

Identificación de Outliers en Encuestas

Introducción¶

En la descripción del dataset "[We are from our childhood](#)" de kaggle.com, se plantean posibles investigaciones que se pueden realizar sobre el dataset...

Una de las investigaciones que nos llamó la atención es "*Detección de Valores anómalos*" (Outliers Detection), ya que el dataset, está basado en **una encuesta** que tuvieron que realizar los amigos de una clase de estadística.

La encuesta es realmente grande, tiene 139 preguntas, en las que hay 5 posibles valores para responder y 11 preguntas más a responder categoricas como la edad, estudios, etc.

Un pequeño(esperemos) número de personas tiene que haber respondido de forma aleatoria.

Nuestra hipótesis se basa en que **Llenar esa encuesta tiene que haber sido una tarea realmente larga y tediosa.** por lo que es posible encontrar a esas personas, y diferenciar entre diferentes tipos de outliers con ayuda de los algoritmos y técnicas usados para detectar valores anómalos.

Outliers o valores atípicos

Los outliers son valores que están alejados de la mayoría de los otros datos.

Son malos porque "ensucian" las muestras, desviando los valores y estadísticas derivadas de las muestras como medias por ejemplo dando lugar a resultados engañosos o de poca calidad, sobre todos cuando se tienen pocos datos.

Causas: Hay varias causas para la aparición de valores anómalos. Algunos son datos erróneos como, errores en la medición, errores en la transcripción. Otros son datos verdaderos, que simplemente son desviaciones naturales en una población, o muestras que pertenecen a una población diferente.

En este caso los outliers a identificar son aquellas personas que para completar la encuesta pusieron valores al azar con el afán de completarla más rápido.

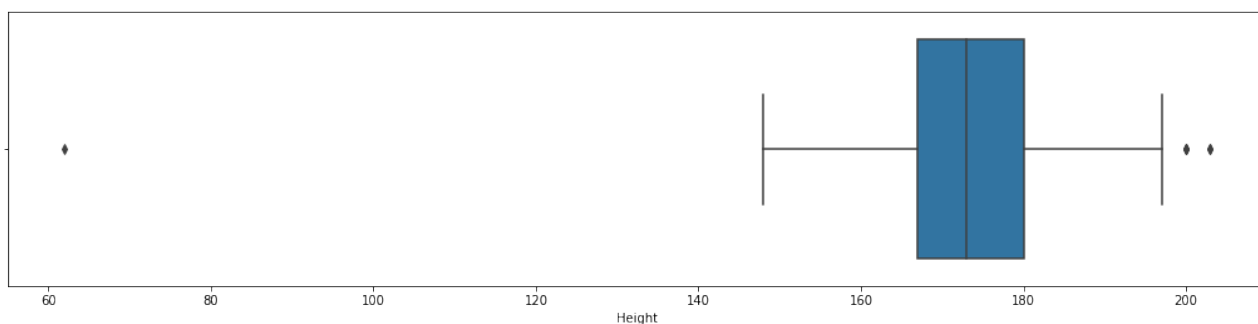
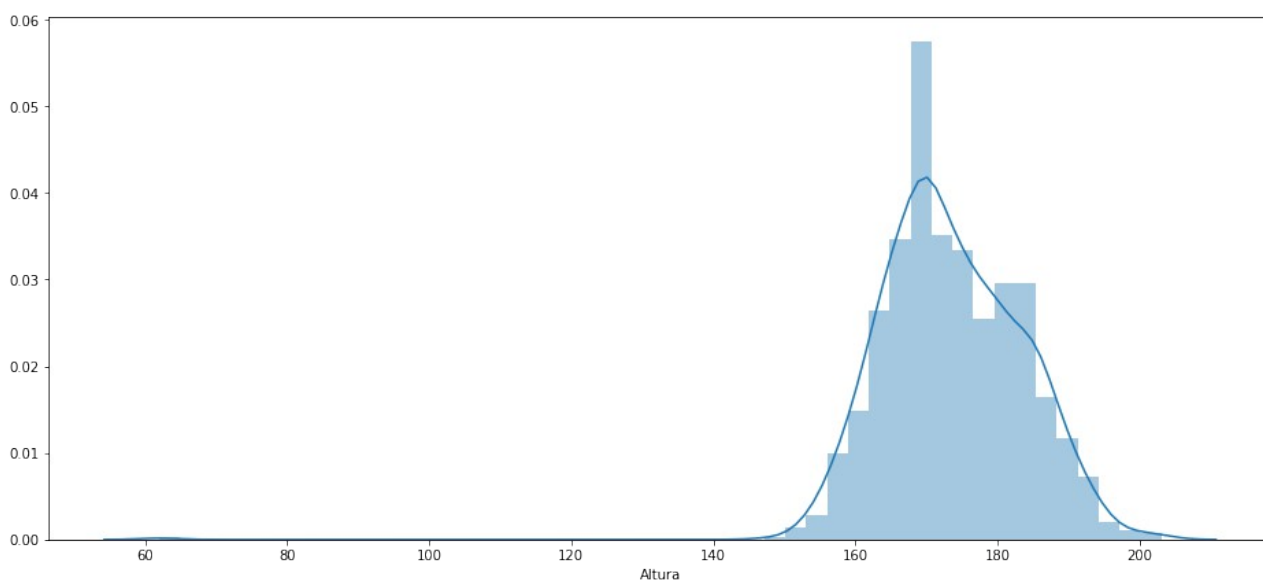
Outliers en una Dimension vs Outliers globales.

Esto es la clave para nuestro análisis. Ya que un valor puede ser muy atípico para una dimensión, pero eso no quiere decir que sea una encuesta falsa. Nuestro objetivo es encontrar encuestas que se alejen en varias dimensiones.

Por ejemplo una persona excepcionalmente alta, pero que en las demás dimensiones sigue a la media no es nuestro objetivo, esa persona es un outlier para esa sola dimension, pero no nos asegura que haya mentido al realizar la encuesta.

Nuestro objetivo son aquellas personas que en varias variables se alejen de la muestra, aunque lo hagan en pequeñas proporciones.

Análisis de outliers en la dimension altura.



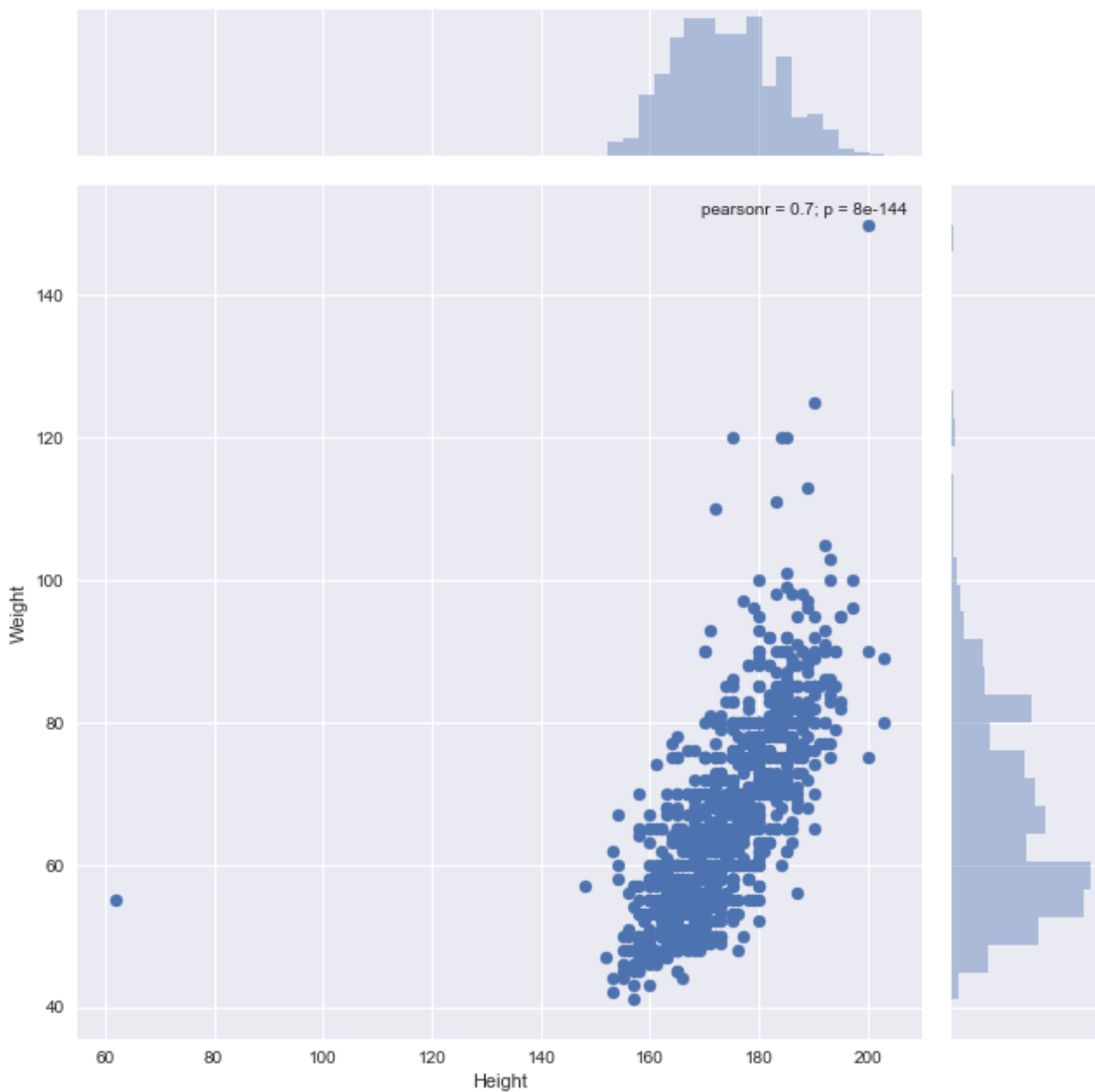
El diagrama de caja y bigotes nos permiten ver los valores atípicos (puntos) fuera de los bigotes (brazos) del gráfico.

En esta caso observamos tres outliers, uno que esta cerca de los 60 cm de altura y otros dos cercanos a los 200 centímetros de altura.

Esos puntos **son outliers**, pero ¿**pertenecen a desviaciones naturales de la población de datos, a un error de cargar o transcripción, o a alguien que respondió la encuesta con valores aleatorios?**

Para eso vamos a tener que analizar más a fondo ese valor atípico.

Para descartar o no una desviación natural de la población. Que en este caso podría ser que alguien sufra de enanismo o sea un niño pequeño, vamos a comparar dos dimensiones al mismo tiempo, peso y altura. Ya que alguien que mida solo 60cm debe pesar menos que las personas que miden como la media.



En el gráfico se aprecia que el punto de la izquierda el outlier que mide un poco más que 60 cm, no es un outlier para la variable peso, el punto se encuentra casi en la media de la distribución de los pesos. Por lo tanto, **podemos descartar que sea un desviación natural.**

Queda averiguar si es un valor atípico producto de un error de transcripción/medición o si es un valor atípico producto de alguien que respondió azarosamente.

Para esto debemos usar algún algoritmo de detección de outliers que nos ayude a ver que tan alejado del resto de los valores está ese sujeto pero tomando en cuenta todas las dimensiones.

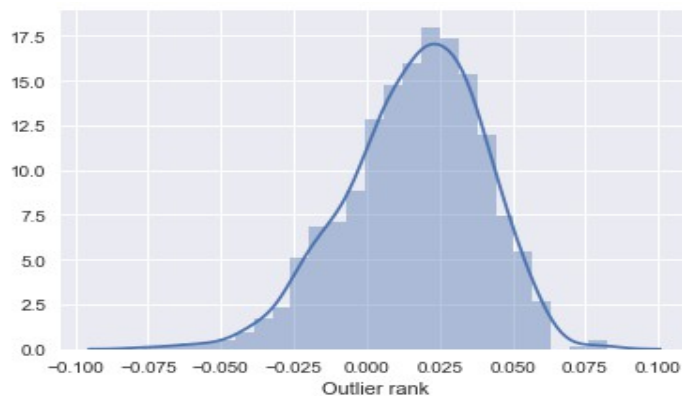
Métodos de Detección de Outliers¶

Hay varios métodos para la detección de outliers. Los aspectos más determinantes a la hora de elegir un método son la distribución que los datos y la cantidad de dimensiones.

En nuestro caso como tenemos 150 dimensiones y asumimos que siguen una distribución normal. Tenemos que usar o el método de **Isolation Forest** o **Local Outlier Factor**

Usamos el método **Isolation Forest**, ya que la implementación de este algoritmo nos devuelve un “ranking” de que tan outlier es una medición.

Ese ranking tiene una distribución normal también.



Si analizamos el valor que nos devolvió Isolation forest para el sujeto de 62 cm, nos da un valor de 0.038994. Por lo que es muy poco probable que sea un outlier global.

Con esto podemos asegurar que el **outlier es un error de carga/transcripción y no una persona que respondió todo al azar.**

Veamos que pasa con el outlier de la persona más alta.

Si observamos nuevamente el grafico que tiene en cuenta las dimensiones altura y peso, vemos que si bien es un outlier en la dimension altura, ya que mide un poco más de 200 cm, no lo es en la dimencion peso. Por lo que en este caso, si puede deberse a una desviacion natural, a una persona que es muy alta, pero es un dato verdadero. Para sacarnos la duda analizamos su ranking de outlier global y nos da un resultado de 0.002005 esto qvalor sigue siendo muy alto y no deberia ser un outlier global. Por lo que **podemos inferir que se trata de un outlier para la dimension altura, pero se trata de una desviacion natural**, simplemente es alguien muy alto, pero con gustos parecidos al resto.

Para encontrar las personas que podrian haber respondido varias preguntas al azar, tenemos que fijarnos solamente en los **outliers globales** y ver cuales tienen peor ranking.

En este caso si analizamos los valores del sujeto que tiene peor ranking y es seguramente un outlier global, no quiere decir que siempre sea un outlier local tambien, puede ser que en la mayoria se aleje de la media de los valores, pero no necesariamente sea un outlier en todas las dimensiones.

Conclusion.

Usando los algoritmos debidamente y haciendo un analisis entre distintas variables, relacionadas, es posible identificar entre los diferentes tipos de valores atípicos. Es este caso identificamos;
Un valor atipico causado por un dato erróneo, como el caso de la persona que media 62cm.
Un valor atipico causado simplemente por desviaciones naturales, la persona alta.
Y por medio del algoritmo personas que seguramente contestaron la encuesta poniendo valores al azar, en este caso son atípicas, pero tomando en cuenta todas las dimensiones a la vez.