ggplot2

Isidio Martins e Tomás Barcellos

27 de janeiro de 2017

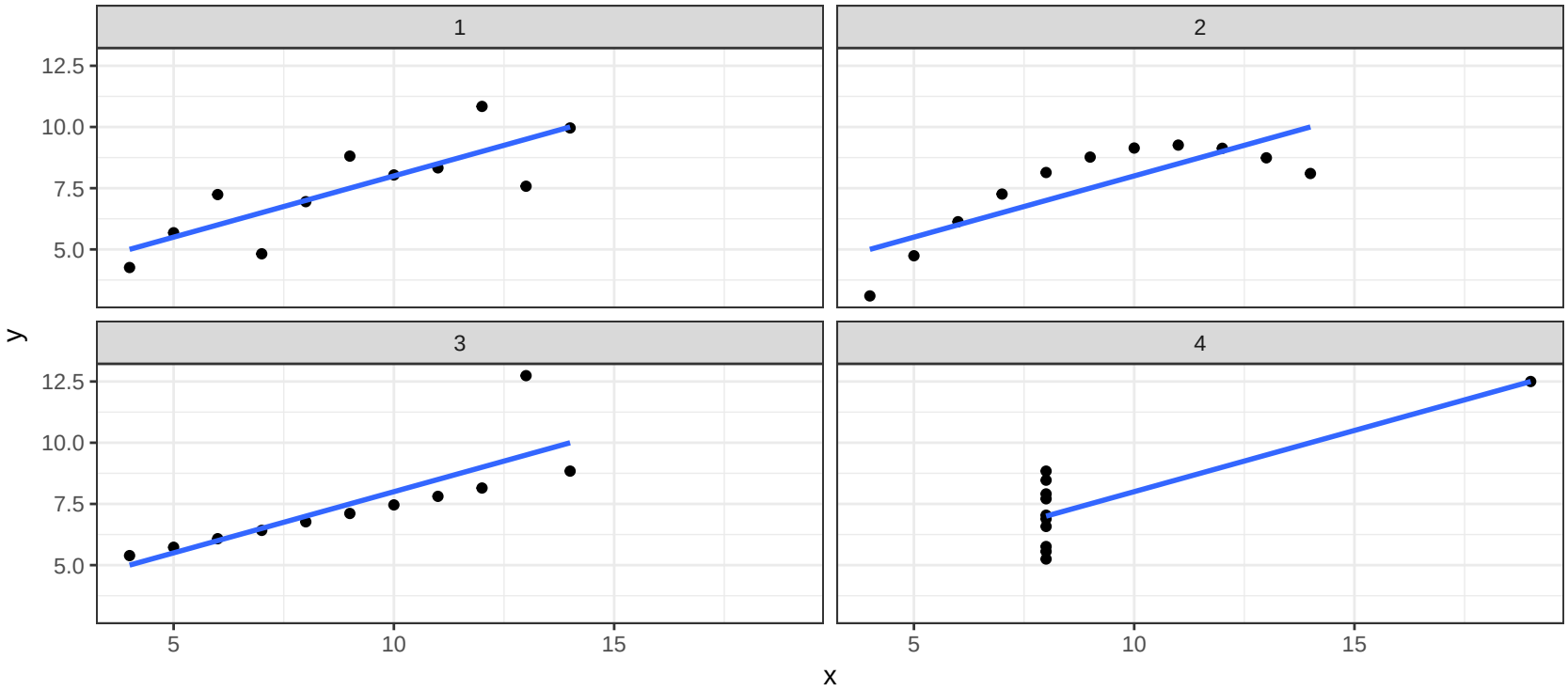
Visualização de dados

A gramática dos gráficos

Quarteto de Anscombe

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Quarteto de Anscombe



Mas antes...

O gráfico é um meio de comunicação e, como tal, deve ser adequado ao seu público. É diferente preparar um **gráfico de apresentação** para o portal do ministério ou fazer um **gráfico exploratório** para você mesmo. Ambos diferem em público e também em objetivo.

Tenha isso em mente quando for preparar os gráficos.

Concepção do ggplot2

O ggplot2 é mais do que um pacote para fazer gráficos; ele é uma tentativa (muito bem sucedida) trazer para o dia-a-dia dos técnicos uma **gramática dos gráficos em camadas**.

Por que uma **gramática** dos gráficos?

Através dela podemos definir **sistematicamente** quais são os componentes de um gráficos e como eles se interelacionam.

Veja mais informações em http://docs.ggplot2.org/.

A gramática dos gráficos

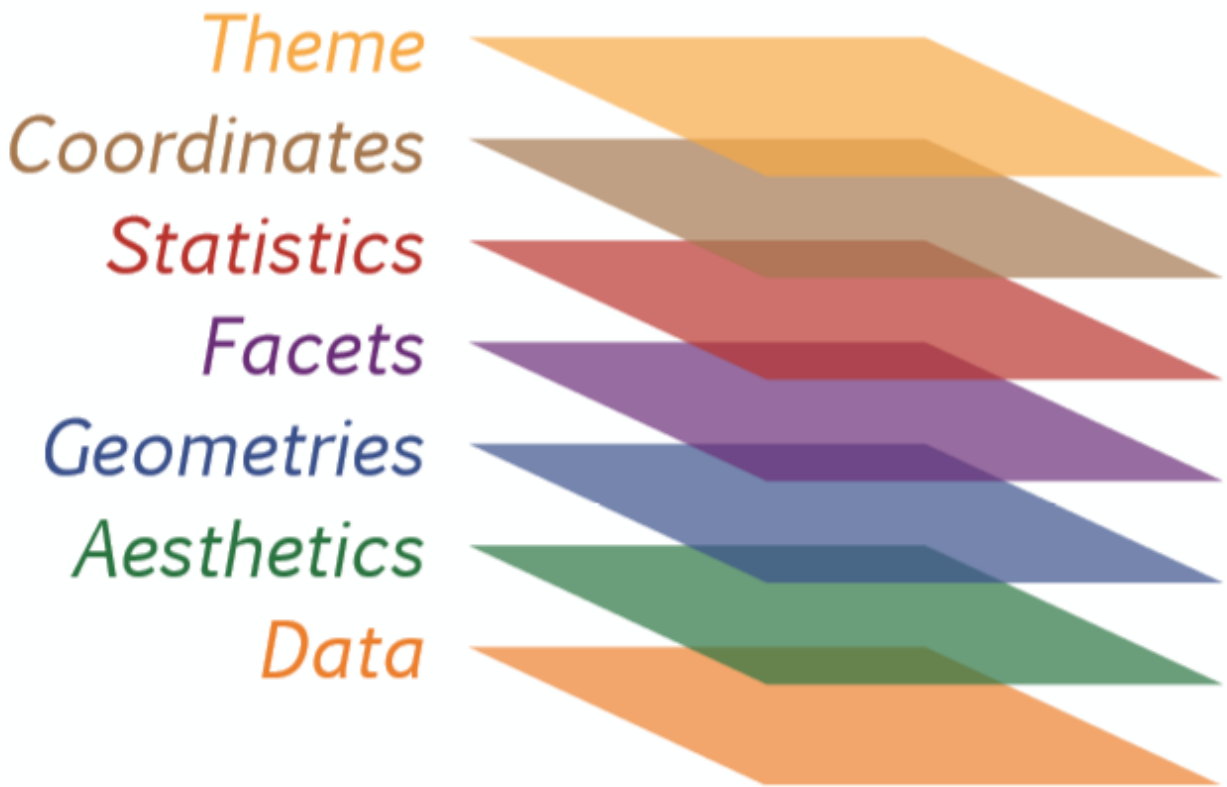


Figure 1

A gramática dos gráficos

elemento	exemplos
dados (informação)*	produção, fiscalizações
(a)estética*	cor, formato
geometrias*	barra, ponto
e**stat**ísticas	mediana, máximo
facetas	facetas
coordenadas	polar, cartesiana
t(h)emas	eixos, título

'* aspéctos estéticos imprescindíveis para criar um gráfico no ggplot2

Sintaxe do ggplot2

```
ggplot(um_data_frame, aes(estética1 = variável1,
                          estética2 = variável2,
                          estética3 = variável3)) +
  geometria(estética4 = "atributo1") +
  facetas +
  tema
```

Note que cada função cria uma (ou mais) camadas e que usamos o + para ir adicionando camadas.

A camada de dados

Carregando os dados

A primeira etapa da construção de um gráfico é ter os dados que serão representados graficamente.

Vamos carregar os dados da Pesquisa Agrícola Municipal (PAM) agregados no nível de grandes regiões para os anos entre 1990 e 2015.

```
# importa dados
dados <- readRDS('amostra-PAM.RDS')
```

Ainda os dados

O arquivo "amostra-PAM.RDS" traz dados de 67 culturas diferentes e é difícil visualizar tantas variáveis categóricas.

Para faciliar as coisas vamos reduzir nossos dados apenas para as culturas de **arroz**, **feijão**, **milho** e **soja** (grãos) com o código abaixo.

```
library(dplyr)
dados <- dados %>%
  filter(cultura %in%
           c("Arroz (em casca)", "Feijão (em grão)",
             "Soja (em grão)", "Milho (em grão)" ))
feijao <- filter(dados, cultura == "Feijão (em grão)")
CO <- filter(dados, regiao == "Centro-Oeste")
```

Aspectos Estéticos

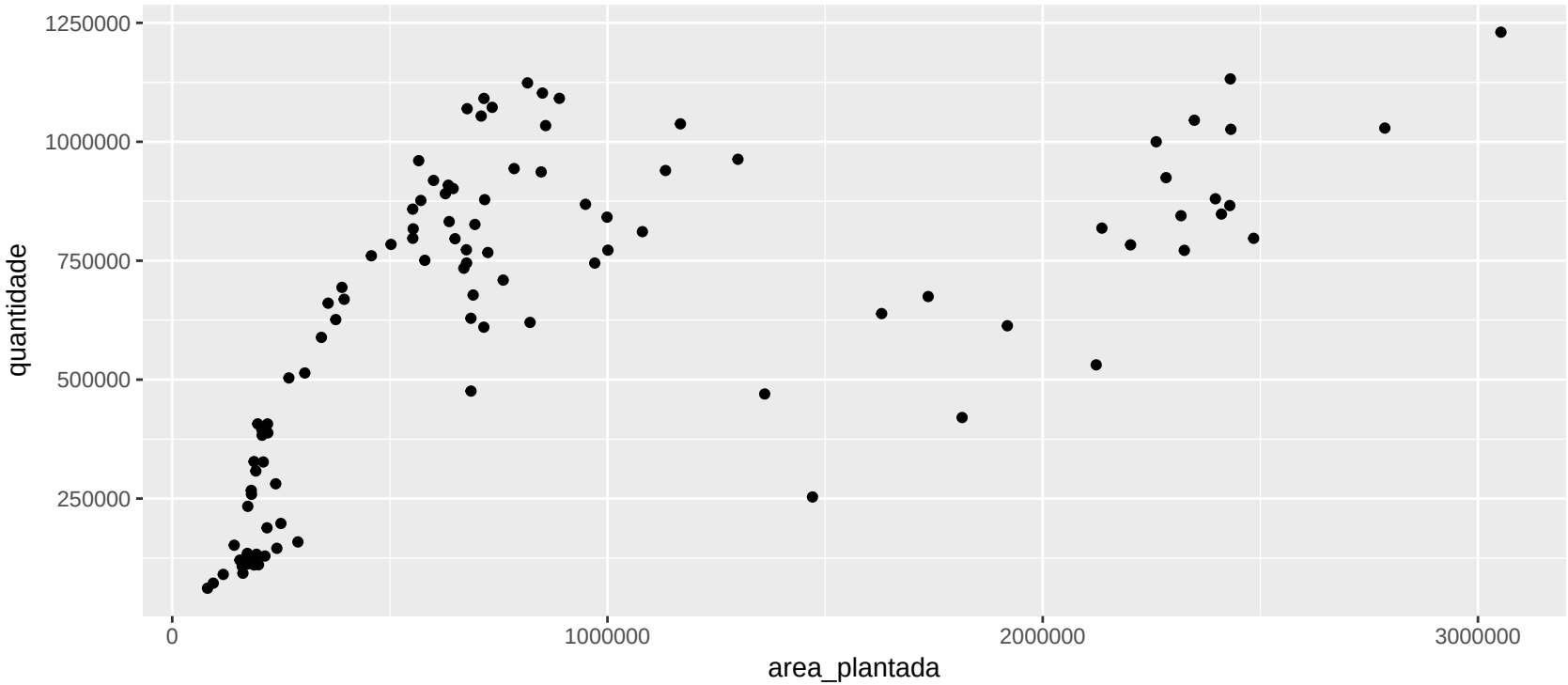
Aspectos Estéticos

Os principais aspectos estéticos são:

Estética	Descrição
x	Eixo horizontal
y	Eixo vertical
colour	Cor dos pontos ou das linhas das formas
fill	Cor de preencimento
size	Diametro dos pontos e espessura das linhas
alpha	Transparência
linetype	Tipo (padrão) da linhas
labels	Texto no gráfico ou nos eixos
shape	Forma

Representando dados

```
ggplot(feijao,
       aes(x = area_plantada, y = quantidade)) +
  geom_point()
```



Representando dados

Imagine que você fosse desenhar um gráfico. Como você decidiria até onde deve ir a barra ou onde ficariam os pontos? O computador também precisa de critérios para decidir como representar os dados, como o Valor Bruto da Produção agropecuária (VBP) de uma região, em um gráfico.

Assim, o VBP pode ser representado no eixo vertical ou os faixas de valores podem aparecer como cores ou formas (até R$ 50 milhões: triângulos; entre 50 e 100: quadrados; e maiores que 100: circulos).

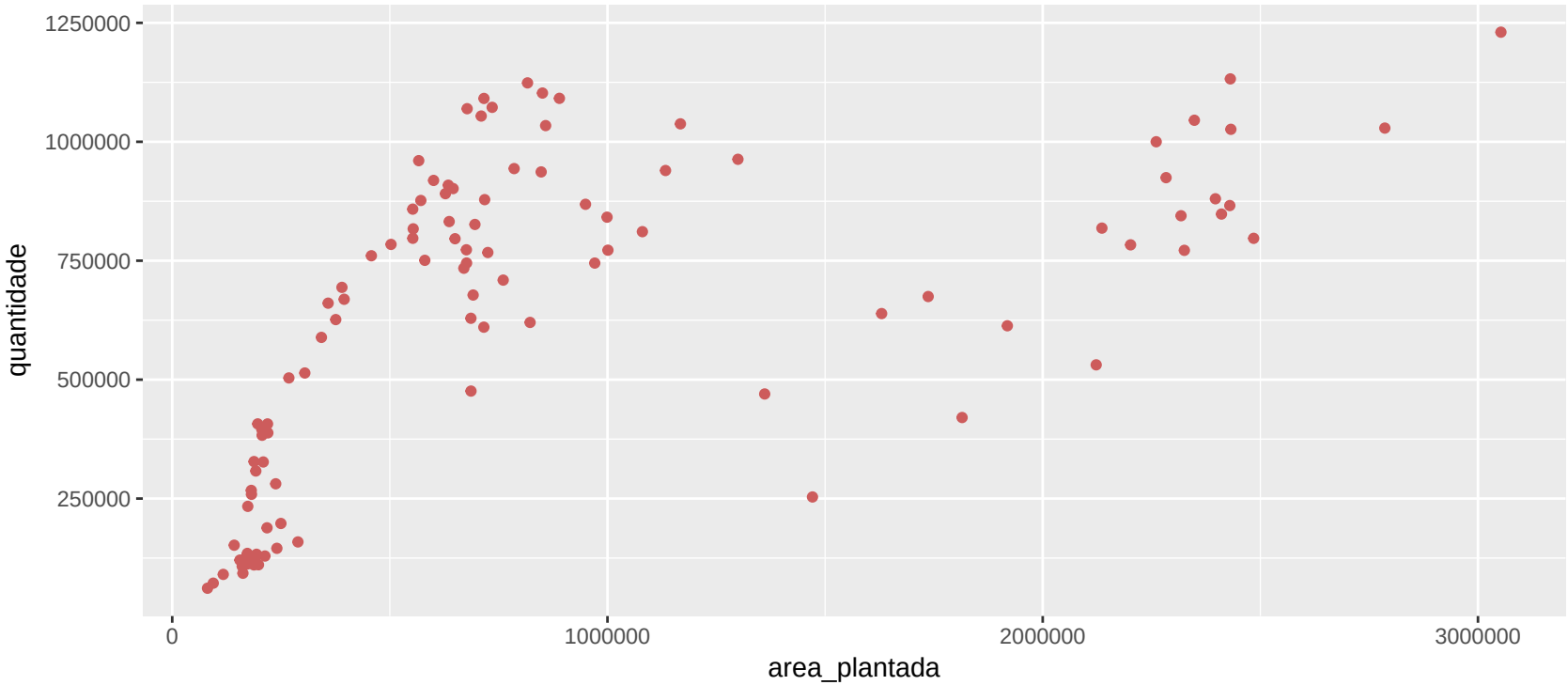
Mapeando variáveis em estéticas

É diferente **mapear** uma estética e **atribuir um valor** a um aspecto estético. Mapear uma variável em uma estética é dizer que a cor **vermelha** representa o Centro-Oeste e a cor **azul** o Sudeste. Isto é diferente de definir a cor de pontos ou barras como **verde**. Para poder fazer isso, precisamos carregar o pacote `ggplot2`

```
libarry(ggplot2)
```

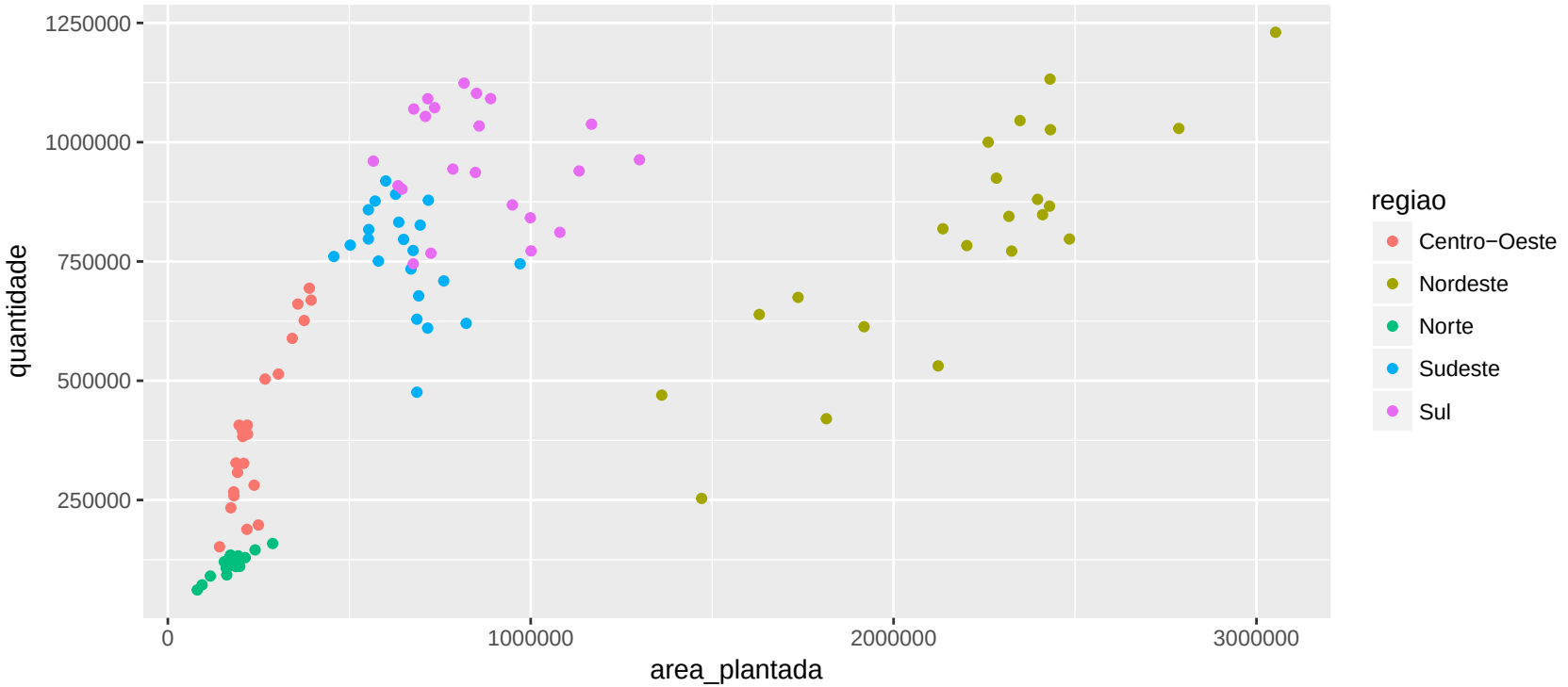
Atributos Estéticos - Atribuir cor à elemento estético

```
ggplot(feijao ,
       aes(x = area_plantada, y = quantidade)) +
  geom_point(col = "indianred")
```



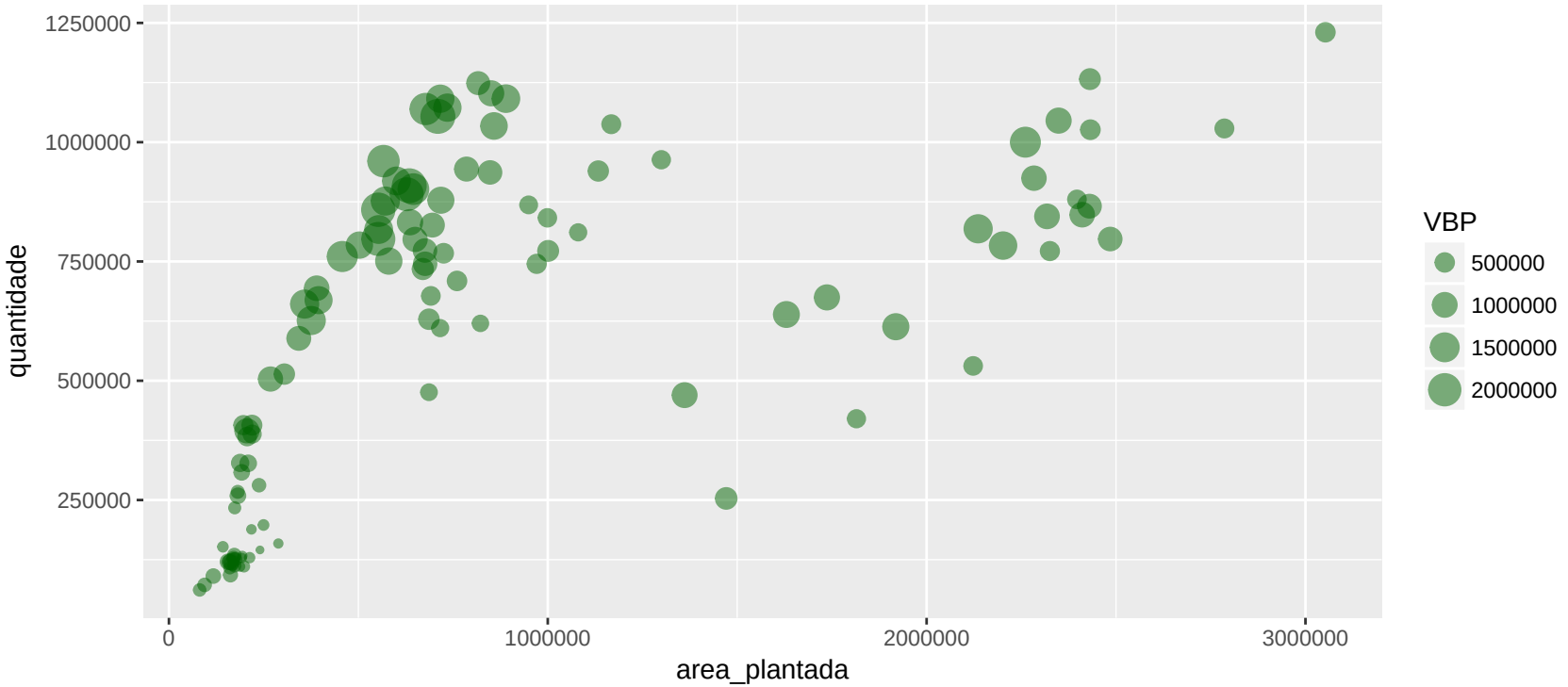
Aspectos Estéticos - Mapear a cultura na cor

```
ggplot(feijao , aes(x = area_plantada,
                    y = quantidade, col = regiao)) +
  geom_point()
```



Aspectos Estéticos - Mapear VBP no tamanho

```
ggplot(feijao , aes(x = area_plantada, y = quantidade)) +
  geom_point(aes(size = VBP),
             col = "darkgreen", alpha = 0.5)
```



Aspectos Estéticos - Variáveis contínuas

Estética	Descrição
x	Eixo horizontal
y	Eixo vertical
colour	Cor dos pontos ou das linhas das formas
fill	Cor de preencimento
size	Diametro dos pontos e espessura das linhas
alpha	Transparência

Aspectos Estéticos - Variáveis contínuas

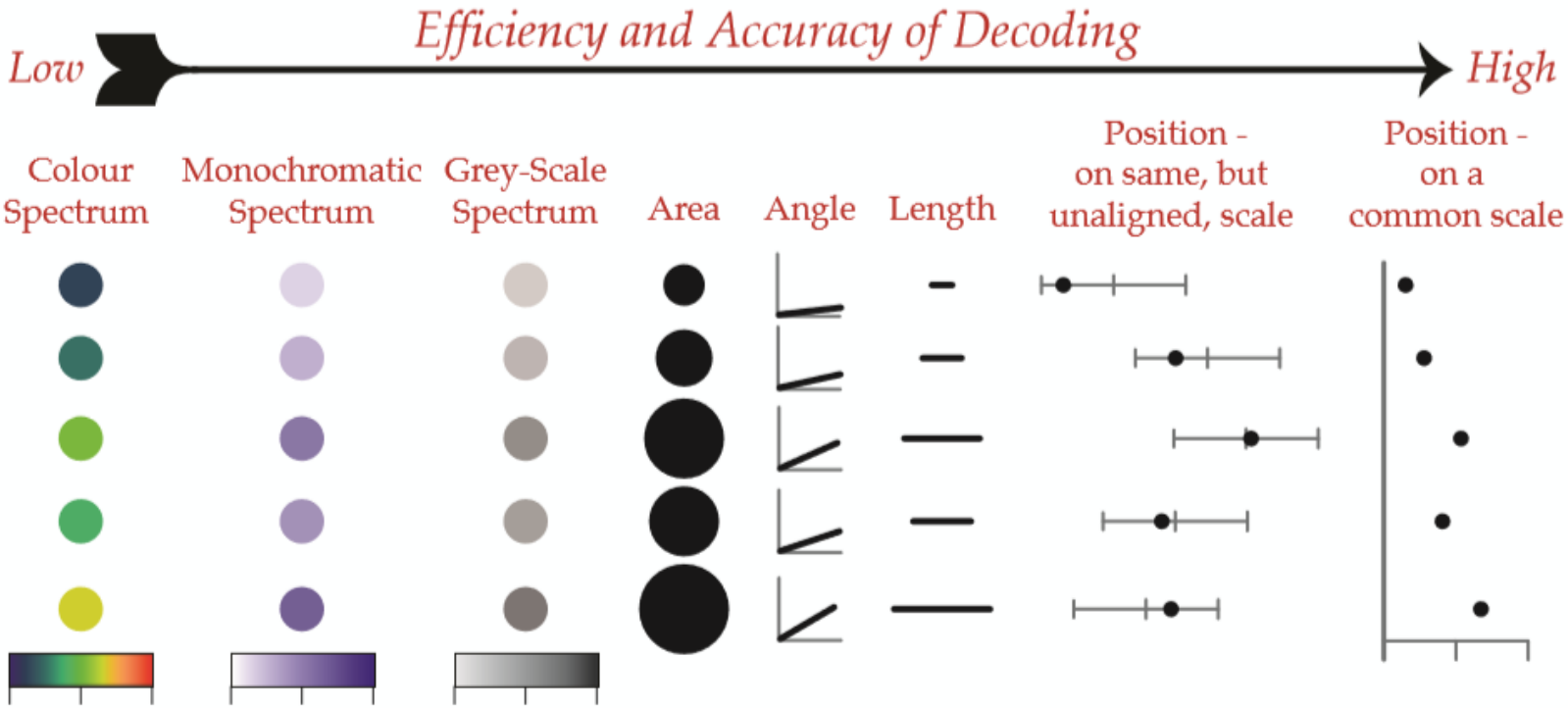
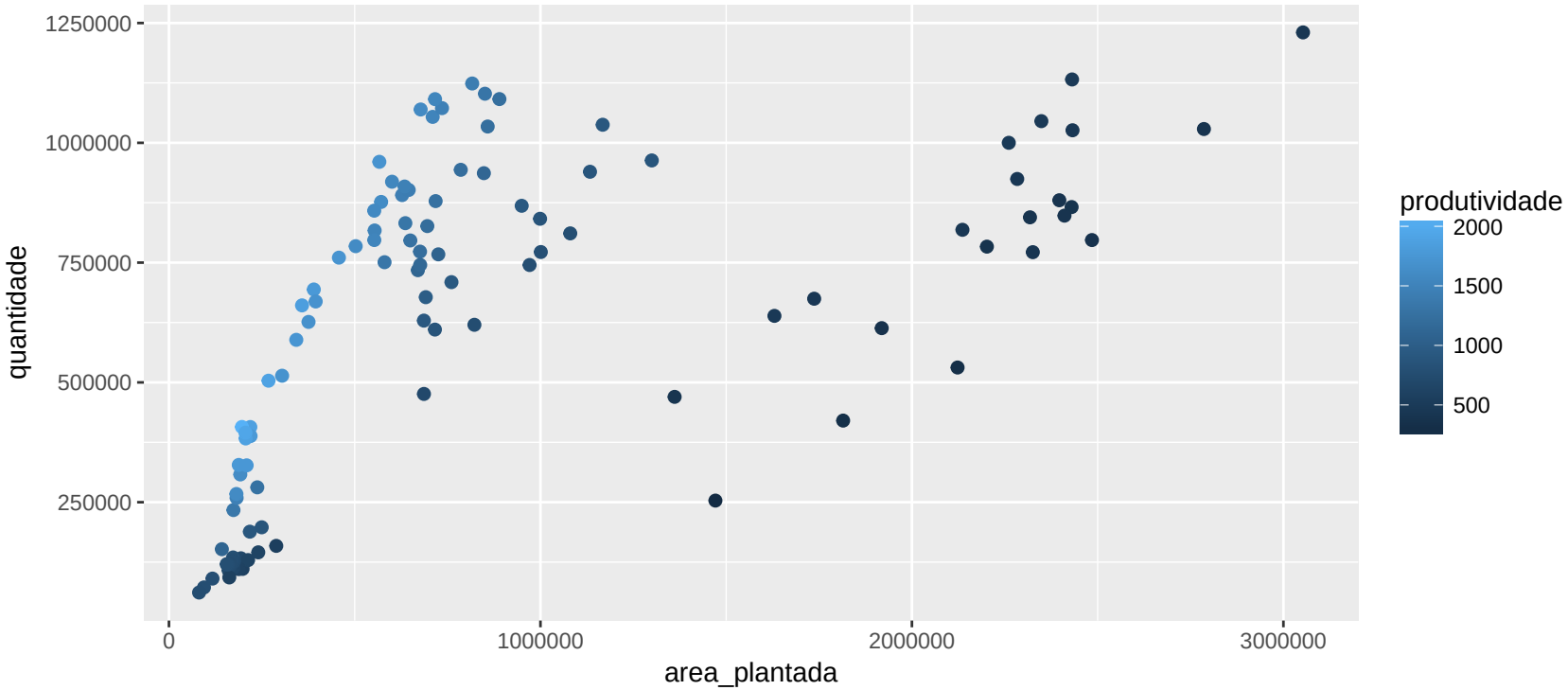


Figure 2

Fonte: www.datacamp.com

Aspectos Estéticos - Variáveis contínuas

```
ggplot(feijao, aes(x = area_plantada, y = quantidade,
                   col = produtividade)) +
  geom_point(size = 2)
```



Aspectos Estéticos - Variáveis categóricas

Estética	Descrição
colour	Cor dos pontos ou das linhas das formas
fill	Cor de preencimento
size	Diametro dos pontos e espessura das linhas
alpha	Transparência
linetype	Tipo (padrão) da linhas
labels	Texto no gráfico ou nos eixos
shape	Forma

Atributos Estéticos - Variáveis categóricas

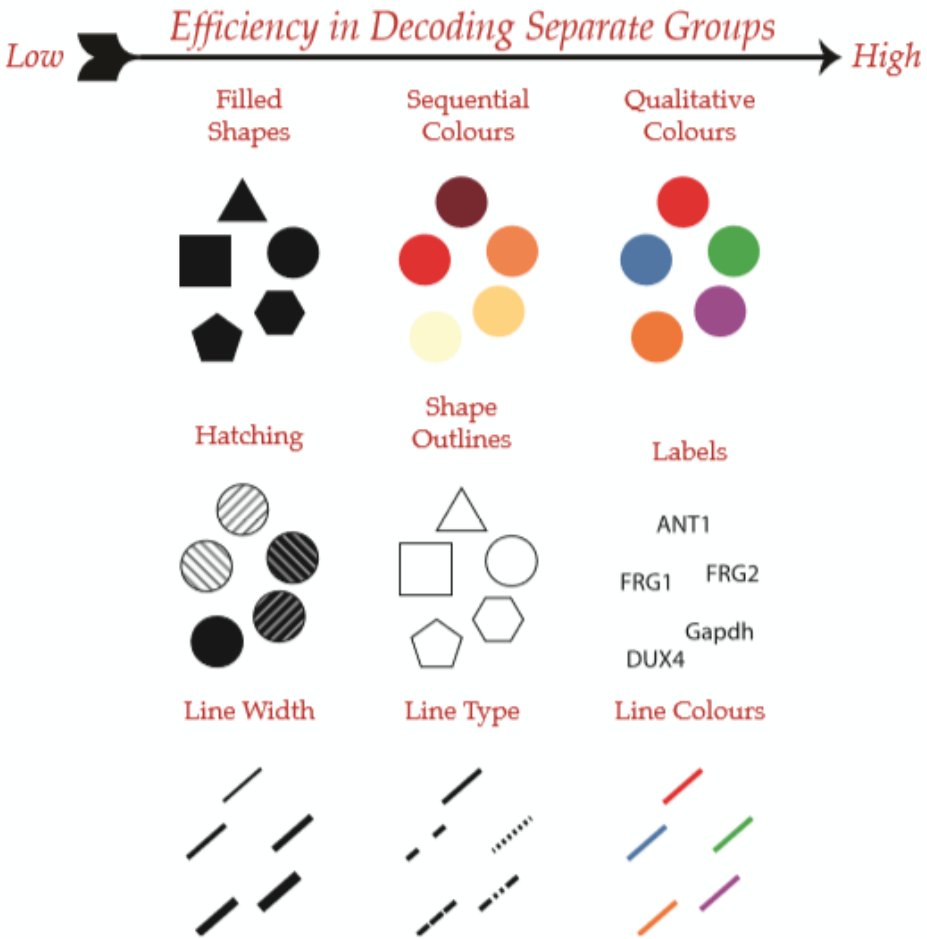
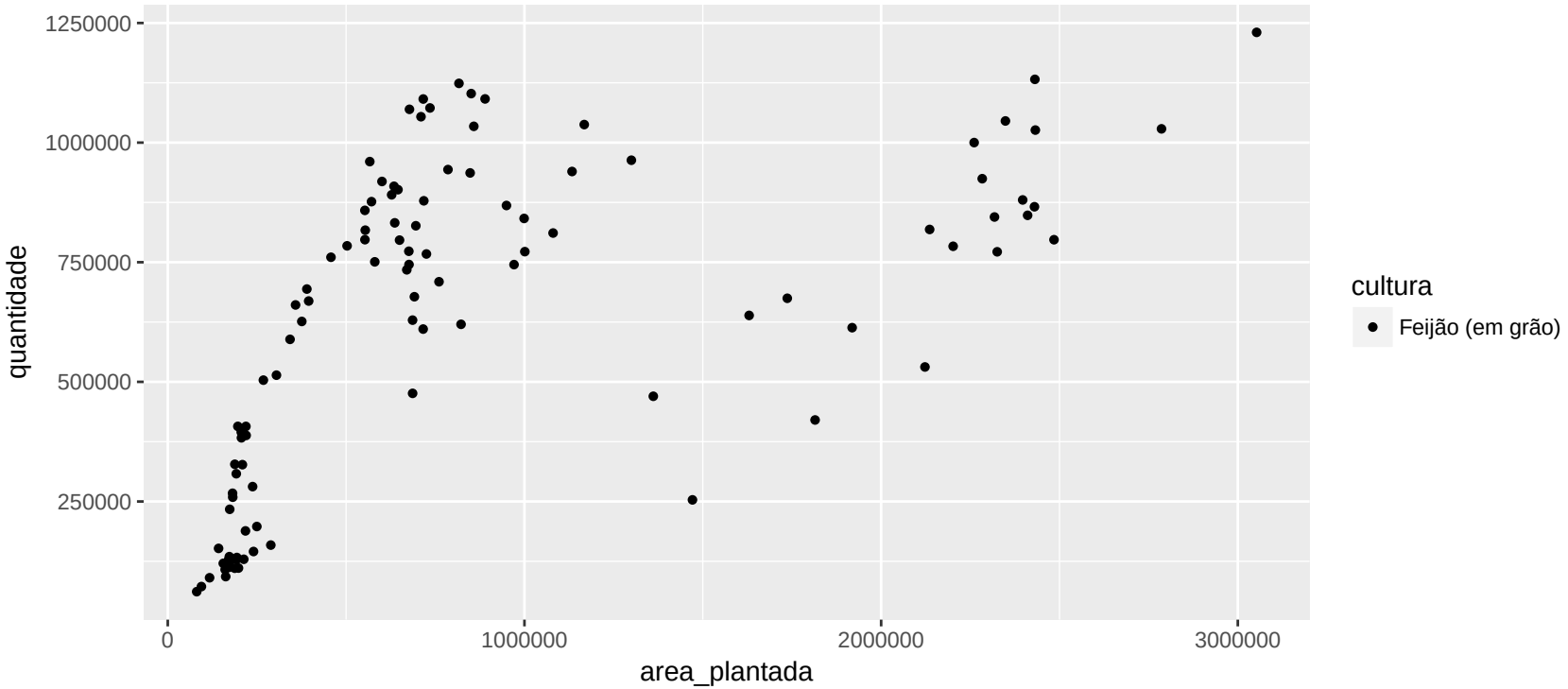


Figure 3

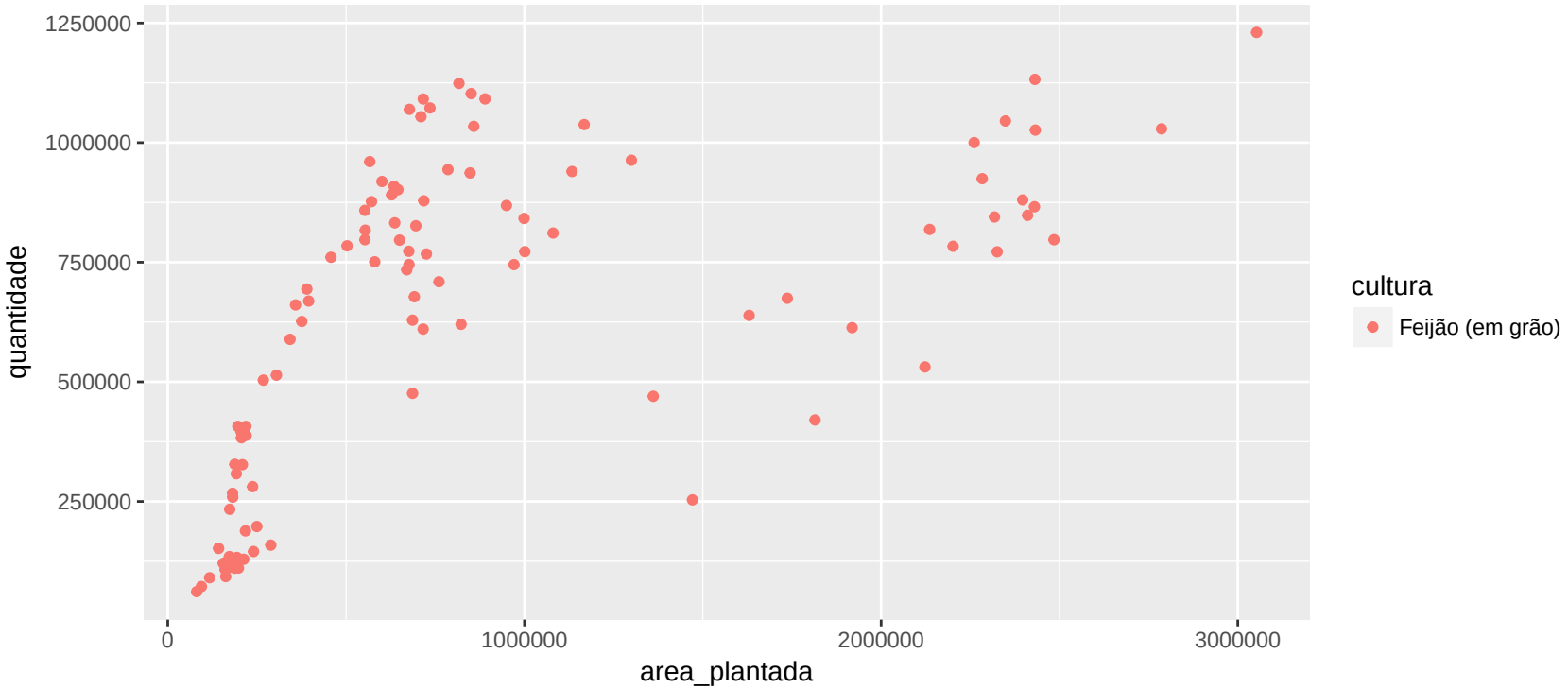
Aspectos Estéticos - Variáveis categóricas

```
ggplot(feijao,aes(x = area_plantada,
                  y = quantidade, shape = cultura)) +
  geom_point()
```



Aspectos Estéticos - Variáveis categóricas

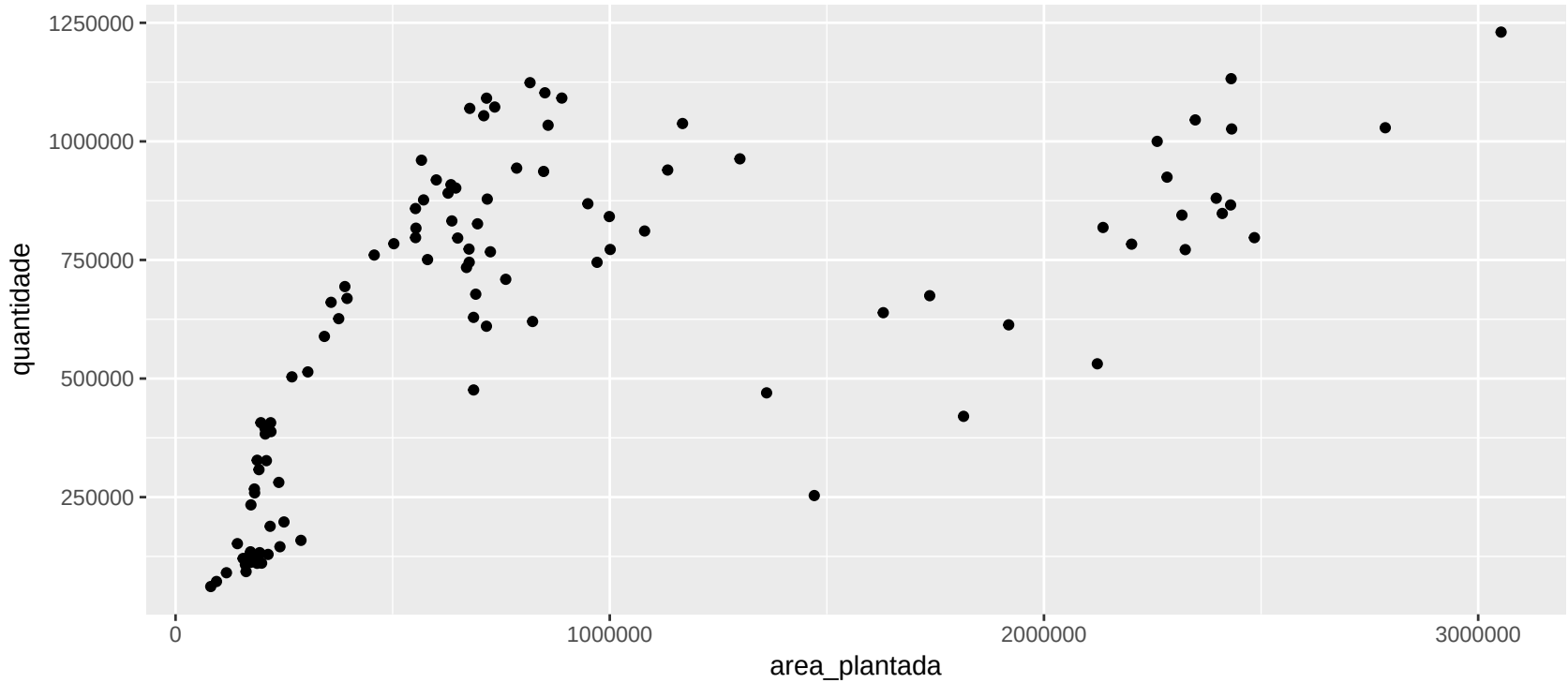
```
ggplot(feijao, aes(x = area_plantada,
                   y = quantidade, col = cultura)) +
  geom_point()
```



Geometrias

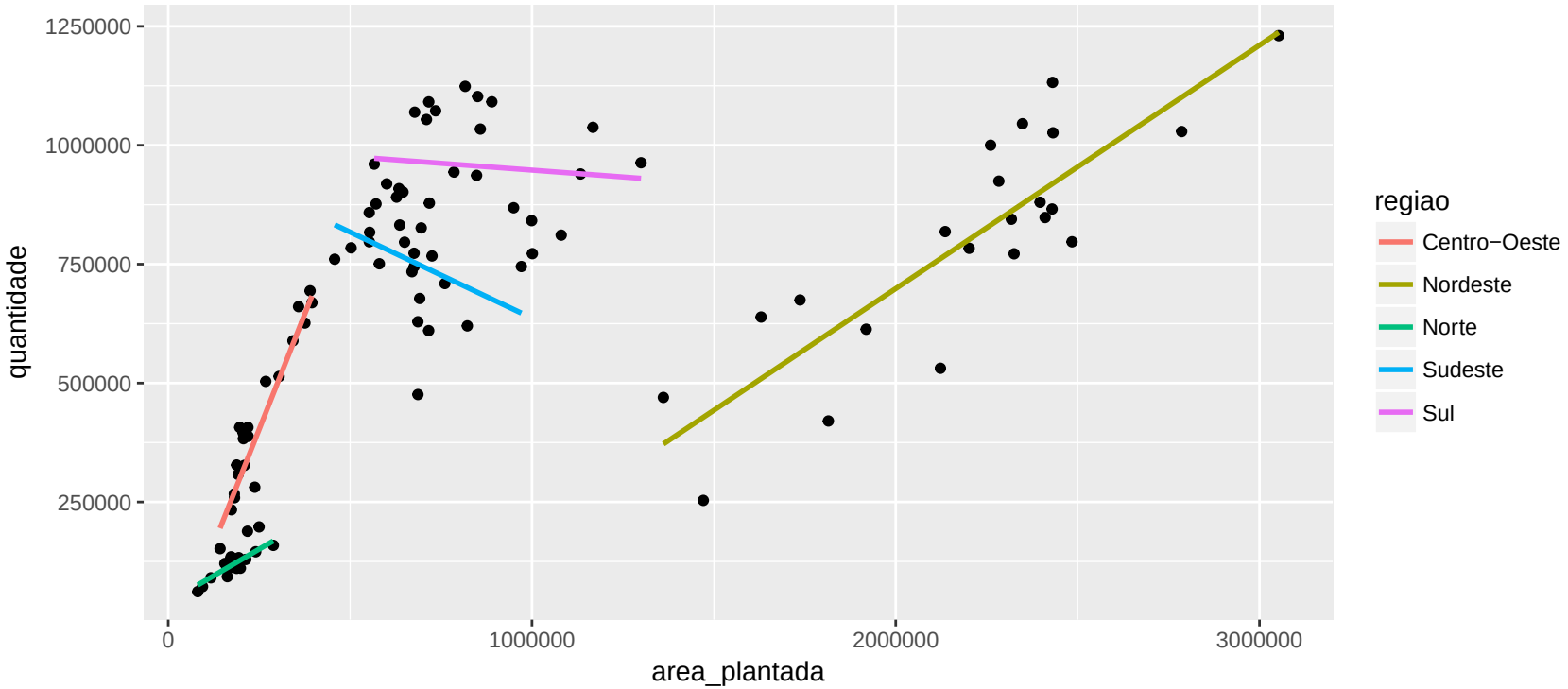
geom_point()

```
ggplot(feijao, aes(x = area_plantada, y = quantidade)) +
  geom_point()
```



geom_smooth()

```
ggplot(feijao, aes(x = area_plantada, y = quantidade)) +
  geom_point() +
  geom_smooth(aes(col = regiao), se = FALSE, method = "lm")
```

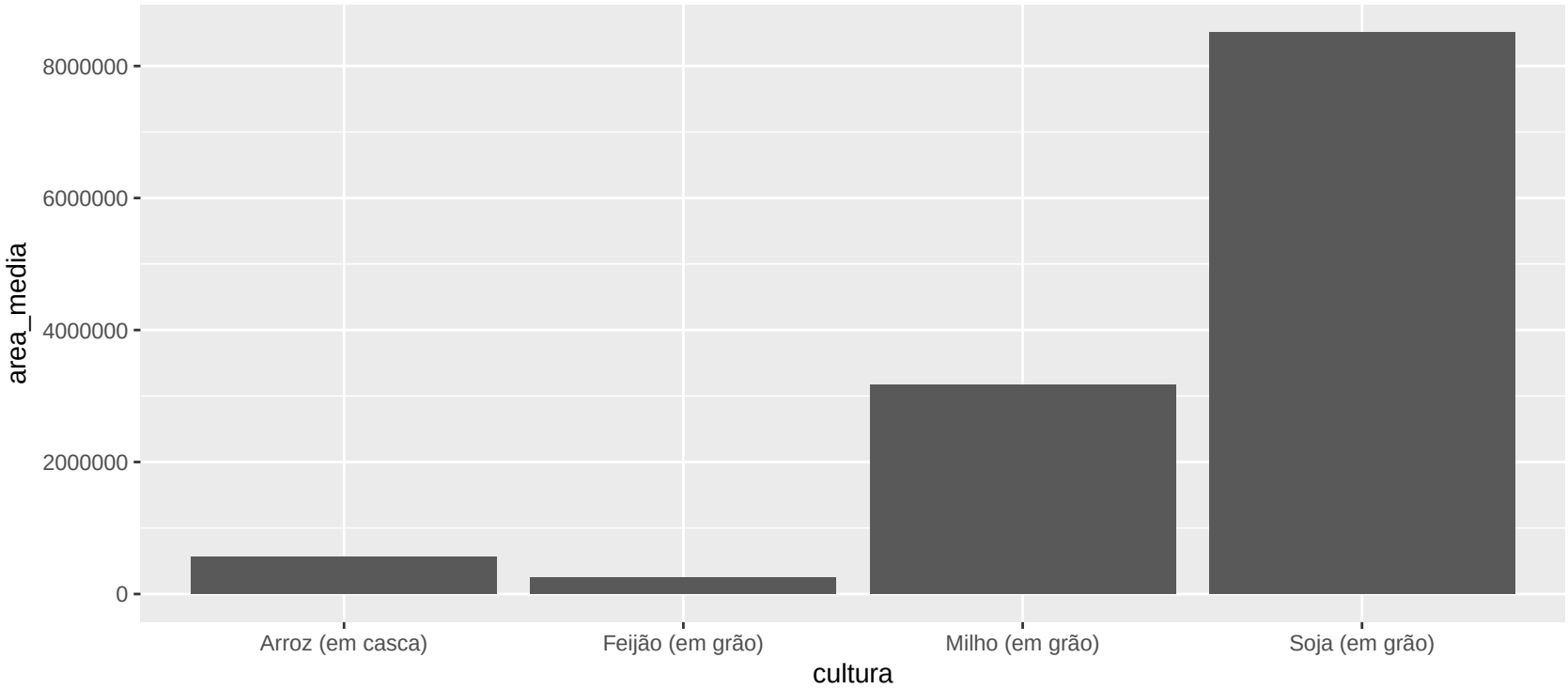


geom_col() ou geom_bar()

```
medias <- CO %>% group_by(cultura) %>%
  summarise(area_media = mean(area_plantada))
```

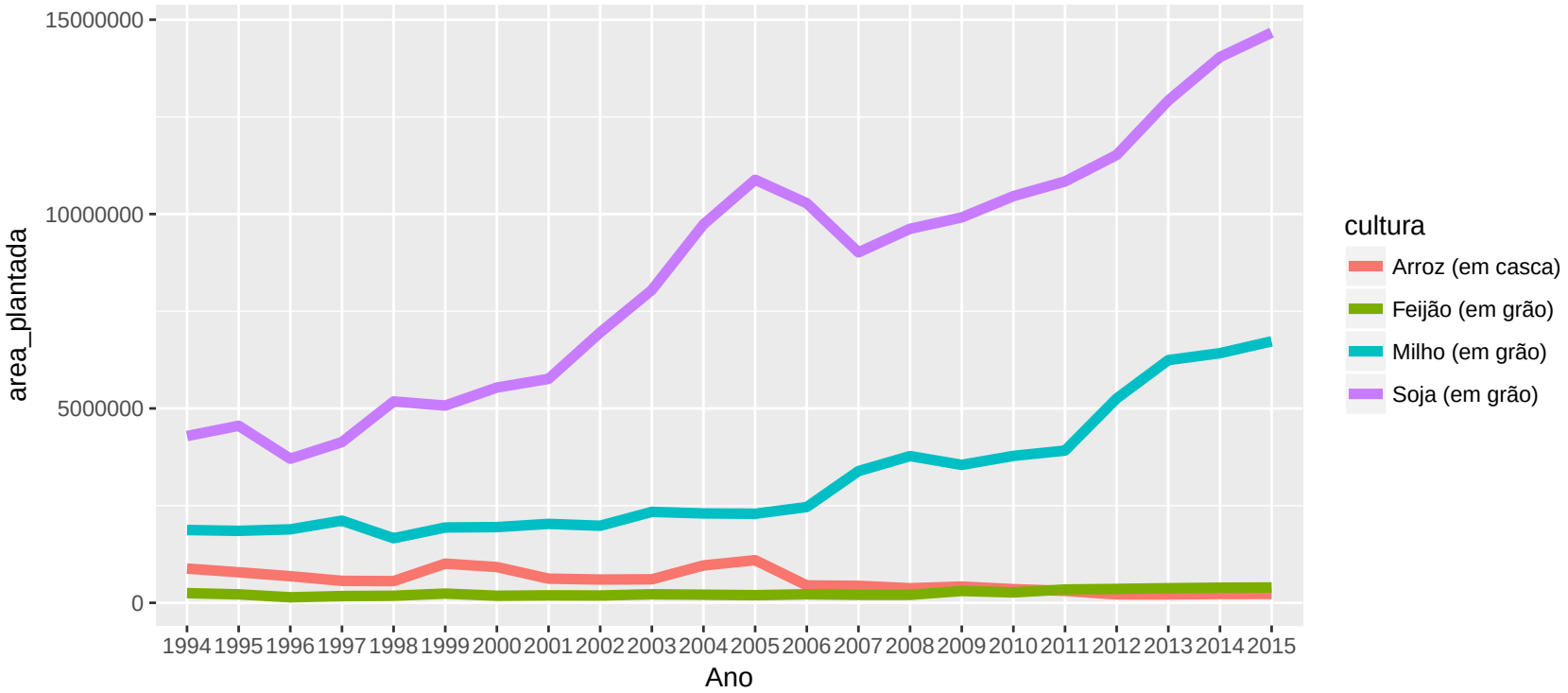
geom_col() ou geom_bar()

```
ggplot(medias, aes(x = cultura, y = area_media)) +
  geom_col() # ou geom_bar(stat = "identity")
```



geom_line()

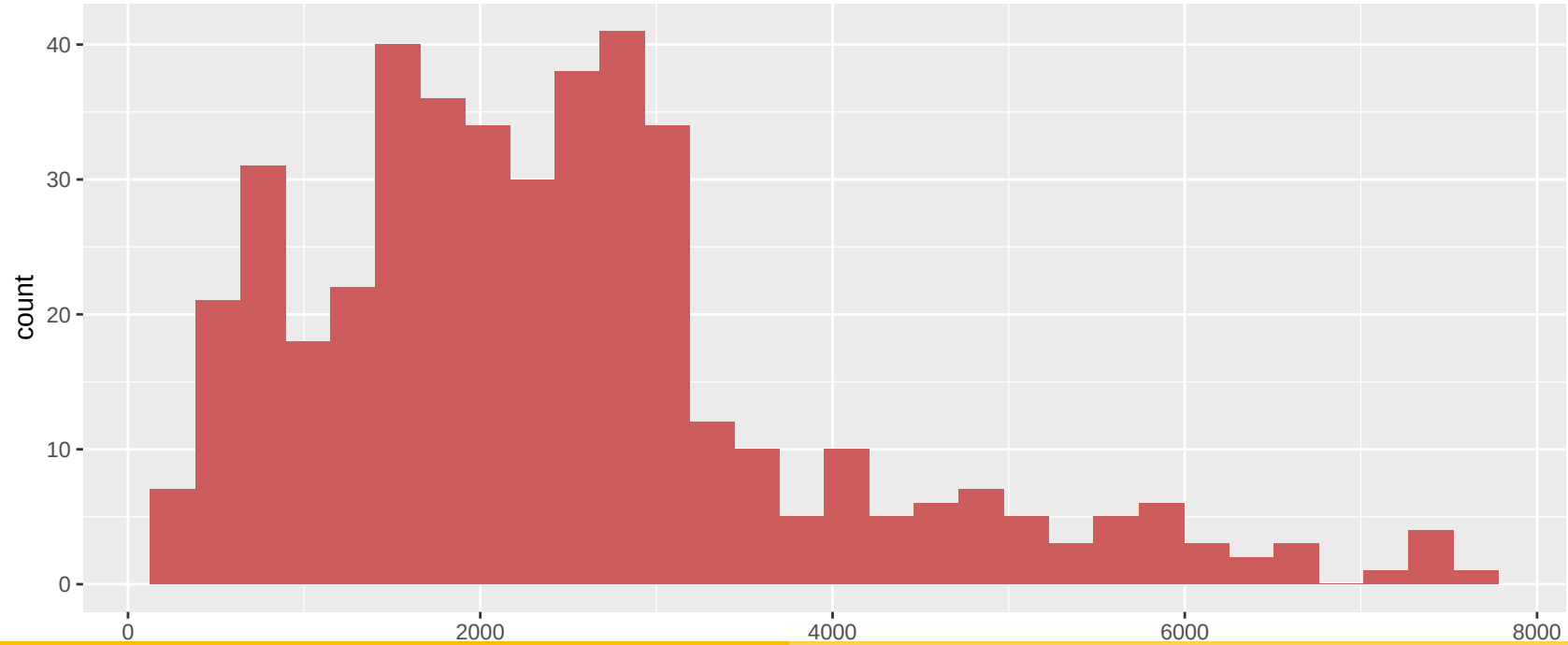
```
ggplot(CO, aes(Ano, area_plantada)) +
  geom_line(aes(col = cultura, group = cultura),
            size = 2)
```



geom_histogram()

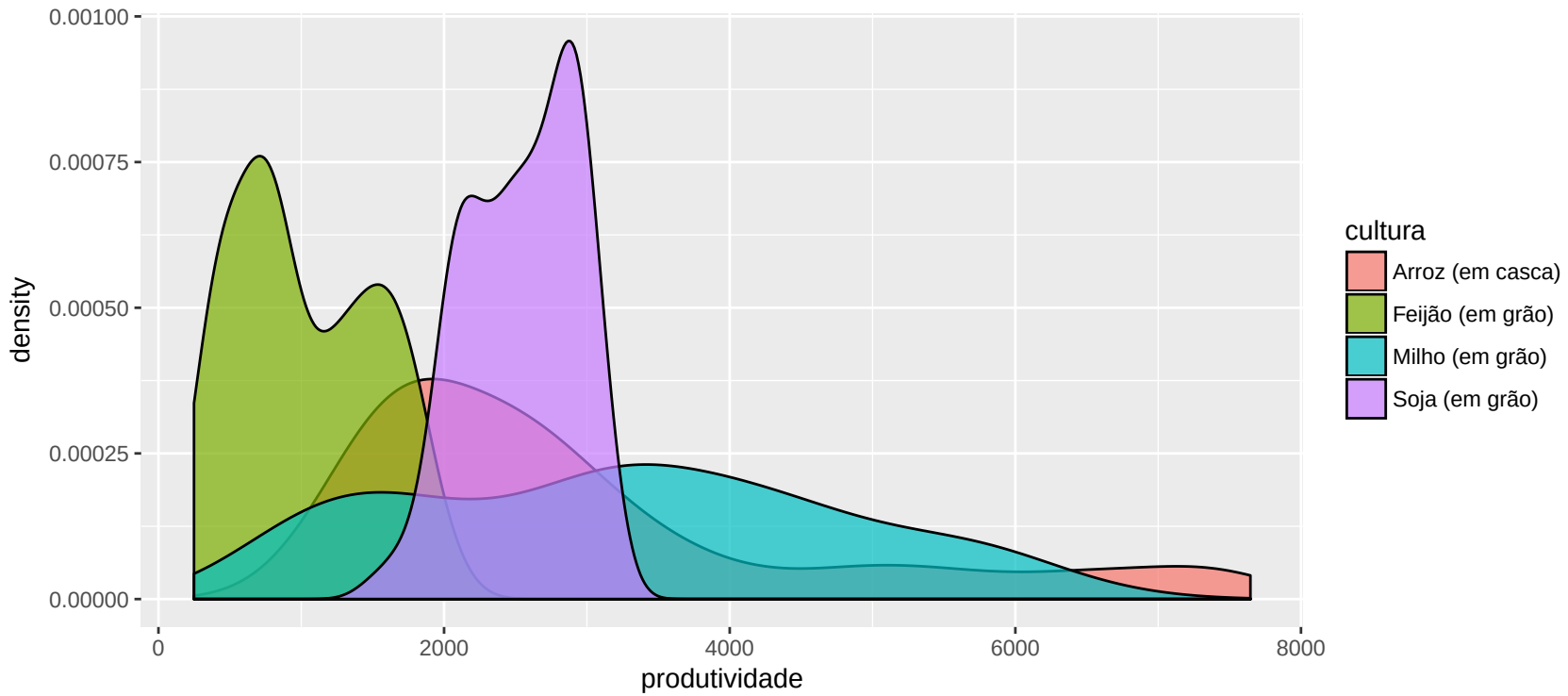
```
ggplot(dados, aes(x = produtividade)) +
  geom_histogram(fill = "indianred") # definir intervalos
```

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



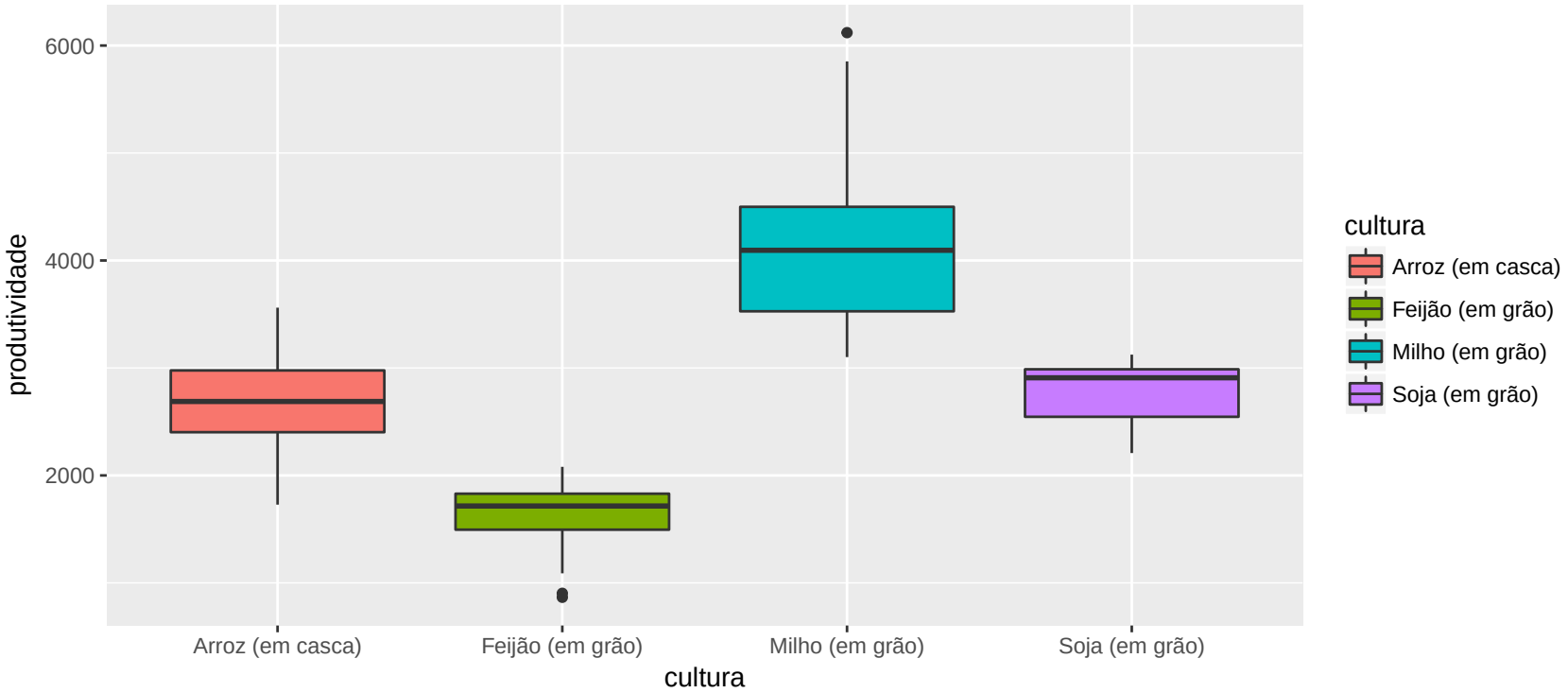
Ou então geom_density()

```
ggplot(dados, aes(x = produtividade, fill = cultura)) +
  geom_density(alpha= 0.7)
```



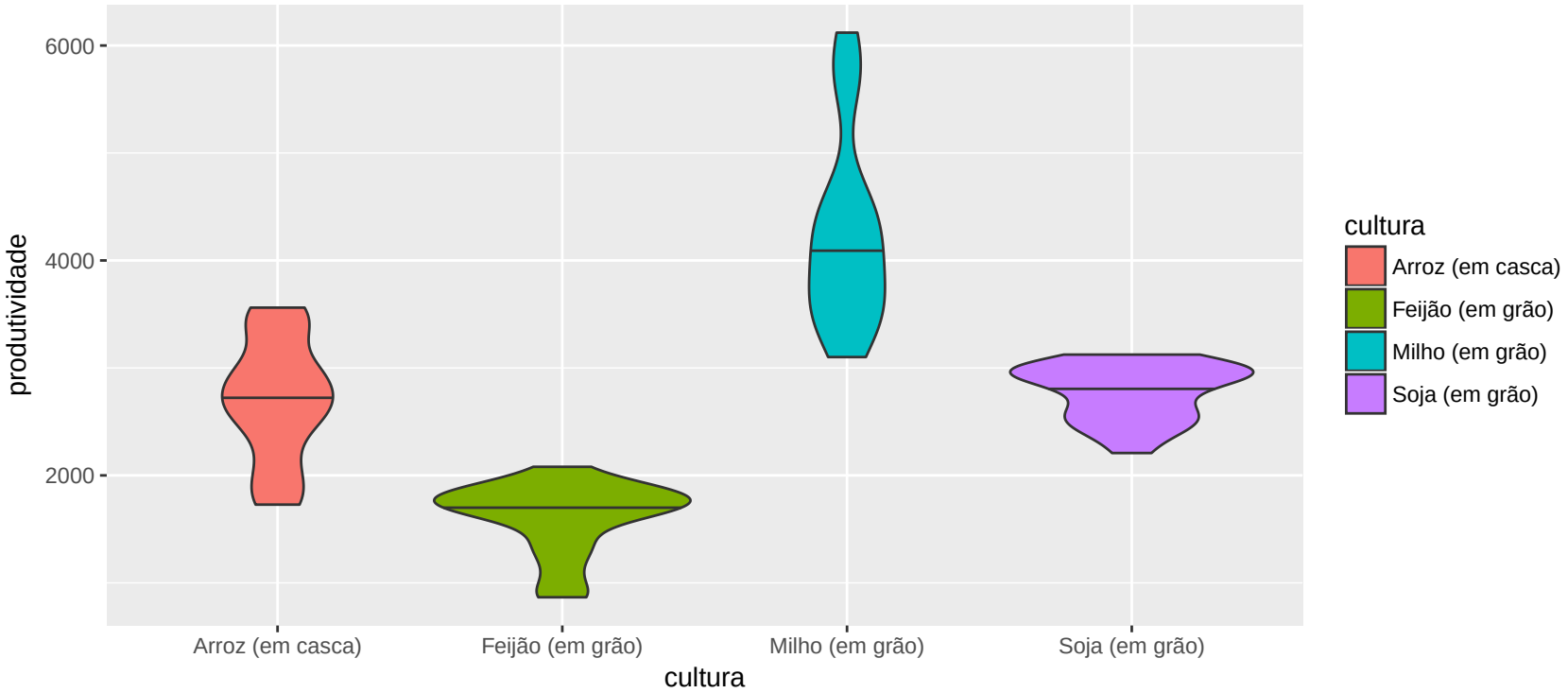
geom_boxplot()

```
ggplot(CO, aes(x = cultura, y = produtividade)) +
  geom_boxplot(aes(fill = cultura))
```



. . . ou geom_violin()

```
ggplot(CO, aes(x = cultura, y = produtividade,
               fill = cultura)) +
  geom_violin(draw_quantiles = 0.5)
```



Exercícios

1. Desenhe um gráfico de violino da produtividade com o conjunto de dados a sua escolha. Adicione uma camada representando as observações com pontos. Dê uma olhada em `?geom_jitter`, pode ser útil.
2. Há grande diferença entre a relação área colhida / área plantada entre as regiões? E Entre as culturas? Responda construindo gráficos. Utilize o conjunto `dados` para esta questão.

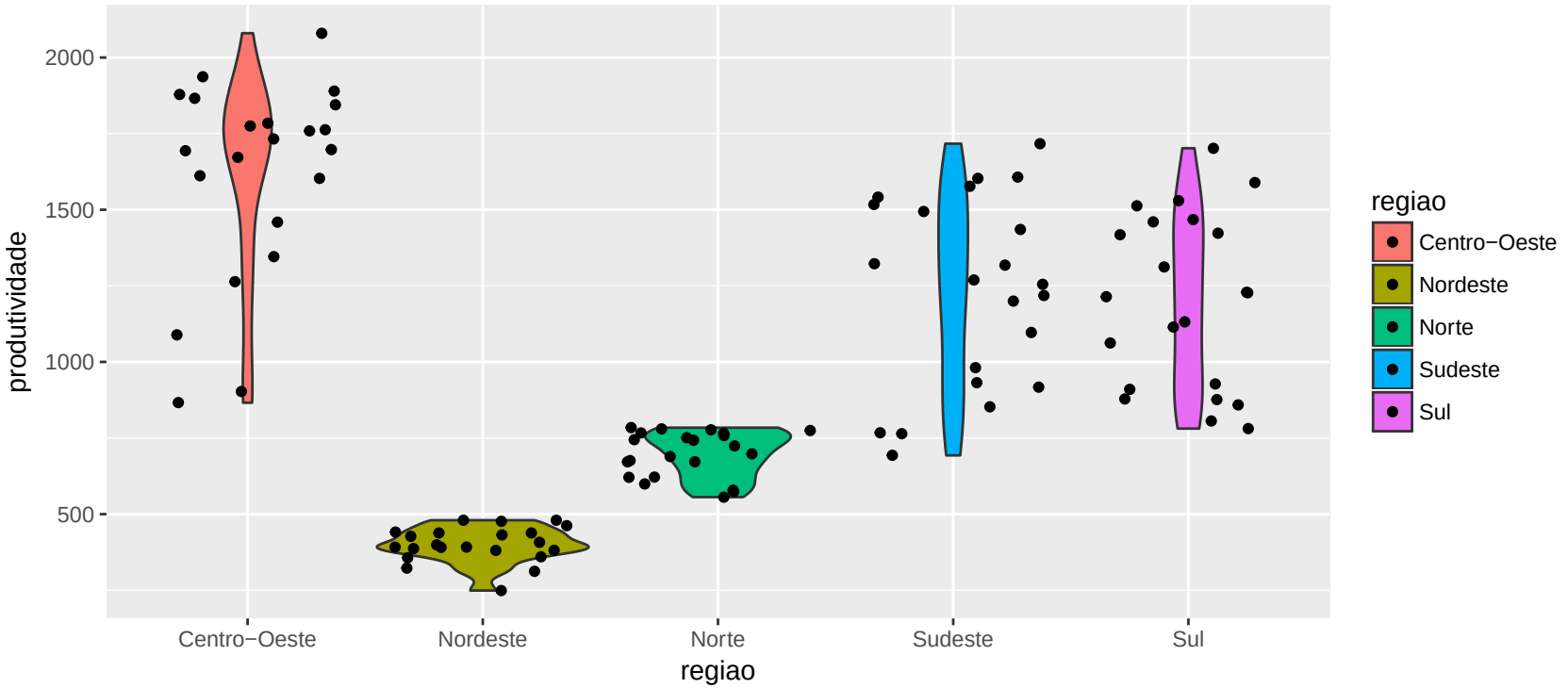
Exercícios

3. Desenhe um gráfico de dispersão com a produtividade em um eixo e a área plantada em outro. Utilize os dados de produção do feijão. Adicione uma reta de regressão para todo o conjunto de dados. Adicione nova camada com uma regressão para região (utilize a cor para diferenciar os grupos). A regressão geral representa bem a relação entre área e produtividade para os subconjuntos?

Extra: Parta do mesmo gráfico de dispersão anterior (produtividade x área). Tente adiconar mais informações no mesmo gráfico mapeando outras variáveis. Busque identificar em que momento as informações idicionais poluem o gráfico.

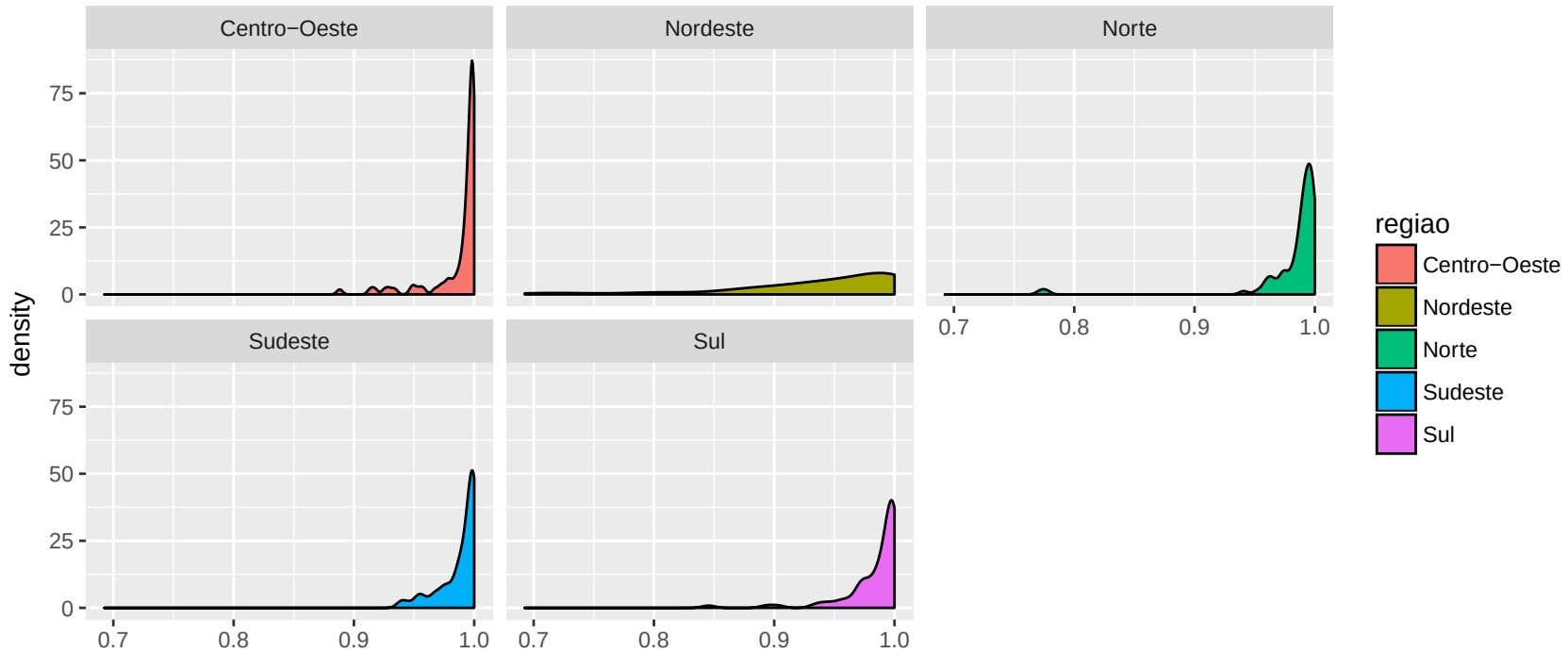
Solução 1

```
ggplot(feijao, aes(x = regiao, y = produtividade, fill = re
  geom_violin() +
  geom_jitter()
```



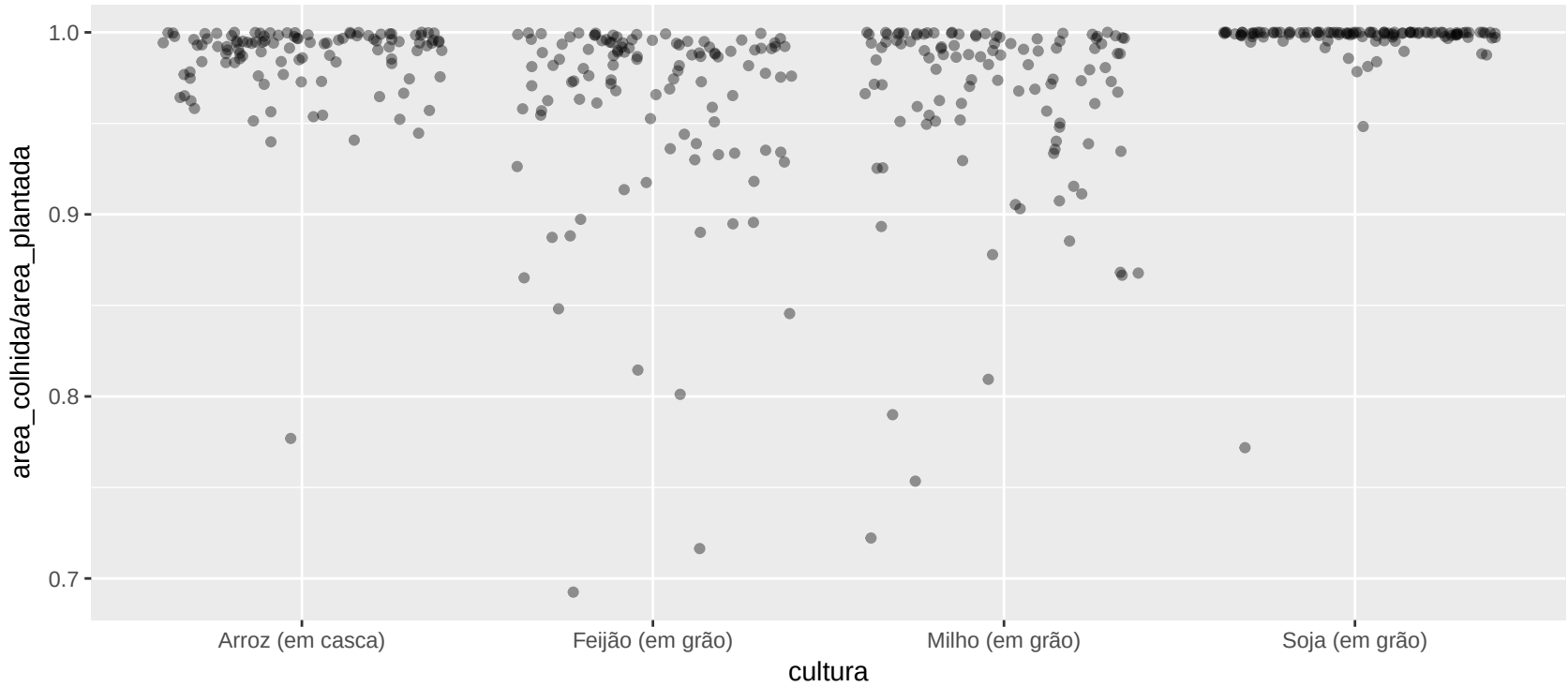
Soluções 2 - a

```
ggplot(dados) +
  geom_density(aes(x = area_colhida / area_plantada,
                   fill = regiao)) +
  facet_wrap(~regiao)
```



Soluções 2 - b

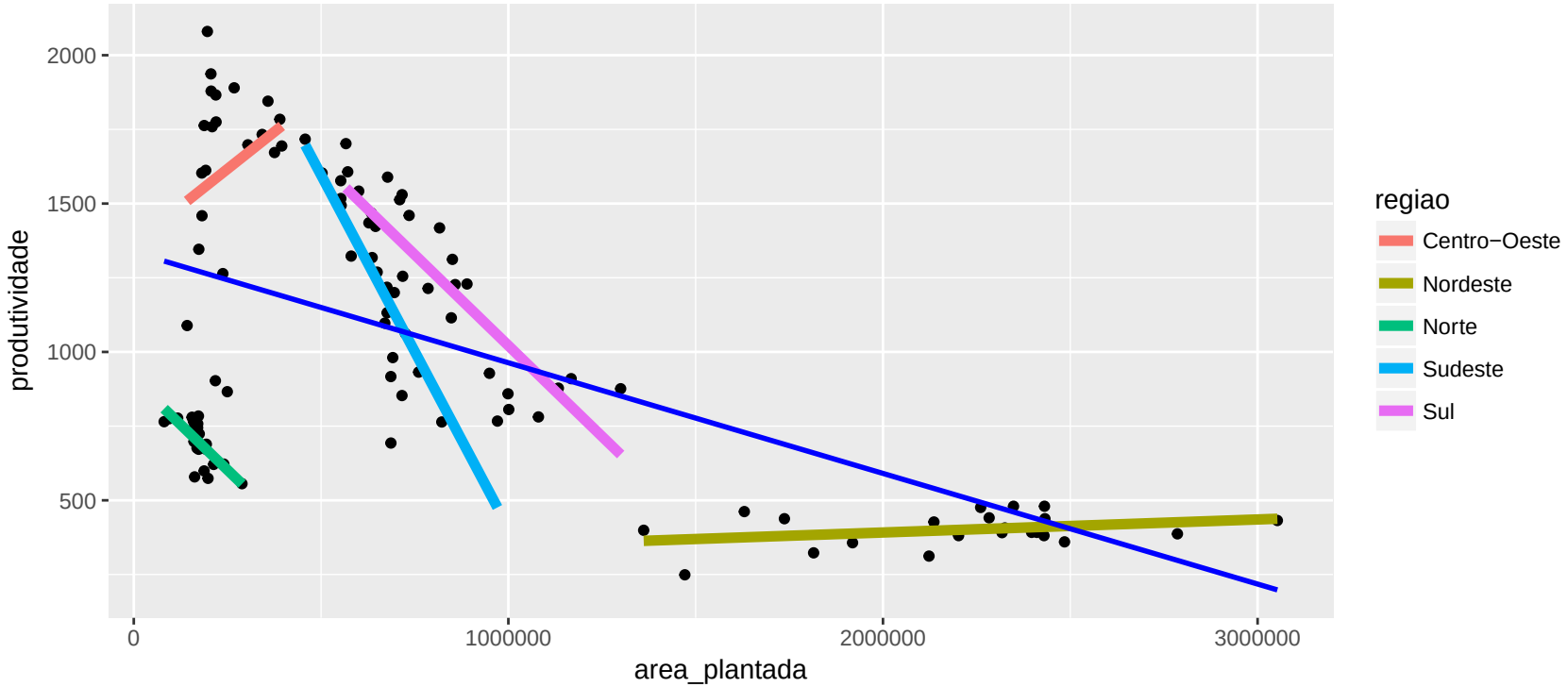
```
ggplot(dados, aes(x = cultura,
                  y = area_colhida / area_plantada)) +
  geom_point(alpha = 0.4, position = "jitter")
```



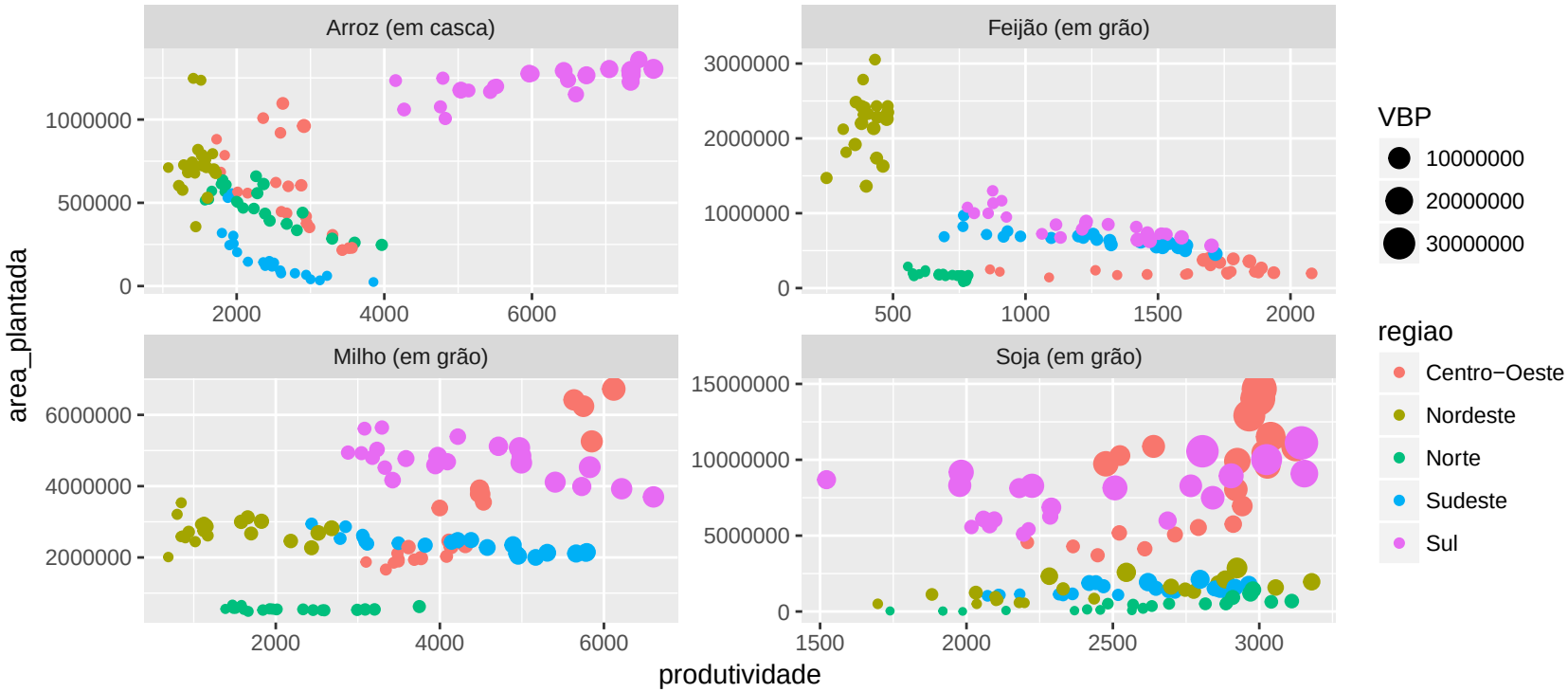
Solução 3

```
ggplot(feijao, aes(x = area_plantada,
                   y = produtividade)) +
  geom_point() +
  geom_smooth(aes(col = regiao),method = "lm",
              se = FALSE, size = 2) +
  geom_smooth(group = 1, col = "blue",
              method = "lm", se = FALSE)
```

Solução 3


produtividade
2000
1500
1000
500
0
1000000
2000000
3000000
area_plantada
regiao
Centro−Oeste
Nordeste
Norte
Sudeste
Sul

Extra


Arroz (em casca)
Feijão (em grão)
Milho (em grão)
Soja (em grão)
area_plantada
produtividade
VBP
10000000
20000000
30000000
regiao
Centro−Oeste
Nordeste
Norte
Sudeste
Sul