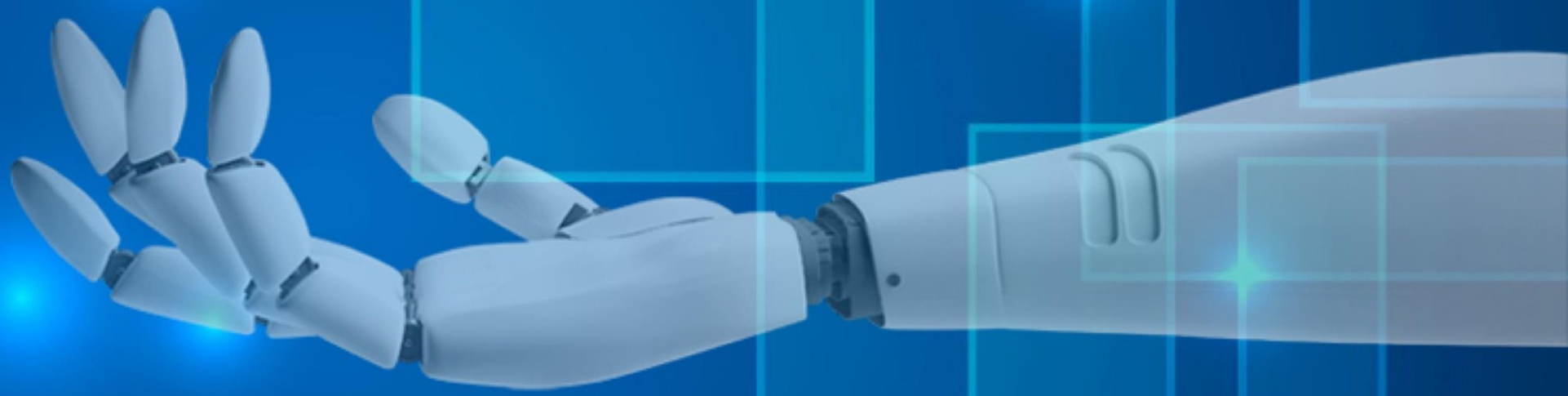
Tedu.cn 达内教育

达内教育研究院

数据分析与可视化

搭建环境



主流的数据分析编程语言

目前主流的数据分析语言有R，Python，MATLAB三种程序语言。

以理服人

	R	Python	MATLAB
语言学习难易程度	入门难度低	入门难度一般	入门难度一般
使用场景	数据分析，数据挖掘，机器学习数据可视化等。	数据分析，机器学习，矩阵运算科学数据可视化，数字图像处理web应用，网络爬虫，系统运维等。	矩阵计算，数值分析，科学数据可视化，机器学习，符号计算，数字图像处理，数字信号处理，仿真模拟等。
第三方支持	拥有大量的Packages，能够调用C，C++，Fortran，Java等其他程序语言。	拥有大量的第三方库，能够简便地调用C，C++，Fortran，Java等其他程序语言。	拥有大量专业的工具箱，在新版本中加入了对C，C++ Java的支持。
流行领域	工业界≈学术界	工业界 > 学术界	工业界≤学术界
软件成本	开源免费	开源免费	商业收费

数据分析编程语言

Tedu.cn 达内教育

- 人工智能、数据分析首选语言—— Python
- 任何人可学，终身受益

眼见为实

1、高效获取及处理互联网海量信息

2、可满足不同业务需求，轻松实现数据可视化

3、搭建多种数据模型，进行数据建模分析，实现业务的智能化分析，做出研究/评估/预测。

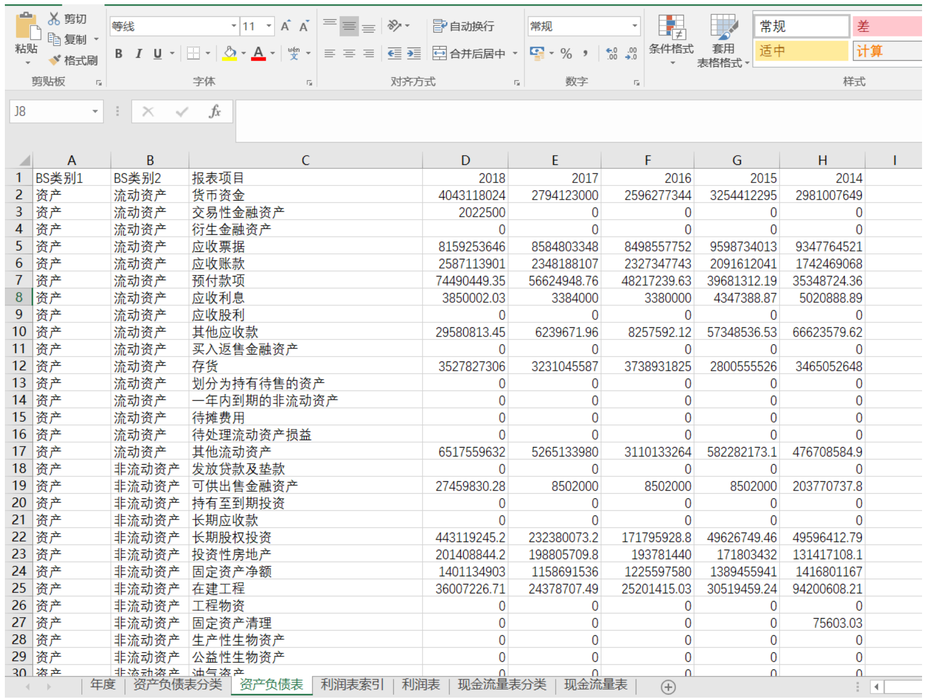
数据分析编程语言

Tedu.cn 达内教育

- 工作高效，流程可控

  - 使用Python代码控制数据分析工作流程简易高效，相较于传统Excel或绘图工具，修改时只需要调整参数即可，即可获得分析结果或可视化效果，而不用大量繁琐易错的手工操作。

眼见为实



	A	B	C	D	E	F	G	H	I
1	BS类别1	BS类别2	报表项目	2018	2017	2016	2015	2014	
2	资产	流动资产	货币资金	4043118024	2794123000	2596277344	3254412295	2981007649	
3	资产	流动资产	交易性金融资产	2022500	0	0	0	0	
4	资产	流动资产	衍生金融资产	0	0	0	0	0	
5	资产	流动资产	应收票据	8159253646	8584803348	8498557752	9598734013	9347764521	
6	资产	流动资产	应收账款	2587113901	2348188107	2327347743	2091612041	1742469068	
7	资产	流动资产	预付款项	74490449.35	56624948.76	48217239.63	39681312.19	35348724.36	
8	资产	流动资产	应收利息	3850002.03	3384000	3380000	4347388.87	5020888.89	
9	资产	流动资产	应收股利	0	0	0	0	0	
10	资产	流动资产	其他应收款	29580813.45	6239671.96	8257592.12	57348536.53	66623579.62	
11	资产	流动资产	买入返售金融资产	0	0	0	0	0	
12	资产	流动资产	存货	3527827306	3231045587	3738931825	2800555526	3465052648	
13	资产	流动资产	划分为持有待售的资产	0	0	0	0	0	
14	资产	流动资产	一年内到期的非流动资产	0	0	0	0	0	
15	资产	流动资产	待摊费用	0	0	0	0	0	
16	资产	流动资产	待处理流动资产损益	0	0	0	0	0	
17	资产	流动资产	其他流动资产	6517559632	5265133980	3110133264	582282173.1	476708584.9	
18	资产	非流动资产	发放贷款及垫款	0	0	0	0	0	
19	资产	非流动资产	可供出售金融资产	27459830.28	8502000	8502000	8502000	203770737.8	
20	资产	非流动资产	持有至到期投资	0	0	0	0	0	
21	资产	非流动资产	长期应收款	0	0	0	0	0	
22	资产	非流动资产	长期股权投资	443119245.2	232380073.2	171795928.8	49626749.46	49596412.79	
23	资产	非流动资产	投资性房地产	201408844.2	198805709.8	193781440	171803432	131417108.1	
24	资产	非流动资产	固定资产净额	1401134903	1158691536	1225597580	1389455941	1416801167	
25	资产	非流动资产	在建工程	36007226.71	24378707.49	25201415.03	30519459.24	94200608.21	
26	资产	非流动资产	工程物资	0	0	0	0	0	
27	资产	非流动资产	固定资产清理	0	0	0	0	75603.03	
28	资产	非流动资产	生产性生物资产	0	0	0	0	0	
29	资产	非流动资产	公益性生物资产	0	0	0	0	0	
30	资产	非流动资产	油气资产	0	0	0	0	0	

年度 | 资产负债表分类 | 资产负债表 | 利润表索引 | 利润表 | 现金流量表分类 | 现金流量表

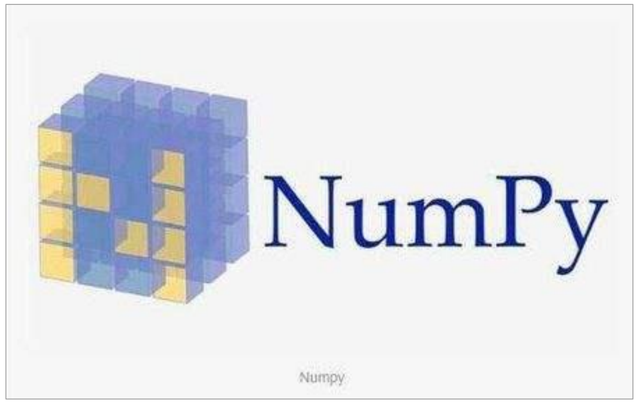
```python
# coding=utf-8
import threading
import MySQLdb
from datetime import datetime
import time
import smtplib
from email.mime.text import MIMEText
from log import logger


def get_con():
    host = "127.0.0.1"
    port = 3306
    logsdb = "logsdb"
    user = "root"
    password = "never tell you"
    con = MySQLdb.connect(host=host, user=user, passwd=password, db=logsdb, port=port, charset="utf8")
    return con


def calculate_time():

    now = time.mktime(datetime.now().timetuple())-60*2
    result = time.strftime('%Y-%m-%d %H:%M:%S', time.localtime(now))
```

数据分析编程语言

Tedu.cn 达内教育

- 工具库丰富，功能庞大、使用方便
  - Python提供丰富多样的工具库用于支持数据分析、交互和探索性计算，如科学计算库numpy，高级科学计算库scipy，数据分析库pandas、数据可视化库matplotlib、机器学习库sklearn等。

眼见为实



Python数据分析与可视化三件套

数据分析编程语言

Tedu.cn 达内教育

- 代码简洁、可读性强、可拓展

  - Python语法简洁，利用python编写的程序代码通常仅为其他语言（如Java、C++等）编写代码行数的五分之一；编程风格非常强调可读性

眼见为实

使用k均值构建模型

```
: import numpy as np
  import pandas as pd
  from sklearn.cluster import KMeans #导入kmeans算法
  k = 5 ## 确定聚类中心数
  #构建模型
  kmeans_model = KMeans(n_clusters = k,random_state=123)
  fit_kmeans = kmeans_model.fit(airline_features_scaled)    #模型训练
  #统计不同类别样本的数目
  r1 = pd.Series(kmeans_model.labels_).value_counts()
  print('最终每个类别的数目为：\n',r1)
  # kmeans_model.labels_[:20] #查看样本的类别标签
  kmeans_model.cluster_centers_ #查看聚类中心
```

```
最终每个类别的数目为：
 3    24618
4     15733
1     12114
2      5337
0      4242
dtype: int64
```

针对聚类结果进行特征分析，如图所示。

聚类类别	聚类中心				
	L	R	F	M	C
客户群1	0.483	-0.8	2.483	2.425	0.309
客户群2	0.056	0	-0.226	-0.229	2.2
客户群3	1.16	-0.38	-0.087	-0.095	-0.156
客户群4	-0.7	-0.42	-0.161	-0.161	-0.254
客户群5	-0.313	1.686	-0.574	-0.537	-0.173

基于特征描述，本项目定义五个等级的客户类别：重要保持客户，重要发展客户，重要挽留客户，一般客户，低价值客户。每种客户类别的特征如图所示。

python-最适合人工智能和数据分析的语言

以理服人

常用库			
序号	扩展库	简介	详情
1	Numpy	提供数组支持，以及相应高效的处理函数	python没有提供数组，列表list可以完成数组，但不是真正的数据，但数据量增大时，他的速度很慢，所以Numpy扩展包提供了数组支持，同事很多高级扩展包以来它，例如：Scipy，Matplotlib,Pandas.
2	Scipy	提供矩阵支持，以及矩阵相关的数值计算模块	提供矩阵支持，以及矩阵相关的数值计算模块，Numpy让python有了Matlab的味道，那么Scipy让python成为第二个Matlab
3	Matplotlib	强大的数据可视化工具，做图库，绘图，绘表	数据分析与可视化--公共基础课程
4	Pandas	数据分析扩展库，强大，灵活的数据分析和探索工具	面板数据（panel data），是python最强大的数据分析和检索工具，因金融分析工具而开发，支持类似SQL的数据增删查改，支持时间序列化，灵活处理缺失数据。
5	StatsModels	统计建模和计量经济学，包括描述统计，统计模型分析和推断	统计学是数学的一个分支，涉及数据的收集，分析，解析，呈现和组织
6	Scikit-Learn	支持回归，分类，聚类等的强大的机器学习扩展库	Scikit-Learn是一个基于python的用于数据挖掘和数据分析的简单而有效的东西，基本功能包括六个部分：分类，回归，聚类，数据降维，模型选择，数据预处理等
7	Keras	深度学习库，用于建立神经网络以及深度学习模型	在底层机器学习框架之上的高级API架构
8	Gensim	用来做文本主题模型的库，文本挖掘可能用的	自然语言课程：文本挖掘
9	OpenCV	开源的跨平台计算机视觉库	众多图像函数能直接对Numpy数组进行处理，编写图像处理，计算机视觉程序变得更加简洁可用于开发实时的图像处理，计算机视觉以及模式识别程序
10		三维可视化库VTK、复杂网络分析库igraph	

数据分析方法介绍

- 数据分析工具介绍
- 搭建数据分析环境

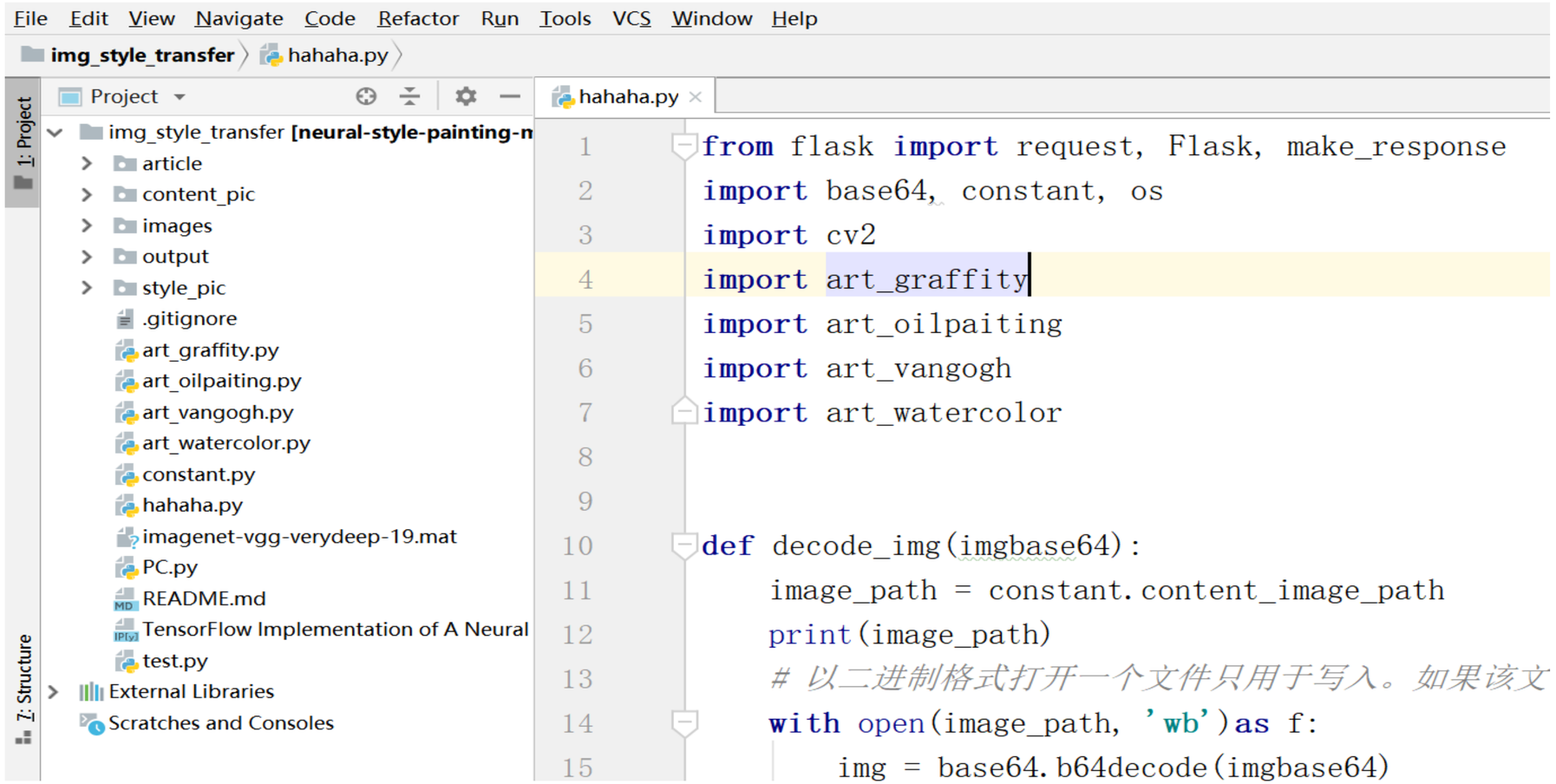
达内教育研究院

python编程工具介绍

以理服人

安装python编程环境，推荐使用编程工具

1. pycharm

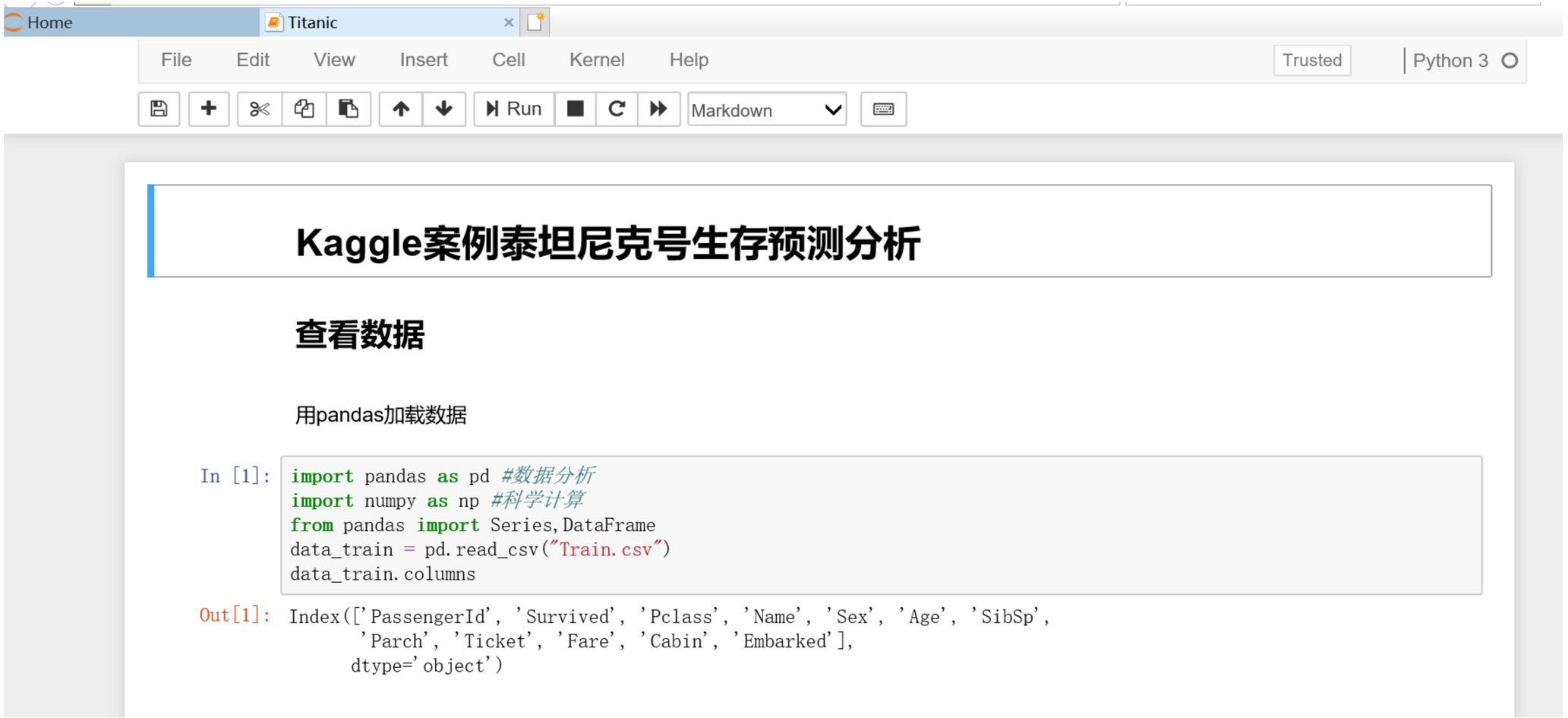


python编程工具介绍

Tedu.cn 达内教育

推荐使用编程工具

2. Jupyter notebook



以理服人

python编程工具介绍

本阶段需要安装的库及安装方式（cmd中）

1. numpy　　pip install numpy
2. scipy　　pip install scipy
3. matplotlib　　pip install matplotlib
4. pandas
6. ipython　　~
7. Jupyter　　~
8. sklearn　　~

以理服人

Tedu.cn
达内教育

谢谢