每日导学 - Day1

> If you change nothing, nothing will change.

Hi，各位同学们，大家好，欢迎踏上数据分析的~~不归路~~列车~在发车之前，有几点需要给大家强调一下：

课程和项目：

- 请使用Chrome或者火狐浏览器。
- 课程以项目为导向，系统判定你是否通关的唯一标准就是是否通过项目
- 试学班以**体验式学习**为目的，是为了让大家**体验什么是数据分析**，**体验python编程**，**体验Udacity的课程形式，项目及审阅**等，学习曲线相对来说比较跳跃（正式班的学习曲线十分平缓），所以大家要秉着**按需知情**的最快学习原则，**先搞懂模块化的知识**，等正式班再去深究细节的知识
- **请时刻保持信心**，我也是从一个小白一步步走过来的，~~我这么菜的人如今都能当助教了？~~所以请大家一定要对自己有**信心**
- **诚信原则**，对于项目，大家一定要重视！切不可抄袭，一定要**对自己负责任**
- **善始善终**，既然花了钱来体验，那就要有一个体验的结果，相信大家都能顺利毕业~
- **保持沟通**，不要害羞，有问题直接来问我呀~，~~我很闲的~~早八点至晚十点都是我的响应时间~

你的学习周如何度过？

- 希望你每天都能抽出一到两个小时，一暴十寒远不如细水长流，养成学习的习惯~~每天不敲点代码手痒痒~~
- 我每天22点前会把次日的**每日导学**（即学习纲要）发到微信群，大家通勤路上~~上厕所~~无聊的时候可以拿出来瞅瞅，晚上下班直接学起来~
- 我每天11点前会把当天的**每日一题**（刁难你们的小问题）发到微信群，大家记得在群里**回答对应自己学号的问题**，另外，有些小问题并不是课程中的，而是需要你们自己去搜索解决的，锻炼你们的搜索能力，期待你们的答案~**每日一题的答案**会在每天22点左右放出~
- **周六晚上的优达日公开课**请一定要腾出时间来（~~如果有约会的话，不如拉上另一半来看优达日？~~），重点就是**带大家过一遍项目**

如何提问：

在课程学习中难免会遇到问题，请按照以下流程进行问题提问：

- **课程知识问题**：
  - 先自行查找问题答案，参考：谷歌/必应搜索、菜鸟教程、CSDN、stack**overflow**
  - 若问题未解决，请将**问题**及其**所在课程章节**发送至微信群，并@助教-Allen即可
- **非课程知识问题**：

  比如账号登录、课程加载等问题，请详细描述问题，反馈给班主任即可；

  有关于后续正式课程、服务的疑问和选课建议，联系你的学习规划师就行~

该说的也说了个差不多，那么，请大家~~系好安全带~~做好觉悟：

- 工作可能会很忙，但是每天至少要抽出一个小时来坚持学习，否则一日废，日日废
- 准备好一个笔记本或者有道云笔记、EverNote这种电子笔记本，用来记录学习笔记/问题
- 自律，自信！

发车！

今日目标

今天不需要接触代码（先舒一口气），只是了解一下**统计学基础**中的基础和**数据分析的基本流程**~

- 学习课程：**数据类型和统计基础**（大概半小时左右）和**数据分析过程**（大概二十分钟左右）
- 每日一题：群内回复每日一题~

知识清单

数据和统计基础

> - 了解数据的概念和重要性，
> - 掌握数据的分类，能举出一些生活中的实例

数据类型可以分为两大类：数值型数据和分类型数据；

数值型数据又可以分为连续型和离散型；

分类型数据又可以分为定序型和定类型。

数据类型		
数值:	**连续**	**离散**
	身高、年龄、收入	书中的页数、院子里的树、咖啡店里的狗
分类:	**定序**	**定类**
	字母成绩等级、调查评级	性别、婚姻状况、早餐食品

针对不同的数据类型要用不同的方法进行分析描述，用不同的可视化图像进行展示。比如说对连续型数据分析时，我们应该使用平均数、分位数、标准差等等进行描述，使用直方图或者箱线图进行可视化；但对于定类型数据而言，我们会分类统计数量，使用柱状图或者饼状图进行可视化。

所以，数据类型算是基础中的基础，是你之后进行分析和可视化的重要依据。

描述统计学基础

> 了解集中趋势测量的三种方式即可。

本节内容对如何数值数据和分类数据进行了概述，并摘选了数值数据中的集中趋势测量进行了详细讲解。

- 数值数据的分析

1. `Center` 集中趋势测量
    - `Mean` 均值 ：即数据的平均值
    - `Median` 中位数：即将数据按照从小至大的顺序排列，对于奇数个数据来说是最中间位置的那个值，对于偶数个数据来说是最中间位置的那两个值的均值。
    - `Mode` 众数 ：即数据组中出现次数最多的那个值。有可能**无众数**（所有数值出现的次数相同），也有可能**多众数**（有多个数值出现相同的最多次）

> 以下为拓展，简单了解下即可，相关概念可自行搜索查阅
>
> 2. `Spread` 离散程度测量
>
>   - 极差：即最大值与最小值之差
>   - 四分位差：第三四分位数与第一四分位数之差
>   - 方差
>   - 标准差
>
> 3. `Shape` 数据的形状（需配合直方图）
>
>   - 左偏态
>   - 右偏态
>   - 对称分布（通常是正态分布）
>
> 4. `Outliers` 异常值：一般为大于最大值或小于最小值1.5倍四分位差的数值，可通过箱线图观察。

理解数据分析过程

> 了解数据分析的基本~~套路~~流程即可。

数据分析不是从上至下一蹴而就的过程，而是需要你不断迭代、重复、完善，最终得到结论的过程。

提出问题

- 数据集中的各个变量之间的相关性如何？是否存在某些联系？
- 变量的统计结果会揭示什么？
- 根据现有掌握的数据，能否对未来走势进行预测？
- 根据你想了解的问题，去收集数据，再对问题进行修缮，如此**迭代**，获取更全面的数据，提出更一阵见血的问题。

  ...

整理数据

- 收集

  数据库提取？直接下载？网络爬虫？

- 评估

  这个过程是对数据产生直观印象的过程，你要尝试了解数据集的大小，基本的统计结果，是否存在数据重复？缺失？数据类型是否正确？是否每个变量成一列&每个观察值成一行？数据是否有统计错误？(严重偏离正常值，比如说气温达到70℃等等)...

- 清理

  对评估出的问题进行逐项排查、清理，直至获取到干净的数据（推荐超级有用且经典的Tidy Data，虽然代码用的是R语言，但代码不就只是工具而已嘛，关键的是**思维方法**）

探索性数据分析

即课程中提到的EDA(Exploratory data analysis)，这是一种分析数据集——尤其是陌生数据集——的方法，具体实施的话可以采用定量、定性的数据分析或者是可视化分析。

这是一个强调**迭代**的过程，在这个阶段你要不断的对数据进行探索（提问、整理、分析、可视化等等），根据你得到的结果再去丰富你的数据或者完善你的问题，最终得出结论。

这是一个考验耐心和细心的繁琐过程，所以一定要**心平气和**，保持**工作的连贯性**。（~~不做完一套不能睡觉？~~）

得出结论

- 通过可视化直接得出结论（描述、总结）
- 统计学（预测）
- 机器学习算法（主要是用来做预测）

传达结果

撰写报告，和别人分享你的研究结果，所以一定要逻辑清晰、结论都要有根有据，让被分享者信服你的结论。

我们会在后天开始的试学班项目——**实战：分析北上广空气质量**中接触到部分数据分析过程，大家到时候体验一下~

最后

第一天的知识理论性较多，大家可以多联系一下自己实际生活或者工作中的数据，加深对理论知识的理解。

明天开始，我们就要接触一点点python编程啦，别太激动，晚上早点休息~