**Trabajo Final**

**1-El estudiante debe aportar el set de datos: puede tomar cualquier set de datos de interés, puede ser un set de datos numérico, categórico o ambos. Plantee al menos 3 hipótesis o preguntas de lo que quisiera resolver o encontrar en los datos.**

Para el presente trabajo de investigación se utiliza el set de datos llamado “gym\_members\_exercise\_tracking” tomado de Kaggle, el cual contiene datos de los clientes de un gimnasio, con un total de 973 observaciones y 15 variables (Edad, Género, Peso, Estatura, Max\_BPM, Avg\_BPM, Resting\_BMP, Session\_Duration, Calories\_Burned, Workout\_Type, Fat\_Percentage, Water\_Intake, Workout\_Frequency, Experience\_Level y BMI).

Hipotesis a resolver:

I-Se desea conocer si existe una relación entre la cantidad de calorías quemadas en el cuerpo y el tiempo o la duración (en horas) de las sesiones en el gimnasio (con los datos disponibles).

III-ANDREY

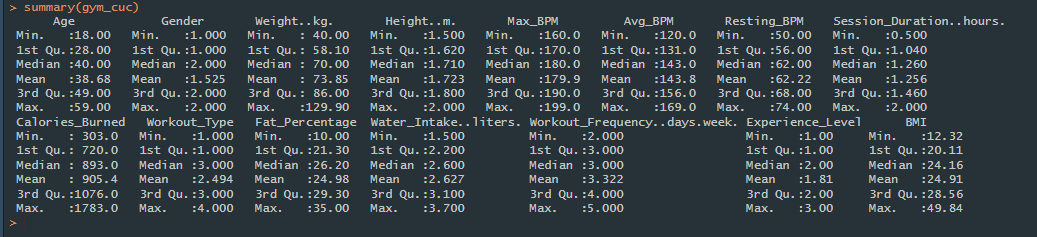
**2- Realice limpieza de datos en caso de ser necesario.**

Se nombra el dataset como gym\_cuc. Posteriormente, se procede a transformar las observaciones tipo character a tipo integer en las columnas de Gender y Workout\_Type con el fin de lograr un mayor alcance en el análisis de los datos.

Para la variable de Gender, las observaciones de Female equivalen a 1 y Male equivalen a 2, para la variable de Workout\_Type, Cardio equivale a 1, HIIT equivale a 2, Strength equivale a 3 y Yoga equivale a 4.

**3.Obtenga las estadísticas básicas del set de datos: moda, media, mediana.**

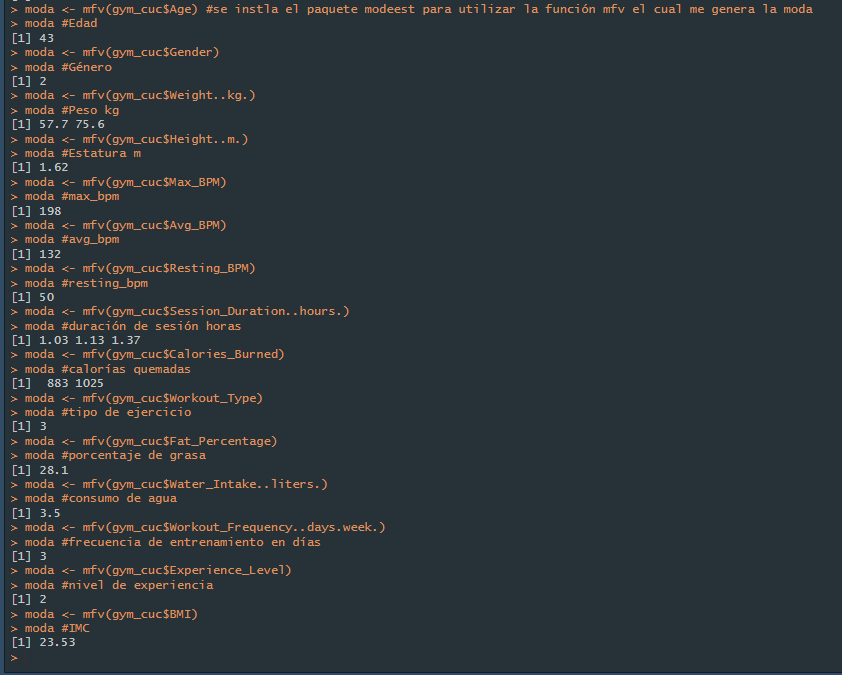
A continuación, se muestran las 15 columnas del dataset gym\_cuc y las estadísticas relacionadas a la media y mediana, adicional de los valores mínimos, máximos, primer cuartil y tercer cuartil.



Estas estadísticas nos aproximan a tener una noción del universo de datos a analizar, donde por ejemplo observamos que la edad mínima de los clientes es de 18 años y la edad máxima de 59 años, teniendo un promedio de 38.68 años.

Mientras que las sesiones de entrenamiento en horas duran en promedio 1 hora y 15 minutos, con sesiones de máximo 2 horas y sesiones mínimas de 30 minutos, relacionado a la observación anterior, la frecuencia en días promedio es de 3 días.

A continuación, se procede a obtener la moda o el valor que mayores veces se repite en las observaciones del dataset.

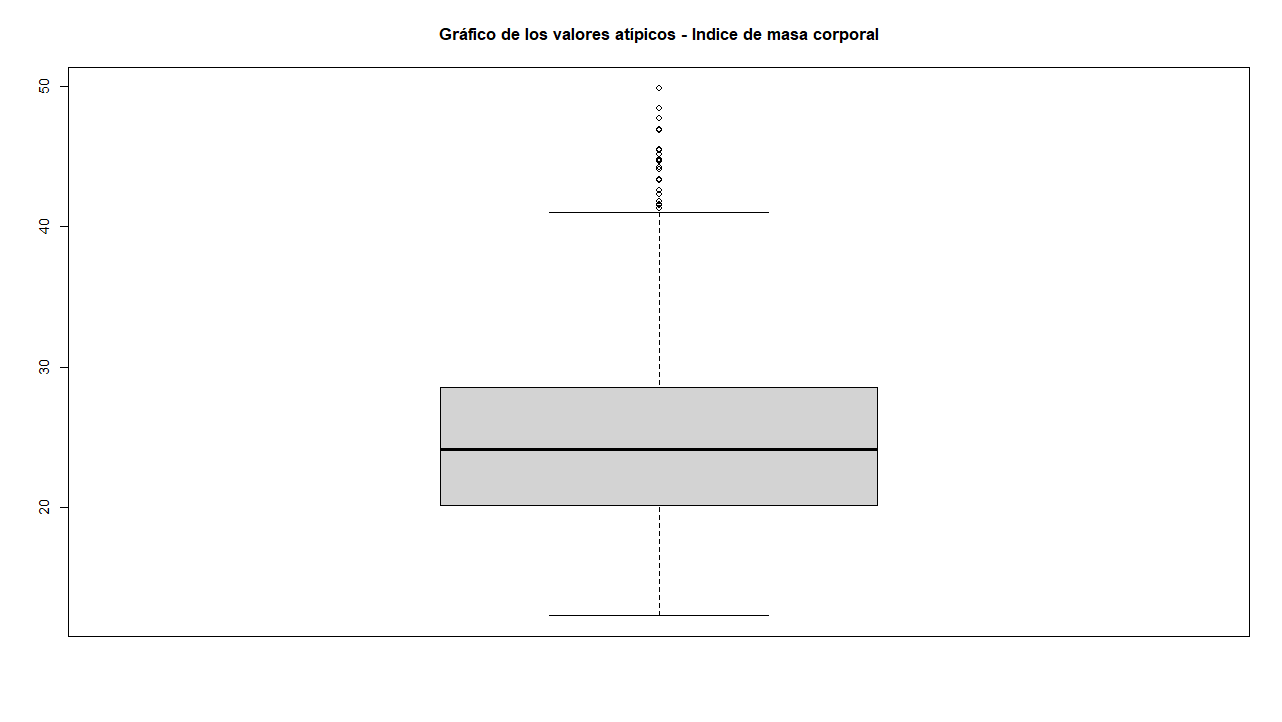


Por ejemplo, se puede observar que la edad de 43 años; es el valor que más se repite entre los clientes que asisten al gimnasio, al igual que el género 2 (masculino), también que el tipo de ejercicio que mayores veces se repite es Strength (fuerza o pesas). Incluso hay pares o tríos de valores que comparten la moda, como lo son el peso con 57.7 y 75.6 y la duración de la sesión en horas con 1.03 - 1.13 y 1.37.

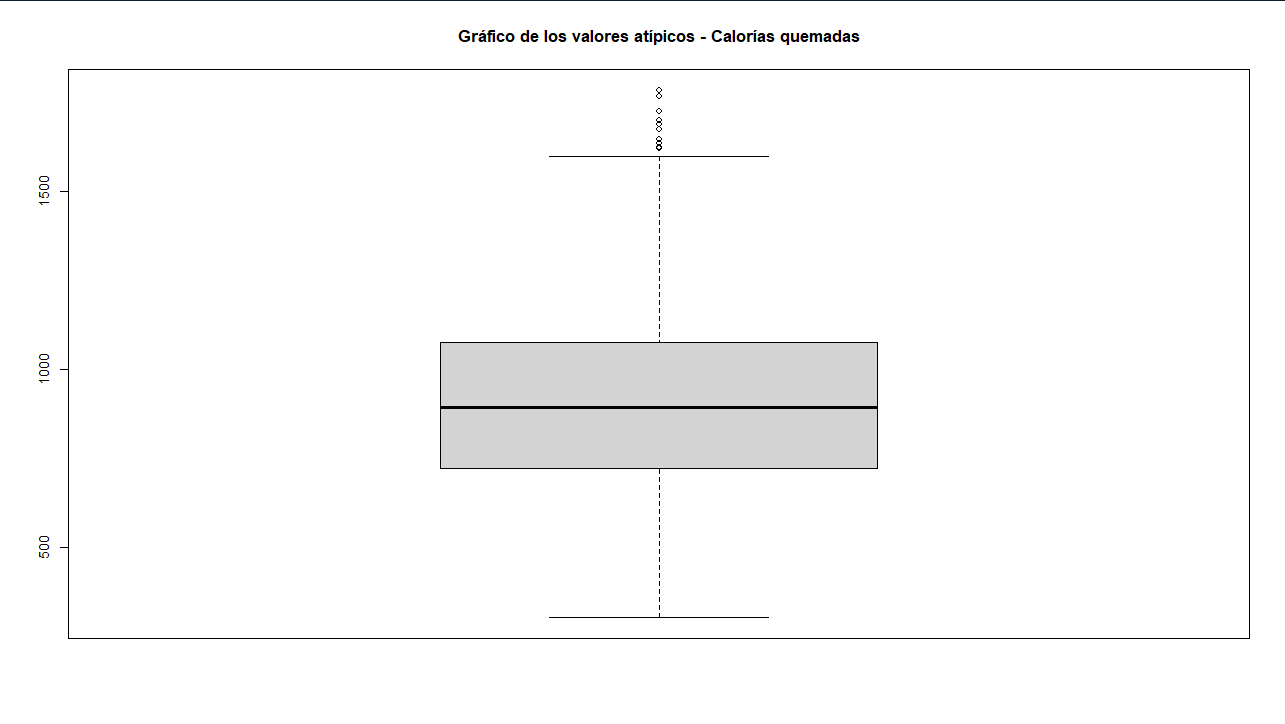
**4.Realice al menos 2 correlaciones entre las variables del set de datos, muestre gráficos y explique con sus palabras cual es la correlación que existe. \*\*Andrey\*\***

**5.Identifique los valores atípicos (outliers) por medio del grafico Box Plot, explique su análisis.**

Respecto a los valores atípicos (outliers), en las observaciones de BMI ó IMC (índice de masa corporal) encontramos que la mediana se encuentra por debajo del límite superior a 25; lo cual significa que la clasificación por debajo de esa mediana se encuentra en un rango saludable, sin embargo, logramos obtener valores atípicos superiores a 40 IMC, lo cual indica de un grupo de clientes que poseen obesidad tipo III, lo cual se podría sugerir un servicio específico para esa población en particular, con rutinas enfocadas y personalizadas, dietas y nutrición complementario y control más estricto de cliente/médico/gimnasio/nutricionista.



Respecto al siguiente gráfico de los valores atípicos en las observaciones de las calorías quemadas, podemos observar que el segundo cuartil se encuentra entre las 720 y 1076 calorías quemadas, con una mediana de 893, sin embargo, en los valores atípicos se tiene observaciones por encima de las 1500 calorías quemadas, lo cual puede relacionarse a esos clientes que le dedican más tiempo de duración del entrenamiento, el nivel de experiencia, la cantidad de agua consumida, entre otros.



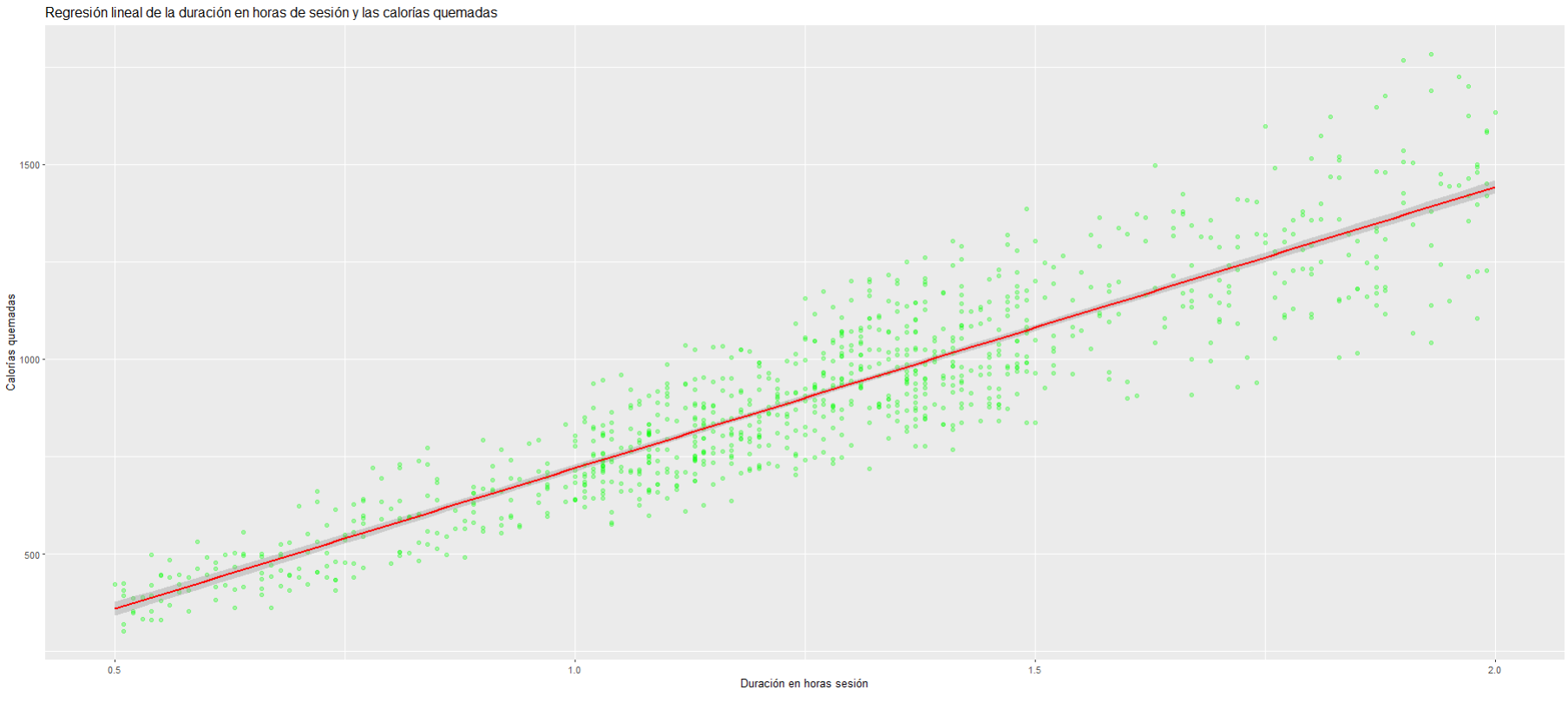
**6.Aplique 2 algoritmos: no supervisado clustering y cualquier otro a escogencia visto en clase para encontrar algún patrón y/o predicción entre los datos. Explique su análisis.**

**Algoritmo de clustering \*\*Andrey\*\***

**Algoritmo de Regresión lineal**

Se desea conocer si existe relación entre la cantidad de calorías quemadas en el cuerpo y el tiempo (en horas) de las sesiones en el gimnasio.

Para poder responder a la hipótesis anterior, se aplica el modelo de regresión lineal simple, con el objetivo de encontrar una relación entre ambas variables; donde la duración de la sesión de entrenamiento en horas hace referencia a las variables independientes del eje “x” y las calorías quemadas hacen referencia a las variables dependientes o incógnitas del eje “y”.



Como se puede visualizar en el gráfico anterior, existe una relación lineal positiva, pues las observaciones del eje “y” van incrementando conforme el eje “x” aumenta, observamos que la pendiente (línea de color rojo) se desplaza hacia la derecha de forma ascendente, con lo que podemos asociar una relación lineal entre las variables de las calorías quemadas y el tiempo en horas del entrenamiento. Lo cual es útil si se desea estimar la cantidad de calorías a quemar durante la sesión de entrenamiento en el gimnasio y el tiempo que dura la rutina, así se puede adecuar la rutina, según los requerimientos de cada persona, ya sea tiempo disponible, intensidad de la rutina, días disponibles, entre otros. Por ejemplo, si una persona solo dispone de 30 minutos diarios para realizar ejercicio en el gimnasio, podrá quemar un estimado de 250 calorías diarias. Si dispone de 1 hora diaria para realizar los ejercicios, podrá quemar un estimado de 740 calorías diarias, mientras que, si dispone de 1 hora y 30 minutos para ejercitarse, podrá quemar un estimado de 1100 calorías, finalmente, si esa persona dispone de 2 horas para realizar su rutina de entrenamiento en el gimnasio, podrá quemar un estimado de 1450 calorías.