《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：

姓 名：

学 号：

完成时间：

1. 实验目的

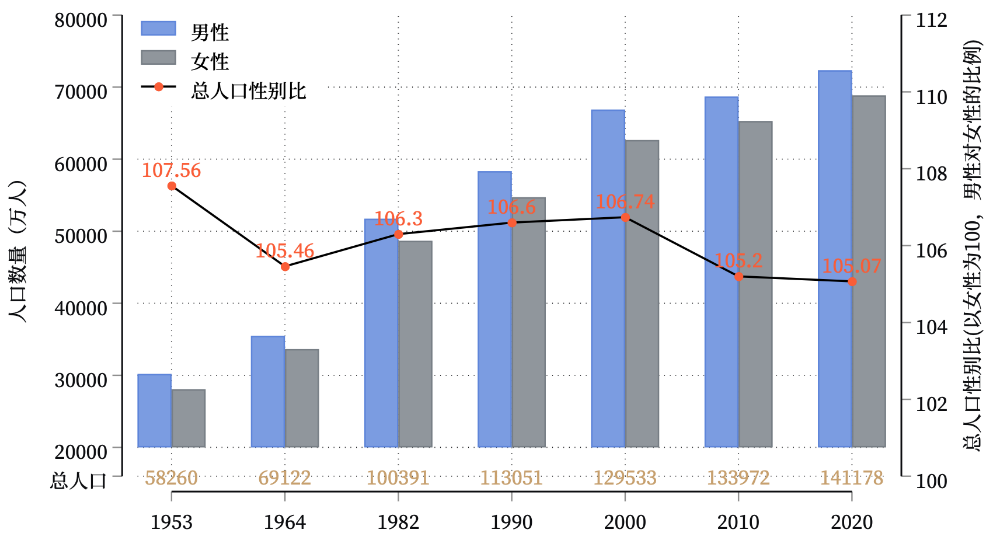
掌握数据读入的各类函数；简单的数据处理：新变量的生成，数据集筛选排序，数据集合并。

掌握dplyr包中的主要函数。

掌握绘图的主要函数。

掌握数据整理、描述性统计分析、回归等，较为完整的数据分析过程。

1. 实验内容
   1. 读入数据集：“(1)2002-2022年网民数量.csv”
2. **计算**半年度网民数量增长率；
3. 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；
4. 根据图形**说明**我国网民规模的变化趋势与特征。
   1. 读入数据集：“(2)2007-2022年接入方式.csv”
5. 与“(1)2002-2022年网民数量.csv”进行**合并**
6. **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）
7. 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。
   1. 读入数据集：“(3)2008-2022年网民结构.csv”
8. 与“(1)2002-2022年网民数量.csv”进行**合并**；
9. **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）；
10. **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）；
11. 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。（bonus：主题设置、颜色设置、细节处理基本跟下图完全一样）



* 1. 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；
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）。
   1. 合并2.1、2.2、2.3、2.4中处理后的数据集，保留的变量如下表中的三级指标所示：

|  |  |  |
| --- | --- | --- |
| 一级指标 | 二级指标 | 三级指标 |
| 互联网发展指数 | 网民规模 | 网民数量 |
| 网民数量增长率 |
| 接入方式 | 手机接入比例 |
| 笔记本电脑接入比例 |
| 网民结构 | 性别比 |
| 城乡比 |
| 年龄比 |
| 学历比 |
| 应用状况 | 基础应用类网民数量 |
| 商务交易类网民数量 |
| 网络娱乐类网民数量 |
| 公共服务类网民数量 |

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。
   1. 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**：
4. 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；
5. 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；
6. 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。
7. 实验要求

完成实验内容中的每一步，请附上代码，并使用“head()”函数对每一个数据集结果前10行进行部分展示，贴出结果。

请直接将内容呈现在“实验报告.docx”中，重命名为姓名学号，并将代码写清楚注释，形成以姓名学号命名的“姓名学号.r”脚本文件，与“实验报告.docx”一起存放在同样以姓名学号命名的文件夹中（不要压缩），提交给学习委员后，一起压缩打包发给我