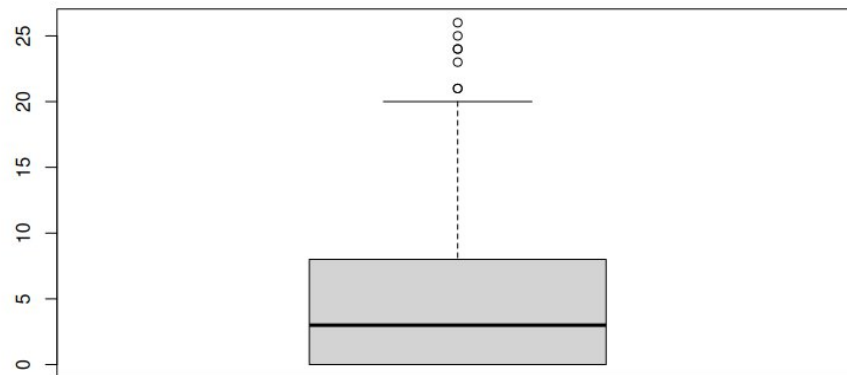


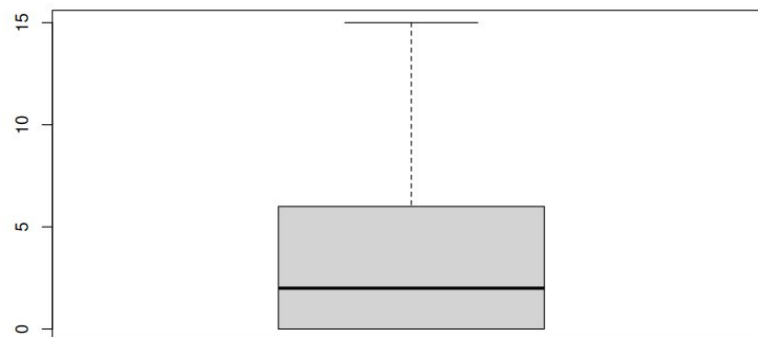
Algumas análises na base de dados do COVID-19 na cidade de São Paulo.

Discrepâncias, *outliers* e óbitos novos.

Primeiro após uma boa limpeza de dados, analisamos via *boxplot* sobre os óbitos novos em São Paulo, verificando as discrepâncias maiores dessa variável, começando a entender conceitos de *outliers*, media, mediana e quartis de forma mais gráfica.

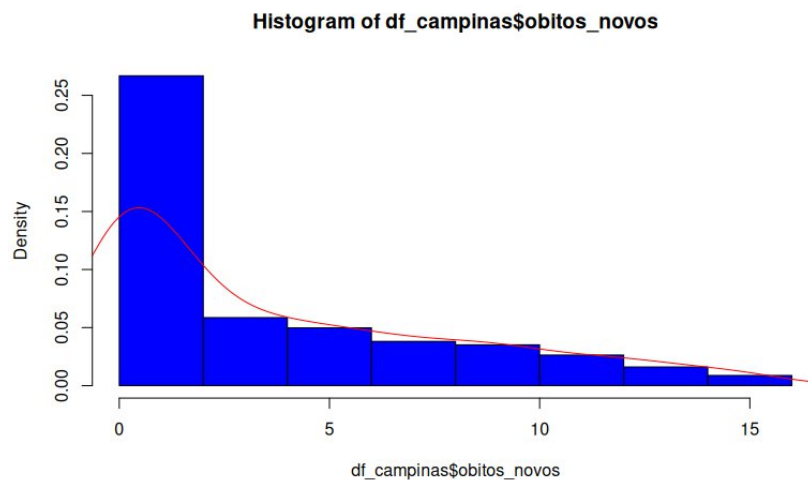


Após isso, fiz uma limpeza nos *outliers*, retirando todos dados menores que o limite inferior, e maiores que os dados do limite superior, chegando a representação:

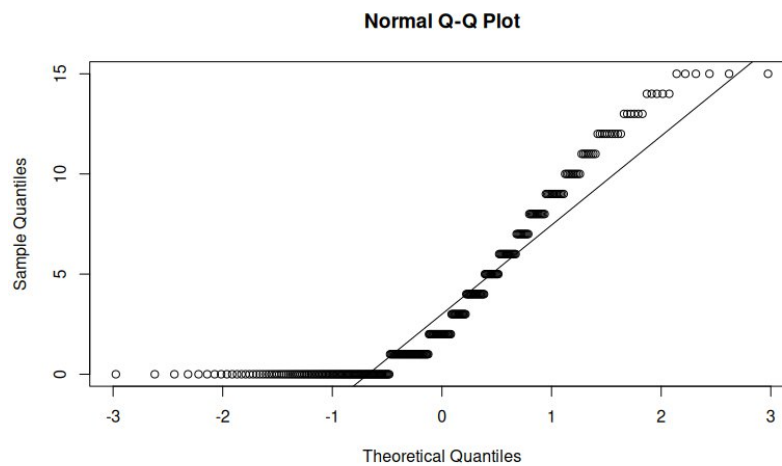


Testes de normalidade

Verificando a variável de óbitos novos, vamos descobrir se ela é uma variável normal simétrica, critério crasso para várias outras funções

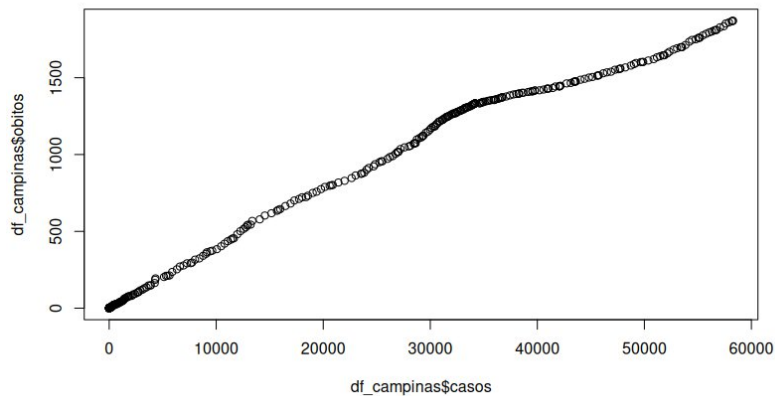


Verificando esse histograma é nítido que não é uma distribuição normal, não segue as regras de desvio padrão nem de mediana. Assim como a linha não forma um sino. Utilizei também o *qqnorm* e o *qqline* para verificar a normalidade, sempre verificando se os pontos seguem as linhas.



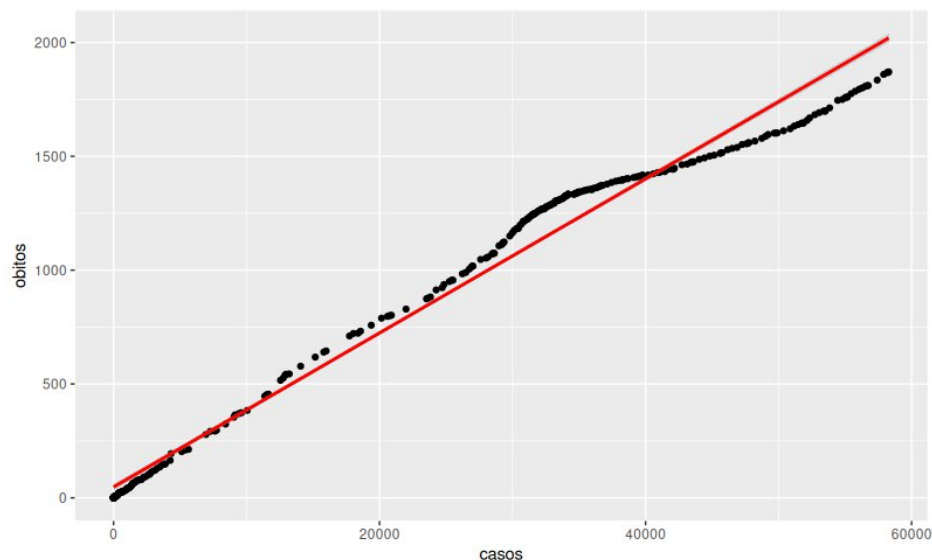
Correlação e regressão linear

Após algumas verificações, queria chegar a algum modelo de regressão linear para a ideia de *casos-óbitos*, verifiquei a normalidade de ambos, e verifiquei o *plot* de linha dessa ideia, sendo possível visualizar que não há normalidade, mas há linearidade.



Dessa forma, utilizei o método de correlação de *Spearman*, para um grande número de dados não paramétricos, chegando a uma correlação de *casos-óbitos* de 99,967%, praticamente perfeita.

A partir disso, utilizei a função *lm* do R para calcular os coeficientes da regressão, conseguindo eles montei a função, retirando o R^2 (ou precisão) de 98% nessa regressão. Após isso utilizei *ggplot* para verificar a linha da minha função de regressão (a vermelha) e os *casos-óbitos* (pontos), verificando a precisão da regressão feita.



Após isso gerei um função da qual calcula, por cidade, quantos óbitos teríamos em determinados casos, utilizando a regressão linear. Como visto no código abaixo.

```
funcao_regressao_obitos_cidade <- function(cidade, casos) {  
  df_regressao <- df %>% filter(municipio == cidade)  
  regressao <- lm(formula = obitos ~ casos, data = df_regressao)  
  coeficientes <- regressao$coefficients  
  
  calculo <- coeficientes[2] * casos + coeficientes[1]  
  r_squared <- summary(regressao)$r.squared  
  return(sprintf("%.2f óbitos com a precisão de %.4f", calculo, (r_squared * 100)))  
}  
funcao_regressao_obitos_cidade("Campinas", 60000)
```