

鲸析

# 资源分享

## Data Science Repository



鲸析



中国大陆



扫一扫上面的二维码图案，加我微信

鲸析



缺氧的小鲸鱼🐳

小红书号：1821360719

🎓美国乔治城大学理学硕士  
✈️中国商飞数据管理中心数据分析师...

小红书

扫描二维码  
在小红书找到我



# Content

1. *Intro*
2. *Math & Stats*
3. *Programming Languages*
4. *Data Collection*
5. *Data Preparation*
6. *Feature Engineering*
7. *Data Modelling*
8. *Data Science Projects*
9. *Interview Resources*
10. *Courses & Cheat Sheets*
11. *Favorite Channels*

## 1. Intro

大家好，我是知识渊博、爱好小酌、嗓音独特，还挺幽默的鲸鲸！

首先，我想说鲸鲸不是专家，鲸析也不是个神奇的数据人才基地。

鲸析的文化是「**终身学习，追求真实的快乐，那便是恒久的努力！**」。

我会在这里分享我的经验、以及从中获得的理解（understanding）和见解（insight）给支持我的粉丝们，所以如果你觉得有用，请帮鲸鲸分享！

鲸鲸本科是【数学与应用数学】专业，没有一点 Python 和 R 的基础，但是一个月速成 Python 什么的，绝不可能，要想真正用 Python 做数据分析、数据科学，并且用这个来吃饭的话，请做一个以月为周期的学习目标。

我希望你认识到：

- 数据分析、数据科学的技术栈的学习不可能一蹴而就。
- 数学、统计很重要！
- Python 要好好学习
- *Don't be a coder, be a solver.*

从这些资料中，你会充分了解各种机器学习算法、数据科学、数据分析以及实用工具的细节。

因此，我建议在学习机器学习或数据科学之前先从【数学&统计】开始。如果你对微积分和积分、线性代数和统计学没有基本的了解，理解各种算法背后的原理是不可能的。

同样，如果对 Python 完全没有了解的话，那可能这篇分享对你用处不大，不必浪费时间。

## 鲸析

在此资源中，您将找到我在整个数据科学之旅中创建和发现的资源库，我认为这是尽可能简单地解释概念的最佳资源。

如果你觉得这些东西有用，不妨了解一下鲸析的实战项目。

- Kaggle DS 端到端实战项目【Porto Seguro Safe Driver Prediction】

7 个阶段 45 天带你玩转数据科学！

**项目介绍：**老司机最痛苦的事就是为高额保险费买单，构建筛选有效因子为车险定价。

**项目仓库：**关注【鲸析】公众号，后台输入：[safe driver prediction](#) 获取公开仓库。

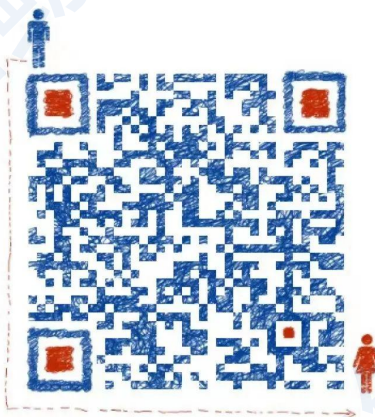
- JHU CSSE DA 端到端实战项目【Covid-19 Data Analysis & Time Series Prediction】

14 天带你玩转数据分析！

**项目介绍：**奥密克戎肆虐，上海沦陷，对比美国疫情现状，探讨上海到底是否应该动态清零。

**项目仓库：**关注【鲸析】公众号，后台输入：[上海疫情](#) 获取公开仓库。

## Contact me



扫一扫上面的二维码图案，加我微信

鲸析

## 2. Math & Stats

实话实说，数学和统计是不得不打好的基础，或许说，这是你能够在这条路走多远的关键，如果只想用数据分析混口饭吃，那么请跳过，看下一章节。

### Calculus & Algebra

当涉及到【概率分布】和【假设检验】时，积分是必不可少的。虽然我不是在说你要成为这方面的 big guy，但是不要忽视。

- [Introduction to Calculus](#) (Video)
- [Introduction to Linear Algebra](#) (Video)
- [Scipy Tutorial](#) (article)
- [Practice Questions](#) (Questions)

### Statistics

如果你非要让我在统计和上面的内容里面二选一的话，那我建议你直接开始搞统计这一块的内容。

统计相比起来更贴合我们的实际应用情况，也更能和实际的数据分析、数据科学的包（package）结合使用，上手更快，而且并没那么难理解！

推荐一些资料给大家！

请在【鲸析】后台回复：**数学统计书单**，即可获取下载链接！

### 3. Languages

之前所提到的内容是在理论层面我们所需要打好的基础，那么如何将以上所说的内容在实际当中应用，或者说体系化，程序化。我们要借助一些工具。

但是，请不要神化这些工具，认为编程能力才是你从事数据分析、数据科学行业的核心，恰恰相反，你需要借助这些工具辅助你去解决问题。

让我们来看看都需要哪些**技术栈**吧！

#### SQL

这个不必多说了，不管你是做 da 还是 ds，你都得掌握 sql，一般一个面试官考考你 sql 就知道你大概的编程水平了，因为编程讲求逻辑，而 sql 就是一个你只要逻辑没问题就能写出来的东西，**所以，你必须要会 sql。**

#### Learn SQL

- [SQL for data science in Coursera](#)
- [SQL for data science in DataCamp](#)
- [SQL from zero to hero in Bilibili](#)
- [Top 20 SQL interview questions NOT SQL QUERIES](#)

#### Practice SQL

- [Leetcode](#)
- [nowcoder](#)
- [Case Studies](#)

## Python

python 这里我们只谈论与数据分析、数据科学相关的部分，因为 python 过于强大，这里不会涉及过于广泛。如果是作 da 的话，numpy/pandas/matplotlib 就可以了，ds 的话就需要更多 sklearn/seaborn/scipy 甚至更多机器学习、深度学习框架 Tensorflow/pytorch 等内容。

推荐大家关注我的 github 账号！

里面有很多 python 在数据分析、数据科学里的辅助学习资料（notebook）。

<https://github.com/datoujinggzj>

请关注我的公众号：**鲸析**

在【**鲸析课堂**】中可以找到 **Numpy** 和 **Pandas** 的相关教程哦！

- [Learn Pandas with Kaggle.](#)
- [Pandas 100 Questions](#)
- [Python for Data Science](#)

## R Language

R 语言是统计学专业的学生经常用的工具，尤其是生物医药相关的 da 和 ds 用的会多一点，R 语言的 tidyverse、ggplot2 以及一些统计检验的包还是很好用的，不幸的是，出了校园，还是 Python 的世界。

- [R for Data Science](#)
- [R for Time Series Prediction](#)
- [Hands on Machine Learning in R](#)
- [ggplot2 Cheatsheet](#)

## 4. Data Collection

恭喜你，能看到这里，说明你已经具备了做一个 da 或者 ds 项目了，那么为了避免你像一只无头苍蝇一样去做项目，我建议你先梳理一下数据分析的[实现流程](#)。

在数据层面，我们要做的第一步就是**数据获取**，其难度是很大的，如何找到适配的数据是我们要攻克的第一个难题！

1. [Get data from public sources](#)
2. [Get data via web scraping](#)
3. [Get data via APIs](#)

## 5. Data Preparation

关于数据准备，其实包含了很多内容，这一环节让我们从 almost understand the data 转变到 completely understand the data。

比如：

- Data cleaning
  - [Data Manipulation in Pandas](#)
  - [Handling Missing Values](#)
  - [Handling Outliers](#)
- [MetaData](#)
- [Imbalanced Data Processing](#)
- [EDA](#)

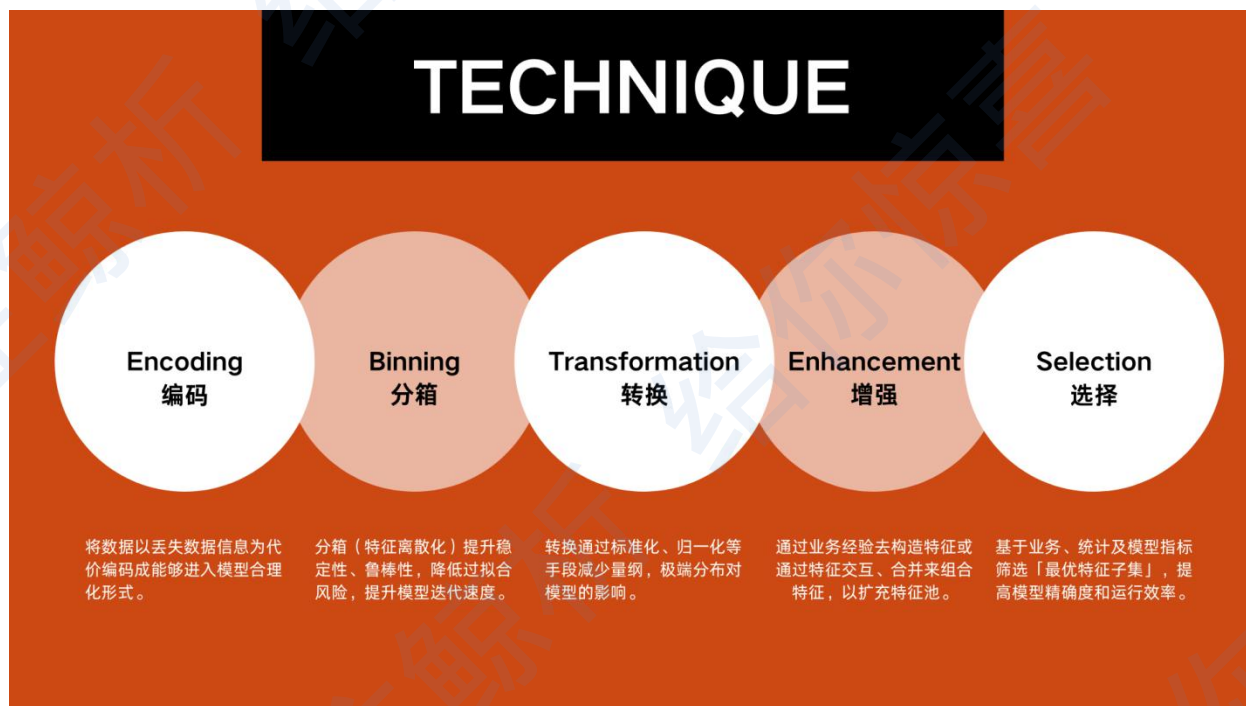
[An Extensive Step By Step Guide for Data Preparation](#)



## 6. Feature Engineering

这一步是成功的关键，让我们从对数据的 understanding 过渡到 insights。我们通过各种手段去挖掘、构建、合并、转换、交互因子，以不断筛选精英因子，为 feed 入模型做最后铺垫。

考虑到企业 PB 级数据的回溯成本，在有限资源和合适的调度下，我们会更倾向把侧重点和绝大部分精力放在【特征工程】上，而不是模型的迭代优化上。



- Feature Encoding
- Data Discretization (Binning)
- Data Transformation
- Data Enhancement
- Data Selection

[A Short Guide for Feature Engineering and Feature Selection](#)

## 7. Data Modelling

### Machine Learning Algorithms

- Linear Regression
  - [Comprehensive Summary](#)
  - [Mathematical Explanation](#)
  - [Video](#)
- Logistic Regression
  - Comprehensive Summary [coming soon]
  - [Video](#)
- K-Nearest Neighbours
  - Comprehensive Summary [coming soon]
  - [Video \(In-depth explanation\)](#)
- Decision Trees [coming soon]
  - [Video](#)
- Naive Bayes
  - [Comprehensive Summary](#)
  - [Video](#)
- Support Vector Machines (SVMs)
  - Comprehensive Summary [coming soon]
  - [Python Implementation](#)
  - [Mathematical Explanation](#)
- Neural Networks

## 鲸析

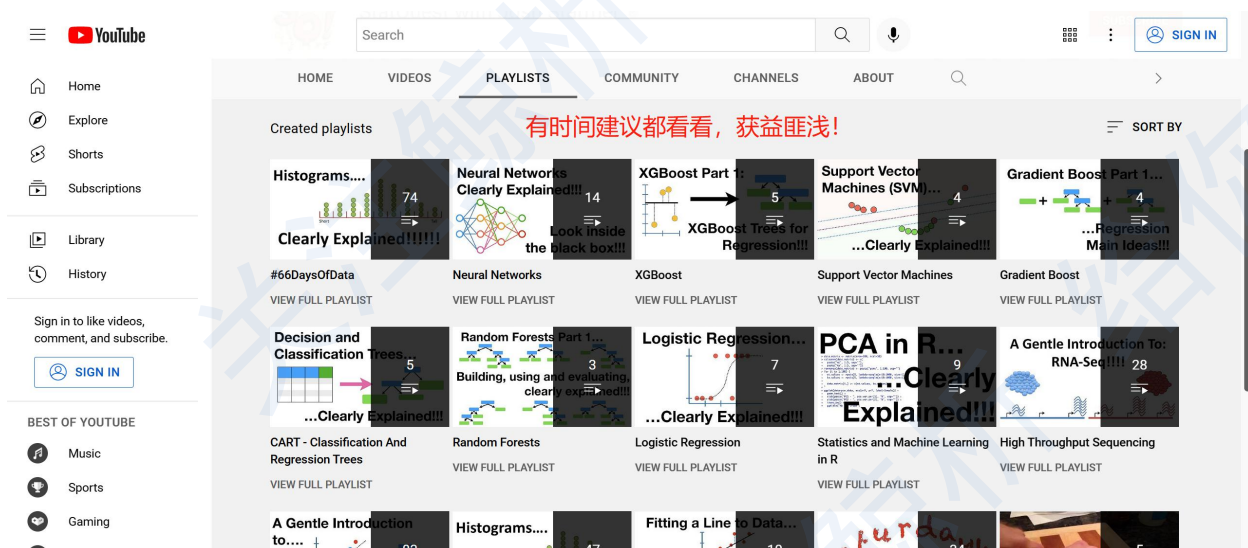
- [Comprehensive Summary](#)
- Random Forests
  - [Video](#)
- AdaBoost
  - [Comprehensive Summary](#)
  - [Video](#)
- Gradient Boost
  - [Video](#)
- XGBoost
  - [Video](#)
- K-Means Clustering
  - [Article](#)
- Hierarchical Clustering
- EM Algorithm
  - [Article](#)
- Hidden Markov Model (HMM)
  - [Article](#)
- AutoEncoder
  - [Article](#)
- Convolutional Variational AutoEncoder
  - [Article](#)
- Principal Component Analysis (PCA)
  - [Video](#)

# Fundamental Machine Learning Concepts

- Bias and Variance Tradeoff
- Regularization
- Confusion matrix and relevant metrics
- AUC and ROC

StatQuest 的 machine learning foundation 是个很好的选择，点击[这里](#)。

- Bootstrap Sampling
- Ensemble Learning, Bagging, and Boosting
- Scaling vs Standardization vs Normalization



## 8. Data Science Projects

### Data Science Projects

















- Porto's Seguro Safe Driver Prediction ([Video](#), [Github](#)) [HARD]
- Shanghai Covid-19 Data Analysis ([Video](#), [Github](#)) [MEDIUM]
- AB testing in Udacity ([Github](#)) [MEDIUM]
- Titanic - Machine Learning from Disaster ([Kaggle](#)) [EASY]
- House Prices - Advanced Regression Techniques ([Kaggle](#)) [EASY]
- Digit Recognizer ([Kaggle](#)) [EASY]
- 5 Data Analytics Projects for Beginners ([Article](#)) [EASY]

### Interview Prep Resources

### Interview Questions and Answers

- Top 50 DS interview Q&A ([Article](#))
- Amazon data scientist interview: the only post you'll need to read ([Article](#))
- Top Amazon Data Scientist Interview Questions and Answers ([Articles](#))
- 数据分析师精选面经合集 ([Article](#))
- 《数据分析面试汇总，持续更新中。。。》 ([Article](#))
- 挫折中成长——数据分析师面经 ([Article](#))

## Courses and Cheat Sheets

 DL_cheatsheet.pdf
 Keras_Cheat_Sheet_Python.pdf
 ML_cheatsheets_compressed.pdf
 Matplotlib.pdf
 Matplotlib_datacamp.pdf
 Numpy.pdf
 Pandas.pdf
 Pandas_datacamp.pdf
 Python_SciPy_Cheat_Sheet_Linear_Algeb..
 README_zh-hans.pdf
 SQL-Cheat-Sheet-websitesetup.pdf
 SQL_whale.pdf
 Seaborn.pdf
 Seaborn_datacamp.pdf
 mysql_cheat_sheet.pdf
 sql-basics-cheat-sheet-letter.pdf

关注【鲸析】，后台回复 **cheatsheet** 获取全部资料哦~

## Favorite YouTube Channels

- [StatQuest](#)
- [3Blue1Brown](#)
- [Luis Serrano](#)
- [Tech with Tim](#)