‘R’ te dá asas...

Neale Ahmed El-Dash neale.eldash@gmail.com

R Day Curitiba (Dez/2019)

Introdução

Minha trajetória...

- Graduação em Estatística - UNICAMP (1995-1999) - ???
- Trabalho no Núcleo de Estudos de Políticas Públicas NEPP (1998-1999) - SAS
- Mestrado em Estatística - UNICAMP (2000-2002) - Matlab
- Trabalho na IPSOS Brasil (2002-2006) - SPSS
- Doutorado em Estatística - USP (2006-2010) - R
- Trabalho na IPSOS EUA (2011-2013) - SPSS
- Consultor (2006 até hoje) - Principalmente R

- Não deveria haver uma distinção tão grande entre *"Academia x Mercado"*...
  - São complementares, não opostos!

"A teoria sem a prática é vazia. A prática sem a teoria é cega."

A importância do R

- Após o Mestrado me tornei *"autodidata"*
- Na teoria ok, mas na prática não conseguia entregar
  - Artigos incríveis, queria aplicar
  - Não havia tempo pra implementar
  - Era refém do software estatístico
  - No *"mercado"*, tem que entregar o melhor possível dentro do prazo (curto)
- Quando conheci o R, fiquei mais confiante que podia fazer qualquer análise...
  - Porém o meu workflow com o R era péssimo...
- Quando conheci o R + RStudio + Bibliotecas, minha vida mudou...
  - **O "R" me deu asas...**
  - **Teoria <=> R <=> Prática**

Essa apresentação

- Aprendendo o R
  - Como migrei para o R
- Depois do R
  - As bibliotecas que me possibilitaram fazer coisas que eu não conseguia fazer antes
- Algumas aplicações

Aprendendo R (entre 2006 e 2013)

Motivação

- Empresas de pesquisa gostam de fazer uma aplicação customizada de uma técnica
  - Batizam com um nome
  - Vendem como um produto
- Precisava replicar algum produto que usava **Regressão Bayesiana**
  - Na Ipsos Brasil usava-se SPSS e Excel
  - Só consegui fazer com R
- Dificuldades:
  - Linha de comando
  - Importar os dados para o R
  - Diferentes formatos de dados:
    - dataframe, list, array e matrix

Doutorado

- Me forcei a fazer em R
  - código em espaguete
  - Biblioteca **base** somente
- Via potencial, mas continuava sofrendo...
  - Análises específicas / novas
  - Gráficos (é possível fazer qualquer coisa)

Trabalhando com SPSS + R

- Uma das maiores demandas com pesquisas de opinião pública é a ponderação dos dados
  - No começo fazia manualmente.

- Ponderação:
  - Dar pesos diferentes para cada observação da base de dados
  - Para que a soma ponderada de variáveis na amostra reproduzam o perfil conhecido da população
  - Ex. Mulheres representando 40% das observações na base e 50% na população
  - Com muitas variáveis é bem mais complexo - **Raking / Post-stratification**

- Workflow inicial (SPSS »> R):
  - Manipulava bancos de dados no SPSS
  - Exportava dados para o R
  - PONDERAR OS DADOS (biblioteca survey)
  - Importava resultados do R

Automatizando workflow SPSS + R

- R Essentials for SPSS (link)
  - Foi possível automozatizar todo o processo
  - Na eleição de 2012 nos EUA, usava todo dia na pesquisa da Reuters/Ipsos
  - usado até hoje
- Ninguém na Ipsos US usava R
  - Indispensável parece bom, porém causa problemas...
  - usar o SPSS como interface me permitiu tirar férias...

Trabalhando com Excel + R

- Workflow inicial (EXCEL »> R):
  - Usava a interface do Excel + VBA + Formulários
  - Exportava dados para o R
  - Ponderar dados
  - Importava resultados do R
- Biblioteca RExcel
  - integração muito legal entre Excel e R
  - instala o R Commander também
- Uso na Ipsos
  - Na eleição de 2012 nos EUA, usei no *"Exit Poll"* Reuters/Ipsos

Livros “cross-over”

- R for SAS and SPSS Users
  - Robert A. Muenchen
- R Through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics
  - Richard M.Heiberger, Erich Neuwirth
- Também dá pra usar o SAS para chamar o R
  - explicação aqui

"R" te dá asas...

- Já usava:
  - análises estatísticas específicas
  - gráficos

- **(1)**) RStudio foi lançado em 2011...

- **(2)** Consegui importar e exportar dados sem erros
  - **openxlsx** (importar e exportar pro excel, fácil instalação)
  - **haven** (importar e exportar do SPSS, inclusive nomes de variáveis e níveis de fatores)

- **(3)** Consegui manipular dados eficientemente
  - **dplyr** (transform data)
  - **tidyr** (reshape data)

"Entendi que o que importava mesmo era escolher a biblioteca certa!!!"

Depois do R (2013 em diante...)

Fazendo o trabalho no R fluir...

- Me ajudaram muito:
  - **str**: mostra a estrutura do objeto (hoje em dia a janela do RStudio já faz isso)
  - **dput**: transforma um objeto em um texto que constrói o objeto no R
  - **CTRL+C** e **CTRL+V**: write.table('clipboard') e read.table('clipboard')
  - Biblioteca datapasta: permite colar objetos direto no R de forma bastante flexível
- Permite customizar o R:
  - Criar um arquivo **.Rprofile**
  - Criar um pacote pessoal. Referências: R Packages e GitHub

Aprendendo mais sobre o R...

- Acessei o libgen e baixei todos os livro com " R " no título.
  - Uso o R para aprender e para aplicar
  - Seymour Papert: **"Knowledge is only part of understanding. Genuine understanding comes from hands-on experience"**
- Cheatsheets do R
- Newsletters como R-Bloggers e sites como o StackOverflow
- Livro *"Advanced R"* do **Hadley Wickham**

Workflow com o R+Rstudio

- Depois que fiquei confortável com o R
  - Uso o R como **interface** para vários outros softwares/linguagens
  - Algumas bibliotecas são justamente desenhadas para ser a interface

- Exemplos:
  - **Linha de comando**
  - rJava
  - C++
  - zip
  - gmail
  - PDFTables API
  - highcharts
  - html widgets
  - Google Big Query
  - Amazon S3
  - Microsoft Office

Integração com Python

- Biblioteca reticulate
  - Permite rodar o Python de dentro do R
  - Sessão interativa ou importar funções ou rodar scripts

- Polêmica Python versus R. Minha opinião/experiência:
  - **R utiliza menos tempo do programador**: Permite pensar mais rápido!
  - Python usa menos tempo do processador
  - Pyhton: Tem mais API's já implementadas

Bibliotecas fundamentais - Análise de dados e amostragem

- dplyr e tidyr
  - Tranformar qualquer base de dados
- ggplot2
  - Visualizar qualquer informação (pode adicionar quantos ni'veis quiser)
- lme4
  - Modelos Hierárquicos entre outros
- survey
  - Ponderação - Pós-estratificação e Raking
- sampling
  - Amostras complexas

Bibliotecas fundamentais - Análise de dados 2

- fuzzyjoin
  - Juntar bases de dados quando as chaves são strings (não idênticas)
- stringr
  - Utilizar expressões regulares (regra de formação da string)
- purrr
  - Loops incrivelmente flexiveis e dataframes com list-columns, entre outros
- multidplyr
  - Análisar bases de dados em paralelo

Bibliotecas fundamentais - Web Scrapping

- rvest
  - Raspar dados de quase qualquer site
  - usar comeandos do **CSS** ou **XPath**
- rSelenium
  - Raspar dados dos sites onde não é possível utilizar o rvest

Bibliotecas fundamentais - Análise Bayesiana

- R2WinBUGS
  - Análise bayesiana usando gibbs sampling
- INLA
  - Análise bayesiana muito rápida
- rstan
  - Esse é o pacote mais ativo hoje...

Bibliotecas fundamentais - Montar um site

- shiny
  - Montar seu próprio site interativo
- rmarkdown
  - Escrever relatórios e códigos no mesmo documento
- blogdown
  - Montar seu próprio site estático
- rplumber
  - transformar sua função em uma API acessível por qualquer um

Aplicações

App da Moving

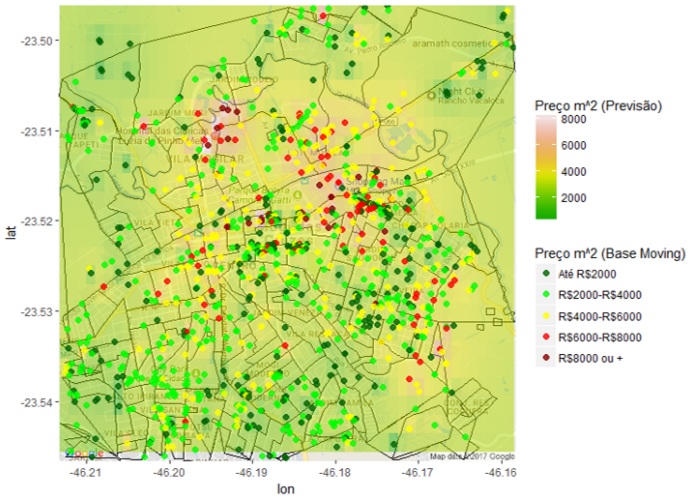
- Prevendo preços de imóveis utilizando um modelo estatístico (link):
  - Toda a parte estatística feita em R (BACK-END)
  - Porém a parte acessível aos analistas (FRONT-END) foi feita em Python / PHP / Angular
  - Estimativas do modelo acessíveis pela API criada usando rplumber
- Os dados de imóveis:
  - WebScrapping de sites de anúncio de imóveis (Zap, ImovelWeb, VivaReal e Moving)
  - Geocodificar endereços de forma **assíncrona** usando api do Google de Geocode e a biblioteca curl

O modelo da Moving

- O sistema permite que um verificador valide a estimativa antes de aprová-la
  - Planilha com os imóveis próximos mais similares (Dist. de Mahalanobis)
  - A estimativa pode ser alterada manualmente (ajustes serão utilizados para melhorar o modelo)

- O modelo têm três níveis de dados:
  - Base de setores censitários
    - Menor unidade geográfica com informações oficiais ()
    - Informações de água, esgoto, iluminação, renda, etc..
  - Informações dos imóveis
    - m^2, banheiros, dormitórios, suítes, vagas na garagem, tipo do imóvel, condomínio
    - Benfeitorias: copa, varanda, academia,etc..
  - Calculamos uma estimativa de preço do m^2 de construção por setor censitário, utilizando **interpolação espacial** (algoritmo *"Inverse Distance Weights"*). Usamos o pacote gstat.

Visualização do método



Site Polling Data

- **Site Polling Data**
  - http://www.pollingdata.com.br/
- **Pesquisas Eleitorais**
  - http://www.pollingdata.com.br/menu_pesquisas/
  - Agregador de pesquisas eleitorais
  - Mais preciso do que as pesquisas eleitorais vista separadamente
  - Entra todo dia no wikipedia e baixa todos os dados de eleições com pesquisas eleitorais
  - Roda um modelo bayesiano para cada eleição.
  - Estima o **house effect** de cada empresa de pesquisa
  - Prevê as chances de vitória de cada partido/candidato
  - Já ganhei dinheiro prevendo qual seria o resultado do datafolha/ibope, e não da eleição em si

Blog

- Novo site (estático)
- **Blog** +http://www.pollingdata.com.br/blog/
  - Utilizando Hugo
  - pacote Rmarkdown
  - pacote Blogdown
  - Adicionar apps do Shiny dentro do site como no post do tutorial inferência bayesiana http://www.pollingdata.com.br/bayes/

Previsão Futebol

- **Previsão Futebol**
  - http://www.pollingdata.com.br/menu_futebol/
  - Utiliza um modelo esperto que extrai os poderes de ataque e defesa de cada time
  - Só leva em conta informações como placar dos jogos e mando de campo.
  - webscrapping diário de jogos de futebol de mais de 30 países e de todas as seleções
  - Apostava no Bet Fair
  - Sulamerica ganhava, mas quando expandi Mundo + Seleções, perdi 500 libras. Parei de apostar...

Outras aplicações do site

- **API de amostragem**
  - http://www.pollingdata.com.br/api_amostra/
  - Com o R, automatizei totalmente a seleção da amostra. Tempo necessário é o de execução apenas.
  - Antigamente para selecionar a mesma amostra, 2 estatísticos levam 3 dias de trabalho cada um.

- **App Tabela Fipe corrigida pela Km**
  - http://www.pollingdata.com.br/2019/10/app-para-ajustar-as-estimativas-da-tabela-fipe/
  - Permite corrigir os dados da tabela FIPE levando em consideração a kilometragem do seu carro
  - Utiliza todos os dados da FIPE mais informações do site Mercado Livre
  - Usa biblioteca fuzzyjoin pra compatibilizar as marcas/veículos dos dois sites

O "R" te dá asas...

- Estatística não é só sobre ganhar $
- Podemos tentar responder qualquer pergunta
- H. G. Wells no seu livro *Mankind in the Making* em 1914:

  "Um dia o pensamento estatístico será tão importante para o pleno exercício da cidadania quanto ler e escrever."

- Hoje, graças ao R, acredito que estamos muito mais perto desse futuro...
  - Acesso gratuíto ao software
  - **pensamento estatístico**: conhecimento + R

Quem sabe o "R" não dará asas, em algum momento, a todos?