# Parkinson's Disease Voice Analysis

## 1. Introduction

Parkinson's Disease, a neurodegenerative disorder, wreaks havoc on movement and speech. Dopamine levels plummet in the brain, causing tremors, stiffness, and vocal difficulties like mumbled words, quiet speech, and a flat tone of voice. Cognitive and mood shifts can emerge, and the specter of dementia looms larger. Traditional diagnosis relies on neurological history and observation of motor skills – a process often fraught with difficulty, especially in early stages where motor issues are subtle. Repeated clinical visits burden patients and healthcare systems. Imagine a better way. Voice recordings, a non-invasive tool, capture the unique vocal fingerprints of Parkinson's. This project explores the potential of machine learning to analyze these recordings, envisioning a future where a simple voice test could be a crucial first step towards diagnosis, paving the way for timely intervention and improved patient care.

## 2. Dataset and Motivation

Our investigation draws upon the Parkinson's Disease dataset hosted by the UCI Machine Learning Repository, a well-known source for research data. This dataset comprises voice recordings from individuals with and without Parkinson's, along with various speech signal measurements. These measurements quantify vocal characteristics like fundamental frequency (pitch), variation in frequency and amplitude (jitter and shimmer), and noise-to-tonal ratios, providing a rich landscape for analysis. Our group chose this project because it bridges the gap between technical innovation and human well-being. Parkinson’s diagnosis often arrives late, delaying access to support and treatment. Developing a pre-screening tool using readily available technology, like voice recordings, could address this critical need. The prospect of contributing to a more accessible and efficient diagnostic process—that's what drives us.

*Table 1: Column present in the dataset*

| Feature | Data Type | Description |
|---|---|---|
| MDVP:Fo(Hz) | float64 | Fundamental frequency (Hz) |
| MDVP:Fhi(Hz) | float64 | Highest frequency (Hz) |
| MDVP:Flo(Hz) | float64 | Lowest frequency (Hz) |
| MDVP:Jitter(%) | float64 | Percent jitter |
| MDVP:Jitter(Abs) | float64 | Absolute jitter |
| MDVP:RAP | float64 | Relative amplitude perturbation |
| MDVP:PPQ | float64 | Phonation quotient |
| Jitter:DDP | float64 | Jitter: Differential perturbation quotient |
| MDVP:Shimmer | float64 | Shimmer |
| MDVP:Shimmer(dB) | float64 | Shimmer in dB |
| Shimmer:APQ3 | float64 | Shimmer: Amplitude perturbation quotient 3 |
| Shimmer:APQ5 | float64 | Shimmer: Amplitude perturbation quotient 5 |
| MDVP:APQ | float64 | Amplitude perturbation quotient |
| Shimmer:DDA | float64 | Shimmer: DDA |
| NHR | float64 | Noise-to-harmonics ratio |
| HNR | float64 | Harmonics-to-noise ratio |

| | | |
|---|---|---|
| status | int64 | 0: Healthy, 1: Parkinson's Disease |
| RPDE | float64 | Recurrence period density entropy |
| DFA | float64 | Detrended fluctuation analysis |
| spread1 | float64 | Spread1 |
| spread2 | float64 | Spread2 |
| D2 | float64 | D2 |
| PPE | float64 | Phonation percentage |

## 3. Data Exploration and Preprocessing

The dataset is structured with each row representing a single voice recording from a participant. The 'name' column serves as a unique identifier for each recording. Here’s a glimpse of the data’s structure:

Columns like 'MDVP:Fo(Hz)' (average fundamental frequency) and 'MDVP:Fhi(Hz)' (maximum fundamental frequency) are numerical (float64 data type), representing continuous measurements of vocal pitch. Other columns, like 'MDVP:Jitter(%)' and 'MDVP:Shimmer,' also numerical, quantify variations in frequency and amplitude—subtle indicators of vocal instability potentially linked to Parkinson’s. 'status,' an integer (int64), flags each recording as belonging to a healthy individual (0) or someone with Parkinson's (1)—our target variable. No major data cleaning was necessary for this initial analysis. Each row encapsulates a snapshot of an individual’s voice, providing the raw material for our investigation.

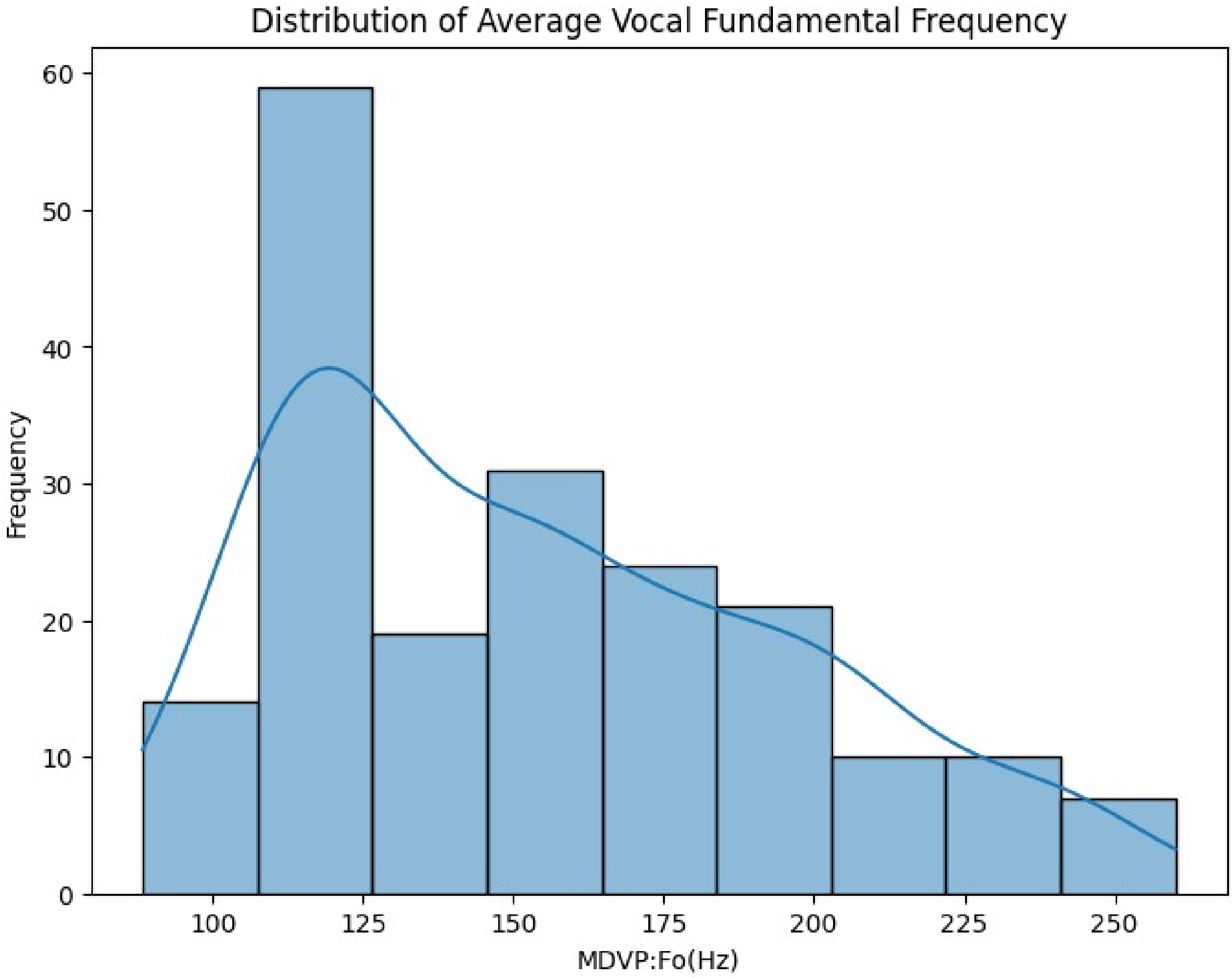

*Figure 1: Distribution of average vocal fundamental frequency*

## 4. Research Question and Hypotheses

Our research question zeroes in on a key vocal characteristic affected by Parkinson’s: Does the average vocal fundamental frequency (MDVP:Fo(Hz)) differ significantly between individuals with and without Parkinson's disease (status)? This question seeks to establish if a quantifiable vocal feature can distinguish between healthy and Parkinson’s-affected individuals.

To test this, we formulated the following hypotheses:

Null Hypothesis (H0): There is no significant difference in the average vocal fundamental frequency between individuals with and without Parkinson's disease.

Alternative Hypothesis (H1): There is a significant difference.

These hypotheses frame our statistical investigation, allowing us to assess the evidence for a relationship between vocal frequency and Parkinson's status.

## 5. Data Visualization

To gain visual insights into the data, we generated a histogram of the 'MDVP:Fo(Hz)' variable. This histogram reveals the distribution of average vocal fundamental frequencies across all participants, showing whether the data are normally distributed, skewed, or have other noticeable patterns. A box plot comparing 'MDVP:Fo(Hz)' for healthy individuals and those with Parkinson's visualizes potential differences in the central tendency (median) and spread (interquartile range) of vocal frequency between the two groups. Clear axis labels and informative titles enhance the clarity of these visualizations, making the patterns readily apparent.

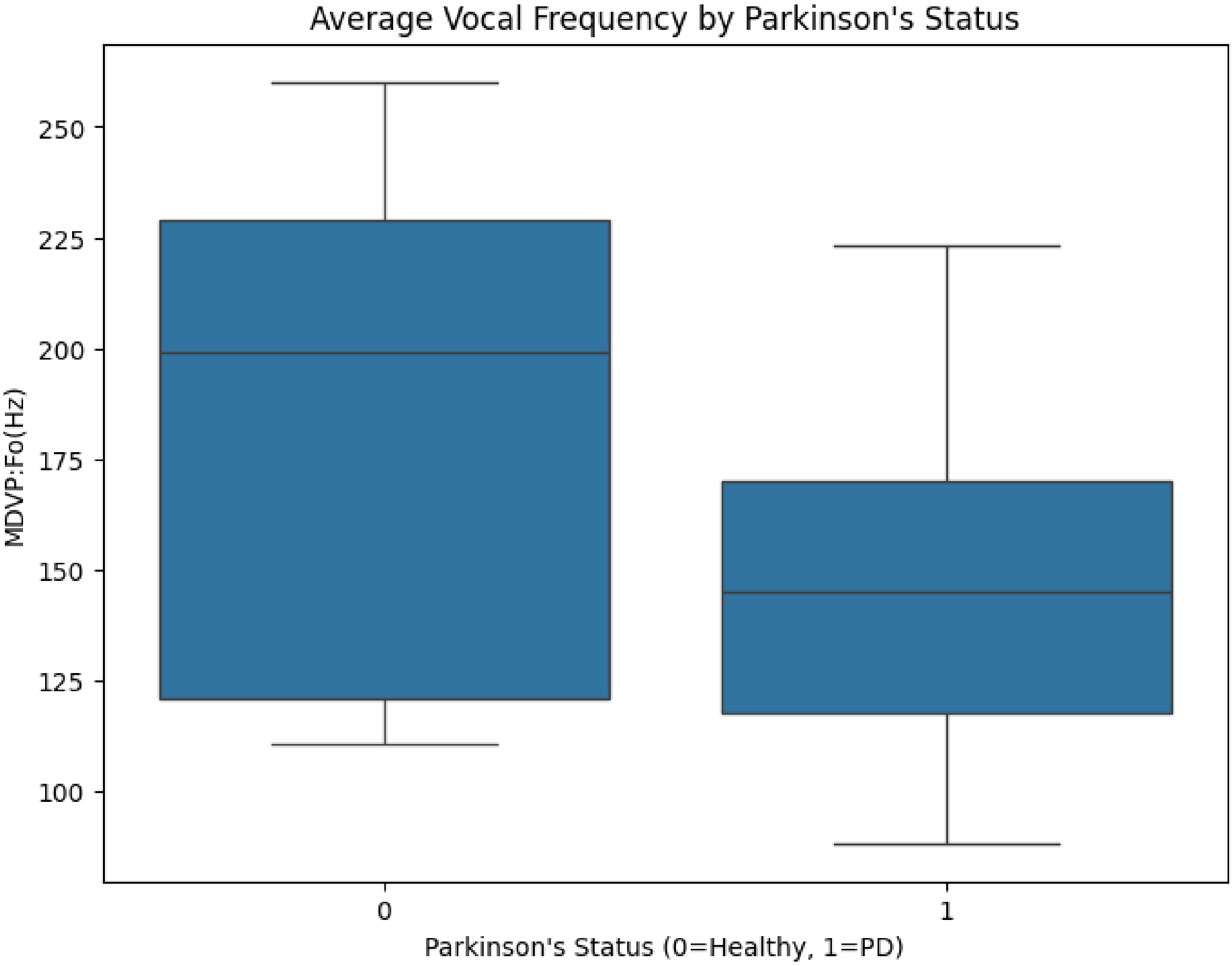

*Figure 2: Average vocal frequency by Parkinson status*

## 6. Statistical Analysis

Given our research question and the nature of our data, we selected the independent samples t-test as our statistical method. This test assesses whether the means of two independent groups (healthy vs. Parkinson's) differ significantly. The t-test is suitable here since we are comparing means of a continuous variable ('MDVP:Fo(Hz)') between two distinct groups. Additionally, before conducting the t-test, we calculated the skewness of the distribution. The skewness, found to be positive, suggests a slight right skew in the data, not deviating so drastically to force us into non parametric test options. Our t-test yielded a t-statistic of 5.77 and a p-value of approximately 3.12e-08, far below the standard significance level of 0.05. This tiny p-value leads us to reject the null hypothesis. The evidence strongly suggests a statistically significant difference in average vocal fundamental frequency between healthy individuals and those with Parkinson's. The substantial t-statistic indicates a notable difference between the group means.

## 7. Reflection and Discussion

Our analysis reveals a compelling link between average vocal fundamental frequency and Parkinson's disease. The statistically significant difference observed suggests that vocal frequency could be a valuable marker for distinguishing between healthy individuals and those with Parkinson’s. This finding has implications for developing non-invasive diagnostic tools. However, this analysis isn’t without limitations. The dataset represents a specific sample, and its generalizability to broader populations needs further investigation. Additional factors, such as age, gender, and other medical conditions, could influence vocal frequency and require careful consideration in a real-world diagnostic setting. Future research could explore the combined effect of multiple vocal characteristics, potentially enhancing diagnostic accuracy. Integrating machine learning algorithms with these findings could lead to the development of automated voice analysis tools, potentially transforming how Parkinson’s is screened and diagnosed.

## 8. Conclusion

Our analysis underscores the potential of voice recordings as a diagnostic tool for Parkinson's disease. The significant difference in average vocal fundamental frequency between healthy individuals and those with Parkinson’s supports further investigation into voice-based screening methods. While further research is essential to refine these findings and address potential limitations, this project highlights the promise of using readily available technology like voice recordings to improve the lives of individuals affected by this debilitating disease. The prospect of readily accessible pre-screening tools holds tremendous potential to change how Parkinson's is identified, potentially impacting countless lives through earlier interventions and improved patient outcomes.