1. **需尽快确认清洗数据结果**

看看top frequent words的结果有没有奇怪的， 比如之前topic里面就还会有数字啊啥的，比如，一般的单词比较有意义的可能都是3-4个character起，所以3个一下character的单词可能就可以筛掉了等，因为清洗结果会影响后面的topic modelling结果

1. **然后确认数据分析结果=>需要发给老师看，因为老师回复会有延时**
2. **完成论文主体draft=>需要发给老师看，因为老师回复会有延时**
3. **完成论文**

注意：通过研究Airbnb的用户评论等数据，研究人们的态度变化从而来映射疫情对共享经济的影响，因为是对共享经济的影响，所以不能单纯只看rating score，要价格和空置也要结合一起看。

-------------------------------------------------------------------------------------

我能想到的数据分析细节如下：

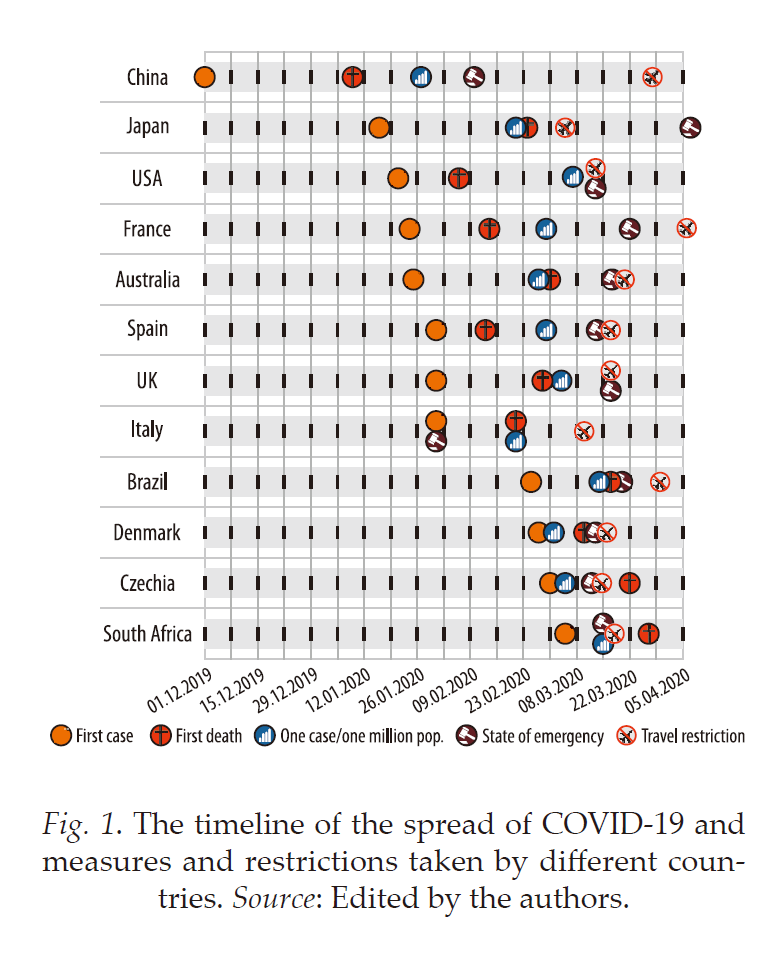
# 数据清洗&预处理

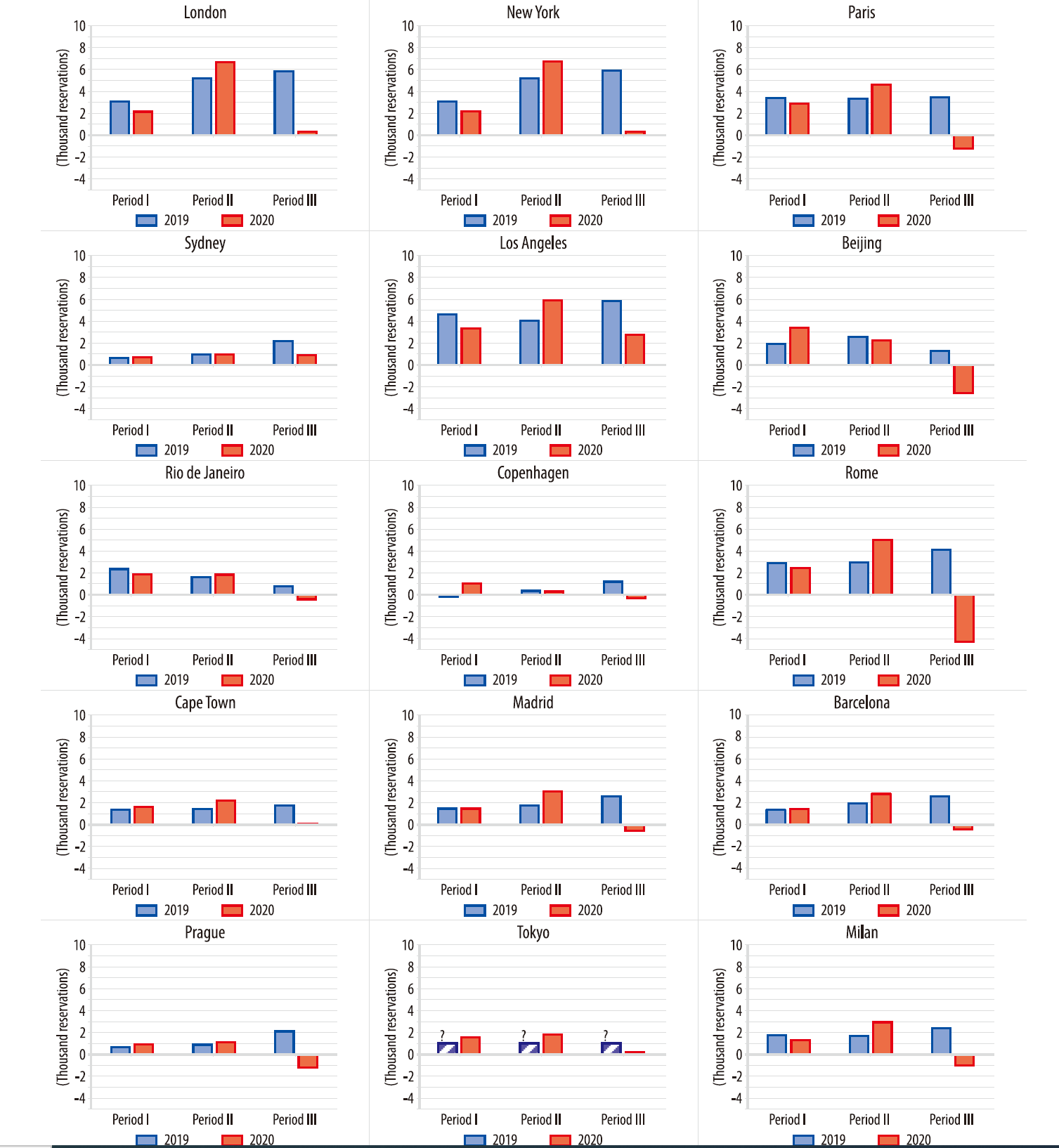
* 过滤出需要的日期区间：
* 去掉空值NA数据
* 去掉字符符号icon等
* 过滤英文
* 根据数据，过滤对应listing\_id下review数量过多和过少（outliers）的listing\_id
* 根据数据（查看characters length之类的），过滤reviews长短，例如某个review 写的特别长，和一些特别短的reviews
* 选出需要的variables
* 词性标注->过滤只留下noun和adj
* 去停用词，除了常用的stopwords list意外，根据数据添加自己的custom stopwords，例如：拼写错误的airbnb，aribnb，Airbnb，AIRBNB，城市名，bedroom bathroom kitchen那些（不知道需不需要）"km", "minutes", "rm", "approx","access", "home","house", “apartment”， “condo”， “flat”， “ time",neighbourhood,host\_name, room\_name, property\_name，property\_type 等
* 去拼写错误
* TFIDF 去掉outliers
* Tokenization
* Stemming & lemmatization
* Dependency parsing
* Consitituency parsing

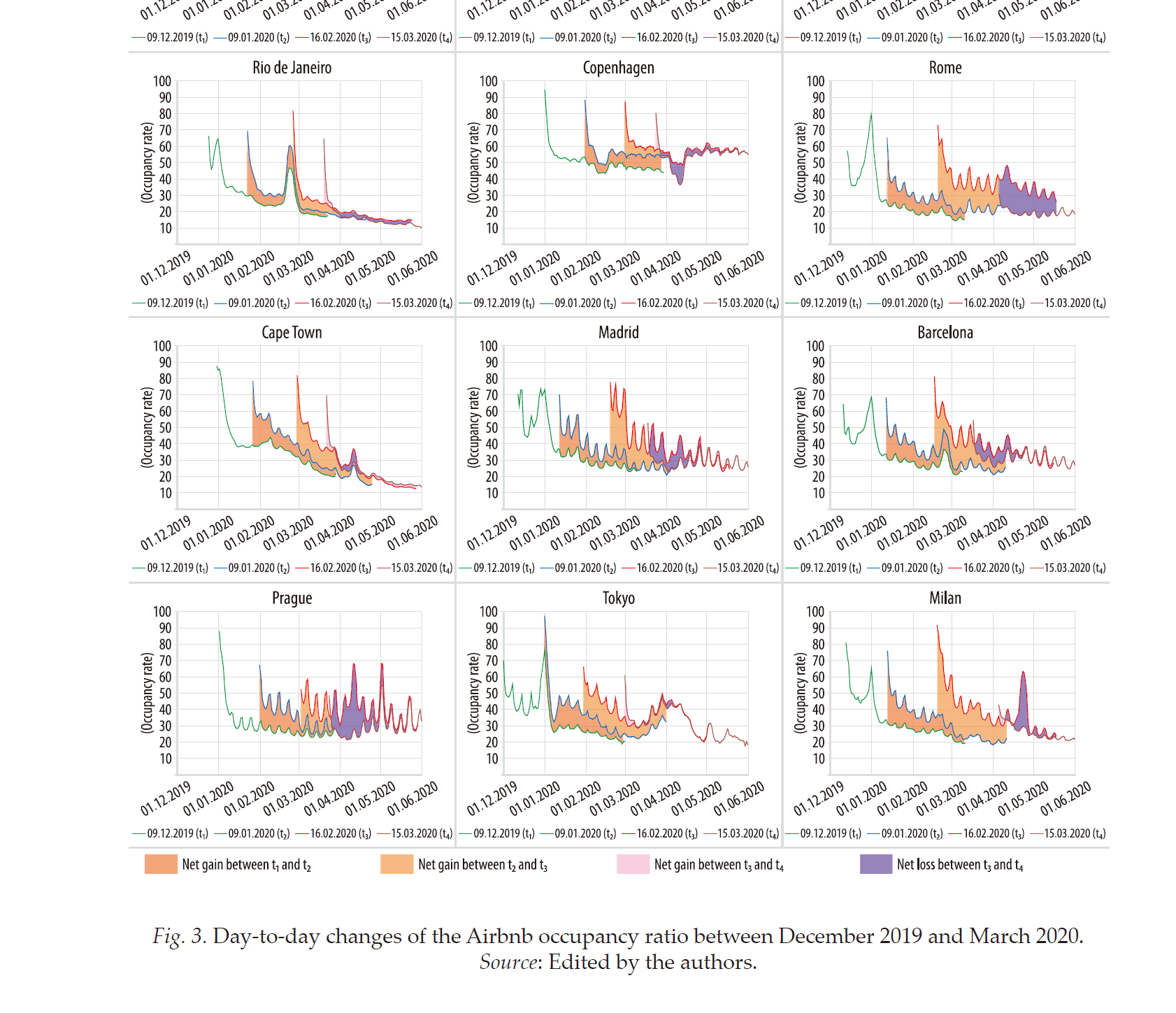
# Overall Analysis

* 各个国家：疫情前后分别：房源数量，评论数量，平均每个房源的评论数量，平均价格，平均rating score，平均空置率等
* 总体的（所有国家加起来）：疫情前后分别：房源数量，评论数量，平均每个房源的评论数量，平均价格，平均rating score，平均空置率（需根据calendar那张表来计算）等
* 需要图来看到一些分布情况之类的。
* 各个国家first case、first death、state of emergency、travel restriction时间线

要保证这些变量和指标要存在于数据中，除此之外，为了与论文主题的一致性，可能会增加一些变量的分析，也可能会删掉一些变量







# Sentiment Analysis

用3个模型（例如naïve bayes，bert或XLnet或xgboost或svm之类的）做sentiment analysis，然后在三个模型得出sentiment scores的基础上增加回归分析：

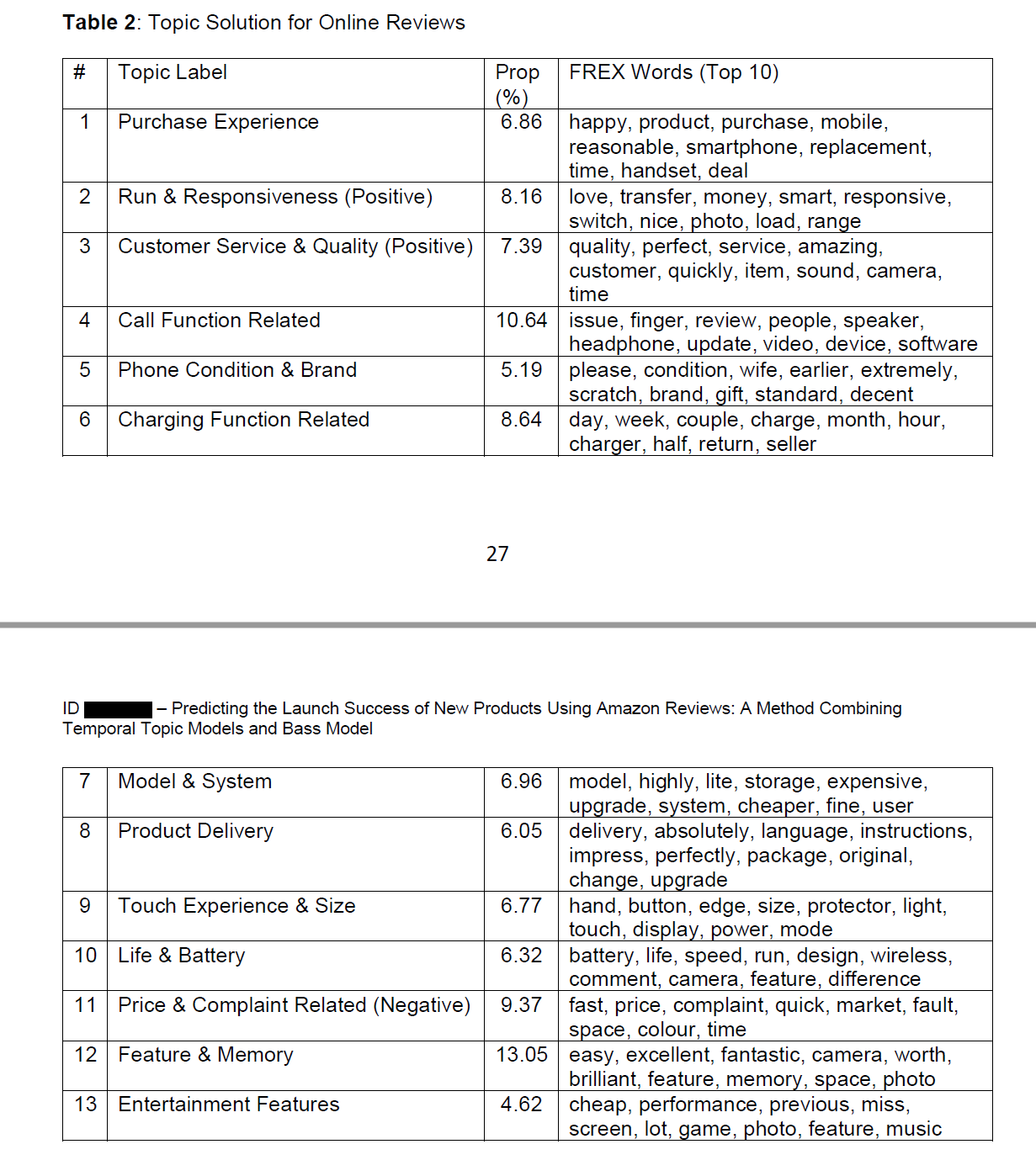
1. 疫情新增cases数据
2. 疫情死亡数量
3. 房源价格
4. 房源rating score
5. 房源空置率

# LDA（下面的点无前后顺序）

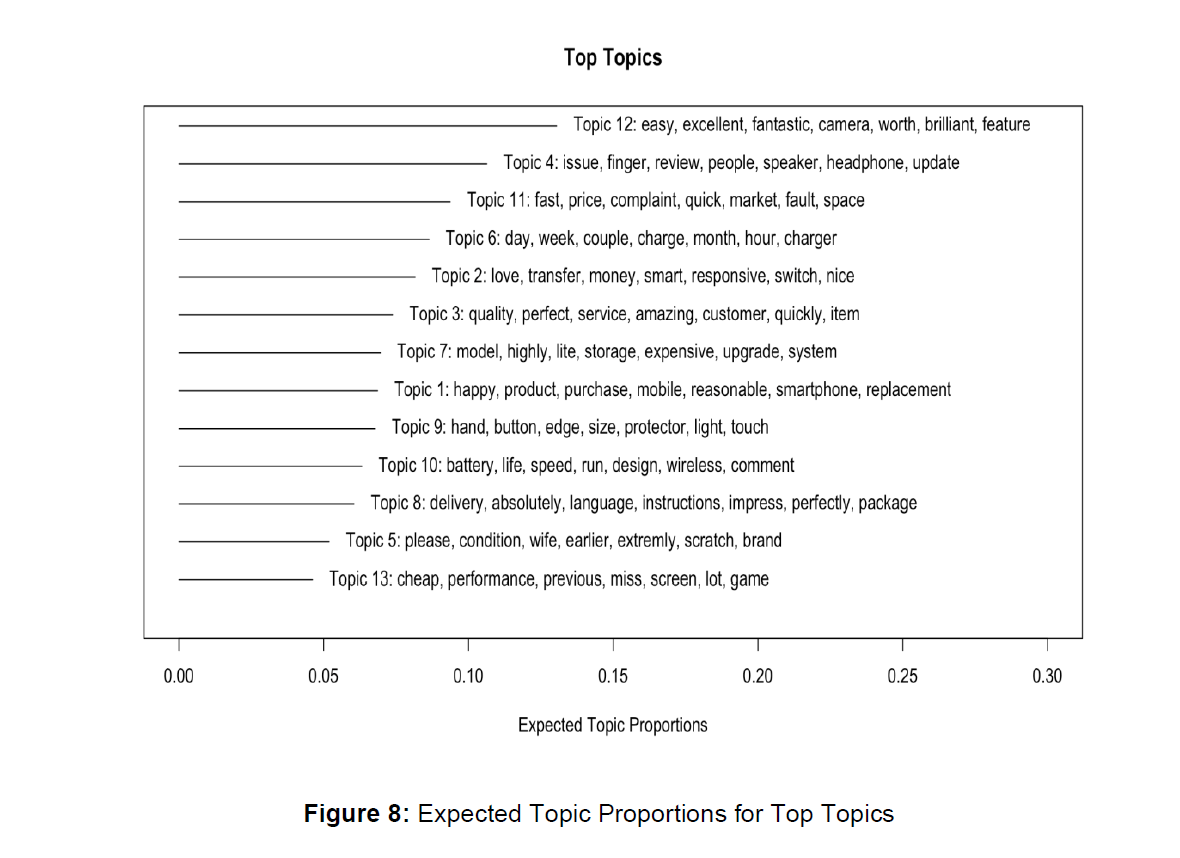
* Define optimal number of topics => search K 之类的



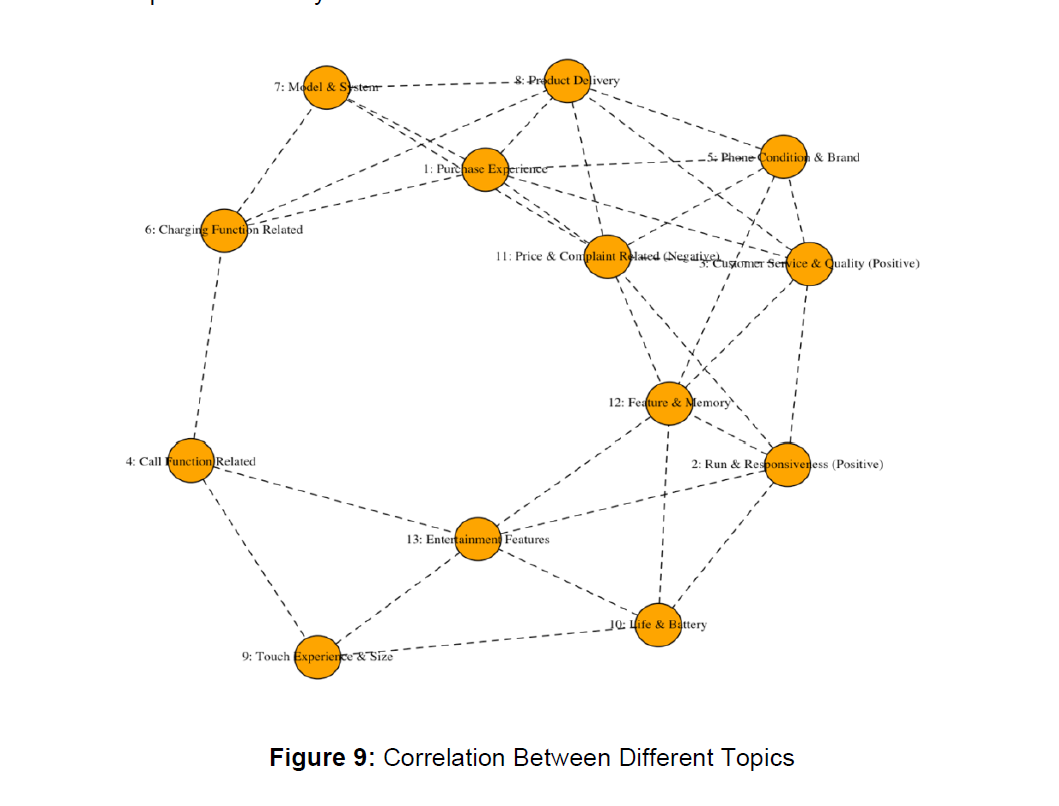
* Topic label & FREX



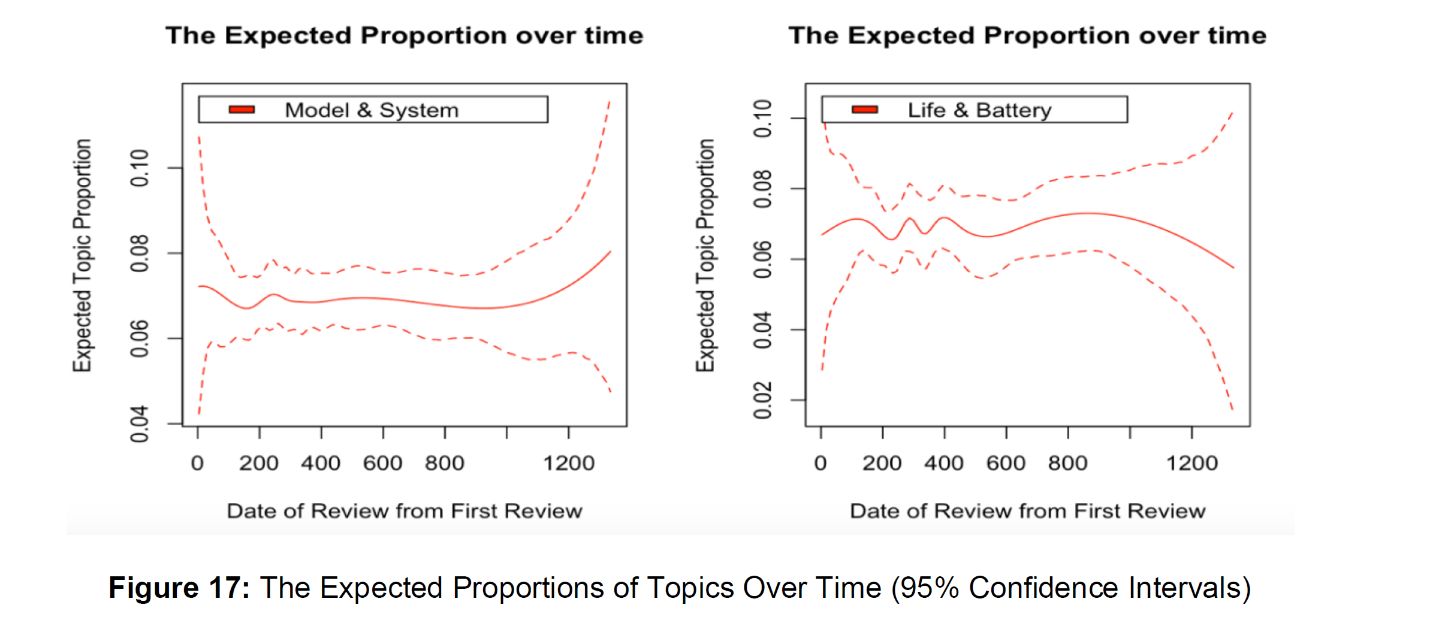
* Plot Expected Topic Proportions

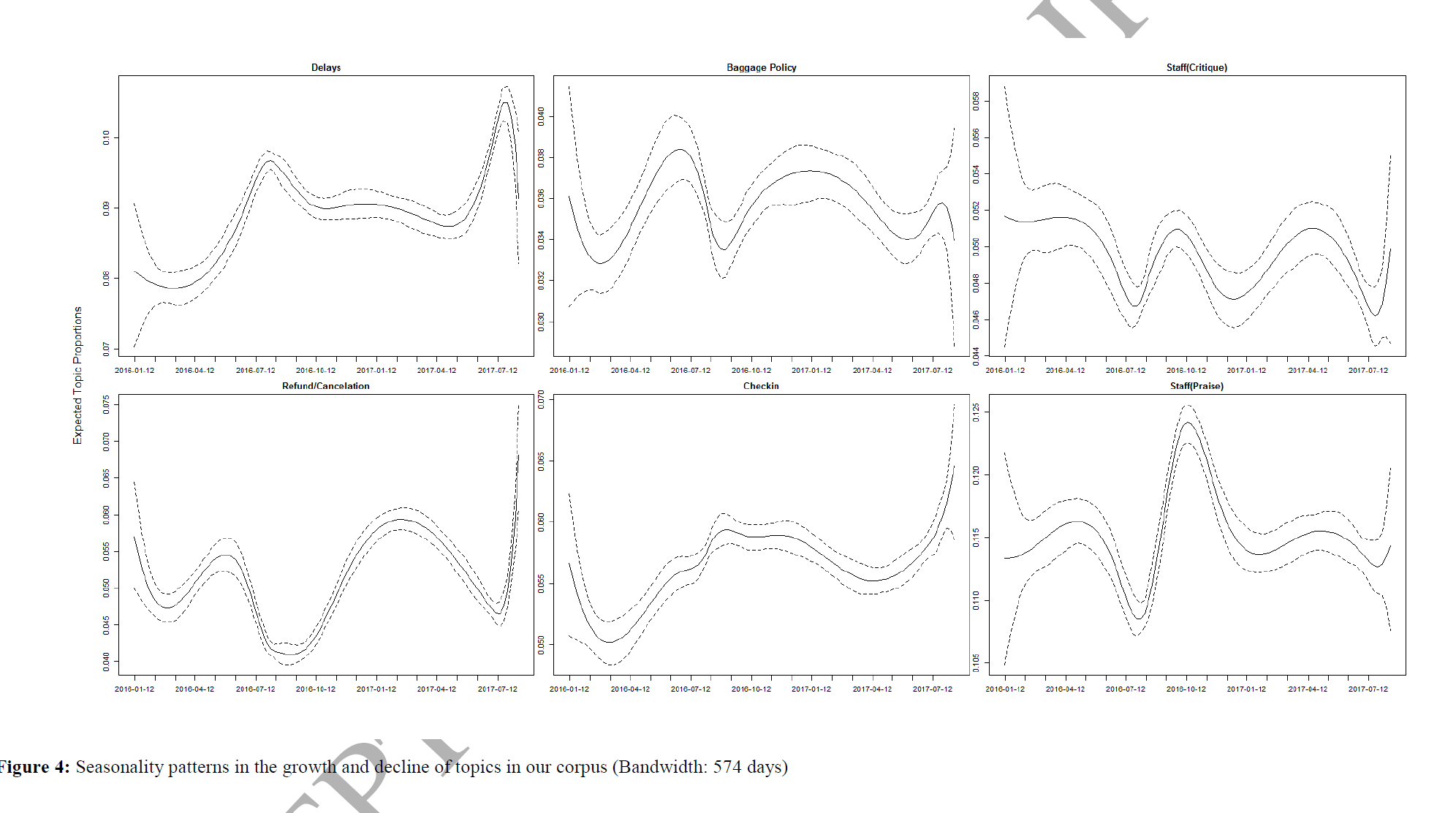


* Topic Correlations



* 根据疫情新增+死亡数，来看topics的变化
* Estimation on Topic change





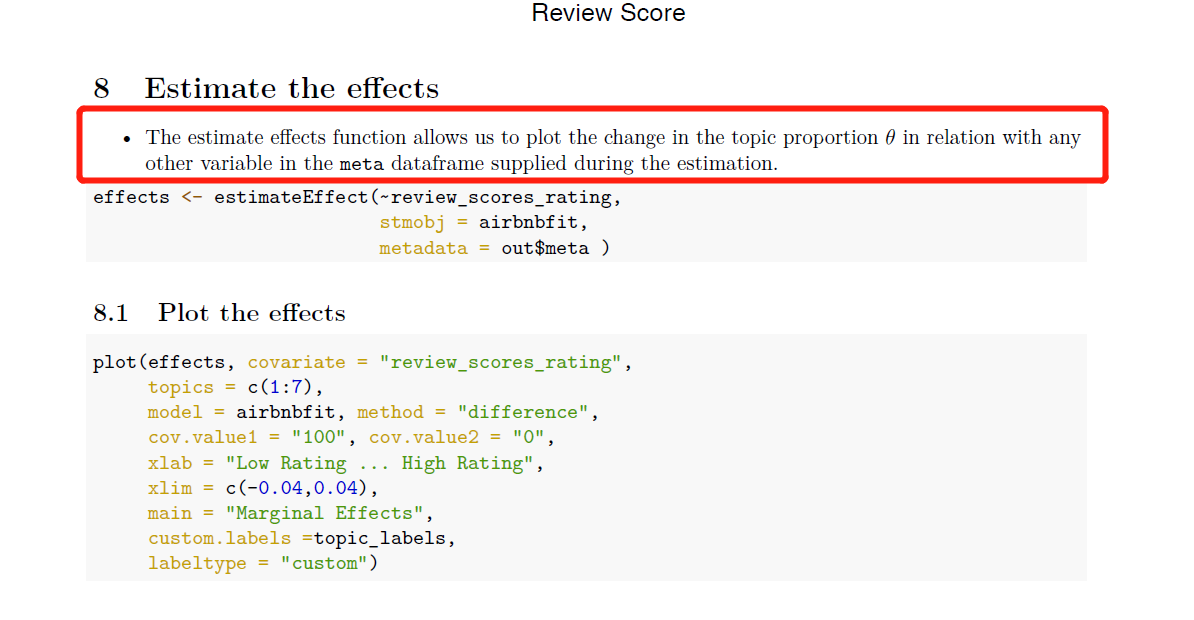
* 把topics和房源价格、空置率、rating score一起建模（回归或者类似）来分析，具体这块需要看我发给你的文章，我有点忘记怎么做了=> 应该是用topic probability
* Topic model Evaluation

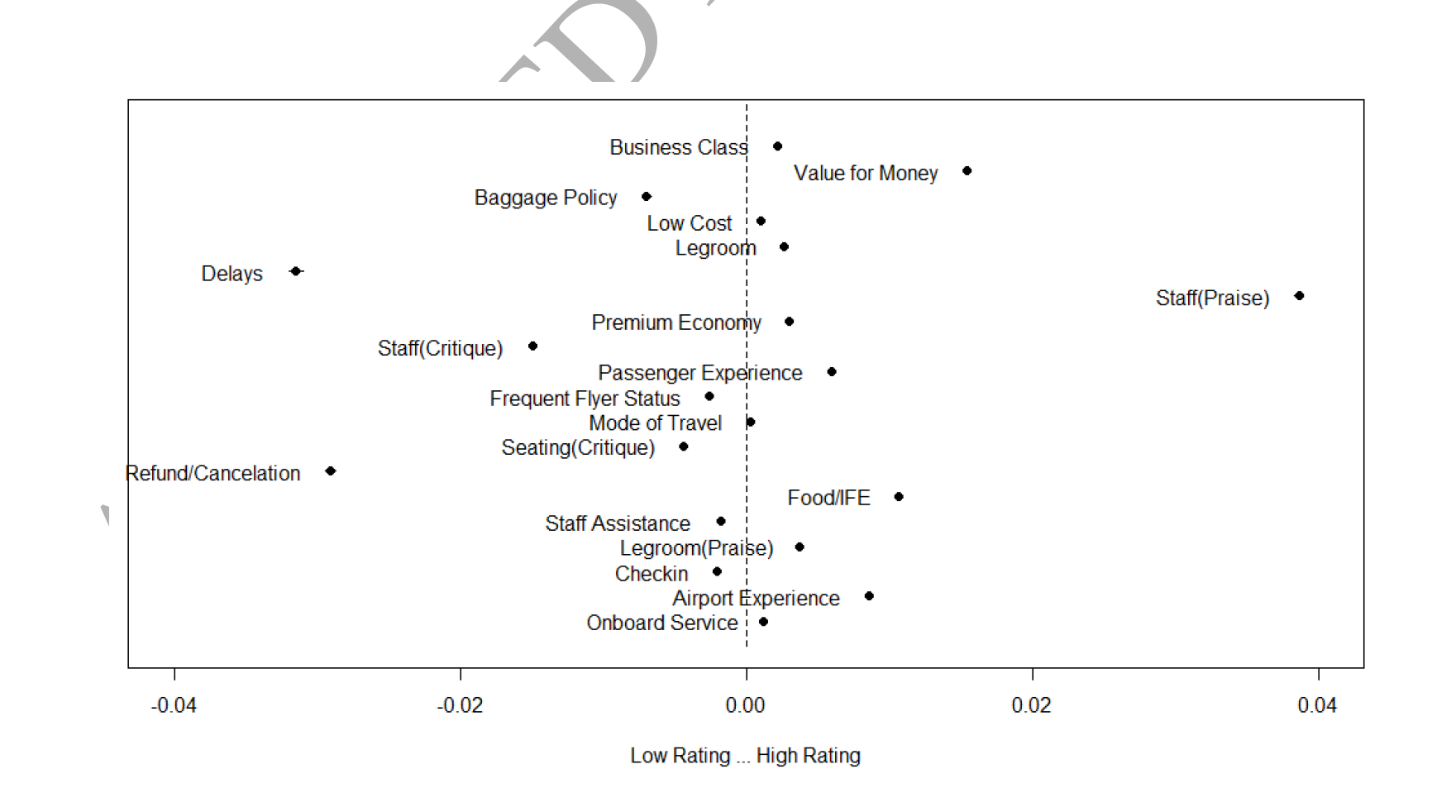
参考链接<https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0>

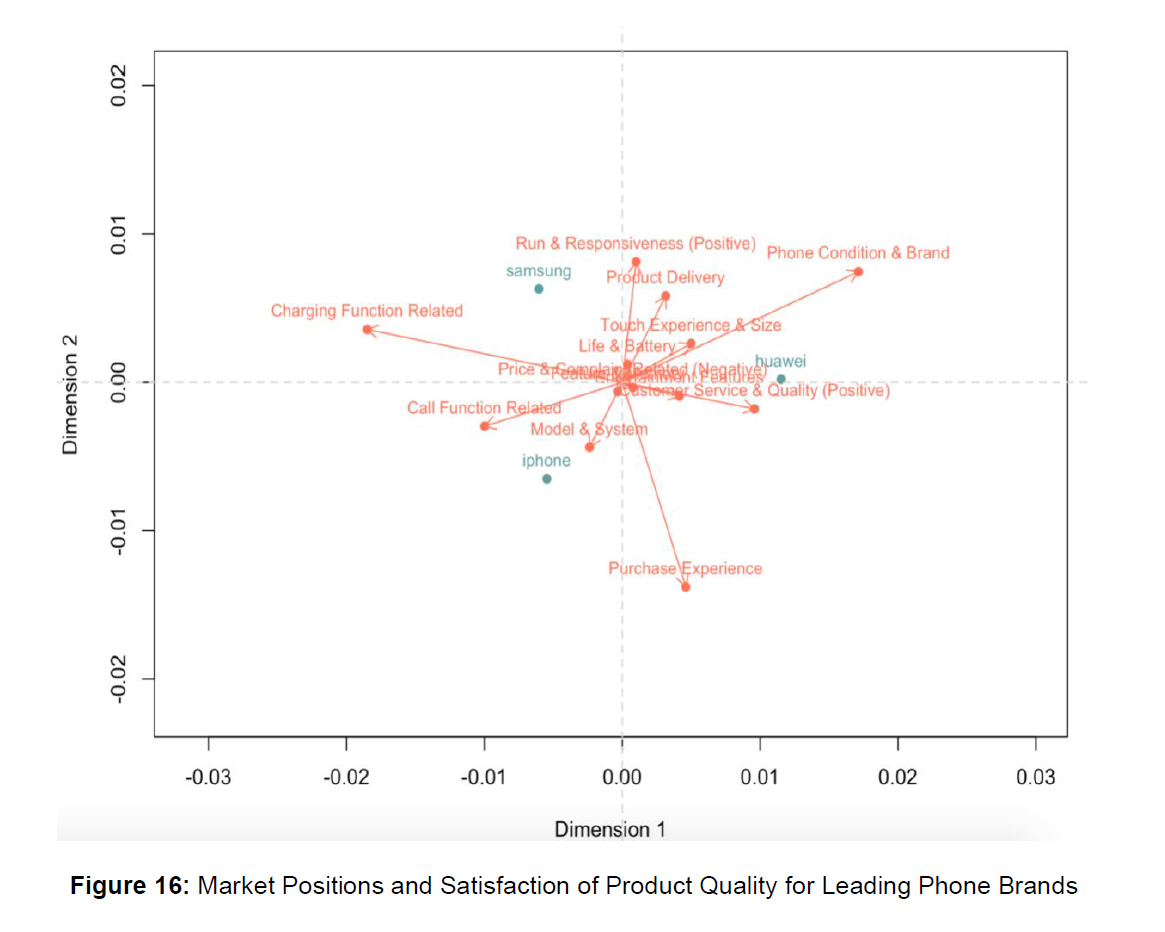
也可参考topic model evaluation相关文献

* Semantic Coherence and Exclusivity
* Estimated Effects

这里应该是可以用reviews rating score，也可以用price/空置率



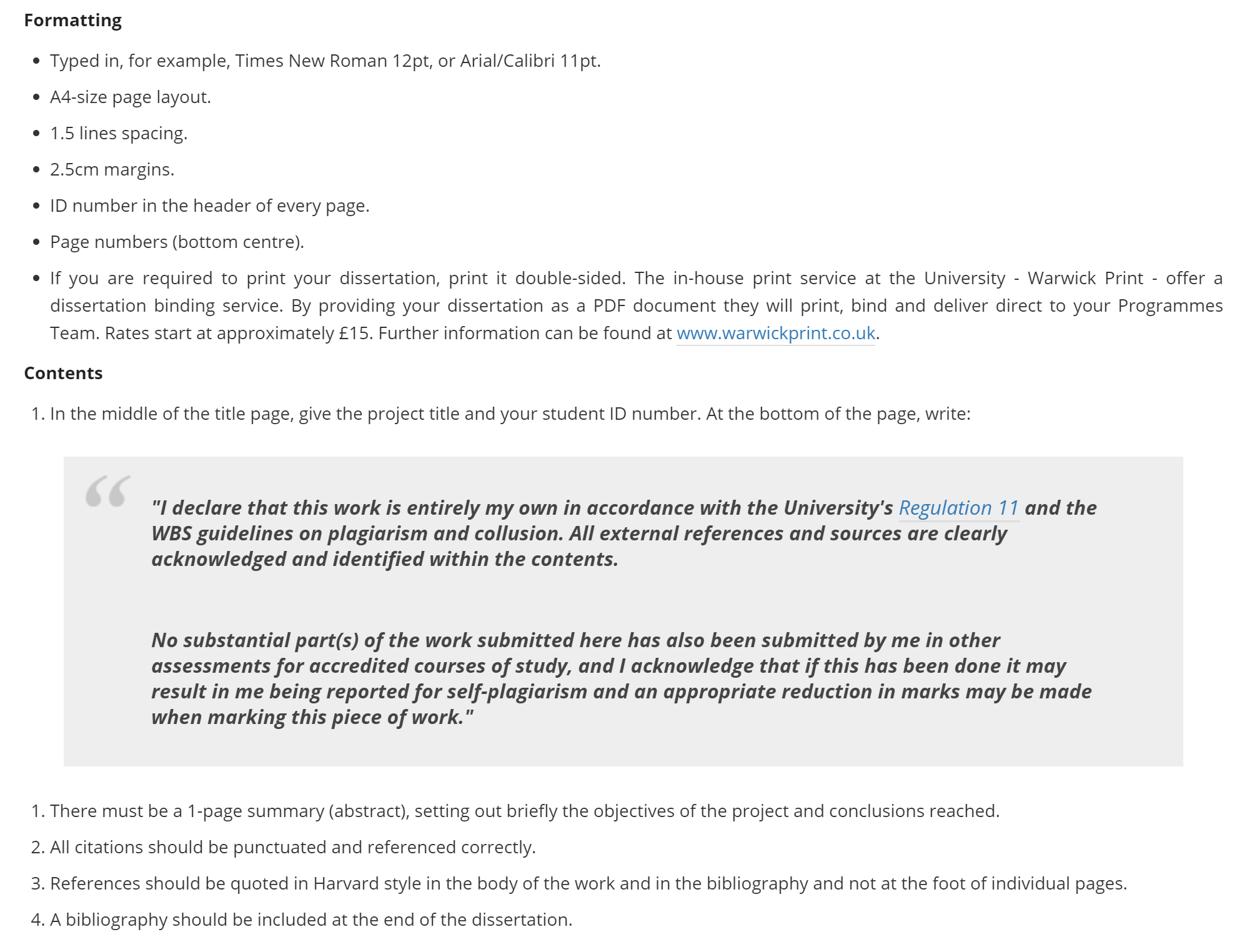




上图根据factoextra看topic relation/dimension

------------------------------------------------------------------------------------------------------------------------------------------

# 论文格式要求：



字数：10000字

代码作为appendix附在最后，代码注释要求每段写清楚