实验2：python数据应用基础操作

廖　军
liaojun@cqu.edu.cn
重庆大学大数据与软件学院

熟悉python的基本数据应用操作

Sorting

edu.sort_values(by = ‘Values’, ascending = False, inplace = True)

Sorted in descending order by the values in the Value column

edu.sort_index(axis=0, ascending = True, inplace = True)

That will return to the original order, i.e. ordered by the index

Z-score transformation

```
def zscoreScaling(data):
    return (data-data.mean())/data.std()

df3 = df2.apply(zscoreScaling)
```

matplotlib.pyplot在显示时无法找到合适的字体，显示乱码
解决方法：添加相应字体包

```
from matplotlib.font_manager import FontProperties
font = FontProperties(fname=r"c:\windows\fonts\simsun.ttc",
size=14)
.....
plt.xlabel(u'x值', fontproperties=font) # x轴名称
plt.ylabel(u'y值', fontproperties=font) # y轴名称
```

How to Find the Correlation coefficient?

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

rxy – the correlation coefficient of the linear relationship between the variables x and y
xi – the values of the x-variable in a sample
$\bar{x}$ – the mean of the values of the x-variable
yi – the values of the y-variable in a sample
$\bar{y}$ – the mean of the values of the y-variable

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

首先计算出均值和方差，检验正态性：

```
u1,u2 = df['Dribbling'].mean(),df['BallControl'].mean()  # 计算均值
std1,std2 = df['Dribbling'].std(),df['BallControl'].std()  # 计算标准差
print('Dribbling正态性检验：\n',stats.kstest(df['Dribbling'], 'norm', (u1, std1)))
print('BallControl正态性检验：\n',stats.kstest(df['BallControl'], 'norm', (u2, std2)))
```

然后再计算每个算式，如

```
df['(x-u1)*(y-u2)'] = (df['Dribbling'] - u1) * (df['BallControl'] - u2)
```

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

pandas相关性方法：（method默认pearson）

data.corr(method='pearson')