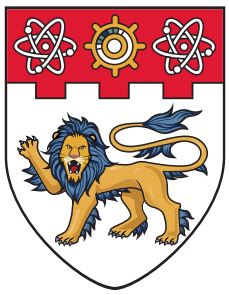


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

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	
Zhang Xintong	XZHANG113@e.ntu.edu.sg	

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

Advancements in artificial intelligence (AI) have revolutionized various domains, including speech synthesis. The ability to generate human-like voices using AI has immense potential applications, from virtual assistants to entertainment media. It's essential to understand the nuances and characteristics of human speech, including variations influenced by gender. By examining datasets encompassing a wide range of voices, we aim to uncover insights into the distinct patterns and distributions of voice frequencies across genders. Though the available voice samples remain unprocessed and unrefined, our objective is to explore the correlation between voice frequency data attributes—such as mean frequency, standard deviation, median frequency, Q25 frequency, Q75 frequency, skewness, mean fundamental frequency—and the gender or age group of the respective voice sample.

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 standard deviation between voices of different genders?
4. Can gender be discerned by examining the quantiles of voice frequency data?

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

*Equal Contribution

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

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

2 Data Preparation

2.1 The DiffVoice Dataset

In order to analyse the relationship between voice features and the information of the speaker, DiffVoice is created for analysis. We collected the raw voice data with captions from Common Voice (cite) English subset, with different genders, regions and ages.

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

1. **Audio Loading**: Raw audio files are loaded and converted into waveforms.
2. **Preprocessing**: Audio is normalized and resampled to a consistent format.
3. **Feature Extraction**: Key features such as Mel-frequency cepstral coefficients (MFCCs), spectral centroid, spectral entropy, spectral flatness, pitch, and magnitude are extracted.
4. **Statistical Aggregation**: Statistical measures like mean, standard deviation, and median are calculated for features extracted.

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.

We organized the dataset into CSV files and uploaded it onto HuggingFace for easier visualization and management.

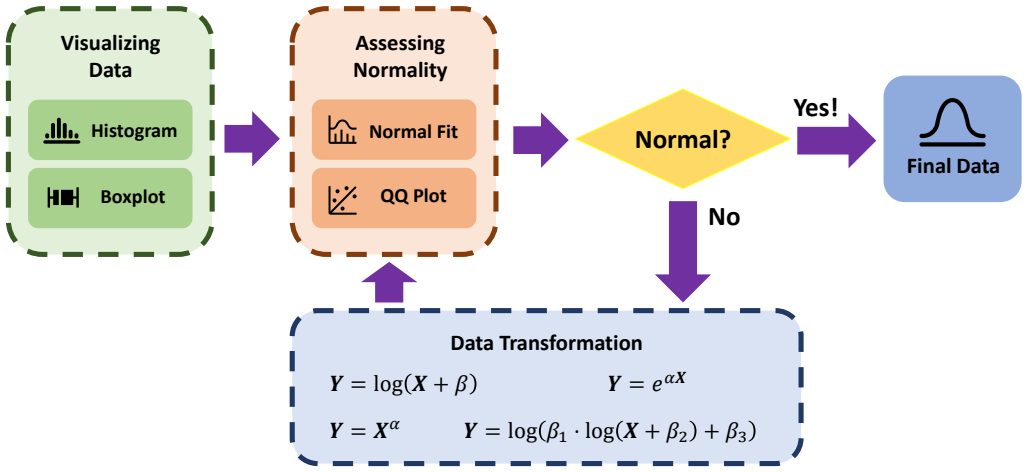


Figure 1: The pipeline for data preparation.

Before proceeding with data analysis, preliminary data cleaning was performed to achieve the following:

1. Irrelevant columns such as "accents" are eliminated from the data set
2. The "country" attribute is dropped as it is out of the scope of our project.

2.2 Data Preparation

The data preparation process starts with:

1. Gaining a basic understanding of the data distribution using histogram. This visual representation easily allows us to assess the skewness or symmetry of these distributions.
2. The mean and standard deviation are then calculated for each attribute, giving us an idea of the central tendency and dispersion of the data.
3. We then proceed to assess the normality of data using QQ-plot, deviation from the reference line indicates that the data is not normal. In that case, we will apply logarithmic transformation to normalize some features of the data.
4. Followed by log transformation, we defined a function `get_outlier` to extract the outliers. We examined the total number of outliers identified across different columns and found that the percentage of outliers is less than 5%. Thus, we decided to remove all the outliers to remove the noise from the data while ensuring the robustness of our analysis.

Table 2. shows a complete summary result of our data preparation stage.

We plot histogram to visualise the variables before and after transformation in Fig. 2

We also plot QQ-plot to visualise the normality of the variables before and after transformation in Fig. 3

2.3 Balancing Data

Upon initial examination, it was evident that the original data is unbalanced, as the proportion of male voice samples is significantly larger than the number of female voice samples. There were also samples categorized under unspecified gender, making it difficult to analyze.

Therefore,we balanced the gender proportions in the dataset to prevent the female data from being treated as a minority and potentially categorized as outliers. To ensure balanced representation of genders, the R script **randomly sample** data from the larger gender group to match the size of the smaller gender group. After combining the sampled rows, the dataset is examined for outliers using

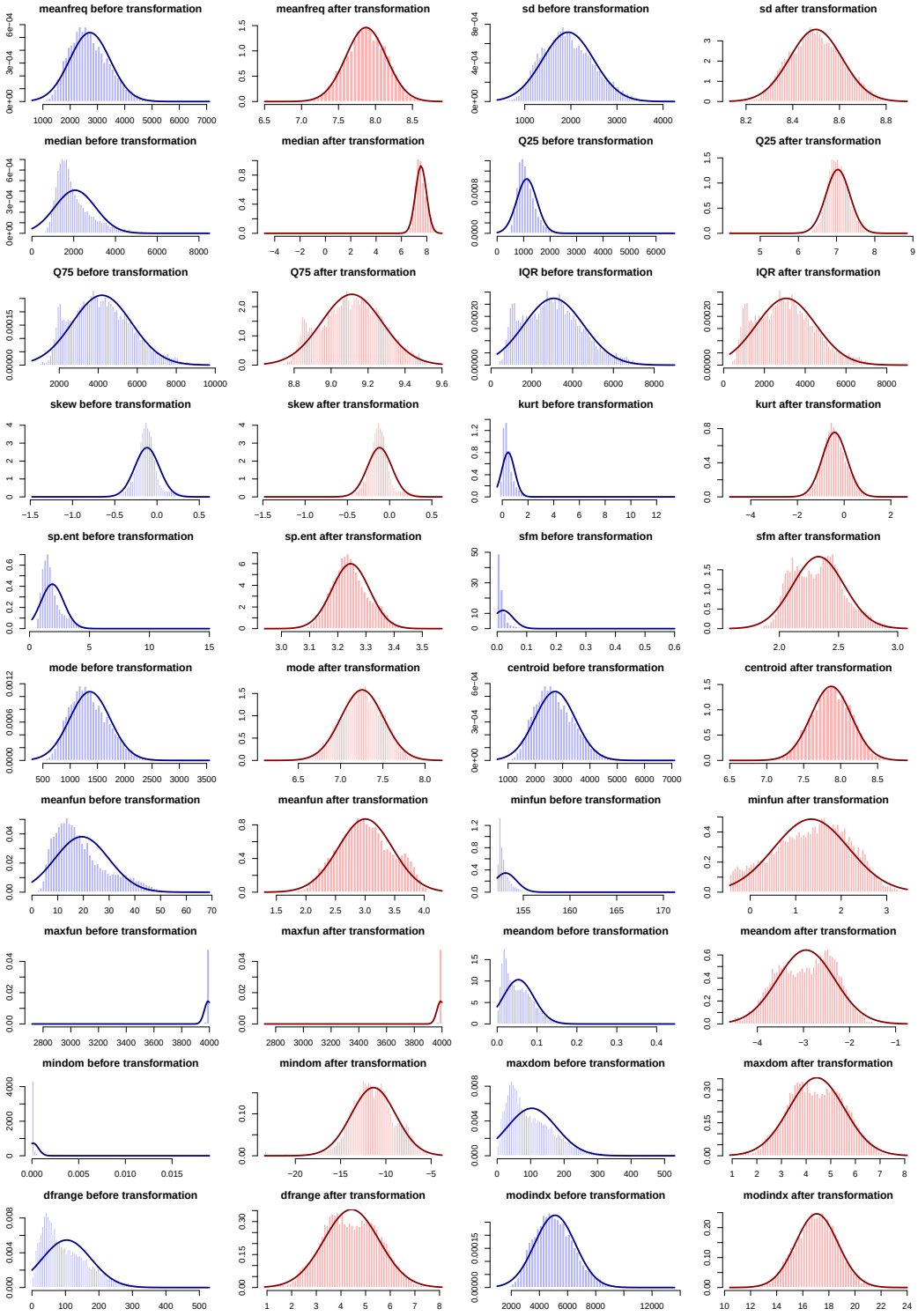


Figure 2: Histogram of variables before and after transformation.

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

the Interquartile Range (IQR) method. This ensures that both genders are adequately represented in the dataset, minimizing the risk of unintentional bias during the cleaning process.

After all the preparation, there are totally 1444 observations from female and male samples. Also, 9 numerical variables and 2 categorical variables (i.e. gender and age) are retained for analysis:

- meanfreq: mean frequency (in kHz)
- sd: standard deviation of frequency
- median: median frequency (in kHz)
- Q25: first quantile (in kHz)
- Q75: third quantile (in kHz)
- skew: skewness
- sp.ent: spectral entropy
- sfm: spectral flatness
- meanfun: average of fundamental frequency measured across acoustic signal

3 Data Analysis and Testing

3.1 Data Analysis By Gender

General steps:

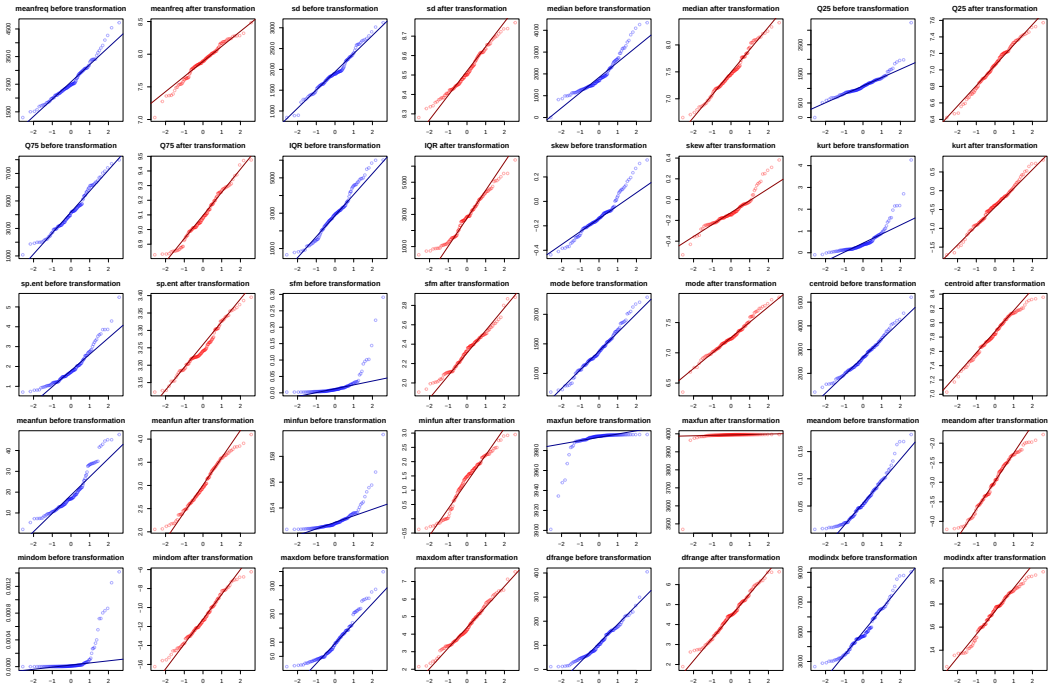


Figure 3: QQ-plot of variables before and after transformation.

- Visualizing the data: We begin by creating a histogram to visually inspect the distribution of the data.
- Checking for normality: Next, we generate a QQ-plot to show whether the data follows a normal distribution.

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

- 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 assuming unequal variances.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 "alt = less" in the Wilcoxon test, we can determine which group has a smaller mean.

The histogram of 9 variables of both genders are given below in Fig. 4:

The histogram and KDEs shows noticeable differences between male and femore voices across various attributes. In general, we can observe lower frequencies (mean, median, Q25, Q75) in male voices and higher in female voices. Skewness and spectral properties like flatness and entropy also differ between those two genders.

The QQ-plot of 9 variables of male dataset are given in Fig. 5:

The QQ-plot of 9 variables of female dataset are given in Fig. 6:

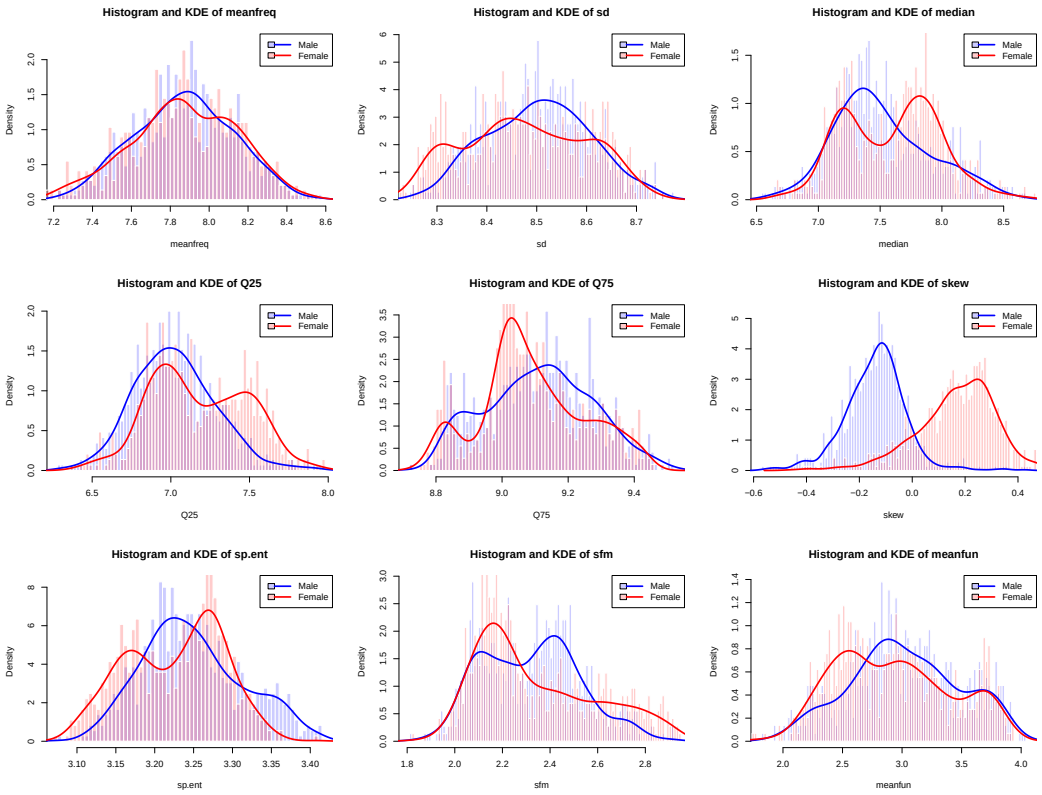


Figure 4: Histogram of variables in both genders.

3.2 Data Analysis By Age

General steps:

- Visualizing the data: **Boxplots** are firstly used to illustrate the distribution of mean frequency across different age groups. Each boxplot displays the median, quartiles, 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: **Kruskal-Wallis tests** are conducted to compare the mean frequency among different age groups to identify significant differences.
- **Pairwise Wilcoxon tests**: If significant differences are found, pairwise Wilcoxon tests are conducted to determine which age groups exhibit differences.

4 Correlation

We visualise the correlation between variables in Fig. 7

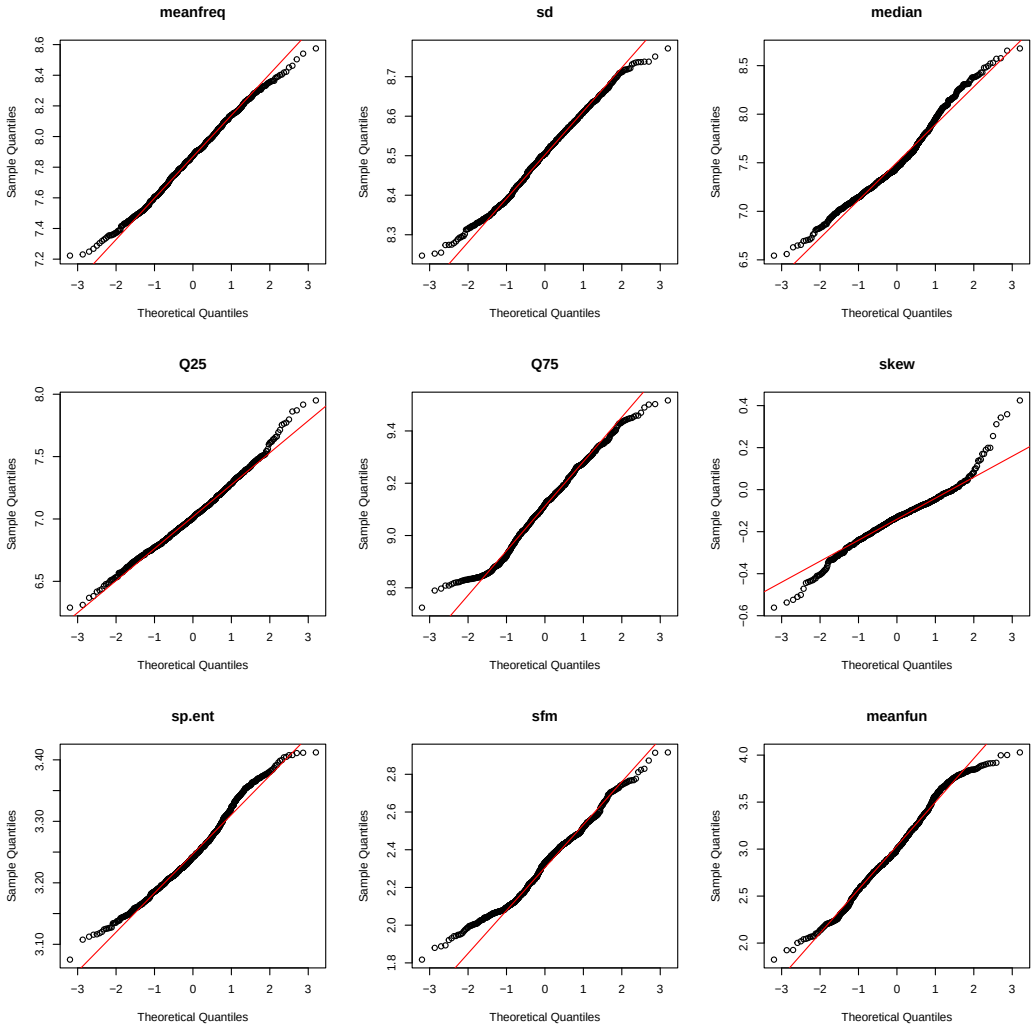


Figure 5: QQ-plot of male dataset.

5 Regression Analysis

5.1 Simple Linear Regression

5.2 Multiple Linear Regression

5.3 Logistic Regression

5.4 K Nearest Neighbour

6 Submission of papers to NeurIPS 2024

Please read the instructions below carefully and follow them faithfully.

6.1 Style

Papers to be submitted to NeurIPS 2024 must be prepared according to the instructions presented here. Papers may only be up to **nine** pages long, including figures. Additional pages *containing only*

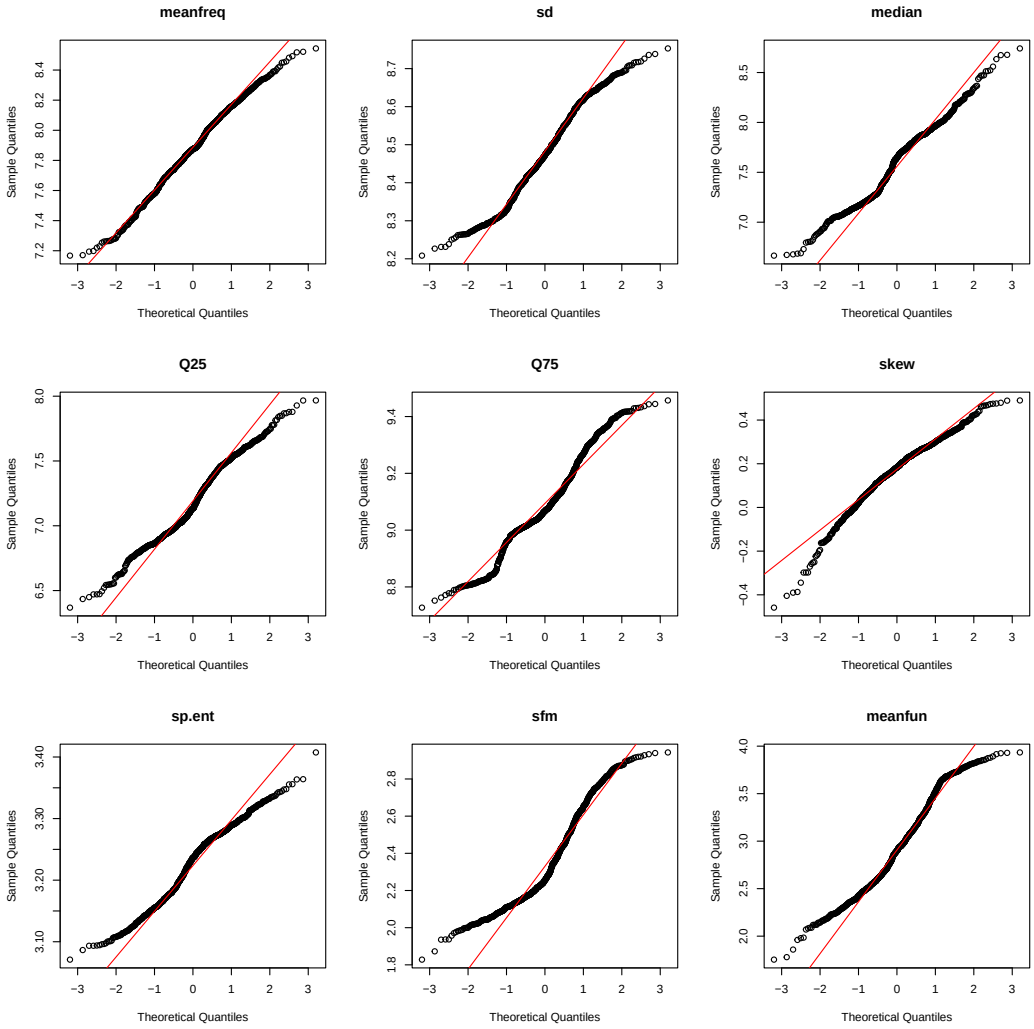


Figure 6: QQ-plot of female dataset.

acknowledgments and references are allowed. Papers that exceed the page limit will not be reviewed, or in any other way considered for presentation at the conference.

The margins in 2024 are the same as those in previous years.

Authors are required to use the NeurIPS LaTeX style files obtainable at the NeurIPS website as indicated below. Please make sure you use the current files and not previous versions. Tweaking the style files may be grounds for rejection.

6.2 Retrieval of style files

The style files for NeurIPS and other conference information are available on the website at

`http://www.neurips.cc/`

The file `neurips_2024.pdf` contains these instructions and illustrates the various formatting requirements your NeurIPS paper must satisfy.

The only supported style file for NeurIPS 2024 is `neurips_2024.sty`, rewritten for LaTeX 2_ε. **Previous style files for LaTeX 2.09, Microsoft Word, and RTF are no longer supported!**

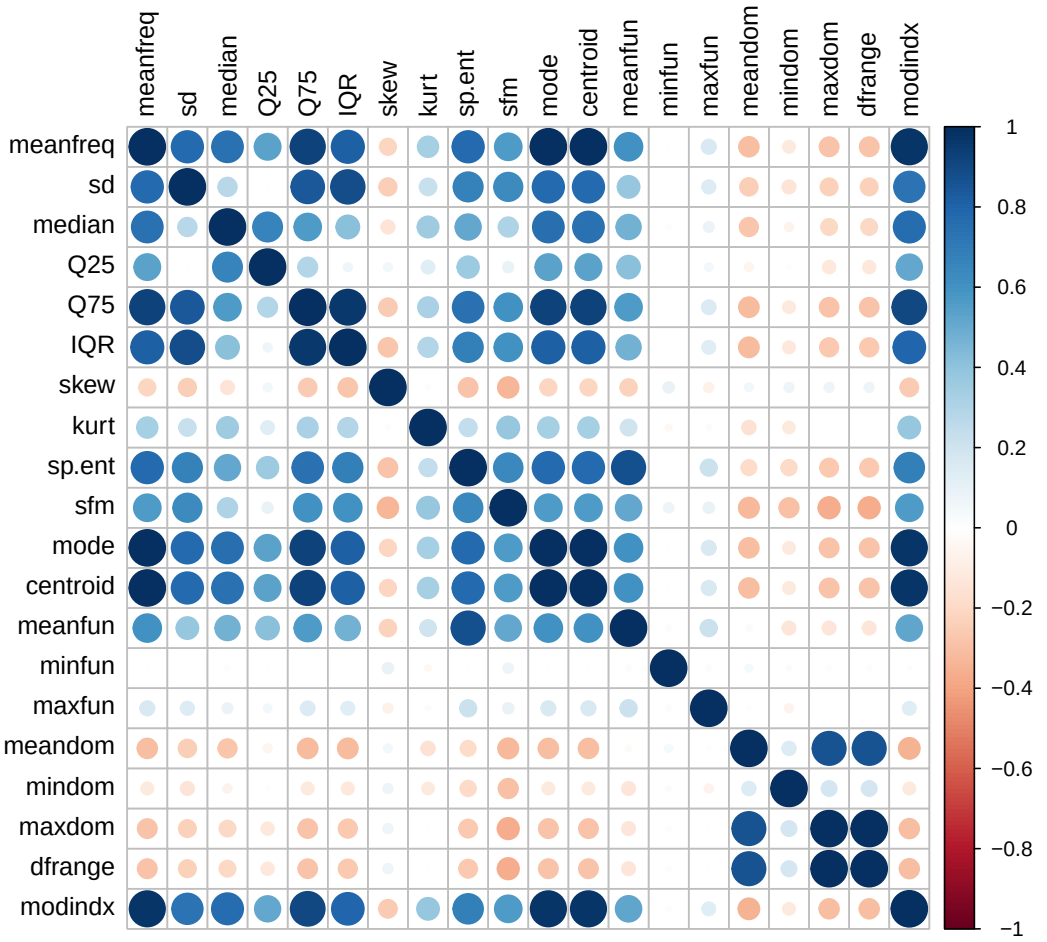


Figure 7: Correlation Matrix.

The LaTeX style file contains three optional arguments: `final`, which creates a camera-ready copy, `preprint`, which creates a preprint for submission to, e.g., arXiv, and `nonatbib`, which will not load the `natbib` package for you in case of package clash.

Preprint option If you wish to post a preprint of your work online, e.g., on arXiv, using the NeurIPS style, please use the `preprint` option. This will create a nonanonymized version of your work with the text "Preprint. Work in progress." in the footer. This version may be distributed as you see fit, as long as you do not say which conference it was submitted to. Please **do not** use the `final` option, which should **only** be used for papers accepted to NeurIPS.

At submission time, please omit the `final` and `preprint` options. This will anonymize your submission and add line numbers to aid review. Please do *not* refer to these line numbers in your paper as they will be removed during generation of camera-ready copies.

The file `neurips_2024.tex` may be used as a "shell" for writing your paper. All you have to do is replace the author, title, abstract, and text of the paper with your own.

The formatting instructions contained in these style files are summarized in Sections 7, 8, and 9 below.

7 General formatting instructions

The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long. The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing (leading) of 11 points. Times New Roman is the preferred typeface throughout, and will be selected for you by default. Paragraphs are separated by ½ line space (5.5 points), with no indentation.

The paper title should be 17 point, initial caps/lower case, bold, centered between two horizontal rules. The top rule should be 4 points thick and the bottom rule should be 1 point thick. Allow ¼ inch space above and below the title to rules. All pages should start at 1 inch (6 picas) from the top of the page.

For the final version, authors' names are set in boldface, and each name is centered above the corresponding address. The lead author's name is to be listed first (left-most), and the co-authors' names (if different address) are set to follow. If there is only one co-author, list both author and co-author side by side.

Please pay special attention to the instructions in Section 9 regarding figures, tables, acknowledgments, and references.

8 Headings: first level

All headings should be lower case (except for first word and proper nouns), flush left, and bold.

First-level headings should be in 12-point type.

8.1 Headings: second level

Second-level headings should be in 10-point type.

8.1.1 Headings: third level

Third-level headings should be in 10-point type.

Paragraphs There is also a `\paragraph` command available, which sets the heading in bold, flush left, and inline with the text, with the heading followed by 1 em of space.

9 Citations, figures, tables, references

These instructions apply to everyone.

9.1 Citations within the text

The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as long as you maintain internal consistency. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

The documentation for `natbib` may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
\citet{hasselmo} investigated\dots
```

produces

> Hasselmo, et al. (1995) investigated…

If you wish to load the `natbib` package with options, you may add the following before loading the `neurips_2024` package:

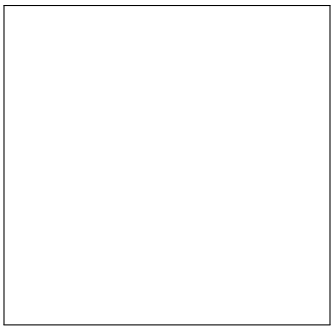


Figure 8: Sample figure caption.

`\PassOptionsToPackage{options}{natbib}`

If `natbib` clashes with another package you load, you can add the optional argument `nonatbib` when loading the style file:

`\usepackage[nonatbib]{neurips_2024}`

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of Jones et al. [4]," not "In our previous work [4]." If you cite your other papers that are not widely available (e.g., a journal paper under review), use anonymous author names in the citation, e.g., an author of the form "A. Anonymous" and include a copy of the anonymized paper in the supplementary material.

9.2 Footnotes

Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number[2] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).

Note that footnotes are properly typeset *after* punctuation marks.[3]

9.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction. The figure number and caption always appear after the figure. Place one line space before the figure caption and one line space after the figure. The figure caption should be lower case (except for first word and proper nouns); figures are numbered consecutively.

You may use color figures. However, it is best for the figure captions and the paper body to be legible if the paper is printed in either black/white or in color.

9.4 Tables

All tables must be centered, neat, clean and legible. The table number and title always appear before the table. See Table 3.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Note that publication-quality tables *do not contain vertical rules.* We strongly suggest the use of the `booktabs` package, which allows for typesetting high-quality, professional tables:

`https://www.ctan.org/pkg/booktabs`

This package was used to typeset Table 3.

[2] Sample of the first footnote.

[3] As in this example.

Table 3: Sample table title

	Part	
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

9.5 Math

Note that display math in bare TeX commands will not create correct line numbers for submission. Please use LaTeX (or AMSTeX) commands for unnumbered display math. (You really shouldn't be using $$ anyway; see `https://tex.stackexchange.com/questions/503/why-is-preferable-to` and `https://tex.stackexchange.com/questions/40492/what-are-the-differences-between-align-equation-and-displaymath` for more information.)

9.6 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

10 Preparing PDF files

Please prepare submission files with paper size "US Letter," and not, for example, "A4."

Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or Embedded TrueType fonts. Here are a few instructions to achieve this.

- You should directly generate PDF files using `pdflatex`.
- You can check which fonts a PDF files uses. In Acrobat Reader, select the menu Files>Document Properties>Fonts and select Show All Fonts. You can also use the program `pdffonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.
- `xfig` "patterned" shapes are implemented with bitmap fonts. Use "solid" shapes instead.
- The `\bbold` package almost always uses bitmap fonts. You should use the equivalent AMS Fonts:

  ```
  \usepackage{amsfonts}
  ```

  followed by, e.g., `\mathbb{R}`, `\mathbb{N}`, or `\mathbb{C}` for $\mathbb{R}$, $\mathbb{N}$ or $\mathbb{C}$. You can also use the following workaround for reals, natural and complex:

  ```
  \newcommand{\RR}{I\!\!R} %real numbers
  \newcommand{\Nat}{I\!\!N} %natural numbers
  \newcommand{\CC}{I\!\!\!\!C} %complex numbers
  ```

  Note that `amsfonts` is automatically loaded by the `amssymb` package.

If your file contains type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

10.1 Margins in LaTeX

Most of the margin problems come from figures positioned by hand using `\special` or other commands. We suggest using the command `\includegraphics` from the `graphicx` package. Always specify the figure width as a multiple of the line width as in the example below:

```
\usepackage[pdftex]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.pdf}
```

See Section 4.4 in the graphics bundle documentation (`http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf`)

A number of width problems arise when LaTeX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command when necessary.

Acknowledgments and Disclosure of Funding

Use unnumbered first level headings for the acknowledgments. All acknowledgments go at the end of the paper before the list of references. Moreover, you are required to declare funding (financial activities supporting the submitted work) and competing interests (related financial activities outside the submitted work). More information about this disclosure can be found at: `https://neurips.cc/Conferences/2024/PaperInformation/FundingDisclosure`.

Do **not** include this section in the anonymized submission, only in the final paper. You can use the `ack` environment provided in the style file to automatically hide this section in the anonymized submission.

References

References follow the acknowledgments in the camera-ready paper. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. Note that the Reference section does not count towards the page limit.

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.

A Appendix / supplemental material

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
```

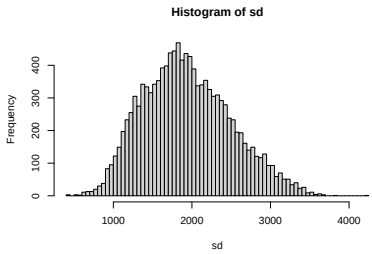
Visualizing the Data

We selected `sd` column to perform the analysis.

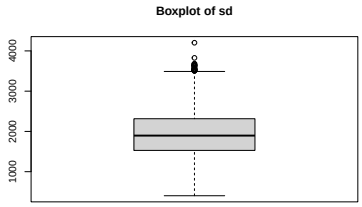
First, we load the data and draw a histogram of the `sd` column to get an initial understanding of the data distribution.

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


Figure 9: Example Code of Data Analysis on `sd`, Page 1




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



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

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

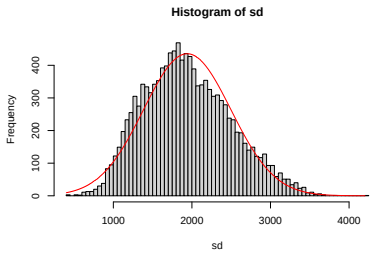
```
##    vars     n   mean     sd  median trimmed    mad    min     max   range skew
## X1    1 12135 1940.3 555.86 1896.32 1918.18 580.72 402.55 4202.62 3800.07 0.33
##    kurtosis   se
## X1    -0.31 5.05
```

Assessing Data Normality

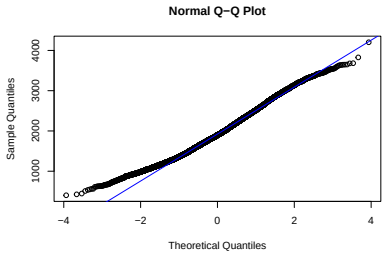
After that, we assess the normality of the `sd` column by drawing a histogram with a normal curve and a Q-Q plot.

```
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")
```


Figure 10: Example Code of Data Analysis on `sd`, Page 2




```
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)")
```

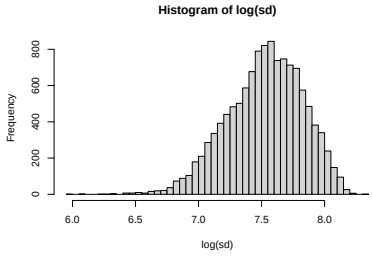
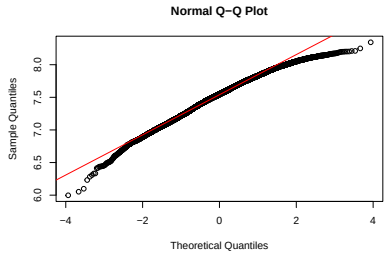


Figure 11: Example Code of Data Analysis on `sd`, Page 3

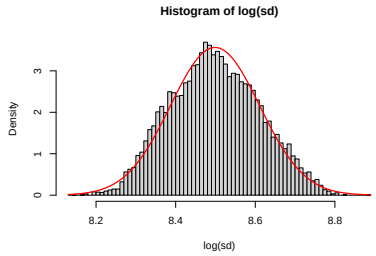
```r
qqnorm(sd_trans)
qqline(sd_trans, col = "red")
```



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: $y = \log(x + 3000)$

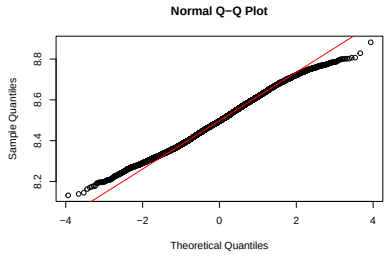
```r
sd_trans <- log(sd + 3000)

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.

Figure 12: Example Code of Data Analysis on `sd`, Page 4

```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
```


Figure 13: Example Code of Data Analysis on `sd`, Page 5