ACÁMICA

¡Bienvenidos/as a Data Science!

Agenda

¿Cómo anduvieron?

Repaso: Outliers

Hands on training

Break

Sabías que

Encuesta

Cierre

¿Cómo anduvieron?

Repaso: Outliers

SUGERENCIA

Siempre es importante pensar por qué hay un outlier en nuestro dataset

¿Qué es un Outlier? ¿Por qué ocurren?

OUTLIER = valor atípico que difiere significativamente del resto de las observaciones.

¿Por qué difiere?

- Error de medición del instrumento.
- Error al introducir un dato.
- Estamos trabajando con muestras/poblaciones que no son tan homogéneas como creíamos.

¿Qué es un Outlier? ¿Por qué ocurren?

¡Muchas veces los OUTLIERS son una manifestación del proceso que estamos estudiando!

Ejemplos:

- Transacción fraudulenta con una tarjeta de crédito.
- Persona enferma en un conjunto de personas sanas.
- Mayor incidencia de una enfermedad en una ciudad. ¿Esperable o outlier?

¿Valores Atípicos vs. Valores Extremos

Valor Extremo = valor distante del resto de las observaciones pero comprendido dentro de los valores esperados en mi distribución.

En general, son más comunes en distribuciones con alta curtosis.

Valores Atípicos vs. Valores Extremos

Valor Extremo

¿Debemos sacarlos? ¿Los podemos considerar outliers?
Eso dependerá del problema que estemos estudiando.

Pero es importante recalcar la diferencia entre:

un valor atípico porque estamos mezclando poblaciones (transacciones no-fraudulentas/fraudulentas, personas sanas/no-sanas, etc.)

valores extremos de una población homogénea (Ejemplo: usamos siempre la tarjeta de crédito para hacer compras pequeñas y un día compramos un pasaje en avión).

Hands-on training

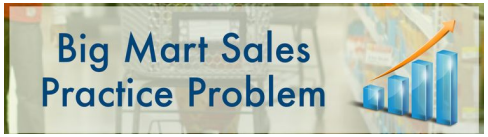
Escenario

Escenario

Los ci para 1559 productos en 10 tiendas en diferentes ciudades. Además, se han definido ciertos atributos de cada producto y tienda.

En el análisis de este conjunto de datos, BigMart intentará comprender las propiedades de los productos y las tiendas que juegan un papel clave en el aumento de las ventas.

Algunos Metadatos

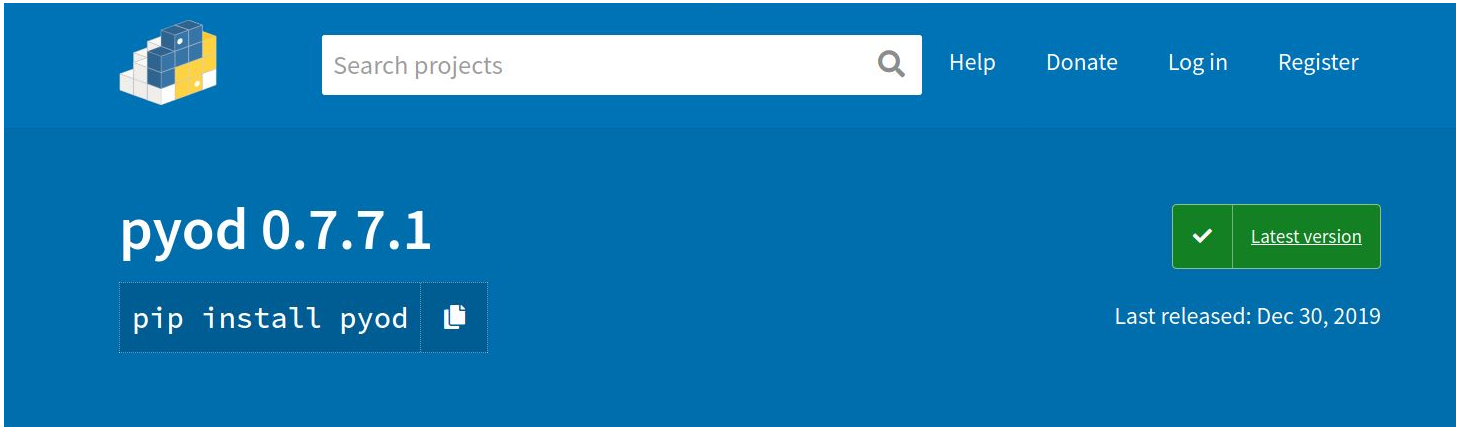


Variable	Descripción
Item_Identifier	Unique product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or not
Item_Visibility	The % of total display area of all products in a store allocated to the particular product.
Item_Type	The category to which the product belongs
Item_MRP	Maximum Retail Price (list price) of the product
Outlet_Identifier	Unique store ID
Outlet_Establishment_	The year in which store was established
Year	The size of the store in terms of ground area covered
Outlet_Size	The type of city in which the store is located
Outlet_Location_Type	The type of city in which the store is located
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket
Item_Outlet_Sales	Sales of the product in the particular store. This is the outcome variable to be predicted.

Librería

PyOD

A Python Toolkit for Scalable Outlier Detection (Anomaly Detection)



Hands-on training

DS_Clase_14_Outliers.ipynb

¡BREAK!

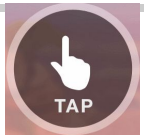
Ph. Credit: Drew Coffmann

¿Sabías cuál es el problema de las ciudades chicas?

Educación

Pruebas Aprender: en Provincia, los mejores rendimientos los tienen los alumnos de las ciudades chicas y del interior

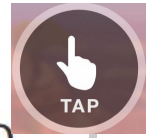
Clarín accedió a los resultados desagregados por municipios. Los distritos del sur y sureste de la Provincia están al tope de la lista, tanto en Lengua como en Matemática. Los más pobres del GBA quedaron relegados.



Santa Fe

Clarín en San Jorge, el pueblo donde los suicidios triplican el promedio de la Argentina

El drama afecta sobre todo a jóvenes y adolescentes. Lo atribuyen a varios factores y ya declararon la emergencia social.

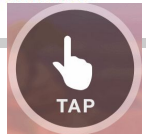


SOCIEDAD

28/04/2015 AGROTÓXICOS

Un pueblo de Entre Ríos en alerta: casi la mitad de su población muere por cáncer

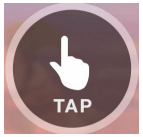
La fuerte movilización de habitantes que denuncian que casi la mitad de su población muere por cánceres generados presumiblemente por los agrotóxicos, motivó que el municipio entrerriano de San Salvador convocara a especialistas de las universidades de Rosario y de La Plata para realizar un estudio epidemiológico-ambiental, cuya primera etapa se cumplió la semana pasada, con científicos encuestando vecinos casa por casa y tomando muestras de aire, tierra y agua.

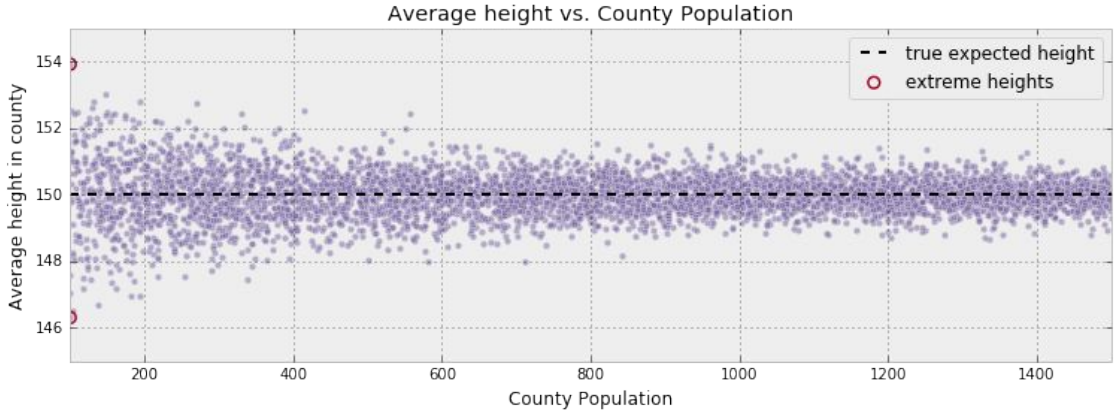


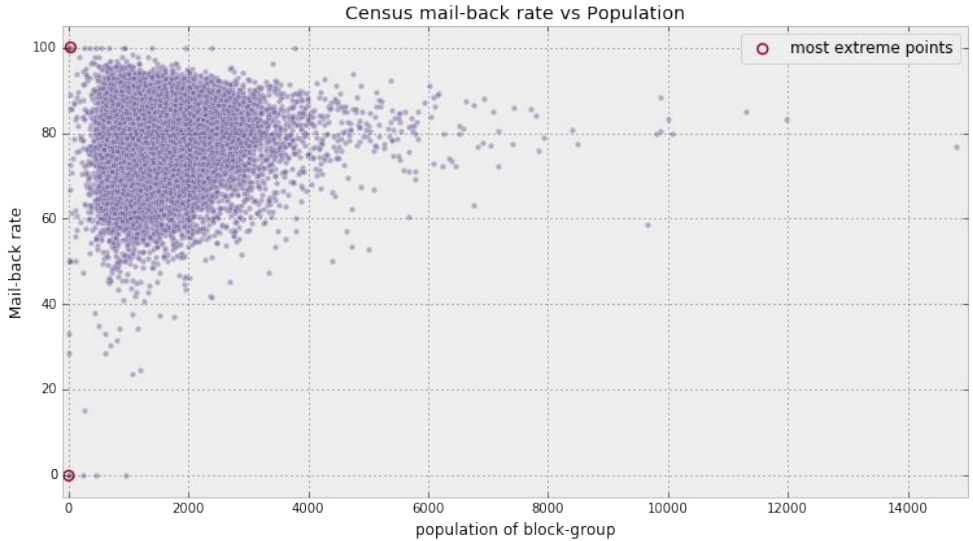
En los ejemplos que veremos a continuación, cabe preguntarnos…

¿Estamos en presencia de fenómenos “reales” o se tratan de desviaciones estadísticas esperables?

Ejemplo 1


TAP


Average height vs. County Population
true expected height
extreme heights
Average height in county
154
152
150
148
146
200
400
600
800
1000
1200
1400
County Population

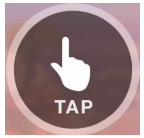

Census mail-back rate vs Population
most extreme points
Mail-back rate
100
80
60
40
20
0
0
2000
4000
6000
8000
10000
12000
14000
population of block-group

Ejemplo 2

The Most Dangerous Equation

Ignorance of how sample size affects statistical variation has created havoc for nearly a millennium

Howard Wainer



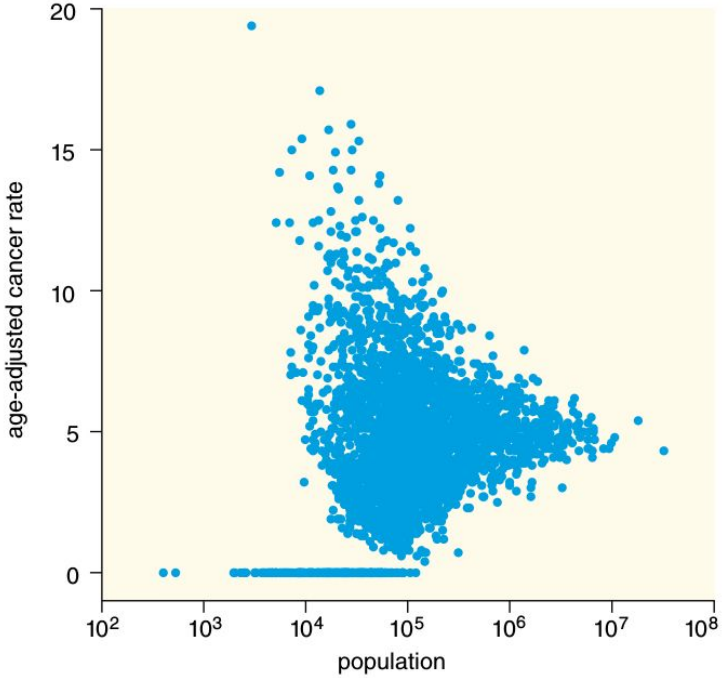
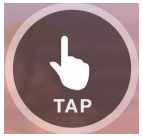


Figure 3. When age-adjusted kidney-cancer rates in U.S. counties are plotted against the log of county population, the reduction of variation with population becomes obvious. This is the typical triangle-shaped bivariate distribution.

[1] Este artículo es muy interesante, pero uno de los ejemplos es discutible.

Ejemplo 3



The most dangerous hospital or the most dangerous equation?

Yu-Kang Tu & Mark S Gilthorpe

BMC Health Services Research 7, Article number: 185 (2007) | Download Citation

5556 Accesses | 8 Citations | 2 Altmetric | Metrics »

Results

A close examination of the information reveals a pattern which is consistent with a statistical phenomenon, discovered by the French mathematician de Moivre nearly 300 years ago, described in every introductory statistics textbook: namely that variation in performance indicators is expected to be greater in small Trusts and smaller in large Trusts. From a statistical viewpoint, the number of deaths in a hospital is not in proportion to the size of the hospital, but is proportional to the square root of its size. Therefore, it is not surprising to note that small hospitals are more likely to occur at the top and the bottom of league tables, whilst mortality rates are independent of hospital sizes.

Conclusion

This statistical phenomenon needs to be taken into account in the comparison of hospital Trusts performance, especially with regard to policy decisions.

Entonces...

Cuando tratamos con **muestras pequeñas**, ¡la **varianza es mucho mayor**!

Entonces hay que tener mucho cuidado al afirmar que hay una mayor (o menor) incidencia de un efecto en alguna de esas muestras *por encima de lo esperado*.

En general, para poder decir que existe un fenómeno de este tipo, vamos a tener que:

Y es muy útil, además, una relación causal que pueda explicar el fenómeno que estamos viendo.
Esto dependerá de las características específicas de cada problemática.
En general, son problemáticas multicausales y existe un riesgo de caer en explicaciones simplistas o, directamente, erróneas.

Actividad:
Dudas comunitarias

Entrega 2: Transformación


¡FELICIDADES!

Encuesta

¡Queremos escucharte!

ENCUESTA

Para la próxima

1. Terminar la Entrega 02.

2. ¡Arrancamos con Machine Learning! Ver los videos “Machine Learning: Qué es Machine Learning”

3. Completar notebooks atrasados y, si lo desean, seguir explorando el dataset que eligieron.

ACÁMICA