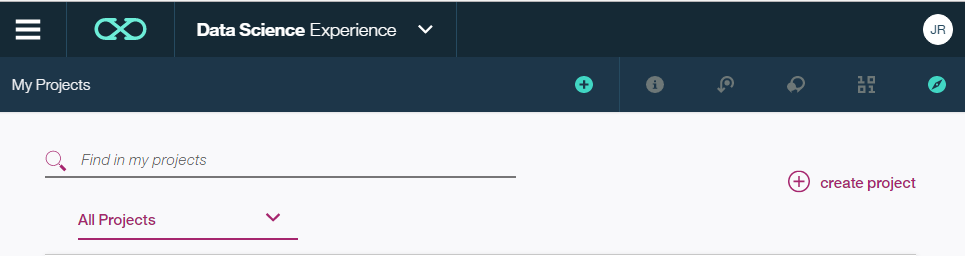
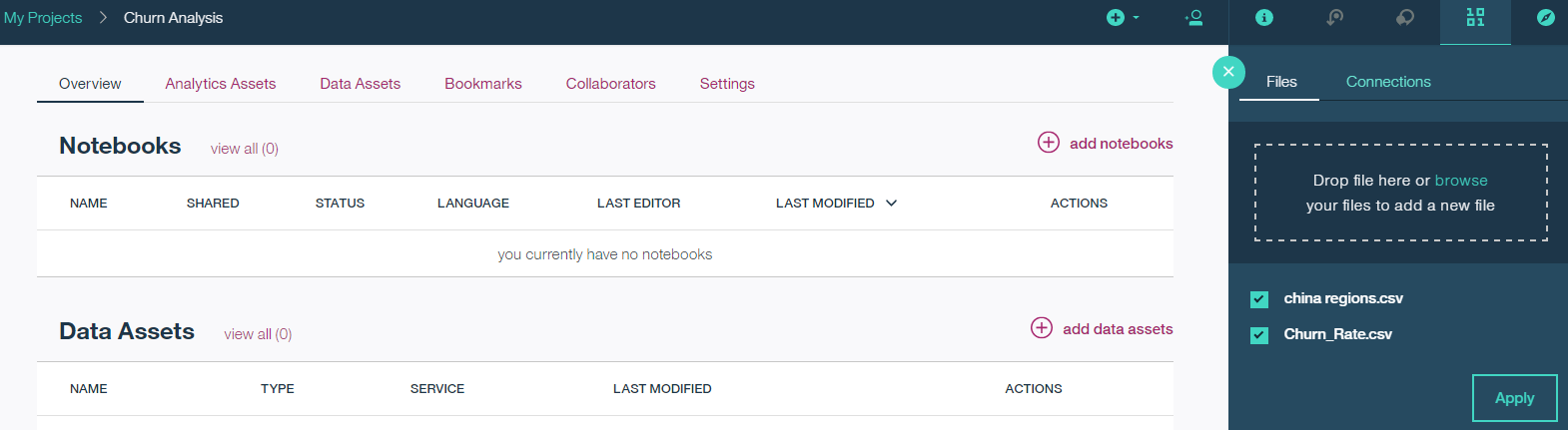
**创建notebook**

1. 登录<http://datascience.ibm.com/registration/stepone>

2. 创建工程，My Projects -> create project，在Name处输入工程名字，点击Create，稍等片刻



3.添加数据文件，点击右上角find and add data，把数据文件拖入右方或者点击browse添加文件，然后点击Apply



5.创建notebook，点击上图add notebooks, 在Name处输入notebook名字，选择你需要用的Spark version和Language（如果选择scala，spark version选1.6 ，如果选python，spark version不限），点击Creat Notebook.

**Python版本**

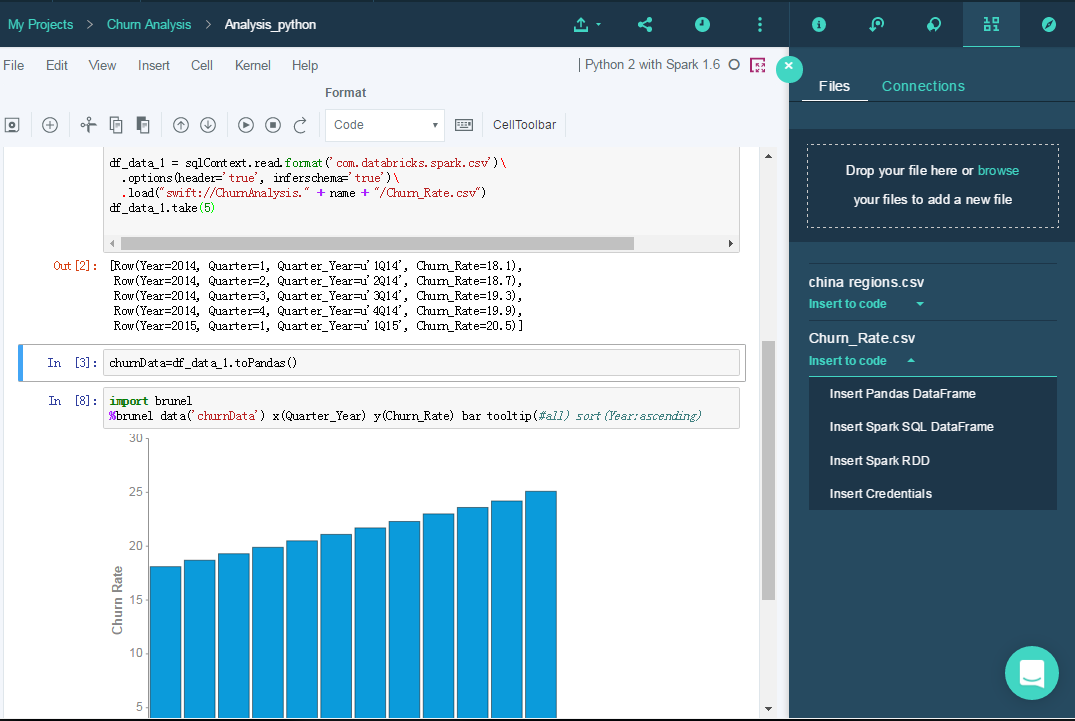
1. 导入数据，在notebook右侧点击创建工程时导入的数据文件，选择Insert Spark SQL DataFrame,则notebook出现下图代码，点击运行Run cell ，输出结果，则生成的df\_data\_1即为Spark DataFrame类型数据。

2.如果使用brunel分析数据，需要选用Insert Pandas DataFrame, 或者把Spark DataFrame类型数据用toPandas方法转成Pandas DataFrame类型

3.导入brunel, 输入Import brunel.

4.用brunel 画图，例：%brunel data('dfData1') x(QUARTER\_YEAR) y(CHURN\_RATE) bar tooltip(#all) sort(YEAR:ascending)

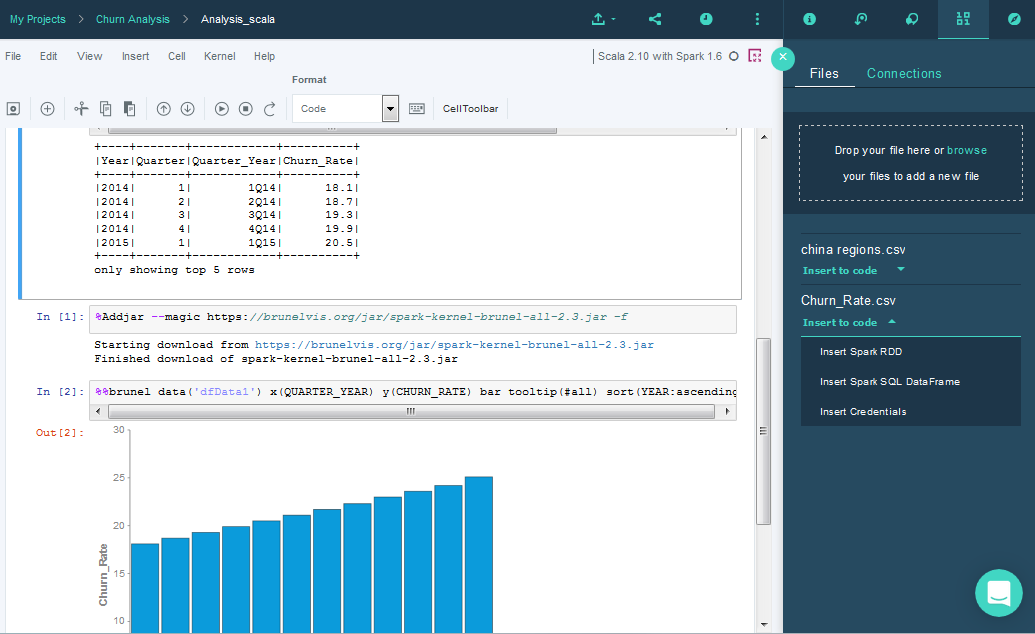
详情参考：Visualizationwith open source package brunel  
<http://datascience.ibm.com/blog/brunel-interactive-visualization-in-jupyter-notebooks-2/>



**Scala版本**

1. 创建工程时注意选择spark1.6版本，如果后续chrome看不到图像，可在chrom右键属性，在target处设置允许不安全内容，参考<https://superuser.com/questions/487748/how-to-allow-chrome-browser-to-load-insecure-content>。或者换用firefox等其他浏览器。

2. 导入数据，在notebook右侧点击创建工程时导入的数据文件，选择Insert Spark SQL DataFrame,则notebook出现下图代码，点击运行Run cell，输出结果，则生成的dfData1即为Spark DataFrame类型数据。

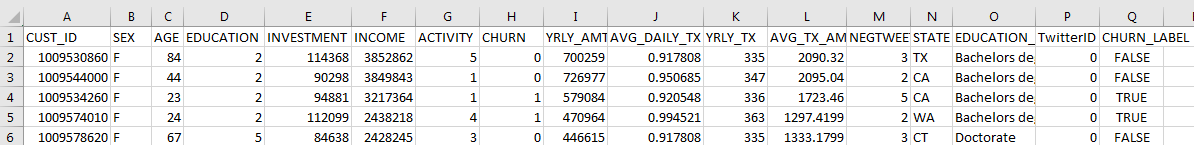


3.导入brunel, 输入%Addjar --magic https://brunelvis.org/jar/spark-kernel-brunel-all-2.3.jar -f, 运行

4.用brunel 画图，例：%%brunel data('dfData1') x(QUARTER\_YEAR) y(CHURN\_RATE) bar tooltip(#all) sort(YEAR:ascending)

详情参考：Visualizationwith open source package brunel  
<http://datascience.ibm.com/blog/brunel-interactive-visualization-in-jupyter-notebooks-2/>

**数据文件CUST\_SUM.csv的数据结构说明**



从左往右数，每列内容依次为：

信用卡用户ID、性别（string）、年龄、教育程度（Int）、一年内投资的费用、收入、社区活动、是否流失（Int）、累计交易额、日均交易笔数、一年总交易笔数、平均每笔交易金额、在社交平台上对银行所做的评价量、所在地区(string)、教育程度(string）、Twitter账户、是否流失（bool）

注：教育程度分Int和string, 是同样的意思。是否流失分bool和Int类型，也是同样的意思。