Data Audit Report: Health Data Analysis

Piotr Paturej

November 23, 2024

1 Executive Summary

This report presents a comprehensive analysis of two interconnected health datasets containing information about 2,000 patients. The analysis focuses on various health metrics, lifestyle factors, and their relationships with blood pressure abnormalities.

2 Dataset Overview

2.1 Dataset Characteristics

The analysis comprises two datasets:

- Dataset 1: Contains 14 variables including demographic information, health metrics, and lifestyle factors
- Dataset 2: Contains physical activity data (steps per day) for each patient over multiple days

2.2 Data Quality Assessment

Missing data analysis revealed:

- Pregnancy data: 1,558 missing values (77.9%)
- Alcohol consumption: 242 missing values (12.1%)
- Genetic Pedigree Coefficient: 92 missing values (4.6%)
- All other variables: Complete data

3 Demographic Analysis

3.1 Age Distribution

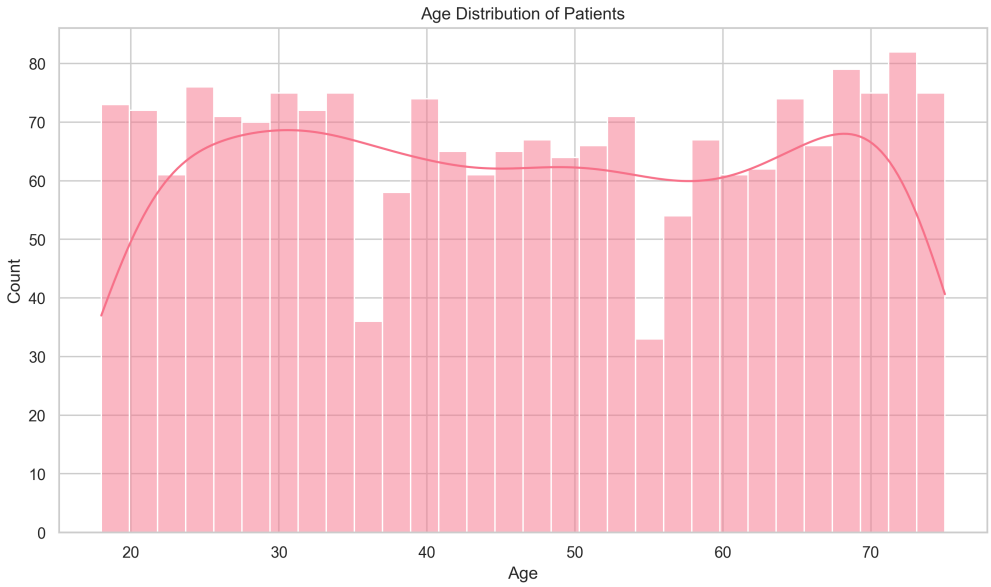

Figure 1: Age Distribution of Patients

The age distribution shows:

- Age range: 18-75 years
- Mean age: 46.56 years
- Relatively uniform distribution across age groups
- Slight increase in frequency for older age groups (65-75 years)

3.2 Gender and Blood Pressure

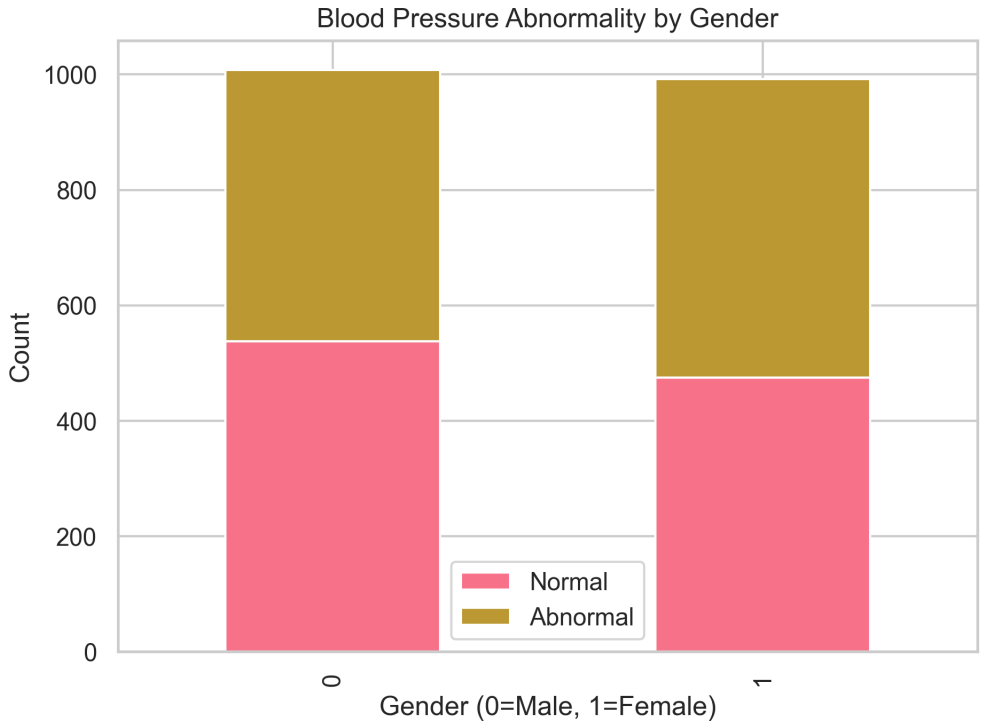

Figure 2: Blood Pressure Abnormality by Gender

Gender analysis revealed:

- Significant gender difference in blood pressure abnormalities (p=0.0159)
- Females show slightly higher prevalence of blood pressure abnormalities
- Males: 46.6% abnormal BP
- Females: 52.1% abnormal BP

4 Health Metrics Analysis

4.1 BMI Distribution

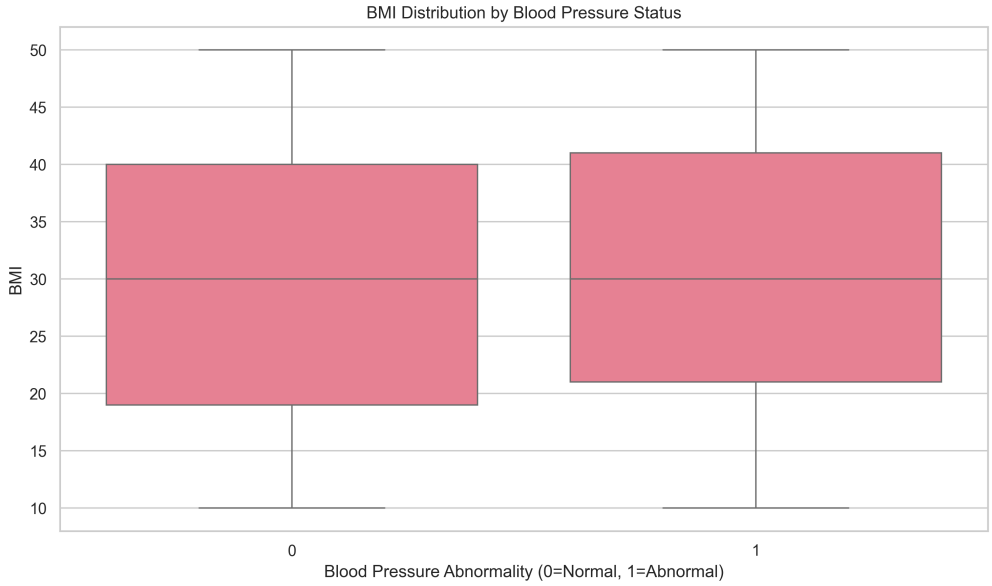

Figure 3: BMI Distribution by Blood Pressure Status

Key findings:

- Mean BMI: 30.08 (indicating overweight population)
- Higher BMI slightly associated with blood pressure abnormalities
- Large proportion (51.2%) of patients classified as obese
- Significant variation in BMI across all blood pressure groups

4.2 Correlation Analysis

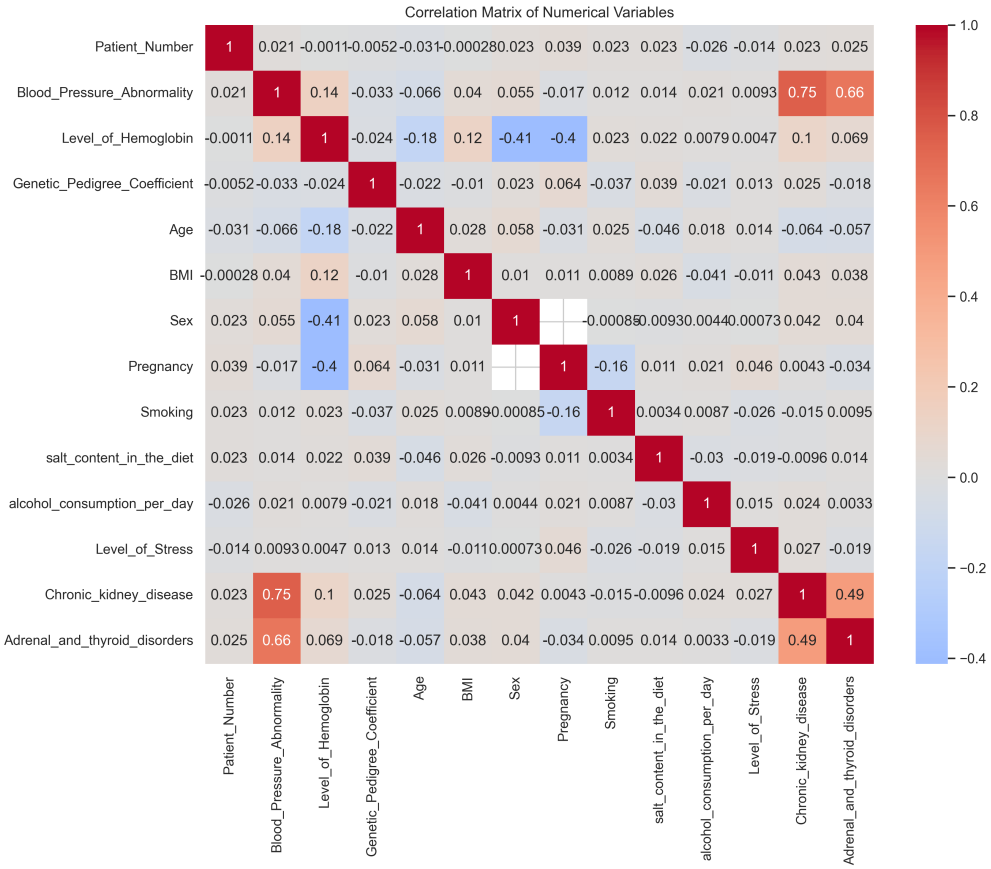

Figure 4: Correlation Matrix of Numerical Variables

Notable correlations:

- Strong correlation between BP abnormality and chronic kidney disease (0.75)
- Strong correlation between BP abnormality and thyroid disorders (0.66)
- Moderate negative correlation between hemoglobin and sex (-0.41)
- Weak correlations between lifestyle factors and BP

5 Lifestyle Factors

5.1 Physical Activity Analysis

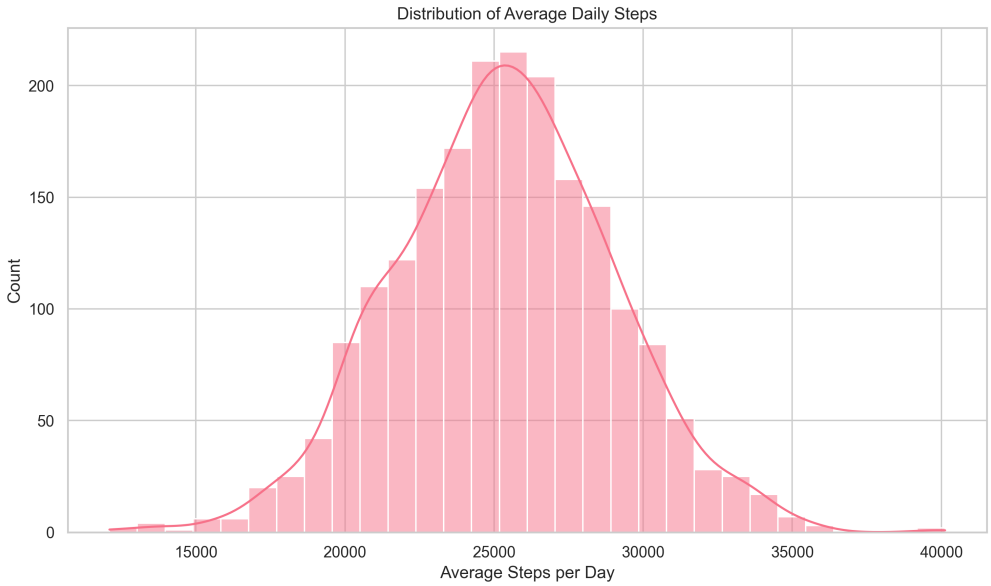

Figure 5: Distribution of Average Daily Steps

Activity patterns:

- Mean daily steps: 25,329.72
- Standard deviation: 3,669.63 steps
- Relatively normal distribution of physical activity
- No significant correlation with blood pressure status

5.2 BMI vs Physical Activity

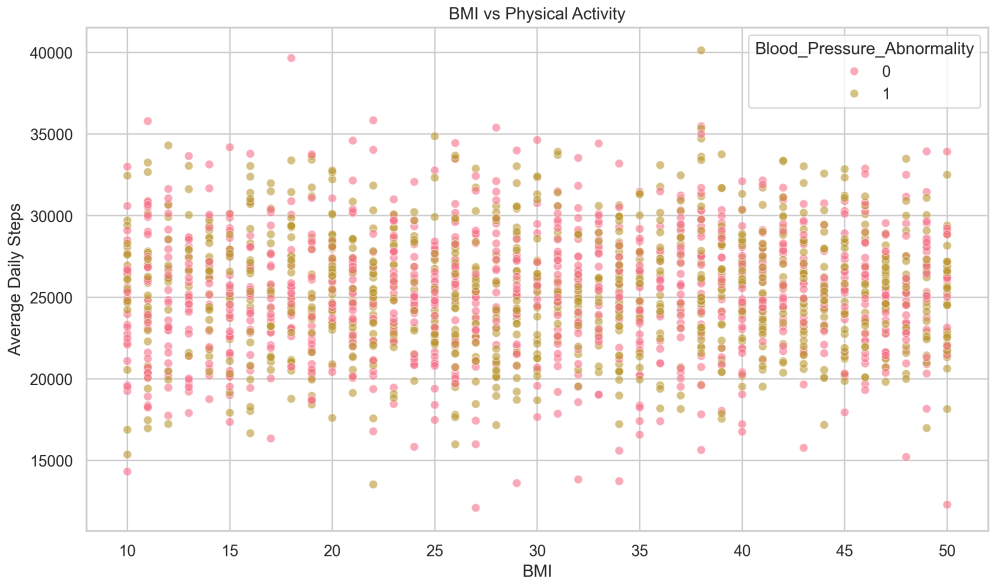

Figure 6: BMI vs Physical Activity by Blood Pressure Status

Analysis shows:

- Weak correlation between BMI and physical activity (r=0.0345)
- No clear separation between BP groups based on activity levels
- Wide range of activity levels across all BMI categories

5.3 Stress and Risk Analysis

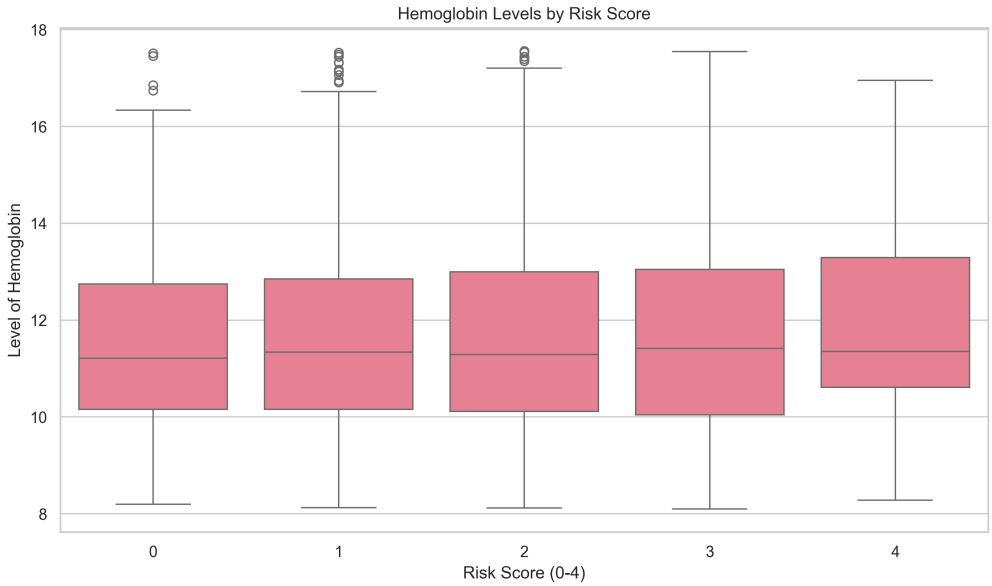

Figure 7: Hemoglobin Levels by Risk Score

Risk factor findings:

- Even distribution across stress levels (Low: 33.3%, Normal: 32.2%, High: 34.5%)
- Higher risk scores associated with slightly elevated hemoglobin levels
- No significant correlation between stress levels and BP abnormalities

6 Key Findings and Recommendations

6.1 Summary of Findings

1. **Medical Conditions:** Strongest predictors of BP abnormalities are chronic kidney disease and thyroid disorders
2. **Demographics:** Age and gender show significant but modest associations with BP abnormalities
3. **Lifestyle Factors:** Surprisingly weak correlations between lifestyle factors (smoking, alcohol, salt intake) and BP
4. **Physical Activity:** No significant relationship between activity levels and BP status
5. **BMI:** High prevalence of obesity in the population, but only weak correlation with BP abnormalities

6.2 Data Quality Recommendations

1. Address high missing data rate in pregnancy variable
2. Implement standardized collection methods for alcohol consumption data
3. Consider collecting additional lifestyle variables
4. Include more detailed physical activity metrics beyond step count

6.3 Clinical Implications

1. Focus on screening for kidney and thyroid disorders in BP management
2. Develop targeted interventions for high-risk demographic groups
3. Investigate why traditional risk factors show weak correlations
4. Consider more comprehensive lifestyle assessment methods