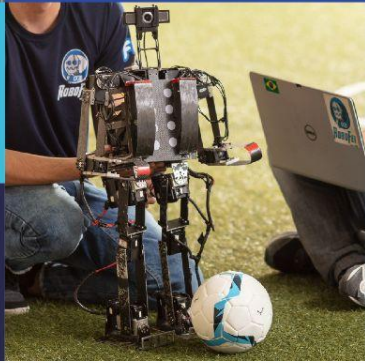




CCP010

Digital Experience



# Análise Exploratória de Dados

Em inglês é conhecido como *EDA - Exploratory Data Analysis*

- É a etapa inicial de um projeto que envolve Dados
- Envolve *analisar* e *visualizar* dados para entender suas principais características
- Descobrir padrões e identificar relacionamentos entre variáveis
- Estudar e explorar conjuntos de dados (*datasets*) para entender suas características, descobrir padrões, localizar outliers e identificar relacionamentos entre variáveis
- A EDA é normalmente realizada como uma etapa preliminar antes de realizar análises mais formais ou modelar o problema com IA

# Análise Exploratória de Dados

Impotância da **EDA - Exploratory Data Analysis**

- **Entender as Estruturas dos Dados:** A EDA ajuda a se familiarizar com o **dataset**, entender o número de atributos e o tipo de dados em cada atributo (**feature**). Esse entendimento é crucial para selecionar técnicas de análise ou previsão apropriadas
- **Identificar Padrões e Relacionamentos:** Por meio de visualizações e resumos estatísticos, a EDA pode revelar padrões ocultos e relacionamentos intrínsecos entre variáveis



# Análise Exploratória de Dados

Importância da **EDA - Exploratory Data Analysis**

- **Detectar Anomalias e Outliers:** A EDA é essencial para identificar erros ou pontos de dados incomuns que podem afetar adversamente os resultados da análise
- **Facilitar a limpeza de dados:** a EDA ajuda a identificar valores ausentes e erros nos dados, que são problemas essenciais de se abordar antes de uma análise mais aprofundada. Primordial para melhorar a qualidade e a integridade dos dados



# Análise Exploratória de Dados

Impotência da *EDA - Exploratory Data Analysis*

- As técnicas de exploração de dados incluem soluções de software de *análise manual* e *exploração automatizada*
- De qualquer forma, a ideia é permitir que os analistas de dados obtenham maior percepção a partir dos dados brutos
- Os dados são frequentemente coletados em grandes volumes não estruturados e de várias possíveis fontes
  - Analistas de dados devem primeiro entender os dados para desenvolver uma visão abrangente antes de extrair dados relevantes para análise posterior



# Análise Exploratória de Dados

## Conjuntos de Dados - *Datasets*

- ***Datasets* - Conjuntos de Dados** - são compostos por uma coleção de dados que compõe a matéria prima dos processos de Análise de Dados e Aprendizado de Máquina, por exemplo



# Análise Exploratória de Dados

## Conjuntos de Dados - *Datasets*

- Os *Datasets* podem conter:
  - **Dados estruturados:** tabelas
  - **Dados não estruturados:** imagens, vídeos, áudios e outros tipos de dados brutos

# Análise Exploratória de Dados

Alguns tipos de Conjuntos de Dados - *Datasets*

- Exemplo de Conjunto de Dados em formato de tabela:

ALUNO

Id	nome	sobrenome	idade
1	João	Pereira	32
2	Carlos	Gonçalves	41
3	Ana	Silva	13



# Análise Exploratória de Dados

## Conjuntos de Dados - *Datasets*

- O mais comum é que os *Datasets* sejam compostos por dados **tabulados (tabelas)**
- A cada elemento do dataset se indicam várias características, atributos (*features*)
- Cada coluna representa uma característica diferente
- Cada linha corresponde a um determinado membro do conjunto de dados em questão
- Cada valor é conhecido como um dado

ALUNO

Id	nome	sobrenome	idade
1	João	Pereira	32
2	Carlos	Gonçalves	41
3	Ana	Silva	13



# Análise Exploratória de Dados

## Conjuntos de Dados - *Datasets*

- Os *Datasets* podem conter dados:
  - **Qualitativos / Categóricos** ou
  - **Quantitativos / Numéricos**

# Análise Exploratória de Dados

## Conjuntos de Dados - *Datasets*

- **Qualitativos ou Categóricos:** Usadas para categorizar/dividir uma variável em grupos específicos
  - **Nominais:** contém códigos simples atribuídos às variáveis
    - A variável *estado civil* pode ser categorizada como solteiro, casado, divorciado
    - Pode ser representado por variáveis binomiais: sim / não, verdadeiro / falso
  - **Ordinais:** similar ao nominal, porém é possível estabelecer um ranking
    - Exemplo: avaliação ruim, regular, bom
  - **Numéricos:** utilizados de forma categórica; não é possível efetuar cálculos
    - Eles também são discretos: valores finitos

# Análise Exploratória de Dados

## Conjuntos de Dados - *Datasets*

- **Quantitativos ou Numéricos:** representam os valores numéricos de variáveis específicas
  - **Discretos:** valores inteiros, contagens
    - Idade, quantidade populacional
  - **Contínuos:** valores em escala contínua
    - Massa, velocidade

Pode-se efetuar cálculos com resultados, como, por exemplo, média e desvio padrão da idade da população no Brasil.



# Análise Exploratória de Dados

Alguns tipos de Conjuntos de Dados - *Datasets*

- Banco de dados relacionais
- Banco de dados não-relacionais
- Arquivos *json*, *csv* ou *xls*
- Arquivos de logs
- API's
- Dados streamings
- Dados de sensores (IoT)

# Análise Exploratória de Dados

## *JSON*

- ***JSON (JavaScript Object Notation)***
  - Arquivo de texto amplamente utilizado para armazenar e trocar dados estruturados
  - Pode ser facilmente lido e interpretado por programas de computador
  - Sintaxe simples baseada em pares chave-valor, onde cada chave é um nome que identifica um valor específico
  - Os valores podem ser números, strings, objetos, arrays, ou booleanos, e podem ser aninhados para criar estruturas mais complexas



# Análise Exploratória de Dados

## JSON

- O arquivo descreve um possível *dataset* de filmes:

```
[
  {
    "year" : 2013,
    "title" : "Turn It Down, Or Else!",
    "info" : {
      "directors" : [ "Alice Smith", "Bob Jones"],
      "release_date" : "2013-01-18T00:00:00Z",
      "rating" : 6.2,
      "genres" : ["Comedy", "Drama"],
      "image_url" : "http://ia.media-imdb.com/images/N/09ERWAW7FS797AJ7LU8HN09AMUP908RL1o5JF90EWR7LJKQ7@@._V1_SX400_.jpg",
      "plot" : "A rock band plays their music at high volumes, annoying the neighbors.",
      "actors" : ["David Matthewman", "Jonathan G. Neff"]
    }
  },
  {
    "year": 2015,
    "title": "The Big New Movie",
    "info": {
      "plot": "Nothing happens at all.",
      "rating": 0
    }
  }
]
```

# Análise Exploratória de Dados

## CSV

- **CSV** (*Comma-Separated Values*)
  - Arquivo de texto simples que é usado para armazenar dados em tabelas
  - Cada linha do arquivo representa uma linha da tabela, enquanto cada valor separado por vírgula representa uma coluna da tabela
  - O formato CSV é amplamente utilizado para importar e exportar dados de programas de planilhas eletrônicas, bancos de dados e outras aplicações que trabalham com tabelas de dados

# Análise Exploratória de Dados

## CSV

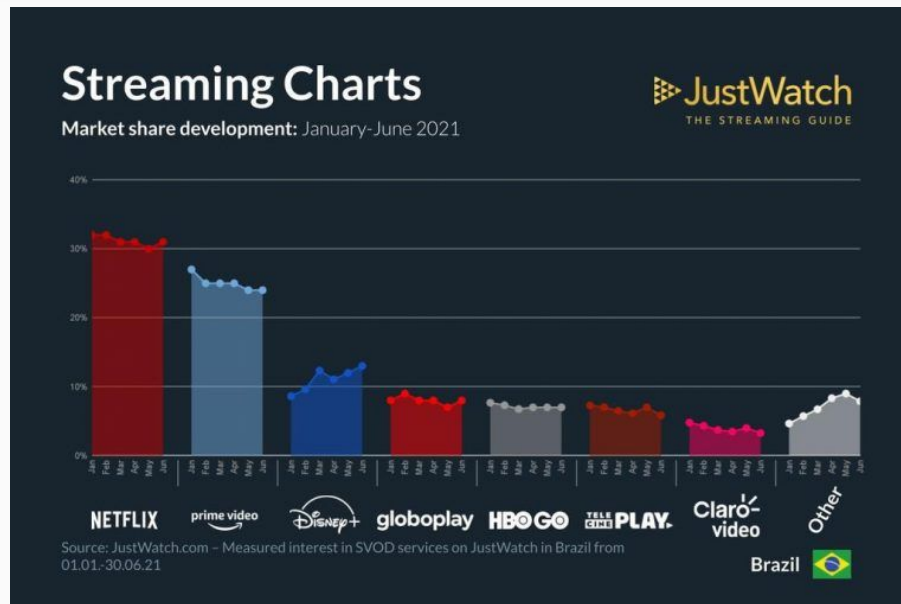
- O arquivo descreve um possível *dataset* de imóveis:

```
city,area,rooms,bathroom,parking spaces,floor,animal,furniture,hoa (R$),rent amount (R$)
São Paulo,70,2,1,1,7,accept,furnished,2065,3300,211,42,5618
São Paulo,320,4,4,0,20,accept,not furnished,1200,4960,1750,63,7973
Porto Alegre,80,1,1,1,6,accept,not furnished,1000,2800,0,41,3841
Porto Alegre,51,2,1,0,2,accept,not furnished,270,1112,22,17,1421
São Paulo,25,1,1,0,1,not accept,not furnished,0,800,25,11,836
São Paulo,376,3,3,7,-,accept,not furnished,0,8000,834,121,8955
Rio de Janeiro,72,2,1,0,7,accept,not furnished,740,1900,85,25,2750
São Paulo,213,4,4,4,4,accept,not furnished,2254,3223,1735,41,7253
São Paulo,152,2,2,1,3,accept,furnished,1000,15000,250,191,16440
Rio de Janeiro,35,1,1,0,2,accept,furnished,590,2300,35,30,2955
São Paulo,26,1,1,0,2,accept,furnished,470,2100,150,27,2747
Campinas,46,1,1,1,10,accept,not furnished,550,580,43,8,1181
São Paulo,36,1,1,0,11,accept,not furnished,359,2100,70,27,2556
São Paulo,55,1,1,1,2,accept,furnished,790,4200,224,54,5268
São Paulo,100,2,2,2,24,accept,furnished,900,4370,17,56,5343
Campinas,330,4,6,6,-,accept,furnished,680,8000,328,121,9129
```

# Análise Exploratória de Dados


## Streaming

- Plataformas que transmitem o conteúdo continuamente, em pequenos pacotes, ao invés de exigir que o usuário faça o download completo antes da reprodução



# Análise Exploratória de Dados

## Dados de Sensores (IoT)

- Dados que são emitidos continuamente, provenientes das medidas de sensores
    - Exemplos:
      - Distância
      - Umidade
      - Temperatura
      - Pressão
      - Velocidade
      - Aceleração
- 





# Análise Exploratória de Dados

## *Dados de Sensores (IoT)*

- Dados que são emitidos continuamente, provenientes das medidas de sensores

- Exemplos:

- Distância
- Umidade
- Temperatura
- Pressão
- Velocidade
- Aceleração





# Análise Exploratória de Dados

## Alguns repositórios de datasets públicos

- Existem muitos repositórios de **datasets**, alguns são públicos, outros privados (*é necessário pagar para usar*)
- Alguns exemplos de repositórios públicos são:
  - Kaggle: <https://www.kaggle.com/>
  - UCI: <https://archive.ics.uci.edu/>

kaggle





# Análise Exploratória de Dados

## HANDS-ON!

- Vamos **explorar** e **analisar** alguns diferentes **datasets** para entender como os dados se comportam em cada um deles, quais são os atributos (features), possíveis problemas e algumas sugestões simples de solução

# Análise de *Dataset*:

## *Significant Earthquakes, 1965-2016*

Dataset inclui registro da *data*, *hora*, *localização*, *profundidade*, *magnitude* e *origem* de abalos sísmicos com magnitude relatada de 5,5 ou superior (entre 1965 e 2016)



# Análise de Dataset:

## *Brazilian houses to rent*

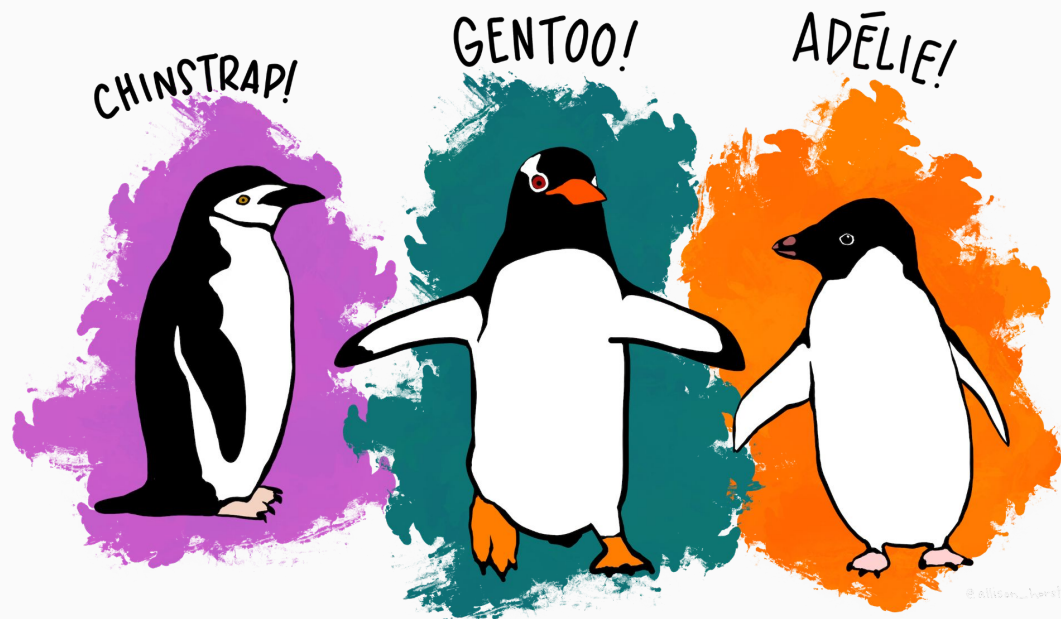
*Dataset* com dados de casas para alugar no ano de 2020

- Possui 10692 casas para alugar



# Análise de Dataset:

## *Dataset Penguins*



# Para próxima aula

- Atividade em dupla
- Buscar um dataset com dados relacionados à alguma ODS (ou a todas)
- Apresentar para sala o dataset encontrado, explicitando:
  - sobre o que é o dataset
  - a quantidade de amostras
  - quantidade de atributos (*features*)
  - apresentar os dados quantitativos e qualitativos
  - possíveis problemas (falta de valores, dados incorretos, formatos diferentes sendo usados na mesma característica)
  - Mostrar possíveis soluções sobre os problemas encontrados
  - Apresentar gráficos e conclusões que a dupla conseguiu tirar a partir dos dados do dataset
  - Fazer tudo utilizando Excel e Powerpoint