Preparando os dados - Limpando os dados com Google Cloud Dataprep

Vou mostrar como usar algumas técnicas para limpar um conjunto de dados que esteja muito sujo. Vou usar um ótimo exemplo que foi criado por Sarah Cohen. Ela ainda estava no New York Times, ela está agora na Universidade do Estado do Arizona. Foi um ótimo exemplo, foi uma aula que eu assisti que ela ministrou durante a conferência para reportagens assistidas por computador, NICAR 2016, em Denver.

É um ótimo exemplo porque este é um conjunto de dados que tem tudo. Tem erros de digitação, tem espaços em branco, tem tudo. Então, este é um conjunto de dados que foi compilado a partir de relatórios de cuidados de longo prazo do Medicaid no Estado de Nova York. E essa é a aparência dele.

Ele está muito bagunçado, ele tem em um lugar aqui o nome dos planos. Você tem os municípios no Estado de Nova York e tem o total de matrículas. Aqui você tem a data, o mês e depois o ano. Mas então você tem padrões diferentes para isso, e é realmente confuso. Cada tabela na planilha corresponde a um mês em determinado ano e então ela construiu uma planilha completa, uma única planilha com todas as informações de que precisamos, certo? Então, nós queremos sair disso e chegar a algo como isto. Está muito mais arrumado. Temos os nomes dos planos em uma coluna, temos o município em outra, temos o número de matrículas, o mês e o ano. Cada variável tem sua própria coluna, cada coluna tem apenas um tipo de dado e temos tudo junto aqui.

Também estamos eliminando os totais, certo? Porque não precisamos deles. Podemos calcular depois, então não precisamos de algo aqui que vai nos dizer o total porque vamos usar um programa de processamento para fazer isso.

O programa que eu vou usar hoje, ela deu uma aula sobre OpenRefine e eu vou ensinar a limpar o mesmo conjunto de dados usando OpenRefine em outro vídeo. Mas a ferramenta que eu vou usar para este aqui se chama Dataprep.

Dataprep, da Trifacta, é uma ferramenta que faz parte da suíte de ferramentas Google Cloud para ajudar você a analisar, processar e limpar dados. Dataprep foi feito sob medida para conjuntos de dados enormes, mas também é um programa muito interessante para explorar e fazer visualização rápidas e limpezas rápidas de conjuntos de dados. Mas é realmente poderoso quando você usa Dataprep para conjuntos de dados realmente enormes, com milhões e milhões de linhas e você quer executar rotinas de limpeza que são as mesmas. Mas também podemos usar para este pequeno conjunto de dados que Sarah reuniu. Este é um pequeno conjunto de dados, só tem cerca de seis mil, quase sete mil linhas, então não é tão grande.

Mas deixe-me mostrar para vocês. Logo que você entra em cloud.google.com/dataprep você pode ir para o console. Ele vai pedir para você criar uma conta e assim que você criar uma conta você vai para uma tela como esta, que carrega o aplicativo e então você pode começar a usar Dataprep. Então, para fazer isso, vamos importar dados aqui e para

importar dados, vamos escolher um arquivo. Aqui está o arquivo Excel, vamos colocar isso para você baixar nos fóruns, não se preocupe. Basta clicar aqui e então ele vai começar a fazer upload para a ferramenta, certo? Ele reconhece automaticamente que é um arquivo do Excel. E ele me diz que há uma aba no arquivo e quero carregar essa aba específica. Eu vou clicar em "Import & Wrangle" para que ele possa começar a trabalhar no arquivo imediatamente.

Assim que eu faço isso, Dataprep me saúda com sua interface e ela é bastante intuitiva. Você não tem que se preocupar com muitas das opções que estão aqui para esta complexidade do conjunto de dados que queremos, mas algumas coisas são dignas de nota. A primeira é esta representação visual das colunas aqui no topo. Dataprep vai sempre dizer se há colunas faltando ou valores incompatíveis, então você sempre pode encontrar isso em um clique. Agora ele me diz que a coluna 2 aqui tem mais de quatro mil valores faltando aqui. Estes são todas essas linhas em branco e ele me diz imediatamente aqui que aqueles estão faltando. Ele também me diz que na coluna 4 há valores incompatíveis. São valores que não são do mesmo tipo de dados. Isso me ajuda a encontrar esses valores.

Então, o que eu quero fazer aqui primeiro é apenas lembrar que isso é o que eu quero. Eu quero uma coluna com os planos, quero outra coluna com os municípios, e então as matrículas, mês e ano. Não tem que estar na mesma ordem, mas isso é o que queremos, certo?

Então, voltando aqui para Dataprep. O que eu vou fazer em primeiro lugar é extrair o ano, certo? Eu quero uma coluna com o ano. Então eu quero pegar esse ano aqui e criar uma coluna com isso. Eu vou selecionar aqui 2009 e Dataprep vai automaticamente aqui à direita sugerir algumas ações que eu posso querer fazer. Então o que eu quero aqui é extrair os valores, certo? Eu quero extrair valores que são correspondentes a dígito quatro, então quatro dígitos aqui. Não explicitamente ou literalmente 2009, eu quero quatro dígitos porque assim eu obtenho todos os anos que estão aqui neste conjunto de dados. Então isso está correto, é isso o que eu queria fazer. E veja que em azul ele destaca a minha coluna original, em amarelo ele dá uma pré-visualização da coluna que vai criar. E ele vai criar uma coluna que tenha valores 2008, 09, 10, 11, 12, 13 e 14 também. Ele inclusive me dá a quantidade de valores, então são doze 2018, que é 0,18% desta coluna e doze 2009, e assim por diante. Então eu vou continuar e clicar em "add" (Adicionar). Ele acrescentou a coluna aqui. Eu não vou preocupar com os nomes, eu vou mudar o nome de todas as colunas depois.

Mas agora eu quero extrair o mês, certo? E para extrair um mês, eu vou selecionar um mês aqui. Mas não vai ser tão fácil assim, certo? Com o ano foi bem fácil, mas com o mês aqui o que nós vamos fazer é realmente selecionar o extrato aqui, mas vamos editar o que está acontecendo, e você pode selecionar qualquer ação. A coluna da qual extrair é a coluna 2. E nós vamos fazer um padrão de texto customizado. Então vou apagar isso e vou dizer que vai começar a extrair a partir de "NYS", certo? Há sempre um NYS aqui antes do mês, e vai ser sempre antes destes quatro dígitos. Lembra quando extraímos os quatro dígitos? Vamos digitar aqui do jeito que estão indicando aqui. Então dígito e depois quatro. Tudo certo. Aqui ele mostra que está extraindo tudo o que eu preciso, na verdade mais do que eu

preciso, porque está extraindo também os espaços aqui que estão em torno dele e também a vírgula. Mas vamos lidar com isso em breve. Eu vou adicionar e daí ele cria outra coluna com os meses. E o que eu vou fazer com este espaço vazio aqui, eu vou apenas selecionar esse espaço vazio. E veja como Dataprep já sugere que a gente substitua os valores aqui, mesmo este espaço antes de janeiro e depois de janeiro que não tem nada. Então vamos deletar aqueles espaços vazios. Então eu vou selecionar este aqui, e agora vou clicar em adicionar. E você vê que ele removeu tudo aquilo. Agora vou remover a vírgula também, apenas encontrar uma vírgula e substituir por nada, então você vai deletar a vírgula.

Tudo certo, então temos uma coluna para meses e uma coluna para anos. Mas agora veja que temos os valores aqui. Então janeiro está aqui e então rolamos para baixo, janeiro está aqui novamente com um ano diferente. Seguimos rolando para baixo, vemos janeiro de novo. Então queremos ter certeza de que todos estes valores aqui estão preenchidos com janeiro e todos estes valores aqui estão preenchidos com 2009, até chegar a 2010 e depois eles são preenchidos de novo até encontrar o outro ano, 2011. E assim por diante. Isso é chamado de preenchimento para baixo e podemos fazer isso automaticamente no Dataprep apenas usando um novo passo.

Então vamos clicar em "new step" (nova etapa) e então vamos digitar aqui "window" (janela). E depois há uma fórmula chamada "fill" (preenchimento), e então abrimos parênteses e selecionamos uma coluna. Vamos fazer primeiro na coluna 5, que é a coluna no momento que estamos usando para meses. Vou digitar "fill" e depois "column5" (coluna5). E então eu vou escrever aqui o número fonte que é o número da linha no arquivo fonte original aqui. Então você digita isso e veja que a coluna azul aqui, que é a fonte, se torna esta coluna amarela aqui. Então janeiro é preenchido abaixo até ele encontrar um valor diferente, que é fevereiro. Se eu seguir rolando para baixo até encontrarmos um mês diferente, ele encontra fevereiro e depois até o fim do conjunto de dados.

Vou apenas adicionar isto aqui e ele vai criar outra coluna aqui para nós. Esta será nossa coluna de mês e eu vou fazer o mesmo com a coluna ano, então "new step", "window" e então eu vou colocar "fill", e essa é a coluna 1. Aqui eu não preciso disso, "column1" e depois número da fonte aqui, "sourcerownumber". E então sim, ele cria aqui outra coluna para mim que vai ser minha coluna de ano. Agora clique em "add".

Certo, eu tenho o município aqui, a matrícula aqui, o mês, o ano, o plano. Ele está começando a se parecer com esta coisa que nós queremos, mas ainda precisamos fazer algumas coisas. A primeira coisa que vamos fazer depois de ter criado essas colunas, podemos querer agora mudar o nome de todas elas. Eu vou renomear essa aqui como plano de saúde. Novamente, eu só cliquei aqui nesta seta apontando para baixo. Cliquei em "rename" (renomear), e então eu vou renomear apenas para "plano". Eu poderia querer fazer outra coisa com essa também, mas só vou renomear esta para "plano". Mesma coisa. Esta, na verdade, eu vou remover porque eu não preciso dela. Também vou remover esta, delete. Renomear esta como "município". Esta renomeie como "matrícula". Tudo certo. E esta como "mês", e esta como "ano".

Tudo certo, então agora veja a representação visual aqui no topo. Podemos encontrar os valores que faltam também e todos os valores incompatíveis. Mas vamos começar a remover coisas como o total, certo? Não precisamos desse total, então vamos destacá-lo e pedir ao Dataprep para deletar todas as linhas que correspondem a "total", porque não precisamos disso.

Agora podemos querer fazer a mesma coisa aqui com os planos de saúde, com os planos Medicaid para preencher para baixo, certo? Porque queremos que isso seja sobre o mesmo plano que está aqui, porque se você der uma olhada na tabela original, eles agruparam os planos Medicaid juntos e há espaços vazios para indicar que eles são os mesmos. Então, precisamos ter certeza de que aqueles sejam preenchidos com os nomes dos planos também. Vamos continuar e fazer isso. E se você se lembra, vamos a "new step", "window", a fórmula é "fill", e então vamos a "plano", que é o nome da coluna agora, e depois "sourcerownumber" (número da coluna fonte). E então ele cria uma nova coluna aqui que tem todos os valores preenchidos. Quero deletar isso também e mover esta para o começo. Eu vou adicionar e mudar o nome desta aqui para "plano" de novo.

Tudo bem, agora eu quero remover todos os valores que faltam aqui porque ou eles têm totais ou eles são sobre linhas que não fazem parte do meu conjunto de dados, meu conjunto de dados final. Vou simplesmente removê-los. Também vou dar uma olhada em valores incompatíveis aqui, esses valores incompatíveis são aqueles que não são números. Então eu vou remover os valores incompatíveis também, e há valores ausentes aqui também que eu quero remover. Isso já está muito mais parecido com o conjunto de dados que temos aqui e completamos isso em apenas vinte e um passos. Então este é o conjunto de dados limpo final.

Agora, o que você deve fazer, se você quiser exportar isso para um arquivo CSV, você seleciona essa opção aqui, "run job" (executar trabalho). Então Dataprep usará sua infraestrutura de nuvem para obter essa receita que você acabou de criar e aplicá-la aos conjuntos de dados que você importou. Por isso ele é muito mais poderoso quando você está lidando com conjuntos de dados enormes, mas também pode fazer isso com um pequeno. Você clica em "run job", e então você verifica se está tudo ok. Há opções aqui que você pode querer selecionar, por exemplo, eu quero selecionar essa que vai criar um arquivo CSV e eu tenho mais opções aqui como "include headers as first row on creation" (incluir cabeçalhos como primeira linha ao criar) e então eu atualizo isto. Vou selecionar um local onde salvar o arquivo em minha conta na nuvem. Eu seleciono as regiões e então eu executo o trabalho. E assim que o trabalho estiver pronto, você terá um arquivo CSV como este, eu tenho um aqui comigo também. Então aqui está um arquivo CSV que você pode querer carregar.

Então é isso para o Dataprep. Vá a cloud.google.dataprep e experimente.