# 第一阶段 业务分析师

# 建模数据分析师

## 网校相关视频

SPSS MODELER数据挖掘

## 明确本次课知识点，明确重点难点

### 【知识点目标】

* 数据挖掘简介
* 什么是数据挖掘
* 数据挖掘的一般步骤
* 数据挖掘与传统数据分析的比较
* 数据挖掘的特点
* 数据挖掘技术简介
* SPSS MODELER软件界面介绍
* 数据流设计区
* 管理区
* 项目区
* 工具面板区
* 选项卡
* SPSS MODELER连接数据源
* 导入excel文件
* 导入变量文件
* 导入statistics文件
* 载入数据源之后的设置
* SPSS MODELER基础操作介绍
* 数据流操作——生成数据流
* 数据流操作——添加和删除节点
* 数据流操作——连接数据流
* 数据流操作——修改连接节点
* 数据流操作——运行数据流
* 图形的制作——散点图
* 图形的制作——直方图
* 图形的制作——分布图
* 图形的制作——网络图
* 图形的制作——图形版

### 【重点】

* 数据流设计区
* 管理区
* 项目区
* 工具面板区
* 选项卡
* SPSS MODELER连接数据源
* 导入excel文件
* 导入变量文件
* 导入statistics文件
* 连接到数据库
* 载入数据源之后的设置
* SPSS MODELER基础操作介绍
* 数据流操作——生成数据流
* 数据流操作——添加和删除节点
* 数据流操作——连接数据流
* 数据流操作——修改连接节点
* 数据流操作——运行数据流
* 图形的制作——散点图
* 图形的制作——直方图
* 图形的制作——分布图
* 图形的制作——网络图
* 图形的制作——图形版

### 【难点】

* SPSS MODELER连接数据源
* 导入excel文件
* 导入变量文件
* 导入statistics文件
* 连接到数据库
* 载入数据源之后的设置

## 复习巩固作业讲解

无

## 本次课程任务讲解

### 【知识点1】数据挖掘简介

近年来，数据挖掘技术在各领域都得到了非常广泛的应用，只要该产业拥有极具分析价值，都可以利用数据挖掘工具进行有目的的分析。一般较常见的应用案例大多发生在零售业、制造业、物流业、金融业、保险业以及医疗服务业等。

案例：银行电话推销信用卡

**国外成功案例**

国外有关数据挖掘的成功应用实例较多，如英国电信选用了IBM公司的数据挖掘产品IBM SPSS Modeler，为其“商业高速公路”活动分析数据和建立探索模型。应用这个系统的结果是：英国电信更好地了解客户和他们在电信市场上的行为特征；向销售人员和营销活动提供了“最佳客户”清单。

### 【知识点2】什么是数据挖掘

数据挖掘，一般是指从大量的数据中自动搜索（算法建立的模型）隐藏于其中的特殊关系性信息的过程。

数据挖掘是目前人工智能和数据库领域研究的热点问题，主要基于人工智能、机器学习、模式识别、统计学、数据库、可视化技术等，高度自动化地分析企业的数据，做出归纳性的推理，从中挖掘出潜在的模式，从而帮助决策者调整市场策略，减少风险。

数据挖掘在技术上的定义是指从大量的、不完全的、有噪声的、模糊的和随机的数据中，提取隐含在其中的、事先不知道的，但又是潜在有用信息和知识的过程。与数据挖掘相近的同义词有数据融合、数据分析和决策支持等。

### 【知识点3】数据挖掘的一般步骤

1. 数据准备：从相关的数据源中选取所需的数据并整合成用于数据挖掘的数据集
2. 规律寻找：用某种方法从大量数据中将数据集所隐含的规律找出来
3. 规律表示：以用户可理解的方式（如可视化）将找出的规律表示出来

### 【知识点4】数据挖掘与传统数据分析的比较

数据挖掘与传统数据分析的本质区别：

数据挖掘是在没有明确假设的前提下进行数据的挖掘，通过挖掘而发现知识，挖掘所得到的信息应具有先前未知、有效和实用三个特征。

先前未知的信息是指该信息是预先未曾预料到的，即数据挖掘是要发现那些不能靠直觉发现的信息，甚至是违背直觉的信息，挖掘出的信息越是出乎意料，就可能越有价值。

在商业应用中最典型的例子是“啤酒尿布”的案例：沃尔玛通过数据挖掘发现，下班后前来购买婴儿尿布的男顾客大都购买啤酒。于是经理重新布置货架，把啤酒类商品布置在婴儿尿布货架附近，并在二者之间放置男士日常用品。这样，上述几种商品的销量大增。

### 【知识点5】数据挖掘的特点

1. 数据处理更强：处理大量实际数据更强势，且无需太专业的统计学背景去使用数据挖掘的工具。
2. 符合企业需求：从大型数据库抓取所需数据并使用专业分析软件、数据挖掘的工具更符合企业需求。
3. 方便用户使用：数据挖掘的目的是方便企业终端用户使用，而并非是给统计学家检测使用。

### 【知识点6】数据挖掘技术简介

1、分类预测（回归、决策树、神经网络、SVM等）

对将来发生的事情进行预测

2、聚类/细分（无导师/无目标）

基于事物的属性进行自动归类

3、关联（寻找多变量间的关联规则）

比如：啤酒尿布

1. 序列相关分析：确定事件发生的序列

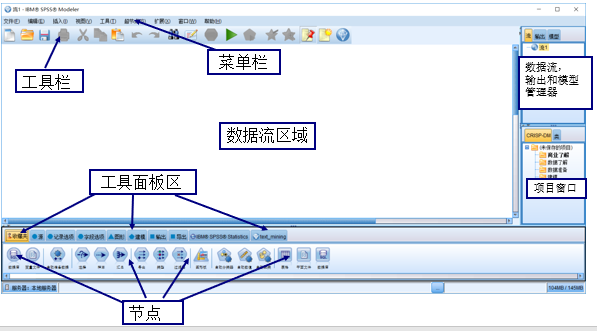
比如：先买手机、再买手机壳

5、异常分析：从正常群体中筛选出行为异常的个体

6、时间序列分析

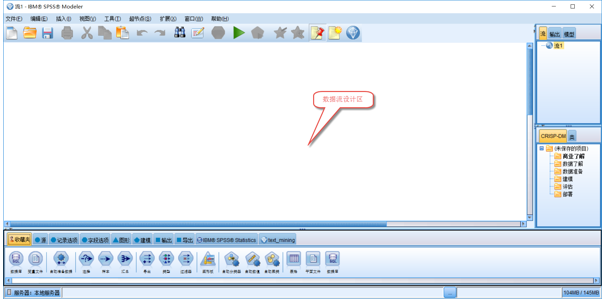
基于事物发展的延续性和随机性预测事务未来的发展

### 【知识点7】SPSS modeler软件界面介绍



### 【知识点8】数据流设计区

IBM SPSS Modeler项目中可以同时开启多个数据流设计区，显示建模过程中的数据流。



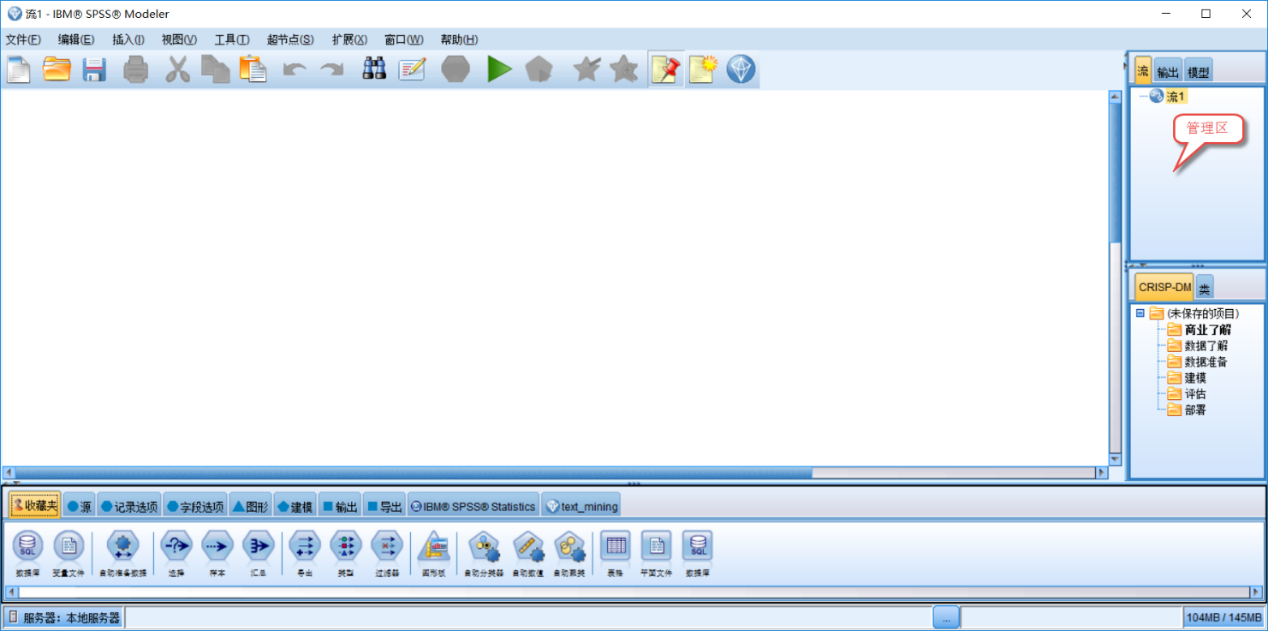
### 【知识点9】管理区

管理区包括“流”、“输出”、“模型”三个选项，主要用于在建模过程中对数据和结果进行有效管理。

**流：**SPSS Modeler界面上可以同时存在多个数据流，通过在管理区的“流”选项中单击可切换不同的数据流。

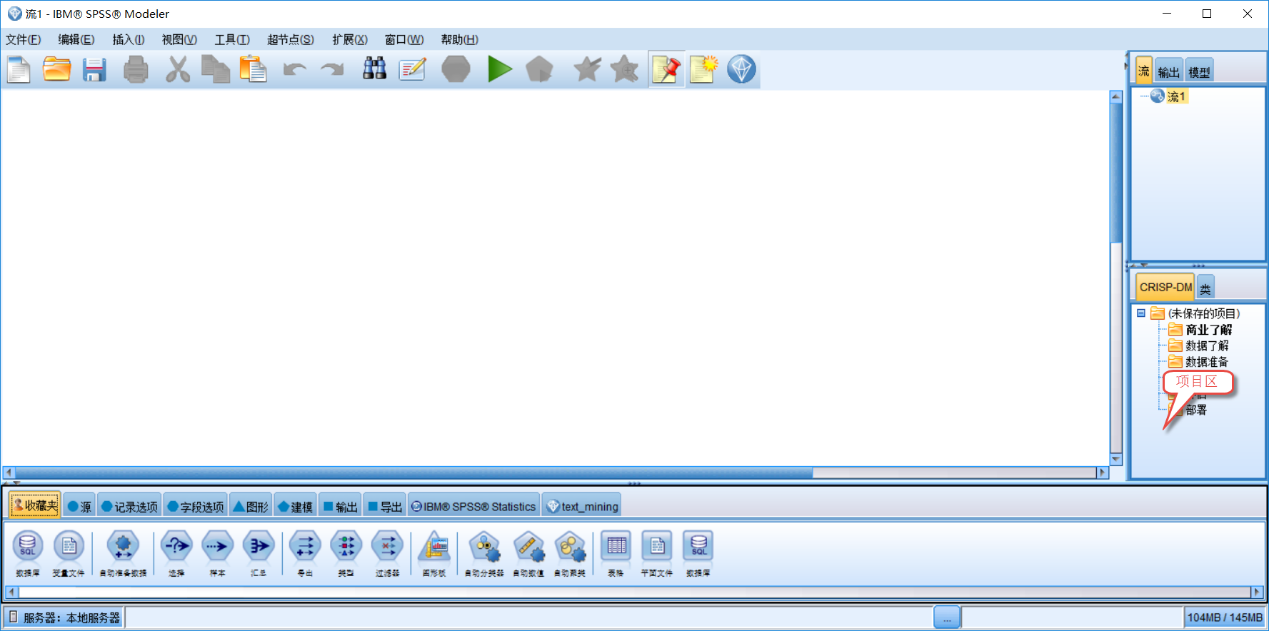
**输出：**不要跟工具栏中的输出相混淆，这里是模型产生的分析结果，例如数据源连接到数据审查、直方图，在执行数据流后，这个工具产生了三个输出，在管理区的“输出”选项卡中双击这些输出，可查看图形或报表。

**模型：**不要跟工具栏中的建模相混淆，这里是模型工具产生的分析结果，例如数据源连接到时间序列、回归等建模节点，在执行数据流后，产生模型输出。在管理区的模型栏中右击这些输出，在弹出的快捷菜单中选择“浏览”就可以看到输出的建模结果。



### 【知识点10】项目区

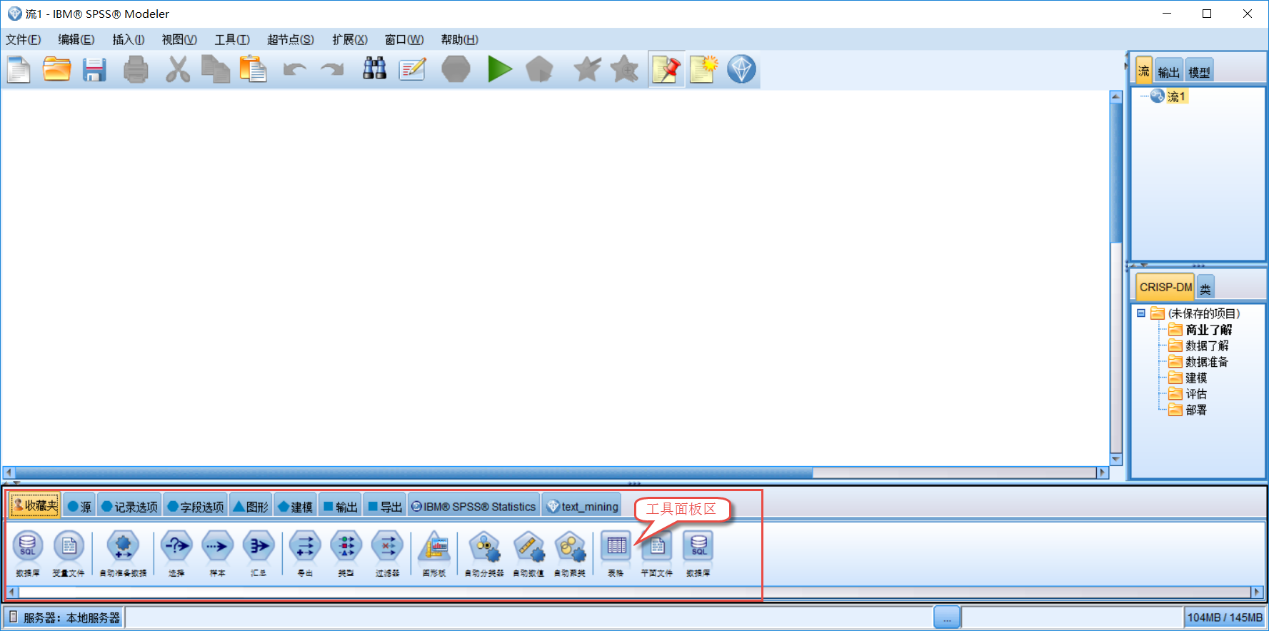
**它是对数据挖掘项目的管理，提供了两种视图：CRISP-DM和类。在SPSS Modeler里通过组织CRISP-DM的6个步骤完成项目，在项目中可以加入流、节点、输出、模型等。**

****

### 【知识点11】工具面板区

工具面板区是IBM SPSS Modeler在建模等过程中可以使用的工具，是我们学习的重点。

工具面板区主要包括“源”选项卡、“记录选项”选项卡、“字段选项”选项卡、“图形”选项卡、“建模”选项卡等9大类工具，后续的数据挖掘都是基于这些工具进行的，因此它是我们学习SPSS Modeler的基础和重点，下面将逐一进行介绍。

****

### 【知识点12】选项卡

1、源选项卡

“源”选项卡包含SPSS Modeler可以直接读取的所有数据源格式，主要有数据库、变量文件、固定文件、Excel文件、SAS文件、Statistics文件等。



2、记录选项

“记录选项”选项卡可用于对数据**行**进行转换，包含选择、汇总、排序、合并、追加、区分等。其中“选择”节点是选出符合我们条件的数据；“汇总”节点是将数据按照特定条件进行汇总统计；“排序”节点是将数据按照一定的规则进行排序；“合并”节点是将两个及以上的文件按照关键字等进行整合；“追加”节点是将两个及以上的文件进行数据的累加；“区分”节点是按照条件将重复数据删除。在记录上进行操作，一条记录是一种“情形”或一“行”数据



3、字段选项

“字段选项”选项卡可用于对**列**进行转换，包含类型、过滤、导出、填充、转置、字段重排等。在字段上进行操作，一个字段是一个变量



4、图形

“图形”选项卡可用于数据的可视化分析，包括SPSS Modeler可以生成的主要图形，如分布图、直方图、多重散点图、网络图、时间散点图、评估图等，在建模之前和之后用来可视化数据。



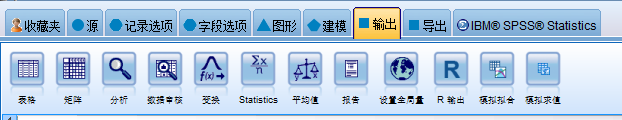
5、建模

SPSS Modeler中包括了丰富的数据挖掘模型，提供了一系列的数据挖掘技术，用来进行预测、聚类、关联、分类等，可满足数据挖掘的应用需求。



6、输出

SPSS Modeler的输出不仅仅是ETL过程（数据仓库技术：用来描述将数据从来源端经过抽取（extract）、交互转换（transform）、加载（load）至目的端的过程），还包括了对数据的统计分析报告输出，如表、矩阵、分析、数据审核、变换、统计量等。



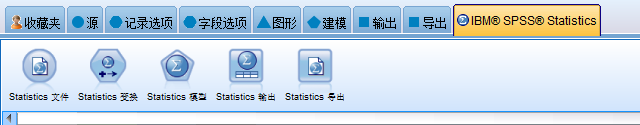
7、导出

SPSS Modeler导出的格式与“源”选项卡类似，包含数据库、Excel、SAS导出、Statistics导出等，用来对处理后的结果输出成相应格式。



8、**Statistics**

SPSS Statistics在数据分析中经常被使用，SPSS Modeler为了提高客户日常工作的效率，设置该节点便于模型结果的再利用，从而实现与SPSS Statistics的兼容。



### 【知识点13】SPSS MODELER连接数据源

每一个数据挖掘工具，都需要支持很多种数据格式，由于数据存储形式，或者存储软件的因素造成了有时候数据的格式会有很多种，为此作为一个好的数据挖掘软件，最重要的就是要支持很多种的数据格式，这样可以避免数据转换格式时耗费大量的时间和成本。

源节点顾名思义，就是数据来源的节点，对应在SPSS Modeler中，由“源”选项卡来完成这个工作。

### 【知识点14】导入excel文件

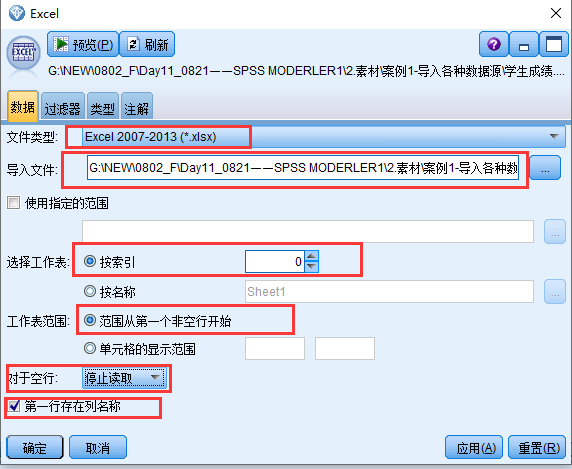
**文件类型**：分为Excel 97-2003（.xls）和Excel 2007-2013（.xlsx）两种类型，需要根据实际情况进行选择，否则数据将无法正常读取。

**选择工作表**：按索引或者名称指定要导入的工作表。

**工作表范围**：可以以第1个非空行作为开始导入的数据，也可通过指定单元格的显示范围导入数据。

**注意：导入的文件不能通过其他软件打开占用**

数据：学生成绩.xlsx



### 【知识点15】导入变量文件

**从文件中读取字段名**：默认是选中的，此时把文件的第1行看作是变量名。每个字段也能被分配一个名字，如果我们的第1行没有列标签，就可以取消选中。

**指定字段数**：指定每个记录中的字段数。只要记录以新行结束，就可以自动检测字段数，也可以手动设置字段数。

**跳过标题字符**：指定要忽略第1个记录开头处的多少个字符，默认为0。

**EOL注解字符**：指定字符以表示数据中的注解。无论这些字符之一出现在数据文件的何处，从该字符起直到下一个新行字符（不包括）之前的所有字符都将被忽略。

**去除开头和末尾的空格**：丢弃导入字符串中开头和结尾的空格。

**无效字符**：空字符或者指定编码中不存在的字符，通过丢弃进而删除，或替换成其他字符。

**小数符号**：选择在数据源中使用的小数分隔符类型。

**字段定界符**：选择将数据进行分列的符号，也可以自己指定符号，此处是复选，也就是说可以使用一个以上的分列手段，这点和Excel的分列大同小异。

**列和类型的扫描行数**：指定数据类型要扫描的行数。

**引号**：如何处理导入数据的单引号和双引号问题。

在此对话框中操作的任何时刻，都可单击“刷新”按钮以从数据源重新载入字段。在更改源节点的数据连接时，或者在对话框的选项之间进行操作时。

**变量文件一般用来导入含有分隔符的数据文件，如CSV,TXT等**

**数据：bank.csv**



### 【知识点16】导入statistics文件

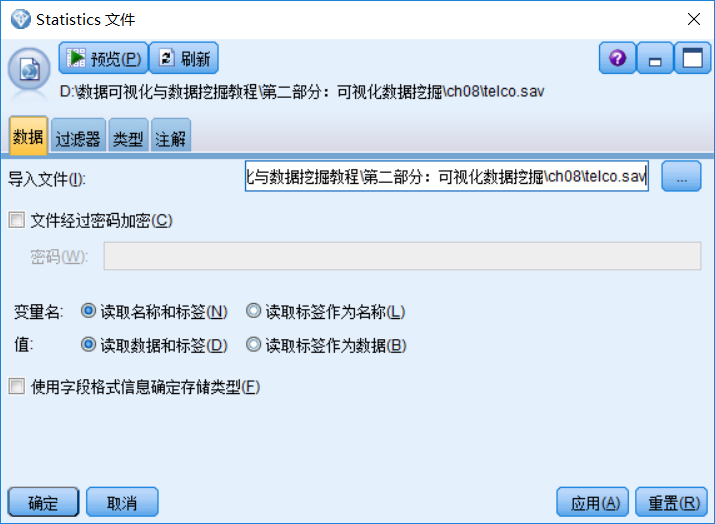
Statistics文件（IBM SPSS Statistics软件保存的文件）是我们进行数据存储的常见数据格式，快速正确的导入SPSS数据对于后期分析具有积极作用，“Statistics文件”对话框如图所示。

对该对话框中主要选项的说明如下。

**读取名称和标签：**选中此选项将变量名称和标签同时读入。默认情况下将选中此选项，并且变量名称将显示在“类型”节点中。根据“流属性”对话框中指定的选项，标签将显示在表达式构建器、图表、模型浏览器和其他类型的输出中。

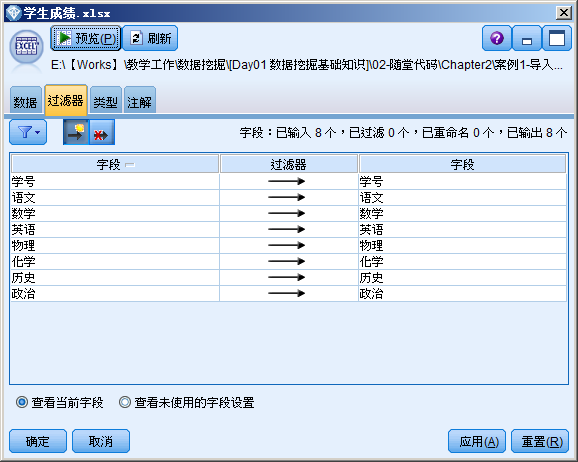
**读取标签作为名称：**选择从文件中读取变量标签而不是字段名，并将这些标签作为变量名称在SPSS Modeler中使用。

注意：值的读取也分为读取数据和标签、读取标签作为数据。



### 【知识点17】载入数据源之后的设置

IBM SPSS Modeler加载数据源之后，可以在数据源节点的“过滤器”、“类型”及“注解”选项卡对该节点的一些属性进行设置，包括字段的控制、字段的类型与角色设置等。



字段的测量级别：

**连续型** – 用于描述数值，如0-100 或者0.75-1.25 内的连续值一个连续值可以是整数、实数或日期/时间

**分类（离散型）**– 用于当一个具体值的精确数量未知时描述字符串，一旦数据被读取，其类型就会转换成标记、名义或者无类型

**名义** – 用于描述带有多个具体值的数据（黄、绿、蓝）

**标记** – 用于只取两个具体值的数据（真、假，1、0）

**无类型** – 用于不符合上述任一种类型的数据或者含有太多元素的集合类型数据

### 【知识点18】SPSS MODELER基础操作介绍

**1、数据流操作：数据流操作是我们进行软件操作和数据挖掘的基础。数据流的操作主要包括生成数据流、添加和删除节点、连接数据流、修改连接节点和运行数据流等5个方面。**

**2、图形的制作：SPSS Modeler可以生成分布图、直方图、集合图、多重散点图、网络图、时间散点图和评估图等，它们在数据挖掘过程中有着广泛的应用。由“图形”选项卡来实现。**

### 【知识点19】数据流操作——生成数据流

在数据挖掘过程中，首先需要生成数据流，数据流是由一系列的节点组成，当数据通过每个节点时，将会对它进行事先设置好的操作。

建立数据流通常遵循以下4步：

1. 向数据流程区添加新的节点；
2. 将这些节点连接到数据流中；
3. 设定数据节点或数据流的功能；
4. 如何运行数据流。

### 【知识点20】数据流操作——添加和删除节点

当向数据流程区添加新的节点时，有下面4种方法。

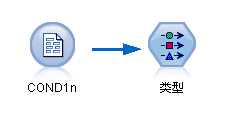
1. 双击面板中待添加的节点。
2. 左键按住待添加节点，将其拖到数据流程区内。
3. 选中面板中待添加的节点，将鼠标放入数据流程区，当鼠标变为十字形时单击流程区。
4. 右击面板中待添加的节点，从弹出的快捷菜单中选择“添加到流”命令。

通过上面4种方法可将选中的节点添加到数据流程区内。当我们不再需要数据流程区内的某个节点时，可以通过以下两种方法来删除：

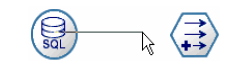
1. 左键单击待删除的节点，按Del键。
2. 右键单击待删除的节点，在弹出的快捷菜单中选择“删除”命令。

### 【知识点21】数据流操作——连接数据流

**方法一**：选中数据流中要连接新节点的节点（起始节点），在工具面板区双击要连接到数据流的节点（目标节点），这样便将数据流中的节点与新节点相连接了，如图所示。

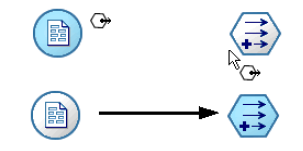


**方法二**：在工作区内选择两个待连接的节点，选中连接的起始节点，按住鼠标滑轮将其拖曳到目标节点放开，连接便自动生成（如果鼠标没有滑轮可选用Alt键和左键代替），如图所示。



**方法三**：右键单击待连接的起始节点，从弹出的快捷菜单中选择“连接”（Connect），选中“连接”（Connect）后鼠标和起始节点都出现了连接标记，用鼠标单击数据流程区内要连接的目标节点，连接便生成，如图所示。

第1种连接方法是将软件界面下方选项卡中的节点与数据流相连接，后两种方法是将已在数据流程区中的节点加入到数据流中。此外，数据源节点不能有前驱节点，即在连接时它只能作为起始节点而不能作为目标节点。

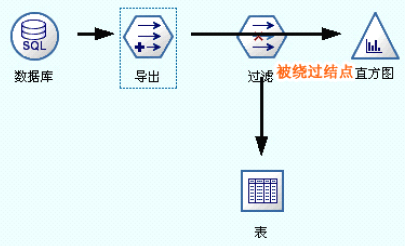


### 【知识点22】数据流操作——修改连接节点

修改连接节点主要包括：绕过数据流中的节点、将节点加入已存在的连接、删除已经存在的连接等3种基本形式的修改。

**1. 绕过数据流中的节点**

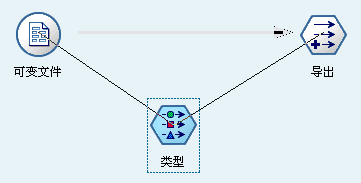
当我们暂时不需要数据流中的某个节点时我们可以绕过该节点，在绕过它时，如果该节点既有输入节点又有输出节点，那么它的输入节点和输出节点便直接相连；如果该节点没有输出节点，那么绕过该节点时与这个节点相连的所有连接便被取消，如图所示。



方法：按住Alt键，通过用鼠标左键单击该节点来完成。

**2. 将节点加入已存在的连接**

当我们需要在两个已连接的节点中再加入一个节点时，可以采用这种方法将原来的连接变成两个新的连接，如图所示。



方法：用鼠标滑轮单击欲插入新节点的两个节点间的连线，按住它并把它拖到新节点时放手，新的连接便可生成（在鼠标没有滑轮时亦可用Alt键代替）。

**3. 删除已经存在的连接**

当某个连接不再需要时，可以通过以下两种方法将它删除：

选择待删除的连接，单击右键，从弹出的快捷菜单中选择删除连接。

选择待删除连接的节点，按F3键，删除所有连接到该节点上的连接。

### 【知识点23】数据流操作——运行数据流

数据挖掘。构建好数据流结构后，只有通过运行数据流，数据才能从读入开始流向各个数据节点。

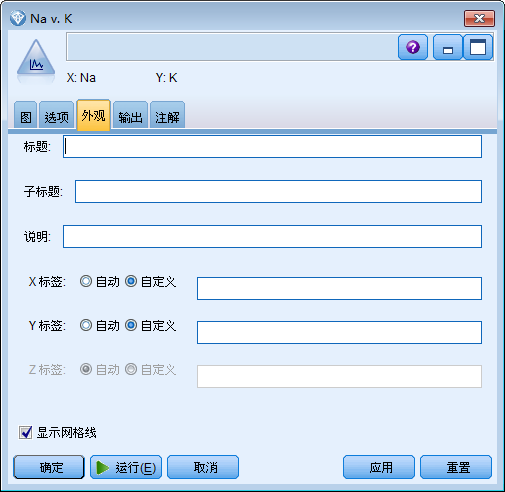
运行数据流的方法有以下3种：

1. 选择菜单栏中的按钮 ，数据流区域内的所有数据流将被执行，具有多个输出时，由于这种方法不便于使用者观察挖掘结果，因此使用不是很多。
2. 先选择要输出的数据流，再单击菜单栏中的按钮，被选择的数据流将被执行，由于是一对一的展示模式，便于查看结果，因此该方法在日常数据挖掘中较常用。
3. 选择要执行的数据流中的输出节点，单击鼠标右键，在弹出的快捷菜单中选择“运行”，这样就可执行被选中的数据流，在日常的数据挖掘工作中较少使用。

### 【知识点24】图形的制作——散点图

在日常数据分析中散点图是经常遇到的图形，尤其是在判断变量的分布时，从散点图发现规律特征，这将有助于后期分析的顺利进行，如图所示。

对于除了条形图以外的所有图来说，用户可以在图创建之前或者探索已生成图时指定外观选项，如图所示为散点图的“外观”选项卡。数据源：历史数据.xlsx

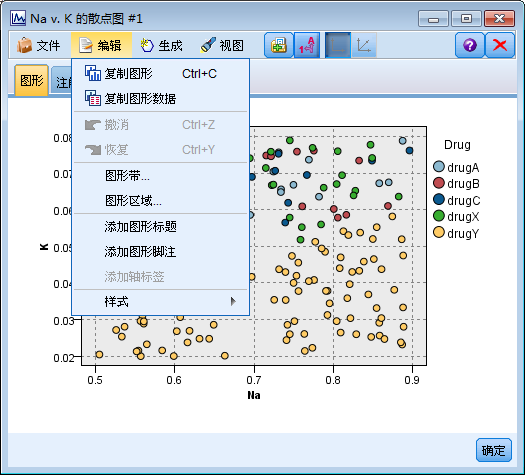
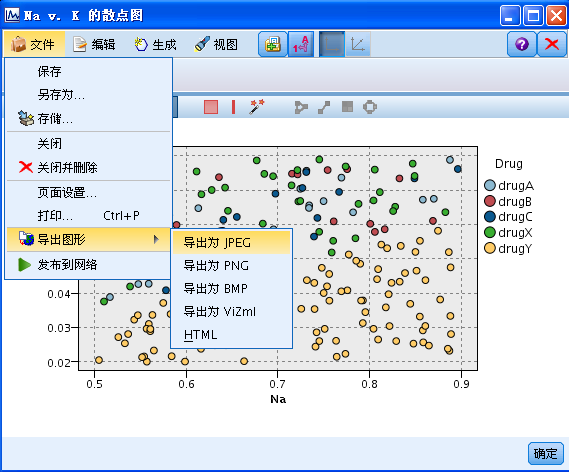


用户创建图形后，可利用一些方法制定和操作我们创建的图形，允许的操作方式包括：

使用鼠标来选择一个图表区域以便进行更深入的操作。

不同的图表可能具有不同类型的可用菜单和选项，在选定区域上右击鼠标，在弹出的快捷菜单中给这个区域提供可用的内容菜单，如图所示。

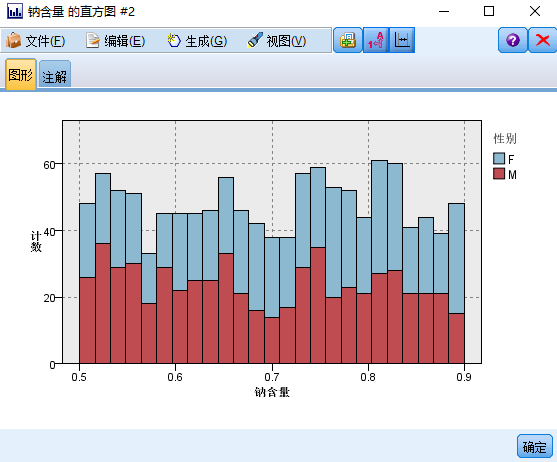
页面设置和导出功能强大。可以对图形的显示进行灵活设置，同时可以导出成多种格式的图形，如JPEG、PNG和BMP等，此外还可以将图形发布到网络，如图所示。

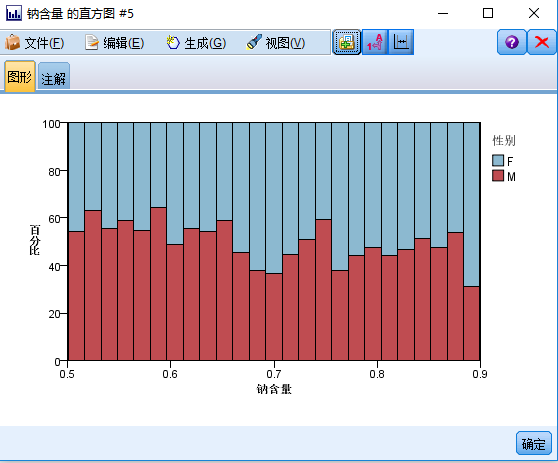
 

### 【知识点25】图形的制作——直方图

直方图又称柱状图、质量分布图，是一种统计报告图，由一系列高度不等的纵向条纹或线段表示数据分布的情况，一般用横轴表示数据类型，纵轴表示分布情况，如图所示。

直方图的优点是：可以解析出数据的规则性，比较直观地看出数据的分布状态，对于资料分布状况一目了然，便于判断其总体分布情况。数据源：历史数据.xlsx

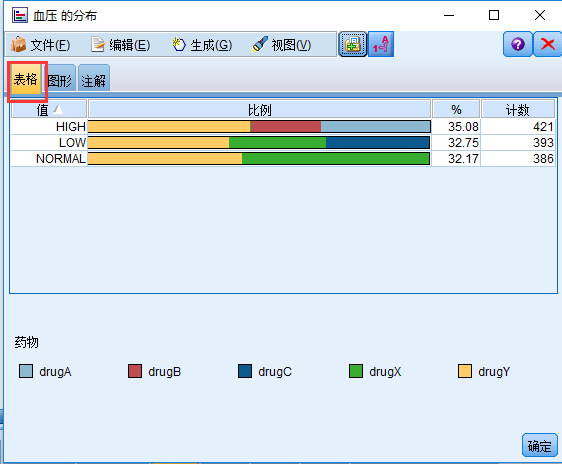
 



### 【知识点26】图形的制作——分布图

分布图可以查看分类的变量对其他变量的影响情况

数据源：历史数据.xlsx



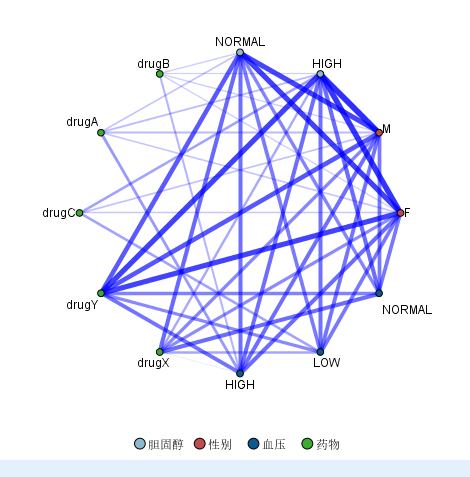
### 【知识点27】图形的制作——网络图

建立网络图的目的是为了让用户可以直观地看到商品之间的关联程度，从而有一个感性认识。双击“网络”节点，打开如图所示的对话框，在“字段”列表中选择需要添加的字段，可以将所有的商品字段添加进来，也可以选中“仅显示真值标志”复选框。

打开“选项”选项卡，用户可以在此设置链接数量的显示范围，不显示一些链接数量低的链接。

单击“运行”按钮，这时会生成一个商品之间关联程度（链接数量）的网络图，用户可以在下方的调节杆上调节链接数量的显示范围，如图所示。在图中，线的粗细和深浅代表着联系的强弱，可以直观地看到beer和frozenmeal、cannedveg的联系程度比较强。

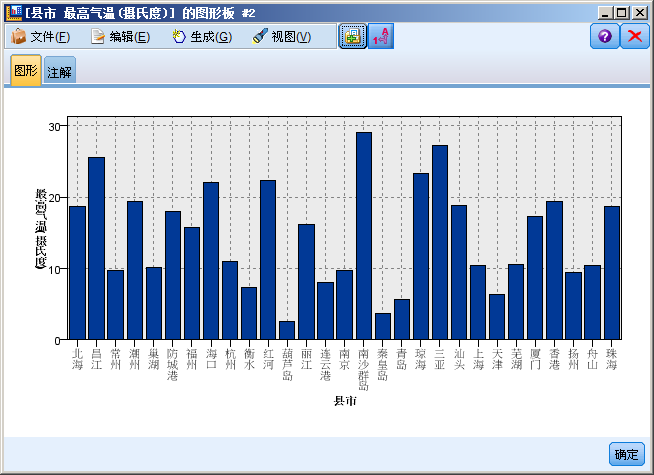
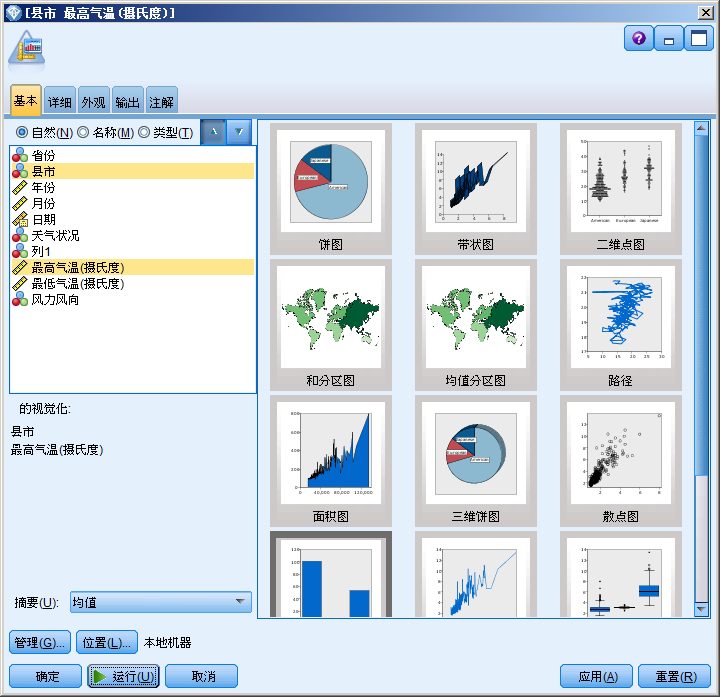
数据源：历史数据.xlsx

### 【知识点28】图形的制作——图形板

通过“图形板”节点，可以从单个节点上的许多不同图形输出（条形图、饼图、直方图、散点图和热图等）中进行选择。从第一个选项卡开始，选择需要探索的数据字段，然后节点将提供一个适用于数据的图形类型的选项。节点将自动过滤出适用于字段选项的所有图形类型。在“详细”选项卡上，可以定义详细的选项或较高级的图形选项。

打开“基本”选项卡，用户可以在此设置需要展示的字段及所需要的图形效果，还可以通过“详细”、“外观”、“输出”等选项卡设置更丰富的效果。



## 自主学习作业讲解

无

## 课程总结

* 数据挖掘简介
* 什么是数据挖掘
* 数据挖掘的一般步骤
* 数据挖掘与传统数据分析的比较
* 数据挖掘的特点
* 数据挖掘技术简介
* SPSS MODELER软件界面介绍
* 数据流设计区
* 管理区
* 项目区
* 工具面板区
* 选项卡
* SPSS MODELER连接数据源
* 导入excel文件
* 导入变量文件
* 导入statistics文件
* 载入数据源之后的设置
* SPSS MODELER基础操作介绍
* 数据流操作——生成数据流
* 数据流操作——添加和删除节点
* 数据流操作——连接数据流
* 数据流操作——修改连接节点
* 数据流操作——运行数据流
* 图形的制作——散点图
* 图形的制作——直方图
* 图形的制作——分布图
* 图形的制作——网络图
* 图形的制作——图形版

## 下次自主学习任务布置

#### 观看预习视频

数据建模分析师——Python语音基础

#### 课后作业

1. 参见直播平台
2. Qq群课后习题