**Validações iniciais**

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

12.202

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

Não. Em nenhuma análise que realizei identifiquei necessidade de excluir registros.

**Etapa 1: AED e limpeza**

Usei SQL para realizar a análise exploratória inicial dos dados (AED), para entendê-los e ver como se cruzavam as tabelas e quais eram os melhores cruzamentos para poder extrair os insights necessários para atender o case. Entre essas análises, verifiquei valores nulos nas tabelas, não encontrando nenhum em nenhuma delas.

**Etapa 2: Preparação de campanhas**

A partir da primeira etapa, verifiquei que uma análise inicial importante, já mirando a solução do problema proposto, seria verificar se havia alguma diferença de engajamento entre as diferentes campanhas.

Para fazer essa análise, foi preciso primeiramente preparar os dados de maneira a diferenciar as campanhas DSP e DSPP, já que há campanhas (campo ‘acao’) que têm o mesmo nome, embora estejam em réguas diferentes – e, portanto, correspondam a momentos bem distintos do negócio, o que torna essa separação fundamental. Além disso, para gerar uma tabela que permitisse a melhor análise, também foi necessário definir adequadamente os campos a serem usados e seus agrupamentos. Assim, fiz uma query com a seguinte lógica (**01\_engajamento\_por\_campanha.sql**):

* Cláusula CASE para diferenciar as campanhas
* JOIN de portfolio\_geral e portfolio\_comunicados com o fim de usar, no CASE, o valor do campo “dsp” e “dspp” junto com o “acao” para que ele pudesse retornar precisamente a qual campanha, especificamente, cada registro se referia.
* Além da campanha, trouxe o campo “tipo\_acao” para posteriormente também poder, na exploração, verificar se havia alguma diferença de engajamento em relação ao canal de disparo
* Agrupamento por campanha, tipo\_acao e status, que permitiria a flexibilidade necessária para, depois, na etapa de exploração, poder fazer os reagrupamentos desejados.

**Etapa 3: Análise de engajamento de campanhas**

Como o número de linhas era pequeno, para maior agilidade fiz a exploração dos dados de campanha (**engajamento\_por\_campanha.csv**) no Google Spreadsheets, usando tabela dinâmica e gráficos simples.

Eu parti de duas hipóteses iniciais:

1. HSM e Email teriam níveis de abertura e resposta diferentes, provavelmente com o primeiro (WhatsApp) tendo maior engajamento
2. Quanto mais avançado o ciclo de inadimplência, menor seria a taxa de engajamento (leitura e resposta)

De imediato, fiquei surpreso ao ver que as duas hipóteses foram refutadas.

Eu já estava preparado para, após esta análise, partir para um aprofundamento da análise no momento da régua em que houvesse uma queda de leitura, ou em relação ao canal com maior engajamento. Como é costume quando os dados refutam hipóteses que consideramos fortes, refiz a preparação e análise diversas vezes e, após confirmar que estava tudo correto, retornei para o enunciado do case e para a AED, buscando quais outros cruzamentos poderiam me levar a insights fortes para o negócio.

O gráfico fruto desta análise foi colocado no topo do dashboard porque é o resultado dele que deriva nas análises seguintes. Além disso, se fosse um dashboard para o time, considero muito importante manter o trackeamento constante deste dado porque pode haver uma mudança de comportamento ao longo do tempo e como é um dado estratégico para a comunicação, é importante reagir rápido caso isso aconteça.

**Etapa 4: Preparação para análise de eficiência das campanhas**

Após uma nova AED, decidi investigar não só o engajamento, mas também ver se os dados poderiam indicar a eficiência de cada uma das réguas (DSP e DSPP) e suas campanhas em termos de gerar pagamentos. Assim, partindo do princípio de que, se o cliente paga, deixa de receber comunicados (ao menos para o contrato em questão), notei que a maneira mais direta de chegar na minha análise era verificando até qual comunicado os ciclos de inadimplência chegavam. Com esse dado, eu poderia verificar quantos (e qual porcentagem de) ciclos se interrompiam depois do primeiro, do segundo, do terceiro, etc, comunicados de cada uma das réguas.

Chamo de “ciclos de inadimplência” cada uma das vezes em que um contrato atinge 5 dias sem pagamento – e, portanto, recebe o primeiro comunicado da régua DSP.

A primeira preparação nesse sentido foi a criação da tabela “ciclos\_comunicados”, puxando o contrato\_id, a campanha (usando os mesmos critérios e nomenclatura da etapa anterior) e usando count para calcular o número de vezes que cada campanha foi enviada (**02\_create\_table\_ciclos\_comunicados**).

Ao mostrar o número de envios de cada campanha, esta tabela ainda não é a ideal para fazer a análise que eu desejava porque uma soma de número de ciclos não me daria a visão que eu buscava – que era onde o ciclo foi finalmente interrompido (pago). Vejamos um exemplo:

| **contrato\_id** | **campanha** | **nr\_ciclos** |
| --- | --- | --- |
| 000c35a61297edadc2842f6d5b4028e1 | DSP\_CONTATO\_01\_05DD | 8 |
| 000c35a61297edadc2842f6d5b4028e1 | DSP\_CONTATO\_02\_10DD | 2 |
| 000c35a61297edadc2842f6d5b4028e1 | DSP\_CONTATO\_03\_15DD | 2 |
| 000c35a61297edadc2842f6d5b4028e1 | DSP\_CONTATO\_04\_30DD | 2 |
| 000c35a61297edadc2842f6d5b4028e1 | DSP\_CONTATO\_05\_60DD | 2 |
| 000c35a61297edadc2842f6d5b4028e1 | DSP\_CONTATO\_06\_90DD | 2 |

A tabela mostra que, para este contrato, o primeiro comunicado DSP foi enviado 8 vezes (ou seja, o contrato ficou inadimplente por 5 dias ou mais em 8 momentos diferentes), e todos os demais da régua DSP foram enviados duas vezes. Portanto, em 6 desses ciclos este contrato fez um pagamento após o primeiro comunicado, e em dois ciclos ele não efetuou pagamento algum antes do último envio (90 dias). Para refletir essa visão para cada contrato, eu precisava ainda criar outra query que gerasse o output dessa forma. De maneira gráfica, eu precisava que o resultado da query fosse assim:

| **contrato\_id** | **campanha** | **nr\_ciclos** |
| --- | --- | --- |
| 000c35a61297edadc2842f6d5b4028e1 | DSP\_CONTATO\_01\_05DD | 6 |
| 000c35a61297edadc2842f6d5b4028e1 | DSP\_CONTATO\_02\_10DD | 0 |
| 000c35a61297edadc2842f6d5b4028e1 | DSP\_CONTATO\_03\_15DD | 0 |
| 000c35a61297edadc2842f6d5b4028e1 | DSP\_CONTATO\_04\_30DD | 0 |
| 000c35a61297edadc2842f6d5b4028e1 | DSP\_CONTATO\_05\_60DD | 0 |
| 000c35a61297edadc2842f6d5b4028e1 | DSP\_CONTATO\_06\_90DD | 2 |

Antes, porém, eu sabia que essa mesma visão poderia ficar ainda mais rica de insights na etapa da exploração se eu criasse segmentações de clientes e ver se haveria comportamentos de pagamentos diferentes dentro desses segmentos. Para isso, além do novo cálculo, seria necessário fazer um JOIN da tabela ciclos\_comunicados com a tabela portfolio\_clientes.

**Etapa 5: Cruzamento com a tabela portfolio\_clientes**

Houve um problema nesta etapa. Por alguma razão que até agora não identifiquei, não consegui montar adequadamente o JOIN da tabela ciclos\_comunicados com a tabela portfolio\_clientes. Mesmo fazendo o JOIN pelas chaves da tabela porfolio\_geral (onde eu precisava buscar o campo “nr\_documento” a fim de conectar com a portfolio\_clientes), ele constantemente gerava uma tabela de mais de 36 milhões de linhas (e que demorava mais de 30 minutos para ser gerada em minha máquina – algo estava, definitivamente, errado).

Após algumas tantas tentativas e pesquisas no StackOverflow que não resolveram o problema, optei por montar uma solução alternativa criando uma tabela que tivesse apenas o contrato\_id e o nr\_documento (**03\_create\_table\_contrato\_id\_nr\_documento.sql**). Embora isto pudesse ser uma tabela temporária dentro de outra query, optei por criá-la e gerar o script em separado para poder deixar a explicação mais clara neste documento.

Enfim, após criada essa tabela, pude fazer a combinação da ciclos\_comunicados com a portfolio\_clientes, gerando uma tabela com todas as colunas de ambas, que chamei de ciclos\_clientes (**04\_create table\_ciclos\_clientes.sql**).

Esta tabela é, finalmente, a que usei para a análise final.

**Etapa 6: Tabela final e exploração**

Uma vez superadas todas as etapas acima, a tabela final é gerada pelo script **05\_calculo\_ciclos\_clientes.sql.** Nele, crio primeiramente a tabela temporária calculo\_ciclos, que equivale à tabela ciclos\_clientes, adicionando uma coluna que puxa o número de ciclos da linha seguinte. Ela é gerada a partir da união de duas tabelas, uma para cada régua – se fosse gerada uma única tabela, sem a cláusula WHERE, ele não quebraria o cálculo quando ocorre a mudança de régua (DSPP para DSP), o que geraria um cálculo equivocado. O último select do script ainda elimina os valores nulos gerados pelo PARTITION BY.

Após todos esses passos, cria-se então a tabela que foi exportada no arquivo **engajamento\_por\_campanha.csv** e que levei para explorar no DataStudio, onde também montei o dashboard apresentado.

A exploração permite um número muito grande de insights. Inclusive, o próprio dashboard, se fosse pensado para o usuário final (e não para entrega em PDF), poderia incluir diversos filtros dinâmicos que permitiriam uma análise customizada. Seguem alguns dos insights que tive a partir da exploração:

**Gráficos Engajamento de campanha**

* Não há diferença de engajamento (entrega, leitura e resposta) dos comunicados ao longo das réguas de relacionamento (insight ainda da etapa 2)

**Gráficos Ciclos de inadimplência**

* Os números da régua de DSP parecem mostrar uma eficiência dos comunicados ou, ao menos, uma relação de Pareto razoável para o trabalho de comunicação: cerca de 75% dos pagamentos são retomados antes de 15 dias e com apenas dois dos seis comunicados enviados. A porcentagem cresce para 87% se considerarmos 30 dias (1 mês). Avalio que é uma performance razoável.
* O mesmo não se pode dizer da régua de DSPP. Nesta, sim, é evidente a **necessidade de ajustes**, já que mais de 40% dos clientes não retomam o pagamento do Principal mesmo após 45 dias e três comunicados – sendo que o segundo comunicado, aos 30 dias, já mostra uma eficiência muito baixa em retomar os pagamentos.

Esta régua tem menos comunicados e com períodos maiores. **Minha sugestão** seria puxar os comunicados para mais próximo da lógica da régua DSP, com mais comunicados no início da inadimplência do Principal. Note-se que o primeiro comunicado do DSPP ocorre apenas após 15 dias de inadimplência. Aos 15 dias, a régua DSP já enviou três comunicados e recuperou 75% dos inadimplentes.

**Gráfico estratificados DSP**

* Os **segmentos** de Posto e Supermercado/Farmácia são os que costumam enfrentar maiores problemas para pagamento:
  + Posto tem a maior proporção (7%) entre os que ultrapassam 90 dias de inadimplência (e, portanto, recebem o sexto comunicado). Supermercado/ Farmácia vem em segundo, com 6,55. Ambos acima da média, que entre todos os segmentos é de 5,0%.
  + Supermecado/Farmácia é o que tem menor proporção (57,6%) pagando antes do 10º dia de inadimplência, e Posto vem em segundo, com 58,7%. Ambos abaixo da média, de 62,6%.
* O segmento de Serviços Recorrentes é o que responde melhor à inadimplência, chegando em menor proporção ao final da régua, com apenas 4% dos inadimplentes chegando ao quinto comunicado e 3% ao sexto. A média dos clientes é de 6,8 (quinto) e 5% (sexto).
* Na estratificação **geográfica**, o destaque é o Acre, que claramente é o estado onde os inadimplentes realizam pagamentos mais rapidamente – 77,3% deles pagam após o primeiro comunicado, sem chegar ao segundo. O número é quase 15% maior que a média (62,6).
* No caso dos **tipos de empresa**, há um equilíbrio muito grande, com os três tipos apresentando números semelhantes.

**Gráficos estratificados DSPP**

* No pagamento do Principal, os **segmentos** de Alimentação, Viagens e Entretenimento e Outros são os que mais chegam ao final da régua (os três na faixa de 48%).
* Na **geografia**, o Acre novamente é o que costuma fazer pagamentos do principal mais rapidamente: 57% pagam entre o primeiro e o segundo comunicados. O número é 15% maior do que a média. O destaque negativo é do Amazonas, onde 47% chegam ao 45º dia sem pagamento.
* No **tipo de empresa**, novamente há um equilíbrio.

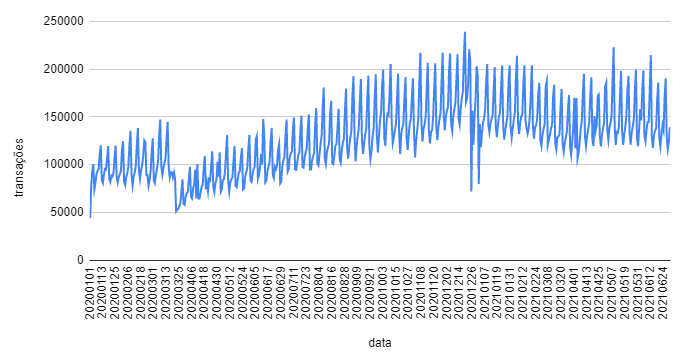
Ao identificar os segmentos e estados com maiores dificuldades, **minha sugestão** é que se use esta estratificação para criar comunicações (ou condições de pagamento) específicas, direcionadas apenas a estes clientes.

**Etapa extra: Impacto da pandemia nas transações da maquininha**

Embora eu tenha acabado por não usar a tabela portfolio\_tpv nas análises que fiz para responder a questão central do case (engajamento das campanhas e seus impactos nos pagamentos), em um dado momento eu me perguntei quanto a pandemia teria afetado o número de transações de cartão de crédito.

O resultado é o gráfico abaixo, onde fica evidente a queda abrupta que ocorreu em março/2020 e uma recuperação para patamares anteriores apenas em julho-agosto do mesmo ano. É também curioso observar os imensos vales que acontecem nos dias 25 de dezembro e 1º de janeiro.

A curva também evidencia um crescimento ao longo do segundo semestre de 2020, inclusive a patamares superiores aos de antes da pandemia. Penso em duas hipóteses: o aumento no número absoluto de clientes da Stone ou um volume maior de transações impulsionado pelo crescimento do comércio online.



**SEMIFINAL**

**Etapa 1: passo atrás**

O email que comunicava a aprovação para a semifinal deixava claro que, pela parte dos dados, era preciso investir pesado na resposta para a pergunta principal e também no cruzamento entre os acionamentos e os não-pagamentos dos contratos (desengajamentos). Dei, então, um passo atrás e voltei um pouco para a AED, criando inclusive rascunhos manuais dos cruzamentos possíveis. De repente, me perguntei quantas pessoas entravam não no ciclo da régua de campanhas (como investi na primeira entrega), mas sim no ciclo de inadimplência. Ou seja, quantas vezes os contratos ficavam 1 dia sem pagamento? E 2? E 3? Embora simples por usar praticamente só 1 coluna (dsp) de uma única tabela (portfolio\_geral), percebi que esse poderia ser um caminho. Contruí, então, a query **sf\_dias\_sem\_pagamento.sql** e gerei o csv de mesmo nome.

Fiz diversas análises no Google Sheets, onde encontrei o insight de trazer a régua para o DSP 3. Mas talvez mais difícil do que encontrar o insight em si tenha sido chegar a um dataviz bem resolvido, que ajudasse a conduzir para essa conclusão. Finalmente – inspirado pela ideia de conduzir o storytelling pela analogia com a pandemia e o “achatar a curva” – consegui chegar no gráfico que abre o meu dashboard.

**Etapa 2: reconstrução do dashboard**

Após a orientação do documento enviado e, também, do workshop, optei por:

1. Focar nas curvas e índices que guiam a resposta principal do case, ou seja, gráficos que demonstrem por que a curva deve-se concentrar entre os DSP 3 e 10.
2. Reduzir o número de valores da dimensão “campanha” nos gráficos de barras. Isso facilitou muito a visualização dos dados, aumentando o impacto e conduzindo melhor o olhar para os insights

**Etapa 3: construção da narrativa da apresentação**

Baseado na importância do storytelling e dos primeiros 90 segundos – aspectos muito bem abordados no workshop –, reconstruí a apresentação, mirando o público da banca. Busquei:

1. Uma analogia com a pandemia que busca ser o esqueleto condutor da história dos dados e traçar um paralelo com o próprio problema apresentado no case, que tem a ver com o aumento de inadimplência justamente durante a pandemia. É uma espécie de meta-história que espero que dê resultado e ajude a cativar o público e a destacar a apresentação. A banca, na avaliação após a apresentação, poderá não lembrar quem é Juan, mas espero que lembre quem é o “cara do ‘achatar a curva’”
2. O gráfico inicial, do Google Trends, que tenta surpreender e captar a atenção do público logo de início. Por ser uma banca técnica de dados, acredito que trazer um gráfico como elemento-surpresa cative e, de uma certa maneira, me confie autoridade. Se o apresentador (eu, no caso) surpreende o público com um elemento que faz parte do território que o público domina (o de gráficos), então é mais provável que esse público levante mais a orelha para o que o apresentador tem a dizer a seguir.

**Etapa 4: a vacina**

Um insight que tive ainda na exploração foi ver quanto tempo os contratos demoravam, em média entre o começo do contrato e a inadimplência ou entre sair de um ciclo de inadimplência e entrar em outro. Fiz, então, a query **sf\_med\_dias\_antes\_inadimplencia.sql** que traz a mediana dessa análise. Ao ver que 50% dos contratos ficam apenas 5 dias adimplentes, percebi que isso poderia ter relação com o dia da semana e, já finalizando o prazo para entrega, fui construir a query **sf\_dsp\_dia\_da\_semana.sql**, que validou a hipótese: o dsp 1 mais comum é o sábado e o dsp 2 é o domingo.Por isso, o número de inadimplência cai tanto no DSP 3. Os estabelecimentos reabrem e voltam a passar a maquininha, impedindo que eles cheguem no DSP 3. Esses últimos achados só reforçam a necessidade de passar um comunicado a todos que atingirem DSP 3.