第 14 期 Datawhale 组队学习

Pandas 教程(下)综合练习

一、端午节的淘宝粽子交易

问题

- (1) 请删除最后一列为缺失值的行,并求所有在杭州发货的商品单价均值。
- (2) 商品标题带有"嘉兴"但发货地却不在嘉兴的商品有多少条记录?
- (3) 请按照分位数将价格分为"高、较高、中、较低、低"5个类别,再将类别结果插入到标题一列之后,最后对类别列进行降序排序。
- (4) 付款人数一栏有缺失值吗? 若有则请利用上一问的分类结果对这些缺失 值进行合理估计并填充。
- (5) 请将数据后四列合并为如下格式的 Series: 商品发货地为 ××, 店铺为 ××, 共计 ×× 人付款,单价为 ××。
- (6) 请将上一问中的结果恢复成原来的四列。

二、墨尔本每日最低温度

问题

- (1) 剔除国庆节、五一劳动节和每月第一个周一,求每月的平均最低气温。
- (2) 季节指数是一种对于周期性变化序列的特征刻画。记数据集中第 k 年 平均最低气温为 $TY_k(k=1,...,10)$,第 k 年第 j 个月平均最低气温为 $TM_{kj}(j=1,...,12)$,定义 $S_j = \frac{\sum_k TM_{kj}}{\sum_k TY_k}$ 。请按照如上定义,计算 12 个月的季节指数 S_i 。

(3) 移动平均法是一种时间序列的常见平滑方式,可分为 k 期移动平均和 k 期中心移动平均,都使用了某一时刻及其周围的数据对该时刻的数据进行平滑修正。设原序列为 $x_1,...,x_n$, 对于 x_t 的 k 期移动平均修正 $\widetilde{x_t}$ 为 $\frac{\sum_{i=0}^{k-1} x_{t-i}}{x_t}$, 对于 k 期中心移动平均修正为

$$\widetilde{x_t} = \begin{cases} \frac{1}{k} (\frac{1}{2} x_{t-\frac{k}{2}} + x_{t-\frac{k}{2}+1} + \ldots + x_t + \ldots + x_{t+\frac{k}{2}-1} + \frac{1}{2} x_{t+\frac{k}{2}}), k \ is \ even \\ \frac{1}{k} (x_{t-\frac{k-1}{2}} + x_{t-\frac{k-1}{2}+1} + \ldots + x_t + \ldots + x_{t+\frac{k-1}{2}-1} + x_{t+\frac{k-1}{2}}) \ , else \end{cases}$$

- (a) 求原序列的 5 期移动平均序列。
- (b) 求原序列的 5 期与 6 期中心移动平均序列。

三、2016 年 8 月上海市摩拜单车骑行记录

问题

- (1) 平均而言,周末单天用车量比工作日单天用车量更大吗?
- (2) 工作日每天的高峰时间段大致为上午 7:30 至 9:30、下午 17:00 至 19:00,请问 8 月里早高峰骑行记录量(以 start_time 为准)高于晚高峰的有几天?
- (3) 请给出在所有周五中(以 start_time 为准),记录条数最多的那个周五 所在的日期,并在该天内分别按 30 分钟、2 小时、6 小时统计摩拜单车 使用时间的均值。
- (4) 请自行搜索相关代码或调用库, 计算每条记录起点到终点的球面距离。
- (5) 摩拜单车的骑行结束时间是以电子锁关闭的记录时间为准,但有时候用户会忘记关锁,导致骑行时间出现异常。同时,正常人的骑行速度往往大致落在一个合理的区间,请结合上一问中的球面距离和骑行起始、结束时间,找出潜在的异常骑行记录。
- (6) 由于路线的曲折性,起点到终点的球面距离往往不能充分反应行程长度,请利用 track 列的路线坐标数据,计算估计实际骑行距离,并重新 仿照上一问的方法找出可能的异常记录。