Relatório II - Tópicos Especiais em Computação

Marcos Antonio, Vandirleya Barbosa

[1]Universidade Federal de Piauí (UFPI)

marcos.brito, vandirleya.barbosa@ufpi.edu.br

Resumo. *Este trabalho apresenta uma análise realizada sobre o conjunto de dados Housing, amplamente utilizado no campo do aprendizado de máquina e originário do UCI Machine Learning Repository. O conjunto de dados contém 506 instâncias e 14 atributos, incluindo o valor mediano das casas ocupadas pelos proprietários. As variáveis incluem taxa de criminalidade, proporção de zonas residenciais, número médio de quartos por habitação, entre outros, utilizados para prever o valor das habitações. Portanto, este estudo propôs análises estatísticas, bem como a utilização de técnicas de pré-processamento e um resumo exploratório dos resultados obtidos, visando compreender melhor as relações entre as características das habitações e seu valor no mercado imobiliário.*

Abstract. *This paper presents an analysis conducted on the Housing dataset, widely used in the field of machine learning and originating from the UCI Machine Learning Repository. The dataset contains 506 instances and 14 attributes, including the median value of owner-occupied homes. Variables include crime rate, residential zone proportion, average number of rooms per dwelling, among others, used to predict housing values. Therefore, this study proposed statistical analyses, as well as the use of preprocessing techniques and an exploratory summary of the results obtained, aiming to better understand the relationships between housing characteristics and their value in the real estate market.*

1. Introdução

O mercado imobiliário é um setor complexo e dinâmico, onde a compreensão necessita da análise de diversas variáveis. Estas variáveis vão além das características físicas das propriedades, como área, número de quartos e banheiros, incluindo também fatores externos como localização geográfica, oferta e demanda, condições econômicas, e até mesmo características sociais e culturais. A interação entre essas variáveis pode ter um impacto significativo no valor das propriedades e nas tendências do mercado.

As tendências do mercado imobiliário são influenciadas por diversos fatores, como mudanças na oferta e demanda, taxas de juros, políticas governamentais e desenvolvimentos econômicos locais e globais. Por exemplo, áreas urbanas em crescimento podem ver um aumento na demanda por imóveis, levando a valorizações significativas. Da mesma forma, mudanças nas políticas fiscais ou econômicas podem afetar a confiança dos investidores e influenciar os preços das propriedades. Portanto, compreender essas tendências e

antecipar suas possíveis repercussões é essencial para tomar decisões informadas no mercado imobiliário, seja para compra, venda ou investimento em propriedades. A análise estatística e o uso de técnicas de aprendizado de máquina tornam-se ferramentas cruciais para entender melhor as dinâmicas do mercado imobiliário e prever seu comportamento futuro.

2. Metodologia

Esta seção descreve a metodologia adotada para a elaboração deste relatório. Serão detalhados o uso do conjunto de dados e os passos seguidos para a análise e obtenção dos resultados. Isso inclui a seleção das variáveis de interesse, a aplicação de técnicas de pré-processamento e estatísticas específicas e a seleção dos gráficos para interpretação dos padrões identificados.

2.1. Códigos

A metodologia adotada neste estudo envolveu várias etapas para o pré-processamento e análise dos dados do conjunto ”Housing”. Primeiramente, na Figura 1, carregamos o conjunto de dados a partir do arquivo CSV usando a biblioteca Pandas do Python. Em seguida, na Figura 2, substituímos os valores categóricos por valores numéricos para facilitar a análise. Posteriormente, tratamos os valores nulos, se houvesse, substituindo-os pela mediana dos respectivos atributos. Em seguida, selecionamos apenas as colunas desejadas para a análise, incluindo ”price”, ”area”, ”stories”e ”airconditioning”. Na etapa seguinte, normalizamos as características numéricas (Figura 4) utilizando a técnica de Min-Max Scaling para que todas as variáveis estejam na mesma escala. Além disso, aplicamos a discretização em algumas variáveis selecionadas, como ”area”e ”stories”, usando o método KBinsDiscretizer com 5 bins.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import MinMaxScaler, KBinsDiscretizer
from matplotlib.backends.backend_pdf import PdfPages
from google.colab import files
```

Figura 1. Bibliotecas.

```
from google.colab import drive
drive.mount('/content/drive')
```

Figura 2. Drive Colab.

```
# Carregar o conjunto de dados, tem que colocar o caminho ate o csv no teu drive
df = pd.read_csv('/content/drive/Meu Drive/Topicos em Computação/atv_pratica2/Housing.csv')
```

Figura 3. Carregando o conjunto de dados.

Para visualizar os efeitos das operações de pré-processamento, geramos histogramas e boxplots antes e depois das transformações (Figura 5). Isso nos permitiu observar como as distribuições e a variabilidade dos dados foram alteradas após

```python
# Converter colunas categóricas em numéricas
df.replace({'yes': 1, 'no': 0, 'furnished': 2, 'semi-furnished': 1, 'unfurnished': 0}, inplace=True)

# Tratar valores nulos, se houver
df.fillna(df.median(), inplace=True)

# Selecionar apenas as colunas desejadas
df_selected = df[['price', 'area', 'stories', 'airconditioning']]

# Normalizar as características numéricas
scaler = MinMaxScaler()
df_normalized = pd.DataFrame(scaler.fit_transform(df_selected), columns=df_selected.columns)

# Aplicar a discretização nas variáveis que necessitam
discretizer = KBinsDiscretizer(n_bins=5, encode='ordinal', strategy='uniform')
df_discretized = pd.DataFrame(discretizer.fit_transform(df[['area', 'stories']]), columns=['area', 'stories'])
```

Figura 4. Tratamento de dados.

o pré-processamento. Adicionalmente, geramos matrizes de dispersão para examinar as relações entre as variáveis antes e depois da normalização (Figura 6). Além disso, calculamos as matrizes de correlação entre as variáveis selecionadas e entre todas as variáveis do conjunto de dados, antes e depois da discretização. Por fim, na Figura 7, todas as visualizações foram salvas em um arquivo PDF chamado ”preprocessing_visualizations.pdf”, e as estatísticas descritivas, incluindo média, mediana, desvio padrão e variância, foram calculadas e salvas em um arquivo CSV denominado ”Housing_Statistics.csv”, para facilitar a referência e a análise futura dos resultados.

```python
# Gerar os histogramas antes e depois das operações
with PdfPages('/content/preprocessing_visualizations.pdf') as pdf_pages:
    # Histogramas antes das operações
    df_selected.hist(bins=50, figsize=(20,15))
    plt.suptitle('Histograms Before Preprocessing')
    pdf_pages.savefig()
    plt.close()

    # Histogramas depois da normalização
    df_normalized.hist(bins=50, figsize=(20,15))
    plt.suptitle('Histograms After Normalization')
    pdf_pages.savefig()
    plt.close()

    # Boxplots antes das operações
    df_selected.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False, figsize=(20,15))
    plt.suptitle('Boxplots Before Preprocessing')
    pdf_pages.savefig()
    plt.close()

    # Boxplots depois da discretização
    df_discretized.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False, figsize=(20,15))
    plt.suptitle('Boxplots After Discretization')
    pdf_pages.savefig()
    plt.close()

    # Matriz de dispersão antes das operações
    pd.plotting.scatter_matrix(df_selected, figsize=(20, 15))
    plt.suptitle('Scatter Matrix Before Preprocessing')
    pdf_pages.savefig()
    plt.close()
```

Figura 5. Geração de gráficos.

3. Resultados

Esta seção apresenta uma discussão acerca dos resultados obtidos. Nela serão exibidos as principais informações derivadas da análise estatística e exploratória dos dados, bem como as tendências identificadas no mercado imobiliário com base nos dados disponíveis. Além disso, serão discutidas as implicações práticas desses resultados e como podem ser utilizados para tomada de decisão no contexto do mercado de habitação.

A Figura 8 exibe histogramas que ilustram a distribuição de diferentes variáveis: preço, área, número de andares e presença de ar condicionado. O primeiro gráfico no canto superior esquerdo representa a distribuição do preço das propriedades. Ele mostra uma distribuição com uma cauda longa à direita, indicando que a maioria dos preços está concentrada em valores mais baixos, com poucos atingindo valores mais altos. O segundo gráfico no canto superior direito mostra a distribuição da área das propriedades. Sua forma é semelhante ao primeiro gráfico, com uma concentração em áreas menores

```
    # Matriz de dispersão depois da normalização
    pd.plotting.scatter_matrix(df_normalized, figsize=(20, 15))
    plt.suptitle('Scatter Matrix After Normalization')
    pdf_pages.savefig()
    plt.close()

    # Matriz de correlação antes das operações (com as 4 variáveis selecionadas)
    correlation_matrix_selected = df_selected.corr()
    plt.figure(figsize=(12, 10))
    sns.heatmap(correlation_matrix_selected, annot=True, cmap='coolwarm', fmt=".2f")
    plt.title('Correlation Matrix Before Preprocessing (Selected Variables)')
    pdf_pages.savefig()
    plt.close()

    # Matriz de correlação entre todas as variáveis
    correlation_matrix_all = df.corr()
    plt.figure(figsize=(12, 10))
    sns.heatmap(correlation_matrix_all, annot=True, cmap='coolwarm', fmt=".2f")
    plt.title('Correlation Matrix Between All Variables')
    pdf_pages.savefig()
    plt.close()

    # Matriz de correlação depois da discretização
    correlation_matrix_discretized = df_discretized.corr()
    plt.figure(figsize=(12, 10))
    sns.heatmap(correlation_matrix_discretized, annot=True, cmap='coolwarm', fmt=".2f")
    plt.title('Correlation Matrix After Discretization')
    pdf_pages.savefig()
    plt.close()

# Salva o PDF com os gráficos gerados no ambiente do Colab
files.download('/content/preprocessing_visualizations.pdf')
```

Figura 6. Salvar pdfs.

```
import pandas as pd
from google.colab import files  # Importa a biblioteca para salvar arquivos no Colab

# Carregar o conjunto de dados
df = pd.read_csv('/content/drive/MyDrive/si/topicos_comp/atv_02/Housing.csv')
df.replace({'yes': 1, 'no': 0, 'furnished': 2, 'semi-furnished': 1, 'unfurnished': 0}, inplace=True)

# Calcular medidas de tendência central
central_tendency = df.describe().loc[['mean', '50%']].transpose()
central_tendency.columns = ['Mean', 'Median']

# Calcular medidas de dispersão
dispersion = pd.DataFrame()
dispersion['Std Deviation'] = df.std()
dispersion['Variance'] = df.var()

# Combinar as medidas de tendência central e dispersão
statistics = pd.concat([central_tendency, dispersion], axis=1)

# Salvar em um arquivo CSV com apenas 3 casas decimais
statistics.to_csv('/content/drive/MyDrive/si/topicos_comp/atv_02/Housing_Statistics.csv', float_format='%.3f')

# Exibir as estatísticas calculadas
print(statistics)

# Baixar o arquivo CSV
files.download('/content/drive/MyDrive/si/topicos_comp/atv_02/Housing_Statistics.csv')
```

Figura 7. Estatísticas.

e poucas propriedades com áreas maiores. O terceiro gráfico no canto inferior esquerdo representa o número de andares das propriedades. A maioria das propriedades tem apenas 1 andar, seguida por aquelas com 2 andares, enquanto muito poucas têm 3 ou 4 andares. Por fim, o quarto gráfico no canto inferior direito mostra se as propriedades possuem ar condicionado ou não. Observa-se que a maioria das propriedades não possui ar condicionado.

A Figura 9 exibe o histograma das mesmas variáveis da figura anterior, porém após o pré-processamento. O primeiro gráfico no canto superior esquerdo representa a distribuição do preço após o pré-processamento. As barras azuis indicam uma distribuição com um pico pronunciado em torno de 0,2. Isso sugere que o pré-processamento pode ter normalizado os dados de preço. O segundo gráfico no canto superior direito mostra a distribuição da área após o pré-processamento. Semelhante ao primeiro gráfico, há um pico pronunciado, mas com várias outras elevações visíveis. Isso pode indicar que a área foi transformada de uma maneira que preserva mais da estrutura original dos dados.

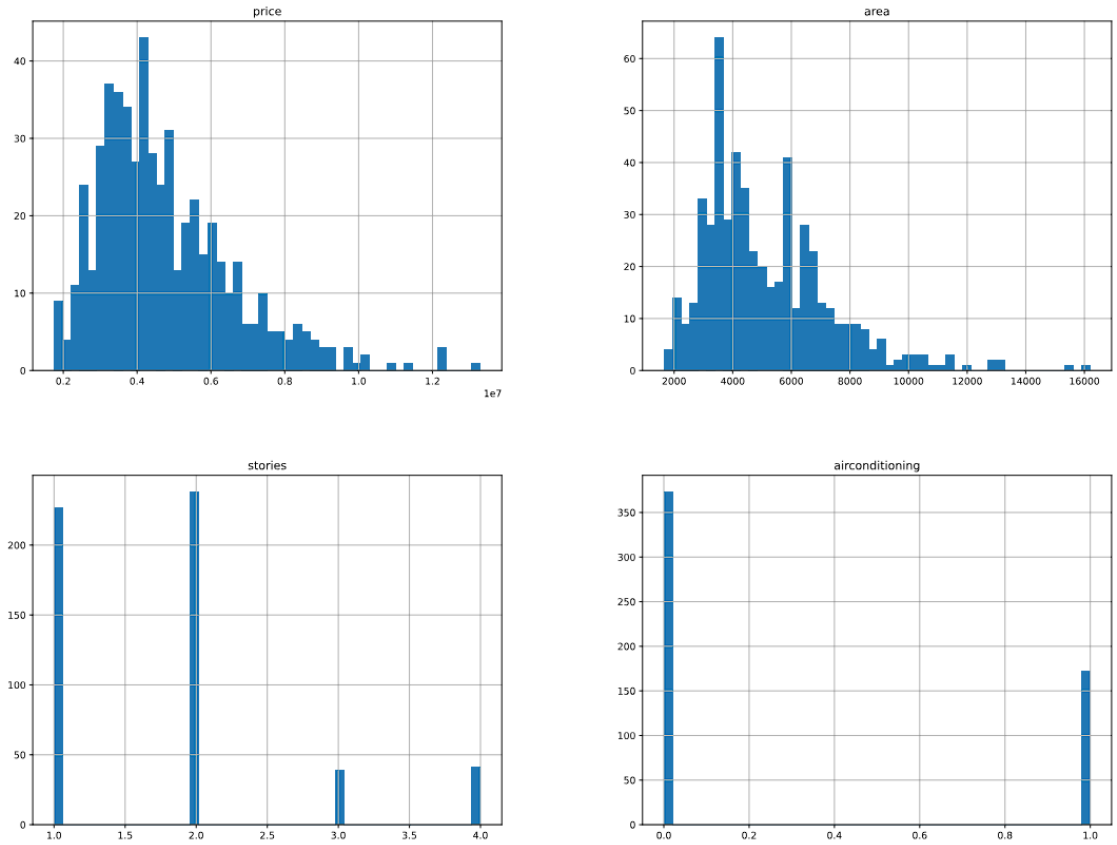

Figura 8. Histogramas antes do pré-processamento.

O terceiro gráfico no canto inferior esquerdo é para o número de andares após o pré-processamento. Ele mostra três barras distintas em 0, 0.2 e 1, indicando possivelmente um número discreto de andares nos dados. O último gráfico no canto inferior direito representa a presença de ar condicionado após o pré-processamento, com duas barras proeminentes em 0 e cerca de 0.9. Isso sugere que o ar condicionado é uma variável binária (sim ou não) nos dados.

A Figura 10 mostra os box plots e a distribuição de diferentes variáveis. O primeiro gráfico à esquerda superior representa a distribuição do preço. A caixa indica que muitos dados estão concentrados em uma faixa específica de preço; há vários pontos acima da caixa representando outliers, que são preços significativamente mais altos. O segundo gráfico à direita superior mostra a distribuição da área. Semelhante ao primeiro gráfico, a área também tem muitos dados concentrados em uma faixa mais baixa com alguns outliers visíveis. O terceiro gráfico à esquerda inferior é para quartos. A variação é menor comparada às duas primeiras variáveis, com poucos outliers visíveis. O último gráfico à direita inferior representa ar condicionado. Esta variável parece ser binária ou categórica, indicando talvez a presença ou ausência de ar-condicionado.

A Figura 11 apresenta o box plots após pré-processamento de discretização dos dados. O primeiro gráfico à esquerda representa a distribuição da área. A caixa indica que muitos dados estão concentrados em uma faixa específica de área; há dois pontos acima da caixa representando outliers, que são áreas significativamente maiores. O gráfico à

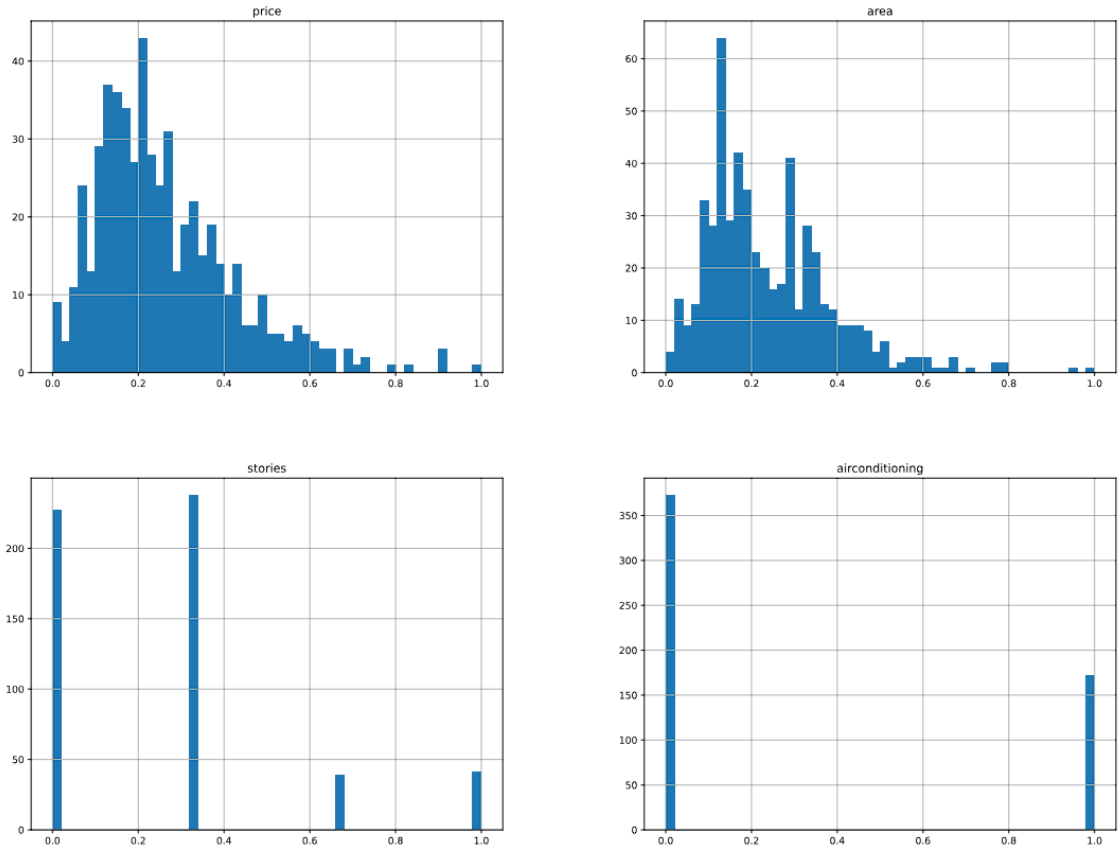

Figura 9. Histogramas depois do pré-processamento.

direita representa os dados relacionados às stories. O alcance interquartil é semelhante ao gráfico da area, mas não há outliers visíveis, indicando que a maioria das propriedades tem um número de andares dentro de um intervalo específico.

A Figura 12 apresenta os gráficos de dispersão que mostram a relação entre diferentes variáveis, como preço, área, número de quartos e banheiros. O primeiro gráfico à esquerda superior representa a relação entre o preço e a área. Os pontos no gráfico indicam que há uma tendência positiva entre essas duas variáveis, ou seja, à medida que a área aumenta, o preço também tende a aumentar.O segundo gráfico à direita superior mostra a relação entre o preço e o número de quartos. Os pontos no gráfico indicam que há uma tendência positiva entre essas duas variáveis, ou seja, casas com mais quartos tendem a ter preços mais altos. O terceiro gráfico à esquerda inferior é para preço vs banheiros. A distribuição dos pontos sugere que casas com mais banheiros tendem a ter preços mais altos. O último gráfico à direita inferior representa área vs número de quartos. Os pontos no gráfico indicam que casas com mais quartos tendem a ter uma área maior.

A Figura 13 exibe os resultados após o pré-processamento dos dados. O primeiro gráfico à esquerda superior representa a relação entre o preço e a área. Os pontos no gráfico indicam que há uma tendência positiva entre essas duas variáveis, ou seja, à medida que a área aumenta, o preço também tende a aumentar. O segundo gráfico à direita superior mostra a relação entre o preço e o número de quartos. Os pontos no gráfico indicam que há uma tendência positiva entre essas duas variáveis, ou seja, casas com mais quartos tendem a ter preços mais altos. O terceiro gráfico à esquerda inferior é para preço

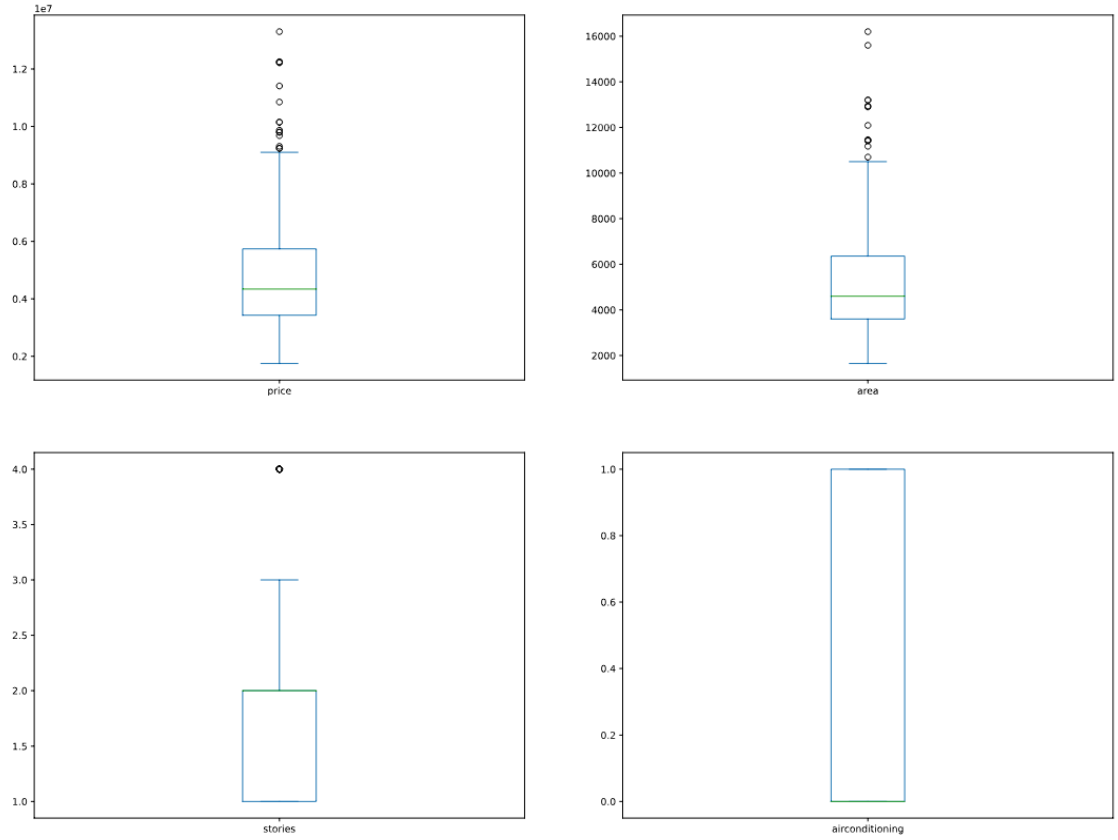

Figura 10. Box plots antes do pré-processamento.

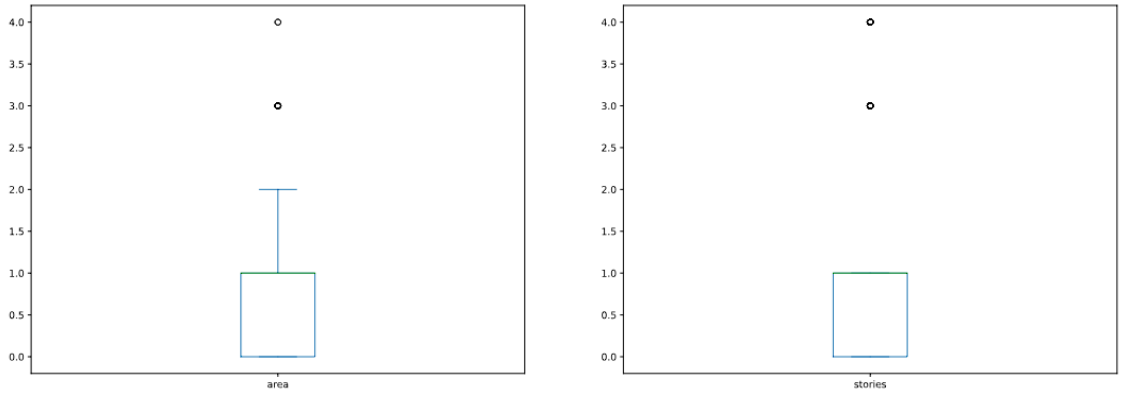

Figura 11. Box plots depois do pré-processamento.

vs banheiros. A distribuição dos pontos sugere que casas com mais banheiros tendem a ter preços mais altos. O último gráfico à direita inferior representa área vs número de quartos. Os pontos no gráfico indicam que casas com mais quartos tendem a ter uma área maior.

A Figura 14 apresenta uma matriz de correlação que mostra as relações entre diferentes variáveis, como preço, área, número de andares e ar condicionado.A primeira linha/coluna representa a correlação do preço com todas as outras variáveis. O valor de 1 na diagonal principal indica que o preço tem uma correlação perfeita consigo mesmo,

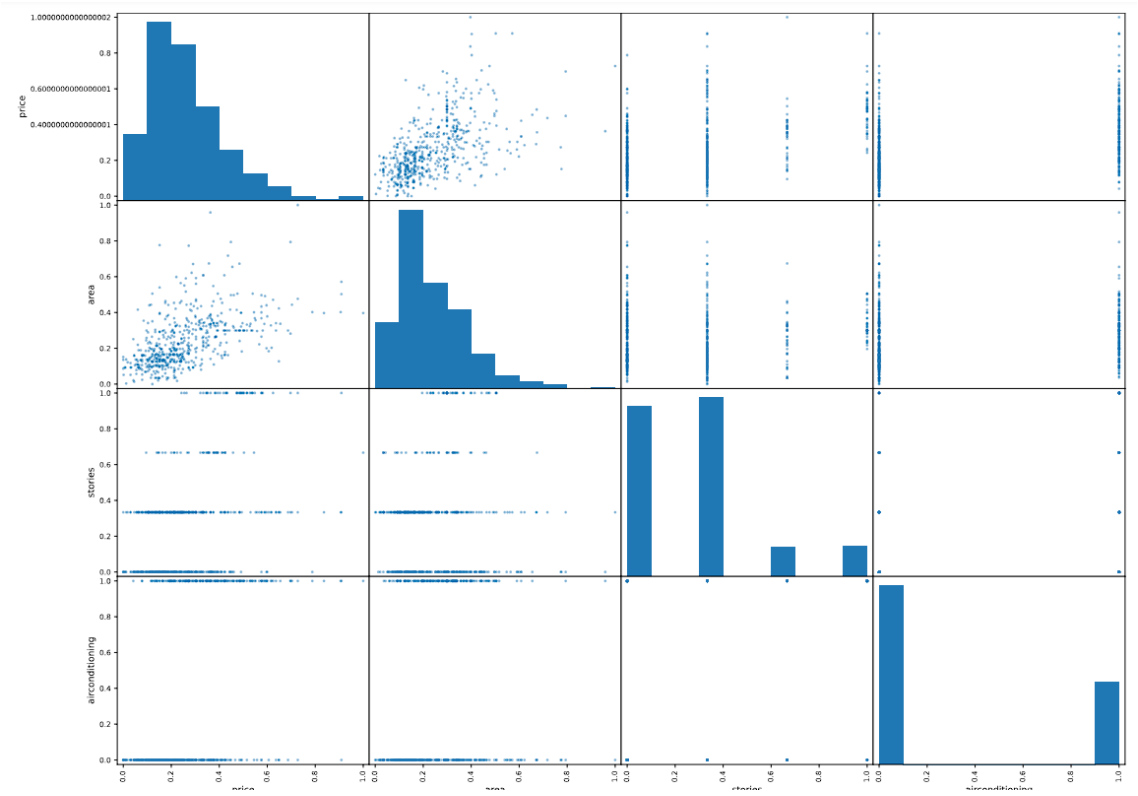

Figura 12. Dispersão antes do pré-processamento.

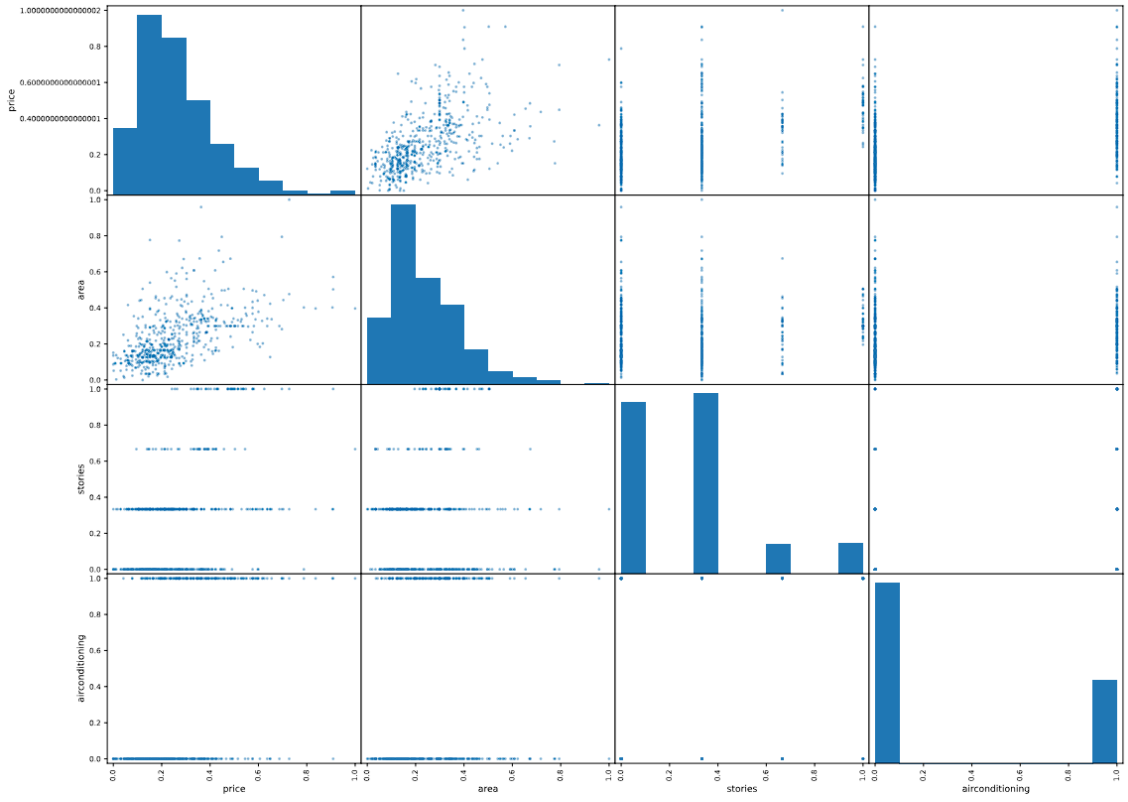

Figura 13. Dispersão depois do pré-processamento.

como esperado. A segunda linha/coluna representa a correlação da área com todas as outras variáveis. A terceira linha/coluna é para número de andares. A última linha/coluna representa ar condicionado. Os valores representam a correlação do ar condicionado com o preço, a área e o número de andares. Os números na matriz representam os coeficientes de correlação entre as variáveis. Um valor de 1 indica uma correlação perfeita positiva, enquanto um valor de -1 indica uma correlação perfeita negativa. Valores próximos a 0 indicam pouca ou nenhuma correlação. E observando a matriz podemos notar que a correlação que apresenta uma melhor combinação é a área com o preço.

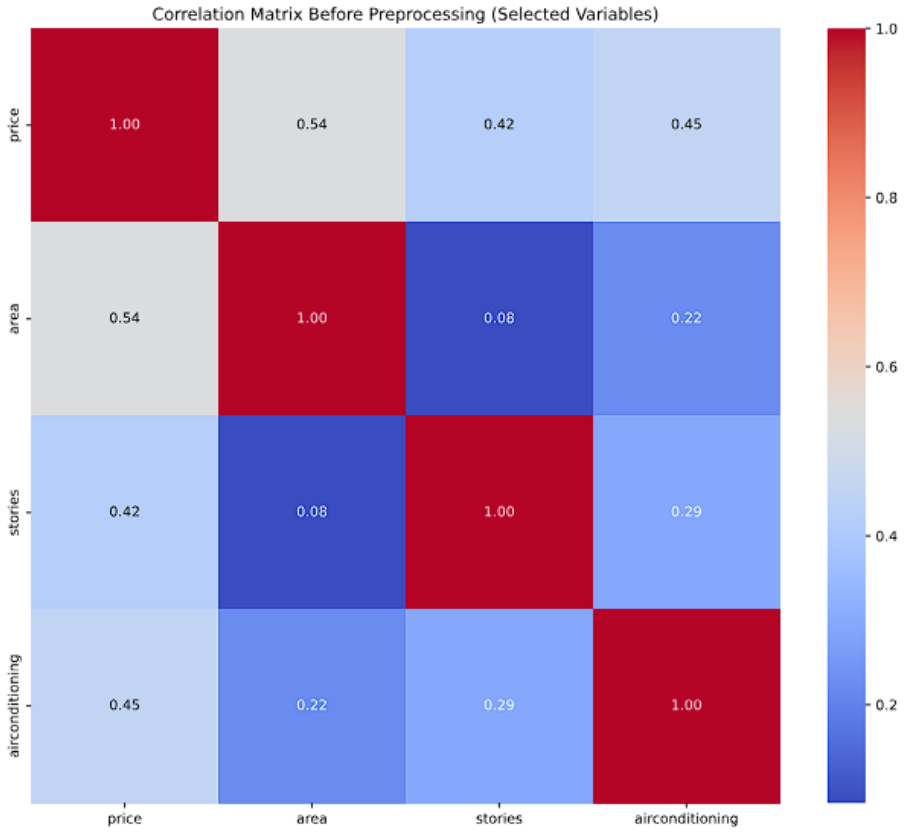

Figura 14. Matriz de correlação antes do pré-processamento.

A Figura 15 apresenta a matriz de correlação com todas as variáveis do conjunto de dados. As cores das células variam do vermelho ao azul; o vermelho indica uma alta correlação positiva e o azul indica uma baixa correlação. Por exemplo, a correlação entre o preço e a área é de 0.54, indicando uma correlação moderada positiva. Isso significa que, em geral, à medida que a área de uma propriedade aumenta, o preço também tende a aumentar. O mesmo padrão se aplica ao preço e aos banheiros com 0.52, andares com 0.42 e ao aquecimento de água quente, com 0.42.

A Tabela 1 fornece um resumo estatístico dos dados relacionados a diversas características de propriedades imobiliárias. Cada linha representa uma característica específica, enquanto as colunas exibem medidas estatísticas cruciais, incluindo média, mediana, desvio padrão e variância. A média representa o valor médio da característica em questão em todo o conjunto de dados. Por exemplo, a média do preço das propriedades é de aproximadamente 4.766.729,248. A mediana é o valor que divide os dados em

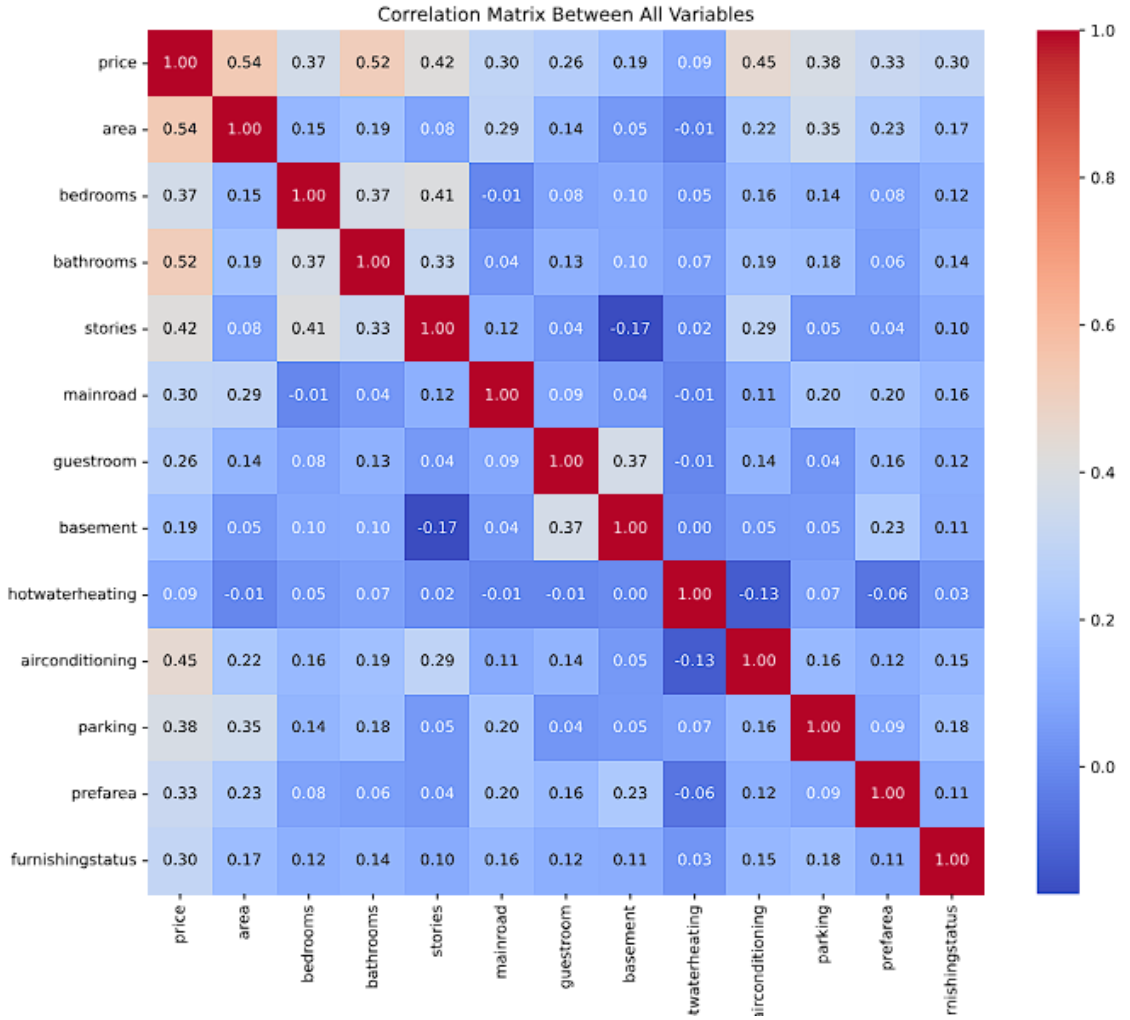

Figura 15. Matriz de correlação depois do pré-processamento.

duas partes iguais quando organizados em ordem crescente. Por exemplo, a mediana do número de quartos (bedrooms) é de 3. O desvio padrão indica a dispersão dos valores em relação à média. Quanto maior o desvio padrão, maior é a dispersão dos dados. Por exemplo, o desvio padrão para a área das propriedades é de aproximadamente 2.170,141. A variância reflete a variabilidade ou a dispersão dos valores em relação à média, sendo o quadrado do desvio padrão. Valores de variância maiores indicam maior dispersão dos dados. Por exemplo, a variância do preço das propriedades é de aproximadamente 3.498.544.355.820,573.

4. Conclusão

Este trabalho propôs a realização de uma análise de dados de um dataset de housing para identificar padrões ou tendências que possam indicar características importantes relacionadas ao mercado imobiliário e suas características. Os histogramas exibidos destacam as distribuições originais e após o pré-processamento das variáveis de preço, área, número de andares e presença de ar condicionado. Antes do pré-processamento, observamos uma concentração de valores mais baixos de preço e área, bem como uma predominância de propriedades com apenas 1 ou 2 andares e sem ar condicionado. No entanto, após o

Tabela 1. Resumo Estatístico dos Dados.

Componentes	Média	Mediana	Desvio Padrão	Variância
price	4766729.248	4340000.000	1870439.616	3498544355820.573
area	5150.541	4600.000	2170.141	4709512.058
bedrooms	2.965	3.000	0.738	0.545
bathrooms	1.286	1.000	0.502	0.252
stories	1.806	2.000	0.867	0.753
mainroad	0.859	1.000	0.349	0.122
guestroom	0.178	0.000	0.383	0.147
basement	0.350	0.000	0.478	0.228
hotwaterheating	0.046	0.000	0.209	0.044
airconditioning	0.316	0.000	0.465	0.216
parking	0.694	0.000	0.862	0.742
prefarea	0.235	0.000	0.424	0.180
furnishingstatus	0.930	1.000	0.761	0.580

pré-processamento, notamos uma normalização dos dados, especialmente no preço e na área, indicando uma possível correção das distribuições originais. Os box plots, oferecem uma visão mais detalhada da distribuição dos dados e a presença de outliers antes e depois do pré-processamento. Observamos uma redução na dispersão dos dados após a discretização, especialmente na distribuição da área e do número de andares, indicando uma maior consistência nos valores dessas variáveis. As dispersões revelam as relações entre preço, área, número de quartos e banheiros. Antes e após o pré-processamento, observamos uma tendência positiva entre preço e área, bem como entre preço e número de quartos, sugerindo que propriedades maiores e com mais quartos tendem a ter preços mais altos. Além disso, as matrizes de correlação confirmam essas relações, mostrando correlações moderadas positivas entre preço e área, quartos e banheiros. Portanto, os resultados da análise exploratória indicam que as variáveis de área, número de quartos e banheiros têm uma influência significativa no preço das propriedades. Além disso, o pré-processamento dos dados parece ter normalizado as distribuições e reduzido a dispersão, tornando-os mais adequados para análises subsequentes. Essas descobertas têm implicações importantes para a tomada de decisões no mercado imobiliário, permitindo aos investidores e compradores entender melhor os fatores que afetam os preços das propriedades

Referências