FUNDAÇÃO ESCOLA DE COMÉRCIO ALVARES PENTEADO

Análise dos dados iniciais

SÃO PAULO

2025

Integrantes:

Isaac Santos RA: 23025417

Caroline Gomes RA: 23024619

Icaro Luiz RA: 23025413

Giovanne Braga RA: 23025648

Sumário

Introdução

O objetivo desse documento é realizar uma análise inicial das bases de dados fornecidas pela empresa Khipo, e nisso entender os dados que compõem essas bases históricas a fim de realizar uma filtragem dos dados que serão utilizados para compor os modelos de Machine Learning.

SCRIPT INICIAL

Para realizar a abertura da base histórica, utilizamos a biblioteca “Pandas” dentro do ecossistema “Colab”, como vemos na imagem abaixo:

```
[3]  import pandas as pd

[7]  # Importação dos arquivos
     df1 = pd.read_csv("product.csv", delimiter=";")
     df2 = pd.read_csv("ride_v2.csv", delimiter=";")
     df3 = pd.read_csv("rideaddress_v1.csv", delimiter=";")
     df4 = pd.read_csv("rideestimative_v3.csv", delimiter=";")


[8]  for i, df in enumerate([df1, df2, df3, df4], 1):
         print(f"\n🔹 Base {i}:")

         print(" Primeiras linhas:")
         print(df.head())

         print("\n Informações gerais:")
         print(df.info())

         print("\n Estatísticas descritivas:")
         print(df.describe())

         print("\n Valores nulos por coluna:")
         print(df.isnull().sum())

         print("\n Valores duplicados:", df.duplicated().sum())

         print("-" * 50)
```

Os comandos utilizados são para exibir um relatório bruto dos dados carregados.

```
if df1 is not None:
    print("\n Tabela Product:")
    display(df1.head(50))

if df2 is not None:
    print("\n Tabela Ride:")
    display(df2.head(50))

if df3 is not None:
    print("\n Tabela RideAddress:")
    display(df3.head(50))

if df4 is not None:
    print("\n Tabela RideEstimative:")
    display(df4.head(50))
```

Esse comando foi utilizado para estruturar os dados em uma tabela para facilitar a análise.

ANÁLISE DOS DADOS:

Tabela Product:

Tabela Product:

	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
5	7da40add1baab37e19cae629f6907b13	4	2	Cabify Lite Corp
6	99COMFORT	3	9	99COMFORT
7	99ENTREGA	3	10	99ENTREGA
8	99ENTREGA MOTO	3	1	99ENTREGA MOTO
9	99ENTREGAMOTO	3	10	99ENTREGAMOTO
10	99PLUS	3	1	99PLUS

Significados dos campos da tabela product:

ProductID: Traz o ID da empresa utilizada no transporte (apesar de ter problemas de estruturação, como exibição dos nomes das empresas e tokens);

ProviderID: ID da empresa que realizou o transporte;

CategoryID: A ideia provavelmente é ser o ID da categoria, porém é uma coluna inconsistente;

Description: Descrição/ empresa responsável pelo serviço.

Considerações: A tabela product traz uma qualidade mediana dos dados, um problema recorrente é na coluna de "CategoryID" no qual não mantém uma consistência, em tese é uma coluna que classifica o ID da categoria, porém nos deparamos com casos como: 99POP com ID 1 e 2, 99Entregas e 99TOP com ID 1, serviços de aluguel de bicicletas com ID 1 entre outros. Do outro lado a tabela "ProviderID" aparenta estar mais estruturada, apesar de algumas inconsistências, com ela deu para compreender que a 99 usa o ID 3, Uber usa o ID 2 e serviços de táxi convencional usam ID 5.

Logo as colunas utilizadas serão as colunas "ProviderID" e "Description" pois se apresentam mais consistentes e estruturadas.

Um adendo importante, é que na próxima etapa será necessário expurgar serviços que não existem mais como o Cabify que encerrou as atividades em 2021.

Tabela Ride:

Tabela Ride:

	RideID	UserID	Schedule	Create	RideStatusID	CompanyID	ProviderID	RideProviderID	price	Updated	CategoryID	TotalUsers	Car	RideDriverLocationID	ScheduledRide
0	1885755	e15b8cc3-5a67-4630-b89f-ee69f302b582	2025-02-10 14:31:10.8858446	2025-02-10 14:31:10.9084221	1.0	2.0	NaN	NaN	0.00	2025-02-10 14:31:10.9084233	NaN	1.0	NaN	NaN	0.0
1	1885754	5c3fb011-8aea-429a-8305-b88953b77df1	2025-02-10 14:26:35.3411403	2025-02-10 14:26:35.4160873	2.0	230.0	5.0	NaN	30.45	2025-02-10 14:28:02.4656963	NaN	1.0	NaN	NaN	0.0
2	1885753	d7e2f4dc-337f-45f5-b762-67b72b077abc	2025-02-10 14:23:45.2540905	2025-02-10 14:24:32.7058722	2.0	52.0	3.0	NaN	11.40	2025-02-10 14:24:46.5037165	NaN	1.0	NaN	NaN	0.0
3	1885752	2125ed9c-89b8-4df6-9be6-53195397a269	2025-02-10 14:23:12.9838635	2025-02-10 14:23:12.9975475	8.0	230.0	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	1885751	72cbebfb-5d70-49ab-ab23-8e3b57c7e399	2025-02-10 14:19:30.5937678	2025-02-10 14:19:30.6117164	2.0	2.0	3.0	NaN	17.28	2025-02-10 14:24:45.9711764	NaN	1.0	NaN	NaN	0.0
5	1885750	2125ed9c-89b8-4df6-9be6-53195397a269	2025-02-10 14:17:31.7575518	2025-02-10 14:18:11.1775140	2.0	230.0	38.0	NaN	43.45	2025-02-10 14:24:33.2966510	NaN	1.0	NaN	NaN	0.0
6	1885749	2125ed9c-89b8-4df6-9be6-53195397a269	2025-02-10 14:16:21.9019781	2025-02-10 14:16:53.7124248	2.0	230.0	5.0	NaN	35.06	2025-02-10 14:17:10.5833443	NaN	1.0	NaN	NaN	0.0
7	1885748	0b43be82-678f-4da3-aa99-2996d327499b	2025-02-10 14:16:07.7452792	2025-02-10 14:16:31.1347796	2.0	266.0	3.0	NaN	10.46	2025-02-10 14:17:10.5716877	NaN	1.0	NaN	NaN	0.0
8	1885747	5fa62073-993a-4aca-9795-c8ec8b68990f	2025-02-10 14:14:28.3109995	2025-02-10 14:14:28.7951597	2.0	265.0	3.0	NaN	9.17	2025-02-10 14:15:23.1372218	NaN	1.0	NaN	NaN	0.0
9	1885746	6d4220c8-1ed0-4d35-afc4-2b8a7c67533d	2025-02-10 14:12:14.5128863	2025-02-10 14:12:15.5246988	2.0	52.0	2.0	NaN	10.30	2025-02-10 14:12:45.1757039	NaN	1.0	NaN	NaN	0.0
10	1885745	5fa62073-993a-4aca-9795-c8ec8b68990f	2025-02-10 14:11:42.1178865	2025-02-10 14:11:42.1360991	2.0	265.0	2.0	NaN	8.46	2025-02-10 14:12:45.3005394	NaN	1.0	NaN	NaN	0.0

Significados dos campos da tabela ride:

RideID: Traz o ID da corrida;

UserID: Apresenta o ID do usuário;

Schedule: Provavelmente se refere a um agendamento de corrida futura, ou uma simples consulta;

Create: Criação da solicitação da corrida;

RideStatusID: O ID do Status da corrida;

CompanyID: O ID da companhia de transporte;

ProviderID: ID do provedor;

RideProviderID: ID do provedor da corrida (aparentou ser redundante);

price: Preço (Não pode ser utilizada);

Updated: Provavelmente indica a data de atualização da cotação;

CategoryID: ID da categoria;

TotalUsers: Total de usuários;

Car: Carro utilizado no serviço;

RideDriverLocationID: ID Localização do motorista da corrida;

ScheduledRide: Agendamento da corrida;

Considerações: Essa tabela pode ser classificada com uma qualidade média/alta, porém não traz tantos dados relevantes, dados como "UserID", "RideStatusID", "CompanyID", "RideProviderID", "price" (proibição do escopo do projeto), "CategoryID", “TotalUsers", "Car", "RideDriverLocationID", "ScheduledRide" não serão utilizados, por conta da ausência de outras fontes para compreensão, falta de utilidade e um alto número de dados nulos.

Portanto a ideia inicial é atuar com os campos: "RideID", "Schedule", "Create", "ProviderID" e "Updated".

Tabela RideAdress

Tabela RideAddress:

	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	-26.329754299999998	-48.840427999999996	1.0	1183200.0
1	2334278	Av. Dr. Nereu Ramos, 450 - Rocio Grande, São F...	Av. Dr. Nereu Ramos, 450 - Rocio Grande, São F...	450	NaN	NaN	NaN	-26.2554657	-48.6434197	2.0	1183200.0
2	2334279	Rodovia Rafael da Rocha Pires, 1883 - Rodovia ...	Rodovia Rafael da Rocha Pires	1883	Rodovia Rafael da Rocha Pires	NaN	Brasil	-27.4919788	-48.528287999999996	1.0	1183201.0
3	2334280	Angeloni Ingleses (Florianópolis) - Supermerca...	Angeloni Ingleses (Florianópolis) - Supermercado	6375	NaN	NaN	NaN	-27.4371486	-48.398243099999999	2.0	1183201.0
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	-19.8495799	-44.019915999999995	1.0	1183202.0
5	2334282	R. Antônio de Albuquerque, 1080 - Funcionários...	R. Antônio de Albuquerque, 1080 - Funcionários...	1080	NaN	Belo Horizonte	Minas Gerais	-19.936899	-43.9401603	2.0	1183202.0
6	2334283	Tv. Duzentos e Sessenta e Um, 72, 72	Tv. Duzentos e Sessenta e Um, 72	72	NaN	NaN	NaN	-23.9624233	-46.25465759999999	1.0	1183203.0
7	2334284	Semar Supermercados Bertioga, 2141	Semar Supermercados Bertioga	2141	NaN	NaN	NaN	-23.8373074	-46.1321725	2.0	1183203.0
8	2334285	Rua Argentina, 160 - Rua Argentina - Brasil	Rua Argentina	160	Rua Argentina	NaN	Brasil	-10.9198019	-37.077441799999995	1.0	1183204.0
9	2334286	R. Simeão Aguiar, 430 - Novo Paraíso, Aracaju ...	R. Simeão Aguiar, 430 - Novo Paraíso, Aracaju ...	430	NaN	Aracaju	Sergipe	-10.9071288	-37.0877194	2.0	1183204.0
10	2334287	Rua João Cirilo de Oliveira, 5 - Rua João Cirí...	Rua João Cirilo de Oliveira	5	Rua João Cirilo de Oliveira	NaN	Brasil	-22.8735015	-43.5714019	1.0	1183205.0

Significados dos campos da tabela Ride Adress:

RideAddressID: ID do endereço da corrida (segue ordem crescente);

Address: Endereço;

Street: Rua/ Avenida (tem atuado de forma redundante com relação ao endereço);

Number: Número do endereço;

Neighborhood: "Vizinho" do endereço (Tem atuado de forma redundante com relação ao endereço);

City: Cidade;

State: Estado;

Lat: Latitude;

Lng: Longitude;

RideAddressTypeID: ID que clássifia origem com o número 1 e destino com número 2;

RideID: Traz o ID da corrida.

Considerações: Essa é a tabela de maior importância, ela traz dados bem completos, porém possui uma ala de campos nulos nas colunas de "City" e "State", os endereços sofrem de erros de padronização, então a decisão mais assertiva é atuar com a latitude e longitude, somado ao "number" pois os dados de Latitude e Longitude estão mais estruturados. Então a ideia inicial é atuar com os campos: "Number", "Lat", "Lng", "RideAddressTypeID" e "RideID".

E consequentemente descartar as colunas: "RideAddressID", "Address", "Street", "Neighborhood", "City" e "State".

Tabela RideEstimative

Tabela RideEstimative:

	RideEstimativeID	RideID	ProductID	WaitingTime	Price	FareID	Selected	RideReasonSelectedEstimativeID	Fee
0	8619946	1183200	Flash	8	89.00	c6aaac64-5f89-4fc4-8b66-0251ec1c78a8	0.0	NaN	0.0
1	8619947	1183200	UberX	6	89.00	ff3cc941-93a8-4d0e-a274-bb988576d7d4	0.0	NaN	0.0
2	8619948	1183200	Comfort	10	116.50	d7708871-2f2c-447d-81e6-a2d121863a2f	0.0	NaN	0.0
3	8619949	1183200	poupa99	5	170.21	NaN	0.0	NaN	0.0
4	8619950	1183200	pop99	7	170.21	NaN	0.0	NaN	0.0
5	8619951	1183200	turbo-taxi	6	151.05	NaN	0.0	NaN	0.0
6	8619952	1183200	regular-taxi	6	151.05	NaN	1.0	4.0	0.0
7	8619953	1183201	Flash	10	31.50	d2c2ad0c-6d23-4807-afb8-884ad65f2918	0.0	NaN	0.0
8	8619954	1183201	Comfort	11	33.50	9925fac9-70a9-4d09-9346-2661fdef50b6	0.0	NaN	0.0
9	8619956	1183201	poupa99	5	26.91	NaN	0.0	NaN	0.0
10	8619957	1183201	pop99	5	29.14	NaN	0.0	NaN	0.0

Significados dos campos da tabela RideEstimative:

RideEstimativeID: ID da estimativa da corrida (em ordem crescente);

RideID: ID da corrida;

ProductID: ID do produto da empresa de transporte (Flash, UberX, Confort);

WaitingTime: Tempo de espera, provavelmente em minutos;

Price: Preço estimado;

FareID: ID da tarifa;

Selected: Não foi possível interpretar;

RideReasonSelectedEstimativeID: ID da razão de escolha da estimativa;

Fee: Não foi possível interpretar.

Considerações: Por último, essa tabela será necessário consultar a viabilidade de alguns campos como por exemplo a utilizar do campo "Price" como base, já que a proibição era para a tabela "ride", portanto, a ideia é utilizar os campos: "RideID", "ProductID" e "Price" (se for permitido).

E no caso as colunas: "RideEstimativeID", "WaitingTime" (pode ser revogado), "FareID", "Selected", "RideReasonSelectedEstimativeID" e "Fee".