机器学习纳米学位

Capstone Proposal

刘伟

2019年4月13日

Proposal

Domain Background

Rossmann是欧洲的一家连锁药店，在7个欧洲国家中用超过3000家连锁药店。Rossmann各个药店经理需要对未来6周内每天销售额进行预测，这样可以提前进行合理的资源分配。影响药店销售额的因素有很多，包括促销，竞争对手，节假日，季节以及地理位置，这个项目的目标是基于各个门店过往的销售数据，以及门店本身的信息，对未来6周的销售额进行预测，项目提供的数据集为1115个Rossmann门店的历史销售记录和这些门店的相关信息。

Problem Statement

项目要求预测各个门店，未来6周内每天的销售额，这个问题属于一个回归问题，预测需要考虑的特征包含各个门店的相关信息，比如地理位置，附件竞争对手，促销活动，历史销售额等。具体来说，大致步骤可分为：

- 数据探索性分析：包括对项目提供的门店历史销售数据，以及门店的相关数据进行探索了解，例如数据丰富度，是否存有大量缺失数据，各个特征的数

据分布情况等。
- 特征预处理：包括空值填充，类别特征处理等。
- 特征工程：包括特征组合，特征筛选等。
- 构建模型：选择合适的模型进行训练，可选用包括树模型（xgboost，randomForest），RNN深度模型等。
- 模型评估：定义模型评价指标：rmspe。
- 模型分析及优化：结合模型预测的结果，分析优化方向，例如提取更丰富的特征，选择不同的模型等。

通过以上各个步骤，不断对模型以及数据特征处理，特征工程等方法的优化，在验证集上达到比较高的准确度，体现在对各个门店的未来6周的销售额预测和实际销售额的准确率。

Datasets and Inputs

数据探索

项目一共有三份数据：用于训练的train.csv和store.csv，以及用于测试的test.csv，其中train.csv包含了各个门店历史每天的销售情况：

- Store：各个门店的ID标识，一共有1115家门店
- Date：日期，从2013-01-01到2015-07-31一共942天的数据
- DayOfWeek：1-7分别表示周一到周日
- Sales：某门店某天的销售额，最小0，最大41551
- Customers：某门店某天的顾客数，最小0，最大7388
- Open：某门店某天是否营业，0表示没有营业，1表示营业
- Promo：某门店某天是否有进行促销，0表示没有促销，1表示有促销
- StateHoliday：当天是否是节假日，以及是哪个节假日，0表示非节假

日，a表示公共假日，b表示复活节，c表示圣诞节

- SchoolHoliday：当天是否是学校假日，0表示不是，1表示是

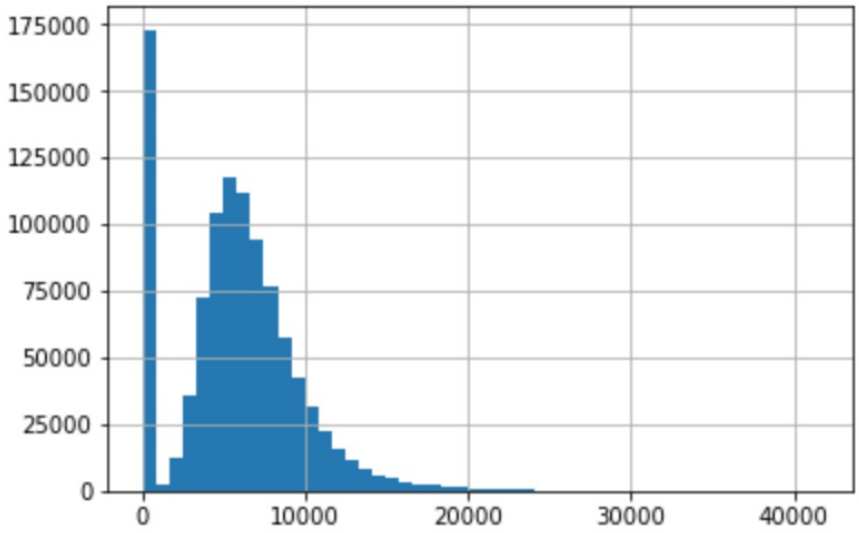
store.csv包含了各个门店的相关信息：

- Store：各个门店的ID标识，和train.csv的Store字段对应
- StoreType：门店的类型，一共有4种门店，分别用a，b，c，d表示，四种门店类型中，b类型的门店数量最少，只有17家
- Assortment：门店的商品类型，一共有3种，分别用a，b，c表示，其中b类型商品的门店数量最少，只有9家
- CompetitionDistance：离门店最近的竞争对手门店距离
- CompetitionOpenSinceMonth：离门店最近的竞争对手门店开业的月份
- CompetitionOpenSinceYear：离门店最近的竞争对手门店开业的年份
- Promo2：表示门店是否有参与进行持续的促销活动，0表示没有参与，1表示有参与
- Promo2SinceWeek：表示门店参与持续促销的周数，NaN表示门店未参与持续促销
- Promo2SinceYear：表示门店参与持续促销的开始年份，NaN表示门店未参与持续促销
- PromoInterval：表示门店每次开始持续促销的月份，NaN表示门店未参与持续促销

探索性可视化

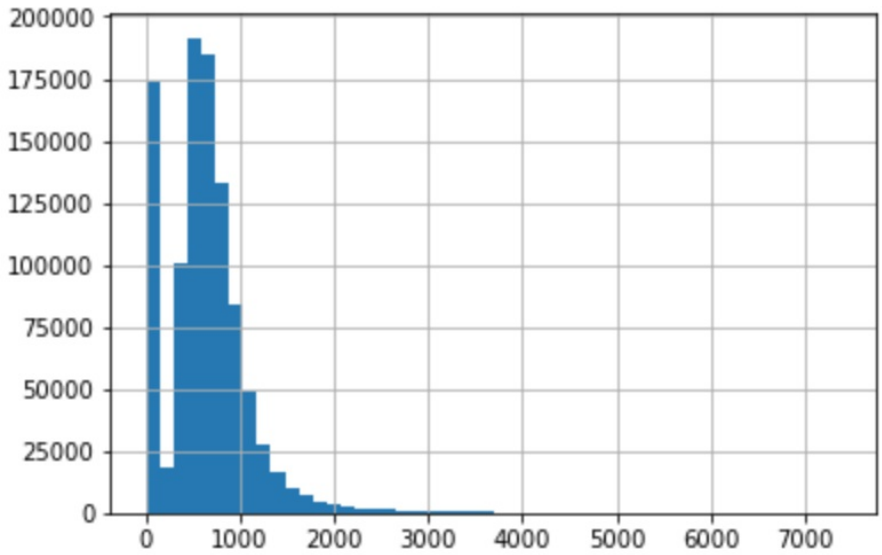
- 训练数据中的Sales，主要集中在4000-8000的范围内，有一部分数据的Sales是0，数值的分布如下图：

```
In [5]: train_df['Sales'].hist(bins=50)
Out[5]: <matplotlib.axes._subplots.AxesSubplot at 0x11193d550>
```



- 训练数据中的Customers，主要集中在300-1000，有一部分数据的Customers是0：

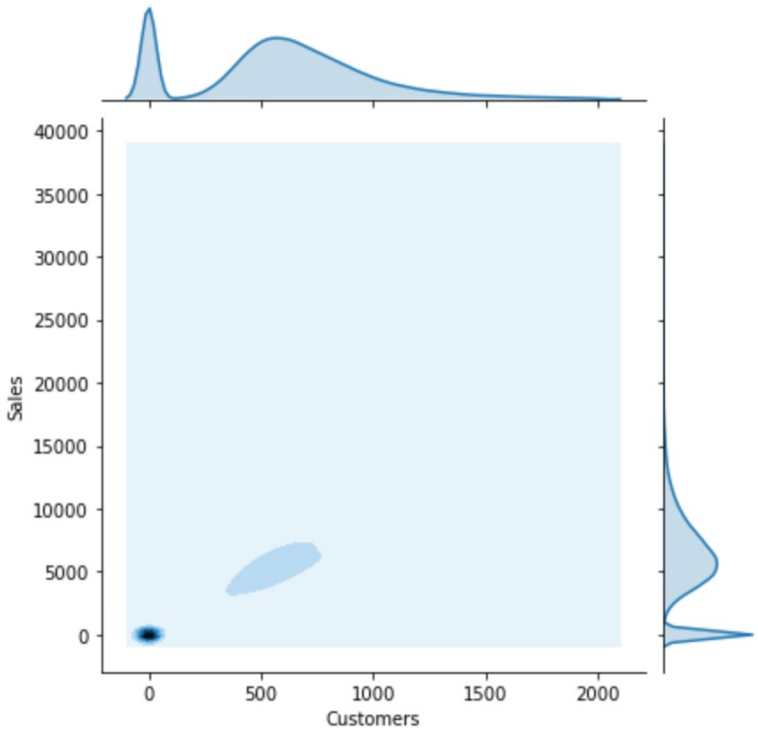
```
train_df['Customers'].hist(bins=50)
Out[3]: <matplotlib.axes._subplots.AxesSubplot at 0x112c5e400>
```



- 训练数据中，Customer人数和Sales销售额成正比关系：

```
In [7]: sample_df = train_df.loc[train_df['Customers'] < 2000].sample(100000)

        sns.jointplot(sample_df['Customers'], sample_df['Sales'], kind="kde")
```

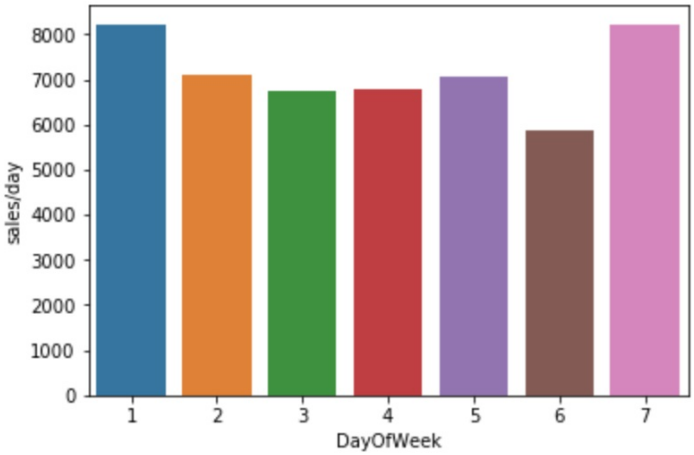
```
Out[7]: <seaborn.axisgrid.JointGrid at 0x111b53518>
```



- DayOfWeek的每一天里销售额的分布也有一定规律，首先看到每天对应的销售额分布，星期一的销售额最高，周日的销售额最低，但发现周日里药店开业的数量也是最低的，因此重新对比周一到周日每天有营业的平均销售额分布，如下图：

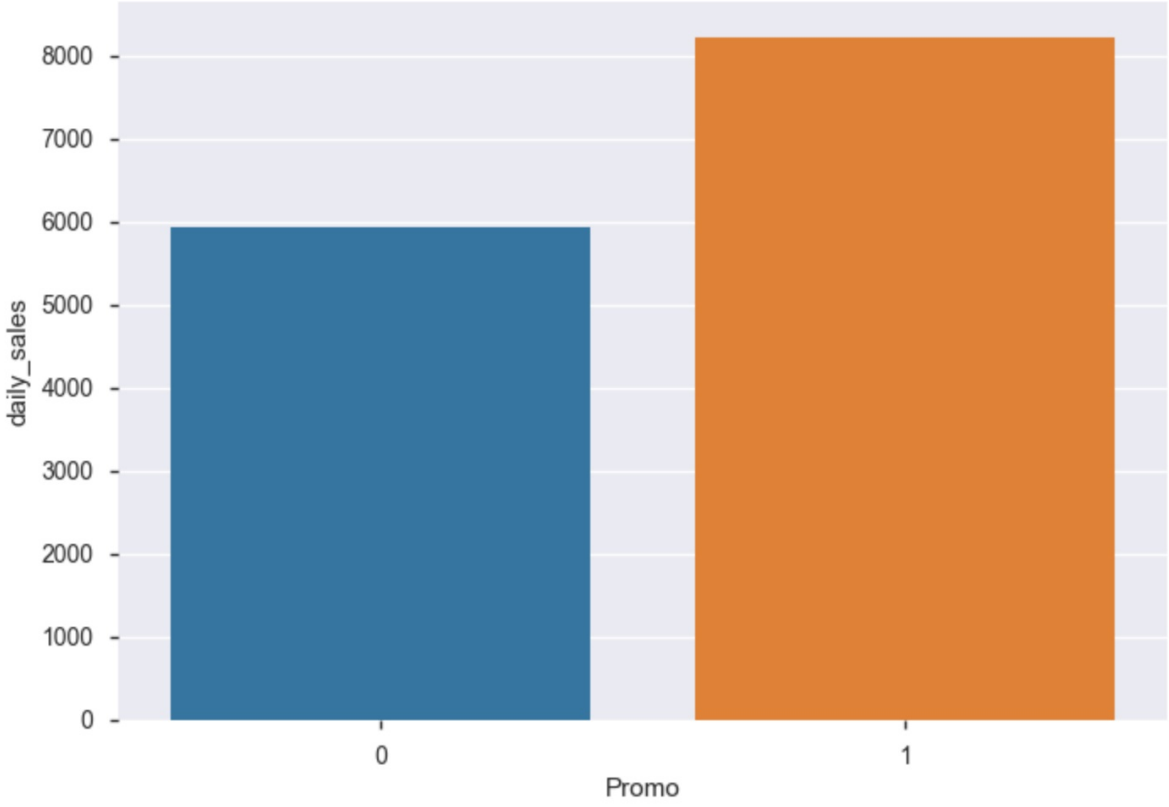
```
dayAvgSales = dayOfWeek_Sales.merge(open_days, how='inner', on = ['DayOfWeek'])
dayAvgSales['sales/day'] = dayAvgSales['Sales'] / dayAvgSales['Open']

sns.barplot(x=dayAvgSales['DayOfWeek'], y=dayAvgSales['sales/day'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x12410a4a8>
```



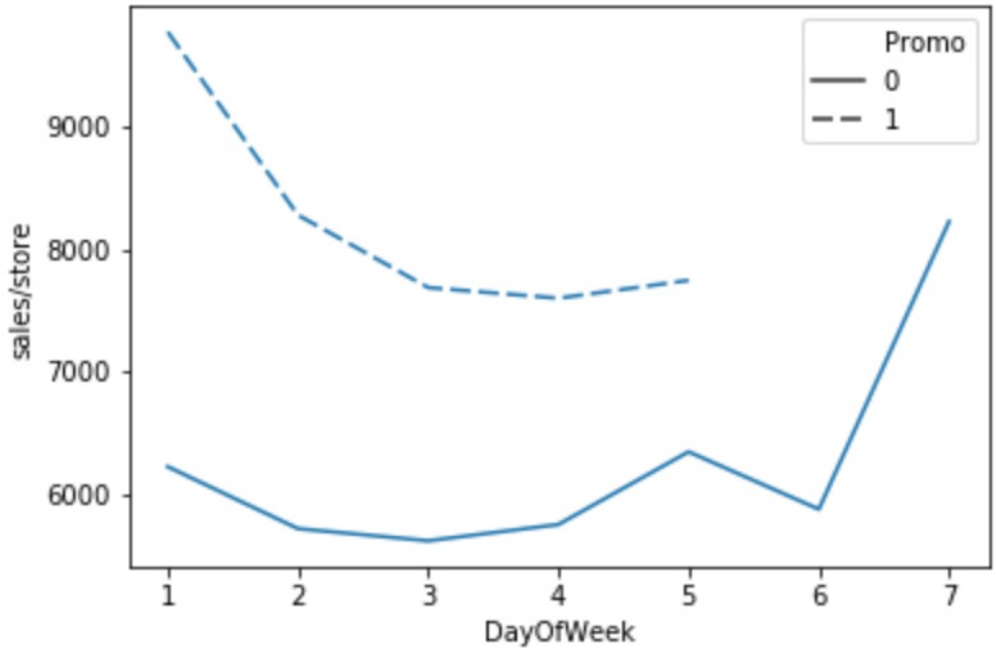
- Promo和销售额的关系，整体平均来看，有促销的销售额比没有促销的销

售额要高：

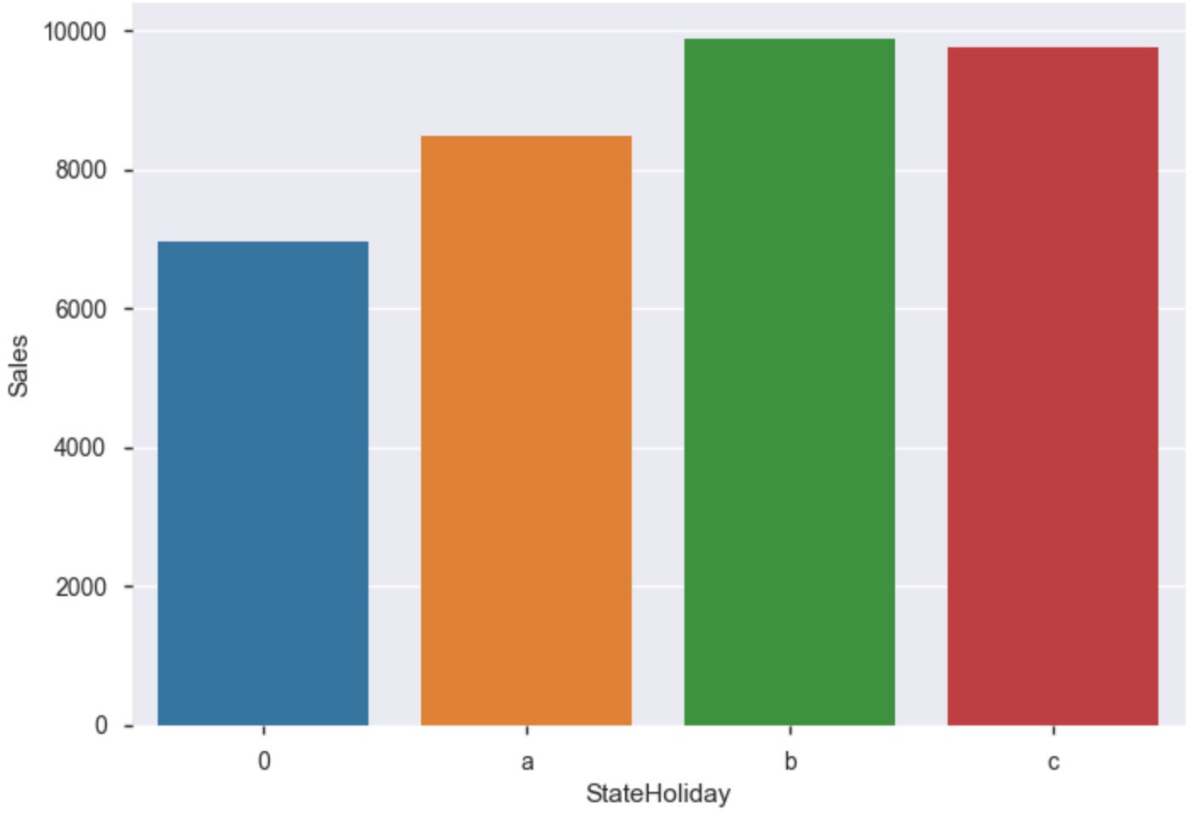


同时，如果把Promo, DayOfWeek结合一起，看有促销和无促销，分别对应每天的平均销售额分布关系，可以发现：

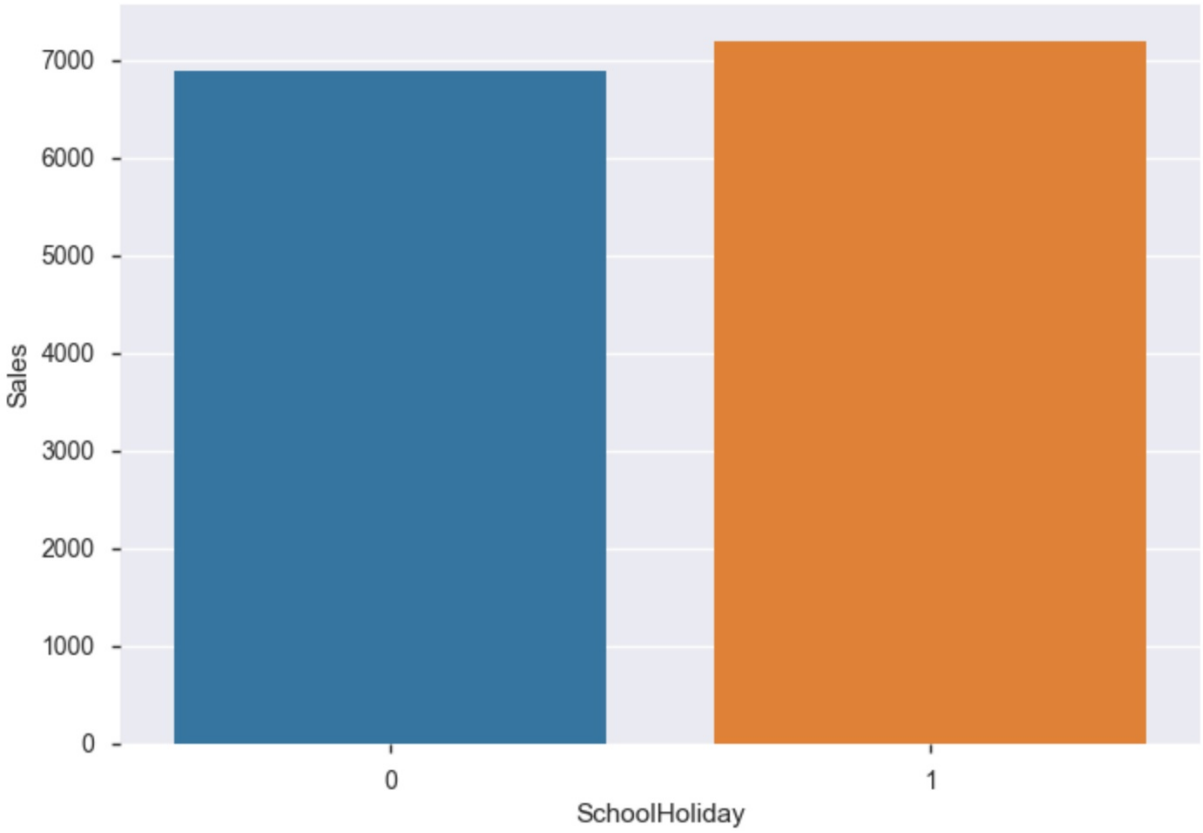
- 周六和周日都没有促销活动；
- 在没有促销活动的时候，如果药店营业，周日的销售额最高；
- 在有促销活动时，周一的销售额最高



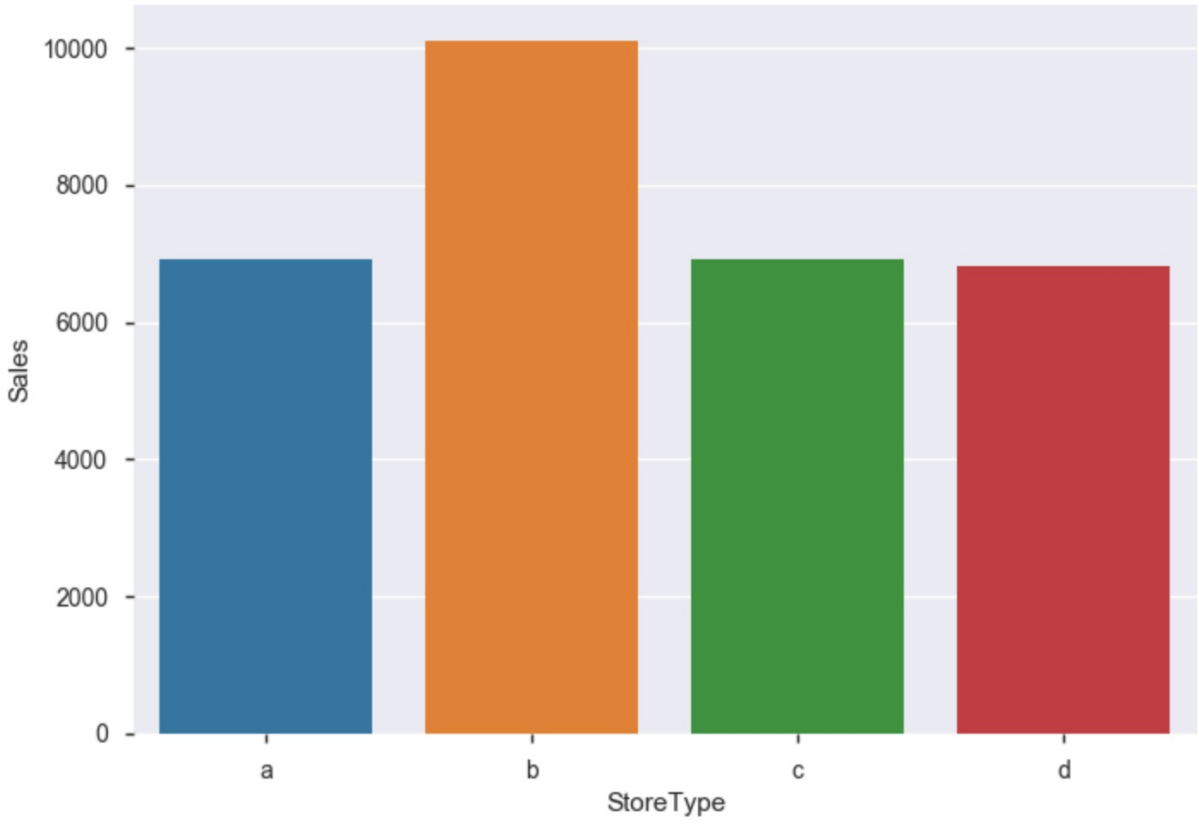
- StateHoliday与销售额的关系，有节假日的药店平均销售额比没有节假日的要高，同时，复活节和圣诞节的平均销售额最高，另外，在训练样本中，StateHoliday存在数字0与字符’0’的数据，需要对他们进行转换，合并为一个类别：



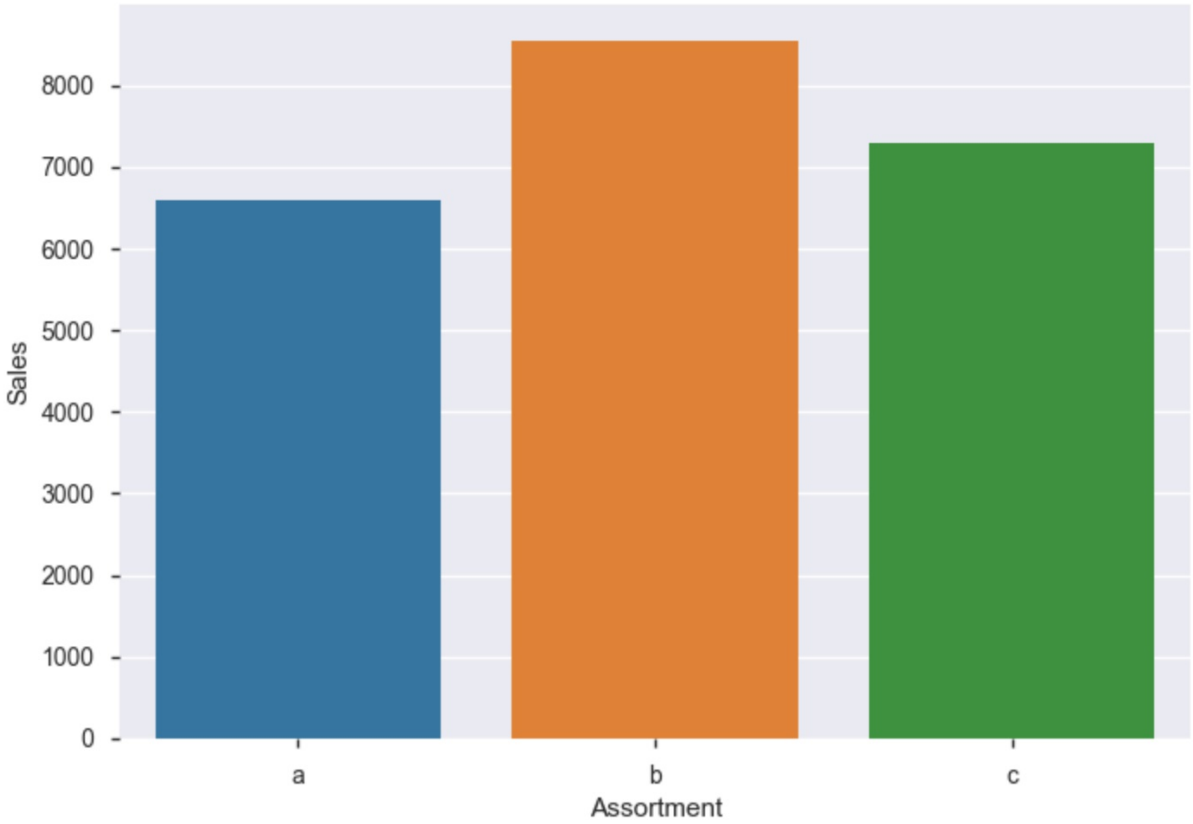
- SchoolHoliday与销售额的关系并不明显，是学校假日的药店平均销售额比非学校假日的高一点：



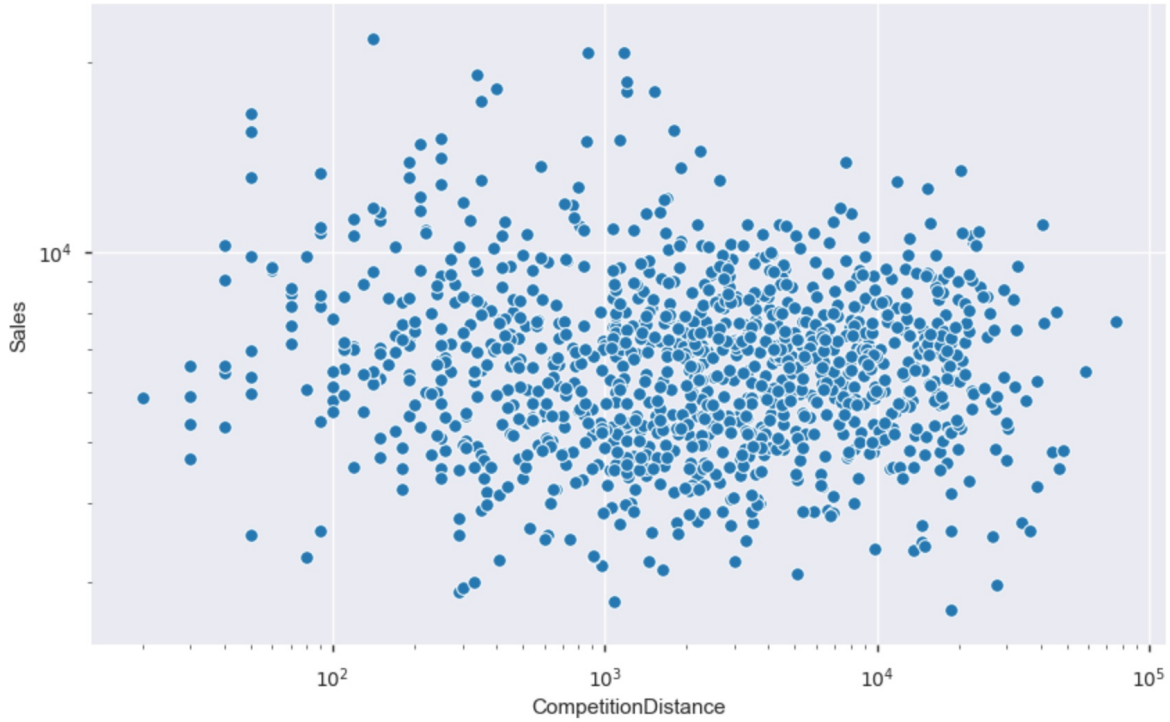
- StoreType一共有四种，统计每种类型的药店的平均日销售额，可发现storeType为b的药店日平均销售额最高：



- Assortment一共有三类，其中b类型对应的日均销售额最高，其次是c：



- 看竞争对手的距离和药店日平均销售额的关系，粗略可看出竞争对手的距离越近，药店的日均销售额较低，竞争对手距离越远，日均销售额高：



Solution Statement

这个项目需要对每个药店未来6周的销售额进行预测，属于一个回归问题。鉴于前面

的特征分析，训练数据中的特征包括数值特征，同时也包含大量的Categorical特征，部分特征可进行特征组合，因此，可以利用的算法包括线性回归，树模型等算法进行拟合，常见的树模型包括随机森林，GBDT，xgboost等。这里对线性回归模型以及xgboost模型进行说明：

线性回归模型，是通过一个多维的线性函数，对一组给定的训练数据，包括多个特征变量，以及一个预测变量进行拟合，在这里是指通过训练数据里的门店自身的特征以及门店的历史销售特征作为特征变量，门店的实际销售额作为待预测的变量，通过各个特征的一个线性函数进行拟合，使得样本中的实际销售额与线性函数预测的销售额尽可能接近，即所有样本的预测销售额与实际销售的均方误差和最小。线性回归模型的优势在于结果非常容易解释，每个特征的权重代表特征对结果的影响大小，权重越高的特征，对结果的影响力最大，因此往往通过一些特征工程的手段，人为制造更多的强特征，使得模型得到优化；而线性回归模型的缺点就是模型只能通过线性方式拟合样本，而往往结果与特征之间并非是线性关系，因此无法用线性回归模型进行有效的拟合。

xgboost模型，是集成模型的一种，通过多个弱的树模型集成而得，每个弱模型的训练是在前一次训练的结果基础上，进一步优化模型预估与实际值之间的残差训练而得，而xgboost的优势在于解决了前面线性回归模型仅限于线性拟合的问题，同时xgboost优化了训练算法，可并行训练，加快模型训练速度。

Benchmark Model

从前面的数据探索部分，我们得知与销售额有强相关的特征有：DayOfWeek，Promo，Open三个特征，因此我们可通过对训练样本的每个门店历史销售额，分别计算每个门店对应每个DayOfWeek，Promo，Open值对应的中位数销售额，作为我们的基准模型数据。

Evaluation Metrics

首先这个问题是销售预测问题，属于回归模型，销售额的准确性是评估模型的指标，采用的评价指标是RMSPE：

RMSPE=\sqrt { \frac { 1 }{ n } \sum { i=1 }^{ n } (\frac { y{ i }-\hat { y{ i } } }{ { y }{ i } })^{ 2 } }

Project Design

首先，需要从项目给出的多个数据集中制作训练集，校验集，其中校验集需尽量和测试集数据保持一致，这里采用的校验集是train.csv中最后6周各门店的销售数据。而训练集则采用之前日期数据。

另外，我们需要对项目给出的train.csv数据进行一定的数据清理，包括缺失值填充，比如CompetitionOpenSinceMonth，CompetitionOpenSinceYear，以及Promo2SinceWeek和Promo2SinceYear。之后，还需要针对性的进行特征工程工作，例如对历史销售额，客户数进行一定维度的统计，从之前数据探索的分析可发现，历史销售数据和DayOfWeek，Promo，Holiday等数据都有强关联，因此可以考虑按这些维度进行统计，制作更能反映数据特点的特征；除了统计维度特征外，还可以对多个原始特征进行合并，例如对Promo2，Promo2SinceWeek，Promo2SinceYear，PromoInterval这几个特征，可以合并为一个Promo2Type的特征，如果门店没有持续促销活动，则为0；若有持续促销活动，但当前时间在活动开始之前，则为1；若正处于促销活动月份期间，则为2；若处于促销活动月份之后，则为3，这样可形成更有表达含义的特征。

最后，通过GridSearch的方法，可以对选用的Model（例如xgboost）进行调参，

最优参数后，通过验证集对Model进行测试验证，比较RMSPE的大小，进一步优化模型，防止overfit。