4.1 - Estatística básica

00:00:06:14 - 00:00:32:26
Olá pessoal! Bem-vindes a mais uma aula do curso de Análise de Dados. Nas últimas aulas, nós fizemos algumas análises com a nossa tabela usando a Biblioteca Pandas do Python. Mas não sei, eu sinto que está faltando alguma coisa, sabe? Está faltando números. Um analista de dados precisa entender sobre estatística e saber fazer a análises estatísticas. Nessa aula, nós vamos focar em aprender alguns conceitos de estatística básica e vamos fazer algumas análises numéricas com a nossa tabela.

00:00:33:10 - 00:00:59:20
Bom, bora lá. Para começar, a gente tem que fazer a importação de mais uma bibliotea. Na verdade, para começar, vamos começar organizando nosso notebook, como a gente tem feito. Vamos colocar aqui mais uma célula de texto. E essa aula é de estatística, então a gente vai colocar aqui o título de “estatística básica” para ficar legal aí. Ok, beleza. Agora sim.

00:01:00:02 - 00:01:24:02
Para começar, a gente vai fazer a importação de mais uma biblioteca. A gente importou a biblioteca pandas, que já utilizamos bastante com algumas análises. Porém, quando a gente fala em análises estatísticas, nós utilizamos muito a biblioteca “numpy”. Essa biblioteca nos permite fazer muitas operações matemáticas e estatísticas com números, listas etc, de uma forma bem rápida e fácil.

00:01:24:02 - 00:01:43:02
Para fazer importação dessa biblioteca, a gente vem aqui na nossa célula de código e faz o import numpy. Bem simples: import numpy. “Numpy com final Y. A gente executa. E a gente importou o numpy, só que toda vez que a gente precisar usar o numpy, a gente vai ter que escrever “numpy”. Então vamos dar um apelido, igual a gente fez a Pandas?

00:01:43:14 - 00:02:08:18
Vamos colocar: “import numpy as np”. Aí, toda vez que a gente for usar o numpy, a gente coloca só “np”. Vamos executar de novo. Agora sim. Vamos começar a entender um pouquinho sobre estatística. E aí eu pensei que a gente poderia, por exemplo, ter uma lista de idades. Eu tenho aqui nesse bloco de notas a lista que eu coloquei, mas para vocês colocarem igual, vocês podem dar uma pausa no vídeo e digitar aqui.

00:02:08:18 - 00:02:33:25
É só abrir uma nova linha de código e digitar: lista idades igual a abre colchetes, os números e fecha colchetes. Beleza? Vocês viram que eu tirei aqui um espaço? Tinha um espaço aqui, eu tirei, mas isso não importa não, tá, gente?! Pode ter espaço aqui, pode ter espaço aqui, é só para poder ter um padrãozinho mesmo, todo mundo próximo da vírgula.

00:02:34:18 - 00:03:00:01
Ok, eu tenho aqui essa lista de idades. Para definir que é uma lista, a gente precisa colocar os valores aí dentro desses colchetes e a gente precisa saber a média das idades dessas pessoas. Mas o que é a média? Média tenta encontrar o pontinho central, o meio mesmo. Se fossem só dois números, por exemplo, o 4 e o 6, o meio entre esses números seria o 5.

00:03:00:21 - 00:03:28:02
Para descobrir isso, somamos todos os números e dividimos pela quantidade de números somados. Vamos fazer isso aqui no nosso notebook. Vamos executar aqui para poder salvar na memória a lista de idades. Aqui vamos abrir mais uma célula de código. E pra fazer a soma dos valores, a gente pode usar uma função do numpy que se chama “sum”, S-U-M, de soma.

00:03:28:02 - 00:03:48:20
Então a gente pode chamar a biblioteca, que é o np, que a gente deu o apelidinho, ponto - porque a gente está chamando uma função dela - e o “sum”. E dentro do sum a gente coloca a nossa lista, que é a lista idades. Vou selecionar aqui, dar um ctrl+c e um ctrl+v, para não precisar digitar de novo. E aí a gente pode executar essa função. E a gente tem que a soma dos valores das idades é 328.

00:03:48:20 - 00:04:10:19
Executando essa função, a gente tem essa soma de todas as idades. Para saber a quantidade de idades, a gente usa o “len”. Lembra que a gente usou essa função de len para poder saber a quantidade de linhas da tabela, numa aula que a gente aprendeu sobre Pandas? Então a gente pode fazer assim: numa outra célula, a gente coloca “len”, lista de idades, ctrl+v aqui.

00:04:11:08 - 00:04:47:06
Executando, a gente tem 11 idades, Então: soma - a gente está somando todas as idades, o len - a gente está pegando a quantidade de dados. Ok. E para saber a média, a gente faz a soma dividido pela quantidade, beleza? E como que a gente faz isso? Numa próxima célula de texto, a gente copia aqui a nossa soma, coloca lá e aí a gente vai dividir pela quantidade.

00:04:47:18 - 00:05:14:26
A divisão é feita por essa barrinha aqui, simbolizada por essa barrinha. Então a gente pega aqui a soma da quantidade e coloca aqui. O que a gente está fazendo? Dividindo a soma das idades pela quantidade das idades. Quando a gente executa, a gente tem aqui que o resultado é 29,8. Isso significa que a média das idades dessa lista é de 29,8 anos.

00:05:15:08 - 00:05:40:22
Temos um outro jeito de fazer isso, que é utilizando funções já prontas, porque aí a gente não precisa colocar o np. No caso, a gente tem uma função numpy, que é o “mean” que a média em inglês. Olha como é mais fácil... Numa próxima célula de texto, a gente pode colocar o np ponto mean, de média, e colocar aqui o lista idades.

00:05:41:07 - 00:06:02:11
Executando, a gente tem o mesmo valor. Para poder ficar mais bonitinho, vamos colocar esse valor em uma variável. Aí, toda vez que a gente chamar a variável, a gente já tem esse valor. Vamos chamar de média. Média é igual a... sumiu o valor. A gente já sabe que é porque a gente atribuiu valor numa variável. Então se a gente quer saber o valor dessa variável, a gente chama a média.

00:06:02:26 - 00:06:27:26
E aí sim apareceu. Bom, vamos fazer de um jeito mais bonitinho? Vamos usar o print? Porque a gente está usando aqui várias coisas que a gente já aprendeu e está implementando. Então vamos colocar "print", tudo na mesma célula mesmo, só pra gente poder ver formas diferentes de colocar. Mas a gente poderia usar células diferentes para isso. Vamos colocar print.

00:06:28:24 - 00:06:53:11
E aí eu quero imprimir a média. Então vou colocar um texto de: média aritmética. Assim... que colocar dois pontos. E aí qual média eu quero mostrar depois? Essa média que a gente está calculando aqui em cima. Pra isso, eu coloco vírgula, fechei as aspas aqui da parte de texto e agora eu coloco vírgula e coloco a média.

00:06:54:01 - 00:07:17:10
E aí, nesse formato, o que está acontecendo? Eu vou printar a média aritmética, dois pontos, o valor da média, que é o nosso número ali, que é o 29,8. Vamos executar e apareceu... média aritmética: 29,8. Beleza, aí a gente não precisa usar mais o sum e o len, é mais importante a gente saber o que está acontecendo ali por trás.

00:07:17:15 - 00:07:45:22
Então assim, a gente está usando a função, mas a gente sabe que essa função está fazendo o quê? Está pegando a soma e dividindo pela quantidade. Existe também a média ponderada, que é quando damos pesos diferentes para valores. E no nosso exemplo de 4 e 6, que eu tinha falado. Entre 4 e 6, a média é 5. E se o 4 valesse duas vezes mais que o 2, a nossa média seria 4,6 e não 5.

00:07:46:08 - 00:08:09:08
Mas vamos deixar um material complementar explicando melhor esses conceitos. Não precisa se preocupar. Outra coisa importante para a gente aprender é o conceito de mediana. A mediana é o valor central, literalmente o valor central, que está no meio. Na média, a gente procura um ponto central ali dos valores em questão de concentração dos valores. Na mediana, a gente está procurando o meio mesmo.

00:08:10:22 - 00:08:32:29
Esses valores de mediana e média quase nunca vão ser o mesmo valor, quase nunca vão ser iguais. Na mediana, a gente vai colocar os nossos valores em ordem crescente e vai procurar o centro dessa lista. É bem tranquilo. Ela é importante pra quando a gente tem "outliers", que a gente vai ter uma aula só sobre isso, que são valores discrepantes, muito distantes dos valores da nossa lista.

00:08:33:06 - 00:09:05:21
Exemplo, se na nossa lista de idades aqui, a gente adicionar alguém com 100 anos, a nossa média sobe de 29,5 para 35,5. Inclusive façam aí: peguem aí, coloca 100 e calcula aí de novo para vocês verem. Mesmo que a maioria das pessoas esteja bem abaixo desse valor: então a gente tem idades de 26, 30 anos, mas se a gente coloca uma pessoa com 100, a nossa média sobe bastante. Já a mediana, como sempre ordenamos e olhamos o que fica no meio, não é afetada por esse outlier e, no nosso caso, ficaria ali por volta de 30.

00:09:05:22 - 00:09:33:11
Calma que a gente vai confirmar esse valor agora. Primeiro a gente precisa ordenar a nossa lista. Aqui na nossa lista estão os valores de 26, depois vai para 22. Vamos ordenar para ficar mais fácil de a gente visualizar. E a gente tem uma função que chama "sort". Então a gente vai abrir aqui, vai adicionar mais uma célula de código, colocar a nossa lista idades, ponto (porque é uma função) sort, abre e fecha parênteses.

00:09:34:17 - 00:10:02:02
E aí a gente executa. Deu um erro. Lista idade não está definida. Por que não está definida? De novo esqueci mais um "s". Na última aula eu também tinha esquecido um "s", né gente? Vamos ficar atentos aí. Lista idades... executa. E aí, beleza, a gente está já ordenou, para a gente poder visualizar vamos colocar aqui embaixo, pode ser na mesma célula, só pra gente poder visualizar a nossa lista: aqui já está ordenada.

00:10:02:15 - 00:10:23:21
Começa no 20, 22, 23... Beleza? Se a gente contar os valores dessa lista, vai ter 11 valores e a mediana a gente já combinou que é o número central. Então, se eu contar são dez. Então vão ficar cinco números de um lado, cinco do outro. E o meio é a nossa mediana. 1, 2, 3, 4, 5...

00:10:24:03 - 00:10:54:15
Aqui a gente tem cinco de um lado e cinco do outro, né? E depois aqui o sexto número, que é a nossa mediana. Então o 30 é a nossa mediana, ele está exatamente no meio. Mas se a gente editar a nossa lista idades e adicionar lá a nossa idade 100 no final... Vamos pegar aqui, copiar a lista de idades lá em cima e adicionar o 100, vou colar aqui... Adicionar a idade de 100...

00:10:55:17 - 00:11:15:19
Aí a gente vai fazer o quê? O sort de novo, porque as idades estão fora da ordem. Aqui... beleza, agora está na ordem. Mas agora a gente não tem 11 valores, a gente tem 12. Cadê o valor central? Como a gente tem 12 valores, vai ficar seis de um lado, seis no outro é o que está no meio? Não vai ter valor no meio.

00:11:16:01 - 00:11:47:17
Mas aí o que a gente vai fazer? Vai pegar os dois valores que estão no centro, ou seja, a gente vai deixar cinco de um lado, cinco do outro. Pegar esses dois valores do meio e tirar a média. A média a gente já aprendeu, que é a soma dos valores dividido pela quantidade. Aqui, a gente consegue fazer rapidinho por cabeça, que a soma é 61 dividido por 2, que são a quantidade de valores, que dá 30,5, Ou seja, adicionando a idade de 100 anos aqui no final da lista, a gente tem que a mediana agora é 30,5.

00:11:48:22 - 00:12:30:27
Igualmente à média, nós temos uma função para mediana, que é "median", de mediana em inglês. Então a gente pode colocar assim... já vamos atribuir o valor: mediana igual a np, ponto median, lista idades. Executamos. Vamos colocar aqui já a mediana embaixo para aparecer o valor da mediana. 30,5, beleza? Bom, como vocês viram, dá para utilizar as funções em listas simples usando o "numpy", mas vamos voltar lá para a nossa tabela e aplicar lá, pra gente poder ver e comparar alguns valores.

00:12:32:02 - 00:13:05:10
Bom, a gente tem algumas colunas com números, por exemplo, a coluna de idade. Vamos aplicar nela, na coluna de idade, e ver a média, a média de idade de quem respondeu o nosso formulário, pode ser? Só pra poder organizar melhor o nosso notebook, vou colocar uma célula de texto, subir ela aqui e vou colocar aqui vários jogos da velha, porque eu não quero texto grandão, quero o texto um pouco menor. Eu vou colocar aqui: voltando para a tabela.

00:13:06:10 - 00:13:32:20
Só pra eu poder saber que nesse ponto eu estou voltando pra minha tabela, ok? E a gente executa aqui só pra poder ficar organizado. Inclusive gente, eu não estou colocando comentários, mas é importante colocar. Então vocês podem voltar e colocar aqui, na primeira vez que usou o sort, coloca em uma linha acima um comentário de: colocando as idades na ordem.

00:13:34:04 - 00:14:05:10
Aí vai ficar um negócio bacana e se outra pessoa pegar o seu código, vai saber que nessa célula o que você está fazendo? Colocando as idades na ordem, ok? Bom, agora indo para nossa tabela mesmo, vamos calcular o quê? A média da coluna de idade. Então vamos lá: tabela dados, coluna idade, qual é a função? O mean. Então vamos executar. Carol, mas aqui você não usou numpy, né?

00:14:06:00 - 00:14:33:15
A gente não precisa usar o numpy pra média porque o próprio Pandas já calcula internamente isso pra gente. A média, no caso da nossa coluna de idade na tabela de dados, é 31,1. Aí você me pergunta: nossa, mas você mostrou com o numpy, usou aqui com o Pandas, por quê? Porque eu queria mostrar pra vocês que as bibliotecas, algumas bibliotecas, já têm algumas funções internas, tipo o Pandas calcula a média.

00:14:33:15 - 00:15:07:28
Mas outras funções estatísticas o Pandas não vai calcular e por isso a gente tem o numpy. Assim como a gente tem outras bibliotecas de estatísticas mais avançadas, que também calculam a média. Então a mesma biblioteca pode ter várias funções diferentes. Ok? Bom, vamos calcular agora a mediana da coluna de idade. Então vamos lá: dados, coluna idade, ponto median, que também é do Pandas, a nossa mediana. Vamos ver executando.

00:15:08:17 - 00:15:34:19
Opa, deu errado. Deu errado porque não coloquei a coluna com o mesmo nome. Vamos ficar atentos, retóricos a isso. A nossa mediana é 30. Então deu para ver, né? Por mais próximo que tenha sido a mediana da média, não são valores iguais. Outro conceito importante da estatística é a moda. Se a gente levar para o nosso dia a dia, a moda é o que está na moda, é o que está todo mundo usando.

00:15:35:03 - 00:16:01:06
E quando a gente fala de estatística, é mais ou menos a mesma coisa. Vai ser aquele número que aparece mais vezes. Vamos ver direto no nosso dado qual dado, qual a idade que aparece mais. Vamos lá. Para isso, vamos usar a função “mode” do Pandas, que é assim: mais uma célula de código, dados idade (sem o S dessa vez) ponto mode.

00:16:02:15 - 00:16:40:06
Ok, vamos executar pra gente poder saber qual é a idade que mais aparece. 27. Legal, então a gente sabe que deve ter muitas pessoas com idade de 27 anos. E um último conceito muito legal de estatística que vamos ver nessa aula é o desvio padrão. O desvio padrão vai dizer o quão distante ou o quão disperso os nossos dados estão em relação à média, ou seja, ele meio que indica se os dados são homogêneos, todo mundo concentradinho - por exemplo, a mesma faixa de idade - ou se estão variando muito - por exemplo, a gente tem jovens e idosos.

00:16:40:18 - 00:17:12:13
Quanto mais próximo de zero o valor de desvio padrão for, mais uniforme, mais homogêneos os nossos dados estão, eles estão mais próximos da média. Quanto mais longe de zero, mais dispersos nossos dados estão, mais espalhados e mais longe da média. Vamos calcular o nosso desvio padrão da nossa coluna de idade. Mais uma célula de código, dados, idade que é a nossa coluna, e a função do desvio padrão é “std”.

00:17:12:28 - 00:17:33:20
Ok? E os nossos parênteses vazios porque a gente não está colocando nenhum parâmetro. Vamos executar? Então nosso desvio padrão é de 6,9, vamos arredondar pra 7. Pensando que estamos falando de idade, em que os nossos valores estão em torno de 30, 50, um valor de 7 significa que a gente tem uma variação alta. Não somente isso.

00:17:33:20 - 00:18:05:13
Vamos levar em contexto também a quantidade de dados que a gente tem. A gente tem uma tabela de 4200 e poucas linhas, e isso, na área de análise de dados, são poucos dados. Então, um desvio padrão de 7 é um valor alto. Mas se tivéssemos uma planilha com 1 milhão de linhas e uma coluna de idades nessa planilha, um valor de desvio padrão de 7 seria baixo porque a gente tem ali várias entradas.

00:18:05:13 - 00:18:27:12
Então a gente tem que estar sempre atento a esse contexto, sempre lembrar sobre aquela parte inicial dos dados onde a gente olha o tamanho da tabela, as colunas e tudo mais. Bom, pra encerrar, vamos ver aqui algumas outras funções, como o "max", que vai retornar o nosso valor máximo e o "min" que vai retornar nosso valor mínimo.

00:18:27:24 - 00:19:04:07
E a gente faz isso assim: nessa célula de código, vamos lá na nossa coluna de idade e vamos colocar o ponto min e os parênteses vazios. Executando. A gente tem aqui o 18, que é a idade mínima de quem preencheu esse formulário, quem estava lá fazendo a pesquisa. Bom, uma coisa que eu vi que eu fiz aqui e eu acho muito legal de explicar é que eu executei e assim que executou já surgiu uma célula embaixo. Vou apagar essa célula aqui só para poder mostrar para vocês como que faz: eu aperto o shift+enter...

00:19:04:23 - 00:19:25:12
Aí, quando eu aperto essas teclas juntas, vai executar a célula que eu estou fazendo e já vai surgir uma embaixo. Isso é muito bom porque já ajuda: a gente não precisa ficar subindo aqui e ficar clicando em código porque tem outro embaixo. Bom, agora calcular o valor máximo. Estou com preguiça. Vou copiar aqui o que está em cima, ctrl+c. Clico na célula de baixo: ctrl+v.

00:19:25:16 - 00:19:49:04
E substituo o "min" pelo "max", M-A-X. Executo de novo... Já vou executar já deixando uma célula embaixo. E aí eu tenho que a idade máxima de quem respondeu o formulário é 54. A gente pode combinar tudo o que a gente viu até aqui, por exemplo, se a gente quer saber a média de idade só das pessoas do gênero feminino.

00:19:49:17 - 00:20:17:27
A gente vai fazer um filtro e a gente vai calcular a média. Bom, vamos fazer essas duas coisas. Primeiro, vamos fazer lá o filtro que a gente já aprendeu, né? Tabela dados, abre colchetes, dados... A gente quer filtrar tudo o que é feminino. Então qual é a coluna? De gênero. A gente não vai fazer atribuição, a gente está fazendo uma comparação: dois sinais de igual. Aí depois do igual, aspas simples ou dupla e feminino.

00:20:18:05 - 00:20:43:19
Feminino igualzinho está lá na coluna de gênero. Ok, a gente tem o filtro do feminino, mas a gente quer a média de idade de quem respondeu que é do gênero feminino. Então a gente quer o quê? A gente quer a coluna de idade. Então depois o filtro aqui, a gente coloca o colchete, a coluna que a gente está querendo, que é de idade, e o ponto min, que é a função que calcula pra gente a média, ok? E aqui a gente executa.

00:20:44:06 - 00:21:07:22
E a média de idade de quem respondeu que é do gênero feminino é 31,31. Não é tão diferente assim, não é tão distante da média das idades em geral, né? Será que a média de idade dos homens, de quem respondeu que é do gênero masculino, é muito diferente? Fica aí o exercício para vocês fazerem. Nós temos outra coluna com valores numéricos, que é o salário.

00:21:08:16 - 00:21:38:09
Vale avisar que essa coluna de salário, nós que adicionamos na planilha pra fazer esse curso exatamente, e não faz parte da base de dados original. Vou deixar para vocês calcularem e responderem qual o valor máximo de salário, qual o valor mínimo, qual a média e qual o desvio padrão. E uma pergunta que a gente pode responder com essa análise da coluna de salário é: a média de salário das pessoas do gênero feminino é maior ou menor que a média salário das pessoas do gênero masculino?

00:21:39:09 - 00:22:00:02
E aí a gente pode fazer o quê? O filtro por gênero e depois calcular as médias e comparar. Vou fazer aqui... Vou criar mais uma célula. E aí é aquela coisa: estou com preguiça, posso copiar aqui o que está em cima, na célula de cima, dou ctrl+c, clico na célula de baixo, que é de código, ctrl+v.

00:22:00:18 - 00:22:26:18
Só que agora eu não quero a coluna de idade, eu quero a coluna de salário. Esqueceu como que está a coluna de salário? Você cria aí uma célula: dados ponto columns, vê lá como está escrito, depois apaga a célula e está tranquilo. Salário... E aí aqui eu estou fazendo o filtro do feminino, colocando o salário que eu quero saber a média. Vou executar... Já vou executar já deixando uma célula embaixo.

00:22:27:06 - 00:23:01:09
E aí eu sei que a média do salário das pessoas que responderam que são do gênero feminino é 8675. Vamos fazer do masculino e dar uma comparada. Vou copiar a de cima, vou colar embaixo e agora eu quero o quê? Eu quero tudo que seja igual a masculino na coluna de gênero. Então vou substituir: masculino, vou executar... Ok... 11.727. É uma diferença de 3.000 reais?!

00:23:02:02 - 00:23:27:12
Com isso, podemos dizer que as pessoas do gênero masculino recebem mais que as pessoas do gênero feminino na área de TI. Uma conclusão triste. Para ter certeza disso, precisamos também olhar para outras variáveis, como o cargo, o Estado, entre outras coisas, que podem também influenciar nesse valor e a gente conseguir ter uma análise mais robusta. Mas a gente vai ver isso mais profundo em outras aulas. E é isso por hoje, gente, até a próxima.