# Tarefa Pré-processamento de Dados
# Aluno: Alex Halatiki Vicente

**Remoção de atributos irrelevantes:**

Id - atributo irrelevante
Name - atributo irrelevante
Host id - atributo irrelevante
Host name - atributo irrelevante
Neighbourhood - atributo complexo pois possui muitos valores categóricos
Country - atributo irrelevante pois todos são United States
Country code - atributo irrelevante pois todos são US
Host_name - atributo irrelevante
Last review - atributo irrelevante e complicado por ser uma data
House rules - atributo muito complexo com inúmeros valores categóricos
License - atributo irrelevante pois não possui nenhum valor em nenhuma linha

**A base não possui valores duplicados.**

**Lidando com valores nulos:**

Linhas em que as colunas Host_identity_verified, Neighbourhood group, Neighbourhood, Lat, Long, Instant_bookable, Cancellation_policy, Room type, Construction year, Number of reviews, Reviews per month, Review rate number e Price apresentam valores nulos são simplesmente removidas da base pois essas colunas podem ser relevantes e atribuir algum valor não é tão óbvio e diminui a confiabilidade dos dados.

Valores da coluna Minimum nights que são nulos podem ser interpretados como 1, representando uma noite mínima para reserva (diária) e as linhas em que essa coluna apresenta valor negativo são removidas, pois não faz sentido ter um valor negativo nessa coluna, o que representa um erro.

Valores da coluna Availability 365 que são nulos podem ser interpretados como 0, representando nenhuma disponibilidade de dias para os próximos 365, e valores negativos não fazem sentido. Portanto, as linhas com valores negativos nesta coluna são removidas.

Valores da coluna Calculated host listings count que são nulos podem ser interpretados como 1, já que o Host tem no mínimo o anúncio em questão na geografia da Cidade/Região. Valores negativos nesta coluna não estão presentes na base e não precisam ser tratados.

Valores da coluna Service fee que são nulos podem ser interpretados como 0, representando nenhuma taxa de serviço. Valores negativos nesta coluna não estão presentes na base e não precisam ser tratados.

**Padronização:**

A coluna Neighbourhood group possui os seguintes valores únicos: Brooklyn, Manhattan, brookln, Queens, Staten Island, Bronx. O valor brookln precisa ser mapeado para Brooklyn.

**One-Hot encoding:**

A coluna Host_identity_verified possui os seguintes valores únicos: unconfirmed, verified. Cada valor (categoria única) pode ser transformado em uma nova variável binária.

A coluna Neighbourhood group possui os seguintes valores únicos: Brooklyn, Manhattan, Queens, Staten Island, Bronx. Cada valor (categoria única) pode ser transformado em uma nova variável binária.

A coluna Instant_bookable possui os seguintes valores únicos: False, True. Cada valor (categoria única) pode ser transformado em uma nova variável binária.

A coluna Cancellation_policy possui os seguintes valores únicos: strict, moderate, flexible. Cada valor (categoria única) pode ser transformado em uma nova variável binária.

A coluna Room type possui os seguintes valores únicos: Private room, Entire home/apt, Shared room. Cada valor (categoria única) pode ser transformado em uma nova variável binária.

**Outliers:**

A coluna Price possui valores como 1,06 e 1,018. Esses valores aparentam ser outliers gerados por um erro de entrada de dados, porém é possível interpretá-los como valores na casa dos milhares, sendo assim 1060,0 e 1018,0. Isto faz mais sentido para um número que representa o preço da diária.

A coluna Long possui valores como -73975,0. Valores assim aparentam ser outliers gerados por um erro de entrada de dados, porém é possível interpretá-los como valores na casa das dezenas, sendo assim -73,975. Isto faz mais sentido em relação aos demais valores.

**Código e base processada em:**

https://github.com/AlexHalatiki/IAR/tree/master/data_processing