**EDA Netflix (Paso a Paso)**

**#Datos de Películas (Títulos) y Fechas#**

**Fase Inicial:**

**Primer Paso: Cargamos las librerías Necesarias**

* **Pandas**
* **Numpy**
* **Recordar: No trabajar encima del DataFrame Original**

**Segundo Paso: Cargar el Archivo .CSV de Netflix**

**Tercer Paso: Analizar las Variables que nos aporta Netflix de nuestros Datos}**

**Fase de Análisis Exploratorio - Recomendable:**

**Primer Paso: ¿Cuál es la dimensión de Nuestros Datos?**

* **¿Cuántos Registros están implicados?**
* **¿Cuántas Variables están implicadas?**

**Segundo Paso: ¿Cuál es la Información más resaltante de la tabla?**

* **Títulos de las Variables**
* **¿De las Entradas cuales son nulas?**
* **¿Cuál es el tipo de nuestras variables?**
* **“Recordar que el tipo Object hace referencia que no lo clasifica con precisión”**

**Tercer Paso: Según el tipo de Dato que tengamos, debemos Castear.**

* **Es común que nuestros datos no vengan con el formato correcto de manera que evaluaremos si especificar nuestros tipos de datos es vital o no.**
* **ATENCIÓN Siempre que vemos un dato que pertenece al tipo de dato FECHA es recomendable castearlo.**

**Cuarto Paso: Intentar Visualizar el Contenido con más claridad. [Los Primeros Registros]**

* **En este paso debemos de intentar descubrir patrones, ¿cuál es el formato visual del contenido de nuestras variables?**
* **Ejemplo: ¿Como está la Fecha? Día mes año, mes día año… etc.**
* **También es común que en una misma variable se estén combinando elementos de categorías distintas, quizás podríamos evaluar el separarlas o no.**
* **Por otra parte, también pueden existir patrones en el contenido mismo, por ejemplo, ¿en una serie como separan las temporadas y el nombre de los episodios que visualizamos?**
* **¿Debemos preguntarnos también, esto aplica con las películas?**
* **Extrapolar está información es Vital.**

**Quinto Paso: Hemos Descubierto un Patrón es hora de Comenzar “:”**

* **Creamos una Lista, el objetivo es visualizar en cuantas partes se separan nuestros contenidos con este patrón.**
* **Sep\_lista = df.Title.str.split(pat = “:”, expand = False).to\_frame()**
* **Se nos creará un DataFrame, separando cada componente, sin crear nuevas columnas.**

**Sexto Paso: Debemos Crear una nueva Columna. ¿En cuántas partes se dividieron nuestros títulos?**

* **Para esto del DataFrame Sep\_lista creamos una nueva columna llamada “Numero de Partes”, la cual tendrá un valor igual a la función largo aplicado a ese misma DataFrame Sep\_Lista.**
* **De Esta forma obtenemos unas columnas de vital importancia**

**Séptimo Paso: Analizamos “La nueva Columna”**

* **En este momento sería ideal preguntarnos, ¿cuál es el máximo número de partes que se podrían presentar?**
* **¿Cuál sería el menor número de partes presentado?**
* **¿Hay alguno que se repita más que el resto?**
* **Evaluamos ello, con la función contando los valores.**
* **Debemos obtener: Los más frecuentes, los Menos Frecuentes, El Rango de partes que puedan aparecer.**

**Octavo Paso: Obtenemos Características de dichos valores**

* **¿Que pasarán con aquellos títulos de solo 1 parte? Haz una muestra de 10 elementos.**
* **¿Qué pasan ahora con aquellos títulos de 2 partes? ¿Qué son? ¿Películas o Series?**
* **¿Los que tienen 3 partes? ¿Seguro son Series o no?**
* **¿¿¿Y las de 4 Y las de 5???**

**Noveno Paso: Cosechar los Frutos del Análisis Exploratorio – ¡Crearemos más Variables!**

* **Podemos Hacer una diferenciación fenomenal ahora que conocemos las características de una serie vs una película o documental.**
* **Esto lo determinaremos por las partes que tienen cada uno.**

**Fase de Implementación de Variables Derivadas o Adicionales:**

**Paso1:**

**-Importante!! Aplicaremos Numpy**

**-Perfecto, entonces nos ponemos en marcha creando una nueva columna llamada tipo, esa columna será igual a un método de Numpy asombroso “np.where”**

**-Entonces dentro, llamaremos a nuestro DataFrame[numero\_de\_partes], si este es menor que 3, será película, sino será una serie.**

**-Así tendríamos nuestro DataFrame, pero sigue sin ser el principal.**

**Paso 2:**

* **Creamos una copia de nuestro DataFrame original y este lo concatenamos con la nueva columna que nos hemos tomado la molestia de obtener.**

**Paso 3:**

* **En esta ocasión si nos tomaremos la molestia de obtener los componentes de los títulos: Sep\_lista = df.Title.str.split(pat = “:”, expand = True).to\_frame()**
* **En el primer nivel es obvio que se trata del nombre de la Serie, en el según es común que se registren las temporadas, mientras que, en el tercero con algunas fallas, el nombre del episodio, sin embargo, en cuanto a niveles finales, existen problemas que deberemos resolver.**
* **Renombramos el DataFrame, las columnas para que correspondan con los niveles.**
* **Todo esto lo concatenamos a nuestra tabla principal.**

**Paso 4:**

* **Manipulamos la columna de las fechas, para convertirlas en el formato adecuado, entonces, creamos una nueva columna, haciendo el cambio,**
* **Dropeando, por supuesto la columna original, y guardando los cambios.**
* **Es mucho más enriquecedor trabajar con fechas.**
* **Una de ellas es usarlas como índices.**