=======================================

============ Curso Udemy ==============

=======================================

DETALHES DAS AULAS ESTÃO NOS SLIDES

Aula 52.

Conversão de tipos de array em Numpy:

array.astype(np.int32)

Método para criação de matriz com valor diagonal 1 e os demais zerdos:

np.eye(5) ##cinco linhas

O método random(), que funciona mais ou menos como o range em questão de passagem de parâmetro porém ele gera os valores POSITIVOS randomicamente:

np.random.random((3,3))

array([[ 0.07878569, 0.7176506 , 0.05662501],

[ 0.82919021, 0.80349121, 0.30254079],

[ 0.93347404, 0.65868278, 0.37379618]])

Já o Método randn gera valores Literais ( positivos e negativos ) randomicamente.

valoresLiterais = np.random.randn(3,3)

Detalhe que o random precisa de duas chaves (()) enquanto o randn apenas uma ().

Método UNIQUE é equivalente ao distinct.

unique = np.array([11,11,23,18,19,17,11,18,16,20])

distintos = np.unique(unique)

array([11, 16, 17, 18, 19, 20, 23])

Aula 63.

Tratativa de dados

Usar o comando "sep=';'" após o caminho do arquivo csv para ele já passar para o dataframe com com os valores separados pro ponto e virgúla.

Comando para agrupar dados:

agrupado = df.groupby(['Estado']).size()

Estado

PR 257

RP 1

RS 478

SC 258

SP 4

TD 1

Podemos utilizar o método "describe()" em um campo para pegar alguns valores importantes dele e já fazer algumas analises iniciais.

df['Saldo'].describe()

count 9.990000e+02

mean 7.164928e+06

std 6.311840e+06

min 0.000000e+00

25% 0.000000e+00

50% 8.958835e+06

75% 1.258684e+07

max 2.117743e+07

Name: Saldo, dtype: float64

Sempre validar antes os valores NaNs (nulos) na base, aqui no caso pode-se utilizar do comando a baixo para fazer essa validação:

df.isnull().sum()

Aula 69.

Acrônimo de coisas a se observar no seu gráfico/dash

A.C.C.E.N.T

Dashboard != Números

-Dashboard não pode ter muitos números 20 ou menos.

-Variar dados e não design.

-Conseguimos guardar de 3 a 9 informações visuais em nosso cerebro.

-Considerar o daltonismo na escolha das cores.

ESTÁTISTICA

Aula. 94

Campos básicos da estatística:

-Probabilidade: Analisar situações sujeitas ao acaso

-Inferência: Obter respostas sobre um fenômeno com dados representativos

( É pegar uma amostra e dela inferir o resultado do todo ) Caso dos eleitores de não precisarem questionar a todos e sim uma parte e dessa parte sugerir o todo.

-Observação e Experimento: Observação é o estudo em que os elementos analisados não são afetados(pesquisa) e experimentos são as condições ou tratamentos que são impostas a grupos, para avaliar o resultado.

variáveis:

Quantitativas - numéricas:

Continuas: Valores reais, podem assumir qualquer intervalo.

Discretas: números fixos, inteiros e em um intervalo

Qualitativas - Categóricas

Nominais: categorias sem hierarquia

ordinais: categoria com hierarquia

EDA ( Exploratory data analysis )

Busca obter informações ocultas sobre os dados

Variações

Anomalias

Distribuição - Como os dados estão distribuídos, mais a esq, dir,

tendencia de subida ...

Tendências

Padrões

Relações

Conceitos:

População: Alvo do estudo

Amostra: Subconjunto da população

Censo: Pesquisa com toda a população

Enviesamento:

Você subestima ou superestima o parâmetro da população

Principais tipos de amostras:

Aleatória simples

Estratificada

Sistemática

Aleatória simples:

-Um determinado número de elementos é retirado da população de forma aleatória ( Usando um algoritmo de aleatoriedade por ex )

-Todos os elementos da população alvo do processo de amostragem, devem ter as mesmas chances de serem selecionados para fazer parte da amostra.

-Ele pode ser com reposição ou sem reposição. Que seria poder ou não repetir a mesma amostra ou número.

Amostra estratificada:

As vezes as populações estão divididas nos chamados estratos.

ex de estratos: Raça, escolaridade ... (Grupos/categoria/classe da amostra)

Amostra sistemática:

-Neste tipo de amostragem, é escolhido um elemento aleatório, e a partir daí, a cada N elementos um novo membro é escolhido.

Seria pegar um elemento a cada x elementos. Então é escolhido um número ex 5 e ao pegar uma amostra pula 5 amostras e seleciona outra ...

Usado por exemplo quando você não sabe o número total da população.

Medidas de Centralidade:

Média

Moda

Mediana

Média:

-A soma de todos e a divisão pela quantidade de elementos somados.

-Média ponderada é cada número está sujeito a um peso:

Prova A peso 1, nota 7

Prova B peso 2, nota 9

Cálculo: p = (((7\*1) + (9\*2))/(1+2)) = 8,3

A nota \* o peso, soma ambas multiplicações e divide pela soma dos pesos.

Moda:

-O valor mais frequente, o que mais ocorre.

{1,2,3,4,5,6,7,8,9} - não tem moda

{1,2,2,5,7,8,8} - 2,8

Mediana:

-Valor médio

-Ordena-se os valores em ordem crescente

-Se a quantidade for par, a mediana será média dos valores nas posições n/2 e n/2+1

-Se a quantidade for impar, a mediana será o valor na posição(n+1)/2

Media x mediana:

10,20,30,40,10.000

Media: 2.020

Mediana: 30

Aula. 99

Variabilidade:

Variância:

É como os dados se alteram entre si, é a regularidade dos dados.

O quanto esses se diferenciam, então mesmo tendo valores na casa dos milhões porém a variação entre esses valores for baixa, significa que a variância será um valor baixo.

Desvio padrão:

É a raiz quadrada da variância.

Amplitude:

A diferença entre o maior e menor valor.

Medidas não centrais:

Q1: 25% dos menores valores

Q2: 50% (Equivalente a mediana)

Q3: 75% dos maiores valores

Pode haver também o Q0 (Menor valor) e o Q4 (Maior valor).

Aula .101

Probabilidade:

0 - Impossível de acontecer.

1 - Com certeza acontecerá.

Conceitos:

Experimento: O que está sendo estudado

Espaço amostral: Todas as possibilidades de ocorrência do evento

Evento: resultados ocorridos

Eventos excludentes:

Quando não podem ocorrer ao mesmo tempo, como cara ou coroa

Eventos não excludentes:

Quando podem ocorrer ao mesmo tempo, quando joga um dado e ele ser 2 e par.

Eventos dependentes:

A ocorrência de um evento afeta o outro. Um tem que ocorrer para depois que o outro ocorra.

Eventos independentes:

A ocorrência de um evento não afeta o outro.

p = Ocorrência esperada/Numero de eventos possíveis

ex: jogar uma moeda: p=1/2, p=0,5

ex: Jogar um dado e dar 6: p=1/6, 0,16

ex: Jogar um dado e dar 1 ou 6: p=2/6, 0,33

ex: Jogar um dado e dar 1,2,3,4,5 ou 6: p=6/6, 1

ex: Jogar um dado e dar impar ou maior que 4: 1,3,5 e 6: p=4/6, 0,66

Eventos excludentes:

Soma-se as probabilidades

ex: jogar um dado e ser 1 ou par: 1/6 + 3/6 = 4/6 =0,66

Eventos não-excludentes:

Soma-se as probabilidades, diminui-se as sobreposições

ex: Jogar um dado e ser 2 ou par: 1/6+3/6 - 1/6 = 3/6 = 0,5

Você faz a soma como no excludentes e depois ve se há valores "repetidos" na soma. No exemplo temos o 2 que cai em ambas as condições ele é 2 e é par, então após somar as condições você subtrai a quantidade de valores repetidos das condições e tem o valor final.

Eventos independentes:

Usa-se a multiplicação.

ex: Qual a probabilidade de jogar dois dados, e dar 1 e 6: (Dois evento independentes) usando dados duas vezes

1/6\*1/6 = 1/36 = 0,027

Eventos dependentes:

Com 6 cartas na mão(A,2,3,4,5,6) qual a probabilidade de primeiro evento tirar A e no segundo evento tirar 4?

(No segundo evento temos 1 carta a menos no baralho)

ex:1/6 \* 1/5 = 1/30 = 0,033

Aula .103

Pequenas amostras:

Probabilidade de conclusões falsas

Não reprodutível

Não comparável

Aula .104

Passeio aleatório:

Sucessão de etapas aleatórias e independentes (uma não influência a outra)

Usadas para estudar vários fenômenos naturais e empresariais

Em finanças é usado para estudar os preços de ações

Não há como prever apenas há duas correntes, mas ambas são apenas teorias de qual melhor.

Aula .106

Conceito do que são os termos Estocástico x Determinístico

Estocástico: Dada uma mesma entrada, a saída pode variar.

Determinístico: Dada uma mesma entrada, apresenta sempre a mesma saída.

Aula .109

Distribuição normal ou gaussiana.

A distribuição normal: É o mais comum entre eles e a maioria dos dados estão distantes em no máximo três desvios padrão da média.

É utilizado para cálculo de probabilidade

Aula .128 estatística 02

Intervalos de confiança

É utilizado para quando de uma mesma população extraímos mais de uma amostra e isso faz com que tenhamos uma relativa diferença nos resultados entre elas e o intervalo de confiança é a medição da variância destas amostras.

É possível inferir características de uma população a partir de uma amostra, porque?

- custo

- viabilidade

etc ...

Preço: erro padrão/nível de confiança

riscos: dados ruins, enviesamento

Margem de erro: variabilidade, para mais ou para menos

Nível de confiança: de 80 a 99%

tamanho da amostra (n)

Queremos estimar o salário médio de um cientista de dados:

- 100 pesquisados

- Intervalo de confiança: 95%

- O desvio padrão é 1100,00

- A média é de 5.800

- Valor de z\* = 1,96

- Margem de erro: +- 215,60

O salário médio de um cientista de dados é entre 5.584 e 6.015 com nível de confiança de 95%

Teste de hipótese

Confirmar ou negar uma premissa usando uma amostra

Esta premissa usa um parâmetro, por exemplo: 50% dos brasileiros não gostam de estatística.

Encontrar diferença não é tudo, é preciso saber se esta diferença é estatisticamente significante