Análisis de series temporales

Práctico de Análisis y Curación de datos

En este práctico se propone continuar analizando el dataset más en detalle y tomar acciones de limpieza y curación sobre los datos cuando sea necesario.

Instalación

Recomiendo instalar un ambiente virtual con las librerías estándar para procesar datos y visualizarlos, para evitar contratiempos.
Las siguientes librerías son más que suficiente para cumplir con los objetivos de este práctico: **jupyterlab, numpy**, **pandas**, **matplotlib**, **scikit-learn**.

Informe

Se deberá realizar un informe en formato jupyter notebook donde se presente la información solicitada, y se explique las decisiones tomadas a lo largo del análisis y los procesos de curación. Se espera que el mismo notebook donde se realizan las acciones, tenga un formato de informe con títulos y texto en markdown en cada sección, como se muestra en el informe de ejemplo.
Una vez terminado el informe se deberá **exportar a HTML** para evitar problemas de lectura y facilitar la corrección.
El informe se puede organizar a su mejor criterio, sin omitir la información solicitada.
Se espera que se realicen gráficos cada vez que se pueda, aunque no se lo pida de forma explícita.

Implementación

Como primer paso es necesario verificar la consistencia de la información.
Para esto debemos verificar al menos lo siguiente:

- ¿Los ids son únicos?
- Si no tuviéramos estos índices, ¿tenemos información para construir una clave primaria?
- ¿Tenemos datos faltantes? Dar detalles.
- ¿Tenemos datos inconsistentes o raros? Dar detalles.
- ¿Tenemos outliers muy lejanos? ¿Conviene quitarlos del dataset?
- ¿Las fechas tienen sentido? Dar detalles.
- ¿Que otras verificaciones básicas podrías hacer?

¿Como se podría imputar las fechas faltantes de la columna **date_sent**? Justificarlo e implementar alguna solución.
¿Qué riesgos existen al imputar datos? ¿Qué riesgos existen al imputar estos en particular?

¿Cómo corregirías las fechas inconsistentes? Implementar alguna solución

¿Es conveniente aplicar normalización o estandarización sobre algunos features? ¿Cuales features? ¿Porqué?
¿Qué técnica utilizarías? Implementar alguna solución.

¿Es necesario reducir la dimensión de los features?
¿Sería útil aplicar el algoritmo de PCA? ¿Sobre qué features? ¿Con qué objetivo?

Actualmente el target tiene granularidad de días, ¿lo podrías refinar? ¿Como? ¿Qué beneficios obtendrías al aumentar la granularidad?

Se espera que utilicen las preguntas como guía para realizar el análisis y las implementaciones correspondientes, todas las respuestas y decisiones tomadas tienen que estar justificadas con datos o gráficos. Se evaluará la legibilidad del notebook, el detalle a la hora de responder las preguntas y mostrar la información solicitada, y además que los gráficos utilizados sean apropiados y correctos. **La información no contemplada en las preguntas que se logre encontrar y presentar correctamente será valorada.**