RA272746_aula4

April 19, 2024

0.1 IA376I – Tópicos em Engenharia de Computação VII

0.1.1 Tópico: Análise de Dados Visual (Visual Analytics)

Professora: Wu, Shin - Ting **Aluno:** Luiz Roberto Albano Junior **RA:** 272746

0.1.2 Exercícios 5.6

Item 1: Faça os itens 1 – 7 da Seção 12.10 da referência [61], usando o conjunto de dados Galton em [86]. Carregamento dos dados e bibliotecas necessárias

[ ]:
```
import pandas as pd
from plotnine import *
from scipy import stats

#Importação dos dados
galton = pd.read_csv("GaltonFamilies.csv")

#Filtra os filhos de gênero masculino
galton = galton[galton['gender'] == "male"]
galton.head()
```

[ ]:
```
    rownames family  father  mother  midparentHeight  children  childNum  \
0          1    001    78.5    67.0            75.43         4         1
4          5    002    75.5    66.5            73.66         4         1
5          6    002    75.5    66.5            73.66         4         2
8          9    003    75.0    64.0            72.06         2         1
10        11    004    75.0    64.0            72.06         5         1

   gender  childHeight
0    male         73.2
4    male         73.5
5    male         72.5
8    male         71.0
10   male         70.5
```

[ ]:
```
#Adiciona os dados de altura dos filhos em um vetor x
x = galton['childHeight']
```

```
x.describe()
```

```
[ ]: count    481.000000
     mean      69.234096
     std        2.623905
     min       60.000000
     25%       67.500000
     50%       69.200000
     75%       71.000000
     max       79.000000
     Name: childHeight, dtype: float64
```

1. Compute the average and median of these data.

```
[ ]: galton_avg = x.agg('mean')
     galton_median = x.agg('median')

     print(f"Média: {galton_avg:.2f}")
     print(f"Mediana: {galton_median:.2f}")
```

```
Média: 69.23
Mediana: 69.20
```

2. Compute the median and median absolute deviation of these data.

```
[ ]: galton_abs_dev = stats.median_abs_deviation(x)

     print(f"Mediana: {galton_median:.2f}")
     print(f"Desvio absoluto da mediana: {galton_abs_dev:.2f}")
```

```
Mediana: 69.20
Desvio absoluto da mediana: 1.80
```

3. Now suppose Galton made a mistake when entering the first value and forgot to use the decimal point. You can imitate this error by typing:

```
[ ]: x_with_error = x.copy()
     x_with_error.iloc[0] = x_with_error.iloc[0] * 10
     x_with_error
```

```
[ ]: 0      732.0
     4       73.5
     5       72.5
     8       71.0
     10      70.5
             …
     918     68.0
     924     64.5
     925     66.0
```

```
929     64.0
932     66.5
Name: childHeight, Length: 481, dtype: float64
```

How many inches does the average grow after this mistake?

[ ]:
```
print( x_with_error.agg('mean') - galton_avg )
```

```
1.369646569646548
```

4. How many inches does the SD grow after this mistake?

[ ]:
```
print( x_with_error.agg('std') - x.agg('std') )
```

```
27.70915374097227
```

5. How many inches does the median grow after this mistake?

[ ]:
```
print( x_with_error.agg('median') - x.agg('median') )
```

```
0.0
```

6. How many inches does the MAD grow after this mistake?

[ ]:
```
print( stats.median_abs_deviation(x_with_error) - stats.median_abs_deviation(x)␣
↪)
```

```
0.0
```

7. How could you use exploratory data analysis to detect that an error was made? Since it is only one value out of many, we will not be able to detect this. We would see an obvious shift in the distribution. **(X) A boxplot, histogram, or qq-plot would reveal a clear outlier.** A scatterplot would show high levels of measurement error.