Comentários e cultura

```
dados = read_csv(
    here::here("data/participation-per-country.csv"),
    col_types = cols(
        .default = col_double(),
        site = col_character(),
        country = col_character(),
        geo = col_character(),
        four_regions = col_character(),
        eight_regions = col_character(),
        six_regions = col_character(),
        `World bank income group 2017` = col_character()
    )
) %>%
    filter(usuarios > 200)
glimpse(dados)
```

```
## Rows: 121
## Columns: 21
## $ site                          <chr> "StackOverflow", "StackOverflow", "S...
## $ country                       <chr> "Argentina", "Australia", "Austria",...
## $ PDI                           <dbl> 49, 36, 11, 80, 65, 69, 70, 39, 63, ...
## $ IDV                           <dbl> 46, 90, 55, 20, 75, 38, 30, 80, 23, ...
## $ MAS                           <dbl> 56, 61, 79, 55, 54, 49, 40, 52, 28, ...
## $ UAI                           <dbl> 86, 51, 70, 60, 94, 76, 85, 48, 86, ...
## $ usuarios                      <dbl> 2798, 12313, 2518, 2558, 4275, 10717...
## $ responderam_prop              <dbl> 0.5357398, 0.6133355, 0.6310564, 0.3...
## $ perguntaram_prop              <dbl> 0.5210865, 0.5897832, 0.5933280, 0.4...
## $ editaram_prop                 <dbl> 0.09256612, 0.14699911, 0.14932486, ...
## $ comentaram_prop               <dbl> 0.25339528, 0.33395598, 0.35027800, ...
## $ GNI                           <dbl> NA, 59570, 48160, 840, 44990, 11630,...
## $ Internet                      <dbl> 51.0, 79.5, 79.8, 5.0, 78.0, 45.0, 5...
## $ EPI                           <dbl> 59.02, NA, 63.21, NA, 61.21, 49.96, ...
## $ geo                           <chr> "arg", "aus", "aut", "bgd", "bel", "...
## $ four_regions                  <chr> "americas", "asia", "europe", "asia"...
## $ eight_regions                 <chr> "america_south", "east_asia_pacific"...
## $ six_regions                   <chr> "america", "east_asia_pacific", "eur...
## $ Latitude                      <dbl> -34.00000, -25.00000, 47.33333, 24.0...
## $ Longitude                     <dbl> -64.00000, 135.00000, 13.33333, 90.0...
## $ `World bank income group 2017` <chr> "Upper middle income", "High income"...
```

#Análise

O bojetivo desse estudo é analisar a relação entre quanto as pessoas de diferentes países comentam em questões dos outros. A proporção das pessoas do país que comentou nas questões de outros está medido na variável `comentaram_prop`.

Considerando essa variável, queremos examinar a relação entre ela e o quão hierárquicas são as relações em um país (`PDI`). Queremos também levar em conta o quanto as pessoas daquele país têm acesso à Internet

(`Internet`) e qual o tamanho da base de dados que detectamos daquele país (`usuarios`).

##Pré-processamento dos dados

Primeiramente, iremos selecionar apenas os dados que serão utilizados na análise. **Removendo linhas com valores nulos** e alterando alguns dados para facilitar a leitura e entendimento das vizualiações. Abaixo podemos ver como ficaram os dados após essa etapa inicial de pré-processamento.

```
dados = dados %>% select(c('country','PDI','Internet','usuarios','comentaram_prop','six_regions')) #Fil

dados = na.omit(dados) #Removendo linhas com valores NA

dados <- dados %>% mutate(six_regions = case_when(
    six_regions == 'america' ~ "América",
    six_regions == 'east_asia_pacific' ~ "Ásia Pacífico",
    six_regions == 'europe_central_asia' ~ "Europa e Ásia Central",
    six_regions == 'south_asia' ~ "Ásia do Sul",
    six_regions == 'middle_east_north_africa' ~ "Oriente Médio e Norte da África",
    six_regions == 'sub_saharan_africa' ~ "África subsariana",
    TRUE ~ six_regions
  )
)

glimpse(dados)
```

```
## Rows: 115
## Columns: 6
## $ country         <chr> "Argentina", "Australia", "Austria", "Bangladesh", ...
## $ PDI             <dbl> 49, 36, 11, 80, 65, 69, 70, 39, 63, 80, 67, 35, 73,...
## $ Internet        <dbl> 51.0, 79.5, 79.8, 5.0, 78.0, 45.0, 51.0, 83.0, 52.3...
## $ usuarios        <dbl> 2798, 12313, 2518, 2558, 4275, 10717, 1463, 17591, ...
## $ comentaram_prop <dbl> 0.25339528, 0.33395598, 0.35027800, 0.15989054, 0.3...
## $ six_regions     <chr> "América", "Ásia Pacífico", "Europa e Ásia Central"...
```

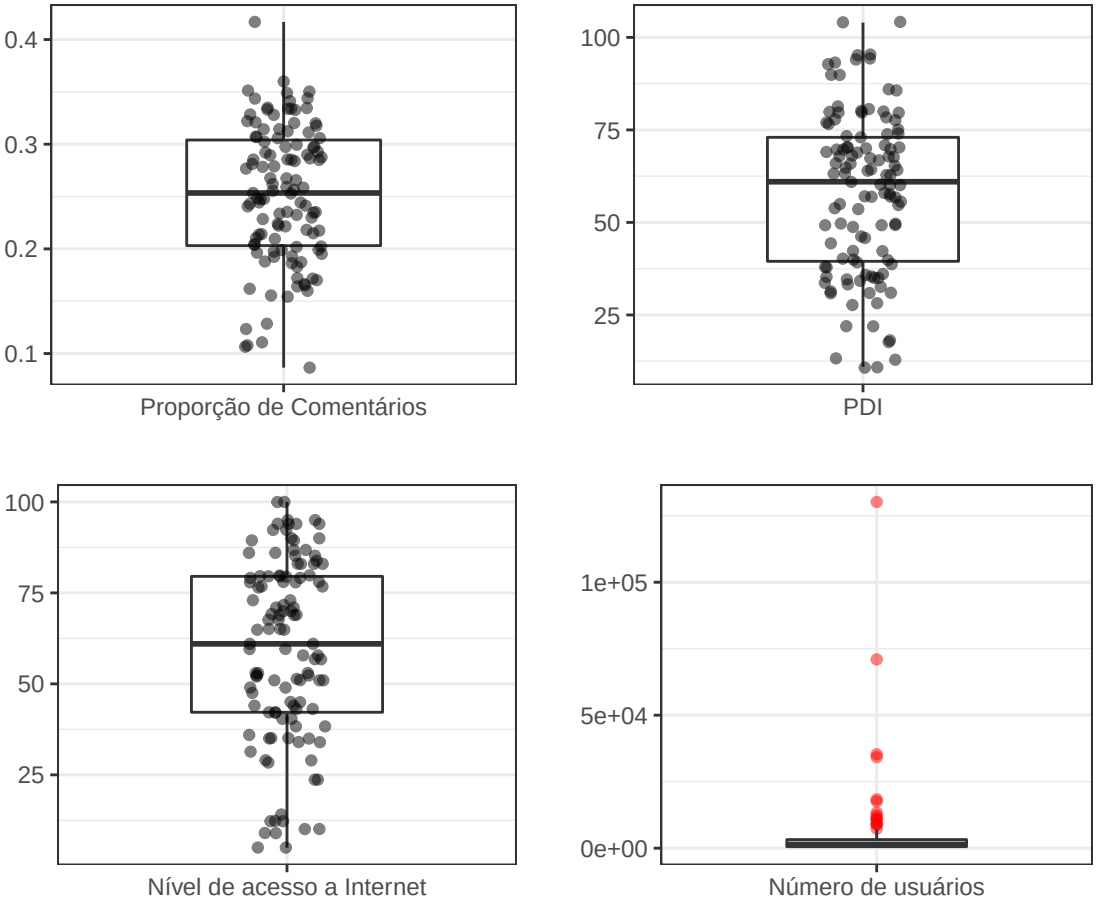
##Análise exploratória dos dados

Finalizada a etapa inicial de pré-processamento dos dados, foi realizado uma análise individual em cada uma das variáveis que será utilizadas na pesquisa, para poder visualizar como se comportam e como é a distribuição dos dados. Sendo assim, é possível descobrir a priori se existe inconsistencia nos dados, valores incomuns, que podem afetar a análise a ser realizada. Abaixo podemos ver os boxsplots de distribuição de cada uma das variáveis. Analisando as figuras abaixo, podemos ver que nas três primeiras variáveis (proporção de comentários, PDI e Nível de acesso a Internet) não aparentam ter valores incomuns e estão bem distribuidas. Porém, a variável "Número de usuários" parece existir muitos valores perto do 0 e poucos valores muito grandes (pontos vermelhos), que seriam *outliers* nos dados.

Naturalmente, o que deveria ser feito é outro processamento para remover esses valores atípicos. Porém, como o conjunto de dados é limitado e cada amostra do conjunto representa uma país, remover esses dados implicaria em remoção daquele país na análise, o que por sua vez, poderia enfraquecer a análise em geral. Outro ponto a ser destacado é que, como a relação principal se dará pelas variáveis "Proporção de comentários" e "PDI", os *outliers* da variável "Número de usuários" não afetará essa relação. Por fim, foi decidido manter todos os dados.

```
pc <- ggplot(dados, aes(x = 'Proporção de Comentários', y = comentaram_prop)) +
  geom_boxplot(width = 0.5, outlier.colour = "red") +
  geom_jitter(width = 0.1, alpha = .5)+
  ylab("")+
  xlab("")
```

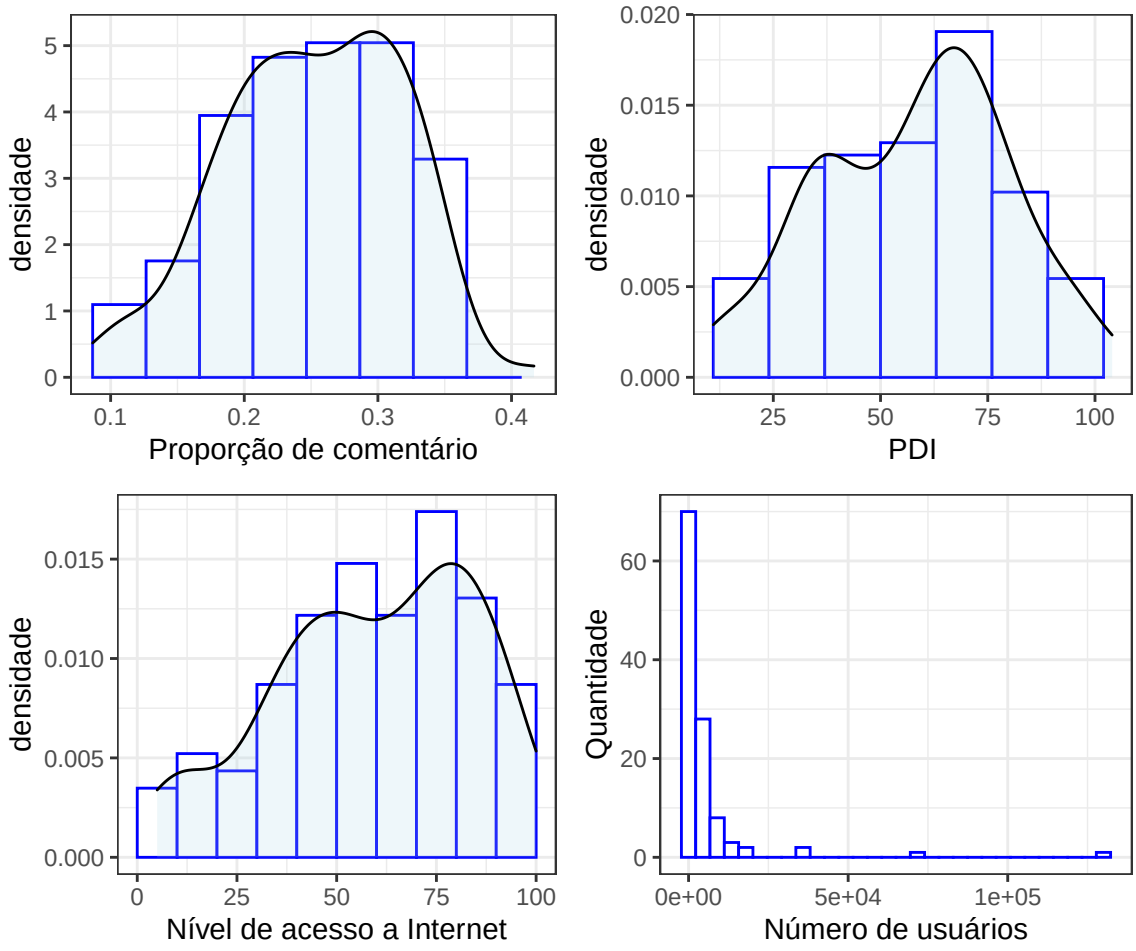
```r
pdi <- ggplot(dados, aes(x = 'PDI', y = PDI)) +
  geom_boxplot(width = 0.5, outlier.colour = "red") +
  geom_jitter(width = 0.1, alpha = .5)+
  ylab("")+
  xlab("")

internet <- ggplot(dados, aes(x = 'Nível de acesso a Internet', y = Internet)) +
  geom_boxplot(width = 0.5, outlier.colour = "red") +
  geom_jitter(width = 0.1, alpha = .5)+
  ylab("")+
  xlab("")

usuarios <- ggplot(dados, aes(x = 'Número de usuários', y = usuarios)) +
  geom_boxplot(width = 0.5, alpha = .5, outlier.colour = "red") +
  ylab("")+
  xlab("")


grid.arrange(pc, pdi, internet, usuarios, ncol = 2)
```



Outra maneira de analisar a distribuição das mesmas variáveis é através de histrogramas e gráficos de densidado. Como podemos ver abaixo, as três primeiras variáveis parecem ter distribuições normais, enquato a quatra variável tem uma alta concetração de valores pequenos e alguns raros valores grandes.

```r
pc <- ggplot(dados, aes(x=comentaram_prop)) +
 geom_histogram(aes(y=..density..),
                breaks=seq(min(dados$comentaram_prop), max(dados$comentaram_prop), by=0.04), fill="white
 geom_density(alpha=.2, fill="lightblue") +
 xlab("Proporção de comentário") +
 ylab("densidade")

pdi <- ggplot(dados, aes(x=PDI)) +
 geom_histogram(aes(y=..density..),
                breaks=seq(min(dados$PDI), max(dados$PDI), by=13), fill="white", col="blue") +
 geom_density(alpha=.2, fill="lightblue") +
 xlab("PDI") +
 ylab("densidade")

internet <- ggplot(dados, aes(x=Internet)) +
 geom_histogram(aes(y=..density..),
                breaks=seq(0, max(dados$Internet), by=10), fill="white", col="blue") +
 geom_density(alpha=.2, fill="lightblue") +
 xlab("Nível de acesso a Internet") +
 ylab("densidade")

usuarios <- ggplot(dados, aes(x=usuarios)) +
 geom_histogram(fill="white", col="blue") +
 xlab("Número de usuários") +
 ylab("Quantidade")


grid.arrange(pc, pdi, internet, usuarios, ncol = 2)
```

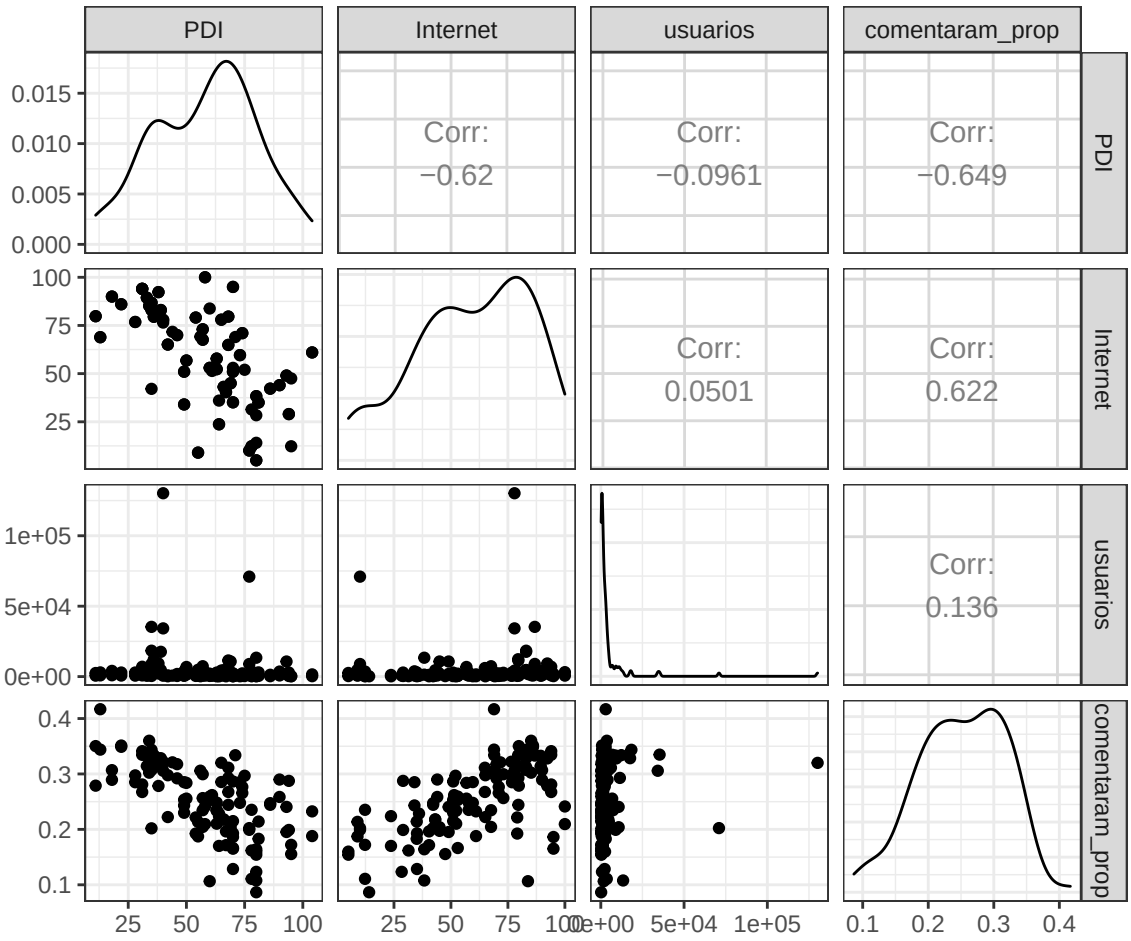
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Examinando essa relação

Uma primeira análise simples que pode ser feita é a relação entre todas as variáveis do conjunto de dados. Logo abaixo, é possível perceber que a correlação mais forte que existe no conjunto, é entre a "Proporção de comentários" e "PDI" com um valor de **-0.649** considerada uma correlação negativa moderada. Outra correlação moderada que podemos destacar é entre "Proporção de comentários" e "Nível de acesso a Internet" com um valor de **0.622** correlação positiva moderada. Intuitivamente faz sentido essa variáveis terem uma correlação moderada ou forte, pois quanto mais acesso a internet tem um país mais ele poderá comentar em questões dos outros.

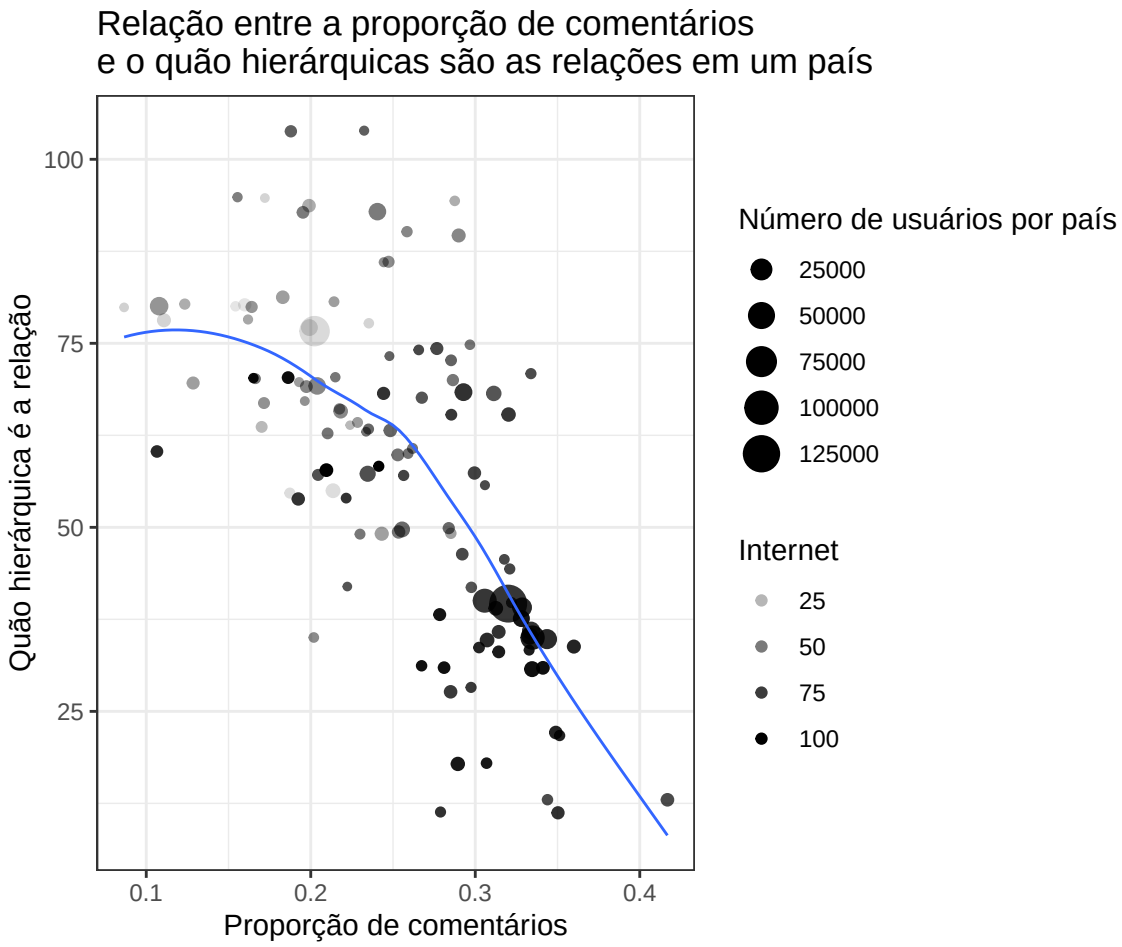
```
dados %>% select(-country, -six_regions) %>% ggpairs(progress = F)
```



Agora, analisando a relação entre as quatro variáveis juntas. A partir da figura apresentada abaixo, podemos perceber obter algumas informação. Primeiro, podemos perceber que quanto maior o nível de acesso a internet de um país menor, menos hierárquica é as relações. Isso se destaca pois, em média, os pontos mais escuros estão mais abaixo, no eixo y, do que os mais claros. Segundo, podemos também perceber, que o número de usuários não parece ter relação com nenhuma das duas outras três variáveis. Pois, além do tamanho dos pontos, que representa o número de usários, estar distribuida uniformente por todo o gráfico (não parece ter pontos maiores ou menores concentrados em uma região específica). Támbém, é possível ver que existem pontos de diferentes tamanhos com cores mais escuras e mais claras, o que indica que não há relação entre as variáveis "Número de usuários" e "Nível de acesso a internet". Por fim, podemos concluir que quanto maior a proporção de comentários, menos hierárquicas são as relações e maior é o acesso a internet.

```
ggplot(data = dados) +
  geom_point(mapping = aes(x = comentaram_prop, y = PDI, alpha = Internet, size = usuarios), position =
  geom_smooth(mapping = aes(x = comentaram_prop, y = PDI), size = .5, se = FALSE) +
  scale_size(name="Número de usuários por país") +
  ggtitle("Relação entre a proporção de comentários\ne o quão hierárquicas são as relações em um país")
  xlab("Proporção de comentários") +
  ylab("Quão hierárquica é a relação")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



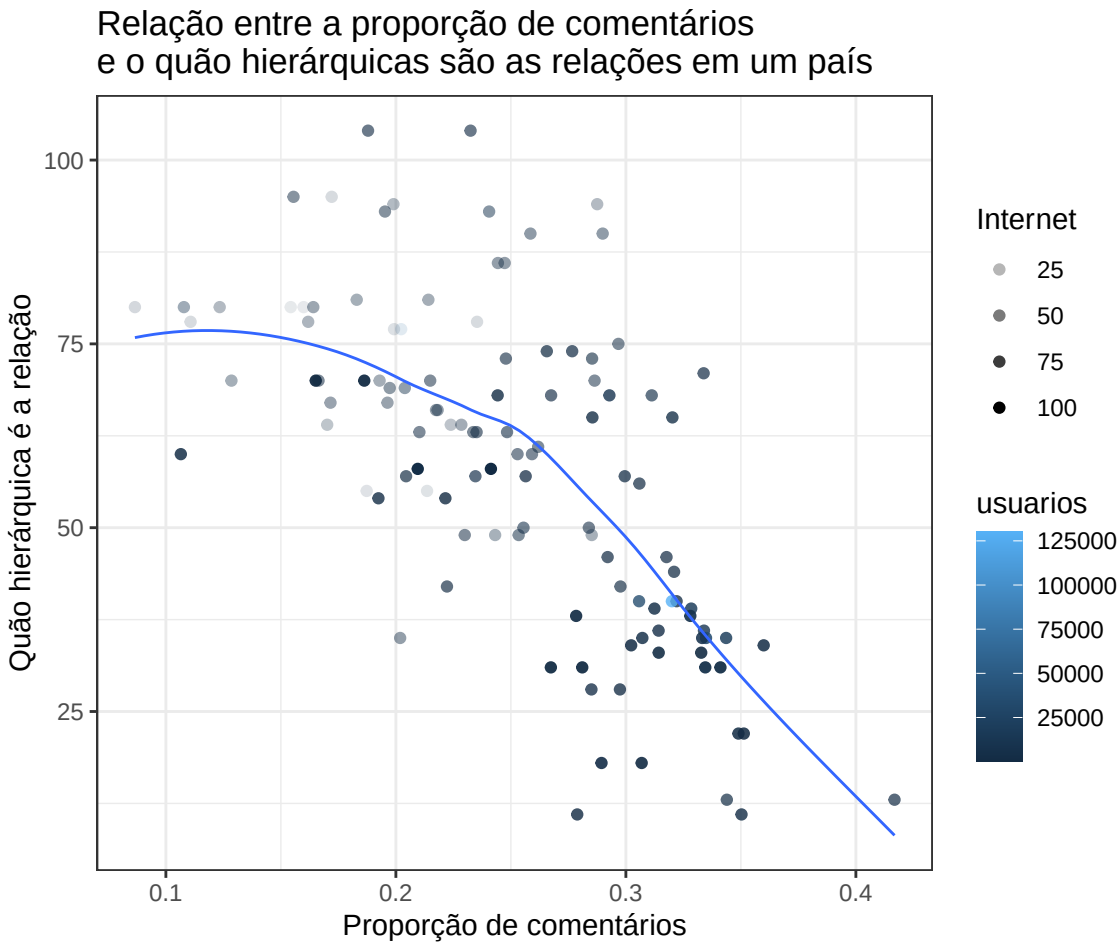
Outras formas de ver

Em seguida, faça 5 visualizações que usem as mesmas variáveis e também pontos, mas que sejam **menos eficazes** que a que você escolheu acima.

Primeira maneira, de ver a mesma análise só que de forma menos eficaz é utilizar uma variável contínua na característica de cor, humanos ter dificuldade de diferenciar tons de cores muito semelhantes. Como podemos ver no gráfico abaixo, se torna mais difícil saber, quais países tem mais ou menos usuários. Cores é mais comumente usado para rotular e categorizar elementos

```
ggplot(data = dados) +
  geom_point(mapping = aes(x = comentaram_prop, y = PDI, alpha = Internet, color = usuarios)) +
  geom_smooth(mapping = aes(x = comentaram_prop, y = PDI), size = .5, se = FALSE) +
  scale_size(name="Número de usuários por país") +
  ggtitle("Relação entre a proporção de comentários\ne o quão hierárquicas são as relações em um país") +
  xlab("Proporção de comentários") +
  ylab("Quão hierárquica é a relação")
```

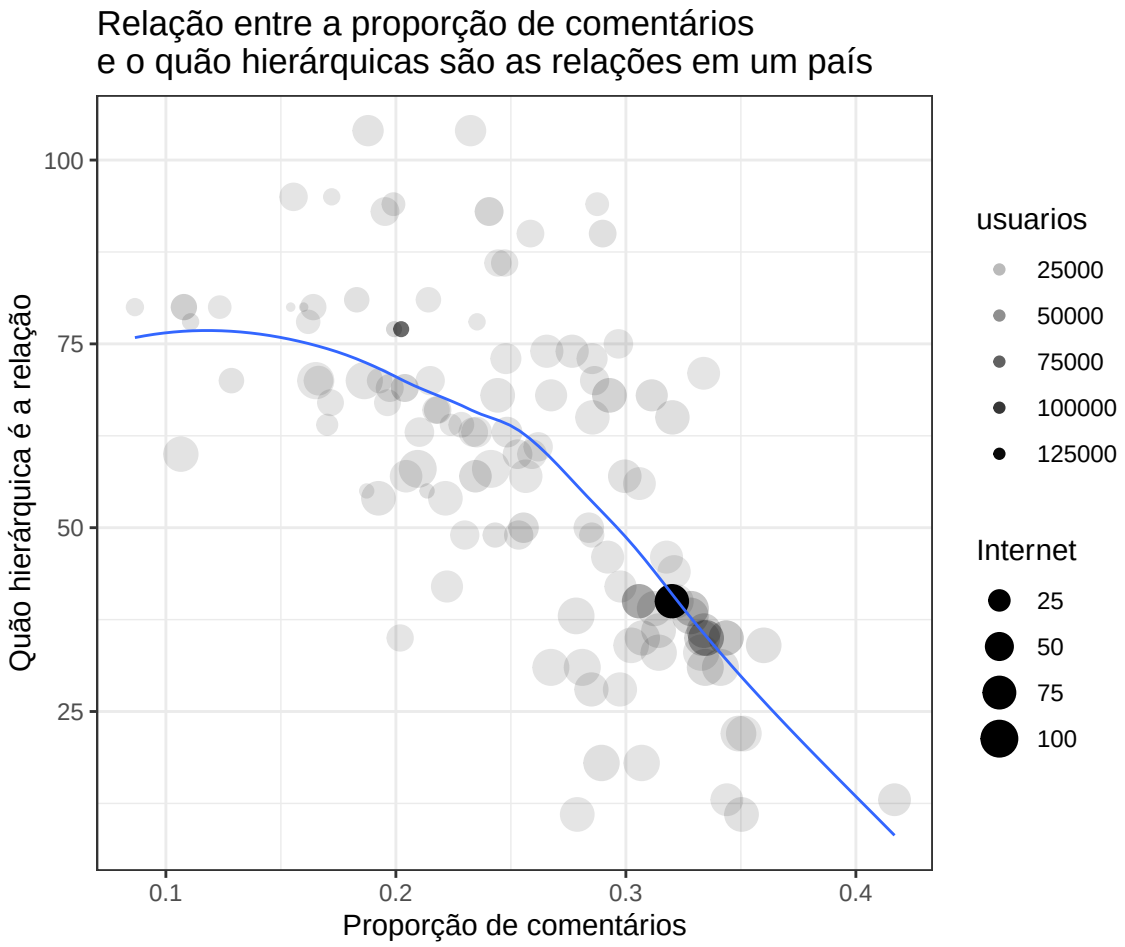
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Segunda maneira, é utilizar o tamanho dos pontos para representar o nível de acesso a Internet. Como a variável "Nível de acesso a internet" tem muitos valores altos o gráfico fica cheio de pontos grandes o que pode dificultar a leitura, pois muitos pontos se sobrepoem. Outro problema, é utilizar a claridade do pontos para representar o "Número de usuários", pois ao contrário do "nível de acesso", essa variável possui muitos valores baixos, fazendo com que a maioria dos pontos fique muito clara.

```
ggplot(data = dados) +
  geom_point(mapping = aes(x = comentaram_prop, y = PDI, alpha = usuarios, size = Internet)) +
  geom_smooth(mapping = aes(x = comentaram_prop, y = PDI), size = .5, se = FALSE) +
  scale_size(name="Internet") +
  ggtitle("Relação entre a proporção de comentários\ne o quão hierárquicas são as relações em um país") +
  xlab("Proporção de comentários") +
  ylab("Quão hierárquica é a relação")
```

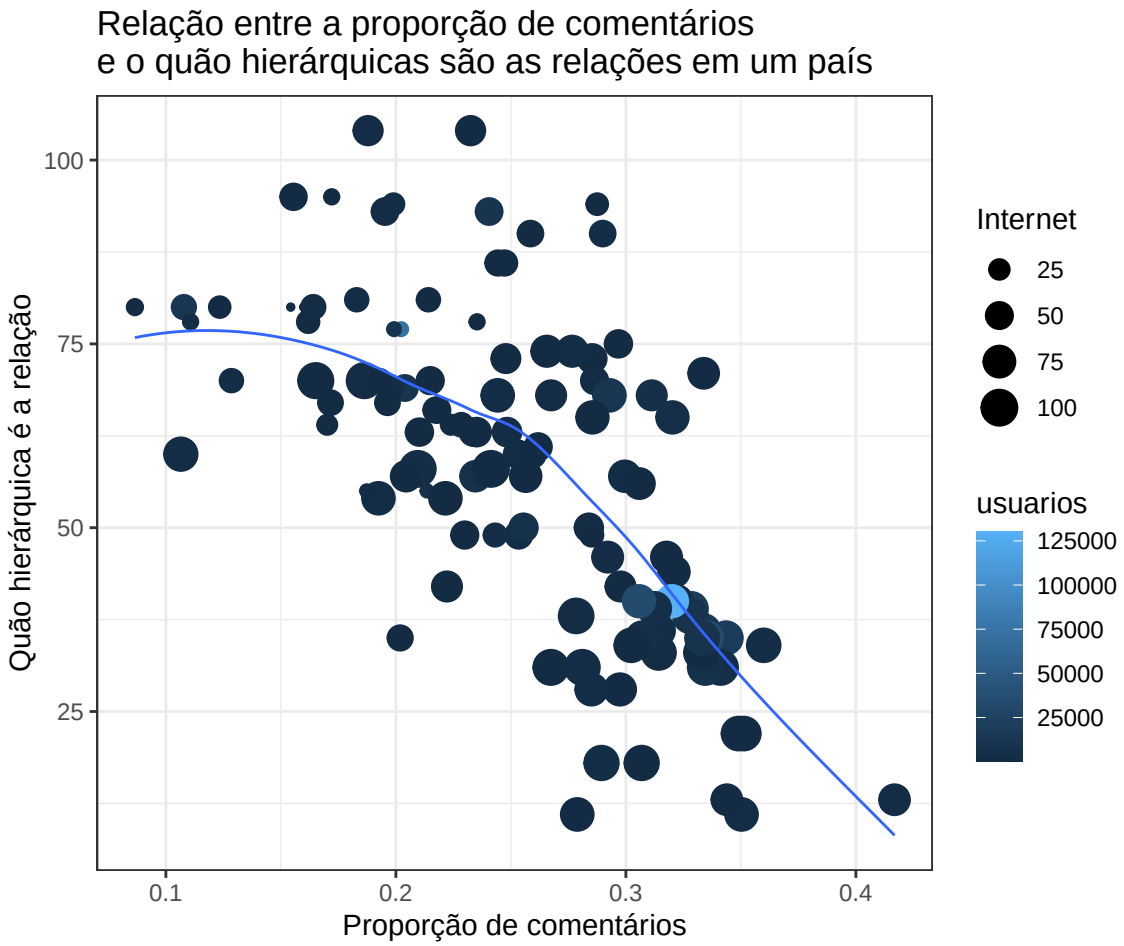
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Terceira maneira, seria utilizar a cor para representar o número de usuários. Como na primeira maneira, um dos problemas é utilizar cores para representar uma variável contínua. O segundo problema é que, a distribuição dos dados da variável "Número de usuários" tem alta concentração de valores pequenos, fazendo com que o gráfico se quase todo preenchido por pontos mais escuros o que deixa uma leitura mais complexa das informações apresentadas.

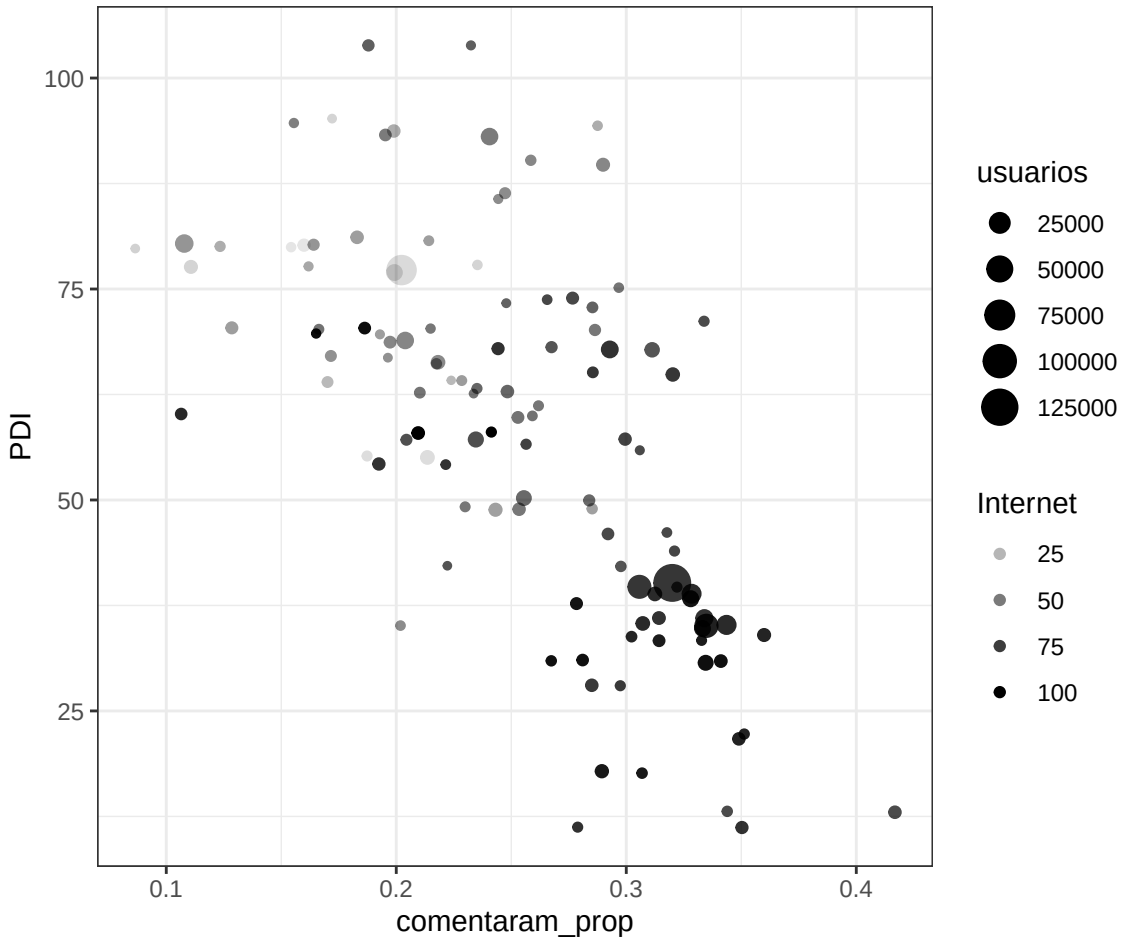
```
ggplot(data = dados) +
  geom_point(mapping = aes(x = comentaram_prop, y = PDI, color = usuarios, size = Internet)) +
  geom_smooth(mapping = aes(x = comentaram_prop, y = PDI), size = .5, se = FALSE) +
  scale_size(name="Internet") +
  ggtitle("Relação entre a proporção de comentários\ne o quão hierárquicas são as relações em um país") +
  xlab("Proporção de comentários") +
  ylab("Quão hierárquica é a relação")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



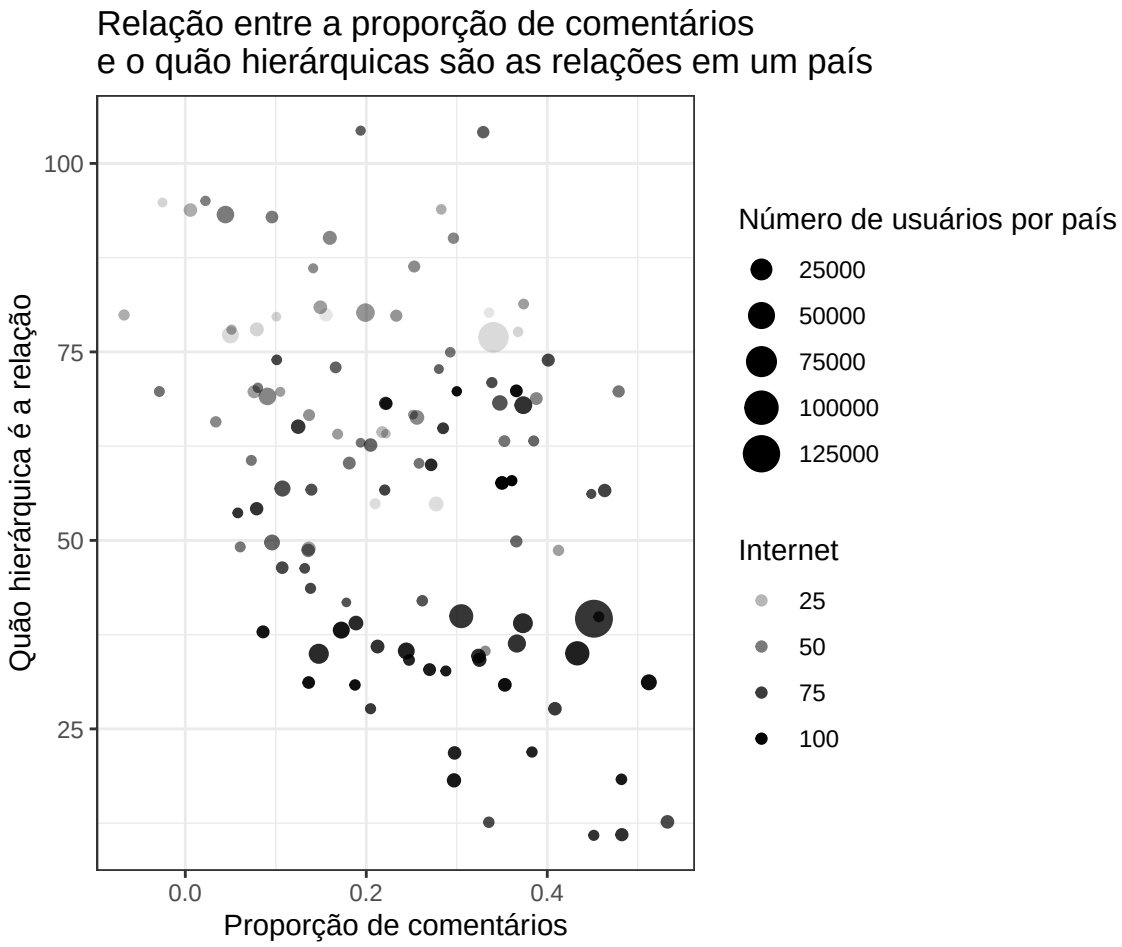
Quarta maneira, nesse caso não foi mexido na representação das variáveis do gráfico original. Porém, foram removidos todas as legendas, isso acaba dificultando a leitura por um público que não conhece bem os dados a serem estudos. Por exemplo, um pessoa não vai saber o que signigica a variável "PDI" no eixo y.

```
ggplot(data = dados) +
  geom_point(mapping = aes(x = comentaram_prop, y = PDI, alpha = Internet, size = usuarios), position =
```



Quinta maneira, essa é uma maneira de poder enganar o público adicionando pequena quantidade de ruído aleatório a cada ponto através da função jitter que espalha dos dados pelo mapa. Essa função embora torne seu gráfico menos preciso em pequenas escalas, como no exemplo abaixo, torna seu gráfico mais revelador em grandes escalas quando existe muita sobreposição de dados. Como nosso exemplo possui poucos dados, não seria uma boa abordagem. Podemos ver que esse ruído, deu a impressão que não existe relação entre as variáveis.

```
ggplot(data = dados) +
  geom_jitter(mapping = aes(x = comentaram_prop, y = PDI, alpha = Internet, size = usuarios), width = 0
  scale_size(name="Número de usuários por país") +
  ggtitle("Relação entre a proporção de comentários\ne o quão hierárquicas são as relações em um país")
  xlab("Proporção de comentários") +
  ylab("Quão hierárquica é a relação")
```



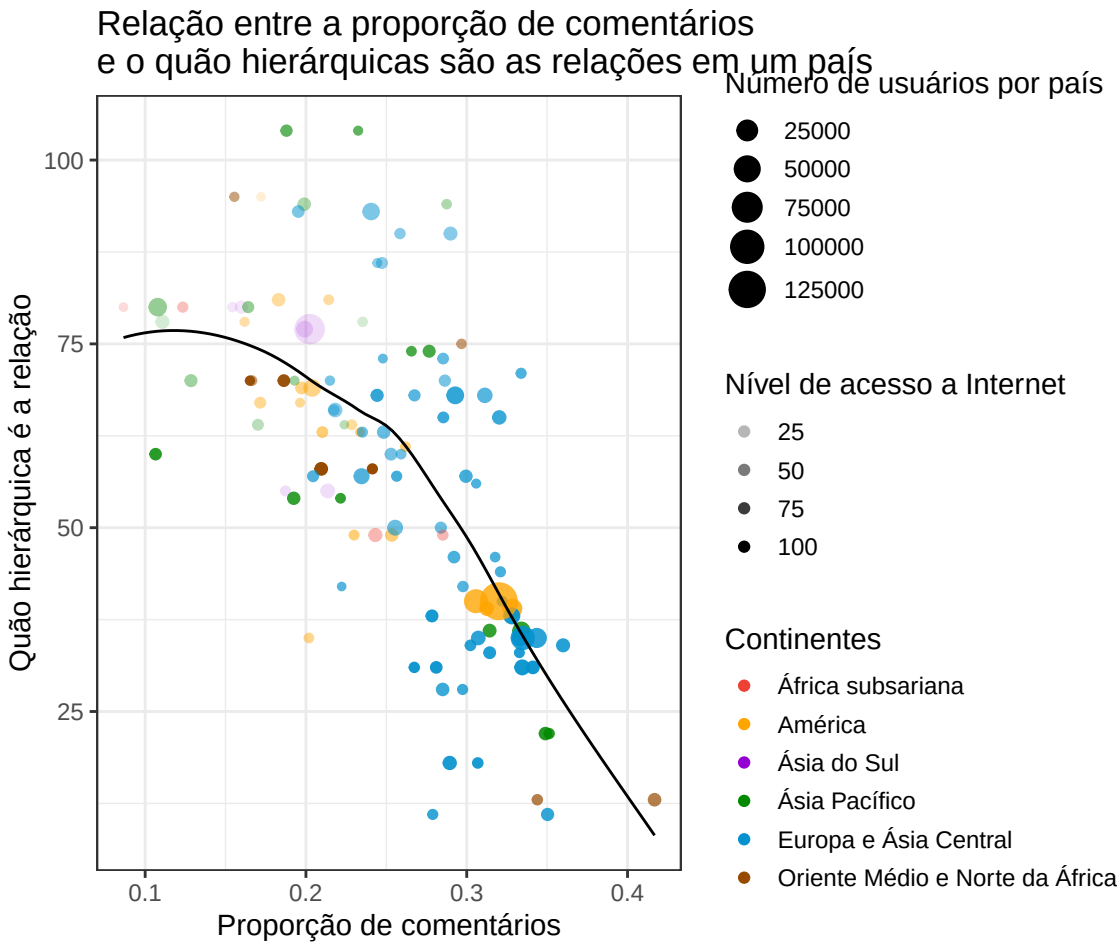
Bônus

Inclua o continente dos países (`six_regions`) na visualização.

A mesma vizualiação só que utilizando cores para representar os continentes. Como é uma variável categoria, é uma boa opção utilizar as cores como forma de apresentação. Além disso, foi escolhido uma palleta de cores que se diferenciassem bem umas das outras.

```
ggplot(data = dados) +
  geom_point(mapping = aes(x = comentaram_prop, y = PDI, alpha = Internet, size = usuarios, color = six
  geom_smooth(mapping = aes(x = comentaram_prop, y = PDI), se = FALSE, color = "black", size = .5) +
  theme(legend.key.size = unit(1,"line"))+
  scale_color_manual(values = c("#ee4035", "#ffa500", "#9400d3", "#028900", "#0392cf","#964b00"))+
  labs(
        x = "Proporção de comentários",
        y = "Quão hierárquica é a relação",
        color = "Continentes",
        alpha="Nível de acesso a Internet",
        size ="Número de usuários por país",
        title = "Relação entre a proporção de comentários \ne o quão hierárquicas são as relações em um
    )
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Outra alternativa de incluir a variável "Continente" seria agrupar os dados por continente e visualizar cada relação separadamente, para ver como cada uma se comporta por continente. Essa forma é mais informativa, pois você sabe como cada "continente" se comporta separadamente, porém os gráficos ficam menores dificultando a visualização.

```
ggplot(data = dados) +
  geom_point(mapping = aes(x = comentaram_prop, y = PDI, alpha = Internet, size = usuarios)) +
  geom_smooth(mapping = aes(x = comentaram_prop, y = PDI), color = "red", size=.5, se = FALSE) +
  theme(legend.key.size = unit(1,"line"))+
  facet_wrap(~six_regions, ncol= 2) +
  labs(
        x = "Proporção de comentário",
        y = "Valor hierarquico das relações",
        alpha="Nível de acesso a Internet",
        size ="Número de usuários por país",
        title = "Relação entre a proporção de comentários \ne o quão hierárquicas são as relações em um
    )
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : span too small. fewer data values than degrees of freedom.
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 0.085545
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 0.15761

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 0.026489

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : Chernobyl! trL>n 6

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : Chernobyl! trL>n 6

## Warning in sqrt(sum.squares/one.delta): NaNs produced
```