■図3-1：データの読み込み結果

ノック21：データを読み込んでみよう

```
In [1]: import pandas as pd
        uselog = pd.read_csv('use_log.csv')
        print(len(uselog))
        uselog.head()
```

197428

Out[1]:

	log_id	customer_id	usedate
0	L00000049012330	AS009373	2018-04-01
1	L00000049012331	AS015315	2018-04-01
2	L00000049012332	AS040841	2018-04-01
3	L00000049012333	AS046594	2018-04-01
4	L00000049012334	AS073285	2018-04-01

```
In [2]: customer = pd.read_csv('customer_master.csv')
        print(len(customer))
        customer.head()
```

4192

Out[2]:

	customer_id	name	class	gender	start_date	end_date	campaign_id	is_deleted
0	OA832399	XXXX	C01	F	2015-05-01 00:00:00	NaN	CA1	0
1	PL270116	XXXXX	C01	M	2015-05-01 00:00:00	NaN	CA1	0
2	OA974876	XXXXX	C01	M	2015-05-01 00:00:00	NaN	CA1	0
3	HD024127	XXXXX	C01	F	2015-05-01 00:00:00	NaN	CA1	0
4	HD661448	XXXXX	C03	F	2015-05-01 00:00:00	NaN	CA1	0

```
In [5]: class_master = pd.read_csv('class_master.csv')
        print(len(class_master))
        class_master.head()
```

3

Out[5]:

	class	class_name	price
0	C01	オールタイム	10500
1	C02	デイタイム	7500
2	C03	ナイト	6000

```
In [6]: campaign_master = pd.read_csv('campaign_master.csv')
        print(len(campaign_master))
        campaign_master.head()
```

3

Out[6]:

	campaign_id	campaign_name
0	CA1	通常
1	CA2	入会費半額
2	CA3	入会費無料

実行すると、それぞれのデータの先頭5行のデータが確認できます。

第1部でも述べましたが、最初は先頭数行を表示させ、どのようなデータ列が存在するのか、それぞれのデータ列の関係性など、データの大枠を掴むことが重要です。また、今回はデータ件数を把握するために、len()を用いてデータ件数の表示も行っています。

利用履歴であるuse_log.csvを読み込んだuselogは、顧客ID、利用日を含んだ3列のみのシンプルなデータであることがわかります。これは、どの顧客がいつジムを利用したのかがわかるデータとなっています。件数は、197428件と縦に長いデータとなっていることがわかります。

次に、会員データのcustomer_master.csvを読み込んだcustomerには、顧客ID、名前、会員クラス、性別、登録日等の情報が含まれていることがわかります。名前は、マスキングされており、名前だけで個人が特定できないようになっています。また、is_deleted列は、2019年3月時点で退会しているユーザーをシステム的に素早く検索するための列となります。これらから、uselogのcustomer_idと紐付けが可能であることがわかります。会員データのデータ件数は、既に退会済みのユーザーも含めて4192人となっていることがわかります。

会員区分、キャンペーン区分データは、それぞれの区分をデータに含んでおり、それぞれ、class、campaign_id列を用いると、会員データと結合できることがわかります。

次は、分析のためのデータ加工に進んでいきますが、そのためには、主とするデータを考える必要があります。

分析の目的によって、主とするデータは違ってきますが、この章のケースでは、どのデータを主とするべきでしょうか。

ここで考えられるのは、顧客データであるcustomerと利用履歴データであるuselogです。

まずは、データ数も少ないので、顧客データを主に考えてみましょう。

後半で、利用履歴データ（uselog）を主とした分析も行っていきます。

まず、利用履歴データは一旦無視して、顧客データを整形し、どのような顧客が何人くらいいるのか等の全体像を掴んでいきましょう。