**Análisis de Datos (Multivariante): Todo numéricas**

Supongamos que por el desarrollo y necesidades del análisis de datos tenemos que cotejar varias variables numéricas a la vez. Además de las posibles visualizaciones que pueda recomendarte el interesante [sitio](https://www.data-to-viz.com/) que ya hemos mencionado otras veces, yo te propongo como punto de partida (luego tendrás que completarlo tú) lo siguiente que viene a continuación en el notebook.

**import** numpy **as** np

**import** matplotlib.pyplot **as** plt

**import** pandas **as** pd

**import** seaborn **as** sns

**from** bootcampviztools **import** plot\_categorical\_numerical\_relationship, plot\_combined\_graphs,

pinta\_distribucion\_categoricas, plot\_grouped\_boxplots, plot\_categorical\_relationship\_fin, plot\_grouped\_histograms, grafico\_dispersion\_con\_correlacion, bubble\_plot

df\_seguros **=** pd.read\_csv("./data/Marketing-Customer-Analysis.csv")

df\_air\_jun **=** pd.read\_csv("./data/dataset\_viajes\_jun23.csv")

**Caso 2. Viajes: Ingresos-Consumo\_kg-Distancia**

Lo primero sería mostrar los valores y diagramas de cada uno, así como las correlaciones dos a dos:

plot\_combined\_graphs(df\_air\_jun,["ingresos","consumo\_kg","distancia"])

(3, 2)

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

grafico\_dispersion\_con\_correlacion(df\_air\_jun,"consumo\_kg", "ingresos", mostrar\_correlacion**=** **True**, tamano\_puntos**=** 30)

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

grafico\_dispersion\_con\_correlacion(df\_air\_jun,"distancia","ingresos", mostrar\_correlacion**=** **True**, tamano\_puntos**=** 30)

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

grafico\_dispersion\_con\_correlacion(df\_air\_jun,"distancia","consumo\_kg", mostrar\_correlacion**=** **True**, tamano\_puntos**=** 30)

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Lo siguiente será mostrar los tres valores a la vez, con un diagrama de burbujas que es como un scatter plot (un diagrama de dispersión) pero en el que el tamaño de los puntos los marca la tercera variable:

bubble\_plot(df\_air\_jun,"distancia","consumo\_kg","ingresos", scale **=** 100000)

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

El tamaño de las burbujas depende de los ingresos, se ve la correlación entre distancia y consumo, aunque claramente hay una línea de vuelos mucho más eficientes y si nos fijamos en los tamaños de las burbujas de una línea y otra son aparentemente similares. Es decir, parece y esta sería una gran línea de trabajo, un conjunto de viajes que son más eficientes en todos los términos que otros.

Pero ahí no acaba el análisis. El siguiente paso sería escoger una de las tres, hacerle un *binning* o convertirla a categórica y estaríamos en el análisis multivariante explicado en la sesión anterior. Este caso lo veremos en los ejercicios

**Analisis multivariante: Matriz de correlación**

Para terminar, pandas permite obtener la matriz de correlación de sus variables numéricas. Es decir una matriz en la que aparecen las correlaciones (de pearson, si no se indica lo contrario) dos a dos:

corr\_matrix **=** df\_seguros.corr(numeric\_only**=** **True**) *# Si no lo pones y tienes variables tipo texto, fallará7*

corr\_matrix

Tabla

Descripción generada automáticamente

Y aún hay más, con seaborn podemos hacer una representación de esta matriz a través de lo que se denomina un "mapa de calor":

plt.figure(figsize**=**(10, 8)) *# Ya lo veremos pero esto permite ajustar el tamaño de las gráficas*

sns.heatmap(corr\_matrix, annot**=True**, fmt**=**".2f", cmap**=**"coolwarm", cbar**=True**, square**=True**, linewidths**=**.5) *# el cmap es el rango de colores usado para representar "el calor"*

plt.title('Matriz de Correlación')

plt.xticks(rotation**=**45) *# Rota las etiquetas de las x si es necesario*

plt.yticks(rotation**=**45) *# Rota las etiquetas de las y si es necesario*

plt.show()

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Y para terminar también es posible completar tu análisis bivariante/multivariante con una visualización de los diagramas de dispersión (pero ojo que si son muchos no verás nada)

df\_solo\_numeros **=** df\_air\_jun.select\_dtypes(include **=** ["float","int"])

df\_solo\_numeros.columns

Index(['distancia', 'consumo\_kg', 'duracion', 'ingresos'], dtype='object')

sns.pairplot(df\_solo\_numeros)

plt.show()

Gráfico

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

**ANALISIS DE DATOS: (humilde) GUIA DE REFERENCIA**

PARTE I

1. **Establecer objetivos**
2. **Completar la definición de los datos/Rellenar fichas datasets**
3. **Tipificar los datos**
4. **Priorizar variables**

PARTE II  
4. **Aplicar tendencia central a todas las variables**  
5. **Analizar las frecuencias de las variables categóricas**: Gráficos de barras  
6. **Analizar las medidas de posición y los rangos de las variables numéricas**: Graficos de Caja  
7. **Analizar las medidas de dispersión de las variables numéricas**  
8. **Analizar las distribuciones de las variables numéricas**: Histogramas y funciones estimadas de densidad (combinados)  
9. **Analizar las anomalías o outliers**: confrontación de histogramas y gráficos de cajas. Umbrales (2-3 desviaciones estándar, 4.5 en tamaño "bigotes")  
10. **Trata los outliers**: separa en dos, elimina, crea una categoría especial...

PARTE III  
11. **Revisa prioridades, preguntas y notas hasta el momento**  
12. **Preparara el análisis bivariante** en función de la revisión anterior.  
13. **Empieza por categóricas importantes/target**: combínalas con categóricas y numéricas  
14. **Sigue con numéricas importantes/target**  
15. **Analiza las parejas de variables que necesites para contestar tus preguntas**  
15.1 **Es directo así que: Matriz de correlación, heatmap y pairplot**  
16. **Recapitula**: Clasifica ya tus hallazgos en posibles mensajes, elementos a profundizar, cosas interesantes, pero no importantes (descártalas). Añade las preguntas contestadas y la lista de preguntas sin contestar.

PARTE IV  
17. **Intenta resolver todas las preguntas pendientes**  
18. Si necesitas análisis más complicados (3 o más variables): Planifica muy bien que grupos  
19. Aplica lo visto en estas sesiones, tendrás que probar variaciones hasta llegar a encontrar el punto adecuado.  
20. **Acaba el análisis**: Cierras las preguntas pendientes, lo que te quede por trabajar -> Líneas de trabajo futuro.