EDA For Task2

Group6

26/11/2019

Load the data

```
data <- read.csv("cleanedData.csv")
```

```
library(ggplot2)
library(ggpubr)
```

```
## Loading required package: magrittr
```

```
theme_set(theme_pubr())
```

Histograms and QQ-Plots for all the variables used in K-N-N

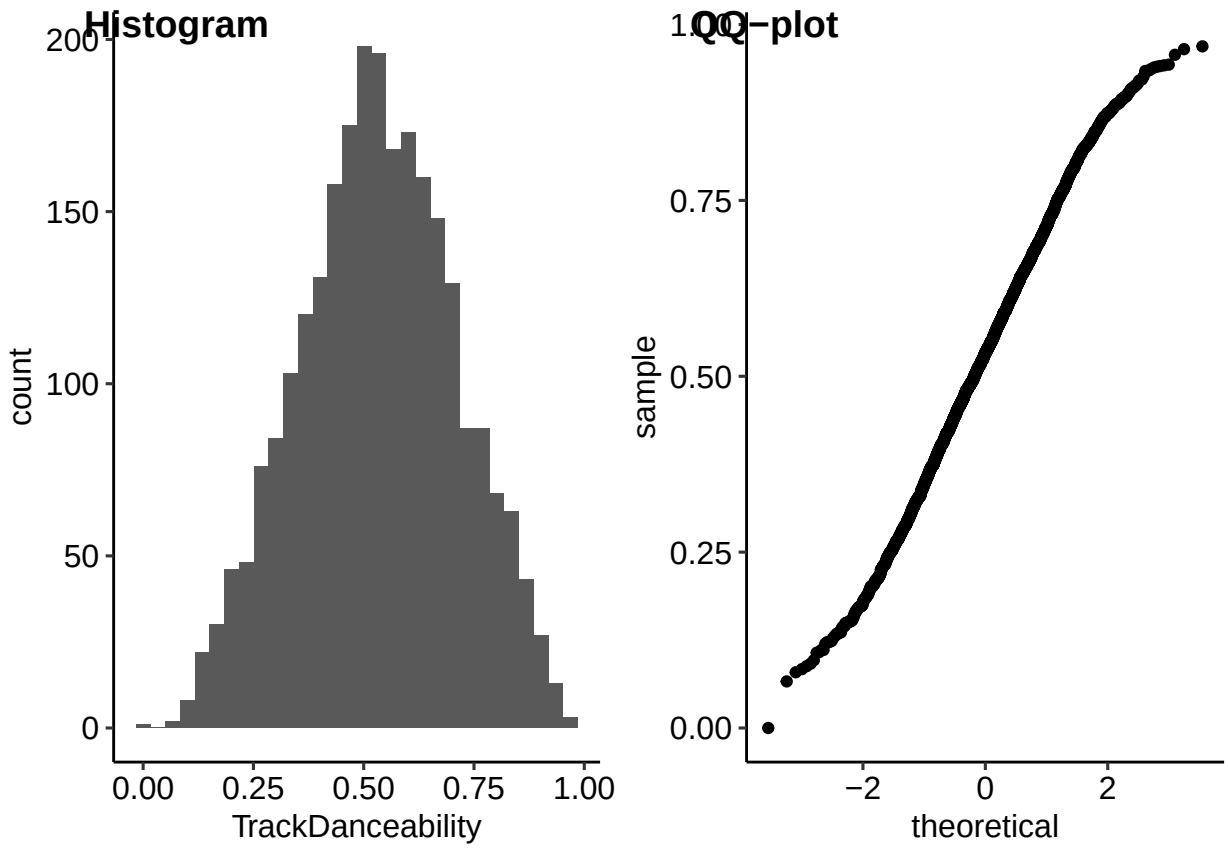
We can use these plots to identify whether the corresponding variables follow Gaussian distribution or not.

TrackDanceability

```
h <- ggplot(data, aes(x=TrackDanceability)) + geom_histogram()
q <- ggplot(data, aes(sample=TrackDanceability))+stat_qq()
figure <- ggarrange(h,q,
                    labels = c("Histogram", "QQ-plot"),
                    ncol = 2, nrow = 1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
figure
```

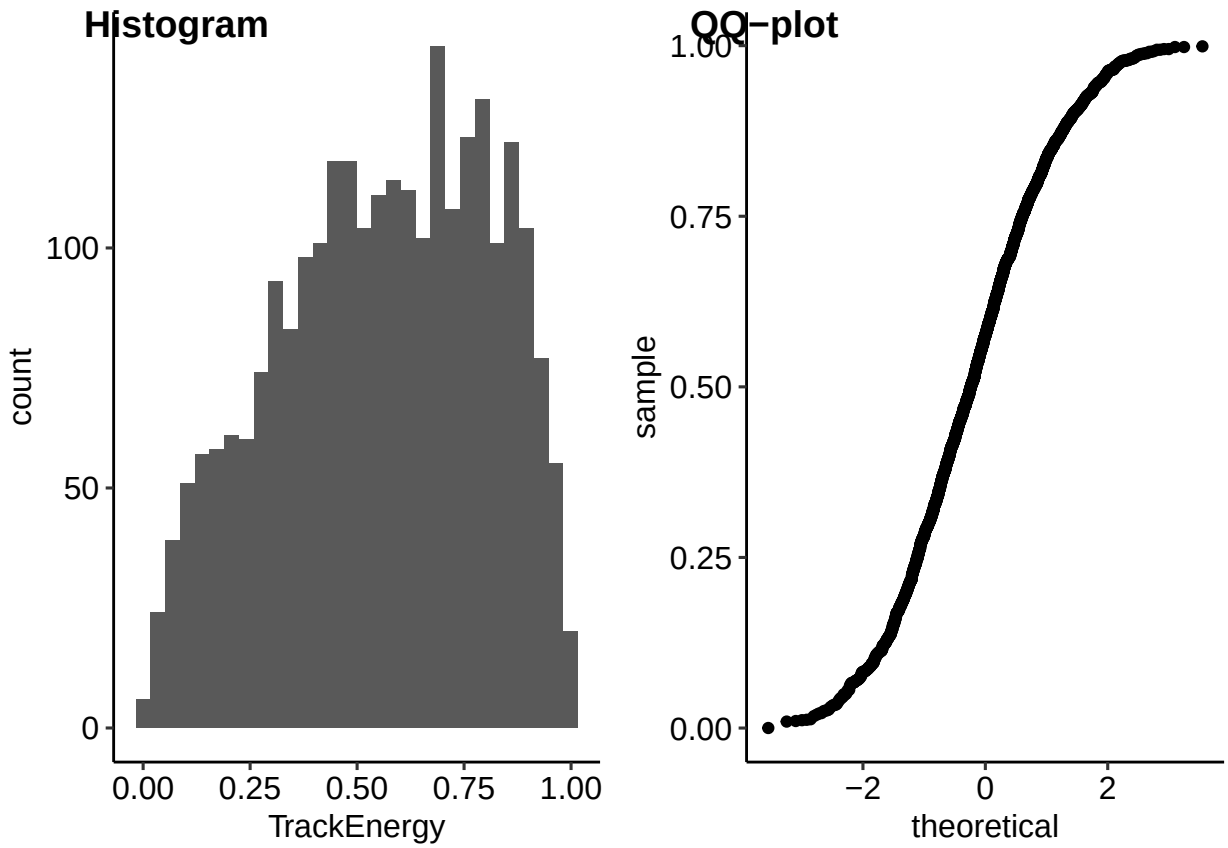


TrackEnergy

```
h <- ggplot(data, aes(x=TrackEnergy)) + geom_histogram()
q <- ggplot(data, aes(sample=TrackEnergy))+stat_qq()
figure <- ggarrange(h,q,
                    labels = c("Histogram", "QQ-plot"),
                    ncol = 2, nrow = 1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
figure
```

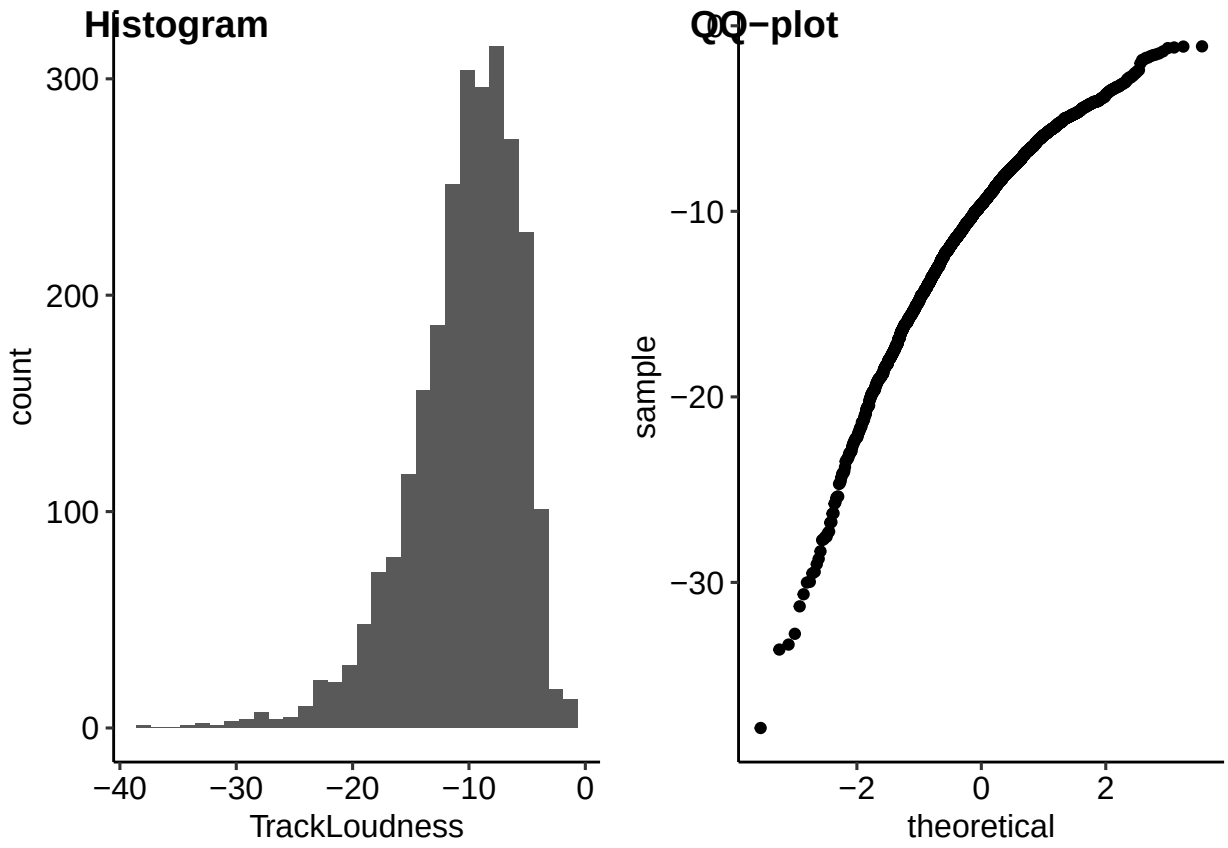


TrackLoudness

```
h <- ggplot(data, aes(x=TrackLoudness)) + geom_histogram()
q <- ggplot(data, aes(sample=TrackLoudness))+stat_qq()
figure <- ggarrange(h,q,
                    labels = c("Histogram", "QQ-plot"),
                    ncol = 2, nrow = 1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
figure
```

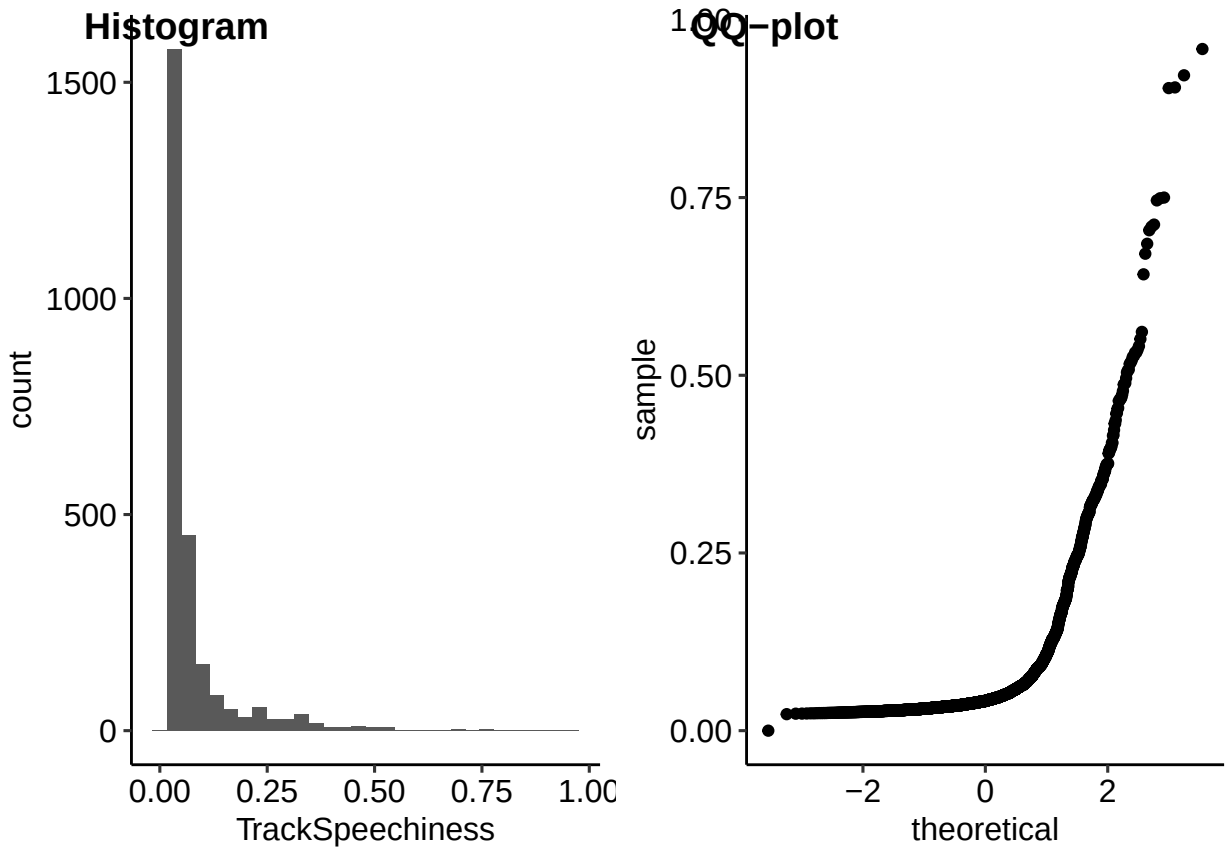


TrackSpeechiness

```
h <- ggplot(data, aes(x=TrackSpeechiness)) + geom_histogram()
q <- ggplot(data, aes(sample=TrackSpeechiness))+stat_qq()
figure <- ggarrange(h,q,
                    labels = c("Histogram", "QQ-plot"),
                    ncol = 2, nrow = 1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
figure
```

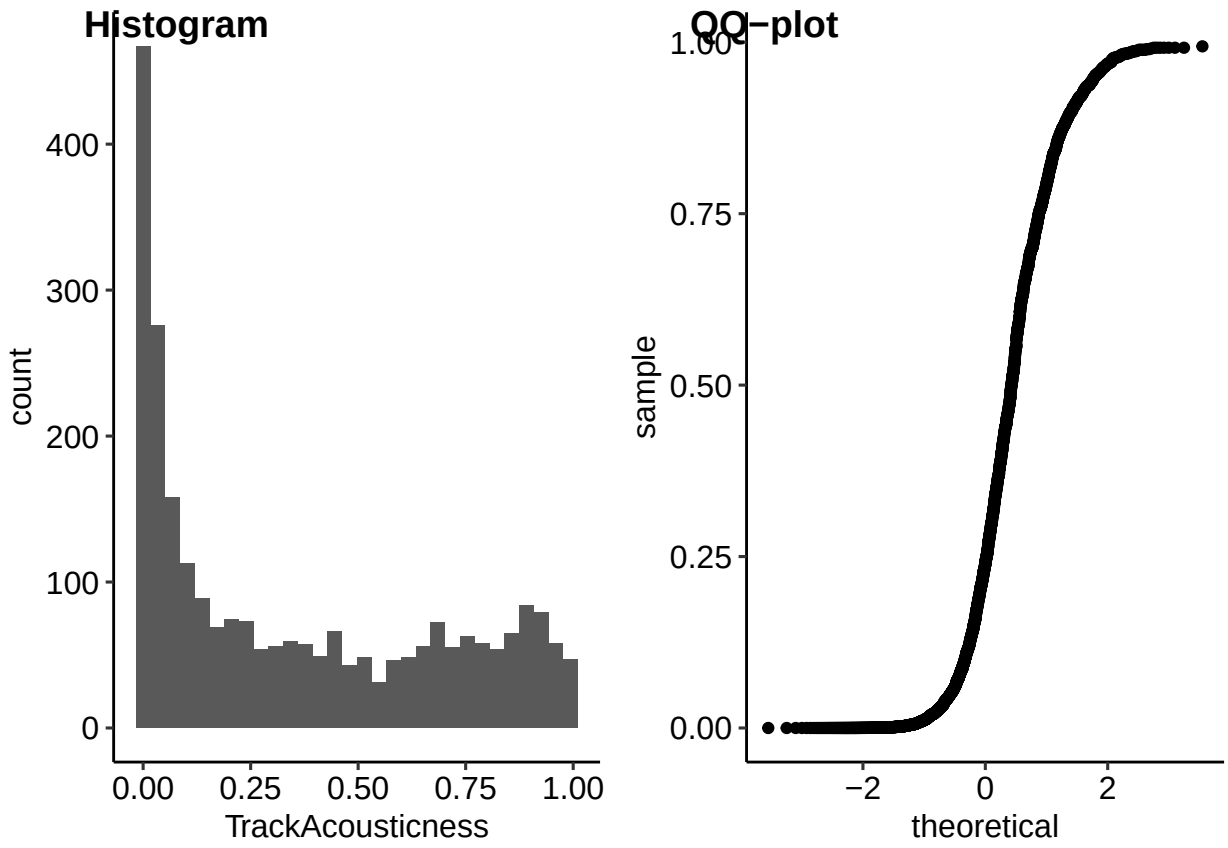


TrackAcousticness

```
h <- ggplot(data, aes(x=TrackAcousticness)) + geom_histogram()
q <- ggplot(data, aes(sample=TrackAcousticness))+stat_qq()
figure <- ggarrange(h,q,
                    labels = c("Histogram", "QQ-plot"),
                    ncol = 2, nrow = 1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
figure
```

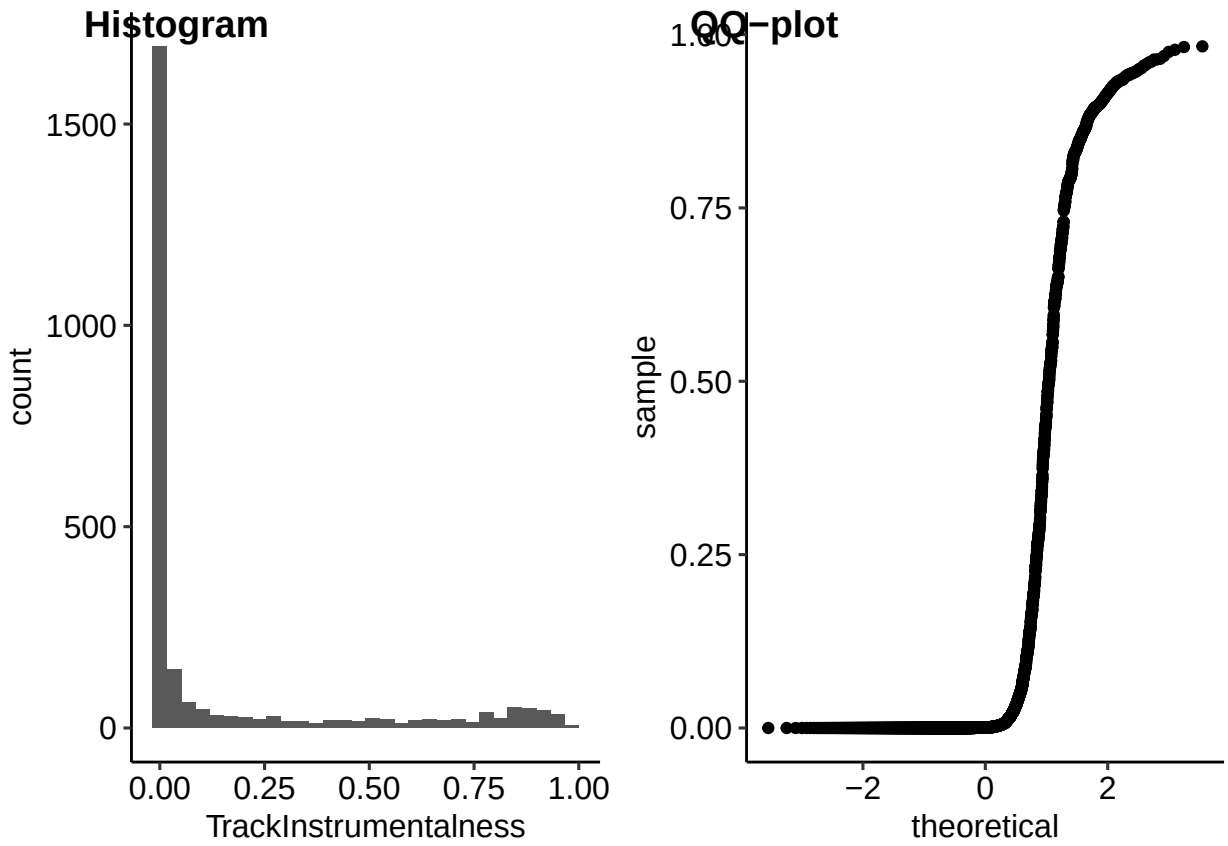


TrackInstrumentalness

```
h <- ggplot(data, aes(x=TrackInstrumentalness)) + geom_histogram()
q <- ggplot(data, aes(sample=TrackInstrumentalness))+stat_qq()
figure <- ggarrange(h,q,
                    labels = c("Histogram", "QQ-plot"),
                    ncol = 2, nrow = 1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
figure
```

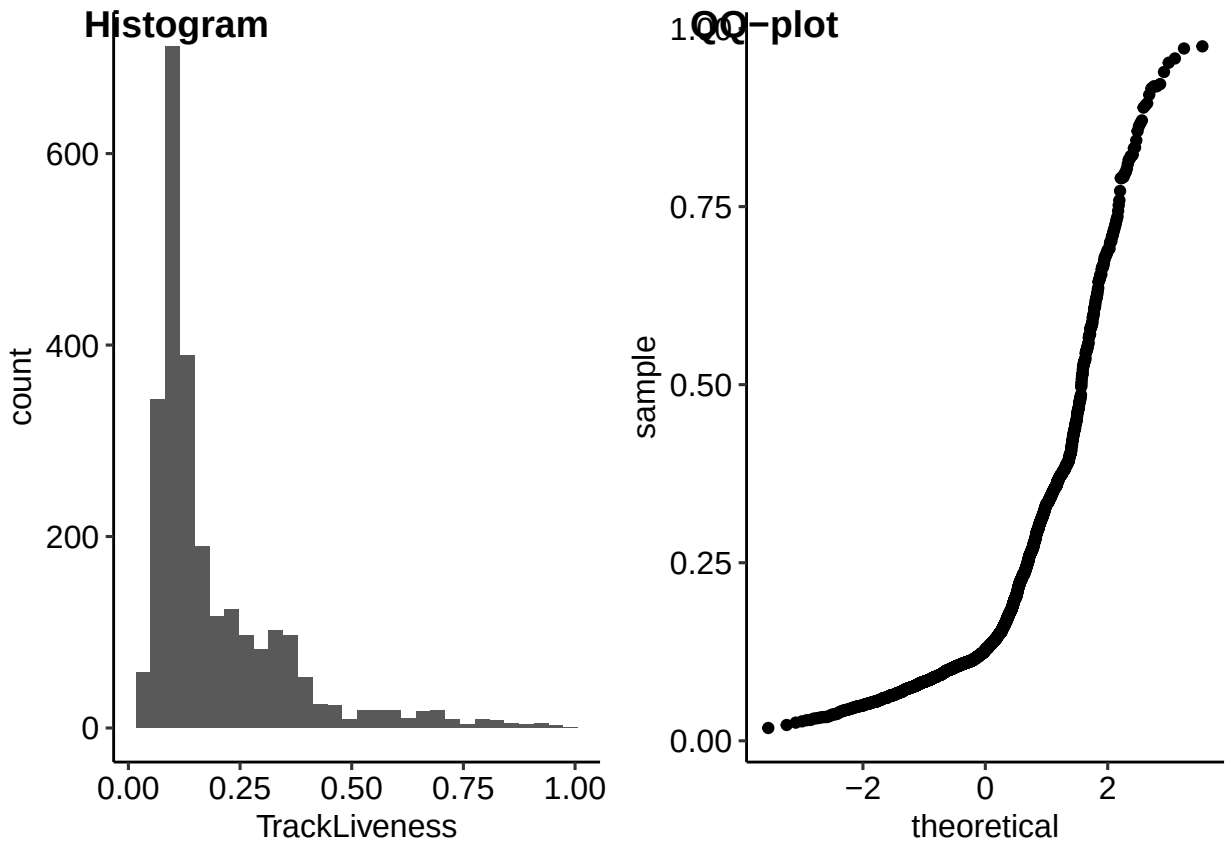


TrackLiveness

```
h <- ggplot(data, aes(x=TrackLiveness)) + geom_histogram()
q <- ggplot(data, aes(sample=TrackLiveness))+stat_qq()
figure <- ggarrange(h,q,
                    labels = c("Histogram", "QQ-plot"),
                    ncol = 2, nrow = 1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
figure
```

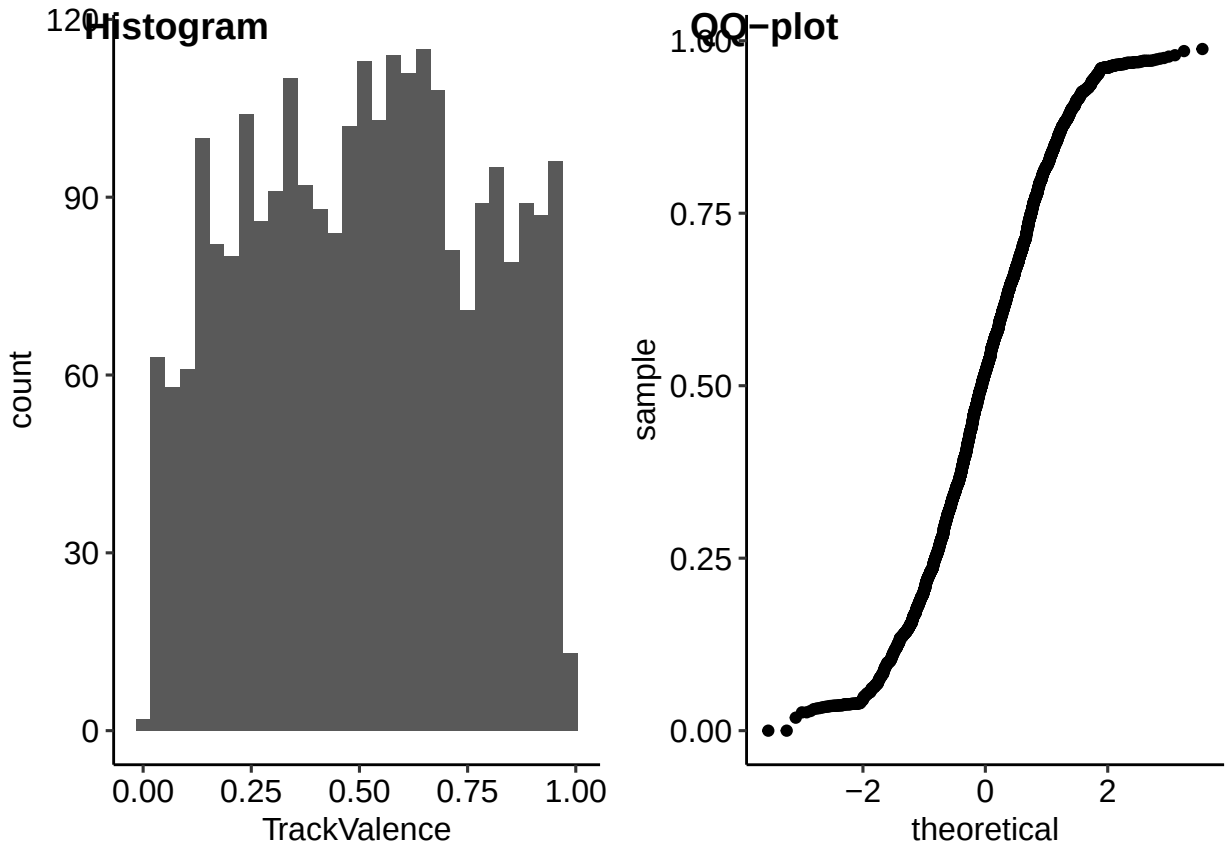


TrackValence

```
h <- ggplot(data, aes(x=TrackValence)) + geom_histogram()
q <- ggplot(data, aes(sample=TrackValence))+stat_qq()
figure <- ggarrange(h,q,
                    labels = c("Histogram", "QQ-plot"),
                    ncol = 2, nrow = 1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
figure
```

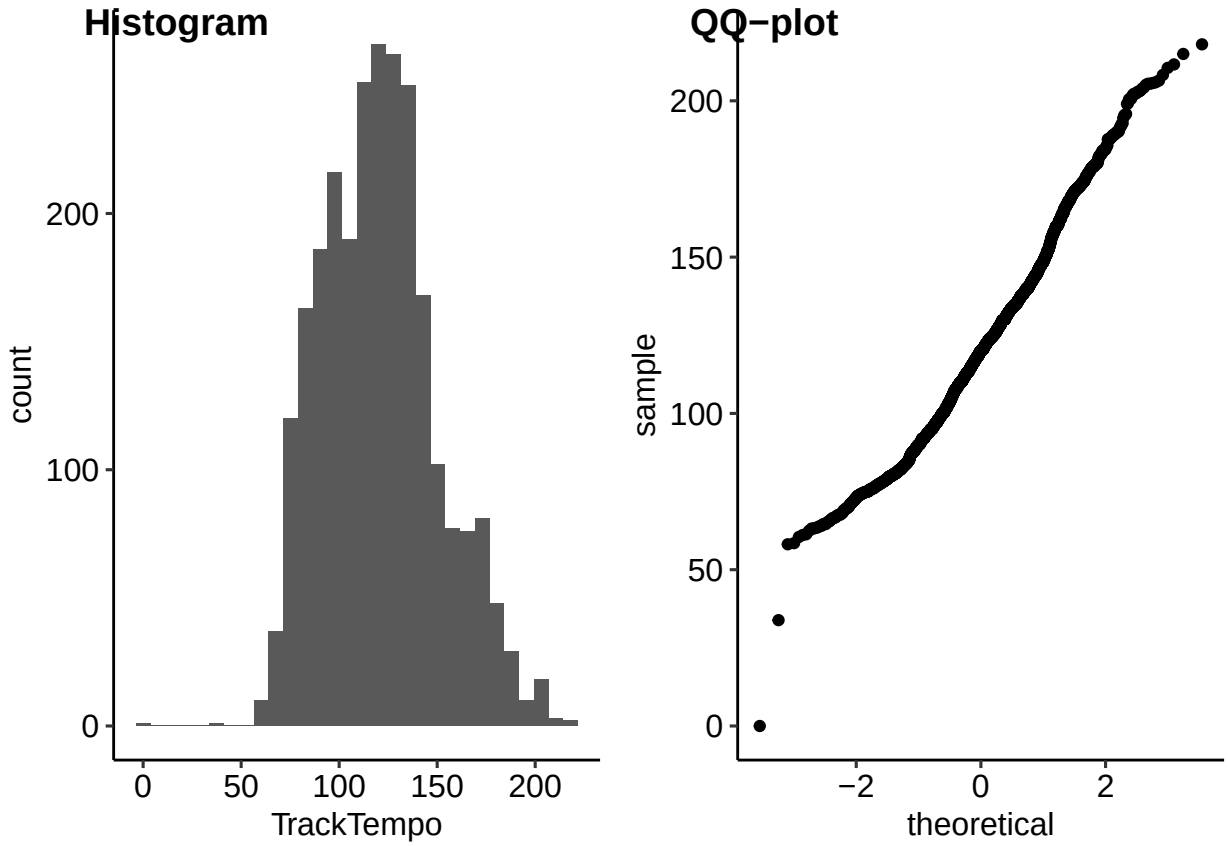


TrackTempo

```
h <- ggplot(data, aes(x=TrackTempo)) + geom_histogram()
q <- ggplot(data, aes(sample=TrackTempo))+stat_qq()
figure <- ggarrange(h,q,
                    labels = c("Histogram", "QQ-plot"),
                    ncol = 2, nrow = 1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
figure
```


Histogram
count
200
100
0
0
50
100
150
200
TrackTempo
QQ−plot
sample
200
150
100
50
0
−2
0
2
theoretical