Tecnológico de Monterrey | Educación Continua

THE LEARNING GATE

Visualización de datos con Python

Explotar las ventajas que tiene Python y las plataformas de visualización Matplotlib, Pandas y Seaborn, para generar diversos tipos de gráficos que permitan analizar e interpretar resúmenes de datos del mundo real

Dra. Grettel Barceló Alonso
Diseñadora de la competencia

2

Tecnológico de Monterrey | Educación Continua

THE LEARNING GATE

Lecciones

2 Plataformas de visualización en Python
Reconoce el panorama de visualización de Python y algunas de las plataformas más empleadas para seleccionar la más adecuada a su contexto de aplicación

3 Estructura de los datos y tipos de gráficos
Resume la información usando representaciones gráficas acordes a la estructura subyacente de los datos para darles significado e identificar tendencias

4 Gráficas para exploración de datos
Selecciona las gráficas adecuadas de exploración de datos para reconocer la distribución, dispersión y correlación de las variables y puntos de datos a visualizar

5 Gráficas, ejes y figuras
Crea subgráficas y gráficas superpuestas que permitan hacer análisis comparativos y de variabilidad

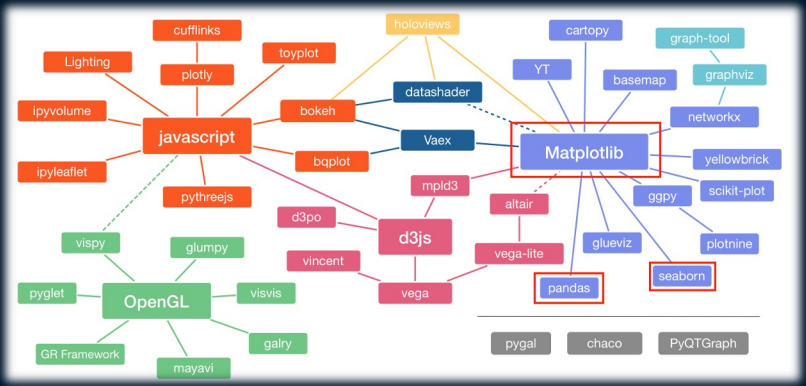
6 Anotaciones en las gráficas
Integra texto, apuntadores y datos tabulares en las gráficas para obtener representaciones más descriptivas

3

Tecnológico de Monterrey | Educación Continua

THE LEARNING GATE

Plataformas de visualización



https://pyviz.org/overviews/index.html

4

Tecnológico de Monterrey | Educación Continua

THE LEARNING GATE

PLATAFORMA	¿CUÁNDO USARLA?
pandas	Si se desea comprender la estructura de los datos, de manera ágil, más que crear gráficos atractivos o con calidad de publicación.
matplotlib	Si se tienen muchos elementos que personalizar de un gráfico, por ejemplo: crear subgráficas, hacer anotaciones textuales o incluir tablas.
seaborn	Si va a crear gráficos atractivos, usando diversidad de paletas de colores y estilos predeterminados, visualmente atrayentes y modernos.

La realidad es que la mayoría de las veces necesitamos combinar:

Seaborn: `sns.countplot(x=countries['Continent'])`
Matplotlib: `plt.xticks(rotation=90)`

5

Tecnológico de Monterrey | Educación Continua

THE LEARNING GATE

	Matplotlib plt	Pandas pd	Seaborn sns
Líneas	`plot()`	`plot()`	`lineplot()`
Barras	`bar()`	`plot(kind = 'bar')` `plot.bar()`	`barplot()`
Circular	`pie()`	`plot(kind = 'pie')` `plot.pie()`	-
Histograma	`hist()`	`plot(kind='hist')` `plot.hist()`	`histplot()`
Caja y bigote	`boxplot()`	`plot(kind='box')` `plot.box()`	`boxplot()`
Dispersión	`scatter()`	`plot(kind='scatter')` `plot.scatter()`	`scatterplot()`

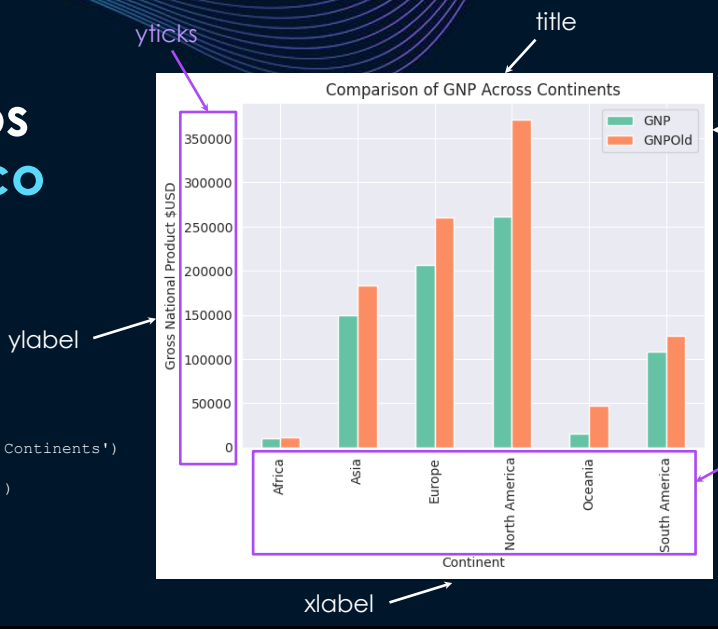
6

Tecnológico de Monterrey | Educación Continua

THE LEARNING GATE

Elementos del gráfico

```
plt.xlabel('Continent')
plt.title('Comparison of GNP Across Continents')
plt.legend(['GNP', 'GNP Old'])
plt.yticks(np.arange(0,400000,50000))
```



7

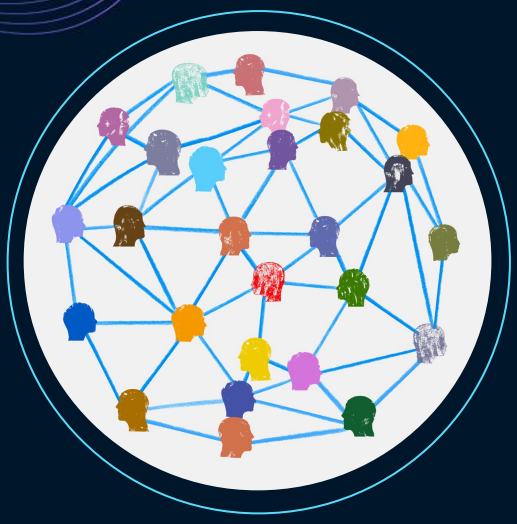
Tecnológico de Monterrey | Educación Continua

THE LEARNING GATE

Integrar conexión

Ingresa a app.gosoapbox.com con el event code que se proporcione en el chat

Duración: 10 minutos



8

Tecnológico de Monterrey | Educación Continua

THE LEARNING GATE

Reflexiones generales

- Familiarízate siempre con tus datos, conoce su estructura y contenido
- Elige la plataforma de visualización que se ajuste a tus necesidades (Pandas es útil para gráficos simples y Seaborn admite perspectivas más complejas, pero en ambos casos debes considerar adentrarte en la complejidad de Matplotlib para personalizar)
- Utiliza diversos tipos de gráficos para explorar tus datos
- Intenta identificar patrones y experimentar con diversos escenarios

9

Tecnológico de Monterrey | Educación Continua

THE LEARNING GATE

Reto
Índice de felicidad

10

Tecnológico de Monterrey | Educación Continua

THE LEARNING GATE

El informe mundial sobre la felicidad es una encuesta que clasifica a 156 países por sus niveles de bienestar, tomando en cuenta 6 factores:

1. Producción económica
2. Apoyo social
3. Esperanza de vida
4. Libertad
5. Ausencia de corrupción y
6. Generosidad

Se toma el estudio del 2019, recuperado del portal kaggle.com, para explorar los datos y puntuación por país o región y obtener gráficas relevantes

11

Tecnológico de Monterrey | Educación Continua

THE LEARNING GATE

La felicidad (Score) se mide en escala de 0 al 10, donde el 10 es lo mejor y el 0 lo peor

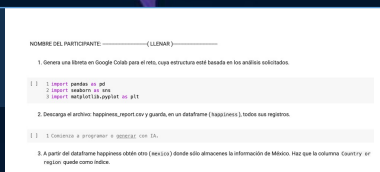
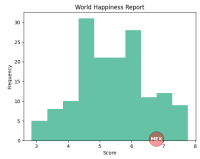
Las siguientes columnas representan la medida en que los seis factores contribuyen a evaluar la felicidad en cada país.

	Overall rank	Country or region	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
0	1	Finland	7.769	1.340	1.587	0.986	0.596	0.153	0.393
1	2	Denmark	7.600	1.383	1.573	0.996	0.592	0.252	0.410
2	3	Norway	7.554	1.488	1.582	1.028	0.603	0.271	0.341
3	4	Iceland	7.494	1.380	1.624	1.026	0.591	0.354	0.118
4	5	Netherlands	7.488	1.396	1.522	0.999	0.557	0.322	0.298
5	6	Switzerland	7.480	1.452	1.526	1.052	0.572	0.263	0.343
6	7	Sweden	7.343	1.387	1.487	1.009	0.574	0.267	0.373
7	8	New Zealand	7.307	1.303	1.557	1.026	0.585	0.330	0.380
8	9	Canada	7.278	1.365	1.505	1.039	0.584	0.285	0.308
9	10	Austria	7.246	1.376	1.475	1.016	0.532	0.244	0.226

12

Tecnológico de Monterrey | Educación Continua

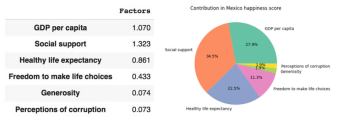
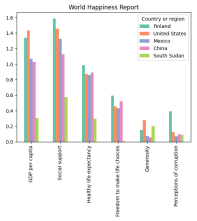
THE LEARNING GATE

PREGUNTA		TIP
1. Genera una libreta en Google Colab para el reto, cuya estructura esté basada en los análisis solicitados.		Utiliza celdas de texto para las preguntas y de código para mostrar los resulltados (Compartir plantilla)
2. Descarga el archivo: happiness_report.csv y guarda, en un dataframe (happiness), todos sus registros.		Sólo hay que leer el archivo una vez
3. A partir del dataframe happiness obtén otro (mexico) donde sólo almacenes la información de México. Haz que la columna Country or region quede como índice.		Este dataframe lo podrás utilizar en las siguientes preguntas
4. Obtén un histograma del puntaje (score) para identificar el rango o clase más frecuente. Ubica el valor de México con una etiqueta de texto.		No es necesario volver a filtrar el dataframe happiness, puedes usar el de mexico obtenido en la pregunta 3 Utiliza el parámetro bbox de la función text() para replicar el círculo rojo

13

Tecnológico de Monterrey | Educación Continua

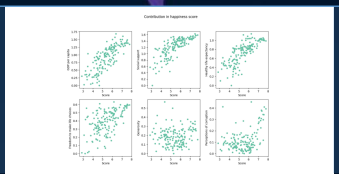
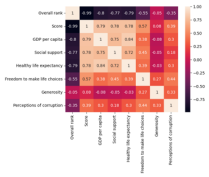
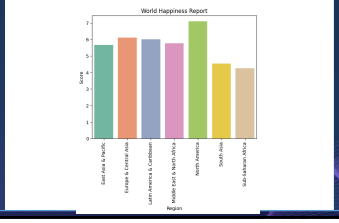
THE LEARNING GATE

PREGUNTA		TIP
5. Construye un box plot para la columna esperanza de vida (Healthy life expectancy). Ubica el valor de México con una anotación.		No es necesario volver a filtrar el dataframe happiness, puedes usar el de mexico obtenido en la pregunta 3 Debes usar una anotación, NO un texto
6. Crea un gráfico circular para analizar en qué medida los factores contribuyen a evaluar la felicidad en México. Para ello, deberás modificar la estructura del dataframe mexico obtenido anteriormente.		Hacer el cambio en de estructura en el dataframe mexico, hará que la graficación sea mucho más sencilla.
7. Filtra el dataframe para quedarte con 5 países (el más feliz, el menos feliz, México y dos más de tu interés) y visualiza en una misma gráfica los 6 factores.		La función isin() permite comparar varios valores en una única instrucción columna.isin(['valor1','valor2',...,'valorN']) en lugar de utilizar condiciones individuales (columna == valor1) | (columna == valor2) | ...(columna == valorN)

14

Tecnológico de Monterrey | Educación Continua

THE LEARNING GATE

PREGUNTA		TIP
8. Crea una matriz de subgráficas de 2 x 3 con scatter plots del puntaje (score) versus los 6 factores para determinar qué factor influye más en la evaluación.		Para crear matrices de subgráficos, sólo se puede usar Matplotlib, pero una vez creada, puedes dibujar los scatter con cualquier plataforma de visualilzación No olvides incluir el título para toda la figura. Utiliza el contenedor (fig)
9. Comprueba lo anterior con un heatmap donde incluyas los índices de correlación.		Para aplicar la función corr() sólo a las columnas numéricas, incluye el parámetro numeric_only = True Puedes redondear los valores con round()
10. Combina con el dataframe metada (Metadata.csv) para graficar la felicidad promedio por región.		Para conjuntar los dataframes, identifica las columnas que poseen los valores comunes. Asegúrate de aplicar la función de agregado correcta.

15

Tecnológico de Monterrey | Educación Continua

THE LEARNING GATE

Gracias.

16