when to keep outliers

القيم المتطرفة هي قيم غير عادية في مجموعة البيانات الخاصة بك ، ويمكن أن تشوه التحليلات الإحصائية وتنتهك افتراضاتها. لسوء الحظ ، سيواجه جميع المحللين القيم المتطرفة وسيضطرون إلى اتخاذ قرارات بشأن ما يجب فعله معهم. نظرًا للمشكلات التي يمكن أن تسببها ، قد تعتقد أنه من الأفضل إزالتها من بياناتك. لكن هذا ليس هو الحال دائمًا. إزالة القيم المتطرفة أمر مشروع فقط لأسباب محددة.

رسم بياني يعرض الخارج.يمكن أن تكون القيم المتطرفة مفيدة للغاية حول مجال الموضوع وعملية جمع البيانات. من الضروري فهم كيفية حدوث القيم المتطرفة وما إذا كان من الممكن حدوثها مرة أخرى كجزء طبيعي من العملية أو منطقة الدراسة. لسوء الحظ ، قد تكون مقاومة إغراء إزالة القيم المتطرفة بشكل غير لائق أمرًا صعبًا. القيم المتطرفة تزيد من التباين في بياناتك ، مما يقلل من القوة الإحصائية . وبالتالي ، يمكن أن يتسبب استبعاد القيم المتطرفة في أن تصبح نتائجك ذات دلالة إحصائية.

في رسالتي السابقة ، عرضت خمس طرق يمكنك استخدامها لتحديد القيم المتطرفة . ومع ذلك ، فإن تحديد الهوية هو مجرد خطوة أولى. يعتمد تحديد كيفية التعامل مع القيم المتطرفة على التحقيق في السبب الكامن وراءها.

في هذا المنشور ، سأساعدك في تحديد ما إذا كان يجب عليك إزالة القيم المتطرفة من مجموعة البيانات الخاصة بك وكيفية تحليل بياناتك عندما لا يمكنك إزالتها. يعتمد الإجراء المناسب على أسباب القيم المتطرفة. في السكتات الدماغية العامة ، هناك ثلاثة أسباب للقيم المتطرفة - إدخال البيانات أو أخطاء القياس ، ومشاكل أخذ العينات والظروف غير العادية ، والتباين الطبيعي.

لنستعرض هذه الأسباب الثلاثة!

أخطاء إدخال البيانات والقياس والقيم المتطرفة

يمكن أن تحدث أخطاء أثناء القياس وإدخال البيانات. أثناء إدخال البيانات ، يمكن أن تنتج الأخطاء المطبعية قيمًا غريبة. تخيل أننا نقيس ارتفاع الرجال البالغين ونجمع مجموعة البيانات التالية.

في مجموعة البيانات هذه ، من الواضح أن قيمة 10.8135 هي قيمة شاذة . إنها لا تبرز فحسب ، بل إنها قيمة ارتفاع مستحيلة. عند فحص الأرقام عن كثب ، نستنتج أن الصفر ربما كان عرضيًا. نأمل أن نتمكن من العودة إلى السجل الأصلي أو حتى إعادة قياس الموضوع لتحديد الارتفاع الصحيح.

هذه الأنواع من الأخطاء هي حالات يسهل فهمها. إذا حددت أن القيمة الخارجة هي خطأ ، فصحح القيمة عندما يكون ذلك ممكنًا. يمكن أن يتضمن ذلك إصلاح الخطأ المطبعي أو ربما إعادة قياس العنصر أو الشخص. إذا لم يكن ذلك ممكنًا ، فيجب حذف نقطة البيانات لأنك تعلم أنها قيمة غير صحيحة.

يمكن أن تسبب مشاكل أخذ العينات القيم المتطرفة

تستخدم الإحصائيات الاستدلالية عينات لاستخلاص استنتاجات حول مجموعة سكانية معينة . يجب أن تحدد الدراسات مجموعة سكانية بعناية ، ثم تسحب عينة عشوائية منها على وجه التحديد. هذه هي العملية التي يمكن للدراسة من خلالها التعرف على السكان.

لسوء الحظ ، قد تحصل دراستك عن طريق الخطأ على عنصر أو شخص ليس من السكان المستهدفين. هناك عدة طرق يمكن أن يحدث هذا. على سبيل المثال ، يمكن أن تحدث أحداث أو خصائص غير معتادة تنحرف عن المجموعة السكانية المحددة. ربما يقوم المجرب بقياس العنصر أو الموضوع في ظل ظروف غير طبيعية. في حالات أخرى ، يمكنك عن طريق الخطأ جمع عنصر يقع خارج المجموعة المستهدفة ، وبالتالي ، قد يكون له خصائص غير عادية.

وظيفة ذات صلة : الاستنتاج مقابل الإحصاء الوصفي

أمثلة على مشاكل أخذ العينات

دعونا نجعل هذا ينبض بالحياة مع العديد من الأمثلة!

افترض أن الدراسة تقيم قوة المنتج. يعرف الباحثون السكان على أنهم ناتج عملية التصنيع القياسية. تتضمن العملية العادية المواد القياسية وإعدادات التصنيع والظروف. إذا حدث شيء غير عادي أثناء جزء من الدراسة ، مثل انقطاع التيار الكهربائي أو انحراف الجهاز عن القيمة القياسية ، فقد يؤثر ذلك على المنتجات. يمكن أن تسبب ظروف التصنيع غير الطبيعية هذه القيم المتطرفة من خلال إنشاء منتجات ذات قيم قوة غير نمطية. لا تعكس المنتجات المصنعة في ظل هذه الظروف غير العادية المجموعة المستهدفة من المنتجات من العملية العادية. وبالتالي ، يمكنك إزالة نقاط البيانات هذه بشكل شرعي من مجموعة البيانات الخاصة بك.

صورة الأشعة السينية للساقين.خلال دراسة لكثافة العظام التي شاركت فيها كعالم ، لاحظت تباينًا في نمو كثافة العظام لموضوع ما. كانت قيمة نموها غير عادية للغاية. اكتشف منسق موضوع الدراسة أن المريض يعاني من مرض السكري ، مما يؤثر على صحة العظام. كان هدف دراستنا هو نمذجة نمو كثافة العظام لدى الفتيات قبل سن المراهقة مع عدم وجود ظروف صحية تؤثر على نمو العظام. وبالتالي ، تم استبعاد بياناتها من تحليلاتنا لأنها لم تكن عضوًا في المجموعة السكانية المستهدفة.

إذا كان بإمكانك إثبات أن عنصرًا أو شخصًا لا يمثل السكان المستهدفين ، فيمكنك إزالة نقطة البيانات هذه. ومع ذلك ، يجب أن تكون قادرًا على عزو سبب أو سبب معين لعدم ملاءمة هذا العنصر النموذجي للسكان المستهدفين.

يمكن أن ينتج التباين الطبيعي القيم المتطرفة

الأسباب السابقة للقيم المتطرفة أشياء سيئة. إنها تمثل أنواعًا مختلفة من المشكلات التي تحتاج إلى تصحيحها. ومع ذلك ، يمكن للتنوع الطبيعي أيضًا أن ينتج قيمًا متطرفة - وليست بالضرورة مشكلة.

لإيجاد القيم المتطرفة.جميع توزيعات البيانات لها انتشار من القيم. يمكن أن تحدث القيم المتطرفة ، لكن Z توزيع درجات احتمالاتها أقل. إذا كان حجم عينتك كبيرًا بما يكفي ، فأنت ملزم بالحصول على قيم غير عادية. في التوزيع الطبيعي ، سيكون ما يقرب من 1 من كل 340 ملاحظة على الأقل ثلاثة انحرافات معيارية عن المتوسط. ومع ذلك ، قد تتضمن الفرصة العشوائية قيمًا متطرفة في مجموعات بيانات أصغر! بمعنى آخر ، قد ينتج عن العملية أو المجتمع الذي تدرسه قيمًا غريبة بشكل طبيعي. لا حرج في نقاط البيانات هذه. إنها غير عادية ، لكنها جزء طبيعي من توزيع البيانات.

الوظيفة ذات الصلة : التوزيع الطبيعي ومقاييس التباين

مثال على التباين الطبيعي الذي يسبب الانحراف

صورة لجريدة ترومان ممسكة.على سبيل المثال ، أنا أوافق نموذجًا يستخدم تقييمات الموافقة الرئاسية الأمريكية التاريخية للتنبؤ بمدى تصنيف المؤرخين اللاحقين لكل رئيس في النهاية. اتضح أن أدنى معدل موافقة للرئيس يتنبأ برتب المؤرخ. ومع ذلك ، هناك نقطة بيانات واحدة تؤثر بشدة على النموذج. الرئيس ترومان لا يناسب النموذج. كان لديه أدنى معدل موافقة سيئ بنسبة 22 ٪ ، لكن المؤرخين في وقت لاحق أعطاه مرتبة جيدة نسبيًا من المرتبة السادسة. إذا قمت بإزالة تلك الملاحظة الفردية ، فإن مربع R يزيد بأكثر من 30 نقطة مئوية!

ومع ذلك ، لا يوجد سبب مبرر لإزالة هذه النقطة. في حين أنها كانت غريبة ، إلا أنها تعكس بدقة المفاجآت المحتملة وعدم اليقين المتأصل في النظام السياسي. إذا قمت بإزالته ، فإن النموذج يجعل العملية تبدو أكثر قابلية للتنبؤ مما هي عليه في الواقع.

على الرغم من أن هذه الملاحظة غير العادية مؤثرة ، إلا أنني تركتها في النموذج. من الممارسات السيئة إزالة نقاط البيانات ببساطة لإنتاج نموذج ملائم أفضل أو نتائج ذات دلالة إحصائية.

إذا كانت القيمة المتطرفة عبارة عن ملاحظة مشروعة تمثل جزءًا طبيعيًا من السكان الذين تدرسهم ، فيجب أن تتركها في مجموعة البيانات! سأشرح كيفية تحليل مجموعات البيانات التي تحتوي على القيم المتطرفة التي لا يمكنك استبعادها قريبًا!

لمعرفة المزيد حول المثال أعلاه ، اقرأ مقالتي حوله ، فهم تصنيفات المؤرخين لرؤساء الولايات المتحدة باستخدام نماذج الانحدار .

إرشادات للتعامل مع القيم المتطرفة

من الأفضل أحيانًا الاحتفاظ بالقيم المتطرفة في بياناتك. يمكنهم الحصول على معلومات قيمة تشكل جزءًا من منطقة الدراسة الخاصة بك. قد يكون الاحتفاظ بهذه النقاط صعبًا ، خاصةً عندما يقلل الدلالة الإحصائية! ومع ذلك ، فإن استبعاد القيم المتطرفة فقط بسبب حدها الأقصى يمكن أن يشوه النتائج عن طريق إزالة المعلومات حول التباين المتأصل في منطقة الدراسة. أنت تجبر منطقة الموضوع على الظهور بشكل أقل تباينًا مما هي عليه في الواقع.

عند التفكير فيما إذا كنت تريد إزالة استثناء ، ستحتاج إلى تقييم ما إذا كان يعكس بشكل مناسب السكان المستهدفين ، ومجال الموضوع ، وسؤال البحث ، ومنهجية البحث. هل حدث أي شيء غير عادي أثناء قياس هذه الملاحظات ، مثل انقطاع التيار الكهربائي ، أو الظروف التجريبية غير الطبيعية ، أو أي شيء آخر خارج عن القاعدة؟ هل هناك أي شيء يختلف اختلافًا جوهريًا في الملاحظة ، سواء كانت شخصًا أو عنصرًا أو معاملة؟ هل حدثت أخطاء في القياس أو إدخال البيانات؟

إذا كان الخارج في السؤال هو:

خطأ في القياس أو خطأ في إدخال البيانات ، قم بتصحيح الخطأ إن أمكن. إذا لم تتمكن من إصلاحها ، فقم بإزالة هذه الملاحظة لأنك تعلم أنها غير صحيحة.

ليس جزءًا من السكان الذين تدرسهم (على سبيل المثال ، خصائص أو ظروف غير عادية) ، يمكنك إزالة الخارج بشكل شرعي.

جزء طبيعي من السكان الذين تدرسهم ، لا يجب إزالتهم.

عندما تقرر إزالة القيم المتطرفة ، قم بتوثيق نقاط البيانات المستبعدة واشرح أسبابك. يجب أن تكون قادرًا على عزو سبب محدد لإزالة القيم المتطرفة. نهج آخر هو إجراء التحليل مع وبدون هذه الملاحظات ومناقشة الاختلافات. تعد مقارنة النتائج بهذه الطريقة مفيدة بشكل خاص عندما لا تكون متأكدًا من إزالة أحد العناصر الخارجية وعندما يكون هناك خلاف كبير داخل المجموعة حول هذا السؤال.

التحليلات الإحصائية التي يمكنها التعامل مع القيم المتطرفة

ماذا تفعل عندما لا يمكنك إزالة القيم المتطرفة بشكل شرعي ، لكنها تنتهك افتراضات تحليلك الإحصائي؟ تريد تضمينهم ولكن لا تريد منهم تشويه النتائج. لحسن الحظ ، هناك العديد من التحليلات الإحصائية حتى المهمة. فيما يلي العديد من الخيارات التي يمكنك تجربتها.

اختبارات الفرضية اللامعلمية قوية بالنسبة للقيم المتطرفة . بالنسبة لهذه البدائل للاختبارات البارامترية الأكثر شيوعًا ، لن تنتهك القيم المتطرفة بالضرورة افتراضاتها أو تشوه نتائجها.

في تحليل الانحدار ، يمكنك محاولة تحويل بياناتك أو استخدام تحليل انحدار قوي متاح في بعض الحزم الإحصائية.

أخيرًا ، تستخدم تقنيات التمهيد بيانات العينة كما هي ولا تضع افتراضات حول التوزيعات.

تسمح لك هذه الأنواع من التحليلات بالتقاط التباين الكامل لمجموعة البيانات الخاصة بك دون انتهاك الافتراضات والنتائج المشوهة.