3.5 - Repetindo a primeira análise (parte 1)

00;00;00;17 - 00;00;25;18
Ei pessoal, estamos de volta. Na última aula nós já começamos a explorar nossa planilha, mas agora usando o Python do Google Colab. Conhecemos uma biblioteca que vai ser a nossa maior aliada na análise de dados, que é a Biblioteca Pandas. E a gente tá aprendendo muitas coisas. É importante dizer que se ficou alguma dúvida, volta na aula anterior e assiste de novo, pode assistir quantas vezes vocês quiserem.

00;00;26;07 - 00;00;46;29
Na aula de hoje, nós vamos refazer algumas análises que a gente fez no Módulo 2. Antes nós usamos ali o Google Planilhas, fizemos filtros e tal. Hoje vamos refazer, mas usando códigos. Vamo que vamo. Mas olhando aqui o nosso notebook, esquecemos de uma coisa muito importante na última aula, que é colocar nome, né? É importante a gente colocar um nome.

00;00;47;24 - 00;01;16;14
Então a gente vai aqui em cima, tira esse “untitled” aí, sem título e vamos colocar “análise de dados”. Enter aqui e pronto. A gente já nomeou aí o nosso notebook, que agora é análise dados. Para ficar organizado também, a gente está começando uma nova aula, vamos criar aqui uma célula de texto e vamos colocar aqui um título só pra gente poder separar o que foi da aula anterior para essa aula.

00;01;17;00 - 00;01;56;27
Vamos colocar aqui “repetindo análise do Excel”. Ok. Enter. Shift+enter aqui para executar. Ótimo. Agora a gente pode começar essa próxima aula com tudo organizadinho. Bom, nós já demos algumas olhadas iniciais. A gente viu tail, head, as primeiras, últimas linhas, olhamos o tamanho da tabela, quais são as colunas etc. A primeira coisa que a gente fez lá no Excel foi um filtro na coluna de gênero para ver as respostas apenas de quem respondeu feminino, correto?

00;01;57;17 - 00;02;24;15
A Biblioteca Pandas permite que a gente realize filtro de forma prática. A gente consegue fazer o filtro de gênero assim... A gente coloca: tabela dados, colchetes, tabela dados, colchetes de novo, aspas simples ou duplas - não tem diferenciação aqui - e aí a gente coloca o nome da coluna dentro dessas aspas. O nome da coluna tem que ser exatamente igual ao nome da coluna da tabela.

00;02;24;15 - 00;02;52;05
Nossa Carol, eu esqueci qual é o nome da coluna. Uma dica aqui: dá mais uma célula de código aqui, acrescenta. Coloca essa célula de código antes só pra poder ficar numa ordem ok. E a gente aprendeu que tem uma funçãozinha bacana do pandas, que a gente consegue ver o nome de todas as colunas, que é o dados, nome da nossa tabela, ponto columns, coluna em inglês, a gente executa e aqui a gente tem o nome de todas as colunas.

00;02;52;17 - 00;03;23;05
Qual que é a coluna que a gente quer mesmo? É gênero. Vamos pegar igualzinho está aqui, olha: gênero, todo maiúsculo, sem acento. Ctrl+C. Vamos lá embaixo, na outra célula que a gente está querendo fazer o filtro, ctlr+v. Tá igual, não há erro. Então o que a gente tem aqui? A gente tem primeiro que, dentro da tabela de dados, a gente está pegando a coluna de gênero e aí a gente quer algo específico dentro dessa coluna, que é todas as pessoas que marcaram o gênero feminino.

00;03;23;09 - 00;03;41;15
Então a gente vai colocar dois sinais de igual porque aqui a gente não está atribuindo valor. Quando a gente atribui valor, a gente coloca um sinal. A gente está querendo comparar, a gente quer tudo aqui seja igual a. Por isso, a gente usa dois sinais de igual. E aqui na frente a gente coloca aspas simples ou dupla e feminino aqui na frente.

00;03;41;15 - 00;04;06;27
E, de novo, esse feminino tem que ser exatamente igual ao que está na coluna. Então, se lá na coluna é feminino tudo maiúsculo, tem que ser maiúsculo. Se está tudo minúsculo, tudo minúsculo. Carol, fiquei na dúvida de como é lá na coluna. Podemos também criar aqui mais uma célula de código, lembrando de colocar antes para poder ter uma ordem bonitinha.

00;04;07;14 - 00;04;31;08
A gente pode visualizar essa coluna como: dados, colchetes. A gente coloca o nome da coluna, a gente sabe agora que é gênero, tudo maiúsculo. A gente executa aqui, aí a gente está chamando apenas essa coluna e a gente consegue ver como está escrito masculino e como está escrito feminino. Olha, o feminino está com F maiúsculo, o restante minúsculo.

00;04;31;17 - 00;05;01;02
A gente coloca exatamente igual. Se a gente colocar diferente e executar, vai dar erro, porque não tem lá esse feminino que a gente escrever diferente. Então aqui a gente tem o nosso filtro. A gente está filtrando dentro da tabela dados, tudo o que está na coluna de gênero e que seja igual a feminino. Executando, a gente tem aqui todas as respostas em que as pessoas colocaram o gênero feminino. Se a gente for nessa tabela, nessa coluna de gênero, está aqui tudo feminino, beleza?

00;05;01;12 - 00;05;31;03
A gente pode usar o símbolo exclamação igual, que significa diferente para filtrar. Por exemplo, a gente quer filtrar tudo que não é do gênero masculino, ou seja, a gente está querendo filtrar o feminino. Mas aí a gente pode colocar de um jeito diferente. Vamos criar mais uma célula de código e vamos filtrar assim: dados (essa tabela dados) colchete, dados de novo.

00;05;31;21 - 00;06;08;23
Qual é a nossa coluna? Gênero. E agora a gente não quer dois iguais, a gente quer exclamação igual, ou seja, que seja diferente. Aqui a gente coloca então exclamação igual, tudo que seja diferente a masculino. Ok. Ou seja, a gente está fazendo um filtro de tudo o que seja diferente do masculino. Executando aqui, a gente pode vir aqui e olhar na nossa coluna de gênero que está tudo feminino, ou seja, filtramos tudo o que é diferente de masculino.

00;06;09;09 - 00;06;35;07
Qual é a diferença desses dois filtros? Uma coisa é eu pegar tudo o que seja igual a feminino, outra coisa é eu pegar tudo que seja diferente de feminino. Porque se alguém respondeu nessa mesma coluna de gênero “prefiro não responder” ou deixou o campo vazio, já é diferente de masculino, então a resposta vai estar aqui. Tanto que se a gente olhar aqui no final do filtro, tudo igual a feminino, a gente tem uma quantidade de linhas de 1056.

00;06;35;24 - 00;07;02;26
E se a gente olhar aqui a quantidade de linhas diferentes de masculino, a gente já vai ter uma quantidade de linhas de 1077. Beleza? São filtros diferentes. Vamos supor que eu esqueci como estava quando não quiseram preencher o campo de gênero, mas lembro que tinha a palavra “não”. Então a gente pode usar a função contains, que é basicamente filtrar tudo que contém alguma coisa que vamos definir.

00;07;02;26 - 00;07;34;22
O formato do uso da função é: str.contains e parênteses. E dentro dos parênteses, a gente coloca a palavra que a gente está buscando e outros parênteses possíveis também, como o “na”. Esse NA preenche os valores nulos pelo que a gente definir. Calma que vamos colocar aqui como que é. A gente coloca mais uma célula de código e a gente coloca: dados, que é a nossa tabela primeiro, depois a nossa coluna, que é de gênero.

00;07;34;26 - 00;08;00;10
Aí, ao invés de a gente colocar aqui: igual igual ou exclamação igual, a gente coloca a função que é: ponto str ponto contains e o parêntese. Dentro dos parênteses, a gente vai colocar qual é a palavra que a gente está buscando. No caso, a gente está buscando o “não”, porque a gente lembra que tinha “não”, a gente não lembra qual era a frase completa.

00;08;00;15 - 00;08;24;06
E aqui a gente coloca uma vírgula e coloca “na” igual a false. O que é esse “na igual a false”? NA é nulos. E aí a gente está querendo dizer que todos os NAs ali vão ser colocados como false. A gente colocou aqui false, mas a gente pode colocar NA e substituir por um texto: não nulos, por exemplo.

00;08;25;21 - 00;08;53;03
É só para não confundir os valores, mesmo. Então vou voltar aqui: NA false. E se a gente executar, qual é a ideia? A ideia é pegar todos os valores da coluna de gênero que contenha a palavra “não”. Vamos executar aqui agora. E executando a célula, aparecem então 12 linhas que a gente pode ver no final: 12 linhas.

00;08;53;03 - 00;09;16;04
E aqui na coluna de gênero, a gente vê a frase “prefiro não informar”. A gente não lembrava a frase, mas a gente sabia que tinha uma palavra ali que era o “não”, e através disso a gente conseguiu filtrar com a função contains. E se tivermos uma coluna com dados numéricos, podemos usar outros critérios de filtro também.

00;09;16;16 - 00;09;43;02
Por exemplo, na coluna de idade, podemos filtrar todas as idades maiores que 30. Basta a gente usar os símbolos matemáticos para maior ou menor. Vamos colocar aqui mais uma célula de código e vamos lá para o nosso filtro. Tabela dados, abre colchetes. Agora a gente vai colocar a nossa coluna. que não é gênero, é idade. Nossa, esqueci como é que está escrito idade na coluna.

00;09;43;14 - 00;10;11;14
Pode colocar lá dados ponto columns, como a gente fez aqui mais pra cima. Vamos dar uma olhada. Vamos subir que não foi tão pra cima, né? Não tinha pensado, mas está aqui nosso dados ponto columns. Vamos dar uma olhada. Olha a idade: idade, tudo maiúsculo. Ok, vamos dar um ctrl+c e vamos lá pro final, onde está o nosso filtro que a gente tá fazendo e aqui eu dou o ctrl+v. Ok.

00;10;12;02 - 00;10;35;26
Idade. E aí o que a gente quer mesmo? A gente quer todas as pessoas que são maiores de 30. A gente usa aquele símbolo matemático para maior de 30, a gente coloca aqui 30. E aí está o nosso filtro, ou seja, dentro da tabela de dados eu quero na coluna de idade tudo que seja maior de 30. Então a gente executa e aqui a gente tem o resultado.

00;10;35;26 - 00;11;14;24
E se a gente der uma olhada aqui na coluna de idade, está aqui: 39, 32. Tudo maior do que 30. Ok, 1956 colunas. Outra coisa, percebam que na célula de cima eu usei aspas simples para gênero aqui. E assim, o legal do Python é que tanto faz: eu posso colocar aspas simples aqui, aspas duplas aqui, não tem problema, beleza? Mas a gente não pode misturar também.

00;11;14;28 - 00;11;48;01
Se eu começar com aspas duplas e terminar com uma simples, vai dar um erro. Então vamos padronizar aí pra não ter esses errinhos assim. E aqui filtramos apenas maiores de 30. Se quiséssemos, a gente podia filtrar também as pessoas com maior de 30, mas incluir o 30 também. Eu quero que seja maior ou igual a 30.

00;11;48;14 - 00;12;11;25
A gente pode colocar o igual aqui depois no símbolo de maior. Aí a gente está incluindo 30. Beleza? Então aqui, se a gente executar de novo, todas as pessoas que têm uma idade igual ou maior que 30 estão no nosso filtro, ok? E se a gente usasse combinar os filtros: eu quero ver todos que responderam que são do gênero feminino e que são maiores de 30 anos.

00;12;12;11 - 00;12;39;01
Para isso, a gente usa o operador “&” (E comercial), que junta dois filtros. E a gente coloca entre parênteses também para poder ficar um negócio organizado, bonitinho. Bom, então vamos lá colocar mais uma célula de código e vamos lá iniciar o nosso filtro. Qual que é a tabela? Tabela dados, colchetes... Qual é a coluna? Vamos começar pela coluna de idade. Opa, estava esquecendo das aspas, pode não.

00;12;39;10 - 00;13;11;02
Idade maior que 30. Ok, aí eu quero idade maior que 30 e gênero feminino. Então a gente usa o & aqui. Ah, Carol, não cria "e", não queria maior que 30 e igual a feminino, queria "ou": ou feminino ou maior. Aí a gente usa um outro símbolo, que é aquela barra retinha. Mas no caso aqui vamos usar o & mesmo, quero maior que 30 e feminino.

00;13;11;05 - 00;13;40;04
Então vamos lá para nossa coluna dados: gênero igual igual a feminino. Não esqueça das aspas e que o nome tem que ser igual. Ok, vamos executar aqui. Bom, deu erro. O erro aqui é porque está parecendo que não suporta, o nosso & aqui, nossa operação não suporta esse tipo aqui.

00;13;40;15 - 00;14;12;14
Enfim, o que é esse erro? Esse erro está basicamente dizendo que não dá para fazer esses dois filtros, está misturando. Então, pra não misturar, a gente vai colocar cada filtro entre parênteses. E aí a gente pode vir aqui no início, colocar o parêntese, ir lá no final, colocar parênteses para o primeiro filtro. Ou a gente pode fazer um negócio muito legal, que é selecionar aqui o que a gente quer colocar entre parênteses e já apertar o shift+parêntese aqui, o parêntese que abre. E aí, automaticamente, vai colocar aquilo que a gente selecionou entre os parênteses.

00;14;12;23 - 00;14;42;11
Rápido e prático, né, gente? E aí a gente pode executar e a gente vai ter um resultado legal. A gente pode combinar quantos filtros a gente quiser e também usar outros operadores, tipo esse "ou", aquela barrinha que eu mostrei para vocês. Vamos deixar eles bem detalhados nos materiais complementares. Façam aí outros filtros. Que tal um filtro aí para a gente ver os resultados de quem respondeu na tabela de cor e raça, que é amarela?

00;14;43;01 - 00;14;50;24
E um filtro pra gente saber quantas pessoas tem abaixo de 40 anos. E um filtro misturando as duas coisas...