# 文件说明

* AnalysisData$：数据分析
* CreateTrain$：创建训练集（用去年数据预测今年数据）
* CreateTrain2$：创建训练集（使用整体数据预测某一周的销量）
* DateUtil$：时间工具类
* LoggerUtil$：日志工具类
* Output$：结果输出类
* test$：测试
* TrainModel$：训练模型（用去年数据预测今年数据）
* TrainModelByWeek$训练模型（使用整体数据预测某一周的销量）
* TrainMoelByTimeSeries$：训练模型（使用时间序列的方法）
* cluster.Output$:结果输出类(集群)
* cluster.TrainModelByWeek$：训练模型（使用整体数据预测某一周的销量/集群）
* cluster.TrainModelLinear$：训练模型（使用整体数据预测某一周的销量/使用线性模型/集群）

# 数据分析

随时间的销量变化： - weekNum 代表离预测点的周数

+-------+------+  
|weekNum| count|  
+-------+------+  
| 6| 86585|  
| 7|112896|  
| 8|106617|  
| 9|114461|  
| 10|116029|  
| 11|116756|  
| 12|121918|  
| 13|135743|  
| 14|128344|  
| 15| 92041|  
| 16| 99624|  
| 17|185007|  
| 18|206716|  
| 19|184790|  
| 20|160563|  
| 21|145898|  
| 22|404553|  
| 23|314275|  
| 24|208830|  
| 25|153468|  
| 26|140348|  
| 27|159576|  
| 28|150702|  
| 29|127234|  
| 30|123003|  
| 31|144482|  
| 32|128280|  
| 33|118033|  
| 34| 82206|  
| 35|119689|  
| 36|152568|  
| 37|163488|  
| 38|130286|  
| 39|132246|  
| 40|116518|  
| 41| 96353|  
| 42| 89834|  
| 43| 73000|  
| 44| 59891|  
| 45|119192|  
| 46|193515|  
| 47|182953|  
| 48|133054|  
| 49|118693|  
| 50|117988|  
| 51|109326|  
| 52|112512|  
| 53|126019|  
| 54|105528|  
| 55| 91584|  
| 56| 93187|  
| 57| 90949|  
| 58| 89053|  
| 59| 75384|  
| 60| 63240|  
+-------+------+

# 特征提取：

1. 思路一：用去年数据预测今年数据

使用17年3月的数据和18年3月的数据形成训练集训练模型，然后用17年5月的数据预测18年5月，得出结果

1. 思路二：时间序列

使用每个商品55周的数据拟合出一个函数，推断出未来十几周的销量

1. 思路三：使用整体数据预测某一周的销量

使用所有商品55周的所有特征结合需要预测的目标周形成训练集，训练模型，将训练周数改为62-66，得出结果

1. 思路四：使用滑动窗口训练模型

建立5个模型，分别预测6周后-10周后的销量，特征的采集使用滑动窗口实现

# 结果分析：

思路1和4尚未完成，暂未完成。其他模型结果如下：

* 思路二

预测准度最高，1/(RMSE+1)的分值达到0.22，但是训练样本较少，最严重的问题是训练的训练样本只有少部分在18年的时候出现了，可用性较低

* 思路三

训练样本的结构如下：

商品id、第1周的销量、第1周的点击量....第1周的浏览量.....第55周的销量....目标周数（1-55） 目标周数销量(label)

这个思路保证了训练样本的数量是一定足够的，但困难的是难以找到合适的模型去拟合它，在训练过程中不断的调整参数，达到的最大分值是0、8，在集群中对模型进行了500轮次的训练，但最终导致过拟合，效果不佳。

* 思路4是我认为最合理的，也是最复杂的，还未实现