Big Data e Machine Learning com Hadoop e Spark

Conteúdo

CONTEÚDO PROGRAMÁTICO

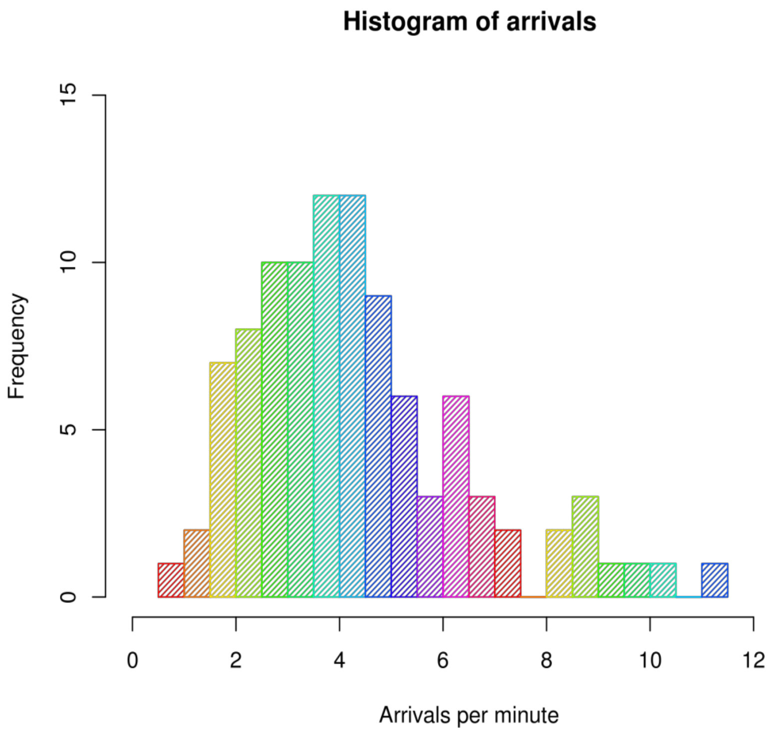
- Visão geral da ciência de dados e aprendizado de máquina em escala
- Visão geral do ecossistema do Hadoop
- Instalação de um Cluster Hadoop
- Trabalhando com dados do HDFS e tabelas do Hive usando o Hue
- Visão geral do Python
- Visão geral do R
- Visão geral do Apache Spark 2
- Leitura e gravação de dados
- Inspeção da qualidade dos dados
- Limpeza e transformação de dados
- Resumindo e agrupando dados
- Combinando, dividindo e remodelando dados
- Explorando dados
- Configuração, monitoramento e solução de problemas de aplicativos Spark
- Visão geral do aprendizado de máquina no Spark MLlib
- Extraindo, transformando e selecionando recursos
- Construindo e avaliando modelos de regressão
- Construindo e avaliando modelos de classificação
- Construindo e avaliando modelos de cluster
- Validação cruzada de modelos e hiperparâmetros de ajuste
- Construção de pipelines de aprendizado de máquina
- Implantando modelos de aprendizado de máquina

MATERIAL DIDÁTICO

- Slides do treinamento em PDF
- GitHub com exercícios e códigos exemplo
- Máquinas virtuais para exercícios simulados
- Gravação das aulas disponível durante 3 meses

Desvio Padrão e Variância

Exemplo de um Histograma



Variância mede quão “espalhados” os dados são.

- Variância (σ^2)é simplesmente a **média das diferenças quadradas da média**
- Exemplo: Qual a variância deste dataset (1, 4, 5, 4, 8)?
  - Calcule a Média: (1+4+5+4+8)/5 = 4.4
  - Agora encontre as diferenças da média: (-3.4, -0.4, 0.6, -0.4, 3.6)
  - Encontre o quadrado das diferenças: (11.56, 0.16, 0.36, 0.16, 12.96)
  - Calcule a média do quadrado das diferenças:

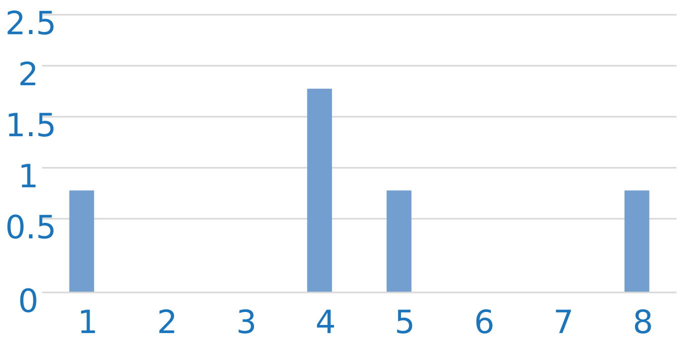
$$\sigma^2 = (11.56 + 0.16 + 0.36 + 0.16 + 12.96) / 5 = 5.04$$

paulo.rodrigues@konfidentia.com

Desvio Padrão é a raiz quadrada da Variância

$\sigma 2 = 5.04$

$\sigma = \sqrt{5.04} = 2.24$



Então o Desvio Padrão de
(1, 4, 5, 4, 8) é 2.24.

Isso é normalmente usado para identificar outliers. Pontos que ficam a mais de um Desvio Padrão da Média podem ser considerados não usuais.

Você pode se referei a quão extremo é um ponto de dados dizendo "quantos sigmas" longe da média ele está.

População vs. Amostra

- Se você está trabalhando com uma Amostra dos dados ao invés de Um dataset completo de dados (a *População* inteira)…
  - Então você vai querer usar a "variância da amostra" ao invés da "variância da população"
  - Para N amostras, você divide a variância quadrada por N-1 ao invés de N.
  - Então, no nosso exemplo, calculamos a variância da população assim:
    - $\sigma^2 = (11.56 + 0.16 + 0.36 + 0.16 + 12.96) / 5 = 5.04$
  - But the sample variance would be:
    - $S^2 = (11.56 + 0.16 + 0.36 + 0.16 + 12.96) / 4 = 6.3$

Fórmulas

- Variância da População:

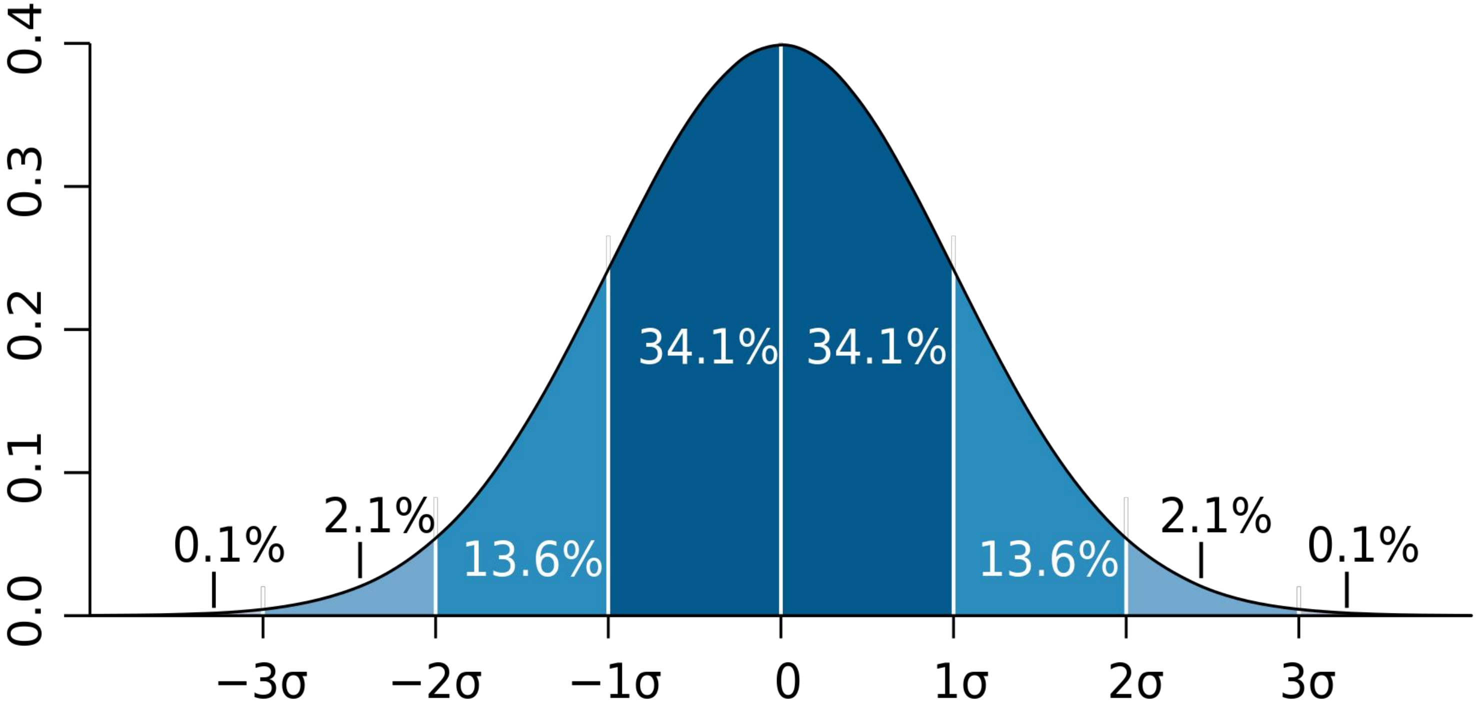
$$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$$

- Variância da Amostra:

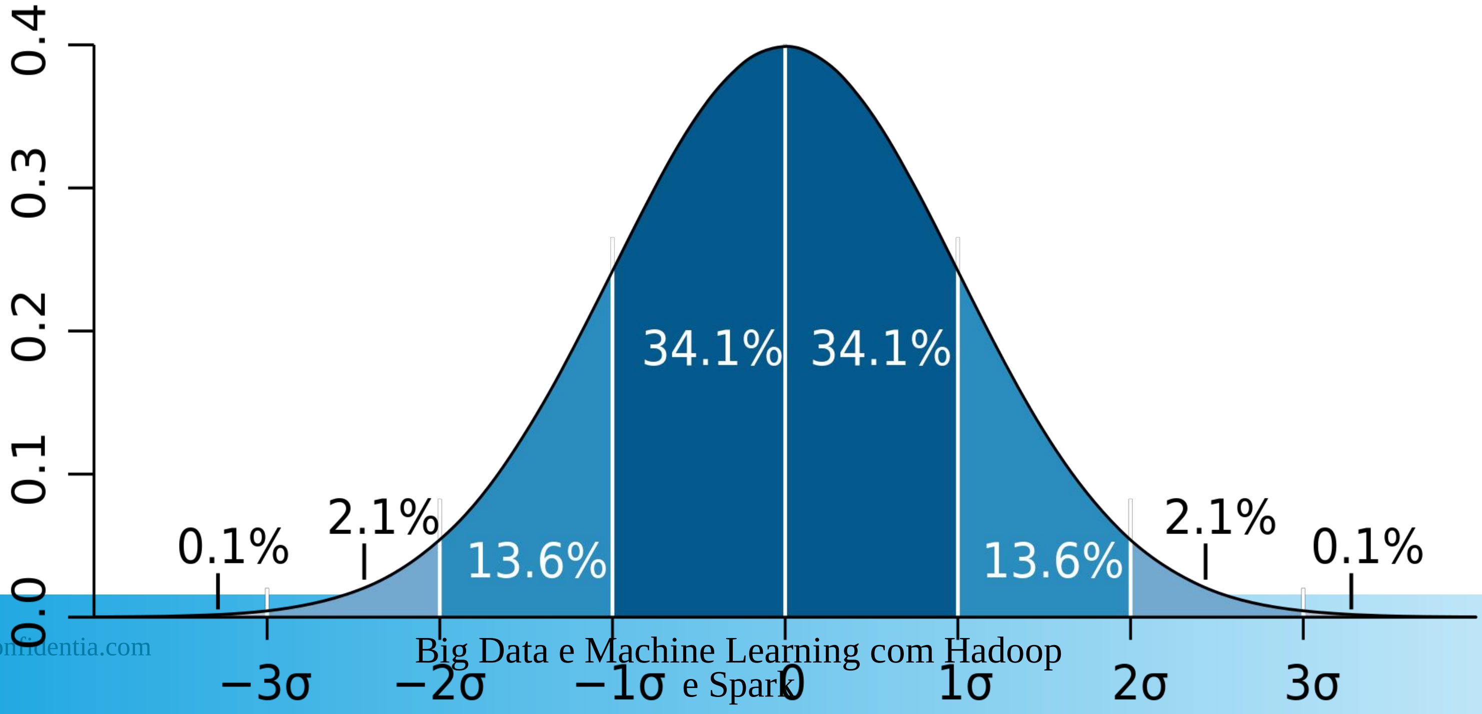
$$s^2 = \frac{\sum(X-M)^2}{N-1}$$

Funções de densidade de probabilidade

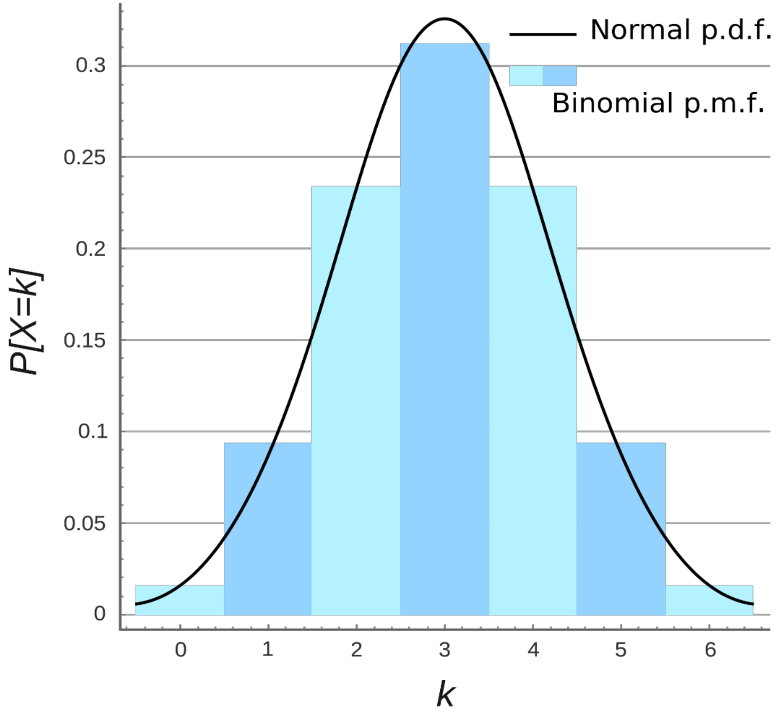
Exemplo: uma “distribuição normal



Dá a probabilidade de um ponto de dados cair dentro de um dado intervalo de um dado valor.



Função de massa de probabilidade

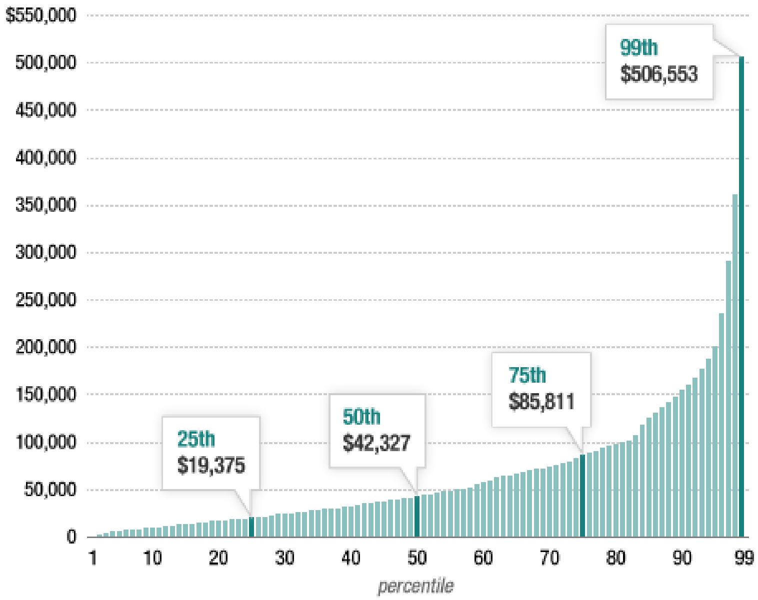


Vamos ver alguns exemplos

Percentis e Momentos

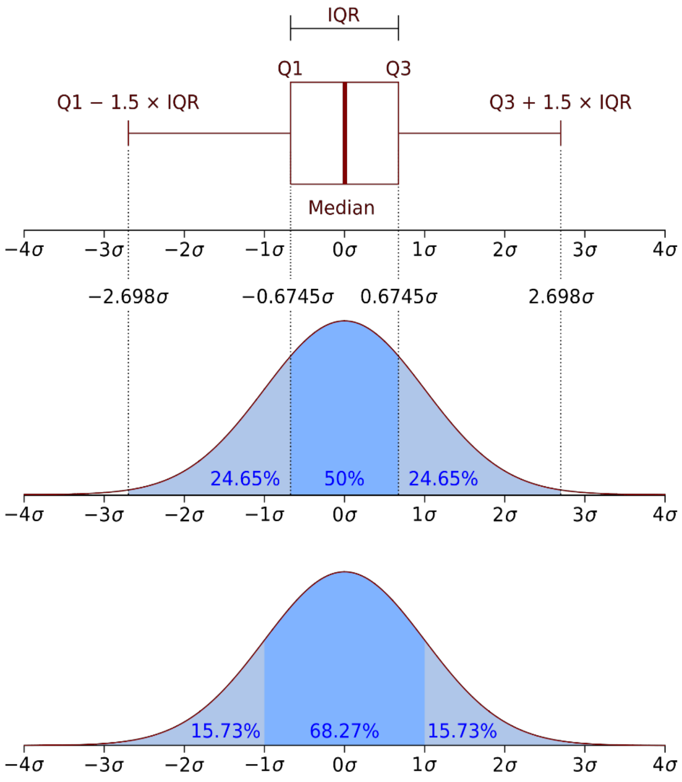
Percentis

- Em um conjunto de dados, qual é o ponto em que X% dos valores são menores que esse valor?
- Exemplo: distribuição de renda



paulo.rodrigues@konfidentia.com

Percentis em uma distribuição normal



paulo.rodrigues@konfidentia.com

Vamos ver alguns exemplos

Momentos

Medidas quantitativas da forma de uma função de densidade de probabilidade Matematicamente elas são um pouco difíceis de entender:

$$\mu_n = \int_{-\infty}^{\infty} (x - c)^n f(x)\,dx$$ (para um momento *n* em torno do valor *c*).

Mas intuitivamente, é muito mais simples em estatística.

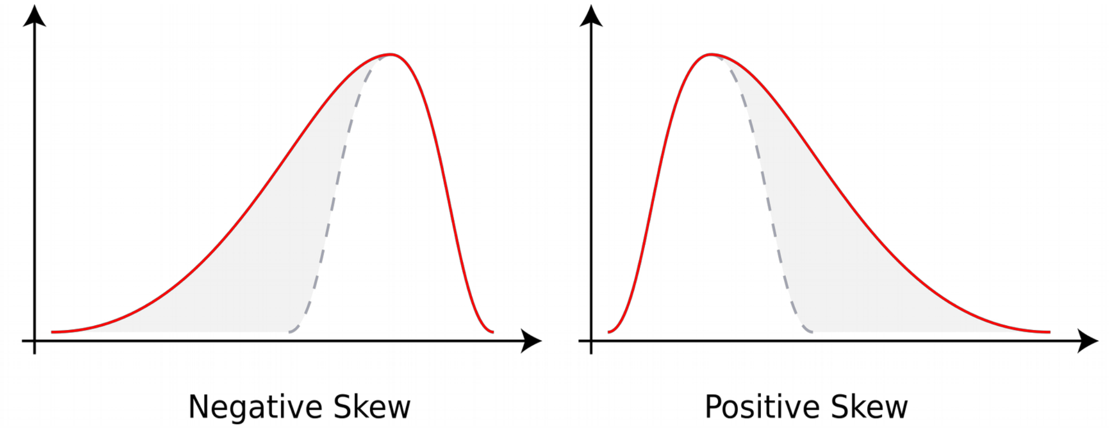
O primeiro momento é a média

O segundo momento é a variância

Simples assim...

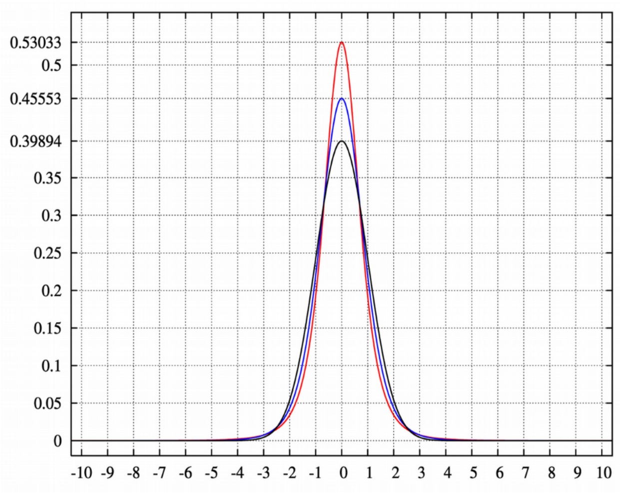
O terceiro momento “inclinação”

Quão “desequilibrada” é a distribuição? Uma distribuição com uma cauda mais longa à esquerda ficará inclinada para a esquerda e terá uma inclinação negativa.



O quarto momento é "curtose"

Quão espessa é a cauda e quão nítido é o pico, comparado a uma distribuição normal? Exemplo: picos mais altos têm maior curtose

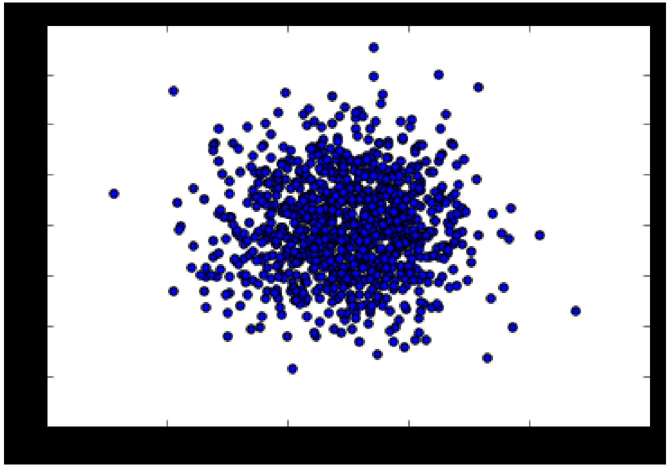


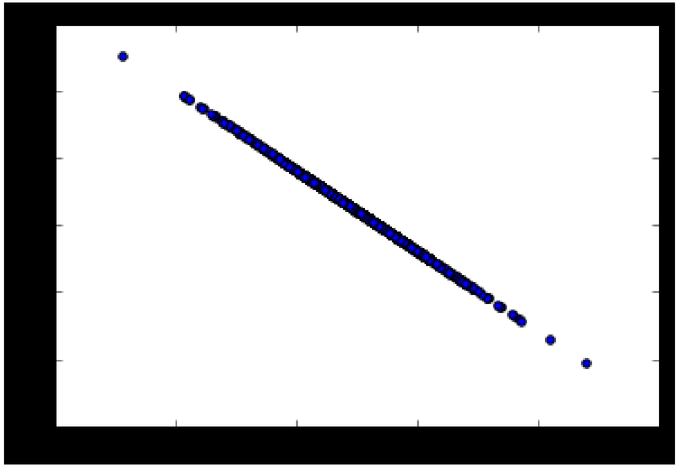
Vamos computar os 4 momentos com Python

Covariância e Correlação

Covariância

Mede como duas variáveis variam em conjunto a partir de suas médias.





paulo.rodrigues@konfidentia.com

Medindo a Covariância

- Pense nos conjuntos de dados para as duas variáveis como vetores de alta dimensionalidade
- Converta-os em vetores de variações a partir da média
- Pegue o produto escalar (cosseno do ângulo entre eles) dos dois vetores
- Divida pelo tamanho da amostra

Interpretar covariância é difícil

- Sabemos que uma pequena covariância, próxima de 0, significa que não há muito correlação entre as duas variáveis.
- E grandes covariâncias - ou seja, longe de 0 (pode ser negativo para inverso relacionamentos) significa que há uma correlação
- Mas quão grande é “grande”?

É aí que entra a correlação!

- Apenas divida a covariância pelos desvios padrão de ambas as variáveis, e isso normaliza as coisas.
- Portanto, uma correlação de -1 significa uma correlação inversa perfeita
- Correlação de 0: sem correlação
- Correlação 1: correlação perfeita

Lembre-se: a correlação não implica causalidade!

- Somente um experimento controlado e randomizado pode fornecer informações sobre causalidade.
- Use a correlação para decidir quais experimentos realizar!

Vamos ver alguns exemplos

Obrigado!!!

Nos vemos amanhã!!!

Bom descanso!