Estudos estatíticos

Sandro Ricardo De Souza

2024-04-26

Table of contents


```
# Caminho
path_file = r'C:\Users\srsouza\Documents\Estudos\Pos-graduacao_Mineracao_Dados_IFES\datasets\
```

```
# Leitura do dataframe
custumer_info = pd.read_csv(path_file)
```

1 Amostra dos dados

Vamos obter uma amostra aleatória dos dados para verificarmos, entre outras coisas, os dados faltantes. Se não fizer diferença, vamos remover da amostra, estes dados. Podemos usar a função $f(x) = x^2$, se for preciso.

A equação

$$h(x) = \int_a^b (1 - x)dx$$

,

simula melhor.

```
cabecalho('Amostra')
print(custumer_info.sample(5))
```

```
Amostra
--------------------------------------------------------------------------------
             ID  Sex  Marital status  Age  Education  Income  Occupation  \
375   100000376    1               1   37          1  170386           1
571   100000572    0               0   38          1  147472           1
947   100000948    0               0   38          1  147760           2
1872  100001873    0               1   25          1   65207           0
792   100000793    0               1   27          1   96323           1

      Settlement size
375                 1
571                 0
947                 1
1872                0
792                 0
```

2 Informações gerais

```
cabecalho('Resumo')
print(custumer_info.info())
```

```
Resumo
--------------------------------------------------------------------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ID               2000 non-null   int64
 1   Sex              2000 non-null   int64
 2   Marital status   2000 non-null   int64
 3   Age              2000 non-null   int64
 4   Education        2000 non-null   int64
 5   Income           2000 non-null   int64
```

```
 6   Occupation       2000 non-null   int64
 7   Settlement size  2000 non-null   int64
dtypes: int64(8)
memory usage: 125.1 KB
None
```

```
# Contagens
num_columns = ['ID','Age','Income']
cat_columns = ['Sex','Marital status','Education','Occupation','Settlement size']
custumer_info[cat_columns] = custumer_info[cat_columns].astype('str')
```

3 Dados categóricos

```
cabecalho('Describe')
print(custumer_info.describe(include='object').T)
```

```
Describe
--------------------------------------------------------------------------------
                count unique top  freq
Sex              2000      2   0  1086
Marital status   2000      2   0  1007
Education        2000      4   1  1386
Occupation       2000      3   1  1113
Settlement size  2000      3   0   989
```

```
print('\nPrograma finalizado com sucesso!!!')
```

```
Programa finalizado com sucesso!!!
```