01_Motivação

James Hunter

23 de abril de 2020

Neste capítulo, vou explicar porque R é uma solução desejável para a análise de dados. Vou contar um pouco da historia de R, como relaciona com uma outra linguagem popular neste época, Python, e como difera dos outros softwares em uso comum para tarefas de análise de dados. Também, vou explicar um pouco como eu uso R e destacar um projeto que estou fazendo que é tópico principal de 2020, SARS-CoV-2, o vírus que causa a doença COVID-19.

Historia de R e Porque Precisamos Ele

Historia

O antigo Bell Labs de ATT em Nova Jersey foi uma fonte rica das linguagens de computação. C e seus derivativos vêm de lá. S, é uma linguagem que permitiu que cientistas do laboratório podem pôr estatística no contexto de computação sem precisar programar do início todas as funções. (Esse foi necessário na linguagems das primeiras gerações como Fortran, Cobol ou Pascal. Quando o Bell Labs foi desmembrado ao final da década dos 1980s, S tornou um produto comercial (que ainda existe).

Nos anos 1990s, Ross Ihaka e Robert Gentleman de Nova Zelândia criou um derivativo *open-source* da S, que eles chamaram “R”, continuando a tradição de Bell Labs de dar às linguagens nomes de uma letra só. Ihaka e Gentleman lançaram a R em 1999. Agora, estamos usando versão 3.6.3 (logo a ser trocado por 4.0.0). R tem uma comunidade extremamente ativa. Desenvolvedores e usuários contribuíram com mais de 15.500 pacotes com funções e conjuntos de dados adicionais. Esses pacotes abrangem todas as áreas em que possamos pensar, de análise de finanças quantitativas até astrofísica até microbiologia. R realmente tornou na última década uma das mais importantes linguagens de computação.

Nas faculdades em quase todo o mundo, R é a ferramenta de pesquisa mais popular hoje. E a presença dela nas empresas está crescendo diariamente.

Uma das razões que explica este crescimento forte da R é que fica *open-source*. Portanto, pode adquirir **de graça**, sim–livre de custo. **Custo zero** (para aqueles que não acreditaram a primeira vez).

Porque o Professor Usa R

Eu tenho uma longa história com programas que executam estatísticas como SPSS, Stata, SAS e mesmo Graphpad Prism. Comecei de usar SPSS quando ainda estava nos cartões IBM e os computadores eram os IBM 360s (aliás, os anos 60). Tudo bem. Usei várias versões de todos esses outros programas, mas sempre eu estava insatisfeito. Em 2012, descobri R e percebi que minhas insatisfações com SPSS, etc. tiveram solução. A vantagem grande de R sobre os outros programas é que você pode controlar absolutamente a sequência das operações (passos) em sua análise. Você escreve um programa num *script* (i.e., pequeno programa) e a R vai executar exatamente o que você a mandou fazer. E amanhã vai executar uma nova análise com exatamente o mesmo resultado que ontem.

Problemas com os Tradicionais Programas da Análise de Dados

Com todas as caixas de diálogo e menus, os programas da análise de dados distanciam você dos detalhes e passos da análise e da limpeza dos dados. E mais, se você não clicar nas opções exatas na terça análise que você fez antes na segunda, o programa vai produzir um resultado diferente. Pior ainda, essas caixas oferecem muitas opções que você não entende (porque você é um administrador, biólogo, arquiteto, etc., mas não um especialista em estatística). Como resultado, a análise não faz sentido e você não pode realmente explicar as conclusões.

Como disse Prof. Sacha Eskamp da Universidade de Amsterdã:

> *You can only do what the buttons say you can do.* [1] [Você pode fazer somente o que os botões dizem que você pode fazer.]

Software – *Open-Source* vs. Comercial

Além disso, o custo desses programas é absurdamente caro. Mesmo para as edições para estudantes, o preço frequentemente fica acima do que os alunos podem pagar. Como efeito colateral, essa encoraja a proliferação das cópias não autorizadas. Onde fica a ética em ciência ou negócios quando alunos precisam começar a carreira roubando software? Faz muito tempo, softwares *open-source* eram de baixa qualidade. Mas, não mais. Existem equivalentes para os softwares pagos em quase todas as áreas em que precisamos programas. O mundo dos softwares *open-source* realmente mudou. Fora do Microsoft Word®, eu raramente uso softwares pagos.

R vs. Python

Uma alternativa para R com um conceito paralelo de programação das análises é a linguagem *Python*. Lançado em 1989 pelo holandês Guido van Rossum, Python tirou o nome do grupo comédico inglês “Monty Python’s Flying Circus”, não da espécie da serpente. Ela é uma linguagem de alto nível interpretada, exatamente como R. Mas, Python foi desenhada como uma linguagem geral em contraste a R, que tem sua origem em estatística. Para executar até estatística básica em Python, precisa também aprender a programação de vários módulos como Pandas e Numpy, necessários para representação dos conjuntos de dados em Python.

Por isto, eu prefiro e esta matéria vai focar em R. Entretanto, R e Python estão tornando mais compatíveis. Já pode executar Python dentro de RStudio e R dentro dos programas equivalentes de Python. Talvez em 2 anos, para aprender análise de dados, vai aprender os dois idiomas juntos.

Para vocês que gostariam de aprender mais um pouco sobre Python, posso recomendar os seguintes recursos:

- Downey, **Think Python** (http://greenteapress.com/wp/think-python-2e/ (http://greenteapress.com/wp/think-python-2e/))
- Severance, **Python for Everybody** (http://do1.dr-chuck.com/pythonlearn/EN_us/pythonlearn.pdf (http://do1.dr-chuck.com/pythonlearn/EN_us/pythonlearn.pdf))

Aprendizagem de R, É Difícil?

É bastante fácil aprender os elementos básicos de R, inclusive, especificar vetores e conjuntos de dados (que chamamos em R, *dataframes* ou *tibbles*), executar funções básicas de estatística e matemática. Na primeira aula, você vai construir e executar vários scripts e antes do final do curso várias simulações Monte Carlo[2], uma tarefa que precisa habilidade de construir e especificar dados e fazer uma análise.

Então, nas fases iniciais de aprender R, você não vai sofrer muito. Mais tarde, quando você quer utilizar pacotes e funções mais avançados, esses precisarão bastante foco e concentração para programar corretamente. Mas, esta fase seria para um outro dia.

Um Pequeno Exemplo - com Uma Pequena Lição

Como as outras linguagens de programação, você escreve comandos numa forma que o programa entende, um depois do outro com uma sintaxe abreviada que segue as regras da função que você está executando. Por exemplo, se queremos achar a média de 100 números aleatórios entre 1 e 1000, podemos dizer o seguinte:

```
set.seed(1)
dados <- runif(100, min = 0, max = 1000)
media <- mean(dados)
```

A primeira linha conta para o gerador dos números aleatórios de usar `1` como base de contagem. Este garanta que todo mundo que roda o programa terá o mesmo resultado. A segunda linha cria uma variável chamada `dados`, que seria 100 números aleatórios (o "r" no comando `runif`) vindo da distribuição estatística uniforme (o resto de `runif`) que tem um valor entre 0 e 1.000. A terceira linha cria uma variável, `media`, que vai assumir o valor da média dos valores de `dados`.

> **VSS** Uma dica sobre nomes das variáveis: não use acentos nos nomes em R. Se você tenta de rodar seu script num computador que não usa o mesmo sistema de "*encoding*" (maneira de representar as letras na tela), vai produzir erros que são difíceis a concertar.

Este exemplo funciona. (O resultado é 517.847.) Quando você instala R, pode executar esses comandos. O ponto aqui é mostrar para vocês que o código precisa ser escrito exatamente como o programa demanda. Como vocês provavelmente já sabem, o computador é uma máquina de adição grande e estúpida (*"a computer is a big, dumb adding machine"*). Ele faz exatamente o que você comanda ele fazer e mais nada.

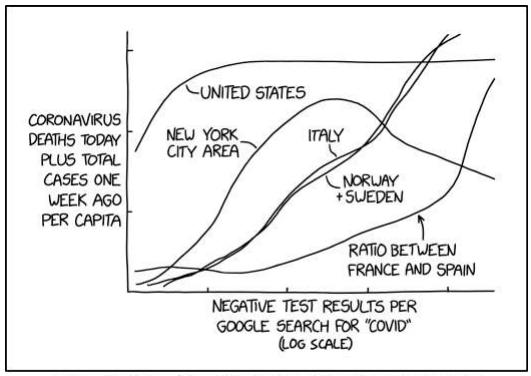
R e a Análise da Pandemia de COVID-19

> **VSS** Nome do vírus - "*SARS-CoV-2*" porque é o primo muito perto do vírus SARS da epidemia da primeira decada do seculo e da família Coronavirus e "2" para não confundir com "SARS". *COVID-19* é o nome da doença que causa.

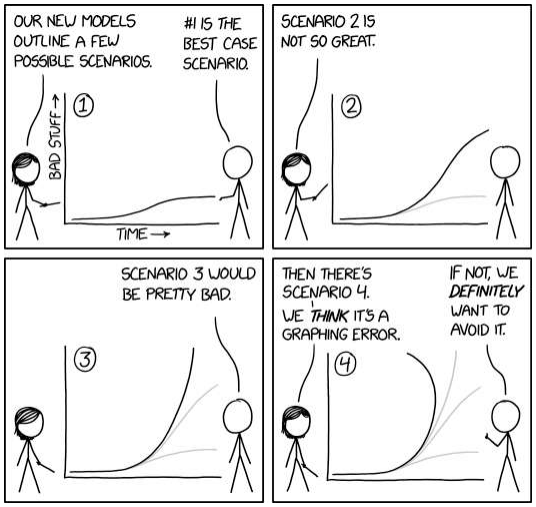
Eu sou um professor de virologia na Escola Paulista de Medicina. Meu doutorado trata do comportamento do vírus HIV-1. Mas em 2020, não importa a doença principal que você estuda, você está trabalhando para conter a pandemia de COVID-19. Minha contribuição aos meus colegas é de focar nos números da evolução da

doença, quem está sendo infectado, como as regiões e estados comparam um ao outro e como Brasil compara aos outros países do mundo. Ainda estamos no início de nosso entendimento do comportamento deste vírus e a sua doença. Sabemos já que não é muito parecido com a influenza (gripe). Tem vetores de transmissão diferentes e tratamentos diferentes. Sabemos que este não é um “gripezinho”. É uma doença muita grave. Os tecidos do corpo humano que o vírus ataca são múltiplos, não só os pulmões.

Nós ainda estamos correndo para colecionar os dados que permitirão que a ciência pode saber o que o vírus realmente faz a vitima e ao nível celular. Até agora, os dados não são muito claros. Qualquer modelo que você vé agora na imprensa é pouco mais de uma extrapolação da nossa experiência com influenza, ou seja, uma aposta. Nosso grupo também prepara modelos preditivos, mas a evolução deles em termos de estrutura tanto quanto os dados usamos para iniciar o modelo é muito rápida. Os charges seguintes darão uma ideia de como cientistas veem os modelos.[3]

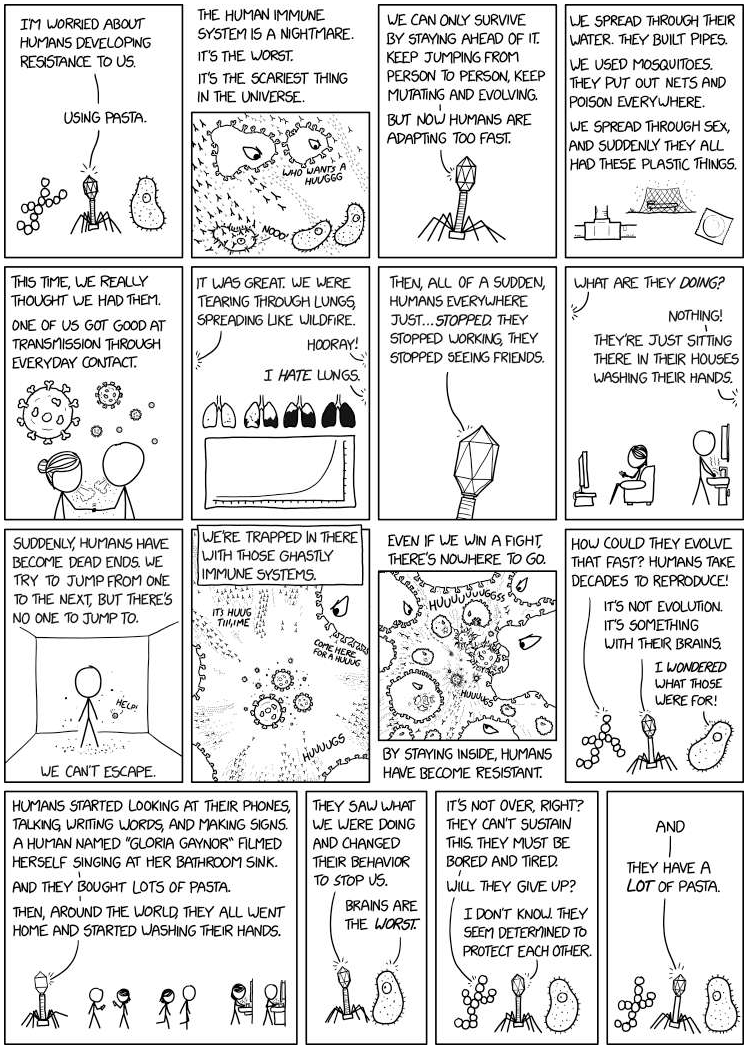


XKCD - Coronavirus Charts



XKCD - Scenario 4

Um último charge de Munroe (ele mesmo um físico) indica a complexidade da resposta a uma pandemia viral melhor que parágrafos de texto. Ele genericamente descreve como funciona o sistema imunológico do corpo e como “distanciamento social” afeta a capacidade dos vírus de espalhar de pessoa em pessoa.[4] O que nós estamos tentando de entender agora é se isolamento e as outras medidas que a sociedade está tomando (ou tentando tomar) são efetivas.



XKCD - Pathogen Resistance

Meus Papers

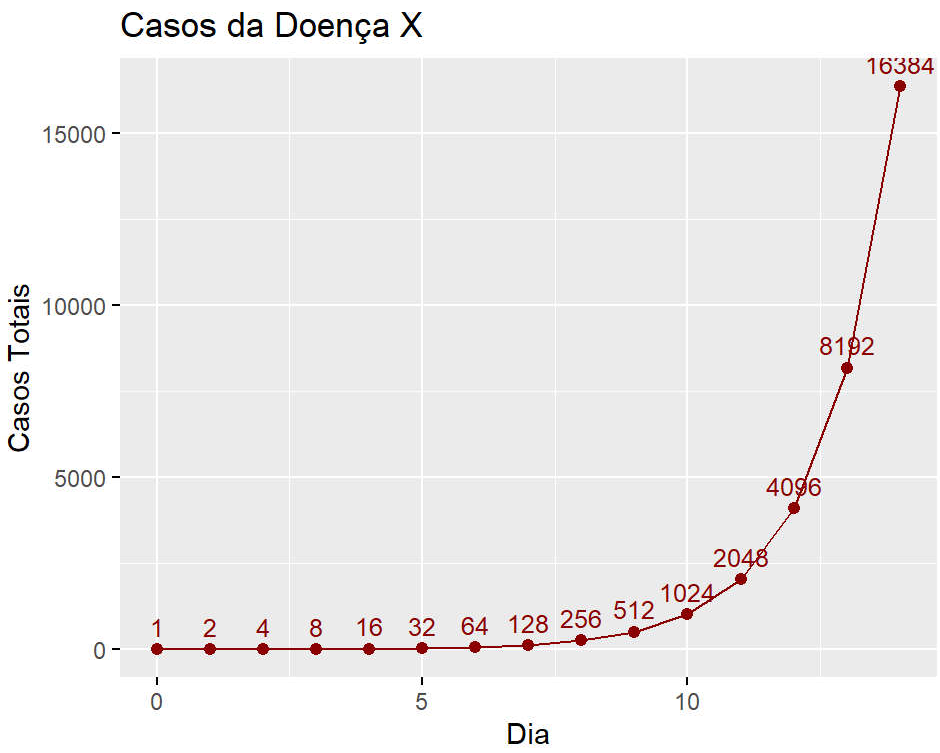
Trabalhando em casa (respeitando a quarentena em São Paulo), procurei conjuntos de dados que eram confiáveis. Periodicamente, eu faço uma atualização de dados para dois papers, um sobre a situação no mundo e um sobre Brasil.[5] Esses papers contêm não só os dados e algumas conclusões mas também toda a programação para meus colegas possam examinar como organizei os dados e analisei eles. Esse procedimento faz parte do dia-a-dia da revisão que fazemos de nosso trabalho com. Apesar não de ter conclusões decisivas sobre a pandemia, algumas coisas são claras.

COVID-19 Segue um Crescimento Exponencial, Sim

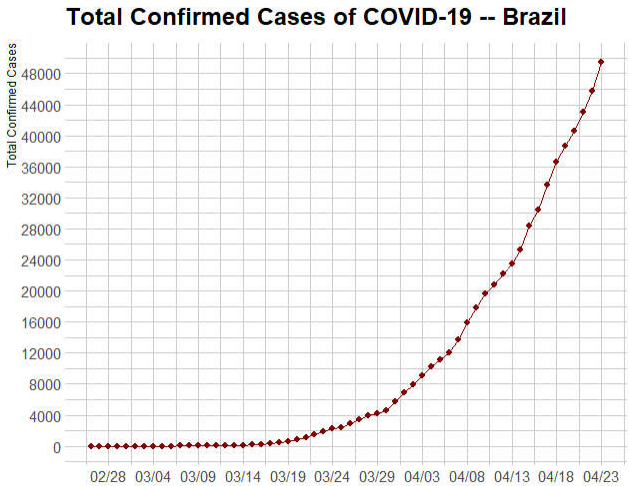
Você leu na imprensa que os casos estão dobrando cada *x* dias e esse fato é muito assustador. Sim e não. Doenças espalham por causa do contato entres pessoas que são candidatos de adquirir a doença. No caso de COVID-19, este incluí quase tudo mundo. De verdade é o resultado inevitável da matemática das doenças. É um tipo de crescimento chamado “exponencial”.

Todas as epidemias começam com uma acumulação inicialmente lento. Mas, depois, o efeito da cada pessoa já infectado causa uma aceleraçao nos totais de doentes porque todos esses doentes estão infectando novas vitimas. Vamos olhar na curva de crescimento de uma doença nova teórica, que vamos chamar Doença X. A primeira pessoa que pega esta doença passa ela para mais duas pessoas por dia (típico para uma doença

comunicável). O gráfico seguinte mostra a rapidez com que esta doença pode espalhar dentro das duas semanas. Apesar de ser assustador, até iniciamos medidas de controle, todas as doenças comportam assim.



COVID-19 segue também uma curva exponencial. Pode ver isso nos casos totais em Brasil desde 24 de fevereiro, o dia do primeiro caso (pensamos).

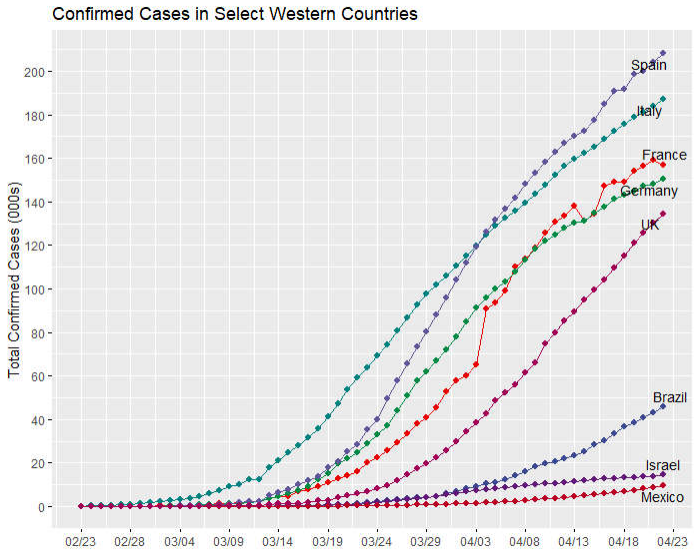


Casos Confirmados - Brasil

A Posicão de Um País, um Estado, uma Região Depende em Quando Teve o Primeiro Caso

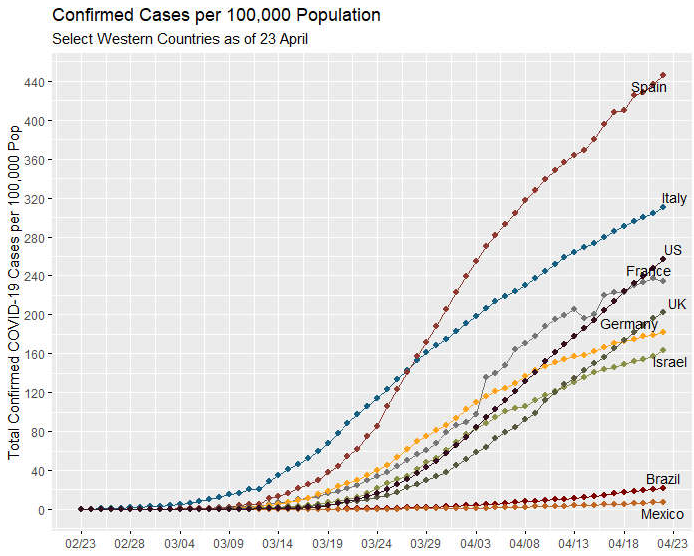
Quando olhamos na resposta do Brasil entre outros países ocidentais, pode ver que em termos absolutos, Brasil ainda fica bem abaixo dos outros países que destaquei no gráfico. Não é necessariamente por causa da excelência da resposta do governo e a população, mas porque Brasil teve os primeiros casos confirmados atrás dos países europeus. Nossa curva ainda não chegou perto do pico. Mas as curvas de Espanha, Itália,

etc. estão chegando perto do pico.



Casos Confirmados - Comparação

Essa contagem dos casos em termos absolutos não conta toda a historia. En termos absolutos, Quando ajustamos o número de casos confirmados por a população do país (de 2018), o padrão geral fica o mesmo, com umas diferenças marcantes.

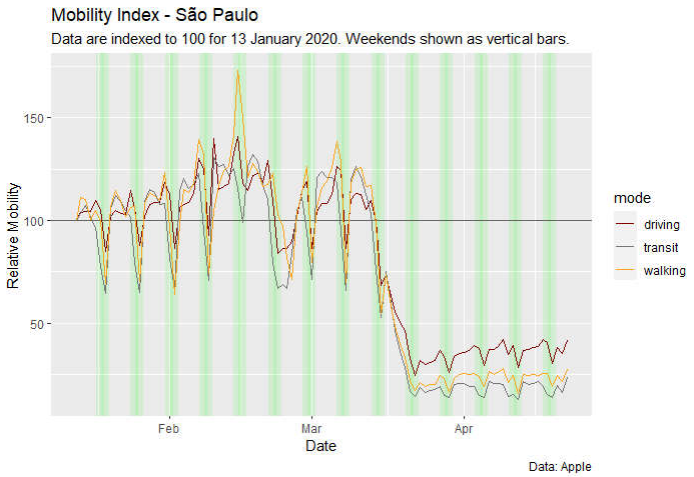


Casos per 100.000 - Comparação

Aqui, o número de casos em Espanha é muito mais que os casos em Itália. Não consegui incluir os Estados Unidos no primeiro gráfico porque o número absoluto de casos foi tão alto, mas aqui podemos ver que a situação nos EUA é entre os piores, mas não tão alto que Espanha e Itália. Também pode ver que o número de casos em pequno Israel é relativamente alto, ao nível dos países europeus. E, o Brasil relativamente tem uma taxa de casos confirmados baixo em relação ao resto dos países. Entretanto, Brasil e Mexico têm taxas baixos porque sofreram os primeiros casos depois dos outros países. Os primeiros casos em Brasil eram pessoas que tinha viajadas em Itália, não a China. Não usei os dados de China porque não estão confiáveis.

Esta É Uma Doença de Contato Alheio ou Próximo?

Uma das téorias que estamos explorando é que a COVID-19 é uma doença que espalha por contato geralizado na praça pública. Essa é típica da Influenza. Você pode passar a gripe pessoa-a-pessoa na rua ou outro lugar pública. Apesar de ser um vírus da mesma fámilia que influenza, o SARS-Cov-2 tem características que sugere que só pode ser comunicado por contato muito próximo, dentro de um raio de menos de um metro. Se for o caso, transmissão entre membros de família seria o meio de transmissão mais comum. Existe evidência sobre isso. Também, uma implicação disso é que mobilidade aumenta com tempo, os novos casos não aumentariam na mesma velocidade. Então, estudamos dados sobre mobilidade nas cidades afetadas. Apple acabou de publicar um índice sobre mobilidade nas cidades e nos páises onde eles têm uma presença forte.[6] No gráfico abaixo, pode ver que o transito de todas as modalidades caiu drasticamente na semana em que as medidas de isolamento social entrarem em vigor. Em abril, pode também perceber que as pessoas começaram de viajar mais, mas pelo menos em termos de número de pedidos de informação ao Apple Maps, o aumento ainda é pequeno. Na primeira aula, nós vamos experimentar um pouco com esses dados.



Mobilidade em São Paulo

Como Essas Análise Aplicam ao Curso de ADcR?

Ainda não temos conclusões sobre COVID-19. Estamos na fase incial de entender como funciona esse vírus. Mas, todas minhas análises e todos os relatórios foram feitos em R com ajuda de RStudio. R permite que eu monto uma análise rapidamente para nós podemos ver a qualidade dos dados e quais tipos de análise podemos aplicar aos dados. Análise de dados não é uma fase separada de uma pesquisa. Com a facilidade de R e RStudio, podemos integrar R e análise de dados em todas as fases de pesquisa.

---

1. Baker, Monya, “Code Alert”, **Nature**, Vol 541, 26/1/2017, p. 563 - 565.↩
2. Irizzary, Rafael, **Introduction to Data Science**, (https://rafalab.github.io/dsbook (https://rafalab.github.io/dsbook)), Ch. 26.3.↩
3. Randall Munroe, “Coronavirus Charts”, https://xkcd.com/2294/ (https://xkcd.com/2294/) e Randall Munroe, “Scenario 4”, https://xkcd.com/2289/ (https://xkcd.com/2289/)↩
4. Randall Munroe, “Pathogen Resistance”, https://xkcd.com/2287/ (https://xkcd.com/2287/)↩
5. Coloquei cópias das versões dos papers preparados no dia 23 de abril no SER para você possa ler eles.↩
6. Cópia de minha análise desses dados também está no SER. Anote que existem problemas com os

dados de Apple que eu comento na minha análise.↩