山东大学 软件 学院

**数据可视化** 课程实验报告

|  |  |  |  |
| --- | --- | --- | --- |
| 学号：  202000300xxx | 姓名：叶昕骅 | | 班级：20大数据x班 |
| 学号：  202000300xxx | 姓名：苏锡良 | | 班级：20大数据x班 |
| 学号：  202000300xxx | 姓名：李超昊 | | 班级：20大数据x班 |
| 学号：  202000300xxx | 姓名：赵尚祁 | | 班级：20大数据1班 |
| 实验编号：大作业3 | | | |
| 实验题目：交互式可视化 | | | |
| 实验学时：16 | | 实验日期：2021.5.30 | |
| 作业主题与分析目的：  北京人口流动可视化分析  世界人口稳定增长带来的迁徙现象越发明显，由此引发新一轮人口流动分析，学者们致力于开采人口流动可视化中的巨额财富。本项目致力于中国首都-北京市的人口流动可视化，期望在世界性人口流动分析趋势中取得丰厚成果。  新型冠状病毒肺炎疫情牵动人心，举国关注。人口的流动聚集，客观上加大了疫情传播的风险和防控的难度。出于对公共卫生、重大公共利益相关影响的为研究目的，为进一步掌握人员流动聚集动向，做好紧急疫情的防控工作。本组成员针对人群密度进行时空数据可视化和预测 | | | |
| 分工与工作量占比：  苏锡良：绘图 25%  李超昊：前后端 25%  赵尚祁：前后端25%  叶昕骅：数据挖掘25% | | | |
| **设计与技术：**  交互式前后端部分：  前端主要使用了html 5、css 3、js，ajax，jQuery等技术。HTML编辑页面文字部分，css页面美化，JavaScript实现动画以及鼠标交互方法等函数的编写。在编写前端时，使用了ajax库来实现页面的动态更新，后续可通过接口拓展实时功能。jQuery库简化JavaScript的编写，自带以及封装好的ajax方法。  后端使用了python、flask框架、mysql等技术。Python的flask框架搭建整个系统的后端，并与前端进行交互提供服务如实现查询功能的方法。后端连接数据库可以调用数据。后端python编写一下数据处理，文件读取等方法。  具体布局如下：            **图表绘制部分**  图标的绘制主要用到了echarts，以及kepler，大体分为四个模块，前三个用到echarts，后两个用到kepler  ① 总览：   1. 2020春节出入京人流信息相关图   此系列图表用到migration\_index.csv里的信息，将2020春节期间北京人口出入的信息展示了出来。图形大都带有时间轴（timeline模块），可以看到数据随着时间的变化。   1. 迁徙流向图   此图用到了echarts的geo模块，展现的是每天出入京top100的市及整体的流动轨迹。  descript   1. 出入京数据总览图   此图采用bar模块，详细展示了各省份的出入京总迁徙数据，用以横向比较不同省份的数据总量  descript   1. 出/入京人流量来源/去向组成图   采用circle模块，以扇形图展示出入京人口的组成  descript   1. 总体数据变化   使用line、river等图来展示总数据的变化，主要侧重于时间而不是省份  descriptdescriptdescriptdescriptdescript   1. 市内人口流动相关图   主要用到area\_passenger\_index.csv里的信息，展示北京市内的人口流动信息和趋势   1. day/hour总人流量统计图   横轴和纵轴分别代表日期和小时，展现的是北京市人口流动的总趋势  descript   1. 不同类型场所人流量雷达图   将场所划分为六个类型，将夜间和白天数据分开，绘制雷达图，以查看区别  descript   1. 不同类型总人流动指数变化图   用到了stackline，既能看到时间变化趋势，也能看到不同类别的对比  descript   1. 同时段人流量横向对比   用以展示每天同一时段人流量的变化  descript   1. 3维图（多场景多维人口流动对比）   descript   1. kepler相关图   descript  ② 制作方式   1. 数据处理   将csv导入sql server，利用sql动态查询，以得到合适的数据。   1. python和ecarts的结合   使用pyecharts包，将python语言翻译为echarts，进而绘出图形   1. 图形设计   整体风格一致，讲究色彩搭配。视觉通道较为突出合理。整体性较强。  ③ 克服的问题   1. 数据庞大，加载负担重   使用适宜的方法进行缩减，如聚集函数等，或按照一定规律进行筛选，减少信息条数。此外，将数据按照日/小时来进行划分也能减少总数据量   1. 视觉通道设计   不停尝试，找到适合的视觉通道   1. 颜色选取   保证和谐性的同时要求颜色能够突出显示。为此登陆各大选色网站一点一点进行调试   1. 布局   图上的元素应该放在哪个位置也应该提前决定好。  **数据挖掘部分：**  我们数据挖掘的目标是通过30天的数据预测未来7天的北京区域人口密度指数   1. 数据读取   首先对数据集的数据使用python的pandas进行读取，转成dataframe数据格式，共使用五个csv文件数据：  area\_passenger\_index.csv（地理区域编号ID，时间time以及人口指数people\_index）  area\_passenger\_info.csv（地理区域编号ID,地理区域名称area\_name,地理区域类型area\_type,区域中心经度Center\_x,区域中心纬度Center\_y，区域中心所在网格经度Grid\_x,区域中心所在网格纬度Grid\_y,区域面积area）  migration\_index.csv （日期date,离开省份departure\_province,离开城市departure\_city,目的地省份arrival\_province,目的地城市arrival\_city,迁徙人口指数index）  grid\_strength.csv（小时hour,起始区域中心所在网格经度start\_grid\_x,起始区域中心所在网格纬度start\_grid\_y,到达区域中心所在网格经度end\_grid\_x,到达区域中心所在网格纬度end\_grid\_y,联系人口指数Index）  test\_submit\_example.csv（地理区域编号ID，时间time以及人口指数people\_index）用作最后预测训练集数据存储格式   1. 数据分析与处理   首先我们对前两个集合df1与df2进行分析，两个数据集之间按ID进行连接，通过EDA可视化分析可以看出人口指数与时间和地点种类进行分类后有一定的对比程度，于是按照它们进行特征处理。对于第三个数据集df3，它主要展示了迁入迁出北京的人口指数，那么对于我们数据挖掘的目标，我们只需要处理迁入北京的人口指数以及迁出北京的人口指数，从热力图可以看出，由于所给的数据正值春运时期，所以对于春节假期前的数据与未来七天的人口有着较大的偏差，为避免其对预测造成影响，我们只选取了后13天的数据进行处理和训练。  对于第一个数据集df1，通过简要查看可看出时间的格式为%Y%m%d%H，于是将其分隔开进行分析，将人口指数按 weekday，hour以及ID来进行取平均处理，创建特征。  对于第二个数据集df2，先将其与处理后的df1按ID进行连接，将人口指数按区域类别area\_type取平均处理，添加特征。然后去除掉一些已使用或不相关的特征。  对于第三个数据集，首先进行时间格式处理，然后对离开城市为北京，目的地城市为北京的数据进行提取和分类，对它们的迁徙人口指数index进行针对时间的取平均与标准差的运算，添加特征。对于这些数据，我们需要对其进行未来七天的预测，但这部分没有其他的特征可供使用，只有时间序列特征，于是我们采用时间序列的AR模型来使用以后数据进行未来七天的迁入迁出人口平均值与标准差的预测。  对于第四个数据集df4，由于我们能力有限，无法对center与grid两类数据进行区别分类，于是只对时间进行了取平均处理，添加特征。  对于前四个数据集处理后的特征，将其按其处理方式对第五个数据集进行处理，构造测试数据。   1. 建模调参   我们一共使用了xgb、随机森林以及lgb进行模型构造，并采用k折交叉验证。最后发现lgb+10折交叉的效果最佳。  使用python类库pandas，numpy，sklearn，statsmodels，matplotlib，seaborn等  平台：数据分析使用百度飞桨，建模调参使用山东大学高性能计算平台  二维码部分：采用github上的开源python类库amzqr进行二维码生成，并使用gitee与腾讯云的静态网页管理生成域名。 | | | |
| **测试与运行结果分析：**  本项目的主题为：疫情及春节期间，北京市各区域以及北京市对外对内人口流动分析，运用一系列大数据可视化技术，帮助用户分析疫情对北京春运期间人口流动的影响。世界人口稳定增长带来的迁徙现象越发明显，由此引发新一轮人口流动分析，学者们致力于开采人口流动可视化中的巨额财富。本项目致力于中国首都-北京市的人口流动可视化，期望在世界性人口流动分析趋势中取得丰厚成果。  模块介绍：   1. 春运可视化：本模块通过全国迁徙流向图（针对北京市）和一系列饼状图帮助用户分析挖掘春运中北京市人口流动的规律以及找寻疫情之下北京市人口流动的相关受体，分析新冠疫情对春运期间北京市人口流动的影响。 2. 市内人口流动可视化：本模块通过以时间、地域为单位分析北京市市内人口流动的规律和现象。针对不同区域的不同人口流动占比我们可以发现北京市人口密集的区域和场所，且疫情对区域人口流动的影响不容忽视，我们可以借助这个模块对不同区域的人口流动进行全方面的分析，包括全天各个时段的针对分析和借助热力图的全面分析。 3. 北京市场所人口流动可视化：本模块包含13个不同的北京市市内具有代表性的场所比如动物园、植物园、火车站、博物馆和少年宫等等。这些区域的人口流动情况将会以柱状图（三维立体）的方式进行可视化展示，以日期和具体时间为单位进行可视化探索，帮助分析春节和疫情对北京市市内的代表性场所的人口流动规律以及深入分析导致这些规律的原因。 4. 区域点对点人口流动地图可视化：本模块针对不同的月份以及未来的时间，以简单直观的方式呈现北京市市内各区域之间的人口流动情况，且以点线结合的方式予以具体的展现。本模块用于具体分析北京市市内人口流动情况，因为其数据非常具体而有利于作出深入分析。 | | | |
| **总结：**  我们小组四人分工明确，有条理有阶段性地顺利完成可视化大作业的制作，并取得了较好的效果。每个人都从零开始学习并上手，都学习了可视化的相关技术，如echarts的使用。同时为了满足需要，我们还针对性学习了前后端以及数据挖掘的相关技术，提高交互性与实用性。小组成员还多次对大作业进行开会讨论，按照预定时间完成，优化，展示课程项目。  从本次大作业设计与制作，深刻感受到了可视化的优点与实用性，并对数据可视化这门课程有了更深刻的认识和思考，对于课上的知识有了较好的实践和复现，大家都感到成就感，受益匪浅。最后感谢老师的指点与帮助。 | | | |