
# Mini Projeto – Roof Imóveis

## 1. Objetivo do Negócio

Nossa equipe, contratada pela empresa Roof Imóveis, realizará uma consultoria estratégica, que por meio de uma análise crítica e aprofundada em negócios imobiliários, recomendará os melhores investimentos para expandir internacionalmente seus negócios.

Analisaremos os dados disponibilizados, colhendo as informações relevantes e, ao final, auferindo os *insights*, fazer um modelo que indicará as 5 (cinco) melhores opções de investimento com maior retorno financeiro.

Com o fim de evitar dispêndios financeiros não lucrativos, serão demonstrados ainda os 5 (cinco) imóveis com os piores indicadores e menor probabilidade de retornos financeiros.

Para o melhor entendimento e esclarecimento das indicações serão expostos com precisão as métricas utilizadas para o desenvolvimento dos modelos de análise e das referidas recomendações, ressaltando as passagens mais complexas e *insights* obtidos durante a resolução do projeto.

## 2. Business Understanding

## 2.1. Apresentação

A Roof Imóveis, uma das maiores empresas do ramo imobiliário brasileiro, deseja expandir sua área de atuação fazendo um investimento internacional.

Como etapa inicial do projeto de expansão desejado pela empresa, foi escolhido o Condado de King (King County), Estado de Washington, nos Estados Unidos, um dos 39 condados do estado americano de Washington.

Fundado em 1852, o Condado de King, em Washington, é também o mais densamente populoso e povoado do estado, além de ser o 12º mais populoso do país.

A sede e cidade mais populosa do condado é Seattle, entretanto, ao contrário dos demais condados, que normalmente abrigam uma cidade de grande porte onde a população costuma se concentrar, no Condado de King a população se divide pelas cidades vizinhas, em áreas urbanas, restando menos de um terço (29,5%), da população residente na cidade sede, Seattle, conforme Censo Nacional de 2020.

Destaca-se ainda o fato de que o Condado de King, em Washington, apresentou um crescimento populacional de 17,5% na última década, entre os anos de 2010 e 2020, 14,6% acima da média estadual no mesmo período, demonstrando uma forte tendência de crescimento no mesmo período a partir de 2020.

Num aspecto habitacional, houve um aumento de 13,9% na densidade populacional de 2020 em relação ao censo anterior. Em 2020, apenas 5,3% das unidades habitacionais se encontravam desocupadas com uma média de 2.3 pessoas por residência.

A renda média anual de uma residência ocupada é de $53.157,00( cinquenta e três mil cento e cinquenta e sete dólares) e a renda média anual de uma família é de $ 66.035,00(sessenta e seis mil e trinta e cinco dólares), dado que pode vir a ter relevância para análise da capacidade financeira média de prováveis compradores ou locatários dos imóveis em investimento pela Roof Imobiliária.

Para uma melhor concretude no resultado da presente análise e consultoria estratégica, é importante destacar que os dados buscados e colhidos não se confundem com o Condado de King no Texas, que diferentemente do seu homônimo, é o segundo menor condado em termos populacionais do Texas e o terceiro menos populoso do país.

Os dados disponibilizados contemplam as informações e características dos imóveis espalhados pelo condado.

## 2.2. Dados

Os dados utilizados foram extraídos das seguintes fontes:

https://www.kaggle.com/harlfoxem/housesalesprediction

https://geodacenter.github.io/data-and-lab/KingCounty-HouseSales2015/

Numa primeira análise é possível verificar que os dados apresentam as vendas de imóveis na região e suas características no período de 12(doze) meses, entre maio de 2014 a maio de 2015.

## 2.3. Dicionários dos dados

As features disponíveis no dataset apresentam um total de 21 atributos e 21613 registros.:

- Id – Identificador único do imóvel
- Date – Data de venda
- Price – Preço de venda

- Bedrooms – Número de quartos
- Bathrooms – Números de banheiros
- Sqft_liv – Tamanho da área habitável em $ft^2$
- Sqft_lot – Tamanho do terreno em $ft^2$
- Floors – Número de andares
- Waterfront – Indicativo se o imóvel é a beira-mar
- View – Grau de quão belo é a vista do imóvel (0 a 4)
- Condition – Condição da casa (1 a 5)
- Grade – Classificação por qualidade de material utilizado na construção
- Sqft_above – Área acima do solo em $ft^2$
- Sqft_basmt – Área abaixo do solo em $ft^2$
- Yr_built – Ano de Construção
- Yr_renov – Ano de restauração (Caso não haja restauração: 0)
- Zipcode – Zip Code 5 (similar ao CEP brasileiro)
- Lat – Latitude
- Long – Longitude
- Squft_liv15 – Média da área habitável dos 15 imóveis mais próximos em $ft^2$
- Squft_lot15 – Média da área do lote dos 15 imóveis mais próximos em $ft^2$

Além dos registros listados acima, foram criados os demais atributos para análise:

- Cities – Cidade na qual o imóvel está localizado
- Profit – Lucro obtido com a revenda do imóvel
- Sqft_lot_price_x – Valor médio do pé quadrado do lote
- Sqft_living_price – Valor médio do pé quadrado da área construída do lote
- Sqft_lot_price_y – Valor médio do pé quadrado da cidade

## 3. Coleta de Dados

```
[1]  # Instalando as Bibliotecas
     !pip install pandas
     !pip install numpy
     !pip install uszipcode
     !pip install matplotlib
     !pip install folium
     !pip instal seaborn
```

```
[2]  #Importando as Bibliotecas
     import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import uszipcode
     import seaborn as sns
     import plotly.express as px
     import folium
     from folium.plugins import HeatMap
```

```
[4]  #Carregando o DataSet
     url = "https://raw.githubusercontent.com/sostenesdev/analise-de-lucro-vendas-king/main/kc_house_data.csv"
     table = pd.read_csv(url)
     DF = pd.DataFrame(table)
```

```
[5]  # Verificando as informações do dataset
     DF.info()
```

```
[6]  # Verificando as dimensões do dataset
     DF.shape
```

## 4. Limpeza dos Dados

```
[7]  #Verificando se há dados faltantes
     DF.isnull().sum()
```

```
[8]  # Verificando se há valores duplicados
     DF.duplicated().sum()
```

```
[9]  # Validação de valores de elementos em tabela
     DF.nunique().sort_values()
```

## 5. Exploração dos Dados

- Quais serão as métricas essenciais utilizadas?
- Com base nos dados, quais imóveis deverão receber o investimento da Roof Imobiliária?

Dentre os atributos diretos, foram analisados primordialmente a data da venda, preço da venda, indicador único do imóvel, tamanho do terreno, localização, quantidade de vendas do mesmo imóvel.

Foram levados em consideração ainda, como forma de refino, os demais fatores relevantes, com alta probabilidade de influência no mercado imobiliário e que possam refletir diretamente na valorização e lucratividade do investimento como condições do imóvel, qualidade do material de construção e ano da última restauração.

As métricas iniciais definidas visam verificar a lucratividade e liquidez do imóvel resultando em negócios com maior probabilidade de retorno do investimento.

Para o aprofundamento da análise, os demais fatores indiretos foram analisados como forma de assegurar sua relevância ou irrelevância no exame dos *insights*.

Inicialmente, foi adicionada a coluna '*cities*' indicando a cidade de cada imóvel de acordo com seu código postal:

```
[10] #Adição da coluna 'Cities' com a cidade de cada imóvel

     cities = []
     from uszipcode import SearchEngine
     engine = SearchEngine()
     for x in table.zipcode.values:
       zipcode = engine.by_zipcode(x)
       cities.append(zipcode.major_city)
     DF['cities'] = cities
     DF.head()
```

Em seguida, para verificar o número de negócios em cada cidade e por consequência obter um *insight* acerca da liquidez, foram criados dois gráficos com o 'Número de Vendas por Cidade' e com o 'Número de Revendas por Cidade' respectivamente.

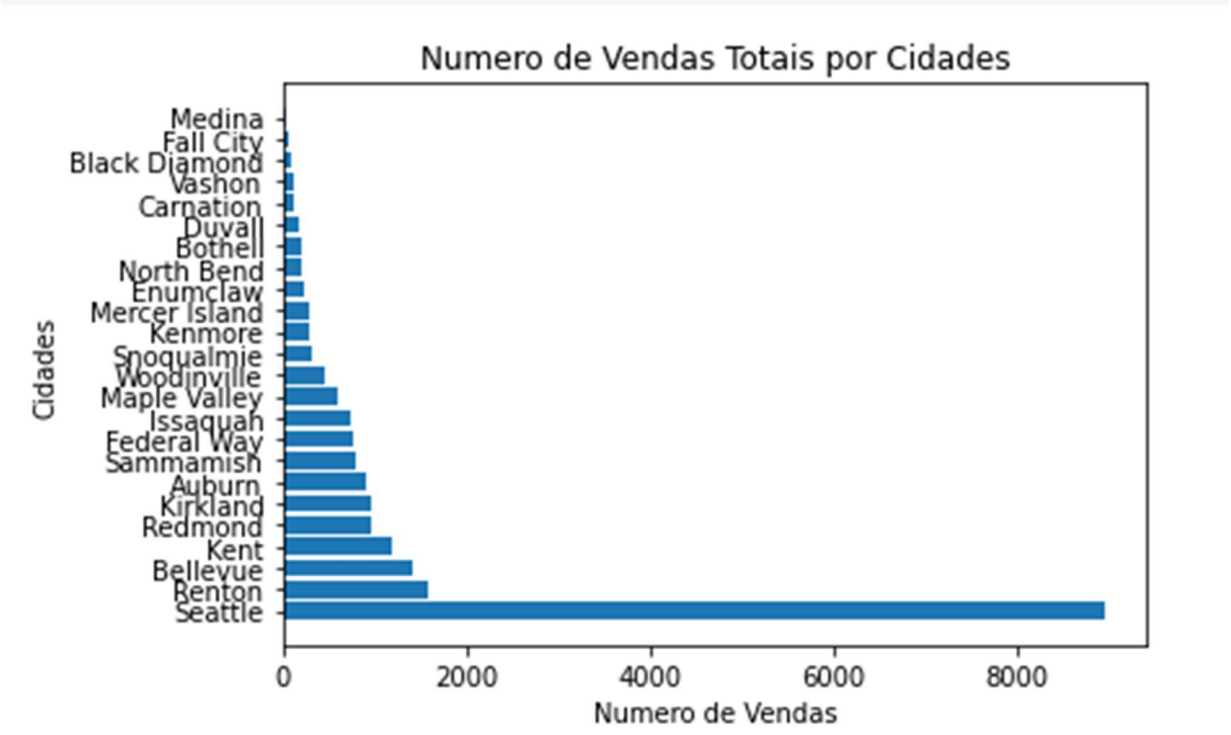
-

```
[11] #Verificação do Número de Vendas e Revendas por Cidade*

     counts = DF['id'].value_counts()
     DFresale=DF[DF.apply(lambda x: counts[x['id']] > 1.0, axis=1)]
     DFresale

     sale_number = DF['cities'].value_counts().rename_axis('Cities').reset_index(name='Numero de vendas')
     plt.barh(sale_number['Cities'].tolist() ,sale_number['Numero de vendas'].tolist())
     plt.ylabel("Cidades")
     plt.xlabel("Numero de Vendas")
     plt.title("Numero de Vendas Totais por Cidades")
     plt.show()

     resale_number = DFresale['cities'].value_counts().rename_axis('Cities').reset_index(name='Numero de vendas')
     plt.barh(resale_number['Cities'].tolist() ,resale_number['Numero de vendas'].tolist())
     plt.ylabel("Cidades")
     plt.xlabel("Numero de Vendas")
     plt.title("Numero de Revendas por Cidades")
     plt.show()
```




Numa breve análise dos dados expostos pelo gráfico supra, é perceptível o destaque que a cidade de Seattle ocupa, enquanto sede e cidade mais populosa do Condado, em comparação com as demais cidades disponíveis para negócio. O destaque no número de vendas e revendas demonstra um aumento na liquidez quista, sendo esse um dos principais parâmetros iniciais, junto a lucratividade, para determinar os imóveis mais indicados para investimento.

Seguindo com a análise, tendo em vista o objetivo principal do negócio, foram investigados os percentuais de lucratividade dentre os imóveis que foram revendidos. A lucratividade, neste caso, levando em consideração os dados disponíveis, é refletida na valorização histórica do imóvel pela comparação entre os valores de venda registrados num mesmo imóvel (id).

```
[12] #Verificação da lucratividade dos imóveis Revendidos

     grouped = DFresale.groupby('id')

     outputlist = []
     for local_id, group in grouped :
         #obtendo as linhas de cada imóvel ordenados por data da mais nova para mais antiga
         # (importante para cálculo do lucro)
         local = DFresale[DFresale["id"] == local_id].sort_values(by="date", ascending=False).copy()
         profit = 0
         #calculando o lucro de cada imóvel
         profit = local.iloc[0]['price']*100/local.iloc[-1]['price'] -100
         #adicionando resultado do lucro em uma lista
         outputlist.append([local_id , local.iloc[0]['cities'],local.iloc[0]['condition'],
                            local.iloc[0]['yr_renovated'], profit , local.iloc[0]['lat'],local.iloc[0]['long']])
     profit_df = pd.DataFrame(outputlist, columns = ['id','cities' ,'condition','yr_renovated','profit','lat', 'long'])
     output = []
     profit_df.sort_values(by="profit", ascending=False)
```

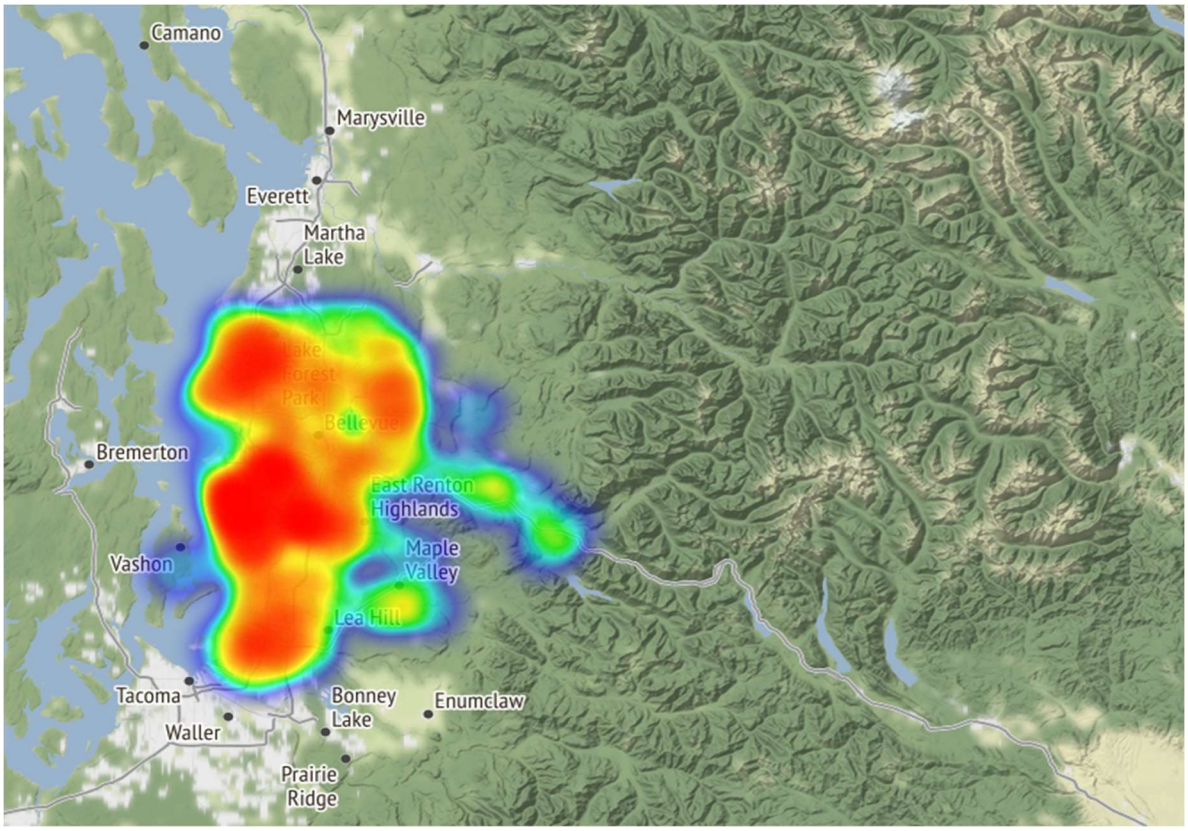
Para uma melhor visualização da concentração de negócios, baseados na lucratividade, foi executado o seguinte mapa de calor, onde as cores mais quentes(Vermelho, Laranja e Amarelo) representam um maior aglutinamento de negócios lucrativos:

```
[13]
      #heatmap do lucro por imóvel

      m2 = folium.Map(location=[47.58766, -121.83072], zoom_start=9,tiles='Stamen Terrain')

      lat = profit_df.lat.tolist()
      lng = profit_df.long.tolist()
      profit = profit_df.profit.tolist()

      HeatMap(list(zip(lat, lng, profit))).add_to(m2)
      m2
```




Somado aos *insights* previamente expostos, o mapa de calor demonstra claramente que os negócios mais lucrativos se concentram primordialmente na cidade de Seattle e seus arredores.

Com objetivo de refinar ainda mais a presente análise, foram determinados os valores médios do $FT^2$ por propriedade, originando uma nova variável de comparação.

-

```
[14] #Verificação do valor medio do FT² por propriedade.

     grouped = DFresale.groupby('id')

     outputlist = []
     for local_id, group in grouped :
         #obtendo as linhas de cada imóvel ordenados por data da mais nova para mais antiga (importante para calculo do lucro)
         local = DFresale[DFresale["id"] == local_id].sort_values(by="date", ascending=False).copy()
         profit = 0
         #calculando o lucro de cada imovel
         profit = local.iloc[0]['price']*100/local.iloc[-1]['price'] -100
         #valor pé quadrado terreno
         sqft_lot_price = local['price'].mean()/local.iloc[0]['sqft_lot']
         #valor pé quadrado area construida
         sqft_living_price = local['price'].mean()/local.iloc[0]['sqft_living']
         #adicionando resultado do lucro em uma lista
         outputlist.append([local_id, profit, sqft_lot_price,sqft_living_price, local.iloc[0]['lat'], local.iloc[0]['long']])
     profit_df2 = pd.DataFrame(outputlist, columns = ['id', 'profit','sqft_living_price','sqft_lot_price','lat', 'long'])
     df_com_profit = pd.merge(DFresale,profit_df2,on='id',how='left')
     grouped_city_ft_mean = df_com_profit.drop(['bedrooms','bathrooms','waterfront','view'],axis=1).sort_values(by=["profit","condition"], ascending=[False,False]).groupby('cities')
     city_ft_mean_lot = grouped_city_ft_mean['sqft_lot_price'].mean()
     city_ft_mean_living = grouped_city_ft_mean['sqft_living_price'].mean()
```

Em sequência, foram definidos ainda os valores médios do FT² divido por cidade, resultando em mais uma variável para comparações.

```
[15] #Verificação do valor médio do FT² por cidade.
     #preço do pé quadrado por cidade (lote)
     df_city_ft_mean_lot = pd.DataFrame({'cities': city_ft_mean_lot.index, 'sqft_lot_price': city_ft_mean_lot.values})
     #preço do pé quadrado por cidade (área construída)
     df_city_ft_mean_living = pd.DataFrame({'cities': city_ft_mean_living.index, 'sqft_living_price': city_ft_mean_living.values})
```

Prosseguindo com a análise, foram relacionados os reflexos da Condição do imóvel no valor de venda.

```
[16] #Gráfico de condição x preço (condição mais vendida)
     fig  = px.scatter(DF, x = "condition", y = 'price', hover_name = "id", log_x = True, width = 800)
     fig.update_traces(marker = dict(size = 20, line=dict(width = 2)), selector = dict(mode = 'markers'))
     fig.update_layout(title = 'Condições da casa X Preço de venda')
     fig.update_xaxes(title = 'Condições da casa')
     fig.update_yaxes(title = 'Preço de venda')
     fig.show()
```


As análises supramencionadas demonstram uma concentração de resultados nos imóveis de Condição 3(três) ou mais. O gráfico apresentado não deixa dúvidas acerca do destaque da Condição 3(três) que se mostra proporcionalmente com o maior número de negócios bem valorizados, mesmo em face de Condições melhores, como 4(quatro) ou 5(cinco).

Diante dos fatores analisados, foram filtrados os imóveis com condição melhor ou igual a 3(três), com o posterior cálculo da média de valor por FT² e desvio padrão.

-

```
[17] # Filtro para selecionar imóveis com a condição melhor ou igual 3
     df_com_profit2 = df_com_profit.drop(['waterfront','view','floors','grade','sqft_basement','yr_renovated','yr_built','sqft_above'],axis=1)
     df_com_profit2 = df_com_profit2[df_com_profit2['cities'] ==  'Seattle' ]
     df_com_profit2 = df_com_profit2[df_com_profit2['condition']>= 3 ]
     df_com_profit2
```

```
[18] #Cálculo de média e desvio padrão(STD - Standard Deviation)
     profit_df_media = pd.merge(df_com_profit2,df_city_ft_mean_lot,on='cities',how='left')
     profit_df_media[['sqft_lot_price_x']].describe().apply(lambda s: s.apply('{0:.2f}'.format))
     # a média de valor por FT² foi 257,87
     # o STD foi de 102,38
```

Em face dos resultados obtidos com os valores médios e o desvio padrão, foi possível a determinação dos valores de $156,00 (cento e cinquenta e seis dólares) a $360,00(trezentos e sessenta dólares) para filtrar os imóveis que se encontram neste *range*.

```
[19] #Explicando o desvio padrão
     #foram utilizados os valores de $ 156 até $ 360 por FT² para filtrar com base no desvio padrão
     profit_df_finish = profit_df_media[profit_df_media['sqft_lot_price_x']>= 156 ]
     profit_df_finish = profit_df_finish[profit_df_finish['sqft_lot_price_x']<= 360 ]
     profit_df_finish
```

Percebendo os resultados apresentados diante das métricas definidas, foi determinado o filtro para imóveis que se apresentassem fora dos valores de desvio padrão, assim como em Condição menor ou igual a 3.

-

```
[20] #Filtro por condição do imóvel menor ou igual a 3
     df_com_profit4 = df_com_profit[df_com_profit['condition']<= 3 ]

     #Calculo de média e desvio padrão
     profit_df_media1 = pd.merge(df_com_profit4,df_city_ft_mean_lot,on='cities',how='left')
     profit_df_media1[['sqft_lot_price_x']].describe().apply(lambda s: s.apply('{0:.2f}'.format))
     # a média de valor por FT² foi 238,88
     # o STD foi de 96.25
```

```
[21] #Explicando desvio padrão
     #foram utilizados os valores de $ 143 ate $ 335 por FT² para filtrar com base no desvio padrão
     profit_df_finish1 = profit_df_media1[(profit_df_media1['sqft_lot_price_x']<= 143) | (profit_df_media1['sqft_lot_price_x']>= 335) ]
```

Por fim, reunindo os valores extraídos, diante das métricas definidas, foram estabelecidos os filtros para exibição dos 5(cinco) imóveis indicados para investimento, assim como dos 5(cinco) imóveis não indicados para investimento.

-

```
[23] #Filtrando imóveis com pior margem dentro dos parâmetros estabelecidos
     grouped1 = profit_df_finish1.groupby('id')

     outputlist = []
     for local_id, group in grouped1 :
         local = profit_df_finish1[profit_df_finish1["id"] == local_id]
         outputlist.append([local_id,local.iloc[0]['profit'],local.iloc[0]['price'],local.iloc[0]['sqft_lot_price_x'],local.iloc[0]['condition'],local.iloc[0]['cities']])
     groupbyid_df = pd.DataFrame(outputlist, columns=['id','profit %','price','sqft_lot_price_x','condition', 'cities' ])
     groupbyid_df.sort_values(by="profit %", ascending=True).head(5)
```

```
[22] #filtrando imóveis com melhor margem dentro dos parâmetros estabelecidos
     grouped = profit_df_finish.groupby('id')

     outputlist = []
     for local_id, group in grouped :
         local = profit_df_finish[profit_df_finish["id"] == local_id]
         outputlist.append([local_id,local.iloc[0]['profit'],local.iloc[0]['price'],local.iloc[0]['sqft_lot_price_x'],local.iloc[0]['condition']])
     groupbyid_df = pd.DataFrame(outputlist, columns=['id','profit %','price','sqft_lot_price_x','condition' ])
     groupbyid_df.sort_values(by="profit %", ascending=False).apply(lambda s: s.apply('{0:.2f}'.format)).head(5)
```

As demais variáveis, tendo em vista o objetivo exclusivo de investimento e inferido lucro da Roof Imobiliária, não demonstraram suficiente relevância para a indicação dos imóveis nesta análise e consultoria estratégica.

## 6. Conclusão

Isto posto, diante da análise supra exposta, levando em consideração as métricas estabelecidas intimamente ligadas ao objetivo do negócio, é possível verificar que para verificação da Lucratividade e Liquidez dos investimentos, os principais atributos utilizados foram, não descartados os demais:

- Id – Identificador único do imóvel
- Date – Data de venda
- Price – Preço de venda
- Sqft_liv – Tamanho da área habitável em ft²
- Sqft_lot – Tamanho do terreno em ft²
- Condition – Condição da casa (1 a 5)
- Zipcode – Zip Code 5 (similar ao CEP brasileiro)

Apesar de outros imóveis analisados apresentarem características semelhantes, é possível observar, levando em consideração as métricas definidas, que os imóveis filtrados apontam destaque quanto aos atributos que resultam numa maior liquidez e lucratividade, fortalecendo e destacando a capacidade de negociação destes para investimento.

Desta forma, em face dos fatores mencionados, demonstrado o alto grau e probabilidade de retorno lucrativo do investimento, optamos pela indicação dos seguintes imóveis, em ordem de prioridade (id 7129304540, id 3883800011, id 1423049019, id 5132000140, id 6141100320):

| | id | profit % | price | sqft_lot_price_x | condition |
|---|---|---|---|---|---|
| **42** | 7129304540.00 | 230.83 | 133000.00 | 200.35 | 3.00 |
| **27** | 3883800011.00 | 168.17 | 82000.00 | 175.52 | 3.00 |
| **10** | 1423049019.00 | 144.44 | 90000.00 | 267.24 | 3.00 |
| **33** | 5132000140.00 | 137.14 | 175000.00 | 215.33 | 3.00 |
| **38** | 6141100320.00 | 132.65 | 245000.00 | 271.67 | 3.00 |

Ao passo que nas mesmas métricas, utilizando a lógica estritamente inversa, é possível verificar que os imóveis menos indicados para investimento, com uma razoável probabilidade de prejuízo são (id 2726049071, id 2767603612, id 1000102, id 8129700644, id 5332200530):

| | id | profit % | price | sqft_lot_price_x | condition | cities |
|---|---|---|---|---|---|---|
| **9** | 2726049071 | -3.931373 | 510000.0 | 609.725610 | 3 | Seattle |
| **11** | 2767603612 | -2.200000 | 500000.0 | 383.333333 | 3 | Seattle |
| **0** | 1000102 | 7.142857 | 280000.0 | 120.833333 | 3 | Auburn |
| **25** | 8129700644 | 9.090909 | 715000.0 | 359.375000 | 3 | Seattle |
| **18** | 5332200530 | 11.538462 | 910000.0 | 409.574468 | 3 | Seattle |

Os imóveis menos indicados para investimento, apesar de conterem características e vizinhança similares aos imóveis sugeridos para investimento, apresentam uma desvalorização contínua, indicando fatores atípicos inapropriados para investimento.

( LINK NOTEBOOK COLAB - https://colab.research.google.com/drive/1bmmEnrZqDOAwMEEGNisuQHwMjS4ImiRv?usp=sharing )