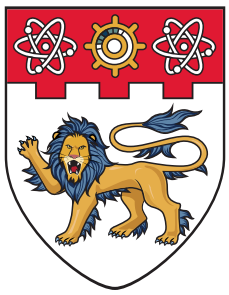


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

MH3511: Data Analysis with Computer

Project Report

Name	Email	Matric Number
Pu Fanyi	FPU001@e.ntu.edu.sg	U2220175K
Jin Qingyang	JINQ0003@e.ntu.edu.sg	U2220239A
Soo Ying Xi	D220001@e.ntu.edu.sg	U2220021D
Shan Yi	SH0005YI@e.ntu.edu.sg	U2222846C
Zhang Xintong	XZHANG113@e.ntu.edu.sg	U2210809G

Course Coordinator: Dr. Yue Mu

School of Computer Science and Engineering
Nanyang Technological University, Singapore

2023/2024 Semester 2

How You Distinguish People by Voice

Pu Fanyi* **Jin Qingyang*** **Soo Ying Xi*** **Shan Yi*** **Zhang Xintong***
School of Computer Science and Engineering
Nanyang Technological University
Singapore 639798
{FPU001, JINQ0003, D220001, SH0005YI, XZHANG113}@e.ntu.edu.sg

Abstract

People can generally distinguish the characteristics of a speaker by their voice. This project investigates how people identify others through specific features of certain vocal signals. We have released a new dataset, `DiffVoice`, and explored the differences in voices emitted by different individuals using various statistical analysis methods. Our code can be found at https://github.com/pufanyi/DiffVoice.

1 Introduction

Gender and age play a significant role in shaping the fundamental characteristics of vocal communication. Recognizing and understanding these differences is crucial for developing AI systems capable of producing voices that resonate authentically with diverse audiences. With more and more open-source voice samples available online today, we extract the data from voice samples, with further analysis to gain more insight into this topic. Although the available voice samples remain unprocessed and unrefined, our objective is to explore the correlation between voice frequency data attributes and gender or age group of the respective voice sample.

In our project, a dataset comprising labels indicating gender and age group alongside various voice frequency attributes is used. Our group downloaded open-source voice samples from the internet and further extracted diverse voice frequency attributes to compile this dataset.

Based on this dataset, we seek to answer the following questions:

1. Is there a notable discrepancy in mean frequency between male and female voices?
2. Does the gender of a voice sample correlate with its mean fundamental frequency?
3. Are there distinct variations in the median of frequency between voices of different genders?
4. Can gender be discerned by examining the quantiles of voice frequency data?
5. Can we identify the age group of the voice sample by inferring from certain voice attributes?

This report will cover the data descriptions and analysis using R language. For each of our research objectives, we performed statistical analysis and draw conclusions in the most appropriate approach, together with explanations and elaborations.

*Equal Contribution

Submission to *NTU MH3511: Data Analysis with Computer*.

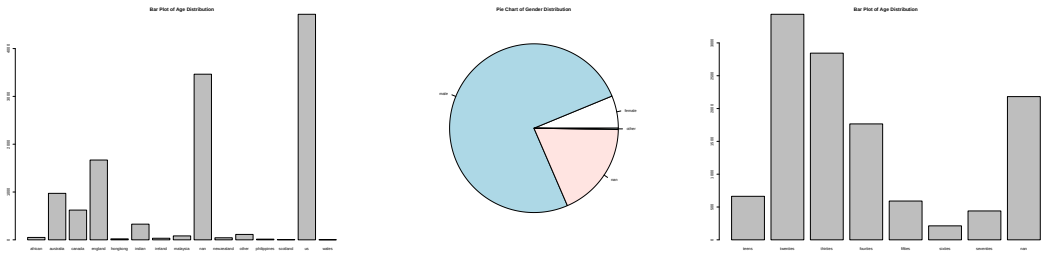


Figure 1: The distribution of different genders, regional accents, and age groups.

2 Data Preparation

2.1 The `DiffVoice` Dataset

To investigate the correlation between vocal attributes and speaker demographics, we established the `DiffVoice` dataset. This dataset was meticulously compiled from the English subset of the Common Voice [1], ensuring a diverse representation of genders, regional accents, and age groups (shown in Fig. 1).

The feature extraction pipeline for voice data involves the following steps:

1. **Audio Loading**: Initially, raw audio files are loaded and transformed into digital waveforms, serving as the foundation for subsequent analysis.
2. **Preprocessing**: The audio waveforms undergo normalisation and resampling procedures to ensure uniformity across the dataset.
3. **Feature Extraction**: Subsequently, a comprehensive set of acoustic features is extracted, including Mel-frequency cepstral coefficients (MFCCs), spectral centroid, spectral entropy, spectral flatness, pitch, and magnitude, each providing insights into different facets of the audio signal.
4. **Statistical Aggregation**: To synthesize the extracted data, statistical metrics such as the mean, standard deviation, and median are computed, offering a condensed yet informative representation of the features.

This pipeline transforms raw voice recordings into a set of numerical descriptors that capture the essential qualities of the audio for analytical tasks.

The descriptions of the extracted features have been listed in Table 1.

For enhanced accessibility, the DiffVoice dataset, along with its comprehensive feature set, has been systematically catalogued into CSV files and the HuggingFace Dataset form [2], providing a centralized and user-friendly repository for data exploration and analysis. The dataset is available for download at `https://huggingface.co/datasets/pufanyi/DiffVoice`.

2.2 Data Preparation

2.2.1 Data Normalization

The data preparation process starts with normalizing data, the steps can be described as follows:

1. **Visualizing Data**: Gaining a basic understanding of the data distribution using histogram and boxplot. This visual representation easily allows us to assess the skewness or symmetry of these distributions.
2. **Assessing Normality**: We then proceed to assess the normality of data by imposing a normal PDF on the histogram and Quantile-Quantile Plot (QQ-plot).
3. **Data Transformation**: If the data is not normal, we tried to transform the data by selecting a proper transformation function, and go to step 2 to check for normality again.

Fig. 2 presents the comprehensive pipeline for data normalisation.

Feature Name	Feature Description	Feature Type
meanfreq	Average frequency (kHz)	Continuous Variable
sd	Frequency standard deviation	Continuous Variable
median	Median frequency (kHz)	Continuous Variable
Q25	First quartile (kHz)	Continuous Variable
Q75	Third quartile (kHz)	Continuous Variable
IQR	Interquartile range (kHz)	Continuous Variable
skew	Skewness of the frequency distribution	Continuous Variable
kurt	Kurtosis of the frequency distribution	Continuous Variable
sp.ent	Spectral entropy	Continuous Variable
sfm	Spectral flatness measure	Continuous Variable
mode	Mode frequency	Continuous Variable
centroid	Frequency centroid	Continuous Variable
meanfun	Mean fundamental frequency across the signal	Continuous Variable
minfun	Minimum fundamental frequency across the signal	Continuous Variable
maxfun	Maximum fundamental frequency across the signal	Continuous Variable
meandom	Mean dominant frequency across the signal	Continuous Variable
mindom	Minimum dominant frequency across the signal	Continuous Variable
maxdom	Maximum dominant frequency across the signal	Continuous Variable
dfrange	Dominant frequency range	Continuous Variable
modindx	Modulation index	Continuous Variable
age	Age of the speaker	Ordinal Variable
gender	Gender of the speaker	Nominal Variable
accent	Accent of the speaker	Nominal Variable

Table 1: Description of Features

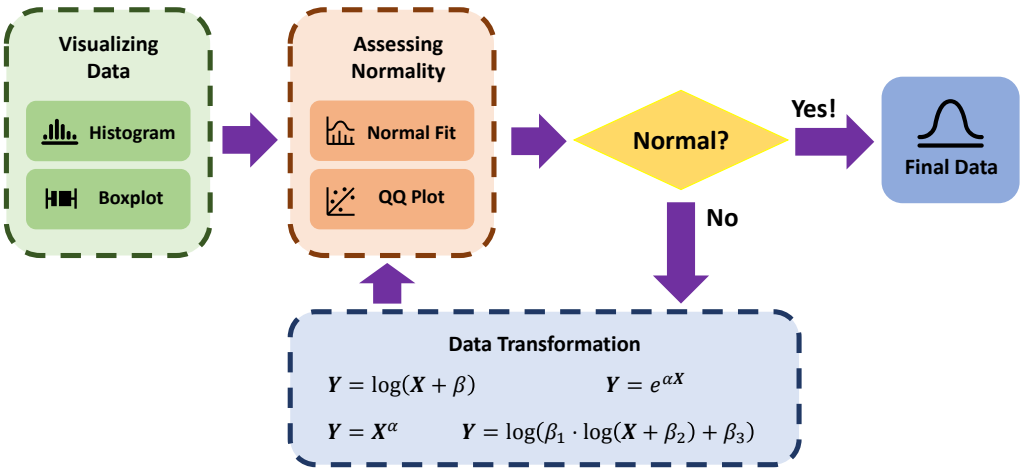


Figure 2: The pipeline for data preparation.

Feature	Before Transformation	Transformation Function	After Transformation
meanfreq	Almost Normal	$\log(x)$	Normal
sd	Not Normal	$\log(x + 300)$	Normal
median	Not Normal	$\log(x + 0.01)$	Almost
Q25	Not Normal	$\log(x + 70)$	Almost
Q75	Not Normal	$\log(x + 5000)$	Normal
IQR	Almost Normal	-	-
skew	Almost Normal	-	-
kurt	Almost Normal	-	-
sp.ent	Not Normal	$\sqrt{\log(x) + 10}$	Normal
sfm	Not Normal	$\sqrt{\log(x) + 10}$	Almost Normal
mode	Almost Normal	$\log(x + 100)$	Normal
centroid	Not Normal	$\log(x)$	Normal
meanfun	Not Normal	$\log(x + 3)$	Almost Normal
minfun	Not Normal	$\log(10 \log(x - 151.35) + 0.7)$	Almost Normal
maxfun	Not Normal	-	Not Normal
meandom	Not Normal	$\log(x + 0.01)$	Almost Normal
mindom	Not Normal	$\log(x)$	Almost Normal
maxdom	Not Normal	$\sqrt[3]{x}$	Normal
dfrange	Not Normal	$\sqrt[3]{x}$	Almost
modindx	Almost Normal	$\sqrt[3]{x}$	Normal

Table 2: Normalization transformations applied to features

Table 2 provides a detailed summary of the results from our data preparation phase.

Histograms are utilized to depict the distribution of variables pre- and post-transformation, as illustrated in Fig. 3.

Additionally, QQ-plots are employed to assess the normality of the variables, with comparisons drawn between their states before and after transformation, as shown in Fig. 4.

For an illustration of the code used to normalize the sd column, see Appendix A.

2.2.2 Feature Selection

After transformation, we selected 9 features for further analysis: meanfreq, sd, median, Q25, Q75, skew, sp.ent, sfm, meanfun.

2.2.3 Balancing Data

An initial data review revealed a significant imbalance, with male voice samples outnumbering female ones, as depicted in Fig. 1. The presence of entries with unspecified gender further complicated the analysis.

To address this, we equalized the gender distribution within the dataset, ensuring that female data were not overshadowed or misclassified as anomalies. We randomly selected samples from the predominant gender category until they matched the count of the less-represented gender.

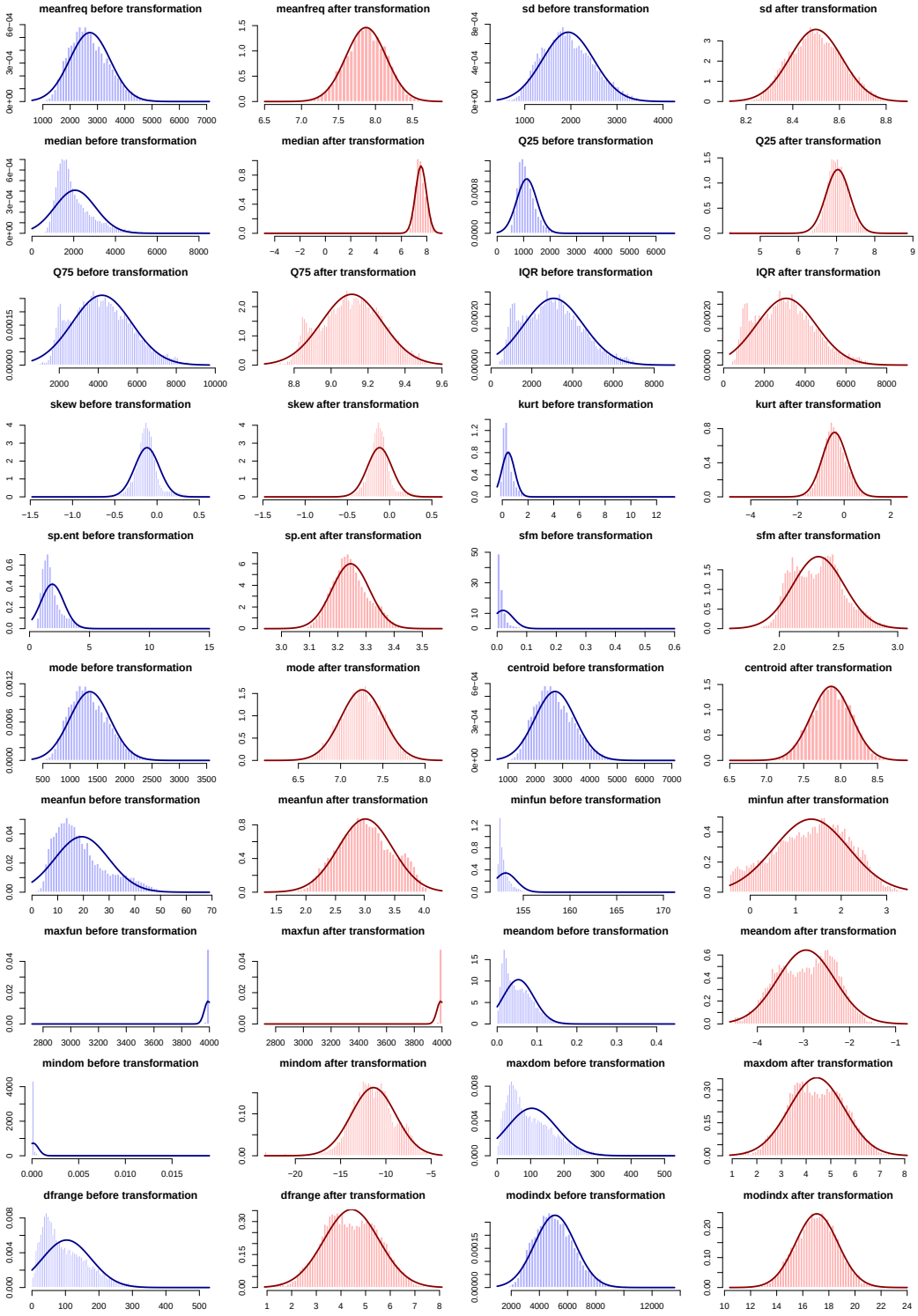


Figure 3: Histogram of variables before and after transformation.

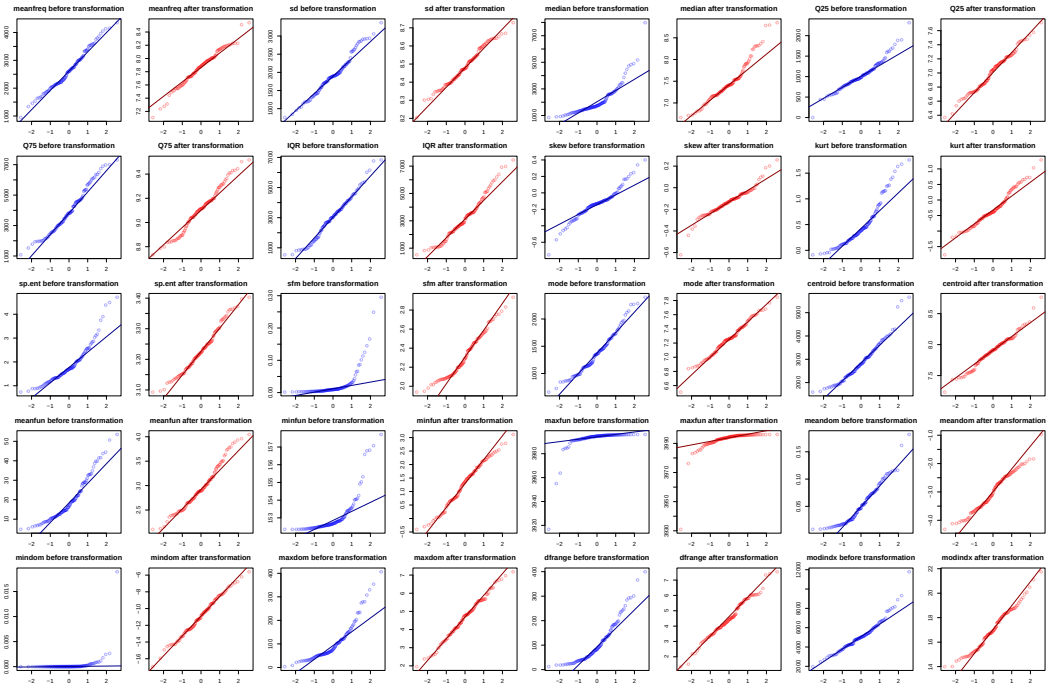


Figure 4: QQ-plot of variables before and after transformation.

2.2.4 Outlier Removal

Following the balancing of the dataset, we conducted a thorough examination to detect outliers, as such will enable us to maintain an equitable and impartial representation of both genders. Outliers were defined as observations with values exceeding 1.5 times the interquartile range (IQR), which is the range between the first and third quartiles. This method helps identify and address data points that are significantly different from the overall pattern, ensuring the integrity of our analysis.

2.2.5 Rebalancing Data

Finally, we rebalance the data to ensure gender equality.

After all the preparation, there are a total of 1444 observations from female and male samples, with 722 male samples and 722 female samples.

3 Data Analysis and Testing

3.1 Data Analysis By Gender

3.1.1 General Description

The analysis methods generally unfold through the following stages:

- **Data Visualization**: We begin by creating a histogram and a boxplot to visually inspect the distribution of the data, seperated by gender, shown in Fig. 5.
- **Assessing Normality**: Next, we check the normality by different methods including the graph overlaid with a normal pdf, QQ-plot (Fig. 6 for male and Fig. 7 female) and Shapiro-Wilk test [3].

If the data is normally distributed, we conduct an F-test to compare the variance of the data.

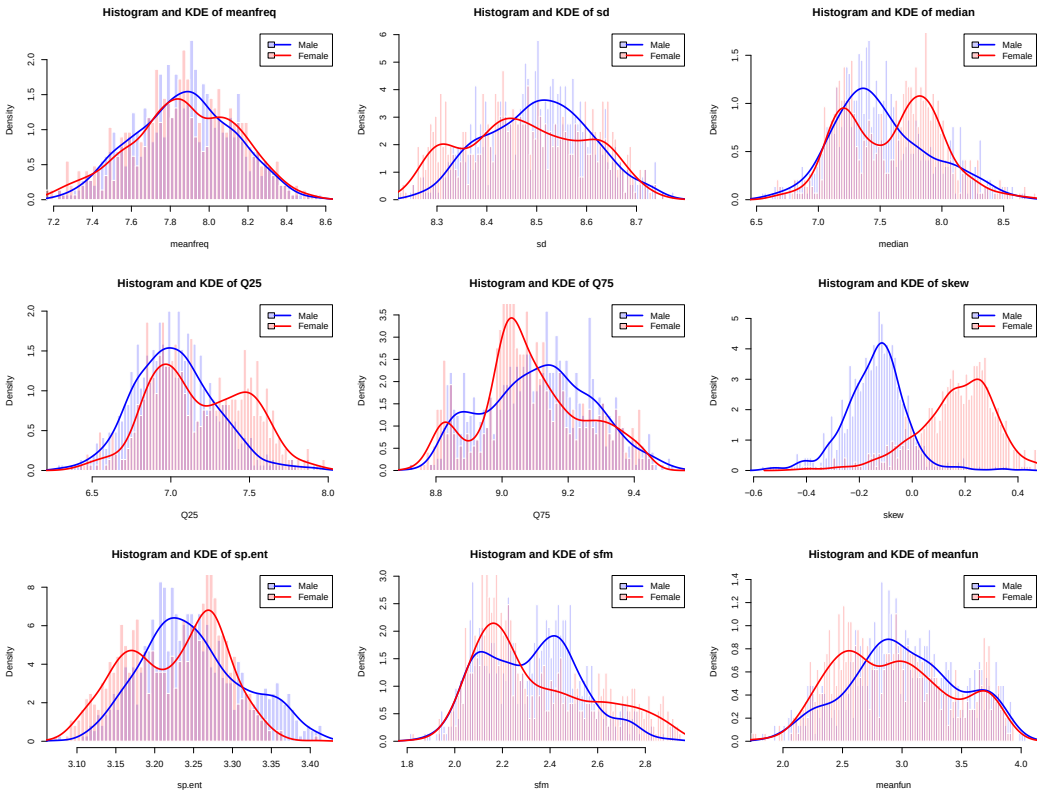


Figure 5: Histogram of variables in both genders.

- Comparing variance using **F-test**: If the p-value from the F-test is less than 0.05, we reject the null hypothesis that the variance of the data is the same across groups. Otherwise, we do not reject the null hypothesis.
- Using **two sample t-test**: We proceed to perform a two-sample t-test. If the p-value from the t-test is less than 0.05, we reject the null hypothesis that the mean of the data is the same across groups. Otherwise, we do not reject the null hypothesis.

If the data is not normally distributed, we employ the Wilcoxon test to compare the mean of the data.

- **Wilcoxon test** for non-normally distributed data: If the p-value from the Wilcoxon test is less than 0.05, we reject the null hypothesis that the mean of the data is the same across groups. Additionally, by specifying the side in the Wilcoxon test, we can determine which group has a smaller mean.

3.1.2 Analysis in Different Features

Mean Frequency Regarding `meanfreq`, we observe that the data for both males and females are normally distributed. Consequently, we first apply an F-test to evaluate whether the variances for male and female data are equivalent. With the null hypothesis H_0: *males and females have the same variance* and the alternative hypothesis H_1: *males and females have different variances*, we obtain a p-value of 0.04445. This result leads us to reject H_0 in favour of H_1. Subsequently, we conduct a t-test with H_0: *males and females have the same means* against H_1: *males and females have different means*. The resulting p-value of 0.849 indicates that we cannot reject H_0. Contrary to intuition, this suggests that there is virtually no difference in mean frequency between males and females. Then, what could be the cause behind the common perception that female voices are more shrill than male voices? We attempt to investigate this further by examining other variables.

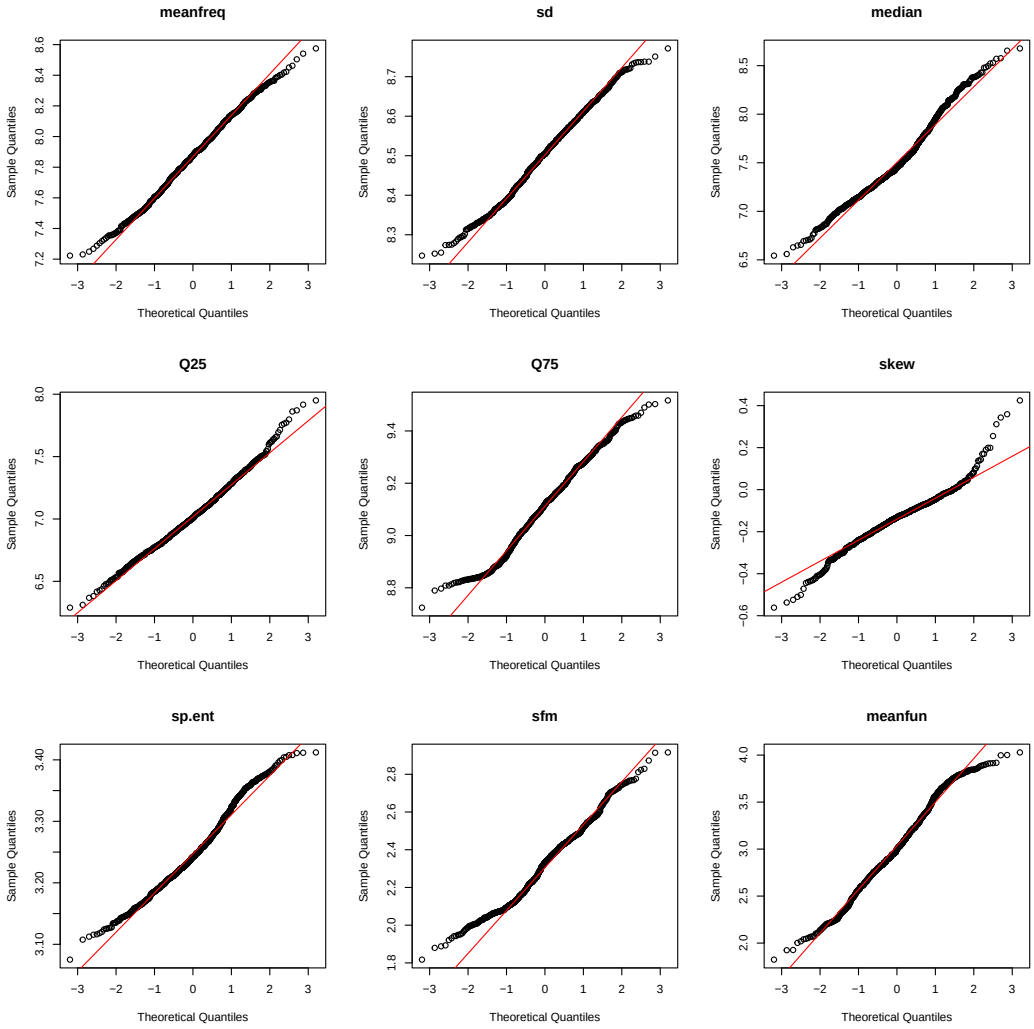


Figure 6: QQ-plot of male dataset.

Median Frequency The non-parametric test – one-sided Wilcoxon Test is applied due to the non-normal distribution, with null hypothesis H_0: *Distribution of male* `median` *and female* `median` *are the same*; alternative hypothesis H_1: *Distribution of male* `median` *has smaller values than female* `median`. The p-value from the test is 8.022×10^{-5} which is less than the significant value. Thus, we reject H_0 in favour of H_1, supporting the distribution of male `median` has smaller values than female `median`. This indicates that males generally have a lower median pitch in their voices compared to females.

First Quartile of Frequency While the `Q25` data for males is normally distributed, the skewness observed in the female `Q25` necessitates the use of a non-parametric test. We employ a one-sided Wilcoxon test, positing the null hypothesis H_0: *The distributions of male and female* `Q25` *are identical*, against the alternative hypothesis H_1: *The distribution of male* `Q25` *skews towards smaller values compared to female* `Q25`. The exceedingly small p-value of 2.2×10^{-16}, well below the threshold of significance, leads us to reject H_0 in favor of H_1. This finding corroborates that the distribution of male `Q25` indeed gravitates towards smaller values than that of females. This suggests that despite similar average frequencies, females tend to have fewer low-frequency components in their voices compared to males.

Third Quartile of the Frequency After justifying the non-normal distribution of the data, a one-sided Wilcoxon test is applied, with null hypothesis H_0: *Distribution of male* `Q75` *and female* `Q75`

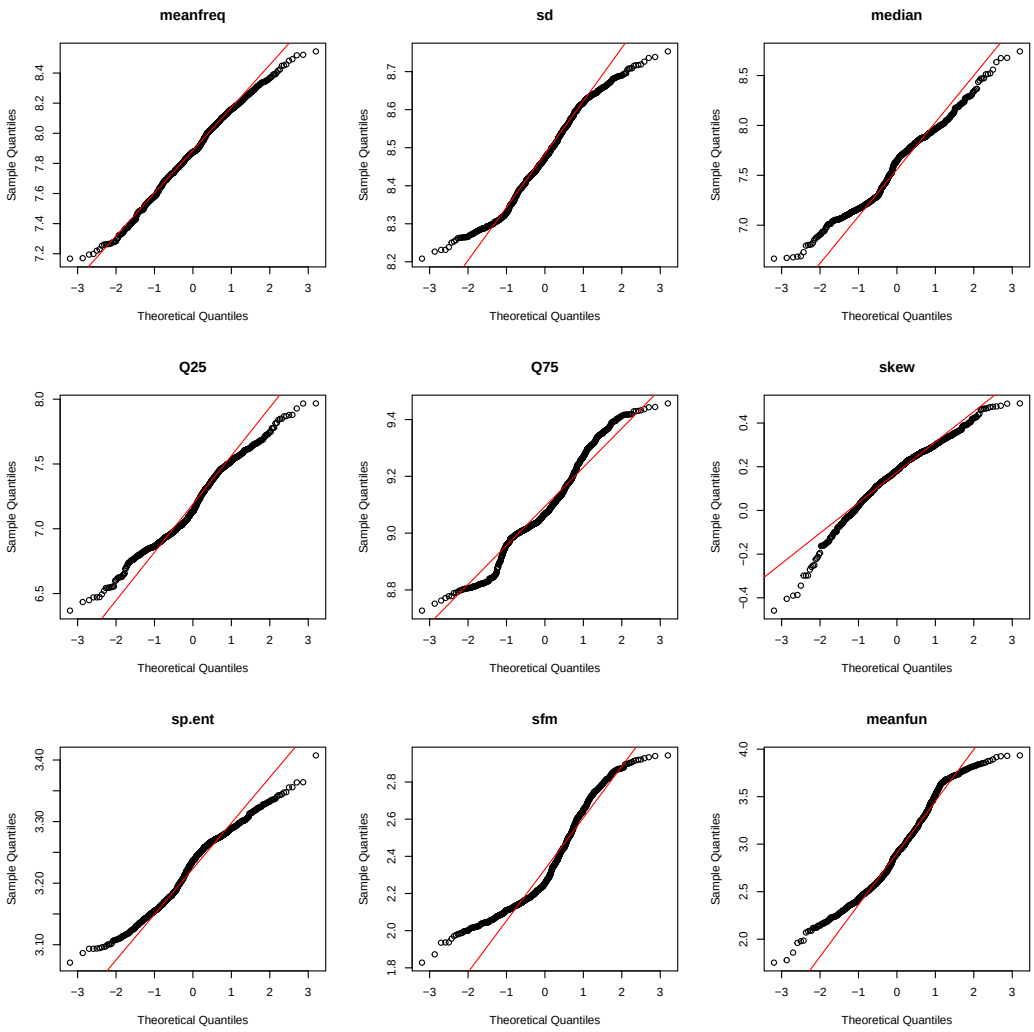


Figure 7: QQ-plot of female dataset.

are the same; alternative hypothesis H_1: *Distribution of male* `Q75` *has smaller values than female* `Q75`. The p-value from the test is 0.0007 which is less than the significant value. Thus, we reject H_0 in favour of H_1, supporting the distribution of male `Q75` has greater values than female `Q75`. This is also quite surprising, which means that in fact, males are capable of producing high-pitched sounds as well.

Skewness of the Frequency Distribution Same as above, a one-sided Wilcoxon test is applied, with null hypothesis H_0: *Distribution of male* `skew` *and female* `skew` *are the same*; alternative hypothesis H_1: *Distribution of male* `skew` *has smaller values than female* `skew`. The p-value from the test is 2.2×10^{-16} which is less than the significant value. Thus, we reject H_0 in favour of H_1, supporting the distribution of male `skew` has smaller values than female `skew`. Another interesting thing we can find in Fig 5 is that most of the females are right-skew and most of the males are left-skew. This could suggest that male voices tend to have a fuller sound in the lower frequency range, while female voices may exhibit a fuller sound in the higher frequency range.

Spectral Entropy One-sided Wilcoxon test is applied, with null hypothesis H_0: *Distribution of male* `sp.ent` *and female* `sp.ent` *are the same*; alternative hypothesis H_1: *Distribution of male* `sp.ent` *has smaller values than female* `sp.ent`. The p-value from the test is 5.564×10^{-8} which is less than the significant value. Thus, we reject H_0 in favour of H_1, supporting the distribution of male `sp.ent` has larger values than female `sp.ent`. Which suggests that there is a greater complexity

or randomness in their sound spectra. This could imply that male voices have a richer variety of frequencies or a less predictable structure compared to female voices, potentially contributing to a perception of a "rougher" or more "textured" quality in male speech.

Spectral Flatness Measure In this case, a two-sided Wilcoxon test is applied, with null hypothesis H_0: *Distribution of male* `sfm` *and female* `sfm` *have no significant difference*; alternative hypothesis H_1: *Distribution of male* `sfm` *has significant difference compared female* `sfm`. The p-value from the test is 0.8317 which is larger than the significant value. Thus, we do not reject H_0 in favour of H_1, supporting the distribution of male `sfm` has no significant difference compared to female `sfm`.

Mean Fundamental Frequency Across the Signal The data in `meanfun` are not normally distributed as suggested from the result of the Shapiro-Wilk Test, we apply the non-parametric test – Wilcoxon Test with null hypothesis H_0: *Distribution of male* `meanfun` *and female* `meanfun` *are the same*; alternative hypothesis H_1: *Distribution of male meanfun has larger values than female* `meanfun`. The p-value from the test is 1.859×10^{-5} which is less than the significant value. Thus, we reject H0 in favor of H1, supporting the distribution of male `meanfun` has higher values than female `meanfun`.

3.1.3 Outcomes

The outcomes of our analysis challenge traditional assumptions. We found that, contrary to popular opinion, the average frequencies produced by both male and female voices are nearly identical. Nonetheless, female voices are more commonly found at higher frequency ranges, whereas male voices predominantly occupy lower frequency ranges, even though they are also capable of reaching higher pitches. Moreover, the sound spectra of male voices demonstrate a higher degree of complexity or variability, potentially leading to the impression that male voices have a "rougher" or more "textured" quality.

3.2 Data Analysis By Age

We investigate whether distinct age groups exhibit varying characteristics. This examination utilizes the same features as those employed in the gender analysis.

3.2.1 Methodology

Similar to the approaches in analysis of gender, we balance the data based on gender equity first. Then we conduct the following step on each features across different age groups:

- Visualizing the data: **Boxplots** are firstly used to illustrate the distribution of the data across different age groups. Each boxplot displays the median, quantiles, and outliers for each age group.
- Checking for normality: **Shapiro-Wilk tests** are performed on the mean frequency data within each age group to determine if the data follows a normal distribution.
- Comparing differences between age groups:
  - For normally distributed data, we employ an **ANOVA** model to discern any significant differences across age groups.
  - For data that does not follow a normal distribution, **Kruskal-Wallis tests** are utilized to identify significant differences between age groups.
- Upon detecting notable differences, we proceed with pairwise comparisons to pinpoint specific age groups that differ:
  - **Pairwise t-tests** are used for data adhering to normality.
  - For non-normal data, **pairwise Wilcoxon tests** are conducted.

We would now use the test for mean frequency across different age groups as an example. We first visualise the distribution of the data in Fig. 8 The Shapiro-Wilk test suggest that most age group does not follow a normal distribution. The result of the Pairwise Wilcoxon Test is shown in Fig. 9

We perform tests for each feature. For each feature, data is not normally distributed based on the age group, and the result of Shapiro-Wilk test suggests that there is no feature such that there is no

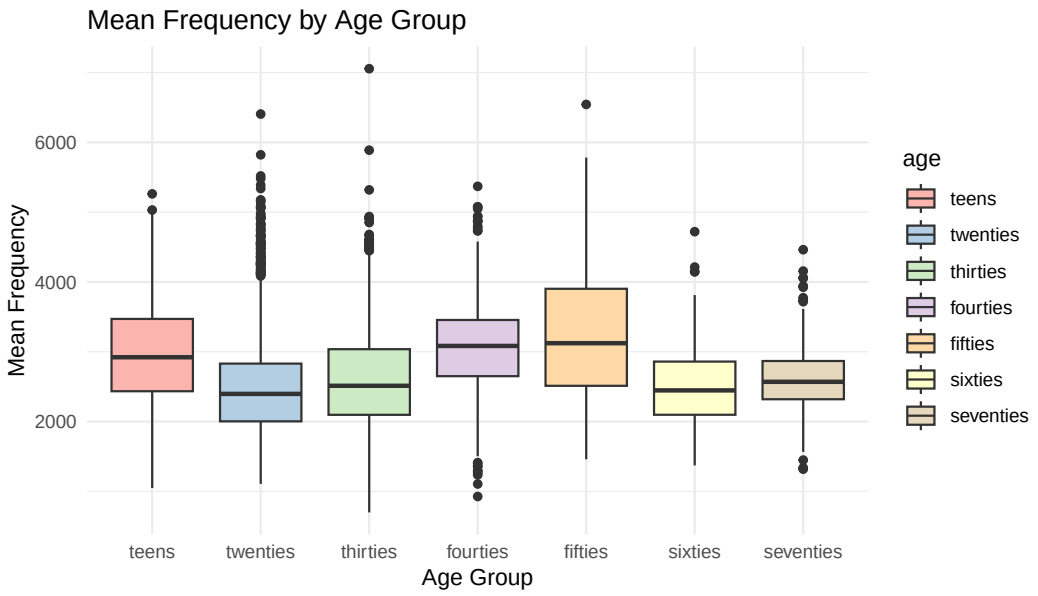


Figure 8: Mean Frequency by Age Group.

```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  voice$meanfreq and voice$age

          teens   twenties thirties fourties fifties sixties
twenties  < 2e-16 -        -        -        -       -
thirties  < 2e-16 1.1e-12  -        -        -       -
fourties  0.00047 < 2e-16  < 2e-16  -        -       -
fifties   1.9e-07 < 2e-16  < 2e-16  0.00083  -       -
sixties   1.7e-14 0.42174  0.04525  < 2e-16  < 2e-16 -
seventies 3.6e-16 2.9e-08  0.23866  < 2e-16  < 2e-16 0.00664
```

Figure 9: Pairwise Wilcoxon Test Result.

significance difference between all age groups. there and the detailed results will be provided in the appendix.

3.2.2 Outcomes

After the tests, we listed our outcomes as below.

meanfreq There is no significant difference between the mean frequency of the thirties and sixties age groups, nor between the thirties and seventies age groups.

sd There is no significant difference between the standard deviation of the teens and fifties age groups, teens and seventies age groups, thirties and sixties age groups, nor fifties and seventies age groups.

Q25 There is no significant difference between the Q25 of the teens, twenties and thirties age groups.

Q75 There is no significant difference between the Q75 of the teens and fifties age groups, thirties and sixties age groups, forties and fifties age groups, nor sixties and seventies age groups.

skew There is no significant difference between the skewness of the teens, twenties and sixties age groups, thirties and fifties age groups, nor thirties and sixties age groups.

sp.ent There is no significant difference between the sp.ent of the twenties and sixties age groups.

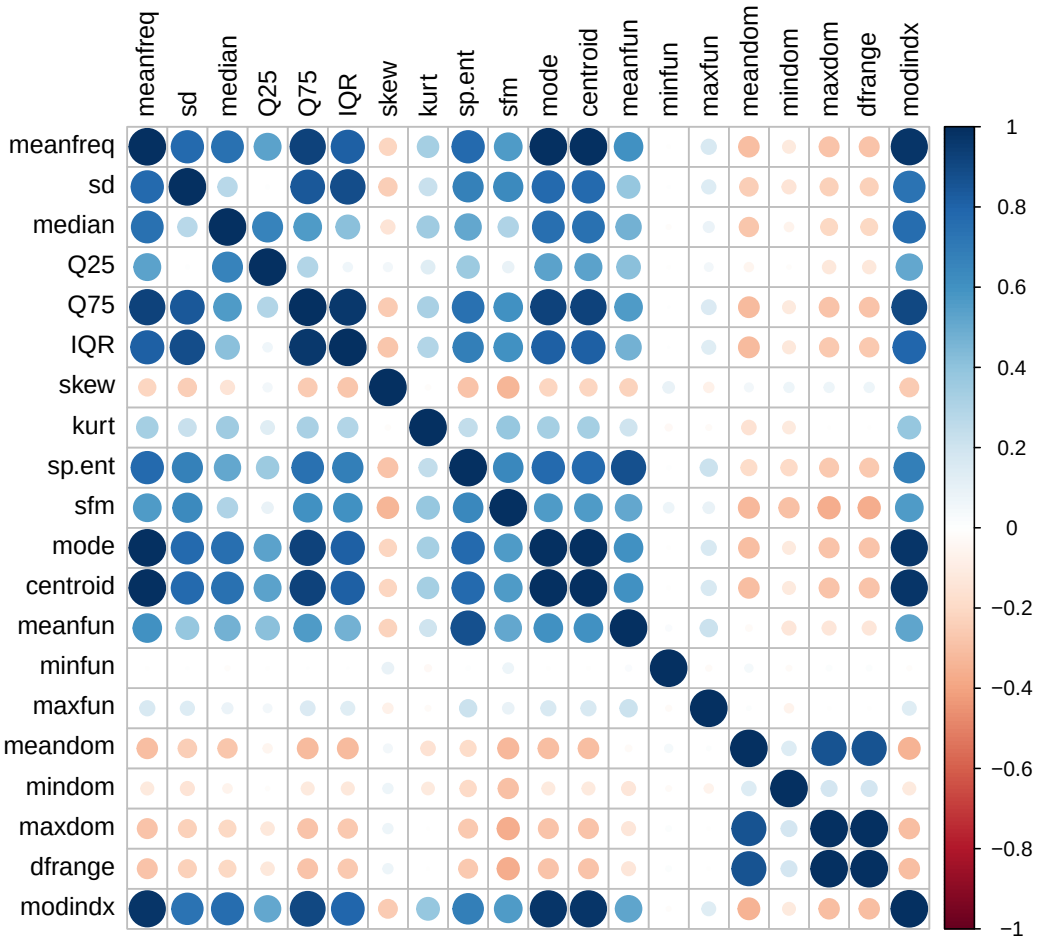


Figure 10: Correlation Matrix.

sfm There is no significant difference between the sfm of the teens and fourties age groups, teens and fifties age groups, twenties and sixties age groups, thirties and seventies, nor sixties and seventies age groups.

minfun There is no significant difference between the minfun of the sixties and seventies age groups.

3.3 Correlation Between Numerical Features

We visualise the correlation between all variables based on the transformed data in Fig. 10. We can get a glimpse of the correlation between variables.

We also plot a scatter plot to visualize the correlation between some variables (`meanfreq`, `median`, `Q25`, `Q75`, `meandom`) based on different genders. As shown in Fig. 11.

4 Regression Analysis

In this section, we will explore regression analysis techniques to understand the relationship between various features extracted from voice samples and the target variable. Additionally, we will delve

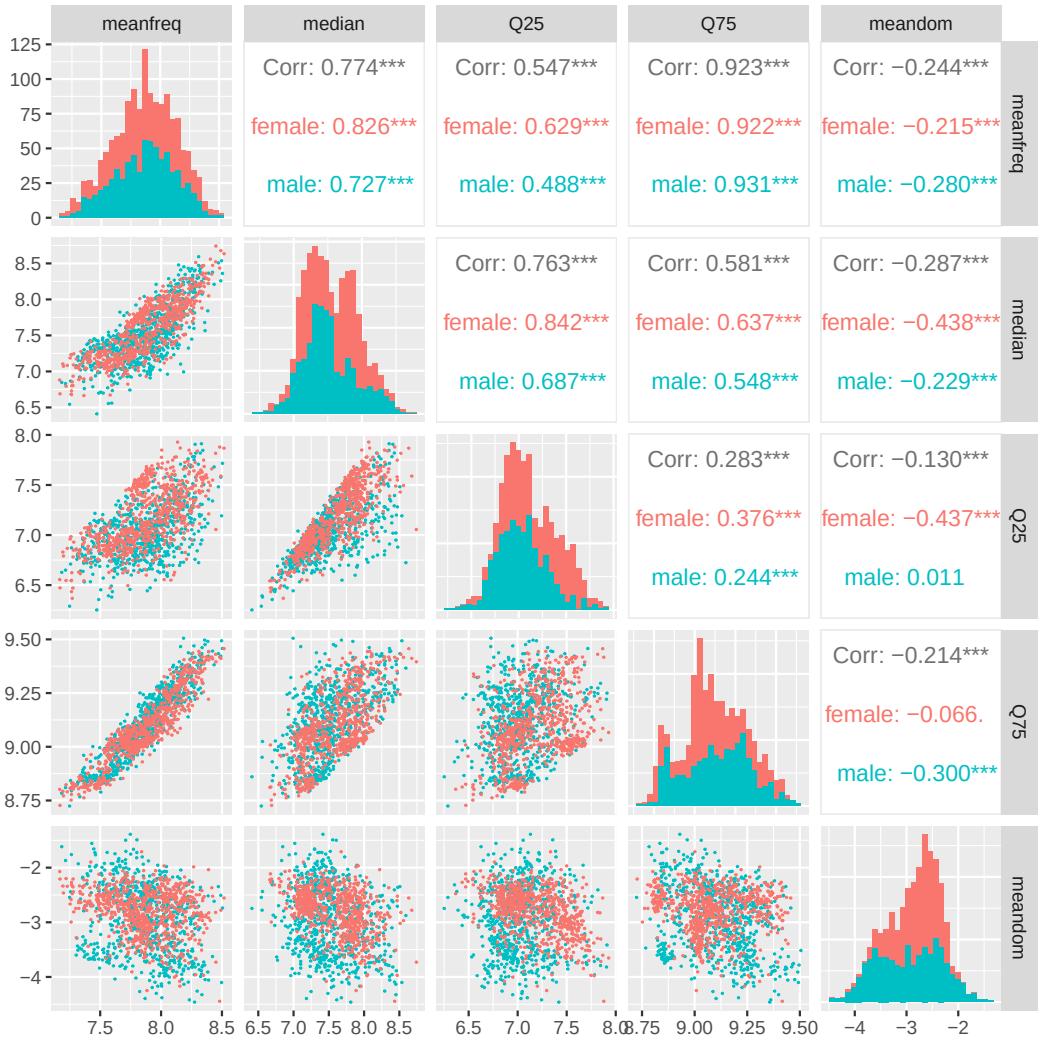


Figure 11: Scatter Matrix.

into logistic regression and k-nearest neighbours (KNN) classification to predict gender based on voice features.

4.1 Relations Amount Features of Voice

4.1.1 Simple Linear Regression

We will begin by examining the relationship between individual features and mean frequency (meanfreq) using simple linear regression. Specifically, we will select a subset of features that are likely to have a significant impact on mean frequency based on prior knowledge or domain expertise. Features such as `meanfun`, `meandom`, `mindom` and `maxdom` are some of the variables we will consider.

Initially, we perform a correlation test to examine the presence of any relationships. If a non-zero correlation is established, we proceed with linear regression to analyze the relationship between `meanfreq` and the chosen feature. Finally, we plot a scatter diagram to visualize and assess the efficacy of the model.

meanfun We conduct a correlation test between `meanfreq` and `meanfun`, since the p-value is smaller than 0.05, we reject H_0. Therefore, the correlation is not equal to 0. We then use sim-

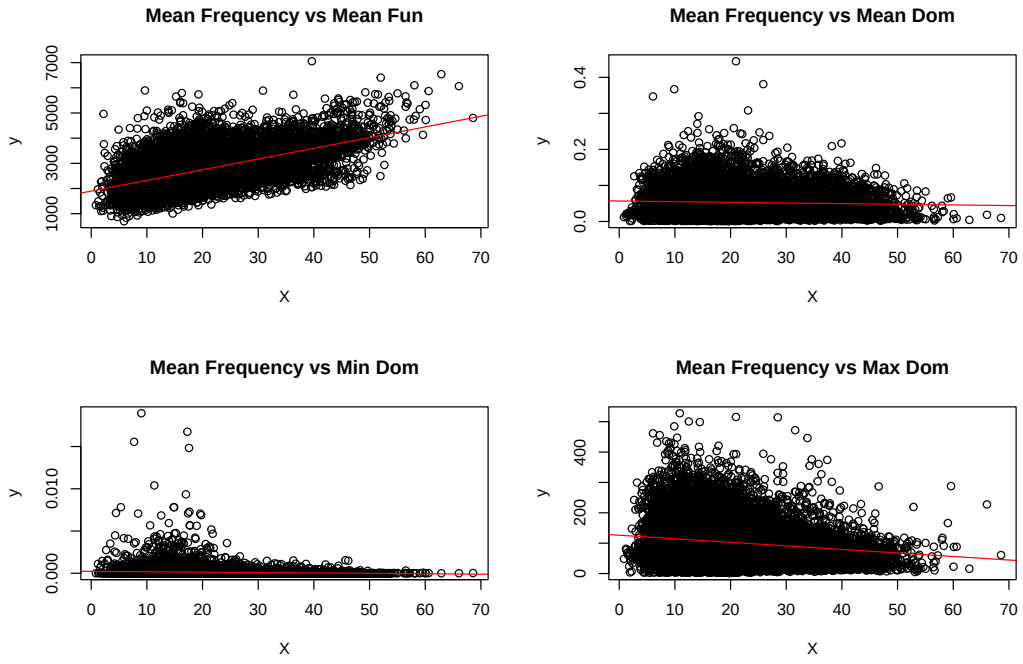


Figure 12: Simple Linear Regression.

ple linear regression model to fit the data, with $\hat{y} = 0.30272 \cdot x + 6.97028$ and R-Sqaure of 0.2849. Thus, we can conclude that `meanfreq` and `meanfun` are positively correlated.

meandom We conduct a correlation test between `meanfreq` and `meandom`, since the p-value is smaller than 0.05, we reject H_0. Therefore, the correlation is not equal to 0. We then use simple linear regression model to fit the data, with $\hat{y} = -0.12007 \cdot x + 7.52414$ and R-Square of 0.2555. Thus, we can conclude that `meanfreq` and `meandom` are negatively correlated.

mindom We conduct a correlation test between `meanfreq` and `mindom`, since the p-value is smaller than 0.05, we reject H0. Therefore, the correlation is not equal to 0. We then use simple linear regression model to fit the data, with $\hat{y} = -0.028755 \cdot x + 7.543237$ and R-Sqaure of 0.05978. Thus, we can conclude that `meanfreq` and `mindom` are negatively correlated.

maxdom We conduct a correlation test between `meanfreq` and `maxdom`, since the p-value is smaller than 0.05, we reject H0. Therefore, the correlation is not equal to 0. We then use simple linear regression model to fit the data, with with $\hat{y} = -0.036590 \cdot x + 8.040413$ and R-Square of 0.02178. Thus, we can conclude that `meanfreq` and `maxdom` are negatively correlated.

We plotted scatter plot with fitted line for these four features in Fig. 12.

4.1.2 Multiple Linear Regression

Moreover, we will extend our analysis to multiple linear regression, where we simultaneously consider multiple predictor variables to predict mean frequency. This will allow us to assess the combined effect of various features on mean frequency, providing a more comprehensive understanding of the relationship. Similar to simple linear regression, we use features such as `meanfun`, `meandom`,

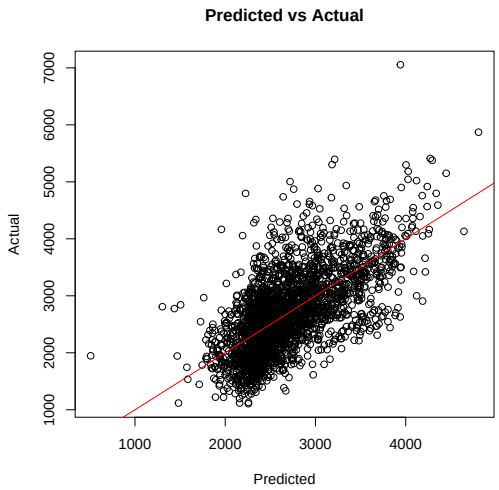


Figure 13: Multiple Linear Regression.

`mindom` and `maxdom` to predict the `meanfreq`.

$$\begin{aligned}\hat{y} &= 2124.7260 \\ &+ 43.0794 \cdot X_{\text{meanfun}} \\ &- 7431.3557 \cdot X_{\text{meandom}} \\ &- 8434.6393 \cdot X_{\text{mindom}} \\ &+ 1.5605 \cdot X_{\text{maxdom}} \\ &+ \epsilon\end{aligned}$$

We utilize this formula to predict the value of `meanfreq`, and generate a diagram as below, Fig. 13.

4.2 Relationship Between Voice Feature And Speaker

4.2.1 Logistic Regression

Moving beyond linear regression, we will explore logistic regression to predict gender based on voice features. Logistic regression is a powerful tool for binary classification problems, such as predicting gender (male/female) based on voice characteristics. By employing logistic regression, we aim to build a robust predictive model that accurately classifies the gender of speakers based on their voice attributes.

$$\frac{1}{1 + e^{-(101.25 + 11.20x)}}$$

We obtain an accuracy of the test set at 94.37%. Which is quite high. Therefore, the model performs well in predicting the gender categories.

'True Positive Rate' represents the proportion of actual "male" cases correctly identified, which is 93.75%.

'True Negative Rate' represents the proportion of actual "female" cases correctly identified, which is 95.00%.

'False Positive Rate' represents the proportion of actual "female" cases incorrectly classified as "male", which is 5.00%.

'False Negative Rate' represents the proportion of actual "male" cases incorrectly classified as "female", which is 6.25%.

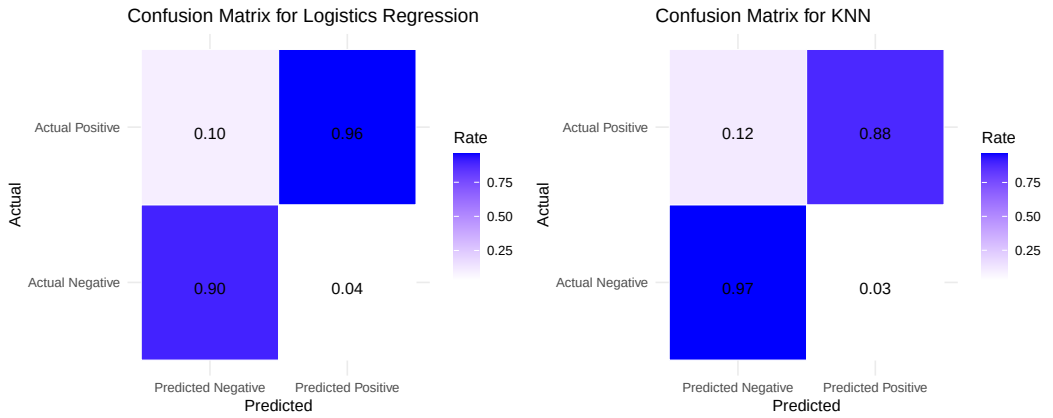


Figure 14: Logistic Regression heatmap and KNN hearmap.

Overall, the logistic regression model appears to perform well, with high accuracy and relatively low error rates.

4.2.2 K Nearest Neighbour

Finally, we will explore KNN classification to predict gender based on voice features. KNN classification leverages the proximity of data points in feature space to make predictions, making it suitable for classification tasks where data points with similar features are likely to belong to the same class. By employing KNN classification, we aim to develop an accurate model for gender prediction based on voice characteristics.

We obtain an accuracy of test set at 92.49%, which is quite high. Therefore, the model performs well in predicting the gender categories.

'True Positive Rate' represents the proportion of actual "male" cases correctly identified, which is 92.96%.

'True Negative Rate' represents the proportion of actual "female" cases correctly identified, which is 88.03%.

'False Positive Rate' represents the proportion of actual "female" cases incorrectly classified as "male", which is 11.98%.

'False Negative Rate' represents the proportion of actual "male" cases incorrectly classified as "female", which is 7.04%.

Overall, the K-Nearest Neighbour model appears to perform well, with high accuracy and relatively low error rates.

We plot a heatmap to illustarte the accuracy of both Logistic Regression Model and K-Nearest Neighbor, in Fig. 14.

5 Conclusion and Discussion

In summary, our project, "How You Distinguish People by Voice", revealed several findings about the correlation between various voice frequency data attributes and demographic factors including gender and age.

We have concluded the following key findings which answer the 4 questions above: 1. Based on our analysis, the mean frequency between male and female voices has NO significant difference, as suggested by the statistical tests conducted. 2. In fact, the male mean fundamental frequency is much higher than the female mean fundamental frequency. By conducting the Shapiro-Wilk Test and Wilcoxon Test, we obtain a relatively small p-value. Therefore, we reject the hypothesis where male mean fundamental frequency and female mean fundamental frequency are the same. 3. As

suggested from our analysis, males generally have a lower median frequency as compared to the female voice. 4. Surprisingly from our analysis, there is enough evidence to support that the third quantile of males has greater values than that of females. However, upon examining the distribution graph, the difference is not very significant. On the other hand, there is enough evidence to support that the first quantile of males has smaller values than that of females. 5. Generally, the features will have different mean values across different age groups.

References

[1] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

[2] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. Datasets: A community library for natural language processing. *arXiv preprint arXiv:2109.02846*, 2021.

[3] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.

A Sample Code for the Project

Data Analysis for SD

Load Data

```
data_path <- "../../data/original/train.csv"
voice <- read.csv(data_path)
head(voice)
```

```
##   id meanfreq       sd   median      Q25      Q75      IQR        skew
## 1  0 3521.667 2332.212 2997.294 1660.408 4621.867 2961.459  0.11656897
## 2  1 4189.998 2430.977 4302.741 1832.028 5901.071 4069.043  0.04560770
## 3  2 3154.455 2150.497 2609.968 1460.612 4053.928 2593.316 -0.16147499
## 4  3 4384.338 3029.302 3426.479 1596.072 7283.314 5687.242  0.02416762
## 5  4 4557.150 3158.111 4543.116 1608.165 8074.335 6466.170  0.11711588
## 6  5 4069.004 2983.199 2565.487 1305.284 6961.581 5656.297  0.13049391
##        kurt   sp.ent         sfm     mode centroid  meanfun   minfun   maxfun
## 1 0.9817728 2.308696 0.008450270 1761.333 3521.667 32.33476 153.1934 3995.790
## 2 0.9214181 3.522410 0.022862796 2095.499 4189.998 42.56545 154.0434 3993.462
## 3 0.3882481 2.027891 0.006853276 1577.728 3154.455 26.15712 153.4610 3995.524
## 4 1.4739316 4.823092 0.084471270 2192.669 4384.338 37.56627 153.6399 3994.671
## 5 1.2885699 3.820815 0.100988194 2279.075 4557.150 29.34924 153.8535 3994.646
## 6 0.7668548 3.726702 0.073939204 2035.002 4069.004 29.89368 153.2515 3995.253
##      meandom       mindom    maxdom   dfrange  modindx      age gender accent
## 1 0.06084856 9.842593e-04 194.17128 194.17029 5914.581 twenties female canada
## 2 0.04495757 7.060266e-04 102.27859 102.27788 7693.945 twenties female canada
## 3 0.08144125 2.950821e-04 164.99316 164.99287 5261.606 twenties female canada
## 4 0.01039643 3.165859e-08  29.66787  29.66787 7942.756      nan    nan    nan
## 5 0.01848914 9.267869e-07  85.19259  85.19259 8383.634      nan    nan    nan
## 6 0.01521549 6.052965e-07  32.57839  32.57839 7575.469      nan    nan    nan
```

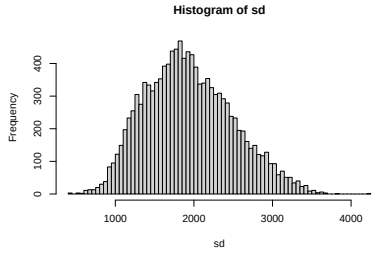
Visualizing the Data

We selected `sd` column to perform the analysis.

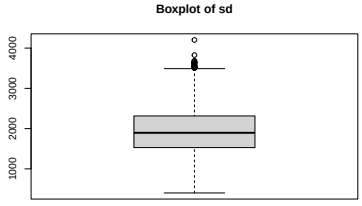
First, we load the data and draw a histogram of the `sd` column to get an initial understanding of the data distribution.

```
sd <- voice$sd
hist(sd, breaks = 80, main = "Histogram of sd", xlab = "sd")
```

1




```
boxplot(sd, main = "Boxplot of sd")
```



Then, we generate the descriptive statistics of the `sd` column.

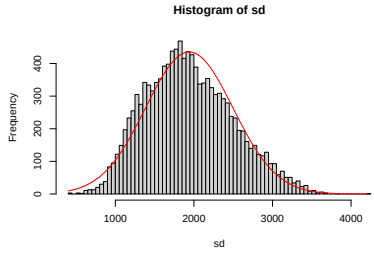
```
library(psych)
describe(sd, type = 1)
```

```
##    vars     n   mean     sd  median trimmed    mad    min     max   range skew
## X1    1 12135 1940.3 555.86 1896.32 1918.18 580.72 402.55 4202.62 3800.07 0.33
##    kurtosis   se
## X1    -0.31 5.05
```

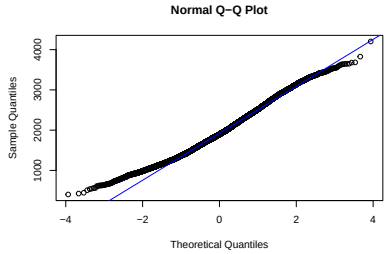
Assessing Data Normality

After that, we assess the normality of the `sd` column by drawing a histogram with a normal curve and a Q-Q plot.

```
hist(sd, breaks = 80, main = "Histogram of sd", xlab = "sd")
# impose a normal curve on the histogram
xpt <- seq(402, 4203, by = 0.1)
n_den <- dnorm(xpt, mean = mean(sd), sd = sd(sd))
ypt <- n_den * length(sd) * 50
lines(xpt, ypt, col = "red")
```



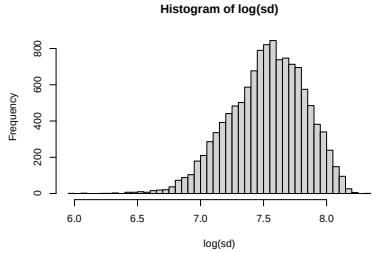

```
qqnorm(sd)
qqline(sd, col = "blue")
```



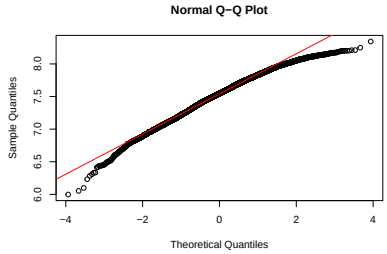
Transformation

We found that the data is almost normally distributed, but not perfect. We tried to log-transform the data to see if it can be improved.

```
sd_trans <- log(sd)
hist(sd_trans, breaks = 80, main = "Histogram of log(sd)", xlab = "log(sd)")
```



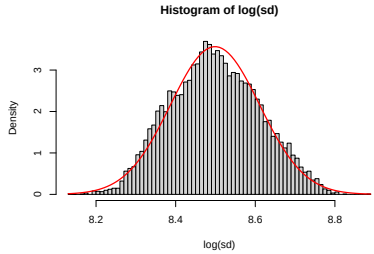
```r
qqnorm(sd_trans)
qqline(sd_trans, col = "red")
```



We observed that after log-transformation, the data's fit to a normal distribution did not improve as expected.

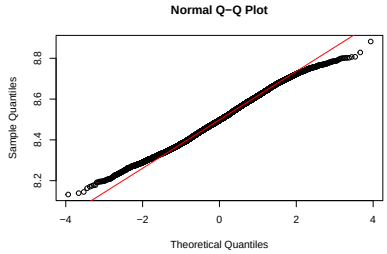
Therefore, we explored an alternative transformation: $y = \log(x + 3000)$

```r
sd_trans <- log(sd + 3000)

hist(
  sd_trans,
  breaks = 80,
  main = "Histogram of log(sd)",
  xlab = "log(sd)",
  probability = TRUE
)

curve(
  dnorm(x, mean = mean(sd_trans), sd = sd(sd_trans)),
  col = "red",
  lwd = 2,
  add = TRUE
)
```



And we checked the Q-Q plot of the transformed data.

```r
qqnorm(sd_trans)
qqline(sd_trans, col = "red")
```



Finally, we calculated the descriptive statistics of the transformed data.

```r
describe(sd_trans)
```

```
##    vars     n mean   sd median trimmed  mad  min  max range skew kurtosis se
## X1    1 12135  8.5 0.11    8.5     8.5 0.12 8.13 8.88  0.75 0.07    -0.44  0
```

Data Analysis by Age on Mean Frequency

Data Preparation

```
voice <- read.csv("../../data/gender/balanced_train.csv")
head(voice)
```

```
##   meanfreq       sd   median      Q25      Q75         skew   sp.ent      sfm
## 1 8.153891 8.570102 8.002904 7.328629 9.291727 -0.199356530 3.369166 2.584677
## 2 7.846562 8.423659 7.689777 7.146453 9.074857 -0.007415137 3.253375 2.378387
## 3 7.637648 8.369293 7.497563 6.976269 8.985667 -0.016312126 3.214666 2.115166
## 4 7.542351 8.426862 7.100093 6.743659 8.928714 -0.054684730 3.160715 2.180566
## 5 7.681082 8.358729 7.607353 7.034379 8.995721 -0.124070090 3.151379 2.457708
## 6 7.584942 8.456333 6.927504 6.782198 8.959107 -0.167355900 3.146243 2.412977
##    meanfun gender
## 1 3.817343   male
## 2 3.183698   male
## 3 3.052549   male
## 4 2.337924   male
## 5 2.251824   male
## 6 2.393773   male
```

```
male_data <- voice[voice$gender == "male", ]
female_data <- voice[voice$gender == "female", ]
head(male_data)
```

```
##   meanfreq       sd   median      Q25      Q75         skew   sp.ent      sfm
## 1 8.153891 8.570102 8.002904 7.328629 9.291727 -0.199356530 3.369166 2.584677
## 2 7.846562 8.423659 7.689777 7.146453 9.074857 -0.007415137 3.253375 2.378387
## 3 7.637648 8.369293 7.497563 6.976269 8.985667 -0.016312126 3.214666 2.115166
## 4 7.542351 8.426862 7.100093 6.743659 8.928714 -0.054684730 3.160715 2.180566
## 5 7.681082 8.358729 7.607353 7.034379 8.995721 -0.124070090 3.151379 2.457708
## 6 7.584942 8.456333 6.927504 6.782198 8.959107 -0.167355900 3.146243 2.412977
##    meanfun gender
## 1 3.817343   male
## 2 3.183698   male
## 3 3.052549   male
## 4 2.337924   male
## 5 2.251824   male
## 6 2.393773   male
```

```
head(female_data)
```

```
##     meanfreq       sd   median      Q25      Q75       skew   sp.ent      sfm
## 726 8.166690 8.581521 8.005468 7.456112 9.171794  0.1165690 3.291912 2.286142
## 727 8.340455 8.599874 8.367010 7.550676 9.296616  0.0456077 3.355465 2.494345
## 728 8.056571 8.546849 7.867097 7.333423 9.110954 -0.1614750 3.272155 2.239860
## 729 8.267929 8.654711 7.734899 7.329590 9.392445  0.1786457 3.301512 2.239053
## 730 7.695245 8.401543 7.485312 6.902178 9.032689  0.2401945 3.160920 2.571130
```

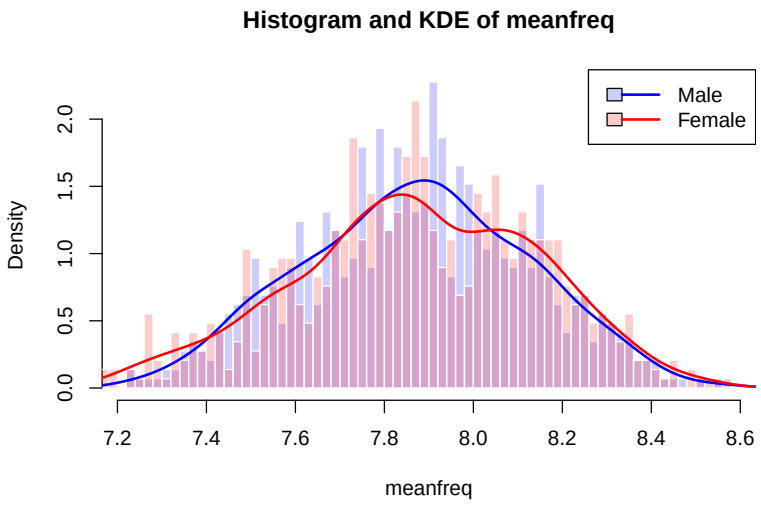
```
## 731 8.058577 8.526025 7.825418 7.218233 9.224732  0.2416200 3.262145 2.565457
##      meanfun gender
## 726 3.564867 female
## 727 3.819150 female
## 728 3.372699 female
## 729 3.160863 female
## 730 2.638777 female
## 731 3.231355 female
```

Visualizing the data

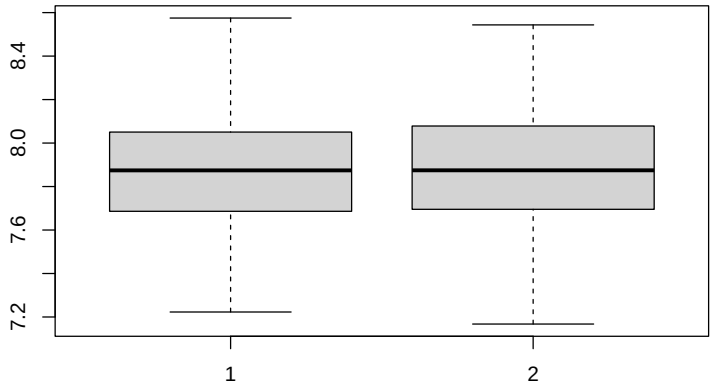
```
visualize_data <- function(column) {
  # return(male_data[column])
  hist(
    male_data[[column]],
    xlab = column,
    col = MALE_COLOR,
    prob = TRUE,
    breaks = 80,
    border = "white",
    main = sprintf("Histogram and KDE of %s", column)
  )
  hist(
    female_data[[column]],
    xlab = column,
    col = FEMALE_COLOR,
    prob = TRUE,
    add = TRUE,
    breaks = 80,
    border = "white"
  )

  # Calculate and plot KDE for male data
  male_density <- density(male_data[[column]])
  lines(male_density, col = "blue", lwd = 2)

  # Calculate and plot KDE for female data
  female_density <- density(female_data[[column]])
  lines(female_density, col = "red", lwd = 2)

  legend(
    "topright",
    legend = c("Male", "Female"),
    col = c("blue", "red"),
    lwd = 2,
    fill = c(MALE_COLOR, FEMALE_COLOR)
  )
}
```

We first visualize the data by plotting the histogram.

```r
visualize_data("meanfreq")
```



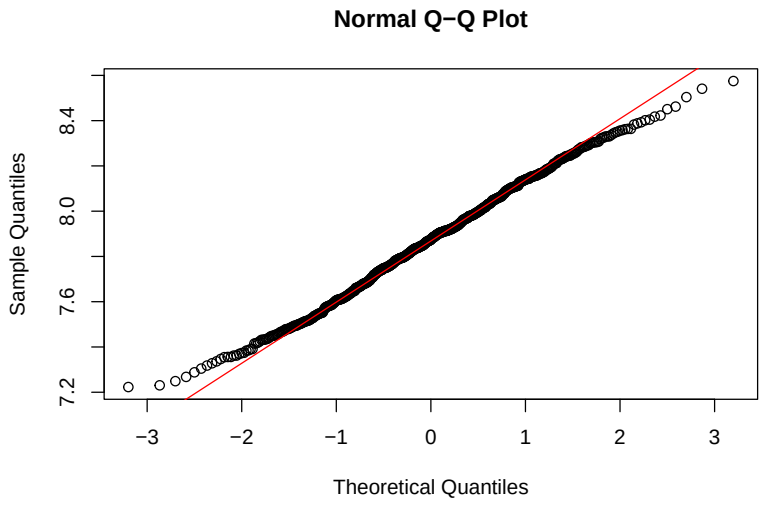

```r
variable <- "meanfreq"
boxplot(male_data[[variable]], female_data[[variable]])
```



QQ-plot

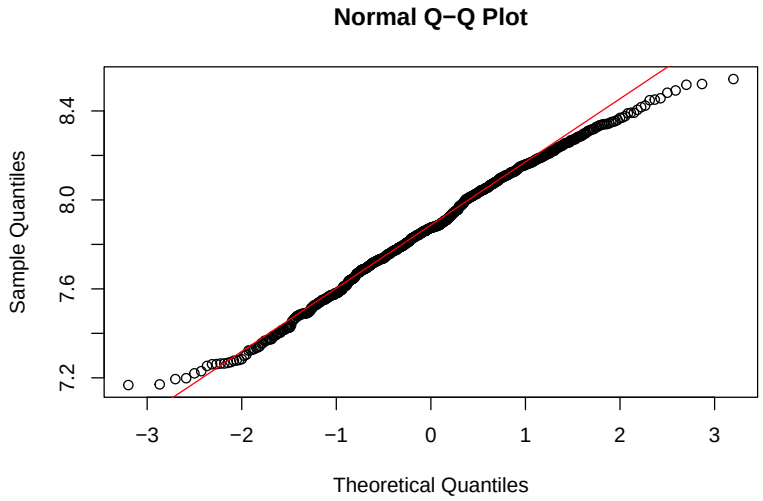
We then plot the QQ-plot to check for normality

```
qqnorm(male_data$meanfreq)
qqline(male_data$meanfreq, col = "red")
```




```
qqnorm(female_data$meanfreq)
qqline(female_data$meanfreq, col = "red")
```

5




```
shapiro.test(male_data$meanfreq)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  male_data$meanfreq
## W = 0.99583, p-value = 0.04956
```

```
shapiro.test(female_data$meanfreq)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  female_data$meanfreq
## W = 0.99301, p-value = 0.0018
```

Based on the QQ-plot, we can see that the data is normally distributed. We would therefore use the F test to compare the variance of the data

F-test

```
var.test(male_data$meanfreq, female_data$meanfreq)
```

```
##
##  F test to compare two variances
##
## data:  male_data$meanfreq and female_data$meanfreq
## F = 0.86113, num df = 724, denom df = 724, p-value = 0.04445
```

```
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7443040 0.9962966
## sample estimates:
## ratio of variances
##          0.8611315
```

Since the p-value is less than 0.05, we reject the null hypothesis that the variance of the data is the same, we would therefore use the two sample t-test with unequal variance

Two Sample T-test

```
t.test(male_data$meanfreq, female_data$meanfreq, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  male_data$meanfreq and female_data$meanfreq
## t = -0.19049, df = 1440, p-value = 0.849
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02973444  0.02447059
## sample estimates:
## mean of x mean of y
##  7.870237  7.872869
```

Since the p-value is greater than 0.05, we do not reject the null hypothesis that the mean of the data is the same.