2017－2018 年第一学期

《大数据统计基础》试题　答题纸

学校：中央财经大学　　学号：2017210785　　姓名：司徒雪颖　　成绩_____

三、数据可视化

1. （1）风玫瑰图

由图 3-1 可以看出，贷款等级越低，贷款期限为 60 个月概率越大

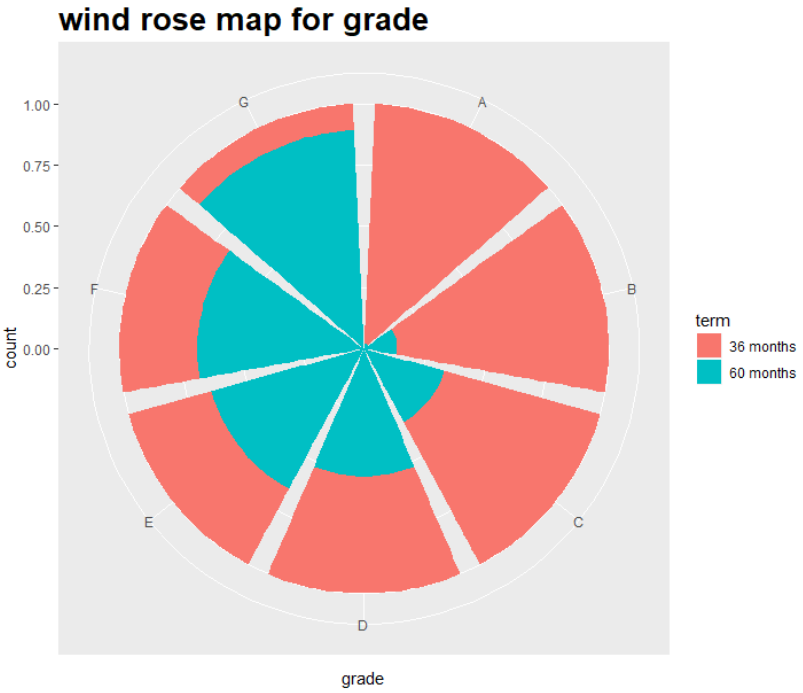

图 3- 1 贷款等级和期限的风玫瑰图

（2）直方图，并加入拟合分布线，直方图的组距和组数自己设定（不要使用默认的），并且每个柱子里面填上相应的组的频数，整个图片加上一个黑色的外框，并且图的底色为浅色，柱子为深色，在密度最高的部分加上文字标注“此处密度最大”

由图 3-2 可以看出 loan_amnt 贷款额在 10000 左右密度最大，频数为 36524.

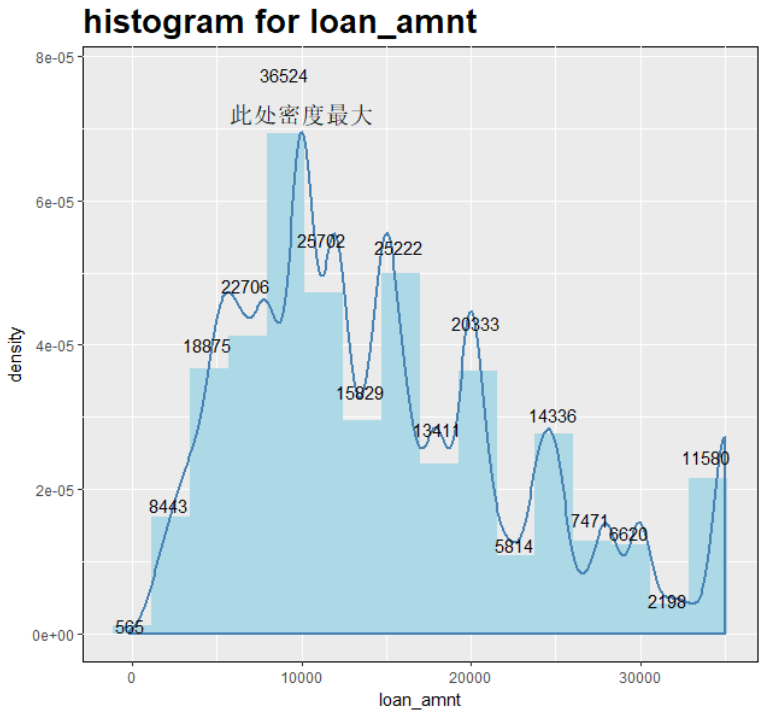

图 3- 2 贷款数额分布直方图

（3）某两个连续型变量的密度图，并且在图中找出一个部分加上一个方框与其他部分区别开来

由双变量直方图 3-3 可以看出，期限为 36 个月的贷款在 5000-10000 元之间达到峰值，跨度从 0 到 35000 元，而期限为 60 个月的贷款在 15000 左右达到峰值，跨度为 5000 到 35000 元。

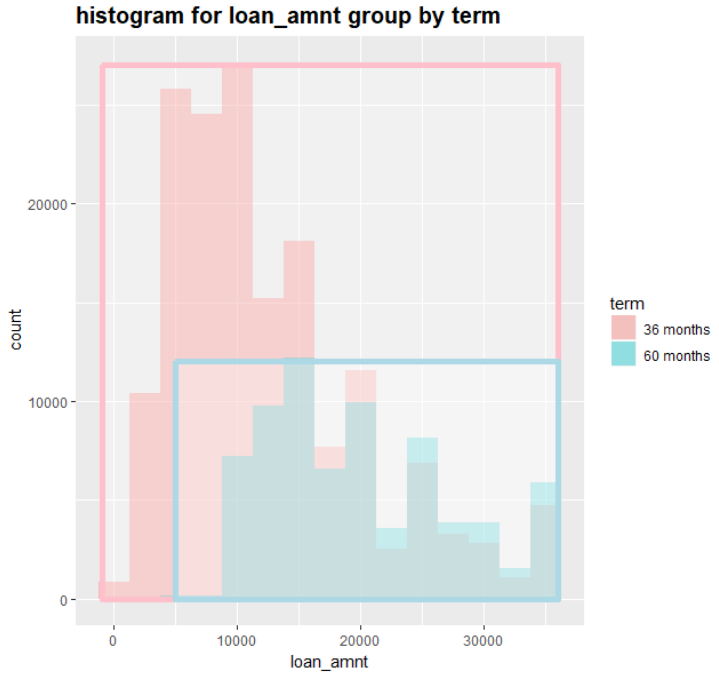

图 3- 3 两种期限的贷款数额分布直方图

（4）挑选多个连续型变量，进行聚类，并且绘制相应的热图，并进行美化（可以不用全部样本）。（2 分）

选取 out_prncp，out_prncp_inv，total_pymnt，total_pymnt_inv，total_rec_prncp，total_rec_int 这 6 个变量，随机抽取 100 个样本，画聚类热力图。从图 3-4 可以看出其中一些样本的距离是很接近的。

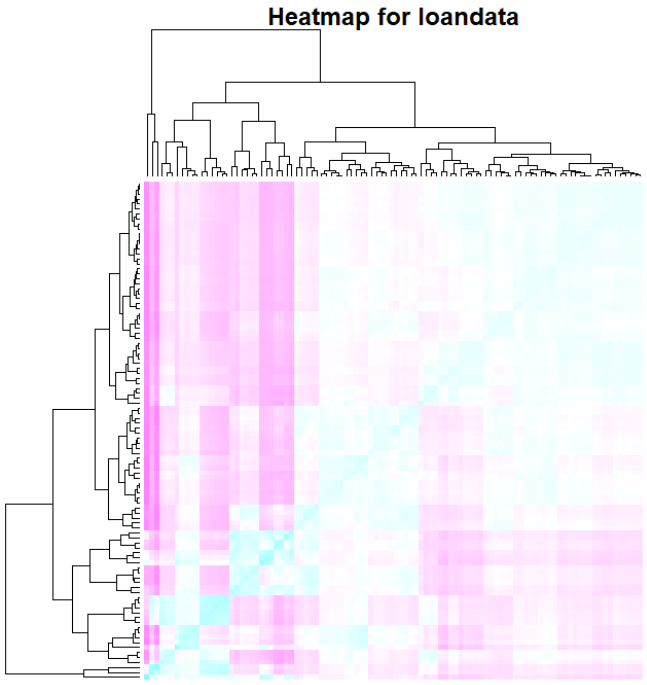

图 3- 4 贷款数据热力图

2、使用 province 数据中合适的变量，绘制两幅不同的图，进行空间数据的展示。（10 分）

注：交互式地图的 html 文件在附件

由图 3-5 热力图可以看出，GDP 最高的区域为北京，长三角。

由图 3-6 色彩图可以看出，人口最多的省份为山东，河南，四川，广东

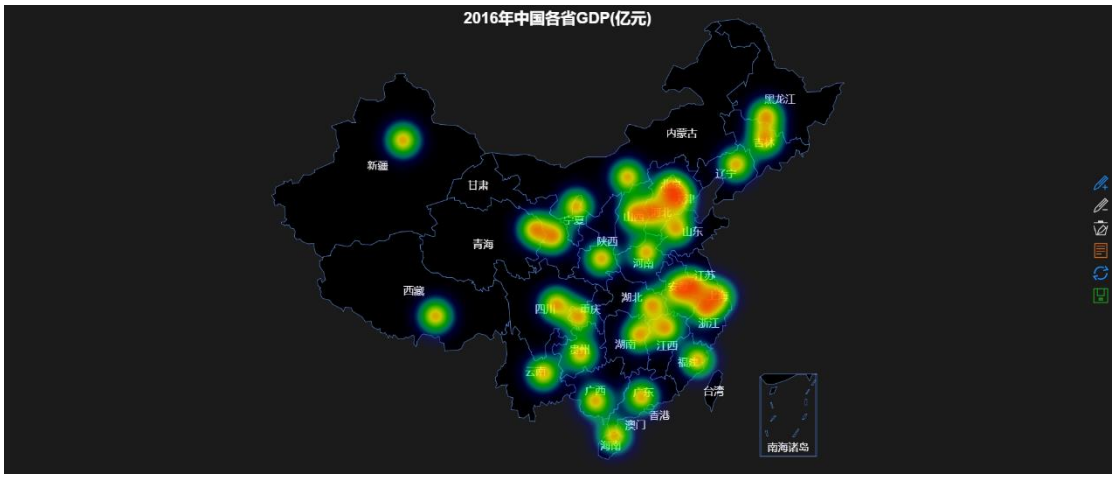

图 3- 5 2016 年中国各省 GDP 热力图

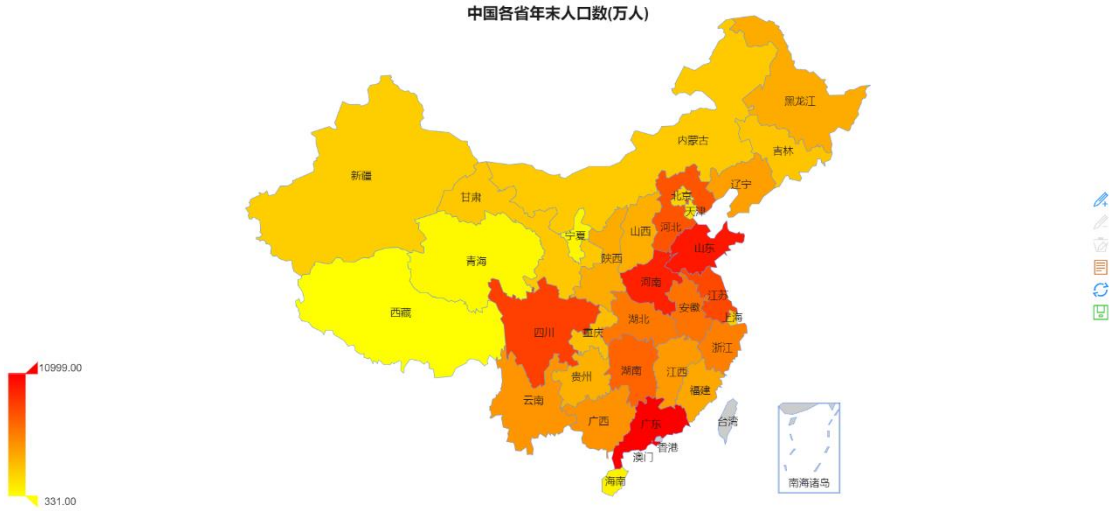

图 3- 6 2016 年中国各省年末人口数色彩图