Final Project Proposal

Stella Koo, Yixin Zheng, Thomas Tang, Yonghao YU,

2024-10-31

Group members (names and UNIs)

Stella Koo (bk2959), Yixin Zheng (yz4993), Thomas Tang (tt3022), Yonghao YU (yy3564)

Project Title

The Influence of Diagnostic Attributes on Heart Disease Across Regions (with Varying on Development Levels or Climate Groups?)

- Challenge: Climate Data: To analyze by climate, we would need climate classification data or average environmental conditions (temperature, humidity) that can be associated with each region. While this is technically possible, finding or linking appropriate climate data for each location in the heart disease dataset may require additional data collection and preprocessing.

Development Level Data: Development level, often assessed using indicators like GDP per capita would similarly require obtaining additional socioeconomic data specific to each region. Matching these indicators to the heart disease dataset's geographic locations would add complexity and require control for multiple factors.

Alternative Title/Questions and Analysis Ideas: 1. Comparing Diagnostic Factors for Heart Disease Across Regions: explore whether certain diagnostic factors (e.g., cholesterol levels, exercise-induced angina) are more predictive of heart disease in one region compared to others. 2. The Influence of Age and Lifestyle Factors on Heart Disease Across International Datasets: narrows down the analysis to age and lifestyle-related variables (e.g., chest pain type, cholesterol, exercise-induced angina). 3. Regional Patterns in Heart Disease Diagnostic Attributes: focusing on comparing the prevalence and distribution of key diagnostic attributes (like age, cholesterol, etc.) across the four datasets. 4. Examining Predictive Power of Clinical Indicators for Heart Disease in Diverse Populations: predictive accuracy of clinical indicators (like cholesterol, blood pressure) for heart disease across various demographics.

Motivation (Ian)

Intended Final Products (Yixin)

1. Website: A user-friendly website to display our findings with interactive features, such as a density map and a R-SHINY tool to predict individual heart disease risk. The website will include things like following:

- Homepage: Introduction to the project, motivation, data sources, and some questions.
- Data Exploration: Interactive visualizations of key diagnostic factors with options to filter by heart disease status and region.

- Comparative Analysis: Visualizations comparing the significance and distribution of heart disease predictors across regions, including correlation heatmaps and scatter plots etc.
- Model Insights: Diagnostic plots to assess model assumptions. Summary of model results, showcasing the most predictive attributes in each region.
- Conclusion: Summary and takeaway of regional differences in heart disease diagnostics.
- Tool: Using R's Shiny to allow for interactive exploration.

2. Report: A comprehensive report within the website to display our project steps and findings. The report will cover things like: Introduction, Data Processing and Cleaning, Exploratory Data Analysis (EDA), Modeling, Results, Conclusion and Limitations etc.
3. Video: A brief video walkthrough to introduce our project, explain our motivations, and highlight key findings. This screencast will guide viewers through the website and its interactive features.
4. Code: A link to our organized GitHub repository will be available on the website, containing all data cleaning, analysis, and visualization code. The repository will include a README file for clear instructions on reproducing our analysis.

Data Sources (Stella)

The Heart Disease datasets were obtained from UC Irvine's Machine Learning Repository that can be accessed here. This directory contains four datasets focused on heart disease diagnosis, each representing a distinct geographic location and with attributes recorded as numeric values. The data was gathered from the following four locations:

```
1. Cleveland Clinic Foundation (cleveland.data)
2. Hungarian Institute of Cardiology, Budapest (hungarian.data)
3. V.A. Medical Center, Long Beach, CA (long-beach-va.data)
4. University Hospital, Zurich, Switzerland (switzerland.data)
```

Although the original datasets contain 76 raw attributes, the source provides processed datasets with 14 carefully selected variables that have been widely utilized and cited in numerous research studies. This project will focus on these processed datasets, which include the following attributes:

```
1. age: age in years
2. sex: sex (1 = male; 0 = female)
3. cp: chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic
4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
5. chol: serum cholestoral in mg/dl
6. fbs: fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7. restecg: resting electrocardiographic results (0 = normal; 1 = having ST-T wave abnormality; 2 = sho
8. thalach: maximum heart rate achieved
9. exang: exercise induced angina (1 = yes; 0 = no)
10. oldpeak: T depression induced by exercise relative to rest
11. slope: slope of the peak exercise ST segment (1 = upsloping; 2 = flat, 3 = downsloping)
12. ca: number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
14. num: diagnosis of heart disease (0 = < 50% diameter narrowing; 1 = > 50% diameter narrowing)
```

While our goal is to explore all the variables listed above, some datasets contain a significant amount of missing data for certain attributes. The approach we will take to address these gaps is still under consideration.

Planned analyses, Visualizations, Coding Challenges (Thomas, Yonghao)

- Exploratory Data Analysis (EDA):
  - Graphic Displays: Use histograms, density plots, and box plots to visualize distributions of key variables, such as cholesterol, age, thalach (maximum heart rate), and compare distributions between those with and without heart disease.
  - Correlation Analysis: Create a correlation heatmap to identify relationships between variables and detect multicollinearity, especially for predictors that may be used in regression models.
- Linear Regression Model:
  - Model Goal: Assess the relationship between continuous clinical features (e.g., cholesterol, age, trestbps) and heart disease risk scores (or likelihood of heart disease as defined by num).
  - Diagnostics and Assumptions:
    - Check assumptions of linear regression (normality, homoscedasticity, independence) to validate model appropriateness. (For example, plot residuals to check for constant variance and use variance inflation factor (VIF) to check for multicollinearity.)
    - Analysis of our model: Perform hypothesis testing on model coefficients to evaluate each predictor's impact on the dependent variable and evaluate the R-squared and adjusted R-squared values to determine how well the model explains variance in heart disease risk scores.
    - QQ Plot: Use a QQ plot to visually inspect normality of residuals, which supports the model's reliability.
- Comparative Approach:
  - Combine Cleveland and Long Beach Data and do model comparison: Since both datasets are from the U.S., merge these two datasets to assess if a larger combined sample improves model performance compared to analyzing each separately, then compare predictive power and variable significance across regions to see if combining datasets yields better insights or predictive accuracy.
- Visualizations
  1. Correlation Heatmap: * Variables to Include: Age, cholesterol (chol), resting blood pressure (trestbps), maximum heart rate achieved (thalach), ST depression induced by exercise (oldpeak), and number of major vessels colored by fluoroscopy (ca). * Purpose: This heatmap will help identify the relationships between these key clinical variables, guiding decisions on which variables to include in predictive models and flagging any multicollinearity issues.
  2. Interactive Variables Visualization: Primary Variables:
     - Age vs. Maximum Heart Rate (thalach): Use this to explore how age relates to heart rate, and color points based on heart disease status (num) for additional insight.
     - Age vs. ST Depression (oldpeak): This can reveal if age affects exercise-induced ST depression differently for those with and without heart disease.
     - Purpose: Scatter plots provide an interactive way to explore relationships between continuous variables, making it easier to detect patterns by heart disease status.
  3. QQ Plot and Residual Plots: Diagnostic plots, including a QQ plot, will assess model assumptions. Residual vs. fitted value plots can further validate if linear model assumptions hold.
  4. Histograms or Density Plots for Distribution Insights
     - Variables:Age, Cholesterol (chol), num
     - Purpose: These plots allow you to assess the general distribution of variables and identify any outliers or unusual patterns that could impact model accuracy.
- Coding Challenges:
  - Convert '?' to NaNs, handle missing values through imputation or deletion, and ensure consistency in variable types across datasets, especially if merging Cleveland and Long Beach data.
  - Create age groups, examine exercise-induced angina, and introduce other potential interaction terms (e.g., age and cholesterol) that might enhance model predictive accuracy.
  -

Explore different metrics (e.g., R-squared, adjusted R-squared) to compare models across datasets and regions, highlighting which approach (e.g., separate vs. merged datasets) performs best.

  - Linear Regression Model
  - Diagnostics: Check assumptions for linear regression
  - Address hypothesis, R squared value
  - QQ Plot

- Correlation Heatmap
- Interactive Variables
- Exploratory Analysis (Graphic Displays) Try different approaches (maybe merge Cleveland and Long Beach, both same country) and check which approach works better.

Planned Timeline (Stella)

November 11th to 17th:

- Meet with TA to address project-related questions and finalize the project title.
- Hold a group meeting to assign tasks and decide on the structure and sections of the website.
- Begin data cleaning and initiate the planned analysis.

November 19th to 24th:

- Finalize the planned analysis and start creating visualizations.
- Draft the report as the analysis progresses.
- Develop the website's core structure in R.
- Conduct a group meeting to review progress and clarify any outstanding issues.

November 25th to December 1st:

- Complete any remaining analysis and finalize visualizations.
- Integrate all components into the website.
- Hold a group meeting to create and review the screencast (explanatory video).

December 2nd to 7th:

- Finalize the report and website, making any necessary revisions.
- Conduct peer assessment and submit final project.