Módulo 2 - Desafío 2

Introducción

En este módulo aprenderemos acerca del “arte” de plantear un problema correctamente, trabajamos con pivot tables y empezaremos a jugar con Pandas. Además, incorporaremos algunos conocimientos y técnicas para el cleaning de un dataset y la generación de variables dummy. Van a ver un dataset bastante desordenado y sucio. La idea es “limpiarlo”, ordenarlo y realizar un análisis exploratorios.

Este desafío tiene una enseñanza importante: entre el 70% y el 80% de la tarea del data scientist está en la limpieza, el ordenamiento… en fin, el grueso del tiempo lo dedicamos en el proceso que prepara los datos para el análisis. En Data Mining suelen llamar a esta etapa del proceso EDA (Exploratory Data Analysis). Los modelos no hacen magia: si los datos no son buenos, los modelos tampoco. Un primer paso es tener datos “ordenados”.

Resumen del proyecto

Un programa de radio los contrata para realizar un análisis acerca de la música de la década del 2000. La idea es poder armar una nota que hable sobre el tema y para eso decidieron realizar una aproximación “data-driven” al problema. Les interesa mucho tratar de entender **cuáles son los motivos por los que una determinada canción llega a los primeros puestos**. Ustedes deben generar tanto el material bruto como el análisis que sustentará la nota del programa.

Luego de hurgar bastante en la web deciden quedarse con los rankings de Billboard y generaron un dataset para abordar los problemas planteados. Para ello, **deberán transformar la pregunta difusa “¿qué es lo que hace a un hit?” a un problema formulado de forma clara y que sea abordable con los datos disponible.**

Objetivos y Requisitos

El trabajo debe:

- formular un problema de forma clara, correcta y relevante
- identificar las ventajas y limitaciones de lo datos disponibles para la resolución de dicho problema
- importar el dataset usando la librería Pandas
- realizar un análisis exploratorio del dataset
- aplicar técnicas de limpieza, ordenado y preprocesamiento
- generar visualizaciones usando los módulos de ploteo de Python
- identificar correlaciones / relaciones causales en los datos
- testear hipótesis usando análisis estadístico

Puntos extra

- escribir un blog o white paper breve que contenga lo siguiente (incluí el link en la notebook):
    - una descripción de la filosofía de limpieza de los datos
    - una descripción más detallada de las técnicas de limpieza usadas en el reporte

Material a entregar

- Un notebook con el código que genera los estadísticos y los gráficos debidamente comentado, explicando los pasos que se siguieron en el análisis
- Una presentación (no técnica) de 5 slides. Recordar entregarla en formato PDF. Deberá exponerse en 5 minutos el día de la presentación.

Fecha de entrega

- El material deberá entregarse en la Clase 15 (.

Código básico

El código básico fue diseñado en formato de notebook jupyter. Por favor, completar el proyecto en este notebook.

Dataset

- billboard.csv

¿Cómo empezar? Sugerencias

- cargá los datos
- corré algunos comandos básicos de numpy para empezar a describir los datos
- escribí un pseudocódigo antes de empezar a codear. Suele ser muy útil para darle un esquema y una lógica generales al análisis
- leé la documentación de cualquier tecnología o herramienta de análisis que uses. A veces no hay tutoriales para todo y los documentos y las ayudas son fundamentales para entender el funcionamiento de las herramientas utilizadas
- documentar todos los pasos, transformaciones, comandos y análisis que realices.

Recursos útiles

- CSVKit: una herramienta para limpiar datos CSV en la línea de comando
- ¿Qué es un white paper?

Evaluación

Los profesores usarán la siguiente escala para calificar tu trabajo y las habilidades técnicas adquiridas en módulo:

Puntaje	Descripción
0	Incompleto
1	No cumple con las expectativas
2	Cumple con las expectativas. Buen trabajo!
3	Excede con creces las expectativas!