Parte2 - Salário dos Profissionais de TI no Brasil

Arthur Sena

04/13/2015

Descrição dos dados:

Os dados foram coletados através de um formulário no google. Tais dados contem informações sobre o salário de algumas profissões de TI no Brasil, onde foram coletados 162 obervações com 12 variáveis cada. Na figura abaixo, é possivel visualizar uma pequena amostra dos dados.

```
salariosTI <- read.csv("salarios-ti-refinado.csv")
head(salariosTI[,c(2,3,4,6,12)])
```

```
##           Cidade UF Salario.Bruto Tempo.de.Empresa       Regiao
## 1 Campina Grande PB         42120              5.0     Nordeste
## 2       Brasilia DF         15000             12.0 Centro-oeste
## 3         Recife PE         13787              1.0     Nordeste
## 4       Brasilia DF         12960              0.5 Centro-oeste
## 5 Belo Horizonte MG          9500              3.0      Sudeste
## 6       Campinas SP          8500              1.0      Sudeste
```

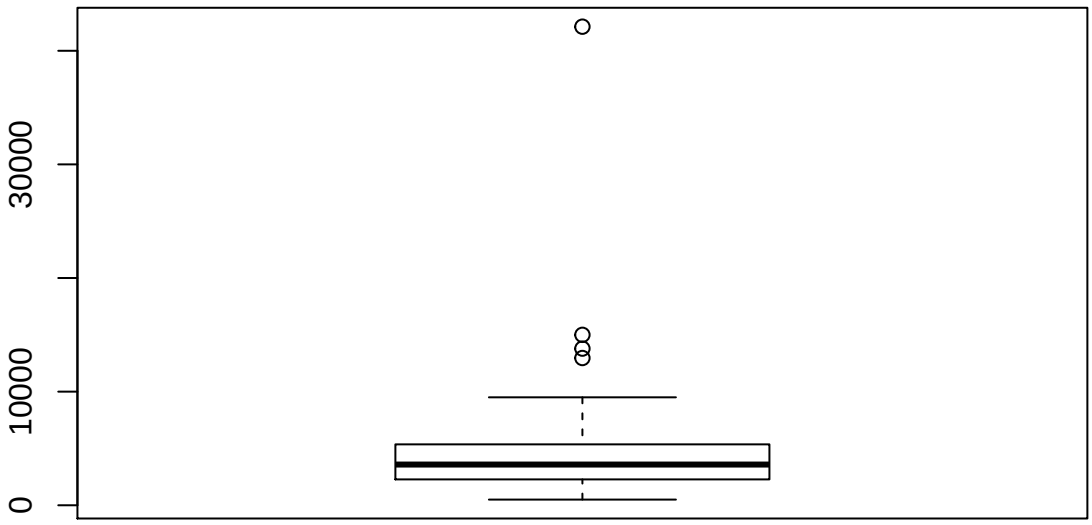
Para uma melhor visualização, algumas colunas/variáveis foram cortada da figura acima, contudo todo o conjunto de colunas/variaveis podem ser vistas abaixo:

```
colnames(salariosTI)
```

```
##  [1] "Column"                           "Cidade"
##  [3] "UF"                               "Salario.Bruto"
##  [5] "Horas.Diarias"                    "Tempo.de.Empresa"
##  [7] "Experiencia.Profissional"         "Iniciativa.Privada.ou.Concursado"
##  [9] "Cargo"                            "Formacao"
## [11] "Pos.Graduacao.ou.Certificacao"    "Regiao"
```

Antes de se fazer uma análise mais aprofundada nos nossos dados, é de bom tom procurar por uma ou mais observações estranhas ou fora do padrão. Seguindo tal conceito, é possível notar que uma da observações apresenta uma comportamento um tanto peculiar à respeito da variável Salario Bruto.

```
boxplot(salariosTI$Salario.Bruto)
```



O gráfico acima apresenta todas as observações da variável "Salario.Bruto". É possível notar que a grande maioria dos dados esta dentro de um determinado "range", contudo existe uma observação que apresenta um comportamento bastante diferenciado, onde a mesma se encontra isolada na parte superior do gráfico. Tal observacao, esta ligada a algum profissional que tem um salário bruto maior do que R$40,000 reais. Por apresentar tal comportamento, essa observacão é chamada de "outlier".

Um bom analista de dados deve ter cuidado com outlier's, pois dependo da pergunta que é feita ao seus conjunto de dados, os outlier's podem influenciar a sua resposta de uma maneira errônea. Assim sendo, é aconselhável usar estratégias que possam dimunuir ou eliminar a influência dessas observações.

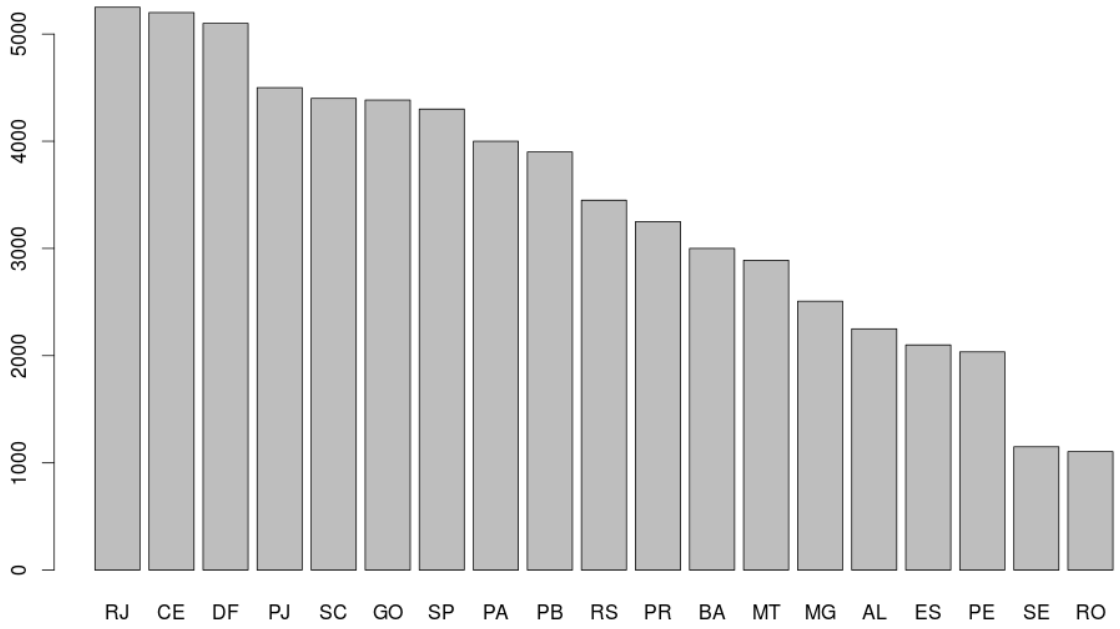
Consultando os dados

Com esses dados em mãos, nós podemos tentar responder algumas perguntas, como por exemplo:

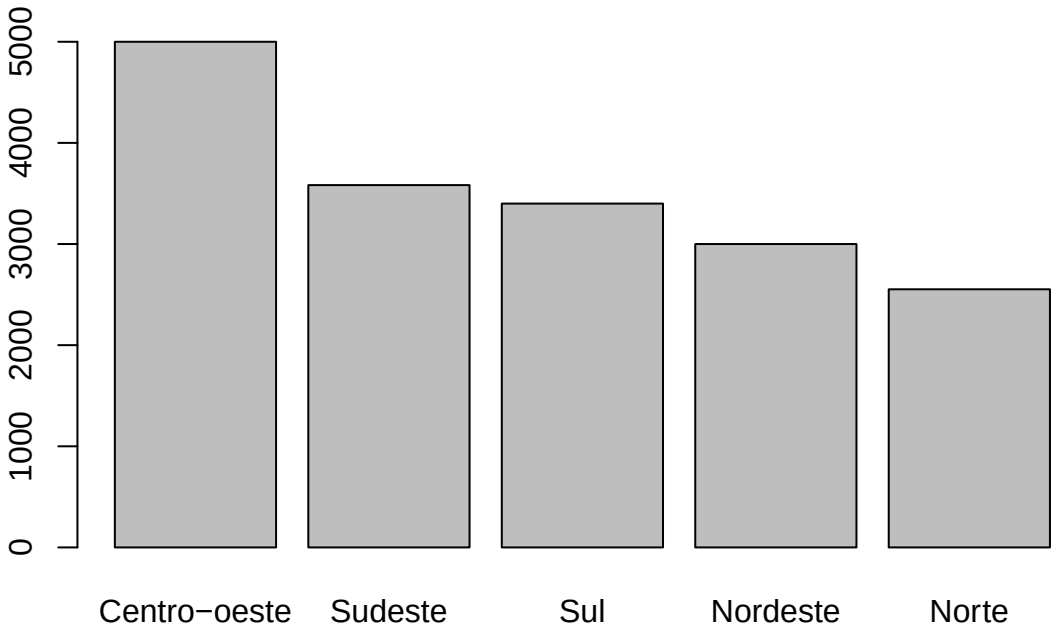
Em qual estado e em qual região estão os melhores salários para profissionais de TI?

Resposta: Para responder esta pergunta, primeiro precisamos filtrar os dados por estados e para cada estado calculamos a mediana da variável "Salario.Bruto". Eu escolhi usar a mediana, porque a média nesse caso não seria uma medida representativa para o conjunto, pois a mesma é, facilmente, afetada por outlier's.

```
#O código abaixo filtra e recupera a mediana dos salários por estados,onde por fim será gerado um gráfi
estados <- levels(salariosTI$UF)
estados_salario <- sapply(estados,function(estado){
  salario_estado <- salariosTI[salariosTI$UF == estado,]
  median(salario_estado$Salario.Bruto)
})

estados_salario <- as.data.frame(estados_salario)
estados_salario <- cbind(Estados = rownames(estados_salario), estados_salario)
colnames(estados_salario)<- c("Estados","Media_Salarial")
estados_salario<-estados_salario[order(-estados_salario$Media_Salarial),]
rownames(estados_salario) <- NULL
```




```
#O código abaixo filtra e recupera a mediana dos salários por regiões,onde por fim será gerado um gráfi

regioes <- levels(salariosTI$Regiao)
regioes_salarios <- sapply(regioes, function(regiao){
   salario_regiao <- salariosTI[salariosTI$Regiao == regiao,]
   median(salario_regiao$Salario.Bruto)
 })

regioes_salarios <- as.data.frame(regioes_salarios)
regioes_salarios <- cbind(Estados = rownames(regioes_salarios), regioes_salarios)
colnames(regioes_salarios)<- c("Regiao","Media_Salarial")
regioes_salarios<-regioes_salarios[order(-regioes_salarios$Media_Salarial),]
rownames(regioes_salarios) <- NULL
barplot(regioes_salarios$Media_Salarial,names.arg = regioes_salarios$Regiao)
```



Quão desiguais são os salários comparando quem ganha muito e pouco no Brasil como um todo? Há regiões mais desiguais?

Resposta: Para responder tal pergunta, primeiros temos que tentar definir o que seria um salário alto. Assim sendo, poderíamos dizer que os salarios dos profissionais que recebem um valor acima do terceiro quartil entrariam dentro do conjunto de salario alto. Por sua vez, os salários que se encontram abaixo do terceiro quartil entram no conjunto de salario baixo.