#MCM

Wealth of Data / 数据之财

评判过程

- 首先一定是看摘要和作文，如果在这里面没有完整地介绍每个模型的建立过程和结果，判 S 奖；如果作文中的内容使得裁判无法理解，或者无法抓住重点，判 S 奖
- 只要漏掉任何一条要求，S 奖

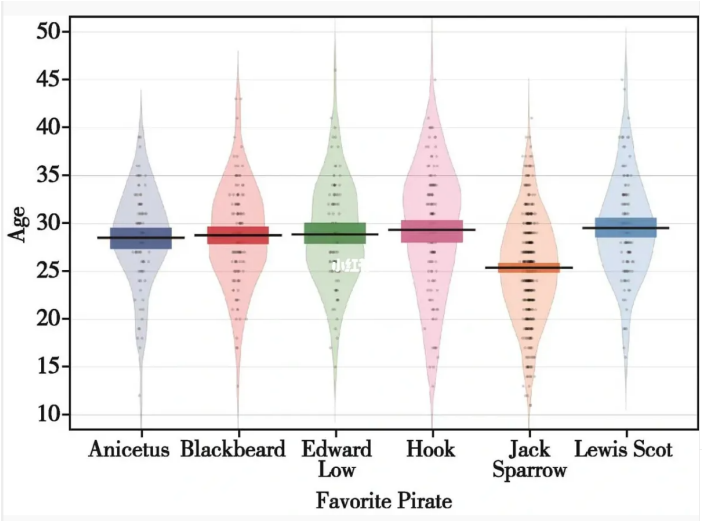
题目解说

如何基于文本进行数据分析

这一部分主要考察学生如何从文本中抽取特征，更具体地，学生需要从文本中抽取一系列随机变量，用作后续分析。R 语言中有许多库可以做情感分析、LDA 等等。裁判并不希望队伍在这一部分中使用多么高级的方法，而是希望所有人可以从文字中提取一个有效的评分，用作后续建模。
成功的队伍通过一些方法识别了刷评论机器人和对手公司的水军等等，例如，通过一个评论是否被 Verified
成功的队伍发现了文本数据和其他特征之间的联系，例如，负面评论通常更长一些；他们还观察了星数和评论数随着时间的变化，来发掘不同的模式；一些队伍注意到一星和五星的评论对于产品的改进十分重要；一些成功的队伍将评论的情感和星数进行了对比，还有一些队伍觉得评论的标题与评论本身相比更能突出产品的特征，这也是可以的
一些队伍寻找了提取特征单词最有效的机器学习算法；一些队伍将词语分成三类（Positive, neutral, negative），并计数了每一类词语的数量；一些队伍在分析中采用了时间序列方法
这一部分没有固定的答案，只要模型可以被解释，并且在下一部分的分析中有效，就是好的模型

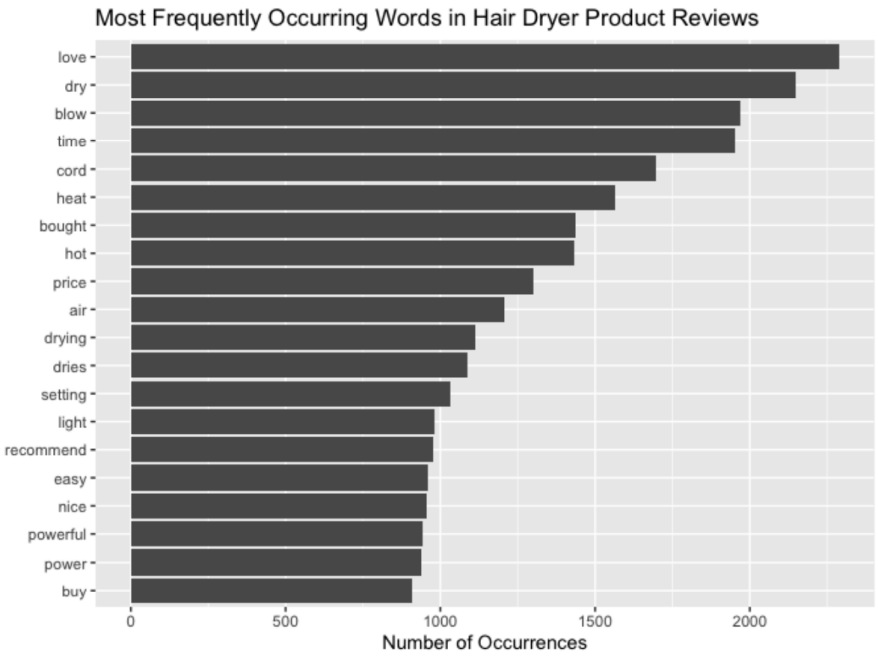
数据的可视化

对于大部分的数据，可以用折线图、箱线图等展示关系。**这里特别提出箱线图，可以很好地用于对比不同分布，例如在特定词语出现时星数的对比**，另外，推荐小提琴图和海盗图



此外，一些很火的可视化方式不一定是最合适的！例如词云图、Tree map 等等，人类的眼睛比较面积大小等等是困难的，最好比较的是长度（注意：扇形统计图因此也有劣势）（词云图中，频率正比于字号，但是，单词的长度也影响视觉效果）

下面的图就是合理的



Handling Uncertainty

任何的统计结果中都要包含对于不确定性的讨论，例如，一些队伍在时间序列预测中给出了置信区间。**（任何没有考虑不确定性的预测都是失效的）** 这种不确定性不仅可以在预测时加以考虑，在其他时候也需要。例如，清华大学的一支队伍在进行二元预测时，使用 AUC 衡量不确定度

部分队伍有问题：缺失重要部分

- 缺失对于问题的描述，比如，问题重述，本队使用的方法，本队的结果简介等等
- 假设和说明
- 模型建立和应用（需要 Well discribe）

- 模型验证
- 敏感性分析
- 优缺点
- **最后的结论**
- 信
  少了任何一部分，基本上就判 S 奖

写信

信的组织结构十分重要，很好的可视化和流程图为信增色不少，信中，你需要

- 提供关于产品的清晰、有效、可行的结论
- 给 Market Director 提供可行的措施
- 这些建议中必须包含**从数据中抽取的正面、负面的词语**，并使用这些词语提出建议

Outstanding 论文的优势

对于过程的完整描述和对于结果的分析使得论文突出。此外，**行至论文结尾，不要忘记自己的初心是做什么**，本题的重点是提供销售建议，因此在小文章中这点应该有所体现。例如，有一支队伍建议，吹风机应该注重效率；此外，销售策略可能包括高峰期的和低谷期的，例如高峰期多打广告，等等
好的论文使用了优秀的可视化、流程图等，清晰地解释了结果
有的论文衡量了预测结果的不确定性