Python 数据分析与数据挖掘

一、课程背景

互联网的飞速发展伴随着海量信息的产生，而海量信息的背后对应的则是海量数据如何从这些海量数据中获取有价值的信息来供人们学习和工作使用，这就不得不用到大数据挖掘和分析技术。数据分析作为大数据技术的核心一环，其重要性不言而喻。

在数据分析领域，Python 语言以其简单易用，并提供了优秀、好用的第三方库和数据分析的完整框架而深受数据分析人员的青睐。可以说，Python 已经当仁不让地成为了数据分析人员的一把“利器”。程序员想要进入数据分析行业，首先要掌握 Python 数据分析技术，只有这样才能在严峻的就业市场中具有较强的竞争力。

二、课程收益

通过本课程学习，达到如下目的：

- 了解 Python 使用场景，能够搭建自己的编程开发环境。
- 掌握 Python 编程的基础语法知识、精髓及编程思想。
- 掌握常用的第三方扩展库的使用，特别是文件夹处理、word excel ppt 文件处理；
- 学会使用 Python 提升职场常见办公场景的工作效率，如邮件自动化、网络爬虫。
- 了解 Numpy 库多维数组的创建、切片和索引方法，以及数组的运算和存取。
- 学会使用 Pandas 库完成数据的导入导出、数据整理和数据多角度分析的方法。
- 学习正则表达式及如何爬取网络数据进行数据分析。
- 学会使用 Matplotlib 模块绘制常用图表和高大上图表，以及如何与 EXCEL 联动；
- 了解机器学习概念，会使用 Sklearn 模块进行线性回归、逻辑回归的分析方法。

能结合课程学习到的方法和工具对本职工作中遇到的场景进行针对性分析。

三、培训对象

本课程适用于职场从事数据分析或和数据分析工作相关的职场人士。

四、 授课讲师

张晓如 老师（微软 OFFICE 大师级认证）

五、 授课时间

4 天（6 小时/天）。

六、课程大纲

PartI、数据分析理念

***了解数据分析的方法、工具和流程。**

1、什么是数据分析？

- 概念与目的

发现数据规律——找到可行方案——指导管理决策。

- 数据分析三阶段

描述性分析，发生了何事

诊断性分析，为何发生

预测性分析，将发生何事

2、数据分析方法

- 对比分析
- 同比分析
- 环比分析

- 回归分析
- 聚类分析
- 时间序列分析

3、数据分析的工具

- 常规工具 VS 高大上工具

4、数据分析流程

1） 步骤 1：明确目的

- 确定分析目的：要解决什么样的业务问题
- 确定分析思路：分解业务问题，构建分析框架

2） 步骤 2：数据收集

- 明确收集数据范围
- 确定收集来源
- 确定收集方法

3） 步骤 3：数据预处理

- 数据质量评估
- 数据清洗、数据处理和变量处理

4） 步骤 4：数据分析

- 选择合适的分析方法
- 构建合适的分析模型
- 选择合适的分析工具

5） 步骤 5：数据展示

- 选择恰当的图表
- 选择合适的可视化工具

6） 步骤 6：报表撰写

- 选择报告种类
- 完整的报告结构

Part2、Python 环境搭建

***搭建自己的 Python 编程开发环境。**

1、 认识 Python 与环境搭建

7） What——什么是 Python

8） Anaconda 如何下载、安装与配置

9） IDLE **VS** Jupyter Notebook **VS** Spyder

5、Python 初体验——十秒钟快速创建 100 个 Excel 工作薄并统一命名

1） Spyder 界面介绍

2） Python 文件的打开、编辑与保存

3） 案例：认识一下 Python 代码的整体构成

6、模块的类别、安装、导入

- 内置模块
- 第三方模块
- 用 PIP 命令安装、卸载、升级模块
- Import 语句导入模块
- From 语句导入模块

实战：搭建并配置自己的 Python 运行环境。

Part3、Python 编程基础

***掌握 Python 编程思想、编程语句、数据结构。**

1、 语法特点

- 缩进
- 注释
- PEP8 编写规范

2、 变量

- 变量的赋值
- 变量命名规则

3、 数据类型

- 数值型：整型与浮点型
- 字符型：字符型的定义
- 逻辑型：1 和 0，或 TRUE 和 FALSE
- 数据类型的查询：TYPE 函数
- 数据类型的运算：数值型/字符型/逻辑型如何运算
- 数据类型的转换：Str()函数、int()函数、float()函数

7、数据结构

- 列表（LIST）：如何定义/访问/增加/修改/删除
- 字典(DICTIONARY)：如何定义/访问/增加/修改/删除
- 元组：如何定义/访问
- 集合：如何定义/访问

8、流程控制语句

- If 语句——选择结构

- For 语句——循环结构
- While 语句——循环结构
- 循环结构中的 break 语句和 continue 语句
- 控制语句的嵌套

9、函数

- 常用内置函数：print()input()replace()/strip()/split()open()……
- 如何自定义函数:def 语句

10、编程中的异常处理

练习：基本 Python 编程语句实战操作。

Part4、NumPy 入门与实战

***学习 NumPy 库对多维数组的创建、切片和索引方法，以及数组的运算和存取。**

1、 ndarray 多维数组

- 创建 ndarray 多维数组
- Ndarray 的对象属性、数据类型及变换

11、数组的索引和切片

- 数组索引方法
- 数组切片方法

12、数组的运

- 数组和标量间的运算
- 数组的条件逻辑运算
- 统计运算

- 数组内如何排序

Part5、数据预处理

***学习 Pandas 库和 xlwing 库对文件的读写操作、数据整理的方法。**

1、 pandas 数据结构

- Series 对象：如何定义/访问/增加/修改/删除
- DataFrame 对象：如何定义/访问/增加/修改/删除

13、读、写数据

- 读、写文本文件
- 读、写 Excel 文件
- 读、写数据库数据
- 读、写网页

14、数据操作

- 数据的增、删、改、查
- NaN 数据处理
- 时间数据的处理
- 数据的抽取：字段拆分、记录抽取、随机抽样

15、数据的预处理

- 处理缺失值
- 去除重复数据
- 处理异常值
- 合并数据：追加合并、匹配合并
- 数据标准化：0-1 标准化

16、数据的分组与聚合

- 数据分组
- 数据聚合

17、使用 xlwing 库批量处理工作薄/工作表/行/列(EXCEL)

- 批量新建、保存、关闭工作薄
- 批量打开一个文件夹下的所有工作薄
- 批量重命名一个工作薄中的工作表名称
- 批量打印工作薄中的指定工作表/指定页
- 按条件将 EXCEL 中的多个工作表合并为一个工作表
- 按条件将 EXCEL 中的一个工作表拆分为多个工作薄

案例实操：超市交易数据清洗、查看员工业绩波动、分析员工业绩。

Part6、Pandas 模块数据分析

***学习 Pandas 中常用的数据分析方法。**

1、 基础数据分析方法

- 批量升序/降序排序一个工作薄中的所有工作表
- 使用描述统计呈现数据的相关指标（如平均值、极值、%分位值、峰度系数、偏度系数等）

18、进阶数据分析方法

- 制作数据透视表进行交叉分析
- 分组对比分析（定性分组与定量分组）
- 使用相关系数判断数据的相关性
- 数据建模回归分析

19、时间序列分析

- Datetime 模块的时间数据类型
- 如何把字符型转为时间型数据
- 时间序列如何索引和切片数据
- 如何创建介于某时间区间的时间数据（天/月/固定天数）

案例实操：超市交易数据清洗、查看员工业绩波动、分析员工业绩。

Part7、爬取网络数据进行分析

***学习正则表达式及如何爬取网络数据进行数据分析。**

1、 认识网页结构和网页源代码

- 查看源代码
- 查看网页结构（区块/列表/标题/链接/元素）

20、正则表达式

- 认识普通字符和元字符
- 使用正则表达式提取数据

21、Request 模块获取网页源代码

22、Selenium 模块获取网页源代码

23、Selenium 模块模拟鼠标和键盘操作

24、爬虫实战：

- 爬取某网站图书销量排行榜数据并分析
- 爬取某网站关于某关键词的实时新闻数据
- 爬取新闻热点排行榜

Part8、Sklearn 机器学习与数据挖掘

*了解机器学习概念，掌握线性回归、逻辑回归的分析方法。

1、 机器学习基本概念

25、机器学习库 sklearn 简介.

- 扩展库 sklearn 常用模块与对象.
- 选择合适的模型和算法

26、线性回归算法的原理与应用

- 线性回归模型的原理.
- sklearn 中线性回归模型的简单应用+
- 岭回归的基本原理与 sklearn 实现
- 使用线性回归模型预测儿童身高

27、逻辑回归算法的原理与应用

- ..逻辑回归算法的原理与应用 sklear 实现
- 使用逻辑回归算法预测.考试能否及格

28、朴素贝叶斯算法的原理与应用

- 基本概念..
- 朴素贝叶斯算法分类的原理与 sklearn 实现
- 使用朴素贝叶斯算法对中文邮件进行分类...

29、案例：

使用线性回归分析对销售收入进行分析和预测

30、案例：

使用 Pandas、sklearn 模块对客户价值进行分析

七、特别注意

学员自备电脑（建议一人一台），老师讲解示范后学员操作练习；

本课程内容及顺序可能根据学员需求及难度而调整。