

Análise Exploratória de Dados

E-mail: identificando spams

Dados coletados em 2012, num levantamento feito durante 3 meses pelo cientista de dados David Diez

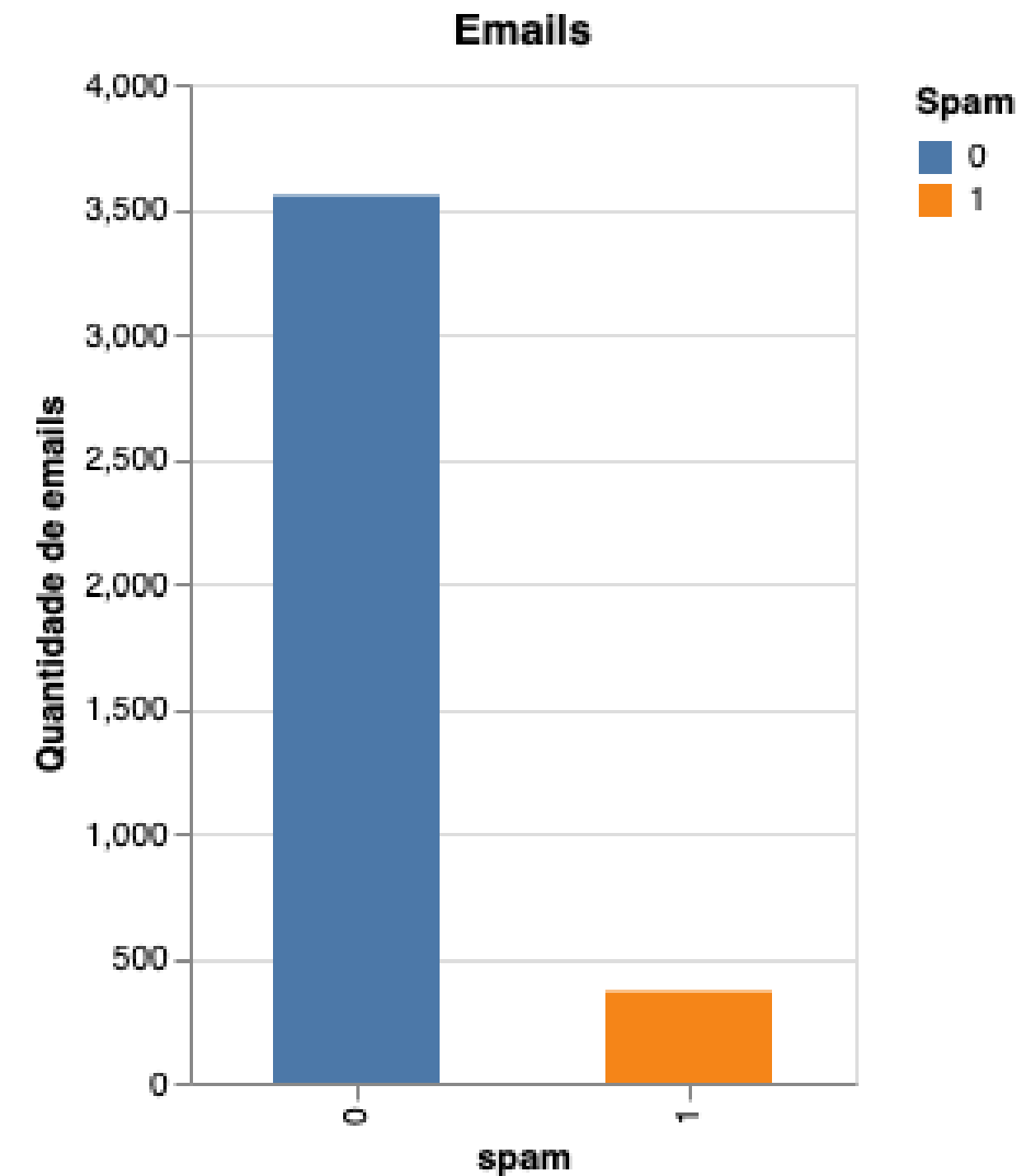
Amanda Audi, Carina Dourado e Jéssica Avelar

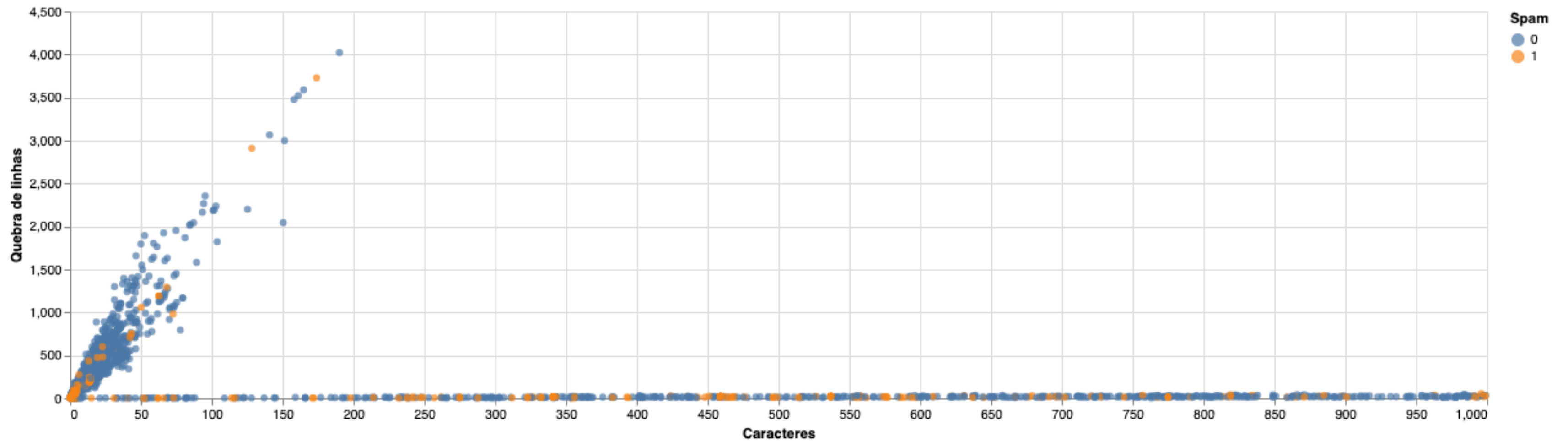


De **3921** e-mails
observados, **367** eram
spams

Dos e-mails recebidos, 9,35% foram
classificados como spams

- . 0 = não (não spams)
- . 1 = sim (spams)



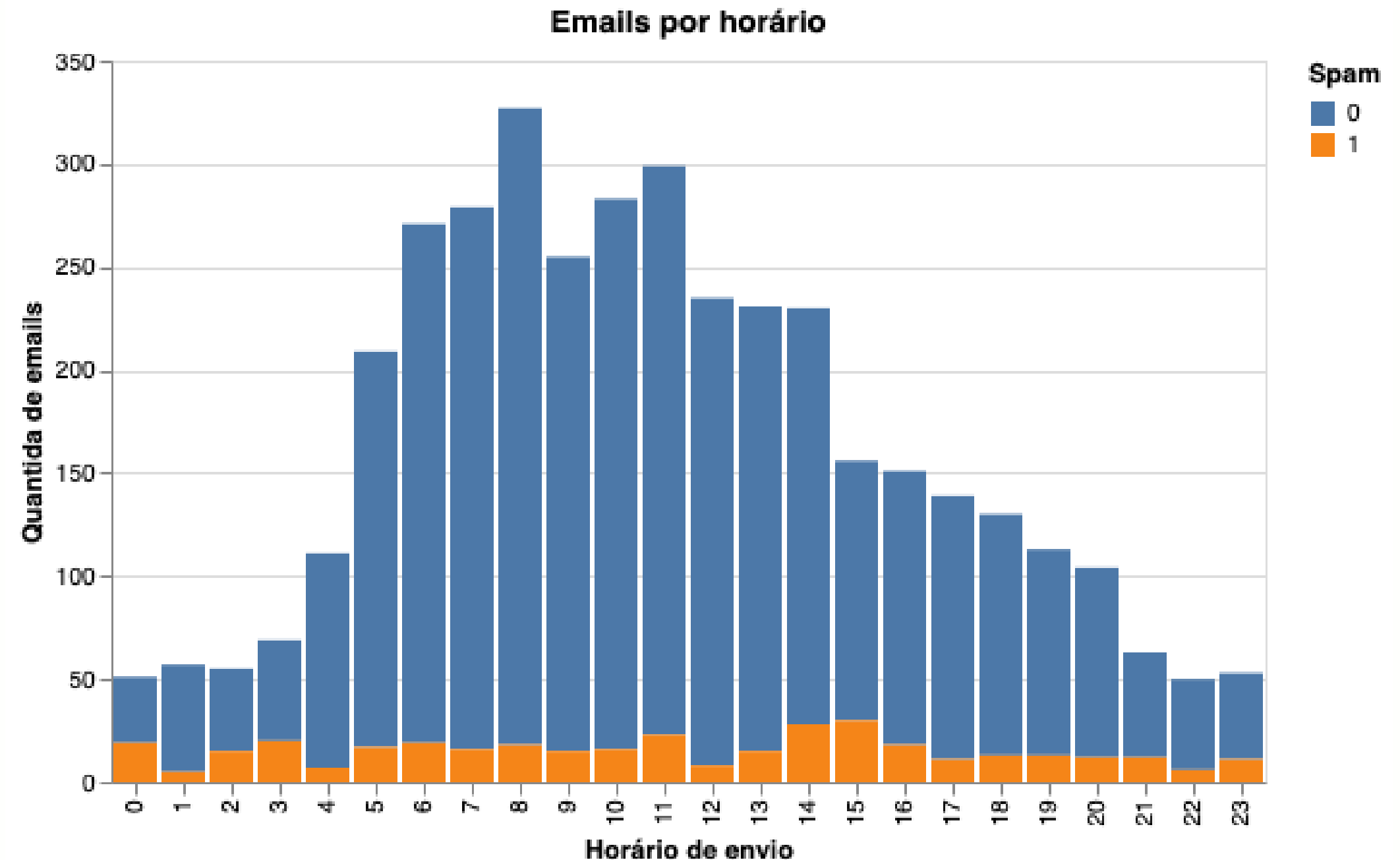


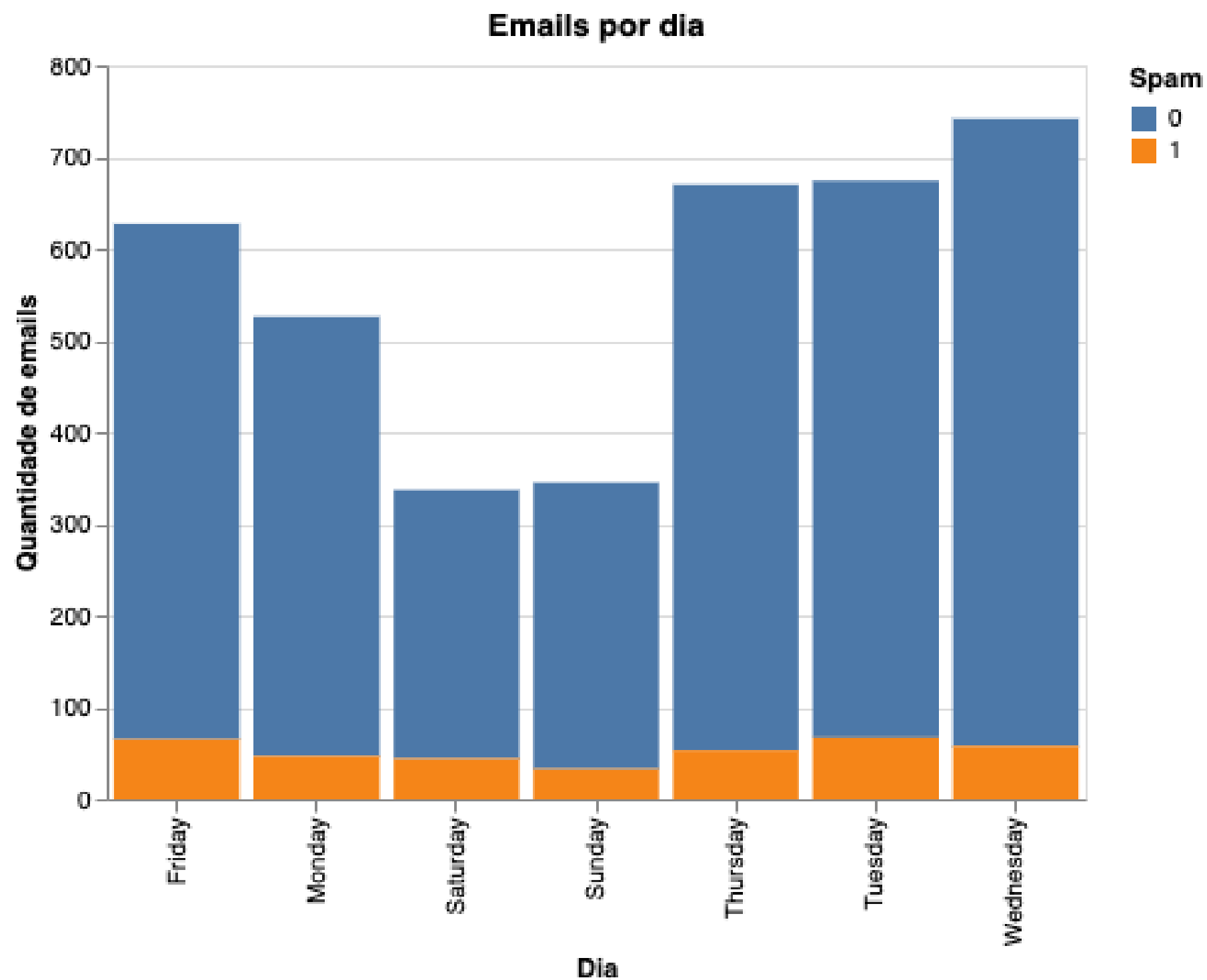
Os **spams** têm, em média, menos quebras de linhas e menos caracteres do que os **comuns**

Picos de horário de envio de **spams** são diferentes dos picos de e-mails **comuns**

E-mails comuns são enviados mais entre 5h e 14h

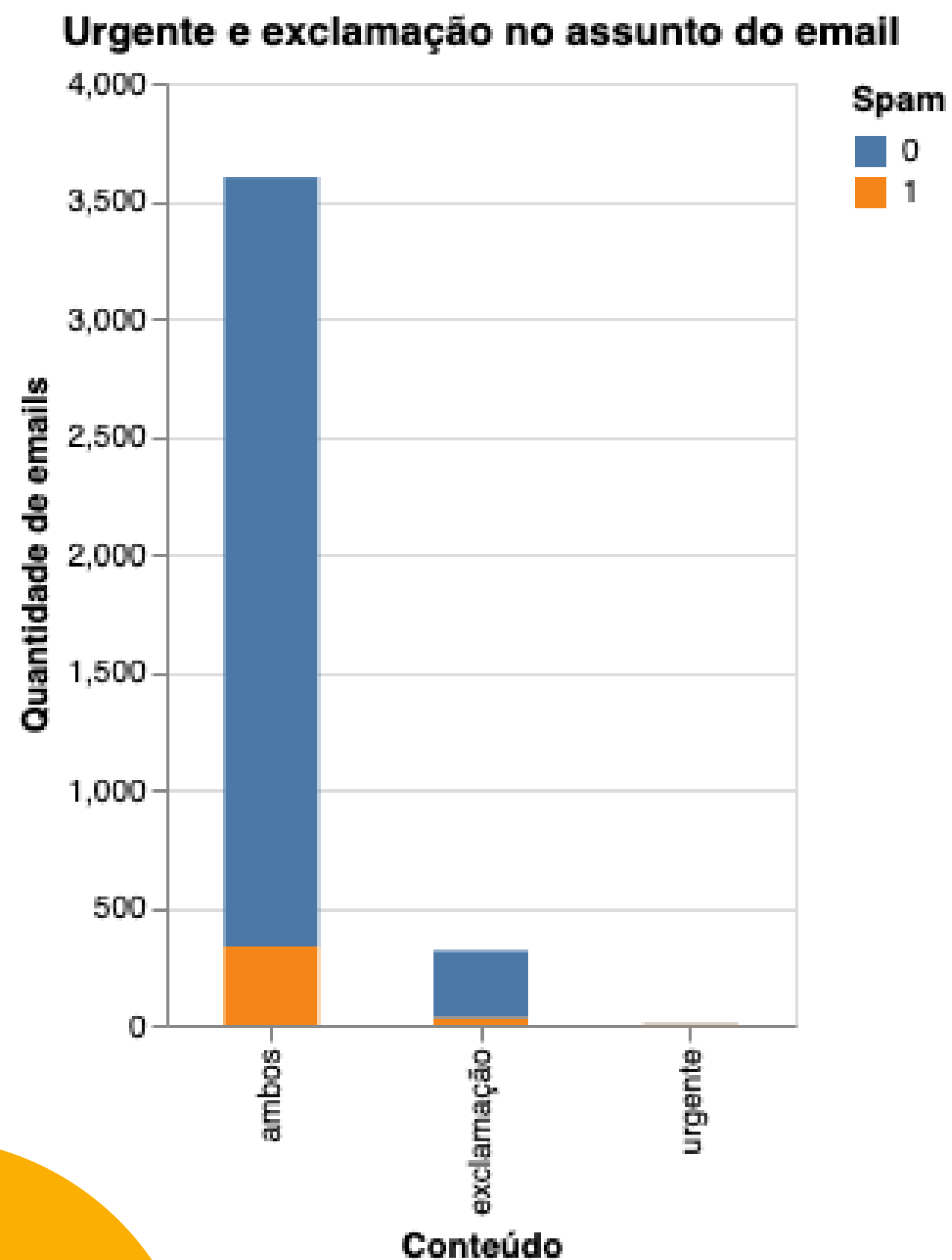
Os envios de spam são mais homogêneos ao longo dia, com picos no fim da manhã, início da tarde e início da madrugada





E-mails **comuns** são mais enviados em dias úteis. **Spams**, apesar de uma redução aos domingos, variam menos durante a semana

Os dias de maior chance de você receber um spam são na **terça** e **sexta-feira**. no **domingo** é menos provável.

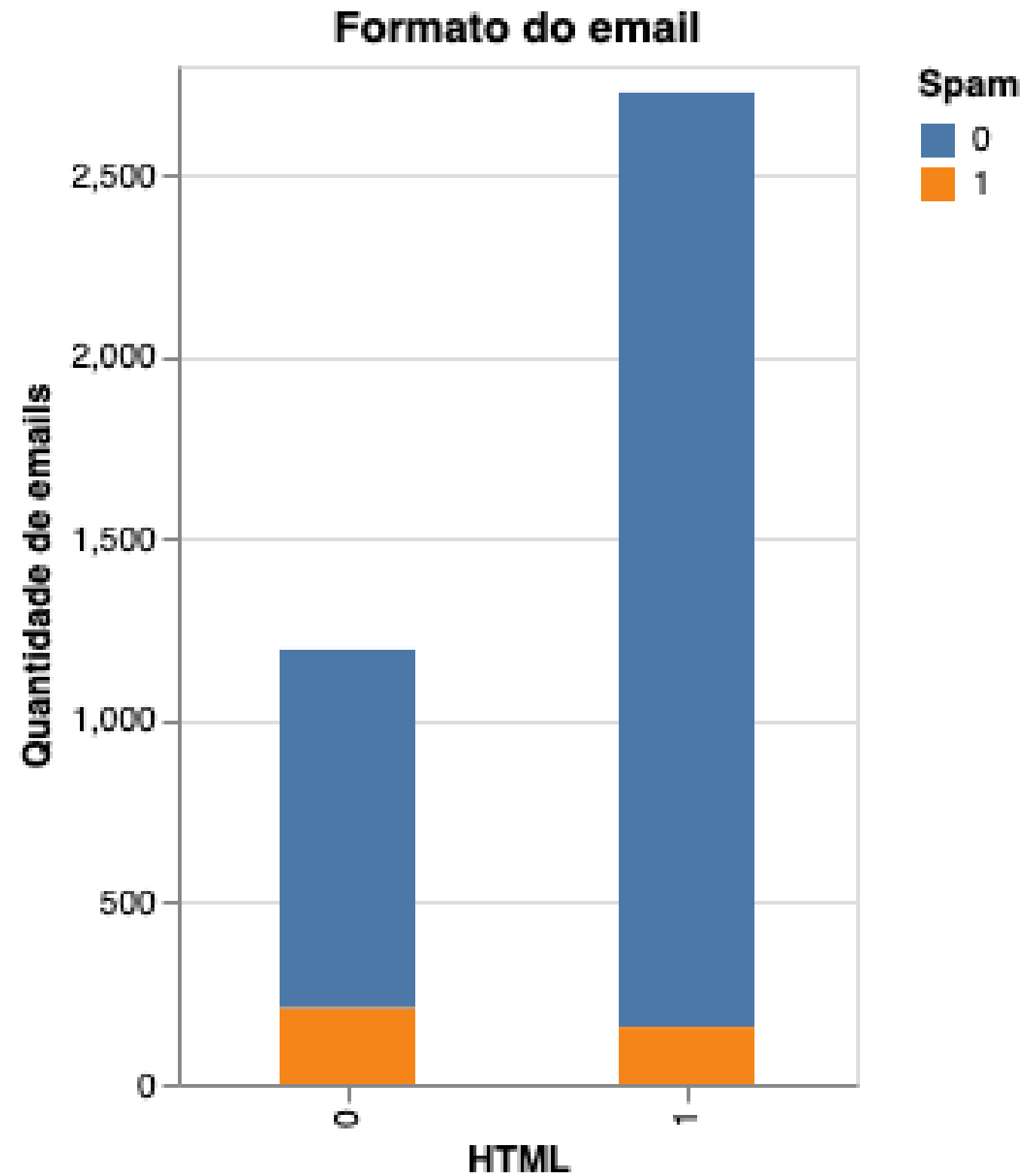


A chance de que um e-mail com "urgente" e "exclamação" no assunto seja **spam** é de 7 a cada 100 e-mails recebidos

Proporcionalmente, este não é o melhor mecanismo para diferenciar um spam: 10% tanto de e-mails comuns quanto de spams possuem exclamação e urgente no assunto

Ponto de exclamação não é algo muito comum nos e-mails tanto comuns, quanto spams



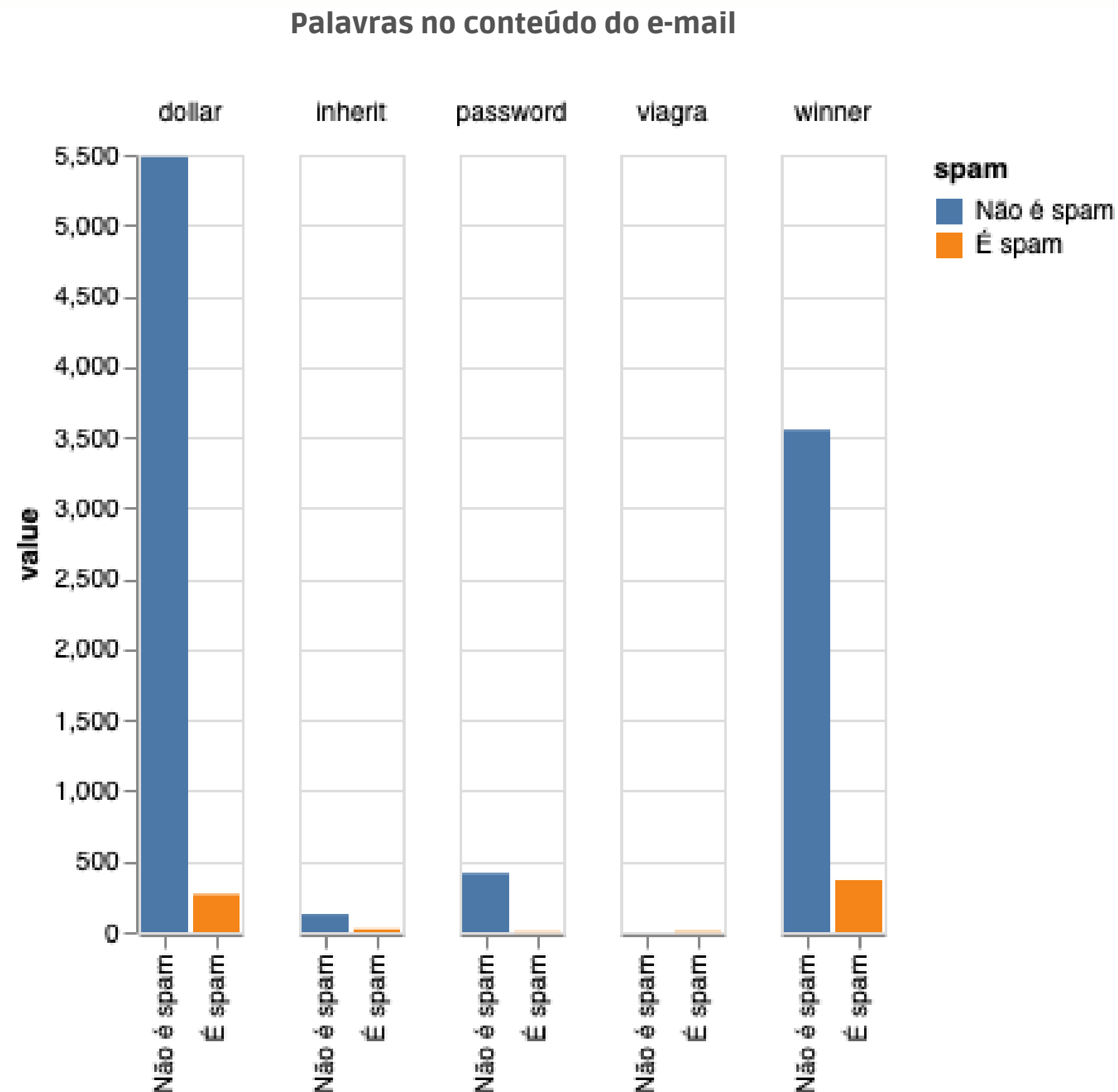


E-mails comuns usam
mais o formato HTML
no conteúdo que spams





Mas é mais provável que um spam
não seja de formato HTML

As palavras "manjadas" não são os melhores filtros de spams

Mesmo assim, palavras mais prováveis de aparecer em um spam são: dollar, winner (vencedor) e inherit (herdar) do que password (senha) e viagra, por exemplo.



OBSERVAÇÕES

-  spams correspondem a cerca de 10% do total de e-mails recebidos
-  geralmente tem o conteúdo menor do que e-mails comuns
-  palavras e exclamações não são bons parâmetros para classificar um e-mail como spam
-  os spams têm frequência de envio mais homogênea em horários e dias de semana

CONSIDERAÇÕES FINAIS

Os dados foram coletados em 2012, desde então, muitas mudanças ocorreram nos servidores de e-mails para classificar um conteúdo como spam - assim como as formas de envio e conteúdos de spam também mudaram.

O conjunto de dados corresponde a apenas um servidor de e-mail, o Gmail, e de apenas uma pessoa. Para uma melhor conclusão, seria preciso analisar uma variedade maior de usuários e de servidores.

Devido às colocações apresentadas acima, não é possível aplicar esta análise para fazer inferências, como um padrão ou demonstrativo da realidade. Ainda assim, a base de dados nos dá uma ideia do que poderia ser feito para uma análise completa de e-mails classificados como spam.