Relatório – Projeto 01

Título do Projeto: O Mercado.

Objetivo: O objetivo desse projeto é realizar a análise exploratória da base de clientes da loja, com foco na segmentação destes clientes utilizando a metodologia RFM, a fim de entender o comportamento de compra destes clientes, identificar padrões, e assim obter insights valiosos a partir dos dados e usá-los para criar planos de ação estratégicos e eficientes.

Responsável pelo Projeto: Gabriela Ferreira Genangelo.

Ferramentas e Tecnologias Utilizadas:

1. Canva
2. Google Docs
3. Google Sheets
4. Google Slides
5. Lighthouse Gauge
6. Looker Studio
7. Loom
8. OpenAI

Processamento e Análises:

5.1.1

Para importar os arquivos criei 3 seções: *clientes, transacoes e resumo_compras.*

Em **clientes** utilizei a fórmula:

=IMPORTRANGE("https://docs.google.com/spreadsheets/d/1-vihlyqeU6WRkGx0PA8YWluLpscjzd9UV71e0hdX998/edit?gid=0#gid=0";"CLIENTES!A:I")

Em **transacoes** utilizei:

=importrange("https://docs.google.com/spreadsheets/d/1pDww7Ofa81VnIC8W88x-uhLxBsIcZoLEZ2intDh5g0U/edit?gid=0#gid=0";"transacoes!A:D")

E em **resumo_compras**:

=IMPORTRANGE("https://docs.google.com/spreadsheets/d/1_YwdkTqhVFRjrrO6lQU_GutMCe5OI7MQswQ9nNqvtfc/edit?gid=1803379532#gid=1803379532";"resumo_compras!A:G")

Após realizar as importações por meio da função **=IMPORTRANGE**, concedi as permissões de acesso necessárias para dar início ao tratamento dos dados.

Cheguei a testar a função **=IMPORTRANGE** com **=QUERY,** deu certo, mas, achei melhor seguir somente com o =**IMPORTRANGE** para de início, ter uma visão mais ampla dos dados.

5.1.2

Para encontrar os valores nulos eu utilizei em cada aba a função:

=Countblank (contar.vazio) Iniciei aplicando a fórmula de contagem de valores nulos em todo o conjunto de colunas e, conforme fui identificando a presença de dados ausentes, refinei a análise utilizando a fórmula individualmente em cada coluna. Realizei esse processo em todas as seções. Isso permitiu localizar exatamente em quais seções e campos os valores nulos estavam concentrados, facilitando a análise e o tratamento posterior.

Na **aba clientes**, com a fórmula: **=CONTAR.VAZIO(E2:E2231)** Identifiquei 24 salários nulos.

Decidi não excluir os dados de salário nulos, optando por ignorá-los temporariamente, pois essa informação ausente não comprometeria minha análise no momento. Futuramente, poderia revisar essa questão com mais cuidado e, se fosse possível eu poderia, imputar os valores utilizando a mediana com base no nível de escolaridade dos clientes correspondentes.

Na **aba transacoes**, com a fórmula: **=CONTAR.VAZIO(B2:B22128)** encontrei 7 id_clientes nulos.

Inicialmente, tentei encontrar alguma relação entre esses registros e a data de cadastro dos clientes, considerando a hipótese de que essas transações poderiam ter ocorrido no mesmo dia do cadastro. No entanto, não foi identificada nenhuma correspondência.

Também analisei a possibilidade de relacionar esses IDs ausentes com os clientes presentes no resumo de compras, partindo da ideia de que, se houve uma transação, o cliente correspondente deveria estar registrado lá. Novamente, não houve correspondência.

Diante disso, optei por remover estes sete ID_cliente nulos. Pois, por se tratarem de registros na aba de transações, onde um mesmo cliente pode ter realizado várias compras, manter esses dados sem um identificador confiável comprometeria toda a análise. Gerar um número aleatório ou criar um nome fictício para esses clientes não garantiria que as transações pertencem a um único cliente, nem permitiria saber quantas compras cada um realmente realizou. Isso impactaria métricas como média, soma e segmentações por cliente, distorcendo os resultados.

Caso esses registros estivessem na aba de clientes, onde cada entrada representa um cliente único, até seria possível considerar a criação de um identificador artificial para

manter o dado. Mas não é o caso aqui, e por isso a exclusão foi a decisão mais adequada para preservar a integridade da análise.

Para limpar estes nulos cheguei a testar o **=importrange** junto com a função **=filter**, ficando a formula:
=filter(importrange("https://docs.google.com/spreadsheets/d/1pDww7Ofa81VnIC8W88x-uhLxBsIcZoLEZ2intDh5g0U/edit?gid=0#gid=0";"transacoes!A:D");(transacoes!B1:B22128<>"")) e também testei o **=importrange** junto com o **=query**, onde utilizei a formula:
=query(importrange("https://docs.google.com/spreadsheets/d/1pDww7Ofa81VnIC8W88x-uhLxBsIcZoLEZ2intDh5g0U/edit?gid=0#gid=0";"transacoes!A:D");"Select Col1,Col2,Col3, Col4 where Col2 is not null";1)

ambas deram certo, mas, acabei optando por abrir uma nova aba, na qual nomeei: **filter_transacoes** e utilizei a fórmula:

=FILTER(transacoes!A1:D22128; transacoes!B1:B22128<>"") Onde os dados foram importados para uma nova aba/planilha, sem os valores nulos**.**

Já em **resumo_compras**, a fórmula me trouxe também os valores de faturamento que estavam como $0, mas, por se tratarem de um valor real, indicando que o cliente de fato não realizou compras em determinado segmento de produto, não considerei estes como nulos.

Em todos os nulos encontrados, eu utilizei a formatação condicional para destacar visualmente as células vazias nas colunas analisadas.

5.1.3

Para identificar as duplicidades acessei a opção “Dados” -> “Limpeza de dados” -> “Remover cópias” para ter uma base das informações que iria retornar, fiz isso nas 3 seções, e somente a aba **resumo_ de_compras** me retornou com a mensagem de que: *“9 linhas duplicadas foram encontradas e removidas”* como eu utilizei a função =importrange quando importei inicialmente os dados, estas duplicidades que supostamente foram encontradas e removidas voltaram para a base.

Utilizei a fórmula **=arrayformula(CONT.SE(A2:A2250;A2:A2250))** para contar quantas vezes cada valor no intervalo de A2:A2250 aparecia dentro do mesmo intervalo, onde se o número fosse maior que 1 significaria que o dado estaria duplicado.

Como já esperado, a fórmula me retornou identificando os nove IDs com registros duplicados. Utilizei a formatação condicional com a fórmula **=CONT.SE(A:A;A2)>1**, destacando os valores duplicados em amarelo para facilitar a comparação.
Como de fato, estes dados eram idênticos em todos os campos, manter as duplicidades poderia comprometer a integridade da análise, gerando distorções e inflando os resultados de forma indevida.

Sendo assim, criei uma nova aba nomeada de: **resumo_compras_unique**, onde na **coluna A** apliquei a fórmula **=UNIQUE(resumo_compras!A1:G2250)** para importar os dados removendo as duplicidades e assim mantendo apenas os registros únicos.
Aproveitei esta nova aba e criei uma nova variável, utilizando a fórmula:
=ARRAYFORMULA(B2:B2241+C2:C2241+D2:D2241+E2:E2241+F2:F2241+G2:G2241)
coloquei para formatar a coluna em dólar americano e assim eu trouxe para a variável: **total_gasto_por_cliente_** o valor total que cada clientes havia gasto na loja, e consegui analisar se havia algum cliente que constava em resumo_compras_unique sem ter realizado compras, no entanto, todos os clientes listados nessa aba apresentavam registros de compras, ou seja, de fato realizaram transações.
Também fiz uma tabela com a informação de quantos clientes não compraram tal segmento de produto, para ter uma base de comportamento de compra, utilizei a função **=cont.se** a fim de contar quantas vezes o valor $0 aparecia em cada segmento, representando a ausência de compras.

5.1.4

Em relação aos dados fora do escopo, mencionados no Checkpoint da guia, eu já havia eliminado estes registros durante o processo de tratamento dos valores nulos.

Além destes que eu já havia removido, associei os clientes que estavam em resumo_compras_unique com as transações, utilizando a fórmula:
=SE(ÉERROS(PROCV(A2;transacoes!B:B;1;FALSO));"Não encontrado";"Encontrado")
Descobri que 10 clientes que compraram determinados produtos, estavam sem dados de transação, cheguei a analisar se estes poderiam estar associados com os 7 id_cliente nulos que eu havia removido da aba transações, mas, na análise não encontrei nenhuma associação.

Decidi considerar esses registros como cadastros e transações de clientes realizadas antes do período analisado, tratando-os como clientes mais antigos. Aplicar imputação nesses casos não seria adequado, pois faltavam dados cruciais para a análise, e, referente a algumas informações, a estimativa seria imprecisa, especialmente considerando que o mesmo cliente poderia ter realizado mais de uma transação, o que comprometeria a consistência dos resultados.

Neste caso utilizei a fórmula: **=UNIQUE(FILTER(resumo_compras!A1:G2250; ÉNÚM(CORRESP(resumo_compras!A1:A2250; transacoes!B:B; 0))))** para deixar apenas os dados dos clientes que constavam tanto em resumo de compras quanto em transacoes e fiz o mesmo com a aba clientes, utilizando a fórmula:

=FILTER(IMPORTRANGE("https://docs.google.com/spreadsheets/d/1-vihlyqeU6WRkGx0PA8YWluLpscjzd9UV71e0hdX998/edit?gid=0#gid=0"; "CLIENTES!A1:I2241"); ÉNÚM(CORRESP(IMPORTRANGE("https://docs.google.com/spreadsheets/d/1-vihlyqe

U6WRkGx0PA8YWluLpscjzd9UV71e0hdX998/edit?gid=0#gid=0"; "CLIENTES!A1:A2241"); transacoes!B:B; 0)))

Ajustei algumas fórmulas.

E identifiquei **três clientes com anos de nascimento discrepantes**: **ID_cliente 11004, 1150 e 7829**. Nesse caso, ou as **informações de data de nascimento** desses clientes foram preenchidas incorretamente, ou é possível que **terceiros, como filhos, parentes ou outras pessoas**, tenham realizado compras utilizando os dados desses clientes.
Para esses, realizei a correção com base na **mediana do ano de nascimento correspondente ao nível de escolaridade** de cada cliente, utilizando as seguintes fórmulas:

1°) **=ÍNDICE(B2:B2231; CORRESP(MÍNIMO(SE(B2:B2231 > 1925; B2:B2231)); B2:B2231; 0))** Que identificou o menor ano de nascimento válido como sendo: 1940, considerando 100 anos.

2°) **=ÍNDICE(B2:B2231; MENOR(SE(B2:B2231<1940; LIN(B2:B2231)-LIN(B2)+1); LIN(A1)))** Que identificou os clientes com ano de nascimento outliers, nascidos antes de 1940, foram encontrados: 1900, 1893 e 1899 eu já tinha analisado essas informações pelo filtro, mas, aproveitei para testar a fórmula.

3°) **=PROCV(P2; B:C; 2; FALSO)** Que identificou a escolaridade desses clientes outliers, sendo 2 deles com escolaridade Secundaria (1900 e 1893), e 1 cliente Posgrado (1899).

4°) **=SEERRO(QUERY(A2:C2231; "SELECT B WHERE C = 'Posgrado' AND B > 1939"; 1); "")** Puxou os anos de nascimento válidos dos clientes com 'Posgrado'."

5°) **=MED(R2:R854)** A mediana dos anos de nascimento dos clientes com 'Posgrado' foi calculada, resultando no valor de 1968

6°) **=SEERRO(QUERY(A2:C2231; "SELECT B WHERE C = 'Secundaria' AND B > 1939"; 1); "")** Puxou os anos de nascimento válidos dos clientes “Secundaria”

7°) **=MED(T2:T257)** A mediana dos anos de nascimento dos clientes com nível de escolaridade “Secundaria” foi calculada, resultando no valor de 1975.

Ou seja, para os dois clientes que estavam com nível de educação 'secundária', a fórmula me trouxe o ano de 1975, e para o cliente 'Posgrado', a fórmula me trouxe o ano de 1968. Assim, segui com a “imputação” desses dados. Na verdade, usei uma fórmula condicional para modificar esses dados. A fórmula utilizada a princípio foi: **=SE(B2=1893; 1975; SE(B2=1900; 1975; SE(B2=1899; 1968; B2)))**. Com a fórmula aplicada na coluna K, a variável **(ano_nascimento_outliers_corrigido),** retornou os valores nas células com os dados ajustados da forma que eu precisava para as análises futuras.

Filtrei os dados da variável salario_anual_dolar para análise e, com isso, identifiquei o **Id_cliente 9432** com um salário que considerei outlier.
Calculei a mediana com base nos outros clientes com o mesmo nível de escolaridade (grado o superior), utilizando a fórmula: **=MED(Z2:Z1109)** e encontrei a mediana de

$52,095.50, visto que também encontrei 11 “Grado o superior” nulos, utilizei essa mesma informação para corrigir estes nulos. Calculei a mediana dos salarios anuais de posgrado utilizando a fórmula: **=MED(X2:X844)** onde a mediana calculada foi de **$54.108,00** que utilizei para corrigir os 10 nulos que tinha como escolaridade “posgrado”. Com a fórmula: **=MED(V2:V255)** calculei a mediana relacionada aos clientes com nível de escolaridade “Secundaria” para estes encontrei o valor de **$36.223,00** e existiam 3 salarios nulos.

Nomeei a coluna L da aba clientes como: salario_anual_outliers_corrigidos e utilizei a principio a formula condicional:

=SE(OU(E2=""; E2=0); SE(C2="Posgrado"; 54108; SE(C2="Grado o superior"; 52095,5; SE(C2="Secundaria"; 36223; E2))); SE(E2>170000; 52095,5; E2))

A fórmula retornou as células com os 24 valores nulos corrigidos, com base nas medianas correspondentes ao nível de escolaridade, além da correção do salário anual do **ID_cliente 9432.** Os dados que não exigiam ajustes retornaram inalterados.

5.1.5

Antes de unir as tabelas, decidi agrupar dados importantes.

Na aba clientes utilizei a fórmula: **=QUERY(A2:I2241;"SELECT A WHERE I > 0")** para saber somente quem eram os clientes que tinham aceitado a campanha de marketing, nomeei a variável de: **clientes_que_aceitaram_a_campanha.**

Na coluna J, nomeei a variável como: **total_filhos** e utilizei a fórmula: **=SOMA(F2;G2)** para calcular o total de filhos de cada cliente.

E utilizei a fórmula: **=SE(I2=0; "Não aceitou a Campanha"; SE(I2=1; "Aceitou a Campanha"))** na coluna M para retornar os dados da campanha de forma nominal, e nomeei esta nova variável como: **campanha_de_marketing.**

Em **filter_transacoes** utilizei a fórmula: **=CONT.SE(B:B; B2)** para agrupar a quantidade de transações realizadas por cliente, **=CONT.SES(B:B; B2; D:D; "en línea")** para contar quantas vezes cada cliente comprou em loja online e: **=CONT.SES(B:B; B2; D:D; "tienda")** quantas vezes comprou na loja física.

Também utilizei a formúla: **=ArrayFormula(ÍNDICE(C:C; CORRESP(MÁXIMO(SE(B:B=B2; C:C)); C:C; 0)))** para retornar a data da última compra de cada cliente, identificando a transação mais recente com base no histórico individual.

A **chave para unir tabelas** que eu identifiquei foi o **Id_cliente.**

Abri uma nova aba, na qual nomeei de: **dados_consolidados_iniciais.**

1°) Com a fórmula: **=QUERY(clientes!A1:M2231; "SELECT A,K,C,D,L,F,G,H,M,J";1)**

Preenchendo da coluna A até a coluna J, filtrei as seguintes variáveis:

Id_cliente, ano_nascimento_outliers_corrigido, nivel_educacao, estado_civil, salario_anual_outliers_corrigidos, criancas_ate_dez_anos, criancas_mais_dez_anos data_entrada, campanha_de_marketing, total_filhos.

Os cabeçalhos das colunas, que representam as variáveis que criei, foram formatados em negrito para facilitar a análise e a identificação das informações.

No início, eu tinha pensando em unir apenas as variáveis que faziam sentido para a análise **RFM**, mas, como o próximo passo seria a análise exploratória, a unificação de todos os dados principais dos clientes seria perfeita para analisar alguns erros que ainda poderiam existir, e visualidade de uma forma mais ampla.

2°) Com a fórmula: **=ÍNDICE(filter_transacoes!E:E; CORRESP(A2; filter_transacoes!B:B; 0))** atribui à **coluna K** a variável: **quantidade_de_transacao_realizada** por cada cliente.

3°) Utilizando a fórmula: **=ÍNDICE(filter_transacoes!F:F; CORRESP(A2; filter_transacoes!B:B; 0))** atribuí a **coluna L** a quantidade de **compras_online** que cada cliente realizou e com a fórmula: **=ÍNDICE(filter_transacoes!G:G; CORRESP(A2; filter_transacoes!B:B; 0))** na **coluna M** a quantidade de **compras_loja_fisica**.

4°) Usando a fórmula: **=ÍNDICE(filter_transacoes!H:H; CORRESP(A2; filter_transacoes!B:B; 0))** atribuí à coluna **N** os dados da **data_da_ultima_transacao** por cliente.

5°) Com as fórmulas:

=ÍNDICE(resumo_compras_unique!B:B; CORRESP(A2; resumo_compras_unique!A:A; 0))

=ÍNDICE(resumo_compras_unique!C:C; CORRESP(A2; resumo_compras_unique!A:A; 0))

=ÍNDICE(resumo_compras_unique!D:D; CORRESP(A2; resumo_compras_unique!A:A; 0))

=ÍNDICE(resumo_compras_unique!E:E; CORRESP(A2; resumo_compras_unique!A:A; 0))

=ÍNDICE(resumo_compras_unique!F:F; CORRESP(A2; resumo_compras_unique!A:A; 0))

=ÍNDICE(resumo_compras_unique!G:G; CORRESP(A2; resumo_compras_unique!A:A; 0))

Atribuí os dados referentes à quantidade de compras realizadas por cada cliente em cada segmento de produto.

6°) Já com a fórmula: **=ÍNDICE(resumo_compras_unique!H:H; CORRESP(A2; resumo_compras_unique!A:A; 0))** atribuí na **coluna U** os dados do **total_gasto_por_cliente_$**.

5.1.6

Eu já havia criado por meio de fórmulas diversas variáveis, como: total_filhos, quantidade_de_transacao_realizada, compras_online, compras_loja_fisica, data_da_ultima_transacao, total_gasto_por_cliente_ $, dados_de_transacao que utilizei para identificar quais clientes possuíam compras e não tinham as informações de transações, ano_nascimento_outliers_corrigido, salario_anual_outliers_corrigidos, também havia utilizado a função **=MIN** e **=MAX** para identificar o período inicial e final dos dados de transações, data mais antiga e data mais recente.

De modo geral, todas as variáveis criadas foram utilizando fórmulas.

Aproveitei essa etapa do marco, e na aba **resumo_compras_unique**, criei uma variável com a função **=SOMA** para cada **segmento de produto**, com o objetivo de ***calcular*** *o valor* ***total faturado*** *por segmento*, as fórmulas ficaram assim:

Coluna J, variável: total_geral_vinho =SOMA(B2:B2231)

Coluna K, variável: total_geral_frutas =SOMA(C2:C2231)

Coluna L, variável: total_geral_carnes =SOMA(D2:D2231)

Coluna M, variável: total_geral_peixes =SOMA(E2:E2231)

Coluna N, variável: total_geral_doces =SOMA(F2:F2231)

Coluna O, variável: total_geral_outros =SOMA(G2:G2231)

5.2

Na análise exploratória, aba: **dados_consolidados_iniciais**, na **coluna V**, com a fórmula **=QUERY(clientes!A:E; "SELECT E";1)** eu puxei os dados de **salario_anual_dolar** sem as correções realizadas anteriormente.

Na **coluna W** *criei uma variável* **(chave_unica)**, utilizando a fórmula: **=B2 & C2 & D2 & F2 & G2 & H2 & I2 & K2 & L2 & M2 & O2 & P2 & Q2 & R2 & S2 & T2** onde criei uma *chave composta a partir da* ***concatenação*** *de variáveis*. E na **coluna X** com a fórmula: **=SE(CONT.SE($W$2:$W$2231; W2) > 1; "Duplicado"; "Unico")** consegui identificar registros *possivelmente duplicados*. Em seguida, criei uma aba nomeada **'Análise de Duplicados com Diferentes IDs'**, com o objetivo de examinar esses dados com mais profundidade.

Com base nessa análise, também consegui identificar os valores de *quatro registros* que anteriormente estavam com *salários nulos*, eu poderia usar uma **fórmula condicional**, para corrigir estes, mas, como eram poucos dados, resolvi fazer estes, de forma manual. Apaguei a fórmula da coluna V que eu havia feito, retornei na aba cliente, copiei a coluna de salários de origem, (antes da correção com a mediana), e colei eles na aba: dados_consolidados_inicial, coluna V, (4correcoes_salario_anual_dolar), a fórmula retornou com os dados anteriores de salários.

De acordo com a *relação dos possíveis duplicados* que eu havia encontrado, fiz a correção dos **4 salários nulos**. Eles ficaram da seguinte forma:

Id_cliente 8557 = $28567

Id_cliente 8268 para = $54456

Id_cliente 3769 para = $38887

Id_cliente 1612 = $ 36038

***Na seção de **'Limitações e Próximos Passos'** deste relatório, abordarei como foi essa análise com mais detalhes.*

Após isso atualizei a fórmula da chave única da coluna W, incluindo a informação de salários **=B2 & C2 & D2 & F2 & G2 & H2 & I2 & K2 & L2 & M2 & O2 & P2 & Q2 & R2 & S2 & T2 & V2**

E a fórmula da **coluna X (unico_ou_duplicado)**, também foi atualizada.

Depois em uma nova aba, na qual nomeei: **dados_consolidados_base_oficial** utilizei a formula:

=UNIQUE(FILTER(dados_consolidados!A2:V; dados_consolidados!X2:X = "Unico"))

Com o objetivo de *reunir todos os dados unificados*, **excluindo aqueles com chaves únicas duplicadas**, a fórmula me retornou **1.862 ID_cliente**, conforme já era esperado.

Em resumo, nesta etapa, decidi que a melhor alternativa era **excluir** alguns dados que **apresentavam indícios de inconsistência**. Após essa exclusão, **refiz o *cálculo* das *medianas*** que foram utilizadas anteriormente para **imputar** os *anos de nascimento* dos que eu considerei *outliers*, e os valores de *salários nulos*.

a fórmula final que corrigiu os dados de ano nascimento foi aplicada na **aba clientes**, na coluna K, *(ano_nascimento_outliers_corrigido)*, e ficou:

=SE(B2=1893; 1976; SE(B2=1900; 1976; SE(B2=1899; 1968; B2)))

e a que corrigiu os salários nulos, foi aplicada na coluna L *(salario_anual_outliers_corrigidos)* e ficou:

=SE(OU(E2=""; E2=0); SE(C2="Posgrado"; 54222; SE(C2="Grado o superior"; 52848,5; SE(C2="Secundaria"; 36595; E2))); SE(E2>170000; 52848,5; E2))

Durante essa etapa, também observei que alguns segmentos de produtos apresentavam valores de compra bastante baixos. Considerei que isso poderia ser resultado de promoções ou cupons de desconto.

Aproveitei as dicas do oráculo e desenvolvi dois gráficos de pizza para análise de comportamento do cliente: o primeiro compara a porcentagem de compras realizadas na loja física versus na loja online; o segundo ilustra a proporção de clientes que aderiram à campanha de marketing em relação aos que não aderiram, ambos estão na aba: **dados_consolidados_base_oficial.**

5.2.1

5.2.2

Para as metas mencionadas acima, criei algumas variáveis na aba *dados_consolidados_base_oficial*.

- ➔ Na coluna X, criei a variável: **idade_dos_clientes_considerando_2025** e utilizei a fórmula: **=2025 - B2.**
- ➔ Na coluna Y, criei a variável: **faixa_etaria** e utilizei a fórmula: **=SE(X2<18;"Menor de 18";SE(X2<=24;"18-24";SE(X2<=34;"25-34";SE(X2<=44;"35-44";SE(X2<=54;"45-54";SE(X2<=64;"55-64";"65+"))))))**

Cheguei a criar outras variáveis.

Em uma nova aba nomeada ***analise_tabelas_e_graficos***, na qual elaborei diversas tabelas dinâmicas e gráficos para a análise.

Com essas análises, foi possível identificar que o **nível de escolaridade** mais predominante na base analisada é **'Grado o superior'**, com **927 clientes**, o que representa aproximadamente **49,79% do total.**

Em relação à **faixa etária**, a maior parte dos clientes, representando **31,74%** da base tem entre **45 e 54 anos.** O grupo **mais jovem** é o *menos representado*: a faixa etária de **25 a 34 anos** corresponde a apenas **1,83%** da base analisada, o que equivale a **34** clientes

Dos clientes analisados, ***724 são casados****, **411 são solteiros** e 4 dos clientes não possuem nenhum estado civil declarado*, nos dados, o estado civil destes 4 clientes está nomeado como “Otros”

A **renda média anual geral** é de **$52.111,78**, e o **número médio de filhos** por cliente é de **0,96.**

84,32% dos clientes *não participaram da campanha de marketing*, apenas **292** *clientes aceitaram a campanha.*

A **frequência média de compra** geral é **0,39.**

O **valor monetário médio** de **vinho** é de **$306,95.** De **peixes $37,43**. De **carnes $167,42**. **Doces $27,56**. De **frutas $26,40**. E para os produtos classificados como **‘Outros’**, o valor médio foi de **$44,18**. Se tivéssemos acesso à quantidade de transações por segmento, seria possível realizar uma análise mais completa.

5.2.3

Para a minha análise, optei por realizar o cálculo com base nos percentis.

Na aba: **dados_consolidados_base_oficial, coluna Z**, com a fórmula: **=DATA(2022;12;31) - N2** calculei a **quantidade_dias_entre_a_ultima_transacao** por cliente.

Na coluna AA com a fórmula: **=Z2 / K2** calculei o **intervalo_medio_entre_compras** dos clientes.

E na coluna AB, utilizando a fórmula: **=SE(Z2=0; 0; K2 / Z2)** foi calculado a frequencia_media_de_compra dos clientes.

Iniciei uma nova aba: **analise_rfm**, e utilizando a função: **=ARRAYFORMULA**, obtive as seguintes variáveis: **Id_cliente, quantidade_dias_entre_a_ultima_transacao, quantidade_de_transacao_realizada, total_gasto_por_cliente.**

Criei variáveis.

- ➔ Na coluna E: **percentil_recency_r** com a fórmula: **=SE(B2<=PERCENTILE.INC(B$2:B$1863; 0,2); 5;SE(B2<=PERCENTILE.INC(B$2:B$1863; 0,4); 4; SE(B2<=PERCENTILE.INC(B$2:B$1863; 0,6); 3;SE(B2<=PERCENTILE.INC(B$2:B$1863; 0,8); 2; 1))))** para me trazer a recência já classificada.
- ➔ Na coluna F: **percentil_frequency_f** com a fórmula: **=SE(C2>=PERCENTILE.INC(C$2:C$1863; 0,8);5;SE(C2>=PERCENTILE.INC(C$2:C$1863; 0,6); 4;SE(C2>=PERCENTILE.INC(C$2:C$1863; 0,4); 3;SE(C2>=PERCENTILE.INC(C$2:C$1863; 0,2);2;1))))** para me trazer a frequência classificada.
- ➔ Na coluna G: **percentil_monetary_m** com a fórmula: **=SE(D2>=PERCENTILE.INC(D$2:D$1863; 0,8); 5; SE(D2>=PERCENTILE.INC(D$2:D$1863; 0,6); 4; SE(D2>=PERCENTILE.INC(D$2:D$1863; 0,4); 3; SE(D2>=PERCENTILE.INC(D$2:D$1863; 0,2); 2; 1))))** para me trazer o percentil de valor monetário também classificado.

5.3

Ainda na aba: **analise_rfm** na coluna H, (rfm_score), utilizei a fórmula: **=CONCATENAR(E2;F2;G2)** para fazer uma chave única da *recência, frequência e valor monetário* classificados, com o objetivo de fazer uma análise.

Criei uma *tabela da coluna M até a coluna S* com os nomes e **descrições** de cada **segmento de cliente,** nessa mesma tabela, após analisar os dados, atribuí notas que seriam mais apropriadas para cada segmento de cliente, de forma que *todos os clientes* analisados *fossem segmentados*, sem deixar nenhum sem classificação.

No total criei **8 segmentos**, que são:

- **Cliente VIP (Gourmets da Casa)**

São os clientes mais fiéis e valiosos. Eles compram com frequência, gastam bastante e fazem compras recentes. Os melhores em termos de receita.

- **Cliente Leal (Chefes de Carteirinha)**

São clientes que compram regularmente e demonstram lealdade à marca, mas o valor de suas compras é médio em comparação aos Clientes VIP. Embora o gasto por compra seja menor, eles têm grande potencial para aumentar seu valor de compra.

- **Cliente Potencial (Candidatos a Gourmet)**

Têm potencial de se tornar grandes consumidores no futuro. Eles compram com uma frequência razoável e têm uma tendência a gastar um valor moderado.

- **Cliente Promissor (Caçadores de Sabores)**

São clientes que estão iniciando sua jornada na marca. Eles fazem compras de maneira esporádica e têm um gasto médio baixo. Embora tenham realizado algumas compras recentes, ainda não apresentam consistência nas aquisições.

- **Cliente em Risco (Sumidos da Despensa)**

Clientes em risco de se tornarem inativos, realizaram compras no passado, mas o engajamento diminuiu significativamente.

- **Cliente Inativo de Médio e Alto Valor (Degustadores Lendários)**

Clientes que têm um histórico de compras de valor significativo, mas que estão inativos no momento.

- **Cliente Inativo de Baixo Valor (Forasteiros do Mercado)**

Clientes que já não compram há um bom tempo, com baixa frequência e baixo valor de compra. Estes clientes são os menos engajados e têm um histórico de gastos baixo.

- **Cliente Novo (Novos Degustadores)**

Clientes que fizeram suas primeiras compras recentemente. Embora suas compras sejam esporádicas, há um grande potencial para fidelizá-los.

Com base nas notas que eu atribuí, na coluna I criei a variável: **segmento_rfm** e utilizei a fórmula:
=SE(E(EXT.TEXTO(H2;1;1)*1>=4;EXT.TEXTO(H2;2;1)*1>=4;EXT.TEXTO(H2;3;1)*1=5);"Cliente VIP";SE(E(EXT.TEXTO(H2;1;1)*1>=4;EXT.TEXTO(H2;2;1)*1>=4;EXT.TEXTO(H2;3;1)*1<=4);"Cliente Leal";SE(E(EXT.TEXTO(H2;1;1)*1>=3;EXT.TEXTO(H2;2;1)*1>=2;EXT.TEXTO(H2;3;1)*1>=2);"Cliente Potencial";SE(E(OU(EXT.TEXTO(H2;1;1)*1=3;EXT.TEXTO(H2;1;1)*1=4);EXT.TEXTO(H2;2;1)*1<=3;EXT.TEXTO(H2;3;1)*1<=2);"Cliente Promissor";SE(E(EXT.TEXTO(H2;1;1)*1=2;EXT.TEXTO(H2;2;1)*1<=5;EXT.TEXTO(H2;3;1)*1<=5);"Cliente em Risco";SE(E(EXT.TEXTO(H2;1;1)*1=1;EXT.TEXTO(H2;2;1)*1<=3;EXT.TEXTO(H2;3;1)*1<

=2);"Cliente Inativo de Baixo

Valor";SE(E(EXT.TEXTO(H2;1;1)*1=1;EXT.TEXTO(H2;2;1)*1>=1;EXT.TEXTO(H2;3;1)*1>=2);"Cliente Inativo de Médio e Alto
Valor";SE(E(EXT.TEXTO(H2;1;1)*1=5;EXT.TEXTO(H2;2;1)*1<=2;EXT.TEXTO(H2;3;1)*1<=4);"Cliente Novo";"Outro")))))))

Aproveitei a mesma tabela que eu havia criado e com a função: **=CONT.SE** contei *quantos clientes* tinha em *cada segmento* que foi criado.

Em relação ao '*Processamento e Análises'*, além das etapas principais descritas, foram também conduzidos outros procedimentos relevantes que contribuíram para a *consolidação* dos **resultados** e **insights obtidos**, como os **dashboards** desenvolvidos no *Google Sheets* e no *Looker Studio* e a **apresentação** dos *principais resultados* e *insights* no *Google Planilhas*.

Resultados e Conclusões:

O *período analisado* foi de **30/07/2020** a **31/12/2022**, totalizando *885 dias.*

Foram segmentados **1862 clientes**.

O *faturamento total* deste período considerando os clientes segmentados foi de **$1.135.711,00,** e o *ticket médio geral* **$61,39.**

Com a análise, foi concluído que a *campanha de marketing* **não gerou conexão suficiente com o público**, dos **1862 clientes analisados**, apenas **292** aderiram a campanha e mais de **84%** da base não aderiu a campanha.

Dos 50 clientes segmentados como **"Cliente Novo"** *(cliente com recência mais alta, baixa frequência e baixo valor monetário)*, apenas 2 aderiu a campanha, ou seja, se o intuito da campanha foi atrair novos clientes, *os resultados demonstram baixa efetividade nesse objetivo*, uma vez que *não houve aumento significativo na base de clientes.*

Outro ponto também é que a campanha teve *baixa aderência principalmente* com o **público mais jovem**, dos clientes que aderiram a campanha apenas **3,42%** tinham entre **25-34 anos.**

O perfil que *mais aceitou a campanha* foram os de estado **civil: solteiro, nível de escolaridade posgrado, faixa etária de 45-54 anos.**

Dos *segmentos*, o perfil que teve um número maior de clientes que aderiu à campanha, foi os **Clientes Potenciais**, que representaram o maior número absoluto de participantes, com **89 clientes** engajados. No entanto, *em termos percentuais dentro do próprio segmento*, os que mais engajaram foram os **Clientes VIPs**, com **27,45%** de *aceitação.*

Já que a **fidelidade** do cliente se tornou um **desafio** para "o Mercado", para estes perfis que tiveram maior aderência à ação de marketing, sugiro Investir em campanhas de **upsell e cross-sell**, oferecendo a eles produtos de qualidade superior e versões exclusivas, uma vez que este público está engajado com a marca. É o momento ideal para fidelizá-lo.

Um ponto de atenção, é que os clientes classificados como **“Cliente Em Risco”** *(com baixa recência)*, e que são **19,8%** da base analisada, sendo o **2° maior segmento**.
Representam o **2º maior volume de transações** e **2º maior faturamento no produto mais lucrativo**, que é o **vinho**, se esses clientes se tornarem inativos, o cenário se tornaria ainda **mais crítico.**
A fim de aumentar a **recência** desses clientes, sugiro ofertar com senso de **urgência** descontos exclusivos e com validade curta, ou benefícios VIP temporários.
A maioria destes clientes tem mais de 65 anos, a maior parte também são casados e destes classificados como “Cliente Em Risco” **18,82%** aderiram à campanha de marketing.
Para os **Clientes Novos**, que representam **2,69%** da base analisada, a maior parte deles tem entre **45-54 anos**, e **50%** possuem ***grado o superior***. A sugestão para estes, é implementar um **programa de pontos** ou **cashback** a partir da *segunda compra,* tendo uma **validade curta** e que preferencialmente, possa ser utilizado como um desconto percentual aplicado sobre o valor da próxima compra, o objetivo aqui é **engajar esse perfil** de clientes desde o começo, transformando eles em *compradores recorrentes*, com foco em *fidelizá-los.*

O produto **mais vendido** é o **vinho**, responsável por **50,32%** do *faturamento*, o que representa **$571.546,00.**
Os *maiores compradores* são os *clientes potenciais*.
Como este é o **produto campeão** em todos os **8 segmentos de clientes**, recomendo utilizar este produto como ***“isca”*** para *vender outros itens,* e assim também *aumentar o ticket médio,* a loja poderá adicionar **amostras de outros produtos** junto ao item campeão para incentivar o cliente a conhecer outros produtos, poderá montar kits com produtos combinados a fim de tentar **impulsionar produtos com menor saída**, como por exemplo **doces e frutas** que são os produtos com **menor faturamento**, *doces* representam **4,52%** de vendas e *frutas* apenas **4,33%** do faturamento.

Em relação ao **canal de compra**, a **loja física** é o **canal favorito** dos clientes representando **58,3%** das **18.501** *transações analisadas.*
Para esse cenário, tenho duas recomendações:

1. *Fortalecer o **marketing digital,** investindo em **tráfego pago** para expandir a marca e **alcançar novos públicos.***
2. *Criar campanhas de marketing e incentivos cruzados entre os canais, como por exemplo: **Descontos na Loja Física** (se o cliente comprar online).*

Para os **Clientes Vips**, que são os melhores em termos de receita, e para os **Clientes Leais** que estão sempre comprando, a fim *preservar a* ***fidelidade*** e *manter a* ***recorrência****,* para estes clientes eu sugiro campanhas com convites exclusivos, onde gastando "X" o cliente ganha uma degustação exclusiva que pode inclusive incluir sommelier, já que o produto mais comprado por eles é o vinho, sugiro também acesso antecipado a produtos especiais, atendimento personalizado, e um programa de pontos para cada amigo que indicarem a fim de *expandir a visibilidade da marca* e *atrair novos clientes.*

Os **Clientes Casados** lideram a base analisada, representando **38,88%** do total, além de serem a maior parcela em todos os segmentos. Também se destacam no faturamento, sendo responsáveis por **37,96%** da *receita gerada*. Recomendo investir em campanhas voltadas para casais, como combos temáticos ex: Jantar Romântico *(vinho+queijo+geléia)*, outra sugestão é ofertar programa de pontos compartilhado entre contas de um casal.

Devido a *inconsistências* e possíveis *dados corrompidos*, dos mais de 2.200 registros de ID do *Dataset*, apenas *1.862 clientes puderam ser segmentados.*

Em **'Limitações e Próximos Passos'** darei *informações complementares* sobre, e *recomendações.*

Outros *resultados* e *conclusões* também foram obtidos ao longo do projeto, no entanto, neste resumo, destaquei aqueles que *considerei mais relevantes.*

Limitações e Próximos Passos:

Durante a fase de **análise exploratória (5.2)**, quando criei uma chave única, um dos principais desafios foi a identificação de **possíveis registros duplicados, mas, de IDs de clientes diferentes.**

Foram identificados os seguintes ID_cliente com informações duplicadas:

9,20,24,49,67,78,92,175,202,217,234,241,271,295,524,544,574,577,663,716,736,762,833,849,868,879,880,933,946,968,973,979,1000,1020,1052,1071,1135,1146,1162,1225,1232,1245,1250,1343,1388,1407,1409,1419,1456,1502,1600,1612,1630,1745,1876,1920,1966,1998,2055,2088,2106,2114,2154,2176,2217,2227,2246,2253,2262,2326,2345,2392,2407,2429,2452,2453,2493,2521,2561,2563,2631,2661,2669,2711,2747,2782,2793,2807,2811,2814,2826,2849,2891,2920,2928,2929,2945,2963,2968,3006,3007,3011,3033,3037,3056,3091,3129,3130,3152,3174,3220,3266,3310,3312,3332,3340,3363,3386,3421,3428,3433,3560,3565,3697,3769,3829,3852,3855,3856,3859,3867,3900,3910,3934,4001,4047,4086,4088,4094,4096,4102,4107,4119,4127,4168,4186,4198,4211,4248,4261,4301,4385,4391,4399,4436,4459,4470,4500,4508,4541,455,4599,4607,4646,4698,4749,4767,4838,4843,486,4939,4944,497

1,4973,4998,5011,5012,5068,5077,5084,5093,5107,5120,5153,5156,5184,520,5232,5299,5314,5386,5396,5424,5430,5462,5491,5513,5517,5529,5536,5538,5547,5731,5751,577,5830,5885,5929,6024,6071,6173,6183,6202,6222,6245,6283,6310,6318,6383,6417,6431,6544,6583,6661,6729,6730,6798,6818,6853,6864,6887,6918,6935,6988,7010,7059,7093,7094,7152,7212,7232,7254,7261,7275,7290,7321,7366,7375,7386,7426,7433,7451,7485,7500,7516,7521,7530,7592,762,7685,7699,7723,7787,7789,7807,7881,8015,8175,8268,8299,8312,8315,8334,8360,8362,8372,8373,8420,8427,8430,8432,8462,8504,8524,8557,8588,8595,8619,8652,8685,8702,8724,8727,8775,8780,8800,8842,8867,8895,8910,8953,9064,9121,9150,9209,9284,9308,9323,933,9347,9381,9384,9396,9460,9477,9495,9543,9589,9733,9738,9757,979,9805,9826,9888,9916,9923,9955,9958,9964,9971,9972,10033,10037,10065,10120,10128,10144,10146,10160,10172,10177,10207,10212,10258,10260,10264,10304,10451,10536,10562,10602,10617,10642,10643,10648,10681,10703,10727,10735,10741,10766,10789,10795,10826,10872,10897,10905,10906,10981,11171,11188.

Os dados analisados eram extremamente idênticos (mesma data de entrada, valores de compras atribuídos a cada tipo de produto iguais, ano de nascimento, nível de educação, estado civil e vários outros dados), diferindo apenas o n° **ID_cliente** e algumas datas de transação, já que **duas transações** as datas sempre se cruzavam entre eles, e essas que se cruzavam se tratavam exatamente da mesma **data de entrada (cadastro)**. Isso me levou a acreditar que pode ter ocorrido um **erro de entrada no sistema**, que acabou gerando **registros duplicados** com **pequenas variações**.
Através destes possíveis duplicados, realizando uma análise mais aprofundada pude identificar os valores de salário que estavam ausentes na base de dados para os seguintes ID_cliente:

1. **O Id_cliente 8268 apresentava os mesmos dados do Id_cliente: 2782, exceto a última data de compra, a do Id_cliente 8268 foi em em 17/05/2022 e do Id_cliente 2782 foi em 26/07/2022.**
2. **O Id_cliente 3769 apresentava os mesmos dados que o: 10146, exceto a última data de compra, a do ID_cliente 3769 foi em 20/09/2022 e a do ID_cliente 10146 foi em 25/11/2022.**
3. **O Id_cliente 1612 apresentava os mesmos dados que o: 4385, exceto a última data de compra, a do ID_cliente 1612 foi em 08/05/2022 e a do ID_cliente 4385 foi em 31/01/2022.**
4. **e O Id_cliente 8557 os mesmos que o Id_cliente 8724, exceto também a última data de compra, a do ID_8557 foi em 21/11/2022 enquanto a do ID_cliente 8724 foi em 12/10/2021.**

Inicialmente, tratei os salários nulos utilizando a mediana salarial correspondente ao nível de escolaridade dos clientes. No entanto, após identificar registros duplicados destes quatro clientes que anteriormente apresentavam salários ausentes, foi possível recuperar esses valores a partir dos dados duplicados, permitindo a atualização dessas informações na base.

Após isso, **atualizei as medianas e refiz as fórmulas condicionais** dos demais, ajustando os valores de forma mais precisa e alinhada à realidade observada na base.

Também verifiquei se esses registros duplicados poderiam estar relacionados a outros outliers presentes na base de dados, mas a única correlação identificada foi com os casos de salários nulos.

O total de duplicados encontrados sendo de Id_cliente diferentes foi de **368 IDs**, que representavam **16,5% da base total**, *considerando alguns dados que já haviam sido tratados*. uma fatia significativa, o que me levou a analisar por bastante tempo **o que fazer com estes duplicados.**

Eu poderia unificar os dados, somando a quantidade de transação, e deixando apenas um ID_cliente na base de dados, que seria no caso, o que tivesse a **última data de compra sendo a mais recente**, mas, e se estes dados se tratarem de clientes diferentes? São muitas coincidências, mas, poderia ser apenas uma grande coincidência.

Eu poderia também unificar, somando os valores das compra entre estes duplicados, considerando que estes clientes compraram utilizando contas diferentes, e por isso os Id's seriam diferentes, mas além de ter datas de transação cruzadas, também existem datas de transação diferentes entre eles, e os valores e itens comprados são exatamente os mesmos. E se foi um erro na entrada de dados onde duplicou as informações? Se eu fizesse a **soma** eu **estaria inflando os dados**, o que não geraria uma precisão na minha análise.

Analisei diferentes perspectivas, mas, diante da **impossibilidade de verificar a veracidade** dessas duplicidades, eu optei por deletá-las, não deixar nenhuma nem outra. Se tratando de **análise RFM** acredito que é melhor ter um volume menor de dados, mas, que sejam confiáveis, do que trabalhar com registros **potencialmente corrompidos** que possam comprometer a qualidade e a precisão da análise.

Com estes desafios apresentados, minha sugestão para "O Mercado", é que implementem validações no processo de entrada de dados, para evitar preenchimentos incorretos e duplicidades. Sugiro a automatização do processo de limpeza e tratamento dos dados para melhorar a consistência das informações. E para garantir a identificação correta de cada cliente, sugiro que implementem a criação de identificadores únicos e rastreáveis, como o CPF.

Implementar essas sugestões não só reduzirá erros e inconsistências, mas também otimizará os processos de análise, permitindo uma tomada de decisão mais

informada e eficaz. Com essas medidas, "O Mercado" estará mais preparado para enfrentar desafios futuros, garantindo maior confiança em suas análises e no desempenho geral dos negócios.

Links de interesse:

- *Spreadsheet*
  https://docs.google.com/spreadsheets/d/1slsWoEBSUyfIUF3UxoPh4zd-aRkJnU94gKc9o_VIMPk/edit?usp=sharing

- *Apresentação de Slides*
  https://docs.google.com/presentation/d/1ylBli39vwyR6keyr_et__7HtcjebEPzt/edit?usp=sharing&ouid=100403582151255039255&rtpof=true&sd=true

- *Dashboard Looker Studio (Marco Adicional)*
  https://lookerstudio.google.com/s/jgDPwAt0m4E

- *Vídeo: Apresentação de Resultados*
  https://www.loom.com/share/9fa6ee4574db4f6aaca115256f52acb4?sid=3ab53939-bcf9-4ab8-b8ca-157cd79f95d1