Trabalho A2 - Parte 1 - Yonathan Rabinovici Gherman

Base de Dados escolhida: https://www.kaggle.com/datasets/iamsouravbanerjee/software-professional-salaries-2022

Essa base de dados refere-se a profissionais da Índia, e os salário são expressos na moeda local.

Análises Unidimensionais

```
# Importando as bibliotecas
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#getwd()

# Importando a base de dados e verificando os tipos de variáveis(int, dbl, char...)
base_dados <- read.csv("salario_tech.csv")

glimpse(base_dados)
```

```
## Rows: 22,770
## Columns: 8
## $ Rating            <dbl> 3.8, 4.5, 4.0, 3.8, 4.4, 4.2, 3.7, 3.1, 3.7, 3.6, 3.~
## $ Company.Name      <chr> "Sasken", "Advanced Millennium Technologies", "Unaca~
## $ Job.Title         <chr> "Android Developer", "Android Developer", "Android D~
## $ Salary            <int> 400000, 400000, 1000000, 300000, 600000, 100000, 192~
## $ Salaries.Reported <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2~
## $ Location          <chr> "Bangalore", "Bangalore", "Bangalore", "Bangalore", ~
## $ Employment.Status <chr> "Full Time", "Full Time", "Full Time", "Full Time", ~
## $ Job.Roles         <chr> "Android", "Android", "Android", "Android", "Android~
```

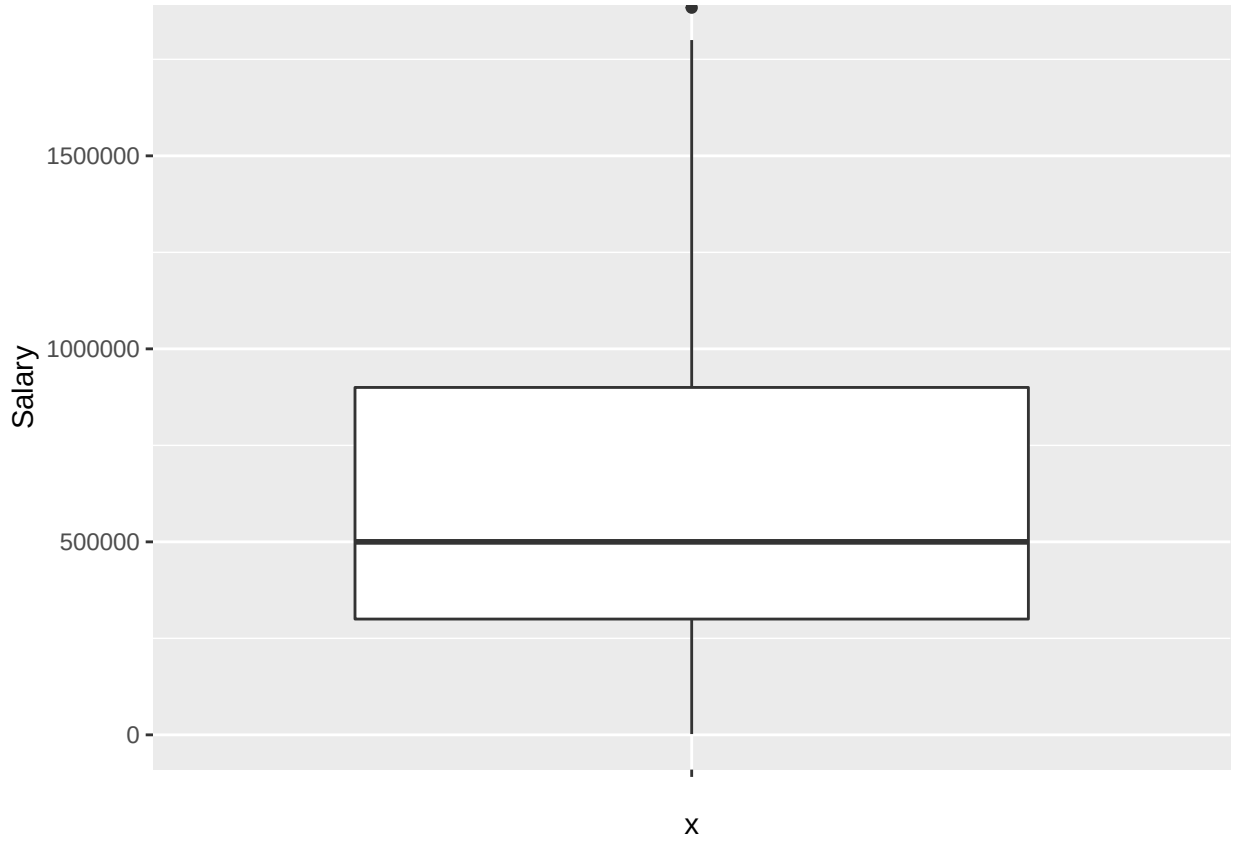
```
# Garantindo que todas as palvras fiquem em maiúsculo para não haver diferenciação de um mesmo dado
base_dados <- base_dados %>% mutate_if(is.character, toupper)
```

Salário

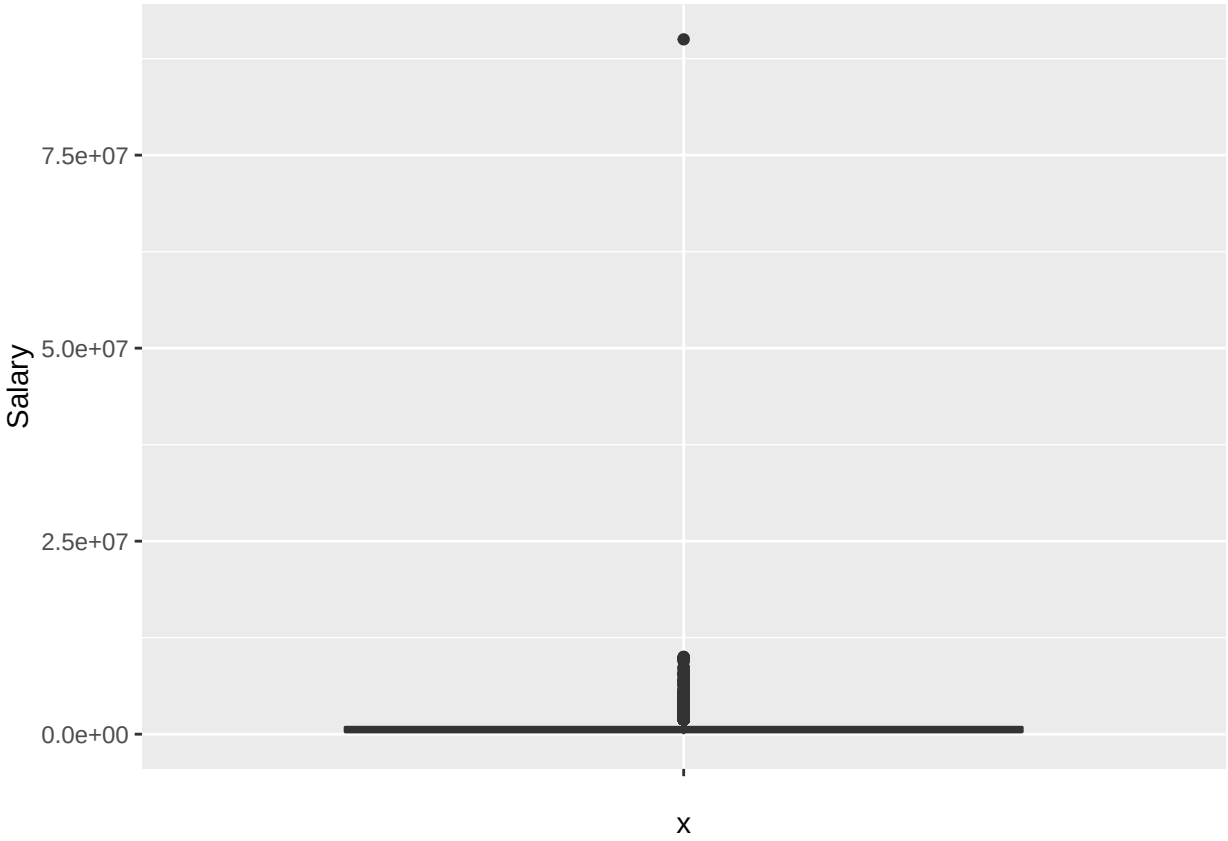
```r
# Salário
# Análise da média, mediana, quartil, mínimo e máximo dos salários na moeda Indian Rupee -
summary(base_dados$Salary)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     2112   300000   500000   695387   900000 90000000
```

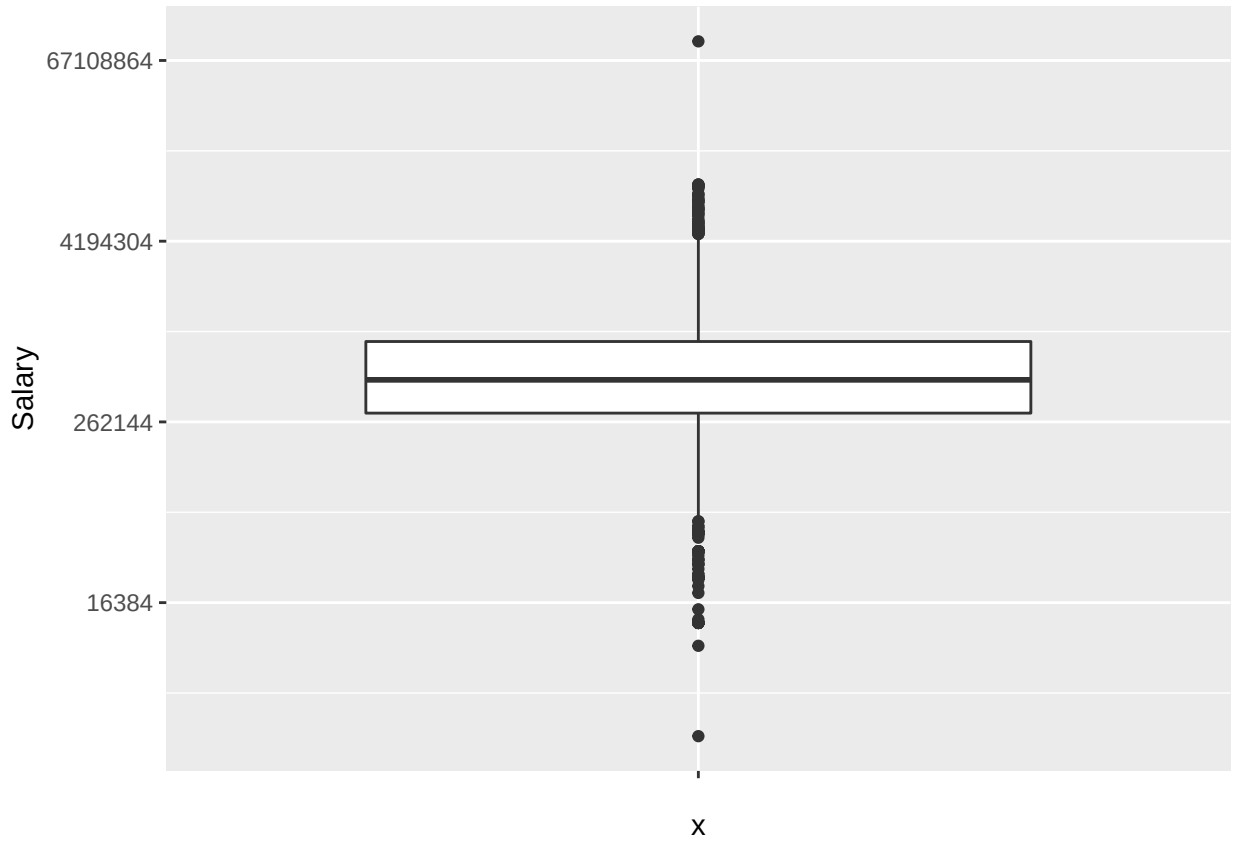
```r
# Criação do box plot de salários limitado a 1800000 para que os outliers(que estão presentes em grande

ggplot(base_dados, aes(y = Salary, x="")) + geom_boxplot() + coord_cartesian(ylim=c(0,1800000))
```



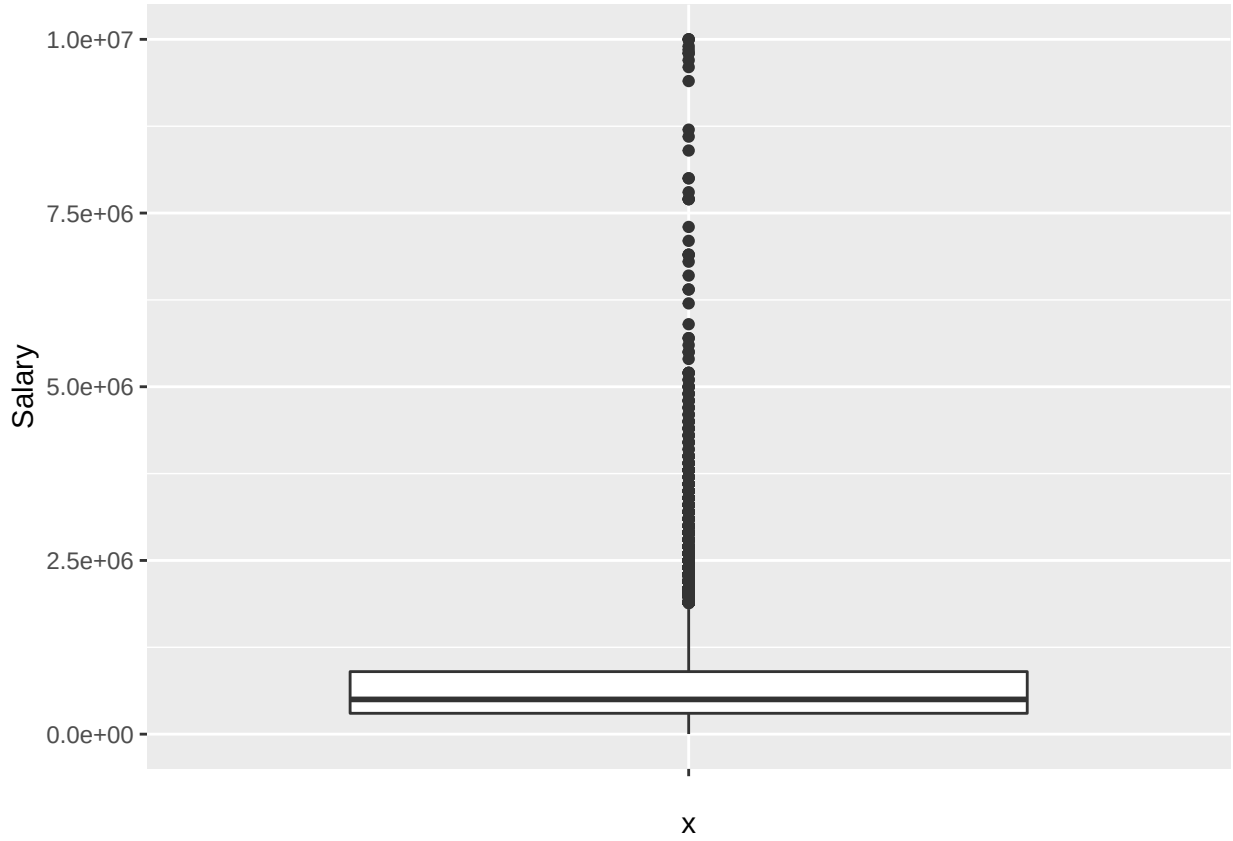

```r
# Caso não houvesse limitação ele seria visto assim:
ggplot(base_dados, aes(y = Salary, x="")) + geom_boxplot()
```




```
# Porém, ao colocá-lo em escala logaítmica é possívelo vê-lo sem ter de limitar
ggplot(base_dados, aes(y = Salary, x="")) + geom_boxplot() + scale_y_continuous(trans = 'log2')
```



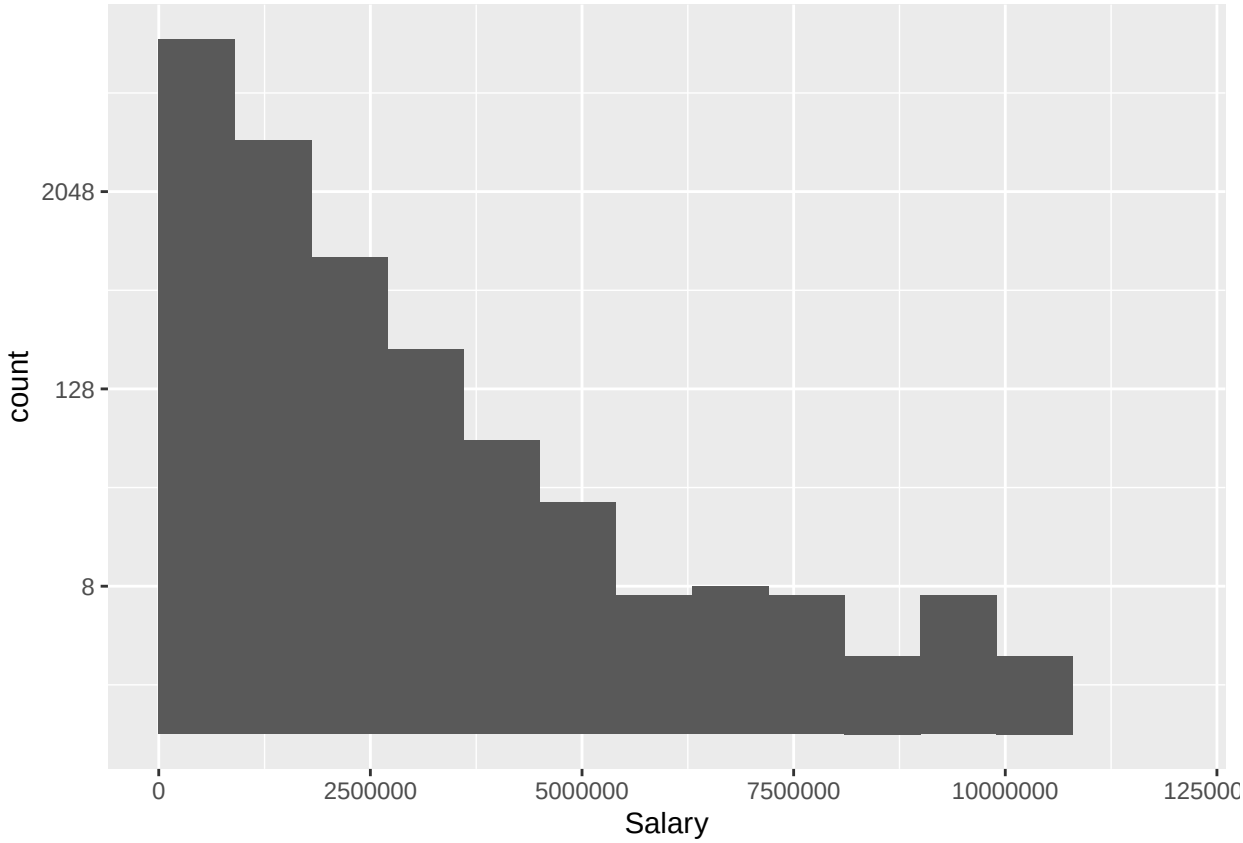

```
# Grande parte dos outliers concentram-se nesse limite

ggplot(base_dados, aes(y = Salary, x="")) + geom_boxplot() + coord_cartesian(ylim=c(0,10000000))
```




```
# Histograma dos salários
ggplot(base_dados, aes(Salary)) + geom_histogram(binwidth = 900000, boundary = 900000) + coord_cartesian
  scale_y_continuous(trans = 'log2')
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 87 rows containing missing values (geom_bar).
```

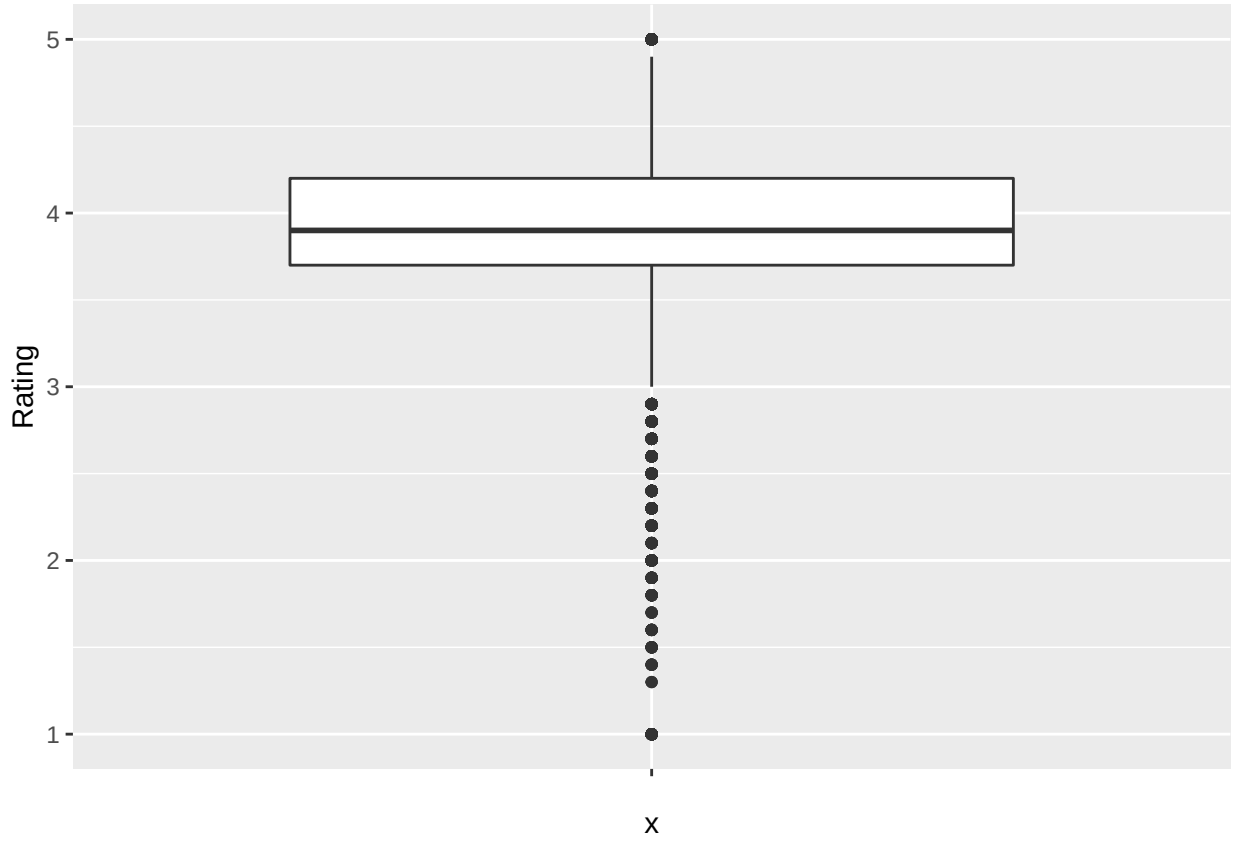


Avaliação da Empresa

```
# Avaliação da Empresa
# Análise da média, mediana, quartil, mínimo e máximo das avaliações dadas as empresas
summary(base_dados$Rating)
```

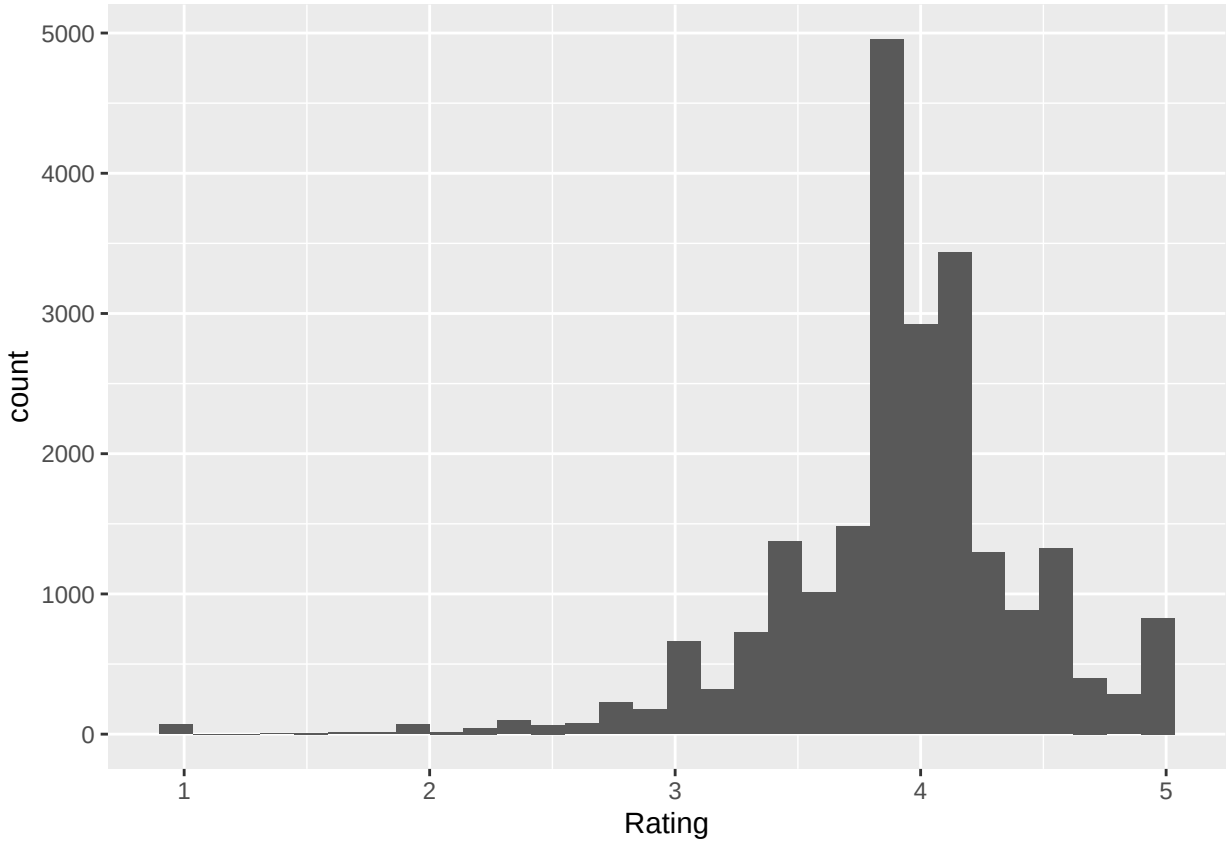
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.700   3.900   3.918   4.200   5.000
```

```
# Box plot das avaliações
ggplot(base_dados, aes(y=Rating, x="")) + geom_boxplot()
```




```
# Histograma das avaliações
ggplot(base_dados, aes(Rating)) + geom_histogram()
```

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

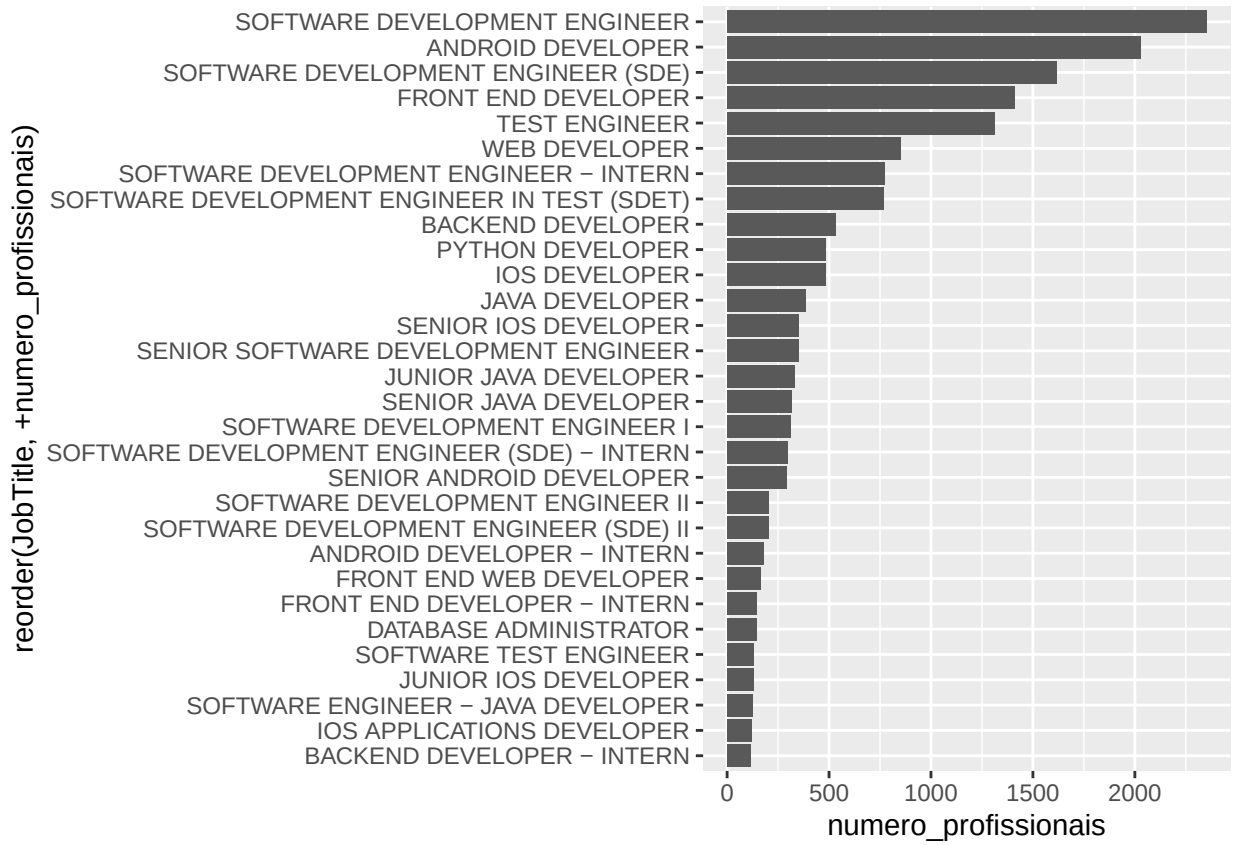


TOP 30 Profissões com mais profissionais

```
contagem <- base_dados %>% count(Job.Title)
colnames(contagem)[1] <- "JobTitle"
p <-contagem %>% arrange(desc(n))
p <- p %>% top_n(n=30)
```

```
## Selecting by n
```

```
colnames(p)[2] <- "numero_profissionais"

p <- ggplot(data = p, aes(x=reorder(JobTitle, +numero_profissionais), y=numero_profissionais)) + geom_ba
p +coord_flip()
```

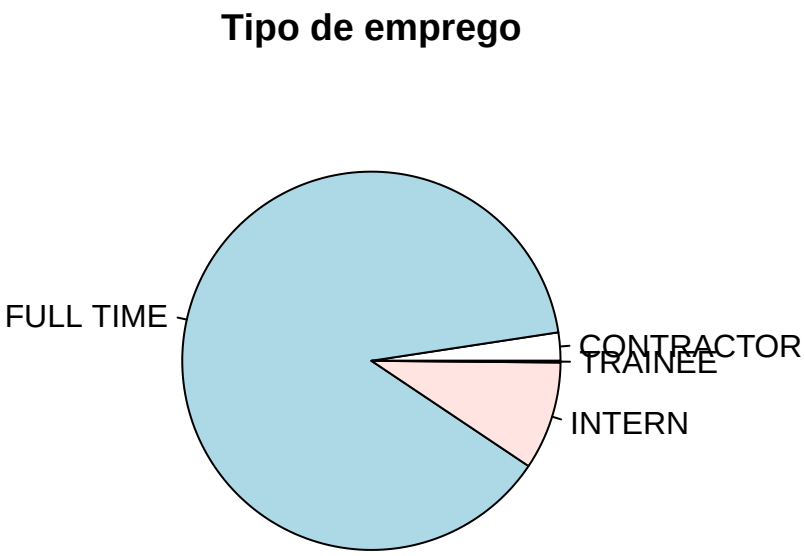


Frequência dos tipos de contrato/Status de emprego

```
#Status de Emprego
p <- base_dados %>% count(Employment.Status)
tibble(p)
```

```
## # A tibble: 4 x 2
##   Employment.Status     n
##   <chr>             <int>
## 1 CONTRACTOR          548
## 2 FULL TIME         20083
## 3 INTERN             2106
## 4 TRAINEE              33
```

```
pie(x=p$n, labels=p$Employment.Status, main = "Tipo de emprego")
```



Análise Bidimensional

Quantidade de pessoas em uma área x média de salários nessa área

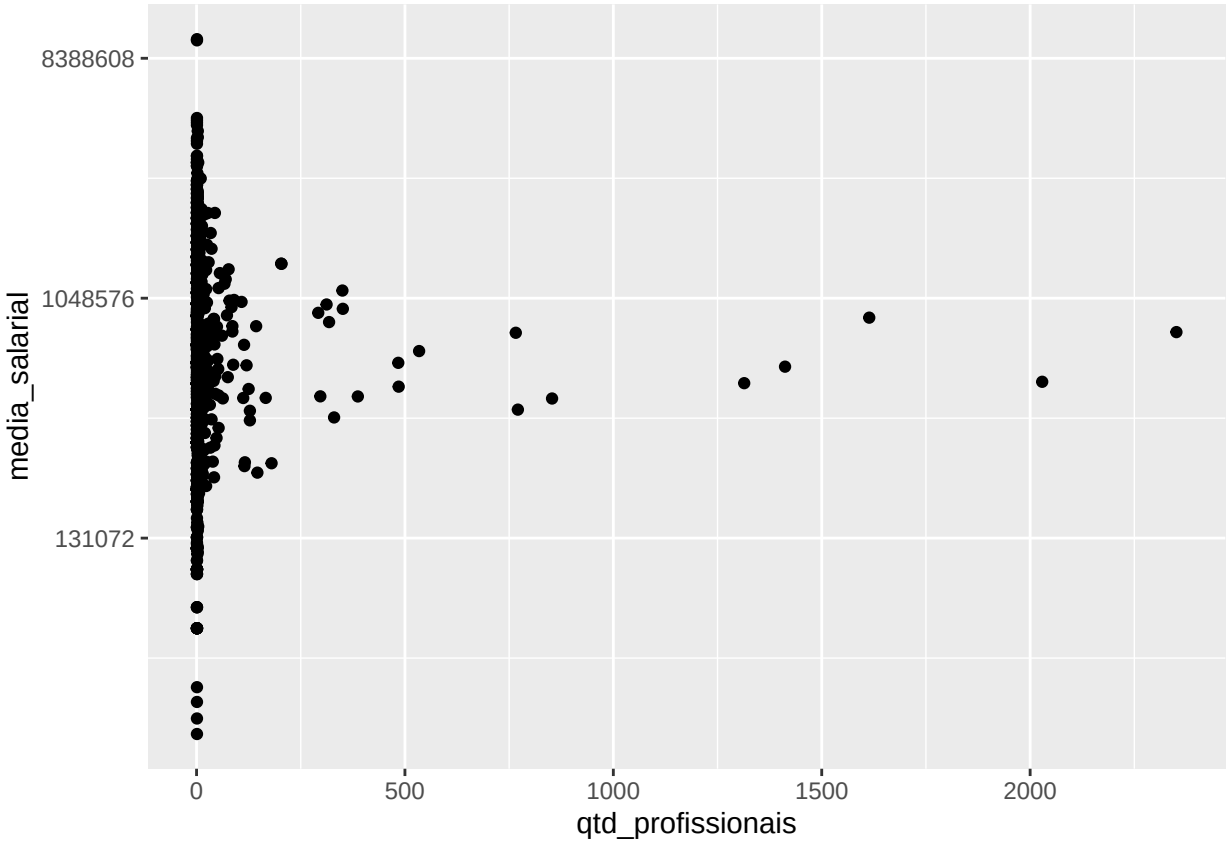
```
# PARTE 2 - Análise Bidimensional
#Análise entre Média Salarial e o nome da posição ocupada
p <- base_dados
base <- as_tibble(p)
p <- base %>% count(Job.Title)
sal <- base %>% group_by(Job.Title) %>% summarise(media=mean(Salary))
dado <- bind_cols(sal,p)
```

```
## New names:
## * `Job.Title` -> `Job.Title...1`
## * `Job.Title` -> `Job.Title...3`
```

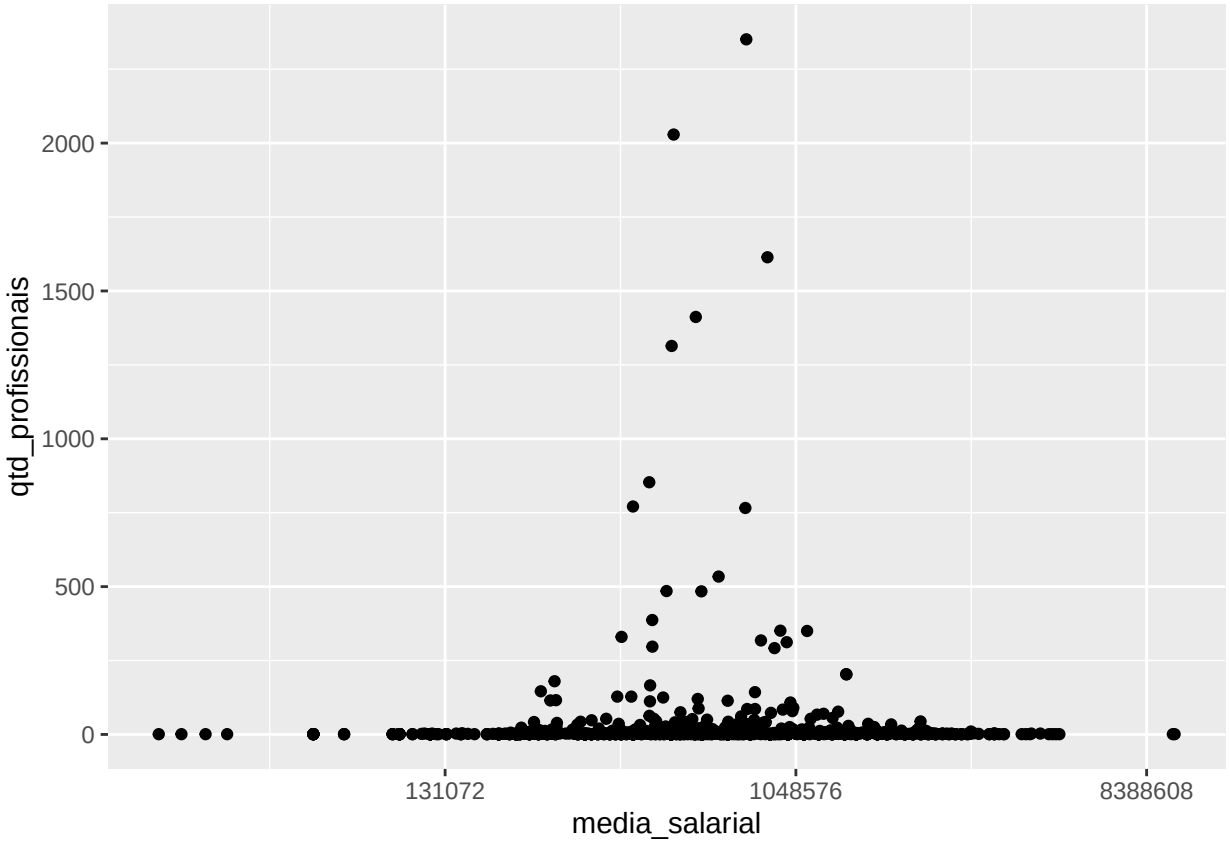
```
dado <- dado[,-c(3)]
dado <- dado %>% arrange(desc(n))
colnames(dado)[2] <- "media_salarial"
colnames(dado)[3] <- "qtd_profissionais"
dado
```

```
## # A tibble: 1,080 x 3
##    Job.Title...1                                media_salarial qtd_profissionais
##    <chr>                                                 <dbl>             <int>
##  1 SOFTWARE DEVELOPMENT ENGINEER                       781570.              2351
##  2 ANDROID DEVELOPER                                   508106.              2029
##  3 SOFTWARE DEVELOPMENT ENGINEER (SDE)                 885917.              1614
##  4 FRONT END DEVELOPER                                 579359.              1412
##  5 TEST ENGINEER                                       502066.              1314
##  6 WEB DEVELOPER                                       439625.               853
##  7 SOFTWARE DEVELOPMENT ENGINEER - INTERN              399219.               771
##  8 SOFTWARE DEVELOPMENT ENGINEER IN TEST (SDET)        777185.               766
##  9 BACKEND DEVELOPER                                   663234.               534
## 10 PYTHON DEVELOPER                                    487069.               485
## # ... with 1,070 more rows
```

```
ggplot(dado, aes(x=qtd_profissionais, y=media_salarial)) + geom_point() +  scale_y_continuous(trans = '
```

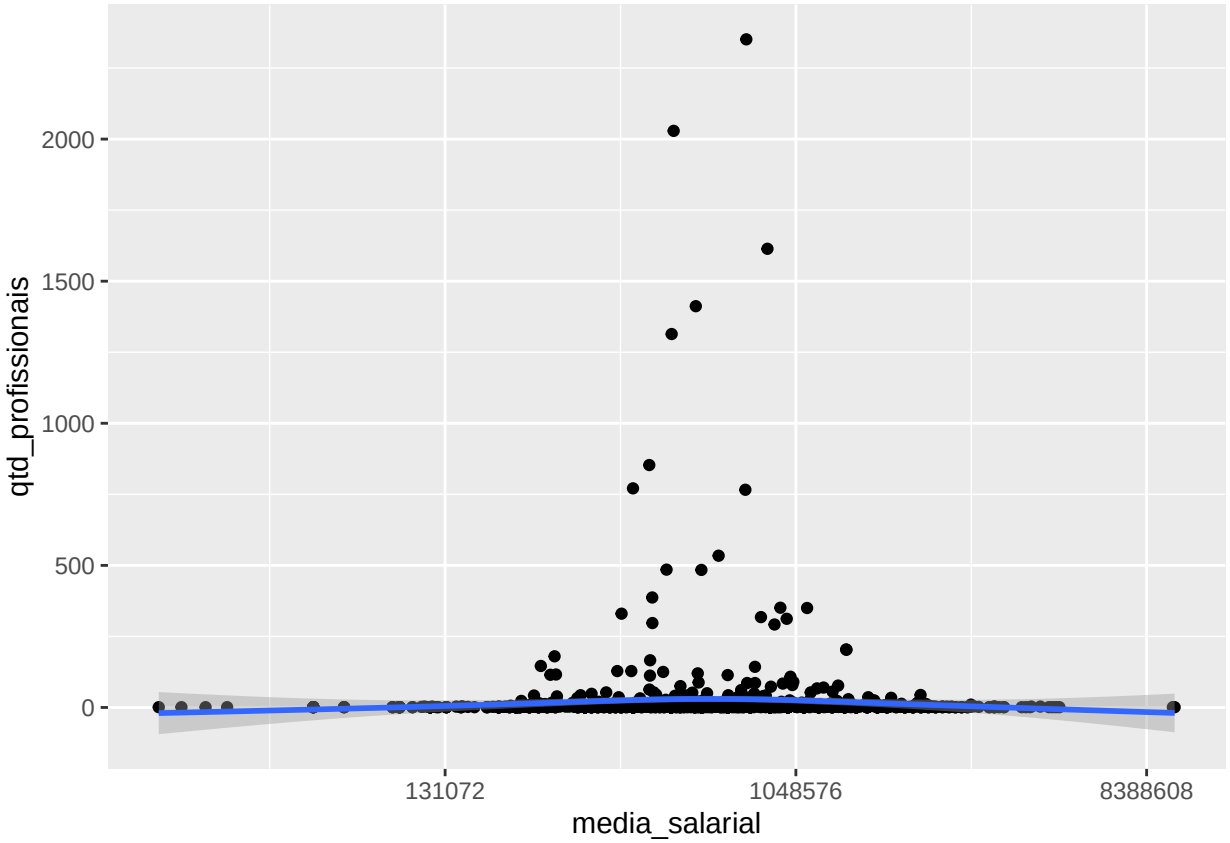



```
ggplot(dado, aes(x=media_salarial, y=qtd_profissionais)) + geom_point() +  scale_x_continuous(trans = '
```

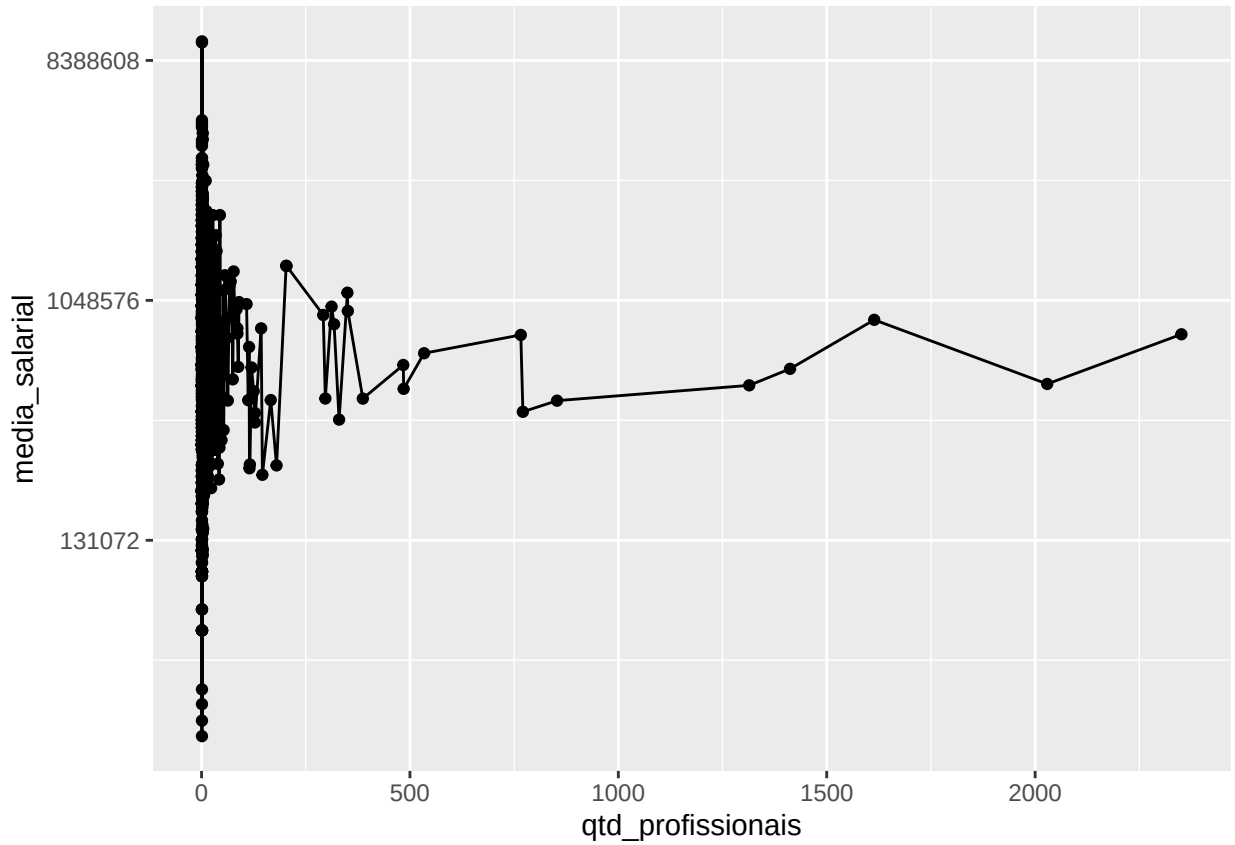



```
ggplot(dado, aes(x=media_salarial, y=qtd_profissionais)) + geom_point() +  scale_x_continuous(trans = '
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
ggplot(dado, aes(x=qtd_profissionais, y=media_salarial)) + geom_line() + geom_point() +  scale_y_continu
```




```
# Teste de Correlação

cor.test(x= dado$qtd_profissionais, y= dado$media_salarial, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  dado$qtd_profissionais and dado$media_salarial
## t = -0.98447, df = 1078, p-value = 0.3251
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08946282  0.02973420
## sample estimates:
##         cor
## -0.02997086
```

```
#cor = 1: Correlação perfeita positiva entre as duas variáveis.
#cor = 0: Não existe correlação linear, a correlação deve ser investigada por outros métodos.
#cor = -1: Correlação perfeita negativa entre as duas variáveis.
```

Tipo de contrato/status de emprego x média salarial

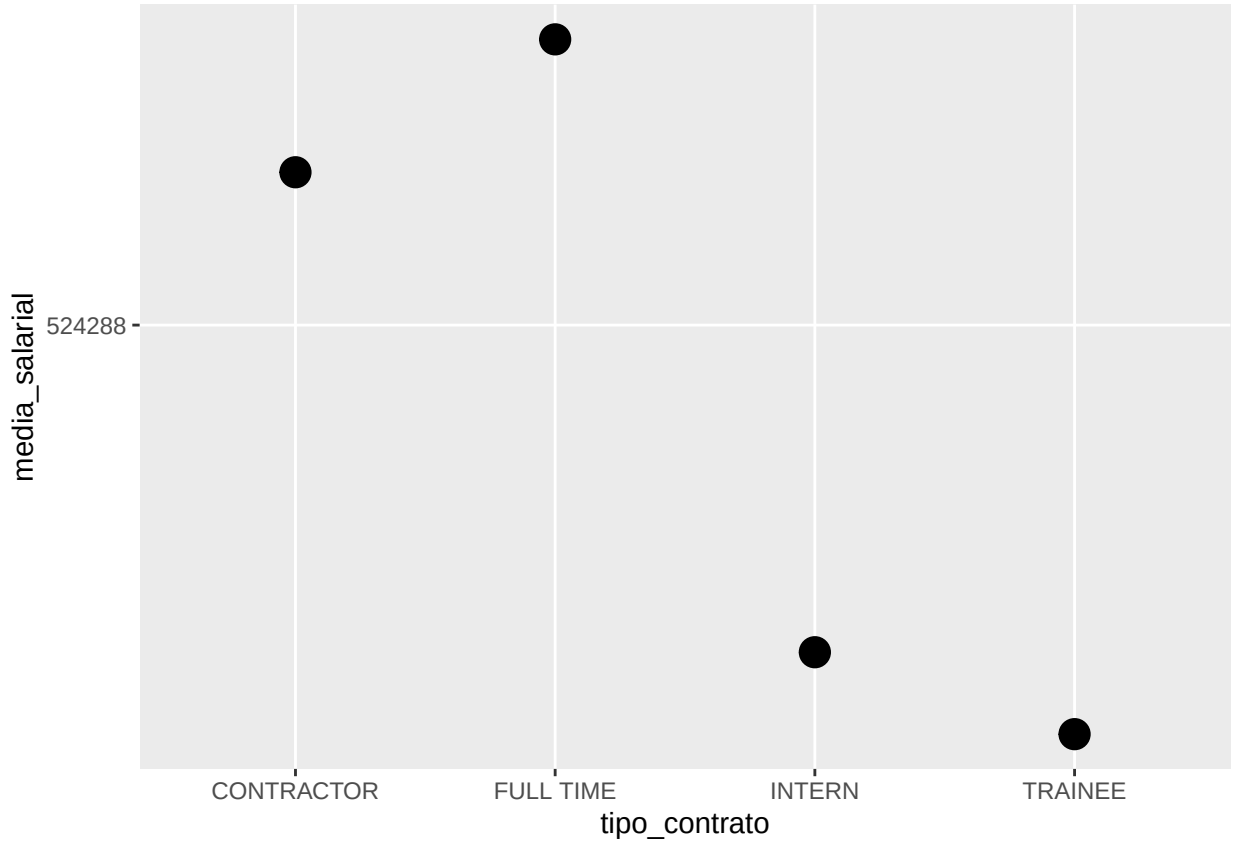
```r
p <- base_dados
base <- as_tibble(p)
p <- base %>% count(Employment.Status)
sal <- base %>% group_by(Employment.Status) %>% summarise(media=mean(Salary))
dado <- bind_cols(sal,p)
```

```
## New names:
## * `Employment.Status` -> `Employment.Status...1`
## * `Employment.Status` -> `Employment.Status...3`
```

```r
dado <- dado[,-c(3)]
colnames(dado)[1] <- "tipo_contrato"
colnames(dado)[2] <- "media_salarial"
colnames(dado)[3] <- "qtd_profissionais"
dado <- dado %>% arrange(desc(media_salarial))
dado
```

```
## # A tibble: 4 x 3
##   tipo_contrato media_salarial qtd_profissionais
##   <chr>                  <dbl>             <int>
## 1 FULL TIME            733332.             20083
## 2 CONTRACTOR           627362.               548
## 3 INTERN               357054.              2106
## 4 TRAINEE              324303.                33
```

```r
ggplot(dado, aes(y=media_salarial, x=tipo_contrato)) + geom_point(size=5) +  scale_y_continuous(trans =
```



Porcentagem dos tipos de contrato nas 5 cidades com mais profissionais(LEMBRE-SE: LEVE EM CONTA A UNIDADE DO EIXO X)

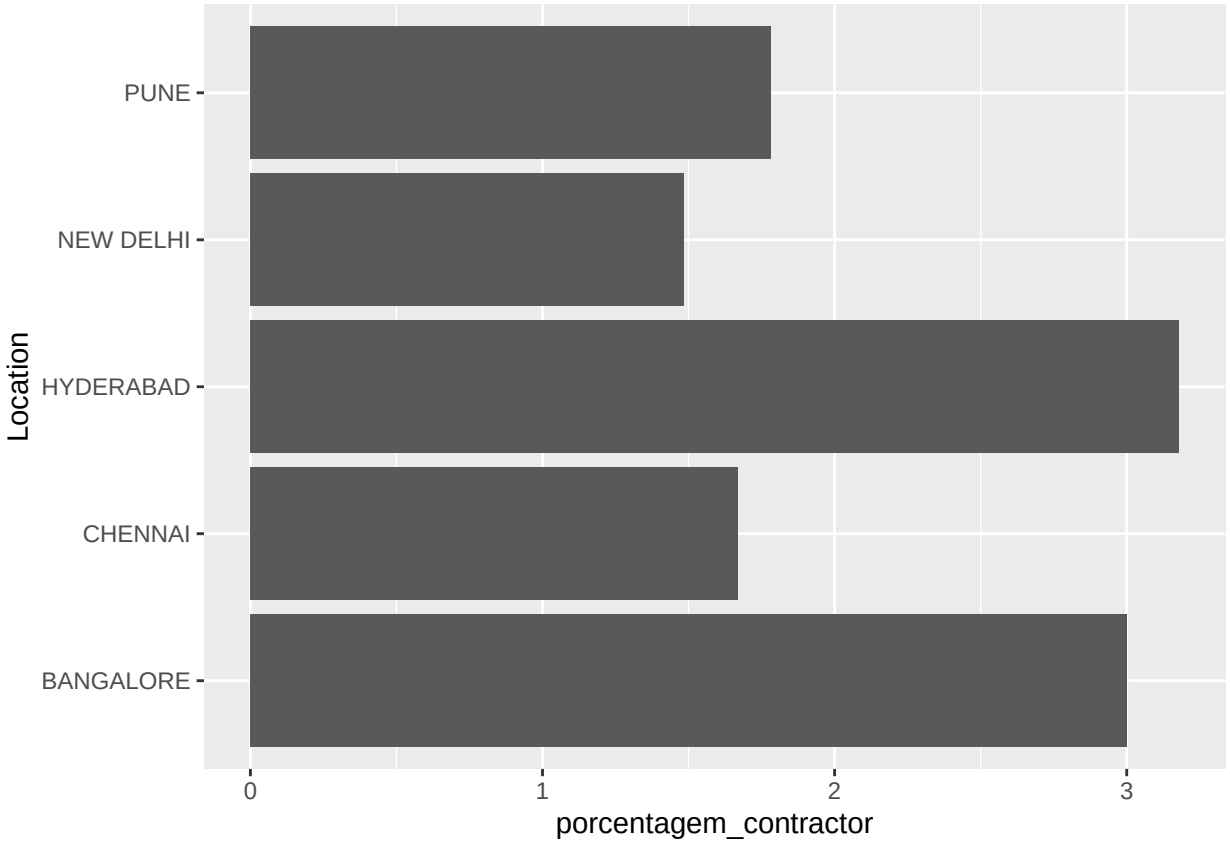
```
# Porcentagem dos tipos de contrato nas 5 cidades com mais profissionais
p <- base_dados
base <- as_tibble(p)

# BANGALORE
status_contractor_bangalore <- (nrow(base %>% filter(str_detect(Location,"BANGALORE")) %>%
                                     filter(str_detect(Employment.Status,"CONTRACTOR")))/nrow(base %>%
                                                                                                filte
status_full_bangalore <- (nrow(base %>% filter(str_detect(Location,"BANGALORE")) %>%
                                 filter(str_detect(Employment.Status,"FULL TIME")))/nrow(base %>%
                                                                                        filter(str_
status_intern_bangalore <- (nrow(base %>% filter(str_detect(Location,"BANGALORE")) %>%
                                 filter(str_detect(Employment.Status,"INTERN")))/nrow(base %>%
                                                                                       filter(str_
status_trainee_bangalore <- (nrow(base %>% filter(str_detect(Location,"BANGALORE")) %>%
                                    filter(str_detect(Employment.Status,"TRAINEE")))/nrow(base %>%
                                                                                          filter(str_d


#CHENNAI

status_contractor_chennai <- (nrow(base %>% filter(str_detect(Location,"CHENNAI")) %>%
```

```
                                  filter(str_detect(Employment.Status,"CONTRACTOR")))/nrow(base %>%
                                                                                    filter(str_d

status_full_chennai <- (nrow(base %>% filter(str_detect(Location,"CHENNAI")) %>%
                                    filter(str_detect(Employment.Status,"FULL TIME")))/nrow(base %>%
                                                                                          filter
status_intern_chennai <- (nrow(base %>% filter(str_detect(Location,"CHENNAI")) %>%
                                    filter(str_detect(Employment.Status,"INTERN")))/nrow(base %>%
                                                                                          filter

status_trainee_chennai <- (nrow(base %>% filter(str_detect(Location,"CHENNAI")) %>%
                                    filter(str_detect(Employment.Status,"TRAINEE")))/nrow(base %>%
                                                                                          filter
#HYDERABAD

status_contractor_hyderabad <- (nrow(base %>% filter(str_detect(Location,"HYDERABAD")) %>%
                                    filter(str_detect(Employment.Status,"CONTRACTOR")))/nrow(base %>%
                                                                                          filter


status_intern_hyderabad <- (nrow(base %>% filter(str_detect(Location,"HYDERABAD")) %>%
                                     filter(str_detect(Employment.Status,"INTERN")))/nrow(base %>%
                                                                                           filt


status_full_hyderabad <- (nrow(base %>% filter(str_detect(Location,"HYDERABAD")) %>%
                                     filter(str_detect(Employment.Status,"FULL TIME")))/nrow(base %>
                                                                                           filt


status_trainee_hyderabad <- (nrow(base %>% filter(str_detect(Location,"HYDERABAD")) %>%
                                     filter(str_detect(Employment.Status,"TRAINEE")))/nrow(base %>%
                                                                                           filt

#NEW DELHI
status_contractor_new <- (nrow(base %>% filter(str_detect(Location,"NEW DELHI")) %>%
                                     filter(str_detect(Employment.Status,"CONTRACTOR")))/nrow(base %
                                                                                           filt

status_intern_new <- (nrow(base %>% filter(str_detect(Location,"NEW DELHI")) %>%
                                   filter(str_detect(Employment.Status,"INTERN")))/nrow(base %>%
                                                                                         filter(

status_full_new <- (nrow(base %>% filter(str_detect(Location,"NEW DELHI")) %>%
                                filter(str_detect(Employment.Status,"FULL TIME")))/nrow(base %>%
                                                                                      filter(str_dete
status_trainee_new <- (nrow(base %>% filter(str_detect(Location,"NEW DELHI")) %>%
                                filter(str_detect(Employment.Status,"TRAINEE")))/nrow(base %>%
                                                                                      filter(str_dete

#PUNE
status_contractor_pune <- (nrow(base %>% filter(str_detect(Location,"PUNE")) %>%
                                 filter(str_detect(Employment.Status,"CONTRACTOR")))/nrow(base %>%
```

```
                                                                                filter(str_
status_full_pune <- (nrow(base %>% filter(str_detect(Location,"PUNE")) %>%
                                filter(str_detect(Employment.Status,"FULL TIME")))/nrow(base %>%
                                                                                  filter(st
status_intern_pune <- (nrow(base %>% filter(str_detect(Location,"PUNE")) %>%
                                filter(str_detect(Employment.Status,"INTERN")))/nrow(base %>%
                                                                                  filter(st
status_trainee_pune <- (nrow(base %>% filter(str_detect(Location,"PUNE")) %>%
                                filter(str_detect(Employment.Status,"TRAINEE")))/nrow(base %>%
                                                                                  filter(st
```

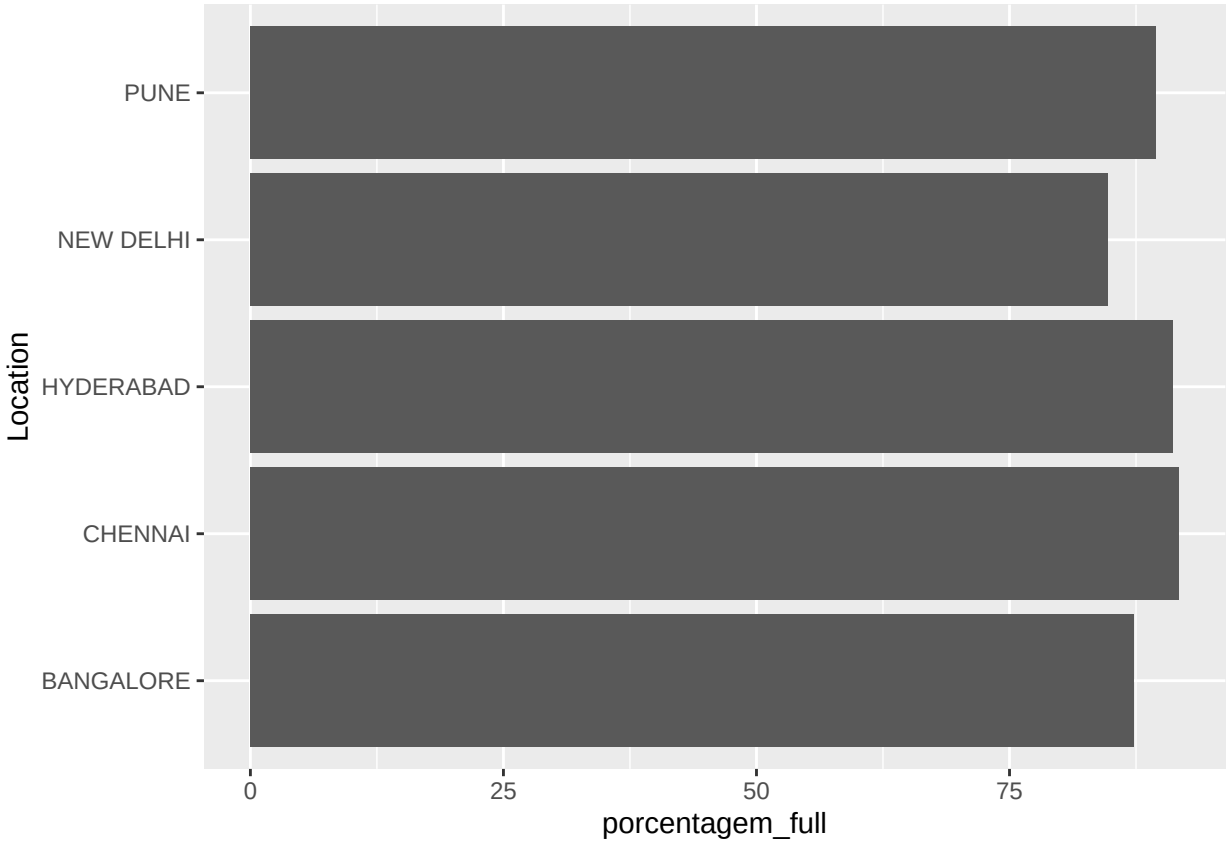
Porcentagem de CONTRACTOR nas 5 cidades com mais profissionais

```
#CONTRACTOR
p <- base_dados
base <- as_tibble(p)
p <- base %>% count(Location)
p <- slice(p,  c(1,2,3,9,10))
p_contractor <- p[,-c(2)]
arr <- c(status_contractor_bangalore, status_contractor_chennai, status_contractor_hyderabad, status_co
df_contractor <-data.frame(arr)
df_contractor <- bind_cols(p_contractor,df_contractor)
colnames(df_contractor)[2] <- "porcentagem_contractor"
ggplot(df_contractor, aes(Location, porcentagem_contractor)) + geom_bar(stat="identity") + coord_flip()
```



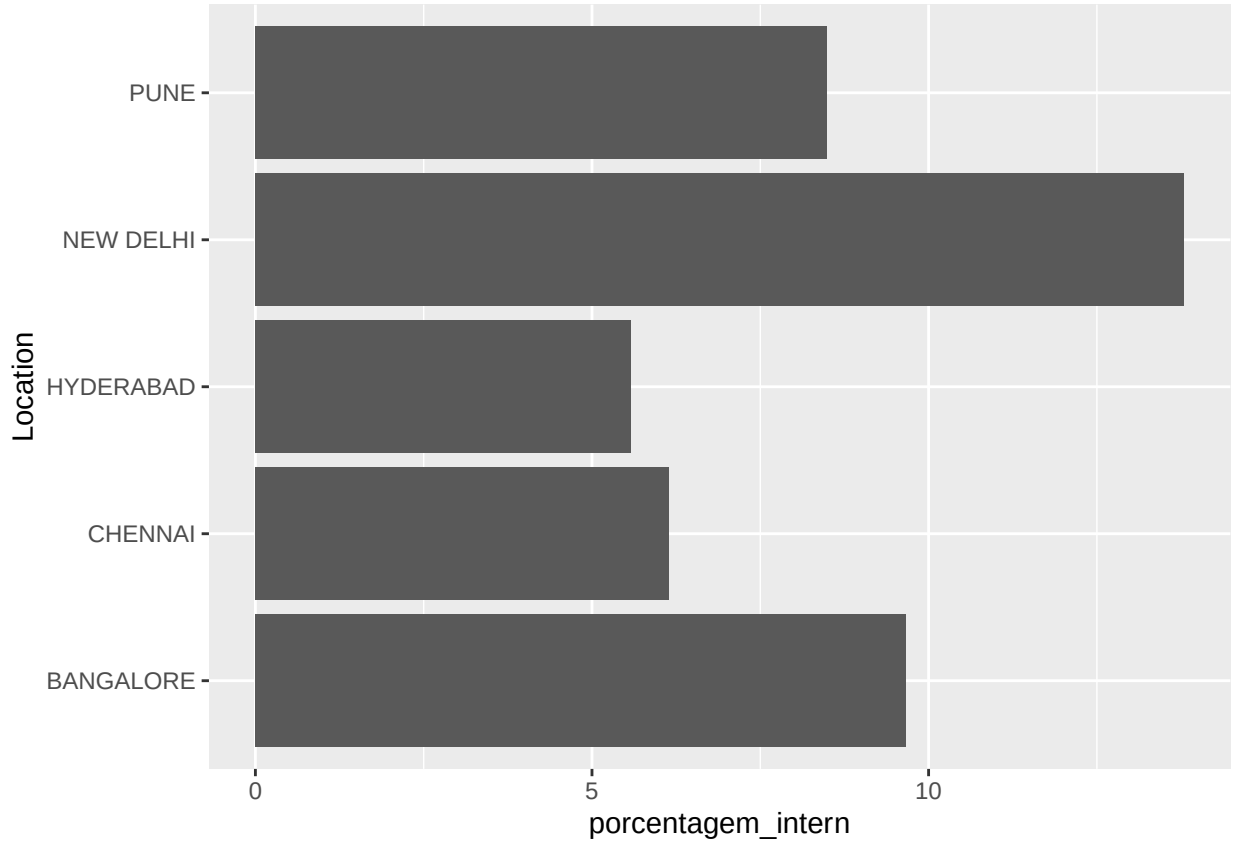
Porcentagem de FULL TIME nas 5 cidades com mais profissionais

```
#FULL TIME
arr <- c(status_full_bangalore, status_full_chennai, status_full_hyderabad, status_full_new, status_full
df_full <-data.frame(arr)
df_full <- bind_cols(p_contractor,df_full)
colnames(df_full)[2] <- "porcentagem_full"
ggplot(df_full, aes(Location, porcentagem_full)) + geom_bar(stat="identity") + coord_flip()
```



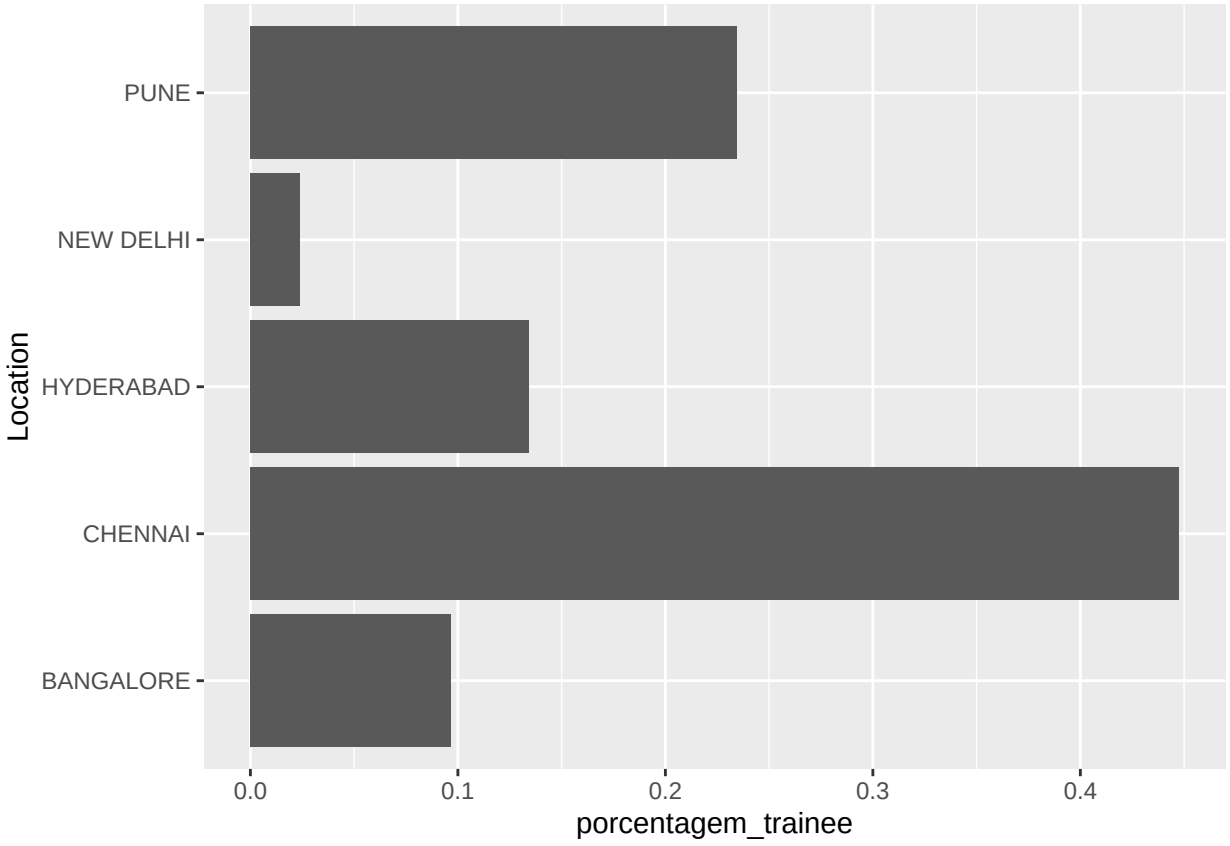
Porcentagem de INTERN nas 5 cidades com mais profissionais

```
# INTERN
arr <- c(status_intern_bangalore, status_intern_chennai, status_intern_hyderabad, status_intern_new, sta
df_intern <-data.frame(arr)
df_intern <- bind_cols(p_contractor,df_intern)
colnames(df_intern)[2] <- "porcentagem_intern"
ggplot(df_intern, aes(Location, porcentagem_intern)) + geom_bar(stat="identity") + coord_flip()
```



Porcentagem de TRAINEE nas 5 cidades com mais profissionais

```
# TRAINEE
arr <- c(status_trainee_bangalore, status_trainee_chennai, status_trainee_hyderabad, status_trainee_new
df_trainee <-data.frame(arr)
df_trainee <- bind_cols(p_contractor,df_trainee)
colnames(df_trainee)[2] <- "porcentagem_trainee"
ggplot(df_trainee, aes(Location, porcentagem_trainee)) + geom_bar(stat="identity") + coord_flip()
```



Observa-se constância nas empresas (que poderia ter sido observada no gráfico de pizza das análises unidimensionais, porém aqui é demonstrado de forma mais detalhada): elas possuem mais contratos Full Time, em seguida Intern, Contractor e Trainee

A partir dessa análise é possível chegar em conlusões como: cidades que dão mais chance de pessoas iniciarem como trainee, por exemplo, Chennai. Mas, é importante considerar outras questões, como espaço amostral daquela cidade, entre outros.