数据科学介绍1.1.1： Exploratory Data Analysis，探索性数据分析

在我们开始之前，我们先了解一下数据科学，机器学习到底是要做什么事情，有哪些基本概念。之后我们要学习一种不需要任何建模，就可以对我们的数据有一定了解的方法，就是Exploratory Data Analysis（EDA）。

什么是数据（Data）

数据是个很广的概念，在数据库科学里面我们一般认为数据是信息的一种表示方法。**数据一般是通过观测获取的，一套数值的、定性或者定量的描述某一个东西的变量。**比如，“他长得好帅”，这个就不是一个很好的数据，如果说，“他的颜值系数=7000”，或者说“他的颜值属于超帅这个等级”这些就是数据了。但是如果你说“我猜小丽肯定喜欢他”这个也不是数据，应为这个不是观测得到了。

样本 Sample

说到数据有个十分重要的概念，就是**样本。** 刚刚说数据是通过观测得到的，这个观测的过程也可以叫做采样，就是Sampling。每**一次Sampling过程，也就是每次观测得到的结果就叫做样本。**采样是统计学里面的技巧，也偶很多种方法来保证你获取的样本等代表整体（population）的规律我们这里就不讲了。

我们的数据可以认为就是样本的集合，如下面的表格，是病人的基本信息和她十年内中风的概率。每一行，就是一个样本。而每一列代表的就是一个描述这个样本的变量，我们就把它叫做特征。有时候除了特征，样本里面还会有标签，这个标签，一般就是我们未来需要预测的变量。

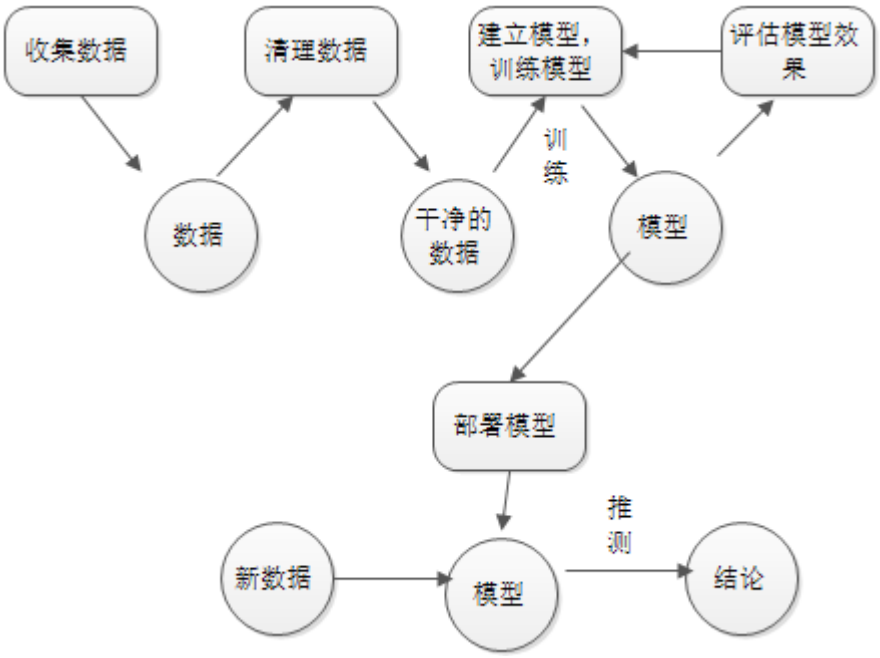
特征

Age x1	Blood Pressure x2	Smoker x3	% Risk of Stroke over Next 10 Years Y
63	129	No	7
75	99	No	15
80	121	No	31
82	125	No	17
60	134	No	14
79	205	Yes	48
79	120	Yes	36

样本

标签

数据科学的基本套路



数据科学的套路基本就如上图，首先我们收集数据，也就是采样，然后我们把这些数据整理一下，因为数据里面可能有很多我们用不上的特征或者是搞错了的样本。然后再把这些干净的数据送到一个算法里面，他可以帮我们用这些数据建立一个模型，这个过程叫做**训练，train**也可以叫做 **Fit or Estimate。**当我们得到满意的模型了，然后把新收集的道数据给这个模型，他就能给我们一些关于这个数据的结论，比如预测未来，或者对奇怪的东西经行分类，这个过程叫做**预测，predict**，也可以叫做**inference， forecast。**

数据的种类

刚刚说了什么是数据，就是对一个东西定性或者定量描述的变量。他有很多类型，我们这里简单介绍一下：

类型数据与连续数据

类型数据，Categorical data： 类型数据，就是他是有限取值的，只有几个有限的类型，比如性别，只有男和女，最多还有个不男不女，但是没有70%男这种说法。类型数据还分为**Nominal**和**Ordinal**，两种。Nominal就是种类是没

有顺序的，比如，刚刚说的性别，男，女，不存在先后。但是**Ordinal**是有顺序的，例如，成绩分为 A，B，C，三等，那么A就是比B和C要成绩好，而B是比C好，比A差。

连续数据，Continuous data： 连续数据就是，和他名字说的一样，他的取值有无限多中可能，它分为两种，**Interval**和**Ratio**。Interval的意思就，不太好解释，你可以理解为就是一个数值，没有绝对0点，但是有绝对的单位。就是假设，Trump智商60，艾因斯坦智商120，我智商180，那么爱因斯坦比Trump聪明多少我就比爱因斯坦聪明多少，但你不能说我比Trump聪明3倍，因为，没有绝对0值。而Ratio和Interval没啥区别，但是他又绝对的0点，比如20%就是10%的2倍。

Exploratory Data Analysis（EDA）

EDA是一种初步的，不需要建立任何模型就可以让我们对我们的数据建立初步的认识，找到总体的规律。通常EDA是通过数据可视化的方法完成的，就是画图。

通过EDA我们可以大概的知道我们数据有什么特性，这能帮助我们后面更有效地建立模型，有时候甚至能直接得到很多有用的规律，得到比模型更有用的信息。

我们下面导入一个1985年汽车大全的数据做一下EDA试试。

描述性统计

我们首先卡看哪些变量属于哪些类型。可以看到里面很多是interval类型的变量，比如length，width，horsepower。然后有些例如make，就是牌子，他是nominal类型的变量，然后horsepower-binned，是马力大小的分类，这个就是ordinal的数据。

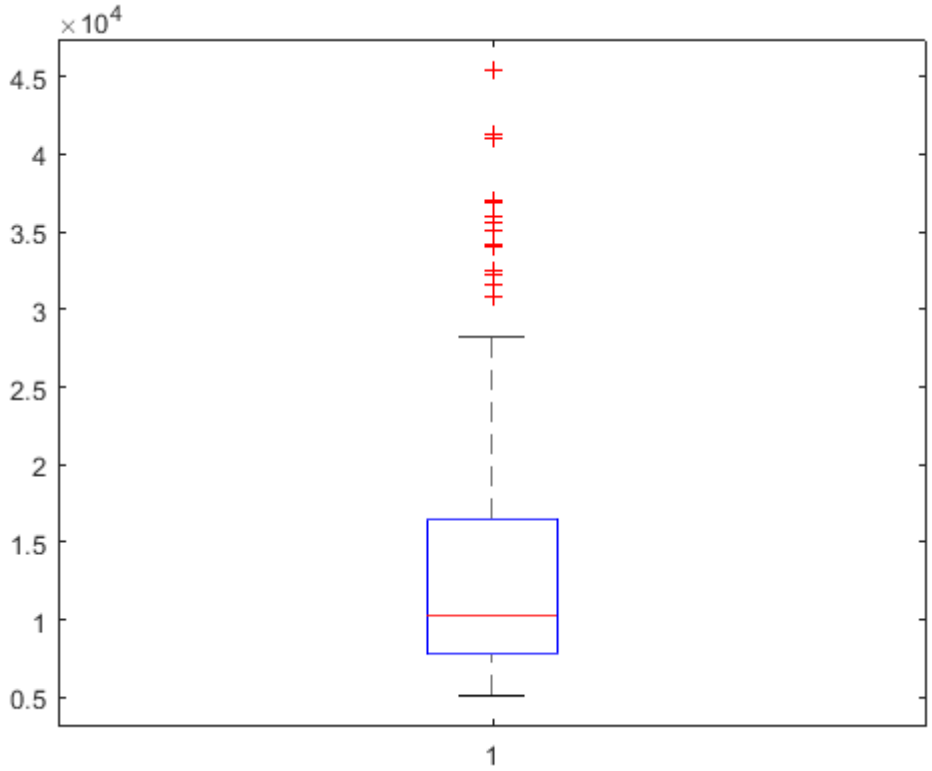
对于interval的变量，这个很简单，大家都知道，对于一个连续数据我们想知道他大概是个什么特性我们可以用最简单的几个变量描述，平均值、中位值、范围、还有Variability。

我们可以读取那个数据文件到一个表格里面，然后运行 `summary(table)` 这个指令，得到一个数据的基本描述，里面有那么一点有用的东西。

```
auto_summary.price
mean(auto.price)
std(auto.price)
```

通过上面的指令就可以呃看到汽车加个的范围（Max，Min），中位数（Median），以及平均值。std是计算标准差的，标准差可以表示数据分布式比较松散还是集中。不过这还不是很直观，我偶们可以用box plot来visualize一下。

```
boxplot(auto.price)
```

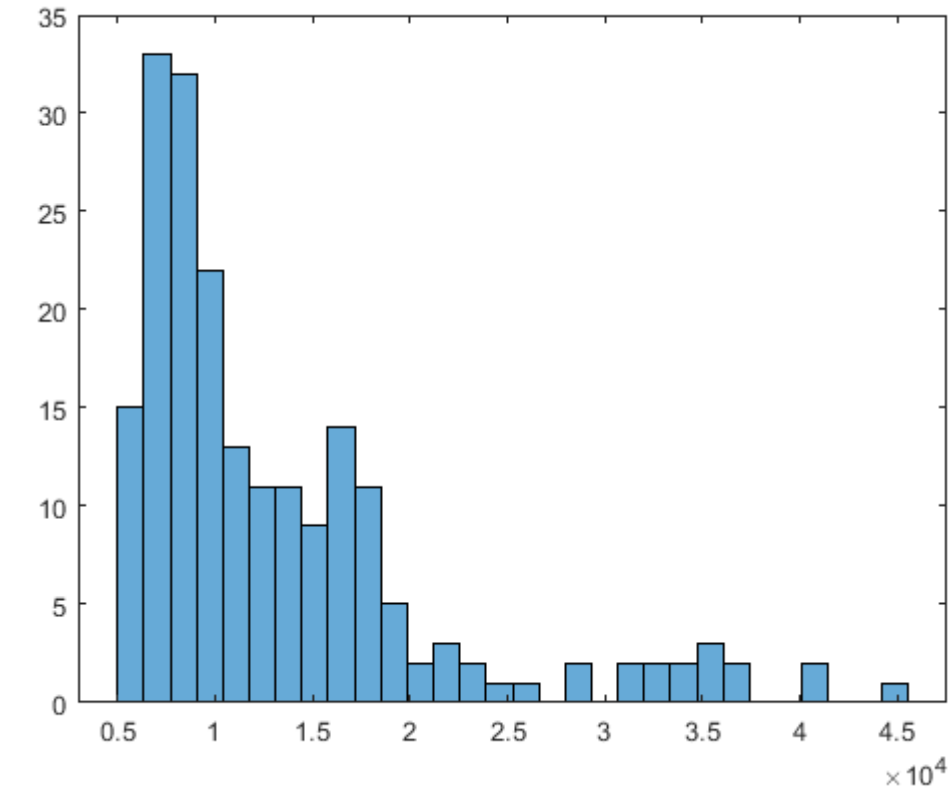


这个图里面中间的线是中位值，也就是50%的值，然后数据按照大小和数量分为4块（Q1~Q4），方框是-+25%的值，就是25%，75%。最外面两条线就是正常数据上下限。但是不含异常值（outlier）。在上下先以外的就是异常值，小于Q1上限1.5倍Q3-Q1（IQR），或者大于Q4下限1.5倍IQR。

这样大概可以看出一些数据的分布，也就是可以了解一下variability，可以看到高端车真的是很贵很贵，比普通车贵的多。

如果想进一步了解他的分布，可以画出直方图：

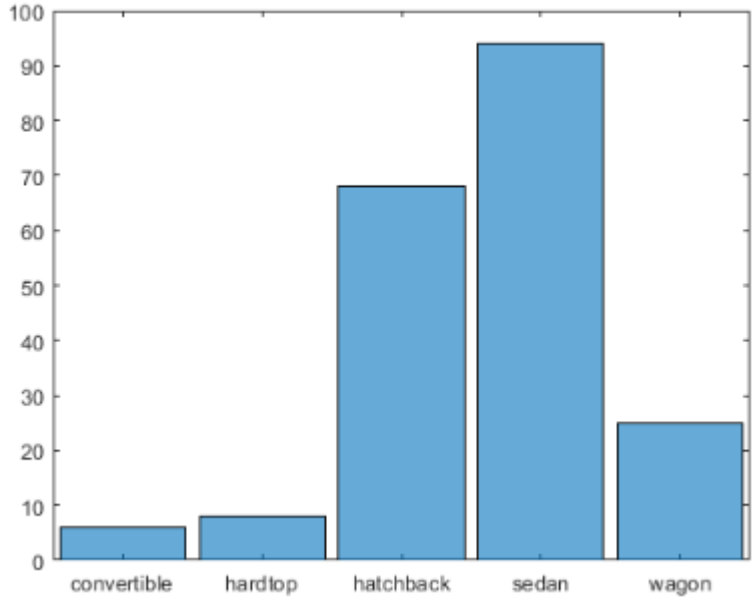
```
histogram(auto.price,30)
```

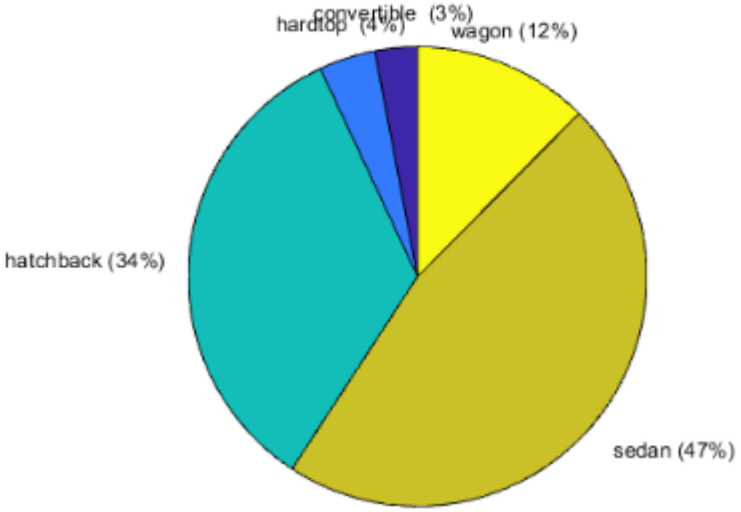


刚刚我们讲的是连续数据，对于类型数据，我们可以用bar chart 或者pie chart来看看。

例如我们想看看车字有哪些body style，例如掀背车、旅行车、敞篷车这些。

```
auto.body_style=categorical(auto.body_style)
histogram(auto.body_style)
pie(auto.body_style)
```





大部分的车都是轿车。

这个对大部分Categorical数据，包括nominal和ordinary都是有效的。

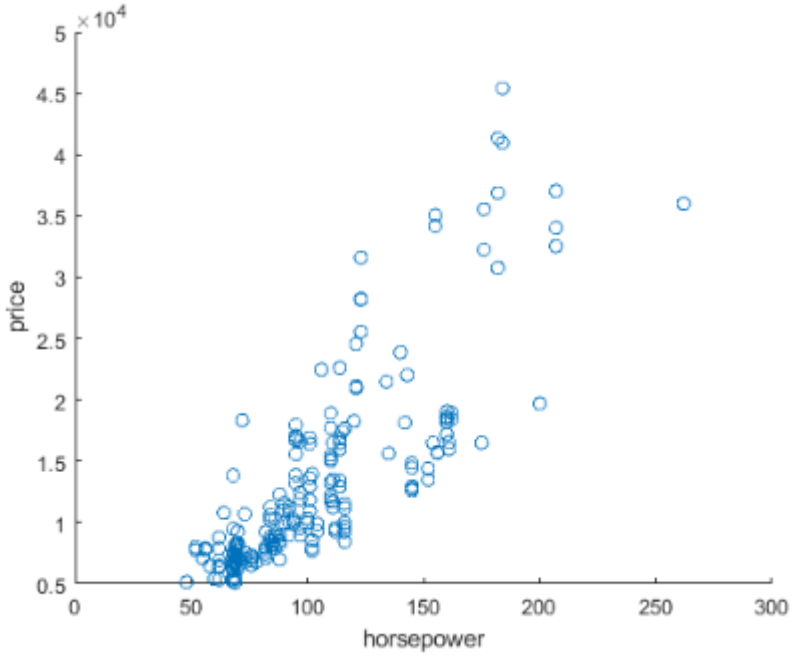
多个变量间的关系

刚刚只看了一个变量，但是数据之间是有联系的，那我们想干多个变量的关系怎么版内？

连续变量

首先对于两个都是连续变量的画就用"相关图就可以了"，他就是把两个变量分别华仔x，y轴上看联系，这个大家都会，例如下面，马力和售价的关系：

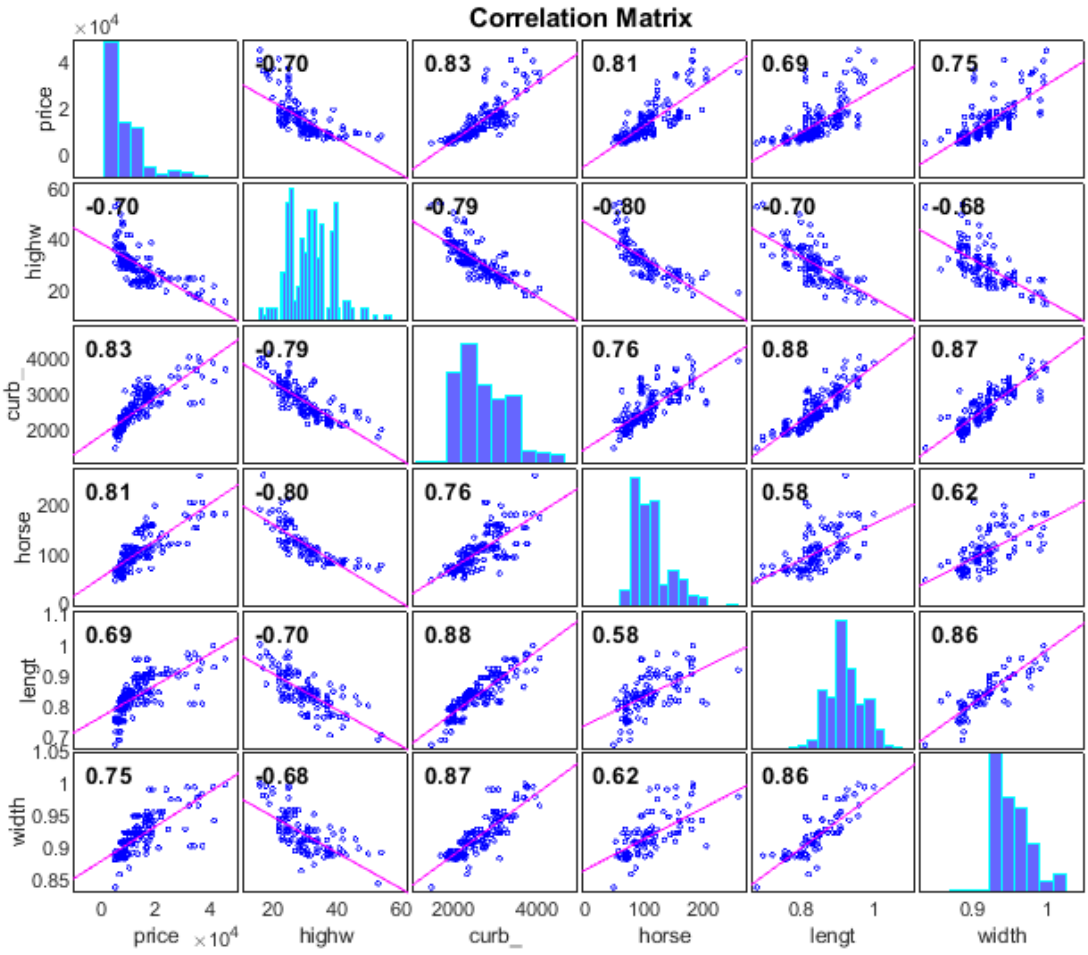
可以看出来玛丽越大的车子价格也越贵。




```
scatter(auto.horsepower,auto.price)
xlabel('horsepower')
ylabel('price')

corrplot(auto(:,["price" "highway_mpg","curb_weight" "horsepower" "length","width"]))
```

我们可以用correplt函数画出来多个变量之间的关系矩阵：



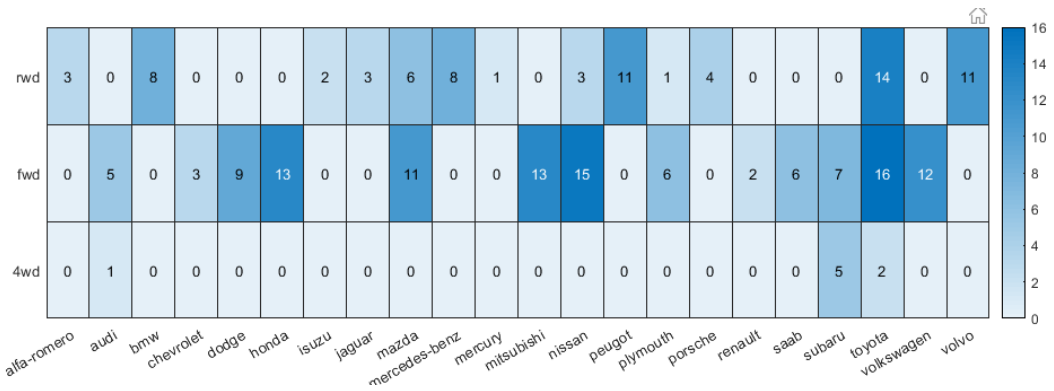
上面这个图，是对角线对称的，对角线上，是自己和自己的关系，肯定是1没有任何意义，他就化了直方图代替。可以看出他给出了相关系数，可一条直线，那条直线的斜率就是县官系数。你可以看书那两个之间有关系是正相关还是负相关。越接近于1月相关。我们可以看到我选的这几个变量之间还是挺相关的，例如越费油的车越贵，月中的车越贵，还有显然的，越宽越长的车越重。

类型变量

那如果是两个类型变量呢？怎么看关系？这个就得用Contingency Table，它用来展示两个类型变量之间的交叉的关系。撅个例子，我们的数据里面有这些车是什么牌子的，和他是前轮还是后轮驱动的。我们最这两个变量做一个contingency table就可以直到，这些拍字，有多少前轮或者后轮驱动的车了。

```
[tbl,chi2,p,labels] = crosstab(auto.drive_wheels,auto.make )

hm=heatmap(tbl)
hm.XDisplayLabels = labels(:,2);
hm.YDisplayLabels = labels(1:3,1);
```

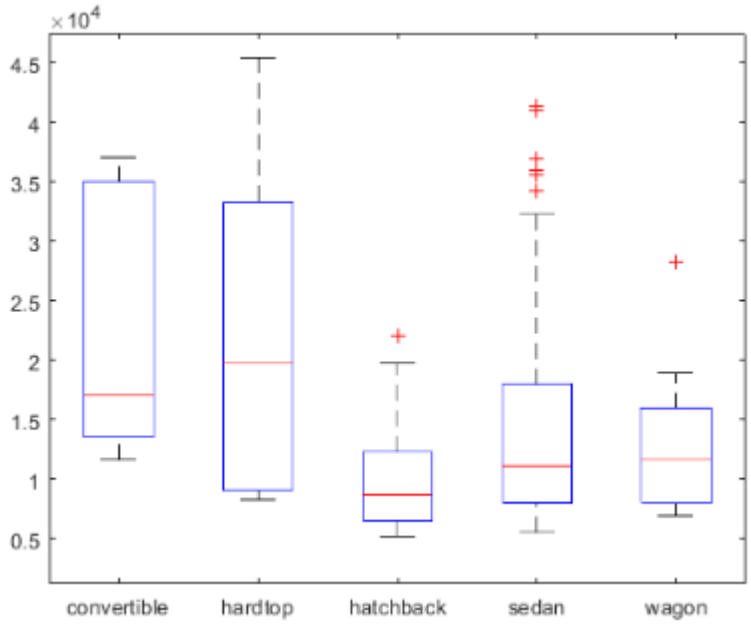


我们可以看出honda，大众这些人名的品牌，全是前轮驱动的车，而宝马、奔驰这种土豪车牌子，全部都是后轮驱动的车。而4驱车只有奥迪和Subaru还有丰田（🤔）在做，这个表很有意思把。

类型变量和连续变量间的关系

这个时候我们就又要回到box plot了。例如我们想看看各种类型的车，他的售价怎么分布的?

```
boxplot(auto.price,auto.body_style)
boxplot(auto.price,auto.drive_wheels)
```



可以看到敞篷车可以买到很贵，而掀背车，就是golf，AE86这些，就是便宜货。

ok，EDA就讲这么多，其实还有很多好玩的东西尤其是数据可视化。

大家还可以看看这些介绍：

https://www.kaggle.com/fazilbtopal/exploratory-data-analysis-with-python/data