Analytics e Inteligência Artificial Data Science

Tema da aula

Análise Exploratória de Dados

FIA BUSINESS SCHOOL

lab.data

FIA Business School

NOSSOS DIFERENCIAIS | QUEM SOMOS

BUSINESS SCHOOL
Graduação, pós-graduação, MBA, Pós- MBA, Mestrado Profissional, Curso In *Company* e EAD

CONSULTING
Consultoria personalizada que oferece soluções baseadas em seu problema de negócio

RESEARCH
Atualização dos conhecimentos e do material didático oferecidos nas atividades de ensino

FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO

Líder em Educação Executiva, referência de ensino nos cursos de graduação, pós-graduação e MBA, tendo excelência nos programas de educação. Uma das principais escolas de negócio do mundo, possuindo convênios internacionais com Universidades nos EUA, Europa e Ásia. +8.000 projetos de consultorias em organizações públicas e privadas.

- Único curso de graduação em administração a receber as notas máximas
- A primeira escola brasileira a ser finalista da maior competição de MBA do mundo
- Única *Business School* brasileira a figurar no *ranking* LATAM
- Signatária do Pacto Global da ONU
- Membro fundador da ANAMBA - Associação Nacional MBAs
- Credenciada pela AMBA - *Association of MBAs*
- Credenciada ao *Executive MBA Council*
- Filiada a AACSB - *Association to Advance Collegiate Schools of Business*
- Filiada a EFMD - *European Foundation for Management Development*
- Referência em cursos de MBA nas principais mídias de circulação

BUSINESS SCHOOL lab.data

O **Laboratório de Análise de Dados** – LABDATA é um Centro de Excelência que atua nas áreas de ensino, pesquisa e consultoria em análise de informação utilizando técnicas de ***Big Data***, ***Analytics*** e **Inteligência Artificial**.


Profª Drª Alessandra Montini

O LABDATA é um dos pioneiros no lançamento dos cursos de *Big Data* e *Analytics* no Brasil.

Os diretores foram professores de grandes especialistas do mercado.

+10 anos de atuação.

+9.000 alunos formados.

Docentes

- Sólida formação acadêmica: doutores e mestres em sua maioria;
- Larga experiência de mercado na resolução de *cases;*
- Participação em congressos nacionais e internacionais;
- Professor assistente que acompanha o aluno durante todo o curso.

Estrutura

- 100% das aulas realizadas em laboratórios;
- Computadores para uso individual durante as aulas;
- 5 laboratórios de alta qualidade (investimento +R$2MM);
- 2 unidades próximas à estação de metrô (com estacionamento).

FIA BUSINESS SCHOOL lab.data

PROFA. DRA. ALESSANDRA DE ÁVILA MONTINI

Diretora do LABDATA-FIA, apaixonada por dados e pela arte de lecionar. Tem muito orgulho de ter criado na FIA cinco laboratórios para as aulas de Big Data e Inteligência Artificial. Possui mais de 20 anos de trajetória nas áreas de Data Mining, Big Data, Inteligência Artificial e Analytics. Cientista de dados com carreira realizada na Universidade de São Paulo. Graduada e mestra em Estatística Aplicada pelo IME-USP e doutora pela FEA-USP. Com muita dedicação chegou ao cargo de professora e pesquisadora na FEA-USP, ganhou mais de 30 prêmios de excelência acadêmica pela FEA-USP e mais de 30 prêmios de excelência acadêmica como professora dos cursos de MBA da FIA. Orienta alunos de mestrado e de doutorado na FEA-USP. Parecerista da FAPESP e colunista de grandes portais de tecnologia.

Currículo do Professor

PROF. ÂNGELO CHIODE, MSc

Bacharel, mestre e candidato ao PhD em Estatística (IME-USP), atua como professor de Estatística Aplicada para turmas de especialização, pós-graduação e MBA na FIA. Trabalha como consultor nas áreas de Analytics e Ciência de Dados há 13 anos, apoiando empresas na resolução de desafios de negócio nos contextos de finanças, adquirência, seguros, varejo, tecnologia, aviação, telecomunicações, entretenimento e saúde. Nos últimos 5 anos, tem atuado na gestão corporativa de times de Analytics, conduzindo projetos que envolviam análise estatística, modelagem preditiva e *machine learning*. É especializado em técnicas de visualização de dados e design da informação (Harvard) e foi indicado ao prêmio de Profissional do Ano na categoria Business Intelligence, em 2019, pela Associação Brasileira de Agentes Digitais (ABRADi).

Conteúdo Programático

DISCIPLINAS

- IA E TRANSFORMAÇÃO DIGITAL
- **ANALYTICS**
- INTELIGÊNCIA ARTIFICIAL: MACHINE LEARNING
- INTELIGÊNCIA ARTIFICIAL: DEEP LEARNING
- EMPREENDEDORISMO E INOVAÇÃO
- COMPORTAMENTO HUMANO E SOFT SKILLS

TEMAS: ANALYTICS E MACHINE LEARNING

- **ANÁLISE EXPLORATÓRIA DE DADOS**
- INFERÊNCIA ESTATÍSTICA
- TÉCNICAS DE PROJEÇÃO
- TÉCNICAS DE CLASSIFICAÇÃO
- TÓPICOS DE MODELAGEM
- TÉCNICAS DE SEGMENTAÇÃO
- TÓPICOS DE ANALYTICS
- MANIPULAÇÃO DE BASE DE DADOS
- AUTO ML

TEMAS: DEEP LEARNING

- REDES DENSAS
- REDES CONVOLUCIONAIS
- REDES RECORRENTES
- MODELOS GENERATIVOS

FERRAMENTAS

- **LINGUAGEM R**
- LINGUAGEM PYTHON
- DATABRICKS

FIA BUSINESS SCHOOL lab.data

Conteúdo da Aula

1. Introdução
2. Objetivo
3. Estruturação de Dados e Tipos de Variáveis
4. Análises de Unicidade e Preenchimento
5. Análises Univariadas: Variáveis Qualitativas
   i. Distribuição de Frequências
   ii. Gráficos: Barras e Setores
6. Análises Univariadas: Variáveis Quantitativas
   i. Medidas Resumo de Posição
   ii. Medidas Resumo de Dispersão
   iii. Gráficos: Histograma e *Boxplot*
7. Análises Bivariadas e Trivariadas
   i. Qualitativas *vs.* Qualitativas
   ii. Quantitativas *vs.* Quantitativas
   iii. Qualitativas *vs.* Quantitativas
8. *Case*

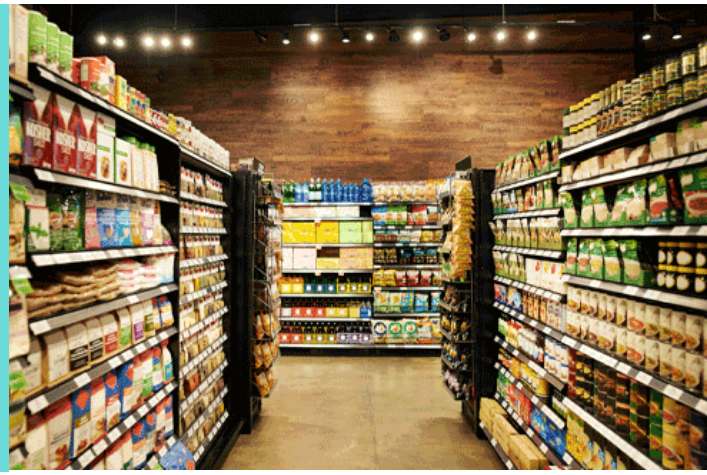
Referências Bibliográficas

1. Introdução

Case: Perfil de Compra em Supermercado

Exemplo:
Analisar características sociodemográficas e aspectos dos produtos comprados por consumidores nas lojas de um varejista nos últimos 6 meses, a fim de entender melhor as suas preferências e otimizar estratégias de marketing.

Aplicação:
Varejo alimentar (ou outros)



Case: Perfil de Clientes *Diamante*

Exemplo:
Descrever os principais aspectos transacionais e sociodemográficos dos clientes portadores da modalidade *diamante* de um cartão de crédito, a fim de compreender o seu perfil e características de consumo.

Aplicação:
Segmento bancário ou emissores de cartão de crédito

FIA BUSINESS SCHOOL lab.data

Case: Padrões de Investimento

Exemplo:
Avaliar os padrões de investimento e o comportamento de risco dos clientes em uma instituição financeira nos últimos 12 meses, para entender suas estratégias financeiras e prever tendências de mercado.

Aplicação:
Segmento bancário ou empresas de investimentos

Case: Preferências de Viagem

Exemplo:
Resumir as principais preferências de destino e os padrões de reserva dos viajantes nacionais no último *réveillon*, a fim de personalizar as ofertas e melhorar a experiência do cliente para o próximo ano.

Aplicação:
Turismo e hotelaria

Case: Perfil de Imóveis Residenciais

Exemplo:
Descrever as características dos imóveis residenciais disponíveis para venda em uma determinada cidade, a fim de compreender melhor as oportunidades imobiliárias existentes no local.

Aplicação:
Área imobiliária

2. Objetivo

Objetivo

O objetivo da **análise exploratória** consiste em **descrever** uma base de dados para extrair conhecimento a respeito do comportamento dos dados.

Trata-se de uma análise **preliminar** que, por boa prática, deve ser realizada antes de outras análises mais sofisticadas ou de modelagens preditivas.

A análise exploratória abrange, principalmente, as seguintes finalidades técnicas:

- avaliação de **consistência** da base de dados (unicidade, preenchimento, *outliers*).
- avaliação de **posição, centralidade** e **dispersão** de dados quantitativos.
- avaliação de **frequências** de dados qualitativos.

Vamos estudar os principais métodos de análise exploratória nesta aula, tanto sob a ótica **univariada** quanto **bivariada** e **trivariada**.

3. Estruturação de Dados e Tipos de Variáveis

Estruturação de Dados

Podem existir diferentes graus de **estruturação** de dados.

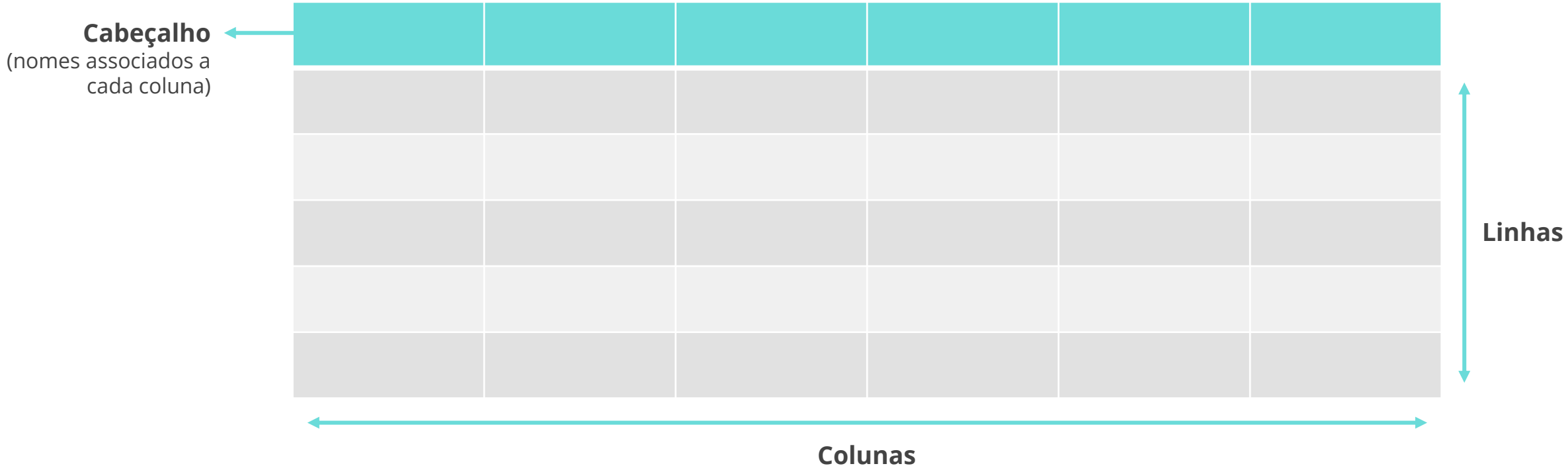
Em geral, costumamos trabalhar com dados **estruturados** para as análises do cotidiano.

Quando lidamos com dados **não estruturados**, técnicas mais avançadas de *machine learning* costumam ser mais efetivas. Ainda assim, a maioria delas envolve algum tipo de "estruturação" dos dados durante o processo.

FIA BUSINESS SCHOOL lab.data

Formato de Base de Dados Estruturados

O formato usual de uma base de dados estruturados é o formato **tabular**, ou seja, uma **tabela** com linhas e colunas.



Cada **linha** representa uma **observação**, de acordo com o nível de granularidade necessário para análise.
Exemplos: cada consumidor, cada cliente, cada respondente de uma pesquisa, cada empresa, cada cidade etc.

Cada **coluna** representa uma **variável**, que corresponde a uma característica das observações.
Exemplos, para clientes: CPF, idade, gênero, cidade de residência, renda mensal, tempo desde a última compra etc.

Formato de Base de Dados Estruturados

O formato usual de uma base de dados estruturados é o formato **tabular**, ou seja, uma **tabela** com linhas e colunas.

Cabeçalho
(nomes associados a cada coluna)

CPF	Idade	Genero	Cidade_Resid	Renda_Mensal	Tempo_Compra
11.111.111-11	49	F	São Paulo	5.500	5
22.222.222-22	56	M	Rio de Janeiro	12.900	1
33.333.333-33	32	F	Ribeirão Preto	*NA*	2
44.444.444-44	41	F	São Paulo	*NA*	2
55.555.555-55	24	M	Londrina	3.800	12

Linhas

Colunas

Cada **linha** representa uma **observação**, de acordo com o nível de granularidade necessário para análise.
Exemplos: cada consumidor, cada cliente, cada respondente de uma pesquisa, cada empresa, cada cidade etc.

Cada **coluna** representa uma **variável**, que corresponde a uma característica das observações.
Exemplos, para clientes: CPF, idade, gênero, cidade de residência, renda mensal, tempo desde a última compra etc.

Formato de Base de Dados Estruturados

As linhas também podem representar **datas distintas**: dias, quinzenas, meses, trimestres, anos etc.

Cabeçalho
(nomes associados a cada coluna)

Mes	Faturamento	Qtde_Vendas	Qtde_Lojas	Ticket_Medio	Faixa_Meta_Atingida
Jan-23	2.720.000	3.420	25	737	Mais de 100%
Fev-23	2.860.000	3.760	29	761	Mais de 100%
Mar-23	2.650.000	3.310	27	760	80% a 100%
Abr-23	2.810.000	3.860	28	751	80% a 100%
Mai-23	2.630.000	3.500	25	728	Menos de 80%

Linhas

Colunas

A essa estrutura, damos o nome de **dados de série temporal**.

As técnicas para análise de dados temporais são mais complexas do que o usual, pois aqui, as linhas da base de dados **não são independentes** entre si. Ou seja, os dados de cada linha estão potencialmente correlacionados com os dados de linhas próximas, por representarem instantes de tempo próximos entre si.

FIA BUSINESS SCHOOL lab.data

Tipos de Variáveis

Os principais **tipos de variáveis** em bases de dados tabulares são:

- Variáveis **identificadoras** (chaves)
- Variáveis **quantitativas**: discretas e contínuas
- Variáveis **qualitativas**: ordinais e nominais

Além desses, outros tipos específicos de **variáveis auxiliares** comuns são:

- Datas/horas
- Textos
- Localização

Tipos de Variáveis

Variáveis Identificadoras *(Chaves)*

Consistem em variáveis que auxiliam a **identificar os registros** de forma **única**.

Exemplos:

- Números de documento (ex.: RG, CPF, CNH)
- Códigos alfanuméricos gerados pelo administrador de um banco de dados (ex.: ID, “hash”)
- Combinações de campos numéricos ou não (ex.: nome da empresa + código da filial)

Em conformidade com a **Lei Geral de Proteção de Dados** (Lei Nº 13.709 de 2018), a fim de preservar a confidencialidade dos indivíduos, deve-se evitar utilizar números de documento como identificadores em bases de dados voltadas para análise estatística. O acesso a essas informações deve ser concedido apenas para times cujas atividades que necessitem estritamente delas, o que geralmente não é o caso dos analistas/cientistas de dados.

As empresas devem assegurar que possuem **base legal** para coletar, armazenar e processar dados potencialmente sensíveis de seus clientes, segundo as especificações da LGPD: consentimento do titular; cumprimento de obrigação legal; legítimo interesse etc.

Tipos de Variáveis

Variáveis Quantitativas Discretas

Representam características mensuráveis como **quantidades**.

Além disso, assumem valores num conjunto **finito** ou **enumerável** (ou seja, contável).

Exemplos:

- Quantidade de indivíduos que se identificam com o gênero feminino, num time de 30 profissionais.
  Valores possíveis: 0, 1, 2, 3, ... , 28, 29, 30 (finito).
- Quantidade de meses do ano em que um cliente comprou produto(s).
  Valores possíveis: 0, 1, 2, 3, ..., 10, 11, 12 (finito).
- Quantidade de clientes ativos de uma empresa.
  Valores possíveis: 0, 1, 2, 3, ... (infinito, mas enumerável).

Tipos de Variáveis

Variáveis Quantitativas Contínuas

Representam características mensuráveis como **quantidades**.

Além disso, assumem valores num conjunto **não enumerável** (ou seja, não contável).

Exemplos:

- Idade de um profissional adulto.
  Valores possíveis: 18, ... , 18,001, ... , 18,01, ..., 19, ..., 19,001, ... (não enumerável).
- Preço do litro da gasolina em determinado dia, em diferentes postos de abastecimento.
  Valores possíveis: R$ 4, ... , R$ 4,01, ... , R$ 4,02, ... , R$ 4,03, ... (não enumerável).

O fato de uma variável quantitativa contínua poder ser **arredondada** (por exemplo: idade **em anos**) não faz com que ela se torne discreta. Na prática, qualquer variável com casas decimais terá que ser arredondada para ser **armazenada** em uma base de dados. A diferenciação entre variáveis discretas e contínuas reside na sua **natureza**, ou seja, na possibilidade de assumir ou não valores em um conjunto não enumerável, ainda que de forma hipotética.

Tipos de Variáveis

Variáveis Qualitativas Ordinais

Representam características mensuráveis como **qualidades**, ou seja, **categorias**.

Além disso, assumem categorias que possuem **ordem natural** entre si.

Exemplos:

- Nível de satisfação de um cliente.
  Valores possíveis: muito satisfeito, satisfeito, neutro, insatisfeito, muito insatisfeito.
- Nível de glicose presente no sangue de um paciente, em jejum.
  Valores possíveis: acima do esperado, dentro do esperado, abaixo do esperado.

Tipos de Variáveis

Variáveis Qualitativas Nominais

Representam características mensuráveis como **qualidades**, ou seja, **categorias**.

Além disso, assumem categorias que **não possuem ordem natural** entre si.

Exemplos:

- Último produto adquirido por um cliente de uma seguradora.
  Valores possíveis: seguro residência, seguro auto, seguro de vida, etc.
- Região de localização de cada filial de uma empresa.
  Valores possíveis: Sul, Sudeste, Nordeste, Norte, Centro-Oeste.

Tipos de Variáveis

Variáveis Auxiliares

- **Datas/horas**

  São utilizadas para realizar filtros e agrupamentos de períodos, ou para cálculo de outras variáveis (exemplos: idade, tempo de relacionamento, tempo desde o último chamado aberto).

- **Textos**

  São utilizados para realização de análises específicas que resumam o seu conteúdo, tais como nuvens de palavras frequentes e modelos de processamento de linguagem natural (NLP).

- **Localização**

  São informações úteis para calcular variáveis relacionadas a distâncias entre duas localizações (exemplo: distância entre o endereço de residência declarado pelo cliente e o endereço da loja mais próxima).
  As mais comuns são *latitude* e *longitude*.

Tipos de Variáveis

Quais os tipos das variáveis envolvidas neste exemplo?

CPF	Idade	Genero	Cidade_Resid	Renda_Mensal	Tempo_Compra
11.111.111-11	49	F	São Paulo	5.500	5
22.222.222-22	56	M	Rio de Janeiro	12.900	1
33.333.333-33	32	F	Ribeirão Preto	*NA*	2
44.444.444-44	41	F	São Paulo	*NA*	2
55.555.555-55	24	M	Londrina	3.800	12
Chave	Quantitativa contínua	Qualitativa nominal	Qualitativa nominal	Quantitativa contínua	Quantitativa contínua

Tipos de Variáveis

Quais os tipos das variáveis envolvidas neste exemplo?

Mes	Faturamento	Qtde_Vendas	Qtde_Lojas	Ticket_Medio	Faixa_Meta_Atingida
Jan-23	2.720.000	3.420	25	737	Mais de 100%
Fev-23	2.860.000	3.760	29	761	Mais de 100%
Mar-23	2.650.000	3.310	27	760	80% a 100%
Abr-23	2.810.000	3.860	28	751	80% a 100%
Mai-23	2.630.000	3.500	25	728	Menos de 80%
Data	Quantitativa contínua	Quantitativa contínua	Quantitativa discreta	Quantitativa contínua	Qualitativa ordinal

Case: Perfil de Imóveis Residenciais

Arquivo: Imoveis.txt

Vamos prosseguir com o *case* de **avaliação de perfil de imóveis residenciais**.

Trata-se de uma base de dados com **561 observações** (linhas) e **9 variáveis** (colunas).

- ✓ **ID_IMOVEL**
  Código de identificação do imóvel (número sequencial)
- ✓ **BAIRRO_IMOVEL**
  Bairro do imóvel *(Santa Rosa, Vila Verde, Recanto Mar, Jardim Sol)*
- ✓ **METRAGEM**
  Metragem do imóvel, em m²
- ✓ **TIPO_IMOVEL**
  Indicação do tipo do imóvel *(casa, apartamento)*
- ✓ **VALOR_VENDA**
  Valor de venda do imóvel, em R$
- ✓ **INCIDENCIA_LUZ**
  Incidência de luz solar ao longo do dia *(nenhuma, pouca, muita)*
- ✓ **VAGAS_GARAGEM**
  Quantidade de vagas de garagem
- ✓ **FLUXO_VEICULOS**
  Nível do fluxo de veículos na rua do imóvel (*baixo*, *intermediário, intenso*)
- ✓ **COMERCIOS_RAIO_1KM**
  Quantidade de estabelecimentos comerciais próximos, num raio de até 1km

FIA BUSINESS SCHOOL lab.data

Case: Perfil de Imóveis Residenciais

Quais são os **tipos das variáveis** no *case* de perfil de imóveis residenciais?

ID_IMOVEL	BAIRRO_IMOVEL	METRAGEM	TIPO_IMOVEL	VALOR_VENDA	INCIDENCIA_LUZ	VAGAS_GARAGEM	FLUXO_VEICULOS	COMERCIOS_RAIO_1KM
#001	Santa Rosa	140	Apartamento	517000	Pouca	2	Intenso	19
#002	Santa Rosa	90	Apartamento	427000	Pouca	2	Intermediário	13
#003	Santa Rosa	130	Casa	558000	Pouca	5	Baixo	11
#004	Recanto Mar	50	Apartamento	354000	Pouca	2	Intenso	
#005	Recanto Mar	70	Apartamento	416000	Nenhuma		Intermediário	26
#006	Jardim Sol	110	Apartamento	468000	Muita		Intenso	27
#007	Santa Rosa	90	Apartamento	436000	Nenhuma		Intermediário	18
#008	Jardim Sol	240	Casa	509000	Nenhuma	1	Intermediário	13
#009	Recanto Mar	210	Casa	692000	Pouca	4	Intenso	24
#010	Santa Rosa	70	Apartamento	395000	Nenhuma	2	Intenso	31
#011	Jardim Sol	110	Apartamento	416000	Pouca	1	Baixo	17
#012	Santa Rosa	260	Casa	462000	Nenhuma	5	Intermediário	11
#013	Jardim Sol	50	Apartamento	403000	Nenhuma		Intenso	11
#014	Vila Verde	260	Casa	576000	Muita	5	Intenso	6
#015	Jardim Sol	110	Casa	407000	Muita	1	Intenso	14
#016	Recanto Mar	260	Casa	615000	Muita	5	Baixo	15
#017	Santa Rosa	290	Casa	765000	Muita		Baixo	7
#018	Recanto Mar	200	Casa	610000	Pouca	2	Intermediário	14
#019	Santa Rosa	50	Apartamento	402000	Pouca	1	Intermediário	
#020	Recanto Mar	210	Casa	462000	Pouca	2	Intenso	28

Arquivo: Imoveis.txt

Case: Perfil de Imóveis Residenciais

Quais são os **tipos das variáveis** no *case* de perfil de imóveis residenciais?

ID_IMOVEL	BAIRRO_IMOVEL	METRAGEM	TIPO_IMOVEL	VALOR_VENDA	INCIDENCIA_LUZ	VAGAS_GARAGEM	FLUXO_VEICULOS	COMERCIOS_RAIO_1KM
#001	Santa Rosa	140	Apartamento	517000	Pouca	2	Intenso	19
#002	Santa Rosa	90	Apartamento	427000	Pouca	2	Intermediário	13
#003	Santa Rosa	130	Casa	558000	Pouca	5	Baixo	11
#004	Recanto Mar	50	Apartamento	354000	Pouca	2	Intenso	
#005	Recanto Mar	70	Apartamento	416000	Nenhuma		Intermediário	26
#006	Jardim Sol	110	Apartamento	468000	Muita		Intenso	27
#007	Santa Rosa	90	Apartamento	436000	Nenhuma		Intermediário	18
#008	Jardim Sol	240	Casa	509000	Nenhuma	1	Intermediário	13
#009	Recanto Mar	210	Casa	692000	Pouca	4	Intenso	24
#010	Santa Rosa	70	Apartamento	395000	Nenhuma	2	Intenso	31
#011	Jardim Sol	110	Apartamento	416000	Pouca	1	Baixo	17
#012	Santa Rosa	260	Casa	462000	Nenhuma	5	Intermediário	11
#013	Jardim Sol	50	Apartamento	403000	Nenhuma		Intenso	11
#014	Vila Verde	260	Casa	576000	Muita	5	Intenso	6
#015	Jardim Sol	110	Casa	407000	Muita	1	Intenso	14
#016	Recanto Mar	260	Casa	615000	Muita	5	Baixo	15
#017	Santa Rosa	290	Casa	765000	Muita		Baixo	7
#018	Recanto Mar	200	Casa	610000	Pouca	2	Intermediário	14
#019	Santa Rosa	50	Apartamento	402000	Pouca	1	Intermediário	
#020	Recanto Mar	210	Casa	462000	Pouca	2	Intenso	28
Chave	**Qualitativa nominal**	**Quantitativa contínua**	**Qualitativa nominal**	**Quantitativa contínua**	**Qualitativa ordinal**	**Quantitativa discreta**	**Qualitativa ordinal**	**Quantitativa discreta**

Arquivo: Imoveis.txt

FIA BUSINESS SCHOOL lab.data

4. Análises de Unicidade e Preenchimento

Análise de Unicidade

Duas análises preliminares importantes em toda base de dados são as análises de **unicidade** e **preenchimento**.

Análise de unicidade

Consiste em verificar se cada observação aparece uma **única vez** na base de dados, ou se existem **repetições indevidas**. Essa análise costuma ser realizada a partir de algum campo **chave** que identifica as observações.

Exemplo: em uma base cadastral de clientes identificados pelo "ID do cliente", cada ID deve aparecer uma única vez.

É possível que haja duplicações em **níveis de granularidade maiores** do que a visão unitária das observações da base de dados.

Exemplo: em uma base de transações, cada "ID do cliente" pode aparecer mais de uma vez, pois um mesmo cliente pode ter realizado várias transações. Entretanto, cada "ID de transação" deve aparecer uma única vez.

No **Excel**, a análise de unicidade pode ser realizada copiar e colando os valores do campo chave em uma planilha à parte, e utilizando a funcionalidade *"Remover Duplicadas"*. Caso nenhuma observação seja removida, isso significa que a base de dados respeita o princípio da unicidade.

Case: Perfil de Imóveis Residenciais

Duas análises preliminares importantes em toda base de dados são as análises de **unicidade** e **preenchimento**.

Análise de unicidade

Realizando a análise de unicidade na coluna “ID_IMOVEL” para o *case* de perfil de imóveis residenciais, verificamos que todos os ID’s são distintos. Ou seja, **não há repetições** de imóveis na base de dados.

Arquivo: Imoveis.txt

Análise de Preenchimento

Duas análises preliminares importantes em toda base de dados são as análises de **unicidade** e **preenchimento**.

Análise de preenchimento

Consiste em contabilizar a quantidade de observações (linhas) que possuem valores **preenchidos** para as variáveis de interesse (colunas); e, consequentemente, os valores **ausentes** (ou ***missing values***).

Os valores ausentes podem ter diferentes **significados** em uma variável:

1. **Indisponibilidade/desconhecimento** acerca do preenchimento.
   →É chamado de "valor ausente não informativo", ou seja, não há o que fazer para obter a informação.
2. **Não aplicabilidade** de preenchimento.
   →É chamado de "valor ausente informativo", ou seja, o fato de não haver preenchimento traz alguma informação.
3. **Inconsistência** no processo de cálculo/extração da base de dados.
   →É um problema que podemos tentar resolver recorrendo ao administrador da base de dados.
4. **Equivalência** com algum valor numérico
   →Pode ser substituído por valor numérico (em geral, zero), caso isso possa ser assumido com segurança.

Case: Perfil de Imóveis Residenciais

Duas análises preliminares importantes em toda base de dados são as análises de **unicidade** e **preenchimento**.

Análise de preenchimento

Por meio da função CONTAR.VAZIO() do **Excel**, realizamos a contagem de valores ausentes nas variáveis do *case* de perfil de imóveis residenciais. Obtivemos os seguintes resultados:

Variável	Qtde. de ausentes
BAIRRO_IMOVEL	0 (0%)
METRAGEM	0 (0%)
TIPO_IMOVEL	0 (0%)
VALOR_VENDA	0 (0%)
INCIDENCIA_LUZ	0 (0%)
VAGAS_GARAGEM	61 (11%)
FLUXO_VEICULOS	0 (0%)
COMERCIOS_RAIO_1KM	96 (17%)

Quais os possíveis **significados** para os *missings* nestas variáveis?

Arquivo: Imoveis.txt

5. Análises Univariadas: Variáveis Qualitativas

Dimensões de Análise

Vamos começar as análises principais para geração de *insights* a partir dos dados.

Antes disso, vale a pena ressaltar as diferentes dimensões de análise que realizaremos ao longo do curso:

Análises univariadas

- ✓ Análises descritivas/exploratórias, envolvendo **uma única variável por vez**

Análises bivariadas e trivariadas

- ✓ Análises descritivas/exploratórias, relacionando **duas ou três variáveis entre si**
- ✓ Análises inferenciais (modelos), para predizer **uma variável a partir de outra** FUTURO

Análises multivariadas

- ✓ Análises inferenciais (modelos), para predizer **uma variável a partir de várias outras** FUTURO

Análises Univariadas: Variáveis Qualitativas

As principais técnicas para análise **univariada** de variáveis **qualitativas** são:

- ✓ Tabelas de **distribuição de frequências** (absolutas e/ou relativas)
- ✓ Gráficos de **barras** e/ou **setores**

Distribuição de Frequências

A tabela de **distribuição de frequências** exibe as frequências **absolutas** (quantidades) de observações em cada categoria de uma variável qualitativa. Pode exibir, também, as frequências **relativas** (percentuais/porcentagens).

No exemplo abaixo, apresenta-se uma tabela de distribuição de frequências acerca da **incidência de luz solar** em 561 imóveis, para o *case* de perfil de imóveis residenciais.

Incidência de luz solar	Frequência absoluta	Frequência relativa	Porcentagem
Nenhuma	155	0,276	27,6%
Pouca	187	0,333	33,3%
Muita	219	0,390	39,0%
Total	**561**	**1,000**	**100%**

→ Porcentagem = frequência relativa x 100

↓ A soma das frequências relativas é igual a 1

No Excel: **tabelas dinâmicas**
CONT.SE()
No R: **table()**
prop.table()

Arquivo: Imoveis.txt

FIA BUSINESS SCHOOL lab.data

Gráficos: Barras e Setores

A mesma noção de **distribuição de frequências** pode ser representada de forma gráfica, em geral, por meio de gráficos de barras (simples ou empilhadas) e/ou de setores.

Gráfico de barras simples

Incidência de luz solar

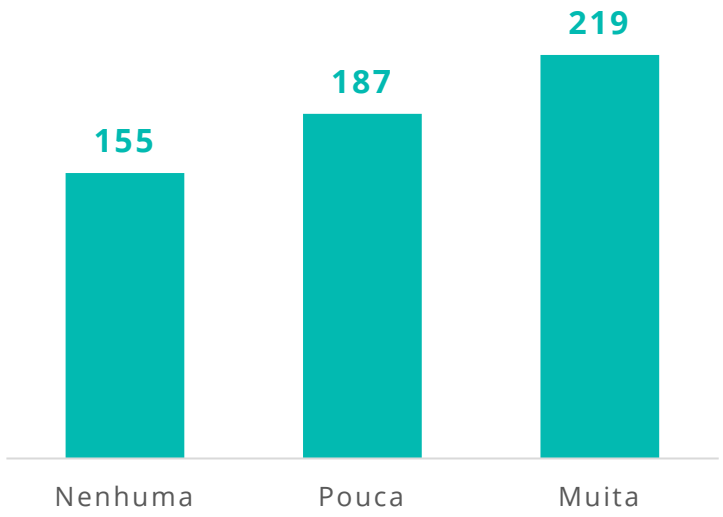


Gráfico de barras empilhadas

Incidência de luz solar

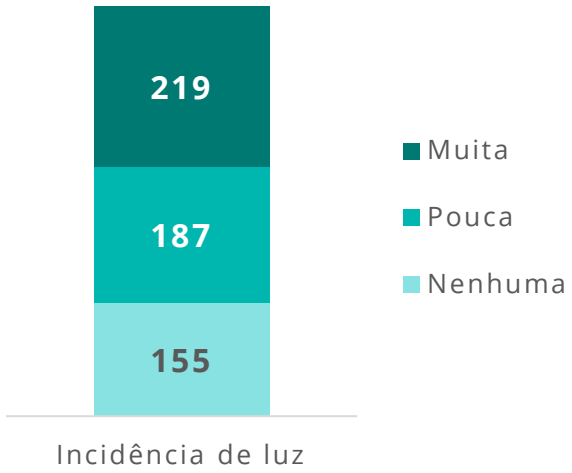
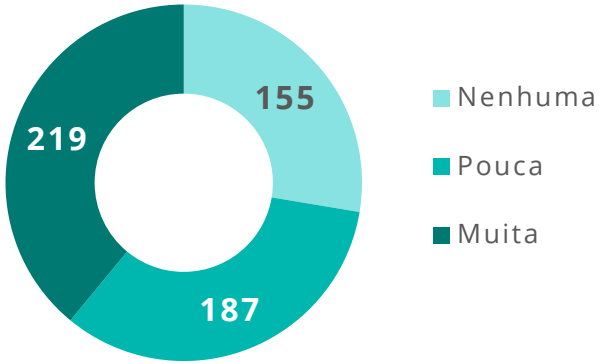


Gráfico de setores

Incidência de luz solar



Em geral, gráficos de **barras** são mais efetivos para comparação de frequências do que gráficos de setores, pois envolvem a noção de **comprimento** em vez de noções de áreas e ângulos.

Arquivo: Imoveis.txt

FIA BUSINESS SCHOOL lab.data

Gráficos: Barras e Setores

A mesma noção de **distribuição de frequências** pode ser representada de forma gráfica, em geral, por meio de gráficos de barras (simples ou empilhadas) e/ou de setores.

Gráfico de barras simples

Incidência de luz solar

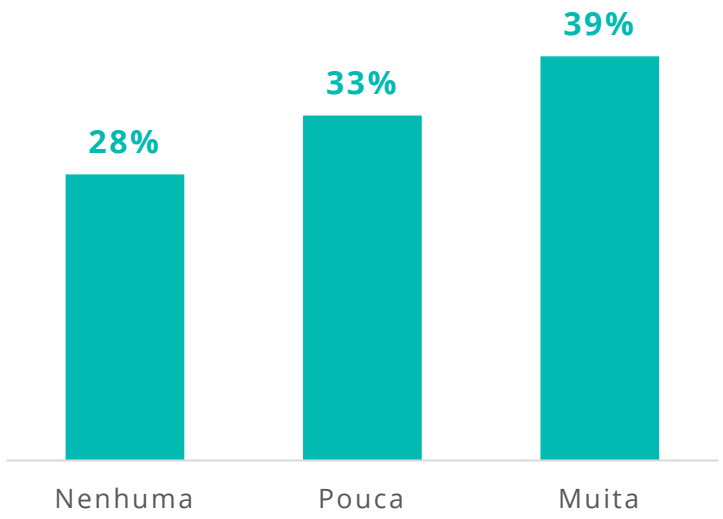


Gráfico de barras empilhadas

Incidência de luz solar

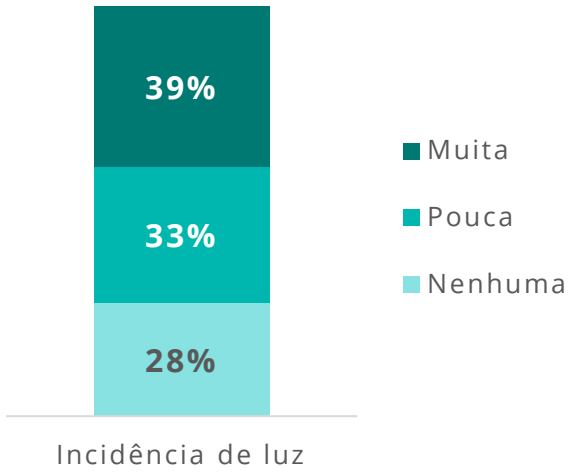
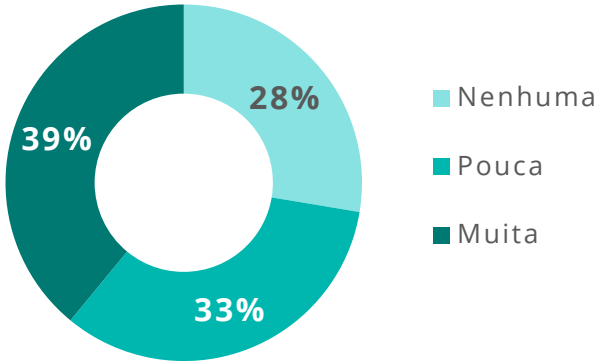


Gráfico de setores

Incidência de luz solar



Em geral, gráficos de **barras** são mais efetivos para comparação de frequências do que gráficos de setores, pois envolvem a noção de **comprimento** em vez de noções de áreas e ângulos.

Arquivo: Imoveis.txt

FIA BUSINESS SCHOOL lab.data

6. Análises Univariadas: Variáveis Quantitativas

Análises Univariadas: Variáveis Quantitativas

Caso a variável seja **quantitativa discreta** e admita **poucos** valores distintos, as mesmas técnicas utilizadas para variáveis qualitativas podem ser aplicadas, ou seja:

- ✓ Tabelas de **distribuição de frequências** (absolutas e/ou relativas)
- ✓ Gráficos de **barras** e/ou **setores**

Análises Univariadas: Variáveis Quantitativas

Já para variáveis **quantitativas discretas** que assumem **muitos** valores possíveis, ou para variáveis **quantitativas contínuas**, as principais técnicas para análise **univariada** são:

- ✓ **Medidas resumo** de **posição**
  - *Média*
  - *Mediana*
  - *Moda*
  - *Mínimo e máximo*
  - *Quartis*
  - *Percentis*
- ✓ **Medidas resumo** de **dispersão**
  - *Variância*
  - *Desvio padrão*
  - *Coeficiente de variação*
  - *Amplitude*
  - *Intervalo interquartil*
- ✓ Gráficos de **histograma** e ***boxplot***

FIA BUSINESS SCHOOL lab.data

Análises Univariadas: Variáveis Quantitativas

Já para variáveis **quantitativas discretas** que assumem **muitos** valores possíveis, ou para variáveis **quantitativas contínuas**, as principais técnicas para análise **univariada** são:

- ✓ **Medidas resumo** de **posição**
  - *Média*
  - *Mediana*
  - *Moda*
  - *Mínimo e máximo*
  - *Quartis*
  - *Percentis*
- ✓ **Medidas resumo** de **dispersão**
  - *Variância*
  - *Desvio padrão*
  - *Coeficiente de variação*
  - *Amplitude*
  - *Intervalo interquartil*
- ✓ Gráficos de **histograma** e ***boxplot***

FIA BUSINESS SCHOOL lab.data

Medidas Resumo de Posição

As **medidas resumo** de **posição** são cálculos que trazem uma informação sumarizada a respeito do patamar de valores que uma variável quantitativa assume.

Motivação: como trazer informações **resumidas** a respeito dos valores que a variável METRAGEM assume, no *case* de perfil de imóveis residenciais?

ID_IMOVEL	METRAGEM
#001	140
#002	90
#003	130
#004	50
#005	70
#006	110
#007	90
#008	240
#009	210
#010	70
#011	110
#012	260
#013	50
#014	260
#015	110
#016	260
#017	290
#018	200
#019	50
#020	210

ID_IMOVEL	METRAGEM
#021	250
#022	40
#023	290
#024	80
#025	270
#026	140
#027	40
#028	140
#029	200
#030	110
#031	90
#032	110
#033	260
#034	180
#035	60
#036	100
#037	270
#038	240
#039	170
#040	140

ID_IMOVEL	METRAGEM
#041	80
#042	140
#043	70
#044	90
#045	90
#046	250
#047	190
#048	130
#049	260
#050	110
#051	280
#052	50
#053	70
#054	130
#055	130
#056	90
#057	130
#058	60
#059	250
#060	110
...	...

Média

A **média** é uma medida resumo acerca da **centralidade** dos dados. Ou seja, é um valor de referência que representa o principal patamar em torno do qual os dados estão concentrados.

Racional de cálculo:

$$\text{Média} = \frac{\text{Soma dos valores de interesse}}{\text{Quantidade de valores de interesse}}$$

Fórmula:

$$\frac{\sum_{i=1}^{n} x_i}{n}$$

i é um índice que representa cada observação
n é a quantidade de observações

Média

Qual é a média de **metragem** (em m²) dos 561 imóveis no *case* de perfis de imóveis residenciais?

$$\text{Média} = \frac{140 + 90 + 130 + \ \ldots}{561} = 123{,}2$$

Em **média**, os imóveis da base de dados têm cerca de **123 m²** de metragem.

ID_IMOVEL	METRAGEM
#001	140
#002	90
#003	130
#004	50
#005	70
#006	110
#007	90
#008	240
#009	210
#010	70
#011	110
#012	260
#013	50
#014	260
#015	110
#016	260
#017	290
#018	200
#019	50
#020	210

ID_IMOVEL	METRAGEM
#021	250
#022	40
#023	290
#024	80
#025	270
#026	140
#027	40
#028	140
#029	200
#030	110
#031	90
#032	110
#033	260
#034	180
#035	60
#036	100
#037	270
#038	240
#039	170
#040	140

ID_IMOVEL	METRAGEM
#041	80
#042	140
#043	70
#044	90
#045	90
#046	250
#047	190
#048	130
#049	260
#050	110
#051	280
#052	50
#053	70
#054	130
#055	130
#056	90
#057	130
#058	60
#059	250
#060	110
...	...

No Excel: **MÉDIA()**
No R: **mean()**

Arquivo: Imoveis.txt

FIA BUSINESS SCHOOL lab.data

Mediana

A **mediana** é outra medida resumo acerca da **centralidade** dos dados, com uma proposta diferente da média. Ela consiste no valor central que subdivide **50%** dos valores mais baixos dos demais **50%** dos valores mais altos.

Racional de cálculo:

$$\text{Mediana} = \text{valor que subdivide os dados em dois conjuntos com } 50\% \text{ do total de valores}$$

Fórmula:

Se n ímpar: $x_{\frac{n+1}{2}}$

Se n par: $\frac{(x_{n/2}+x_{n/2+1})}{2}$

x_i representa o i-ésimo valor de interesse, após ordenação do menor para o maior

Mediana

Qual é a mediana de **metragem** (em m²) dos 561 imóveis no *case* de perfis de imóveis residenciais?

$$\text{Mediana} = 110$$

Em **mediana**, os imóveis da base de dados têm **110 m²** de metragem.

ID_IMOVEL	METRAGEM
#001	140
#002	90
#003	130
#004	50
#005	70
#006	110
#007	90
#008	240
#009	210
#010	70
#011	110
#012	260
#013	50
#014	260
#015	110
#016	260
#017	290
#018	200
#019	50
#020	210

ID_IMOVEL	METRAGEM
#021	250
#022	40
#023	290
#024	80
#025	270
#026	140
#027	40
#028	140
#029	200
#030	110
#031	90
#032	110
#033	260
#034	180
#035	60
#036	100
#037	270
#038	240
#039	170
#040	140

ID_IMOVEL	METRAGEM
#041	80
#042	140
#043	70
#044	90
#045	90
#046	250
#047	190
#048	130
#049	260
#050	110
#051	280
#052	50
#053	70
#054	130
#055	130
#056	90
#057	130
#058	60
#059	250
#060	110
...	...

No Excel: **MED()**
No R: **median()**

Arquivo: Imoveis.txt

FIA BUSINESS SCHOOL lab.data

Mediana

Qual é a principal **vantagem** da mediana em relação à média?

Exemplo 1: qual é a média e a mediana para o conjunto de valores $x = (1, 2, 3, 4, 5, 6, 7, 8, 9)$?

$$\text{Média} = \mathbf{5} \text{ e } \text{Mediana} = \mathbf{5}$$

Exemplo 2: qual é a média e a mediana para o conjunto de valores $x = (1, 2, 3, 4, 5, 6, 7, 8, \mathbf{18})$?

$$\text{Média} = \mathbf{6} \text{ e } \text{Mediana} = \mathbf{5}$$

Exemplo 3: qual é a média e a mediana para o conjunto de valores $x = (1, 2, 3, 4, 5, 6, 7, 8, \mathbf{900})$?

$$\text{Média} = \mathbf{104} \text{ e } \text{Mediana} = \mathbf{5}$$

O valor da média é muito influenciado por **valores extremos**, ao contrário da mediana, que é mais **resistente**. Por isso, durante a análise exploratória, é conveniente calcular tanto a **média** quanto a **mediana** das variáveis quantitativas.

Moda

A **moda** é uma medida que denota o **valor mais comum** no conjunto dos dados, ou seja, aquele que mais ocorre. Pode ser calculada para quaisquer variáveis quantitativas que tenham **baixa diversidade** de valores observados.

Racional de cálculo:

$$\text{Moda} = \text{valor que mais aparece no conjunto}$$

Moda

Qual é a moda de **metragem** (em m²) dos 561 imóveis no *case* de perfis de imóveis residenciais?

$$\text{Moda} = 70$$

A **moda** de metragem dos imóveis da base de dados é de **70 m²**.

30	40	50	60	70
1	23	39	35	55
0.002	0.041	0.070	0.062	0.098

80	90	100	110	120
47	37	37	30	38
0.084	0.066	0.066	0.053	0.068

130	140	150	160	170
38	37	11	12	8
0.068	0.066	0.020	0.021	0.014

180	190	200	210	220
8	10	11	13	7
0.014	0.018	0.020	0.023	0.012

230	240	250	260	270
5	8	10	14	9
0.009	0.014	0.018	0.025	0.016

280	290	300
10	7	1
0.018	0.012	0.002

No Excel: **MODO.ÚNICO()**
No R: **names(sort(-table()))[1]**

Arquivo: Imoveis.txt

FIA BUSINESS SCHOOL lab.data

Mínimo e Máximo

O **mínimo** e o **máximo** são medidas resumo que representam os valores **mais extremos** acerca dos dados. Correspondem ao menor e ao maior valor de uma variável quantitativa, respectivamente.

Racional de cálculo:

$$\text{Mínimo} = \text{menor valor do conjunto} \qquad \text{Máximo} = \text{maior valor do conjunto}$$

Mínimo e Máximo

Qual é o mínimo e o máximo de **metragem** (em m²) dos 561 imóveis no *case* de perfis de imóveis residenciais?

$$\text{Mínimo} = 30 \qquad \text{Máximo} = 300$$

Ou seja, o menor imóvel da base de dados possui **30 m²** de metragem, e o maior imóvel possui **300 m²** de metragem.

ID_IMOVEL	METRAGEM
#371	300
#384	30

No Excel: **MÍNIMO()**
MÁXIMO()
No R: **min()**
max()

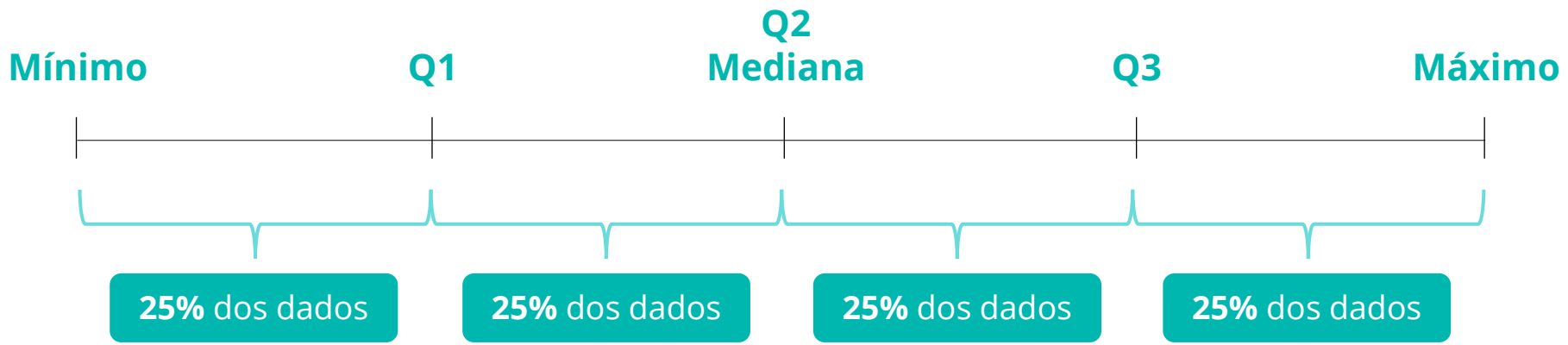
Arquivo: Imoveis.txt

Quartis

Os **quartis** são generalizações da mediana que avaliam outros cenários de **fatiamento** dos dados. Em vez de dividir os dados em duas partes com 50% das observações, divide-se em **quatro partes** com **25% das observações**.

Racional de cálculo:

$$\text{Quartil 1 (ou Q1)} = \text{valor que separa os 25\% menores valores em relação aos 75\% maiores valores}$$
$$\text{Quartil 2 (ou Q2)} = \text{valor que separa os 50\% menores valores em relação aos 50\% maiores valores}$$
$$\text{Quartil 3 (ou Q3)} = \text{valor que separa os 75\% menores valores em relação aos 25\% maiores valores}$$



Note que o **quartil 2** é correspondente à **mediana**, por definição.

Quartis

Quais são os quartis de **metragem** (em m²) dos 561 imóveis no *case* de perfis de imóveis residenciais?

$$\text{Quartil } 1 = 70 \qquad \text{Quartil } 2 = 110 \qquad \text{Quartil } 3 = 150$$

Ou seja, os 25% menores imóveis possuem metragens **até 70 m²**; e os 25% maiores imóveis possuem metragens **a partir de 150 m²**.

ID_IMOVEL	METRAGEM
#001	140
#002	90
#003	130
#004	50
#005	70
#006	110
#007	90
#008	240
#009	210
#010	70
#011	110
#012	260
#013	50
#014	260
#015	110
#016	260
#017	290
#018	200
#019	50
#020	210

ID_IMOVEL	METRAGEM
#021	250
#022	40
#023	290
#024	80
#025	270
#026	140
#027	40
#028	140
#029	200
#030	110
#031	90
#032	110
#033	260
#034	180
#035	60
#036	100
#037	270
#038	240
#039	170
#040	140

ID_IMOVEL	METRAGEM
#041	80
#042	140
#043	70
#044	90
#045	90
#046	250
#047	190
#048	130
#049	260
#050	110
#051	280
#052	50
#053	70
#054	130
#055	130
#056	90
#057	130
#058	60
#059	250
#060	110
...	...

Arquivo: Imoveis.txt

No Excel: **QUARTIL.INC(..., 1)**
QUARTIL.INC(..., 2)
QUARTIL.INC(..., 3)
No R: **quantile()**

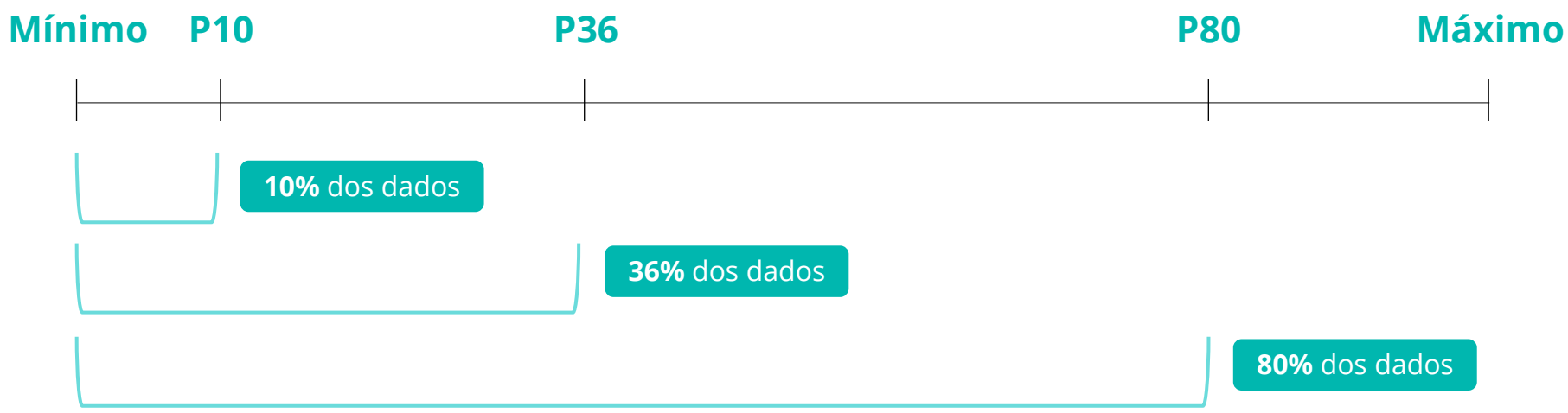
FIA BUSINESS SCHOOL lab.data

Percentis

Os **percentis** são generalizações da mediana que avaliam outros cenários de **fatiamento** dos dados. Agora, podemos dividir os dados em partes com **quaisquer representatividades** que se tenha interesse.

Racional de cálculo:

Percentil 10 (ou P10) = valor que separa os 10% menores valores em relação aos 90% maiores valores
Percentil 36 (ou P36) = valor que separa os 36% menores valores em relação aos 64% maiores valores
Percentil 80 (ou P80) = valor que separa os 80% menores valores em relação aos 20% maiores valores



Percentis

É comum utilizar os **percentis 1** e **99** como medidas alternativas para os valores **mínimo** e **máximo**, por serem menos influenciadas por valores **extremos** (tal como na comparação entre média e mediana).

Racional de cálculo:

$$\text{Percentil } 1 \text{ (ou P01)} = \text{valor que separa os } 1\% \text{ menores valores em relação aos } 99\% \text{ maiores valores}$$
$$\text{Percentil } 99 \text{ (ou P99)} = \text{valor que separa os } 99\% \text{ menores valores em relação aos } 1\% \text{ maiores valores}$$

BUSINESS SCHOOL lab.data

Percentis

Quais são os percentis 1 e 99 de **metragem** (em m²) dos 561 imóveis no *case* de perfis de imóveis residenciais?

$$\text{Percentil } 1 = 40 \qquad \text{Percentil } 99 = 290$$

Ou seja, os 1% menores imóveis possuem metragens **até 40 m²**; e os 1% maiores imóveis possuem metragens **a partir de 290 m²**.

Lembre-se que os valores mínimo e máximo de metragem eram:

$$\text{Mínimo} = 30 \qquad \text{Máximo} = 300$$

Neste caso, os valores dos percentis 1 e 99 não foram muito distantes dos valores mínimo e máximo, respectivamente. Porém, se alterássemos a metragem de **um único imóvel** de **300 m²** para **3.000 m²** (supondo um erro de digitação no registro), os novos valores seriam:

$$\text{Percentil } 99 = 290 \text{ (sem alteração)}$$
$$\text{Máximo} = 3.000 \text{ (distorcido)}$$

No Excel: **PERCENTIL.INC(...)**
No R: **quantile(..., probs = .)**

Arquivo: Imoveis.txt

Análises Univariadas: Variáveis Quantitativas

Já para variáveis **quantitativas discretas** que assumem **muitos** valores possíveis, ou para variáveis **quantitativas contínuas**, as principais técnicas para análise **univariada** são:

- ✓ **Medidas resumo** de **posição**
  - *Média*
  - *Mediana*
  - *Moda*
  - *Mínimo e máximo*
  - *Quartis*
  - *Percentis*
- ✓ **Medidas resumo** de **dispersão**
  - *Variância*
  - *Desvio padrão*
  - *Coeficiente de variação*
  - *Amplitude*
  - *Intervalo interquartil*
- ✓ Gráficos de **histograma** e ***boxplot***

FIA BUSINESS SCHOOL lab.data

Medidas Resumo de Dispersão

As **medidas resumo** de **dispersão** são cálculos que trazem uma informação sumarizada a respeito do grau de variabilidade dos valores que uma variável quantitativa assume.

Motivação: como trazer informações **resumidas** a respeito da variabilidade de valores da variável METRAGEM, a fim de sabermos se existe **pouca** ou **muita** heterogeneidade de metragens entre os imóveis?

ID_IMOVEL	METRAGEM
#001	140
#002	90
#003	130
#004	50
#005	70
#006	110
#007	90
#008	240
#009	210
#010	70
#011	110
#012	260
#013	50
#014	260
#015	110
#016	260
#017	290
#018	200
#019	50
#020	210

ID_IMOVEL	METRAGEM
#021	250
#022	40
#023	290
#024	80
#025	270
#026	140
#027	40
#028	140
#029	200
#030	110
#031	90
#032	110
#033	260
#034	180
#035	60
#036	100
#037	270
#038	240
#039	170
#040	140

ID_IMOVEL	METRAGEM
#041	80
#042	140
#043	70
#044	90
#045	90
#046	250
#047	190
#048	130
#049	260
#050	110
#051	280
#052	50
#053	70
#054	130
#055	130
#056	90
#057	130
#058	60
#059	250
#060	110
...	...

Variância

A **variância** é uma medida resumo acerca da **variabilidade** dos dados em torno do valor da **média**.
Quanto maior a variância, maior a heterogeneidade de valores que a variável assume.

Racional de cálculo:

$$\text{Variância} = \text{Média das diferenças, ao quadrado, entre: (i) cada valor do conjunto e (ii) a média geral}$$

Fórmula amostral:

$$\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

i é um índice que representa cada observação da amostra
n é a quantidade de observações na amostra
$\overline{x}$ é a média amostral

Variância

A **variância** é uma medida resumo acerca da **variabilidade** dos dados em torno do valor da **média**.
Quanto maior a variância, maior a heterogeneidade de valores que a variável assume.

Racional de cálculo:

$$\text{Variância} = \text{Média das diferenças, ao quadrado, entre: (i) } \ldots \text{ média geral}$$

Fórmula amostral:

$$\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

Caso se trate do desvio padrão populacional, substituímos o denominador por n. O uso do denominador $n-1$ para amostras está embasado na teoria da estatística inferencial.

i índice que representa cada observação da amostra
n é a quantidade de observações na amostra
$\overline{x}$ é a média amostral

FIA BUSINESS SCHOOL lab.data

Variância

Qual é a variância de **metragem** (em m²) dos 561 imóveis no *case* de perfis de imóveis residenciais?

$$\text{Variância} = \frac{(140 - 123{,}2)^2 + (90 - 123{,}2)^2 + (130 - 123{,}2)^2 + \ \ldots}{561 - 1} = 4.346{,}2$$

A variância de metragem dos imóveis é de cerca de **4.346,2 (m²)²**.

Para amostras

No Excel: **VAR.A()**
No R: **var()**

Para populações

No Excel: **VAR.P()**
No R: **var() * (n-1)/n**

Arquivo: Imoveis.txt

Variância

Qual é a variância de **metragem** (em m²) dos 561 imóveis no *case* de perfis de imóveis residenciais?

$$\text{Variância} = \frac{(140 - 123{,}2)^2 + (90 - 123{,}2)^2 + (130 - 123{,}2)^2 + \ \ldots}{561 - 1} = 4.346{,}2$$

A variância de metragem dos imóveis é de cerca de **4.346,2 (m²)²**.

Por que o valor da variância é **tão alto**, e sua escala está elevada **ao quadrado**?

→ Porque o cálculo envolve soma de diferenças ao quadrado. Logo, o resultado final está numa escala **potencializada** em relação à escala original dos dados, o que dificulta a interpretação da medida de variância.

Arquivo: Imoveis.txt

Desvio Padrão

O **desvio padrão** é uma medida de **padronização da variância** dos dados, em relação ao valor da média. Ele consiste, simplesmente, na **raiz quadrada** do valor da variância.

Racional de cálculo:

$$\text{Desvio padrão} = \text{Raiz quadrada da variância}$$

Fórmula amostral:

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

i é um índice que representa cada observação da amostra
n é a quantidade de observações na amostra
$\overline{x}$ é a média amostral

Desvio Padrão

Qual é o desvio padrão de **metragem** (em m²) dos 561 imóveis no *case* de perfis de imóveis residenciais?

$$\text{Desvio padrão} = \sqrt{\frac{(140 - 123{,}2)^2 + (90 - 123{,}2)^2 + (130 - 123{,}2)^2 + \ \ldots}{561 - 1}} = 65{,}9$$

O desvio padrão de metragem dos imóveis é de cerca de **65,9 m²**.

Tendo em vista que a média de metragem é de 123,2 m², o desvio padrão nos indica que é muito **comum** observar imóveis com metragens desde 123,2 – 65,9 = **57,3 m²** até 123,2 + 65,9 = **189,1 m²**.

Ou seja, existe grande heterogeneidade de metragens entre os imóveis da base de dados.

No Excel: **DESVPAD.A()**
DESVPAD.P()
No R: **sd()**
sd() * (n-1)/n

Arquivo: Imoveis.txt

FIA BUSINESS SCHOOL lab.data

Coeficiente de Variação

O **coeficiente de variação (CV)** é uma medida que **relativiza o desvio padrão** a partir do valor da **média**.

Racional de cálculo / **Fórmula**:

$$\text{Coeficiente de variação} = \frac{\text{Desvio padrão}}{\text{Média}}$$

Coeficiente de Variação

Qual é o coeficiente de variação da **metragem** (em m²) dos 561 imóveis no *case* de perfis de imóveis residenciais?

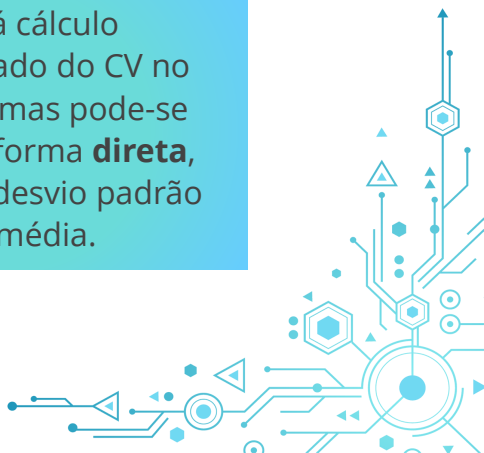
$$\text{CV} = \frac{65{,}9}{123{,}2} = 0{,}53$$

O coeficiente de variação dos imóveis é de cerca de **0,53**. Também pode ser interpretado como uma porcentagem, ou seja, **53%**.

Isso significa que o desvio padrão assume valor corresponde a aproximadamente **metade** do valor da média.

Não há cálculo implementado do CV no Excel ou R, mas pode-se calcular de forma **direta**, dividindo o desvio padrão pela média.

Arquivo: Imoveis.txt



Coeficiente de Variação

O coeficiente de variação é especialmente útil para comparar variabilidade entre subgrupos que possuam **diferentes valores médios**.

Exemplo: considere os seguintes valores de média e desvio padrão de salários dos funcionários de uma empresa, a depender do seu nível hierárquico *(dados fictícios).*

	Estagiário	Analista Jr.	Analista Pl.	Analista Sr.	Coordenador	Gerente
Média	R$ 2.000	R$ 4.000	R$ 6.000	R$ 8.000	R$ 12.000	R$ 20.000
Desvio padrão	R$ 400	R$ 400	R$ 400	R$ 400	R$ 400	R$ 400

Podemos afirmar que todos os níveis apresentam o mesmo padrão de **variabilidade** de salários?

Coeficiente de Variação

O coeficiente de variação é especialmente útil para comparar variabilidade entre subgrupos que possuam **diferentes valores médios**.

Exemplo: considere os seguintes valores de média e desvio padrão de salários dos funcionários de uma empresa, a depender do seu nível hierárquico *(dados fictícios).*

	Estagiário	Analista Jr.	Analista Pl.	Analista Sr.	Coordenador	Gerente
Média	R$ 2.000	R$ 4.000	R$ 6.000	R$ 8.000	R$ 12.000	R$ 20.000
Desvio padrão	R$ 400	R$ 400	R$ 400	R$ 400	R$ 400	R$ 400

Podemos afirmar que todos os níveis apresentam o mesmo padrão de **variabilidade** de salários?

Não! Apesar de os desvios padrão serem iguais, os coeficientes de variação (CV) são **muito distintos**:

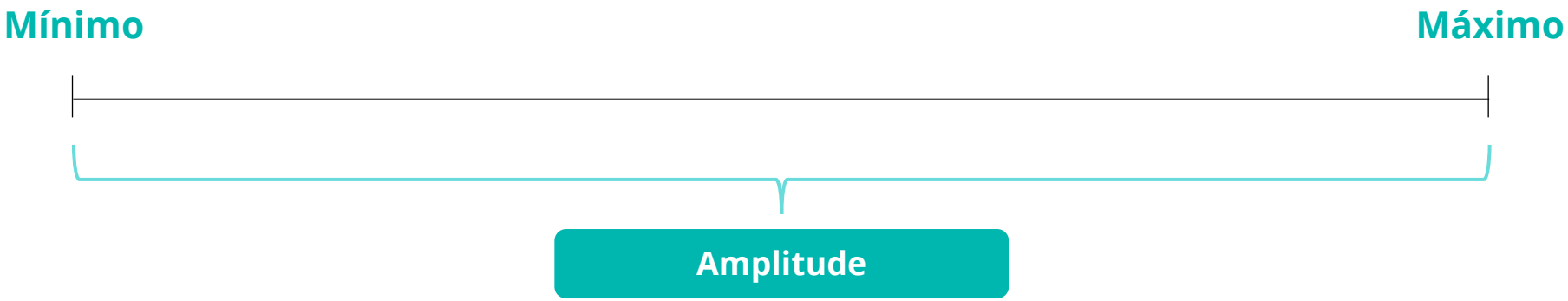
	Estagiário	Analista Jr.	Analista Pl.	Analista Sr.	Coordenador	Gerente
Média	R$ 2.000	R$ 4.000	R$ 6.000	R$ 8.000	R$ 12.000	R$ 20.000
CV	0,2	0,1	0,07	0,05	0,03	0,02

Amplitude

A **amplitude** é uma medida de variabilidade que mensura o **intervalo total de variação** dos dados.

Racional de cálculo:

$$\text{Amplitude} = \text{Diferença entre o valor máximo e o valor mínimo}$$



Amplitude

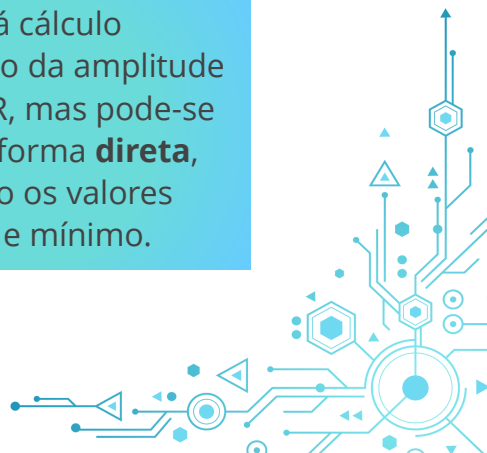
Qual é a amplitude da **metragem** (em m²) dos 561 imóveis no *case* de perfis de imóveis residenciais?

$$\text{Amplitude} = 300 - 30 = 270$$

Ou seja, a amplitude de variação total dos imóveis é de **270 m²**.

Não há cálculo implementado da amplitude no Excel ou R, mas pode-se calcular de forma **direta**, subtraindo os valores máximo e mínimo.

Arquivo: Imoveis.txt

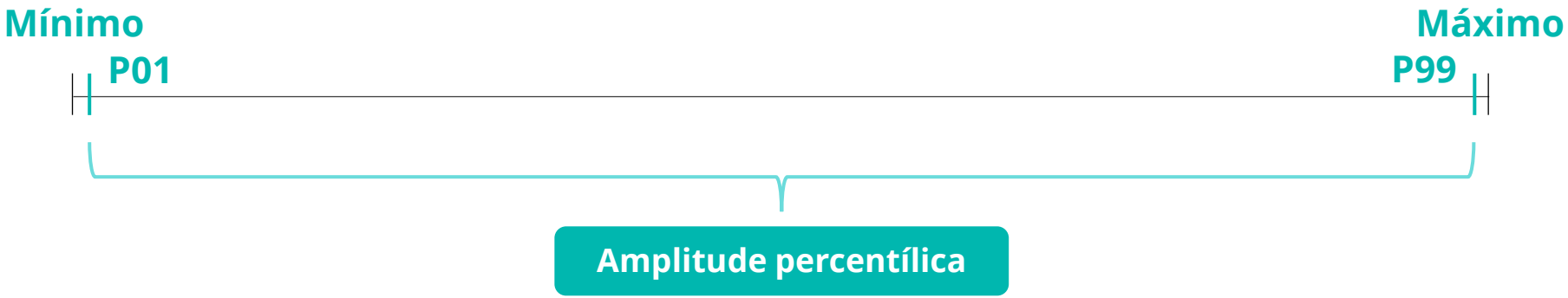


Amplitude Percentílica

A **amplitude percentílica** é uma medida de variabilidade que mensura o **intervalo total de variação** dos dados, desconsiderando 1% dos valores mais baixos e 1% dos valores mais altos.

Racional de cálculo:

$$\text{Amplitude percentílica} = \text{Diferença entre o percentil 99 e percentil 1}$$



Amplitude Percentílica

Qual é a amplitude percentílica da **metragem** (em m²) dos 561 imóveis no *case* de perfis de imóveis residenciais?

$$\text{Amplitude} = 290 - 40 = 250$$

Ou seja, a amplitude percentílica de variação dos imóveis é de **250 m²**.

Não há cálculo implementado da amplitude percentílica no Excel ou R, mas pode-se calcular de forma **direta**, subtraindo os valores de P99 e P01.

Arquivo: Imoveis.txt

Outliers

As seguintes comparações podem ser uma ferramenta útil para detectar a presença de valores atípicos (ou ***outliers***) em variáveis quantitativas:

- valor **mínimo** *versus* **percentil 1**
- valor **máximo** *versus* **percentil 99**
- valor de **amplitude** *versus* de **amplitude percentílica**

Os *outliers* podem representar:

- valores **procedentes** (corretos), mas que são naturalmente **atípicos** em relação à maior parte de valores;
- valores **improcedentes** (incorretos), que precisam ser revistos e **corrigidos** para prosseguir com as análises.

Exemplos:

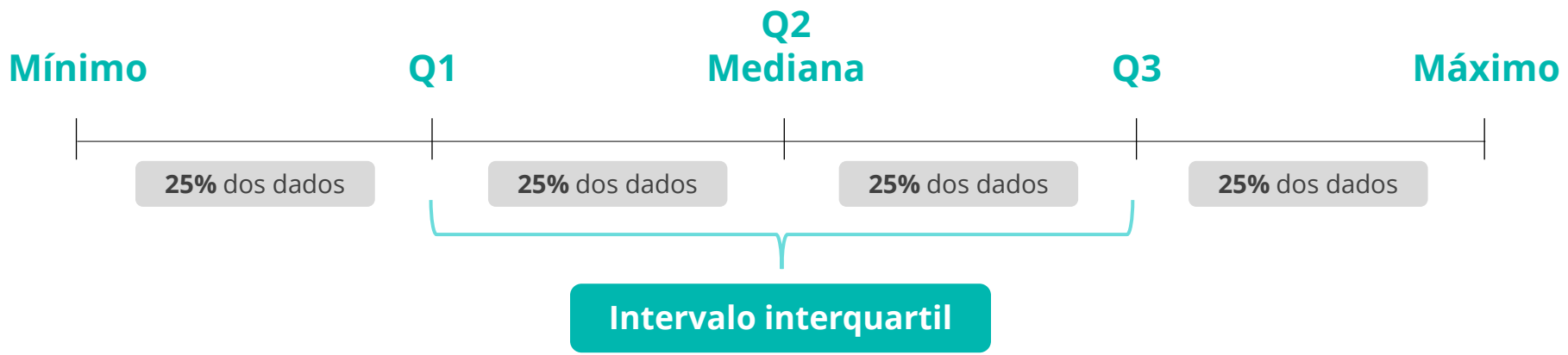
- Base de clientes com idade mínima = **18 anos**, percentil 99 = **85 anos** e idade máxima = **105 anos**.
  Ou seja, 99% dos clientes possuem idades entre 18 e 85 anos, mas naturalmente há clientes com idades atípicas, que chegam até 105 anos.
- Base de produtos alimentares com preço mínimo = **R$ 2**, percentil 99 = **R$ 350** e preço máximo = **R$ 1 milhão**.
  Ou seja, 99% dos produtos alimentares custam entre R$ 2 e R$ 350, mas há produtos com preços discrepantes e potencialmente incorretos.

Intervalo Interquartil

O **intervalo interquartil (IIQ)** é uma alternativa à amplitude que mensura o **intervalo de variação** apenas da metade dos dados mais **centrais**, ao redor da mediana.

Racional de cálculo:

$$\text{Intervalo interquartil} = \text{Diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1)}$$



Intervalo Interquartil

Qual é o intervalo interquartil da **metragem** (em m²) dos 561 imóveis no *case* de perfis de imóveis residenciais?

$$\text{Intervalo interquartil} = 150 - 70 = 80$$

O intervalo interquartil dos imóveis é de **80 m²**, ou seja, selecionando **metade dos imóveis** que possuem os valores mais "centrais", ao redor da mediana, a amplitude de variação desses imóveis é de 80 m².

Não há cálculo implementado do IIQ no Excel ou R, mas pode-se calcular de forma **direta**, subtraindo os valores de Q3 e Q1.

Arquivo: Imoveis.txt

Quadro Final de Medidas Resumo

Sumarizando as medidas resumo de **posição** e **dispersão** que foram calculadas para a **metragem** no *case* de perfil de imóveis residenciais, temos o panorama a seguir.

Medidas resumo da variável METRAGEM

Média	Moda
123,2 m²	70 m²

Mínimo	P01	Q1	Mediana	Q3	P99	Máximo
30 m²	40 m²	70 m²	110 m²	150 m²	290 m²	300 m²

Desvio padrão	CV	Amplitude	Amplitude percentílica	IIQ
65,9 m²	0,53	270 m²	250 m²	80 m²

Arquivo: Imoveis.txt

FIA BUSINESS SCHOOL lab.data

Análises Univariadas: Variáveis Quantitativas

Já para variáveis **quantitativas discretas** que assumem **muitos** valores possíveis, ou para variáveis **quantitativas contínuas**, as principais técnicas para análise **univariada** são:

- ✓ **Medidas resumo** de **posição**
  - *Média*
  - *Mediana*
  - *Moda*
  - *Mínimo e máximo*
  - *Quartis*
  - *Percentis*
- ✓ **Medidas resumo** de **dispersão**
  - *Variância*
  - *Desvio padrão*
  - *Coeficiente de variação*
  - *Amplitude*
  - *Intervalo interquartil*
- ✓ Gráficos de **histograma** e ***boxplot***

FIA BUSINESS SCHOOL lab.data

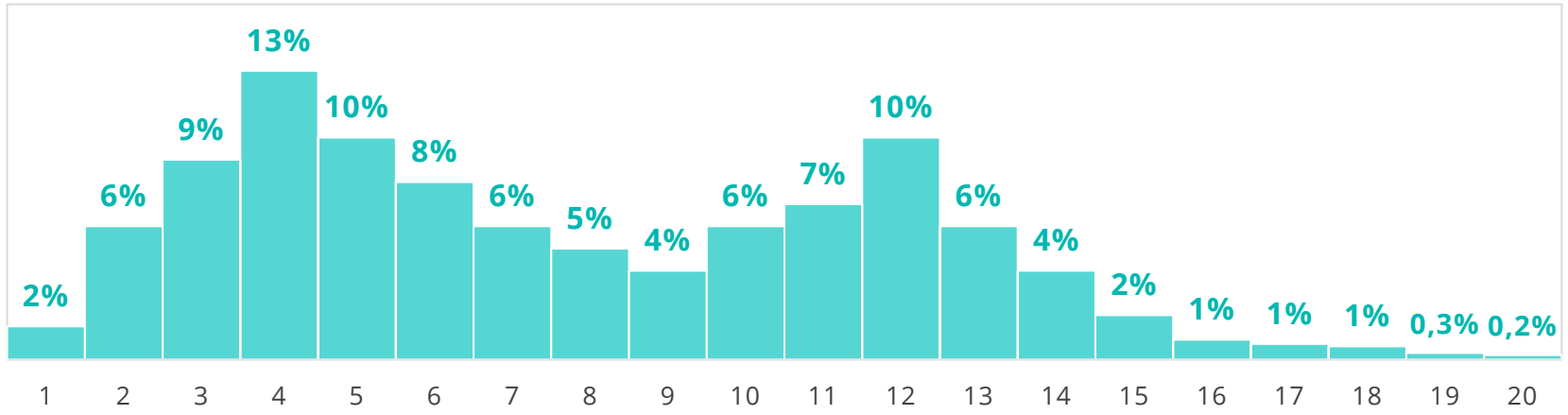
Gráficos: Histograma e *Boxplot*

Os dois gráficos mais apropriados para análise univariada de variáveis quantitativas são o **histograma** e o ***boxplot***.

Histograma

- ✓ Representa a distribuição de frequências associadas à variável quantitativa, fatiada em **faixas de valores**.
- ✓ Diferencia-se do gráfico de barras pelo aspecto de **continuidade**, a partir da ausência de espaço entre as barras.
- ✓ Denota aspectos de **posição** (centralidade, mínimo e máximo, moda), **dispersão** (desvio padrão, amplitude) e **simetria** (decaimento análogo na esquerda e na direita).

Exemplo: Distribuição de renda dos clientes ativos, em salários mínimos



Gráficos: Histograma e *Boxplot*

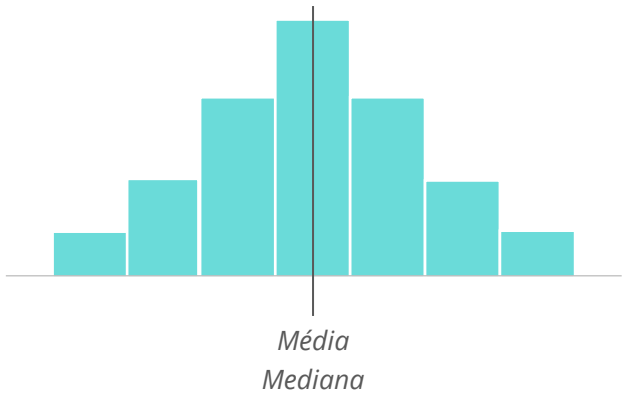
Os dois gráficos mais apropriados para análise univariada de variáveis quantitativas são o **histograma** e o ***boxplot***.

Histograma

- ✓ Representa a distribuição de frequências associadas à variável quantitativa, fatiada em **faixas de valores**.
- ✓ Diferencia-se do gráfico de barras pelo aspecto de **continuidade**, a partir da ausência de espaço entre as barras.
- ✓ Denota aspectos de **posição** (centralidade, mínimo e máximo, moda), **dispersão** (desvio padrão, amplitude) e **simetria** (decaimento análogo na esquerda e na direita).

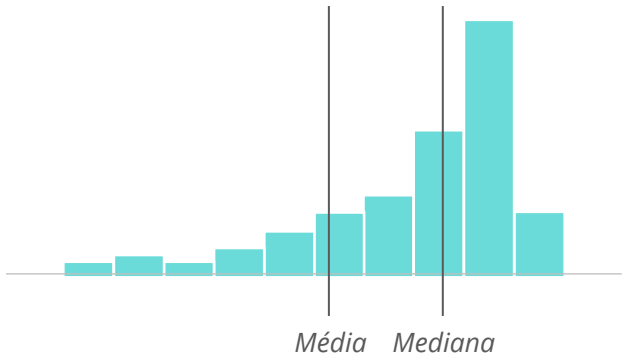
Distribuição simétrica

Média = Mediana



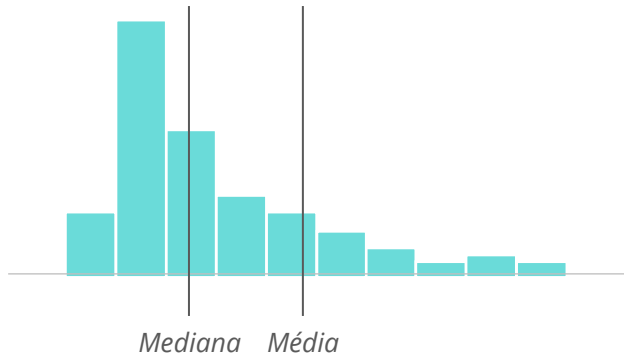
Distribuição assimétrica à esquerda

Média < Mediana



Distribuição assimétrica à direita

Média > Mediana



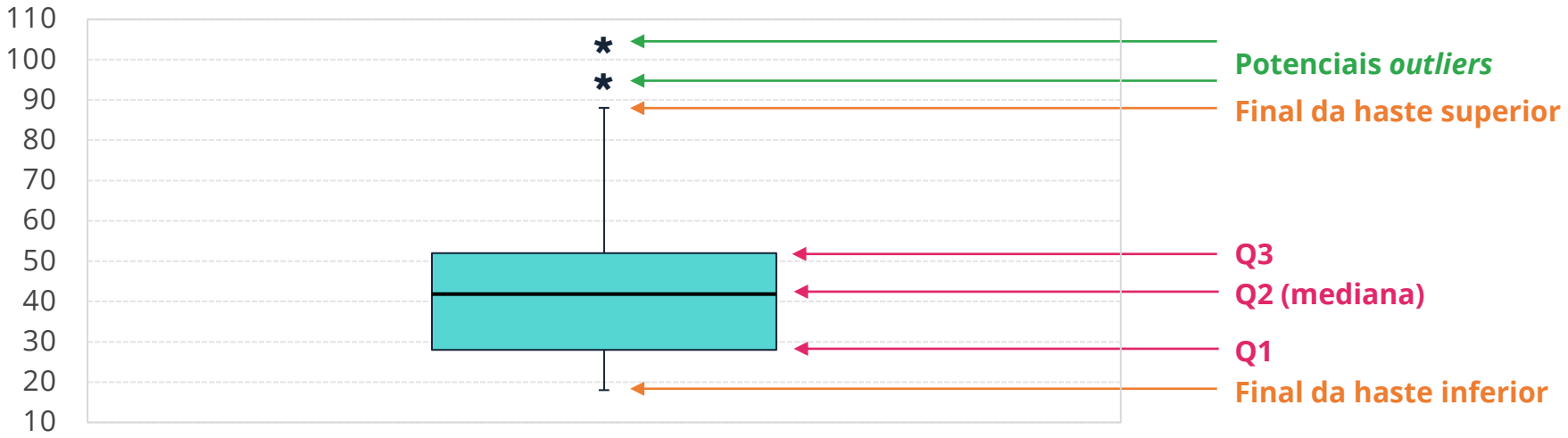
Gráficos: Histograma e *Boxplot*

Os dois gráficos mais apropriados para análise univariada de variáveis quantitativas são o **histograma** e o ***boxplot***.

Boxplot

- ✓ Representa a distribuição de frequências associadas à variável quantitativa, resumida por meio dos **quartis**.
- ✓ A **haste inferior** se estende até o valor mínimo, ou até Q1 - 1,5 * IIQ (o que vier antes).
- ✓ A **haste superior** se estende até o valor máximo, ou até Q3 + 1,5 * IIQ (o que vier antes).
- ✓ Os valores compreendidos fora das hastes são **potenciais *outliers***, representados com pontos (.) ou asteriscos (*).

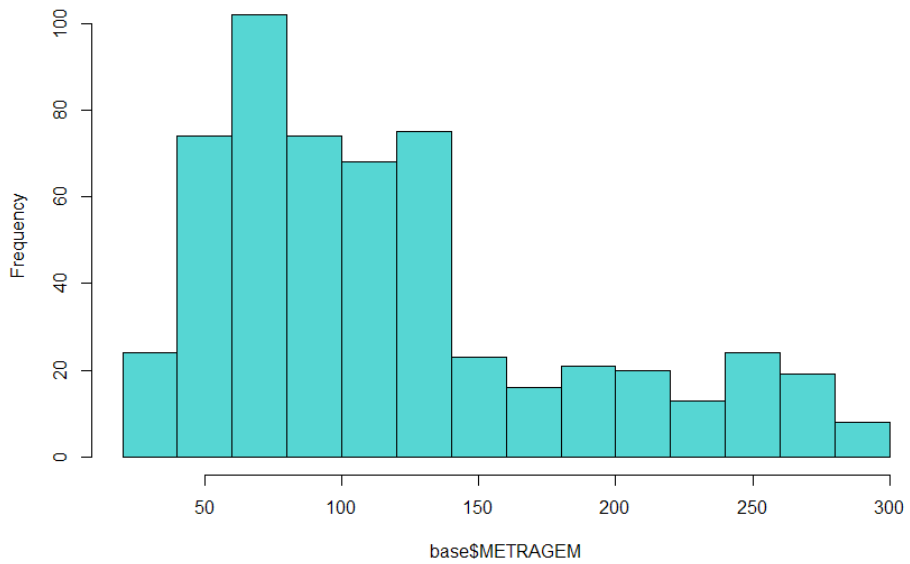
Exemplo: Distribuição de idade dos clientes, em anos



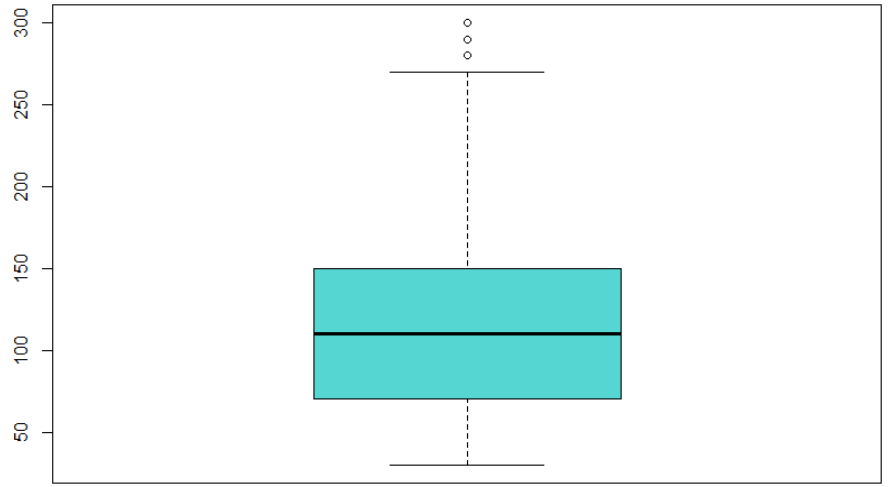
Gráficos: Histograma e *Boxplot*

Gráficos de histograma e *boxplot* para a **metragem** (em m²) dos 561 imóveis no *case* de perfis de imóveis residenciais:

Histograma



Boxplot



A implementação do histograma e *boxplot* no Excel é recente e não muito funcional. Recomenda-se utilizar o R, com as funções **hist()** e **boxplot()**.

Arquivo: Imoveis.txt

7. Análises Bivariadas e Trivariadas

Análises Bivariadas e Trivariadas

Podemos enriquecer ainda mais a nossa análise exploratória realizando a descrição do comportamento de **duas variáveis entre si** (análise **bivariada**), ou de **três variáveis entre si** (análise **trivariada**).

Novamente, os melhores métodos de análise dependem dos **tipos das variáveis** envolvidas. Podemos ter três situações:

- ✓ Variáveis **qualitativas** *versus* variáveis **qualitativas**
- ✓ Variáveis **quantitativas** *versus* variáveis **quantitativas**
- ✓ Variáveis **qualitativas** *versus* variáveis **quantitativas**

Caso as variáveis quantitativas sejam **discretas** e assumam uma quantidade **pequena** de valores, elas podem ser tratadas como variáveis **qualitativas** para fins de escolha das técnicas de análise mais adequadas.

Qualitativas *versus* Qualitativas

Para analisar a **associação** entre **duas ou três variáveis qualitativas**, pode-se utilizar:

- ✓ tabelas de **frequências absolutas e/ou relativas**
- ✓ gráficos de **barras empilhadas**

Qualitativas *versus* Qualitativas

Qual á a relação entre **incidência de luz solar** e **tipo de imóvel**?

Incidência de luz solar	Tipo de imóvel		Total
	Apartamento	Casa	
Nenhuma	33% (105)	21% (50)	28% (155)
Pouca	33% (106)	34% (81)	33% (187)
Muita	35% (112)	45% (107)	39% (219)
Total	**100% (323)**	**100% (238)**	**100% (561)**

Conclusão: Existe maior frequência de imóveis com muita incidência solar entre casas do que entre apartamentos.

Arquivo: Imoveis.txt

Qualitativas *versus* Qualitativas

Qual á a relação entre **incidência de luz solar** e **tipo de imóvel**?

Incidência de luz solar	Tipo de imóvel		Total
	Apartamento	Casa	
Nenhuma	68% (105)	32% (50)	**100% (155)**
Pouca	57% (106)	43% (81)	**100% (187)**
Muita	51% (112)	49% (107)	**100% (219)**
Total	**58% (323)**	**42% (238)**	**100% (561)**

Sempre avalie se é mais adequado somar 100% nas linhas ou nas colunas.

Conclusão: Existe maior frequência de imóveis com muita incidência solar entre casas do que entre apartamentos.

Arquivo: Imoveis.txt

FIA BUSINESS SCHOOL lab.data

Qualitativas *versus* Qualitativas

Qual á a relação entre **incidência de luz solar** e **tipo de imóvel**?

Gráfico de barras empilhadas

Incidência de luz solar versus tipo de imóvel

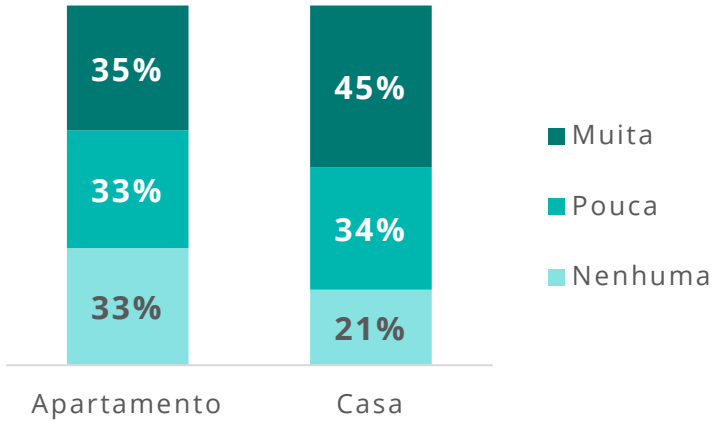
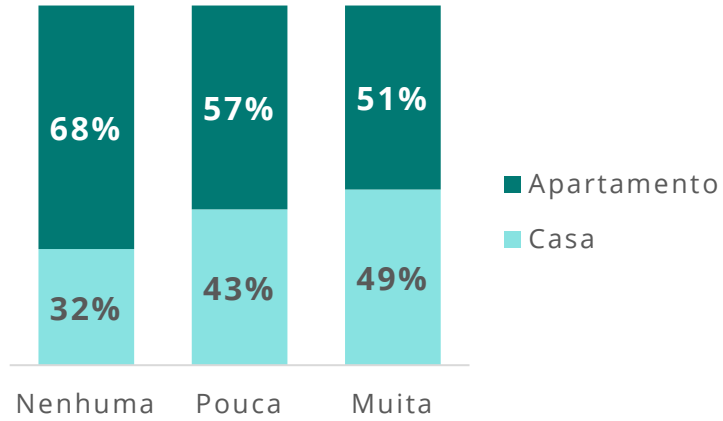


Gráfico de barras empilhadas

Tipo de imóvel versus incidência de luz solar



Arquivo: Imoveis.txt

FIA BUSINESS SCHOOL lab.data

Qualitativas *versus* Qualitativas

Qual é a relação entre **incidência de luz solar, fluxo de veículos** e **tipo de imóvel**?

Incidência de luz solar	Apartamento			Casa			Total
	Fluxo baixo	Fluxo interm.	Fluxo intenso	Fluxo baixo	Fluxo interm.	Fluxo intenso	
Nenhuma	36% (38)	29% (33)	33% (34)	19% (14)	26% (22)	18% (14)	28% (155)
Pouca	36% (38)	32% (36)	31% (32)	43% (32)	31% (27)	29% (22)	33% (187)
Muita	29% (31)	39% (44)	36% (37)	39% (29)	39% (37)	53% (41)	39% (219)
Total	**100% (107)**	**100% (113)**	**100% (103)**	**100% (75)**	**100% (86)**	**100% (77)**	**100% (561)**

Arquivo: Imoveis.txt

Qualitativas *versus* Qualitativas

Qual é a relação entre **incidência de luz solar, fluxo de veículos** e **tipo de imóvel**?

Incidência de luz solar	Apartamento			Casa			Total
	Fluxo baixo	Fluxo interm.	Fluxo intenso	Fluxo baixo	Fluxo interm.	Fluxo intenso	
Nenhuma	36% (38)	29% (33)	33% (34)	19% (14)	26% (22)	18% (14)	28% (155)
Pouca	36% (38)	32% (36)	31% (32)	43% (32)	31% (27)	29% (22)	33% (187)
Muita	29% (31)	**39% (44)**	36% (37)	39% (29)	**39% (37)**	53% (41)	39% (219)
Total	**100% (107)**	**100% (113)**	**100% (103)**	**100% (75)**	**100% (86)**	**100% (77)**	**100% (561)**

Conclusão 1: Tanto para casas quanto apartamentos situados em locais de fluxo intermediário de veículos, a maioria dos imóveis possui muita incidência de luz solar.

Arquivo: Imoveis.txt

Qualitativas *versus* Qualitativas

Qual é a relação entre **incidência de luz solar, fluxo de veículos** e **tipo de imóvel**?

Incidência de luz solar	Apartamento			Casa			Total
	Fluxo baixo	Fluxo interm.	Fluxo intenso	Fluxo baixo	Fluxo interm.	Fluxo intenso	
Nenhuma	36% (38)	29% (33)	33% (34)	19% (14)	26% (22)	18% (14)	28% (155)
Pouca	36% (38)	32% (36)	31% (32)	43% (32)	31% (27)	29% (22)	33% (187)
Muita	29% (31)	39% (44)	**36% (37)**	39% (29)	39% (37)	**53% (41)**	39% (219)
Total	**100% (107)**	**100% (113)**	**100% (103)**	**100% (75)**	**100% (86)**	**100% (77)**	**100% (561)**

Conclusão 2: Para imóveis situados em locais de fluxo intenso de veículos, a maioria possui muita incidência de luz solar; mas esse destaque é mais expressivo para casas do que apartamentos.

Arquivo: Imoveis.txt

FIA BUSINESS SCHOOL lab.data

Qualitativas *versus* Qualitativas

Qual é a relação entre **incidência de luz solar, fluxo de veículos** e **tipo de imóvel**?

Incidência de luz solar	Apartamento			Casa			Total
	Fluxo baixo	Fluxo interm.	Fluxo intenso	Fluxo baixo	Fluxo interm.	Fluxo intenso	
Nenhuma	**36% (38)**	29% (33)	33% (34)	19% (14)	26% (22)	18% (14)	28% (155)
Pouca	**36% (38)**	32% (36)	31% (32)	**43% (32)**	31% (27)	29% (22)	33% (187)
Muita	29% (31)	39% (44)	36% (37)	**39% (29)**	39% (37)	53% (41)	39% (219)
Total	**100% (107)**	**100% (113)**	**100% (103)**	**100% (75)**	**100% (86)**	**100% (77)**	**100% (561)**

Conclusão 3: Para apartamentos situados em locais de fluxo baixo de veículos, a maioria possui nenhuma ou pouca incidência de luz solar; já para casas, a maioria possui pouca ou muita incidência de luz solar.

Arquivo: Imoveis.txt

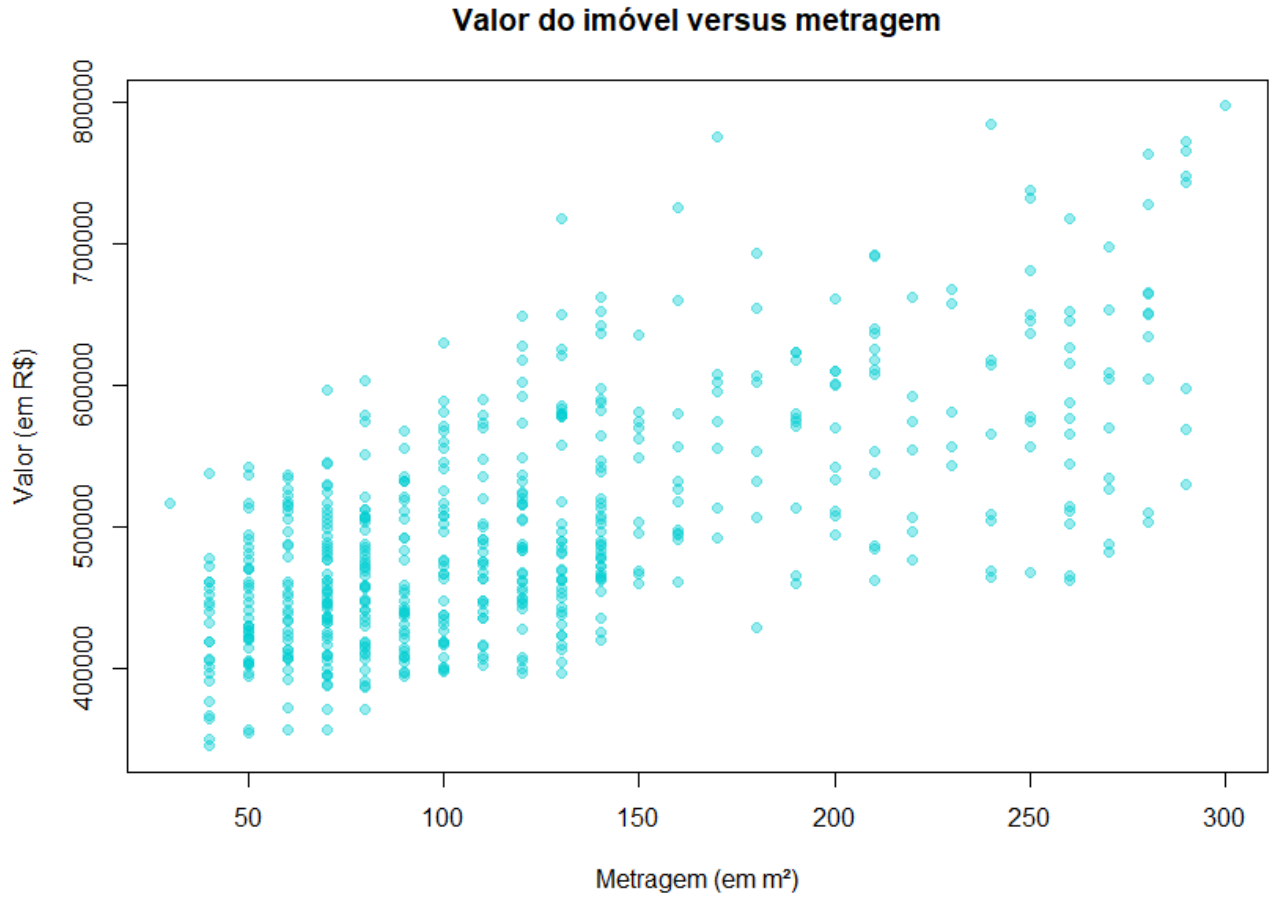
FIA BUSINESS SCHOOL lab.data

Quantitativas *versus* Quantitativas

Para analisar a **associação** entre **duas variáveis quantitativas**, pode-se utilizar **gráficos de dispersão**.

Quantitativas *versus* Quantitativas

Qual é a relação entre **valor da venda** e **metragem** dos imóveis?



Parece existir uma **associação** entre as duas variáveis, de forma que, quanto maior a metragem, maior tende a ser o preço de venda do imóvel. Entretanto, os preços apresentam **bastante variabilidade**.

Arquivo: Imoveis.txt

FIA BUSINESS SCHOOL lab.data

Qualitativas *versus* Quantitativas

Para analisar a **associação** entre uma **variável qualitativa** e uma **quantitativa**, pode-se utilizar:

- ✓ tabelas de **medidas resumo**
- ✓ gráficos de **histograma** ou ***boxplot***

Caso se trate de **duas variáveis quantitativas** *versus* uma **qualitativa**, novamente, podemos utilizar gráficos de **dispersão**.

Qualitativas *versus* Quantitativas

Qual é a relação entre **metragem** e **tipo** dos imóveis?

Medidas resumo da variável METRAGEM, por tipo de imóvel

Apartamentos

Mínimo	P01	Q1	Mediana	Média	Q3	P99	Máximo
40 m²	40 m²	70 m²	90 m²	89 m²	120 m²	140 m²	150 m²

Casas

Mínimo	P01	Q1	Mediana	Média	Q3	P99	Máximo
30 m²	40 m²	110 m²	170 m²	169 m²	230 m²	290 m²	300 m²

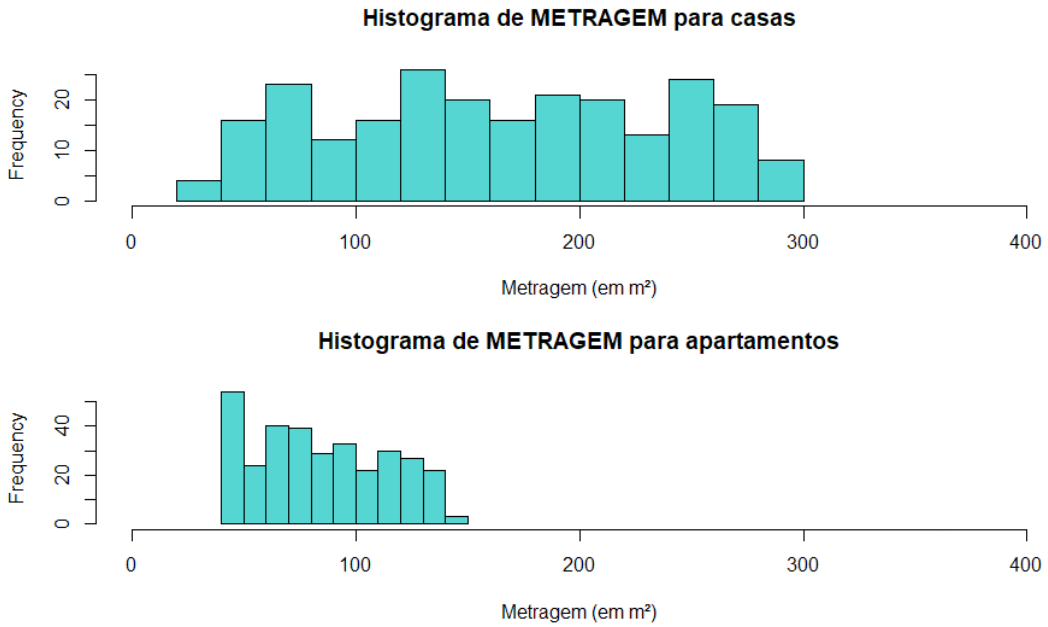
Naturalmente, as casas tendem a possuir **maiores metragens** do que os apartamentos.

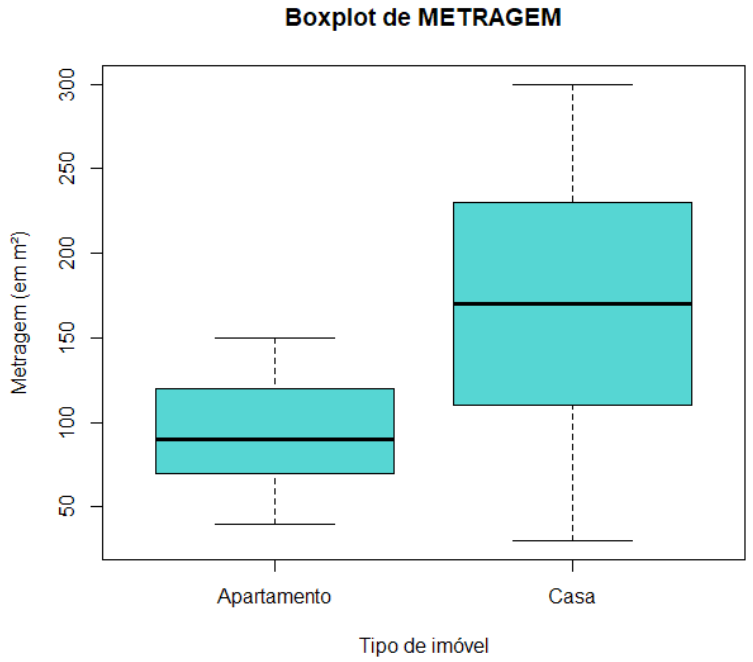
Arquivo: Imoveis.txt

FIA BUSINESS SCHOOL lab.data

Qualitativas *versus* Quantitativas

Qual é a relação entre **metragem** e **tipo** dos imóveis?





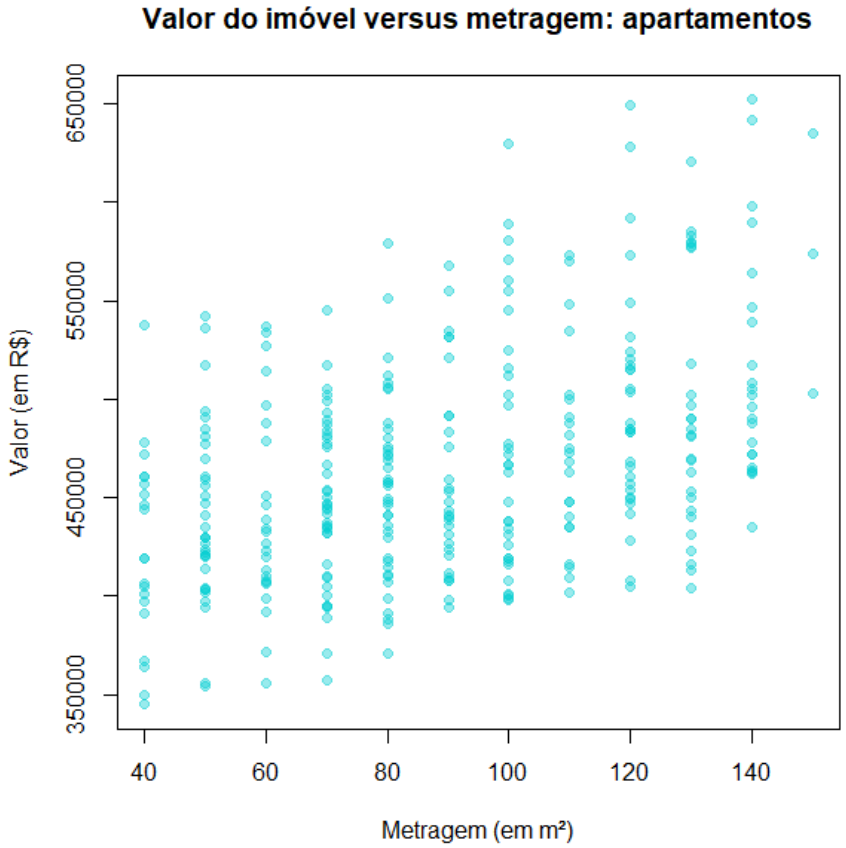
Naturalmente, as casas tendem a possuir **maiores metragens** do que os apartamentos.

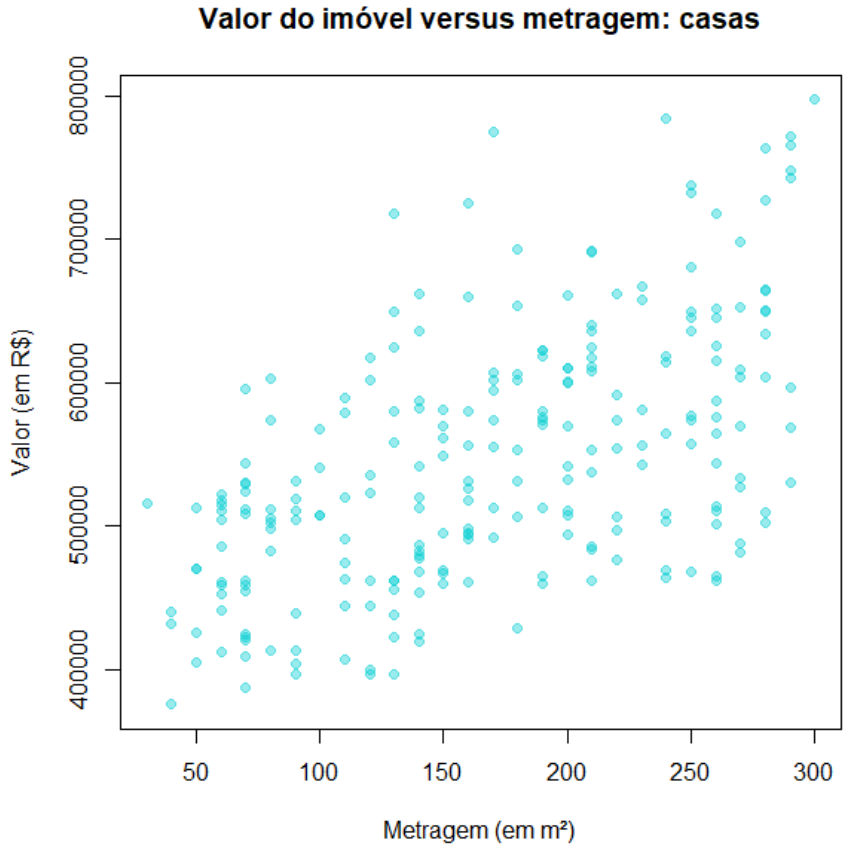
Arquivo: Imoveis.txt

FIA BUSINESS SCHOOL lab.data

Qualitativas *versus* Quantitativas

Qual é a relação entre **valor da venda**, **metragem** e **tipo** dos imóveis?





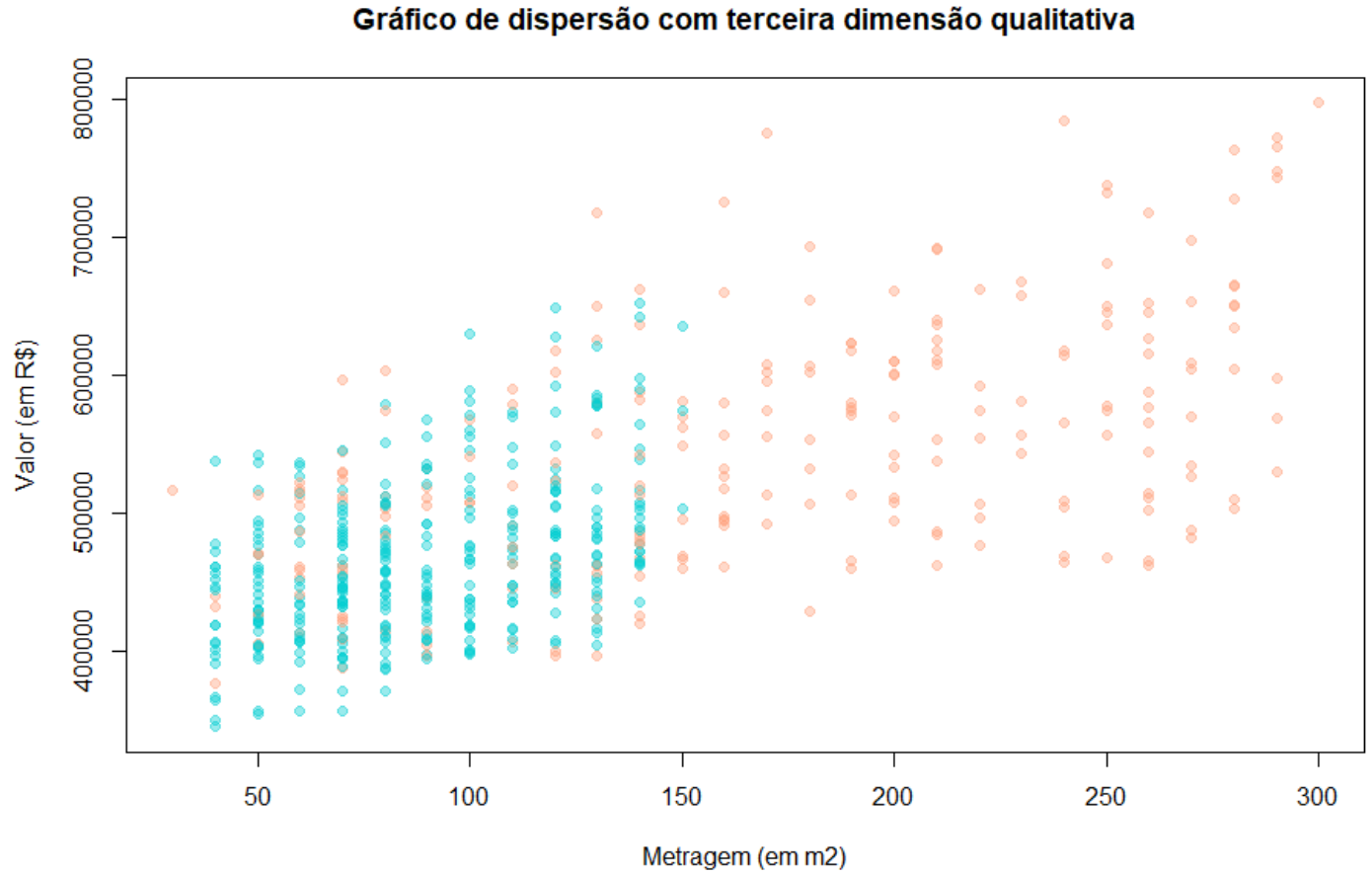
Parece existir uma **associação** entre as duas variáveis, de forma que, quanto maior a metragem, maior tende a ser o valor de venda do imóvel. Isso vale tanto para **casas** quanto **apartamentos**.

Arquivo: Imoveis.txt

FIA BUSINESS SCHOOL lab.data

Qualitativas *versus* Quantitativas

Qual é a relação entre **valor da venda**, **metragem** e **tipo** dos imóveis?



Parece existir uma **associação** entre as duas variáveis, de forma que, quanto maior a metragem, maior tende a ser o valor de venda do imóvel. Isso vale tanto para **casas** quanto **apartamentos**.

Arquivo: Imoveis.txt

FIA BUSINESS SCHOOL lab.data

8. *Case*

Case: Perfil de Imóveis Residenciais

Arquivo: Imoveis.txt

Temos o objetivo de **descrever as características dos imóveis residenciais** disponíveis para venda em uma determinada cidade, a fim de compreender melhor as oportunidades imobiliárias existentes no local.

Responda às seguintes questões de negócio:

a) Quais bairros possuem menor e maior oferta de imóveis disponíveis para venda? Esse comportamento muda a depender do tipo de imóvel (casa/apartamento)?
b) Qual é a menor e a maior qtde. de vagas de garagem existentes em um imóvel? Qual a qtde. mais comum de vagas? Esse comportamento é diferente a depender do bairro ou do tipo de imóvel?
c) A metragem dos imóveis está relacionada com a qtde. de vagas de garagem?
d) Como é a distribuição da quantidade de estabelecimentos comerciais num raio de até 1km dos imóveis? Existem bairros com maior apelo comercial?
e) A incidência de luz solar sobre os imóveis é análoga em todos os bairros?
f) O comportamento de preço dos imóveis é simétrico em relação ao preço médio?
g) O comportamento de preço dos imóveis varia por bairro?
h) O comportamento de preço dos imóveis varia por qtde. de estabelecimentos comerciais num raio de até 1km?

FIA BUSINESS SCHOOL lab.data

Referências Bibliográficas

ANÁLISE EXPLORATÓRIA

- Anderson, R. A. et al. *Estatística Aplicada a Administração e Economia.* 5ª edição. Cengage, 2021.
- Bussab, W. O., Morettin, P. A. *Estatística Básica*. 9ª edição. Saraiva Uni, 2017.
- Illowsky, B., Dean, S. *Introductory Statistics*. Open Stax, 2018.

  Download gratuito em *https://openstax.org/details/books/introductory-statistics*

lab.data

http://labdata.fia.com.br
Instagram: @labdatafia
Facebook: @LabdataFIA