# 题目四

1. **SQL test**

请完成SQL test.xlsx中的SQL题目

*使用PPT汇报时请略去所有建表的代码，只展示查询代码！*

1. **时间序列预测**

在data.csv中，您将找到产品A在上海的每日销量。字段说明如下：

1. **calendar\_date：日期，起始时间是2016-01-01，结束时间是2022-02-27。**
2. **day\_of\_week：该天为一周中的哪一天，取值为1~7。**
3. **holiday：节假日信息，如果为空，则代表没有节假日。**
4. **quantity：销量。**

您的任务是根据历史数据预测产品A的未来**每日**销量。

1. 向我们展示您如何进行数据清洗。比如如何进行异常值缺失值的识别和处理？
2. 处理数据，并向我们展示2〜3个数据中你发现的有趣的点，比如数据的周期性、特殊日的表现等。
3. 请您构建模型来预测第二天的每日销量。使用2022/1/1之前的数据作为训练集，使用2022/1/1之后的数据作为测试集。您可以使用任何您认为合适的模型（比如传统的线性模型：ARIMA；机器学习模型LightGBM/random forest等；也可以CNN或RNN模型，如**背景知识.pdf**，第9页和第11页中所示）请构建模型并回答以下问题：
   1. 您选择哪种模型？该模型如何工作，为什么适合该任务？
   2. 如何训练模型？超参数如何设置？
   3. 该模型对测试数据的性能如何？请使用MAPE（平均绝对百分比误差）评估模型。
   4. 测试数据中哪5天的错误百分比最高？您如何改善它们？
4. 如果您的任务是一次性预测所有测试时间段的每日销量，那么您将如何构建模型？
5. 请猜测一下2022年3月-5月产品A的每日销量，并说明您是如何参考历史数据给出猜测的？（提示：需要结合客观因素）