# Step1：整合数据

【代码文件】data.R

【需要的package】dplyr：处理数据；

【说明】

getwd()语句会输出一个目录，表示工作区的位置，请将所有csv文件存储在该目录中，再运行所有代码。

运行完毕会将整合后的数据输出至工作区，文件名为[all\_data]product&device，还有一些附属文件是以[data]，[subsetdata]开头的csv文件，这些数据仅作参考，不会被用到后续的绘图过程。

# Step2：特征工程

【代码文件】feature.R

【需要的package】

corrplot：做相关系数图；polycor：计算相关系数；ggplot2：画图；ggthemes：修改图；lubridate：处理时间和日期数据；jiebaR：分词；jiebaRD：分词；wordcloud2：绘制词云

【说明】

由于数据有一个重要特性：可用于直接绘图的数值型数据太少，因此这些数据绘图的第一步是根据数据的特点，提取出各种数值和关联关系，再基于这些内容绘图。特征工程的主要思路有：

**因子变量**

首先在整合表中，对Factor类型较少的Factor变量，统计其每个Factor对应的频数，Factor和频数对应在一个dataframe中，dataframe一般以count\_开头，Var1列是Factor,Freq列是对应的频数。

**二分变量**

对某些int类型的变量，他们的意义一般是该device是否满足某特征，如该device是否有third party flag，对这种变量我们统计其是或否的频数，并将所有该类型的变量的统计结果组织在一个dataframe中：data\_summary\_int，由于频数差距太大（mix:98，max:350000+）,因此为便于画图我们加入对频数就行对数运算，结果存储在data\_log\_summary\_int中。

**日期的特殊处理**

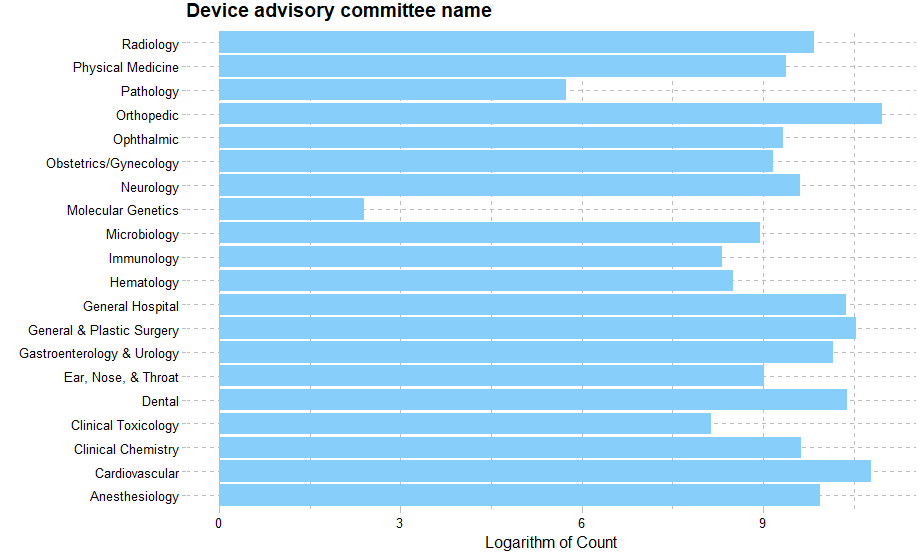
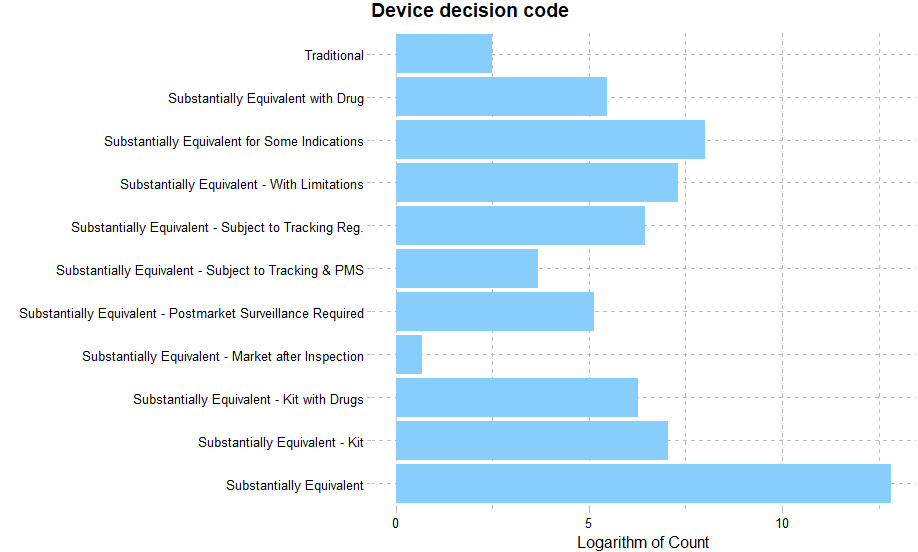
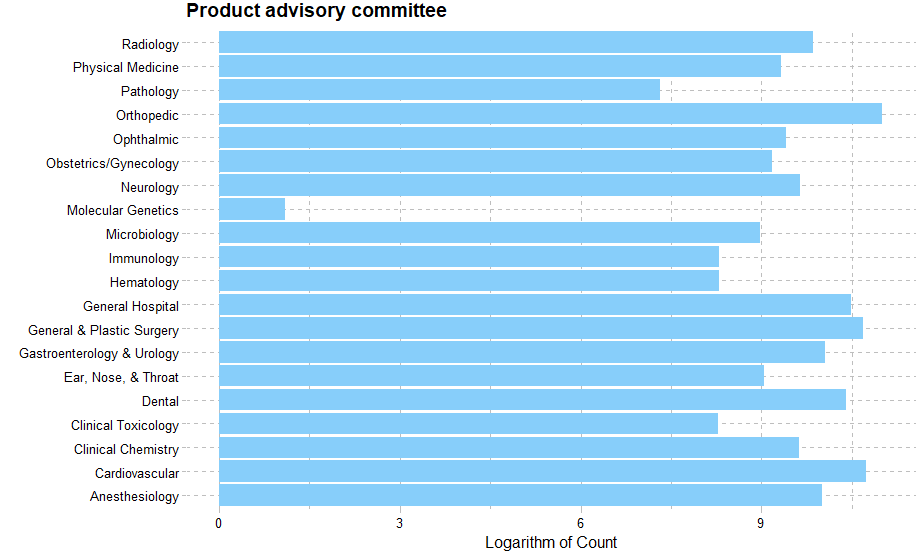
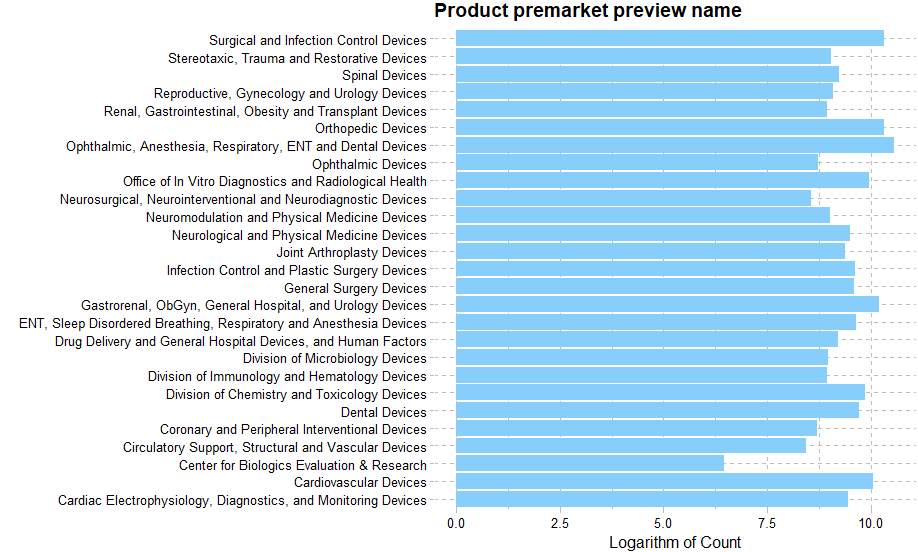
Alldata中共有两列时间数据，receive date和decision date，将两列时间数据相减，差值即为该device的spent time，进一步做两方面的统计，第一是根据月份统计频数，方便后续做box plot，第二是根据对每个Date统计，后续做密度变化统计图。

# Step3：画图

## Plot1：barplot

上述因子变量的结果一般做barplot，结合数据特征，先对所有数据进行对数运算（防止极值影响太大），再做barplot，这些barplot的意义较简单，分析起来也比较容易，但要结合实际意义（这就是你所擅长了）。

下面的图作图方式都类似，纵轴为类型，横轴为频数统计的对数值。这里需要注意的是，由于横轴取了对数，因此横轴每相差1，实际数值即有约2.7倍的差距：

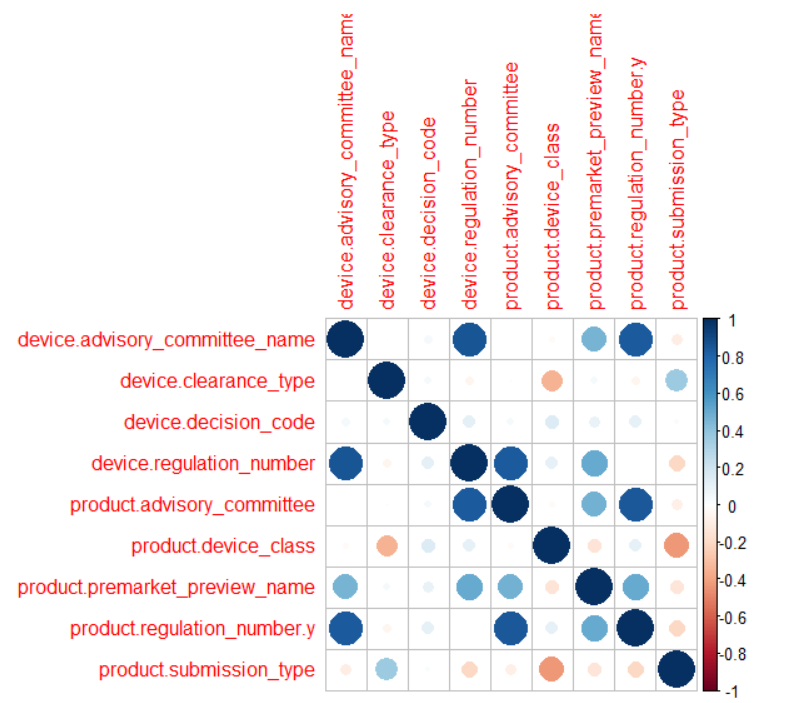


## Plot2：corplot

Corplot可以清楚的揭示变量间的相关性，其具体做法是先选出需要参加相关性计算的变量，这里我采取的选择标准是选出所有factor变量，将它们整合在表cor\_data中，然后使用hetcor函数计算各个变量间的相关性，再利用corrplot作图如下：

简而言之两个变量间的颜色越深表示两个变量相关关系越强，要进一步了解相关性图的知识请浏览：

<https://blog.csdn.net/weixin_33751566/article/details/92288962?utm_medium=distribute.pc_relevant.none-task-blog-BlogCommendFromMachineLearnPai2-1.nonecase&depth_1-utm_source=distribute.pc_relevant.none-task-blog-BlogCommendFromMachineLearnPai2-1.nonecase>

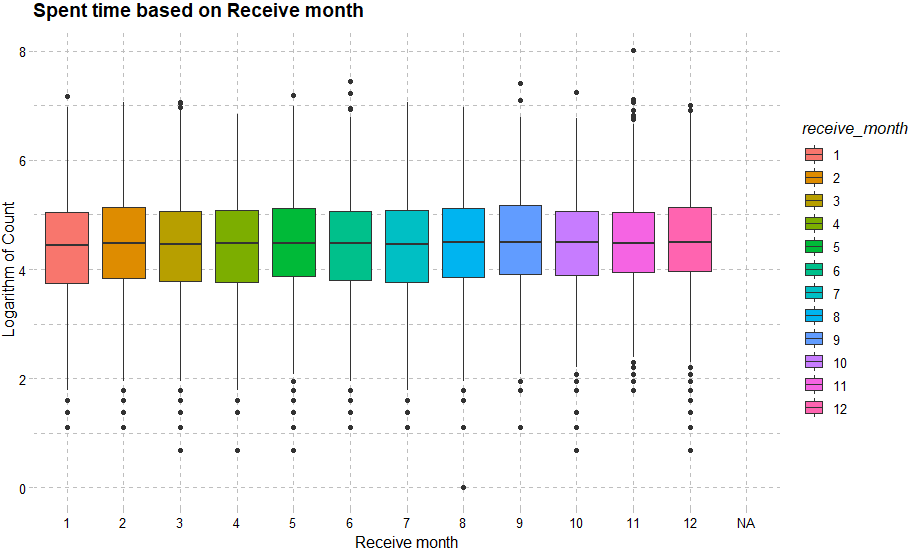
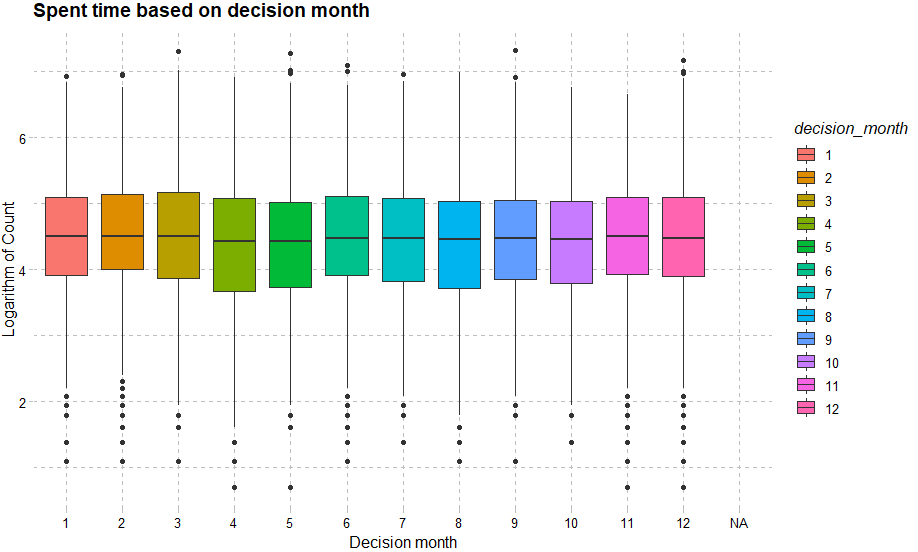


## Plot3：boxplot

上文提到的每个device的spent time在本例中被用于画图，该图展示了以月份作为分类单位，该月份的device的receive和decision情况（我个人理解receive和decision分别是接受申请和获取批准），boxplot的特点是，每个长方形中间的横线，表示该月的平均申请和批准数，长方形的下沿和上沿分别表示前25%和后25%的水平，上下的散点表示改组的极值。

【注】这里每个月份的频数也取了对数，否则受极值影响太大

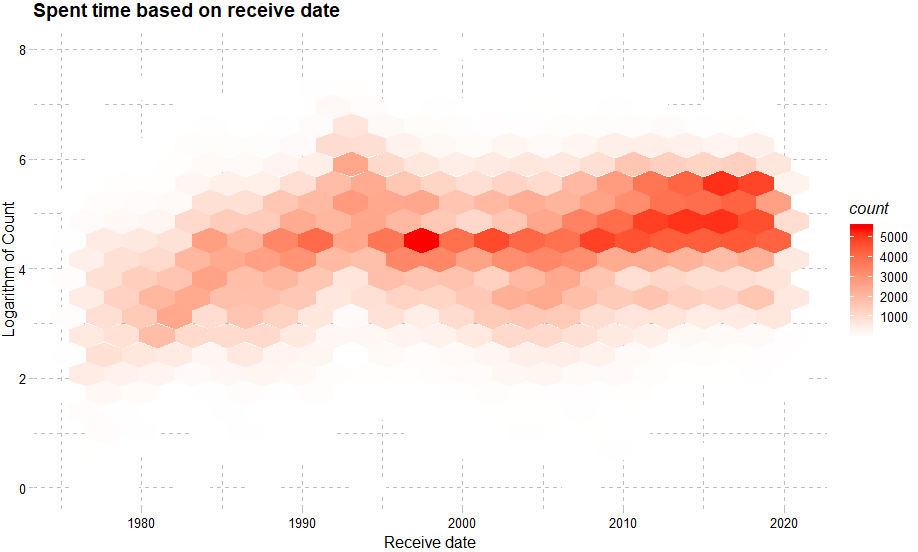
【boxplot相关知识】<https://zhuanlan.zhihu.com/p/110580568?from_voters_page=true>



## Plot4：hexplot

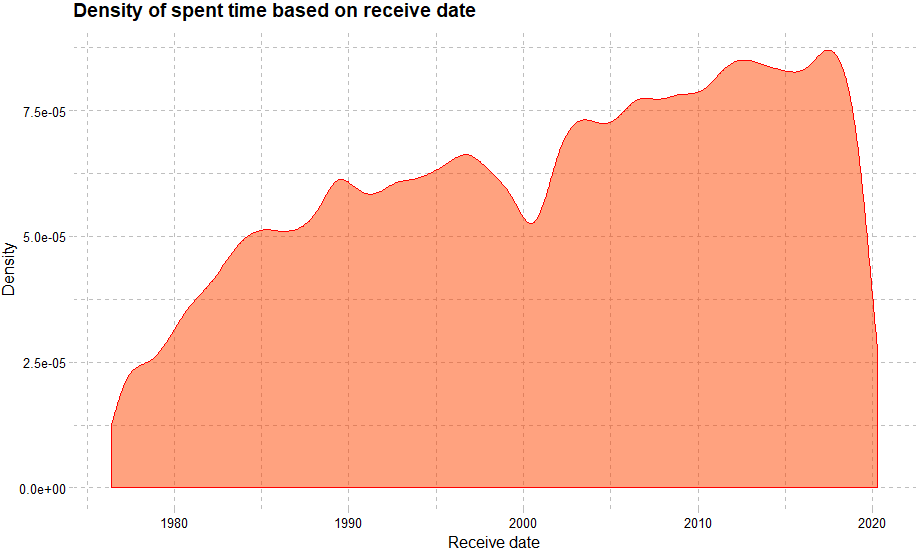
该图横轴表示每个device receive的date，纵轴表示该date的频数（取对数），六边形是该部分的热力学统计，即颜色越红，该部分的点越多。如果这张图用散点图，则整个图层会被充满，hexplot可以有效避免这一点，并清晰展示数据的特征。

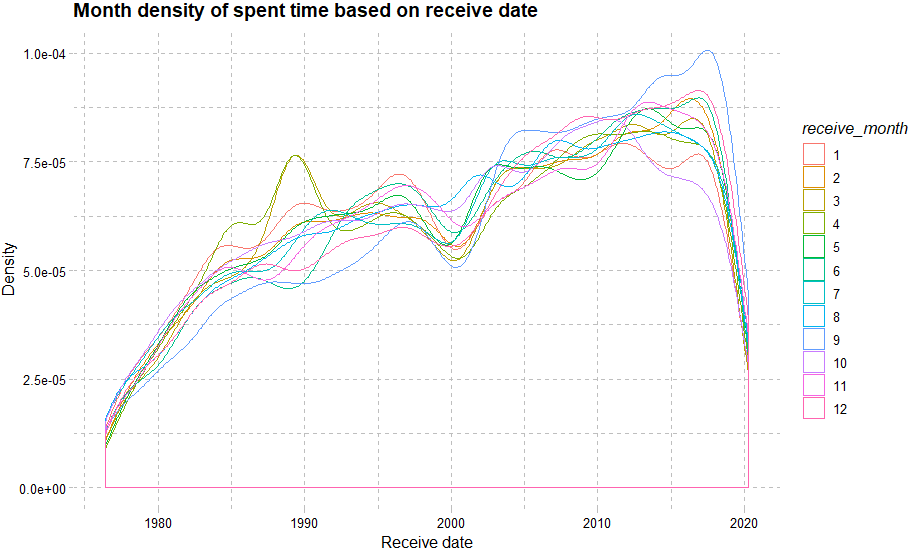
简而言之该图有两个特征：六边形越考上，该段时间内每天平均receive的数目越多，六边形颜色越红，这段时间内的receive越密集



## Plot5: densityplot

Densityplot是针对date对reveive的密度统计，这里有两幅图，一个是总计，一个是按月统计，意义较为明确，可以着手分析。

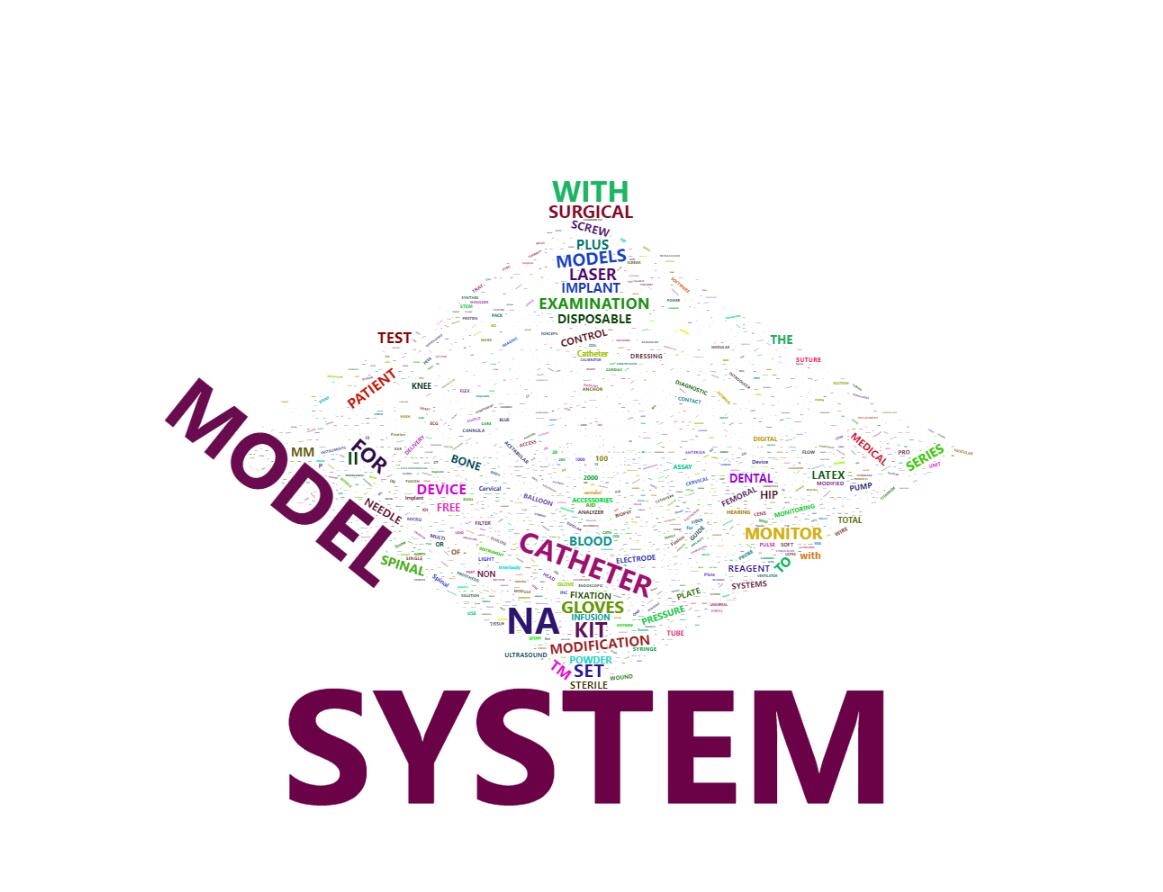




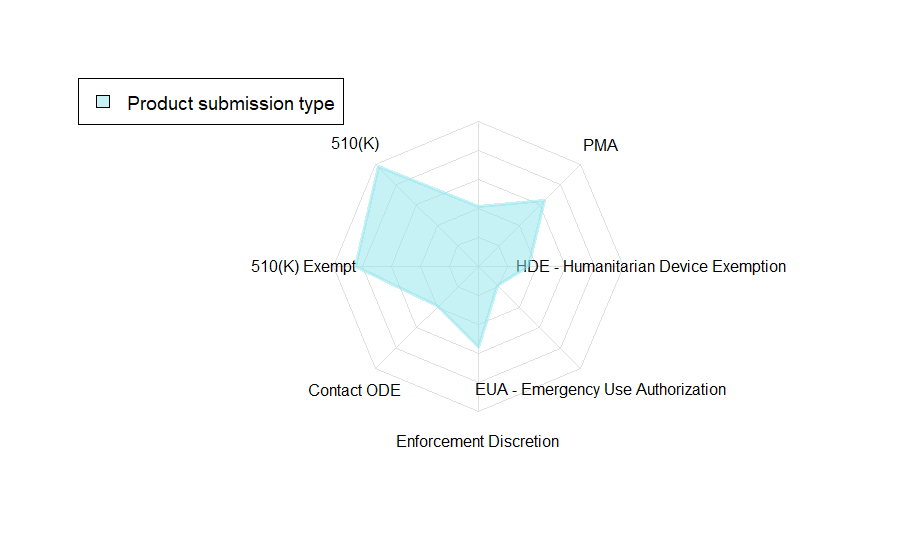
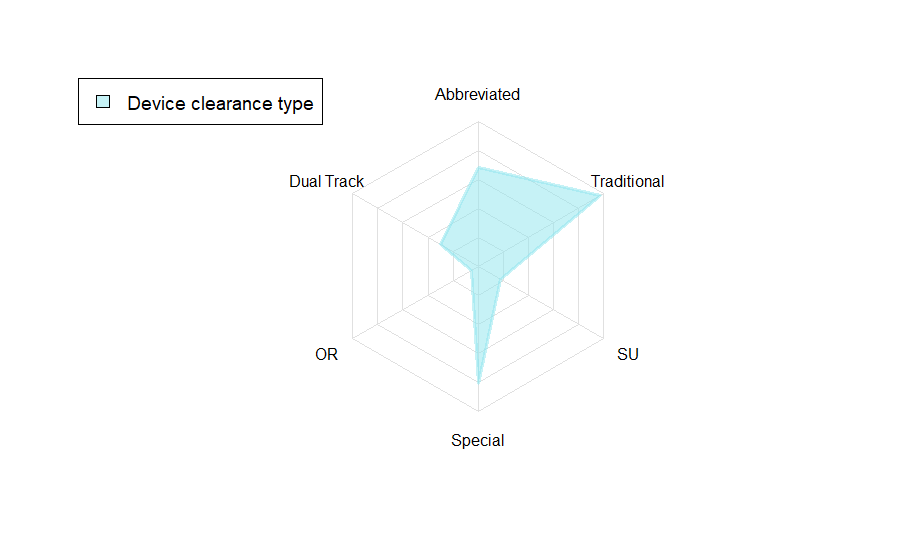
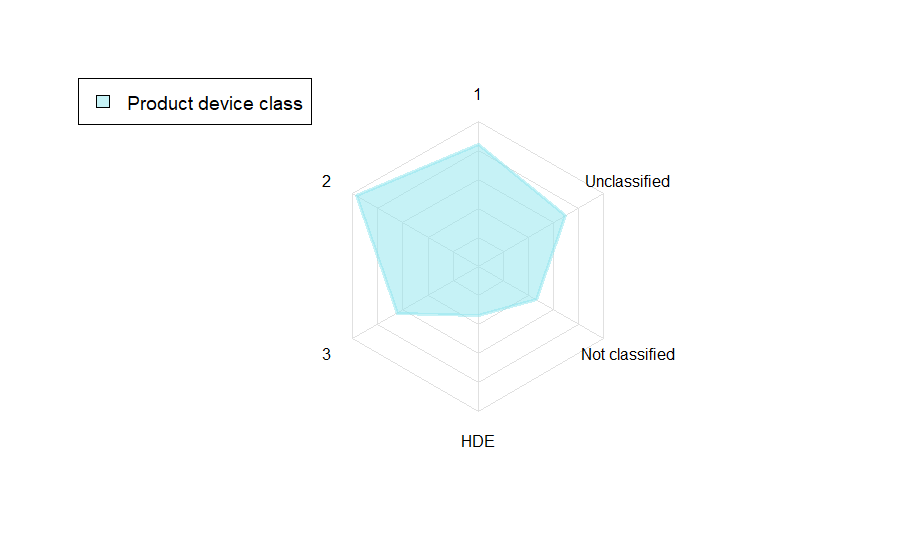
## Plot6：words cloud

对alldata中的device name列绘制了词云，词云为交互图，能显示该单词在所有device name中的出现次数，在R中运行完毕后，光标可以选中查看具体频数。

对于不想要的词，可以在stopwords.txt中添加，注意每行一个



## Plot7：radar plot



## Plot8：lollipop plot

