# Framework de avaliação - Data Challenge 2022

## Validações iniciais

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

R: Após os merges e tratamentos, a quantidade de contratos distintos foram 14756.

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

R: Sim, foram aplicados alguns filtros, são eles:

* Remoção de valores Nan, os quais não foram encontrados na base, e por isso não teve nenhum contrato removido.
* Remoção de valores duplicados, que ao analisar todas as tabelas foi visto que apenas no dataset de Clientes houve duplicados, sendo removidas 26 linhas de contratos duplicados.
* Ainda no dataset de Clientes foram observados que algumas colunas apresentavam valores “estranhos” e esses valores também foram removidos, como por exemplo nas colunas de Subsegmento e Estado, onde respectivamente tinham valores como None e ND, o que não fazem sentido para a análise e por isso foram removidos, e somando ao todo foram 89 linhas.

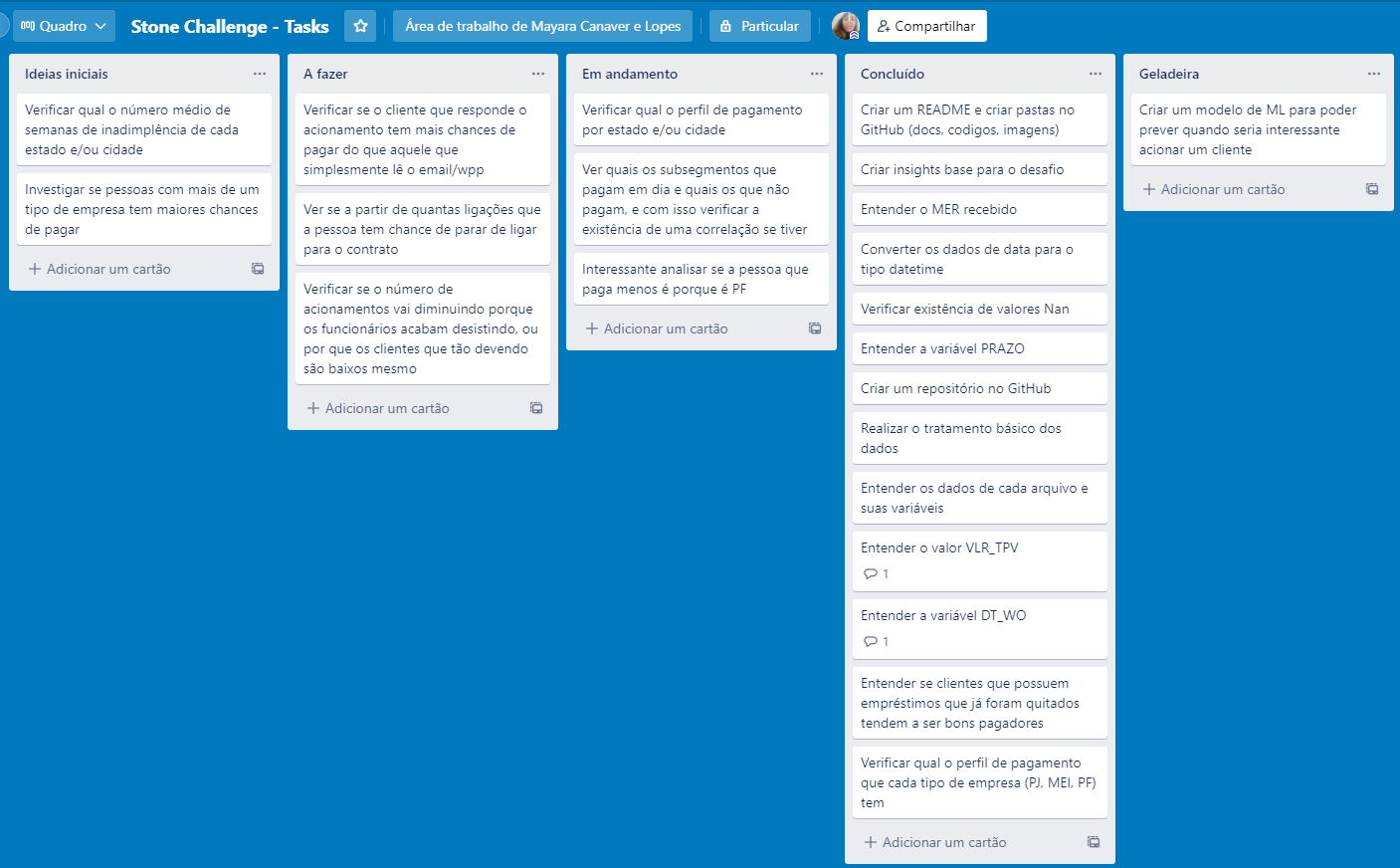
## Fluxo de etapas realizadas

#### Etapa 1 - Análise dos Dados:

Foi realizada uma observação e análise inicial dos dados, fazendo assim a compreensão de cada um dos dados e anotações e pesquisa de dados não entendidos, o mapa de dados foi utilizado principalmente nos dados de TPV e Geral, algumas anotações a mais foram realizadas também para dar mais significado ao dataset, experiências de trabalho anteriores também ajudaram a identificar e criar alguns insights para os dados.

#### Etapa 2 - Planejamento Organizacional:

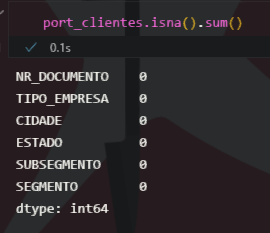
Foi feito um planejamento e separação das atividades que seriam feitas e possíveis insights do que poderia ser útil ou não ao projeto, a ferramenta Trello foi utilizada como base para o escopo dos insights e tarefas que precisavam ser feitas, e abaixo está um exemplo do andamento das atividades:



Foram definidas etapas de ideias, tarefas a se fazer, que estão em andamento, concluídas e uma parte de tarefas “congeladas”, onde caso desse tempo poderia ser realizado.

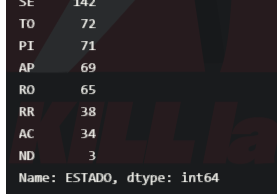
Além de criar também um documento contendo resumos e insights de cada tratamento realizado por coluna.

#### Etapa 3 - Limpeza e Tratamento dos Dados:

O tratamento como foi dito na parte de validação foi realizado ao observar cada coluna separadamente.

Alguns pontos a serem citados são por exemplo:

A existência de Nan, que foi abordado na etapa de validação, foi tratada individualmente para cada dataset e criada uma visão como a do lado, onde mostra para cada coluna se há ou não valores Nan para os dados.

Sobre os tratamentos adicionais de Estado e Subsegmento, o exemplo ao lado mostra sobre a coluna de Estado o que apareceu como dados “estranhos”.

Há 3 valores em ESTADO que estão listados como ND, como no Brasil não há um estado com a sigla de ND, foi considerado que esse valor possa ser “NÃO DEFINIDO”, e como há apenas 0.02% dessa variável neste dataset foi preferível remover esse valor.

#### Etapa 4 - Sumarização dos Dados:

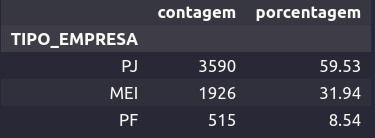
Como forma de guiar a definição da quantidade de acionamentos ideal, foram elaboradas estratégias para identificar perfis de bons e maus pagadores. Depois disso, características mais granulares (tipo de empresa, segmento, estado, status do contrato, dentre outras) foram analisadas a fim de caracterizar subgrupos e assim alcançar resultados específicos para cada tipo de perfil.

A caracterização de bons e maus pagadores se deu da seguinte forma: bons pagadores consistem de pessoas que possuem pelo menos 1 contrato anterior quitado e que não possui nenhum contrato prévio suspenso, enquanto que maus pagadores consistem de pessoas que possuem pelo menos um contrato pregresso que fora suspenso e que não possui nenhum contrato quitado.

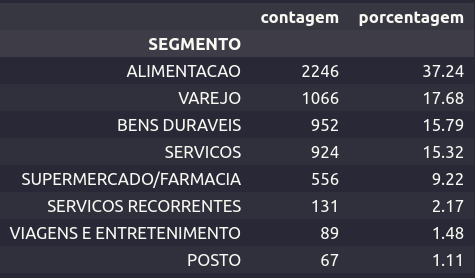
Os dados levaram em consideração todas as tabelas em sua totalidade, com exceção da tabela de portfólio geral - onde foi extraída uma amostra aleatória composta de 200 mil linhas a fim de facilitar a manipulação de dados e evitar problemas de memória.

Em seguida foi iniciada uma análise mais granular a fim de entender os nuances de cada subgrupo (definido por tipo de empresa, estado, segmento, status de contrato, etc) e como eles se relacionam com os perfis de pagadores.

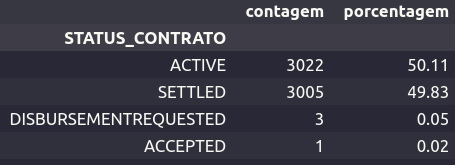
A partir dessa análise foi possível observar que a maioria dos bons pagadores são pessoas jurídicas seguidas dos Micro Empreendedores Individuais (MEI). As pessoas físicas são a parcela minoritária entre os bons pagadores no que tange o tipo de empresa.



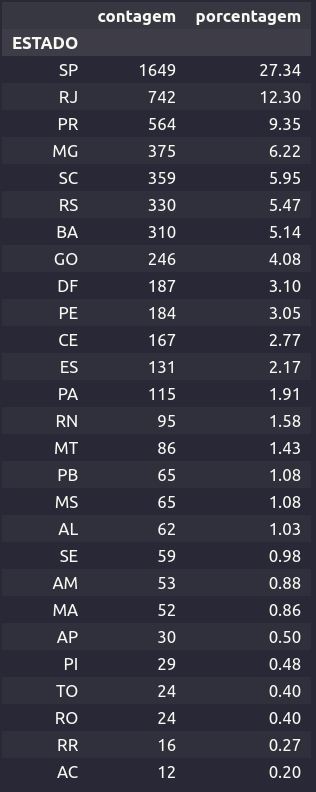
Além disso, os segmentos de alimentação, varejo, serviços e bens duráveis constituem 86.03% dos bons pagadores. Os segmentos de serviços recorrentes, viagens e entretenimento e posto, por outro lado, constituem apenas 4.76% da totalidade de bons pagadores. Isso pode ser um indicativo de que esses segmentos tendem a ser maus pagadores ou simplesmente há uma menor quantidade de clientes com esse tipo de segmento.



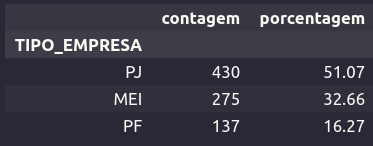
Ademais, há a predominância entre contratos ativos e quitados entre os bons pagadores - o que já era esperado.



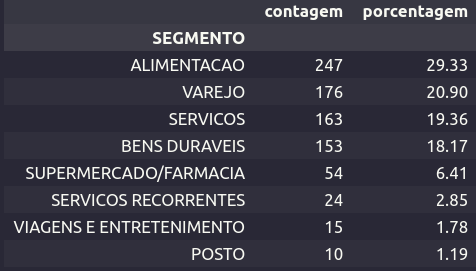
Por fim, em termos absolutos, os bons pagadores estão concentrados no eixo Rio - São Paulo. Esse resultado já era esperado, dado a elevada densidade populacional nesses locais. Uma análise interessante a ser feita em trabalhos futuros, seria analisar a porcentagem de bons e maus pagadores associados a cada estado. Dessa forma seria possível indicar quais estados de fato costumam pagar melhor ou não.



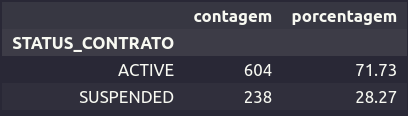
Quanto aos maus pagadores, a distribuição de tipos de empresa se assemelha à dos bons pagadores no que se refere ao tipo de empresa - apresentando apenas pequenas variações em termos percentuais.



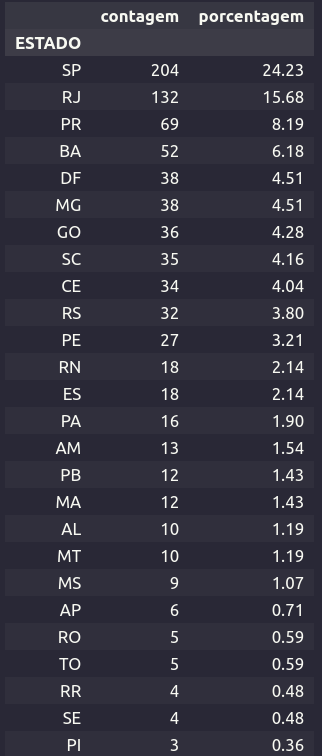
O mesmo acontece com segmento, percentualmente os resultados são parecidos, porém diferem consideravelmente em termos absolutos.



No status de contrato é possível observar a primeira diferença significativa em termos percentuais, pois o que outrora era uma distribuição equilibrada entre o status ativo e o status associado a cada perfil de pagador (quitado ou suspenso) agora apresenta uma predominância de status ativo.



Por fim, mais uma vez a distribuição se assemelha entre bons e maus pagadores - dessa vez em relação ao estado. O eixo Rio - São Paulo segue representando a porção majoritária, seguido de outras regiões metropolitanas.



## Conclusão:

### Qual é a curva ideal de vezes que devemos acionar um cliente?

A partir dos insights, das análises conduzidas e dos perfis de pagadores definidos chegou-se à conclusão de que a quantidade ideal de vezes que um bom pagador deve ser acionado é 6 vezes, enquanto que um mau pagador deve ser acionado 14 vezes. Essas quantidades foram obtidas através de quantis observados em cada perfil, no caso dos bons pagadores foi considerado o 90-percentil e no para os maus pagadores foi considerado o 99-percentil (o percentil dos maus pagadores foi maior, pois considero que eles de fato precisam de mais atenção). De maneira geral, um cliente qualquer deve ser acionado 12 vezes (para isso foi considerado uma média ponderada e um peso maior (7) foi atribuído aos maus pagadores em relação aos bons pagadores (3).