
Tera

Aula #17

Regressão Linear – **Feature Engineering**

Contexto
ESTUDO DE CASO

- Fonte da base de dados : kaggle
- Link: https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings
- Notas de videogame compiladas pelo Metacritic
- Objetivo do Caso: Utilizar os conceitos aprendidos na aula e no e-learning para explicar a relação entre as vendas de jogos e a crítica de notas, tentar ajustar um modelo de regressão com uma variável explicativa.

Crítica especializada afeta a venda de produtos de entretenimento?

Pratica

Conclusões das Análises das Bases de Videogames??

Como tratar as variáveis

Conceito:

O que fazer com as variáveis categóricas?

Variáveis Ordinais - Existe uma relação de Ordem? (Classificação A, B, C e outras?)

R: Transformar em Números Ordenados

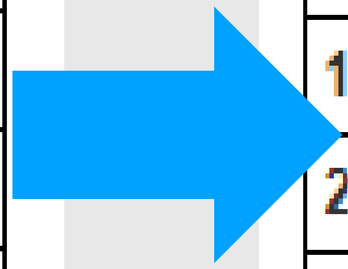
Variáveis Nominais – Sem relação de Ordem? (Sexo, Nomes e Cidades)

R: Transformar em Dummy Variables

Conceito:

O que fazer com as variáveis categóricas?

	first_name	last_name	sex
0	Jason	Miller	male
1	Molly	Jacobson	female
2	Tina	Ali	male
3	Jake	Milner	female
4	Amy	Cooze	female



Dummy

	first_name	last_name	sex	female	male
0	Jason	Miller	male	0.0	1.0
1	Molly	Jacobson	female	1.0	0.0
2	Tina	Ali	male	0.0	1.0
3	Jake	Milner	female	1.0	0.0
4	Amy	Cooze	female	1.0	0.0

Conceito:

Utilizar Variável Categórica no estudo?

	first_name	last_name	sex
0	Jason	Miller	male
1	Molly	Jacobson	female
2	Tina	Ali	male
3	Jake	Milner	female
4	Amy	Cooze	female

```
# Create a set of dummy variables from the sex variable
df_sex = pd.get_dummies(df['sex'])
```

```
# Join the dummy variables to the main dataframe
df_new = pd.concat([df, df_sex], axis=1)
df_new
```

	first_name	last_name	sex	female	male
0	Jason	Miller	male	0.0	1.0
1	Molly	Jacobson	female	1.0	0.0
2	Tina	Ali	male	0.0	1.0
3	Jake	Milner	female	1.0	0.0
4	Amy	Cooze	female	1.0	0.0

Conceito:

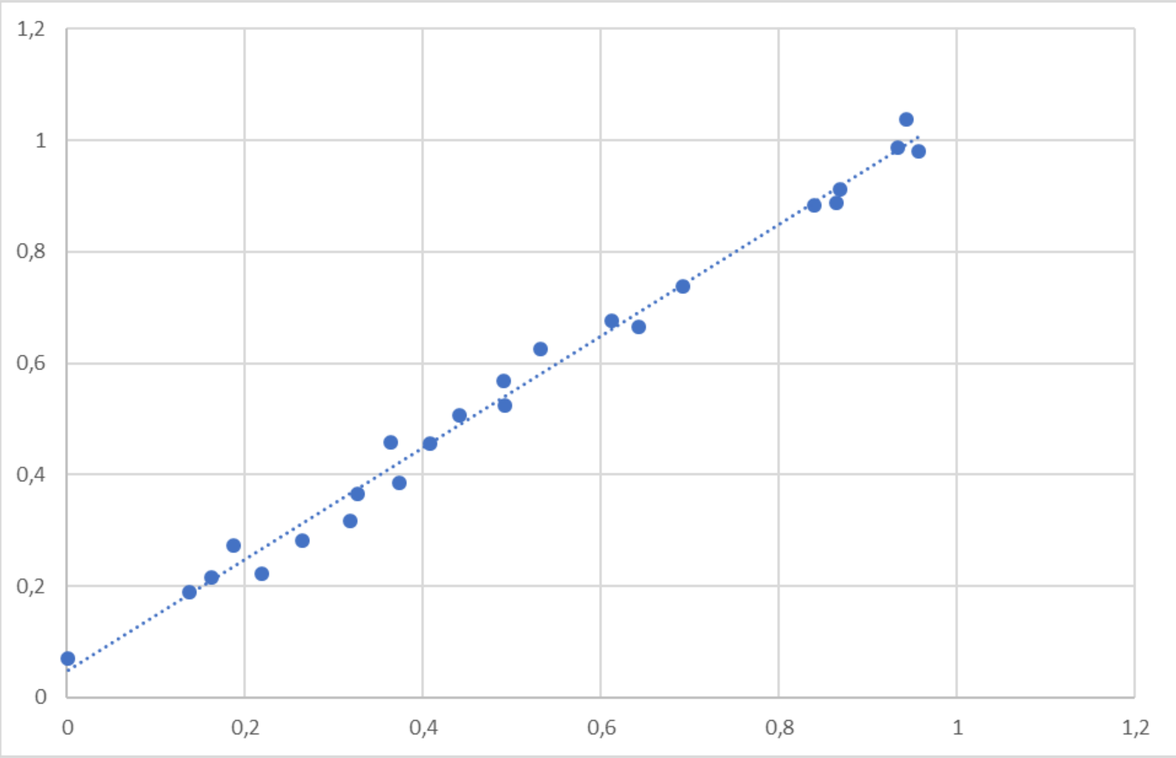
Utilizar Variável Missing no estudo?

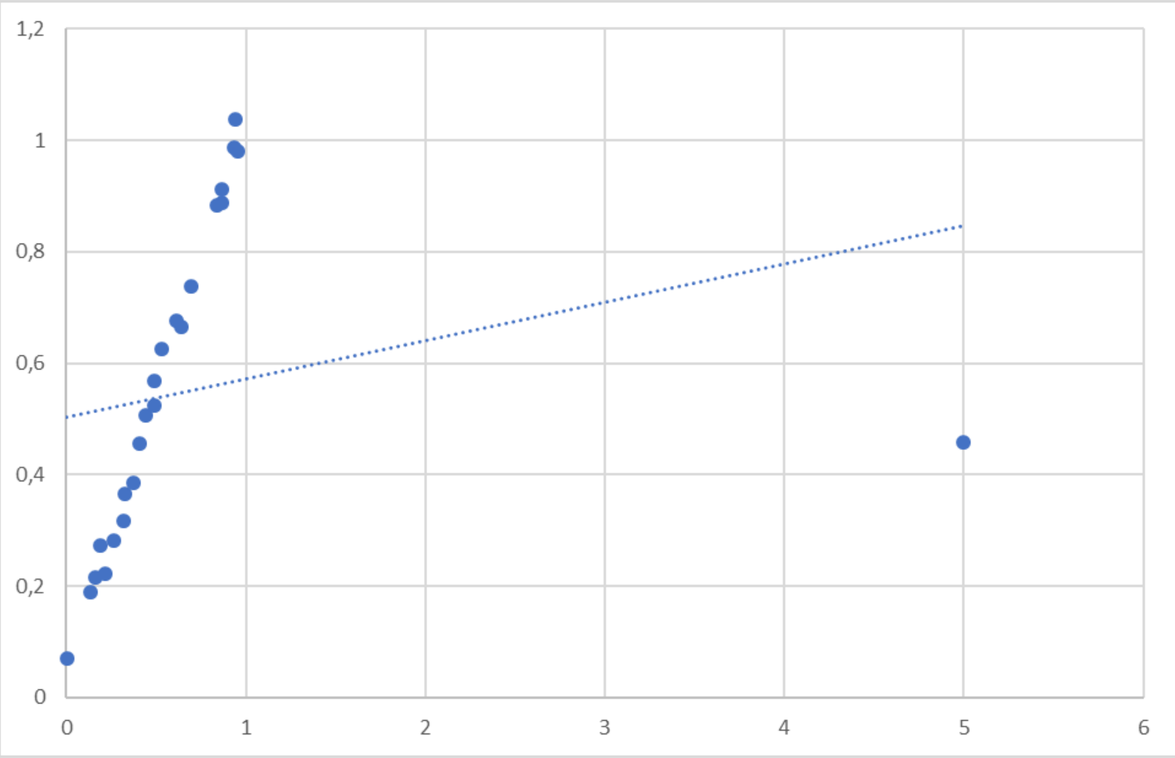
- **Criar Flag (dummy) de Missing**
- **Inputar com valor imparcial (moda e mediana)**
- **Criar Modelo para prever o Valor da Variável Missing Xi <-Modelo**

  (x1,x2,...xi-1,xi+1,...Xn)

Conceito:

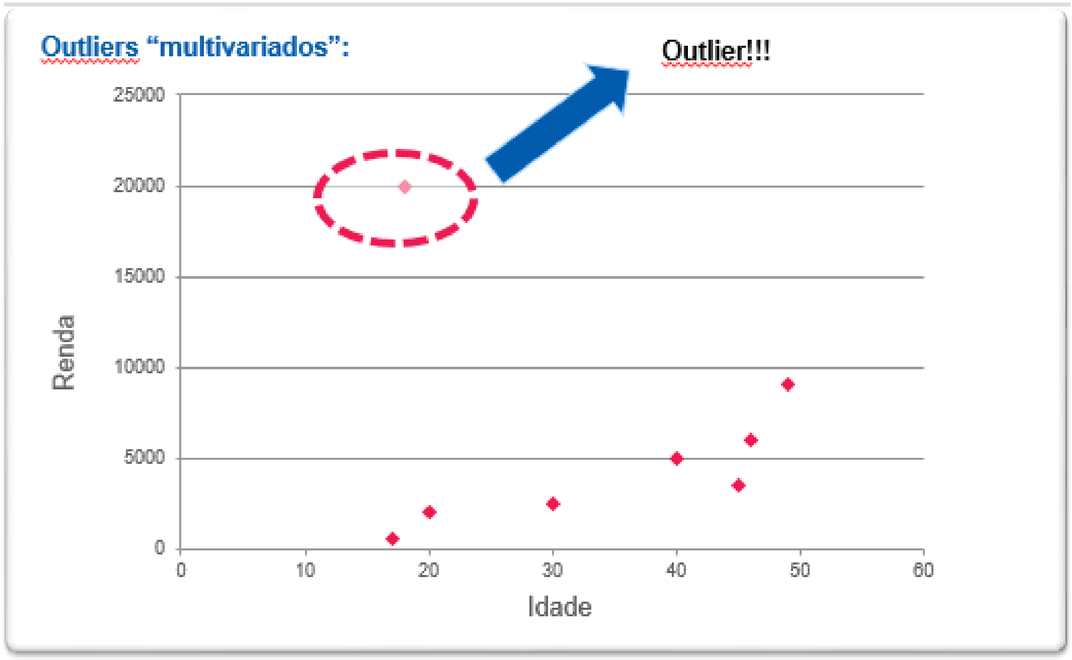
Outliers e Regressão Linear?

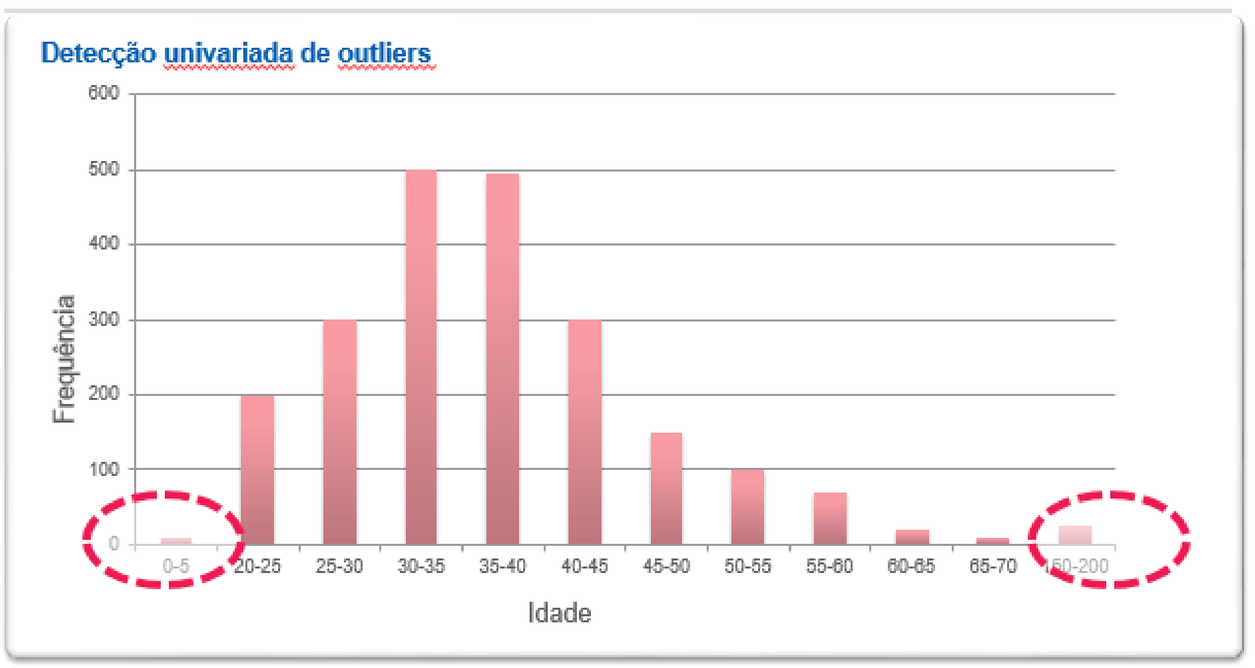




Conceito:

Outliers e Regressão Linear?



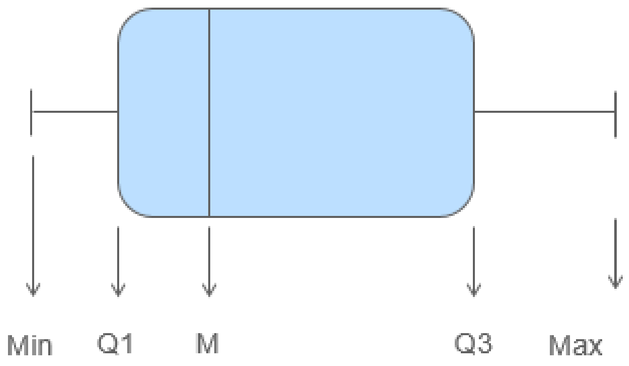


Conceito:

Outliers e Regressão Linear?

Detecção univariada de outliers

- Box Plot é uma representação visual de 5 números
  - Mediana
  - Quartil1 $P\ (X < Quartil1) = 0{,}25$
  - Quartil3 $P\ (X < Quartil3) = 0{,}75$
  - Mínimo
  - Máximo
  - IQR = Quartil 3 - Quartil1
  - Outliers > 1,5 IQR



Detecção multivariada de outliers

- Distância de Mahalanobis

$$D^2 = (\ x_i\ - vetor_{médias})^T \Sigma^{-1} (x_i - vetor_{médias})$$

- Métodos de Cluster : analisar elementos fora dos clusters
- Métodos de regressão : ajuste linerar e busca pelos maiores erros ou gráfico de residuos
- Conselho Prático : Foque mais em outliers univariados

Conceito:

Alternativas para colocar variáveis categóricas

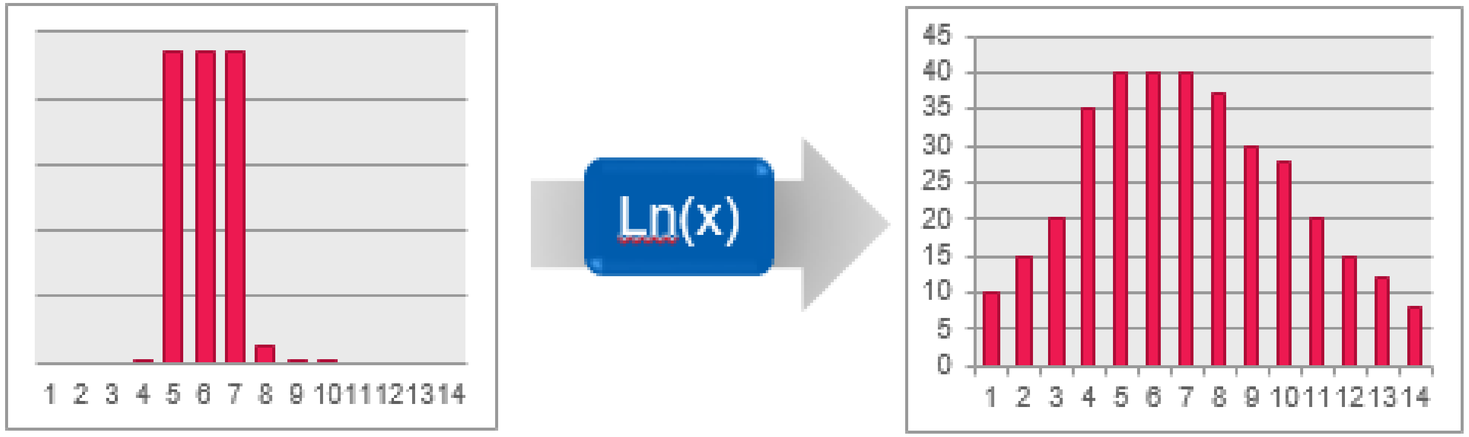
- **Label Encoding**
- **Atribuir um valor numérico para a variável categórica (válido para ordinais)**
- **Criar uma terceira variável que é a frequencia da categoria**
- **Criar novas categorias com maior poder preditivo/ correlacionadas com a variável resposta (comum em regressão logística), cada categoria se torna uma dummy ao final. Aumenta a perfomance do Modelo**

Conceito:

Alternativas para colocar variáveis numéricas

Transformação Log:

- Contribui na obtenção de uma distribuição mais simétrica e normal



- É comumente aplicado em variáveis financeiras, referentes a tamanho e etc
- É possivel utilizar outras transformações (potencia, raiz, exponencial e etc) lembrando que é importante preservar a INTUIÇÃO da variável

Conceito:

Alternativas para colocar variáveis numéricas

- **Variáveis numéricas não significativas para modelos lineares, categorizar de acordo otimizando a correlação entre as a variável resposta e variável explicativa**
- **Transformação LOG**
- **Padronização Z, MinMax (evita distorções em escala. Serve para “Clusterização”**

$$\boldsymbol{Exemplo}: \boldsymbol{X_{novo}} = (\boldsymbol{X_{antigo}} - \boldsymbol{media\ X\_antigo}) / \boldsymbol{Desvio\_padrao\_X}$$

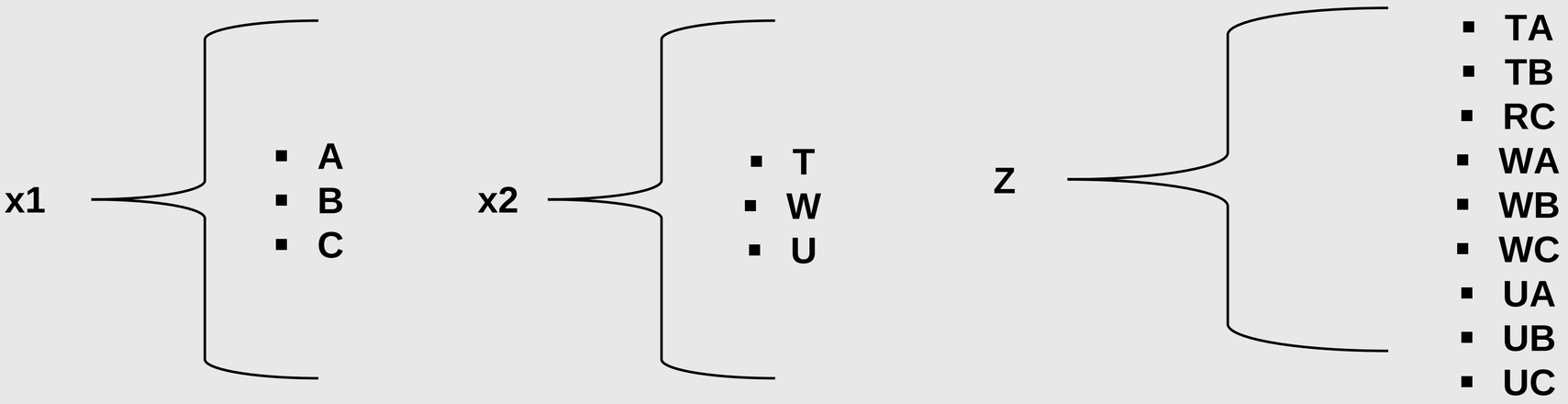
$$X_{new} = \frac{X_{old} - MIN(X_{old})}{\max(X_{old}) - \min(X_{old})}$$

Conceito:

Outros Tratamentos – Combinações de Variáveis

- **Z = X1+ X2**

- **Z = X1 x X2**



DÚVIDAS?!