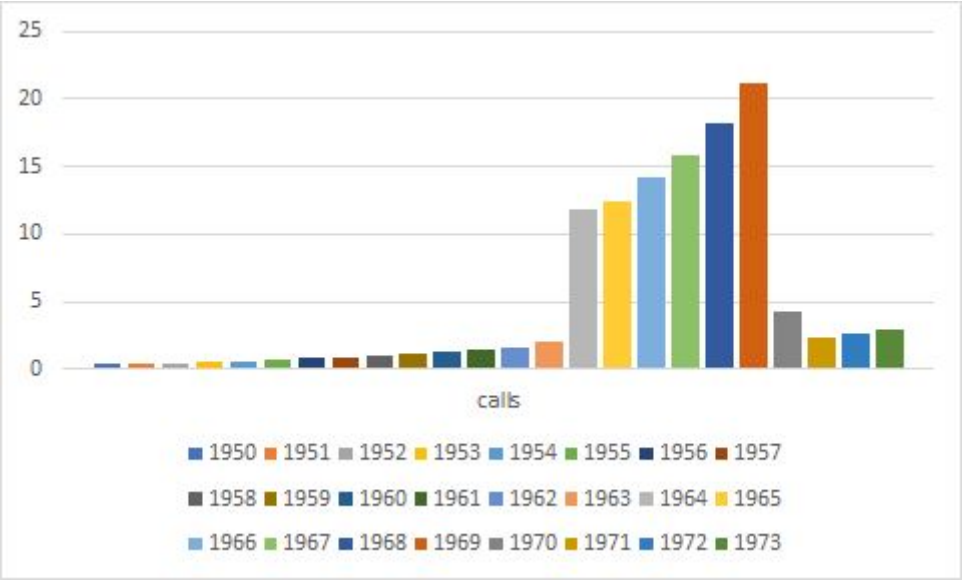
TA3 - Enzo Cozza - Agustín Fernández

Ejercicio 1



Se puede observar que entre 1960 y 1963, la cantidad de llamadas rondaba valores bajos y aumentaba de manera lineal. Luego, en 1964, se produce un gran salto y comienza a crecer de manera exponencial hasta 1970, donde vuelve al nivel presentado hasta 1963.
Los valores entre el 1964 y el 1969 parecen ser outliers del problema planteado.

Luego de buscar información del problema, se descubrió que en los años 1964-1969, el proceso de medición de los mismo fue distinto al del resto. En esos años, lo que se midió no fue la cantidad de llamadas internacionales realizadas, sino la cantidad de minutos de duración de cada una de esas llamadas, debido a eso se produce esta anomalía en los datos.

Se decide eliminar los outliers de este dataset, ya que se pudo comprobar que el error se encuentra en la toma de datos, y no es simplemente una anomalía. Tampoco se quiere reemplazar los datos por unos que se asemejen al resto, ya que se podría estar contaminando el dataset con datos ficticios, cuando todo el resto de los datos sí son datos reales.

Ejercicio 2

Detect Outlier (Distances): se basa en la distancia a su k-ésimo vecino (señalado en number of neighbours), y los valores con las mayores distancias son los calificados como outliers.
Parámetros:
Number of neighbours: número de vecinos a ser analizado.
Number of outliers: número de outliers a ser identificados.
Distance function: euclidian distance, squared distance, cosine distance, inverted cosine distance, angle.

Detect Outlier (LOF): la distancia a los k vecinos más cercanos se utiliza para calcular la densidad de cada dato. Comparando la densidad local con la de los vecinos, se identifican zonas de densidades similares, y los puntos que tienen densidades más bajas que las de sus vecinos son los que se consideran outliers.
Parámetros:
Minimal points to lower bound: cota inferior para el test de outliers.
Minimal points to upper bound: cota superior para el test de outliers.
Distance function

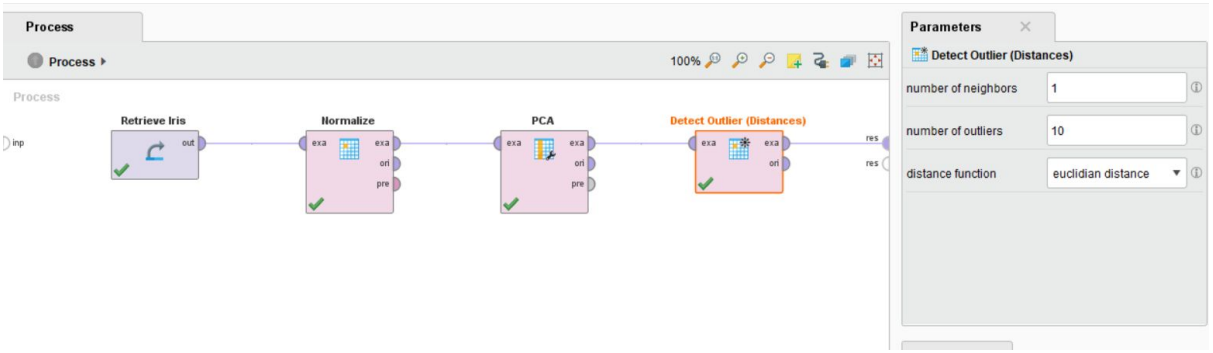
Detect Outlier (Densities): identifica outliers basado en la densidad de los datos. Todos los objetos que tengan por lo menos una proporción (p) de todos los objetos más alejados que una distancia (D), son considerados outliers.
Parámetros:
Distance: Especifica la distancia D mencionada anteriormente.
Proportion: Especifica la proporción p mencionada anteriormente.
Distance function

Detect Outlier (COF): identifica outliers basado en Class Outliers Factor, lo que se basa en clasificar cada instancia dados los parámetros N (valores atípicos superiores de clase N) y K (el número de vecinos más cercanos).
Parámetros:
Number of neighbors
Number of class outliers
Measure types: Selecciona el tipo de medida a utilizar cuando se efectúa la medición de las distancias entre puntos.
Mixed measure: Este parámetro se encuentra disponible cuando 'Measure types' se establece a 'Mixed measures'. La única opción disponible es 'Mixed Euclidean Distance'.

Ejercicio 3

Atributos

Largo del sépalo en cm: Largo de la hoja que forma el cáliz de una flor. Valor real, con rango variando de 4.3 a 7.9, media de 5.84 y desviación estándar 0.83. Cuenta con una distribución similar a la normal.
Ancho del sépalo en cm: Ancho de la hoja que forma el cáliz de una flor. Valor real, con rango de 2.0 a 4.4, media de 3.05 y desviación estándar de 0.43. Cuenta con una distribución normal.
Largo del pétalo en cm: Largo de la hoja que forma la corola de la flor. Valor real, con rango de 1 a 6.9, media de 3.76 y desviación estándar de 1.76. No cuenta con una distribución conocida.
Ancho del pétalo en cm: Ancho de la hoja que forma la corola de la flor. Valor real, con rango de 0.1 a 2.5, media de 1.20 y desviación estándar de 0.76. No cuenta con una distribución conocida.
Clase: Variable objetivo y polinomial, toma el valor de 'Iris Setosa', 'Iris Versicolour' o 'Iris Virginica'.



Outliers obtenidos del dataset:

Row No.	id	label	outlier	pc_1	pc_2
16	id_16	Iris-setosa	true	-0.550	-0.519
23	id_23	Iris-setosa	true	-0.737	-0.095
33	id_33	Iris-setosa	true	-0.672	-0.349
42	id_42	Iris-setosa	true	-0.612	0.410
61	id_61	Iris-versicolor	true	-0.116	0.492
80	id_80	Iris-versicolor	true	-0.070	0.184
88	id_88	Iris-versicolor	true	0.194	0.238
110	id_110	Iris-virginica	true	0.722	-0.333
114	id_114	Iris-virginica	true	0.362	0.243
119	id_119	Iris-virginica	true	0.872	-0.007