ノック7：各種統計量を把握しよう

データ分析を進めていく上で、まずは大きく2つの数字を知る必要があります。

1つ目は欠損している値の状況、2つ目は全体の数字感です。

欠損値は多くのデータに含まれる可能性があります。集計や機械学習に欠損値は大きく影響するので、除去や補間をする必要がありますので、数字を抑えておきましょう。

一般的にデータ分析では、商品毎、顧客属性毎など、様々な切り口で集計を行っていきます。

その際に、今月の商品Aが10万円だったとわかったとしても、全体の売上が10億円単位規模なのか100万円規模なのかによって意味が大きく違ってきます。そこで、全体の数字感を掴むのが重要となります。

Jupyter-Notebookのセルにそれぞれ分けて書きましょう。

```
join_data.isnull().sum()
```

```
join_data.describe()
```


■図1-7：各種統計量

ノック7：各種統計量を把握しよう

```
In [14]: join_data.isnull().sum()
Out[14]: detail_id          0
         transaction_id     0
         item_id            0
         quantity           0
         payment_date       0
         customer_id        0
         customer_name      0
         registration_date  0
         customer_name_kana 0
         email              0
         gender             0
         age                0
         birth              0
         pref               0
         item_name          0
         item_price         0
         price              0
         dtype: int64
```

```
In [15]: join_data.describe
```

Out[15]:

	detail_id	quantity	age	item_price	price
count	7144.000000	7144.000000	7144.000000	7144.000000	7144.000000
mean	3571.500000	1.199888	50.265677	121698.628219	135937.150056
std	2062.439494	0.513647	17.190314	64571.311830	68511.453297
min	0.000000	1.000000	20.000000	50000.000000	50000.000000
25%	1785.750000	1.000000	36.000000	50000.000000	85000.000000
50%	3571.500000	1.000000	50.000000	102500.000000	120000.000000
75%	5357.250000	1.000000	65.000000	187500.000000	210000.000000
max	7143.000000	4.000000	80.000000	210000.000000	420000.000000

1つ目のセルでは欠損値の数を出力しています。isnull()を用いると、欠損値をTrue/Falseで返してくれて、そのTrueの数をそれぞれの列毎にsum()で計算しています。今回のデータは綺麗なデータのため、欠損値はありませんでした。

2つ目のセルでは全体感を把握するための各種統計量を出力しています。

describe()を用いると、データ件数(count)、平均値(mean)、標準偏差(std)、最小値(min)、四分位数(25%、75%)、中央値(50%)、最大値(max)を簡単に出力できます。例えば、priceを見ると、平均は135937円となっています。最高金額は420000円となっており、これは単価210000円のPCを2台買ったユーザーがいると想像できます。また、quantityを見ると、最高は4ですが、75%数でも1なので、ほとんどの顧客が数量1で購入していることがわかります。さらに、ageを見ると、20歳から80歳までの範囲の顧客像が見えてきます。