# 数据分布

通过观察发现

1. 缺失数据大多数为非随机缺失，缺失有随时间点左截断的现象（大部分银行数据缺失是由于当时没有成立）
2. 时间序列有趋势性，且绝大多数变量之间有相关性

# 数据挖掘中的处理方案

数据挖掘中的模型常常需要等长的模型，数据出现上述现象，不利于处理。为了变成模型好处理的数据，大致有三种方法：

## 丢弃缺失时间点的数据，换来可信度

适用于模型对数据敏感的情景。

## 缺失值补全

1. 局部插值，不大合适，因为数据是截断性的缺失，局部插值不大适合外推法，因为局部插值对于趋势的把握并不好，在离开了数据分布的范围后置信度很低
2. 以均值填充，不合适，非随机性缺失
3. 通过对单变量时间序列建模（如ARIMA等模型）后预测，（这里是时间序列倒叙后向前预测），由于缺失值较多，该方法未能利用整体的趋势和相关性，结果未必很好
4. 通过对多变量时间序列建模（如VAR模型）后预测，（这里是时间序列倒叙后向前预测）
5. 通过建模+EM算法填充
6. 通过建模（线性回归、随机森林、CART、SVM、神经网络等）挖掘出缺失序列和其他非缺失序列之间的关系，预测后填充

适用于模型对数据不敏感的情景，且缺失不要太多。

## 对模型进行改造，让其适合非等长序列

例如其中一个思路：可以加入滑动时间窗口等之类的，在每个窗口内选择非缺失的数据进行挖掘，之后再根据滑动窗口整合每个窗口的模型。

该方法适合数据比较多的情景，适用于序列非随机缺失的情景，适用于缺失序列较多的情景，且对模型有一定要求。

## 综上，选择缺失值补全方案中的第6种方法

# 通过机器学习方法挖掘含缺失序列和其他序列之间的关系并通过预测补全缺失值

## 步骤

Step1，尝试数据挖掘方法：SVM/CART/随机森林/线性回归，构建一个较好的回归模型

Step2，利用回归模型预测缺失序列

Step3，可视化查看效果

## 实践