

Projeto 1: Segmentação. (=Customer Segmentation)

Realizar uma análise descritiva e segmentação de clientes utilizando RFM para uma loja especializada em produtos alimentícios importados chamada "O Mercado".

Segmentação

É uma estratégia essencial na análise de dados que consiste em dividir um conjunto de dados como registro de compras por clientes de uma empresa, em grupos mais homogêneos com base em características ou comportamentos semelhantes.

Pasos para el análisis de datos

- | | |
|-------------------------------|--------------------------------------|
| 1- Definir el problema | Hacer preguntas |
| 2- Preparar los datos | Reunir los datos |
| 3- Procesar/Limpieza de datos | Limpia datos |
| 4- Analizar | Responder preguntas |
| 5- Confortar | Tablero |
| 6- Actuar | Toma decisiones con base a los datos |

Objetivo

- Identificar posibles diferencias significativas entre los clientes
- Analizar las ventas y segmentar la base de clientes usando RFM. En cuales grupos la empresa puede concentrar esfuerzo y estrategias de fidelización.
- Comprender los comportamientos de compra de los clientes

Etapas análise de datos

4. Processo de trabalho

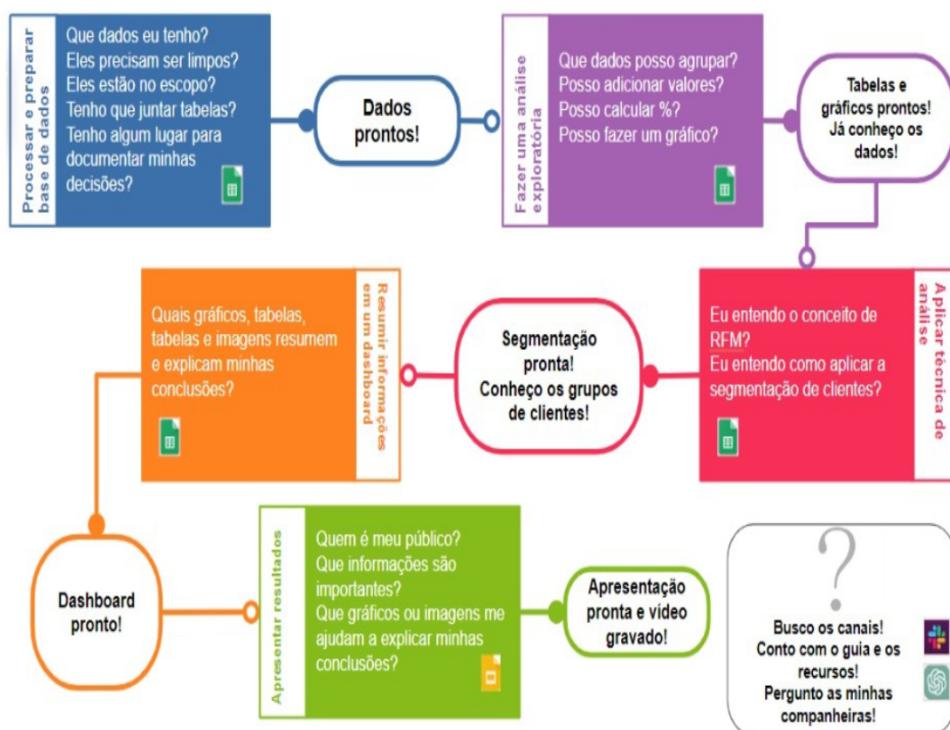


Marco 5

5. Mão à obra! - Marco 1

Em geral, em todos os marcos você passará por todas as etapas da análise de dados (com exceções), desde a compreensão do problema, processamento e limpeza dos dados, até a análise propriamente dita e a apresentação dos resultados.

Neste marco vamos explorar os dados de uma loja especializada em produtos alimentícios importados para segmentar clientes e fornecer informações importantes para o negócio. O contexto descrito anteriormente nos ajuda a entender o problema a ser resolvido, que perguntas devem ser respondidas e o que se espera ao final do projeto, mas não se limite ao exposto acima. Se você encontrar alguma informação interessante ou alguma métrica que possa complementar a análise, desafie-se e vá em frente.



Tres puntos nuevos y críticos:

- Tomada de decisão na etapa "Processar e preparar banco de dados". As perguntas que podem surgir são: O que fazer quando encontrar valores nulos? O que fazer quando encontrar valores duplicados? Como identificar valores fora do escopo?
- Como segmentar clientes na etapa "Aplicar técnica de análise". As questões que podem surgir aqui são: Como decidir quantos e quais perfis de clientes criar? Como faço para criar as regras?
- Como montar uma apresentação na etapa "Apresentar resultados". As perguntas podem ser: Que informações são realmente importantes? E em que investir o pouco tempo que tenho para apresentar?

Nestes três passos o conselho é o mesmo: **fique calma, use todos os recursos disponíveis** (canais Slack, sessões colaborativas, perguntar para uma IA, pesquisar no Google, etc.) e **tome decisões com confiança**. Isso é uma parte importante do seu aprendizado. Se você não tem dúvidas ou comete erros, como você aprende? 😊

→ Busca melhorar a qualidade e a confiabilidade de dados conjuntos de dados confiáveis.

I: Procesar e preparar base de datos.

Após coletar os dados, eles não podem para serem analisados. Primeiro, limpa os dados:

- 1: Remover erros, informações que podem estar duplicados e discrepâncias
- 2: Remover dados que não são úteis para a análise
- 3: Corrigir erros de digitação
- 4: Preencher os principais lacunas nos dados

* Conectar/importar los datos:

• Usé la formula IMPORTARANGE()
para importar los datos de los planillas en una sola planilla de google sheets

* Identificar y tratar valores nulos:

• Usé la formula CANTBLANQ() /
CONTAR.VAZIO() para saber el numero de celdas o datos vacios.

• Usé la formula =SE(ÉCÉL.VAZIA(clientes!E2),
ARRED(MED(FILTER(clientes!E:E, clientes!E:E<>""))),
clientes!EZ) en la planilla clientes, donde

Creé otra Pestaña llamada "clientes_not-null" y allí copie los datos de la pestaña "clientes" y en la Columna subtotal_anual_dollar colocar la fórmula para llenar los campos vacíos, (24) usandola como medida de tendencia central la mediana ya que hay datos muy dispersos y la mediana es mejor que la media.

- En la planilla "transacciones" cree una Pestaña llamada "transacciones_not-null" y coloque la fórmula =FILTER(transacciones!A:D, NÚM. CARACT(transacciones!B:B)>0) que me traerá los datos de la pestaña "transacciones" pero sin los valores vacíos de la columna Id-cliente.

* Identificar y tratar valores duplicados:
Por medio de herramientas de Planillas google, identif. cci
valores duplicados.

- Para identificar valores repetidos se usa la
fórmula: $=IF(COUNTIF(A:A, "<>" & "") > COUNTUNIQUE(A:A), TRUE, FALSE)$
- Para sacar los valores repetidos/duplicados,
se usa la fórmula =FILTER(A:A, COUNT.SE(A:A, A:A) > 1),
esta formula la usé en la florilla resumo_compras,
donde, free una copia de resumo_compras
aplique la formula para saber que Id-cliente
está duplicado.

* Identificar e gerenciar datos para do
escopo da análise

- En esta parte se hicieron los clientes de
la florilla transacción que no tienen Id-cliente
porque no se podían segmentar/categorizar
esos transacciones y no son útiles para
el análisis.

• En la tabla clientes, en la columna Salario_anual, usando la medida de tendencia central: mediana para tener los espacios vacíos de esa columna. Se tomó como referencia la mediana porque no es afectada por los valores atípicos.

• En la tabla resumo_compras, en la columna Id-cliente, se eliminaron los Id-cliente que estaban repetidos, esto con el fin de que no afecte nuestro análisis con Id-cliente repetidos.

• Hay clientes que no tienen transacción pero si tienen registro de compras. Decidi dejar esos clientes fuera del grupo de análisis porque si quiero realizar algún análisis comparativo entre las informaciones de los clientes con más ventas en línea o en linea, entonces hay que tener en cuenta a esos clientes. También esos clientes son solo 10 y el porcentaje es muy bajo en ventas ($\$376$).

• Haciendo clientes sin transacciones se dirigió esos clientes sin transacciones de la tabla clientes. Se unió clientes y resumo de compras usando como base la tabla clientes. (Al usar la tabla clientes como base solo se toman en cuenta los clientes con transacciones).

• En la planilla de transacciones,
encontré 4394 transacciones que
están fuera del escopo de estudio
del Supermercado en el periodo
de 30/07/2020 a 29/06/2020,
motivos por los cuales consideré esos
registros de compra fuera del
escopo, al analizar es decir
eliminar esos registros de compra.

* Unir tabelas

Refere-se ao processo de combinar ou relacionar dados de dois ou mais conjuntos de dados que estão em planilhas ou intervalos de células diferentes em uma das Planilhas Google.

Se puede usar fórmulas como: VLOOKUP(FRAC), INDEX+MATCH(ÍNDICE-CORRESP), QUERY y otros más...

- Para unir os 3 tabelas partiendo como base la tabla transacciones, porque al contiene más registros que las otras, tabla 2 no se pide tanto información repetitiva, los otros tablas. Entiendo que podría usar la tabla clientes como base y traer un resumen de la tabla transacciones por cada cliente (por ejemplo la cantidad de transacciones de compra) para esto combinaría la unión de las tablas.
- Fue:
la fórmula que use para unir las tablas
mihabise
2231

```

=ArrayFormula(
{
    transacoes!A:D,
    PROCV(transacoes!B:B, clientes!A:I, COL(INDIRETO("R1C2:R1C"&COLS(clientes!A:I),0)),0),
    PROCV(transacoes!B:B, resumo_compras!A:G, COL(INDIRETO("R1C2:R1C"&COLS(resumo_compras!A:G),0)),0)
}
)

```

Named Ranges

- table1 : Sheet1!A1:C3
- table2 : Sheet2!A1:C3
- ID : Sheet1!A1:A3

Formula

```

=ArrayFormula(
{
    table1,
    vlookup(ID, table2, COLUMN(Indirect("R1C2:R1C"&COLUMNS(table2),0)),0)
}
)

```

Remarks:

- Using open ended ranges is possible but this could make the spreadsheet slower.
- To speed up the recalculation time :
 1. Replace `Indirect("R1C2:R1C"&COLUMNS(table2),0)` by an array of constants from 2 to number of columns of table2.
 2. Remove the empty rows from the spreadsheet

En el paso anterior de unir tablas, seguí las variables de la tabla de "transacoes", la tabla clientes y luego uní los tablas clientes y resuma de compras. (minha box de dados 223)

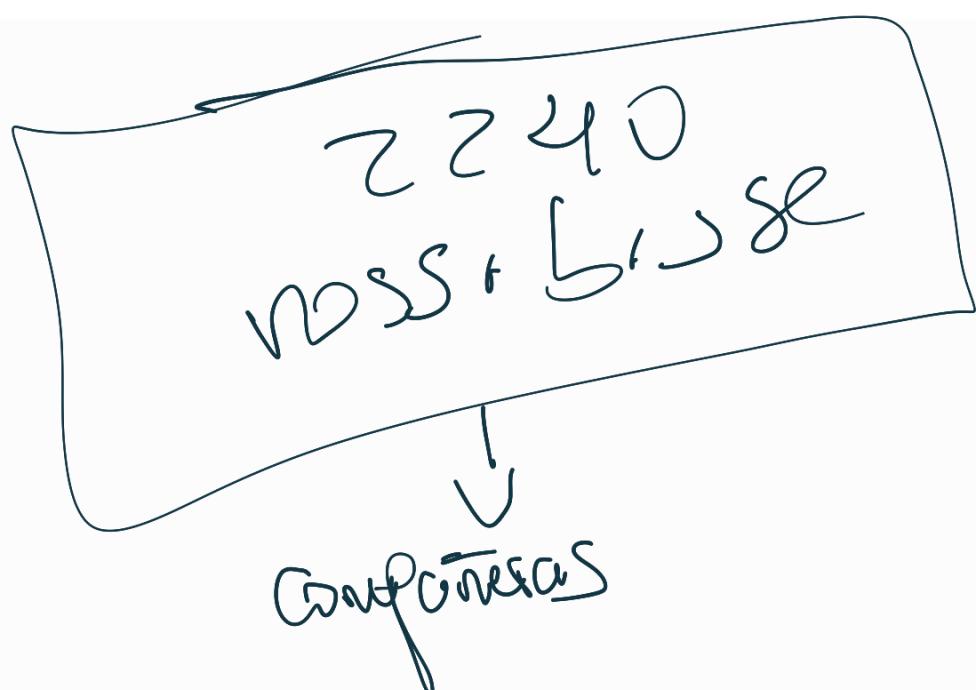
unir los tableros de las formas (usando las fórmulas)

→ Primera Fórmula

```
fx =ARRAYFORMULA({  
    clientes!A:J,  
    SEERRO(POCV(clientes!A:A, {resumo_compras!A:A, resumo_compras!B:B}, 2, FALSO), ""),  
    SEERRO(POCV(clientes!A:A, {resumo_compras!A:A, resumo_compras!C:C}, 2, FALSO), ""),  
    SEERRO(POCV(clientes!A:A, {resumo_compras!A:A, resumo_compras!D:D}, 2, FALSO), ""),  
    SEERRO(POCV(clientes!A:A, {resumo_compras!A:A, resumo_compras!E:E}, 2, FALSO), ""),  
    SEERRO(POCV(clientes!A:A, {resumo_compras!A:A, resumo_compras!F:F}, 2, FALSO), ""),  
    SEERRO(POCV(clientes!A:A, {resumo_compras!A:A, resumo_compras!G:G}, 2, FALSO), "")  
})
```

→ Segunda Fórmula

```
fx =ArrayFormula(  
{  
    clientes!A:J,  
    PROCV(clientes!A:A,resumo_compras!A:G,COL(INDIRETO("R1C:R1C"&COLS(resumo_compras!A:G),0)), FALSO)  
})
```



* Criar novas Variáveis

- Criei a variável `idade`, para saber la edad de los clientes
- Criei a variável `total_compras` & `compras que hizo` la suma de total de compras que hizo cliente

2º Fazer uma análise exploratória

É uma etapa crucial para obter una comprensión inicial y una visión general de los datos antes de aplicar técnicas o métricas.

- faz gráficos para tener una visión general del negocio.

Days (Today): M5

"Quantidade que gasto em fritas" =
Total gas 105

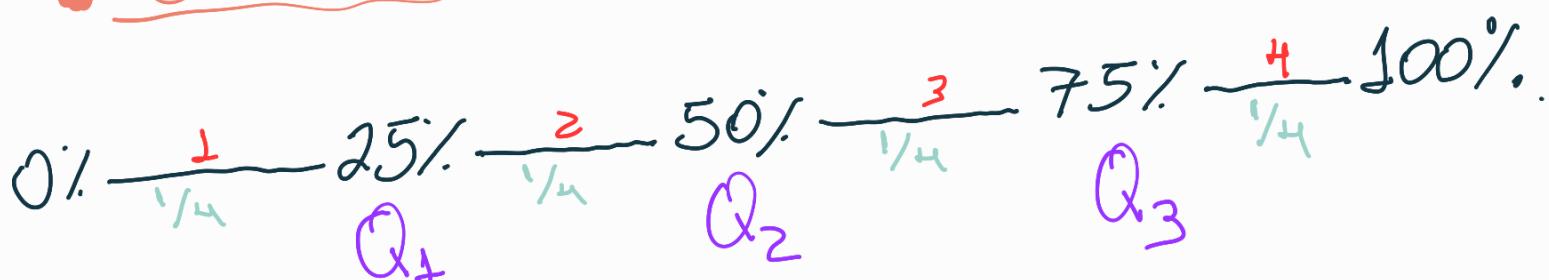
$$\frac{\frac{1}{4} + \frac{1}{7}}{2} = \frac{5}{8}$$

Calcular quartil, deciles o Percentil

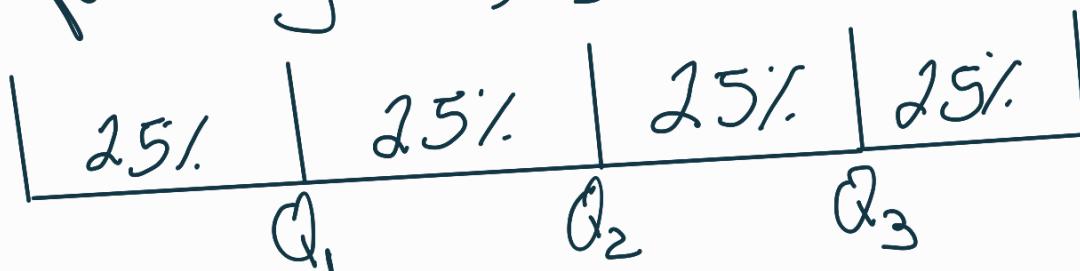
Existen otros estadígrafos que dividen a los datos en otras proporciones y no sólo en mitades como la mediana. Estos mididos se llaman Cuartiles o Cuantiles. Los más usados son: Cuartiles, deciles y percentiles. Se usan para describir el comportamiento de una población, los valores se dan con menudos en tantos por ciento.

Quartil

$$\frac{10}{5} = 2$$



Son valores que dividen a un conjunto de datos ordenados en forma ascendente en cuatro partes iguales, y se denota por Q_i : $i=1, 2, 3$



→ Primer Quartil (Q_1): valor situado de tal modo en la serie de datos que 25% de las observaciones son menores que él e 75% son mayores.

→ Segundo Quartil (Q_2): Valor situado de tal modo na série de dados que 50% das observações são menores que ele e 50% são maiores.

→ Terceiro Quartil (Q_3): Valor situado de tal modo na série de dados que 75% das observações são menores que ele e 25% são maiores.

Decil.

Denominados decis os valores de uma série que o dividem em 10 partes iguais

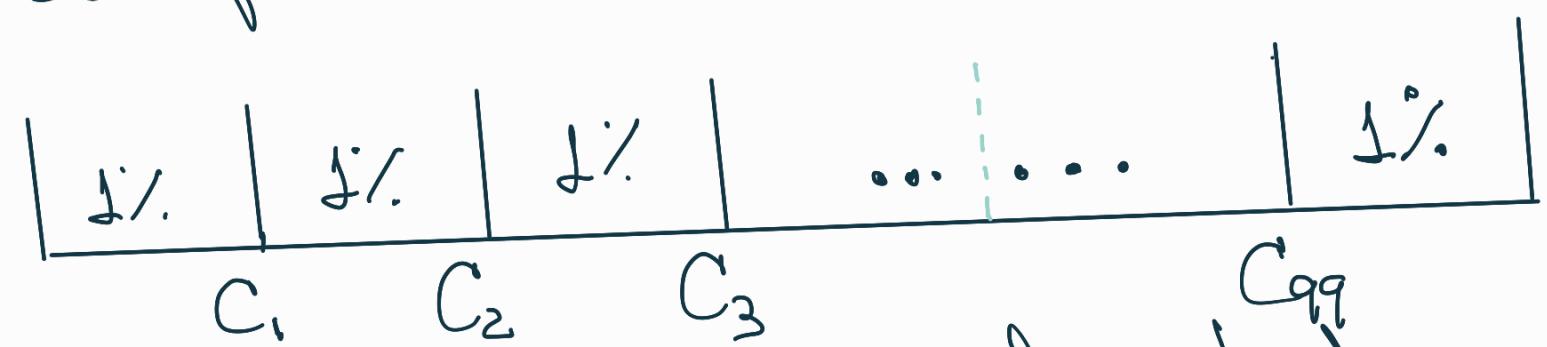
D₁ D₂ D₃ ... D₉

→ Primeiro Decil (D_1): Valor situado de tal modo na série de dados que 30% das observações são menores que ele e 90% são maiores.

→ Segundo Decil (D_2): valor situado de tal modo na série de dados que 20% das observações são menores que ele e 80% são maiores

→ Nono Decil (D_9): valor situado de tal modo na série de dados que 90% das observações são menores que ele e 10% são maiores

• Centil (ou Percentil):
Denominamos percentis os valores de uma série que a dividem em 100 partes iguais.



→ Primeiro Percentil (C_1): valor situado de tal modo na série de dados que 1% das observações são menores que ele e 99% são maiores

→ Segundo Percentil (C_2): Valor situado de tal modo na Série de dados que 2% das observações são menores que ele e 98% são maiores.

→ Nonagésimo Nono Percentil (C_{99}): Valor situado de tal modo na Série de dados que 99% das observações são menores que ele e 1% são maiores.

3º Aplicar técnica de análisis

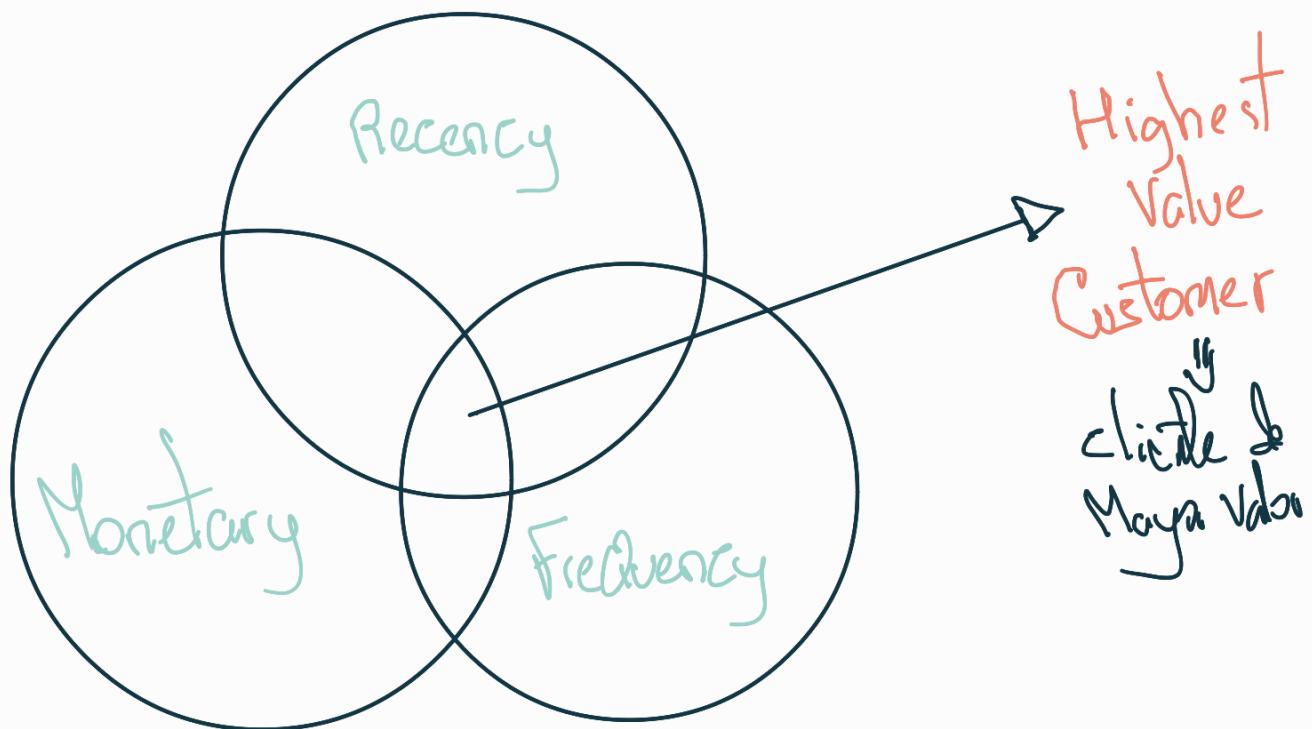
Segmentación

La segmentación es una estrategia esencial en análisis de datos que consiste en dividir grupos de datos, como registros de compras por clientes de una empresa en grupos menores y más homogéneos con características y comportamientos semejantes. Pueden incluir características demográficas, comportamiento de compra, intereses, preferencias y/o histórico de interacciones con la empresa.

Segmentación de clientes

Es una estrategia de Marketing y análisis de datos que consiste dividir los clientes o usuarios de una empresa en grupos homogéneos y distintos con características y comportamientos semejantes. El principal objetivo es comprender mejor las necesidades, preferencias y comportamientos de los diferentes grupos de clientes, de la forma de ofrecerles

Productos, servicios y comunicaciones más personalizados y adecuados a sus intereses



Segmentação RFM

• R (Recencia): Tiempo para (desde) se refiere al tiempo transcurrido desde la última interacción o transacción del cliente con la empresa. Es una medida de cuándo fue la última vez que un cliente realizó una compra, visitó la página, tuvo cualquier otro tipo de interacción (devante con la empresa). Cuanto más reciente fué la última interacción, mayor será el valor atribuido a la dimensión "R".

→ R: Hace santos días fue la última compra do cliente.

• F(Freqüência): Se refere a la frecuencia o número de veces que un cliente ha interactuado o realizado una compra con la empresa en un periodo determinado. É uma medida de quão ativo um cliente é em termos do número de vezes que ele faz transações ou interações.

→ Quanto maior a freqüência, maior será o valor atribuído à dimensão "F".

→ Quanto compras esse cliente já fez na sua empresa, desde que cadastrou?

• M(Monetariedade): Representa o valor monetário total que um cliente gastou na empresa durante um período específico

→ Quanto maior o valor monetário, maior será o valor atribuído à dimensão "M".

→ Quanto esse cliente já gastou em dinheiro na sua empresa?

→ Quanto mais recente for a compra do cliente,
maior será o português de R - ou score.

→ Quanto mais compras o cliente realizar,
maior será o score de F.

→ Quanto maior for o gasto dele, maior
será o score de M.

MARCO 2

1= Procesar y preparar base de datos

* Construir tablas auxiliares

Se creó una tabla auxiliar llamada "análisis_coorte" donde importé los datos de la tabla de transacciones (datos limpios) y de la tabla de clientes la columna "data_entrada". Después se creó las variables "ano_mes_data_entrada", "meses_data_entrada", "ano_mes_data_transacción", "meses_data_transacciones". Esas variables van a servir para hacer el análisis o crear tabla de coorte.

2- Fazer uma análise exploratória.

*Aplicar Medidas de tendência central

- Encontrar os clientes de maior
4 meses em realizar sua primeira
compra/transação

30/07/2020
29/06/2022

Conclusões

* Visualizar Distribución

- Se usó un gráfico llamado Histograma para visualizar la distribución

Conclusiones

3º Aplicar técnica de análise

* Análise por cohorte

O termo "cohorte" refere-se a um grupo de indivíduos que compartilham uma característica ou experiência comum dentro de um período definido. Na prática, isso significa agrupar clientes com base em critérios como a data de adquisição, o primeiro uso de um serviço ou produto, ou outros eventos significativos.

* Churn

Taxa de churn = (Quantidade de clientes perdidos / total clientes novos)