A Statistical Analysis of Finnish Vowel Duration

Loading data:

read.delim2() is used as the data is in .txt format:

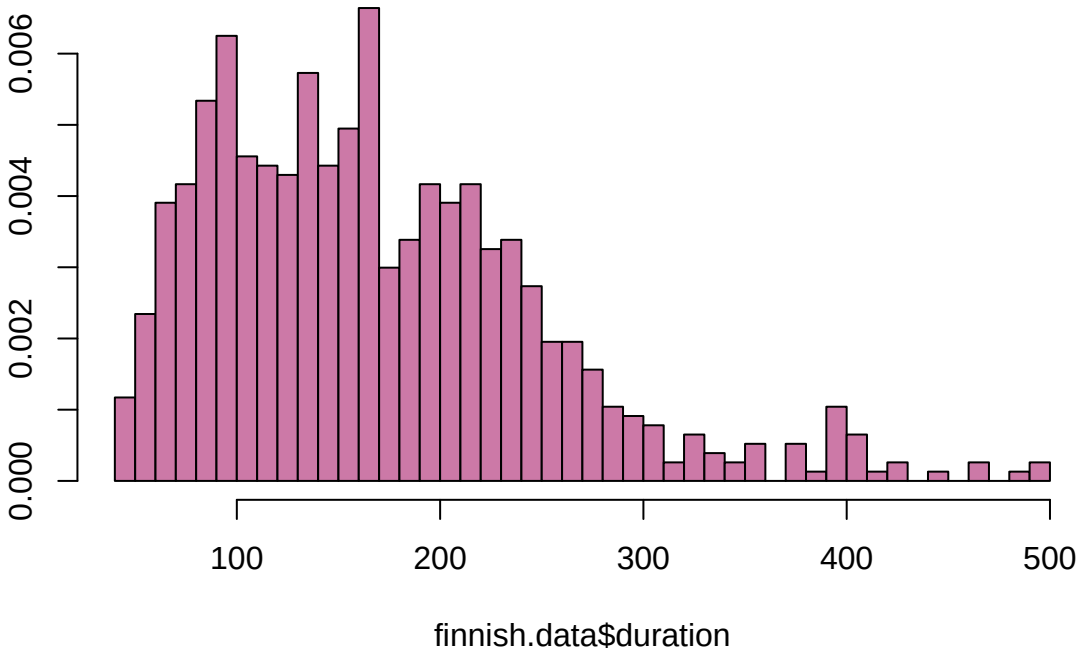
```
##   token duration      length syllaType position    speaker sanoiDur vcdDur
## 1   j1a      109 short vowel        CV    final Speaker #1      385     68
## 2   j1b       78 short vowel        GV    final Speaker #1      374     51
## 3   j1c      145  long vowel        GV nonfinal Speaker #1      344    145
## 4   j1d      162  long vowel        CV nonfinal Speaker #1      333    162
## 5   j1e       88 short vowel         V nonfinal Speaker #1      349     88
## 6   j1f      136 short vowel        GV    final Speaker #1      361     91
```

The devoiced durations can be obtained by subtracting the values of the 'vcdDur' vector from the 'duration' vector:

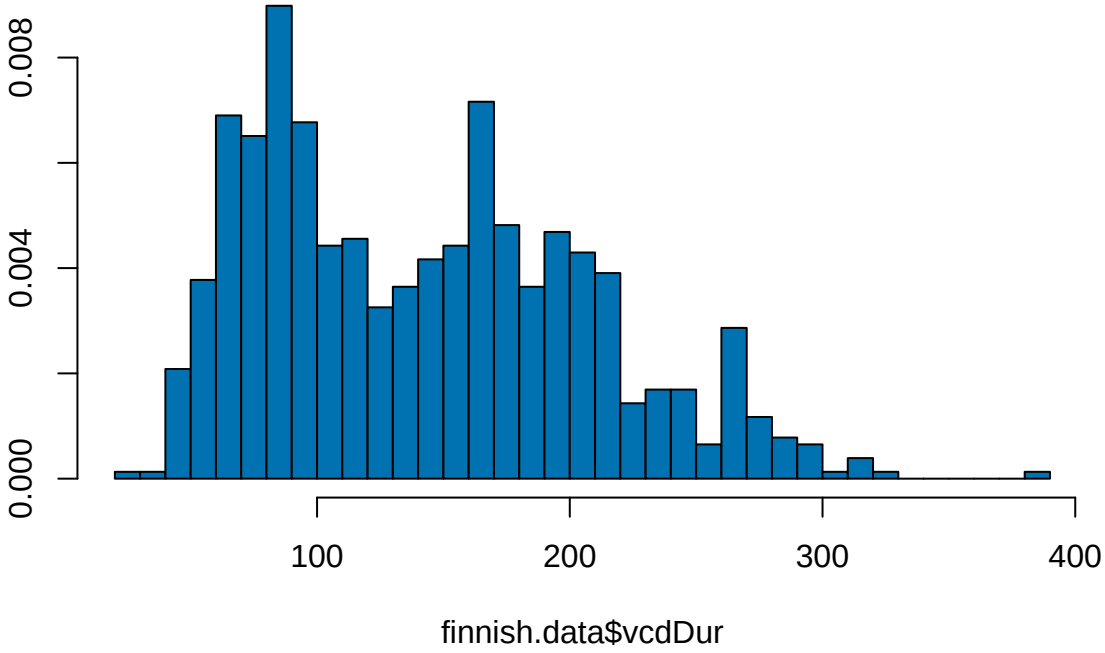
```
##   token duration      length syllaType position    speaker sanoiDur vcdDur
## 1   j1a      109 short vowel        CV    final Speaker #1      385     68
## 2   j1b       78 short vowel        GV    final Speaker #1      374     51
## 3   j1c      145  long vowel        GV nonfinal Speaker #1      344    145
## 4   j1d      162  long vowel        CV nonfinal Speaker #1      333    162
## 5   j1e       88 short vowel         V nonfinal Speaker #1      349     88
## 6   j1f      136 short vowel        GV    final Speaker #1      361     91
##   dvcdDur
## 1      41
## 2      27
## 3       0
## 4       0
## 5       0
## 6      45
```

Plotting Histograms:

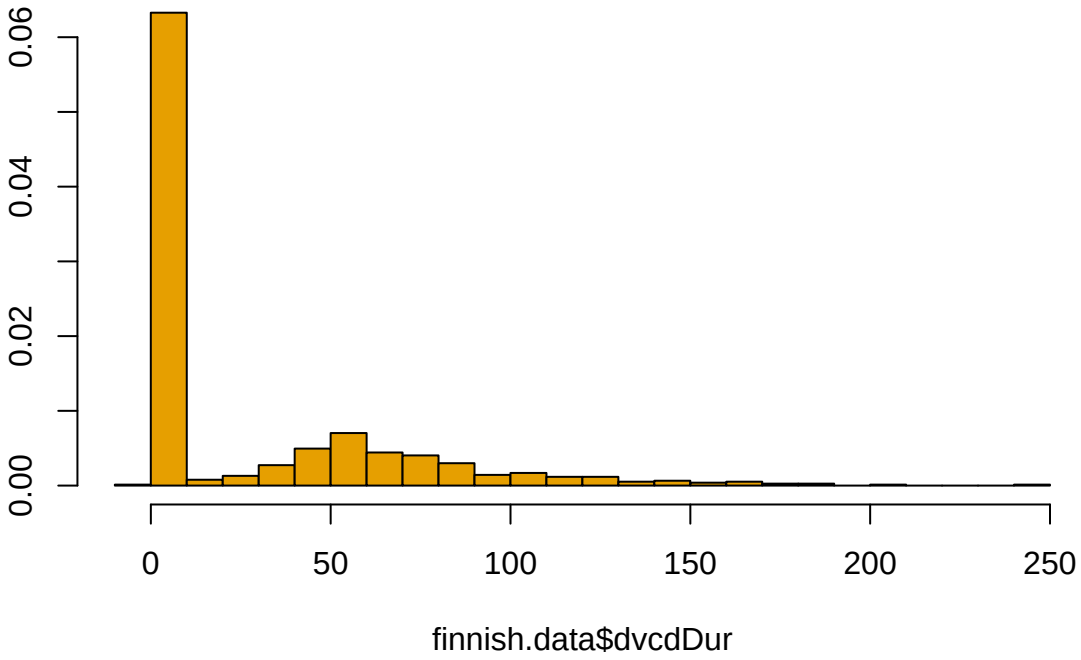
Plot for 'duration' vector:



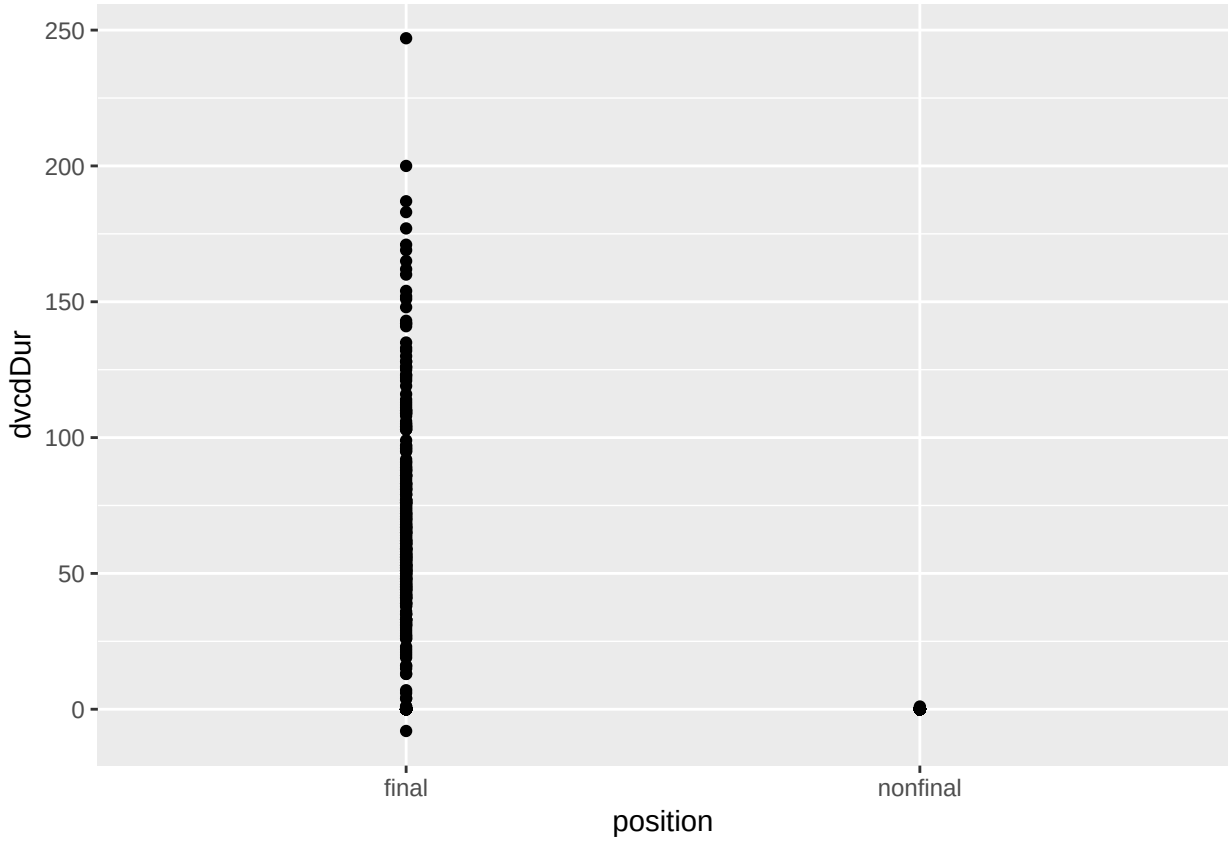
Plot for ‘vcdDur’ vector:



Plot for ‘dvcdDur’ vector:



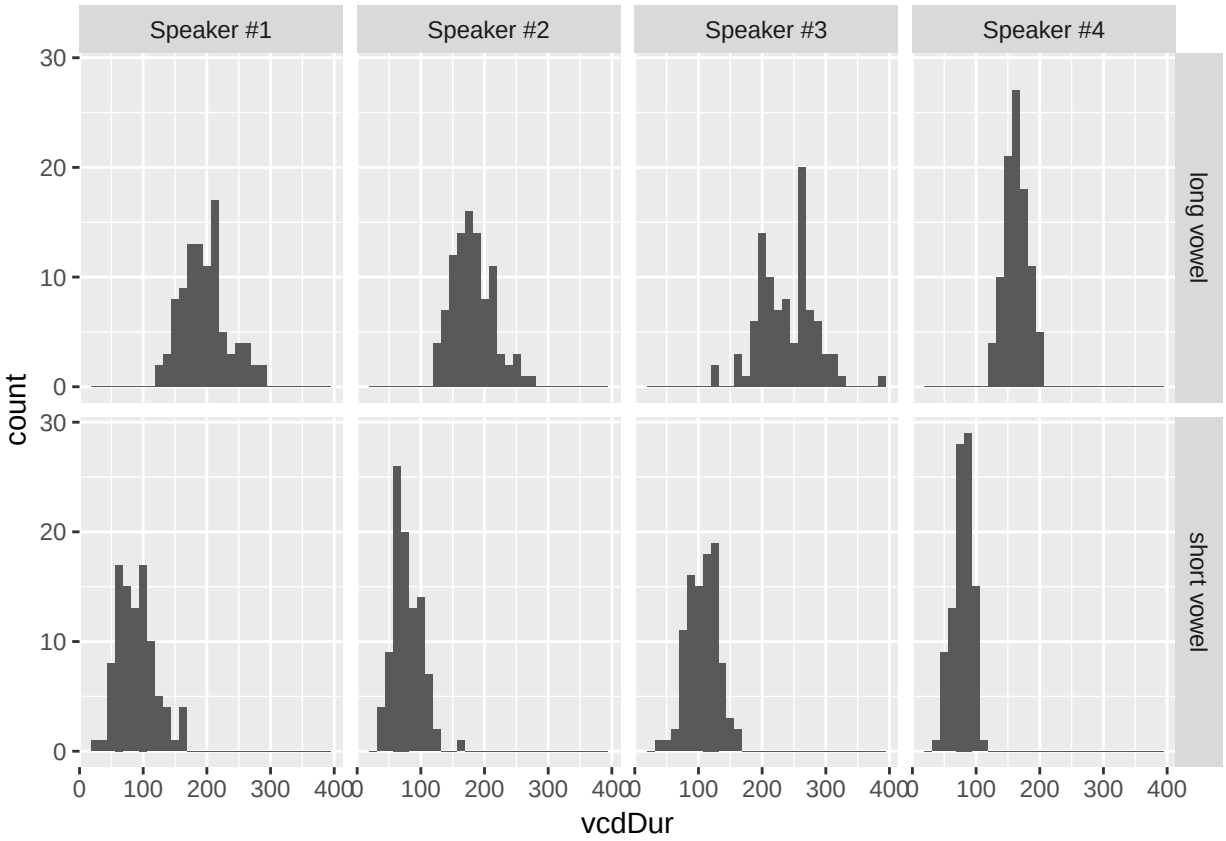
We notice in the plot for dvcdDur that the most common value by far is 0. Looking over the data in finnishVowelDuration.txt, it appears that when vowel position is nonfinal, we tend to get dvcdDur values of 0. This suggests that vowels are devoiced only in the final position in Finnish.



Analysis by speaker:

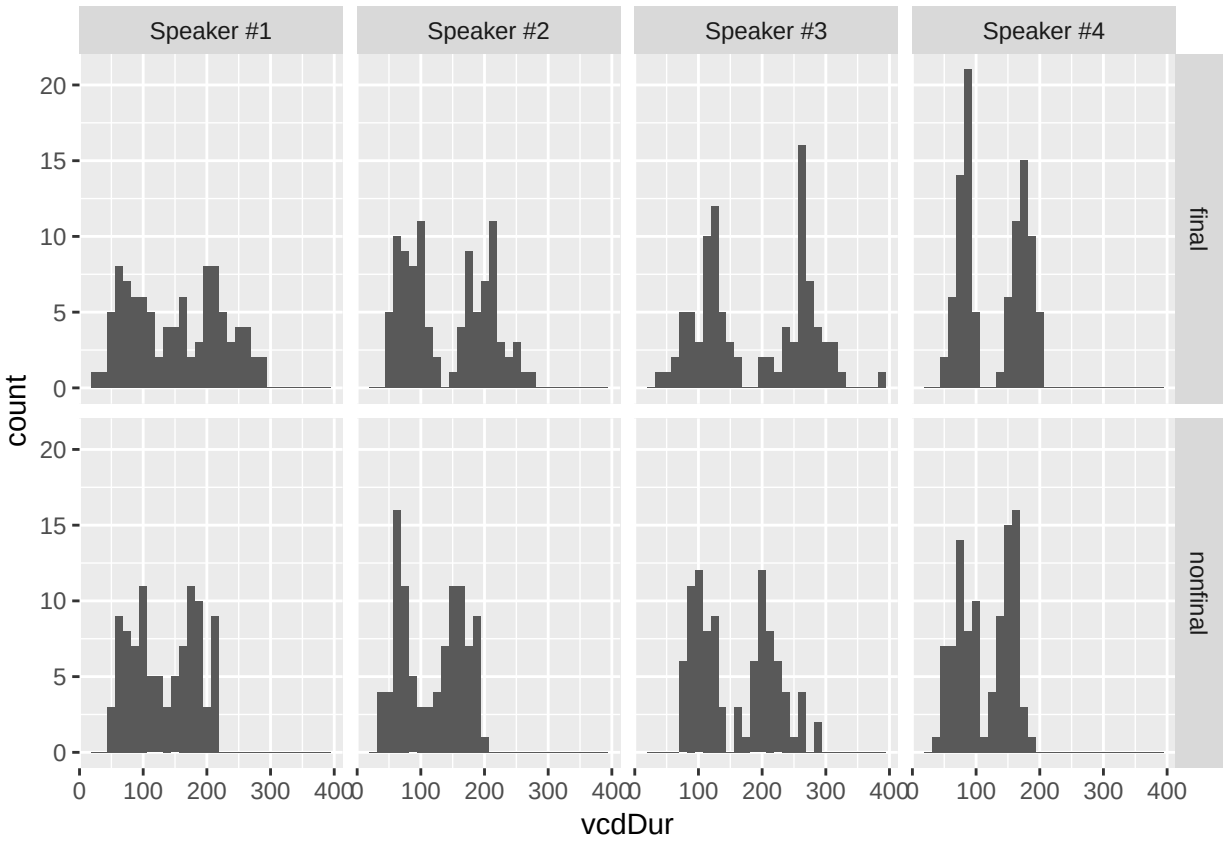
We now plot the values for 'vcdDur', as they differ according to vowel length, for each of the 4 speakers:

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



Doing the same with position instead of length:

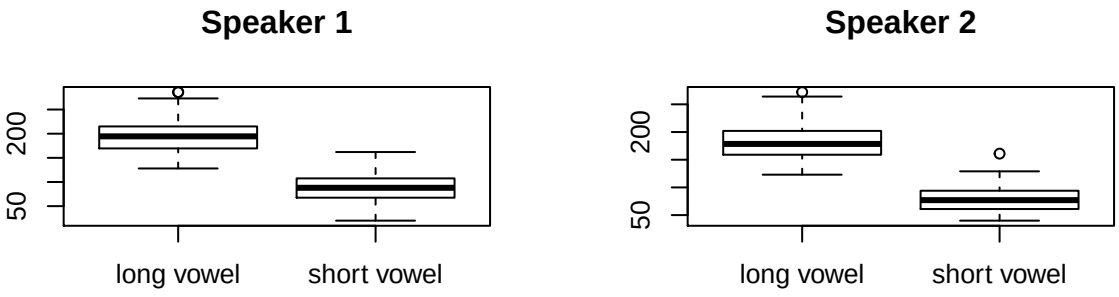
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

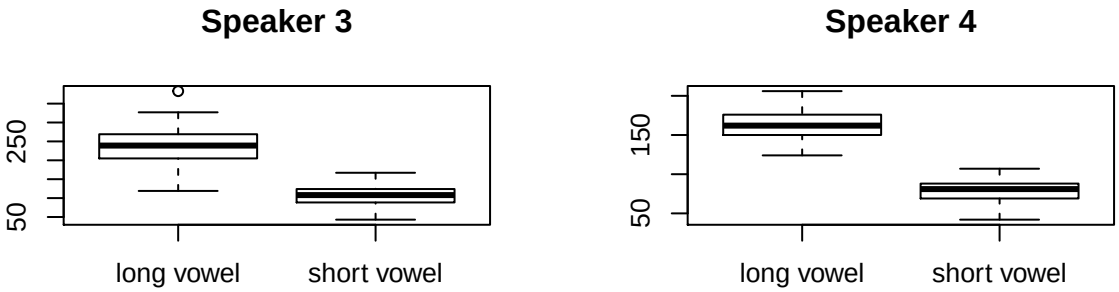


Vowel length seems to be a good indicator for the values of 'vcdDur'; short vowels correlate to 'vcdDur' values of ~0-150 where as long vowels correlate to values between 150 and 300.

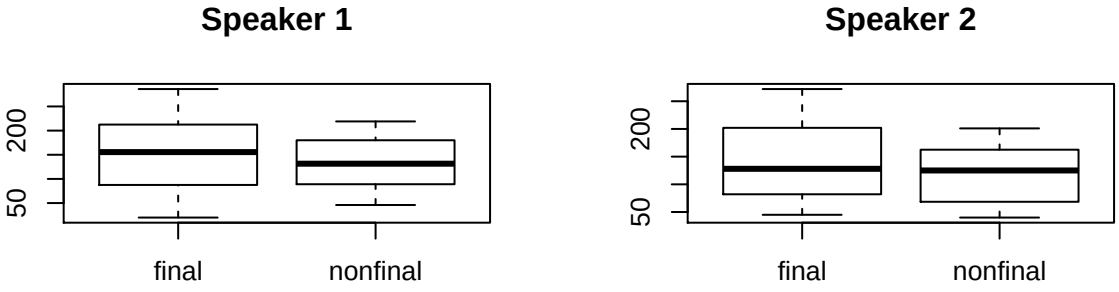
The figures for position show that the values for voiced durations extend to longer durations in final position than those in non-final position.

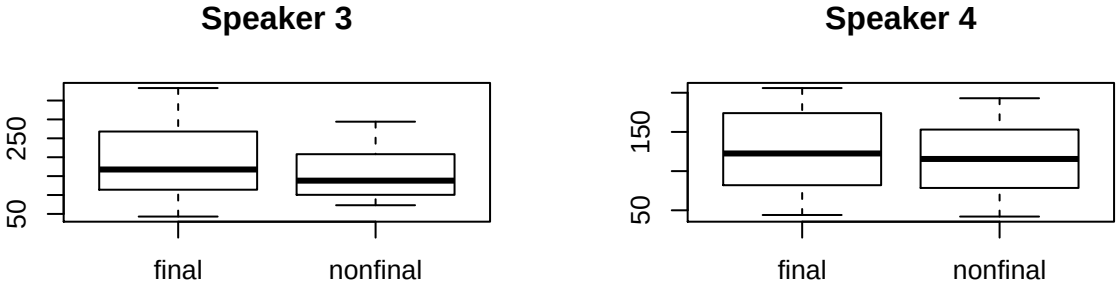
We can also plot boxplots for this data:





Doing the same with position instead of length:





Like the histograms, the box plots show voiced durations extending to longer values in final than non-final position.

Central tendancies and spread:

Creating a table summarising mean, median and standard deviation of vcdDur:

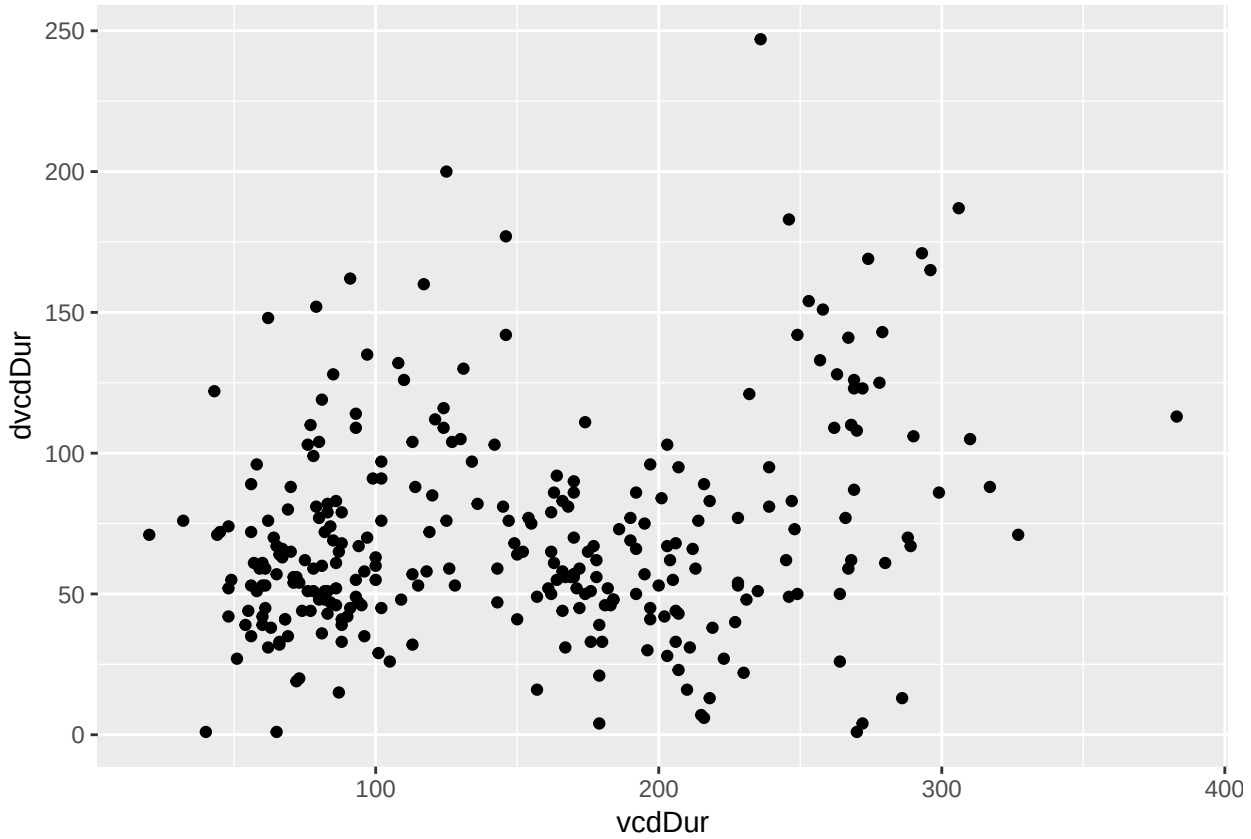
speaker	length	mean	median	sd
Speaker #1	long vowel	197.10417	194.5	36.05157
Speaker #1	short vowel	89.39583	88.0	29.50093
Speaker #2	long vowel	181.71875	178.5	32.13154
Speaker #2	short vowel	78.31250	77.0	22.38294
Speaker #3	long vowel	239.42708	239.0	44.47418
Speaker #3	short vowel	106.72917	108.0	23.74136
Speaker #4	long vowel	162.67708	162.0	18.39136
Speaker #4	short vowel	77.89583	81.0	14.98665

The speakers do not differ too much, perhaps with the exception of speaker 3 who has a noticably higher mean than the other three. The mean and median values coincide almost perfectly. The standard deviation is higher for long vowels than for short vowels across all four speakers.

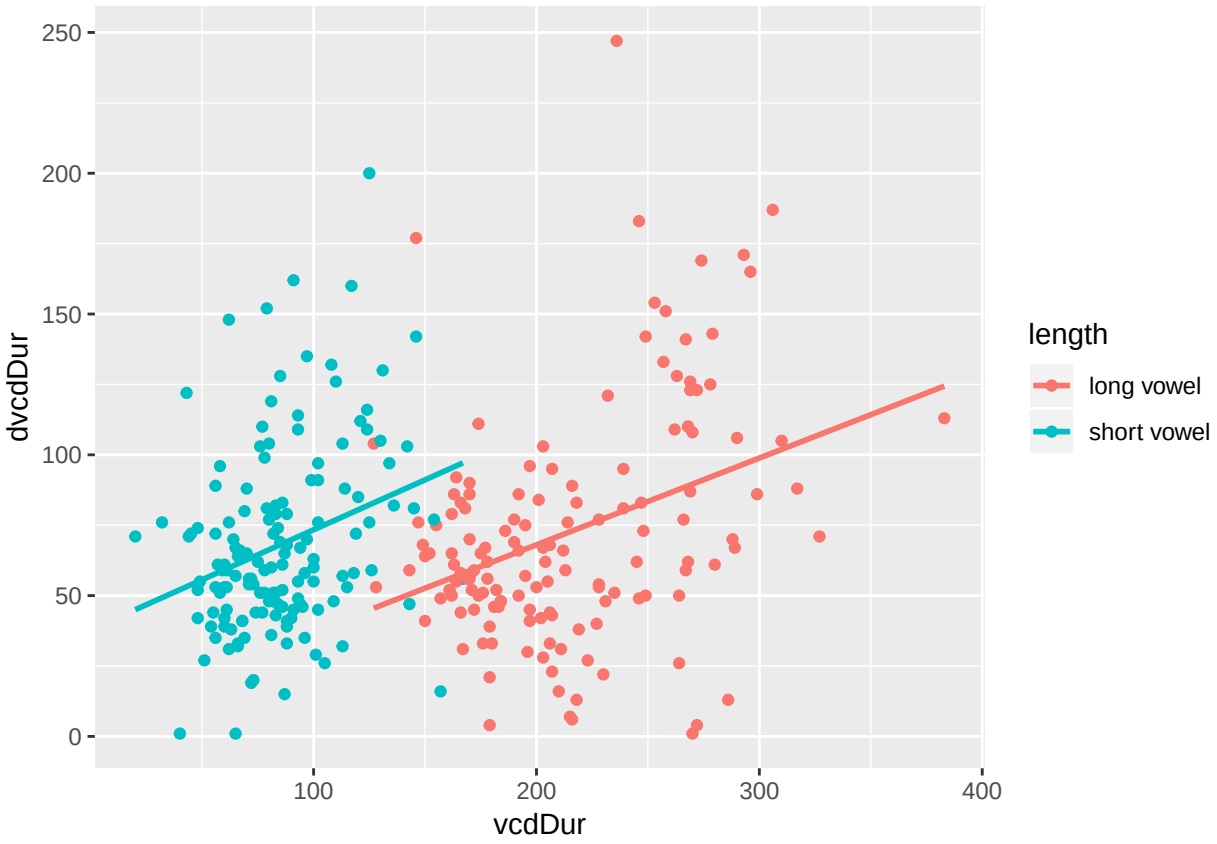
Relationships:

Before exploring relationships between vcdDur and dvcdDur, we need to clean up the data a little bit. First we must get rid of all the entries with dvcdDur = 0:

Now we can create a scatterplot showing the relationship between vcdDur and dvcdDur:

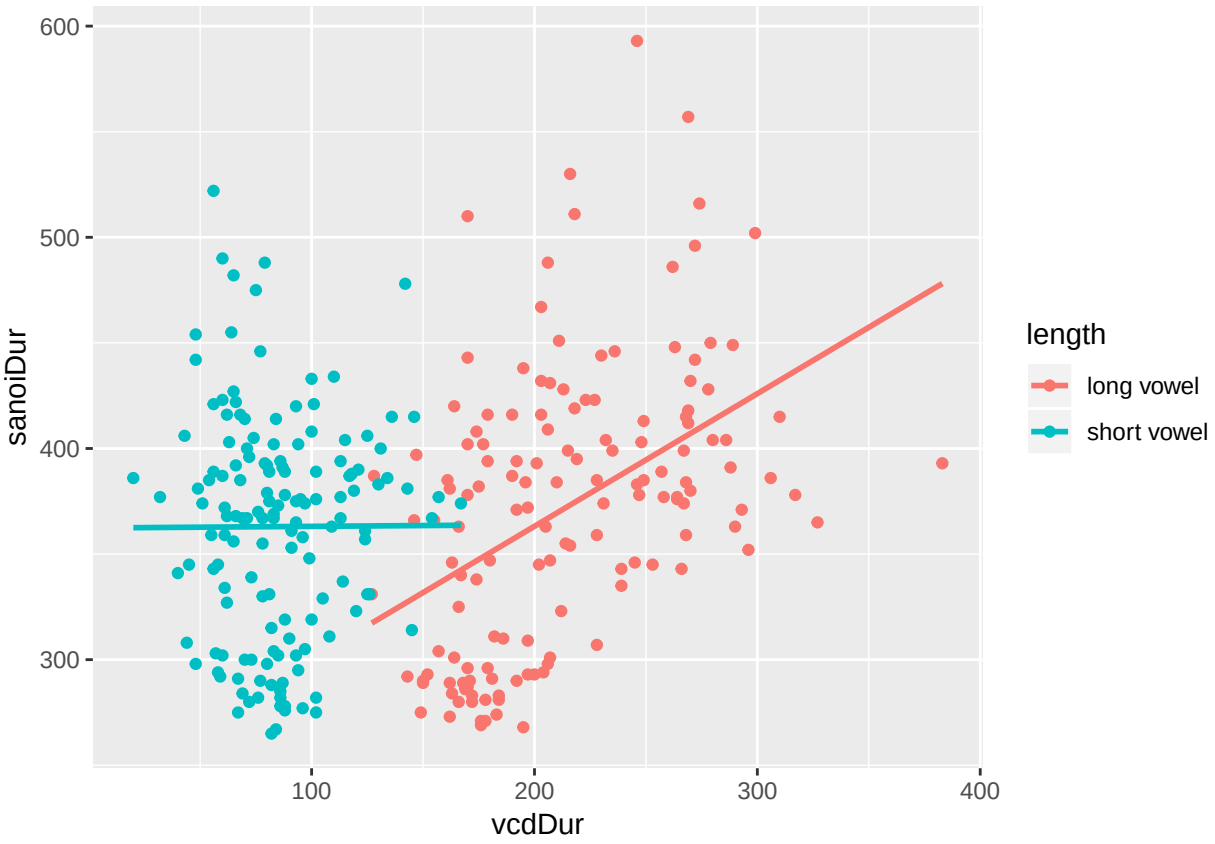


Adding a smoother and separating points by length:



From these plots it appears that there is a slight positive correlation between vcdDur and dvcdDur. This correlation is slightly stronger for short vowels than it is for long vowels.

We can now do the same for sanoiDur and vcdDur:



This plot shows a much stronger correlation between vcdDur and sanoiDur for long vowels. There is almost no correlation for the short vowels.

Making a table for correlation coefficients:

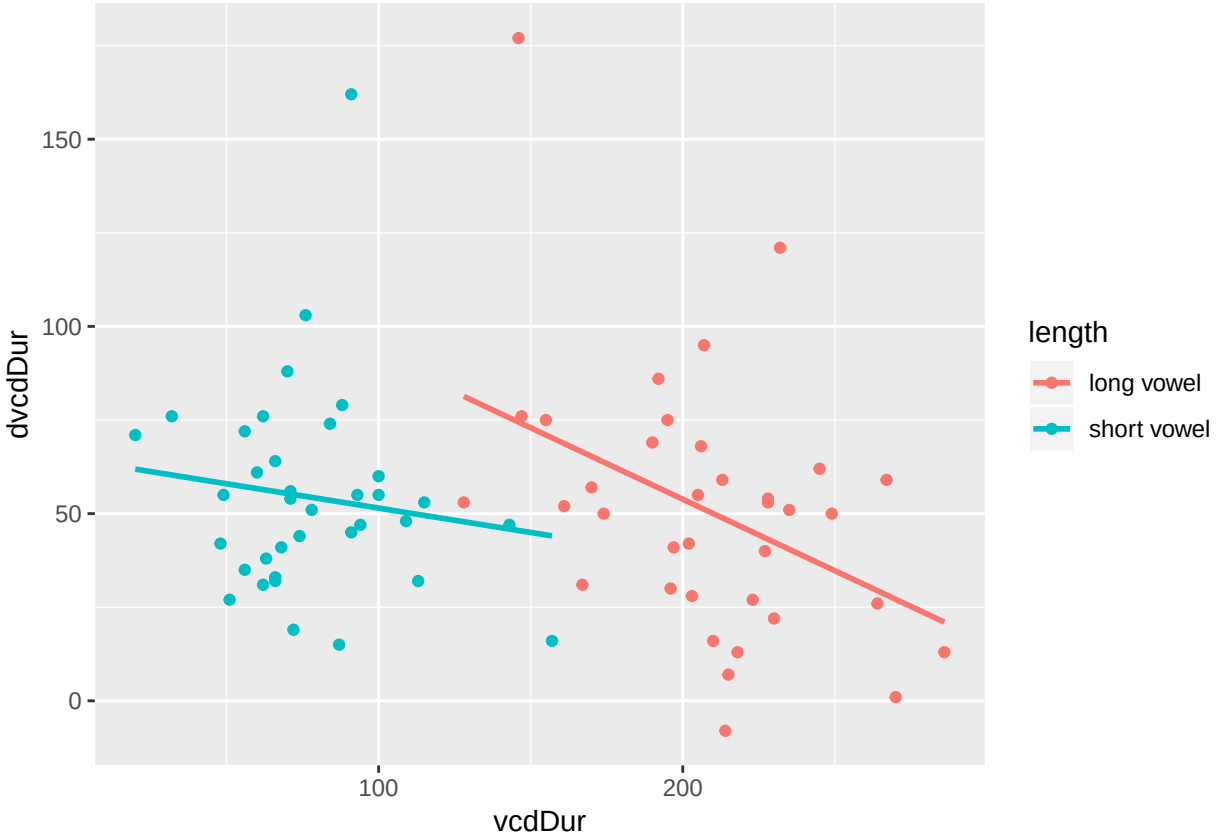
length	dvcd.by.vcd	vcd.by.sanoi
long vowel	0.3581531	0.4522661
short vowel	0.2872989	0.0039189

These numbers reflect the information provided by the plots: the highest correlation is shown between vcdDur and sanoiDur for long vowels whereas the correlation for short vowel is almost 0. We can also take into account the different speakers and obtain the following table:

length	speaker	dvcd.by.vcd	vcd.by.sanoi
long vowel	Speaker #1	-0.4133443	0.0842344
long vowel	Speaker #2	-0.3319361	0.0516208
long vowel	Speaker #3	-0.1411867	-0.3646200
long vowel	Speaker #4	-0.1881483	0.0074608
short vowel	Speaker #1	-0.1315856	-0.1448145
short vowel	Speaker #2	0.2371369	-0.1634136
short vowel	Speaker #3	-0.1769520	-0.1058575
short vowel	Speaker #4	-0.0420906	-0.1890529

These numbers are a bit puzzling. The correlations for individual speakers are mostly negative. However in the table and plots above there is a clearly positive correlation between vcdDur and dvcdDur as well as

vcdDur and sanoiDur. If we create a scatterplot between vcdDur and dvcdDur for only speaker 1, we get the following plot:



I am not quite sure what to make of this. There is a positive correlation when the data for all 4 speakers is pooled together, but not for each individual's data.

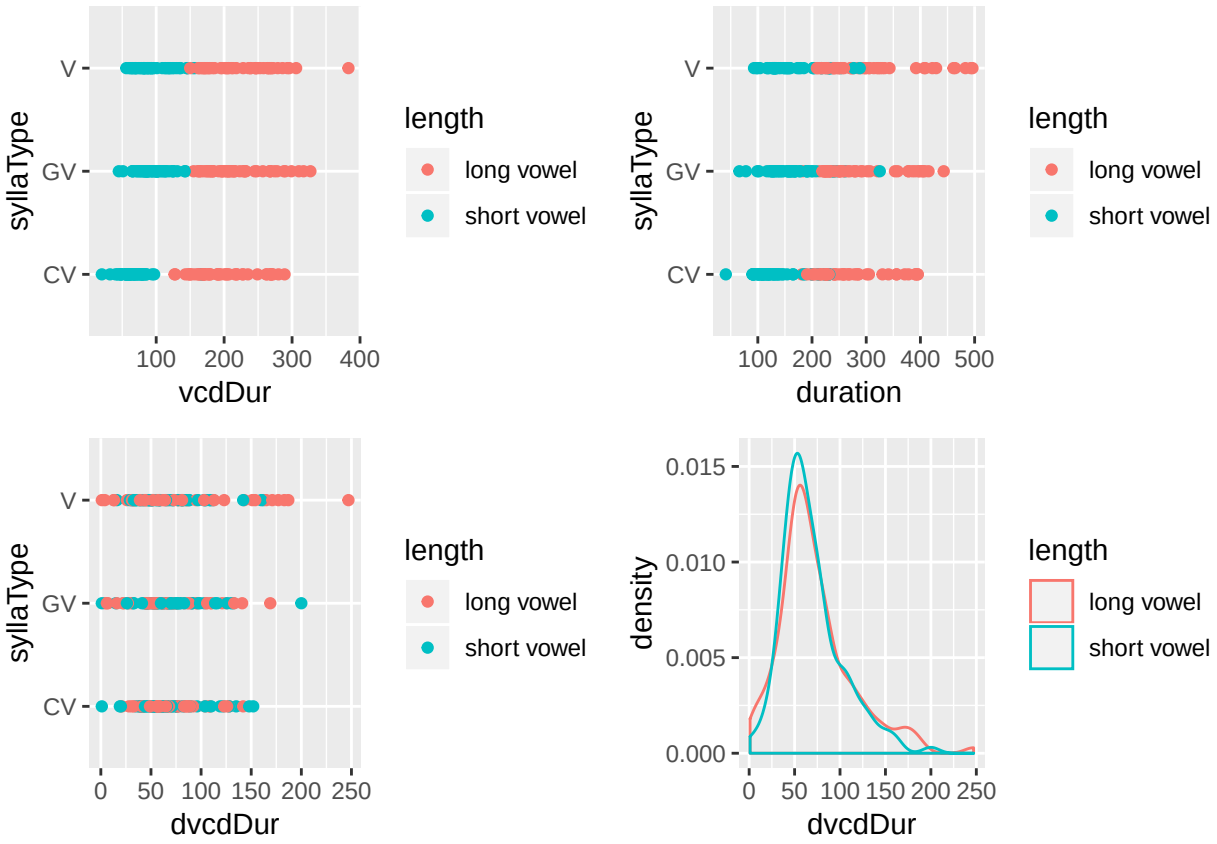
Open-ended:

We can explore the relationship between syllaType and vcdDur, duration and dvcdDur in a similar fashion:

```
## Loading required package: gridExtra
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```



From these plots it is clear that for duration and vcdDur the distribution is divided clearly into long and short categories. dvcdDur does not show this. Instead, the distribution is concentrated between the range ~0-150 regardless of vowel length. If we take the mean of each of these measures and split them along vowel length we get the following table:

length	mean_vcdDur	mean_duration	mean_dvcdDur
long vowel	214.25874	286.6853	72.42657
short vowel	85.31724	153.4897	68.17241

This table shows that dvcdDur does not appear to be a funciton of duration or vcdDur.