Associações e correlações: Os Elementos Essenciais

Uma estratégia holística para descobrir a história dos seus dados

LEE BAKER -CEO Chi-Squared Innovations

Capítulo 1: Coleta e limpeza de dados

O primeiro passo em qualquer projeto de análise de dados é coletar e limpar os dados. Se você tiver a sorte de ter recebido um conjunto de dados perfeitamente limpo, parabéns - você está no bom caminho ...

Para o resto de nós, porém, há sempre um pouco de trabalho duro a ser feito antes que se possa chegar à alegria da análise (sim, eu sei, realmente tenho que arrumar coisa melhor para fazer na vida .....).

Neste capítulo, você aprenderá as características que um bom conjunto de dados deve ter e como ele deve ser formatado para torná-lo acessível à análise por associação e realizar testes de correlação.

Mais importante ainda, você vai aprender por que não é necessariamente uma boa idéia coletar dados de vendas de sorvete e creme para hemorróidas no mesmo conjunto de dados...

Se você está satisfeito com o seu conjunto de dados e tem certeza de que ele não precisa de limpeza, então você pode, certamente, ignorar este capítulo. Não levarei isso para o lado pessoal - sério!

1.1: Coleta de Dados

A primeira pergunta que se deve fazer antes de se começar qualquer projeto é "Qual é a minha pergunta?". Se você não sabe qual é a sua pergunta, então você não saberá como obter uma resposta. Em ciência e estatística, isso é chamado de ter uma hipótese. Hipóteses típicas podem ser:

- O tabagismo está relacionado com o câncer de pulmão?
- Existe alguma associação entre vendas de sorvete e creme de hemorroidas?
- Existe correlação entre o consumo de café e a insônia?

É importante começar com uma pergunta, porque isso irá ajudá-lo a decidir quais dados você deve coletar (e quais você não deve).

Não é normal que você possa responder a esses tipos de perguntas coletando dados apenas nessas variáveis. É muito mais provável que haja outros fatores que possam influenciar a resposta e todos esses fatores devem ser levados em consideração. Se você quiser responder à pergunta 'o fumo está relacionado ao câncer de pulmão?', Você normalmente também coletará dados sobre idade, altura, peso, histórico familiar, fatores genéticos, fatores ambientais, etc., e seu conjunto de dados irá se tornar bem grande em comparação com sua hipótese.

Portanto, quais dados você deve coletar? Bem, isso depende da sua hipótese, a sabedoria percebida do pensamento atual e pesquisas anteriores realizadas, mas, em última análise, se você coletar dados com sensibilidade, você normalmente terá resultados sensatos e vice-versa, então é uma boa ideia levar algum tempo pensando com cuidado antes de começar.

Não vou entrar nos detalhes profundos de coleta e limpeza de dados aqui, mas é importante que o seu conjunto de dados esteja em conformidade com alguns padrões antes de começar a analisá-lo.

A propósito, se você quiser uma cópia do meu livro "Limpeza de dados práticos" - '*Practical Data Cleaning*'-, você pode obter uma cópia gratuita seguindo as instruções no minúsculo anúncio no final desta seção ...

1.1.1: Lista de verificação do conjunto de dados

Para iniciar o processo, você deve ter um conjunto de dados bem formatado que esteja armazenado em uma única planilha Excel, cada coluna é uma única variável, a linha 1 contém os nomes das variáveis, e abaixo desta, cada linha é uma amostra ou paciente distinto. Deve ser algo como a Tabela 1.

Tabela 1. Conjunto típico de dados para associação e análise de correlação.

UniqueID	Gender	Age	Weight	Height	…
1	Male	31	76.3	1.80	…
2	Female	57	59.1	1.91	…
3	Male	43	66.2	1.86	…
4	Male	24	64.1	1.62	…
…	…	…	…	…	…

Ok, então vamos lá. Aqui estão as características essenciais que um conjunto de dados deve ter para estar pronto para análise de associação e correlação:

- Seu conjunto de dados é uma matriz retangular de dados. Se os seus dados estão espalhados em diferentes planilhas ou tabelas, não é um conjunto de dados (*dataset*), é um banco de dados, e não está pronto para análise.
- Cada coluna de dados é uma única variável correspondendo a uma única informação (como idade, altura, peso, etc.).
- A coluna 1 é uma lista de números consecutivos únicos a partir de um. Isso permite que você identifique de forma exclusiva uma determinada linha e recuperar a ordem original do seu conjunto de dados com um único comando de classificação.
- A linha 1 contém os nomes das variáveis. Se você usar as linhas 2, 3, 4, etc., como os nomes das variáveis, você não poderá inserir seu conjunto de dados em um programa de estatísticas.
- Cada linha contém os detalhes de uma única amostra (paciente, caso, tubo de ensaio, etc.).
- Cada célula deve conter uma única informação. Se você inseriu mais de uma informação em uma célula (como data de nascimento e sua idade), você deve separar a coluna em duas ou mais colunas (uma para a data de nascimento, outra para a idade).
- Não insira o número zero em uma célula, a menos que o que foi medido, contado ou calculado resulte na resposta *zero*. Não use o número zero como um código para significar *'sem dados'*.

Para o resto deste livro, é assim que eu vou assumir que seu conjunto de dados está estabelecido, então posso usar os termos *Variável* e *Coluna* de forma intercambiável, da mesma forma os termos *Linha*, *Amostra* e *Paciente*.

1.2: Limpeza de dados

O próximo passo é a limpeza desses dados. Você pode ter cometido alguns erros na digitação e alguns de seus dados podem não ser utilizáveis. Você precisa encontrá-los e corrigi-los. Uma possibilidade é que seus dados podem não ser adequados para o propósito em si e podem enganá-lo na busca das respostas às suas perguntas.

Mesmo depois de corrigir os erros óbvios de entrada, ainda podem existir outros tipos de erros nos seus dados que são mais difíceis de encontrar.

1.2.1: Verifique se seus dados têm sentido.

Apenas porque seu conjunto de dados está limpo, isso não significa que eteja correto - a vida real segue regras, e seus dados também devem seguir. Há limites nos níveis de participantes em seu estudo, portanto, verifique se todos os dados se encaixam dentro de limites razoáveis. Calcule os valores mínimo, máximo e médio das variáveis para ver se todos os valores têm sentido.

Às vezes, juntar duas ou mais peças de dados podem revelar erros que, de outra forma, poderiam ser difíceis de detectar. A diferença entre a data de nascimento e a data do diagnóstico lhe dá um número negativo? Seu paciente tem mais de 300 anos de idade?

1.2.2: Verifique se as suas variáveis têm sentido

Uma vez que você tenha um conjunto de dados perfeitamente limpo, é relativamente fácil comparar variáveis entre si para descobrir se há uma relação entre elas (o assunto deste livro). Mas só porque você pode, isso não significa que você deve. Se não há uma boa razão pela qual deve haver relacionamento entre vendas de sorvete e creme de hemorroidas, então você deve considerar expulsar uma ou ambas variáveis do seu conjunto de dados. Se você coletou seus próprios dados de fontes originais, então você terá de considerar de antemão quais dados são sensatos colecionar (você não tem ?! ??), mas se o seu conjunto de dados é um pastiche de dois ou mais conjuntos de dados, então você pode encontrar combinações estranhas de variáveis.

Você deve verificar suas variáveis antes de fazer qualquer análise e considerar se é sensato fazer essas comparações.

Enfim, agora você coletou, limpou e verificou se seus dados são sensatos e adequados.

No próximo capítulo, analisaremos os conceitos básicos de classificação de dados e apresentamos os quatro tipos de dados.

Capítulo 2: Classificação de Dados

Neste capítulo, você aprenderá a diferença entre dados *quantitativos* e *qualitativos*. Você também aprenderá sobre os tipos de dados proporcão, intervalar, ordinal e nominal e quais as operações que você pode executar em cada um deles.

2.1: Dados quantitativos e qualitativos

Os dados ou são *quantitativos*; medidos com algum tipo de instrumento de medição, tipo régua, jarro, balanças, cronômetro, termômetro, e assim por diante, ou são *qualitativos*; dizem respeito a uma característica observada de interesse que é colocada em categorias, por. ex.: gênero (masculino, feminino), saúde (saudável, doente), opinião (concordar, neutro, discordar).

Os dados quantitativos e qualitativos podem ser subdivididos em quatro outras classes de dados:

Quantitativo (medido)

- Proporção
- Intervalar

Qualitativo (categorizada)

- Ordinal
- Nominal

A diferença entre eles pode ser estabelecida fazendo apenas três perguntas:

Ordenados

- Pode ser detectado algum tipo de progresso entre pontos ou categorias de dados adjacentes ou os dados podem ser ordenados de forma significativa?

Equidistantes

- A distância entre pontos ou categorias de dados adjacentes é consistente?

Zero Significativo

- A escala de medição inclui um valor zero único e não arbitrário?

Passemos por cada uma das quatro classes para ver como elas se encaixam nessas perguntas.

2.1.1: Dados nominais

Dados nominais são observados, não medidos, não ordenados, não equidistantes e sem zero significativo.

Podemos diferenciar entre categorias baseadas apenas em seus nomes, daí o título 'nomen' (do nome latino, que significa 'nome').

A Figura 2.1 pode ajudá-lo a decidir se seus dados são nominais.

Figura 2.1: Características dos dados nominais

Exemplos de dados nominais incluem:

- Gênero (masculino, feminino)
- Nacionalidade (britânica, americana, espanhola, ...)
- Gênero / Estilo (Rock, Hip-Hop, Jazz, Clássico, ...)
- Cor favorita (vermelho, verde, azul, ...)
- Animal favorito (porco-da-terra, coala, bicho-preguiça, ...)
- Ortografia favorita para 'pista' (rasto, rastro)

As únicas operações matemáticas ou lógicas que podemos realizar nos dados nominais são dizer se uma observação é (ou não é) a mesma que outra (igualdade ou desigualdade), e podemos determinar o item mais comum ao encontrar a *moda* (você se lembra disso nas aulas do ensino médio?).

Outras formas de encontrar o meio da classe, como a *mediana* ou a *média*, não têm sentido, porque **ranking** - *uma "classificação ordenada de acordo com critérios determinados* - não tem significado para dados nominais.

Se as categorias são *descritivas* (nominais), como 'Porco', 'Carneiro' ou 'Cabra', pode ser útil separar cada categoria em sua própria coluna, como Porco [Sim; Não], Ovelha [Sim; Não], e Cabra [Sim; Não]. Estas são chamadas de variáveis "falsas" ou “artificiais” (dummy variables) e podem ser muito úteis de forma analítica. Mais sobre isso mais tarde ...

2.1.2: Dados ordinais

Os dados ordinais são observados, não medidos, são ordenados, mas não equidistantes e não têm zero significativo.

Suas categorias podem ser encomendadas (1º, 2º, 3º, etc. - daí o nome 'ordinal'), mas não há consistência nas distâncias relativas entre categorias adjacentes. A Figura 2.2 mostra as características dos dados ordinais.

Figura 2.2: Características dos Dados Ordinais

Exemplos de dados ordinais incluem:

- Saúde (saudável, doente)
- Opinião (concordar, em grande parte concordar, neutro, em grande parte discordar, discordar)
- Grau de Tumor (1, 2, 3)
- Estágio tumoral (I, IIA, IIB, IIIA, IIIB, etc.)
- Hora do dia (manhã, meio dia, noite)

Matematicamente, podemos fazer comparações simples entre as categorias, como mais (ou menos) saudáveis / graves, concordar mais ou menos, etc., e uma vez que há uma ordem para os dados, podemos classificá-los e calcular a mediana (ou moda, mas não a média) para encontrar o valor central.

É interessante notar que, na prática, alguns dados ordinais são tratados como dados intervalares - a graduação dos tumores é um exemplo clássico nos cuidados de saúde - porque os testes estatísticos que podem ser usados em dados de intervalo (eles atendem a exigência de intervalos iguais) são muito mais poderosos do que aqueles usados em dados ordinais. Isso só estará correto se seu método de coleta de dados garantir que a regra equidistante não está "forçada" demais.

2.1.3: Dados Intervalares

Dados Intervalares são medidos e ordenados como itens equidistantes, mas não têm o zero significativo. Os dados de intervalo podem ser contínuos (têm um número infinito de passos) ou discretos (organizados em categorias) e o grau de diferença entre itens é significativo (os intervalos são iguais), mas não a proporção deles.

A Figura 2.3 ajudará na identificação de dados de intervalo.

Figura 2.3: Características dos dados de intervalo

Exemplos de dados de intervalo incluem:

- Temperatura (° C ou F, mas não Kelvin)
- Datas (1066, 1492, 1776, etc.)
- Intervalo de tempo em um relógio de 12 horas (6h, 18h)

Embora os dados de intervalo possam parecer muito semelhantes aos dados de porcentagem (razão), a diferença está em seus pontos zero definidos.

Se o ponto zero da escala foi escolhido arbitrariamente (como o ponto de fusão da água ou de uma época arbitrária, como AD), os dados não podem estar na escala de proporção e devem ser intervalos.

Matematicamente, nós podemos comparar os graus dos dados (igualdade / desigualdade, mais / menos) e podemos adicionar e/ou subtrair os valores, como “20° C é 10 graus mais quente do que 10° C” ou “6 da tarde é 3 horas depois do que 15:00”. No entanto, não podemos multiplicar ou dividir os números por causa do zero arbitrário, então nós não podemos dizer que "20° C é duas vezes mais quente do que 10 °C" ou "6hs da tarde é duas vezes mais tarde que 15h".

O valor central dos dados de intervalo geralmente é a média (mas pode ser a mediana ou a moda), e também podemos expressar a propagação ou variabilidade dos dados usando medidas como intervalo, desvio padrão, variação e / ou intervalos de confiança.

2.1.4: Dados de Razão

Os dados de razão são medidos e ordenados como itens equidistantes e possuem um zero significativo. Tal como acontece com os dados intervalares, os dados de razão podem ser contínuos ou discretos e diferem dos dados do intervalo, na medida em que existe um ponto zero não arbitrário para os dados. As características dos dados de razão são mostradas na Figura 2.4.

Figura 2.4: Características dos Dados de Razão

Exemplos incluem:

- Idade (de 0 a 100 anos)
- Temperatura (em Kelvin, mas não ° C ou F)
- Distância (medida com uma régua ou outro dispositivo de medição)
- Intervalo de tempo (medido com um cronômetro ou similar)

Para cada um desses exemplos, existe um ponto zero real e significativo - a idade de uma pessoa (um filho de 12 anos é o dobro do de idade de 6 anos), zero absoluto (matéria a 200K tem o dobro da energia da matéria a 100K), a distância medida a partir de um ponto pré-determinado (a distância de Barcelona a Berlim é metade da distância entre Barcelona e Moscou) ou tempo (demoro duas vezes mais tempo para correr os 100 metros como Usain Bolt, mas apenas metade do tempo do meu avô).

Os dados de razão são os melhores para lidar matematicamente (note que eu não disse mais fácil ...) porque todas as possibilidades estão na mesa. Podemos encontrar o ponto central dos dados usando qualquer técnica: moda, mediana ou média (aritmética, geométrica ou harmônica) e utilizar também todos os métodos estatísticos mais poderosos para analisar os dados.

Desde que escolhidos corretamente, podemos realmente ficar confiantes de que não estamos sendo enganados pelos dados e que nossas interpretações provavelmente terão mérito.

Portanto, isto conclui a primeira parte deste livro.

Até agora, você coletou, limpou, pré-processou e classificou seus dados. É um bom começo, mas ainda há muito o que fazer. A próxima seção trata de construir uma caixa de ferramentas estatística que você pode usar para extrair informações reais e significativas de seu novo e brilhante conjunto de dados.