

Instituto Tecnológico y de Estudios Superiores de Monterrey

Arturo Caballero Ortega

A01369947

05/11/2020

Proyecto de Ciencia de Dato

Índice

[Introducción 3](#_Toc56073623)

[Fase 1. Entendimiento del negocio 4](#_Toc56073624)

[Fase 2. Entendimiento de los datos 5](#_Toc56073625)

[Fase 3. Preparación de los datos 6](#_Toc56073626)

[Fase 4. Modelación de los datos 7](#_Toc56073627)

[Fase 5. Reflexión Final 8](#_Toc56073628)

[Anexos 9](#_Toc56073629)

[Referencias 16](#_Toc56073630)

# Introducción

En los últimos años la ciencia de datos ha tomado un mayor auge en México y en el mundo. Este es un instrumento de gran uso para las empresas principalmente por el análisis de información para usos de venta de mercado.

La ciencia de datos es una disciplina conformada por estadística, métodos científicos y análisis de datos para saber cómo interpretar los datos y representar dichos datos para darle un uso. Esto quiere decir que al usar un banco de datos la ciencia de datos permite realizar análisis estadísticos siguiendo una metodología científica con herramientas matemáticas y computacionales para poder procesar la información que se tiene y así darle un sentido que le permita a las empresas, organizaciones o grupo de personas que les permita entender y predecir la información que tienen lo que les permite tomar decisiones.

La intención de este proyecto es darle a conocer al estudiante los usos que tienen la ciencia de datos para darle la información para la toma de decisiones. Este proyecto se enfoca en poderle mostrar al alumno como es la tendencia de alimentación y a través del análisis de los datos poder entender como es la alimentación y la predicción en el consumo de calorías y de nuestra alimentación.

# Fase 1. Entendimiento del negocio

Para la realización de un proyecto de negocio con el análisis de datos para una toma de decisiones es importante tener claro la situación de los objetivos que se quiere llegar a realizar, es preciso que se tengan datos que nos permita realizar una base por donde comenzar a realizar un estudio dependiendo de lo que se quiere llegar.

Para poder realizar este proyecto se necesita tener claro cuál es el objetivo que se intenta llegar a predecir, en este caso para este proyecto se necesita intentar predecir el consumo de calorías en un alimento para predecir una alimentación correcta. Una vez identificada claramente que es lo que se intenta obtener como resultado en el trabajo o proyecto se necesitará tener los datos con los que se van a trabajar, para este caso va a ser necesario conocer de preferencia el origen de los datos y todos los detalles de estos, como es la forma en cómo se obtuvo la información, la manera en cómo se clasifica, entre otros datos para luego el grupo que tendrá que analizar los datos pueda realizar su trabajo.

Para esta parte es necesario que se identifiquen y dividan las diferentes tareas que se van a realizar, así como tener una planeación bien desarrollada sobre el desarrollo del trabajo, así como tener un criterio para poder decir que los datos obtenidos están entre lo esperado o que los resultados se encuentren bajo un criterio de discriminación que les permita tener un criterio de la calidad de los datos para su manejo.

Luego de tener todas las condiciones previas listas se pude continuar con la parte realizar el trabajo con las herramientas necesarias para conseguir la solución a nuestra problema o necesidad.

Para este caso se sabe que lo que se intenta es saber el consumo alimenticio del sujeto en este caso seré yo Arturo Caballero Ortega el cliente, así como el profesor que imparte la optativa quien podrá visualizar y cuestionar mis datos y resultados para la obtención de lo requerido en el criterio de evaluación de esta asignatura.

Se identifica como el reto de esta asignatura sobre la predicción alimentaria del estudiante por lo que la variable dependiente o variable a predecir es la cantidad del consumo de calorías en cada alimento.

Para esta parte del proyecto se continua a realizar un análisis estadístico de toda la información identificando las variables independientes y como estas afectan a la variable dependiente, luego en caso de que se identifiquen variables o datos que afecten o no se encuentre una relación con la variable dependiente se proseguirá a excluir del modelo planteado los datos que en este no se encuentre relación. Por lo que se necesitará realizar la desviación estándar, un análisis de regresión, obtener una distribución normal, realizar una gráfica como la campana gaussiana, saber que son las variables y diferentes tipos de variables, se necesita saber cómo tratar los errores, como realizar una regresión lineal, el planteamiento de una hipótesis y como negarla o aceptarla, así como el análisis de una tabla ANOVA y validar el modelo de mi hipótesis.

Durante el desarrollo del trabajo primero se necesitará registrar la alimentación diaria durante todas las semanas en que se lleve a cabo este trabajo, luego esos datos van a ser copiados para trabajar con una copia y no con los originales, luego se realizaran los diferentes análisis descritos previamente para así obtener un resultado, en caso de ser necesario repetir el mismo procedimiento cuantas veces sea necesario para así obtener un mejor resultado.

Todo esto es para que la con el uso de la ciencia de datos pueda analizar y entender el comportamiento de mi alimentación y predecir la forma en cómo me alimento, como identificar el nivel de calorías de cada alimento.

# Fase 2. Entendimiento de los datos

En el proyecto he utilizado datos provenientes de la información nutrimental de cada alimento que he comido desde hace varias semanas proporcionadas por la aplicación Samsung Healt la cual contiene datos sobre varios alimentos y se ha usado como punto de referencia sobre cada alimento que he comido pues me he visto con la obligación de considerar estos datos no tan certeros ya que es una información genérica sobre lo que contiene cada alimento en cierta cantidad de masa o que la forma en cómo se prepararon los alimentos puede variar y con eso cambiar los valores de cierta manera de las variables a considerar de cada alimento para este trabajo. Por lo que se ha proseguido de una manera en considerarlos no tan certeros, pero sí aproximados a lo real. En este trabajo los datos manejados representan la información nutrimental cómo las calorías, carbohidratos, los lípidos, las proteínas y el sodio de cada uno de los alimentos consumidos cada día de la semana durante los últimos meses, por lo cual se ha considerado actualizar la tabla cada semana de la información de cada alimento consumido en cada día el domingo de cada semana por lo cual se van a seguir registrando alimentos semanalmente hasta la fecha de conclusión de registro de datos de este proyecto de investigación.

Las columnas de la base de datos que se lleva registrando parecen más prometedoras las de carbohidratos, lípido y proteína por la relación que tienen y que se ha presentado durante el análisis de los datos que son los que tienen una mayor relación con la variable dependiente que es calorías y la variable que al parecer podemos descartar de nuestro modelo es el sodio.

En todo estudio estadístico es siempre mejor tener la mayor cantidad de datos posibles para poder realizar un modelo más preciso que permita explicar el comportamiento de algún fenómeno por lo que con una mayor cantidad de datos más exacto es el modelo y en este caso con la cantidad de datos que se tienen son una buenos para realizar conclusiones y poder explicar cuan posibles es que algo ocurra, pero siempre con más datos es mejor.

En cuanto a los atributos que hay para realizar un modelo en este caso se ha reflejado que con los que se manejan han sido muy buenos para realizar el trabajo ya que han mostrado una gran relación en el análisis y entendimiento de las calorías que hay en cada alimento por lo que ha sido más que suficiente para este modelo en lo que se ha visto reflejado.

Todos los datos obtenidos provienen de la misma fuente que es Samsung Healt y son actualizados cada semana, todos los datos vienen completos. Hasta el momento la cantidad de datos a los que se puede acceder son 190 y algunos son con valores aproximados a lo que realmente se ha estado consumiendo.

Una de las hipótesis puede ser que si falta una de las variables ya escogidas disminuye la calidad en la predicción de las calorías.

# Fase 3. Preparación de los datos

La fusión de datos no fue necesario en el punto de vista que se sólo se usó una aplicación para la recopilación de datos y que en el procesos se usaron los mismos estándares de evaluación para el registro de los valores de los alimentos, por otra parte se podría considerar que la aplicación Samsung Healt generó diferentes archivos csv para registrar los datos por lo que en el proceso se tuvo que relacionar los datos de los diferentes archivos para poder registrar los alimentos con sus respectivos valores de información nutrimental ya que todo la información estaba con claves y se tuvo que procesar todos los datos.

No hubo necesidad de realizar subconjuntos ya que toda la información que se presenta es de la misma agrupación de datos y toda está en la misma sintonía para realizar este trabajó por lo que lo justifico en que el proyecto necesita datos del registro nutrimental como el alimento calorías, carbohidratos, sodio, lípidos, proteínas y todos esos datos son parte de la información de los alimentos por lo que los datos con los que se trabajan son del mismo conjunto de datos ya que es alimento con información del alimento.

En cuanto a los datos se puede agregar más datos ya que requiere un registro de alimentos diarios por lo que se puede adicionar la cantidad de datos con su respectiva información cuanto sea necesario ya que una ventaja de esto es que va a adquirir una mayor precisión nuestro análisis al poder tener más datos con que trabajar porque tienen más información para analizar.

No hubo necesidad en realizar cambios en atributos de los datos ya que todos contienen la información necesaria para trabajar y todos los atributos de todos los datos fueron obtenidos de la misma aplicación con los mismos entandares por lo que se asegura una relación correcta entre los atributos con los datos.

No hubo necesidad ordenar los datos para el análisis en general a excepción en la obtención de la moda y mediana estadística donde sí se necesitaba tener el orden de mayor a menor de cada uno de los atributos de los datos.

Los valores en blanco que se tuvieron que eliminar fueron con respecto al agua ya que este no proporcionaba información que ayudara al análisis de los datos por lo que quedó fuera de la realización del trabajo.

# Fase 4. Modelación de los datos

En la realización de este trabajo se tuvo que realizar varias veces todos los procesos durante el desarrollo hasta esta entrega final en la cual se tuvo que realizar los mismos procedimientos con las gráficas para marcar la regresión lineal considerando cada una de las variables con las que se decidió trabajar por lo que durante el trabajo general se realizó por lo menos unas 10 veces y en esta última parte se tuvo que repetir unas 5 veces.

Podemos darnos cuenta de que los lípidos afectan en gran medida a la variable dependiente, seguido de las proteínas, los carbohidratos y al final la constante.

Podemos percatarnos con la gráfica que todos los datos tienden a irse a la izquierda a excepción de unos que se separan de la tendencia y que tienen relación con los alimentos chatarra. A diferencia del análisis inicial mi alimentación era muy variada, pero en el último mes he estado repitiendo alimentos de forma sincronizada semanalmente y por días por lo que se ve muy reflejada la gráfica actual con la inicial.

# Fase 5. Reflexión Final

La ciencia de datos es una disciplina que ha demostrado ser de gran utilidad al poder ofrecer herramientas que permiten entender el comportamiento de nuestro entorno o del comportamiento en una actividad especifica que puede ser necesidad del cliente. Tiene una gran utilidad en la vida diaria al utilizar herramientas matemáticas y estadísticas para poder cuantificar los fenómenos que se presentan y con estos realizar modelos matemáticos y estadísticos que nos permitan entender la tendencia en el comportamiento de este fenómeno y explicar porque se presenta cierta cosa y con base en todos estos resultados poder tomar una decisión que sea conveniente o la más acertada para nosotros o nuestros clientes.

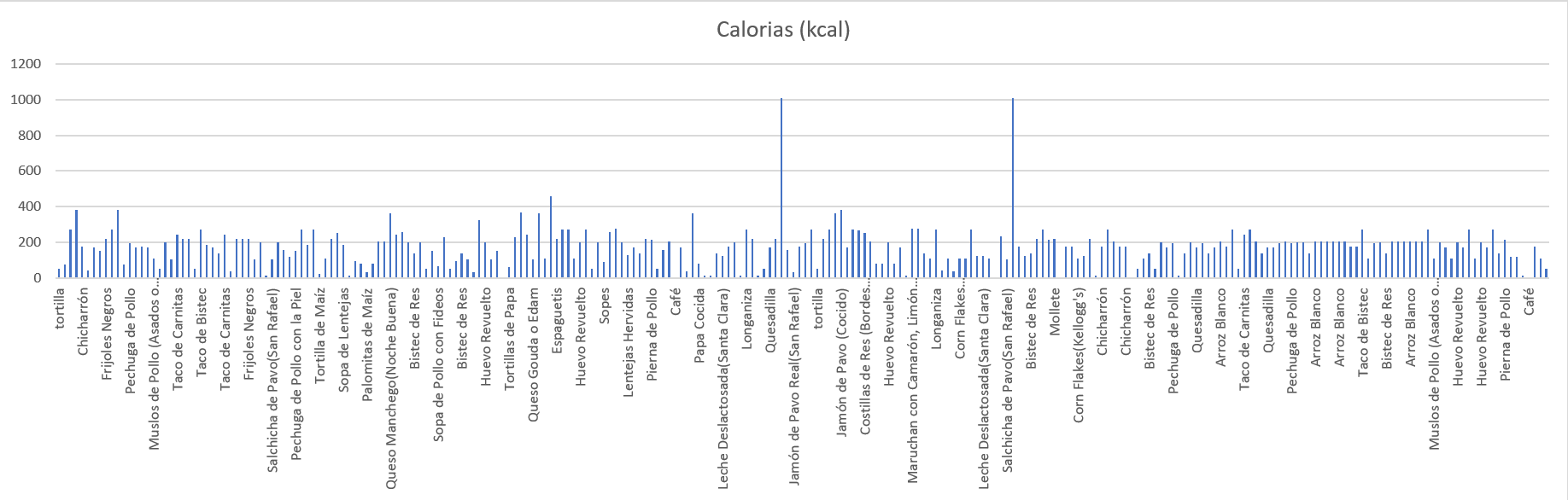
La ciencia de datos al final nos ofrece herramientas que nos permiten entender la información que se está recopilando, así como descartar grupos de datos que no muestren una relación a los objetivos que se propongan en el modelo o en el propósito del proyecto. También con el uso de las herramientas computacionales se puede realizar este trabajo de una manera más rápida y eficiente ya que se puede automatizar los procesos facilitando el análisis y la interpretación de los datos.

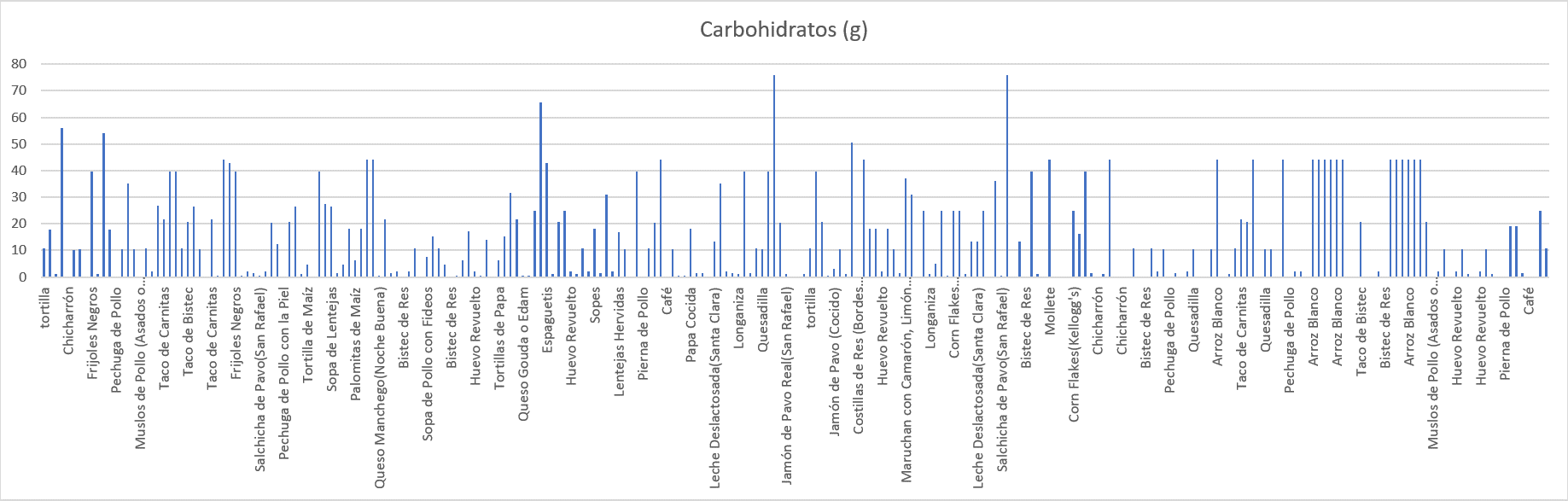
La ética en la ciencia de datos es muy importante pues el manejo de la información delicada se vuelve algo muy peligroso al no usarla con responsabilidad o con fines malos. La información a la que se llega a tener acceso permite tener una muestra en el comportamiento en los individuos por lo que se puede llegar a entender que le gusta y hasta como piensa una persona por lo que se le puede empezar a ofrecer cierta información como el caso mostrar información a favor o en contra de un candidato a la presidencia, así como lucrar indiscriminadamente con la información que se recopila de las personas violando la privacidad de los individuos.

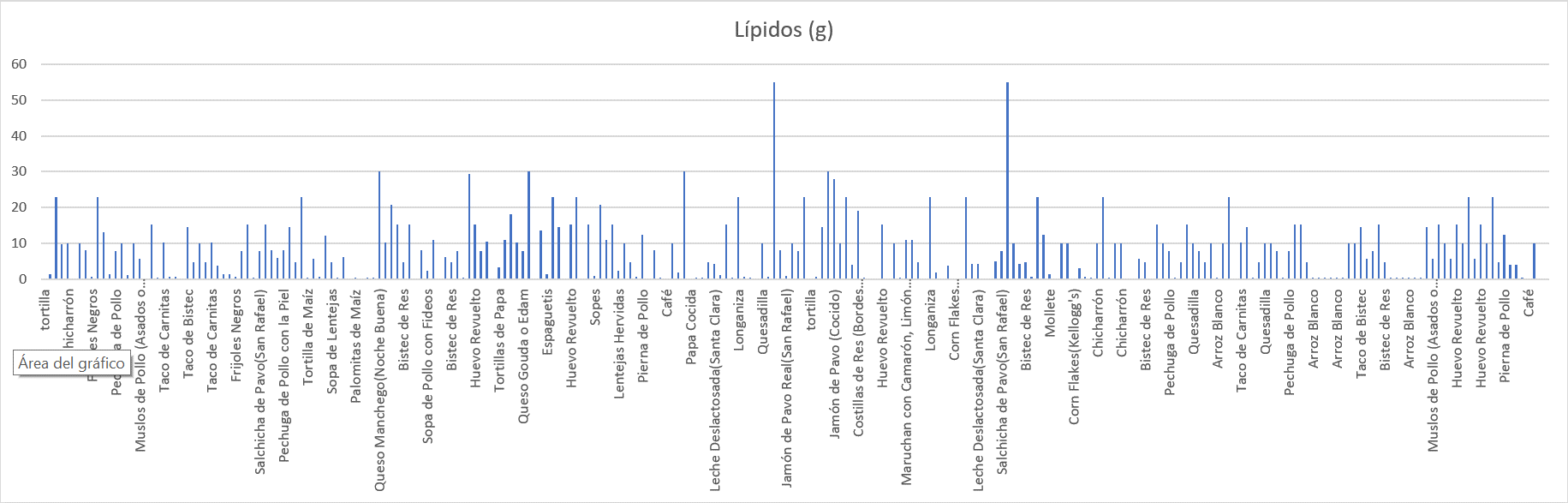
La ciencia de datos es una gran herramienta que nos facilita en la comprensión de los fenómenos que queramos estudiar pero su mal uso puede llegar a vulnerar la privacidad de los individuos y alterar la opinión pública así como poner en riesgo la libertad de las personas por lo que hay que ser conscientes en como usamos los datos que poseemos para un beneficio personal o moral así como para poderle brindar beneficios a las personas de las que se recopila los datos.

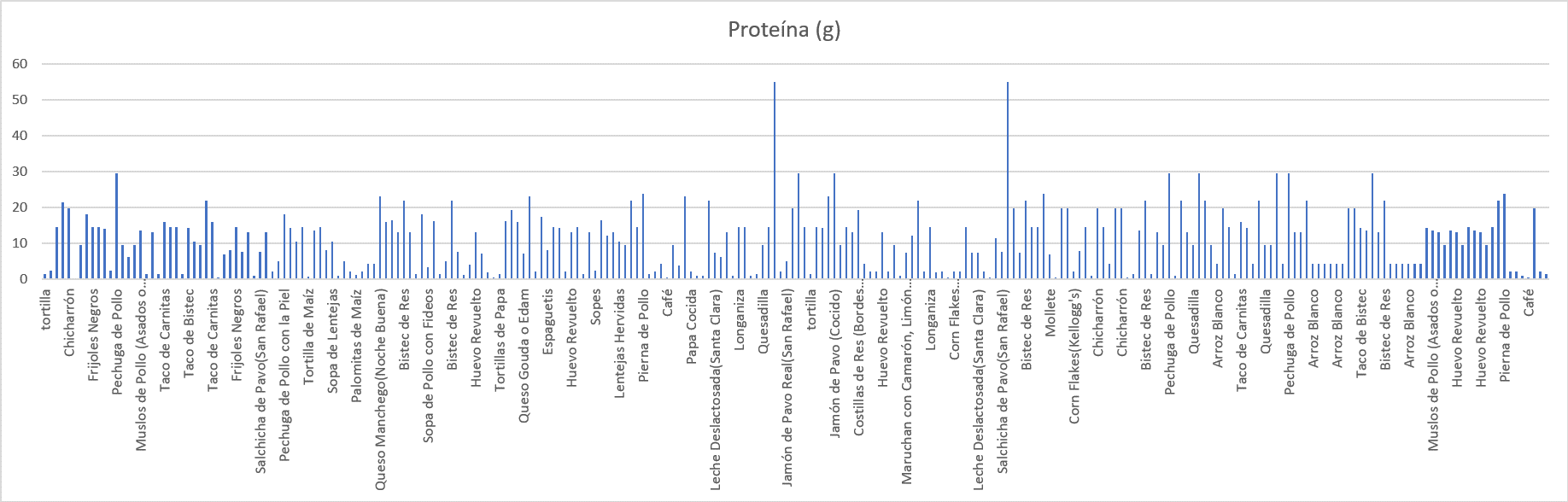
# Anexos

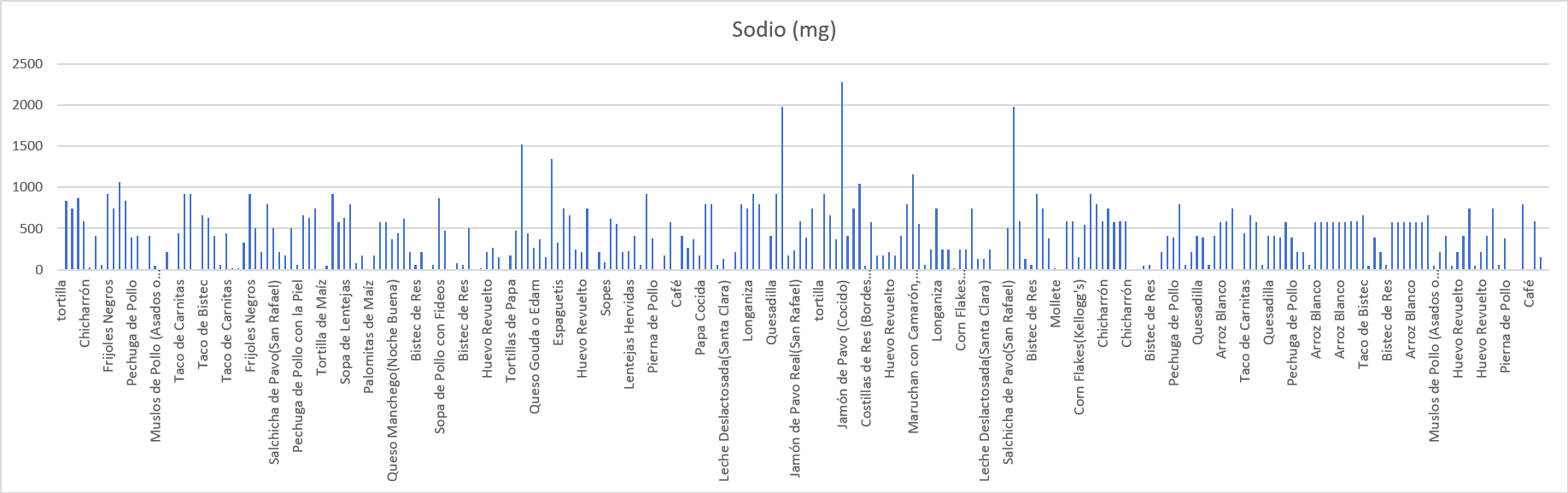
Se muestran los histogramas de cada variable:











A continuación, se mostrará el código en Python para realizar este trabajo:

Este código muestra la gráfica de Gauss

import statistics as st

import numpy as np

import matplotlib.mlab as mlab

from scipy.stats import norm

import sys

df = pd.read\_csv("A01369947\_residuo.csv")

df.head(5)

data = df['Residuos'].values.tolist()

print(data)

mean,std=norm.fit(data)

print(mean,std)

plt.hist(data, bins=20, density=True, alpha=0.6, color='g')

xmin, xmax = plt.xlim()

x = np.linspace(xmin, xmax, 100)

p = norm.pdf(x, mean, std)

plt.plot(x, p, 'k', linewidth=2)

title = "Fit results: mu = %.2f, std = %.2f" % (mean, std)

plt.title(title)

print(x)

plt.show()

#sys.exit()

#plt.hist(data, bins=30, normed=True)

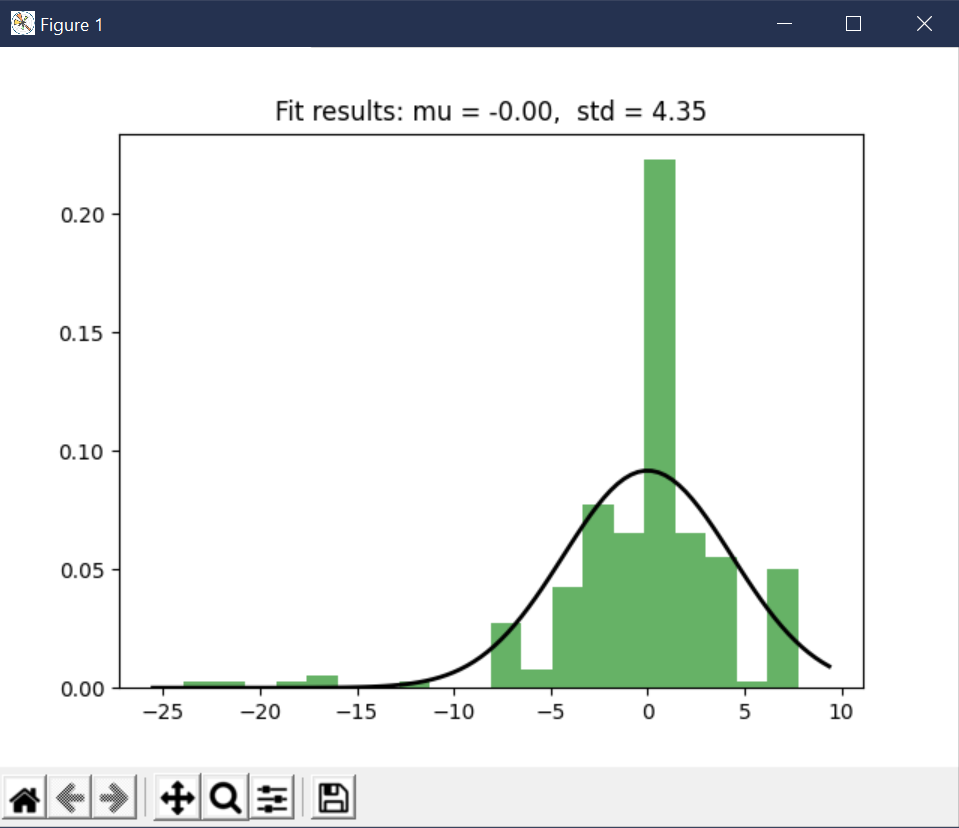
xmin, xmax = plt.xlim()

x = np.linspace(xmin, xmax, 100)

y = norm.pdf(x, mean, std)

plt.plot(x, y)

plt.show()



Para el análisis de regresión:

import pandas as pd

import statsmodels.api as sm

import statsmodels.formula.api as smf

import matplotlib.pyplot as plt

from sklearn.linear\_model import LinearRegression

dt = pd.read\_csv("A01369947\_EtiquetasNutrimentales.csv")

calorias = dt["Calorias (kcal)"]

carbohidratos = dt["Carbohidratos (g)"]

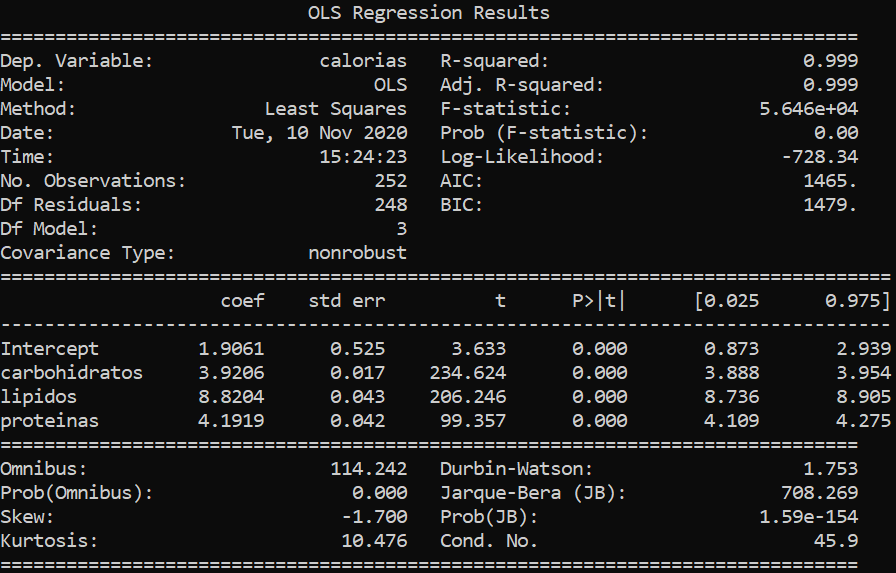
lipidos = dt["Lípidos (g)"]

proteinas = dt["Proteína (g)"]

#sodio = dt["Sodio (mg)"]

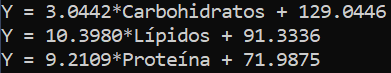
mod = smf.ols('calorias ~ carbohidratos + lipidos + proteinas', data=dt).fit()

print(mod.summary())



Para generar las gráficas de regresión de todas las variables:

Se realizó la regresión lineal de cada variable dando como resultado



# Para predecir el comportamiento de los Carbohidratos

x = carbohidratos.values.reshape(-1, 1)

y = calorias.values.reshape(-1, 1)

linear\_regression = LinearRegression().fit(x,y)

pred = linear\_regression.predict(x)

m = linear\_regression.coef\_[0][0]

c = linear\_regression.intercept\_[0]

label = r'Y = %0.4f\*Carbohidratos + %0.4f'%(m,c)

print(label)

plt.scatter(carbohidratos,calorias)

#plt.plot(carbohidratos, calorias, label = "Calorias")

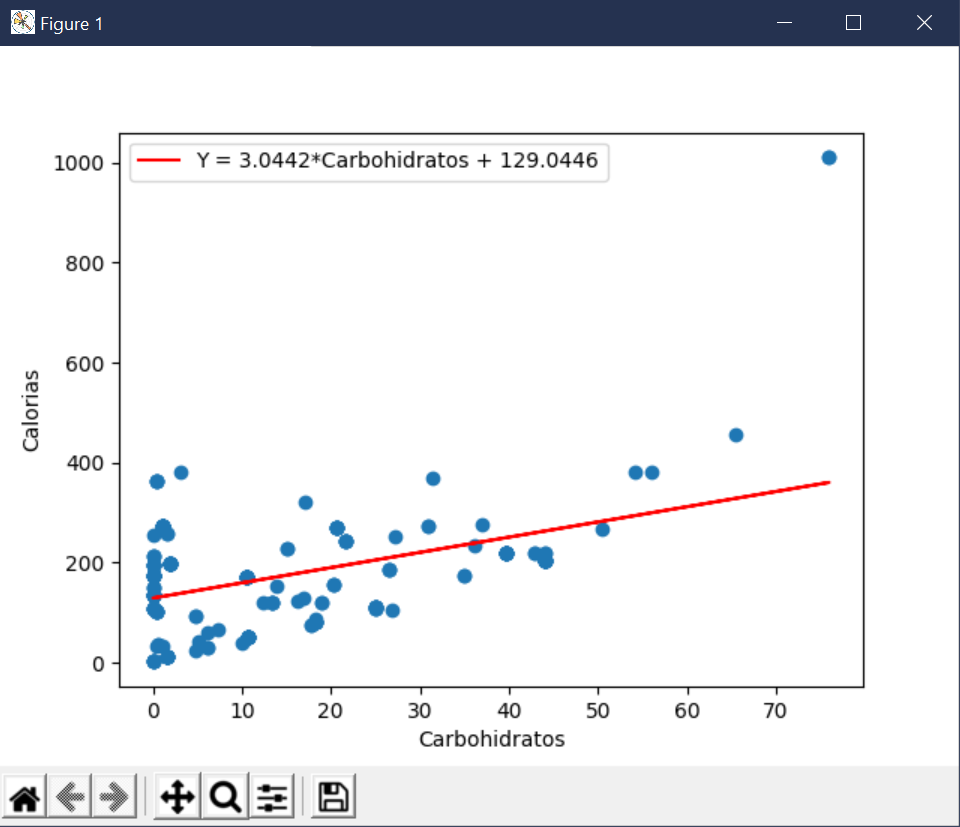
plt.plot(x, pred, color = 'red', label = label)

plt.xlabel("Carbohidratos")

plt.ylabel("Calorias")

plt.legend()

plt.show()



# Para predecir el comportamiento de los Lípidos

x = lipidos.values.reshape(-1, 1)

y = calorias.values.reshape(-1, 1)

linear\_regression = LinearRegression().fit(x,y)

pred = linear\_regression.predict(x)

m = linear\_regression.coef\_[0][0]

c = linear\_regression.intercept\_[0]

label = r'Y = %0.4f\*Lípidos + %0.4f'%(m,c)

print(label)

plt.scatter(lipidos,calorias)

#plt.plot(lipidos, calorias, label = "Calorias")

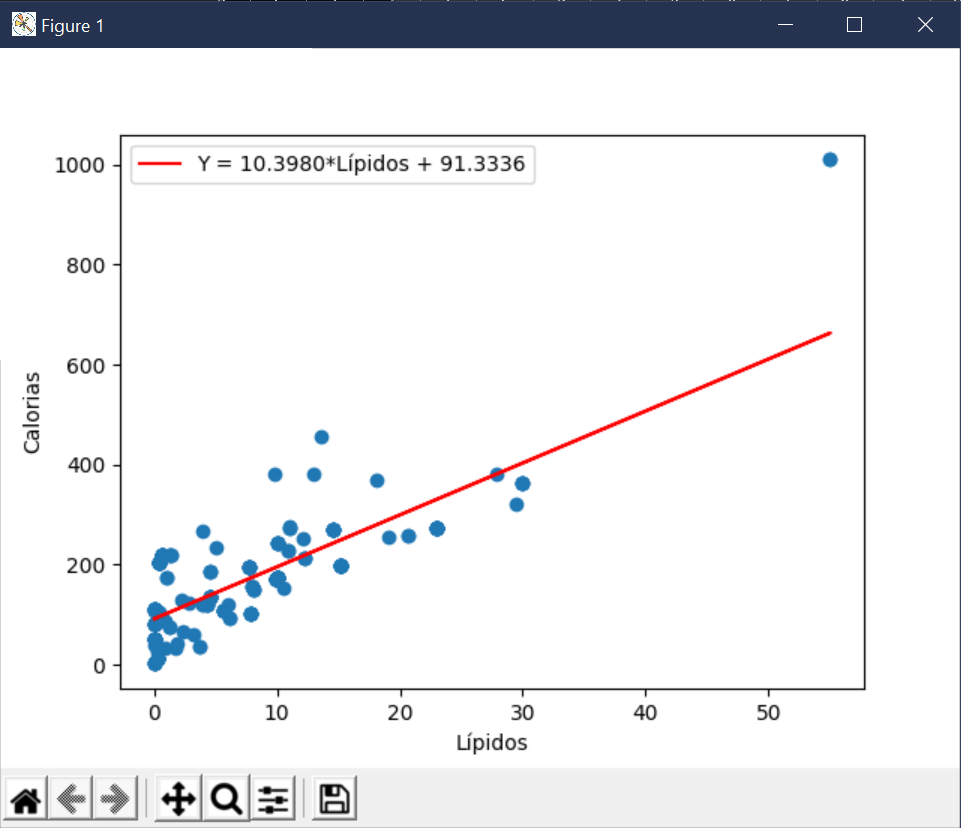
plt.plot(x, pred, color = 'red', label = label)

plt.xlabel("Lípidos")

plt.ylabel("Calorias")

plt.legend()

plt.show()



# Para predecir el comportamiento de la Proteína

x = proteina.values.reshape(-1, 1)

y = calorias.values.reshape(-1, 1)

linear\_regression = LinearRegression().fit(x,y)

pred = linear\_regression.predict(x)

m = linear\_regression.coef\_[0][0]

c = linear\_regression.intercept\_[0]

label = r'Y = %0.4f\*Proteína + %0.4f'%(m,c)

print(label)

plt.scatter(proteina,calorias)

#plt.plot(proteina, calorias, label = "Calorias")

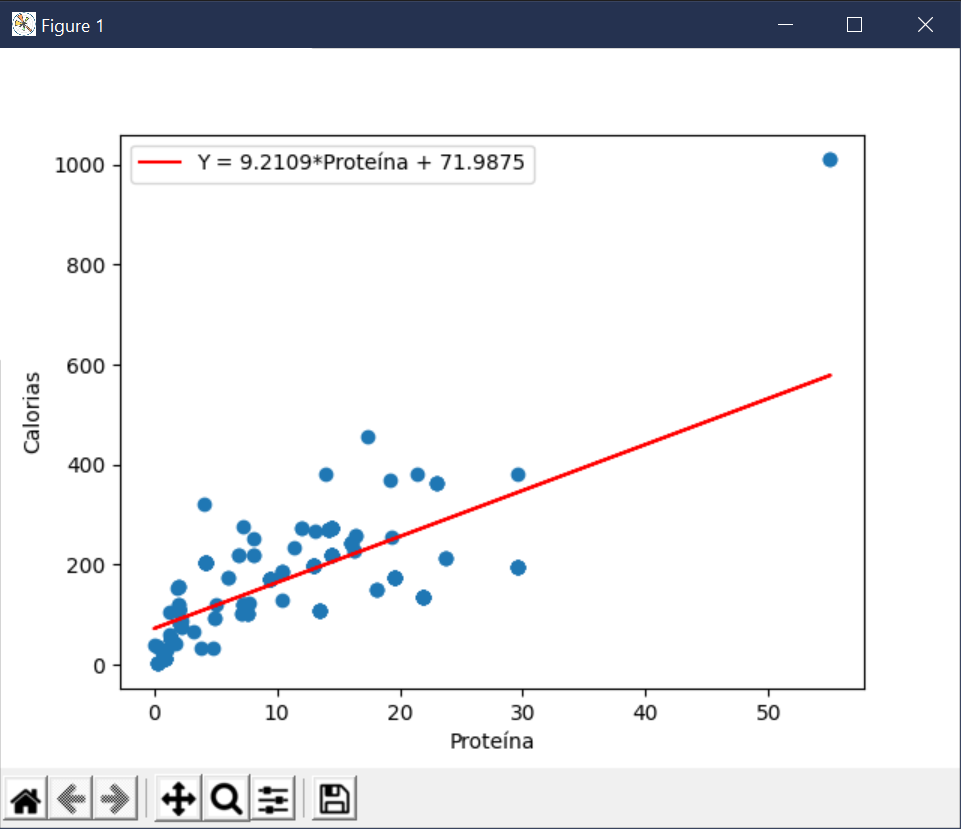
plt.plot(x, pred, color = 'red', label = label)

plt.xlabel("Proteína")

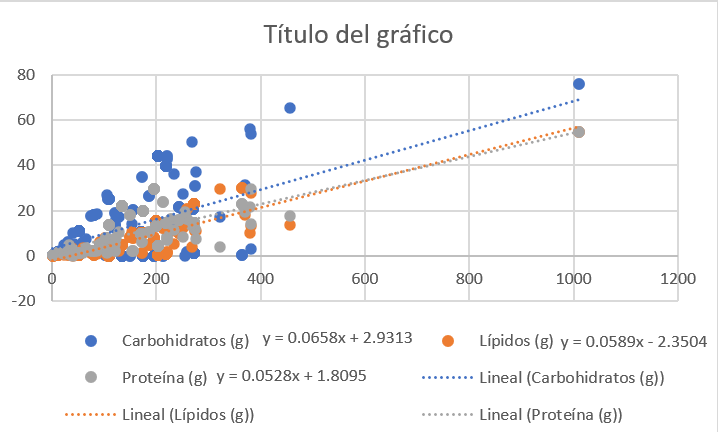
plt.ylabel("Calorias")

plt.legend()

plt.show()



A continuación, se muestran las gráficas resultantes del análisis de cada variable usando Excel:



# Referencias

Seabold, Skipper, and Josef Perktold. “[statsmodels: Econometric and statistical modeling with python.](http://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf)” Proceedings of the 9th Python in Science Conference. 2010.

Enlace a la tabla de alimento: <https://drive.google.com/file/d/1kxhkSoTMyLuctu-CgGRvMj8xHigInIU-/view?usp=sharing>