Configurações:
> Título: Sans Bold 13
> Descrição: Sans Regular 11

DATA SCIENCE: SEMANA 1 - Introdução a conceitos básicos

O que é ciência?

> Começa com a observação
> Formular hipóteses
> Inicialmente a observação é neutra
> Raciocínio indutivo
> René Descartes - PLANO CARTESIANO
    -O que é verdade? Negar o que lee acreditava dentro de si
    -A única verdade é o que ele pensava, assim, penso logo existo
> Edgard Morin - PENSAMENTO COMPLEXO
    -Não estudar fenômenos separados
> Estatística
    -O viés da variável omitida
> Senso crítico
    -Ir além das nossas análises
> Todo bebê conhece o método científico
    -Fazer uma observação
    -Formular uma hiótese
    -Realizar um experimento
    -Reportar as descobertas
    -Convidar terceiros para replicar os resultados
> Utilizar métodos e metodologias científicas
> Don't follow the hype
    -Questionar sempre

MÉTODO CIENTÍFICO

Big Data - Um grande problema

> Explosão do volume de dados
> IoT
> Problema de transferência de dados
> 5 v's of data
    - Velocidade = velocidade que os dados se transformam/chegam para você
    - Volume = quantos dados são gerados a cada segundo
    - Variedade = quais tipos de dados eu tenho
    - Veracidade = como garantir que os dados são confiáveis
    - Valor = se não ta gerando valor com o dado porque está sendo armazenado?
> Governança de dados
    - Para onde vou, como vou, quando vou?
    - Que dados são necessários?
    - Como manter?
    - Quais áreas são mais importantes e que eu devo manter os dados e dar uma melhor curadoria

Papéis dentro de um projeto de dados

> Domínio no que vamos trabalhar
    - Marketing
    - Medicina
> CIentista de dados (Data Scientist)
> Gerente de Data Science (Data Science Manager)
> Arquiteto de dados (Data Architect)
> Engenheiro de dados
> Estatístico

Tipos de soluções de análise

> Solução descritiva (Descriptive Analytics)
    - Caracterizar o passado
    - Descrever o que existe
    - O que aconteceu? Quantas peças eu vendi?
> Solução diagnóstica
    - Por que isso aconteceu?
    - Passado/Atual
    - Descreve os porquês
> Solução preditiva (Predictive Analytics)
    - O que vai existir
    - Usa o passado para prever/predizer o futuro
    - O que vai acontecer?
        * Previsão: subconjunto da predição vinculado ao tempo
        * Predição: forma genérica de como um fato seria antes de sua ocorrência
> Solução prescritiva (Prescriptive Analytics)
    - Recomenda tomada de decisão
    - O que vou fazer?
    - Usando uma predição/previsão recomenda uma ação

Tipos de problemas - o que a ciência de dados resolve

> Classificação binária
    - Prevê resultado binário de uma classe
    - Regressão logística = atacar um problema menos complexo para um mais complexo
        * O cliente comprará este produto?
        * Este e-mail é spam ou não é?
        * Este produto é um livro ou um animal de fazenda?
> Classificação multiclasse
    - Permite gerar várias classes
    - Regressão logística multinomial
        * Este produto é um livro, animal ou roupa?
        * Este filme é uma comédia romântica, um documentário ou um suspense?
> Regressão
    - Prever um valor numérico
        * Temperatura amanhã
        * Quantas unidades serão vendidas

>Agrupamento ou clusterização
- Modelo retorna um grupo
- K-means
* Quantos perfis de cliente minha empresa tem?
* Qual será o preço de venda desta casa?

> Sistemas de recomendaçãp
- Modelo sugere/recomenda algo a um usuário
- Retorna sugestão
- Collaborative filtering = criar recomendação a partir de perfis parecidos com o do usuário
- Content Based = produto mais propenso a consumir
* Qual filme recomendar a um usuário baseado em seu gosto?
* Qual imóvel comprar?

Tipos de aprendizado

> Aprendizado supervisionado: apresento a saída que eu quero que ele aprenda
- Objetivo: aprender uma regrageral que mapeia as entradas para entender as saídas
- Exemplo: imagem do Tom e Jerry --> Ao inserir uma entrada identificar se é o Tom ou o Jerry
- São apresentadas ao computador exemplos de entradas e saídas desejadas, fornecidas por um "professor". O objetivo é aprender uma regra geral que mapeia as entradas para as saídas

> Aprendizado não supervisionado: não se passa nenhum tipo de etiqueta para o meu dado.
- O objetivo principal é o algoritmo encontrar padrões nos dados
- Nenhum tipo de etiqueta é dado. Encontrar estrutura nas entradas fornecidas sozinho.
- Pode ser um objetivo em si mesmo(descobrir novos padrões nos dados) ou um meio para atingir um fim.
- Exemplo
* Mandar as fotos do Tom e Jerry
* Identificar o que separa os dois

Maturidade com analytics

> O sucesso ou o fracasso de um modelo de DS está ligado diretamente a maturidade com analytics

> Maturidade com os dados
- Como essa empresa tem trabalhado com esses dados
- Dados Crus (Raw Data) = pouca maturidade com os dados
- Gerão informação (Information) = já usa software, relatório automatizado, ainda é pouco o que a empresa utiliza os dados; não há padronização dos dados
- Bussiness intelligence = empresa entendeu que os dados são importantes e manipulam coisas básicas com os dados
- Advanced analytics = empresa já tem conhecimento profundo quanto aos seus dados

> Maturidade analítica: tipos de soluções de análise, empresa que entende o que aconteceu, o que vai acontecer, o que fazer dado que eu sei o que vai acontecer

Data Science: Semana 2 - Manipulação de dados

→ Comandos importantes para a manipulação dos códigos

- Criar ambiente: conda create -n nome
- Para ativar o ambiente criado: conda activate nome
- Instalar pacotes: pip install nomepacote

→ Trabalhando com Streamlit App

↳ Cria uma app web com estruturação pronta

Para executar um arquivo .py que seja um streamlit: streamlit run nome.py

↘ Para que dê tudo certo, precisa ativar o ambiente anaconda com o comando anterior

Para ativar ambiente virtualenv => source venv/bin/activate

Arquivos módulo 2

- Minimally Sufficient Pandas
  https://medium.com/dunder-data/minimally-sufficient-pandas-a8e67f2a2428
- Why and How to Use Pandas with Large Data
  https://towardsdatascience.com/why-and-how-to-use-pandas-with-large-data-9594dda2ea4c
- Getting started with Data Analysis with Python Pandas
  https://towardsdatascience.com/getting-started-to-data-analysis-with-python-pandas-with-titanic-dataset-a195ab043c77
- Python Pandas: Tricks & Features You May Not Know
  https://realpython.com/python-pandas-tricks/
- Essential basic functionality
  https://pandas.pydata.org/pandas-docs/stable/getting_started/basics.html
- Pandas Tutorial: Essentials of Data Science in Pandas Library
  https://medium.com/@shakasom/pandas-tutorial-essentials-of-data-science-in-pandas-library-9b0c81dbfcb1
- Python Pandas Tutorial: A Complete Introduction for Beginners
  https://www.learndatasci.com/tutorials/python-pandas-tutorial-complete-introduction-for-beginners/
- Basic Time Series Manipulation with Pandas
  https://towardsdatascience.com/basic-time-series-manipulation-with-pandas-4432afee64ea
- Tidy Data
  https://r4ds.had.co.nz/tidy-data.html
- Python For Data Science - Cheat Sheet Pandas Basics
  https://assets.datacamp.com/blog_assets/PandasPythonForDataScience.pdf

Diferença entre padronizar e normalizar dados

E a diferença básica é que padronizar as variáveis irá resultar em uma média igual a 0 e um desvio padrão igual a 1. Já normalizar tem como objetivo colocar as variáveis dentro do intervalo de 0 e 1, caso tenha resultado negativo -1 e 1.

Padronizar os dados normalmente é feita usando a fórmula *z-score*:

$$z = \frac{x - \mu}{\sigma}$$

z-score fórmula

Normalizar os dados usando *Min-Max:*

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Min-Max fórmula

Erros no terminal quando submete teste

Possíveis soluções:

- 0)Sair do antigo virtualenv conda deactivate
- 1) Criar um novo virtualenv virtualenv venv -p python3
- 2) Acessar ambiente virtualenv: source nome/bin/activate
- 3)Instalar pip install jupyter
- 4)Instalar pip install pytest
- 5)Instalar pip install pandas
- 6) remover ambiente virtual env: rm -r nome

Semana 3: Análise de dados exploratória

→ Análise de dados exploratória (EDA)

- Análise de dados exploratória é extremamente importante para o desenvolvimento de um modelo de Machine Learning.
- Procedimentos para analisar dados, técnicas para interpretar os resultados de tais proce dimentos, formas de planejar a reunião dos dados para tornar sua análise mais fácil.

---

* Sugerir hipóteses sobre as causas dos fenômenos observados;
* Avaliar pressupostos sobre os quais a inferência estatística se baseará;
* Apoiar a seleção de ferramentas e técnicas estatísticas apropriadas;
* Oferecer uma base para coleta posterior de dados por meio de pesquisas e experimentos

→ Estatística descritiva univariada

- Média: soma dos valores pela contagem
  M = SUM(valores)/COUNT(valores)
- Mediana: valor que separa a metade das informações em dois conjuntos de quantidades iguais
  Exemplo: [1, 2, 3, 4, 5, 6, 7, 8, 9] = Mediana é 6
  [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] = Mediana é 4,5
- Quartis:
  * Divide os dados em 4 conjuntos de dados
  * 25% em cada conjunto
- Percentis:
  * Divide os dados em 100 parte do todo
  * 1% acumulado em cada segmento
- Amplitude Interquartil: o 50% central dos valores quando ordenados do menor para o maior
  * Encontra-se a mediana(valor do meio) da menor e da maior metade dos dados
  * São o quartil 1(Q1) e o quartil 3(Q3). A amplitude interquartil é a diferença entre Q3 e Q1

  INTERQUARTILE RANGE = Q3 - Q1
- Desvio Padrão:
  * Medida que expressa o grau de dispersão de um conjunto de dados
  * Indica o quanto um conjunto de dados é uniforme
  * Quanto mais próximo de 0 for o desvio padrão, mais homogêneo são os dados

Desvio Padrão (Dp)

$$Dp = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

x_i = valor individual
$\bar{x}$ = média dos valores
n = número de valores

- Assimetria:
  * É o grau de distorção da curva simétrica a distribuição normal
  * Ele mede a falta de simetria entre os dados
  * Uma distribuição simétrica terá uma assimetria igual a zero
- Curtose: é uma medida de dispersão que caracteriza o "achatamento" da curva da função de distribuição
  * Mesórcuticas = 0: achatamento da distribuição normal
  * Leptocúrtica > 0: possui a curva da função de distribuição mais afunilada; pico mais alto do que a distribuição normal

* Platicúrtica < 0: função de distribuição é mais achatada do que a distribuição normal

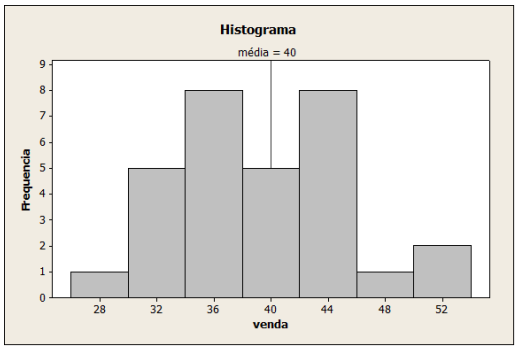
- Normalizar & Padronizar: transformar todas as variáveis na mesma ordem de grandeza
    * Padronizar as variáveis irá resultar em uma média igual a 0 e um desvio padrão igual a 1.
    * Normalizar tem comon objetivo colocar as variáveis dentro do intervalo de 0 e 1, caso tenha resultado negativo -1 e 1

> Estatística descritiva bivariada/multivariada

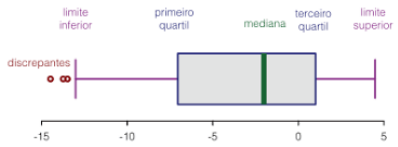
- Amostra possui mais de uma variável
- Relacionamento entre os pares ou conjunto de variáveis
    * Tabulações cruzadas e tabelas de contigência
    * Representação gráfica via gráfico de dispersão
    * As medidas quantitativas de dependência
    * As descrições de distribuição condicionais
- Correlação
    * Medida padronizada da relação entre duas variáveis
    * Indica a força e a direção do relacionamento de duas variáveis aleatórias
    * Está entre -1 >= correlação =< 1
        + Correlação próxima a zero = as duas variáveis não estão relacionadas
        + Correlação positiva: variáveis movem juntas; forte quanto mais se aproxima de 1
        + Correlação negativa:variáveis se movem em direções opostas; forte quanto mais se aproxima de -1
    * Duas variáveis que estão perfeitamente correlacionadas positivamente (r=1) se move essencialmente em perfeita proporção na mesma direção
    * Dois conjuntos que estão perfeitamente correlacionados negativamente (r=-1) se move em perfeita proporção em direções opostas
- Correlação de Spearman: avalia a relação monotônica entre duas variáveis contínuas ou ordinais
    * Relação monotônicas variáveis:
    ~ Mudar juntas mas não necessariamente a uma taxa constante
    ~ O coeficiente de correlação de Spearman baseia-se nos valores classificados de cada variável, em vez de dados brutos
- Correlação de Pearson: relação linear entre duas variáveis continuas
    * Relação linear: mudança em uma variável é associada a uma mudança proporcional na outra variável.

> Visualização

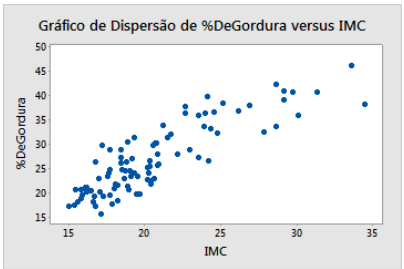
- Histogramas
    * Distribuição de frequências
    * Representação gráfica em colunas ou em barras
    * Dividido em classes uniformes ou não uniformes
    * A base de cada retângulo representa uma classe
    * A altura de cada retângulo:
        - Quantidade
        - Frequência absoluta com que o valor da classe ocorre no conj de dados p/ classes uniformes
        - Densidade de frequência para classes não uniformes



- Diagrama de caixa (Box Plot)
    * Ferramenta gráfica para representar a variação de dados observados de uma variável númerica por meio de quartis
    * Reta que estende-se verticalmente ou horizontalmente a partir da caixa, indicando a variabilidade fora do quartil superior e do quartil inferior
    * Outliers ficam como pontos individuais



- Dispersão
    * Representa duas(normalmente) ou mais variáveis
    * São exibidos como uma coleção de pontos, cada um com o valor de uma variável



- Matriz de correlação
    * Colocar os valores de correlação em uma matriz

Material de apoio da Semana 3

- How to self-learn statistics of data science
  https://medium.com/ml-research-lab/how-to-self-learn-statistics-of-data-science-c05db1f7cfc3

- Statistics Done Wrong
  https://www.statisticsdonewrong.com/index.html

- Exploratory Data Analysis | R for Data Science
  https://r4ds.had.co.nz/exploratory-data-analysis.html

- Exploratory Data Analysis
  https://itl.nist.gov/div898/handbook/eda/eda.htm

- A Gentle Introduction to Exploratory Data Analysis
  https://towardsdatascience.com/a-gentle-introduction-to-exploratory-data-analysis-f11d843b8184?gi=3450a93dbd2f

- A Simple Tutorial on Exploratory Data Analysis
  https://kaggle.com/pavansanagapati/a-simple-tutorial-on-exploratory-data-analysis

- Introduction to Hypothesis Testing
  https://us.sagepub.com/sites/default/files/upm-binaries/40007_Chapter8.pdf

- The Power of Visualization in Data Science
  https://towardsdatascience.com/the-power-of-visualization-in-data-science-1995d56e4208

- 15 Stunning Data Visualizations (And What You Can Learn From Them)
  https://visme.co/blog/examples-data-visualizations/

- 15 Insane Things That Correlate With Each Other
  http://tylervigen.com/spurious-correlations

Semana 4 - Continuação da análise de dados exploratória

> Amostragem
- é o processo de obtenção de amostrar, que são uma pequena parte de uma população
> População
- se refere a todos os membros de um grupo o qual queremos estudar
> Amostra
- se refere a um pequeno grupo de membros pertencentes a uma população

---

Alguns problemas

>Undercoverage Bias
- Acontece ao analisar muito poucas observações ou omitir segmentos inteiros de uma população
=> Ex.: conduzir uma pesquisa de satisfação de funcionários de um hospital durante o dia
>Funcionários do turno da noite não foram contemplados na pesquisa
>Self-selection Bias
- Acontece quando as pessoas que se dispuseram a participar da pesquisa diferem muito da real população
=> Ex.: fazer uma pesquisa online sobre um time de futebol
>Somente pessoas que provavelmente já torcem para o time respoderam a pesquisa
>Health-users Bias
- Acontece quando retiramos uma amostra somente da parte mais sadia da população
=> Ex.: realizar uma pesquisa sobre hábitos saudáveis em uma academia
>Pessoas que frequentam academia provavelmente tem hábitos mais saudáveis que a maioria da população
>Survivor Bias
- Consiste no erro lógico de nos concentrar em coisas ou pessoas que sobreviveram a algum processo enquanto ignoramos aqueles que foram eliminados devido a sua falta de visibilidade
=> Ex.: focar em partes do avião em que sofreram muitos tiros ao invés de onde não há tiros e os tripulantes não sobreviveram
>Se tripulantes não voltaram, aonde precisa ser reforçado a segurança é nesta parte do avião, não aonde há vários tiros e os tripulantes retornaram

---

Tipos de amostragens

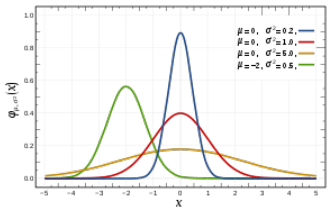
> Amostragem aleatória: retirar elementos aleatórios de dentro de uma população
> Amostragem não aleatória: retirar elementos selecionados de dentro de uma população
> Amostragem estratificada:
1) proporcional: a amostra deverá obter camadas que obtenham as mesmas proporções observadas na população
2) uniforme: a atribuir o mesmo tamanho de amostra para todas as camadas, independentemente do peso dos estratos da população

- O que precisamos entender para utilizar uma amostra estratificada?
+ Qual a população relevante?
+ Qual a população afetada?
+ Qual a subdivisão dessa população?
+ Qual o tamanho de cada sub divisão?
+ Todas as subdivisões são afetadas?

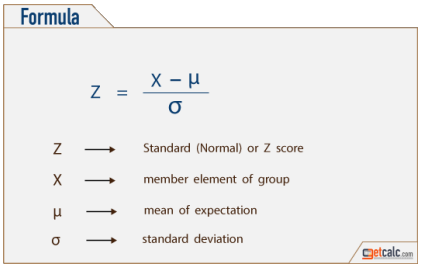
Revisão de probabilidades
> Variam de 0 a 1
> O resultado da soma das probabilidades resulta em 1
> A distribuição relaciona x com a probabilidade p(x)
> Podem ser funções discretas ou contínuas
+ Ex 1.: qual a probabilidade de se obter cara jogando uma moeda(justa) para cima? =>DISCRETA
+ Ex 2.: qual a probabilidade de uma pessoa tirar uma nota acima da média no Enem? => CONTÍNUA

> Distribuição normal ou Gaussiana
- Área sob essa curva determina a probabilidade de ocorrer o evento por ela correlacionado
- Soma da área sob a curva de densidade igual a 1
- Uma curva simétrica em torno do seu ponto médio, apresentando assim seu famoso formato de sino
- Média, mediana e moda dos dados possuem o mesmo valor

Fenômenos que são representados por uma distribuição gaussiana:
1) Altura e peso de uma população
2) Tamanho do crânio de recém nascidos
3) Pressão sanguínea

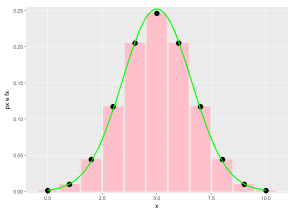


Cálculo do z-score



> Distribuição binomial
- Distribuição de probabilidade e estatística discreta(inteiros e contáveis) de uma determinada sequência de tentativas
- Espaço amostral finito
- Apenas dois resultados possíveis => sucesso ou fracasso
- Todos os elementos devem possuir possibilidades iguais de ocorrência
- Eventos devem ser independentes um dos outros

Exemplos:
1) Jogar uma moeda para o alto
2) Rolar um dado de 6 faces



Funções de PDF e CDF

> Função densidade de probabilidade(PDF):
- Descreve a probabilidade relativa de uma variável aleatória tomar um valor dados
- É não negativa sempre
- E sua integral sobre todo o espaço é igual a 1
- Informa a probabilidade de a variável X assumir um valor naquele intervalo

> Função distribuição acumulada (CDF):
- Descreve a probabilidade acumulada de uma variável aleatória tomar um conjunto de valores dado;
- É não negativa sempre
- E sua integral sobre todo o espaço é igual a 1
- Informa a probabilidade de a variável X assumir um valor naquele intervalo

> Função distribuição acumulada empírica(ECDF):
- Um estimador da função de distribuição cumulativa
- Modelo empírico
- Funciona com sua observação

Material de apoio semana 4

Probability Theory Review for Machine Learning
https://see.stanford.edu/materials/aimlcs229/cs229-prob.pdf

Understanding Probability Distributions
https://statisticsbyjim.com/basics/probability-distributions/

Probability Distribution
https://en.wikipedia.org/wiki/Probability_distribution

Statistical Modeling: The Two Cultures
http://www2.math.uu.se/~thulin/mm/breiman.pdf

Variáveis Aleatórias Unidimensionais
http://www.professores.uff.br/anafarias/wp-content/uploads/sites/32/2017/08/GET00182-DistNomal.pdf

Probability and Information Theory
https://www.deeplearningbook.org/contents/prob.html

Slide Módulo 4 I
https://acceleration-assets-highway.s3-us-west-1.amazonaws.com/ds-online-1/slide_modulo_4_I.pdf

Slide Módulo 4 II
https://acceleration-assets-highway.s3-us-west-1.amazonaws.com/ds-online-1/slide_modulo_4_II.pdf

Resumo sobre a utilização de GROUPBY

A função groupby() retorna um objeto do tipo GroupBy que descreve como as linhas do conjunto de dados originais foram divididas.

Toda operação envolvendo o groupby segue os seguintes procedimentos:
- Separar os dados;
- Aplicar uma função;
- Combinar os resultados.

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

DATAFRAME QUE SERÁ UTILIZADO PARA EXEMPLIFICAR OS COMANDOS

EXEMPLO 1: E se quiséssemos descobrir a média dos estudantes nos três testes realizados, podemos agrupar os estudantes pelo sexo e aplicar a função mean()

```
import pandas as pd

data = pd.read_csv('StudentsPerformance.csv')

grouped = data[['math score', 'reading score', 'writing score']].groupby(data['gender'])
grouped.mean()
```

gender	math score	reading score	writing score
female	63.633205	72.608108	72.467181
male	68.728216	65.473029	63.311203

EXEMPLO 2:O agrupamento não é limitado apenas para uma chave, podemos agrupar usando mais de uma.

```
grouped_keys = data[['math score', 'reading score', 'writing score']].groupby([data['gender'],
                                                data['test preparation course']])
grouped_keys.mean()
```

gender	test preparation course	math score	reading score	writing score
female	completed	67.195652	77.375000	78.793478
	none	61.670659	69.982036	68.982036
male	completed	72.339080	70.212644	69.793103
	none	66.688312	62.795455	59.649351

APLICANDO FUNÇÕES USANDO GROUPBY

Para usar suas próprias funções de agregação, passe elas para agg() ou aggregate().

EXEMPLO 3:agrupei os dados de acordo com race/ethnicity e parental level of education e em seguida calculei a média da nota de matemática para os grupos.

```
grouped = data.groupby(['race/ethnicity', 'parental level of education'])
grouped_math = grouped['math score']
grouped_math.agg('mean')
```

```
race/ethnicity  parental level of education
group A         associate's degree             61.000000
                bachelor's degree              67.166667
                high school                    60.444444
                master's degree                57.666667
                some college                   63.888889
                some high school               58.916667
group B         associate's degree             66.097561
                bachelor's degree              69.300000
                high school                    59.791667
                master's degree                67.166667
                some college                   63.189189
                some high school               61.815789
group C         associate's degree             66.730769
                bachelor's degree              68.150000
                high school                    60.906250
                master's degree                67.052632
                some college                   65.130435
                some high school               60.551020
group D         associate's degree             67.600000
                bachelor's degree              67.571429
                high school                    62.863636
                master's degree                72.521739
                some college                   68.731343
                some high school               66.760000
group E         associate's degree             74.897436
                bachelor's degree              76.555556
                high school                    70.772727
                master's degree                74.625000
                some college                   73.828571
                some high school               72.111111
Name: math score, dtype: float64
```

CRIANDO AMBIENTE VIRTUAL P/ PROGRAMAÇÃO

Para criar um novo ambiente virtual:
- virtualenv venv -p python3.7
- conda create -n env python3.7

Instalar pacotes recomendados do pacote:
- pip install -r requirements.txt

Para acessar o virtualenv criado:
- Entra na pasta codenation que contem a pasta venv
- digita no cmd: source venv/bin/activate

OBSERVAÇÃO: Tem que entrar no ambiente virtualenv para poder instalar os requirements :)

SEMANA 5

Método científico

> Complexidade

- contexto = o conhecimento das informações ou dos dados isolados é insuficiente. É preciso situar as informações em seu contexto para adquirirem sentido
- global = o global é mais que o contexto, é o conjunto das diversas partes ligadas a ele de modo inter-retroativo ou organizacional. Dessa maneira, uma sociedade é mais que um contexto: é o todo organizador do qual fazemos partes
- multidimensional = unidades complexas como o ser humano ou a sociedade são multidimensionais. O ser humano é ao mesmo tempo biológico, psíquico, social, afetivo e racional. A sociedade comporta as dimensões histórica, econômica, religiosa.. O conhecimento pertinente deve agregar esses dados e não favorecer nenhuma redução sintética.
- complexo = há complexidade quando elementos diferentes são inseparáveis constitutivos do todo (como o econômico, o político, o sociológico, o psicológico, o afetivo, o mitológico). Por isso a complexidade é a união entre a unidade e a multiplicidade

> Pesquisa

- Pode ser definida como o procedimento racional e sistemático que tem por objetivo proporcionar respostas aos problemas que são propostos.
- O que um pesquisar precisa ser/ter?

1) conhecimendo do assunto a ser pesquisado
2) curiosidade
3) criatividade
4) integridade intelectual
5) atitude autocorretiva
6) sensibilidade social
7) imaginação disciplinada
8) perseverança e paciência
9) confiança na experiência

- O início da pesquisa começa com um problema

> O problema científico

- questão não resolvida e que é objeto de discussão
- Como elaborar um problema científico?
  = Precisa relacionar coisas/variáveis
- CARACTERÍSTICAS DO PROBLEMA CIENTÍFICO

1) Deve ser formulado como pergunta
2) Deve ser claro e preciso

3) Deve ser empírico
4) Deve ser suscetível de solução
5) Deve ser delimitado a uma dimensão viável

> Construção de hipóteses
- A pesquisa científica se inicia com a descrição de um problema solucionável
- O próximo passo consiste em oferecer uma possível solução, mediante uma proposição, ou seja, uma expressão verbal suscetível de ser declarada como verdadeira ou falsa. //A isso denominamos hipóteses
- Assim, a hipótese é a proposição verificável de uma possível solução para o problema

COMO ELABORAR UMA HIPÓTESE?

- É de maneira criativa
- A qualidade do pesquisador é o domínio do tema
> Observação assistemática: o estabelecimento assistemático de relações entre os fatos do dia a dia é que fornece indícios para a solução dos problemas propostos pela ciência
> Intuição: hipóteses derivadas de simples palpites ou de intuições. Porém, as intuições, por sua própria natureza, não deixam claro as razões que determinaram, torna-se difícil, a priori, avaliar a qualidade dessas hipóteses.
> Observação sistemática: parte de uma observação interessada e metodicamente orientada. Dela extrai-se as características e/ou a frequência que ocorrem determinados fenômenos.
> Resultados de outras pesquisas = revisão bibliográfica
> Teorias

- Uma hipótese simples é sempre preferível a uma mais complexa, desde que tenha o mesmo poder explicativo.

PENSAMENTO ESTATÍSTICO EM PYTHON

Teorema do limite central: a distribuição das médias retiradas de uma amostra cuja população tenha qualquer distribuição terão uma distribuição normal.

Teste estatístico(hipótese):
- Para que serve? usaremos como ferramenta de validação de que o caminho que estamos seguindo possui alguma significância ou não a nível estatístico
=> Permite a tomada de decisão entre 2 ou + hipóteses
=> Define-se uma hipótese nula (H0) e uma ou mais hipóteses alternativas (H1,H2,...HN)
- H0 é assumida como verdadeira; aquilo que se quer testar
- HN é considerada quanto H0 é falsa (não possui relevância estatística)

* Testes de média: a garrafa de cerveja padrão tem 600ml com uma amostra de 50 garrafas podemos continuar dizendo que uma garrafa possui 600ml?
* Testes de proporção: uma fábrica declara que no máximo 5% da sua produção vem com defeito, em uma amostra de 100 unidades encontramos 7 defeituosas. Os números da fábrica estão corretos?

* Erro tipo I = falso positivo //exemplo do homem estar grávido
* Erro tipo II = falso negativo //exemplo da mulher que está com sinais de gravidez não estar grávida

* P - VALOR: probabilidade de obter uma estatística de teste igual ou maior que a observada em uma amostra para H0(hipótese nula)
=> Quanto maior o valor de P maior a chance de se rejeitar a hipótese nula
=> ALPHA é definido antes do experimento //p value hacking

TIPOS DE TESTES

1) T-test (Student T-Test)
- Baseia-se na distribuição t student
+ Dist. simétrica
+ Semelhante a curva normal
+ Grau de liberdade

- Comparar 2 grupos
    + Utilizando a média entre os valores
    + Desvio padrão e variância

2) Shapiro-Wilk
- Teste de normalidade
    + Utiliza uma amostra de uma população para validar se a mesma está distribuida normalmente
    + Calcular a probabilidade usando essa distribuição normal
- Variância, média e desvio padrão
- H0: amostra provém de uma normal
- H1: amostra não provém de uma normal
- Ela não funciona bem com + de 5000 amostras

3)Jarque-Bera
- Teste de normalidade
    + Validar se existe desvio padrão
- Curtose e assimetria
- H0: amostra provém de uma normal
- H1: amostra não provém de uma normal

4) Gráfico Q-Q Plot
- Gráfico quantil-quantil
- compara a distribuição de duas probabilidades
    + entre duas variáveis
    + entre uma variável e "quartis teóricos"
- ajuda a validar se uma distribuição é normal

MATERIAL DE APOIO - SEMANA 5

A Gentle Introduction to Statistical Hypothesis Testing
https://machinelearningmastery.com/statistical-hypothesis-tests/

How to Correctly Interpret P Values
https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-correctly-interpret-p-values

A Dirty Dozen: Twelve P-Value Misconceptions
http://www.perfendo.org/docs/BayesProbability/twelvePvaluemisconceptions.pdf

An investigation of the false discovery rate and the misinterpretation of p-values
https://royalsocietypublishing.org/doi/pdf/10.1098/rsos.140216

Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations
https://link.springer.com/content/pdf/10.1007%2Fs10654-016-0149-3.pdf

Why Are P Values Misinterpreted So Frequently?
https://statisticsbyjim.com/hypothesis-testing/p-values-misinterpreted/

Statistical Significance Explained
https://towardsdatascience.com/statistical-significance-hypothesis-testing-the-normal-curve-and-p-values-93274fa32687

Definition of Power
https://newonlinecourses.science.psu.edu/stat414/node/304/

The Math Behind A/B Testing with Example Python Code
https://towardsdatascience.com/the-math-behind-a-b-testing-with-example-code-part-1-of-2-7be752e1d06f

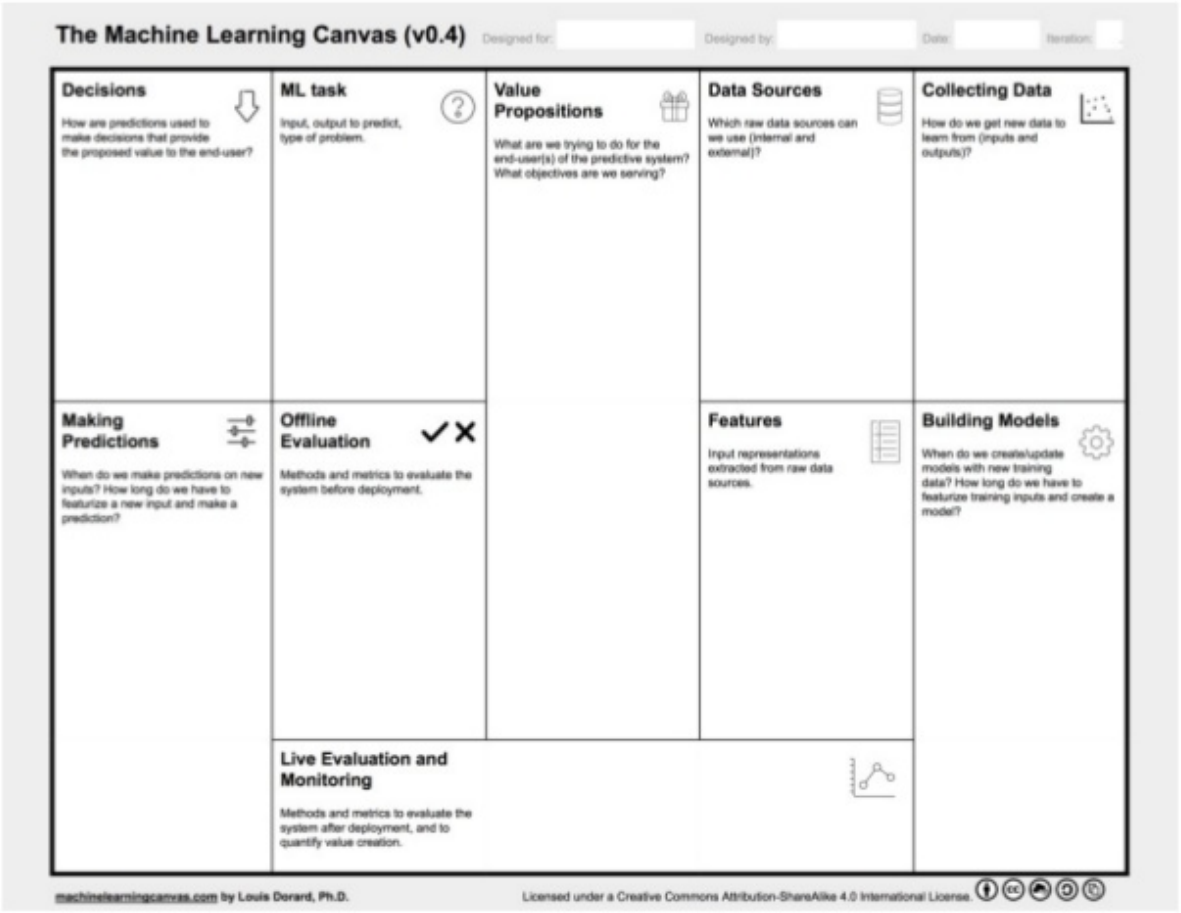
Handy Functions for A/B Testing in Python
https://medium.com/@henryfeng/handy-functions-for-a-b-testing-in-python-f6fdff892a90

Slide da aula
https://acceleration-assets-highway.s3-us-west-1.amazonaws.com/ds-online-1/pensamento_estatistico_python_aula_1.pdf

SEMANA 6 - Seleção de variáveis

COMEÇANDO MEU PROJETO DE DADOS

Utilizando Machine Learning Canvas



EXEMPLO DE UM MACHINE LEARNING CANVAS

> Confecção da base para treinamento
  - Mapa Mental

MATERIAL DE APOIO

StackExchange - Relationship between SVD and PCA. How to use SVD to perform PCA?
https://stats.stackexchange.com/questions/134282/relationship-between-svd-and-pca-how-to-use-svd-to-perform-pca

In Depth: Principal Component Analysis
https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html

In-Depth: Manifold Learning
https://jakevdp.github.io/PythonDataScienceHandbook/05.10-manifold-learning.html

Recursive Feature Elimination
https://bookdown.org/max/FES/recursive-feature-elimination.html

A Tutorial on Principal Component Analysis
http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Principal Component Analysis Explained
https://www.kaggle.com/nirajvermafcb/principal-component-analysis-explained

Step Forward Feature Selection: A Practical Example in Python
https://www.kdnuggets.com/2018/06/step-forward-feature-selection-python.html

Semana 7 - Engenharia de feautures

Material de apoio

- Feature Engineering
  https://jakevdp.github.io/PythonDataScienceHandbook/05.04-feature-engineering.html

- Feature Scaling with scikit-learn
  http://benalexkeen.com/feature-scaling-with-scikit-learn/

- Anthony Goldbloom gives you the secret to winning Kaggle competitions
  https://www.import.io/post/how-to-win-a-kaggle-competition/

- What are some best practices in Feature Engineering?
  https://www.quora.com/What-are-some-best-practices-in-Feature-Engineering

- Machine Learning Mastery
  https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/

- Fundamental Techniques of Feature Engineering for Machine Learning
  https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114

- Feature Engineering Cookbook for Machine Learning
  https://medium.com/@michaelabehsera/feature-engineering-cookbook-for-machine-learning-7bf21f0bcbae

- Outlier detection with Scikit Learn
  https://www.mikulskibartosz.name/outlier-detection-with-scikit-learn/

- Working With Text Data
  https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html

- WTF is TF-IDF?
  https://www.kdnuggets.com/2018/08/wtf-tf-idf.html

-

Semana 8

Material de apoio

- Gentle Introduction to the Bias-Variance Trade-Off in Machine Learning
  https://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning/

- Understanding the Bias-Variance Tradeoff
  http://scott.fortmann-roe.com/docs/BiasVariance.html

- Introduction to Machine Learning Algorithms: Linear Regression
  https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a

- 7 Classical Assumptions of Ordinary Least Squares (OLS) Linear Regression
  https://statisticsbyjim.com/regression/ols-linear-regression-assumptions

- Statistics By Jim
  https://statisticsbyjim.com/regression/gauss-markov-theorem-ols-blue

- Tikhonov regularization
  https://en.wikipedia.org/wiki/Tikhonov_regularization

- Ridge Regression for Better Usage
  https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db

- Lasso (statistics)
  https://en.wikipedia.org/wiki/Lasso_(statistics)

- Understanding Linear Regression and Regression Error Metrics
  https://www.dataquest.io/blog/understanding-regression-error-metrics

- Understand Regression Performance Metrics
  https://becominghuman.ai/understand-regression-performance-metrics-bdb0e7fcc1b3

-

Semana 9

- TensorFlow on Google
  https://acceleration-assets-highway.s3-us-west-1.amazonaws.com/ds-online-1/TensorFlow_on_Google_Cloud.pdf

- Confusion matrix and other metrics in machine learning
  https://medium.com/hugo-ferreiras-blog/confusion-matrix-and-other-metrics-in-machine-learning-894688cb1c0a

- Let's learn about AUC ROC Curve!
  https://medium.com/greyatom/lets-learn-about-auc-roc-curve-4a94b4d88152

- Classification Algorithms Comparison
  https://www.kaggle.com/metetik/classification-algorithms-comparison

- Having an Imbalanced Dataset? Here Is How You Can Fix It
  https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb

- FOUNDATIONS OF IMBALANCED LEARNING
  https://pdfs.semanticscholar.org/1678/7e213ed0a5c0cf9baabdb45f9df631248a91.pdf

- DATA MINING FOR IMBALANCED DATASETS: AN OVERVIEW
  https://www3.nd.edu/~dial/publications/chawla2005data.pdf

- An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain
  https://pdfs.semanticscholar.org/3305/2b1d2363aee3ad290612109dcea0aed2a89e.pdf

- Explaining the Success of Nearest Neighbor Methods in Prediction
  https://devavrat.mit.edu/wp-content/uploads/2018/03/nn_survey.pdf

- Classification: Basic concepts, decision trees, and model evaluation
  https://www-users.cs.umn.edu/~kumar001/dmbook/ch4.pdf