项目背景

清洗 WeRateDogs 推特数据，创建有趣且可靠的分析和可视化。这份推特档案很棒，但是只包含基本的推特信息。要达到 "Wow!" 的效果，在分析和可视化前，还需要收集额外的数据、然后进行评估和清洗。

我的探索

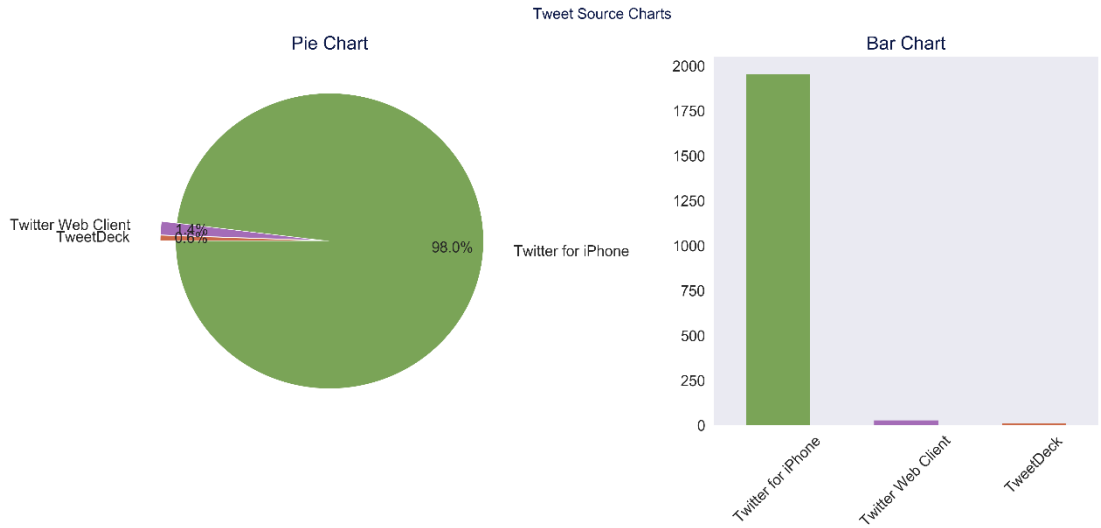
我通过数据整理（收集数据，评估数据，清洗数据），之后对所得数据进行可视化分析，探索了以下几个问题

1. 推主哪一类的客户端使用较多？
2. 哪种 stage 的狗狗最多？
3. 评分，喜爱和转发三者两两关系是怎样？
4. 哪种品种的狗狗的评分次数较多？
5. 常见的狗狗名字？
6. 狗狗名字几个字符的居多？
7. 评分最高的狗狗？
8. 不同种类的狗狗的平均评分

问题探索

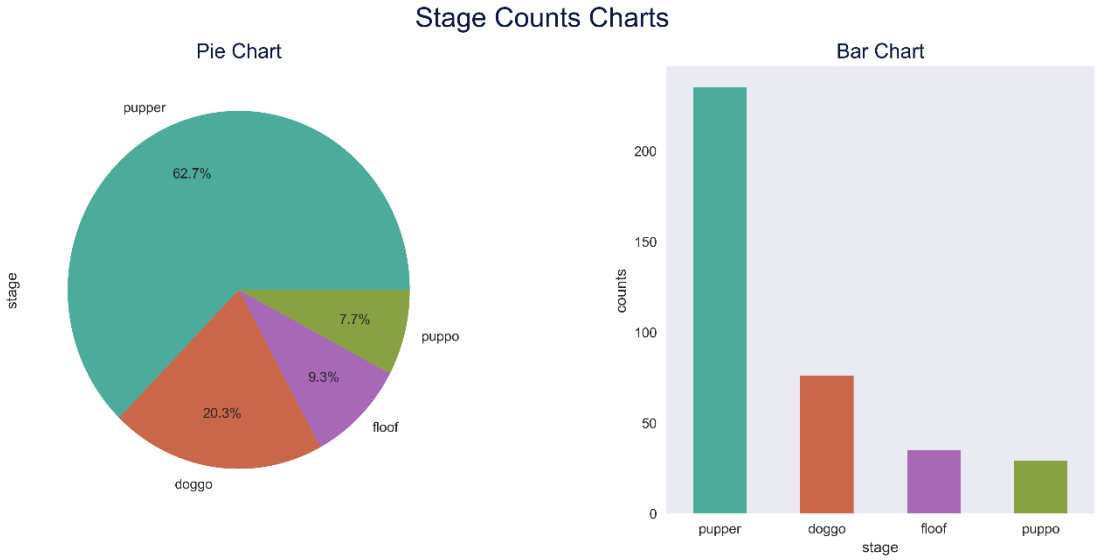
1. 推主哪一类的客户端使用较多？

我对推主每次发送 tweet 所用的客户端进行了一个统计计数，作图如下



可以看出，推主 tweet 的主要方式是使用 iphone 手机客户端，占比高达 98%，这是很好理解的，手机轻巧，方便外出携带，自带拍照上网功能，看到有趣的狗狗，随时拍照记录，即拍完可上传评分，十分便捷

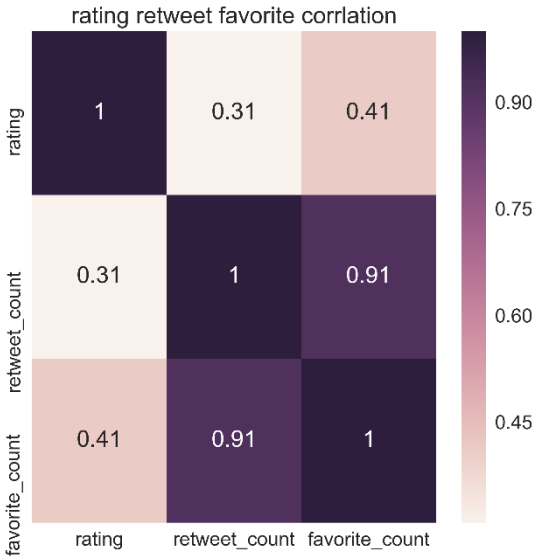
2. 哪种 stage 的狗狗最多？



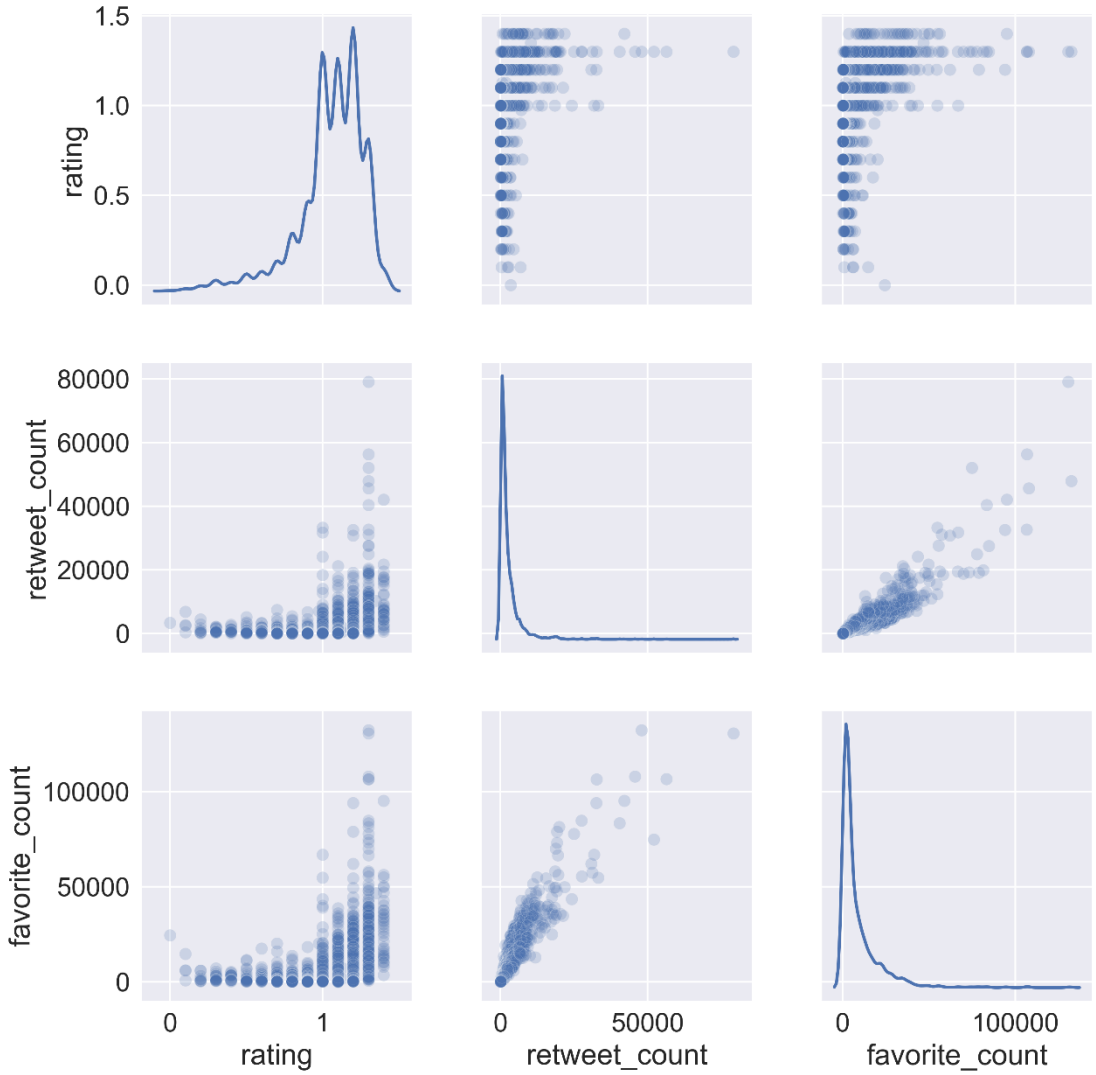
可以看出狗狗 stage 为 pupper 的居多 ，占比达 62.7%,其次是 doggo，20.3%，最后两位分别是 puppo， 占比 9.3%， 还有 floof， 占比为 7.7%

3. 评分，喜爱和转发三者两两关系是怎样？

我求得评分，喜爱和转发三者之间两两的皮尔逊相关系数，然后将其放置于下图中。数字越大，代表这两者的相关性越强，也就是说，某一方增加，另外一方也会跟着增加的概率更大。



下面这张图则是以上三者两两之间一个散点图，斜对角线则是一个核密度估计图。



从热力图或者散点图,可以看出，转发数和喜爱数是呈正相关的，两者的皮尔逊相关系数高达 0.91。

转发数和喜爱数相关性极高，这个很好理解，因为一般喜爱的非敏感话题，大家更有可能会彼此分享到各自的圈子，让其扩散，而各自的圈子里的人，大多与自己喜好相同，继而被自己圈子的人喜爱和转发，而扩散之后，又可以被点赞和转发，接着循环如上。喜爱和转发两者的按钮设计的又十分靠近，方便快捷。让点赞和转发几乎同时发生。

评分（这里指分子/分母的比值）与喜欢和转发数之间关系相关性虽然没有没有转发数和喜爱数之间的那么强烈，但是也是较高的，分别为 0.411237 和 0.308543。

关于评分与转发数的相关性较低原因，我的理解是人与人之间的差异性和相似性造就了这个结果。

* 差异性，对狗狗评分的不同，这里其实可以通过查看一些 tweet 的评论发

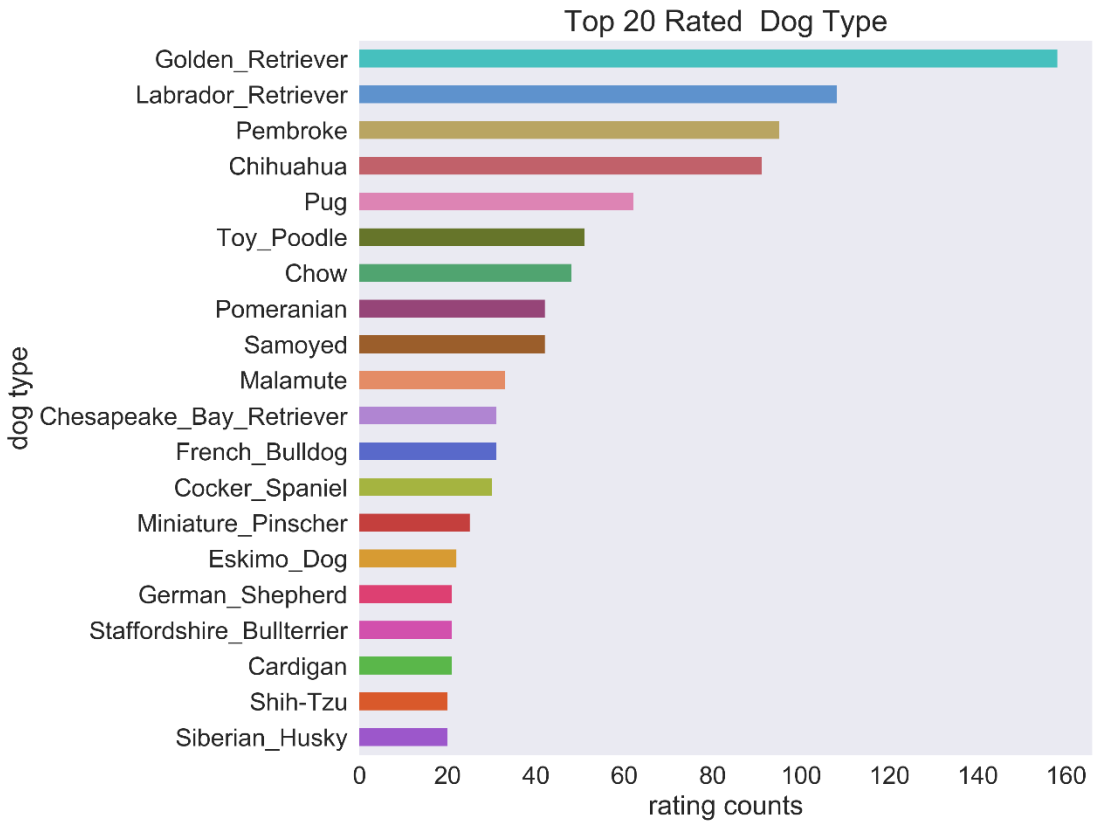
现，多数人的评分都各不一样。

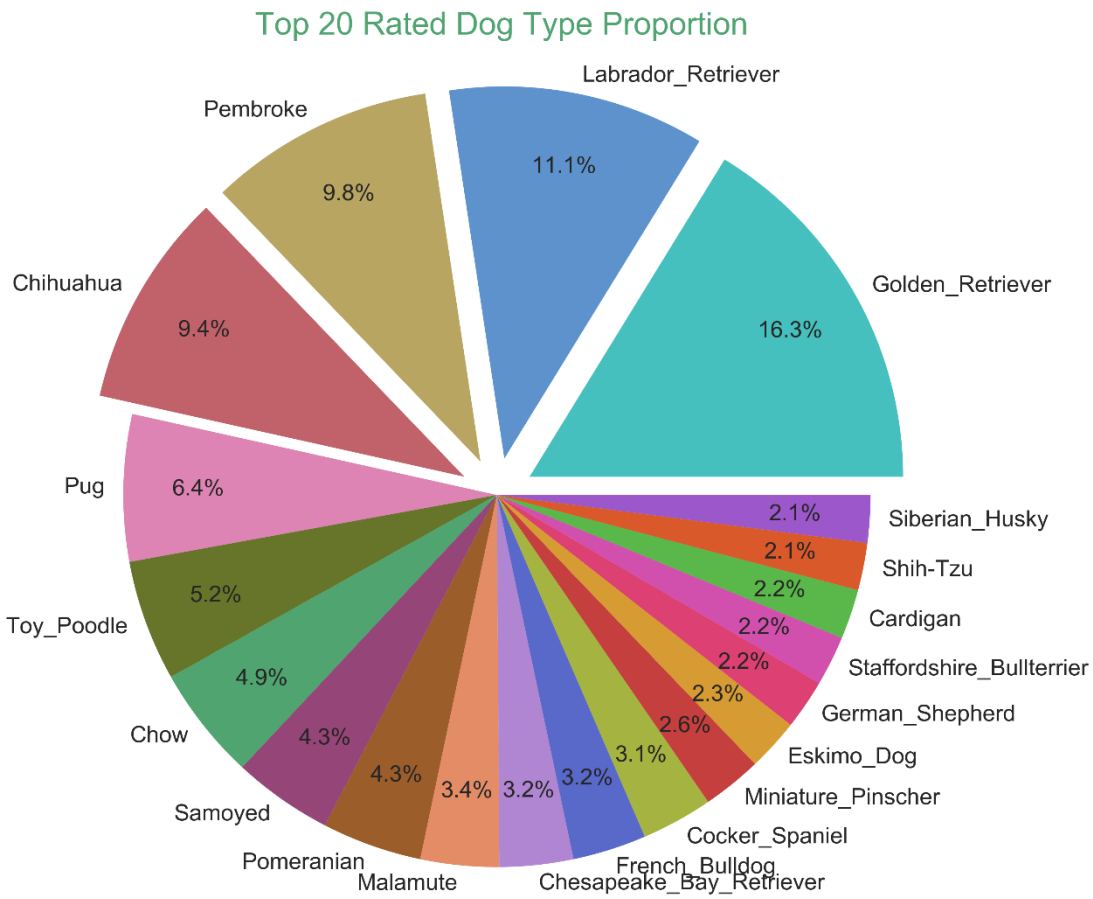
* 转发 tweet 的人跟推主的评分可能不一直，比如，同一张，图片，推主评分 12/10，那么有的人可能较为喜欢的 14/10，低一些的 10/10，都有可能

* 相似性，对狗狗某些特征的有共同的喜爱，推主的评分代表了一部分人的评分，这里我称之为相似。

具体点说：人们转发是出于自己的喜爱，而喜爱，是因人而异的。这里做个思想实验，假设人们的喜爱同推主的一致，那么，根据上面喜爱和转发数的关系，相关系数高达 0.91，再假设评分反应了推主的喜爱程度，那么可以推断出，评分越高，喜欢的程度就越高，转发的概率也就更高，也就表明，评分和转发相关性很强。现在只改变一个条件，假设人们的喜爱不大一致，那么，推主的评分高的话，只能代表推主的评分，不能代表全体人们内心的评分，反应在相关性上，比之前弱了些，也就是相关系数要更小些。

4. 哪种品种的狗狗的评分次数较多？





这里我作了评分次数排名前 20 的狗狗的柱状图和饼状图

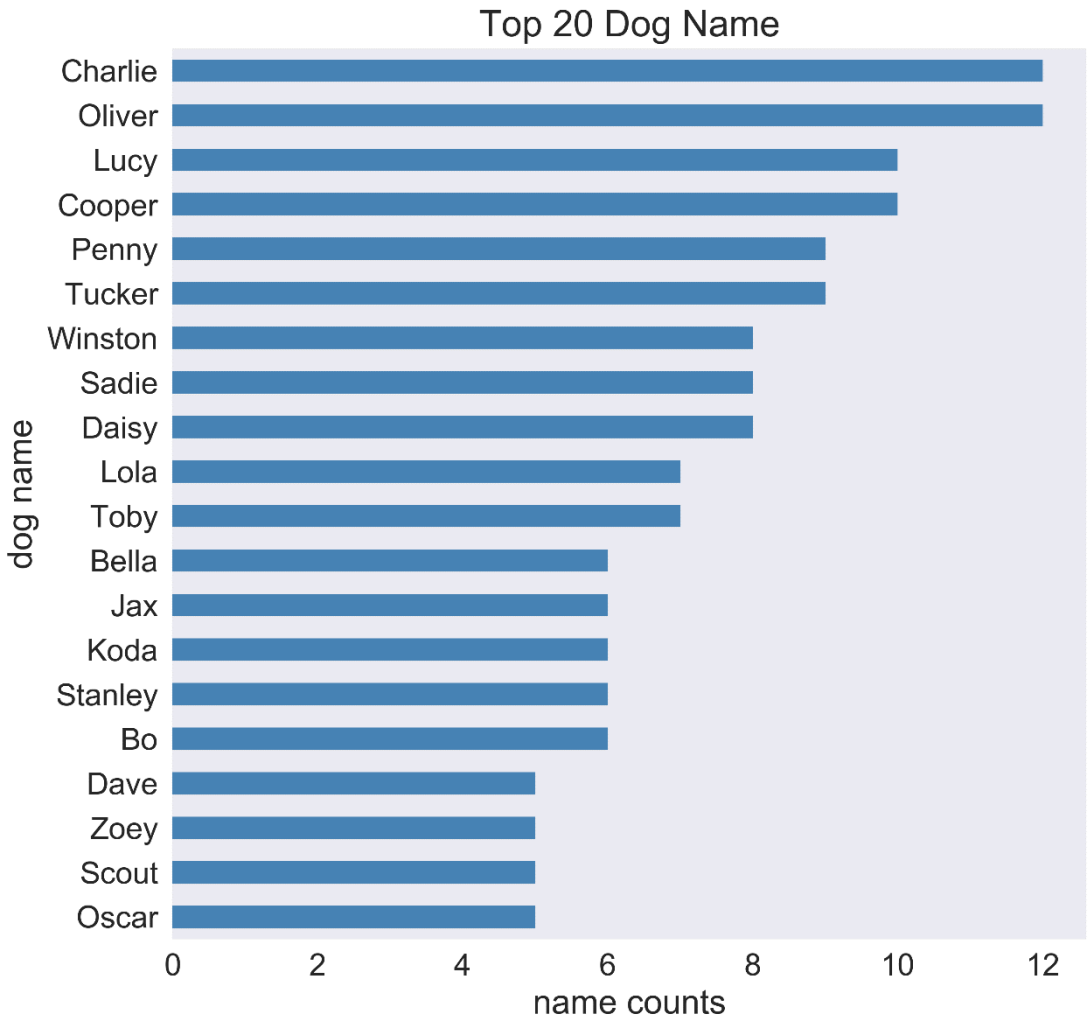
评分次数排名前四的分别为：Golden Retriever，Labrador_Retriever，Pembroke，Chihuahua,评分次数分别为 158，108，95，91 而不同品种的狗狗评分次数的平均评分次数为 14.92，标准差为 22.4，可以看出分布还是较为离散的。

我的理解如下

- 可能推主更喜欢这些种类的狗狗，也可能推主遇见此类狗狗较多
- 还可能是，预测狗狗的程序，对这类狗狗更好预测，导致预测结果中，这类狗狗的占比较大

5. 常见的狗狗名字？

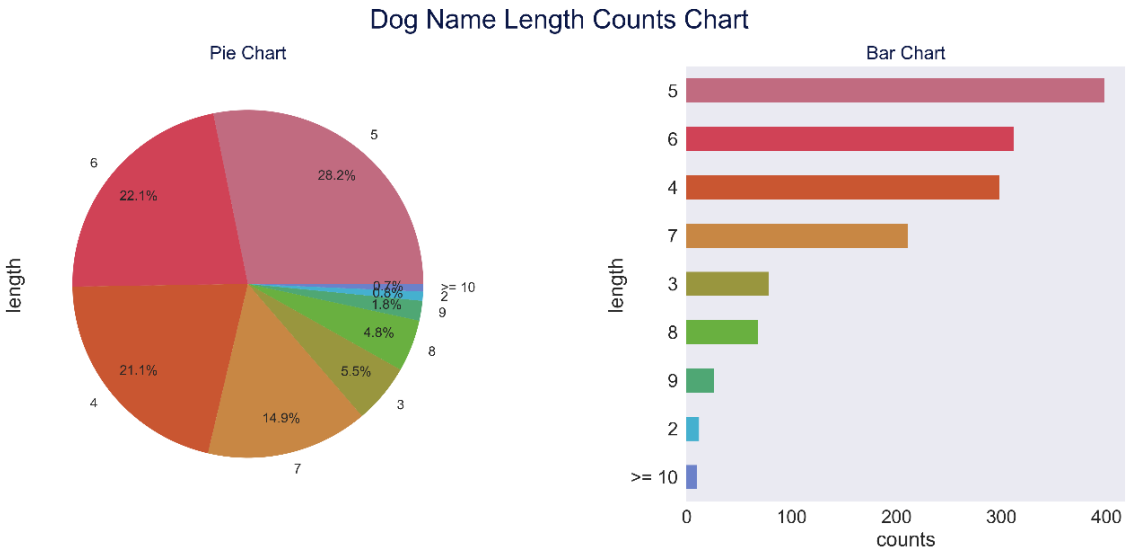
常见的狗狗名字排序如下，前四名（降序排序）分别为：Charlie, Oliver, Lucy 和 Cooper

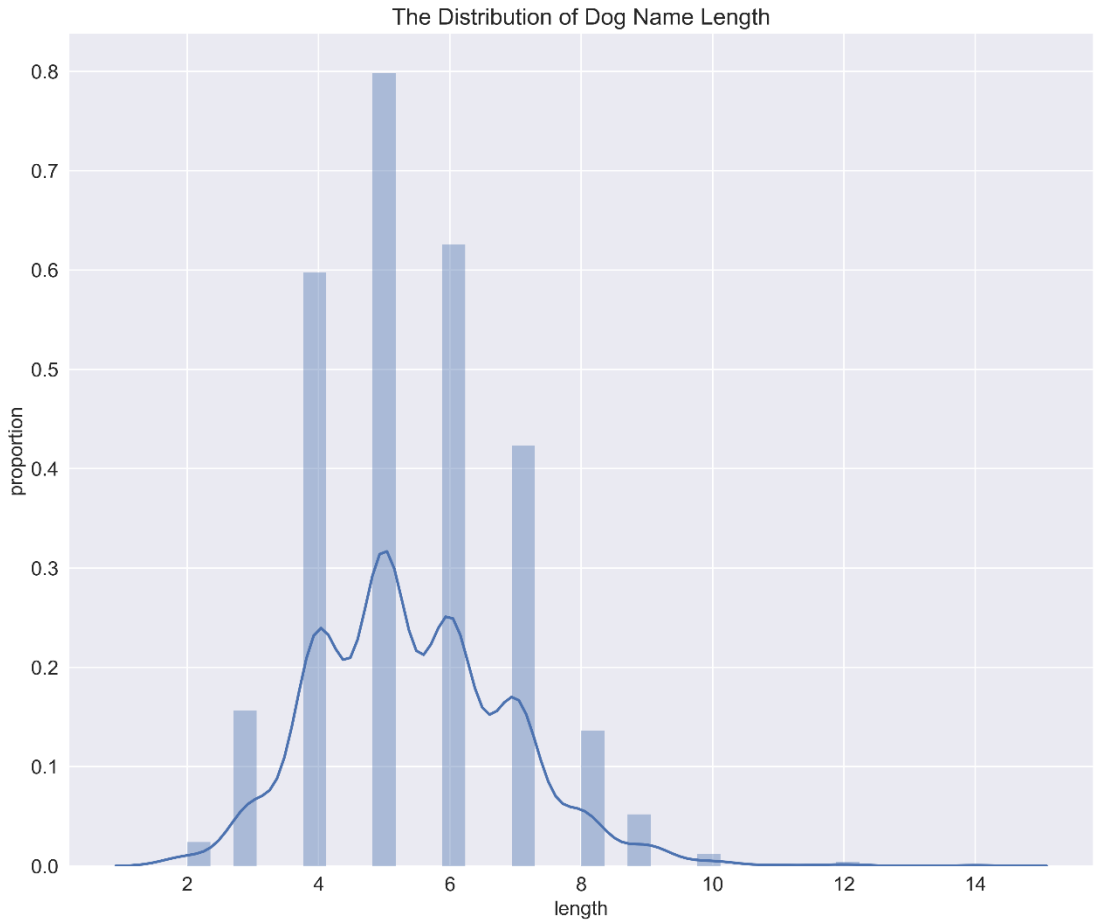


6. 狗狗名字几个字符的居多？

通过下面两图，可以发现

- 名字长度的分布图呈现出钟形
- 由大到小顺序排列，名字长度集中于 5，6，4，7 这三种长度中
- 对应占比分别是 0.281670，0.220807，0.210899，0.149328
- 共占比高达 86%，囊括了绝大多数名字，这可能由命名与发音习惯有关，名字太短，不宜发音，太长，则发音起来太麻烦，适中刚好





7. 评分最高的狗狗？

最高评分为 1776/10，这个比较特殊，发送 tweet 的当天是 7.4 ，是美国的独立日，具有特别意义。这里 1776 我猜推主是想表明美国的独立宣言发表年。
注：1776 年 7 月 4 日，美利坚合众国在费城发表独立宣言，正式宣布独立。

此条 Tweet 对应地址如下：
https://twitter.com/dog_rates/status/749981277374128128/photo/1

8. 不同种类的狗狗的平均评分各是多少？

在评分分值排名前 30 的狗狗种类中，评分数量>10 的评分中，只有 13 个品种入围
这 13 个品种的狗狗平均评分的偏差为 0.016543，差别十分细微，平均评分的均值为 1.136

平均评分排名前三由大到小排列

* 狗狗品种分别为 Samoyed，Golden_Retriever，Great_Pyrenees
* 平均评分分别为 1.169048，1.162975，1.146667