Power BI批量获取网页数据——阳光高考全国高校基础数据分析

我们日常利用PowerBI数据单个网页数据时非常简单，但是批量获取网页数据就显得相对麻烦一点。在这里我们可以用PowerBI的Power Query组件批量获取多个网页的数据。同样，也可以利用高版本Excel自带的Power Query进行获取。

本文以阳光高考网站为例，获取2019年度全国普通高等学校名单。

详细操作步骤如下：

（一）找到目标网页

（二）分析网址结构

（三）采集第一页的数据

（四）根据页码参数设置自定义函数

（五）批量调用自定义函数

（一）找到目标网页

打开阳光高考网站，找到“2019年度全国普通高等学校名单”网页。



点开单个网页查看，如下：



（二）分析网址结构

这里选取前三个省份网址

北京市：https://gaokao.chsi.com.cn/gkxx/zszcgd/dnzszc/201906/20190619/1799402422-2.html

天津市：https://gaokao.chsi.com.cn/gkxx/zszcgd/dnzszc/201906/20190619/1799402422-3.html

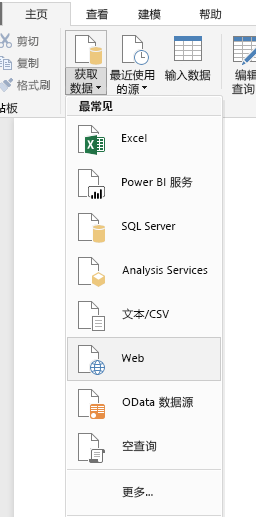
河北市：https://gaokao.chsi.com.cn/gkxx/zszcgd/dnzszc/201906/20190619/1799402422-4.html

可以看出网址中只有最后一个数字是变量，这里我们把它看做页码ID

（三）采集第一页的数据

（这里页码ID从北京市的“2”开始）

打开PowerBI Desktop，通过“获取数据”中的“web”选项获取数据，这里“web”界面选择“基本”选项卡即可。

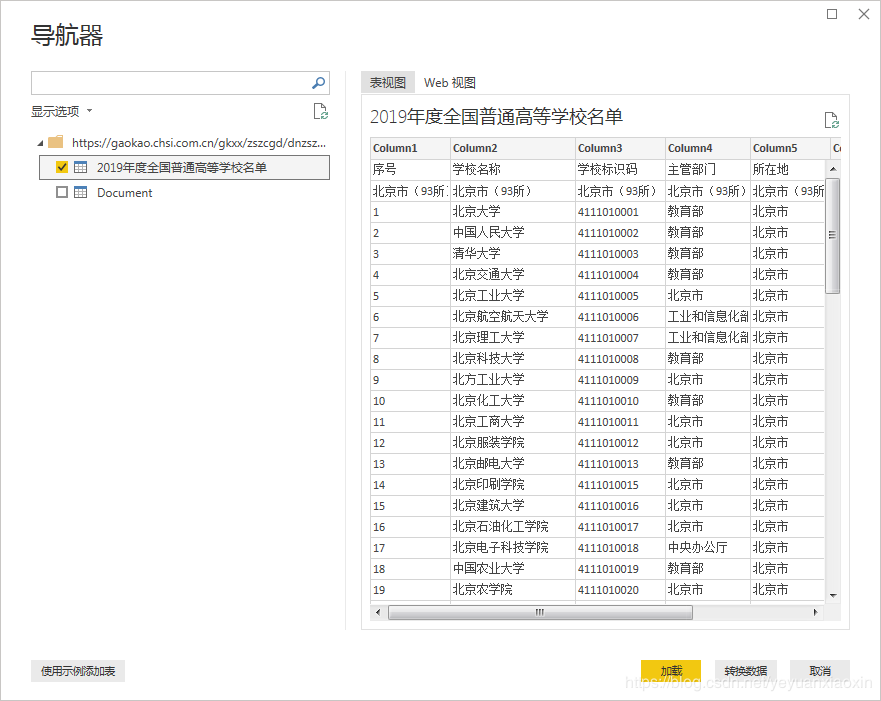


这里我们在基本选项卡中输入目标网址

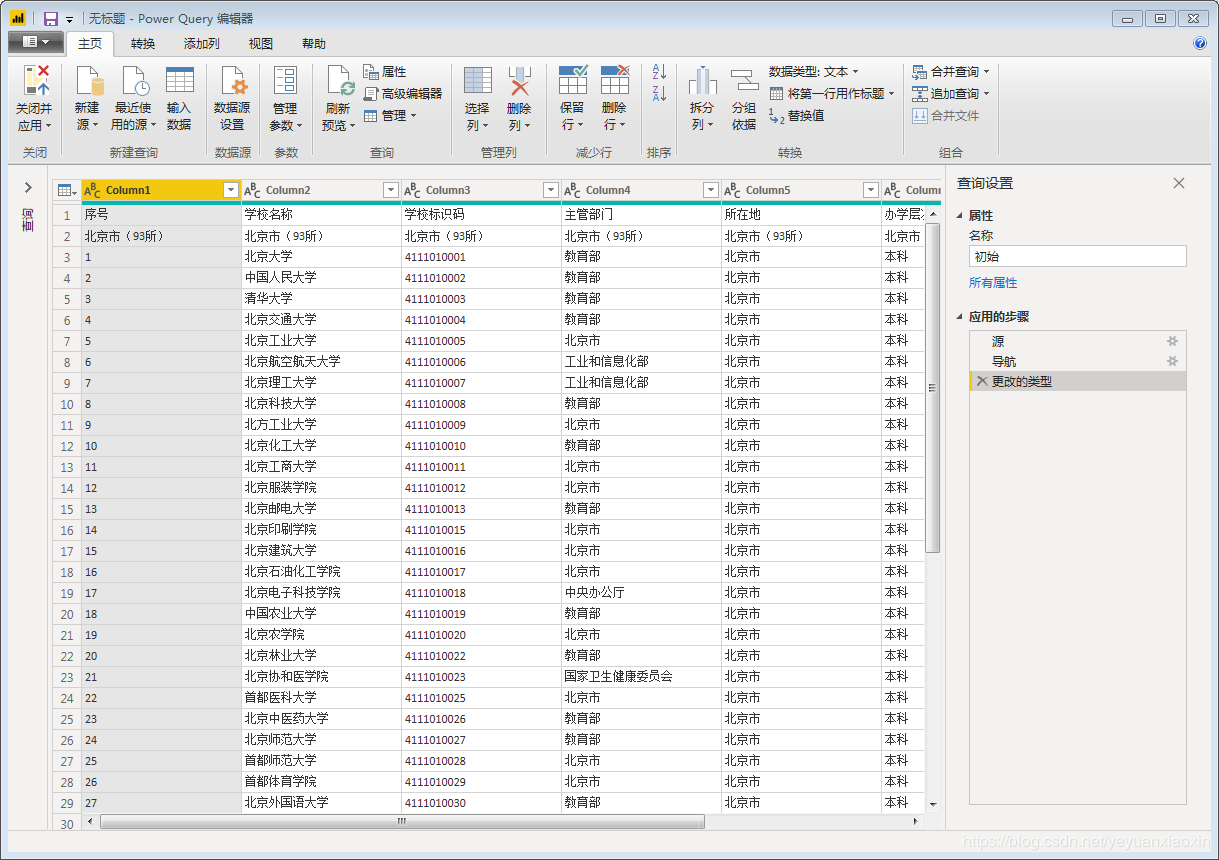
（北京市<https://gaokao.chsi.com.cn/gkxx/zszcgd/dnzszc/201906/20190619/1799402422-2.html>）



获取数据源信息如下，这里只有第一个表是我们想要的，勾选，然后点击右下角转换数据进行数据处理。



进入页面如下：



这样我们就简单地采集到第一页数据。然后对这一页的数据进行整理，删掉无用信息，添加字段名。整理好后，下面采集其他页面时，数据结构都会和第一页整理后的数据结构一致，采集的数据可以直接拿来用。

如果要大批量的抓取网页数据，为了节省时间，对第一页的数据可以先不整理，直接进入下一步。

这里我们先不处理。

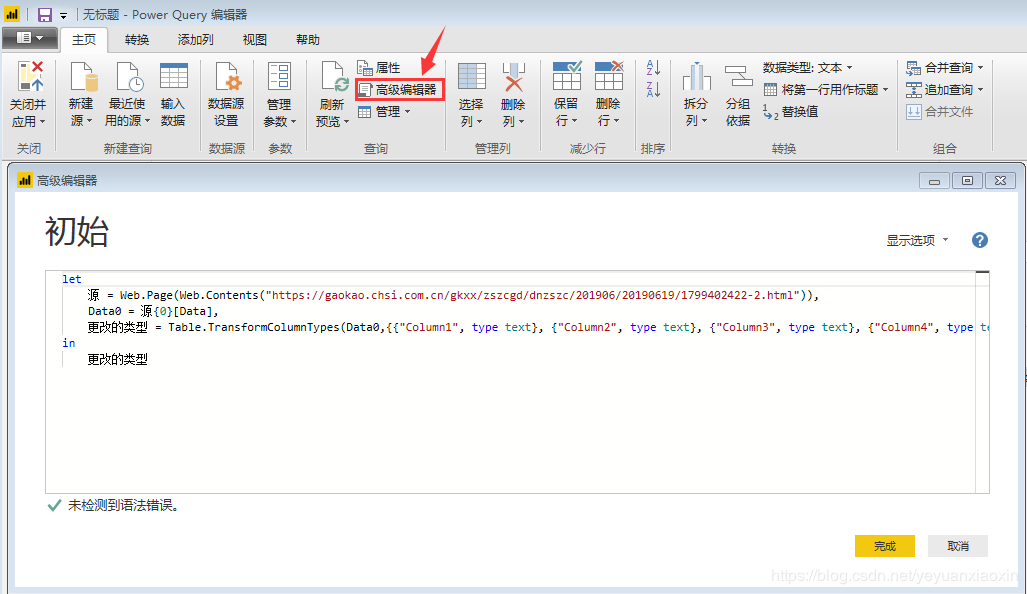
（四）根据页码参数设置自定义函数

这是最重要的一步

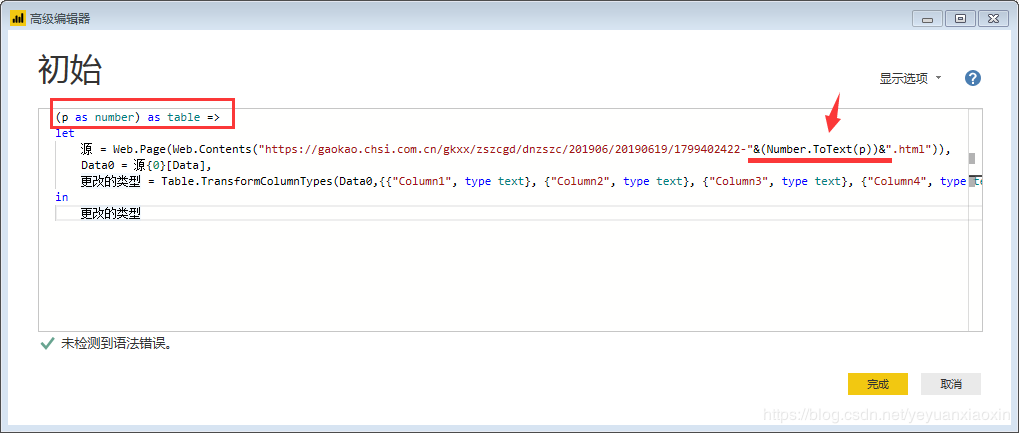
在当前数据的编辑器窗口，打开【高级编辑器】，在let前输入：

(p as number) as table=>

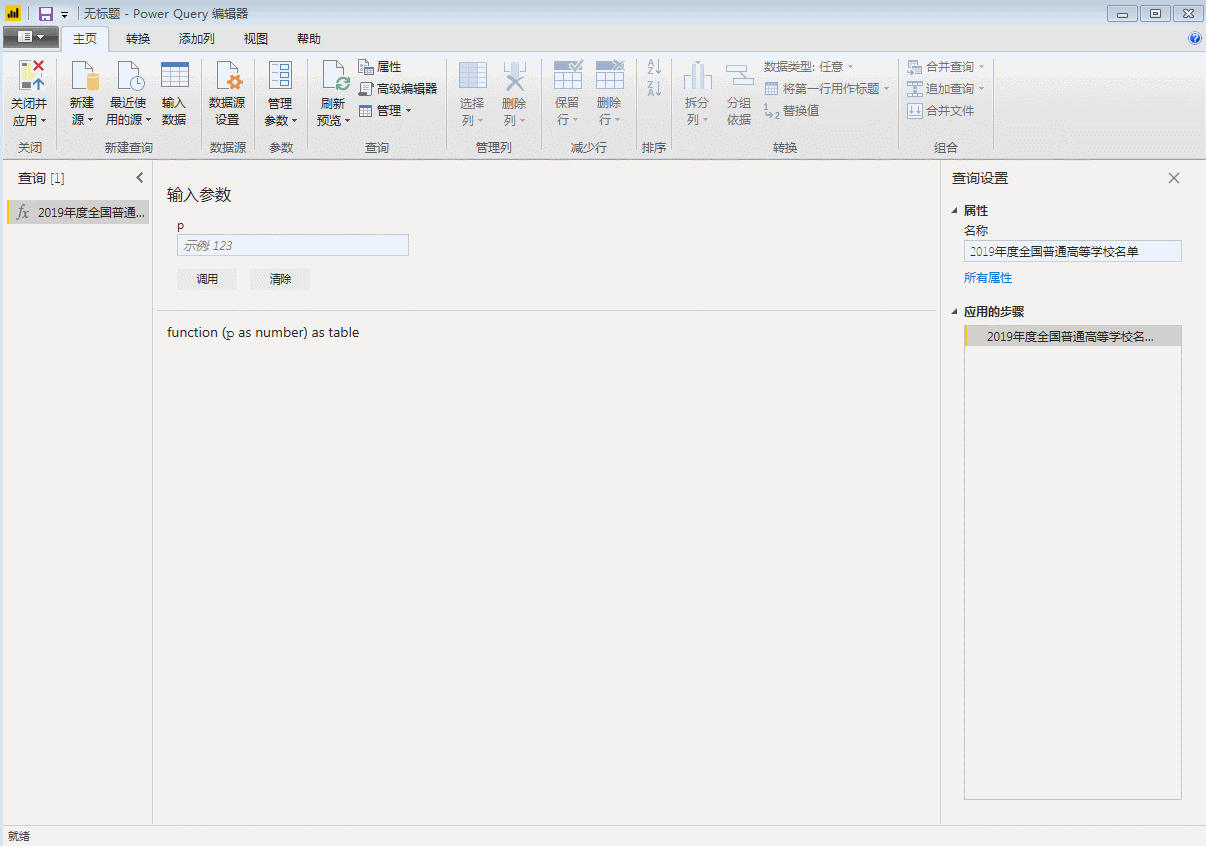
并且在链接中把网页页码，即上文提到的“1,2”等数字修改为"&Number.ToText(p)&"。



更改后变为：



点击“完成”，刚才第一页数据的查询窗口直接变成了自定义函数的输入参数窗口，表格也变成了函数的样式。自定义函数完成，p是该函数的变量，用来控制页码，**随便输入一个数字，比如3，将抓取第3页的数据**。



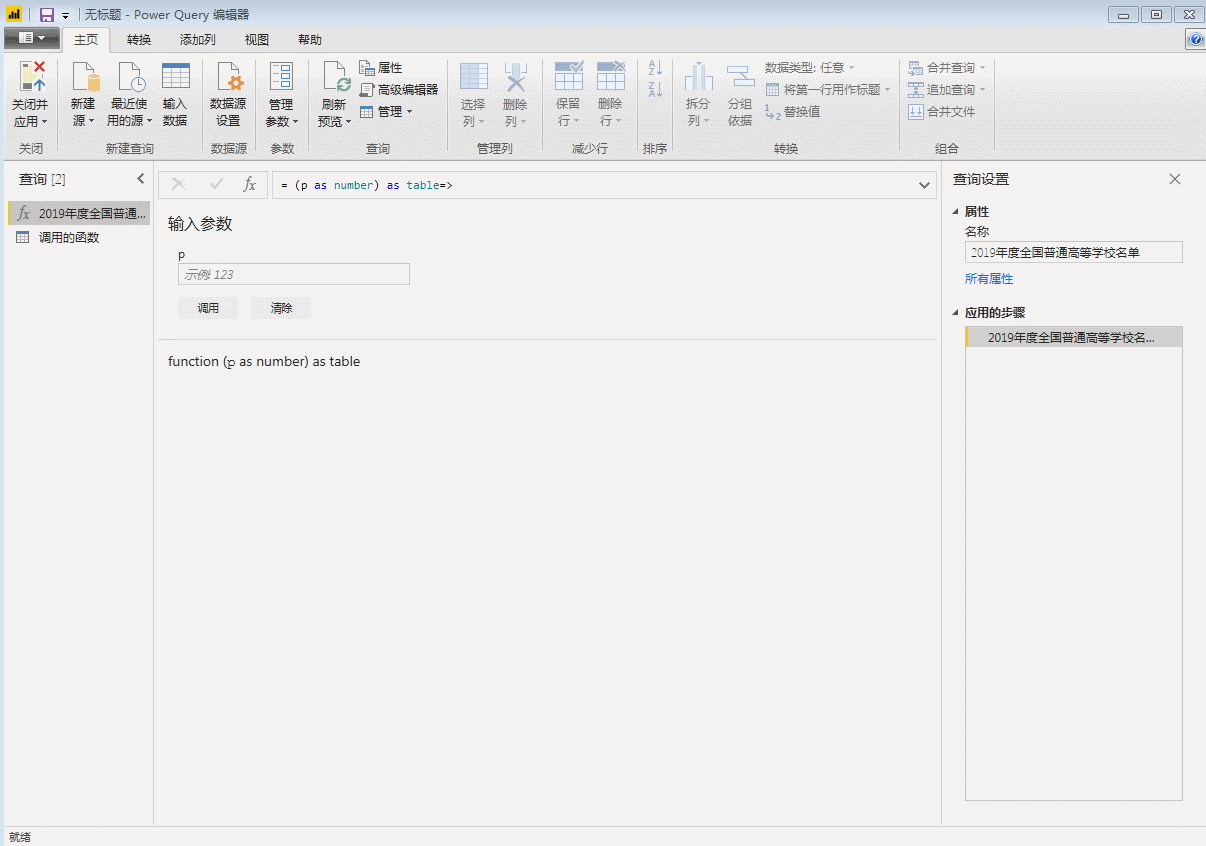
**输入参数只能一次抓取一个网页，要想批量抓取，还需要下面这一步。**

**（五）批量调用自定义函数**

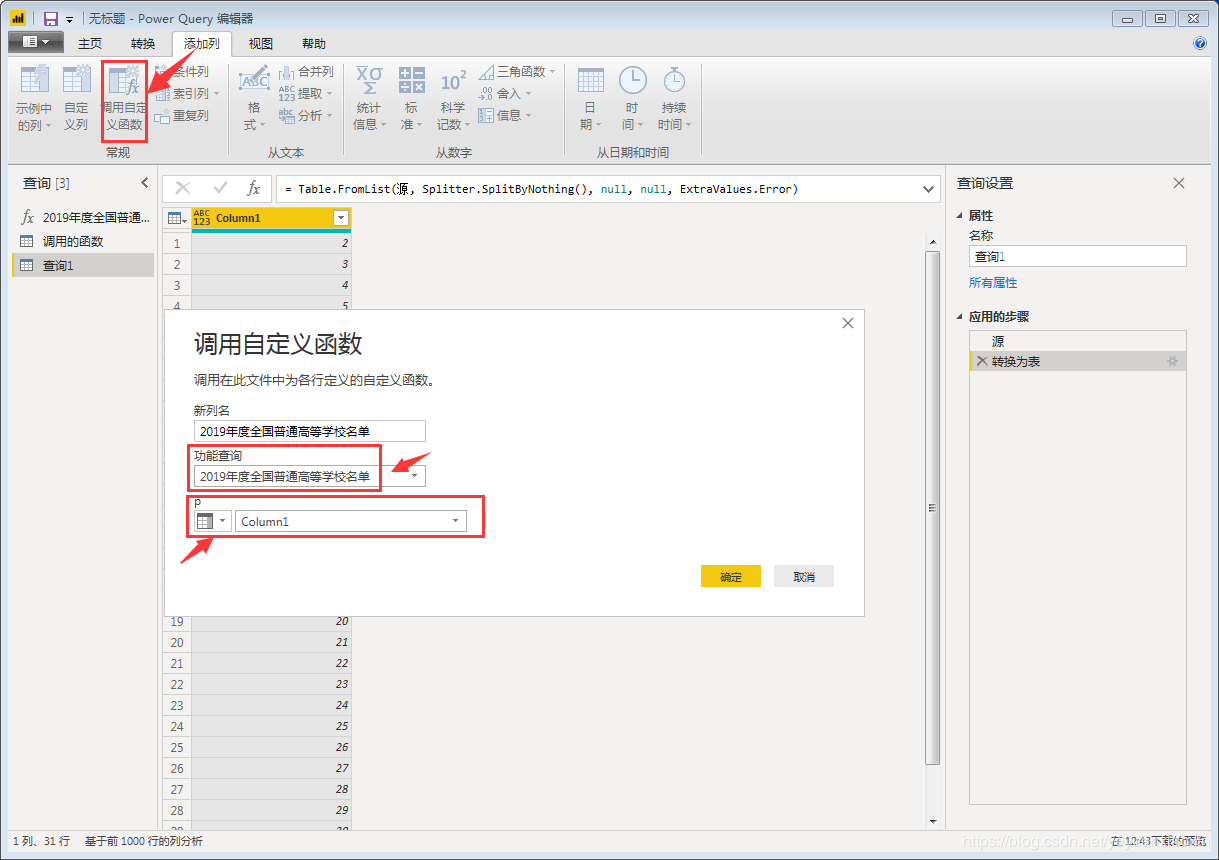
首先使用空查询建立一个数字序列，这里因为要获取从2页到32页的数据，所以我们建立从2到32的序列，在空查询中输入：

**={2..32}**

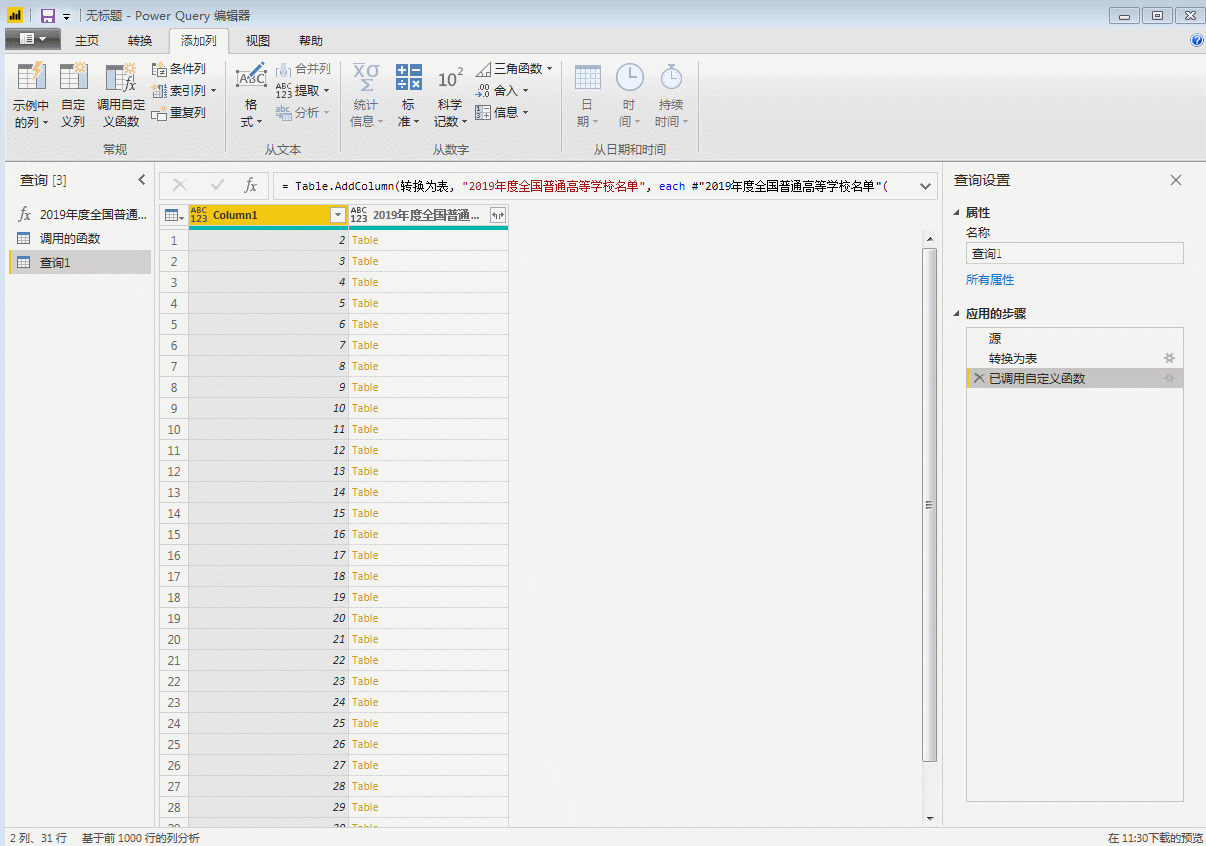
回车就生成了从2到32的序列，然后转为表格。



然后调用自定义函数



点击确定，就开始批量抓取网页了，这里如果采集页数量较大，不推荐在前面获取第一个网页后马上进行数据整理，否则可能会造成采集时间过长。

这里我们展开表格，就是全部31页的数据。

然后这里我们就就看进行后续的数据整理和可视化了。

备注1：如果采集的网页数据是不断更新的，在操作完以上的步骤之后，在PQ中点击刷新，可以随时一键提取网站实时的数据，可以说非常方便了。

备注2：以上主要使用的是PowerBI中的Power Query功能，在可以使用PQ功能的Excel中也是可以同样操作的。