Insper

Ciência dos Dados

Aula 01

Introdução à disciplina

Professores:

Maria Kelly Venezuela

1º semestre de 2017

Fonte: http://www.cienciaedados.com/carreiras-em-data-science/

O que é Ciência dos dados?

Cientistas de dados são os grandes mineradores de dados. Eles recebem uma enorme massa de dados desorganizados (estruturados e não estruturados) e usam suas habilidades em matemática, estatística e programação para limpar, tratar e organizá-los. Em seguida, eles aplicam suas capacidades analíticas – conhecimento da indústria, compreensão contextual, ceticismo de suposições existentes – para descobrir soluções para os desafios de negócios ocultos. Entre suas principais responsabilidades estão:

- Realizar pesquisas sem direção e formular perguntas abertas aos dados
- Extrair grandes volumes de dados de múltiplas fontes internas e externas
- Empregar os programas de análise sofisticadas, aprendizado de máquina e métodos estatísticos para preparar os dados para uso em modelagem preditiva e prescritiva
- Explorar e analisar dados de uma variedade de ângulos para determinar fraquezas escondidas, tendências e / ou oportunidades
- Conceber soluções orientadas a dados para os desafios mais prementes
- Inventar novos algoritmos para resolver problemas e criar novas ferramentas para automatizar o trabalho
- Comunicar previsões e resultados para a gestão e os departamentos de TI através de visualizações de dados eficazes
- Recomendar mudanças econômicas aos procedimentos e estratégias existentes

Insper

Fonte: http://www.sas.com/pt_br/insights/analytics/cientistas-de-dados.html

O que é Ciência dos dados?

Sobre o SAS

Funções típicas dos cientistas de dados

Não há uma descrição de trabalho definitiva quando se trata de um cientista de dados. Mas aqui estão algumas coisas que você provavelmente terá de fazer:

- Coletar grandes quantidades de dados "unruly" ou desafiadores e transformá-los em um formato mais prático.
- Solucionar problemas de negócios com técnicas de orientação à dados.
- Trabalhar com uma variedade de linguagens de programação, incluindo SAS, R e Python.
- Ter uma sólida compreensão de estatísticas, incluindo testes estatísticos e distribuições.
- Manter-se a par das técnicas analíticas, como a aprendizagem de máquinas, ou *machine learning*, a aprendizagem profunda, ou *deep learning* e análise de dados textuais, ou *text analytics*.
- Comunicar-se e colaborar com TI e área de negócios.
- Procurar por ordens e padrões nos dados, bem como detectar tendências que podem ajudar os resultados de uma empresa.

Cientista de dados: perfil

Qual a diferença do Cientista de Dados e do Cientista Tradicional?

Convencional

- Laboratórios e experimentos
- Nichos de aplicação: controle de qualidade, pesquisas de opinião

Futuro, atual

- Volume grande de dados, muitas vezes não aproveitados
- Oportunidades

O Cientista de Dados deve ter o conhecimento "old school".

Fonte: http://www.cienciaedados.com/usando-data-science-no-combate-a-fraudes/

Usando Data Science no combate a fraudes

Dados x Fraudes

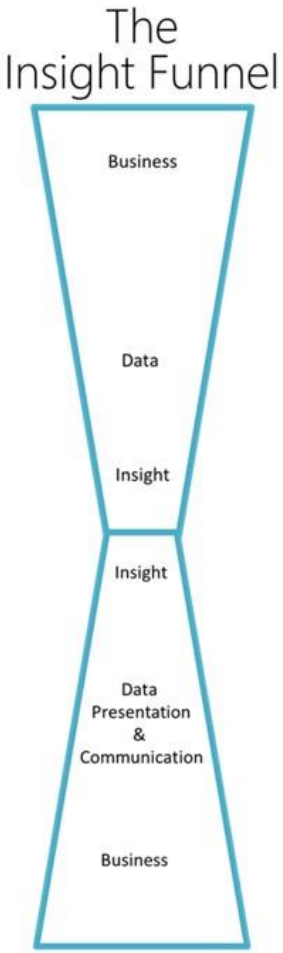
... A fraude em si é um fenômeno dinâmico, que muda e se adapta ao longo do tempo e as pessoas que cometem fraudes, são normalmente experts naquilo que fazem, o que torna o desafio de combate à fraude ainda maior. E por isso mesmo, os métodos tradicionais de análises de dados não têm sido capazes de identificar e prever as fraudes, mesmo com os dados disponíveis. É quando entra em ação a Ciência de Dados, principalmente o campo de Machine Learning. ... Operadoras de cartão de crédito, de telefonia, bancos, indústrias. Todos estão criando seus departamentos de combate à fraude. E estão usando Data Science para isso.

Técnicas de Detecção de Fraudes

.... Um sistema de análise de dados para detecção e prevenção de fraudes tem de estar equipado com uma quantidade substancial de conhecimento e ser capaz de executar tarefas de raciocínio envolvendo esse conhecimento com novos dados fornecidos. No esforço para atingir esse objetivo, os Cientistas de Dados voltaram-se para o Machine Learning (Aprendizado de Máquina). Basicamente, o objetivo da aprendizagem de máquina é converter dados e exemplos (entrada) em conhecimento (saída).

Fonte: https://courses.edx.org

Interpretação e comunicação de insights de dados em um negócio



Business Group/Function	Pense na necessidade/oportunidade do negócio.
Process Area	
Business Objective	E afunile suas decisões.
Decision	
Business Question	
Data	Busque os dados que auxiliem nesse objetivo.
Measurements	
Insight	E comunique os resultados.
Data Presentation	
Data Communication	
Business Opportunity	Faça recomendações...
Decisions	
Business Objective	

Insper

Exemplo: NYC Open Data

Hidrante campeão de arrecadação

Blog "I Quant NY" identificou um problema baseado nos dados públicos de NY

Fontes:

https://data.cityofnewyork.us/

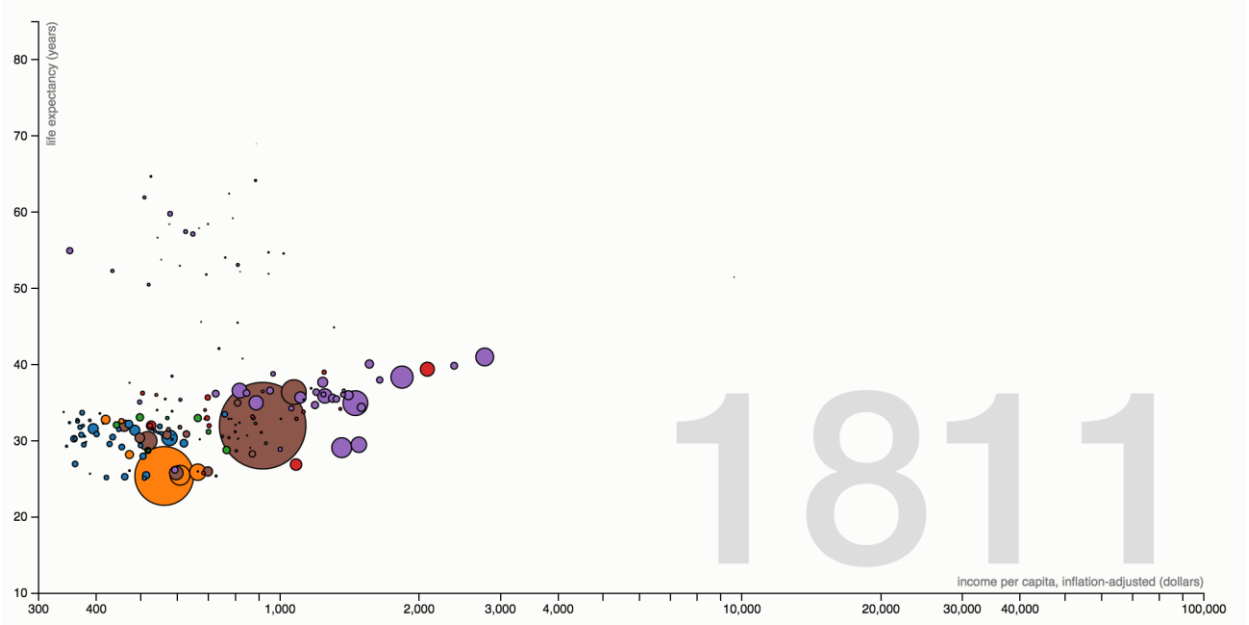
Exemplo: NYC Open Data

Notificada pela comunidade de Data Science e pelo Reddit, a prefeitura de NY corrigiu o problema

Fonte: http://iquantny.tumblr.com/post/87573867759/success-how-nyc-open-data-and-reddit-saved-new

TED Talk sobre o caso: The Worst Parking Spot in NY
http://tinyurl.com/tedworsepark

Visualização de dados

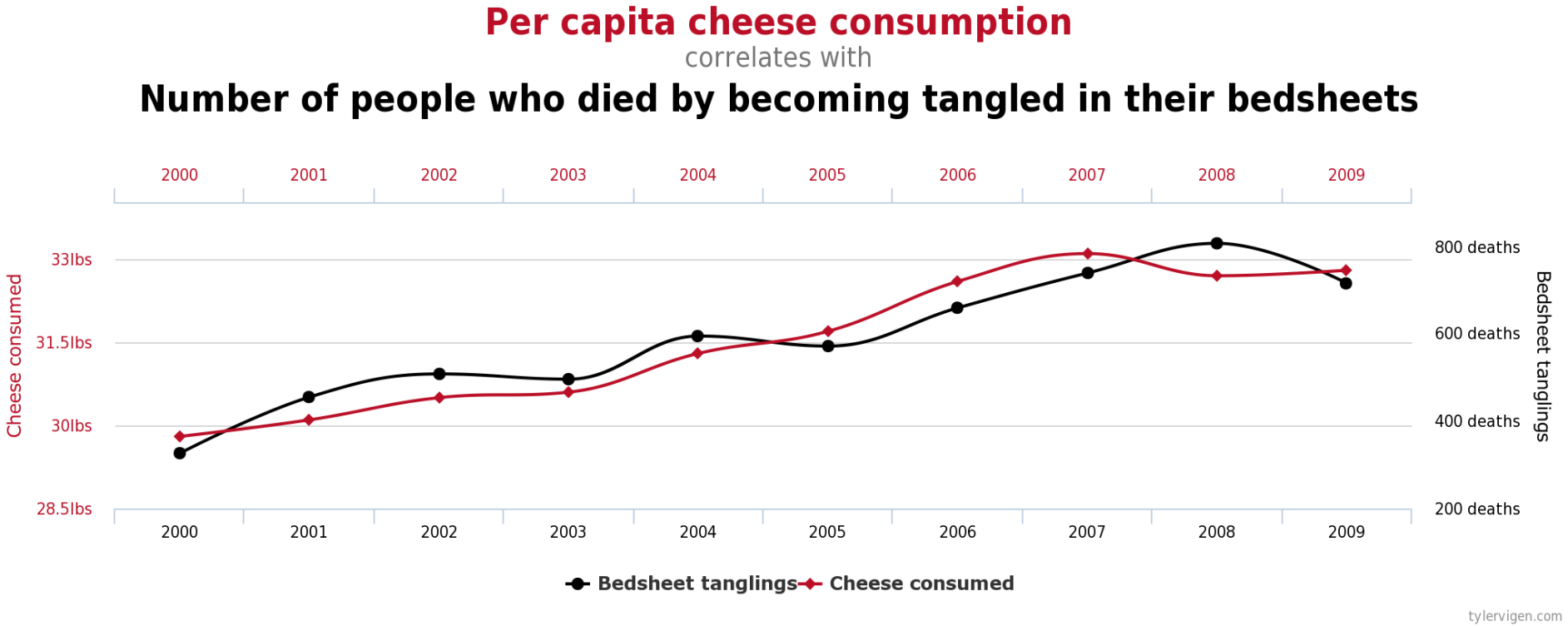


Assistam: The Best Statistics You've Ever Seen – TED
http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen

Usar aplicativo:
http://tinyurl.com/gosling-visualization

Fonte: http://www.tylervigen.com/spurious-correlations

Exemplo: Correlações?



Como o próprio nome diz: uma correlação espúria!

Cuidado para não fazer interpretações/conclusões espúrias nas suas análises!

Insper

Áreas relacionadas

Big Data

Trata de questões de escalabilidade, alto volumes de dados e ambientes de produção (em que os dados participam do negócio).

Machine Learning

Estuda como treinar algoritmos a partir dos dados – ex.: filtro de spams, detector de fraudes,…

Visualização de dados

Procura tornar os dados mais fáceis de entender para amparar a tomada de decisão.

Objetivos de aprendizado

Ao final do semestre, o aluno deverá ser capaz de:

- Elaborar análises exploratórias de dados (univariadas e multivariadas), utilizando ferramentas estatísticas e computacionais adequadas;
- Especificar as distribuições de probabilidades adequadas para as variáveis quantitativas discretas e contínuas;
- Conduzir testes inferências adequados que possam dar base à tomada de decisão; e
- Analisar relações entre as variáveis, utilizando ferramentas estatísticas inferenciais adequadas.

Bibliografia básica

1. MAGALHÃES, M.N; DE LIMA, A. C. P. **Noções de Probabilidade e Estatística**. 7.a Ed. Edusp
2. MONTGOMERY, D.; RUNGER, G. C.; HUBELE, N. **Engineering Statistics**. 5.a Ed. John Wiley and Sons, 2011.
3. DOWNEY, A.B. **Think Stats**. O'Reilly Media, 2011.

Ciência dos dados

AI: Avaliação Individual

Prova 1:
06/04/207 (quinta-feira)

Prova 2:
04/05/2017 (quinta-feira)

Prova 3
01/06/2017 (quinta-feira).

Ciência dos dados

AP: Avaliação em Projetos

Projeto 1 (individual): Análise Descritiva

Descrever um perfil de domicílios (a ser escolhido pelo aluno) e escrever uma matéria noticiando fato. **PNAD**

Projeto 2 (individual): Dados + Teoria

Ajustar uma distribuição probabilística mais adequada aos dados.

Projeto 3 (dupla): Simulação

TLC + Distribuição de média amostral padronizada.

Projeto 4 (trio): Experimento

Comparar duas grupos com objetivo de identificar alguma eficácia ou diferença entre eles.

Projeto 5 (trio): Modelo de Regressão Múltipla

Analisar relação entre variáveis via modelos de regressão, considerando um problema real.

Suporte ao curso

1. **Blackboard**
2. **Github**
3. **Anaconda – Jupyter notebook**

Do que lembramos?

Atividade: Explorando dados reais

Fast Food.xlsx

Dom2014.txt

Próxima aula...

1. Leitura prévia necessária: Magalhães e Lima (7ª. Edição):

   pág. 9 a 16 – destacando para variáveis qualitativas.

2. **INSTALAÇÃO do ANACONDA**

   (https://www.continuum.io/downloads).

3. Importação dos arquivos Dom2014.csv para Python.