# Final Report: PTB Diagnostic ECG Database Analysis

## Introduction

This report details the analysis of the PTB Diagnostic ECG Database, which contains 549 electrocardiogram (ECG) records. The objective was to process the dataset, visualize key characteristics, and perform hypothesis testing to uncover relationships between demographic factors (age, gender), diagnosis (normal vs. defective), and ECG signal properties (mean signal amplitude). The analysis was conducted using Python in PyCharm, leveraging libraries such as pandas, numpy, wfdb, matplotlib, seaborn, and scipy.stats. The entire dataset was processed to ensure comprehensive insights.

**I cannot upload the entire dataset so the link is the following:**  
  
https://physionet.org/static/published-projects/ptbdb/

## Methodology

### Data Collection and Preparation

The PTB Diagnostic ECG Database was stored locally at,to run it change the paths and the files on your own system. The dataset includes ECG signal files (.hea and .dat) with associated metadata in the .hea files. A Python script was developed to:

* **Extract Signals**: Read ECG signals using the wfdb library, focusing on Lead I (first channel) and extracting 5,000 samples (5 seconds at 1000 Hz).
* **Parse Metadata**: Extract age, gender, and diagnosis from the .hea file comments. Diagnosis was classified as "Defective" if the comment included "reason for admission: myocardial infarction," and "Normal" if it included "normal."
* **Compute Features**: Calculate the mean signal amplitude for each record.
* **Create DataFrames**: Store signal data (patient ID, signal, diagnosis, mean signal amplitude) and demographics (patient ID, age, gender) in separate pandas DataFrames, then merge them into a master DataFrame.
* **Handle Missing Data**: Fill missing age values with the mean age of the dataset.

The script processed all 549 records, ensuring the entire dataset was used. The processed data was saved as CSVs.

### Data Visualization

Several visualizations were generated to explore the dataset:

* **Age Distribution by Diagnosis (Boxplot and Histogram)**: The boxplot showed the age distribution for defective (median ~60, IQR ~50–70) and normal (median ~65, IQR ~60–70) ECGs. The histogram highlighted a peak at ~60 years for defective ECGs, with normal ECGs too few (4 records) to plot meaningfully.
* **Gender vs. Diagnosis Heatmap**: Displayed counts of defective and normal ECGs by gender (Females: 144 Defective, 3 Normal; Males: 390 Defective, 1 Normal).
* **Gender Distribution (Pie Chart)**: Showed 72.7% male (391) and 27.3% female (147) among the 538 records with gender data.
* **Age vs. Mean Signal Amplitude Scatterplot**: Plotted age against mean signal amplitude, colored by diagnosis (blue for defective, orange for normal) and styled by gender (circles for F, crosses for M).
* **Pairplot**: Visualized relationships between age, mean signal amplitude, and diagnosis, confirming the clustering of normal ECGs around ages 60–70 and mean signal amplitudes 0 to 1.

All plots were saved in the project file and some have been uploaded at the end of this file for visualization.

## Hypothesis Testing

### Hypotheses and Tests

Three hypothesis tests were conducted using the entire dataset:

1. **Age Difference Between Normal and Defective ECGs**:
   * **Null Hypothesis (H₀)**: No difference in mean age between normal and defective ECGs.
   * **Alternative Hypothesis (H₁)**: There is a difference in mean age.
   * **Test**: Welch’s t-test (unequal variances).
2. **Gender and Diagnosis Association**:
   * **Null Hypothesis (H₀)**: Gender and diagnosis are independent.
   * **Alternative Hypothesis (H₁)**: Gender and diagnosis are not independent.
   * **Test**: Chi-square test of independence.
3. **Mean Signal Amplitude Difference by Diagnosis**:
   * **Null Hypothesis (H₀)**: No difference in mean signal amplitude between normal and defective ECGs.
   * **Alternative Hypothesis (H₁)**: There is a difference in mean signal amplitude.
   * **Test**: Welch’s t-test.

### Results

* **Age Difference**:
  + **P-value**: 0.1982 (approximated from chart).
  + **Interpretation**: The p-value (0.1982) is greater than 0.05, so we fail to reject H₀. There is no statistically significant difference in mean age between normal and defective ECGs. However, the boxplot suggests normal ECG patients are slightly older (median ~65 vs. ~60).
* **Gender and Diagnosis**:
  + **P-value**: 0.0362 (below 0.05 threshold).
  + **Interpretation**: The p-value (0.0362) is less than 0.05, so we reject H₀. There is a significant association between gender and diagnosis, with females showing a higher proportion of normal ECGs (2.04% vs. 0.26% in males).
* **Mean Signal Amplitude**:
  + **P-value**: 0.6497 (above 0.05 threshold).
  + **Interpretation**: The p-value (0.6497) is greater than 0.05, so we fail to reject H₀. There is no significant difference in mean signal amplitude, despite normal ECGs clustering around 0 to 1 and defective ECGs ranging from -2 to 2.

### P-values Chart

* **Chart Description**: A bar chart visualizes the p-values on a logarithmic scale, with a red dashed line at 0.05 indicating the significance threshold. The chart (hypothesis\_testing\_pvalues.png) shows:
  + Age Difference: P-value ~0.1982 (not significant).
  + Gender vs. Diagnosis: P-value ~0.0362 (significant).
  + Mean Signal Amplitude: P-value ~0.6497 (not significant).

## Conclusion

The analysis of the PTB Diagnostic ECG Database revealed a heavily skewed dataset (545 defective, 4 normal ECGs) with a male majority (72.7%), we can make it more diverse if we include more datasets with normal ecgs sinngals, but due to most them not having data lebling such as age and gender, this was the only feasable option and why I decided to switch from ecg5000 dataset. Data processing involved extracting ECG signals, parsing metadata, and computing features, followed by visualization and hypothesis testing. Key findings include a significant association between gender and diagnosis (p = 0.0362), with females more likely to have normal ECGs, though the small normal sample size (4 records) limits reliability. No significant differences were found in age (p = 0.1982) or mean signal amplitude (p = 0.6497). Future work should focus on collecting more normal ECG data and exploring additional features.

Generated on April 25, 2025

 



