Análise breve dos dados

Antes de continuar para uma análise mais aprofundada dos dados, é importante primeiro entender o contexto do problema e o caráter mais superficial dos dados.

Com a primeira visão dos dados, é possível notar algumas colunas que, aparentemente, não influenciam na decisão do modelo sobre o preço estimado. São elas: "*id"*, "*host_id"* e "*host_name"*. Esses dados são úteis apenas para gerenciamento no site de aluguéis.

Também é interessante notar que casas com possuem em seu *"nome"* referências a eventos famosos, com SuperBowl, ou eventos artísticos como "Film", "Event", "Photo" possuem uma média de preços muito superior.

Os dados da coluna "*bairro_group"* apresentam o que, inicialmente, parece a característica mais determinante no preço. Dos 48893 dados, apenas 373 pertencem ao grupo "Staten Island", o que pode indicar uma possível média de preços maior. Com o método *info(),* colunas que possuem alguns valores nulos são expostas, indicando a necessidade de um tratamento futuro para o treinamento. Com o método *describe(),* alguns indicativos estatísticos interessantes devem ser considerados: 75% dos valores têm um preço menor ou igual a 175.00, ao mesmo tempo que temos um valor máximo de 10000. Essa discrepância pode indicar a presença de *outliers (*amostras de dados que estão fora do comum em relação ao todo).

Análise de correlações

Uma correlação no dado indica uma interdependência em duas ou mais características. No caso desse projeto, é interessante olhar se existe essa interdependência em relação ao preço. Para isso, a matriz de correlação é muito útil.

```
price                             1.000000
disponibilidade_365               0.081833
bairro                            0.062057
calculado_host_listings_count     0.057472
bairro_group                      0.044246
minimo_noites                     0.042799
latitude                          0.033939
reviews_por_mes                  -0.030608
numero_de_reviews                -0.047954
longitude                        -0.150020
room_type                        -0.249351
Name: price, dtype: float64
```

Matriz de correlação

Curiosamente, a matriz indica que a característica "*bairro_group"* influencia muito pouco no preço (0.044246), sendo praticamente negligenciável. Temos que o atribuito que mais

se relaciona com o preço é "*room_type",* no caso, uma relação negativa. O mesmo para *longitude*.

No caso de "*room_type",* um encoding foi realizado. 'Entire home/apt', 'Private room', 'Shared room' agora estão em suas representações númericas 0, 1 e 2, respectivamente. A correlação indica que, quanto mais aumentamos essa categoria, menor o preço. Ou seja, imóveis em que o usuário terá proveito de toda a casa/apartamento são normalmente mais caros que aqueles que o usuário terá um quarto compartilhado.

Já para *longitude,* quanto mais aumentamos (nos deslocamos para leste), menor o preço do aluguel.

Feature Engineering

Nossos dados descrevem imóveis em Nova York. Um dos pontos mais famosos em Nova York é o Central Park. Imóveis localizados nas proximidades do Central Park tendem a ter preços mais altos.

Uma nova feature pode surgir imaginando um retângulo que cobre o Central Park e verificando se o imóvel está dentro desse retângulo. Chamarei essa nova feature de "*near_central_park",* e os resultados são:

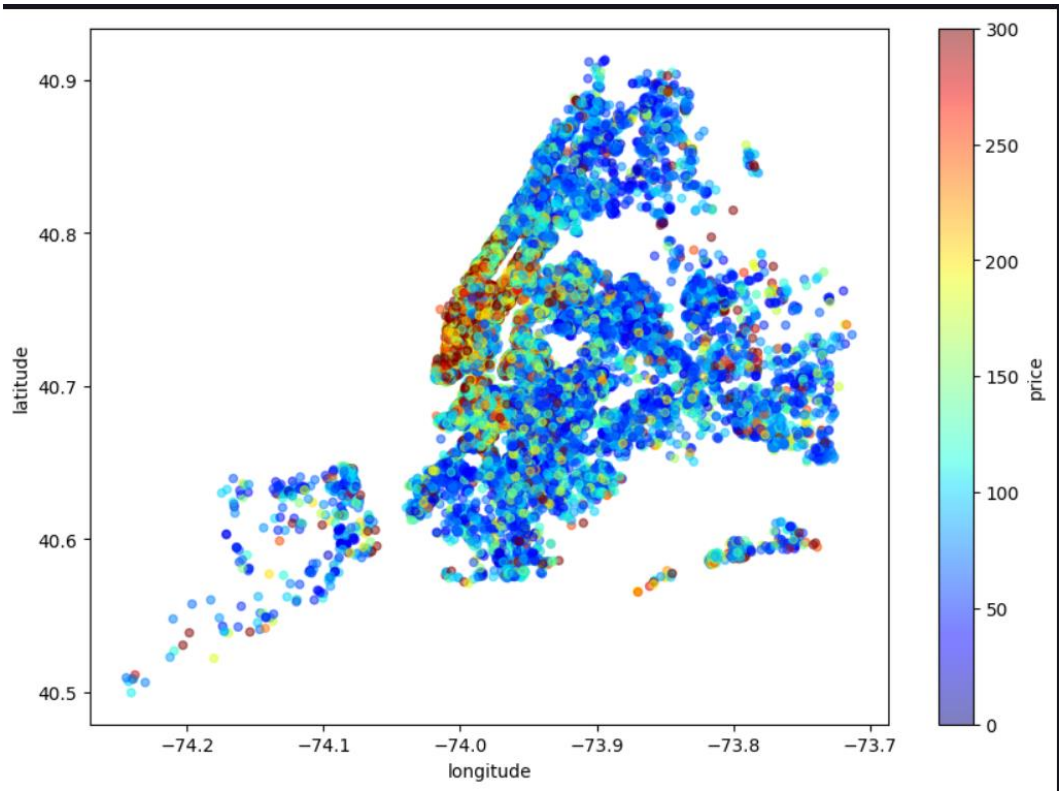
```
price                             1.000000
near_central_park                 0.093264
disponibilidade_365               0.081833
bairro                            0.062057
calculado_host_listings_count     0.057472
bairro_group                      0.044246
minimo_noites                     0.042799
latitude                          0.033939
reviews_por_mes                  -0.030608
numero_de_reviews                -0.047954
longitude                        -0.150020
room_type                        -0.249351
Name: price, dtype: float64
```

Entre as features que se relacionam positivamente com o preço, "*near_central_park*" tem sua relação maior que todas as outras!

Isso pode ser útil no treinamento do modelo.

Gráfico de dispersão

Por último, uma análise no gráfico de dispersão para analisar a relação da longitude e latitude com o preço.



Há uma concentração de aluguéis mais caros entre as latitudes 40.7 e 40.8 e longitude –74.0 e –73.9. Isso indica que Manhattan tem aluguéis mais caros, o que pode ser confirmado com os dados abaixo:

```
bairro_group
Bronx            87.496792
Brooklyn        124.381983
Manhattan       196.875814
Queens           99.517649
Staten Island   114.812332
Name: price, dtype: float64
```