宠物新店选址

阮祥炬 2019/07

Step1:业务理解和数据理解

问题 1：需要做什么决定？

Pawdacity 是怀俄明州一家领先的宠物连锁店，在全州拥有 13 家商店。今年，Pawdacity 想扩大并开设第 14 家店。做决定之前需要根据现有数据来预测的年度销售情况，以便为 Pawdacity 的最新商店推荐合适的城市。

问题 2：需要什么数据来支持这些决策？

在做出正确的决策前，我们需要得到包含 Pawdacity 店的当前的数据集，其中所需的信息包含以下部分：

- 2010 年所有 Pawdacity 商店的月销售数据
- 怀俄明州 2010 年人口普查数据
- 土地面积
- 人口密度
- 18 岁以下家庭数量
- 家庭总数

当然除了这些信息还有下面的信息需要考虑：

1：城市中其他分店的销售额数据

2：城市人口数据总数

Step2:创建训练集

当前的所有的数据来源于下面三个数据集

1. 所有宠物店 2010 年的月销售数据
2. 美国人口普查工作室发布的人口数据
3. 怀俄明州每个市县的人口数据

观察所有宠物店 2010 年的月销售数据表，将 city 字段进行聚合，可以得到已经开设的 13 宠物家店分布在 11 个城市，然后将和这些城市相关最重要的六个字段进行聚合操作，可以计算出它们的总值和均值（下表所示），它们将对接下来的模型构建起到重要的作用。

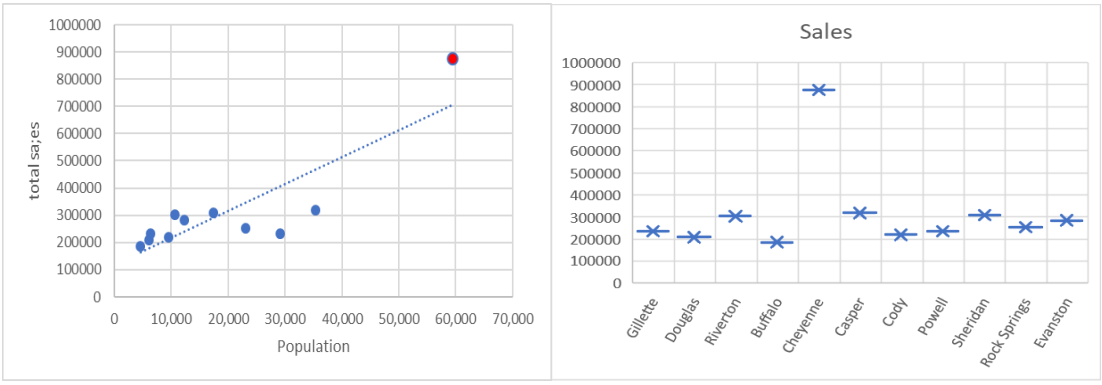
columns	sum	Avg
2010 Census Population	213862	19442.00
Total Pawdacity Sales	3773304	343027.64
Households with Under 18	34064	3096.73
Land Area	33071	3006.49
Population Density	63	5.71
Total Families	62653	5695.71

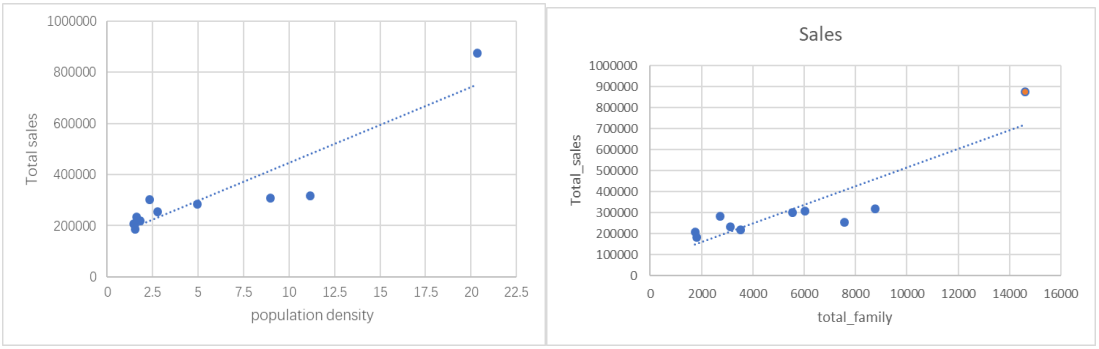
Step3:处理异常值

按照 step2 中得到的六个字段构建关于十一个城市相关的汇总表。然后通过计算四分位距（IQR）的方法,可以得到以下异常信息：

- 2010 年人口普查数据 Cheyenne 市的数据超出上限值
- 总销售额 Cheyenne 市的数据超过上限值
- 人口密度 Cheyenne 市的数据超过上限值
- 家庭总数 Cheyenne 市的数据超过上限值
- 总销售额 Gillette 市的数据超过上限值
- 土地面积 Rock Springs 市的数据超过上限值

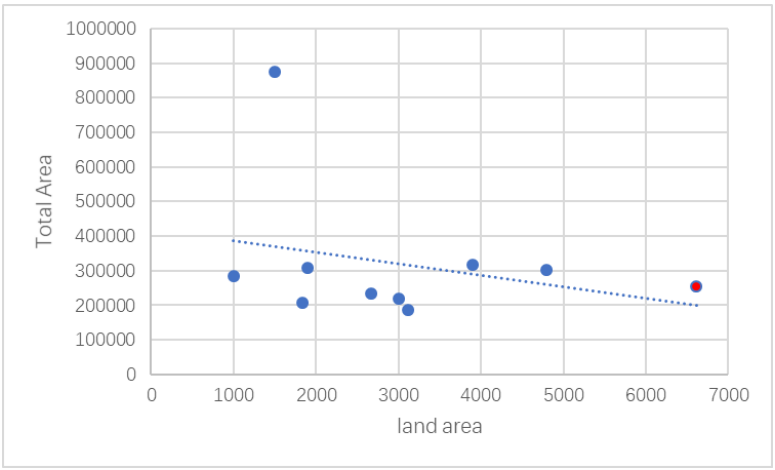
通过分析异常值我们发现。城市 Cheyenne（夏延）有四项指标高于平均标准，其中包括 2010 年的人口数量，人口密度，家庭总数，总销售额。但是由于 Cheyenne 是首府城市，所以它的总人口数，总人口密度，总家庭数比其他城市的范围高是正常的。因此在上面三个因素的影响下，总的销售额也是偏高的，但是和上面的三个因素是线性关系。因此保留和其相关的条目，以便为后续的人口密集的城市提供参考数据。





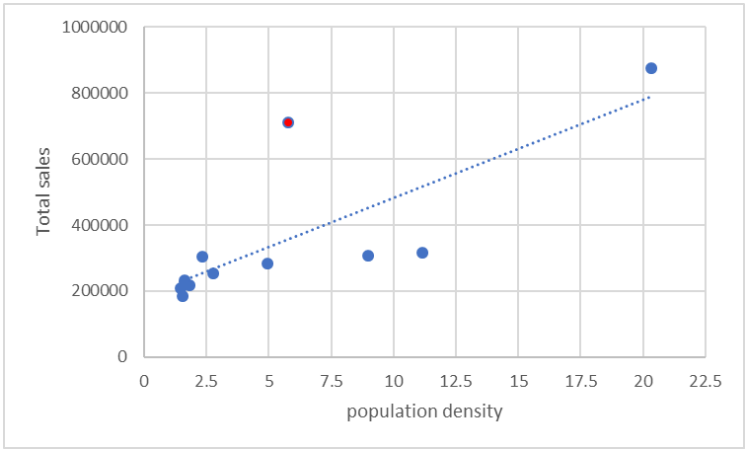
上面的四张图显示了 Cheyenne 在销售额，人口密度，家庭总数，人口数量远高于其他的城市。

Rock Springs 的城市面积有可能是一项异常值。但是，尽管他有一个更大的面积，但是他和其他的数据一样同总销售额之间具有线性关系。而且每个城市的面积存在客观差异，所以这不是一个异常值。



Gillete 在人口密度和销售额关系上可能是一个异常值。其他的城市的人口密度和总销售额都保持着线性关系，但是从散点图上可以看出，Gillete 的人口密度相对较低的情况下却有一个相当高的销售额，因此它的存在可能对于后续模型的构

建可能造成影响，所以应该将它剔除。

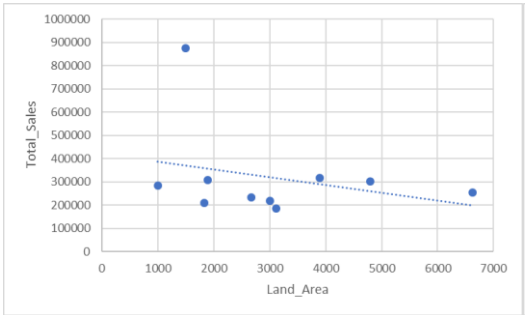


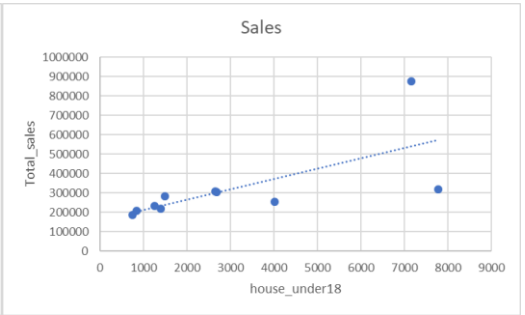
Step4:构造线性回归模型

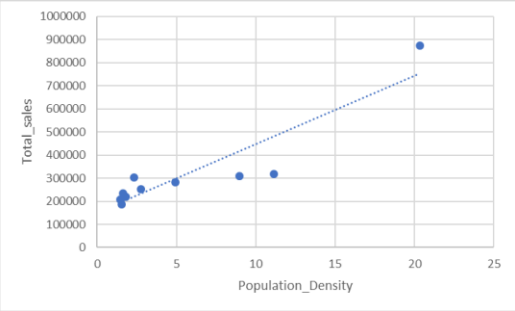
移除了 Cillete 的数据后得到下面这张用于后续预测的数据表，一共有 10 个城市。

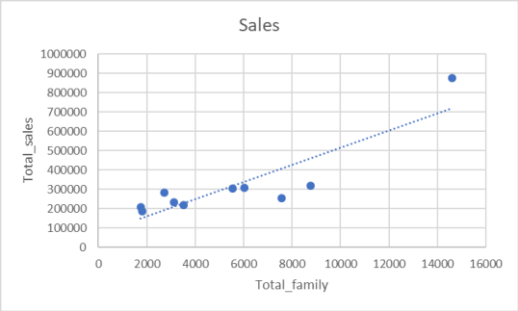
CITY	Land Area	Households with Under 18	Population Density	Total Families	Population	Sales
Douglas	1829.4651	832	1.46	1744.08	6,120	208008
Riverton	4796.859815	2680	2.34	5556.49	10,615	303264
Buffalo	3115.5075	746	1.55	1819.5	4,585	185328
Cheyenne	1500.1784	7158	20.34	14612.64	59,466	874908
Casper	3894.3091	7788	11.16	8756.32	35,316	317736
Cody	2998.95696	1403	1.82	3515.62	9,520	218376
Powell	2673.57455	1251	1.62	3134.18	6,314	233928
Sheridan	1893.977048	2646	8.98	6039.71	17,444	308232
Rock Sprir	6620.201916	4022	2.78	7572.18	23,036	253584
Evanston	999.4971	1486	4.95	2712.64	12,359	283824

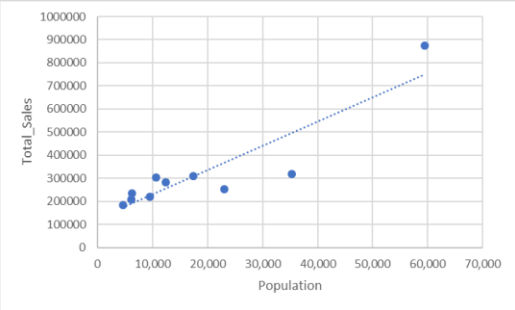
第一步需要做的就是确定上面的表格中哪些字段作为影响因子会对线性回归模型的构建有较大影响。为了得到最有效的预测，自变量必须和我们需要预测的因变量（总销售额）之间有相关联的线性关系。使用双变量分析法来分析各变量和销售额之间的关系。











通过双变量分析法得到上面的散点图，但是发现这些自变量和销售额之间都存在着线性关系，无法排出任何一个自变量。

但是由于一些自变量之间本身可能存在着关联，所以可以使用双变量分析法研究一下其他变量之间的关系，以便进一步排除无关变量。

	Land Area	Households with Under 18	Population Density	Total Families	Population	Sales
Land Area	1	0.189375819	-0.317419204	0.107304205	-0.0524699	-0.28708
Households with Under 18	0.189375819	1	0.82198575	0.90566006	0.91156245	0.674652
Population Density	-0.3174192	0.82198575	1	0.891680268	0.94438856	0.90618
Total Families	0.107304205	0.90566006	0.891680268	1	0.96919023	0.874663
Population	-0.05246991	0.911562446	0.944388558	0.969190231	1	0.898755
Sales	-0.28707758	0.674651999	0.90618038	0.874663379	0.89875464	1

虽然看起来土地面积好像不受其他变量的影响，所以先选择土地面积作为一个自变量。但是其他的变量之间有很强的相关性，最后选择家庭总数作为另一个自变量。最终得到下面所示的一个线性回归模型。

Predicted sales = 197330.41 - 48.42*(land area) + 49.14 * (total families)

Step5:执行分析

通过以上几步，我们找到了用于分析的线性回归方程，下面我们需要结合下面的具体要求，来选择合适的开店地址。

1. 新店应该位于一个新的城市。
2. 新城市其他商店的总销售额低于$500000。
3. 新城市人口必须超过 4000 人口（2014 年的数据）
4. 预估的年销售额必须超过$200000。
5. 选择预测合集中有最高销售额的城市。

根据以上要求筛选出了下面四个城市：Jackson，Lander，Laramie，Worland。
下面是他们的基本情况：

City	2014 Estimate Pop	SALES VOLUME	Land Area
Jackson	10,449	110000	1757
Lander	7,642	108197	3346
Laramie	32,081	76000	2513
Worland	5,366	100000	1294

结合线性回归模型可以得到下表中的预测销售额，四个城市的销售额都大于$200000,因此选择金额最高的城市 Laramie，最高的预测销售金额为$305036.47。

City	2014 Estimate Pop	SALES VOLUME	Land Area	Total Fami	Predict Sales
Jackson	10,449	110000	1757	2313	225917.29
Lander	7,642	108197	3346	3876	225783.73
Laramie	32,081	76000	2513	4668	305036.47
Worland	5,366	100000	1294	1364	201701.89