Módulo 2 : Data Analysis Python

Preprocesamiento de datos

↳ conversión de data cruda para preparar los datos para futuro análisis

- NA
- Data formatting
- Normalización de Datos
- Data binning (categorías)
- Categóricas - numéricos

Missing Values

↳ eliminar data
→ reemplazar → la mejor opción
↓ con promedio
- frecuencia / dejar data como NA

→ En pandas se hace drop.na para eliminar registros

→ replace ⇒ cambia valores con nuevos utilizado para NA

Data formatting

⇒ Estandarizar datos para comprender los mismos
→ Aplicando cálculos para estandariza cada columna

Datos ⇒ Objetos
↳ dtypes()
↳ astype() → cambia formato

Data Normalization

→ Método para comparar variables de distintos valores

→ útil para realizar análisis (regresiones, PCA)

→ Simple feature (todo dividido para el Max)
→ Min-Max → divide la diferencia con el rango
→ z-score → dividir el error con la desviación estandar

Binning (Python)

→ agrupar datos en grupos (bins)
↳ datos numéricos en categóricos

Python ⇒ np.linspace (min (df[]), max (df[], 4)
↓
para crear cuantos grupos

Categórica a numérica

→ Para modelos podemos cambiar texto a números (string) (numerical)

→ Creación de Dummy (1, 0)

→ pd.get_dummies()

Review Questions (Evidencia)