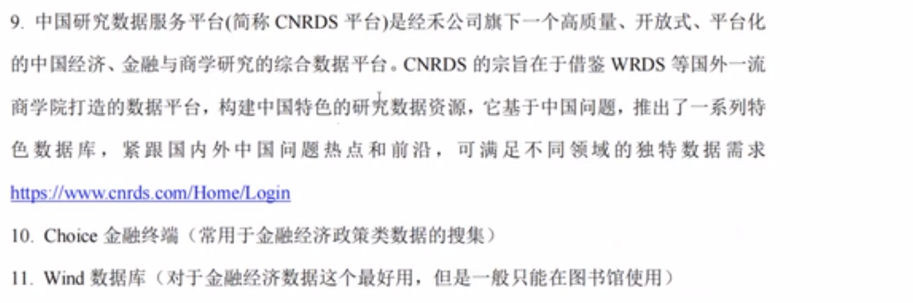
1. **数据处理与检验**

****

**统模**

常见数据预处理方法（缺失值、异常值、数据标准化、归一化etc）

描述性统计后和考虑数据与现有文献数据的一致性和可比性（eg新质生产力几项核心变量和现有文献基本一致，确保了研究结果的可比性）

不要只放统计结果描述，要注重对结果的解释

观察描述性统计中结果有无明显错误/异常值（3sigma原则）

**二、统计检验理论**

case：女士可以判断8杯奶茶先加奶还是茶，检验是否真能辨别还是猜的

原假设H0:乱猜50%对

判断结果是否合理：如猜对6杯以上概率>5%,不是靠猜；否则可能是猜的

思想：利用已知样本信息导出分布从而检验假设

步骤：

1. 提出假设 原假设H0:the one we want to challenge eg.女士没有辨别能力是靠猜的；

备择假设H1: the one we want to prove eg.女士有辨别能力

1. 选择显著性水平alpha：通常取5%，表示我们允许犯错的概率。（有5%的风险会错误的拒绝原假设）
2. 计算检验统计量和p值
   1. 检验统计量：根据实验数据计算的一个值，用来衡量结果与原假设的偏离程度
   2. p值：在原假设成立的前提下，出现当前结果（或更极端结果）的概率
   3. p越小，越不支持原假设
3. 比较p值和显著性水平 （统模比赛中主要看的）
   1. 如果p值小于alpha，可以拒绝原假设，备择假设可能成立
   2. 如果p值大雨alpha，结果还算合理，无法拒绝原假设
4. 根据p值和显著性水平比较，得出统计结论

假设检验的核心：根据p值判断是接受原假设还是拒绝原假设

**正态性检验（知道如何解释）**

H0:数据服从正态分布（p大于alpha） H1:数据不服从正态分布

常用场景：描述性统计中用于说明数据分布（成绩、身高etc是否符合正态分布）、回归分析中残差检验

方法一：

Shapiro-Wilk检验：考察数据点是否围绕样本均值对称分布。p<0.05，拒绝原假设，数据不服从正态分布 python中stats.shapiro函数包

方法二：

QQ图 调用python stats.probplot函数包 根据绘图结果判断是否正态分布

**独立性检验**

H0: 变量A与变量B之间相互独立

H1:变量A与变量B之间不相互独立

常用场景：判断两个分类变量之间是否独立（吸烟与肺癌的关系）

方法：卡方列联检验 调包python

也是看p值，p<0.05拒绝原假设，两个变量不相互独立

**分布拟合检验**

H0:某组数据服从\*\*分布 H1:某组数据服从\*\*分布

使用情况：用于检验数据分布或者通过随机数进行概率仿真；正态性检验呢往往用于检验数据是否服从正态分布，分布拟合检验是任何一个分布 eg灯泡使用寿命一般服从指数分布，排队人数一般服从泊松分布

思想：检验样本数据分布是否与总体分布相符合（连续性数据需要离散化，分组）

spss求解，

同时服从多个分布--》看如何解释，除了考虑统计合理性，还要考虑现实合理性

非参数检验

Wilcoxon符号检验 （python）

不能判断是什么分布，用更粗糙的办法对假设结果进行评判

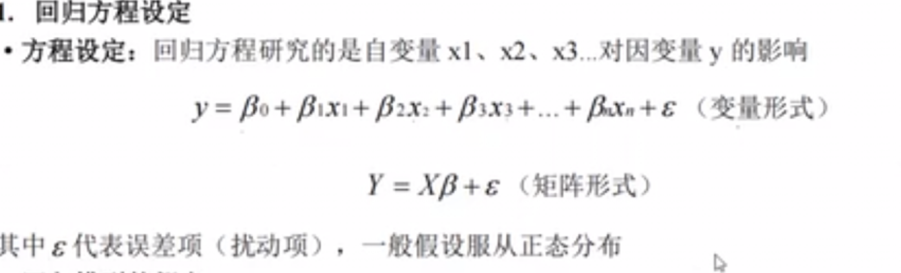
H0:A组和B组不存在显著差异 H1:A组和B组存在显著差异

看p值-->原假设是否成立-->是否存在显著差异

使用：国家出台政策，检验政策出台前和出台后数据是否有显著差别，判断政策是否有效果

**三、回归分析**

**回归方程设定**



方程的参数估计：最小二乘法stata

SSE可解释平方和

SST总平方和

SSR总参差平方和

SST=SSE+SSR

SSE越大，SSR越小，回归效果越好

R方=SSE/SSR

统计建模中一般采用调整自由度后的R方，

回归方程的假设检验：

t检验、F检验

1 单个系数的检验：t检验 （判断每个自变量的显著性）

t检验用于检验回归模型中每个自变量系数是否显著不同于0

原假设H0: 系数=0，备择假设 H1:系数不等于0

t统计量计算方法

结果解释：如p值小于显著性水平0.05，则拒绝原假设，认为该系数显著不为0

回归方程的总体检验：F检验

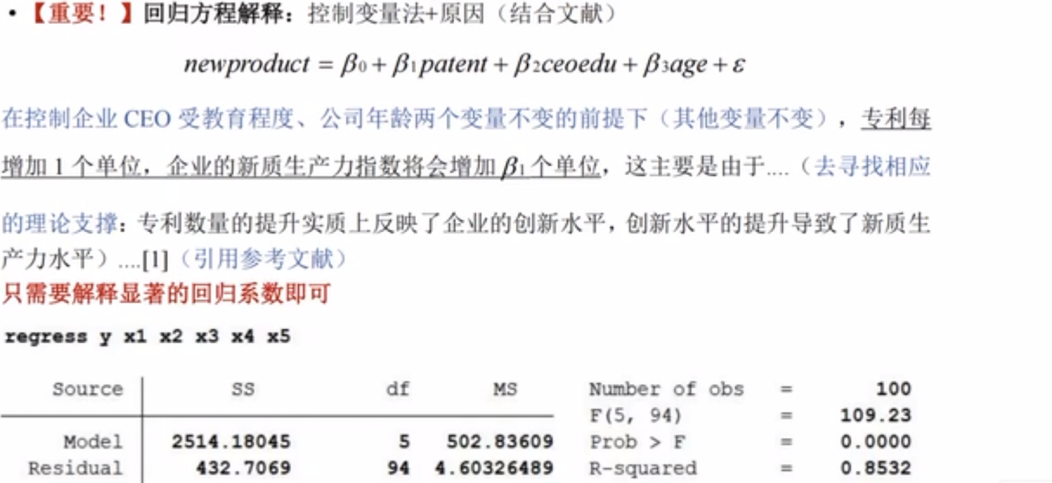
F检验用于检验回归模型整体的显著性，即判断所有自变量对因变量的影响是否显著。

原假设H0:系数全为0；备择假设H1；系数不全为0

结果解释：如p值小于显著性水平0.05，则拒绝原假设，认为至少有一个自变量对因变量有显著影响，使整个模型是显著的

重要！

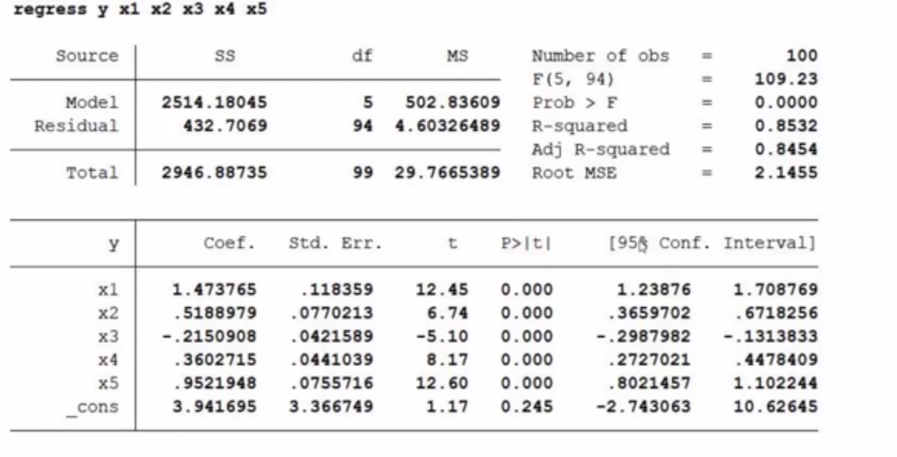
回归方式解释：



研究专利patent 对新质生产力newproduct的影响，有被解释变量（newproduct）、核心变量（patent）、控制变量（ceo教育程度、公司年龄）、残差

解释自变量影响因变量发生的变化+原因

只需要解释显著的回归系数即可



问题：

1.多重共线性

vif判断

危害&解决方案（删除冗余自变量、合并自变量、增加样本量、ridge regression、标准化自变量）

1. 自相关与异方差问题
   1. 自相关：一个时间序列数据中，当前值和过去值之间存在相关性问题

解决办法：A时间序列模型 B费用调整 （带有自相关调整的标准误）C差分：对时间序列数据进行差分操作，消除自相关

1. 异方差：误差项的方差不恒定，导致标准误偏差

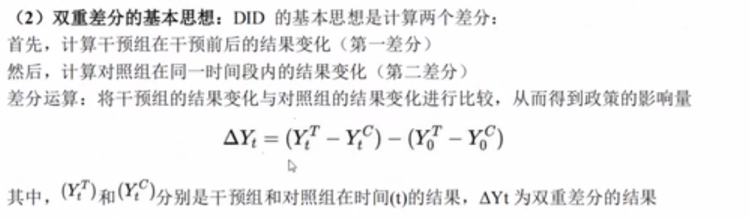
解决办法：A加权最小二乘法：为每一个观测值分配权重，逆向调整其方差 B使用稳健标准误（Stata）C数据变换：对变量进行变换（如对数变换）来稳定方差

1. 聚类方法（统模中比较常用）在时间序列或面板数据上聚合数据（例如按时间段用户分组），可能会减少自相关和异方差的影响。因为聚合后，每个组的估计可能更稳定。采用聚类稳健标准误（如对样本中某一分组的聚合）

固定效应

固定效应将个体效应视为固定的，使研究者控制不随时间变化的个体特征。stata

**实验设计与机制检验（**把回归做的很细）

1. 双重差分法（Difference in Differences)：比较干预组和对照组在干预前后结果的变化，控制未观察到的异质性和时间趋势对结果的影响

应用：研究不同地区最低工资政策影响。A地区（干预组）提高了最低工资，B（对照组）没有，比较两地区在政策实施前后的就业变化。

使用时要进行稳健性检验：平行趋势检验、安慰剂检验、倾向匹配得分、面板Probit等

1. 中介效应（慎用）
   1. 一个自变量（X）通过影响一个或多个中介变量（M）影响因变量（Y）
   2. eg教育水平-->工作经验-->收入
   3. 回归分析
   4. 完全中介效应（x只通过m影响y）&部分中介效应（x还可以影响y，系数显著）
2. 调节效应

某一变量的影响因第三个变量的变化而变化 eg受教育程度对收入的影响会收到年龄的影响而变化。

构造交互项（X1\*Z)

1. 分组回归（也可以用来验证异质性）

**回归模型拓展**

1. Lasso回归 常用于高维数据的分析，如基因表达数据分析 数据集包含大量数据
2. 岭回归 解决多重共线性问题 eg经济学模型中多个宏观经济指标可能彼此高度相关，稳定模型估计
3. 逻辑回归：预测事件发生的概率 预测患病概率

**四、方差分析**

**1.单因素方差分析**

比较三个或更多组样本的均值是否存在显著差异。通过比较组内方差和组间方差来判断组与组之间的差异是否显著。

原假设H0:所有组均值相等（没有显著差异）

备择假设H1：至少有一组的均值与其他组不同

组内方差：每组内部的数据波动程度

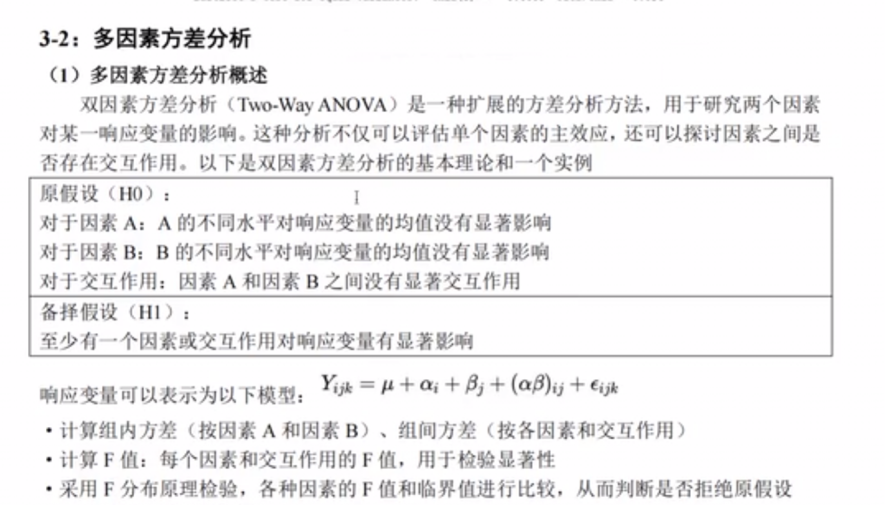
组间方差：计算各组均值与总体均值之间的波动程度

F统计量：计算F值，该值为组间方差/组内方差。F值较大，表明组间差异比较大，可能拒绝原假设，F值与临界值比较，决定是否拒绝原假设。



**2.多因素方差分析：**

研究两个因素对某一相应变量的影响，可以评估单个因素的主效应，还可以探讨因素之间是否存在交互作用，





1. **时间序列分析**

对数据的预测

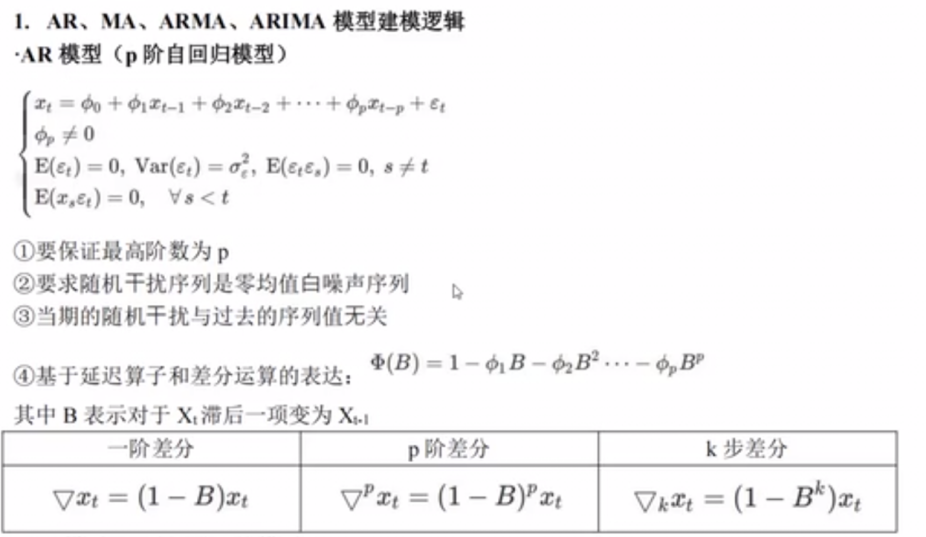
**时间序列的预处理：**

1. 自协方差函数、自相关系数
2. 序列平稳性与ADF检验
3. 纯随机性检验

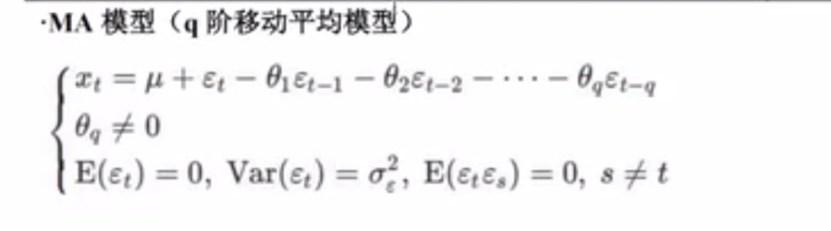
能建模的一定是平稳、非随机的序列。非平稳序列需要通过差分的方法变为平稳序列后才能建模。

**ARIMA模型**

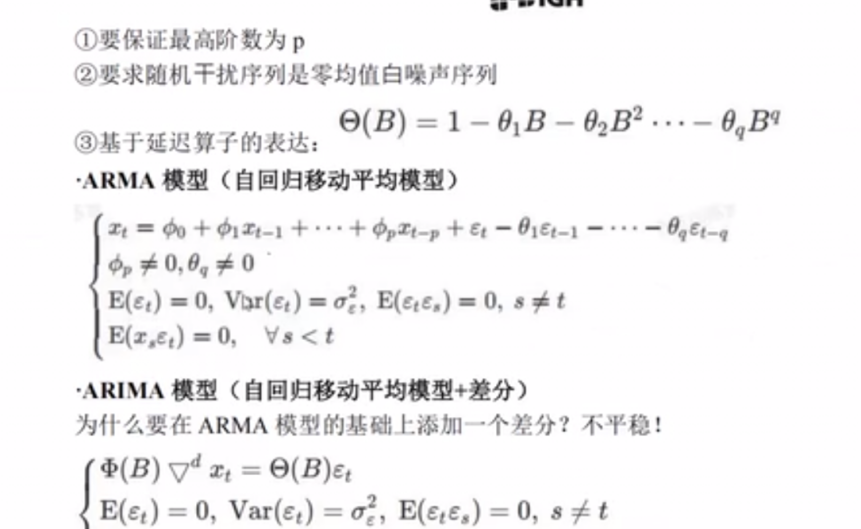
1.AR模型（p阶自回归模型）用已知数据预测现在的



1. MA模型（q阶平均移动模型）用不能观测到的数据预测



1. ARMA模型（自回归移动平均模型）
2. ARIMA模型（自回归移动平均模型+差分）



ARCH&GARCH模型

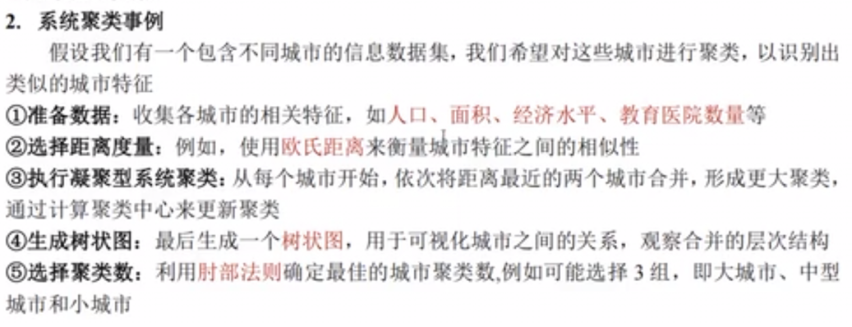
1. **聚类分析 （都是python制图.?)**
2. **k-means聚类**

k簇，均值作为聚类中心

eg市场细分和用户画像

注意：对于聚类算法，不同方法可能有不同结果，关键是对结果的解释

1. 系统聚类



1. 谱聚类 适用于图像聚类
2. **机器学习**

1.KNN算法

2.朴素贝叶斯算法 交叉验证&网格搜索

3.支持向量机（SMV）

4.随机森林

**真题：**

1.数据新动能和新经济的统计测度研究（测度实证类）

政策性呼应

**前言**：背景、现状（严峻）、所以要研究数据新动能

**研究内容、研究意义**（理论意义：为学界；现实意义：为个人、政府、企业）

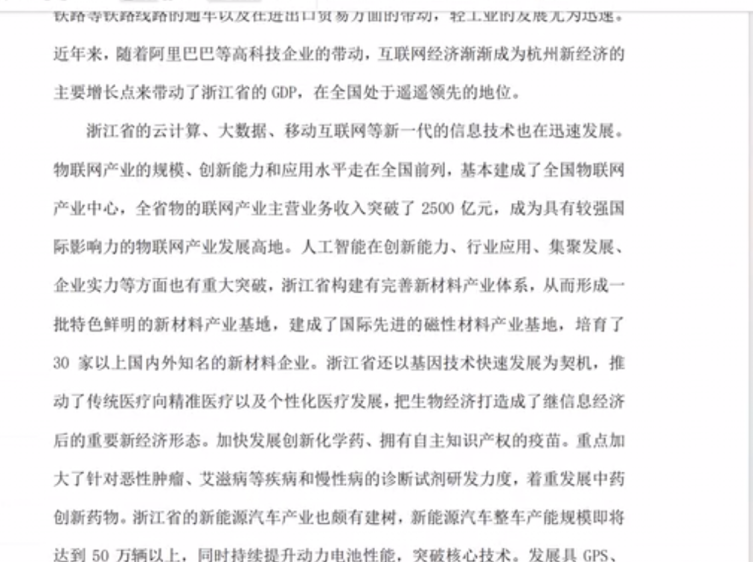
**文献综述（**注意要有评述）（主要概念（新经济、新动能）这个文献解决了什么问题，可以怎样借鉴；未解决什么问题，我怎样解决；研究方法）



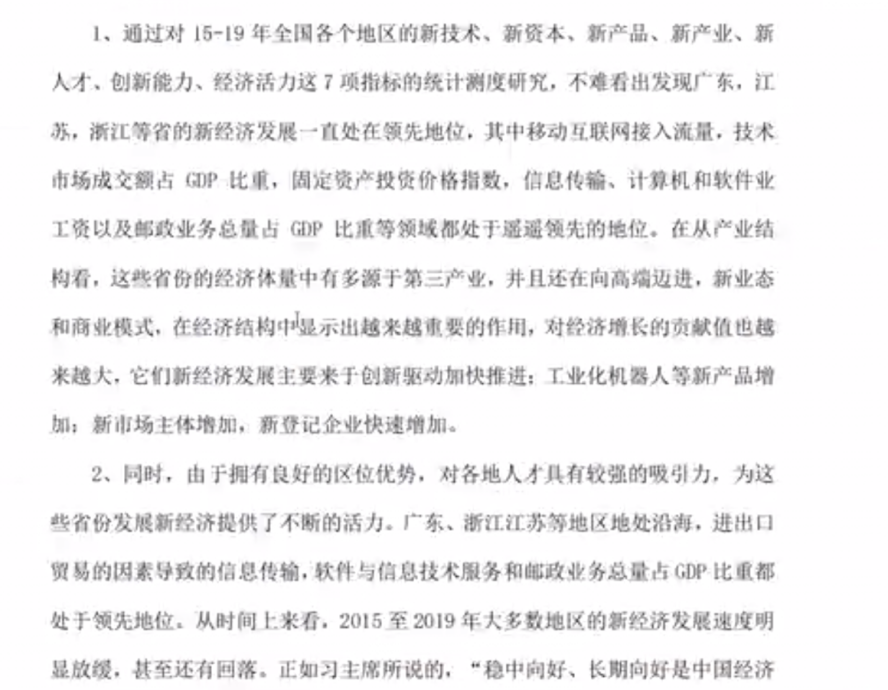
**数据来源**

**建模介绍**

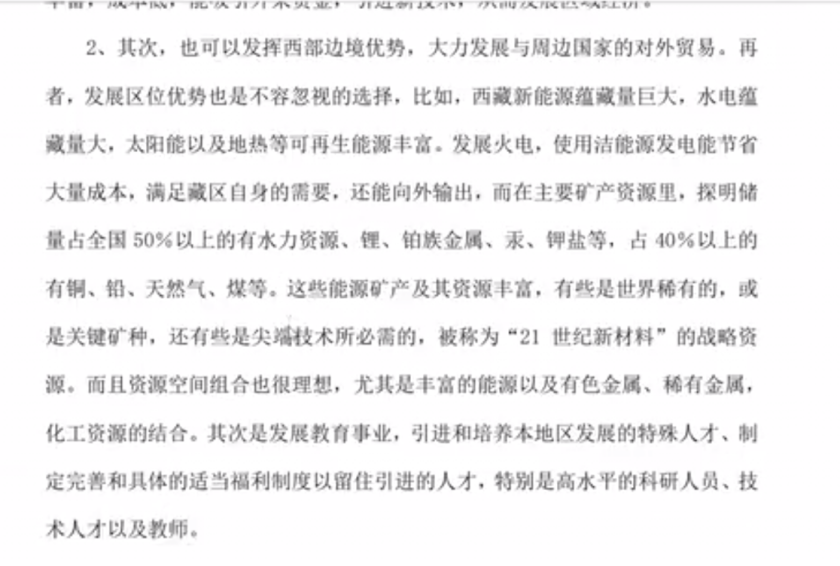
**结果解释（重要）**



**结论**



**建议**



**附录**

2.基于多源数据融合的个体商户信用评估研究（机器学习）

选题背景（目前问题及如何解决）及意义

研究内容及创新

研究流程图推荐



代码尽量不要放在正文里

关于比赛：

统计创新应用 数据引领未来

应用领域的创新，解决社会问题，有社会价值和意义

坚持下去！

选题70%重要

论文写作水平和模型方法，决定部分下限

选题：模仿，

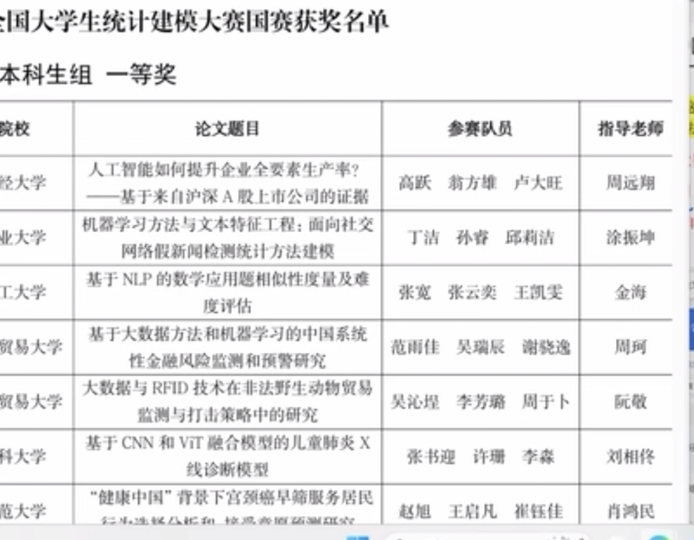
统计：计量经济、实证

ai是潮流

情感分析自然语言处理nlp 机器学习

新质生产力需要细化

中国各省人工智能综合活力指数测度研究--基于AHP-CRITIC混合加权法（方法旧，问题新）



避雷：碳中和、环境

关注前沿论文：国内高水平期刊，高水平SCI期刊，两会热词、国内政策etc（经济类可行性较高）当年新领域、新方向 结合地区，结合学校优势学科

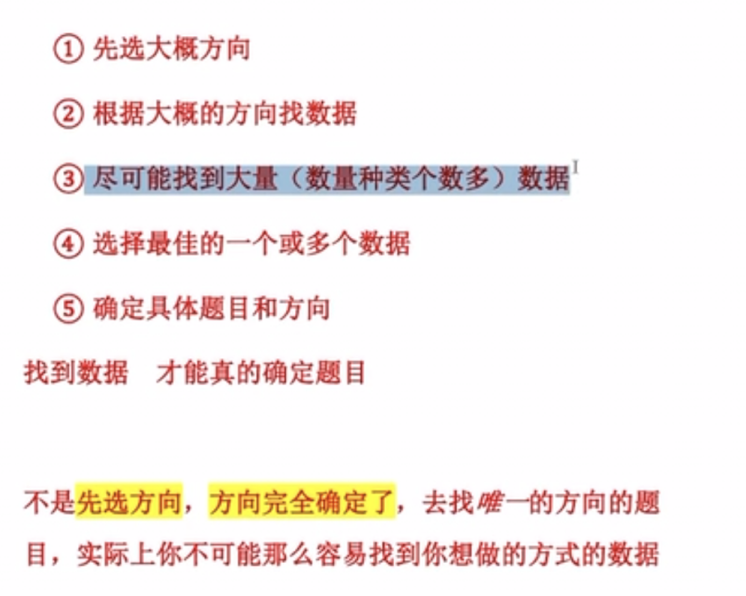


题目结合十四五、、写大一点

提振消费（公司企业，eg消费提振对消费意愿度提振的效果

科技（高效 科研经费支出





评价定性问题：哪个方案最好，怎么评价这个好坏

分类识别判断问题：病患识别，台风路径方向识别，洪水情况判断

预测回归问题：天气、经济、二氧化碳、新质生产力、等预测

相关性、影响分析问题：影响房价因素、影响AIGC发展因素