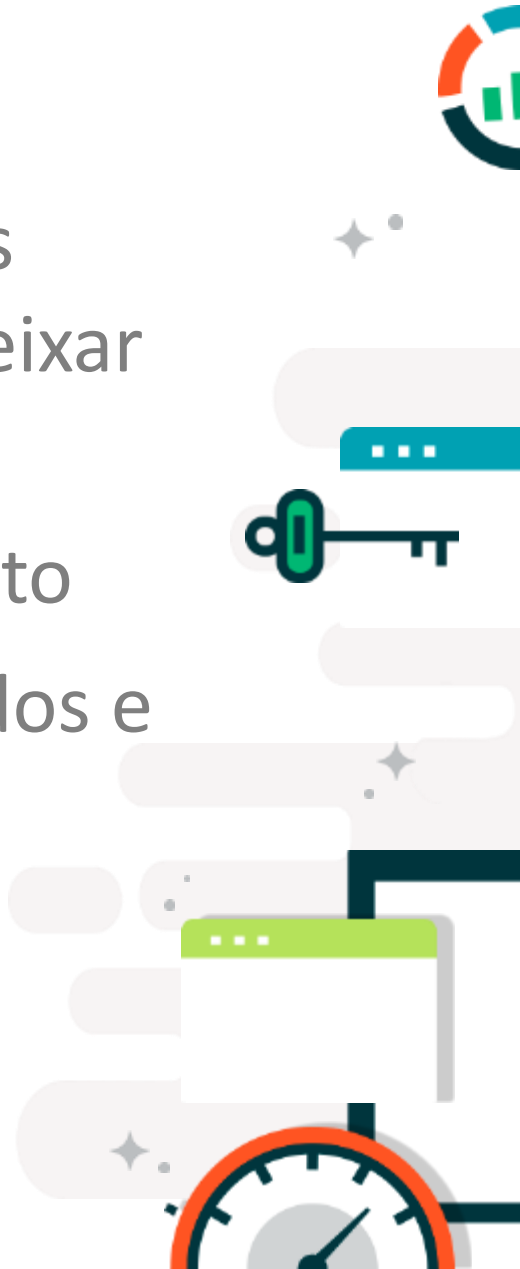
# Preparação de Dados

Prof. Sandro Jerônimo de Almeida

# Preparação de dados

- Conjunto de tarefas de pré-processamento, realizadas manualmente ou com uso de ferramentas, visando deixar os dados aptos para a serem explorados
- Geralmente consome de 70 a 80% do tempo do projeto
- **Qualidade**: devemos garantir que estejam padronizados e estatisticamente consistentes.
- Exemplos: formatos datas, moedas, medidas.

  **dd**/mm/aaaa  mm/**dd**/aaaa  01/**13**/2020



# Tarefas de Pré-Processamento

- Eliminação manual de atributos
- Integração de dados
- Amostragem de dados
- Balanceamento de dados
- Limpeza de dados
- Redução de dimensionalidade
- Transformação de dados

| Atrib. 1 | Atrib. 2 | Atrib. 3 | Atrib. 4 | ... |
|---|---|---|---|---|
| Reg.1 | | | | |
| Reg. 2 | | | | |
| .... | | | | |

**Resumo**

- Atributos: remoção e seleção de colunas
- Registros: remoção e seleção de linhas
- Transformação de dados



# Atributos: remoção e seleção de colunas

- Alguns atributos podem não ser importantes e devem ser eliminados
- A remoção com base no conhecimento da área
- A remoção com base em análises estatísticas que determinem a relevância do atributo
- Redução de dimensionalidade / *feature engineering*

*Principal Component Analysis – Pearson, K. (1901)*

# Atributos: remoção e seleção de colunas

kaggle

## Base dados exemplo

- kaggle.com
- Airplane Crash Data Since 1908
- Aircraft Accidents from 1908 to 2019

Acidente: 22 de Agosto de 2020.
Juba, Sudão do Sul. Fonte: http://planecrashinfo.com/

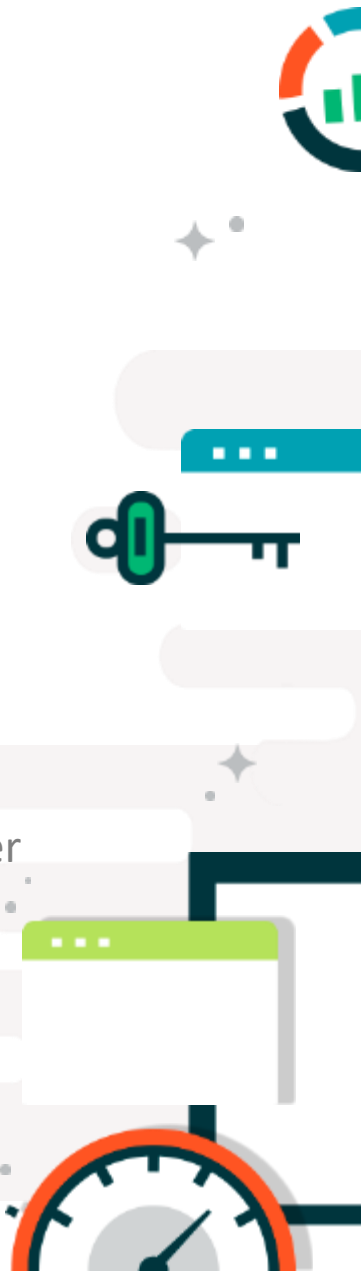
# Atributos: remoção e seleção de colunas

## Objetivo

- Analisar os impactos dos acidentes em termo e vítimas
- Podemos remover colunas não relacionadas ao objetivo

**Colunas/Atributos da base de dados**

- **Date**: Date of accident, in the format - January 01, 2001
- **Time**: Local time, in 24 hr. format unless otherwise specified
- **Airline/Op**: Airline or operator of the aircraft
- **Flight #**: Flight number assigned by the aircraft operator
- **Route**: Complete or partial route flown prior to the accident
- **AC Type**: Aircraft type
- **Reg**: ICAO registration of the aircraft
- **cn / ln**: Construction or serial number / Line or fuselage number
- **Aboard**: Total aboard (passengers / crew)
- **Fatalities**: Total fatalities aboard (passengers / crew)
- **Ground**: Total killed on the ground
- **Summary**: Brief description of the accident and cause if known



# Atributos: remoção e seleção de colunas

- Quais atributos são relevantes para minha análise?

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Date | Time | Location | Operator | Flight # | Route | AC Type | Registratic | cn/ln | Aboard | Aboard Passangers | Aboard Crew | Fatalities | Fatalities Passangers | Fatalities Crew | Ground | Summary | | |
| 4943 | 01/29/2018 | | Zhengchang, Suiyang ( | People's Liberation Army Air Force | | Training | Shaanxi Y-8GX-3 | 30513 | | 12 | 7 | 5 | 12 | 7 | 5 | 0 | While completing a training miss | | |
| 4944 | 02/10/2018 | 17:31 | Grand Canyon, Arizona | Papillon | | Sightseeing | Eurocopter EC 130B4 | N155GC | 7091 | 7 | 6 | 1 | 3 | 3 | 0 | 0 | The helicopter was observed ma | | |
| 4945 | 02/11/2018 | 14:31 | Argunovo, Russia | Saratov Airlines | 703 | Moscow - Orsk | Antonov AN-148-100 | RA-61704 | 2,7E+10 | 71 | 65 | 6 | 71 | 65 | 6 | 0 | The regional airliner took off fro | | |
| 4946 | 02/18/2018 | 09:32 | Kohangan village, Iran | Iran Aseman Airlines | EP3704 | Tehran - Yasuj | ATR 72-212 | EP-ATS | 391 | 66 | 60 | 6 | 66 | 60 | 6 | 0 | The airliner was on approach to Y | | |
| 4947 | 03/06/2018 | 14:51 | Latakia, Syria | Military - Russian Air Force | | Kuweires Air Ba | Antonov An-26 | RF-92955/ | 10107 | 39 | 33 | 6 | 39 | 33 | 6 | 0 | While on approach to Latakia-Kh | | |
| 4948 | 03/11/2018 | 19:08 | New York, New York | Liberty Helicopters | | Sightseeing | Eurocopter AS 350B2 | N350LH | 7654 | 6 | 5 | 1 | 5 | 5 | 0 | 0 | The sightseeing helicopter sudde | | |
| 4949 | 03/12/2018 | 14:15 | Kathmandu, Nepal | US-Bangla Airlines | 221 | Dhaka - Kathmar | de Havilland Canada DHC-8 | S2-AGUÂ | BS4041 | 71 | 67 | 4 | 51 | 47 | 4 | 0 | After getting clearance to land, t | | |
| 4950 | 04/11/2018 | 08:00 | Boufarik AB, Algeria | Military - Algerian Air Force | | Boufarik AB - Be | Ilyushin 76-TD | 7T-WIP | 1,04E+09 | 257 | 247 | 10 | 257 | 247 | 10 | 0 | The Algerian military plane crash | | |
| 4951 | 04/17/2018 | 10:04 | NW of Philadelphia, P | Southwest Airlines | 1380 | New York - Dalla | Boeing 737-7H4 | N722SW | 27880/601 | 149 | 144 | 5 | 1 | 1 | 0 | 0 | While climbing to FL320, the No. | | |
| 4952 | 05/02/2018 | 11:30 | Port Wentworth, Geor | Military - US Air Force | | Savannah - Tusc | Lockheed HC-130H Hercule | 65-0968 | 382-4110 | 9 | 0 | 9 | 9 | 0 | 9 | 0 | A Porto Rico Air National Guard p | | |
| 4953 | 05/18/2018 | 12:08 | Havana, Cuba | Cubana (leased from Glo | 972 | Havana- Holguin | Boeing 737-201 | XA-UHZ | 21816/592 | 113 | 107 | 6 | 112 | 106 | 6 | 0 | After taking off from runway 06 a | | |
| 4954 | 07/10/2018 | 07:44 | Pretoria, South Africa | Rovos Air | | Test Flight | Convair CV-340 | ZS-BRV | 215 | 19 | 16 | 3 | 1 | 0 | 1 | 1 | Shortly after takeoff from runwa | | |
| 4955 | 08/04/2018 | 16:55 | Flims, Switzerland | Ju Air | | Locorno - Duben | Junkers JU-52 | HB-HOT | 6595 | 20 | 17 | 3 | 20 | 17 | 3 | 1 | The vintage aircraft crashed onto | | |
| 4956 | 09/28/2018 | 10:10 | Chuuk, Micronesia | Air Niugini | | Pohnpei - Chuuk | Boeing 737-8BK | P2-PXE | 33024/168 | 47 | 35 | 12 | 1 | 1 | 0 | 0 | The aircraft was approaching for | | |
| 4957 | 10/29/2018 | 06:31 | Off Jakarta, Indonesia | Lion Air | 610 | Jakarta - Pangka | Boeing 737-MAX 8 | PK-LQP | 43000/705 | 189 | 181 | 8 | 189 | 181 | 8 | 0 | The airliner crashed into the Jaka | | |
| 4958 | 11/06/2018 | 02:53 | Georgetown, Guyana | Fly Jamaica Airways | | Georgetown - Tc | Boeing 757-N23 | N524ATÂ | 30233/895 | 128 | 120 | 8 | 1 | 1 | 0 | 0 | After taking off and reaching FL2 | | |
| 4959 | 01/14/2019 | 08:30 | Karaj, Iran | Saha Air | | Bishkek - Payam | Boeing 707-3J9C | EP-CPP | 21128/917 | 16 | 13 | 3 | 15 | 13 | 2 | 0 | The cargo plane was operated by | | |
| 4960 | 02/23/2019 | 12:45 | Houston, Texas | Atlas Air serving Amazon | 3591 | Miami - Houstor | Boeing 767-375ER | N1217A | 25865 | 3 | 0 | 3 | 3 | 0 | 3 | 0 | ATC lost radar contact with the ca | | |
| 4961 | 03/09/2019 | 10:40 | Vereda La Bendicion, ( | Laser Aereo Colombia | | San Jose - Villav | Douglas DC-3 | HK-2494 | :33105/163 | 14 | 11 | 3 | 14 | 11 | 3 | 0 | While on approach to land, the cr | | |
| 4962 | 03/10/2019 | 08:44 | Bishoftu, Ethiopia | Ethiopian Airlines | 302 | Addis Ababa - N | Boeing 737 Max 8 | ET-AVJ | 63450/724 | 157 | 149 | 8 | 157 | 149 | 8 | 0 | The internationally scheduled air | | |

# Registros: remoção e seleção de linhas

- Alguns registros podem não ser consistentes
  > Dados conflitantes: registros duplos
  > Dados incorretos: entradas errôneas, ruídos
  > Dados omissos ou faltantes
  > Outliers
- Uma limpeza na base de dados pode ser requerida

# Registros: remoção e seleção de linhas

- Ruídos

- Modificação dos valores originais

- Ruídos de leitura

- Ruídos de medição

- Ruídos de fundo

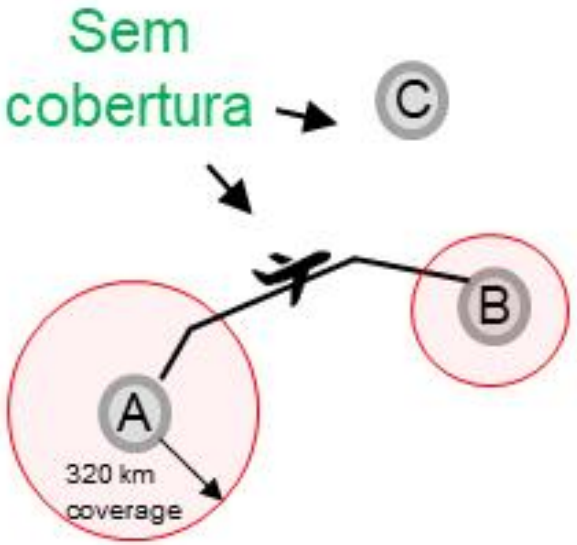
- Outliers

Dados com características consideravelmente diferentes da maioria dos dados no dataset

# Registros: remoção e seleção de linhas

- Dados omissos ou faltantes
- Voos com trajetórias incompletas (apenas 4% completas)

| Antena coletora ADS-B | Coletor de dados ADS-B FlightAware | Cobertura de antenas ADS-B pelos usuários da empresa FlightAware |
|---|---|---|
|  |   |   |



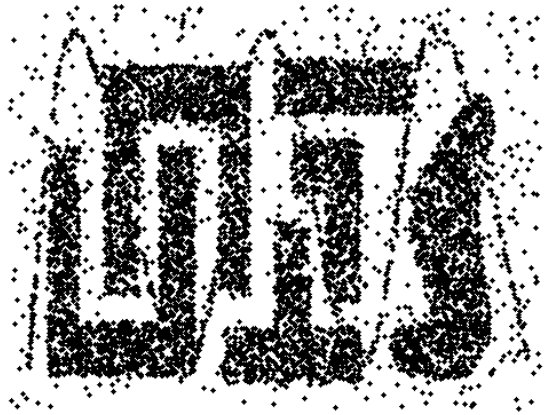

Fonte: Almeida, S.J et al (2018) - A method for estimating flight paths missing data

# Registros: remoção e seleção de linhas

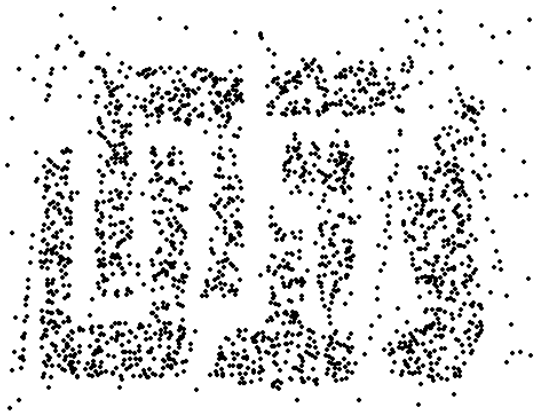
- Estratégias para tratar dados omissos ou incompletos:
  > Remoção dos registros
  > Novos dados podem ser inferidos/estimados
  > Dados pode ser tratados como casos especiais

# Registros: remoção e seleção de linhas

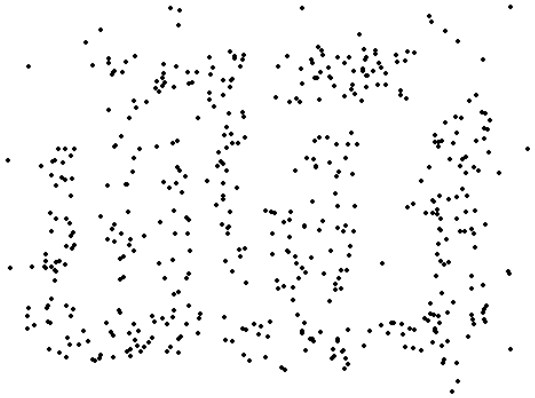
- Seleção usando amostragem de dados
- Uma amostra pode ser significativa para a análise



8000 pontos



2000 pontos



500 pontos

# Transformação de Dados

## Agregação de Colunas

- Combina dois ou mais atributos em um único atributo.
- Exemplo

Agregação em nova coluna

| N_Criancas | N_Adultos | Total_Pessoas |
|---|---|---|
| 2 | 2 | 4 |
| 2 | 1 | 3 |
| 1 | 2 | 3 |
| … | … | … |

## Agregação de Linhas

- Resumir registros em função de alguns atributos
- Nota ENEM de cada cidade pode oferecer um valor médio da nota por Estado
- Funções de agregação: contagem, soma, média, moda, desvio padrão, contagem, min, max

# Transformação de Dados

- Algoritmos e ferramentas pode ser limitados quanto ao tipo e formato dos dados
- Algumas conversões pode ser necessárias
  > Categórico não ordinal para binominal
  > Categórico ordinal para numérico
  > Numérico para numérico
  (mudança de escala ou normalização de dados)

# Ferramentas de Pré-processamento

- Editor de arquivos texto (ex. bloco de notas)
- Planilhas eletrônicas (ex. Google Planilhas, Excel Online)
- Ferramentas ETL
- Ferramentas de Aprendizado de Máquina

  (Ex. Orange Data Mining)
- Programação: algoritmos/próprio código/scripts


PUC Minas
Virtual