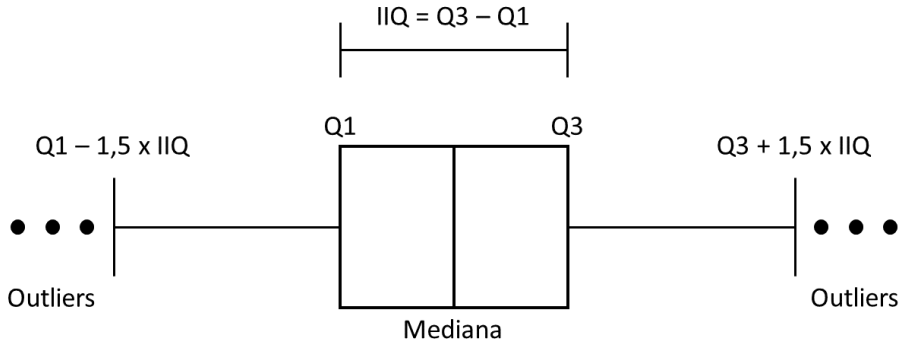
# Relatório de Análise VIII

## Identificando e Removendo Outliers

In [1]:

```
%matplotlib inline
import pandas as pd
import matplotlib.pyplot as plt
plt.rc('figure', figsize = (14, 6))
```




**Box-plot**

In [2]:

```
dados = pd.read_csv('dados/aluguel_residencial.csv',sep=';')
dados.head(10)
```

Out[2]:

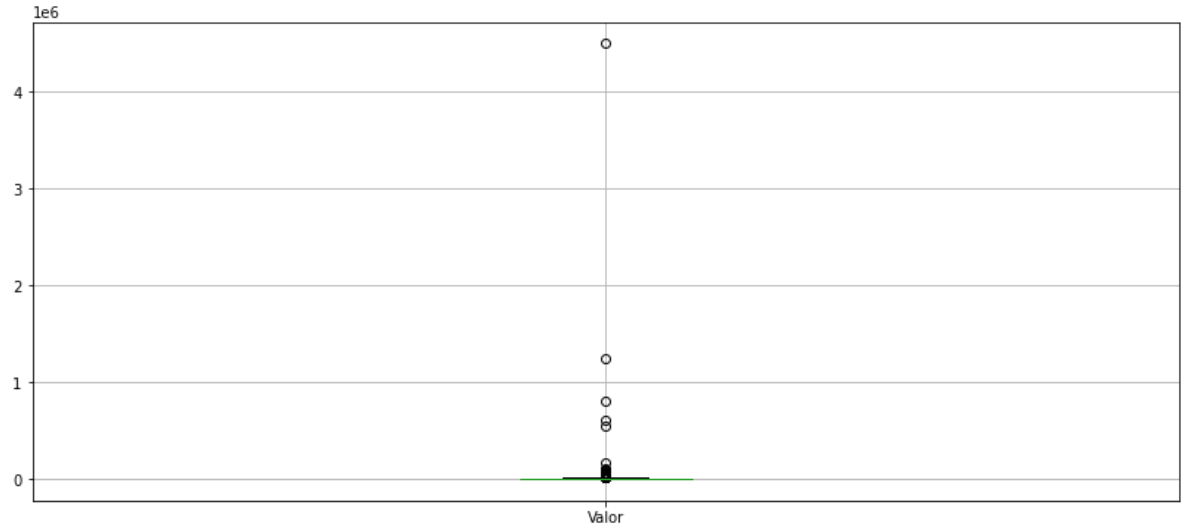
| | Tipo | Bairro | Quartos | Vagas | Suites | Area | Valor | Condominio | IPTU | Valor_i |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Quitinete | Copacabana | 1 | 0 | 0 | 40 | 1700.0 | 500.0 | 60.0 | 42 |
| **1** | Casa | Jardim Botânico | 2 | 0 | 1 | 100 | 7000.0 | 0.0 | 0.0 | 70 |
| **2** | Apartamento | Centro | 1 | 0 | 0 | 15 | 800.0 | 390.0 | 20.0 | 53 |
| **3** | Apartamento | Higienópolis | 1 | 0 | 0 | 48 | 800.0 | 230.0 | 0.0 | 16 |
| **4** | Apartamento | Cachambi | 2 | 0 | 0 | 50 | 1300.0 | 301.0 | 17.0 | 26 |
| **5** | Casa de Condomínio | Barra da Tijuca | 5 | 4 | 5 | 750 | 22000.0 | 0.0 | 0.0 | 29 |
| **6** | Casa de Condomínio | Ramos | 2 | 2 | 0 | 65 | 1000.0 | 0.0 | 0.0 | 15 |
| **7** | Apartamento | Grajaú | 2 | 1 | 0 | 70 | 1500.0 | 642.0 | 74.0 | 21 |
| **8** | Apartamento | Lins de Vasconcelos | 3 | 1 | 1 | 90 | 1500.0 | 455.0 | 14.0 | 16 |
| **9** | Apartamento | Copacabana | 1 | 0 | 1 | 40 | 2000.0 | 561.0 | 50.0 | 50 |

In [3]:

```
dados.boxplot(['Valor'])
```

Out[3]:

```
<AxesSubplot:>
```



In [4]:

```
selecao = dados['Valor'] >= 500000
dados_selecao = dados[selecao]
dados_selecao
```

Out[4]:

| | Tipo | Bairro | Quartos | Vagas | Suites | Area | Valor | Condominio | IPTU |
|---|---|---|---|---|---|---|---|---|---|
| **7629** | Apartamento | Barra da Tijuca | 1 | 1 | 0 | 65 | 600000.0 | 980.0 | 120.0 |
| **10636** | Casa de Condomínio | Freguesia (Jacarepaguá) | 4 | 2 | 3 | 163 | 800000.0 | 900.0 | 0.0 |
| **12661** | Apartamento | Freguesia (Jacarepaguá) | 2 | 2 | 1 | 150 | 550000.0 | 850.0 | 150.0 |
| **13846** | Apartamento | Recreio dos Bandeirantes | 3 | 2 | 1 | 167 | 1250000.0 | 1186.0 | 320.0 |
| **15520** | Apartamento | Botafogo | 4 | 1 | 1 | 300 | 4500000.0 | 1100.0 | 0.0 |

In [5]:

```
dados_selecao_dois = dados[dados['Valor'] >= 500000]
dados_selecao_dois
```

Out[5]:

| | Tipo | Bairro | Quartos | Vagas | Suites | Area | Valor | Condominio | IPTU |
|---|---|---|---|---|---|---|---|---|---|
| **7629** | Apartamento | Barra da Tijuca | 1 | 1 | 0 | 65 | 600000.0 | 980.0 | 120.0 |
| **10636** | Casa de Condomínio | Freguesia (Jacarepaguá) | 4 | 2 | 3 | 163 | 800000.0 | 900.0 | 0.0 |
| **12661** | Apartamento | Freguesia (Jacarepaguá) | 2 | 2 | 1 | 150 | 550000.0 | 850.0 | 150.0 |
| **13846** | Apartamento | Recreio dos Bandeirantes | 3 | 2 | 1 | 167 | 1250000.0 | 1186.0 | 320.0 |
| **15520** | Apartamento | Botafogo | 4 | 1 | 1 | 300 | 4500000.0 | 1100.0 | 0.0 |

In [6]:

```
valor = dados['Valor']
valor.shape
```

Out[6]:

```
(21826,)
```

In [7]:

```
(valor.sum() / 21826)
```

Out[7]:

```
5046.172821405663
```

In [8]:

```
Q1 = valor.quantile(.25)
Q3 = valor.quantile(.75)
IIQ = Q3 - Q1
limite_inferior = Q1 - 1.5 * IIQ
limite_superior = Q3 + 1.5 * IIQ
print(f" Q1:{Q1} Q3:{Q3} IQQ:{IIQ} Limite Inferior:{limite_inferior} Limite Superior: {limi
```

```
 Q1:1600.0 Q3:5500.0 IQQ:3900.0 Limite Inferior:-4250.0 Limite Superior: 113
50.0
```

In [9]:

```
selecao = (valor >= limite_inferior) & (valor <= limite_superior)
dados_new = dados[selecao]
dados_new
```

Out[9]:

| | Tipo | Bairro | Quartos | Vagas | Suites | Area | Valor | Condominio | IPTU | Val |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Quitinete | Copacabana | 1 | 0 | 0 | 40 | 1700.0 | 500.0 | 60.0 | |
| **1** | Casa | Jardim Botânico | 2 | 0 | 1 | 100 | 7000.0 | 0.0 | 0.0 | |
| **2** | Apartamento | Centro | 1 | 0 | 0 | 15 | 800.0 | 390.0 | 20.0 | |
| **3** | Apartamento | Higienópolis | 1 | 0 | 0 | 48 | 800.0 | 230.0 | 0.0 | |
| **4** | Apartamento | Cachambi | 2 | 0 | 0 | 50 | 1300.0 | 301.0 | 17.0 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **21821** | Apartamento | Méier | 2 | 0 | 0 | 70 | 900.0 | 490.0 | 48.0 | |
| **21822** | Quitinete | Centro | 0 | 0 | 0 | 27 | 800.0 | 350.0 | 25.0 | |
| **21823** | Apartamento | Jacarepaguá | 3 | 1 | 2 | 78 | 1800.0 | 800.0 | 40.0 | |
| **21824** | Apartamento | São Francisco Xavier | 2 | 1 | 0 | 48 | 1400.0 | 509.0 | 37.0 | |
| **21825** | Apartamento | Leblon | 2 | 0 | 0 | 70 | 3000.0 | 760.0 | 0.0 | |

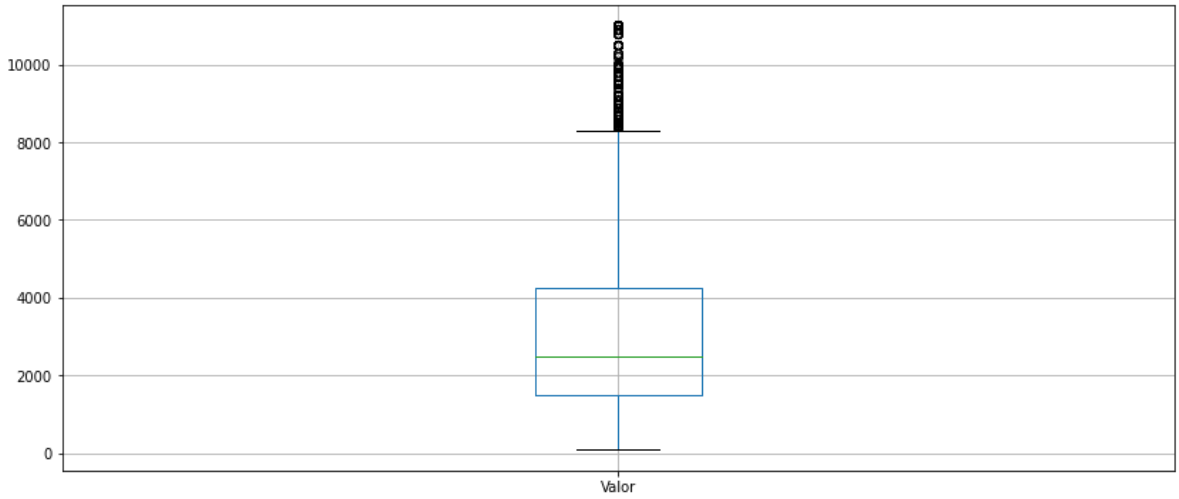
19814 rows × 11 columns

In [10]:

```
dados_new.boxplot(['Valor'])
```

Out[10]:

```
<AxesSubplot:>
```

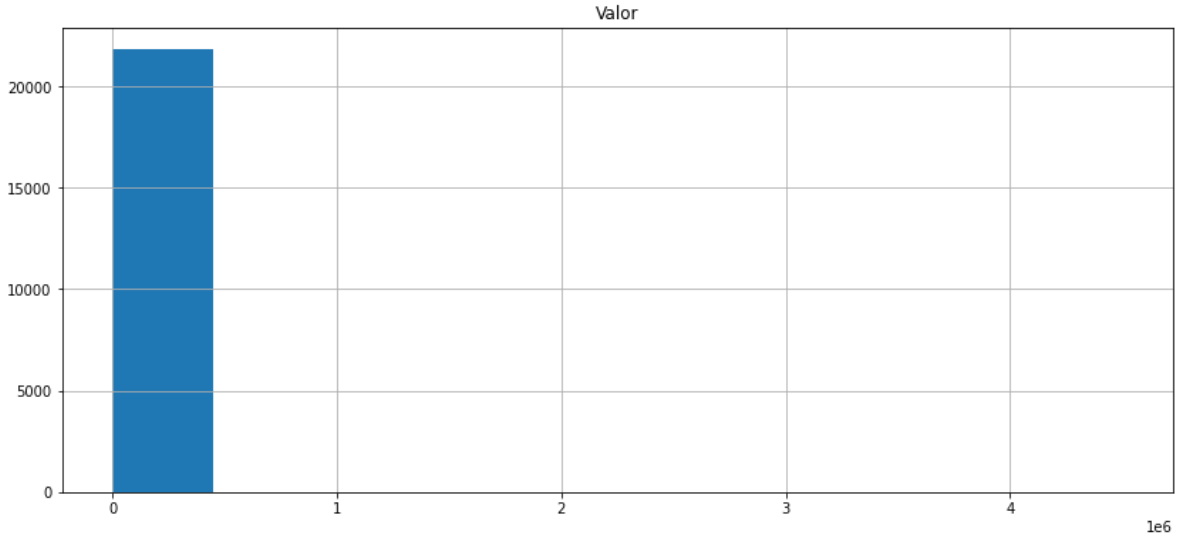


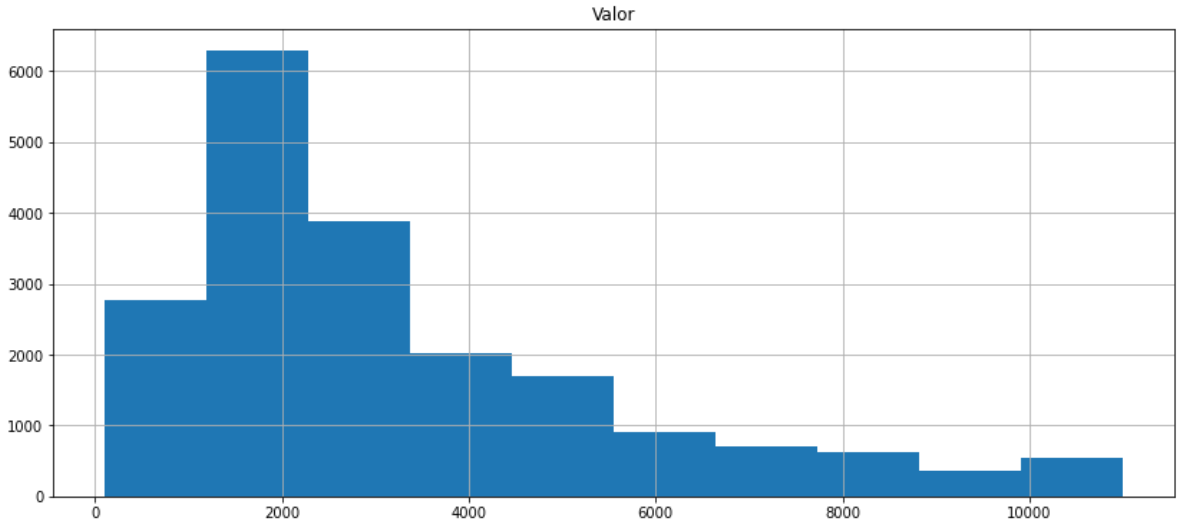
In [11]:

```
dados.hist(['Valor'])
dados_new.hist(['Valor'])
```

Out[11]:

```
array([[<AxesSubplot:title={'center':'Valor'}>]], dtype=object)
```



In [12]:

```
valor_m2 = dados['Valor_m2']
valor_m2.head(10)
```

Out[12]:

```
0    42.50
1    70.00
2    53.33
3    16.67
4    26.00
5    29.33
6    15.38
7    21.43
8    16.67
9    50.00
Name: Valor_m2, dtype: float64
```

In [13]:

```
Q1_m2 = valor_m2.quantile(.25)
Q3_m2 = valor_m2.quantile(.75)
IIQ_m2 = Q3_m2 - Q1_m2
limite_inferior_m2 = Q1_m2 - 1.5 * IIQ_m2
limite_superior_m2 = Q3_m2 + 1.5 * IIQ_m2
print(f" Q1:{Q1_m2} Q3:{Q3_m2} IQQ:{IIQ_m2} Limite Inferior:{limite_inferior_m2} Limite Sup
```

```
 Q1:21.12 Q3:42.0 IQQ:20.88 Limite Inferior:-10.2 Limite Superior: 73.32
```

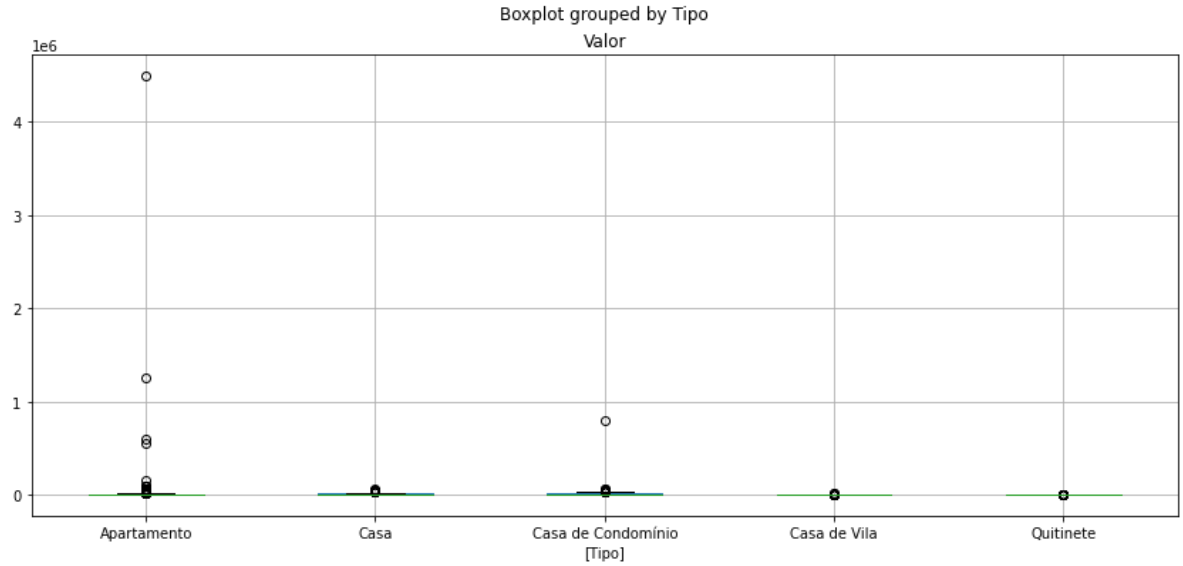
# Identificando e Removendo Outliers(Continuação)

In [14]:

```
dados.boxplot(['Valor'], by = ['Tipo'])
```

Out[14]:

```
<AxesSubplot:title={'center':'Valor'}, xlabel='[Tipo]'>
```



In [23]:

```
grupo_tipo = dados.groupby('Tipo')['Valor']
grupo_tipo
```

Out[23]:

```
<pandas.core.groupby.generic.SeriesGroupBy object at 0x0000012FD1C1A288>
```

In [24]:

```
grupo_tipo.groups
```

Out[24]:

```
{'Apartamento': [2, 3, 4, 7, 8, 9, 11, 13, 14, 15, 17, 18, 19, 20, 21, 23, 2
4, 25, 26, 27, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 43, 44, 4
5, 46, 47, 48, 49, 50, 51, 52, 53, 55, 56, 59, 60, 61, 62, 63, 64, 65, 66, 6
7, 68, 69, 70, 72, 73, 74, 75, 76, 77, 79, 80, 82, 83, 84, 85, 87, 88, 89, 9
0, 91, 92, 93, 94, 95, 97, 98, 99, 102, 103, 104, 105, 106, 107, 108, 109, 1
10, 111, 112, 113, 114, 115, 116, 117, 118, 119, ...], 'Casa': [1, 22, 54, 5
7, 96, 100, 144, 160, 180, 238, 250, 253, 264, 286, 310, 316, 334, 339, 357,
378, 385, 399, 424, 434, 474, 475, 481, 511, 542, 543, 567, 571, 572, 618, 6
21, 630, 635, 636, 660, 676, 700, 721, 724, 760, 771, 780, 856, 873, 937, 98
2, 1029, 1036, 1118, 1123, 1125, 1157, 1178, 1249, 1256, 1316, 1335, 1350, 1
371, 1412, 1426, 1430, 1440, 1445, 1472, 1475, 1488, 1586, 1604, 1656, 1662,
1666, 1671, 1684, 1709, 1717, 1762, 1810, 1835, 1875, 1905, 1933, 1942, 196
0, 2019, 2039, 2056, 2075, 2101, 2107, 2108, 2133, 2170, 2201, 2204, 2211,
...], 'Casa de Condomínio': [5, 6, 12, 16, 42, 58, 166, 168, 183, 207, 222,
246, 259, 265, 279, 291, 308, 336, 391, 401, 440, 445, 449, 502, 556, 609, 6
22, 657, 663, 673, 707, 723, 781, 807, 809, 863, 883, 887, 934, 958, 961, 97
9, 986, 992, 1004, 1008, 1028, 1082, 1095, 1112, 1129, 1148, 1158, 1182, 122
0, 1227, 1229, 1239, 1246, 1308, 1312, 1320, 1341, 1356, 1406, 1438, 1439, 1
467, 1495, 1531, 1560, 1582, 1601, 1615, 1646, 1713, 1722, 1728, 1756, 1764,
1770, 1802, 1860, 1880, 1883, 1899, 1938, 2031, 2033, 2071, 2152, 2168, 220
0, 2224, 2246, 2248, 2327, 2333, 2357, 2371, ...], 'Casa de Vila': [81, 212,
220, 303, 332, 697, 822, 844, 918, 1012, 1353, 1362, 1447, 1491, 1553, 1639,
1669, 1703, 1769, 2087, 2249, 2267, 2446, 2533, 2547, 2605, 2641, 2727, 284
0, 2872, 2977, 2984, 3017, 3025, 3300, 3426, 3523, 3703, 3823, 3855, 3858, 3
863, 4094, 4146, 4153, 4165, 4340, 4444, 4826, 5151, 5170, 5175, 5198, 5294,
5410, 5535, 5597, 5724, 5751, 5911, 5950, 5995, 6008, 6031, 6049, 6201, 623
6, 6300, 6348, 6402, 6429, 6754, 6795, 6939, 6957, 7033, 7091, 7146, 7296, 7
697, 7712, 7778, 7837, 7843, 7968, 8004, 8136, 8427, 8452, 8578, 9229, 9234,
9319, 9476, 9619, 9624, 9716, 9739, 9784, 9867, ...], 'Quitinete': [0, 10, 2
8, 71, 78, 86, 101, 120, 146, 174, 191, 206, 223, 248, 301, 314, 327, 344, 3
55, 425, 426, 427, 460, 486, 532, 633, 650, 680, 808, 870, 917, 919, 924, 92
8, 939, 944, 970, 1001, 1016, 1044, 1070, 1156, 1170, 1172, 1184, 1192, 119
6, 1212, 1217, 1261, 1274, 1334, 1351, 1360, 1393, 1404, 1407, 1483, 1496, 1
510, 1543, 1595, 1611, 1613, 1633, 1696, 1697, 1706, 1733, 1753, 1772, 1824,
1839, 1853, 1910, 2013, 2085, 2098, 2125, 2142, 2149, 2156, 2160, 2227, 223
7, 2239, 2258, 2272, 2326, 2362, 2382, 2383, 2384, 2394, 2445, 2457, 2462, 2
493, 2507, 2630, ...]}
```

In [25]:

```
Q1 = grupo_tipo.quantile(.25)
Q3 = grupo_tipo.quantile(.75)
IIQ = Q3 - Q1
limite_inferior = Q1 - 1.5 * IIQ
limite_superior = Q3 + 1.5 * IIQ
print(f" Q1:{Q1} Q3:{Q3} IQQ:{IIQ} Limite Inferior:{limite_inferior} Limite Superior: {limi
```

```
 Q1:Tipo
Apartamento           1700.0
Casa                  1100.0
Casa de Condomínio    4000.0
Casa de Vila           750.0
Quitinete              900.0
Name: Valor, dtype: float64 Q3:Tipo
Apartamento           5000.0
Casa                  9800.0
Casa de Condomínio    15250.0
Casa de Vila           1800.0
Quitinete              1500.0
Name: Valor, dtype: float64 IQQ:Tipo
Apartamento           3300.0
Casa                  8700.0
Casa de Condomínio    11250.0
Casa de Vila           1050.0
Quitinete               600.0
Name: Valor, dtype: float64 Limite Inferior:Tipo
Apartamento           -3250.0
Casa                 -11950.0
Casa de Condomínio   -12875.0
Casa de Vila           -825.0
Quitinete                 0.0
Name: Valor, dtype: float64 Limite Superior: Tipo
Apartamento           9950.0
Casa                  22850.0
Casa de Condomínio    32125.0
Casa de Vila           3375.0
Quitinete              2400.0
Name: Valor, dtype: float64
```

In [29]:

```
limite_superior['Casa']
```

Out[29]:

```
22850.0
```

In [34]:

```
dados_new = pd.DataFrame()

for tipo in grupo_tipo.groups.keys():
    eh_tipo = dados['Tipo'] == tipo
    eh_dentro_limite = (dados['Valor'] >= limite_inferior[tipo]) & (dados['Valor'] <= limit
    selecao = eh_tipo & eh_dentro_limite
    dados_selecao = dados[selecao]
    dados_new = pd.concat([dados_new, dados_selecao])
```

In [35]:

```
dados_new
```

Out[35]:

| | Tipo | Bairro | Quartos | Vagas | Suites | Area | Valor | Condominio | IPTU | Va |
|---|---|---|---|---|---|---|---|---|---|---|
| **2** | Apartamento | Centro | 1 | 0 | 0 | 15 | 800.0 | 390.0 | 20.0 | |
| **3** | Apartamento | Higienópolis | 1 | 0 | 0 | 48 | 800.0 | 230.0 | 0.0 | |
| **4** | Apartamento | Cachambi | 2 | 0 | 0 | 50 | 1300.0 | 301.0 | 17.0 | |
| **7** | Apartamento | Grajaú | 2 | 1 | 0 | 70 | 1500.0 | 642.0 | 74.0 | |
| **8** | Apartamento | Lins de Vasconcelos | 3 | 1 | 1 | 90 | 1500.0 | 455.0 | 14.0 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **21687** | Quitinete | Glória | 1 | 0 | 0 | 10 | 400.0 | 107.0 | 10.0 | |
| **21728** | Quitinete | Flamengo | 1 | 0 | 0 | 23 | 900.0 | 605.0 | 0.0 | |
| **21748** | Quitinete | Centro | 1 | 0 | 0 | 24 | 1100.0 | 323.0 | 0.0 | |
| **21815** | Quitinete | Copacabana | 1 | 0 | 0 | 22 | 1500.0 | 286.0 | 200.0 | |
| **21822** | Quitinete | Centro | 0 | 0 | 0 | 27 | 800.0 | 350.0 | 25.0 | |

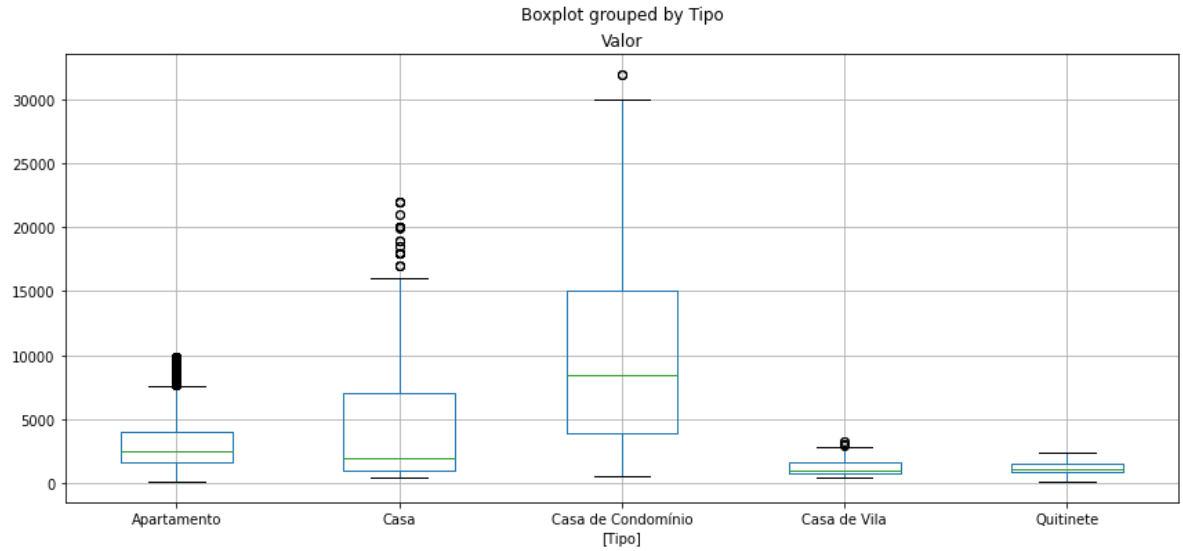
19831 rows × 11 columns

In [36]:

```
dados_new.boxplot(['Valor'], by = ['Tipo'])
```

Out[36]:

```
<AxesSubplot:title={'center':'Valor'}, xlabel='[Tipo]'>
```




In [37]:

```
dados_new.to_csv('dados/aluguel_residencial_sem_outliers.csv', sep=';', index=False)
```

**Um outlier ou uma anomalia, seria um valor atípico, ou seja, uma observação que se apresenta bastante distante dos demais valores da distribuição. Estes valores podem ser geradas por diversos fatores, uma forma mais comum é por conta de ruídos na coleta de dados ou erros de transformações.**