# 1.1.x知识点总结

## 1、数据的加载和采集

**读取数据集**
data = pd.read_csv('文件名.csv')

## 2、数据的统计和清洗

### 新建列

**按条件新建列**
data['新列列名'] = np.where(条件, 条件为真时的值, 条件为假时的值)

**新建划分区间列**
data['新列列名'] = pd.cut(data['划分依据列'],区间边界bins, 区间标签labels, 区间开闭)

**新建布尔值标记列**
1、可以按条件新建列，设置条件为真的值为True，条件为假的值为False

2、data['新列列名'] = data['判断依据列'].between(需要大于的值，需要小于的值)

3、data['新列列名'] = data['判断依据列'] > 需要大于的值

4、data['新列列名'] = data['判断依据列'] < 需要小于的值

### 数据统计

**统计指定列各个值的数量**
data['列名'].value_counts()

**统计所有行的数量**
len(data)

**分组统计**

**按指定列分组之后统计另一列对应值**
data.groupby(data['分组依据列'])['待统计列'].统计方法()

**.统计方法()**

**取平均值**
.mean()

**取数量**
.count()

**取平均值和数量**
.agg(['mean','count'])

**按指定列分组之后统计其他多列对应值**
data.groupby(data['分组依据列']).agg({'待统计列1':'统计方法','待统计列2':'统计方法'})

**数据行筛选**
data[筛选条件]

**筛选出指定列里指定值所在的所有行**
data[data['指定列'].isin(['指定值'])]

**筛选出指定列的值满足指定条件的所有行**
1、data[data['指定列'].between(需要大于的值，需要小于的值)]

2、data[data['指定列']>需要大于的值]

3、data[data['指定列']<需要小于的值]

### 异常值处理

**填充缺失值**
data['待填充列'].fillna(method='填充方法', inplace = True)

**异常值统计**

**缺失值统计**
data.isna().sum()

**重复值统计**
data.duplicated().sum()

**异常值标记**

**新建布尔值标记列，标记列里的值均为False和True**

**删除缺失值所在行**
data.dropna()

### 数据标准化

**数据类型统一**

**数据类型转换**
data['待转换的列'] = data['带转换的列'].astype(目标数据类型)

**标准化公式**
(data['待标准化列']-data['待标准化列'].mean())/data['待标准化列'].std()

### 删除列

**删除列**
data.drop(columns=['待删除列名'])

## 3、数据的保存

**数据保存**
待保存数据.to_csv('保存文件名.csv', index = False)