数据分析和可视化

我对数据集进行了评估，清洗与整理之后，我提出了感兴趣的问题，并针对问题进行行了数据分析和可视化。

我提出了了以下问题:

1.在一天中的哪个时段，该用户发推特比较多?

2.哪些种类的狗狗获得了较高的转发数?

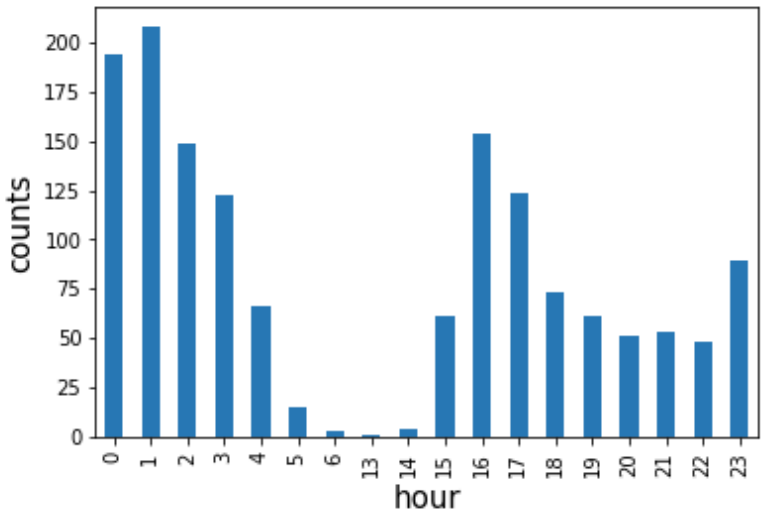
3.得到了高评分的狗狗得到的转发数就一定多吗?

4.数据中狗狗评分的高低和所处于的地位有什么关系?

针对第 1 个问题进行分析:

1.在一天中的哪个时段，该用户发推特比较多?

首先我们把时间栏提取出来，并单独提取数据中的 '小时'数据， 用value_counts() 查看该数据特殊值的分布，并用条形图可视化数据。

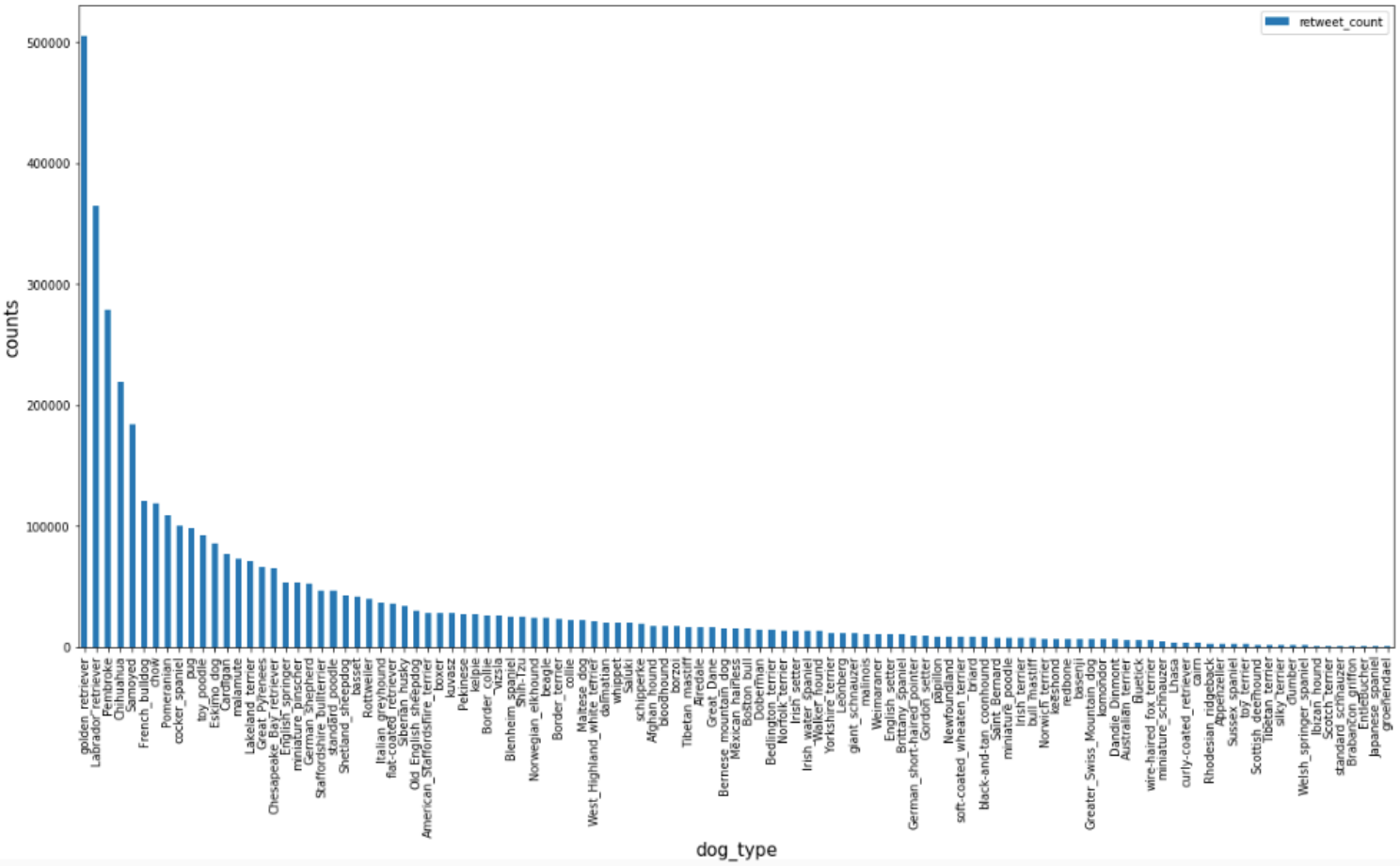


从图中我们可以看出， 凌晨时段和下午傍晚时段， 该用户发推特比较多。

针对第 2 个问题进行分析:

2.哪些种类的狗狗获得了较高的转发数?

提取预测结果狗的种类列与转发数这一列，根据种类分组，用条形图可视化，观察每种获得多少次转发。



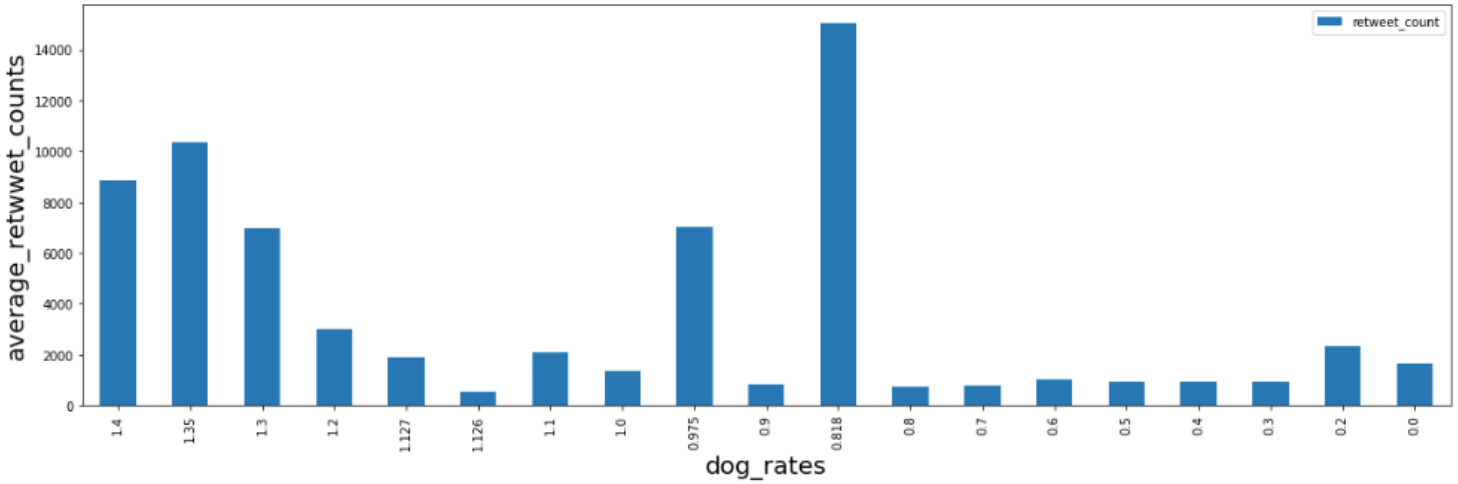
我们可以看到golden_retriever比比较受欢迎， 得到了了最多的转发数。

针对第 3 个问题进行分析:

3.得到了高评分的狗狗得到的转发数就一定多吗?

清理质量问题后，根据数据中的分子和分母，相除计算评分的最终结果， 提取该列和转发数的列。

以评分数的各个特殊值，对两列组成的dataFrame进行分组， 求出每个评分所对应转发数的均值，并以柱状图可视化。



并且，我们计算这两类的相关值为0.33

综合得出，狗狗的评分高低和得到多的转发数多少关系不大

针对第 4 个问题进行分析:

4.数据中狗狗评分的高低和所处于的地位有什么关系?

我们提取出地位栏和评分栏， 观察每个评分下所对应的狗的地位分布情况。

```
: temporary_3.groupby('rates_final')['status'].value_counts()

: rates_final  status
  0.700        pupper             2
  0.800        pupper             5
               doggo              2
               doggo & pupper     1
  0.900        pupper            11
               doggo & pupper     1
               puppo              1
  1.000        pupper            34
               doggo              3
               puppo              3
               doggo & pupper     2
               floofer            1
  1.100        pupper            38
               doggo             12
               doggo & pupper     1
               floofer            1
               puppo              1
  1.127        pupper             1
  1.200        pupper            39
               doggo             15
               puppo              5
               floofer            2
               doggo & pupper     1
  1.300        doggo             16
               pupper            12
               puppo              9
               floofer            3
  1.400        doggo              5
               pupper             4
               puppo              1
  Name: status, dtype: int64
```

观察数据发现，数据中狗狗评分的高低和所处于的地位关系不大，评分中的地位大多被pupper和doggo占据。