

Datascraping e EDA de Imóveis em São Paulo/SP

Agenda

1. Introdução
2. Objetivos
3. Coleta de Dados
4. Limpeza dos Dados
5. Feature Engineering e EDA
6. Conclusões

1. Introdução

Nesse post vamos fazer uma coleta de dados de imóveis, mais precisamente, apartamentos da cidade de São Paulo e uma análise completa deles. Os dados foram extraídos no dia 22/03/2023 do [site](#) da VivaReal, um site de aluguel/venda de imóveis. Ressaltando aqui que os dados foram utilizados exclusivamente para essa análise e **sem nenhum fim financeiro**. Foram coletados mais de 86 mil dados de apartamentos na cidade de São Paulo, dentre eles, bairro, tamanho do apto em m², número de quartos e valor do aluguel. Essas informações serão usadas para verificar quais regiões da cidade possuem os maiores valores de aluguel, maiores e menores apartamentos, quais são os bairros mais caros de se morar e etc. Toda a coleta foi feita por meio de webscraping na linguagem de programação Python.

2. Objetivos

Os principais objetivos dessa análise são:

- Coletar os dados por meio de webscraping
- Qual a zona da cidade com mais aptos para alugar
- Valor médio de tamanho e aluguel de cada zona e bairros da Cidade de São Paulo
- Quais são os bairros mais caros e baratos para alugar um apartamento
- Se há uma relação entre tamanho do apto, valor do aluguel e bairro

3. Coleta de Dados

Os dados foram coletados por meio da biblioteca BeautifulSoup. Por meio dela, foi possível obter o bairro do apartamento, o seu tamanho em metros quadrados, o número de quartos e o valor do aluguel. Como os dados estão dispostos em várias páginas, foi necessário iterar sobre cada uma. Foram coletados dados de mais de 86 mil apartamentos.

Figura 1 – Tabela dos dados coletado

	bairro	tamanho_m2	n_quartos	valor_aluguel
0	Rua Ytaipu - Mirandópolis, São Paulo - SP	75	2	R\$ 4.000 /Mês
1	Rua do Cristal - Vila Mariana, São Paulo - SP	65	2	R\$ 4.000 /Mês
2	Cambuci, São Paulo - SP	54	2	R\$ 2.550 /Mês
3	Rua da Independência, 706 - Cambuci, São Paulo...	98	3	R\$ 5.000 /Mês
4	Jardim Jaú, São Paulo - SP	60	2	R\$ 1.500 /Mês
...
86683	Avenida Miguel Yunes, 540 - Usina Piratininga,...	70	3	R\$ 3.200 /Mês
86684	Vila Nova Manchester, São Paulo - SP	78	3	R\$ 2.500 /Mês
86685	Rua Melo Alves, 265 - Cerqueira César, São Pau...	120	2	R\$ 4.000 /Mês
86686	Rua Coronel Cabrita - Jardim da Glória, São Pa...	124	3	R\$ 6.700 /Mês
86687	Rua Adalberto Kemeny, 82 - Parque Industrial T...	53	1	R\$ 3.800 /Mês

Com isso, temos na figura 1, os dados numa 'dataframe'. Podemos ver o bairro, tamanho e etc.

4. Limpeza dos dados

Para podermos trabalhar com os dados, primeiramente precisamos limpá-los. Na figura 1, vemos que 'bairro' está com mais informações além do bairro, como a rua e 'valor_aluguel' está com formatado com 'R\$', '.' e '/Mês'. Além disso, o datatype das colunas estão como object e não como int ou float. Vamos transformar os dados numéricos pra float para essa análise.

Figura 2 – Tabela limpa

	bairro	tamanho_m2	n_quartos	valor_aluguel
0	Mirandópolis	75	2	4000
1	Vila Mariana	65	2	4000
2	Cambuci	54	2	2550
3	Cambuci	98	3	5000
4	Jardim Jaú	60	2	1500
...
86683	Usina Piratininga	70	3	3200
86684	Vila Nova Manchester	78	3	2500
86685	Cerqueira César	120	2	4000
86686	Jardim da Glória	124	3	6700
86687	Parque Industrial Tomas Edson	53	1	3800

Podemos ver o resultado na figura 2. Temos apenas as informações necessárias e os datatypes estão corretos. Por fim, vamos salvar essa df no formato parquet para, além de outras coisas, preservar o dtype das colunas.

5. Feature Engineering e EDA

Com os dados obtidos na figura 2, a única feature nova que vamos incluir é a zona que o bairro se encontra, pois, essa informação vai ajudar na análise.

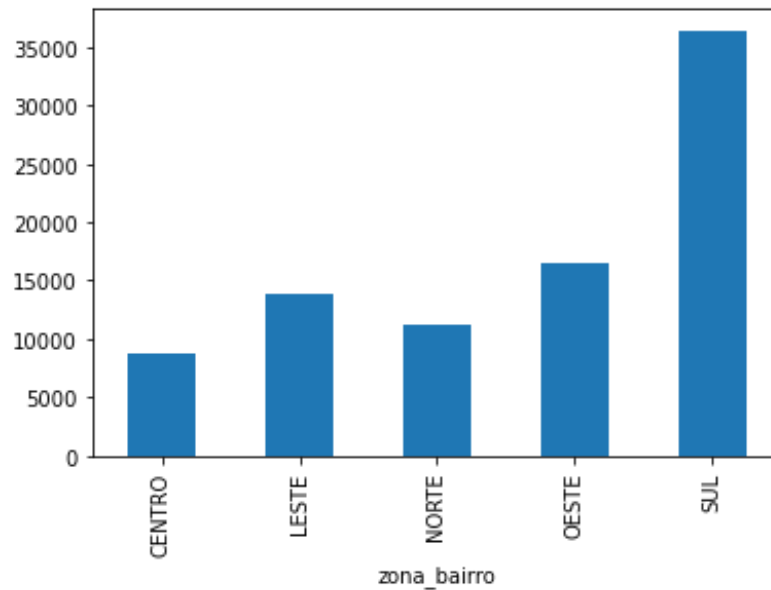
Figura 3 – Tabela EDA

	bairro	tamanho_m2	n_quartos	valor_aluguel	zona_bairro
0	Mirandópolis	75.0	2.0	4000.0	SUL
1	Vila Mariana	65.0	2.0	4000.0	SUL
2	Cambuci	54.0	2.0	2550.0	CENTRO
3	Cambuci	98.0	3.0	5000.0	CENTRO
4	Jardim Jaú	60.0	2.0	1500.0	LESTE
...
86683	Usina Piratininga	70.0	3.0	3200.0	SUL
86684	Vila Nova Manchester	78.0	3.0	2500.0	LESTE
86685	Cerqueira César	120.0	2.0	4000.0	OESTE
86686	Jardim da Glória	124.0	3.0	6700.0	SUL
86687	Parque Industrial Tomas Edson	53.0	1.0	3800.0	OESTE

Podemos ver a nova feature, 'zona_bairro', na figura 3.

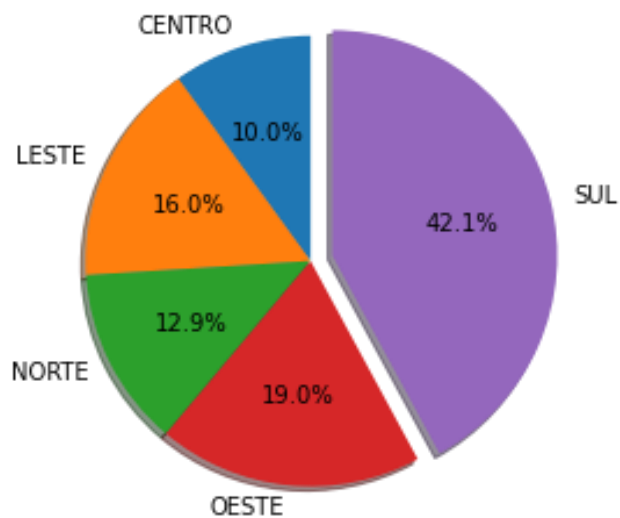
Agora podemos ir para a EDA.

Figura 4 – Quantidade de aptos por zona



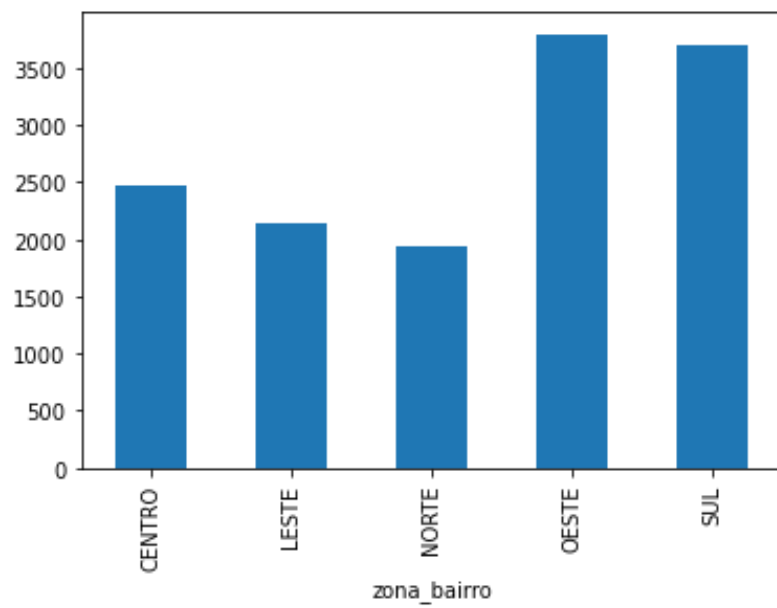
Na figura 4, vemos que a zona sul é a que possui mais apartamentos anunciados para aluguel o centro com a menor quantidade. Zona oeste e leste possuem quase o mesmo número de aptos anunciados.

Figura 5 – Regiões com mais vagas



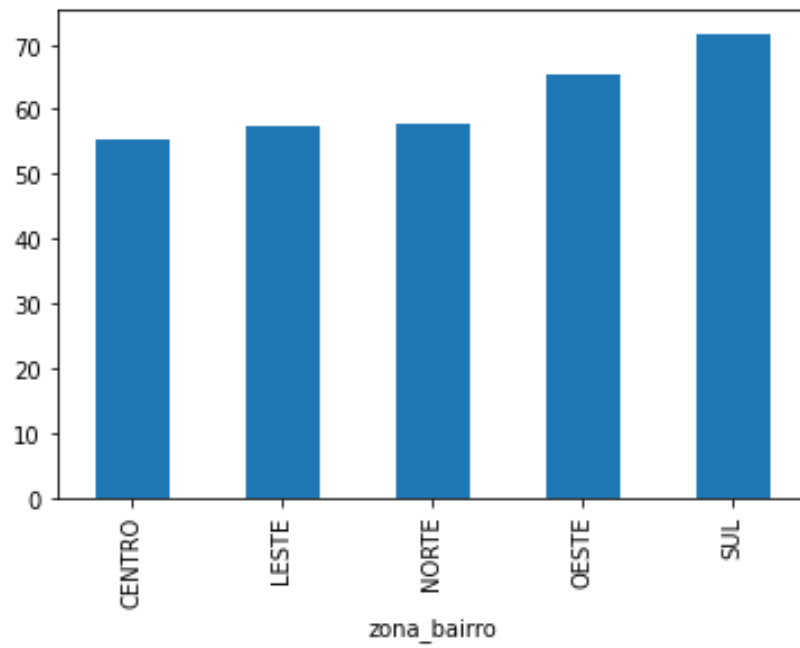
Temos que a zona sul possui 42.1 % dos apartamentos anunciados, figura 5, ou seja, mais que todos os aptos das zonas norte, leste e centro juntas.

Figura 6 – Valor médio do aluguel por zona



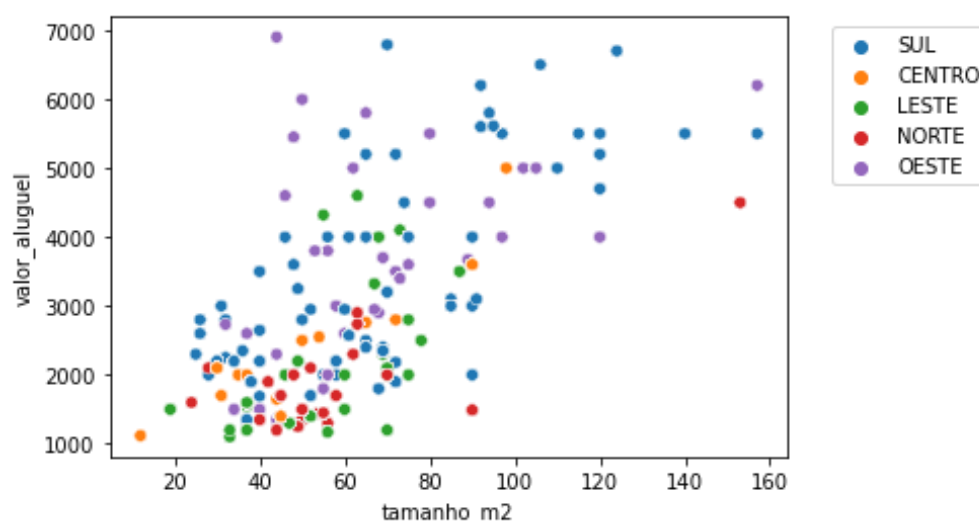
Na figura 6, temos que as 2 zonas que mais possuem apartamentos para aluguel são as mesmas que possuem as maiores médias de aluguel na cidade de São Paulo. Tanto a zona oeste como a zona sul possuem valores médios de aluguel maior que 3500 reais. A zona norte é a mais barata com um valor médio de, aproximadamente, 2000 reais.

Figura 7 – Tamanho médio do apto por zona



Como é visto na figura 7, a zona sul e oeste, como na figura 6, são as que possuem a maior média de tamanho de apartamento com 72 m² e 65 m², respectivamente, e, as outras 3 zonas, estão praticamente empatadas.

Figura 8 – Gráfico de dispersão de aluguel x tamanho



A figura 8 apresenta interessantes resultados, por exemplo, a zona oeste (pontos roxos) possui diversos apartamentos com pequenos ($\sim 45 \text{ m}^2$ - 60 m^2) com um valor de aluguel bem alto. Já na zona sul, temos pontos em todas as partes do gráfico, aptos pequenos com aluguel baixo, aptos grandes com aluguel caro e aptos de tamanho médio com valor médio. Na zona leste temos os pontos concentrados com metragem baixa ($\sim 20 \text{ m}^2$ - 65 m^2) com valores baixos de aluguel, assim como as zonas norte e centro da cidade.

Figura 9 – Boxplot de valor_aluguel

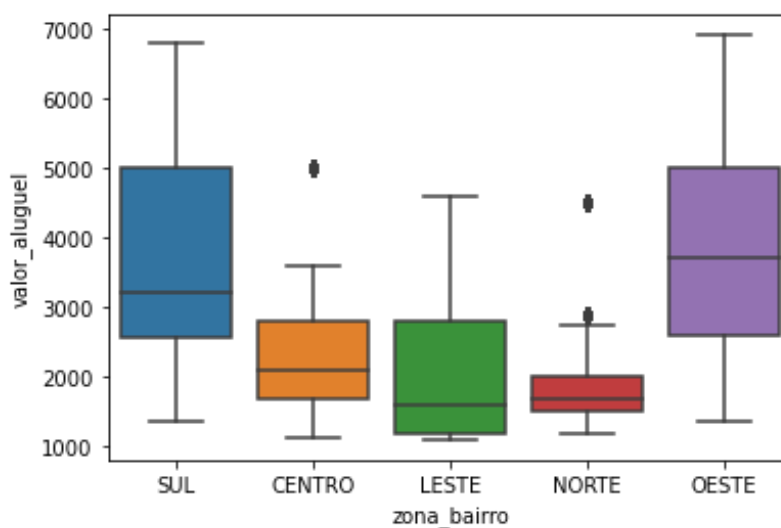
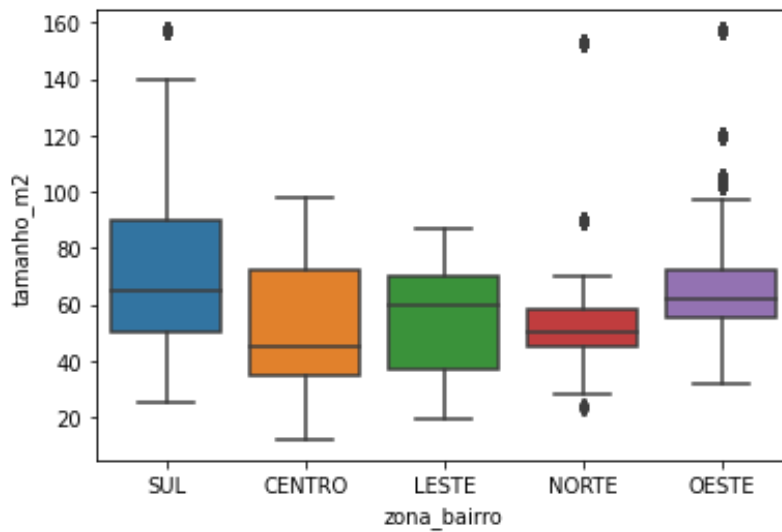
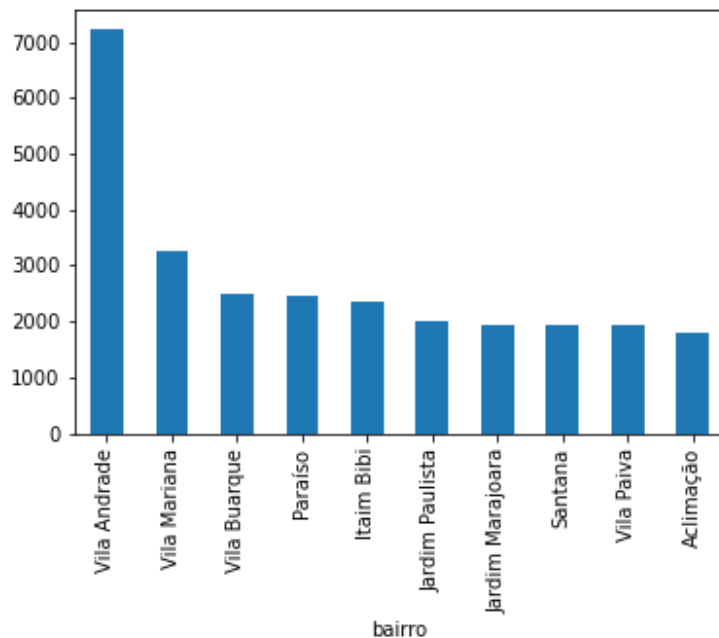


Figura 10 – Boxplot de tamanho_m2



Com as figuras 9 e 10, vemos que a zona sul e oeste possuem a maior amplitude de valores de aluguel, começando por 1500 até pouco mais de 7000 reais. A zona norte possui a menor variação de preços para aluguel. Em relação ao tamanho dos aptos, a zona oeste, mesmo com os aluguéis mais caros da cidade, possui metade dos aptos com aproximadamente 65 m². A zona sul possui a maior variação de tamanhos de apartamentos, de 30 m² até 140 m².

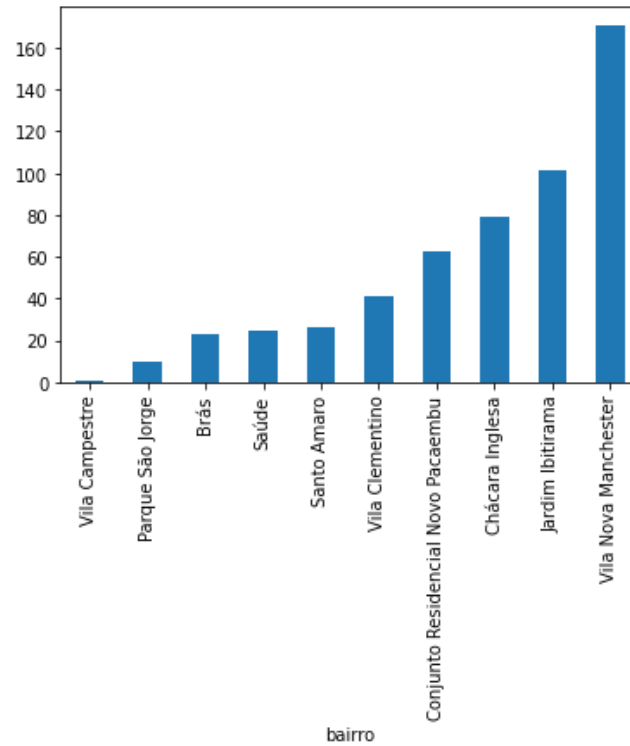
Figura 11 – Bairros com mais aptos anunciados



Com a análise das zonas da cidade feita, vamos fazer um estudo dos bairros de São Paulo.

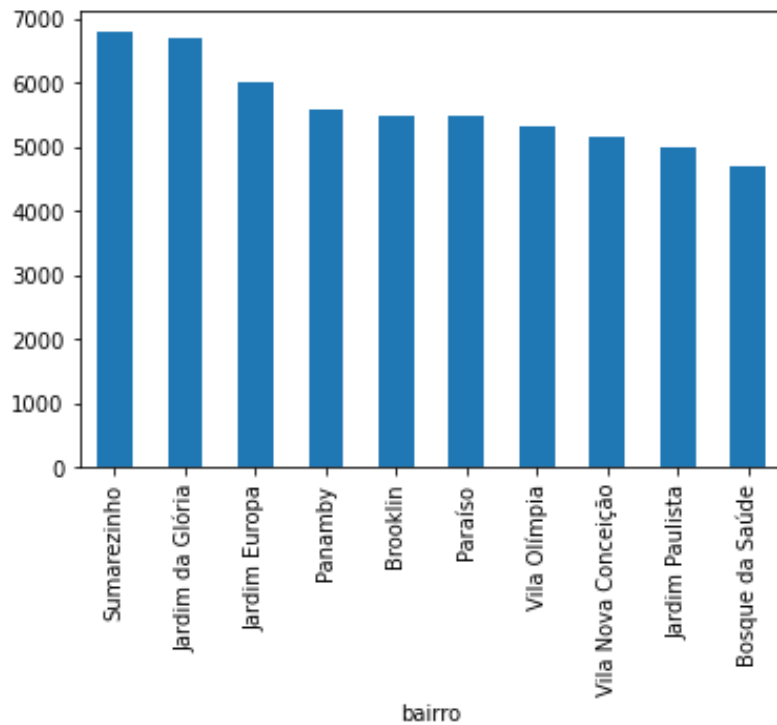
O bairro com mais aptos anunciados é a Vila Andrade (zona sul) com mais de 7000 anúncios, como visto na figura 11. Em seguida temos Vila Mariana e Vila Buarque, zonas sul e centro, respectivamente.

Figura 12 – Bairros com menos aptos anunciados



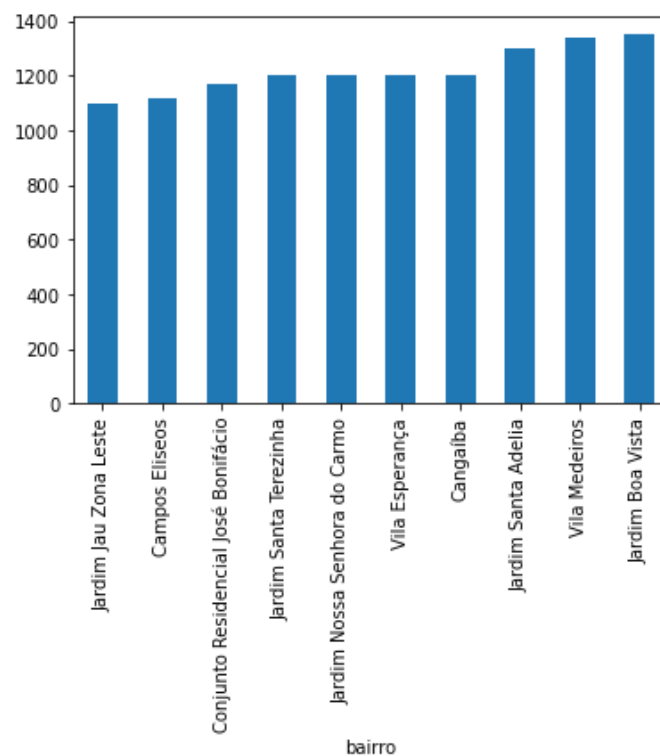
Por outro lado, na figura 12, temos que os bairros com a menor quantidade de anúncios são: Vila Campestre, Parque São Jorge e Brás.

Figura 13 – Bairros com maiores médias de aluguel



Na figura 13, vemos que os bairros mais caros de São Paulo são: Sumarezinho, Jardim da Glória e Jardim Europa. Todos com valores de mais de 6000 reais de aluguel. Os 2 primeiros ficam localizados na zona sul e o ultimo na zona oeste. Todo top 10 bairros mais caros ficam localizados nas zonas sul e oeste.

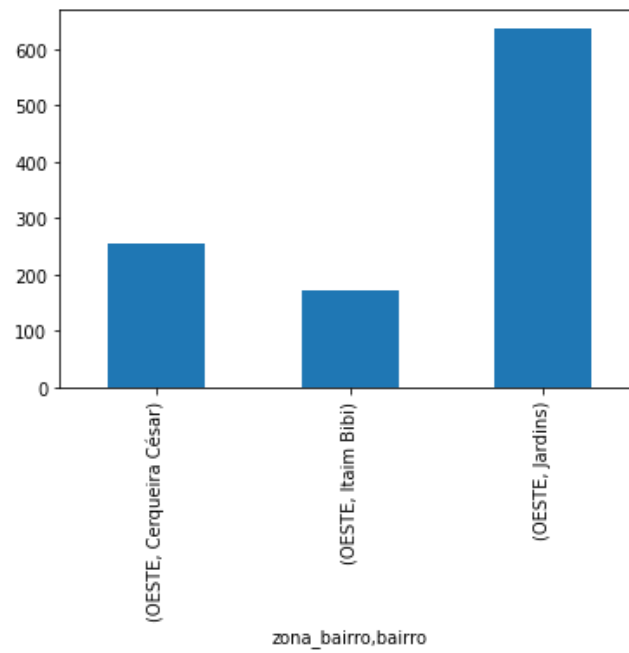
Figura 14 – Bairros com menos médias de aluguel



Por outro lado, na figura 14, temos os bairros mais baratos da cidade. São eles: Jardim Jau Zona Leste, Campos Eliseos e Conjunto Residencial José Bonifácio. A maioria está localizada na zona leste.

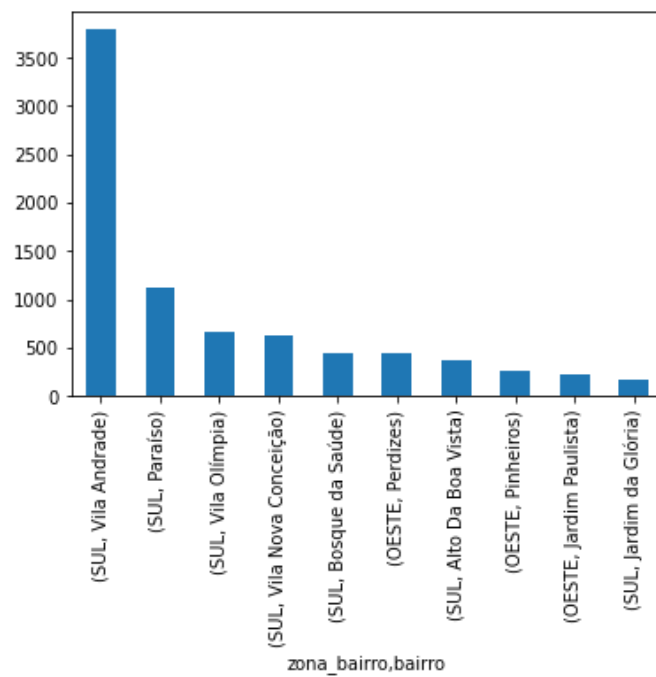
Foi feita uma análise para tamanhos e aluguéis mais específicos.

Figura 15 – Aptos com tamanho_m2 < 50 e aluguel > 4500



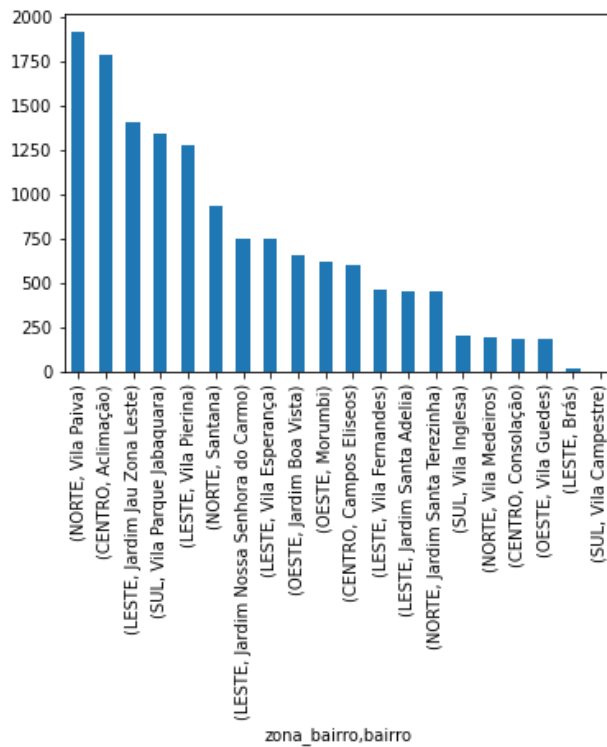
Todos os apartamentos listados com um tamanho menor que 50 m² e valor de aluguel maior que 4500, figura 15, encontram-se na zona oeste. Jardins é o bairro com mais apartamentos disponíveis.

Figura 16 – Bairros com tamanho_m2 > 100 e aluguel > 4500



Já para apartamentos com tamanho maior que 100 m² e aluguel maior que 4500 reais, praticamente todos se encontram na zona sul da cidade, como mostra a figura 16.

Figura 17 – Bairros com tamanho_m2 < 50 e aluguel < 2000



Por fim, para aptos com tamanho menor que 50 m² e aluguel menor que 2000, temos que quase todos se encontram ou na zona leste ou na zona norte da cidade, visto na figura 17.

Não há apartamentos maiores que 100 m² com aluguel menor que 4000 reais disponíveis na cidade.

6. Conclusões

Com os dados coletados usando webscraping, temos que a zona sul é a que possui mais apartamentos anunciados no site. A zona sul e oeste são as que possuem as maiores médias de valor de aluguel e tamanho. Os bairros mais caros são Sumarezinho e Jardim da Glória, ambos localizados na zona sul e os bairros mais baratos são Jardim Jau e Campos Eliseos, ambos na zona leste. Vemos que, por exemplo, na zona oeste, há apartamentos pequenos com valor de aluguel alto, mostrando que é zona altamente valorizada.