2023 年“泰迪杯”数据分析技能赛

A 题

档案数字化加工流程数据分析

一、 背景

档案数字化是随着扫描、OCR、数字摄影、数据库、多媒体、存储等技术的发展而产生的一种新型档案信息处理技术，它把各种载体的档案资源转化为数字化档案信息，以数字化的形式存储，网络化的形式互相连接，利用计算机系统进行管理，形成一个有序结构的档案信息库。我国档案工作采取“存量数字化、增量电子化”的信息化战略。当前我国各行业的存量档案数量巨大，档案数字化的需求不断增加，档案数字化加工行业的市场规模呈现逐年增长的趋势。

二、 目标

对加工流程数据进行统计分析，并作可视化展示，便于管理人员及时了解档案加工处理动态。

1. 统计档案数字化流程的耗时和进度情况。
2. 统计操作人员的工作量和工作效率情况。

三、 案卷加工流程说明

1. 加工流程按先后顺序分为以下几个工序：扫描、图像处理、自检全检、PDF 处理。

2. 操作人员领取、提交案卷：一个操作人员可以胜任多个工序的工作。启动某个工序时，操作人员首先在系统上批量领取一定数量的加工任务，文件 data.xlsx 中的字段“dUPDATE_TIME”记录了每份案卷的领取时间；档案处理完成后，在系统上进行批量提交，文件 data.xlsx 中的字段“dNODE_TIME”记录了每份案卷的提交时间。当领取的案卷数量较多时，通常会在中午休息前或下午下班前提交已完成的部分案卷。允许操作人员在未完成已领取的任务前领取新任务。

3. 工作效率按批进行计算，将同一批案卷的最后提交时间减去这批案卷的最早领取时间作为该批案卷的总耗时，以此计算该批案卷的平均耗时。所谓“批”是对同一个操作人员在同一个工序中，从领取第一份案卷开始，直到该操作人员在该工序中所有案卷都提交完成，在这段时间内处理的所有案卷。文件 data.xlsx 中的字段“sBatch_number”记录了批的编号。

4. 文件 data.xlsx 中的字段“iNODE_STATUS”（工序状态）为 2，表明案卷已完成并提交，且不需要返工；该字段为 5，表明案卷经过返工，已完成并提交。

5. 工作时间为周一至周六上午 8:30-12:00，下午 13:00-18:00，案卷的处理时间和操作人员的工作时长应去掉非工作时间。注意：实际工作中，可能有提前上岗或推迟下岗的情况。

四、 任务

data.xlsx 记录了某档案数字化加工单位 2020 年 7 月加工处理过程中各个工序的管理数据。请**编程完成**以下任务并撰写报告，**在报告中**详细描述各项任务的**处理思路、过程及必要的结果**。结果的模板文件在文件夹“result”中。

任务 1　数据预处理与统计

任务 1.1　统计完成四道工序的案卷数量，在报告中列出统计结果。汇总各案卷各工序的开始时间及各案卷的完成时长，以表 1 的格式将汇总结果保存到文件“result1_1.xlsx”中，同时在报告中列出案卷完成时长最长的三个案卷的结果。

注 1　每个案卷的完成时长是扫描、图像处理、自检全检三个工序的耗时之和，PDF 处理无需计算耗时，各工序的耗时是该工序的开始时间至结束时间的时长。

注 2　完成时长应去掉非工作时间（“三、案卷加工流程说明”第 5 条），单位：h，保留 3 位小数。

表 1

案卷号	扫描 开始时间	扫描 结束时间	图像处理 开始时间	图像处理 结束时间	自检全检 开始时间	自检全检 结束时间	PDF 处理 开始时间	PDF 处理 结束时间	完成 时长
XXX	2020-07-14 16:56:14	2020-07-15 10:12:47	2020-07-15 14:48:42	2020-07-15 15:58:46	2020-07-22 11:19:04	2020-07-22 15:08:05	2020-07-23 17:24:54	2020-07-23 17:43:37	6.762
……	……	……	……	……	……	……	……	……	……

任务 1.2　统计需要返工的案卷数量及其占完工案卷总数的百分比，在报告中列出结果。汇总返工案卷的返工工序和返工开始时间，以表 2 的格式将汇总结果保存到文件“result1_2.xlsx”中，同时在报告中列出返工案卷号“托 40606-册六”“托 40606-册七”“托 5901_1-册三”的结果。

注　未返工工序的时间为空。

表 2

案卷号	扫描	图像处理	自检全检	PDF 处理
XXX			2020-07-08 15:27:00	
……	……	……	……	……

任务 1.3　对自检全检工序，汇总每个操作人员的返工案卷数，计算其占该操作人员该工序工作总量的百分比，按百分比降序排列，以表 3 的格式将结果保存到文件“result1_3.xlsx”中，同时在报告中列出前三位操作人员的结果。结果保留 3 位小数，例如：返工案卷占比为 1%，在结果表中填写“1.000”。

表 3

操作人员 ID	返工案卷占比 (%)
……	……

任务 1.4　按工序分别统计完成案卷的数量、总耗时和平均耗时，以表 4 的格式将结果保存到文件“result1_4.xlsx”中，并在报告中列出结果。结果保留 3 位小数。

注　按工序计算总耗时，是该工序各个批次的案卷集最早开始时间至案卷集最晚结束时间之和，而不是各个案卷完成时长的总和。

表 4

工序	完成案卷的数量	总耗时 (h)	平均耗时 (h/卷)
扫描	……	……	……
……	……	……	……

任务 1.5　按操作人员、工序统计工作时长、完成案卷的数量和每个案卷的平均耗时（h/卷），以表 5 的格式将结果按操作人员 ID 升序排列保存到文件“result1_5.xlsx”中，同时在正文中列出操作人员 ID“10”“33”“48”的结果。结果保留 3 位小数。

注　按操作人员、工序统计工作时长是按批进行的（“三、案卷加工流程说明”第 3 条），应去除非工作时间（“三、案卷加工流程说明”第 5 条）。

表 5

操作人员 ID	工序	工作时长 (h)	完成案卷的数量	每个案卷的平均耗时 (h/卷)
001	扫描	……	……	……
001	图像处理	……	……	……
002	……	……	……	……

任务 2　数据分析与可视化

任务 2.1　计算并绘制每天不同工序完成案卷数量的簇状柱形图：x 轴表示时间，y 轴表示完成案卷的数量，用不同颜色标记不同工序。

任务 2.2　计算并绘制各工序每天投入工作量（单位：人·小时）的多重折线图：x 轴表示时间，y 轴表示每天投入的工作量，用不同颜色标记不同工序。

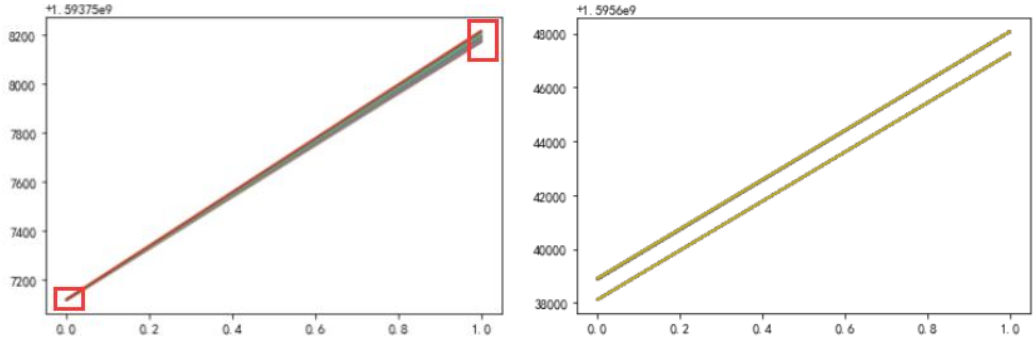
任务 2.3　绘制每天各工序返工案卷数占当天返工案卷总数的百分比堆积面积图：x 轴表示时间，y 轴表示百分比，用不同颜色标记不同工序。

任务 2.4　对图像处理工序，汇总每个操作人员返工案卷数，计算其占该工序返工案卷总数的百分比，并按百分比进行排序，绘制饼图，其中排名第 10 位及以后的合并成一个扇区。

任务 3　领取提交模式分析

档案的数字化加工中，操作人员在某个工序中的正常领取提交模式是在相对集中的时间内领取若干案卷，全部加工完成后在相对集中的时间内按照领取的顺序提交该批案卷。但在现实中会出现更多类型的领取提交模式，例如操作人员有时会分多次领取案卷，处理完后一起集中提交；或在未完全提交已领取案卷的情况下又领取了新的案卷。出现这种情况的可能原因是，在处理已领取的案卷时发现这些案卷的处理难度低于平均难度，操作人员出于提高个人工作效率的考虑通过多次领取的方式“囤积”易处理案卷。

可通过可视化的方法分析案卷领取提交时序，例如在图 1 所示的桑基图中，每份案卷对应图中一条直线，左端点的纵坐标表示领取时间，右端点的纵坐标表示提交时间。



(a) 正常领取提交模式　　(b) 一种非正常领取提交模式

图 1 案卷领取提交时序桑基图

图 1(b)对应于分两次领取、分两次提交，不同的案卷集在处理时序上出现交叉的领取提交模式。也可以使用图 2 中的模式示意图来表达。

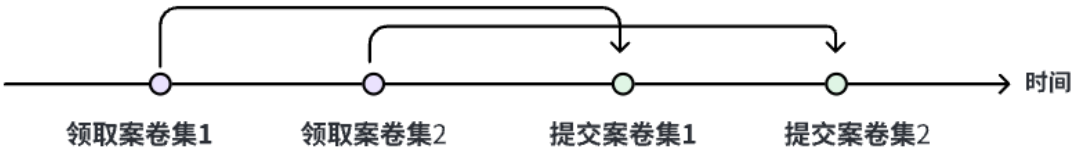


图 2　对应于图 1(b)的领取提交模式示意图

请根据文件 data.xlsx 的批次数据，通过可视化的方法分析批次内案卷的领取提交时序，总结有哪几种领取提交模式。对每一种模式给出一个实际例子，以表 6 的格式保存到文件“result3.xlsx”中，同时在报告中参照图 1 和图 2 的方式分别绘制两种不同的示意图。

表 6

模式序号	操作人员 ID	批号	时间	操作	案卷号
1	33	500	2020/7/22 16:27:48	领取	托 693976-册一
1	33	500	2020/7/23 08:20:17	提交	托 694000-册一
1	33	501	2020/7/23 08:37:36	领取	托 695826_2-册一
1	33	501	2020/7/23 10:40:07	提交	托 695850-册一
……	……	……	……	……	……

附录　数据说明

数据文件 data.xlsx 中的每条记录对应于一个案卷某个工序的处理记录，其中的字段名及其含义如表 7 所示。

表 7

字段名	含义	取值说明
iID	记录 ID（主键）	
iPID	主 ID	
uFILE_FLAG	案卷标识（GUID）	
sARCH_ID	案卷号	
sFLOW_NAME	工作流名称	
sNODE_NAME	工序节点名称	
iNODE_STATUS	工序状态	2：案卷已完成，且已提交； 5：案卷需要返工，且已提交
iUSER_ID	操作人员 ID	
iWF_ID	工作流 ID	
iWN_ID	流程节点 ID	12：扫描；13：图像处理；22：自检全检； 15：PDF 处理
sPIC_PATH	图片路径	
iFLOW_NODE_NO	工序号	1：扫描；2：图像处理；3：自检全检； 4：PDF 处理
iPROC_USERID	返工操作人员 ID	
sPIC_SERVER_PATH	图片路径	
sPDF_SERVER_PATH	PDF 路径	
iARCH_TYPE	案卷类型	
sORDER_ARCH_ID	排序档号	
dUPDATE_TIME	工序开始时间	该工序节点开始时间
dNODE_TIME	工序结束时间	该工序节点结束时间
dPROC_TIME	返工时间	返工开始时间
sBatch_number	批次编号	