**QUAL O PAPEL DO CIÊNTISTA DE DADOS?**

- DEFINIR O PROJETO  
- COLETA OS DADOS (COMO OS DADOS VÃO CHEGAR ATÉ MIM?)  
- SELECIONAR FERRAMENTAS E TÉCNICAS   
- COMUNICAÇÃO COM O CLIENTE

**EPICICLOS DA ANÁLISE**

01 - DEFINIR A PERGUNTA  
02 - ONDE ESTÃOS OS DADOS, QUAL A FONTA? ANÁLISE DOS DADOS (CATEGÓRICOS, CONTÍNUOS)  
03 - CONSTRUÇÃO DE MODELOS (RESPOSTAS E EXPLICATIVAS)  
04 - INTERPRETAR OS DADOS  
05 - TRADUZIR PARA A AUDIÊNCIA

**DEFININDO A PERGUNTA DE TRABALHO E OBJETIVO:**

QUAL A PERGUNTA QUE VOCÊ QUER RESPONDER?  
QUAL O PROBLEMA VOCÊ QUER SOLUCIONAR?  
TEM QUE SER ALGO MENSURÁVEL!

CASO:  
CORRETORA DE EMPRÉSTIMOS:  
“Queremos melhorar a nossa escolha de solicitantes, evitando empréstimos ruins” (ruin)

“Queremos reduzir nossa taxa de perdas em 10%, utilizando um modelo preditivo dos solicitantes de empréstimo” (ok)

Tipos de perguntas:

* Descritivas:
  + Fazem um sumário de uma característica de um determinado banco de dados (Proporção de pagadores ruins, proporção entre os diferentes sexos)
  + Diagnóstico
* Exploratórias
  + Fazer análise dos dados em busca de padrões, tendências ou correlações
* Inferenciais
  + Testar responder pergunta com banco de dados (Solicitantes do sexo masculino são, no geral mal pagadores
* Preditivas
  + Uma predição futura (Qual a probabilidade de um futuro solicitante não pagar o empréstimo)
* Mecanicistas
  + O mecanismo por trás do padrão encontrado (Solicitantes do sexo masculino não pagam porque são divorciados e precisam pagar pensão)

REFINANDO O RESULTADO

EXPECTATIVAS

O que você acha que vai aconteceu? Porque? (HIPÓTESE)

TRANSFORMAR A PERGUNTA EM UM PROBLEMA DE DADOS

O que é ser um “pagador ruim”? Qual a informação mensura isso?

**COMO ORGANIZAR PLANILHA DE DADOS**

Cada coluna é uma variável  
Cada célula é um valor  
Cada linha é a repetição

**COLETA E ORGANIZAÇÃO DE DADOS**

**VALOR INEXISTENTE**

O que fazer com valores inexistentes?  
 NaN (Not Avaliable Number) Usado para binários  
 NA (Not Avaliable) Modo Geral

**DICA GERAIS**  
 - Evitar nomes com espaços  
 - Acentos e símbolos  
 - Utilize ponto no lugar de vírgulas

**ANÁLISE EXPLORATÓRIA**

- SUMÁRIO  
 - TIPOS DE VARIÁVEIS

**CONVERTENDO VARIÁVEIS**var2<-as.factor(var)  
var2<-as.numeric(var)

**HISTOGRAMAS E DISTRIBUIÇÕES**hist()  
plot()

**GRÁFICOS EXPLORATÓRIOS**abline(lm(x~y))

**PERSONALIZANDO UM GRÁFICO**pch: tipo de ponto  
ylim: limite do eixo y  
xlim: limite do eixo x  
main: título principal  
cex.lab: tamanho do título do eixos  
col: cores  
sub: Subtítulos  
bty: bordas  
cex: tamanho dos pontos  
log: transformar em logo o eixo  
lwd: largura da linha  
como exportar...

**GRÁFICOS DE CORRELAÇÃO**

Obs: cuidado com correlações ao acaso.

# Corrgram  
corrgram(Airq, upper.panel = panel.cor)

# PerformanceAnalytics  
Airq2<-Airq[,-5] #Removendo coluna categórica do banco de dados  
chart.Correlation(Airq2)

**CONSTRUÇÃO DE MODELOS ESTATÍSTICOS**

VARIÁVEIS RESPOSTAS – Sempre será uma única, a que recebe o efeito das demais  
VARIÁVEIS EXPLOCATIVAS – As que influenciam a variável reposta

A densidade populacional exerce efeito sobre a qualidade do ar, logo, Densidade Populacional é a variável explicativa e qualidade do ar a variável resposta.

Essa relação pode ser uma relação positiva ou negativa, quando a densidade populacional diminui a variável resposta, essa é uma **expectativa** em relação a essas atividades.

Variável resposta ~ Variável Explicativas ( ~ em função de)

Ex: airq~dens

Para adicionar mais variações podemos incluir o + Ex:

Airq ~ dens + vala + coas + medi + rain

# Montando um modelo estatístico

modelo<-glm(airq~vala+rain+coas+dens+medi, data=Airq)

# Selecionando as variáveis significativas

library(MuMin)

options(na.action = "na.fail")

dredge(modelo, rank = "AIC") # Critério de Akaike

**INTRODUÇÃO AO DATAVIZ**

Tipo de Gráficos  
- Dispersão (Variáveis X e Y contínuas)  
- Barras (ANOVA) (Variável continua Y, categoria X)  
- Diagrama de caixa (boxplot) (eixo y contínua e x categóricas) com base na mediana.  
- Circular (pizza ou pie) (Evitar Pizza em função da Barra)

SITE : datavizcatalogue.com, data-to-viz.com

OUTRAS DICAS

Fontes claras e Legendas  
02 Gráficos diferentes  
Apresentação ou Impresso?

CORES

- A cor permite humor  
- Você pode usar cores para guiar o espectador  
- Contraste entre valores   
- Gradientes

SITE: projects.susielu.com/viz-palette, colorbrewer2.org

Acessibilidade, Poucas Cores, Baixo Contraste  
Forma, Tamanho, Tipo, Espessuras

O QUE NÃO FAZER

- muita variável não usar gráfico de pizza  
- valores em cima do gráfico, achatado e sem legenda  
- esquecer nomes nos eixos  
- trabalhe escalas no eixos  
- esquecer legendas