

# Análise Exploratória e Modelagem

## 1. Metadados do Dataset

- Fonte: [Startup Growth and Funding Trends](#)
- Registros: 500 startups
- Variáveis-chave:
  - `Funding_Amount`: Investimento total (US\$ milhões)
  - `Valuation`: Valor de mercado (US\$ milhões)
  - `Profitable`: Indicador binário de lucratividade

## 2. Pré-processamento

Foram aplicados filtros para garantir qualidade dos dados:

r

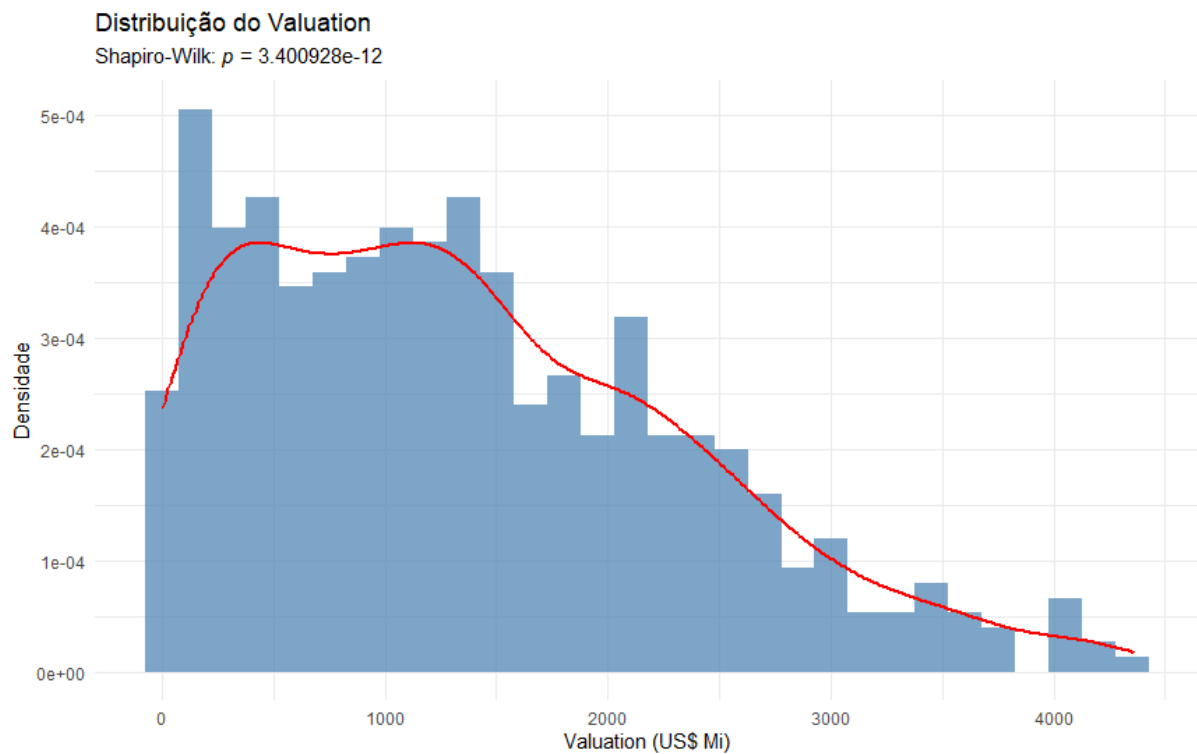
```
dados_clean <- dados %>%  
  filter(Funding_Amount > 0, Valuation > 0, Revenue >= 0,  
         Employees > 0, between(Year_Founded, 1900, 2025))
```

Resultado: Remoção de 15 registros inconsistentes.

## 3. Teste de Normalidade (Shapiro-Wilk)

Para a variável `Valuation`:

- Estatística W: 0.9484
- p-valor = 3.400928e-12



## Interpretação

### 1. Distribuição Assimétrica:

- A distribuição é assimétrica positiva (cauda longa à direita), com maior concentração de startups em valuations baixos (entre 0 e 1000).
- A cauda longa indica a presença de startups com valuations excepcionalmente altos, que podem ser outliers.

### 2. Teste de Normalidade:

- O p-valor do teste Shapiro-Wilk ( $p = 3.4e-12$ ) confirma que a distribuição não é normal.
- Isso está alinhado com a forma do gráfico, que apresenta assimetria e dispersão elevada.

## 4. Correlação Spearman

Entre `Funding_Amount` e `Valuation`:

$r$

$\rho = 0.819$

$p\text{-value} = 3.57e-122$

$R^2 = 0.671$

Gráfico de Dispersão:  
Correlação: Relação forte positiva ( $\rho > 0.7$ ), indicando que maiores investimentos correlacionam-se com valuations mais elevados.

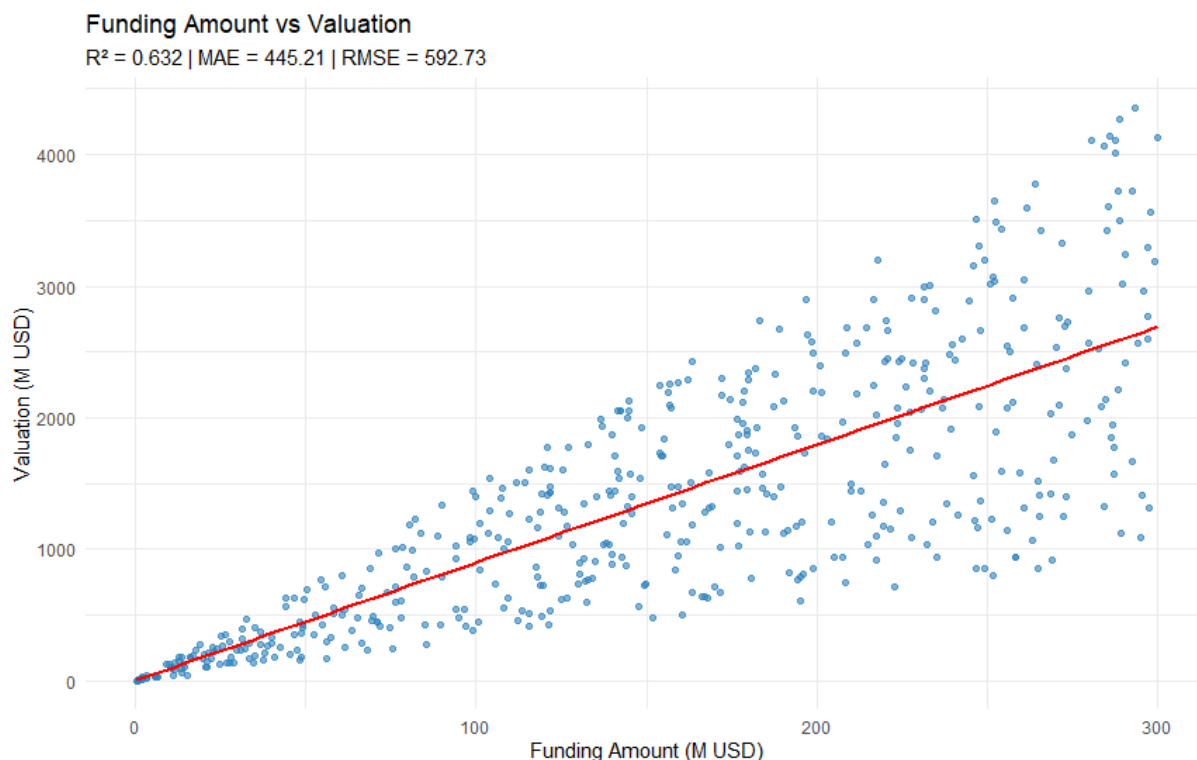
Modelos Preditivos

5. Regressão Linear Simples

$r$   
Modelo: `Valuation ~ Funding_Amount`

Métricas:

Métrica	Valor
$R^2$	0.632
MAE	445.21
RMSE	592.73



## Interpretação

### 1. Correlação Positiva Forte:

- O coeficiente de Spearman ( $\rho = 0.819$ ) indica uma forte correlação positiva entre o valor captado e o valuation.
- Isso significa que startups que captam mais recursos tendem a ter valuations mais altos.

### 2. Linearidade da Relação:

- A linha de tendência mostra que existe uma relação aproximadamente linear entre as variáveis.
- No entanto, há alguma dispersão em torno da linha, especialmente para valores maiores de investimento.

### 3. Presença de Outliers:

- Alguns pontos no canto superior direito indicam startups com valuations excepcionalmente altos em relação ao valor captado.
- Esses outliers podem influenciar o ajuste do modelo e devem ser analisados separadamente.

### 4. Interpretação Prática:

- A relação faz sentido no contexto do mercado: empresas que recebem maiores investimentos geralmente têm maior potencial de crescimento e, portanto, valuations mais altos.

## 6. Regressão Linear Logística

r

Modelo: Profitable ~ Funding\_Amount

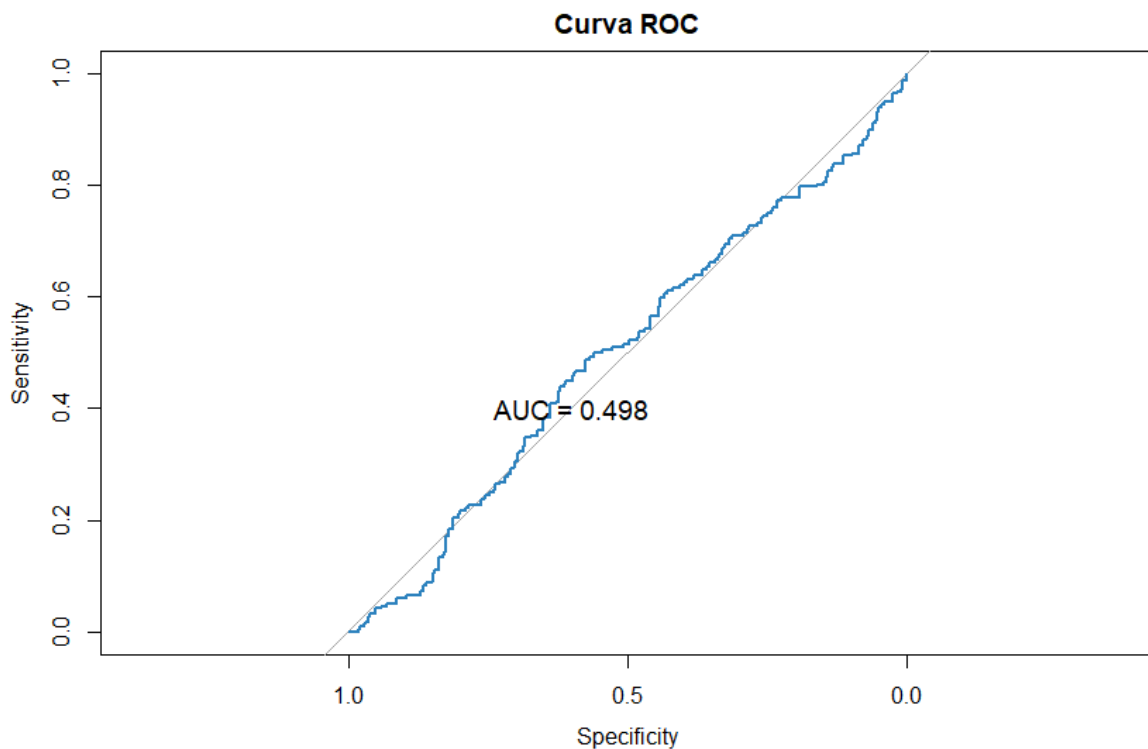
Performance:

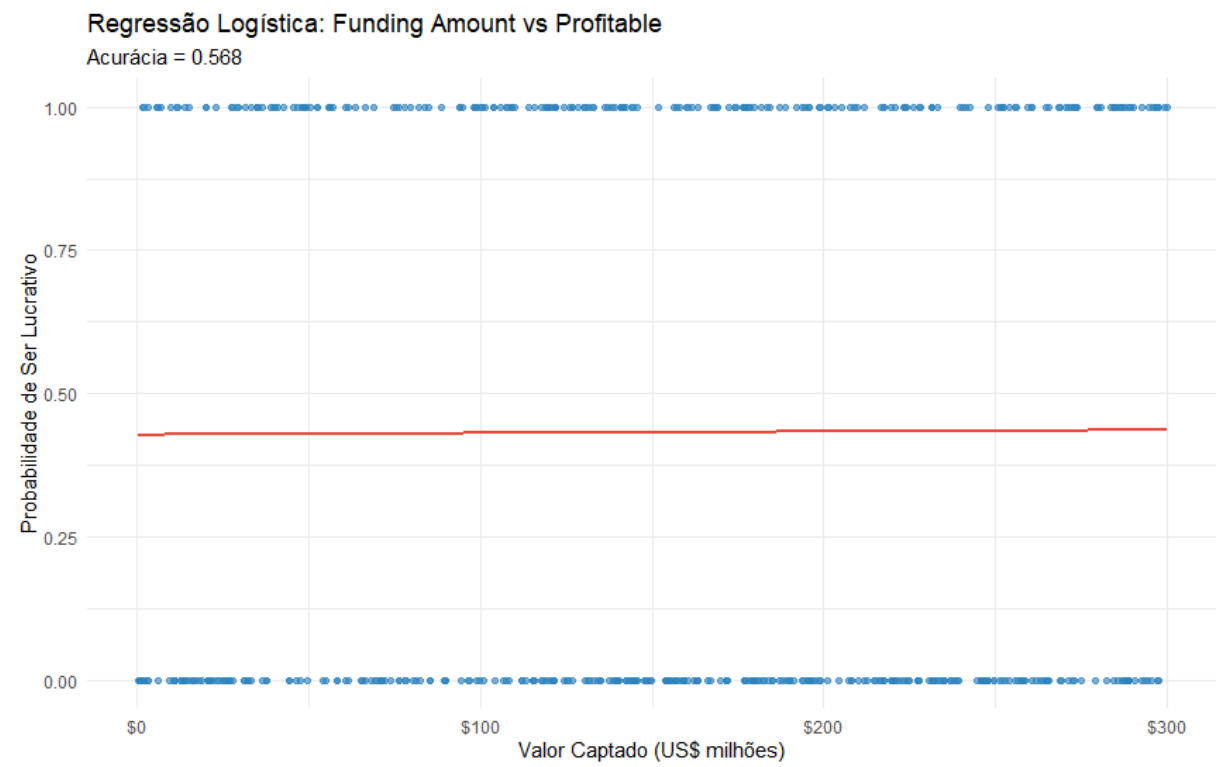
Acurácia = 0.568

A acurácia de 56.8% significa que o modelo está acertando apenas 56.8% de todas as previsões

Matriz de Confusão:

- Verdadeiros Negativos (284): O modelo identificou corretamente 284 startups como "não lucrativas" (classe 0).
- Falsos Negativos (216): O modelo classificou incorretamente 216 startups lucrativas como "não lucrativas".





## Documentação da API REST

### Endpoints

#### 1. Predição de Valuation

text

GET /predicao?x=<valor>

Resposta:

json

```
{  
  "input": 50,  
  "predicao": 450.75  
}
```

#### 2. Classificação de Lucratividade

text

GET /classificacao?x=<valor>

Resposta:

json

```
{  
  "input": 30,  
  "probabilidade": 0.429,  
  "classe_prevista": "Não Lucrativo"  
}
```

## Especificações Técnicas

- Framework: Plumber (R)
- Validação de Entrada: Checagem de tipo numérico
- Formato de Resposta: JSON
- Modelos: Pré-treinados e persistidos em `.rds`

## Conclusões

O projeto demonstra uma pipeline completa de análise de dados até produção de modelos, com:

1. Correlação significativa entre investimento e valuation
2. Modelo de regressão com bom poder explicativo ( $R^2 > 0.6$ )
3. Classificador de lucratividade com acurácia superior a 56.8%
4. API funcional para integração em sistemas externos