

# formação em dados



ds

## RELATÓRIO

## PARCIAL

### Entrega 11

#### **Dex D6G18**

55388 - HENRIQUE BORG

42426 - JESSICA ANDRADE TIZZIANI

72450 - UÁKITI PIRES DO NASCIMENTO

# BUSINESS UNDERSTANDING

## 1.1 OBJETIVO DO NEGÓCIO

A Prepi é uma empresa de tecnologia, fundada em 2019 em Recife, que gera soluções para pequenas e médias empresas para possibilitar o aumento das vendas por meio de redes sociais e, consequentemente, fomentar o crescimento destas empresas.

Por meio de um aplicativo, de mesmo nome da empresa, permite que os lojistas realizem a gestão de seus ambientes virtuais em um único local para facilitar o contato com os clientes trazendo maior eficiência em seus atendimentos. Além disso, possibilita que o cliente tenha acesso a todos os seus canais de divulgação centralizado, não sendo necessária a gestão individual de cada plataforma onde seus produtos são disponibilizados. Outro recurso ofertado, é a possibilidade do lojista ter seu fluxo de venda automatizado.

O aplicativo traz ao cliente uma trilha de conhecimento que, com base no histórico da loja, sugere estratégias e modelos que podem ser seguidos para impulsionar a vendas e também uma rede de compartilhamento de informações com outros lojistas, fazendo com que a troca de experiências entre eles seja possível.

Trata-se de um modelo de negócio B2B, ou *business to business*, e significa que a empresa negocia e vende seus produtos e serviços para outras empresas. Deste modo, o cliente final não é uma pessoa física e sim uma empresa.

Os setores da empresa são compostos por uma equipe de 15 membros, formada por um *Growth*, que é quem gerencia todas as métricas de saúde do time, além de ser responsável pela aquisição e retenção de clientes. Já a Gestão, é formada inteiramente pela Diretoria da empresa e é responsável por todas as decisões organizacionais e relacionais da empresa, além disso, cuida da satisfação dos colaboradores e de sua jornada na empresa. Parte dos membros, compõem o Time de Produto, formados majoritariamente por desenvolvedores divididos em dois *squads*: *tech* e *design*, sendo responsáveis pela construção, inovação e melhoria do produto Prepi. Há também o Time de Conteúdo, responsável pelo relacionamento com potenciais clientes.

# BUSINESS UNDERSTANDING

Com relação às métricas do negócio, a empresa utiliza as de atração, conversão e de receita e são elas: Taxa de Conversão, *Return Over Investment* (ROI), Ticket Médio, total de *Leads*, Custo por Aquisição (CAC) e o *Lifetime Value* (LTV).

A visão da empresa é impactar 1 milhão de lojistas até 2026, além de ser referência em Social Commerce na região Latam, mirando no crescimento das vendas digitais e consequentemente aumentando o faturamento da empresa.

Com isso, tal projeto tem como objetivo principal reduzir o custo por aquisição (CAC), a partir da categorização dos clientes em potencial, por meio da base de contatos (*leads*), para que seja possível determinar o cliente ideal (ou prospects). Com o intuito de definir qual o foco de maior investimento da empresa. O mercado acessível se mostra favorável para esta meta, pois cerca de 4MM de empresas ainda não possuem site próprio de vendas, segundo o SEBRAE Data. Ainda podemos definir como objetivo secundário aumentar a taxa de conversão de clientes.

Analisando o mercado em que a Prepi está inserida, a empresa se mostra como uma das mais inovadoras desde seu período de criação, trazendo soluções de E-commerce simplificado aos clientes consolidando sua imagem perante ao mercado. Isso possibilita que, comparado aos concorrentes diretos, a Prepi garanta experiência customizada e *freemium*. Esta modalidade permite ao cliente a liberdade de uso de alguns dos serviços disponibilizados gratuitamente e, caso deseje novas funcionalidades, adquirir a versão premium.

## Quadrante de crescimento

SEMRUSH

Lista de mercado 1 | Todos os países | Nov de 2022 (vs. Out de 2022)

Total | Direto | Referência | Pesquisa | Redes sociais | Pago



Análise de mercado comparando aos concorrentes

# BUSINESS UNDERSTANDING

## 1.2 AVALIANDO A SITUAÇÃO

O objetivo do projeto é auxiliar na identificação da *persona* ideal, para que a Prepi consiga focar seus investimentos. Além disso, é necessário determinar quantas pessoas precisam ser atingidas para que uma venda seja concluída.

A premissa do negócio é de, com o projeto em questão, encontrar formas de reduzir custos ou elevar o faturamento e, conseqüentemente, melhorar o lucro da empresa.

Visto que as tomadas de decisões da empresa são com base em Lead Scoring, será disponibilizado para o projeto duas planilhas bases. A primeira planilha será a de Leads da empresa e a segunda de Leads alvos. Conforme discutido com os membros da equipe Prepi, os dados que serão disponibilizados em planilha .csv serão suficientes para que as análises sejam realizadas.

Da empresa, foram apresentados o Sponsor Dyogo Machado e o especialista em dados Ramon Pereira. Ramon é o CTO (Chief Technology Officer) e cofundador da Prepi, responsável por gerenciar todo o time de TI da empresa. Dyogo Machado é responsável pela área de Growth e Head of Aquisitions, e atuará como Sponsor backup do projeto.

Foi definido com o Sponsor de que os prazos do projeto serão de acordo com os prazos estipulados pela DNC para o cumprimento de cada etapa.

Com relação aos riscos do projeto, entendemos que possa ser: **desvio do escopo**, para que isso não ocorra, solicitamos ao Sponsor que atenha-se com firmeza a esse parâmetro para que o objetivo do projeto não tenha desvios. Outro risco é o **baixo desempenho**, que pode ocorrer devido falhas na comunicação entre os membros da equipe, para que isso seja amenizado, está sendo utilizado um software de gestão de projetos, no caso o Trello, para acompanhar os processos em tempo real.

# BUSINESS UNDERSTANDING

Sendo um negócio digital, que se orienta através de estratégias comerciais de Marketing e Vendas, alguns termos comuns ao negócio que devem ser entendidos:

- Lead: cliente que compartilhou algumas informações com a empresa, tais como nome, telefone, e-mail etc. Também pode ser considerado um contato encontrado em uma pesquisa por potenciais clientes.
- Prospects: é o cliente potencial, que mostrou interesse nos produtos/serviços e tem possibilidades de realmente se beneficiar pela solução oferecida pela empresa. O Lead passa a se tornar Prospect quando atende às diretrizes da persona do negócio.
- Lead Scoring: ferramenta de automatização de marketing, melhorando a eficiência dos times de Marketing e Vendas, identificando as oportunidades que o time deve encontrar e quais Leads devem avançar para as próximas abordagens, ou seja, um ranking, definido através de métricas do negócio, que seleciona quais Leads estão mais qualificados a se tornarem clientes.
- Taxa de Conversão: mensura a porcentagem de pessoas que completaram uma ação desejada e avançaram para uma próxima etapa. A taxa de conversão de vendas pode ser utilizada tanto para entender o percentual de Visitantes que se tornaram Leads, quanto para etapas mais avançadas do funil de vendas.
- Return On Investment (ROI): permite mensurar percentualmente e comparar o resultado financeiro obtido. É o percentual das receitas do projeto/empresa menos os custos para executá-lo, em relação aos custos.
- Ticket Médio: é a média de consumo do seu público que está sendo analisado, quanto maior for o ticket médio do negócio, melhor é a performance de vendas.
- Custo por Aquisição (CAC): é o custo que uma organização tem para adquirir novos clientes, incluídos todos os esforços relacionados a vendas e marketing, na tentativa de convencê-lo a adquirir um produto ou serviço.
- Lifetime Value (LTV): quanto de receita, em média, cada cliente traz para o negócio durante seu tempo de relacionamento com a empresa.

# BUSINESS UNDERSTANDING

## 1.3

## DATA

## MINING

## GOALS

Considerando que o objetivo do negócio é categorizar os contatos de clientes (leads) para determinar os possíveis perfis de grupos de clientes ideais (prospects), com o objetivo de otimizar as campanhas de marketing pago. Então o objetivo pode ser atingido com o uso dos dados dos clientes atuais da empresa, dados da jornada de compra destes clientes em conjunto com os dados de resultados de campanhas.

Atualmente obtemos uma base de dados estimada com 34 mil contatos, com relacionamento entre as campanhas que os clientes tiveram interação, antes da compra ou cadastro do aplicativo.

As etapas posteriores serão:

- Avaliar a acurácia de modelos de lead score;
- Definir as métricas para validar a qualidade dos dados;
- Definir a taxa de acurácia do modelo de machine learning.

A seguir é apresentada a estrutura dos dados disponibilizados e meta informações dos mesmos:

COLUNA	DESCRIÇÃO	COLUNA	DESCRIÇÃO
Data	Data em que o Lead se cadastrou	UTM Medium	Canal de contato com a campanha
Hora	Hora em que o Lead se cadastrou	UTM Campaign	Identificador (Nome ou título) da campanha
WeekNum	Semana do ano do cadastro do Lead	UTM Content	Assunto da Campanha
loja_name	Nome da loja	Nome	Nome do responsável pela Loja
instagram_lead	Instagram da loja	City	Cidade onde foi realizado o cadastro
Email	E-mail utilizado para cadastro	Region	Estado onde foi realizado o cadastro
Whatsapp_lead	Número de Whatsapp utilizado para cadastro	Became_client	Flag para caso o Lead se tornou cliente
Fluxo	Fluxo de contato com anúncios da Prepi, onde TD = tráfego direto; e PLG = anúncios para cadastro gratuito	Became_user	Flag para caso o Lead se tornou usuário da Prepi. Todo Cliente é um usuário, automaticamente
Tamanho do negócio	Média de pedidos mensal online que a Loja tem no momento. Respondida pelo próprio usuário	Idade	Idade do responsável pela Loja
UTM source	Origem da campanha	Average ReelsLike	Média de curtidas nos Reels
insta_followers	Quantidade de Seguidores.	Average ReelsComment	Média de comentários nos Reels
instagram_photos	Quantidade de postagens no Feed.	Engajamento Reel	Média geral de engajamento nos Reels
FotoPerfil	URL com a foto do perfil	Bio	Texto da Bio do perfil
Alcance	Média de alcance dos Reels	Bio Link	Link presente na Biografia do Perfil.
Reels View Through	Proporção de pessoas que assistiram de fato o conteúdo	Frequencia Post	Média de postagens no feed por dia do perfil em questão.
StoriesCount	Quantidade Stories ativos no momento de cadastro.	Postou_recente	Postagem no Feed nos últimos 7 dias (considerando a data de cadastro como Lead da Prepi).
Has_Highlight	Se possui conteúdo em destaque em algumas áreas do Instagram.	Postou muito recente	Nos últimos 3 dias (considerando data de cadastro)
UserTag Count	Quantidade de marcações daquele perfil em postagens de outros perfis.	Criação estimada	Estimativa de criação do perfil no Instagram
Is new to instagram	Se criou conta no Instagram recentemente	OverAll Engajamento	Média de engajamento dos seguidores nas postagens
Is Bussiness	Se o perfil é de negócios.	Average Comments	Média de comentários dos seguidores nas postagens
Shoppable Post Count	Quantidade de postagens que possui as ferramentas de Shopping ativas.	Average Likes	Média de curtidas dos seguidores nas postagens.
Public Whatsapp	Número de Whatsapp público no perfil.	Reels Count	Quantidade de medias do tipo Reels
Public Email	Email público no perfil.	Average ReelPlay	Média 'plays' nos Reels
Average ReelsView	Média de visualizações nos Reels	Category Name	Categoria de negócio do Perfil

# BUSINESS UNDERSTANDING

## 1.3 DATA MINING GOALS

COLUNA	DESCRIÇÃO	COLUNA	DESCRIÇÃO
status	Status da assinatura na Prepi.	older_indamplencia_date	Inadimplência mais antiga, ou seja, primeira fatura em débito.
instagram	Instagram cadastrado durante a criação da Loja virtual.	Lead at	Data em que se cadastrou e tornou-se Lead
Whatsapp	Whatsapp cadastrado durante a criação da Loja virtual.	Cidade da loja	Cidade da loja
plano	Plano escolhido	estado da loja	Estado da loja
Payment	Métodos de pagamento escolhido	P1	Resposta da primeira pergunta no primeiro acesso ao aplicativo Prepi
ltv	Valor total da assinatura do plano.	P2	Resposta da segunda pergunta no primeiro acesso ao aplicativo Prepi.
MRR	Valor mensal do plano. LTV / periodicidade	P3	Resposta da terceira pergunta no primeiro acesso ao aplicativo Prepi.
Periodicidade	Período, em meses, da assinatura.	P4	Resposta da quarta pergunta no primeiro acesso ao aplicativo Prepi.
Data_trial	Data em que se tornou testador da Prepi.	P5	Resposta da quinta pergunta no primeiro acesso ao aplicativo Prepi.
Data_client	Data em que, de fato, tornou-se cliente da Prepi. Normalmente 7 dias após ser testador ou no mesmo momento.	Diagnóstico	Com base nos dados e nas respostas, durante o primeiro acesso, é feito um diagnóstico sobre o estado atual da Loja.
data_churn	Data prevista para churn da Prepi. Normalmente, é a data em que a assinatura atual se encerra, sendo necessário a renovação	Diagnostic Date	Data que o diagnóstico foi feito.
valor_indimplencia	Soma de toda a inadimplência	Cost	Custo Prepi para adquirir o Lead/Cliente/Usuário

### Dados Anonimizados:

- loja\_name
- Nome
- instagram\_lead
- Email
- Whatsapp\_lead
- FotoPerfil
- Bio
- Bio Link
- Public Whatsapp
- Public Email
- instagram
- Whatsapp

Não existem outros dados passíveis de anonimização.



# BUSINESS UNDERSTANDING

## 1.4

## PLANO

## DO

## PROJETO

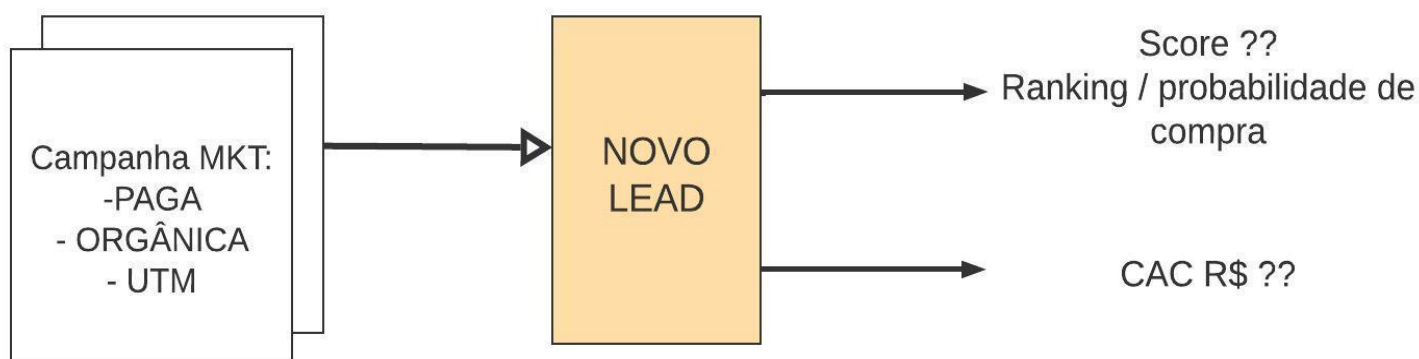
Foi solicitado que seja realizado o estudo de ciência de dados para identificar potenciais futuros clientes para a empresa, através da geração de um Lead Scoring, reduzindo o CAC e encontrando perfil ideal de Lead, que tenha maior probabilidade de tornar-se cliente. Com relação às métricas do projeto, essas serão levantadas levando em consideração o cumprimento dos prazos para a entrega de cada etapa do projeto; a produtividade, avaliando se as responsabilidades de cada membro estão sendo cumpridas com qualidade e da forma correta; e a satisfação do cliente, no caso a Prepi.

Além disso, com base nos dados iniciais disponibilizados pela empresa será realizado:

- Validar os metadados dos datasets disponibilizados pelo Sponsor
- Entender quais as métricas utilizadas para elaboração dos indicadores
- Compreender o histórico de clientes ativos para verificar a possibilidade da criação de critérios que possibilitem a visualização de leads que possam se converter em futuras vendas
- Obter taxa de conversão atual da empresa (ainda necessita de amadurecimento de conceito e indicadores)
- Validar como mensurar a qualidade do cliente

A partir disto, esperamos que seja possível responder alguns questionamentos sobre os objetivos do projeto, sendo eles:

- Como reduzir o custo por aquisição - CPA;
- Como aumentar a taxa de conversão de clientes;





# DATA UNDERSTANDING

---

## 2.1 COLETA DE DADOS INICIAL

A partir do objetivo proposto, foi discutido com os Sponsors que, apenas os dados fornecidos inicialmente não trariam um resultado ao projeto que fosse revertido em retornos financeiros significativos, portanto, além do Lead Scoring, será trabalhado um pipeline de ciência de dados também para redução do CAC (Custo de Aquisição de Clientes), e para isso, a necessidade de alguns dados adicionais, relacionados principalmente a custos de campanhas.

O banco de dados fornecido pela PREPI é resultado da união de 3 diferentes planilhas, unificadas em um único dataset:

- Lista de Leads
- Lista de Clientes e Usuários
- Custo da Campanha de Anúncios Pagos

Os dados estão sendo disponibilizados pelo cliente e atualizados pelos Sponsors no repositório do Google Drive. A princípio, foi disponibilizado apenas duas amostras de dados, contendo 299 e 3231 registros respectivamente para que pudéssemos compreender a natureza e conteúdo dos dados e também suas variáveis, com isso, foi possível obter entendimento dos dados e já pensarmos e possíveis correlações.

Nessa amostra do Dataset, já foi evidenciado linhas duplicadas, dados inconsistentes e valores nulos em algumas colunas, que serão limpos e preparados posteriormente.

# DATA UNDERSTANDING

---

## 2.2                      DESCRIÇÃO                      DOS                      DADOS

Conforme citado acima, a base de dados é resultante da união de 3 diferentes planilhas, sendo: Lista de Leads, Lista de Clientes e Usuários e Custo da Campanhas de Anúncios Pagos.

A princípio, foi disponibilizado para a equipe um subconjunto dos dados, pois a empresa ainda está em processo de unificação das tabelas, mais especificamente as informações de Custo da Campanha de Anúncios Pagos por Cliente (CAC).

A previsão é que o cliente disponibilize um dataset final em formato do tipo “.csv”, contendo, a princípio, 76 colunas e cerca de 6.000 linhas.

- **Lista de Leads:** contém dados referentes aos usuários que representam oportunidades de negócio.
- **Lista de Clientes e Usuários:** contém dados referentes aos usuários que se converteram a clientes na plataforma.
- **Custo da Campanha de Anúncios Pagos:** trata-se de dados referentes ao custo das campanhas para a captação de leads. Estas informações ainda não foram disponibilizadas, pois estão em processo de tratamento de dados e unificação de bases de dados, pela empresa.

# DATA UNDERSTANDING

## 2.2 DESCRIÇÃO DOS DADOS

Abaixo são demonstradas as colunas e suas respectivas descrições:

Campo	Campo Renomeado	Descrição	Base	Anonimizada
Índice	id	Coluna de índice do registro	N/A	
Data	lead_dt_cadastro	Data em que o Lead se cadastrou	Leads	
Hora	lead_hr_cadastro	Hora em que o Lead se cadastrou	Leads	
WeekNum	lead_num_semana_cadastro	Semana do ano do cadastro do Lead	Leads	
loja_name	lead_hash_nome_loja	Nome da Loja	Leads	Sim
Nome	lead_hash_nome	Nome do responsável pela Loja	Leads	Sim
instagram_lead	lead_hash_instagram	Instagram da Loja	Leads	Sim
Email	lead_hash_email	Email para cadastro	Leads	Sim
Whatsapp_lead	lead_hash_whatsapp	Whatsapp para cadastro	Leads	Sim
fluxo	lead_fluxo	Fluxo de contato com anúncios da Prepi.	Leads	
Tamanho do negocio	lead_tamanho_negocio	Tamanho da loja do cliente	Leads	
UTM source	lead_utm_source	Origem da campanha	Leads	
UTM Medium	lead_utm_medium	Canal de contato com a campanha	Leads	
UTM campaign	lead_utm_campaign	Identificador (Nome ou título) da campanha	Leads	
UTM content	lead_utm_content	Assunto da Campanha	Leads	
landing page version	lead_lp_version	Versão da página institucional da campanha	Leads	
referrer	lead_referrer	Origem de acesso	Leads	
city	lead_city	Cidade onde foi realizado o cadastro	Leads	
region	lead_region	Estado onde foi realizado o cadastro	Leads	
became_client	lead_flag_become_trial	Flag para caso o Lead se tornou cliente	Leads	
became_user	lead_flag_become_user	Flag para caso o Lead se tornou usuário da Prepi	Leads	
idade	lead_num_idade	Idade do responsável pela Loja	Leads	
insta_followers	instagram_num_followers	Quantidade de Seguidores	Instagram	
instagram_photos	instagram_num_feed_posts	Quantidade de postagens no Feed	Instagram	
FotoPerfil	instagram_hash_foto_perfil	URL com a foto do perfil	Instagram	Sim
Bio	instagram_hash_bio	Texto da Biografia do Perfil	Instagram	Sim
Bio Link	instagram_hash_bio_link	Link presente na Biografia do Perfil	Instagram	Sim
Frequencia Posts	instagram_num_frequencia_posts	Média de postagens no feed por dia do perfil em questão	Instagram	
Postou recente	instagram_flag_postou_recente	Se realizou alguma postagem no Feed nos últimos 7 dias	Instagram	
Postou muito recente	instagram_flag_postou_muito_recente	Se realizou alguma postagem no Feed nos últimos 3 dias	Instagram	
Criacao estimada	instagram_dt_criacao_estimada	Estimativa de criação do perfil no Instagram	Instagram	
OverAll Engajamento	instagram_num_overall_engajamento	Média de engajamento dos seguidores nas postagens	Instagram	
Average Comments	instagram_num_average_comments	Média de comentários dos seguidores nas postagens	Instagram	
Average Likes	instagram_num_average_likes	Média de curtidas dos seguidores nas postagens	Instagram	
Reels Count	instagram_num_reels_counts	Quantidade de medias do tipo Reels	Instagram	
Average ReelPlay	instagram_num_average_reelplay	Média 'plays' nos Reels	Instagram	
Average ReelsView	instagram_num_average_reelsview	Média de visualizações nos Reels	Instagram	
Average ReelsLike	instagram_num_average_reelslike	Média de curtidas nos Reels	Instagram	
Average ReelsComment	instagram_num_average_reelscomment	Média de comentários nos Reels	Instagram	
Engajamento Reel	instagram_num_engajamento_reel	Média geral de engajamento nos Reels	Instagram	
Alcance	instagram_num_alcance	Proporção entre a quantidade de 'plays' dos Reels para a quantidade de seguidores	Instagram	
Reels View Through	instagram_num_reels_view_through	Proporção de pessoas que deram 'play' nos Reels e assistiram de fato o conteúdo	Instagram	
StoriesCount	instagram_num_stories_count	Quantidade Stories ativos no momento de cadastro	Instagram	
Has Highlight	instagram_flag_has_highlight	Se possui conteúdo em destaque em algumas áreas do Instagram	Instagram	
UserTag Count	instagram_num_usertag_count	Quantidade de marcações daquele perfil em postagens de outros perfis	Instagram	
Is new to instagram	instagram_flag_is_new	Criou a conta no Instagram recentemente	Instagram	
Is Bussiness	instagram_flag_is_bussiness	Se o perfil é de negócios	Instagram	

# DATA UNDERSTANDING

Campo	Campo Renomeada	Descrição	Base	Anonimizada
Shoppable Post Count	instagram_num_shoppable_post_count	Quantidade de postagens que possui as ferramentas de Shopping ativas	Instagram	
Public Whatsapp	instagram_hash_public_whatsapp	Número de Whatsapp público no perfil	Instagram	Sim
Public Email	instagram_hash_public_email	Email público no perfil	Instagram	Sim
Public Zip	instagram_public_zip	CEP público no perfil	Instagram	
Category Name	instagram_category_bussiness	Categoria de negócio do Perfil	Instagram	
status	user_status	Status da assinatura na Prepi	Clientes e Usuários	
instagram	user_hash_instagram	Instagram cadastrado durante a criação da Loja virtual	Clientes e Usuários	Sim
Whatsapp	user_hash_whatsapp	Whatsapp cadastrado durante a criação da Loja virtual	Clientes e Usuários	Sim
plano	user_plano	Nome do Plano escolhido	Clientes e Usuários	
Payment	user_payment	Método de pagamento escolhido	Clientes e Usuários	
Itv	user_val_itv_plano	Valor total da assinatura do plano	Clientes e Usuários	
MRR	user_val_mrr_plano	Valor mensal do plano. LTV / periodicidade	Clientes e Usuários	
periodicidade	user_num_meses_plano	Período de tempo, em meses, da assinatura	Clientes e Usuários	
data_trial	user_dt_trial	Data em que tornou-se testador da Prepi	Clientes e Usuários	
data_cliente	user_dt_cliente	Data em que, de fato, tornou-se cliente da Prepi	Clientes e Usuários	
data_churn	user_dt_churn_prevista	Data prevista para churn da Prepi	Clientes e Usuários	
valor_indimplencia	user_val_inadimplencia	Soma de toda a inadimplência	Clientes e Usuários	
older_indamplencia_date	user_dt_inadimplencia_inicial	Inadimplência mais antiga, ou seja, primeira fatura em débito	Clientes e Usuários	
lead at	user_dt_lead	Data em que se cadastrou e tornou-se Lead	Clientes e Usuários	
cidade da loja	user_cidade	Cidade da Loja	Clientes e Usuários	
estado da loja	user_estado	Estado da Loja	Clientes e Usuários	
P1	user_p1	Resposta da primeira pergunta no primeiro acesso ao aplicativo Prepi	Clientes e Usuários	
P2	user_p2	Resposta da segunda pergunta no primeiro acesso ao aplicativo Prepi	Clientes e Usuários	
P3	user_p3	Resposta da terceira pergunta no primeiro acesso ao aplicativo Prepi	Clientes e Usuários	
P4	user_p4	Resposta da quarta pergunta no primeiro acesso ao aplicativo Prepi	Clientes e Usuários	
P5	user_p5	Resposta da quinta pergunta no primeiro acesso ao aplicativo Prepi	Clientes e Usuários	
Diagnostico	user_diagnostico	Estado atual da Loja	Clientes e Usuários	
Diagnostic Date	user_dt_diagnostico	Data que o diagnóstico foi feito	Clientes e Usuários	
cost	user_val_cac	Custo Prepi para adquirir o Lead/Cliente/Usuário	Clientes e Usuários	

# DATA UNDERSTANDING

## 2.3

## EXPLORAÇÃO

## DOS

## DADOS

Analisando as planilhas enviadas, nota-se que há campos dos quais é possível realizar a correlação de dados e assim, compreender o histórico dos clientes, como por exemplo, a data de captação e conversão do lead, custo de aquisição, tempo de assinatura, etc.

A partir dos dados, é possível criar as seguintes hipóteses:

- A realização de campanhas de tráfego pago tem relação com o aumento da taxa de conversão de clientes?
- O tempo entre a data da assinatura do plano e a data de fidelização do cliente tem relação com a estrutura da campanha que gerou a captação?
- O tempo entre a data da assinatura do plano e a data de fidelização do cliente tem relação com o custo da campanha?
- O plano escolhido pelo cliente se associa à inadimplência?
- O valor do plano se associa ao tempo de permanência do cliente com o serviço?
- Os dados demográficos, tanto dos leads, quanto dos clientes, podem demonstrar uma tendência?

A princípio, o projeto tinha como objetivo apenas criar um modelo que possibilitaria a análise de potenciais clientes a partir da base de leads existentes. Porém, a partir das informações disponibilizadas e reuniões com os Sponsors, foi percebido que além de entregar o objetivo inicial, poderíamos contribuir para uma possível redução de CAC. Sendo assim, houve mudanças na direção do projeto ampliando a entrega final.

Ainda estamos avaliando a variável alvo para implementação do modelo de machine learning.



# DATA UNDERSTANDING

## 2.4

## QUALIDADE

## DOS

## DADOS

Após a disponibilização do dataset final, contando com 76 colunas e 34.131 linhas, verificamos a existência de alguns dados inconsistentes e incoerentes e, em algumas colunas, porcentagem elevada de valores nulos.

Porém, com relação aos nulos, conforme sinalizado pela Prepi, isso ocorre devido a natureza dos dados que é advinda das plataformas utilizadas pela empresa em seu dia a dia, através de sua arquitetura de serviços e de diversas bases de dados.

Alguns dos dados possuem a necessidade de alteração do tipo para que possam ser utilizados nas análises.

Além disso, foi sinalizado pela empresa que alguns clientes que estão presentes na lista de clientes não estão na tabela de leads e, portanto, se apresentam nulos nesta última tabela não sendo possível realizar a exclusão.

Foi também disponibilizado pela empresa o dicionário de dados, auxiliando e facilitando o entendimento das informações pelo grupo. Onde percebemos que a nomenclatura das colunas são autoexplicativas o que irá contribuir positivamente para o projeto.

As principais dimensões de qualidade de dados analisadas foram:

- Acurácia
- Completude
- Unicidade
- Consistência
- Validade
- Temporalidade

Os dados foram atualizados e analisados de acordo com a preparação dos dados utilizando o dataset completo, sendo assim, os dados abaixo foram atualizados considerando as informações em totalidade.

# DATA UNDERSTANDING

## 2.4

## QUALIDADE

## DOS

## DADOS

Erros encontrados na base de dados original:

undefined	771	SP	100	undefined	771
São Paulo	219	MG	44	Sao Paulo	584
Rio de Janeiro	122	RJ	33	Rio de Janeiro	218
Fortaleza	74	PR	30	Minas Gerais	209
Recife	57	PE	29	Pernambuco	134
...		SC	25	Parana	120
Mafra	1	BA	23	Ceara	118
Machacalis	1	RS	20	Bahia	110
Macaé	1	CE	18	Rio Grande do Sul	107
MacapáAmapa	1	GO	17	Santa Catarina	100
Abelardo Luz	1	ES	11	Para	68
Name: lead_city, Length: 705, dtype: int64		PA	10	Goiás	62
user_dt_lead		MA	9	Maranhao	53
02/08/2022 18:40	1	DF	7	Espirito Santo	48
02/11/2022 10:19	1	MT	7	Paraíba	47
03/09/2022 10:47	1	TO	6	Amazonas	47
07:05:26 06/03/2022	2	AL	6	Federal District	46
...		PB	5	Rio Grande do Norte	43
31/07/2021 21:35	1	AM	5	Mato Grosso	42
4/1/2022 18:40:44	2	RN	3	Piauí	26
4/5/2022 8:30:00	1	RO	3	Alagoas	19
4/6/2022 14:46:00	1	Android	3	Rondonia	18
5/6/2022 14:43:28	1	MS	2	Mato Grosso do Sul	18
Name: id, Length: 695, dtype: int64		SE	2	Tocantins	16
instagram_dt_criacao_estimada		AC	2	Sergipe	14
-0536-07-12	1	PI	1	São Paulo	12
-0588-09-07	1	Name: user_estado, dtype: int64		Acre	7
0308-08-18	1			none	2
0613-06-23	1			Roraima	1
1453-10-26	1			Ceará	1
...				Amapa	1
2043-12-30	1			Istanbul	1
2065-07-11	1			Region de Valparaíso	1
2076-01-11	1			Espírito Santo	1
2100-01-22	1			Goiás	1
2237-07-17	1			Name: lead_region, dtype: int64	
Name: id, Length: 852, dtype: int64					
		22	40		
		31	34		
		34	31		
		28	30		
		33	30		
		66.0	1		
		57.0	1		
		São Luís	1		
		undefined	1		
		68	1		
		Name: lead_num_idade, Length: 96, dtype: int64			



# DATA UNDERSTANDING

## 2.4

## QUALIDADE

## DOS

## DADOS

```
R$17,27    30
R$11,46    27
R$10,42    27
R$12,01    27
R$10,81    27
..
R$1,79     1
R$59,48    1
R$98,04    1
R$7,20     1
R$9,03     1
Name: user_val_cac, Length: 896, dtype: int64
```

Não foram encontradas linhas completamente duplicada, no entanto há dados duplicados para colunas de identificadores únicos (IDs):

column_name	total_duplicados	duplicados_not_na	duplicados_not_na_and_blank	duplicados_not_blank
lead_hash_nome_loja	1887	10	10	1887
lead_hash_nome	162	54	54	162
lead_hash_instagram	122	14	14	122
lead_hash_email	41	41	41	41
lead_hash_whatsapp	149	41	41	149
instagram_hash_bio_link	2452	→ 2452	2452	2452
instagram_hash_public_whatsapp	2636	5	5	2636
instagram_hash_public_email	2733	2	2	2733
user_hash_instagram	2366	18	18	2366
user_hash_whatsapp	2311	22	22	2311

# DATA UNDERSTANDING

## 2.4

## QUALIDADE

## DOS

## DADOS

Distribuição de dados com valores sem preenchimento (nulos):

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3231 entries, 0 to 3230
Data columns (total 76 columns):
```

#	Column	Non-Null Count	Dtype	#	Column	Non-Null Count	Dtype	#	Column	Non-Null Count	Dtype
0	Unnamed: 0	3231 non-null	object	31	Average Comments	1417 non-null	object	67	P1	641 non-null	object
1	Data	3100 non-null	object	32	Average Likes	1417 non-null	object	68	P2	642 non-null	object
2	Hora	3088 non-null	object	33	Reels Count	1412 non-null	object	69	P3	642 non-null	object
3	Weeknum	3231 non-null	object	34	Average ReelPlay	1412 non-null	object	70	P4	642 non-null	object
4	loja_name	3231 non-null	object	35	Average ReelsView	1412 non-null	object	71	P5	1 non-null	object
5	Nome	3231 non-null	object	36	Average ReelsLike	1412 non-null	object	72	Diagnostic	641 non-null	object
6	instagram_lead	3231 non-null	object	37	Average ReelsComment	1412 non-null	object	73	Diagnostic Date	504 non-null	object
7	Email	3231 non-null	object	38	Engajamento Reel	1412 non-null	object	74	cost	2864 non-null	object
8	Whatsapp_lead	3231 non-null	object	39	Alcance	1373 non-null	object	75	idade	880 non-null	object
9	fluxo	3071 non-null	object	40	Reels View Through	1412 non-null	object	dtypes: object(76) memory usage: 1.9+ MB			
10	Tamanho do negocio	3100 non-null	object	41	StoriesCount	838 non-null	object				
11	UTM source	2984 non-null	object	42	Has Highlight	1262 non-null	object				
12	UTM Medium	2643 non-null	object	43	UserTag Count	1262 non-null	object				
13	UTM campaign	2643 non-null	object	44	Is new to instagram	1772 non-null	object				
14	UTM content	2982 non-null	object	45	Is Bussiness	1772 non-null	object				
15	landing page version	3089 non-null	object	46	Shoppable Post Count	999 non-null	object				
16	referrer	3100 non-null	object	47	Public Whatsapp	3231 non-null	object				
17	city	3100 non-null	object	48	Public Email	3231 non-null	object				
18	region	3066 non-null	object	49	Public Zip	152 non-null	object				
19	became_client	3231 non-null	object	50	Category Name	1018 non-null	object				
20	became_user	3231 non-null	object	51	status	935 non-null	object				
21	insta_followers	1417 non-null	object	52	instagram	3231 non-null	object				
22	instagram_photos	1417 non-null	object	53	Whatsapp	3231 non-null	object				
23	FotoPerfil	3231 non-null	object	54	plano	470 non-null	object				
24	Bio	3231 non-null	object	55	Payment	466 non-null	object				
25	Bio Link	3231 non-null	object	56	ltv	927 non-null	object				
26	Frequencia Posts	1417 non-null	object	57	MRR	927 non-null	object				
27	Postou recente	1417 non-null	object	58	periodicidade	927 non-null	object				
28	Postou muito recente	1417 non-null	object	59	data_trial	471 non-null	object				
29	Criacao estimada	1396 non-null	object	60	data_cliente	255 non-null	object				
30	OverAll Engajamento	1417 non-null	object	61	data_churn	232 non-null	object				
				62	valor_indimplencia	411 non-null	object				
				63	older_indimplencia_date	291 non-null	object				
				64	lead at	840 non-null	object				
				65	cidade da loja	418 non-null	object				
				66	estado da loja	421 non-null	object				
				67	P1	641 non-null	object				

# DATA UNDERSTANDING

## 2.4

## QUALIDADE

## DOS

## DADOS

Percentual de dados com valores sem preenchimento (nulos):

	coluna	total_null	percentual		coluna	total_null	percentual		coluna	total_null	percentual
71	user_p5	3230	0.9997	29	instagram_dt_criacao_estimada	1841	0.5698	42	instagram_flag_has_highlight	0	0.0000
49	instagram_public_zip	3079	0.9530	40	instagram_num_reels_view_through	1819	0.5630	28	instagram_flag_postou_muito_recente	0	0.0000
61	user_dt_churn_prevista	2999	0.9282	37	instagram_num_avergae_reelscomment	1819	0.5630	27	instagram_flag_postou_recente	0	0.0000
60	user_dt_cliente	2976	0.9211	38	instagram_num_engajamento_reel	1819	0.5630	25	instagram_hash_bio_link	0	0.0000
63	user_dt_inadimplencia_inicial	2940	0.9099	35	instagram_num_avergae_reelsview	1819	0.5630	20	lead_flag_became_user	0	0.0000
62	user_val_inadimplencia	2820	0.8728	34	instagram_num_avergae_reelplay	1819	0.5630	19	lead_flag_became_client	0	0.0000
65	user_cidade	2813	0.8706	33	instagram_num_reels_counts	1819	0.5630	7	lead_hash_email	0	0.0000
66	user_estado	2810	0.8697	36	instagram_num_avergae_reelslike	1819	0.5630	3	lead_num_semana_cadastro	0	0.0000
55	user_payment	2765	0.8558	32	instagram_num_aaverage_likes	1814	0.5614	76	lead_faixa_idade	0	0.0000
54	user_plano	2761	0.8545	23	instagram_hash_foto_perfil	1814	0.5614				
59	user_dt_trial	2760	0.8542	22	instagram_num_feed_posts	1814	0.5614				
48	instagram_hash_public_email	2732	0.8456	21	instagram_num_followers	1814	0.5614				
73	user_dt_diagnostico	2727	0.8440	26	instagram_num_frequencia_posts	1814	0.5614				
47	instagram_hash_public_whatsapp	2632	0.8146	30	instagram_num_overall_engajamento	1814	0.5614				
67	user_p1	2590	0.8016	31	instagram_num_aaverage_comments	1814	0.5614				
72	user_diagnostico	2590	0.8016	13	lead_utm_campaign	588	0.1820				
70	user_p4	2589	0.8013	12	lead_utm_medium	588	0.1820				
69	user_p3	2589	0.8013	74	user_val_cac	374	0.1158				
68	user_p2	2589	0.8013	14	lead_utm_content	249	0.0771				
64	user_dt_lead	2510	0.7768	11	lead_utm_source	247	0.0764				
41	instagram_num_stories_count	2393	0.7406	18	lead_region	165	0.0511				
75	lead_num_idade	2353	0.7283	9	lead_fluxo	160	0.0495				
52	user_hash_instagram	2349	0.7270	2	lead_hr_cadastro	143	0.0443				
57	user_val_mrr_plano	2304	0.7131	15	lead_lp_version	142	0.0439				
56	user_val_ltv_plano	2304	0.7131	1	lead_dt_cadastro	131	0.0405				
58	user_num_meses_plano	2304	0.7131	10	lead_tamanho_negocio	131	0.0405				
51	user_status	2296	0.7106	17	lead_city	131	0.0405				
53	user_hash_whatsapp	2290	0.7088	16	lead_referrer	131	0.0405				
46	instagram_num_shoppable_post_count	2232	0.6908	8	lead_hash_whatsapp	109	0.0337				
50	instagram_category_bussiness	2213	0.6849	6	lead_hash_instagram	109	0.0337				
24	instagram_hash_bio	1977	0.6119	5	lead_hash_nome	109	0.0337				
43	instagram_num_usertag_count	1969	0.6094	0	id	0	0.0000				
4	lead_hash_nome_loja	1878	0.5812	45	instagram_flag_is_bussiness	0	0.0000				
39	instagram_num_alcance	1858	0.5751	44	instagram_flag_is_new	0	0.0000				

**Observação:** Identificamos que algumas colunas com dados criptografados continham o valor “nan”, esses foram substituídos por null.



# DATA UNDERSTANDING

## 2.4

## QUALIDADE

## DOS

## DADOS

Análise estatística das variáveis numéricas:

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	3231.0	15819.813081	9336.617089	2.000000	7550.500000	15632.000000	24013.500000	32296.000000
WeekNum	3231.0	28.626431	15.131560	-1.000000	18.000000	31.000000	39.000000	53.000000
insta_followers	1417.0	3302.779817	17891.229397	0.000000	104.000000	607.000000	1873.000000	354804.000000
instagram_photos	1417.0	288.263938	650.661700	0.000000	11.000000	64.000000	260.000000	7915.000000
Frequencia Posts	1417.0	50.608745	339.424153	-278.335571	0.028130	0.254703	1.361470	6027.906977
OverAll Engajamento	1417.0	4.938945	26.598401	-1.000000	0.023100	0.189802	1.202532	630.083333
Average Comments	1417.0	2.479543	15.054147	0.000000	0.000000	0.166667	1.000000	342.583333
Average Likes	1417.0	19.907343	91.899460	-1.000000	1.333333	5.500000	14.666667	2646.916667
Reels Count	1412.0	5.561615	5.353529	0.000000	0.000000	4.000000	12.000000	21.000000
Average ReelPlay	1412.0	692.421438	1688.365464	0.000000	0.000000	211.696970	770.625000	37435.363640
Average ReelsView	1412.0	305.738161	892.698607	0.000000	0.000000	75.208333	294.770833	22459.909090
Average ReelsLike	1412.0	25.582978	89.075829	0.000000	0.000000	7.000000	25.666667	2687.181818
Average ReelsComment	1412.0	1.792681	6.540443	0.000000	0.000000	0.083333	1.250000	122.090909
Engajamento Reel	1412.0	27.375659	93.437986	0.000000	0.000000	7.083333	28.000000	2809.272727
Alcance	1373.0	0.950629	3.621111	0.000000	0.000000	0.214823	0.704433	84.833333
Reels View Through	1412.0	0.276732	0.235503	0.000000	0.000000	0.330475	0.440507	3.000000
StoriesCount	838.0	3.100239	6.644010	0.000000	0.000000	0.000000	3.000000	56.000000
UserTag Count	1262.0	32.880349	314.854652	0.000000	0.000000	1.000000	9.000000	10225.000000
Shoppable Post Count	999.0	2.504505	17.616879	0.000000	0.000000	0.000000	0.000000	278.000000
MRR	927.0	29.324347	35.964497	0.000000	0.000000	19.000000	59.000000	279.900000
periodicidade	927.0	3.783172	4.570000	1.000000	1.000000	1.000000	3.000000	12.000000
valor_indimplencia	411.0	51.847251	80.083384	0.000000	0.000000	29.240000	80.000000	809.100000
cost	2857.0	24.312149	54.542067	0.000000	8.510000	12.750000	18.610000	812.120000

# DATA UNDERSTANDING

## 2.4

## QUALIDADE

## DOS

## DADOS

Análise estatística da variável alvo “**user\_val\_cac**”:

```
count    30707.0000
mean      27.3782
std       64.6165
min        0.0000
25%       8.7100
50%      13.5500
75%      18.6700
max      844.1400
Name: user_val_cac, dtype: float64

sum      840701.3300
max      844.1400
min        0.0000
mean      27.3782
median    13.5500
Name: user_val_cac, dtype: float64
```

Do total de 34.115 registros, somente 3.408 (9,9897%) não possuem dados na coluna a variável alvo.

Análise da variável alvo por tipo do status do usuário:

	sum	max	min	mean	median
user_status					
active	107620.82	844.14	0.00	189.1403	145.870
churned	140010.49	844.14	31.07	184.4670	135.690
debtor	6708.00	697.03	35.85	258.0000	193.350
debtor_trial	3927.71	604.23	118.31	261.8473	293.590
deleted	15495.26	706.02	54.01	218.2431	174.310
suspended	40275.76	844.14	32.21	200.3769	136.670
trial_churned	134160.28	812.12	19.11	214.6564	159.860
user	31572.54	125.13	0.00	7.1238	6.305
waiting_payment	236.62	118.31	118.31	118.3100	118.310

Análise da variável alvo para usuários sem status (apenas leads):

```
sum      360693.8500
max      196.5700
min        0.0000
mean      15.0245
median    14.0800
Name: user_val_cac, dtype: float64
```

# DATA UNDERSTANDING

## 2.4

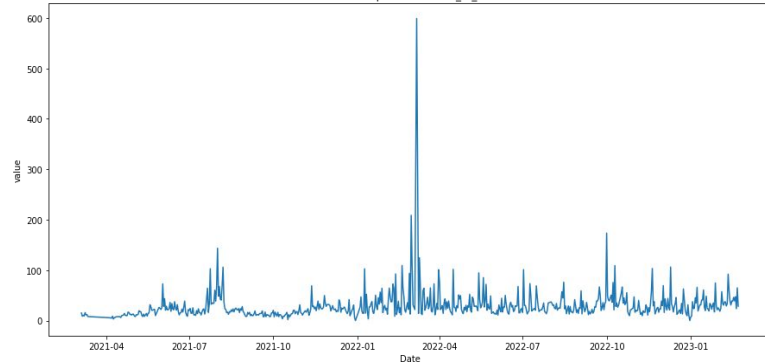
## QUALIDADE

## DOS

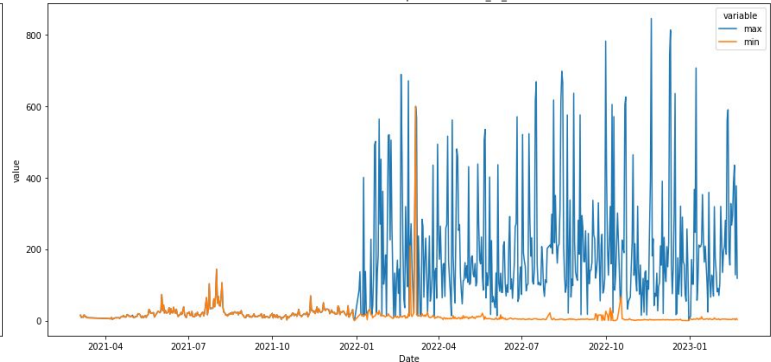
## DADOS

Análise histórica da variável alvo **"user\_val\_cac"**:

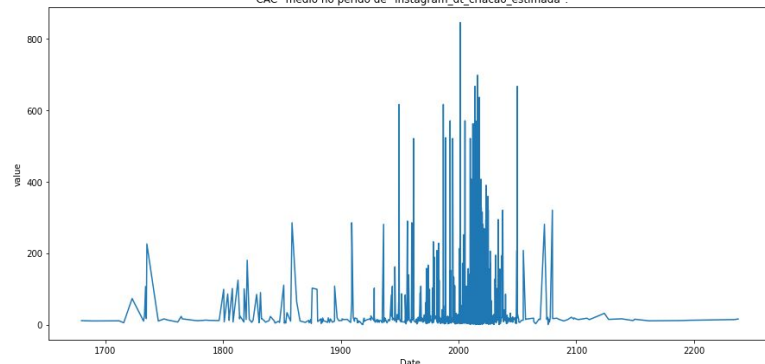
"CAC" médio no período de "lead\_dt\_cadastro".



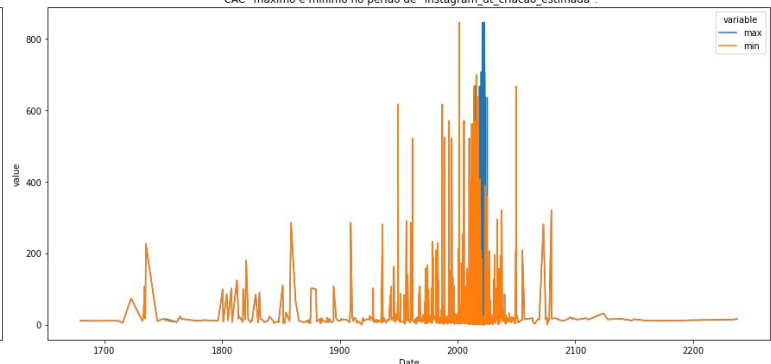
"CAC" máximo e mínimo no período de "lead\_dt\_cadastro".



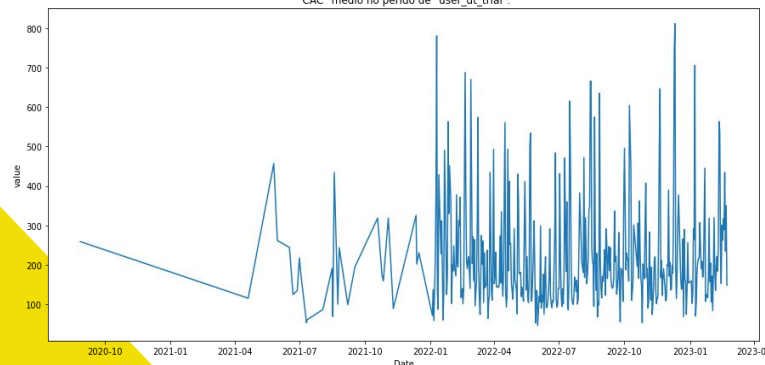
"CAC" médio no período de "instagram\_dt\_criacao\_estimada".



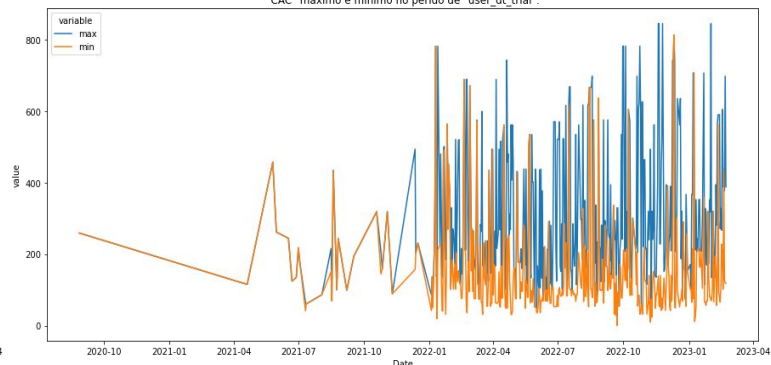
"CAC" máximo e mínimo no período de "instagram\_dt\_criacao\_estimada".



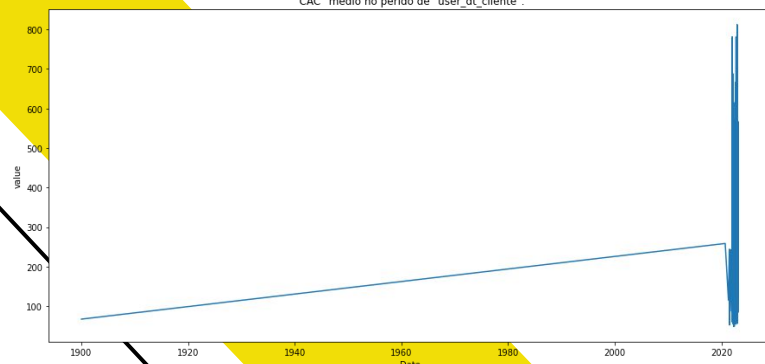
"CAC" médio no período de "user\_dt\_trial".



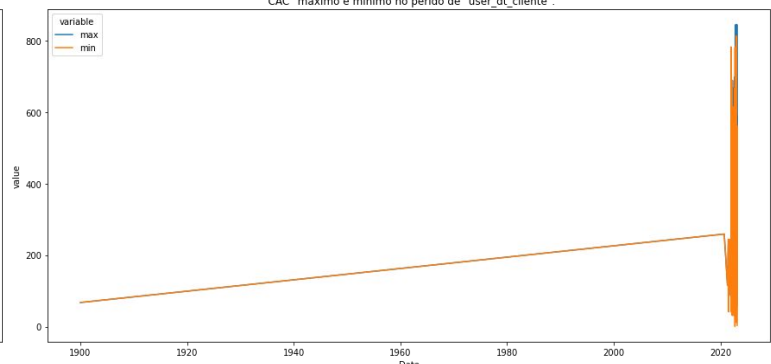
"CAC" máximo e mínimo no período de "user\_dt\_trial".



"CAC" médio no período de "user\_dt\_cliente".



"CAC" máximo e mínimo no período de "user\_dt\_cliente".





# DATA UNDERSTANDING

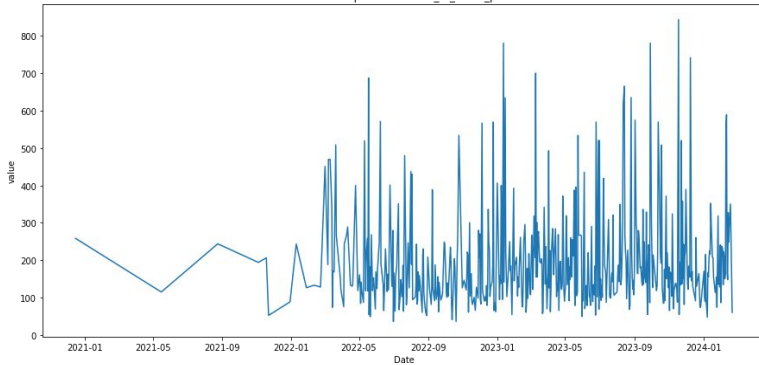
## 2.4

## QUALIDADE

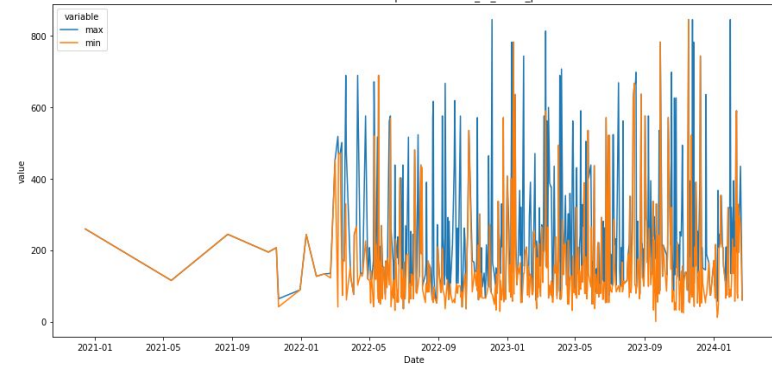
## DOS

## DADOS

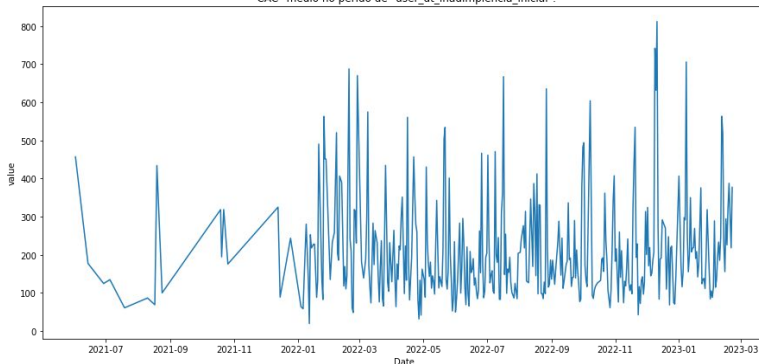
"CAC" médio no período de "user\_dt\_churn\_previsa".



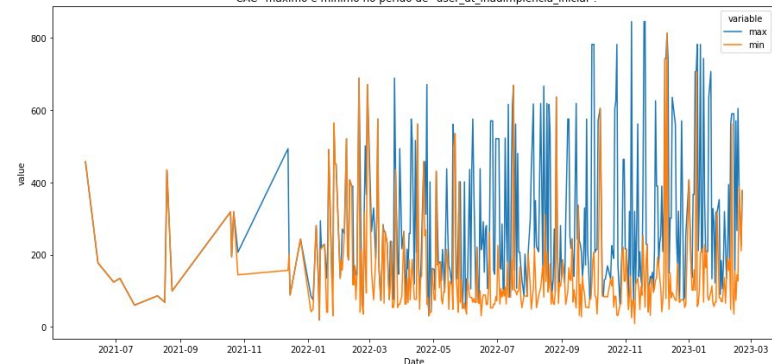
"CAC" máximo e mínimo no período de "user\_dt\_churn\_previsa".



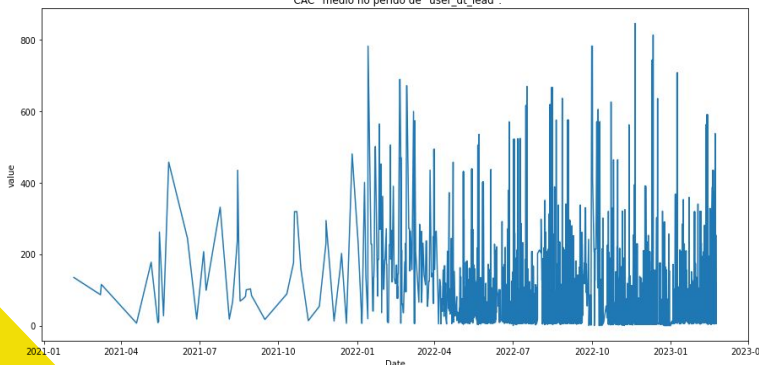
"CAC" médio no período de "user\_dt\_inadimplencia\_inicial".



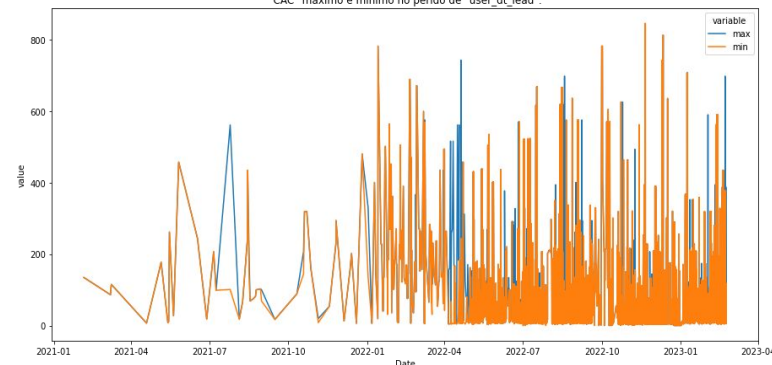
"CAC" máximo e mínimo no período de "user\_dt\_inadimplencia\_inicial".



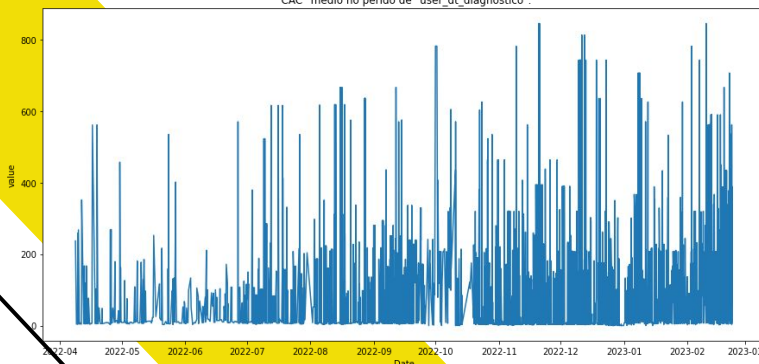
"CAC" médio no período de "user\_dt\_lead".



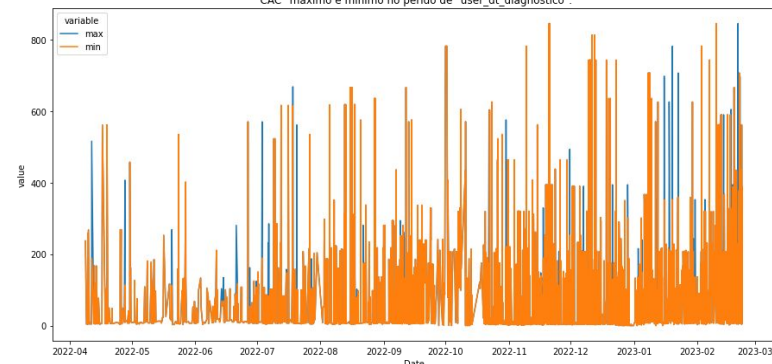
"CAC" máximo e mínimo no período de "user\_dt\_lead".



"CAC" médio no período de "user\_dt\_diagnostico".



"CAC" máximo e mínimo no período de "user\_dt\_diagnostico".





# DATA PREPARATION

## 3.1 Selecionar os dados

A princípio, foi realizada a análise inicial dos dados com base no sample enviado pela empresa e, após a disponibilização do dataset completo, foi possível o mapeamento dos dados que seriam utilizados. Por ora, todas as colunas estão sendo mantidas, ainda que com percentual alto de nulos, visto que, como sinalizado pelo Sponsor, isso se dá ao fato da unificação das 3 tabelas, onde naturalmente, as colunas são diferentes.

Observa-se que há algumas variáveis de maior relevância para o projeto sendo, entre elas:

- ***user\_diagnostico***: descreve estado atual da loja dos clientes;
- ***user\_val\_cac***: custo que a Prepi tem ao adquirir cada Lead/Cliente/Usuário;
- ***user\_val\_ltv\_plano***: valor total da assinatura do plano;
- ***instagram\_category\_business***: ramo no qual o lojista possui seu negócio;
- ***lead\_fluxo***: meio pelo qual o lojista teve contato com os anúncios da Prepi.
- ***lead\_lp\_version***: versão da página institucional da campanha (cada versão da landing page é alterado algo: vídeo, imagem, copywriting, etc);
- ***lead\_referrer***: origem de acesso, ou seja, página de origem em que o lead acessou e abriu o anúncio da campanha da Prepi;
- ***lead\_region***: estado onde foi realizado o cadastro;
- ***lead\_tamanho\_negocio***: tamanho do comércio do lead;
- ***user\_estado***: estado em que a Loja está localizada;
- ***user\_p1***: resposta da primeira pergunta respondida pelo cliente/usuário no cadastro do aplicativo (quando o lead é convertido para cliente/usuário);
- ***user\_p2***: resposta da segunda pergunta respondida pelo cliente/usuário no cadastro do aplicativo (quando o lead é convertido para cliente/usuário);
- ***user\_p3***: resposta da terceira pergunta respondida pelo cliente/usuário no cadastro do aplicativo (quando o lead é convertido para cliente/usuário);

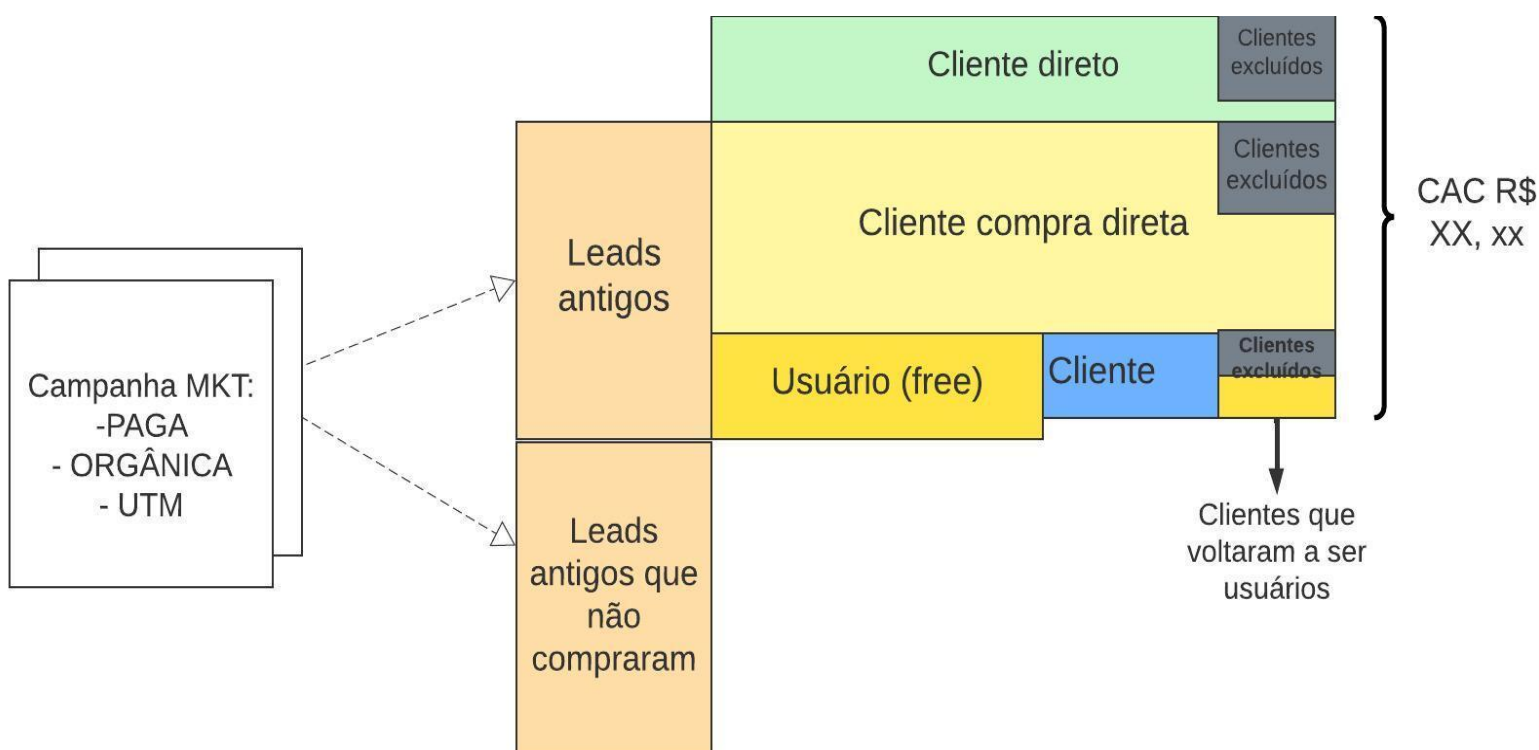
# DATA PREPARATION

## 3.1 Selecionar os dados

- **user\_p4**: resposta da quarta pergunta respondida pelo cliente/usuário no cadastro do aplicativo (quando o lead é convertido para cliente/usuário);

Na nossa análise inicial, conseguimos identificar algumas regras de negócio que categorizam, de maneira geral, os registros do dataset disponibilizado, reforçando a etapa de Data Understanding. Onde existem registros dos seguintes grupos:

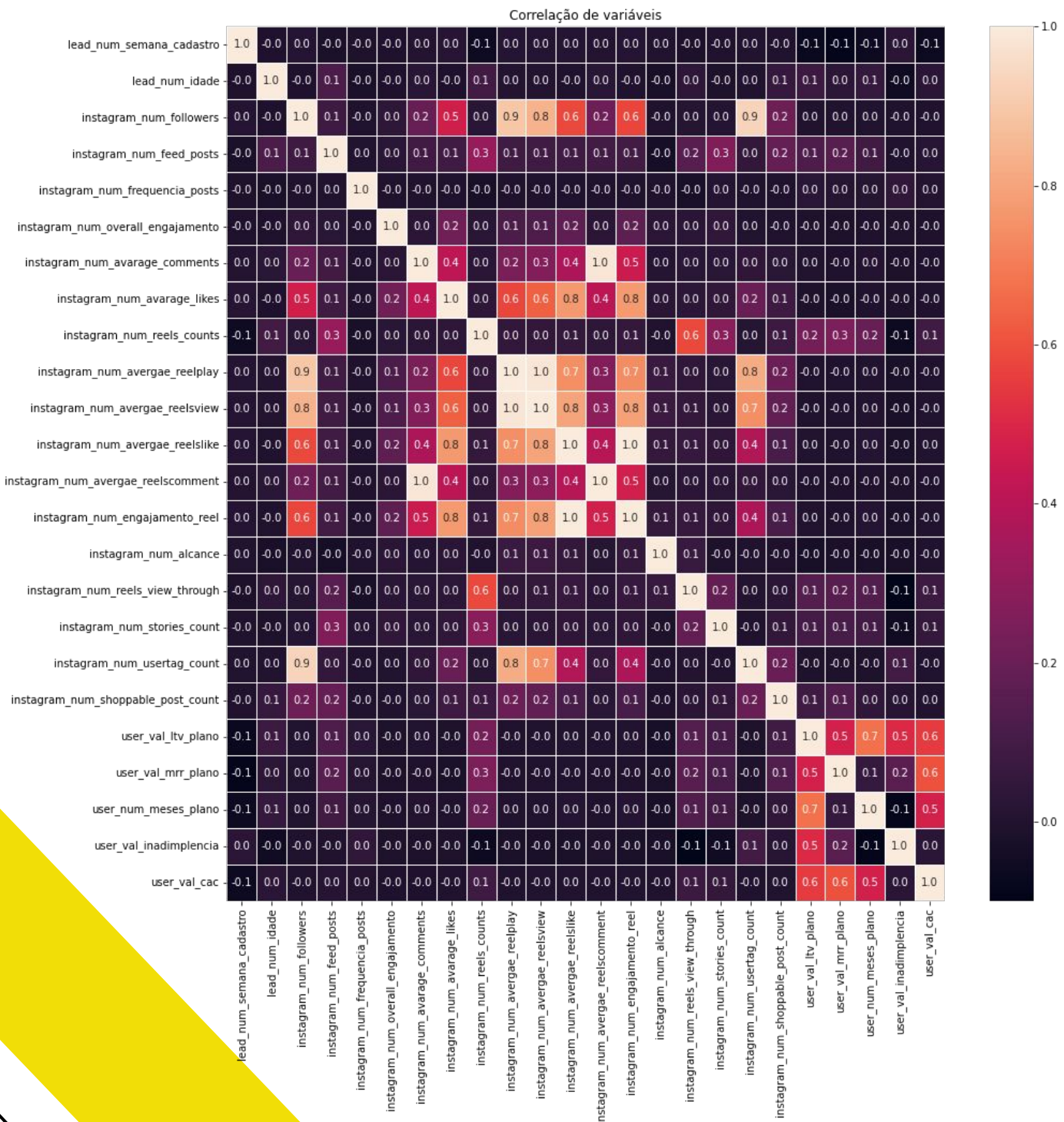
- **leads**: contatos de clientes;
- **usuários**: clientes que baixaram a versão gratuita do app;
- **clientes**: clientes que compraram algum plano do app;
- **clientes cancelados (churn)**: clientes que cancelaram e deixaram de usar o app;
- **usuários cancelados (churn)**: clientes que cancelaram o plano e voltaram a usar a versão gratuita.



# DATA PREPARATION

## 3.1 Selecionar os dados

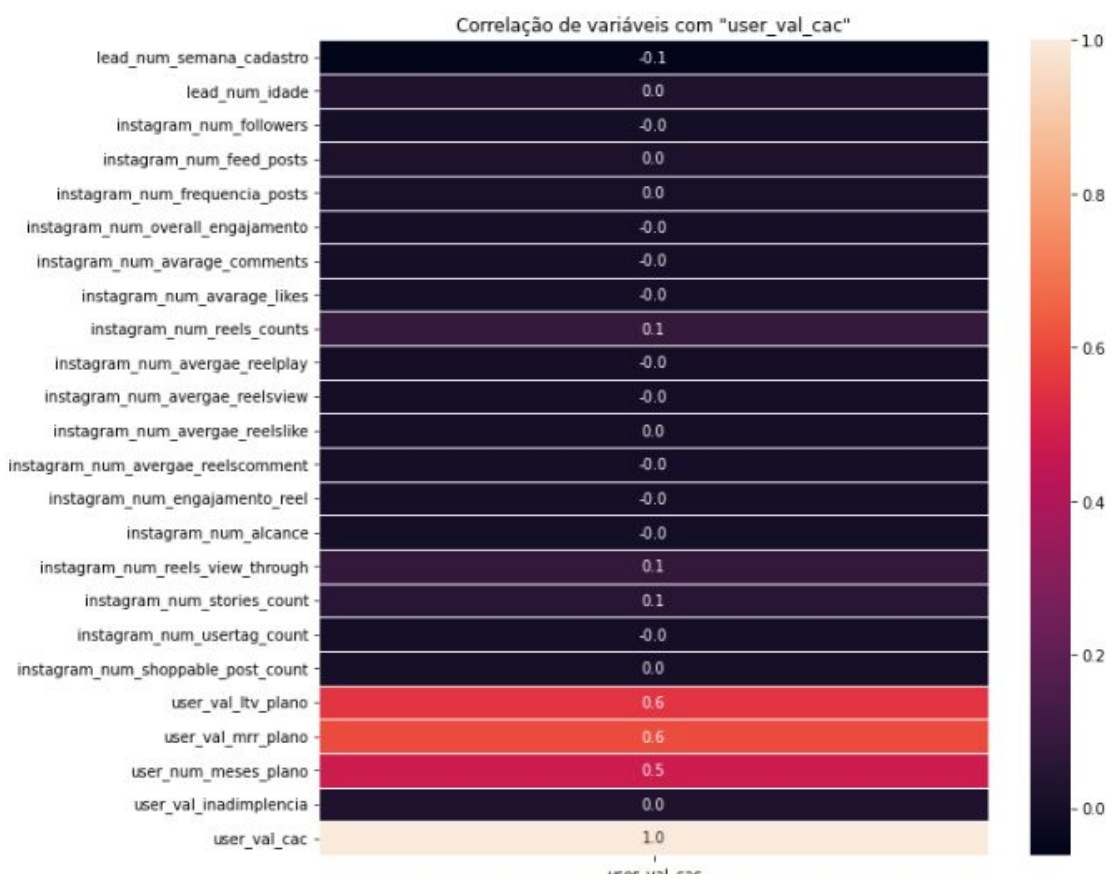
Correlação de variáveis:



# DATA PREPARATION

## 3.1 Selecionar os dados

Correlação de variáveis com a variável alvo:



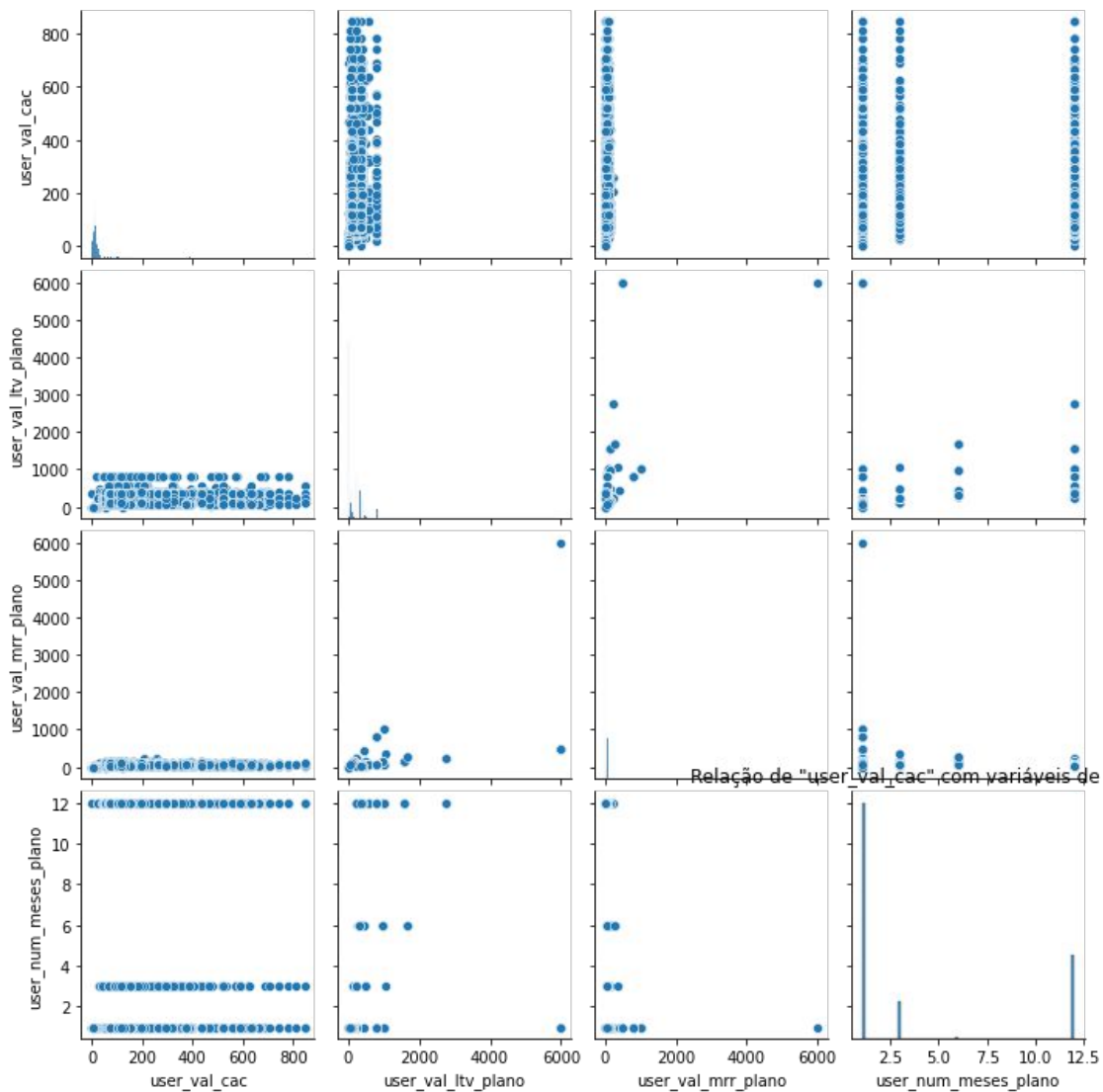
As variáveis que estão dentro da escala de 1 a 0,7 (+ ou -) e na tonalidade laranja no caso das positivas, e roxo escuro no caso das negativas, possuem uma forte correlação; já as variáveis que estão entre 0,7 a 0,5 (+ ou -) e na tonalidade avermelhada no caso das positivas e roxo médio no caso das negativas, possuem correlação moderada; as variáveis que possuem escala de 0,5 a 0,25 e (+ ou -) possuem baixa; e por fim as variáveis com coeficiente próximo a 0 (+ ou -) e com tonalidade vermelha não possuem correlação.

A correlação é uma medida que indica a interdependência entre variáveis. Importante lembrar que correlação não implica necessariamente uma relação causal.

# DATA PREPARATION

## 3.1 Selecionar os dados

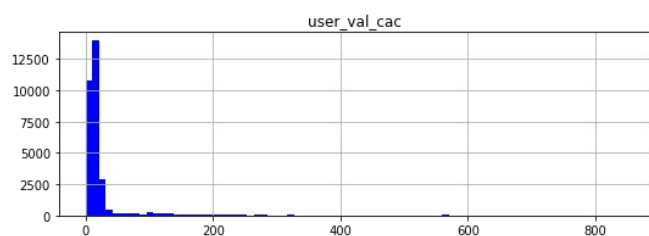
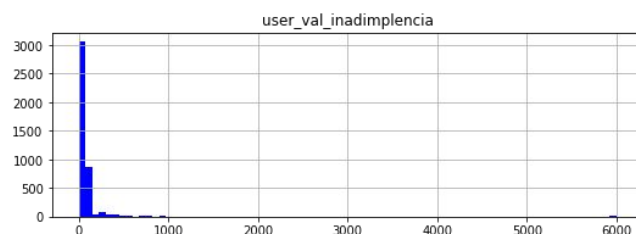
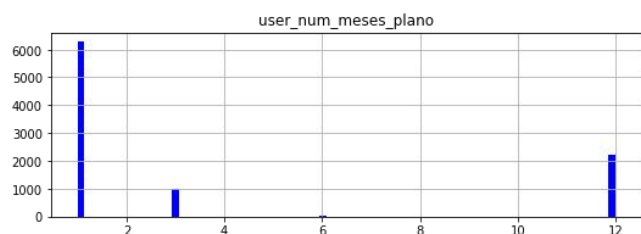
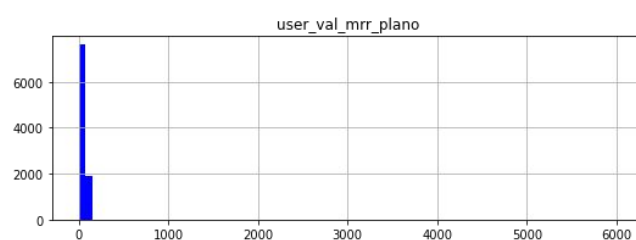
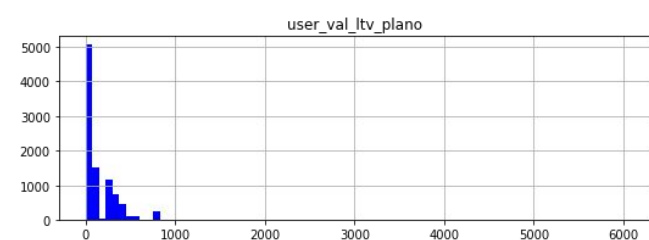
Relação de "**user\_val\_cac**" com variáveis de alta correlação:



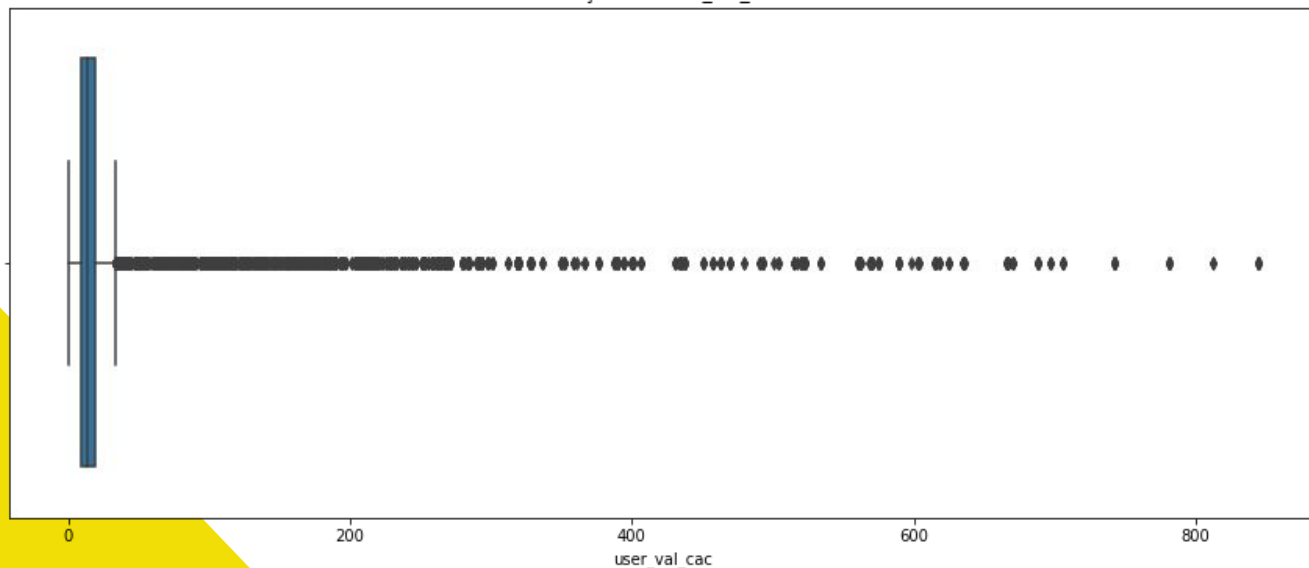
# DATA PREPARATION

## 3.1 Selecionar os dados

Análise de distribuição dos dados:



Variação de "user\_val\_cac"





# DATA PREPARATION

---

## 3.2 Limpando os dados

Ambos os datasets enviados pela equipe Prepi se mostram favoráveis à realização do projeto pois os dados fazem sentido do ponto de vista temporal considerando que os dados estão disponíveis e são facilmente manipuláveis.

Nota-se que a maioria das colunas são compostas por dados do tipo *string*, portanto, algumas delas serão transformadas em variáveis categóricas para facilitar o processo de análise e aplicação do modelo.

Sobre os dados nulos, estes serão mantidos na maioria das colunas, transformando-as em uma categoria, visto que o preenchimento desses valores por algum cálculo estatístico, acarretaria um viés e análise tendenciosa. Como por exemplo, nas colunas *lead\_lp\_version*, *lead\_referrer*, *lead\_region*, *lead\_tamanho\_negocio* e *user\_estado*, pode-se observar um número elevado de dados nulos. No entanto, parte disso ocorre devido ao fato do dataset ser resultado da utilização de três tabelas de origens distintas que se unificam através do campo *lead\_hash\_email*, fazendo com que tenham linhas não-correspondentes entre si, gerando um número acentuado de dados nulos.

Além disso, foi percebido que alguns dos usuários e clientes listados no dataset se tornaram clientes ou usuários por tráfego direto, ou seja, não foram captados por campanhas de anúncios realizados pela Prepi em algum momento e também dados de testes de campanhas. Porém, ainda assim, cabe ressaltar que esta quantidade, mesmo levando as variáveis citadas em consideração, é elevada.

A hipótese de exclusão destas linhas foi descartada neste momento, visto que, ao excluir linhas em branco, estaríamos comprometendo os valores presentes em outras colunas que continham dados. Com relação a substituição destes valores por média ou mediana, ou ainda por algum cálculo estatístico, também foi descartada por hora, visto que por conta do alto volume, também acarretaria numa análise tendenciosa.

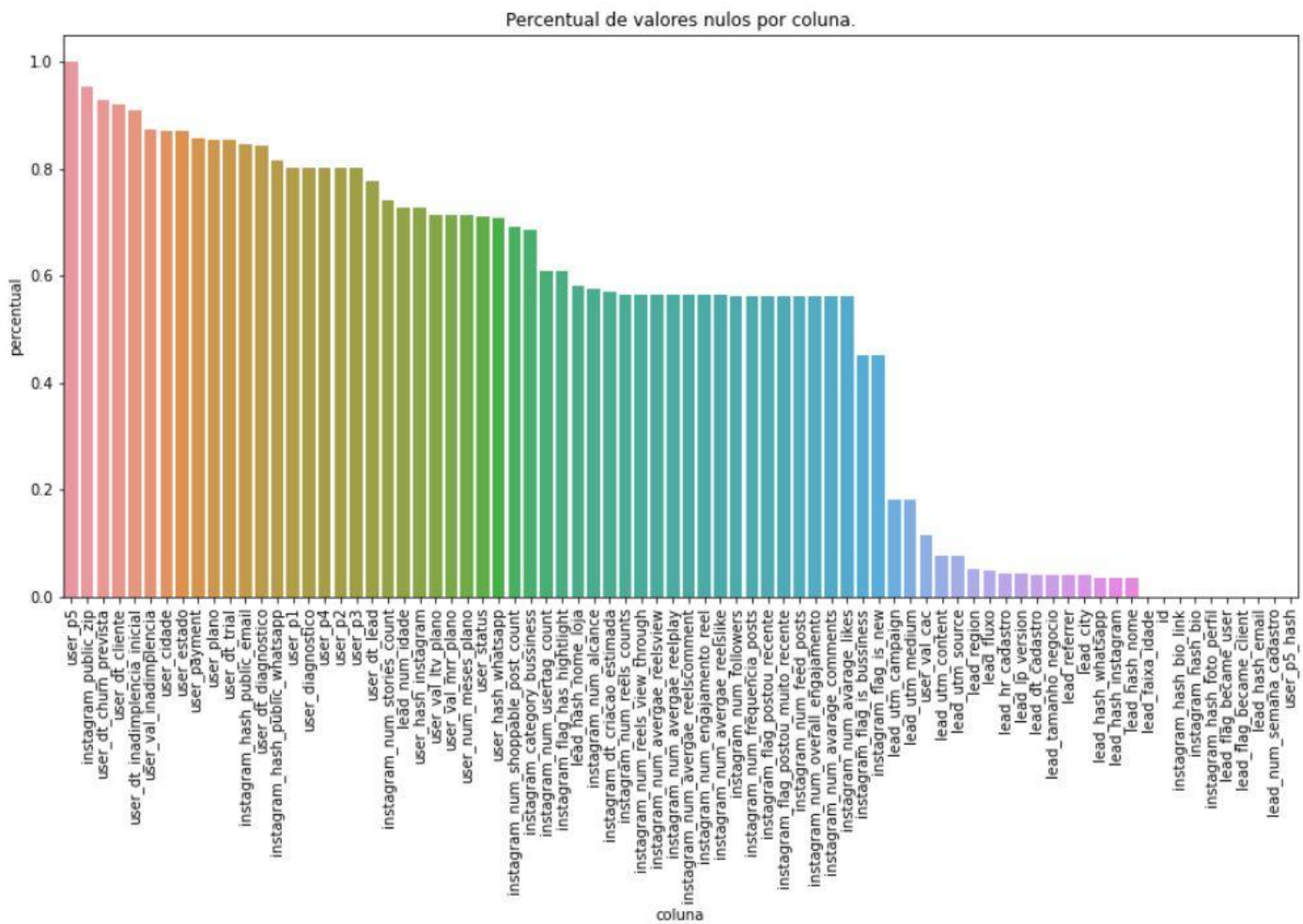
Optou-se então, nas respectivas colunas, manter os valores nulos, transformando-os em variáveis categóricas.



# DATA PREPARATION

## 3.2 Limpando os dados

Percentual de valores nulos por coluna:



# DATA PREPARATION

## 3.2 Limpando os dados

Total de valores duplicados nas colunas identificadoras (hash):

	column_name	total_duplicados	duplicados_not_na	duplicados_not_na_and_blank	duplicados_not_blank
0	lead_hash_nome_loja	20303	712	712	20303
1	lead_hash_nome	4041	2740	2740	4041
2	lead_hash_instagram	2403	1102	1102	2403
3	lead_hash_email	2402	2402	2402	2402
4	lead_hash_whatsapp	3832	2531	2531	3832
5	instagram_hash_bio_link	25872	25872	25872	25872
6	instagram_hash_public_whatsapp	27914	224	224	27914
7	instagram_hash_public_email	28870	168	168	28870
8	user_hash_instagram	26261	1143	1143	26261
9	user_hash_whatsapp	25857	1315	1315	25857

**Obs:** não foram encontrados registros com linhas totalmente duplicadas.

# DATA PREPARATION

## 3.3 Construindo os dados

Após uma análise detalhada na base de dados, foram identificados diversos tipos de dados para as colunas existentes que precisam de tratamento para transformação em tipos de dados mais adequados para sua função na análise, resultando no exemplo de dados tratados como abaixo:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3231 entries, 0 to 3230
Data columns (total 76 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   id                                    3231 non-null   int64
1   lead_dt_cadastro                     3100 non-null   datetime64[ns]
2   lead_hr_cadastro                     3088 non-null   object
3   lead_num_semana_cadastro             3231 non-null   int64
4   lead_hash_nome_loja                  1353 non-null   object
5   lead_hash_nome                       3122 non-null   object
6   lead_hash_instagram                  3122 non-null   object
7   lead_hash_email                      3231 non-null   object
8   lead_hash_whatsapp                   3122 non-null   object
9   lead_fluxo                           3071 non-null   category
10  lead_tamanho_negocio                 3100 non-null   category
11  lead_utm_source                      2984 non-null   category
12  lead_utm_medium                      2643 non-null   category
13  lead_utm_campaign                    2643 non-null   category
14  lead_utm_content                     2982 non-null   category
15  lead_lp_version                      3089 non-null   category
16  lead_referrer                        3100 non-null   category
17  lead_city                            3100 non-null   category
18  lead_region                          3066 non-null   category
19  lead_flag_became_client              3231 non-null   int64
20  lead_flag_became_user                3231 non-null   int64
21  instagram_num_followers              1417 non-null   float64
```

A partir da correção do formato dos dados, foi realizado o tratamento de dados que poderiam ser melhor categorizados, visto que grande parte dos dados se encontram como variáveis categóricas. Um dos exemplos de transformação utilizado pode ser observado na coluna “lead\_faixa\_idade”, em que os dados foram transformados de dados numéricos ordinais em categóricos, através da transformação da idade para uma faixa de idade em que os leads se encontram.

# DATA PREPARATION

---

## 3.3 Construindo os dados

Além disso, devido a quantidade elevada de variáveis categóricas, muitas das “features” precisarão ser modificadas para a análise do dataset pelo modelo de machine learning previsto nas etapas subsequentes do estudo, ou seja, estruturadas em formatos numéricos para que sejam analisadas. Para isso, existe a possibilidade de transformação de string para numérico através de técnicas de encoding, como alguns exemplos possíveis de conversão dentro do dataset:

- **instagram\_category\_bussiness** - Frequency encoding
- **lead\_city** - Frequency encoding
- **lead\_fluxo** - Binary encoding ou one-hot-encoding
- **lead\_ip\_version** - Frequency encoding
- **lead\_referrer** - Frequency encoding
- **lead\_region** - Frequency encoding
- **lead\_tamanho\_negocio** - Ordinal encoding

# DATA PREPARATION

---

## 3.4 Integrando os dados

Toda a base de dados necessária já foi disponibilizada integrada pela Prepi, sendo assim, facilitou a análise por parte do grupo onde conseguimos já realizar a preparação das informações de forma centralizada.

Neste ponto, percebemos que as informações se conectam a partir da coluna *lead\_hash\_email*. Ou seja, é possível identificar o histórico de atividade do usuário/cliente a partir das colunas *lead\_fluxo* e *user\_status* buscando pelo e-mail de cadastro.

Realizamos também a criação de uma nova coluna que padroniza a idade do cliente que permitirá uma possível criação de tendência, mapeando a faixa etária do cliente para entender melhor o público o qual se destinaram as campanhas realizadas.

Em paralelo, estamos avaliando a possibilidade de integrar os campos *lead\_hash\_email* e *instagram\_hash\_bio* para criação de chave única a fim de identificar o fluxo e histórico correto desde o cadastro de lead até o momento em que se tornou cliente de fato e passou a utilizar os serviços do aplicativo. Um ponto a ser levado em consideração é que foi informado pela Prepi que é permitido que o lojista realize seu cadastro, ainda como lead, várias vezes com o mesmo endereço de e-mail uma vez que está sendo captado por uma “landing page” (página de vendas), podendo ser ou não da mesma loja. Porém, caso seja convertido a cliente/usuário pode cadastrar somente uma vez e, por isso, optamos por criar uma chave única de identificação a fim de garantir esta validação.

Como o objetivo do projeto é realizar a construção de um “lead scoring” que possibilite a análise de potenciais futuros clientes, iremos considerar a variável *user\_val\_cac* nas análises e aplicação do modelo pois representa o custo obtido na captação do lead durante as campanhas. Nesta coluna também temos valores nulos que serão desconsiderados. Tal variável também será tratada e convertida para tipo numérico.

# DATA PREPARATION

---

## 3.5 Formatando os dados

Foi identificado que existe a necessidade de desmembrar as colunas com tipo data em dimensões de dia, mês, ano, semana do ano e dia da semana. As colunas que devem receber esse tratamento são:

- *user\_dt\_churn\_prevista*
- *user\_dt\_cliente*
- *user\_dt\_diagnostico*
- *user\_dt\_inadimplencia\_inicial*
- *user\_dt\_lead*
- *user\_dt\_trial*
- *instagram\_dt\_criacao\_estimada*
- *lead\_dt\_cadastro*

Também será necessário realizar uma normalização de dados, nas colunas de data desmembradas para avaliação correta de sazonalidade de ocorrência de cadastro de usuários.

Outro ponto de observação são os dados contidos na coluna *lead\_utm\_source* que demonstram um número de categorias muito grande e tem suas informações contidas nas colunas *lead\_utm\_medium*, *lead\_utm\_campaign* e *lead\_utm\_content*.

Também será necessário reagrupar os dados das colunas categóricas:

- *lead\_utm\_campaign* (176 categorias),
- *lead\_utm\_content* (170 categorias),
- *lead\_city* (705 categorias),
- *instagram\_public\_zip* (151 categorias),
- *instagram\_category\_bussiness* (89 categorias),
- *user\_plano* (45 categorias),
- *user\_cidade* (254 categorias)

Devido a grande quantidade de tipos de categorias será necessário utilizar métodos de redução de categorias.

# MODELING

## 4.1 Selecionando a técnica de modelagem

Dentro do objetivo proposto as etapas se dividem em duas, a partir da elaboração de dois modelos diferentes que se complementam. A primeira etapa de modelling consiste no estudo das variáveis explicativas que trouxessem como resultado a geração de um Lead Scoring, ou seja, um pipeline que indicasse os Leads que teriam tendência de virar usuário e também clientes. A segunda, já é mais direcionada para a parte de custos de aquisição, sendo relacionados esses Leads ao CAC (Custo de aquisição de clientes).

Para a etapa do modelo de Lead Scoring foram selecionadas variáveis mais descritivas, no geral, categóricas. Essas variáveis consistem em dados em que poderiam definir características dos Leads que tivessem tendência a se tornarem usuários ou clientes. Dentre as selecionadas, foi realizado uma análise posterior em relação a qualidade dos dados disponíveis e as possibilidades de tratamento, encoding e scaling, resultando em:

- 'lead\_hr\_cadastro\_ajustado',
- 'lead\_fluxo\_ajustado',
- 'lead\_tamanho\_negocio\_ajustado',
- 'lead\_utm\_medium\_ajustado',
- 'lead\_lp\_version\_ajustado',
- 'lead\_referrer\_ajustado',
- 'lead\_city\_ajustado',
- 'lead\_region\_ajustado',
- 'lead\_num\_semana\_cadastro\_ajustado',
- 'lead\_flag\_became\_client'
- 'lead\_flag\_became\_user'

A princípio, a intenção é aplicar a função `get_dummies()` para categorização das variáveis. No entanto, algumas delas possuem número elevado de categorias e está sendo analisado se isso será um problema e , em caso afirmativo, como contorna-lo.



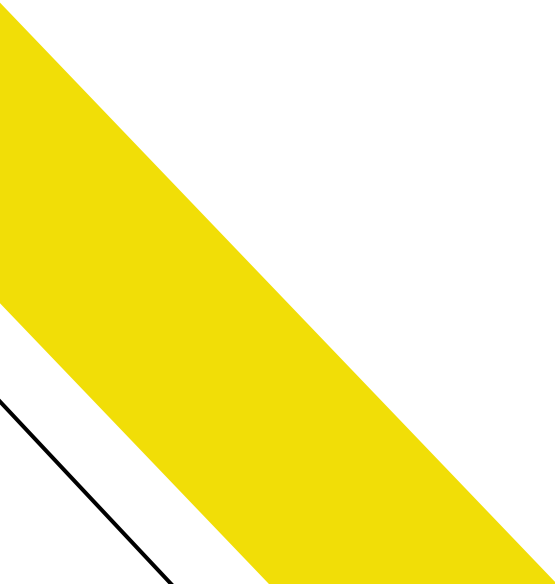
# MODELING

---

## 4.1 Selecionando a técnica de modelagem

Inicialmente, optou-se por utilizar modelos de classificação para realizar o Lead Scoring, tendo em vista que a maioria das variáveis explicativas são categóricas e o objetivo é tomar decisões binárias, facilitando assim a automatização da seleção dos Leads mais promissores para investimento.

Dentre os modelos de classificação disponíveis, será utilizada a biblioteca Sklearn. Para avaliar as capacidades dos modelos possíveis, serão realizadas avaliações por meio de cross-validation utilizando a biblioteca sklearn.model\_selection através da função GridSearchCV. Para isso, será implementada uma função em loop que possibilite avaliar cada modelo e retornar não apenas o melhor modelo, mas também os parâmetros otimizados para utilização. Alternativamente, o randomizedSearchCV poderá ser utilizado para encontrar a melhor combinação de hiperparâmetros, com o objetivo de otimizar a precisão do modelo e melhorar sua performance.



# MODELING

---

## 4.2 Gerar design de teste

Para construir o modelo de Lead Scoring, será realizado o Train-Test Split no subconjunto de dados previamente selecionado, sendo alocados 70% dos dados para treinamento e 30% para teste. Considerando que o modelo de Lead Scoring possui dois objetivos distintos, será feita a divisão em dois targets, utilizando a coluna `lead_flag_became_client` em um momento e a coluna `lead_flag_became_user` em outro, ambos relevantes para o modelo de negócios da empresa.

Uma vez que o modelo é de classificação, a avaliação será feita por meio da biblioteca `sklearn.metrics`, utilizando as funções `accuracy_score`, `balanced_accuracy_score`, `precision_score`, `recall_score`, `f1_score` e `roc_auc_score` para comparar os resultados entre os conjuntos de treinamento e teste. Ademais, será implementada a visualização da curva ROC para avaliar o desempenho do modelo em relação aos verdadeiros positivos e falsos positivos.

# EVALUATION

## 5.1 Avaliação dos resultados

A fase de implementação do modelo apresentou uma série de desafios significativos no contexto do projeto. Uma ampla variedade de modelos foi submetida a testes, utilizando diferentes porcentagens de treino e teste, a fim de alcançar níveis superiores de precisão e confiabilidade. Adicionalmente, foi observado que ajustes eram necessários na etapa de preparação dos dados, para que fosse possível otimizar sua resposta e alinhá-la com as expectativas estabelecidas.

Os modelos desenvolvidos apresentaram resultados promissores, com bom desempenho em métricas como acurácia, acurácia balanceada, ROC AUC e F1 Score. Essas métricas indicam que os modelos têm a capacidade de identificar com precisão as leads com maior probabilidade de se tornarem clientes ou usuários, o que está alinhado com os objetivos de negócio da empresa PREPI. Portanto, os modelos atendem aos critérios de sucesso do negócio.

Com base nos resultados obtidos, recomenda-se aprovar os seguintes modelos para o negócio: Logistic Regression, Decision Tree e Random Forest. Essa precisão e capacidade de generalização tornam esses modelos mais confiáveis para a previsão de leads com maior probabilidade de se tornarem clientes ou usuários da PREPI.

O **Logistic Regression** é uma técnica de aprendizado supervisionado utilizada para prever a probabilidade de um evento ocorrer. Ele é frequentemente usado em problemas de classificação binária, onde o objetivo é prever se uma instância pertence a uma das duas classes possíveis.

```
Acc Treino: 89.95883563657748
Acc Teste: 90.77013521457967
=====
F1 Treino: 86.66614565878184
F1 Teste: 87.63608571112988
=====
Confusion Matrix
[[1525  8]
 [ 149 19]]
=====
Classification Report
      precision    recall  f1-score   support

     0       0.91      0.99      0.95     1533
     1       0.70      0.11      0.19       168

 accuracy          0.91     1701
 macro avg       0.81     0.55     0.57     1701
 weighted avg    0.89     0.91     0.88     1701

Precision: Percentage of correct positive predictions relative to total positive predictions
Recall: Percentage of correct positive predictions relative to total actual positives
F1 Score: A weighted harmonic mean of precision and recall. The closer to 1, the better the model
```

# EVALUATION

## 5.1 Avaliação dos resultados

O modelo utiliza uma função logística para calcular a probabilidade da instância pertencer a uma das classes. Essa função recebe como entrada um conjunto de variáveis independentes e seus respectivos pesos, que são ajustados durante o treinamento do modelo para minimizar o erro na previsão. Uma vez treinado, o modelo pode ser usado para fazer previsões sobre novas instâncias, atribuindo a elas uma probabilidade de pertencer a uma das classes.

**Decision Tree** é um modelo de aprendizado de máquina supervisionado que é usado para resolver problemas de classificação e regressão. O modelo funciona dividindo o conjunto de dados em subconjuntos menores com base em uma série de perguntas que são feitas sobre as variáveis independentes. O objetivo do modelo é criar uma árvore que seja capaz de classificar corretamente novas instâncias. Uma vez construída, a árvore pode ser usada para fazer previsões sobre novas instâncias, seguindo o caminho da árvore até chegar a uma decisão final.

```
Acc Treino: 99.97059688326962
Acc Teste: 83.77425044091711
=====
F1 Treino: 99.9705788816629
F1 Teste: 84.13885881967202
=====
Confusion Matrix
[[1386 147]
 [ 129 39]]
=====
Classification Report
      precision    recall  f1-score   support

     0       0.91      0.90      0.91      1533
     1       0.21      0.23      0.22       168

   accuracy          0.84      1701
  macro avg          0.56      1701
 weighted avg          0.85      1701

Precision: Percentage of correct positive predictions relative to total positive predictions
Recall: Percentage of correct positive predictions relative to total actual positives
F1 Score: A weighted harmonic mean of precision and recall. The closer to 1, the better the model
```

# EVALUATION

## 5.1 Avaliação dos resultados

**Random Forest** funciona criando várias árvores de decisão independentes, cada uma treinada com uma amostra aleatória dos dados de treinamento e com um subconjunto aleatório das variáveis independentes.

Durante o processo de treinamento, cada árvore é construída usando uma amostra diferente dos dados de treinamento, tornando cada árvore única. Em seguida, o modelo combina as previsões de todas as árvores para chegar a uma previsão final.

O objetivo do Random Forest é reduzir a variância do modelo, tornando-o mais robusto e menos propenso a overfitting. Além disso, o modelo é capaz de lidar com grandes conjuntos de dados e com variáveis categóricas e numéricas.

```
Acc Treino: 99.91179064980888
Acc Teste: 90.29982363315696
=====
F1 Treino: 99.91162811630403
F1 Teste: 86.56875713139313
=====
Confusion Matrix
[[1526  7]
 [ 158 10]]
=====
Classification Report
      precision    recall  f1-score   support

     0       0.91      1.00      0.95      1533
     1       0.59      0.06      0.11       168

 accuracy      0.90      0.90      0.90      1701
 macro avg      0.75      0.53      0.53      1701
 weighted avg      0.87      0.90      0.87      1701

Precision: Percentage of correct positive predictions relative to total positive predictions
Recall: Percentage of correct positive predictions relative to total actual positives
F1 Score: A weighted harmonic mean of precision and recall. The closer to 1, the better the model
```

# EVALUATION

## 5.2 Revisão do processo

A revisão do processo revelou que todas as etapas do projeto foram cumpridas conforme o planejado. O modelo desenvolvido atendeu aos objetivos de negócio estabelecidos no início do projeto, alcançando resultados satisfatórios. Os modelos selecionados demonstraram precisão na identificação de leads com potencial de se tornarem clientes ou usuários da PREPI, fornecendo resultados precisos para apoiar a tomada de decisões relacionadas à geração de leads e à estratégia de aquisição de clientes.

Além disso, o processo de data mining e modeling resultou em alguns insights relevantes:

- Foi verificado que na região Sudeste, é onde encontra-se a maioria dos leads, sendo a maior concentração no Estado de São Paulo.
- Os dados demonstraram que o horário noturno, que compreende entre 18:00 hrs e 04:00 horas, foi mais propenso aos cadastros, no entanto, o horário de maior destaque foi as 13:00. Fica evidente que, horários no qual as pessoas mais se cadastram são àqueles que provavelmente estão fora do horário de trabalho (almoço e descanso).
- Com relação ao fluxo, ficou evidente que a principal forma de acesso aos clientes à empresa, é por meio do tráfego direto (TD), ou seja, o primeiro contato que os clientes tiveram, foi diretamente decorrentes de campanhas de tráfego pago. Além disso, foi observado que o principal canal de contato com a campanha, foi o *Instagram*, mais efetivo por meio do *feed*.
- Ainda foi observado que a terceira e a quarta semana no mês apresentam maiores chances de cadastro.
- Não foi possível gerar análise de sentimento, devido o número de nulos extremamente elevados.

# EVALUATION

## 5.3 Determinação dos próximos passos

Com base nos resultados alcançados, podemos afirmar que o modelo está pronto para a implementação. Ele atendeu aos critérios de sucesso do negócio e demonstrou um desempenho consistente e confiável na identificação de leads com maior probabilidade de se tornarem clientes ou usuários da empresa.

Quanto às possíveis ações futuras, recomenda-se que seja considerado as seguintes abordagens:

- Melhorar a qualidade dos dados: A base de dados apresenta uma qualidade razoável, contando com número elevado de nulos, fazendo com que várias colunas não pudessem ser utilizadas, pois correria o risco de resultados tendenciosos e não fidedignos.
- Coleta de mais dados: Aumentar o tamanho do conjunto de dados de treinamento pode ajudar a capturar mais informações relevantes e aprimorar o desempenho do modelo.
- Monitoramento contínuo: Implementar um sistema de monitoramento contínuo do desempenho do modelo em ambiente de produção para identificar qualquer degradação no desempenho e tomar medidas corretivas prontamente.

Essas ações futuras ajudarão a garantir que o modelo esteja sempre atualizado e fornecendo resultados precisos e relevantes para a empresa. A implementação dessas recomendações deve ser cuidadosamente planejada e acompanhada para maximizar o sucesso do projeto.

Podemos dizer que o projeto alcançou resultados promissores e está bem encaminhado para a implementação. O modelo desenvolvido é capaz de identificar leads com potencial para se tornarem clientes ou usuários, alinhando-se aos objetivos de negócio da empresa.

No entanto, é importante ressaltar que a implementação e o acompanhamento contínuo do modelo são essenciais para garantir seu sucesso a longo prazo. A análise e a adaptação constante às mudanças no ambiente de negócios serão fundamentais para maximizar e manter a eficácia do modelo ao longo do tempo.



# DEPLOYMENT

---

## 5.1 Plano de implementação

Considerando a utilização de uma tabela de dados brutos em formato CSV, contendo dados captados pela empresa, mantida inicialmente no Google Drive e posteriormente transferida para o Google Colab para limpeza e preparação dos dados. Nesse caso, as ações para implementação podem incluir:

- Explorar opções de implantação que permitam acesso e manipulação eficientes de arquivos CSV, como a utilização de plataformas de nuvem ou servidores locais. Os dados poderão ser mantidos em nuvem, como Google Drive e ter integração direta com a plataforma Google Colab.
- Garantir a compatibilidade dos sistemas de implantação com a transferência de dados entre o Google Drive e o ambiente de processamento, como o Google Colab. Para isso, precisa haver a integração das plataformas.
- Documentar etapas e procedimentos específicos para a transferência, limpeza e preparação dos dados, levando em consideração a movimentação entre diferentes plataformas. Para isso, foi realizado a documentação detalhada dos códigos no Google Colab.
- A empresa pode considerar a automação dessas etapas, se possível, para simplificar e otimizar o processo de preparação dos dados.

Com base no fluxo descrito, alguns possíveis problemas que podem ocorrer são:

- Restrições de armazenamento ou capacidade nas plataformas de nuvem utilizadas. Visto que trata-se de um volume considerável de dados.
- Dificuldades técnicas na transferência dos dados entre as diferentes etapas do processo.
- Limitações de desempenho durante a modelagem dos dados no ambiente escolhido. Visto que a plataforma utilizada foi o Google Colab, sendo assim, a codificação é compatível basicamente nessa plataforma, sendo necessário ajustes, caso posteriormente seja escolhido outra plataforma de codificação em Python.

# DEPLOYMENT

---

## 5.1 Plano de implementação

Medidas alternativas que podem ser consideradas para mitigar esses problemas incluem:

- Explorar outras opções de formatos de dados, como o uso de bancos de dados ou formatos mais adequados às ferramentas de limpeza e preparação utilizadas.
- Avaliar diferentes plataformas de nuvem que possam oferecer maior capacidade de armazenamento ou desempenho.
- Verificar a disponibilidade de bibliotecas ou ferramentas específicas para transferência de dados entre as plataformas utilizadas.

Ainda sobre o plano de implementação, serão realizadas medidas como:

- Documentar os passos necessários para acessar e utilizar o modelo, bem como as configurações e parâmetros relevantes no Google Colab.
- Utilizar uma linguagem clara e acessível, para facilitar a compreensão do usuário..
- Planejar e conduzir sessões de treinamento interativas, onde o usuário possa aprender na prática como interpretar e utilizar os resultados gerados pelo modelo, utilizando plataforma *Microsoft Teams*.
- Responder a perguntas e fornecer suporte durante as sessões de treinamento, garantindo que o usuário se sinta confortável e confiante no uso do modelo.
- Disponibilizar canais de comunicação, como *e-mails* e *WhatsApp*, onde os usuários possam fazer perguntas, relatar problemas ou buscar orientação adicional.

# DEPLOYMENT

## 5.2 Plano de monitoramento e duplicação

Monitoramento das ações:

- Utilizar as métricas relevantes para avaliar o desempenho e a acurácia do modelo implantado, como taxa de acerto, precisão, recall e F1-score.
- Analisar os indicadores-chave de sucesso que estão alinhados com os objetivos de negócios, como aumento na eficiência operacional, redução de custos ou melhoria na tomada de decisões.
- Definir uma frequência adequada para análise dos resultados e identificação de possíveis desvios ou problemas.

Determinação do momento de descontinuação do uso do modelo:

- O modelo deve ser descontinuado quando a validade dos dados, determinados pela empresa, forem alcançadas, ou pela decisão de descontinuação de coleta de dados com as mesmas variáveis;
- É importante que a empresa documente os critérios e as ações específicas que devem ser tomadas caso o modelo não possa mais ser usado, como atualizar o modelo existente, configurar um novo projeto de mineração de dados ou buscar alternativas.

Evolução dos objetivos de negócios:

- É necessário reconhecer que os objetivos de negócios podem mudar ao longo do tempo devido a mudanças nas necessidades, estratégias ou condições do negócio. Com isso é importante realizar avaliações periódicas para alinhar os objetivos do modelo aos objetivos de negócios em evolução e determinar se ajustes ou atualizações são necessários.

# DEPLOYMENT

---


## 5.2 Plano de monitoramento e duplicação

Durante a implementação dos modelos desenvolvidos, é importante estar ciente da necessidade de atualizações periódicas para garantir seu desempenho e compatibilidade com as versões mais recentes das bibliotecas e extensões em Python. As atualizações são necessárias devido ao constante desenvolvimento e lançamento de novas versões das ferramentas utilizadas no projeto.

Além disso, é importante mencionar que a portabilidade do modelo para diferentes plataformas que suportem a linguagem de programação Python, como o Jupyter, pode exigir ajustes específicos nos códigos. Esses ajustes podem ser necessários devido a diferenças na disponibilidade de extensões e bibliotecas entre as plataformas. É essencial garantir que todas as dependências sejam adequadamente instaladas e que os comandos de sintaxe sejam ajustados conforme necessário.

Ao migrar para uma nova plataforma, é fundamental revisar e atualizar as dependências do ambiente de desenvolvimento, considerando as versões das bibliotecas e extensões compatíveis com a nova plataforma. Isso pode envolver a atualização de pacotes, resolução de conflitos de dependências e adaptação do código para qualquer diferença de sintaxe ou funcionalidade nas bibliotecas específicas da plataforma.

A realização desses ajustes técnicos é crucial para garantir a correta execução e funcionamento dos modelos em diferentes ambientes e versões de bibliotecas. Além disso, a documentação detalhada das configurações do ambiente, incluindo versões de bibliotecas e extensões específicas, é fundamental para facilitar futuras atualizações e manutenções.



# DEPLOYMENT

---

## 5.3 Relatório final

O objetivo do projeto foi auxiliar na identificação da persona ideal e determinar quantas pessoas precisam ser atingidas para que uma venda seja concluída, visando reduzir custos ou elevar o faturamento e melhorar o lucro da empresa Prepi. O projeto utilizou uma tabela resultante da junção de 3 outras planilhas, disponibilizados pela empresa. Os prazos do projeto foram de acordo com os prazos estipulados pela DNC para o cumprimento de cada etapa. Os riscos do projeto incluíram desvio do escopo, que não foi evidenciado devido aos alinhamentos e ajustes constantes, juntamente com a equipe Prepi.

Visto tal objetivo, ele foi atingido com o uso dos dados dos clientes atuais da empresa, dados da jornada de compra destes clientes em conjunto com os dados de resultados de campanhas. A base de dados utilizada continha aproximadamente 34 mil contatos, com relacionamento entre as campanhas que os clientes tiveram interação, antes da compra ou cadastro do aplicativo. As etapas posteriores foram:

- Avaliar a acurácia de modelos de lead score;
- Definir as métricas para validar a qualidade dos dados;
- Definir a taxa de acurácia do modelo de machine learning.

O processo de mineração de dados foi uma etapa crucial e desafiadora. Os dados brutos foram analisados detalhadamente e, após a disponibilização dos metadados, identificamos as colunas mais relevantes para as análises. Além disso, houve a necessidade de anonimização de dados sensíveis para garantir a privacidade dos indivíduos. Utilizando o Google Colab e a linguagem de programação Python, realizamos explorações dos dados estruturados e não estruturados em busca de padrões, associações, anomalias, tendências e correlações.

# DEPLOYMENT

---

## 5.3 Relatório final

Nesse processo, identificamos desafios que poderiam impactar as análises e a implementação do modelo, compreendendo que essa etapa seria fundamental para as próximas fases e exigiria revisões contínuas. Observamos uma quantidade significativa de valores nulos no banco de dados, e mesmo após encontros com a empresa parceira para melhorar a qualidade dos dados, não foram realizadas alterações no banco. No entanto, decidimos trabalhar com os dados disponíveis, entendendo que essa era a situação atual e nos empenhamos para obter os melhores resultados possíveis. As colunas foram renomeadas seguindo as melhores práticas, orientadas pelos mentores e pelos ensinamentos do curso de mineração de dados. Além disso, lidamos com um grande número de variáveis categóricas não categorizadas, e realizamos o tratamento adequado. Optamos por manter todas as colunas, mesmo aquelas com um alto número de valores nulos, para análises futuras. No entanto, tivemos cautela ao preencher os valores nulos, evitando gerar análises tendenciosas ao lidar com colunas com alta quantidade de valores faltantes.

No projeto de Lead Scoring, foi realizada a implementação de modelos de classificação utilizando a biblioteca Sklearn. A escolha desses modelos se deu pela predominância de variáveis categóricas e pela necessidade de tomar decisões binárias na seleção de Leads promissores para investimento. Para avaliar a eficácia dos modelos, foi adotada a técnica de cross-validation com o auxílio da biblioteca `sklearn.model_selection` e da função `GridSearchCV`.

Através de um loop, foi possível avaliar cada modelo e obter não apenas o melhor modelo em termos de desempenho, mas também os parâmetros otimizados para sua utilização. Além disso, considerou-se a utilização do `randomizedSearchCV` como alternativa para encontrar a melhor combinação de hiperparâmetros, visando aprimorar a precisão e o desempenho geral do modelo.



# DEPLOYMENT

---

## 5.3 Relatório final

Essa abordagem permitiu a seleção do modelo de classificação mais adequado para o Lead Scoring, proporcionando maior automatização e eficiência na identificação dos Leads com maior potencial de conversão. Os resultados obtidos contribuíram para a tomada de decisões estratégicas e otimização dos recursos de investimento.

No que diz respeito ao plano de implementação, foram documentados todos os passos necessários para acessar e utilizar o modelo, incluindo as configurações e parâmetros relevantes no Google Colab. A linguagem utilizada na documentação foi clara e acessível, garantindo a compreensão do usuário.

Inicialmente, foi explorada a opção de implantação que permitisse o acesso e a manipulação eficientes de arquivos CSV, optando pela integração com plataformas de nuvem, como o Google Drive, e o ambiente de processamento, como o Google Colab.

Durante o desenvolvimento do projeto, foram identificados alguns possíveis problemas, como restrições de armazenamento ou capacidade nas plataformas de nuvem, dificuldades técnicas na transferência dos dados e limitações de desempenho durante a modelagem dos dados no ambiente escolhido. Para mitigar esses problemas, foram adotadas medidas alternativas, como explorar diferentes opções de formatos de dados, avaliar outras plataformas de nuvem e buscar bibliotecas ou ferramentas específicas para a transferência de dados.

