Insper

Ciência dos Dados

Aula 01

Introdução à disciplina

Professora:

Kelly Venezuela

1º semestre de 2018

Aula de hoje

1. **O que Ciência dos dados?**
2. **Aplicações**
3. **Programa de ensino (conteúdo e tarefas)**
4. **Atividade: Socrative**

Cientista de dados: perfil

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.bigdatabusiness.com.br/aprenda-mais-sobre-big-data-vendo-esses-7-cases-de-sucesso/

Aplicações

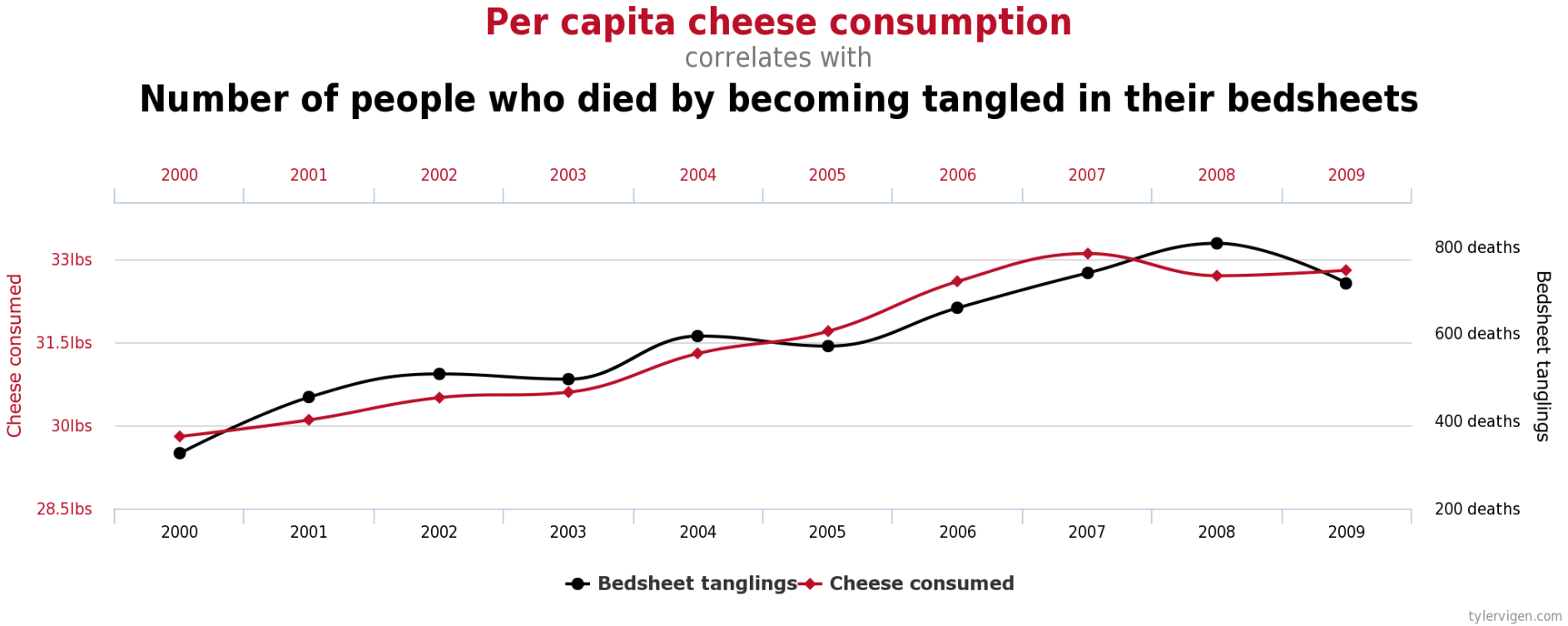
Até algumas décadas atrás, a simples contratação de uma empresa de pesquisas de mercado era suficiente para desvendar o comportamento do consumidor. Mas esse **universo empresarial estático ficou para trás**.

O dinamismo da troca de informações das mídias sociais, a formação de um novo consumidor, exigente, antenado e conectado 24 horas à web, ... fizeram com que a compreensão plena do mercado só pudesse ser alcançada por meio do trabalho aprofundado com centenas de variáveis internas e externas, a serem coletadas e processadas por poderosos softwares.

Big data (Ciência dos dados) permite às empresas mapearem e compreenderem plenamente seus consumidores, otimizarem seus processos de negócios ou enxergarem antes da concorrência uma eventual mudança de tendência.

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

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

Ciência dos dados

Define conceito (NP): nota dos projetos

Projeto 1 (individual): Análise Descritiva

Descrever perfil dos brasileiros no olhar socioeconômico e escrever uma matéria noticiando fato. **PNAD Pessoas.**

Projeto 2 (dupla): Filtro AntiSpam

Construir filtro antispam considerando teorema de Bayes.
Avaliar se algoritmo Naive Bayes tem overfitting e aprimorar algoritmo considerando várias bases treinamento/teste.

Projeto 3 (trio): Modelos Preditivos considerando Predição pela Média, k-Vizinhos mais Próximos, Regressão Linear e Árvore de Decisão

Dentre alguns temas, desenvolver projeto que seja de interesse do grupo (papel mais ativo e de maior engajamento do grupo).

Ciência dos dados

Define conceito (NA):

nota das avaliações

Avaliação Intermediária (AI):
05/04 ou 10/04

Avaliação Final (AF):
07/06 ou 12/06

Avaliação Substitutiva (AS):
13 ou 14/06 (apenas se faltou em uma das avaliações anteriores).

Ciência dos dados

Define conceito (APS):

Atividade Prática Supervisionada

- **APS 1: Check no próximo atendimento – 22/02**
- **APS 2: Check no próximo atendimento – 01/03**
- **APS 3: Check no próximo atendimento – 15/03**
- **APS 4: Exercícios – 22/03**
- **APS 5: Exercícios – 29/03**
- **APS 6: Exercícios – 17/04**
- **APS 7: Exercícios – 03/05**
- **APS 8: Exercícios – 22/05**

Regra de avaliação da disciplina

A disciplina terá basicamente três notas: **NA**, **NP** e **APS**, as quais serão formadas da seguinte forma:

- **NA – Nota das avaliações:**

Nota numérica (0 a 10) composta pela média simples das notas obtidas nas duas avaliações (AI e AF).

- **NP – Nota dos projetos:**

Nota numérica composta pela média simples dos três projetos, transformando antes o conceito de cada projeto em valor numérico.

Conceito	A+	A	B+	B	C+	C	D	I
Valor numérico	10	9	8	7	6	5	4	2

- **APS – Nota final das APSs:**

A APS poderá receber uma das duas notas: 10 (dez), se houver a entrega de pelo menos 50% das atividades prática supervisionada; ou 0 (zero), caso contrário.

Regra de avaliação da disciplina

Cada uma dessas notas refletirá o aprendizado do aluno nos objetivos já especificados anteriormente. Assim, para **obter uma nota satisfatória nessa disciplina, precisa ser aprovado (nota ≥ 5) de forma independente nos subconjuntos descritos como NA, NP e APS**.

A **nota final (NF) na disciplina** será expressa de forma numérica e se dará da seguinte forma:

- **Média simples entre NA e NP,** se todas as notas (NA, NP e APS) forem maiores ou igual a 5; ou
- **Menor nota entre NA e NP,** se NA ou NP for menor do que 5 e APS for maior ou igual a 5; ou
- **Nota da APS,** se APS for menor do que 5, não importando os valores das notas NA e NP.

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.

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:

CD2018

Insper

Definições

Definições

População e Amostra

População: é a coleção completa de todos os elementos (escores, pessoas, medidas, animais, índices, etc) a serem estudados.

Amostra: é um subconjunto de membros de uma população.

Dados: são as informações obtidas de uma unidade experimental ou de uma observação.

Variável: é toda característica que, observada em uma unidade experimental, pode variar de uma unidade para outra.

Definições

Tipos de variáveis

Cada variável tem um tipo de classificação, o qual auxilia buscar técnicas estatísticas mais adequadas. Esses tipos de classificação são:

Qualitativa (ou **categórica** ou **de atributo**): as respostas da variável podem ser separadas em diferentes classes (categorias) que se distinguem por alguma característica não numérica.

Cada variável **qualitativa** pode ser definida:

NOMINAL ou ORDINAL

Quantitativa ou Numérica: as respostas consistem em números que representam, em geral, contagem ou medidas.

Cada variável **quantitativa** pode ser definida:

DISCRETA ou CONTÍNUA

Definições

Tipos de variáveis

As variáveis **Qualitativas** são classificadas em:

Ordinal: a sequência dos resultados dessas variáveis *tem ordem* natural

Exemplos:

Qualidade de atendimento: Ruim, Bom, Ótimo e Excelente

Perfil de fundos de investimentos: Conservador, Moderado e Arrojado

Nominal: a sequência dos resultados dessas variáveis *não tem ordem* natural (apenas um rótulo)

Exemplos:

Cor do veículo: Prata, Branco, Azul, Preto …

Sexo: Feminino e Masculino

Resposta de pesquisa: Sim, Não e Indeciso

Definições

Tipos de variáveis

As variáveis **Quantitativas** são classificadas em:

Discreta: uma característica desse tipo de variável é que entre dois resultados consecutivos não existe nenhum valor intermediário, geralmente é uma contagem.

Exemplos:

Número de acidentes de trabalho;

Número de vezes que perdeu em um jogo;

Contínua: resultam de infinitos valores possíveis que correspondem a alguma escala contínua que cobre um intervalo de valores sem vazios, interrupções ou saltos.

Exemplos:

Inflação (em %);

Salário (em R$);

Preço de venda de um determinado bem (em R$).

Próxima aula...

1. **Leitura prévia necessária:**

   i. **Tutorial via Jupyter**

   ii. Magalhães e Lima (7ª. Edição): pág. 9 a 16 – destacando para variáveis qualitativas.

2. **INSTALAÇÃO do ANACONDA (se não tem) (https://www.continuum.io/downloads).**

3. **Importação dos arquivos da PNAD para Python.**