Project Part 2: Relating Research Questions, Data, and Figures

Section 1: Data and Research Questions

The dataset used for this project comes from Liang et al. (2018), published in *Scientific Data*. The study created a dataset that links short photoplethysmogram (PPG) recordings with cuff-measured blood pressure readings. A photoplethysmogram measures changes in blood flow using light, which allows researchers to analyze pulse-wave patterns that reflect systolic and diastolic blood pressure measured by a traditional cuff. This study is important because it provides the type of data needed to evaluate whether PPG could eventually serve as an accurate, easy-to-use method for cuff-less blood pressure monitoring. Once this technology is fully developed, it could be included in devices such as smartwatches to give users continuous blood pressure insights. Data was collected from 219 adults between the ages of 21 and 86 at Guilin People's Hospital in China, including both healthy participants and individuals with hypertension or diabetes.

Each subject rested for ten minutes before measurements began. The researchers then recorded three short PPG signals, each about two seconds long, while simultaneously taking a standard cuff blood pressure reading. This produced matched pairs of PPG waveforms and blood pressure values. Additional variables included age, sex, heart rate, body mass index, and general health information.

Features

The dataset contains fourteen main features describing each participant: age, sex, height, weight, body mass index (BMI), heart rate, systolic blood pressure, diastolic blood pressure, hypertension category, diabetes status, history of cerebral infarction, history of cerebrovascular disease, and a subject ID number that links each participant to their waveform files. These features provide demographic, clinical, and physiological information that helps interpret each participant's PPG recordings.

Observations

The observations in this dataset are the individual adult participants. There are 219 observations total. Each observation includes one cuff blood pressure reading and three short PPG waveform recordings collected during the same measurement window.

Research Question

The authors' research question was whether short PPG recordings could accurately estimate systolic and diastolic blood pressure, making continuous cuff-less monitoring possible. Since each waveform is directly paired with an actual blood pressure reading, the dataset allows this question to be tested precisely. Another possible question is whether factors such as age or hypertension status influence how PPG features relate to blood pressure.

Section 2: Description of the Data Files

The dataset was downloaded directly from the Figshare link provided in the article. The download contained three main items: the PPG-BP Database folder, the published paper as a PDF, and a small Excel file labeled “Table 1.” Inside the database folder is another subfolder named “Data File,” which includes an Excel sheet called PPG-BP dataset and a folder named 0_subject.

The metadata Excel sheet contains 219 rows and 11 columns, with each row representing a participant. The columns include demographic information, clinical blood pressure category, heart rate, and the cuff-measured systolic and diastolic blood pressure values. The 0_subject folder contains hundreds of small text files, each representing an individual PPG waveform segment labeled by subject number and recording (for example, “2_1,” “2_2,” “2_3”). These files contain the raw PPG signal data used in the study.

Together, these files form a complete dataset with both summary information and detailed waveform recordings. The dataset structure allows each PPG segment to be easily linked back to participant characteristics and their corresponding blood pressure measurement.

Section 3: Visualizations and Data Summaries

Figure 1: Measurement Protocol.

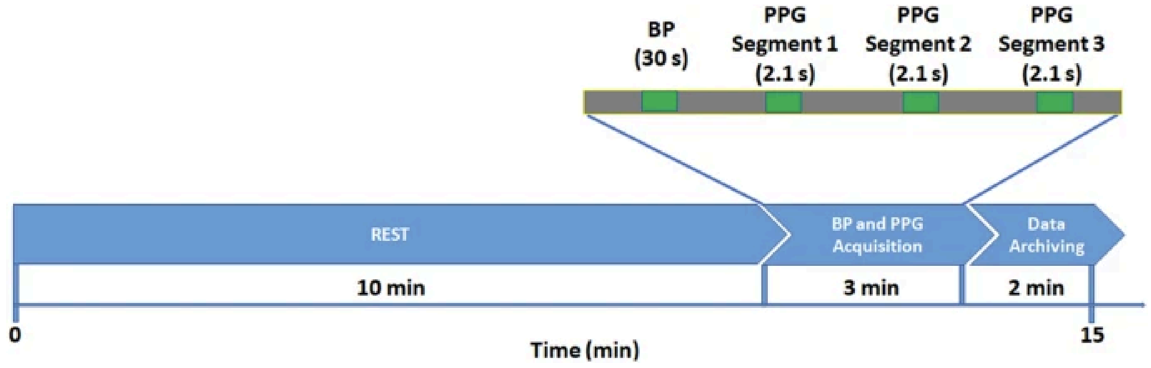


Figure 1: Measurement Protocol
This figure shows the timing of each stage in the experiment. Participants rested for ten minutes, followed by three minutes of PPG and blood pressure recording, and then two minutes of data archiving. This demonstrates that the procedure was standardized and controlled, which helps ensure consistent data quality.

Figure 2: Statistics about the PPG-BP dataset.

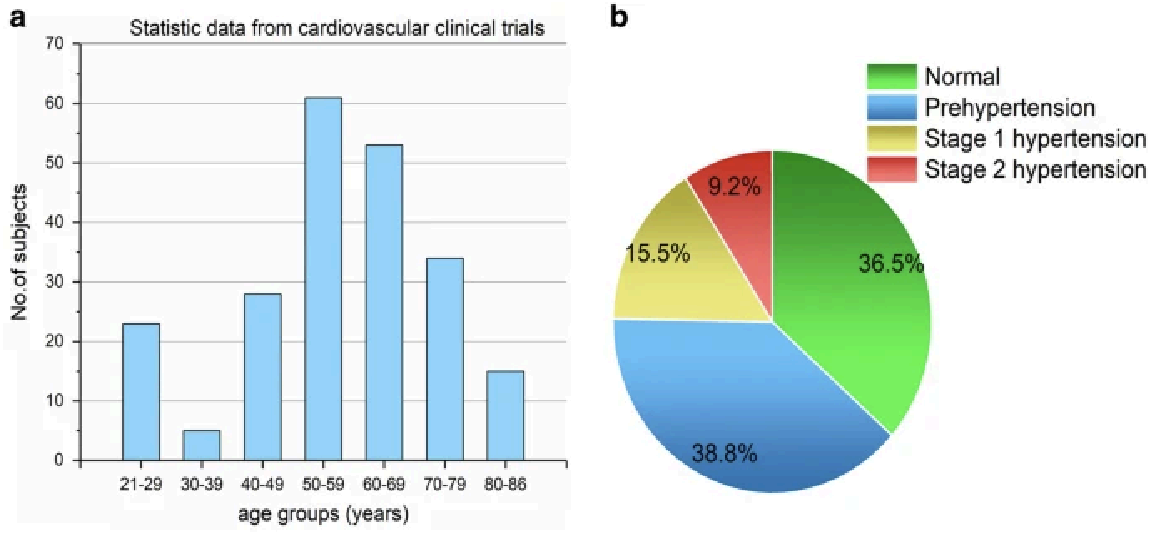
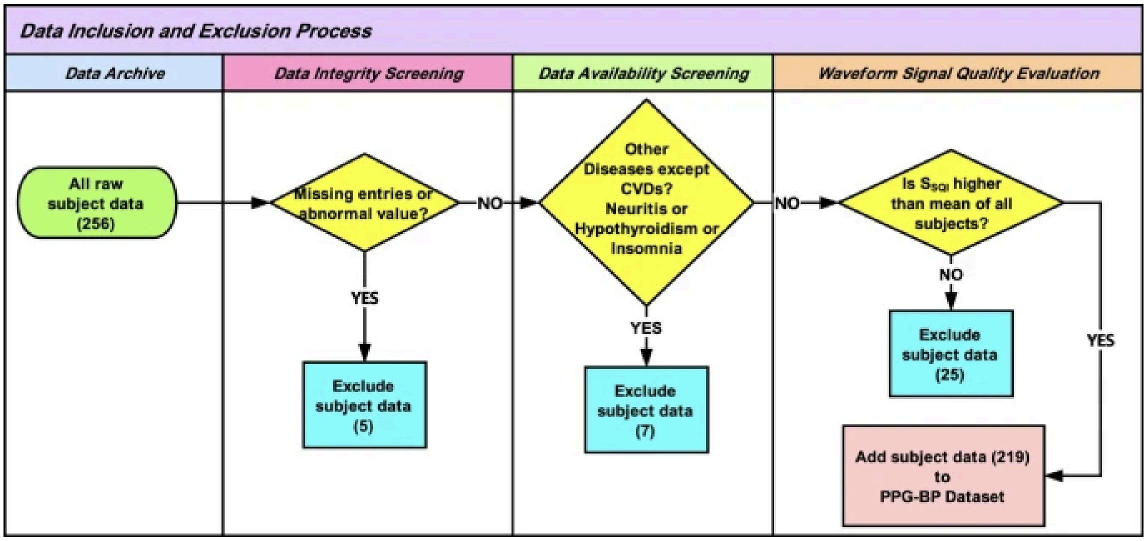


Figure 2: Summary of the PPG-BP Dataset
This figure presents participant statistics. The histogram shows the distribution of subjects across age groups, and the pie chart shows the proportion of normal, prehypertensive, and hypertensive individuals. Most participants were between ages 50–69, and nearly half had elevated blood pressure. This confirms that the dataset includes a broad range of cardiovascular conditions.

Figure 3: A process flowchart of data inclusion and exclusion.



Data Inclusion and Exclusion Process
Data Archive
Data Integrity Screening
Data Availability Screening
Waveform Signal Quality Evaluation
All raw subject data (256)
Missing entries or abnormal value?
NO
YES
Exclude subject data (5)
Other Diseases except CVDs? Neuritis or Hypothyroidism or Insomnia
NO
YES
Exclude subject data (7)
Is S_{SQI} higher than mean of all subjects?
NO
Exclude subject data (25)
YES
Add subject data (219) to PPG-BP Dataset

Figure 3: Data Inclusion and Exclusion Process
This flowchart shows how the authors selected the final sample. Out of 256 total subjects, 219 were retained after removing cases with missing values, unrelated diseases, or poor-quality signal recordings. This demonstrates that careful screening was used to maintain accurate and usable data.

Figure 4

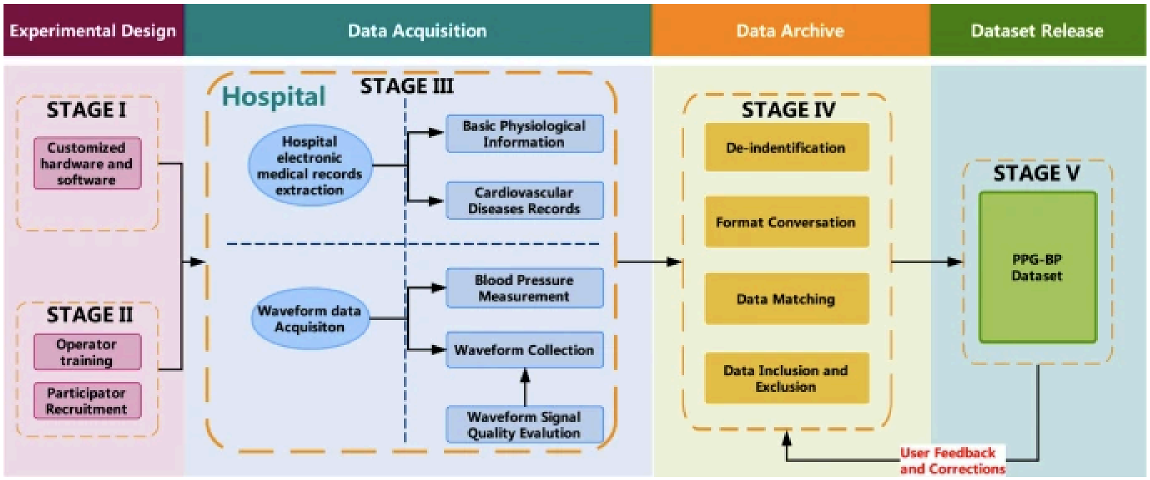


Figure 4: Overview of Dataset Creation
This figure provides an overview of how the dataset was produced. It begins with experimental design and participant recruitment, continues through data collection and validation, and ends with de-identification and public release. It highlights the study's structured and ethical approach.

Figure 5

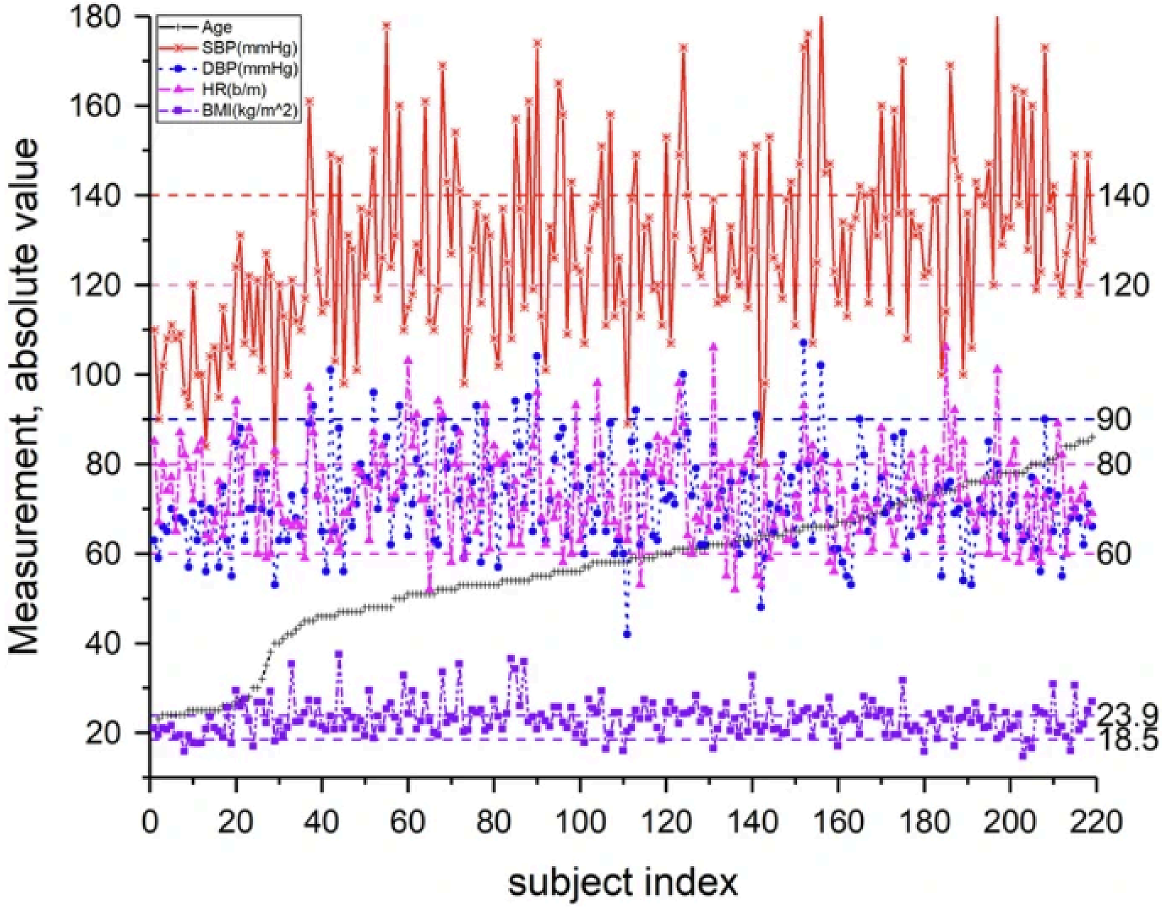


Figure 5: Physiological Measurements
This plot displays age, systolic and diastolic blood pressure, heart rate, and body mass index for all 219 participants. Systolic blood pressure increases with age, while diastolic blood pressure remains relatively stable, and heart rate does not show a strong upward trend. This demonstrates realistic physiological patterns in the sample.

Additional Figure Suggestion
An additional scatter plot showing age on the x-axis and systolic blood pressure on the y-axis could make the age-related trend easier to see. Coloring points by hypertension category would visually show how blood pressure increases with age and how participants cluster into categories.

Improvements from Part 1

From the feedback I received on project part 1 I realized that none of the datasets fully met the criteria for an A project. The feedback also noted that higher-quality projects should use original data that I download and organize myself rather than a dataset from Kaggle or other repositories. Based on this advice, I searched for new papers and found a new, short-recorded photoplethysmogram dataset for blood pressure monitoring in China. This paper collects its own biomedical data rather than reusing another dataset and provides full public access to the raw files and metadata through Figshare. This paper seems to include high quality data that is accessible for later use as I build upon this project.

Improvements from Part 2

In this revision, I clarified the dataset features by listing all eleven variables and explicitly defining the observations as the 219 participants. I added the number of rows and columns in the metadata file and explained how it relates to the waveform files in the 0_subject folder. I also corrected my interpretation of the physiological measurements, noting that systolic, but not diastolic, blood pressure increases with age. These updates fully address the feedback I received and make the explanation clearer for readers.