as unidades da federação.

Fonte para maiores esclarecimentos: [IBGE - Orçamento familiar: pesquisa de orçamentos familiares.](#)

10.3. RAIS e CAGED

Microdados da RAIS (Relação Anual de Informações Sociais) e do CAGED (Cadastro Geral de Empregados e Desempregados) são conjuntos de informações detalhadas sobre o emprego formal no Brasil. A RAIS coleta dados anuais, incluindo informações sobre empregados, utilizados para estatísticas e controle de benefícios. O CAGED registra admissões e demissões de trabalhadores com carteira assinada, sendo essencial para análises do mercado de trabalho e políticas públicas. Ambos fornecem uma visão detalhada da dinâmica do emprego no país. Esse critério pode ser utilizado para especular o potencial de consumo em cada região do Brasil. Os dados disponibilizados são anuais, começando em 1985 até 2016. Todos os estados brasileiros são contemplados nessa análise. Os critérios e colunas são diversos, voltados para caracterizar o trabalhador / desempregado brasileiro.

Selecione as variáveis: - Select one or more

- Faixa Hora Contrat (1994-16)
- Faixa Remun Dezem {SM}
- Faixa Remun Média {SM}
- Faixa Tempo Emprego
- Grau Instrução 2005-1985 (1985-05)
- Escolaridade após 2005 (2006-16)
- Qtd Hora Contr (1994-16)
- Idade (1994-16)
- Ind CEI Vinculado (1999-16)
- Ind Simples (2001-16)**
- Mês Admissão
- Mês Desligamento
- Mun Trab (2002-16)
- Município
- Nacionalidade
- Natureza Jurídica (1994-16)
- Ind Portador Defic (2007-16)
- Qtd Dias Afastamento (2002-16)
- Raça Cor (2006-16)
- Regiões Adm DF (1996-16)
- VI Remun Dezembro Nom (1999-16)
- VI Remun Dezembro {SM}
- VI Remun Média Nom (1999-16)
- VI Remun Média {SM}
- CNAE 2.0 Subclasse (2004-16)
- Sexo Trabalhador
- Tamanho Estabelecimento
- Tempo Emprego
- Tipo Admissão (1994-16)
- Tipo Estab1
- Tipo Estab2
- Tipo Defic (2007-16)
- Tipo Vínculo
- VI Rem Janeiro CC (2015-16)
- VI Rem Fevereiro CC (2015-16)
- VI Rem Março CC (2015-16)
- VI Rem Abril CC (2015-16)
- VI Rem Maio CC (2015-16)
- VI Rem Junho CC (2015-16)
- VI Rem Julho CC (2015-16)
- VI Rem Agosto CC (2015-16)
- VI Rem Setembro CC (2015-16)
- VI Rem Outubro CC (2015-16)
- VI Rem Novembro CC (2015-16)
- Ano Chegada Brasil (2016)

OK

Cancelar

Figura 131: Variáveis RAIS e CAGED

Fonte: RAIS e CAGED

10.4. Receita Federal Dados Abertos

Para o objetivo do projeto, dos dados disponibilizados pelo site do governo na seção da receita federal (Dados da Receita Federal), os dados que fazem mais sentido para a ingestão são aqueles referentes à distribuição de renda. Essa escolha se baseia exatamente na necessidade do cliente de examinar o mercado consumidor de alimentos ao longo do Brasil e avaliar em qual região há um maior potencial consumidor. No caso específico das bases, essa análise do mercado consumidor se basearia no parâmetro do poder aquisitivo percebido por meio da distribuição de renda.

Tabela distribuição-renda.csv:

- Linhas: 46350
- Colunas: 24

Tabela distribuição-renda-sócios.csv:

- Linhas: 46350
- Colunas: 24

Tabela distribuição-renda-sócios-exclusiva.csv:

- Linhas: 46350
- Colunas: 24

Fonte: <https://www.gov.br/receitafederal/pt-br/acesso-a-informacao/dados-abertos>

10.5. MEC

O Programa Universidade para Todos (Prouni) do Ministério da Educação do Brasil oferece bolsas de estudo em instituições de ensino superior privadas. Os dados abertos do Prouni incluem informações detalhadas sobre as bolsas concedidas, perfis dos beneficiários, instituições de ensino participantes e outros dados relevantes. Os conjuntos de dados abrangem os anos de 2016 a 2020 e estão disponíveis em formato CSV.

10.5.1 Descrição dos Conjuntos de Dados

Os dados do Prouni oferecem uma visão ampla sobre a distribuição de bolsas de estudo no ensino superior privado no Brasil. Esses dados são cruciais para compreender as tendências educacionais, demográficas e socioeconômicas. A seguir, segue a tabela de relevância dos dados do prouni.

Relevância	Nome do Conjunto	Tipo de Arquivo
Alta	Dados do Prouni 2016	.CSV
Alta	Dados do Prouni 2017	.CSV
Alta	Dados do Prouni 2018	.CSV
Alta	Dados do Prouni 2019	.CSV
Alta	Dados do Prouni 2020	.CSV

Tabela 01: Relevância dos dados do prouni

Fonte: RAIS e CAGED

10.5.2 Dados Relevantes para Análises Estratégicas

Distribuição Regional de Bolsas: Informações sobre a localização geográfica dos beneficiários, incluindo estado e município, podem indicar áreas com maior potencial de mercado e poder de compra.

Perfil Demográfico e Socioeconômico: Dados sobre sexo, raça, idade e tipo de bolsa (integral ou parcial) dos beneficiários fornecem insights sobre a diversidade e o perfil socioeconômico dos estudantes.

10.5.3 Potencial de Uso dos Dados

Os dados do Prouni são uma fonte valiosa para análises de mercado, desenvolvimento de políticas educacionais e estratégias de marketing. Eles podem ser utilizados para:

- Identificar regiões com alta demanda por educação superior e potencial de mercado para produtos e serviços relacionados;
- Compreender o perfil dos estudantes beneficiários para campanhas de marketing direcionadas;

- Avaliar as tendências do mercado educacional e adaptar estratégias de negócios de acordo.

10.5.4 Análises e Insights dos Dados do Prouni

A análise dos dados do Prouni revela várias tendências e padrões significativos que são cruciais para o entendimento do mercado educacional brasileiro e para o desenvolvimento de estratégias de marketing eficazes.

10.5.4.1 Análises e Insights dos Dados do Prouni

A distribuição desigual de bolsas entre as regiões e estados brasileiros indica áreas com maior concentração de estudantes beneficiados pelo Prouni.

Aplicação Estratégica: Identificar estados e regiões-chave para campanhas de marketing direcionadas, desenvolvimento de produtos específicos e parcerias com instituições locais.

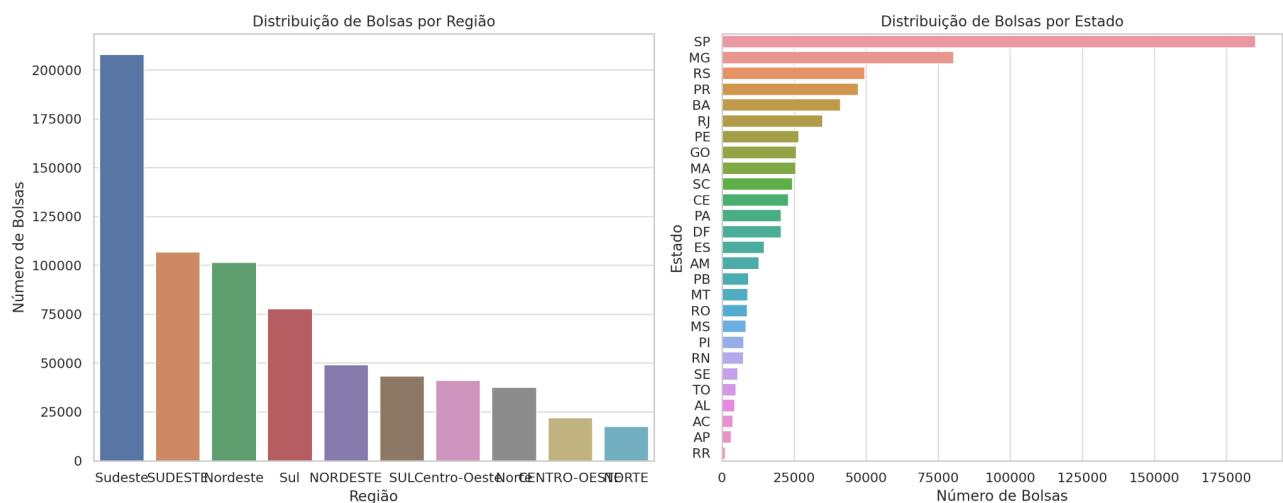


Figura 132: Gráfico de distribuição (Região e Estado)

Fonte: Prouni

10.5.4.2 Perfil dos Beneficiários

A análise do perfil dos beneficiários mostra a proporção de estudantes em cursos presenciais versus EAD, tipos de bolsa (integral ou parcial), distribuição racial e de gênero.

Aplicação Estratégica: Desenvolver estratégias de marketing inclusivas e diversificadas que atendam às necessidades de um público variado.

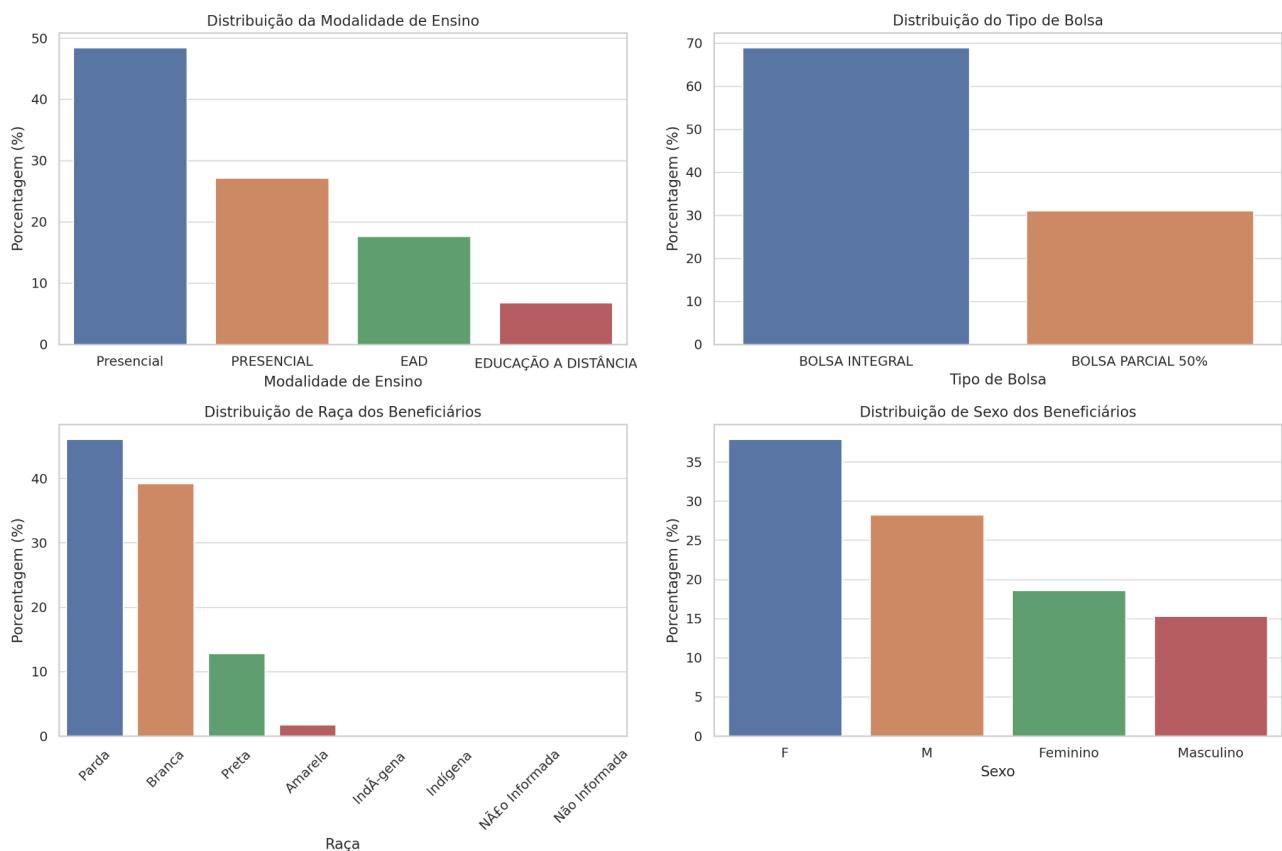


Figura 133: Gráfico de distribuição (Ensino, Bolsa, Raça e Sexo)

Fonte: Prouni

10.5.4.3 Descrição dos dados

Cada arquivo CSV contém várias colunas (ou índices), cada uma representando um tipo específico de informação. Aqui está uma descrição de alguns dos índices chave:

ANO_CONCESSAO: Ano em que a bolsa foi concedida.

CODIGO_EMECIES: Código de identificação da instituição de ensino superior.

NOMEIES: Nome da instituição de ensino superior.

MUNICÍPIO: Município da instituição de ensino.

TIPO_BOLSA: Tipo da bolsa concedida (integral ou parcial).

MODALIDADE_ENSINO: Modalidade do curso (presencial ou EAD).

NOME_CURSO: Nome do curso para o qual a bolsa foi concedida.

NOME_TURNO_CURSO: Turno do curso.

CPF_BENEFICIARIO: CPF do beneficiário (anonimizado).

SEXO_BENEFICIARIO: Sexo do beneficiário.

RACA_BENEFICIARIO: Raça declarada pelo beneficiário.

DATA_NASCIMENTO: Data de nascimento do beneficiário.

BENEFICIARIO_DEFICIENTE_FISICO: Indica se o beneficiário é deficiente físico.

REGIAO_BENEFICIARIO: Região do beneficiário.

UF_BENEFICIARIO: Unidade Federativa do beneficiário.

MUNICIPIO_BENEFICIARIO: Município de residência do beneficiário.

Os dados são extensos e oferecem uma visão abrangente do programa Prouni ao longo dos anos mencionados. Abaixo está a quantidade de registros (linhas) e variáveis (colunas) em cada ano:

2016: O dataset contém 239.262 registros e 15 colunas.

2017: O dataset contém 236.636 registros e 15 colunas.

2018: O dataset contém 241.032 registros e 15 colunas.

2019: O dataset contém 241.032 registros e 15 colunas.

2020: O dataset contém 166.830 registros e 17 colunas.

10.5.5 Considerações Finais

A análise detalhada dos dados do Prouni oferece uma visão abrangente do setor educacional privado no Brasil. As informações obtidas são indispensáveis para organizações que buscam compreender melhor o mercado educacional, desenvolver e implementar estratégias de go-to-market focadas e bem-sucedidas.

10.6. INEP

10.6.1 ANA - Avaliação Nacional de Alfabetização

A Avaliação Nacional de Alfabetização, também conhecida como ANA, é uma ferramenta de avaliação utilizada no contexto educacional brasileiro para medir o nível de alfabetização e letramento de crianças em séries do ensino fundamental. Para a educação, esses dados são utilizados para acompanhar o progresso dos estudantes e para a formulação de políticas públicas e a tomada de decisões em relação ao sistema educacional, direcionando recursos e esforços para aprimorar a qualidade.

Para o projeto, esse dado pode ser utilizado pensando na compreensão do mercado, como a compreensão do nível de educação dos consumidores pode ajudar a

adaptar estratégias de marketing e etiquetagem de produtos. Além disso, podemos cruzar as informações para compreender como a alfabetização numérica afeta as decisões de compra e consumo, e com isso ajudar na previsão de tendências de mercado e no desenvolvimento de estratégias de vendas mais eficientes. No site do INEP, os únicos dados disponibilizados são de 2014 e 2016 e esses dois são utilizados no projeto.

10.6.2 Censo da Educação Superior

O Censo da Educação Superior é uma pesquisa anual que tem como objetivo a coleta dos dados abrangentes sobre instituições de ensino superior, estudantes, cursos e programas acadêmicos em todo o território nacional. Esses dados são importantes para identificar tendências e preferências alimentares em diferentes grupos demográficos. Por exemplo, alunos de cursos relacionados à saúde podem preferir opções alimentares mais saudáveis. Além disso, é comprovado que pessoas formadas em algum curso da educação superior tem um maior poder de compra, então o setor alimentício se favorece com essas regiões. Foram selecionados os dados do IES, de 2009 até 2022.

10.6.3 ENCEJA

O Encceja, ou Exame Nacional para Certificação de Competências de Jovens e Adultos, é um exame brasileiro que certifica a conclusão do Ensino Fundamental ou Médio para jovens e adultos que não tiveram oportunidade de concluir seus estudos na idade apropriada. Para o projeto, esses dados podem ser utilizados para identificar tendências e preferências alimentares em diferentes grupos demográficos. Foi selecionado os dados de 3 anos diferentes, 2014, 2020 e 2022.

10.7. Open Datasus

O DATASUS disponibiliza 31 conjuntos de dados coletados e fornecidos pelo Ministério da Saúde, acessíveis através do link [DATASUS - Open DataSUS](#). Esses conjuntos de dados abrangem uma variedade de informações relacionadas à saúde da população brasileira, incluindo dados recentes e antigos. Nos próximos tópicos, exibe-se detalhes sobre esses conjuntos de dados, suas especificações e quais deles têm relevância para o projeto.

10.7.1 Descrição dos conjuntos

O DATASUS disponibiliza uma variedade de informações relacionadas à saúde, que podem ser úteis para entender os padrões de consumo de alimentos. No entanto,

nem todos esses dados são igualmente relevantes para o nosso projeto. A seguir, apresentamos uma tabela dos conjuntos de dados do DATASUS, classificando-os com sua relevância para a análise de consumo alimentício. A classificação foi feita considerando a relação direta ou indireta entre os dados e o comportamento de consumo alimentar. Além disso, indicamos os tipos de arquivo associados a cada conjunto de dados para facilitar a sua identificação e acesso.

10.7.2 Dados menos relevantes para análise de consumo alimentício

SISAGUA - Vigilância em Parâmetros Básicos e outras fontes relacionadas à qualidade da água: Embora a qualidade da água seja importante para a saúde pública, esses dados podem ter uma influência indireta no consumo alimentar.

SRAG - Banco de Dados de Síndrome Respiratória Aguda Grave: Se você já está usando os dados mais recentes da SRAG, essas versões anteriores podem não ser necessárias para a análise de consumo alimentar.

Febre Amarela em humanos e primatas não-humanos: A febre amarela é uma doença transmitida por mosquitos e não está diretamente relacionada ao consumo de alimentos.

10.7.3 Dados relevantes para análise de consumo alimentício

Notificações de Síndrome Gripal: Esses dados podem indicar surtos de doenças semelhantes à gripe em diferentes regiões, o que pode ser um indicativo de problemas de saúde na população. A ocorrência de surtos de doenças gripais pode afetar o comportamento do consumidor, levando a mudanças nos padrões de compra e consumo de alimentos, como maior demanda por alimentos saudáveis e suplementos vitamínicos.

Campanha Nacional de Vacinação contra Covid: O sucesso das campanhas de vacinação contra a COVID-19 em diferentes áreas pode influenciar a confiança da população na segurança de sair e consumir alimentos fora de casa, como em restaurantes. Uma alta taxa de vacinação pode levar a uma maior sensação de segurança e estimular o consumo em estabelecimentos alimentícios.

Registro de Ocupação Hospitalar COVID: Se os hospitais estiverem sobrecarregados, as pessoas podem optar por evitar locais com aglomerações, como restaurantes e bares, impactando o consumo nesses estabelecimentos.

SRAG 2021 a 2023 - Banco de Dados de Síndrome Respiratória Aguda Grave (incluindo dados da COVID-19): O aumento dos casos de SRAG, especialmente os

graves, pode resultar em medidas de isolamento e restrições de movimento, afetando o comportamento de consumo, como o aumento do uso de serviços de entrega de alimentos.

Sistema de Informação sobre Nascidos Vivos – Sinasc: O aumento na taxa de nascimentos pode indicar uma necessidade de planejamento de suprimentos alimentícios nas áreas onde a população está crescendo rapidamente (demanda por alimentos em escolas, creches e outras instituições.)

Sistema de Informação sobre Mortalidade – SIM: Os dados de mortalidade podem fornecer informações sobre as principais causas de morte em diferentes regiões, incluindo doenças relacionadas à dieta, como doenças cardiovasculares e diabetes.

Unidades Básicas de Saúde - UBS:

- **Acesso a Alimentos Saudáveis:** Quando as UBS estão ausentes ou insuficientes em uma região, as pessoas podem enfrentar dificuldades no acesso a cuidados de saúde e, indiretamente, a informações sobre práticas de alimentação saudável.
- **Desertos Alimentares:** A falta de UBS em uma área pode indicar uma possível carência de infraestrutura de saúde e, por extensão, a presença de desertos alimentares. Os desertos alimentares são áreas onde os residentes têm dificuldade em acessar alimentos frescos e saudáveis devido à falta de supermercados ou lojas que ofereçam esses produtos.

CNES - Cadastro Nacional de Estabelecimentos de Saúde: Esses dados podem fornecer informações sobre a disponibilidade de serviços de saúde em diferentes áreas, que estão diretamente ligados ao acesso da população a cuidados de saúde.

10.7.4 Dados carregados na AWS S3

Os dados provenientes do Sistema Único de Saúde (SUS), constituem uma fonte rica sobre a saúde no Brasil. Esses conjuntos abrangem informações hospitalares, natalidade e vigilância epidemiológica, oferecendo uma perspectiva abrangente ao longo do tempo. Abordaremos detalhes de cada conjunto para destacar suas características.

10.7.4.1 Hospitais e Leitos - 2007 a 2023

Este conjunto de dados fornece informações sobre estabelecimentos hospitalares, leitos gerais e complementares, incluindo dados de contato, como endereço, telefone e

e-mail. Os arquivos são disponibilizados mensal e anualmente. É importante destacar que, para garantir a conformidade com a Lei de Acesso à Informação e a Lei Geral de Proteção de Dados, os dados são divulgados de forma agregada, preservando o sigilo das informações pessoais.

Limitações: As limitações podem surgir devido à natureza da coleta de dados, dependendo das informações fornecidas pelos gestores locais de saúde.

10.7.4.2 Sistema de Informação sobre Nascidos Vivos – Sinasc - 1996 - 2023

O Sistema de Informações sobre Nascidos Vivos (Sinasc) foi estabelecido em 1990 para coletar dados sobre nascimentos em todo o território nacional. Os registros incluem informações socioeconômicas, local de residência, ocorrência, anomalias congênitas, parto e pré-natal. Esse sistema contribui para o conhecimento da saúde da população e avaliação de políticas relacionadas à saúde materno-infantil.

10.7.4.3 SRAG - Banco de Dados de Síndrome Respiratória Aguda Grave - incluindo dados da COVID-19

Este banco de dados contém informações epidemiológicas sobre a Síndrome Respiratória Aguda Grave (SRAG) no Brasil desde 2009, incorporando dados da COVID-19 a partir de 2020. A vigilância é realizada pela Secretaria de Vigilância em Saúde e o sistema oficial para registro é o Sistema de Informação da Vigilância Epidemiológica da Gripe (SIVEP-Gripe). Os dados são disponibilizados semanalmente, sujeitos a alterações decorrentes de investigações e correções de erros, com anonimização em conformidade com a Lei 13.709/2018.

Limitações: Os dados podem estar sujeitos a alterações devido a investigações em curso. A anonimização é aplicada para cumprir as regulamentações de privacidade.

Observação: Pelo grande número de glossário para esse conjunto de dados, pode-se acessar o arquivo completo [Dicionário de Dados SRAG Hospitalizado - 19.09.2022](#)

10.7.4.4 Registro de Ocupação Hospitalar COVID-19

Implementado devido à pandemia, este banco de dados fornece informações sobre a ocupação de leitos clínicos e de UTI SUS destinados a pacientes com casos suspeitos ou confirmados de COVID-19. A coleta é realizada por meio do Sistema ESUS Notifica-Módulo Internações SUS, iniciado em abril de 2020. A partir de 2022, novos

campos foram acrescentados para descrever a ocupação dos leitos. Não se oferece glossário dos dados.

Limitações: Devido ao grande volume de registros, alguns estados têm mais de um milhão de entradas, registros anteriores a 2022 podem não conter os novos campos preenchidos.

10.8. Códigos postais mundiais

O B4a (BackForApp), cloud de dados, disponibilizou uma base de dados abrangente com informações sobre códigos postais de todos os países do mundo. Link: <https://www.back4app.com/database/back4app/zip-codes-all-countries-in-the-world>.

Os dados foram extraídos do site <http://www.geonames.org/>, no qual cada conjunto de dados representa um país e inclui as seguintes informações:

- Número do Código Postal em formato de string;
- Nome do Local em formato de string;
- Geolocalização em formato de ponto geográfico;
- Precisão da Geolocalização em formato de string;
- Código Administrativo em formato de string;
- Nome Administrativo em formato de string.

O conjunto de dados disponibilizado abrange os códigos postais de 97 países diferentes. No entanto, em atendimento a solicitação específica de nosso parceiro, estaremos concentrando nossos esforços na análise exploratória dos dados referentes exclusivamente ao Brasil. Dessa forma, os dados relativos aos demais países serão desconsiderados durante este processo de análise.

Esta abordagem visa encontrar insights precisos e valiosos que atendam às necessidades e expectativas específicas delineadas em nosso projeto. Essa estratégia permite uma análise mais aprofundada e personalizada dos dados, maximizando o impacto de nossos esforços de análise exploratória no contexto brasileiro.

No conjunto de dados há um total de 11 colunas distintas, cada uma oferecendo informações específicas sobre os códigos postais.

10.9. Estudo de Correlações

10.9.1 Correlações entre os dados - Views

10.9.1.1 - 1º correlação: Análise de consumo de mercado - POF e CNPJ

Pela POF mostrar o consumo das famílias de determinados produtos, é possível verificar quais empresas vendem esse tipo de produto.

10.9.1.2 - 2ª correlação: CNPJ e Data Sus

Ao ter acesso aos dados do DATASUS, temos a capacidade de disponibilizar uma ampla gama de informações relacionadas à saúde, fornecendo insights valiosos para a compreensão dos padrões de consumo alimentar. Ao vincular esses dados aos CNPJs, torna-se possível discernir diversos padrões de consumo no mercado e identificar tendências. Isso estabelece uma conexão essencial entre a saúde da população e os hábitos alimentares predominantes. A análise desses dados não só proporciona uma compreensão aprofundada da saúde em si, mas também lança luz sobre os diversos fatores que influenciam as escolhas alimentares.

10.9.1.3 - 3ª correlação: Áreas com maior demanda - POF E CNPJ

Os dados da POF apresentam as regiões em que determinado produto tem maior demanda. Esse dado ajuda a compreender as empresas que são mais fortes naquele segmento em dada região. Além disso, ajuda a entender sobre padrões de consumo em diferentes regiões do país.

10.9.1.4 - 4ª correlação: Expansão empresarial - POF E CNPJ

Empresas podem usar dados da POF para identificar áreas com alta demanda por seus produtos. O cruzamento desses dados com as informações do CNPJ pode ajudar a identificar regiões com pouca concorrência, que podem ser locais potenciais para expansão.

10.9.1.5 - 5ª Correlação: "Localização do CNPJ" com "Dados de Rendimento e Despesas" da POF

Identificar áreas com maior potencial de mercado com base no poder aquisitivo e despesas com alimentos.

10.9.1.6 - 6ª Correlação: Correlação com Dados Econômicos Mais Amplos (Dados da POF e CNPJ)

Tendências Econômicas Regionais: Correlacionar a saúde econômica de uma região (renda, desemprego) com a presença de empresas do setor alimentício.

10.9.1.7 - 7ª Categorias de Produtos Mais Vendidas por Estado

Identificar quais categorias de produtos são mais vendidas em cada estado, correlacionando vendas, categorias de produtos e informações de CNPJ.

Fornece insights sobre preferências de consumo em diferentes estados, ajudando as empresas a direcionar sua oferta de produtos.

Pode ser usado para estratégias de marketing regionalizadas e para otimizar a gestão de estoque.

10.9.1.8 - 8ª Empresas com Maior Valor Total de Vendas

Determinar quais empresas têm o maior valor total de vendas, agregando dados de vendas e informações empresariais. Identifica líderes de mercado e tendências de vendas em diferentes setores. Útil para análise de concorrência e para identificar potenciais parceiros de negócios ou aquisições.

10.9.1.9 - 9ª Cidade com Maior Número de Estabelecimentos

Determinar qual cidade tem o maior número de estabelecimentos comerciais, utilizando dados de CNPJ. Revela centros comerciais com alta densidade de negócios, indicando áreas potencialmente lucrativas para novos investimentos. Pode influenciar decisões de expansão empresarial e planejamento urbano.

10.9.1.10 - 10ª Empresas com o Mesmo CNAE Fiscal Principal

Agrupar empresas com o mesmo CNAE para identificar setores com maior ou menor concentração de negócios. Fornece uma visão do nível de concorrência em diferentes setores. Pode ser usado para identificar oportunidades de mercado em setores menos saturados.

10.9.1.11 - 11ª Cidade com Maior Número de Vendas

Identificar a cidade com o maior número de vendas, utilizando dados de CNPJ e registros de vendas. Destaca áreas com alta atividade comercial, útil para estratégias de marketing e expansão de negócios. Pode indicar regiões com alto potencial de consumo para determinados produtos ou serviços.

10.9.1.12 - 12ª Correlação entre Número de Estabelecimentos e Valor Total de Vendas

Explorar a relação entre o número de estabelecimentos de uma empresa e seu valor total de vendas. Ajuda a entender se uma maior presença física se traduz em maiores vendas. Importante para estratégias de expansão física e investimentos em novas localidades.

10.9.1.13 - 13ª Distribuição de Vendas por Categoria de Produtos em Cada Estado

Analisa como as vendas de diferentes categorias de produtos se distribuem por estado. Oferece insights sobre preferências regionais, ajudando na segmentação de mercado e estratégias de publicidade localizadas. Pode auxiliar na tomada de decisões sobre quais produtos estocar em diferentes localidades.

10.9.1.14 - 14ª Empresas com Maior Quantidade de Estabelecimentos

Identificar empresas com o maior número de estabelecimentos, utilizando dados de estabelecimentos e CNPJ. Revela quais empresas têm maior presença física, indicando potencial de mercado e força de marca. Útil para análises de expansão de rede e compreensão da distribuição geográfica de negócios.

10.9.1.15 - 15ª Empresas com Maior Quantidade de Vendas em um Período Específico

Determinar quais empresas tiveram o maior volume de vendas em um período específico. Permite entender quais empresas tiveram melhor desempenho em determinadas condições de mercado. Importante para análise de tendências de mercado e planejamento estratégico.

10.10. Views

No Amazon Redshift, as views são estruturas que permitem aos usuários criar consultas predefinidas e reutilizáveis. Elas são essencialmente consultas salvas como objetos no banco de dados, proporcionando uma maneira de organizar e simplificar consultas complexas.

A `view` `alimentos_mais_consumidos_por_estado` agrupa dados da tabela `consumo_alimentar_pof`, contabilizando a quantidade de alimentos consumidos e o total de gramas por estado. A ordenação é feita pela quantidade de alimentos consumidos em ordem decrescente.

A `view` `categorias_mais_vendidas_por_estado` conecta informações de três tabelas: `cnpjs` (deve ser aplicado nos 5 arquivos), `dados_sale`, e `dados_categoria`. Esta apresenta o número de vendas por categoria, agrupadas por estado (`sigla_uf`), com ordenação descendente pelo número de vendas. Esta view calcula a soma de vendas para cada CNPJ e nome fantasia da tabela `cnpjs` (deve ser aplicado nos 5 arquivos) e `dados_sale`. A condição WHERE filtra as vendas para o intervalo de datas entre janeiro de 2021 e dezembro de 2022. Os resultados são ordenados pela soma de vendas em ordem decrescente.

A view `potencial_expansao_empresarial` é projetada para identificar áreas com potencial para expansão empresarial no setor de supermercados. Ela correlaciona o consumo de alimentos em gramas com o número de supermercados existentes em cada unidade federativa. Esta análise é feita cruzando dados da tabela `consumo_alimentar_pof` com informações de empresas da tabela `cnpj1`, filtrando especificamente pelo CNAE de supermercados. Esta view identifica as empresas com o maior valor total de vendas. A análise é realizada ao agregar os dados de vendas de cada empresa e somar os valores totais, proporcionando uma visão clara das empresas com maior volume de vendas.

A view `empresas_maior_quantidade_estabelecimento` identifica as empresas com a maior quantidade de estabelecimentos. Utiliza dados da tabela `cnpj1` para contar e listar as empresas com base no número de estabelecimentos que possuem, proporcionando uma visão clara das empresas com a maior presença física.

A view `correlacao_economica_setor_alimenticio` analisa a correlação entre a saúde econômica das regiões (como representada pela renda média) e a presença de empresas do setor alimentício. Utiliza dados do consumo alimentar da POF e informações do CNPJ para investigar essa relação.

A view `correlacao_economica_setor_alimenticio` tem como objetivo identificar os alimentos mais consumidos por domicílios da classe A, onde a renda bruta mensal total é superior a R\$ 22.000,00. A consulta utiliza dados da tabela `consumo_alimentar_pof` e outros rendimentos.

A view `quantidade_cnpjs_por_cnae` tem como objetivo contabilizar a quantidade de CNPJs por categoria de CNAE fiscal principal. A consulta utiliza dados de várias tabelas, como `cnpj1`, `cnpj2`, `cnpj3`, `cnpj4`, e `cnpj5`.

A view `quantidade_estabelecimentos_por_estado` visa contabilizar a quantidade total de estabelecimentos por estado. A consulta utiliza dados de várias tabelas, como `cnpj1`, `cnpj2`, `cnpj3`, `cnpj4`, e `cnpj5`.

A view `top_3_cnaes_por_estado` identifica os três CNAEs (Classificação Nacional de Atividades Econômicas) principais com a maior quantidade de

estabelecimentos por estado. A consulta utiliza dados de várias tabelas, como cnpj1, cnpj2, cnpj3, cnpj4, e cnpj5.

11. Armazenamento, Organização e Acesso aos Dados

11.1. Data Lake

"Data Lake" é um conceito que se refere a um repositório de dados que pode armazenar uma grande quantidade de dados brutos, estruturados e não estruturados, em sua forma nativa, até que sejam necessários para análise. É uma abordagem que permite armazenar dados em sua forma bruta, sem a necessidade de estruturá-los previamente, oferecendo flexibilidade na análise posterior.

11.1.2 Importância do Data Lake

Armazenamento Versátil: Data Lakes, como o S3 da AWS, oferecem um local para armazenar dados em diversos formatos, como texto, imagens, vídeos e outros, sem a necessidade imediata de organização. Isso permite que as organizações armazenem uma grande variedade de dados de maneira econômica.

Análise Pós-Fato: Ao armazenar dados brutos, as organizações podem realizar análises mais profundas e avançadas quando surgem novas questões ou métodos analíticos. Isso contrasta com abordagens mais tradicionais que requerem estruturação antecipada dos dados.

Integração com Ferramentas de Big Data: Data Lakes são frequentemente integrados a ecossistemas de Big Data, permitindo o processamento eficiente de grandes volumes de dados usando ferramentas como Apache Spark ou Apache Hive.

11.1.3 Aplicação data lake no projeto

No contexto do projeto, a opção por utilizar o Amazon S3 (Amazon Simple Storage Service) como Data Lake reflete uma estratégia que se fundamenta na capacidade escalável de armazenamento, sendo ele um serviço de armazenamento na nuvem escalável, projetado para armazenar e recuperar qualquer quantidade de dados a qualquer momento. Essa abordagem permite armazenar dados em sua forma bruta, sem a necessidade imediata de estruturação, alinhando-se aos princípios de flexibilidade e eficiência característicos da computação em nuvem.

Essa escolha proporciona a flexibilidade necessária para futuras análises, permitindo uma integração com outras ferramentas e serviços dentro do ecossistema da AWS. Dessa

maneira, ao utilizar o S3 como Data Lake, não apenas atende às necessidades imediatas de armazenamento, mas também estabelece uma base sólida para a exploração e extração de insights a partir dos dados armazenados.

11.1.4 Quantidade de dados armazenados

A gestão dos dados é um aspecto crucial para análises. A quantificação precisa da quantidade de dados armazenados no S3 oferece insights valiosos sobre o escopo e a complexidade do repositório. Abaixo, detalhamos a distribuição por buckets, a quantidade de arquivos e a totalidade em gigabytes.

Quantidade de Buckets: 8

- cnpj-datadream
- dadosinep-datadream
- dadosmec-datadream
- datasus-datadream
- ibge-datadream
- pofmain-datadream
- receita-datadream
- zipcode-datadream

Quantidade de arquivos por buckets: 80

- cnpj-datadream: 5 arquivos
- dadosinep-datadream: 3 arquivos
- dadosmec-datadream: 8 arquivos
- datasus-datadream: 41 arquivos
- ibge-datadream: 3 arquivos
- pofmain-datadream: 16 arquivos
- receita-datadream: 3 arquivos
- zipcode-datadream: 1 arquivo

Quantidade Total de Gigabytes: 51,3 GB

- cnpj-datadream : 4,55GB
- dadosinep-datadream : 1,04GB
- dadosmec-datadream : 7,28GB
- datasus-datadream : 14,5GB
- ibge-datadream : 0,99MB

- pofmain-datadream : 900MB
- receita-datadream : 18,27MB
- zipcode-datadream : 409.2 kb

11.2. OLAP (Online Analytical Processing)

OLAP é uma categoria de software que permite aos usuários analisar dados multidimensionais de forma rápida e eficiente. Ele é projetado para consultas e análises complexas, fornecendo uma visão multidimensional dos dados. Alguns conceitos-chave associados ao OLAP:

Cubo OLAP: Os dados em um ambiente OLAP são organizados em cubos. Cada cubo contém métricas ou medidas que podem ser analisadas, bem como dimensões que representam as várias maneiras de visualizar os dados.

Dimensões: São as categorias pelas quais os dados podem ser analisados. Por exemplo, em um cubo de vendas, as dimensões podem incluir tempo, localização geográfica, produtos e clientes.

Medidas: São os dados numéricos que podem ser analisados. No contexto de vendas, as medidas podem incluir receita, quantidade vendida e lucro.

11.2.1 OLAP na AWS

Na AWS, a implementação da solução OLAP foi realizada usando o serviço Amazon Redshift. Ele oferece suporte ao processamento OLAP por meio de consultas SQL e pode integrar-se a ferramentas de visualização de dados.

11.2.2 Benefícios da Utilização de OLAP na AWS

Desempenho Rápido: O processamento OLAP na AWS, especialmente com serviços como o Amazon Redshift, oferece desempenho rápido para consultas analíticas complexas em grandes volumes de dados.

Elasticidade: Os serviços da AWS são altamente escaláveis, permitindo ajustes nos recursos conforme necessário para lidar com cargas de trabalho variáveis.

Integração com Ferramentas de Visualização: É fácil integrar soluções OLAP na AWS com ferramentas de visualização de dados populares, como Tableau, Power BI e outras, para criar relatórios e painéis interativos.

11.3. Amazon RedShift e Escalabilidade dos Dados

O Amazon Redshift é um serviço de data warehouse, ou seja, ele consolida e armazena um grande volume de dados de diferentes fontes em um único local centralizado. Esse serviço fornece uma solução de armazenamento de dados na nuvem (o que contribui para esse acesso e reunião dos dados), ele é projetado para possuir escalabilidade (é uma solução escalável e dimensionável, sendo possível começar com um cluster menor e aumentar o tamanho conforme as necessidades crescem), segurança (a AWS oferece recursos de segurança, backup e recuperação em todo ambiente fornecido pelo Redshift) e desempenho. Com todos esses atributos, a solução permite que as organizações armazenem e consultem grandes volumes de dados de maneira eficiente, o que facilita análises e relatórios em ambientes empresariais (Business Intelligence (BI)). Para o contexto do atual projeto, o Redshift facilitará a criação do ambiente OLAP (Online Analytical Processing) a partir do seguintes recursos:

SQL Aplicado: É possível usar a linguagem SQL (Structured Query Language) no Redshift para a manipulação e consulta de bancos de dados relacionais, o que permite escrever consultas analíticas complexas. A medida que as necessidades de armazenamento e processamento de dados de uma organização aumentam, é possível expandir o cluster do Redshift para lidar com essas novas demandas. Essa capacidade de escala permite que as empresas ajustem seus recursos de acordo com o volume de dados e as complexidades das consultas.

Ferramentas BI: As ferramentas de BI (como Tableau, Power BI e Looker) podem se conectar diretamente ao Redshift, permitindo aos usuários criar relatórios interativos, painéis de controle e visualização de dados sem a necessidade de manipular diretamente consultas complexas no SQL. Essa integração facilita a exploração e a interpretação dos dados para a tomada de decisão. O armazenamento colunar é uma abordagem em que os dados são organizados e armazenados por coluna em vez de por linha. Isso significa que os valores de uma coluna são armazenados juntos, o que proporciona eficiência significativa em consultas analíticas, afinal, elas envolvem a recuperação de um conjunto limitado de colunas em vez de todas as colunas de uma tabela, assim, o armazenamento colunar reduz o tempo necessário para acessar e processar os dados relevantes.

Paralelismo Massivo: o Redshift possui a capacidade de distribuir uma única consulta entre vários nós de computação, fazendo com que cada um processe parte dos dados, permitindo que a consulta seja executada em paralelo. A vantagem dessa

característica fica muito aparente em consultas complexas que envolvem grandes volumes de dados, pois o processamento simultâneo em várias máquinas acelera significativamente o tempo de resposta.

11.3.1 Possibilidades Além do Amazon Redshift

No dinâmico universo da computação em nuvem, a análise de dados é fundamental, e o Amazon Redshift, da AWS, destaca-se. Contudo, o leque de opções vai além. Apresenta-se alternativas ao Redshift, destacando três soluções competitivas. Ao examinar essas opções, as organizações podem encontrar alternativas que atendam melhor às suas necessidades específicas de análise de dados, proporcionando flexibilidade na otimização das operações em nuvem.

11.3.2. Google BigQuery

Desenvolvido pela Google, o Google BigQuery, é um serviço de data warehouse baseado em nuvem que oferece escalabilidade automática e consultas SQL rápidas para análises de dados em larga escala.

Vantagens em Relação ao AWS Redshift:

- Escalabilidade automática, permitindo consultas rápidas em grandes conjuntos de dados, sem a necessidade de gerenciar a infraestrutura;
- A integração eficiente com outros serviços da Google Cloud;
- Melhor modelo de precificação (pay-as-you-go) do BigQuery, simplificando a previsão de custos;

Desvantagens em Relação ao Amazon Redshift:

- Maior custo para Grandes Volumes de Dados;
- Dependência de Conectividade com a Internet;
- Menos Adequado para Cargas de Trabalho Transacionais (não é a melhor escolha para transações ou atualizações frequentes).

11.3.3. Snowflake

O Snowflake, desenvolvido pela Snowflake Computing, é um serviço de data warehouse na nuvem conhecido por sua arquitetura multi-cluster única.

Vantagens em Relação ao AWS Redshift:

- Arquitetura multi-cluster que permite escalabilidade horizontal automática, otimizando o desempenho conforme a demanda;
- Dimensionamento de recursos de forma independente, otimizando custos conforme as necessidades específicas de armazenamento e processamento;

- Suporte à SQL padrão ANSI, facilitando a migração de bancos de dados existentes e uma experiência ainda mais familiar para os usuários.

Desvantagens em Relação ao Amazon Redshift:

- Modelagem de Custos Complexa;
- Menos integrações diretas com algumas ferramentas de BI e ecossistemas de nuvem;
- Curva de aprendizado maior para compreender completamente sua arquitetura única e aproveitar ao máximo seus recursos.

11.3.4. Azure Synapse Analytics

O Azure Synapse Analytics também é um data warehouse que oferece recursos para consultas analíticas rápidas em grandes conjuntos de dados, além de ser escalável e projetado para lidar com cargas de trabalho analíticas complexas.

Vantagens em Relação ao AWS Redshift:

- Integração com Ecossistema Azure: O Azure Synapse Analytics tem uma integração profunda com o ecossistema Azure, permitindo uma colaboração eficiente com outros serviços e ferramentas da plataforma. Isso pode facilitar a criação de soluções end-to-end e a implementação de arquiteturas de dados mais abrangentes;
- Flexibilidade na Escala: O Azure Synapse Analytics oferece uma flexibilidade significativa na escala, permitindo que você dimensione recursos de armazenamento e computação conforme necessário. Isso possibilita ajustar a capacidade de acordo com as demandas específicas do seu projeto, otimizando custos e desempenho;
- Análise em Tempo Real: A capacidade de realizar análises em tempo real é uma vantagem do Azure Synapse Analytics. Ele permite a ingestão e a análise de dados em tempo real, fornecendo insights mais recentes e suportando casos de uso que exigem tomada de decisões em tempo hábil.

Desvantagens em Relação ao Amazon Redshift:

- Custo: Em comparação com outras soluções, como o Amazon Redshift, o Azure Synapse Analytics pode ter custos mais elevados em algumas situações. O modelo

de preços pode ser complexo, e os usuários precisam entender completamente como os recursos são alocados e utilizados para otimizar os custos;

- Curva de Aprendizado: A curva de aprendizado para usar efetivamente o Azure Synapse Analytics pode ser íngreme para usuários inexperientes. Isso se deve à sua gama de recursos e à necessidade de compreender a integração com outros serviços Azure, o que pode demandar tempo e esforço;
- Maturidade do Serviço: Dependendo dos requisitos específicos do seu projeto, a maturidade do Azure Synapse Analytics em comparação com soluções mais consolidadas, como o Amazon Redshift, pode ser uma consideração. A estabilidade e a maturidade da plataforma podem variar de acordo com as necessidades individuais e a evolução do serviço ao longo do tempo.

11.3.5 Criação Redshift AWS

Criar um cluster Amazon Redshift na AWS Lab envolve vários passos. Abaixo está um guia passo a passo que você pode seguir. Certifique-se de ter permissões adequadas para criar recursos no AWS Lab.

Passo 1: Acesse o Console da AWS Lab e faça login na sua conta.

Passo 2: No Console da AWS, pesquise em serviços por "Amazon Redshift". No menu lateral, selecione a opção "Painel de clusters provisionados".

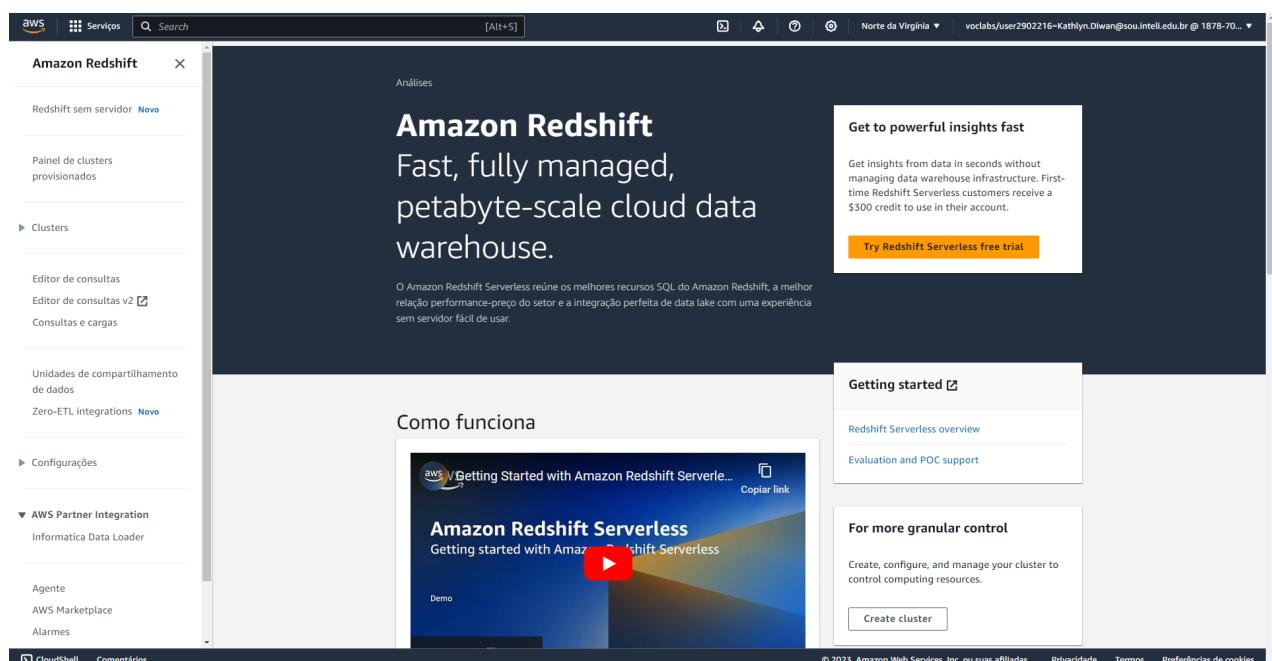


Figura 134: Painel da AWS

Fonte: Autoria Própria

Passo 3: Clique no botão "Create cluster" (Criar cluster).

The screenshot shows the AWS Amazon Redshift console under the 'Painel de clusters provisionados' (Provisioned Clusters Dashboard). The left sidebar includes sections for 'Redshift sem servidor', 'Clusters', 'Editor de consultas', 'Consultas e cargas', 'Unidades de compartilhamento de dados', 'Zero-ETL Integrations', 'Configurações', 'AWS Partner Integration', 'Informatica Data Loader', 'Agente', 'AWS Marketplace', and 'Alarms'. The main dashboard displays metrics like 'Total de nós' (0), 'Nós sob demanda' (0), 'Nós reservados' (0), 'Nós reservados disponíveis (0 de 0 usados)' (0), 'Snapshots automatizados' (0), and 'Snapshots manuais' (0). A central section titled 'Visão geral do cluster (0)' shows a table with columns 'Cluster' and 'Status', indicating 'Nenhum cluster' (No cluster). Below this is a 'Criar um cluster do Amazon Redshift' (Create a cluster) button. To the right, there are sections for 'Unidades de compartilhamento de dados' (Data sharing units), 'Alarmes' (Alarms), and other navigation links.

Figura 135: Console da AWS

Fonte: Autoria Própria

Passo 4: Configuração do Cluster

- *Cluster identifier*: Insira um nome único para o seu cluster.
- *Node type*: Escolha o tipo de nó para o seu cluster.
- *Number of nodes*: Configure o número de nós no seu cluster.

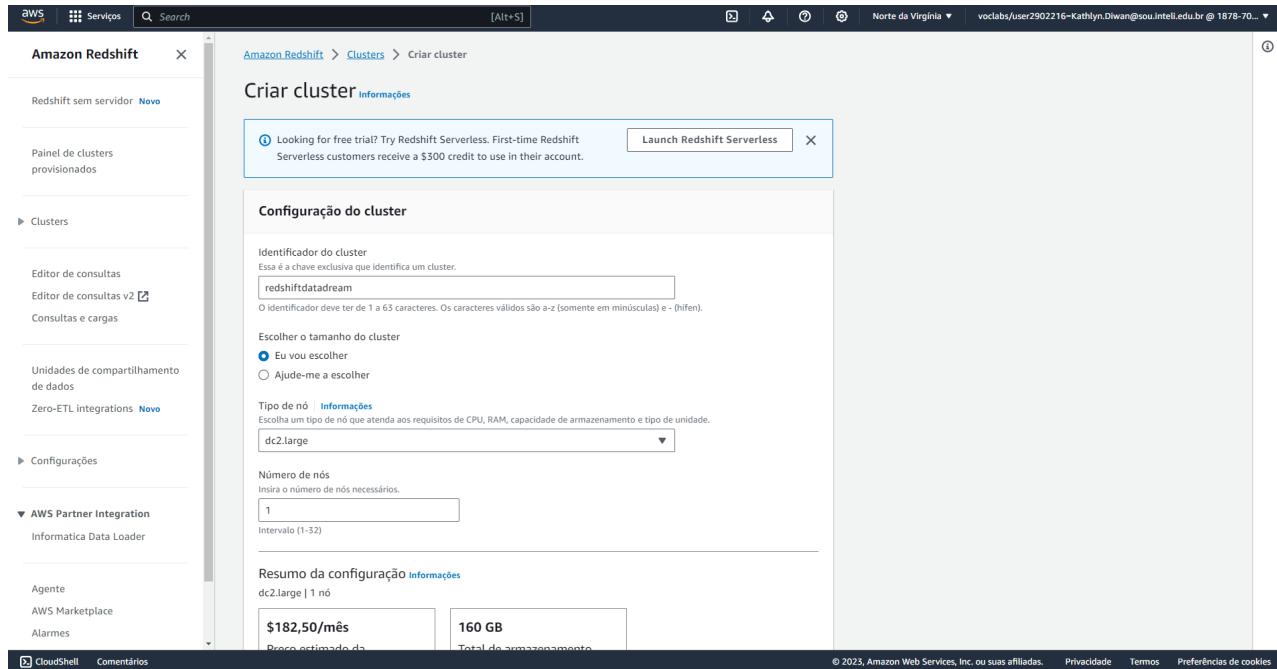


Figura 136: Console da AWS

Fonte: Autoria Própria

- **Database name:** Escolha um nome para o seu banco de dados.
- **Master username e Master password:** Defina um nome de usuário e uma senha para o usuário mestre.

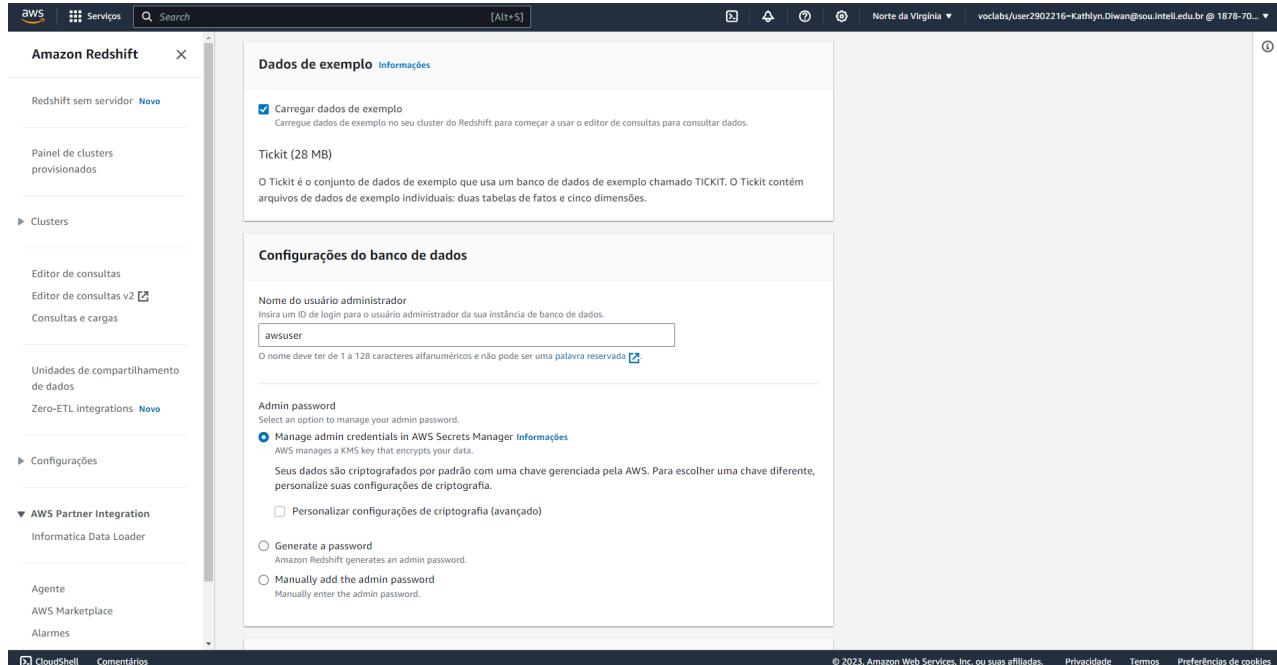


Figura 137: Console da AWS

Fonte: Autoria Própria

Passo 5: Criação IAM

- No Console da AWS, vá para a seção "Services" (Serviços);
- Selecione "IAM" sob a categoria "Security, Identity, & Compliance";
- No painel de navegação à esquerda, escolha "Roles" e clique em "Create role" (Criar função);
- Selecione "AWS service" como o tipo de entidade de confiança e escolha "Redshift" como o serviço que será confiável.

Observação: No processo de criação da função, escolha as políticas do IAM que concederam as permissões necessárias ao Redshift. Isso pode incluir políticas como "*AmazonRedshiftFullAccess*" ou políticas personalizadas que você configurou.

Passo 5.1: Adicione a Função ao Cluster Redshift

- Volte ao Console do Amazon Redshift;
- Selecione o cluster que você criou anteriormente;
- Função do IAM Associados;
- Clique em "Add IAM role" (Adicionar função IAM) e selecione a função IAM que você criou.

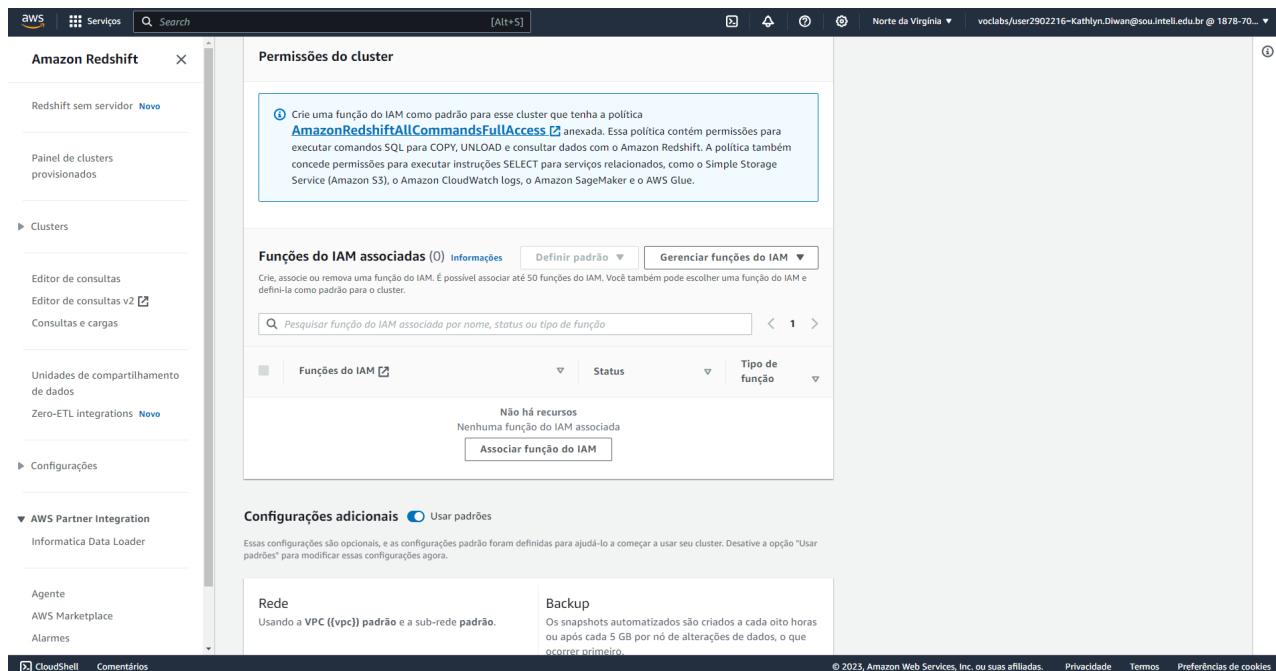


Figura 138: Função ao Cluster Redshift

Fonte: Autoria Própria

Passo 6: Configurações Avançadas (Opcional)

Virtual Private Cloud (VPC): Escolha a VPC na qual o cluster será lançado.

- *Publicly accessible*: Decida se o cluster será acessível publicamente ou apenas internamente;
- *VPC security groups*: Configure os grupos de segurança VPC para controlar o tráfego de rede para o cluster;
- *Automated snapshots*: Escolha se deseja habilitar snapshots automáticos;
- *Snapshot retention period*: Configure o período de retenção para os snapshots automáticos.

Observação: Revise as configurações do seu cluster. Clique em "Create cluster" para iniciar a criação do cluster.

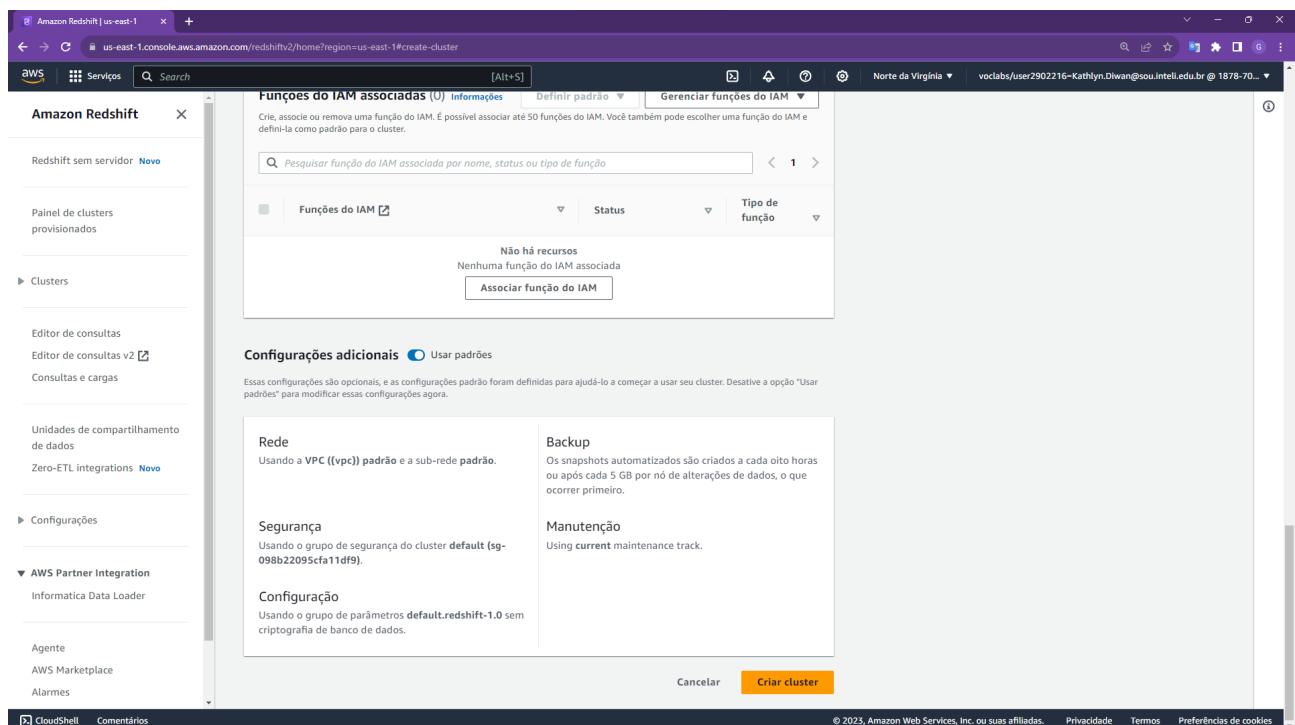


Figura 139: Configurações do cluster

Fonte: Autoria Própria

Passo 7: Aguarde a Criação

Aguarde enquanto o cluster é criado. Isso pode levar alguns minutos.

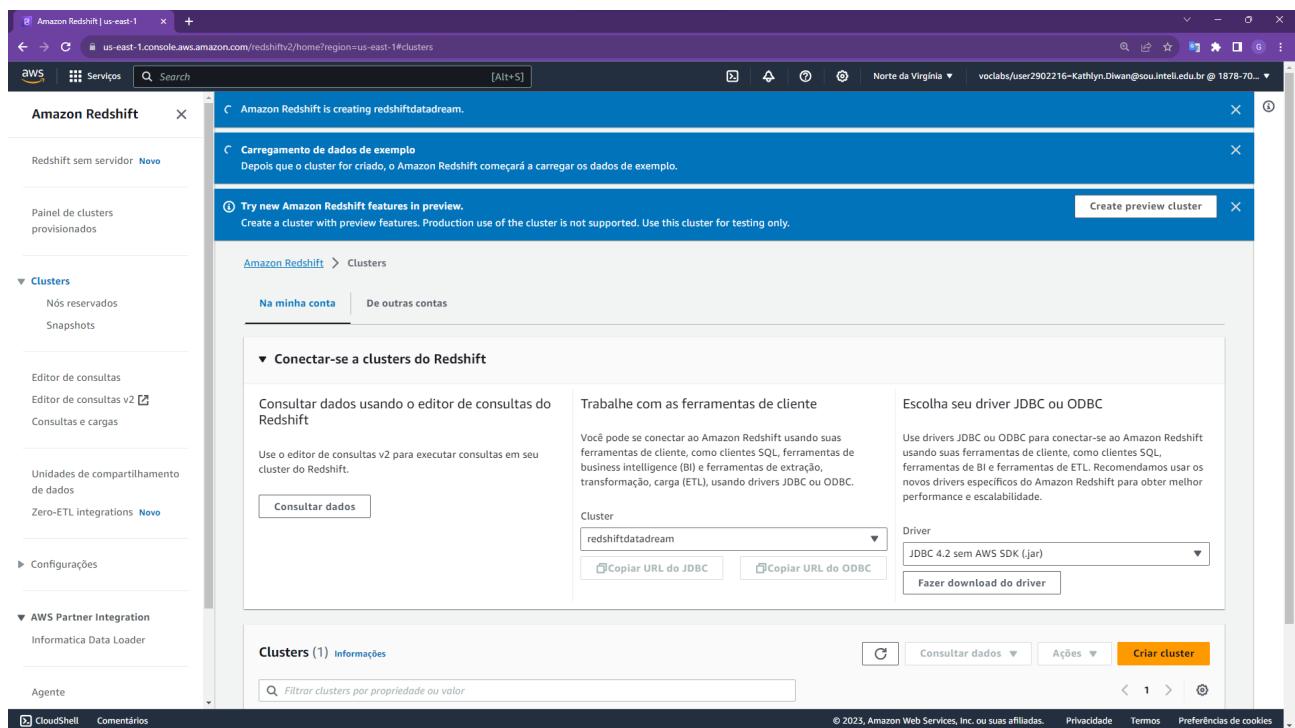


Figura 140: Configurações do cluster

Fonte: Autoria Própria

Passo 8: Acesse o Cluster Após a conclusão, vá para a lista de clusters no Console do Redshift.

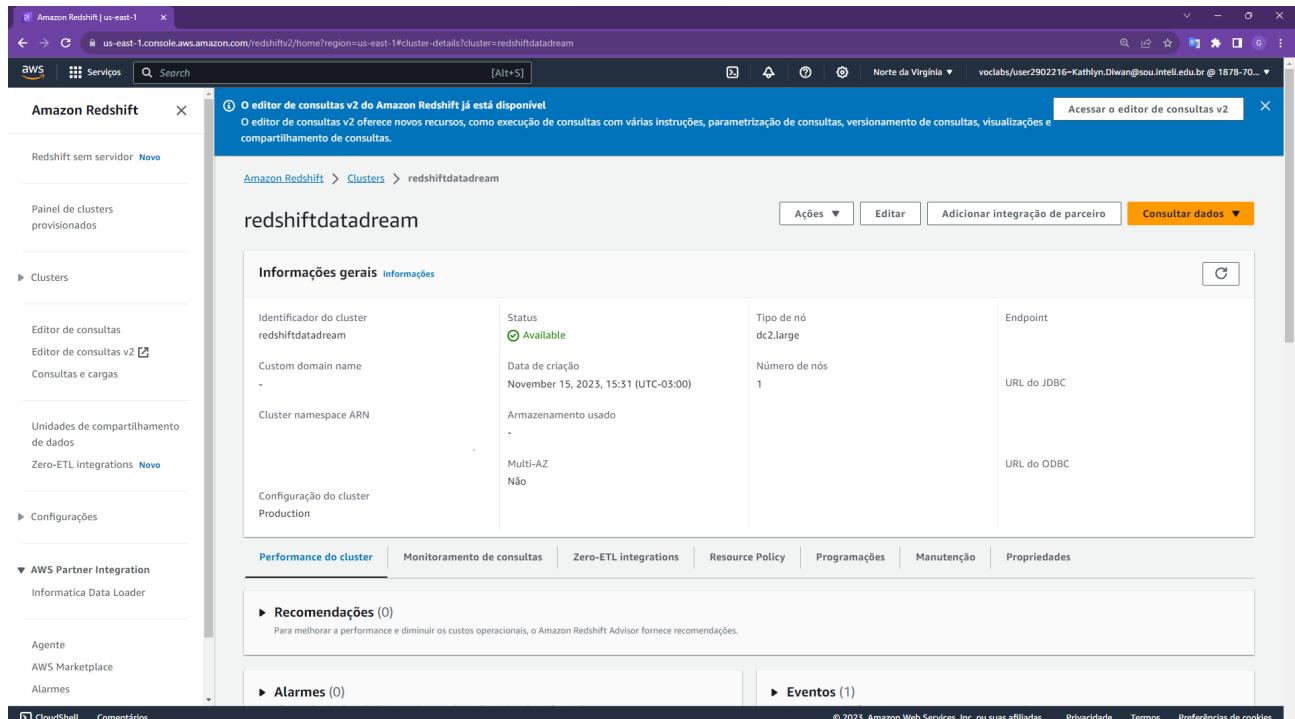


Figura 141: Lista de clusters no Console do Redshift

Fonte: Autoria Própria

11.5. Data Warehouse

Um Data Warehouse é um sistema de gerenciamento de banco de dados projetado para armazenar e analisar grandes volumes de dados, provenientes de várias fontes dentro de uma organização. Ele é otimizado para consulta e análise de dados, oferecendo uma visão integrada e consolidada para facilitar a tomada de decisões.

11.5.1. Importância do Data Warehouse

Consolidação de Dados: O Data Warehouse permite a integração de dados de diferentes fontes, proporcionando uma visão unificada e consistente das informações. Isso facilita a análise e gera insights mais precisos.

Análise de Dados Históricos: Ao armazenar grandes volumes de dados ao longo do tempo, o Data Warehouse possibilita a análise de tendências e padrões históricos, contribuindo para uma compreensão sobre o desempenho da organização.

Suporte à Tomada de Decisões: Ao fornecer um ambiente centralizado para análise de dados, o Data Warehouse capacita os tomadores de decisão a extrair informações rapidamente, promovendo decisões mais informadas e estratégicas.

11.5.2. Data Warehouse - Contexto do Projeto

No contexto do projeto, destaca-se a implementação de um Data Warehouse no Amazon Redshift. A fase de carga (Load) engloba a ingestão eficiente de dados no Redshift, seguida pela organização (Organize) para assegurar que os dados estejam formatados e otimizados para consultas. Na etapa final, Descoberta (Discover), ocorre a exploração e análise de dados para extrair insights.

Essa integração de um Data Warehouse no Redshift não apenas simplifica a consolidação e análise de dados, mas também tira proveito da infraestrutura da AWS para garantir escalabilidade. Abaixo, encontra-se um passo a passo detalhado de como realizar o load desses dados no Redshift, oferecendo uma orientação prática para a execução bem-sucedida dessa fase do processo. Depois do último passo descrito no tópico 4.2, você deve clicar em: “Consultar dados”, isso abrirá uma aba nova com o “Query Editor” do Redshift.

The screenshot shows the 'Informações' (Information) tab of a workspace named 'workspace-cubo-data-dream'. The 'Informações gerais' (General Information) section displays the following details:

Namespace	Status	Nome do usuário administrador
workspace-cubo-data-dream	Available	admin
Namespace ID	Data de criação	Nome do banco de dados
2b23c6b1-fdf0-4011-aae6-a6e56387d063	November 21, 2023, 22:52 (UTC-03:00)	dev
Namespace ARN	Storage used	Total table count
arn:aws:redshift-serverless:us-east-1:185405266895:namespace/2b23c6b1-fdf0-4011-aae6-a6e56387d063	153,5 GB	92

Below this, there are tabs for 'Workgroup', 'Backup de dados', 'Segurança e criptografia', 'Unidades de compartilhamento de dados', and 'Zero-ETL integrat'. The 'Workgroup' tab is selected. The 'Workgroup name' section shows a single entry:

Workgroup	Status
cubo-data-dream	Available

Figura 142: Workspace

Fonte: Autoria Própria

No canto superior esquerdo, é possível visualizar um botão escrito “Load Data”, é necessário clicar nele para subir os dados.

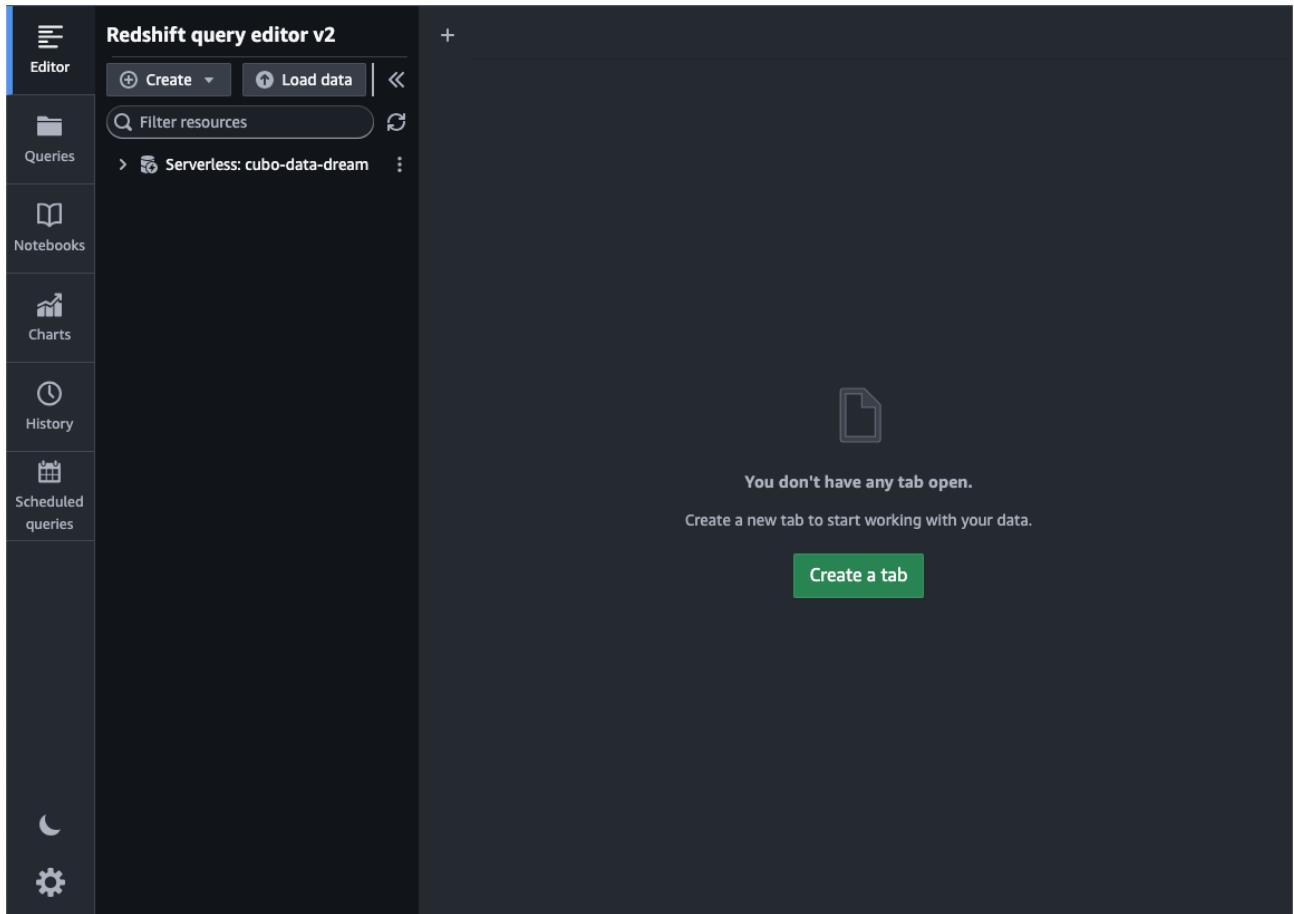


Figura 143: Query Editor - RedShift

Fonte: Autoria Própria

Com o pop-up aberto, você deve selecionar se quer subir um dados que está em um Bucket S3 ou no seu computador, neste caso vamos com a primeira opção. Selecionado, podemos buscar o bucket com o botão: “Browse S3” e selecionar o arquivo único que você deseja subir no momento. Além disso, é importante verificar as informações do formato do arquivo e do seu delimitador. Com tudo selecionado e confirmado, podemos clicar em “Next”.

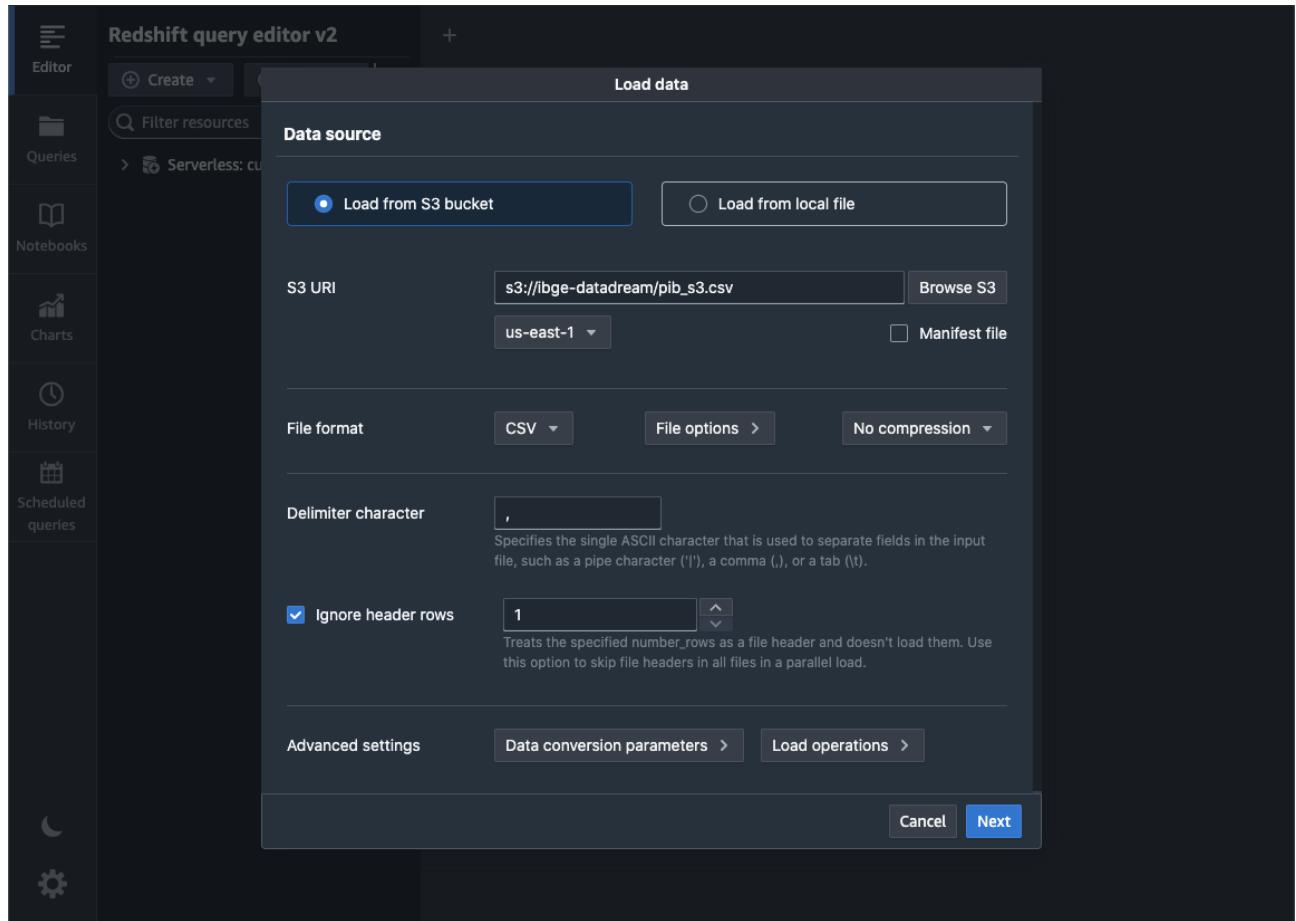


Figura 144: Load Data - Parte 1

Fonte: Autoria Própria

Agora, vamos escolher ou criar a tabela que este arquivo será inserido. Caso a tabela já esteja criada, e você deseja somente atualizar, é necessário clicar em “Load existing table”. Neste caso, vamos criar uma nova tabela, então devemos clicar em “Load New Table”. É necessário selecionar o nome do cluster, criado no tópico 4.2, o database e o schema, este último deve ser sempre o “public”, além disso deve-se criar o nome para a sua tabela e o IAM criado. Com essas configurações, podemos clicar em “Create Table”, ela será criada e você deve clicar em “Load Data”.

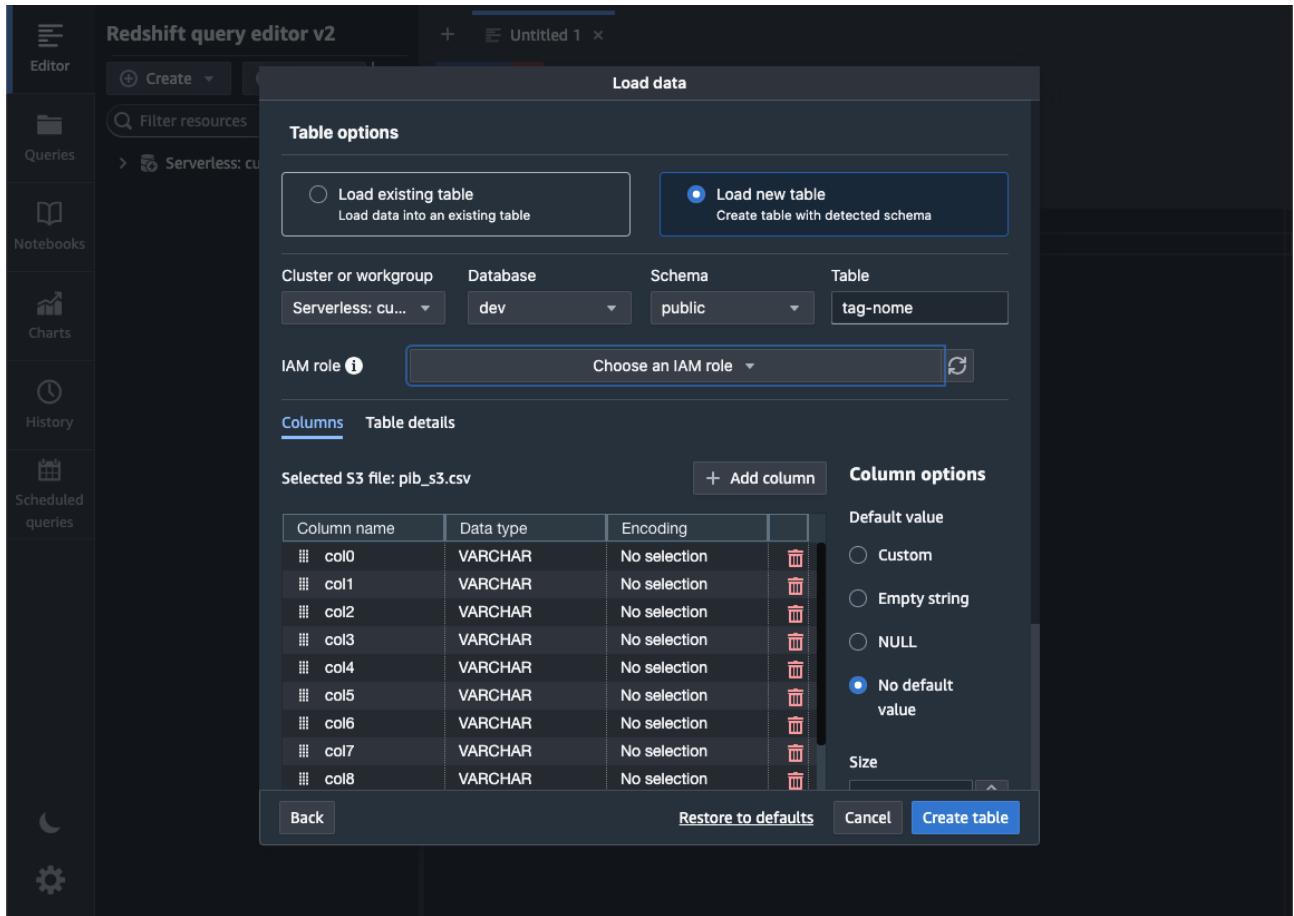


Figura 145: Load Data - Parte 2

Fonte: Autoria Própria

11.6 Views

Em OLAP, uma "view" refere-se a uma visão ou perspectiva específica dos dados armazenados em um banco de dados multidimensional. Uma view em OLAP é uma representação virtual dos dados que pode ser personalizada para atender às necessidades específicas dos usuários. Essas visualizações podem incluir subconjuntos específicos de dados, agregações, cálculos e hierarquias que facilitam a análise de tendências, padrões e resumos de informações.

No Amazon Redshift, pode-se criar uma view usando a linguagem SQL padrão. Aqui está um exemplo de como criar uma view no Amazon Redshift:

```
CREATE VIEW nome_da_view AS
SELECT coluna1, coluna2, SUM(coluna3) AS soma_coluna3
```

```
FROM tabela  
GROUP BY coluna1, coluna2;
```

Neste exemplo, você está criando uma view chamada "nome_da_view" que seleciona algumas colunas da tabela e calcula a soma da coluna3, agrupando pelos valores das colunas 1 e 2.

11.6.1 Importância das views

Simplicidade: As views podem ser projetadas para atender às necessidades específicas dos usuários finais, tornando a análise de dados mais fácil e mais compreensível.

Desempenho: As views podem ser otimizadas para consultas frequentes, permitindo que os usuários acessem dados agregados ou resumidos sem a necessidade de consultar diretamente as tabelas subjacentes, o que pode impactar em termos de desempenho.

Segurança: Views podem ser usadas para restringir o acesso aos dados, permitindo que os usuários vejam apenas as informações relevantes para suas funções e responsabilidades.

Consistência: Ao criar views predefinidas, você garante que os usuários estejam trabalhando com conjuntos de dados consistentes e padronizados, o que ajuda a evitar interpretações divergentes dos dados.

Facilidade de Manutenção: Se houver mudanças na estrutura dos dados, as views podem ser ajustadas sem afetar diretamente as consultas dos usuários, desde que a lógica das views seja mantida.

11.6.2 Entendimento dos dados para as views

Nestas views, serão utilizados três conjuntos de dados:

Dados de Localização Geográfica:

- IBGE;
- Códigos postais.

Dados com Indicadores Socioeconômicos:

- Pesquisa de Orçamento Familiar (POF);
- MEC;

- Receita Federal;
- SUS;
- INEP;
- IBGE.

Dados sobre Canais de Atendimento Disponíveis:

- CNPJ.

11.6.3 Primeiras hipóteses criadas

11.6.3.1. Distribuição de Canais de Atendimento a Partir dos CNPJs Cadastrados

A análise visa identificar a distribuição dos canais de atendimento (como mercado, café, bar, etc.) com base nos CNPJs cadastrados nos Buckets (cnpjs_1, cnpjs_2, cnpjs_3, cnpjs_4, cnpjs_5). A comparação será feita considerando a situação regional, características do canal e período.

Colunas sobre situação Regional:

- sigla_uf
- id_municipio

Características do Canal de Atendimento:

- identificador_matriz_filial
- cnae_fiscal_principal
- cnae_fiscal_secundaria

Período:

- data (data completa)

11.6.3.2. Renda Por Região a Partir do Imposto de Renda (POF)

A análise compara a soma dos valores das colunas de "Renda" com o estado indicado (Ente Federativo) e período, utilizando dados da Receita Federal nos Buckets (distribuicao-renda, distribuicao-renda-socios, distribuicao-renda-socios).

Colunas de renda:

- Rendimentos Tributáveis,
- Isentos, etc.

Estado:

- Ente Federativo

Período:

- Ano-Calendário
- Análise limitada a nível estadual.

11.6.3.3. Renda Por Estado a Partir das Despesas Coletivas (POF)

A análise utiliza a tabela "despesa_coletiva" (Bucket POF) para identificar a condição socioeconômica de cada estado a partir do valor da despesa coletiva, considerando situação regional, características da despesa e período.

Colunas da situação Regional:

- UF (Estado)
- TIPO_SITUACAO_REG (Situação Urbana ou Rural)

Características da Despesa Coletiva:

- NUM_UC (Número de pessoas incluídas no conjunto de consumo)
- V9002 (Forma de Aquisição, Peso Final, Renda Total)

Período:

- V9010 (mês)
- Dados de 2010 até 2018 e análise limitada a nível estadual.

11.6.3.4. Renda Por Estado a Partir das Despesas Individuais (POF)

A análise utiliza a tabela "despesa_individual" (Bucket POF) para identificar a condição socioeconômica de cada estado a partir do valor da despesa individual, considerando situação regional, características da despesa individual e período.

Colunas da situação Regional:

- UF (Estado)
- TIPO_SITUACAO_REG (Situação Urbana ou Rural)

Características da Despesa Individual:

- NUM_UC (Número de pessoas incluídas no conjunto de consumo)
- V9002 (Forma de Aquisição, Peso Final, Renda Total)

Período:

- V9010 (mês)

11.6.3.5. Poder Socioeconômico Regional de Estados a Partir do Desenvolvimento da Indústria (IBGE)

A análise utiliza a tabela "gini_industria_s3" (Bucket IBGE) para identificar a condição socioeconômica de cada estado a partir do desenvolvimento da indústria local, comparando os estados brasileiros ao longo dos anos.

Colunas dos estados Brasileiros:

- Unidade da Federação (estado)

Desenvolvimento Industrial por Ano (Dados disponíveis para anos específicos):

- 2002, 2003, ..., 2017, 2018
- Análises Compostas (Comparando uma ou mais tabelas)

11.6.3.6. Potencial Mercado Consumidor em uma Região

Observação: Considerando a concentração de Canais de Atendimento a Partir dos CNPJs Cadastrados e Renda a Partir do Imposto de Renda (Receita Federal).

Análise envolve a soma do número de cada canal de atendimento e a soma dos rendimentos para cada estado, normalizando datas e regiões. Para fazer essa análise, é interessante fazer:

1: uma soma do número de cada canal de atendimento (cnae_fiscal_principal) para cada estado e cidade

2: soma dos rendimentos (Rendimentos_....) para cada estado

Por fim, é necessário normalizar as datas (ano) e regiões de cada tabela, para que ambos se refiram à mesma região respectiva. Comparar duas tabelas simples lado a lado também é uma opção válida.

11.6.4 Primeiras views criadas

11.6.4.1 View para produtos consumidos por estado (considerando insegurança alimentar)

Duas views foram criadas com o intuito definir quais são os cinco produtos mais consumidos pelas famílias que consideram ter algum grau de insegurança alimentar em cada estado do Brasil, bem como aqueles consumidos por famílias que não possuem esse problema. A partir dessa análise, o parceiro de projeto conseguirá visualizar quais são aqueles produtos que atendem tanto ao público que tem níveis de insuficiência alimentar, quanto aquele que não possui essas condições. Diante disso, é possível

estabelecer estratégias comerciais que gerem lucro para a empresa e impacto social nos locais em que forem aplicadas.

Tabelas correlacionadas:

consumo_alimentar: fornece informações sobre os principais hábitos alimentares das famílias, com dados referentes ao horário de consumo, produtos consumidos, local da alimentação, etc.

qualidade_vida: fornece informações sobre os principais tópicos sobre qualidade de vida, como lazer, moradia, alimentação, situação financeira, etc.

Relevância para o projeto:

As views acima apresentadas são relevantes no processo de identificação dos principais produtos consumidos nos estados brasileiros, fazendo uma correlação entre os produtos consumidos e as situações alimentares da família que o consumiu. Diante disso, é possível fazer uma análise de localidades em que é possível atuar a partir de determinados nichos e mercados, com o intuito de atender, também, às famílias carentes daquele lugar. Além disso, é possível identificar nichos ainda inexplorados que podem ser agregados à linha de produtos ou serviços principal.

Exemplo de representação gráfica:

O gráfico a seguir apresenta os produtos mais consumidos por famílias que sofrem com insegurança alimentar em todo o território nacional.

Alimentos mais consumidos nos estados brasileiros

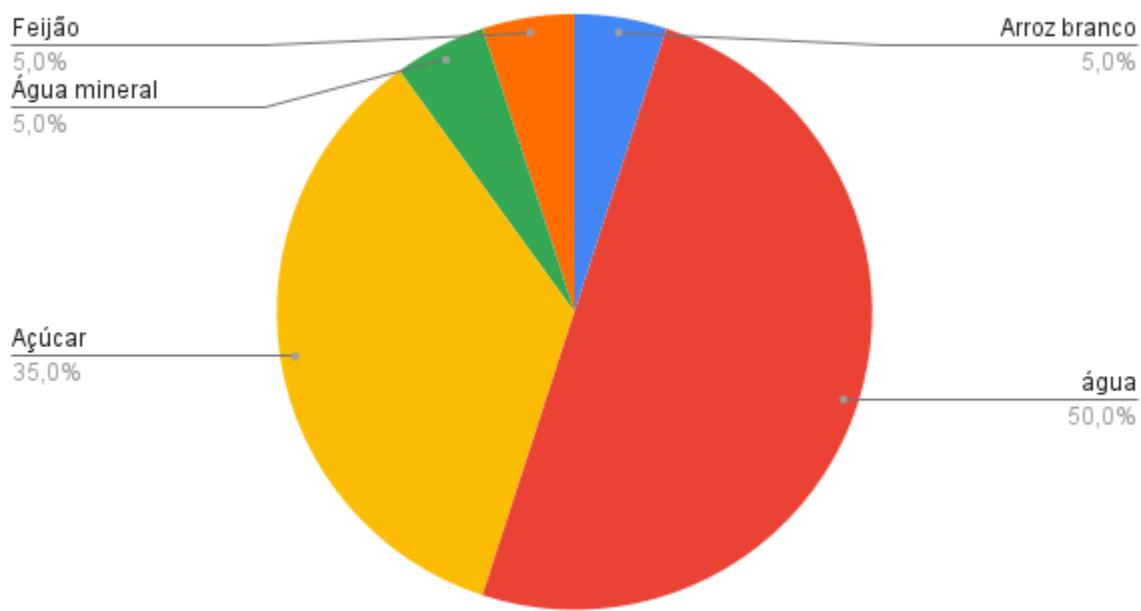


Figura 146: Gráfico - View de famílias que possuem insegurança alimentar

Fonte: Autoria própria.

A partir desses dados, é possível perceber que 50% dos estados brasileiros possuem a água como primeiro produto consumido por famílias com insegurança alimentar. Seguido do açúcar e dos demais produtos analisados.

Abaixo é possível visualizar como fica a tabela cujos dados são referentes às famílias que não possuem insegurança alimentar. Na imagem é possível identificar a tradução do código do produto em uma coluna denominada "Alimento":

Unidade federativa	Alimento	Código do produto	Posição no ranking estadual	Frequência
Rondônia	FEIJAO (PRETO, MULATINHO, ROXO, ROSINHA, ETC)	6303102	1	3336096
Rondônia	AGUA MINERAL	8201004	2	3139744
Rondônia	ARROZ BRANCO	6300113	3	2713056
Rondônia	ACUCAR	6906602	4	2363776
Rondônia	CAFE	8501302	5	1657664
Roraima	ARROZ BRANCO	6300113	1	1705233
Roraima	CAFE COM LEITE	8501303	2	1162872
Roraima	PAO FRANCES	8000101	3	1022202
Roraima	ACUCAR	6906602	4	915918
Roraima	MARGARINA COM OU SEM SAL	7901602	5	773685
Pará	AGUA	8216301	1	22071731
Pará	ACUCAR	6906602	2	8575626
Pará	FARINHA DE MANDIOCA	6501401	3	5758646
Pará	CAFE	8501302	4	5105199
Pará	ARROZ BRANCO	6300113	5	4269341
Amapá	AGUA	8216301	1	2819754
Amapá	ACUCAR	6906602	2	1505028
Amapá	CAFE COM LEITE	8501303	3	1093512
Amapá	MARGARINA COM OU SEM SAL	7901602	4	990150
Amapá	ARROZ BRANCO	6300113	5	833658
Piauí	AGUA	8216301	1	52141012
Piauí	ACUCAR	6906602	2	14869470
Piauí	ARROZ BRANCO	6300113	3	9139515
Piauí	FEIJAO (PRETO, MULATINHO, ROXO, ROSINHA, ETC)	6303102	4	8145060
Piauí	CAFE	8501302	5	6396082
Paraíba	ACUCAR	6906602	1	10949479
Paraíba	CAFE	8501302	2	7035977
Paraíba	AGUA	8216301	3	4898000
Paraíba	FEIJAO (PRETO, MULATINHO, ROXO, ROSINHA, ETC)	6303102	4	4518405
Paraíba	ARROZ BRANCO	6300113	5	3862073
Sergipe	ACUCAR	6906602	1	8111202

Figura 147: Tabela - View de famílias que possuem insegurança alimentar

Fonte: Autoria própria.

Além disso, na coluna "Frequência" é possível visualizar a quantidade de vezes que aquele alimento apareceu no respectivo estado, demonstrando a frequência de consumo dentre as pessoas que não possuem insegurança alimentar.

Diante disso, é possível perceber, por exemplo, que o açúcar está entre os produtos mais consumidos no país (atrás somente da água, que é um item essencial para a sobrevivência humana), mesmo entre as famílias com algum nível de insegurança alimentar.

11.6.4.2. View para Características de Dieta:

Ao unir informações sobre hábitos alimentares, preferências e características específicas de diferentes domicílios, esta view proporciona entender as escolhas alimentares em diversos estratos populacionais. Essa visão é crucial para o cliente, permitindo uma segmentação mais precisa do mercado e a adaptação de estratégias de produtos para atender às demandas específicas de diferentes grupos.

6.4.3. View para Características de potencial de consumo

A query abaixo cria uma view chamada consumo_potencial que mescla informações das tabelas pof_domicilio e pof_consumo_alimentar.

Tabelas Mescladas:

pof_domicilio (d): Fornece informações sobre o domicílio, como localização (uf), número do domicílio (num_dom), características do domicílio (v0201, v0202, v0217), e peso final do domicílio (peso_final).

pof_consumo_alimentar (ca): Contém informações sobre o consumo alimentar, incluindo o número do domicílio (num_dom), número da unidade consumidora (num_uc), quantidade consumida (qtd), dados nutricionais e informações sobre o peso final do consumo (peso_final) e a renda total (renda_total).

Relevância para o Projeto:

A view fornece uma visão unificada do potencial de consumo, relacionando características do domicílio com detalhes específicos do consumo alimentar. Isso é essencial para entender o contexto em que o consumo ocorre.

A análise desses dados pode ajudar o cliente a direcionar suas ações estrategicamente, como ajustar a oferta de produtos com base nas preferências do consumidor em diferentes regiões.

Exemplo de visualização gráfica:

O gráfico mostra o consumo médio de cálcio por pessoa por dia nas duas unidades federativas do Brasil: Rondônia (UF 11) e Acre (UF 12). O consumo é medido em miligramas (mg) por dia.

UF X Consumo de Cálcio

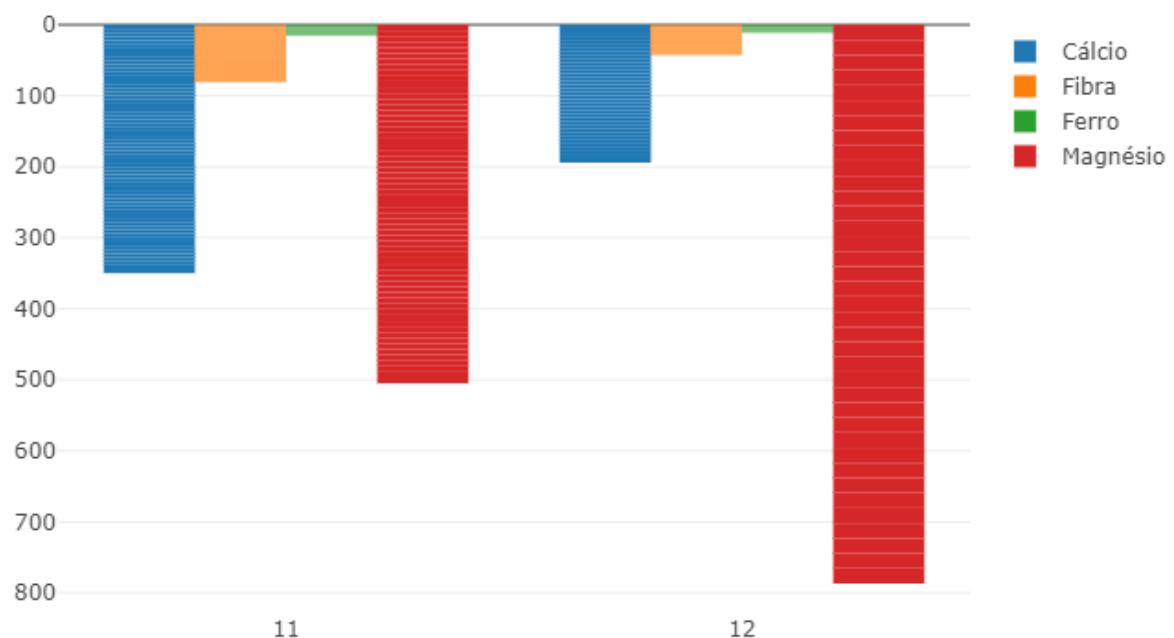


Figura 148: Gráfico - consumo médio de cálcio

Fonte: Autoria Própria

O gráfico mostra que o consumo médio de cálcio em Rondônia é de 570 mg por dia, enquanto no Acre é de 620 mg por dia. Isso significa que os habitantes do Acre consomem, em média, 50 mg de cálcio a mais por dia do que os habitantes de Rondônia.

A Organização Mundial da Saúde (OMS) recomenda que os adultos consumam pelo menos 1.000 mg de cálcio por dia. Portanto, tanto os habitantes de Rondônia quanto os do Acre estão abaixo da recomendação da OMS.

Alguns fatores que podem contribuir para o baixo consumo de cálcio nas duas unidades federativas são:

- A dieta predominante é baseada em alimentos ricos em carboidratos e pobres em nutrientes, como frutas, vegetais e laticínios.
- A falta de acesso a alimentos ricos em cálcio, como leite, queijo e iogurte.
- A falta de educação nutricional sobre a importância do consumo de cálcio.

Para aumentar o consumo de cálcio nas duas unidades federativas, é importante promover mudanças na dieta e na educação nutricional. Isso pode ser feito por meio de campanhas de conscientização, programas de educação alimentar e ações de incentivo ao consumo de alimentos ricos em cálcio.

Aqui estão algumas hipóteses para aumentar o consumo de cálcio:

- Incluir leite, queijo e iogurte na dieta diária.
- Consumir frutas secas, como amêndoas, nozes e castanhas.
- Adicionar vegetais às refeições.

Observação: Devido à restrição de aproximadamente 3000 linhas ao executar a view no Redshift, o gráfico atual exibe apenas dados de duas unidades federativas brasileiras. No entanto, na próxima sprint, quando aplicamos isso em uma ferramenta como o Grafana, a comparação será estendida para incluir todos os estados na análise.

11.7 API do Parceiro (Integration)

Esta seção do documento delineia uma abordagem abrangente para a gestão automatizada de dados, explorando o uso de tecnologias avançadas como Amazon EventBridge e AWS Lambda. Através de um processo detalhadamente descrito, ilustramos como a atualização e exclusão automatizadas de dados podem ser eficientemente implementadas para assegurar a relevância e segurança das informações. Além disso, apresentamos códigos específicos e simulações de funcionamento, fornecendo uma visão prática de como essas tecnologias podem ser aplicadas no gerenciamento de dados.

11.7.1 Consulta e atualização automática

A integração eficiente e segura de dados é um aspecto crucial no gerenciamento de informações provenientes de APIs de parceiros. Neste contexto, a atualização automática de dados desempenha um papel vital, assegurando que as informações estejam sempre atualizadas e disponíveis para consulta e análise. Para atingir este objetivo, implementamos um processo automatizado utilizando as capacidades avançadas do Amazon EventBridge e AWS Lambda. Este processo não só simplifica a gestão de dados, mas também aumenta a confiabilidade e a eficiência do sistema.

11.7.2 Criação da Função Lambda

Passo 1: Acesse AWS Console.

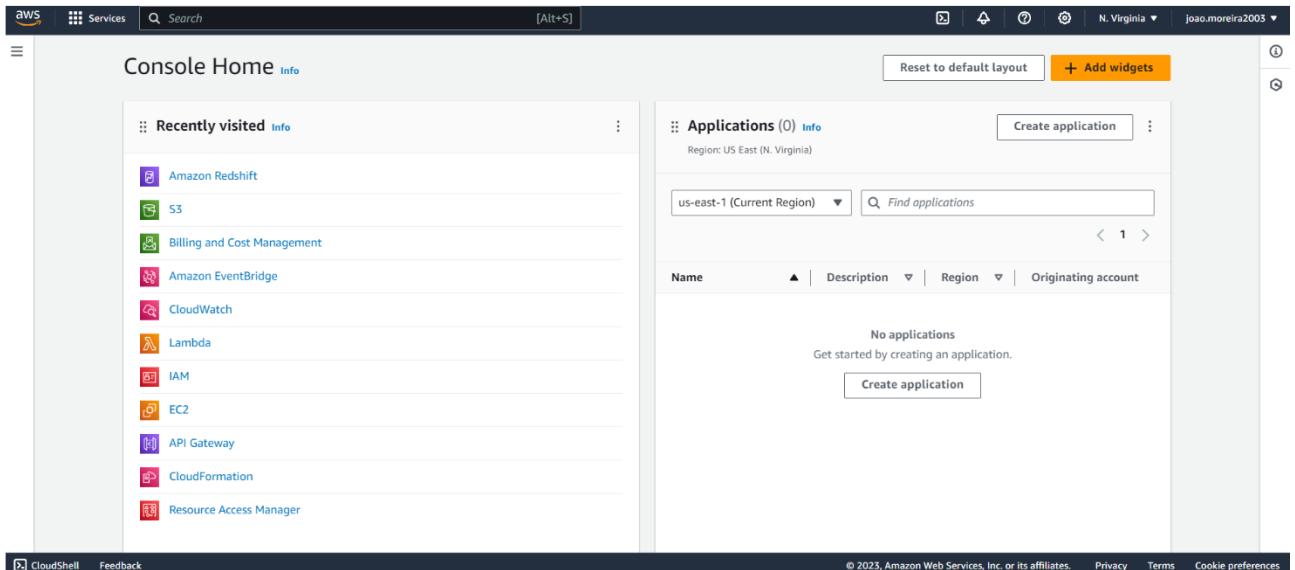


Figura 149: AWS Console Home

Fonte: Autoria Própria

Passo 2: No menu "Serviços", selecione "Lambda".

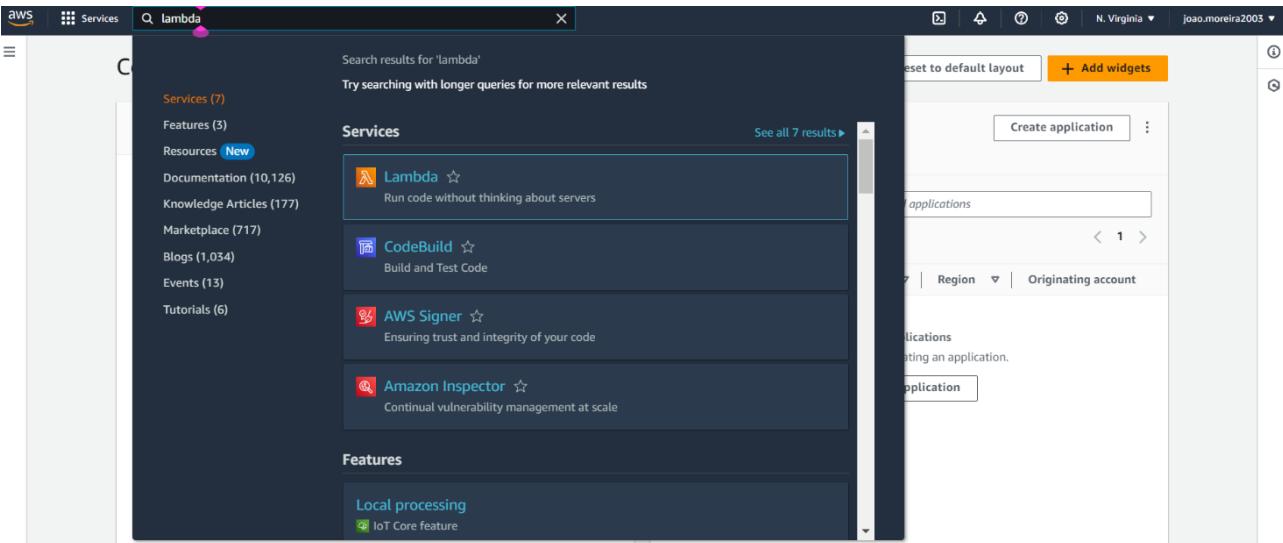


Figura 150: AWS Console Serviços

Fonte: Autoria Própria

Passo 3: Configuração dos Detalhes da Função - Clique em "Criar função".

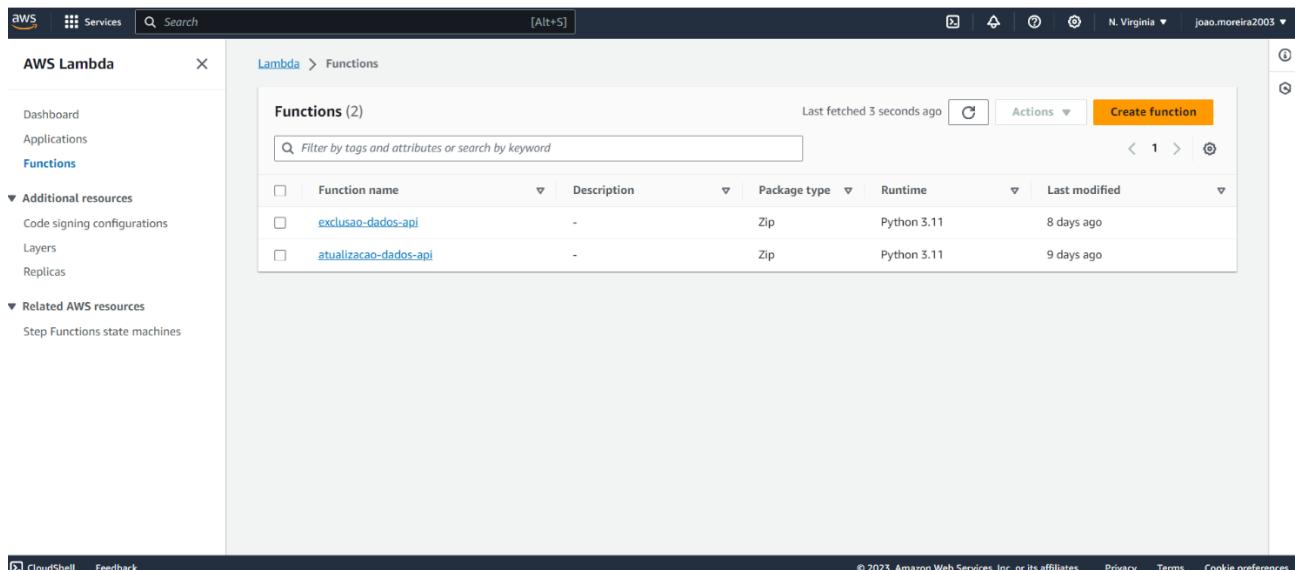


Figura 151: AWS Lambda

Fonte: Autoria Própria

Passo 4: Faça os tópicos abaixo

- Defina um nome para a função.
- Escolha a runtime desejada (por exemplo, Node.js, Python).
- Revisão e Criação da Função:
- Revise todas as configurações.
- Clique em "Criar função" para finalizar a criação da função Lambda.

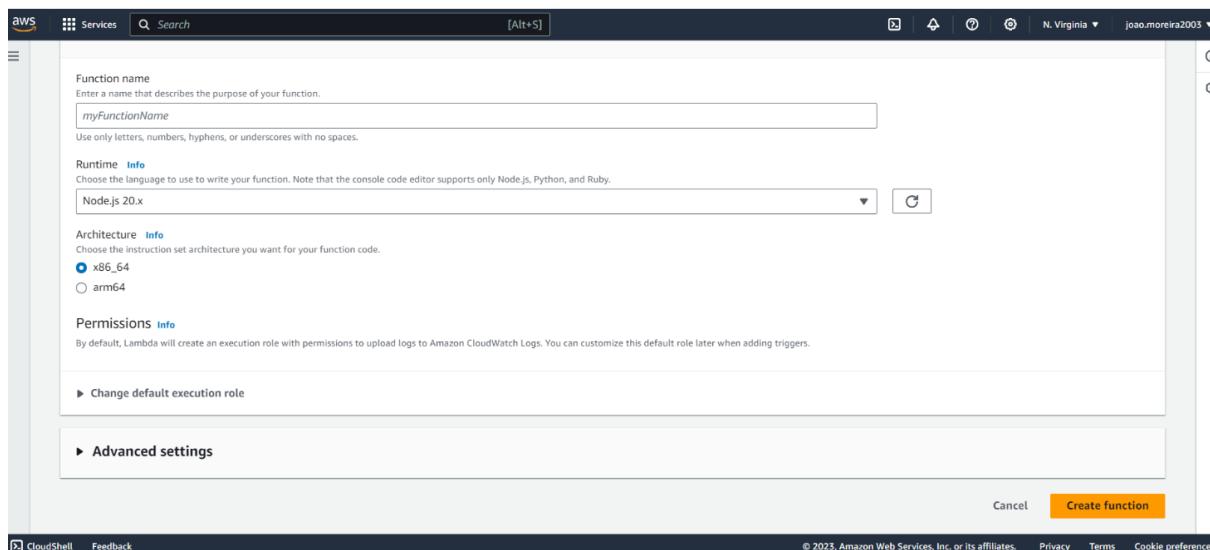


Figura 152: AWS Lambda

Fonte: Autoria Própria

Passo 4: No Amazon EventBridge, clique em "Programação do EventBridge" e, em seguida, "Criar programação".

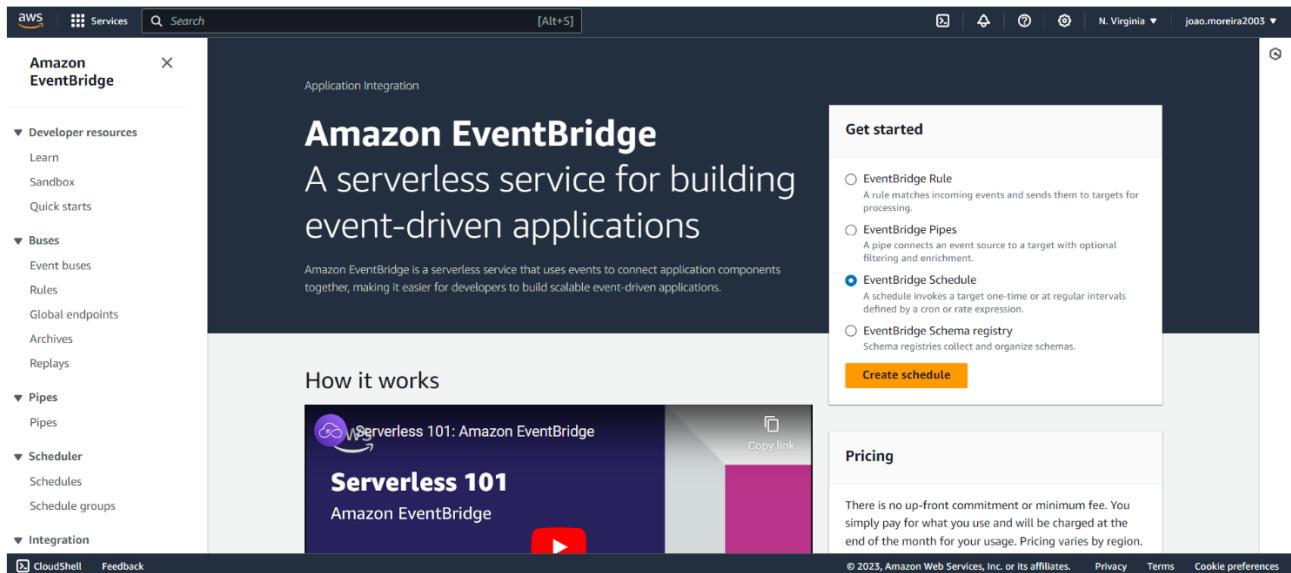


Figura 153: AWS EventBridge

Fonte: Autoria Própria

Passo 5: Especificação de Detalhes do Cronograma

- Nome do cronograma: atualizacao-dados-api-bucket.
- Descrição: Atualizar dados da api do cliente nos buckets S3 AWS.

Figura 154: AWS EventBridge

Fonte: Autoria Própria.

Passo 6: Especificação de Detalhes do Cronograma

- Ocorrência: Escolha "cronograma recorrente".
- Tipo de Cronograma: Selecione "cronograma baseado em cron".
- Expressão Cron: Configure para (0 minutos, 00 horas * dia do mês, * mês, ? dia da semana, * ano).
- Janela de Tempo Flexível: Defina como "Desligado".

- Período de Tempo: Deixe em branco.

Passo 7: Detalhes do Destino

- Escolha "destinos modelados".
- Selecione "AWS Lambda Invoke".

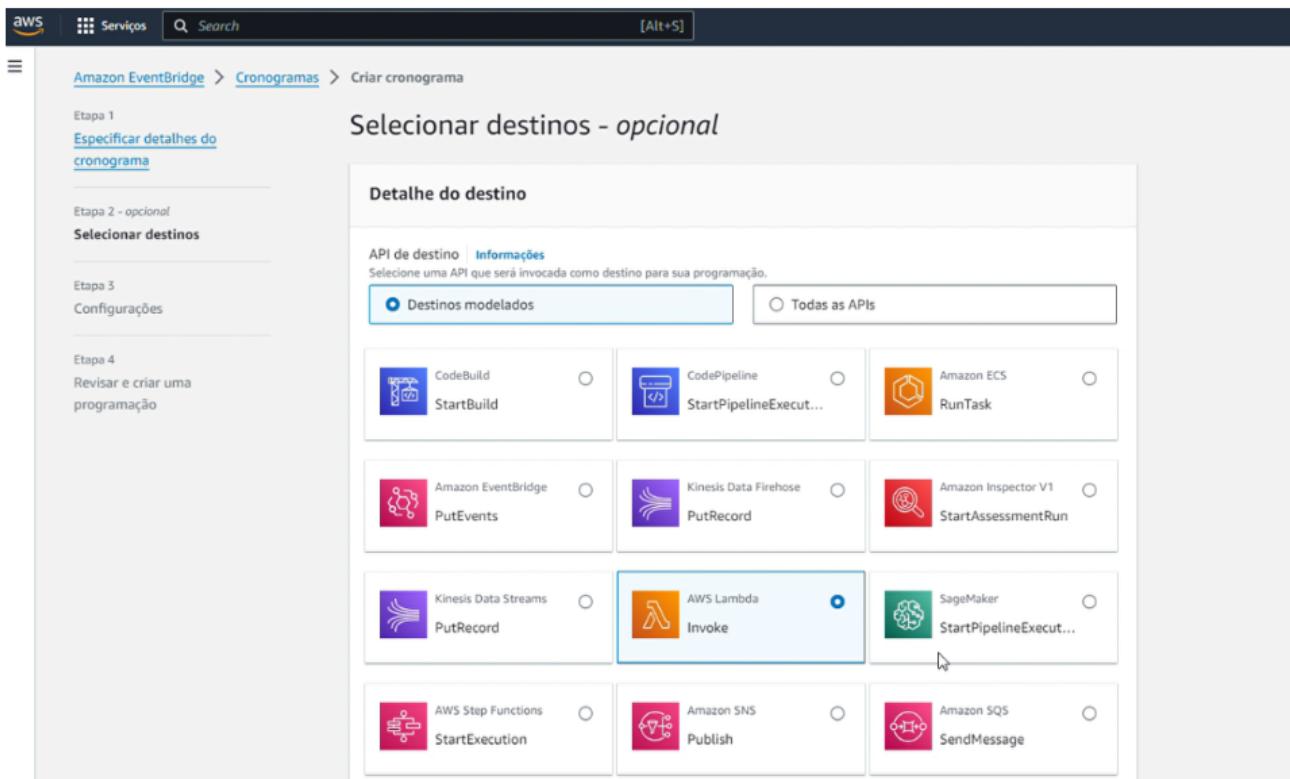


Figura 155: AWS EventBridge

Fonte: Autoria Própria.

Passo 8: Configuração do Invoke

- Selecione a função Lambda criada: atualização-dados-api.
- Clique em "Próximo".

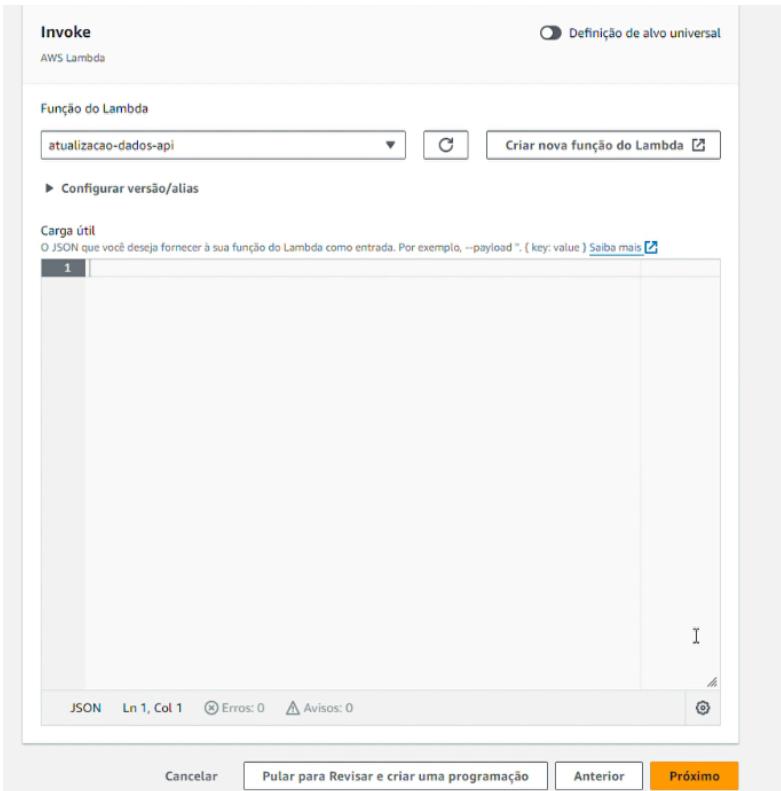


Figura 156: Invoke

Fonte: Autoria Própria.

Passo 9: Ação Após a Conclusão do Agendamento

- Selecione a opção "NONE".

A screenshot of the 'Ação após a conclusão do agendamento' (Action after scheduling completion) configuration section. It includes a 'Habilitar cronograma' (Enable cronogram) section with a toggle switch set to 'Habilitar' (Enable) and a note about enabling it later. Below this is a large section titled 'Ação após a conclusão do agendamento' with a dropdown menu currently set to 'NONE'. A note below the dropdown states: 'Se você escolher DELETE, o Agendador do EventBridge excluirá automaticamente o agendamento após concluir sua última invocação se não tiver nenhuma invocação de destino futura planejada.' (If you choose DELETE, the EventBridge Scheduler will automatically delete the scheduled execution after completing its last invocation if there is no future destination invocation planned.)

Figura 157: Invoke

Fonte: Autoria Própria.

Passo 10: Finalização do Cronograma

- Nome do perfil: Atualizacao-Dados-API-Cliente.

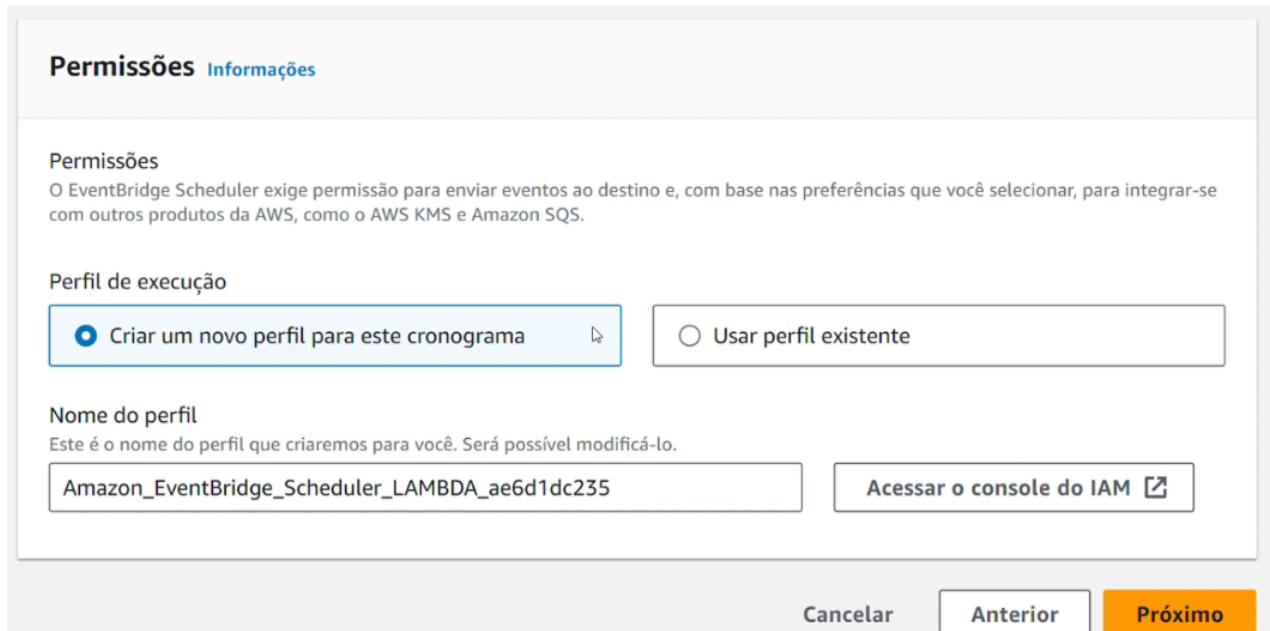


Figura 158: Cronograma

Fonte: Autoria Própria.

Passo 11: Clique em "Criar cronograma".

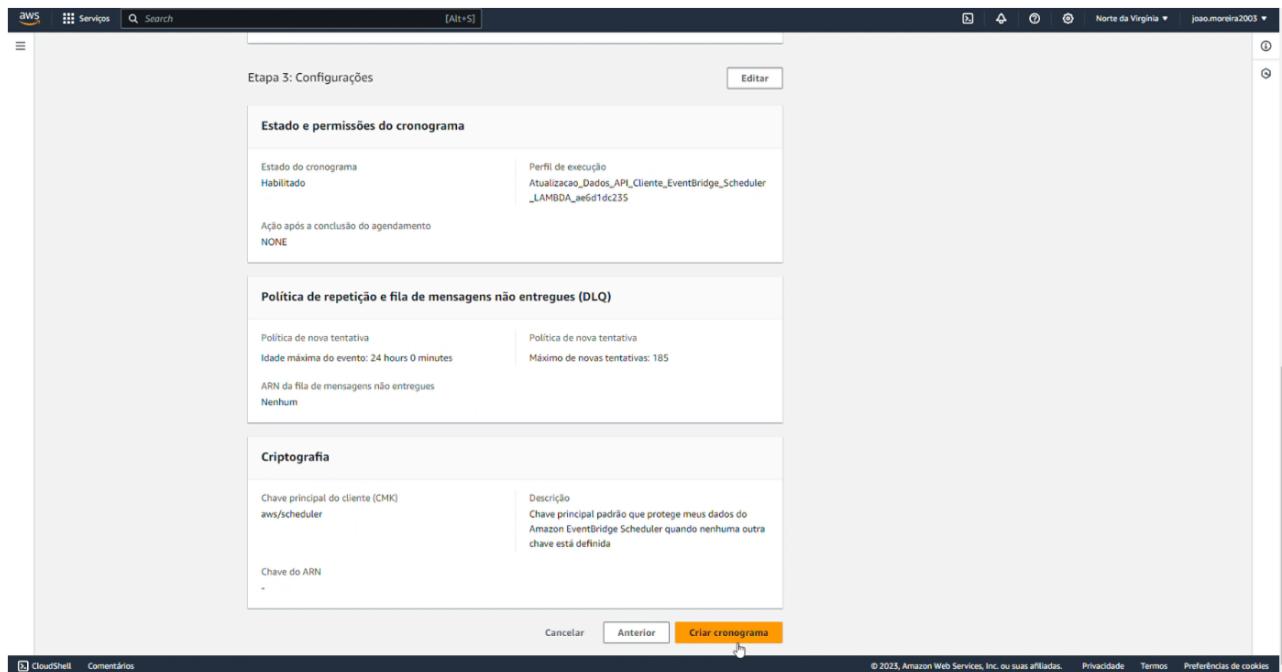


Figura 159: Configuração

Fonte: Autoria Própria.

Passo 12: Ao finalizar com êxito esta etapa, é exibida a notificação "Cronograma Criado".

The screenshot shows the AWS EventBridge console with a green header bar indicating 'Seu cronograma atualizacao-dados-api-bucket está sendo criado.' (Your cron schedule 'atualizacao-dados-api-bucket' is being created). The main area displays the 'Detalhes do cronograma' (Cron schedule details) for the schedule 'atualizacao-dados-api-bucket'. The schedule is enabled ('Habilitado') and has the ARN 'arn:aws:scheduler:us-east-1:1463714952561:schedule/default/atualizacao-dados-api-bucket'. It runs at 00:00 every day. The execution region is set to 'America/Sao_Paulo'. The creation date is Nov 29, 2023, at 16:37:46 UTC-03:00. The last modified date is also Nov 29, 2023, at 16:37:46 UTC-03:00. Below the details, there are tabs for 'Cronograma' (Schedule), 'Destino' (Destination), 'Política de nova tentativa' (New attempt policy), 'Fila de mensagens não entregues' (Unsent message queue), and 'Criptografia' (Encryption).

Figura 160: AWS Eventbridge

Fonte: Autoria Própria.

Passo 13: Configuração da Função Lambda atualizacao-dados-api

O código da função Lambda é responsável por processar e salvar os dados atualizados. Abaixo, disponibilizamos o código implementado:

```

import csv
from datetime import datetime, timedelta
import boto3

def save_to_csv(data, file_name):
    if not data:
        return

    fieldnames = list(data[0].keys())
    with open('/tmp/' + file_name, 'w', newline='', encoding='utf-8') as csvfile:
        writer = csv.DictWriter(csvfile, fieldnames=fieldnames)
        writer.writeheader()
        for row in data:
            writer.writerow(row)

def upload_to_aws(local_file, bucket, s3_file):
    s3 = boto3.client('s3')
    try:
        s3.upload_file(local_file, bucket, s3_file)
        return True
    except FileNotFoundError:
        return False
    except Exception as e:
        print(f"Erro ao enviar arquivo para o S3: {e}")
        return False

def process_and_save(table_name, date, csv_file_path, s3_bucket):
    s3 = boto3.client('s3')
    bucket = s3_bucket
    key = f"{table_name}/{date}.csv"

    try:
        data = s3.get_object_content(Bucket=bucket, Key=key)
        data = data.decode('utf-8')
        data = csv.DictReader(data.splitlines(), delimiter=',')
        save_to_csv(data, csv_file_path)
        print(f'Dados de {table_name} processados e salvos.')
    except Exception as e:
        print(f"Erro ao processar dados de {table_name}: {e}")

def lambda_handler(event, context):
    sale_date = (datetime.now() - timedelta(days=1)).strftime('%Y-%m-%d')

    categories = ['category', 'establishment', 'sale']
    s3_bucket = 'apiparceiro'

    for category in categories:
        csv_file_path = f'dados_{category}.csv'
        process_and_save(category, date=sale_date, csv_file_path=csv_file_path, s3_bucket=s3_bucket)

    return 'Dados processados e salvos com sucesso.'

```

Figura 161: Código

Fonte: Autoria Própria.

Passo 14: Teste da Função Lambda - Configuração do Evento de Teste:

- Na função Lambda, clique em "Test".
- Nome do evento de teste: atualizacao-dados-api.
- Compartilhamento: Defina como "Privado".
- JSON do evento: {}.

Execução do Teste: Clique em "Criar Teste" e execute o teste para validar o funcionamento da função.

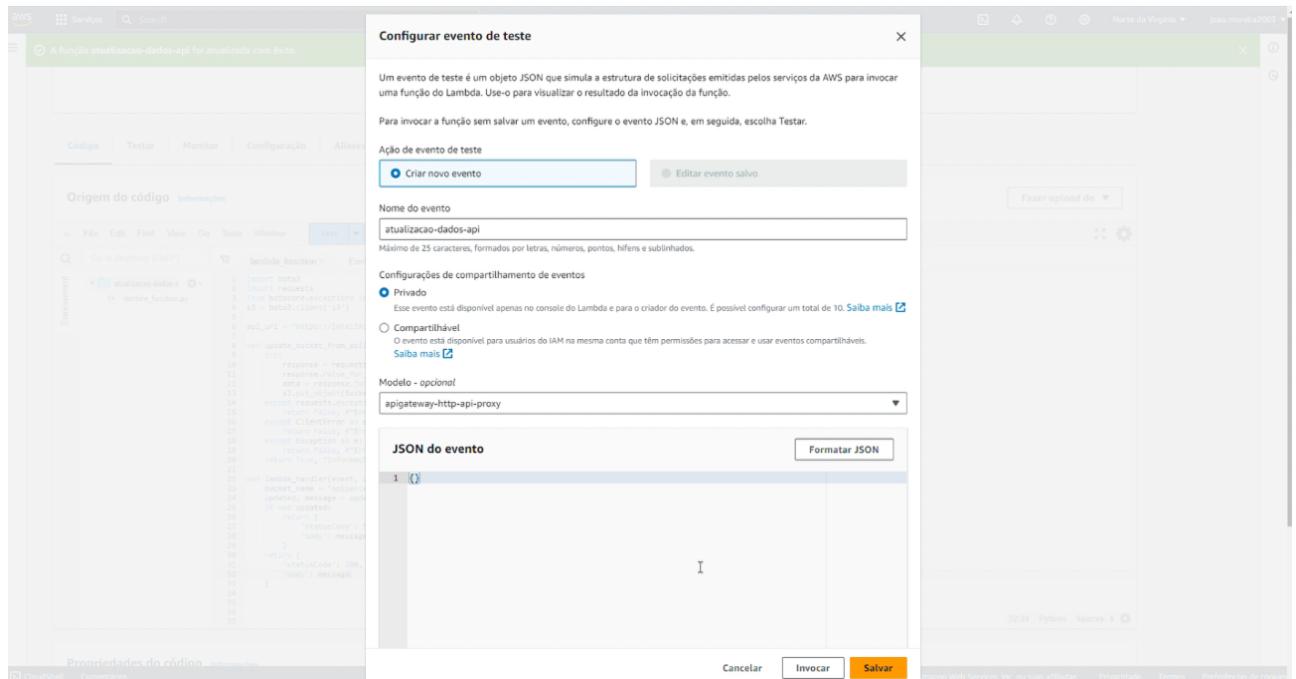


Figura 162: Evento - Teste

Fonte: Autoria Própria.

Resultados Esperados

Após a execução bem-sucedida do teste, os dados da API do parceiro são processados e salvos automaticamente no S3, conforme configurado na função Lambda. Este processo garante a atualização contínua e automática dos dados.

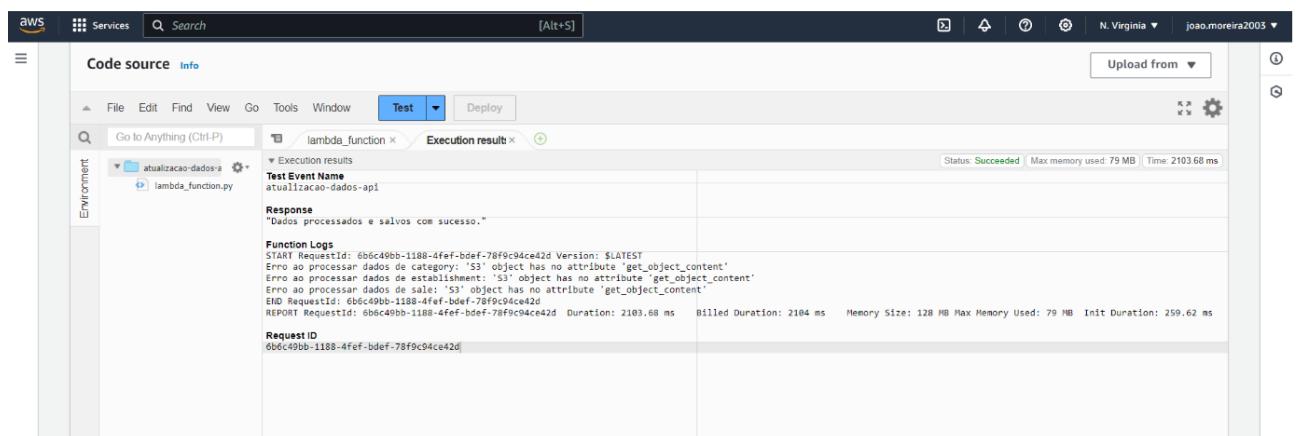


Figura 163: Resultados

Fonte: Autoria Própria.

11.7.3 Exclusão automatizada dos dados

Dado que as informações provenientes da API do parceiro são confidenciais, não é apropriado mantê-las em buckets públicos por períodos prolongados. Portanto, implementou-se uma exclusão automatizada desses dados.

Por motivos de segurança, a exclusão requer a apresentação, em formato JSON, do nome do bucket e da senha cadastrada para exclusão. Isso devido ao acesso concedido a vários colaboradores e à variação nos períodos de retenção dos dados, dependendo do projeto em questão. A seguir, exibe-se o processo de automação da exclusão, incluindo os códigos pertinentes e uma simulação do seu funcionamento.

Passo 1: Criação da AWS Lambda

Acesse o Console da AWS: Faça login na sua conta AWS e acesse o Console da AWS em <https://aws.amazon.com/>.

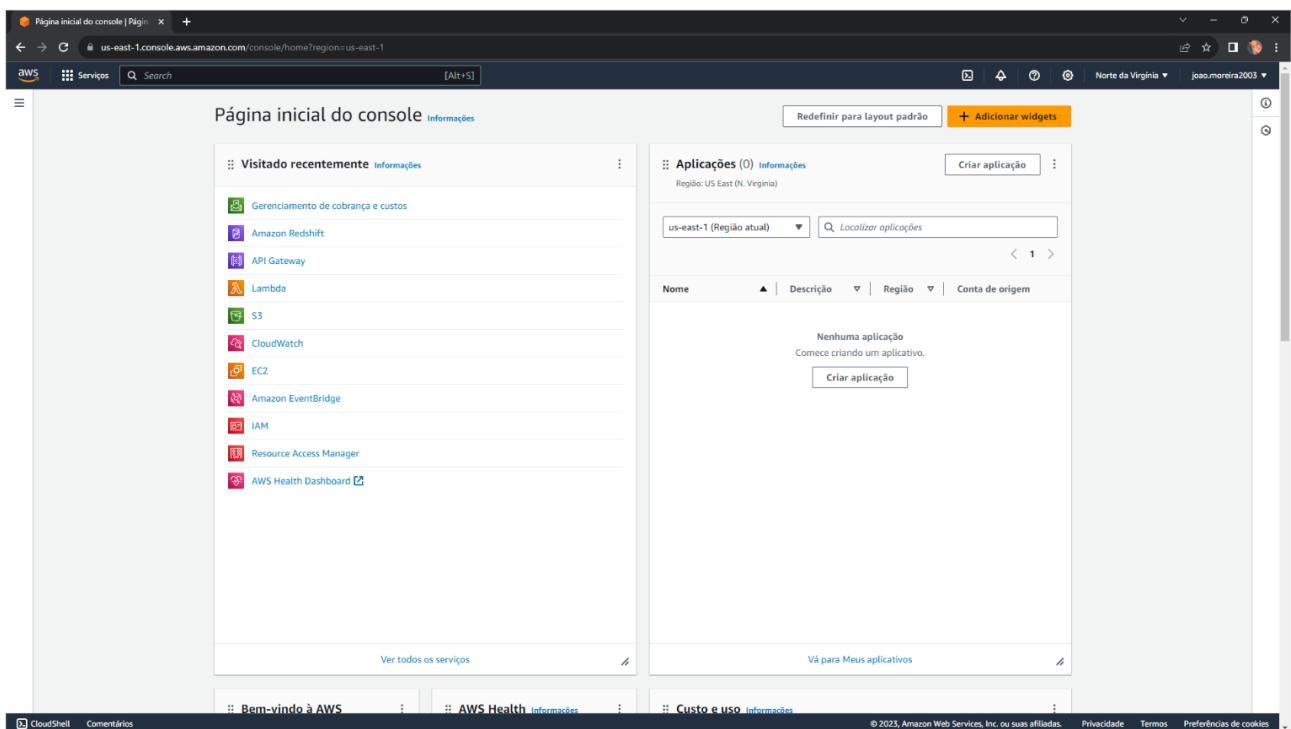


Figura 164: AWS Console

Fonte: Autoria Própria.

Passo 2: Navegue até o AWS Lambda - No Console da AWS, vá para o serviço "Lambda" localizado no menu "Serviços".

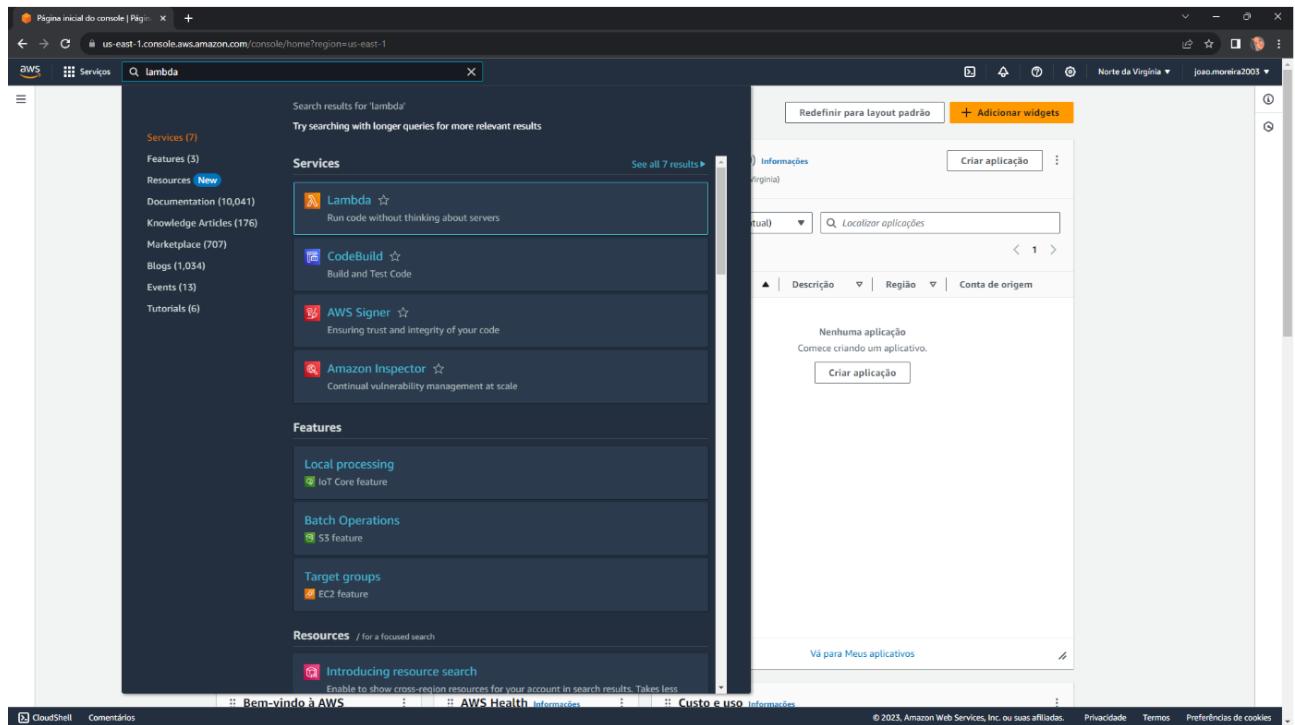


Figura 165: AWS Console - Lambda

Fonte: Autoria Própria.

Passo 3: Crie uma Nova Função Lambda - Clique em "Funções" no painel de navegação à esquerda. Em seguida, clique no botão "Criar função".

Figura 166: AWS Lambda

Fonte: Autoria Própria.

Passo 4: Configure os Detalhes da Função - Dê um nome para a função. Escolha uma runtime, como Node.js, Python, etc.

Figura 167: AWS Lambda

Fonte: Autoria Própria.

Passo 5: Reveja e Crie - Revise todas as configurações e clique em "Criar função" para criar sua função Lambda.

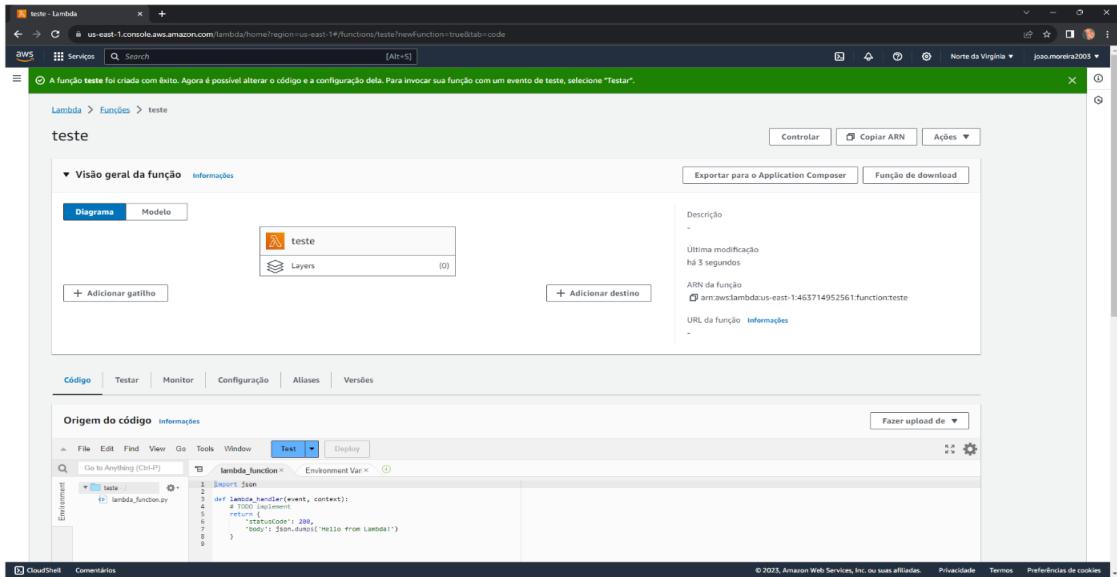


Figura 168: AWS Lambda

Fonte: Autoria Própria.

Passo 6: Adição do gatilho "API Gateway"

- Selecione sua Função Lambda:
 - Encontre e selecione a função Lambda à qual deseja adicionar o gatilho API Gateway.
- Adicione um Gatilho:
 - No tópico "Visão geral da função", clique em "Adicionar Gatilho";
 - Escolha "API Gateway" como o tipo de gatilho.

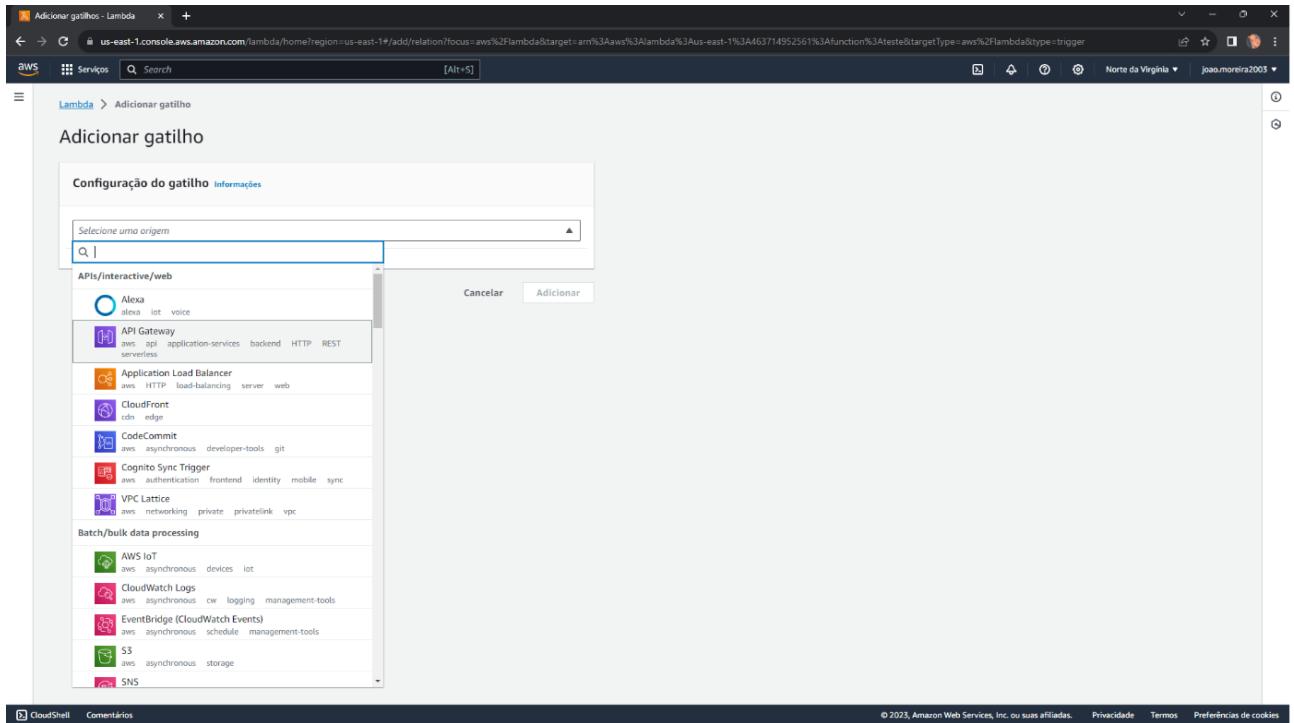


Figura 169: AWS Lambda

Fonte: Autoria Própria.

Passo 6: Configure o Gatilho API Gateway

- Escolha a opção "Criar uma nova API" ou selecione uma API Gateway existente,
- Escolha "REST API",
- Configure a segurança conforme necessário,
- Clique em "Adicionar";

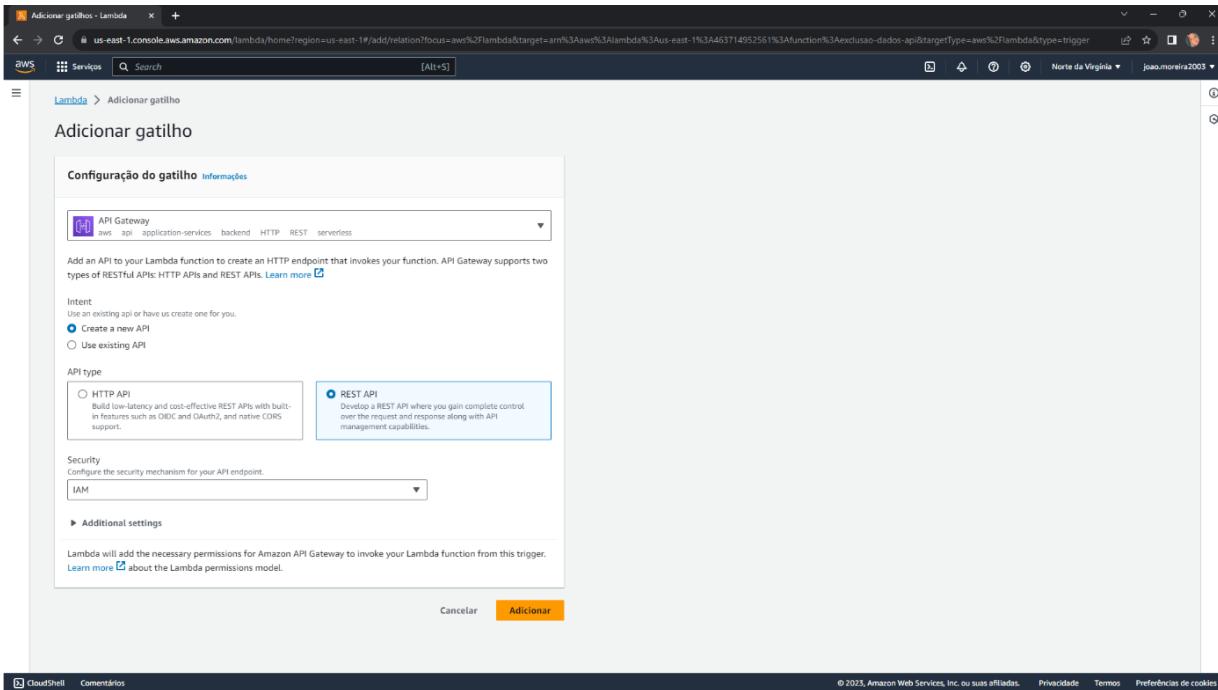


Figura 170: AWS Lambda

Fonte: Autoria Própria.

Passo 7: Salve as Configurações: Certifique-se de salvar as configurações da função Lambda após adicionar o gatilho.

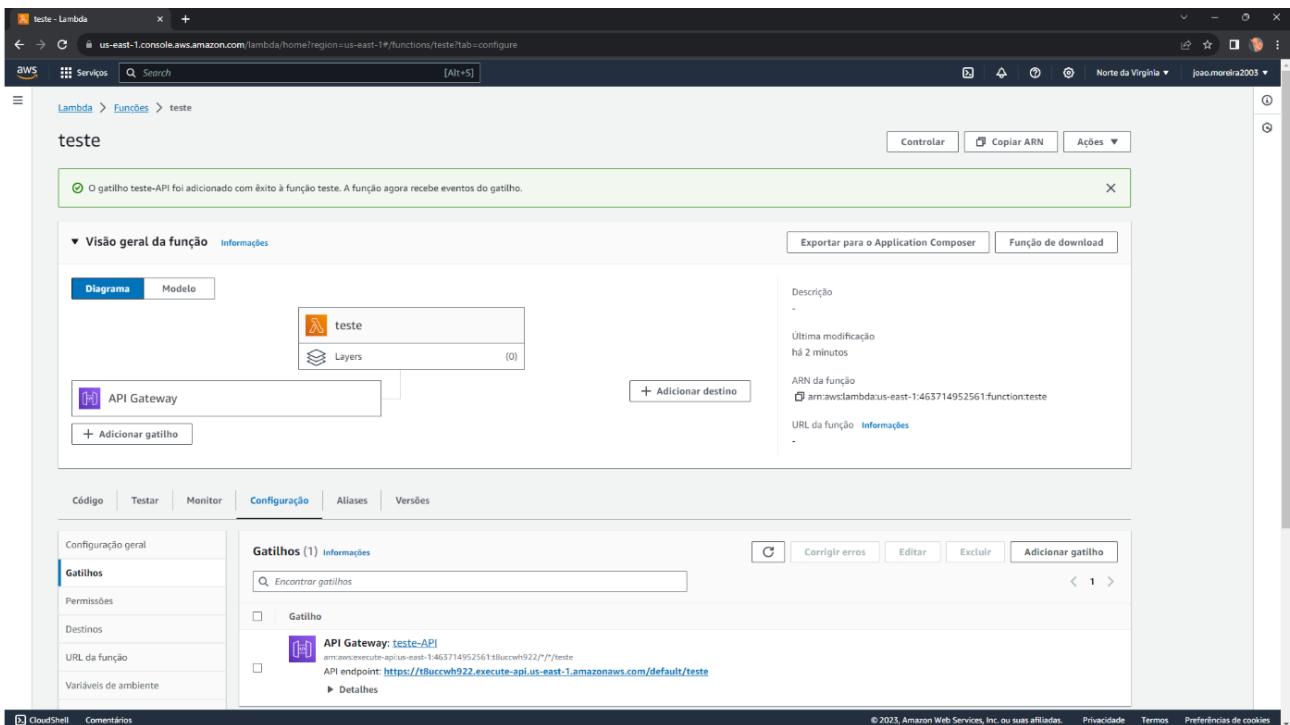


Figura 171: AWS Lambda

Fonte: Autoria Própria.

Passo 8: Criação do método:

- Clique no bloco "API Gateway" criado,

- Na guia "Configuração", clique na url com o nome da api criada,
- Na nova janela que abrir, clique no botão "Ações",
- Selecione "Criar método";

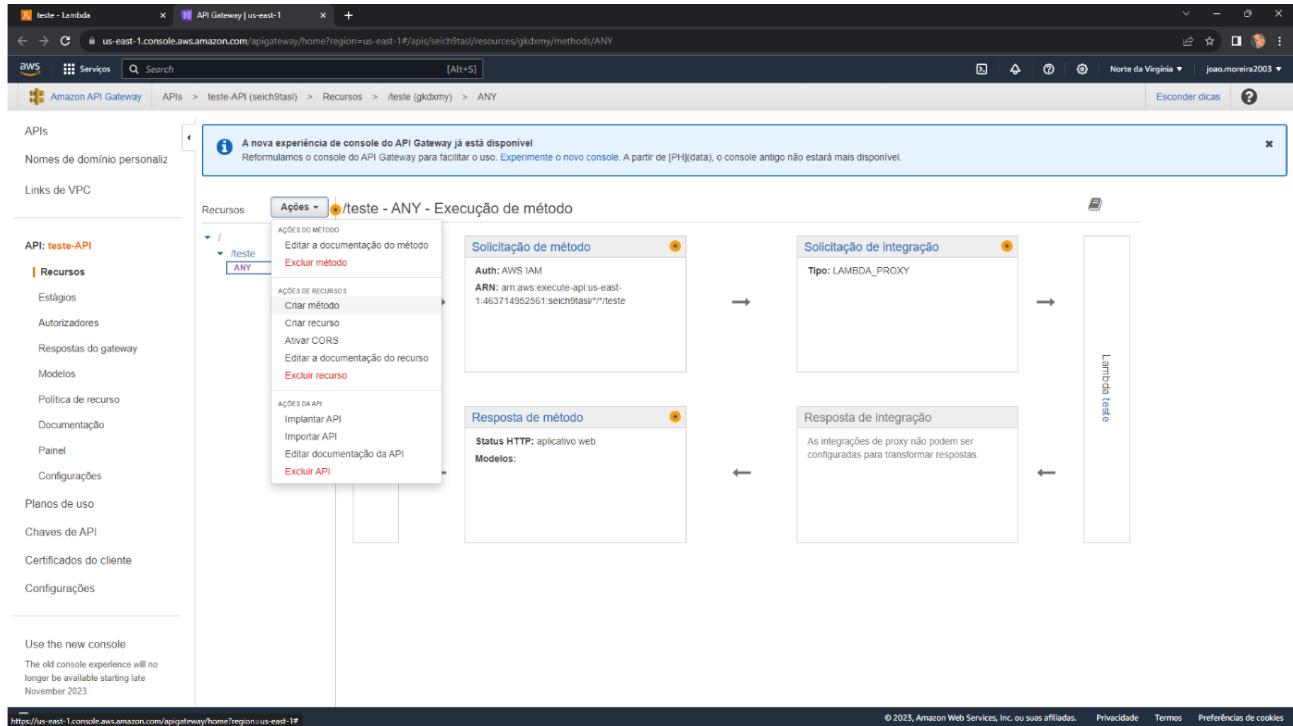


Figura 172: API Gateway

Fonte: Autoria Própria.

Passo 9: Configuração do método:

- Selecione o método "POST",
- Clique no ícone "✓" para confirmar a criação,
- No tópico "Tipo de integração" selecione "Função Lambda",
- Abra outra guia na página inicial da sua função lambda, e copie a "ARN da função",
- No tópico "Função Lambda" cole o "ARN da função" copiado,
- Clique no botão "Salvar";

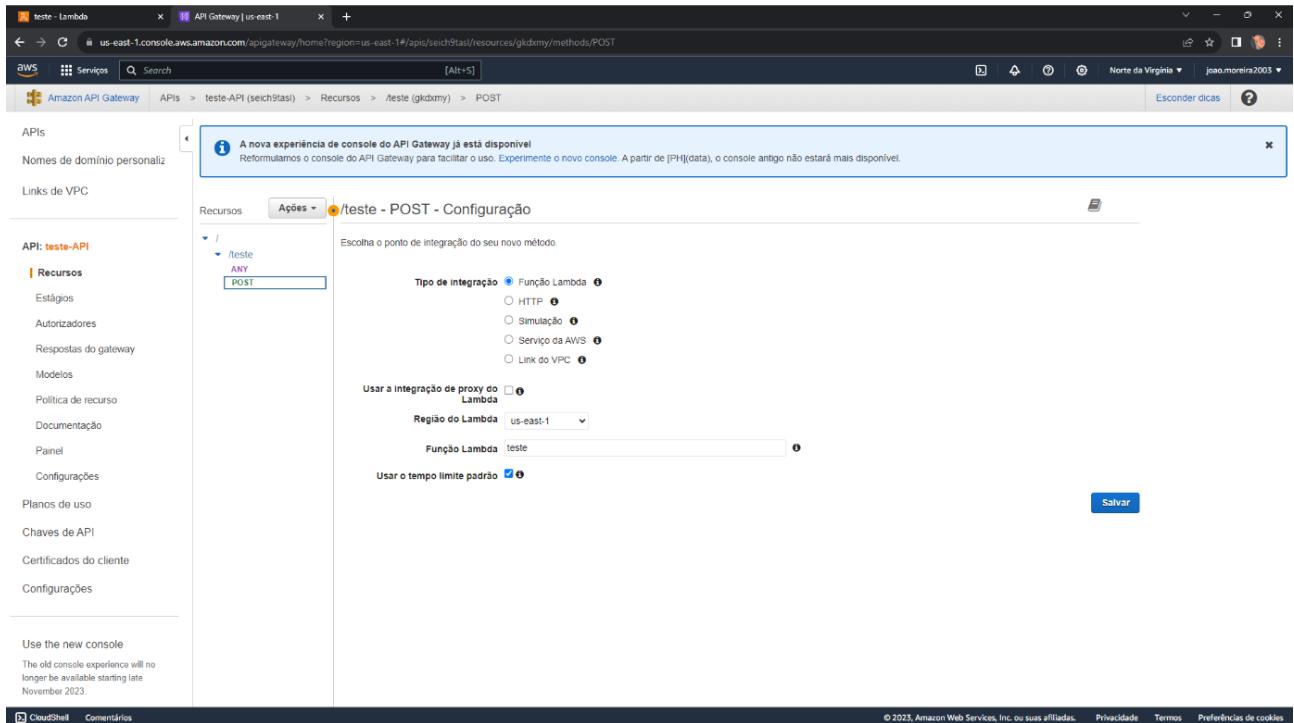


Figura 173: API Gateway

Fonte: Autoria Própria.

Passo 9: Criação do "Estágio"

- Encontre e selecione a API à qual deseja adicionar um estágio,
- No menu lateral, clique em "Estágios",
- Clique no botão "Criar",
- Dê um nome ao seu estágio,
- Adicione uma "Descrição do estágio",
- Selecione uma "Implantação" para associar ao estágio,
- Clique em "Criar";

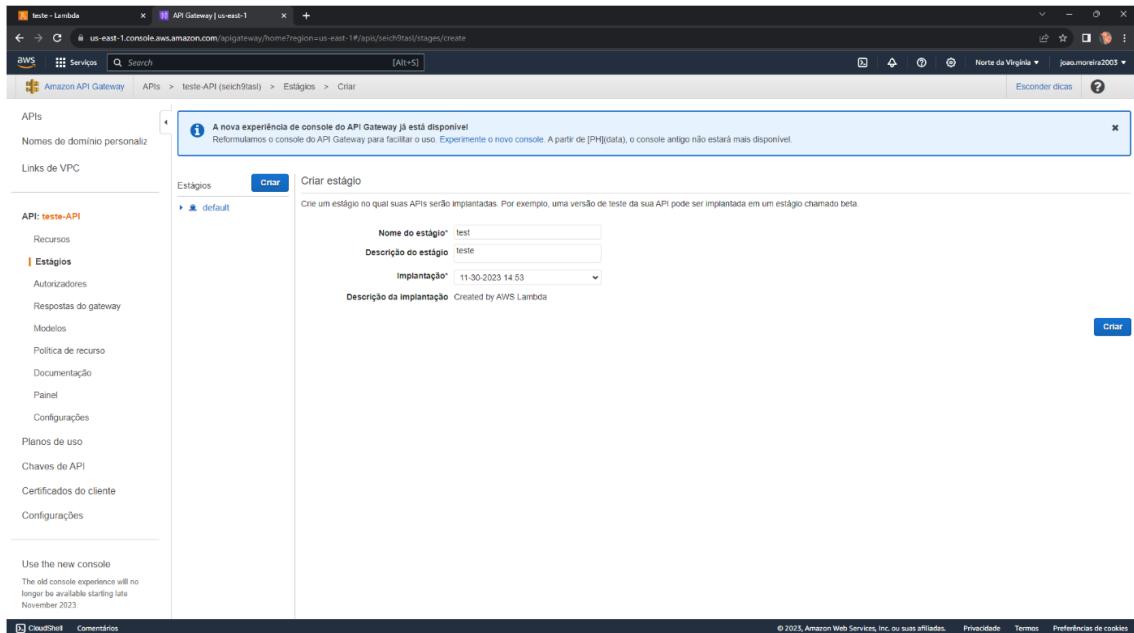


Figura 174: API Gateway

Fonte: Autoria Própria.

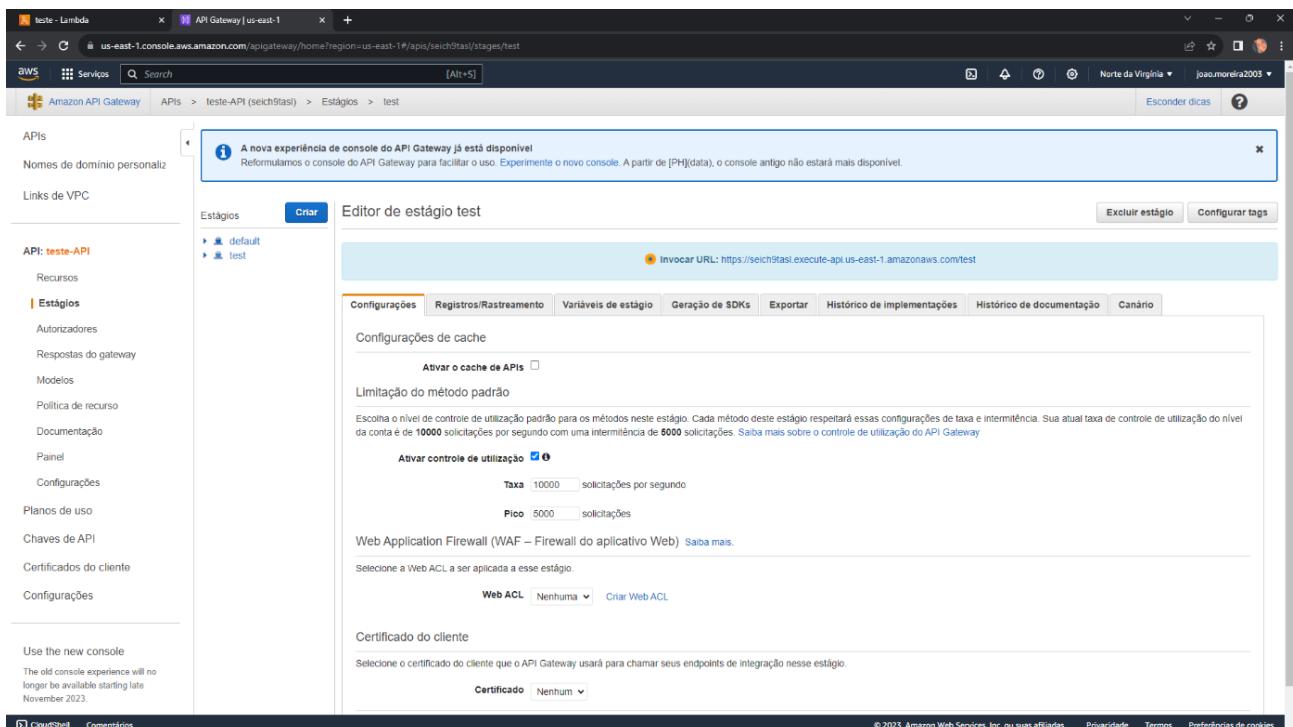


Figura 175: API Gateway

Fonte: Autoria Própria.

Passo 10: Implante as Mudanças:

- Retorne a aba "Recursos" e clique no método que está sendo utilizado,
- Clique em "Ações" e escolha "Implantar API".
- Escolha o "Estágio de implantação" desejado e clique em "Implante".

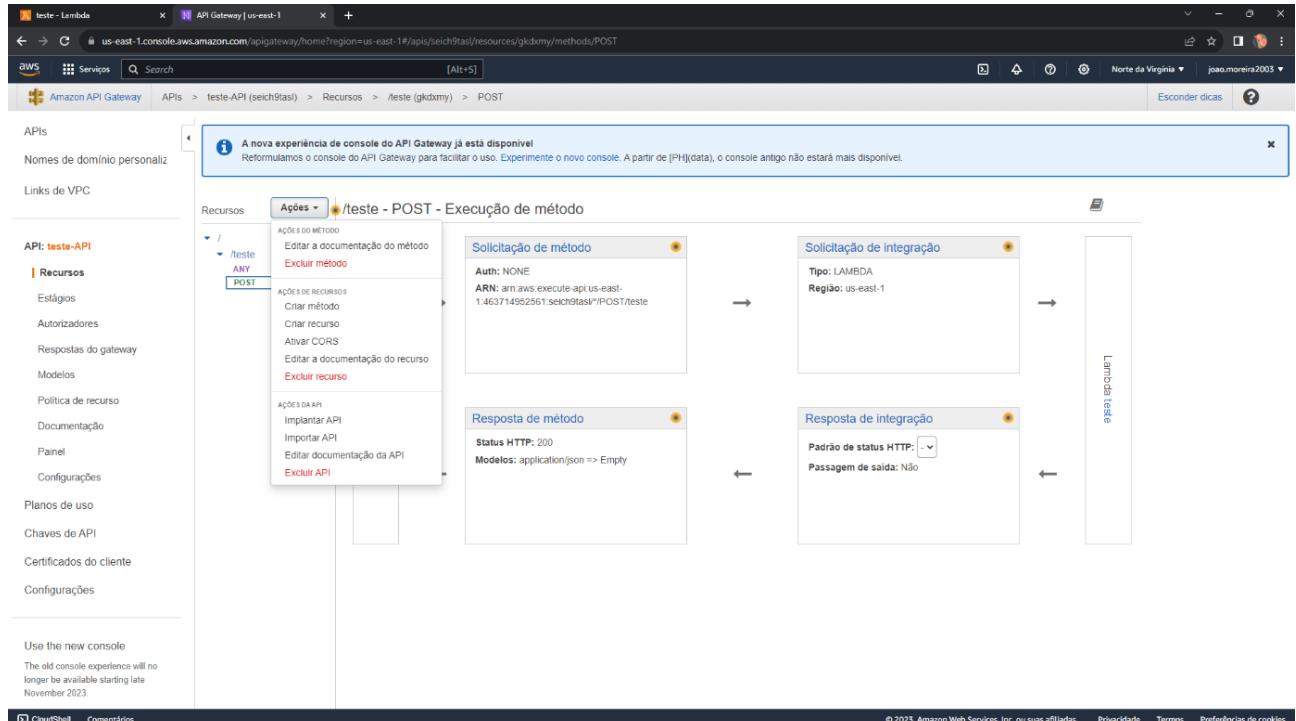


Figura 176: API Gateway

Fonte: Autoria Própria.

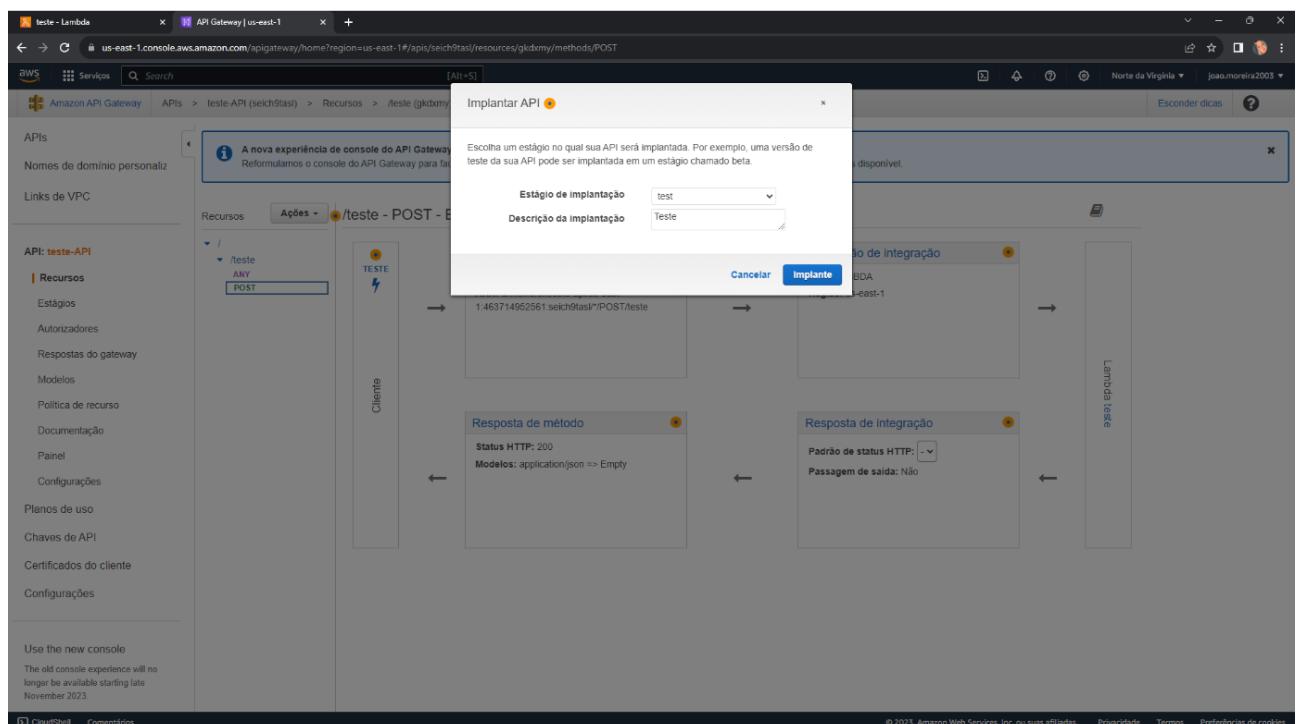


Figura 177: API Gateway

Fonte: Autoria Própria.

Passo 11: Obtenha a URL do Estágio:

- Na página de detalhes da API, clique na URL do estágio criado.
- Essa URL agora representa o endpoint da sua API no estágio especificado.

The screenshot shows the AWS API Gateway console with the URL <https://us-east-1.console.aws.amazon.com/apigateway/home?region=us-east-1#/apis/seich9tasl/stages/test>. The left sidebar shows the API named 'teste-API'. The 'Estágios' tab is selected, showing two stages: 'default' and 'test'. The 'test' stage is currently active. The main panel is titled 'Editor de estágio test' and displays the configuration for this stage. At the top, there is a message: 'A nova experiência do console do API Gateway já está disponível! Reformulamos o console do API Gateway para facilitar o uso. Experimente o novo console. A partir de [HH]data), o console antigo não estará mais disponível.' Below this, the 'Invoker URL' is listed as <https://seich9tasl.execute-api.us-east-1.amazonaws.com/test>. The configuration tabs include 'Configurações', 'Registros/Rastreamento', 'Variáveis de estágio', 'Geração de SDKs', 'Exportar', 'Histórico de implementações', 'Histórico de documentação', and 'Canário'. Under 'Configurações', there are sections for 'Configurações de cache' (with 'Ativar o cache de APIs' checked) and 'Limitação do método padrão' (with a note about a rate limit of 10,000 requests per second). There is also a section for 'Web Application Firewall (WAF – Firewall do aplicativo Web)' and a 'Web ACL' dropdown set to 'Nenhuma'. The 'Certificado do cliente' section shows a dropdown set to 'Nenhum'. The bottom of the page includes standard AWS footer links like CloudShell, Comentários, and links to terms and conditions.

Figura 178: API Gateway

Fonte: Autoria Própria.

Passo 12: Código "lambda_function.py"

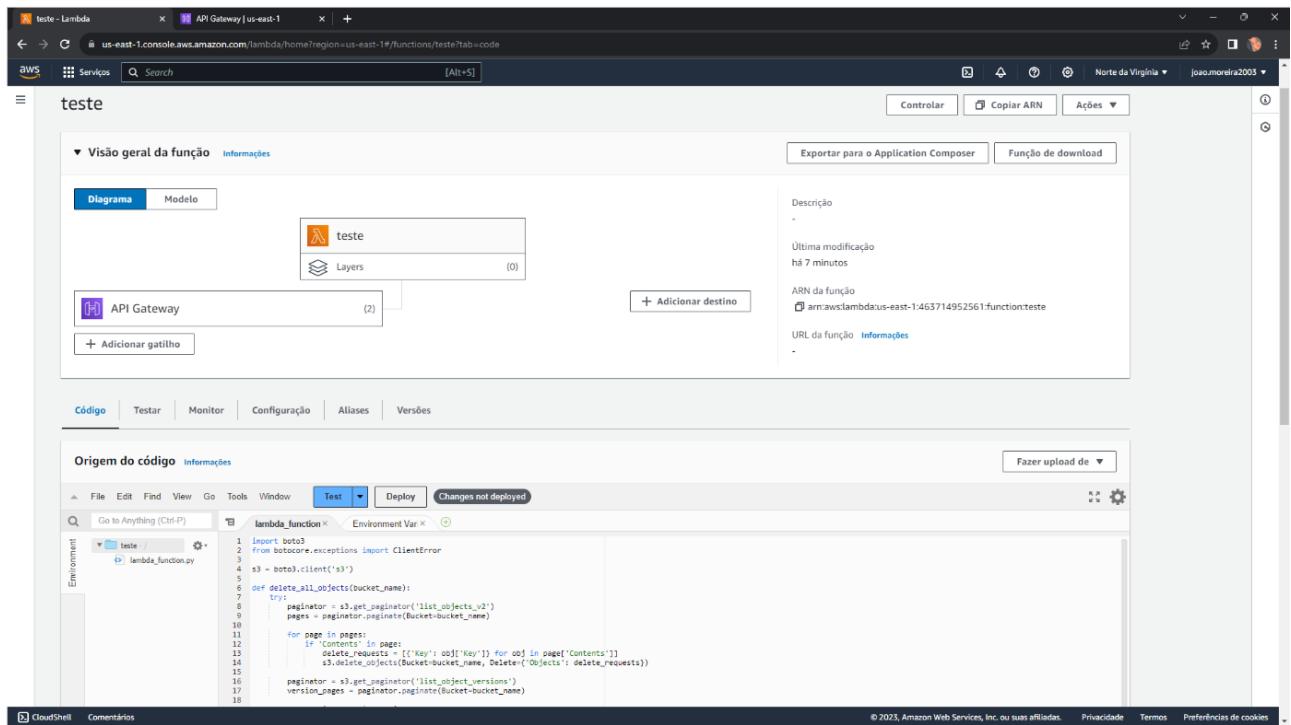


Figura 179: Função

Fonte: Autoria Própria.

```

import boto3
from botocore.exceptions import ClientError

s3 = boto3.client('s3')

def delete_all_objects(bucket_name):
    try:
        paginator = s3.getPaginator('list_objects_v2')
        pages = paginator.paginate(Bucket=bucket_name)

        for page in pages:
            if 'Contents' in page:
                delete_requests = [{'Key': obj['Key']} for obj in page['Contents']]
                s3.delete_objects(Bucket=bucket_name, Delete={'Objects': delete_requests})

        paginator = s3.getPaginator('list_object_versions')
        version_pages = paginator.paginate(Bucket=bucket_name)

        for version_page in version_pages:
            delete_requests = []
            if 'Versions' in version_page:
                delete_requests.extend([{'Key': v['Key'], 'VersionId': v['VersionId']} for v in version_page['Versions']])
            if 'DeleteMarkers' in version_page:
                delete_requests.extend([{'Key': v['Key'], 'VersionId': v['VersionId']} for v in version_page['DeleteMarkers']])
            if delete_requests:
                s3.delete_objects(Bucket=bucket_name, Delete={'Objects': delete_requests})

    except ClientError as e:
        return False, e.response['Error']['Message']
    return True, "Objects deleted"

def lambda_handler(event, context):
    bucket_name = event['bucket_name']
    password = event['password']

    if password == ' ': # Adicione a senha desejada

        objects_deleted, message = delete_all_objects(bucket_name)
        if not objects_deleted:
            return {
                'statusCode': 500,
                'body': message
            }
        return {
            'statusCode': 200,
            'body': "All objects in bucket '{}' deleted successfully".format(bucket_name)
        }
    else:
        return {
            'statusCode': 401,
            'body': "Unauthorized"
        }

```

Figura 180: Código

Fonte: Autoria Própria.

Resultados esperados

O formato do JSON esperado pelo endpoint no BODY é:

```
{  
    "bucket_name": "nomeBucket",  
    "password": "senhaCadastrada"  
}
```

11.8. Pacote Python

11.8.1 CNPJ

Entre os cinco arquivos específicos sobre CNPJ, é possível encontrar informações variadas, como dados cadastrais das empresas, histórico de alterações contratuais, situação cadastral, meio de contato. A combinação desses arquivos proporciona uma base sólida para a tomada de decisões estratégicas. A seguir, é apresentado as premissas e as restrições das funções utilizadas para tratar esses dados. Importante ressaltar que o código está presente neste repositório, na pasta src/scripts/ScriptCNPJ.

Arquivo lib.py

A função carregar_dados_delimitador_correto tem como objetivo carregar um arquivo CSV, permitindo a especificação do delimitador utilizado. Essa função opera em chunks para otimizar a leitura de grandes conjuntos de dados, lidando adequadamente com erros de parsing.

Premissa: existência de arquivo em formato CSV na pasta especificada no código, por exemplo: ./csv/nome_arquivo.csv'.

Além disso, a função não apresenta nenhum tipo de restrição ou dependência por ser a primeira função a ser aplicada no arquivo. Abaixo, é possível visualizar como o arquivo pode ser aplicado. Este, utiliza o ponto e vírgula como delimitador, como segundo argumento da função. O DataFrame resultante é armazenado na variável df.

```
df = funcao.carregar_dados_delimitador_correto('./csv/nome_arquivo.csv', ';')
```

A função limpar_dados realiza a limpeza de dados em um DataFrame, removendo caracteres especiais de determinadas colunas. A limpeza também pode ser efetuada em chunks, neste caso, se no momento da aplicação não for especificado o tamanho do chunks, ele vai utilizar o "padrão" dito na função. Como premissa é necessário que o arquivo tenha sido carregado, ou seja, aplicado na função carregar_dados_delimitador_correto, e a mesma variável definida anteriormente (df) seja aplicada na função. A segunda premissa é que as colunas devem ser passadas como arrays, já que neste caso é mais de uma. A única restrição dessa função é que deve ser uma variável a ser aplicada na função, e não um arquivo csv, por exemplo.

A aplicação da função limpar_dados é feita ao DataFrame df, removendo caracteres especiais e ignorando o array passado em colunas_a_ignorar, que incluem 'email', 'data', 'data_situacao_cadastral', 'data_inicio_atividade' e 'data_situacao_especial'. Essas colunas foram ignoradas já que os caracteres especiais significam algo, como o '@' na coluna 'email'.

```
df = funcao.limpar_dados(df, colunas_a_ignorar=['email', 'data',  
'data_situacao_cadastral', 'data_inicio_atividade', 'data_situacao_especial'])
```

A função tratar_valores_nulos aborda valores nulos em um DataFrame, substituindo-os por um valor especificado. O tratamento também pode ser efetuada em chunks, neste caso, se no momento da aplicação não for especificado o tamanho do chunks, ele vai utilizar o "padrão" dito na função. A premissa dessa função é utilizar o df criado na função anterior, além disso, é necessário definir um substituto para os valores nulos, neste exemplo: "nan". A única restrição dessa função é que deve ser uma variável a ser aplicada na função, e não um arquivo csv, por exemplo.

Aqui, a função tratar_valores_nulos é utilizada para substituir valores nulos no DataFrame (df) pelo valor especificado ('nan'). A operação é realizada em chunks de tamanho 100 para otimizar o tratamento em grandes conjuntos de dados.

```
df = funcao.tratar_valores_nulos(df, 'nan', chunk_size=1000)
```

A função `remover_coluna` remove uma coluna do DataFrame. A premissa dessa função é utilizar o df criado na função anterior, além disso, é necessário definir a coluna a ser removida, como mostra o exemplo. A única restrição dessa função é que deve ser uma variável a ser aplicada na função, e não um arquivo csv, por exemplo.

Abaixo, é empregado a função `remover_coluna` para excluir a coluna 'cnae_fiscal_secundaria' do DataFrame df. Essa coluna foi excluída pois não foi encontrada nenhuma utilidade no momento para ela. Note que a coluna deve ser passada em ".

```
df = funcao.remover_coluna(df, 'cnae_fiscal_secundaria')
```

A função `ajustar_amazon_s3` é projetada para ajustar um DataFrame antes de carregá-lo no Amazon S3. A premissa desta função é receber como primeiro argumento a variável aplicada na função anterior e é necessário definir um caminho para que a função crie esse arquivo CSV, como mostra o código abaixo. Além disso, a restrição é que deve ser uma variável a ser aplicada na função, e não um arquivo csv, por exemplo.

A aplicação da função `ajustar_amazon_s3` é feita para criar um arquivo CSV na pasta `./csv_s3/cnpj_5_s3.csv`.

```
funcao.ajustar_amazon_s3(df, './csv_s3/cnpj_5_s3.csv')
```

Aplicação do arquivo `send_s3.ipynb`

O script em Python apresentado tem como objetivo automatizar a transferência de arquivos CSV de um diretório local para o Amazon S3, seguido pela exclusão desses arquivos locais. Na primeira parte, são definidas variáveis que armazenam as credenciais de acesso à AWS, a região do Amazon S3, os diretórios locais dos arquivos CSV e o nome do bucket S3. Abaixo é demonstrado uma tabela que explica as variáveis do código abaixo.

Nome da variável	Explicação	Premissa	Onde deve ser preenchido
aws_access_key_id	String alfanumérica única que identifica uma conta ou usuário AWS para propósitos de autenticação	Deve ter uma conta na AWS com as credenciais criadas ou acessar no AWSLabs	Arquivo <code>.env tmpl</code> por questão de segurança
aws_secret_access_key	Chave secreta utilizada em conjunto com a anterior para assinar as solicitações	Deve ter uma conta na AWS com as credenciais criadas ou acessar no AWSLabs	Arquivo <code>.env tmpl</code> por questão de segurança
aws_session_token	Componente das credenciais temporárias	Deve ter uma conta na AWSLabs	Arquivo <code>.send_s3</code>
region_name	Identifica a região da AWS onde você deseja que seus recursos e operações estejam localizados	Deve ser a mesma região do bucket	Arquivo <code>.send_s3</code>
bucket_name	Nome do Bucket que você deseja inserir na AWS	O Bucket deve estar criado na mesma conta das credenciais	Arquivo <code>.send_s3</code>
csv_directory_s3	Diretório que se localiza os arquivos tratados	Os arquivos tratados devem estar neste diretório	Arquivo <code>.send_s3</code>
csv_directory	Diretório que se localiza os arquivos iniciais	Os arquivos devem estar neste diretório	Arquivo <code>.send_s3</code>
s3	Criação de um cliente para o serviço Amazon S3 usando o Boto3	Preenchimento de todas as variáveis acima	Arquivo <code>.send_s3</code>

Figura 181: Tabela do Send

Fonte: Autoria Própria.

Por último, é definida uma função `delete_csv_files` para excluir todos os arquivos CSV de um diretório local. A função é então chamada para os diretórios `csv_directory_s3` e `csv_directory`.

```
def delete_csv_files(directory):
    csv_files = [file for file in os.listdir(directory) if file.endswith('.csv')]
    for csv_file in csv_files:
        file_path = os.path.join(directory, csv_file)
        os.remove(file_path)
        print(f'O arquivo {csv_file} foi excluído de {directory}.')
```

11.8.2 DataSUS

O DataSUS é o departamento de informática do Sistema Único de Saúde (SUS) no Brasil e é responsável por coletar, processar e disseminar informações sobre saúde no país. Entre as inúmeras bases de dados mantidas pelo DataSUS, duas categorias importantes para o projeto são: ocupação de leitos e os arquivos de leitos em si. Importante ressaltar que o código está presente neste repositório, na pasta `src/scripts/ScriptDataSUS`.

Arquivo lib.py

A função carregar_csv carregar um arquivo CSV a partir de um caminho especificado. O parâmetro opcional delimiter permite a personalização do delimitador utilizado no arquivo CSV.

Premissa: existência de arquivo em formato CSV na pasta especificada no código, por exemplo: ./csv/nome_arquivo.csv'.

Além disso, a função não apresenta nenhum tipo de restrição ou dependência por ser a primeira função a ser aplicada no arquivo. Abaixo, é possível visualizar como o arquivo pode ser aplicado. O DataFrame resultante é armazenado na variável df.

```
df = funcao.carregar_csv('./csv/nome_arquivo.csv') # Se o delimitador for ',',  
não precisa especificar na aplicação
```

A função acima é utilizada nos arquivos de 'Leito' e de 'Leito Ocupação', a única exceção é 'Leitos 2023' que precisa ser aplicado com a função abaixo.

A função carregar_csv_com_codificacao carregar um arquivo CSV a partir de um caminho especificado, mas oferece suporte a diferentes codificações. Ela tenta carregar o arquivo usando várias codificações comuns até encontrar a que funciona. O parâmetro adicional chunk_size permite o carregamento em blocos, útil para arquivos muito grandes.

Premissa: existência de arquivo em formato CSV na pasta especificada no código, por exemplo: ./csv/nome_arquivo.csv'.

Além disso, a função não apresenta nenhum tipo de restrição ou dependência por ser a primeira função a ser aplicada no arquivo. Abaixo, é possível visualizar como o arquivo pode ser aplicado. O DataFrame resultante é armazenado na variável df.

```
df = funcao.carregar_csv_com_codificacao('./csv/nome_arquivo.csv')
```

A função remover_coluna remove uma coluna do DataFrame.

Premissa: utilizar o df criado na função anterior, além disso, é necessário definir a coluna a ser removida, como mostra o exemplo.

A única restrição dessa função é que deve ser uma variável a ser aplicada na função, e não um arquivo csv, por exemplo. Abaixo, é empregado a função remover_coluna para excluir a coluna 'cnae_fiscal_secundaria' do DataFrame df. Essa coluna foi excluída pois não foi encontrada nenhuma utilidade no momento para ela. Note que a coluna deve ser passada em ''.

```
df = funcao.remover_coluna(df, 'cnae_fiscal_secundaria')
```

A função limpar_dados realiza a limpeza de dados em um DataFrame, removendo caracteres especiais de determinadas colunas. A limpeza também pode ser efetuada em chunks, neste caso, se no momento da aplicação não for especificado o tamanho do chunks, ele vai utilizar o "padrão" dito na função.

Premissa: é necessário que o arquivo tenha sido carregado, ou seja, aplicado na função carregar_dados_delimitador_correto, e a mesma variável definida anteriormente (df) seja aplicada na função. A segunda premissa é que as colunas devem ser passadas como arrays, já que neste caso é mais de uma.

A única restrição dessa função é que deve ser uma variável a ser aplicada na função, e não um arquivo csv, por exemplo. A aplicação da função limpar_dados é feita ao DataFrame df, removendo caracteres especiais e ignorando o array passado em colunas_a_ignorar, que incluem 'email', 'data', 'data_situacao_cadastral', 'data_inicio_atividade' e 'data_situacao_especial'. Essas colunas foram ignoradas já que os caracteres especiais significam algo, como o '@' na coluna 'email'.

```
df = funcao.limpar_dados(df, colunas_a_ignorar=[ 'dataNotificacao' , '_created_at' , '_updated_at' , 'origem' , '_id' , '_p_usuario'])
```

A função tratar_valores_nulos aborda valores nulos em um DataFrame, substituindo-os por um valor especificado. O tratamento também pode ser efetuada em chunks, neste caso, se no momento da aplicação não for especificado o tamanho do chunks, ele vai utilizar o "padrão" dito na função.

Premissa: utilizar o df criado na função anterior, além disso, é necessário definir um substituto para os valores nulos, neste exemplo: "nan".

A única restrição dessa função é que deve ser uma variável a ser aplicada na função, e não um arquivo csv, por exemplo. Aqui, a função tratar_valores_nulos é utilizada para substituir valores nulos no DataFrame (df) pelo valor especificado ('nan'). A operação é realizada em chunks de tamanho 100 para otimizar o tratamento em grandes conjuntos de dados.

```
df = funcao.tratar_valores_nulos(df, 'nan', chunk_size=1000)
```

Além disso, é aplicada uma função da biblioteca 'pandas' .split("). Nesta caso, a função substitui os valores na coluna por suas versões com a parte da hora removida, mantendo apenas a parte da data.

Premissa: utilizar o df criado na função anterior, além disso, é necessário definir qual é o caractere que separa as duas informações.

A única restrição dessa função é que deve ser uma variável a ser aplicada na função, e não um arquivo csv, por exemplo. Aqui, a função .split(") é utilizada na coluna de data.

```
df[ 'dataNotificacao' ] = df[ 'dataNotificacao' ].str.split('T').str[0]
```

A função ajustar_amazon_s3 é projetada para ajustar um DataFrame antes de carregá-lo no Amazon S3. A premissa desta função é receber como primeiro argumento a variável aplicada na função anterior e é necessário definir um caminho para que a função crie esse arquivo CSV, como mostra o código abaixo. Além disso, a restrição é que deve ser uma variável a ser aplicada na função, e não um arquivo csv, por exemplo.

A aplicação da função ajustar_amazon_s3 é feita para criar um arquivo CSV na pasta ./csv_s3/cnpj_5_s3.csv.

```
funcao.ajustar_amazon_s3(df, './csv_s3/cnpj_5_s3.csv')
```

Aplicação do arquivo send_s3.ipynb

O script em Python apresentado tem como objetivo automatizar a transferência de arquivos CSV de um diretório local para o Amazon S3, seguido pela exclusão desses arquivos locais. A explicação deste arquivo se deu no tópico anterior.

11.8.3 IBGE

O Instituto Brasileiro de Geografia e Estatística (IBGE) é uma instituição responsável por coletar, analisar e divulgar informações estatísticas sobre o Brasil. No que diz respeito ao Produto Interno Bruto (PIB), o IBGE realiza pesquisas e censos econômicos que fornecem uma visão abrangente da atividade econômica do país. Além do PIB, o IBGE também é responsável por calcular o Índice de Gini, uma medida de desigualdade econômica que avalia a distribuição de renda em uma sociedade. O Índice

de Gini varia de 0 a 1, sendo 0 representativo de uma distribuição totalmente igualitária, enquanto 1 indica extrema desigualdade. Importante ressaltar que o código está presente neste repositório, na pasta src/scripts/ScriptIBGE.

Arquivo lib.py

A função `load_data_with_correct_delimiter` carregar um arquivo CSV a partir de um caminho especificado. OTenta inicialmente ler o arquivo usando o delimitador ';'. Se ocorrer um erro de análise, assume que o delimitador real é ',' e substitui todas as ocorrências de ';' por ',' no conteúdo do arquivo antes de tentar novamente a leitura.

Premissa: existência de arquivo em formato CSV na pasta especificada no código, por exemplo: `./csv/nome_arquivo.csv`.

Além disso, a função não apresenta nenhum tipo de restrição ou dependência por ser a primeira função a ser aplicada no arquivo. Abaixo, é possível visualizar como o arquivo pode ser aplicado. O DataFrame resultante é armazenado na variável `df`.

```
df = funcao.load_data_with_correct_delimiter("./csv/nome_arquivo.csv")
```

A função `clean_data` recebe um DataFrame como entrada e retorna um novo 'df' após a remoção de linhas que contêm valores nulos. Esta função é utilizada somente porque o modelo do arquivo do IBGE contém uma fonte no final do arquivo, e devemos excluir essa parte.

Premissa: utilizar o `df` criado na função anterior.

A única restrição dessa função é que deve ser uma variável a ser aplicada na função, e não um arquivo csv, por exemplo. Abaixo, é empregado a função `clean_data` para excluir os valores nulos encontrados.

```
df = funcao.clean_data(df)
```

A função `ajustar_amazon_s3` é projetada para ajustar um DataFrame antes de carregá-lo no Amazon S3. A premissa desta função é receber como primeiro argumento a variável aplicada na função anterior e é necessário definir um caminho para que a função crie esse arquivo CSV, como mostra o código abaixo. Além disso, a restrição é que deve ser uma variável a ser aplicada na função, e não um arquivo csv, por exemplo.

A aplicação da função ajustar_amazon_s3 é feita para criar um arquivo CSV na pasta ./csv_s3/cnpj_5_s3.csv.

```
funcao.ajustar_amazon_s3(df, './csv_s3/cnpj_5_s3.csv')
```

Aplicação do arquivo send_s3.ipynb

O script em Python apresentado tem como objetivo automatizar a transferência de arquivos CSV de um diretório local para o Amazon S3, seguido pela exclusão desses arquivos locais. A explicação deste arquivo se deu no tópico anterior.

11.8.4 POF

A Pesquisa de Orçamentos Familiares (POF), conduzida pelo Instituto Brasileiro de Geografia e Estatística (IBGE), é uma iniciativa fundamental para compreender os padrões de consumo e a estrutura orçamentária das famílias brasileiras. Realizada periodicamente, a POF coleta dados detalhados sobre os gastos das famílias em diversos itens, como alimentação, habitação, transporte, saúde e educação. Importante ressaltar que o código está presente neste repositório, na pasta src/scripts/ScriptPOF.

Arquivo lib.py

A função load_data_with_correct_delimiter renomeia as colunas em um DataFrame, utilizando um arquivo de mapeamento. O DataFrame é carregado a partir de um arquivo CSV, juntamente com um segundo DataFrame que contém um mapeamento entre os códigos das variáveis e suas traduções correspondentes.

Premissa: existência dos arquivos em formato CSV e devem ser correspondidos em variáveis chamadas na função.

Além disso, a função não apresenta nenhum tipo de restrição ou dependência por ser a primeira função a ser aplicada no arquivo. Abaixo, é possível visualizar como o arquivo pode ser aplicado. O DataFrame resultante é armazenado na variável df.

```
arquivo_csv = './csv/nome_arquivo.csv' # Arquivo da POF
```

```
arquivo_excel = './tradutores/Dicionários de variáveis.xlsx' # Dicionário das variáveis - utilizado depois

arquivo_mapeamento = './tradutores/traducao_domicilio.csv' # Dicionário das colunas

column_name = 'Categorias' # Utilizado depois

df_traduzido = funcao.traduzir_nomes_colunas(arquivo_csv, arquivo_mapeamento)
```

A função `encontrar_codigos_com_branco` recebe um DataFrame como entrada e retorna um novo 'df' após o encontro dos valores nulos, e com esse valores nulos ele substitui para os valores especificados no Dicionário (`arquivo_excel`).

Premissa: utilizar o df criado na função anterior.

A única restrição dessa função é que deve ser uma variável a ser aplicada na função, e não um arquivo csv, por exemplo. Abaixo, é empregado a função `encontrar_codigos_com_branco` para excluir os valores nulos encontrados.

```
df = funcao.preencher_valores_nulos_csv(df, arquivo_excel, column_name)
```

A função `encontrar_codigos_com_branco` aborda valores nulos em um DataFrame, substituindo-os por um valor especificado.

Premissa: utilizar o df criado na função anterior, além disso, é necessário definir um substituto para os valores nulos, neste exemplo: "N/A".

A única restrição dessa função é que deve ser uma variável a ser aplicada na função, e não um arquivo csv, por exemplo. Aqui, a função `tratar_valores_nulos` é utilizada para substituir valores nulos no DataFrame (df) pelo valor especificado ('N/A').

```
df_nulos = funcao.tratar_valores_nulos(df_com_nulos_preenchidos, 'N/A')
```

A função `translate_numeric_values_in_df_uf` traduz os valores numéricos em uma coluna específica de um DataFrame, utilizando um arquivo CSV de mapeamento para os estados brasileiros. Ao mapear os códigos numéricos para os estados correspondentes, ela cria uma nova coluna no DataFrame original com os valores traduzidos, facilitando a interpretação dos dados.

Premissa: utilizar o df criado na função anterior, além disso, é necessário especificar o caminho do arquivo com a tradução. Essa função deve ser aplicada em todos os arquivos da POF.

A única restrição dessa função é que deve ser uma variável a ser aplicada na função, e não um arquivo csv, por exemplo. Aqui, a função `translate_numeric_values_in_df_uf` é utilizada para traduzir os estados brasileiros.

```
df = funcao.translate_numeric_values_in_df(df, './tradutores/traducao_uf.csv',
'unidade_federativa')
```

A função `translate_numeric_values_in_df_alimento` traduz os valores numéricos em uma coluna específica de um DataFrame, utilizando um arquivo CSV de mapeamento para os códigos dos alimentos. Ao mapear os códigos numéricos para os alimentos correspondentes, ela cria uma nova coluna no DataFrame original com os valores traduzidos, facilitando a interpretação dos dados.

Premissa: utilizar o df criado na função anterior, além disso, é necessário especificar o caminho do arquivo com a tradução. Essa função deve ser aplicada somente no arquivo de consumo_alimentar.

A única restrição dessa função é que deve ser uma variável a ser aplicada na função, e não um arquivo csv, por exemplo. Aqui, a função `translate_numeric_values_in_df_alimento` é utilizada para traduzir os estados brasileiros.

```
df = funcao.translate_numeric_values_in_df_alimento(df, './tradutores/Produtos
do Consumo Alimentar.csv', 'codigo_tipo_alimento')
```

A função ajustar_amazon_s3 é projetada para ajustar um DataFrame antes de carregá-lo no Amazon S3. A premissa desta função é receber como primeiro argumento a variável aplicada na função anterior e é necessário definir um caminho para que a função crie esse arquivo CSV, como mostra o código abaixo. Além disso, a restrição é que deve ser uma variável a ser aplicada na função, e não um arquivo csv, por exemplo.

A aplicação da função ajustar_amazon_s3 é feita para criar um arquivo CSV na pasta ./csv_s3/cnpj_5_s3.csv.

```
funcao.ajustar_amazon_s3(df, './csv_s3/cnpj_5_s3.csv')
```

Aplicação do arquivo send_s3.ipynb

O script em Python apresentado tem como objetivo automatizar a transferência de arquivos CSV de um diretório local para o Amazon S3, seguido pela exclusão desses arquivos locais. A explicação deste arquivo se deu no tópico 1.2 deste mesmo documento.

12. ETL

12.1 Mapeamento do Fluxo

O processo ETL, Extract, Transform, Load (Extrair, Transformar, Carregar), é utilizado no gerenciamento de dados, especialmente em ambientes de data warehousing. Ao lidar com informações provenientes de diversas fontes, como arquivos CSV e APIs, a execução eficiente do ETL é crucial para garantir a qualidade e a utilidade dos dados no contexto de um projeto.

12.1.1 Extração

Na fase inicial, a extração é realizada a partir de fontes heterogêneas, abrangendo desde bancos de dados SQL e NoSQL até arquivos CSV e sistemas de CRM ou ERP. É imperativo atentar para a integridade dos dados durante esse processo, considerando a possibilidade de inconsistências ou erros nas fontes. Essa etapa estabelece a base para as transformações subsequentes.

12.1.2 Transformação

A transformação representa a refinada arte de moldar dados brutos em informações úteis e coerentes. Inicia-se com a limpeza, abrangendo correção de erros, eliminação de duplicatas e tratamento de valores ausentes. Além disso, os dados são formatados de modo a garantir consistência e conformidade, incorporando regras de negócios específicas do projeto.

12.1.3 Carregamento

A etapa final, o carregamento, conduz os dados transformados ao seu destino final: um data warehouse. Inicialmente, um carregamento completo é efetuado, seguido pela possibilidade de atualizações incrementais. O data warehouse organiza os dados de maneira eficiente, oferecendo a base sólida necessária para análises avançadas e a geração de relatórios significativos.

Essas fases do processo ETL não apenas formam a espinha dorsal de projetos de Big Data, mas também garantem que os dados, desde sua origem até o armazenamento final, estejam preparados para análises robustas, promovendo insights valiosos para a tomada de decisões estratégicas.

12.2 Serviços utilizados

A Amazon Web Services (AWS) é amplamente reconhecida por sua ampla gama de serviços, proporcionando a capacidade de construir pipelines de ETL (Extração, Transformação e Carga) e sistemas de armazenamento de dados altamente escaláveis e eficientes. O processo de ETL abrange as fases de extração, transformação e carga de dados, e a AWS oferece diversas ferramentas destinadas a otimizar cada uma dessas etapas, resultando em pipelines de dados robustos e flexíveis. Esses serviços foram estrategicamente empregados na implementação de um pipeline de ETL personalizado, ajustado às necessidades específicas do projeto em questão.

Amazon S3: O Amazon S3 é um serviço de armazenamento de objetos escalável que pode ser utilizado no processo de armazenar dados brutos e processados. Dessa forma, podem servir como um local intermediário para dados processados antes de serem carregados em um data warehouse.

O Amazon EC2 (Elastic Compute Cloud): oferece capacidade de computação na nuvem, desempenhando um papel essencial na fase de transformação, especialmente para dados complexos que demandam recursos computacionais significativos. Permitindo com que os usuários criem e executem máquinas virtuais em servidores da AWS, proporcionando flexibilidade e escalabilidade.

AWS Lambda: O AWS Lambda, projetado para execução eficiente de código sem a necessidade de gerenciar infraestrutura, é útil para transformações de dados simples e menos complexas, permitindo a criação e implantação de funções que respondem a eventos e executam tarefas específicas.

Amazon Redshift: O Amazon Redshift é um serviço de data warehouse gerenciado, que oferece alta performance para análise de dados. Após a etapa de ETL, os dados podem ser carregados no Redshift, possibilitando a execução de consultas analíticas para obter insights valiosos a partir dos dados processados

12.3 Processo de ETL

12.3.1 Visão geral

A seguir, descreveremos as práticas de monitoramento e gerenciamento do processo de ETL em um ambiente AWS, utilizando o serviço AWS Redshift Serverless e as ferramentas de monitoramento do Amazon CloudWatch.

12.3.2 Informações do ambiente

General information		
Namespace workspace-cubo-data-dream	Status Available	Admin user name admin
Namespace ID 2b23c6b1-fdf0-4011-aae6-a6e56387d063	Date created November 21, 2023, 22:52 (UTC-03:00)	Database name dev
Namespace ARN arn:aws:redshift-serverless:us-east-1:185405266895:namespace/2b23c6b1-fdf0-4011-aae6-a6e56387d063	Storage used 153.5 GB	Total table count 92

Figura 181: Informações sobre o ambiente

Fonte: Autoria Própria.

- Namespace: workspace-cubo-data-dream
- ID do Namespace: 2b23c6b1-fdf0-4011-aae6-a6e56387d063
- ARN do Namespace: arn:aws:redshift-serverless:us-east-1:185405266895:namespace/2b23c6b1-fdf0-4011-aae6-a6e56387d063
- Status: Disponível
- Data de Criação: 21 de Novembro de 2023
- Usuário Admin: admin
- Nome do Banco de Dados: dev
- Armazenamento Utilizado: 153.5 GB
- Contagem Total de Tabelas: 92

12.3.3 Monitoramento com amazon cloudwatch

O Amazon CloudWatch é utilizado para monitorar métricas vitais relacionadas ao processo de ETL, as quais incluem o número de objetos e o tamanho do bucket em bytes. As métricas são visualizadas em gráficos que demonstram a média diária, proporcionando uma análise de tendências e comportamento dos dados ao longo do tempo.

12.3.3.1 Métricas monitoradas

NumberOfObjects: Esta métrica representa a contagem média de objetos em diferentes buckets de dados, categorizados por fontes de dados como cnjp-data-dream, dadosinep-data-dream, dadosmec-data-dream, entre outros.

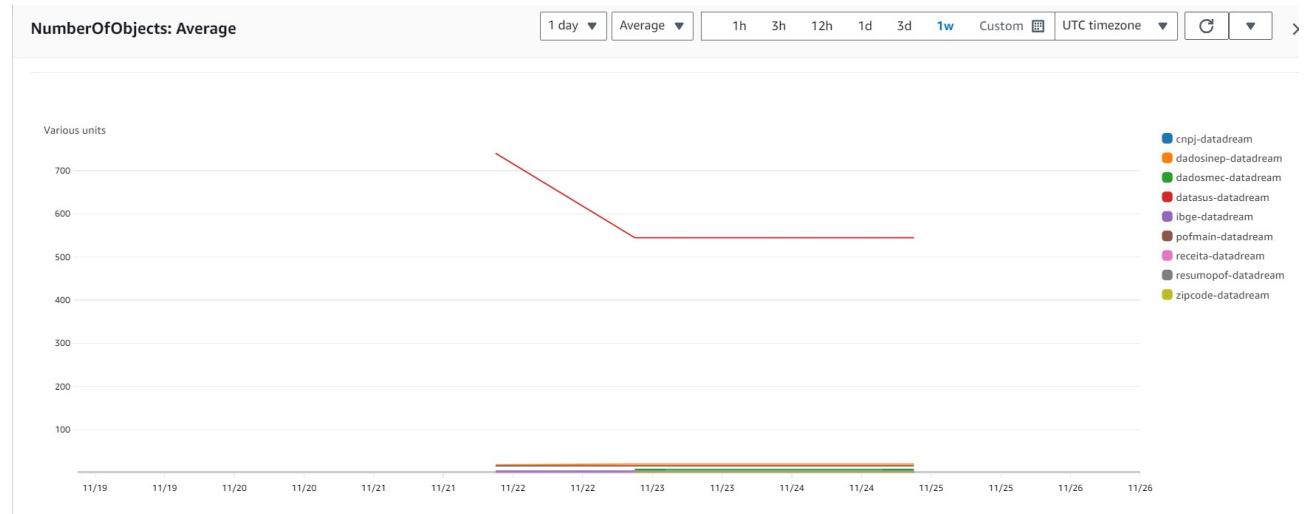


Figura 182: Number Of Objects

Fonte: Amazon CloudWatch.

BucketSizeBytes: Esta métrica representa o tamanho médio em bytes dos buckets mencionados, permitindo a avaliação da utilização do armazenamento. A métrica BucketSizeBytes revela uma disparidade notável entre os tamanhos dos buckets. Especificamente, o bucket associado ao datasus-data-dream destaca-se por ter um tamanho significativamente maior em comparação com os demais. Isso indica que os dados provenientes do DATASUS são substancialmente maiores, o que pode refletir a complexidade e a riqueza dos conjuntos de dados de saúde, exigindo uma atenção especial em termos de processamento e análise.

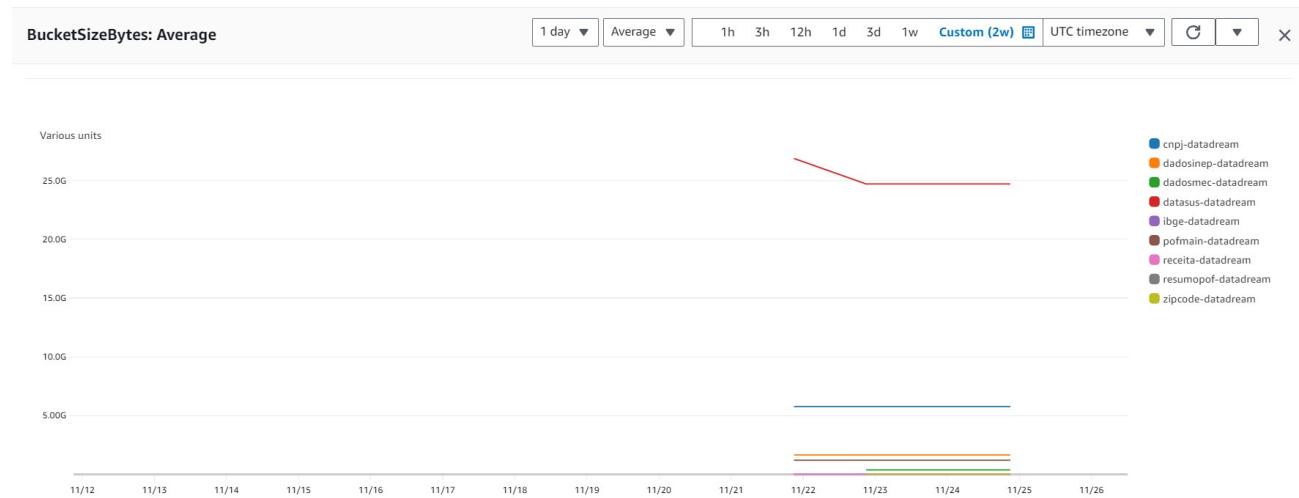


Figura 183: Bucket Size Bytes

Fonte: Amazon CloudWatch.

12.3.3.2 Implicações e ações:

Avaliação de Capacidade: A discrepância no tamanho dos buckets, especialmente para datasus-data-dream, sugere a necessidade de uma avaliação de capacidade e performance, para assegurar que o ambiente está dimensionado adequadamente.

Análise de Custo: A gestão do tamanho dos buckets é essencial para otimizar os custos. O bucket datasus-data-dream pode necessitar de estratégias de compactação de dados ou arquivamento para controlar os gastos.

Priorização de Processamento: Dados do datasus-data-dream podem requerer mais recursos durante as etapas de transformação devido ao seu volume, o que deve ser considerado no planejamento de capacidade do ETL.

12.3.3.3 Alarmes e notificações

- **Alarmes de Utilização:** Configurados para notificar a equipe caso o número de objetos diminua ou aumente drasticamente, indicando potenciais problemas de ingestão ou exclusão de dados.
- **Alarmes de Capacidade:** Acionados se o tamanho do bucket exceder um limite específico, o que pode indicar uma necessidade de escalar recursos ou investigar a eficiência do processo de transformação de dados.

12.3.4 Gerenciamento e otimização

O gerenciamento proativo do ambiente é realizado por meio da revisão contínua das métricas coletadas, garantindo que o processo de ETL esteja otimizado para performance e custo.

12.3.4.1 Ações de gerenciamento

Análise de Tendências: Realizada semanalmente para identificar padrões de crescimento de dados e ajustar processos de ETL conforme necessário.

Otimização de Recursos: Baseada nas métricas, recursos como capacidade de armazenamento e computação são ajustados para atender à demanda dinâmica.

12.3.5 Conclusão

Em resumo, o ambiente workspace-cubo-data-dream está estruturado para garantir uma operação de ETL segura. Através do monitoramento constante e das práticas de gerenciamento adotadas, buscamos assegurar a integridade dos processos. A análise detalhada das métricas no Amazon CloudWatch, especialmente no que diz respeito ao notável aumento no tamanho do bucket datasus-data-dream, destaca a importância de adaptar dinamicamente nossa capacidade para lidar com conjuntos de dados desafiadores, como os relacionados à saúde. Ao realizar análises semanais de tendências e ajustes contínuos com base nas métricas coletadas, procuramos otimizar o desempenho e controlar efetivamente os custos. Em suma, nosso ambiente de ETL está preparado para evoluir conforme necessário, mantendo a qualidade e a confiabilidade dos dados ao longo do tempo.

13. Ensemble

13.1 Modelo RandomForest com CRISP-DM

13.1.1 Entendimento do negócio

Durante essa etapa, é crucial obtermos uma compreensão profunda dos objetivos do negócio. Estabelecer uma comunicação transparente e eficaz com nossos parceiros é essencial. É fundamental definir os critérios que determinarão o sucesso de nosso projeto e entender como agir com base neles. Em outras palavras, ao entendermos o perfil de nossos clientes e usuários, podemos, por meio de análises, extrair insights valiosos e relevantes.

13.1.2 Entendimento dos dados

A fase de compreensão dos dados tem início na coleta inicial dos dados. Nesse estágio, conduzimos uma análise abrangente dos dados fornecidos pelo nosso parceiro, abrangendo informações do POF, bem como a coleta de dados considerados cruciais de fontes como Datasus, IBGE, Mec e Inep. Identificamos correlações significativas, como aquelas entre o consumo de itens do mercado e CNPJs, permitindo insights valiosos sobre áreas com maior demanda, expansão comercial, entre outros.

13.1.3 Preparação dos dados

Na etapa de preparação dos dados, realizamos a limpeza e o pré-processamento dos dados coletados. Selecioneamos os dados e arquivos considerados mais relevantes, realizando a escolha e transformação de variáveis essenciais. Além disso, desenvolvemos um script padronizado para todos os arquivos de dados, garantindo que passassem por um processo consistente de e processamento. Esse script envolvia a leitura de arquivos CSV, aplicação de tratamentos aos dados, remoção de linhas duplicadas, eliminação de valores nulos, entre outros. Esse procedimento foi repetido para todos os arquivos de dados, assegurando um carregamento consistente, limpeza e tratamento.

Na imagem abaixo, é possível observar um exemplo de código empregado no estágio de pré-processamento de dados, executando as devidas etapas para a realização dos testes com base nos dados.

```
df = funcao.limpar_dados(df, colunas_a_ignorar=['email', 'data', 'data_situacao_cadastral', 'data_inicio_atividade'])
df = funcao.tratar_valores_nulos(df, 'nan', chunk_size=1000)
df = funcao.remover_coluna(df, 'cnae_fiscal_secundaria')
funcao.ajustar_amazon_s3(df, './csv_s3/cnpjs_1_s3.csv')
```

Figura 184: Pré-processamento dos dados

Fonte: Autoria Própria

13.1.4 Modelagem

O processo de modelagem abrange as fases de seleção e aplicação de técnicas, como algoritmos de aprendizado de máquina, envolvendo a configuração de parâmetros e o ajuste do treinamento dos modelos. Nessa etapa, é crucial avaliar o desempenho do modelo por meio dos dados de treinamento.

No exemplo a seguir, utilizamos um notebook para ler dois arquivos de tabelas distintas da POF, empregando uma abordagem analítica com diferentes tabelas para testar duas teorias. Inicialmente, selecionamos dois índices amplamente utilizados: o IPM-NM (Índice de Pobreza Multidimensional Não Monetária) e o IVM-PN (Índice de Vulnerabilidade Municipal - Pobreza Não Monetária). O IPM-NM avalia a pobreza em diversas áreas da vida, como saúde, educação e moradia, enquanto o IVM-PN mede a vulnerabilidade e pobreza em uma cidade, considerando fatores além da renda, como acesso a serviços e qualidade de vida.

13.1.5 IPM-NM

Para a análise do IPM-NM, escolhemos um algoritmo bem conhecido para compreender e interpretar os dados: Random Forest (Floresta Aleatória), pertencente à categoria de métodos ensemble, para combinar vários modelos a fim de melhorar o desempenho e a generalização do modelo final, utilizando árvores de decisão como base.

Selecionamos uma tabela da POF que abrange aspectos cruciais nos estados do Brasil, como proporção de pessoas com algum grau de pobreza, nível de educação, saúde e moradia, para auxiliar nas análises.

A seguir, conduzimos o processo de construção, treinamento e avaliação do modelo escolhido. Inicialmente, carregamos os dados, dividindo-os em conjuntos de treinamento e teste para representar valores, variáveis e características. Em seguida, instalamos o modelo para garantir reproduzibilidade e iniciamos o treinamento com o conjunto de treinamento. Posteriormente, realizamos a previsão do modelo no conjunto de testes. Por fim, definimos várias métricas de regressão para avaliar o desempenho do

modelo, incluindo Erro Quadrático Médio (MSE), Erro Quadrático Médio Raiz (RMSE), Erro Absoluto Médio (MAE) e Coeficiente de Determinação (R^2).

13.1.6 IVM-NM

Para a implementação do modelo na análise do índice IVM-NM, empregamos o algoritmo Gradiente Boosting Regressor, reconhecido como uma técnica de aprendizado de máquina voltada para problemas de regressão. A concepção fundamental por trás do Gradient Boosting envolve a construção sequencial de modelos fracos, nos quais cada novo modelo é treinado para corrigir os erros do modelo anterior. Esse processo iterativo ajusta os pesos dos modelos anteriores com base nos erros residuais, sendo a otimização do gradiente descendente, onde o modelo é ajustado na direção oposta ao gradiente da função de perda em relação aos parâmetros do modelo, denominada como "Gradient".

No contexto de regressão, o Gradient Boosting Regressor demonstra eficácia especialmente em situações com relacionamentos complexos e não lineares nos dados. Sua capacidade de lidar com outliers e resistência ao overfitting contribuem para um desempenho robusto em diversas circunstâncias.

Para essa análise específica, selecionamos uma tabela da POF fornecida pelo parceiro, contendo informações diversas, como moradia, acesso aos serviços de utilidade pública, saúde e alimentação, educação, acesso a serviços financeiros e padrão de vida, entre outros. Posteriormente, procedemos com a preparação e instanciamento da aplicação do modelo, incluindo o treinamento do modelo, a predição no conjunto de teste e a escolha das mesmas métricas utilizadas no outro algoritmo para análise, visando obter insights e informações relevantes.

As métricas utilizadas para avaliação foram as mesmas do caso anterior: Erro Quadrático Médio (MSE), Erro Quadrático Médio Raiz (RMSE), Erro Absoluto Médio (MAE) e Coeficiente de Determinação (R^2). Essas métricas são essenciais para compreender a qualidade do modelo na previsão do índice IVM-NM, proporcionando uma análise comparativa e aprofundada de seu desempenho.

13.1.7 Avaliação

O processo de avaliação inclui do modelo foi realizado com base nos critérios de sucesso definidos para compreendermos caso atingimos o objetivo do projeto.

13.1.7.1 IPM-NM

O Modelo de regressão utilizando a técnica de Floresta Aleatória foi avaliado com base em diversas métricas para compreender sua eficácia na previsão do índice de

pobreza multidimensional normalizado (IPM-NM). E os seguintes valores foram encontrados: Mean Squared Error (MSE): é a média dos quadrados dos erros entre as previsões do modelo e os valores reais. Quanto menor o MSE, melhor. Neste caso, o valor de MSE indica que, em média, os quadrados dos erros são cerca de 20.44. Root Mean Squared Error (RMSE): O RMSE é a raiz quadrada do MSE. Ele fornece uma interpretação mais intuitiva, pois está na mesma unidade que a variável de resposta. Neste caso, o RMSE indica que, em média, os erros são cerca de 4.52 unidades. Mean Absolute Error (MAE): O MAE é a média dos valores absolutos dos erros entre as previsões e os valores reais. Ele mede a magnitude média dos erros sem considerar a direção.

Neste caso, o MÃE indica que, em média, os erros são cerca de 4.19 unidades. R-squared (R^2): O R-squared, ou coeficiente de determinação, mede a proporção da variabilidade na variável de resposta que é explicada pelo modelo. Um valor de 1 indica um ajuste perfeito, enquanto um valor de 0 indica que o modelo não explica nada da variabilidade. Neste caso, o R^2 de aproximadamente 0.55 sugere que o modelo explica cerca de 55% da variabilidade nos dados, o que pode ser considerado um ajuste razoável. Após essa etapa, procedemos à criação de um gráfico de dispersão para visualizar o ajuste do modelo de regressão com a linha ideal, utilizando valores reais e previsões. A análise do gráfico revela a presença de pontos próximos à linha zero, indicando previsões precisas, mas também a existência de pontos distantes, sugerindo possíveis desbalanceamentos ou falhas no modelo.

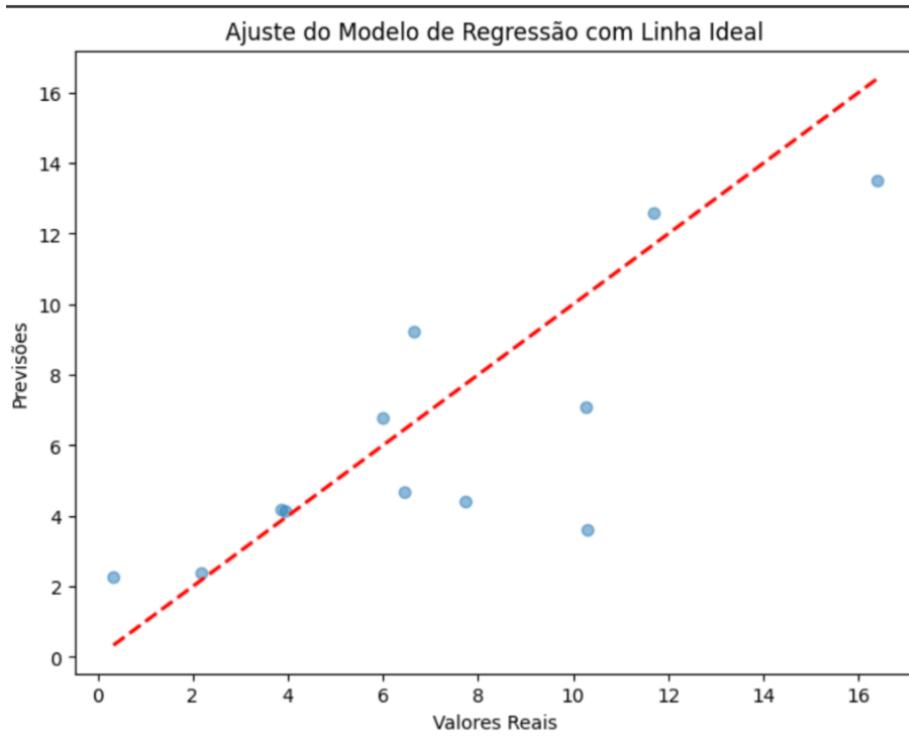


Figura 185: Gráfico de dispersão

Fonte: Autoria Própria

Subsequentemente, realizamos a plotagem de um gráfico de resíduos, visando identificar padrões nos erros do modelo. O eixo X representa os valores reais do conjunto de testes, o eixo Y representa as diferenças entre os valores reais e as previsões do modelo, e a linha horizontal vermelha ($y=0$) representa a linha zero. Observamos que os resíduos estão distribuídos aleatoriamente, mas a presença de pontos distantes da linha horizontal pode indicar desafios ou falhas.

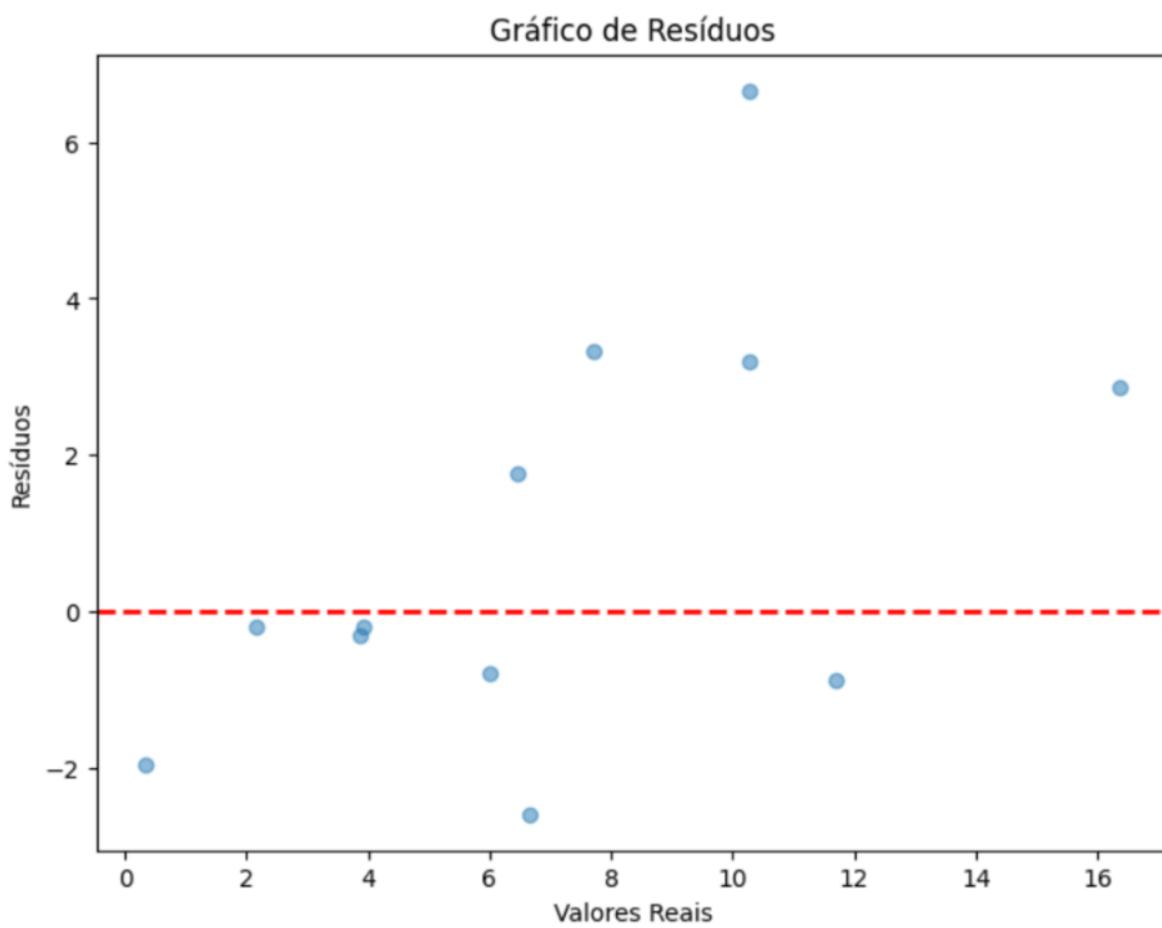


Figura 186: Gráfico de resíduos

Fonte: Autoria Própria

Finalmente, criamos um histograma de resíduos para examinar a distribuição dos resíduos de maneira mais detalhada. O eixo X (resíduos) representa os diferentes valores dos resíduos, e o eixo Y (frequência) indica a frequência com que cada valor de resíduo ocorre. Espera-se uma forma de sino, indicando uma distribuição normal dos resíduos, mas a presença de outliers ou valores extremos pode sugerir problemas ou padrões no modelo. A interpretação cuidadosa desses gráficos é fundamental para compreender se o modelo está capturando efetivamente os padrões nos dados ou se há áreas de melhoria necessárias.

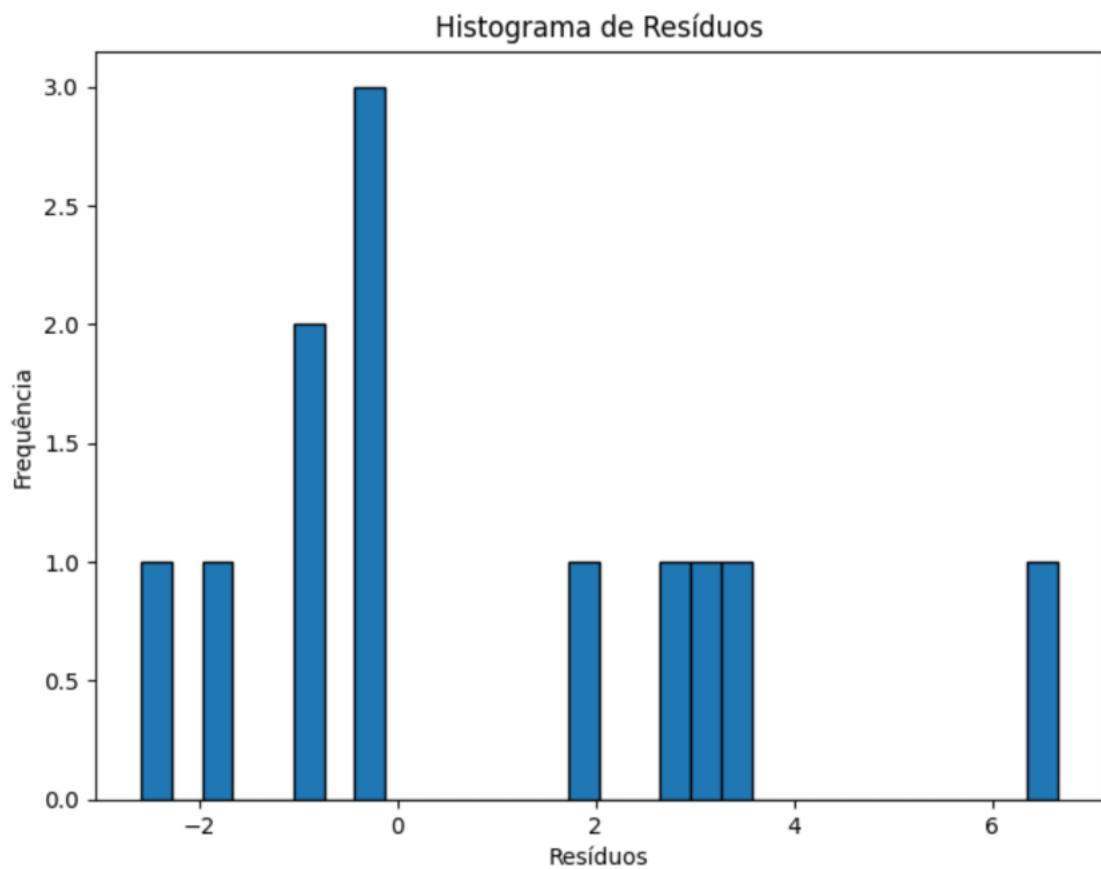


Figura 187: Histograma de resíduos

Fonte: Autoria Própria

13.1.7.1 IVM-NM

A análise do modelo para o Índice de Vulnerabilidade Municipal (IVM) foi conduzida com base nos critérios de sucesso definidos, visando compreender a eficácia na consecução dos objetivos do projeto.

O Modelo de Regressão, utilizando a técnica de Gradient Boosting, foi submetido a uma análise detalhada, empregando diversas métricas para avaliar sua capacidade de previsão do índice de vulnerabilidade municipal (IVM). Os resultados obtidos foram os seguintes: Mean Squared Error (MSE): 7.57; Root Mean Squared Error (RMSE): 2.75; Mean Absolute Error (MAE): 2.15; R-squared (R^2): 0.59.

Essas métricas proporcionam insights cruciais sobre a qualidade do modelo. O MSE, ao representar a média dos quadrados dos erros, indica que, em média, os quadrados dos erros são aproximadamente 7.57. O RMSE, sendo a raiz quadrada do MSE, interpreta que os erros médios são cerca de 2.75 unidades. O MAE, que é a média dos valores absolutos dos erros, revela que, em média, os erros são de aproximadamente 2.15 unidades. O R^2 , ou coeficiente de determinação, sugere que o modelo explica cerca de 59% da variabilidade nos dados.

Posteriormente, foi realizada uma análise visual por meio do gráfico intitulado 'Ajuste do Modelo de Regressão (Gradient Boosting) com Linha Ideal'. Esse gráfico evidenciou que os pontos estão próximos da linha zero, indicando que, apesar da presença de valores extremos e outliers, muitos pontos estão bem ajustados ao modelo. A proximidade à linha zero sugere previsões precisas, enquanto a presença de valores distantes pode indicar desbalanceamento ou falhas no modelo.

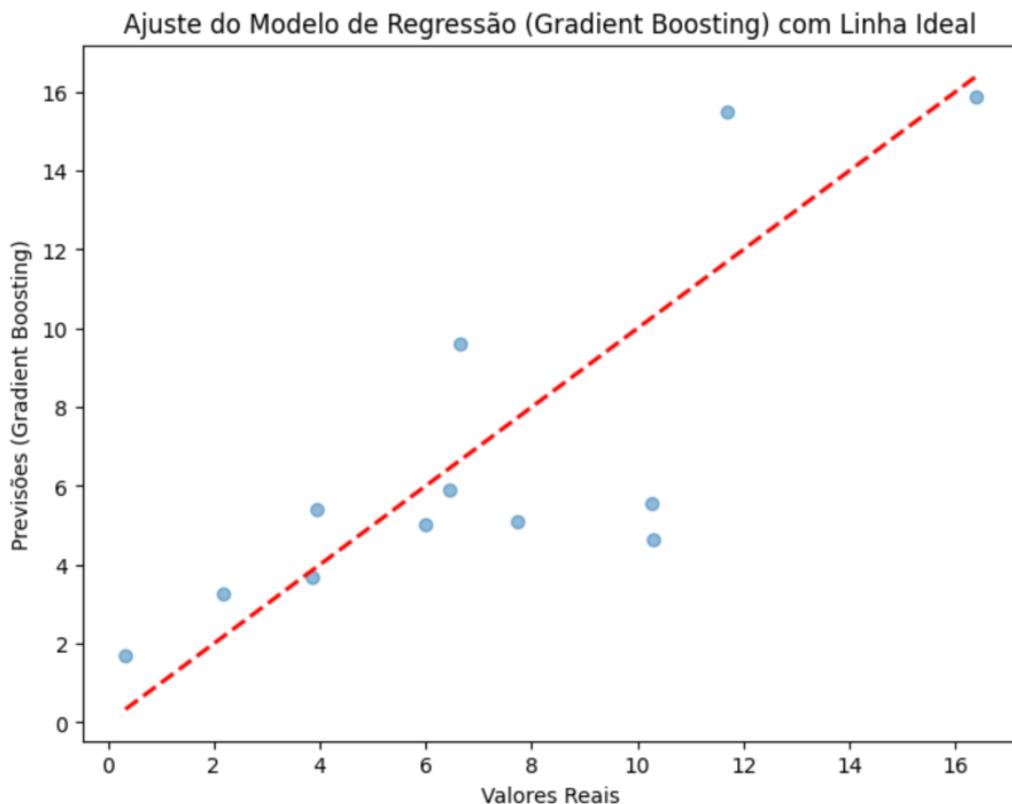


Figura 188: Ajuste do modelo de Regressão - Gradient Boosting

Fonte: Autoria Própria

Adicionalmente, foram gerados gráficos de resíduos e um histograma de resíduos utilizando o Gradient Boosting. O gráfico de resíduos possibilitou visualizar a diferença entre os valores reais e as previsões do modelo, destacando áreas de possível desbalanceamento. O histograma de resíduos, por sua vez, contribuiu para examinar a distribuição dos resíduos, identificando padrões ou outliers que podem indicar possíveis problemas no modelo.

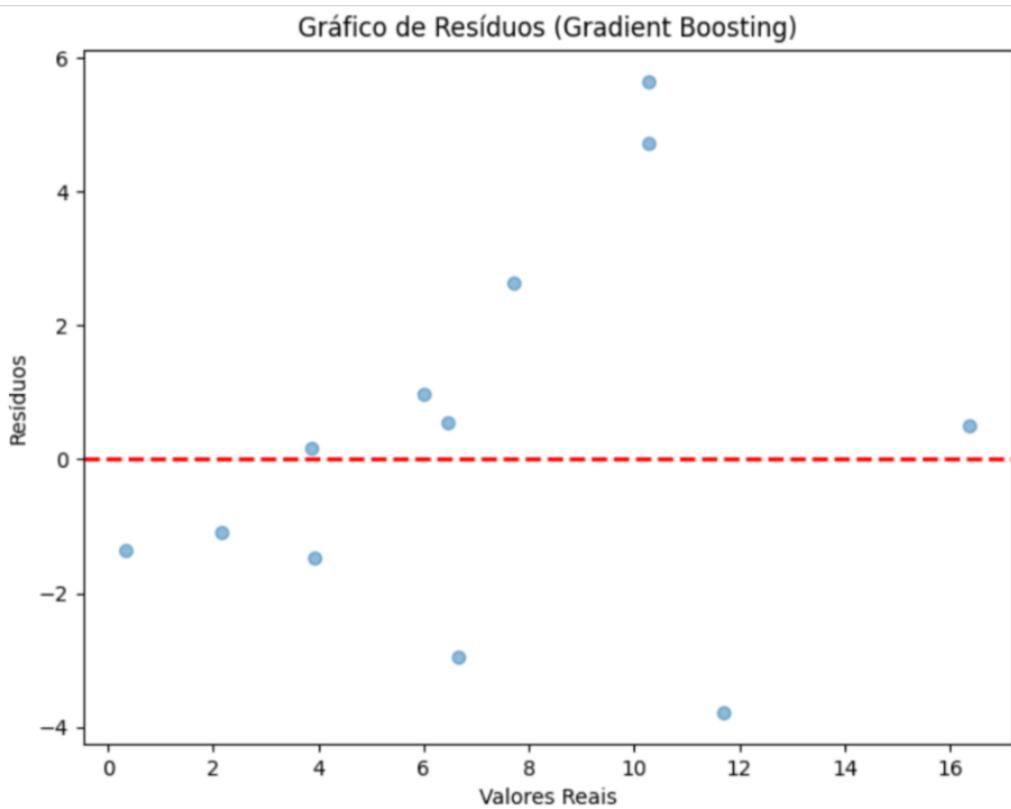


Figura 189: Gráfico de resíduo - Gradient Boosting

Fonte: Autoria Própria

13.7.8 Implementação e Deploy

Este processo é conhecido por realizar a integração do modelo em ambiente de produção. A partir disso, compreendemos que com base nos dados que foram preparados, e modelados a partir dos algoritmos escolhidos, desenvolvemos um infográfico que servirá como uma forma de representarmos em um formato visual a partir do perfil esperado pelo nosso cliente, o que pode ser tomado como conclusão.

13.7.9 Avaliação e validação do modelo de Floresta Aleatória com o CRISP-DM

Hiperparametrização do gradient boosting O objetivo do código a seguir é realizar a hiperparametrização do modelo de Regressão por Gradient Boosting para melhorar seu desempenho na previsão de um índice específico ("IVM-NM") com base em um conjunto de dimensões fornecidas.

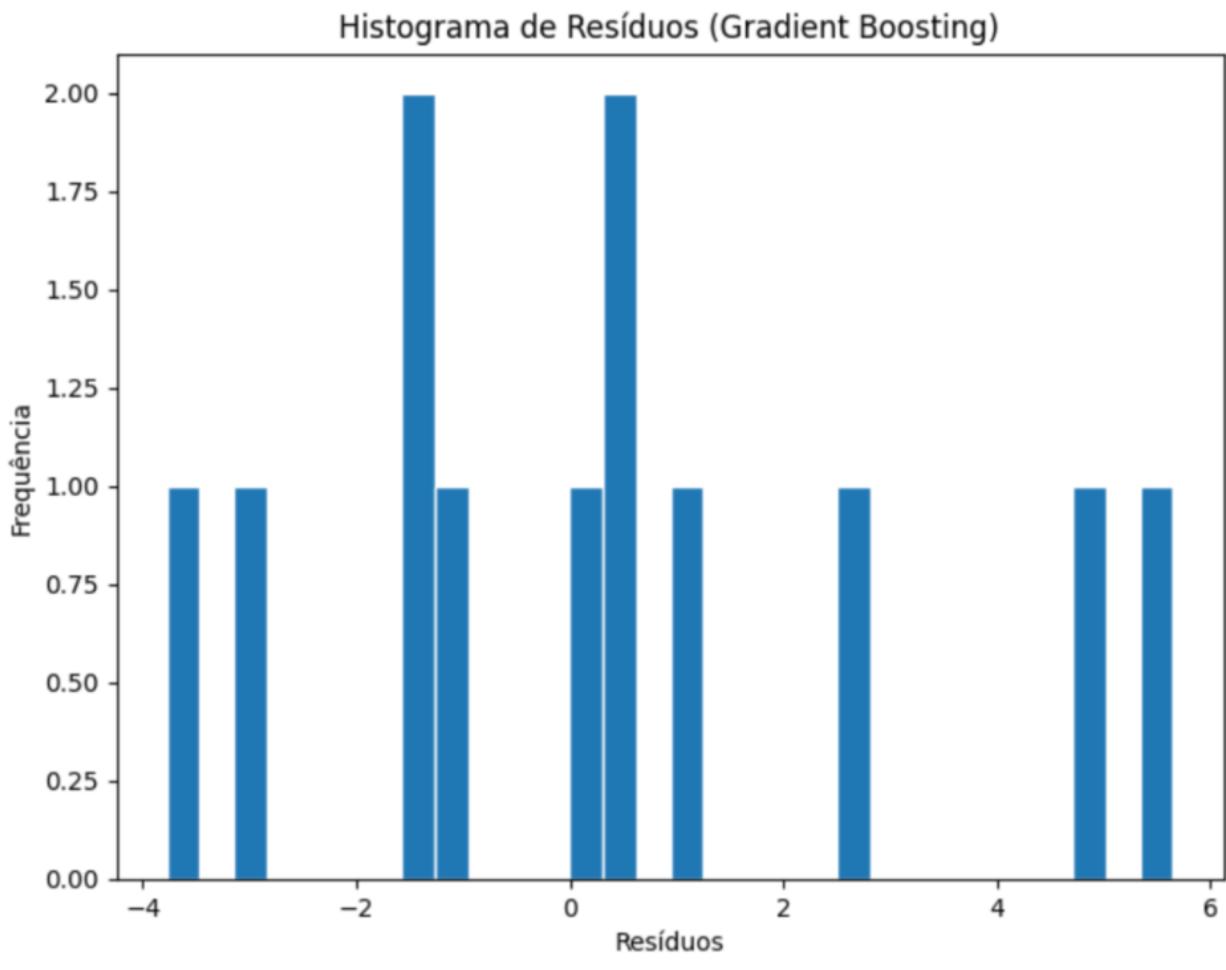


Figura 190: Histograma de resíduo - Gradient Boosting

Fonte: Autoria Própria

Este processo é conhecido por realizar a integração do modelo em ambiente de produção. A partir disso, compreendemos que com base nos dados que foram preparados, e modelados a partir dos algoritmos escolhidos, desenvolvemos um infográfico que servirá como uma forma de representarmos em um formato visual a partir do perfil esperado pelo nosso cliente, o que pode ser tomado como conclusão.

13.7.10 Avaliação e validação do modelo de Floresta Aleatória com o CRISP-DM

Hiperparametrização do gradient boosting O objetivo do código a seguir é realizar a hiperparametrização do modelo de Regressão por Gradient Boosting para melhorar seu desempenho na previsão de um índice específico ("IVM-NM") com base em um conjunto de dimensões fornecidas.

```

import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score

# Carregamento dos dados (substitua 'X' e 'y' pelos seus dados)
X = pof6[['Moradia', 'Acesso aos serviços de utilidade pública', 'Saúde e alimentação',
'Educação', 'Acesso a serviços financeiros e padrão de vida', 'Transporte e lazer']]
y = pof6['IPM-NM']

# Divisão dos dados em treino e teste
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Definição do modelo RandomForestRegressor
rf_model = RandomForestRegressor(n_estimators=500, max_depth=30, min_samples_split=2, min_samples_leaf=1, random_

# Treinamento do modelo
rf_model.fit(X_train, y_train)

# Predição no conjunto de teste
y_pred_rf = rf_model.predict(X_test)

# Métricas de avaliação
mse_rf = mean_squared_error(y_test, y_pred_rf)
rmse_rf = np.sqrt(mse_rf)
mae_rf = mean_absolute_error(y_test, y_pred_rf)
r2_rf = r2_score(y_test, y_pred_rf)

# Exibindo as métricas
print("Mean Squared Error (MSE):", mse_rf)
print("Root Mean Squared Error (RMSE):", rmse_rf)
print("Mean Absolute Error (MAE):", mae_rf)
print("R-squared (R2):", r2_rf)

```

Figura 191: Código

Fonte: Autoria Própria

Mean Squared Error (MSE): 6.961276826408555

Root Mean Squared Error (RMSE): 2.6384231704577936

Mean Absolute Error (MAE): 2.0661330444566675

R-squared (R²): 0.6207753914149101

13.7.11 Carregamento e Preparação dos Dados

No início, são carregados os dados referentes às variáveis independentes (X) e à variável dependente (y). As features escolhidas incluem informações sobre moradia, acesso a serviços públicos, saúde, educação, acesso a serviços financeiros, padrão de vida, transporte e lazer.

13.7.12 Divisão dos Dados

Os dados são divididos em conjuntos de treino e teste utilizando a função `train_test_split` da biblioteca scikit-learn. 80% dos dados são destinados ao treino, enquanto 20% são reservados para teste.

13.7.13 Definição do Modelo RandomForestRegressor

O modelo de Regressão RandomForestRegressor é escolhido devido à sua capacidade de lidar com relações não-lineares e complexas nos dados. São definidos hiperparâmetros específicos para o modelo, como o número de estimadores (500), a profundidade máxima da árvore (30), e os critérios de divisão (min_samples_split=2, min_samples_leaf=1).

13.7.14 Treinamento do Modelo

O modelo é treinado utilizando o conjunto de treino. Durante esse processo, o algoritmo ajusta os parâmetros internos da floresta de árvores de decisão para melhor se adaptar aos padrões nos dados.

13.7.15 Predição e Avaliação

Após o treinamento, o modelo é utilizado para realizar previsões no conjunto de teste. Em seguida, diversas métricas de avaliação são calculadas para avaliar o desempenho do modelo. Essas métricas incluem o Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) e o R-squared (R^2).

13.7.16 Resultados

Ao comparar os resultados obtidos anteriormente com a hiperparametrização, observamos melhorias nas métricas de avaliação. O MSE, RMSE e MAE diminuíram, indicando uma redução nos erros de previsão. Além disso, o R^2 aumentou, sugerindo uma melhor capacidade de explicar a variabilidade nos dados.

Esse processo de otimização é fundamental para ajustar o modelo de modo a alcançar o melhor desempenho possível para o problema em questão. Experimentações adicionais e ajustes finos podem ser realizados para buscar melhorias contínuas no modelo.

Resultados Iniciais (Sem Hiperparametrização):

Mean Squared Error (MSE): 7.42497

Root Mean Squared Error (RMSE): 2.72488

Mean Absolute Error (MAE): 2.06260

R-squared (R^2): 0.59552

Resultados Após Hiperparametrização:

Mean Squared Error (MSE): 6.96128

Root Mean Squared Error (RMSE): 2.63842

Mean Absolute Error (MAE): 2.06613

R-squared (R^2): 0.62078

13.7.17 Análise Comparativa:

MSE e RMSE: Houve uma melhoria significativa no MSE e RMSE após a hiperparametrização, indicando uma redução nos erros médios quadráticos e na dispersão dos resíduos. Isso sugere uma melhor precisão nas previsões.

MAE: Embora tenha havido um ligeiro aumento no MAE após a hiperparametrização, a diferença é mínima. O MAE continua em um nível aceitável, indicando que os erros absolutos médios permanecem razoáveis.

R-squared (R^2): Houve um aumento no R^2 após a hiperparametrização, indicando uma melhoria na capacidade do modelo explicar a variabilidade nos dados. Este é um sinal positivo de melhor ajuste do modelo aos dados de teste.

Considerações Finais:

A hiperparametrização é uma abordagem eficaz para otimizar o desempenho do modelo e ajustar seus parâmetros para um melhor ajuste aos dados. O Random Forest demonstrou melhorias significativas nas métricas após a hiperparametrização. A análise comparativa fornece insights sobre como as mudanças nos hiperparâmetros impactaram as métricas de avaliação do modelo Random Forest.

13.2 Spark

Apache Spark é um framework de processamento de dados distribuído, projetado para lidar com grandes volumes de dados de forma eficiente e escalável. Ele fornece APIs em várias linguagens, incluindo Python, para processamento de dados em lote e em tempo real.

Funcionalidades do Spark:

- Processamento Distribuído: Permite processar dados em clusters para maior velocidade e escalabilidade.

- Suporte a Diversos Tipos de Dados: Pode lidar com dados estruturados e não estruturados.
- Módulos Integrados: Inclui módulos para processamento de SQL, machine learning, streaming, etc.
- Resiliência e Tolerância a Falhas: Oferece alta disponibilidade e capacidade de recuperação em caso de falhas.

Importância do Spark:

- Processamento Rápido: O Spark é conhecido por seu processamento rápido de dados, sendo até 100 vezes mais rápido que o Hadoop MapReduce.
- Facilidade de Uso: Oferece APIs simples e abstratas para facilitar o desenvolvimento.
- Versatilidade: Pode ser usado para uma variedade de casos, incluindo análise de dados, machine learning e processamento de streaming.
- Escalabilidade: Pode escalar horizontalmente para lidar com grandes volumes de dados.

13.3 Correlação de Dados por Ensemble

O objetivo dessa correlação é identificar, entre os dados incluídos no cubo, quais conexões seriam mais interessantes e importantes para o desenvolvimento das views. A solução que está sendo desenvolvida se propõe a oferecer ao cliente uma ferramenta que permita compreender o potencial de consumo de categorias de produtos alimentares em um nível altamente granular, incluindo informações geográficas e detalhes dos canais de atendimento. Em suma, é necessário se atentar a três critérios de análise: categoria, canal e região. Baseado nessas premissas é que as correlações foram desenvolvidas.

POF características dieta

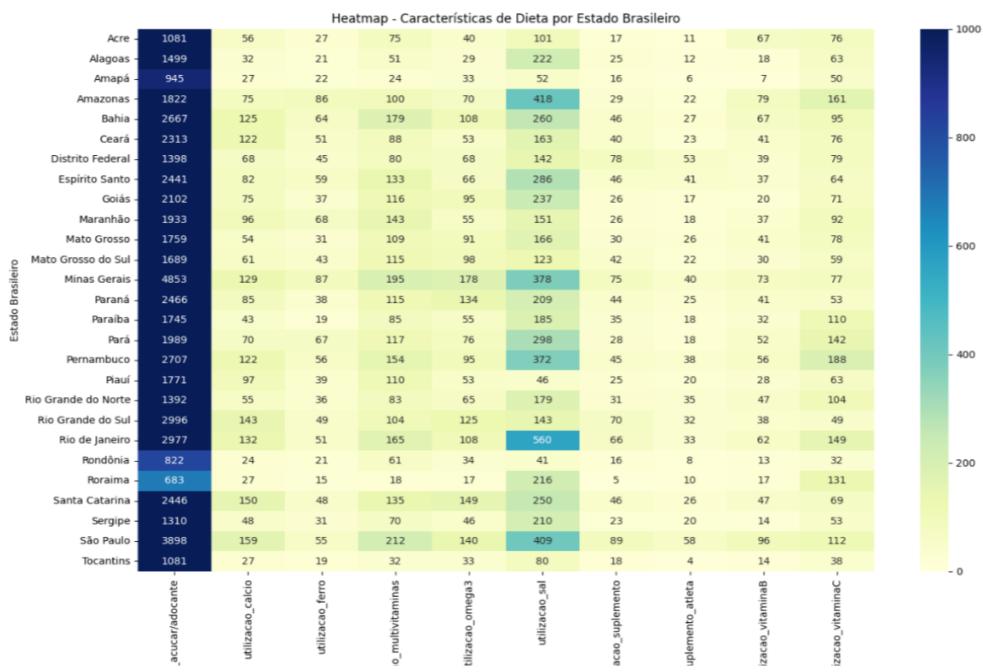


Figura 192: HeatMap - Dieta por estado

Fonte: Autoria Própria

Nessa correlação ensemble por heatmap foi buscado identificar correlações entre características da dieta e regionalidade (estado brasileiro).

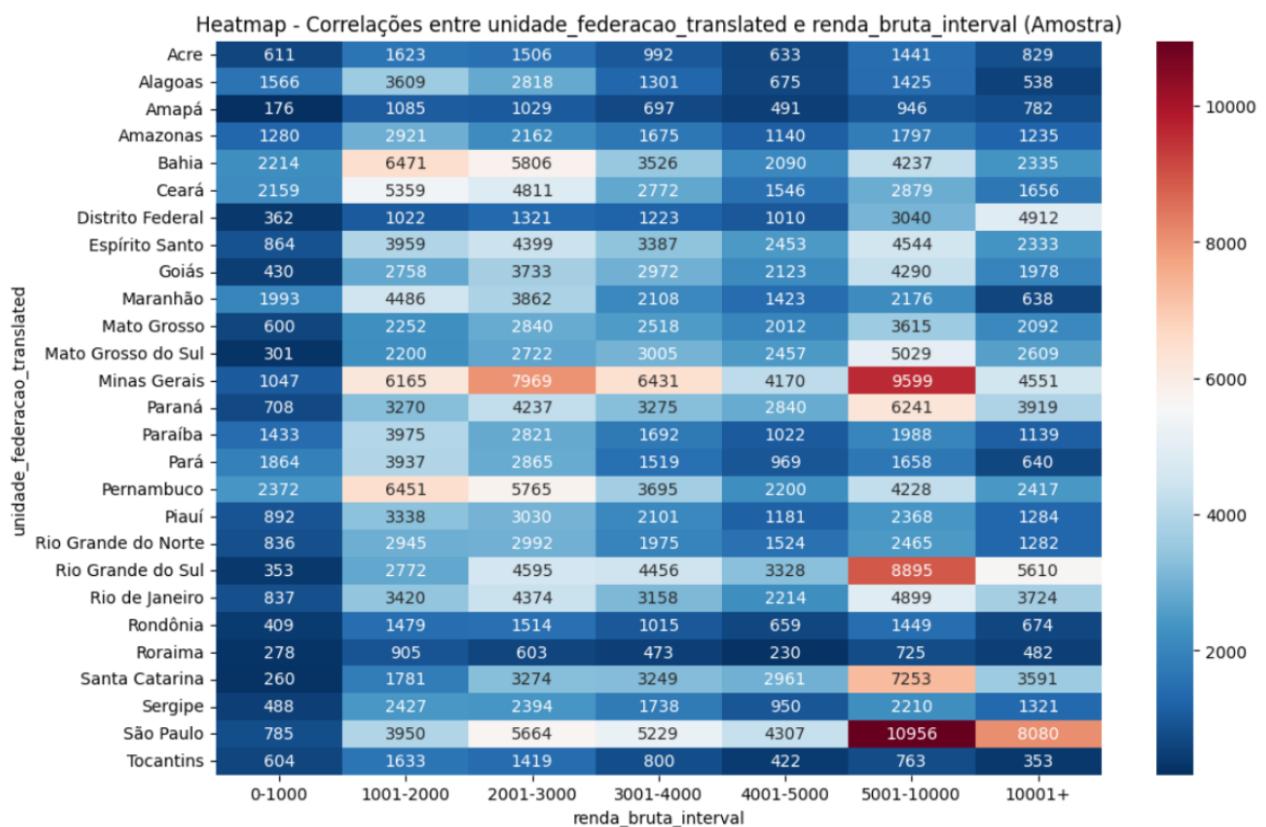


Figura 193: HeatMap - Renda por estado

Fonte: Autoria Própria

Nessa correlação ensemble por heatmap foi buscado identificar correlações entre a renda coletiva e regionalidade (estado brasileiro).



Figura 194: HeatMap - Motivo da restrição alimentar por estado

Fonte: Autoria Própria

Nessa correlação ensemble por heatmap foi buscado identificar correlações entre as restrições de produtos e regionalidade (estado brasileiro).

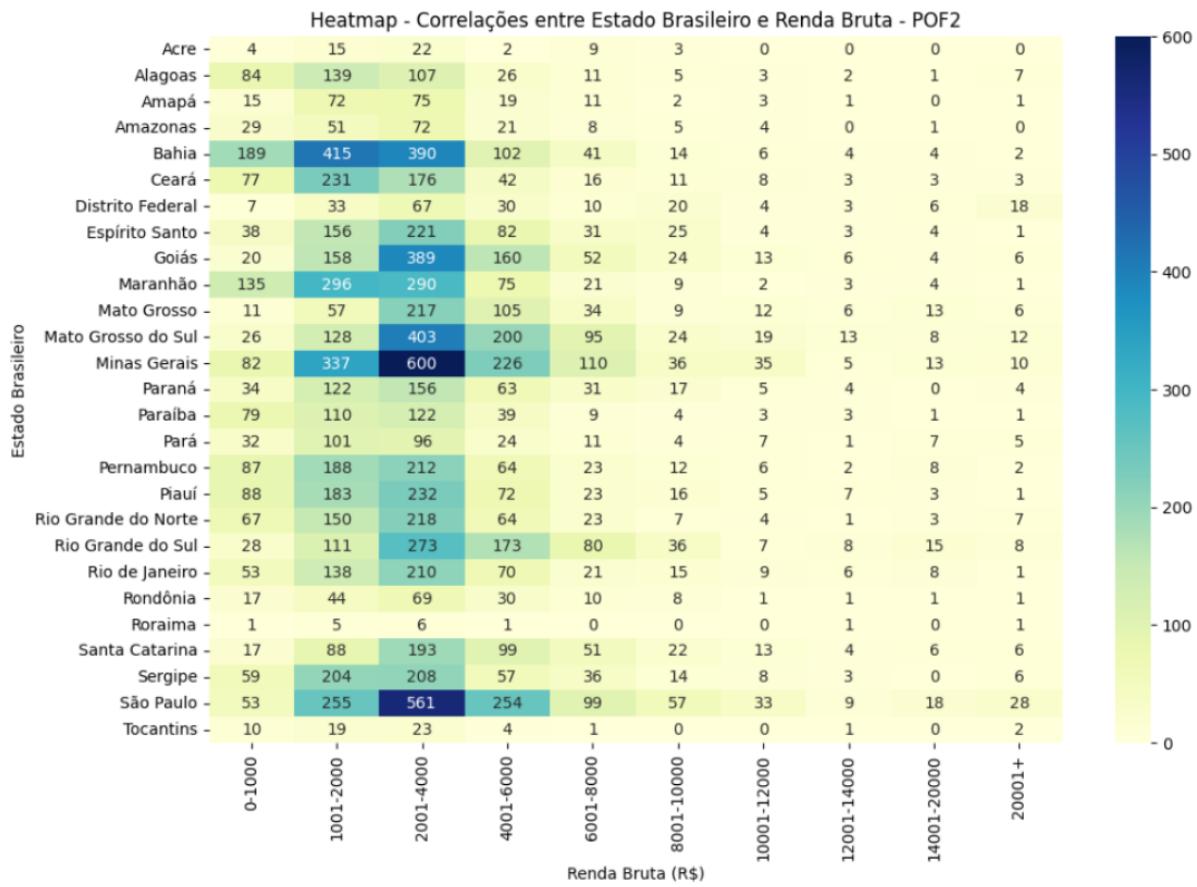


Figura 195: HeatMap - Renda por estado

Fonte: Autoria Própria

Nessa correlação ensemble por heatmap foi buscado identificar correlações entre renda bruta e regionalidade (estado brasileiro).

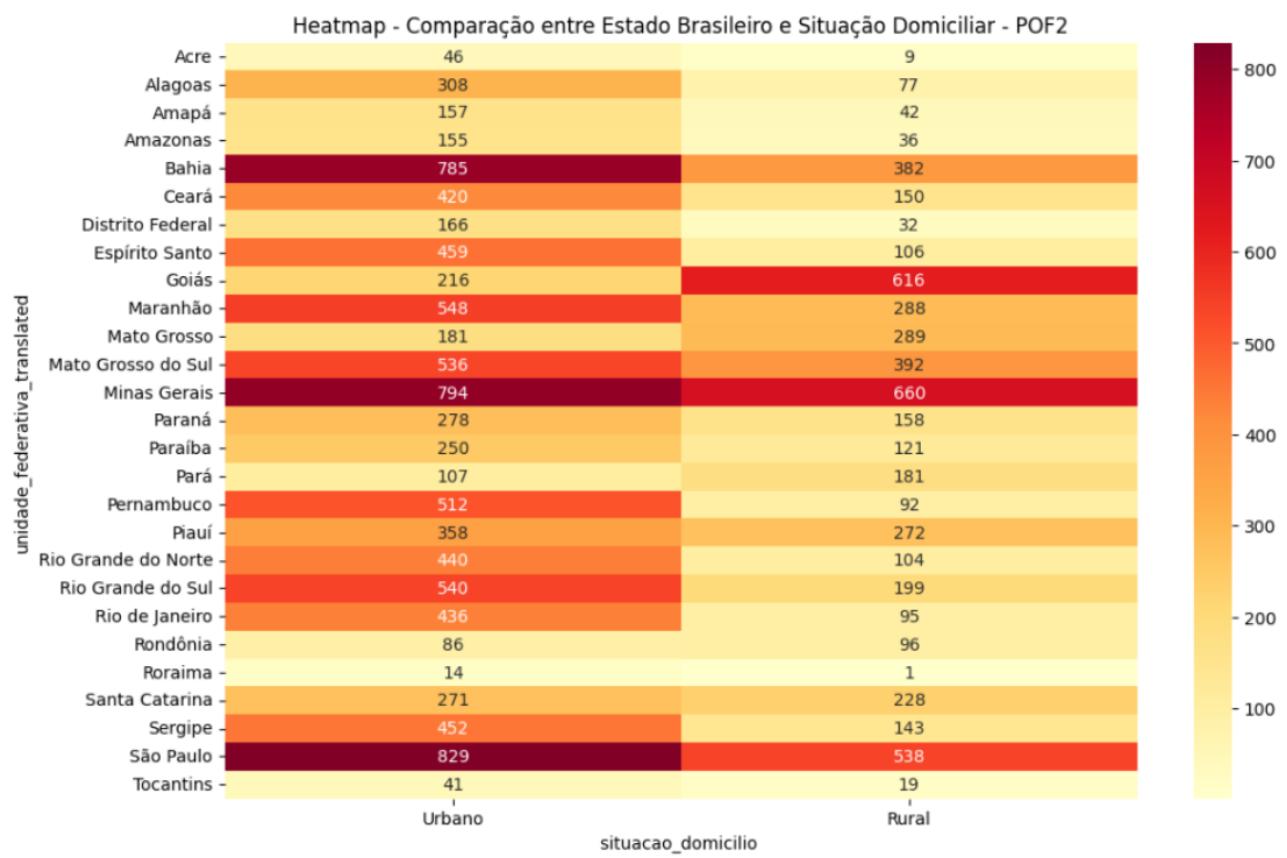


Figura 196: HeatMap - Situação Alimentar por estado

Fonte: Autoria Própria

Ainda na POF2, correlações entre situação rural/urbana versus estado brasileiro.

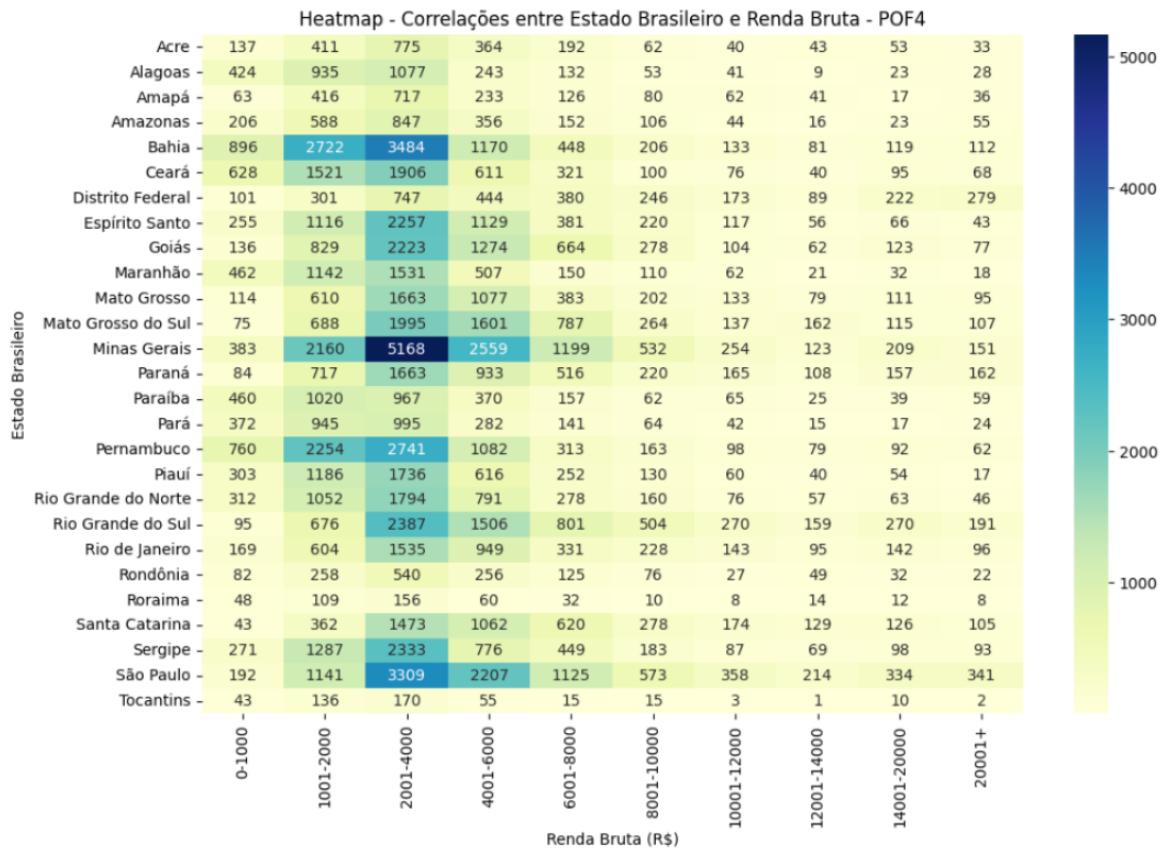


Figura 197: HeatMap - Renda por estado

Fonte: Autoria Própria

Nessa correlação ensemble por heatmap foi buscado identificar correlações entre renda bruta e regionalidade (estado brasileiro).

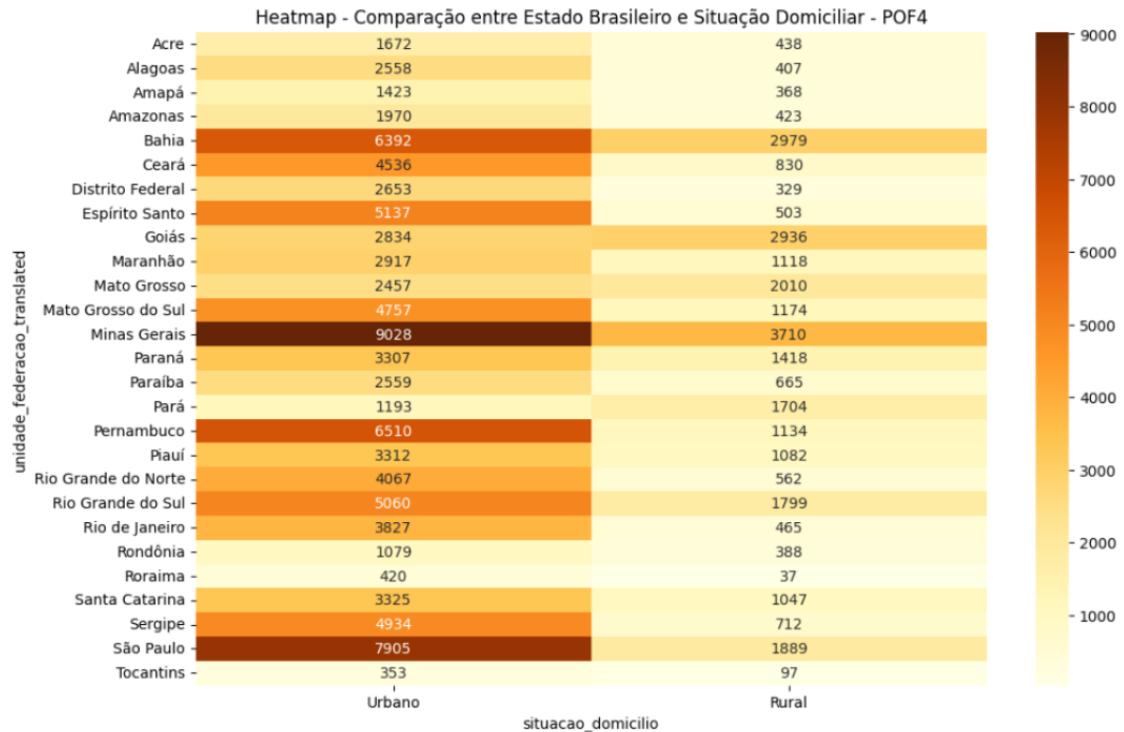


Figura 198: HeatMap - Situação Alimentar por estado

Fonte: Autoria Própria

Ainda na POF4, correlações entre situação rural/urbana versus estado brasileiro.

13.4 K-Means

O método K-means é um algoritmo de clustering utilizado em análise de dados e aprendizado de máquina. Sua finalidade é agrupar dados similares em clusters, facilitando a identificação de padrões e insights nos conjuntos de dados. O algoritmo funciona particionando os dados em K clusters, onde K é um número pré-definido.

Neste tópico, abordaremos a aplicação do método K-means em um contexto prático, utilizando dados relacionados a rendimentos e empregos por estado, pib e gini. Primeiramente, será apresentada a conexão com um banco de dados Redshift e a criação de uma view no ambiente de desenvolvimento Colab (aplicado somente para rendimento_emprego). Em seguida, a criação do gráfico do cotovelo para determinar o número ideal de clusters. Por fim, o K-means em conjunto com a Análise de Componentes Principais (PCA) para visualizar e interpretar os resultados do clustering. Todos os dados citados estão em `src/ensemble/kmeans`.

13.4.1 Conexão com o RedShift

Este trecho de código estabelece uma conexão com um banco de dados Redshift, utilizado neste projeto para estruturar os dados e criar as views. Utilizando a biblioteca psycopg2, o código realiza uma consulta SQL para extrair dados específicos relacionados aos rendimentos e empregos por estado, este é uma view que foi criado no RedShift. Em seguida, os resultados são transformados em um DataFrame do pandas e salvos em um arquivo CSV. O código está disponível em: `src/ensemble/kmeans`

Obs: este código somente é utilizado para views, já que não existe um arquivo csv para elas.

13.4.2 Criação do Gráfico do Cotovelo

O código abaixo aborda a determinação do número ideal de clusters (K) através da criação do gráfico do cotovelo. A soma dos quadrados das distâncias dos pontos ao centróide (inércia) é calculada para diferentes valores de K, e o gráfico resultante ajuda a identificar o ponto de inflexão, indicando o número ótimo de clusters. O código abaixo está disponível em: `src/ensemble/kmeans`.

13.4.3 K-means e PCA

Por último, o código aplica o método K-means em conjunto com a Análise de Componentes Principais (PCA). Os dados são escalados e reduzidos para duas dimensões usando PCA, facilitando a visualização. Em seguida, o K-means é aplicado aos componentes principais para realizar a clusterização. O código abaixo está disponível em: `src/ensemble/kmeans`.

14. Análise de Custo

Empresas podem usar dados da POF para identificar áreas com alta demanda por seus produtos. O cruzamento desses dados com as informações do CNPJ pode ajudar a identificar regiões com pouca concorrência, que podem ser locais potenciais para expansão.

14.1 Primeiros passos

Passo 1: Acesse a AWS Calculator - Abra o site da AWS Calculator em <https://calculator.aws/#/>.

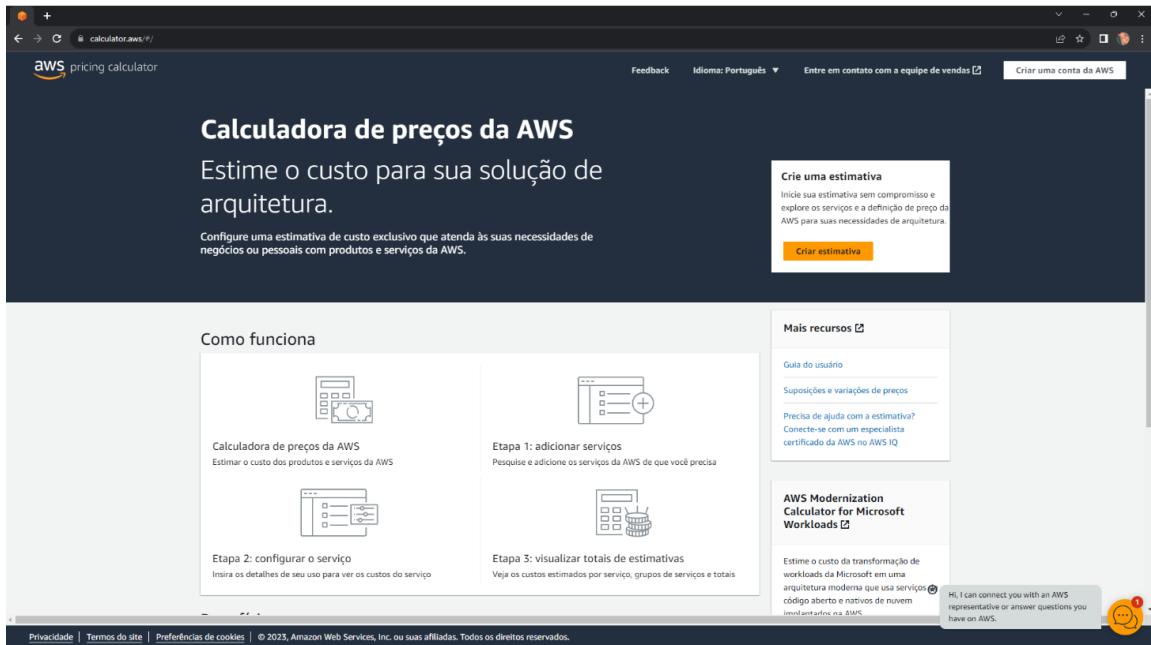


Figura 199: AWS Calculator

Fonte: Autoria Própria

Passo 2: Crie uma Estimativa - Clique em "Criar estimativa" para iniciar o processo de análise de custo.

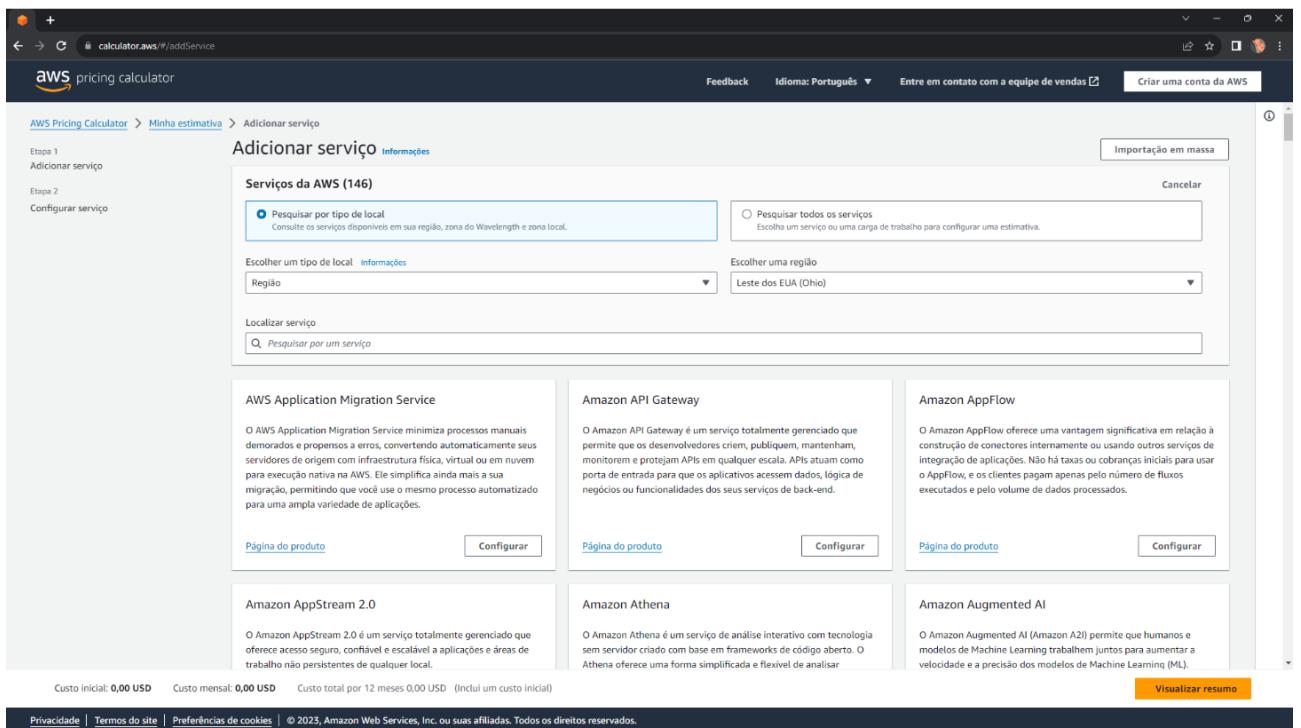


Figura 200: AWS Calculator

Fonte: Autoria Própria

Passo 3: Pesquise Todos os Serviços - Em "Serviços da AWS (146)", clique na caixa de seleção e troque para "pesquisar todos os serviços". Isso permitirá que você encontre e adicione os serviços específicos necessários para o projeto.

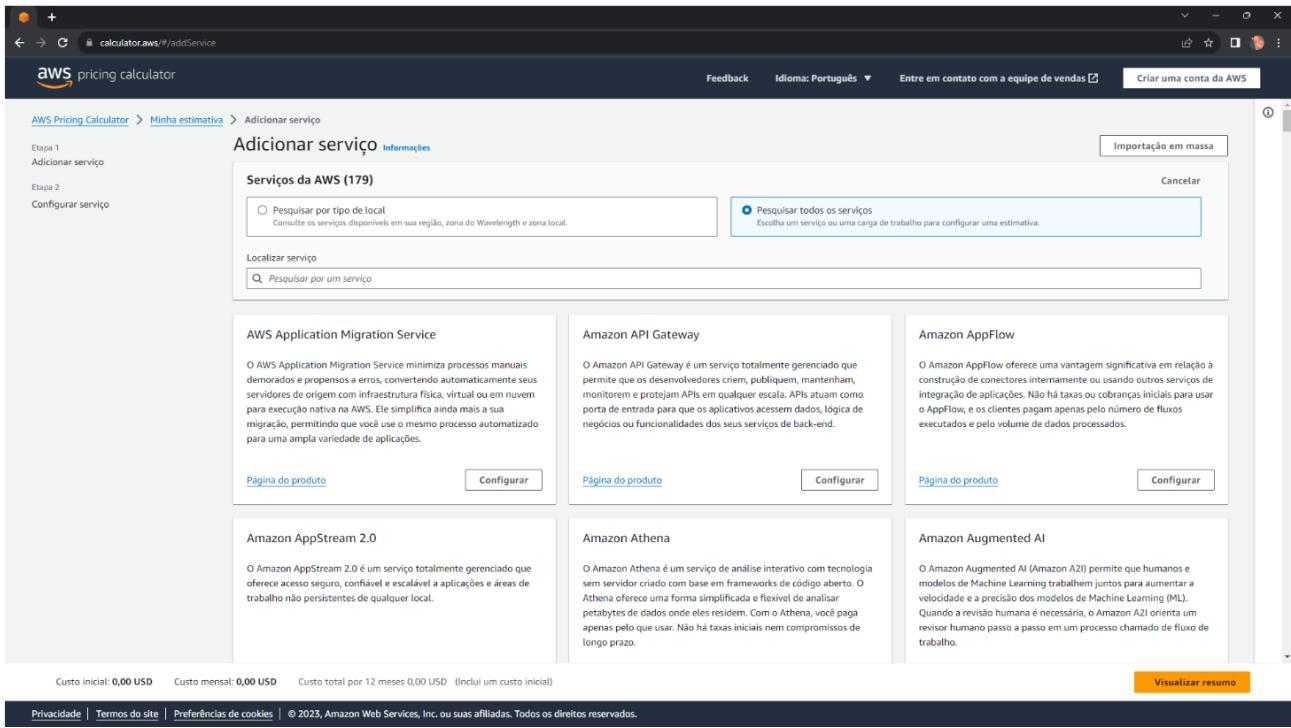


Figura 201: AWS Calculator

Fonte: Autoria Própria

14.2 Configurando serviços

Amazon API Gateway

O primeiro passo na análise de custos é adicionar o serviço "Amazon API Gateway". Siga os passos abaixo:

Passo 1: Pesquisar e Configurar o Amazon API Gateway - Na barra de pesquisa, digite "Amazon API Gateway" e clique em "configurar".

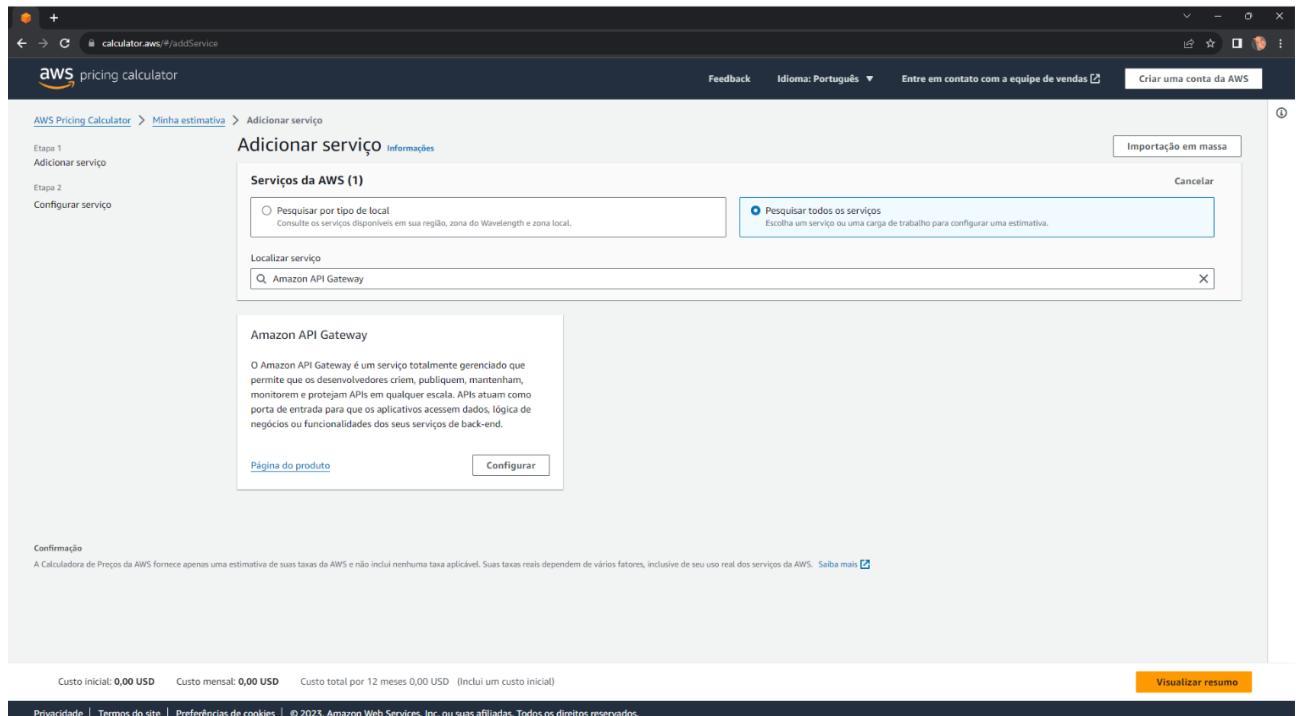


Figura 202: API Gateway

Fonte: Autoria Própria

Passo 2: Configurações Regionais e API REST - Na nova janela, selecione a região desejada, por exemplo, "leste dos EUA (N. da Virgínia)". Para este exemplo, deixaremos os valores relacionados às APIs HTTP zerados. Desça até a seção "API REST".

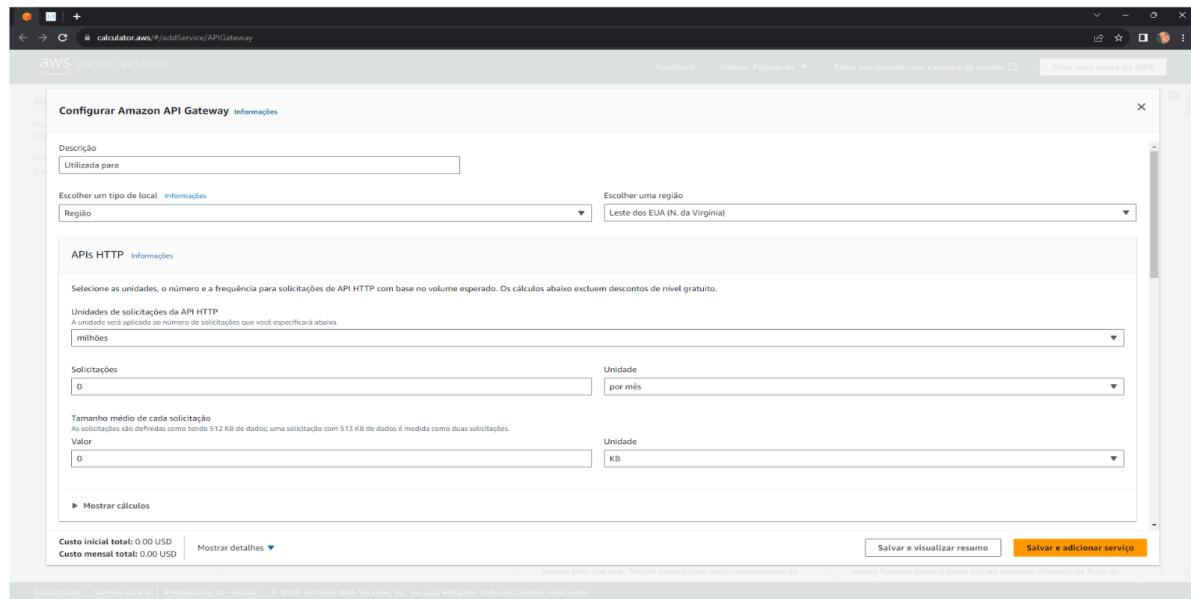


Figura 203: API Gateway

Fonte: Autoria Própria

Passo 3: No campo "Unidades de solicitação da API REST", defina como "Número exato" e insira 8 solicitações mensais (2 por semana). O campo "Tamanho da memória do cache (GB)" é opcional e pode ser deixado em branco por enquanto. Desça até a seção da API WebSocket.

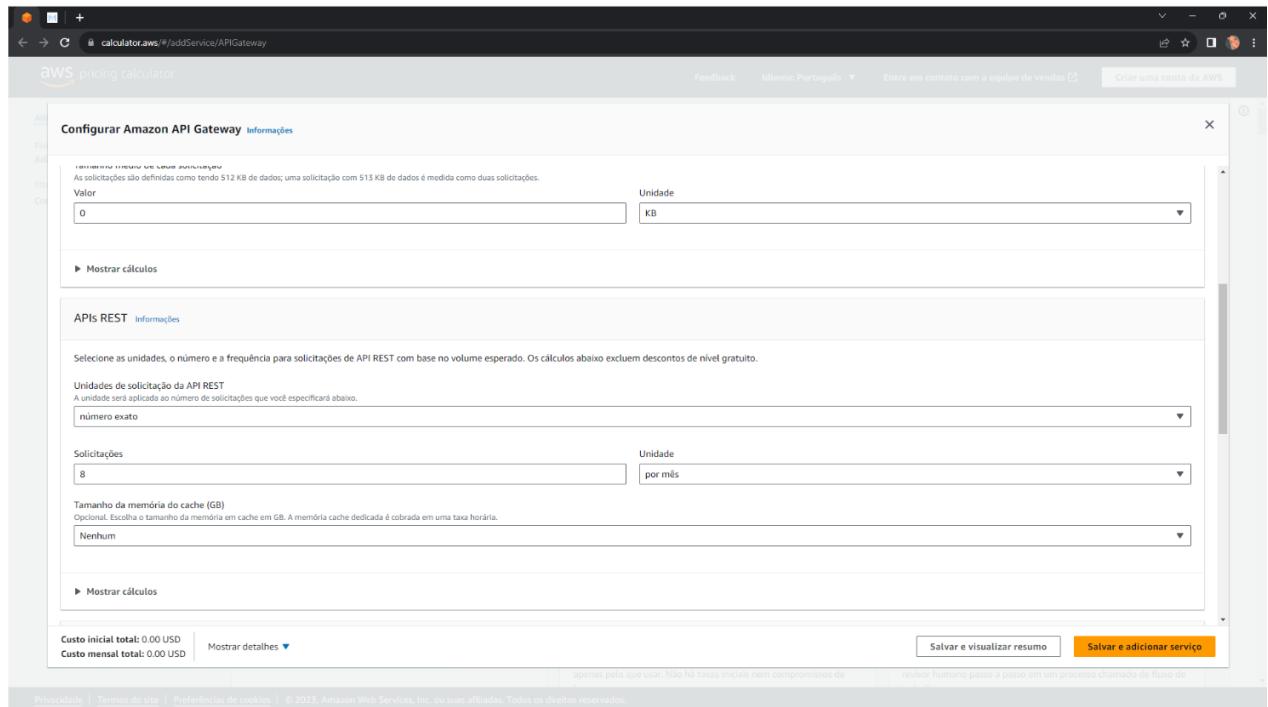


Figura 204: API Gateway

Fonte: Autoria Própria

Passo 4: Configuração da API WebSocket - Como não utilizaremos a API WebSocket neste projeto, preencha todos os campos numéricos com zero.

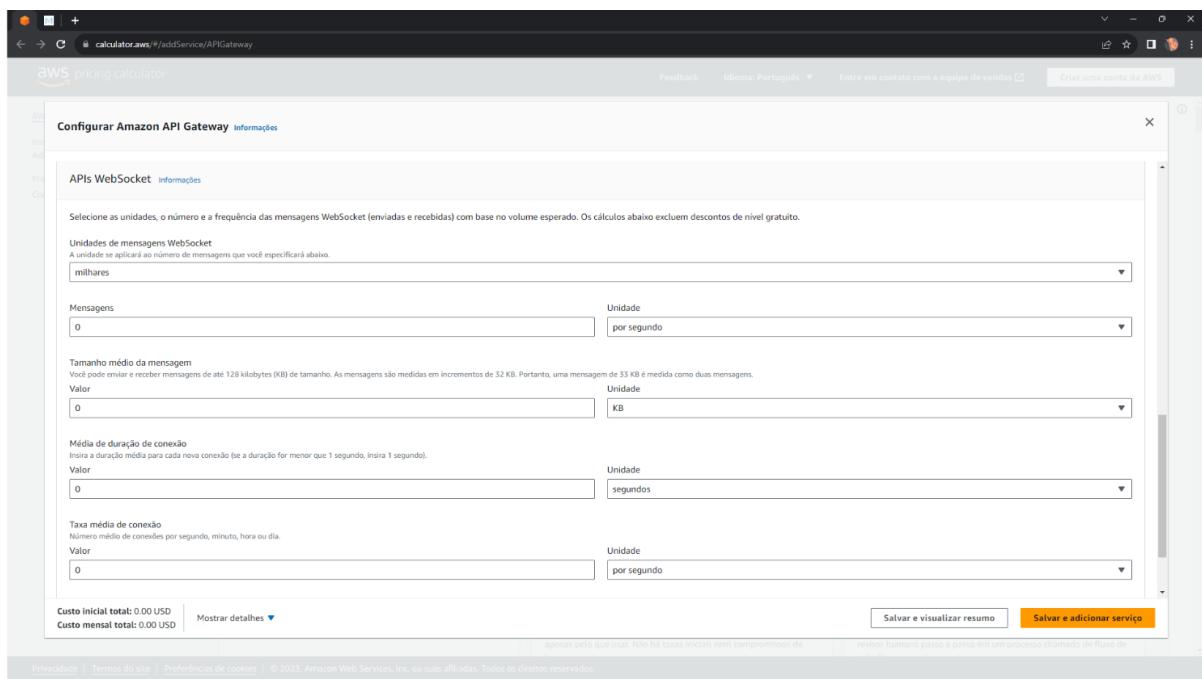


Figura 205: API Gateway

Fonte: Autoria Própria

Passo 5: Clique em "Salvar e adicionar serviço" para concluir a adição do Amazon API Gateway com as configurações especificadas.

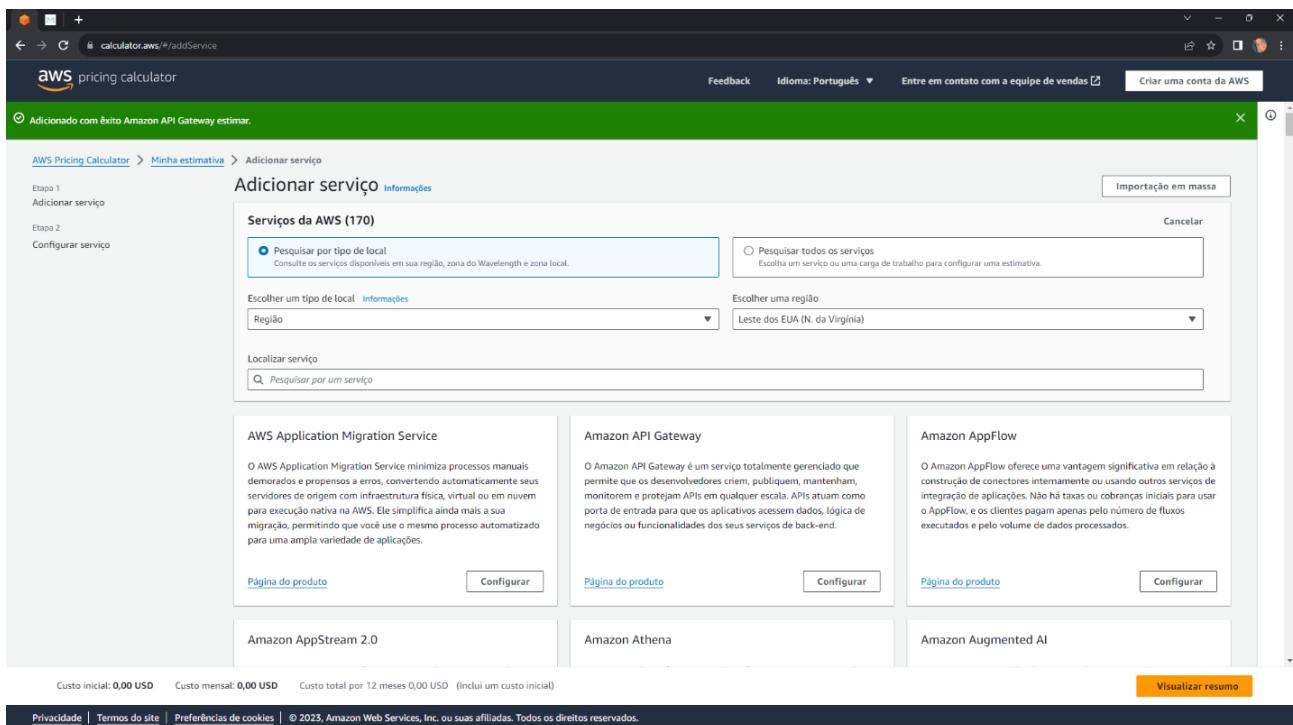


Figura 206: API Gateway

Fonte: Autoria Própria

Passo 6: Agora, você adicionou com sucesso o Amazon API Gateway à sua estimativa de custos na AWS Calculator.

Amazon Simple Storage Service (S3)

O segundo passo na análise de custos é adicionar o serviço "Amazon Simple Storage Service (S3)". Siga os passos abaixo:

Passo 1: Pesquisar e Configurar o Amazon S3 - Na barra de pesquisa, digite "Amazon Simple Storage Service (S3)" e clique em "configurar".

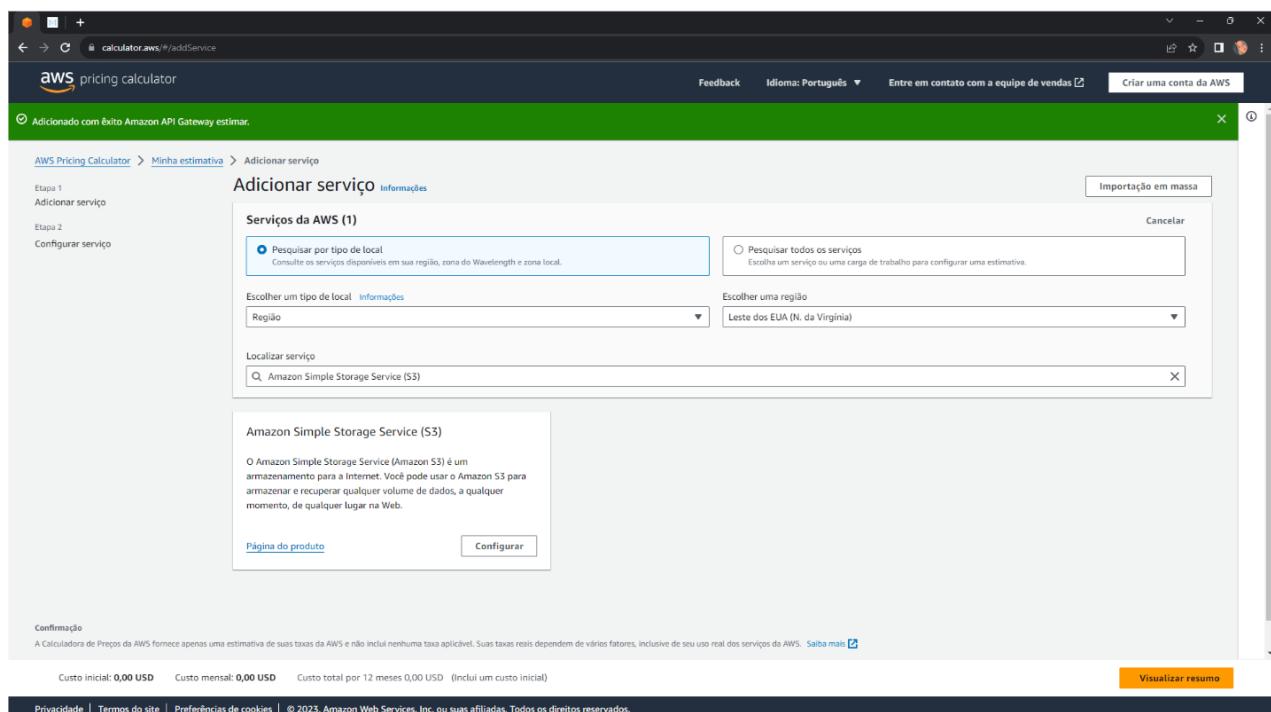


Figura 207: AWS S3

Fonte: Autoria Própria

Passo 2: Seleção da Classe de Armazenamento - Escolha a região desejada, por exemplo, "leste dos EUA (N. da Virgínia)". No campo "Selecionar classes de armazenamento no S3 e outros recursos", avalie as opções disponíveis:

- S3 Standard: Padrão para armazenamento de dados na nuvem. Oferece alta durabilidade, disponibilidade e desempenho, sendo adequado para dados frequentemente acessados.
- S3 Intelligent - Tiering: Automatiza a movimentação de dados entre camadas de armazenamento com base em padrões de acesso. Os dados mais frequentemente

acessados ficam em camadas mais rápidas, enquanto os menos acessados vão para camadas mais econômicas.

- S3 Standard - Infrequent Access: Semelhante ao S3 Standard, mas projetado para dados que são acessados com menos frequência. Oferece custos mais baixos de armazenamento, mas com taxas um pouco mais altas para acesso aos dados.
- S3 One Zone - Infrequent Access: Armazena dados em uma única zona de disponibilidade, tornando-o mais econômico. No entanto, não oferece a mesma durabilidade que o S3 Standard, pois está em uma única localização.
- S3 Glacier Flexible Retrieval: Parte do serviço Glacier, permite recuperar dados de maneira flexível, adaptando a velocidade de recuperação aos requisitos específicos.
- S3 Glacier Deep Archive: Também parte do serviço Glacier, é a opção mais econômica, mas projetada para dados que são acessados muito raramente.
- S3 Management and Insights: Fornece ferramentas para gerenciar e entender melhor o uso e os custos de armazenamento no Amazon S3.
- S3 Object Lambda: Executa código personalizado em resposta a solicitações de leitura de objetos no Amazon S3, possibilitando manipulação dinâmica durante a recuperação.
- S3 Glacier Instant Retrieval: Uma opção do Glacier que permite recuperar dados quase instantaneamente, adequada para situações em que a rapidez na recuperação é crítica.
- Data Transfer: Refere-se à transferência de dados para dentro e para fora do Amazon S3, que pode envolver custos adicionais dependendo da quantidade de dados transferidos.

Para este projeto, escolhemos "S3 Standard" devido ao acesso frequente aos dados.

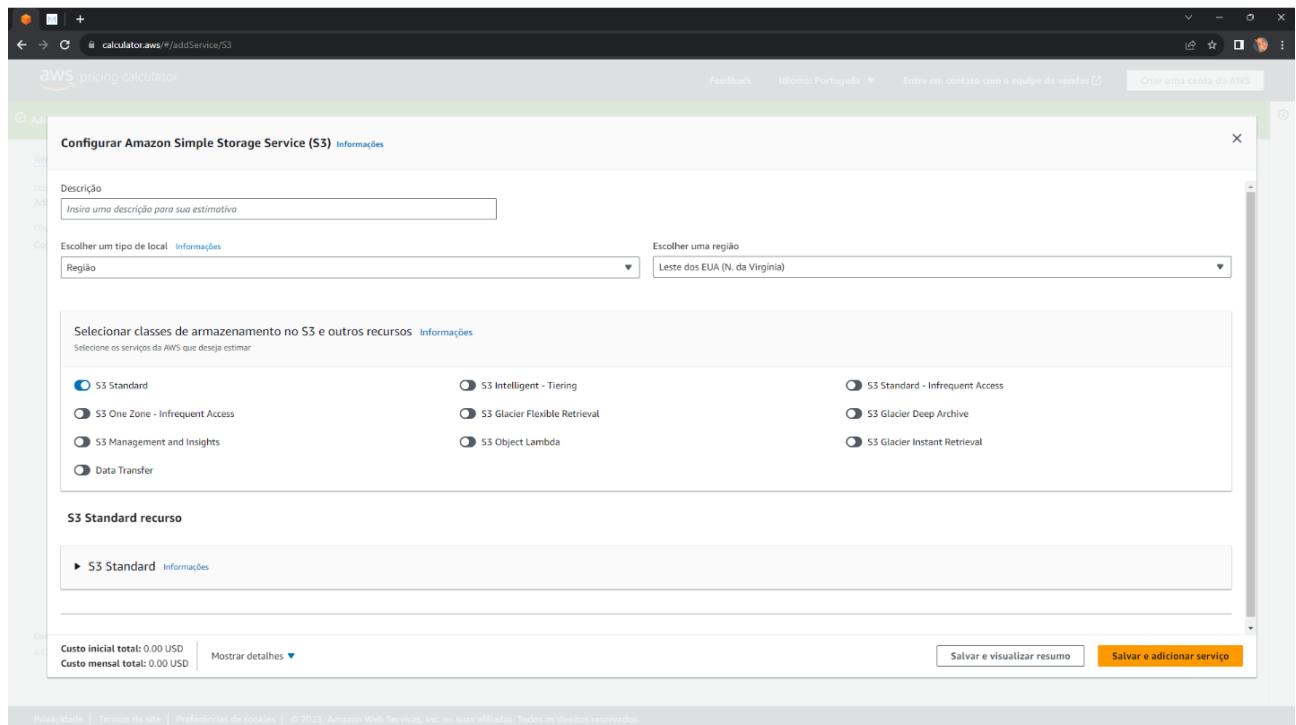


Figura 208: AWS S3

Fonte: Autoria Própria

Passo 3: Configuração do S3 Standard Recurso - Abra o toggle do serviço "S3 Standard recurso" e preencha os dados:

- "Armazenamento S3 Standard": 20 GB por mês.
- "A quantidade especificada de dados já está armazenada no S3 Standard".
- "Solicitações PUT, COPY, POST, LIST para S3 Standard": 50 solicitações.
- "GET, SELECT e todas as outras solicitações do S3 Standard": 40 solicitações.

Os campos "Dados retornados pelo S3 Select" e "Dados verificados pelo S3 Select" podem ser deixados em branco por enquanto, já que ainda não temos previsão específica para essas solicitações.

Configurar Amazon Simple Storage Service (S3) [Informações](#)

Os cálculos abaixo excluem os descontos do nível gratuito.

Armazenamento S3 Standard

20	Unidade	GB por mês
----	---------	------------

Como os dados serão movidos para S3 Standard?

Calcula automaticamente os custos de PUT, COPY e POST para mover dados para a categoria S3 Standard inicialmente. Para comparar o custo do armazenamento atual em S3 Standard com o custo da transferência do ciclo de vida desses dados para uma nova classe de armazenamento, especifique que o seu armazenamento já está em S3 Standard ao selecionar Lifecycle na nova classe de armazenamento para capturar o custo inicial da movimentação dos dados.

A quantidade especificada de dados já está armazenada no S3 Standard

Solicitações PUT, COPY, POST, LIST para S3 Standard

Número mensal contínuo de solicitações PUT, COPY, POST ou LIST	50
--	----

GET, SELECT e todas as outras solicitações do S3 Standard

Número mensal contínuo de solicitações GET, SELECT e todas as outras solicitações	40
---	----

Dados retornados pelo S3 Select

Volume mensal contínuo de dados retornados por solicitações do S3 Select

Valor	Unidade	GB por mês
Inserir quantidade		

Dados verificados pelo S3 Select

Volume mensal contínuo de dados verificados por solicitações do S3 Select

Valor	Unidade	GB por mês
Inserir quantidade		

Custo inicial total: 0,00 USD | Custo mensal total: 0,46 USD | [Mostrar detalhes](#) ▾

[Salvar e visualizar resumo](#) | [Salvar e adicionar serviço](#)

Figura 209: AWS S3

Fonte: Autoria Própria

Clique em "Salvar e adicionar serviço" para concluir a adição do Amazon S3 com as configurações especificadas.

Adicionado com êxito Amazon Simple Storage Service (S3) estimar.

[AWS Pricing Calculator](#) > [Minha estimativa](#) > Adicionar serviço

Adicionar serviço [Informações](#)

Serviços da AWS (170)

Pesquisar por tipo de local Consulte os serviços disponíveis em sua região, zona do Wavelength e zona local.

Pesquisar todos os serviços Escolha um serviço ou uma carga de trabalho para configurar uma estimativa.

Importação em massa [Cancelar](#)

Escolher um tipo de local [Informações](#)

Região	Escolher uma região
Leste dos EUA (N. da Virginia)	

Pesquisar por um serviço

AWS Application Migration Service

O AWS Application Migration Service minimiza processos manuais demorados e propensos a erros, convertendo automaticamente seus servidores de origem com infraestrutura física, virtual ou em nuvem para execução nativa na AWS. Ele simplifica ainda mais a sua migração, permitindo que você use o mesmo processo automatizado para uma ampla variedade de aplicações.

[Página do produto](#) [Configurar](#)

Amazon API Gateway

O Amazon API Gateway é um serviço totalmente gerenciado que permite que os desenvolvedores criem, publiquem, mantenham, monitorem e protejam APIs em qualquer escala. APIs atuam como porta de entrada para que os aplicativos acessem dados, lógica de negócios ou funcionalidades dos seus serviços de back-end.

[Página do produto](#) [Configurar](#)

Amazon AppFlow

O Amazon AppFlow oferece uma vantagem significativa em relação à construção de conectores internamente ou usando outros serviços de integração de aplicações. Não há taxas ou cobranças iniciais para usar o AppFlow, e os clientes pagam apenas pelo número de fluxos executados e pelo volume de dados processados.

[Página do produto](#) [Configurar](#)

Amazon AppStream 2.0

[Página do produto](#)

Amazon Athena

[Página do produto](#)

Amazon Augmented AI

[Página do produto](#)

Custo inicial: 0,00 USD | Custo mensal: 0,46 USD | Custo total por 12 meses 5,52 USD (Inclui um custo inicial)

[Visualizar resumo](#)

Figura 209: AWS S3

Fonte: Autoria Própria

Agora, você adicionou com sucesso o Amazon S3 à sua estimativa de custos na AWS Calculator

Amazon Redshift

O terceiro passo na análise de custos é adicionar o serviço "Amazon Redshift". Siga os passos abaixo:

Passo 1: Pesquisar e Configurar o Amazon Redshift - Na barra de pesquisa, digite "Amazon Redshift" e clique em "configurar".

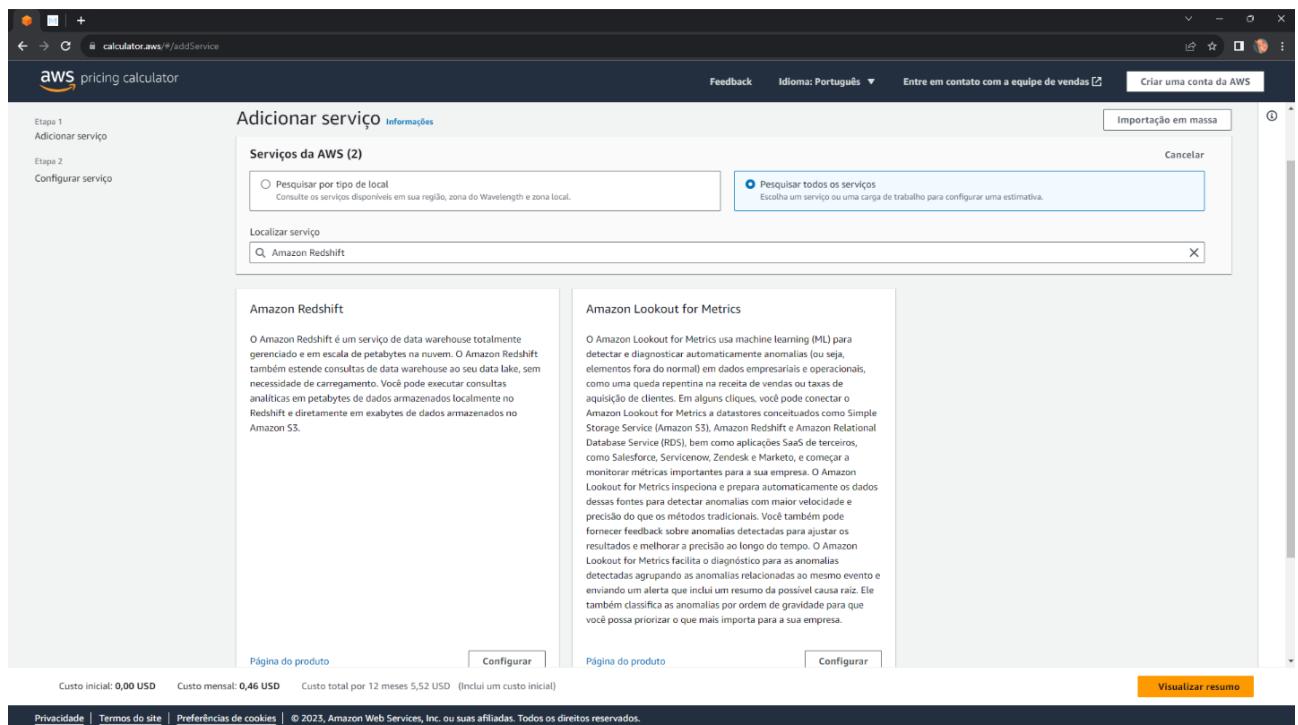


Figura 210: Amazon RedShift

Fonte: Autoria Própria

Passo 2: Configurações Regionais e Escolha da Modalidade "Redshift sem servidor" - Na nova janela, selecione a região desejada, por exemplo, "leste dos EUA (N. da Virgínia)". Escolha a opção "Redshift sem servidor".

Passo 3: Escolha do Tamanho da Workload - Avalie e escolha o tamanho da workload com base nas características abaixo:

- Pequeno: Volume de dados relativamente baixo, consultas simples, poucos usuários simultâneos, baixa demanda computacional.
- Médio: Volume de dados moderado, consultas mais complexas, número moderado de usuários simultâneos, maior demanda computacional.

- Grande: Grandes volumes de dados, consultas altamente complexas, muitos usuários simultâneos, requer capacidade computacional substancial.

Para este exemplo, escolheremos "Pequeno", pois o volume de dados é baixo.

Passo 4: Escolha da RPU Base

A RPU (Redshift Processing Unit) é uma unidade de medida que representa a capacidade de processamento. Escolha a RPU Base de acordo com a capacidade necessária. Para 15GB, selecionaremos a RPU Base de 16.

Passo 5: Definição do Tempo de Execução Diário Esperado

Informe o tempo de execução diário esperado, considerando a jornada de trabalho. Para este exemplo, consideraremos 3 horas diárias. Clique em "Salvar e adicionar serviço" para concluir a adição do Amazon Redshift com as configurações especificadas.

The screenshot shows the AWS Pricing Calculator interface. The user is on Step 1, 'Adicionar serviço'. They have selected 'Amazon Redshift' from a list of 179 services. The calculator displays the following estimated costs:

Custo inicial	Custo mensal	Custo total por 12 meses
0,00 USD	549,46 USD	6.595,52 USD (Inclui um custo inicial)

At the bottom, there are links for 'Privacidade', 'Termos de site', 'Preferências de cookies', and a copyright notice: '© 2023, Amazon Web Services, Inc. ou suas afiliadas. Todos os direitos reservados.'

Figura 211: Amazon RedShift

Fonte: Autoria Própria

Agora, você adicionou com sucesso o Amazon Redshift à sua estimativa de custos na AWS Calculator.

AWS Lambda

O quarto passo na análise de custos é adicionar o serviço "AWS Lambda". Siga os passos abaixo:

Passo 1: Pesquisar e Configurar o AWS Lambda - Na barra de pesquisa, digite "AWS Lambda" e clique em "configurar".

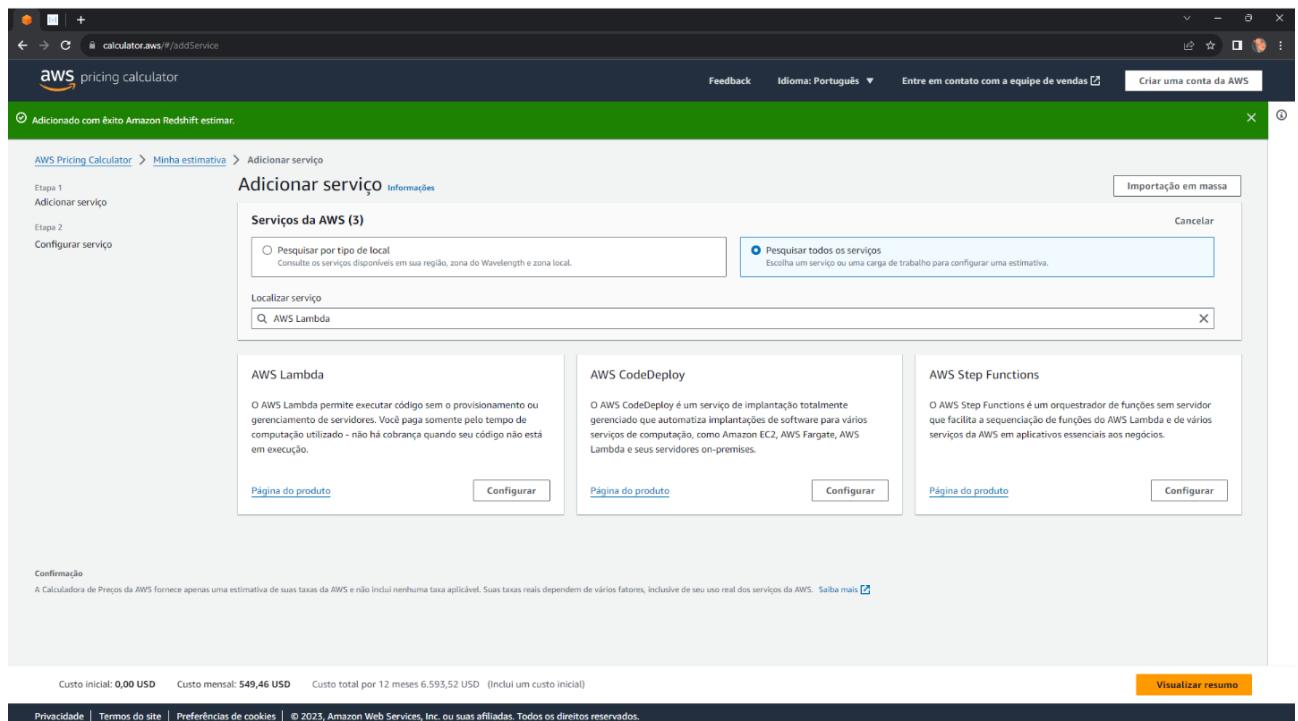


Figura 212: AWS Lambda

Fonte: Autoria Própria

Passo 2: Configurações Regionais - Na nova janela, escolha a região desejada, por exemplo, "leste dos EUA (N. da Virgínia)".

Passo 3: Utilização do Nível Gratuito - Para este exemplo, utilizaremos o nível gratuito que oferece 1 milhão de solicitações gratuitas por mês e 400.000 GB/segundo de tempo de computação por mês. Este serviço será utilizado para extrair ou excluir dados da API do cliente.

The screenshot shows the AWS Lambda Pricing Calculator interface. At the top, there's a navigation bar with links for Feedback, Idioma: Português, Entre em contato com a equipe de vendas, and Criar uma conta da AWS. Below the navigation, the title 'Configurar AWS Lambda' is displayed, along with a 'Informações' link.

Description: A text input field with placeholder text 'Insira uma descrição para sua estimativa'.

Escolher um tipo de local: A dropdown menu with 'Informações' and a 'Região' option selected, showing 'Leste dos EUA (N. da Virginia)'.

Função do Lambda - Incluir nível gratuito: Descrição: 'O Lambda contabiliza uma solicitação cada vez que começa a executar em resposta a uma notificação de evento ou chamada de invocação, incluindo invocações de testes com origem no console. Você é cobrado pelo número total de solicitações em todas as suas funções. O preço depende da quantidade de memória alocada para a sua função. O nível gratuito do Lambda inclui 1 milhão de solicitações gratuitas por mês e 400.000 GiB/segundos de tempo de computação por mês.'.

Escolher uma região: A dropdown menu with 'Leste dos EUA (N. da Virginia)' selected.

Função do Lambda - Sem nível gratuito: Descrição: 'O Lambda contabiliza uma solicitação cada vez que começa a executar em resposta a uma notificação de evento ou chamada de invocação, incluindo invocações de testes com origem no console. Você é cobrado pelo número total de solicitações em todas as suas funções. O preço depende da quantidade de memória alocada para a sua função. Os descontos do nível gratuito do Lambda estão excluídos.'

Configurações de serviços: A section with 'Informações' and a 'Arquitetura' dropdown set to 'x86'.

Número de solicitações: An input field with '1' and a dropdown 'Unidade' set to 'por mês'.

Duração de cada solicitação (em ms): A note stating 'A duração é calculada a partir do momento em que o código começa a ser executado até retornar ou encerrar de outra forma.' followed by an input field with placeholder 'Insira a duração em ms'.

Quantidade de memória alocada: A note stating 'Insira a quantidade entre 128MB e 10GB.' followed by an input field with placeholder '...'. Below it, 'Custo inicial total: 0.00 USD' and 'Custo mensal total: 0.00 USD' are shown.

Buttons: 'Mostrar detalhes', 'Salvar e visualizar resumo', and 'Salvar e adicionar serviço'.

Figura 213: AWS Lambda

Fonte: Autoria Própria

Passo 4: Preenchimento dos Campos

- Arquitetura: x86
- Número de Solicitações: 10.000
- Duração de Cada Solicitação (em ms): 3.000
- Quantidade de Memória Alocada: 5
- Quantidade de Armazenamento Temporário Alocada: 5

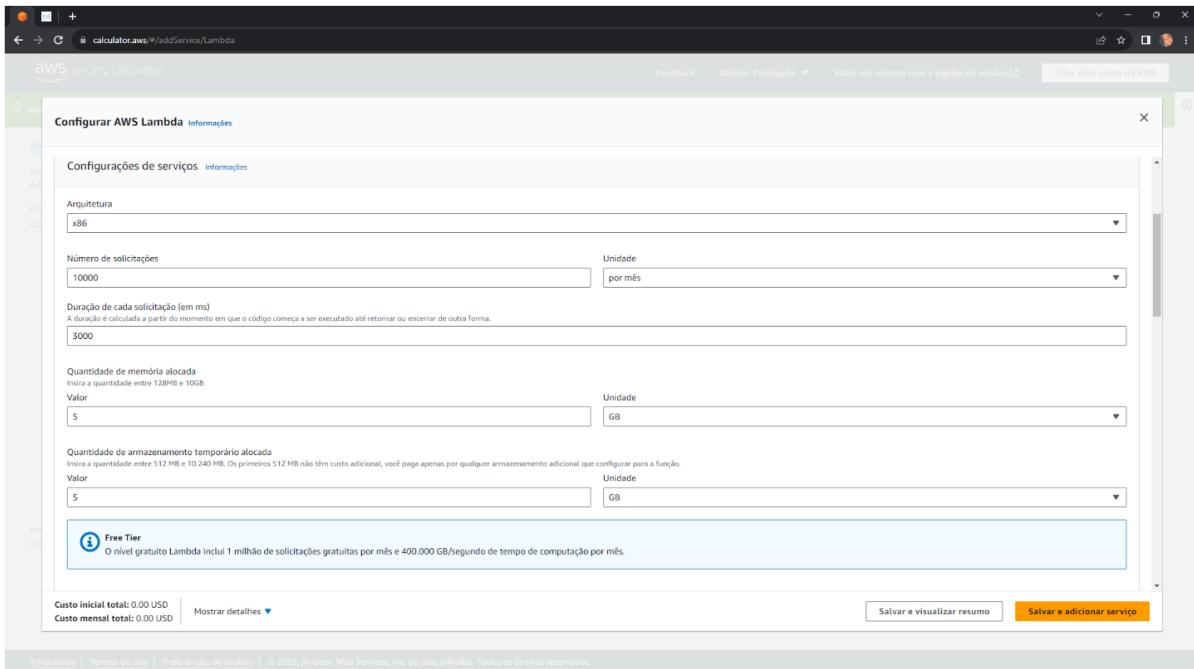


Figura 214: AWS Lambda

Fonte: Autoria Própria

Clique em "Salvar e adicionar serviço" para concluir a adição do AWS Lambda com as configurações especificadas.

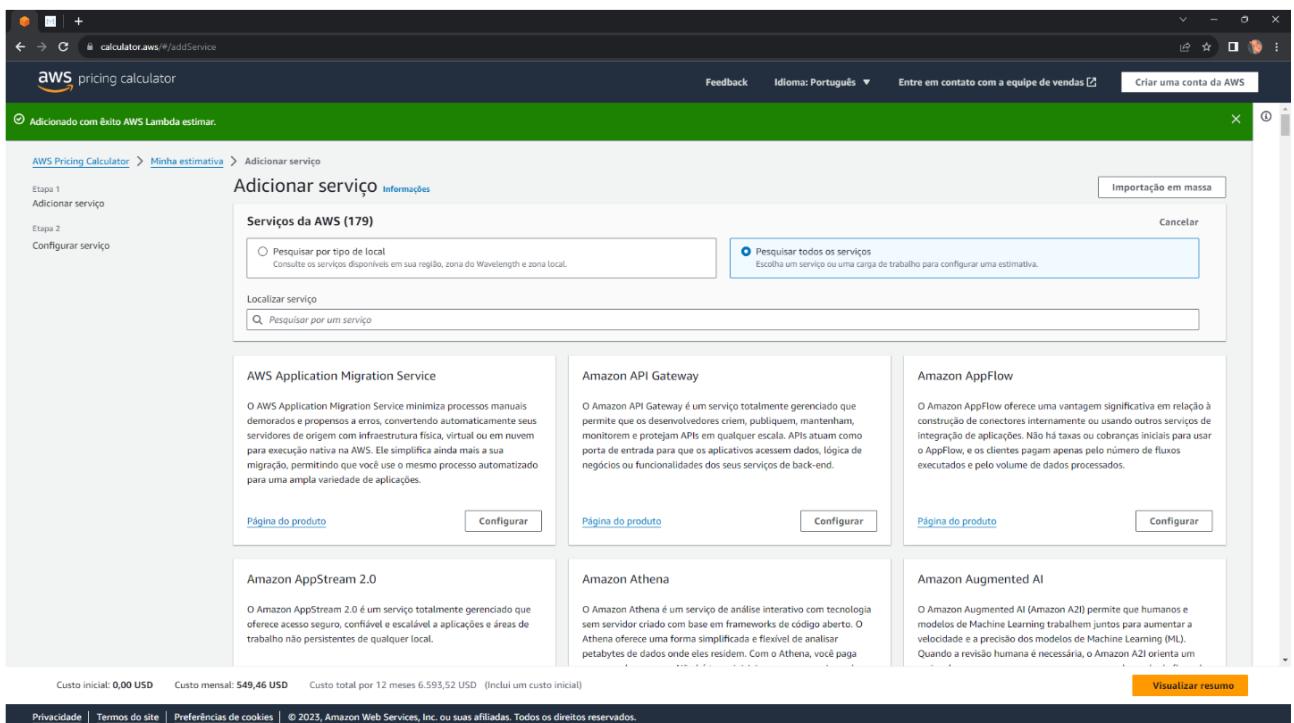


Figura 215: AWS Lambda

Fonte: Autoria Própria

Agora, você adicionou com sucesso o AWS Lambda à sua estimativa de custos na AWS Calculator.

Amazon EventBridge

O quinto passo na análise de custos é adicionar o serviço "Amazon EventBridge". Siga os passos abaixo:

Passo 1: Pesquisar e Configurar o Amazon EventBridge - Na barra de pesquisa, digite "Amazon EventBridge" e clique em "configurar".

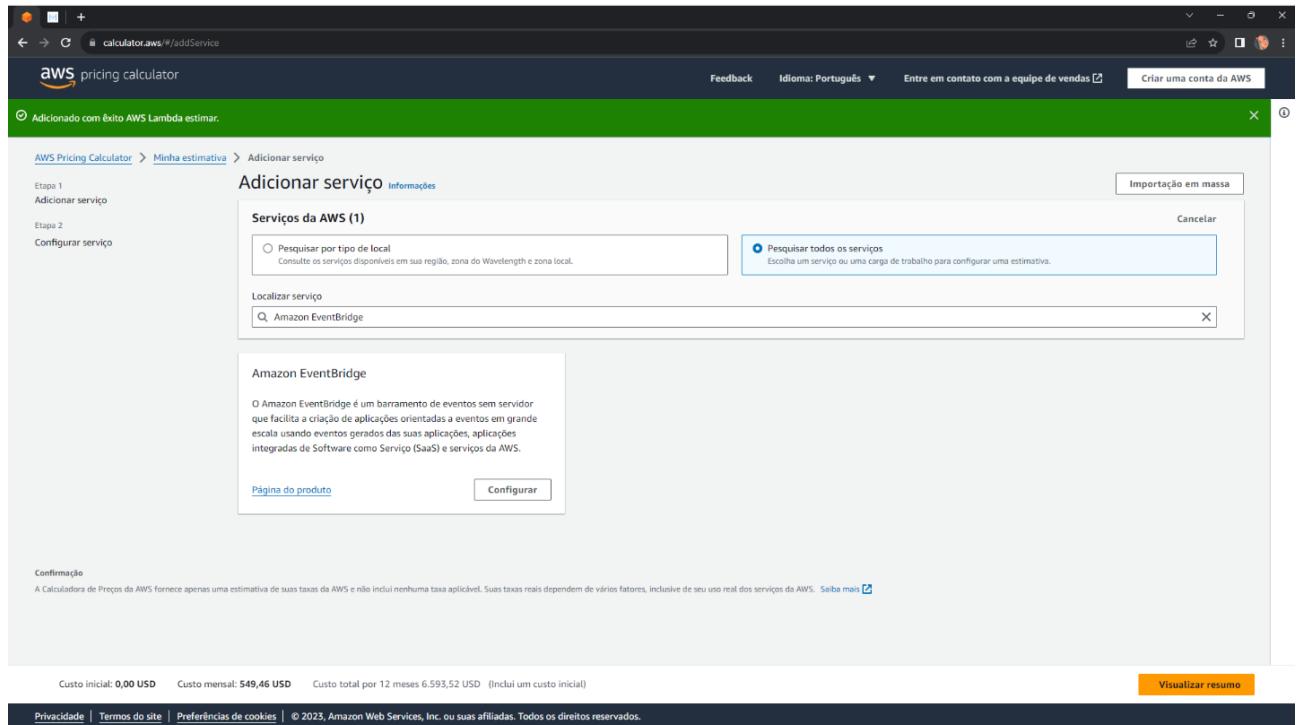


Figura 216: Amazon EventBridge

Fonte: Autoria Própria

Passo 2: Configurações Regionais - Na nova janela, escolha a região desejada, por exemplo, "leste dos EUA (N. da Virgínia)".

Passo 3: Preenchimento dos Campos

- Tamanho da Carga Útil (Payload): 800 KB: Refere-se ao tamanho máximo permitido para os dados incluídos em um evento. Neste caso, é limitado a 800 kilobytes.
- Número de Eventos Personalizados: 8 por Mês: Indica a quantidade de eventos que você pode criar e enviar de forma personalizada para o EventBridge em um período de um mês. Neste caso, você pode enviar até 8 eventos personalizados por mês.

- Número de Eventos de Parceiros: 8 por Mês: Representa a quantidade de eventos provenientes de parceiros ou serviços integrados ao EventBridge que você pode receber em um período de um mês. Limite de 8 eventos de parceiros por mês.
- Número de Eventos entre Regiões: 0 por Mês: Indica o número de eventos que podem ser enviados entre regiões. Neste caso, o limite é zero, o que significa que não há eventos permitidos entre diferentes regiões do AWS.
- Número Faturável entre Barramentos nos Eventos da Mesma Conta (Mensal): 0 por Mês: Refere-se à quantidade de eventos que podem ser faturados entre barramentos do EventBridge na mesma conta AWS. Neste caso, o limite é zero, indicando que não há eventos faturáveis entre esses barramentos.

The screenshot shows the AWS Pricing Calculator interface for configuring Amazon EventBridge. Key settings visible include:

- Location Type:** Escolher um tipo de local (Informações) - Região
- Region:** Escolher uma região - Leste dos EUA (N. da Virginia)
- Utilization:** Tamanho da carga útil - O tamanho da carga útil: 800 KB
- Amazon EventBridge Configuration:**
 - Número de eventos personalizados: 8 por mês
 - Número de eventos de parceiros: 8 por mês
 - Número de eventos entre regiões: 0 por mês
 - Número faturável entre barramentos nos eventos da mesma conta (mensal): 0 por mês
- Total Costs:** Custo inicial total: 0.00 USD | Custo mensal total: 0.00 USD

Figura 217: Amazon EventBridge

Fonte: Autoria Própria

- Número de Invocações: 20 por Mês: Indica a quantidade de vezes que um target (como uma função Lambda) pode ser invocado em resposta a eventos no EventBridge. Limite de 20 invocações por mês.
- Número de Eventos: 22 por Mês: Representa a quantidade total de eventos (soma de eventos personalizados, eventos de parceiros, etc.) que você pode gerenciar no EventBridge em um período de um mês. Neste caso, o limite é de 22 eventos por mês.

- Número de Eventos Reproduzidos: 22 por Mês: Indica quantos eventos você pode reproduzir (reenviar) no EventBridge em um período de um mês. Limite de 22 eventos reproduzidos por mês.

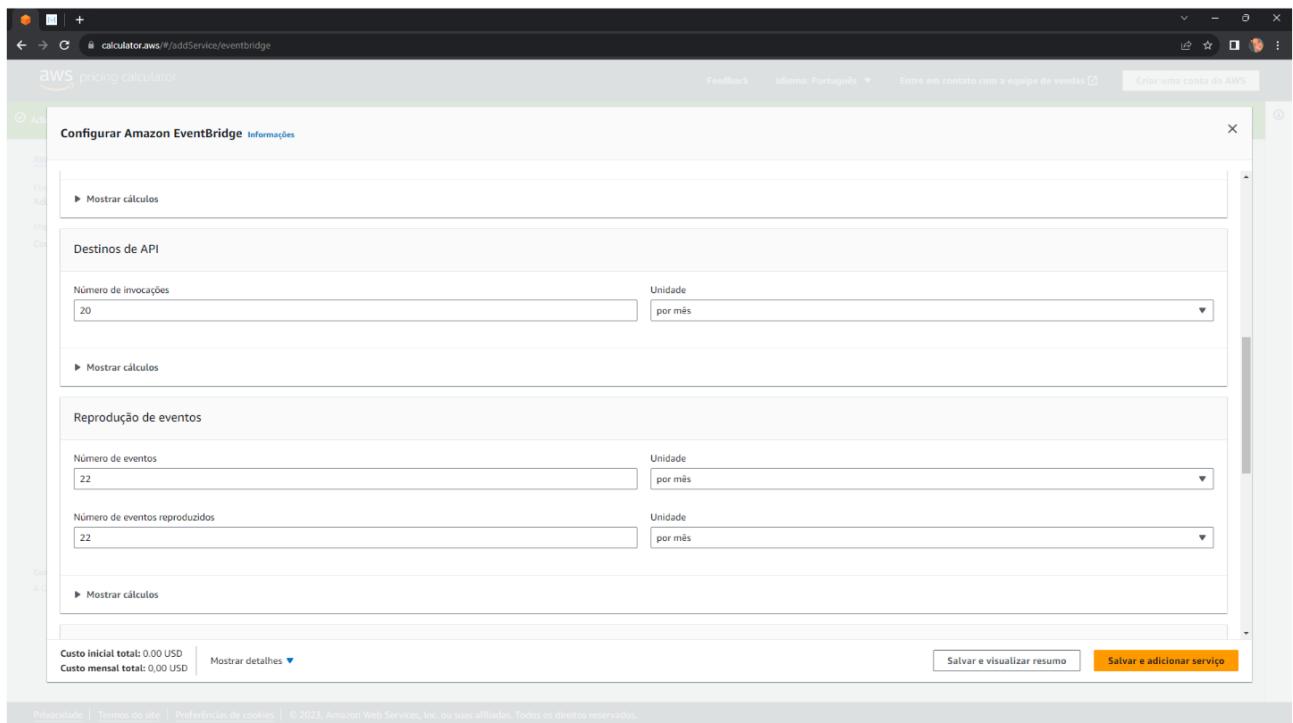


Figura 218: Amazon EventBridge

Fonte: Autoria Própria

Clique em "Salvar e adicionar serviço" para concluir a adição do Amazon EventBridge com as configurações especificadas.

The screenshot shows the AWS Pricing Calculator interface. The top navigation bar includes links for Feedback, Idioma: Português, Entre em contato com a equipe de vendas, and Criar uma conta da AWS. The main title is "Adicionar com êxito Amazon EventBridge estimar." Below this, the breadcrumb trail is "AWS Pricing Calculator > Minha estimativa > Adicionar serviço". The left sidebar has two tabs: "Etapa 1 Adicionar serviço" (selected) and "Etapa 2 Configurar serviço". The main content area is titled "Adicionar serviço" with a "Informações" link. It features a search bar with options "Pesquisar por tipo de local" and "Pesquisar todos os serviços" (selected). Below the search bar are two input fields: "Localizar serviço" and "Pesquisar por um serviço". A large grid of service cards is displayed, each with a "Página do produto" and "Configurar" button. The cards include:

- AWS Application Migration Service
- Amazon API Gateway
- Amazon AppFlow
- Amazon AppStream 2.0
- Amazon Athena
- Amazon Augmented AI

At the bottom of the page, there are links for "Privacidade", "Termos do site", "Preferências de cookies", and a copyright notice: "© 2023, Amazon Web Services, Inc. ou suas afiliadas. Todos os direitos reservados.". A summary at the bottom indicates a "Custo inicial: 0,00 USD", "Custo mensal: \$49,46 USD", and "Custo total por 12 meses: 6.593,52 USD (Inclui um custo inicial)". A "Visualizar resumo" button is also present.

Figura 219: Amazon EventBridge

Fonte: Autoria Própria

Agora, você adicionou com sucesso o Amazon EventBridge à sua estimativa de custos na AWS Calculator.

Amazon EC2

O sexto e último passo na análise de custos é adicionar o serviço "Amazon EC2". Siga os passos abaixo:

Passo 1: Pesquisar e Configurar o Amazon EC2 - Na barra de pesquisa, digite "Amazon EC2" e clique em "configurar".

AWS Pricing Calculator > Minha estimativa > Adicionar serviço

Adicionar serviço

Serviços da AWS (7)

Pesquisar por tipo de local
Pesquisar todos os serviços

Localizar serviço
Q. Amazon EC2

Amazon EC2
O Amazon EC2 oferece uma ampla seleção de tipos de instância, otimizados para atender a diferentes casos de uso. Tipos de instância abrangem combinações variadas de CPU, memória, armazenamento e capacidade de rede e oferecem a flexibilidade de escolher a combinação adequada de recursos para suas aplicações.

[Página do produto](#) [Configurar](#)

Windows Server and SQL Server on Amazon EC2
A calculadora do Windows Server e do SQL Server no Amazon EC2 fornece uma estimativa de preço para cargas de trabalho específicas. Ela recomenda opções de implantação de nuvem adequadas e modelos de definição de preço econômicos com base em informações de licenciamento e infraestrutura.

[Página do produto](#) [Configurar](#)

Amazon Elastic Block Store (EBS)
O Amazon Elastic Block Storage (EBS) permite criar volumes de armazenamento em bloco persistentes e anexá-los a instâncias do Amazon EC2.

[Página do produto](#) [Configurar](#)

Amazon EMR
O Amazon EMR é a plataforma de big data na nuvem líder do setor para processar grandes quantidades de dados usando ferramentas de código aberto, como Apache Spark, Apache Hive, Apache HBase, Apache Flink, Apache Hudi e Presto. O Amazon EMR facilita a configuração, a operação e a escalabilidade de ambientes de big data.

[Página do produto](#) [Configurar](#)

AWS Audit Manager
O AWS Audit Manager ajuda você a auditar continuamente seu uso da AWS para simplificar a forma como você avalia o risco e a conformidade. Quando você definir e iniciar uma avaliação com base em uma estrutura de avaliação, o Audit Manager executará uma avaliação de recursos para cada recurso individual, como as instâncias

[Página do produto](#) [Configurar](#)

AWS CodeDeploy
O AWS CodeDeploy é um serviço de implantação totalmente gerenciado que automatiza implantações de software para vários serviços de computação, como Amazon EC2, AWS Fargate, AWS Lambda e seus servidores on-premises.

[Página do produto](#) [Configurar](#)

Custo inicial: 0,00 USD Custo mensal: 549,46 USD Custo total por 12 meses: 6.593,52 USD (Inclui um custo inicial)

Visualizar resumo

Privacidade | Termos do site | Preferências de cookies | © 2023, Amazon Web Services, Inc. ou suas afiliadas. Todos os direitos reservados.

Figura 220: Amazon EC2

Fonte: Autoria Própria

Passo 2: Configurações Regionais - Na nova janela, escolha a região desejada, por exemplo, "leste dos EUA (N. da Virgínia)".

Passo 3: Preenchimento dos Campos

- Locação: Instâncias Compartilhadas: Refere-se ao fato de as instâncias estarem sendo executadas em hardware físico compartilhado com outras instâncias na mesma máquina física.
- Sistema Operacional: Linux: Indica que as instâncias estão utilizando o sistema operacional Linux.
- Carga de Trabalho: Consistente, Número de Instâncias: 2: Descreve a natureza consistente da carga de trabalho e especifica que há duas instâncias em execução.

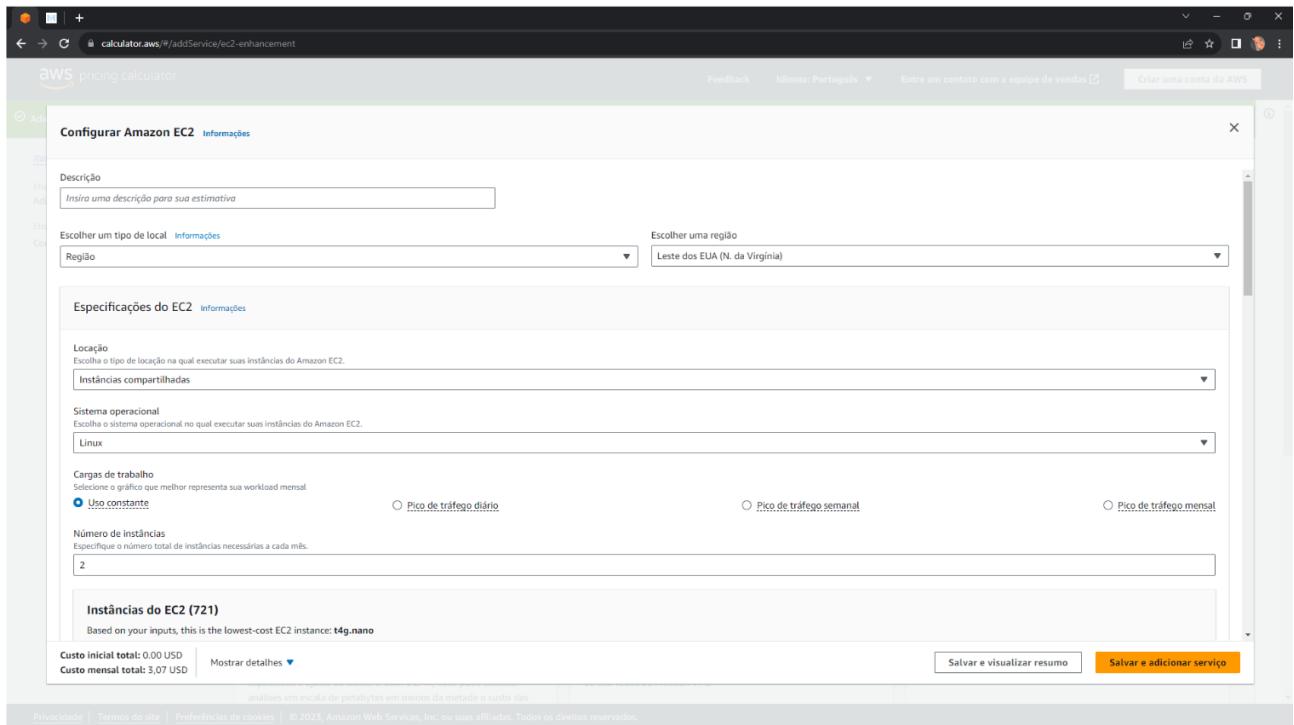


Figura 221: Amazon EC2

Fonte: Autoria Própria

- Instância do EC2 Avançada: t4g.medium: Memória 4 GB: Essa instância pertence à família t4g, que é uma família de instâncias baseadas em ARM (Amazon EC2 Graviton2). As instâncias Graviton2 são projetadas pela AWS e são baseadas na arquitetura ARM, oferecendo desempenho eficiente e otimizado para cargas de trabalho variadas. Para processar os dados são necessários minimamente 4GB.

Figura 222: Amazon EC2

Fonte: Autoria Própria

- Estratégia de Preços: Utilização On-Demand: 100% Utilizado/Mês: Refere-se à estratégia de preços, indicando que as instâncias estão sendo pagas sob demanda e estão 100% utilizadas durante todo o mês.

Opção de pagamento	Preço mensal (USD)
Sob demanda	12.19/Mês
Compute Savings Plans	12.19 (66% desconto)
Reserva (3 anos)	12.19 (desconto)
Instances Spot	12.19 (54% desconto)

Figura 223: Amazon EC2

Fonte: Autoria Própria

Clique em "Salvar e adicionar serviço" para concluir a adição do Amazon EC2 com as configurações especificadas.

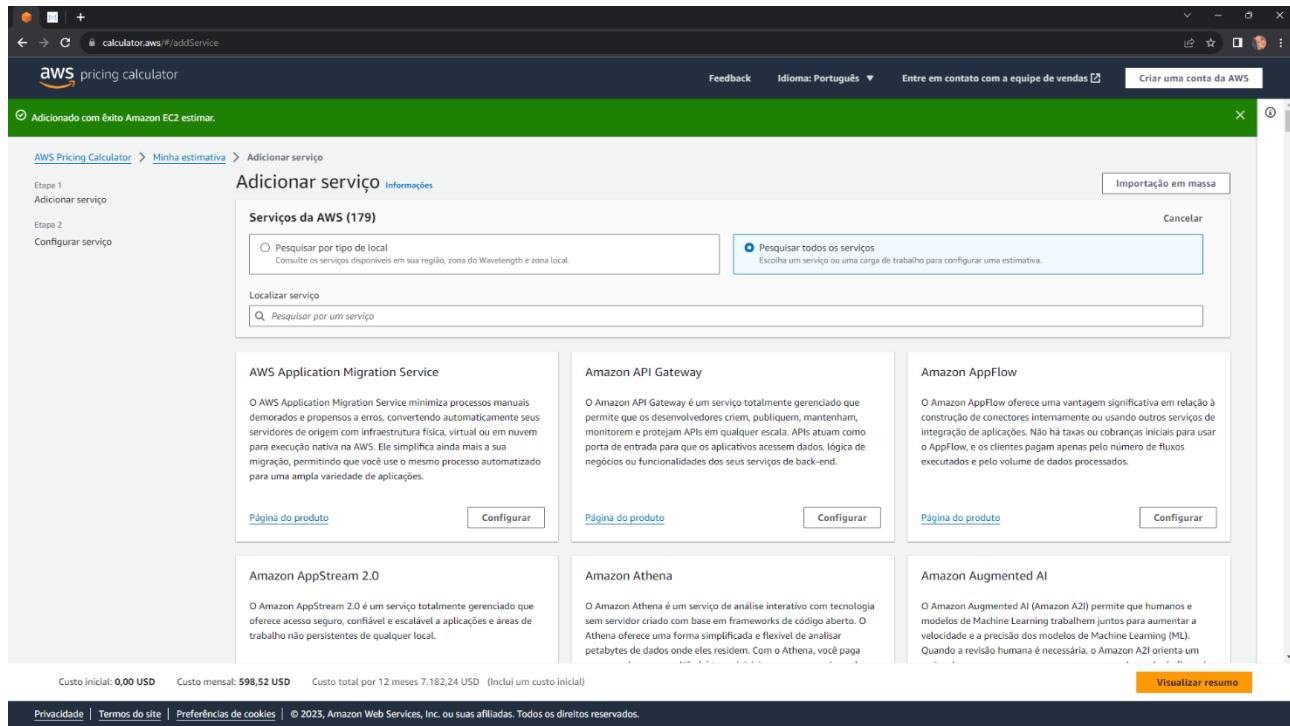


Figura 224: Amazon EC2

Fonte: Autoria Própria

Agora, você adicionou com sucesso o Amazon EC2 à sua estimativa de custos na AWS Calculator.

Visualizar - Para visualizar o resumo da estimativa, clique em "Visualizar resumo".

Custo inicial	Custo mensal	Custo total de 12 months
0,00 USD	598,52 USD	7.182,24 USD Inclui um custo inicial

Nome do serviço	Status	Custo inicial	Custo mensal	Descrição	Região	Resumo da configuração
Amazon API Gateway	-	0,00 USD	0,00 USD	Utilizada para	Leste dos EUA (N. da Virgínia)	Unidades de solicitação da API REST (nú...
Amazon Simple Storage Service (S3)	-	0,00 USD	0,46 USD	-	Leste dos EUA (N. da Virgínia)	Armazenamento S3 Standard (20 GB po...
Amazon Redshift	-	0,00 USD	549,00 USD	-	Leste dos EUA (N. da Virgínia)	RPU base (16), Tempo de execução diári...
AWS Lambda	-	0,00 USD	0,00 USD	-	Leste dos EUA (N. da Virgínia)	Arquitetura (x86), Arquitetura (x86), Mo...
Amazon EventBridge	-	0,00 USD	0,00 USD	-	Leste dos EUA (N. da Virgínia)	O tamanho da carga útil (800 KB), Núm...
Amazon EC2	-	0,00 USD	49,06 USD	-	Leste dos EUA (N. da Virgínia)	Locação (Instâncias compartilhadas), Sis...

Figura 225: Amazon EC2

Fonte: Autoria Própria

- Amazon API Gateway:
 - Utilizada para manipulação da API do parceiro (GET).
- Amazon Simple Storage Service (S3):
 - Utilizado para armazenamento dos dados em CSV (Data Lake).
- Amazon Redshift:
 - Utilizado para o carregamento dos dados e criação das views (Data Warehouse).
- AWS Lambda:
 - Utilizada para funções que auxiliam na manipulação da API do parceiro.
- Amazon EventBridge:
 - Utilizado para atualizar a API do parceiro semanalmente.
- Amazon EC2:
 - Utilizado para manter o Metabase conectado (Infográfico/Dashboard).

The screenshot shows the AWS Pricing Calculator interface. At the top, there are navigation icons, a search bar, and links for Feedback, Idioma: Português, Entre em contato com a equipe de vendas, and Criar uma conta da AWS. Below the header, the page title is "My Estimate" with an "Editar" link. On the left, a sidebar titled "Resumo da estimativa" shows "Custo inicial: 0,00 USD" and "Custo mensal: 598,52 USD". To the right, a summary box displays "Custo total de 12 months: 7.182,24 USD" (Inclui um custo inicial). A sidebar titled "Conceitos básicos da AWS" includes links to "Comece a usar gratuitamente" and "Entre em contato com a equipe de vendas". The main content area is titled "My Estimate" and contains a table with columns: Nome do serviço, Status, Custo i..., Custo m..., Descrição, Região, and Resumo da configuração. The table lists several AWS services with their respective details. At the bottom of the page, there are links for "Confirmação", "Privacidade", "Termos do site", "Preferências de cookies", and a copyright notice: "© 2023, Amazon Web Services, Inc. ou suas afiliadas. Todos os direitos reservados."

Figura 226: Cálculo Final

Fonte: Autoria Própria

Para visualizar o PDF criado pela AWS com a estimativa, [clique aqui](#).

14.3 Comparação - AWS X Azure

Pensando na eventualidade de o cliente desejar migrar os dados do pipeline de big data para a Azure, estimamos os custos relacionados à infraestrutura na Azure, considerando a mesma quantidade de dados (15GB), serviços e configurações utilizados atualmente na AWS.

Observação: Utilizamos o Pricing Calculator da Azure para realizar a estimativa, seguindo os mesmos passos da AWS. Abaixo apresenta-se a estimativa gerada.

Para visualizar o PDF criado pela Azure com a estimativa, [clique aqui](#) ou acesse pelo link <https://azure.com/e/84215a92bd334ddc85d8037607144494>.

14.4 Custos - Time de desenvolvimento

O projeto em questão tem uma duração estimada de 10 semanas, envolvendo a colaboração de uma equipe composta por 6 desenvolvedores. Para a realização dessa estimativa salarial, consideramos a média salarial de desenvolvedores juniores nos

Estados Unidos, que é de \$5,942.00 por mês, de acordo com informações obtidas no [Glassdoor](#).

A seguir, apresenta-se uma tabela sumarizando os principais dados do projeto:

Item	Descrição	Valor
Duração do Projeto	10 Semanas	-
Número de Desenvolvedores	6 Pessoas	-
Salário mensal dev júnior (\$)	Média	\$5,942.00
Horas trabalhadas totais	-	162
Remuneração por hora	2 meses de salário / Quant. Horas trabalhadas	\$73.36

Figura 227: Tabela Principais dados

Fonte: Autoria Própria

Fonte da média salarial: [Glassdoor](#)

Dólar (taxa de câmbio): \$1 = 4.92

14.5 Cálculo do Custo Total

1. Custo total dos salários dos desenvolvedores:

Salário mensal de um desenvolvedor júnior: \$5,942.00

Número de desenvolvedores: 6

Duração do projeto: 10 semanas (considerando 4 semanas por mês)

Custo total dos salários = $\$5,942.00 * 6 * 10/4 = \$89,130.00$

2. Custo total da infraestrutura:

Custo mensal da infraestrutura: \$598.52

Custo total da infraestrutura = $\$598.52 * 10/4 = \$1,496.30$

Custo total do projeto:

Soma do custo total dos salários e do custo total da infraestrutura.

Custo total do projeto = \$89,130.00 + \$1,496.30 = \$90,626.30

Cálculo do Custo Total do Projeto com Azure

1. Custo total dos salários dos desenvolvedores:

Salário mensal de um desenvolvedor júnior: \$5,942.00

Número de desenvolvedores: 6

Duração do projeto: 10 semanas (considerando 4 semanas por mês)

Custo total dos salários = \$5,942.00 * 6 * 10/4 = \$89,130.00

2. Custo total da infraestrutura:

Custo mensal da infraestrutura: \$530.53

Custo total da infraestrutura = \$530.53 * 10/4 = \$1,326.32

Custo total do projeto:

Soma do custo total dos salários e do custo total da infraestrutura.

Custo total do projeto = \$89,130.00 + \$1,326.32 = \$90,456.32

15. Plano de Comunicação

Obejtivo | 1

Definir o objetivo da comunicação para assegurar o alinhamento e compreensão entre todos envolvidos

+

Stakeholders | 5

Usuários finais do sistema em desenvolvimento

Datadream (Grupo desenvolvedor da solução) e outros grupos da Turma 04.

Consultores de marketing e vendas da Integration

Analista de dados da integration

Equipe docente de apoio do Inteli

+

Mensagens chaves | 4

Escopo e benefícios esperados pelos usuários finais e colaboradores da Integration

Atualizações regulares sobre o progresso do projeto

Desafios e possíveis dificuldades que podem ser encontrados

Incentivo de troca de feedback, incentivos e sugestões por parte de todos envolvidos.

+

Figura 228: Plano de Comunicação 1

Fonte: Autoria Própria

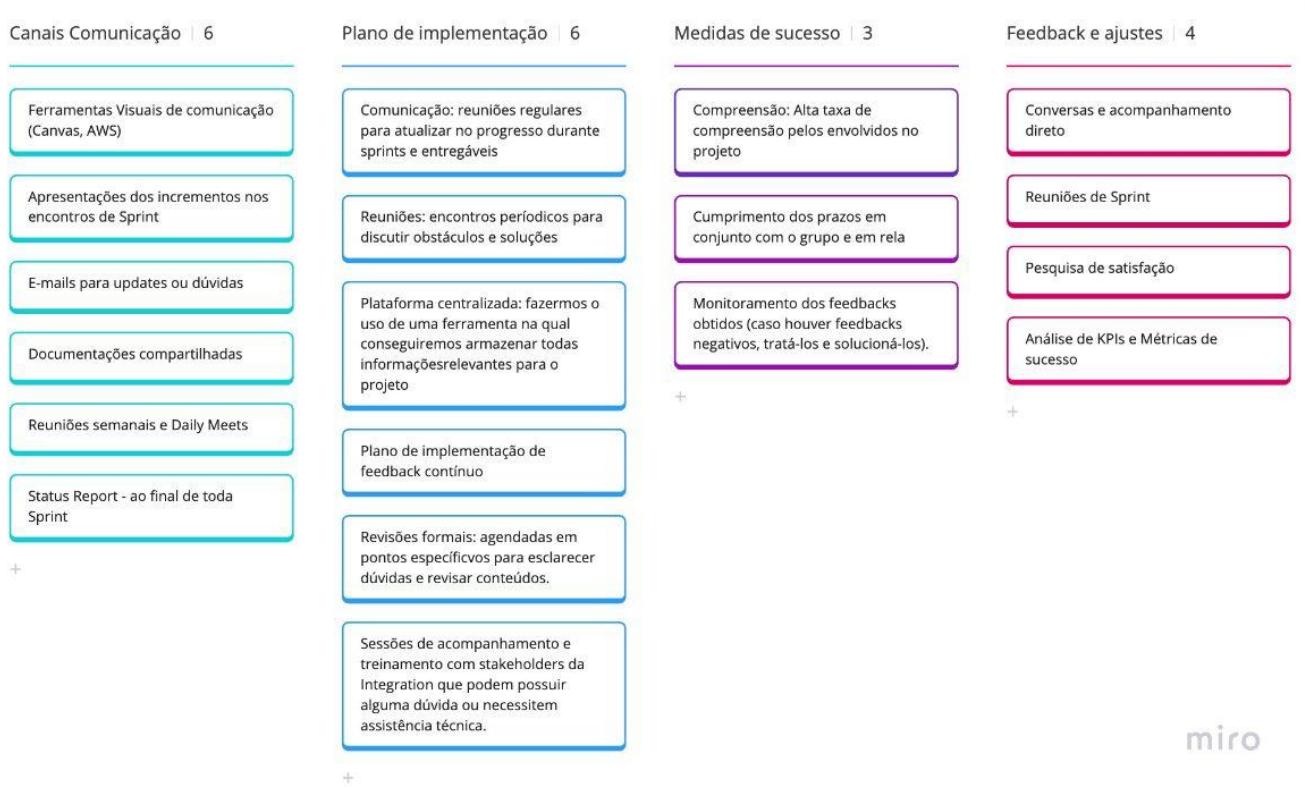


Figura 229: Plano de Comunicação 2

Fonte: Autoria Própria

O Plano de Comunicação é uma peça-chave para o sucesso de um projeto, e tem como intuito fornecer e estabelecer as bases de como as informações serão transmitidas, recebidas e gerenciadas dentro de um projeto ou organização. Dessa forma, serve como um guia abrangente para todas atividades de comunicação e visa garantir com que as mensagens sejam efetivamente entregues aos públicos-alvo, que a compreensão seja alcançada e que o feedback seja coletado de maneira sistemática.

15.1 Objetivo

O propósito claro do plano de comunicação é assegurar o alinhamento e compreensão entre todos os participantes do projeto, estabelecendo uma base sólida para um progresso coordenado e eficiente. Em projetos extensos, diversas peças fundamentais desempenham papéis cruciais em várias fases da gestão e desenvolvimento. Este plano destaca os principais stakeholders que desempenham um papel vital.

15.2 Stakeholders

Destacamos inicialmente o impacto significativo da equipe da Data Dream, que, com o suporte da equipe docente do Inteli, é responsável por criar a solução mais

impactante e valiosa para os usuários finais, um grupo também crucial para o projeto. Além disso, os consultores de marketing e o analista de dados da Integration desempenham papéis essenciais, pois a solução está sendo desenvolvida para atender às suas necessidades. É imperativo ouvir seus feedbacks, opiniões e sugestões para garantir a criação da melhor solução possível. Compreender a relevância e o impacto desses stakeholders no projeto é fundamental para uma comunicação eficaz.

15.3 Mensagens chaves

No cerne do nosso Plano de Comunicação residem as nossas mensagens chave, delineadas para fornecer uma direção clara e relevante a cada stakeholder envolvido no projeto. Dessa forma, ao considerar os usuários finais, focamos nas mensagens que abrangem o escopo do projeto e nos benefícios esperados. Reconhecemos a importância de manter esses stakeholders informados sobre o progresso do projeto, destacando as atualizações regulares como uma prioridade para garantir sua compreensão e envolvimento contínuo. Além disso, conscientes dos desafios e dificuldades que podem surgir, incluímos esses aspectos em nossas mensagens, demonstrando transparência e preparando os usuários finais para possíveis obstáculos.

Ao nos voltarmos para a equipe da Data Dream e a equipe docente do Inteli, adaptamos as mensagens chave para refletir não apenas os objetivos gerais do projeto, mas também a relevância específica de seu impacto na concepção e desenvolvimento da solução. Enfatizando sempre a importância do seu papel na criação da solução.

Reconhecemos que a solução que estamos desenvolvendo visa atender às necessidades específicas dos consultores de marketing e dos analistas de dados da Integration. Portanto, adaptamos nossas mensagens chave para incluir um incentivo explícito à troca de feedback, opiniões e sugestões, reconhecendo a sua experiência como valiosa para aprimorar o produto final.

15.4 Canais de comunicação

Os canais de comunicação selecionados em nosso plano foram cuidadosamente escolhidos para atender de forma precisa e eficaz às demandas específicas de cada grupo de stakeholders. Ao direcionar nossa atenção para os usuários finais, optamos por ferramentas visuais, como Canvas e AWS, e apresentações durante os encontros de Sprint. Essa abordagem visa não apenas proporcionar informações de maneira

visualmente atrativa, mas também criar uma interação que promova uma compreensão mais profunda do progresso do projeto.

No contexto da equipe da Data Dream e da equipe docente do Inteli, empregamos uma variedade de canais para garantir uma cobertura completa. Os e-mails são utilizados para comunicações detalhadas e esclarecimento de dúvidas, enquanto as documentações compartilhadas servem como uma fonte centralizada e contínua de referência. As reuniões semanais e os Daily Meets desempenham um papel fundamental ao facilitar a comunicação em tempo real, proporcionando um ambiente para colaboração e resolução de desafios de forma eficiente.

Quanto aos consultores de marketing e ao analista de dados da Integration, adaptamos a estratégia de comunicação para atender às suas necessidades específicas. Utilizamos ferramentas visuais, reuniões formais e e-mails, abrangendo diversas formas de interação. Essa abordagem multifacetada garante que as mensagens chave sejam transmitidas de maneira abrangente e compreensível.

Em relação à frequência da comunicação, cada canal foi ajustado cuidadosamente para atender às expectativas de cada grupo de stakeholders. Reuniões semanais e Status Reports ao final de cada Sprint garantem uma atualização regular, ao passo que a flexibilidade proporcionada pelos e-mails permite respostas rápidas e esclarecimento de dúvidas.

15.5 Plano de implementação

O plano de implementação é claro, viável e reflete uma consideração cuidadosa dos recursos disponíveis e dos possíveis obstáculos. Sua estrutura organizada e abordagem proativa para antecipar desafios destacam sua importância para o sucesso da execução do projeto.

As reuniões regulares desempenham um papel crucial ao fornecer atualizações durante as sprints, revisando entregáveis e criando um fórum eficiente para manter todos os stakeholders informados sobre o progresso. A utilização de uma plataforma centralizada para armazenar informações facilita o acesso e promove a colaboração ao consolidar dados cruciais em um local acessível a todos.

A consideração cuidadosa dos recursos disponíveis é evidente no plano, que inclui reuniões formais agendadas em pontos específicos para otimizar o uso eficiente do tempo

dos stakeholders. Além disso, a programação de revisões formais em momentos específicos demonstra uma abordagem proativa para esclarecer dúvidas e revisar conteúdos, maximizando a utilização dos recursos. A análise dos possíveis obstáculos é uma característica distintiva, com reuniões periódicas dedicadas a discutir obstáculos e soluções, refletindo uma abordagem preventiva para mitigar possíveis complicações.

15.6 Medidas de sucesso

As medidas de sucesso delineadas para o plano de comunicação são fundamentais para avaliar sua eficácia e impacto ao longo do projeto. O primeiro indicador aborda a compreensão, buscando alcançar uma alta taxa de entendimento entre os envolvidos. A mensuração da compreensão proporciona insights valiosos sobre a eficácia das mensagens chave transmitidas.

Outro ponto-chave é o cumprimento dos prazos em conjunto com o grupo. Esta métrica é não apenas mensurável, mas também tangível, fornecendo uma base sólida para avaliar se a comunicação está resultando em ações coordenadas e eficientes, no qual o monitoramento dos feedbacks torna-se uma prática crucial. A relevância e mensurabilidade desses indicadores são evidentes, pois cada um fornece uma visão específica do progresso e eficácia do plano de comunicação. Eles não apenas refletem diretamente os objetivos definidos para a comunicação, mas também oferecem uma estrutura clara para avaliar o impacto das mensagens chave nos comportamentos e ações dos stakeholders.

15.7 Feedback e Ajustes

A abordagem de feedback e ajustes no plano de comunicação é vital para garantir sua eficácia contínua. Com métodos como conversas diretas, reuniões de Sprint, pesquisas de satisfação e análise de KPIs, o plano demonstra um compromisso sólido em capturar percepções dos stakeholders. Esta abordagem pró-ativa não apenas reconhece a importância do feedback, mas também evidencia a disposição de ajustar as estratégias de comunicação conforme necessário. A viabilidade do plano é notável, destacando-se pela inclusão de reuniões formais agendadas para esclarecer dúvidas e revisar conteúdos. Este compromisso com a revisão contínua reflete a adaptabilidade do plano. Ao enfrentar possíveis desafios, a análise de KPIs e métricas de sucesso oferece dados

tangíveis para identificar áreas que podem exigir ajustes. Assim, o plano não só é viável, mas também se orienta proativamente para lidar com desafios potenciais.

16. Análise de Impacto Ético

16.1 Introdução

Nos últimos anos, a revolução digital transformou radicalmente a maneira como o mundo coleta, armazena, processa e utiliza informações. Um dos catalisadores mais significativos dessa transformação é o fenômeno conhecido como Big Data, que se refere ao crescente volume, variedade e velocidade com que os dados são gerados e analisados em nossa sociedade moderna. Este trouxe consigo uma série de oportunidades e desafios, moldando profundamente tanto a sociedade humana quanto o meio ambiente em que vivemos. Além disso, a utilização de Big Data também tem implicações diretas no meio ambiente, pois o processamento de dados requer vastas quantidades de energia e recursos naturais. A seguir, serão descritos os possíveis impactos de Big Data na sociedade e no meio ambiente, enfocando cinco dimensões críticas.

16.2 Segurança e Proteção de Dados

O Big Data revolucionou a forma como as informações são coletadas e usadas, muitas vezes envolvendo uma grande quantidade de dados pessoais. Com isso, empresas, governos e organizações utilizam tecnologias avançadas de coleta de dados, como sensores, mídias sociais, dispositivos IoT (Internet das Coisas) e câmeras de vigilância para capturar informações.

16.2.1 Dados Pessoais

A coleta de dados pessoais envolve informações como nome, endereço, número de telefone, histórico de compras, preferências de pesquisa na web e até mesmo dados biométricos, como impressões digitais e reconhecimento facial.

16.2.2 Conformidade com Regulamentações

À medida que a coleta de dados pessoais se tornou mais recorrente, os legisladores em várias jurisdições responderam com regulamentações rigorosas destinadas a proteger

a privacidade e os direitos dos indivíduos. O exemplo mais notável é o Regulamento Geral de Proteção de Dados (GDPR) na União Europeia, que estabelece diretrizes estritas para a coleta e processamento de dados pessoais. No Brasil, por exemplo, existe a LGPD (Lei Geral de Proteção de Dados) que regulamenta a proteção de dados pessoais no país, muito influenciada pela GDPR.

16.2.3 Minimização de Dados

Uma abordagem ética na coleta de dados pessoais envolve a minimização de dados, onde apenas as informações estritamente necessárias devem ser coletadas e usadas para a finalidade específica pré comunicada. A coleta excessiva de dados, que não está diretamente relacionada à finalidade declarada, é vista como uma invasão da privacidade e pode ser prejudicial para os indivíduos.

16.2.4 Consentimento Informado

Um princípio fundamental na coleta de dados pessoais é o consentimento informado. Os indivíduos devem ser devidamente informados sobre quais dados estão sendo coletados, como serão usados e com quem serão compartilhados, além de ter opção de consentir ou não com a coleta de seus dados. Isso promove a transparência e dá às pessoas o controle sobre suas informações pessoais.

16.2.5 Segurança de Dados

As organizações devem implementar medidas de segurança robustas para proteger os dados contra acessos não autorizados e violações, a perda ou vazamento de dados pessoais pode ter sérias consequências para a privacidade das pessoas e a confiança nas instituições que coletam esses dados.

16.2.6 Dados Não Pessoais

Dados não pessoais incluem informações que não podem ser usadas para identificar diretamente um indivíduo específico, podendo ser agregados, anonimizados ou simplesmente não conter informações pessoais identificáveis.

16.2.7 Transparência e Responsabilidade

Apesar de não serem dados pessoais, é fundamental que as organizações sejam transparentes e responsáveis em relação à coleta e ao uso desses dados. Por isso, a empresa deve informar como os dados estão sendo coletados, processados e utilizados. Já a responsabilidade envolve que as práticas de coleta e uso de dados não prejudiquem os interesses de terceiros, como por exemplo processo de tomada de decisão.

16.2.8 Benefícios Sociais e Econômicos

Dados não pessoais desempenham um papel crucial em impulsionar benefícios sociais e econômicos. Ao analisar padrões de comportamento em larga escala, as organizações podem identificar oportunidades de melhoria, otimizar processos e contribuir para o desenvolvimento sustentável, sem comprometer a privacidade individual. Com isso, a sociedade se beneficia e esta não precisa se preocupar com a privacidade.

16.2.9 Desafios Éticos

Apesar das vantagens, a coleta e o uso de dados não pessoais também apresentam desafios éticos. A possibilidade de inferir informações pessoais a partir de dados aparentemente não pessoais (re-identificação) é uma preocupação. Portanto, é necessário que organizações adotem práticas éticas, mesmo que os dados aparentem não serem pessoais, como dito no tópico 2.1.1.

16.3 Equidade e justiça

Existem várias formas do atual projeto ter um impacto social em grupos específicos, priorizando alguns em detrimento de outros. Assim, é necessário visualizar formas de minimizar essas disparidades.

16.3.1 Problemas que o Projeto Pode Gerar para Questões de Equidade e Justiça:

Conforme os estudos desenvolvidos na Sprint 1 (entendimento do usuário e negócio), o atual projeto se propõe a desenvolver uma ferramenta que identifica cidades do Brasil com maior potencial de consumo, visualizando categorias de produtos e os

canais de atendimento disponíveis. Identificando esses mercados potencialmente lucrativos, a princípio (pela regra de oferta e demanda simples), as empresas que consultaram a Integration tenderão a fazer investimentos exclusivamente nessas regiões com maior potencial de consumo e compra imediato. Essa visão de investimento favorece regiões mais ricas e desfavorece ainda mais regiões mais pobres, com baixo potencial consumo. Um relevante autor que comenta esse tipo de cenário é o economista regional Albert O. Hirschman, que explorou o fenômeno do desenvolvimento desigual. Ele argumenta que as disparidades econômicas entre regiões podem resultar em um ciclo vicioso, onde as regiões mais ricas continuam a atrair mais recursos e investimentos, enquanto as mais pobres enfrentam dificuldades crescentes para alcançar um desenvolvimento equitativo.

A curto e médio prazo pode parecer racionalmente econômico investir exclusivamente apenas nos mercados aquecidos, contudo, a chances desse mercado saturar são muito grandes, sendo um mercado grande e diversificado muito mais vantajoso e sustentável a longo prazo. Além disso, alguns mercados de nicho lucrativos podem ser desconsiderados nessa lógica utilitarista imediata, é necessário explorar e assumir riscos para não perder esse tipo de mercado. Ultrapassando o critério econômico, também não é justo esse cenário tendo em vista a igualdade de direitos e potenciais de equidade.

16.3.2 Soluções e Propostas para Problemas de Equidade e Justiça:

A partir desse cenário, algumas ações são possíveis para minimizar os impactos citados. O primeiro é a possibilidade de redirecionar o uso da ferramenta criada: o potencial de identificar os mercados mais fortes também é o mesmo para identificar os mais fracos, possibilitando uma atuação especial nas regiões menos favorecidas. Esse novo tipo de uso pode ser aplicado tanto por empresas privadas quanto estatais, afinal, todas as soluções desenvolvidas no Inteli são “open source”, qualquer um pode utilizar e aplicar.

Outra forma de minimizar essas disparidades é incluir na solução uma “carta de intenção”, estimulando que as organizações que utilizaram a presente solução também invistam, diretamente ou indiretamente, nas regiões mais pobres.

Em último caso, as regiões menos favorecidas serão beneficiadas por essa solução indiretamente: a Integration, nossa parceira de projeto, atualmente apoia o

Instituto Arca+, uma startup de tecnologia que foca em pessoas de baixa renda em situação vulnerável, às conectando com doadores, fornecedores de serviços, produtos e programas sociais.



Figura 230: Logo Arca +

Fonte: Autoria Própria

Mais sobre a Arca+: https://integrationconsulting.com/en/testimonial_trusted/arca/

16.4 Transparência e Consentimento Informado em Projetos de Big Data

A era do Big Data trouxe mudanças significativas na forma como dados são coletados, processados e utilizados. Com isso, a transparência e o consentimento informado emergem como pilares fundamentais para garantir o uso ético e responsável desses dados, especialmente quando se tratam de informações sensíveis ou pessoais.

16.4.1 Transparência na Coleta e Uso de Dados

A transparência em projetos de Big Data significa comunicar abertamente as práticas de coleta e uso de dados. Isso inclui fornecer informações detalhadas sobre que tipos de dados são coletados (sejam eles dados demográficos, comportamentais, transacionais, etc.), as técnicas utilizadas para coleta (como análise de redes sociais, rastreamento por GPS, sensores IoT), como esses dados são processados e analisados, e as finalidades para as quais são utilizados. Também é importante esclarecer como os dados são protegidos, a duração do armazenamento e as políticas de descarte de dados após o uso.

16.4.2 Consentimento Informado

O consentimento informado é um processo dinâmico e contínuo em Big Data. Deve-se garantir que os usuários estejam cientes não apenas do tipo de dados coletados, mas também do contexto e do escopo dessa coleta, as metodologias de análise utilizadas, e especialmente, como esses dados podem influenciar decisões que os afetam diretamente ou indiretamente. Isso pode incluir implicações em termos de privacidade, potenciais riscos de segurança e o direito de retirar o consentimento a qualquer momento.

16.4.3 Privacidade e Segurança dos Dados

A segurança e privacidade dos dados em projetos de Big Data são desafios complexos e multifacetados. Eles exigem a implementação de medidas avançadas de segurança cibernética, como criptografia forte, autenticação multifatorial, e sistemas de detecção e prevenção de intrusões. Além disso, é crucial considerar aspectos de privacidade ao projetar sistemas de coleta e análise de dados, adotando práticas como anonimização de dados, privacidade diferencial e avaliações regulares de impacto à privacidade.

16.4.4 Conformidade com Regulamentações

Manter a conformidade com regulamentações internacionais e locais é essencial. Isso inclui não apenas o entendimento e a implementação de leis como o GDPR na União Europeia e a LGPD no Brasil, mas também uma compreensão aprofundada de como essas leis se aplicam a diferentes tipos de dados e situações. Além disso, deve-se estar atento às atualizações e mudanças na legislação, garantindo que as práticas de coleta e uso de dados permaneçam em conformidade ao longo do tempo.

16.4.5 Minimização de Dados

A minimização de dados é uma estratégia que vai além de evitar a coleta excessiva de informações. Envolve uma avaliação crítica da necessidade real de cada tipo de dado coletado, questionando se cada elemento é essencial para a finalidade pretendida. Isso não só ajuda a proteger a privacidade dos usuários, mas também contribui para a

eficiência do processamento e análise de dados, reduzindo o ruído e aumentando a relevância das informações coletadas.

16.4.6 Governança de Dados

Uma governança de dados eficiente é um aspecto chave na gestão de projetos de Big Data. Isso envolve desenvolver um quadro de políticas e procedimentos que regulem todos os aspectos da gestão de dados, desde a coleta até a análise, uso e compartilhamento. Deve incluir práticas para assegurar a qualidade e integridade dos dados, mecanismos para lidar com erros e imprecisões, e canais para que os usuários possam reportar preocupações ou abusos.

16.4.7 Participação e Empoderamento do Usuário

Empoderar os usuários no contexto de Big Data significa assegurar que eles tenham compreensão clara e controle sobre como seus dados são utilizados. Isso inclui facilitar o acesso às suas próprias informações, fornecer opções claras e acessíveis para gerenciar consentimentos e preferências, e assegurar direitos como a portabilidade e exclusão de dados.

16.4.8 Dados Não Pessoais

Os dados não pessoais, embora não identifiquem indivíduos diretamente, podem ainda assim ter impactos significativos quando coletados e analisados em larga escala. É essencial abordar as implicações éticas e sociais da utilização desses dados, como o potencial de influenciar padrões de mercado, políticas públicas e comportamentos sociais, mantendo a transparência e a responsabilidade ética no seu uso.

16.4.9 Desafios Éticos com Dados Não Pessoais

Os desafios éticos associados à coleta e uso de dados não pessoais em Big Data são vastos e complexos. Um dos maiores riscos é a re-identificação, onde dados anônimos podem ser cruzados com outras fontes para revelar informações pessoais. Outras questões éticas incluem a potencial discriminação ou viés algorítmico que pode emergir da análise desses dados. As organizações devem adotar uma abordagem

proativa para identificar e mitigar esses riscos, garantindo que os benefícios do Big Data sejam alcançados sem comprometer a ética e os direitos dos indivíduos.

16.5 Responsabilidade Social

O impacto social pode ser definido como as consequências observáveis e mensuráveis das ações de indivíduos, organizações ou governos sobre a sociedade. Isso inclui mudanças positivas ou negativas em diversas áreas, como educação, saúde, economia, meio ambiente, justiça social, entre outras.

Diante disso, o projeto em questão possui implicações significativas em termos de responsabilidade social, considerando tanto os impactos positivos quanto os potenciais efeitos adversos sobre diferentes comunidades e o meio ambiente. Os parágrafos que se seguem abordam aspectos relevantes para a avaliação do impacto social do projeto, além de apresentar os possíveis alinhamentos com diferentes objetivos de desenvolvimento sustentável e as oportunidades para mitigar os efeitos negativos.

16.5.1 Impacto Social Positivo

No que tange aos impactos positivos, a implementação do pipeline de Big Data se apresenta como uma solução capaz de potencializar o surgimento de empregos diretos e indiretos. Ao implementar uma solução baseada em um grande volume de dados, a Integration poderá aumentar a lucratividade e, consequentemente, o tamanho de seus clientes. Com a crescente demanda de mão de obra para comportar o aumento de seus negócios, os clientes da consultoria terão que contratar mais pessoas, gerando mais empregos nos locais em que atuam. Logo, essa implementação impulsionará o desenvolvimento econômico local, impactando positivamente o meio em que estará inserida.

Além disso, a análise detalhada dos padrões de consumo das diferentes camadas da sociedade poderá resultar em operações logísticas mais eficientes e no desenvolvimento de práticas sustentáveis. Isso porque é possível que a consultoria e seus clientes desenvolvam estratégias capazes de diminuir o impacto ambiental gerado pela cadeia de

distribuição dos produtos, a partir da implementação de rotas mais eficientes, da adoção de embalagens eco-friendly e da promoção de práticas de transporte sustentável.

Por fim, a análise detalhada dos dados permite entender de maneira mais profunda as preferências dos consumidores, possibilitando o desenvolvimento de produtos e serviços mais alinhados com as necessidades específicas de diferentes clientes. Diante disso, é possível que a Integration incentive seus parceiros a customizarem e inovarem continuamente em resposta às demandas locais mapeadas pela solução, permitindo também que eles adentrem em novos mercados inexplorados.

16.5.2 Impacto Social Negativo

É possível que, através da implementação extensiva de recursos tecnológicos, os clientes da Integration se destaquem significativamente em relação aos seus concorrentes locais. Embora esse diferencial possa ser percebido positivamente, é provável que os competidores do mercado não disponham da mesma capacidade de adotar medidas tão avançadas quanto aquelas adotadas pelos clientes da consultoria. Essa disparidade potencial pode resultar na dominância do mercado por parte dos clientes da Integration, comprometendo a dinâmica do livre mercado e limitando as opções disponíveis para os usuários finais.

Além disso, a assimetria na adoção de tecnologias avançadas pelos clientes da Integration pode criar barreiras de entrada significativas para novos competidores, restringindo a competição no setor. A possibilidade de um domínio exacerbado por parte dos clientes da consultoria levanta preocupações quanto à diversidade e inovação no mercado, uma vez que competidores locais podem se encontrar em desvantagem para acompanhar o ritmo das tecnologias emergentes.

Para mitigar esses impactos, estratégias inclusivas e parcerias colaborativas podem ser consideradas, visando promover a equalização de oportunidades tecnológicas e preservar um ambiente de mercado mais dinâmico e competitivo. Essa abordagem não apenas fortaleceria a integridade do livre mercado, mas também beneficiaria os consumidores ao garantir uma gama mais ampla de escolhas e inovações.

16.5.3 Correlação com os Objetivos de Desenvolvimento Sustentável

O projeto em análise apresenta conexão com diferentes Objetivos de Desenvolvimento Sustentável (ODS), destacando seu potencial para impulsionar progressos significativos em direção a metas globais de sustentabilidade.

16.5.4 Alinhamento com os ODS

ODS 8: Trabalho Decente e Crescimento Econômico

A implementação do pipeline de Big Data abre oportunidades concretas para o alcance do ODS 8 ao estimular o surgimento de empregos diretos e indiretos. Ao impulsionar o crescimento econômico, esse projeto contribui para a promoção do trabalho decente, favorecendo a estabilidade e o bem-estar dos trabalhadores impactados.

ODS 9: Indústria, Inovação e Infraestrutura

A implementação de tecnologias avançadas, como aquelas utilizadas no projeto de Big Data, apoia diretamente o ODS 9 ao promover inovação e infraestrutura sustentável. Ao otimizar operações logísticas e impulsionar a customização de projetos, o projeto contribui para o avanço tecnológico e a eficiência na utilização de diferentes recursos.

16.5.5 Afastamento dos ODS

ODS 10: Redução das Desigualdades

O destaque tecnológico dos clientes da Integration em relação aos concorrentes locais pode criar disparidades significativas, possivelmente afastando-se do ODS 10. Essa assimetria na adoção de tecnologias avançadas pode contribuir para desigualdades econômicas entre empresas do setor, demandando atenção específica para garantir que os benefícios sejam distribuídos de maneira equitativa.

16.6 Viés e discriminação

O viés de dados refere-se à presença de distorções sistemáticas ou desigualdades em conjuntos de dados que podem influenciar resultados analíticos. Essas distorções podem surgir de diversas fontes e manifestar-se de várias maneiras. Por outro lado, a discriminação de dados ocorre quando as decisões tomadas com base em análises de dados resultam em tratamento desigual ou injusto para diferentes grupos.

16.6.1 Formas de Viés de Dados

Viés de Amostragem: Resulta de uma amostra não representativa do conjunto populacional, levando a conclusões não generalizáveis.

Viés de Seleção: Ocorre quando certos grupos são sub-representados ou excluídos, influenciando a validade das conclusões.

Viés Temporal: Reflete mudanças nas condições ao longo do tempo, afetando a relevância de dados históricos.

16.6.2 Características do Viés

Involuntário: Pode ocorrer sem intenção, muitas vezes sendo uma consequência não prevista da coleta ou processamento de dados.

Sistemático: Afeta consistentemente certos grupos ou características, introduzindo padrões previsíveis nas análises.

Contextual: Depende do contexto específico do conjunto de dados e do problema em questão.

Social: Reflete preconceitos existentes na sociedade, como estereótipos culturais e históricos.

Discriminação Ética: Envolve dilemas éticos associados ao tratamento justo e equitativo de grupos vulneráveis.

16.6.3 Formas de Discriminação de Dados

Discriminação Direta: Resulta em tratamento explícitamente diferenciado para grupos específicos.

Discriminação Indireta: Ocorre quando certos critérios ou características, aparentemente neutros, têm impactos negativos em grupos específicos.

16.6.4 Características da Discriminação

Explícita: Quando as decisões são baseadas explicitamente em características sensíveis, como raça, gênero ou origem.

Implícita: Pode ocorrer devido a vieses não intencionais incorporados em algoritmos ou processos analíticos.

16.6.5 Medidas de Mitigação:

Diversidade de Dados: Garantir que o conjunto de dados utilizado seja representativo de todas as categorias, regiões e canais, evitando viéses específicos.

Avaliação Contínua: Implementar mecanismos de monitoramento constante para identificar e corrigir viéses ao longo do tempo. Isso pode envolver a revisão regular dos resultados do pipeline para detectar disparidades e ajustar os algoritmos conforme necessário.

Transparência e Explicabilidade: Tornar o processo de tomada de decisão do algoritmo transparente, permitindo que os usuários compreendam como as conclusões são alcançadas. Isso facilita a identificação de possíveis viéses.

16.6.6 Abordagem no Desenvolvimento do Pipeline:

Ao projetar e implementar o pipeline de Big Data, é vital incorporar práticas que minimizem riscos de discriminação e exclusão. Isso inclui:

Equidade no Processamento: Garantir que cada categoria, região e canal seja tratado com imparcialidade durante as fases de coleta, processamento e análise de dados.

Testes: Realizar testes extensivos para identificar possíveis pontos de discriminação. Isso pode envolver a criação de casos de teste específicos para cenários diversos.

Envolvimento de Stakeholders Diversos: Incluir representantes de diferentes grupos nos processos de desenvolvimento e validação para garantir perspectivas diversas.

16.6.7 Exemplo de Viés no Projeto

Viés Coleta de Dados de Canais de Atendimento

Suponha que, ao coletar dados sobre o desempenho das categorias nos canais de atendimento, haja uma sobre-representação de informações de supermercados em comparação com outros canais, como food service. Se a ferramenta der um peso desproporcional aos dados de supermercados, pode resultar em uma análise enviesada, subestimando o potencial de consumo em canais menos representados. Isso pode levar a uma alocação inadequada de recursos e estratégias, impactando negativamente a eficácia das ações direcionadas a diferentes canais.

Discriminação na Representação Geográfica:

Suponha que, durante a criação do infográfico para visualização dos resultados estatísticos, a representação geográfica seja mais detalhada para áreas urbanas em comparação com áreas rurais. Se a visualização não refletir de maneira equitativa a distribuição geográfica dos dados, pode criar uma percepção distorcida do potencial de consumo em diferentes regiões. Isso poderia levar a decisões discriminatórias na alocação de recursos, com ênfase excessiva em áreas urbanas e negligência das oportunidades em áreas rurais.

Medidas de Mitigação:

Para evitar esses problemas, é importante implementar medidas no pipeline de Big Data:

Garantir Representatividade: Certificar-se de que a coleta de dados abrange todos os canais de atendimento e regiões geográficas de maneira proporcional, evitando distorções na análise.

Equidade na Visualização: Ao criar visualizações, garantir que a representação geográfica e de categorias seja equitativa, destacando de forma igual o potencial de consumo em diferentes regiões e canais.

Revisão Regular: Realizar revisões periódicas do pipeline para identificar possíveis desvios e ajustar o processo conforme necessário para manter a imparcialidade.

16.7 Conclusão

A conclusão deste estudo sobre o impacto do Big Data na sociedade e no meio ambiente destaca a dualidade de seu papel. Por um lado, o Big Data oferece

oportunidades extraordinárias para avanços em diversos setores, como saúde, educação, e comércio. A capacidade de analisar grandes volumes de dados pode levar a insights mais profundos, inovações significativas e uma maior eficiência operacional. Por outro lado, os desafios éticos, de privacidade, e de segurança associados ao Big Data são inegáveis e exigem atenção contínua.

O Big Data carrega desafios significativos em termos de equidade e justiça social, com potencial para perpetuar ou intensificar desigualdades. É indispensável que empresas e instituições que utilizam o Big Data adotem uma responsabilidade social ativa, focando não só no cumprimento das normas de proteção de dados, mas também na promoção de inclusão, justiça e práticas sustentáveis. Além disso, é crucial manter a transparência e garantir o consentimento informado, assegurando que os indivíduos compreendam como seus dados são usados e protegidos contra violações de privacidade. A segurança dos dados, tanto pessoais quanto não pessoais, deve ser uma prioridade constante para prevenir abusos e garantir a confiança na utilização do Big Data.

Por fim, o Big Data deve ser visto como uma ferramenta poderosa, mas que carrega consigo uma grande responsabilidade. Seus benefícios podem ser vastos, mas apenas se forem gerenciados de maneira ética e responsável. As organizações que utilizam Big Data devem estar na vanguarda da criação de um futuro onde a tecnologia serve a humanidade de maneira justa e equitativa, ao mesmo tempo em que respeita a privacidade e promove o desenvolvimento sustentável. A adoção de uma abordagem holística e ética em relação ao Big Data é essencial para assegurar que seus benefícios sejam amplamente compartilhados e que seus riscos sejam minimizados, contribuindo assim para um futuro mais justo e sustentável para todos.

17. Referências

COMMUNITY REVELO. Arquitetura de Big Data: o que é? [S.I.], 2021. Disponível em:
<https://community.revelo.com.br/arquitetura-de-big-data-o-que-e>
Acesso em: 27 out. 2023.

MICROSOFT. Big Data Architecture Guide. [S.I.], 2023. Disponível em:
<https://learn.microsoft.com/pt-br/azure/architecture/data-guide/big-data>
Acesso em: 27 out. 2023.

MICROSOFT. Big Data Architecture Styles. [S.I.], 2023. Disponível em:
<https://learn.microsoft.com/pt-br/azure/architecture/guide/architecture-styles/big-data>
Acesso em: 27 out. 2023.

LOPES, Fábio Augusto de Carvalho. Arquitetura Big Data: escolha a canalização correta para sua empresa! [S.I.], 2019. Disponível em:
<https://www.linkedin.com/pulse/arquitetura-big-data-escolha-canalizacao-correta-para-lopes/?origin.alSubdomain=pt>
Acesso em: 27 out. 2023.

AWARI EDUCATION. Arquitetura de Big Data: modelos e implementações [S.I.], 2019. Disponível em: <https://awari.com.br/arquitetura-de-big-data-modelos-e-implementacoes-11>
Acesso em: 27 out. 2023.

AMAZON WEB SERVICES. Arquitetura de dados moderna na AWS. [S.I.], 2023. Disponível em: <https://aws.amazon.com/pt/big-data/datalakes-and-analytics/modern-data-architecture> Acesso em: 27 out. 2023.

AMAZON WEB SERVICES. Arquitetura referência de análise de dados sem servidor na AWS. [S.I.], 2023. Disponível em:
<https://aws.amazon.com/pt/blogs/aws-brasil/arquitetura-referencia-de-analise-de-dados-sem-servidor-na-aws>
Acesso em: 27 out. 2023.