Introdução à Ciência de Dados e Decisão

Análise Exploratória de Dados com R

4intelligence - Inatel

Thiago Curado e Alejandro Padrón

April 3, 2020

Instruções de Realização e Entrega

- Cada aluno deve realizar sua própria resolução da lista de exercícios;
- A entrega deve ser feita até o final do dia **13 de abril**;
- **A resolução deve ser feita, obrigatoriamente, em formato de *notebook***, contendo tanto o arquivo-base como seu output final (arquivo html ou pdf). A resolução deverá ser enviada por e-mail aos professores do módulo. Aceitaremos envios de arquivos zipados ou links para pastas armazenadas na nuvem (Dropbox, Google Cloud, etc), mas dê preferencia pela entrega via compartilhamento de repositório GIT (aproveite para se familiarizar mais com essa ferramenta!);
- Os alunos têm liberdade de escolher a linguagem utilizada para resolução do problema, desde que esta seja open source. Não haverá discriminação nas notas por conta da escolha, mas incentiva-se que os alunos utilizem o R para esse módulo, em linha com o programa do curso. *Resoluções feitas em linguagens proprietárias, como Matlab e Stata, não serão consideradas.*

Divirta-se!

Parte 1

A primeira seção desta lista trata de séries de escopo relativamente específicos, quais sejam, séries temporais de natureza econômica. Tais tipos de séries constituem bons exemplos para avaliação dos conceitos vistos na segunda aula do curso, como ciclo, sazonalidade, tendência, e visualização de dados temporais. As questões dessa seção também são relativamente mais direcionadas, tendo em mente os conceitos vistos em aula

Na sequência, a Parte 2 das questões trará datasets com séries de outra natureza, abrindo também o escopo das questões de forma a permitir a livre exploração do rico conjunto de dados disponibilizado.

1. **Visualização básica de dados**

   Leia o arquivo RDS "us_change". Trate-se de um tibble de variáveis trimestrais contendo as variações percentual no gastos privados com consumo, renda disponível, produção, popuplção e taxa de desemprego no Estados Unidos entre 1970 e 2016. As taxas de variação foram obtidas a partir de em valores reais medidos em dólares americanos de 2012.

   a) Construa um novo tibble no qual todas as variáveis sejam disponibilizadas em número índice, assumindo valor 100 no primeiro trimestre do ano 2000 (ie 2010Q1 = 100).

   b) Explore a correlação entre as variáveis. Qual a diferença entre se calcular a correlação das variáveis em número índice e em taxa de variação?

   c) Construa gráficos que contribuam em seu entendimento sobre a dinâmica de cada variável do dataset, bem como as relações entre elas. Assim, por exemplo, como ponto de partida plote gráficos de dispersão conjunta das variáveis, bem como suas evoluções ao longo do tempo. Sinta-se livre para complementar tal caracterização com todo e qualquer arsenal analítico que julgue interessante.

   d) A partir das visualizações obtidas no item anterior, que tipo de aprendizado você consegue extrair acerca de (i) evolução das variáveis ao longo do tempo e (ii) das correlações nas dinâmicas das diversas variáveis?

   e) Você consegue identificar, visualmente, alguns movimentos bruscos/atípicos/anômalos na evolução das séries? Tente destacar tais pontos nos gráficos construídos

anteriormente. A quais eventos concretos esses momentos atípicos estão relacionados?

2. **Séries de tempo, ciclo, sazonalidade e tendência**

   O arquivo "retail.xlsx" contém informações sobre vendas mensais de varejo para diversos estados da Austrália.

   a) Leia os dados contidos no arquivo "retail.xlsx". Qual cuidado adicional você precisou ter ao realizar essa importação?

   b) Selecione uma das variáveis e as converta para o formato "time series".

   c) Explore a série escolhida por meio da construção de gráficos. Em particular, se estiver utilizando o R, teste as funções *ggseasonplot* e *ggmonthplot*. O que você consegue identificar em termos de ciclo, sazonalidade e tendência?

   d) Decomponha a série utilizando o método X11. Ele revela algum outlier ou padrões anômalos não identificados anteriormente?

Parte 2

A ideia desta segunda parte da avaliação é propiciar aos alunos oportunidade de aplicar todo o ferramental aprendido em datasets razoavelmente ricos e propícios à analises descritivas. Aqui não será pedido nenhum tipo de análise específica, mas sim que o aluno explore ao máximo as bases, de modo a transformar dado em informação útil e de fácil absorção! Todo tipo de insight e análise que puder ser retirado das bases é útil, pois ajuda a compreender fenômenos implícitos nos dados. Usem e abusem dos pacotes e funções aprendidas, do Google e do material complementar recomendado no material.

Ambos datasets fazem parte do chamado "Tidy Tuesday", um evento semanal onde a cada terça-feira um novo dataset e disponibilizado e membro da comunidade R fazem análises e/ou aplicam visualizações interessantes e novas.

3. **Dataset Spotify - package "spotifyr"**

   Os autores do package compilaram mais de 5.000 músicas de gêneros e subgêneros distintos. O descritivo do dataset, bem como a obtenção dos dados em si, está toda no seguinte repositório: <https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-01-21/readme.md>

a) Use e abuse de todo o ferramental aprendido (e também do que será aprendido, por ventura, em consultas ao Google). A avaliação será feita tanto em cima da riqueza do código em si (em termos do ferramental usado) quanto do aprofundamento analítico na exploração dos dados e obtenção de informações e relações úteis.

4. **Video Games Dataset**

   O dataset contém dados como a data de lançamento, desenvolvedor, tempo médio jogado, etc. O descritivo do dataset, bem como a obtenção dos dados em si, está toda no seguinte repositório: <https://github.com/rfordatascience/tidytuesday/tree/master/data/2019/2019-07-30>

   a) Use e abuse de todo o ferramental aprendido (e também do que será aprendido, por ventura, em consultas ao Google). A avaliação será feita tanto em cima da riqueza do código em si (em termos do ferramental usado) quanto do aprofundamento analítico na exploração dos dados e obtenção de informações e relações úteis.