1_groupby_hierarchical_index

May 16, 2019

In [ ]:
```
import pandas as pd
```

Group by - Basic

In [ ]:
```
# data from:
ipl_data = {'Team': ['Riders', 'Riders', 'Devils', 'Devils', 'Kings',
         'kings', 'Kings', 'Kings', 'Riders', 'Royals', 'Royals', 'Riders'],
         'Rank': [1, 2, 2, 3, 3,4 ,1 ,1,2 , 4,1,2],
         'Year': [2014,2015,2014,2015,2014,2015,2016,2017,2016,2014,2015,2017],
         'Points':[876,789,863,673,741,812,756,788,694,701,804,690]}

df = pd.DataFrame(ipl_data)
df
```

In [ ]:
```
df.groupby("Team")["Points"].sum()
```

0.0.1 Hierarchical index

In [ ]:
```
df
```

In [ ]:
```
h_index = df.groupby(["Team", "Year"])["Points"].sum()
h_index
```

In [ ]:
```
h_index.index
```

In [ ]:
```
h_index["Devils":"Kings"]
```

In [ ]:
```
h_index.unstack()
```

In [ ]:
```
h_index.swaplevel()
```

In [ ]:
```
h_index.swaplevel().sortlevel(0)
```

In [ ]:
```
h_index.head()
```

In [ ]:
```
h_index.sum(level=0)
```

In [ ]:
```
h_index.sum(level=1)
```

0.0.2 Groupby - gropuped

```
In [ ]: grouped = df.groupby("Team")
        grouped

In [ ]: for name,group in grouped:
            print (name)
            print (group)

In [ ]: for name,group in grouped:
            print (type(name))
            print (type(group))

In [ ]: grouped.get_group("Riders")
```

0.0.3 Aggregation

```
In [ ]: df

In [ ]: grouped.agg(min)

In [ ]: import numpy as np
        grouped.agg(np.mean)

In [ ]: grouped['Points'].agg([np.sum, np.mean, np.std])
```

0.0.4 Transofrmation

$$z_i = \frac{x_i - \mu}{\sigma}$$

```
In [ ]: score = lambda x: (x - x.mean()) / x.std()
        grouped.transform(score)

In [ ]: df.groupby('Team').filter(lambda x: len(x) >= 3)

In [ ]: df.groupby('Team').filter(lambda x: x["Points"].max() > 800)

In [ ]: # !wget https://www.shanelynn.ie/wp-content/uploads/2015/06/phone_data.csv

In [ ]: df_phone = pd.read_csv("./data/phone_data.csv")
        df_phone.head()

In [ ]: df_phone.info()

In [ ]: import dateutil
        df_phone['date'] = df_phone['date'].apply(dateutil.parser.parse, dayfirst=True)
        df_phone.head()

In [ ]: df_phone.info()

In [ ]: df_phone.groupby('month')['duration'].sum()
```

```
In [ ]: df_phone[df_phone['item'] == 'call'].groupby('month')['duration'].sum()

In [ ]: df_phone.groupby(['month', 'item'])['duration'].sum()

In [ ]: df_phone.groupby(['month', 'item'])['date'].count()

In [ ]: df_phone.groupby(['month', 'item'])['date'].count().unstack()

In [ ]: df_phone.groupby('month', as_index=False).agg({"duration": "sum"})

In [ ]: df_phone.groupby(['month', 'item']).agg({'duration':sum,      # find the sum of the du
                                                'network_type': "count", # find the number of net
                                                'date': 'first'})    # get the first date per gro

In [ ]: df_phone.groupby(['month', 'item']).agg({'duration': [min],      # find the min, max,
                                                'network_type': "count", # find the number of net
                                                'date': [min, 'first', 'nunique']})    # get the

In [ ]: grouped = df_phone.groupby('month').agg( {"duration" : [min, max, np.mean]})
        grouped

In [ ]: grouped.columns = grouped.columns.droplevel(level=0)
        grouped

In [ ]: grouped.rename(columns={"min": "min_duration", "max": "max_duration", "mean": "mean_du

In [ ]: grouped = df_phone.groupby('month').agg( {"duration" : [min, max, np.mean]})
        grouped

In [ ]: grouped.columns = grouped.columns.droplevel(level=0)
        grouped

In [ ]: grouped.add_prefix("duration_")

In [ ]: df_phone

In [ ]:
```