Entrega 1 - Estatística Inferencial

Objetivo do Projeto

O projeto de Ciência de Dados do grupo consiste em analisar a diversidade e as tendências do mercado de TI, ligado à área de Dados no Brasil, focando em como as vagas de emprego estão sendo distribuídas entre diferentes grupos, como por exemplo, gênero, raça, orientação sexual e deficiência. Além disso, queremos filtrar quais tecnologias estão sendo mais demandadas e quais níveis de senioridade estão sendo mais buscados.
O objetivo geral do nosso projeto é gerar análises eficientes para entender como as empresas filtram candidatos, gerando assim maior entendimento na hora dos processos seletivos. Além disso, queremos saber como esses profissionais desempenham seu trabalho, seja por senioridade, horas trabalhadas e entre outros aspectos.
Nas entregas da matéria de Análise Inferencial de Dados, resolvemos filtrar nossa pesquisa em faixa salarial e idade dos participantes.

Cálculo média

Calculando a média etária dos participantes da pesquisa, podemos perceber que na área da tecnologia, especificamente na área da dados, temos uma média de pessoas com aproximadamente 32 anos. Com isso, podemos concluir que provavelmente existe um mercado relativamente jovem, já que é uma área que cresceu consideravelmente nas últimas décadas.

```
> # Calcular a média da coluna "Idade"
> media_idade <- mean(dados_diversidade_2$`Idade`, na.rm = TRUE)
>
>
> # Mostrar os resultados
> print(media_idade)
[1] 31.99717
>
```

Calculando moda

Definimos a moda observando nos dados o número que mais se repete. Como já analisado acima, o mercado jovem é uma possível conclusão para o nosso resultado, sendo a idade mais presente, 27 anos.

```
> # Calcular a moda da coluna "Idade"
> moda_idade <- calcular_moda(dados_diversidade_2$`Idade`)
>
>
> # Mostrar os resultados
> print(moda_idade)
[1] 27
>
```

Calculando variância e desvio padrão

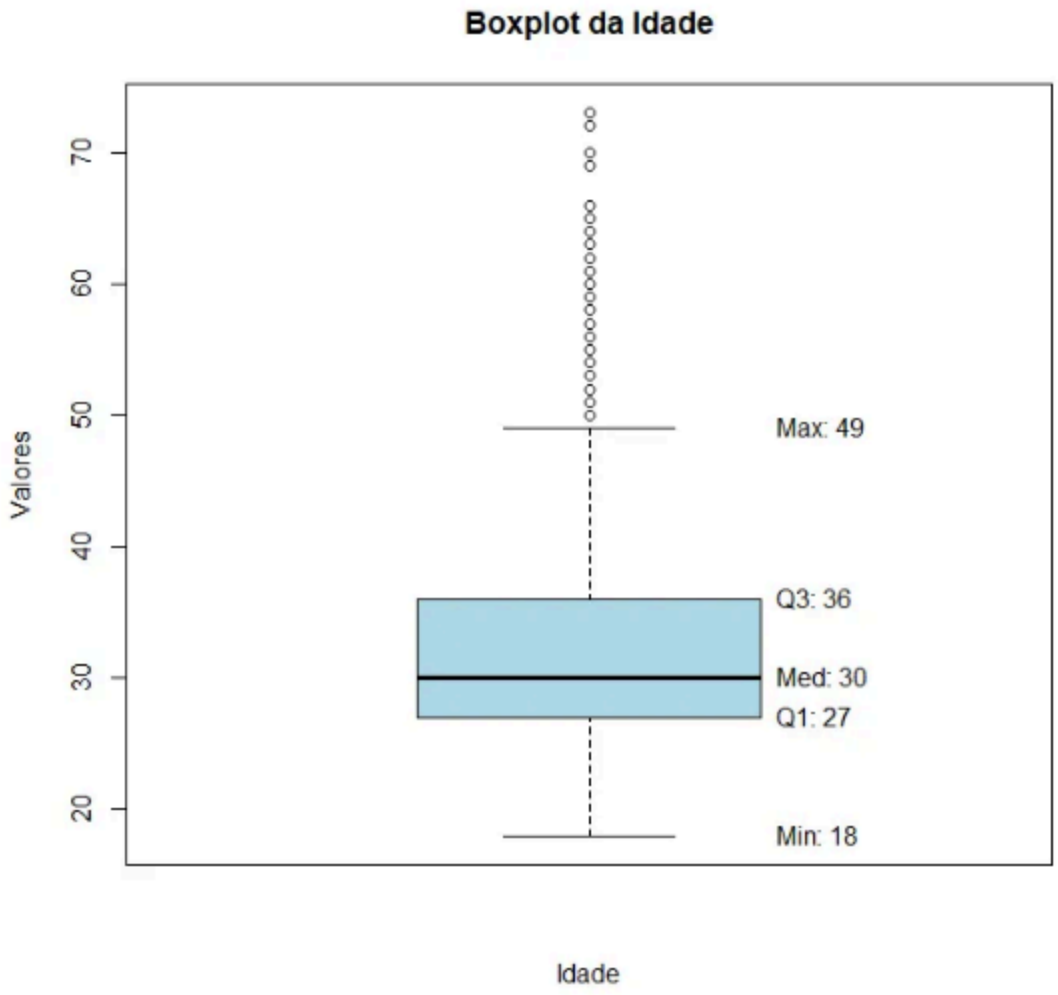
Tendo um desvio padrão de aproximadamente 7 anos, podemos analisar que a maioria das idades está relativamente próxima da média. Com isso observamos que a maioria dos profissionais possuem idade entre 25 e 39 anos aproximadamente, e que a faixa etária predominante na área não possui extremos muito distantes.

```
> # Calcular a variância da coluna "Idade"
> variancia_idade<- var(dados_diversidade_2$`Idade`, na.rm = TRUE)
>
>
> # Exibir a variância da coluna "Idade"
> print(variancia_idade)
[1] 58.11772
>
```

```
> # Calcular o desvio padrão da coluna "Idade"
> desvio_idade <- sd(dados_diversidade_2$`Idade`, na.rm = TRUE)
>
>
> # Mostrar os resultados
> print(desvio_idade)
[1] 7.623498
>
```

Boxplot

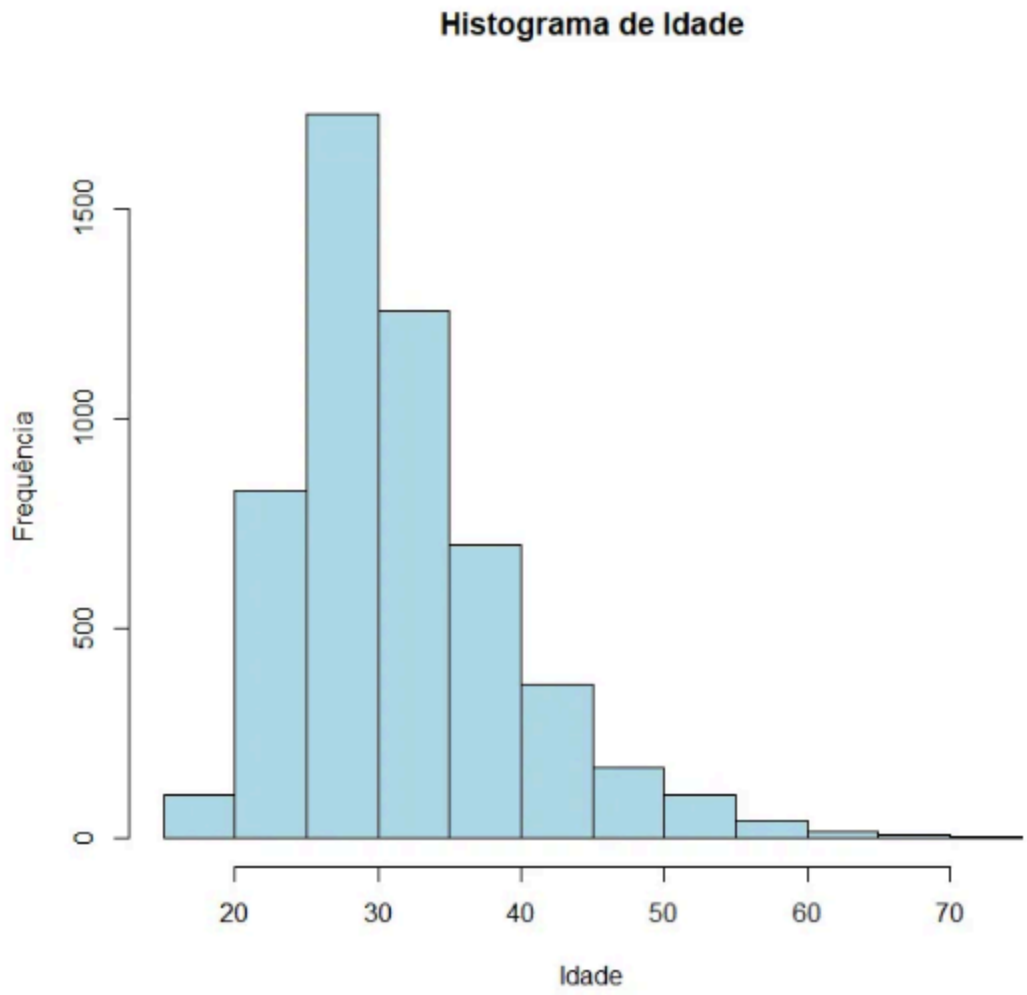
A análise do Boxplot da idade nos permite observar que a idade máxima dos participantes da pesquisa não ultrapassam os 50 anos. Com isso, podemos reafirmar nossa teoria de que o público atuante na área de dados se reflete em um público mais jovem.



Histograma

O Histograma e quartis abaixo, apenas reafirmam toda a conclusão que obtivemos nos gráficos e dados acima, em que a idade de pessoas que trabalham na área de dados que foram entrevistadas, além de não ultrapassar os 50 anos, se reflete em um público relativamente jovem.

```
> # Calcular os quartis da coluna "Idade"
> quartis_idade <- quantile(dados_diversidade_2$`Idade`, probs = c(0.25, 0.5, 0.75), na.rm = TRUE)
>
> # Mostrar os resultados
> print(quartis_idade)
25% 50% 75%
 27  30  36
>
```


Histograma de Idade
Frequência
0
500
1000
1500
20
30
40
50
60
70
Idade