

введение

2 переменные

колич+колич

колич+катег

1 переменная

количественная

категориальная

3 переменные

колич+колич+колич

фасетизация

Визуализация данных: базовые функции R и пакет ggplot2

Г. Мороз

Р для визуализации? Совсем не обязательно...

Взято [отсюда](#). Куча ресурсов, которые скоро устареют.

Matplotlib

Bokeh

- AnyChart
- Chart Tool
- Chart.js
- Chartbuilder
- Chartbuilder 2.0
- ChartGo
- Chiasm
- D3plus
- Datahero
- Datamatic
- Datavisual
- Datawrapper
- Diagrammer

- Diychart
- Dygraphs
- Echarts
- Envision.js
- filtergraph
- Flare
- Google Charts
- Highcharts
- iCharts
- Infogr.am
- JS Charts
- JavaScript InfoVis Toolkit
- Livegap Charts
- Lyra
- Plotly

Processing

- Qlik
- Raw
- Lumira
- Slemma
- Spotfire
- Sprites
- Tableau
- VIDI
- Vega
- Visage
- Vizydrop
- Weave
- Zingchart

Элементы визуализации

введение

○ система координат

2 переменные

колич + колич

колич + катег

1 переменная

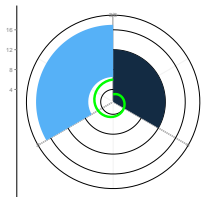
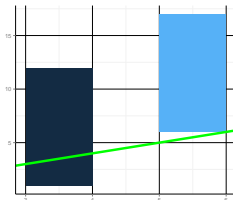
количественная

категориальная

3 переменные

колич + колич + колич

фасетизация



Элементы визуализации

введение

2 переменные

колич+колич

колич+катег

1 переменная

количественная

категориальная

3 переменные

колич+колич+колич

фасетизация

- позиция
- длина
- форма
- угол
- направление
- размер
- цвет
- ...

Данные

введение

2 переменные

колич + колич

колич + катег

1 переменная

количественная

категориальная

3 переменные

колич + колич + колич

фасетизация

В данной презентации все примеры будут приводиться на примере датасета из работы [Chi-kuk 2007] (доступна по ссылке <http://goo.gl/ZjrgaF>). В работе исследовались речь 7 гомосексуальных и 7 гетеросексуальных носителей кантонского диалекта языка юэ. В датасете есть следующие переменные:

- долгота s (`s.duration.ms`)
- долгота гласных (`vowel.duration.ms`)
- среднее значение ЧОТ (`average.f0.Hz`)
- диапазон ЧОТ (`f0.range.Hz`)
- сколько носителей воспринимает говорящего как гомосексуала (`perceived.as.homo`)
- сколько носителей воспринимает говорящего как гетеросексуала (`perceived.as.hetero`)
- ориентация говорящего (`orientation`)
- возраст говорящего (`age`)

презентация доступна: <https://goo.gl/4T12Ar>

В R визуализация реализована по-разному:

- core R (всегда будет слева)
- библиотекой ggplot2 (всегда будет справа)

Все примеры ggplot2 в данной презентации реализованы в связке с пакетом dplyr. Так что, для того чтобы код из данной презентации работал следует включить библиотеки и скачать датасет:

```
library(ggplot2)
library(dplyr)
chi.kuk <- read.csv("http://goo.gl/Zjr9aF") # закачивает датасет
```

scatterplot

введение

2 переменные

колич+колич

колич+катег

1 переменная

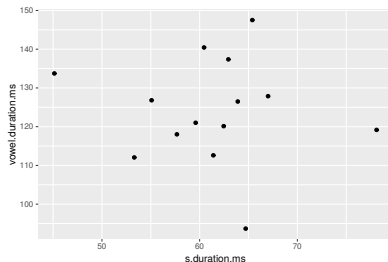
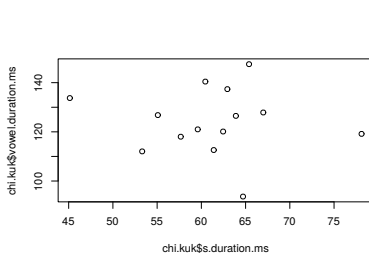
количественная

категориальная

3 переменные

колич+колич+колич

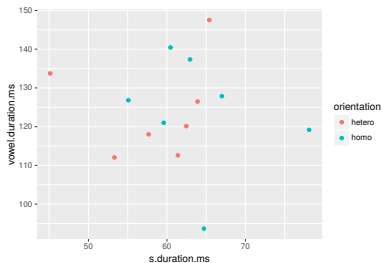
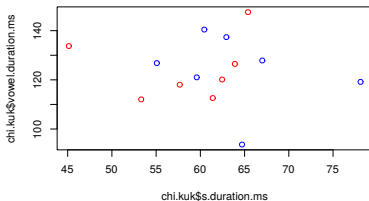
фасетизация



```
# base R  
plot(chi.kuk$s.duration.ms, chi.kuk$vowel.duration.ms)
```

```
# dplyr, ggplot2  
chi.kuk %>%  
  ggplot(aes(s.duration.ms, vowel.duration.ms)) +  
  geom_point()
```

scatterplot: цвет



```
# base R
```

```
plot(chi.kuk$s.duration.ms, chi.kuk$vowel.duration.ms,  
     col = c("red", "blue")[as.numeric(chi.kuk$orientation)])
```

```
# dplyr, ggplot2
```

```
chi.kuk %>%
```

```
  ggplot(aes(s.duration.ms, vowel.duration.ms, color = orientation)) +  
  geom_point()
```


scatterplot: форма

введение

2 переменные

колич+колич

колич+катег

1 переменная

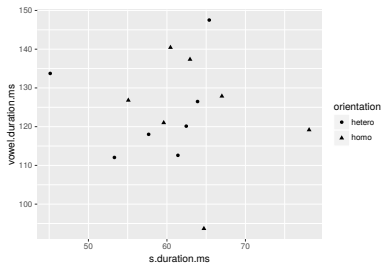
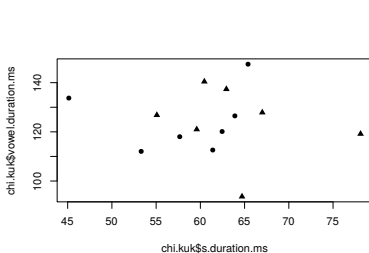
количественная

категориальная

3 переменные

колич+колич+колич

фасетизация



```
# base R
```

```
plot(chi.kuk$s.duration.ms, chi.kuk$vowel.duration.ms,  
     pch = c(16, 17)[as.numeric(chi.kuk$orientation)])
```

```
# dplyr, ggplot2
```

```
chi.kuk %>%
```

```
  ggplot(aes(s.duration.ms, vowel.duration.ms, shape = orientation)) +  
  geom_point()
```

scatterplot: размер

введение

2 переменные

колич+колич

колич+катег

1 переменная

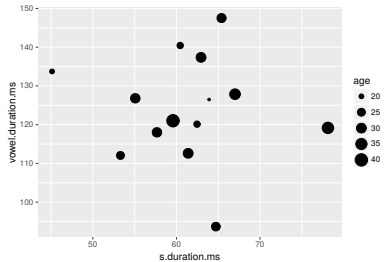
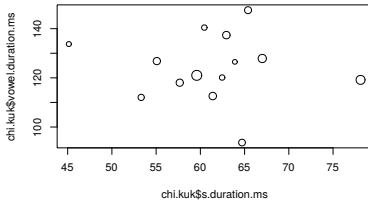
количественная

категориальная

3 переменные

колич+колич+колич

фасетизация



```
# base R
```

```
plot(chi.kuk$s.duration.ms, chi.kuk$vowel.duration.ms,  
     cex = chi.kuk$age/20)
```

```
# dplyr, ggplot2
```

```
chi.kuk %>%
```

```
  ggplot(aes(s.duration.ms, vowel.duration.ms, color = orientation)) +  
  geom_point()
```

scatterplot: текст

введение

2 переменные

колич+колич

колич+катег

1 переменная

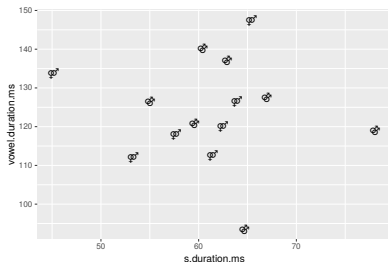
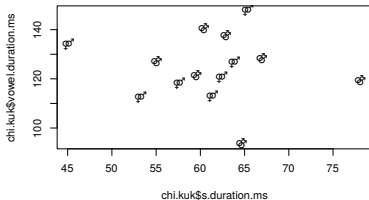
количественная

категориальная

3 переменные

колич+колич+колич

фасетизация



```
# base R
```

```
plot(chi.kuk$s.duration.ms, chi.kuk$vowel.duration.ms,  
     pch = c("ø", "ɘ")[as.numeric(chi.kuk$orientation)])
```

```
# dplyr, ggplot2
```

```
levels(chi.kuk$orientation) <- list("homo"="ɘ", "hetero"="ø")
```

```
chi.kuk %>%
```

```
  ggplot(aes(s.duration.ms, vowel.duration.ms, label = orientation)) +  
  geom_text()
```

scatterplot: заголовок

введение

2 переменные

колич+колич

колич+катег

1 переменная

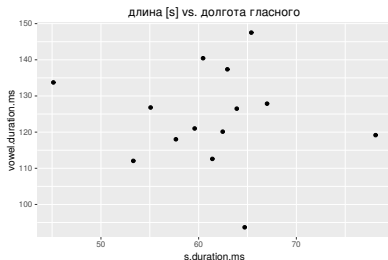
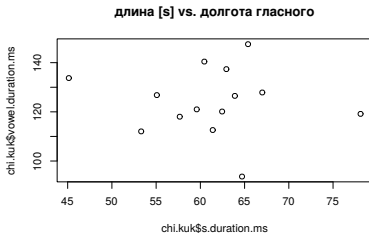
количественная

категориальная

3 переменные

колич+колич+колич

фасетизация



```
# base R
```

```
plot(chi.kuk$s.duration.ms, chi.kuk$vowel.duration.ms,  
     main = "длина [s] vs. долгота гласного")
```

```
# dplyr, ggplot2
```

```
chi.kuk %>%
```

```
  ggplot(aes(s.duration.ms, vowel.duration.ms)) +
```

```
  geom_point() +
```

```
  ggtitle("длина [s] vs. долгота гласного")
```

scatterplot: подписи осей

введение

2 переменные

колич+колич

колич+катег

1 переменная

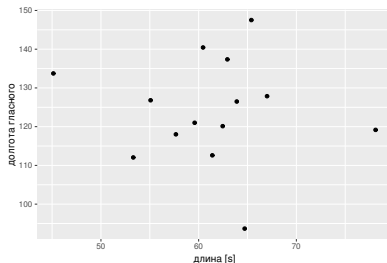
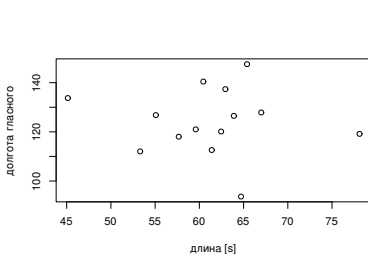
количественная

категориальная

3 переменные

колич+колич+колич

фасетизация



```
# base R
```

```
plot(chi.kuk$s.duration.ms, chi.kuk$vowel.duration.ms,  
     xlab = "длина [s]" ylab = "долгота гласного")
```

```
# dplyr, ggplot2
```

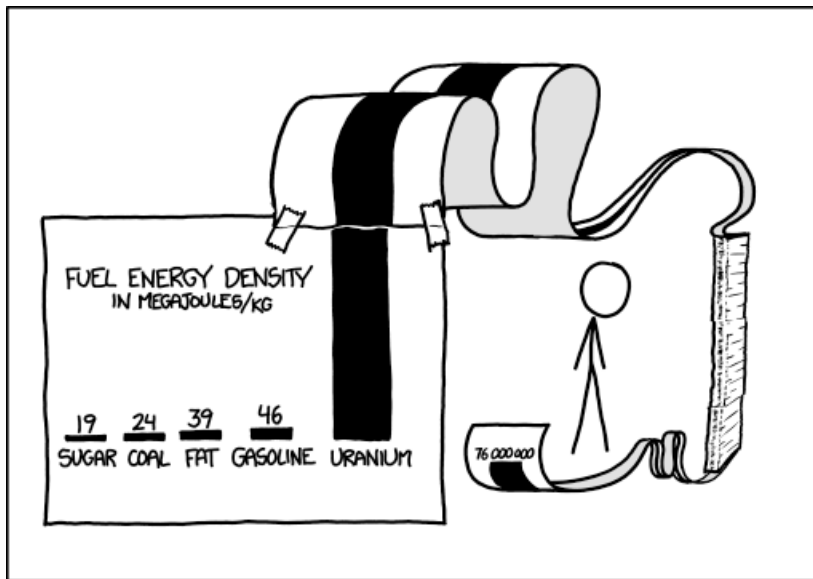
```
chi.kuk %>%
```

```
  ggplot(aes(s.duration.ms, vowel.duration.ms)) +
```

```
  geom_point() +
```

```
  xlab("длина [s]") +
```

```
  ylab("долгота гласного")
```



SCIENCE TIP: LOG SCALES ARE FOR QUITTERS WHO CAN'T
FIND ENOUGH PAPER TO MAKE THEIR POINT *PROPERLY*.

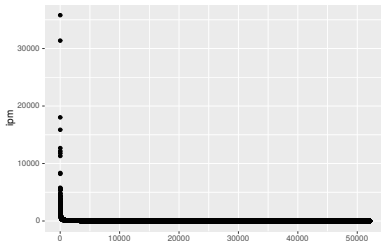
scatterplot: логарифмирование осей

Воспользуемся частотным словарем [Ляшевская, Шаров 2009] (топ-50000 лемм СЛРЯ) и посмотрим на параметр частоты слова (Freq.ipm., среднее на миллион словоупотреблений).

```
freq <- read.csv("https://goo.gl/TIX7xW", sep = "\t")
```

Если оси не логарифмировать, то получится следующее:

```
freq %>%  
  ggplot(aes(1:52138, Freq.ipm.)) +  
  geom_point() +  
  xlab("") +  
  ylab("ipm")
```



scatterplot: логарифмирование осей

введение

2 переменные

колич+колич

колич+катег

1 переменная

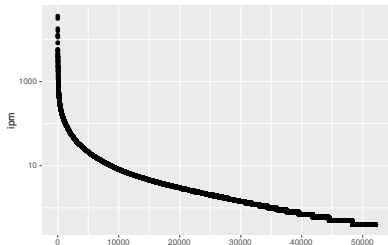
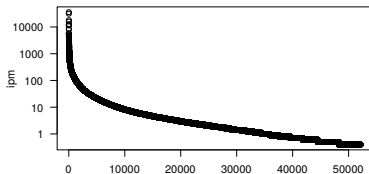
количественная

категориальная

3 переменные

колич+колич+колич

фасетизация



```
# base R
plot(1:52138, freq$Freq.ipm.,
     xlab = NA, ylab = "ipm",
     las = 1,
     log = "y")
```

поворот значений на оси y

```
# dplyr, ggplot2
freq %>%
  ggplot(aes(1:52138, Freq.ipm.)) +
  geom_point() +
  xlab("") +
  ylab("ipm") +
  scale_y_log10()
```

презентация доступна: <https://goo.gl/4T1zAr>

введение

2 переменные

колич+колич

колич+катег

1 переменная

количественная

категориальная

3 переменные

колич+колич+колич

фасетизация

Спасибо что долистали!

Пишите письма

agricolamz@gmail.com