

2015-11-19 大数据文摘



关于转载授权

大数据文摘作品，欢迎个人转发朋友圈，自媒体、媒体、机构转载务必申请授权，后台留言“机构名称+转载”，申请过授权的不必再次申请，只要按约定转载即可，但文末需放置大数据文摘二维码。

编译 | 黄念

校对 | 丁一

引言

艺术之美根植于其所传达的信息。有时候，现实并非我们所看到或感知到的。达芬奇（Da Vinci）和毕加索（Picasso）等艺术家都通过其具有特定主题的非凡艺术品，试图让人们更加接近现实。

数据科学家并不逊色于艺术家。他们用数据可视化的方式绘画，试图展现数据内隐藏的模式或表达对数据的见解。更有趣的是，一旦接触到任何可视化的内容、数据时，人类会有更强烈的知觉、认知和交流。

在数据科学中，有多种工具可以进行可视化。在本文中，我展示了使用Python来实现的各种可视化图表。

怎样才能Python中实现可视化？

涉及到的东西并不多！Python已经让你很容易就能实现可视化——只需借助可视化的两个专属库（libraries），俗称matplotlib和seaborn。听说过吗？

Matplotlib：基于Python的绘图库为matplotlib提供了完整的2D和有限3D图形支持。这对在跨平台互动环境中发布高质量图片很有用。它也可用于动画。

Seaborn：Seaborn是一个Python中用于创建信息丰富和有吸引力的统计图形库。这个库是基于matplotlib的。Seaborn提供多种功能，如内置主题、调色板、函数和工具，来实现单因素、双因素、线性回归、数据矩阵、统计时间序列等的可视化，以让我们来进一步构建复杂的可视化。

我能做哪些不同的可视化？

刚出版不久的《A comprehensive guide on Data Visualization》中，介绍了最常用的可视化技术。在进一步深入学习前，如果你尚未阅读此书，我们建议你参考此书。

以下是Python代码与其输出结果。我就是用下面的数据集来创建这些可视化的。

EMPID	Gender	Age	Sales	BMI	Income
E001	M	34	123	Normal	350
E002	F	40	114	Overweight	450
E003	F	37	135	Obesity	169
E004	M	30	139	Underweight	189
E005	F	44	117	Underweight	183
E006	M	36	121	Normal	80
E007	M	32	133	Obesity	166
E008	F	26	140	Normal	120
E009	M	32	133	Normal	75
E010	M	36	133	Underweight	40

导入数据集

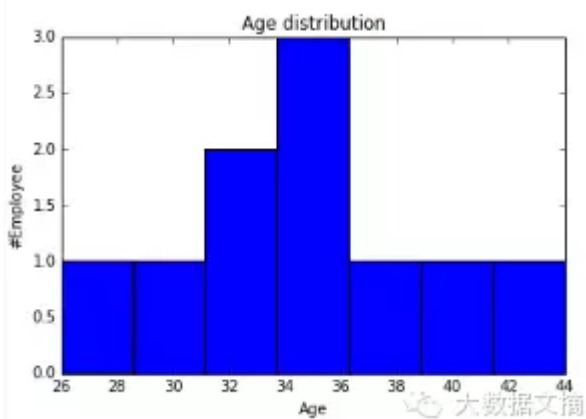
```
import matplotlib.pyplot as plt
import pandas as pd
df=pd.read_excel("E:/First.xlsx", "Sheet1")
```

大数据文摘

1. 直方图

```
fig=plt.figure() #Plots in matplotlib reside within a figure object, use plt.figure to create new figure
#Create one or more subplots using add_subplot, because you can't create blank figure
ax = fig.add_subplot(1,1,1)
#Variable
ax.hist(df['Age'],bins = 7) # Here you can play with number of bins
Labels and Tit
plt.title('Age distribution')
plt.xlabel('Age')
plt.ylabel('#Employee')
plt.show()
```

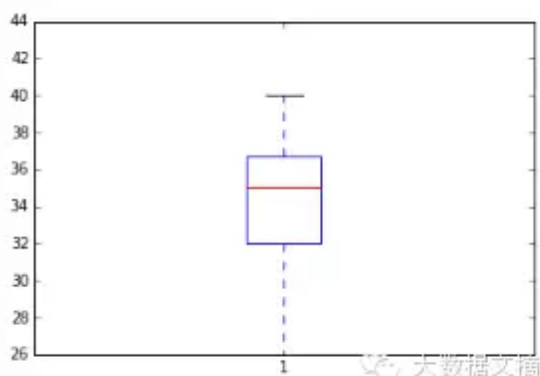
大数据文摘



2. 箱线图

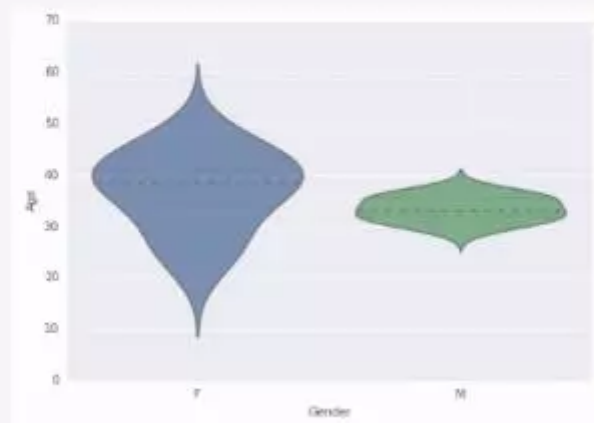
```
import matplotlib.pyplot as plt
import pandas as pd
fig=plt.figure()
ax = fig.add_subplot(1,1,1)
#Variable
ax.boxplot(df['Age'])
plt.show()
```

大数据文摘



3. 小提琴图

```
import seaborn as sns
sns.violinplot(df['Age'], df['Gender']) #Variable Plot
sns.despine()
```

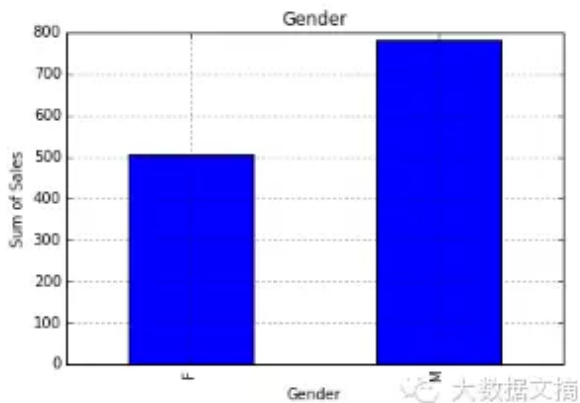


大数据文摘

4. 条形图

```
var = df.groupby('Gender').Sales.sum() #grouped sum of sales at Gender level
fig = plt.figure()
ax1 = fig.add_subplot(1,1,1)
ax1.set_xlabel('Gender')
ax1.set_ylabel('Sum of Sales')
ax1.set_title("Gender wise Sum of Sales")
var.plot(kind='bar')
```

大数据文摘



5. 折线图

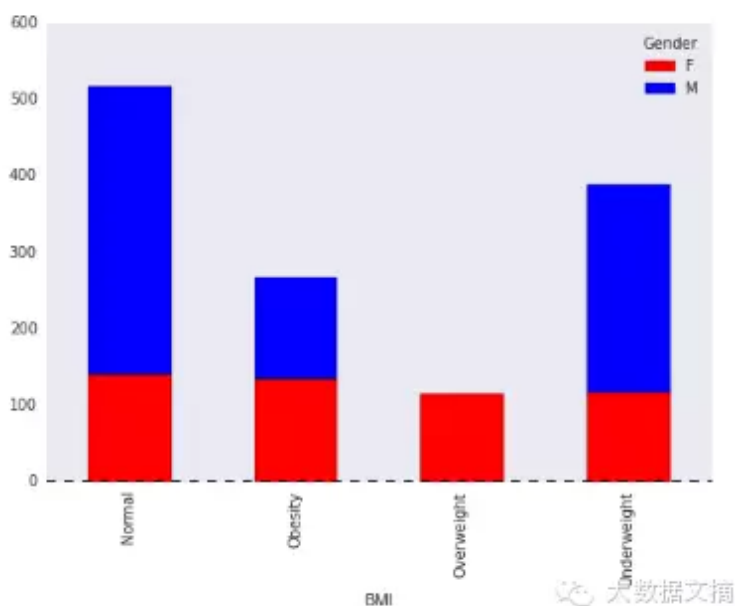
```
var = df.groupby('BMI').Sales.sum()
fig = plt.figure()
ax1 = fig.add_subplot(1,1,1)
ax1.set_xlabel('BMI')
ax1.set_ylabel('Sum of Sales')
ax1.set_title("BMI wise Sum of Sales")
var.plot(kind='line')
```

大数据文摘



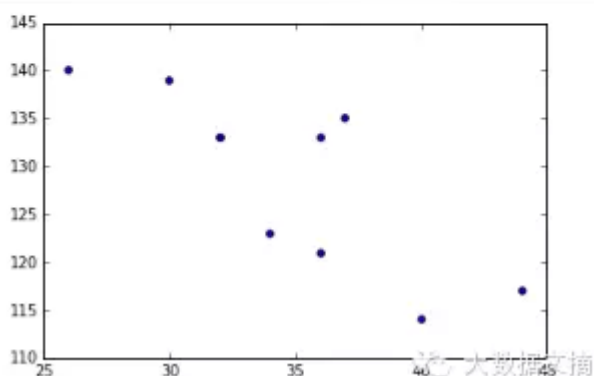
6. 堆积柱形图


```
var = df.groupby(['BMI', 'Gender']).Sales.sum()
var.unstack().plot(kind='bar', stacked=True, color=['red', 'blue'], id='false')
```



7. 散点图

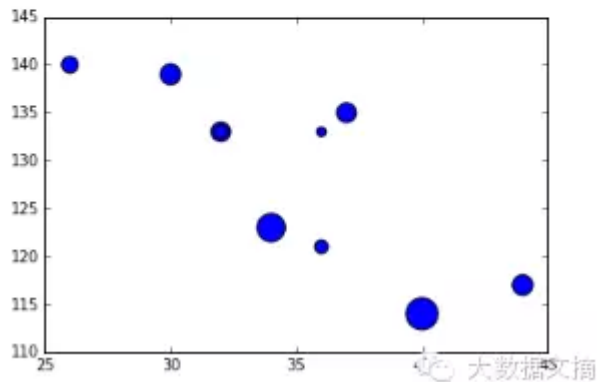
```
fig = plt.figure()
ax = fig.add_subplot(1,1,1)
ax.scatter(df['Age'], df['Sales']) #You can also add more variables here to represent
color and size.
plt.show()
```



8. 气泡图

```
fig = plt.figure()
ax = fig.add_subplot(1,1,1)
ax.scatter(df['Age'],df['Sales'], s=df['Income']) # Added third variable income as si
ze of the bubble
plt.show()
```

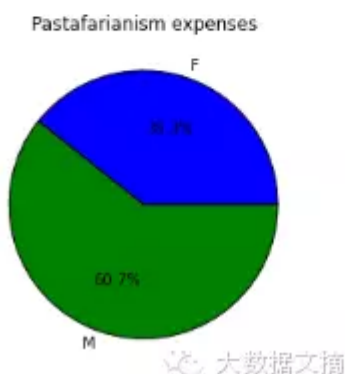
大数据文摘



9. 饼图

```
var=df.groupby(['Gender']).sum().stack()
temp=var.unstack()
type(temp)
x_list = temp['Sales']
label_list = temp.index
pyplot.axis("equal") #The pie chart is oval by default. To make it a circle use pyplo
t.axis("equal")
#To show the percentage of each pie slice, pass an output format to the autopctparame
ter
plt.pie(x_list,labels=label_list,autopct="%1.1f%%")
plt.title("Pastafarianism expenses")
plt.show()
```

大数据文摘

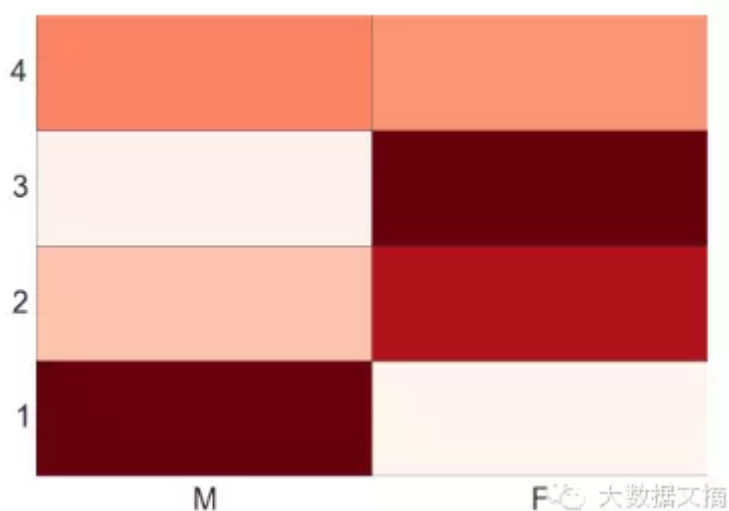


大数据文摘

10. 热图

```
import numpy as np
#Generate a random number, you can refer your data values also
data = np.random.rand(4,2)
rows = list('1234') #rows categories
columns = list('MF') #column categories
fig,ax=plt.subplots()
#Advance color controls
ax.pcolor(data,cmap=plt.cm.Reds,edgecolors='k')
ax.set_xticks(np.arange(0,2)+0.5)
ax.set_yticks(np.arange(0,4)+0.5)
# Here we position the tick labels for x and y axis
ax.xaxis.tick_bottom()
ax.yaxis.tick_left()
#Values against each labels
ax.set_xticklabels(columns,minor=False,fontsize=20)
ax.set_yticklabels(rows,minor=False,fontsize=20)
plt.show()
```

大数据文摘



你可以尝试绘制基于两个变量的热图，如X轴为性别，Y轴为BMI，数据点为销售值。

结语

现在，你肯定已经意识到了数据可视化的美妙，为什么不自己动手试试呢？在以后的文章中，我们还将探讨用Python实现地图可视化和词云。

大数据文摘也曾经发布过用R进行数据可视化的文章，《用R语言进行数据可视化的综合指南（一）》和《用R语言进行数据可视化的综合指南（二）》。大家可以参考一下，做个对比。


用R语言进行数据可视化的综合指南（一）

用R语言进行数据可视化的综合指南（二）



黄念

上海长海医院在读硕士，对生物医药大数据挖掘的及其应用很感兴趣，愿意借助本平台认识更多的小伙伴。


 大数据文摘

大数据文摘 大数据文摘 大数据文摘



丁一

杜克大学药理系在读博士，对生物信息学和临床药学的大数据挖掘很感兴趣。

 大数据文摘

【限时干货下载】

2015/11/30前

2015年10月干货文件打包下载，请点击大数据文摘底部菜单：下载等--10月下载

大数据文摘精彩文章：

回复【金融】 看【金融与商业】专栏历史期刊文章

回复【可视化】 感受技术与艺术的完美结合

回复【安全】 关于泄密、黑客、攻防的新鲜案例

回复【算法】 既涨知识又有趣的人和事

回复【谷歌】 看其在大数据领域的举措

回复【院士】 看众多院士如何讲大数据

回复【隐私】 看看在大数据时代还有多少隐私

回复【医疗】 查看医疗领域文章6篇

回复【征信】 大数据征信专题四篇

回复【大国】 “大数据国家档案”之美国等12国

回复【体育】 大数据在网球、NBA等应用案例

长按指纹，即可关注“大数据文摘”

专注大数据，每日有分享

覆盖千万读者的WeMedia联盟成员之一