

"Ciencia de datos aplicada al estudio de la Obesidad y otras enfermedades crónicas en Córdoba"

Práctico N°1 Ejercicios de análisis y visualización

Integrantes del grupo X:


Basmadjian, Osvaldo Martín

Fernández, María Emilia

Romero, Fernando

1. Elección de variables

Como primera aproximación al conjunto de datos elegimos las siguientes variables:

- **IMC**, una variable cuantitativa continua que representa el índice de masa corporal. Además, exploramos su versión categórica (**clasifIMC**). El índice de masa corporal es una razón matemática que asocia la masa (en kg) y la talla (en mts) de un individuo. Estos índices permiten clasificar a las personas en bajo peso, peso normal, preobesidad, obesidad grado I, obesidad grado II, obesidad grado III. Así, ambas variables constituyen índices nutricionales que brindan información directa acerca de cómo la obesidad (nuestro problema de estudio) se distribuye en nuestra muestra. Aunque son índices ampliamente utilizados, es valedero mencionar que no tienen en cuenta rasgos como la edad, el sexo o la ascendencia (etnia) de las personas clasificadas.
- **cc**, una variable cuantitativa continua que representa la circunferencia de cintura. Además, exploramos su versión categórica (**riesgo**). La circunferencia de cintura constituye una medida antropométrica utilizada en la práctica clínica para valorar la grasa visceral y por tanto el riesgo cardiovascular. Esta variable podría ayudar a predecir el estado nutricional de un sujeto de la población de Córdoba. Respecto a este último punto, el análisis conjunto de esta variable junto al IMC, nos permitiría dilucidar cuál de ellas es más sensible o mejor predictora del estado nutricional de los individuos de la población de Córdoba.
- **mets**, una variable cuantitativa continua que da cuenta del nivel de actividad física de los individuos y que hemos categorizado (según bibliografía) en los siguientes niveles: Sedentario (valor de mets de 0-599), Activo (valor de mets de 600-1499), Muy Activo(valor de mets igual o mayor a 1500). Incluimos esta variable en el análisis dado que el nivel de actividad física modula el metabolismo energético de un individuo, y por tanto podría impactar no sólo en el peso de los mismos (y por ende en el IMC y la cc) si no también en el metabolismo de los alimentos consumidos, acumulación de grasa y afectar en última instancia indicadores fisiológicos como la presión arterial. Así el nivel de actividad física podría constituir un factor ambiental de riesgo o protector frente a la obesidad o enfermedades cardiovasculares.
- **fgr2**, una variable cuantitativa continua que representa la ingesta de grasas (sin distinguir por tipo de grasa) consumidas por los individuos. Incluimos esta variable a fin de valorar si este grupo alimentario representa un factor ambiental de riesgo capaz de promover el desarrollo de la obesidad, considerando que la ingesta de grasas tiene impacto sobre el metabolismo energético y la acumulación de grasa visceral.
- **tenmax**, una variable cuantitativa continua, calculada como promedio de las tensiones arteriales máximas de los individuos. Incluimos esta variable en el análisis como indicadora del riesgo cardiovascular de una persona, una condición fisiológica/comorbilidad de la obesidad. De manera complementaria, exploramos además, la variable hipertensión (hta) 

- **Sexo** y **Edad**, como covariables que circunscriben el contexto fisiológico-endocrinológico particular en el que las demás variables se desenvuelven.

2. Análisis Univariado

2.1. Sexo y Edad de los encuestados

Para la variable sexo se cuentan con 4292 registros en el dataset. Esta variable categórica, tomó dos valores posibles: masculino (1) o femenino (2). El 58% de las personas encuestadas fueron de sexo “femenino”, representando una mayor proporción que aquellas de sexo “masculino” que constituyeron el 42% restante (Fig. 1).

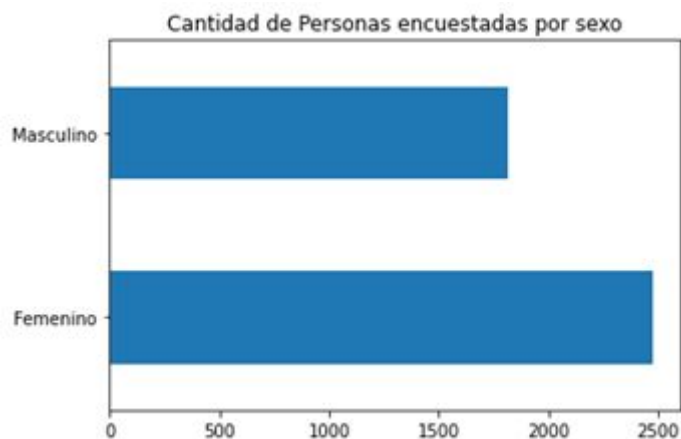


Fig. 1: Personas encuestadas por sexo

Para la variable edad se cuentan con 4292 registros en el dataset. Los individuos encuestados tuvieron entre 18 y 97 años. El promedio de los individuos encuestados fue 42 años, aunque la mayoría de los individuos encuestados (moda) tenían 23 años. Esta diferencia entre media y moda nos indica que la edad no se encuentra normalmente distribuida en nuestra muestra (Fig. 2).



Fig. 2: Distribución de las edades

2.2. Índice de Masa Corporal

Los estadísticos descriptivos correspondientes a la variable índice de masa corporal (IMC) se presentan en la Tabla 1. Para esta variable se cuentan con 4292 registros en el dataset. Esta variable cuantitativa continua tomó valores entre 14.76 y 51.11 kg / m², indicando que en la muestra se encontraron representados individuos de “bajo peso” (menos de 18.5 kg / m²) a individuos con “obesidad grado III” (Mayor o igual a 40 kg / m²). La media y a mediana para el IMC en la muestra se encontraron próximas entre sí con valores de 25.82 y 25.10 kg / m², respectivamente, lo cual indicaría que en promedio y al menos la mitad de las personas encuestadas tiene “preobesidad” (aunque con un valor cercano al umbral inferior de esta categoría). Sin embargo, el valor de IMC más frecuente en la muestra (la moda) fue de 22.77 kg / m², correspondiente a la categoría “peso normal”. El hecho de que media, mediana y moda no coincidan en la distribución de frecuencias del IMC (Tabla 1 y Fig. 3A), probablemente indica que esta variable no se encuentra normalmente distribuida en la muestra (valor del estadístico W = 0.94 y valor p = 1.33e-36). En este sentido, la media muestral anteriormente mencionada no resultaría un buen estimador de la media poblacional, siendo más apropiado utilizar la mediana como estimador de la misma (que justamente tomó un valor un poco inferior al de la media muestral).

Las categorías de IMC más frecuentes en la muestra fueron, en orden decreciente, “peso normal”, “preobesidad” y “obesidad grado I”. Mientras que las categorías “peso bajo”, “obesidad grado II” y “obesidad grado III”, presentaron frecuencias muy bajas.

El IMC no se encuentra igualmente distribuido entre individuos de sexo masculino y femenino de la muestra (Fig. 3C y 3D). En este sentido, la distribución de frecuencia del IMC para el sexo femenino es más asimétrica que la del sexo masculino, con mayor frecuencia de valor bajos de la variable IMC (Fig. 3C). Particularmente a nivel de la categoría “peso normal” la frecuencia de individuos de sexo femenino es mayor que la de sexo masculino (Fig. 3D).

Totales	4292.00
Media	25.82
Desviacion Estandar	4.78
Minimo	14.76
Cuartil 25%	22.49
Cuartil 50%	25.10
Cuartil 75%	28.38
Maximo	51.11
Moda	22.77

Tabla 1: Descriptores IMC

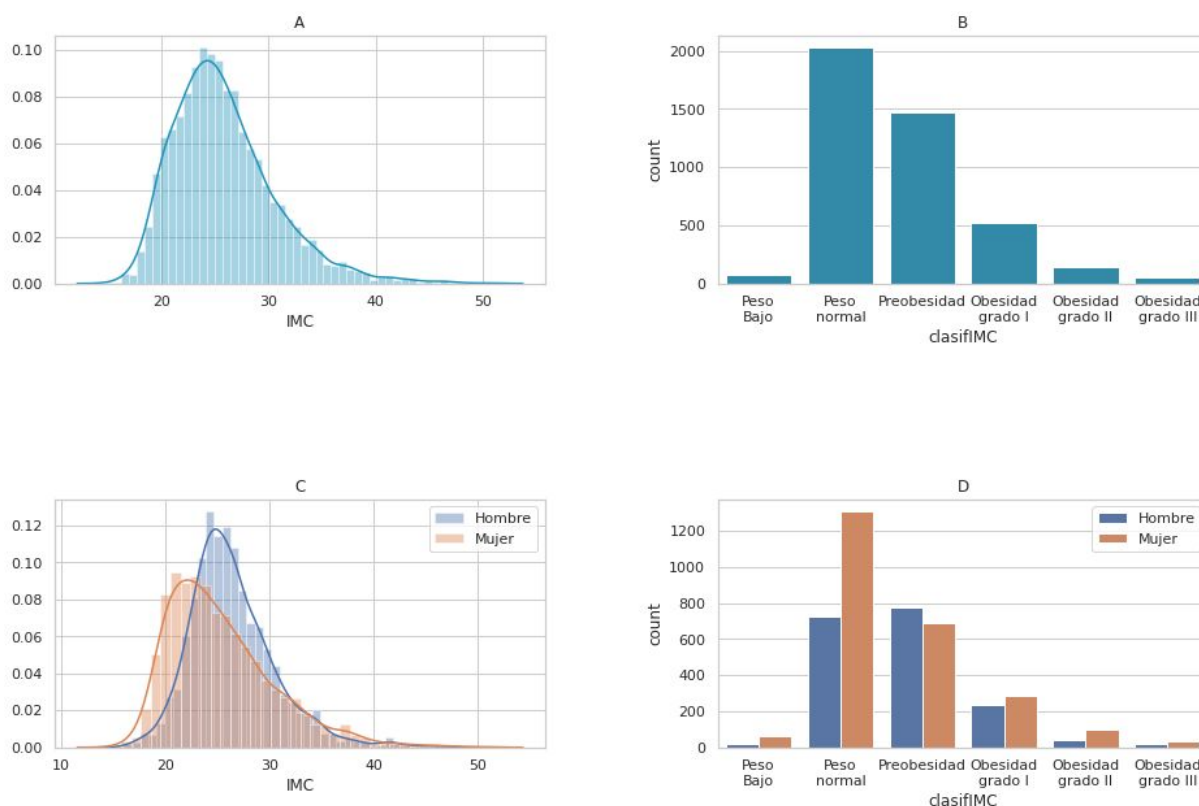


Fig. 3: Análisis IMC y clasifIMC

2.3. Circunferencia de cintura

Los estadísticos descriptivos correspondientes a la variable circunferencia de cintura (cc) se presentan en la Tabla 2. Para esta variable se cuentan con 4292 registros en el dataset. Esta variable cuantitativa continua tomó valores entre 52.00 y 196.00 cm, indicando que en la muestra se encontraron representados individuos de “bajo riesgo” (sexo femenino: <80 cm; sexo masculino: < 94 cm) a individuos con “alto riesgo” (sexo femenino: ≥ 88 cm; sexo masculino: ≥ 102 cm). La media, mediana y moda muestral no fueron coincidentes (Tabla 2, Fig. 4A). La circunferencia de cintura no se encuentra normalmente distribuida en la muestra (Fig. 4A; valor del estadístico W = 0.96 y valor p = 7.55e-31).

La circunferencia de cintura no se encuentra igualmente distribuida entre individuos de sexo masculino y femenino de la muestra (Fig. 4C y 4D). Particularmente, la distribución de frecuencia de la circunferencia de cintura para el sexo femenino presenta mayor frecuencia de valores bajos de la variable que en el sexo masculino (Fig. 3C). Sin embargo, considerando las categorías de riesgo (Fig. 4D), observamos que la frecuencia de individuos con cintura de “bajo riesgo” es semejante entre individuos de sexo femenino y masculino. Esto viene dado por el hecho de que los umbrales de cada categoría de riesgo de la circunferencia de cintura para el sexo femenino están siempre unos puntos por debajo que los umbrales en el sexo masculino. La categoría de circunferencia de cintura de “bajo riesgo” es la que presenta mayor frecuencia en ambos sexos. En los individuos de sexo masculino las categorías “riesgo incrementado” y “alto riesgo” presentaron frecuencias iguales entre sí, y 3 veces menores que la categoría “bajo riesgo”. A diferencia de los

encuestados de sexo masculino, en individuos de sexo femenino, la frecuencia de la categoría de circunferencia de cintura de “riesgo alto” fue mayor que la frecuencia de la categoría “riesgo incrementado”.

Totales	4292.00
Media	86.72
Desviacion Estandar	15.07
Minimo	52.00
Cuartil 25%	75.00
Cuartil 50%	85.00
Cuartil 75%	96.00
Maximo	196.00
Moda	80.00

Tabla 2: Descriptores cc

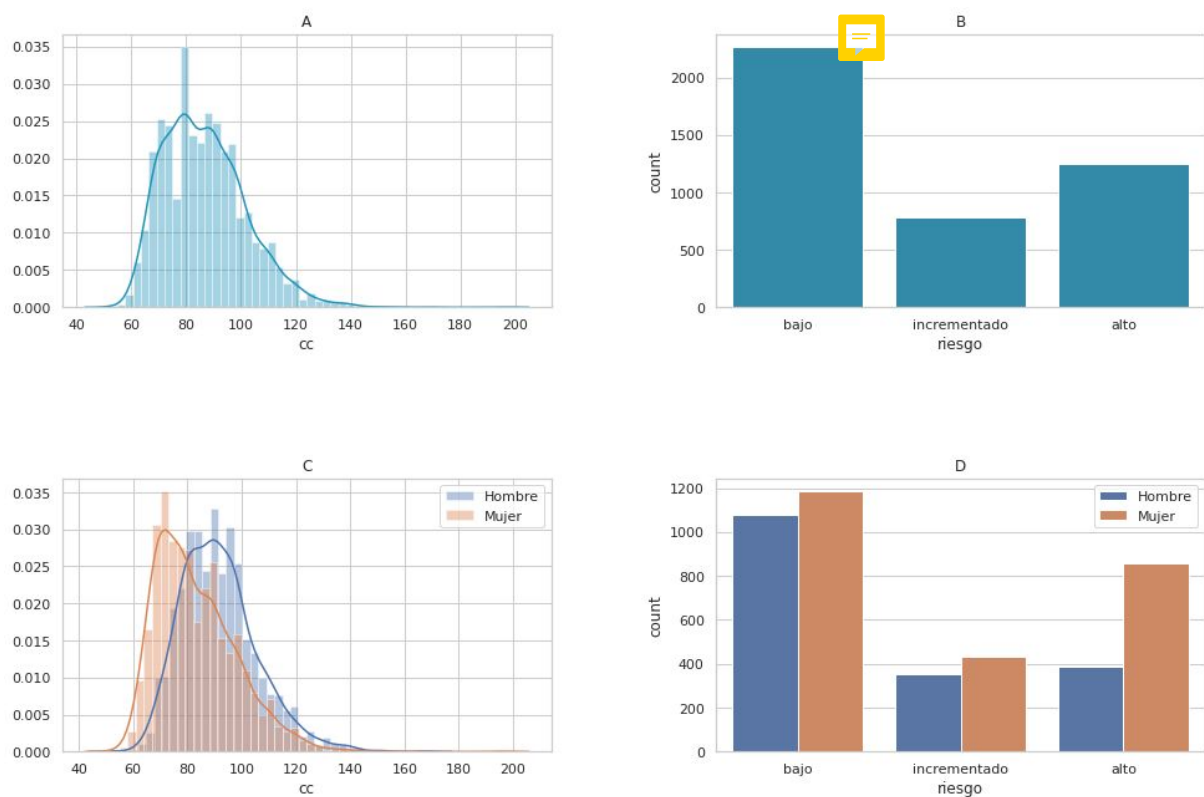


Fig. 4: Análisis cc y riesgo

2.4. Nivel de Actividad Física

Los estadísticos descriptivos correspondientes a la variable continua nivel de actividad física (mets) se presentan en la Tabla 4. Se cuentan con 4292 registros en el dataset para esta variable. Esta variable tomó valores entre 0 y 10000, indicando que en la muestra se encontraron representados individuos con niveles de actividad física equivalentes a la

categoría “Sedentarios” (de 0 a 599 mets) hasta individuos con niveles de actividad correspondientes a la categoría “Muy Activos” (más de 1500 mets).

La media muestral tomó un valor de mets 663. Sin embargo, considerando la gran asimetría de la distribución de esta variable en la muestra (Fig. 5A; valor del estadístico $W = 0.59$ y valor $p = 0$), debemos destacar que dicho valor promedio está muy afectado por valores extremos y no es representativo del comportamiento general de los datos. En este sentido, la mediana y la moda son mejores estimadores de la media poblacional, y a partir ellos observamos que al menos la mitad de los individuos de la muestra son sedentarios y que el valor de mets más frecuente en la muestra es 0 (Tabla 4). Así, la categoría de nivel de actividad física (mets_cat) más frecuente en la muestra para ambos sexos fue “Sedentario”, seguido por las categorías “Activo” y “Muy Activo” que mostraron frecuencias semejantes entre sí y entre sexos.

Es valedero destacar la gran dispersión que presentan los valores de esta variable (asociado probablemente a la presencia de algunos outliers con valores extremos muy altos, lo cual se observa en el histograma de frecuencias), con una desviación estándar de casi el doble de la media (Tabla 3).

Totales	4292.00
Media	663.69
Desviacion Estandar	1189.19
Minimo	0.00
Cuartil 25%	0.00
Cuartil 50%	198.00
Cuartil 75%	852.00
Maximo	10000.00
Moda	0.00

Tabla 3: Descriptores Nivel de actividad física

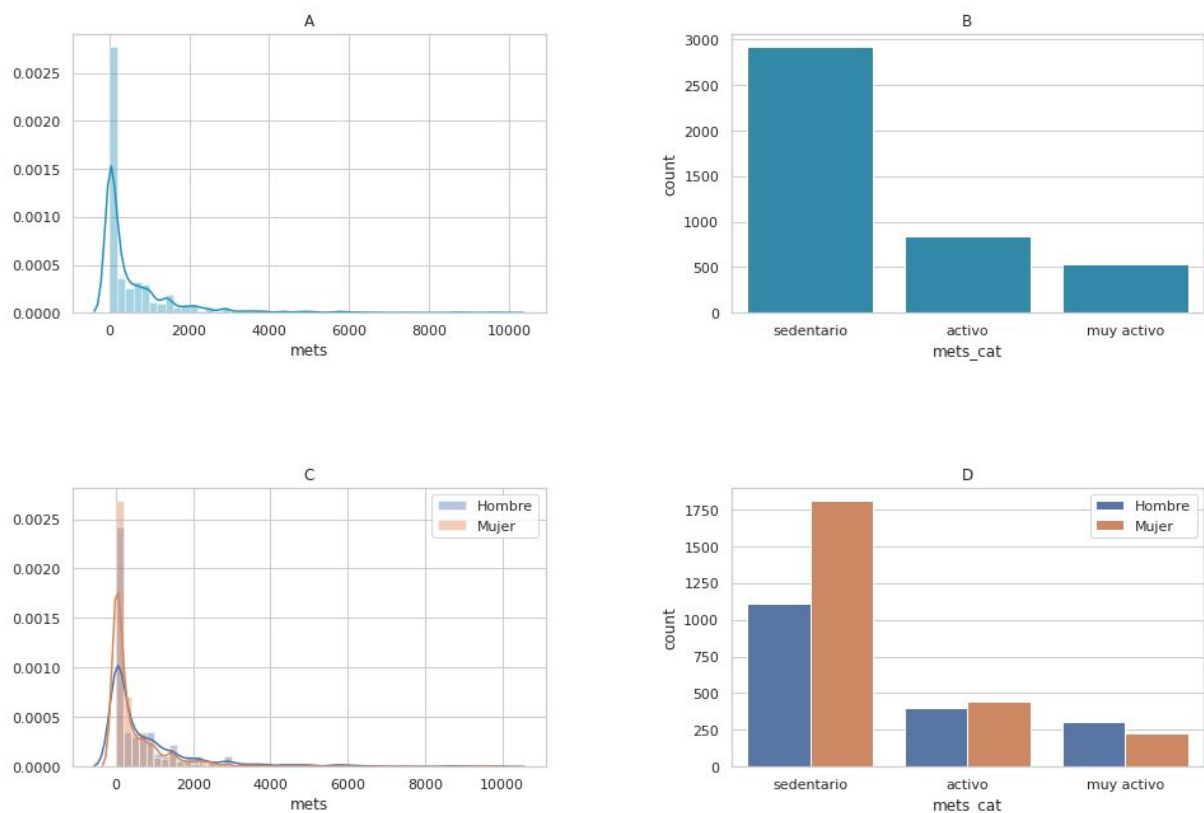


Fig. 5: Análisis nivel de actividad física

2.5. Consumo de Grasas

Los estadísticos descriptivos correspondientes a la variable consumo de grasas (fgr2) se presentan en la Tabla 4. Para esta variable se cuentan con 4292 registros en el dataset. Esta variable cuantitativa continua tomó valores entre 3.32 y 1152.00 g.

La media, mediana y moda no son coincidentes, reflejando que esta variable no se distribuye normalmente en la muestra (valor del estadístico W = 0.81 y valor p = 0). Interesantemente, existen tres modas aparentes para esta variable (Tabla 4), lo cual podría estar asociado a la forma en la que se registró esta variable. En este contexto, la mediana resultaría un mejor estimador de la media poblacional.

De modo equivalente entre individuos de ambos sexos, la categoría consumo de grasas “alto” presentó la mayor frecuencia, seguida de las categorías “adecuado” y “bajo”, en orden decreciente.

Es valedero destacar la gran dispersión (asociado probablemente a la presencia de algunos outliers con valores extremos muy altos, lo cual se observa en el histograma de frecuencias) que presentan los valores de esta variable, con una desviación estándar de aproximadamente la mitad del valor la media (Tabla 3).

Totales	4292.00
Media	102.08
Desviacion Estandar	48.23
Minimo	3.32
Cuartil 25%	72.97
Cuartil 50%	94.90
Cuartil 75%	122.68
Maximo	1152.00
Moda_1	84.76
Moda_2	96.88
Moda_3	112.53

Tabla 4: Descriptores consumo de grasa

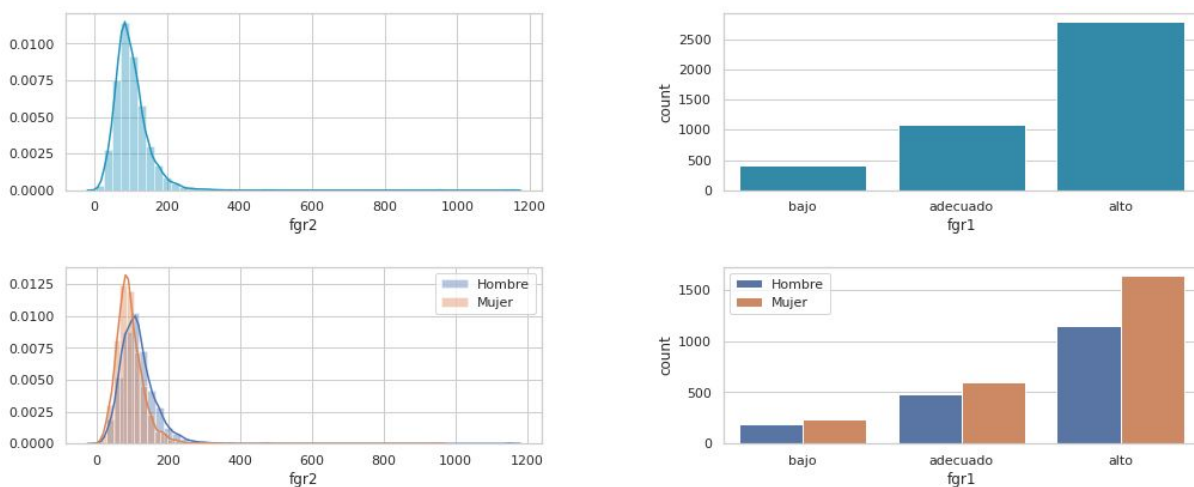


Fig. 6: Analisis consumo de grasa

2.6. Tension Arterial Maxima

Los estadísticos descriptivos correspondientes a la variable tensión arterial máxima (tenmax, como promedio de ten1max y ten2max) se presentan en la Tabla 5. Para esta variable se cuentan con 4292 registros en el dataset. Esta variable cuantitativa continua tomó valores entre 90 y 170.

La media, mediana y moda muestrales fueron coincidentes y tomaron un valor de 120 (Tabla 5), indicando que la tensión arterial máxima promedio se encuentra, probablemente, normalmente distribuida en la muestra (Fig. 7A; valor del estadístico $W = 0.97$ y valor $p = 7.63e-27$), el fallo en la prueba de normalidad, se puede atribuir a los valores definidos, no continuos de las mediciones por la naturaleza del instrumento de medición.

La mayoría de los individuos de ambos sexos no tienen hipertensión arterial (Fig. 7B y 7D).

Totales	4292.00
Media	120.92
Desviacion Estandar	10.76
Minimo	90.00
Cuartil 25%	115.00
Cuartil 50%	120.00
Cuartil 75%	126.50
Maximo	170.00
Moda	120.00

Tabla 5: Descriptores tensión máxima

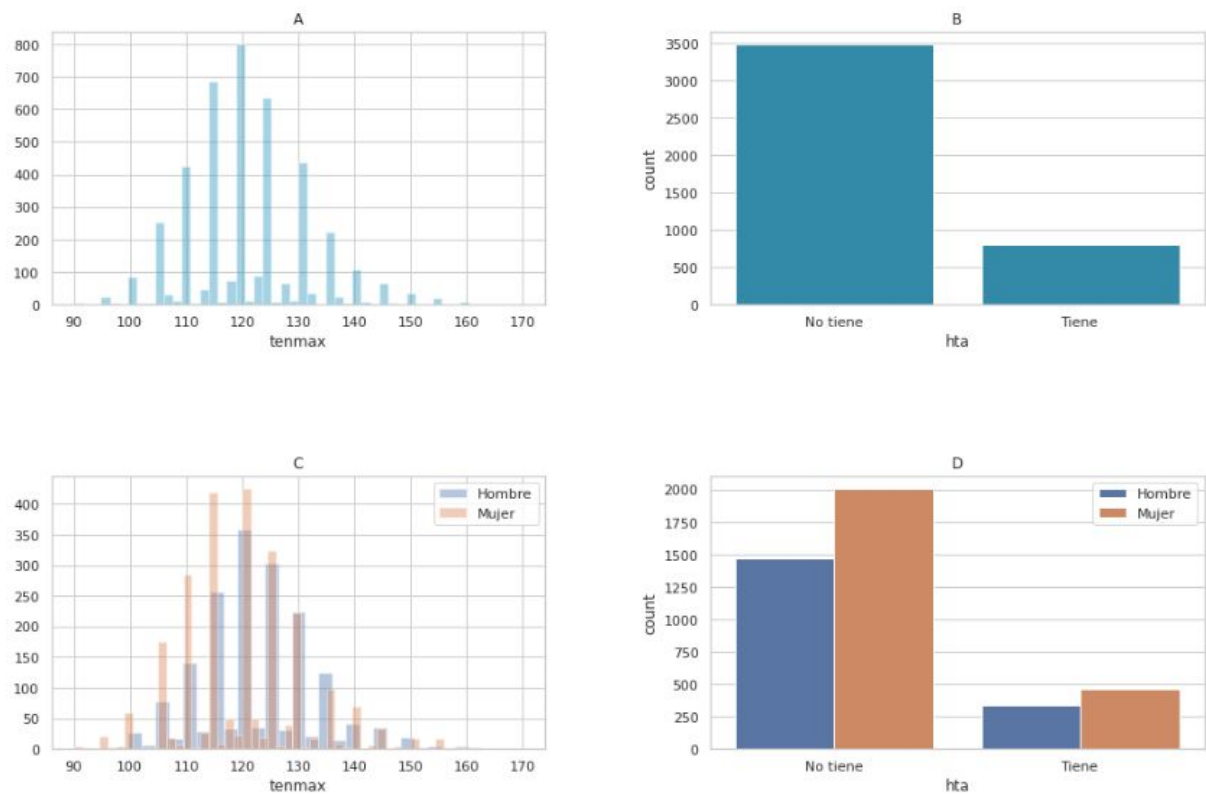


Fig. 7: Análisis tensión máxima

3. Relaciones entre variables

3.1. IMC y CC

El índice de masa corporal y la circunferencia de cintura se encuentran correlacionadas linealmente de manera significativa (valor $p= 0.0$, test de Spearman). Específicamente, dicha correlación es positiva con valor de 0.82 (Fig. 8). Debido a que ambas variables no tienen distribución normal se utilizó el test de spearman para calcular su correlación.

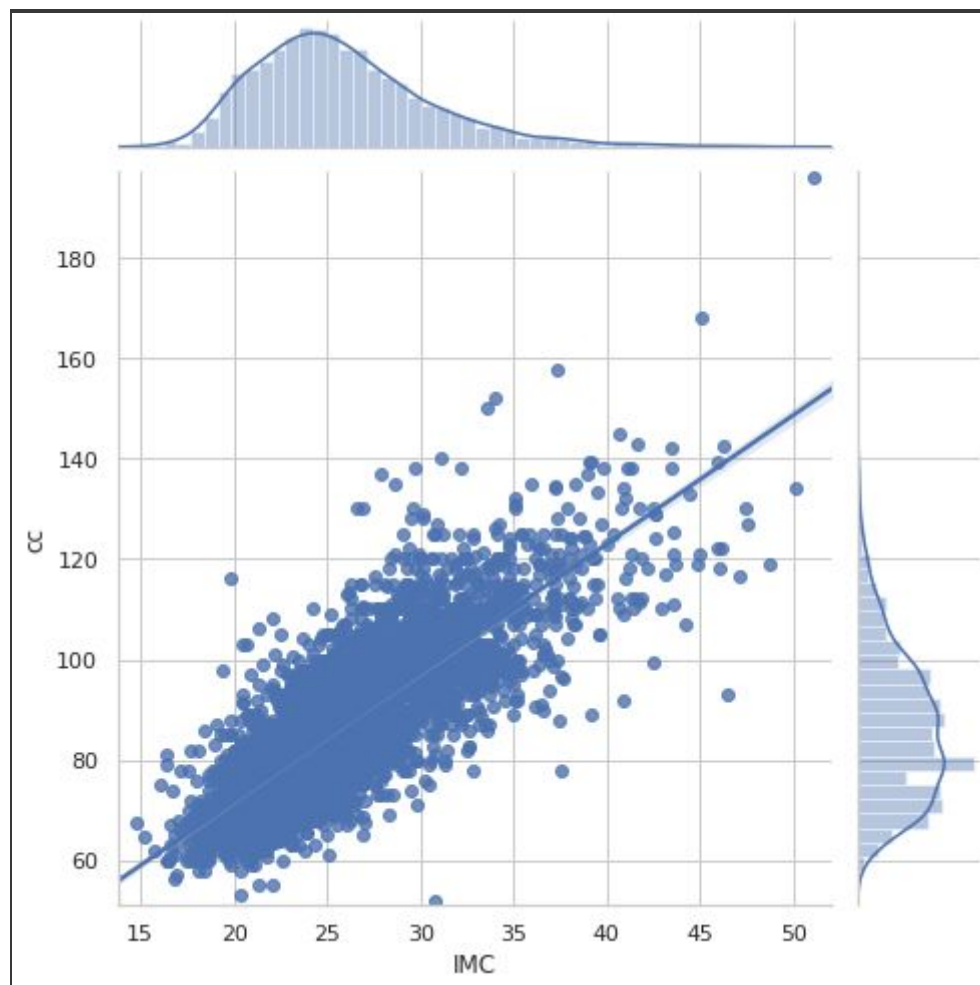


Fig. 8: Scatterplot cc e IMC

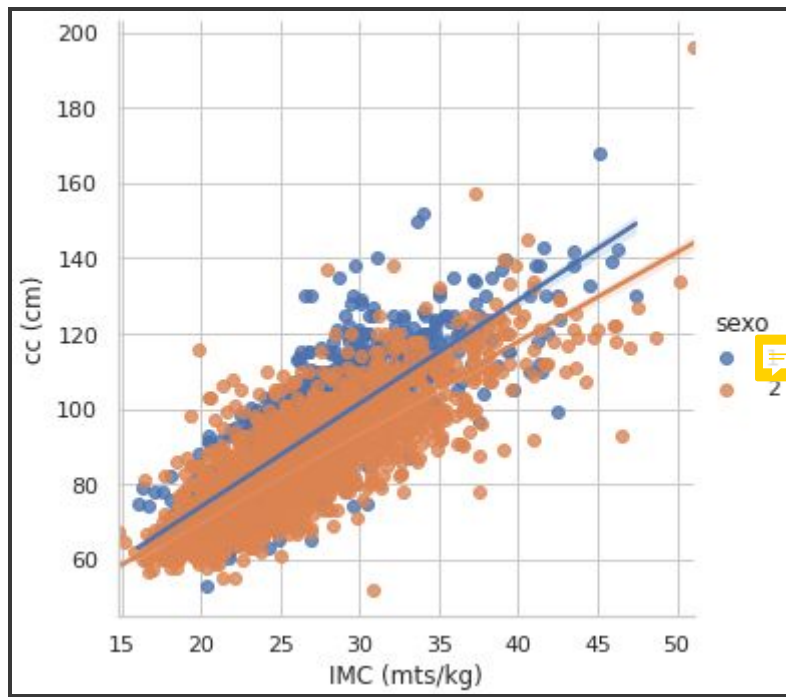


Fig. 9: Scatterplot diferenciado por sexo

Esta correlación fue significativa y positiva en individuos de ambos sexos, siendo ligeramente superior en los encuestados de sexo femenino (correlación para sexo femenino = 0.80; R sexo femenino = 0.82; Fig. 9).

3.2. Grasas y Tensión Máxima Promedio

Las relaciones entre variables que se presentan a continuación serán visualizadas a partir de mapas de calor. En todos los casos las variables fueron discretizadas en función de sus cuartiles, obteniéndose cuatro intervalos ordinales para cada una ellas. Por lo anterior, se utilizó el test de kendall para medir la correlación entre cada par de variables.

consumo de grasa en gramos por día vs tensión máxima promedio

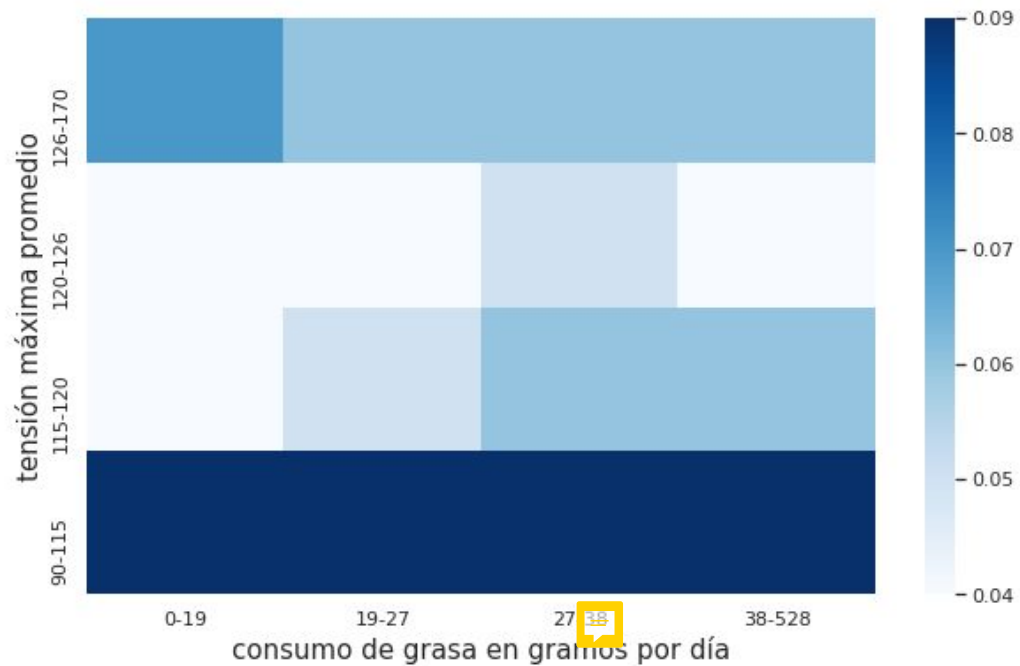


Fig. 10: Heatmap tensión máxima y consumo de grasas

La tensión máxima promedio y el consumo de grasa por día es son prácticamente independientes ya que presentan una correlación significativa pero muy baja (correlación de kendall = -0.03, valor p de la correlación = 0.026).

3.3. Edad y Tensión Máxima Promedio

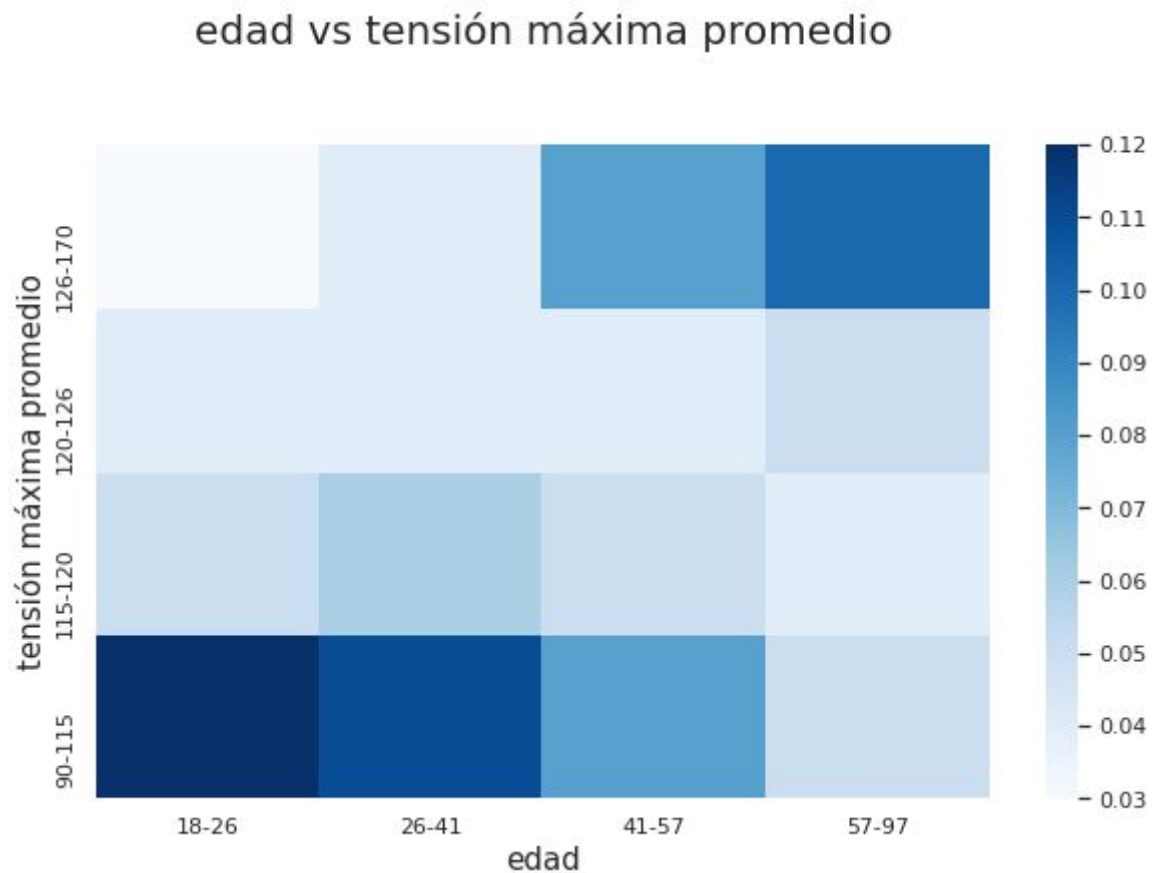


Fig. 11: Heatmap tensión máxima promedio y edad

La tensión máxima promedio y la edad presentan una correlación significativa baja (correlación de kendall = 0.23, valor p de la correlación = $2.02e-72$).

Por ejemplo, a medida que aumenta la edad de los individuos, disminuye la frecuencia de la tensión máxima promedio de la categoría 90-115. Por otro lado, mientras que los individuos jóvenes presenta tensiones promedio bajas (rango 90-115 de tensión vs rango de edad 18-26), los individuos de mayor edad presentan tensiones promedio más elevadas (rango 126-170 de tensión vs rango de edad 57-97).

3.4. Índice de Masa Corporal vs Consumo de Grasas

consumo de grasa en gramos por día vs índice de masa corporal

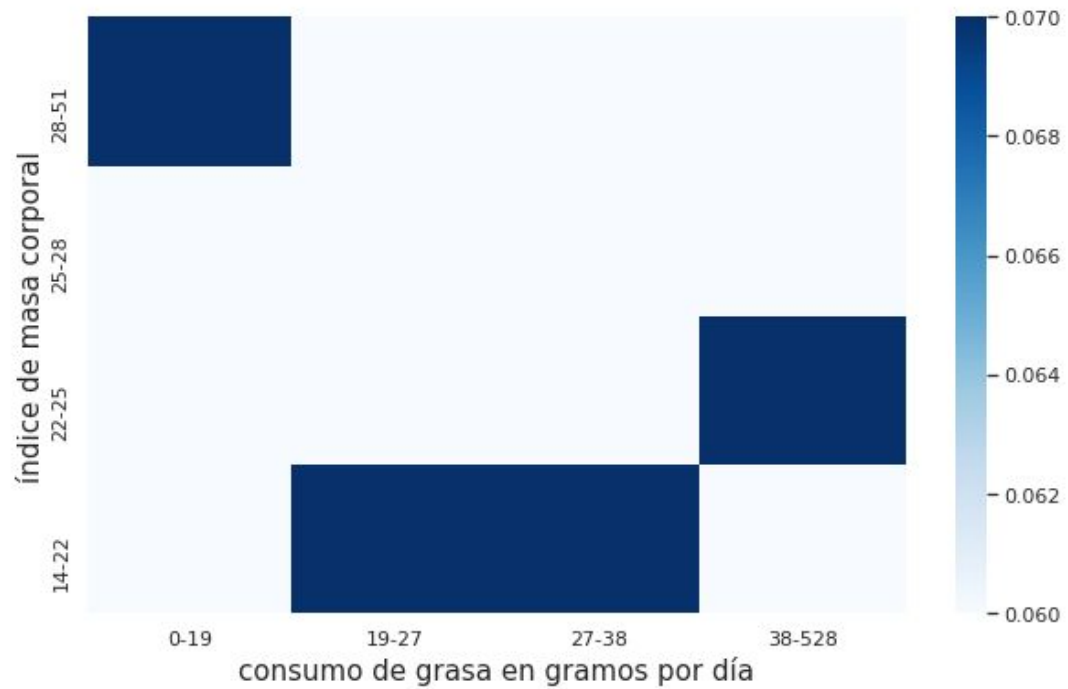


Fig. 12: Heatmap IMC y consumo de grasas.

El índice de masa corporal y el consumo de grasa por día son independientes (correlación de kendall = -0.004, valor p de la correlación = 0.77).

3.5. Edad vs Índice de Masa Corporal

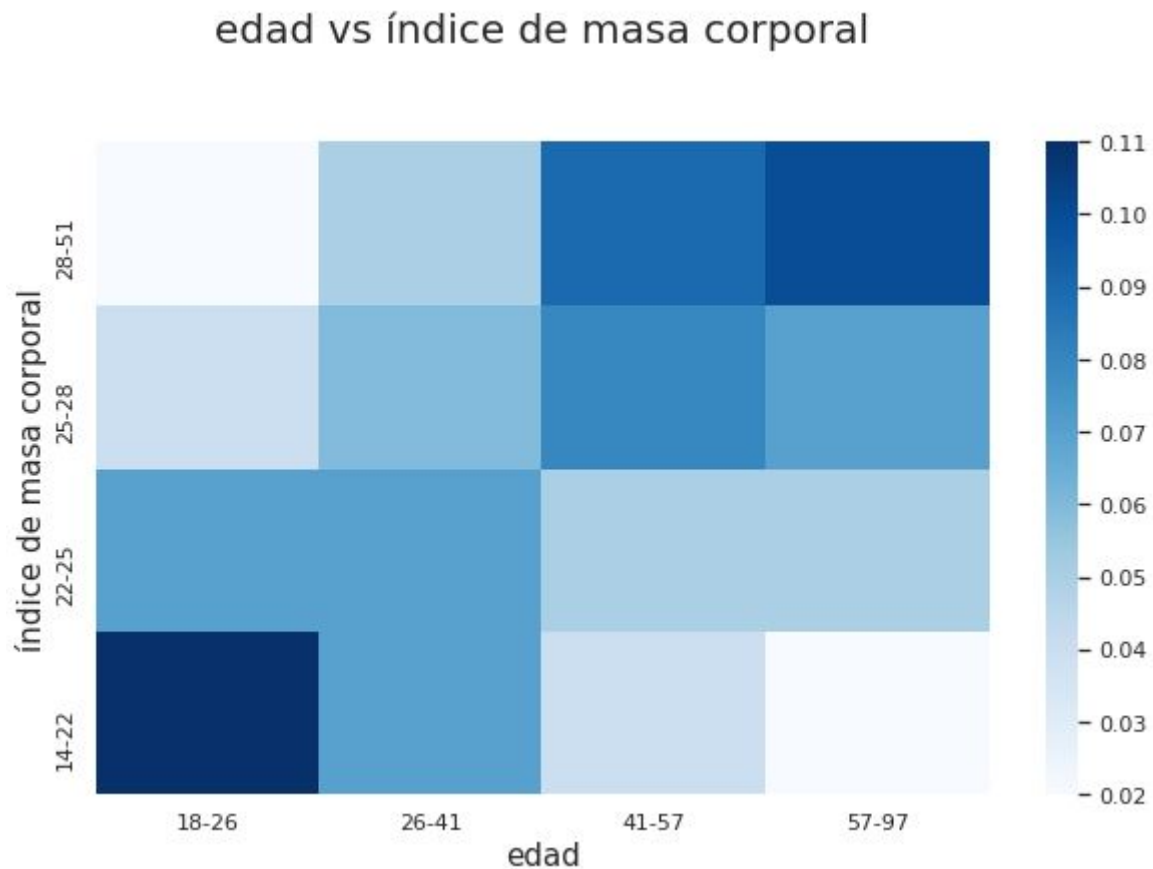


Fig. 13: Heatmap edad e IMC

El índice de masa corporal y la edad se encuentran positivamente correlacionadas (correlación de kendall = 0.33, valor p de la correlación = $1.45e-142$). Es decir, a medida que aumenta la edad, el índice de masa corporal también aumenta.

3.6. Edad vs Circunferencia de Cintura

edad vs circunferencia de cintura

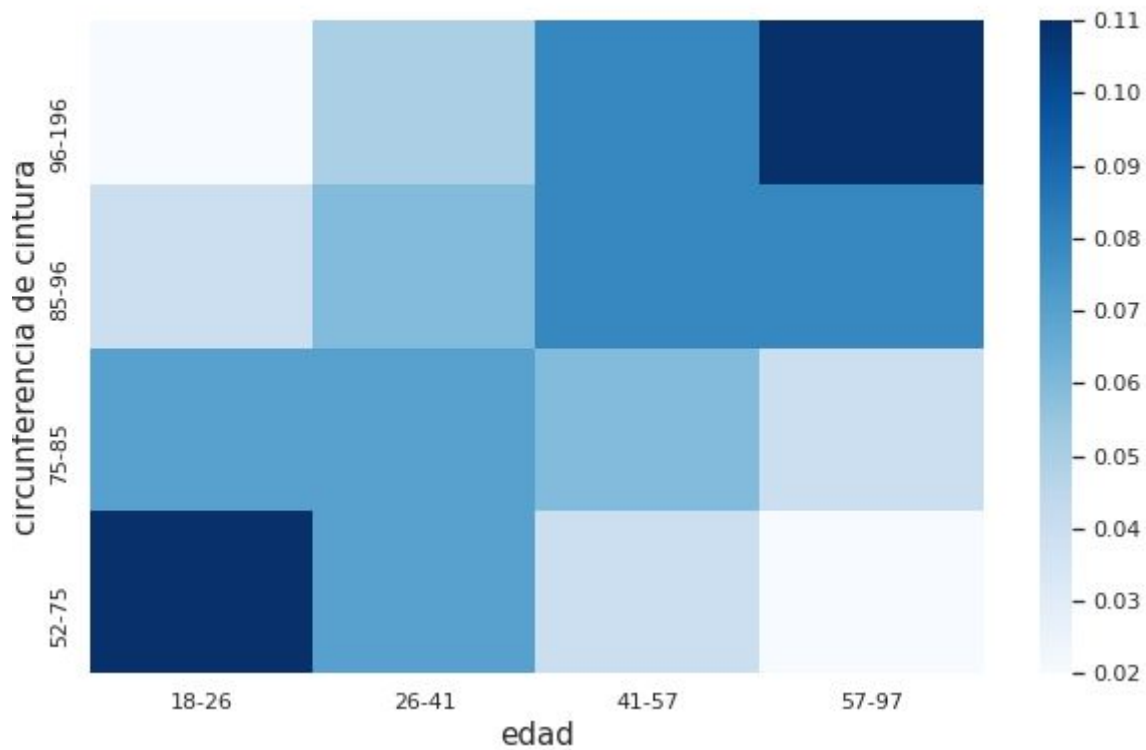


Fig. 14: Heatmap edad y circunferencia de cintura

Como se esperaba, la circunferencia de cintura y la edad también se encuentran positivamente correlacionadas (correlación de kendall = 0.37, valor p de la correlación = $3.28e-180$). Interesantemente, el valor de esta correlación es ligeramente superior al que existe entre el índice de masa corporal y la edad.

3.7. Edad vs Consumo de Grasa

edad vs consumo de grasa en gramos por día

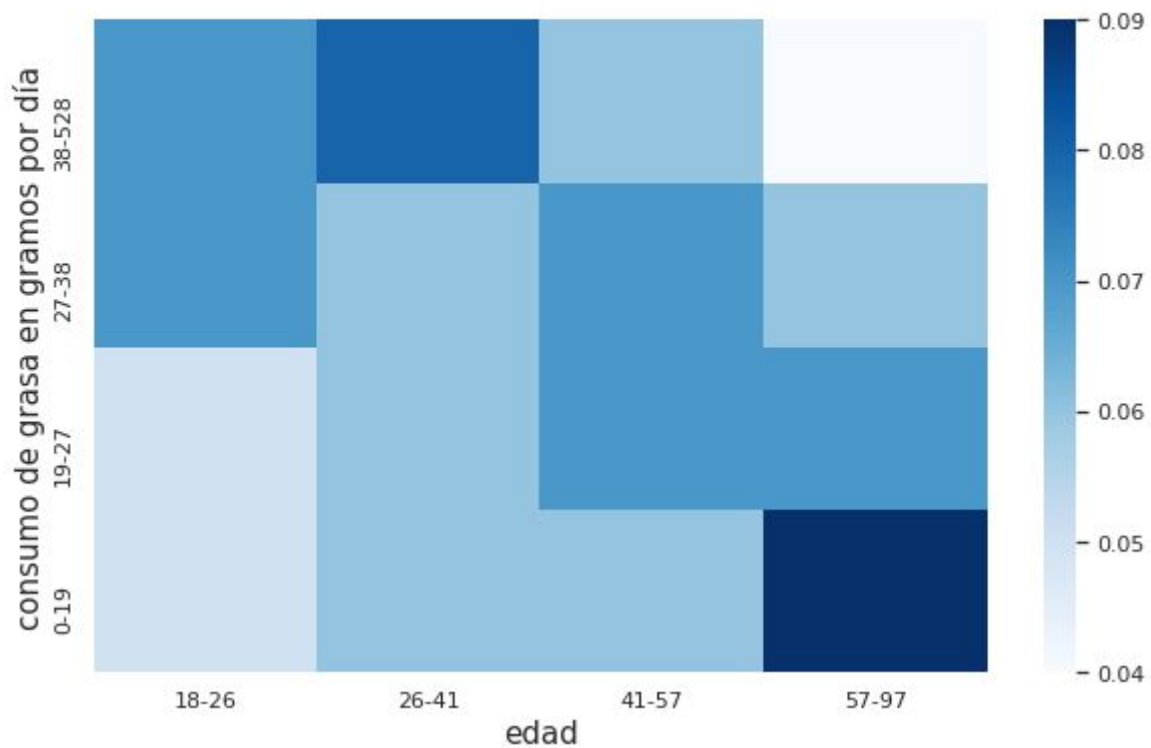


Fig. 15: Heatmap edad y consumo de grasa

El consumo de grasa por día se encuentra negativamente correlacionado con la edad (correlación de kendall = -0.14, valor p de la correlación = 8.09e-28). Por ejemplo, los individuos del rango de mayor edad presentaron menor consumo de grasa por día.

3.8. Índice de Masa Corporal vs Tensión Máxima Promedio

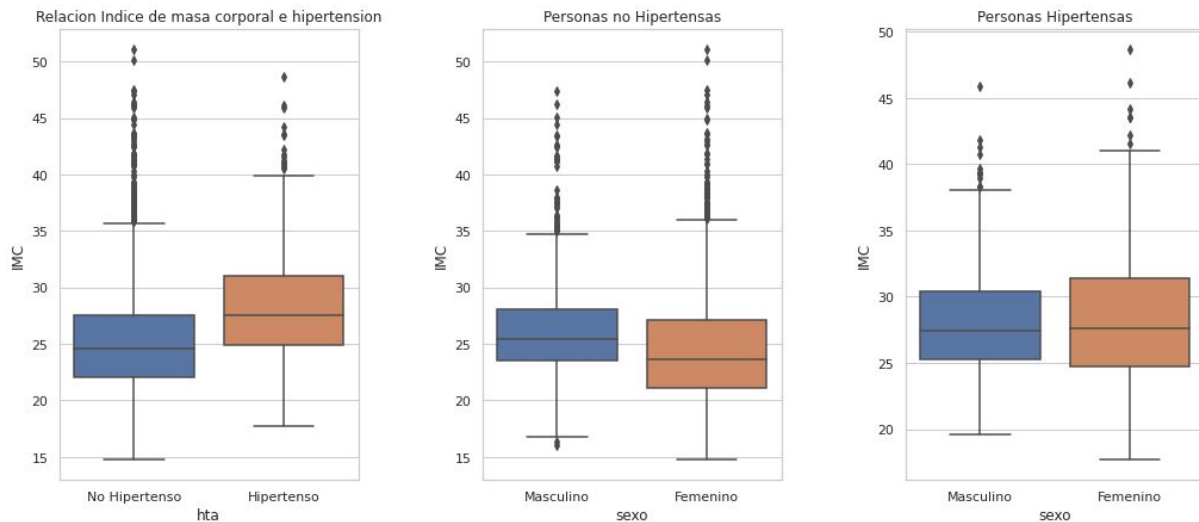


Fig. 16: Boxplot IMC y tensión máxima promedio

Al analizar la relación entre el índice de masa corporal (IMC) y la hipertensión arterial (HTA) se observa que los individuos que padecen HTA tienen mayores IMC comparados con los individuos no hipertensos. Teniendo en cuenta que el IMC en la muestra en estudio no posee una distribución normal, para realizar la comparación estadística se utilizó la prueba de Mann Whitney obteniéndose un p-valor < 0.05 (figura A). A su vez, entre los individuos no hipertensos, los de sexo femenino presentaron menores IMC comparados con los de sexo masculino (prueba de Mann Whitney con un p-valor < 0.05 , figura B). Por otro lado, en individuos hipertensos, no se encontraron diferencias para el IMC entre sexos (prueba de Mann Whitney con un p-valor = 0.81).

En ambos sexos, se detectaron outliers con valores extremos de IMC muy elevados, lo cual es consistente con lo descrito en relación a la Fig. 3B y 3D, en donde se observa que las categorías más frecuente de la distribución del IMC son “peso normal” y “preobesidad”.

3.9. Circunferencia de Cintura vs Tensión Máxima Promedio

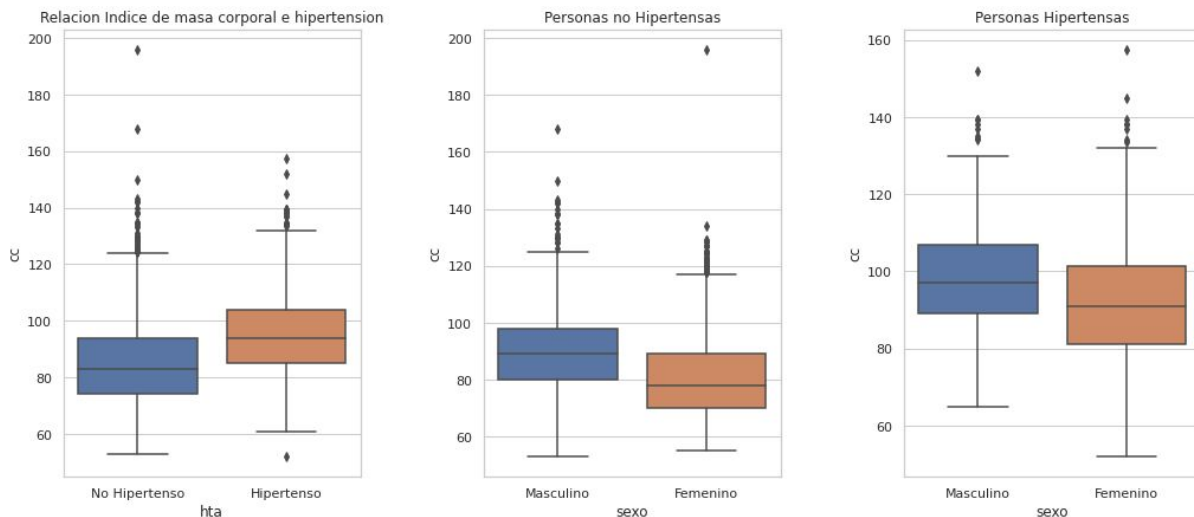


Fig. 17: Boxplot circunferencia de cintura y tensión máxima

Al analizar la relación entre circunferencia de cintura (CC) y la hipertensión arterial (HTA) se observa que los individuos que padecen HTA tienen mayores valores de CC comparados con los individuos no hipertensos. Teniendo en cuenta que el CC en la muestra en estudio no posee una distribución normal, para realizar la comparación estadística se utilizó la prueba de Mann Whitney obteniéndose un p-valor < 0.05 (figura A). A su vez, las mujeres no hipertensas e hipertensas presentaron menores valores de CC comparados con los varones no hipertensos e hipertensos respectivamente (prueba de Mann Whitney con un p-valor < 0.05 en ambos análisis, figuras B y C).

3.10. Nivel de Actividad Física, Índice de Masa Corporal y Circunferencia de Cintura

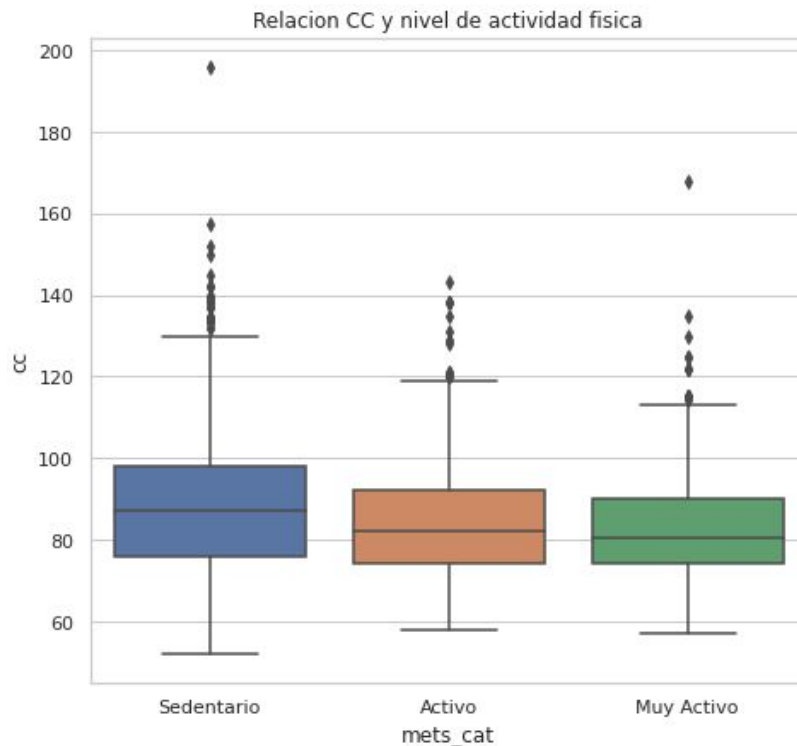


Fig. 18: Boxplot nivel de actividad física y circunferencia de cintura

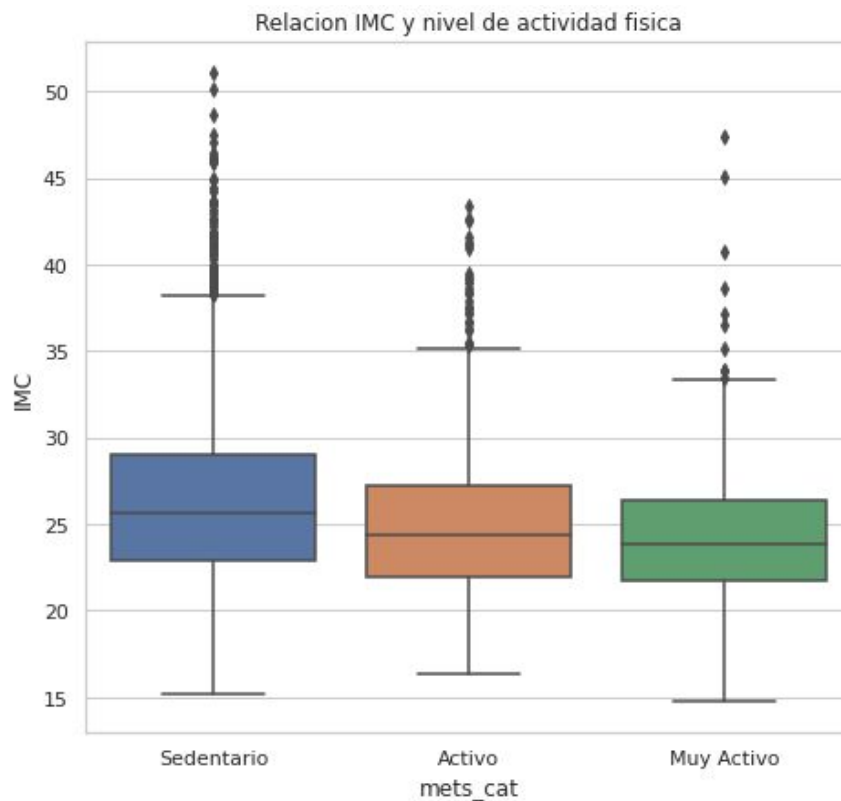


Fig. 19: Boxplot IMC y nivel de actividad física

Al analizar la relación entre la actividad física (mets_cat) con la circunferencia de cintura (CC) se observa que los individuos activos y muy activos (mets entre 600 - 1499 y mets > 1500 respectivamente) tienen menores valores de CC comparados con los individuos sedentarios. Teniendo en cuenta que los valores de CC en la muestra en estudio no posee una distribución normal y que se realizó una comparación de tres grupos independientes, para realizar la comparación estadística se utilizó la prueba de Kruskal-Wallis obteniéndose un p-valor < 0.05 (figura A). Para identificar las diferencias individuales entre los grupos se realizó una prueba a posteriori de Mann Whitney con corrección de Bonferroni observándose diferencias significativas entre el grupo sedentario con los grupos activo y muy activo y sin diferencias entre el grupo activo y muy activo.

Por otro lado, la comparación entre la actividad física y el IMC arrojó resultados similares a los anteriormente descritos con un p-valor en la prueba de kruskal-wallis < 0.05 mostrando diferencias significativas entre el grupo sedentario con los grupos activos y muy activos en la prueba a posteriori de Mann Whitney (valor p < 0.05).

