程序设计实训 2023 大作业：交互式医学数据分析平台

助教：凌精望

邮箱：lingjw20@mails.tsinghua.edu.cn

作业目标

本次大作业的目标是运用 C++和 Qt 编程技能，实现一个交互式医学数据分析平台。医学数据分析在现代医疗领域具有重要意义，它能够帮助医生们深入探究患者的病情和健康状况，根据以往患者的数据推断影响病情的因素，从而做出更准确的诊断和治疗方案。在这个作业中，你将学会如何运用面向对象编程的理念，结合 Qt 框架的优势，创造出一个功能强大、界面美观的医学数据分析平台。

医学数据分析是将现代医学中积累的海量数据与计算机技术相结合的一门关键领域。在这个作业中，我们将关注一个人是否患上某种特定疾病，以及与该疾病相关的各种指标，如生化数据、体征等。通过对这些医学数据的分析，我们可以揭示患病与各种生理指标、生活方式等因素之间的关系。这种分析有助于确定疾病的发生原因，深入了解潜在的健康风险，并为医生提供更精确的诊断和治疗建议。

你需要实现一个交互式医学数据分析平台，满足以下要求：

- 数据导入与存储：允许用户将医学数据导入系统中，并支持数据的存储和管理。
- 数据可视化：利用 Qt 框架的优势，实现多种数据可视化方式，如折线图、散点图等，以清晰展示不同指标之间的关系。
- 数据分析：提供数据分析算法，帮助用户发现患病与各种指标之间的模式与关联，实现归因分析。
- 用户交互：设计友好的用户界面，使医生能够轻松导入数据、进行可视化和分析，并直观地理解数据结果。
- 界面美观与易用性：参考已有数据可视化软件和界面设计准则，确保界面美观程度和用户体验，提升医生使用平台的满意度。

我们在课上介绍的 Qt 框架非常适合这类可视化分析应用。许多广泛使用的数据可视化和分析软件（Visualization Toolkit (VTK)、ParaView、LabPlot 等）都是基于 Qt 框架实现的。我们的可视化分析任务更简单一些，主要围绕 Qt Widgets、Qt Charts、Qt Data Visualization 模块。助教会提供的相关算法代码，也鼓励引入开源项目实现统计分析、机器学习等算法。此

外，你还可以查阅相关软件的视频演示、文档和截图，从中汲取灵感，以确保你的医学数据分析平台在功能性、可用性和界面美观程度方面达到最佳水平。

通过完成这个作业，你将不仅掌握了面向对象编程和 Qt 框架的应用，还为医学数据分析领域的进展贡献了一份有现实意义的项目。你的平台有望为医生们提供更深入的数据洞察，从而为医疗决策提供更可靠的支持。祝你在这个有意义的项目中取得成功！

数据

文件来源：

本次作业提供的医疗数据来源为 kaggle 竞赛网站，感兴趣的同学可以通过以下链接详细了解：

https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset

程序需要处理的输入即作业文件中的 breast-cancer.csv。

文件说明：

Breast Cancer（乳腺癌）数据集是一个常用的医学数据集，用于乳腺癌的诊断和预测研究。该数据集收集了乳腺肿块的一组特征，包括临床观察结果和细胞核的形态学特征，以帮助医生确定肿块的良性（benign）或恶性（malignant）。数据集来自 kaggle，包含 569 个样本。

数据中 diagnosis 列为类别，B 代表良性，M 代表恶性。其余的列为特征，包括以下特征：半径（radius）、纹理（texture）、周长（perimeter）、面积（area）、光滑度（smoothness）、紧密度（compactness）、对称性（symmetry）、分形维度（fractal dimension）。

功能需求

助教在 “附录：算法示例代码” 中提供了算法代码及其相关说明。在完成作业时，同学们可以将注意力集中在 Qt 相关代码的编写和已有代码的调用上。建议大家可以参考作业要求中提供的 “提示” 小节中的信息，同时也鼓励尝试采用与 “附录” 和 “提示” 不同的全新方法来实现作业所需的功能。

需求点一：直方图 (20%)

能够选择一列数据，计算均值方差和直方图。

具体要求

- 能够导入数据，在 Qt 图形界面中以类似 Excel 表格的形式呈现 (4%)。
- 数据分组：对于离散非数值变量，可以自动映射为数值，比如将"A", "B", "C"映射为 0, 1,

2，将"B"和"M"映射为 0 和 1 等，从而可以进行后续计算 (2%)。

- 选择一列数据，计算该列数据的均值方差。需要提供一个展示框，以及一个能够点击展示的按键，按键允许点击后在展示框内显示该列的均值和方差 (4%)。
- 要求能够选择绘制直方图，需要规划一个板块进行展示。注意离散数据和连续数据区分，自动计算合适区间大小或者允许用户编辑的数据分段区间大小，保证数据直方图能够给到有效的分布信息（如果所有数据在一个区间，则无法给到有效信息） (5%)。
- 拟合数据的正态分布曲线，并规划一个版块进行绘制。实现一个按钮，点击后可以展示正态分布曲线（推荐和直方图叠加显示） (5%)。

提示

- 算法实现可参考《附录：算法示例代码》中均值和方差的计算一节。
- 类似 Excel 的表格：QTableWidget，参考 https://doc.qt.io/qt-6/qtwidgets-itemviews-spreadsheet-example.html
- CSV 文件导入：QTextStream
- 直方图：可以基于柱状图修改 https://doc.qt.io/qt-6/qtcharts-barchart-example.html
- 绘制正态分布曲线：计算不同 x 坐标下的函数值，绘制插值曲线，参考 https://doc.qt.io/qt-6/qtcharts-overview.html#line-and-spline-charts
- 在同一个图中显示正态分布曲线和直方图：参考 https://doc.qt.io/qt-6/qtcharts-lineandbar-example.html

需求点二：散点图/曲线拟合 (25%)

选择两列数据，一列作为横轴，一列作为纵轴，绘制散点图；对散点进行曲线拟合，绘制曲线图，并计算两列数据的相关性。

具体要求

- 能够绘制散点图 (4%)。
- 能够绘制曲线图 (6%)。
- 支持设置拟合多项式次数： 目前我们已经给出能够拟合任意次多项式的代码，希望同学实现能够根据设置的多项式次数，完成曲线绘制 (4%)。
- 绘制散点图时，散点图给出横轴名称，纵轴名称 (3%)。
- 部分数据可能会有多点重合的现象，如果点有重叠， 给出重叠点数量，在一个合适的地方显示 (2%)。
- （可选）鼠标悬停至点上方，可见该点对应横纵坐标值 (额外 5%)。

- 选出两列，计算拟合曲线的 p 值和 r^2，并在醒目的输出窗口进行显示 (6%)。

提示

- 算法实现可参考《附录：算法示例代码》中曲线拟合（最小二乘法）的计算一节。
- 散点图：https://doc.qt.io/qt-6/qtcharts-scatterchart-example.html
- 折线图：https://doc.qt.io/qt-6/qtcharts-linechart-example.html
- 曲线图：和折线图类似，参考 https://doc.qt.io/qt-6/qtcharts-overview.html#line-and-spline-charts
- p 值，r^2可参考：https://zhuanlan.zhihu.com/p/110886609

需求点三：相关性和协方差矩阵 (20%)

协方差衡量了两个变量之间的线性关系程度，帮助我们理解它们如何随着彼此的变化而变化，从而揭示变量之间的相互影响和趋势。

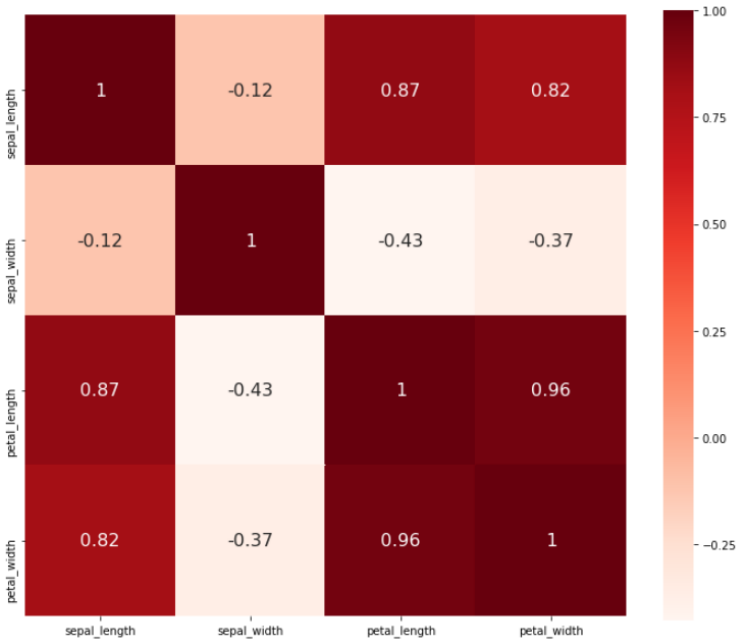
能够选择多列计算皮尔逊系数和协方差矩阵，绘制结果热图。

具体要求

- 选出多列，计算选中列之间的协方差，通过热图方式绘制出来 (4%)。
- 选出多列，计算选中列之间的相关性系数，通过热图方式绘制出来。推荐用一个按钮切换协方差/相关系数显示 (4%)。
- 热图需要有颜色和数值显示 (4%)。
- colorbar 需要有刻度显示 (4%)。
- 排版美观。需要显示列名称 (4%)。

提示

- 算法实现可参考《附录：算法示例代码》中**协方差矩阵和相关系数矩阵**的计算一节。
- 协方差矩阵：QTableWidget
- Color gradient 参考：https://doc.qt.io/qt-5/qtdatavisualization-surface-example.html#
- 热图：可以考虑直接修改 QTableWidgetItem 的背景颜色，也可以 QPainter 自己画。
- 协方差矩阵参考样式：



需求点四：降维绘制 (15%)

降维可视化的作用是通过减少数据的维度，将复杂的高维数据转化为更易理解和展示的低维表示，帮助发现数据中的模式和关系。

作业要求能够选择不同病例的特定列，进行降维分析，绘制到平面图，3D 界面，降维采用 PCA（主成分分析）方法。

具体要求

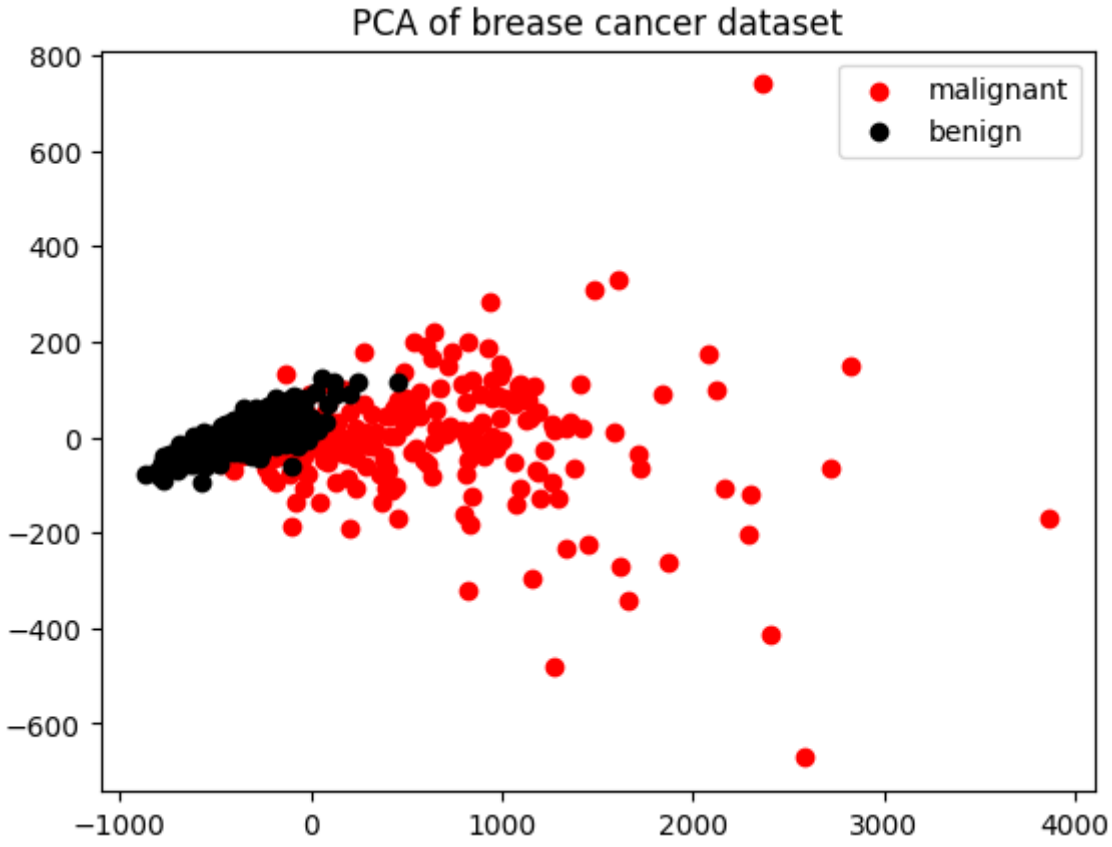
- 能够支持选择任意多列数据，利用 pca 降维方法后，将每一行数据作为一个点绘制到散点图上，如果降维到 2 维，则绘制平面散点图 (5%)；
- 如果降维到 3 维，则绘制立体散点图 (5%)。
- 能够根据类别信息（"B"和"M"）,将散点图中的散点赋予不同的颜色进行展示 (5%)。
- （选做）在降维可视化界面中鼠标悬停、或者点击能够显示详细属性信息 (额外 5%)。

提示

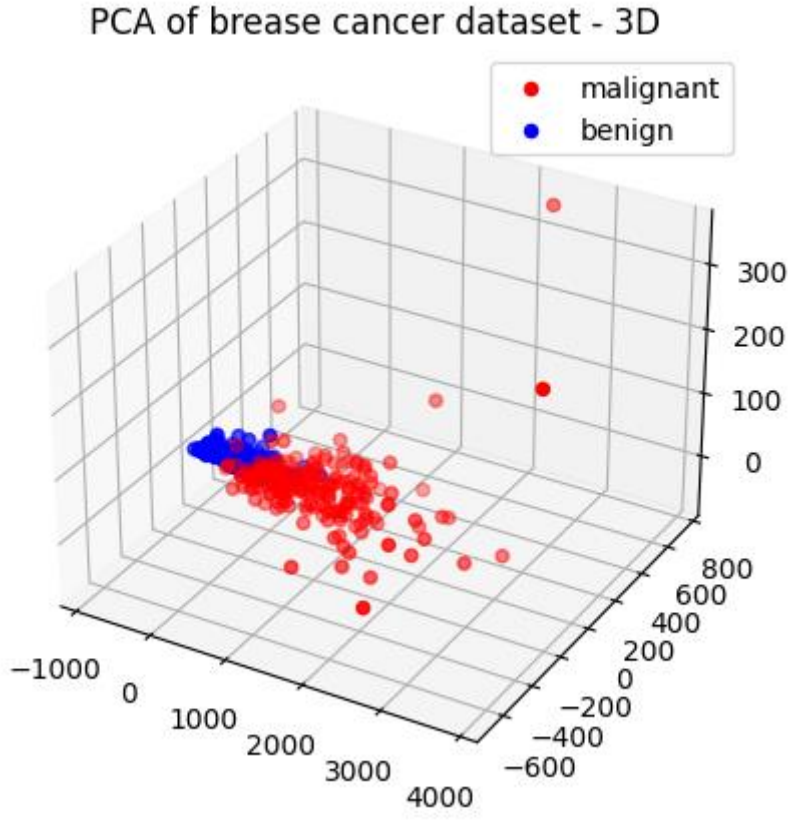
- 算法实现可参考《附录：算法示例代码》中 **PCA 降维算法**一节。
- 二维降维分析可视化：还是参考散点图 scatter chart
- 三维降维分析可视化：考虑 https://doc.qt.io/qt-6/qtdatavisualization-overview.html#3d-scatter-graphs
- PCA 介绍：https://en.wikipedia.org/wiki/Principal_component_analysis；https://zhuanlan.zhihu.com/p/37810506?utm_id=0

参考样例：

平面图：


PCA of brease cancer dataset
malignant
benign
800
600
400
200
0
−200
−400
−600
−1000
0
1000
2000
3000
4000

3D 图：

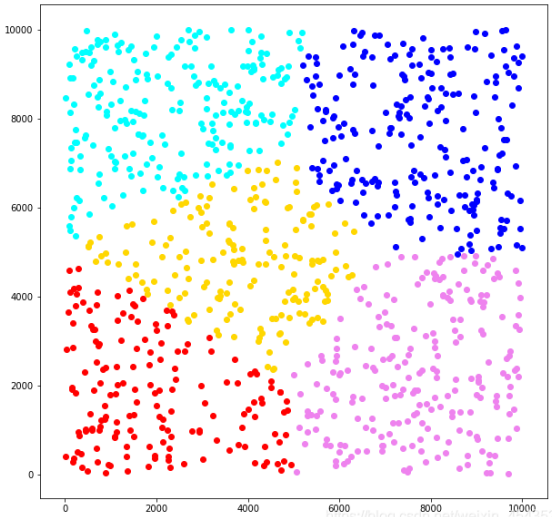

PCA of brease cancer dataset - 3D
malignant
benign
300
200
100
0
800
600
400
200
0
−200
−400
−600
−1000
0
1000
2000
3000
4000

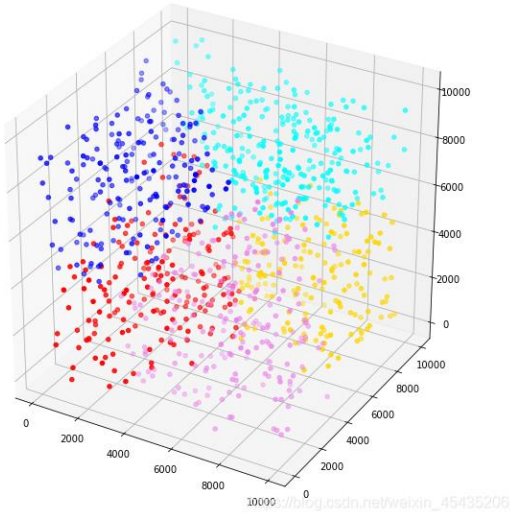
需求点五：聚类分析 (20%)

聚类分析是指将数据对象的集合**分组**为由类似的对象组成的多个**类别**的分析过程。实际使用中，我们常用聚类分析发现无标注数据中一些潜在的分类信息。

作业要求能够选择不同病例的某些列，进行聚类分析，聚类算法至少实现一种，实现多种可以加分。聚类完成后结果降维绘制到平面图和 3D 界面。

聚类结果绘制样例：





具体要求

- 至少实现 **KMeans 聚类（下面将给出）**(5%)。
- 实现其他聚类算法为加分项 (额外 5%每个，最多额外 3 个)。
- 完成聚类后能够给每一行数据分配聚类类别，新增一列，列名为聚类算法名称，用于将每一行数据进行聚类后的归类 (5%)。
- 能够实现每行按照聚类结果将同一种类别的行给到同一颜色，并实现通过按钮控制颜色是否显示 (5%)。
- 聚类完成后，对于（四）中降维后生成的 2D、3D 可视化结果，也需要根据聚类生成的类别对散点赋予不同的颜色，方便医生观察数据分布。推荐将需求点（四）中根据真实类别划分颜色的散点同时显示，方便对比实际标签和聚类算法结果 (5%)。

提示

- 关于聚类分析的更多介绍，可参考 https://zhuanlan.zhihu.com/p/139924042
- 其他聚类算法可选 dbscan、漂移聚类、谱聚类，可参考：DBSACN（https://zhuanlan.zhihu.com/p/185623849）、漂移聚类（https://zhuanlan.zhihu.com/p/580994780）、谱聚类

（https://zhuanlan.zhihu.com/p/392736238）

- 算法实现可参考《附录：算法示例代码》中 **KMeans 聚类算法**一节。

可选需求点：机器学习模型归因分析 (额外 20%)

xgboost 作为常用的**机器学习模型**，可以根据医学表征实现病种的良恶性分类，根据学习结果可以观察特征重要性，本节需要实现常见机器学习算法 xgboost 的集成，并能够可视化归因分析结果。

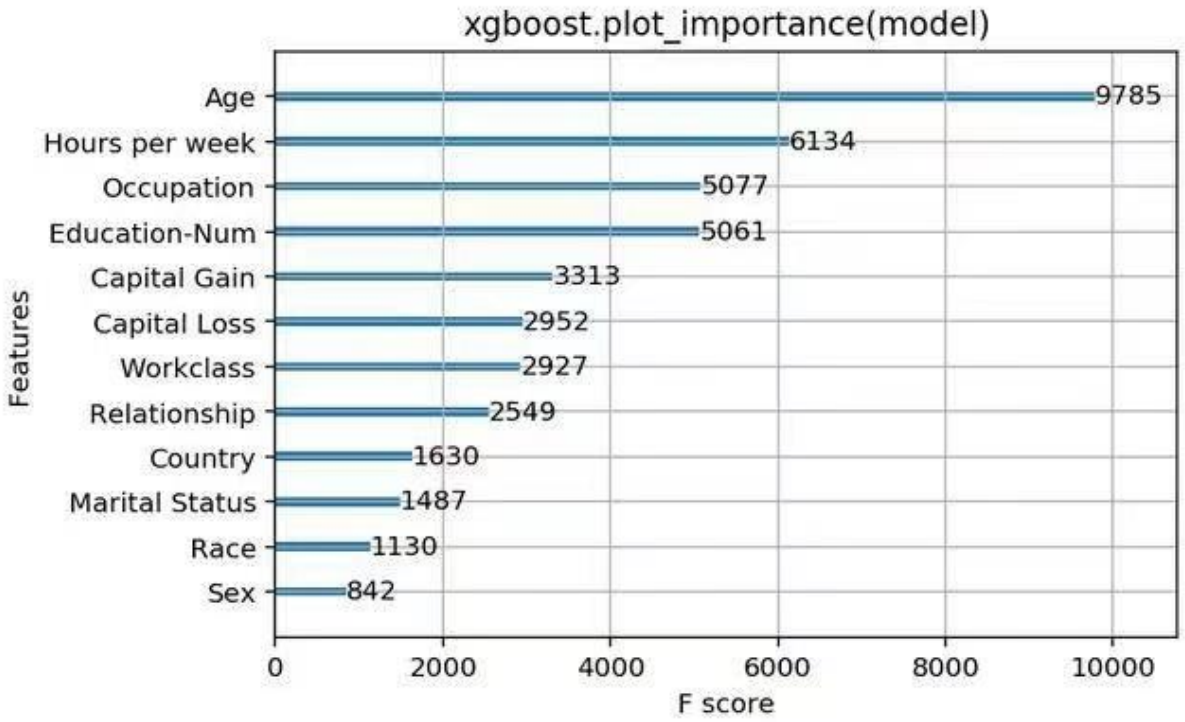
实现目标：针对一个可以分类的问题，划分训练集、测试集，进行机器学习模型的训练和预测。助教提供了 xgboost 相关说明，也可以任选一个库。

具体要求

- 为机器学习部分新建一个界面，使用原表格中的数据，按照固定比例随机划分成为训练集和测试集。在训练集上训练，预测测试集上数据的良恶性分类并计算结果指标。功能点需要包括：
  - 能够给到按照 10%/90%，20%/80%，30%/70%，50%/50%比例划分的测试集/训练集划分选项，通过设置随机种子进行随机数据划分 (额外 3%)。
  - 可视化测试集数据的预测结果 (额外 3%)。
  - 对应每一个表格规划合适的区域，能够显示训练后的模型，在训练集和测试集上测试结果指标，指标包括 AUC、F1Score (额外 4%)。
- 在训练集和测试集划分的表格中，能够实现新增一列，体现在训练集和预测的 B 和 M 结果 (额外 3%)。
- 支持简单的参数调整，具体可以根据附件以及简介中的 xgboost 接口设置，至少能够调整的参数包括：训练迭代次数，树最大深度 (额外 2%)。
- 支持特征贡献度的可视化，特征贡献度按照表格给出，横坐标是贡献度大小，纵坐标为每一个特征名，按照由大到小排序后展示在图上，可参考下面给出的示例图 (额外 5%)。

提示

- AUC、F1Score，参考：（https://zhuanlan.zhihu.com/p/404453388AUC）、（https://zhuanlan.zhihu.com/p/595712347F1Score）
- 算法实现可参考《附录：算法示例代码》中 **xgboost** 一节。
- 对于可视化 xgboost 中所选特征贡献度大小，可参考以下样例：



提交说明

代码要求

- 使用 C++、Qt 完成（不能使用 Python 和 PyQt 完成。学习 Python 会增加额外的课程负担，并且使用 Python 不容易融入课程社区）
- 基于面向对象编程思想
- 要求使用相对路径访问工程资源
- 代码风格统一
- 包含必要注释，代码具有可读性
- 具有良好的运行效率

文档要求

文档非常重要，至少要包括以下内容：

- 模块之间逻辑关系
- 程序运行流程
- 完成了哪些要求？（可以用运行截图说明）
- 参考文献、引用代码出处（使用了哪些库）

提交格式

以 学号_姓名 为名称的压缩包提交，压缩包内应包含三个文件夹：

- src，包含完整的工程文件，要求工程可以直接编译（注意使用相对路径配置代码以及资源文件），并清理编译过程中产生的中间文件
- doc，包含你的作业文档
  - 需要提交一份**大作业自评表**（见作业压缩包），如实填写对各个功能点的完成情况（助教会检查自评表的真实性，如果没有提交自评表或者自评表不真实，分数为实际卷面分数*0.6）。
- bin，包含可以在 Windows 10 64 位下可以运行的可执行文件

提交截止时间

2023 年 9 月 17 日（周日）23:59

提交说明：

- 迟交作业的分数按每天 10%的递减。
- 为避免因迟交带来更大的损失，建议大家提前上传一个版本到网络学堂上。

课程汇报

- **汇报日期**：**2023 年 9月15日（周五）**
- 汇报顺序：助教会提前在网络学堂公布《2023 年程序设计实训大作业汇报时间安排》，如果分配时间无法汇报，可自行找同学交换时间，交换之后告知助教。如果找不到同学交换，可直接联系助教申请**提前**答辩，**不能申请延后答辩**。
- 汇报材料：
  - **自己的电脑设备**，不需准备转接头
  - 程序**源代码**：通过现场运行程序演示的方式来展现实现的功能。
- **汇报时间：6 分钟以内**
- **线下答辩地点**：四教 4202

※请按时到达教室，并提前将设备准备好。**每一组同学请于自己所在时间段的起始时间到达教室**，如 A 同学被分到 8-9 点这一时间段进行汇报，则无论他在该时间段内的汇报顺序如何，都应该于 8 点或之前到达教室并做好相关准备。

评分细则

课程汇报占大作业成绩中的 20%，汇报成绩主要从以下三点进行评估，最终三项分数取平均得到最终的汇报成绩。

- 表达（10 分）
  - 要求能够结合实时程序演示流畅地表述大作业的完成情况，演示时间把控在 4-5 分钟，能够理解老师和助教的提问并且给予相应的回答。
- 功能（10 分）
  - 要求能够实时演示展示出程序完整的流程以及实现的功能点。
- 界面（10 分）
  - 要求界面美观不混乱、用户交互友好。

（可选参加）中期技术分享

为了帮助同学们更好地掌握项目规模，激发模块编写的思路，以及鼓励那些进度较快的同学分享他们的实现经验，我们将在本课程中举行一场中期技术分享。这个分享环节旨在为大家提供一个互相学习和交流的平台，同时也为那些在项目中取得较快进展的同学提供展示和分享的机会。

- 时间与地点： **2023 年 9 月 8 日（周五）下午 13:30~16:00**，地点为三教 3300（上课教室）。
- 分享内容： 我们计划邀请 6 位进度较快的同学作为讲者，每位同学将有 10 分钟的时间进行界面演示和 PPT 讲解，展示他们的设计思路和实现心得。之后，将有 10 分钟的答疑环节，其他同学可以提出问题或寻求进一步的解释。
- 技术顾问： 我们还将邀请两位有医学项目合作经验的研究生同学作为技术顾问。他们将评价各位讲者当前进展的功能和可用性，并提出进一步改进的建议。
- 评选规则： 讲者完成分享后，参与同学和技术顾问可以按照技术高度、收获大小等要素进行评分。技术顾问的评分占 20%，同学的评分占 80%。根据评分，第一至第三名同学将获得 500 元额度的电子产品奖品，而第四至第六名同学将获得 200 元额度的电子产品奖品。
- 报名方式： 同学们可以在 2023 年 9 月 7 日（周四）下午 17:00~22:00 之间，将分享提纲（不超过 300 字）发给助教凌精望完成报名。如果报名人数超过 6 人，我们将根据分享提纲估计的完成度来选择 6 位讲者。被选中的 6 位讲者的分享顺序将会随机排列。

完成作业所需的能力建议

- 充分利用面向对象编程，深刻理解面向对象编程的概念，以便在你的项目中实现代码复用。将常见的功能封装成类，以便在多处地方使用，从而提高代码的可维护性和效率。项目中，你会成为自己代码的用户，合理的面向对象可以帮助后续功能的开发。
- 官方英语文档自学： 在解决技术问题和学习新知识时，英语文档是重要的资源。推荐阅读 Qt 官方文档以了解 Qt 框架提供的功能和特性，这将有助于你获得设计思路和解决方案。借助划词翻译浏览器插件，可以更轻松地理解英文文档。推荐阅读：
  - Qt 图表总览：https://doc.qt.io/qt-6/qtcharts-overview.html
  - Qt 数据可视化模块总览：https://doc.qt.io/qt-6/qtdatavisualization-overview.html
- 充分利用搜索引擎收集信息，如 Google、Bing 等。
- 在项目中遇到困难时，你可以与同学们保证学术诚信下分享自己的思路。通过交流思路，你可能会得到新的见解和解决方案。
- 参与中期技术分享。不要错过中期技术分享的机会。这是一个展示你的进展、分享你的实现经验的平台。通过分享和答疑，你将更深入地理解你的项目，并从其他同学和技术顾问的反馈中受益。