1. Analytical questions

Question 1

Question 2

Question 3

To answer this question, I computed a series of descriptive statistics to understand the characteristics of the datasets. This included measures of central tendency (mean, median) and dispersion (standard deviation, coefficient of variation). I examined controls and Parkinson’s patients separately in each language.

Starting with the Czech dataset, the mean and median ages are relatively close within each group, indicating that the data are symmetrically distributed as evidenced by the kernel density plots. I also conducted a Shapiro-Wilk test due to the fact that the distribution of Parkinson’s patients seems slightly bimodal. In both groups, dispersion is low, as evidenced by the coefficient of variation which is lower than 1. However, the range of age is quite large. The control group has a larger IQR compared to the PD group, as evidenced by the box plots. The IQR of the control group can be considered moderately high, indicating variability in the middle 50% of ages among healthy controls, while the IQR of the PD group can be considered low. Both subgroups have a male-to-female ratio of 3:2, making the dataset slightly imbalanced towards males.

The Spanish dataset shows similar patterns. The mean and median ages are relatively close within each group, indicating that the data are symmetrically distributed (see kernel density plots). The coefficients of variation point to low variation (coefficient of variation < 1). The ranges are larger in the Spanish dataset, but the IQR values are smaller, indicating low variation in the middle 50% (see boxplots). The dataset is completely balanced in terms of gender.

2. Open question

First of all, creating a well-balanced dataset is critical in order to account for age and gender differences. Speech features may vary between genders and change over time. For instance, women on average tend to speak at higher pitch whereas men’s voices are generally deeper. Since there are intrinsic differences based on gender and age both genders and different age groups (relevant to research question) should be equally represented in the dataset. A dataset where a group (e.g. elderly men) dominates will give a wrong impression and could lead to inaccurate conclusions.

Furthermore, analyzing the speech features by dividing participants into subgroups based on gender and/or age can provide a deeper understanding of their relationship with the dependent variable (e.g., the participant's health status). This approach enables us to assess how the relationship between the speech features and the dependent variable varies across different subgroups, contributing to a more comprehensive analysis of the data.

A different approach is to include gender and age as covariates alongside the dependent and independent variables. This way we can examine the relationship between speech features and different conditions such as Parkinson’s while controlling for the potential confounding effects of age and gender. Ultimately, this allows us to assess, analyze and draw more accurate conclusions.