06

Correção do projeto

Transcrição

[0:00] Ok, pessoal. Agora que vocês já fizeram todo o projeto, eu sei que você se esforçou bastante, vão comparar o que você fez com o que eu fiz. Caso você tenha ficado com alguma dúvida, ou então não tenha conseguido fazer algum tópico.

[0:13] A gente vai ver aqui agora, eu deixei esse notebook com a resposta, está aqui: Análise Descritiva Respostas. Está lá, até vou ver o download então. Deixei os dois, o vazio e o com as respostas, caso você queira fazer novamente.

[0:26] Então, vamos lá. Aqui, vamos baixar até chegar ao ponto. Aqui, no import das bibliotecas eu puxei o Pandas, o NumPy e o Seaborn. O Seaborn a gente puxa lá no desafio, que eu deixei para vocês.

[0:40] Aqui, por padrão, é tudo o que a gente fez no nosso curso. Dados, pd com o read_csv dentro do Pandas. Mostramos ele aqui, só para a gente ter uma noção.

[0:49] Aqui, a gente já começa criando a distribuição de frequências, onde a gente cria as classes para a variável renda. Seria aquele A, B, C, D e E, só que eu mudei a classe para você tentar fazer sozinho aqui.

[1:02] Então, eu criei aqui dentro dessa fórmula, eu criei uma lista e coloquei os valores dentro dela. Sendo que o primeiro, é o valor mínimo.

[1:11] Aqui, a gente já começa com duas vezes o salário mínimo, esse sete oito oito, eu deixei aqui já como dica para você. Porque aí a gente tem até dois salários mínimos, então duas vezes o sete oito oito.

[1:19] Aqui, o próximo valor seria esse aqui, aí vai de dois à cinco. Então, você tem que entregar para ele o cinco aqui, cinco vezes o sete oito oito. E assim sucessivamente, o 15, depois o 25 e depois o máximo.

[1:32] É o que a gente fez aqui, importei, criei uma classe que vai de zero até dois salários mínimos. Agora, de dois salários mínimos, aí pega esse aqui novamente até cinco. De cinco salários mínimos até 15 salários mínimos. E de 15 até 25, e de 25 até o máximo. Então, até aqui.

[1:50] Aí, a gente coloca isso aqui. Criei os labels também E, D, C, B, A. Por que está ao contrário? Porque o E é a classe de renda mais baixa. Está nessa ordem aqui da menor para a maior, então aqui também tem que ser da menor para a maior, para poder entender.

[2:04] Passamos por ele aqui. Esse aqui é padrão, a gente fez nas nossas aulas. Value counts, aqui a gente usou o cut para criar as classes. Aqui também, a mesma coisa, só que a gente fez para pegar em valores percentuais e não absolutos.

[2:21] A gente usa o normalize é igual a true, multiplicamos por 100, para poder escrever assim: porcentagem abre e fecha parênteses, tem um percentual aqui e aí a gente não precisa formatar esse número. Fica mais simples.

[2:33] Está aqui, a tabela já formatada e já organizada por A, B, C, D, E. Dessa forma aqui, sort_index. Tudo isso a gente fez no nosso curso.

[2:43] Aqui é uma representação gráfica dessa tabela, não é um programa, porque o programa a gente faz com variáveis quantitativas. E como a nossa tabela, é assim, de uma quantitativa a gente criou uma qualitativa. E atribuímos essas classes A, B, C, D e E. Aí, a gente usa um gráfico de barras, fica semelhante ao histograma. Ou outra solução para esse caso, poderia ser um gráfico de pizza, aquela bolinha com as fatias, também dá para representar numa boa.

[3:10] Então, vamos lá. Seguindo em frente, aqui a gente tem o histograma para as variáveis quantitativas do nosso dataset. Aí, você tinha que saber quais são, eu falei quais eram.

[3:18] Eu peguei aqui a idade, a altura e a renda. O histograma, a gente faz com o distplot no Seaborn, que a gente fez no nosso curso. Não tem segredo, é só você colocar lá para colar, se tiver alguma dúvida.

[3:35] As conclusões, eu deixei aqui para você, tirar suas conclusões. Mas, é basicamente aquilo que a gente vem falando na aula. Aqui é uma assimetria, aqui é uma distribuição perfeitamente simétrica, e por aí vai.

[3:50] Aqui, você pode falar dos tipos de barras, tentar entender o porquê dessa coisa. Porque tem uns vales aqui, uns picos, que provavelmente são as idades cheias, como 40. Aí, depois a pessoa diz "Eu tenho 40", mas na verdade ela tem 39. Pode ter esse tipo de problema, isso a gente tem que ir analisando no nosso dataset.

[4:09] Enfim, passando aqui, eu pedi para você construir o mesmo histograma. Só que esse aqui está muito difícil de visualizar. Para facilitar a nossa visualização e o comportamento das medidas, eu pedi para você construir para pessoas com renda até R$20.000,00, porque você corta essa parte toda aqui, que têm uma renda absurda. E aí, a gente consegue ver bem aquela coisa da assimetria aqui para esse lado, que seria a direita

[4:39] Então, vamos lá. Aqui, eu pedi para construir tabelas de frequências, cruzando as variáveis sexo e cor. E aqui, o que eu tinha dado para vocês, são os dicionários que passam a descrição dos valores das variáveis, que é justamente o que a gente viu aqui no começo. Aqui está a variável F, por exemplo, o 41 é o Paraná, o 33 é o Rio de Janeiro, o 35 é o São Paulo e por aí vai. O sexo, zero é o masculino e o um o feminino. E foi isso que eu coloquei lá naquela variável, aqueles dicionários que estão aqui, para você poder utilizar se quiser, para gerar os seus resultados.

[5:18] Então, vamos lá. Esse aqui é só rodar, aqui é o exemplo de utilização dele. Aqui, eu peço para fazer uma tabela com a frequência cruzada entre duas variáveis, sexo e cor, ok? Aqui, se você não tivesse usado essas duas informações, você teria colunas com valores numéricos. Aqui seria zero um, aqui dois, quatro, seis, se eu não me engano, era isso.

[5:41] Aqui, é a forma de criar essa tabela, com o crosstab, como vimos no nosso vídeo. Aqui é a mesma coisa, percentual. E aí, a gente consegue fazer algumas análises. Por exemplo, a cor branca aqui tem 28%, o sexo masculino 12%. No feminino, aqui a cor parda é mais até do que a cor branca. Aí, a gente pode tentar separar essas duas cores, fazer algumas análises com elas. De renda, de educação,e por aí vai.

[6:10] Aqui, eu peço para fazer uma análise descritiva especificamente com a renda. Aí, a gente faz o cruzamento das outras variáveis com a renda. Aqui, pura e simplesmente a renda.

[6:23] A gente tem a média, a mediana, coisas que aprendemos ao longo do nosso curso. A moda, a gente conheceu a tendência central. Aqui, a gente já vem com as dispersões. Aqui, o desvio médio absoluto. Aqui é a variância e aqui o desvio padrão. Tudo isso foi calculado lá também.

[6:42] Aqui, eu peço para você obter a média, a mediana e o valor máximo que é o top lá. Para a variável renda novamente, a gente continua com a análise da renda segundo sexo e cor.

[6:56] E aí, a gente utiliza o crosstab, aqui eu deixei uma dica para você. E aqui eu mostro como fazer isso. Aqui você vê o value, que é a variável que a gente vai fazer a conta da média, da mediana e do valor, cruzando sexo e cor. Então, eu vou usar a renda para obter a média.

[7:13] E aqui eu passo nesse aggfunc, dentro desse, não é um dicionário, mas entre chaves nessa versão do Pandas. Algumas versões aceitam, a mais nova se eu não me engano, colocar isso aqui como se fosse uma lista. Ou seja, entre colchetes, aí você testa e vê se dá certo.

[7:34] E aqui você passa as funções que você quer. Eu quero a média, que é o mean, a mediana vai ser a median e o máximo é o max. Aí vai criar para você uma tabela aqui, sexo e as cores. Aqui é o máximo, aqui é a média e aqui a mediana, ok?

[7:55] Aqui, eu peço para fazer basicamente a mesma coisa, só que com medidas de dispersão, eu peço elas aqui. O mad, que é o desvio médio absoluto, a variância e o desvio padrão. É o mesmo destino, só que com estatísticas de dispersão. Aqui também, a tabela pronta já. Lembrando, que isso aqui em baixo eu coloquei simplesmente para a gente nomear linhas e colunas. O index está nomeando linha e cor, e columns, colunas.

[8:32] Continuando, aqui a gente vem com a construção do boxplot da variável renda segundo sexo e cor. Duas variáveis, do jeito que a gente fez lá, as tabulações. Só que acho que a gente não viu isso na aula.

[8:44] Então, eu mostrei uma dica, para você incluir a terceira usando esse parâmetro aqui. Publiquei duas vezes, para colocar mais um na variável. O que ele faz? O boxplot está aqui, eu passo para ele o X que é a renda e o eixo Y, que é a cor. E eu quero que ele faça uma diferenciação por sexo, para cada cor. É o que ele faz aqui.

[9:18] Vou fechar aqui, porque não precisa disso agora. Ou seja, aqui a cor indígena eu tenho os homens e no verde as mulheres, com a legenda do lado. Isso é incrível, gente.

[9:29] Aqui, eu coloquei todas as configurações que eu acho necessárias para você entender, daqui para baixo é só configuração do gráfico. O da legenda é um pouco mais complicado, esse aqui por exemplo, o ax get legend handles labels. Ele vai te dar duas saídas, é como se ele tivesse uma tupla, e eu só quero o handles. O que ele passa para você? Ele passa o handles, que é esse quadradinho aqui, um tipo de quadrado. E o próximo é a legenda, ele passa esses dois aqui.

[10:05] Com underscores, significa que eu não me interesso por esse valor, ele vai deixar vazio. Então, só vou pegar o handles, coloco vírgula e underscore, e vamos soltar dois valores. Eu pego só o primeiro e coloco aqui na legenda.

[10:16] Isso é só técnica de construção, como o nosso curso não é de Seaborn, então eu deixei só para você saber se virar aqui. Depois você pode procurar no google, tem aqui em cima.

[10:27] Então como eu estava dizendo, eu também pedi para que a gente faça uma separação da renda só para renda até R$ 10.000, para poder visualizar melhor nosso gráfico. Porque se você colocar a renda toda, isso aqui vai ficar muito no cantinho e a gente não vai conseguir ver essas construções no meio, onde tem a maior densidade de informações.

[10:55] Aqui a gente já pode perceber uma diferença bastante interessante da renda, tanto por cor quanto por sexo. A gente vê que até dentro de um agrupamento de cor de pele, a branca por exemplo, tem uma diferença forte entre homens e mulheres. Isso, a gente já tinha visto lá somente com sexo e aqui a gente está fazendo uma dupla visualização. Aí, a gente pode ver uma diferença bastante interessante aqui. Por exemplo, a cor amarela, você vê que eles têm uma renda maior, a mediana deles é maior. Mas, existe uma diferença de rendimento também por sexo, na cor amarela.

[11:38] Seguindo, para o nosso desafio. O que eu pedi aqui? O percentual de pessoas, no caso aqui eu pedi um percentual exato. Quando a gente estava falando do percentil, eu falei se não me engano, que 28% abaixo eram as pessoas que ganhavam até um salário mínimo.

[11:57] O que eu estou querendo aqui é, "Qual o percentual de pessoas de nosso dataset ganham um salário mínimo ou menos?" Ou seja, tem a renda menor ou igual a um salário mínimo. Eu quero saber o percentual exato e isso aqui eu não consigo usar com o percentil.

[12:13] Eu teria que continuar dividindo em mil partes ou sabe lá quantas partes, até eu conseguir chegar a um valor mais exato. Aqui, eu não consigo e lá também eu ia ter que visualmente, ou então, criar uma função para poder pegar esse valor para mim. Isso é mais complicado, aqui ele já faz tudo isso para a gente.

[12:33] Fazendo o que? Eu passo para essa função, que eu deixei para vocês aqui, o percentileofscore. Eu passo para ele qual é o dado que eu quero, ou seja, a renda que é o que a gente está estudando. Passo o valor de corte, que é justamente o salário mínimo, o sete oito oito. E passo o tipo de corte que eu quero. Por que esse tipo?

[12:55] Vamos lá. Aqui ele está me dizendo, no final do resultado, temos 28,87% das pessoas da nossa amostra, da nossa população, que ganham até um salário mínimo. Que tem renda menor ou igual a um salário mínimo, quase 30%.

[13:13] Vamos entrar aqui rapidinho, só para mostrar para vocês essa coisa do kind rank, que é o padrão dessa função. O que ele faz? Aqui, ele tem uns exemplos.

[13:29] Por padrão, ele vai fazer o que? Aqui tem um exemplo, ele já passou um dado para ele, o um, dois, três, três e quatro. E ele passou para a gente aqui, o três, que é como se isso fosse a minha renda e aqui o salário mínimo, que eu estou falando. E aqui visualmente, a gente consegue ver legal.

[13:45] O três aqui, por exemplo, nesse tipo de visão do rank, ele vai identificar esse aqui e vai passar para a gente 70%. Porque ele pegou a média, aqui, eu pego 20, 40, 60% a distribuição está aqui, 80% está aqui, e aqui 100%. Correto? 20, 40, a gente vai de 20 em 20.

[14:06] Chegamos aqui, a gente tem 60 e aqui 80. O que ele faz? Ele tira a média desses dois, o rank. Ou seja, ele fica com 70%, esse é o rank. Quando ele usa o strict, ele vai pegar o que? Ele vai chegar e verificar que o três está aqui, então ele vai informar que 20, 40, chegou aqui no 60, ele já está aqui. Então, ele vai falar que 40% da distribuição está abaixo desse valor, e ele não inclui o três. É isso que o strict faz.

[14:43] O weak, que foi o que a gente usou já inclui esse valor, por isso que eu usei ele. E a renda menor ou igual à sete oito oito, ou seja, o salário mínimo. Então, vamos supor que o nosso salário mínimo seja três, aí vai ficar 20, 40, 60, 80. 80% ganha, no caso aqui, tem o valor menor ou igual à três.

[15:09] E o último é a média, desses dois aqui. Ele identifica qual é a média, que é como se estivesse calculando a mediana. Na verdade, não é nem a média, ele pega aqui o 20, o 40, o 60 e tem o 80. Ele parte isso no meio aqui, ou seja, vai cortar na metade. Por isso, que é média. A metade ele vai jogar para cá, ou seja, 60% abaixo desse valor que é a média, que cortou no meio.

[15:40] Ok? Então, esse era o desafio. Eu quis introduzir o SciPy, porque a gente não usou em outros cursos de estatística, e a gente vai começar a mexer com ele. É interessante você começar a pegar o traquejo de todos esses, porque cada um tem uma ferramenta que você pode utilizar, ou então é mais especializado em determinada coisa. Esse SciPy tem muito teste e a gente vai utilizar os testes dele, no próximo curso, teste de posse. Então, é isso.

[16:05] Aqui, o negócio do 99, é bem simples. Foi justamente aquilo que a gente viu no nosso curso. "Qual o valor máximo ganho por 99% das pessoas de nosso dataset?" Ou seja, a gente pega o quantile ponto 99, e vê qual a renda que corta isso em 99% abaixo e 1% acima. Então, é 15 mil. É bem simples, é o que a gente fez na aula.

[16:31] Aqui, "Obtenha a média, mediana, o valor máximo e o desvio padrão da renda segundo anos de estudo e sexo". É a mesma coisa que lá, só a gente utilizou anos de estudo aqui. Então, anos de estudo e sexo, aqui os nomes são diferentes. Aqui, eu agrupei mais, eu tenho o máximo, média, mediana e desvio padrão. Isso só para você treinar a sua capacidade analítica. O legal é que em cada tabela dessa, você observe o que está acontecendo ali, características interessantes e importantes. E aí, faça anotações lá em baixo.

[17:01] Aqui, a pessoa construiu um boxplot também. Esse é bem parecido com aquele da renda, que a gente viu agora há pouco. Só que aqui, eu fiz uma coisa um pouco diferente. Eu passei uma lista, no primeiro eu deixei, aqui para você visualizar que é esse mesmo, esse ticklabels. A gente passa uma lista, na ordem que aparece, dos itens que a gente quer nomear aqui. Por exemplo: indígena, branca, preta, amarela, parda.

[17:32] Eu, como tenho um dicionário lá, voltando aqui, esse aqui é a mesma funcionalidade. Como eu criei um dicionário lá, o que eu vou fazer? Deixa eu só mostrar para vocês o código acima, não sei se dá para ver direito aqui. Mas, a gente volta lá, é porque esse aqui é o que mandei para vocês e estou visualizando.

[17:59] Mas, rodando essa célula aqui a gente consegue ter um resultado lá em baixo, onde eu criei. Está aqui, muito bem. O que ele faz? Como a gente tem um dicionário, que tem chave e valor, vamos colocar ele aqui em cima, Code e anos de estudo. Aí, a gente percebe no cenário temos a chave, que eu escrevo o valor. E aqui o valor, chave valor, chave valor. É assim que funciona.

[18:36] Eu coloquei dentro de uma lista um for, a gente já viu na nossa aula isso, onde eu queria passar o valor. Então, eu coloquei o key aqui, mas está errado. Porque o key seria como se fosse isso aqui, mas não tem problema. Nesse nome aqui, você pode colocar o que quiser. Pode pôr I, I.

[18:55]Aqui que eu importo, no caso se eu quisesse a chave que é o um, dois, três, quatro, cinco, seis, sete, até o 17. Não é o key esse, desculpe. Agora sim, eu vou passar as chaves. Quando eu quis aqui os valores, eu criei uma lista. Com os valores eu fiz isso aqui, para já passar para ele esses valores. Aí ele passa para cá, para esse aqui, que amigavelmente completa aqui para a gente. Isso que foi feito. A única diferença do outro é essa aqui. O resto é tamanho do gráfico, título, label e por aí vai. Aqui, a legenda do masculino e o feminino.

[19:43] Conclusões, é para você concluir. Aqui, eu fiz uma coisa um pouco diferente, usando o groupby para a variável renda segundo unidades da federação. É a única variável que a gente ainda não tinha usado no curso, que são as UFS, os estados do Brasil. Então, eu não quis fazer cruzamentos para não ficar umas tabelas enormes. Mas, se você quiser, fique à vontade.

[20:05] Aqui, eu só fiz a média da renda por estado. A média, a mediana, o máximo e o desvio padrão, usando esse macete aqui do groupby. Basicamente, eu fiz aqui, para você entender como que funciona. Aqui, eu passo o groupby UF. E esse agg, é para fazer as agregações da variável renda, e eu quero essas estatísticas aqui. Eu tenho a média, a mediana, máximo e desvio padrão para a variável renda, agrupado por UF. É dessa forma que funciona. Aqui eu tenho o UF.

[20:42] E aqui para finalizar, se eu não me engano acho que é o último, é o nosso gráfico boxplot. Para a gente ter uma ideia também, porque esses de renda, todos que você deve ter reparado. Aqui o 10 mil, só para você visualizar. Se você rodar esse aqui sem o 10 mil, vai ficar um gráfico imprensado, por causa da assimetria, dessa variável.

[21:02] E aqui eu queria mostrar para vocês, para você visualizar e perceber os efeitos regionalmente. Porque o índice, o intervalo do primeiro quartil, porque aqui a diferença entre o primeiro e o terceiro quartil é uma estatística de variação, de dispersão. Então, você dá uma olhada, aqui o Distrito Federal, se há uma dispersão muito grande. Aí você vê os comportamentos aqui, em cada região. O tamanho de renda também. Aqui Rio de Janeiro, São Paulo maior um pouco, o de Brasília, Distrito Federal bem maior, e por aí vai.

[21:42] Não se prenda a isso aqui, a gente finalizou agora. Você pode fazer mais análises, testar mais cruzamentos, usar um outro dataset. E treinar bem essas conclusões, que isso é bem importante. Era isso que eu queria mostrar, no próximo vídeo a gente faz uma conclusão da aula, de todo o curso que a gente fez. Abraço.