O Histograma como ferramenta de Análise de Dados

A capacidade de contar uma história através de gráficos é uma necessidade implícita na atividade do Cientista de Dados.

Além de extrair informações através da construção de modelos em Machine Learning, a montagem de visualizações capazes de adicionar valor ao projeto é extremamente importante.

Nesta aplicação trabalharemos o gráfico chamado histograma, e sua interpretação diante de uma demanda.

Histograma “o histograma, também conhecido como distribuição de frequências, é a representação gráfica em colunas ou em barras de um conjunto de dados previamente tabulado e dividido em classes uniformes ou não uniformes. A base de cada retângulo representa uma classe” Fonte: Wikipédia.

Notebook do artigo

Este trabalho em Python (Jupyter Notebook/Anaconda) usou a biblioteca Pandas, para auxiliar e tornar mais amigável a manipulação e análise de dados. Para tanto foi utilizado uma massa de dados que está a disposição no Kaggle, chamado athlete_events.csv. O mesmo representa um dataset histórico dos Jogos Olímpicos Modernos (Atenas (1896) até Rio de Janeiro (2016)), e é disponibilizado na extensão csv. Abaixo está demonstrado a captura e apresentação do dataset em seu formato original.

```
In [1]: import pandas as pd
```

```
In [2]: dados = pd.read_csv('C:/Users/prate/Documents/Lidiano/Estudos/000DidáticaTech/PythonparaML/DadosKaggle/athlete_events
```

```
In [3]: dados.head()
```

Out[3]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN

```
In [ ]:
```

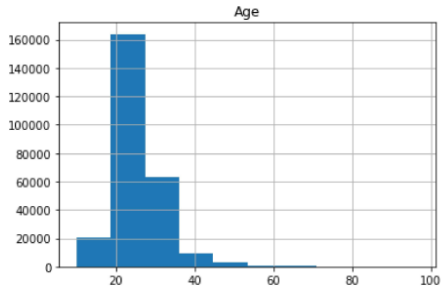
```
In [4]: import matplotlib.pyplot as plt
```

O estudo abaixo visa analisar a idade dos participantes dos Jogos Olímpicos Modernos ao longo do tempo, com a ferramenta Pandas numa amostragem de 10 colunas. Nesta configuração (dados.hist(column = ‘Age’, bins=10)), são demonstrados grupos compactados, onde com certeza não conseguiremos tirar informações mais detalhadas, como segue.

- A idade das pessoas varia de zero (0) a cem (100) anos.
- A idade dos participantes varia entre 10 e 90 anos.

- Calculando a idade máxima menos a idade mínima, temos um range de oitenta (80) anos (90 – 10 = 80). Como estamos criando dez (10) barras em nosso histograma, cada barra terá o tamanho de oito (8) anos, de acordo com o cálculo a seguir:
    - Range / número de barras solicitadas = 80 / 10 = 8....então:
        - Barra1 = 10 + 8 = 18 anos > média de 20.000 atletas
        - Barra2 = 18 + 8 = 26 anos > 160.000 atletas
        - Barra3 = 26 + 8 = 34 anos > 60.000 atletas
        - Barra4 = 34 + 8 = 42 anos > menor que 20.000 atletas

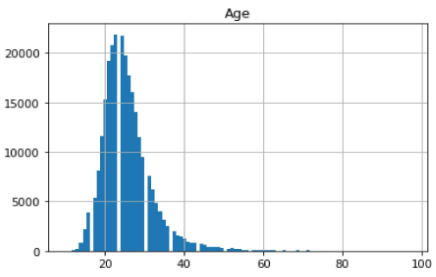
```
In [5]: dados.hist(column = 'Age', bins=10)
Out[5]: array([[<matplotlib.axes._subplots.AxesSubplot object at 0x000001AADFAD3CC8>]],
              dtype=object)
```



O estudo abaixo, com uma amostragem muito melhor (dados.hist(column = 'Age', bins=100)), nos ajuda a explicar com mais exatidão os fenômenos que os dados nos proporcionam.

- A idade das pessoas varia de zero (0) a cem (100) anos.
- A idade dos participantes varia entre 10 e 90 anos.
- Calculando a idade máxima menos a idade mínima, temos um range de oitenta (80) anos (90 – 100 = 80). Como estamos criando cem (100) barras em nosso histograma, cada barra terá o tamanho de 0.8 anos, de acordo com o cálculo a seguir:
    - Range / numero de barras solicitadas = 80 / 100 = 0,8..então:
        - Barra1 = 10 + 0,8 = 10,8 ou 11 anos > bem poucos atletas, o que demonstra que no gráfico anterior tínhamos uma visão errônea das quantidades, pois sua distribuição não proporcionava determinadas conclusões.
        - Podemos verificar que existe uma maior densidade em idades entre dezoito (18) e vinte e cinco (25) anos.

```
In [9]: dados.hist(column = 'Age', bins=100)
Out[9]: array([[<matplotlib.axes._subplots.AxesSubplot object at 0x000001AAE14E7DC8>]],
              dtype=object)
```



Conclusão: o histograma em sua função básica, conta a quantidade de ocorrências, criando faixas de valores que quanto maior for sua granularidade melhor será a análise da distribuição, na maioria dos casos.