U N I V E R S I D A D D E P I U R A

Curso Análisis de Datos con Python - Nivel 1

Profesor:

Ing. Pedro Rotta.

Integrantes:

Paz Echevarría, Luis.

Paz Echevarría, Nathali.

Porras Talledo, Fabricio Sebastián.

Piura, 23 de enero del 2022

1. Introducción

La cerveza es una de las bebidas más democráticas y consumidas del mundo. No sin razón, es perfecto para casi todas las situaciones, desde pequeñas reuniones hasta grandes bodas. En el 2015 la bebida antes mencionada obtuvo alrededor de 10.841 millones de litros en venta en Brasil. En el 2019 fue la bebida alcohólica más vendida. Ese año, se comercializaron alrededor de 12.600 millones de litros de dicha bebida en el país sudamericano, lo que la convierte en la bebida alcohólica preferida de los brasileños. Bebidas destiladas tales como el vodka, el whisky, la cachaca y la ginebra ocupaban la segunda posición, con un volumen de ventas de más de 700 millones de litros. Además, las ventas de vino ascendían a cerca de 330 millones de litros. Por su parte, los refrescos fueron el tipo de bebidas sin alcohol con mayor volumen de ventas al por menor en el país ese año.

El objetivo de este trabajo será demostrar los impactos de las variables sobre el consumo de cerveza en una determinada región y la previsión de consumo para determinados escenarios. Los datos (muestra del año 2015) fueron recolectados en São Paulo — Brasil, en un área universitaria, donde hay algunas fiestas con grupos de estudiantes de 18 a 28 años (promedio).

2. Análisis del problema

El consumo de alcohol es un problema de salud pública que requiere de acciones preventivas inmediatas y de promoción de la salud. Esto al considerarse un factor determinante para algunos trastornos neuropsiquiátricos y de enfermedades no transmisibles como las afecciones cardiovasculares, cirrosis hepática y diversos tipos de cánceres. En la actualidad el consumo de alcohol es considerado una práctica socialmente aceptada, y se le reconoce como vehículo de socialización en diversos grupos sociales como en los adolescentes; el alcohol es la droga legal de inicio y su consumo incrementa el riesgo de involucrarse con otro tipo de sustancias ilícitas como la marihuana, la cocaína, entre otros. Para comprender el problema del consumo de alcohol en la población en general es importante partir de aspectos básicos como lo son las definiciones y el panorama general de esta problemática en el mundo entero, además de conocer las consecuencias de su consumo y la función de enfermería en la prevención. El consumo de alcohol puede describirse en términos de gramos de alcohol consumido o por el contenido alcohólico de las distintas bebidas

La siguiente investigación tiene como objetivo general analizar el comportamiento del consumidor de cerveza en Brasil enfocándonos en el año 2015.

3. Análisis de datos

La data de nuestro código rondará entre los datos de temperatura media, temperatura mínima, temperatura máxima, precipitado y consumo en litros de la cerveza
del país sudamericano de Brasil.

Podemos observar que en todo el año 2015 el consumo en litros de cerveza es de 10 millones de litros aproximadamente.

Las temperaturas más altas que tiene el país (entre 30-40°C) son en los meses de diciembre a enero, estas son las fechas del verano en dicho país, el consumo en litros de cerveza es el más alto; agregándole que son las fechas en donde los jóvenes están más libres y tienen más tiempo para su propia diversión. Agregándole que en dichos meses se celebra el carnaval más famoso a nivel mundial, estos meses sobrepasan a los meses festivos como año nuevo y navidad, en consumo de esta bebida; lo cual tiene sentido, ya que llegan una elevada cantidad de turistas de todo el mundo a celebrar el mejor carnaval del mundo, agregándole lo antes ya mencionado que son los meses de vacaciones y la temperatura del país es la más alta, como se muestra en la Figura 1 y Figura 2.

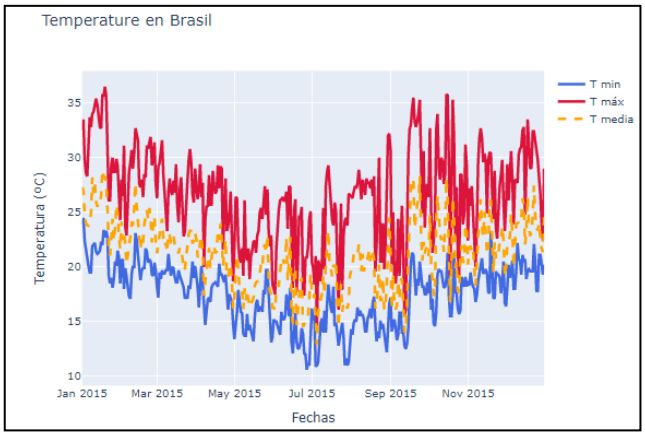

Figura. 1 Gráfico de Temperatura (°C) vs Fecha.

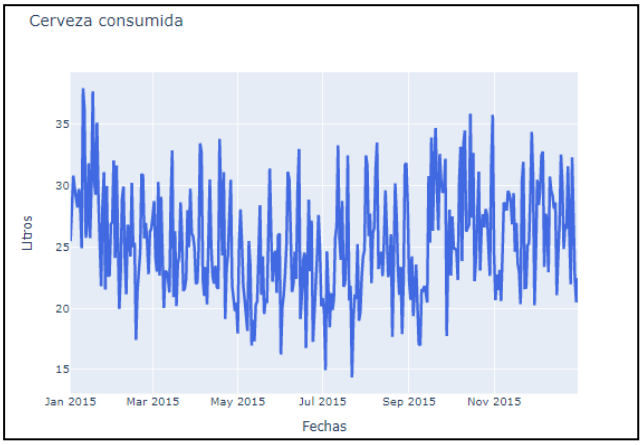

Figura. 2 Gráfico Litros vs Fechas.

3. Análisis de datos

```
import numpy as np
import pandas as pd
import plotly.express as px
import plotly.graph_objects as go
%matplotlib inline
```

El código empieza con la importación de algunas de las ya conocidas librerías como numpy y pandas, y también de unas partes integradas de la librería ploty como lo son ploty.express y plotly.graph_objects que usaremos para generar los objetos gráficos.
La línea nº5 “%matplotlib inline” servirá para que las imágenes de las gráficas permanezcan estáticas.

```
def regresion_lineal(x, y):
    n = len(x)
    x_mean = x.mean()
    y_mean = y.mean()
    B1_numerador = sum(((x-x_mean) * (y-y_mean)))
    B1_denominador = sum((x-x_mean)**2)
    B1_hat = B1_numerador/B1_denominador
    B0_hat = y_mean - B1_hat * x_mean
    return B0_hat, B1_hat
```

Las siguientes líneas de código son para crear una función que resuelva una equivalencia al modelo de regresión final que ya todos conocemos (Ver figura 3). Más adelante se hará llamado de esta función.

$$\begin{cases} y = a + bx & \Rightarrow \quad x = \dfrac{y-a}{b} \\ Encontrar \quad b & \Rightarrow \quad b = \dfrac{(cantidad * xy) - (x * y)}{(cantidad * xx) - (x * x)} \\ Encontrar \quad a & \Rightarrow \quad a = \dfrac{y - (b * x)}{cantidad} \end{cases}$$

```
def estadistica(datos):  #Función para el análisis estadístico
    return pd.Series([datos.min(), datos.max(), datos.mean(), datos.var(),
            datos.std(), datos.mad()], index =['Valor minimo', 'Valor Maximo', 'Valor promedio',
            'Varianza', 'Desviación estandar', 'Desviacion absoluta media'])
```

En la siguiente parte del código haremos otra función pero esta vez será para poder realizar el análisis estadístico, aquí utilizaremos ‘pd.Series’ para convertir nuestras listas en series y poder mostrar los datos de una forma visualmente entendible, aquí la secuencia de índices serán los nombres de los parámetros que queremos hallar y la lista serán las operaciones ya resueltas de dichos parámetros.
Los parámetros que consideramos evaluar son:

- Valor mínimo
- Valor Máximo
- Valor Varianza
- Desviación estándar
- Desviación absoluta media

```
ruta_archivo = '/Consumo_cerveja.csv'
df =  pd.read_csv(ruta_archivo,  decimal=",")
print("El dataset contiene {} filas y {} columnas \n".format(df.shape[0], df.shape[1]))
```

En esta parte del código se hace llamado al archivo del cual son los datos con los que trabajaremos. Deberemos importar el archivo desde google colab y con la ruta de acceso del archivo podremos referirnos a los datos de Excel con una variable definida, en nuestro caso será “df”.
La última línea de la parte 4 del código es para hacer un conteo del total de los espacios utilizados en el archivo Excel, para contar las filas y columnas utilizadas se hace uso de **“df.shape”** que devuelve la tupla de forma (filas, columnas) de un archivo. Y para mostrar al usuario estos valores se hará uso de **“.format”**, así los valores ocuparan los lugares entre las llaves.

```
df.columns = ['Fecha', 'Temperatura_media', 'Temperatura_min', 'Temperatura_max', 'Precipitacion', 'Fin_semana',
              'Litros_consumidos']
print('la cantidad de registros nulos para cada columna es \n' + str(df.isnull().sum()) )
print('\n')
print('Se borrarán registros nulos ... \n')

df =  df.dropna()
print('Registros nulos borrados  \n')
```

Aquí lo que se busca es cambiar los nombres del encabezado (1era fila) que tenía nuestro archivo Excel, ya que los datos que hemos elegido fueron realizados en Brasil y por consecuencia, estos están en idioma portugués. Después se cuenta los valores faltantes de las columnas, para esto haremos uso de **“df.isnull().sum()”** y para “eliminarlos” y no tener problemas en las operaciones haremos uso de **“df.dropna()”** que elimina los valores nulos del Dataframe.

```
for i in range(df.shape[1]):
    print(str(i) + ' ' + str(df.columns[i]))

print('\n Elija una columna del 1 al {} para realizar el análisis estadístico'.format(str(df.shape[1]-1)))

num = int(input('Ingrese el número seleccionado \n'))

print('La columna seleccionada para el análisis es {} \n'.format(df.columns[num]))

#print('Los datos estádisticos de la columna seleccionada se muestrana a continuación \n')

columna_seleccionada = df[df.columns[num]]
analisis = pd.DataFrame(estadistica(columna_seleccionada))

print('Resultados del análisis \n')
print(analisis)
```

En esta parte del código empezamos con un for que analizara “df.shape [1]” que como lo vimos anteriormente se encarga de contar las columnas utilizadas, y los print que les continúan serán para hacer uso de los nombres que cambiamos para el encabezado, de esta manera se lograra tener cada nombre del encabezado al lado de su índice correspondiente para que el usuario puede elegir que encabezado desea mediante un número.

```
fig = go.Figure()
fig.add_trace(go.Scatter(x=df['Fecha'], y=df['Temperatura_min'], name='T min',
                         line=dict(color='blue', width=3)))
fig.add_trace(go.Scatter(x=df['Fecha'], y=df['Temperatura_max'], name = 'T máx',
                         line=dict(color='red', width=3)))
fig.add_trace(go.Scatter(x=df['Fecha'], y=df['Temperatura_media'], name='T media',
                         line=dict(color='orange', width=3,
                             dash='dash')))
fig.update_layout(title='Temperature en Brasil',
                  xaxis_title='Fechas',
                  yaxis_title='Temperatura (ºC)')
fig.show()


fig_1 = go.Figure()
fig_1.add_trace(go.Scatter(x=df['Fecha'], y=df['Litros_consumidos'], name='Litros de cerveza consumidos',
                         line=dict(color='royalblue', width=3)))
fig_1.update_layout(title='Cerveza consumida',
                  xaxis_title='Fechas',
                  yaxis_title='Litros')
fig_1.show()


fig_2 = go.Figure()
fig_2.add_trace(go.Scatter(x=df['Fecha'], y=df['Precipitacion'], name='Litros de cerveza consumidos',
                         line=dict(color='orange', width=3)))
```

```
fig_2.update_layout(title='Cantidad de lluvia por día',
                    xaxis_title='Fechas',
                    yaxis_title='Cantidad de lluvia (mm)')
fig_2.show()
```

Con esta parte del código fueron generados los gráficos. Aquí primero llamamos al módulo “go” de la librería “plotly.graph_objects” para crear un objeto de figura. Luego agregamos trazos a la figura con “fig.add_trace (go.Scatter ()), con esto agregamos valores a los ejes (ordenadas y abscisas), definición de la leyenda, color de las gráficas y ancho de estas. Para ponerle títulos a las gráficas y a los ejes se hizo uso de “.update_layout(). Finalmente las gráficas se mostraran con fig.show ().

```
x = df['Temperatura_media']
y = pd.to_numeric(df['Litros_consumidos'])

[b, m] = regresion_lineal(x, y)

print('La ecuación de la línea recta ajustada es  y = {}x + {}'.format(m,b))
```

La parte final del código será para el análisis de regresión lineal a la temperatura media (x) Vs Litros consumidos de cerveza (y). Aquí se hace llamado de la función que hicimos al principio "regresion_lineal(x,y)" en la que B ordenada al origen, m pendiente; serán los parámetros de línea.

Ojo: pandas.to_numeric () es una de las funciones generales en Pandas que se utiliza para convertir un argumento en un tipo numérico.

4. Conclusiones

- A partir del *dataset* escogido sobre el impacto que tienen ciertas variables en el consumo de cerveza, para la población de 18 a 28 años en un área universitaria brasileña, hemos podido hacer un análisis estadístico, usando distintas librerías de Python, que nos han permitido hallar valores mínimos, máximos, promedios, varianza, desviación estándar y desviación absoluta media, así como representar en gráficos nuestros datos para un mejor análisis.
- Podemos observar en la gráfica de Temperatura (°C), donde distinguimos temperatura mínima, máxima y media, que coindice que, a lo largo del año 2015, en los meses donde hubo mayor temperatura, aumentó el consumo de cerveza, ocurría de manera contraria cuando disminuía la temperatura.
- Se muestra además, en la gráfica Cantidad de lluvia (mm) por día, que en los meses de menor cantidad de lluvia (mayo-septiembre) hay una ligera disminución del consumo de cerveza. Los meses de mayor consumo de cerveza, no coinciden con los de mayor precipitación como muestran las gráficas.
- Donde se evidencia una influencia muy clara en la cantidad de cerveza ingerida por la muestra de personas analizada, es en la variable fin de semana. En la gráfica Litros vs Fechas, se observan picos, los cuales representan en su mayoría a los días domingos del año 2015, siendo los días martes donde se presentan la mayoría de picos bajos. La gráfica además nos muestra que los meses con mayor consumo de cerveza son en enero y octubre.

Referencias

[1] (s.f). "Consumo de Cerveza – Sao Paulo" [online]. Disponible en: https://www.kaggle.com/jhous92/consumo-cerveja/data

[2] Guzman, R. (2019) "Utilizando El Tipo Series de Python Pandas". Disponible en: https://relguzman.blogspot.com/2015/06/usando-el-tipo-series-de-python-pandas.html

[3] Covantec. (s,f) "Programación en Python. Nivel Básico" , https://entrenamiento-python-basico.readthedocs.io/es/latest/leccion5/funciones_integradas.html#:~:text=format()%C2%B6,tipo%20de%20cadena%20de%20caracteres

[4] Acervo Lima (s.f). "Python/Pandas df.size, df.shape y df.ndim" Disponible en: https://es.acervolima.com/python-pandas-df-size-df-shape-y-df-ndim/

[5] QA Strack. (2019). "Cómo contar los valores NaN en una columna pandas en DataFrame". Disponible en: https://qastack.mx/programming/26266362/how-to-count-the-nan-values-in-a-column-in-pandas-dataframe#:~:text=2-

[6] Programador Click. (s.f). "Uso de Python para realizar de visualización de datos del banco de motores". Disponible en: https://programmerclick.com/article/84961594801/