小作业(2) 数据预处理

本次作业含两道题目，分别为7分和3分，均为必做题。

作业1：爬取并存储链家的新房数据，并进行预处理（本题7分）

（1）爬取起始网页：https://bj.fang.lianjia.com/loupan/

（2）爬取信息的提取及存储要求（单条数据示例在第4页）

- 信息以csv文件存储，应包括以下字段：名称，类别，地理位置（3个字段分别存储），房型（只保留最小房型），面积（区间取中值并取整），均价（元，整数），总价（万元，区间取中值并取整）
- 对于所有字符串字段，要求去掉所有的前后空格
- 删除面积缺失的房屋数据

（3）数据统计

- 找出总价最贵和最便宜的房子，以及总价的中位数
- 找出均价最贵和最便宜的房子，以及均价的中位数

作业1：爬取并存储链家的新房数据，并进行预处理（本题7分）

（4）异常值处理

- 列出总价在均值三倍标准差以外的房屋，展示其基本信息（如果太多可以只展示一部分），并分析其原因（找4条数据即可）
- 通过箱型图原则判断并列出均价为异常值的房屋，展示其基本信息（如果太多可以只展示一部分），并分析其原因（找4条数据即可）

（5）离散化处理

- 对房屋的均价进行离散化处理，自行设定每个区间的长度并给出设置的理由，给出每个区间的房屋数量和所占比例

水岸壹号 别墅 在售

房山 / 良乡 / 良乡大学城西站地铁南侧800米，刺猬河旁

3室 / 4室

建面 185-199㎡

50000 元/㎡(均价)

总价1100-1300(万/套)

新房顾问：张红松 沟通

绿金案场 车位充足 人车分流 国央企

建邦•顺颐府 别墅 在售

顺义 / 后沙峪 / 空港B区裕民大街30号

3室

建面 270㎡

55583 元/㎡(均价)

总价1300(万/套)

新房顾问：曹莉 沟通

3天无理由退房 绿金案场 限竞房 低总价

名称，类别，地理位置（3个字段分别存储），房型（只保留最小房型），面积（区间取中值并取整），均价（元，整数），总价（万元，区间取中值并取整），结果示例如下：

水岸壹号,别墅,房山,良乡,良乡大学城西站地铁南侧800米， 刺猬河旁,3室,192,50000,1200

建邦·顺颐府,别墅,顺义,后沙峪, 空港B区裕民大街30号,270,55583,1300

作业2：分析处理2015年北京市PM2.5指数数据集空值（本题3分）

（1）原始数据集：BeijingPM20100101_20151231.csv（列信息见第6页说明）

（2）数据抽取及存储：从原始数据集中抽取2015年度数据，存储为新的csv文件

（3）找出空值：对新的csv文件，找出存在的空值列及相应的空值数量

（4）空值处理方法：对所有存在空值的列，给出空值的处理方法及理由，要求处理方法必须可在本数据集范围内执行

（5）空值处理并存储：按照自己的处理方法，通过pandas、numpy或python方法对空值进行处理，完成后给出新的空值列信息，并将处理后的数据（不涉及空值的列应原样保留）存储为新的csv文件

作业2：分析处理2015年北京市PM2.5指数数据集空值（本题3分）

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
No	year	month	day	hour	season	PM_Dongs	PM_Dongs	PM_Nongz	PM_US Pos	DEWP	HUMI	PRES	TEMP	cbwd	Iws	precipitatic	Iprec
1	2010	1	1	0	4	NA	NA	NA	NA	-21	43	1021	-11	NW	1.79	0	0
2	2010	1	1	1	4	NA	NA	NA	NA	-21	47	1020	-12	NW	4.92	0	0
3	2010	1	1	2	4	NA	NA	NA	NA	-21	43	1019	-11	NW	6.71	0	0
4	2010	1	1	3	4	NA	NA	NA	NA	-21	55	1019	-14	NW	9.84	0	0
5	2010	1	1	4	4	NA	NA	NA	NA	-20	51	1018	-12	NW	12.97	0	0
6	2010	1	1	5	4	NA	NA	NA	NA	-19	47	1017	-10	NW	16.1	0	0
7	2010	1	1	6	4	NA	NA	NA	NA	-19	44	1017	-9	NW	19.23	0	0
8	2010	1	1	7	4	NA	NA	NA	NA	-19	44	1017	-9	NW	21.02	0	0
9	2010	1	1	8	4	NA	NA	NA	NA	-19	44	1017	-9	NW	24.15	0	0
10	2010	1	1	9	4	NA	NA	NA	NA	-20	37	1017	-8	NW	27.28	0	0
11	2010	1	1	10	4	NA	NA	NA	NA	-19	37	1017	-7	NW	31.3	0	0

- No: 行号
- year: 年份
- month: 月份
- day: 日期
- hour: 小时
- season: 季节
- PM_xx: PM2.5浓度 (ug/m^3)(地点： xx)
- DEWP: 露点 (摄氏温度) 指在固定气压之下，空气中所含的气态水达到饱和而凝结成液态水所需要降至的温度
- HUMI: 湿度 (%)
- PRES: 气压 (hPa)c
- TEMP: Temperature (摄氏温度)
- cbwd: 组合风向
- Iws: 累计风速 (m/s)
- precipitation: 降水量/时 (mm)
- Iprec: 累计降水量 (mm) mm)

作业提交方式及要求

作业要求以报告形式在北邮教学云平台https://ucloud.bupt.edu.cn/ 上提交。

报告中除题目要求的各项统计分析数据/信息外，还需给出题目中要求存储的csv内容（每个csv截取至少50条数据即可，其中作业2应分别截取包含空值数据及进行了空值处理的数据），并将核心代码贴在报告中。核心代码应清楚展示题目中要求进行的处理。

文件格式要求为pdf，文件名要求为“<学号>_小作业2”，
例如：2021211353_小作业2

作业提交截止时间：2023.12.24 24点