基于faithful数据集的数据可视化分析

小组成员

2052320 朱凡

数据集介绍

A data frame with 272 observations on 2 variables.
老忠实喷泉，美国黄石公园中最著名的间歇泉，每隔一段时间（45-125分钟）就喷发一次，历时约4分钟，高度达40-50米

```
> names(faithful)
[1] "eruptions" "waiting"
```

数据集有两种数据，eruptions为喷发的时间，waiting为两次间隔的时间。

```
> dim(faithful)
[1] 272   2
```

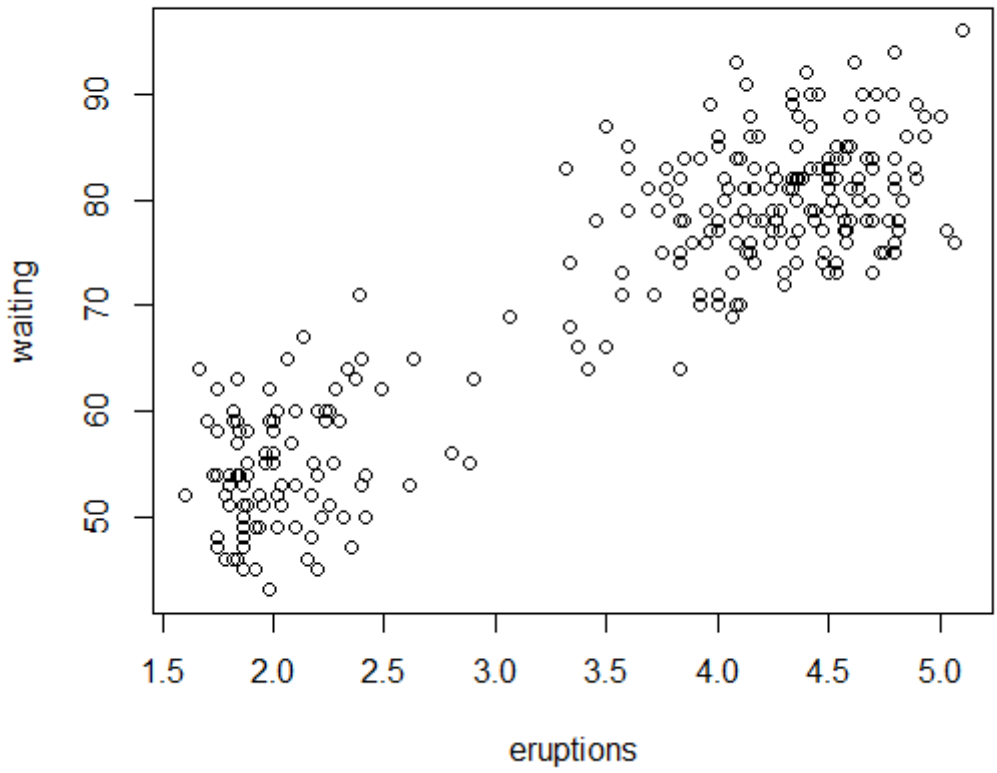
数据集共有两维，有272组数据。

```
> head(faithful)
  eruptions waiting
1     3.600      79
2     1.800      54
3     3.333      74
4     2.283      62
5     4.533      85
6     2.883      55
```

数据可视化分析

图1

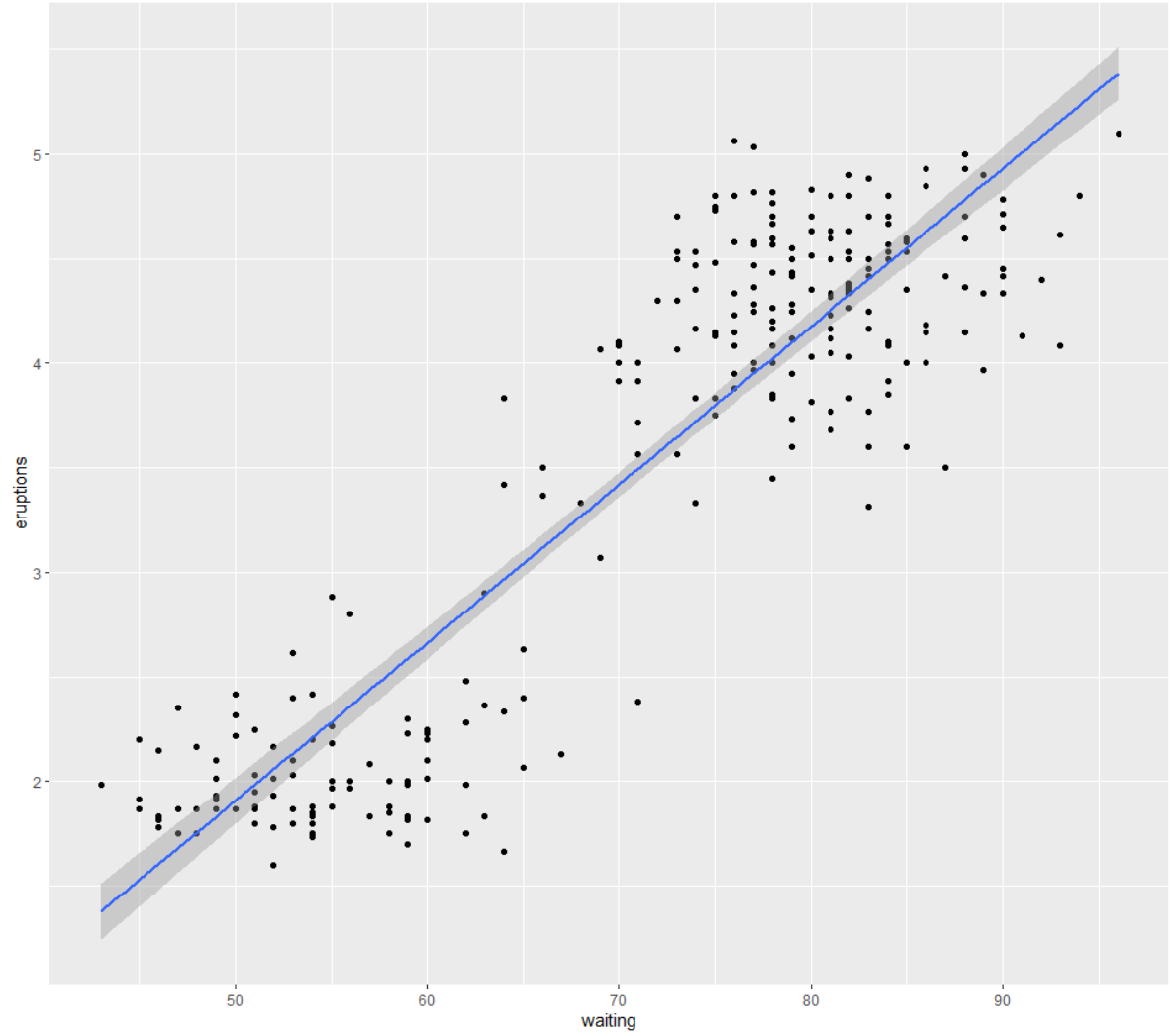
```
plot(faithful)
```



作默认散点图，从中可以粗略看出等待时间和喷射时间的关系的，大致是喷射时间越长，两次喷发的间隔时间越长；也可以说是两次喷射的间隔时间越长，那么下次喷射的时长越长。

图2

```
ggplot(data=faithful,aes(x=waiting,y=eruptions))+
  geom_point()+
  geom_smooth(method=lm)
```

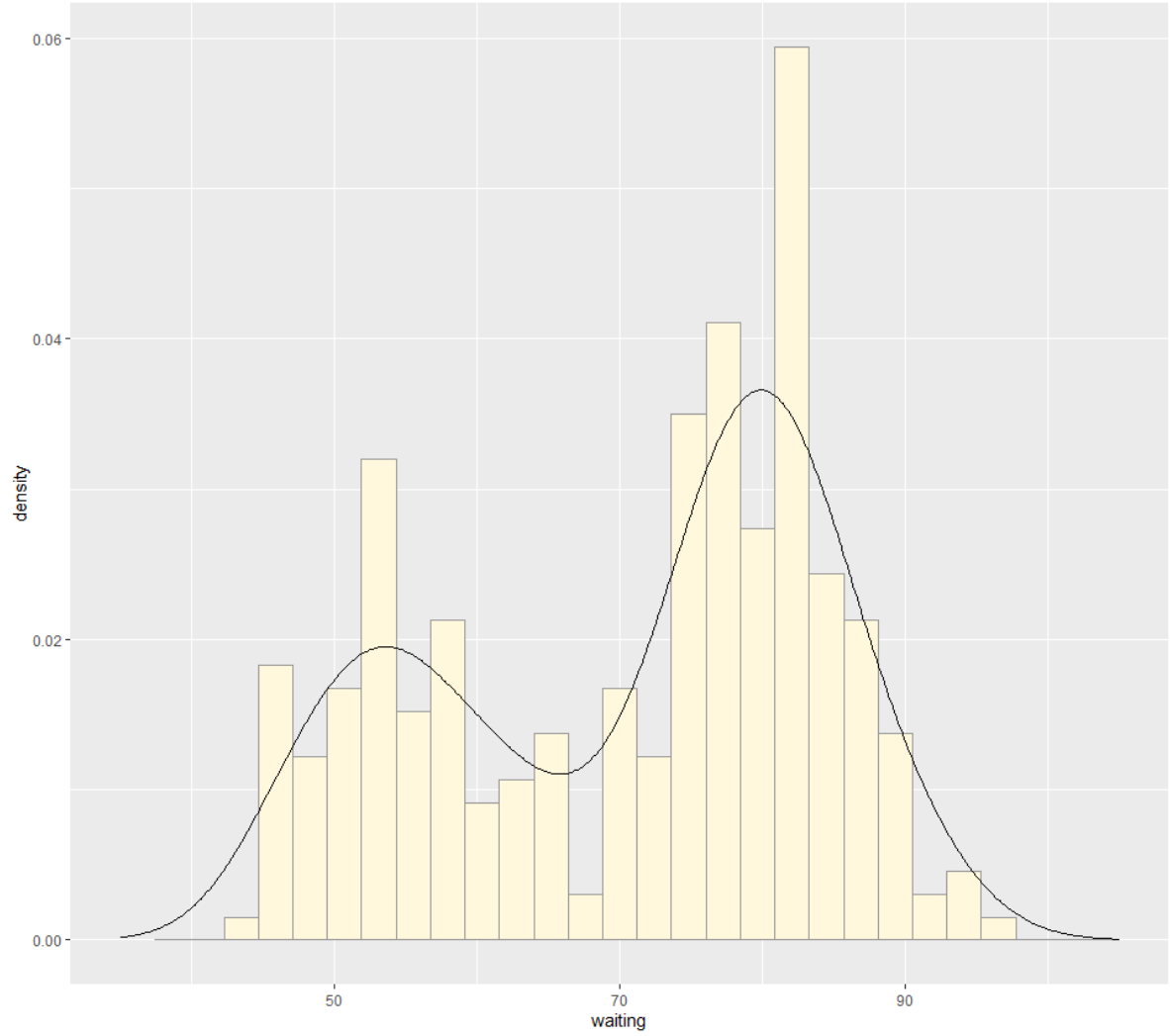


作者认为蓄力越久，喷射越久的逻辑更加合理，因此将等待时间作为横轴进行调研。

用平滑直线连接后，可以发现喷射时间的大小基本与等待时间成正比。这个结论也比较符合人们的常识。

图3

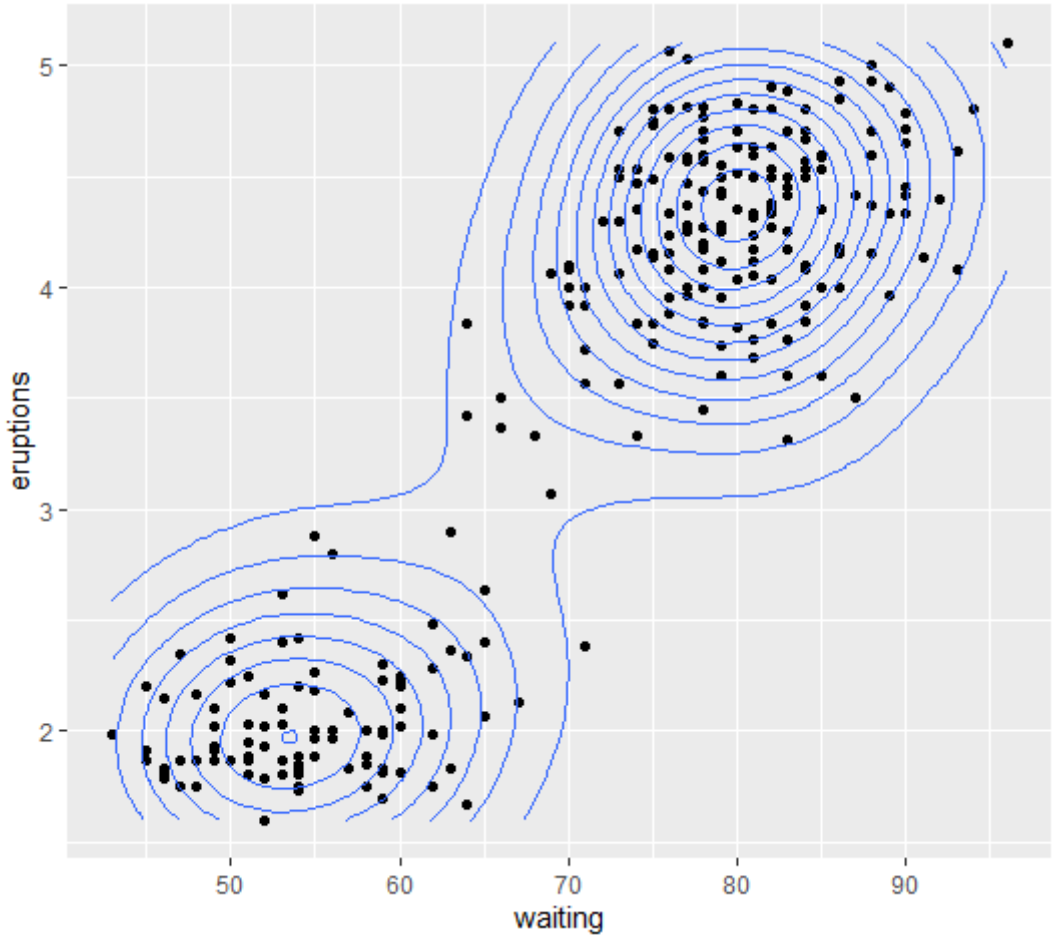
```
ggplot(data=faithful,aes(x=waiting,y=..density..))+
  geom_histogram(fill="cornsilk",color="grey60",size=0.2)+
  geom_density()+
  xlim(35,105)
```



作waiting的频率直方图和密度曲线，可以看出等待时间在55分钟左右和80分钟左右有一个峰值，等待时间大多在这两个值周围。

图4

```
ggplot(data=faithful,aes(x=waiting,y=eruptions))+
  geom_point()+
  stat_density2d()
```



从这图更能看出数据分成两个簇，围绕这两个簇中心的数据更加密集，除此之外的数据就比较稀疏了。

调查结论

- 图2 说明，喷射时长与间隔时长基本成正比；也就是说，等的越久，老忠泉喷的也就也久。导游可以根据等待时长，判断是否是更加激烈的喷射，以便游客观赏到更加美丽的喷泉。
- 图3和图4说明，间隔时长分布在55分钟和80分钟左右的居多。而且间隔时长没有少于四十分钟的或者大于100分钟的。那么游客在喷泉结束后，至少要四十分钟后才可能看到喷泉。如果只是想随缘，那么在50-60分钟或者75-85分钟的时候来看比较合适。
- 综合来看，游客若不想错过喷射，可以选择早点到来，在上次喷射后四十分钟后就前来观赏，这样基本不会错过喷射奇观。当距离上次喷射55分钟后还没喷射后，可以选择70分钟的时间点到达，这样大概率能够看到激烈的喷射景象。
- 老忠泉作为自然景观来说，喷射的时间还是比较具有规律的。