Iniciando a brincadeira com R

1. **Assista** [**esse vídeo**](https://www.ted.com/playlists/56/making_sense_of_too_much_data) **com atenção. Perceba que a grande pergunta que Hans Rosling tenta responder é: temos dois mundos separados; o mundo ocidental norte rico e o resto do mundo pobre? Talvez você precise pausar o vídeo para entender, assistir mais de uma vez, etc.** 
   1. **Trace um esboço do caminho que ele percorreu analisando os dados para responder essas perguntas.**
   2. **Anote os termos ou ideias que você não entende ainda.**

Ele começa demostrando que as ideias preconcebidas influenciam no julgamento dos alunos. Em seguida ele mostra (através de um gráfico de bolhas) que nos anos 60 a ideia dos alunos sobre os países “em Desenvolvimento” poderia estar certa, porém já não se aplica ao ano de 2003. Hans analisa os dados sobre a distribuição de renda no mundo, onde demostra que aos poucos as pessoas estão saindo da faixa de pobreza.

Hans mostra os dados sobre a mortalidade infantil, onde chega a concluir que um pais mais rico possui uma mortalidade infantil menor, porém é preciso investir bastante. Em seguida ele comenta que os países podem ser “quebrados” em partes menores para que se possam analisar melhor e formar estratégias melhores, como por exemplo definir uma estratégia para combater a mortalidade infantil para a população pobre e outra estratégia para a população rica.

Comenta que dados como os apresentados são muito importantes e que vários deles estão sendo coletados. Porém ainda existe um trabalho a ser feito para que tais dados fiquem totalmente disponíveis. Onde ele mostra que a China está chegando perto da USA.

1. **Qual o panorama geral dos profissionais de TI no Brasil sem ainda se preocupar com seus salários, apenas com sua localização, função, expertise...?**

A maioria dos profissionais de TI do Brasil estão na iniciativa privada e trabalhando na região Sudeste do Brasil, principalmente em Belo Horizonte. Metade dos profissionais atuantes não possuem uma Pós Graduação ou uma Certificação.

Porém podemos perceber que possuir

Comandos Usados:

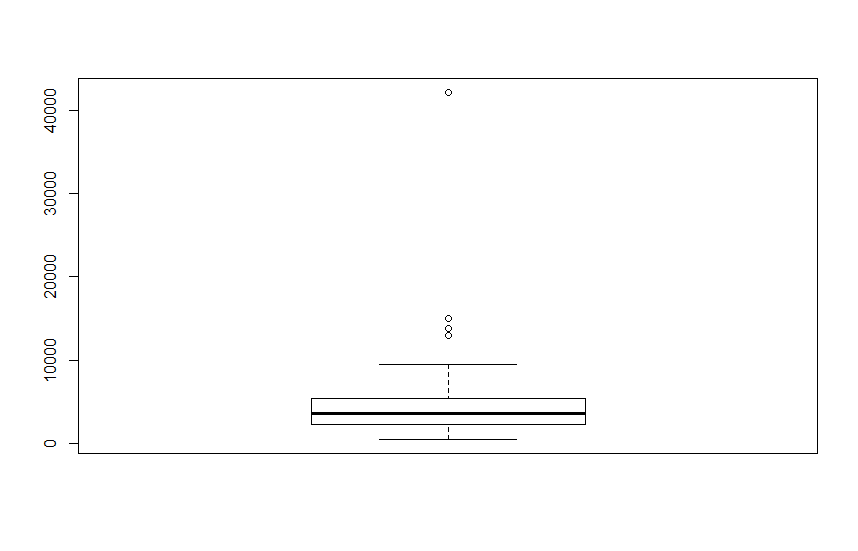
|  |
| --- |
| > mosaicplot(table(salarios.ti.regioes$Regiao))  > pie(table(salarios.ti.regioes$Iniciativa.Privada.ou.Concursado))  > mosaicplot(table(salarios.ti.regioes$Cidade))  > mosaicplot(table(salarios.ti.regioes$Pos.Graduacao.ou.Certificacao)) |

1. **Os salários estão bem distribuídos no intervalo que vai do menor salário para o maior salário? Ou será que os salários são muito concentrados em um intervalo pequeno de valores? Será que existem salários atípicos (*outliers*)?**

Como podemos observar no Box Plot dos salários, a maioria está concentrado em um intervalo pequeno entre 2281 e 4233. Também podemos observar que existem um salário atípicos de 42120.

**Summary**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** |
| 500 | 2281 | 3583 | 4233 | 5332 | 42120 |

**Box Plot - Salários**

Comandos Usados:

|  |
| --- |
| > summary(salarios.ti.regioes$Salario.Bruto)  > boxplot(salarios.ti.regioes$Salario.Bruto) |

1. **A distribuição que você encontrou é a que você esperava sem olhar os dados? Se não é, você tem alguma suspeita da razão para a diferença?**

Não, já que eu esperava um intervalo de salário mais alto. O fato de vários profissionais não terem formação acadêmica deve ter influenciado negativamente no valor dos salários.

1. **É interessante ver a média (*mean*) salarial para cada região. Como resultado, apresente uma tabela com o nome da região e a sua média salarial ordenada da região com a maior média salarial para a menor. Se você tivesse usado a mediana (*median*) em vez da média como métrica de interesse, o resultado seria o mesmo? Analise a situação e justifique. Aproveite e salve esta tabela em um arquivo do tipo csv.** Dica: use *tapply, write.csv* e *sort*.

|  |  |
| --- | --- |
| **Região** | **Média Salarial** |
| Centro-Oeste | 5620.474 |
| Nordeste | 4565.706 |
| Sudeste | 3962.227 |
| Sul | 3657.538 |
| Norte | 2552.500 |

Podemos ver abaixo como ficaria a tabela se tivesse usado a mediana.

|  |  |
| --- | --- |
| **Região** | **Mediana Salarial** |
| Centro-Oeste | 5000.00 |
| Sudeste | 3582.95 |
| Sul | 3500.00 |
| Nordeste | 2900.00 |
| Norte | 2552.500 |

Podemos observar o Nordeste “caiu” para quarto lugar. Isso se deve ao fato dos salários dos outros estados estarem mais agrupados e não terem grandes outliers como o caso do profissional que tem um salário de 42mil no Nordeste.

*Comandos Usados:*

|  |
| --- |
| > sort(tapply(salarios.ti.regioes$Salario.Bruto, salarios.ti.regioes$Regiao, mean), decreasing=TRUE)  > sort(tapply(salarios.ti.regioes$Salario.Bruto, salarios.ti.regioes$Regiao, median), decreasing=TRUE)  > mean\_salario = sort(tapply(salarios.ti.regioes$Salario.Bruto, salarios.ti.regioes$Regiao, mean), decreasing=TRUE)  > mean\_salario  > write.csv(mean\_salario, "C:\\Users\\matheus\\git\\analise-dados-1\\problem1\\data.csv") |

1. **Se um profissional de TI deseja ter um salário melhor, qual a(s) dica(s) para ele?** Dica: Use sua criatividade para brincar com os dados e descobrir informações relevantes/interessantes.

Ele possuir uma Pós Graduação ou Certificado e estudar para passar em um Concurso para trabalhar no Centro-Oeste (Brasília).

*Comandos Usados:*

|  |
| --- |
| > salario\_greater\_4000 = subset(salarios.ti.regioes, Salario.Bruto > 4000)  > sort(tapply(salario\_greater\_4000$Salario.Bruto, salario\_greater\_4000$Regiao, mean), decreasing=TRUE)  > centrooeste = subset(salarios.ti.regioes, Regiao == 'Centro-oeste')  > pie(table(centrooeste$Pos.Graduacao.ou.Certificacao))  > tapply(centrooeste$Salario.Bruto, centrooeste$Iniciativa.Privada.ou.Concursado, mean) |