Documento de Entendimento

Sprint Session 4 - Ana Health
Grupo - ANA CODA

Leonardo Scarlato, Matheus Aguiar, João Alfredo, Gustavo Antony, Erik Soares, Marcelo Marchetto

A empresa parceira nessa Sprint Session é a Ana Health, uma startup no ramo da saúde que funciona como uma clínica online. Ela atende tanto pessoas físicas quanto empresas, agindo como benefício nesse último caso. A empresa tem aproximadamente dois (2) anos de existência, sendo fundada em 2021.

O problema que enfrentaremos neste projeto está relacionado ao *churn*, uma métrica de perda de clientes pela empresa. Nesse caso, estamos interessados em observar quais os tipos de clientes que são mais propícios a cancelar ou desistir do plano.

Desafios envolvendo o DataSet

O DataSet possui 73 colunas e 1202 amostras (linhas), com informações que vão desde a data de início do contrato até a quantidade de mensagens da equipe que foram direcionadas ao usuário. Cada linha de usuário pode ser uma pessoa física, pessoa jurídica ou acolhimento desemprego.

Ademais, percebemos que muitas colunas contam com valores nulos, e outras com diversos valores para a mesma célula. Isso é provavelmente o resultado de uma junção de diferentes tabelas, problemas esses que teremos que lidar para que não atrapalhe o nosso algoritmo final. Além disso, algumas colunas aparentam estar com o tipo de dado diferente do que é esperado, como por exemplo o status de casado (que é um número inteiro), ou valores binários em colunas que representam quantidades (como é o caso da coluna de datas de atendimentos médicos).

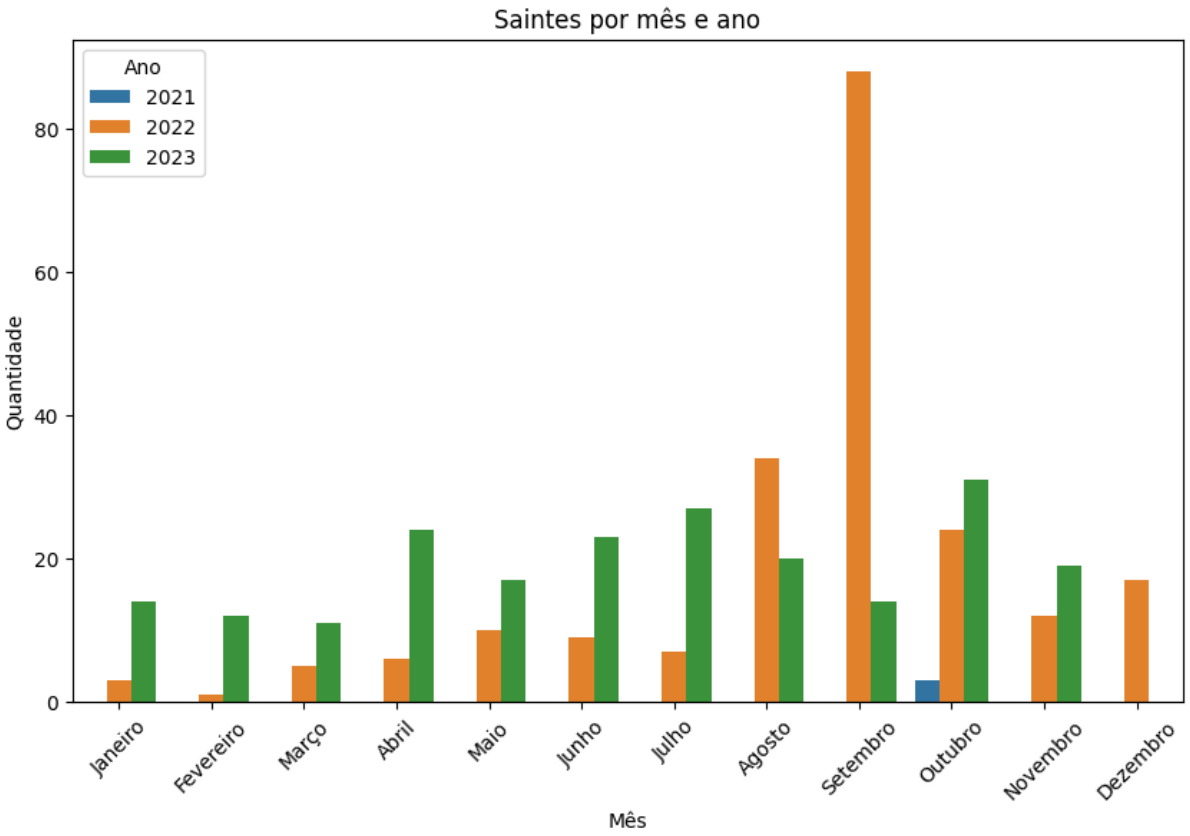
Quando tratamos das pessoas físicas, na maioria das vezes, não observamos o problema relacionado a diversos valores para uma mesma célula. As únicas, possíveis exceções, são relacionadas a notas do teste de WHOQOL (teste de qualidade de vida da World Health Organization), onde uma pessoa pode fazer o mesmo teste mais de uma vez, portanto, tendo mais de uma nota.

Em relação a empresas, podemos diferenciá-las de duas principais maneiras: o valor da célula de *id_org* náo é nulo e há mais de um valor (separado por ponto e vírgula) na coluna de *id_stage*.

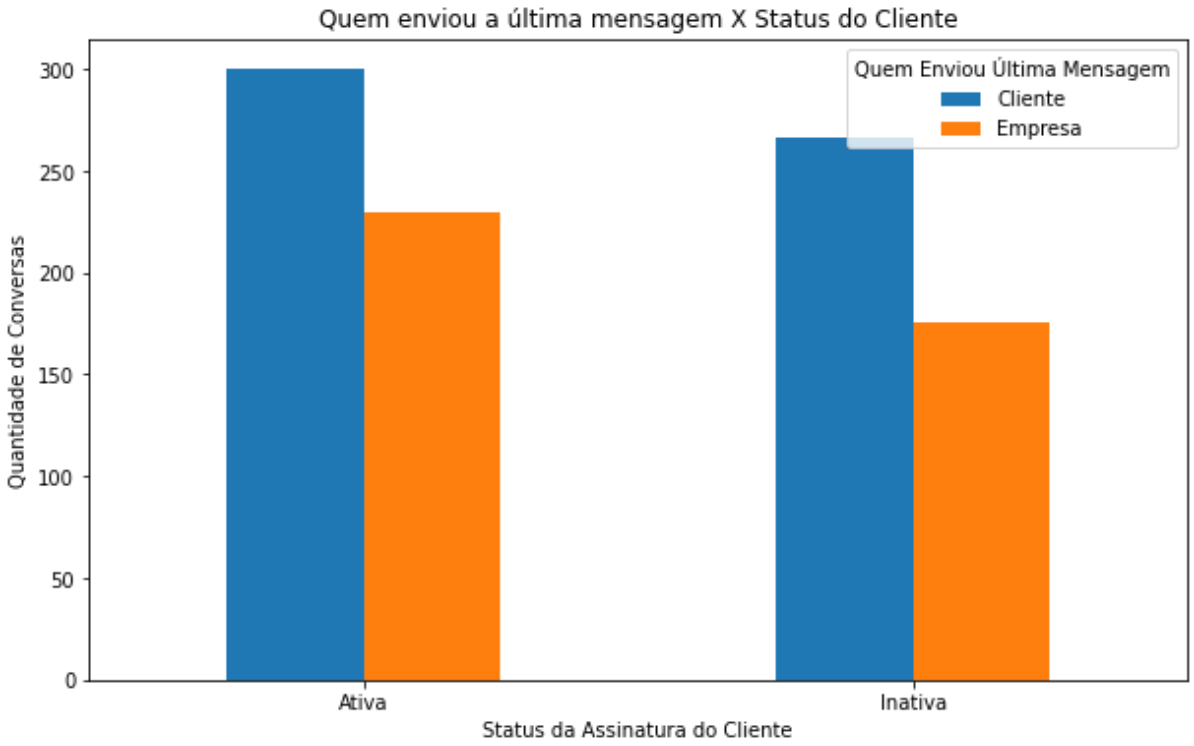
Quatro colunas principais nos indicam possível cancelamento. A primeira é a *contract_end_date*, onde a presença de uma data significa que houve um cancelamento de contrato por parte do cliente. Esse cancelamento pode ser definitivo ou mudança de plano. O status da assinatura nos diz se a assinatura do cliente está em atividade (won) ou se foi interrompida (lost). Há também o *status* do processo de onboarding, que nos diz se o processo foi concluído ou não (algo importante já que também são contemplados como perda de cliente). Por último, a coluna de *lost_time* nos traz informações sobre as datas de cancelamento do cliente, uma para cada assinatura que foi cancelada.

Além dessas, temos também que considerar possíveis faltas de sessões (psicoterapia, atendimento médico e acolhimento) e seu números, se a última mensagem entre o cliente e a equipe de saúde foi feita pela equipe de saúde (na situação onde o cliente for o último a ter enviado a mensagem isso pode indicar uma falha na hora de atender ao pedido), tempo de resposta médio da equipe de saúde e as médias dos teste de WHOQOL (além das diferenças entre testes consecutivos).

Durante a nossa análise exploratória criamos dois gráficos que nos ajudam a entender o comportamento dos clientes da Ana Health. Sendo eles um gráfico que mostra quantos clientes cancelaram o plano a cada mês, e outro que procura uma relação do status da assinatura com o fato do cliente ter sido respondido pela empresa na sua última mensagem ou não.



No gráfico acima, podemos concluir algumas coisas. A mais gritante é a falta de saíntes (pessoas que cancelaram o plano) no ano de 2021. Outra coisa é que no mês de setembro de 2022 houve um aumento expressivo no número de pessoas que cancelaram o seu plano com a Ana Health. Isso já foi explicado durante uma das conversas como a saída de uma empresa parceira. Apesar disso, nos demais meses não observamos uma divergência grande no número de saintes nesses períodos. O ano atual apresenta uma taxa de saíntes mais consistente quando comparada aos demais, com maiores picos em outubro, julho e abril.



Ao propor a análise do gráfico acima acreditávamos que havia a existência de uma correlação entre as mensagens dos usuários serem respondidas e a permanência deles no plano. No entanto, observamos que não é possível chegar nessa conclusão com base nos dados que possuímos por alguns motivos. Ambos os possíveis status do cliente tem uma distribuição parecida, além de que não conseguimos inferir o conteúdo dessa última mensagem. Dessa forma, no momento não encontramos evidências que indiquem que o fato de ter tido sua última mensagem respondida ou não seja um dado relevante para dizermos se o usuário vai cancelar o seu plano.