



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: 24 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

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. Compreensão do Problema	8
3.1 Proposta de Valor	8
3.2 Matriz de Risco	8
3.3 TAM SAM SOM	16
4. Lean Inception	18
4.1 O Produto (É – Não É – Faz – Não Faz)	18
4.2 Funcionalidades	19
4.3 Modelo de dados	19
5. Análise de Experiência do Usuário	19
5.1 Personas	20
5.2 Jornada do Usuário	21
5.3 User Stories	24
5.3.1 US00 - Configuração do Ambiente AWS	24
5.3.2 US01 - Ingestão de Dados	25
5.3.3 US02 - Análise Estatística Inicial	27
5.3.4 US03 - Load de Dados	28
5.3.5 US04 - Configuração da estrutura dos dados	30
5.3.6 US05 - Análise de Consumo	31
5.3.7 US06 - Filtros para Visualização da Distribuição de Consumo	33
6. Identificação dos tipos de dados e suas características.	35
6.1. Dados CSV	35
6.1.1 Tipos de Dados e suas Características:	35
6.1.2 Importância para o Projeto:	37
6.2 Dados CNPJ	37
6.2.1 Tipos de Dados e suas Características:	37
6.2.2 Importância para o Projeto:	40
6.3 API	40
7. Arquitetura Macro	42
7.1. Requisitos do pipeline de dados	42
7.2. Identificação dos dados de entrada e saída	43
7.2.1 Dados de Entrada	43
7.2.2 Dados de Saída	43

7.3. Análise das necessidades e objetivos do pipeline	44
7.3.1 Necessidades	44
7.3.2 Objetivos	44
7.4. Escolha de serviços adequados para cada etapa do pipeline	45
7.4.1 Fonte de Dados	45
7.4.2 Automação de Ingestão	45
7.4.3 Preparação e Armazenamento:	45
7.4.4 Análise e Infográfico	46
7.4.5 Segurança	46
7.5. Justificativa para a escolha dos serviços	47
7.5.1 Justificativas específicas	47
7.6. Representação visual do pipeline	48
7.7. Consideração de boas práticas para garantir resiliência e escalabilidade	50
7.8. Uso de serviços ou recursos da AWS que suportem resiliência e escalabilidade	50
7.9. Calculadora financeira	51
7.10. Arquitetura e a Integration	51
8. Mockup Interface (Preliminar)	52
8.1 Tela de Login (Autenticação de Consultor de Marketing e Vendas)	52
8.2 Tela de Dashboard	52
8.3 Tela de Dados das Fontes (Governo, Parceiro e CNPJ)	53
9. Análise Exploratória	54
9.1. CNPJs	54
9.1.1 CNPJ 1	54
9.1.2 CNPJ 2	57
9.1.3 CNPJ 3	60
9.1.4 CNPJ 4	63
9.1.5 CNPJ 5	66
9.2. Dados do Governo	69
9.2.1 Aluguel Estimado	69
9.2.2 Domicílio	73
9.2.3 Inventário	79
10. Referências	83

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 datalake ou 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 de renome 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 & Pagamentos, Industrial, Agronegócio e Farmacêutico /Assistência médica.

1.2 Definição do Problema

O problema essencial consiste na falta de uma ferramenta que permita à Integration 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

uma base de dados analisável e uma representação visual dos dados torna o parceiro incapaz de oferecer análises estratégicas com eficiência.

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 datalake ou data warehouse. Além disso, busca-se a criação de um infográfico que represente 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 (cidade 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. Compreensão do Problema

Apresenta-se nessa sessão 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.

3.1 Proposta de Valor

A seguir, apresentamos o Canvas Proposta de Valor, elaborado para a empresa Integration, parceira neste módulo do projeto. Através do Canvas Proposta de Valor, buscamos entender e mapear de maneira estruturada as soluções oferecidas pela Integration, especialmente no que tange ao mercado consumidor de alimentos brasileiro.

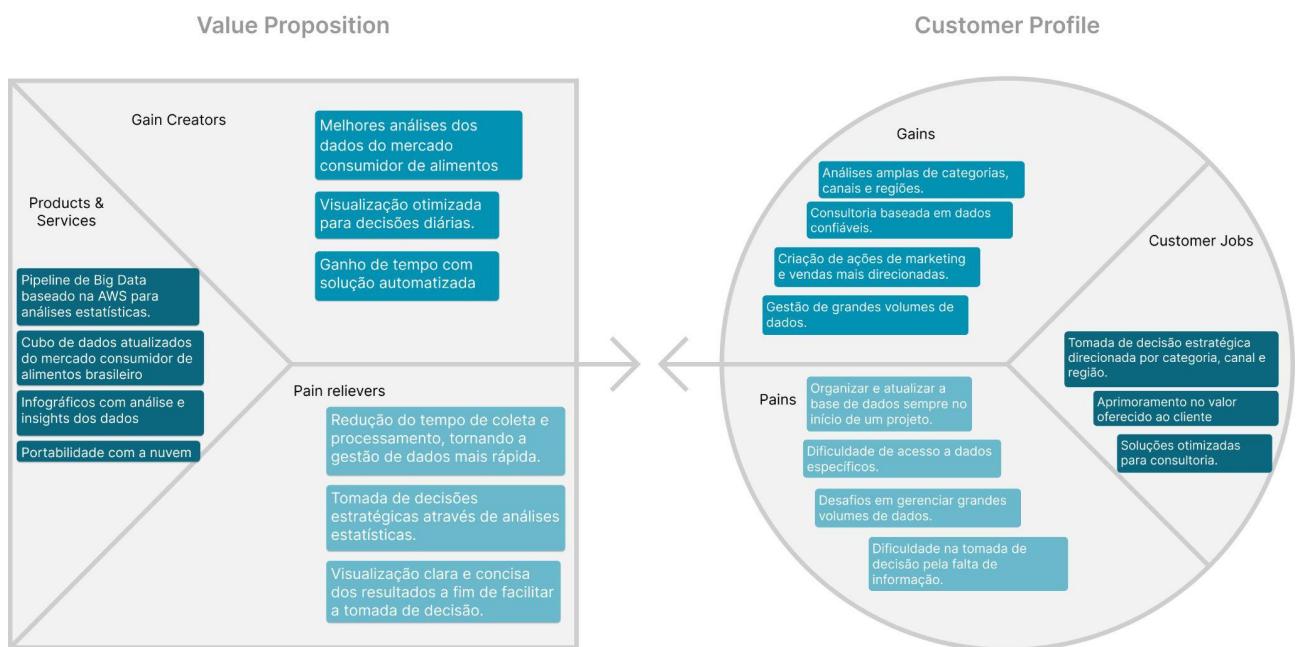


Imagen 01: Canvas Proposta de Valor

Fonte: Criação própria.

3.2 Matriz de Risco

É uma das principais ferramentas na análise de negócios, utilizada para o gerenciamento de riscos operacionais existentes na empresa. A Figura 3, ilustra a construção da matriz de risco para o projeto.

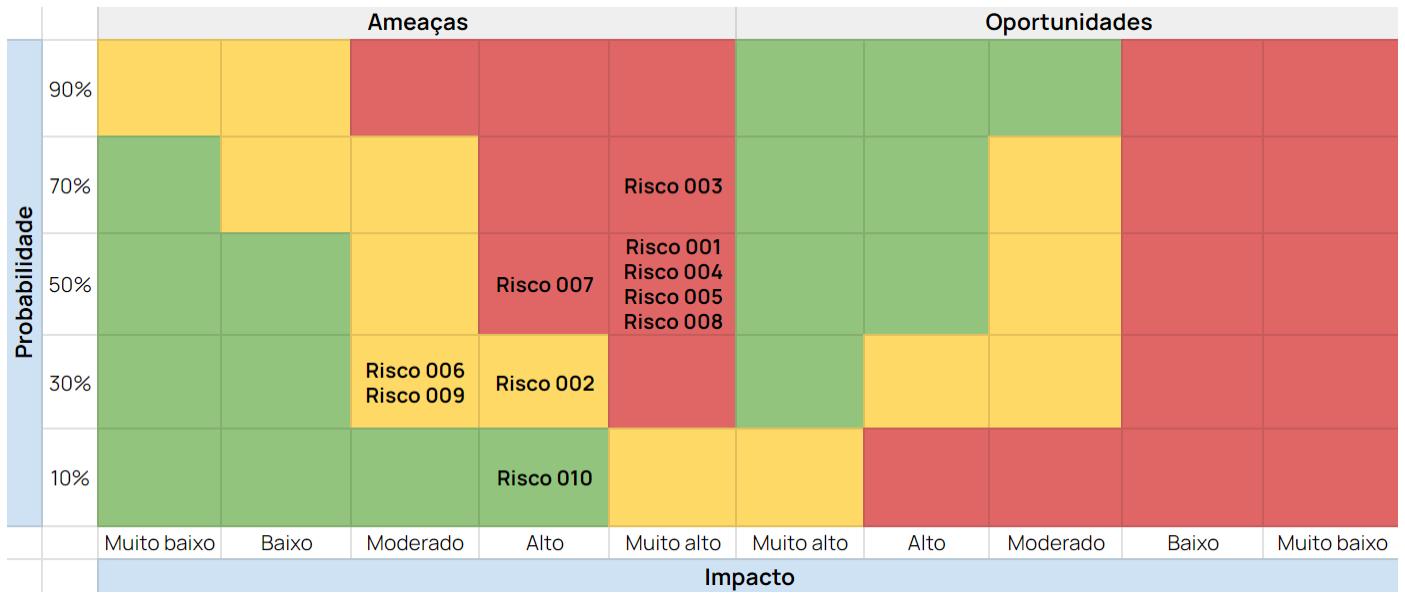


Imagen 02: matriz de risco desenvolvida pelo grupo.

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 ineficiente 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, elaboramos 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

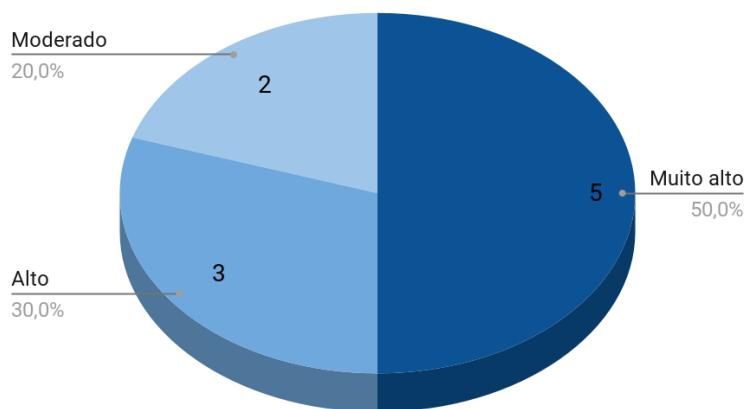
às necessidades e desafios da equipe, a fim de prevenir que esses riscos se transformem em problemas reais no futuro.

Distribuição dos riscos em categorias:



Além disso, procedemos à classificação dos riscos com base em seu impacto no projeto. Nessa análise, constatamos 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:



3.3 TAM SAM SOM

O Total Addressable Market, Service Addressable Market e Service Obtainable Market (TAM, SAM e SOM) são utilizados no contexto de análise de mercado e

estratégias de negócios. Estes, são utilizados para definir e dimensionar quais são os mercados-alvo, auxiliando as empresas a compreender o potencial de seus produtos / serviços. 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 XX.



Figura 05: TAM SAM SOM

Fonte: Elaboração 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 Empresometro. 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.

4. 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.

4.1 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;
 - Saber 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;

4.2 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 (CNAE) dos dados em palavras factíveis;
- Flexível o input do dado do cliente, conseguir customizar as colunas (Utilizar em vários setores);
- Conseguir identificar no infográficos os seguintes dados:
 - Número de CNPJ;
 - Número clientes atendidos;
 - Vendas realizadas;
 - Potencial de consumo;
- Mapeamento de público (Canal, categoria e região):
 - Brasil : Estados, Cidades e Bairro;
 - Onde o consumidor compra? (Canais)

4.3 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

5. Análise de Experiência do Usuário

Nesta sessão, 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. No caso deste projeto, que visa desenvolver uma consultoria de marketing e vendas baseada em Big Data, duas personas distintas foram criadas para melhor ilustrar os usuários que utilizarão nossa solução: 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, a empresa, se encontra. Essas personas não apenas ajudam a equipe a visualizar os usuários finais, mas também a definir estratégias e recursos que atendam às necessidades e expectativas de cada um desses perfis.



AMANDA BRAZ
CONSULTORA
MARKETING E VENDAS

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.

KPI

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

INTERESSES

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

CENÁRIO DE INTERAÇÃO

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

LETRAMENTO DIGITAL

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

BACKGROUND

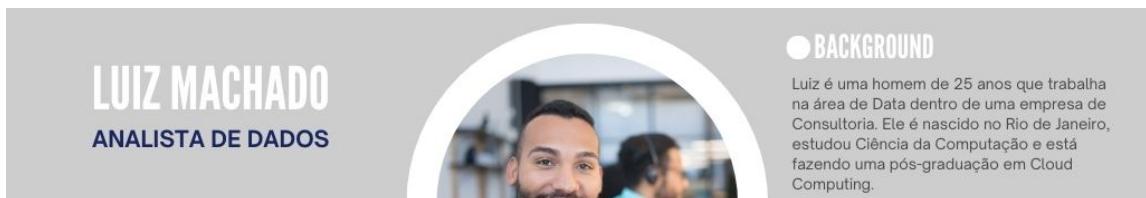
Amanda é uma mulher de 27 anos que trabalha com a área de Marketing cuidando das empresas alimentícias que a consultoria presta serviço. Nascida em São Paulo, formada na ESPM em Publicidade e Propaganda.

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

"Perder informações valiosas é meu maior pesadelo"

Figura 06: Persona - Consultora Marketing e Vendas

Fonte: Elaboração própria



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.

BACKGROUND

Luiz é uma 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.

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 - Analista de Dados

Fonte: Elaboração própria

5.2 Jornada do Usuário

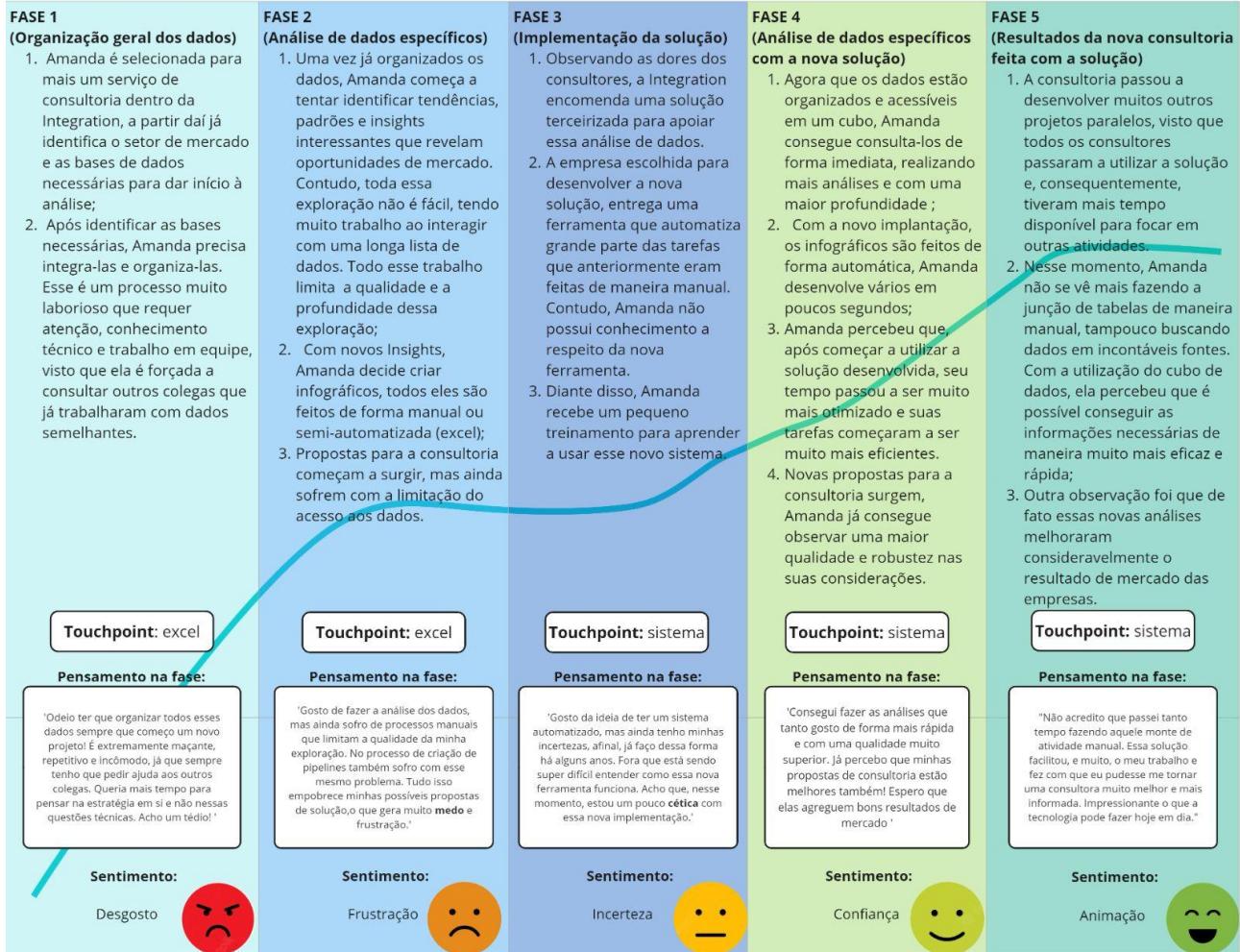
A jornada do usuário construída consiste na representação das etapas principais que envolvem os consultores de marketing e vendas e o analista de dados. Ambos começam com o acesso à plataforma por meio de autenticação. A consultora se envolve com a análise de dados, usando um painel interativo no qual 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. Se exibe nas figuras 8 e 9, sendo elas respectivamente:



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 de Usuário - Consultora de Marketing e Vendas

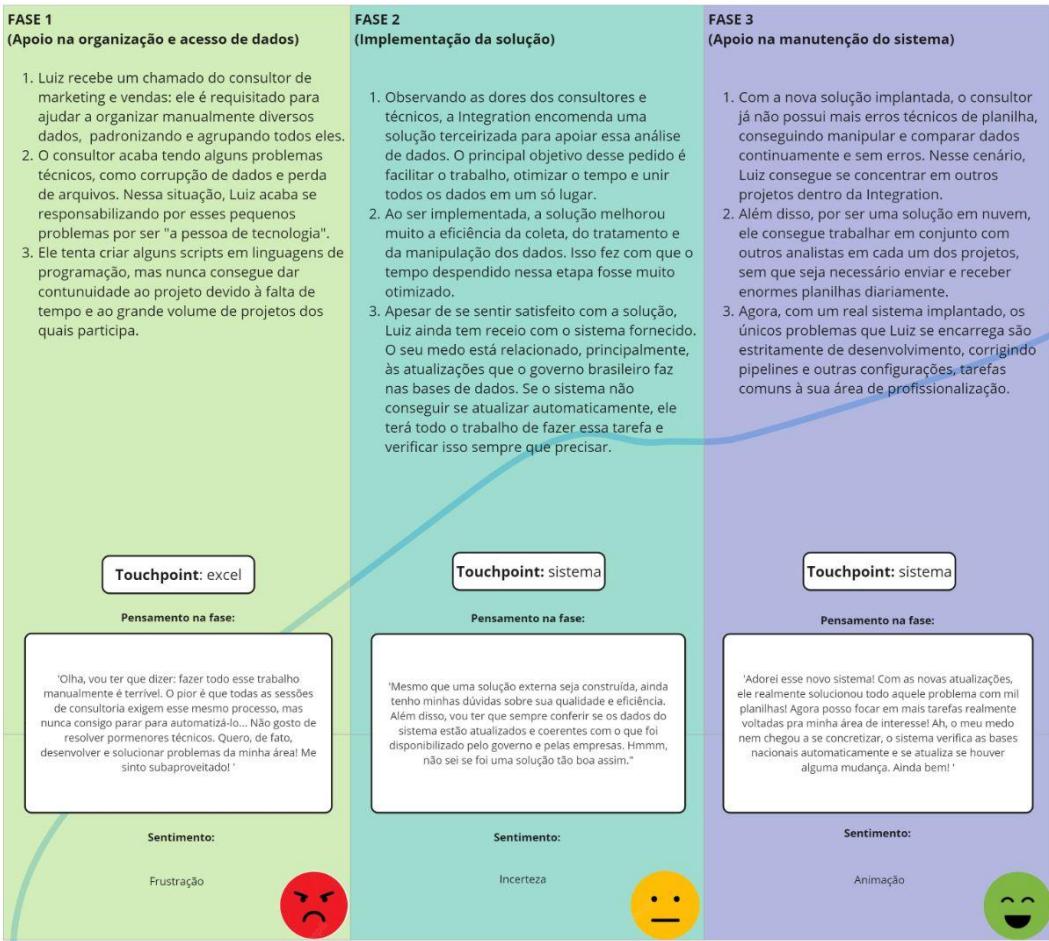
Fonte: Elaboração própria



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 de Usuário - Analista de Dados

Fonte: Elaboração 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 foi 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 rápida 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 - US003

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 macro regiões, apresentada em formato de infográfico.
- Critério de avaliação :

- Critério 1 : A plataforma deve estar acessível para o consultor, com login e acesso seguro.
 - 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: US003 - US004 - US006

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: { Aplataforma 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: US005 - US004

6. 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.

6.1. Dados CSV

O conjunto de dados disponível na pasta “Dados CSV” que pode ser acessada através [deste link](#), apresenta os microdados 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 microdados 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. Apresentaremos a seguir uma análise aprofundada destes microdados, elucidando sua relevância e características.

6.1.1 Tipos de Dados e suas Características:

A. Informações Geográficas e Estratificação Social:

- UF: Identifica a unidade federativa, ajudando a localizar geograficamente os dados.
- ESTRATO_POF: Indica a estratificação social, o que pode ser crucial para entender diferentes comportamentos de consumo e necessidades.

B. Identificação e Situação Residencial:

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

C. Dados Econômicos:

- Colunas como *V9001*, *V9002*, *V8000*, e *RENTA_TOTAL* fornecem informações sobre o poder aquisitivo e condições econômicas, que são fundamentais para entender a demanda e a capacidade de pagamento dos grupos pesquisados.

D. Fatores de Correção e Peso:

- *DEFLATOR*, *FATOR_ANUALIZACAO*, *PESO*, e *PESO_FINAL* ajudam na normalização ou ajuste dos dados para garantir precisão e relevância.

E. Tipo dos dados:

Nome da Coluna:	Tipo do Dado:
UF	int64
ESTRATO_POF int64	int64
TIPO_SITUACAO_R EG	int64
COD_UPA	int64
NUM_DOM	int64
NUM_UC	int64
QUADRO	int64
V9001	int64
V9002	int64
V8000	float64
V9010	int64
V9011	int64
DEFLATOR	float64
V8000_DEFLA	float64

COD_IMPUT_VALOR	int64
FATOR_ANUALIZACAO	int64
PESO	float64
PESO_FINAL	float64
RENDAS_TOTAL	float64

6.1.2 Importância para o Projeto:

Esses dados podem proporcionar insights valiosos sobre variações regionais no comportamento do consumidor, o que é crucial para formular estratégias de *Go to Market*.

6.2 Dados CNPJ

O conjunto de dados disponível na pasta “CNPJ” que pode ser acessada através [deste link](#). 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, visamos não apenas entender a estrutura e operações dessas entidades, mas também estabelecer uma fundação sólida para análises futuras que podem impulsionar decisões de negócios.

6.2.1 Tipos de Dados e suas Características:

A. Identificação da Empresa:

- cnpj: Número completo do CNPJ que identifica de maneira única cada empresa.
- cnpj_basico: Parte básica do CNPJ.
- cnpj_ordem: Número de ordem do CNPJ.
- cnpj_dv: Dígito verificador do CNPJ.

- identificador_matriz_filial: Indica se o CNPJ pertence a uma matriz ou filial.

B. Status Cadastral:

- situacao_cadastral: Status atual do cadastro da empresa.
- motivo_situacao_cadastral: Motivo pelo qual a empresa se encontra na situação cadastral informada.

C. Informações de Atividade e Localização:

- id_pais: Identificação do país.
- cnae_fiscal_principal: Classificação Nacional de Atividades Econômicas principal da empresa.
- id_municipio: Identificação do município.
- id_municipio_rf: Identificação do município na Receita Federal.

D. Contato:

- ddd_1, ddd_2: Códigos de área para telefonia.
- telefone_2: Número de telefone secundário.
- ddd_fax: Código de área para fax.

E. 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.

F. Tipo dos dados:

Nome da Coluna:	Tipo de Dado:
data	object
cnpj	int64
cnpj_basico	int64
cnpj_ordem	int64
cnpj_dv	int64
identificador_matriz_filial	int64
nome_fantasia	object
situacao_cadastral	int64
data_situacao_cadastral	object
motivo_situacao_cadastral	int64
nome_cidade_exterior	object
id_pais	float64
data_inicio_atividade	object
cnae_fiscal_principal	int64
cnae_fiscal_secundaria	object
sigla_uf	object
id_municipio	float64
id_municipio_rf	int64
tipo_logradouro	object
logradouro	object

numero	object
complemento	object
bairro	object
cep	object
ddd_1	float64
telefone_1	object
ddd_2	float64
telefone_2	float64
ddd_fax	float64
fax	object
email	object
situacao_especial	object
data_situacao_especial	object
dtype	object

6.2.2 Importância para o Projeto:

Identificação precisa das empresas é fundamental para qualquer análise ou operação corporativa. Compreender o status cadastral e os motivos associados pode ser crucial para análises legais e de compliance. As colunas relacionadas à localização e atividade principal podem ajudar a entender o ambiente operacional e a natureza dos negócios.

6.3 API

A. **Endpoint:** <https://intelfunctiongetdata.azurewebsites.net/api/InteliFunctionGetData>

- **Método Aceito:** GET

- **Token de uso:**

pZh3gmJW_87epswrWDuB7CvQle-KqjsVh2ZJUaifiXd4AzFuOEy98w==

B. 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.

C. 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.

D. 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.

E. 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. 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.

7.1. 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 que solicitou a análise.
- **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. O uso de contêineres Docker permite manutenção individualizada, com as requisições sendo armazenadas em filas de mensagens e eventos.

7.2. 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.

7.2.1 Dados de Entrada

Fontes de Dados: Os dados de entrada provêm de três principais fontes:

- 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 em streaming.

7.2.2 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.

7.3. 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:

7.3.1 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.

7.3.2 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.

7.4. Escolha de serviços adequados para cada etapa do pipeline

7.4.1 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.

7.4.2 Automação de Ingestão

- **Web Scraping (Python):** Extração de dados da web usando scripts Python para coletar informações relevantes. Utilizado para consultar se existe atualizações nos dados do governo e baixar os novos arquivos. (É um desejável para a aplicação não sendo contemplado no escopo inicial).
- **AWS EC2:** Servidor virtual da Amazon Web Services hospedando a API do parceiros.
- **Kafka:** Plataforma de streaming de dados para coleta, processamento e distribuição em tempo real.

7.4.3 Preparação e Armazenamento:

- **AWS S3 (DataLake):** Serviço de armazenamento escalável da AWS para dados brutos antes do processamento.
- **ETL com Apache Spark:** Realiza operações de Extração, Transformação e Carga nos dados armazenados no AWS S3.
- **PostgreSQL:** Banco de dados relacional usado para armazenar dados processados e deixá-los disponíveis para consulta.
- **Banco de Dados OLAP:** Banco de dados otimizado para análise de dados no formato OLAP (Online Analytical Processing).

7.4.4 Análise e Infográfico

- **AWS Lambda:** Executa o processo de Ensemble, que envolve a combinação de modelos para análise estatística.
- **Grafana:** Plataforma de visualização e análise de métricas, utilizada para a criação de dashboards interativos e relatórios.

7.4.5 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.

7.5. Justificativa para a 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:

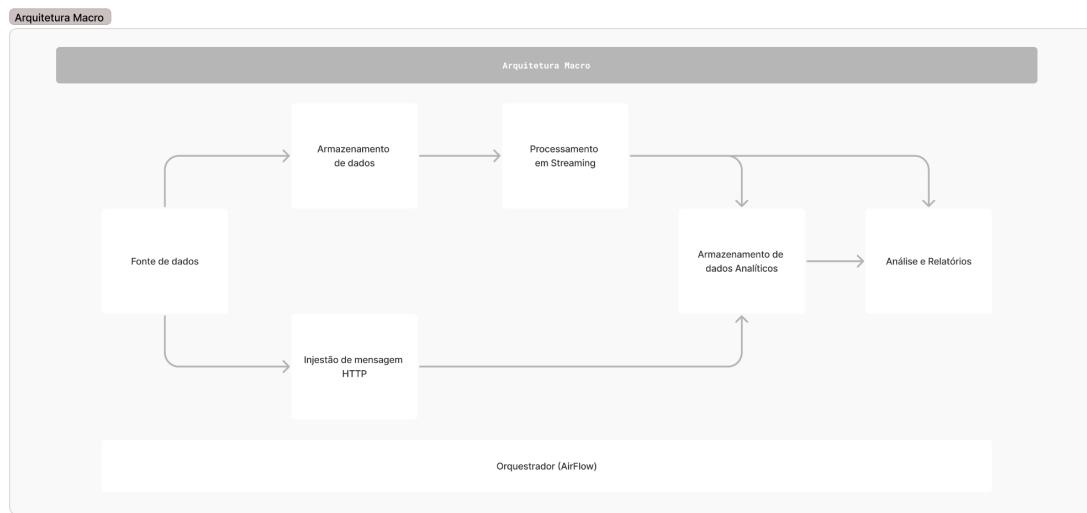
- **Portabilidade:** A maior parte dos serviços foi escolhida com a portabilidade em mente, evitando depender de soluções proprietárias. Isso permite que a arquitetura seja facilmente migrada para outras plataformas de nuvem, se necessário.
- **Uso de Open Source:** A preferência por tecnologias de código 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.

7.5.1 Justificativas específicas

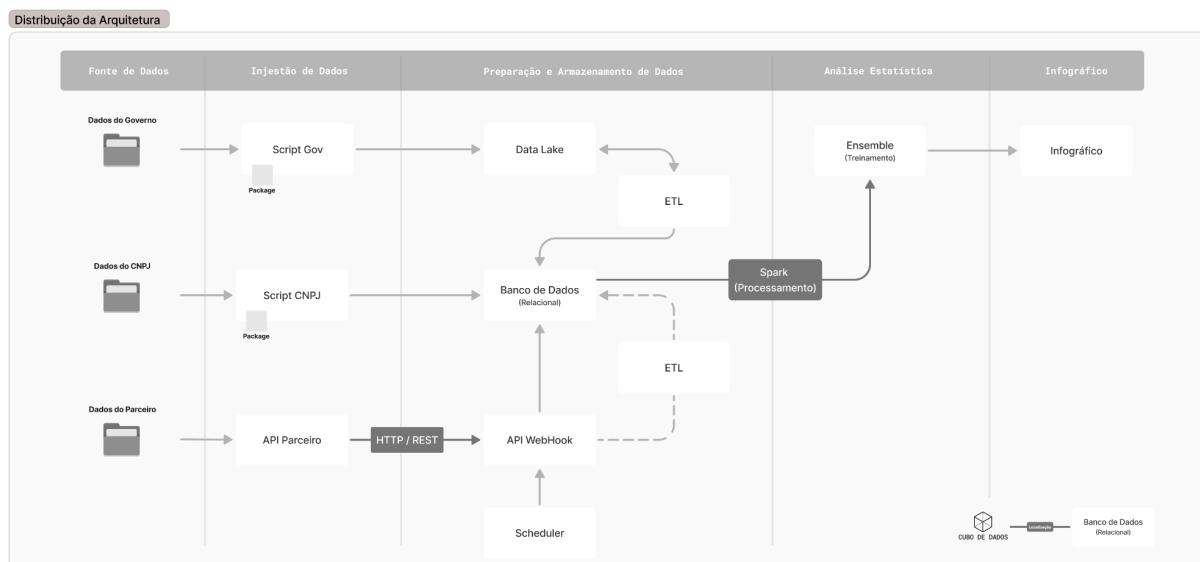
- **Automação de Ingestão:** O uso do Web Scraping e do Kafka oferece flexibilidade e capacidade de escalonamento, atendendo à necessidade de coletar dados em streaming.
- **Preparação e Armazenamento:** A combinação do AWS S3 para armazenamento e do Apache Spark para ETL permite processar dados de forma escalável, enquanto o PostgreSQL oferece um curto custo para armazenamento de dados processados.
- **Análise:** O AWS Lambda e o Grafana foram escolhidos para análise e visualização de dados devido à sua flexibilidade na criação de dashboards e relatórios.
- **Segurança:** O uso do AWS IAM atende à necessidade de gerenciamento de autenticação e controle de acesso, garantindo a segurança dos recursos.

7.6. Representação visual do pipeline

Neste contexto, se detalha a distribuição dos estágios da arquitetura de forma macro, abordando a conexão desde a aquisição de dados, a fase de ingestão, o armazenamento, o processamento, a análise estatística até a geração de relatórios, destacando a interconexão e fluxo contínuo entre essas etapas.

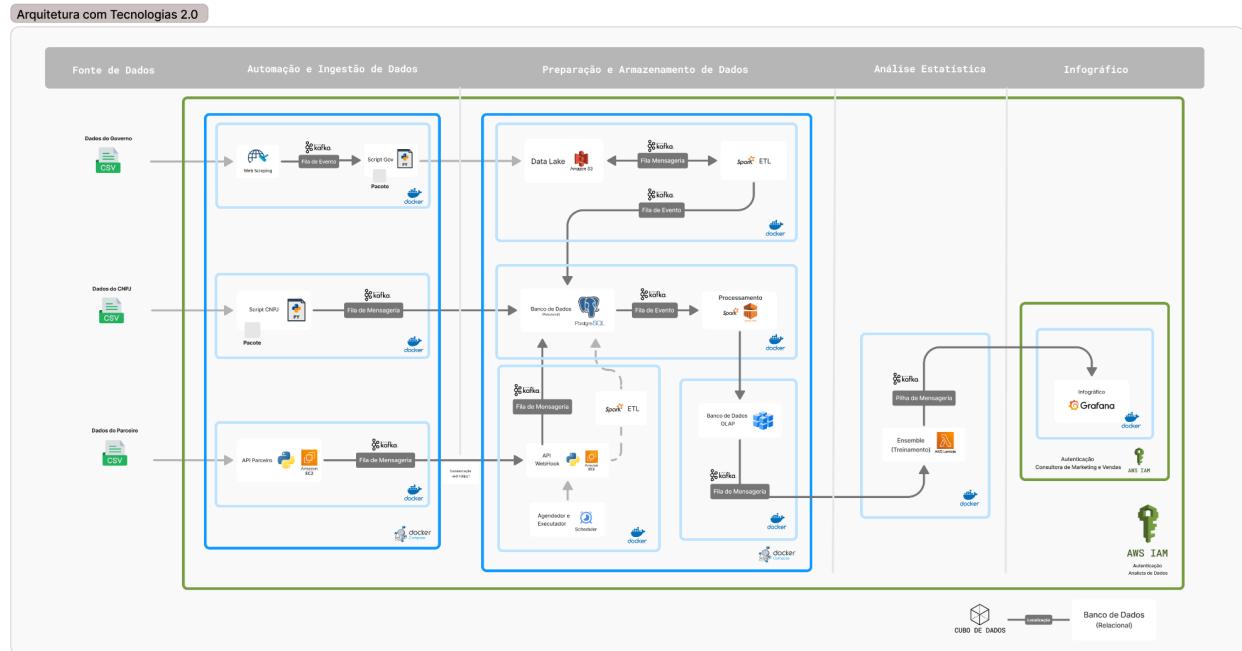


Aqui, se descreve as responsabilidades de cada estágio, fornecendo uma visão detalhada do que será executado em cada um. A imagem ilustra a estrutura minuciosa de cada estágio, como os blocos se encaixam no sistema e o processo que leva à criação do infográfico, sem entrar em detalhes sobre as tecnologias específicas utilizadas.

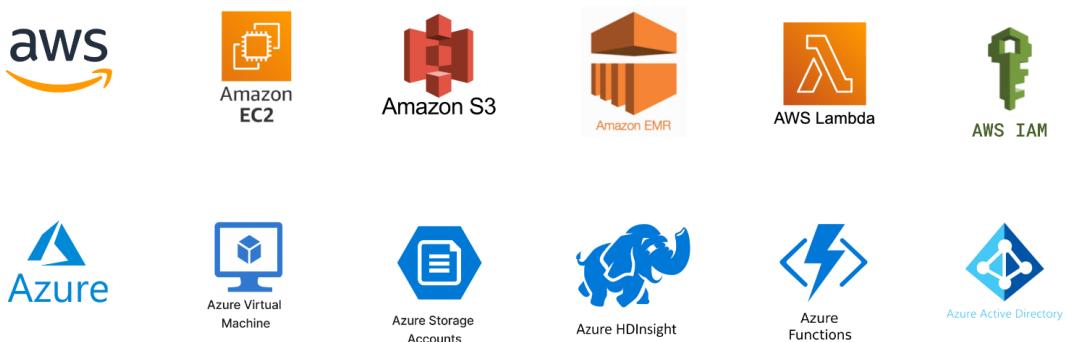


Finalmente, demonstra-se a composição da arquitetura construída, incluindo as especificações de todas as tecnologias mencionadas nos tópicos apresentados. Além disso, a arquitetura é modularizada, com a utilização de contêineres Docker, separando

os serviços da AWS, tecnologias de código aberto, filas de mensagens e detalhando as especificações das requisições.



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.



[Clique aqui, para acessar a Arquitetura Big Data \(Figma Jam\)](#)

7.7. Consideração de boas práticas para garantir 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.
- **Escalabilidade Horizontal:** A capacidade de escalabilidade horizontal foi incorporada na seleção de serviços. Os containers Docker, por exemplo, facilitam a adição de recursos à medida que a demanda aumenta.
- **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.

7.8. Uso de serviços ou recursos da AWS que suportem 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 EC2 (Elastic Compute Cloud):** Serviço de computação em nuvem que fornece servidores virtuais escaláveis. Ele oferece a capacidade de dimensionar horizontalmente as instâncias para atender à demanda.
- **Amazon EMR (Elastic MapReduce):** Serviço de gerenciamento de clusters que facilita a execução de frameworks de processamento de big data, como o Spark.
- **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.

7.9. 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".

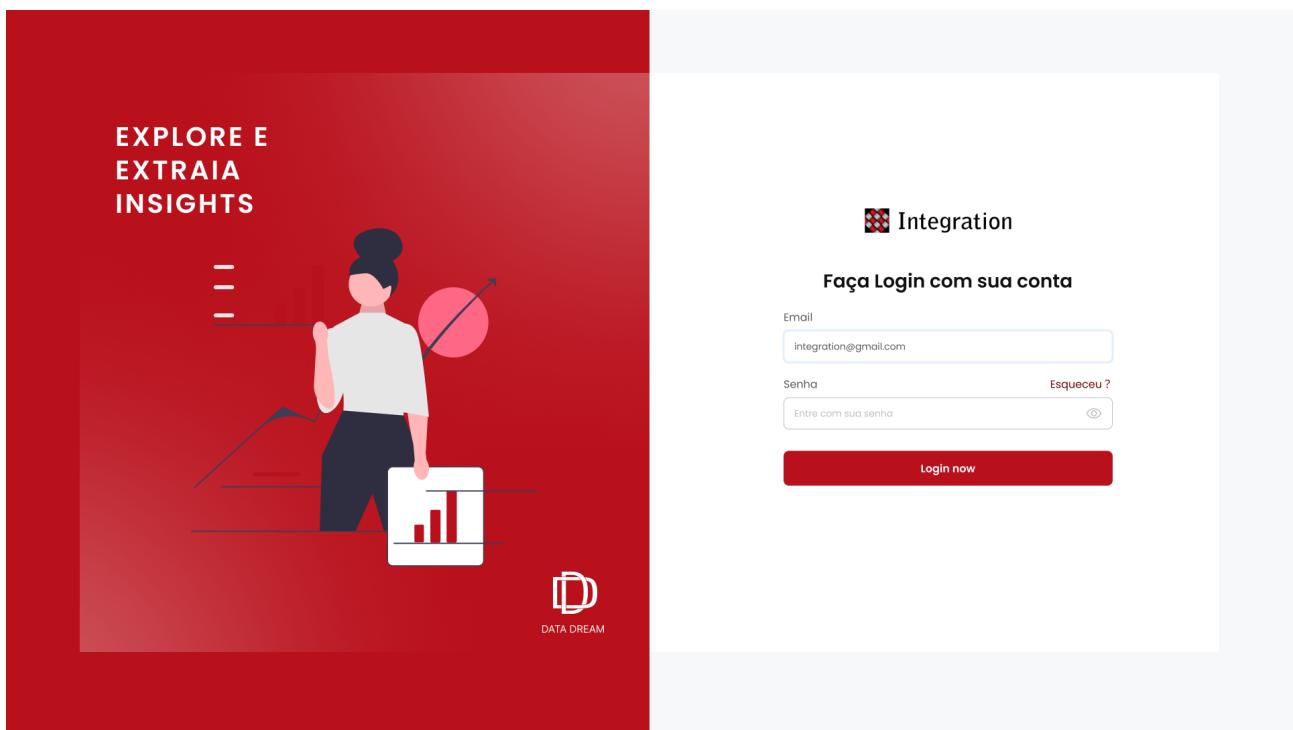
7.10. Arquitetura e a Integration

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. Essa 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.

8. Mockup Interface (Preliminar)

8.1 Tela de Login (Autenticação de Consultor de Marketing e Vendas)

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.



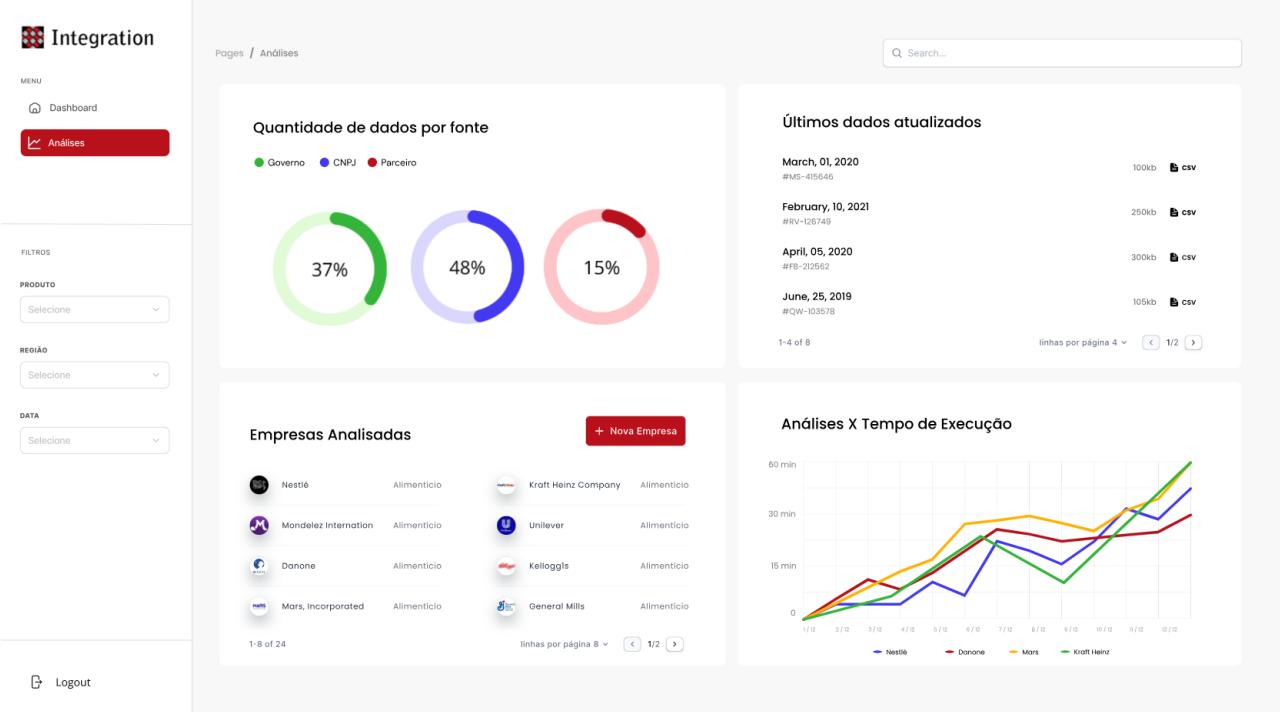
8.2 Tela de Dashboard

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.



8.3 Tela de Dados das Fontes (Governo, Parceiro e CNPJ)

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.



9. Análise Exploratória

9.1. CNPJs

A primeira análise exploratória feita foi a dos dados do Cadastro Nacional de Pessoas Jurídicas, ou CNPJ, sendo um processo fundamental para qual é a estrutura dos csv disponibilizado pelo parceiro. O CNPJ é um registro obrigatório para todas as empresas ativas, o que o torna uma fonte rica de informações sobre a economia e o mercado de trabalho de uma nação. Para este projeto, foi disponibilizado 5 arquivos csv com informações de empresas de 10 CNAEs diferentes. 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.

9.1.1 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 feito 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()
```

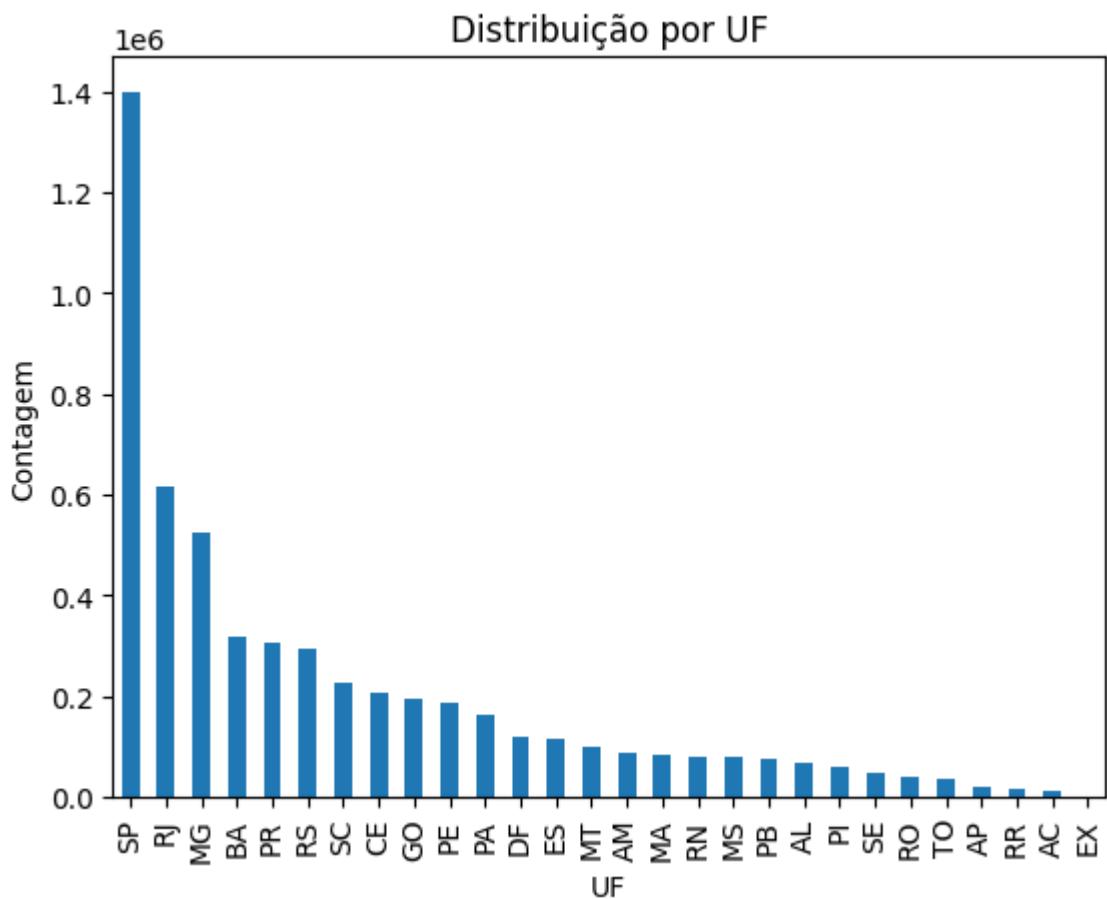


Figura XX: Gráfico “Distribuição por UF” - CNPJ1

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.

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

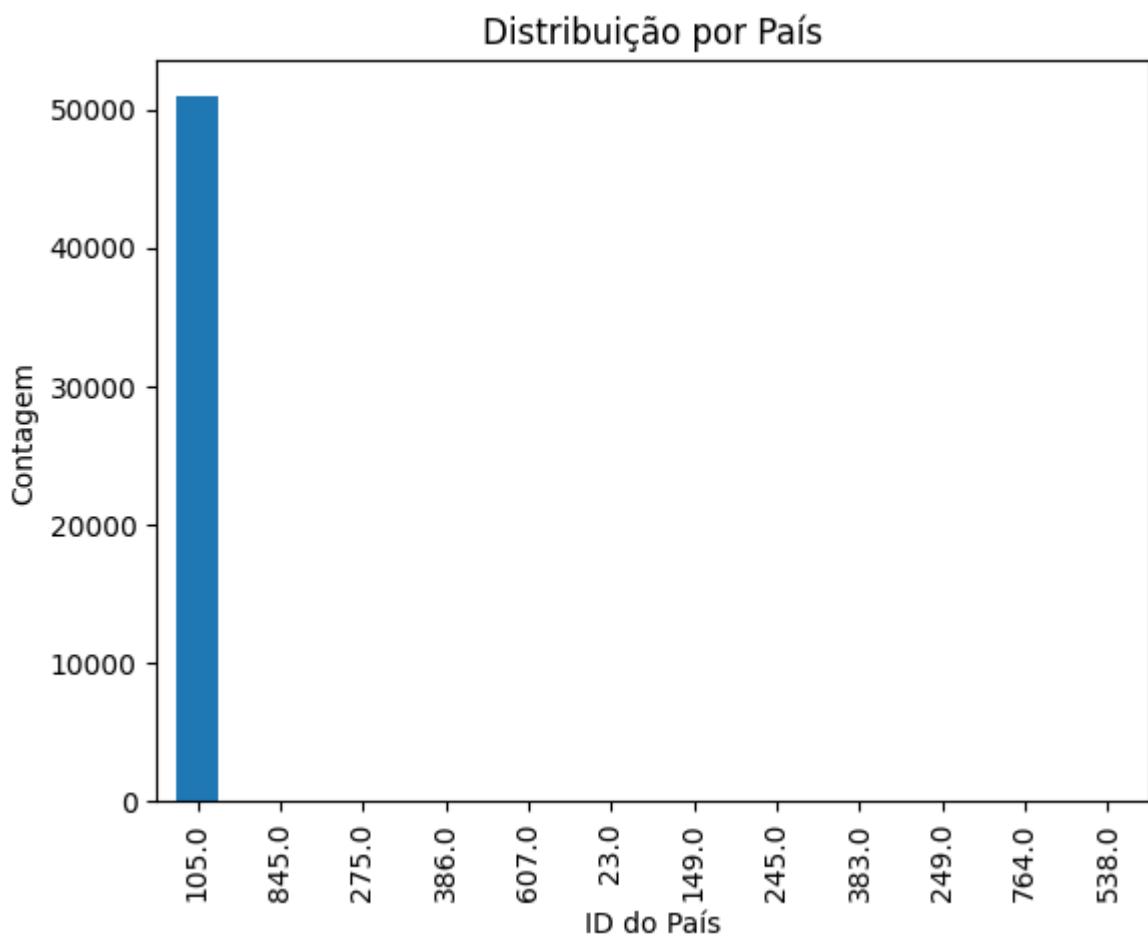


Figura XX: Gráfico “Distribuição por País” - CNPJ1

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 o Brasil. O gráfico a seguir mostra quais empresas são matrizes e quais são filiais.

```
cnpj1['identificador_matriz_filial'].value_counts().plot(kind='bar')
plt.title("Matriz e Filial")
plt.xlabel("ID")
plt.ylabel("Contagem")
plt.show()
```

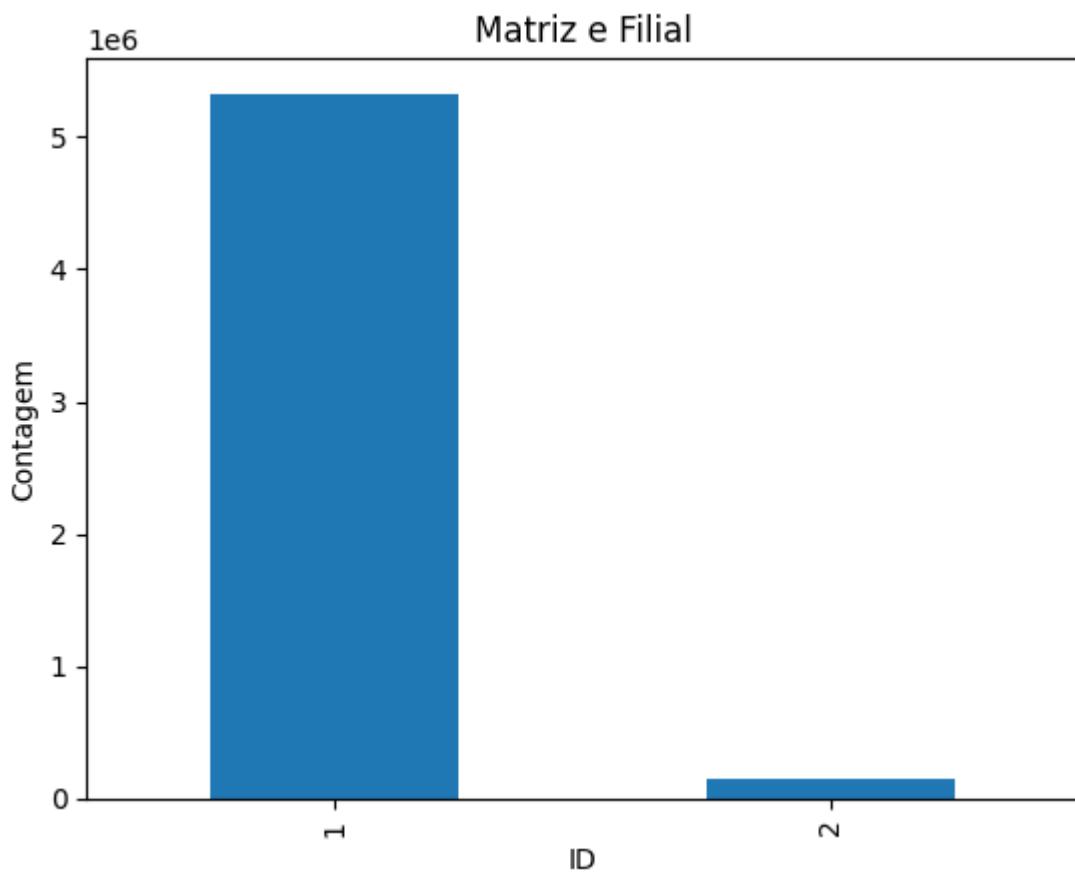


Figura XX: Gráfico “Matriz e Filial” - CNPJ1

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.

9.1.2 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 feito 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()
```

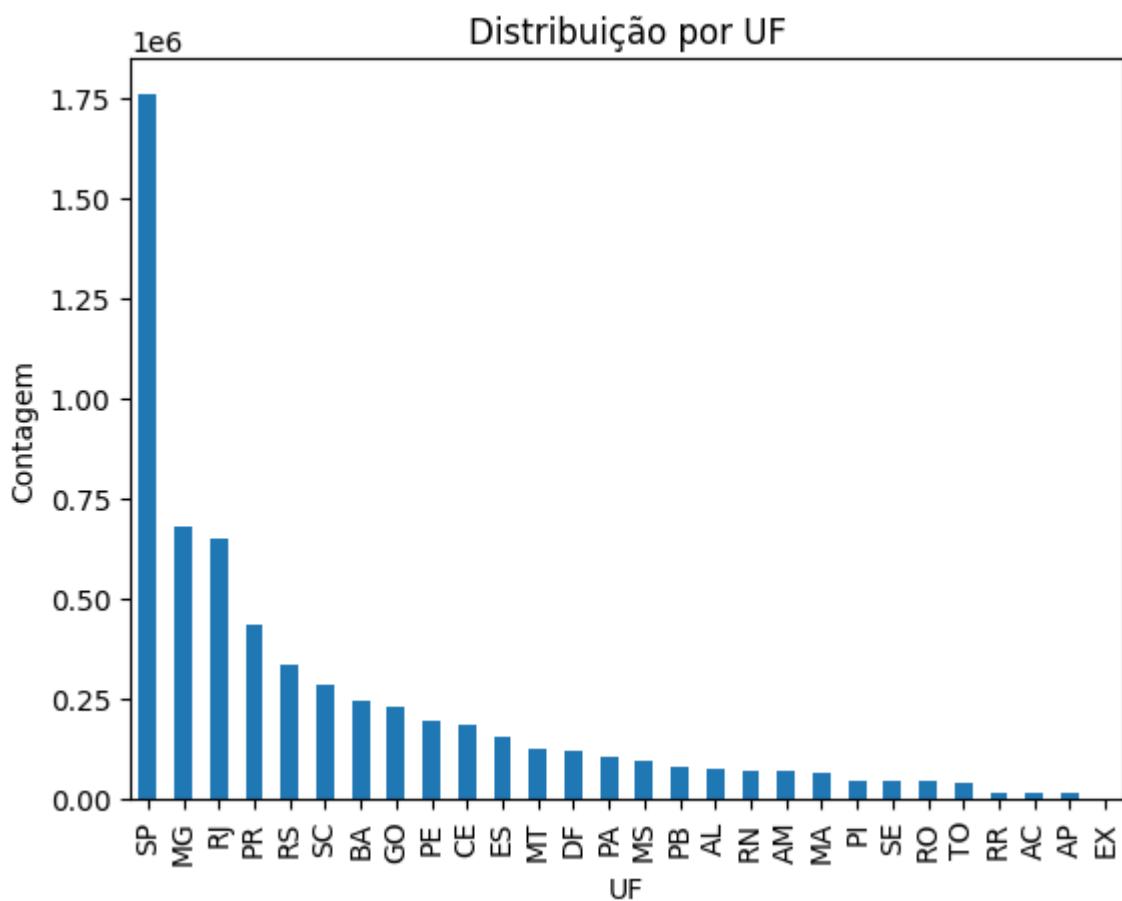


Figura XX: Gráfico “Distribuição por UF” - CNPJ2

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()
```

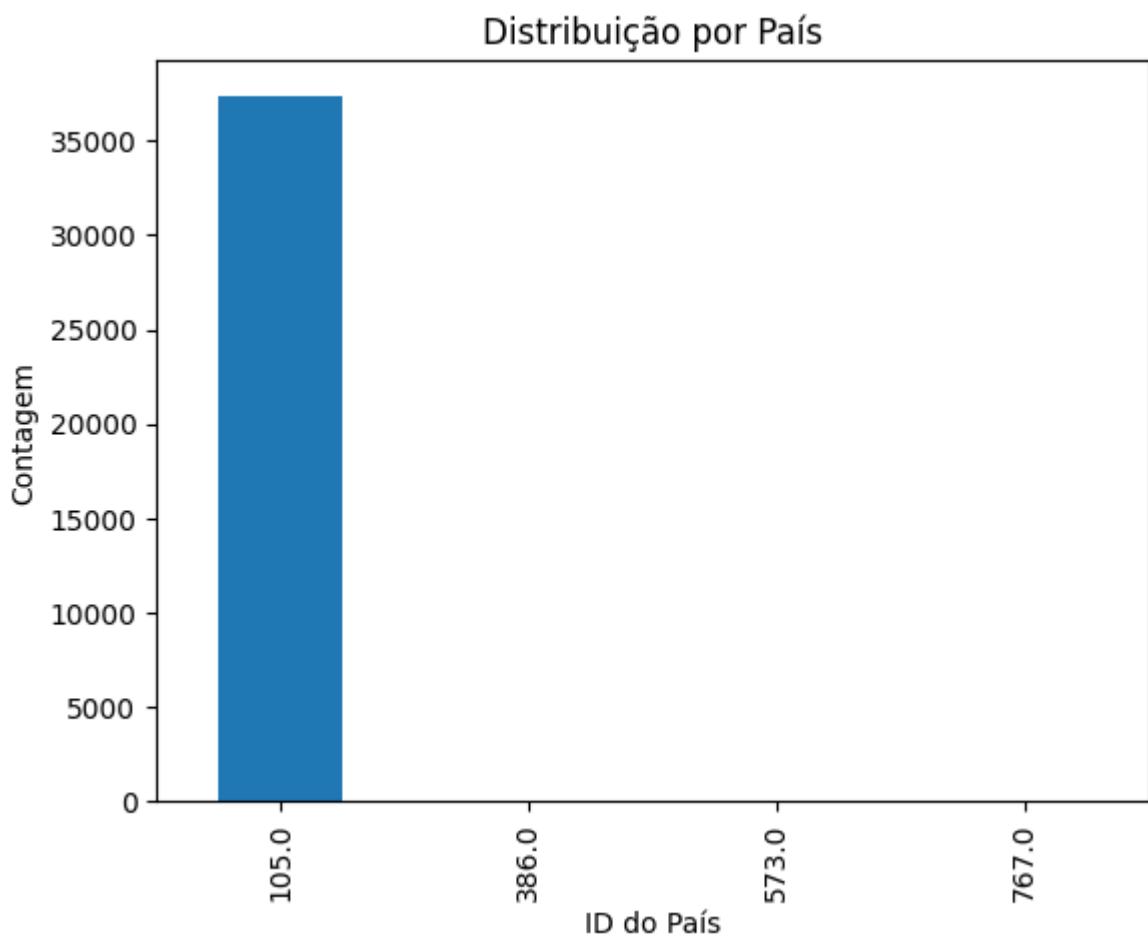


Figura XX: Gráfico “Distribuição por País” - CNPJ2

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 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()
```

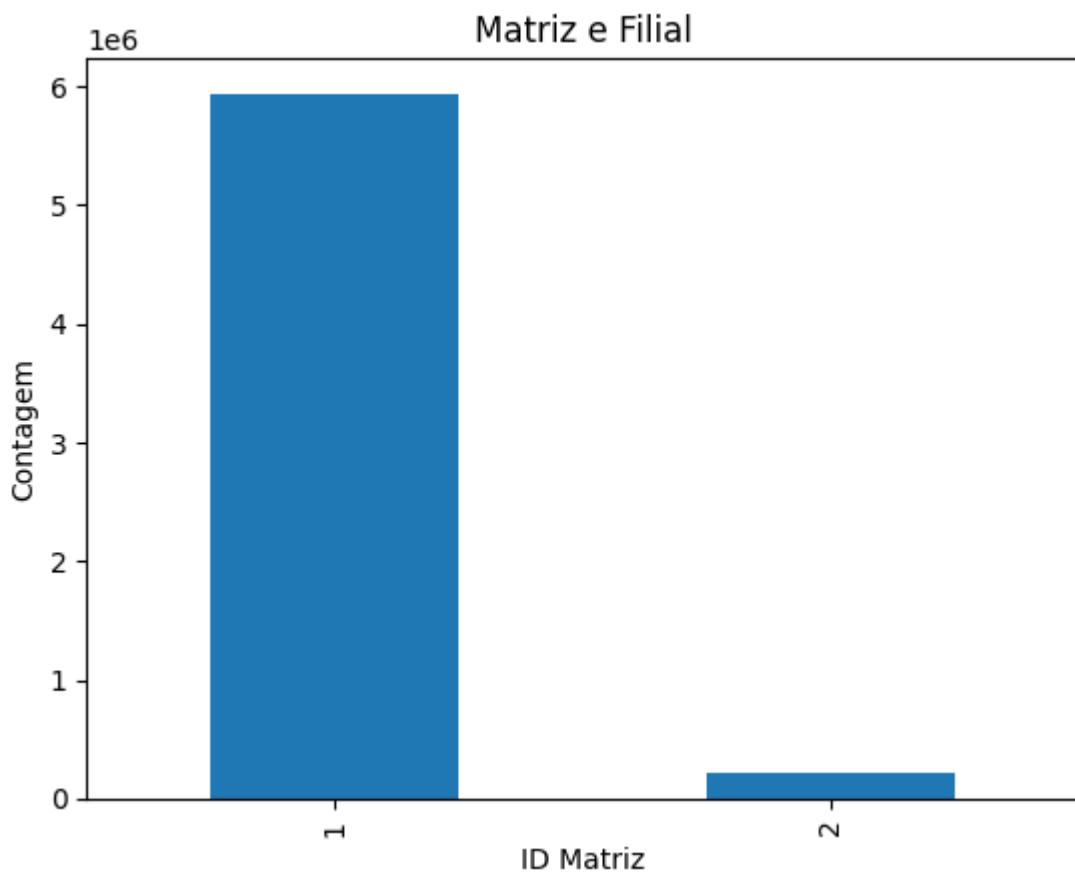


Figura XX: Gráfico “Matriz e Filial” - CNPJ2

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 matrizes do CNPJ.

9.1.3 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 feito 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()
```

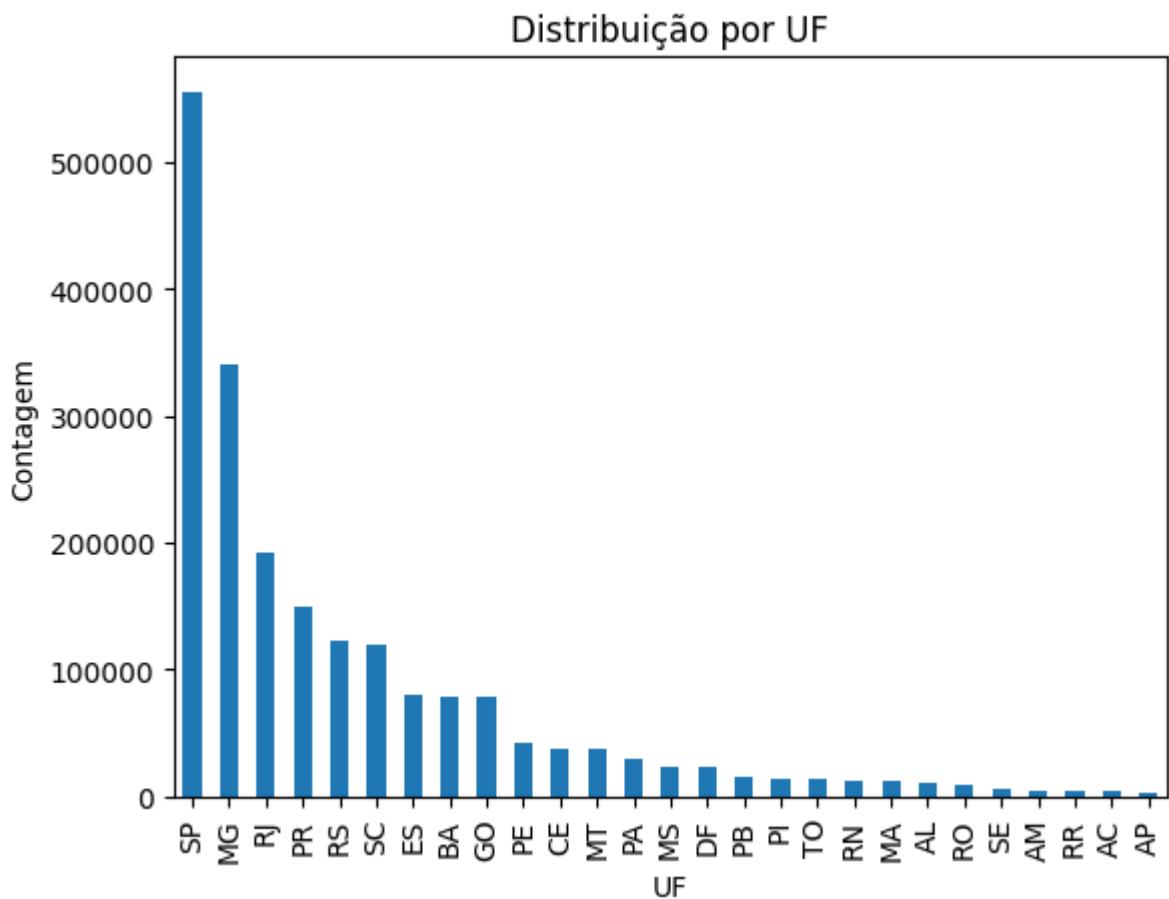


Figura 7: Gráfico “Distribuição por UF” - CNPJ3

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()
```

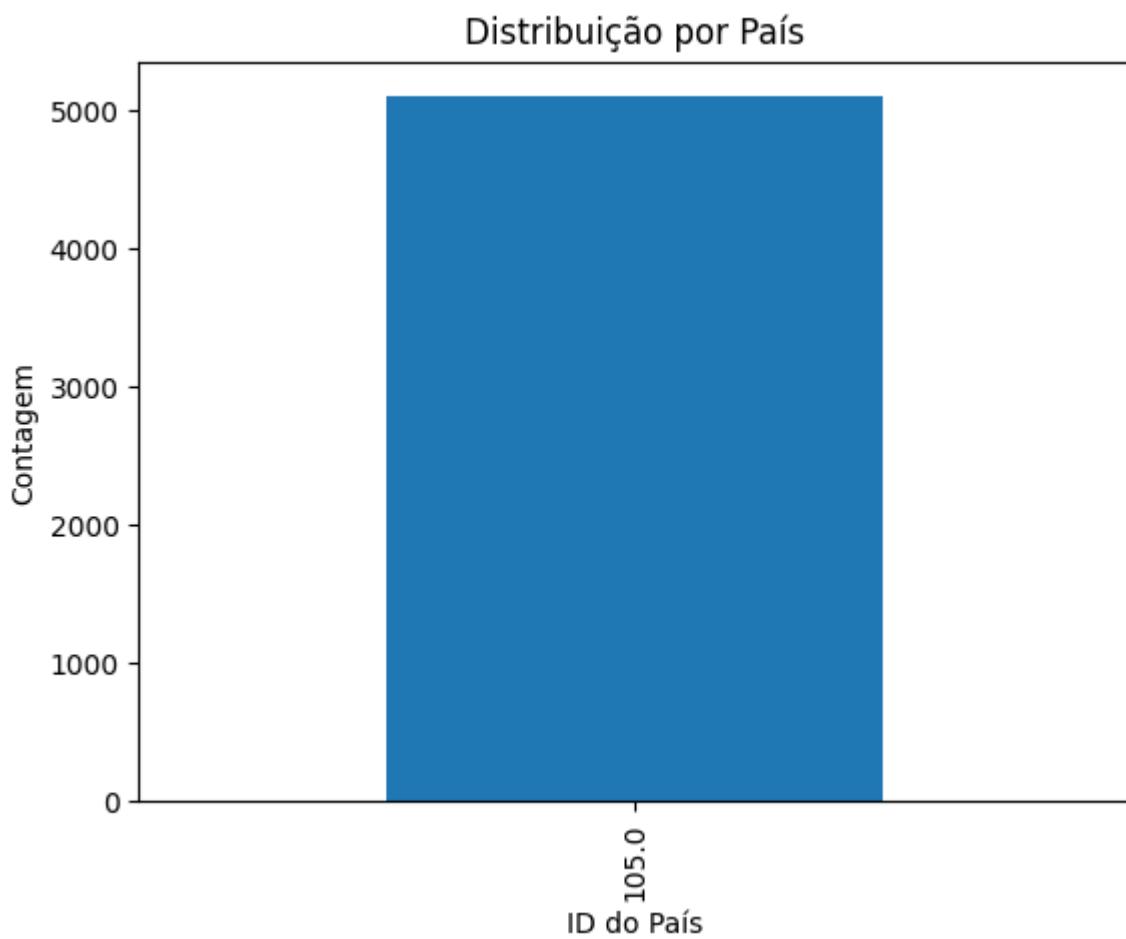


Figura XX: Gráfico “Distribuição por País” - CNPJ3

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()
```

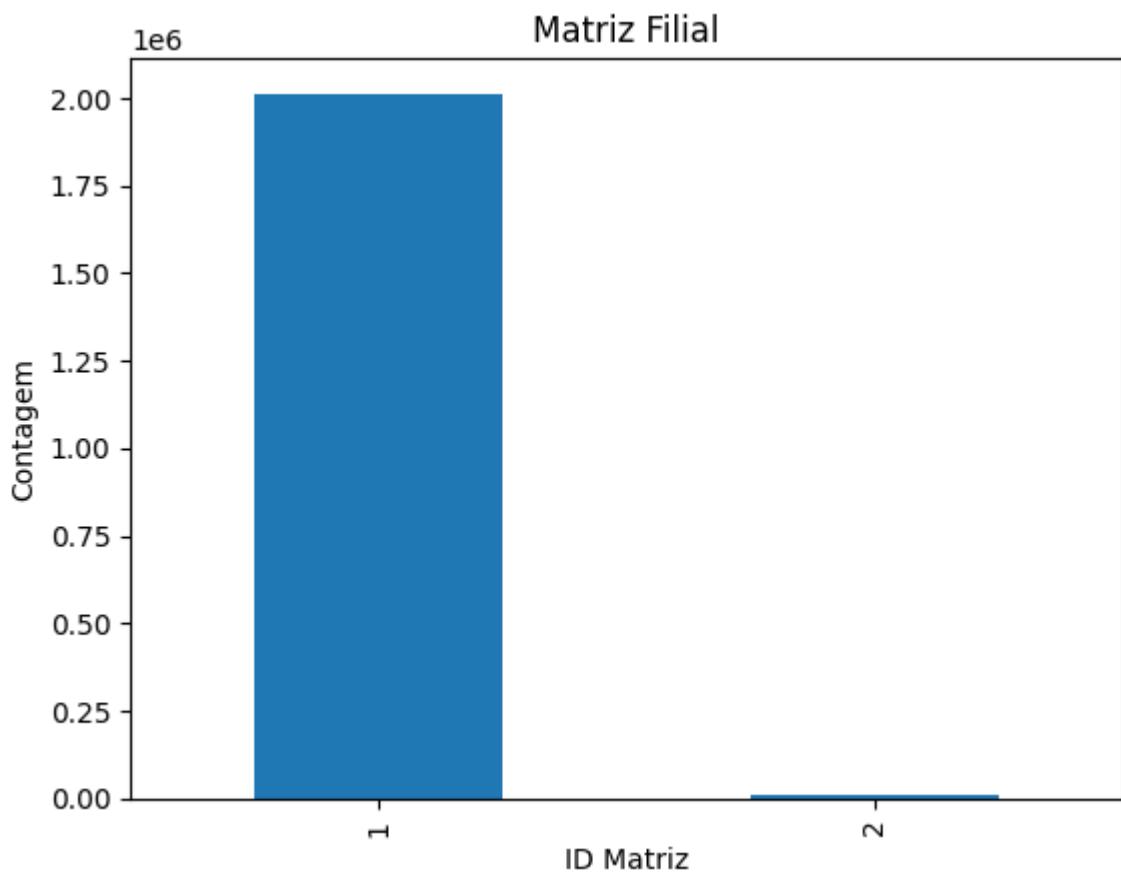


Figura XX: Gráfico “Matriz e Filial” - CNPJ3

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

9.1.4 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.

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

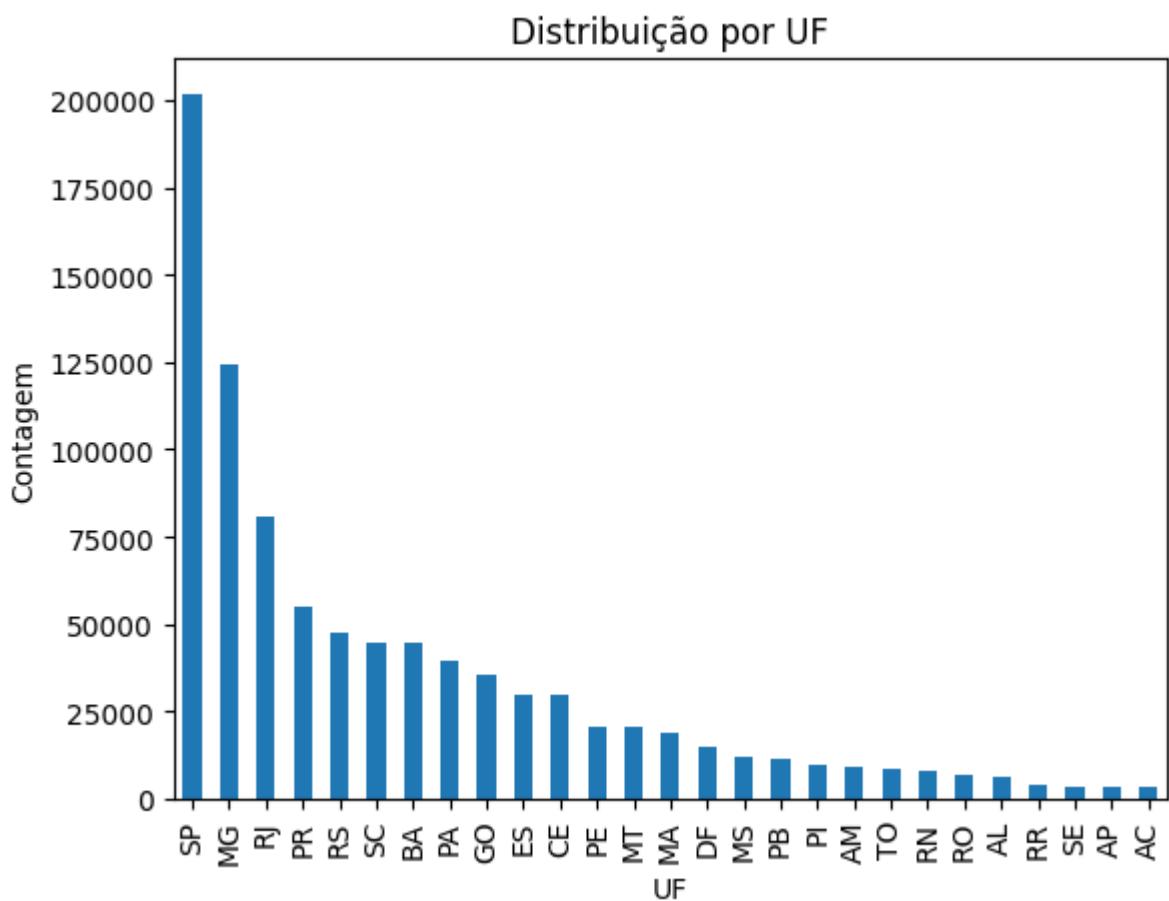


Figura XX: Gráfico “Distribuição por UF” - CNPJ4

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 encontrado 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()
```

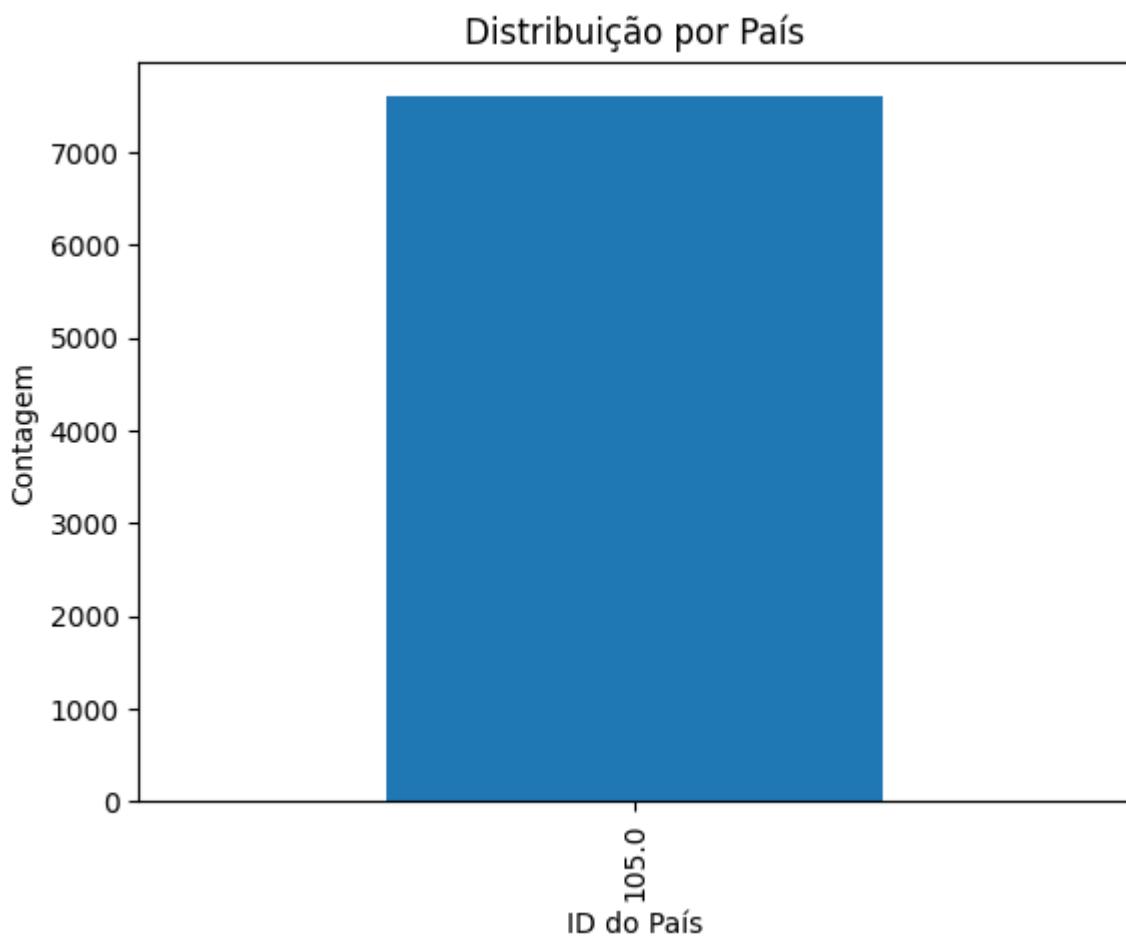


Figura XX: Gráfico “Distribuição por País” - CNPJ4

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()
```

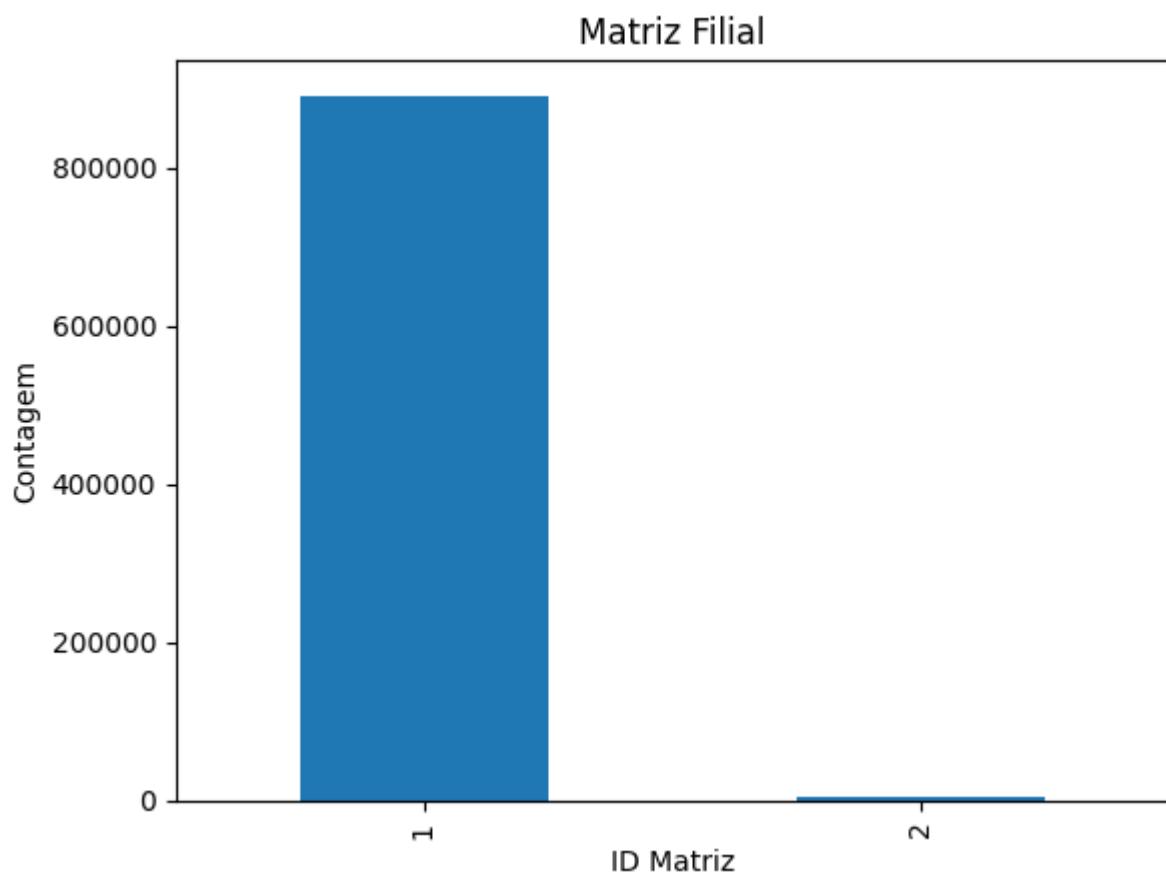


Figura XX: Gráfico “Matriz e Filial” - CNPJ4

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

9.1.5 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 feito foi feito de acordo com a coluna *sigla_uf*, com o código abaixo.

```
cnpj5_1['sigla_uf'].value_counts().plot(kind='bar')  
plt.title("Distribuição por UF")  
plt.xlabel("UF")  
plt.ylabel("Contagem")  
plt.show()
```

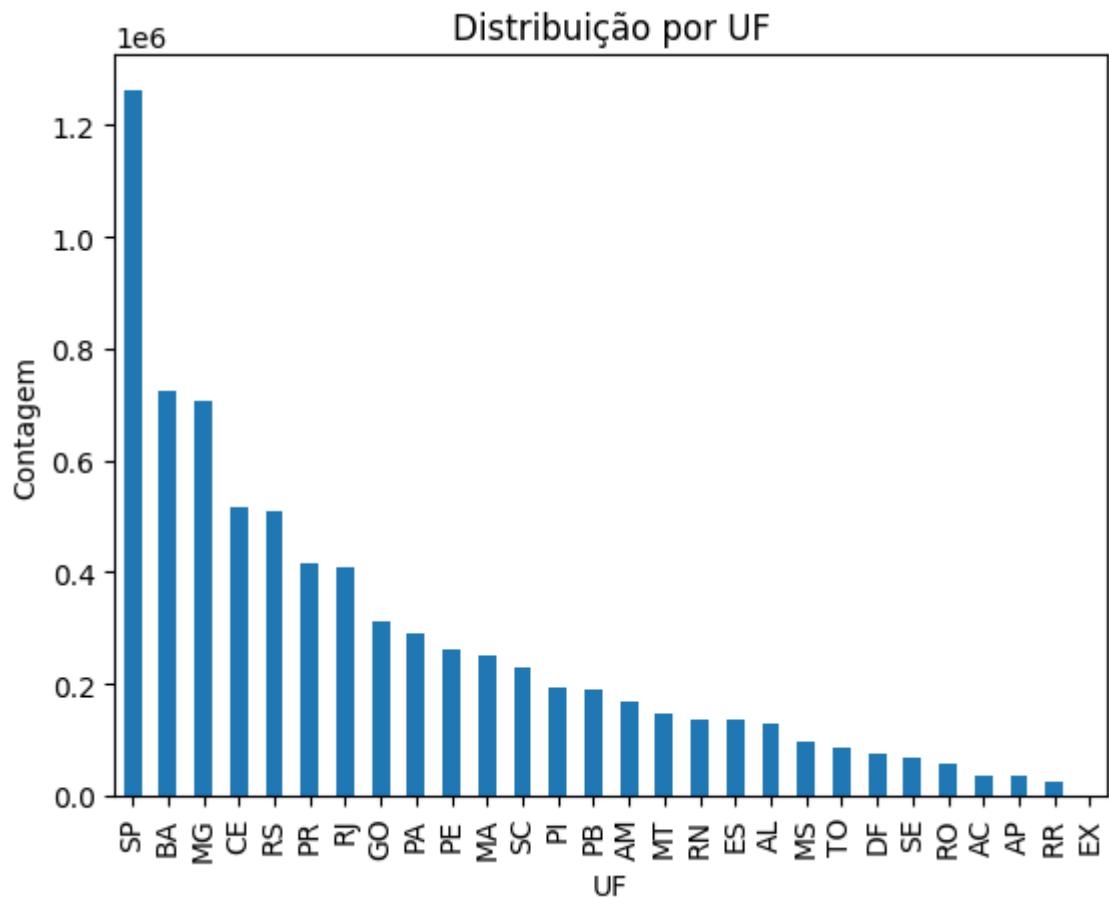


Figura XX: Gráfico “Distribuição por UF” - CNPJ5

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.

```
cnpj5_1['id_pais'].value_counts().plot(kind='bar')  
plt.title("Distribuição por País")
```

```
plt.xlabel("ID do País")
plt.ylabel("Contagem")
plt.show()
```

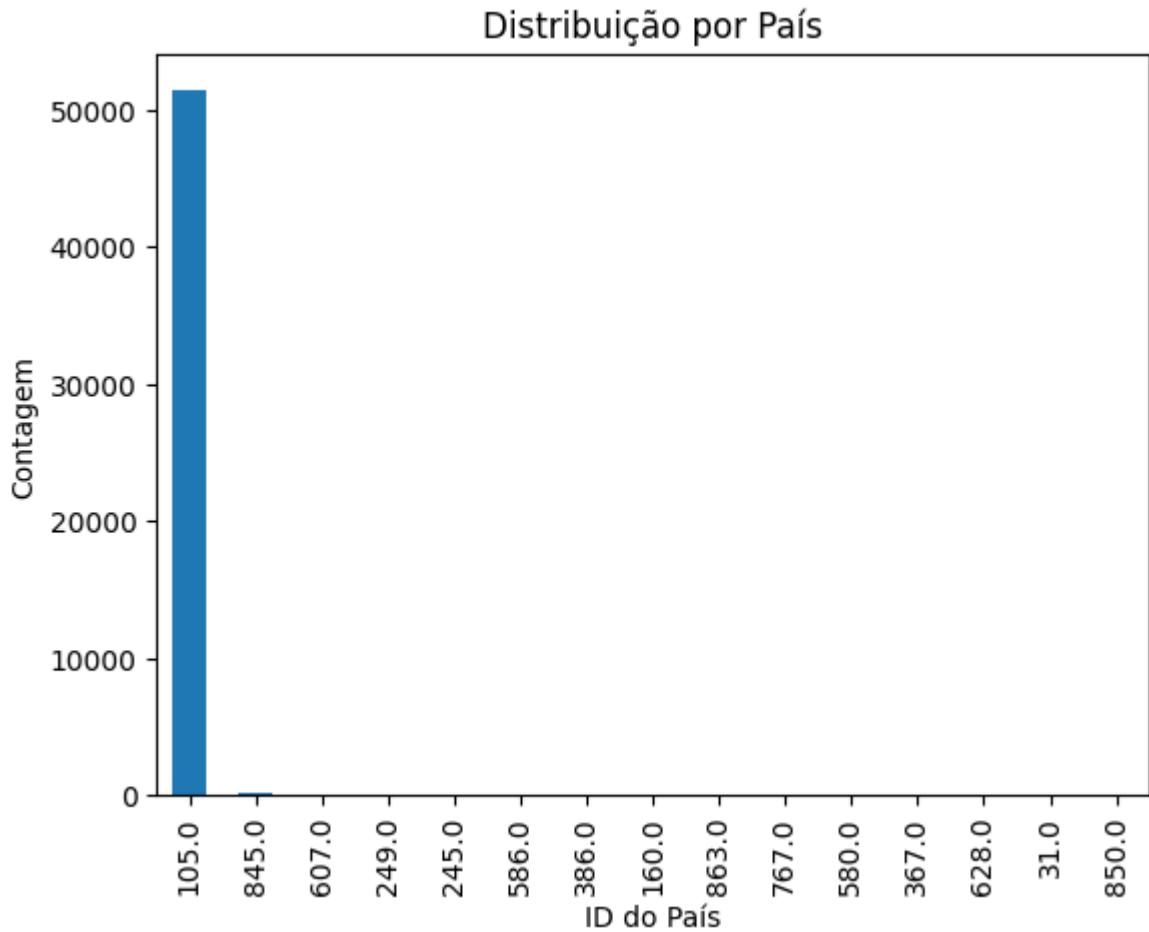


Figura XX: Gráfico “Distribuição por País” - CNPJ5

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 o Brasil. O gráfico a seguir mostra quais empresas são matriz e quais são filiais.

```
cnpj5_1['identificador_matriz_filial'].value_counts().plot(kind='bar')
plt.title("Matriz e Filial")
plt.xlabel("ID")
plt.ylabel("Contagem")
plt.show()
```

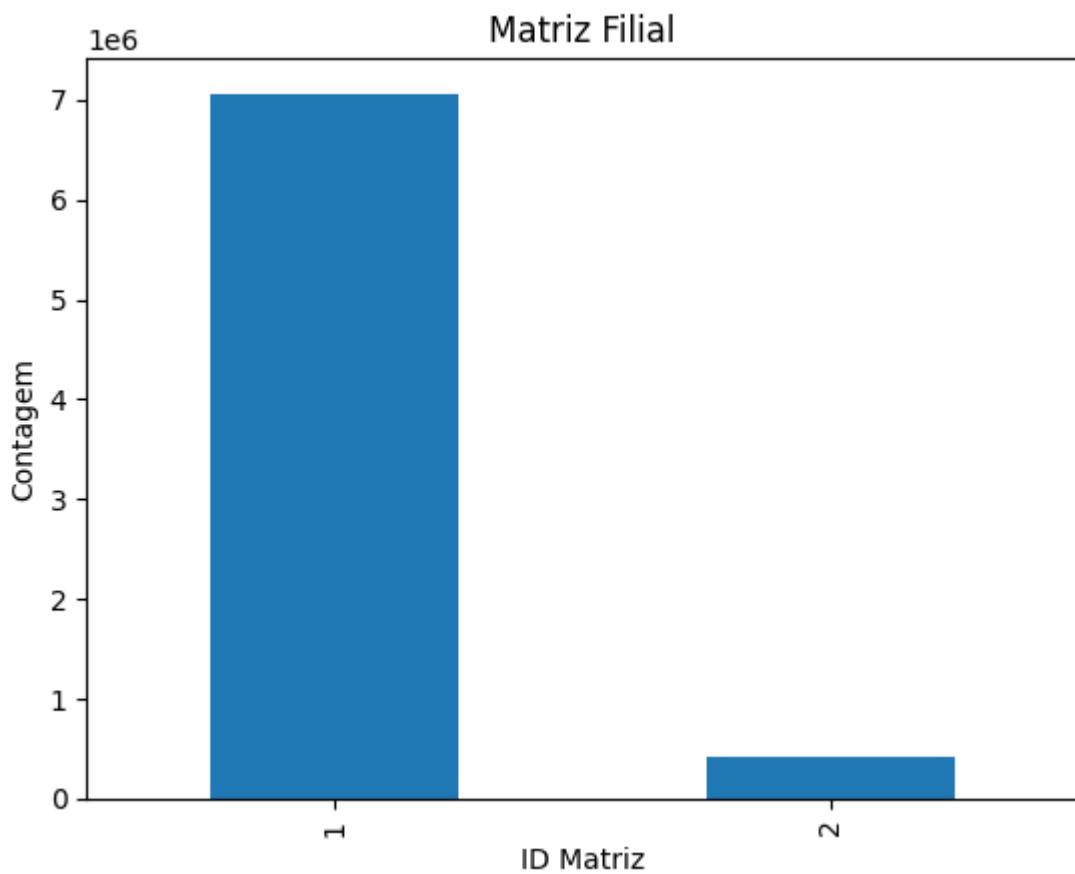


Figura XX: Gráfico “Matriz e Filial” - CNPJ5

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.

9.2. 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.

9.2.1 Aluguel Estimado

O dataset *aluguel_estimado* contém informações sobre os domicílios do Brasil que são alugados e informações sobre. Este é composto por 19 colunas e quase 50 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'})
```

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()
```

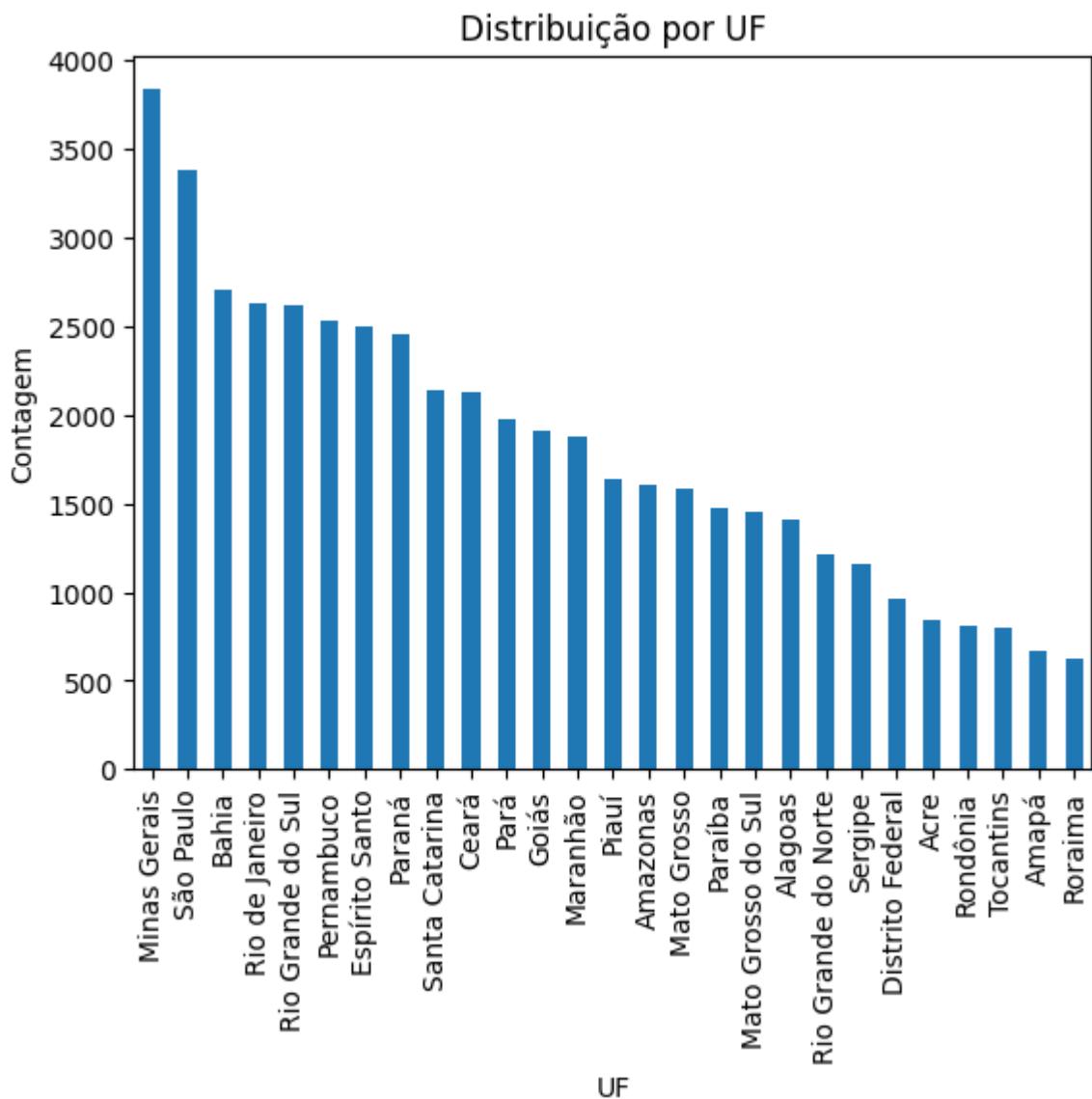


Figura XX: Gráfico “Distribuição por UF” - Aluguel Estimado

Com esse gráfico, pode-se classificar que os estados que mais tem 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'})
```

```
aluguel_estimado['TIPO_SITUACAO_REG'].value_counts().plot(kind='pie',
 autopct='%.1f%%')
plt.title("Tipo de situação regional")
```

```
plt.gca().set_ylabel('')
plt.show()
```

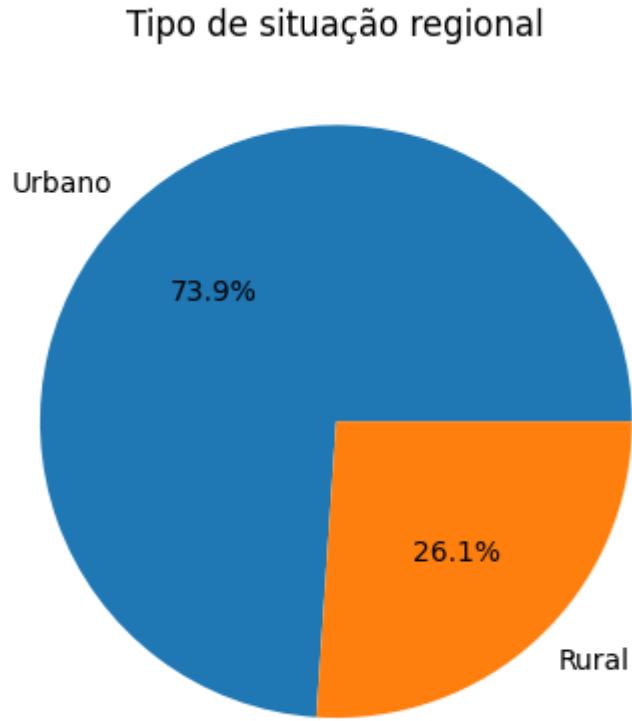


Figura XX: Gráfico “Tipo de situação regional” - Aluguel Estimado

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 que 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()
```

O valor do aluguel estimado foi imputado

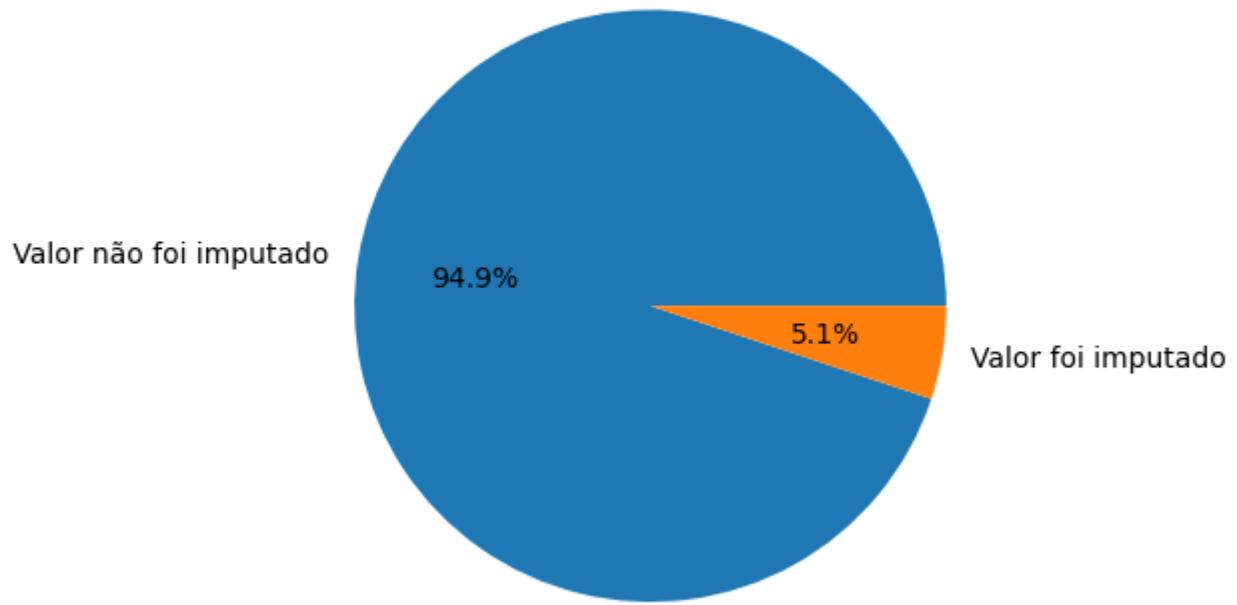


Figura XX: Gráfico “O valor do aluguel estimado foi imputado” - Aluguel Estimado

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

9.2.2 Domicílio

O dataset *domicilio* 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'})
```

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()

```

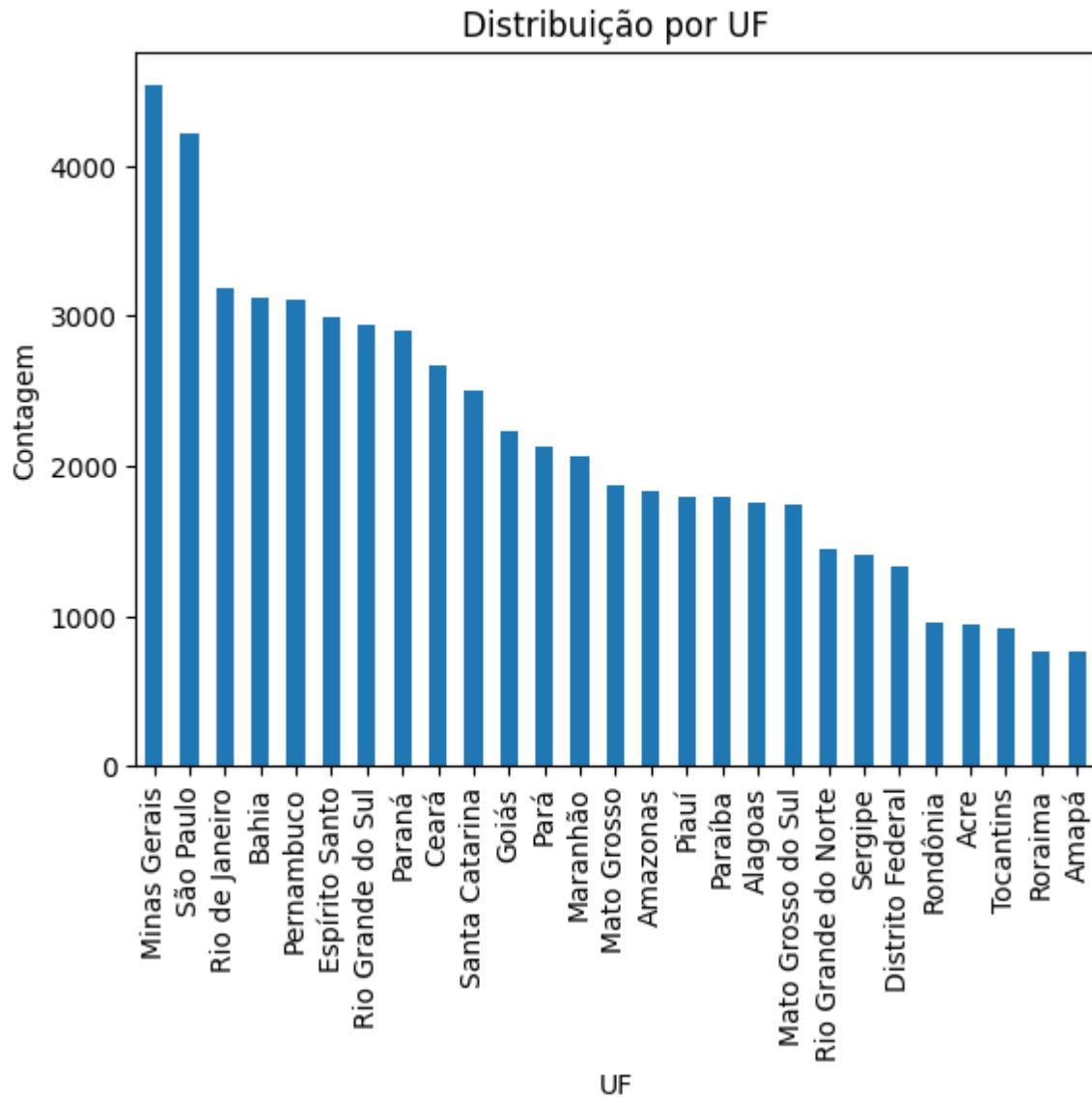


Figura XX: Gráfico “Distribuição por UF” - Domicílio

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'})

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

```

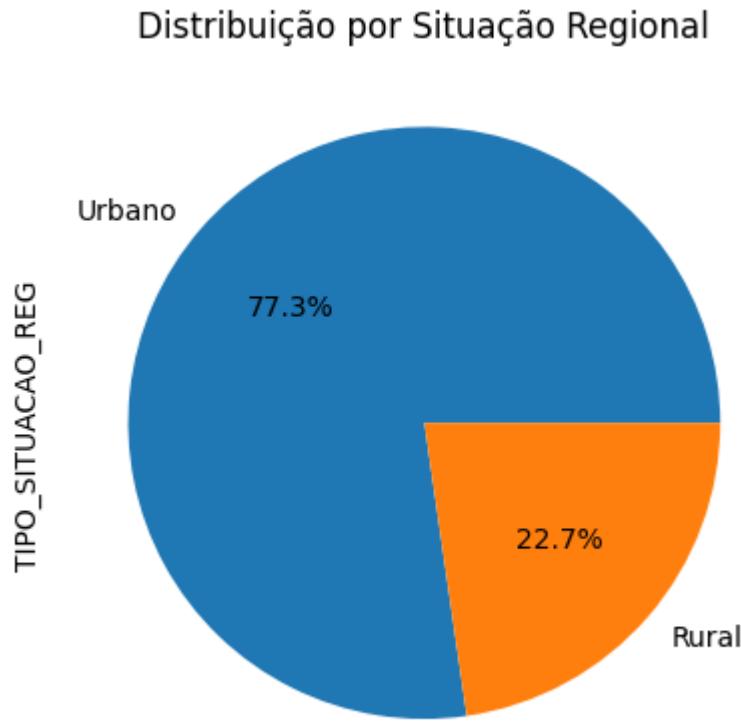


Figura XX: Gráfico “Tipo de situação regional” - Domicílio

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, mas 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'})

domicilio['V0201'].value_counts().plot(kind='pie', autopct='%1.1f%%')

```

```
plt.title("Distribuição por tipo de domicilio")
plt.show()
```



Figura XX: Gráfico “Distribuição por tipo de domicilio” - Domicílio

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'})
```

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

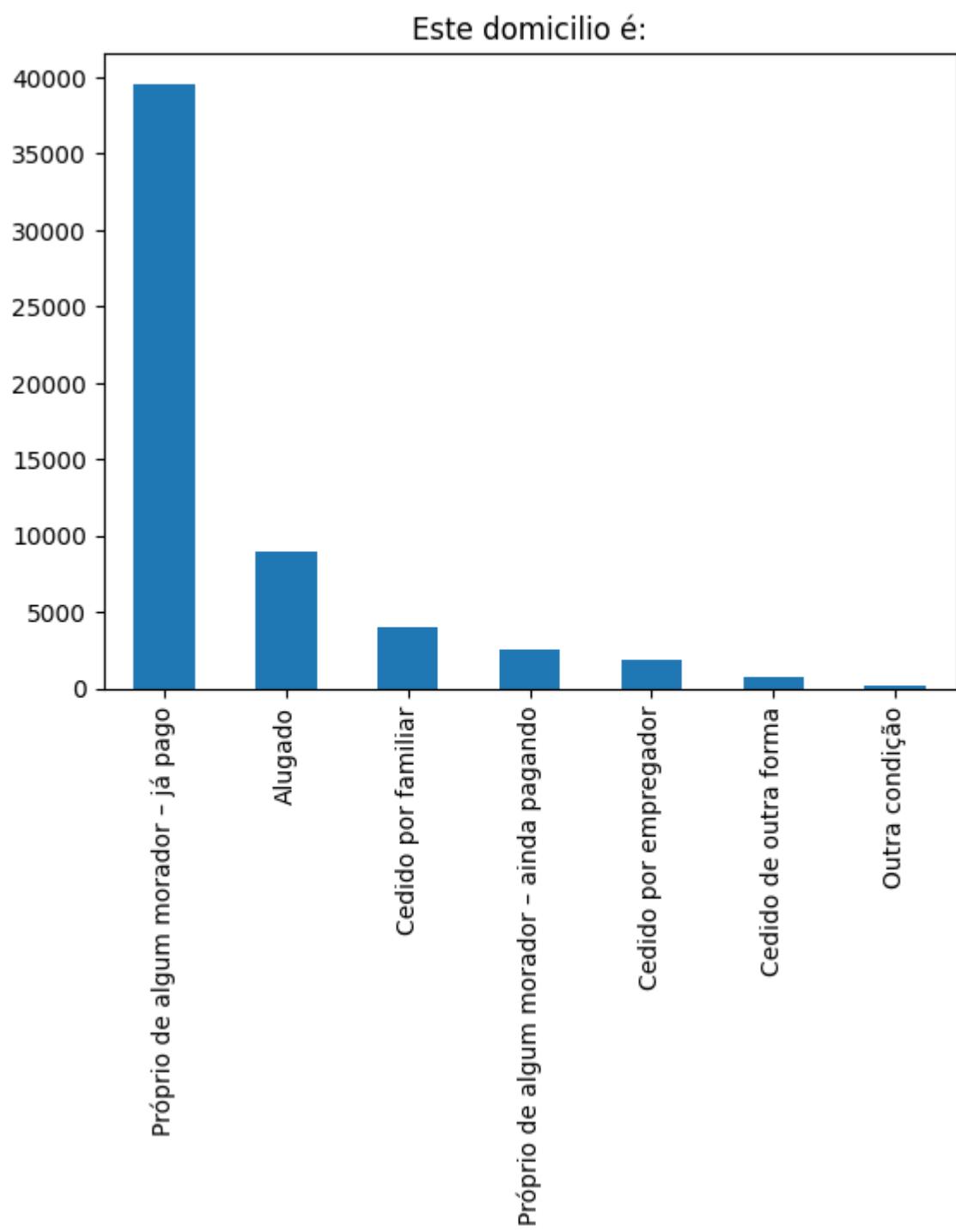


Figura XX: Gráfico “Propriedade do Domicílio” - Domicílio

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()

```

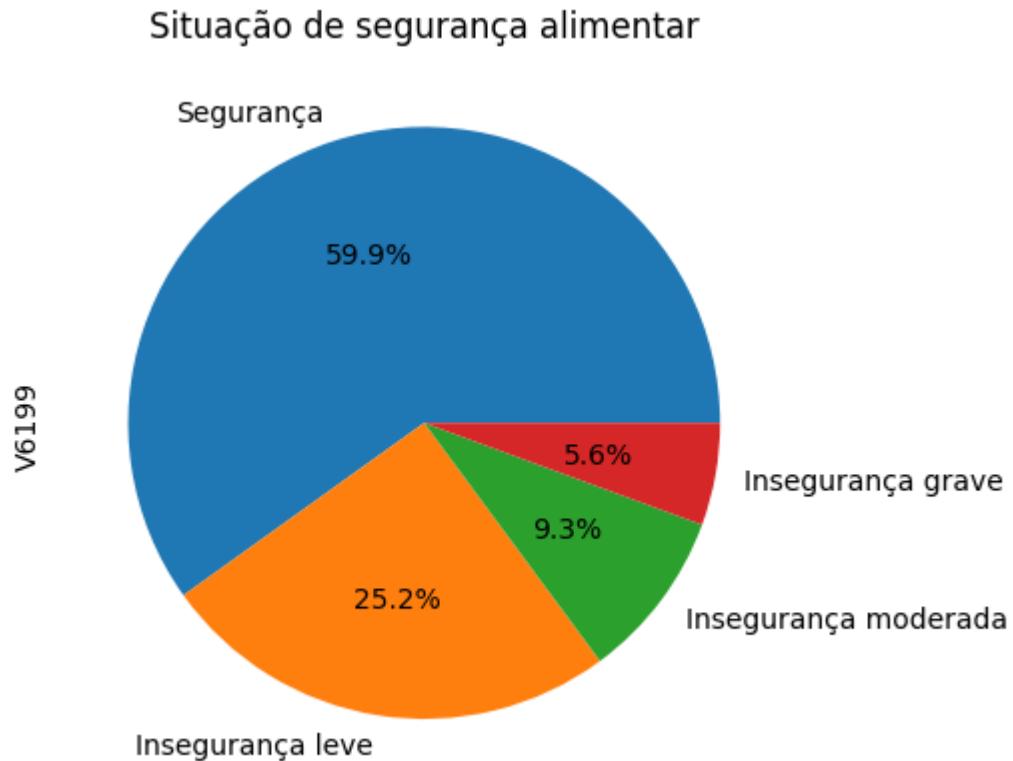


Figura XX: Gráfico “Situação de segurança alimentar” - Domicílio

Esse gráfico demonstra como as pessoas se sentem seguras em trazer comida dentro de casa, e pode-se observar que mais de metade se sente confortável, esse valor apesar de ser expressivo, ainda não é o ideal, porque os outros 40% ainda sofrem de alguma tipo de insegurança. O último processo é aplicar o código `.fillna` em colunas que tem 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)
```

9.2.3 Inventário

O dataset *inventario* 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'})
```

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()
```

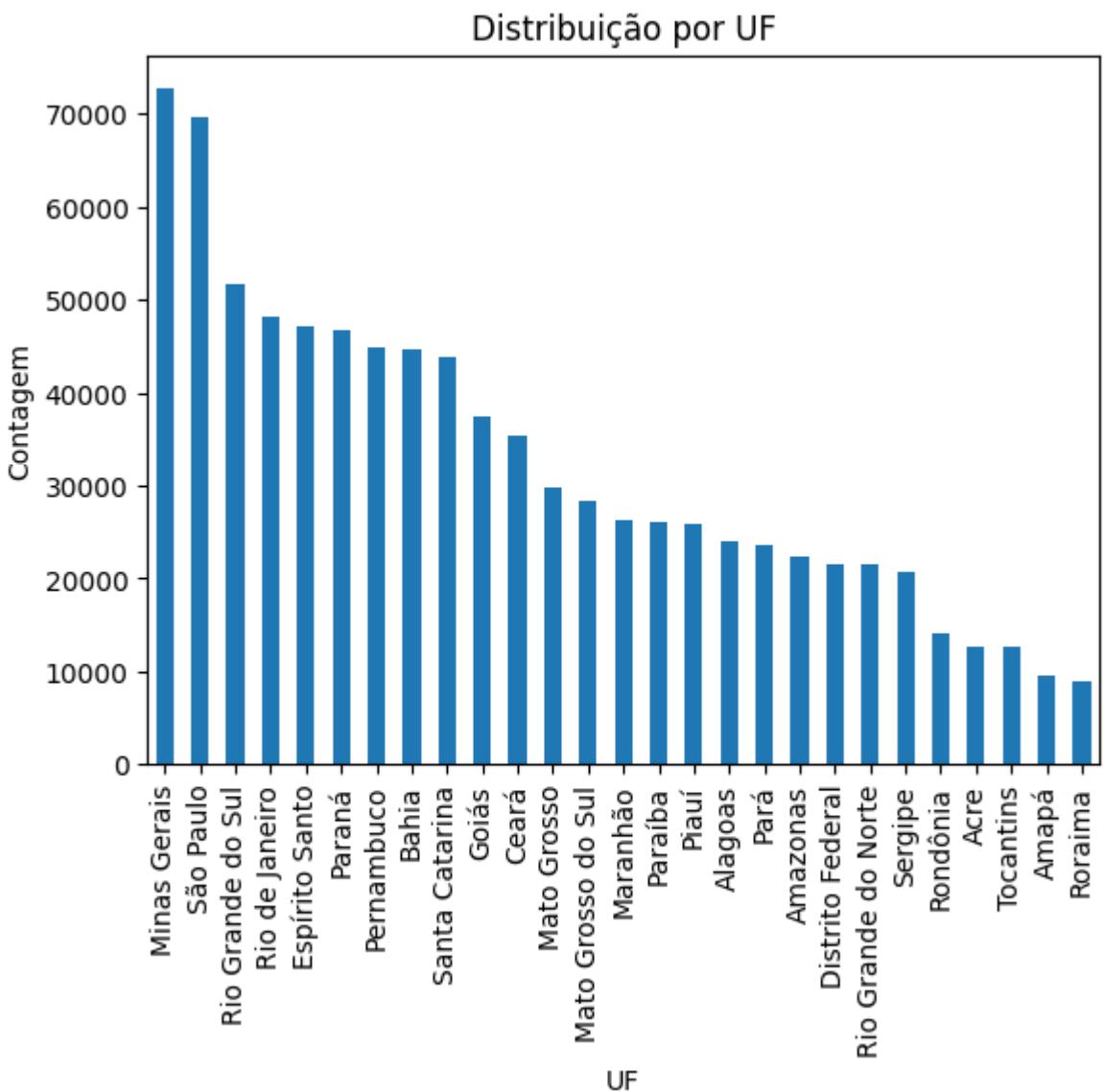


Figura XX: Gráfico “Distribuição por UF” - Inventário

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()
```

Distribuição por Situação Regional

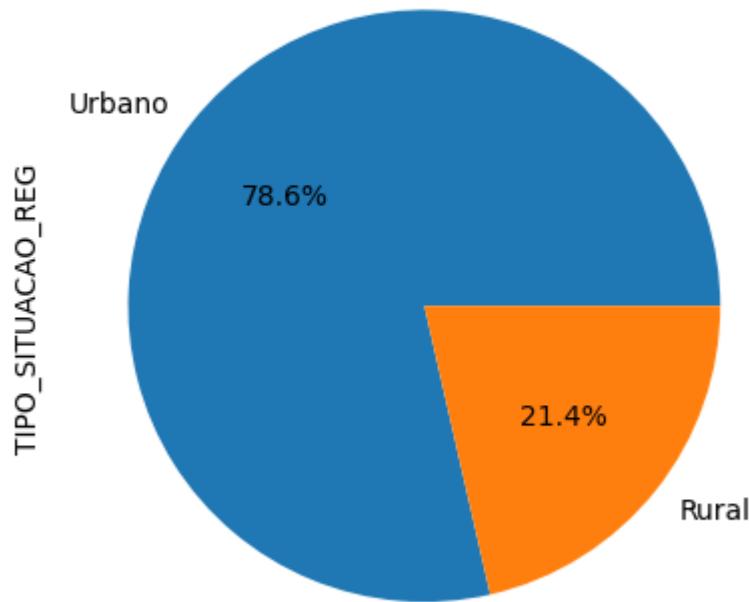


Figura XX: Gráfico “Tipo de situação regional” - Inventário

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 domicilio")
plt.show()
```

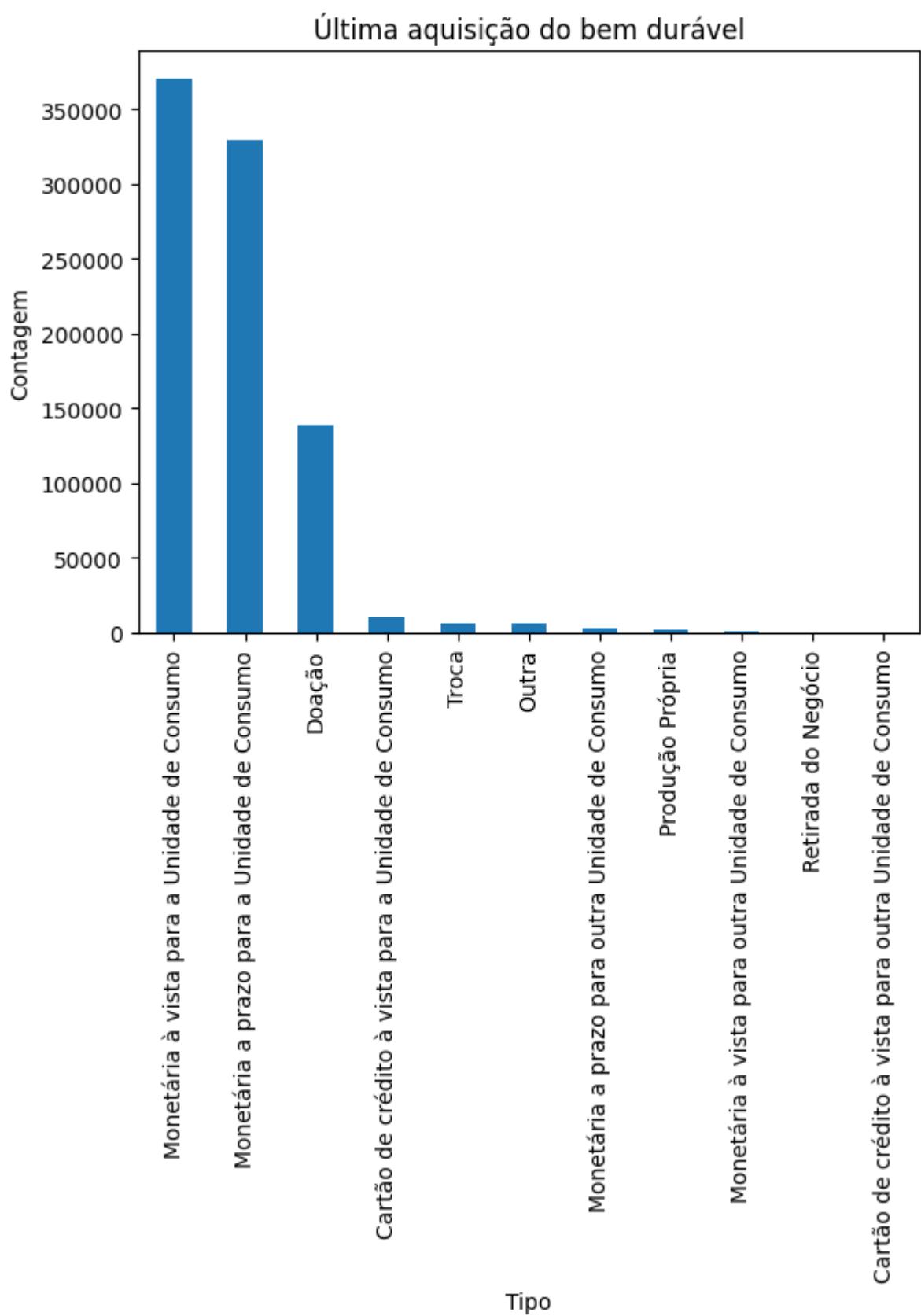


Figura XX: Gráfico “Distribuição por tipo de domicílio” - Domicílio

10. 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-canaliza%C3%A7%C3%A3o-correta-para-lopes/?originalSubdomain=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.