Relatório

Previsão de Preço de Imóveis de São Caetano do Sul

São Paulo

2º/2024

Atividade em Grupo em ordem alfabética (Máximo 3 alunos)

|  |  |
| --- | --- |
| Nome | RA |
|  |  |
|  |  |
|  |  |

Sumário

[1 INTRODUÇÃO 4](#_Toc160784884)

[2 INTEGRAÇÃO E DATA UNDERSTANDING 7](#_Toc160784885)

[2.1 Base de dados 7](#_Toc160784886)

[2.2 Atributos da Base de Dados Final 7](#_Toc160784887)

[2.3 Tabela de Correlação de Pearson 8](#_Toc160784888)

[2.4 Tabela de Frequência 8](#_Toc160784889)

[2.5 Skewness 9](#_Toc160784890)

[2.6 Kurtosis 9](#_Toc160784891)

[2.7 Heatmap Correlação 10](#_Toc160784892)

[2.8 Scatter plot 10](#_Toc160784893)

[2.8.1 area total x preco 11](#_Toc160784894)

[2.8.2 quartos x preco 11](#_Toc160784895)

[2.9 Histograma Geral 12](#_Toc160784896)

[2.10 Histograma com Curva: area\_total 12](#_Toc160784897)

[2.11 Histograma com Curva: preco 13](#_Toc160784898)

[2.12 Histograma e Scatter Geral 14](#_Toc160784899)

[2.13 Boxplot preco 15](#_Toc160784900)

[2.14 Boxplot área\_total 15](#_Toc160784901)

[2.15 Boxplot quartos 16](#_Toc160784902)

[2.16 Problemas Identificados 17](#_Toc160784903)

[2.17 Questões – Compreendendo os Dados 17](#_Toc160784904)

[3 DATA PREPARATION 18](#_Toc160784905)

[3.1 Eliminação Manual de Atributos e Limpeza 18](#_Toc160784906)

[3.2 Atributos Dataset: Antes da Preparação 20](#_Toc160784907)

[3.3 Normalização MaxMin/Escore-Z, One-Hot Encoding e Ordenação do dataset 21](#_Toc160784908)

[3.4 Atributos Dataset: Depois da Preparação e Ordenação das Colunas 21](#_Toc160784909)

INTRODUÇÃO

Considere que fazemos parte de uma imobiliária que tem imóveis localizados UNICAMENTE em São Caetano do sul. Ela quer disponibilizar um preditor aos seus clientes para sugerir o valor de venda ou compra de um imóvel.

Algumas perguntas de negócio poderiam ser:

* Quero comprar um imóvel seguindo algumas características, é possível prever qual o preço ou valor do imóvel?
* Quais dados devem ser capturados para avaliar o valor a pagar?
* Posso montar um conjunto de dados para gerar um modelo preditivo que permite um vendedor avaliar o preço que deve ser ofertado do imóvel? Com esse mesmo conjunto de dados, caso eu queira comprar, é possível utilizá-lo para prever o valor de compra?

A coleta de dados foi realizada por alunos de um curso de Ciência da Computação e Sistemas de Informação em março de 2021 de uma instituição acadêmica fazendo uso de coleta de dados manuais e *webscraping* nos sites:

* [www.vivareal.com.br/](http://www.vivareal.com.br/)
* [www.creditas.com](http://www.creditas.com/)
* [www.imovelweb.com.br](http://www.imovelweb.com.br/)

Nesse cenário, foram coletados e montados 23 datasets por diferentes elementos, dos quais 20 deles constam em uma pasta de nome bd\_SCS.

Os atributos considerados durante essa coleta de dados foram:

* Tipo: Casa (1), Apto (2).
* Tempo do imóvel em anos: (0 = novo)
* Localização: Rua Luís Cavana, 70 - Centro, São Caetano do Sul – SP.
* Área total (m2):
* Área útil (m2):
* Número de quartos:
* Suíte: Sim (1) ou Não (0)
* Número de Banheiros
* Número de Vagas para Carros
* Academia: Sim (1) ou Não (0)
* Churrasqueira: Sim (1) ou Não (0)
* Elevador: Sim (1) ou Não (0)
* Salão de Festas: Sim (1) ou Não (0)
* Piscina: Sim (1) ou Não (0)
* Valor do IPTU: em R$
* Valor do Condomínio: em R$.
* Valor do Imóvel: em R$.

Em um primeiro momento, o professor da disciplina fez uma integração de três *datasets*, sendo dois arquivos texto com extensão “csv”: bdSCS\_1\_csv.csv e bdSCS\_2\_csv.csv, e um do Excel com extensão xlsx: bdSCS\_3\_excel.xlsx, gerando ao final único dataset com nome bdSCS\_final\_csv.csv.

No processo de integração, foram realizadas manipulações nos conteúdos internos de cada um dos arquivos fazendo uso de algumas bibliotecas do Python, dentre elas: pandas, numpy, matplotlib, seaborn e re (expressões regulares), chegando-se a um resultado final de 13 atributos para bdSCS\_final\_csv.csv, sendo 12 deles candidatos a atributos de entrada e 1 o atributo alvo (Valor do Imóvel). Após esse processo, os atributos resultantes foram:

* Tipo: Casa (1), Apto (2).
* Bairro.
* Área total (m2)
* Número de quartos
* Número de Banheiros
* Número de Vagas para Carros
* Academia: Sim (1) ou Não (0)
* Churrasqueira: Sim (1) ou Não (0)
* Elevador: Sim (1) ou Não (0)
* Salão de Festas: Sim (1) ou Não (0)
* Piscina: Sim (1) ou Não (0)
* Valor do Condomínio: em R$.
* Valor do Imóvel: em R$.

A partir do exposto e dos resultados obtidos na etapa Data Understanding do processo CRISP-DM, realizar a etapa Data Preparation com o dataset obtido, preencher o solicitado no relatório.

INTEGRAÇÃO E DATA UNDERSTANDING

* 1. Base de dados

Professor

* Nome do arquivo csv utilizado: bdSCS\_1\_csv e bdSCS\_2\_csv
* Nome do arquivo Excel utilizado: bdSCS\_3\_excel
* Nome da base de dados final: bdSCS\_final\_csv.csv
* Total de Registros (Instâncias ou Observações) do arquivo de imóveis de São Caetano do Sul Final: 6381.
* Total de Atributos Final: 13.

Aluno

* Nome do arquivo csv utilizado: (Preencher)
* Nome do arquivo Excel utilizado: (Preencher)
* Nome da base de dados final: bdSCS\_final\_csv.csv
* Total de Registros (Instâncias ou Observações) do arquivo de imóveis de São Caetano do Sul Final: . (Preencher)
  1. Atributos da Base de Dados Final

Professor

Nesta etapa, foram considerados 13 atributos de cada imóvel, sendo eles:

* Tipo: Casa (1), Apto (2).
* Bairro.
* Área total (m2):
* Número de quartos:
* Número de Banheiros:
* Número de Vagas para Carros:
* Academia: Sim (1) ou Não (0)
* Churrasqueira: Sim (1) ou Não (0)
* Elevador: Sim (1) ou Não (0)
* Salão de Festas: Sim (1) ou Não (0)
* Piscina: Sim (1) ou Não (0)
* Valor do Condomínio: em R$.
* Valor do Imóvel: em R$.

Aluno

Nesta etapa, foram considerados (Preencher) atributos de cada imóvel, sendo eles:

(Preencher)

* 1. Tabela de Correlação de Pearson

Professor

A screenshot of a cell phone

Description automatically generated

Aluno

(Inserir aqui)

* 1. Tabela de Frequência

Professor

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Aluno

(Inserir aqui)

* 1. Skewness

Professor

A screenshot of a cell phone

Description automatically generated

Aluno

(Inserir aqui)

* 1. Kurtosis

Professor

A screenshot of a cell phone

Description automatically generated

Aluno

(Inserir aqui)

* 1. Heatmap Correlação

Professor

A picture containing drawing

Description automatically generated

Aluno

(Inserir aqui)

* 1. Scatter plot
     1. area total x preco

Professor

A close up of a map

Description automatically generated

Aluno

(Inserir aqui)

* + 1. quartos x preco

Professor

A screenshot of a cell phone

Description automatically generated

Aluno

(Inserir aqui)

* 1. Histograma Geral

Professor

A screenshot of a cell phone

Description automatically generated

Aluno

(Inserir aqui)

* 1. Histograma com Curva: area\_total

Professor

A screenshot of a social media post

Description automatically generated

sk = 6.659796686013957, Ck = 110.77760469964831

Média=96.0025074439743, Mediana=76.0, Moda= 90

Aluno

(Inserir aqui)

* 1. Histograma com Curva: preco

Professor

A screenshot of a cell phone

Description automatically generated

sk = 2.8330314405741173, Ck = 7.515772619574491

Média=1193589.0520294625, Mediana=510000.0, Moda= 990000

Aluno

(Inserir aqui)

* 1. Histograma e Scatter Geral

Professor

A picture containing cake, rain, large, person

Description automatically generated

Aluno

(Inserir aqui)

* 1. Boxplot preco

Professor

Rótulo valor\_baixo quartil\_hum mediana quartil\_tres valor\_alto

preco 540.0 450000.0 510000.0 990000.0 1800000.0

A picture containing boat, parked

Description automatically generated

Aluno

(Inserir aqui)

* 1. Boxplot área\_total

Professor

Rótulo valor\_baixo quartil\_hum mediana quartil\_tres valor\_alto

area\_total 38.0 65.0 76.0 106.0 162.0

A picture containing boat

Description automatically generated

Aluno

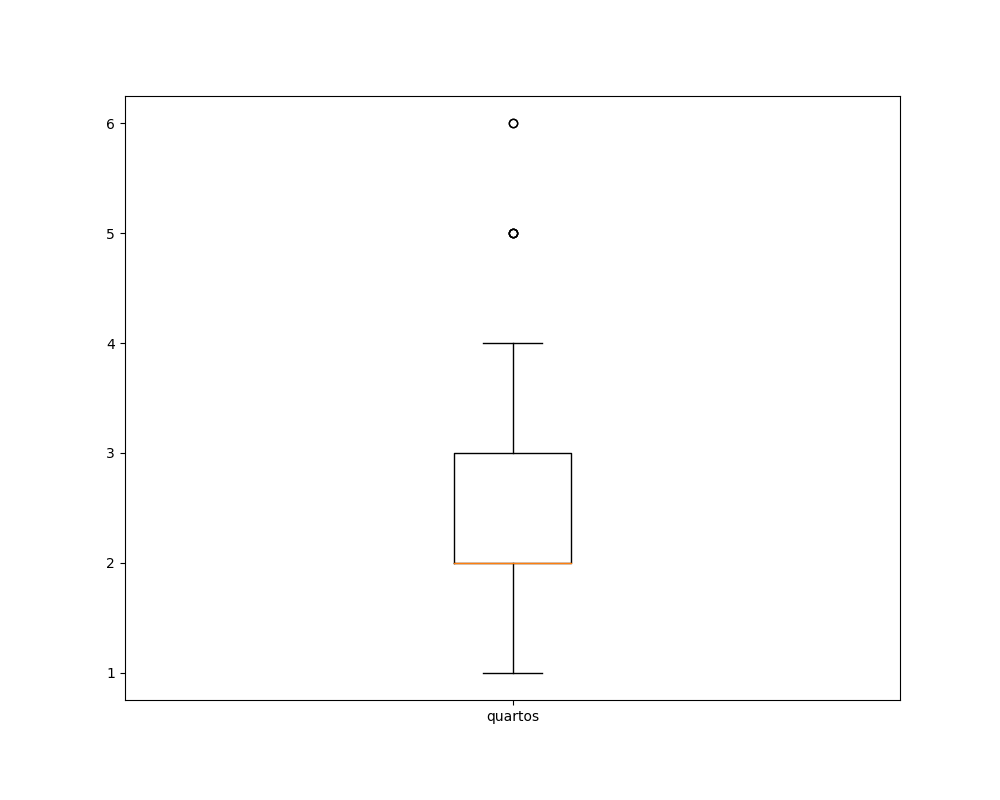
(Inserir aqui)

* 1. Boxplot quartos

Professor

Rótulo valor\_baixo quartil\_hum mediana quartil\_tres valor\_alto

quartos 1.0 2.0 2.0 3.0 4.0



Aluno

(Inserir aqui)

* 1. Problemas Identificados

Dos 17 atributos inicialmente selecionados, o tempo do imóvel, suíte, área útil e valor do IPTU foram removidos por não apresentar valores significativos na maior parte das observações.

O atributo localização continha o endereço completo, mas foi considerado somente o valor do bairro. Por este fato, o atributo foi renomeado para bairro.

O atributo condomínio não apresenta dados em alguns registros. Logo, algo deve ser feito na etapa de preparação dos dados.

O atributo banheiros e vagas apresentam valores inconsistentes internamente. Logo, algo deve ser feito na etapa de preparação dos dados.

* 1. Questões – Compreendendo os Dados

A resposta a essas questões deve constar no Notebook (IA\_EAD\_Atividade\_DataUnderstanding.ipynb) em local apresentado no seu interior. Os resultados obtidos do código elaborado deve ser apresentado como respostas às questões elencadas abaixo.

1. Qual a quantidade de imóveis por bairro do tipo 2?

Resposta:

1. Qual o valor médio dos preços dos imóveis do tipo 1 do bairro Santa Paula?

Resposta:

1. Quantos imóveis do tipo 1 que tem menos do que 2 banheiros com área total maior do que 100 m2

Resposta:

1. Qual o valor médio do condomínio dos imóveis do tipo 1 que tem academia, elevador e piscina?

Resposta:

1. Qual a área média dos imóveis (tipo 1 e 2) que tem 2 quartos ou mais e valor do imóvel maior do que R$ 600000,00?

Resposta:

1. Qual a quantidade de imóveis do tipo 1 que pelo menos 2 quartos, ao menos 2 vagas de garagem, com salão de festas e churrasqueira?

Resposta:

1. Qual a quantidade de imóveis do tipo 1 possui seu preço maior do que a média dos preços dos imóveis (também do tipo 1)?

Resposta:

1. Quantos imóveis do tipo 2 são do bairro Barcelona, Santa Maria e Nova Gerty e tem pelo menos 3 quartos?

Resposta:

1. Apresentar a relação de imóveis de qualquer tipo com preço maior do que R$ 600.000,00 com pelo menos 2 vagas de garagem?

Resposta:

DATA PREPARATION

* 1. Eliminação Manual de Atributos e Limpeza

**Professor**

|  |  |  |
| --- | --- | --- |
| Atividade |  | Valores/Ação realizada |
| Base de dados antes da preparação | Nome da Base | bdSCS\_final\_profIvan.csv |
|  | Total de objetos | 6381 |
|  | Total de Atributos | 13 |
| Eliminação Manual de Atributos | Removido o atributo | Tipo, pois havia poucos valores com Tipo = 1 (casa), podiam afetar a sua predição. |
|  | Antes de remover o atributo Tipo (coluna) | Foram removidos os objetos com Tipo = ‘1’ (casa) |
|  | Total de Objetos restantes | 6272 |
|  | Total de Atributos restantes | 12 |
| Limpeza de Dados – Dados Incompletos | Objetos contendo o atributo condomínio | 11 deles com valores nulos. Foram todos removidos. |
|  | Total de Objetos restantes | 6261 |
| Limpeza de Dados – Dados Inconsistentes | Atributo banheiros contendo 179 “—” e 45 “2-3” | Removidos todos os objetos com esses valores |
|  | Total de Objetos restantes | 6037 |
|  | Os atributos banheiros e vagas estavam com o tipo de dados object | Foram convertidos para inteiro |
|  | O conteúdo do atributo bairro apresentavam dois problemas | 1) acentos nos nomes dos campos  2) Bairro Jardim São Caetano tem com acento e sem acento. |
|  | Ação realizada no atributo bairro | Os acentos foram removidos, resolvendo os dois problemas. |
|  | Remoção dos objetos nos quais o atributo condomínio | tem o valor 0 (757 objetos) |
|  | Total de Objetos restantes | 5280 |
| Limpeza de Dados – Dados Redundantes - Objetos | Foram identificados alguns objetos redundantes | Total 809. Todos foram removidos. |
|  | Total de Objetos restantes | 4471 |
| Limpeza de Dados - Dados Redundantes - Atributos | Inexistente |  |
| Limpeza de Dados - Ruídos | Com base no boxplot, | foi selecionado o atributo preço para remoção de outliers |
|  | Técnica utilizada para remoção dos outliers do atributo preço | outliers baixos estão abaixo de Q1 - 1,5⋅FIQ e outliers altos estão acima de Q3 + 1,5⋅FIQ, onde FIQ = Q3 – Q1. |
|  | Total de Objetos restantes | 3831 |

**Aluno**

|  |  |  |
| --- | --- | --- |
| Atividade |  | Valores/Ação realizada |
| Base de dados antes da preparação | Nome da Base | bdSCS\_final\_csv.csv |
|  | Total de objetos | <Preencher> |
|  | Total de Atributos | <Preencher> |
| Eliminação Manual de Atributos | Removido o atributo | Tipo, pois havia poucos valores com Tipo = 1 (casa), podiam afetar a sua predição. |
|  | Antes de remover o atributo Tipo (coluna) | Foram removidos os objetos com Tipo = ‘1’ (casa) |
|  | Total de Objetos restantes | <Preencher> |
|  | Total de Atributos restantes | <Preencher> |
| Limpeza de Dados – Dados Incompletos | Objetos contendo o atributo condomínio | <Preencher> deles com valores nulos. Foram todos removidos. |
|  | Total de Objetos restantes | <Preencher> |
| Limpeza de Dados – Dados Inconsistentes | Atributo banheiros contendo 179 “—” e 45 “2-3” | Removidos todos os objetos com esses valores |
|  | Total de Objetos restantes | <Preencher> |
|  | Os atributos banheiros e vagas estavam com o tipo de dados object | Foram convertidos para inteiro |
|  | O conteúdo do atributo bairro apresentavam dois problemas | 1) acentos nos nomes dos campos  2) Bairro Jardim São Caetano tem com acento e sem acento. |
|  | Ação realizada no atributo bairro | Os acentos foram removidos, resolvendo os dois problemas. |
|  | Remoção dos objetos nos quais o atributo condomínio | tem o valor 0 (<Preencher> objetos) |
|  | Total de Objetos restantes | <Preencher> |
| Limpeza de Dados – Dados Redundantes - Objetos | Foram identificados alguns objetos redundantes | Total <Preencher>. Todos foram removidos. |
|  | Total de Objetos restantes | <Preencher> |
| Limpeza de Dados - Dados Redundantes - Atributos | Inexistente |  |
| Limpeza de Dados - Ruídos | Com base no boxplot, | foi selecionado o atributo preço para remoção de outliers |
|  | Técnica utilizada para remoção dos outliers do atributo preço | outliers baixos estão abaixo de Q1 - 1,5⋅FIQ e outliers altos estão acima de Q3 + 1,5⋅FIQ, onde FIQ = Q3 – Q1. |
|  | Total de Objetos restantes | <Preencher> |

* 1. Atributos Dataset: Antes da Preparação

**Professor**

Nesta etapa, foram considerados 13 atributos de cada imóvel, sendo eles:

* Tipo: Casa (1), Apto (2).
* Bairro.
* Área total (m2):
* Número de quartos:
* Número de Banheiros:
* Número de Vagas para Carros:
* Academia: Sim (1) ou Não (0)
* Churrasqueira: Sim (1) ou Não (0)
* Elevador: Sim (1) ou Não (0)
* Salão de Festas: Sim (1) ou Não (0)
* Piscina: Sim (1) ou Não (0)
* Valor do Condomínio: em R$.
* Valor do Imóvel: em R$.

**Aluno**

Nesta etapa, foram considerados (Preencher) atributos de cada imóvel, sendo eles:

(Preencher)

* 1. Normalização MaxMin/Escore-Z, One-Hot Encoding e Ordenação do dataset

**Professor**

|  |  |  |
| --- | --- | --- |
| Atividade |  | Atributo |
| Normalização MaxMIn | Convertidos dentro do intervalo: [0, 1] | Quartos, Banheiros e Vagas |
| Normalização (Padronização) Escore-Z | Convertidos para média 0 e desvio padrão 1 | Área\_total e condomínio |
| One-Hot Encoding | Transformação de Variáveis Categóricas para Numéricas | Bairro |
|  | O atributo categórico apresentava | 13 valores diferentes |
|  | Logo, aumentou o número de atributos | Em mais 13 |
| Ordenação do dataset | O atributo alvo não estava no final, então | Preco teve a ordem modificada para o final do dataset |

* 1. Atributos Dataset: Depois da Preparação e Ordenação das Colunas

**Professor**

Ao final foram obtidos 3831 objetos contendo 24 atributos, sendo:

* Candidatos a Atributos Preditores (23):
  1. area\_total: Trasnformação Escore-Z.
  2. quartos: Transformação MaxMin (0-1).
  3. banheiros: Transformação MaxMin (0-1).
  4. vagas: Transformação MaxMin (0-1).
  5. academia: Sim (1) ou Não (0).
  6. churrasqueira: Sim (1) ou Não (0).
  7. elevador: Sim (1) ou Não (0).
  8. salao\_festas: Sim (1) ou Não (0).
  9. piscina: Sim (1) ou Não (0).
  10. condomínio: Trasnformação Escore-Z.
  11. bairro\_Barcelona: Sim (1) ou Não (0).
  12. bairro\_Boa\_Vista: Sim (1) ou Não (0).
  13. bairro\_Centro: Sim (1) ou Não (0).
  14. bairro\_Ceramica: Sim (1) ou Não (0).
  15. bairro\_Fundacao: Sim (1) ou Não (0).
  16. bairro\_Jardim\_Sao\_Caetano: Sim (1) ou Não (0).
  17. bairro\_Maua: Sim (1) ou Não (0).
  18. bairro\_Nova\_Gerty: Sim (1) ou Não (0).
  19. bairro\_Olimpico: Sim (1) ou Não (0).
  20. bairro\_Osvaldo\_Cruz: Sim (1) ou Não (0).
  21. bairro\_Santa\_Maria: Sim (1) ou Não (0).
  22. bairro\_Santa\_Paula: Sim (1) ou Não (0).
  23. bairro\_Santo\_Antonio: Sim (1) ou Não (0).
* Atributo Alvo
  + preco: em R$.

**Aluno**

Ao final foram obtidos (Preencher) objetos contendo (Preencher) atributos, sendo:

* Candidatos a Atributos Preditores ((Preencher)):
* Atributo Alvo
  + preco: em R$.