

Análise Exploratória de Dados

Formulação de perguntas de dados preliminares

Felipe Figueiredo

- 1 Análise Exploratória
 - EDA
 - Gráfico de Dispersão
 - Histogramas
 - Boxplot
 - Referências

1 Análise Exploratória

- EDA
- Gráfico de Dispersão
- Histogramas
- Boxplot
- Referências

Paradigmas de Análises de Dados

../logo

Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória

EDA

Gráfico de Dispersão

Histogramas

Boxplot

Referências

- EDA – Análise Exploratória de Dados
- CDA – Análise Confirmatória de Dados

Paradigmas de Análises de Dados

../logo

Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória

EDA

Gráfico de Dispersão

Histogramas

Boxplot

Referências

- EDA – Análise Exploratória de Dados
- CDA – Análise Confirmatória de Dados

- Formalizado por John W. Tukey nos anos 1970
- Objetivo: formular perguntas com base nos dados disponíveis
- Perguntas que podem ser respondidas pela análise dos dados

- Formalizado por John W. Tukey nos anos 1970
- Objetivo: formular perguntas com base nos dados disponíveis
- Perguntas que podem ser respondidas pela análise dos dados

- Formalizado por John W. Tukey nos anos 1970
- Objetivo: formular perguntas com base nos dados disponíveis
- Perguntas que podem ser respondidas pela análise dos dados

O que é

EDA é uma filosofia/approach para

- maximizar a compreensão/insight sobre um dataset
- descobrir estruturas
- identificar variáveis importantes
- detectar outliers e anomalias

Fonte: NIST Handbook (1998)

O que é

EDA é uma filosofia/approach para

- maximizar a compreensão/insight sobre um dataset
- descobrir estruturas
- identificar variáveis importantes
- detectar outliers e anomalias

Fonte: NIST Handbook (1998)

O que é

EDA é uma filosofia/approach para

- maximizar a compreensão/insight sobre um dataset
- descobrir estruturas
- identificar variáveis importantes
- detectar outliers e anomalias

Fonte: NIST Handbook (1998)

O que é

EDA é uma filosofia/approach para

- maximizar a compreensão/insight sobre um dataset
- descobrir estruturas
- identificar variáveis importantes
- detectar outliers e anomalias

Fonte: NIST Handbook (1998)

O que é

EDA é uma filosofia/approach para

- maximizar a compreensão/insight sobre um dataset
- descobrir estruturas
- identificar variáveis importantes
- detectar outliers e anomalias

Fonte: NIST Handbook (1998)

Tukey, 1980

“Ideas come from previous exploration more often than from lightning strokes. Important questions can demand the most careful planning for confirmatory analysis. (...) Finding the question is often more important than finding the answer. Exploratory data analysis is an attitude, (...) NOT a bundle of techniques (...).”

Tukey (1980) - We Need Both Exploratory and Confirmatory

Para Tukey:

- Não basta fazer a análise confirmatória, nem a exploratória: **ambas** são necessárias
- Paradigma: pergunta → resposta é inadequado
- “Encontrar a pergunta é por vezes mais importante que encontrar a resposta”

Para Tukey:

- Não basta fazer a análise confirmatória, nem a exploratória: **ambas** são necessárias
- Paradigma: pergunta → resposta é inadequado
- “Encontrar a pergunta é por vezes mais importante que encontrar a resposta”

Para Tukey:

- Não basta fazer a análise confirmatória, nem a exploratória: **ambas** são necessárias
- Paradigma: pergunta → resposta é inadequado
- “Encontrar a pergunta é por vezes mais importante que encontrar a resposta”

Paradigma linear

../logo

Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória

EDA

Gráfico de Dispersão

Histogramas

Boxplot

Referências

(*) question → design → collection →

analysis → answer

- 1 Como as perguntas são geradas?
- 2 Como os desenhos (experimentais) são guiados?
- 3 Como a coleta de dados é monitorada?

Paradigma linear

../logo

Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória

EDA

Gráfico de Dispersão

Histogramas

Boxplot

Referências

(*) question → design → collection →

analysis → answer

- 1 Como as perguntas são geradas?
- 2 Como os desenhos (experimentais) são guiados?
- 3 Como a coleta de dados é monitorada?

Paradigma linear

../logo

Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória

EDA

Gráfico de Dispersão

Histogramas

Boxplot

Referências

(*) question → design → collection →

analysis → answer

- 1 Como as perguntas são geradas?
- 2 Como os desenhos (experimentais) são guiados?
- 3 Como a coleta de dados é monitorada?

Questões sobre o paradigma confirmatório

../logo

Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória

EDA

Gráfico de Dispersão

Histogramas

Boxplot

Referências

Respostas: Geralmente por ...

- 1 *insights* teóricos e a exploração de dados anteriores (e.g., pesquisa bibliográfica)
- 2 informação qualitativa disponível obtida da exploração de dados anteriores
- 3 exploração dos dados, conforme são obtidos, buscando comportamento “inesperado”

Questões sobre o paradigma confirmatório

../logo

Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória

EDA

Gráfico de Dispersão

Histogramas

Boxplot

Referências

Respostas: Geralmente por ...

- 1 *insights* teóricos e a exploração de dados anteriores (e.g., pesquisa bibliográfica)
- 2 informação qualitativa disponível obtida da exploração de dados anteriores
- 3 exploração dos dados, conforme são obtidos, buscando comportamento “inesperado”

Questões sobre o paradigma confirmatório

../logo

Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória

EDA

Gráfico de Dispersão

Histogramas

Boxplot

Referências

Respostas: Geralmente por ...

- 1 *insights* teóricos e a exploração de dados anteriores (e.g., pesquisa bibliográfica)
- 2 informação qualitativa disponível obtida da exploração de dados anteriores
- 3 exploração dos dados, conforme são obtidos, buscando comportamento “inesperado”

- A chave então é explorar os dados
- Explorar antes, durante e depois da análise confirmatória
- Busca de pistas, idéias e eventualmente conclusões preliminares (*hipóteses!*)

- A chave então é explorar os dados
- Explorar antes, durante e depois da análise confirmatória
- Busca de pistas, idéias e eventualmente conclusões preliminares (*hipóteses!*)

- A chave então é explorar os dados
- Explorar antes, durante e depois da análise confirmatória
- Busca de pistas, idéias e eventualmente conclusões preliminares (*hipóteses!*)

Paradigma alternativo

../logo

Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória

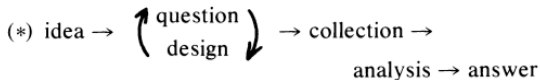
EDA

Gráfico de Dispersão

Histogramas

Boxplot

Referências



Tukey sugere que:

- Antes de termos uma pergunta, temos uma idéia (a ser formalizada)
- Pergunta formal depende dos dados disponíveis
- Questão pragmática, independe do desejo ou vontade

```
../logo
```

Felipe
Figueiredo

EDA

Histogramas

Referências

Paradigma alternativo

../logo

Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória

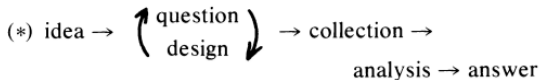
EDA

Gráfico de Dispersão

Histogramas

Boxplot

Referências



Tukey sugere que:

- Antes de termos uma pergunta, temos uma idéia (a ser formalizada)
- Pergunta formal depende dos dados disponíveis
- Questão pragmática, independe do desejo ou vontade

Example

- Idéia: uma certa droga ajuda em uma doença
- Queremos testar/confirmar isso...
- ... com consistência estatística na resposta

- Idéia preliminar informal, vaga
- Geralmente em termos de linguagem coloquial
- Não pode ser avaliada com suporte estatístico

Example

- **Idéia: uma certa droga ajuda em uma doença**
 - Queremos testar/confirmar isso...
 - ... com consistência estatística na resposta
-
- Idéia preliminar informal, vaga
 - Geralmente em termos de linguagem coloquial
 - Não pode ser avaliada com suporte estatístico

Example

- Idéia: uma certa droga ajuda em uma doença
- Queremos testar/confirmar isso...
- ... com consistência estatística na resposta

- Idéia preliminar informal, vaga
- Geralmente em termos de linguagem coloquial
- Não pode ser avaliada com suporte estatístico

Example

- Idéia: uma certa droga ajuda em uma doença
- Queremos testar/confirmar isso...
- ... com consistência estatística na resposta

- Idéia preliminar informal, vaga
- Geralmente em termos de linguagem coloquial
- Não pode ser avaliada com suporte estatístico

Example

- Idéia: uma certa droga ajuda em uma doença
- Queremos testar/confirmar isso...
- ... com consistência estatística na resposta

- Idéia preliminar informal, vaga
- Geralmente em termos de linguagem coloquial
- Não pode ser avaliada com suporte estatístico

Example

- Idéia: uma certa droga ajuda em uma doença
- Queremos testar/confirmar isso...
- ... com consistência estatística na resposta

- Idéia preliminar informal, vaga
- Geralmente em termos de linguagem coloquial
- Não pode ser avaliada com suporte estatístico

Example

- Idéia: uma certa droga ajuda em uma doença
 - Queremos testar/confirmar isso...
 - ... com consistência estatística na resposta
-
- Idéia preliminar informal, vaga
 - Geralmente em termos de linguagem coloquial
 - Não pode ser avaliada com suporte estatístico

Desejo: pergunta geral, de amplo espectro e implicações profundas

Example

“Dos pacientes que morreriam em até três anos desta doença, que fração poderia ser salva por este tratamento?”

- Dificuldade técnica. . .
- . . . nenhum design pode isolar essas pessoas para um experimento

Desejo: pergunta geral, de amplo espectro e implicações profundas

Example

“Dos pacientes que morreriam em até três anos desta doença, que fração poderia ser salva por este tratamento?”

- Dificuldade técnica. . .
- . . . nenhum design pode isolar essas pessoas para um experimento

Desejo: pergunta geral, de amplo espectro e implicações profundas

Example

“Dos pacientes que morreriam em até três anos desta doença, que fração poderia ser salva por este tratamento?”

- Dificuldade técnica. . .
- . . . nenhum design pode isolar essas pessoas para um experimento

O que **pode** ser perguntado está limitado por:

- Idade e sexo dos pacientes
- conjunto mínimo de sintomas
- ausência de outras condições potencialmente fatais
- tipos de pacientes que podem ser encontrados/observados
- etc.

O que **pode** ser perguntado está limitado por:

- Idade e sexo dos pacientes
- conjunto mínimo de sintomas
- ausência de outras condições potencialmente fatais
- tipos de pacientes que podem ser encontrados/observados
- etc.

O que **pode** ser perguntado está limitado por:

- Idade e sexo dos pacientes
- conjunto mínimo de sintomas
- ausência de outras condições potencialmente fatais
- tipos de pacientes que podem ser encontrados/observados
- etc.

O que **pode** ser perguntado está limitado por:

- Idade e sexo dos pacientes
- conjunto mínimo de sintomas
- ausência de outras condições potencialmente fatais
- tipos de pacientes que podem ser encontrados/observados
- etc.

O que **pode** ser perguntado está limitado por:

- Idade e sexo dos pacientes
- conjunto mínimo de sintomas
- ausência de outras condições potencialmente fatais
- tipos de pacientes que podem ser encontrados/observados
- etc.

Formulação da pergunta

../logo

Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória

EDA

Gráfico de Dispersão

Histogramas

Boxplot

Referências

- o que pode concretamente ser perguntado
- que desenhos são viáveis
- chance de um certo design resultar em resposta útil
- “Como eu estudo o que está acontecendo aqui?”

Formulação da pergunta

../logo

Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória

EDA

Gráfico de Dispersão

Histogramas

Boxplot

Referências

- o que pode concretamente ser perguntado
- que desenhos são viáveis
- chance de um certo design resultar em resposta útil
- “Como eu estudo o que está acontecendo aqui?”

Formulação da pergunta

../logo

Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória

EDA

Gráfico de Dispersão

Histogramas

Boxplot

Referências

- o que pode concretamente ser perguntado
- que desenhos são viáveis
- chance de um certo design resultar em resposta útil
- “Como eu estudo o que está acontecendo aqui?”

Formulação da pergunta

../logo

Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória

EDA

Gráfico de Dispersão

Histogramas

Boxplot

Referências

- o que pode concretamente ser perguntado
- que desenhos são viáveis
- chance de um certo design resultar em resposta útil
- “Como eu estudo o que está acontecendo aqui?”

Por onde começar?

../logo

Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória

EDA

Gráfico de Dispersão

Histogramas

Boxplot

Referências

- gráficos dos dados brutos
- estatísticas descritivas simples
- procurar padrões

Por onde começar?

../logo

Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória

EDA

Gráfico de Dispersão

Histogramas

Boxplot

Referências

- gráficos dos dados brutos
- estatísticas descritivas simples
- procurar padrões

Por onde começar?

../logo

Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória

EDA

Gráfico de Dispersão

Histogramas

Boxplot

Referências

- gráficos dos dados brutos
- estatísticas descritivas simples
- procurar padrões

- 1 **Análise Exploratória**
 - EDA
 - **Gráfico de Dispersão**
 - Histogramas
 - Boxplot
 - Referências

Gráficos de dispersão (ou scatter-plots):

- visualizar os dados pontuais diretamente
- identificar possíveis padrões ou tendências
- identificar visualmente possíveis outliers
- desenhar possíveis relações (modelos) sobre os dados

Gráficos de dispersão (ou scatter-plots):

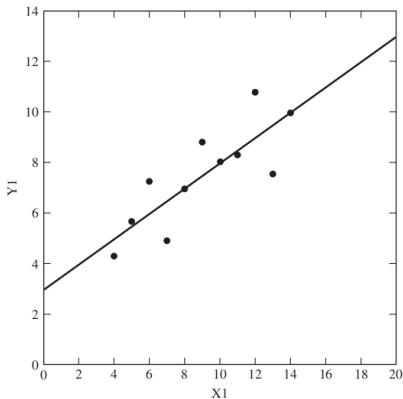
- visualizar os dados pontuais diretamente
- identificar possíveis padrões ou tendências
- identificar visualmente possíveis outliers
- desenhar possíveis relações (modelos) sobre os dados

Gráficos de dispersão (ou scatter-plots):

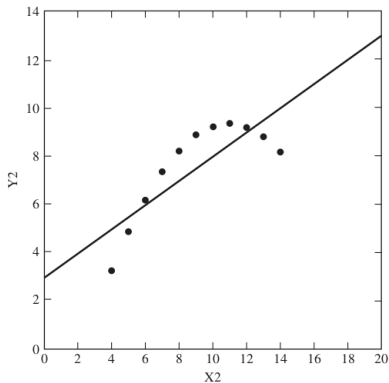
- visualizar os dados pontuais diretamente
- identificar possíveis padrões ou tendências
- identificar visualmente possíveis outliers
- desenhar possíveis relações (modelos) sobre os dados

Gráficos de dispersão (ou scatter-plots):

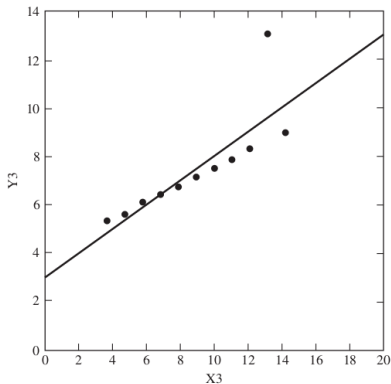
- visualizar os dados pontuais diretamente
- identificar possíveis padrões ou tendências
- identificar visualmente possíveis outliers
- desenhar possíveis relações (modelos) sobre os dados



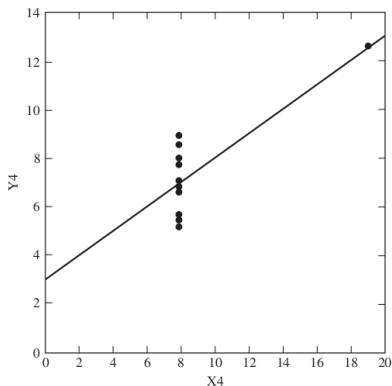
Fonte: Behrens, Yu (2003)



Fonte: Behrens, Yu (2003)



Fonte: Behrens, Yu (2003)



Fonte: Behrens, Yu (2003)

- 1 **Análise Exploratória**
 - EDA
 - Gráfico de Dispersão
 - **Histogramas**
 - Boxplot
 - Referências

- Gráfico de barras com frequências dos dados
- visualização prática da distribuição dos dados
- identificar simetria, tendência central, dispersão, etc

Histograma

../logo

Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória

EDA

Gráfico de Dispersão

Histogramas

Boxplot

Referências

- Gráfico de barras com frequências dos dados
- visualização prática da distribuição dos dados
- identificar simetria, tendência central, dispersão, etc

Histograma

../logo

Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória

EDA

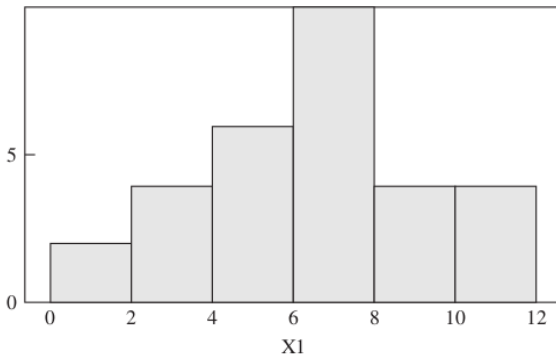
Gráfico de Dispersão

Histogramas

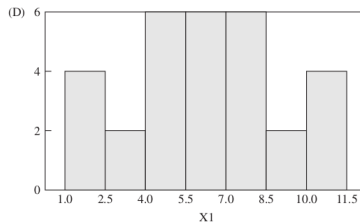
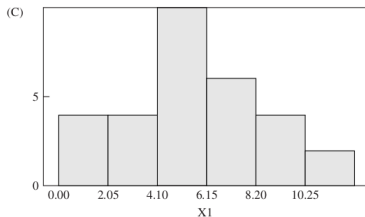
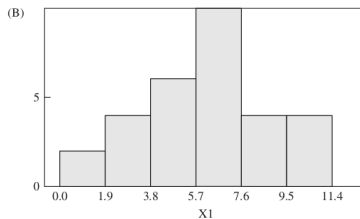
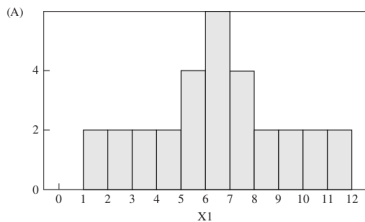
Boxplot

Referências

- Gráfico de barras com frequências dos dados
- visualização prática da distribuição dos dados
- identificar simetria, tendência central, dispersão, etc



Fonte: Behrens, Yu (2003)



- 1 **Análise Exploratória**
 - EDA
 - Gráfico de Dispersão
 - Histogramas
 - **Boxplot**
 - Referências

- caixa que contém 50% dos dados. . .
- . . . e segmentos verticais que englobam a maior parte dos dados
- elementos fora deste intervalo devem ser investigados como possíveis outliers

```
../logo
```

Felipe
Figueiredo

FDA

Gráfico de Dispersão

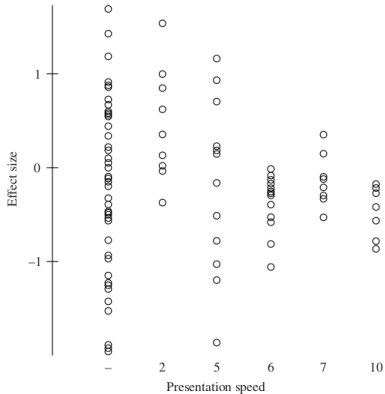
Histogramas

Boxplot

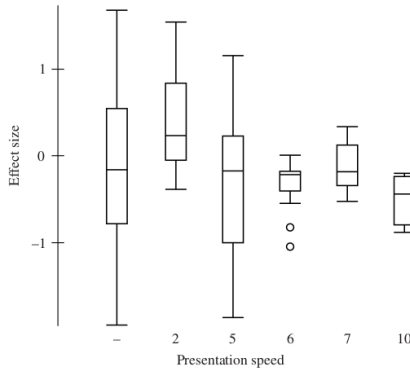
Referências

- ◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡

- caixa que contém 50% dos dados. . .
- . . . e segmentos verticais que englobam a maior parte dos dados
- elementos fora deste intervalo devem ser investigados como possíveis outliers



Fonte: Behrens, Yu (2003)



Fonte: Behrens, Yu (2003)

- 1 **Análise Exploratória**
 - EDA
 - Gráfico de Dispersão
 - Histogramas
 - Boxplot
 - **Referências**

- NIST Handbook (1998), Exploratory Data Analysis, cap 1 - <http://www.itl.nist.gov/div898/handbook/eda/section1/eda1.htm> (Acessado em 10/09/2015)
- Tukey (1980), We need both exploratory and confirmatory, <http://www-ece.rice.edu/~fk1/classes/ELEC697/TukeyEDA.pdf> (Acessado em 10/09/2015)
- Behrens, Yu (2003), Exploratory Data Analysis, cap 2 - Research Methods in Psychology