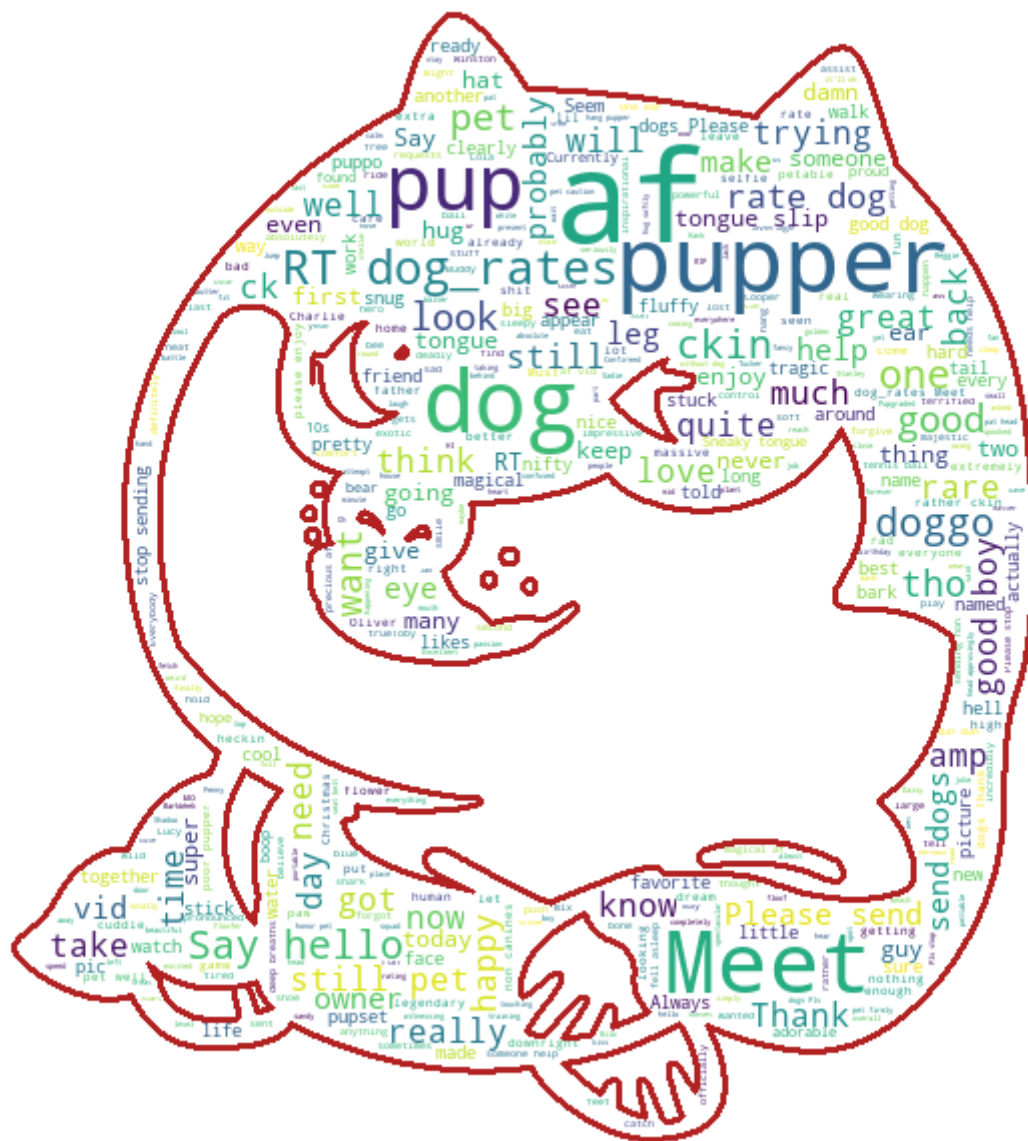


在分析与可视化阶段，我提出了几个自己感兴趣的问题并寻求解答，下面是关于这几个问题的一些总结。

在@WeRateDogs推文中，哪些词出现的次数较多？

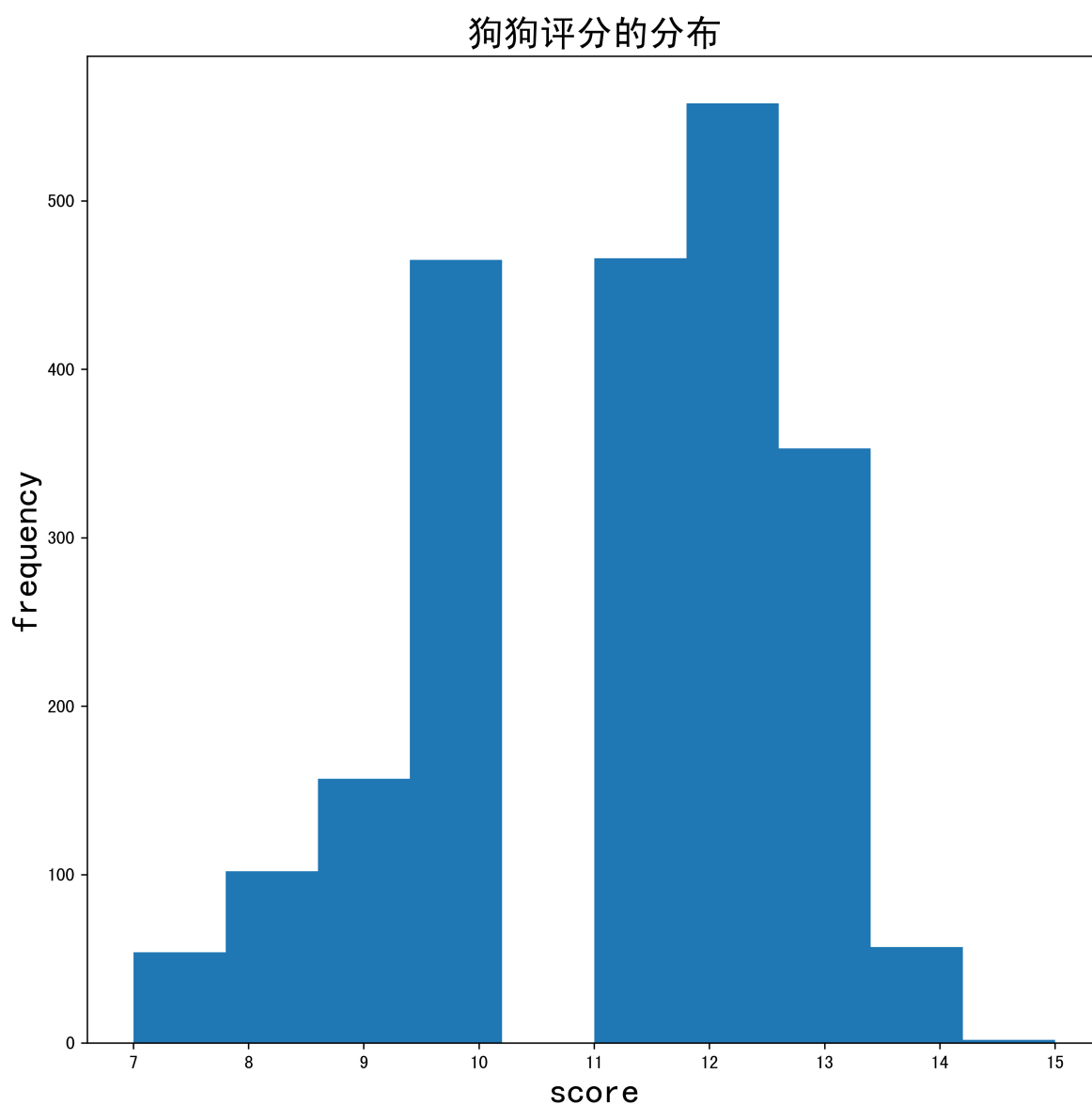
这是一个充满趣味的问题。推文以文本形式保存在 DataFrame 的一个变量中，wordcloud 模块是解决此问题的一个有效方法。这是我用 wordcloud 模块获得的可视化：



在该图中可以看到，推文中出现较多的词有：'Meet'、'dog'、'pupper'、'af'、'pup'等。

推主对狗狗的评分集中在哪个区间?

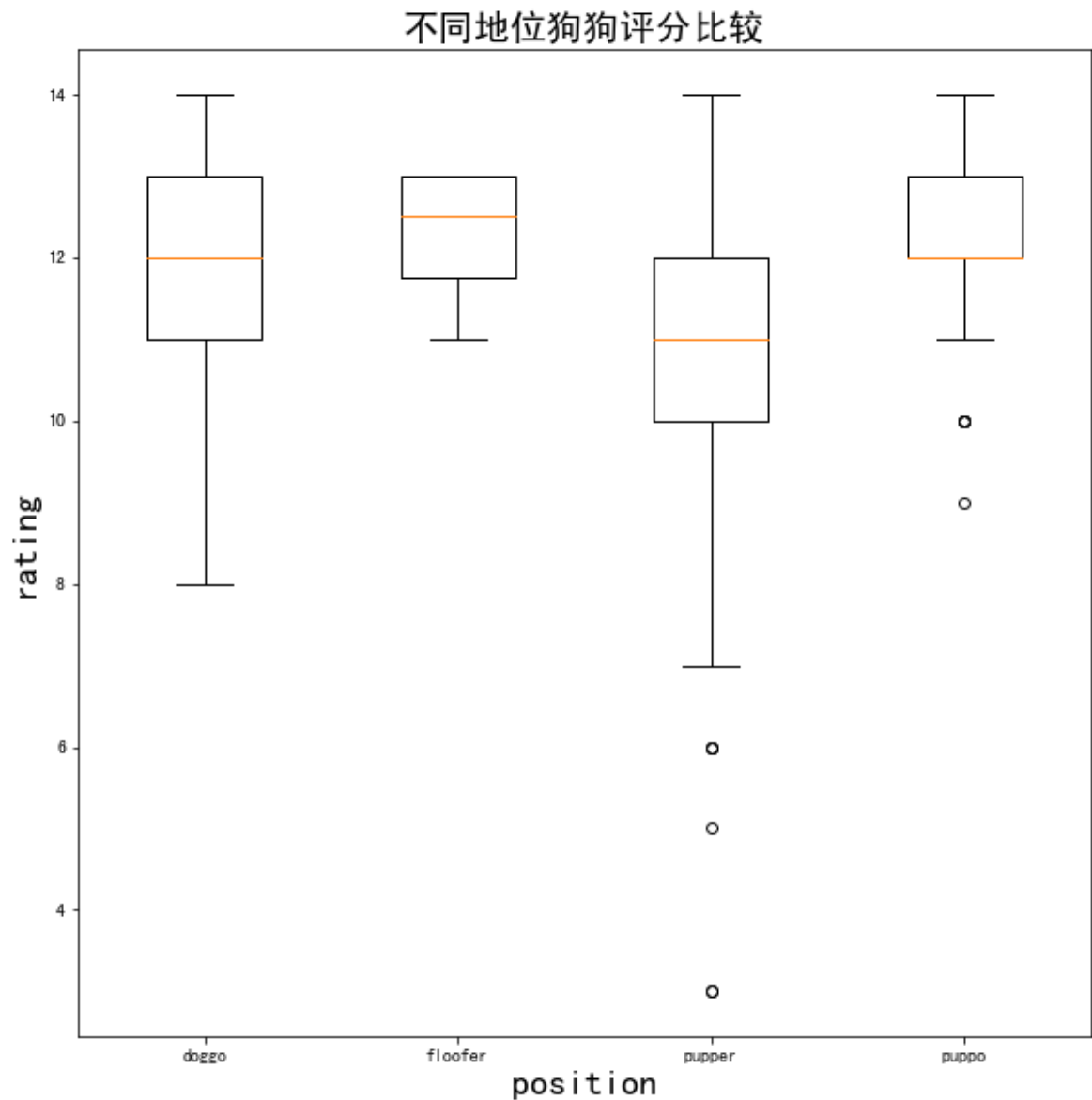
这里选用直方图表示频数分布。但一个问题是狗狗评分中可能存在一些或高或低的异常值。这里我将与25%分位数和75%分位数距离1.5个IQR以上的数据点定义为异常值。由于大部分数据点落在25%分位数和75%分位数之间，这种处理方法得出的结果应该还是比较有代表性的。这是我得到的图表：



狗狗评分集中在10分、11-12分之间。9分以下、14分以上的评分就比较少了。

推主推文中哪种地位狗狗的评分较高？

这里借助盒须图将四种地位的评分进行并列比较，对比效果很明显：



从图中看到：

1. floofer 平均得分最高（但 floofer 记录数很少），其次是 doggo，pupper 记录数最高，得分也最少
2. puppo 的25%分位数和中位数重合，说明大量该地位狗狗拿到了同一评分
3. pupper 评分变异程度最高，floofer 评分变异程度最低

此外还可以观察到其他的现象，但需要指出的是，由于有地位记录的狗狗数量较少且不均衡（floofer有4个，pupper有248个），以上结论可能会有偏颇。同样的思路在数据充足的情况下解释能力更好。