Pontifícia Universidade Católica do Rio de Janeiro
Pós-graduação Lato Sensu em Ciência de Dados e Analytics
Sprint: Engenharia de Dados (40530010057_20240_01)

Aluno: Matheus Costa da Rocha Gonçalves
Matrícula: 4052024001298

Documentação do MVP de Engenharia de Dados

Objetivo geral:

A partir de uma base de dados contendo informações fictícias sobre as avaliações de serviços e produtos de um restaurante fornecidas por seus clientes, o objetivo deste projeto é realizar uma análise descritiva e diagnóstica capaz de produzir informações sobre quais fatores influenciam a experiência dos clientes e impactam o seu índice de satisfação. Serão avaliados os perfis de consumo dos clientes do estabelecimento e serão identificadas correlações entre as notas de satisfação atribuídas pelos clientes aos serviços e produtos com os diferentes atributos para identificar pontos de melhorias, bem como traçar estratégicas com o intuito de aumentar o faturamento e atrair o público alvo do estabelecimento.

Perguntas gerais:

- Quais são os fatores predominantes que estão influenciando a baixa satisfação do cliente?
- Como aumentar a frequência dos clientes fidelizados no estabelecimento?

1) Coleta e Modelagem dos Dados

Os dados para o projeto foram coletados em 10/07/2024 a partir de um *dataset* disponibilizado pelo usuário identificado como RABIE EL KHAROUA no website Kaggle. Os dados foram obtidos por meio do endereço https://www.kaggle.com/datasets/rabieelkharoua/predict-restaurant-customer-satisfaction-dataset e seu uso é permitido sob a licença ATTRIBUTION 4.0 INTERNATIONAL (CC BY 4.0), conforme os termos disponíveis no link a seguir: https://creativecommons.org/licenses/by/4.0/.

O *dataset* utilizado para este projeto é um arquivo CSV e a modelagem utilizada para sua construção foi um esquema Flat. A partir deste *dataset*, foi criada uma tabela chamada de

“restaurant_customer_satisfaction” em um Notebook da plataforma Databricks Community Edition, que possui dezenove colunas, classificadas conforme o catálogo de dados abaixo:

Catálogo de Dados:

Nome da Coluna	Tipo do Dado	Descrição	Unidade	Valor Mínimo	Valor Máximo	Categorias
CustomerID	string	Identificador único para cada cliente	N/A	N/A	N/A	N/A
Age	int	Idade do cliente em anos	Anos	0	N/A	N/A
Gender	string	Gênero do cliente (masculino/feminino)	N/A	N/A	N/A	"Male", "Female"
Income	float	Renda anual do cliente em USD	USD	R$ 0,00	N/A	N/A
VisitFrequency	string	Frequência de visitação do cliente ao estabelecimento (diária, semanal, mensal, raramente)	N/A	N/A	N/A	"Daily", "Weekly", "Monthly", "Rarely"
AverageSpend	float	Quantia média gasta pelo cliente por visita em USD	USD	R$ 0,00	N/A	N/A
PreferredCuisine	string	Tipo de culinária preferida pelo cliente	N/A	N/A	N/A	"Italian", "Chinese", "Indian", "Mexican", "American"
TimeOfVisit	string	Período do dia em que o cliente normalmente visita o estabelecimento	N/A	N/A	N/A	"Breakfast", "Lunch", "Dinner"
GroupSize	int	Número de pessoas que compõem o grupo do cliente durante a visita	Pessoas	1	N/A	N/A
DiningOccasion	string	Tipo de ocasião para a visita ao estabelecimento	N/A	N/A	N/A	"Casual", "Business", "Celebration"
MealType	string	Tipo de pedido (para comer no estabelecimento ou para retirar)	N/A	N/A	N/A	"Dine-in", "Takeaway"
OnlineReservation	string	Valor binário que representa se o cliente fez uma reserva online (0: Não, 1: Sim)	N/A	N/A	N/A	"0", "1"
DeliveryOrder	string	Valor binário que representa se o cliente fez um pedido delivery (0: Não, 1: Sim)	N/A	N/A	N/A	"0", "1"
LoyaltyProgramMember	string	Valor binário que representa se o cliente é um membro do programa de fidelidade (0: Não, 1: Sim)	N/A	N/A	N/A	"0", "1"
WaitTime	float	Média de tempo de espera do cliente para ser atendido no estabelecimento	Minutos	0	N/A	N/A
ServiceRating	int	Indicador de avaliação do serviço pelo cliente	N/A	1	5	"1", "2", "3", "4", "5"
FoodRating	int	Indicador de avaliação da comida pelo cliente	N/A	1	5	"1", "2", "3", "4", "5"
AmbianceRating	int	Indicador de avaliação da ambientação do estabelecimento pelo cliente	N/A	1	5	"1", "2", "3", "4", "5"
HighSatisfaction	string	Valor binário que representa se o cliente está altamente satisfeito (0: Não, 1: Sim)	N/A	N/A	N/A	"0", "1"

3) Carga e Transformação dos Dados

Para realizar o processo de ETL na plataforma Databricks Community Edition, foram realizados os seguintes passos:

1) Acessar o Databricks Community Edition;
2) Clicar em “Compute” e criar um novo *cluster*;
3) Clicar em “New” e em “Table” para criar uma nova tabela;
4) Importar o CSV para dentro do Databricks File System (DBFS);
5) Clicar em “Create Table with UI”;
6) Selecionar o cluster;
7) Clicar em “Preview Table”
8) Configurar os tipos de dados das colunas da seguinte forma:
    a. CustomerID: STRING
    b. Age: INT
    c. Gender: STRING
    d. Income: FLOAT
    e. VisitFrequency: STRING
    f. AverageSpend: FLOAT
    g. PreferredCuisine: STRING
    h. TimeOfVisit: STRING
    i. GroupSize: INT
    j. DiningOccasion: STRING
    k. MealType: STRING
    l. OnlineReservation: STRING
    m. DeliveryOrder: STRING
    n. LoyaltyProgramMember: STRING
    o. WaitTime: FLOAT
    p. ServiceRating: INT
    q. FoodRating: INT
    r. AmbianceRating: INT
    s. HighSatisfaction: STRING

9) Clicar em “Create Table”;
10) Criar um novo *Notebook* e renomear para “MVP Engenharia de Dados”;

Análise: Qualidade dos Dados

Cada linha da tabela “restaurant_customer_satisfaction” representa um cliente distinto identificado por um código de identificação na coluna “CustomerID”.

Os atributos da tabela apresentam características sobre os hábitos de consumo e perfil de cada cliente do estabelecimento.

É importante destacar que a base de dados “restaurant_customer_satisfaction” não representa fatos de visitas dos clientes, mas sim um panorama geral sobre a experiência dos clientes neste restaurante, fornecidas pelos próprios clientes por meio de formulários de avaliação, possivelmente.

Dessa forma, os dados representados na tabela “restaurant_customer_satisfaction” não são apropriados para realizar análises descritivas baseadas em visitas individuais e os indicadores construídos a partir deles podem não representar fielmente a situação real do restaurante, mas sim prover *insights* sobre como a percepção dos clientes sobre os serviços e produtos impactam na sua satisfação geral.

Por não haver dados temporais sobre a data da visita do cliente ao estabelecimento ou a data do registro da pesquisa de satisfação, não é possível realizar uma análise histórica ou com um recorte de tempo definido.

Não é possível mensurar, através dos dados coletados, quantas visitas o estabelecimento recebe periodicamente, pois cada fato da tabela corresponde a uma avaliação de um cliente único.

Com isso em mente, pode-se avaliar cada um dos atributos da seguinte forma:

a. **CustomerID:** existem mil e quinhentos registros únicos nesta coluna, cada um representando um cliente distinto. Nenhum número se repete, todos estão no formato de números inteiros e não há qualquer inconsistência no conjunto de dados.
b. **Age:** representa a idade, em anos, de cada cliente na tabela. O valor mínimo presente na base de dados é 18, o máximo é 69, e não há qualquer inconsistência no conjunto de dados.
c. **Gender:** representa o gênero (masculino ou feminino) do cliente na tabela. Há apenas duas categorias para este atributo e não há qualquer inconsistência no conjunto de dados.
d. **Income:** representa a renda anual dos clientes em dólares americanos (USD). O valor mínimo presente na base de dados é 20012.00 e o máximo é 149875.00. Não há qualquer inconsistência no conjunto de dados.
e. **VisitFrequency:** representa quatro tipos de dados categóricos e não há qualquer inconsistência no conjunto de dados.
f. **AverageSpend:** representa o valor médio gasto pelos clientes do estabelecimento, por visita, em dólares americanos (USD). O valor mínimo presente na base de dados é 10.31 e o máximo é 199.97 e não há qualquer inconsistência no conjunto de dados.

g. **PreferredCuisine:** representa cinco tipos de dados categóricos e não há qualquer inconsistência no conjunto de dados.
h. **TimeOfVisit:** representa 3 tipos de dados categóricos e não há qualquer inconsistência no conjunto de dados.
i. **GroupSize:** representa, em números inteiros, o tamanho do grupo que acompanha o cliente em uma visita ao estabelecimento. Esta informação não é útil neste conjunto de dados pois o mesmo não representa fatos de visitas únicas de clientes.
j. **DiningOccasion:** representa 3 tipos de dados categóricos sobre o motivo da visita do cliente ao estabelecimento. Esta informação não é útil neste conjunto de dados pois o mesmo não representa fatos de visitas únicas de clientes.
k. **MealType:** representa 2 tipos de dados categóricos sobre o tipo de comida que o cliente solicitou no restaurante (para levar ou para comer no local). Esta informação não é útil neste conjunto de dados pois o mesmo não representa fatos de visitas únicas de clientes.
l. **OnlineReservation:** representa, em valores binários (1 ou 0) se o cliente solicitou uma reserva online. Esta informação não é útil neste conjunto de dados pois o mesmo não representa fatos de visitas únicas de clientes.
m. **DeliveryOrder:** representa, em valores binários (1 ou 0) se o cliente solicitou um pedido para entrega em domicílio (*delivery*). Esta informação não é útil neste conjunto de dados pois o mesmo não representa fatos de visitas únicas de clientes.
n. **LoyaltyProgramMember:** representa, em valores binários (1 ou 0), se o cliente é membro do programa de fidelidade do estabelecimento. Esta informação está de acordo com a proposta do conjunto de dados e é útil para descrever o cliente que está sendo analisado na tabela fato e não há qualquer inconsistência no conjunto de dados.
o. **WaitTime:** representa o tempo médio, em minutos, de espera para o cliente ser atendido no estabelecimento. O valor mínimo presente na base de dados é de 0.00 minutos e o máximo é de 59.97 minutos e não há qualquer inconsistência no conjunto de dados.
p. **ServiceRating:** representa, em números inteiros, e em uma escala de 1 a 5, a nota atribuída pelo cliente ao serviço de atendimento do estabelecimento. O valor mínimo presente na base de dados é 1 e o máximo é 5 e não há qualquer inconsistência no conjunto de dados.
q. **FoodRating:** representa, em números inteiros, e em uma escala de 1 a 5, a nota atribuída pelo cliente à comida do estabelecimento. O valor mínimo presente na base de dados é 1 e o máximo é 5 e não há qualquer inconsistência no conjunto de dados.
r. **AmbianceRating:** representa, em números inteiros, e em uma escala de 1 a 5, a nota atribuída pelo cliente ao ambiente do estabelecimento. O valor mínimo presente na base de dados é 1 e o máximo é 5 e não há qualquer inconsistência no conjunto de dados.
s. **HighSatisfaction:** representa, em valores binários (1 ou 0) se o cliente está altamente satisfeito com o estabelecimento ou não. Não há qualquer inconsistência no conjunto de dados.

Análise: Solução do problema

1) Quais são os fatores predominantes que estão influenciando a baixa satisfação do cliente?

Foi utilizada a *query* abaixo para selecionar da tabela "restaurant_customer_satisfaction" a quantidade de clientes altamente satisfeitos (*Highly Satisfied*) e não altamente satisfeitos (*Not Highly Satisfied*), o tempo médio de espera (*Average_Wait_Time*) as médias das notas de avaliação de serviços (*Average_Service_Rating*), comida (*Average_Food_Rating*) e ambiente (*Average_Ambiance_Rating*) e a média geral de avaliação (*General_Average_Rating*).

```
SELECT
DISTINCT count(CustomerID) as Customers,
IF(HighSatisfaction = 1, "Highly Satisfied", "Not Highly Satisfied") as
Satisfaction,
CAST(AVG(WaitTime) as DECIMAL(10,2)) as Average_Wait_Time,
CAST(AVG(ServiceRating) as DECIMAL(10,2)) as Average_Service_Rating,
CAST(AVG(FoodRating) as DECIMAL(10,2)) as Average_Food_Rating,
CAST(AVG(AmbianceRating) as DECIMAL(10,2)) as Average_Ambiance_Rating,
CAST((AVG(ServiceRating) + AVG(FoodRating) + AVG(AmbianceRating))/3 as
DECIMAL(10,2)) as General_Average_Rating
FROM restaurant_customer_satisfaction
GROUP BY HighSatisfaction
ORDER BY General_Average_Rating DESC
```

	Customers	Satisfaction	.00 Average_Wait_Time	.00 Average_Service_Rating	.00 Average_Food_Rating	.00 Average_Ambiance_Rating	.00 General_Average_Rating
1	201	Highly Satisfied	23.87	3.37	3.50	3.27	3.38
2	1299	Not Highly Satisfied	31.14	2.99	2.92	2.94	2.95

O resultado demonstrou que apenas uma pequena porção dos clientes (13,4%) consideram estar altamente satisfeitos com o estabelecimento. As médias das notas de avaliação de serviços (*Average_Service_Rating*), comida (*Average_Food_Rating*) e ambiente (*Average_Ambiance_Rating*) apresentaram pouca variação. No entanto, entre os clientes altamente satisfeitos, a média de avaliação da comida se destaca. É possível notar, também, que a média geral de avaliação dos clientes altamente satisfeitos supera em 0,43 pontos a média dos clientes não altamente satisfeitos. Fica evidente, através desta consulta, que o tempo médio de espera é um dos fatores que influencia no índice de alta satisfação do cliente. Enquanto os clientes altamente satisfeitos apresentam uma média de 23,87 minutos de tempo de espera no estabelecimento, os clientes não altamente satisfeitos possuem um tempo médio de espera de 31,14 minutos.

A *query* abaixo indica que, apesar de ser o turno (*TimeOfVisit*) com a melhor média de tempo de espera, o turno do almoço (*Lunch*) é o pior avaliado e isso se reflete na sua avaliação em relação à comida.

```
SELECT
TimeOfVisit,
CAST(AVG(WaitTime) as DECIMAL(10,2)) as Average_Wait_Time,
CAST(AVG(ServiceRating) as DECIMAL(10,2)) as Average_Service_Rating,
CAST(AVG(FoodRating) as DECIMAL(10,2)) as Average_Food_Rating,
CAST(AVG(AmbianceRating) as DECIMAL(10,2)) as Average_Ambiance_Rating,
CAST((AVG(ServiceRating) + AVG(FoodRating) + AVG(AmbianceRating))/3 as
DECIMAL(10,2)) as General_Average_Rating
FROM restaurant_customer_satisfaction
GROUP BY TimeOfVisit
ORDER BY General_Average_Rating DESC
```

	TimeOfVisit	.00 Average_Wait_Time	.00 Average_Service_Rating	.00 Average_Food_Rating	.00 Average_Ambiance_Rating	.00 General_Average_Rating
1	Dinner	30.69	3.06	3.07	2.99	3.04
2	Breakfast	30.90	3.01	3.02	3.00	3.01
3	Lunch	28.90	3.06	2.90	2.97	2.98

As duas *querys* a seguir demonstram que o tempo médio de espera influencia diretamente na avaliação dos clientes e no seu grau de satisfação com o estabelecimento.

```
SELECT
TimeOfVisit,
CAST(AVG(WaitTime) as DECIMAL(10,2)) as Average_Wait_Time,
CAST(AVG(ServiceRating) as DECIMAL(10,2)) as Average_Service_Rating,
CAST(AVG(FoodRating) as DECIMAL(10,2)) as Average_Food_Rating,
CAST(AVG(AmbianceRating) as DECIMAL(10,2)) as Average_Ambiance_Rating,
CAST((AVG(ServiceRating) + AVG(FoodRating) + AVG(AmbianceRating))/3 as
DECIMAL(10,2)) as General_Average_Rating
FROM restaurant_customer_satisfaction
WHERE HighSatisfaction = 1
GROUP BY TimeOfVisit
ORDER BY General_Average_Rating DESC
```

	TimeOfVisit	.00 Average_Wait_Time	.00 Average_Service_Rating	.00 Average_Food_Rating	.00 Average_Ambiance_Rating	.00 General_Average_Rating
1	Breakfast	24.37	3.58	3.61	3.29	3.50
2	Lunch	22.01	3.44	3.39	3.44	3.42
3	Dinner	25.04	3.07	3.48	3.10	3.22

Os clientes altamente satisfeitos apresentam uma média de tempo de espera significantemente menor quando comparados aos clientes que não estão altamente satisfeitos. Além disso, as médias das notas de avaliação tendem a ser maiores para os clientes com tempo de espera menor.

```
SELECT
TimeOfVisit,
CAST(AVG(WaitTime) as DECIMAL(10,2)) as Average_Wait_Time,
CAST(AVG(ServiceRating) as DECIMAL(10,2)) as Average_Service_Rating,
CAST(AVG(FoodRating) as DECIMAL(10,2)) as Average_Food_Rating,
CAST(AVG(AmbianceRating) as DECIMAL(10,2)) as Average_Ambiance_Rating,
CAST((AVG(ServiceRating) + AVG(FoodRating) + AVG(AmbianceRating))/3 as
DECIMAL(10,2)) as General_Average_Rating
FROM restaurant_customer_satisfaction
WHERE HighSatisfaction = 0
GROUP BY TimeOfVisit
ORDER BY General_Average_Rating DESC
```

	TimeOfVisit	.00 Average_Wait_Time	.00 Average_Service_Rating	.00 Average_Food_Rating	.00 Average_Ambiance_Rating	.00 General_Average_Rating
1	Dinner	31.58	3.06	3.00	2.97	3.01
2	Breakfast	31.99	2.92	2.92	2.95	2.93
3	Lunch	29.87	3.00	2.84	2.91	2.92

Dessa forma, os indicadores apontam para o fato de a não satisfação dos clientes do estabelecimento estar ligada, predominantemente, ao tempo de espera e à qualidade da comida do turno do almoço.

2) Como aumentar a frequência dos clientes fidelizados no estabelecimento?

Os clientes, de forma geral, frequentam o estabelecimento semanalmente e mensalmente de forma majoritária, conforme pode ser observado através da *query* abaixo.

```
SELECT
VisitFrequency,
count(CustomerID) as Customers
FROM restaurant_customer_satisfaction_csv
GROUP BY VisitFrequency
ORDER BY 2 DESC
```

	VisitFrequency	Customers
1	Weekly	606
2	Monthly	428
3	Rarely	313
4	Daily	153

E de forma geral, a culinária preferida dos clientes do estabelecimento é a italiana, seguida pela chinesa, mexicana, indiana e americana.

```sql
SELECT
PreferredCuisine,
count(CustomerID) as Customers
FROM restaurant_customer_satisfaction_csv
GROUP BY PreferredCuisine
ORDER BY 2 DESC
```

	PreferredCuisine	Customers
1	Italian	325
2	Chinese	310
3	Mexican	299
4	Indian	296
5	American	270

Dentre os clientes que frequentam o estabelecimento com maior frequência (diariamente ou semanalmente), a culinária preferida é a chinesa, seguida pela mexicana, indiana, italiana e americana.

```sql
SELECT
PreferredCuisine,
count(CustomerID) as Customers
FROM restaurant_customer_satisfaction_csv
WHERE VisitFrequency = "Daily" or VisitFrequency = "Weekly"
GROUP BY PreferredCuisine
ORDER BY 2 DESC
```

	PreferredCuisine	Customers
1	Chinese	163
2	Mexican	161
3	Indian	160
4	Italian	151
5	American	124

A culinária preferida dos membros do programa de fidelidade é a chinesa, seguida da italiana, mexicana, indiana e americana.

```sql
SELECT
PreferredCuisine,
count(CustomerID) as Customers
FROM restaurant_customer_satisfaction_csv
WHERE LoyaltyProgramMember = 1
GROUP BY PreferredCuisine
ORDER BY 2 DESC
```

	PreferredCuisine	Customers
1	Chinese	159
2	Italian	148
3	Mexican	140
4	Indian	137
5	American	136

Finalmente, para os membros do programa de fidelidade e que frequentam o estabelecimento diariamente ou semanalmente, a culinária preferida é a mexicana, seguida da chinesa, indiana, italiana e americana.

```
SELECT
PreferredCuisine,
count(CustomerID) as Customers
FROM restaurant_customer_satisfaction_csv
WHERE LoyaltyProgramMember = 1 and VisitFrequency = "Daily" or VisitFrequency
= "Weekly"
GROUP BY PreferredCuisine
ORDER BY 2 DESC
```

	PreferredCuisine	Customers
1	Mexican	145
2	Chinese	143
3	Indian	141
4	Italian	139
5	American	114

Através da consulta a seguir, é possível identificar que os clientes membros do programa de fidelidade visitam o estabelecimento semanalmente ou mensalmente, em maioria. Muitos poucos frequentam o estabelecimento diariamente.

```
SELECT
VisitFrequency,
avg(AverageSpend),
count(CustomerID) as Customers
FROM restaurant_customer_satisfaction_csv
WHERE LoyaltyProgramMember = 1
GROUP BY VisitFrequency
ORDER BY 3 DESC
```

	VisitFrequency	1.2 avg(AverageSpend)	Customers
1	Weekly	105.79278510224586	306
2	Monthly	103.73814149041777	206
3	Rarely	107.22125004999566	132
4	Daily	103.98995115882472	76

É possível perceber que uma visita diária ao estabelecimento custa aos clientes em torno de 103 dólares. Para atrair mais clientes que já são membros do programa de fidelidade, a gerência pode optar por disponibilizar pratos mais acessíveis como um menu executivo, ou oferecendo descontos, convênios com empresas e promoções diárias, reduzindo o custo individual de um cliente e aumentando o tráfego diário. Além disso, pode ser interessante disponibilizar mais refeições de culinária mexicana, chinesa ou italiana, sendo essas as preferidas entre os clientes que já visitam diariamente e semanalmente o estabelecimento, e que são membros do programa de fidelidade.

Análise: Conclusão

Ao longo deste trabalho, foram realizadas diversas análises que produziram *insights* valiosos para aprimorar o índice de satisfação de clientes de um restaurante e para elaboração de estratégias de marketing para prospecção de clientes e aumento da receita e do movimento do estabelecimento.

Através de uma análise simples, porém concisa, utilizando a linguagem SQL foi possível identificar os principais fatores que podem influenciar na satisfação (ou falta dela) de um cliente com o estabelecimento. A utilização de operadores lógicos como o "Avg" juntamente com os comandos GROUP BY e WHERE são muito eficientes para resumir e agrupar informações e, dessa forma, comparar diferentes cenários.

Para esta análise e para a resolução dos dois problemas propostos neste trabalho, foram cruzados mais de dois dados em quase todos os exemplos, permitindo realizar análises descritivas e diagnósticas validadas por diferentes cenários. O auxílio de ferramentas de visualização de dados, como o Microsoft Power BI, embora não seja parte do escopo deste trabalho, seria de grande valia para o cruzamento de ainda mais dados e para proporcionar uma visão gráfica mais ampla, permitindo-se aprofundar ainda mais em outras questões relevantes para uma análise de dados em um ambiente de Business Intelligence.

Porém, pelo fato de a base de dados apresentar algumas informações inconsistentes com o propósito do fato a ser analisado, não foi possível aprofundar ainda mais a análise. Uma tabela onde o fato analisado fosse as vendas de um restaurante, unidas com uma pesquisa de satisfação dos clientes após cada visita, apoiada por tabelas dimensão com informações de clientes, produtos, serviços e mais seria o ideal para a elaboração de um modelo estrela eficiente e robusto o suficiente para uma análise mais aprofundada e que seria capaz de retratar a situação real do estabelecimento em questão.