4.7 - Resumo da ópera

00:00:01:02 - 00:00:20:19
Olá pessoal! Bem-vindes a mais uma aula do curso de Análise de Dados da PrograMaria. Nessa aula nós vamos recapitular tudo que fizemos desde o início da nossa análise de dados com Python utilizando Colab. Bom que aproveitamos para colocar alguns comentários e organizar melhor, né? Então vamos lá. Bom, temos aqui o nosso notebook análise dados, já colocamos o nome, ok.

00:00:21:02 - 00:00:43:16
Cada tópico que a gente foi avançando, nós criamos aqui uma espécie de índice. Então a gente tem o uso da biblioteca pandas e tal. E o legal é que podemos recolher as células dos índices e assim conseguir visualizar o nosso notebook de forma geral. A gente pode clicar aqui. Olha que bacana! Eu amo isso, gente, de conseguir recolher, porque às vezes você não quer ver.

00:00:43:24 - 00:01:19:27
Por exemplo, em valores faltantes a gente ficou brincando em 49 células, então assim fica mais fácil de a gente ver quais foram os itens que a gente passou. Ok. É muito importante que uma pessoa analista de dados organize seu código, pois nunca sabemos quem pode pegar o código depois ou até mesmo daqui alguns anos a gente vai ver esse notebook e precisamos entender o que foi feito aqui. Então nós temos a estrutura do nosso notebook assim, ok? Mas seria legal se a gente colocasse um índice inicial na introdução explicando o que vamos fazer nesse notebook, explicando de onde vieram os dados que vamos analisar.

00:01:20:11 - 00:01:54:08
Vamos criar, então, um índice chamado Introdução. Vou tentar colocar aqui uma célula de texto, mas vai ficar lá no final. Acaba sempre criando no final. A gente usa essas setas para cima para mover lá pro início. Então vamos mover lá pro início. Aqui... Agora sim. Beleza, ficou aqui antes do uso da biblioteca. E aí a gente clica aqui nela e vamos criar um índice chamado Introdução para ficar bonito.

00:01:55:09 - 00:02:14:20
Eu já deixei um textinho pronto para vocês, de uma introdução que eu vou colocar aqui, mas seria legal se cada pessoa criasse sua própria introdução do que entendeu dos dados. Enfim, fizesse sua própria busca e escrevesse com suas próprias palavras o que a gente vai fazer e tudo mais. Então vou copiar o meu texto e colocar aqui embaixo desse título de introdução.

00:02:15:02 - 00:02:40:17
E assim o nosso notebook já ganha um contexto legal e quem ler já vai entender qual foi o nosso objetivo aqui. Então vamos lá... Selecionei meu textinho que está nesse bloco de notas e vou colocar aqui embaixo. Logo depois dessa introdução aqui vem esse item “uso da biblioteca Pandas”. Nós temos aqui células de importações de dados.

00:02:40:18 - 00:02:59:17
Então aqui a gente já importou o drive e fica aí uma dica de ter uma célula única para os imports utilizados no notebook. Assim fica mais fácil de a gente administrar as bibliotecas que estamos usando e, além disso, não precisamos ficar importando duas ou três vezes a mesma biblioteca porque a gente importa aqui pandas aí lá embaixo a gente fazer import do pandas de novo.

00:03:00:00 - 00:03:29:27
Então vou colocar um comentário, uma célula aqui de código, colocando o comentário para colocar todas as importações. Aqui já tem essa aqui, então vou dar um enter. Pra comentários a gente usa esse símbolo de jogo da velha, então: importações dois pontos... Aqui embaixo já tem pandas, já vou tirar daqui e colocar aqui embaixo. E aí essa célula que ficou vazia, a gente apaga ela.

00:03:30:17 - 00:04:09:01
Ok? A gente fez outras importações ao longo do notebook. Eu vou deixar essa tarefa para vocês procurarem e colocarem todos os imports aqui nessa célula de importações. Se quiserem, a gente pode até criar um índice chamado Imports e fica bem embaixo da célula para poder ficar mais organizada. Aliás, antes das célula de uso da biblioteca pandas, a gente adiciona uma célula de texto e coloca o índice aqui: importações. Ok? E aí dentro dela, abaixo dela, a gente cria uma célula de código e copia todos os imports pra cá.

00:04:09:01 - 00:04:27:29
A gente apaga essa célula aqui pra não ficar vazia ali no limbo e pronto. Olha só como é que fica mais organizado. A gente tem um índice só para as nossas importações. Temos uma célula aqui de montar o drive e depois de ler os dados. É legal a gente ir colocando os comentários também sobre o que a gente está fazendo. Por exemplo, aqui a gente está montando o drive.

00:04:28:19 - 00:04:49:25
Então clicamos nessa célula, vamos dar enter, e aqui em cima a gente coloca o comentário montando o drive ou criando a conexão com drive. Logo abaixo a gente faz essa leitura dos dados. Então a gente coloca aqui mais um comentário: lendo os dados. Claro que não precisa ser um texto grande, só um pequeno comentário para saber o que está acontecendo.

00:04:50:10 - 00:05:25:03
Estou dando exemplos, mas vou deixar pra vocês comentarem aí no notebook da forma que ficar mais fácil para vocês, ok? Então tá. Aqui no módulo do Pandas, nós aprendemos a explorar os dados, lemos o arquivo, visualizamos a tabela, aprendemos a usar o head, o tail. Descendo aqui: o head, o tail... Aprendemos o shape para ver o tamanho da tabela aqui. Que mais? Lembra que quando vamos chamar uma função, a gente usa o ponto.

00:05:25:07 - 00:05:59:24
Então, por exemplo, ponto shape, ou ponto columns, ponto info (a gente aprendeu o ponto info). E aí a sintaxe fica sempre o nome da tabela ponto a função. E vai colocando esses comentários: olha, a função tem o ponto, com o info aqui, por exemplo, a gente tem informações sobre as colunas, os tipos de dados por coluna etc. Aqui embaixo a gente viu o describe. No describe a gente consegue ver algumas informações de colunas que contêm dados numéricos, informações como média, mínima e máxima, etc.

00:06:00:07 - 00:06:34:02
Vocês podem ir colocando essas coisas também. Bom, legal essa parte. No próximo tópico nós repetimos a análise que a gente viu no Excel aprendemos aqui... Vamos primeiro ver aqui, estender aqui o índice... Aprendemos como visualizar uma única coluna: aqui embaixo a gente visualizou só gênero. Aprendemos a usar os colchetes também; que para poder chamar uma coluna a gente usa as aspas; que o nome da coluna tem que ser exatamente igual nos dados da tabela.

00:06:34:17 - 00:06:58:25
Para visualizar todas as colunas, a gente usa esse "ponto columns" também. Quê mais? Aprendemos a fazer filtro. Aí pra isso precisamos primeiro identificar a tabela de dados. Depois disso, aprendemos a identificar qual é a coluna que queremos filtrar e aí a condição para fazer o filtro. Por isso que usamos sempre o nome da tabela, colchete, a coluna e dois símbolos de igual.

00:06:58:26 - 00:07:24:13
A gente não está fazendo a atribuição, a gente está comparando. Aqui, por exemplo a gente está fazendo o filtro de feminino, então a gente usa dados, nome da coluna, dois símbolos de igual e o que a gente está procurando, o que a gente está querendo filtrar. Bom, para usar números, filtrar com números, a gente usa aquele símbolos matemáticos de maior ou menor.

00:07:24:27 - 00:07:47:24
A gente filtrou todas as pessoas que a idade é maior que 30. Bom, aprendemos a agrupar os valores também, aqui mais para baixo a gente agrupou usando o group by. Lembre-se que aqui é a função, então por isso a gente usa o ponto: ponto group by. E aí algumas funções tem parâmetros que podemos ou não preencher.

00:07:48:10 - 00:08:20:03
No caso do group by informamos a coluna, em parênteses, uma função. Vimos bastante coisa nesse tópico. Depois a gente foi pra estatística básica, onde inicialmente tínhamos uma lista exemplo e aprendemos então sobre média, mediana, moda, que a gente tinha essa lista de idades. Aprendemos média. Quando a gente voltou para a nossa tabela, nós analisamos a coluna de idade e vimos que a média das idades...

00:08:21:01 - 00:08:48:11
Vamos ver aqui... A média das idades é 31.16, a gente viu que a mediana é 30, que a moda é 27. A gente calculou aqui o desvio médio padrão, o mínimo, o máximo, calculamos a média com o filtro também. Vocês podem colocar comentários nas células indicando o que estamos calculando e até mesmo colocando os conceitos de cada item, como por exemplo, o que a média que é mediana, etc.

00:08:48:22 - 00:09:22:19
Além disso, é um ótimo momento caso queiram complementar a análise, calculando outras coisas, calculando outras médias. Bom, então a gente foi para o tópico de valores faltantes. E em valores faltantes nós trabalhamos com três colunas diferentes: com a coluna de gênero, idade e salário. Na coluna de gênero, nós olhamos aqui que tinham nove linhas não preenchidas. Então a gente aprendeu a usar o fillna, que é uma função que pega todos os nulos e substitui por um valor único.

00:09:23:16 - 00:09:47:04
Depois a gente trabalhou com a coluna de idade, que é uma coluna numérica, e aí nós calculamos a média. Para a coluna de salário, nós calculamos a média, mas também teve um caso em que a gente substituiu um nulo pela mediana, correto? Aconselho vocês a reassistir a aula de valores discrepantes e colocar comentários comentando o porquê de usar esse ou aquele método de preenchimento dos valores faltantes.

00:09:47:18 - 00:10:14:21
Beleza? Bom, depois nós tivemos a aula de valores discrepantes, os famosos outliers. Aprendemos a plotar gráficos do tipo box plot para visualizar esses outliers; aprendemos a calcular limite superior e inferior. Esses limites indicam a partir de qual ponto estarão os outliers. Mas eu lembro que foi muito importante a gente discutir o contexto dos dados e entender o que seria ou não outlier para nós.

00:10:14:21 - 00:10:44:18
Que tal se vocês colocassem uma célula de texto mesmo nessa parte, explicando qual foi o contexto analisado e as conclusões que chegamos: Fica aí a tarefa. Depois nós tivemos uma aula sobre distribuição amostral e intervalo de confiança. Nós utilizamos a coluna salário aqui como amostra e calculamos o desvio amostral e depois calculamos o intervalo de confiança para o nível de 95% de confiança.

00:10:45:04 - 00:11:07:11
Que tal comentar a cada célula o que estamos fazendo e ao final colocar uma célula de texto resumindo o que foi feito nesse tópico? Acho que seria muito interessante. Depois a gente teve o featuring engineering, e nessa parte nós criamos novas colunas baseadas em colunas que a gente já tinha, colunas já existentes. Lembra da importância do featuring engineering?

00:11:07:27 - 00:11:36:08
Coloquem aí nos comentários um comentário e indiquem aí essa importância. Se vocês tiverem ido além e trabalhado em outras colunas na nossa tabela, coloquem textos indicativos do porquê exploraram outras colunas. E pra finalizar, depois do featuring engineering, a gente falou sobre correlação. Inicialmente nós olhamos a correlação entre duas colunas com valores numéricos: idade e salário. Qual a conclusão que a gente tirou disso mesmo?

00:11:36:18 - 00:12:02:05
Eu acho que vale aqui também colocar uma célula de texto, colocando os comentários sobre o resultado do cálculo de correlação. Nós esperamos uma alta correlação? Baixa? E o que o resultado representa? Depois nós calculamos a correlação entre variáveis discretas e valores categóricos. E de novo, eu falo para vocês comentarem o que esperavam, qual foi o resultado e o que a gente pode concluir.

00:12:03:10 - 00:12:34:16
Gente, para concluir tudo só falta aprendermos a salvar o nosso arquivo. Então, vamos lá para o final, porque a gente já fez muitas modificações nos dados. Praticamente aqui temos uma nova tabela, então a gente precisa salvar. A gente salva os dados usando a biblioteca Pandas. Assim: a gente fala qual é o dado, que no caso se chama “dados” mesmo, ponto to csv e parêntese. Ou seja, estamos pegando os dados e salvando em csv, com o “to csv”. Podemos salvar em outros formatos, até mesmo no próprio excel.

00:12:34:26 - 00:12:58:05
Mas vamos salvar o nosso arquivo em CSV por algumas razões, como para evitar problemas de compatibilidade que podem existir em arquivos Excel por causa de diferentes versões. Além disso, a facilidade de ler e manipular os arquivos CSV por meio de scripts, a questão de arquivos CSVs serem mais compactos também e terem formatos simples e facilmente lidos por uma variedade de programas e linguagens de programação.

00:12:58:21 - 00:13:19:15
E cá entre nós, já vamos falar em CSV para acostumarmos a manipular esse tipo de formato que é muito utilizado na área de análise de dados. Continuando, dentro dos parênteses, nós vamos colocar onde queremos salvar o nosso arquivo e o nome dele. A gente pode ir lá mesmo no lugar que a gente está salvando o notebook e os outros arquivos do nosso curso e colocar aqui o caminho.

00:13:20:00 - 00:13:44:21
Bom, a gente pode pegar esse caminho aqui nessa aba à esquerda, no drive, my drive (no caso, o nosso drive) e aí eu vou lá na pastinha que eu estou salvando o notebook. Eu estou salvando aqui mesmo: análise de dados. Então, aqui na minha pasta eu vou clicar nesses três pontos, copiar caminho. E aqui dentro desse parêntese eu coloco aspas e colo o caminho, ok?

00:13:45:04 - 00:14:17:05
E aí eu vou colocar um nome da minha tabela. Como eu sou uma pessoa muito criativa, original, eu vou colocar aqui: análise dados ponto csv. A gente tem sempre que colocar o formato, no caso a gente está salvando como CSV: ponto csv. E aqui no final eu coloco uma vírgula (já vou fechar aqui no canto pra ficar melhor de visualizar) e aí eu coloco: index igual a false.

00:14:17:05 - 00:14:44:21
Bom, esse index false aqui é para não criar um index a mais. A gente não precisa salvar uma coluna só com os indexes. Vocês se lembram que a gente que index é aquele numerozinho que fica na tabela? Se a gente colocar aqui: dados, em negrito aqui do lado a gente tem o index. Se a gente não colocar esse index igual false, vai criar uma coluna só para esses números.

00:14:45:08 - 00:15:19:10
Então a gente coloca aqui que não precisa. Então a gente coloca false, ok? Então a gente vai salvar. Ao executar essa célula, podemos olhar na pasta que o arquivo vai estar lá. Deixa eu ver aqui do lado lá na pasta PrograMaria já está aqui meu arquivo: análise dados ponto csv. Bom, então a gente salvou os nossos dados e seria legal também vocês criarem uma célula de texto pra concluir e colocar o que vocês entenderam dos dados, o que a gente conseguiu tirar de observações, até mesmo quais perguntas que a gente conseguiu fazer e que foram respondidas ou que não foram respondidas.

00:15:20:10 - 00:15:39:04
Usar as funções que a gente aprendeu para descobrir mais coisas também, podem aprofundar mais nos dados, olhar em outras colunas que a gente não analisou. E é isso. Bom, resumindo até aqui o que usamos de função, os cálculos e tudo mais. Porém, mais importante do que entender as funções é entender os dados e ser crítico em relação a eles.

00:15:39:16 - 00:16:10:14
Por isso, nós sempre estamos falando de contexto em cada análise feita. Vou deixar para vocês usarem as funções que aprendemos para explorar mais, tirando outras conclusões. Vejam, nós aprendemos muita coisa e em cada tópico podemos retirar informações importantes dos nossos dados que vão contribuir para a análise geral. Fazendo os comentários, deixando os textos de análise, eu tenho certeza que o notebook vai ficar ótimo e que qualquer um, mesmo alguém que não seja da área, vai conseguir entender o que você, o que a gente quis discutir nessa análise.

00:16:10:29 - 00:16:19:17
E é isso, pessoal, peço que vocês compartilhem o notebook de vocês entre vocês. Vai ser um atividade bem legal e nos vemos na próxima.