Insper

Ciência dos Dados

Aula 01

Introdução à disciplina

Professora

Kelly Venezuela

1º semestre de 2019

Aula de hoje

1. **O que Ciência dos dados?**
2. **Aplicações**
3. **Programa de ensino (conteúdo e tarefas)**
4. **Quiz: Socrative**
5. **Atividade: Análise Exploratória com *Jupyter Notebook***

Cientista de dados: perfil

O que é Ciência dos Dados?

Sobretudo, um jeito de pensar:

- Associação
- Causa e efeito
- Previsões
- Identificar tendências

Quando:

- Não temos um modelo mecanístico completo
- Temos um modelo, mas há incerteza nos dados

Usos: Ciências aplicadas (engenharia), medicina, biologia, gestão, marketing, etc, etc

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

O que é Ciência dos dados?

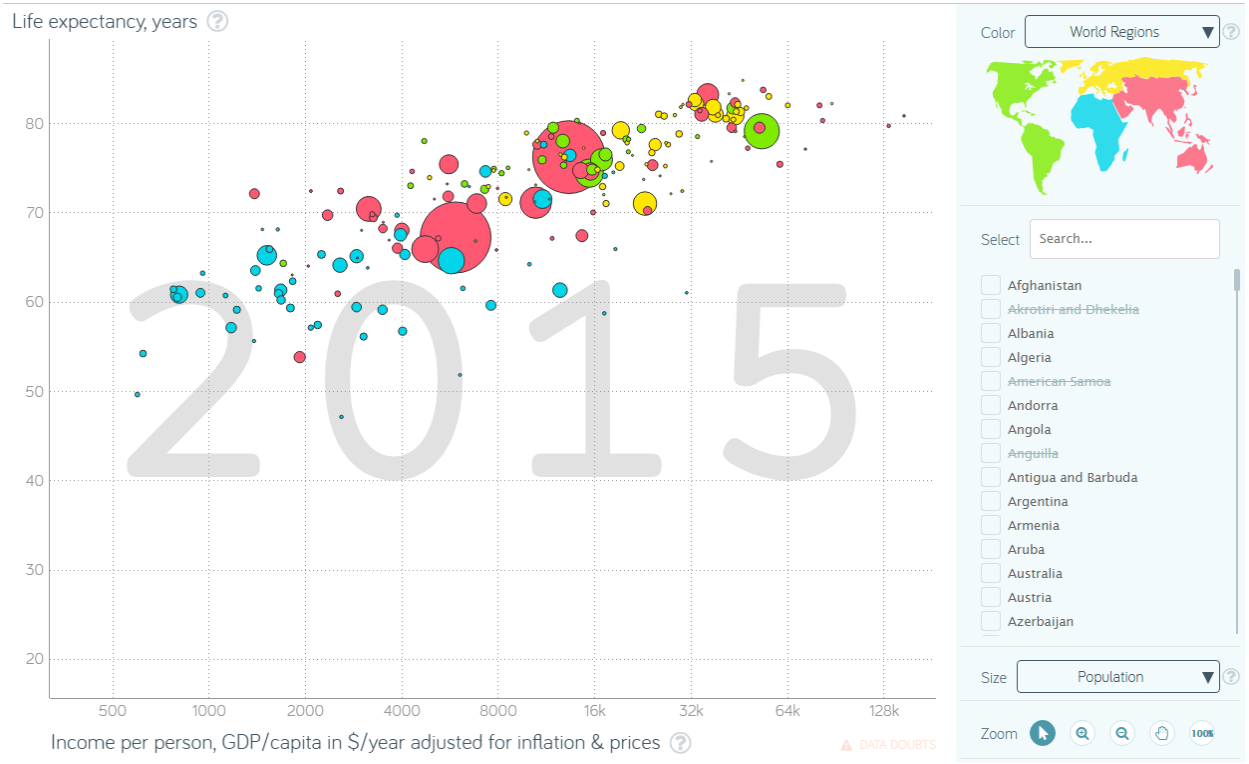
Funções típicas dos cientistas de dados

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.

Fonte: http://www.gapminder.org/tools/

Visualização de dados



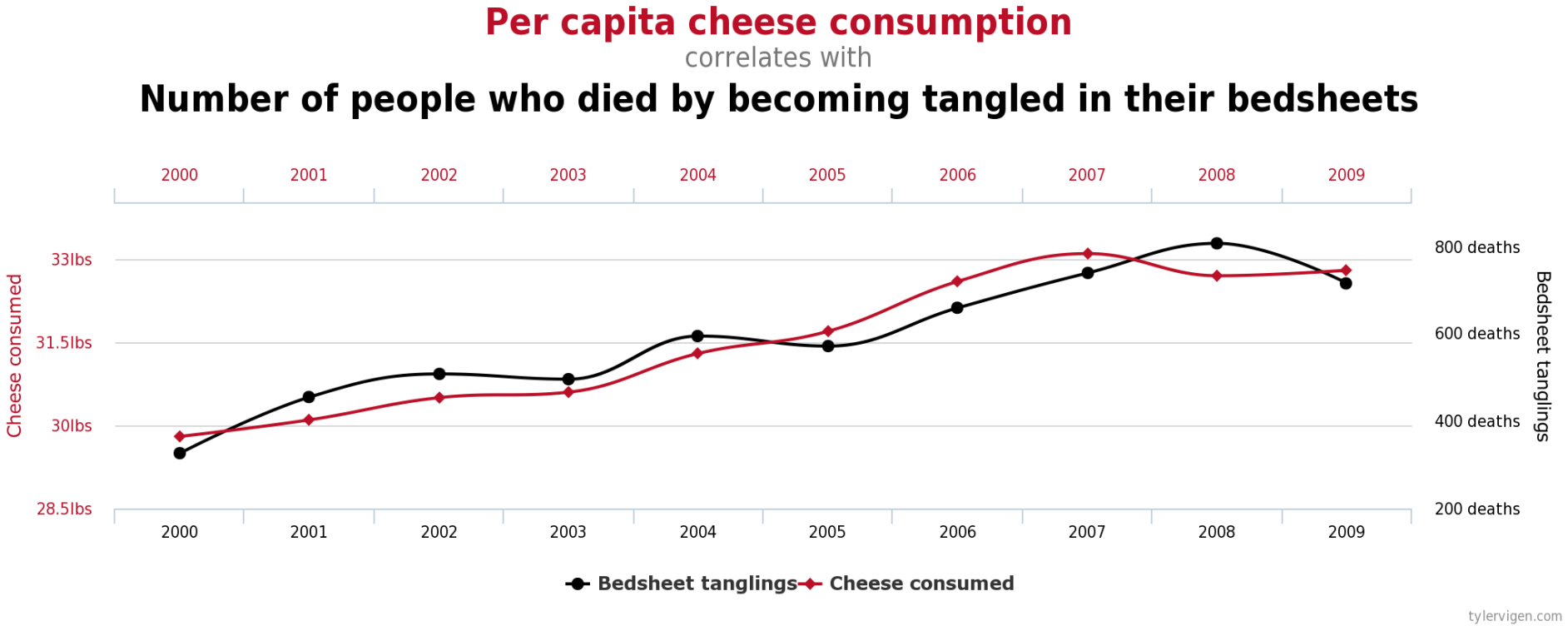
Assistam: The Best Statistics You've Ever Seen – TED
http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen

Usa os dados do World Bank que também usaremos no Projeto 1!

Insper

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

Insper

Critérios

Disciplina: Ciência dos dados

O que teremos / faremos, neste semestre?

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 inferenciais** adequados que possam dar base à tomada de decisão; e
- **Analisar relações entre as variáveis**, utilizando ferramentas estatísticas inferenciais adequadas.

NP - Nota de Projeto

Média simples entre os 3 projetos da disciplina:

Projeto 1 (DUPLA): Análise Descritiva

Análise exploratória sobre dados reais. **PNAD.**

Projeto 2 (TRIO): Filtro AntiSpam

Aplicação de Probabilidade usando Bayes.

Projeto 3 (QUARTETO): Inferência

"Livre", usando o ferramental de inferência (papel mais ativo e de maior engajamento do grupo).

ATENÇÃO: Todos os projetos devem ser entregues e nenhum deles pode ser considerado com conceito I.

Insper

NA - Nota de Avaliação

Média simples entre as duas avaliações:

Avaliação Intermediária (AI):
04/04 ou 09/04

Avaliação Final (AF):
06/06 ou 11/06

Avaliação Substitutiva (AS):
17/06 a 19/06

As datas das avaliações seguem o calendário do Insper.

O critério para a realização da avaliação substitutiva (AS) é o procedimento padrão adotado pelo Insper.

A AS irá englobar todo o conteúdo.

Insper

APS – Atividade Prática Supervisionada

- **APS 1: Check do Projeto 1– 14/02**
- **APS 2: Check do Projeto 1 – 21/02**
- **APS 3: Exercícios – 07/03**
- **APS 4: Check do Projeto 2 - 14/03**
- **APS 5: Exercícios – 28/03**
- **APS 6: Exercícios – 16/04**
- **APS 7: Exercícios – 25/04**
- **APS 8: Exercícios – 14/05**
- **APS 9: Exercícios – 23/05**

Nota Final da Disciplina

A **nota final** da disciplina será calculada da seguinte forma:

- média(NA,NP), se NA **e** NP forem maiores ou iguais a C ou 5 simultaneamente.
- min(NA, NP), caso contrário

Será usada a tabela oficial do Blackboard para converter conceito para nota numérica.

IMPORTANTE: é preciso ter mais de 50% das APSs entregues e validadas pelos ninjas para serem considerados satisfatórios. **A entrega com atraso terá tolerância de uma semana para marcação máxima ser amarela se 100% correto.**

Bibliografia básica

- MAGALHÃES, M. N.; DE LIMA, A. C. P. Noções de Probabilidade e Estatística (7a edição). Edusp, 2013.
- MONTGOMERY, D. Estatística Aplicada e Probabilidade para Engenheiros (6a edição). LTC, 2016.
- GRUS, J. Data Science do Zero: Primeiras Regras com Python. Alta Books, 2016.

Suporte ao curso

1. **Blackboard**

2. **Github**

?

3. **Anaconda – Jupyter Notebook**

Fonte: https://www.datascienceacademy.com.br

Jupyter Notebook

Ferramenta	Função
Jupyter Notebook	Shell interativo
NumPy	Arrays e matrizes
SciPy	Computação científica e álgebra linear
Matplotlib	Visualização de dados
Pandas	Series e Dataframes
Seaborn	Visualização de dados estatísticos
Scikit-Learn	Machine Learning
Bokeh	Visualização interativa
StatsModels	Bibliotecas para processamento estatístico
Scrapy	Web Crawler

Socrative

Vamos relembrar de alguns conceitos importantes para Análise Descritiva?

Entre em:

https://b.socrative.com/login/student/

Room Name:

INSPER

Atividade

Explorando dados reais:

- **Blackboard**
- **Fazer individual e discutir em grupo**
- **Submeter via Blackboard no final da aula**

Próxima aula...

1. **Leitura prévia necessária:**

    i. **Tutorial de Pandas via Jupyter**

    ii. Magalhães e Lima (7ª. Edição): pág. 1 a 13 – destacando para variáveis qualitativas.

2. **INSTALAÇÃO do ANACONDA**

    - **https://www.anaconda.com/distribution/**
    - **https://repo.continuum.io/archive/**