**PUNTOS IMPORTANTES**

* Reconocer EL TIPO DE DATOS, ya sea para el análisis o el modelo predictivo es vital para determinar el tipo de **representación visual.**
* El tipo de datos se relaciona con el rendimiento, determinar su tipo también determina los cálculos para esa variable. ¿Cómo debe procesarlos?
* DataFrame es lo que conocemos como “Marco de Datos”
* Debemos tener cuidado con los datos no estructurados, como los textos ya que estos requieren de procesarse y manipularse para convertirse en un DataFrame.
* Definiciones
  + DataFrame → Estructura básica para modelamiento estadístico y aprendizaje automático.
  + Característica → Columna del DataFrame
  + Resultado → Variable dependiente o variable objetivo
  + Registros → Una fila del DataFrame
  + Variables independientes → Variables Predictoras
  + Variable respuesta → Variable dependiente.
  + Muestra o Sample → Un registro o conjunto de registros.
* Existen estructuras de datos no rectangulares como por ejemplo las series de tiempo.
* También las estructuras de datos en forma de grafos o redes. (Sistemas de recomendación)
* El libro se especializa en datos rectangulares para modelos predictivos.
* Gráficos en estadística =/ Grafos en informática.
* Tener cuidado con la terminología entre los diferentes campos que conforman la ciencia de datos.
* Tendencia Central:
* La media es relativamente sencilla de obtener, pero no siempre significa un buen medidor de tendencia central. Revisemos algunas opciones.
* Los estadísticos “estiman” y los científicos de datos “miden”
* N -Mayúscula → Población, n – minúscula → Muestra
* Lo anterior no aplica tanto en ciencia de datos.
* La Media Truncada omite los valores más extremos del conjunto. En la práctica suele verse a menudo (Ejemplo Jueces Buceo Internacional)
* La Media ponderada sirve para otorgar peso un determinado subconjunto de datos.
* La Mediana solo es posible de obtener ordenando los datos.
* En una importante cantidad de casos la mediana tiene una representación importante.
* Se emplea cuando se conoce la existencia de ciertos valores anómalos que pueden perjudicar a la media.
* ¿Qué son valores atípicos? Son valores extraños en nuestro conjunto de datos, de ser erróneos es posible que hayamos combinado tablas de diferentes unidades o lecturas incorrectas debido a un sensor dañados, aún así es posible que estos datos existan de forma natural.
* En el caso anterior emplear la media sería un suicidio, esto sería deficiente pero la mediana probablemente siga siendo válida. Los valores atípicos entonces deben identificarse y tratarse solo después de investigarse.
* En todo caso usar media truncada, 10% inferior & 10% superior también suele ser un buen indicador.

**Estimación de la Variabilidad:**

* También conocida como dispersión
* Dicho de otro modo, tenemos en cuenta la variabilidad.
* Nos encontramos con:
  + Desviaciones
  + Varianza
  + Desviación estándar
  + Desviación media absoluta
  + Desviación absoluta mediana de la mediana
  + Rango
  + Estadísticos Ordinales
  + Percentil
  + Rango Inter cuartil
* La estimación de variación mas empleada es la desviación.
* Responde a la pregunta de qué tan dispersos están los datos
* Debemos conocer la desviación media absoluta.
* Las mas populares son la varianza y la desviación estándar.
* Los científicos de datos no se deben de preocupar por los grados de libertad
* La varianza y la desviación estándar son especialmente sensibles a los valores atípicos, ya que se basan en las desviaciones al cuadrado.
* Una estimación robusta de la variabilidad es la desviación absoluta mediana de la mediana.
* También es posible calcular la desviación estándar truncada.
* Siempre es recomendable de un conjunto de datos obtener el máximo y mínimo registrado de esta forma tendremos una mejor idea de los datos atípicos.
* El rango no es muy útil como medida de dispersión, a menos que sea exploratoria.
* Cuantil 0.8 es igual a Percentil 0.8
* El rango intercuartílico es igual Percentil 75 – Percentil 25
* Recuerda que para esto debes ordenar los datos y en ocasiones puede resultar costoso.