# ***Outlier’s Detection***

Nota 09-08-2023 avances Tesis Gauto Sand detección de Outliers.

Entre los algoritmos que se probaron están.

Método de desvió estándar

Método de algoritmo KNN (vecinos más cercanos)

Método de LOF (parecido a vecinos más cercanos)

Outlier definición: algún dato que se cree fue creado de otra manera especial distinto a los otros datos y no representa a los datos comunes reales del conjunto de datos, por lo tanto las filas con outlier deben ser borradas o reemplazadas por la media o algún valor representativo.

Antes de detectar Outlier Agrego columnas de auto correlación trimestral,anual, desvío estándar, promedioy predictor Flag.

## **Aplicando el Método de desvió estándar:**

Calculo el promedio y desvío estándar de todas las columnas

Por cada columna comparo si supera 3 Desvíos estándar ya sea por arriba o por debajo del promedio, y todas las columnas que superan +-3 desvíos se quitan

EL shape del dataset inicial es (1734, 34) y al aplicarlo a todas las columnas es en la Salida (1138, 34)

¿No sé si aplicarlo sobre todas las columnas o solo sobre algunas?

## **Fuentes Método Desvío Estándar**

Difícil de citar

<https://www.youtube.com/watch?v=NUoDfjlBT8M> (Universitat Politècnica de València - UPV )

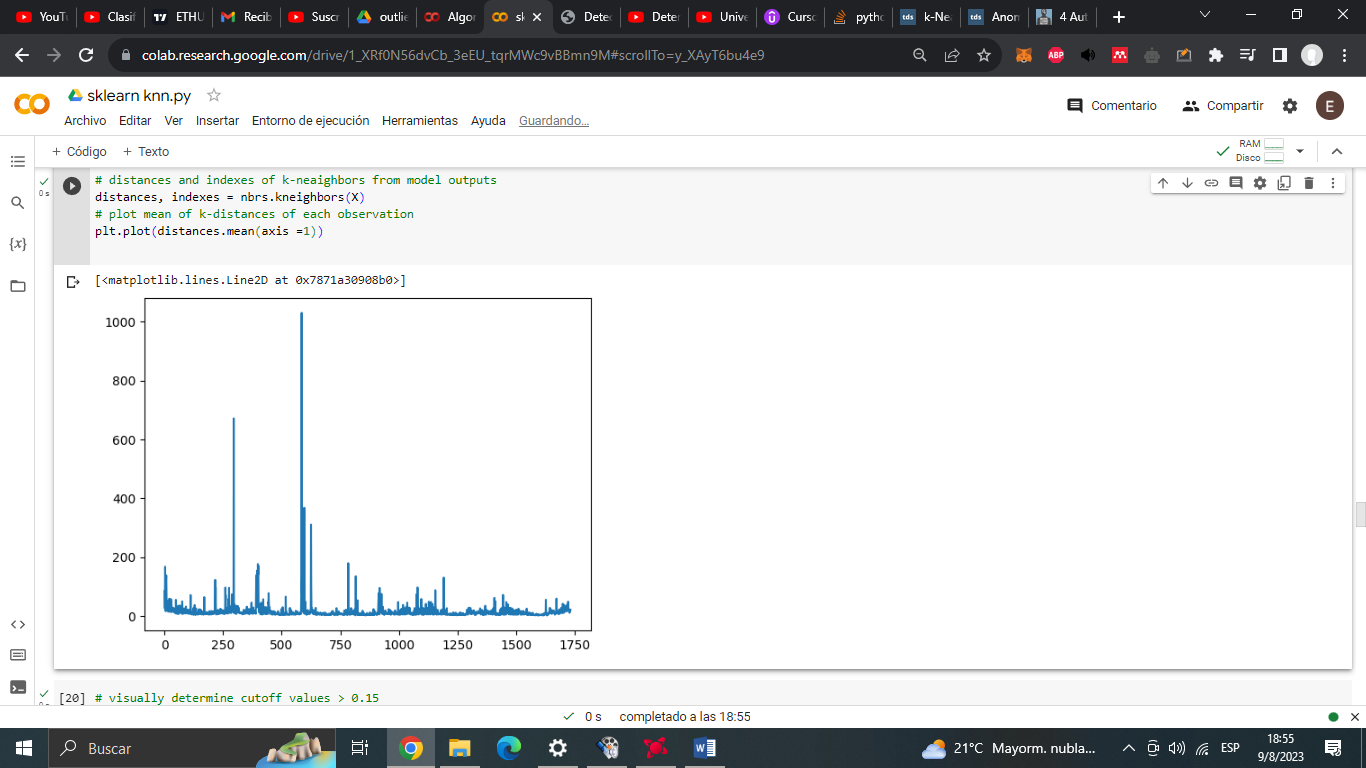
<https://www.sciencedirect.com/science/article/abs/pii/S0022103113000668> (Este paper menciona a D.C. Howell Statistical methods in human sciences(1998) pero no tengo acceso a ese paper, se menciona que en el paper de Howell +-3 desvíos estándar)

## **Aplicando el Método de algoritmo KNN:**

El algoritmo consta de detectar los k vecinos más cercanos a cada punto de dato en cuestión a través de la distancia euclidiana.

Se ejecutó el algoritmo con k=5

Según ví es un algoritmo supervisado, por lo que tengo la duda, según investigue debería de ejecutar un algoritmo de clustering como k-means y luego el knn, usted que opina?



Grafica de promedio de distancias para cada fila del dataset

Procedo a filtrar por lo que en distancia estén por arriba de 200, porque los de arriba de 200 sobresalen mucho.

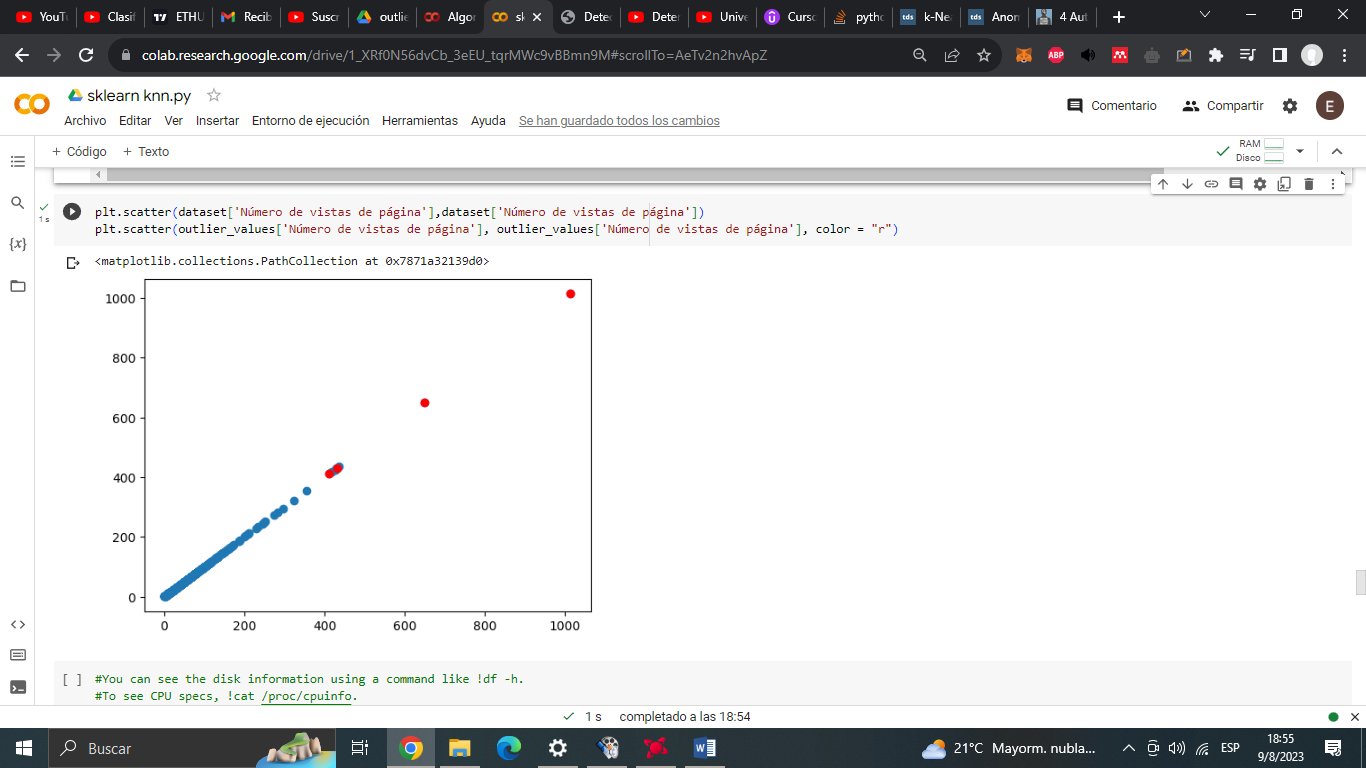


Gráfico de outlier en rojo, normales en azul, eje x e y son el número de vistas de página

## **Fuentes Método de algoritmo KNN:**

<https://www.redalyc.org/journal/5722/572262052006/html/> (usa rapidminer con un a knn especial que no encontré en python)

<https://towardsdatascience.com/k-nearest-neighbors-knn-for-anomaly-detection-fdf8ee160d13> (tutorial de codigo)

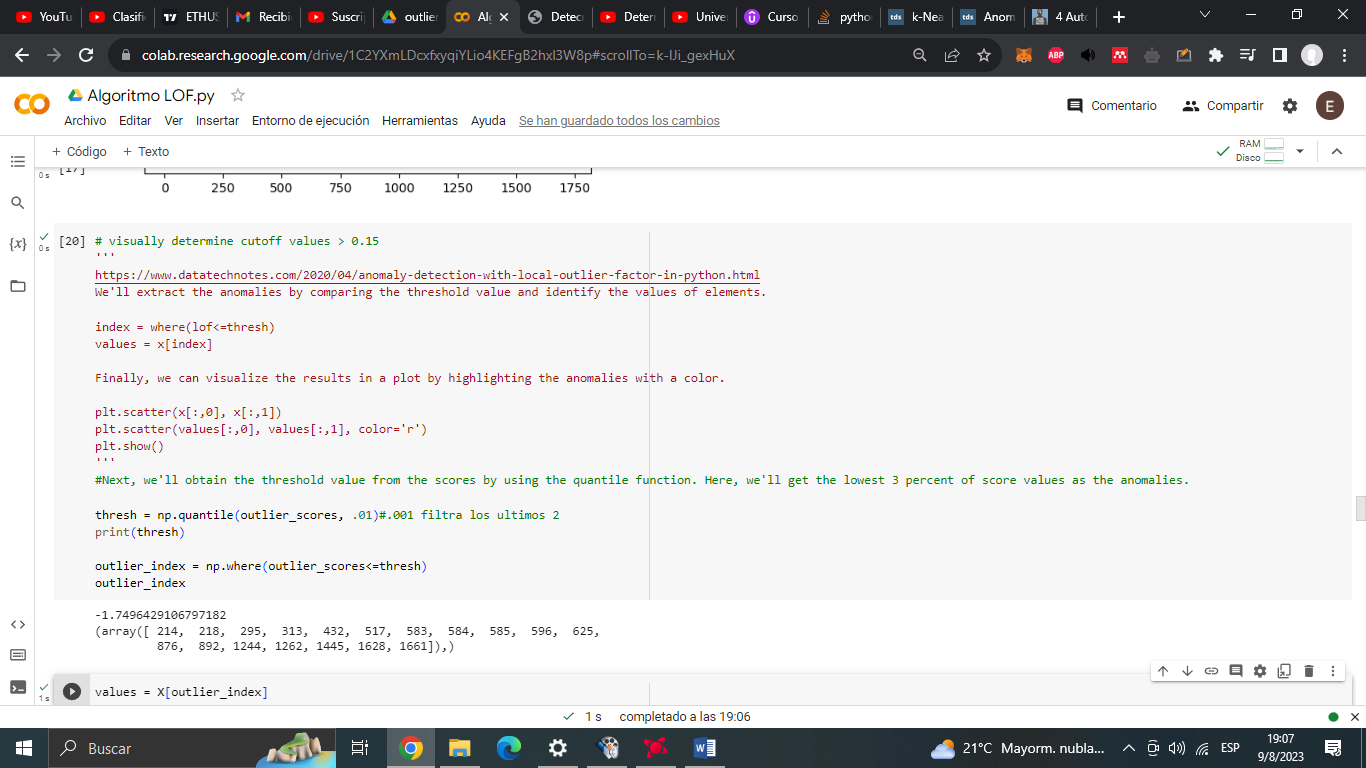
## **Aplicando el Método de algoritmo LOF:**

LOF [5]: This technique detects outliers based on data density. The deviation of a given data with respect to its neighbors is measured by the so called Local Outlier Factor (LOF). This algorithm is unsupervised; hence, it does not require a prior classification of the data

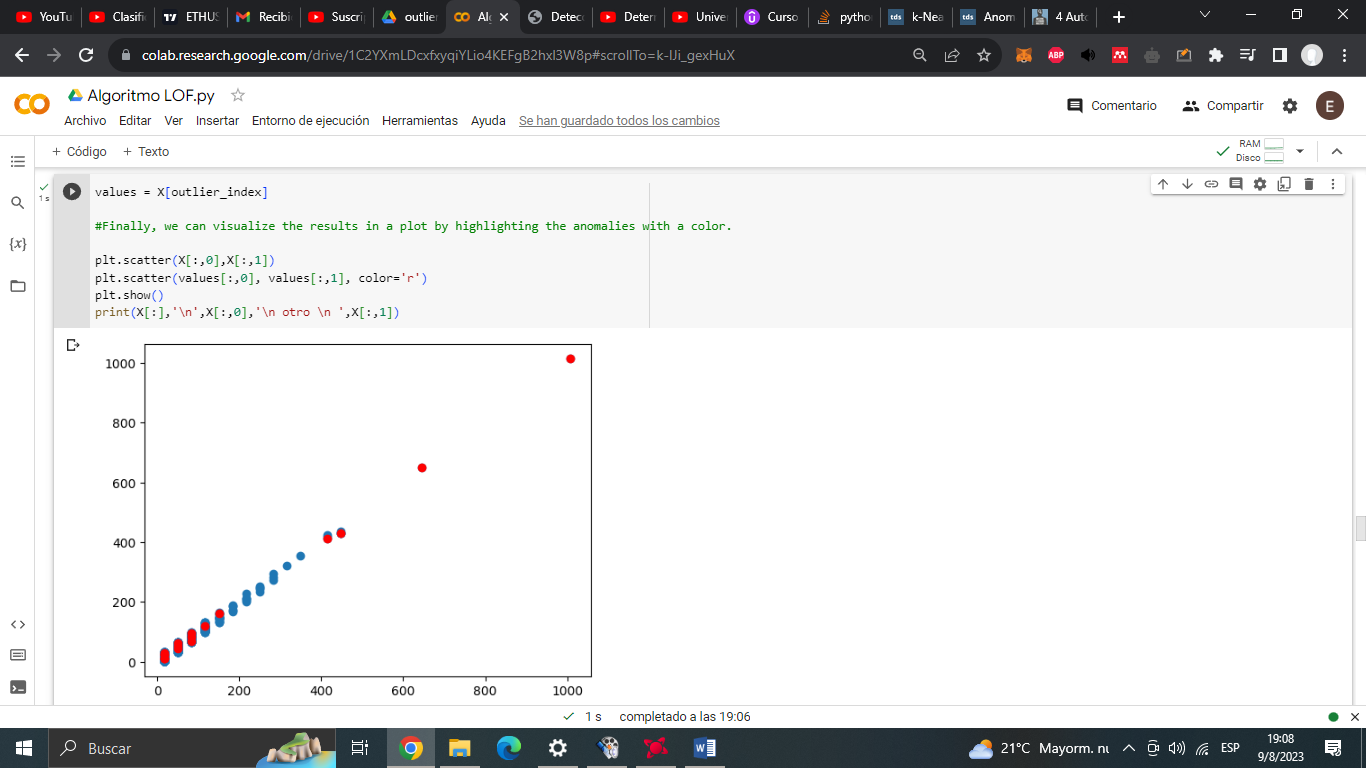
LOF parecido al knn

EL algoritmo lof me genera outlier scores negativos que la documentación oficial dice que ya se consideraría outlier

Con k=5, mismo problema que con knn necesito la cantidad de vecinos



De los scores negativos filtro el cuantil 0.01 o 10% (no tengo justificación para esto es solo ver el gráfico y poner el filtro el codigo origina usaba 0.03=3%, ademas en el paper de Kuna-Alice usan un score 1.2 que seria -1.2 para identificar los outliers, me recomienda usar eso en vez de cuantiles? )



Si pusiera el cuantil 0.001 o 0.1% solo saldrían los 2 de arriba

The opposite LOF of the training samples. The higher, the more normal. Inliers tend to have a LOF score close to 1 (negative\_outlier\_factor\_ close to -1), while outliers tend to have a larger LOF score.

The local outlier factor (LOF) of a sample captures its supposed ‘degree of abnormality’. It is the average of the ratio of the local reachability density of a sample and those of its k-nearest neighbors.

**Fuentes Método de algoritmo LOF:**

Se menciona en los paper

<https://www.datatechnotes.com/2020/04/anomaly-detection-with-local-outlier-factor-in-python.html> (codigo)

<https://www.researchgate.net/publication/265920144_AVANCES_EN_PROCEDIMIENTOS_DE_LA_EXPLOTACION_DE_INFORMACION_PARA_LA_IDENTIFICACION_DE_DATOS_FALTANTES_CON_RUIDO_E_INCONSISTENTES?enrichId=rgreq-ff9126e65410f65f3504d45e49ef35a1-XXX&enrichSource=Y292ZXJQYWdlOzI2NTkyMDE0NDtBUzoxNjczOTAwMzM3NTIwNjRAMTQxNjkyMDMwMjEyNg%3D%3D&el=1_x_3&_esc=publicationCoverPdf>

<https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html>