レシピ62 連続変数でグループ分け（P176）

- pandasでグループ分けを行う場合、通常は**離散的で重複値のあるカラムを使う**
  - 重複値がなければ、1グループ1行でグループ分けの意味がないため
- 一方、**連続数値になっているカラム**は通常グループ分けに使われないことが多いが
- **ビニング処理（ビン分割）**を行うケースではグループ分けを利用することもある

- ビニング･･･連続値を任意の境界で区切り、カテゴリ分けして離散値に変換する処理のこと
  - 例）年齢データを10代、20代の層ごとに分ける、など
  - `cut()` や `qcut()`

- このレシピでは、flightsデータセットを探索して、航空会社の飛行距離の分布を調べる
- 例）500~1000マイルを飛行する便が最も多い航空会社がどこか？など

（1）flightsデータセットを読み込む

In [177…
```
flights = pd.read_csv('flights.csv')
flights.head()
```

Out[177…

	MONTH	DAY	WEEKDAY	AIRLINE	ORG_AIR	DEST_AIR	SCHED_DEP	DEP_DELAY	AIR_TIME	DIST	SCHED_ARR	ARR_DELAY	DIVERTED	CANCELLED
0	1	1	4	WN	LAX	SLC	1625	58.0	94.0	590	1905	65.0	0	0
1	1	1	4	UA	DEN	IAD	823	7.0	154.0	1452	1333	-13.0	0	0
2	1	1	4	MQ	DFW	VPS	1305	36.0	85.0	641	1453	35.0	0	0
3	1	1	4	AA	DFW	DCA	1555	7.0	126.0	1192	1935	-7.0	0	0
4	1	1	4	WN	LAX	MCI	1720	48.0	166.0	1363	2225	39.0	0	0

（2）

- 航空会社の飛行範囲の距離を調べたい。'DIST'カラムを利用
- `cut` を使い、データを5つのビンに分割する
- 5つのビンに分ける場合、ビンは、6つ必要（左右のnp.inf2つ）
- `cut` は5つの順序付きカテゴリのSeriesを戻す

In [178…
```
# np.inf：無限大（負・正）
bins = [-np.inf, 200, 500, 1000, 2000, np.inf]

cuts = pd.cut(flights['DIST'], bins=bins)
cuts.head()

# 下の結果が解釈しにくい...Categories～の行はわかる。
```

Out[178…
```
0     (500.0, 1000.0]
1    (1000.0, 2000.0]
2     (500.0, 1000.0]
3    (1000.0, 2000.0]
4    (1000.0, 2000.0]
Name: DIST, dtype: category
Categories (5, interval[float64]): [(-inf, 200.0] < (200.0, 500.0] < (500.0, 1000.0] < (1000.0, 2000.0] < (2000.0, inf]]
```

（3）順次付きのカテゴリのSeriesが作られた。各カテゴリの値を数える

- 並びがごちゃごちゃなのか。

In [179…
```
cuts.value_counts()
```

Out[179…
```
(500.0, 1000.0]     20659
(200.0, 500.0]      15874
(1000.0, 2000.0]    14186
(2000.0, inf]        4054
(-inf, 200.0]        3719
Name: DIST, dtype: int64
```

（4）

- Seriesのcutsをグループ分けに使う
- pandasでは、自由にグループ分けが可能
- cutsをgroupbyに渡し、'AIRLINE'カラムでvalue_countsメソッドを呼び出し、各距離グループの分布を把握する
- 200マイル以下でSkyWest(OO）が33%近くを占め1位だが、200~500マイルでは16%（3位）しかないことが読み取れる

In [180…
```
flights.groupby(cuts)['AIRLINE'].value_counts(normalize=True)\
                                .round(3).head(15)
```

Out[180…
```
DIST            AIRLINE
(-inf, 200.0]   OO         0.326
                EV         0.289
                MQ         0.211
                DL         0.086
                AA         0.052
                UA         0.027
                WN         0.009
(200.0, 500.0]  WN         0.194
                DL         0.189
                OO         0.159
                EV         0.156
                MQ         0.100
                AA         0.071
                UA         0.062
                VX         0.028
Name: AIRLINE, dtype: float64
```

（補足）割愛（P178）