191250133 陶泽华

1. 数据集合说明

数据集

本次实验采用的数据集合是开源的美国的医疗费用个人数据集，共有1338条数据。数据集合包含数据由以下几列构成：

- 年龄：主要受益人的年龄
- 性别：保险承包商性别，女，男
- bmi：体重指数，提供对体重的理解，体重相对于身高相对较高或较低，使用身高与体重比的客观体重指数（kg / m ^ 2），理想情况下为18.5至24.9
- 儿童：健康保险覆盖的儿童人数
- 吸烟者：吸烟
- 地区：受益人在美国的居住区，东北，东南，西南，西北。
- 费用：由健康保险收取的个人医疗费用

我们对数据进行了格式化，将不是文本的数据全部替换成对应的数字：

1. 将性别为男的male替换为1，性别为女的替换为0
2. 将吸烟者对应的yes替换为1，不吸烟者对应的no替换为0
3. 将居住在东北的替换为0，东南的替换为1，西南的替换为2，西北的替换为3

2. OLAP分析工具

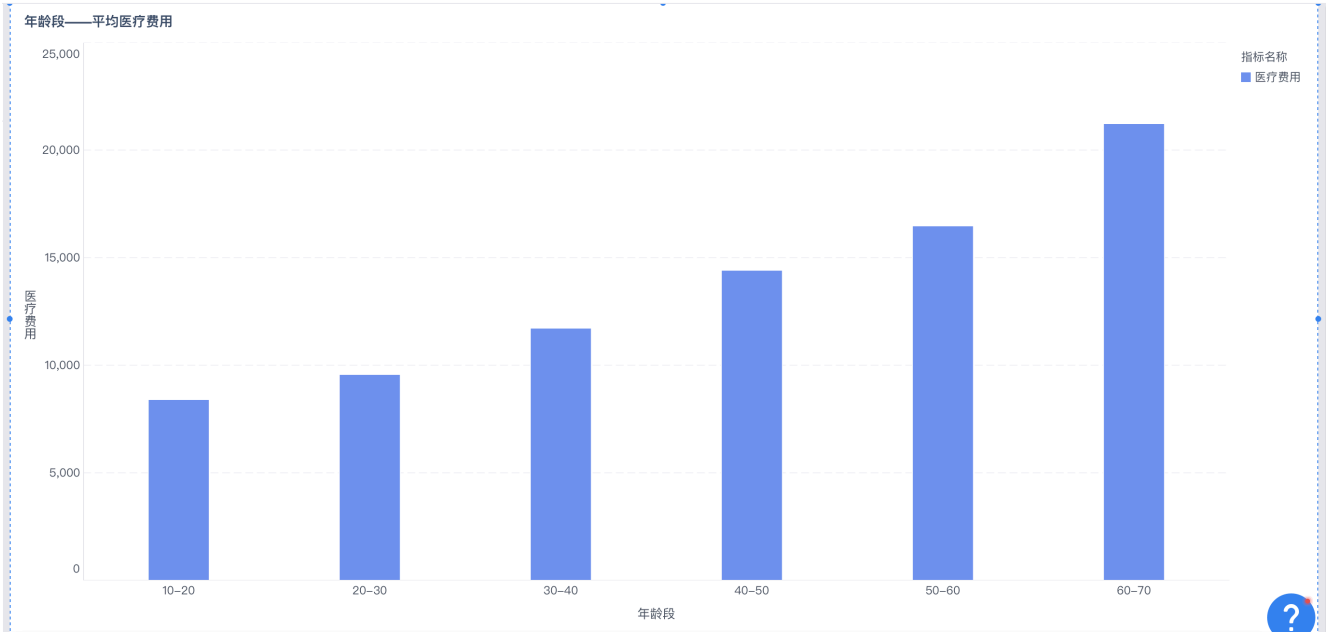
选择了 FineBI 作为 OLAP 分析工具，FineBI 是一款自助大数据分析的 BI 软件，该软件提供了诸多 OLAP 相关操作，使用图形化的方式完成 OLAP 分析。我使用 FineBI 作为工具完成对医疗费用个人数据集的分析。

3. 分析过程、结论和决策

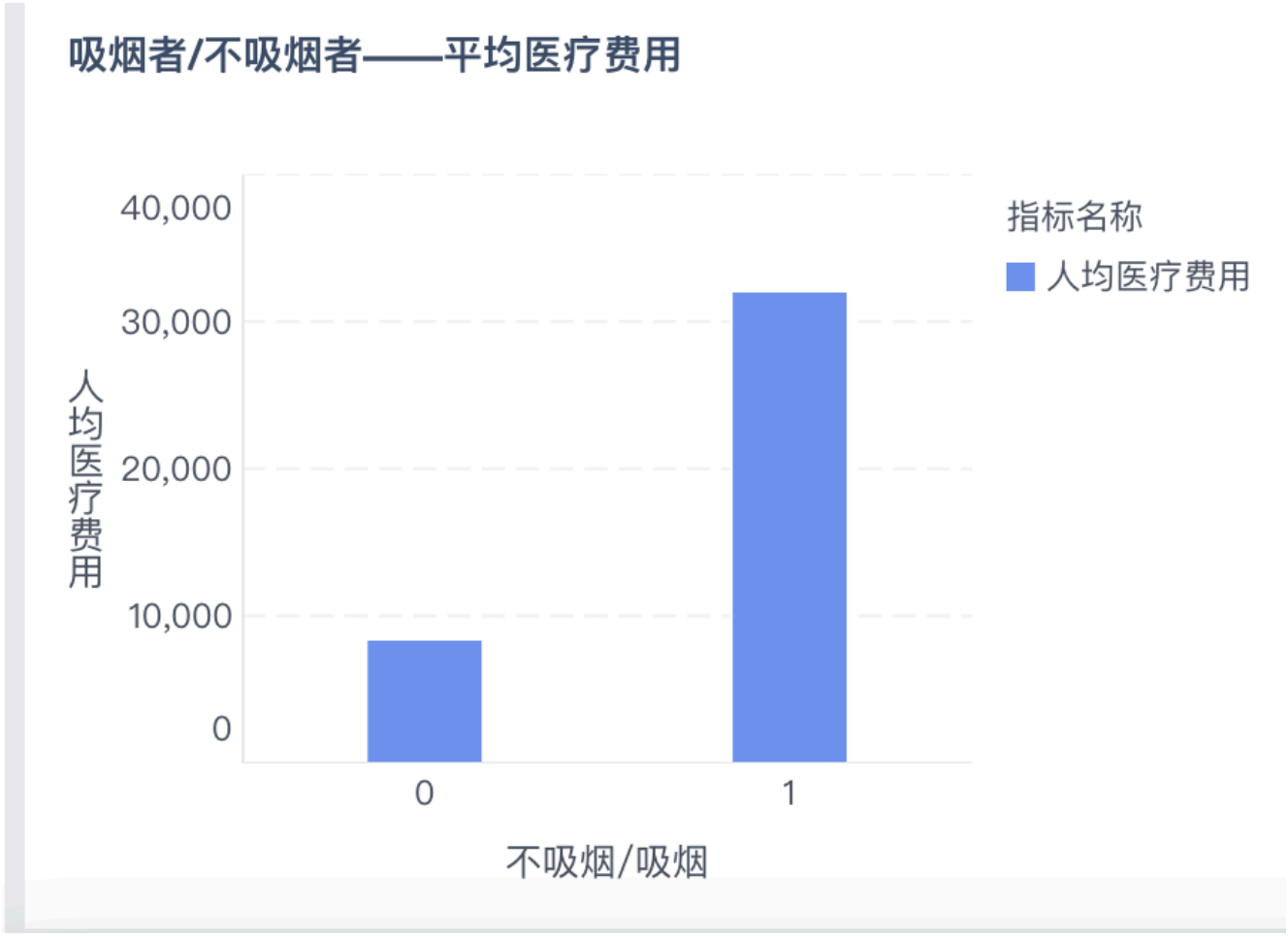
3.1 分析过程

分别对数据进行了以下几类分析

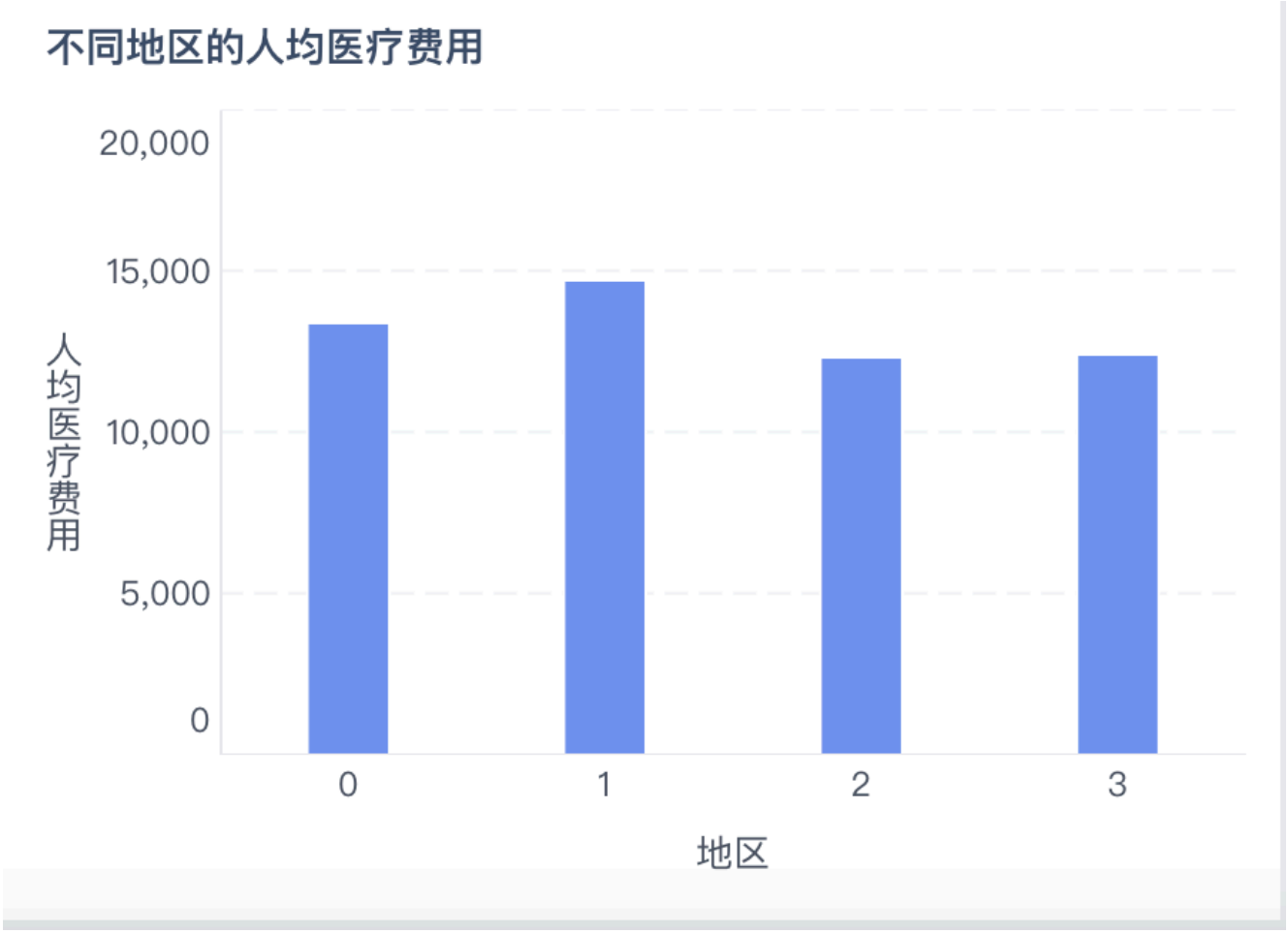
1. 对各个年龄段的人的花费的平均医疗费用进行分析，发现随着年龄的上升，花费的医疗费用也越来越高。



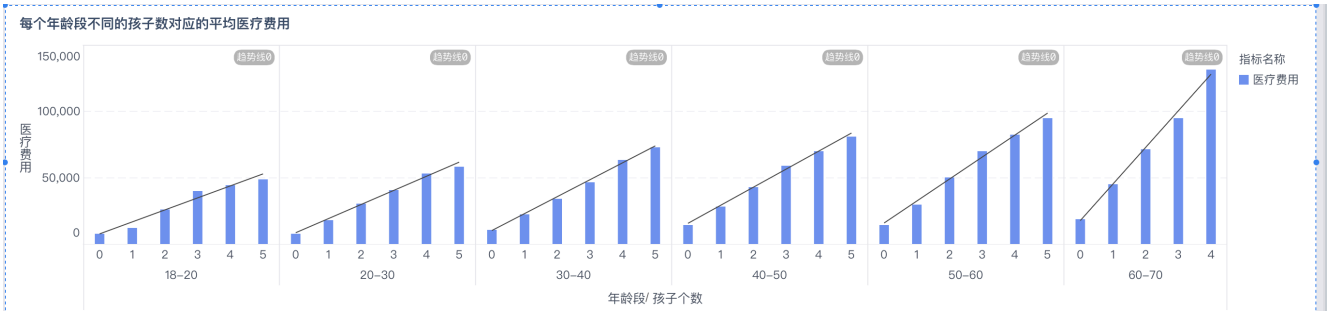
2. 对吸烟者与不吸烟者的话费的平均医疗费用进行分析对比，吸烟者花费的医疗费用显著高于不吸烟者花费的医疗费用。



3. 将不同地区的人均医疗费用，表格中0代表美国东北地区，1代表美国东南地区，2代表美国西南地区，3代表美国西北地区。各个地区的人均花费的医疗费用没有显著的差异。



4. 分析不同孩子个数对医疗费用的影响。总体来说，孩子个数越多，该家庭花费的医疗费用就越多。



3.2 分析结论

综合上面的分析过程，得到了如下结论：

1. 随着年龄的上升，花费的医疗费用也越来越高。平均医疗费用和年龄段是正相关的。
2. 吸烟者花费的平均医疗费用显著高于不吸烟者花费的平均医疗费用。
3. 不同地区的人均医疗费用没有较大的区别，较为平均。
4. 总体来说，孩子个数越多，该家庭花费的医疗费用就越多。孩子个数与家庭花费的医疗费用正相关。

3.3 决策

综合上面得到的结论，做出了如下决策：

1. 医疗保险承保商应该重点关注年龄较大、人数较多和有吸烟者存在的家庭，为他们提供更好的医疗保险服务，也能够增加医疗保险承保商的收益。
2. 若各地区的人均医疗费用产生较大差异，应该关注医疗保险承包商是否存在恶意竞争的行为，保证医疗保险能

够真正的为人民服务。