Muuve Now

Integrantes:

Vinicius Miranda Andrade Piovesan RA: 23025544

Matheus de Medeiros Takaki RA: 23024683

Sérgio Ricardo Pedote Junior RA:23747441

Felipe Ribeiro Almeida RA: 23025143

Analise Inicial dos Dados

Colab :
https://colab.research.google.com/drive/1SAsnJBhogY6gy1PVk6PFVvuwNTLKnT7p?usp=sharing

Tabela Product:

- **ProductID:**
  - **Identificador único do produto.**
- **Observações**:
  - Contém uma mistura de valores alfanuméricos curtos ("99POP", "99TAXI") e hashes longos, cujo hashes não pertencem a essa coluna. ("46cec5c4d23e57bcba2122677eb8c759").
  - Os valores curtos parecem ser nomes de produtos ou serviços.
  - Não há valores ausentes nesta coluna.
  -]
- **Qualidade:**
  - Falta de padronização: alguns valores são legíveis, enquanto outros são hashes criptográficos, o que pode indicar diferentes sistemas de geração de IDs.
  - Qualidade não muito boa.
- **ProviderID:**

    - **Identificador do provedor do produto/serviço.**

- **Observações:**
    - Muitos valores estão ausentes (vazios). Apenas algumas linhas têm valores preenchidos (e.g., "1", "4").
    - Os valores presentes são numéricos e são IDs de provedores.

- **Qualidade:**

    - Alta quantidade de valores ausentes,
    - Qualidade não muito boa.

- **CategoryID:**
    - **Identificador da categoria do produto/serviço.**

- **Observações:**
    - Todos os valores estão preenchidos.
    - Os valores são numéricos e parecem representar categorias específicas ("1", "2", "3", "4").
    - A distribuição das categorias parece variar, com algumas categorias aparecendo com mais frequência em outras

- **Qualidade:**
  Qualidade boa, todos os dados preenchidos

- **Description:**

    - **Descrição do produto/serviço.**

- **Observações:**

    - Todos os valores estão preenchidos.

    - Contém uma mistura de descrições claras ("Taxi Comum", "Cabify Lite") e descrições repetitivas ou pouco informativas ("99POP", "99ENTREGA")

- Algumas descrições parecem ser variantes de uma mesma categoria como: ("99ENTREGA" e "99ENTREGAMOTO").
- Há descrições que incluem informações adicionais, como descontos ("Easy Taxi Corp (-15%)").

- **Qualidade:**
  - Falta de padronização nas descrições, o que pode dificultar a análise e a categorização.
  - Qualidade razoável.

```
                          ProductID  ProviderID  CategoryID  \
0                             99POP           3           1
1                                 1           1           5
2                                 2           1           6
3  46cec5c4d23e57bcba2122677eb8c759           4           5
4  5fc141256dc70a394d0ce4c5c1444dfc           4           5

             Description
0                  99POP
1              Taxi Comum
2               Executivo
3  Easy Taxi Corp (-15%)
4                   Taxi
```

	ProductID	ProviderID	CategoryID	Description
0	99POP	3	1	99POP
1	1	1	5	Taxi Comum
2	2	1	6	Executivo
3	46cec5c4d23e57bcba2122677eb8c759	4	5	Easy Taxi Corp (-15%)
4	5fc141256dc70a394d0ce4c5c1444dfc	4	5	Taxi
...	...	...	...	...
232	WPP-45-1	5	1	Green
233	WPP5	5	4	Black
234	WPP-5-5	5	4	Black
235	WPP7	5	1	Economy
236	WPP-7-6	5	1	Economy

237 rows × 4 columns

Tabela Ride_v2:

- **RideID**
    - **(Identificador da Corrida)**
- **Observações**
    - O campo contém identificadores únicos para cada corrida, sem valores duplicados ou nulos. Isso garante que cada registro seja individualmente rastreável.
- **Qualidade do dado:** Boa.

- **UserID**
    - (Identificador do Usuário)
- **Observações**
    - Os valores seguem um formato UUID válido e não há registros vazios. Isso assegura que cada usuário tenha um identificador único, permitindo análises individualizadas.
- **Qualidade do dado:** Boa.

- **Schedule**
    - (Data e Hora Agendada da Corrida)
- **Observações**
    - O formato da data e hora está consistente (YYYY-MM-DD HH:MM:SS.ssssss), e os dados parecem corretamente preenchidos. Essa informação é essencial para entender padrões de uso do serviço.
- **Qualidade do dado:** Boa.

- **Create**

- (Data e Hora de Criação da Corrida)

- **Observações**

  - Os registros seguem o mesmo formato correto da coluna "Schedule". Esse campo pode ser usado para comparar o tempo entre o pedido e o início da corrida.

- **Qualidade do dado:** Boa.

- **RideStatusID**

  - (Status da Corrida)

- **Observações**

  - A coluna contém apenas valores numéricos sem registros vazios, o que indica que cada corrida tem um status definido. É fundamental para monitoramento e análises operacionais.

- **Qualidade do dado:** Boa.

- **CompanyID**
  - (Identificador da Empresa)

- **Observações**

  - Os valores estão limitados a apenas dois identificadores distintos (2 e 3), sem registros vazios. Isso indica que o serviço de transporte é operado por duas empresas diferentes.

- **Qualidade do dado:** Boa.

- **ProviderID**

  - (Identificador do Provedor do Serviço)

- **Observações**
    - Existem registros nulos nesta coluna, o que pode indicar que algumas corridas não tiveram um provedor de serviço identificado corretamente. Isso pode gerar dificuldades na rastreabilidade das corridas.
- **Qualidade do dado:** Regular.

- **RideProviderID**
    - (Identificador da Corrida no Provedor)
- **Observações**
    - Há valores nulos, o que pode significar que algumas corridas não possuem uma associação clara com um provedor. Caso seja um dado obrigatório, essa inconsistência precisa ser corrigida.
- **Qualidade do dado:** Regular.

- **Price**
    - (Valor da Corrida)
- **Observações**
    - Os valores numéricos estão geralmente corretos, mas existem registros com **0.00**, o que pode indicar falhas no sistema de tarifação, corridas gratuitas ou testes internos. Também há valores altos que podem ser anomalias. É importante revisar os critérios de precificação.
- **Qualidade do dado:** Regular.

- **Updated**
    - (Última Atualização do Registro)

- **Observações**
    - Os registros seguem o formato correto de data e hora, indicando quando a última modificação foi feita no sistema. Esse dado é útil para auditorias e rastreamento de mudanças nos registros.

- **Qualidade do dado:** Boa.

- **CategoryID**
    - (Categoria da Corrida)

- **Observações**
    - A coluna contém muitos valores nulos, sugerindo que essa informação não está sendo preenchida de forma consistente. Caso seja uma informação importante para segmentação ou relatórios, recomenda-se torná-la obrigatória no sistema.

- **Qualidade do dado:** Ruim.

- **TotalUsers**
    - (Número de Usuários na Corrida)

- **Observações**
    - Os valores estão consistentes, com a maioria das corridas registradas com **1** usuário e algumas com **2**. Isso indica que o sistema pode suportar corridas compartilhadas.

- **Qualidade do dado:** Boa.

- **Car**
    - (Modelo do Carro Usado na Corrida)

- **Observações**

- Há muitos registros nulos, o que pode indicar que essa informação não é obrigatória ou não está sendo registrada corretamente. Caso seja relevante para o controle de frota ou monitoramento da qualidade do serviço, é recomendável padronizar o preenchimento.

- **Qualidade do dado:** Ruim.

- **RideDriverLocationID**
  - (Localização do Motorista)

- **Observações**
  - A maioria dos registros contém valores nulos, sugerindo que essa informação só é registrada em certas circunstâncias ou categorias de corrida. Pode ser um problema caso o acompanhamento da localização seja necessário para auditorias ou relatórios operacionais.

- **Qualidade do dado:** Ruim.

- **ScheduledRide**
  - (Indica se a Corrida foi Agendada)

- **Observações**
  - Todos os valores são **0**, o que sugere que nenhuma corrida foi marcada com antecedência. Se o sistema suporta corridas agendadas, isso pode indicar um problema na coleta dos dados ou que essa funcionalidade não está sendo utilizada pelos usuários.

- **Qualidade do dado:** Boa.

	RideID	UserID	Schedule	Create	RideStatusID	CompanyID	ProviderID	RideProviderID	price	Update	CategoryID	TotalUsers	Car	RideDriverLocationID	ScheduledRide
0	1685755	e15b8cc3-5a67-4630-b89f-ee69f302b582	2025-02-10 14:31:10.8858446	2025-02-10 14:31:10.9084221	1	2	NaN	NaN	0.00	2025-02-10 14:31:10.9084233	NaN	1.0	NaN	NaN	0.0
1	1685754	5c3fb011-0aea-429a-8305-b08953b77df1	2025-02-10 14:26:35.3411403	2025-02-10 14:26:35.4169873	2	230	5.0	NaN	30.45	2025-02-10 14:28:02.4656963	NaN	1.0	NaN	NaN	0.0
2	1685753	d7e2f4dc-337f-45f5-b762-67b72b077abc	2025-02-10 14:23:45.2540905	2025-02-10 14:24:32.7058722	2	52	3.0	NaN	11.40	2025-02-10 14:24:46.5037165	NaN	1.0	NaN	NaN	0.0
3	1685752	2125ed9c-89b8-4df6-9be6-53195397a269	2025-02-10 14:23:12.9838635	2025-02-10 14:23:12.9975475	8	230	36.0	1589157	45.79	2025-02-10 14:30:30.6031123	5.0	1.0	VW VIRTUS CL / BRANCA	18361.0	0.0
4	1685751	72cbebfb-5d70-49ab-ab23-8e3b57c7e399	2025-02-10 14:19:30.5937678	2025-02-10 14:19:30.6117184	2	2	3.0	NaN	17.28	2025-02-10 14:24:45.9711764	NaN	1.0	NaN	NaN	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
24872	1660883	1fc86578-3770-4ba7-a05f-afc68102f75f	2024-11-01 19:57:31.8324637	2024-11-01 19:57:31.8563497	2	142	2.0	NaN	50.16	2024-11-01 19:57:42.7406408	NaN	1.0	NaN	NaN	0.0
24873	1660882	1fc86578-3770-4ba7-a05f-afc68102f75f	2024-11-01 19:56:52.4273040	2024-11-01 19:56:52.4503192	2	142	3.0	NaN	43.42	2024-11-01 19:57:14.0060087	NaN	1.0	NaN	NaN	0.0
24874	1660881	e44a08e1-2946-43af-b265-eda7ff159518	2024-11-01 19:56:12.7656081	2024-11-01 19:56:12.7834294	2	211	12.0	NaN	30.00	2024-11-01 19:58:46.1997229	NaN	1.0	NaN	NaN	0.0
24875	1660880	c17e81ee-4db0-4be5-8865-ed518af7ea75	2024-11-01 19:55:08.9414797	2024-11-01 19:55:08.9812784	2	78	3.0	NaN	22.21	2024-11-01 19:55:42.9463469	NaN	1.0	NaN	NaN	0.0
24876	1660879	e44a08e1-2946-43af-b265-eda7ff159518	2024-11-01 19:52:54.5077395	2024-11-01 19:52:54.5336227	2	211	3.0	NaN	25.15	2024-11-01 19:	NaN	NaN	None	NaN	NaN

Tabela RideEstimative_v3

- **RideEstimativeID**
  - (Identificador da Estimativa de Corrida)

- **Observações**
  - O campo contém identificadores únicos para cada estimativa de corrida, sem valores duplicados ou nulos. Isso garante que cada registro seja individualmente rastreável.

• **Qualidade do dado:** Boa.

- **RideID**
  - (Identificador da Corrida)

- **Observações**
  - Os valores são numéricos e parecem estar corretamente preenchidos, associando cada estimativa a uma corrida específica. Não há registros vazios.

- **Qualidade do dado:** Boa.

- **ProductID**
  - (Tipo de Serviço ou Produto da Corrida)

- **Observações**
  - Contém diferentes tipos de serviços, como "Flash", "UberX", "Comfort", "pop99", "regular-taxi" e "Moto". A categorização parece consistente.

- **Qualidade do dado:** Boa.

- **WaitingTime**
  - (Tempo de Espera Estimado)

- **Observações**
  - O tempo de espera está expresso em minutos e não há valores vazios. Porém, há variações grandes, sendo importante validar se os valores refletem corretamente os tempos médios esperados.

- **Qualidade do dado:** Boa.

- **Price**
  - (Preço da Corrida Estimada)

- **Observações**

- Os valores numéricos estão corretamente preenchidos, sem valores nulos. No entanto, há algumas estimativas de preços muito baixas, o que pode indicar promoções, erros ou tarifas especiais.

- **Qualidade do dado:** Boa.

- **FareID**
  - (Identificador da Tarifa Aplicada)

- **Observações**
  - Muitos registros possuem valores **NULL**, o que pode indicar que nem todas as estimativas possuem uma tarifa específica associada.
- **Qualidade do dado:** Regular.

- **Selected**
  - (Indica se a estimativa foi selecionada pelo usuário)

- **Observações**
  - O campo apresenta apenas os valores **0** e **1**, onde **1** indica que a estimativa foi escolhida. O preenchimento está correto, permitindo identificar quais estimativas resultaram em uma corrida confirmada.

- **Qualidade do dado:** Boa.

- **RideReasonSelectedEstimativeID**
  - (Motivo da Escolha da Estimativa)

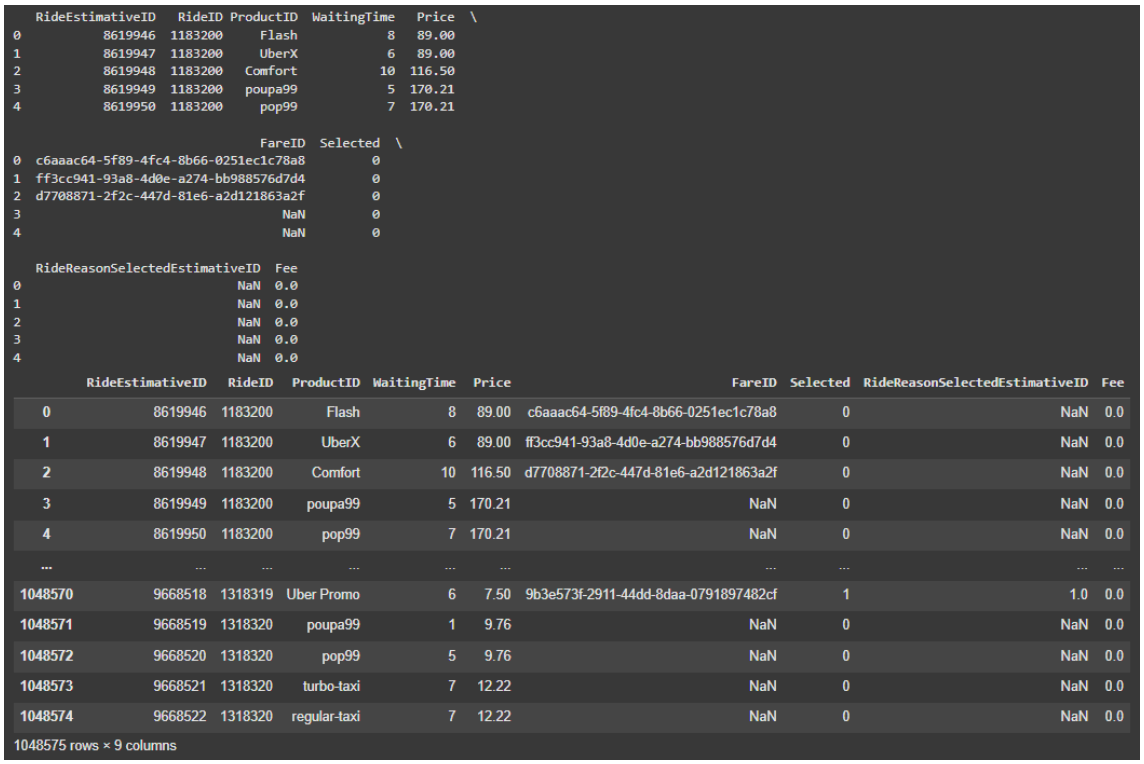
- **Observações**
  - A maioria dos registros contém valores **NULL**, o que indica que raramente há um motivo registrado para a escolha da estimativa. Se for um dado relevante para análise do comportamento do usuário, pode ser necessário incentivar o preenchimento desse campo.

- **Qualidade do dado:** Ruim.

- **Fee**
  - (Taxa Adicional)

- **Observações**
  - Todos os registros possuem **0.00**, indicando que nenhuma corrida teve taxa adicional aplicada. Caso o sistema permita taxas extras, pode ser interessante validar se esse campo está sendo atualizado corretamente.

- **Qualidade do dado:** Regular.



```
   RideEstimativeID   RideID ProductID  WaitingTime   Price  \
0           8619946  1183200     Flash            8   89.00
1           8619947  1183200     UberX            6   89.00
2           8619948  1183200   Comfort           10  116.50
3           8619949  1183200   poupa99            5  170.21
4           8619950  1183200     pop99            7  170.21

                                FareID  Selected  \
0  c6aaac64-5f89-4fc4-8b66-0251ec1c78a8         0
1  ff3cc941-93a8-4d0e-a274-bb988576d7d4         0
2  d7708871-2f2c-447d-81e6-a2d121863a2f         0
3                                   NaN         0
4                                   NaN         0

   RideReasonSelectedEstimativeID  Fee
0                             NaN  0.0
1                             NaN  0.0
2                             NaN  0.0
3                             NaN  0.0
4                             NaN  0.0
```

	RideEstimativeID	RideID	ProductID	WaitingTime	Price	FareID	Selected	RideReasonSelectedEstimativeID	Fee
0	8619946	1183200	Flash	8	89.00	c6aaac64-5f89-4fc4-8b66-0251ec1c78a8	0	NaN	0.0
1	8619947	1183200	UberX	6	89.00	ff3cc941-93a8-4d0e-a274-bb988576d7d4	0	NaN	0.0
2	8619948	1183200	Comfort	10	116.50	d7708871-2f2c-447d-81e6-a2d121863a2f	0	NaN	0.0
3	8619949	1183200	poupa99	5	170.21	NaN	0	NaN	0.0
4	8619950	1183200	pop99	7	170.21	NaN	0	NaN	0.0
...	...	...	...	...	...	...	...	...	...
1048570	9668518	1318319	Uber Promo	6	7.50	9b3e573f-2911-44dd-8daa-0791897482cf	1	1.0	0.0
1048571	9668519	1318320	poupa99	1	9.76	NaN	0	NaN	0.0
1048572	9668520	1318320	pop99	5	9.76	NaN	0	NaN	0.0
1048573	9668521	1318320	turbo-taxi	7	12.22	NaN	0	NaN	0.0
1048574	9668522	1318320	regular-taxi	7	12.22	NaN	0	NaN	0.0

1048575 rows × 9 columns

Tabela RideAddres_v1

- **RideAddressID**
  - (Identificador do Endereço da Corrida)

- **Observações**
  - O campo contém identificadores únicos para cada endereço registrado. Ele garante que cada localização possa ser rastreada corretamente dentro do banco de dados.

- **Qualidade do dado:** Boa.

- **Address**
    - (Endereço Completo)

- **Observações**
    - Contém os endereços completos das corridas, incluindo rua, número, cidade, estado e país. Algumas inconsistências podem ser encontradas, como endereços incompletos ou com formatos diferentes. Pode ser necessário padronizar os dados para evitar problemas de análise.

- **Qualidade do dado:** Regular.

- **Street**
    - (Nome da Rua)

- **Observações**
    - O nome da rua está bem preenchido, mas pode haver variações na grafia ou casos em que a rua está ausente. Uma limpeza e padronização podem melhorar a qualidade dos dados.

- **Qualidade do dado:** Boa.

- **Number**
    - (Número do Endereço)

- **Observações**
    - O campo parece estar preenchido corretamente na maioria dos casos, mas há alguns valores ausentes ou inválidos (como "s/n" para sem número).

- **Qualidade do dado:** Regular.

- **Neighborhood**
  - (Bairro)

- **Observações**
  - Algumas células possuem valores ausentes ou inconsistentes. O preenchimento pode estar incompleto em determinados casos, o que pode dificultar a segmentação por região.

- **Qualidade do dado:** Regular.

- **City**
  - (Cidade)

- **Observações**
  - O campo de cidade está geralmente bem preenchido, mas há alguns registros ausentes ou inconsistentes. Algumas cidades podem estar duplicadas com diferentes grafias, exigindo normalização.

- **Qualidade do dado:** Boa.

- **State**
  - (Estado)

- **Observações**
  - O campo está geralmente correto, mas há algumas células vazias ou com abreviações inconsistentes. Idealmente, todos os estados devem seguir um padrão único (exemplo: "SP" ou "São Paulo").

- **Qualidade do dado:** Regular.

- **Lat**
  - (Latitude)

- **Observações**
  - Os valores de latitude devem ser coordenadas numéricas válidas. Há casos de valores inválidos, como números extremamente altos ou "999.999". Esses registros devem ser tratados para garantir a precisão geográfica.

- **Qualidade do dado:** Ruim.

- **Lng**
  - (Longitude)

- **Observações**
  - Assim como a latitude, a longitude deve conter coordenadas numéricas reais. Alguns valores parecem estar incorretos ou ausentes, o que pode comprometer a análise geoespacial dos dados.

- **Qualidade do dado:** Ruim.

- **RideAddressTypeID**
  - (Tipo de Endereço da Corrida)

- **Observações**
  - O campo parece estar preenchido corretamente, categorizando os endereços de acordo com sua função. Não foram identificados problemas aparentes.

- **Qualidade do dado:** Boa.

- **RideID**
    - (Identificador da Corrida)

- **Observações**
    - Está corretamente preenchido e permite a associação dos endereços com as corridas registradas. Não foram encontrados valores nulos ou duplicados.

- **Qualidade do dado:** Boa.

	RideAddressID	Address	Street	Number	Neighborhood	City	State	Lat	Lng	RideAddressTypeID	RideID
0	2334277	Rua João Pinheiro, 585 - Rua João Pinheiro - B...	Rua João Pinheiro	585	Rua João Pinheiro	NaN	Brasil	-26.329.754.299.999.900	-48.840.427.999.999.900	1	1183200
1	2334278	Av. Dr. Nereu Ramos, 450 - Rocio Grande, São F...	Av. Dr. Nereu Ramos, 450 - Rocio Grande, São F...	450	NaN	NaN	NaN	-262.554.657	-486.434.197	2	1183200
2	2334279	Rodovia Rafael da Rocha Pires, 1883 - Rodovia ...	Rodovia Rafael da Rocha Pires	1883	Rodovia Rafael da Rocha Pires	NaN	Brasil	-274.919.788	-48.528.287.999.999.900	1	1183201
3	2334280	Angeloni Ingleses (Florianópolis) - Supermerca...	Angeloni Ingleses (Florianópolis) - Supermercado	6375	NaN	NaN	NaN	-274.371.486	-4.839.824.309.999.990	2	1183201
4	2334281	Rua Barão do Rio Branco, 12 - Rua Barão do Rio...	Rua Barão do Rio Branco	12	Rua Barão do Rio Branco	NaN	Brasil	-198.495.799	-44.019.915.999.999.900	1	1183202
...	...	...	...	...	...	...	...	...	...	...	...
999995	3336834	Av. Paulista - Bela Vista, São Paulo - SP, Bra...	Av. Paulista - Bela Vista, São Paulo - SP, Brasil	NaN	NaN	São Paulo	SP	-235.657.393	-466.512.379	2	1685755
999996	3336835	Secretaria de Orçamento Federal, - Sepn 516 s...	Secretaria de Orçamento Federal	NaN	Sepn 516 s/n - Asa Norte, Brasília - DF, 70770...	Brasília	DF	-157.399.458	-478.980.851	1	1685756
999997	3336836	Bloco K - Esplanada dos Ministérios, Bloco K -...	Bloco K - Esplanada dos Ministérios	Bloco K	Bloco K - Esplanada dos Ministérios, Bloco K	Brasília	DF	-157.947.356	-478.738.645	2	1685756
999998	3336837	Óticas Brasiliense | Lentes Multifocais Brasil...	Óticas Brasiliense | Lentes Multifocais Brasil...	NaN	SHCS CRS 504 Bloco A Loja 06 - Asa Sul, Brasil...	NaN	DF	-158.046.238	-4.789.689.480.000.000	1	1685757
999999	3336838	Hospital Sírio-Libanês | Brasília, - SGAS II ...	Hospital Sírio-Libanês | Brasília	NaN	SGAS II SGAS 613 s/n Lote 94 - Asa Sul, Brasil...	NaN	DF	-158.346.166	-4.791.132.940.000.000	2	1685757

1000000 rows × 11 columns

Qualidade dos Dados

- A maioria dos identificadores únicos (**RideID, RideEstimativeID, RideAddressID**) estão bem estruturados e não apresentam valores duplicados ou nulos. Isso garante rastreabilidade e confiabilidade para cruzamento de informações.

- Os endereços (**Address, Street, Neighborhood, City, State**) possuem alguns problemas de padronização e inconsistências, como valores ausentes ou variações de escrita.
- Os preços e tempos de espera (**Price, WaitingTime**) parecem estar corretamente preenchidos, mas seria interessante verificar a distribuição dos valores para identificar possíveis outliers.
- Os dados geográficos (**Lat, Lng**) apresentam falhas graves, com valores incorretos ou placeholders inválidos (ex: "999.999"). Isso compromete a análise espacial e exige uma correção ou enriquecimento dos dados.

Conclusão Final

Os dados possuem boa estrutura e rastreabilidade, mas apresentam desafios relacionados à padronização e qualidade, especialmente nos endereços e coordenadas geográficas.