任务1.1：

import pandas as pd

import os

import numpy as np

import matplotlib.pyplot as plt

os.chdir(r'G:\数据')

order = pd.read\_csv('data2.csv', engine = 'python', encoding = 'gbk')# 数据读取

#查看数据形状

print('消费记录表数据形状为：', '\n', order.shape)

#查看数据类型

print('消费记录表数据的类型为：', '\n', order.dtypes)

#查缺失值

print('缺失值数量:\n', order.isnull().sum())

data=order.drop(['TermSerNo','conOperNo'],axis=1)#删除缺失值所在的列

print('去除缺失值之后： ',data.shape)

data.describe()#描述统计

data['Date']=pd.to\_datetime(data['Date'])#转换数据类型

#检测并处理重复值

print('去重前数据的形状为：',data.shape)

#直接对数据表中重复的数据去除

data.drop\_duplicates(inplace = True)

print('数据去重后的形状为：',data.shape)

plt.boxplot(data['Money'].values,whis=3) #消费金额箱线图

plt.show()

data['year']=[i.year for i in data['Date']] # 提取年份

data['month']=[i.month for i in data['Date']] # 提取月份

data['weekday'] = data['Date'].apply(lambda x: x.weekday()+1) # 提取星期

data['day']=[i.day for i in data['Date']] # 提取天

data['hour'] = data['Date'].dt.hour # 提取时

data['minute'] = data['Date'].dt.minute # 提取分

data['second'] = data['Date'].dt.second # 提取秒

data= data.iloc[data['hour'].values>5,[0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15]]#删除0-5点的消费记录

data.to\_csv('task1\_X1.csv',index = False, encoding = 'gbk')#保存文件