**5/22 1组进度报告**

**原始数据处理**

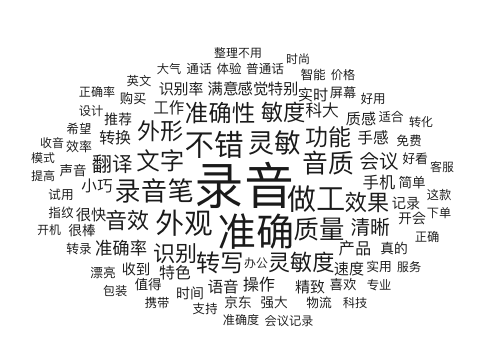
本周，我们在对原始的站内天猫和京东广告数据，进行了格式的统一化，并引入了哑变量来标识不同类别的广告，形成了方便统一使用数据名称、数据格式以及数据的排布等，并且将所需要的因变量全部进行了计算输出，后续可以直接利用这些数据来进行回归建模之类的。

**评论数据**

通过分析销售出的产品在京东平台上的评论数据，得出以下发现：

首先，判断评价性质（好评/差评）。从星级上看，五星好评率为100%；对评论文本做情感分析，得出评论情感值全部为正向，星级与实际评价一致性很高，评论数据较为可靠地反映出真实的用户体验。

随后，对评论文本内容进行分析。将整合后的评论中常出现的词做成词云，词语显示大小由其在评论中的出现频次数决定，得到下图：



从图中可以看出用户对录音笔产品的“敏度”，“准确性”，“做工”，“音质”，“转写”，“识别”等品质方面关注度较高，并给出了“不错”，“灵敏”“准确”，“清晰”等对应评价，可以看出产品本身的高识别敏度，高转写准确度和高录音音质为用户满意度高的关键点。

词云中出现了“录音”，“录音笔”等因与主题相关度过高导致频数大但并不含有效信息的词汇。为消除此影响，我们将其删除。

计算单个词汇的tf-idf值【1】，并以此为基准，重新制作词云，展示评论中重要性高的字段：

1. TF-IDF是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。



从新生成的词云中能够更明显地看出，用户的评价集中在“转写”，“准确性”，“质量”，“做工”，“外观“，“效果“，“功能”，“灵敏度”，“文字”，“方便”，等字眼。

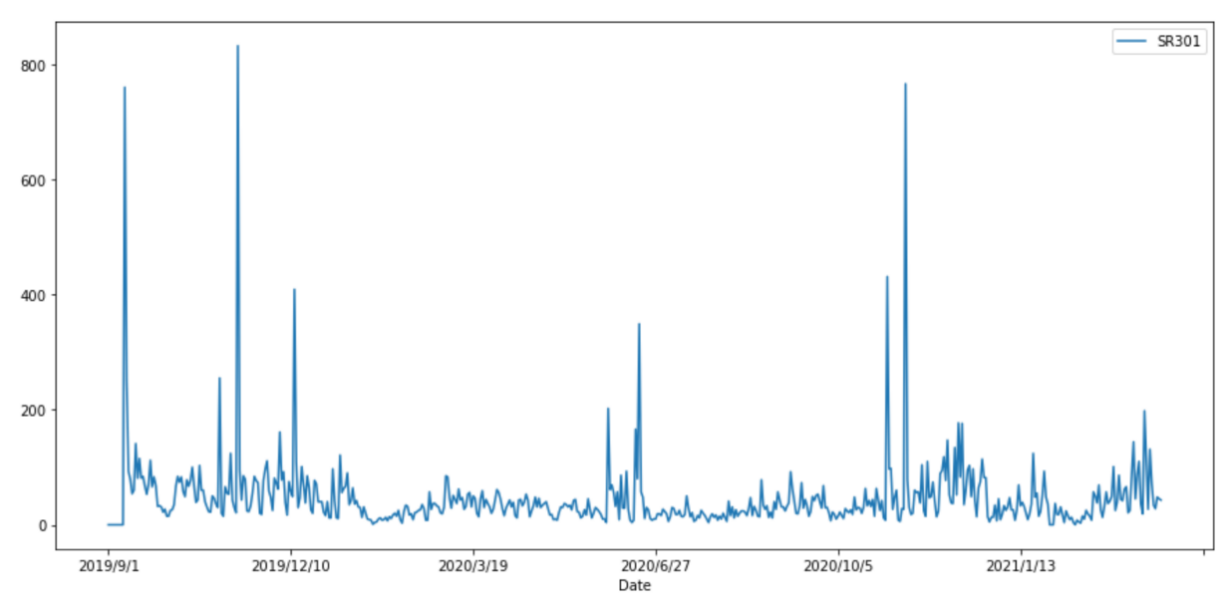
**产品周期**

对于单个产品SR301 19-21的销量数据进行了：拐点识别与匹配

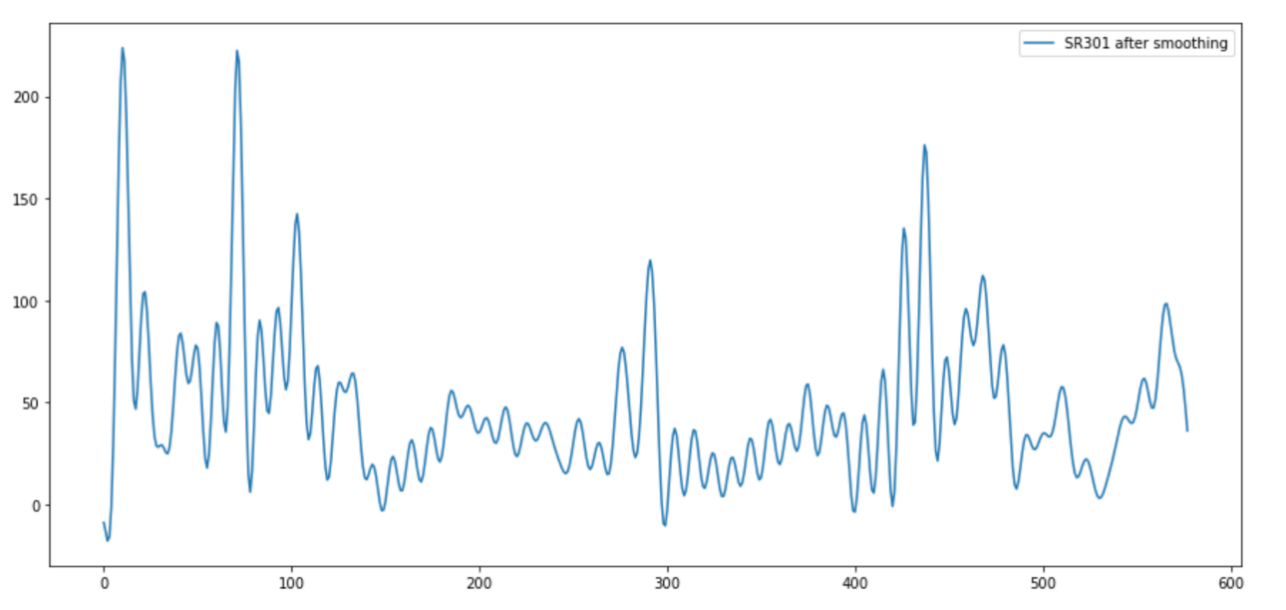
为了更好的研究产品的生命周期，销量的拐点数据相对于其他时点数据更加重要，能够更好的表现出周期的更迭。在判断拐点时，需要使用平滑后的时间序列数据，否则拐点识别会受到噪声的影响，识别到异常波动点。

主要分为以下步骤：

1. 生成原始数据折线图，尝试不同的滤波方法，尽可能保留数据本身的特性，过滤掉部分噪声。

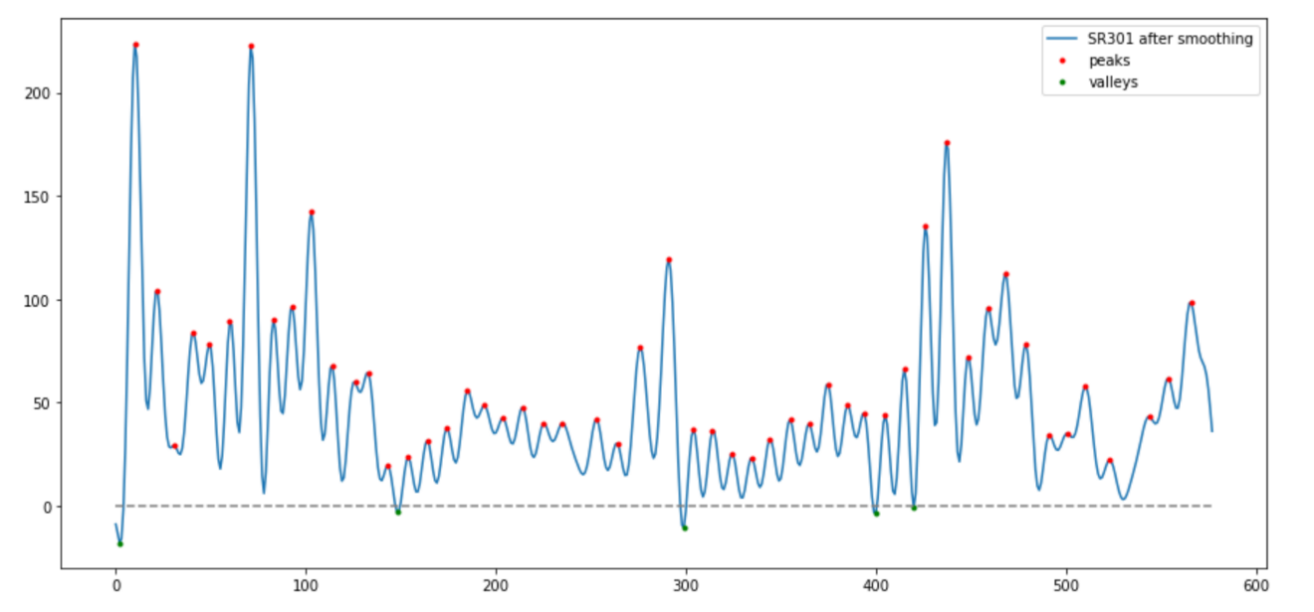


原始数据图



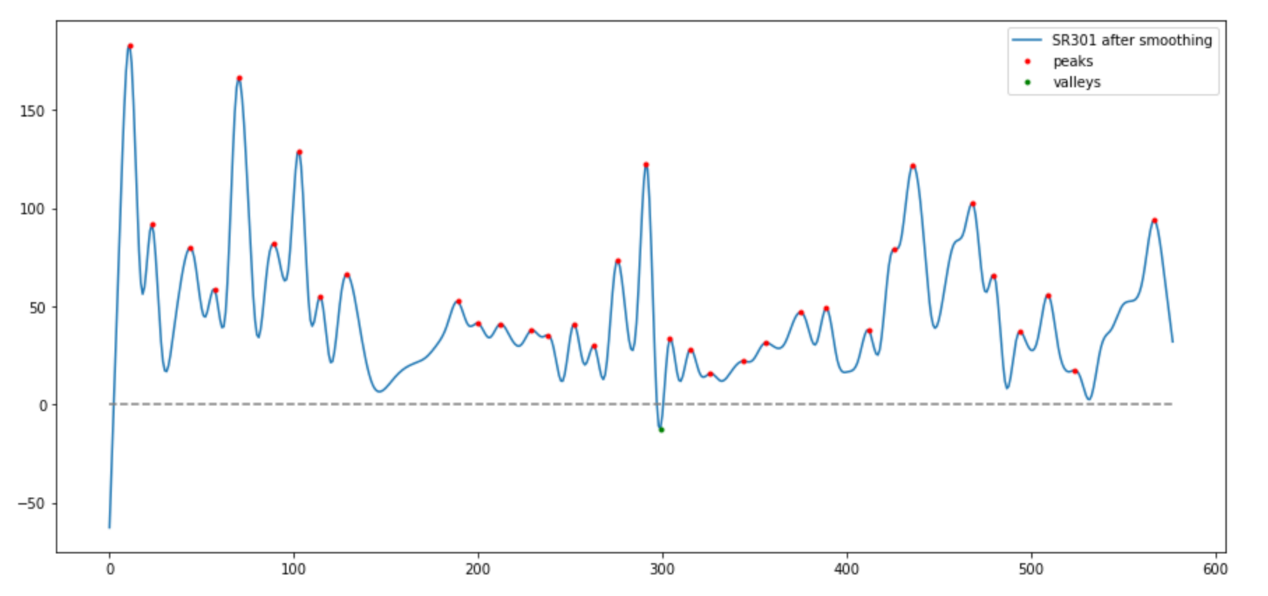
筛选后使用的滤波数据图

1. 识别图中局部极大（小）值点，作为拐点的候选。



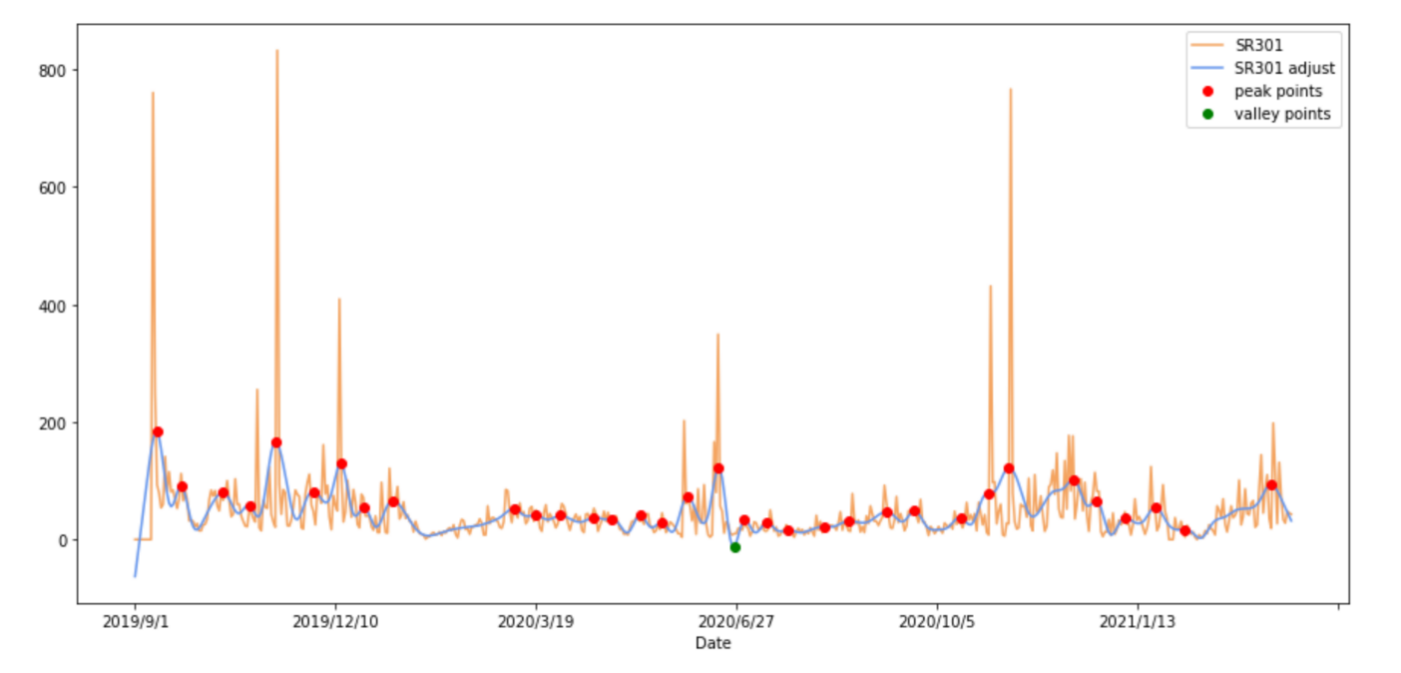
如图所示找出局部极值

1. 拐点筛选，保证峰与峰，谷与谷之间的距离间隔，去除掉间隔过近的拐点，以保证周期时长不会过短。



筛选过后的拐点数据图

1. 与源数据进行对比，查看数据及拐点筛选效果。



与原始数据对比图

**下周预期成果**

**产品周期**

### 根据拐点分析，将数据分为上市前期以及成熟期两段，并对两段的产品销量以及各类数据进行描述性分析，加以对比。

对于不同的产品可能会进行一个综合评价，得到一个比较靠谱的分割比例。

同时，后续的回归模型等也会基于这两段数据，分别进行回归。

**模型与回归**

与上次会议探讨可能有些不同，我们在引入dummy variable之后，决定将所有类型的数据同时进行建模分析。现在经过我们的处理，我们所有的数据，自变量主要包括渠道类型，总费用，对于部分数据还有推行设备等这些变量，都已经数字化且可以直接放入模型进行初步建模。

下周的主要任务就是不同的模型开始进行回归与预测，目前打算使用的模型有：线性回归模型，支持向量机（SVM），随机森林以及XGBoost，利用python实现应该不难，Tree模型可能涉及调参问题。

我们目前的目标是实现广告效果的预测，通过初步建模形成不同广告对广告效果影响的函数，进行预测，从而对我们最终的目标——最优化投放方案起到指导作用。

按目前我们的想法，我们最后得到的可能是只有一个包含哑变量的预测函数，进行预测。而根据上次与投手的沟通，可能有部分渠道的投放比重有一定的范围，于是就可以进行一个规划求解，得到能最大化广告效果的方案。

**后续计划**

接下来的主要剩余工作，一是可能存在的，人工的对数据的可视化分析，二是模型的优化等，三是对站外数据，可能需要加入一些思索，但可能难以加入广告效果的量化分析，最终可能跟评论数据一样，脱离整个的大框架，但可以提供一定的结论。