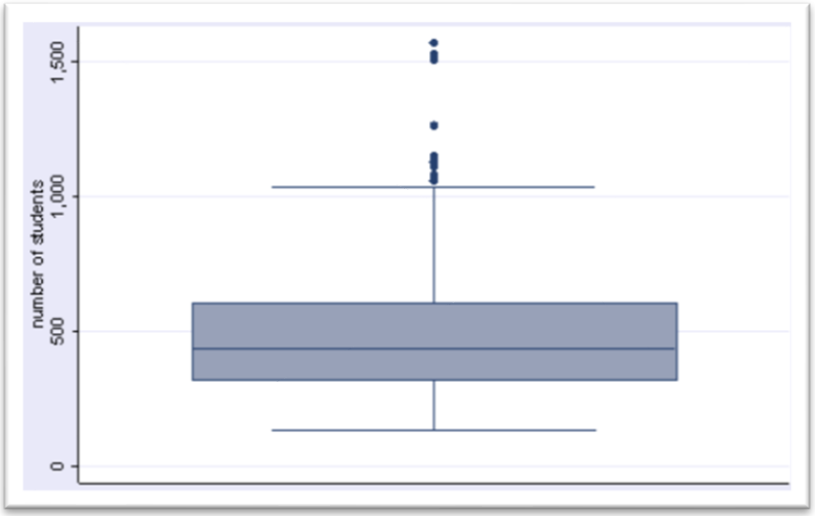
OUTLIERS

MINERIA DE DATOS

Víctor Hugo Cantú Chávez 1806169
Damián Atilano Martínez Alvarado 1735532
Javier Eduardo Salazar Segura 1723152

¿Qué es un outlier?

- Un outlier es una observación que se desvía mucho de otras observaciones y despierta sospechas de ser generada por un mecanismo diferente.



Tipos de outliers

CASOS ATIPICOS

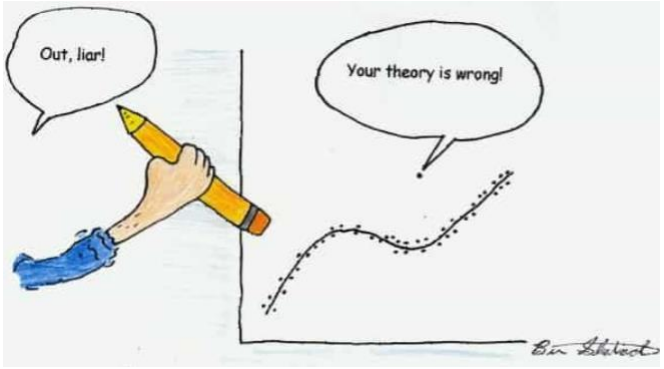
- Surgen de un error de procedimiento, tales como la entrada de datos o un error de codificación. Estos casos atípicos deberían subsanarse en el filtrado de los datos, y si no se puede, deberían eliminarse del análisis o recodificarse como datos ausentes.

OBSERVACION

- Ocurre como consecuencia de un acontecimiento extraordinario. En este caso, el outlier no representa ningún segmento valido de la población y puede ser eliminado del análisis.

Datos extraordinarios

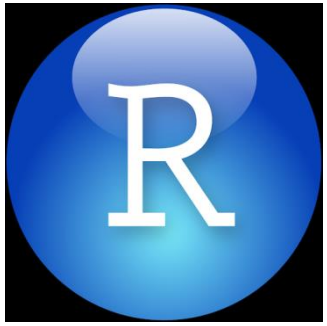
- En estos casos lo mejor que se puede hacer es replicar el análisis con y sin dichas observaciones con el fin de analizar su influencia sobre los resultados. Si dichas observaciones son influyentes el analista debería reportarlo en sus conclusiones y debería averiguar el por que de dichas observaciones.



TECNICAS DE DETECCION DE OUTLIERS

- Prueba de grubbs
- Prueba de Dixon
- Prueba de tukey (Diagrama de caja)
- Análisis de valores atípicos de Mahalanobis
- Regresion simple

PROGRAMAS PARA IDENTIFICAR OUTLIERS



Minitab®

APLICACIÓN DE OUTLIERS EN MINERIA DE DATOS

- Detección fraudes financieros
- Tecnología y telecomunicaciones
- Nutrición y salud
- Negocios

DISTINTOS SIGNIFICADOS DE OUTLIERS

- Error: Si tenemos un grupo de personas y sus pesos y hay una de 350 kilos lo mas seguro es que sea un error.
- Limites: Valores que se escapan por mucho del grupo medio.
- Puntos de interes: Puede que sean los datos que buscamos en el caso de datos anomalos.

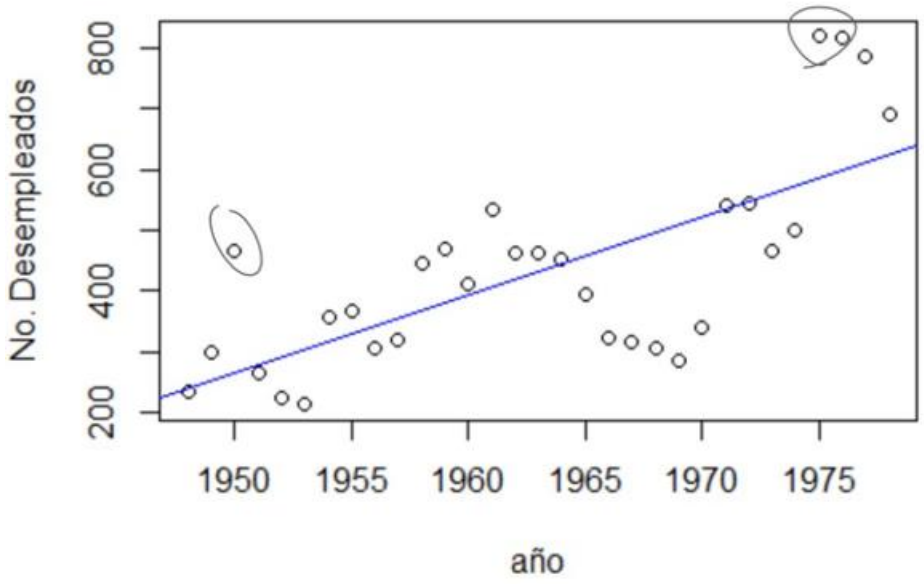
Ejemplo

- Se tiene una base de datos de desempleo por cada mes de 1948 hasta 1978, encuentra que meses hubo datos atípicos.

Año	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
1948	235.1	280.7	264.6	240.7	201.4	240.8	241.1	223.8	206.1	174.7	203.3	220.5
1949	299.5	347.4	338.3	327.7	351.6	396.6	438.8	395.6	363.5	378.8	357	369
1950	464.8	479.1	431.3	366.5	326.3	355.1	331.6	261.3	249	205.5	235.6	240.9
1951	264.9	253.8	232.3	193.8	177	213.2	207.2	180.6	188.6	175.4	199	179.6
1952	225.8	234	200.2	183.6	178.2	203.2	208.5	191.8	172.8	148	159.4	154.5
1953	213.2	196.4	182.8	176.4	153.6	173.2	171	151.2	161.9	157.2	201.7	236.4
1954	356.1	398.3	403.7	384.6	365.8	368.1	367.9	347	343.3	292.9	311.5	300.9
1955	366.9	356.9	329.7	316.2	269	289.3	266.2	253.6	233.8	228.4	253.6	260.1
1956	306.6	309.2	309.5	271	279.9	317.9	298.4	246.7	227.3	209.1	259.9	266
1957	320.6	308.5	282.2	262.7	263.5	313.1	284.3	252.6	250.3	246.5	312.7	333.2
1958	446.4	511.6	515.5	506.4	483.2	522.3	509.8	460.7	405.8	375	378.5	406.8
1959	467.8	469.8	429.8	355.8	332.7	378	360.5	334.7	319.5	323.1	363.6	352.1
1960	411.9	388.6	416.4	360.7	338	417.2	388.4	371.1	331.5	353.7	396.7	447
1961	533.5	565.4	542.3	488.7	467.1	531.3	496.1	444	403.4	386.3	394.1	404.1
1962	462.1	448.1	432.3	386.3	395.2	421.9	382.9	384.2	345.5	323.4	372.6	376
1963	462.7	487	444.2	399.3	394.9	455.4	414	375.5	347	339.4	385.8	378.8
1964	451.8	446.1	422.5	383.1	352.8	445.3	367.5	355.1	326.2	319.8	331.8	340.9
1965	394.1	417.2	369.9	349.2	321.4	405.7	342.9	316.5	284.2	270.9	288.8	278.8
1966	324.4	310.9	299	273	279.3	359.2	305	282.1	250.3	246.5	257.9	266.5
1967	315.9	318.4	295.4	266.4	245.8	362.8	324.9	294.2	289.5	295.2	290.3	272
1968	307.4	328.7	292.9	249.1	230.4	361.5	321.7	277.2	260.7	251	257.6	241.8
1969	287.5	292.3	274.7	254.2	230	339	318.2	287	295.8	284	271	262.7
1970	340.6	379.4	373.3	355.2	338.4	466.9	451	422	429.2	425.9	460.7	463.6
1971	541.4	544.2	517.5	469.4	439.4	549	533	506.1	484	457	481.5	469.5
1972	544.7	541.2	521.5	469.7	434.4	542.6	517.3	485.7	465.8	447	426.6	411.6
1973	467.5	484.5	451.2	417.4	379.9	484.7	455	420.8	416.5	376.3	405.6	405.8
1974	500.8	514	475.5	430.1	414.4	538	526	488.5	520.2	504.4	568.5	610.6
1975	818	830.9	835.9	782	762.3	856.9	820.9	769.6	752.2	724.4	723.1	719.5
1976	817.4	803.3	752.5	689	630.4	765.5	757.7	732.2	702.6	683.3	709.5	702.2
1977	784.8	810.9	755.6	656.8	615.1	745.3	694.1	675.7	643.7	622.1	634.6	588
1978	689.7	673.9	647.9	568.8	545.7	632.6	643.8	593.1	579.7	546	562.9	572.5

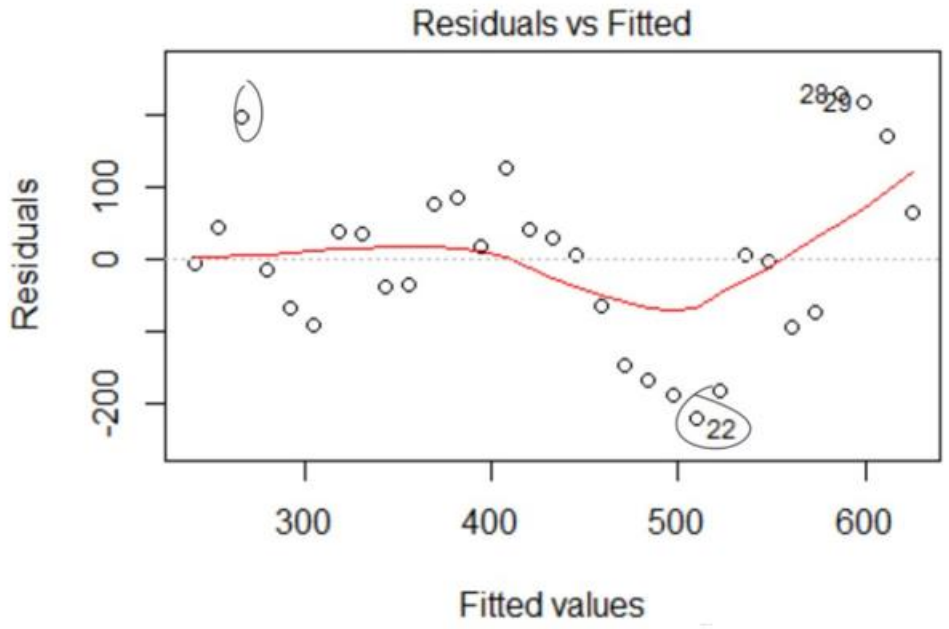
Regresión Lineal

- Podemos observar los datos atípicos graficando la recta de regresión lineal sobre el grafico de dispersión, dichos datos son los que se encuentran mas alejados de la línea de regresión.



Residuales

- Otra forma de verlo es con la grafica de residuales y valores ajustados.

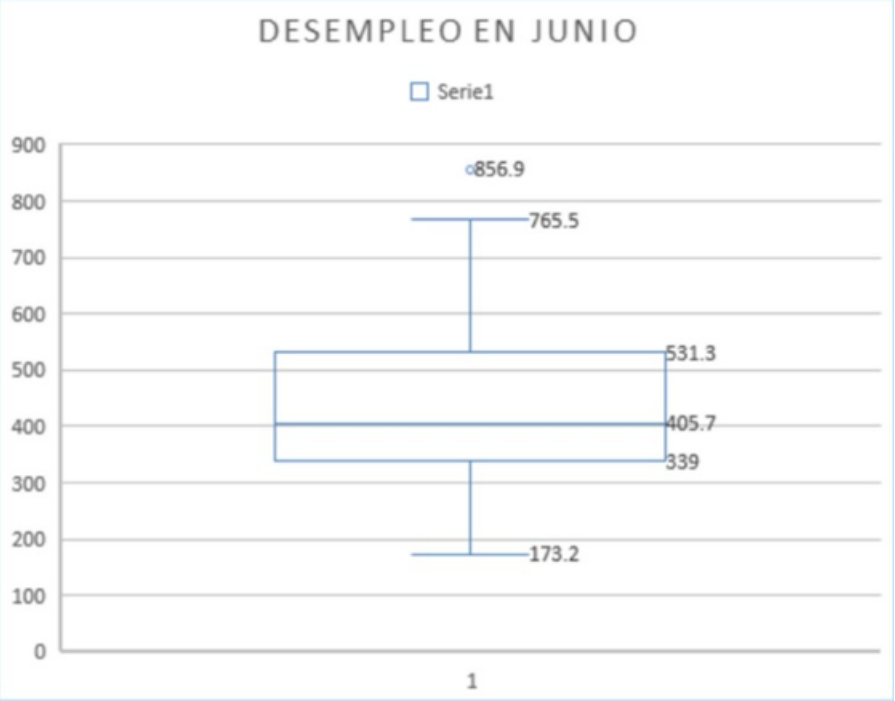


Diferencia en pruebas

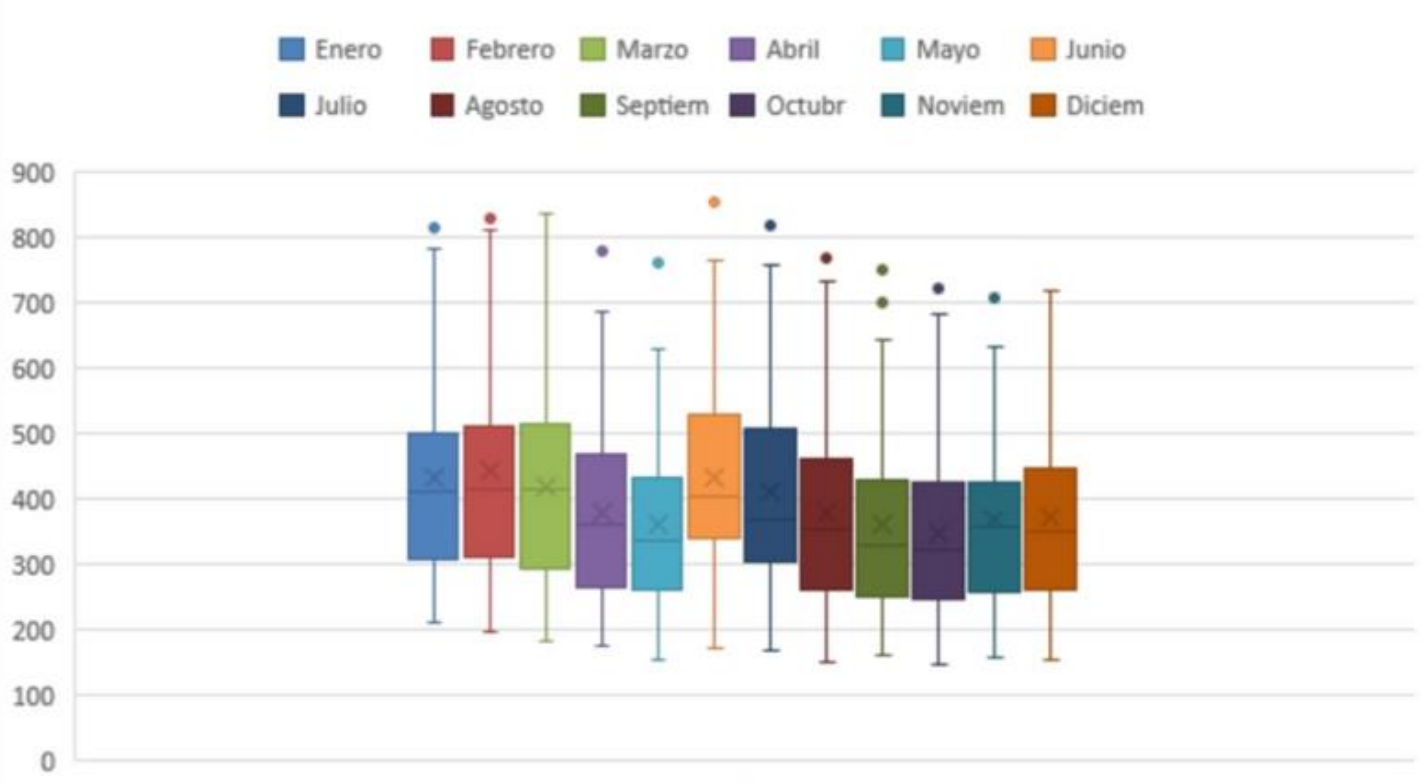
- Dato sospechoso: 856.4 (Junio)
- Prueba de Grubbs:
- Estadístico T = |(Xo – Xm)| / S
- (Xo = Dato sospechoso, Xm = media, S = desviación estándar)

- Según la prueba de Grubbs si T>Gtab entonces dicho dato es atipico (Gtab: valor critico en tablas para pruebas de Grubbs)

- Xm=433.92
- S= 161.26
- Xo=856.4
- T=2.632
- Gtab=2.76 (n=31, a=0.05)
- NO ES ATIPICO



Resumen con diagrama de caja.



Método de detección DBScan Clustering

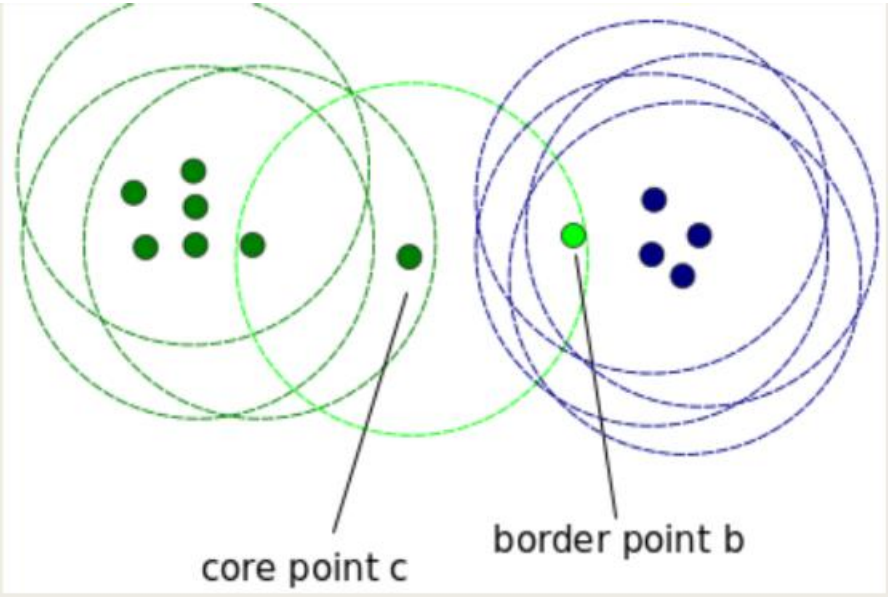
- DBScan es un algoritmo de agrupación en clústeres que utiliza datos agrupados en grupos. También se utiliza como un método de detección de anomalías basado en la densidad con datos unidimensionales o multidimensionales. También se pueden utilizar otros algoritmos de agrupación como k-medias y agrupación jerárquica para detectar valores atípicos.

Método de detección DBScan Clustering

- **Core Points**: Para comprender el concepto de puntos centrales, se deben revisar unos hiperparametros utilizados para definir el trabajo DBScan. El hiperparametro (HP) es **min.simples**. Este es simplemente el numero mínimo de puntos centrales necesarios para formar un grupo. El segundo HP importante es **eps.eps** es la distancia máxima entre dos muestras para que se consideren como en el mismo grupo.

- **Border Points**: se encuentran en el mismo grupo que los puntos centrales, pero mucho mas lejos del centro del grupo.

Método de detección DBScan Clustering



Método de detección DBScan Clustering

- Todos los demás se denominan **Puntos de ruido(Noise Points)**, son puntos de datos que no pertenecen a ningún grupo. Pueden ser anómalos o no anómalos y necesitan mas investigación.

Ejemplo

```
import numpy as np
from sklearn.cluster import DBSCAN
np.random.seed(1)
valoresAleatorios = np.random.randn(50000,2)  * 20 + 20

outlierDetection = DBSCAN(min_samples = 2, eps = 3)
clusters = outlierDetection.fit_predict(valoresAleatorios)
numeroDePosiblesAnomalias = list(clusters).count(-1)
print('\033[95mNumero de posibles anomalias: \033[0m', numeroDePosiblesAnomalias)
```

```
Numero de posibles anomalias:  94
```