《大数据统计分析软件（R语言）》

实 验 报 告

指导教师： 范云菲

专 业： 经济统计学

班 级： 经济统计学二班

姓 名： 张义举

学 号： 202131100137

完成时间： 2023年6月9日

2. 实验内容

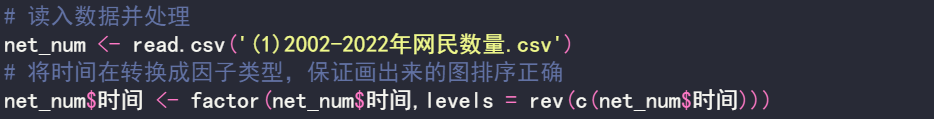
首先导入所用到的R包



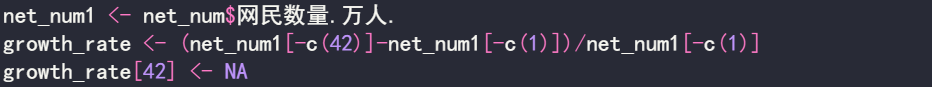
说明：首先修改数据集的编码格式为'utf-8'，防止读取数据集错误，每一个数据集中19年12月的数据是缺失的，这里用20年3月的数据代替19年12月的数据。



2.1 读入数据集：“(1)2002-2022年网民数量.csv”



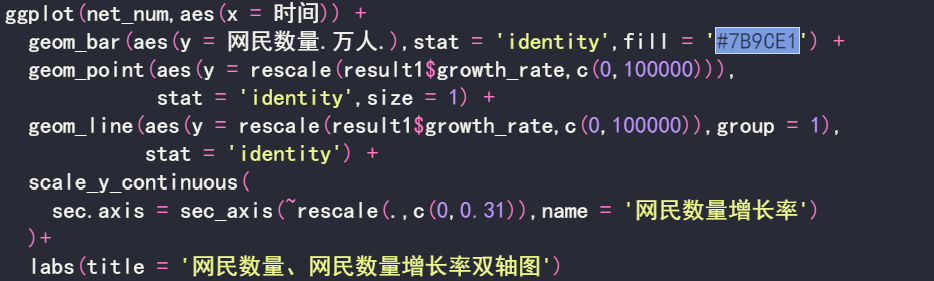
(1) 计算半年度网民数量增长率；



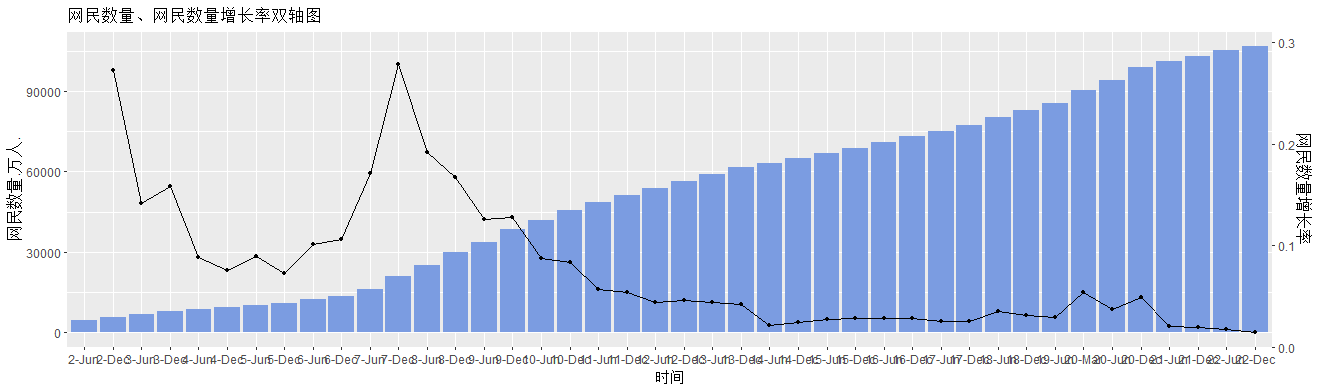
网民数量增长率结果如下：



(2) 根据网民数量、网民数量增长率绘制双轴图，网民数量用柱状图表示，网民数量增长率用折线图表示；



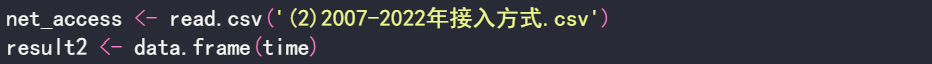
结果如下图所示



(3) 根据图形说明我国网民规模的变化趋势与特征。

总体上看，我国网民数量呈增长趋势，增速在2002年到2008年达到高峰，接着增速整体上逐年递减。

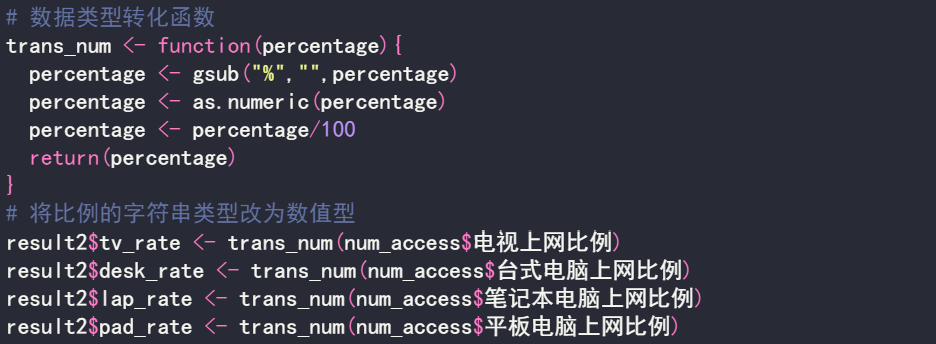
2.2 读入数据集：“(2)2007-2022年接入方式.csv”

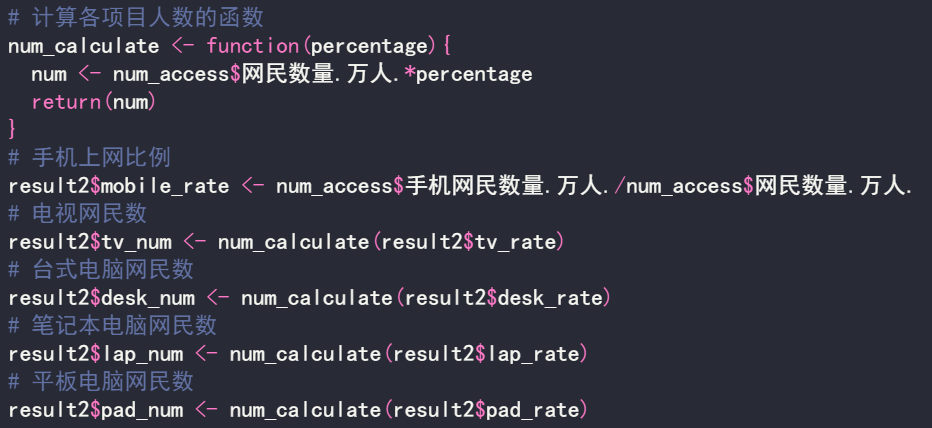


1. 与“(1)2002-2022年网民数量.csv”进行合并

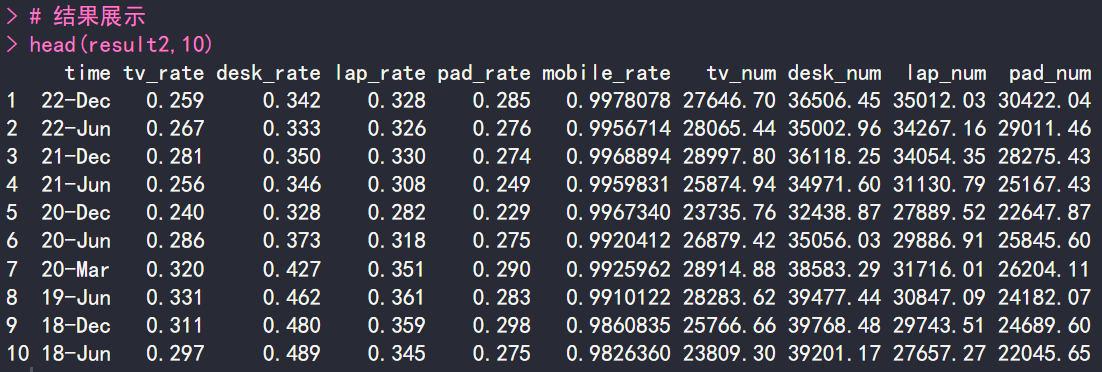


1. 计算出手机上网比例、电视网民数、台式电脑网民数、笔记本电脑网民数、平板电脑网民数（不同接入方式人数有重合，总数会大于真实网民数量）



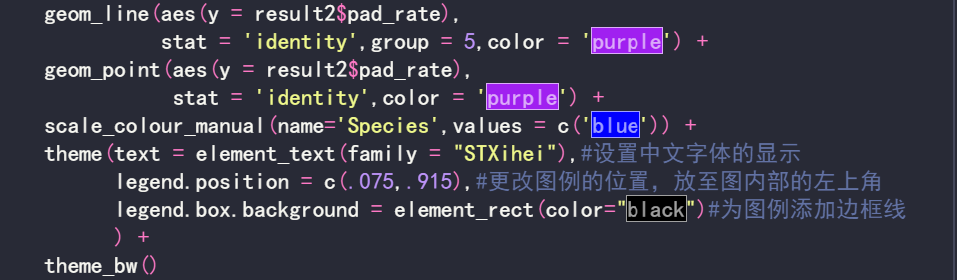


结果展示

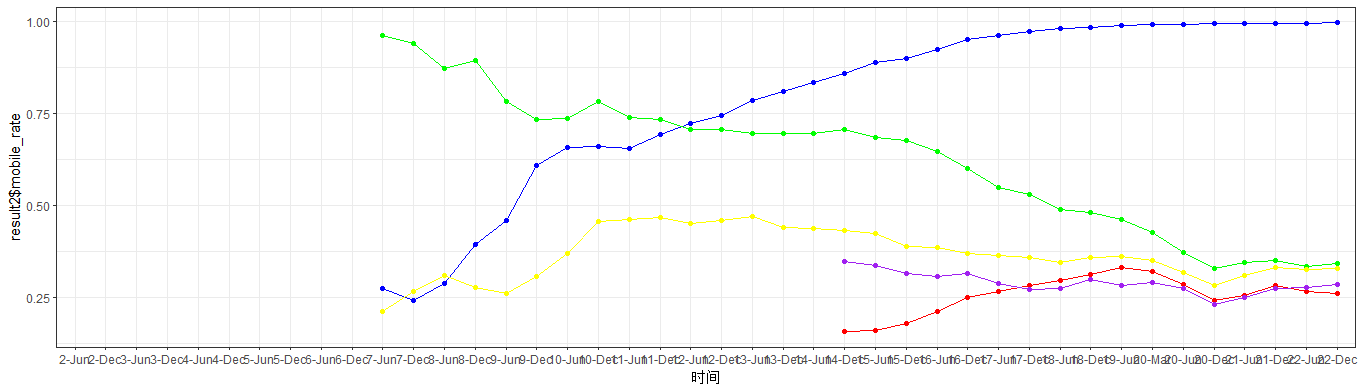


1. 根据数据绘制各种接入方式比例的折线图，并比较说明每种接入方式的变化趋势与特征。



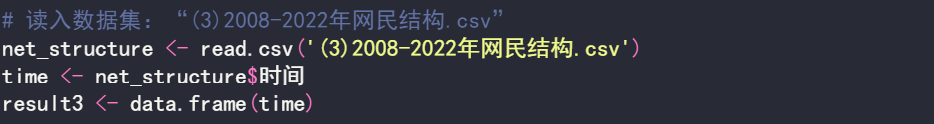


绘图结果如下

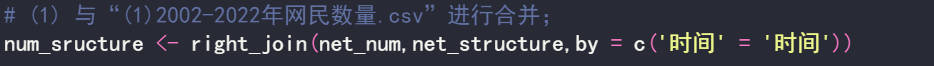


特征趋势：由折线图可以看出，手机上网比例整体上逐年上升，到2012年年底成为上网比例最高的接入方式；电视上网比例由2015年至2019年逐年上升，之后有所下降，2020年年底又上升；台式电脑整体上逐年下降；笔记本电脑上网所占比例最低，此外呈逐年下降趋势，平板电脑分别在2017年之前，2018年年底，2020年年底之后这几个时间节点之间下降，上升，下降，上升，总体上下起伏。

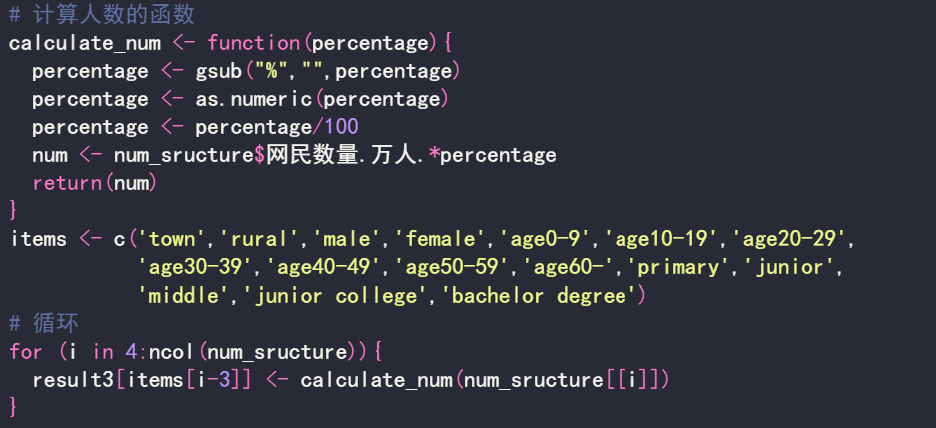
2.3 读入数据集：“(3)2008-2022年网民结构.csv”



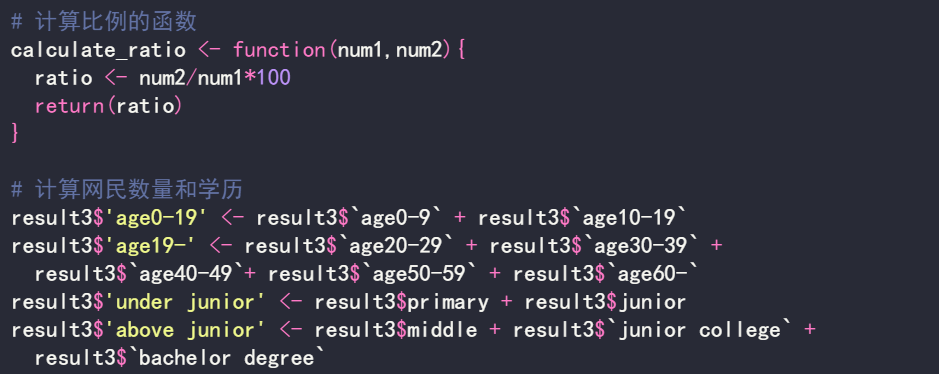
1. 与“(1)2002-2022年网民数量.csv”进行合并；

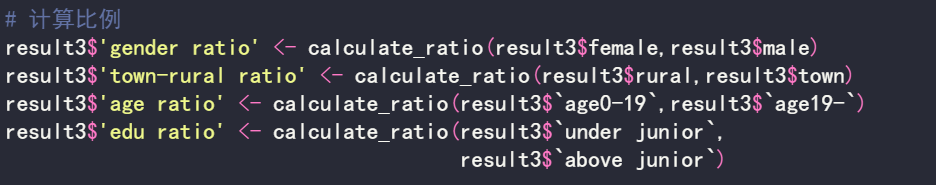


(2) 编写函数计算出城镇网民数量、农村网民数量、男性网民数量、女性网民数量、各年龄层网民数量、各学历网民数量（可以编写函数通过apply族函数进行批量处理，也可以直接编写循环函数直接一次性得到所有结果）；

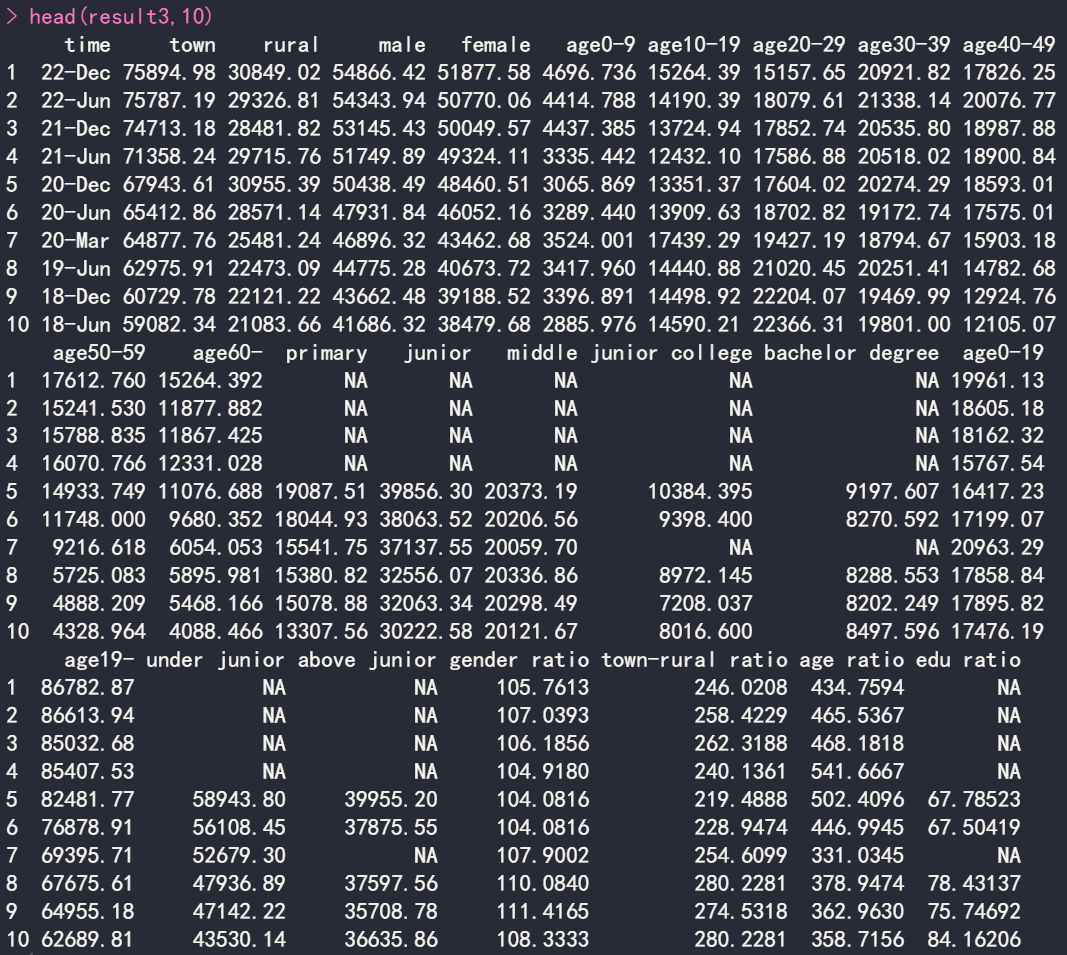


(3) 进一步计算网民性别比（以女性网民为100，男性对女性的比例）、城乡比（以农村网民为100，城镇对农村的比例）、年龄比（以19岁以下网民为100，其他年龄层对19岁以下的比例）、学历比（以初中及以下网民为100，其他学历对初中及以下的比例）；

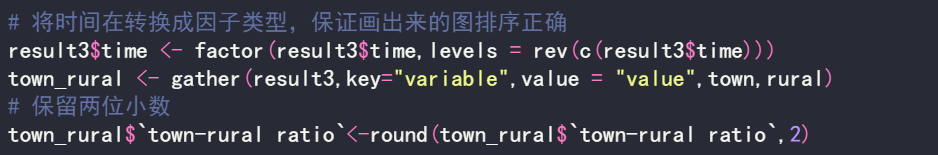


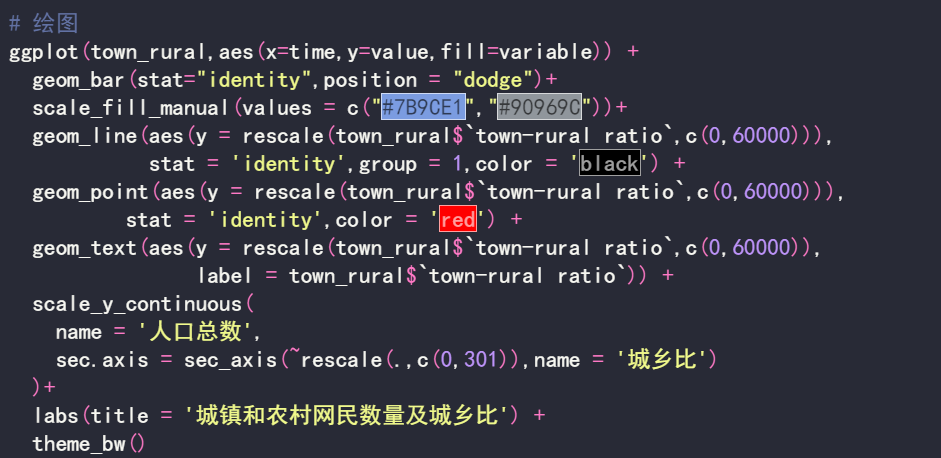


显示结果：

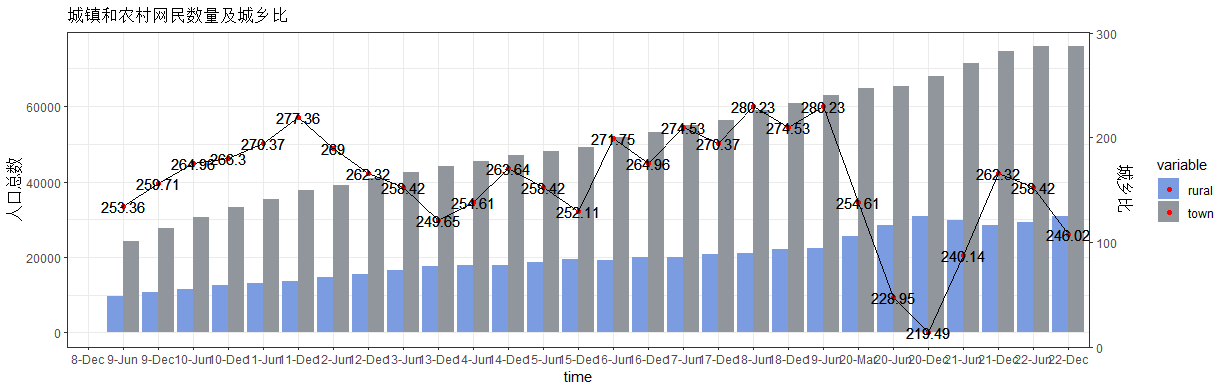


(4) 根据城镇和农村网民数量以及城乡比绘制柱状图和折线图（类似下图所示），最后说明城乡网民的变化。





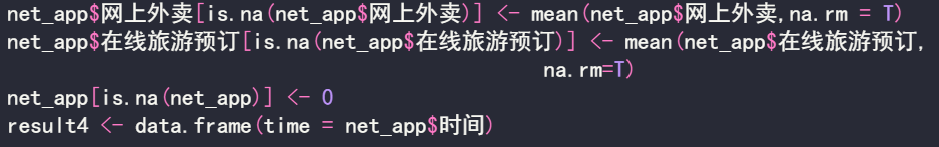
绘图结果：



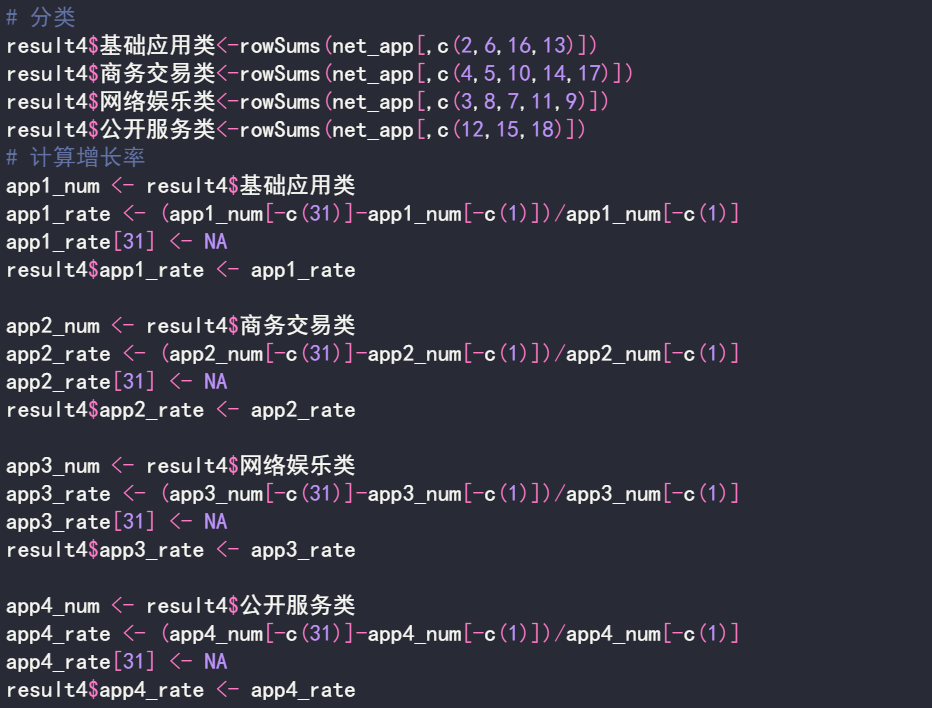
2.4 读入数据集：“(4)2007-2022年各应用网民数.csv”（删除前面的说明部分）

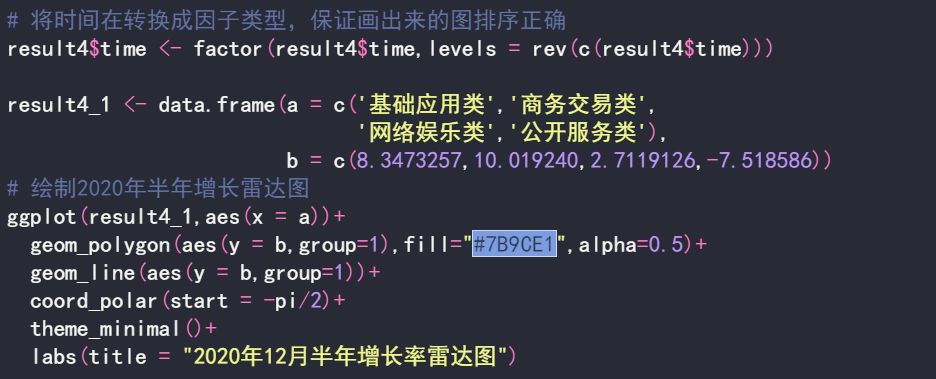


(1) 将“网上外卖”“在线旅游预订”这两列中的空缺值进行填充（其他列中的空缺值不需要处理）；

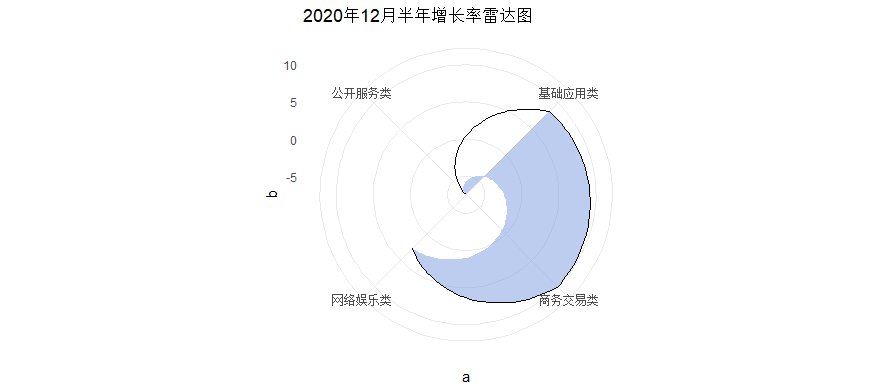


(2) 根据该数据集前4行的说明部分，将17类应用综合为4个大类：基础应用类、商务交易类、网络娱乐类、公共服务类，汇总计算4大类各自的网民数量、半年度网民数量增长率，绘制2020年12月这4大类半年度网民数量增长率的雷达图；





绘图结果：

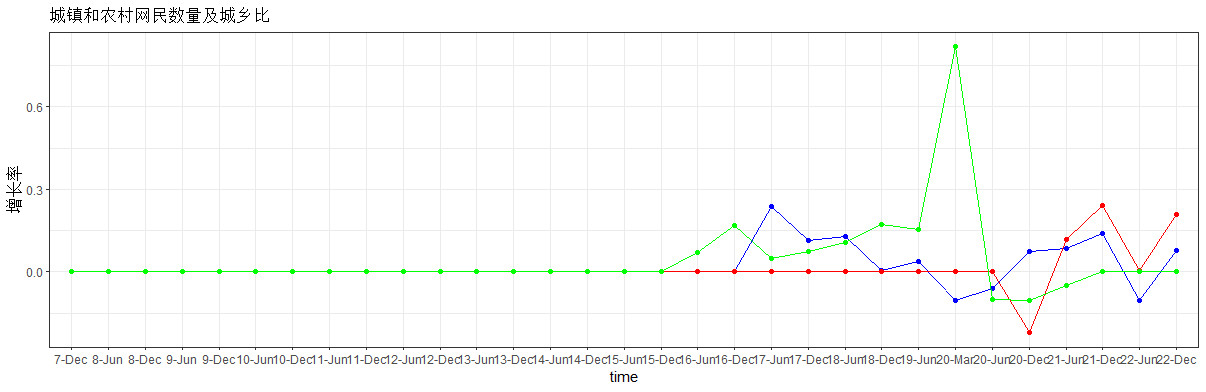


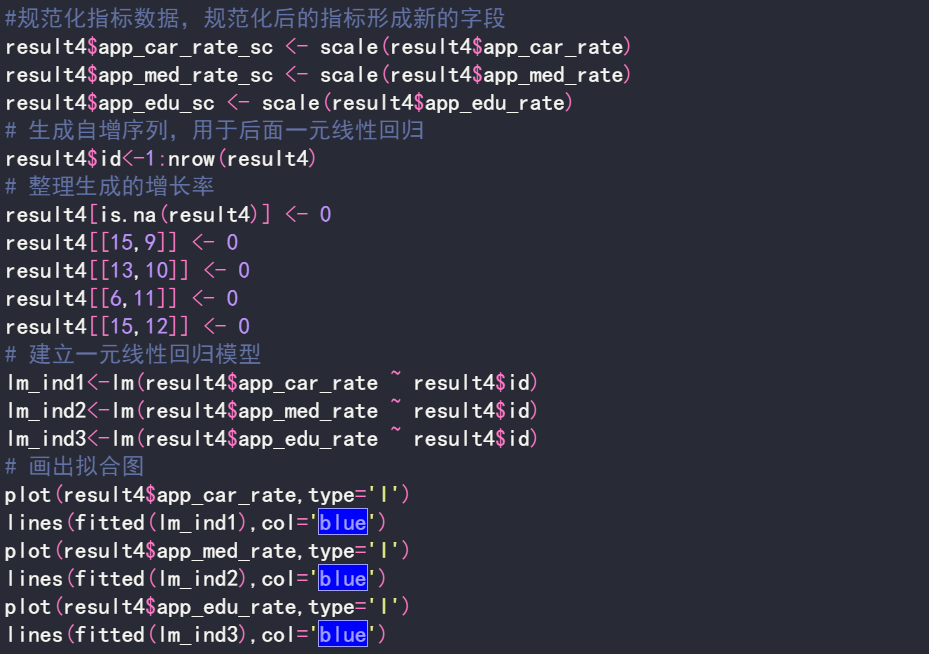
(3) 从4大类中任意选择1类进行详细分析，例如：公共服务类中包含网约车、在线医疗、在线教育，绘制这3类应用的半年度网民数量增长率，说明各自的特点；绘制网民数量和时间的拟合线（参照“10互联网运营指标建立”中的代码）；结合社会现象说明这3类应用网民数量变化的特征（不需要局限于这里的举例，请自行进行多角度分析）。



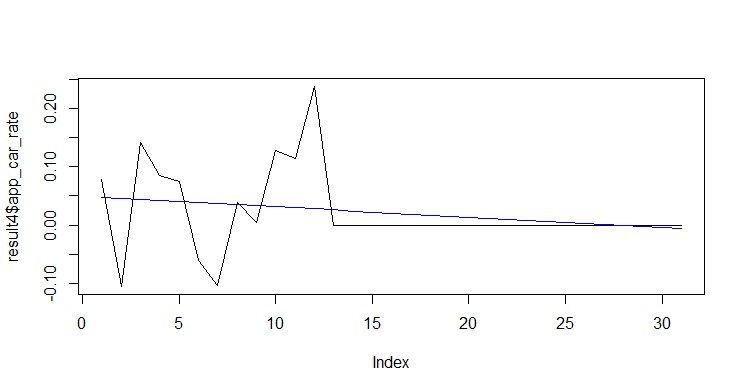


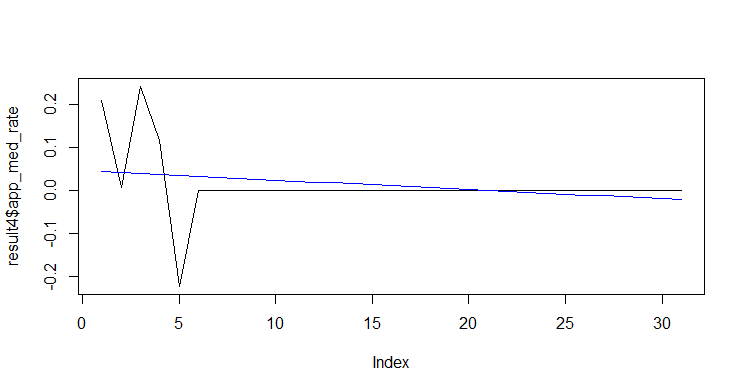
绘图结果：

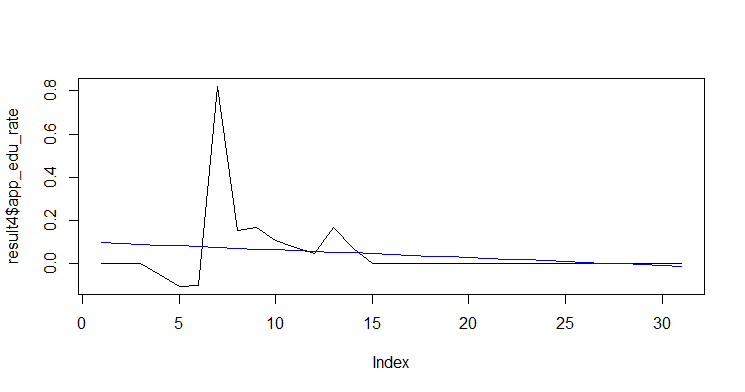




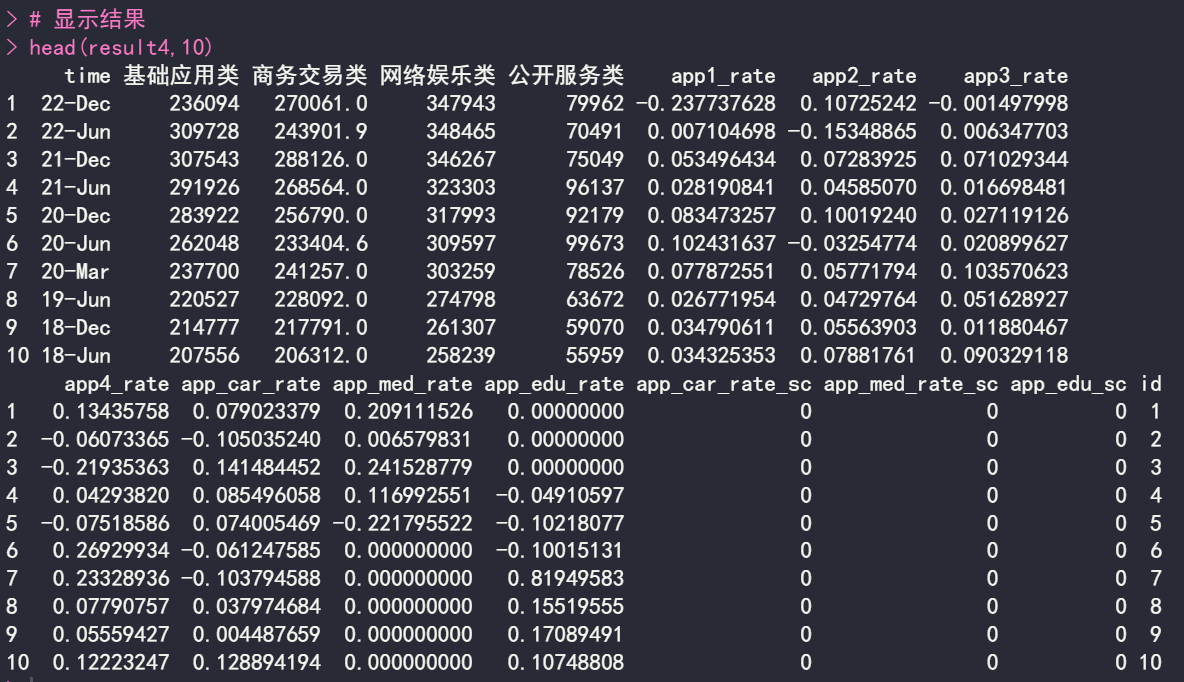
绘图结果：



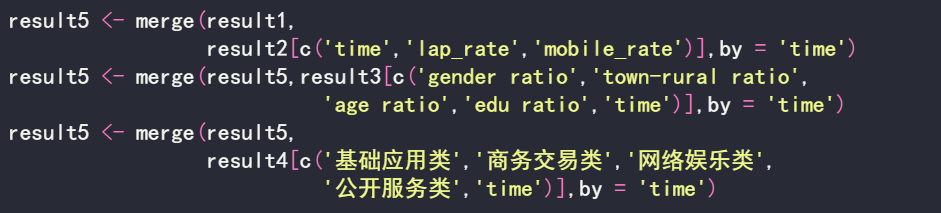




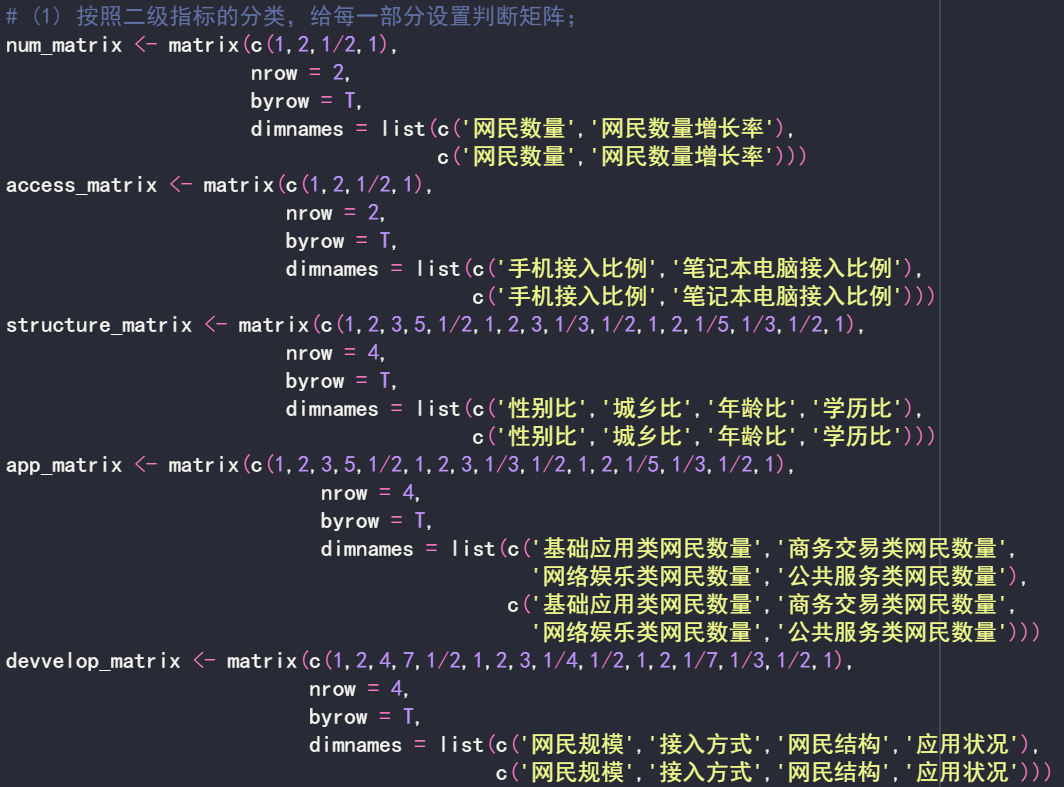
特征：由图像可以看出，网约车，在线教育，在线医疗都经历过负的增长率。在线教育在19年之前增长率上升，19年增长率最大，是因为19年12月爆发疫情，线下授课改为线上，之后增长速度放缓；在线医疗开始增长率为负值，接着逐年增大，到21年12月达到最大值，之后有所下降；网约车17年6月最大值，之后下降，19年到达最低，以为疫情，网约车停止运行，到20年12月开始上升，21年12月下降，22年6月下降到负值。

显示结果：

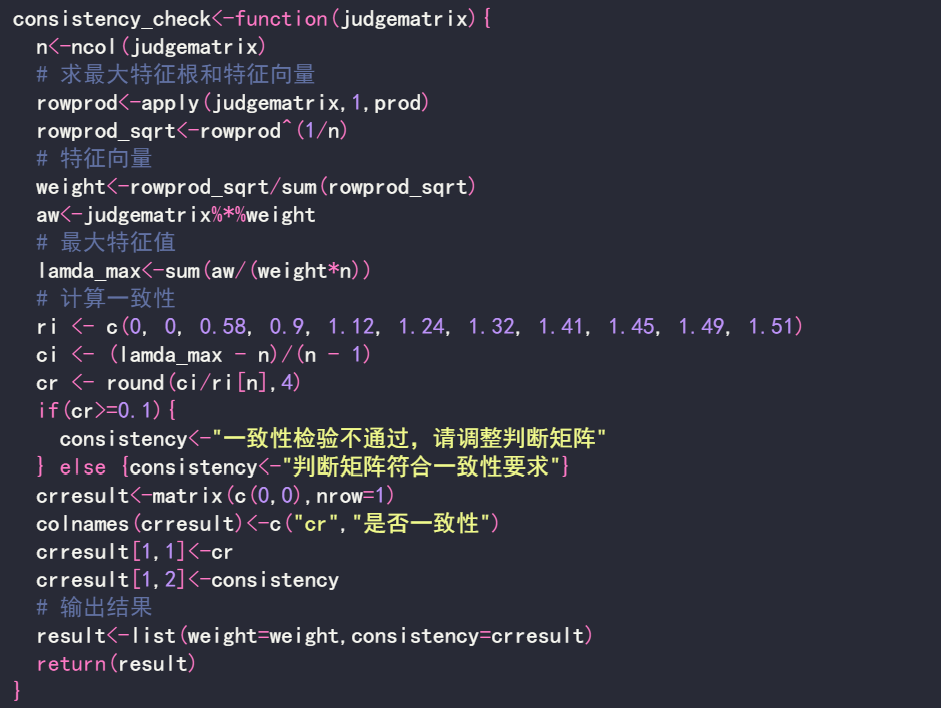
2.5 合并2.1、2.2、2.3、2.4中处理后的数据集



(1) 按照二级指标的分类，给每一部分设置判断矩阵；



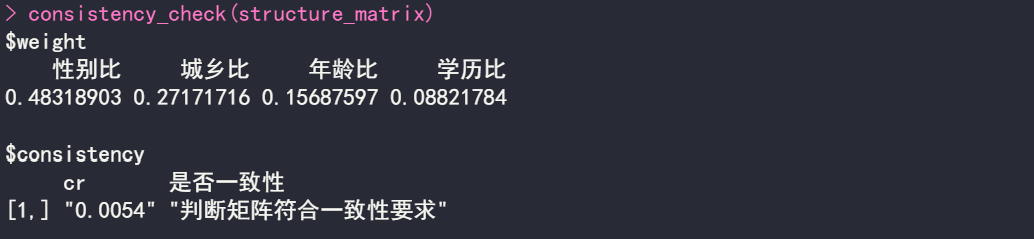
(2) 编写函数对判断矩阵进行一致性检验（函数中不需要对是否正方阵、是否元素数大于11进行检验，最后的结果中也不需要汇报最大特征值，只需要汇报权重和判断结论；

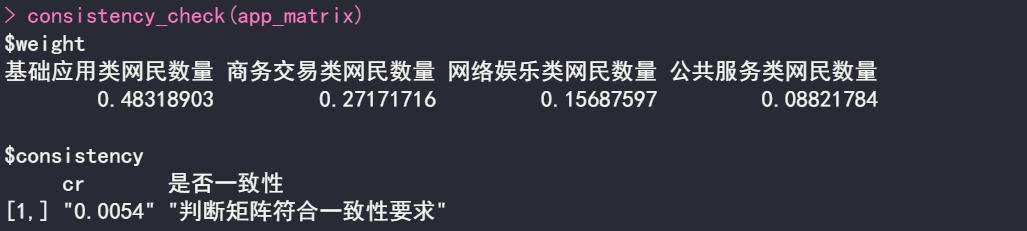


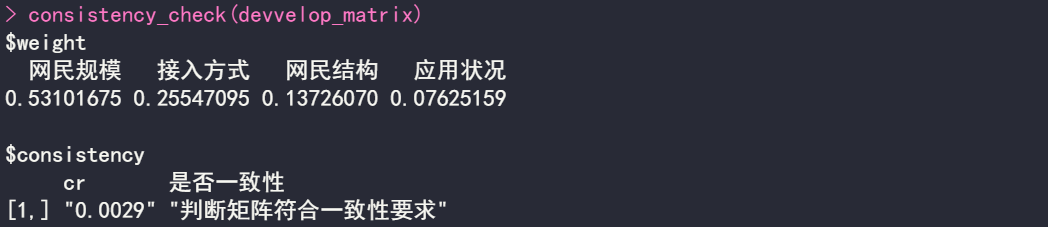
(3) 应用上一步骤中函数计算各级权重，再与实际数据进行加权平均，计算出我国互联网发展指数的时间序列数据。



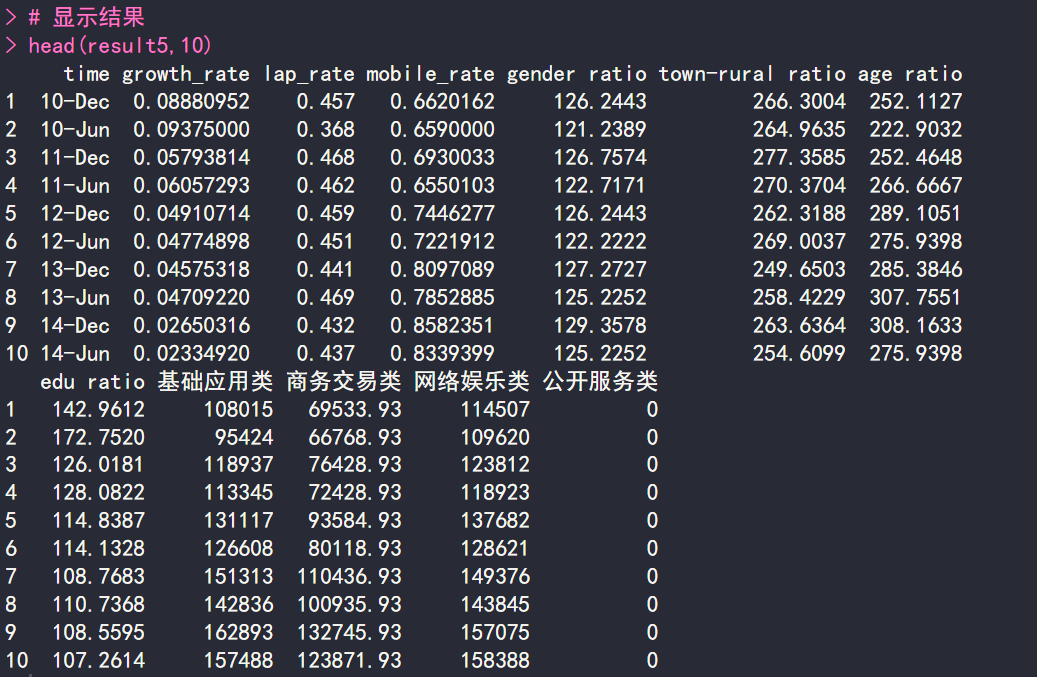
输出的结果：







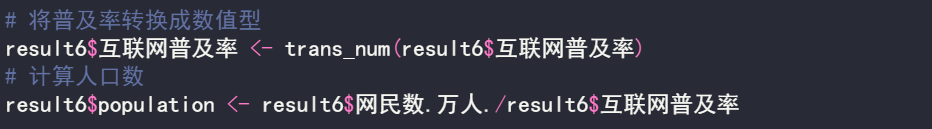
显示结果：



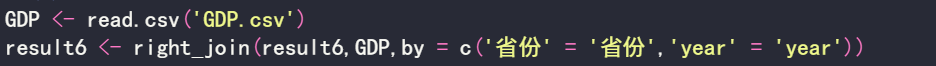
2.6 编写函数批量读入文件夹“各省历年网民规模”中的所有数据集的同时，在每一个数据集中新增一列表示年份的变量，最后把所有数据集进行行合并



(1) 根据数据中各省历年网民规模和互联网普及率，大概计算各省历年人口数；



(2) 收集并整理2007年—2016年我国31各省自治区（港澳台地区除外）的GDP，并与各省历年网民规模数据集进行合并；



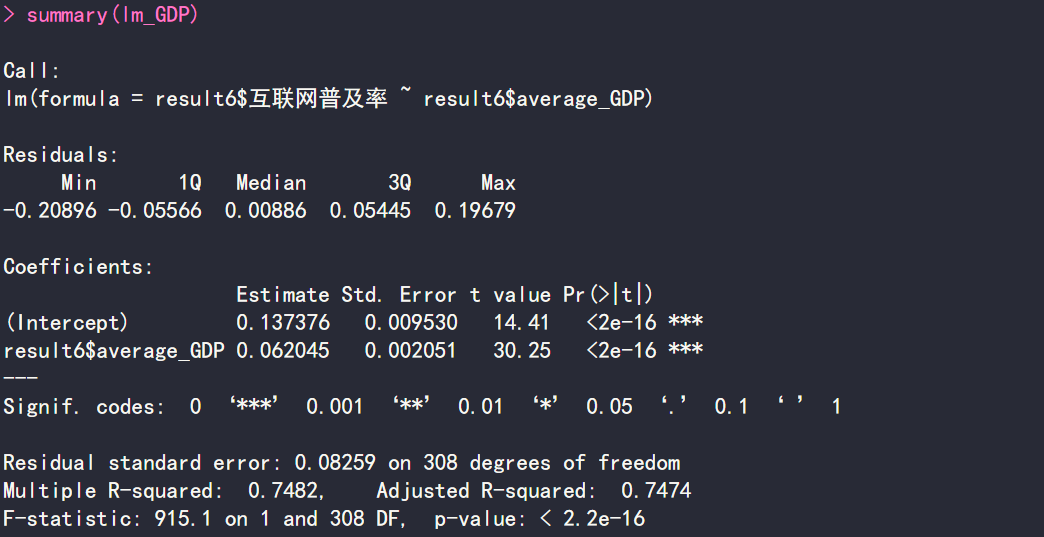
显示结果：



(3) 将各省历年互联网普及率作为被解释变量，人均GDP作为解释变量，进行一元线性回归（Hint：lm(互联网普及率~人均GDP)），说明人均GDP是否会对互联网普及率起到正向作用。



查看回归结果：



统计显著性检验显示：

人均GDP的系数估计值在95%的置信水平下是显著的（p-value < 0.001），说明人均GDP对互联网普及率的影响是具有统计学意义的。调整后的决定系数为0.7474，说明回归模型可以解释约74.74%的互联网普及率的变异。综上所述，根据该线性回归分析，人均GDP对互联网普及率具有正向影响，即随着人均GDP的增加，互联网普及率也有所增加。