Introducción

Tarea perteneciente al módulo de **Estadística** del Máster de Big Data y Data Science por la Universidad Complutense de Madrid.Para la realización de la siguiente tarea haremos uso del lenguaje de programación **Python** y de 2 de sus principales librerías para el calculo :

- **Numpy** : NumPy es una librería de Python especializada en el cálculo numérico y el análisis de datos, especialmente para un gran volumen de datos. Par mas informacion acerca de la librería Numpy, haga click en el siguiente enlace que lo llevara a la documentación.
- **Scypy** : SciPy es una biblioteca libre y de código abierto para Python. Para obtener mas informacion acerca de Scypy, acceda a su documentación.

Por razones de legibilidad se mostrara solamente el resultado(**output**) de las operaciones a realizar en cada uno de los apartados de los ejercicios, sin embargo si se desea ver el trabajo completo en su totalidad,incluyendo todo el código(**input**), haga click en el siguiente enlace hacia Github.

Out[56]:

	Grupo de control	Nivel glucosa basal	Nivel glucosa 60 min
0	1	90	159
1	1	82	151
2	1	80	148
3	1	75	138
4	1	74	141
...	...	...	...
60	2	95	169
61	2	99	172
62	2	88	173
63	2	84	188
64	2	92	160

65 rows × 3 columns

Ejercicio 1

a)

Obtener, usando algún programa estadístico, las medidas de centralización y dispersión para cada uno de los dos grupos de control para el nivel de glucosa basal, especificando para cada uno de los casos si la media es o no representativa,

	Medidas de Centralización		
Grupo de Control	Media	Mediana	Moda
Jovenes(1)	84.686	82.0	75,79,82,90
Adultos(2)	89.4	90.0	88

Coeficiente de variación en grupo de control jóvenes 10.293 %
Coeficiente de variación en grupo de control adultos 8.08 %

Basándonos en la siguiente tabla que muestra los grados de representatividad de la media y teniendo en cuenta los resultados del coeficiente de variación para ambos grupos de control:

- **Jóvenes** - 10.293%
- **Adultos** - 8.08%

Valor del coeficiente de variabilidad	Grado en que la media representa a la serie
De 0 a menos del 10 %	La media es altamente representativa
De 10 a menos del 20 %	La media tiene representatividad.
De 20 a menos del 30 %	La media tiene representatividad
De 30 a menos del 40 %	La media tiene representación dudosa.
De 40 % o más	La media carece de representatividad.

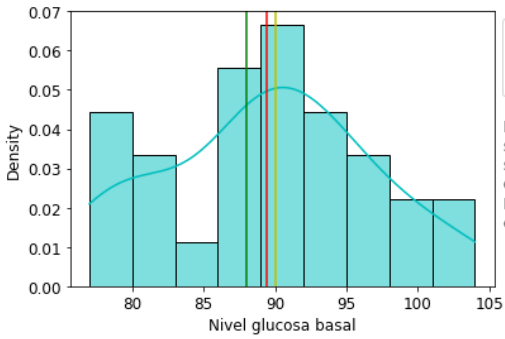
Podemos concluir que para ambos casos **la media es representativa** e incluso para el **grupo de control** de adultos la media es **altamente representativa,** estando dentro del rango del 0 al 10%.

	Medidas de Dispersión		
Grupo de Control	Rango	Varianza	Desviación
Jovenes(1)	38	78.222	8.717
Adultos(2)	27	53.972	7.223

b)

Estudiar la simetría y la curtosis del nivel de glucosa basal en los adultos (grupo de control 2)

Coeficiente de la curtosis en el grupo de control 2: -0.613
Coeficiente de asimetría en el grupo de control 2 : -0.0075



Para el grupo de estudio 2 ,la simetria es ligeramente asimétrica negativa, siendo la media(89.4) menor a la mediana(90) ademas del coeficiente de simetría siendo este -0.0075, asi como se hayan mas valores a la derecha de la media que a la izquierda.
En cuanto a la curtosis, dado que el coeficiente es menor a 0(-0.612), la curtosis es del tipo platicúrtica, con una curva mas plana que la distribución normal

c)

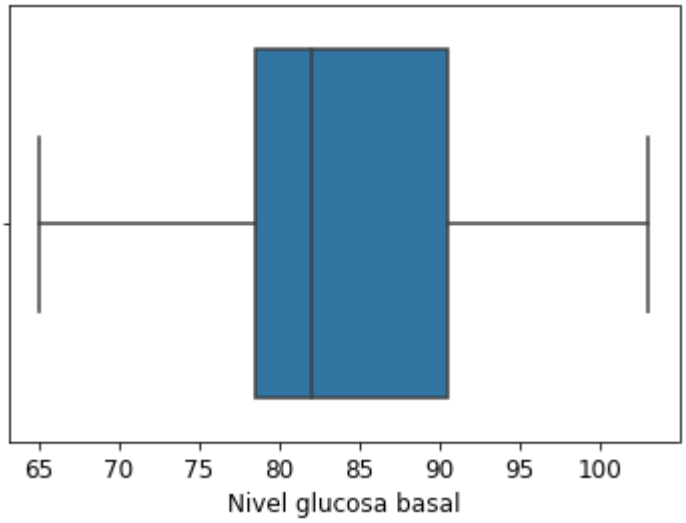
Indicar para cada una de las variables de estudio y en el grupo de control 1 el valor de los cuartiles y su significado y obtener el box- plot (diagrama de cajas) correspondiente. Estudiar la presencia de valores atípicos.

	Valor de los cuartiles		
Variables de estudio	Q1	Q2	Q3
Nivel de glucosa Basal	78.5	82.0	90.5
Nivel de glucosa 60 min	146.5	150.0	154.5

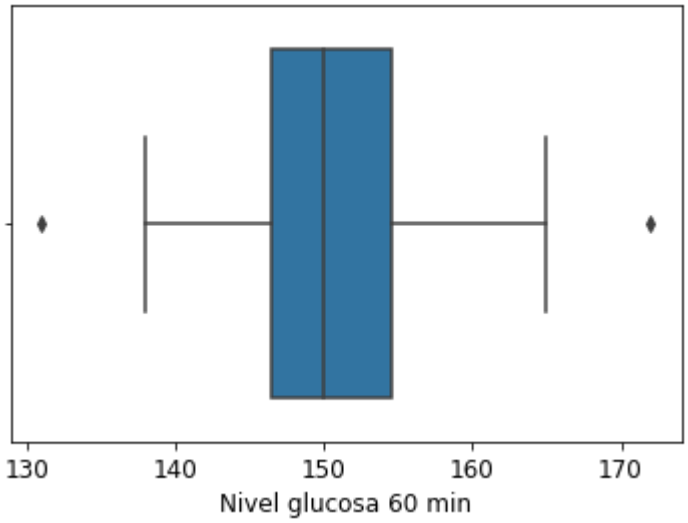
Los cuartiles son valores que dividen una muestra de datos en 4 partes iguales :

- Q1: El primer cuartil Q1, el 25% de los datos es menor o igual a este valor
- Q2: El segundo cuartil Q2 , el valor que divide el 50% de los datos, es también la **mediana**.
- Q3: El tercer cuartil Q3, el 75% de los datos es menor o igual a este valor

La diferencia entre el Q1 y el Q3, que forman la parte inferior y superior de la caja , es también llamado rango intercuartílico.



En la primera figura, del **grupo de control 1** el nivel de glucosa basal, no se observa ningun valor atipíco, es decir ,fuera del límite superior y inferior del diagrama de caja y bigote. Los valores de los bigotes son el mínimo y el máximo de la muestra de datos, **65 y 103** al no poseer valores atípicos.

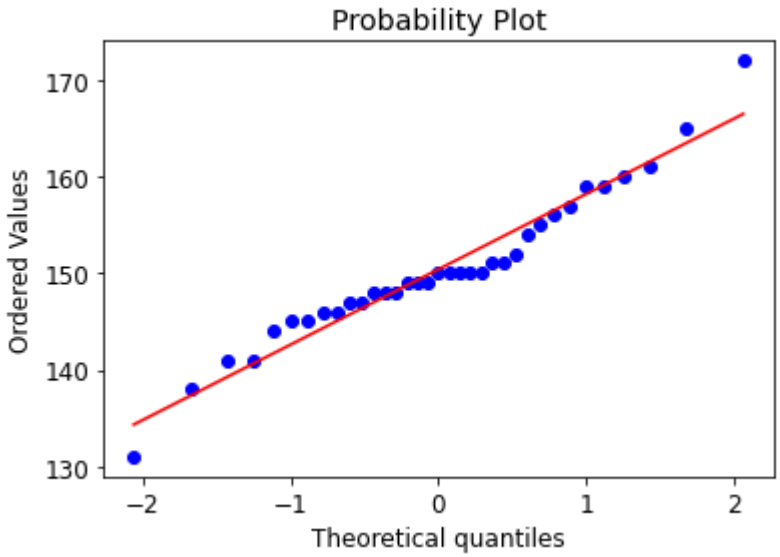


En la segunda figura para el **nivel de glucosa pasados 60 min**, se observan 2 valores atípicos, outliers fuera de los límites de los bigotes, estos son **131** y **172**.

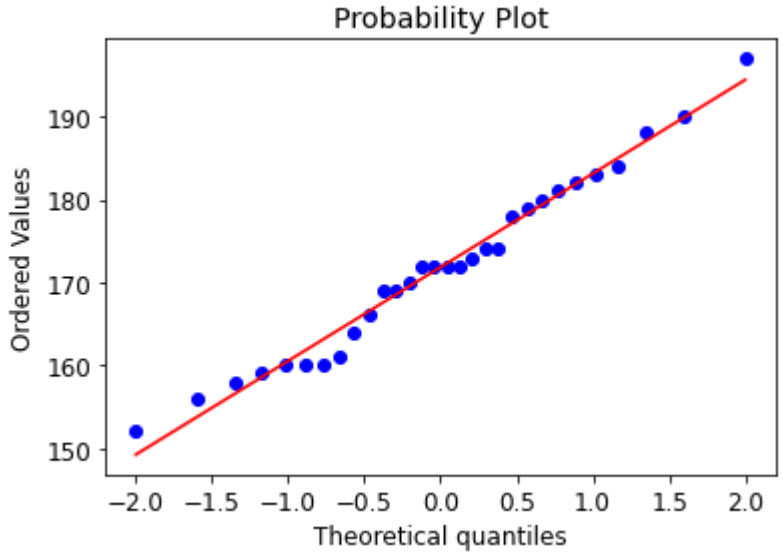
d)

Estudiar la normalidad de los datos de cada uno de los grupos de control estudiados para el nivel de glucosa pasados 60 minutos.

Q-Q Plot Normal Grupo de control 1(Jovenes)



Q-Q Plot Normal Grupo de control 2(Adultos)



Para estudiar la normalidad en ambos grupos de estudio utilizando la variable Nivel glucosa 60 min, utilizamos en primer lugar un método de estudio grafico, **QQ PLOT DISTRIBUCIÓN NORMAL** donde observamos en ambos gráficos que los puntos se acercan a la recta diagonal roja , representando esta los datos teóricos de la distribución normal, por lo tanto podemos deducir que las 2 muestras de datos siguen una distribución normal.

Test Shapiro-Wilk

Además de la prueba gráfica de **QQ PLOT DISTRIBUCIÓN NORMAL** para el estudio de la normalidad, haremos uso del **Test Shapiro-Wilk** ideal para muestras pequeñas. Este consiste en contrastar si un conjunto de datos siguen una distribución normal o no, donde:

- H0: los datos provienen de una distribución normal
- H1: los datos no provienen de una distribución normal

ShapiroResult(statistic=0.9626079201698303, pvalue=0.27341291308403015)
ShapiroResult(statistic=0.9776136875152588, pvalue=0.7591814398765564)

Para ambos casos como el el p-valor es mayor que 0.05 (0.27341 > 0.05 y 0.759181 > 0.05) aceptamos la hipótesis nula (H0), por lo que podemos afirmar que nuestros datos se distribuyen siguiendo una **distribución normal**.

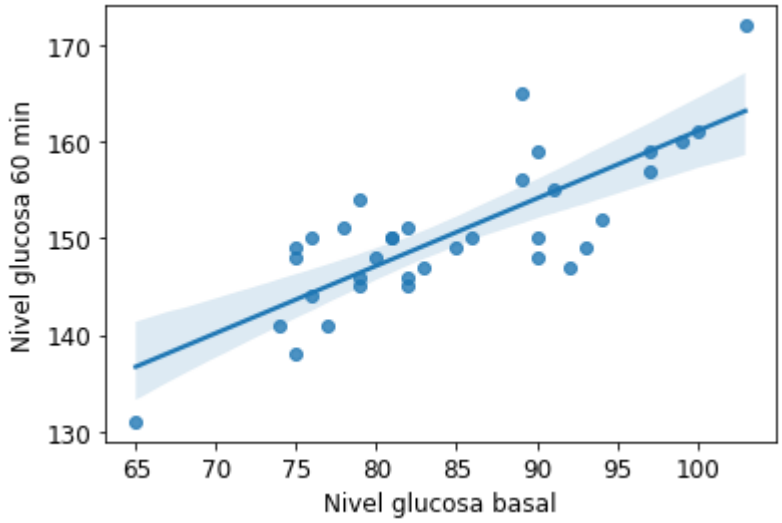
Ejercicio 2

Con los datos del fichero anterior, se quiere estudiar la relación existente entre el nivel basal y el nivel de glucosa que tienen los pacientes sanos jóvenes(grupo 1) una hora después de tomar el preparado de glucosa. Se pide:

a)

Estudiar la relación lineal existente entre estas dos variables de estudio.

Coeficiente de corelacion de Pearson : 0.796405



En el caso de las 2 variables de estudio, Nivel de glucosa basal y Nivel de glucosa pasados 60 min en el grupo de pacientes sanos jovenes (Grupo 1) **existe una corelacion positiva**, como podemos apreciar en el diagrama de dispersion y por el resultado del coeficiente de corelacion de Pearson(0.796405).

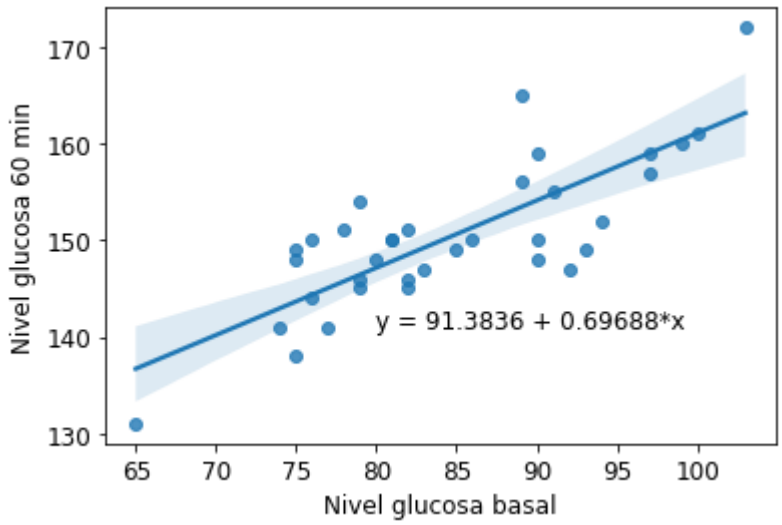
b)

Obtener un modelo lineal que explica el nivel de glucosa en sangre a los 60 minutos en función del nivel basal del paciente y realizar la estimación para un paciente cuyo nivel basal es 83 mg/Dl

Out[73]:

array([0.69688668, 91.38365347])

Coeficiente de corelacion de Pearson : 0.796405



El valor de la pendiente es 0.69688, este valor nos indica que por cada mg/Dl adicional en el **nivel de glucosa basal**, el **nivel de glucosa pasados 60 min** aumentaria en 0.69688

91.3836+0.69688*83 = 149.22

Para un paciente cuyo nivel basal es de 83mg/Dl, su estimacion pasados los 60 min es de : 149.22 mg/Dl

c)

¿Qué tanto por ciento del nivel de glucosa en sangre pasados 60 minutos queda no queda explicado por el anterior modelo?

Para conocer el tanto por ciento no explicado del modelo lineal entre el nivel de glucosa basal y el nivel de glucosa pasados 60 min, es necesario conocer primero el porcentaje explicado que en este caso es del seria 0.796405 elevado al cuadrado: **0.634**, por siguiente el no explicado seria 1-0.634

Porcentaje explicado : 63.4 %
Porcentaje no explicado : 36.6 %

d)

Si aumentásemos el nivel basal de un paciente en 5 mg/Dl ¿Qué variación experimentaría su nivel de glucosa al cabo de 60 minutos?

0.69688*5 = 3.484

Si aumentásemos el nivel basal de un paciente en 5 mg/Dl ,la variacion que experimentaria el nivel de glucosa al cabo de 60 min seria de : 3.484 mg/Dl

Ejercicio 3

a)

Se quiere estudiar si se puede admitir que el nivel medio de glucosa en sangre en el momento de la ingestión en los jóvenes es 88 mg/Dl. Obtener el intervalo de confianza al 95% y al 99% para el nivel medio de glucosa en sangre de los jóvenes y posteriormente contesta a la cuestión planteada con los resultados obtenidos o con un contraste de hipótesis.

Intervalo de confianza del 95%

Intervalos de confianza:
(81.7388139524244, 87.63261461900417)

Intervalo de confianza del 99%

Intervalos de confianza:
(80.85474385243744, 88.51668471899113)

Teniendo en cuenta los resultados obtenidos, los 2 intervalos de confianza de :

Para el 95 % : (81.7388139524244, 87.63261461900417)

Para el 99% : (80.85474385243744, 88.51668471899113)

Solo se podria admitir que el nivel medio de glucosa en la sangre en el momento de la ingestión en los jóvenes es del 88mg/Dl dentro del **intervalo de confianza del 99%** debido a que esta dentro de sus intervalos, sin embargo para el intervalo de confianza del 95% no se puede admitir.

b)

Obtener los intervalos de confianza al 95% y al 99 % para la diferencia de medias en el nivel basal de glucosa entre adultos y jovenes e interpreta los resultados. ¿Se puede concluir que el nivel basal de glucosa de los jóvenes y los adultos es el mismo? . ¿Se cumplen las hipótesis iniciales para determinar los intervalos de confianza?

- **Intervalo de confianza para el 95%**

[1.8644545971824464 **,** 7.564116831388999 **]**

Como este intervalo **no contiene** el valor 0, no podemos aceptar con una probabilidad del 95% que las medias sean iguales.

- **Intervalo de confianza para el 99%**

[0.9263603836797851 **,** 8.50221104489166 **]**

Como este intervalo **no contiene** el valor 0, no podemos aceptar con una probabilidad del 95% que las medias sean iguales.(Aunque este muy cerca del 0)

Al ser ambas muestras aleatorias extraídas de cada población y independientes del tamaño muestral(n), podemos concluir que se **cumplen las hipótesis inciales** para la realización de los intervalos de confianza.

c)

Se quiere estudiar la proporción de la población con un nivel basal de glucosa superior a 95 mg/Dl (prediabetes). A partir de la muestra del fichero (tomando todos los datos) obtener un intervalo de confianza al 98% y contrastar la hipótesis que la proporción de la población con glucosa superior a 95 mg/Dl es 0,15 con nivel de significación del 5%.

Intervalo de Confianza

Intervalo de Confianza al 98%

[0.061 , 0.277 **]**

Podemos afirmar que la proporción de la población con glucosa basal es superior al 95mg/Dl debido a que 0,15 se encuentra dentro de los **intervalos de confianza del 98%**

Contraste de Hipótesis

- **Nivel de Confianza al 95%**
- **H0** -> p = 0.15
- **H1** -> p ≠ 0.15

Se **acepta** la hipótesis nula si **Z** es menor a **Z Crítico**

Se **rechaza** la hipótesis nula si **Z** es mayor a **Z Crítico**

- **Valor de Z** : 0.41349
- **Valor de Z critico** : 1.96

Al ser el valor de z(0.413) menor al valor de z critico(1.96) **se acepta** la hipótesis nula al 95% de confianza y se puede afirmar que la proporción con glucosa basal es superior al 95 mg/dl de 0.15 de proporcion.