Análise dos Dados do Bot no Carnaval de Belo Horizonte

Processo Seletivo para Vaga de Engenharia de Dados **Candidato:** Renan Santos Mendes

June 1, 2018



Sumário

- Pergunta 1
- Pergunta 2
- Pergunta 3
- Pergunta 4
- Pergunta 5
- Novas hipóteses
- Mipóteses não validadas



Pergunta 1

Qual a quantidade de mensagens dos usuários?



Filtrando as mensagens, temos que há um total de 11022 mensagens enviadas pelos usuários.



Pergunta 2

Quantos são os usuários únicos do Bot?



Filtrando as mensagens, temos que há um total de 22744 usuários do Bot.



Pergunta 3

Qual a quantidade de mensagens trafegadas?



Somando as mensagens enviadas pelos usuários e pelo bot, temos um total de 33766 mensagens.



Pergunta 4

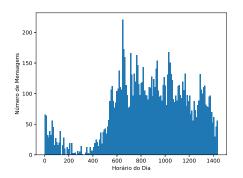
Apresente, da forma como achar melhor, as interações que os foliões tiveram com o Bot.



- Uma forma de ver como os usuários interagiram com o bot, é analisando os horários em que os usuários mais utilizaram o bot;
- Para isso, será feito um histograma contando o número de mensagens por hora;
- Uma pequena alteração nos dados será feita, o horário em que a mensagem foi enviada será convertido para minutos para podermos plotar os dados;
- Assim, o dado não perderá o valor contido.



 Histograma do número total de mensagens enviadas por horário em todos os dias do carnaval.



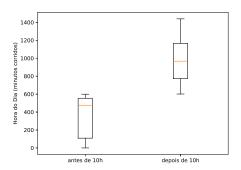




- Percebe-se uma diferença no número de mensagens enviadas entre os intervalos de 0h às 10h e 10h às 23h;
- Algo que é esperado de ante-mão, uma vez que a programação do carnaval de BH é basicamente nesta faixa de horário;
- Apenas para termos certeza, vamos fazer um boxplot para comparar.



• Boxplot para os dois intervalos em análise no carnaval.



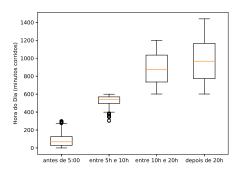


- Como não há intercessão entre as duas caixas, podemos dizer que há diferença estatisticamente significante entre os dois intervalos;
- Para fazer uma segunda análise, vamos dividir os dados em mais intervalos e fazer um boxplot:
 - **1** t < 5h
 - **2** 5h < t < 10h

 - **4** 20h < t



• Boxplot para os novos intervalos em análise no carnaval.





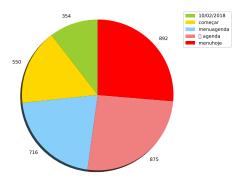
- Fica claro novamente que há diferença na utilização do bot entre os horários de 0h às 5h, 5h às 10h e 10h às 20h;
- Como aparentemente não há diferença na utilização de 10h até 23h59 é necessário aplicar um teste estatístico para se comparar estes valores;
- Vale ressaltar que esta sendo feita apenas uma análise macro da utilização do bot;
- Para entendermos melhor sua utilização é necessário fazer um splot dos dados referentes a cada dia da semana, ver se há diferença entre os intervalos em um mesmo dia e entre cada dia.

Pergunta 5

Quais botões foram os que tiveram maior interação?



 Quantidade de vezes que os 5 botões mais utilizados foram pressionados





- Análise foi feita com base nos botões apresentados no Fluxo Figma;
- O botão HOJE, denominado nos dados como MenuHoje, foi o mais utilizado, sendo apertado 892 vezes;
- Isso mostra que os usuários se interessaram pela programação do dia apesar da grande utilização da agenda (875 vezes).



Novas Hipóteses

Novas Hipóteses a serem inferidas sobre os dados



Novas Hipóteses

- Quais usuários mais utilizaram o bot?
- Quais dias do carnaval o bot foi mais utilizado? Como foi a utilização durante o mês?
- é possível descrever o comportamento da utilização do bot no intervalo de 10h até 23h por meio de um modelo de regressão linear?



Hipótese 1 - Quais usuários mais utilizaram o bot?

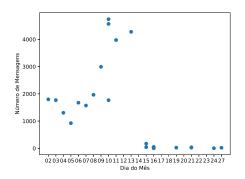
 Fazendo uma contagem dos dados dos usuários temos que os 5 usuários que mais usaram o bot são:

Qtd. Mensagens
550
135
81
77
76

 Parte da string (@messenger.gw.msging.net) foi retirada do ID pra facilitar a visualização da tabela anterior.

Hipótese 2 - Quais dias do carnaval o bot foi mais utilizado? Como foi a utilização durante o mês?

• Fazendo um gráfico de dispersão dos dados temos:







Hipótese 2 - Quais dias do carnaval o bot foi mais utilizado? Como foi a utilização durante o mês?

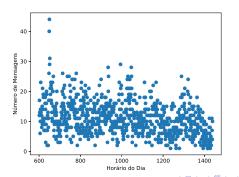
- Com base no gráfico anterior, percebe-se que:
 - O maior uso do bot foi entre os dias 9 e 13, que corresponde, respectivamente, ao período da sexta-feira de carnaval à terça-feira
 - 4 Houve um uso moderado do bot na semana de pré-carnaval (período entre os dias 2 e 8 de fevereiro);
 - 3 Após o carnaval, o uso do bot foi baixo.



- Qual seria a utilidade de um modelo de regressão neste caso?
 - Obter uma estimativa de quantas mensagens estariam sendo trafegadas caso fosse necessário dar alguma manutenção no bot ou ele parasse de funcionar;
 - Pensar na melhor forma, ou melhor horário, para dar manuteção - pensando no caso de manutenção preditiva



 Para termos uma maior noção do que acontece com os dados, vamos fazer um scatterplot para ver a dispersão.





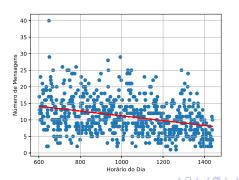
- Apesar da grande variabilidade (dispersão) dos dados, aparenta existir uma relação linear negativa;
- Calculando coeficientes de correlação entre o horário e o número de mensagens:

Método	Coef. Correlação
Pearson	-0.3295969
Spearman	-0.3290413
Kendall	-0.2302467

 Apesar de serem negativos, seus valores são pequenos e representam uma correlação fraca.



- Os dados foram separados em teste e treino;
- Os dados de treino foram utilizados para gerar a seguinte reta:





• Considerando x como horário do dia e y como o número de mensagens, a equação para o modelo linear é a seguinte:

$$y = 18.44918802 - 0.00723053^*x$$

 Usando os dados de teste, temos que o valor de RMSE é igual a 24.63, o que podemos considerar um erro relativamente grande.



- Não houve um bom ajuste do modelo;
- Uma alternativa seria alterar os dados, uma vez que estão apresentados em minutos, seria viável agregar eles por faixa horária (5 em 5 minutos, 10 em 10 minutos, dentre outras);
- Uma segunda alternativa seria em abrir os dados para cada dia, considerando os dias mais utilizados do bot, assim evitando influência de dias em que o bot foi pouco utilizado.



Hipóteses não validadas

Outras hipóteses formuladas para serem testadas futuramente



Hipóteses não validadas

- O bot atendeu a necessidade do usuário pensar em alguma forma de feedback ao fechar o bot
- 2 Haveria como localizar de onde o usuário utilizou o bot? IP?
- Maveria diferença nos resultados obtidos ao se comparar cada dia do carnaval? (pensando numa análise micro dos dados).
- Seria possível fazer uma análise de sentimento da interação do usuário com o bot? (para entender o comportamento do usuário).



Hipóteses não validadas

- Com base no comportamento do usuário, seria possível fazer alguma recomendação de evento ou local?
- ② Como os coeficientes de correlação do modelo não mostraram uma forte correlação, uma alternativa seria em organizar os dados, dividindo por intervalos de tempo (por exemplo, de meia em meia hora) e refazer o modelo linear.



Sumário Perguntas Novas hipóteses Hipóteses não validadas

Obrigado pela atenção!

