DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

O que é um portfólio?

É basicamente um **conjunto dos trabalhos / projetos** feitos por um profissional durante a sua carreira

Em geral, os **mais relevantes** e alinhados
com o **objetivo atual** da pessoa

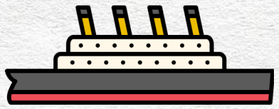
É uma forma de **provar o seu conhecimento** para recrutadores / empresas

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

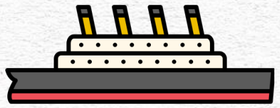
“Lucas, como eu consigo **datasets** para **começar a criar** meu portfolio?”

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

“Lucas, como eu consigo **datasets** para **começar a criar** meu portfolio?”



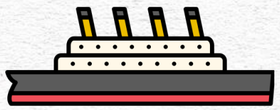
DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO



mais simples ← → mais complexos

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

RH / Time de Seleção

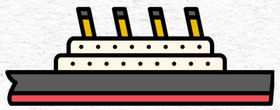


mais simples ← → mais complexos

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

RH / Time de Seleção

Especialistas

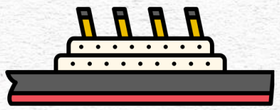


mais simples ← → mais complexos

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

RH / Time de Seleção

Especialistas



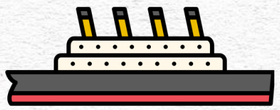
mais simples ← → **mais complexos**

- Mostrar **conhecimento da base** das bibliotecas (ex: importar bases no pandas)
- Explicar **conceitos teóricos** importantes em Ciência de Dados
- Apresentar **detalhes de um método específico** (ex: OneHotEncoder)

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

RH / Time de Seleção

Especialistas



mais simples ← → **mais complexos**

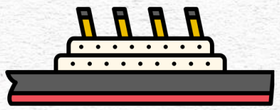
- Mostrar **conhecimento da base** das bibliotecas (ex: importar bases no pandas)
- Explicar **conceitos teóricos** importantes em Ciência de Dados
- Apresentar **detalhes de um método específico** (ex: OneHotEncoder)

- Criar **projetos completos**, de ponta a ponta
- Mais focado nas **conclusões do projeto** do que em explicar os métodos usados
- Utilização de **vários conhecimentos de forma conjunta** e criação de um **storytelling** do que foi feito

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

RH / Time de Seleção

Especialistas



airbnb

mais simples ← → **mais complexos**

- Mostrar **conhecimento da base** das bibliotecas (ex: importar bases no pandas)
- Explicar **conceitos teóricos** importantes em Ciência de Dados
- Apresentar **detalhes de um método específico** (ex: OneHotEncoder)

- Criar **projetos completos**, de ponta a ponta
- Mais focado nas **conclusões do projeto** do que em explicar os métodos usados
- Utilização de **vários conhecimentos de forma conjunta** e criação de um **storytelling** do que foi feito

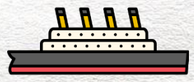
LinkedIn

Medium

GitHub

kaggle

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO



airbnb

Independente de qual base estivermos usando, alguns projetos sempre podem ser feitos:

- Apresentação dos primeiros passos nas bibliotecas do Python
- Análise Exploratória da base
- Visualização e apresentação dos dados
- Se aprofundar em algum método para resolver determinado problema
- Relacionar problemas da sua base com casos reais de empresas
- Utilizar dados da sua base para apresentar conceitos estatísticos
- ...

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

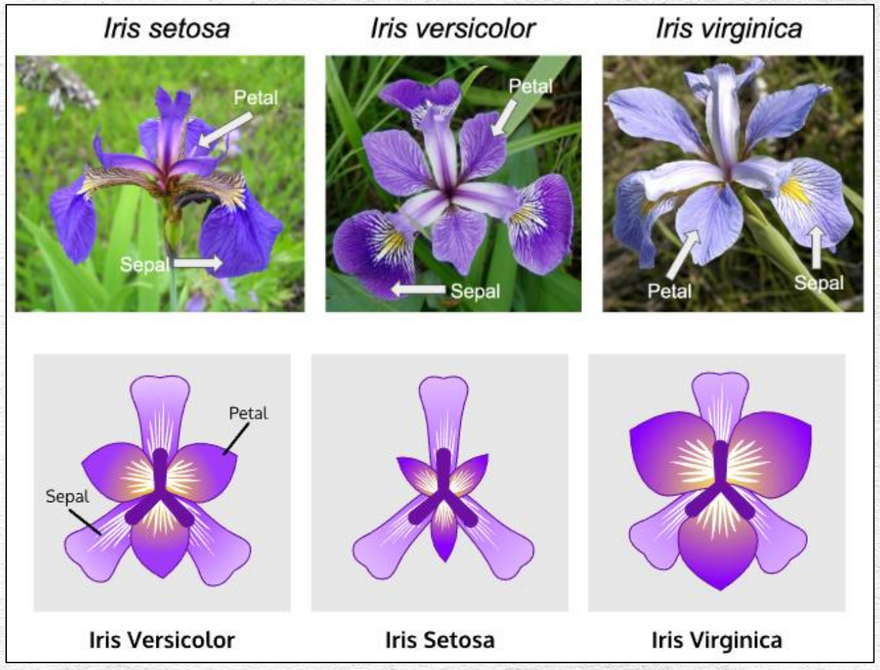
DATASET IRIS, DO SCIKIT-LEARN

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0
5	5.4	3.9	1.7	0.4	0
6	4.6	3.4	1.4	0.3	0
7	5.0	3.4	1.5	0.2	0
8	4.4	2.9	1.4	0.2	0
9	4.9	3.1	1.5	0.1	0

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

DATASET IRIS, DO SCIKIT-LEARN

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0
5	5.4	3.9	1.7	0.4	0
6	4.6	3.4	1.4	0.3	0
7	5.0	3.4	1.5	0.2	0
8	4.4	2.9	1.4	0.2	0
9	4.9	3.1	1.5	0.1	0

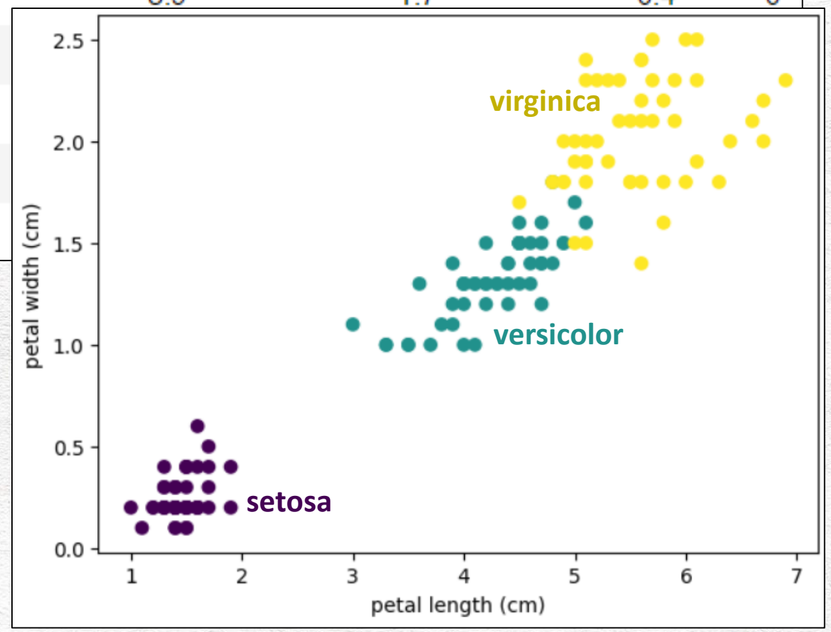


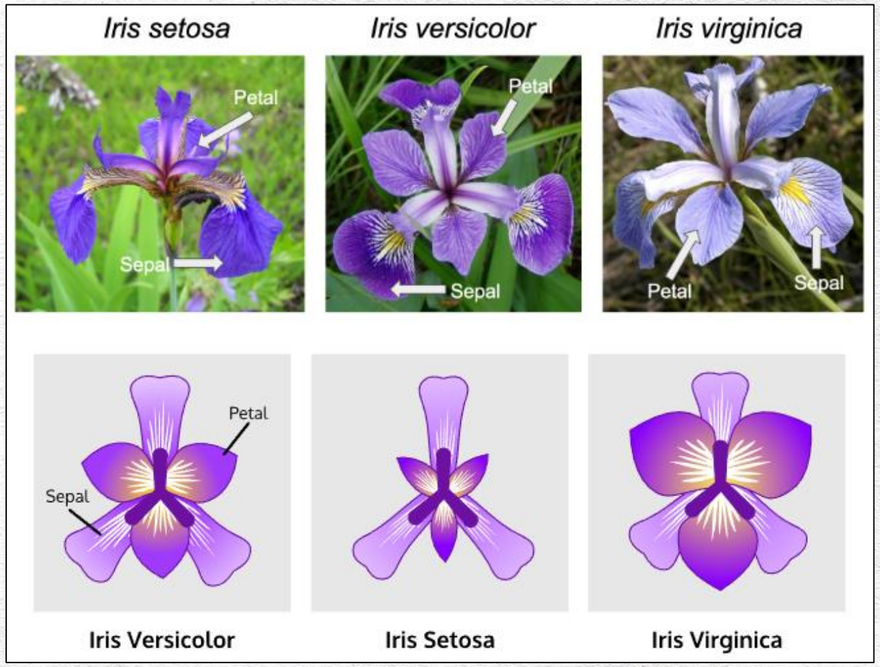
https://www.kaggle.com/code/necibecan/iris-dataset-eda-n/notebook

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

DATASET IRIS, DO SCIKIT-LEARN

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0
5	5.4	3.9	1.7	0.4	0
6	4.6				
7	5.0				
8	4.4				
9	4.9				



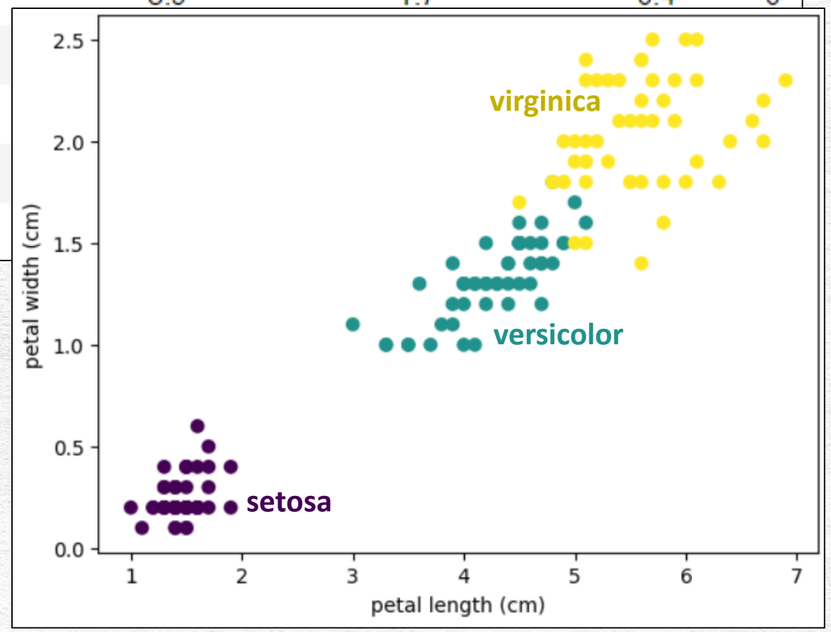


https://www.kaggle.com/code/necibecan/iris-dataset-eda-n/notebook

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

DATASET IRIS, DO SCIKIT-LEARN

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0
5	5.4	3.9	1.7	0.4	0
6	4.6				
7	5.0				
8	4.4				
9	4.9				



- Por ser um **dataset conhecido**, tem **muito conteúdo disponível** na internet (no caso de dúvidas)
- A **separação das classes pode ser feita de forma visual**, então a explicação para alguém leigo se torna muito mais fácil
- É ótimo para quem está começando pela sua **simplicidade**, já que **todas as variáveis são numéricas**
- É um dataset muito inicial, então **não pode ser o único projeto no seu portfólio**
- É possível utilizar bases mais simples para **explicar sobre o Pandas, falar sobre passos básicos e teorias importantes**
- Podemos **comparar diferentes algoritmos** de aprendizado de máquinas (e até **visualizar algoritmos que utilizam distância**)
- Também conseguimos apresentar a **análise de erro na classificação**

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

FETCH_CALIFORNIA_HOUSING, TAMBÉM DO SCIKIT-LEARN

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	target
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23	4.526
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22	3.585
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24	3.521
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25	3.413
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25	3.422
5	4.0368	52.0	4.761658	1.103627	413.0	2.139896	37.85	-122.25	2.697
6	3.6591	52.0	4.931907	0.951362	1094.0	2.128405	37.84	-122.25	2.992
7	3.1200	52.0	4.797527	1.061824	1157.0	1.788253	37.84	-122.25	2.414
8	2.0804	42.0	4.294118	1.117647	1206.0	2.026891	37.84	-122.26	2.267
9	3.6912	52.0	4.970588	0.990196	1551.0	2.172269	37.84	-122.25	2.611

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

FETCH_CALIFORNIA_HOUSING, TAMBÉM DO SCIKIT-LEARN

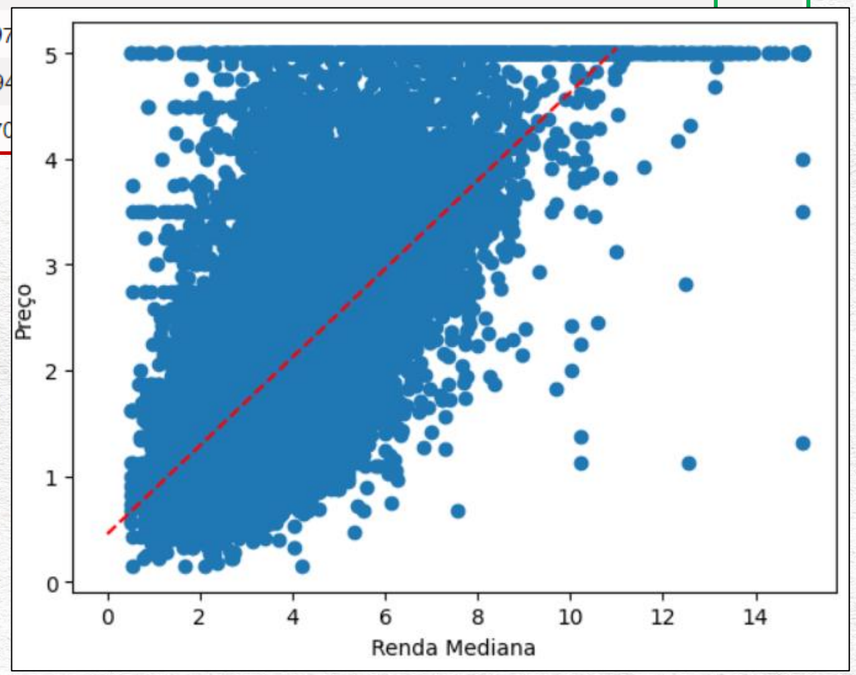
	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	target
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23	4.526
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22	3.585
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24	3.521
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25	3.413
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25	3.422
5	4.0368	52.0	4.761658	1.103627	413.0	2.139896	37.85	-122.25	2.697
6	3.6591	52.0	4.931907	0.951362	1094.0	2.128405	37.84	-122.25	2.992
7	3.1200	52.0	4.797527	1.061824	1157.0	1.788253	37.84	-122.25	2.414
8	2.0804	42.0	4.294118	1.117647	1206.0	2.026891	37.84	-122.26	2.267
9	3.6912	52.0	4.970588	0.990196	1551.0	2.172269	37.84	-122.25	2.611

Temos várias informações sobre as casas da California e queremos **prever qual é o valor de cada uma delas**

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

FETCH_CALIFORNIA_HOUSING, TAMBÉM DO SCIKIT-LEARN

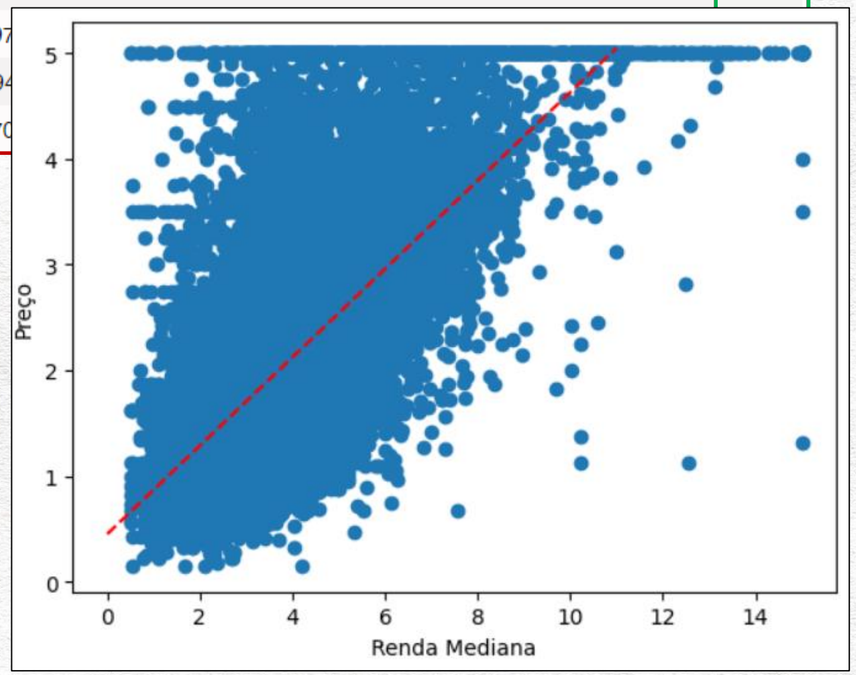
	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	target
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23	4.526
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22	3.585
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24	3.521
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25	3.413
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25	3.422
5	4.0368	52.0	4.761658	1.103627	413.0	2.139896	37.85	-122.25	2.697
6	3.6591	52.0	4.931907	0.951362	1094.0	2.128405	37.84	-122.25	2.992
7	3.1200	52.0	4.797						
8	2.0804	42.0	4.294						
9	3.6912	52.0	4.970						



DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

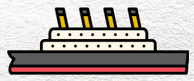
FETCH_CALIFORNIA_HOUSING, TAMBÉM DO SCIKIT-LEARN

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	target
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23	4.526
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22	3.585
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24	3.521
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25	3.413
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25	3.422
5	4.0368	52.0	4.761658	1.103627	413.0	2.139896	37.85	-122.25	2.697
6	3.6591	52.0	4.931907	0.951362	1094.0	2.128405	37.84	-122.25	2.992
7	3.1200	52.0	4.797						
8	2.0804	42.0	4.294						
9	3.6912	52.0	4.970						



- Assim como o dataset iris, também é bastante **conhecido** (bastante conteúdo disponível) e **simples** (variáveis numéricas)
- Também é um dataset muito inicial, então **não pode ser o único projeto no seu portfólio**
- Para esse dataset, podemos **comparar diferentes algoritmos de regressão** (target é numérico)
- Também podemos fazer a **análise do erro**, só que agora para a **regressão**
- Algumas informações estão em **diferentes escalas**
- Como os dados estão em escalas diferentes, podemos falar de **padronização e normalização dos dados**

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO



TITANIC - MACHINE LEARNING FROM DISASTER

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

TITANIC - MACHINE LEARNING FROM DISASTER

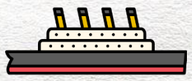
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C

O próprio Kaggle sugere esse dataset para quem está começando e também incentiva que você **escreva o seu código e o submeta para uma avaliação**:

3. **Faça um envio**
   Carregue sua previsão como um envio no Kaggle e receba uma pontuação de precisão.
4. **Verifique a tabela de classificação**
   Veja como seu modelo se classifica em relação a outros Kagglers em nossa tabela de classificação.
5. **Melhore sua pontuação**
   Confira o fórum de discussão para encontrar muitos tutoriais e insights de outros concorrentes.

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO





TITANIC - MACHINE LEARNING FROM DISASTER

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C

- Você pode utilizar essa dataset para ser o seu **primeiro projeto feito no Kaggle** e receber um **feedback do seu resultado**
- Existe **muito conteúdo disponível sobre o titanic na internet**, então é possível tirar qualquer dúvida que exista
- Muitas pessoas já fizeram esse desafio do Kaggle, então existe **muita referência para você pesquisar**
- Existem **colunas com valores de texto** que não podem ser usados no modelo e **dados faltantes**

- Podemos falar sobre **cardinalidade dos dados** (será que o nome do passageiro ajuda na previsão?)
- Como existem colunas de texto, é possível abordar diferentes técnicas de **transformação (encoding) de variáveis categóricas (de texto)**
- Como a base possui **valores nulos,** também podemos falar sobre **diferentes formas de fazer esse tratamento**

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

OCORRÊNCIAS AERONÁUTICAS NA AVIAÇÃO CIVIL BRASILEIRA

	codigo_ocorrencia	codigo_ocorrencia1	codigo_ocorrencia2	codigo_ocorrencia3	codigo_ocorrencia4	ocorrencia_classificacao	ocorrencia_latitude	ocorrencia_
0	81027	81027	81027	81027	81027	INCIDENTE GRAVE	NaN	
1	81030	81030	81030	81030	81030	INCIDENTE	NaN	
2	81023	81023	81023	81023	81023	INCIDENTE GRAVE	NaN	
3	81029	81029	81029	81029	81029	INCIDENTE	NaN	
4	81025	81025	81025	81025	81025	INCIDENTE	NaN	
5	81019	81019	81019	81019	81019	ACIDENTE	NaN	
6	81022	81022	81022	81022	81022	INCIDENTE	NaN	
7	81021	81021	81021	81021	81021	ACIDENTE	NaN	
8	81009	81009	81009	81009	81009	INCIDENTE GRAVE	NaN	
9	81012	81012	81012	81012	81012	INCIDENTE	NaN	

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

OCORRÊNCIAS AERONÁUTICAS NA AVIAÇÃO CIVIL BRASILEIRA

	codigo_ocorrencia	codigo_ocorrencia1	codigo_ocorrencia2	codigo_ocorrencia3	codigo_ocorrencia4	ocorrencia_classificacao	ocorrencia_latitude	ocorrencia_
0	81027	81027	81027	81027	81027	INCIDENTE GRAVE	NaN	
1	81030	81030	81030	81030	81030	INCIDENTE	NaN	
2	81023	81023	81023	81023	81023	INCIDENTE GRAVE	NaN	
3	81029	81029	81029	81029	81029	INCIDENTE	NaN	
4	81025	81025	81025	81025	81025	INCIDENTE	NaN	
5	81019	81019	81019	81019	81019	ACIDENTE	NaN	
6	81022	81022	81022	81022	81022	INCIDENTE	NaN	
7	81021	81021	81021	81021	81021	ACIDENTE	NaN	
8	81009	81009	81009	81009	81009	INCIDENTE GRAVE	NaN	
9	81012	81012	81012	81012	81012	INCIDENTE	NaN	

	codigo_ocorrencia1	ocorrencia_tipo	ocorrencia_tipo_categoria	taxonomia_tipo_icao
0	81030	DESCOMPRESSÃO NÃO INTENCIONAL / EXPLOSIVA	FALHA OU MAU FUNCIONAMENTO DE SISTEMA / COMPON...	SCF-NP
1	81029	ESTOURO DE PNEU	FALHA OU MAU FUNCIONAMENTO DE SISTEMA / COMPON...	SCF-NP
2	81027	ESTOURO DE PNEU	FALHA OU MAU FUNCIONAMENTO DE SISTEMA / COMPON...	SCF-NP
3	81027	EXCURSÃO DE PISTA	EXCURSÃO DE PISTA	RE
4	81026	FALHA OU MAU FUNCIONAMENTO DE SISTEMA / COMPON...	FALHA OU MAU FUNCIONAMENTO DE SISTEMA / COMPON...	SCF-NP

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

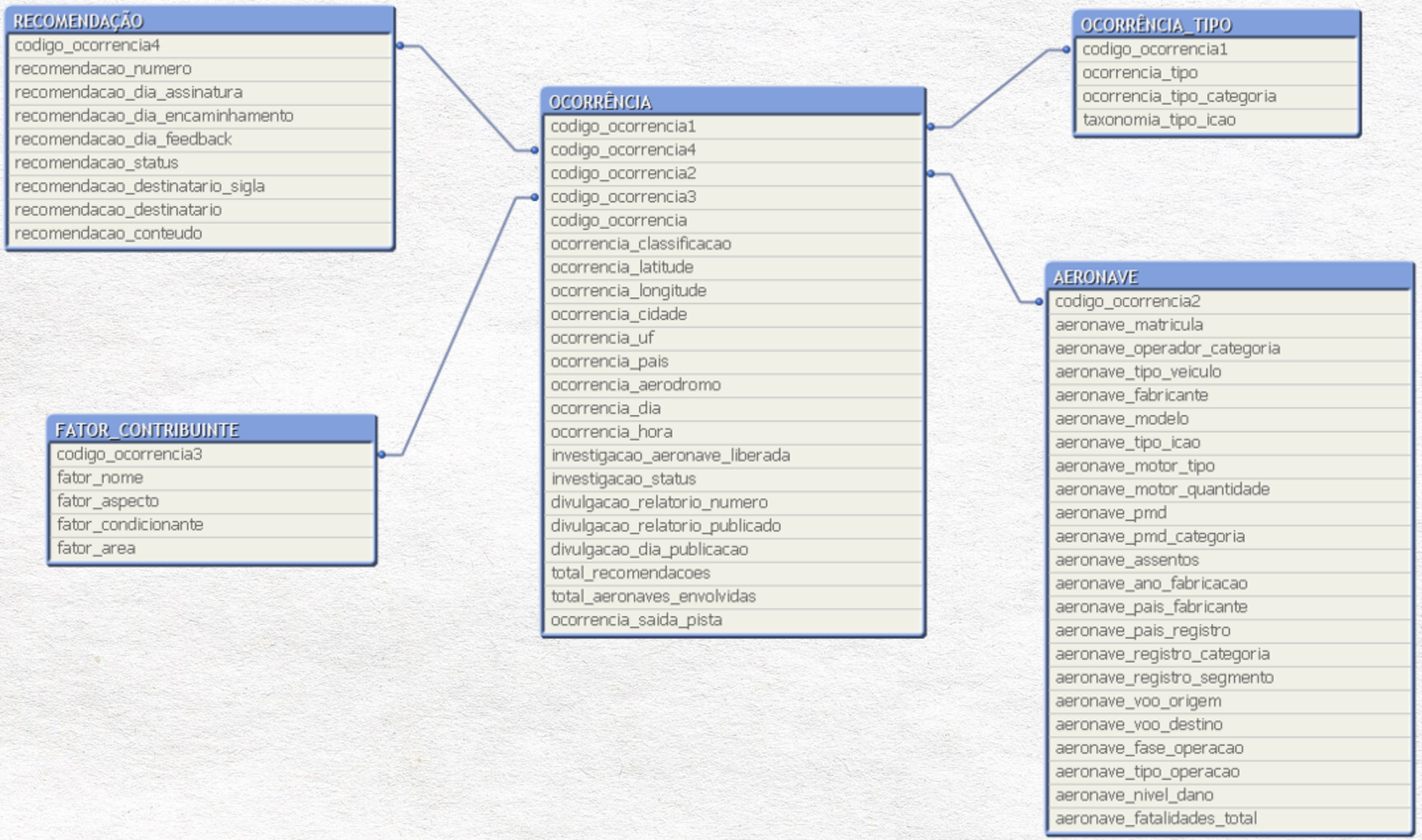
OCORRÊNCIAS AERONÁUTICAS NA AVIAÇÃO CIVIL BRASILEIRA

	codigo_ocorrencia	codigo_ocorrencia1	codigo_ocorrencia2	codigo_ocorrencia3	codigo_ocorrencia4	ocorrencia_classificacao	ocorrencia_latitude	ocorrencia_
0	81027	81027	81027	81027	81027	INCIDENTE GRAVE	NaN	
1	81030	81030	81030	81030	81030	INCIDENTE	NaN	
2	81023	81023	81023	81023	81023	INCIDENTE GRAVE	NaN	
3	81029	81029	81029	81029	81029	INCIDENTE	NaN	
4	81025	81025	81025	81025	81025	INCIDENTE	NaN	
5	81019	81019	81019	81019	81019	ACIDENTE	NaN	
6	81022	81022	81022	81022	81022	INCIDENTE	NaN	
7	81021	81021	81021	81021	81021	ACIDENTE	NaN	
8	81009	81009	81009	81009	81009	INCIDENTE GRAVE	NaN	
9	81012	81012	81012	81012	81012	INCIDENTE	NaN	

	codigo_ocorrencia2	aeronave_matricula	aeronave_operador_categoria	aeronave_tipo_veiculo	aeronave_fabricante	aeronave_modelo	aeronave_tipo_icao	aerc
0	43628	PTEHG	***	AVIÃO	EMBRAER	EMB-820C NAVAJO	PA31	
1	43629	PTHVW	***	HELICÓPTERO	ROBINSON HELICOPTER	R22 BETA	R22	
2	43630	PTXRK	ESPECIALIZADA	AVIÃO	AIR TRACTOR	AT-401B	AT3P	
3	43631	PRGGM	REGULAR	AVIÃO	BOEING COMPANY	737-8EH	B738	
4	43633	PRPSK	***	AVIÃO	EMBRAER	EMB-145LR	E145	

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

OCORRÊNCIAS AERONÁUTICAS NA AVIAÇÃO CIVIL BRASILEIRA



DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

OCORRÊNCIAS AERONÁUTICAS NA AVIAÇÃO CIVIL BRASILEIRA

	codigo_ocorrencia	codigo_ocorrencia1	codigo_ocorrencia2	codigo_ocorrencia3	codigo_ocorrencia4	ocorrencia_classificacao	ocorrencia_latitude	ocorrencia_
0	81027	81027	81027	81027	81027	INCIDENTE GRAVE	NaN	
1	81030	81030	81030	81030	81030	INCIDENTE	NaN	
2						CIDENTE GRAVE	NaN	
3						INCIDENTE	NaN	
4						INCIDENTE	NaN	
5						ACIDENTE	NaN	
6						INCIDENTE	NaN	
7						ACIDENTE	NaN	
8						CIDENTE GRAVE	NaN	
9						INCIDENTE	NaN	

```python
# Importando e criando a conexão
import sqlite3
con = sqlite3.connect("ocorrencias.db")
```

```python
# Enviando a tabela ocorrencias
ocorrencias.to_sql('ocorrencias',con,if_exists='replace',index=False)
```

```
6769
```

```python
# Enviando a tabela tipo
tipo.to_sql('tipo',con,if_exists='replace',index=False)
```

```
7100
```

```python
# Enviando a tabela aeronave
aeronave.to_sql('aeronave',con,if_exists='replace',index=False)
```

```
6339
```

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

OCORRÊNCIAS AERONÁUTICAS NA AVIAÇÃO CIVIL BRASILEIRA

	codigo_ocorrencia	codigo_ocorrencia1	codigo_ocorrencia2	codigo_ocorrencia3	codigo_ocorrencia4	ocorrencia_classificacao	ocorrencia_latitude	ocorrencia_
0	81027	81027	81027	81027	81027	INCIDENTE GRAVE	NaN	
1	81030	81030	81030	81030	81030	INCIDENTE	NaN	
2						CIDENTE GRAVE	NaN	
3						INCIDENTE	NaN	
4						INCIDENTE	NaN	
5						ACIDENTE	NaN	
6						INCIDENTE	NaN	
7						ACIDENTE	NaN	
8						CIDENTE GRAVE	NaN	
9						INCIDENTE	NaN	

```python
# Importando e criando a conexão
import sqlite3
con = sqlite3.connect("ocorrencias.db")
```

```python
# Enviando a tabela ocorrencias
ocorrencias.to_sql('ocorrencias',con,if_exists='replace',index=False)
```

```
6769
```

```python
# Enviando a tabela tipo
tipo.to_sql('tipo',con,if_exists='replace',index=False)
```

```
7100
```

```python
# Enviando a tabela aeronave
aeronave.to_sql('aeronave',con,if_exists='replace',index=False)
```

```
6339
```

```python
# Verificando a tabela ocorrencias
sql = "SELECT * \
        FROM ocorrencias o \
        LEFT JOIN tipo t \
            ON o.codigo_ocorrencia1 = t.codigo_ocorrencia1 \
        LEFT JOIN aeronave a \
            ON o.codigo_ocorrencia2 = a.codigo_ocorrencia2"
resumo = executa_consulta(sql)
```

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

OCORRÊNCIAS AERONÁUTICAS NA AVIAÇÃO CIVIL BRASILEIRA

	codigo_ocorrencia	codigo_ocorrencia1	codigo_ocorrencia2	codigo_ocorrencia3	codigo_ocorrencia4	ocorrencia_classificacao	ocorrencia_latitude	ocorrencia_
0	81027	81027	81027	81027	81027	INCIDENTE GRAVE	NaN	
1	81030	81030	81030	81030	81030	INCIDENTE	NaN	
2								
3								
4								
5								
6								
7								
8								
9								

```
# Importando e criando a conexão
import sqlite3
con = sqlite3.connect("ocorrencias.db")
```

```
# Enviando a tabela ocorrencias
ocorrencias.to_sql('ocorrencias',con,if_exists='replace',
```

```
6769
```

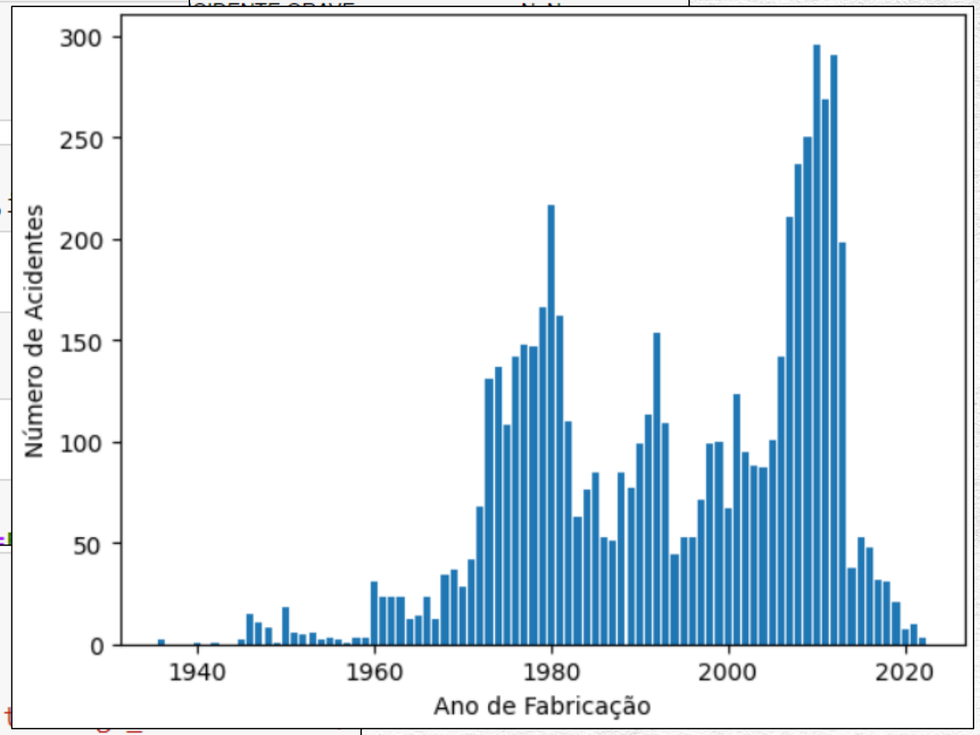
```
# Enviando a tabela tipo
tipo.to_sql('tipo',con,if_exists='replace',index=False)
```

```
7100
```

```
# Enviando a tabela aeronave
aeronave.to_sql('aeronave' con if exists='replace' index=
```

```
6339
```

```
# Verificando a tabela ocorrencias
sql = "SELECT * \
        FROM ocorrencias o \
        LEFT JOIN tipo t \
            ON o.codigo_ocorrencia1 = t
        LEFT JOIN aeronave a \
            ON o.codigo_ocorrencia2 = a.codigo_ocorrencia2"
resumo = executa_consulta(sql)
```



DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

OCORRÊNCIAS AERONÁUTICAS NA AVIAÇÃO CIVIL BRASILEIRA

	codigo_ocorrencia	codigo_ocorrencia1	codigo_ocorrencia2	codigo_ocorrencia3	codigo_ocorrencia4	ocorrencia_classificacao	ocorrencia_latitude	ocorrencia_
0	81027	81027	81027	81027	81027	INCIDENTE GRAVE	NaN	
1	81030	81030	81030	81030	81030	INCIDENTE	NaN	
2								
3								
4								
5								
6								
7								
8								
9								

```python
# Importando e criando a conexão
import sqlite3
con = sqlite3.connect("ocorrencias.db")
```

```python
# Enviando a tabela ocorrencias
ocorrencias.to_sql('ocorrencias',con,if_exists='replace',
```

```
6769
```

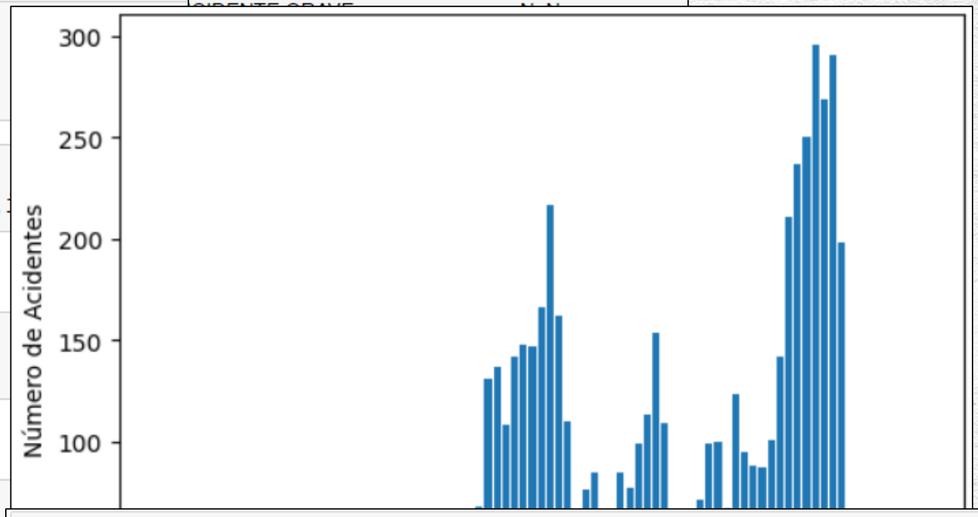
```python
# Enviando a tabela tipo
tipo.to_sql('tipo',con,if_exists='replace',index=False)
```

```
7100
```

```python
# Enviando a tabela aeronave
aeronave.to_sql('aeronave', con, if_exists='replace', index=
```

```
6339
```

```python
# Verificando a tabela ocorrencias
sql = "SELECT * \
        FROM ocorrencias o \
        LEFT JOIN tipo t \
            ON o.codigo_ocorrencia1 =
        LEFT JOIN aeronave a \
            ON o.codigo_ocorrencia2 =
resumo = executa_consulta(sql)
```




```python
# Verificando valores duplicados
resumo['codigo_ocorrencia'].value_counts().head(4)
```

```
80959    4
80862    4
80837    4
80176    4
Name: codigo_ocorrencia, dtype: int64
```

```python
# Verificando no tipo
tipo[tipo.codigo_ocorrencia1 == 80959]
```

	codigo_ocorrencia1	ocorrencia_tipo	ocorrencia_tipo_categoria	taxonomia_tipo_icao
65	80959	GERENCIAMENTO DE TRÁFEGO AÉREO (ATM) / SERVIÇO...	GERENCIAMENTO DE TRÁFEGO AÉREO (ATM) / SERVIÇO...	ATM/CNS
66	80959	INCURSÃO EM PISTA	INCURSÃO EM PISTA	RI

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

OCORRÊNCIAS AERONÁUTICAS NA AVIAÇÃO CIVIL BRASILEIRA

	codigo_ocorrencia	codigo_ocorrencia1	codigo_ocorrencia2	codigo_ocorrencia3	codigo_ocorrencia4	ocorrencia_classificacao	ocorrencia_latitude	ocorrencia_
0	81027	81027	81027	81027	81027	INCIDENTE GRAVE	NaN	
1	81030	81030	81030	81030	81030	INCIDENTE	NaN	
2	81023	81023	81023	81023	81023	INCIDENTE GRAVE	NaN	
3	81029	81029	81029	81029	81029	INCIDENTE	NaN	
4	81025	81025	81025	81025	81025	INCIDENTE	NaN	
5	81019	81019	81019	81019	81019	ACIDENTE	NaN	
6	81022	81022	81022	81022	81022	INCIDENTE	NaN	
7	81021	81021	81021	81021	81021	ACIDENTE	NaN	
8	81009	81009	81009	81009	81009	INCIDENTE GRAVE	NaN	
9	81012	81012	81012	81012	81012	INCIDENTE	NaN	

- Como temos vários arquivos diferentes, podemos **aproveitar para criar uma arquitetura básica de um banco de dados**
- Tendo os arquivos em um banco de dados, é possível **apresentar todo o nosso conhecimento em SQL** (básico e avançado)
- Também podemos **relacionar** desde **informações** básicas do acidente até coisas específicas e fazer **análises temporais**
- Existem erros na base que podemos usar para **criar processos de tratamento** muito similares a **bases de empresas reais**
- Utilizar **análise de Pareto** para busca de principais ofensores, principais ocorrências, etc

- As informações dessa base não são senso comum e, por isso, podem acabar **despertando a curiosidade de recrutadores**
- Os dados vão **precisar de vários tratamentos**, o que é **muito comum em projetos de empresas reais**
- A **base possui vários arquivos**, que podem ser trabalhados utilizando métodos como o merge ou até o próprio SQL

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

OCORRÊNCIAS AERONÁUTICAS NA AVIAÇÃO CIVIL BRASILEIRA

	codigo_ocorrencia	codigo_ocorrencia1	codigo_ocorrencia2	codigo_ocorrencia3	codigo_ocorrencia4	ocorrencia_classificacao	ocorrencia_latitude	ocorrencia_
0	81027	81027	81027	81027	81027	INCIDENTE GRAVE	NaN	
1	81030	81030	81030	81030	81030	INCIDENTE	NaN	
2	81023	81023	81023	81023	81023	INCIDENTE GRAVE	NaN	
3	81029	81029	81029	81029	81029	INCIDENTE	NaN	
4	81025	81025	81025	81025	81025	INCIDENTE	NaN	
5	81019	81019	81019	81019	81019	ACIDENTE	NaN	
6	81022	81022	81022	81022	81022	INCIDENTE	NaN	
7	81021	81021	81021	81021	81021	ACIDENTE	NaN	
8	81009	81009	81009	81009	81009	INCIDENTE GRAVE	NaN	
9	81012	81012	81012	81012	81012	INCIDENTE	NaN	

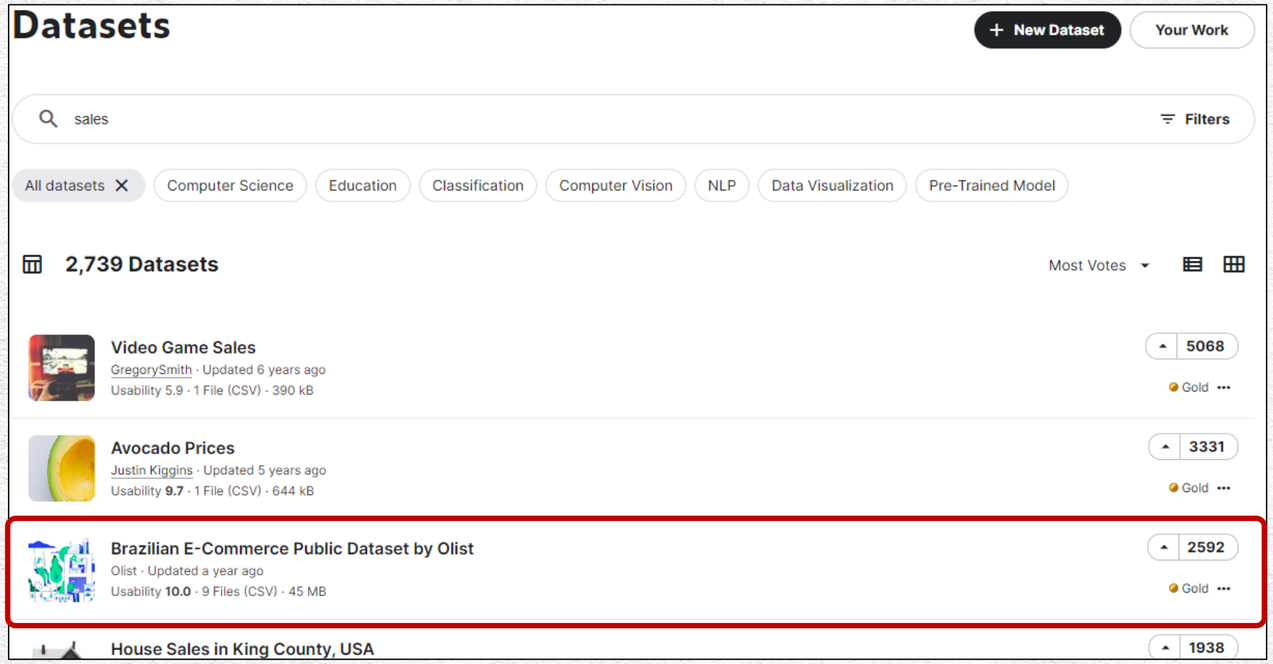
dados.gov.br

data.gov

open.canada.ca

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

CONJUNTO DE DADOS PÚBLICOS DE COMÉRCIO ELETRÔNICO BRASILEIRO



Escolha datasets de assuntos que você gosta

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

CONJUNTO DE DADOS PÚBLICOS DE COMÉRCIO ELETRÔNICO BRASILEIRO

Eu gosto de vendas, então busquei por “Sales” nos datasets do Kaggle (buscar em inglês é mais fácil para achar as bases)

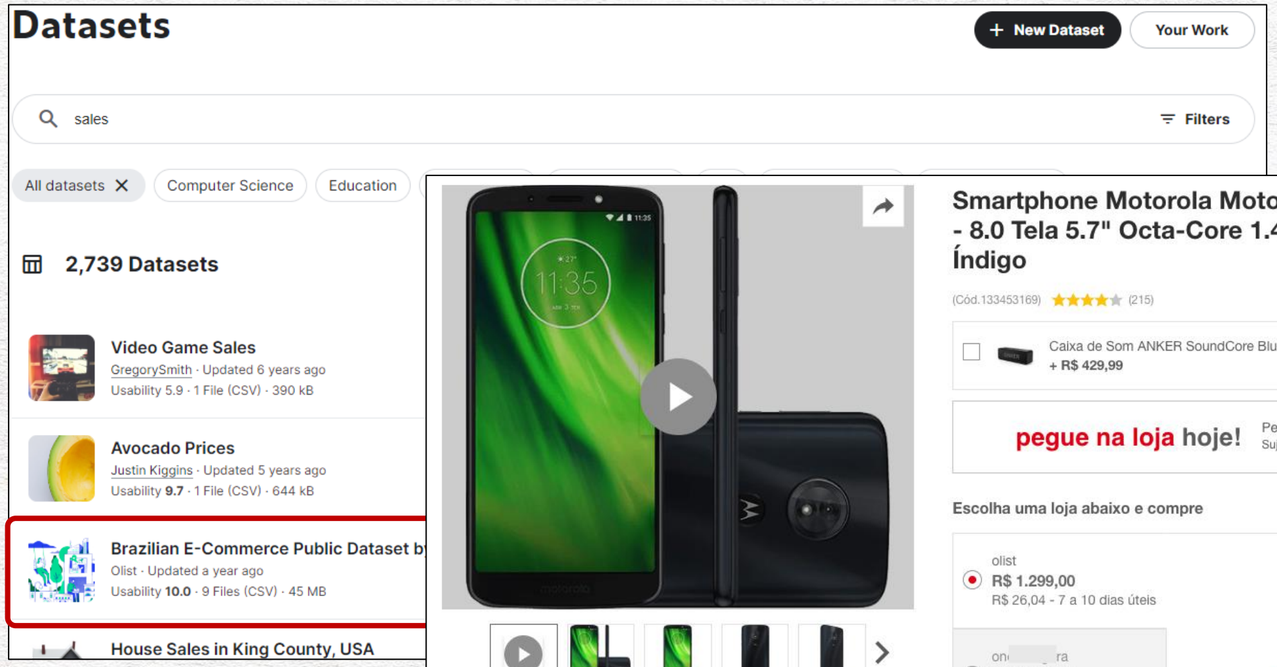
Filtrando pelos mais votados

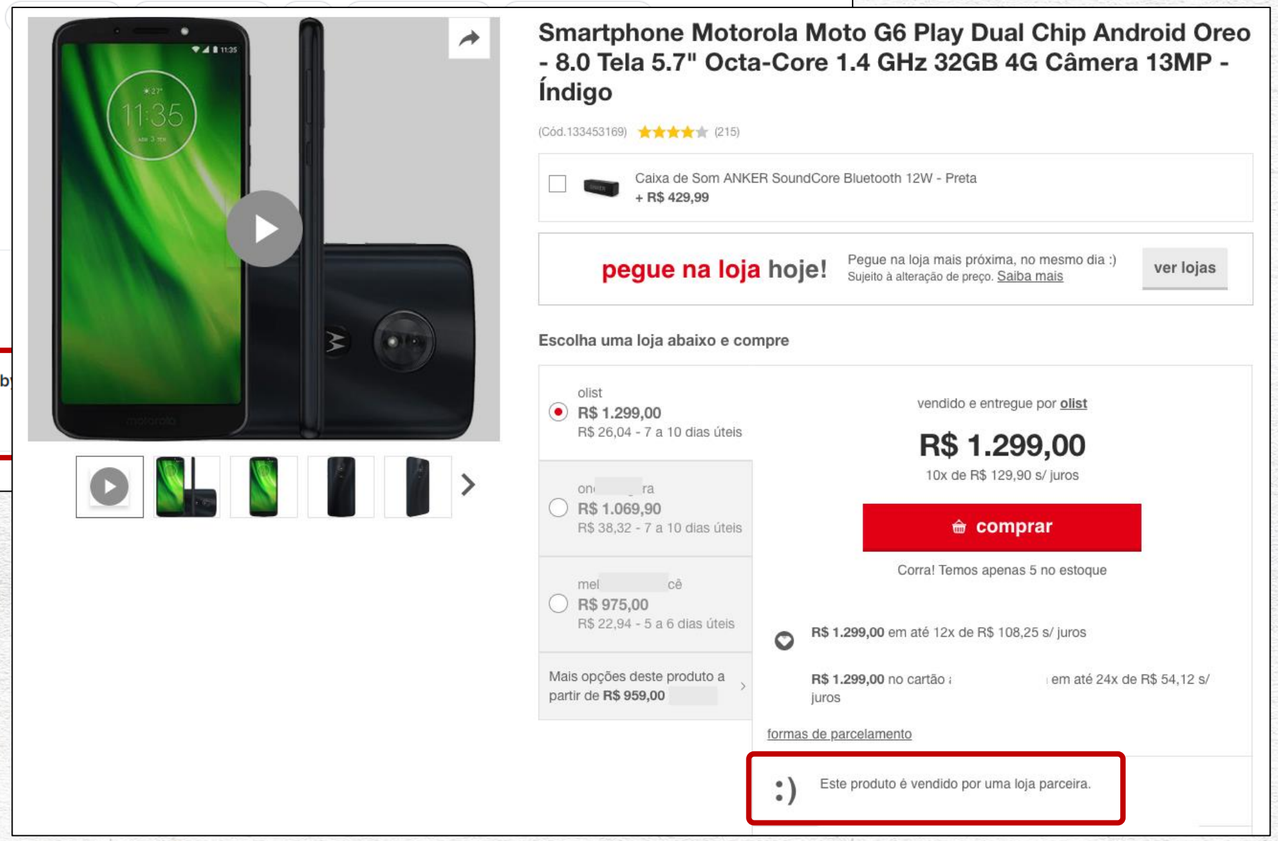
Como era do comércio eletrônico brasileiro, eu achei mais legal e fui me aprofundar nessa base

Escolha datasets de assuntos que você gosta ou até das áreas para as quais você está se candidatando!

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

CONJUNTO DE DADOS PÚBLICOS DE COMÉRCIO ELETRÔNICO BRASILEIRO





Escolha datasets de assuntos que você gosta ou até das áreas para as quais você está se candidatando!

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

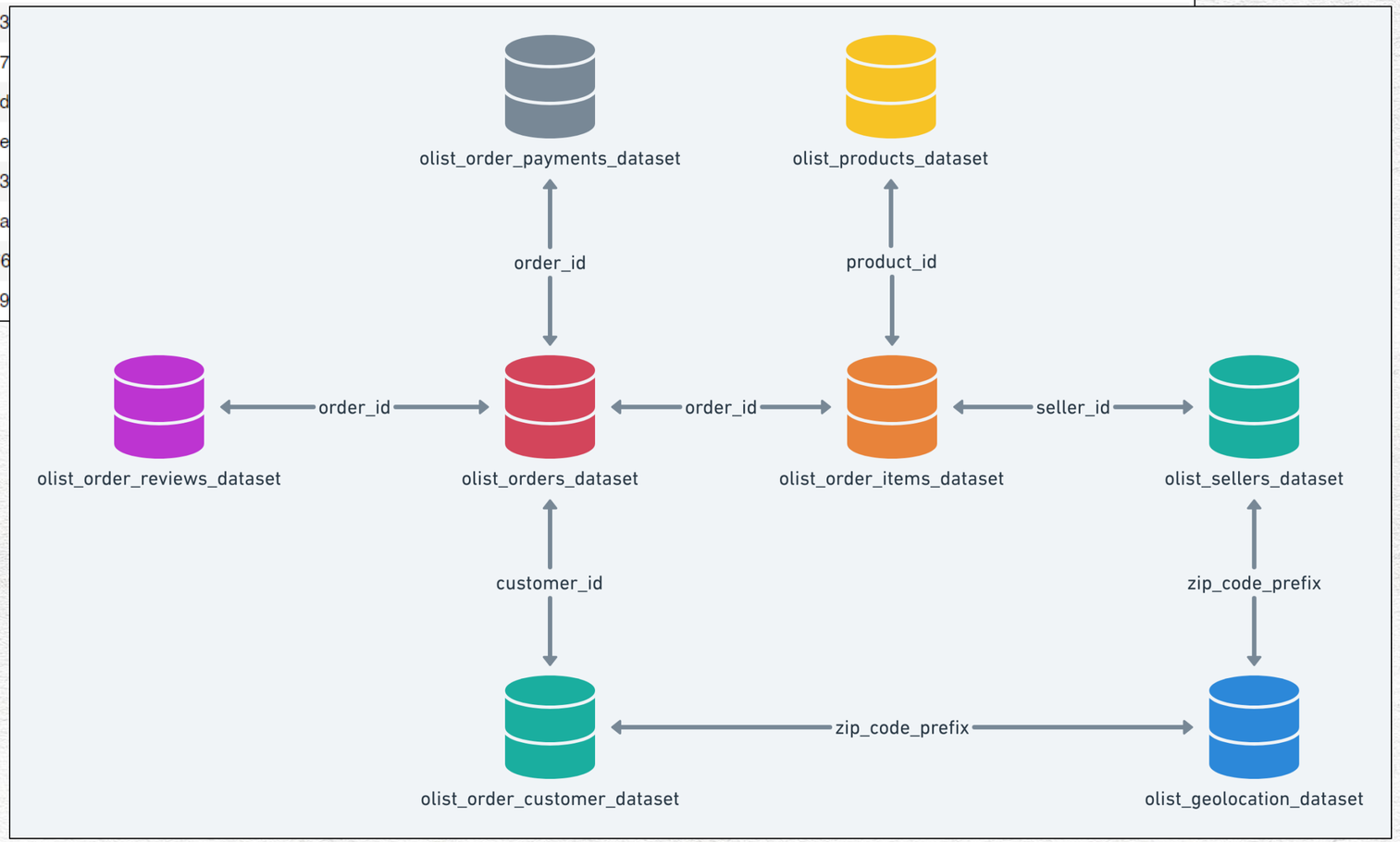
CONJUNTO DE DADOS PÚBLICOS DE COMÉRCIO ELETRÔNICO BRASILEIRO

	order_id	order_item_id	product_id	seller_id	shipping_limit_date	price	freig
0	00010242fe8c5a6d1ba2dd792cb16214	1	4244733e06e7ecb4970a6e2683c13e61	48436dade18ac8b2bce089ec2a041202	2017-09-19 09:45:35	58.90	
1	00018f77f2f0320c557190d7a144bdd3	1	e5f2d52b802189ee658865ca93d83a8f	dd7ddc04e1b6c2c614352b383efe2d36	2017-05-03 11:05:13	239.90	
2	000229ec398224ef6ca0657da4fc703e	1	c777355d18b72b67abbeef9df44fd0fd	5b51032eddd242adc84c38acab88f23d	2018-01-18 14:48:30	199.00	
3	00024acbcdf0a6daa1e931b038114c75	1	7634da152a4610f1595efa32f14722fc	9d7a1d34a5052409006425275ba1c2b4	2018-08-15 10:10:18	12.99	
4	00042b26cf59d7ce69dfabb4e55b4fd9	1	ac6c3623068f30de03045865e4e10089	df560393f3a51e74553ab94004ba5c87	2017-02-13 13:57:51	199.90	
5	00048cc3ae777c65dbb7d2a0634bc1ea	1	ef92defde845ab8450f9d70c526ef70f	6426d21aca402a131fc0a5d0960a3c90	2017-05-23 03:55:27	21.90	
6	00054e8431b9d7675808bcb819fb4a32	1	8d4f2bb7e93e6710a28f34fa83ee7d28	7040e82f899a04d1b434b795a43b4617	2017-12-14 12:10:31	19.90	
7	000576fe39319847cbb9d288c5617fa6	1	557d850972a7d6f792fd18ae1400d9b6	5996cddab893a4652a15592fb58ab8db	2018-07-10 12:30:45	810.00	
8	0005a1a1728c9d785b8e2b08b904576c	1	310ae3c140ff94b03219ad0adc3c778f	a416b6a846a11724393025641d4edd5e	2018-03-26 18:31:29	145.95	
9	0005f50442cb953dcd1d21e1fb923495	1	4535b0e1091c278dfd193e5a1d63b39f	ba143b05f0110f0dc71ad71b4466ce92	2018-07-06 14:10:56	53.99	

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

CONJUNTO DE DADOS PÚBLICOS DE COMÉRCIO ELETRÔNICO BRASILEIRO

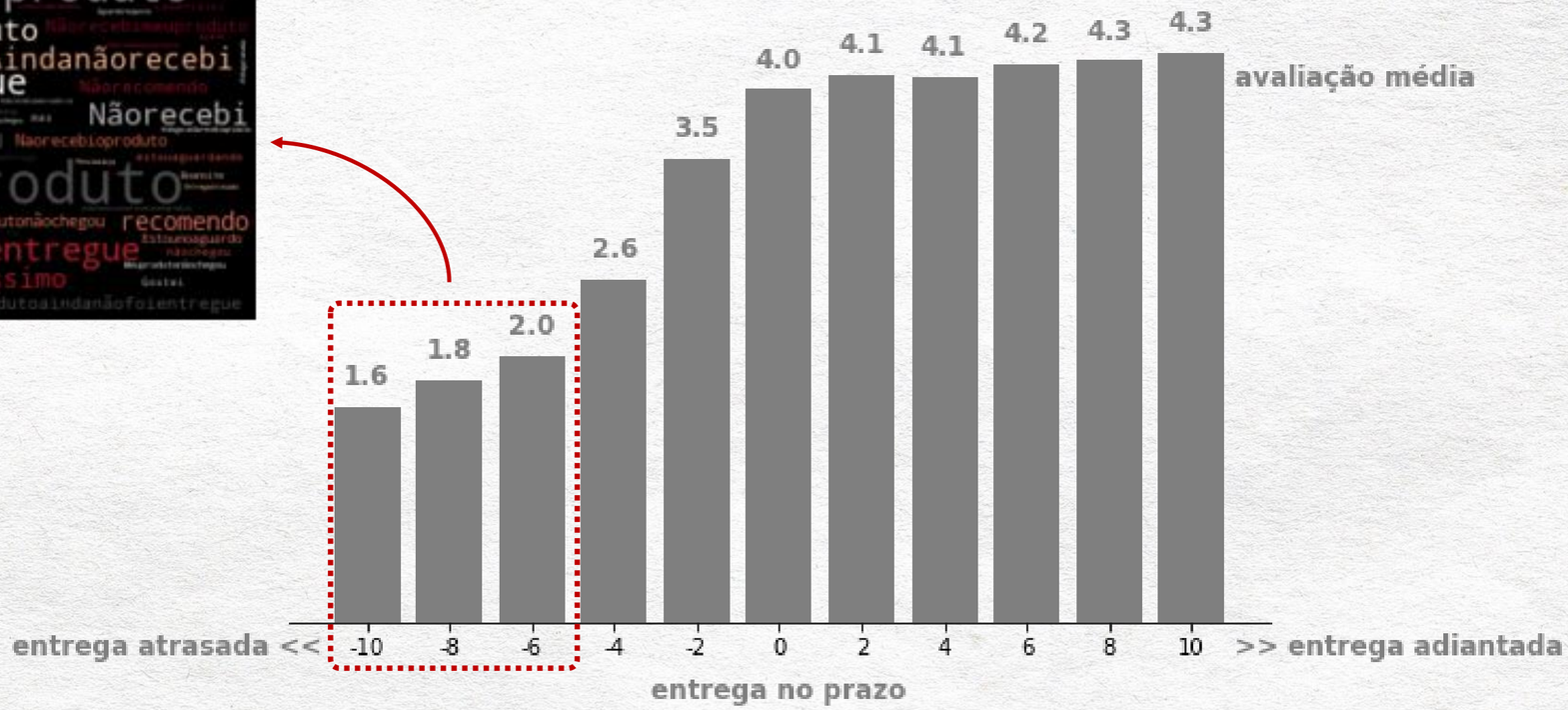
	order_id	order_item_id	product_id	seller_id	shipping_limit_date	price	freig
0	00010242fe8c5a6d1ba2dd792cb16214	1	4244733e06e7ecb4970a6e2683c13e61	48436dade18ac8b2bce089ec2a041202	2017-09-19 09:45:35	58.90	
1	00018f77f2f0320c557190d7a144bdd3	1	e5f2d52b802189ee658865ca93d83a8f	dd7ddc04e1b6c2c614352b383efe2d36	2017-05-03 11:05:13	239.90	
2	000229ec398224ef6ca0657da4fc703						
3	00024acbcdf0a6daa1e931b038114c7						
4	00042b26cf59d7ce69dfabb4e55b4fd						
5	00048cc3ae777c65dbb7d2a0634bc1e						
6	00054e8431b9d7675808bcb819fb4a3						
7	000576fe39319847cbb9d288c5617fa						
8	0005a1a1728c9d785b8e2b08b904576						
9	0005f50442cb953dcd1d21e1fb92349						



DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

CONJUNTO DE DADOS PÚBLICOS DE COMÉRCIO ELETRÔNICO BRASILEIRO

Quanto maior o tempo de atraso, pior a avaliação do cliente.



DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

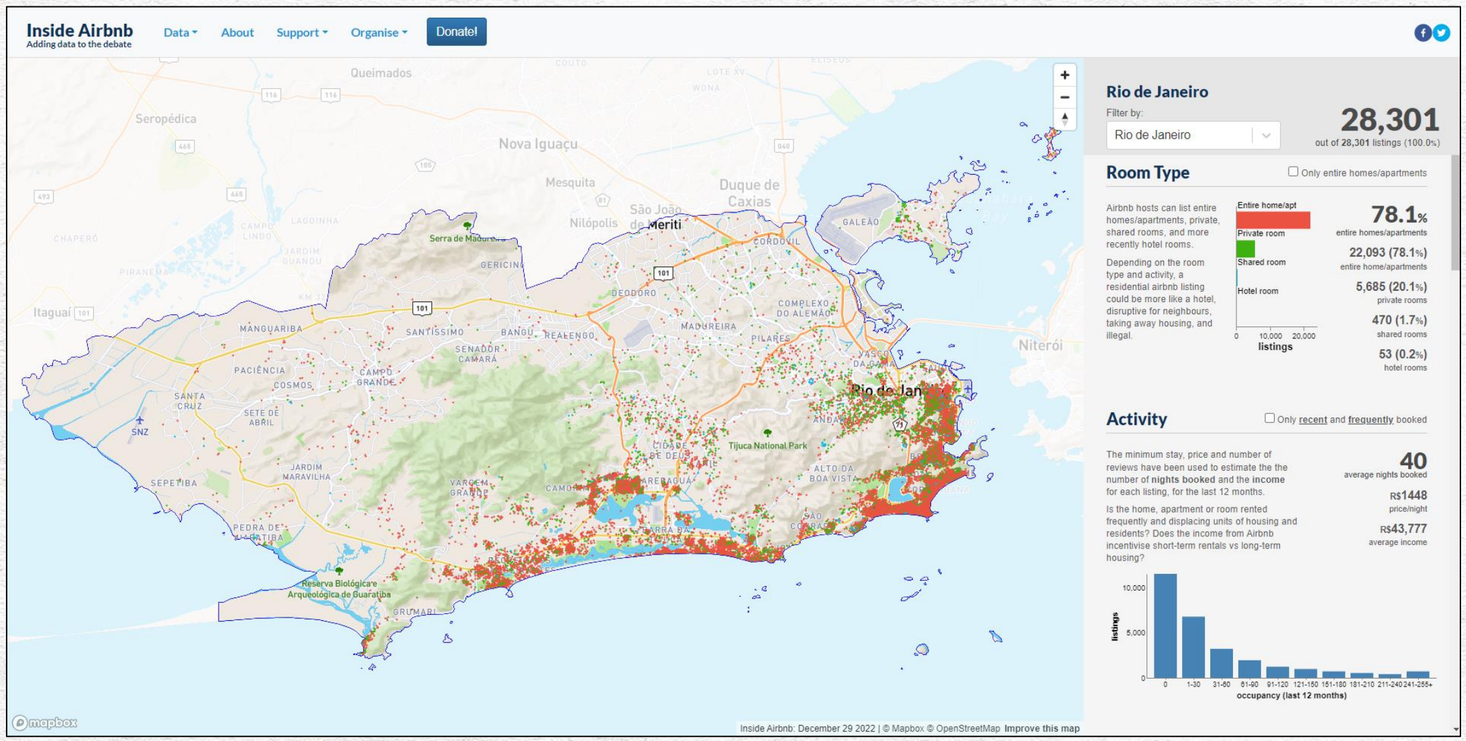
CONJUNTO DE DADOS PÚBLICOS DE COMÉRCIO ELETRÔNICO BRASILEIRO

	order_id	order_item_id	product_id	seller_id	shipping_limit_date	price	freig
0	00010242fe8c5a6d1ba2dd792cb16214	1	4244733e06e7ecb4970a6e2683c13e61	48436dade18ac8b2bce089ec2a041202	2017-09-19 09:45:35	58.90	
1	00018f77f2f0320c557190d7a144bdd3	1	e5f2d52b802189ee658865ca93d83a8f	dd7ddc04e1b6c2c614352b383efe2d36	2017-05-03 11:05:13	239.90	
2	000229ec398224ef6ca0657da4fc703e	1	c777355d18b72b67abbeef9df44fd0fd	5b51032eddd242adc84c38acab88f23d	2018-01-18 14:48:30	199.00	
3	00024acbcdf0a6daa1e931b038114c75	1	7634da152a4610f1595efa32f14722fc	9d7a1d34a5052409006425275ba1c2b4	2018-08-15 10:10:18	12.99	
4	00042b26cf59d7ce69dfabb4e55b4fd9	1	ac6c3623068f30de03045865e4e10089	df560393f3a51e74553ab94004ba5c87	2017-02-13 13:57:51	199.90	
5	00048cc3ae777c65dbb7d2a0634bc1ea	1	ef92defde845ab8450f9d70c526ef70f	6426d21aca402a131fc0a5d0960a3c90	2017-05-23 03:55:27	21.90	
6	00054e8431b9d7675808bcb819fb4a32	1	8d4f2bb7e93e6710a28f34fa83ee7d28	7040e82f899a04d1b434b795a43b4617	2017-12-14 12:10:31	19.90	
7	000576fe39319847cbb9d288c5617fa6	1	557d850972a7d6f792fd18ae1400d9b6	5996cddab893a4652a15592fb58ab8db	2018-07-10 12:30:45	810.00	
8	0005a1a1728c9d785b8e2b08b904576c	1	310ae3c140ff94b03219ad0adc3c778f	a416b6a846a11724393025641d4edd5e	2018-03-26 18:31:29	145.95	
9	0005f50442cb953dcd1d21e1fb923495	1	4535b0e1091c278dfd193e5a1d63b39f	ba143b05f0110f0dc71ad71b4466ce92	2018-07-06 14:10:56	53.99	

- Podemos fazer **tratamento dos dados, criação do banco de dados, uso do SQL** e tudo que falamos anteriormente (e muito mais)
- Em projetos como esse, o mais importante é **trazer boas conclusões relativas ao negócio em si** (e um menor foco em “como” fazer isso)
- É possível buscar nos dados informações e insights que não são óbvios e **apresentar essas conclusões contando uma história envolvente**
- Como essa base possui **comentários em texto dos clientes**, podemos **analisar o que os clientes estão mais reclamando / gostando**
- Outra análise interessante é **relacionar o tempo de atraso com a satisfação do cliente** e até criar um **modelo de previsão de atrasos**

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

AIRBNB: RIO DE JANEIRO



Dados reais disponibilizados pelas empresas em seus próprios sites!

DATASETS PARA COMEÇAR A CRIAR SEU PORTFÓLIO

AIRBNB: RIO DE JANEIRO

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights
0	10463735	Sobrado aconchegante e espaçoso	53918534	Quiá	NaN	Laranjeiras	-22.935550	-43.191070	Entire home/apt	581	1
1	53887789	Quadra da praia	333527901	Lucas	NaN	Copacabana	-22.970320	-43.180810	Entire home/apt	898	5
2	783493769216852616	Leme, Brasil	491704706	Felipe	NaN	Leme	-22.964210	-43.171600	Entire home/apt	720	1
3	703973293620197060	Suíte com entrada independente em casarão 1	20362236	Júlio Cesar	NaN	Botafogo	-22.957920	-43.182226	Private room	599	4
4	782895997622988215	Apartamento próximo ao metrô	302417043	Laís	NaN	Laranjeiras	-22.931960	-43.180180	Entire home/apt	240	3
5	23768085	Vamos a praia	86611015	Mauro	NaN	Barra da Tijuca	-23.011040	-43.320340	Entire home/apt	494	3
6	21568335	Copacabana, perto de tudo	55254246	Ronaldo	NaN	Copacabana	-22.960380	-43.173720	Entire home/apt	657	2
7	784798816581009420	Leblon Luxo Apartamento Inteiro	491910985	Licia	NaN	Leblon	-22.985718	-43.233937	Entire home/apt	3509	4
8	47943201	Diversão, turismo e conforto é em Ipanema	386420902	Wagner	NaN	Ipanema	-22.980910	-43.198140	Entire home/apt	1300	4
9	783267737701368911	apartamento em Lapa!	6355551	Francisco	NaN	Centro	-22.910020	-43.183590	Entire home/apt	232	3

Podemos fazer várias das coisas que já citamos anteriormente, porém agora voltado para o **mercado imobiliário**