1 数据库基本原理

1.1 创建数据库并读入

在 excel 中打开数据，从其他源里面选取 Microsoft query。在 Query 里面选择新的数据源可以从创建新的数据源，并在创建的过程中注意驱动程序的选择和数据库文件。这样就可以创建一个新的数据库。同时也可以从 query 里面选择需要读入的数据。在具体读入的时候，可以选择具体的要读入的表。

1.2 条件查询

不同列之间的条件做并，不同行之间的条件作或。针对日期字段的查询要写为 #1996/7/1#。其他的数字值可以直接用 >,< 之类的符号进行筛选。

1.3 多表查询

1.3.1 内连接

内连接：是将多个表中符合条件的记录挑出来组成一个结果。在内连接查询中，几个表之间一定要有关系。如果没有表表和之间没有关系，需要使用中间过渡的表来使得我们需要读取数据的表相互连接。也可以手动将两个字段关联起来，但是要保证字段中的类型和长度要一致，字段名字可以不同，相当于同种数据直接根据自身的值的到其他表中进行对应以查到进一步的信息。

连接的条件和查询之间的关系：每一个连接会有一个等式表示要对应的数据关系，这个等式的意思是：其连接字段值同时出现在两个表中的（比如「2」这个数字都在两个表中的相连接字段中出现了，那么在查询 1 表的某些值时会出现连接字段值为 2 的数据）每一个对象会作为查询的总范围。

书上的两个例子：不同表中表示同一个东西的不同名字的「运货商」,「运货商 id」；也可以时表示层级关系的上下级关系对应（具体是其上级栏位中对应了上级的雇员 id，这样就可以和另一个复制表中的雇员 id 建立关系）。但是如果没有寻找到对应关系的字段将不会显示出来。

1.3.2 外连接

外连接适用于将不符合连接条件的记录一并查询出来。在连接的选项中，可以选择将表中没有列入查询范围的对象可以在此表中查询到（所有的此表对象均显示）。

1.4 计算字段

Query 中也可以使用类似 excel 的方式进行简单的计算。在计算的时候在选择变量的时候需要具体指定在哪个表的字段。例如：数量 * 订单明细. 单价。在 excel 中计算的时候注意公式需要从字段中拖过来保证能够用 (0. 0)。

2 数据分类汇总

这个是在 excel 里面进行操作。我们只用 D 函数和透视表 (0. 0)

2.1 一些汇总操作之前的操作

存储在 excel 表中的数据也可能作为一种数据源的方式，然后里面里面的主表可能是有 ID 构成然后有字表来解释 ID 的意义。这时候我们也可以用 query 然后在新的 excel 文档中打开。注意选取正确的驱动程序来打开新的数据源。在 query 里面打开的 excel 作为的数据源时要添加表需要在「选项」里面打开「系统表」。excel 之间的表需要手动建立关系。

2.2 数据透视表

在「插入」选项中可以插入数据透视表（图）。数据区域不包含字段名，会自动忽略然后将他们加入到字段列表里面。由「报表筛选」「列标签」「行标签」「数值」来管理各种的标签和最终汇总表的样式。透视表完成之后在光标放在数据透视表中时可以在「数据透视工具」里面可以做存储数据透视图。也可以直接按 Alt + D + P 启动数据透视表创建向导然后直接从外部数据源导入数据建立数据透视图。

将字段放在「报表筛选」之中表示下面表格中出现的项都属于报表筛选中规定的属性。在「行标签」中，最直观显示出来的就是最上面的分类项，其下面的则是对于其父项的更细的分类。

2.2.1 按月汇总

随便点击一个在数据透视表当中的一个单元格，右键「创建组」就就可以对于不同的条件进行汇总。得到数据透视图的折线图时，点击图右键可以使用「趋势线」进行预测。在「预测线」界面可以选择显示方程和 R 值。R 值处于 [0, 1)，越靠近 1 越贴合。由于结束的月份只包含很少的天数，所以我们不包括最后结束的月份以保证预测的正确性。在标签里面可以选择不显示隐藏的月份。对于月份的缺失（当月数据为 0 会不出现在透视表当中），可能需要插入这些月份。在字段列表里面点击下面区域中的相应字段右键「字段设置」，在布局和打印中恢复无数据项。对于无数据项进行值修改，右键此字段点击「数据透视表选项. 布局」，对于空单元格显示 0，注意设置为数字 0 而非字符 0。在数据透视表透视里面可以点击设计可以更高改显示放法（比如表格形式）。在将需要的数据提取出来的时候，可以在第一个位置指定对应的位置，然后将框拖下去就可以得到下面的对应的数据的复制。复制的时候注意单元格的格式设置。

在 query 里面的合并：在 query 里面的字段名双击可以「求和」。有两种用法：通过一个没啥用的字段来合并一二关键字相同的项，用于计数。第二种是合并的按照一二关键字的某个有意义数值，比如销售额求和。

2.2.2 对于销售次数的频率分布的汇总

首先得到一个含有订单时间，销量和产品名称的汇总表。将数量分别拖到行标签和数量中表示「销售数量为 x 的销售有多少次」。注意到在最开始的默认值字段设置里面指汇总方式的求和表示的是「销售量的求和」，是对于字段的数值的求和。那么对于销售量为 2 的产品进行统计的时候就会得到要求答案的 2 倍。将值字段设置为「计数」就是数值出现的次数的求和，满足要求。要得到每一种销量占所有销量的比，在值字段中「值显示方式」修改为列汇总百分比。这里可以指导一个大包装里面装入的最优的数量。

2.2.3 计算百分比

目标；统计某一个范围内数据的数目占比。在 query 里面直接对于第一关键字（客户）和第二关键字（订单 id）相同的订单直接进行合并（对于合并字段右键求和）。excel 对于列进行排序（右键或者上面的「栏位」之中）。对于所选的一块创建组就是直接拉一个组（区别与点一个然后创建组）。

在得出了柱形图之后，在同时有两种不同单位的数据的时候，可以使用副轴来作为一种单位的轴。可以通过鼠标来电或者用键盘上下去选择比教不容易用鼠标选中的图。右键或者在上面数据透视工具「数据系列格式」「系列选项」中可以设置在副轴上。通过样在系列选项里面可以设计柱形图的间隔。对于坐标轴的显示范围和格式也可以右键「设置坐标轴格式」「坐标轴选项」里面进行设置

用散点图可以使选中的第一列为 x 坐标，其余的列都是 y 轴。

2.3 D 函数模拟运算表控件

2.3.1 模拟运算表

模拟运算表在数据-数据工具-模拟分析。拉出一块有一列为自变量，另一列为应变量（最上面额外以个值用于指定在旁边的函数定义），然后将应用列的单元格为函数定义中选择自变量对应的符号定义的例值（也是通过在 excel 表中的位置来关联的）。模拟运算表中的单格内容无法被编辑修改，只能被引用到其他的位置。如果自变量和应变量放在两行，则是应用列的单元格函数定义。如果是二元函数，行列分别设置为自变量和应变量，然后函数值放在行列相交的地方。

2.3.2 D 函数

D 函数用于数据库操作。D 函数的一般形式：

D 函数名称（数据列表，汇总字段，条件区域）

在罗列数据列表的时候，字段名要注意在里面。数据区域可以自定义命名：左上角的编辑栏。

D 函数结合模拟运算表：函数在 excel 里面被定义在每一个单元格上。应用的时候就要指定是引用带有函数表达式的单元格作为函数定义的参考。在 D 函数中，变量为条件区域的时候，那么行和列的标签就应该指向这些条件自变量。在构建模拟数据表的时候，可以使用数据高级筛选利用空的条件加上不显示相同的值就可以贴出来列自变量。行自变量可以通过提取到列后通过复制时开始复制选项中选择转置。

D 函数给出汇总的具体内容，然后通过模拟数据表罗列出所有的汇总的范围的变量然后将它们作为 D 函数条件区域的变量得到一个汇总。如果列出的变量没有出现在原表中的时候，会默认为是现在计算的销售额的那个变量。通过对于模拟数据表进行值复制就可以对它进行操作如排序。对于汇总值进行季度分类，要用到一下函数

left 函数：从左往右截取 x 个函数

left(文本，字符数目)

mid 函数：从左往右数到 x 位置截取 y 个字符

mid(文本，开始位置，截取字符数)

这两个函数可以从字符串中提取信息，如从「1996 年 3 季度」中提取出 1996 和 3。

date 函数：将字符转为日期格式

date(年，月，日) 对于每一个月的最后一天，由于最后一天会随着月份不同，可以用下个月的第一天-1 来得到。

通过以上函数可以设置函数条件：

```
"<=" & date(x, y, z)
">" & date(x, y, z)
```

x, y, z 就是通过给出的文字提取的年月来计算具体的季度上下限。

2.3.3 控件

控件就是下拉按钮！ 在文件-选项-自定义功能区 = 主选项卡-勾选开发工具。对于需要的单元格做控件，用表单控件中的「列表框控件」。在随便一个位置画好框框以后，右键可以设置控件格式设置数据源区域和单元格连接。点击控件里面的选项会在单元格链接中显示序号。

Index 函数：在矩形的数据区域中，根据行号和列号取出行列交叉点数据。

Index(数据区域，行号，列号)

通过将行号或者列号指向控件的单元格链接，就可以通过控件修改对应单元格链接的值来修改单元格的显示内容。对于修改条件区域的控件，可能会出现将「全部」转为条件区域中的空来代表全部的集合。这里要用到 IF 函数

if(条件, 条件 T 的表达式, 条件 F 的表达式)

对应例 11，12；自己回家做.cpp

2.3.4 3-例 13 的一些简单说明

目标：根据在控件之中所选择的总销售额前 10 的公司来展示一张该公司不同季度的图

第一步：调入数据

第二步：构造 d 函数，通过模拟运算表做出所有公司的销售额，并进行排序，选出前 10。

第三步：在某个公司为和某个季度间隔中为条件下的销售额汇总。通过 d 函数的条件区域来设定以上条件，最终通过汇总的模拟运算表得到折线图。注意，季度的条件是由有一个由函数得出的单元格得出的，而这个单元格的函数又上溯于一个原始条件。在使用模拟运算表汇总的时候注意要追溯到原始条件来设定变量。

第四步：通过控件来控制能够得出季度汇总 d 函数的公司条件。如上面所讲通过链接单元格能够加上 index 函数来满足以上要求。

一些扯白：更加复杂的汇总都可以像第三步一样明确条件之后再来考虑具体的 d 函数设计，特别是条件本身的设计方式决定了模拟运算表选择列/行的单元格的例子条件。控件可以用于条件的控制。

柱形图选定的客户有不同的颜色：通过正常的模拟运算表得出的数据并复制一份，其中复制的数据中选定客户的数据不变，其余都为 0；用这两列数据取做柱形图，然后可以用选中客户的柱形去覆盖所有客户的柱形。

2.3.5 作业总结

通过单元格的数字确定其他的单元格：首先先用一些列单元格计算出这个对应的所有的单元格的位置，这里可以使用 INDIRECT(”A1”) 函数来实现。然后再在目标位置通过 INDIRECT(A2) 来访问引用地址得到目标值。

3 预测模型

MSE: 均方误差，越小越好

3.1 移动平均

跨度为 n 的情况下，那么第 t_x 的预测结果就是 $t_{x-n}, t_{x-n+1}, \cdots t_{x-1}$ 的数据平均得到。在 excel 里面，就是使用 average 函数。微调控件可以用来调整数值，需要设置最小值，最大值，步长值，然后可以通过某一个单元格输出。数据分析可以在加载选项中设置的到。

在计算 a^b 的时候，最写成连乘的形式：由于指数运算先要转化为 $e^{bln(a)}$，所以 1 是对于指数有要求，而是由于使用泰勒级数计算所以误差更大。在引用单元格的值的时候，以 H1 为例，如果在下拖的时候想要始终引用这个值，那么表达式要写成 H1(绝对引用)。offset 函数:

```
OFFSET(Reference, Rows, Cols, Height, Width);
```

Reference: 选定单元格作为基点
Rows / Cols: 基点的行列偏移，可正（右下）可负（左上）
Height/Width: 偏移后的基点为左上角形成的矩阵的大小。

对于显示为空要在图表中显示为 0：通过选择数据强制将选择的数据范围扩大到值为空的部分。图表中原本没有显示的原因是自动忽略掉了显示为空的数据而没有加入表的范围之中。

3.2 平滑指数

我们认为离预测天数更近的数据会有更高的权值，而不像移动平均直接做平均数

$$\begin{aligned} Y_t &= \alpha F_{t-1} + \alpha(1-\alpha)F_{t-2} + \alpha(1-\alpha)^2 F_{t-3} + \cdots \\ &= \alpha F_{t-1} + (1-\alpha)Y_{t-1} \\ &= Y_{t-1} + \alpha(F_{t-1} - Y_{t-1}) \end{aligned}$$

在没有初始的预测值的情况下，设置 $Y_1 = F_1$

```
SUMXMY2(Reference1, Reference2)
```

这个函数用于计算对应两数组的数据值差的平方和，用于计算均方误差。

```
Count(Area);
```

这个函数用于计算某一个数据区域中的值的个数，用于求平均的分母。

```
Match(Content, Area, Flag)
```

Content: 要查找的对昂

Area: 要查找的数据范围

Flag: 1, 0, -1; 缺省为 1; 1 代表找到一个小于或者等于 content, 同时要求数据区域升序; -1 则是大于等于 content，同时要求数据区域降序；0 找到的第一个 =content 的值，不要求排序

这个函数可以返回查找到的位置序号，我怀疑他是二分和 for 循环。例子：找到数组中最小指的下标序号 (实际上可以用这个来查找最小 MSE 的平滑常数，通过缩小平滑常数的范围来得到更加精确的值)：

```
INDEX(SEQUENCE_AREA, MATCH(MIN(VALUE_AREA), VALUE_AREA, 0))
```

3.3 趋势预测

通过图表中的趋势性预测来获得数学公式表达。通过 R 平方值来计算拟合度，R 越大越接近。对于线性预测，也可以手动来求线性方程。

```
LINST(x, y)
```

可以得到线性预测的斜率。如果是选中两个水平的单元格而且在确认的时候用 ctrl + shift + enter 来使用数组公式来计算同时得到斜率和截距。也可以用 INTERSECT 和 SLOPE 的组合来分别算截距和斜率。

练习 2: 2009-2011 年

4 回归分析

4.1 一元线性回归分析

可以直接在图上用趋势分析获得一元拟合函数。

回归分析需要在选项，加载项中加载数据分析器，可以直接用数据分析来做回归预测，其中的「标志」勾选表示选择的区域的第一行是字段名。「输出区域」是输出相应的分析报告。其中的调整 r 平方适用于多变量函数。Intercept 和下面的分别就是截距和斜率

也可以先用一个初始的 k=1, b=1 来先做出预测值和 MSE; 然后使用规划求解，设置目标必须制定一个包含了一个函数的单元格。约束条件可以是目标函数的约束和自变量的约束，缺点是没有什么参数来显示预测的准确性。

4.2 多元线性回归分析

多元回归分析在使用数据分析的时候，要计算所有的自变量组合对应的拟合曲线。而针对一组特定的自变量，直接将多个自变量的所有取值作为自变量。观察所有自变量组做出的预测的调整后的 R 平方值，取这个值最大的一组自变量组合。

使用规划求解的话注意预测值的公式选择来达到不同自变量组合之间的选择。

预测后的值加入图标：使用选择性粘贴复制到图中。

4.3 非线性回归分析

将非线性的转化为线性的。注意相应的参数有可能的变换回去。