

DATA CHALLENGE 2022

Uma iniciativa Stone

*“You must gain control over your money or the lack of it will forever control you”*

*Dave Ramsey*

# Sumário

[Instruções de preenchimento deste documento](#_gjdgxs)

[Preenchimento framework de avaliação](#_30j0zll) [Validações iniciais](#_1fob9te)

[Fluxo de etapas realizadas](#_3znysh7) [Conclusões e insights](#_2et92p0)

[Checklist de entregas](#_tyjcwt)

# Instruções de preenchimento deste documento

Este documento funciona como um **template a ser preenchido durante a execução do case** com o intuito de auxiliar e nos contar quais etapas você desenvolveu até chegar à sua conclusão final!

O principal intuito aqui é representar de forma sumarizada sua **linha de raciocínio e análise durante a resolução do case e garantir uma avaliação mais completa do seu trabalho.**

# Preenchendo o framework de avaliação

Abaixo, você encontrará todas as etapas que deverão ser preenchidas. Preste muita atenção durante o preenchimento e não se esqueça de revisar este documento antes de enviá-lo. É através deste documento que boa parte da avaliação será conduzida **então, capricha! ;)**

# Validações iniciais

Com intuito de validar algumas premissas básicas de junção e preparação dos dados, responda às seguintes perguntas:

* + 1. Quantos contratos distintos existem em sua tabela de observação final, após todos os merges e etapas de ETL?

# 14754 Contratos distintos

* + 1. Você considerou ou aplicou algum filtro de exclusão? Caso afirmativo, descreva-o e informe quantos contratos foram filtrados.

# Foram dois critérios de exclusao

1 - eliminei as linhas com estado com descriçao ‘ND’

2 - Após análise sazonalidade considerei dados dos útimos 6 meses

2 **Contratos filtrados em relacao ao número de contratos\_id inicial**

* 1. **Fluxo de etapas realizadas**

Conforme os padrões e exemplos dados abaixo, crie uma etapa para cada processo realizado por você durante a solução do case, registrando sua **linha de raciocínio** com todos os detalhes necessários para que a banca avaliadora **entenda a sua abordagem na resolução do problema.** Utilize quantas etapas achar necessário, porém **seja objetivo!**

Use o exemplo abaixo como um guia para o preenchimento. Sugerimos que você **preencha este documento simultaneamente com o desenvolvimento** do problema proposto para que você não esqueça de registrar nenhuma etapa.

Lembre-se, capriche nas descrições, **a defesa do seu case começa aqui!**

A estratégia usada foi principalmente feita em python e uma análise complementar visual em Power Bi, importante ressaltar que as duas análises se complementar nao dizem a mesma coisa , obejtivo foi clusterizar a base e responder a pergunta de negócio.

Etapas:

**1 - BUSINESS UNDERSTANDING**

Nesa primeira etapa o objetivo e entender o contexto dos dados e entender também qual pergunta de negócio que tenho que responder com a análise de dados

**2 - DATA UNDERSTANDING**

Processo de entendimento dos arquivos de dados de forma separada e através do diagrama MER, entendendo o relacionamento entre as tabelas

2.1 - LOAD DATASETS

Carregamento e visualização dos dados de cada base

2.2 - EXPLORATORY DATA ANALYSIS

Nesta etapa fiz um overview das metricas básicas sobre os datasets, Ex: shape, missing, duplicated.

**3 - ETL**

Etapa fundamental aqui fiz todos os arranjos e transformações em cada dataset com o objetivo de criar uma base única, com todas as dimensões do problema a ser resolvido.

Ex: criação de chaves únicas, agrupamentos, label encoding, remoção de linhas duplicadas, exportação de outputs de cada dataset

3.1 - PORTIFÓLIO COMUNICADOS

Após a analise do dataset comunicados, decidi agrupar todas as informacoes de ação e status em uma linha só, antes cada tipo de acao ocupava uma linha e o contrato\_id nao era um indexador único, criei entao uma PK com a uniao do contrato\_id + dt\_ref\_portfolio, criando uma chave para o merge com portifolio geral.

3.2 - PORTIFÓLIO TPV

Apos a analise do dataset, corrigi o formato da data e criei uma PK com a uniao do 'nr\_documento' + 'dt\_transacao', criando uma chave para o merge com portifolio geral.

3.3 - PORTIFÓLIO CLIENTES

Apos a analise do dataset, removi as linhas duplicadas

3.4 - CRIACAO DA BASE ÚNICA

Foi feito o carregamentos dos resultados dos etls das bases de comunicados/tpv/clientes para fazer o join com portifolio geral

gerando assim a base única, foram feito os seguintes passsos:

- Criação de chaves estrageiras das tabelas dimensão no portofolio geral

- foi feitos os joins separadamente

- rename das colunas e drop das colunas desnecessárias

- exportação do output do etl geral para proxima parte de análise

3.5 - ENRIQUECIMENTO COM FONTES EXTERNAS

- foram eliminadas as colunas UF e código

- Correção dos tipos de dados

- Feito o Join com o dataset df

- foi exportado o 1º data point

**portifolio\_full.csv**

**4 - DATA ANALYSIS**

Foi feito o carregamento da base de dados resultante do Etl abaixo:

*portfolio\_full = p\_geral + p\_comunicados + p\_cliente + p\_ptv + indices\_por\_uf\_ibge*

4.1 - MÉTRICAS (BASE LINE)

- O objetivo aqui é encontrar os parâmetros de controle para os insights

4.1.1 - Curva de sazionalidade por data de transacao (portifolio tpv)

- Aqui eu gerei um gráfico com o objetivo de encotrar a curva de sazonalidade e escolher o melhor intervalo de dados para se gerar insights olhando para o cenário futuro mais aproximado da atual situaçao, foi gerado também 3 gráficos que expressam claramente que o intervalo dos últimos 6 meses refletem o ponto de ruptura, onde a análise será com a base do problema de negócio.

- Após essa análise o novo df estará filtrado no seguinte espaco de tempo ***19-10-2021 a 19-04-2022***.

4.1.2 - Agrupamentos e filtros da base somente linhas com acao

- Aqui foi filtrado gerando o df1 para extrair as métricas de campanhas

4.1.3 - Gráfico engajamento por canal

-Esse gráfico mostra de maneira geral como está o engajamento por todas as açoes por email e hsm

4.1.4 - Gráfico Geral de engajamento por canal e tipo de ação

-Esse gráfico mostra de maneira geral como está o engajamento por tipo de ação nas canais de email e hsm

4.1.5 - rankinq geral de dsp/dspp por segmento/tipo de empresa

- Esse gráfico mostra o ranking

4.1.6 - Criação da janelas de efetividade das ações

- Nessa parte eu fiz dois grandes loops com o ojetivo de criar duas colunas importantes para minha análise a efetividade e acao2

- efetividade - irá informar se houve pagamento apos um acao ter sido lida/ respondida e estiver houver um pagamento em no máximo 5 dias após a campanha.

- acao 2 - irá indeficar quais acoes com base no dsp e dspp

Nesse ponto é criado o 2º datapoint

**df\_efetividade.csv**

4.1.7 - gráfico de efetividade de pagamento

- Sao dois gráficos que serao a base line para análise futura da base em clusters

- Curva de efetividade por acionamentos x efetividade da acao

- indexes % de contratos e % efetividade de contratos

- Curva de efetividade por acao

- Curva de efetividade por acao na zona de desengajamento

4.2 - CLUSTERIZAÇÃO DA BASE

O objetivo aqui foi encontrar insights para diferentes tipos de clientes por meio de um algoritmo de machine learning de clusterizaçao

O Kmeans. doi selecionado as variaveis númericas mais relevantes para essa acao, uma vez que o Algoritmo usa centroídes(média)

4.2.1 - Gráfico de Elbon - Kmeans

- Grádico de Elbom ele mostra o nº de cluster ideal (hiperparemetro do modelo)

- Através do gráfico foi definido que 2 clusters seria o ideal para a nossa análise

4.2.2 - Construindo e Treinando o Modelo - Kmeans

- Aplicando os hiperametros necessários e cronstruindo o modelo

- Foi feito o treinamento do modelo

- Após o treinamento fois extraídos e atribuido a um objeto as labels dos rótulos

- Foi feito o objeto cluster (dim da base clusterizada) e foi feito um join com as labels) (contrao\_id (PK) e labels)

4.2.3 - Construindo base cluserizada - Kmeans

- foi feito o join da base df + cluster para atribuir um tag de cluster em toda a base

- Após isso a base foi dividida em 2 clusters

Nesse ponto tem se o 3º datapoint

**df\_clusterizada.csv**

4.2.4 - Cluster 0 - Kmeans

- foram criadas todas as bases para gerar os gráficos que analisa dois aspectos - - Curva de efetividade por acionamentos x efetividade da acao

- indexes % de contratos e % efetividade de contratos

- Curva de efetividade por acao

- Curva de efetividade por acao na zona desengajamento

4.2.5 - Cluster 1 - Kmeans

- foram criadas todas as bases para gerar os gráficos que analisa dois aspectos - - Curva de efetividade por acionamentos x efetividade da acao

- indexes % de contratos e % efetividade de contratos

- Curva de efetividade por acao

- Curva de efetividade por acao na zona desengajamento

**5 - SUGESTAO DE RÉGUAS DE ACIONAMENTOS**

Nessa parte faço um comparativo dos indexes de base line x clusters 0 e 1 e definir a estratégia a nova régua de cobrança, que por sua vez irá gerar a curva ideal para cada cluster.

5.1 - Base line

Tabela com indexes dsp/dspp/geral de efetividade das açoes geral e na zona de desengajamento.

5.2 - Cluster 0 - Régua de Cobrança

Tabela com indexes dsp/dspp/geral de efetividade das açoes geral e na zona de desengajamento

5.2.1 - Insights Cluster 0

faço um resumo dos insights encontrados

5.3 - Cluster 1 - Régua de Cobrança

Tabela com indexes dsp/dspp/geral de efetividade das açoes geral e na zona de desengajamento

5.3.1- Insights Cluster 1

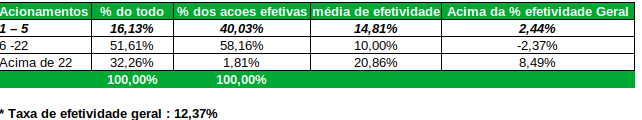
faço um resumo dos insights encontrados

**6 - QUAL É A CURVA IDEAL DE ACIONAMENTOS?**

A curva ideal de acionamentos dos contratos após todos os tratamentos feitos é a seguinte

#### **é de 1 a 5 vezes**

pelos seguintes indicadores e pelo gráfico abaixo:



1

# Conclusões e insights

Para finalizar o preenchimento deste documento, você irá compilar os insights encontrados e a proposta de solução do problema em forma de uma dissertação curta, **com até 20 linhas**.

Lembre-se que estamos em um hackathon de dados, então **use e abuse de números, análises e correlações encontradas** nos dados **para embasar sua conclusão**!

Não se esqueça também de apresentar as **análises complementares** que serão o **diferencial na sua resolução**.

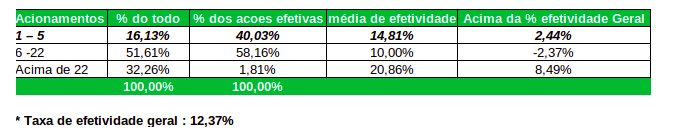
# Conte pra nós, qual é a curva ideal de vezes que devemos acionar um cliente?

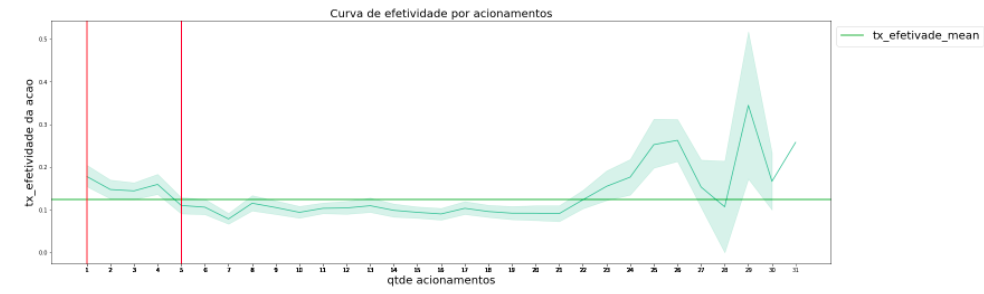
**Após todo o Etl onde agrupei todos os datasets na granularidade linha pela data de referencia dos contratos , fiz o enriquecimento da base de dados, em sequencia analisei a sazonalidade e decidi considerar somente os últimos 6 meses (*19-10-2021 a 19-04-2022*), por considerar esse período o mais adequado para responder a pergunta de negócio.**

**Após definir o critério de efetivade (Acao enviada, lida e/ou respondida e com pagamento nos 5 dias sequintes ao do envio da acao) e apliquei a base já filtrada pelos útlimos 6 meses, chegando a sequinte curva ideal:**

#### **é de 1 a 5 vezes**

**pelos seguintes indicadores e pelo gráfico abaixo:**

****

******

# Checklist de entregas

Para te ajudar no processo de entrega final, separamos este checklist para garantir que você não deixe nada para trás:

* Editar a permissão de seu projeto Github para **público**
* Upload dos códigos e scripts no Github
* Upload do arquivo de dados final no Github
* Upload do documento ‘Framework de avaliação’ no Githubo
* Upload do dashboard no Github
* Upload da apresentação final (opcional)
* Enviar o link do seu projeto Github para o [tech@stone.com.br](mailto:tech@stone.com.br) até o dia 04/05 às 23h59.

Nós do time Stone agradecemos a sua participação!

# Boa sorte!