Aula 5 - Remoção de outliers

UNIFESP

Universidade Federal de São Paulo

1933

AGENDA

- O que é um outlier?
- Identificação de outliers
- Remoção de outliers
- Exemplo

O que são *outliers*?

- Dentro de um conjunto de dados, os *outilers*, são as amostras que mais se diferenciam das demais.
- Os *outliers* estão fora do padrão que se observa para os demais dados do conjunto.
- A existência desse tipo de amostra no conjunto pode causar anomalias nas análises efetuadas e na execução dos algoritmos de ML

O que são *outliers*?

- No conjunto de dados abaixo, São Paulo e Rio de Janeiro estão mais distantes dos demais, ou seja, possivelmente são outliers.

Nº	Município	População	Nº	Município	População
1	São Paulo (SP)	988,8	16	Nova Iguaçu (RJ)	83,9
2	Rio de Janeiro (RJ)	556,9	17	São Luís (MA)	80,2
3	Salvador (BA)	224,6	18	Maceió (AL)	74,7
4	Belo Horizonte (MG)	210,9	19	Duque de Caxias (RJ)	72,7
5	Fortaleza (CE)	201,5	20	São Bernardo do Campo (SP)	68,4
6	Brasília (DF)	187,7	21	Natal (RN)	66,8
7	Curitiba (PR)	151,6	22	Teresina (PI)	66,8
8	Recife (PE)	135,8	23	Osasco (SP)	63,7
9	Porto Alegre (RS)	129,8	24	Santo André (SP)	62,8
10	Manaus (AM)	119,4	25	Campo Grande (MS)	61,9
11	Belém (PA)	116,0	26	João Pessoa (PB)	56,2
12	Goiânia (GO)	102,3	27	Jaboatão (PE)	54,1
13	Guarulhos (SP)	101,8	28	Contagem (MG)	50,3
14	Campinas (SP)	92,4	29	São José dos Campos (SP)	49,7
15	São Gonçalo (RJ)	84,7	30	Ribeirão Preto (SP)	46,3

Fonte: Estatística básica - Wilton de Oliveira Bussab, Pedro Alberto Morettin

Verificando a existência de *outliers*

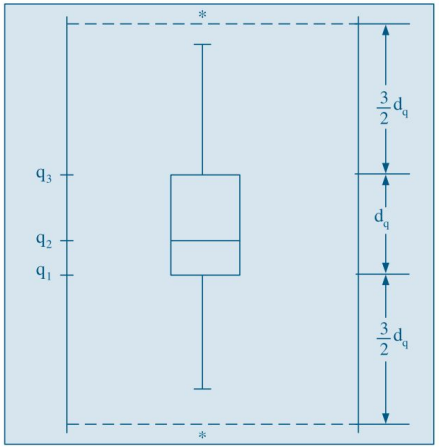
- Uma ferramenta bastante útil para verificar a existência de outliers em um conjunto é o Box-Plot
- O Box-plot é um representação gráfica da variação dos dados por meio de quartis.
- Os quartis são três valores que dividem um conjunto de dados em quatro subconjuntos cada um contendo 25% dos dados originais de maneira ordenada.

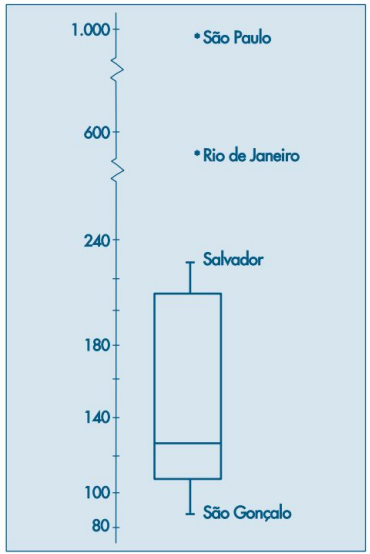
Verificando a existência de *outliers*

- Com os dados ordenados os quartis são obtidos da seguinte forma:
    - Primeiro quartil: item na posição 0.25*(N+1)
    - Segundo quartil:
        - Em caso de N ímpar: item na posição (N+1)
        - Em caso de N par: média dos itens (N/2) e ((N+1)/2)
    - Terceiro quartil: item na posição 0.75*(N+1)

Verificando a existência de *outliers*

- Com os quartis calculados o Box-plot pode ser construído.
- A partir dos dados apresentados anteriormente temos o Box-plot abaixo:





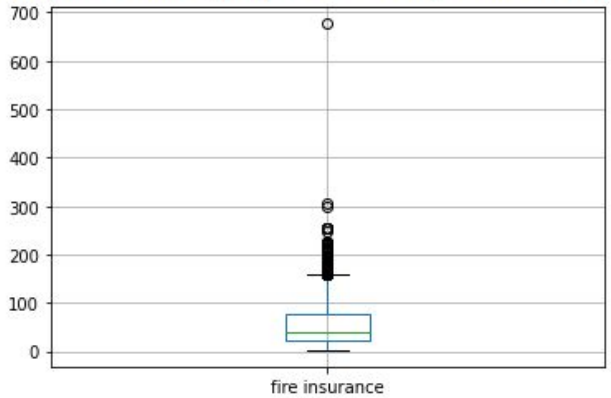
Fonte: Estatística básica - Wilton de Oliveira Bussab, Pedro Alberto Morettin

Box-plot no Pandas

- Com o conjunto de dados carregado em um *DataFrame* pode-se utilizar a biblioteca Pandas para plotar o Box-plot
- Sintaxe: *df.boxplot(column=['A'], return_type='axes')*

```
[6]  dataset.boxplot(column=['fire insurance'], return_type='axes')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7af6698e50>
```



Remoção de outliers

- Após a análise e a detecção dos outliers, dependendo do tipo de problema que se está abordando e a resposta buscada é recomendada a remoção dos outliers.
- Antes de remover os *outliers*:
  - Eles são anomalias ou são valores possíveis de serem observados?
  - O quanto o valor médio é afetado pela presença dos outliers?
  - Os *outliers* são possibilidades reais que devem ser detectadas pelo algoritmo de ML?

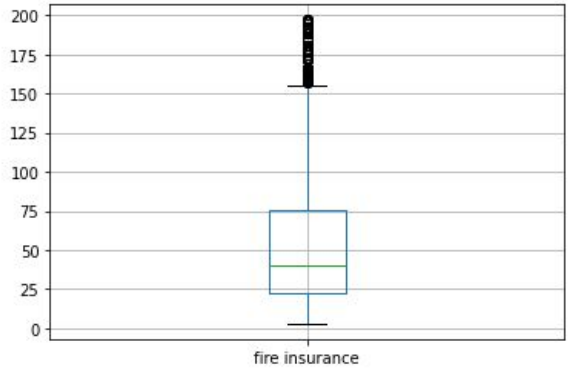
Remoção de outliers - Z Score

- O Z Score fornece um ideia de quão longe um dado está da média do conjunto[1].
- Segundo [2] pode ser entendido como “o número de desvios padrão em relação à média de um ponto de informação”
- O Z Score para a amostra i do conjunto é calculado por: $z_i = (x_i - \text{média})/dp$
- Para literatura estatística o valor aceitável de z score está dentro -3 a 3, portanto, amostras com z score fora de intervalo possivelmente são outliers

Remoção de outliers - Z Score

```
[13] from scipy import stats

     #calcula o z-score para a coluna fire insurance e cria uma nova coluna para armazenar o resultado
     dataset['fire insurance_zscore'] = np.abs(stats.zscore(dataset['fire insurance']))

     #percorre o dataset removendo as linhas onde o z score calculado é maior que 3
     for index,item in dataset.iterrows():
       if(item['fire insurance_zscore']>3):
         dataset.drop(labels=[index],axis=0,inplace=True)

     #remove a coluna com o z score
     dataset.drop(labels=['fire insurance_zscore'],axis=1,inplace=True)
     dataset.boxplot(column=['fire insurance'], return_type='axes')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fdc9f84d9d0>
```



Remoção de outliers - IQR Score

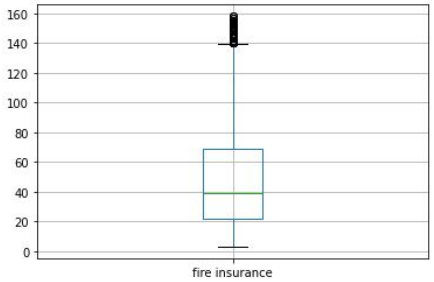
- O IQR(distância interquartil) definida como $Q_3 - Q_1$
- O IQR é calculado para a coluna de dados na qual pretende-se remover outliers
- Com o IQR calcula verifica se a condição:
    - se valor < (Q_1 -1.5*IQR) ou valor > (Q_3 +1.5*IQR) é outlier

Remoção de outliers - IQR Score

- No exemplo abaixo são removidos os outliers da coluna ‘fire insurance’

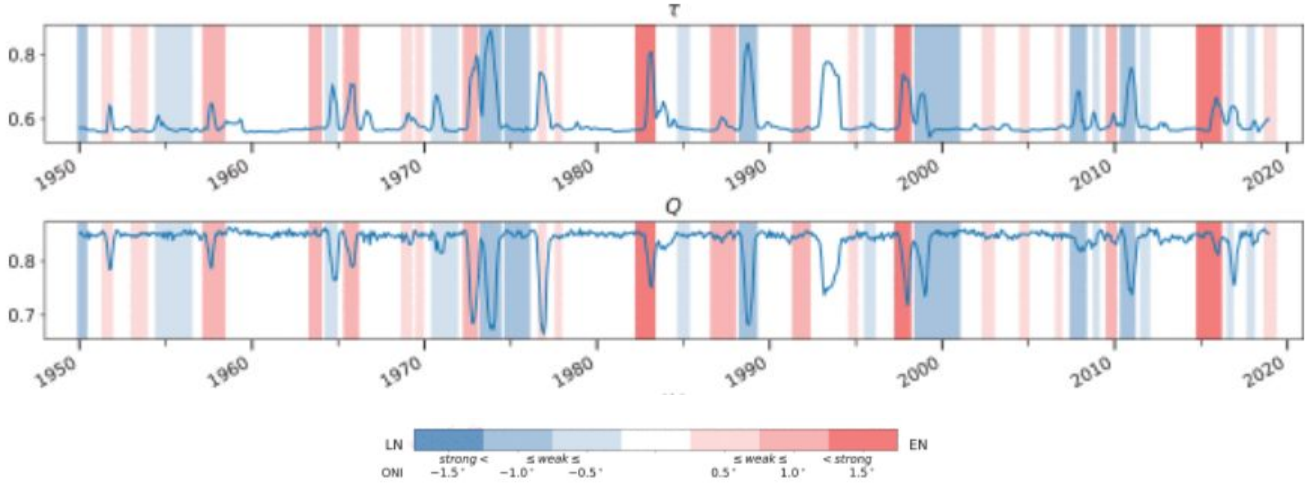
```
[16] Q1 = dataset['fire insurance'].quantile(0.25)
     Q3 = dataset['fire insurance'].quantile(0.75)

     IQR = Q3 - Q1
     print(IQR)

     #percorre o dataset removendo as linhas onde o z score calculado é maior que 3
     for index,item in dataset.iterrows():
       if((item['fire insurance']<(Q1-1.5*IQR)) or (item['fire insurance']>(Q3+1.5*IQR))):
         dataset.drop(labels=[index],axis=0,inplace=True)

     #remove a coluna com o z score
     dataset.boxplot(column=['fire insurance'], return_type='axes')
```

```
54.0
<matplotlib.axes._subplots.AxesSubplot at 0x7fdca150c150>
```



Nem sempre remover os outliers é correto

- Exemplo abaixo é referente a utilização de métricas de redes complexas para dectação de El Niño
- Nesse caso os pontos considerados outliers eram exatamente aqueles que importavam



M. A. De Castro Santos, D. A. Vega-Oliveros, L. Zhao and L. Berton, "Classifying El Niño-Southern Oscillation Combining Network Science and Machine Learning," in *IEEE Access*, vol. 8, pp. 55711-55723, 2020, doi: 10.1109/ACCESS.2020.2982035.

Referências bibliográficas

[1] Z-Score: Definition, Formula and Calculation - Statistics How To

[2] O que é um Z-Score? — Matemática e Estatística — DATA SCIENCE

GRATIDÃO!

UNIFESP

UNIVERSIDADE FEDERAL DE SÃO PAULO

1933