



# T11

## SQL高级查询+ Pandas分析 ——Kaggle信贷风控Vintage分析



 **bilibili** @BILI-AFAN

 **知识星球** @AFAN的费曼生活

 **GitHub** @BILI-AFAN

 **gitee** @BILI-AFAN



T11

# SQL高级查询+ Pandas分析 ——Kaggle信贷风控Vintage分析



 **bilibili** @BILI-AFAN

 **知识星球** @AFAN的费曼生活

 **GitHub** @BILI-AFAN

 **gitee** @BILI-AFAN

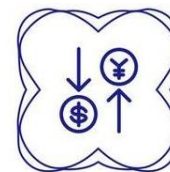


# 回顾：T10

## Python数据库基础操作 ——基于Akshare存储股票行情数据



- 数据库的安装部署流程
- 增删改查等数据库基础操作
- 免费金融数据接口使用方法
- 股票行情数据的存储和更新



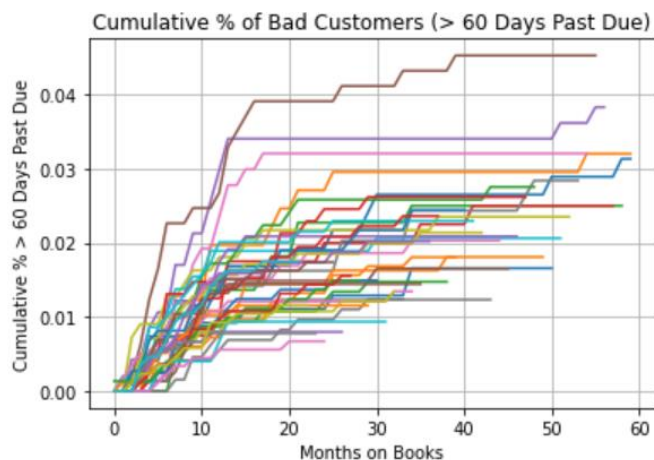
**AkShare**  
Data Science

**数据库的高级查询操作留到下期**





# YOU CAN GET



- Vintage分析方法
- Kaggle平台学习
- Pandas实战分析
- 高级SQL查询操作

kaggle™

 pandas

 MySQL™

# 本教程依赖项

## 依赖情况

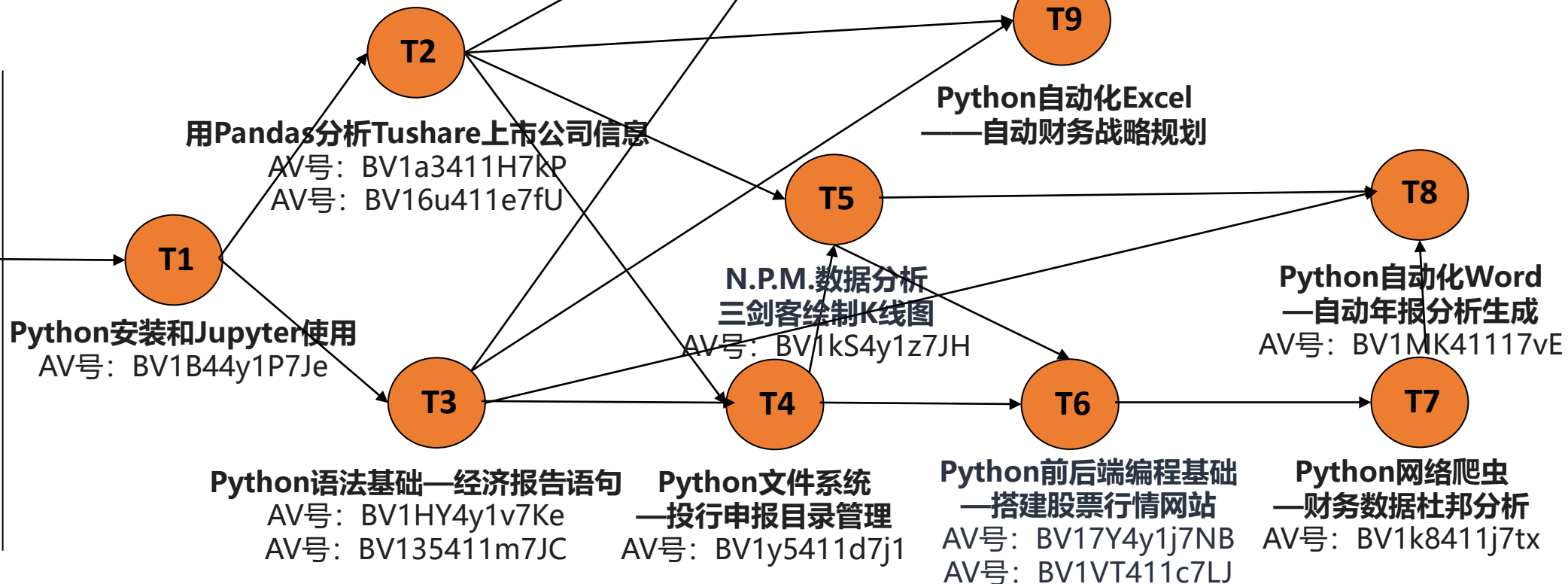
### 全局环境变量

会使用计算机操作

了解金融基本概念

具备基本逻辑思维

了解点金融的高中生



## 首次关注AFAN

A1

第一个B站视频

AV号: BV1j3411W7z5



GitHub

<https://github.com/bili-afan/bili-afan>



Gitee

<https://gitee.com/bili-afan/bili-afan>

资料下载见T1-P9: <https://www.bilibili.com/video/BV1B44y1P7Je?p=9>



 bilibili @BILI-AFAN



# THANKS FOR WATCHING !



GitHub @BILI-AFAN



gitee @BILI-AFAN

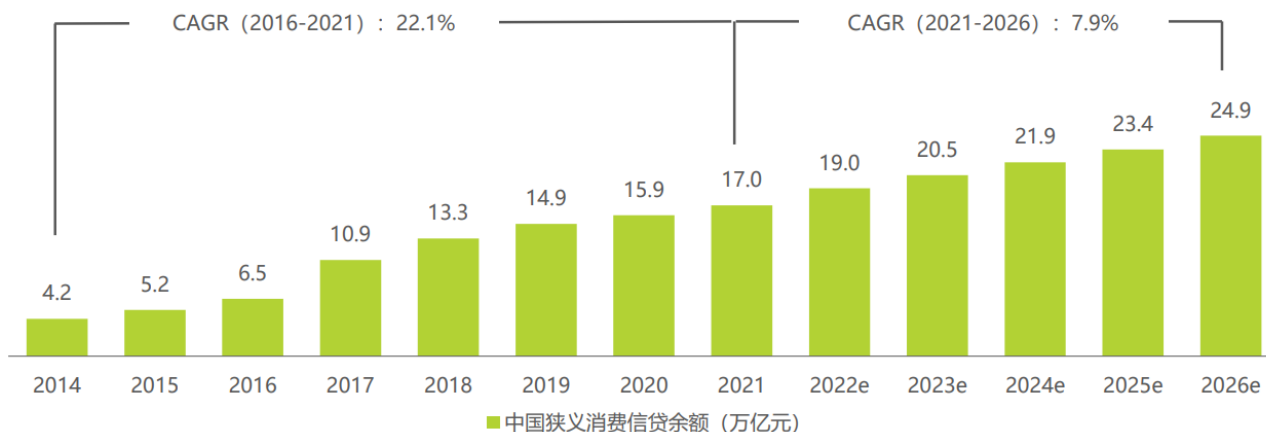
# 消费金融市场情况

## 消费金融近几年发展迅速

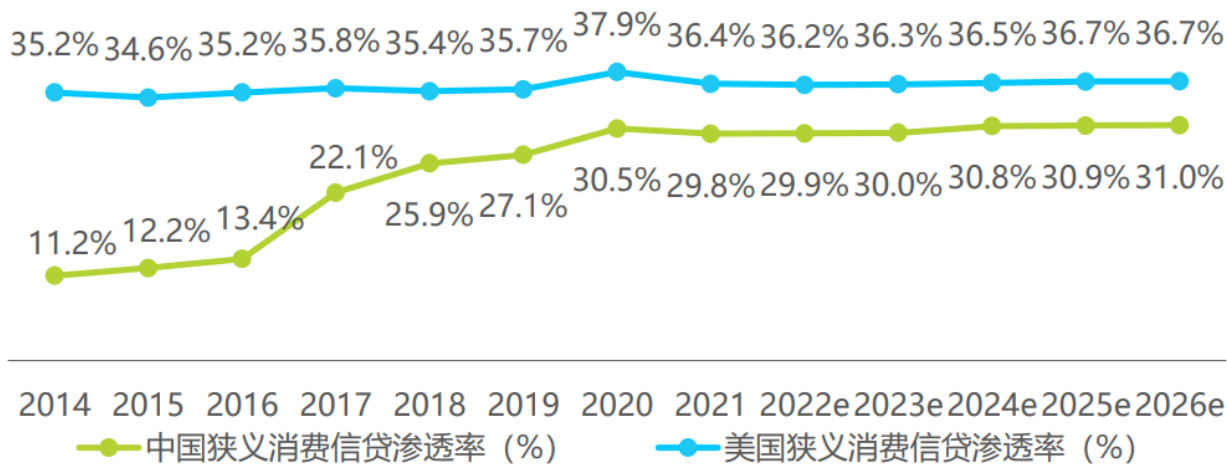
近年来，中国国民经济持续高速增长，为居民生活水平的提升奠定了基础，也使得居民对未来生活水平抱有乐观期待，逐步提升的居民消费水平为消费金融行业的蓬勃发展提供了沃土。从狭义消费信贷余额规模上看，这一指标已经从2014年的4.2万亿上升到2021年末的17万亿，年复合增长率达22.1%。

中国国内整体的信贷行为比较保守，所以中国金融行业尤其在个人信贷端始终存在不可逾越的发展瓶颈。但实际上，美国从1960年的22.1%增长到2021年的36.4%，61年时间增长了14.3%。但中国从2014年的11.2%增长到2021年的29.8%，短短7年的时间增长了18.6%，已经逼近美国现有水平，我国与美国普惠金融进程之间的差距已经越来越小

2014-2026年中国狭义消费信贷余额

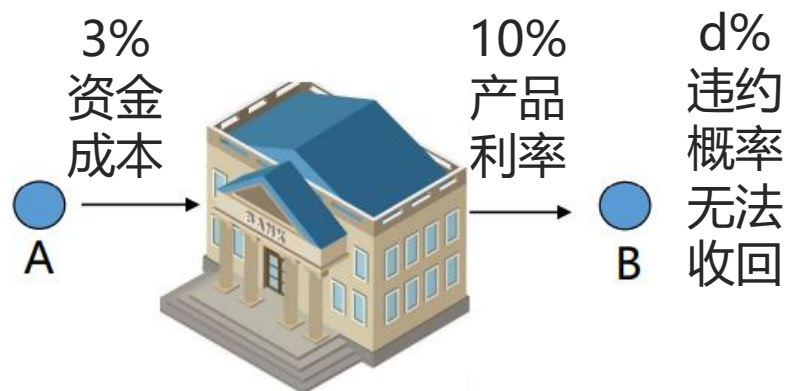


2014-2026年中美狭义消费信贷渗透率



# 消费金融和信贷风控

消费金融是金融科技的重要应用场景



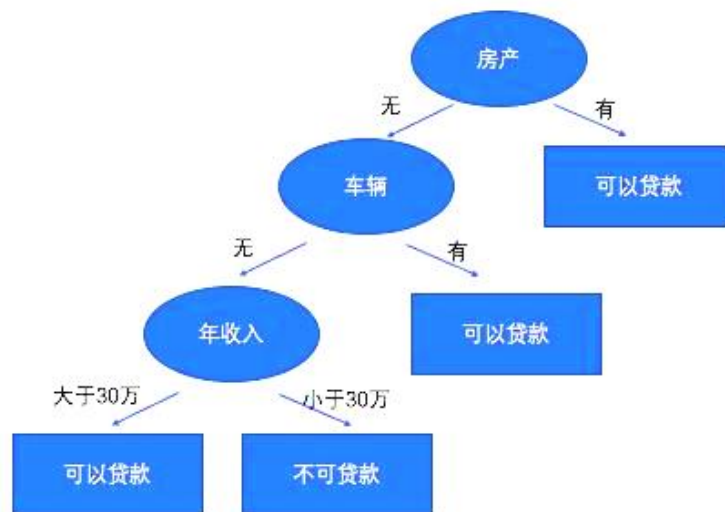
1年后总收入：

$$\begin{aligned} &= (1-d\%)*(1+10\%) + \text{不违约收入} + \text{d\% * 0} + \text{违约收入} - \text{103\% 成本} \\ &= 110\% - 1.1d\% - 103\% \\ &= 7\% - 1.1d\% \end{aligned}$$

违约概率越低  
机构损失越小  
利润空间越大



大数据尽可能描绘人物画像  
人脉、资产、履约、行为、身份



人工智能算法尽可能学习数据表现  
决策树、LightGBM、神经网络



参考知乎：求是汪在路上

《从自动控制理论到信贷风控系统》

<https://zhuanlan.zhihu.com/p/149978781/>



# 信贷风控的链路体系

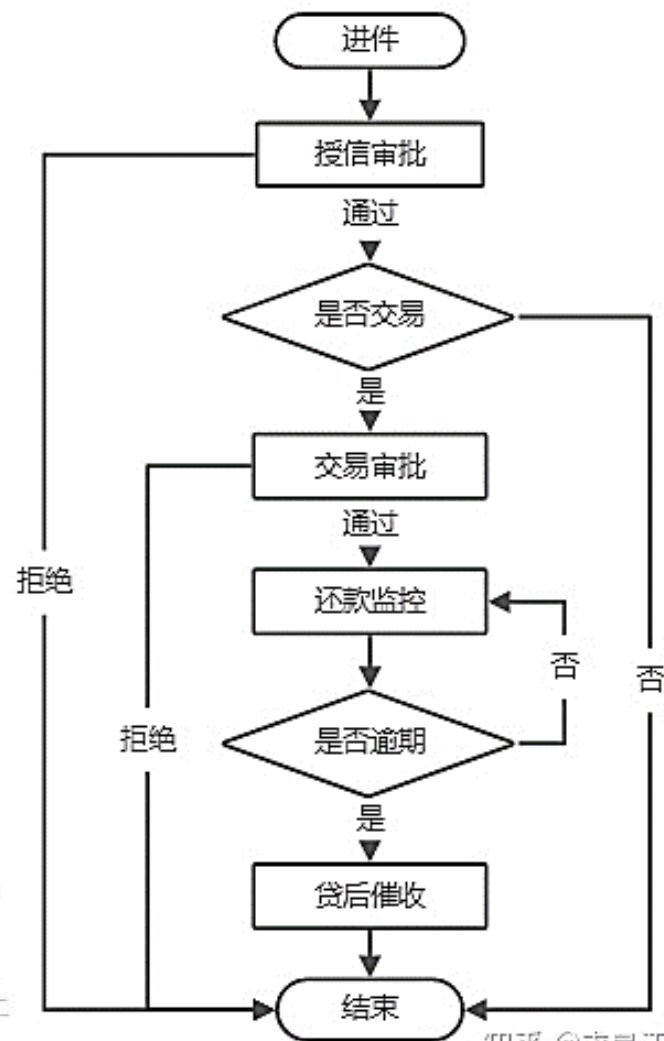
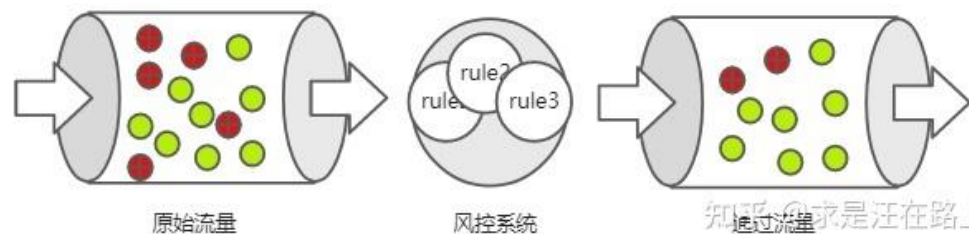
信贷风控流程是一个串行结构，环节较多，链路较长

信贷风控系统大致可分为**贷前**、**贷中**、**贷后**三个阶段。

- 贷前授信审批阶段，主要包括准入、反欺诈等常见拒绝策略，并对授信通过客户给予额度、定价策略。贷中交易审批阶段，主要包括准入、反欺诈等拒绝策略。
- 贷中又需要对存量客户进行经营，包括睡眠唤醒、额度管理、交叉销售等。
- 当放款后，我们又要监控客户的还款行为，对高风险客户提前入催；对已经逾期的客户，在不同阶段采用不同的催收手段，提高回款率。

风控是上游的阀门，本质是剔除坏人、保留好人，其影响到最常见的**两个指标**：

- **通过率**表示阀门拧得有多紧
- **逾期率**表示通过流量有多纯净

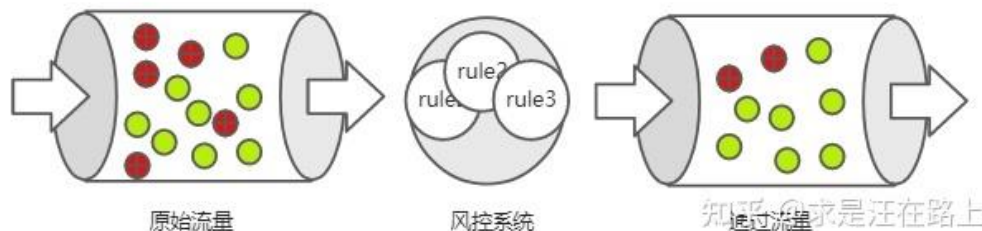


# 还款监控的必要性

## 还款监控是监控“水质”情况的重要手段

风控是上游的阀门，本质是剔除坏人、保留好人，其影响到最常见的两个指标：

- **通过率**表示阀门拧得有多紧
- **逾期率**表示通过流量有多纯净

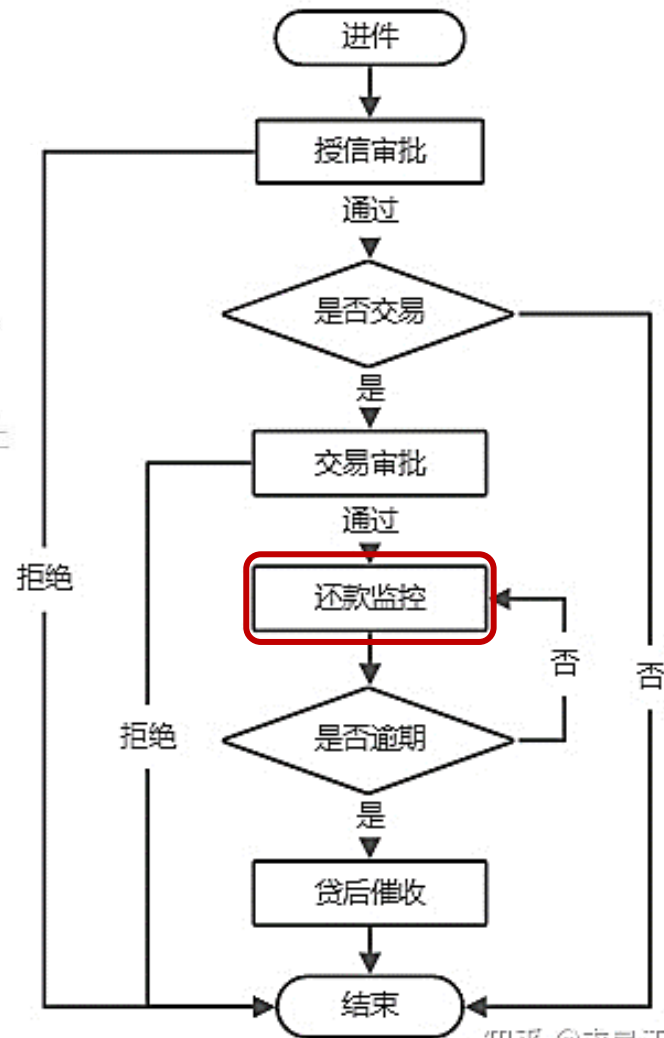


阀门的松紧并没有好坏之分，很多时候和企业战略有关：

- 资方限制，资方提供的额度有限，导致上游流量不足
- 市场扩张，为了获取更多的市场份额，需要让更多的流量通过
- 坏账管理，为了减少贷后的坏账损失，需要减少劣质流量的通过

但是通过流量的质量一定是越高越好，影响质量的主要原因有：

- 过滤系统，风控算法越出色，能更精准的识别“坏客户”
- 上游水质，上游的市场投放规则获取的客群决定了源头的水质





参考知乎：求是汪在路上

《信贷风控中Vintage、滚动率、迁移率的理解》

<https://zhuanlan.zhihu.com/p/81027037>

# 什么是Vintage分析

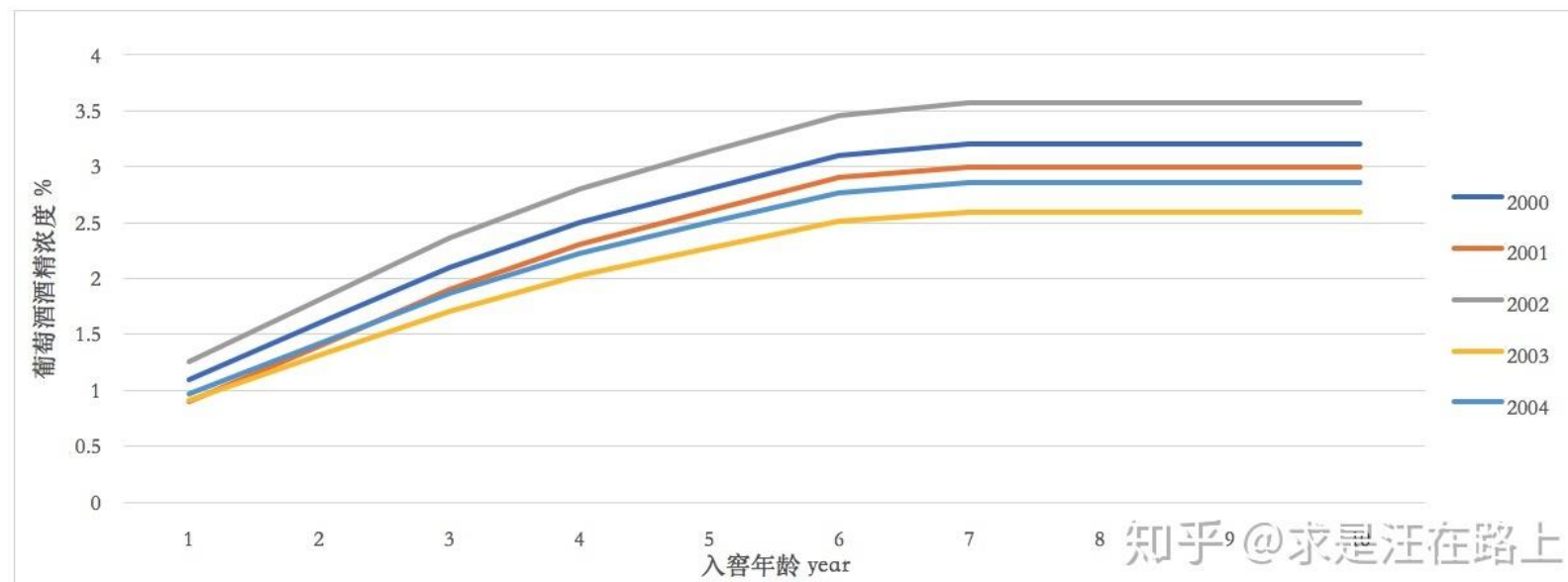
## 不同年份葡萄酒成分和工艺分析的必要

假设AFAN在2000年开了一个红酒酒庄，想利用数据分析技能对葡萄酒的生产进行精细化运营，监控和记录是必不可少的，我们约定：**酒精浓度越高，品质越好。**

由于不同入窖年份的葡萄酒，其指标无法简单比较，所以我们需要将**不同入窖年份的葡萄酒的同一入窖年龄对齐**，才能分析出近几年葡萄酒品质的变化。

葡萄酒入窖年龄及酒精浓度测量记录表

酒精浓度%	入窖年龄									
入窖年份	1	2	3	4	5	6	7	8	9	10
2000	1.10	1.60	2.10	2.50	2.80	3.10	3.20	3.20	3.20	3.20
2001	0.90	1.40	1.90	2.30	2.60	2.90	3.00	3.00	3.00	3.00
2002	1.26	1.81	2.36	2.80	3.13	3.46	3.57	3.57	3.57	3.57
2003	0.91	1.31	1.71	2.03	2.27	2.51	2.59	2.59	2.59	2.59
2004	0.97	1.42	1.87	2.23	2.50	2.77	2.86	2.86	2.86	2.86







参考知乎：求是汪在路上

《信贷风控中Vintage、滚动率、迁移率的理解》

<https://zhuanlan.zhihu.com/p/81027037>

# 什么是Vintage分析

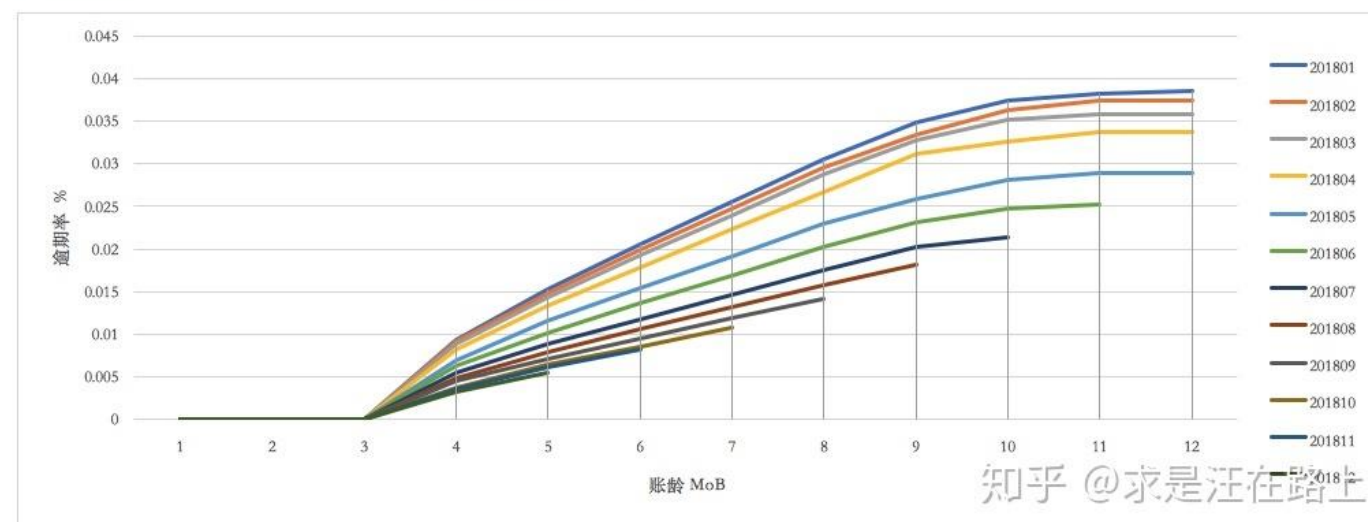
## 信贷领域的资产质量分析工具

行业	葡萄酒	信贷	备注
批次	入窖年份	放贷月份	批次标签
品质	酒精浓度	逾期风险	纵坐标
时间	入窖年龄	账龄MOB	横坐标
成熟期	品质成熟期	账户成熟期	Vintage曲线特征
影响因素	气温、日照、降水等	风控策略、客群、市场环境等	原因分析

用Vintage曲线分析资产质量，可以：

- **确定资产质量**：一般以逾期率来定义资产质量，也就是曲线平缓后对应的逾期率。
- **分析变化规律**：资产质量（例如逾期率指标）的变化情况，如果前几期逾期率上升很快，那么说明短期风险没有捕捉住，欺诈风险较高；反之，如果曲线一直在上升，说明信用风险识别能力不佳。
- **确定账户成熟期**：用来判断客户展现好坏的时间因素，从而帮助定义表现期。
- **分析影响因素**：风控策略收紧或放松、客群变化、市场环境、政策法规等都会影响资产质量。分析影响因素，可以用来指导风控策略的调整。

M4+ %	账龄(MOB)											
放款月份	1	2	3	4	5	6	7	8	9	10	11	12
201801	0.00%	0.00%	0.00%	0.93%	1.52%	2.05%	2.55%	3.06%	3.49%	3.74%	3.83%	3.85%
201802	0.00%	0.00%	0.00%	0.91%	1.48%	1.99%	2.47%	2.96%	3.35%	3.63%	3.74%	3.74%
201803	0.00%	0.00%	0.00%	0.88%	1.43%	1.93%	2.40%	2.87%	3.28%	3.52%	3.59%	3.59%
201804	0.00%	0.00%	0.00%	0.82%	1.33%	1.79%	2.23%	2.67%	3.12%	3.27%	3.38%	3.38%
201805	0.00%	0.00%	0.00%	0.70%	1.15%	1.54%	1.92%	2.30%	2.58%	2.81%	2.90%	2.90%
201806	0.00%	0.00%	0.00%	0.62%	1.01%	1.36%	1.69%	2.02%	2.31%	2.48%	2.52%	
201807	0.00%	0.00%	0.00%	0.55%	0.88%	1.18%	1.47%	1.76%	2.03%	2.14%		
201808	0.00%	0.00%	0.00%	0.49%	0.79%	1.06%	1.32%	1.58%	1.81%			
201809	0.00%	0.00%	0.00%	0.45%	0.71%	0.95%	1.19%	1.42%				
201810	0.00%	0.00%	0.00%	0.37%	0.64%	0.86%	1.07%					
201811	0.00%	0.00%	0.00%	0.35%	0.61%	0.82%						
201812	0.00%	0.00%	0.00%	0.32%	0.55%							



知乎 @求是汪在路上

# Kaggle

## 数据分析建模方案的众包平台

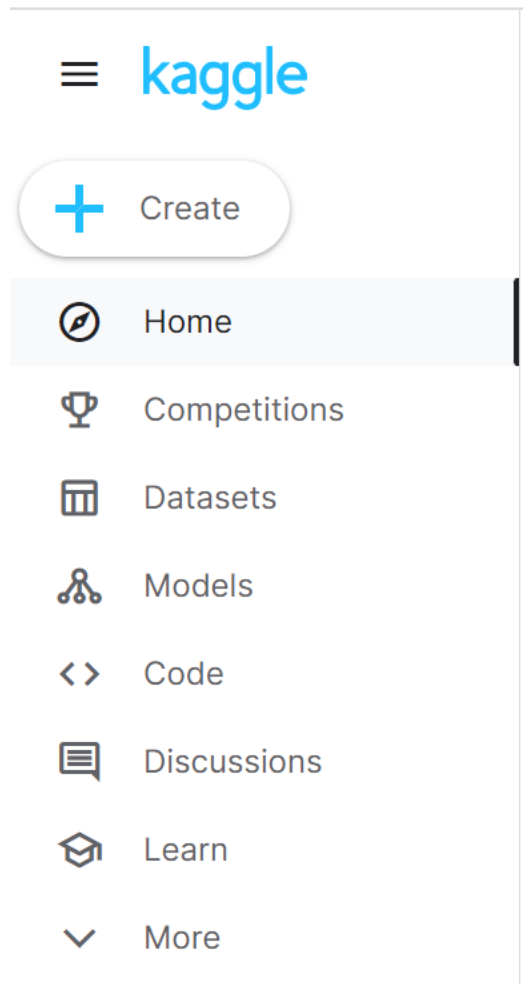
Kaggle是由联合创始人、首席执行官安东尼·高德布卢姆 (Anthony Goldbloom) 2010年在墨尔本创立的, 主要为开发商和数据科学家提供举办机器学习竞赛、托管数据库、**编写和分享代码**的平台。该平台已经吸引了80万名数据科学家的关注, 这些用户资源或许正是吸引谷歌的主要因素。

企业和研究者可在其上**发布数据**, 统计学者和数据挖掘专家可在其上**进行竞赛**以产生最好的模型。这一**众包模式**依赖于这一事实, 即有众多策略可以用于解决几乎所有预测建模的问题, 而研究者不可能在一开始就了解什么方法对于特定问题是最为有效的。Kaggle的目标则是试图通过众包的形式来解决这一难题, 进而使数据科学成为一场运动。2017年3月8日谷歌官方博客宣布收购Kaggle。

TIPS: 能学习代码、能打比赛赚钱、能写在简历上的好社区

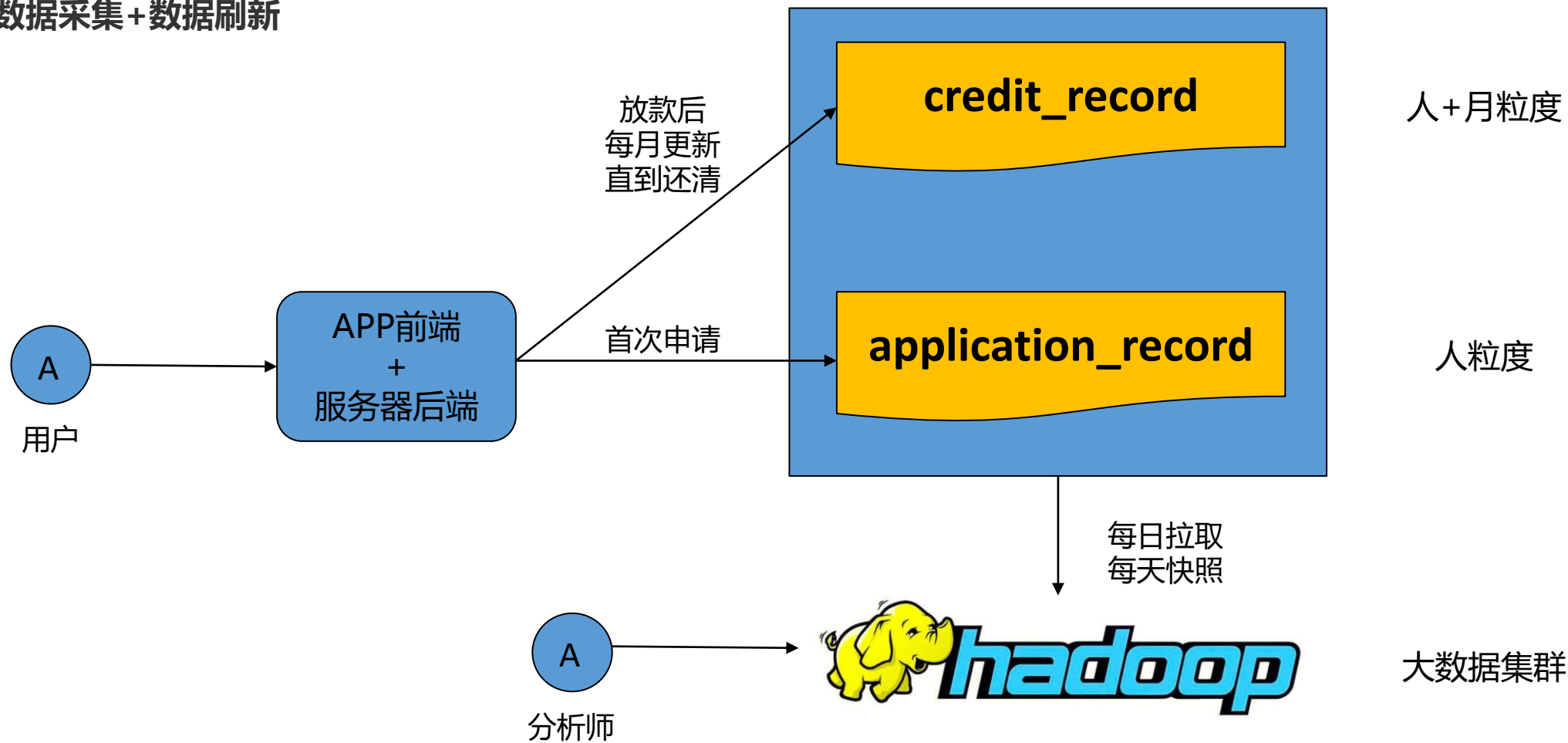
国内对标: 阿里云天池 <https://tianchi.aliyun.com/competition/activeList>

Vintage分析数据集: <https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction>



# 数据表的业务背景

## 用户数据采集 + 数据刷新





 **bilibili @BILI-AFAN**



# THANKS FOR WATCHING !



**GitHub @BILI-AFAN**



**gitee @BILI-AFAN**